Antal Péter – Arany Ádám – Bolgár Bence – Gézsi András – Hajós Gergely – Hullám Gábor – Marx Péter – Millinghoffer András – Poppe László – Sárközy Péter

BIOINFORMATIKA: MOLEKULÁRIS MÉRÉSTECHNIKÁTÓL AZ ORVOSI DÖNTÉSTÁMOGATÁSIG

A molekuláris biológiai méréstechnikai fejlődés a nagy adattömegeket, majd a hipotézismentes kutatási paradigma megjelenését hozta el az orvosbiológiába. Az ezredforduló előtti genetikai-genomikai korszakot a posztgenomikai korszak követte egyre szaporódó omikai szintekkel és leíró hálózati megközelítésekkel. Egy évtized után azonban egyre inkább a nagyléptékű adat- és tudásfúzió került a központba. A jegyzet ezen új kihívásokat tekinti át. Az első két fejezet a genetikai méréstechnika alapjait foglalja össze. A genetikai variánsok hatásainak megértését a fehérjék szerkezetének tárgyalása, ill. a génszabályozási hálózatok bemutatása segíti a következő két-két fejezetben. Ezután az alapvető fontosságú statisztikai asszociációs elemzéseket mutatja be. Az értelmezés támogatására összefoglaljuk az oksági következtetés egy Bayes-hálókon alapuló formalizálását, ill. a szövegbányászati módszereket. A kísérletek szekvencialitása mellett az adatok heterogenitása és így integrált elemzése is központi kihívás, amely kihívást még nehezebbé teszik az egyre elérhetőbb „mély”, azaz részleteiben gazdag fenotípus- és környezeti leírások. Az adatmegosztás hatékonysága miatt és a nagy számításigény miatt is egyre fontosabbá válnak az általánosan elérhető, közmű jellegű informatikai szolgáltatások, amelyek működését példákkal is illusztráljuk. Az áttekintést egy gyógyszerkutatási összefoglaló zárja, amelyben a személyre szabott medicina szempontjai is megjelennek, ill. egy metagenomikai összefoglaló, amely az epigenetikai szint megjelenése után korunk egy új ígéretes omikai szintje.

Kulcsszavak: genotipizálás, új generációs szekvenálási módszerek, fehérjemodellezés, génszabályozási hálózatok, omikai hálózatok, dinamikus rendszerek, kísérlettervezés, munkafolyamat-rendszerek, asszociációs elemzések, biomarker-elemzések, adat- és tudásfúzió, oksági következtetés, orvosi döntéstámogató rendszerek, nagy adatttömegek, szemantikus publikálás, hasonlósági alapú gyógyszerkutatás, metagenomika.

Budapesti Műszaki és Gazdaságtudományi Egyetem és **Semmelweis Egyetem**

Typotex Kiadó
2014

Szakmai lektorok: Molnár Viktor, Antos András

ISBN 978 963 279 180 7

Készült a Typotex Kiadó gondozásában
Felelős vezető: Votisky Zsuzsa

Készült a TÁMOP-4.1.2/A/1-11/1-2011-0079 számú, „Konzorcium a biotechnológia aktív tanulásáért” című projekt keretében.

Tartalomjegyzék

1. fejezet

DNS rekombináns méréstechnológiák, zaj- és hibamodellek

A DNS méréstechnológiák az ezredforduló után rendkívüli sebességgel fejlődtek, de a klinikai gyakorlatba történő beágyazás és az eredmények feldolgozása és értelmezése nem követte a méréstechnika fejlődését. Áttekintjük a DNS szekvenálás módszereinek fejlődését, és a mérések feldolgozásának menetét, annak különböző aspektusait, valamint zaj- és hibakarakterisztikáját. Már a mérések tervezése és előkészítése is speciális szakértelmet kíván, valamint az adatok feldolgozása során ügyelni kell a megfelelő eszközválasztásra, és a mérések megbízhatóságára. Az új generációs DNS szekvenálási eljárások klinikai elterjedésének előfeltétele a pontos, megismételhető mérési munkafolyamat kidolgozása.

1.1. Történelmi áttekintés

A Humán Genom Projekt 1990-ben kezdődött, és 2003-ban fejeződött be. A projekt eredményeként megfejtették a teljes emberi genom szekvenciáját. A projekt kezdetén akkora jelentőséget tulajdonítottak az emberi genom megismerésének az orvostudomány fejlődésére nézve, mint annak idején az anatómiának. A teljes szekvenciához a Sanger-szekvenálás fokozott párhuzamosításának felhasználásával jutottak. A 2000-es évek elején egy teljes emberi genom szekvenálásának törekvéseivel párhuzamosan hívták életre a HapMap projektet, az emberi genetikai variánsok különböző populációkban történő feltérképezését célozva meg. A HapMap projekt abból a feltételezésből indul ki, hogy a gyakori betegségek esetén – különösen azok, amelyek gyermeknemzési kor után jelentkeznek – nagy valószínűséggel közös variánsok azonosíthatók a genomban.

A több mint 3 Gb méretű humán genom szekvenálása jelentős előkészületeket igényelt. A kromoszómákat megközelítőleg 50.000–200.000 bázispár közötti hosszra tördelték. Ezeket a hosszú fragmenseket baktériumokba ültették, hogy ilyen módon a baktérium DNS replikációs mechanizmusát felhasználva másolatokat készíthessenek. Ezeket a másolatokat tartalmazó klónokat egyenként izolálták és a leolvasásra szánt szakaszokat elkülönítették a bakteriális DNS-től, majd néhány száz bázispár hosszú részekre törték, végül Sanger-

szekvenálással olvasták le. A leolvasott szakaszokból (readekből) állították össze a végső, eredeti genomot. Ezt a módszert „hierarchikus shotgun" módszernek hívjuk.

A számítási kapacitás kezdetben nem volt elégséges nagy genomok összerakásához, különösen a 3 milliárd bázispár hosszú emberi genom random shotgun readek esetében. Itt előzetes szelekció nélkül, a genom különböző pozícióiból származó, egymással átfedő (!) szakaszok leolvasása történik, ezért a readek összerakásához új módszerek kifejlesztésére volt szükség.

Az 1000 Genomes projekt nagyban közrejátszott a új generációs szekvenálási eljárások elterjedéséhez, és a mérési sajátosságok megismerésében. Ennek keretében az ezredforduló után 1000 ember teljes DNS-ét szekvenáltak. Az új generációs szekvenálási technológiákkal napjainkra lehetővé vált egy teljes emberi genom szekvenálása mindössze néhány ezer dolláros költséggel.

1.1.1. A genomszekvenálás klinikai aspektusai

A kutatók eleinte talán túlzottan nagy reményeket fűztek az emberi genom szekvenálásával megismert adatokhoz. A célok között szerepelt a gyakori betegségek hátterében álló különbségek azonosítása, ami alapján hatékonyan fejleszthetnek ki új gyógyszereket. Sajnos a genomszekvenálás eredményeit sokkal nehezebb értelmezni. Napjainkban a szekvenálást kutatási célokon kívül ismert gének vizsgálatára, elsősorban monogénes betegségek diagnosztizálására, valamint a megfelelő kezelés kiválasztásának segítésére használják. A klinikai felhasználhatósághoz fontos a mérendő genomiális szakasz gyakori mutációs jellegének ismerete.

1.1.2. Részleges genetikai asszociációs vizsgálatok (PGAS)

A részleges genetikai asszociációs vizsgálatokban kiválasztják az emberi genom egy alhalmazát, amelyben korábbi kisebb felbontású genetikai vizsgálatok vagy egyszerűen hipotézis alapján sejtik, hogy egy betegséggel összefüggésbe hozható variánsok találhatóak. Ezek után meghatározzák a kiválasztott alhalmazban a variánsokat az eset- és kontrollpopulációkon, majd statisztikai módszerekkel vizsgálják, hogy mely variánsoknak van hatása a fenotípusra, és hogy melyek ezek között az okozati variánsok. A részleges genetikai asszociás kísérletek tervezéséről a könyv további fejezeteiben olvashat.

1.1.3. Genomszintű asszociációs vizsgálatok (GWAS)

A HapMap projektben elkészült az emberi rekombinációs hotspotok, valamint az emberi genom variációinak térképe. A genomszintű asszociációs vizsgálatokban a HapMap projektből ismert egynukleotidos polimorfizmusokat (SNP-k) úgy választják ki, hogy maximalizálják az általuk kapcsoltságban levő SNP-k számát, majd meghatározzák az SNP-ket nagy populációkon. A populációra vonatkozó adatok ismeretében lehetőség van a jelleggel asszociált, vagyis kapcsoltsági egyenlőtlenségben álló régiók, akár csökkentett számú SNP-készlet (tag-SNP) meghatározása ellenére, történő nagy hatékonyságú szűrésére. A

GWAS vizsgálatban általában kiválasztanak egy betegséget, és meghatároznak több mint egymillió SNP-t többezer eset- és kontrollmintán. Az eredményeket statisztikai vizsgálatok alá vetik és meghatározzák, hogy mely betegségek asszociálnak a kérdéses betegséggel.

1.2. Első generációs automatizált Sanger-szekvenálás

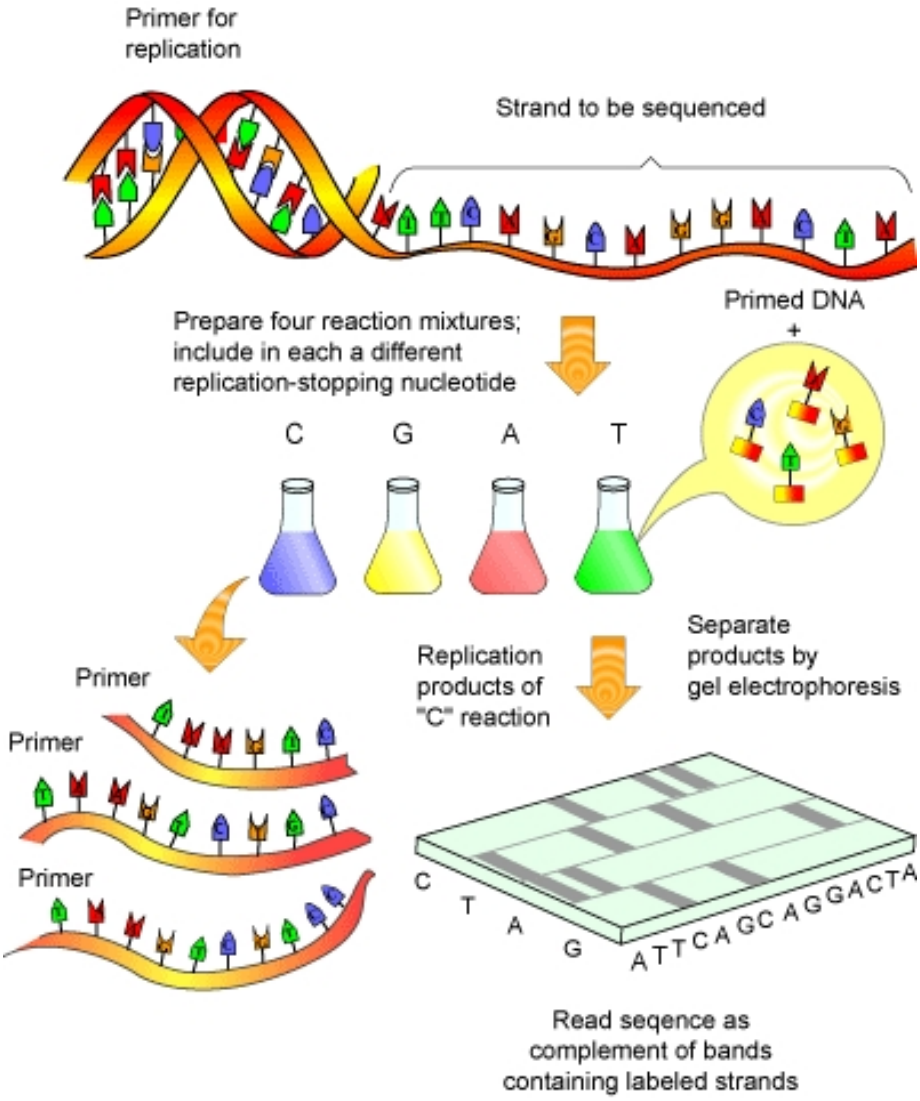
A Sanger-szekvenálás módszere, a nukleotidok szelektív beépülésén alapul. Ezt a módszert 1977-ben fejlesztették ki, és ma is a legelterjedtebb módszernek számít. A Sanger-szekvenálás hosszú, akár 800 bázispáros readeket eredményez, és a HGP után leggyakrabban megerősítő vizsgálatokra (validálásra) és kisebb léptékű kutatásokban használják. A négy különböző nukleotidot négy eltérő fluoreszcens terminátorral jelölik, majd a fragmenseket elektroforézis segítségével gélen megfuttatják, ezután a keletkező fluoreszcens képet rögzítik. Ezek után a fluoreszcens sávokat a végső szekvenciává dekódolják. A Sanger-szekvenálás kifejezetten lassú és drága az újabb módszerekhez képest, ellenben a megbízhatósága és a hibakarakterisztikája jól ismert.

1.3. Új generációs szekvenálási technológiák

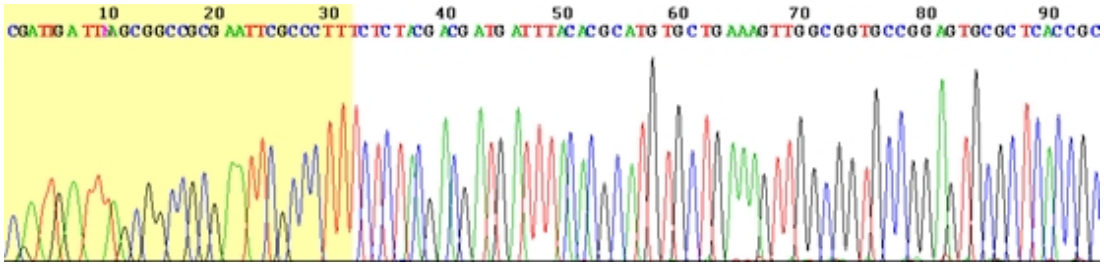
Ezek a technológiák mind a rövid DNS szakaszok leolvasásának nagymértékű párhuzamosításán alapulnak. Általában lehetővé teszik több minta egyidejű vizsgálatát, az egyes minták térbeli elkülönítésével. A nagyobb mértékű párhuzamosítás nem növelte jelentősen az egyes leolvasások pontosítását, de egy szakasz redundáns, nagyobb lefedettséggel (coverage) a mérés végeredménye pontosítható. A gyakorlatban technológiától és a mért szakasz jellegétől függően a kb. 30-szoros lefedettségtől akár több ezerszeres lefedettségig terjedhet a mérés. A jelenlegi vezető technológia elterjedtség alapján az IlluminaMiSeq platformja. Az átlagos read-hossz azonban relatíve alacsony.

1.3.1. Piroszekvenálás és pH alapú szekvenálás

2005-ben a 454 Life Sciences kifejlesztett egy szintézisalapú szekvenálási módszert, amelyben a leolvasni kívánt egyszálú DNS-t egy felszínhez rögzítik és enzimatikus úton szintetizálják a komplementer szálat. A piroszekvenálás alapja az, hogy a DNS-t szintetizáló DNS polimeráz enzim aktivitását egy másik, fénykibocsátó (luciferáz) enzim segítségével detektálják. A szekvenálás folyamán a négy lehetséges nukleotid közül egyszerre egyet juttatnak a reakciótérbe, és ha ez komplementer a következő nukleotiddal, akkor a DNS polimeráz beépíti a szálba. A beépülés következtében, egy kapcsolt biokémiai reakciósor végén a luciferáz enzim által katalizált átalakulási folyamat fénykibocsátással jár. A fény intenzitása arányos a inkorporálódó nukleotidok számával. Ezek után a be nem épült nukleotidokat kimossák, és egy másik nukleotidot adnak a rendszerhez. Ezt a folyamatot ciklikusan ismétlik. Az adott pozícióban bekövetkező fényvillanások megadják a kérdéses DNS szál nukleotidszekvenciáját.



1.1. ábra. A Sanger-szekvenálás folyamata



1.2. ábra. A Sanger-szekvenálás eredménye: a flowgram

A pH alapú szekvenálás elve a nukleotid beépülésekor detektált esemény vonatkozásában tér el, amikor is egy kilépő proton megváltoztatja a reakcióelegy pH-ját. Ezt a

Platform	Library/ template preparation	NGS chemistry	Read length (bases)	Run time (days)	Gb per run	Machine cost (US$)	Pros	Cons	Biological applications	Refs
Roche/454's GS FLX Titanium	Frag, MP/ emPCR	PS	330*	0.35	0.45	500,000	Longer reads improve mapping in repetitive regions; fast run times	High reagent cost; high error rates in homo-polymer repeats	Bacterial and insect genome *de novo* assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics	D. Muzny, pers. comm.
Illumina/ Solexa's GA_x	Frag, MP/ solid-phase	RTs	75 or 100	4[§], 9[§]	18[§], 35[§]	540,000	Currently the most widely used platform in the field	Low multiplexing capability of samples	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Life/APG's SOLiD 3	Frag, MP/ emPCR	Cleavable probe SBL	50	7[§], 14[§]	30[§], 50[§]	595,000	Two-base encoding provides inherent error correction	Long run times	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Polonator G.007	MP only/ emPCR	Non-cleavable probe SBL	26	5[§]	12[§]	170,000	Least expensive platform; open source to adapt alternative NGS chemistries	Users are required to maintain and quality control reagents; shortest NGS read lengths	Bacterial genome resequencing for variant discovery	J. Edwards, pers. comm.
Helicos BioSciences HeliScope	Frag, MP/ single molecule	RTs	32*	8[§]	37[§]	999,000	Non-bias representation of templates for genome and seq-based applications	High error rates compared with other reversible terminator chemistries	Seq-based methods	91
Pacific Biosciences (target release: 2010)	Frag only/ single molecule	Real-time	964*	N/A	N/A	N/A	Has the greatest potential for reads exceeding 1 kb	Highest error rates compared with other NGS chemistries	Full-length transcriptome sequencing; complements other resequencing efforts in discovering large structural variants and haplotype blocks	S. Turner, pers. comm.

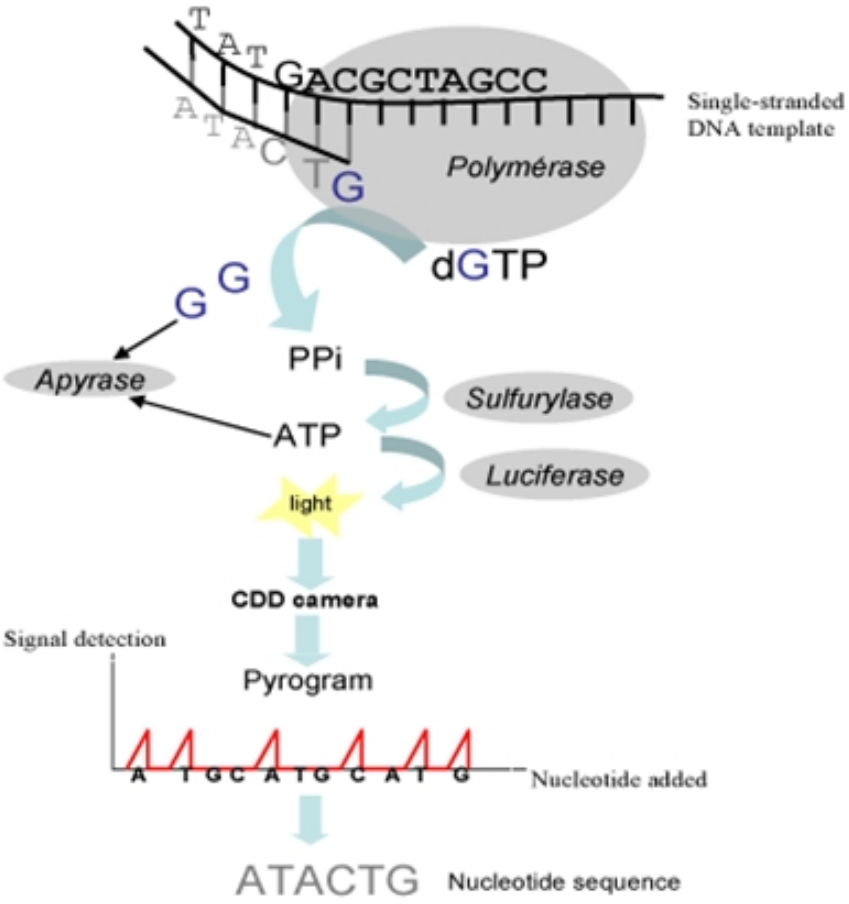
1.3. ábra. Szekvenálási technológiák összehasonlítása

pH-változást egy CMOS (Complementary Metal-Oxide Semiconductor) felületen detektálják. A pH-változás mértéke arányos a beépülő nukleotidok számával. Ezek alapján könnyen belátható, hogy a piroszekvenálás és a pH alapú szekvenálás hibakarakterisztikája rendkívül hasonló.

1.3.2. Reverzibilis terminátor alapú szekvenálás

A technológiát 2006-ban mutatta be az Illumina. A DNS szálakat lemezeken rögzítik, majd helyben sokszorosítják híd-amplifikációval. A mérési folyamat során a következő ciklust ismétlik. 1. A négy különböző terminált és különböző fluoreszcens festékkel jelölt nukleotidot adnak a lemezhez, ahol minden DNS szál következő szabad helyére a megfelelő komplementer nukleotid épül be. 2. A felesleges nukleotidokat kimossák, majd a lemezről rögzítik a négy különböző fluoreszcens festéknek megfelelő hullámhosszhoz tartozó képet. 3. A termináló csoportokat levágják a szálakról és kimossák.

A különbség az Illumina és a Helicos BioSciences megoldása között az, hogy a Helicos

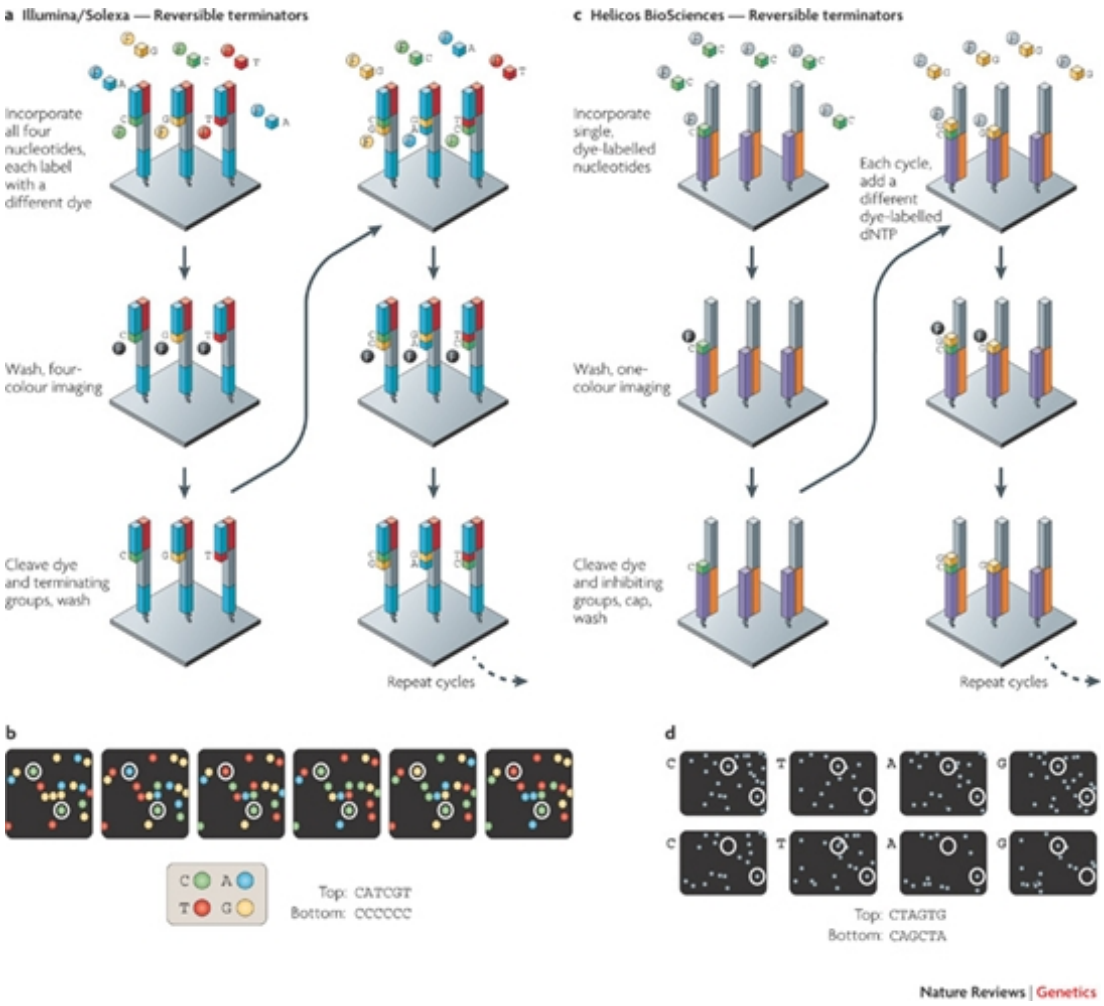


1.4. ábra. A piroszekvenálás menete

megoldása esetén csak egy adott nukleotid van jelen és épül be az egymás után következő ciklusokban, míg az Illumina platform esetében egyszerre mind a négy különböző, négy különböző festékkel. A rögzített képekből megállapítják a klonális klaszterek pozícióit, majd az egyes színekből és intenzitásokból meghatározzák a nukleotidszekvenciákat.

1.3.3. Nanopórus alapú szekvenálás

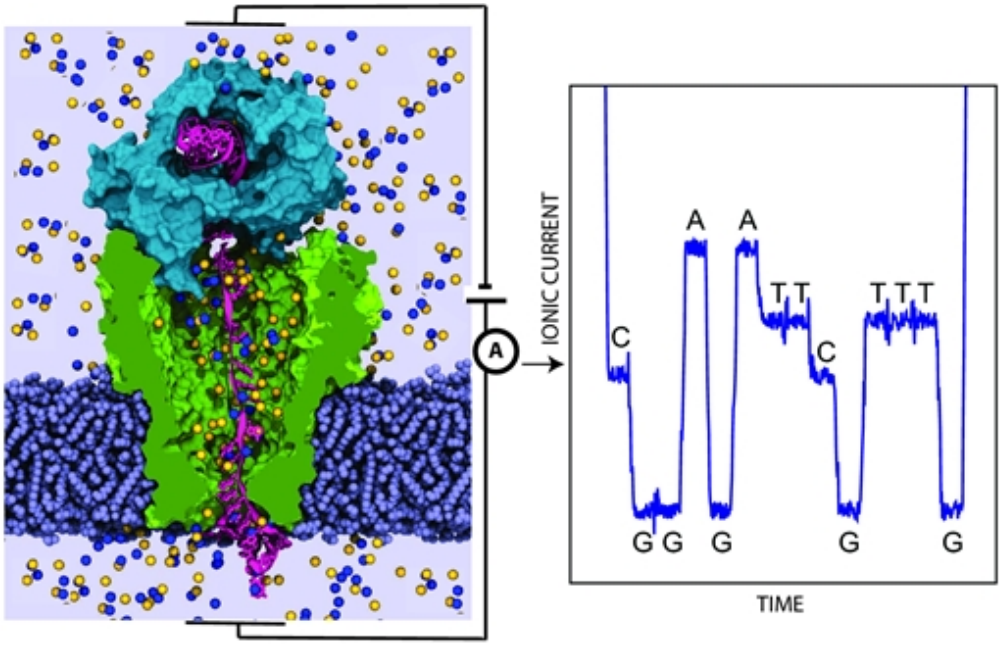
A nanopórus alapú szekvenálást 1995 óta fejlesztik. Jelenleg még nincs kereskedelmi forgalomban, de a technológia 2014-es bevezetésétől hosszú és pontos readeket várnak. A nanopórusok speciális fehérjék, amelyeket egy lemezen rögzítenek. Egy egyszálú DNS szálat vezetnek át a póruson, és megmérik a lemez egyik oldaláról a másikra folyó áram erősségét. A pórus rendkívül kicsi – kevesebb, mint 1 nm átmérőjű –, amelyen keresztül az egyes nukleotidok áthaladását az adott nukleotidra jellemző áramerősség kíséri.



1.5. ábra. A reverzibilis terminátor alapú szekvenálás folyamata

1.4. Új generációs szekvenálási technológiák hibakarakterisztikája

Gyakran egy új generációs szekvenálási mérés nem azt az áttörő eredményt hozza, amire számítottak. A leolvasásra szánt könyvtárak előkészítése egy rendkívül bonyolult laboratóriumi folyamat, amely esetenként akár tíz órát is kitevő labormunkát igényel, és ezért könnyű hibát ejteni az előkésztés során. A kiindulási DNS mennyiség nanogramm-picogramm nagyságrendű. A különböző előkészítésbeli hibák eltérő jellegű torzulást okozhatnak kimeneti adatokban. Három fő hatás van, amely NGS mérések során a legnagyobb hányadban járulnak hozzá a mérési hibákhoz: rendszerszintű hibák, lefedettségi egyenetlenség és a mintaelőkészítési hibák, amelyek mind függenek a technológiai platformtól, a mért szekvencia jellegétől, és a kísérleti variabilitástól.



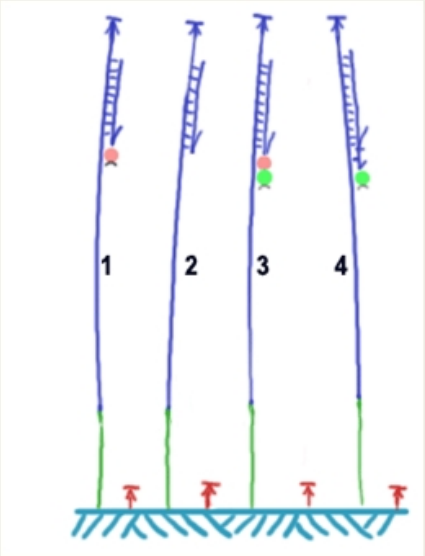
1.6. ábra. A nanopórus alapú szekvenálás illusztrációja és mérési eredménye

1.4.1. Carry forward/incomplete extension

A carry forward/incomplete extension (előrevitel/elégtelen beépülés) hiba akkor fordul elő, ha az egy gyöngyön vagy azonos helyen levő klonálisan azonos szekvenciák nem teljesen szinkron módon szintetizálódnak. Például ha néhány szál nem a megfelelő darabszámú nukleotidot építi be (mert nem volt elégséges számú nukleotid a flow során), vagy ha reziduális nukleotidok maradnak a szálaknál (mert nem megfelelően mosták ki az egyes flow-k közötti mosási ciklusokban), akkor a szálak nem teljesen szinkron módon növekednek. Ennek az eredménye egy jelszint-degradáció, ami addicionális zajt hoz a rendszerbe. Ezáltal nemcsak a readek minősége csökken, hanem halmozott esetben a readek rövidebbek lesznek. Ez a hiba különösen az Illumina eszközét érinti, a piroszekvenálás technológiája ellenállóbb ezzel a fajta hibával szemben. A szűrési algoritmusok tervezése során kiemelten fontos, hogy észlelni (és adott esetben korrigálni) tudják ezt a hibát.

1.4.2. Homopolimer hibák

A piroszekvenálás és a pH alapú szekvenálás során egy ciklusban több azonos nukleotid épülhet be a szekvenciába, ha a célszálon is azonos nukleotidok következnek. A kibocsátott fény mennyisége/pH-változás mértéke arányos a beépülő bázisok számával. Hosszabb homopolimer-régiók esetén, valamint a beépült nukleotid – fény mennyisége között fennálló nemlineáris összefüggés miatt a zaj és variáció növekszik és ezért egyre nehezebbé válik meghatározni a beépülő bázisok pontos számát. Ha nagyobb lefedettséggel szekvenáljuk a target-régiónkat, akkor következtethetünk a homopolimer-régió pontos hosszára.



1.7. ábra. Carry forward/incomplete extension

1.5. Capture technológiák

A legtöbb kísérleti kérdésben a DNS szekvenálást nem a teljes genomon hajtják végre, így nem shotgun módon végzik. A legtöbbször célzott régiókat szeretnének a genomból leolvasni. A könyvtárkészítés első lépése a célrégiónak („target" régió) a kinyerése („capture") és sokszorosítása. Erre több módszert is kidolgoztak.

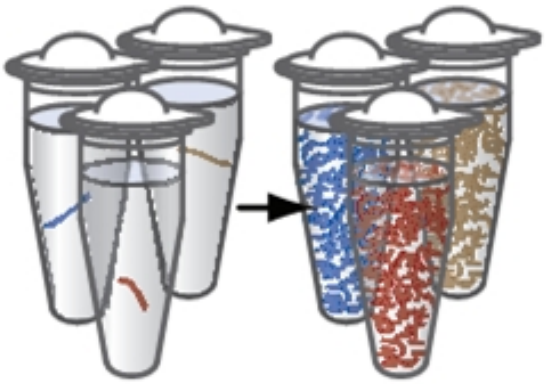
1.5.1. PCR capture

Amennyiben polimeráz láncreakciót használnak egy bizonyos célrégió kinyerésére, olyan primert kell tervezni, amely csakis a target-régió elejére vagy végére köt be. Ennek a primernek egyedinek kell lennie, ami azt jelenti, hogy csak egyetlen helyre köthet be a genomban. Valamint arra is figyelni kell, hogy ne essen SNP vagy más variáns a primer kötőhelye alá, mert ez rontja a specificitást. A primernek költséghatékonysági okokból minél rövidebbnek kell lennie, a gyakorlatban emberi genomnál ez egy 20–25 hosszú oligonukleotid szekvenciát jelent. A DNS mintát felmelegítik, hogy a hélixet alkotó szálak kettéváljanak, ezután a primert hozzáadják enzimekkel és szabad nukleotidokkal, majd az elegyet lehűtik. Ez indítja be a PCR első lépését. A komplementer szálat a DNS polimeráz enzim építi fel a szabad nukleotidokból. Ezt a melegítés-PCR-hibridizáció lépést egymás után többször hajtják végre, és ideális körülmények között a cél régió (termék) minden lépésben megduplázódik.

Uniplex PCR

Az uniplex PCR során a reakcióelegyben csak egyetlen cél régiót szaporítanak fel. Az uniplex PCR kompatibilis minden új generációs szekvenálási platformmal, és ma már ru-

tineljárásnak számít. A leghosszabb régió, ami még amplifikálható PCR-rel, az kb. 10000 bázis, mert ennél hosszabb régiók esetén a DNS polimeráz kevésbé robusztus, és fennáll a korai láncterminációó veszélye. Ha a szál pl. az első lépésben túl korán terminálna, akkor a következő ciklusoktól már minden terméknek legalább a fele csak ez a rövid szekvencia lenne. Több, egyedileg felszaporított PCR termék vegyítése megengedhető, ám fontos figyelembe venni, hogy ezt csak pontos kvantitatív elemzés után ajánlatos tenni. Ha nagyon eltérő koncentrációjú elegyeket vegyítenek, akkor a nagyobb koncentrációban az oldatban levő szálak többségbe kerülnek. Ez az egyenletes lefedettség elérése miatt rendkívül fontos.



1.8. ábra. Uniplex PCR

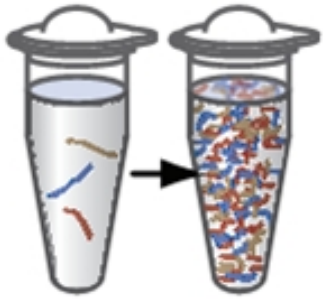
Multiplex PCR

Multiplex PCR reakciókban több primert egyszerre adnak egyetlen reakcióelegyhez, és közös templát jelenlétében sokszorosítják a targeteket. Az egyedi PCR-reakciók térben nem kerülnek elválasztásra, így figyelembe kell venni a különböző szekvenciák eltérő olvadási és hibridizációs hőmérsékletét a primerek tervezése során. A tapasztalatok szerint a lefedettség egyenetlensége kb. 10 target-régióig (amplikon) biztosítható. Az egyes target-régiók hossza közelítőleg azonos kell, hogy legyen az egyenletes lefedettség érdekében.

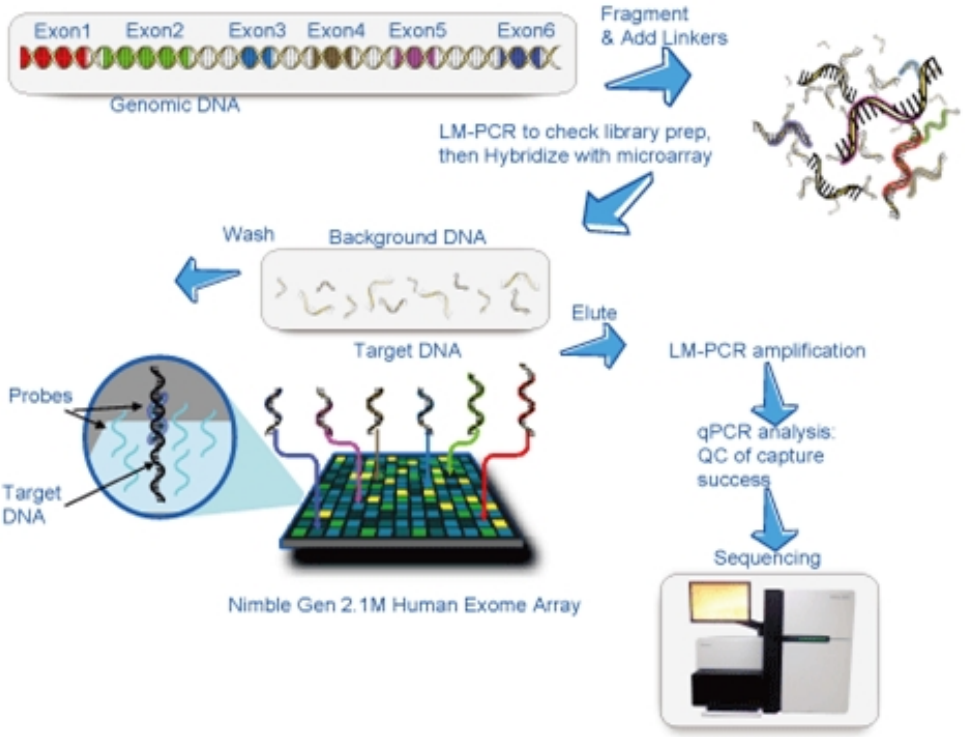
Ha a primerek interakcióba lépnek egymással (átfedések miatt „összeakadnak"), akkor nagyon egyenetlen lefedettség várható, vagy akár egy amplikon egyáltalán nem kerül sokszorosításra. Előfordulhat nem célzott régiók felszaporodása is. A módszer előnye a fajlagosan alacsonyabb anyag- és munkaidőköltség.

Microarray capture

Egy microarray lemezen több millió rögzített oligonukleotid szekvencia lehet, amelyek a célrégiókra specifikusak. A teljes hosszában amplifikált genomi DNS szekvenciákat hibridizálják a rögzített oligonukleotidokra. Azokat a szekvenciákat, amelyek nem kötődnek a lemezen levő helyekre, lemossák, majd a megmaradt célszekvenciákat eluálják (leoldják) a lemezről. A kiválasztott régiókat tartalmazó, eluált DNS-t opcionálisan tovább amplifikálják, majd adapterszekvenciák ligálása után feltöltik a szekvenáló eszközre.



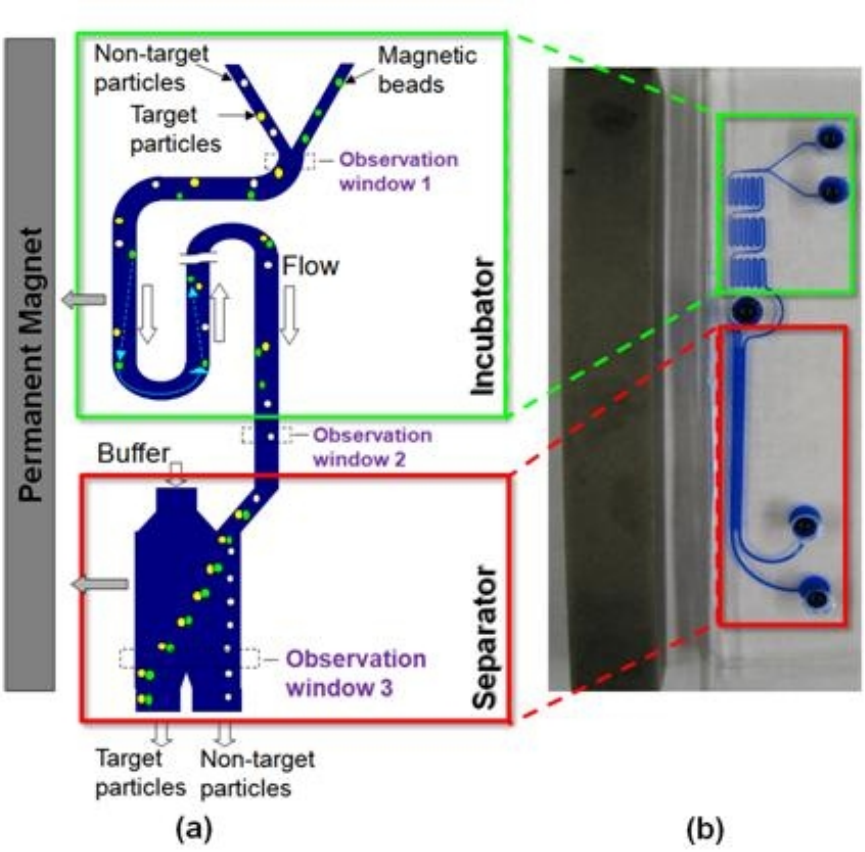
1.9. ábra. Multiplex PCR



1.10. ábra. Microarray capture

Microfluidic capture

A microfluidic capture során apró vízcseppeket hoznak létre egy olaj közegben (emulzió). Minden csepp egy mikroméretű reakciós tartálynak felel meg, ahol a reakciók egymástól elszigetelve futnak. Az egyes primereket és templátokat tartalmazó cseppeket vizuális vagy automatizált ellenőrzés után elektrosztatikus térrel válogatják össze. Ezzel a módszerrel egyszerre több millió reakciót lehet végrehajtani elválasztott cseppekben.

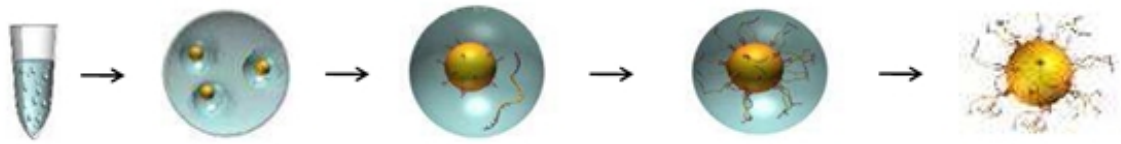


1.11. ábra. Microfluidic capture

1.6. Emulziós PCR

Az emulziós PCR elve hasonló az előző fejezetben bemutatott cseppeknél bemutatottakhoz, a különbség az, hogy itt a vizes reakcióelegy-olaj emulzió minden cseppjében egy mágneses gyöngy van. Először a mágneses gyöngyökön levő oligókkal komplementer adaptereket ligálnak (kapcsolnak) a kiválasztandó DNS szálakra, majd a mágneses gyöngyöket hozzáadják az alacsony koncentrációjú DNS templátot tartalmazó és a PCR-reagenseket (polimeráz, nukleotidok, ionok) tartalmazó elegyhez. Fontos, hogy a mágneses gyöngyök száma megfelelő legyen a DNS szálak számához, mert így biztosítható, hogy a legtöbb gyöngyhöz csak egyetlen egyedi DNS szál fog csatlakozni. Ezt az oldatot emulzifikálják úgy, hogy egy cseppbe legfeljebb egy gyöngy kerüljön, majd a PCR-reakciónak megfelelő hőmérsékleti ciklusokat ismétlik. Ennek eredményeképp minden gyöngyön egy egyedi DNS szál több ezer másolata. Az emulziót „megtörését” követően a gyöngyöket feltöltik a egy olyan lapkára, amelyen éppen akkora lyukak vannak, amiben csak egy gyöngy fér el. Ezek után kezdődhet meg a több százezer térben részlegesen elválasztott kompartmentben a szekvenálás.

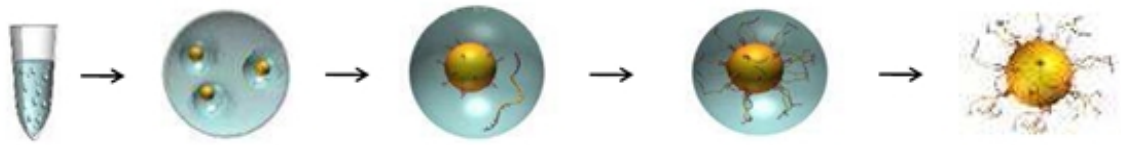
Az emulziós PCR-t mind a piroszekvenálásban, mind a pH alapú szekvenálásban használják.



1.12. ábra. Emulziós PCR

1.7. Híd- (bridge-) amplifikáció

Két különböző, egy nagyobb üveg lemezere kötöttekkel komplementer adapter szekvenciákat ligálnak (a szekvenciák végeire rögzítenek) az előkészített egyszálú DNS fragmensek végeire. A DNS oldatot ráhibridizálják az üveglemezre. Az egyes molekulák véletlenszerű helyekre kötnek a lemezen. A molekulák a hűtés során hidakat képeznek a lemezen úgy, hogy mind a két végükön levő adapterek a lemezhez kötődnek, majd a szekvenciákról másolatok készülnek. A másolatokat ezután denaturálják, majd megismétlik a hídképzés–másolás–denaturálás ciklust, amíg kellő méretű és sűrűségű klonális klaszterek nem képződnek. Egy lemezen több százmillió egymástól elkülönülő klonális klaszter alakul ki, és ha elégséges közöttük a távolság, akkor nem is keverednek össze. Ezt a lemezt utána beillesztik a szekvenáló eszközbe. Az Illumina megoldása ezt a módszert használja.



1.13. ábra. Híd-amplifikáció

1.8. Célzott újraszekvenálás

Az új generációs szekvenálást leggyakrabban olyan régiók vizsgálatában alkalmazzák, amelynek (referencia) szekvenciája már ismert. Többféleképp is kiválasztható és amplifikálható egy organizmus célszekvenciája. A célzott újraszekvenálás az ismeretlen genomok de novo összerakásának feladatát leegyszerűsíti egy egyszerű illesztési problémává, azáltal hogy a referenciaszekvencia már ismert és felhasználható mint térkép. Minden szekvenálási technológiának vannak hibaforrásai, és a rendkívül nagy párhuzamosság miatt ez rendre meg is jelenik a readekben. A legjobban a lefedettség növelésével lehet kiküszöbölni a hibák hatását, de ez sajnos nem túl költséghatékony, valamint szekvencia-specifikus hibáknál nem segít sokat.

1.9. De novo szekvenálás

A de novo szekvenálást abban az esetben alkalmazzák, ha az adott szervezet örökítőanyagát első ízben olvassák le, és így nem áll rendelkezésre referenciaszekvencia. A módszerei ugyanakkor nagyobb léptékű átrendeződések, például tumorsejtek mutációi esetében, ezek keresése során is bevethetők. A nem specifikusan felamplifikált teljes DNS állományt a választott szekvenálási technológiának megfelelő méretűre tördelik (akár ultrahang alkalmazásával vagy enzimes emésztéssel). A leolvasás során előállított szekvencia-fragmensek összeillesztése nagyon komplex feladat, és nagy lefedettség is szükséges ahhoz, hogy folytonos, nagyobb egységekké, akár kromoszómákká lehessen rendezni a readeket. Az emberi DNS 3 milliárd bázispárjának de novo illesztéséhez kb. 100 millió readre van szükség, bár ez erősen függ a technológiától. Az egyszerűbb organizmusok genomja, mint például a baktériumoké vagy a vírusoké, nagyságrendekkel kisebb, és akár egyetlen futásból is teljesen összeilleszthetőek.

1.10. Új generációs szekvenálási munkafolyamatok

Manapság már nemcsak kutatási célokra, hanem rutinszerű diagnosztikában is használnak szekvenálást, így léteznek ajánlások arra, hogyan lehet a mérések eredményét felhasználni döntéstámogatásra és diagnosztikára.

1.10.1. Szűrés

Minden szekvenáló platform empirikus adatokon és méréseken kalibrált módon hozzá rendel minden read minden bázisához egy Phred pontszámot. Ez a pontszám annak a log10 valószínűségét adja meg, hogy a bázishívás hibás. Minden readet szűrni kell a minőségbiztosítás érdekében, például a túl rövid readeket és az alacsony minőségű readeket el kell dobni. Lehetőség van a readek kevésbé jól sikerült végeinek a levágására is, valamint számtalan feltétel felállítására.

1.10.2. Illesztés

Az illesztés, más nevén mapping vagy alignment az újraszekvenálás egyik fontos lépése. Itt a célgenomhoz illesztjük a readeket egyesével, majd ezekből összeállítjuk a lemért szakaszunk konszenzusos szekvenciáját. Több algoritmus is létezik a legjobb illesztési pozíció megkeresésére.

1.10.3. Összerakás

Amennyiben referenciaszekvencia nélkül illesztünk össze rövid readeket egy folytonos szekvenciává, ezt összerakásnak (assembly) nevezzük. Az összerakási probléma szemléltethető úgy, hogy egy könyv több példányát véletlenszerűen apró darabokra szabdaljuk, majd össze kell rakni az eredeti könyvet a kis darabokból. A leggyakrabban használt algoritmus

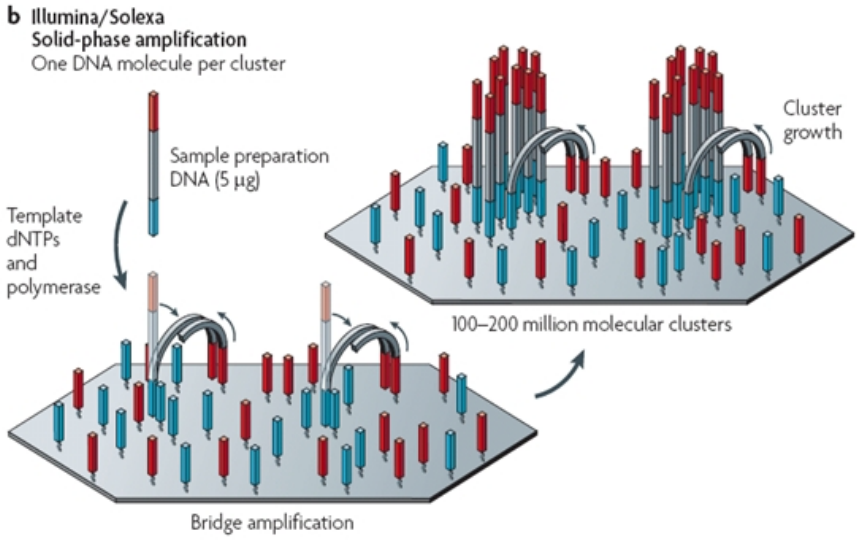
a mohó algoritmus, ahol a cél a legrövidebb közös szekvencia összeállítása, amelynek minden helyét fedi legalább egy read. Ehhez először readek páronkénti illesztését számítják ki, majd a két leginkább hasonló readet összevonják. Ezt addig ismétlik, amíg végül csak egy fragmens marad.

1.10.4. Variánshívás

A variánshívás folyamata az, amikor több readet, amelyek ugyanarra a genomi pozícióra illeszkednek, megvizsgálunk, és megvizsgáljuk, hogy bárhol eltér-e a referencia szekvenciától. Többféle variáns létezik, az egynukleotidos polimorfizmusoktól az inzerciókon és deléciókon, valamint kópiaszám változásokon át a nagyméretű strukturális átrendeződésekig.

1.10.5. Paired-end szekvenálás

Gyakran a rövid readek nem szolgáltatnak elegendő pozicionális információt ahhoz, hogy a nagyméretű kópiaszám-változásokat és átrendeződéseket egyértelműen meg tudjuk határozni, valamint a de novo összerakás is nagyon nehéz rövid readekből. A paired-end szekvenálás során sokkal nagyobb, pár kilobázis hosszú fragmenseket szekvenálunk, de a szekvenálás technológiai korlátai miatt csak a fragmensek két végét izoláljuk. Ilyenkor további információt nyerünk azzal, hogy van két rövid readünk, amelyeknek a hozzávetőleges távolságát ismerjük.



1.14. ábra. Paired-end szekvenálás adatainak összeállítási folyamata

1.11. Több minta párhuzamos szekvenálása

Amennyiben több egyedből származó mintákat szekvenálunk egyetlen futtatással, nagyon fontos hogy meg tudjuk állapítani, hogy melyik read melyik egyedekből származik. Minden új generációs szekvenálási technológia valamilyen formában támogatja a minták azonosító szekvenciákkal való ellátását. Ilyenkor a fragmensek elejére azonosítókat kapcsolnak, amelyek a fragmenssel együtt kerülnek leolvasásra. Ezek az azonosítók körülbelül tíz bázis hosszúak, és többnyire tartalmaznak valamilyen kódolt redundanciát arra az esetre, ha az azonosító leolvasásában hibát ejtenénk. Ugyanakkor egyes platformok támogatják a lemezek külön részekre osztását, így pedig az adott read lemezen elfoglalt helye egyértelműen azonosítja a mintát.

2. fejezet

Genetikai mérések és utófeldolgozásuk, haplotípus-rekonstrukció, imputálás

2.1. A genom fogalma

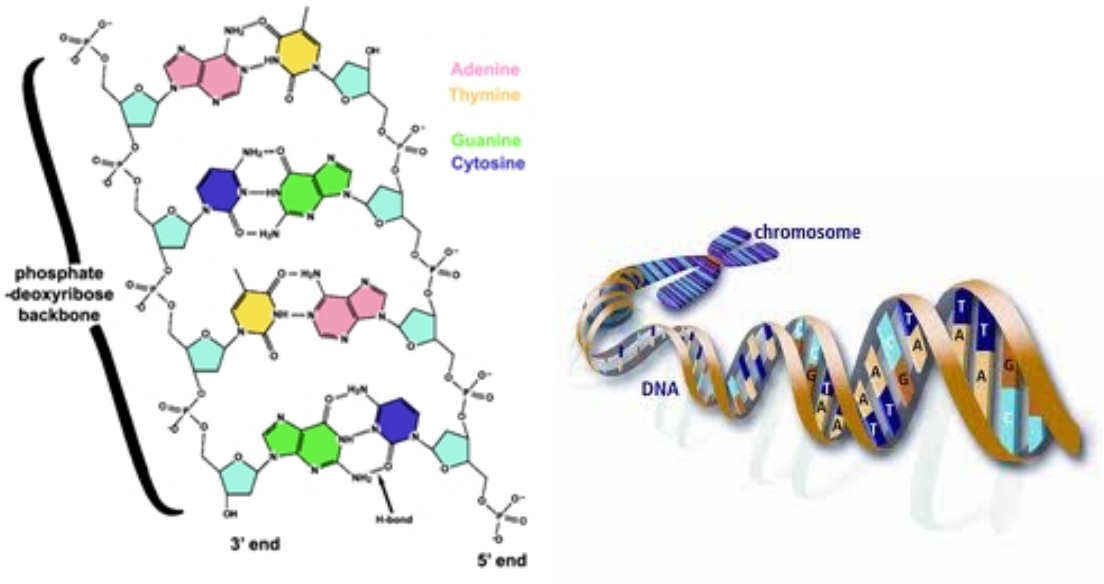
A genom egy szervezet teljes örökítő információját tartalmazza, amely legtöbb esetben a DNS-ben van kódolva.

A DNS kettős hélix szerkezetű molekula. A két összekapcsolódó DNS szál vázát egy cukor-foszfát gerinc alkotja, a cukor egységekhez kapcsolódó bázisok – a bázispárképzés szabályainak megfelelően – másodlagos kötésekkel tartják össze a két, ellentétes lefutású DNS szálat.

A DNS-t felépítő nukleotidok négyféle bázist tartalmazhatnak, két purinbázist: adenint és guanint, illetve két pirimidinbázist: citozint és timint. Az adenin és a timin két, ugyanakkor a guanin és a citozin három hidrogénkötést tud a szálak között létesíteni. A DNS-ben található cukor az ötszénatomos dezoxiribóz. A cukoregységek egymáshoz foszfodiészter kötéssel kapcsolódnak, az egyik dezoxiribóz 3'OH-csoportja és a következő cukorkomponens 5'OH-ja foszfátcsoport közvetítésével kapcsolódik egymáshoz (2.1. ábra).

A genetikai információt az egymást követő négyféle bázis sorrendje, a bázissorrend határozza meg.

A DNS szálak és kapcsolódó fehérjék alkotta kromatin a sejtosztódás során a méret csökkentése érdekében hiszton fehérjék segítségével csomagolódnak össze, és fénymikroszkópban látható kromoszómákba tömörülnek. Az emberi sejtek többségében minden kromoszóma két példánya található meg, az ilyen sejteket diploid sejteknek nevezzük. A két egyforma alakú és nagyságú, összetartozó kromoszóma közül az egyik apai, a másik anyai eredetű (homológ kromoszómák). Az ivarsejtek egyetlen kromoszómakészlettel rendelkeznek (haploidok). Az emberi sejtekben a sejtmag 23 különböző kromoszóma alkotta párt (összesen 46) tartalmaz, amelyből 22 pár homológ testi kromoszóma, míg egy pár nemi kromoszóma (X, illetve Y). A homológ kromoszómák meghatározott pozíciókban ugyanazokat a géneket hordozzák, azonban ez nem azt jelenti, hogy azok feltétlenül bázisról-bázisra azonos genetikai információt tartalmaznak, hiszen a sejtben az adott génnek két különböző (apai és anyai) változata is jelen lehet. Ezek az allélok, a kromoszóma



2.1. ábra. A DNS és a kromoszómák struktúrája

egy adott pontján (lokuszán) elhelyezkedő gén variációi. Egy olyan egyedet, amely két homológ kromoszómáján egymással teljesen azonos génkópiát hordoz, homozigótának, amely különbözőt, azt heterozigótának nevezzük. A fenotípussal összefüggő allél lehet domináns vagy recesszív. Egy domináns és egy recesszív allél hordozása esetén a dominánsnak megfelelő fenotípus fog érvényre jutni; ugyanakkor egy recesszív allélhoz tartozó jelleg csak homozigóta genotípus esetén tud megjelenni.

Egy egyed fenotípusán teljes fizikai megjelenését, vagy bármely megfigyelhető vagy kimutatható (szerkezeti, biokémiai, élettani vagy akár viselkedési) jellemzőjét, amelyet genotípusa és a környezeti hatások együttesen határoznak meg.

2.2. A genotípus „az egyed genetikai identitása”

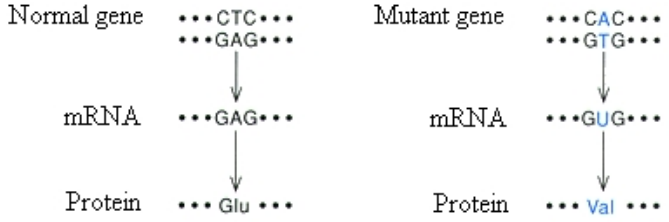
A genotípus tágabb értelemben egy egyed genetikai összetételét írja le, melyet a genomjában, a DNS-szekvenciában található információ határoz meg. Szűkebb értelemben, a genotípus fogalma egy adott gén (vagy gének csoportjához tartozó) változataira vonatkozik, amelyek kombinációja képezhet egy genotípust, amely a hordozók fenotípusának kialakításában szerepet játszik.

A genotipizálás során az egyed genotípusát határozzuk meg, általában részlegesen. A populációk genetikai összetételét, az egyes genotípusok populációbeli gyakoriságával és az ebből kiszámítható allélgyakorisággal lehet jellemezni.

2.2.1. Egypontos nukleotid-polimorfizmus (SNP)

A DNS replikációja során létrejött eltéréseket, hibákat nevezzük mutációnak. A legtöbb mutáció spontán jön létre, de bizonyos mutagén szerek is okozhatják kialakulásukat. A szabályos, leggyakoribb, a jellemző fenotípust kialakító allélt vad típusnak nevezzük. Mutáció következtében a genom szekvenciája kisebb-nagyobb mértékben változhat meg: a génhiba kiterjedésétől függően egyetlen bázist, egy gén bizonyos szakaszát, de akár egy egész kromoszómát vagy -készletet is érinthet.

Azt a variációt a DNS szekvenciában, amely akkor jön létre, ha a genomban egy nukleotid megváltozik, egypontos nukleotid-polimorfizmusnak, az angol kifejezés (single-nucleotide polymorphism) rövidítéseként SNP-nek nevezzük. Ha a populáció több, mint 1%-a a DNS-szekvencia egy adott pozíciójában eltérő nukleotidot hordoz, akkor ez a variáció SNP-nek tekinthető. A legfontosabb fogalmi különbség tehát az egyénekben azonosítható pontmutáció és a populáció akár jelentős részében is megjelenő SNP között a gyakoriságukban van.



2.2. ábra. Egyetlen nukleotid változása egy másik mRNS kodon transzlációját eredményezi, amely végül egy eltérő peptidlánc szintéziséhez vezet.

SNP-k előfordulhatnak a gének kódoló és nem kódoló régióiban, valamint a DNS gének között elterülő intergenikus területein is. Amennyiben a mutáció egy gén kódoló részében fordul elő, a megváltozott szekvencia hatással lehet a termék aminosav- sorrendjére és ezáltal a fehérje szerkezetére, funkciójára (2.2. ábra). A gének nem kódoló régióiban található báziseltérés befolyással lehet pl. a splicing-ra, transzkripciós faktorok kötődésére vagy az mRNS degradációjára.

2.2.2. A pontmutációk lehetséges változatai

A DNS molekula egyetlen bázisa megváltozhat kicserélődéssel, kieséssel vagy beékelődéssel.

A báziscsere során egyetlen bázis cserélődik ki egy másikra. Tranzíció során purinbázis cserélődik purinbázisra, vagy pirimidin pirimidinre (pl. A-G vagy T-C csere). Transzverzió esetében purinbázis cserélődik pirimidinre (pl. A-T vagy A-C csere) vagy pirimidin purinbázisra (pl. C-G vagy C-A csere).

A DNS szekvenciájának egy pontban történő megváltozásának következményeit nagymértékben meghatározza a pozíció fontossága az információ átadásában.

Csendes mutáció esetén báziscsere történik ugyan, de az nem okoz változást az érintett fehérje aminosavsorrendjében. Ennek oka a genetikai kód degeneráltságában rejlik, egy aminosavat többféle bázis-triplet, vagyis kodon is kódolhat, így előfordulhat, hogy a kodon harmadik (lötyögő) pozíciójában történt változás nem eredményez aminosav cserét, amikor egy másik, ugyanazt az aminosavat kódoló tripletre változik meg a szekvencia.

Nonszensz mutáció esetében egy, egyébként aminosavat kódoló triplet egy stop kodonra változik, aminek következtében a fehérjeszintézis megáll ennél a kodonnál és teljes hosszában nem kerül leolvasásra.

Misszensz mutáció esetében a nukleotidcsere aminosavcserét is eredményez, amely hatással lehet a képződő fehérje szerkezetére, funkciójára.

A legdrasztikusabb változást a genetikai kód információtartalmában egy bázis változása esetén annak kiesése (deléció) vagy egy új bázis beékelődése (inzerció) okozhatja. Mindkettő eset a leolvasási keret (reading frame) eltolódását eredményezi. Ennek következtében az soron következő aminosavak nagy valószínűséggel megváltoznak. Gyakran az is előfordul, hogy egy báziskiesés érvénytelenít egy stop kodont, vagy akár új létrehozását is okozhatja. Az ilyen mutáció következtében a fehérjeszintézis során nem megfelelő hosszúságú, illetve szerkezetű fehérjetermék keletkezik.

2.2.3. Mutációk hatása

Az emberi DNS-ben található variánsok hatással vannak arra, hogy hogyan reagál az emberi szervezet a betegségekre, baktériumokra, vírusokra, kemikáliákra. Sejtjeinkben számos mutáció alakul ki életünk során, amelyek jó része kijavításra kerül. A DNS-ben bekövetkező változások jelentik az evolúció folyamán a szervezet adaptációjának lehetőségét is a környezeti változásokhoz.

A nem örökölt, testi sejtekben kialakuló mutációkat, testi vagy szomatikus mutációknak nevezzük. A szomatikus, az egyed életében jelentkező génhibák nagyrészt nem járnak fenotípusos következménnyel. Az ellenkezőjére is jócskán akad példa, ha a daganatokra gondolunk, ahol a sejtosztódásban szerepet játszó gének túlműködése vagy éppen kikapcsolása a szabályozási felügyelet elvesztéséhez vezethet. Az öröklődő vagy csíravonalmutációk esetében nem a testi sejtek, hanem az ivarsejtek genetikai anyagában következik be a változás, amely így már örökölhető lesz és kimutatható a populációban.

Összességében az ismert SNP száma jelenleg több mint 70 millióra tehető az emberi populációkban. Ezek a pontszerűen elhelyezkedő variánsok többsége nem okoz semmilyen ismert káros következményt vagy betegséget az élőlényekben, csak a genetikai változatosságot növeli. Emellett azonban számos olyan SNP-ről tudunk, amely betegség – vagy legalábbis a betegségre való hajlam – kialakulásáért tehetők felelőssé. Ilyen betegségek például a sarlósejtes anémia, diabetes, szív- és érrendszeri betegségek, látásproblémák, rákos megbetegedések (emlő- és petefészekrák), mentális leépüléssel járó betegségek (Alzheimer-kór).

Az SNP-k a személyre szabott gyógyászat kulcsszereplői, nagy segítséget jelentenek az orvosi kutatásokban, gyógyszerek kifejlesztésében, mivel ezek nem sokat változnak generációról generációra, azaz a populációkban való SNP követés lineáris következtetéseket tesz

lehetővé.

2.3. Haplotípusok

A haplotípus a haploid (jelentése egyszeres) és a genotípus szavakból származik. A biológiai definíció szerint a haplotípus egyik szülőtől és egy kromoszómáról származó, egymáshoz szorosan kapcsolódó genetikai markerek halmaza. Egy másik gyakran használt definíció szerint a haplotípus egy homológ kromoszómapárról származó markerek azonos gametikus fázisú nukleotidjait jelöli (az egymás mellett a kromoszómán elhelyezkedő variánsok segítségével meghatározott haplotípus fázis segít annak megadásában, melyik szakasz származik az apai és melyik az anyai homológ kromoszómáról). Ez a megközelítés a szorosan kapcsolódó markereket haplotípus-blokkoknak nevezi. A fejezetben ez utóbbi értelmezést használjuk.

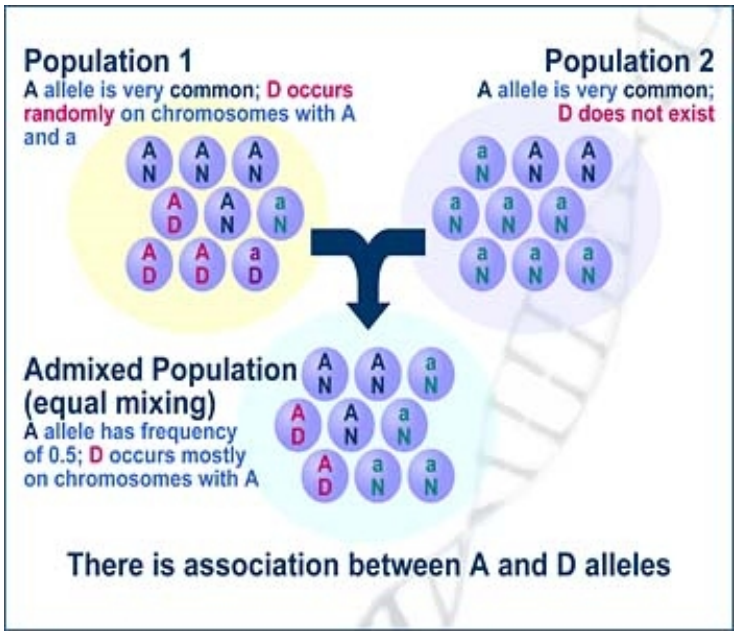
A haplotípusok vizsgálatának számos előnye van. A htSNP-ekre alapozva leszűkíthető a továbbiakban vizsgálandó SNP-ek halmaza. Emellett bizonyos fenotípusjegyeket, különösen a komplex betegségek esetében, több variáns együttesen határoz meg. Ekkor a haplotípus-szintű eredmények jóval erőteljesebbek lehetnek, mint a SNP alapúak.

A haplotípusok struktúrájának meghatározásához szükséges fázisos genotípus adatok legegyszerűbben családfaelemzésekből származhatnak. Családfaelemzések mellett különböző PCR technikákkal, vagy új generációs szekvenálási módszerekkel méréseket is végezhetünk, amelynek eredményeként szintén előáll a kívánt fázisos adat. Ebben az esetben hátrányként lehet említeni a magas költségeket, illetve a méréshez szükséges sok időt. Ezen vizsgálatoknál a kellő mintaszám előállítása jelenti a legfőbb gondot, ugyanis a vizsgálati személyen kívül a szülők mintájára is szükségünk van. Emellett az idős korban megjelenő betegségeknél nyilvánvalóan nem használható ez az eljárás. A harmadik lehetőséget a számítógépes algoritmusok jelentik, amelyek a nyers genotípus-adatból statisztikai módszerrel közvetetten állítják elő a haplotípusokat. Ekkor a rekonstrukció bizonytalansága jelenti a legnagyobb akadályt.

Jelenleg is sokan vizsgálják, hogy közvetlenül vagy közvetve érdemes-e előállítani a haplotípusokat. Általánosságban elmondható, hogy bár a közvetlenül előálló haplotípusokkal végzett elemzések erőteljesebbek, de a növekvő mintaszám, magasabb fokú genetikai kapcsoltság (linkage disequilibrium, LD) és kevesebb marker esetén a nyers genotípusokból megbecsült haplotípusok is megfelelően használhatóak.

2.4. Kapcsoltsági egyensúlytalanság

A kapcsoltsági egyensúlytalanságot leggyakrabban egy csúcsára állított korrelációs mátrix formájában ábrázolják. A mátrix átlójában az egyes markerek szerepelnek, és a pontos kapcsoltság leolvasható a sorok és oszlopok leolvasásával. A mátrix minden elemét a kapcsoltság erősségétől függően színezik.

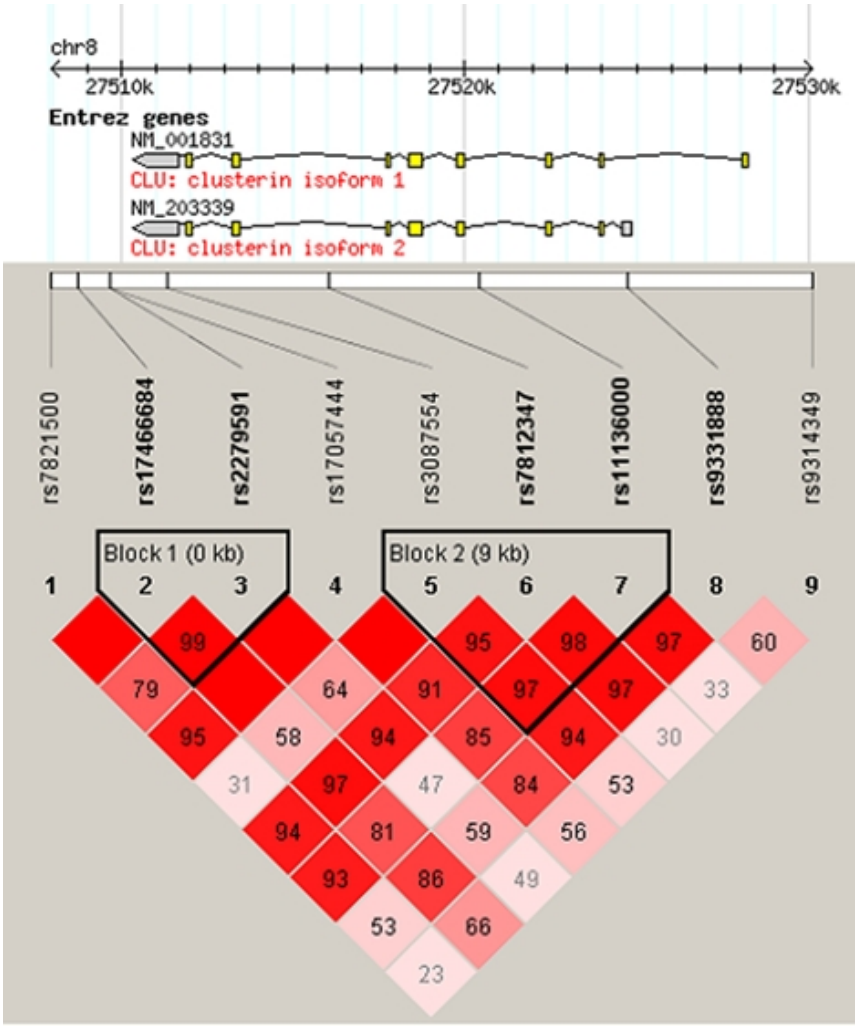


2.3. ábra. Kapcsoltsági egyensúlytalanság: Populáció 1-ben az A allél nagyon gyakori, D pedig véletlenszerűen fordul elő A és a mellett. Populáció 2-ben A nagyon gyakori, de D nem fordul elő. A kevert populációban az A allél frekvenciája 0.5 körüli, és D többnyire A-val együtt fordul elő. Ezt nevezzük kapcsoltsági egyensúlytalanságnak.

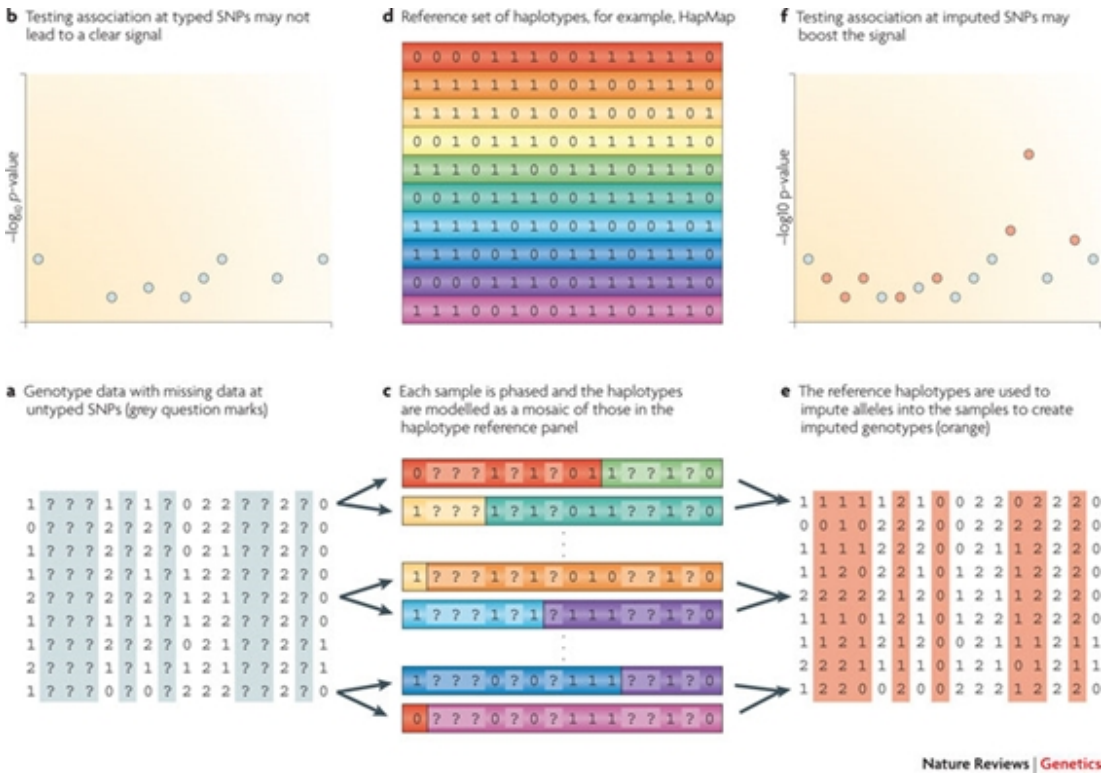
2.5. Haplotípus-rekonstrukció

A legtöbb SNP méréstechnológia nem ad lehetőséget arra, hogy haplotípusokat pontosan meghatározzunk, mivel csak diszkrét pontokon határozzák meg a genotípusokat, és nem képesek annak azonosítására, hogy az apai vagy anyai kromoszómákra vezethetőek-e vissza. A haplotípusok azonosítása azért fontos feladat, mert ha például két variánsnak csak akkor van a fenotípusban megjelenő hatása, ha egy szálon jelennek meg, akkor egy kettős (compound) heterozigóta egyed érintettsége csak a haplotípusok meghatározásával állapítható meg.

Több elterjedt megoldás is született a haplotípus-rekonstrukció problémájának megoldására, a legelterjedtebbek rejtett Markov-modelleken hajtanak végre következtetést. A legpontosabb és leggyakrabban használt módszer a PHASE, amelyik Gibbs-mintavételezéssel becsüli a lehetséges haplotípusokat, feltéve az ismert (megmért) genotípusokat és ismerve a rekombinációs rátát. A haplotípus-rekonstrukciós módszerek általában fel vannak készítve a hiányos adatok kezelésére.



2.4. ábra. Egy génen belüli kapcsoltságok haplotípus blokk ábrája

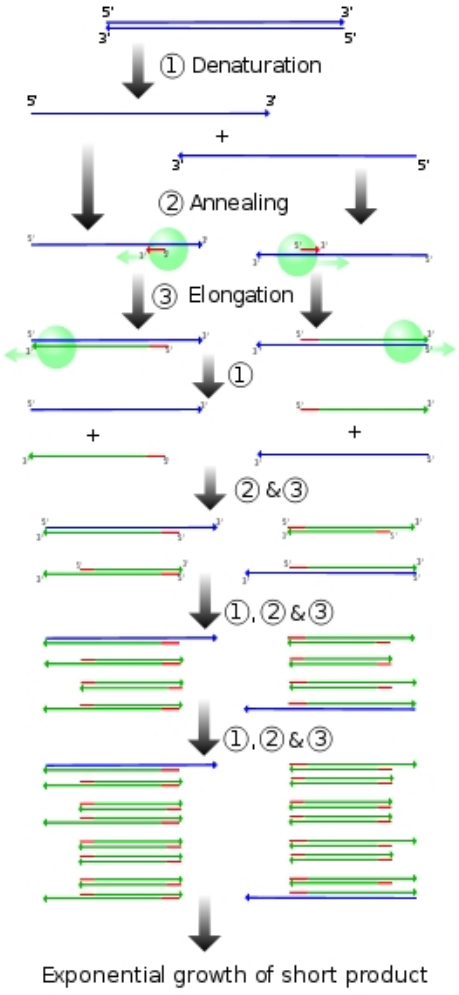


2.5. ábra. Hiányzó adatok imputálása haplotípus blokkok segítségével: A) Genotípusos adat hiányzó mérésekkel. B) A mért SNP-k vizsgálata nem biztos, hogy szignifikáns eredményt ad. C) Minden mintának modellezik a (gametikus) fázisát a referenciapanelben levők alapján. D) Haplotípus referencia, például a HapMap alapján. E) A referencia-haplotípusok segítségével imputálják (következtetve becsülik), előállítva a meg nem mért allélokat. F) Az imputált SNP-k vizsgálata növelheti az asszociációs vizsgálat statisztikai erejét.

2.6. Imputálás

Az SNP mérések eredményeit nemcsak a haplotípusok rekonstruálására használhatjuk fel a kapcsoltsági egyensúlytalanság segítségével, hanem a hiányos vagy esetleg alacsony megbízhatóságú mérések esetén adatpótlásra. Ez egy gyakori feladat genetikai asszociációs vizsgálatokban. A hiányzás mértéke gyakran 1–20% is lehet.

A hiányos adat megnehezíti a későbbi statisztikai elemzést, ezért fontos a rendelkezésre álló genotípus-információ maximalizálása. Az imputálás során külső adatforrásokat is felhasználunk, ideális esetben egy azonos populáción végzett nagyobb (akár teljes) genetikai asszociációs vizsgálat eredményét is.



2.6. ábra. A PCR ciklus lépései: 1. denaturálás, 2. primer hozzákötés, 3. elongáció

2.7. Genotipizálási módszerek

Több platform és módszer létezik a genotípusok meghatározására, amelyek mind áteresztőképességben, mind pontosságban jelentősen eltérnek egymástól.

A mérési módszerek nagy részében mindenképpen szükséges nehány mintaelőkészítési lépes. A polimeráz láncreakciót (PCR) arra használják, hogy felszaporítsanak egy cél DNS régiót a teljes genomból. Általában 100 és 10000 közötti bázis hosszúságú szakaszokat amplifikálnak. A reakció exponenciálisan növeli a PCR primer által megcélzott régió DNS koncentrációját. A felszaporított szakasz mennyiségére felső korlátot jelent a rendelkezésre álló szabad reakcióelegy mennyisége.

A PCR során 20-40 megismételt hűtési-melegítési ciklust hajtanak végre, és minden ciklusban az alábbi lépéseket hajtják végre:

2.7.1. Sanger-szekvenálás

A lánctermináló szekvenálás (más néven Sanger-szekvenálás) segítségével is meghatározható egy DNS szakasz pontos bázisszekvenciája, erről további információk elérhetőek a következő fejezetben. Nagy költségigénye és kis áteresztőképessége miatt nem terjedt el, felhasználása a genotipizáló eljárások kapcsán inkább a nagy megbízhatóságot igénylő, diagnosztikai területen jelentős.

2.7.2. Valós idejű kvantitatív PCR

Ez a módszer lehetővé teszi nem csak az SNP-k pontos azonosítását, hanem egy esetleges heterogén populációban az SNP-k arányának eloszlását is. A módszer az általános PCR lépéseit követi, azzal a különbséggel, hogy itt minden PCR ciklus között meghatározzák a keletkező DNS másolatok relatív mennyiségét is. Ez a módszer azonban nagy pontossága ellenére alacsony áteresztőképességű.

2.7.3. DNS chipek

A DNS chipek egy szilárd hordozó lapocskához (üveg, szilícium, speciális műanyag) kötött, nagyszámú, különböző nukleotidszekvenciájú DNS-próbából állnak. A próbák 25–50 nukleotid hosszúságúak oligonukleotidok. A jelenlegi technikai lehetőségek akár több millió, különböző próbát tartalmazó pont (spot) kialakítását teszik lehetővé egy chip 1-2 cm^2-es felületén.

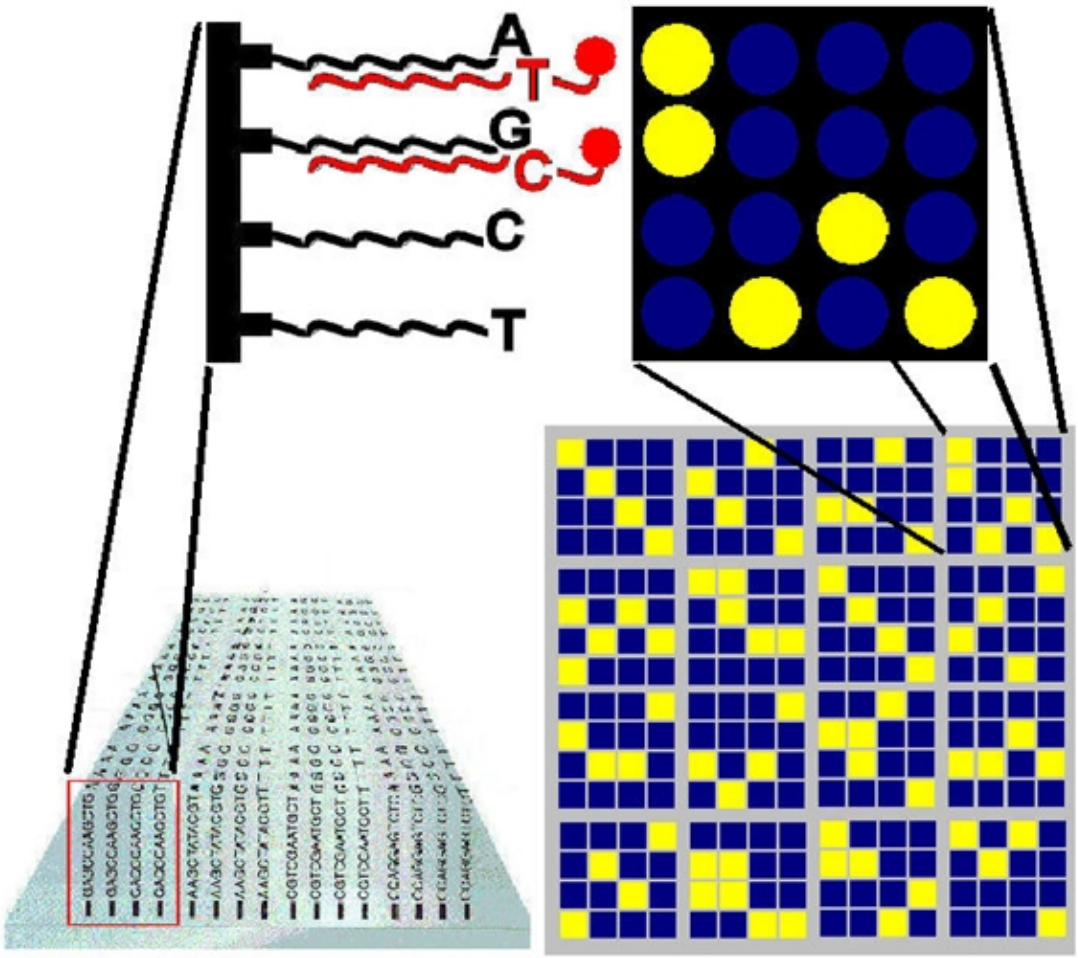
A vizsgálat során először a vizsgálandó mintákból DNS-t izolálunk, majd a számunkra érdekes területekről – az egynukleotidos polimorfizmusok 150 bázispárnyi környezetéről – másolatokat szaporítunk fel polimeráz láncreakció (PCR) segítségével.

Ezután a felszaporított vad és mutáns allélokat tartalmazó DNS láncokról különböző színű fluoreszcens festékkel jelölt kópiát készítünk. A jelölés úgy történik, hogy olyan primert adunk a PCR-el felszaporított DNS darabokhoz, amelynek 3' végi utolsó bázisa az SNP 5' irányú közvetlen szomszédságú bázisával képez párt. Ez után következik a primer 3' végének meghosszabbítása egyetlen fluoreszcensen jelölt, módosított nukleotid beépítése által (az SNP helyén előforduló nukleotidokkal homológ kétféleképpen jelölt aciklonukleotidokkal, melyek beépülése egyrészt megjelöli a primert a 3' végen, másrészt pedig a lánc további növekedését megakadályozza a módosított nukleotid). Az SNP határozza meg, melyik nukleotid kerül beépülésre. Így a vad, ill. mutáns allélek két különböző festékkel jelölhetők meg.

Az előkészítés befejeztével a mintákat olyan előre elkészített DNS chipre visszük fel, amelyen fizikailag kötve olyan DNS láncok találhatók, amelyek komplementerei a vad, ill. az SNP-t tartalmazó mutáns DNS szálaknak.

A minták felszaporított és festékkel megjelölt DNS szakaszai ezekhez a komplementer szálakhoz kötődnek (hibridizálnak). A nem kötődött szálakat mosással eltávolítjuk.

Ezután a két fluoreszcens festék elnyelési tartományának megfelelő hullámhosszú fénnyel (lézerrel) bevilágítva a mintákat – vagyis magát a DNS chipet – az eltérő fluoreszcens



2.7. ábra. Oligonukleotid SNP chip

festékkel megjelölt vad és mutáns allélokat tartalmazó DNS láncok a festékre jellemző hullámhosszú fényt fognak kibocsátani, amelyet detektálni tudunk.

Ekkor készítünk a két színcsatorna alatt egy-egy felvételt, majd a későbbiekben részletezett képfeldolgozási eljárással megfigyeljük az egyes pontok fényességét, valamint a pontok további jellemzőit is rögzítjük.

Ezután az egyes SNP-khez tartozó mintákat összegyűjtjük és egy diagramon ábrázoljuk. A diagram X tengelye a minta színarányát jelöli, az Y tengelyen a pontok összegzett intenzitása szerepel. Attól függően, hogy az adott mintapont a diagram mely oldalára kerül, megállapítható, hogy a vizsgált DNS tartalmazta-e a keresett mutációt vagy sem.

Végül több chipen elvégzett számos kísérlet eredményeit hierarchikus csoport- (cluster) analízis segítségével értékeljük.

Többféle eljárás is elterjedt DNS chipekkel történő SNP meghatározásra, a fentiekben egy lehetséges megközelítést ismertettünk.

2.8. Genotipizálás és génexpresszió

A genotipizálás során egy organizmus örökítőanyagának variabilitását térképezzük fel különböző polimorfizmusok meghatározásával. Fontos kiemelni, hogy ezek a mérések a kvalitatív jellegűek. A génexpressziós mérések ellenben kvantitatív jellegűek, mert itt az egyes génekről átíródó RNS szálak mennyiségét állapítjuk meg. A genotipizálás végeredménye egy konkrét genotípus, míg a génexpressziós mérésé pedig egy mért RNS koncentráció, amelyet sok külső paraméter befolyásol, például a vizsgált szöveti típus, valamint a minta izolálásának körülményei.

2.8.1. Sikeres mérések és pontosságuk

A sikeresen lemért SNP-k és az összes megmért SNP aránya a call rate. A pontosság pedig a sikeresen lemért SNP-kből azok aránya, amelyekhez a valós genotípust rendeltük hozzá. Általánosságban elmondható, hogy a magasabb áteresztőképességű mérési módszereknek alacsonyabb mind a sikeraránya, mind a pontossága. A diagnosztikai tesztekhez általában alacsony áteresztőképességű rendszereket alkalmaznak, mert itt a mérés várható hasznossága magasabb, és mindenképpen a pontosabb mérésre kell törekedni. A modern teljes genom asszociációs vizsgálatokban egyszerre több millió SNP-t is lehet mérni. Többféle általános hibajelenség tapasztalható a mérések során. Amennyiben a kiindulási DNS minta nem megfelelő mennyiségű vagy minőségű, akkor az összes hozzá tartozó SNP mérése sikertelen lehet. Ha az SNP-re jellemző primer nem eléggé specifikus, akkor pedig az összes mintán az adott SNP mérése lehet sikertelen.

3. fejezet

Összehasonlító fehérjemodellezés és molekuladokkolás

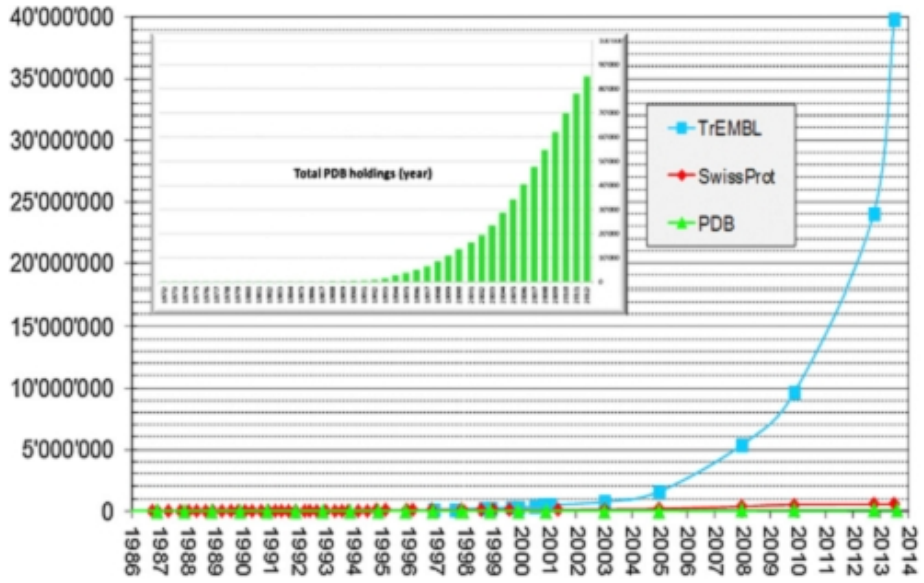
3.1. Bevezetés

A fehérjék szerkezetének meghatározása a molekuláris biológia és a szerkezeti genomika fontos kutatási területe. A fehérjék harmadlagos és negyedleges szerkezetének ismeretében a kutatók megismerhetik és elemezhetik a fehérjék funkcióját és aktív helyeit. Ez nagymértékben megkönnyíthet olyan fontos proteomikai feladatokat, mint például a fehérjemérnökség vagy szerkezet alapú gyógyszertervezés.

A kísérleti módszerek segítségével meghatározott szerkezeteket tartalmazó Protein Adatbank (PDB) [1] képezi az elsődleges alapját a szerkezet alapú proteomikai vizsgálatoknak. A fehérjeszerkezetek meghatározása különböző kísérleti módszerekkel (mint például a röntgensugár-krisztallográfia vagy NMR spektroszkópia, lásd „Fehérjeszerkezet-meghatározás kísérleti módszerei" fejezet) azonban továbbra is nehéz és költséges folyamat. Az emberi proteom mintegy 30.000 jellemzett humán fehérjét tartalmaz (a humán fehérjék referencia adatbázisában, Human Protein Reference Database) [2], de csak mintegy 5.000 humán fehérje vagy domén található a PDB-ben.

Ezért alakult ki igény olyan módszerekre, melyek lehetővé teszik háromdimenziós atomi szintű szerkezetek előállítását szekvencia-adatok alapján. E feladat megoldására olyan számítási módszerek alakultak ki, melyek alkalmasak a fehérje szerkezetének előrejelzésére elsődleges szerkezeti információk (pl. szekvencia adatok) felhasználásával [3, 4].

Az első fehérjeszerkezeti modell [5] megjelenése óta számos további fehérjemodellezési tanulmány is napvilágot látott. E fejezet célja a fehérjemodellezési technikák és a modellek pontosságának áttekintése. Modellezési módszerekre még akkor is szükség van, ha röntgen- vagy NMR-szerkezet áll rendelkezésre, mivel a szerkezetekben szükség lehet helyi javításokra vagy módosításokra (pl. a szerkezet alapú gyógyszertervezés során a nagyszámú lehetséges ligandum-receptor kombináció mindegyikének kísérleti szerkezetmeghatározása a gyakorlatban nem megvalósítható).

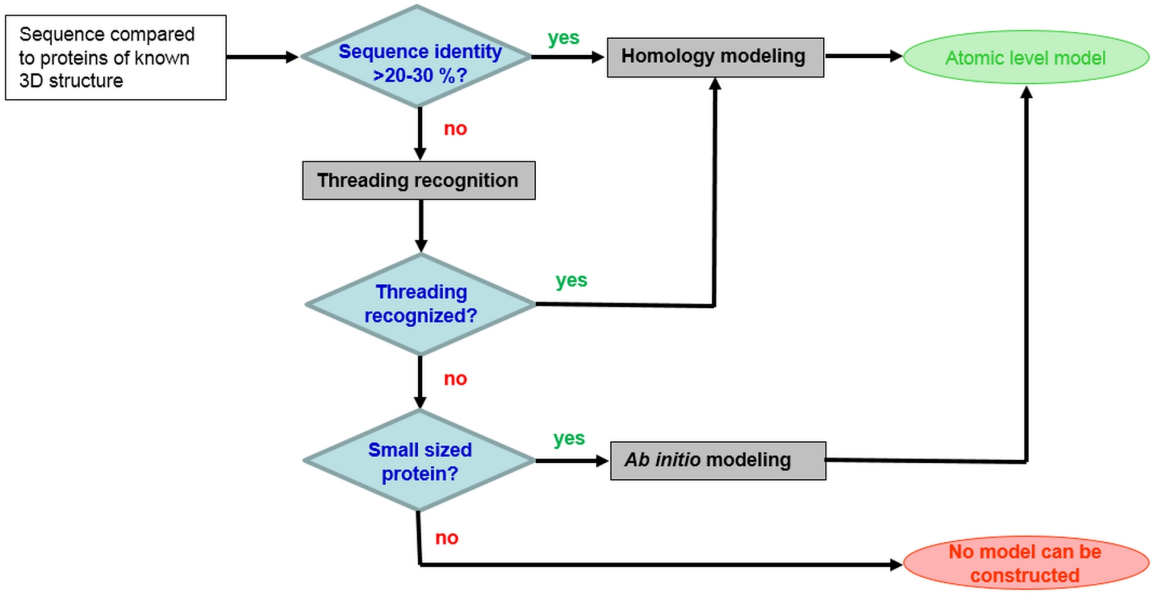


3.1. ábra. A szekvencia-szerkezeti szakadék. A SwissProt és trEMBL szekvencia-adatbázisok [6] és a PDB [1] rekordjainak száma exponenciálisan nő, ennek ellenére a fehérjeszerkezeti szakadék a szekvenciák és a szerkezetek között drámaian nő. Betét: a PDB szerkezetek számának növekedése 1972 és 2013 között. [A T. Schwede összefoglalójában [8] közölt ábra Elsevier kiadó által engedélyezett reprodukciója]

3.1.1. A fehérjeszekvencia-szerkezeti szakadék

A genom szekvenálási programok eredményeképpen ma már több ezer élő szervezet, így az ember teljes genetikai adatai (lásd a Genome adatbázist) ismertek. Az emberiség előtt álló feladat jelenleg e genomok fehérjéinek jellemzése, megismerése és akár módosítása. Ezt elsősorban a fehérjék háromdimenziós szerkezetének megismerése könnyítheti meg, amelyre a kísérleti módszerek (mint például a röntgensugár-krisztallográfia vagy NMR-spektroszkópia, lásd „Fehérjeszerkezet-meghatározás kísérleti módszerei" fejezet) a legalkalmasabbak. A kísérleti módszerek jelentős fejlődése ellenére sok fehérje szerkezetének kísérleti meghatározása azonban még mindig hiányzik különböző okokból.

Az elmúlt évtizedekben a nyilvános, nagy szekvencia-adatbázisokban, mint például az UniProt (SwissProt / TrEMBL) [6] vagy NCBI Gene [7] megtalálható szekvenciák száma hatalmas mértékben nőtt, ezek most közel 50 millió szekvenciát tartalmaznak. Ezzel szemben a szerkezeti genomika fejlődésének ellenére a kísérletileg meghatározott szerkezetek száma a Protein Adatbankban (PDB) lassabban nőtt, és most (2013 végén) is csak 95.000 körüli szerkezetet tartalmaz. Az ismert szekvenciák és szerkezetek száma közti különbség továbbra is növekszik (3.1. ábra) [8]. Ezt a szakadékot próbálják áthidalni a fehérjeszerkezet-előrejelzési módszerek [4, 9].



3.2. ábra. Hogyan válasszunk fehérjemodellezési módszert? Templát alapú modellezés esetén azonosítani kell a homológiamodellezést lehetővé tevő templátot (akár a > 20–30%*-os szekvenciaazonosság, akár hajtogatásfelismerés alapján). Templátmentes* ab initio *modellezés olyan kisméretű fehérje esetében használható, ahol nem lehetett megfelelő templátot azonosítani.*

3.1.2. A fehérjemodellezés módszerei

A fehérjeszerkezetek szekvencia-adat alapú, atomi szintű modellezésére alkalmas módszerek a célfehérje méretétől, valamint a vizsgálandó fehérje és egy homológ, kísérletileg meghatározott szerkezetű fehérje közötti szekvenciaazonosság fokától függnek (3.2. ábra).

A módszerek első csoportja, az úgynevezett *ab initio* (vagy *de novo*) fehérjemodellezés a szerkezetet kizárólag a szekvenciából jósolja meg, anélkül, hogy a modellezett szekvencia és bármilyen ismert szerkezet közötti hasonlóságra támaszkodna [10]. Ezek a módszerek a fehérje 3D modelljének „a semmiből", vagyis a fizikai elvek alapján történő megoldására törekszenek előzetesen megoldott szerkezeti adatok felhasználása nélkül. A de novo módszerek feltételezik, hogy a natív szerkezet megfelel a fehérje globális szabadenergia-minimumának, amit sok megvalósítható fehérjekonformáció előállításával és vizsgálatával próbálnak megtalálni. A de novo módszerek két fő eleme a hatékony konformációkereső eljárás, és a lehetséges konformációk szabadenergia-függvény kiértékelési jósága. Ezek az eljárások általában hatalmas számítási erőforrásokat igényelnek, és így csak kisméretű fehérjék esetében alkalmazhatóak.

A fehérjeszerkezet-modellezési módszerek második osztálya az összehasonlító fehérje modellezése (vagy homológiamodellezés). Ez a számítási módszer a fehérje szerkezetét annak aminosav-szekvenciája és egy azzal homológ, kísérletileg meghatározott szerkezetű templát segítségével nyeri, összehasonlító hajtogatással és modellezéssel. A módszer alapja

az a megfigyelés, hogy a fehérjék 3D-s szerkezete jobban konzerválódott, mint szekvenciáik, és ezért két, szekvenciaszinten csak részben azonos fehérje hajtogatása még mindig ugyanaz lehet [4, 11].

Atomi felbontású modell építése csak akkor megvalósítható a „cél"-fehérje aminosav-szekvencia és egy rokon, homológ „templát"-fehérje egy kísérleti háromdimenziós szerkezete segítségével, ha a „cél" – és a „template"-fehérje közötti szekvenciaazonosság meghaladja a 20–30%-ot. Mivel 20%-os szekvenciaazonosság alatt rendkívül eltérő szerkezetek lehetségesek, ezért homológiamodellezés csak akkor valósítható meg, ha felismerhető az adott szekvenciának megfelelő hajtogatás. Ha van ilyen hajtogatás, ebből homológiamodellt lehet kialakítani.

3.2. Összehasonlító fehérjemodellezés

Az *ab initio* fehérjeszerkezet-előrejelzésben történt haladás [10] ellenére az összehasonlító fehérjemodellezés továbbra is a legmegbízhatóbb módszer a fehérjék atomi szintű 3D szerkezet-előrejelzésére. Sok esetben homológiamodellezéssel nyert szerkezetek pontossága összevethető a kísérletileg meghatározott kisfelbontású szerkezetekével. Emiatt vált mára az összehasonlító fehérjemodellezés a fehérjék atomi szintű szerkezet-előrejelzésének elsődleges eszközévé [4, 11].

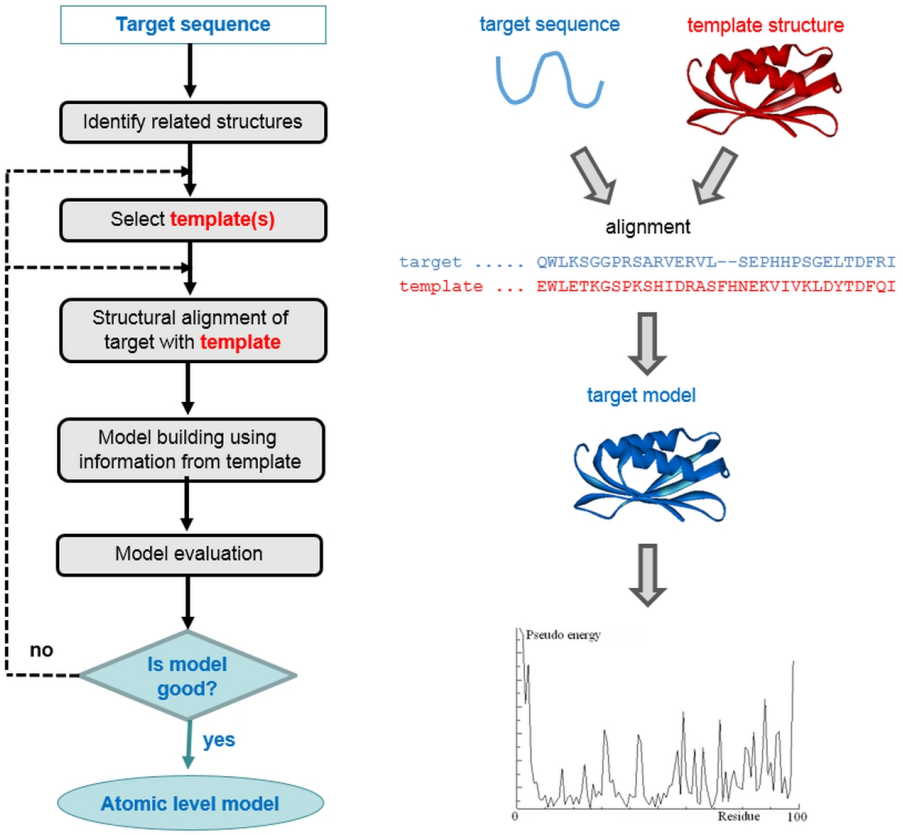
3.2.1. A homológiamodellezés lépései

Az összehasonlító fehérjemodellezés főbb lépései a templát kiválasztása, összerendelés, főlánc-, hurok- és oldallánc-előrejelzés, szerkezet optimalizálása és értékelés (3.3. ábra).

A megfelelő templát(ok) választása igen fontos, mivel nem megfelelő „templát" hibás modellhez vezet. Ezért a „cél"-szekvenciával kellő fokú hasonlóságot mutató „templát"-ot különös gonddal kell azonosítani. Egyes esetekben még alacsony szekvenciahomológiájú „templát"- és a „cél"-szekvenciák esetében is felismerhető a hajtogatás. A „cél"-szekvenciát ezután összerendezzük a „templát"-szekvenciával, majd ezt úgy finomítjuk, hogy a homológ régiók optimális egyezést mutassanak. Miután elértük az optimális összerendezést, a „cél"-szerkezet főlánc atomjait a „templát" 3D-szerkezetére modellezzük, majd előre jelezzük a nem megfelelő hurokrégiókat és a nem konzervált oldallánc-elrendeződéseket. Ezután megfelelő erőtérrel végzett optimalizálással eltávolítjuk a modellből sztérikus ütközéseket, és javítjuk a szerkezeti szempontból fontos kölcsönhatásokat, mint az atomok közötti hidrogén-híd hálózat. Ezután a végső modellt elsősorban a hibás vagy hiányzó régiók szempontjából értékeljük (pl. a nem-konzervált hurkok, amelyeket általában a konzervált régióktól függetlenül szükséges modellezni). Az értékelés a végső minőség eléréséig a modell iteratív finomítását eredményezheti.

Templátkiválasztás és kezdeti illesztés

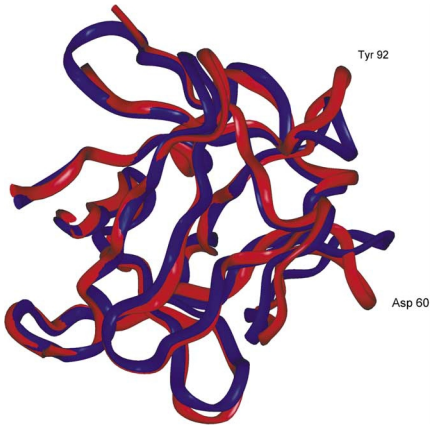
Az összehasonlító homológiamodellezés kezdeti lépése a megfelelő templát(ok) kiválasztása. Az egyszerű helyi összerendezés keresésén alapuló (BLAST) eszközök [13] révén váltak



3.3. ábra. Az összehasonlító fehérjemodellezés lépései. A modellezés további elemi lépéseinek részleteit a szöveg mutatja be. (Az ábra Fiser és mtsai. összefoglalójának [12] ábrája alapján készült.)

évtizedekkel ezelőtt a szekvencia-adatbáziskeresések hatékonyan automatizálhatóvá. Ilyen eszközökkel választhatjuk ki a szerkezeti adatbázisokból (pl. a PDB) a templátot (még a hajtogatás felismerése esetén is) a további modellezési lépések előtt. Mind a hagyományos, mind a hajtogatásfelismerésen alapuló homológiamodellezés eredménye erősen függ e keresés eredményétől.

Első közelítésként a legnagyobb szekvenciaazonosságú találatot választhatjuk templátként. Ne feledjük, hogy még a röngenkrisztallográfiával nyert fehérjeszerkezetek sem tökéletesek (a kristályosodás közbeni részleges bomlás, az alacsony felbontású elektronsűrűség-térkép, vagy egyszerűen csak az emberi hibák miatt, lásd még „Fehérjeszerkezet-meghatározás kísérleti módszerei” fejezet) [14]. Több, mint egy szerkezeti találat esetén kézenfekvő megoldás, hogy (pl. a PDBREPORT szerint) a legkevesebb hibát tartalmazót választjuk templátként. Ezen kívül más szempontokat (egy fehérjének lehetnek aktív és inaktív szerkezetei; kofaktorok/ligandumok jelenléte fontos lehet a szerkezetben stb.) is figyelembe kell venni a templát kiválasztása során. A napjainkban rendelkezésre álló számítási kapacitás lehetővé teszi több templát használatát is, és a legjobb eredményt



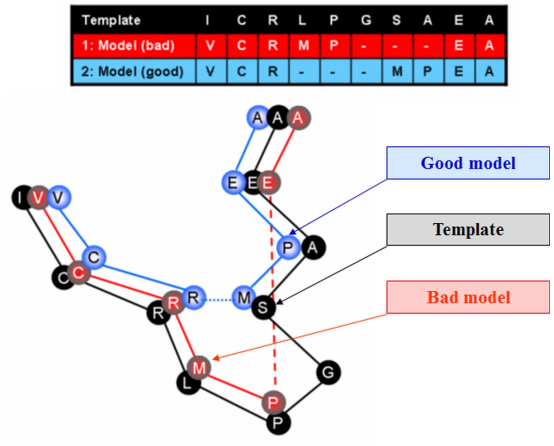
3.4. ábra. Az egyszerű fibroblaszt növekedési faktor (bFGF) kísérleti szerkezete és elméleti modellje. A kísérleti szerkezet (PDB kód 1BFC) kék szalagként, míg az elméleti modell piros szalagként látható. A modell és kísérleti szerkezet közti legnagyobb eltérést mutató két régiót a jelölt aminosavak jelzik. A modellt a nyilvánosan hozzáférhető Swiss-Model szerver felhasználásával készítették. [Az MJ Forster összefoglalójában [15] közölt ábra Elsevier kiadó által engedélyezett reprodukciója]

kiválasztását a további finomításra. Több templát kombinációjával nyert átlagszerkezet segítségével is megvalósítható modellezés. Azok az esetek, amikor egy „templát"-szerkezet több mint 25%-os szekvenciaazonosságot mutat a „cél"-szekvenciával, képviselik azt a szintet, amely felett a homológiamodellezést sikerrel meg lehet kísérelni. Ezt demonstrálja az egyszerű fibroblaszt növekedési faktor (bFGF) modellezését bemutató 3.4. ábra is. A bFGF homológiamodellje a patkány keratinocita növekedési faktor (PDB-kód: 1QQK, lánc B) 41%-os szekvenciaazonosságú (53% hasonlóság) templátszerkezete alapján készült. A modellszerkezet (piros szalag) és a később meghatározott kísérleti szerkezet (kék szalag, PDB kód 1BFC) fehérjefőláncai láthatóan igen hasonlóak, két kevésbé egyező régiótól eltekintve.

Szekvenciaillesztés finomítása

A templát kiválasztása és a kezdeti összerendezése után számos eszköz áll rendelkezésre a „modell"- és a „templát"-szerkezeti illesztések kiválasztására és finomítására, beleértve a három-dimenziós szerkezetmegjelenítési és szerkesztési eszközöket is. Manapság csak néhány eszköz képes a szekvenciaillesztések problémáinak automatikus finomítását kezelni, de ígéretes módszereket is közöltek [16].

Egy adott összerendezés jósága ellenőrizhető a „templát"-, illetve a „cél"-szekvenciához elegendően hasonló új szekvenciák, vagy más, a templát szerkezetére jól illeszkedő kísérleti szerkezetek hozzáadásával. Távoli rokonságban álló fehérjék esetében az is fontos, hogy ellenőrizzük a „cél"-szekvencia másodlagos szerkezet-előrejelzéseinek egyezését a templát másodlagos szerkezetével [17]. Ezek a szerkezeti összerendezés-adatok megjeleníthetők a



3.5. ábra. A modellszekvencia és a templát szerkezeti összerendezésének háromdimenziós értékelése. A kísérleti templátszerkezet főláncának (fekete) összerendezését láthatjuk egy az E és P aminosavak közti megfelelő távolságú jó modellel (kék) és egy, az E és P aminosavak közti túl nagy távolságot eredményező és így rossz modellel (piros).

JOY formátum segítségével [18]. Kisebb mértékű szekvenciakonzerváltság esetén a szerkezeti összerendelést pontosabban lehet elvégezni háromdimenziós szinten (3.5. ábra). Az összerendelés kézi szerkesztése az összehasonlító fehérjemodellezés legidőigényesebb és legkritikusabb része. A modellben akár egyetlen aminosavnyi elcsúszás is a végső szerkezet mintegy 4 Å-ös hibáját eredményezi, mivel a jelenlegi homológiamodellezési algoritmusok általában nem képesek kiigazítani az összerendezés során elkövetett hibákat [19].

Fehérjefőlánc modellezése

A szekvenciaillesztés végeztével következik a főlánc modellezése. A főlánc generálása a legtöbb modell esetében triviális: a templátszerkezet összerendezésben szereplő aminosavainak főláncbeli atomkoordinátáit egyszerűen át kell másolni a modellbe.

Ha egy bizonyos helyzetben a modell és templát összerendezésében az aminosavak eltérnek, akkor csak a főlánc N, C_α, C és O koordinátái (és egyes esetekben a C_β is) másolható. Ha egy adott pozícióban az aminosavak megegyeznek, sok esetben még az oldallánc atomkoordinátái is a modellbe másolhatóak.

Hurokmodellezés

A modell- és a templátszerkezet összerendezése beszúrásokat és törléseket is tartalmazhat. Törlések esetén egyszerűen kihagyjuk a templát felesleges részeit, és a képződő hiányt összekötjük. A beszúrások esetén a templát folyamatos láncát elhasítjuk, majd beszúrjuk az extra aminosavak alkotta hurkot. Belátható, hogy mindkét eset a főlánc konformációváltozásával jár.

Amikor beszúrások vagy törlések vannak a templát/célszerkezet összerendezésében, a hiányzó részek modellezésének pontossága a fehérje különböző részein jelentősen eltérő. A jól definiált másodlagos szerkezeti elemek (α-hélixek és β-szálak) esetén, ahol a merev főláncközelítés általában elfogadható, a modellezés pontosabb, míg kisebb pontosság várható a kevésbé strukturált, így mozgékonyabb hurkok esetében. Sok homológiamodellezési módszer képes hurkok elfogadható kovalens geometriával történő modellezésére, jellemzően hurok-adatbázisbeli keresésekkel. A natív hurokkonformációkkal közel megegyező szerkezetek modellezése azonban nehéz, és megfelelő templát hiányában következetesen a hurkok a homológiamodellek leginkább pontatlan részei [20].

Oldallánc-modellezés

Az oldallánc-modellezés nehézsége egyéb tényezők között erősen függ a cél és a templátszekvencia hasonlóságának fokától és a templátszerkezet minőségétől is. Hasonló fehérjék esetén gyakori, hogy a C_α–C_β torziós szögek is megegyeznek. Sőt, erősen homológ ($> 40\%$-os szekvenciaazonosság) fehérjéknél gyakran (kb. 75% esetben) még a C_γ is hasonló orientációjú.

Következésképpen magas szekvenciaazonosság ($> 40\%$) esetén a konzervált aminosavak gyakran teljesen átmásolhatóak a templátból a modellszerkezetbe. Sok esetben ez a megközelítés pontosabb, mint a főláncatomok átmásolása és oldalláncok *ab initio* módszerekkel való előrejelzése.

Azonban ha szekvenciaazonosság alacsony ($< 35\%$), az oldalláncok a modellek és a templátok 45%-ában különbözőek. Ezekben az esetekben az oldallánc-orientáció modellezése szükséges. A legtöbb, oldallánc előrejelzésére rendelkezésre álló eszköz tudás alapú könyvtárakra támaszkodik. Ezek sok esetben *„fix" könyvtárakat* alkalmaznak, amelyek egy adott oldallánc összes lehetséges állását tárolják. Más módszerek *„helyzetspecifikus"* könyvtárakat használnak, és az oldallánc állását a főlánc szerkezete/konformációja szerint választják ki. Ezek egyszerű változatai az oldallánc-elrendeződéseket a főlánc másodlagos szerkezete (hélix vagy redő) alapján osztályozzák, míg a kifinomultabbak az oldallánckonformációkat a megfelelő, nagy felbontású szerkezetekben találhatóak (5–9) közül választják ki az eltérő főláncgörbületeknek megfelelően.

Az oldallánc-konformáció előrejelzése általában pontosabb a belső, hidrofób részeknél, mint a felszíni oldalláncok esetében. Ez annak a ténynek köszönhető, hogy a mozgékony hurkok oldalláncai – amelyek többnyire a felszínen vannak jelen – többféle konformációt vehetnek fel.

Modelloptimalizálás

A főlánc templátszerkezethez képesti beszúrásokkal és törlésekkel való kiegészítése és az oldallánc-modellezés után a modellszerkezet normalizálásához további lépésekre van szükség, főleg a beszúrások és törlések közelében (lásd 3.2.1. fejezet). A megfelelő erőterekkel végzett molekulamechanikai energiaminimalizálás eltávolíthatja a súlyos van der Waals ütközéseket és javíthatja a kötéshossz- és vegyértékszög-értékeket is. Ez azonban nem

hozza közelebb az atomokat tényleges helyzetükhöz. Az energiaminimalizálások az energiafelület szabálytalansága miatt könnyen megragadhatnak a helyi minimumokban. Az energiaminimalizált szerkezetek tehát gyakran mutatnak kis mértékben megnövekedett globális strukturális eltérést a nem minimalizált modellekhez vagy a kiindulási templátokhoz képest.

Az energiaminimalizálás mellett trajektóriaszimuláció (molekuladinamika, MD) is végezhető hasonló erőterekkel. Az MD-módszerek hasznosak lehetnek a konformációs tér feltérképezésére. A trajektória különböző pontjain vett mintákkal további, a kiindulási modellekkel megegyező jóságú modellek nyerhetőek [21]. Az MD-elemzés alkalmas lehet a modellek pontosságának (vagy hibájának) jellemzésére is.

Modellértékelés

A háromdimenziós szerkezetek értékelése különböző szintű pontosságot igényelhet. Magas szekvenciaazonosság ($> 50\%$) esetén a valós koordinátáktól csak kisebb mértékben eltérő szerkezetek nyerhetőek, így az értékelésre alkalmasak lehetnek a kísérleti szerkezetek esetében használható eszközök (pl.: WHAT-CHECK [14]). Kisebb szekvenciaazonosság (25–50%) esetén a modell általános minősége nem korrelál, eltérések lehetnek a normál sztereokémiától (különösen energiaminimalizálás után, lásd 3.2.1. fejezet). A nem-kötő atomi kölcsönhatások értékelésére atomi statisztikai potenciálok, például ERRAT [22], ANOLEA alkalmasabbak lehetnek. A modellezési eredmények értékelésére további hasznos eszközök a ProSA [23] és Verify3D [24].

25% alatti szekvenciaazonosságok esetében a modell értékelését inkább aminosavanként kell elvégezni. Egyes esetekben pontos helyi elemzésre lehet szükség. A háromdimenziós szerkezetértékelő pontszámokkal egyszerre történő megjelenítése hasznos lehet. Egyedi értékeket figyelhetünk meg az aktív helyek (vagy kötő helyek) vagy ionokkal érintkező (különösen fémek koordinációjában részt vevő) oldalláncok és/vagy mélyen eltemetett ligandumok (különösen a kofaktorok) környezetében, mert ilyen esetekben az aminosavoldalláncok nem-klasszikus környezetben vannak. Hasonlóan egyedi sajátságok figyelhetőek meg hőstabil fehérjék esetében, amelyeket eltemetett környezetben lévő sóhidak stabilizálhatnak. Ha ilyen sajátságokat észlelünk, a modell minőségi értékelése kiterjeszthető a templát szerkezetének értékelésére is.

3.2.2. Homológiamodellezési eszközök

Napjainkban számos módon végezhetőek összehasonlító modellezési feladatok. Homológiamodellező eszközök léteznek önálló (mind kereskedelmi, mind szabadon felhasználható) programokként, valamint automatizált, Web alapú szolgáltatásokként is, amelyek ezeket a technológiákat elérhetővé teszik a bioinformatikában nem szakértő közönség számára is.

Web alapú homológiamodellező eszközök

Csaknem két évtizeddel ezelőtt vált az Interneten elérhetővé az első automatizált modellező szerver, a SWISS-MODEL [25].

Az az igény volt a homológiamodellezési lépések [templátkiválasztás, cél-templát összerendezés, modellezés és a modell minőségértékelés (3.3. ábra)] automatizálásának fő hajtóereje, hogy váljanak ezek a technológiák nyilvánosan elérhetővé a szélesebb közönség számára is. Azóta számos további, fehérjék automatizált homológiamodellezését lehetővé tevő eszközöket kínáló szolgáltatás jött létre [26].

A következő részben az on-line rendelkezésre álló összehasonlító fehérjemodellező eszközök listáját mutatjuk be.

SWISS-MODEL. Teljesen automatizált homológiamodellező szerver (elérhető az ExPASy Web-oldalról, vagy a DeepView – Swiss-PdbViewer programból).

ModWeb. Proteinmodellező szerver. (A MODELLER programot használja; licenszkulcs szükséges.)

Robetta. Rosetta homológiamodellező szoftvert használó Web-szerver (ab initio fragmens-összeállítás Ginzu domén predikcióval).

HHpred. A HHpred szerver a templát alapú szerkezetmodellezések egyik legjobbjának bizonyult (No 1 szervernek ítélve a CASP9 során).

I-TASSER. Web-szerver fehérjeszerkezet és funkció predikciójához. A modellek LOMETS által végzett többszörös szerkezeti összerendezések és iteratív TASSER szimulációk segítségével készülnek. (No 1 szervernek ítélve a CASP8 és CASP10 során.)

Phyre^2. Fehérjehomológia/analógia felismerés (Protein Homology/analogY Recognition Engine).

M4T. Összehasonlító modellező szerver, többszörös templáttechnika, iteratív optimalizálás és alternatív összerendezések ötvözésével.

3D-JIGSAW. Proteinek 3D modelljeit építő szerver ismert szerkezetű homológok felhasználásával és fragmens alapú modellezéssel.

RaptorX szerkezet predikció. Web-szolgáltatás másodlagos szerkezet, oldószer elérhetőség, rendezetlen régiók és harmadlagos szerkezetek előrejelzésére szekvencia alapján. (Kifejezetten alkalmas fehérjeszekvenciákból 3D szerkezetek előrejelzésére közeli homológok nélkül. RaptorX csomag formájában is elérhető.)

QUARK. On-line szolgáltatás, elsősorban megfelelő templát nélküli szerkezetek modellezésére (ab initio fehérjehajtogatás és fehérjeszerkezet-predikció. No 1 szervernek ítélve a templátmentes modellezésben (FM) a CASP9 és CASP10 során).

GeneSilico Metaserver. Hozzáférést biztosít különböző fehérjeszerkezeti előrejelzési módszerekhez: elsődleges szerkezet, másodlagos szerkezet, transzmembrán hélix, rendezetlen régiók, diszulfid kötések, fehérjék nukleinsavkötő helyei, harmadlagos szerkezet.

Proteinmodell-adatbázisok

Ez a fejezet olyan nyilvánosan elérhető adatbázisokat sorol fel, amelyek proteinmodellezési módszerekkel elkészített fehérjemodell-szerkezeteket gyűjtenek össze.

SWISS-MODEL Repository. Leírásokkal ellátott fehérjeszerkezeti modellek, melyeket automatizáltan készítettek az összehasonlító modellezést végző SWISS-MODEL szerverrel.

ModBase. Leírásokkal ellátott fehérjeszerkezeti modellek adatbázisa, melyeket a modellező automata ModPipe (valamint a PSI-BLAST és MODELLER programok) segítsé-

gével készítettek.
(További adatok hajtogatás-hozzárendelésről, feltételezhető ligand-kötőhelyekről és protein-protein kölcsönhatásokról.)

Protein Model Portal (PMP). Hozzáférést biztosít különböző összehasonlító modellezési módszerekkel partneroldalak által számított modellekhez, és elérhetővé tesz különböző modellépítésre és értékelésre alkalmas interaktív szolgáltatásokat.

A homológiamodellezés szoftverei

MODELLER. Szoftver fehérje-homológiamodellek előállítására térbeli korlátozások legjobb kielégítésének felhasználásával. Ingyenes tudományos használatra. Kereskedelmi változata grafikus felhasználói felülettel elérhető az Accelrys-től.

ProModel. Szoftveregyüttes homológiamodellezéshez akár egy kiválasztott templát, akár a felhasználó által megadott templát segítségével. Modellezés kézi üzemmódban (mutáció, kimetszés, törlés, beillesztés vagy hurokbeillesztés), vagy automata módban. A célfehérje szerkezetének, aktív helyének és csatornáinak elemzésére alkalmas. Elérhető a Vlife-tól.

Prime. Teljesen integrált fehérjeszerkezeti előrejelzés-program grafikus felülettel: szekvenciaillesztés, másodlagos szerkezet előrejelzése, homológiamodellezés, proteinfinomítás, hurok- valamint oldallánc-előrejelzés. A Schrödinger cég fejlesztése.

DeepView – Swiss-PdbViewer. Önálló programegyüttes, amely együttműködik az ExPASy web site teljesen automatizált SWISS-MODEL homológiamodellező szerverével.

TASSER-Lite. Fehérjeszerkezetet összehasonlító modellező eszköz, csak a célprotein/templát párok $> 25\%$ szekvenciaazonossága esetén működik. Egydoménes, 41–200 aminosav hosszúságú fehérjék modellezésére optimalizált. Non-profit használatra ingyenes.

Rosettahome. Önálló program a Rosetta algoritmus használatára (ab initio fragmens összeállítás Ginzu domén becsléssel). Csak nem kereskedelmi használatra.

Rosetta CM. A Rosetta kiváló szoftvercsomag makromolekuláris szerkezetek modellezésére. Rugalmas, többcélú alkalmazás, amely a fehérjék és nukleinsavak szerkezet-előrejelzésére, tervezésére és átalakítására alkalmas eszközöket tartalmaz. Nem kereskedelmi használatra ingyenes.

Molide. Nyílt forráskódú, többplatformos grafikus környezet homológiamodellezésre. Alkalmas a modellezés leggyakoribb lépéseinek megvalósítására. Nem kereskedelmi használatra ingyenes.

3.3. Molekuladokkolás

Ha egy fehérje atomi szintű háromdimenziós szerkezete elérhető, vizsgálhatóvá válnak olyan jellemzői, mint alakja, felületi tulajdonságai, üregek jelenléte. A fehérje saját tulajdonságainak vizsgálata mellett az adott fehérje más molekulákkal (mint például különböző kisméretű ligandumok vagy más biológiai makromolekulák, fehérjék vagy nukleinsavak) történő kölcsönhatásaira vonatkozó információk is igen fontosak.

A molekuláris modellező eszközök közül a molekuladokkolás olyan módszer, amely megjósolja egy molekula (általában egy ligandum vagy akár egy biológiai makromolekula) előnyös elrendeződését egy másikhoz (általában egy biológiai makromolekula) kötődve alkotott stabil komplexében. Az előnyös elrendeződés ismeretében a két molekula közötti asszociáció vagy kötéserősség becsülhető. Ezek az adatok felhasználhatóak például funkció-előrejelzések, enzimmechanizmus-vizsgálatok, in silico gyógyszertervezés vagy rendszerbiológiai vizsgálatok során.

A dokkolási módszereket két osztályba sorolhatjuk [27]: i) az egyik empirikus értékelést alkalmaz, így gyorsabb; ii) a másik szabadenergia-számításokat használ, így nagyobb számításigényű. Az első megközelítés a térbeli megfelelés technikáját használja, a célfehérjét és a dokkoló molekulát egymást kiegészítő felületekként kezeli. A második megközelítés a tényleges dokkolási folyamatot szimulálja a célfehérje-dokkoló molekula páronkénti kölcsönhatási energiáit számítva. Egy adott dokkolóprogram sikeressége két fő tényezőtől függ: a keresési algoritmustól és az értékelő módszertől [27].

A ligandumra/célmolekulára különböző keresési stratégiák alkalmazhatóak, mint például

i) szisztematikus vagy sztochasztikus torziós keresések elforgatható kötések körül;
ii) molekuladinamikai szimulációk vagy
iii) genetikus algoritmusok új, alacsony energiájú konformációk „evolúciójára”.

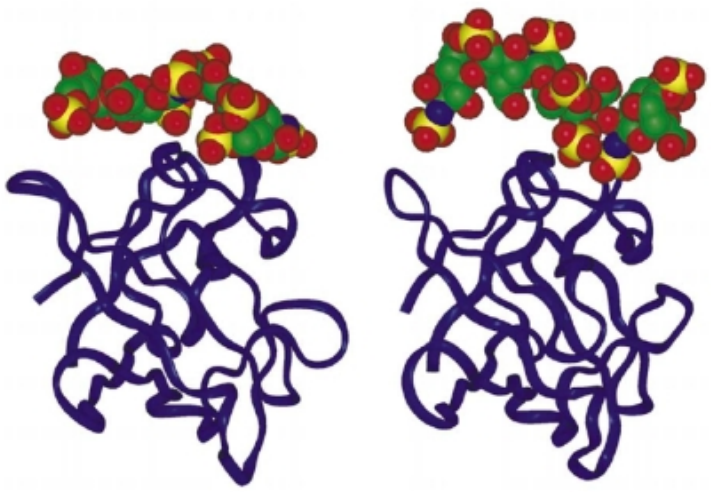
A dokkoló molekula természete szerint is osztályozhatjuk a dokkolási módszereket:

i) fehérje/kismolekula;
ii) fehérje/peptid;
iii) fehérje/fehérje vagy
iv) fehérje/nukleinsav dokkolás.

3.3.1. Fehérje–ligandum kölcsönhatás-előrejelzések

A molekuláris felismerés kulcsfontosságú szerepet játszik az alapvető biomolekuláris történések, mint például az enzimszubsztrát, a gyógyszerfehérje és gyógyszernukleinsav kölcsönhatások során. A fehérje-ligandum dokkolás alkalmas molekuláris modellező eszköz ilyen kölcsönhatások tanulmányozására [28]. A 3.6. ábra azt mutatja, hogy dokkolási módszerek még akkor is sikeresen alkalmazhatóak, ha nem áll rendelkezésre kísérleti fehérje szerkezete.

A dokkoló módszerek a ligandum és a célfehérje flexibilitásától függően különbözhetnek [27]–[29]. A legtöbb dokkoló módszer lehetővé teszi a ligandum flexibilitását, és annak több konformációját is figyelembe veszi. Ezzel szemben a jelenleg használt dokkolási módszerek többsége a célfehérjét egy adott konformációban rögzítettként kezeli. Ezt a megközelítést általában a sebesség és egyszerűség miatti megfontolásból alkalmazzák, elkerülve ezzel a kötőhely flexibilitásának pontos kezelésével járó jelentősen megnövekedett számításigényt. Vannak fehérjeflexibilitást megengedő sikeres erőfeszítések is, ezek pontatlanabb módszerek javításában segíthetnek (pl. pontosíthatóak a receptormodellekbe történő dokkolások).



3.6. ábra. A bFGF/heparin komplex kísérleti szerkezet (jobbra) és egy dokkolási módszerrel nyert modell (balra) összehasonlítása. Ez a dokkolási probléma komoly tesztje, mivel a dokkoláshoz használt fehérjeszerkezet nem kísérleti szerkezet, hanem homológiamodell. Emellett a dokkolás során használt heparin-próbamolekula a modellben egy pentaszacharid, míg az ismert szerkezetű komplexben hexaszacharid. Ez jelzi, hogy a növekedési faktorok heparin kötőhelyeit általános próbamolekulák és fehérje-homológiamodellek segítségével is azonosítani lehet. [Az MJ Forster összefoglalójában [15] közölt ábra Elsevier kiadó által engedélyezett reprodukciója]

Számos, több lehetőséget kínáló dokkoló eszköz áll rendelkezésre, a kis ligandumok merev fehérjékbe történő egyszerű dokkolásától a flexibilis ligandum / flexibilis kötőhely párosítást akár fehérje-fehérje kölcsönhatások esetében is megengedőkig. Ezek például az AutoDock, DOCK, Gold, FlexX, VLifeDock, and ArgusLab. AutoDock, DOCK, Gold, FlexX, VLifeDock vagy az ArgusLab.

3.3.2. Fehérje–biomakromolekula kölcsönhatás-előrejelzések

Dokkolási módszerekkel fehérjék és további biomakromolekulák kölcsönhatásai is vizsgálhatóak. Bár a fehérje-fehérje [29] vagy a fehérje-nukleinsav [30] dokkolás is megvalósítható, a legsikeresebb megközelítések az ilyen dokkolásokat további kísérleti adatok – pl. NMR vagy elektronmikroszkópia (lásd „Fehérjeszerkezet-meghatározás kísérleti módszerei" fejezet) – felhasználásával egészítik ki [31].

A jelenlegi biomakromolekuláris dokkoló módszerek rengeteg dokkolt konformációt értékelnek ki a felületek komplementaritásának mértékét minősítő egyszerű módszerekkel. E módszerek azonban a natív-közeli állapotok mellet sok hamis pozitív találatot adnak, azaz a szerkezetek felületi komplementaritása jó, de a négyzetes középérték-eltérések (RMSD) nagyok. Jelentős erőfeszítések történtek olyan módszerek fejlesztésére, melyek alkalmasak a hamis pozitív találatok kiszűrésére. Bár ezek az eljárások javítják ezt a helyzetet, és

így már általában található a legjobb 10–100 szerkezet között olyan konformáció, melynél az RMSD kevesebb, mint 5 Å, a legjobbaknak sorolt legtöbb komplex szerkezete még továbbra is messze a van a natívtól [32].

A többnyire kis molekula-fehérje kölcsönhatásokat kezelni képes dokkoló eszközök (3.3.1. fejezet) mellett biomakromolekuláknak (többnyire fehérjéknek) a célfehérjékre történő dokkolását lehetővé tévő eszközök is elérhetőek. Ilyenek pl. a HADDOCK, ClusPro, RosettaDock, ZDOCK, GRAMM-X vagy a Hex.

Irodalomjegyzék

[1] Berman H, Henrick K, Nakamura H, Markley JL (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucl Acids Res. 35(suppl 1): D301–D303.

[2] Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A. (2009) Human Protein Reference Database – 2009 update. Nucleic Acids Res. 37(Database issue): D767–D772.

[3] (a) Kopp J, Schwede T (2004) Automated protein structure homology modeling: a progress report. Pharmacogenomics. 5(4): 405–416; (b) Jaroszewski L (2009) Protein structure prediction based on sequence similarity Meth Mol Biol. 569: 129–156.

[4] Orry AJ, Ruben Abagyan R (Eds.) (2012) Homology Modeling: Methods and Protocols (Meth Mol Biol. 857, ISBN: 978-1-61779-587-9), Humana Press, Totowa.

[5] Browne WJ, North AC, Phillips DC, Brew K, Vanaman TC, Hill RL (1969) A possible three dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme. J Mol Biol. 42:65–86.

[6] (a) Magrane M, UniProt Consortium (2011) UniProt Knowledgebase: a hub of integrated protein data. Database. bar009; (b) UniProt Consortium (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic Acids Res. 41(Database issue): D43–D47.

[7] Maglott D, Ostell J, Pruitt KD, Tatusova T (2011) Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. 39(Database issue): D52–D57.

[8] Schwede T (2013) Protein Modeling: What Happened to the „Protein Structure Gap"? Structure 21, 1531–1540.

[9] Baker D, Sali A (2001) Protein structure prediction and structural genomics. Science 294(5540): 93–96.

[10] (a) Baker D (2000) A surprising simplicity to protein folding. Nature 405: 39–42; (b) Bonneau R, Baker D (2001) Ab initio protein structure prediction: progress and prospects. Annu Rev Biophys Biomol Struct. 30: 173–189.

[11] Marti-Renom MA, Stuart A, Fiser A, Sanchez R, Melo F, Sali A (2000) Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct. 29: 291–325.

[12] Fiser A, Sanchez R, Melo F, Sali A (2001) Comparative protein structure modeling. In: Watanabe M, Roux B, MacKerell AD, Jr, Becker O, eds. Computational Biochemistry and Biophysics. New York: Marcel Dekker. pp 275–312.

[13] (a) Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410; (b) Altschul SF, Madden TL, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

[14] Hooft RWW, Vriend G, Sander C, Abola EE (1996) Errors in protein structures. Nature 381: 272–272.

[15] Forster MJ (2002) Molecular modelling in structural biology. Micron 33: 365–384.

[16] (a) Deane CM, Blundell TL (2001) Improved protein loop prediction from sequence alone. Protein Eng 14: 473–478; (b) Deane CM, Kaas Q, Blundell TL (2001) SCORE: predicting the core of protein models. Bioinformatics 17: 541–550; (c) Pei J, Sadreyev R, Grishin NV (2003) PCMA: fast and accurate multiple sequence alignment based on profile consistency. Bioinformatics 19: 427–428.

[17] Errami M, Geourjon C, Deleage G (2003) Detection of unrelated proteins in sequences multiple alignments by using predicted secondary structures. Bioinformatics 19: 506–512.

[18] Mizuguchi K, Deane CM, Blundell TL, Johnson MS, Overington JP (1998) JOY: protein sequence-structure representation and analysis. Bioinformatics. 14: 617–623.

[19] Fiser A, Sali A (2003) Comparative protein structure modeling. In: Chasman D, ed. Protein Structure – Determination, Analysis, and Applications for Drug Discovery. New York: Marcel Dekker, pp. 167–206.

[20] Moult J, James MN (1986) An algorithm for determining the conformation of polypeptide segments in proteins by systematic search, Proteins 1: 146–163.

[21] Flohil JA,Vriend G,Berendsen HJC (2002) Completion and refinement of 3-D homology models with restricted molecular dynamics: Application to targets 47, 58, and 111 in the CASP modeling competition and posterior analysis. Proteins 48: 593–604.

[22] Colovos C, Yeates TO (1993) Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci. 2(9): 1511–1509.

[23] Sippl MJ (1993) Recognition of Errors in Three-Dimensional Structures of Proteins. Proteins 17, 355–362; (b) Wiederstein M, Sippl MJ (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Research 35, W407–W410.

[24] Eisenberg D, Luthy R, Bowie JU (1997) VERIFY3D: assessment of protein models with three-dimensional profiles. Meth Enzymol. 277: 396–404.

[25] Guex N, Peitsch MC, Schwede T (2009) Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: a historical perspective. Electrophoresis 30(Suppl 1): S162–S173.

[26] (a) Battey JN, Kopp J, Bordoli L, Read RJ, Clarke ND, Schwede T (2007) Automated server predictions in CASP7. Proteins 69: 68–82; (b) Brazas, M.D., J.T. Yamada, and B.F. Ouellette (2010) Providing web servers and training in Bioinformatics: 2010 update on the Bioinformatics Links Directory. Nucleic Acids Res. 38(Suppl), W3–W6.

[27] Halperin I, Ma BY, Wolfson H, Nussinov R (2002) Principles of docking: An overview of search algorithms and a guide to scoring functions. Prot Struct Func Genetics 47: 409–443.

[28] (a) Mohan V, Gibbs AC, Cummings MD, Jaeger EP, DesJarlais RL (2005) Docking: Successes and Challenges. Current Pharmaceutical Design, 2005, 11, 323–333; (b) Huang SY, Zou X (2010) Advances and challenges in protein-ligand docking. Int J Mol Sci. 11: 3016–3034; (c) Yuriev E, Agostino M, Ramsland PA (2011) Challenges and advances in computational docking: 2009 in review. J Mol Recogn. 24: 149–164.

[29] (a) Pons C, Grosdidier S, Solernou A, Perez-Cano L, Fernandez-Recio J (2010) Present and future challenges and limitations in protein–protein docking. Proteins 78: 95–108; (b) Li B, Kihara D (2012) Protein docking prediction using predicted protein–protein interface. BMC Bioinform 13: 7.

[30] Roberts VA, Pique ME, Ten Eyck LF, Li S (2013) Predicting protein–DNA interactions by full search computational docking. Prot Struct Funct Bioinf, doi: 10.1002/prot.24395.

[31] Melquiond ASJ, Bonvin AMJJ (2010) Data-driven docking: using external information to spark the biomolecular rendez-vous. In: Protein–protein complexes: analysis, modelling and drug design. Ed.: Zacharrias M, Imperial College Press, London, pp. 183–209.

[32] Zacharias M (2010) Accounting for conformational changes during protein–protein docking. Curr Opin Struct Biol 20(2), 180–186.

4. fejezet

Fehérjeszerkezet-meghatározás kísérleti módszerei és egyszerű fehérjeszerkezet-predikciók

4.1. Bevezetés

A bioinformatika legfontosabb célja, hogy ismeretlen szerkezetű és/vagy funkciójú fehérjék szekvenciáihoz szerkezeti és/vagy funkcionális adatokat rendeljen a hozzá hasonló, ismert szerkezetű és/vagy funkciójú szekvenciák közötti kereséssel. E cél elérése érdekében hatékony és megbízható módszerek szükségesek ahhoz, hogy a fehérjékhez szerkezeti adatokat rendelhessünk. E fejezet a fehérjék másodlagos szerkezetének jellemzésére és háromdimenziós szerkezetük atomi szintű meghatározására alkalmas kísérleti módszereket mutatja be.

A különböző bioinformatikai eljárások során a fehérjeszekvenciák azonosítását és elemezését különböző szinteken végezhetjük.

4.1.1. A fehérjeazonosítás eszközei

A fehérjék azonosítása a proteomikai kutatás fontos kérdése. A fehérjék azonosítására több módszer áll rendelkezésre, a „kis felbontású" technikáktól (pl. azonosítás izoelektromos pontja, molekulatömege és/vagy aminosav-összetétel) kezdve a pontosabb azonosításra és jellemzésre alkalmas peptid MS-ujjlenyomat-adatokon át az olyan „nagy felbontású" technikákig, mint a kapcsolt tömegspektrometriai eljárások.

Számos web-alapú fehérjeazonosítási szolgáltatás érhető el az ExPASY proteomikai szerveren „kis felbontású" fehérjeazonosítási célokra. Ilyen az AACompIdent (a fehérje azonosítása aminosav-összetételéből), az AACompSim (egy UniProtKB/Swiss-Prot szekvencia aminosav-összetételének összehasonlítása a többi szekvenciával), a TagIdent vagy a MultiIdent (fehérje azonosítása izoelektromos pont, pI; molekulatömeg, MW; szekvenciacímke vagy MS-ujjnyomatadatok alapján az adott pI és MW értékekhez közeli fehérjék felsorolásával).

Sok peptidazonosítási szolgáltatás alapul MS-ujjlenyomatokon (fehérjék nem specifikus hasításával képződő peptidek elemzése és azonosítása kísérleti tömegeik alapján), például a Mascot, a PepMAPPER, a FindMod, a ProFound, a FindPept vagy a ProteinProspector. E szolgáltatások általában képesek figyelembe venni vagy előre jelezni a peptidekben lehetséges fehérje poszt-transzlációs módosításokat, az egy-aminosav helyettesítéseket vagy proteázok autolitikus hasítását. A kísérletileg meghatározott peptidtömegeket hasonlítják össze az adott adatbázis-szekvencia vagy a felhasználó által bevitt szekvencia alapján kiszámított elméleti peptidekkel, és a tömegkülönbségeket használják az adott fehérje jobb jellemzésére.

Bonyolultabb fehérje azonosítást/elemzést tesz lehetővé a kapcsolt tömegspektrometriai (MS/MS) módszerek használata. Az ExPASY proteomikai szerveren több web-alapú fehérje és peptid azonosítási/jellemzési szolgáltatás áll rendelkezésre MS/MS adatok alapján, például a hrefhttp://web.expasy.org/quickmod/QuickMod, a Phenyx, a Mascot, az OMSSA, a PepFrag vagy a ProteinProspector. Ezek a szolgáltatások az MS/MS peptidspektrumok azonosítását általában ismert proteinszekvenciák tömegspektrum-könyvtárakban történő keresésekkel végzik.

4.1.2. Egyszerű fehérjeanalízis

A fehérjeazonosítási eszközök mellett továbbiak állnak rendelkezésre fehérjeszekvenciák statisztikai elemzésére (pl. aminosav- és atomösszetétel), egy fehérjeszekvencia által kódolt fehérje egyszerű fiziko-kémiai paramétereinek előrejelzésére (pI, hidrofobicitás, extinkciós együttható stb.), ismétlődő proteinszekvenciák felismerésére vagy domének/régiók előrejelzésére (mint pl. cink-ujjlenyomat vagy peptidkötő régiók).

Számos web-alapú szolgáltatás áll rendelkezésre a ExPASY proteomikai szerveren ilyen egyszerű fehérjeelemzésekre, mint például a ProtParam (fehérjeszekvencia alapján fizikai-kémiai paramétereket számol: aminosav- és atomösszetétel, pI, extinkciós együttható stb.), a Compute pI/Mw (kiszámítja a felhasználó vagy egy MW SWISS-PROT/TrEMBL szekvenciájára az elméleti pI és MW értékeket) vagy a ProtScale (aminosav szintű adatok: hidrofobicitás, egyéb konformációs paraméterek stb.).

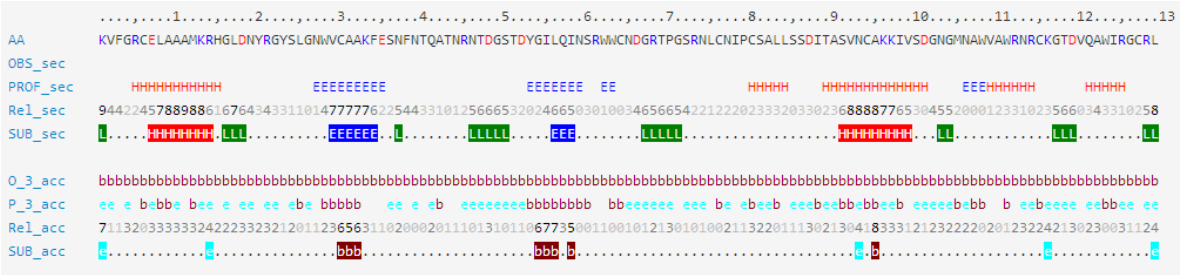
4.1.3. A fehérjeszerkezet-előrejelzés szintjei és nehézségei

A fehérjeszerkezet-előrejelzés általános célja, hogy egy fehérje(szekvencia) esetén meghatározza a szabadentalpia globális minimumának megfelelő konformációt. Kis modellekkel igazolható volt, hogy ez a probléma ún. NP-nehéz. Mivel a megoldáshoz a szükséges idő nem polinomiálisan (hanem jobban) nő a (fehérje)mérettel, egy bizonyos méret felett a problémát nem lehet megoldani. Valós fehérjék esetében azonban a probléma kezelhető, mivel a valós fehérjék szekvenciái meglehetősen specifikusak (evolúció által kiválasztottak), így a már ismert szerkezetek felhasználhatóak például a tudás-bázis alapú előrejelzések során.

A fehérjeszerkezet-előrejelzések szintje eltérő lehet az 1D előrejelzésektől a 2D szerkezeti adatokon át az atomi szintű 3D szerkezetekig.

Egydimenziós előrejelzések esetében a jellemzők egyedi aminosavakhoz rendelhetőek és az eredményt 1D karaktersorral lehet leírni. Ilyen esetek a másodlagos szerkezet, az oldószer-hozzáférhetőség, a hidrofób transzmembrán hélix vagy rendezetlen régiók előrejelzése.

Több web-alapú szolgáltatás létezik különböző egydimenziós előrejelzésekre az ExPASY proteomikai szerveren. Ezekkel fehérjeszekvenciákban jósolható például a fehérje másodlagos szerkezete (APSSP, CFSSP, GOR, Porter, SOPMA), a fehérjefelületi elérhetőség (NetSurfP), β-kanyarok (NetTurnP) vagy helikális transzmembrán régió (HTMSRAP) jelenléte. Egyes szerverek többféle előrejelzést, valamint konszenzus-előrejelzéseket is lehetővé tesznek (Jpred, PredictProtein, PSIpred, Scratch Protein Predictor) (4.1. ábra).



4.1. ábra. Többszörös / konszenzus-előrejelzések a tyúktojás-lizozim fehérje másodlagos szerkezetére és az oldószer-hozzáférhetőségre (az elemzés a PredictProtein szolgáltatással készült).

A ***fehérjék 2D előrejelzéséhez*** aminosavpárok közötti távolságok, kölcsönhatások előrejelzése szükséges. Ugyanakkor ha minden oldallánc-kölcsönhatást előre tudnánk jelezni, lehetővé válna a 3D-s szerkezet építése (lásd később a fehérje NMR módszereknél).

Ahhoz, hogy megbecsüljük az oldallánc-kölcsönhatásokat, a következő adatokat lehet figyelembe venni: a szekvenciában egymástól távoli aminosavak közti korrelált mutációk; statisztikai adatok; átlagos térpotenciálok. A fehérje-2D-előrejelzések során gyakran neurális hálózatokat alkalmaznak. Az eddigi erőfeszítések ellenére mind a mai napig nem igazán sikerült hatékony fehérje-2D-előrejelzési módszereket fejleszteni.

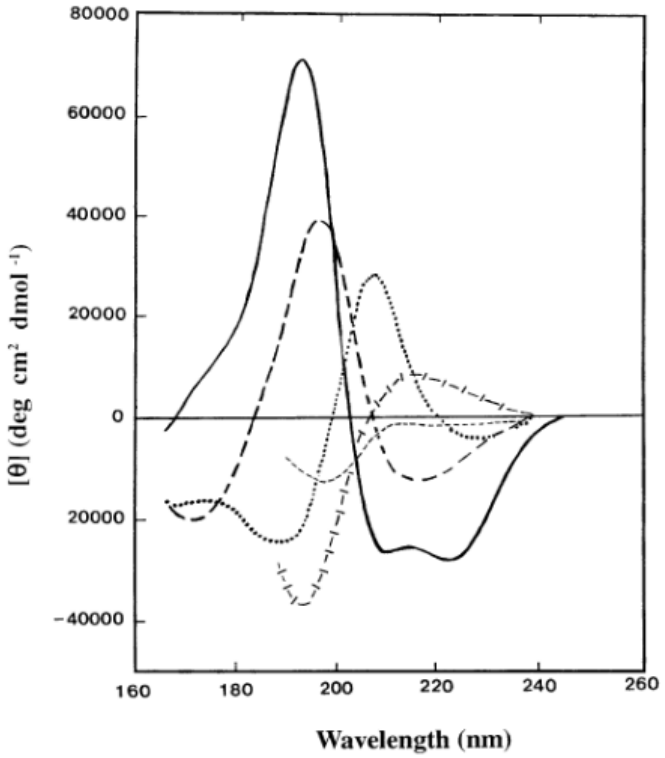
4.2. Fehérjék másodlagos szerkezetének kísérletes vizsgálata

A cirkuláris dikroizmus (CD) széles körben használt technika fehérjék konformációjának és stabilitásának spektroszkópiai vizsgálatához olyan változó környezeti feltételek mellett, mint a hőmérséklet, az ionerősség, vagy más oldott anyagok, illetve kis molekulák jelenléte [1, 2]. A CD-spektroszkópia roncsolásmentes, viszonylag könnyen kezelhető, gyors és csak kis mennyiségű mintát és adatgyűjtést igényel. A szinkrotron sugárzásos cirkuláris dikroizmus (SRCD) spektroszkópia (a szinkrotron nagyobb fluxusa lehetővé teszi az

adatgyűjtést alacsonyabb hullámhosszon) kiterjeszti a hagyományos CD-spektroszkópia (a laboratóriumi alapú eszközök) alkalmazási lehetőségeit [3].

4.2.1. Fehérje cirkuláris dikroizmus (CD)

A CD-spektroszkópia a saját kiralitású vagy királis környezetben lévő kromofórok által a balra és jobbra cirkulárisan polarizált sugárzás elnyelése közötti különbségen alapul. A fehérjékben számos, CD-jeleket eredményező kromofór van jelen [1, 2]. A peptidkötések elnyelésének megfelelő távoli UV-régióban (160–260 nm) a CD-spektrum információt nyújt az olyan másodlagos szerkezeti elemekről, mint például az α-hélix és a β-redő (4.2. ábra).



4.2. ábra. A távoli UV-CD-spektrum kapcsolata a különböző típusú másodlagos szerkezeti elemekkel. Folytonos vonal: α-hélix, hosszú szaggatott vonal: anti-paralell β-redő, szaggatott vonal: I. típusú β-kanyar, áthúzott szaggatott vonal: kibővített 31-hélix vagy poli (Pro) II hélix, rövid szaggatott vonal: szabálytalan szerkezet. [Az S. M. Kelly és munkatársai összefoglalójában [2] közölt ábra Elsevier kiadó által engedélyezett reprodukciója]

A közeli UV régióban (320–260 nm) a CD-spektrum az aromás aminosav-oldalláncok környezetétől függ, és így információt szolgáltat a fehérje harmadlagos szerkezetéről. A CD-jelek olyan más, nem fehérje eredetű kromofóroktól is eredhetnek, mint például flavin- és hemcsoportok, tehát a teljes spektrum az összes érintett kromofór környezeti állapotától függ. Viszonylagos egyszerűsége miatt a CD alkalmas arra, hogy adatokat szolgáltasson a fehérje szerkezetéről, a szerkezetváltozások és ligandkötés mértékéről és sebességéről.

CD-módszerek használhatóak fehérjék vagy fehérjefragmensek szerkezeti stabilitásának és tekeredési jelenségeinek tanulmányozására. A CD rendkívül hasznos technikának bizonyult membránfehérjék szerkezeti integritásának vizsgálatára. Látható, hogy a CD a szerkezeti biológia egy sokoldalú módszere, melyet ennek megfelelően egyre szélesebb körben alkalmaznak [1, 2].

4.2.2. Szinkrotron besugárzásos cirkuláris dikroizmus (SRCD)

Amellett, hogy a laboratóriumi eszközök alapú CD-spektroszkópia a strukturális biológia jól bevált módszere, a szinkrotronsugárzásos cirkuláris dikroizmus (SRCD) spektroszkópia kiterjeszti a hagyományos CD-spektroszkópia alkalmazhatóságát. A szinkrotron nagy fluxusa lehetővé teszi a CD-mérést alacsonyabb hullámhosszon (így nagyobb információtartalom érhető el), nagyobb jel-zaj szintű spektrumok felvételét, valamint vizsgálatokat elnyeléssel rendelkező komponensek (pufferek, sók, lipidek és detergensek) jelenlétében [3]. Az SRCD-spektroszkópia tehát fontos statikus és dinamikus szerkezeti információkat adhat az oldott fehérjékről és olyan fehérje-kölcsönhatásokról, mint például az akár merevtest-, akár indukált-illeszkedési mechanizmussal képződő fehérje-fehérje vagy fehérje-lipid komplexek [3].

A CD- és SRCD- spektrumok és a hozzájuk tartozó metaadatok archiválására, elérésére és elemzésére jött létre nyilvánosan elérhető web-alapú bioinformatikai forrásként a Protein Circular Dichroism Data Bank (PCDDB) [4].

4.2.3. Kísérleti módszerek fehérjék atomi szintű szerkezetének meghatározására

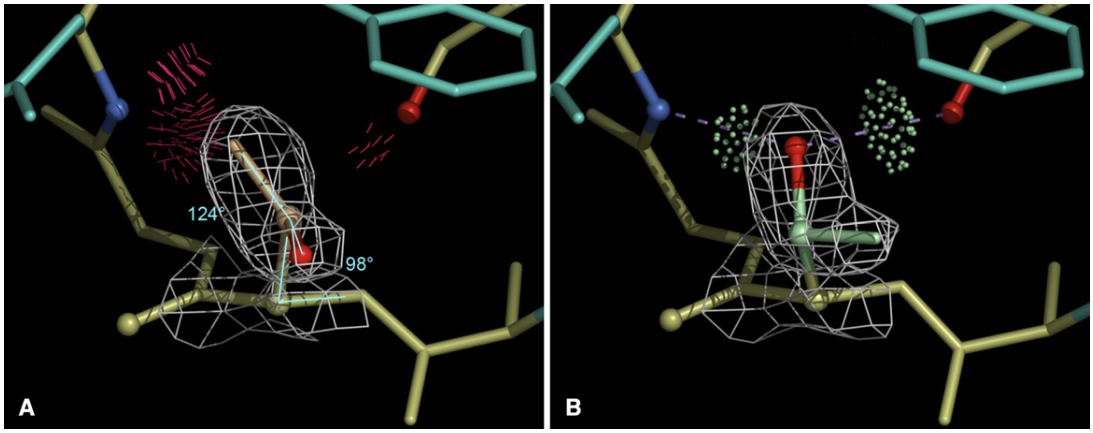
Egy fehérje atomi szintű szerkezetének meghatározására több módszer is alkalmazható. Ilyenek a röntgenkrisztallográfiai, neutrondiffrakciós, elektronmikroszkópiai és elektrondiffrakciós módszerek (ezek kristályos állapotú fehérjeszerkezeteket szolgáltatnak), és az NMR spektroszkópia (ez mind oldat, mind szilárd állapotú szerkezeteket adhat).

Szem előtt kell tartani, hogy minden egyes módszernek vannak előnyei és hátrányai. Az atomi pontosságú végső modellt a tudósok minden esetben több részinformáció összeállításával nyerik. Kiindulásként a tudósok kísérleti adatokat gyűjtenek a molekula szerkezetéről. Az NMR-spektroszkópia esetében az egymáshoz közel elhelyezkedő atomok közötti távolságok nyújtanak információt a helyi konformációról. Röntgenkrisztallográfia esetén a kiindulási adat a röntgendiffrakciós mintázat. Elektronmikroszkópnál a molekula teljes formájának képe a kiindulási pont.

Ezért a kezdeti, kísérleti információ szinte egyetlen esetben sem elegendő önállóan a szerkezet atomi pontosságú meghatározására. A szerkezet meghatározásához a molekulára vonatkozó további információkra is szükség van. A fehérje már ismert aminosavszekvenciája vagy az atomok fehérjékben megszokott geometriája (pl. a kötéshosszak és kötésszögek) gyakran szolgálnak ilyen adatként. A hasonló kiegészítő adatok birtokában a tudósok képessé válnak olyan modellek létrehozására, amelyek összhangban állnak mind

a kezdeti kísérleti adatokkal, mind az ismert szekvenciával és a fehérjék szokásos geometriájával.

Következésképpen a „kísérleti" makromolekuláris szerkezetek mindig kísérleti adatokat és számítógépes predikciókat különböző arányban tartalmazó modellek. A nagyfelbontású kristályszerkezetekben a nehézatomok atomi koordinátáit túlnyomórészt a diffrakciós adatok határozzák meg [5], míg a kevesebb kísérleti megfigyelésre támaszkodó módszerek sokkal nagyobb mértékben alapulnak olyan számítástechnikai eszközökön, melyek a térbeli adatok értelmezésével készítenek szerkezeti modelleket (pl. magmágneses rezonancia [NMR], elektronmikroszkópia [EM], kisszögű röntgenszórás [SAXS], fluoreszcenciarezonancia-energiatranszfer [FRET]) [6]. Nem meglepő tehát, hogy még a viszonylag jó minőségű kísérleti röntgenszerkezetek is tartalmaznak kijavítandó hibákat (4.3. ábra) [5]. Ha tehát kísérleti szerkezetek alapján szeretnénk következtetéseket levonni, legyünk mindig egy kicsit kritikusak. Ne feledjük, hogy a PDB adatbázis [7] szerkezeteit is kísérleti adatok és a tudás alapú modellezés együttes alkalmazásával határozták meg. Ezért mindig tanácsos ellenőrizni, hogy az adott szerkezetre vonatkozó kísérleti adatok támogatják-e az adott szerkezeti modellt, és hogy a tudományos következtetéseket megfelelő modell alapján vontuk-e le.



4.3. ábra. A Thr 32 lokális hibájának kijavítása egy régebbi, 1,7Å felbontású szerkezetben (1SBP). (A) Az 1SBP [8] ezen oldallánca komoly többatomos térbeli ütközést (vörös tüskék) okozott, nem voltak hidrogénkötései, az N-Cα-Cβ és Cγ2-Cβ-Oγ1 tetraéderes szögek (jelölve) rosszak. (B) Az oldallánc 180° elfordítás után már jó geometriájú, ütközésmentes, van két hidrogénkötése, és jobban illeszkedik az elektronsűrűséghez. [Az R. J. Reed és munkatársai által [5] közölt ábra Elsevier kiadó által engedélyezett reprodukciója]

A kísérleti szerkezet megléte lehetővé teszi számunkra a szerkezet elemzését. Atomi szintű szerkezetek alapján lehetséges a fehérjeszerkezet minőségének elemzése, töltések, felületek, üregek vagy másodlagos szerkezet vizsgálata. Emellett szerkezeti motívumok azonosíthatóak vagy vizsgálható kölcsönhatás ligandumokkal, ill. más biomolekulákkal.

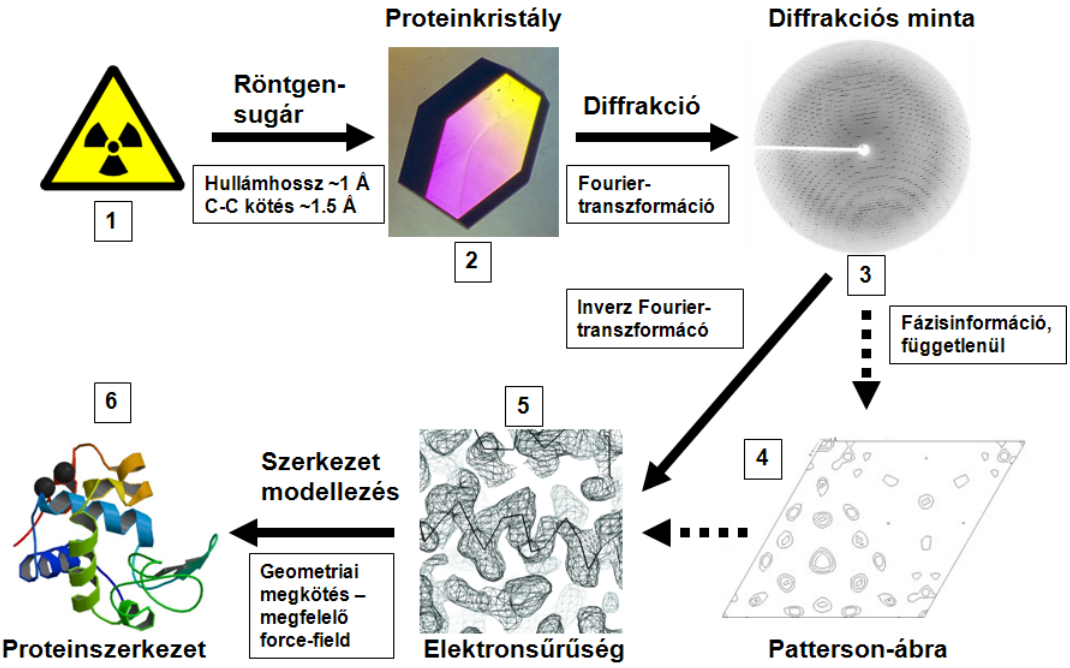
4.2.4. Fehérje-röntgenkrisztallográfia

A PDB adatbázisban [7] elhelyezett szerkezetek többségét röntgenkrisztallográfia [9] segítségével, a 4.4. ábrán látható lépéseken át határozták meg.

Szerkezetük röntgenkrisztallográfiai módszerekkel történő meghatározásához a fehérjéket először elő kell állítani, majd tisztítani és kristályosítani. Ha megvan a megfelelő kristály, azt röntgensugarakkal több irányból intenzíven besugározva elektronikus detektorokkal diffrakciós minták nyerhetők. Mivel a kristályok három dimenzióban periodikusan tartalmazzák a molekulákat, a diffrakciós mintázat folytonos függvény helyett inkább foltok sorozata. A foltok elemzésével meghatározzuk az elektronok eloszlását a fehérjében. Az elemi cella atomi tartalmának képét az eltérített röntgensugárzáson alkalmazott „matematikai lencse" segítségével (inverz Fourier-transzformáció) nyerjük. A kép újjáépítési folyamata bonyolult, mivel a diffraktált röntgensugárzásnak csak az intenzitása mérhető, de az egyes eltérített hullámok relatív fáziseltolódása nem. Ez a hiányzó információ jelenti a „kristálytani fázisproblémát". A hiányzó fázisadatok különböző kísérleti/számítási módszerekkel nyerhetőek (izomorf csere, nehézatom rendellenes szóródása vagy részlegesen ismert szerkezetek alkalmazása) [9]. Mivel a röntgenkrisztallográfiás vizsgálatban a röntgendiffrakciót az elektronok és a röntgensugarak kölcsönhatása okozza, az eredményül kapott kép az elektronsűrűség eloszlása a kristály elemi cellájában. Interaktív és iteratív számításokkal a kísérleti elektronsűrűség-térképhez legjobban illeszkedő atomi helyzeteket meghatározva nyerhető a végső atomi modell. A PDB adatbázisban az így meghatározott kristályszerkezet kétféle adatot tartalmaz. A PDB fájlok a végső modell atomi koordinátáit és a szerkezetmeghatározás szerkezeti tényezőit (a röntgendiffrakciós minta foltjainak intenzitása és fázisa) tartalmazzák. Ezekből az adatokból az elektronsűrűség eloszlás-képe létrehozható olyan eszközökkel, mint például az Astex viewer.

A biológiai molekulakristályok egészen különbözőek lehetnek: egyes esetekben tökéletes, rendezett kristályok, míg máskor csak a gyenge kristályok nyerhetőek. A meghatározható atomi szerkezet pontossága tehát függ a kristályok minőségétől. Egy kristályszerkezet pontossága két fontos paraméterrel jellemezhető, mint a felbontás (amely megszabja milyen részletességgel tehetők láthatóvá a kísérleti adatok) és az R-érték (amely azt mutatja, hogy mennyire jól támasztják alá a szerkezeti tényező fájl kísérleti adatai az atomi modellt). Az 4.5. ábra mutatja be a felbontás jelentőségét. Látható, hogy a nagy felbontású (∼1,0 Å) szerkezet pontos atomi pozíciókat ad, míg 3 Å felbontásnál vagy az alatt csak a fehérje alapvető alakja ábrázolható, és az egyedi atomi pozíciók pontatlanok.

A röntgenkrisztallográfia nagyon részletes atomi információkat szolgáltató szerkezeteket nyújthat, melyek a fehérje vagy nukleinsav minden nehézatomját tartalmazzák, és részleteket szolgáltatnak olyan ligandumok, inhibitorok, ionok és más molekulák jelenlétéről és elrendeződéséről, amelyek megtalálhatóak a kristályban. A kristályosodási folyamat azonban nehéz, és ez korlátozza, hogy milyen típusú fehérjéket lehet tanulmányozni ezzel a módszerrel. Például a szép, jól rendezett kristályokat alkotó merev fehérjék szerkezetének meghatározására ideális a röntgenkrisztallográfia. Ezzel ellentétben sokkal nehezebb a flexibilis fehérjék tanulmányozása ily módon, mivel a krisztallográfia módszere azon alapul, hogy igen sok molekulánk van pontosan azonos elrendeződésben. A fehérje flexibilis



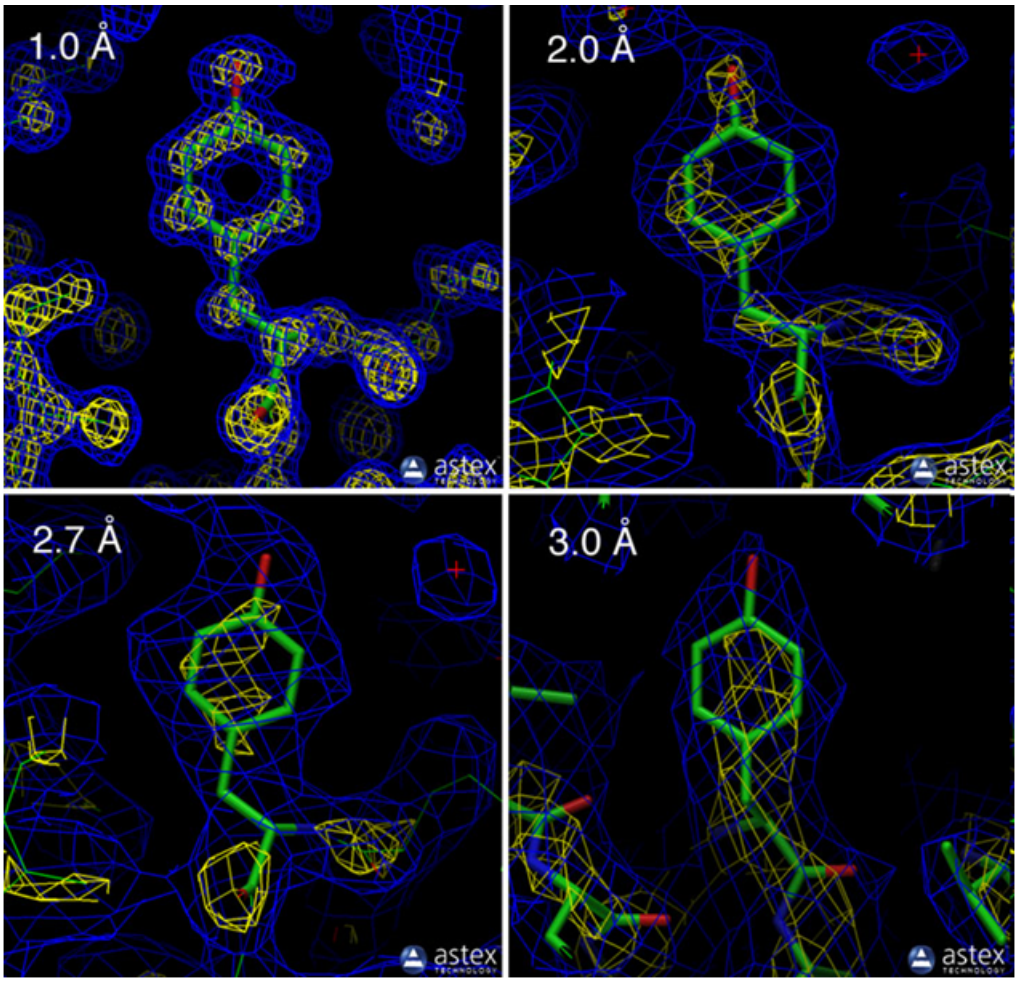
4.4. ábra. A fehérjeszerkezet-meghatározás lépéseinek áttekintése egykristály-diffrakcióval a rendellenes szórást kihasználva. A röntgensugárforrásból (1) röntgensugarakkal besugározva a fehérjekristály (2) eltéríti a sugarakat. Az így nyert rezgési képeket (3) használjuk a szerkezet inverz Fourier-transzformációval végzett „megoldásához". Ezen az ábrán a fázisprobléma megoldását Patterson-térkép (4) segíti a nehézatom-alszerkezet meghatározása során. A fázis- és diffrakciós adatok lehetővé teszik elektronsűrűség-térkép (5) kiszámítását és a kezdeti modell-nyomvonal kialakítását. Modellezési, ellenőrzési, a diffrakciós adatokon és geometriai korlátozásokon alapuló modell újrajavítási, finomítási lépéseket tartalmazó többszörös iteráció után nyerhető a fehérjeszerkezet-modell (6), melyet itt a főlánc tekeredésének megfelelő szalagmegjelenítés mutat be.

részei gyakran láthatatlanok a röntgenkrisztallográfia számára, mivel ezek elektronsűrűségei nagy térben oszlanak el. Ez látszólag hiányzó koordinátákat tartalmazó szerkezeteket eredményezhet.

4.2.5. Fehérje-NMR-spektroszkópia

Magmágneses rezonancia (NMR) -spektroszkópiai módszerekkel oldott fehérjékről juthatunk adatokhoz [10], eltérően azoktól a módszerektől, amelyek fehérjéket kristályban vagy mikroszkopikus rácshoz kötve igényelik. Flexibilis fehérjék atomi szerkezetének tanulmányozására tehát az NMR-spektroszkópia a leginkább alkalmas módszer. Az NMR-spektroszkópiát fehérjeszerkezet-meghatározáshoz a 4.7. ábrán látható módon használják.

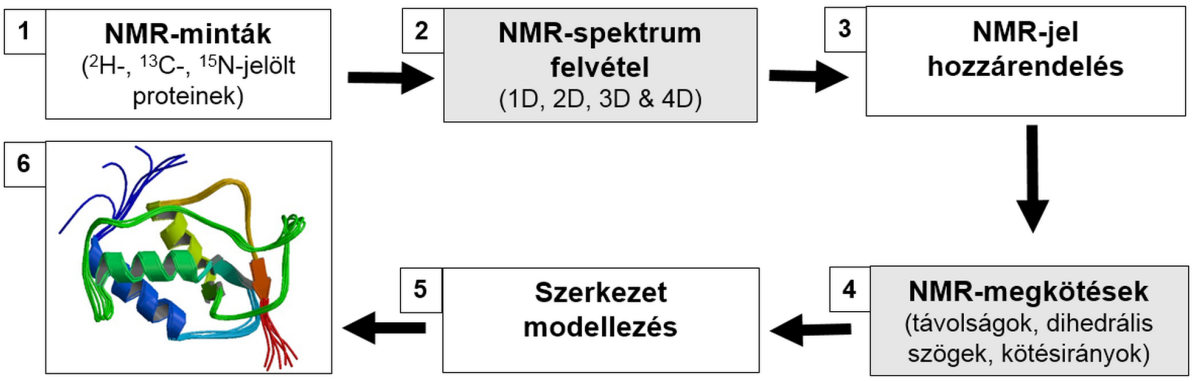
Az NMR-szerkezeti vizsgálatokhoz a kérdéses fehérje tisztított formájának oldata szükséges. Mivel csak a ^{1}H magok (ám a ^{12}C és ^{14}N nem) NMR-aktívak, a nagyobb polipepti-



4.5. ábra. A fehérje röntgenkrisztallográfia felbontásának jelentősége. Az első három példa (A), (B) és (C) a mioglobin Tyr103 egységét mutatja 1,0 Å (1A6M), 2,0 (106M), és 2,7 Å (108M) felbontással. Az utolsó példa (D) a hemoglobin Tyr103 egységét (B lánc) ábrázolja 3,0 Å felbontással (1S0H).

dek és proteinek szerkezeti vizsgálataihoz ^{2}H-, ^{13}C- és ^{15}N-izotóppal jelölt fehérjemintákra van szükség. A stabil, NMR-aktív ^{13}C és ^{15}N izotópok túltermelt fehérjékbe építésére alkalmas hatékony molekuláris biológiai technikák a többdimenziós heteronukleáris spektroszkópiás technikák tervezésének és megvalósításának drámai fejlődését eredményezték [11]. Ennek nyomán a szerkezeti vizsgálat maximális fehérjemérete a homonukleáris 1H-NMR-spektroszkópiával vizsgálható ∼10 kDa méretről heteronukleáris ^{13}C- és ^{15}N-NMR-spektroszkópia használatával a ∼30 kDa méretre és ^{13}C és ^{15}N heteronukleáris NMR-spektroszkópia részleges ^{2}H-gazdagítással kombinálásával kb. ∼40-50 kDa méretre nőtt. A technika jelenleg ilyen fehérjeméretekre korlátozódik, mivel a nagyobb méretű proteinek NMR-spektrumában problémát jelentenek az átfedő csúcsok.

Az NMR-kísérlet során a fehérjeminta oldatát erős mágneses térbe helyezve vizsgál-



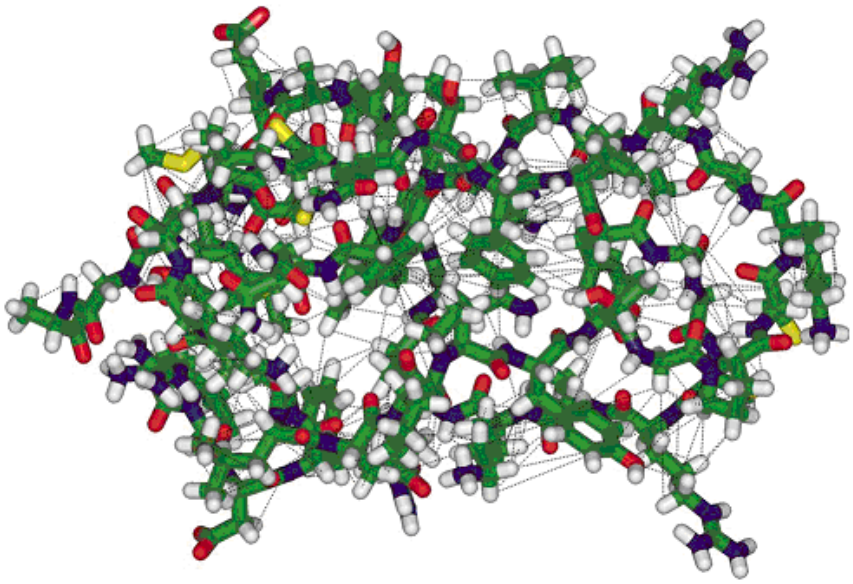
4.6. ábra. Az oldatfázisban végzett NMR-fehérjeszerkezet-meghatározás lépései (az árnyékolt téglalapok képviselik az NMR-kísérletekkel végzett lépéseket). Az izotópjelzett fehérjeminták (1) oldatából különféle NMR-spektrumok (2) készülnek. Az NMR-jelek kézi vagy automatizált módon fehérjeatomokhoz történő hozzárendelését (3) követően különböző típusú korlátozásokat (4) határoznak meg. Ezután szerkezeti korlátozásokat kielégítő modellezési módszerekkel (5) egy sor fehérjeszerkezet (6) készül, melyet itt a főlánc tekeredésének megfelelő szalagmegjelenítés mutat be.

ják rádióhullámokkal. Az NMR-spektrumok kézi vagy automata elemzésével elkészül a rezonanciák atommagokhoz történő részletes hozzárendelése, melyet további speciális NMR-kísérletek követnek (pl. amelyek a nukleáris Overhauser-hatást, NOE alkalmazzák), melyekkel meghatározhatóak olyan atommagok, amelyek közel vannak egymáshoz (4.7. ábra).

Ezek a különböző távolság-, szög- és kötés-torziósszög-adatok jellemzik az egymáshoz kötött atomok helyi konformációját. A korlátozások listája alapján készül el a korlátozásoknak legjobban megfelelő és az egyes atomok helyzetét megadó fehérje modell.

Egy tipikus NMR-szerkezet nem csupán az egyedi fehérje szerkezete, hanem olyan fehérjeszerkezetek együttese, amelyek mindegyike bizonyos mértékig összhangban van a megfigyelt kísérleti korlátozásokkal. Az NMR-szerkezetek ilyen együttese tartalmaz olyan régiókat, amelyekben a szerkezetek az erős korlátozások miatt nagyon hasonlítanak egymásra, és tartalmaz kevésbé korlátozott részeket, ahol azok nagyon különbözőek. Ezek a kevésbé korlátozott részek a molekula flexibilis részei, amelyek az NMR-kísérletekben nem adnak erős jeleket. Az NMR-szerkezet bizonyos mértékig tehát utalhat a fehérje dinamikus viselkedésére.

Az PDB-adatbázisban jellemzően kétféle koordináta-készletű NMR-szerkezet található. Az egyik esetben az NMR-szerkezet több olyan különálló modellszerkezet készlete, amelyek mindegyike megfelel kísérletek során nyert szerkezeti korlátozásoknak. A másik esetben a PDB-fájl egy olyan minimalizált átlagszerkezet, amely a fehérje átlagos tulajdonságainak kíván megfelelni. A PDB-fájlok tartalmazzák az NMR-kísérletben meghatározott korlátozások listáját (például hidrogénkötések és a diszulfid-kötések, egymáshoz közeli hidrogénatomok közötti távolságok, valamint korlátozások a helyi konformáció és a lánc sztereokémiája tekintetében).



4.7. ábra. Példa a fehérjeszerkezet NMR-meghatározására. A hasnyálmirigy tripszininhibitor-szerkezetét szimulált hűtési eljárással nyerték Discover-erőtér alkalmazásával (Accelerys Ltd, ld. http://www.accelerys.com). A szaggatott vonalak hidrogénatom párok közötti, kísérleti NMR-adatokból nyert távolságkorlátozások, amelyek alapján határozták meg a szerkezetet. A kép InsightII-vel (Accelerys) készült. [Az M. J. Forster összefoglalójában [12] közölt ábra Elsevier kiadó által engedélyezett reprodukciója]

4.2.6. Fehérje-elektronmikroszkópia, elektrondiffrakció és elektronkrisztallográfia

Az ***elektronmikroszkópia (EM)*** nagy makromolekuláris komplexek szerkezetének meghatározására alkalmazható. Az EM során a molekuláris objektum képe különböző módszerek segítségével közvetlenül nyerhető az elektronsugarakkal. Ha a fehérjék kisméretű koaxiális kristályokat képeznek, vagy ha szimmetrikusan rendeződnek el egy membránban, ***elektrondiffrakció (ED)*** használható 3D-sűrűség térkép létrehozására a röntgendiffrakciósokhoz hasonló módszerek alkalmazásával. Ha a molekula nagyon szimmetrikus, (mint pl. a vírus kapszidokban), sok különálló diffrakciós kép alkotható különböző nézetekből. E nézetek összerendezése és átlagolása után nyerhetőek ki a 3D adatok. Ezeken túl az ***elektrontomográfia*** egyetlen objektum elforgatásával készít több képet különböző nézetekből elektronmikroszkópos felvételekkel. E nézetek feldolgozásával képezhetőek a 3D-s adatok.

Jellemzően az EM-kísérletek nem teszik lehetővé atomi szintű szerkezet meghatározását, hanem a molekula teljes 3D alakját adják. Néhány különösen jól viselkedő rendszer esetében, mint például egyes membránfehérjék, az EM-mérések atomi szintű adatokat is szolgáltathatnak [13]. Atomi részletek meghatározásához az EM-vizsgálatokat gyakran ötvözik röntgendiffrakciós vagy NMR-spektroszkópiai információkkal, és a röntgen- vagy NMR-kísérletek atomi struktúráit az ED-elektronsűrűség-térképekbe dokkolva nyerik a

komplex modelljét. Ez a kombinált megközelítés különféle multi-biomolekuláris együttesek esetében is sikeresnek bizonyult.

Az e technikákkal nyert kísérleti adatok az Elektronmikroszkóp Adatbankban (EMDB) – ez a makromolekuláris komplexek és szubcelluláris struktúrák elektronmikroszkópos sűrűségtérképeinek nyilvános adattára – találhatóak meg. Olyan különböző technikákkal nyert adatokat tartalmaz, mint az egyrészecske-elemzés, elektrontomográfia és elektron-(2D)-krisztallográfia.

Számos membránfehérje atomi felbontású szerkezetét (<3Å felbontás) határozták meg nemrégiben ***elektronkrisztallográfiával (EC)*** [14]. Bár ezt a módszert több mint 40 évvel ezelőtt dolgozták ki, még mindig gyerekcipőben jár a kétdimenziós (2D) kristályosodás, adatgyűjtés, elemzés és fehérje-szerkezetmeghatározás tekintetében. Az adatokat illetően az elektronkrisztallográfia magába foglalja mind a képalkotást, mind az elektrondiffrakciós adatgyűjtést [14].

Az EC kiegészítheti a röntgenkrisztallográfiás vizsgálatokat olyan, kis kristályokat (<0,1 mikrométer) adó fehérjék esetében (mint például a membránfehérjék), amelyek nem könnyen képeznek a röntgenmódszerekhez szükséges nagy 3D kristályokat. EC-módszerekkel a fehérjeszerkezetek meghatározhatóak akár a 2-dimenziós kristályokból (lapok vagy hélixek), poliéderekből (például virális kapszid) vagy diszpergált egyedi fehérjékből. Míg az elektronok alkalmazhatóak ilyen esetekben, a röntgensugárzás nem, mivel az elektronok kölcsönhatása az atomokkal erősebb, mint a röntgensugaraké. A röntgenkrisztallográfiával szemben, ahol nincs röntgenlencse, és így fennáll a fázisprobléma, az elektronmikroszkópok elektronlencséket tartalmaznak, és így a krisztallográfiai szerkezet faktor-fázisinformációja az EC- vizsgálatban kísérletileg meghatározható.

4.2.7. Fehérje-neutronkrisztallográfia

A neutron-fehérjekrisztallográfia (NC) hatékony kiegészítője lehet a röntgenkrisztallográfiának, mivel lehetőséget ad a biológiai szerkezetekben olyan kulcsfontosságú hidrogénatomok helyzetének meghatározására, amelyek csupán röntgenkrisztallográfiai módszerekkel nem láthatóak. A teljes mértékben deuterált fehérjék elkészíthetősége bakteriális expressziós rendszerekkel megszünteti a háttérhez nagyban hozzájáruló inkoherens hidrogénszórást.

Jellemző, hogy a fehérjék röntgenszerkezetei nem adják meg a hidrogénatomok pontos helyzetét. Bár a nagy felbontású röntgen-kristályszerkezetekben néhány hidrogénatom észlelhető, a funkcionálisan fontos hidrogénatomok gyakran nem láthatók. Együttes röntgen- és neutrondiffrakciós vizsgálatok jelezték a NC alkalmazhatóságát a funkcionálisan fontos hidrogénatomok atomi helyzetének pontos meghatározására (pl. az egyes aminosavak protonálódási/deprotonálódási állapota) a fehérjeszerkezetekben [15].

A protein NC fő akadálya, hogy szokatlanul nagy kristályokra ($\sim$1 mm^3) van szükség a rendelkezésre álló neutronsugárzás gyenge fluxusának ellensúlyozásához.

Irodalomjegyzék

[1] S. M. Kelly and N. C. Price, The Use of Circular Dichroism in the Investigation of Protein Structure and Function. *Curr Prot Peptide Sci* 1:349–338, 2000.

[2] S. M. Kelly, T. J. Jess, and N. C. Price, How to study proteins by circular dichroism. *Biochim Biophys Acta Prot Proteom* 1751:119–139, 2005.

[3] (a) A. J. Miles and B. A. Wallace, Synchrotron radiation circular dichroism spectroscopy of proteins and applications in structural and functional genomics. *Chem Soc Rev* 35:39–51 2006; (b) B. A. Wallace and R. W. Janes, Synchrotron radiation circular dichroism (SRCD) spectroscopy: an enhanced method for examining protein conformations and protein interactions. *Biochem Soc Trans* 38(4):861–873, 2010.

[4] L. Whitmore, B. Woollett, A. J. Miles, R. W. Janes, and B. A. Wallace, The protein circular dichroism data bank, a Web-based site for access to circular dichroism spectroscopic data. *Structure* 18(10):1267–1269, 2010.

[5] R. J. Read, P. D. Adams, W. B. Arendall, A. T. Brunger, P. Emsley, R. P. Joosten, G. J. Kleywegt, E. B. Krissinel, T. Luetteke, Z. Otwinowski, A. Perrakis, J. S. Richardson, W. H. Sheffler, J. L. Smith, I. J. Tickle, G. Vriend, and P. H. Zwart, A new generation of crystallographic validation tools for the protein data bank. *Structure* 19:1395–1412, 2011.

[6] T. Schwede, Protein Modeling: What Happened to the "Protein Structure Gap"? *Structure* 21:1531–1540, 2013.

[7] H. Berman, K. Henrick, H. Nakamura, and J. L. Markley, The worldwide Protein Data Bank (wwPDB), ensuring a single, uniform archive of PDB data. *Nucl Acids Res.* 35(suppl 1):D301–D303, 2007.

[8] J. J. He and F. A. Quiocho, Dominant role of local dipoles in stabilizing uncompensated charges on a sulfate sequestered in a periplasmic active transport protein. *Protein Sci* 2:1643–1647, 1993.

[9] E. E. Lattman and P. J. Loll, *Protein Crystallography: A Concise Guide*. The John Hopkins University Press, Baltimore, Maryland, 2008, 152 pp.

[10] P. R. Markwick, T. Malliavin, M. Nilges, Structural biology by NMR: structure, dynamics, and interactions. *PLoS Comp Biol* 4:e1000168, 2008.

[11] J. Cavanagh, W. J. Fairbrother, A. G. Palmer, M. Rance, and N. J. Skelton, *Protein NMR Spectroscopy* (2nd edition), Academic Press, Burlington, 2007.

[12] M. J. Forster, Molecular modelling in structural biology, *Micron* 33:365–384, 2002.

[13] Y. Fujiyoshi, Electron crystallography for structural and functional studies of membrane proteins. *J Electron Micr* 60(Suppl. 1):S149–S159, 2011.

[14] T. Gonen, The collection of high-resolution electron diffraction data, *Methods Mol Biol* 955:153-–169, 2013.

[15] (a) S. Yamaguchi, H. Kamikubo, N. Shimizu, Y. Yamazaki, Y. Imamoto, and M. Kataoka, Preparation of large crystals of photoactive yellow protein for neutron diffraction and high resolution crystal structure analysis. *Photochem Photobiol.* 83(2):336–338, 2007; (b) E. I. Howard, M. P. Blakeley, M. Haertlein, I. Petit-Haertlein, A. Mitschler, S. J. Fisher, A. Cousido-Siah, A. G. Salvay, A. Popov, C. Muller-Dieckmann, T. Petrova, and A. Podjarny, Neutron structure of type-III antifreeze protein allows the reconstruction of AFP-ice interface. *J Mol Recognit.* 24(4):724–732, 2011.

5. fejezet

Genetikai variánsok funkcionális hatásainak kvantitatív modelljei

5.1. Bevezetés

A gének kifejeződése határozza meg a sejt identitását és ezzel működését és képességeit. A DNS által kódolt RNS-ek és fehérjék folyamatos egyensúly fenntartására törekszenek a termelés és a lebontás között, amire több szinten megvalósuló, sokrétű szabályozási körök adnak lehetőséget. Az örökítőanyag tartalmazza az élő szervezetek használati útmutatásait. A DNS-ben található variánsok számos módon képesek a gének expresszióját és aktuális mennyiségét befolyásolni, ami természetesen a fenotípusban is megjelenhet. Ennek megfelelően nagyon sok kutatás foglalkozik a transzkripciós faktorokkal, de a génexpresszió szabályozása többszintű, és csak a teljes képet vizsgálva érthetjük meg pontosan, hogyan jutunk el a DNS-től a fehérjéig, és azt, hogy egy adott pillanatban egy adott sejtben az expresszált fehérje mennyiségének változása miért történik, és ez a változás mit jelent a fenotípusra nézve. Ebben a fejezetben a genetikai szabályozás különböző szintjeit és típusait tekintjük át. Megvizsgáljuk az egyes variánsok lehetséges funkcionális hatását is.

A fejezetben elsősorban a micro-RNS-ekre és transzkripciós faktorokra helyezzük a hangsúlyt, ugyanakkor az említés szintjén foglalkozunk további szabályozó mechanizmusokkal is (pl. epigenetika). Míg most csak egy-egy variáns lehetséges hatását tekintjük át, egy későbbi fejezetben már hálózat szintű modellezéssel is foglalkozunk.

5.2. Variánsok

Ahhoz, hogy variánsok funkcionális hatásáról beszélhessünk, fontos tisztázni, mit értünk variánsok és funkcionális hatás alatt. Egy rövid áttekintést adunk a genetikai variánsok típusairól és azok lehetséges funkcionális hatásáról.

5.2.1. SNP, indel

A Single Nucleotide Polymorphism (SNP) azaz egy pontos polimorfizmusok a legelterjedtebb genetikai variációk. Ilyenkor a genom egy bázisa felcserélődik a referenciához képest egy másik bázisra. A kérdéses bázis pozíciója alapján megkülönböztetünk:

- kódoló
  - kódoló, aminosavcserét nem okozó (szinonim)
  - kódoló, aminosavcserét okozó (nem szinonim)
    - missense
    - nonsense
- nem kódoló
  - nem transzlálódó regióba (*untranslated region*, UTR) eső
  - intronba eső
  - intergenikus területen elhelyezkedő

SNP-ket. A nem kódoló régióba eső SNP-k az egyes génekről átíródó fehérjének nem változtatják meg az aminosavak sorrendjét, de hatással lehetnek elsősorban a közelükben található gének expressziójára. A kódoló szakaszba eső SNP-k közül a szinonim polimorfizmusok nem változtatják meg az aminosavak sorrendjét, de ritkán közvetlen hatással lehetnek a protein szerkezetére. Ezek mellett a kódoló szakaszba eső és aminosavcserét okozó SNP-k fejtik ki a legkönnyebben leírható hatást. Két típusukat különböztetjük meg: a missense aminosavcserét okoz, de nem stop codonra cseréli ki az adott aminosavat, míg a nonsense típusú SNP stop codonra cseréli az eredeti aminosavat, ezzel sok esetben jelentősen lerövidítve a fehérjelánc hosszát, aminek további erős hatása lehet a fehérje expressziójára. Az úgynevezett UTR SNP-k, ahogy azt a későbbiekben látni fogjuk, szintén fontos szerephez juthatnak a génexpresszió megváltozásában, ugyanis elsődlegesen ezeken a szakaszokon találhatjuk a miRNS kötőhelyeket. Az intronikus szakaszokra eső SNP-k esetében hasonló megfigyelések tehetők, mint a nem kódoló szakaszokon taláható polimorfizmusok esetén.

Az egy bázist érintő polimorfizmusok mellett léteznek még egyéb hasonlóan kis kiterjedésű variánsok, melyek akár több bázist és érinthetnek, mint az inszerciók és a deléciók. Inszerció és deléció esetén egy vagy több bázis illesztődik be, illetve esik ki a genom egy adott pontjáról. Ezek az eltérések az SNP-khez hasonló módon érinthetik a fenotípust. Kódoló régióba eső mutáció esetén további kérdés, hogy okoz-e az aminosav átfordításakor ún. leolvasási kereteltolódást (*frame shift*). Ez abban az esetben fordul elő, ha nem (az aminosavakat kódoló kodonokban lévő bázisoknak megfelelően) 3 vagy ennek valamilyen egész számú többszöröse a kiesett vagy hozzáadott darab hossza.

5.2.2. Alternatív splicing

Egy DNS szakaszból a transzkripció során hírvivő RNS (*messenger*, mRNS) képződik. Már az átíródás alatt megkezdődik a fehérjét kódoló RNS-ek érése: csak az exonok kerülnek be az mRNS-be, az pedig intronok kivágásra kerülnek. Ezt a folyamatot nevezzük splicingnak. A több exonból álló gének esetén sokszor több változat készülhet: vagy az exonok sorrendje cserélődik fel, vagy egyes exonok ki is maradhatnak az mRNS-ből. A gyakran sejt- vagy szövetspecifikusan szabályozott folyamat eredményeként más-más fehérjét kapunk végtermékként.

5.3. A szabályozás szintjei

A folyamatot, melynek során a DNS-ben kódolt információ alapján fehérje keletkezik, bonyolult szabályozási hálózatok befolyásolják. Az egyes szabályozó elemeket el lehet különíteni az alapján, hogy hatását a DNS-ről mRNS-re történő átíráskor (transzkripcionálisan vagy kotranszkripcionálisan pl. transzkripciós faktorok), vagy az érett mRNS-hez kapcsolódva (poszttranszkripcionális szinten, pl. miRNS-ek), esetleg a fehérjéhez kötődéssel (poszttranszlácionálisan, pl. foszforiláció) fejti ki. Az különböző szabályozási szintek között gyakori a kapcsolat több vissza- és előrecsatolással. Egy miRNS gátolhatja egy transzkripciós faktor transzlációját, ahogy egy transzkripciós faktor is gátolhatja egy miRNS expresszióját. Az egyes szabályozó elemek építik fel a génregularizációs hálókat, melyekkel a következő fejezetben részletesebben foglalkozunk.

5.4. Különböző szabályozó elemek

5.5. microRNS

A microRNS (miRNA) egy átlagosan 22 bázispár (bp) hosszú egyszálú RNS darab, amely az mRNS-ekhez kötődve – jellemzően negatívan – befolyásolni tudja az mRNS transzlációját. Először Caenorhabditis elegansban sikerült kimutatni miRNS gének funkcionális jelentőségét. A miRNS-ek szabályozási szerepét számos életfolyamatban igazolták eukariótákban. A sejtosztódásban, az apoptózisban (programozott sejthalál), jelátviteli útvonalak regulációjában, fejlődési programok végrehajtásában, pl. a szív- és érrendszer, vagy az idegrendszer fejlődésében résztvevő gének különösen gyakran esnek a miRNS-ek közvetítette szabályozás alá.

Egy miRNS molekulához pár száz célkötőhely tartozik. Az eddigi ismereteink alapján a miRNS az 5' végén lévő seed szakasz (2-8 bp hosszú) alapján ismeri fel az mRNS 3' végén található kötőhelyét. Ugyanakkor a miRNS közödhet a mRNS 5' UTR régiójába és a mRNS kódoló szakaszába is. Kísérletben kimutatták, hogy a kötőhelytől függően más-más erősségű hatást okoz a miRNS. A miRNS hatásmechanizmusai alapvetően a következők:

- transzláció gátlás
- mRNS deadenyláció
- mRNS tárolás

A miRNS mindig gátolja a mRNS átírását. Az 5.1. ábrán látható a miRNS különböző hatásainak összefoglalása.

5.5.1. miRNS érés

A miRNS érésének folyamata különbözik állatokban és növényekben. A jelen fejezetben az állatokra, így emberekre jellemző folyamatot ismertetjük [1]. A miRNS érése a sejtmagban kezdődik, ahol az elsődleges miRNS-t (pri-miRNS) az RNS-polymerase II enzim átírja a DNS-ről. A pri-miRNS több száz bp hosszú lehet és több miRNS-t is tartalmazhat. Ezt követően a Drosha enzim kimetszi a pri-miRNS-ből a hajtűre emlékeztető prekurzor miRNS-t (pre-miRNS). A pre-miRNS kijut a citoplazmába és itt egy Dicer enzim vágja ki a hajtű törzsének megfelelő kettősszálú szakaszt, amiből érett miRNS keletkezik [1]. Az érett miRNS-nek megfelelő szakasz egy összeszerelődő fehérjekomplexbe (miRISC, miRNA induced silencing complex) épül be, majd egyszálúvá válva, „molekuláris címzésként" irányítja a komplexet a komplementer szekvenciát tartalmazó célpontok felé.

5.5.2. miRNS által mediált szabályozási formák

Transzláció gátlása

A miRNS sok esetben már a transzláció elindulását (iniciáció) is gátolja, de a transzláció elindulását követően is több módon tudja a fehérje keletkezését gátolni. Kísérletes adatok szerint előfordul a riboszóma idő előtti leválása, máskor a miRNS az aminosavlánc hosszabbodását (elongáció) lassíthatja le, esetleg teljesen meg is állíthatja. Ezekben az esetekben kevesebb fehérjetermék keletkezik, viszont a mRNS mennyisége változatlan marad.

mRNS deadenyláció

Az mRNS deadeniláció során a miRNS-től függően az mRNS mennyisége is csökken. Ilyenkor miRNS által vezetett komplex kapcsolódása destabilizálja az mRNS molekulát. A deadenilációt az mRNS 5' végén található sapka (cap) leválasztása követheti, ami az mRNS degradálódásához vezet. Habár sok esetben a deadeniláció előfeltétele a degradációnak, megfigyelések szerint az mRNS nem minden esetben kerül lebontásra. Egy kísérletben a deadenilációt követően találtak stabil, részben stabil mRNS molekulákat is. Annak ellenére, viszont, hogy a deadenylációt követően az mRNS stabil maradt, az expresszió erősen gátolt maradt a miRNS kapcsolódásának eredményeként.

mRNS szekvesztrációja

A miRNS szabályozásnak egy közvetett formája a cél-mRNS-ek kivonása az genetikai információáramlás folyamatából. Ilyenkor a miRNS a szokásos szabályszerűségek szerint hozzákötődik a mRNS-hez, majd a citoplazma ún. P-testébe irányítja az mRNS-t. Itt történhet deadeniláció és a mRNS degradációja is előfordul, de sok esetben csak a kompartment csak ideiglenesen „tárolja" az mRNS-t. Mivel a P-testekben egyáltalán nincsen riboszóma, ezért itt nem tud végbemenni transzláció.

5.6. Transzkripciós faktorok

A transzkripciós faktorok (TF) a génexpresszió szabályozásába a DNS RNS-re történő átírás folyamatának szintjén avatkozhatnak be. Nagyszámú fehérje tartozik ide, amely képes a gének transzkripciójának iniciálására és szabályozására (általános és specifikus TF-ok). Különlegességük, hogy rendelkeznek egy DNS-kötő doménnel (fehérjerészlettel), amely képessé teszi őket a gének promóter, illetve silencer és enhancer szakaszaihoz való kötődésre. A transzkripciós faktorok a miRNS-ekkel szemben nem csak gátolni (represszálni), hanem serkenteni is tudják a gének átírását. A gén környezetében szinte bárhol előfordulhatnak transzkripciós faktor kötőhelyek (*transcription factor binding site*, TFBS): a promóter régióban, távolabb a promóter régión kívül, intronokban és az UTR szakaszokban is, nemegyszer több ezer bp-nyi távolságra a upstream vagy downstream a transzkripciós start helytől. A kötőhelyek általában klaszterekbe szerveződnek, ahova egyszerre több TF is kötődhet. A génekben vagy azok közelében elhelyezkedő, a gének megfelelő kifejeződését biztosító, nem kódoló DNS-szekvenciákat összefoglaló néven *cis*-szabályozó elemeknek nevezzük. Az elnevezés arra utal, hogy a szabályozó elem a DNS-en szorosan a génnel együtt lokalizálódik, szemben a trans – szabályozó elemekkel, amelyek szabályozó hatásukat távoli, pl. más kromoszómán elhelyezkedő génekre fejtik ki. Egy génhez egyszerre több transzkripciós faktor is kötődhet (kombinatorikus szabályozás), és igény szerint a transzkripciós faktorok különböző kombinációkban kötődhetnek az adott génhez.

5.7. Epigenetika

Az epigenetikai vizsgálatok a XXI. század elején lettek igazán népszerűek, jóllehet maga a kifejezés a XX. század első feléből származik. Az epigenetika azokkal a molekuláris mechanizmusokkal foglalkozik, amelyeknek köszönhetően kialakuló örökölhető állapotok nem a DNS szekvencia eltéréseire vezethetők vissza. Amellett, hogy sejt- és szövetspecifikus génexpressziós-szabályozást valósítanak meg, lehetővé teszik a sejtek gyorsabb alkalmazkodását a környezet változásaihoz. Két főbb epigenetikai mechanizmust járunk körbe: a hiszton módosulások és a metiláció segítségével történő szabályozást.

5.7.1. Metiláció

A DNS metilációja során a citozin bázisokhoz, a metil-transzferáz enzimek segítségével egy metil (-CH3) csoport kötődhet, amely így metil-citozinné alakul. A metiláció mértéke fordítottan arányos az érintett kódoló szekvenciák aktivitásával. Az emlősök nagyszámú GC-ismétlődést tartalmazó, jellemzően a gének promóter régiója környezetében előforduló CpG-szigeteinek jó része, 70-80%-a metilált állapotban található a genomban, csendesítve az adott gént. Daganatok esetében rendszerint rendellenes metilációs mintázat figyelhető meg.

Megjegyzés. A génekhez tartozó cisz-szabályozó régiók metiláltsága és a róluk folyó transzkripció mértéke közötti összefüggés nem minden esetben egyértelmű: a gének kódoló régiójában levő, gyakran szövetspecifikusan kialakuló metiláció egyes esetekben éppen fokozza az transzkripció hatékonyságát. A DNS metilációjának és a hiszton-fehérjék kovalens módosulásainak jelentőségét a kromatin denzitásának szabályozásában és ezzel a DNS hozzáférhetőségében feltételezik. A közelmúltban felismert duonok (dual-use codons) a gének kódoló, exonikus, fehérjévé lefordítódó szakaszainak másodlagos (kettős) szerepére mutatnak rá, amikor ezek a szekvenciák amellett, hogy a fehérje aminosavsorrendjét is meghatározzák, transzkripciós faktorok számára szolgálnak kötőhelyként. Az átfogó vizsgálatok, a mintegy 81 különböző sejttípusban végzett genomszintű TF-kötőhely térképezés módszerét használva, megdöbbentő megfigyelésekhez vezettek: a gének több, mint 85%-ában előfordulóan, a genom összes kodonjának, azaz fehérjére lefordítódó szekvenciájának, 15%-ának transzkripciós faktorokkal történő lefedése igazolható. A jelenség a kodonok használatának preferenciáját alakító tényezők közé, a fehérjék aminosavsorrendje mellett a transzkripciós faktorok kötődését lehetővé tévő motívumok kialakítását vetik fel. Ugyanakkor a szinoním, aminosavcserét nem eredményező variánsok génexpresszióra és ezzel a fenotípusra gyakorolt hatását is szükséges átértékelni.

Jóllehet megfigyelhető, hogy a transzkripciós faktorok felülreprezentáltak a magasabb szinten expresszálódó gének exonjaiban, egyelőre tisztázatlan, milyen módon képesek befolyásolni a transzkripció folyamatát. A jelenlegi általános tankönyvi modellbe az eredmény mindenesetre egyelőre nehezen illeszthető be. Lehetséges, hogy ezek a transzkripciós faktorok más, szomszédos gének átírására gyakorolnak hatást, és az is lehet, hogy nem is hagyományos módon működnek, hanem egyszerűen „nyitva tartják” a kromatinszerkezetet, és ezzel a géneket is az átírás számára.

A közelmúltban tett megfigyelés ismételten felhívja a figyelmet arra, hogy a genom még számos rejtett kódot hordozhat magában, és hasonlóak felfedezése tovább diverzifikálja az amúgy is összetett elképzelésünket a működéséről.

5.7.2. Hisztonmódosulások

A beavatkozás setjmagban található DNS magasabb rendű szerveződésébe szintén szabályozásra ad lehetőséget. A kettős hélix hiszton fehérjék alkotta komplexekre feltekert formája elősegíti az érintett szakaszok hozzáférhetőségének befolyásolását, valamint a sejtosztódás folyamán a kromatin kromoszómákká tömörítését. Transzkripció során a hisztonfehérjék (pl. hiszton deacetilázok által katalizált) módosulásainak következtében a megfelelő szekvenciarészletek letekerednek és hozzáférhetővé válnak.

5.8. Modellezés

A technológiai újításoknak köszönhetően, egyre több genetikai információ válik elérhetővé. Ezeket az adatokat felhasználva egyre pontosabban megismerhetjük és modellezhetjük az egyes génszabályozási mechanizmusokat vagy akár teljes génszabályozási hálózatokat. A biológiai szabályozás komplexitása miatt jelenleg nincs olyan általánosan használható modell, amelynek segítségével az egyes mutációk hatását lehetne több szinten vizsgálni. Leginkább a prokarióták alap szabályozó mechanizmusait ismerjük, erre mutatunk egy példát, a laktóz operont.

Egy SNP-nek jelentős hatása lehet egy gén expressziójára, és nem csak akkor, ha aminosavat kódoló régióba esik. Bemutatunk egy módszert, mellyel meg lehet határozni, hogy egy-egy transzkripciós faktor kötőhelyre (TFBS) eső SNP-nek milyen hatása lehet a TF kötési energiájára. Végül adunk egy általános útmutatót arra nézve, hogy milyen típusú matematikai modellekkel lehet jellemezni a transzkripciótól akár a keletkező fehérje mennyiségéig az egyes szabályozó mechanizmusok hatását. Ezek a példák általában az egyes esetekben jelentős megszorításokkal alkalmazhatóak. A paraméterezésük pedig nagyban függ a rendelkezésre álló információktól [3].

5.8.1. regSNP

Az egyes variánsok lehetséges hatásait röviden bemutattuk az 5.2.1. alfejezetben, elsősorban az aminosav sorrendre és a fehérje szerkezetére gyakorolt változásokra koncentráltva. Ezek mellett hasonlóan fontos a keletkező fehérje mennyiségét befolyásoló variánsok hatása. A gének promóter régiójában található transzkripciós faktor kötőhelyek és a hozzájuk kapcsolódó TF-ek köcslönhatását jelentősen befolyásolhatja akár egy SNP is. Ugyanis a TFBS-on található SNP-k módosíthatják a kötési energiát a DNS szakasz és a TF között. A regSNP [4] algoritmust arra fejlesztették, hogy a TFBS és a TF közötti kötési energiát és az adott gén egy fenotípusban (jellemzően egyfajta betegségben) várt szerepe alapján felállítsanak egy sorrendet a kötőhelyeken elhelyezkedő SNP-k között.

A kötési energia kiszámításához felhasználták a TRANSFAC [5] adatbázisban szereplő ún. *positional weight* mátrixokat (PWM). Az allél gyakoriságát, az összes – a TRANSFAC adatbázisban előforduló adott TFBS-hez kötődő – TF számát és az adott allél adott pozicióban lévő PWM-ből vett számosságát felhasználva adják meg a referencia és az alternatív allél esetén számolt kötési energiát a TFBS-re és az adott TF-re nézve. Ezt felhasználva megállapítható, hogy az adott SNP mekkora hatással van a TF kötődésére. A p-érték számításhoz véletlenszerűen választanak SNP-ket a HapMapből. A végső sorrendet az előbb említett módszer és az Endavourrel [6] végzett génprioritizálás sorrendjének fúziójából számítják.

5.8.2. Boolean modellek

Sok biológiai folyamat leírható be/ki jellegű kapcsolókkal, például a géntranszkripció is. Ilyenkor a transzkripciós faktorokat tekintjük a kapcsolóknak, melyek szablyozzák, hogy

egy génről történik átírás. Az egyes szabályozó elemek között pedig ÉS (AND, $\wedge$), VAGY (OR, $\vee$) és NEM (NOT, $\neg$) jellegű kapcsolatokat használhatunk. Ezzel a módszerrel kvalitatívan jól leírható egy biológiai hálózat. Például egy adott génről a fehérje átíródását lehet jellemezni a következőképp.

A gént egy transzkripciós faktor gátolja és egy „bekapcsolja", emellett egy miRNS gátolja, akkor fehérje akkor keletkezik, ha

$$TF_{i,g,be} \wedge \neg TF_{j,g,ki} \wedge \neg MIRNS_{k,g,ki}, \tag{5.1}$$

$$TF_{1,g,be} \vee TF_{2,g,be}, \tag{5.2}$$

ahol g jelöli az adott gént és a be/ki, hogy ki- vagy bekapcsolja az adott szabályozó elem a kérdéses gént. Ezek a modellek azokban az esetekben használhatóak jól, ahol az egyes elemek közötti kapcsolatok ismertek és a rendszer dinamikus működését akarjuk vizsgálni. A leírás egyszerűsége miatt nagy, sok szabályozó elemet tartalmazó hálózatok modellezésére is alkalmas. Ugyanakkor jelentős korlát, hogy kvantitatív jellemzést nem tesz lehetővé.

5.8.3. Termodinamikai modellek

A gének expresszióját a hozzájuk kötődő transzkripciós faktorok kombinatorikusan szabályozzák. A TFBS-hez kötött transzkripciós faktor megakadályozhatja de segítheti is egy újabb TF kötődését a kérdéses génhez. Ezt a folyamatot (a *cis*-szabályozást) jellemezhetjük termodinamikai modellekkel [3]. A jelenlegi modellek nem veszik ugyan figyelembe a kromatinszerkezetet vagy a metiláltságot, de így is kielégítő leírást adnak. A modell felállítása két lépésben történik. Először meghatározzuk és súlyozzuk, az összes lehetséges állapotát a szabályozó régiónak, a kötőhelyeket és az oda kötődő molekulákat figyelembe véve. Ha egy kötőhely van, akkor kettő állapot lesz: amikor beköt egy TF és amikor nem. A súlyozást elsősorban a TF-ek koncentrációja és a kötési energia befolyásolja. Minél magasabb a koncentráció és nagyobb a kötési energia, annál valószínűbb lesz, hogy az adott TF kötődik a génhez. Egy állapot súlyát aztán elosztjuk az összes állapot súlyának összegével. A második lépésben az egyes állapotokhoz rendelünk expressziós mintázatot, azaz meghatározzuk, hogy a transzkripciós faktorok adott kombinációja milyen mértékű gén expressziót okoz.

A termodinamikai modellek az állapottérben folytonos leírást adnak a szabályozó hálózatról. A Boolean modellekhez viszonyítva, pontosabban tudjuk modellezni az adott szabályozó hálózatot, emellett viszont a számítási igény is növekszik.

5.8.4. Differenciálegyenletek

Differenciálegyenleteket akkor használunk modellek leírására, ha tipikusan időben és/vagy térben változó mennyiségeket akarunk jellemezni. Ilyenkor minden egyes elem a többi elem függvénye. Például az mRNS koncentrációját meg lehet adni a miRNS koncentrációjának függvényében. Az egyes mennyiségekhez pedig paraméterként megadjuk a lebomlási időt

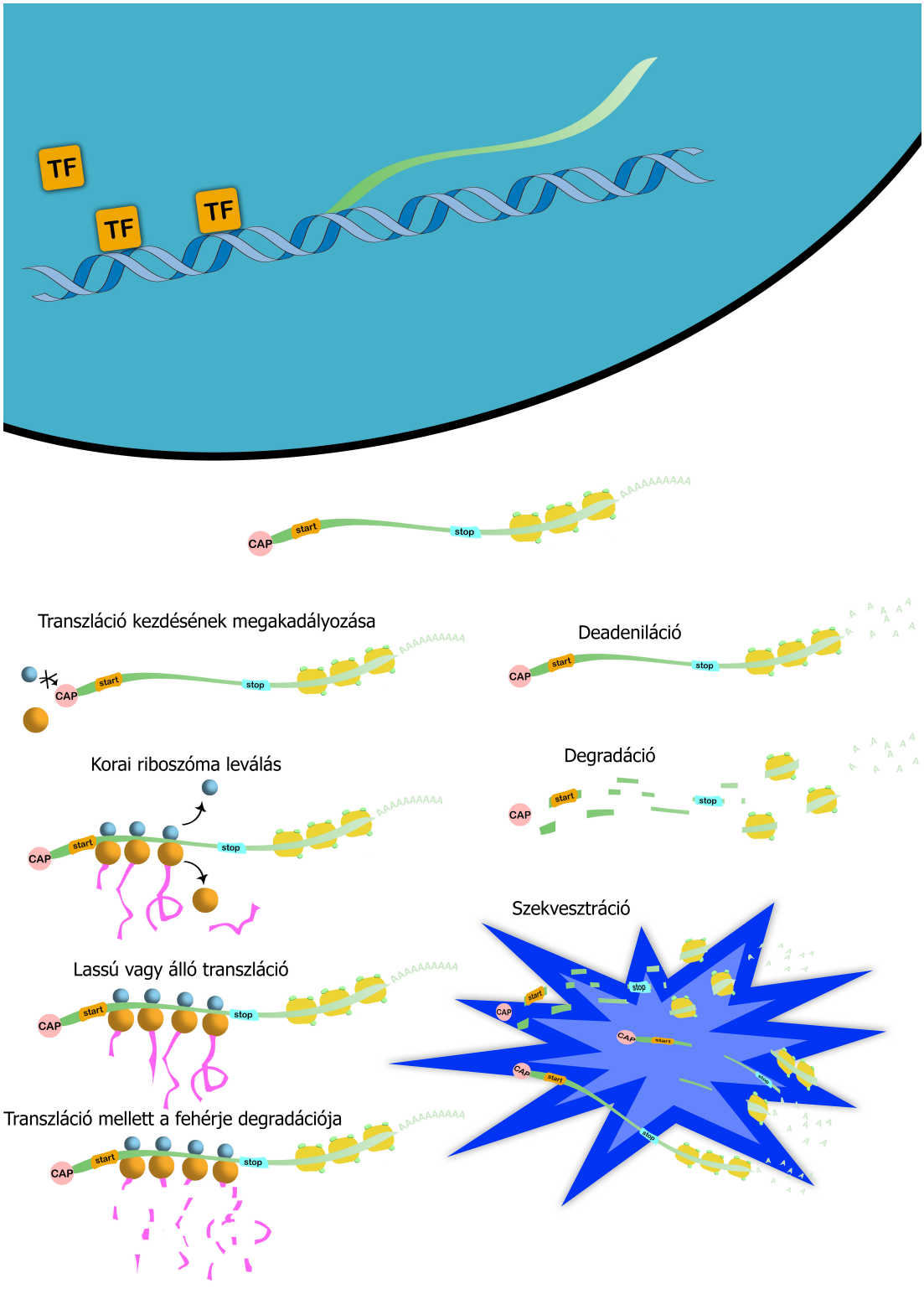
vagy az átírás időtartamát. Két részre bonthatjuk a differenciálegyenleteket: közönséges (ordinary differential equation, ODE) és parciális differenciálegyenletekre (partial differential equation, PDE). ODE csak egy változótól például az időtől függnek, míg PDE esetén több függő változónk van. Ezek a modellek pontos leírását adják a szabályozó hálózatnak, ugyanakkor már pár szabályozó elem esetén bonyolutak lehetnek és analitikusan nehéz a megoldásuk. Jóval nagyobb a számításigényük is, de léteznek numerikus módszerek, melyekkel jó megoldások adhatóak a differenciálegyenlet-rendszerekre. Az első ilyen modellek az operonok voltak, például a laktóz operon.

5.8.5. Lac operon

Az első génszabályozási mechanizmust, a laktóz enzim átírását szabályozó lac operont 1961-ben írta le először Jacob és Monod [7], akik 4 évvel később Nobel-díjat kaptak ezért az eredményért. Az operonok olyan egységei a DNS-nek, ahol több a kromoszómán egymás mellett elhelyezkedő gént egy közös promóter szabályoz. Egy operonnak a következő elemekből áll:

- Szabályozó gén: ez a gén szabályozza az operon strukturális génjeinek a transzkripcióját
- Promóter: a közös promóter régiója a strukturális géneknek
- Operátor(ok): a szabályozó gén az operon operator régiójába kötődik
- Strukturális gének: az operon fehérjét kódoló génjei
- Terminátor: az operon végét jelző DNS szakasz

A lac operont *E. coli* baktériumban írták le. Ez az operon 3 strukturális gént tartalmaz (*lacY,lacZ, lacA*). A működését a *lacI* gén szabályozza, amelynek átírása folyamatos, amíg nincs laktóz a sejtben. Ilyenkor nincs szükség a laktózt feldolgozó enzimekre sem, ezért a *lacI* gén által kódolt represszor az operátorrégióba kötődve megakadályozza az enzimek transzkripcióját. Laktóz megjelenését követően a represszor fehérjének megváltozik a szerkezete, ezért leválik az operátorról, így lehetővé válik az enzimeket kódoló gének transzkripciója. Későbbi kutatások kimutatták, hogy a lac operon további 2 operátort tartalmaz [8], és ezek kombinatorikusan szabályozzák a már ismert operátorral együtt a transzkripciót. A teljes gátláshoz szükséges, hogy minden operátorrégióba kössön gátló fehérje. A korábban már leírt operátor ugyan a legfontosabb, de önmagában csak gyengébben gátolja az DNS átírását. Továbbá egy szabályozó fehérje egyszerre több operátorrégióba köthet, hurok formába kényszerítve a DNS-t.



5.1. ábra. A miRNS-mediált különböző szabályozási mechanizmusok [2]

Irodalomjegyzék

[1] K. Chen and N. Rajewsky, The evolution of gene regulation by transcription factors and microRNAs. *Nat Rev Genet*, 8(2):93–103, 2007.

[2] T. W. Nilsen, Mechanisms of microRNA-mediated gene regulation in animal cells. *Trends in Genetics*, 23(5):243–249, 2007.

[3] A. Ay and D. N. Arnosti, Mathematical modeling of gene expression: a guide for the perplexed biologist. *Critical reviews in biochemistry and molecular biology*, 46(2):137–151, 2011.

[4] M. Teng, S. Ichikawa, L. R. Padgett, Y. Wang, M. Mort, D. N. Cooper, D. L. Koller, T. Foroud, H. J. Edenberg, M. J. Econs, et al., regSNPs: a strategy for prioritizing regulatory single nucleotide substitutions. *Bioinformatics*, 28 (14):1879–1886, 2012.

[5] V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, et al., TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic acids research*, 34(suppl 1):D108–D110, 2006.

[6] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L.-C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, et al., Gene prioritization through genomic data fusion. *Nature biotechnology*, 24(5):537–544, 2006.

[7] F. Jacob and J. Monod, On the Regulation of Gene Activity. *Cold Spring Harbor Symposia on Quantitative Biology*, 26:193–211, 1961.

[8] S. Oehler, E. R. Eismann, H. Krämer, and B. Müller-Hill, The three operators of the lac operon cooperate in repression. *The EMBO journal*, 9(4):973, 1990.

[9] L. Cerulo, C. Elkan, and M. Ceccarelli, Learning gene regulatory networks from only positive and unlabeled data. *BMC Bioinformatics*, 11(1):228, 2010.

[10] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. D. Favera, and A. Califano, ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context, *BMC bioinformatics*, 7(Suppl 1):S7, 2006.

[11] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner, Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biology*, 5(1):e8, 2007.

[12] S. Liang, S. Fuhrman, R. Somogyi, et al., REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. *Pacific symposium on biocomputing*, vol. 3, pp. 18–29, 1998.

[13] C. Elkan and K. Noto, Learning Classifiers from Only Positive and Unlabeled Data. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pp. 213–220, New York, NY, USA, 2008. ACM.

[14] T. D. Le, L. Liu, B. Liu, A. Tsykin, G. J. Goodall, K. Satou, and J. Li, Inferring microRNA and transcription factor regulatory networks in heterogeneous data. *BMC Bioinformatics*, 14:92, 2013.

6. fejezet

Génszabályozási hálózatok matematikai modelljei

6.1. Bevezetés

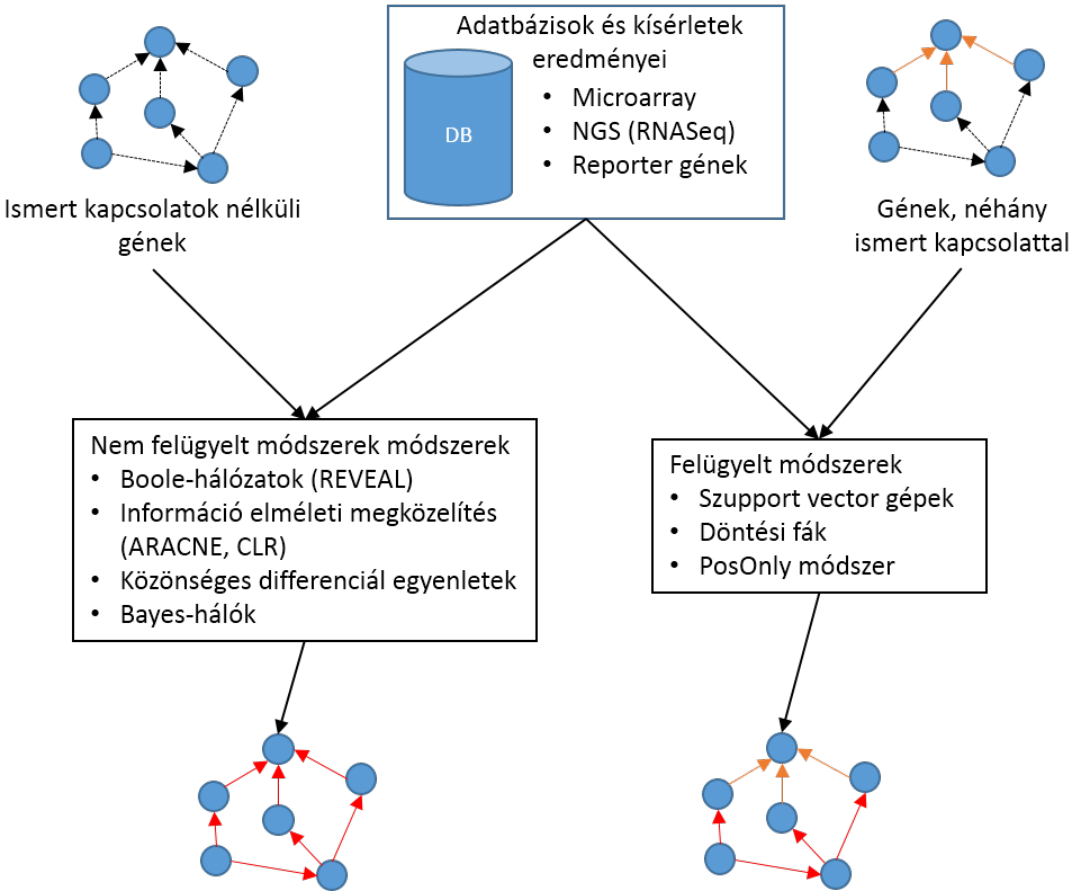
Az 5. fejezetben bemutattuk az egyes genetikai mutációkat és azok lehetséges hatásait. Továbbá körbejártuk azokat az alapvető módszereket, melyekkel ezeket a hatásokat akár polimorfizmusok szintjén lehet modellezni. Végül röviden összefoglaltunk egy-két módszercsaládot (termodinamikai, differenciálegyenlet, Boole-módszerek), melyek a magasabb szintű modellezést tesznek lehetővé. Ebben a fejezetben folytatjuk és részletesebben tárgyaljuk a genetikai szabályozási hálózatok tanulását. Áttekintjük, hogy az egyes algoritmusokat milyen adatforrásokkal tudjuk tanítani, végül néhány módszert részletesen bemutatunk.

6.2. Hálók tanulása

Sok tanulási algoritmus létezik (a 6.1. ábrán szerepel pár a teljesség igénye nélkül) hálózatok tanulására. Ezek két fő osztályba sorolhatók: felügyelt és nem felügyelt tanulási algoritmusok. A nem felügyelt algoritmusok esetén nincsen címkézett adatunk (nincs információnk arról, hogy az adott elem milyen osztályba tartozhat), így nincsen hibamodellünk sem, tehát nem lesz ilyen típusú visszacsatolás a rendszerben, ami megnehezíti a kapott eredmények értékelését. Felügyelt tanulás esetén rendelkezünk tanító pontokkal és az algoritmust az alapján paraméterezzük fel, hogy a tanító vagy a teszt halmazon az adott hibafüggvényre a legkisebb hibát adja. Bioinformatikai alkalamzásokban általában két fő kihívással kell felügyelt tanulás esetén megküzdeni. Sokszor jelentősen eltér a negatív (pl. kontroll) és pozitív (pl. beteg) minták száma. Ilyenkor vagy kiegyensúlyozzuk a halmazt tanító pontok elhagyásával, vagy korrigálunk az eltérő mintaszámra. A másik probléma a negatív minták hiánya. Egy adott fenotípus vagy betegség esetén a korábbi vizsgálatok eredményei alapján van ismeretünk arról, hogy milyen gének állhatnak kapcsolatban a külső jegyekkel. Ezek lehetnek a pozitív tanítópontok. Negatív mintát viszont nehéz ta-

lálni a publikációk hiánya miatt. Eddig nem asszociált génekről nem tudhatjuk biztosan, hogy nem állnak kapcsolatban az adott fenotípussal. Ez jelentős torzítást okoz a tanuló rendszerben, amit figyelembe kell venni.

A génszabályozási hálókat tanuló gépi tanulási technikák általában irányított gráfnak tekintik a szabályozási hálót. Az egyes csomópontok a szabályozási háló elemei, például gének vagy fehérjék, míg az élek az egyes elemek közötti kapcsolatot jelenítik meg.



6.1. ábra. Különböző tanulási algoritmusok [9]

6.3. Nem felügyelt tanulási módszerek

Négy nagy csoportba sorolhatóak a nem felügyelt háló tanulási módszerek.

- Információelméleti modellek
- Boole-hálózati modellek
- Differenciál- és differenciaegyenletekből építkező modellek
- Bayes-i modellek

Az információelméleti modellek, mint az ARACNE [10] és CLR [11] az expressziós szinteket használják kapcsolatok megtalálására az egyes szabályozó elemek között. Ha a génexpresszió szintjének korrelációja két gén esetén egy küszöb fölé esik, akkor a két gén ezen módszerek szerint valamilyen kapcsolatban áll egymással.

A Boole-hálózatok bináris változókat használnak az irányított gráf csomópontjaiként, hogy a gén aktuális állapotát kódolják, és Boole-függvényeket a kapcsolatok reprezentálására. Ilyen módszer a REVEAL [12].

A differencia- és differenciálegyenletek egy génexpressziós szintjét a többi gén expressziójának függvényében definiálják. Ez egy differenciálegyenlet-rendszert ad meg, aminek a megoldása adja meg a hálózatot. Ezek a módszerek általában közönséges differenciálegyenlet-rendszereket használnak a modell készítésre.

Egy Bayes-i módszer minden expressziós szintet random változónak tekint és Bayes-szabályok rendszerét oldja meg. A legnagyobb előnye ezen módszereknek, hogy egyszerű előzetes (prior) információt beépíteni a rendszerbe. Ilyen prior lehet például egy már ismert interakció.

6.3.1. ARACNE

Egy információelméleti módszer az ARACNE [10], amely génpároknak a kölcsönös információ- (*mutual information*, MI) tartalmát számítja ki az expressziós mérésekből. Az egyes mérések alapján meghatározák a génekhez tartozó valószínűséget $P(g_i)$-t. Majd ezt felhasználva számolják a kölcsönös információt:

$$I(x, y) = S(x) + S(y) - S(x, y), \tag{6.1}$$

ahol $S(t)$ a Shannon-entrópia

$$S(t) = -\sum_i p(t_i) \log p(t_i), \tag{6.2}$$

és $p(t_i) = P(t = t_i)$. $I(x, y) = 0$ akkor és csak akkor, ha $P(g_i, g_j) = P(g_i)P(g_j)$. $I(g_i, g_j)$ a két gén közötti statisztika összefüggést méri. A rendelkezésre álló expressziós adatból készítenek egy becslést I_0-t és számítják a hozzá tartozó p-értéket. Ez a becslés lesz a minimum MI érték, ami alatt a kezdeti hálózatba sem kerül be egy gén–gén kapcsolat. Ha két gén (g_1 és g_3) egy harmadik génen (g_2) keresztül van csak kapcsolatban, akkor

$$I(g_1, g_3) \leq \min[I(g_1, g_2), I(g_2, g_3)]. \tag{6.3}$$

Végül a kezdeti hálózatban megvizsgálnak minden hármast, és eltávolítják a legkisebb MI értékkel rendelkező párt.

6.3.2. REVEAL

Sok esetben nem a hálózat dinamikájára vagyunk kíváncsiak, esetleg nincs elegendő adatunk, vagy számítási kapacitásunk bonyolultabb hálók esetén, hanem csak egy hálózat

struktúráját szeretnénk meghatározni. Ilyen esetekben alkalmazhatunk Boole-módszereket, amelyek csak on, off kapcsolóként kezelik az egyes gének közötti kapcsolatokat. A Reveal [12] az ARACNE-hoz hasonlóan a kölcsönös információt használja fel a gének közötti kapcsolathoz. Minden hálózatban szereplő gén szerepel az input és az output rétegben is. Először csak egy-egy gén közötti kapcsolatot vizsgálja. Amennyiben talál olyan input gént (A), amely megmagyarázza az adott gén kimenetét (B'), akkor megalkotja ez alapján a szabályt. Amennyiben nem talál megfelelő egy-egy kapcsolatot, akkor kettő input–egy output kapcsolatot keres, és addig folytatja, amíg nem kap eredményként megfelelő leírást.

A megállási feltétel a következő:

$$M(Y', X)/H(Y') = 1, \tag{6.4}$$

ahol Y' egy tetszőleges gén kimenete, míg X egy vagy több gén bemenete.

6.4. Felügyelt módszerek

A felügyelt tanulás esetén nemcsak az expressziós mérésből származó adatra van szükség, hanem már bizonyítottan ismert szabályozó kapcsolatokra is. Több adatbázis is létezik, melyekben ilyen kapcsolatok találhatóak. A teljesség igénye nélkül a legjelentősebbek:

- TRANSFAC transzkripciós faktorok és kötőhelyeik
- miRNA adatbázisok kísérletileg validált és jósolt miRNS–cél párokkal
  - mirTarBase
  - miRanda
  - TarBase
- String fehérje–fehérje interakciós adatbázis
- KEGG
- IPA

Ezen módszerek alapötlete intuitív. Amennyiben A elem $e(A)$ expressziós profillal rendelkezik és ismert, hogy szabályozza B elemet $e(B)$ expressziós szinttel, akkor a hasonló expressziós profillal rendelkező elemek között is feltételezhetjük, hogy hasonló szabályozó kapcsolat áll fenn. Annak ellenére, hogy ezek az adatbázisok sok információt tartalmaznak, a különböző interakciókról csak pozitív példák szerepelnek bennük, ami a legtöbb osztályozó algoritmusnak gondot okoz. Több megoldás is létezik ennek a hatásnak a kiküszöbölésére, de ezek közül pár erősen alkalmazásfüggő.

A legegyszerűbb módszer a negatív tanító pontok véletlenszerű kiválasztása a nem osztályozott halmazból. Ebben az esetben viszont az algoritmus teljesítményét nagyban befolyásolhatja, ha a random választott pontok közé hamis negatív pontok kerülnek. Ahhoz, hogy jobban tudjunk választani a nem osztályozott tanítópontok közül, használhatunk

szövegbányászatot. Első lépésben választunk negatív tanítópontokat a tf–idf módszer segítségével, majd több osztályozó algoritmust lefuttatva a legjobb eredményt vesszük. Egy másik lehetőség, hogy standard osztályozót tanítunk az eredeti csak pozitív mintákat tartalmazó tanító halmazon, és ennek eredményét használjuk fel arra, hogy meghatározzuk, mekkora valószínűséggel tartozik egy tanítópont a pozitív osztályba. A PosOnly [13, 9] módszer használja ezt a megközelítést.

6.4.1. PosOnly

Rövid betekintést nyújtunk az algoritmusba, a téma iránt mélyebben érdeklődők a [13, 9] cikkekben találnak több információt.

Az adatot a szokásos módon egy tulajdonságokat tartalmazó vektorral, x, és az osztályok címkéjét tartalmazó vektorral, $y = 0, 1$ írjuk le. Emelett bevezetünk egy újabb bináris vektort, s:

$$s = \begin{cases} 1, & \text{ha x-hez tartozik y,} \\ 0 & \text{egyébként.} \end{cases}$$

A tanulás célja itt a következő függvény: $f(x) = p(y = 1|x)$. Megmutatták, hogy ebben az esetben ez ekvivalens a

$$f(x) = p(s = 1|x)/p(s = 1|y = 1)$$

függvénnyel, ahol $p(s = 1|y = 1)$ egy konstans faktor. Ezt a konstanst egy validációs halmaz segítségével lehet becsülni. Ez azt jelenti, hogy ilyenkor a kapott feltételes valószínűség egy konstans faktorral különbözik csak az eredetileg kiszámítandó feltételes valószínűségtől. A [9] cikkben a szerzők mutatnak egy lehetséges becslést $p(s = 1|y = 1)$-re.

$$p(s = 1|y = 1) = \frac{\sum\limits_{x \in P} p(s = 1|x)}{\sum\limits_{x \in V} p(s = 1|x)}, \tag{6.5}$$

ahol P a már osztályozott alhamlaza a validációs halmaznak V.

6.4.2. SIRENE

SIRENE egy szupport vektor gép (*support vector machine*, SVM) alapú tanítási algoritmus. A feladatot felbontják sok kisebb részre, és minden egyes TF esetén tanítanak egy SVM-et. A Gauss-féle radiális bázisfüggvény kernelt használják fel.

$$K(x, y) = \exp\left(-\frac{\|x - y\|}{2\sigma^2}\right). \tag{6.6}$$

Az osztályozás megadja, hogy melyik gének hasonlítanak leginkább a TF által ismerten szabályozott génekhez. A tanítóhalmaz létrehozásához a korábban kísérletekkel validált

TF–gén párokat használja. Mivel elsősorban pozitív mintákat publikálnak (a TF kötődik az adott génhez és befolyásolja a gén expresszióját) ezért itt is kezelni kell a negatív mintákkal való egyensúlyozást. Erre a következő megoldást használják. Veszik az összes olyan gént, amiről nem ismert, hogy a TF kötődik-e a gén szabályozó régiójához. Ezt a halmazt 3 csoportra osztják. Háromszor végzik el a tanítást, és minden esetben az egyik halmaz tesztként funkcionál a másik kettő alkotja a negatív tanítókészletet. Így annak ellenére, hogy az esetleges hamis negatívok rosszul lesznek osztályozva, van esély arra, hogy kiszűrjék őket, és megfelelően osztályozzák.

6.5. TF, miRNS, mRNS szabályozó hálózatok

A genetikai szabályozás komplex hálózatokat eredményez, mivel az egyes szabályozó elemek, mint a miRNS-ek vagy a TF-ok nemcsak egyéb géneket szabályoznak, hanem egy miRNS hatással lehet egy transzkripciós faktor fehérje expressziójára is, míg egy TF serkentheti vagy gátolhatja egy miRNS érését is. Ebben a részben bemutatunk egy módszert [14], melynek segítségével komplex szabályozó hálózatokat lehet tanulni expressziós adatból 3 lépésben.

1. adatelőkészítés
2. hálózattanulás és integráció
3. hálózatinferencia

Az első lépésben az expressziós adatot normalizáljuk és meghatározzuk az egyes fenotípusok között különbözőképp expresszált géneket, miRNS-eket és TF-okat. A hálózat kezdeti struktúráját az egyes adatábzisok alapján becsült kapcsolatokból építjük fel, ehhez szükséges a kérdéses szabályozó elemek és gének kapcsolatáról az adatbázisokból információt letölteni. Teljesen nem hagyatkozhatunk az adatbázisokra, mivel általában szekvencia alapján becsült szabályozó–cél párok szerepelnek bennük, amelyek csak részben adnak megbízható eredményt.

A hálózattanulásához az expressziós adatot fenotípusonként felbontjuk és minden fenotípusból egy feltételt készítünk. Azért, hogy ne egy NP-nehéz keresést kelljen végrehajtani a gráfok terében, csak a páros gráfok terében keresünk, ahol a következő párokat nézzük: miRNS–TF, miRNS–mRNS, TF–TF, TF–miRNS, TF–mRNS. A prior, kezdeti hálózat struktúráját az adatbázisok alapján építjük fel, és a tanuló folyamat során minden kapcsolatot kiértékelünk egy Bayes-i pontozással. A pontozás alapján megbízható kapcsolatok kerülnek felhasználásra a bootstrap és integrációs fázisban. Bootstrap algoritmusra az esetek általában kis száma miatt van szükség a statisztikailag magasabb szignifikancia elérésére. Emellett ebben a lépésben integráljuk az egyes korábban kialakított és eddig külön tanult feltételeket. A $p < 0.05$ szignifikancia szinttel rendelkező kapcsolatokat vesszük be a teljes hálózatba.

Végül a hálózatinferenciát alkalmazunk motívumkereséssel. Azok a motívumok, amelyek a random gráfokban szignifikánsan kisebb valószínűséggel fordulnak elő, lesznek az eredmény fő építőelemei.

Irodalomjegyzék

[1] K. Chen and N. Rajewsky, The evolution of gene regulation by transcription factors and microRNAs. *Nat Rev Genet*, 8(2):93–103, 2007.

[2] T. W. Nilsen, Mechanisms of microRNA-mediated gene regulation in animal cells. *Trends in Genetics*, 23(5):243–249, 2007.

[3] A. Ay and D. N. Arnosti, Mathematical modeling of gene expression: a guide for the perplexed biologist. *Critical reviews in biochemistry and molecular biology*, 46(2):137–151, 2011.

[4] M. Teng, S. Ichikawa, L. R. Padgett, Y. Wang, M. Mort, D. N. Cooper, D. L. Koller, T. Foroud, H. J. Edenberg, M. J. Econs, et al., regSNPs: a strategy for prioritizing regulatory single nucleotide substitutions. *Bioinformatics*, 28 (14):1879–1886, 2012.

[5] V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, et al., TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic acids research*, 34(suppl 1):D108–D110, 2006.

[6] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L.-C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, et al., Gene prioritization through genomic data fusion. *Nature biotechnology*, 24(5):537–544, 2006.

[7] F. Jacob and J. Monod, On the Regulation of Gene Activity. *Cold Spring Harbor Symposia on Quantitative Biology*, 26:193–211, 1961.

[8] S. Oehler, E. R. Eismann, H. Krämer, and B. Müller-Hill, The three operators of the lac operon cooperate in repression. *The EMBO journal*, 9(4):973, 1990.

[9] L. Cerulo, C. Elkan, and M. Ceccarelli, Learning gene regulatory networks from only positive and unlabeled data. *BMC Bioinformatics*, 11(1):228, 2010.

[10] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. D. Favera, and A. Califano, ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context, *BMC bioinformatics*, 7(Suppl 1):S7, 2006.

[11] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner, Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biology*, 5(1):e8, 2007.

[12] S. Liang, S. Fuhrman, R. Somogyi, et al., REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. *Pacific symposium on biocomputing*, vol. 3, pp. 18–29, 1998.

[13] C. Elkan and K. Noto, Learning Classifiers from Only Positive and Unlabeled Data. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pp. 213–220, New York, NY, USA, 2008. ACM.

[14] T. D. Le, L. Liu, B. Liu, A. Tsykin, G. J. Goodall, K. Satou, and J. Li, Inferring microRNA and transcription factor regulatory networks in heterogeneous data. *BMC Bioinformatics*, 14:92, 2013.

7. fejezet

Genetikai asszociációs vizsgálatok standard elemzése

7.1. Bevezetés

A genetikai asszociációs vizsgálatok célja, hogy feltárja a különféle méréstechnikák által mért genotípusok gyakorisága és a vizsgált fenotípusok közötti statisztikai függőségeket. A leggyakoribb az eset–kontroll vizsgálat, ahol egynukleotidos polimorfizmusok (single nucleotide polymorphism - SNP) és egy bináris, betegségstátuszt leíró változó közötti statisztikai függőség elemzésére kerül sor. Ha egy adott SNP lehetséges genotípusainak eloszlása szignifikánsan eltér betegeknél a kontrollokhoz képest, akkor az annak a jele, hogy az adott SNP valamilyen szerepet játszik az adott betegség mechanizmusában. A méréstechnikák gyors fejlődése jelentős változást eredményezett a genomikai vizsgálatok kialakításában és az eredményének feldolgozásában. A kezdetben néhányszor 10-100 SNP együttes mérését, melyet manapság *kandidáns génasszociációs vizsgálatnak* (Candidate Gene Association Study - CGAS) nevezünk, felváltotta az 1000-10000 nagyságrendű teljes genom asszociációs vizsgálatok (Genome-Wide Association Study - GWAS) sora. Ezek azonban sok esetben nem váltották be a hozzájuk fűzött reményeket, azaz számos multifaktoriális betegség (pl.: asztma, obezitás) genetikai hátterének megfejtése továbbra is várat magára. Ennek egyik lehetséges oka a környezeti tényezők, fenotípusok nem megfelelő mérése, vizsgálatának hiánya, a másik a rendelkezésre álló statisztikai eszközök korlátai, legfőképp a többszörös hipotézistesztelés miatti korrekció. Mindezek miatt újra előtérbe kerültek az olyan CGAS-ok, melyek részletes környezeti és fenotípus-leírók figyelembevétele mellett vizsgálják a statisztikai függőségeket. Ebben a fejezetben olyan statisztikai módszereket és eszközöket mutatunk be, melyeket gyakran alkalmaznak génasszociációs vizsgálatok elemzéséhez.

7.2. Genetikai adattranszformáció

A megfelelő elemzés előfeltétele egy jól előkészített adathalmaz, amit genetikai adatok esetében nem lehet eléggé hangsúlyozni. Számos hibaforrás lehetséges (úgymint mérési hibák, nem megfelelő minőségű biológiai minta, adatfeldolgozási hibák), ezért fontos az adathalmaz alapos vizsgálata.

7.2.1. Szűrés

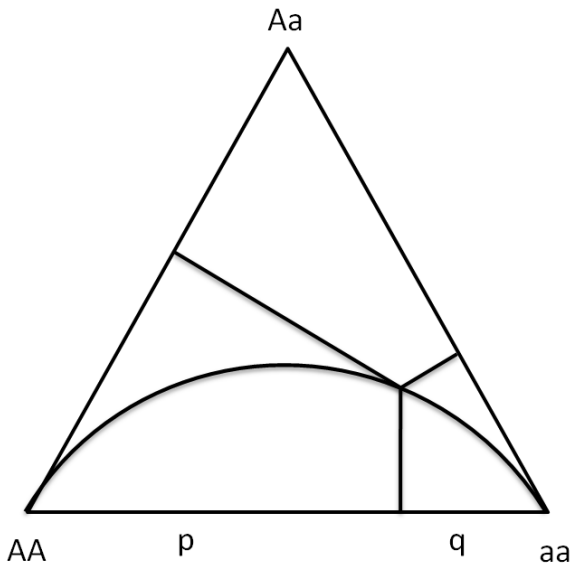
Feltételezve, hogy az adathalmaz már átesett egy alapszintű feldolgozáson egy genotipizáló műszer által (a mérési hibák jelölése megtörtént az adathalmazban), az adathalmaz vizsgálatát a hibás elemek szűrésével kezdjük. A szűrés célja a nem megfelelő adatcellák eltávolítása minták elhagyásával vagy változók kizárásával. Ehhez két küszöbértéket kell meghatároznunk: egyfelől a hiányzás arányát változónként (HAV), másfelől a hiányzás arányát mintánként (HAM). Első lépésben a (majdnem) teljesen hiányzó SNP-eket távolítjuk el, melyeknél a HAV$>95\%$. Ezt követően az adathalmaz mérete és a minták minőségének függvényében végezzük a szűrést. Egy nagyméretű adathalmaz esetén, amely jó minőségű mintákat tartalmaz, szigorú szűrési küszöbértéket alkalmazhatunk a minták szűrésére, úgymint HAM: $5-10\%$. A gyakorlatban jellemzően ennél jóval engedékenyebb küszöbértéket kell alkalmaznunk, HAM: $20-25\%$. Olyan esetben azonban, mikor a mintaszám alacsony, illetve a minták minősége közepesnél nem jobb, akkor akár 50% is lehet ez az érték. Mindezek mellett, ha a célváltozó vagy valamelyik központi fontosságú leíró értéke hiányos, akkor a mintát ki kell zárni az elemzésből, függetlenül a további hiányzás mértékétől. A választott küszöbérték feletti hiányzással rendelkező minták elhagyását követően a változók szűrésére kerül sor. Ezt a küszöbértéket szintén az adathalmaz minőségének függvényében kell megválasztanunk. Az $5\%, 10\%$ és 20%-os értékeket sorrendben szigorú, közepes és engedékeny küszöböknek tekinthetjük. Egy további lépésben figyelmet kell fordítanunk a változók értékkészletére. Mindazon változókat, melyek csak egy lehetséges értékkel rendelkeznek, mint például monomorf SNP-eket, el kell távolítanunk. Általánosságban az 1% alatti variabilitást mutató változókat (azaz a változó egyik értéke az adathalmaz kevesebb, mint 1%-ban vagy kevesebb, mint 10 mintában szerepel) el kell távolítani.

A hiányzó genotípus-értékek pótlására (imputációjára) több módszer alkalmas, közülük a legegyszerűbb az adott genotípus eloszlásán alapuló véletlen mintavétel.

7.2.2. Hardy–Weinberg-egyenlőség vizsgálata

Az adathalmaz szűrését követő lépésben a Hardy–Weinberg-féle egyenlőségi állapot (HWE) vizsgálatára kerül sor minden egyes SNP esetében. A HWE kimondja, hogy az allél, illetve genotípus-frekvenciák nem változnak generációk között amennyiben nincsenek jelen evolúciós hatások, úgymint mutáció, genetikai sodródás, illetve nem véletlenszerű párosodás. Egy kétallélú (A és a) genetikai jegy esetében, melynek allél-gyakorisága p és q, a genotípusok várható gyakorisága p^2 gyakori homozigóta genotípusra (AA), $2pq$ heterozi-

góta genotípusra (Aa) és q^2 a ritka homozigóta genotípusra (aa). Ezeket a gyakoriságokat $p^2, 2pq, q^2$ Hardy–Weinberg-hányadoknak nevezzük, melyek összege egyet tesz ki, azaz kielégítik a $p^2+2pq+q^2 = 1$ egyenletet. Egy kétallélú jegy genotípus hányadait ábrázolhatjuk egy de Finetti-diagram segítségével is (7.1. ábra). A háromszögben ábrázolt ív a Hardy–Weinberg-parabolának felel meg, mely azon pontok összességét fedi le, melyeknél a HWE fennáll.



7.1. ábra. De Finetti-diagram

A HWE-től való eltérés kimutatható a Pearson-féle khi-négyzet-teszttel (részletek az asszociációs teszteket ismertető alfejezetben találhatók), melyhez az adathalmazban lévő megfigyelt értékeket és a HWE által diktált várható értékeket kell alkalmazni (lásd [J. E. Wigginton et al. 2005]). Szignifikáns eredmény esetében a HWE-t felételező nullhipotézist el kell vetnünk. Mindazon SNP-eket, melyeknél szignifikáns p-érték adódik a kontrollpopuláción végzett HWE-teszten, ki kell zárni az elemzésből, mivel kontrollok esetében ez legtöbbször mérési hibát jelez.

7.3. Fenotípus-adattranszformáció

A rendelkezésre álló fenotípus leíróktól, klinikai és környezeti faktoroktól függően további adatfeldolgozásra, transzformációra lehet szükség. A genetikai faktorok értékeivel szemben a fenotípus-, klinikai, illetve környezeti leírók értékei alapértelmezés szerint nem pótolhatók. Ebből kifolyólag e változók megfelelő előfeldolgozása alapvető fontosságú lehet az elemzés sikeressége szempontjából.

7.3.1. Transzformáció

Abban az esetben, ha több kvantitatív fenotípus-leíró adott, melyek célváltozóként (függő változóként) szolgálhatnak az elemzésben, akkor döntenünk kell, hogy egymástól függetlenül kezeljük, vagy egy komplex fenotípus-leíróvá transzformáljuk őket. Az első esetben annyi különálló elemzést kell elvégezni, ahány célváltozónak választott változó adott. Ennek következményeként a többszörös tesztelés miatt szigorúbb p-érték-küszöbök alkalmazására lesz szükség, ami ellehetetlenítheti az eredmények értelmezését (részletekért lásd az asszociációs tesztek alfejezetet). Mindez elkerülhető megfelelő változószelekcióval és transzformációval. Egy lehetséges megoldás, hogy főkomponens-analízissel (PCA) kiválasztjuk a lényeges fenotípus-elemeket, melyekből egy komplex fenotípus-leírót alakítunk ki [Zhang et al. 2012]. Ekkor az elemzésekben már ezt az összetett leírót használhatjuk. Megjegyezzük, hogy egy Bayes-i keretrendszerben ilyen összevonásra nincs szükség, az egyes célváltozók együttesen is vizsgálhatóak.

7.3.2. Diszkretizálás

Számos frekventista és Bayes-i módszer csak diszkrét (kategorikus) változókon alkalmazható, ezért szükség lehet a folytonos jellegű, kvantitatív fenotípus-, környezeti és klinikai faktorok diszkretizálására. Erre számos módszer áll rendelkezésre, köztük a legegyszerűbb az egyenlő szélességű kategóriákat alkalmazó módszer. A nagyobb statisztikai programcsomagokban (pl.: R) jellemzően több összetett diszkretizáló algoritmus elérhető.

7.4. Egyváltozós statisztikai módszerek

Az egyváltozós módszerek alapvető feltételezése az, hogy minden vizsgált faktor független egymástól, és emiatt a célváltozóval való függőségi kapcsolat vizsgálatára faktoronként külön kerül sor. Habár a faktorok egymástól való teljes függetlenségének feltételezése kis valószínűséggel állná meg a helyét, ez a megközelítés mégis elfogadható abban az esetben, ha csak a legszignifikánsabb faktorok azonosítása a célunk, melyek várhatóan hatékony biomarkerekhez vezetnek. A biomarkerek egy-egy betegség jelenlétét, illetve jellegük, súlyosságuk fokát képesek jelezni. Ilyen esetben az interakciók, függőségi mintázatok és más jegyek azonosítása háttérbe szorulhat. Az egyváltozós módszerek alkalmazásának egy másik oka lehet, hogy az összetett, nagy számítási igényű, többváltozós módszerekhez képest relatíve egyszerűek és hatékonyak. Sokféle egyváltozós statisztikai módszer alkalmazható GAS eredmények elemzésére, kezdve az általános asszociációs tesztektől, a hatáserősséget mérő odds ratioig [Balding 2006].

7.4.1. Standard asszociációs tesztek

A konvencionális (frekventista) keretrendszerben a statisztikai módszerek alapjául a hipotézistesztelés szolgál. Adott egy nullhipotézis, amely függetlenséget tételez fel a függő (célváltozó) és a független (magyarázó) változó között, illetve egy alternatív hipotézis,

amely vagy egy általános modellt, vagy GAS esetében speciális genetikai öröklési modellt (additív, domináns, recesszív) alapul véve asszociációt feltételez. Az asszociációs tesztek alapvető eleme a tesztstatisztika, amin a hipotézisek kiértékelése alapszik. Általánosan, a nullhipotézis akkor utasítható el, ha a kiszámított statisztikához tartozó szignifikanciaszint alacsonyabb egy előre meghatározott α küszöbértéknél. A leggyakrabban az $\alpha = 0{,}05$ értéket alkalmazzuk küszöbértékként. GAS esetében a Pearson-féle khi-négyzet-statisztika egy gyakran alkalmazott módszer, ami lehetővé teszi kategorikus változók (például betegségleírók és genetikai faktorok) közötti függőség vizsgálatát. A számítások elősegítésére a változók kardinalitásának (értékeik számosságának) megfelelő méretű kontingenciatáblázatot hozhatunk létre [Agresti 2002]. Például ha adott két bináris változó X (egy adott allél) és Y (egy vizsgált fenotípus), akkor egy 2×2 táblát hozunk létre.

7.1. táblázat. 2×2 kontingenciatábla

	Y=0	Y=1	
X=0	n_{00}	n_{01}	r_0
X=1	n_{10}	n_{11}	r_1
	c_0	c_1	t

A khi-négyzet-statisztikát az $X : Y$ változó értékpárok megfigyelt gyakorisága és a függetlenséget feltételező nullhipotézisnek megfelelő elvárt gyakoriság alapján számoljuk:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}, \tag{7.1}$$

ahol $O_{i,j}$ jelöli a megfigyelt és $E_{i,j}$ a várható gyakoriságát az i-adik sorban és a j-edik oszlopban lévő cellához tartozó értéknek. A várható gyakoriságot a megfigyelt értékek sor (r) és az oszlop (c) részösszegei alapján számolhatjuk:

$$E_{i,j} = \frac{(\sum_{m=1}^{c} O_{i,m}) \cdot (\sum_{n=1}^{r} O_{n,j})}{N}, \tag{7.2}$$

ahol N az összmintaszám. Ez a tesztstatisztika aszimptotikusan megközelíti a $(r-1)(c-1)$ szabadságfokú χ^2 eloszlást. Ha a számított Pearson-féle khi-négyzet-statisztika magasabb, mint a χ^2 eloszlás $\alpha = 0{,}05$ szignifikanciaszinthez tartozó kritikus értéke, akkor a függetlenséget feltételező nullhipotézis elvethető. Más megfogalmazásban, ha a számított statisztikához tartozó p-érték kisebb, mint $\alpha = 0{,}05$, akkor a nullhipotézis elvethető.

Tekintsük példaként a 2×2 kontingenciatáblát, melynek elemei, a megfigyelt gyakoriságok, illetve a sor és oszlop részösszegek a 7.2. táblázatban láthatóak. A feladatunk az, hogy megvizsgáljuk, hogy fennáll-e függőség X genetikai faktor és Y célváltozó között. A nullhipotézis szerint X és Y független egymástól, míg az alternatív hipotézis szerint X és Y függ egymástól. Az első lépés a várható gyakoriságok számítása a megfigyelt gyakoriságok alapján a nullhipotézis szerint.

Például az $X = 0, Y = 0$ értékpár esetén a megfigyelt gyakoriság 60, a várható gyakoriság pedig a sor és oszlop részösszegek, illetve a teljes mintaszám alapján számítható

7.2. táblázat. Mintapélda

	Y=0	Y=1	Σ
X=0	60	50	110
X=1	45	70	115
Σ	105	120	225

$105 \cdot 110 \setminus 225 = 51{,}33$. A második lépés a Pearson-féle khi-négyzet-statisztika számítása a megfigyelt és a várható gyakoriságok alapján:

$$\chi^2 = \frac{(60-51{,}33)^2}{51{,}33} + \frac{(50-58{,}67)^2}{58{,}67} + \frac{(45-53{,}67)^2}{53{,}67} + \frac{(70-61{,}33)^2}{61{,}33} = 5{,}37. \qquad (7.3)$$

A harmadik lépés a χ^2 eloszláshoz tartozó szabadságfokok (df) meghatározása a $df = (r-1)(c-1)$ összefüggés alapján. Mivel mindkét változó bináris, így mind a sorok (r), mind az oszlopok (c) száma 2, tehát a teljes szabadsági fok: 1. Az utolsó lépés a számított khi-négyzet-statisztika összevetése az $df = 1$ szabadságfokú χ^2 eloszlással, valamint a hozzátartozó p-érték meghatározása. A 5,37 khi-négyzet-értékhez 0,0205 p-érték tartozik. Ez a szignifikanciaszint kisebb az általánosan alkalmazott 0,05 szignifikancia-küszöbértéknél, másképp közelítve pedig a 0,05 szignifikancia-szinthez tartozó kritikus érték $df = 1$ esetén 3,84, amit meghalad a számított statisztika. Tehát összességében elvethetjük a függetlenséget feltételező nullhipotézist, és azt állíthatjuk, hogy a függés X és Y között szignifikáns. Fontos megjegyezni, hogy a 0,05 szignifikancia szint jelentése az, hogy annak a valószínűsége, hogy helytelenül vetjük el a nullhipotézist az pontosan 0,05. Egymást követő többszöri asszociációs vizsgálat elvégzésének azonban az a következménye, hogy összességében nő annak az esélye, hogy hamis pozitív eredményeket kapjunk (I. fajú hiba). Például egy 1000 SNP-et tartalmazó vizsgálatban, ha mindegyiket asszociációs tesztnek vetjük alá egy adott célváltozóval, akkor legalább 50 SNP esetében a véletlennek lesz köszönhető az, hogy szignifikáns lett a függés. Tehát a hamis pozitívak aránya elfogadhatatlanul magas lesz. Ezt a jelenséget többszörös (hipotézis) tesztelési problémának nevezzük. Feloldására különféle korrekciós módszereket dolgoztak ki. A legelfogadottabb megközelítés a p-értékek korrigálása például Bonferroni-korrekcióval [Dunn 1961] vagy a Benjamini–Hochberg-módszer alkalmazásával [Benjamini and Hochberg 1995], ami egyúttal a hamis felfedezési arányt (false discovery rate) hivatott kontrollálni. Egy másik lehetséges megközelítés szerint permutációs tesztekkel ellenőrizhető az eredmények validitása.

GAS esetében ezek a korrekciók jellemzően túl konzervatívak, és jelentősen megnehezítik az eredmények elemzését. Mindez új, a GAS eredmények elemzésére alkalmas statisztikai módszerek kialakítására sarkallta a kutatókat. A Bayes-i módszerek növekvő népszerűségnek örvendenek ezen a területen és előszeretettel alkalmazzák őket, mivel a többszörös tesztelési problémát normatív módon kezelik, egyfajta beépített korrekció segítségével.

7.4.2. Cochran–Armitage-trendteszt

A Cochran–Armitage-trendteszt a Pearson-féle khi-négyzet-próba egy speciális változata, melyben egy bináris és egy többértékű kategorikus változó közötti függőség vizsgálatára kerül sor [Cochran 1954, Armitage 1955]. A teszt lényege, hogy a többértékű változó kategóriái között sorrendezettséget (trendet) feltételez, tehát például a 0, 1, 2 kategóriák egy lehetséges értelmezése rendre alacsony, közepes, magas. Eset–kontroll típusú génasszociációs vizsgálatok esetén a Cochran–Armitage-trendtesztben szereplő bináris változó a vizsgált betegség státuszát leíró célváltozó (Target: T), amely megadja, hogy egy adott minta kontroll vagy eset. A többértékű változó S pedig egy vizsgált SNP-nek felel meg 0, 1, 2 értékekkel, melyek jellemző értelmezése rendre gyakori homozigóta, heterozigóta, ritka homozigóta (három lehetséges genotípust feltételezve).

7.3. táblázat. 2×3 kontingenciatábla

	S=0 (aa)	S=1 (ab)	S=2 (bb)	
T=0	n_{00}	n_{01}	n_{02}	r_0
T=1	n_{10}	n_{11}	n_{12}	r_1
	c_0	c_1	c_2	N

A 7.3. táblában szereplő mennyiségek alapján Cochran–Armitage-trendteszt ($CATT$) statisztikája a következőképpen számítható

$$\mathrm{CATT} = \sum_{j=1}^{k} w_j \cdot (n_{0,j} \cdot r_1 - n_{1,j} \cdot r_0), \tag{7.4}$$

ahol a w_j súlyok segítségével különböző típusú asszociációk detektálására hangolható a teszt. Génasszociációs vizsgálatok esetében a feltételezett öröklési módnak megfelelő beállítást célszerű használni, azaz ha

- b allél *domináns* a allélra nézve: $w = (0, 1, 1)$,
- b allél *recesszív* a allélra nézve: $w = (1, 1, 0)$,
- a és b allél *additív* (kodomináns): $w = (0, 1, 2)$.

A $CATT$ statisztika saját szórásával vett hányadosa aszimptotikusan a normális eloszláshoz közelít, ezért a Cochran–Armitage-trendteszt az alábbi hányadosra vonatkoztatott normalitás vizsgálattal is megvalósítható.

$$\frac{\mathrm{CATT}}{\sqrt{var(\mathrm{CATT})}} \sim N(0, 1), \tag{7.5}$$

ahol $var(\mathrm{CATT})$ a következő kifejezéssel adható meg:

$$var(\mathrm{CATT}) = \frac{r_0 \cdot r_1}{N}\left(\sum_{i=1}^{k} w_i^2 \cdot c_i \cdot (N - c_i) - 2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} w_i \cdot w_j \cdot c_i \cdot c_j\right). \tag{7.6}$$

Ha az elvárt trend (domináns, recesszív, additív) teljesül, akkor abban az esetben a trendteszt statisztikai ereje nagyobb lesz az általános khi-négyzet-tesztnél. A vizsgálni kívánt trendtől eltérő trend detektálására azonban nem lesz alkalmas. Génasszociációs vizsgálatoknál, különösen GWAS estében legtöbbször additív (lineáris) trend vizsgálatára alkalmazzák [Purcell et al. 2007].

7.4.3. Hatáserősség

Amíg az asszociációs tesztek célja feltárni, hogy két változó között szignifikáns-e a függés, addig a hatáserősség-mércék a függés erősségét határozzák meg kvantitatív módon. Az odds ratio a leginkább alkalmazott hatáserősség-mutató, ami megmutatja egy adott betegség vagy állapot kontextusában, hogy egy adott genetikai jegy hogyan befolyásolja az eset és kontroll populáció arányát [Balding 2006]. Tehát voltaképpen azt számszerűsíti, hogy az adott jegy védő (OR<1), kockázatnövelő (OR>1) vagy semleges (OR$=1$) szerepet tölt be az adott betegség szempontjából. A standard odds ratio kizárólag a populációk arányát veszi figyelembe; a többváltozós kapcsolatokat nem veszi számításba.

Jelölje $X_1, X_2, ..., X_n$ azon diszkrét változókat, melyek SNP értékeket (0, 1, 2) kódolnak, melyek a gyakori homozigóta, a heterozigóta és a ritka homozigóta genotípusoknak felelnek meg. Ekkor $X_i^{(s)}$ jelölje az X_i SNP-et s értékkel. Továbbá egy Y betegségleíró esetén (ahol $Y^{(0)}$: kontroll, $Y^{(1)}$: eset) az *odds* a következőképp definiálható:

$$o_{X_i^{(s)}} = \frac{p(Y^{(1)}|X_i^{(s)})}{p(Y^{(0)}|X_i^{(s)})}. \tag{7.7}$$

Ennek alapján az *odds-ratio* (OR) például egy heterozigóta (1) versus gyakori homozigóta (0) esetben ily módon adható meg:

$$OR_{X_i^{(1,0)}} = \frac{o_{X_i^{(1)}}}{o_{X_i^{(0)}}}. \tag{7.8}$$

Következésképpen a log OR a következő alakban állítható elő:

$$\log OR_{X_i^{(1,0)}} = \log O_{X_i^{(1)}} - \log O_{X_i^{(0)}} = \log \frac{p(Y^{(1)}|X_i^{(1)})}{p(Y^{(0)}|X_i^{(1)})} - \log \frac{p(Y^{(1)}|X_i^{(0)})}{p(Y^{(0)}|X_i^{(0)})}. \tag{7.9}$$

A megfigyelt adathalmazból számított odds ratiora tekinthetünk úgy, mint egy genetikai jegy hatáserősségének teljes populációra vonatkozó becslésére. E tekintetben érdemes megvizsgálni e becslés megbízhatóságát. A *konfidenciaintervallum* az az értéktartomány, ahol az odds ratio értéke található, ha a vizsgálatot megismétlik más mintával. Az intervallumhoz tartozó konfidenciaszint azt a gyakoriságot adja meg, amilyen gyakran az odds ratio az adott tartományban tartózkodik a vizsgálatok ismétlése során. A leggyakrabban vizsgált tartomány a 95%-os konfidenciaintervallum, aminek tehát az a jelentése, hogy 100 ismétlésből 95 esetben ebbe a tartományba fog esni az odds ratio. A konfidenciaintervallum a megközelítőleg normális eloszlást követő ($N(\log(\text{OR}),\sigma^2)$) log odds ratio standard hibájának segítségével számítható.

$$SE = \sqrt{\frac{1}{n_{00}} + \frac{1}{n_{01}} + \frac{1}{n_{10}} + \frac{1}{n_{11}}}, \tag{7.10}$$

ahol n_{jk} jelöli azon esetek számát, ahol X_i^j és Y^k. Erre építve a log odds ratio (L) 95% konfidenciaintervalluma (CI) megadható úgymint $CI = L \pm 1{,}96 \cdot SE$. Tehát a CI nem más, mint [OR\ exp(1,96·SE), OR· exp(1,96·SE)]. Tekintsük példaként a 7.2. táblázatban ismertetett adathalmazt. Ennek odds ratioja és konfidenciaintervalluma a következőképp számítható:

$$OR_{X^{(1,0)}} = \frac{o_{X^{(1)}}}{o_{X^{(0)}}} = \frac{50/60}{70/45} = \frac{0{,}833}{1{,}556} = 0{,}536 \tag{7.11}$$

$$SE = \sqrt{1/60 + 1/50 + 1/45 + 1/70} = 0{,}2705 \tag{7.12}$$

$$95\%CI_{Low} = OR_{X^{(1,0)}} / \exp(1{,}96 \cdot SE) = 0{,}536/1{,}699 = 0{,}3154 \tag{7.13}$$

$$95\%CI_{High} = OR_{X^{(1,0)}} \cdot \exp(1{,}96 \cdot SE) = 0{,}536 \cdot 1{,}699 = 0{,}9108. \tag{7.14}$$

Ez azt jelenti, hogy X-nek védő hatása van Y betegségre nézve $OR = 0{,}536$ értékkel, és (0,3154 – 0,9108) közötti 95%-os konfidenciaintervallummal. Mivel a 95% CI ez esetben nem tartalmazza a semleges hatáserősséget jelentő 1-es odds ratiot, ezért ez a hatáserősség szignifikánsnak tekinthető.

7.4.4. Egyváltozós Bayes-i módszerek

A Bayes-i módszerek alapvető paradigmája, hogy egy *a priori eloszlás* P(A) és egy *likelihood* P(B|A) alapján az *a posteriori valószínűség* P(A|B) számítható a Bayes-tétel segítségével. Az a priori valószínűség (prior) lehetőséget ad az a priori tudás, illetve egyéb előzetes feltevések felhasználására. Míg ezzel szemben a likelihood kizárólag az adatra épülő mennyiség.

Egyváltozós Bayes-i módszerek esetében gyakori a normális eloszlású vagy kevert normális eloszlású priorok alkalmazása. Egy további eshetőség a normális exponenciális gamma (NEG) priorok használata [Stephens and Balding 2009]. A priorokat a hatáserősségek függvényében is lehetséges definiálni úgy, hogy a nem semleges hatású SNP-ek arányát adjuk meg a teljes vizsgált SNP halmazhoz képest (π), például $\pi = 10^{-4}$ vagyis 1 a 10.000-ből [Stephens and Balding 2009].

A log Bayes-faktor egy egyváltozós mutató, amelyet egyre gyakrabban alkalmaznak GAS eredmények elemzésénél. Különféle implementációi léteznek, mint például a SNPtest [Marchini et al. 2007] programban. A Bayes-faktor voltaképpen két különböző modellhez tartozó marginális likelihoodok aránya. Ha a vizsgált modellek (melyek tartalmazzák X és Y változókat) közül az egyik a függetlenséget feltételező nullmodell (M_0), a másik pedig egy függőséget megengedő alternatív modell (M_1), akkor ez a modellkiválasztás alapú mutató lehetővé teszi X és Y változók közötti függőség vizsgálatát. A modellek közötti különbséget kvantifikálja ez a mennyiség a megfigyelt adaton D, a modellek feltevésein

(M_0, M_1) és azok paraméterezésein (θ_0, θ_1) alapulva:

$$BF = \frac{P(D|M_1)}{P(D|M_0)} = \frac{\int P(\theta_1|M_1)P(D|\theta_1, M_1)d\theta_1}{\int P(\theta_0|M_0)P(D|\theta_0, M_0)d\theta_0}, \tag{7.15}$$

ami a Laplace-approximáció segítségével közelíthető [Marchini et al. 2007].

Látható, hogy ezek a módszerek a SNP-eket egymástól független entitásoknak tekintik, ami egyrészt nem valósághű, másrészt az interakciókban és a komplex függőségi hálózatokban lévő értékes információ így elvész.

7.5. Többváltozós módszerek

A többváltozós módszerek egyfelől lehetővé teszik a komplex függőségi mintázatok vizsgálatát, másfelől rendszerint nagy számítási igénnyel rendelkeznek. A kétértékű kategorikus fenotípus-változók esetében a logisztikus regresszió egy gyakran alkalmazott elemzési eszköz, amely használható mind egy-, mind többváltozós elemzéshez.

7.5.1. Logisztikus regresszió

A logisztikus regresszió egy bináris célváltozó esetén alkalmazott regressziós elemzési módszer [Agresti 2002]. A magyarázó változók (faktorok) értékein alapulva létrehozható egy logisztikus regressziós modell, ami lehetővé teszi azon esély (odds) jóslását, miszerint egy adott minta az esetek közé tartozik. A logisztikus regresszió alapja a logisztikus függvény, ami 0 és 1 között vesz fel értékeket.

$$F(z) = \frac{1}{(1 + e^{-z})}, \tag{7.16}$$

ahol z jelöli az $X_1, X_2, \ldots X_k$ magyarázó változók lineáris kombinációját oly módon, hogy

$$\pi(x) = \frac{1}{(1 + e^{\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_kx_k})}, \tag{7.17}$$

ahol $\pi(x)$ annak a valószínűsége, hogy a célváltozó „eset”. A β_0-át konstansnak (intercept) a többi β_i-t pedig *regressziós koefficiensnek* nevezzük. Felhasználva $\pi(x)$-et a log odds (lo) felírható ebben az alakban:

$$lo(x) = \frac{\pi(x)}{1 - \pi(x)} = \beta_0 + \beta_1x_1 + \beta_2x_2 + \ldots + \beta_kx_k, \tag{7.18}$$

melyet logit függvények nevezünk (bal oldal), és ami jelen esetben ekvivalens egy lineáris regressziós kifejezéssel (jobb oldal). Ez az átalakítás teszi lehetővé lineáris regresszió illesztését a log odds-ra. Legtöbbször maximum likelihood becslés segítségével kerül sor a β_i regressziós koefficiensek megadására. Ehhez egy több lépésből álló iteratív folyamat szükséges, mivel nincs zárt alakja a koefficiensek likelihood függvényre történő maximalizálásának. Egy kezdeti megoldást javít ez a folyamat iteratív módon, amíg el nem ér

egy konvergens állapotot, azaz ahonnan már nem lehet javítani, vagy az is lehetséges, hogy ilyen állapot egyáltalán nem érhető el. A logisztikus regressziós modell részét képzik mindazok a magyarázó változók, amelyek nem nulla regressziós koefficienssel rendelkeznek. Bár ez ebben a formában egy többváltozós modell, az egyes faktorok egyéni hozzájárulása a modellhez mérhető Wald- vagy „likelihood ratio" teszt által. A Wald-teszt a Wald-statisztikára épül $W = (\beta_i^2/SE_{\beta_i}^2)$, melynek eloszlása közelíthető χ^2 eloszlással. Ennek megfelelően a teszt szignifikanciájának meghatározása a khi-négyzet-tesztéhez hasonlóan történik.

7.5.2. Haplotípus-asszociáció

A SNP-k együttes vizsgálatára kézenfekvő választás a haplotípus szintű asszociációs elemzés. Ekkor a haplotípust formáló SNP-ek ($H_1 : \{S_1, S_2, S_3\}$) lehetséges allélvariánsainak (pl.: $S_1 : A/G, S_2 : C/T, S_3 : G/A$) kombinációjaként állnak elő a haplotípus lehetséges értékei (pl.: ACG,ACA,ATA,ATG,GTG,GTA,...). Az így létrehozott többértékű változó célváltozóval vett függőségének vizsgálatára különféle módszerek alkalmazásával nyílik lehetőség, melyeknek alapvetően két lényeges problémát kell kezelniük: a (1) haplotípus fázisinformáció hiánya és (2) a haplotípus értékkészletének nagysága [Liu et al. 2008].

A fázisinformáció megadja, hogy az adott allél az anyai vagy az apai kromoszómán található, ennek hiányában minden kombinációs lehetőséget figyelembe kell venni. A haplotípus-asszociációs metódusok egy része feltételezi, hogy rendelkezésre áll a fázisinformáció (akár mérés, akár becslés által), a módszerek egy másik része pedig integráltan tartalmazza a fázisinformáció becslését.

A haplotípus értékkészletének számossága azért jelenthet gondot, mert jellemzően nem elegendő a mintahalmaz ahhoz, hogy a legritkább haplotípus-variánsokat is statisztikailag elégséges mértékben tartalmazza. Például ahhoz, hogy egy 4 biallélikus SNP-et (pl.: A/G esetében AA, AG, GG genotípust) tartalmazó haplotípus (melynek kardinalitása: $3 \times 3 \times 3 \times 3 = 81$) minden lehetséges variánsához elégséges mintaszám (> 10) álljon rendelkezésre, 810 mintára lenne szükség egyenletes gyakoriságot feltételezve. Valójában azonban nem helytálló az egyenletesség feltételezése, ehelyett jellemzően egy pár gyakori haplotípusérték mellett a lehetséges variánsok nagy része ritka, azaz 1% alatti gyakoriságú. A ritka haplotípusok kezelésének egy lehetséges módja a hasonlóság alapú összevonás például hierarchikus klaszterezéssel [Durrant et al. 2004] vagy evolúciós fa alapú valószínűségi klaszterezéssel [Tzeng 2005]. Egy további lehetséges módszer a súlyozott log-likelihood alapú megközelítés [Souverein et al. 2006].

Haplotípus-asszociációs teszt

A haplotípus asszociációs tesztek legegyszerűbb változata azt vizsgálja, hogy a haplotípus eloszlása az eseteknél és kontrolloknál különbözik-e (ez az ún. goodness-of-fit teszt). Ehhez egy likelihood-arány statisztika (LHR) készíthető, melynek általános formája

$$LHR = 2(\ln L_H^{eset} + \ln L_H^{kontroll} - L_H^{total}), \tag{7.19}$$

amely aszimptotikusan χ^2 eloszlást követ $H-1$ szabadságfokkal nullhipotézis esetén, ahol H a lehetséges haplotípusok száma. Ennek hátránya, hogy nagyszámú haplotípus esetén kicsi lesz a teszt statisztikai ereje egy lehetséges asszociáció detektálására, továbbá előfordulhat, hogy olyannyira kevés a minta, hogy a nullhipotézishez tartozó eloszlás nem χ^2 eloszlást követ.

Egy lehetséges megoldás a nemlineáris transzformációk alkalmazása a haplotípusok eloszlásán oly módon, hogy a transzformáció felnagyítsa a különbséget az eset és a kontroll haplotípusok között. Ennek következtében az alkalmazott χ^2 teszt statisztikai ereje megnövekszik [Zhao et al. 2006].

Mivel egy GAS során rendszerint több lókusz vizsgálatára kerül sor egyszerre, így nem elhanyagolható a többszörös hipotézistesztelés okozta probléma, amelyet megfelelő korrekcióval kezelni kell. Erre a célra az egyik gyakran alkalmazott módszer a permutációs tesztelés, amelyet az egyik népszerű haplotípus-asszociációt vizsgáló programcsomag, a Haploview is alkalmaz [Barrett et al. 2005].

Haplotípus-megoszlás

A haplotípus-megoszlást vizsgáló módszerek arra fókuszálnak, hogy az egyes mintahalmazokon belül mennyire hasonlóak a haplotípust alkotó allélok. Tehát egy adott L lókusz és $s(.)$ hasonlósági mérce esetén $U_1, U_2, \ldots U_N$ kontroll haplotípusokat és $V_1, V_2, \ldots V_M$ eset haplotípusokat vizsgálva négyféle haplotípus-megoszlást mérő metrika adható meg [Nolte et al. 2007].

A kontrollcsoporton belüli haplotípus-megoszlás:

$$\mathrm{HS}_{kontroll}(L) = \frac{2}{N(N-1)} \cdot \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} s(U_i, U_j, L). \tag{7.20}$$

A betegcsoporton belüli haplotípus-megoszlás:

$$\mathrm{HS}_{eset}(L) = \frac{2}{M(M-1)} \cdot \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} s(V_i, V_j, L). \tag{7.21}$$

A beteg és kontroll csoportok közötti haplotípus-megoszlás:

$$\mathrm{HS}_{kereszt}(L) = \frac{1}{NM} \cdot \sum_{i=1}^{N} \sum_{j=1}^{M} s(U_i, V_j, L). \tag{7.22}$$

Összesített haplotípus-megoszlás:

$$\mathrm{HS}_{total}(L) = \frac{N^2 \cdot HS_{kontroll}(L) + M^2 \cdot HS_{eset}(L) + 2NM \cdot HS_{kereszt}(L)}{(N+M)^2}. \tag{7.23}$$

Ezek segítségével különböző haplotípus-megoszlást tesztelő statisztikák hozhatók létre, úgymint a *HSS*-teszt és a *CROSS*-teszt [Nolte et al. 2007]. A *HSS*-teszt az eset és a

kontroll haplotípusok összehasonlításán alapszik, azzal a feltevéssel, hogy az eset haplotípusok közötti megoszlás nagyobb, mint a kontrollok közötti megoszlás. Ennek oka az, hogy jellemzően egy adott betegségre hajlamosító haplotípusok egymáshoz hasonlóak, míg a kontrollokhoz tartozó haplotípusok változatosabbak.

$$t_{HSS}(L) = \frac{HS_{eset}(L) - HS_{kontroll}(L)}{\sqrt{(\sigma(HS_{eset}(L)))^2 + (\sigma(HS_{kontroll}(L)))^2}}, \tag{7.24}$$

ahol $\sigma(.)$ az adott haplotípus-megoszlásokhoz tartozó becsült szórást jelöli. Nagy mintaszám esetén $HS_{eset}(L)$ és $HS_{kontroll}(L)$ normál eloszlást követ, a köztük lévő eltérés szignifikanciája egy $N + M - 2$ szabadságfokú t-teszttel adható meg.

A *CROSS*-teszt ehhez képest azon alapszik, hogy az esetek és a kontrollok közötti haplotípus-megoszlás kisebb, mint két véletlenszerűen választott haplotípus között:

$$z_{CROSS}(L) = \frac{HS_{kereszt}(L) - HS_{total}(L)}{\sigma(HS_{kereszt}(L) - HS_{total}(L))}, \tag{7.25}$$

ahol $\sigma(.)$ jelöli a szórást. A $z_{CROSS}(L)$ statisztika eloszlása normál eloszlással közelíthető az extrém L értékeket leszámítva, ahol egy transzformációt követően χ^2 eloszlással becsülhető [Nolte et al. 2007].

További statisztikák is kialakíthatóak az ismertetett metrikák felhasználásával, melyek többsége az alábbi kvadratikus formában írható fel:

$$Q = \mathbf{H_v^t} \cdot A_v \cdot \mathbf{H_v} - \mathbf{H_u^t} \cdot A_u \cdot \mathbf{H_u}, \quad T = \frac{Q}{\sigma(Q)}, \tag{7.26}$$

ahol $\mathbf{H_v} = (h_{v1}, \ldots, h_{vr})$ és $\mathbf{H_u} = (h_{u1}, \ldots, h_{ur})$ haplotípus-eloszlást jelöl a beteg és a kontroll csoportok esetében, A egy szimmetrikus mátrix, melyet a tetszőleges i és j haplotípus közötti hasonlóságot leíró $K(H_i, H_j)$ szimmetrikus kernelfüggvény definiál, $\sigma(Q)$ pedig Q szórását jelöli. Amennyiben $\mathbf{H_u}$, illetve $\mathbf{H_v}$ szingularitástól mentes, akkor T megközelítőleg standard normális eloszlást követ [Tzeng et al. 2003].

Haplotípus-asszociáció vizsgálata regressziós modellekkel

A regressziós modellek egy előnye, hogy egyszerre teszik lehetővé egy adott haplotípus rekonstruálását (fázisinformáció nélküli adathalmaz esetén), illetve hatásának vizsgálatát. A regresszión alapuló módszereket *prospektív* illetve *retrospektív likelihood* számítást végző csoportokba sorolhatjuk.

Jelölje G_i a megfigyelt genotípus-információt, $H_i(h_i, h_i*)$ egy lehetséges haplotípust (anyai és apai haplotípuspárt) az i-adik mintánál. $P(H_i)$ jelölje a $H_i(h_i, h_i*)$ haplotípus apriori valószínűségét, $\mathbf{Z}$ jelölje a betegségre való hajlamot befolyásoló környezeti tényezőket (pl.: életkor, nem, dohányzás), Y pedig a betegség jelenlétét tükröző változót. Továbbá $S(G_i)$ legyen azon haplotípusok halmaza, melyek konzisztensek az i-adik mintánál megfigyelt $G_i = g_i$ genotípussal. Mindezek segítségével a vizsgált adaton alapuló prospektív

likelihood a következőképp számítható [Schaid 2004]:

$$L_{pro} = \prod_{i=1}^{N^D} \sum_{H_i \in S(G_i)} P(Y_i|\mathbf{Z_i}, H_i, \beta) \cdot P(H_i), \tag{7.27}$$

ahol β a regressziós koefficiensek vektorát jelöli, N^D pedig a teljes mintaszámot. E prospektív regressziós modell illesztése történhet maximum-likelihood [Lake et al. 2003], illetve *EM* alapú módszerekkel [Zhao et al. 2003].

A prospektív szemlélet lényege, hogy az adathalmazból kiindulva, a genotípus (G_i), haplotípus (H_i) és a környezeti faktorok ($\mathbf{Z_i}$) által hordozott információt felhasználva kerül sor a betegség megléte (Y_i) valószínűségének vizsgálatára. Ezzel szemben a retrospektív megközelítésnél a betegség leíró állapotából kiindulva vizsgáljuk a haplotípusok valószínűségét. Ennek megfelelően a retrospektív likelihood az alábbiak szerint fejezhető ki [Epstein and Satten 2003]:

$$L_{ret} = \prod_{g} [\sum_{H_i \in S(g)} P(H_i|Y_i = 0)]^{u_g} \cdot [\sum_{H_i \in S(g)} P(H_i|Y_i = 1)]^{v_g}, \tag{7.28}$$

ahol u_g és v_g a g genotípussal rendelkező kontroll-, illetve betegminták számát jelöli. A retrospektív likelihood előnye, hogy legalább akkora vagy nagyobb statisztikai erővel rendelkezik, mint a prospektív likelihood, azonban hátránya, hogy kevésbé robusztus a Hardy–Weinberg-egyenlőségtől való eltérésekre [Satten and Epstein 2004].

Egy további lehetőség a regressziós modellek általánosítása, a generalizált lineáris modell (GLM), mint statisztikai keretrendszer alkalmazása. A *GLM* alapvető feltevése, hogy a függő változó (esetünkben a betegségleíró) Y eloszlása megadható egy az exponenciális eloszlások családjába tartozó eloszlással, melynek várható értéke μ a független $\mathbf{X}$ változóktól (pl.: genotípus, környezeti faktorok) függ. A független $\mathbf{X}$ változók egy lineáris prediktort (η) alkotnak a nekik megfelelő β paraméterek lineáris kombinációjaként, azaz $\eta = \mathbf{X} \cdot \beta$. A prediktor η és az eloszlás várható értéke μ közötti kapcsolatot a $\mathcal{L}$ link függvény adja meg $\mu = \mathcal{L}^{-1}(\eta)$. Mindezek alapján tehát a GLM általános egyenlete az alábbi formát veszi fel:

$$E(Y) = \mu = \mathcal{L}^{-1}(\mathbf{X} \cdot \beta), \tag{7.29}$$

ahol E(.) a várható érték számítását jelöli. Megjegyezzük, hogy Y varianciája szintén a várható érték (μ) függvényeként fejezhető ki. A GLM mint keretrendszer felhasználható haplotípus-asszociációt mérő statisztika kialakítására az alábbi formában [Schaid 2004]:

$$W = \sum_{i=1}^{N^D} \frac{y_i - \bar{y}_i}{f(\phi)} \cdot E[H_i|G_i], \tag{7.30}$$

ahol y_i az i-adik minta betegségleíró értéke, $\bar{y}_i$ pedig a GLM-mel illesztett becslés, kizárólag környezeti faktorok alkalmazásával, $f(\phi)$ pedig egy normalizációs faktor a GLM-ben használt eloszlásnak megfelelően. $E[H_i|G_i]$ a haplotípusok eloszlása felett számított feltételes várható értéket jelöli az adathalmaz által megadott genotípus függvényében. A W

statisztika voltaképp a környezeti faktorokat használó GLM modell reziduálisainak (a $\bar{y}_i$ becsléseknek a valós y_i értékekhez képest mért hibái) és a haplotípusok várható értékének kovarianciáját méri [Schaid 2004].

7.5.3. Statisztikai erő vizsgálata

A statisztikai erő (Pwr) azt fejezi ki, hogy egy statisztikai teszt mekkora valószínűséggel veti el a nullhipotézist($T^{H_0} = 0$), amikor az valóban hamis ($H_0 = 0$), azaz $Pwr = p(T^{H_0} = 0|H_0 = 0)$. Ez voltaképpen a II. fajú hiba, vagyis a hamis negatív ráta (FNR) ellentéte ($Pwr = 1 - FNR$). A statisztikai erőt alapvetően három fő faktor befolyásolja:

1. *Mintaszám.* A rendelkezésre álló minta nagysága lényeges tényező, hiszen minél több minta áll rendelkezésre, annál kisebb a mintavételezési hiba (a teljes populációhoz képest), azaz annál megbízhatóbb következtetéseket vonhatunk le.

2. *Hatáserősség.* A vizsgált genetikai vagy környezeti faktor hatáserőssége azért fontos szempont, mivel egy relatíve kis hatáserősségű faktor vizsgálatához több minta szükséges, mint egy hozzá képest nagy hatást mutató faktoréhoz.

3. *Szignifikanciaszint.* A statisztikai teszteknél alkalmazott küszöbérték, amely megadja annak a valószínűségét, hogy a statisztika alapján elvethető a nullhipotézis, holott valójában az igaz (I. fajú hiba, hamis pozitív ráta). Egyik leggyakoribb választás az $\alpha = 0{,}05$.

Számos más tényező befolyásolhatja ezeken kívül a statisztikai erőt, azonban ezek jellemzően kisebb hatásúak és az adott vizsgálat jellemzőitől függenek.

A statisztikai erő elemzésére sor kerülhet a priori, a vizsgálat (mintagyűjtés) elvégzése előtt, illetve post-hoc jelleggel a vizsgálatot (mintagyűjtést) követően. Az előbbi esetben az erőelemzés célja - adott szignifikanciaszint és hatáserősség mellett - a kitűzött statisztikai erőhöz szükséges mintaszám meghatározása. Míg post-hoc esetben a cél a ténylegesen rendelkezésre álló mintaszám alapján adódó statisztikai erő kiszámítása. Az erőelemzés a priori alkalmazása teljes mértékben elfogadott, a post-hoc felhasználás azonban vitatott, mivel a statisztikai erő függ a statisztikai teszttel elért p-értéktől. Különösen akkor adódhatnak félrevezető eredmények, amikor a minta eleve nem volt megfelelően nagy egy adott nagyságú hatás vizsgálatához.

A statisztikai erő számításának egy módja a bemutatott főbb tényezőket tartalmazó regressziós modell maximum-likelihood módszerrel történő illesztése. Ezt valósítja meg például a *Quanto* program [Gauderman and Morrison 2006] vagy az online elérhető *Genetic Power Calculator* [Purcell et al. 2003], de számos más statisztikai programcsomag is alkalmas a statisztikai erő számítására.

Irodalomjegyzék

[Agresti 2002] A. Agresti, *Categorical Data Analysis*. Wiley-Interscience, New York, 2002.

[Armitage 1955] P. Armitage, Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3):375–386, 1955.

[Balding 2006] D. J. Balding,A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.*, 7(10):781–791, 2006.

[Barrett et al. 2005] J. C. Barrett, B. Fry, J. Maller, and M. J. Daly, Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, 21(2):263–265, 2005.

[Benjamini and Hochberg 1995] Y. Benjamini and Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.*, 57(1):289–300, 1995.

[Cochran 1954] W. G. Cochran, Some methods for strengthening the common chi-squared tests. *Biometrics*, 10(4):417–451, 1954.

[Dunn 1961] O. J. Dunn, Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961.

[Durrant et al. 2004] C. Durrant, K. T. Zondervan, L. R. Cardon, S. Hunt, P. Deloukas, and A. P. Morris, Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. *Am. J. Hum. Genet.*, 75(1):35–43, 2004.

[Epstein and Satten 2003] M. P. Epstein and G. A. Satten, Inference on haplotype effects in case-control studies using unphased genotype data. *Am. J. Hum. Genet.*, 73(6):1316–1329, 2003.

[Gauderman and Morrison 2006] W. J. Gauderman and J. Morrison, QUANTO 1.1: A computer program for power and sample size calculations for genetic-epidemiology studies. 1–48, `http://hydra.usc.edu/gxe`, 2006.

[J. E. Wigginton et al. 2005] J. E. Wigginton, D. J. Cutler, and G. R. Abecasis, A note on exact tests of Hardy–Weinberg equilibrium, *Am J Hum Genet*, 76:887–893, 2005.

[Lake et al. 2003] S. L. Lake, H. Lyon, K. Tantisira, E. K. Silverman, S. T. Weiss, N. M. Laird, and D. J. Schaid, Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. *Hum. Hered.*, 55(1):56–65, 2003.

[Liu et al. 2008] N. Liu, K. Zhang, and H. Zhao, Haplotype-association analysis. *Adv Genet.*, 60:335–405, 2008.

[Marchini et al. 2007] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly, A new multipoint method for genome-wide association studies via imputation of genotypes, *Nature Genetics*, 39:906–913, 2007.

[Nolte et al. 2007] I. M. Nolte, A. R. deVries, G. T. Spijker, R. C. Jansen, D. Brinza, A. Zelikovsky, and G. J. teMeerman, Association testing by haplotype-sharing methods applicable to whole-genome analysis. *BMC Proc.*, 1(Supp 1):S129, 2007.

[Purcell et al. 2003] S. Purcell, S. S. Cherny, and P. C. Sham, Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics*, 19(1):149–150, 2003.

[Purcell et al. 2007] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. deBakker, M. J. Daly, and P. C. Sham, PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, 81(3):559–575, 2007.

[Satten and Epstein 2004] G. A. Satten and M. P. Epstein, Comparison of prospective and retrospective methods for haplotype inference in case-control studies. *Genet. Epidemiol.*, 27(3):192–201, 2004.

[Schaid 2004] D. J. Schaid, Evaluating associations of haplotypes with traits. *Genet. Epidemiol.*, 27(4):348–364, 2004.

[Souverein et al. 2006] O. W. Souverein, A. H. Zwinderman, and M. W. T. Tanck, Estimating haplotype effects on dichotomous outcome for unphased genotype data using a weighted penalized log-likelihood approach. *Hum. Hered.*, 61(2):104–110, 2006.

[Stephens and Balding 2009] M. Stephens and D.J. Balding, Bayesian statistical methods for genetic association studies. *Nature Review Genetics*, 10(10):681–690, 2009.

[Tzeng et al. 2003] J. Y. Tzeng, B. Devlin, L. Wasserman, and K. Roeder, On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *Am. J. Hum. Genet.*, 72(4):891–902, 2003.

[Tzeng 2005] J. Y. Tzeng, Evolutionary-based grouping of haplotypes in association analysis. *Genet. Epidemiol.*, 28(3):220–231, 2005.

[Zhang et al. 2012] F. Zhang, X. Guo, S. Wu, J. Han, and Y. M. Liu, Genome-wide pathway association studies of multiple correlated quantitative phenotypes using principle component analyses. *PLoS ONE*, 7(12):e53320, 2012.

[Zhao et al. 2003] J. Zhao, S. S. Li, and N. L. Khalid, A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. *Am. J. Hum. Genet.*, 72(5):1231–1250, 2003.

[Zhao et al. 2006] J. Zhao, L. Jin, and M. Xiong, Nonlinear tests for genomewide association studies. *Genetics*, 174(3):1529–1538, 2006.

8. fejezet

Génexpressziós adatok standard asszociációs elemzése

8.1. Bevezetés

A DNS molekula kettős hélixet alkot. A hélix szálai egymás tökéletes komplementerei: minden adeninnel szemben egy timin és minden guaninnal szemben egy citozin áll a másik szálon. A *hibridizáció* folyamata során a két komplementer DNS (vagy RNS) szál összekapcsolódik. A microarray-technológiák ezt használják ki: egy microarray-chip felszínéhez rengeteg egyszálú génszekvencia darabka (ún. próba) van hozzácsatolva, amellyel egy adott mintában található komplementer RNS molekula mennyiségét mérhetjük meg. Az RNS a DNS-ből származó genetikai üzenetet továbbítja (a gének megfelelő szakaszainak lemásolásával) a citoplazmába, ahol a fehérjék készülnek a génmásolatok aminosav-szekvenciákra való lefordításával. A microarray-k egyetlen kísérletben több tízezer gén expressziós szintjét (az RNS formájában tárolt üzenet mennyiségét) képesek megmérni. Megfestett RNS-t öntenek a microarray felületére, majd ha az RNS megtalálja a komplementer szekvenciáját az array felületén, akkor hibridizálódik hozzá. A mérés során a kibocsátott fény mennyisége elárulja, hogy az adott génhez mennyi RNS készült a mintában. Ez lehetővé teszi a kutatók számára, hogy hipotézismentes módon összehasonlítsák különböző biológiai rendszerek, folyamatok és betegség-állapotok transzkripciós profilját [1].

A mikroarray-ket a sok különféle célra használják: betegségek csoportosítására, illetve besorolására; egy adott kezelés *in vivo* vagy *in vitro* hatásainak azonosítására; betegséggének, vagy bizonyos folyamatokban részt vevő gének keresésére [2].

Ebben a fejezetben megpróbálunk egy rövid ízelítőt nyújtani abból, hogy hogyan történik egy mikroarray-kísérlet elemzése. A létező számítási módszerek és eszközök kimerítő áttekintése helyett arra fókuszálunk, hogy bemutassuk a leggyakrabban használt módszereket és az általános megközelítéseket. Először is, a próbák nyers intenzitásértékeinek megmérésétől hosszú út vezet a gének, illetve transzkriptumaik genomszintű expressziós szintjének meghatározásáig. A gyakorlatban számos forrásból származó variabilitás lép be, amelyet figyelembe kell venni, illetve a megfelelő módon kezelni kell: számos módosítást kell végezni, hogy megfelelően pontos eredményeket kapjunk. Ezeket a lépéseket összefog-

lalóan *előfeldolgozás*nak nevezzük, amelyről az 8.2. alfejezetben beszélünk részletesebben. A 8.3. alfejezetben az adatok és a biológiai kérdések közötti kapcsolatra koncentrálunk. Olyan kérdésekre keressük a választ például, hogy: Milyen gének fontosak egy adott szituációban? Két (vagy több) állapot között milyen gének expresszálódnak különbözőképpen? Milyen biológiai folyamatok játszódnak le egy adott szituációban?

Megjegyzés: a fejezet során egycsatornás mikroarray-kkel foglalkozunk, amelyekben egyetlen mintából származó RNS-t vizsgálnak egyszerre egy array-n. A kétcsatornás mikroarray-kkel nem foglalkozunk. (A két mintából származó RNS-t két különböző színnel festik meg, és egyszerre hibridizálják az array felszínéhez. A két szín intenzitásának aránya egy adott pontban a megfelelő két gén differenciális expressziójáról árulkodik a mintákban.)

8.2. Előfeldolgozás

Az előfeldolgozás öt lépésből áll [3]: (1) képelemzés, amely során a szkennelt képeken lévő képpontok intenzitásértékeit próba-szintű adatokká konvertáljuk, (2) háttérkorrekció, amelyben a lemért próba-intenzitások nem-specifikus hibridizációját és a háttérzajt kiszűrjük az intenzitásadatokból, (3) normalizáció, amely során több forrásból származó variabilitást korrigálunk annak érdekében, hogy a különböző array-kből származó mérések összehasonlíthatóak legyenek egymással, (4) összegzés, amelyben a próbák háttérzajra korrigált és normalizált intenzitásadatit összegezzük minden transzkripthez, amelyből az adott próba származik; és így egy olyan értéket kapunk, amely megbecsüli az adott transzkriptnek megfelelő RNS mennyiségét a mintában, végül (5) minőségellenőrzési lépés, amely során a kilógó mérési eredményeket, amelyek az elfogadhatónál nagyobb mértékű fluktuációval rendelkeznek, kiszűrjük.

8.2.1. Háttérkorrekció

A képelemzési lépés után (amellyel jelen fejezetben nem foglalkozunk) az előfeldolgozás első lépése, a háttérzaj hatásainak kiszűrése következik. Ez azért nagyon fontos, mert a háttérzaj erősen befolyásolja a differenciális expresszióra vonatkozó becsléseinket. Képzeljük el a következő esetet: Két különböző mintában egy adott gén valódi expressziójának mértéke legyen s_1 illetve s_2. A képpontok körül azonban közel egyenlő mértékben pozitív háttérzajt is érzékelünk, amelyek torzítják a méréseinket, legyen ezek szintje b_1 illetve b_2. Ebben az esetben a két gén expressziójának valódi aránya s_1/s_2, azonban a megfigyelt $(s_1 + b_1)/(s_2 + b_2)$ arány közelebb van 1-hez mint a valódi arány, és minél közelebb van a valódi expressziós szint a háttérzajhoz, annál inkább közelebb lesz a mért arány 1-hez.

Többféle háttérkorrekciós módszer létezik, például az RMA algoritmus háttérkorrekciós része, amelyet Irizarry és munkatársai fejlesztettek ki [4], vagy például az Affymetrix által kifejlesztett MicroArray Suite 5.0 (MAS) szoftver háttérkorrekciós algoritmusa [5].

8.2.2. Normalizáció

A normalizáció fő célja az, hogy a háttérzajra korrigált intenzitásadatokat módosítsa úgy, hogy a különböző mérésekből származó array-k összehasonlíthatóak legyenek. Általában a normalizációs módszerek a következő kategóriák valamelyikébe sorolhatók [6]: (1) skálázás, amely azt feltételezi, hogy minden egyes array-n az intenzitásoknak hasonló átlagúaknak kell lennie, vagy hasonló medián értékkel kell rendelkeznie; (2) kvantil-normalizáció, amely feltételezi, hogy minden egyes array-n a jelintenzitás-értékeknek azonos eloszlásúaknak kell lennie; (3) lokális regressziós (loess) normalizáció, amely azt feltételezi, hogy a technikai forrásból származó torzulás intenzitásfüggő, és egy loess-görbét illeszt ennek kiküszöbölésére és (4) modell-alapú normalizáció, amely bizonyos technikai forrásból származó varianciákra explicit módon modelleket illeszt, és ezek segítségével szűri ki a nem megfelelő varianciákat.

Skálázás. Válasszunk ki egy alap array-t, és a többi array-t skálázzuk át úgy, hogy a jelintenzitások átlagos vagy medián értéke legyen ugyanakkora, mint a kiválasztott alap array-n. Példaként lásd az 8.1. ábrát.

	Array1	Array2	Array3
Próba1	6,2	11,9	3,9
Próba2	4,8	7,8	9,2
Próba3	12,5	4,6	12,1
Próba4	6,3	3,9	4,5
Próba5	7,2	5,8	6,3
Átlag	7,4	6,8	7,2

$\Rightarrow$

Array1	Array2	Array3
6,2	$11{,}9*\frac{7{,}4}{6{,}8}$	$3{,}9*\frac{7{,}4}{7{,}2}$
4,8	$7{,}8*\frac{7{,}4}{6{,}8}$	$9{,}2*\frac{7{,}4}{7{,}2}$
12,5	$4{,}6*\frac{7{,}4}{6{,}8}$	$12{,}1*\frac{7{,}4}{7{,}2}$
6,3	$3{,}9*\frac{7{,}4}{6{,}8}$	$4{,}5*\frac{7{,}4}{7{,}2}$
7,2	$5{,}8*\frac{7{,}4}{6{,}8}$	$6{,}3*\frac{7{,}4}{7{,}2}$
7,4	$6{,}8*\frac{7{,}4}{6{,}8}$	$7{,}2*\frac{7{,}4}{7{,}2}$

$\Rightarrow$

Array1	Array2	Array3
6,2	12,95	4,0
4,8	8,49	9,46
12,5	5,0	12,44
6,3	4,25	4,63
7,2	6,31	6,48
7,4	7,4	7,4

8.1. ábra. Skálázás alapú normalizáció. Bal oldalon: Az eredeti adatmátrix: 5 próba jelintenzitás-értéke 3 array-n. Az első array-t választjuk ki alapként. Középen: A második és a harmadik array-t átskáláztuk, hogy a jelintenzitások átlaga ugyanakkora legyen, mint az első array-n. Jobb oldalon: A normalizált adatmátrix

Kvantil-normalizáció. Először minden egyes array-n sorba rendezzük a jelintenzitás-értékeket. Majd, minden egyes sorszámra kiszámítjuk az átlagos jelintenzitást. Végül minden array-n minden próba normalizált értéke a sorszámának megfelelő átlagos érték lesz. Példaként lásd a 8.2. ábrát.

8.2.3. Összegzés

Mivel az array-n minden egyes génhez több próba is hozzá van rendelve, ezért ezeket a technikai replikátumokat (ún. próbahalmazokat, *probe set*) összegezni kell annak érdekében, hogy a génhez egyetlen expressziós értéket kapjunk. Ezt többféleképpen is megtehetjük, például a logaritmikusan transzformált expressziós értékek átlagolásával, az eredeti expressziós értékek átlagának logaritmikus transzformációjával, a logaritmikus skála medián-jával, a medián értékek logaritmusával, vagy kifinomultabb, modell-alapú módszerekkel [3].

	Array1	Array2	Array3
Próba1	6,2	11,9	3,9
Próba2	4,8	7,8	9,2
Próba3	12,5	4,6	12,1
Próba4	6,3	3,9	4,5
Próba5	7,2	5,8	6,3

⇒

Array1	Array2	Array3
12,5 (3)	11,9 (1)	12,1 (3)
7,2 (5)	7,8 (2)	9,2 (2)
6,3 (4)	5,8 (5)	6,3 (5)
6,2 (1)	4,6 (3)	4,5 (4)
4,8 (2)	3,9 (4)	3,9 (1)

⇒

Array1	Array2	Array3
12,17 (3)	12,17 (1)	12,17 (3)
8,07 (5)	8,07 (2)	8,07 (2)
6,13 (4)	6,13 (5)	6,13 (5)
5,1 (1)	5,1 (3)	5,1 (4)
4,2 (2)	4,2 (4)	4,2 (1)

⇒

PróbaID	Array1	Array2	Array3
Próba1	5,1	12,17	4,2
Próba2	4,2	8,07	8,07
Próba3	12,17	5,1	12,17
Próba4	6,13	4,2	5,1
Próba5	8,07	6,13	6,13

8.2. ábra. Kvantil-normalizáció. Balra fent: Az eredeti adatmátrix: 5 próba jelintenzitás-értéke 3 array-n. Jobbra fent: A jelintenzitás-értékeket minden egyes array-n egymástól függetlenül csökkenő sorrendbe rendezzük (miközben az eredeti próba-azonosítókat feljegyezzük – itt zárójelben látható). Balra lent: Minden egyes sorszámra (itt: sorra) kiszámítjuk az átlagos jelintenzitást. Jobbra lent: A normalizált adatmátrix

8.2.4. Szűrés

A normalizációs lépések után bevett gyakorlat, hogy a próbahalmazok egy részét a további adatelemzési lépések elvégzése előtt kiszűrjük. Ennek számos oka van: Először is az array-k feldolgozásának, ill. kezelésének számos olyan technikai aspektusa van, amelynek következtében zavaró hatások és potenciális variabilitás léphet fel, ami kilógó vagy megbízhatatlan expressziós értékekhez vezet. Másodszor általánosan elvárt, hogy a kísérlettől függően a gének egy nagy része várhatóan nem expresszálódik egyik kísérleti körülményben (állapotban) sem. A szűrések során megpróbáljuk azonosítani és kizárni a megbízhatatlan, nem változó expressziójú vagy nem expresszálódó próbahalmazokat annak érdekében, hogy pontosabb, megbízhatóbb eredményeket kapjunk a további statisztikai elemzések során [6].

Az előbbiek szemléltetésére leírjuk Kaminski és Friedman [2] szűrési javaslatait: Első lépésben meghatározzák az ún. „legális gének" halmazát; ezek azok a gének, amelyeknek expressziója legalább egy array-ben meghalad egy bizonyos előre meghatározott küszöbértéket. Ez utóbbit úgy határozzák meg, hogy ugyanazt a mintát két mikroarray-re is felviszik és összehasonlítják az expressziós szinteket. Mivel ezek konzisztenciája függ az értéküktől (a nagyobb intenzitásértékű tartományokban kisebb mértékben különböznek a két array-n mért értékek, mint a kisebb jelintenzitások esetén), gyakran megállapítható egy olyan küszöbérték, amely fölött az array-k konzisztenciája meggyőző. Ez a lépés általában harmadával vagy felével csökkenti a gének számát. Ezt követően meghatározzák az ún. „aktív gének" halmazát, amelybe azok a gének tartoznak, amelyek megváltoztak valamely kísérleti körülmények (állapotok) között. A gyakorlatban ez azt jelenti, hogy kiszűrik azokat a géneket, amelyek expressziója nem változott legalább másfélszeres mértékben a kísérletek legalább 5%-ában. Ez a lépés rendszerint jelentős mértékben lecsökkenti a gének számát a további elemzési lépések előtt.

8.3. Adatelemzés

8.3.1. Klaszterezés

A klaszterezés főleg „felderítő" jellegű célokat szolgál a mikroarray-k elemzése során. Ezek a módszerek sokkal inkább az intuíción, mintsem valamiféle formális elméleten alapulnak. Az alapötletük az, hogy meghatározzák gének vagy minták olyan csoportjait, amelyek valamilyen módon elkülönülnek egymástól, miközben a csoport elemei között belső kohézió, hasonlóság van. Ezek a klaszterek általában természetes módon is adódnak a kísérletünk tárgyából eredően. A különféle klaszterező módszerek száma zavarba ejtő; ebben a fejezetben röviden összefoglaljuk a leggyakrabban használtakat és a hátterükben rejlő elgondolásokat.

Minták klaszterezése

A mintáink klaszterezésének célja a kísérletünk típusától függ.

Az időbeli változásokat követő (*time-course*) kísérletekben egy organizmust különböző fejlődési állapotokban mintavételezünk. Ebben az esetben a mintáink klaszterezésével felderíthetjük ezeknek az állapotoknak a hasonlóságát vagy különbözőségét. Például ha asztmás személyeket vizsgálunk az asztmarohamok kialakulása előtt, alatt és után, akkor megbecsülhetjük, hogy mennyi időre van szükség ahhoz, hogy a sejtek visszanyerjék az eredeti állapotukat.

Összehasonlító vizsgálatokban különböző személyeket vizsgálunk eltérő kísérleti körülmények között annak érdekében, hogy a körülményeknek a gének expressziójára gyakorolt hatásait felderítsük. Ezekben a kísérletekben egy adott kísérleti körülményhez általában több személyből és egyénenként több technikai ismétléssel veszünk mintát. Ilyenkor a klaszterezés segíthet a minőségellenőrzésben, ugyanis ha egy minta nem ugyanabba a klaszterbe kerül, mint a technikai vagy biológiai replikátumai (míg a többi minta igen), akkor ez fényt deríthet az adott minta normalizációs vagy hibridizációs problémáira.

Klinikai kísérletekben hasonló fenotípusos jeggyel rendelkező (pl. mellrákos) egyéneket mintavételezünk azzal az *a priori* tudással, hogy az egyes személyek genetikailag különböznek egymástól. Ebben az esetben a minták klaszterezése nagyon fontos, ugyanis segíthet meghatározni az egyének különálló csoportjait, amelyek hasonló genotípussal (azaz jelen értelemben hasonló génexpressziós profillal) rendelkeznek.

A klaszterezés előtt két dolgot kell meghatároznunk: (1) Mit értünk az alatt, hogy a csoportok elemei között „belső kohézió" van? és (2) Mit értünk az alatt, hogy a különféle csoportok „elkülönülnek" egymástól?

A minták közötti távolság Először is, definiáljuk az adatpontjaink közötti távolság fogalmát. Ha a célunk a minták klaszterezése, akkor tekinthetjük ezeket úgy, mint olyan pontokat, amelyeket a génexpressziós értékek reprezentálnak a gének nagy-dimenziós terében. Ezek után a minták közötti távolságot definiálhatjuk geometriai távolságok (L_p

normák) segítségével:

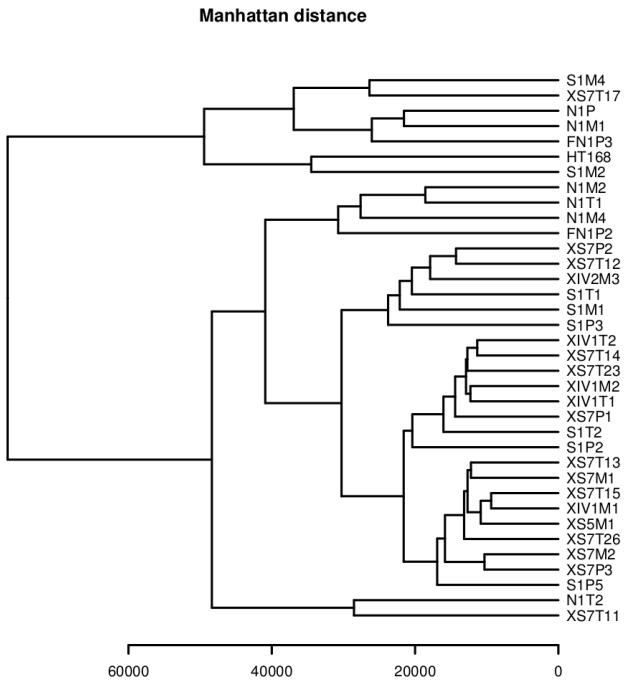
$$d_p(x,y) = \sqrt[p]{\sum_{i=1}^{n_{genes}} |x_i - y_i|^p}, \tag{8.1}$$

ahol x_i és y_i az i-edik gén expressziós szintjeit jelentik az x, illetve y mintában. Minél nagyobb a p értéke, annál érzékenyebb az L_p mérték a kilógó adatpontokra. A legrobosztusabb a Manhattan-távolság (d_1). Ez nem más, mint a két különböző array-n mért, azonos gének közötti távolságok abszolút értékének összege. Az euklideszi-távolság (d_2) érzékenyebb a kilógó értékekre, emiatt gyakrabban használják minőségellenőrzésre, amikor a cél a kilógó array-k azonosítása.

Klaszterek közötti távolság Ezután definiálnunk kell a megfigyeléseink csoportjai közötti távolságot. Mit jelent a „közeli", amikor nem egyedi adatpontokat, hanem adatpontok csoportjait hasonlítjuk össze? Ez attól függ, hogy az egy klaszterbe tartozó adatpontokat hogyan tömörítjük egyetlen, reprezentatív adatpontba. A leggyakrabban használt módszerek: az *átlagos távolság* (average linkage, a két csoport közötti távolság a páronkénti távolságok átlaga), *median távolság* (median linkage, a páronkénti távolságok mediánja), *centroid távolság* (centroid linkage, a két csoport – valamilyen értelemben – középpontjai közötti távolság), *egyszerű távolság* (single linkage, a páronkénti távolságok közül a legkisebb) and *teljes távolság* (complete linkage, a páronkénti távolságok közül a legnagyobb).

Agglomeratív hierarchikus klaszterezés A mikroarray-kísérletekben az egyik leggyakrabban használt klaszterezési algoritmus az agglomeratív hierarchikus klaszterezés. Számos előnye van, pl. a vizualizációja (a jól ismert *dendrogram*) könnyen értelmezhető, és számos olyan kapcsolatra deríthet fényt, amely egyébként rejtve maradna. Különösen hasznos azokban az esetekben, amikor a mintáknak eleve hierarchikus természetük van. Például rákos szövetek vizsgálatakor a különböző ráktípusok jól elkülönülő klaszterekbe tömörülnek. Ezekben számos – különböző genotipikus profilnak megfelelő – további alcsoportok lehetnek, és a legalsó szinten az egyének technikai replikátumai tömörülnek egy-egy klaszterbe. Az agglomeratív hierarchikus klaszterezés folyamata során első lépésben kiszámítjuk az összes minta közötti távolságot. Ezt követően a két legközelebbi adatpontot egy csoportba soroljuk, így kialakítva egy klasztert. Mindig, amikor egy új klasztert hozunk létre, kiszámítjuk a távolságát az összes többi klasztertől. Ezután megkeressük a két, egymáshoz legközelebb álló klasztert, és összevonjuk. Ez egy folytonosan összefésülő folyamatot eredményez, amelynek során egyelemű klasztereket vonunk össze, hogy nagyobb klasztereket kapjunk. Az így kialakuló hierarchiát egy dendrogrammal ábrázolhatjuk (lásd az 8.3. ábrát).

Főkomponens-elemzés A főkomponens-elemzés (Principal Component Analysis, PCA) egy jól ismert dimenziócsökkentő módszer, ami arra (is) használható, hogy egy nagy-dimenziós adatot kettő vagy három (vagy több) dimenzióban ábrázoljunk. A PCA olyan új, egymásra ortogonális tengelyeket hoz létre, amelyek az eredeti tengelyek lineáris kombinációi (azaz az adatunk eredeti dimenziói, amit a génexpressziós értékek reprezentálnak). Az első tengelyt (az első főkomponenst) úgy határozza meg az algoritmus, hogy



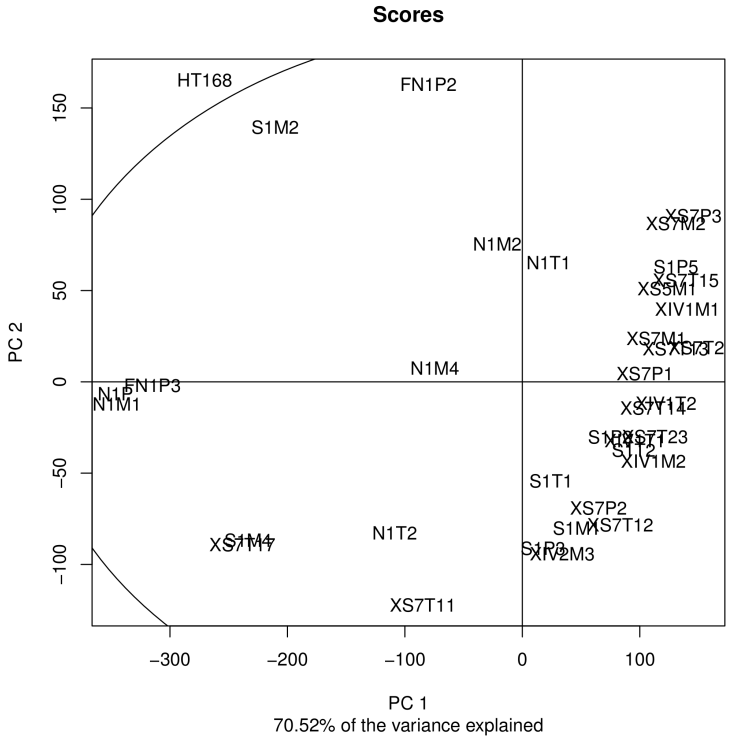
8.3. ábra. Példa egy agglomeratív hierarchikus klaszterezés eredményére

az adatunkban rejlő legnagyobb varianciájú komponenseket foglalja magába. A második komponenst úgy alakítja ki, hogy az első tengelyre ortogonális legyen, és a megmaradt variancia legnagyobb részét magyarázza meg. A harmadik tengely ortogonális lesz az első kettőre, és szintén a megmaradt variancia legnagyobb részét foglalja magába, és így tovább. Ha a gének között korreláció van, akkor az első pár tengely az adatban rejlő variancia legnagyobb részét képes lesz megmagyarázni; így ha a mintáinkat az első pár tengely alapján kirajzoljuk, akkor ez képes lesz feltárni a köztük lévő hasonlóságokat, illetve különbözőségeket (lásd a 8.4. ábrát).

Gének klaszterezése

A mintáink klaszterezése mellett érdekes lehet a hasonló expressziójú géncsoportok azonosítása (azaz a gének klaszterezése) is. Ennek a fő mozgatórugója az, hogy az együttes expresszió (co-expresszió) a gének közös szabályozására deríthet fényt (co-reguláció). Azaz az olyan gének, amelyek különböző körülmények között is hasonló módon viselkednek, valószínűleg közös jegyeket mutatnak, például közös szabályozási mechanizmusokkal rendelkeznek, vagy közös funkciókat látnak el. Tehát a gének esetén a hasonlósági és távolsági mértékek jellemzően mások, mint a minták esetén. A leggyakrabban használt távolsági metrika az együttes expresszión alapul:

$$d_\rho = 1 - \rho(x, y), \tag{8.2}$$



8.4. ábra. Példa egy főkomponens-elemzés eredményére

ahol $\rho(x, y)$ az ún. *Pearson korrelációs koefficiens*, amit a következő képlet ad meg:

$$\rho(x, y) = \frac{cov(x, y)}{\sigma_x \sigma_y}, \tag{8.3}$$

ahol $cov(x, y)$ a kovariancia és σ_x és σ_y az x, illetve y expressziós profilok standard eloszlása.

A korábban részletezett hierarchikus klaszterezésen kívül számos módszer létezik gének klaszterezésére, például a k-közép algoritmus [7], az önszerveződő térképek (self organising map, SOM) [8], vagy különböző gráfelméleti megközelítések [9]. Ezek közül a továbbiakban röviden bemutatjuk a k-közép klaszterezést.

k-közép klaszterezés A k-közép klaszterezés iteratív folyamatában első lépésben eldöntjük, hogy hány darab elkülönülő klasztert várunk. Ezután az algoritmus véletlen módon kiválaszt ennyi számú klaszterközéppontot, és minden gént a hozzá legközelebb álló klaszterhez rendel. Ezt követően az algoritmus módosítja minden klaszter középpontját úgy, hogy a klaszterbe tartozó pontok középponttól való távolságának összege minimális legyen. Ezután a módszer minden gént újra hozzárendel ahhoz a klaszterhez, amelynek középpontja hozzá legközelebb esik. Ezt az iteratív eljárást addig folytatjuk, amíg konvergenciát nem érünk el, azaz a középpontok és a klaszterbe sorolások nem lesznek állandóak. A módszer hátránya, hogy a klaszterek számát előre definiálni kell, illetve nem lehetséges az eredmények szemléletes megjelenítése [2].

8.3.2. Differenciális expresszió

A gének expressziós szintjének különböző körülmények hatására történő megváltozását az ún. differenciális expresszió számszerűsíti. Például ha egy gén transzkripciójának mértéke különbözik egészséges és beteg egyének között, akkor elképzelhető, hogy az adott gén szerepet játszik a betegség patomechanizmusában.

Klasszikus hipotézistesztelés

A differenciálisan expresszálódó gének meghatározására a leggyakrabban használt statisztikai technika a *klasszikus hipotézistesztelés* [1]. Ennek során minden egyes génre teszteljük azt a hipotézist, hogy az adott gén *nem* expresszálódik differenciálisan. Ez az ún. *nullhipotézis*, H_0. Hacsak nincs elegendő bizonyítékunk arra, hogy ez a hipotézis nem igaz, akkor nem tudjuk elvetni, azaz nem tudjuk elfogadni az ún. *alternatív hipotézist*, H_1-et, ami azt állítja, hogy az adott gén differenciálisan expresszálódik. Hipotézistesztelésnek nevezzük azt a módszert, amivel összegezzük az adatainkban található bizonyítékokat (az ún. *teszt-statisztika* kiszámításával) annak érdekében, hogy választani tudjunk a két hipotézis közül. A tesztstatisztika kiszámításának eredménye egy valószínűség (az ún. *p-érték*), ami a nullhipotézis abszurditásának mértékét jelzi. Más szóval, ha a p-érték közel van nullához, az azt jelzi, hogy a nullhipotézis nagyon valószínűtlen, abszurd, így el kell vetnünk, és helyette el kell fogadnunk az alternatív hipotézist. A hipotézistesztelés folyamatát összefoglalva a 8.5. ábrán láthatjuk.

Feltételezés	A nullhipotézis, H_0 igaz, azaz a g gén expressziója nem különbözik lényegesen a két állapot között
Ezután	Kiszámítjuk a tesztstatisztikát, z_g-t, és azt találjuk, hogy a p-érték (annak valószínűsége, hogy legalább z_g értéket figyelünk meg abban az esetben, ha a nullhipotézis H_0 igaz) nagyon közel van nullához
De	Éppen az előbb figyeltük meg z_g-t
Tehát	A nullhipotézis hamis, és az alternatív hipotézis (majdnem biztosan) igaz, azaz a g gén differenciálisan expresszálódik

8.5. ábra. A hipotézistesztelés menete a differenciális expresszió meghatározására

Két átlag közötti eltérés (pl. két különböző állapot során mért expressziós értékek átlagának eltérése) tesztelésére a legnépszerűbb statisztika az ún. *t-statisztika*. Ennek értéke egy g gén esetén valójában a két állapot közötti átlagos eltérés standardizáltja:

$$z_g = \frac{\bar{x}_g - \bar{y}_g}{\sqrt{\frac{s^2_{xg}}{n_x} + \frac{s^2_{yg}}{n_y}}}, \tag{8.4}$$

ahol $\bar{x}_g$ és $\bar{y}_g$ a g gén expressziós értékeinek átlaga az x, illetve y állapotokban; s^2_{xg} és s^2_{yg} a varianciák; és n_x és n_y a két állapotban megfigyelt minták száma.

A nullhipotézis mellett belátható [10], hogy a t-statisztika megközelítőleg követi a t-eloszlást, így a p-érték kiszámítható a z_g érték és a Student t-eloszlás összehasonlításából a megfelelő szabadsági fok mellett.

A standard t-teszt nagyon sok féle variációját vezették be és használják rendszeresen mikroarray-kísérletekben. Ezek vagy bootstrap-pet, permutációs vagy varianciapoolozásos megközelítéseket alkalmaznak, hogy az eredeti t-teszt erős megkötéseit enyhítsék. A leggyakrabban használt módszerek a *limma* [11] és a *Significance Analysis of Microarrays*, SAM [12].

Többszörös hipotézistesztelési probléma

A mikroarray-k statisztikai elemzésének egy súlyos problémával kell szembenéznie, ami akkor jelentkezik, ha egyszerre párhuzamosan több hipotézist is tesztelünk. Ez az ún. „többszörös hipotézistesztelési probléma" [1]. Nem számít, hogy milyen statisztikai módszert is használunk, minél nagyobb számú hipotézisünk van, annál nagyobb annak valószínűsége, hogy véletlenül extrém tesztstatisztika-értékeket figyelünk meg, így egyre valószínűbb, hogy tévesen el fogjuk utasítani a nullhipotézist (és ezzel hamis pozitív kijelentést teszünk, ún. elsőfajú hibát követünk el). Sokféle megközelítés létezik ennek a problémának a kezelésére, amik abban különböznek, hogy milyen hibát próbálnak meg kontrollálni és mennyire konzervatívak.

A legkonzervatívabbnak tartott módszer az ún. *Bonferroni eljárás*, amely a *családiszintű hibát* (familywise error rate, FWER) kontrollálja. Ez annak a valószínűsége, hogy az összes gén közül, amelyek nem differenciálisan expresszálódnak, legalább egyről tévesen azt állítjuk, hogy differenciálisan expresszálódik. A Bonferroni módszer során egyszerűen elosztjuk *alpha*-t (a megkívánt FWER szignifikanciaküszöböt) a hipotézisek számával. Például annak biztosítására, hogy 10 000 statisztikai teszt elvégzése esetén is a családiszintű hiba aránya kisebb legyen, mint 0,05, az elfogadási küszöböt 10^{-6}-ra kell állítanunk.

Mindazonáltal egy mikroarray-kísérlet inkább felderítő jellegű, mintsem megerősítő jellegű eszköz. Így a *hamis felfedezési hibaarány* (false discovery rate, FDR) kontrollálása talán bölcsebb döntés. Az FDR azoknak a géneknek a várható aránya, amelyek nem expresszálódnak differenciálisan azok közül, amelyekről azt állítjuk, hogy differenciálisan expresszálódnak. Más szóval, ha a célunk az, hogy előálljunk hipotézisek egy olyan halmazával, amelynek a legnagyobb része igaz, akkor az FDR-t érdemes kontroll alatt tartani. Benjamini és Hochberg javasolt [13] erre egy lefelé lépegető eljárást: a géneket sorrendezzük a p-értékük szerint, majd egy folyamatosan növekvő küszöbértékhez viszonyítsuk. Ez egy kevésbé konzervatív korrekciós eljárást eredményez, amit előszeretettel használnak mikroarray-kísérletek elemzése során.

8.3.3. Az eredmények biológiai értelmezése

A statisztikai analízis gyakran differenciálisan expresszálódó gének (hosszú) sorát eredményezi, amelyek egy része ismerős lesz a kísérletet végző kutató számára, más része viszont nem. Mindazonáltal nem feltétlenül egyszerű szemmel meghatározni a gének értelmes

biológiai kontextusát. Ebben az alfejezetben röviden bemutatjuk azokat a koncepciókat, amelyek segíthetnek megtölteni az eredményeket biológiai értelemmel.

Gene Ontology elemzés

Egy alapvető kérdés lehet, hogy „Mit csinálnak az alul-, illetve felülexpresszálódó gének a sejtben?" vagy „Milyen biológiai folyamatokban vesznek részt?". Ezeknek a kérdéseknek a megválaszolásában a Gene Ontology adatbázis jöhet a segítségünkre. A Gene Ontology (GO) [14] egy standardizált és strukturált szótár (ontológia) biológiai kifejezések: molekuláris funkciók, biológiai folyamatok és sejtes komponensek leírására; és a közöttük lévő kapcsolatok definiálására [15]. Emellett minden génhez hozzá vannak rendelve azok a kifejezések, amelyek a legjobban leírják annak funkcionalitását. Így ha a korábbi statisztikai elemzések előálltak (két állapot között) alul- vagy felülexpresszálódó gének listájával, akkor az ún. hipergeometrikus tesztet használhatjuk annak eldöntésére, hogy mely Gene Ontology kifejezések vannak alul- vagy felülreprezentálva bennük.

Tekintsük azt az esetet, hogy ki akarjuk számítani annak valószínűségét, hogy egy adott biológiai folyamat felülreprezentált egy számunkra érdekes génlistában. Képzeljünk el egy urnát, amelyben minden egyes génnek egy golyó felel meg (a mikroarray-n lévő N darab gén), és képzeljük el, hogy azok a golyók, amelyek az adott biológiai funkciót ellátó géneknek felelnek meg, fehérek (K darab gén), míg a többi golyó, amelyeknek megfelelő gének nem asszociáltak az adott funkcióval, feketék (N-K darab gén). Ezután húzunk n darab golyót az urnából; méghozzá azokat, amelyek a számunkra érdekes géneknek felelnek meg (pl. felülexpresszálódnak egy adott állapotban egy másikhoz képest). Ezek közül azt látjuk, hogy k darab golyó fehér; ezek azoknak a géneknek felelnek meg, amelyek érdekesek is, és asszociáltak is a kérdéses biológiai funkcióval. Ezek után annak a valószínűségét, hogy pontosan k darab ilyen golyót húztunk, a hipergeometrikus eloszlás adja meg:

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}. \tag{8.5}$$

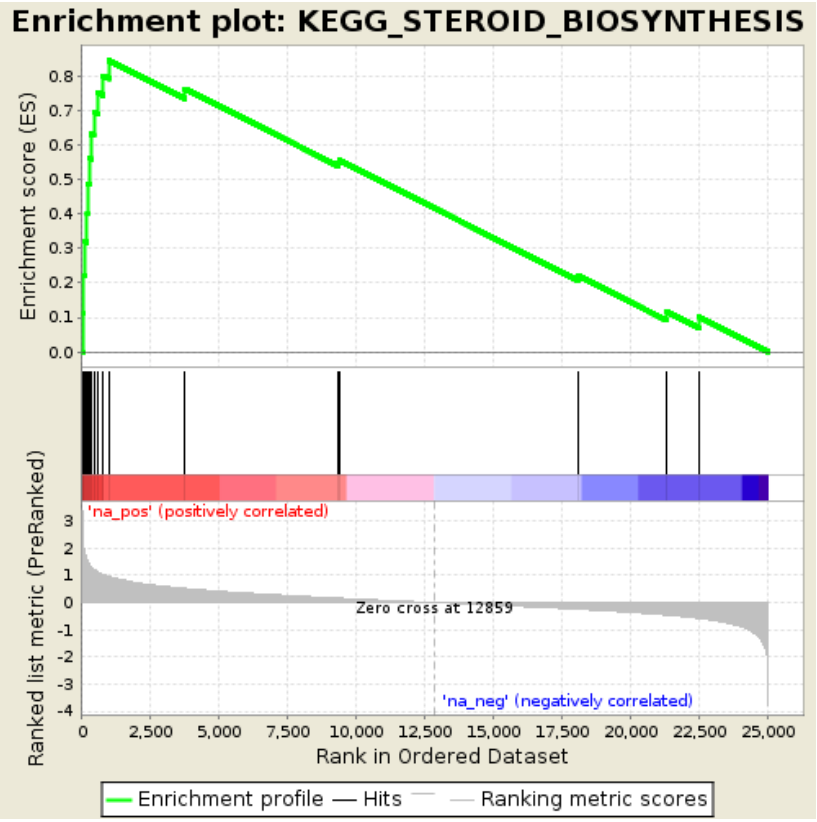
Ebből eredően, azon feltételezés mellett, hogy nincs asszociáció a biológiai funkció és az érdekes génlista között, az adott funkcióval bíró érdekes gének számának a hipergeometrikus eloszlást kell követnie. A megfigyelt érték alapján kiszámítható a nullhipotézis abszurditását jelző p-érték, és a nullhipotézist elvethetjük, ha ez a p-érték közel van nullához. Ha egyszerre több tesztet is végrehajtunk, akkor szükséges valamilyen korrekció is a többszörös hipotézistesztelési probléma kezelésére a korábban ismertetett módok valamelyikén. Ez az elemzés több szoftverben is készen elérhető, pl. a Cytoscape [17] szoftver BiNGO [16] beépülő moduljában.

Génhalmazok feldúsulásának elemzése

A génhalmazok feldúsulásának elemzése (Gene Set Enrichment Analysis, GSEA) [18] fontos kiegészítő módszer, ha génlistákat szeretnénk megtölteni biológiai értelemmel. Ennek segítségével azt határozhatjuk meg, hogy egy előre definiált génhalmaz (pl. egy adott

biológiai funkciót ellátó gének halmaza) mennyire mutat statisztikailag szignifikáns, konkordáns különbségeket két állapot között [19]. A legfontosabb különbség a fent ismertetett hipergeometrikus teszt és a GSEA között az, hogy az utóbbi nem kívánja a gének érdekes és érdektelen csoportokba sorolását. Ehelyett a gének egy teljes sorrendjét használja, ahol a géneket valamilyen folytonos értékű pontszám (pl. a t-statisztika értéke) alapján sorrendezzük. Ez alapján kiszámít egy ún. feldúsulási pontszámot (enrichment score, ES), ami arról nyújt információt, hogy egy előre definiált génlista milyen mértékben van felülreprezentálva a sorrend elején vagy végén. Ha a feldúsulási pontszám pozitív, akkor a génlista a sorrend elején csoportosul (lásd a 8.6. ábrát); ha pedig negatív, akkor a sorrend végén.

A GSEA alapvető elgondolása az, hogy például egy adott metabolikus útvonalba eső gének expressziójának 20%-os megnövekedése drámai módon fogja befolyásolni az adott útvonalon átmenő fluxust, és ez valószínűleg sokkal fontosabb, mint egyetlen gén expressziójának 20-szoros megnövekedése [18].



8.6. ábra. Példa egy génhalmaz-feldúsulási elemzés eredményére

A GSEA módszer szabadon elérhető egy szoftvercsomagban [19] a MSigDB nevű, több mint 8500 előre definiált génhalmazt tartozó adatbázissal együtt (a v3.1-es verzió szerint).

Irodalomjegyzék

[1] Ernst Wit and John McClure, *Statistics for Microarrays: Design, Analysis and Inference*. Wiley, 1st ed., July 2004.

[2] Naftali Kaminski and Nir Friedman, Practical approaches to analyzing results of microarray experiments. *American journal of respiratory cell and molecular biology*, 27(2):125–132, August 2002. PMID:12151303.

[3] *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. `http://www.springer.com/computer/bioinformatics/book/978-0-387-25146-2`

[4] Rafael A. Irizarry, Bridget Hobbs, Francois Collin, Yasmin D. Beazer-Barclay, Kristen J. Antonellis, Uwe Scherf, and Terence P. Speed, Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England)*, 4(2):249–264, April 2003. PMID: 12925520.

[5] Affymetrix Web Site. `http://www.affymetrix.com`

[6] S. B. Pounds, C. Cheng, and A. Onar, Statistical Inference for Microarray Studies. In: D. J. Balding, M. Bishop, and C. Cannings, editors, *Handbook of Statistical Genetics*, pages 231–266. John Wiley & Sons, Ltd, 2008.

[7] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, and V. Sondak, Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406(6795):536–540, August 2000. PMID: 10952317.

[8] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub, Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences of the United States of America*, 96(6):2907–2912, March 1999.

[9] R. Sharan and R. Shamir, CLICK: a clustering algorithm with applications to gene expression analysis. Proceedings / ... International Conference on Intelligent Systems for Molecular Biology; ISMB. International Conference on Intelligent Systems for Molecular Biology, 8:307–316, 2000. PMID: 10977092.

[10] F. E. Satterthwaite, An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2(6):110–114, December 1946.

[11] Gordon K. Smyth, Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, vol. 3, issue 1, 2004. PMID: 16646809.

[12] V. G. Tusher, R. Tibshirani, and G. Chu, Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9):5116–5121, April 2001. PMID: 11309499.

[13] Yoav Benjamini and Yosef Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, January 1995.

[14] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1)25–29, May 2000. PMID: 10802651.

[15] Louis du Plessis, Nives Skunca, and Christophe Dessimoz, The what, where, how and why of gene ontology–a primer for bioinformaticians. *Briefings in bioinformatics*, 12(6):723–735. November 2011. PMID: 21330331.

[16] Steven Maere, Karel Heymans, and Martin Kuiper, BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics (Oxford, England)*, 21(16):3448–3449, August 2005. PMID: 15972284.

[17] Michael E. Smoot, Keiichiro Ono, Johannes Ruscheinski, Peng-Liang Wang, and Trey Ideker, Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics (Oxford, England)*, 27(3):431–432, February 2011. PMID: 21149340.

[18] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric. S. Lander, and Jill P. Mesirov, Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, October 2005.

[19] GSEA. `http://www.broadinstitute.org/gsea/index.jsp`

9. fejezet

Biomarker-elemzés

Elsőként összefoglaljuk a biomarker-kutatás legfőbb kihívásait. Majd ismertetjük a feltételes valószínűségi megközelítésből származó relevancia-fogalmakat és az ezekhez kapcsolódó strukturális tulajdonságait a Bayes-hálóknak. Ismertetjük az ilyen strukturális jegyeken alapuló, utófeldolgozásában skálázható Bayes-háló alapú relevancia-elemzést.

Jelölések*

x,$\underline{x}$,$\underline{\underline{x}}$	skalár, (oszlop)vektor vagy halmaz, mátrix
$X, x, p(X)$	véletlen változó X, érték x, valószínűségi tömegfüggvény/sűrűségfüggvény X
$\mathrm{E}_{X,p(X)}[f(X)]$	$f(X)$ várható értéke $p(X)$ szerint
$\mathrm{var}_{p(X)}[f(X)]$	$f(X)$ varianciája $p(X)$ szerint
$I_p(\underline{X}\|\underline{Z}\|\underline{Y})$	$\underline{X}$ és $\underline{Y}$ megfigyelési függetlensége $\underline{Z}$ feltétellel p esetében
$(X \perp\!\!\!\perp Y \| Z)_p$	$I_p(\underline{X}\|\underline{Z}\|\underline{Y})$
$(X \not\perp\!\!\!\perp Y \| Z)_p)$	$\neg I_p(\underline{X}\|\underline{Z}\|\underline{Y})$
$CI_p(\underline{X};\underline{Y}\|\underline{Z})$	$\underline{X}$ és $\underline{Y}$ beavatkozási függetlensége $\underline{Z}$ feltétellel p esetében
$\prec$	(részleges) sorrendezés
$\prec^c$	a változók egy teljes sorrendezése
$\prec^G$	adott G irányított körmentes gráffal kompatibilis sorrendek halmaza
$\prec(n)$	n objektum sorrendjeinek (permutációinak) a halmaza
$G, \underline{\theta}$	Bayes-háló struktúrája és paraméterei
$G^{\sim}$	G irányított körmentes gráf esszenciális gráfja
$\mathcal{G}(n)/\mathcal{G}^k(n)$	n csomópontú maximum k szülőjű DAG-ok halmaza
$\mathcal{G}^{\prec}$	adott $\prec$ sorrenddel kompatibilis DAG-ok halmaza
$\mathcal{G}^G$	adott G DAG-gal megfigyelési ekvivalens DAG-ok halmaza
$\sim$	kompatibilitási reláció
$\mathrm{pa}(X_i, G) \sim \prec$	$\mathrm{pa}(X_i, G)$ szülői halmaz kompatibilis $\prec$ sorrendezéssel
$\mathrm{MB}_p(X_i)$	Markov-takarója X_i-nek p-ben

*További konvenciók az egyes fejezetekben jelöltek.

$\mathrm{pa}, \mathrm{pa}(X_i, G)$	szülői változók halmaza, X_i szüleinek halmaza G-ben
pa_{ij}	a j. konfigurációja a szülői értékeknek egy sorrendben
$\mathrm{bd}(X_i, G)$	X_i szüleinek, gyerekeinek és gyerekei egyéb szüleinek halmaza G-ben
$\mathrm{MBG}(X_i, G)$	a Markov-takaró algráfja X_i-nek G-ben
$\mathrm{MBM}(X_i, X_j, G)$	a Markov-takaróbeliség relációja
n	valószínűségi változók száma
k	maximális szülőszám DAG-okban
N	mintaszám
V	összes valószínűségi változók száma
Y	válasz, kimeneteli, függő változó
$N_+/N_{\ldots,+,\ldots}$	$N_i/N_{\ldots,i,\ldots}$ megfelelő összegei
$D\|X$	X változóhalmazra szűkített adathalmaz
$\|\|$	kardinalitás
$1()$	indikátorfüggvény
f', f''	f függvény első és második deriváltjai
A^T	A mátrix transzponáltja
$\underline{x} \cdot \underline{y}$	$\underline{x}$ és $\underline{y}$ vektorok skalárszorzata
ξ^+/ξ^-	informatív/nem informatív információs kontextus
$\neg, \wedge, \vee, \neq, \rightarrow$	standard logikai operátorok
$\cap, \cup, \setminus, \Delta$	standard halmazműveletek
$KB \vdash_i \alpha$	α bizonyíthatósága KB-ből
Γ	a Gamma függvény
$\mathrm{Beta}(x\|\alpha, \beta)$	a Béta eloszlás sűrűségfüggvénye (pdf)
$\mathrm{Dir}(x\|\underline{\alpha})$	a Dirichlet-eloszlás sűrűségfüggvénye
$\mathrm{N}(x\|\mu, \sigma)$	az egyváltozós normál eloszlás sűrűségfüggvénye
$\mathrm{N}(x\|\underline{\mu}, \underline{\underline{\Sigma}})$	a többváltozós normál eloszlás sűrűségfüggvénye
BD,BD_e	Bayesian Dirichlet-prior, megfigyelési ekvivalens BD-prior
BD_{CH}	Bayesian Dirichlet (BD) prior 1 hiperparaméterekkel
$\mathrm{BD}_{\mathrm{eu}}$	megfigyelési ekvivalens és uniform BD prior
$L(\underline{\theta}; D_N)$	$p(D_N\|\underline{\theta})$ likelihood függvénye
$H(X, Y)$	X és Y entrópiája
$I(X; Y)$	X és Y kölcsönös információja
$\mathrm{KL}(X\|\|Y)$	X és Y Kullback–Leibler-divergenciája
$H(X\|\|Y)$	X és Y keresztentrópiája
$L_1(,)$, $L_2(,)$	az abszolútértékbeli (Manhattan) négyzetes (euklidészi) távolságok
$L_0(,)$	0-1 veszteség
$\mathcal{O}()/\Theta()$	aszimptotikus, nagyságrendi felső és alsó határ

Rövidítések

ROC	Receiver Operating Characteristic (ROC) görbe
AUC	ROC-görbe alatti terület

BMA	Bayes-i modell átlagolás
BN	Bayes-háló
DAG	irányított körmentes gráf
FSS	jegykiválasztási probléma
MAP	maximum a posteriori
MI	kölcsönös információ
ML	maximum likelihood
MBG	Markov-határ gráf
MB	Markov-takaró
MBM	Markov-takaróbeliség
(MC)MC	(Markov-láncos) Monte Carlo
NBN	naiv Bayes-háló

9.1. Bevezető

Az élettani tudományok terén a közelmúltban végbement technikai fejlődés lehetővé tette a genomok szekvenálását, és a nagy áteresztőképességű genomikai, proteomikai, metabolikai technikák újradefiniálták a biológiát és az orvostudományt, továbbá megnyitották a genomikai és poszt-genomikai korszakot. E korszak nagy ígéretei a személyre szabott megelőzés, diagnózis, hatóanyagok és kezelés. A klinikum nézőpontjából azonban ezek az„átmeneti" ígéretek még mindig beváltásukra várnak, és folyamatosan mind későbbi időpontokra tolódtak. Adatelemzési nézőpontból sem magyarázó jellegű, diagnosztikai biomarkerek, sem új oki célpontok és új hatóanyagok, sem objektív klinikai végpontok felfedezése nem váltotta be a várakozásokat, amint azt olyan hírhedt problémák és cikkek példázzák, mint a „missing heritability", „missing the mark" és a „production gap" a gyógyszerészetben.

Az utóbbi két évtizedben egyre gyorsuló ütemben felhalmozódó rendkívül sokrétű, heterogén és nagy mennyiségű orvosbiológiai adatra és tudásra gondolva valóban paradoxonnak tűnik a gyógyszerkutatások egyre romló költséghatékonysága, vagy akár a személyre szabott medicina reméltnél lassabb fejlődése. A reméltől elmaradó teljesítménye az oki, diagnosztikai, leírói biomarkereknek azért is meglepő, mert a hatóanyagokhoz, génekhez és betegségekhez tartozó felhalmozódó információforrások gazdagsága megdöbbentő: ez tartalmaz olyan gyógyszerészeti információkat, mint a hatóanyag taxonómiák, kémiai ujjlenyomatok, célfehérjék, hatóanyagok és betegségek génexpressziós profiljai, mellékhatások, indikációk, off-label gyógyszeralkalmazás. Továbbá növekszik a mennyisége a betegségek molekuláris biológiai hátteréről rendelkezésre álló információknak, úgymint útvonalinformációk, génregulációs mechanizmusok, fehérje-fehérje hálózatok, gén-betegség hálózatok és a genetikai, epigenetikai variációk hatásai. Megoldást az új, egyre részletesebb és kiterjedtebb molekuláris biológiai adatok mellett legalább annyira az egyre hatékonyabb, tudásgazdag informatikai és statisztikai elemzésektől is várnak a szakértők, különösen a betegségek genetikai hátterének felderítése kapcsán.

Paradox módon azonban a potenciális biomarkerek nagy száma is statisztikai kihívást jelent, illetve az információkészlet sokfélesége is komoly kihívást támaszt az integrált elem-

zés, fúzió szempontjából. Ezek következtében a biomarker-felfedezés több szempontból is tekinthető a transzlációs kutatások egyik kritikus szűk keresztmetszetének. Új biomarker-elemzési módszerek ennek megfelelően a nagy mennyiségű háttértudás befogadását, rendszerszemléletű integrációt, értelmezhetőséget és döntéselméleti felhasználást próbáltak biztosítani. A fejezetben összefoglaljuk a Bayes-hálók felhasználását a biomarkerek következő négy tulajdonságának jellemzésére:

1. *Közvetlenség.*
2. *Oksági szerep.*
3. *Hatáserősség.*
4. *Interakciók.*

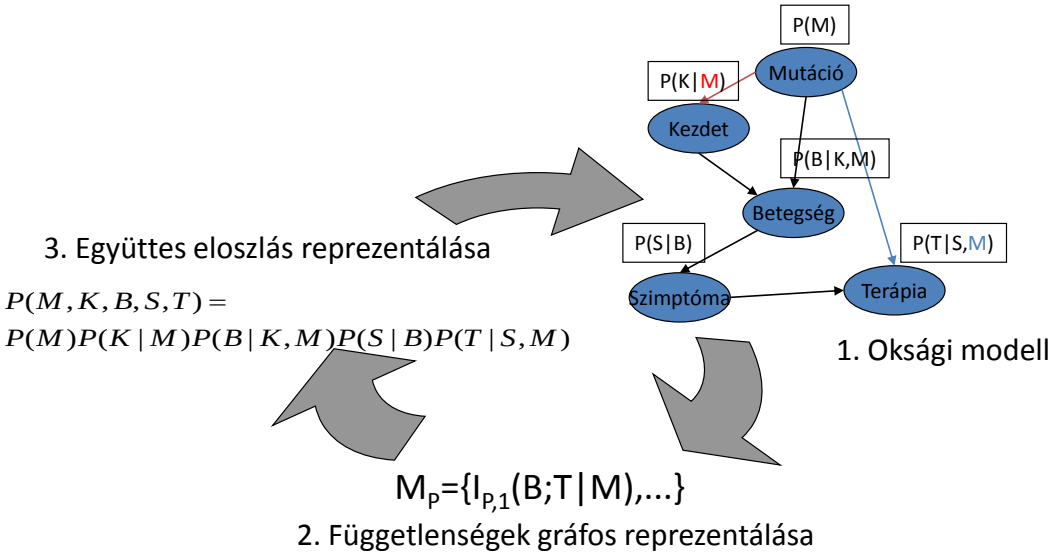
Az adatok és a tudás integrált elemzésére több keretrendszerben is folynak kutatások, mint például logikai, valószínűségi logikai vonalon az adatok relációs voltára tekintettel lévő módszerek, vagy (leíró) hálózati vonalon a nagyléptékű hálózatok szabályszerűségeit vizsgáló módszerek. A fejezetben tárgyalt rendszeralapú megközelítés a Bayes-statisztikai keretrendszerben az úgynevezett Bayes-háló alapú Bayes-i többszintű relevancia elemzés (Bayesian network-based Bayesian Multilevel Analysis of relevance, BN-BMLA). Ez komplex modellek felett átlagolva származtat a változók erős relevanciájára és azok egyre magasabb szintű interakcióihoz a posteriori valószínűségeket. A rendszeralapú megközelítést és a Bayes-statisztikai keretrendszert integráló módszertanok népszerűségét az magyarázza, hogy egyrészt gyakran a komplex modell identifikációhoz nincs elég adat, de a Bayes-i megközelítés lehetővé teszi érdekes modelltulajdonságok kikövetkeztetését is, másrészt maga a rendszeralapú megközelítés biztosítja, hogy a priori ismeretek elérhetőek legyenek az induktív következtetésbe való integráláshoz. A rendszeralapú megközelítésben a Bayes-hálózatok használata azért indokolt, mivel unikális, háromféle értelmezést is lehetővé tevő modellezést kínálnak, nevezetesen egy tárgyterület valószínűségi eloszlásának hatékony algebrai reprezentálását, a feltételes függetlenségek átfogó rendszerének reprezentálását és az oksági modell leírását, lásd a 9.1. ábra.

A Bayes-háló-modellosztály további előnye, hogy adott esetekben a modellparaméterek feletti átlagolás analitikusan kezelhető, amely analitikus kezelést részben lehetséges a modellstruktúrák feletti átlagolásra is kiterjeszteni többváltozós relevancia-elemzések esetében is. A fennmaradó mintavételi eljárásokat pedig Monte Carlo-módszerek párhuzamosításával tehetők hatékonnyá, kihasználva a számítástechnikai ilyen irányú fejlődését.

A jegyzet valószínűségszámítási és valószínűségi gráfos modellekkel kapcsolatos háttere a Valószínűségi döntéstámogatás jegyzetben tárgyalt.

9.2. Elméleti háttér

Az orvosbiológiai kutatások egyik alapkérdése egy vagy több kimeneteli változó esetén azon változók beazonosítása, amelyek prediktív (diagnosztikai) vagy beavatkozási (terápiás) lehetőségeket kínálnak.



9.1. ábra. Bayes-hálók reprezentációjának három aspektusa

Többváltozós megközelítésben mind a diagnosztikai, mind az oksági aspektus optimalitása többféleképpen is formalizálható. Diagnosztikai aspektusban nyilvánvaló követelmények a prediktív erő, bináris esetben az érzékenység, specificitás, pozitív és negatív prediktív érték, de fontos követelmény a redundanciamentesség is, amit mind a prediktorok minimális száma, de a prediktorok egymáshoz viszonyított egyedisége is jelezhet. Oksági aspektusban szintén nyilvánvaló követelmény a hatáserősség, illetve itt is a rendszerszintű egyedisége a változóknak. Mindkét esetben közös szempont lehet az elérhetőség és a költség aspektusa. A formalizálás kidolgozásához tekintsük a következő fogalmakat.

A feltételes vagy prediktív megközelítésben, amikor egy vagy több kimeneti változót befolyásoló bemeneti (vagy prediktor) változót keresünk, a jegyrészhalmaz kiválasztási probléma (Feature Subset Selection, FSS) és a relevancia fogalma definiálható a modellosztály és a predikcióban használt veszteségfüggvény felhasználásával, sőt akár a rendelkezésre álló mintaméret és az optimalizáció is ebbe belefoglalható (ezen „csomagoló" megközelítés leírását lásd [9]). Az axiomatikusabb „szűrő" megközelítésben az FSS fogalmai és módszerei a következő valószínűségi, együttes eloszlásra támaszkodó definícióra támaszkodnak [14].

9.1. Definíció. Egy változóhalmazt, $MB_P(X_i)$-t X_i *Markov-takaró*jának nevezünk $P(X_1, \ldots, X_n)$ eloszlásban, ha $(X_i \perp\!\!\!\perp V \setminus MB(X_i) | MB(X_i))_P$ (egyértelműség esetén P nem jelölt). A minimális Markov-takarót Markov-határnak nevezzük és $MBo(X_i)_P$ jelöli.

Ha a Markov-takaró egyértelműen létezik, akkor bevezethető egy szimmetrikus páronkénti reláció a Markov-takaróbeliségre: $MBM(X_i, X_j)_P$ fennáll X_i és X_j között P-ben,

ha

$$MBM(X_i, X_j)_P \leftrightarrow X_j \in MBo(X_i)_P \tag{9.1}$$

A Markov-határbeliségen belül definiálható egy szigorúbb kategória is, amelyet *közvetlen függésnek* nevezünk, ha minden diszjunkt $Z \subseteq V$ halmazra $(X \not\perp\!\!\!\perp Y|Z)$ fennáll (ebben az esetben a függés két változó között is létezik, amikor $Z = $, ami nem feltétlenül igaz a Markov-határbeli változópároknál).

A feltételes valószínűségi analógja, amely modellosztálytól, veszteségfüggvénytől, adathalmaztól, optimalizációtól független, a következő:

9.2. Definíció. Egy X_i bemeneti (prediktor) változó vagy jegy erősen releváns Y-ra, ha létezik egy olyan $X_i = x_i$, $Y = y$ és

$$s_i = x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n, \quad S_i = \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\},$$

hogy $p(x_i, s_i) > 0$ és $p(y|x_i, s_i) \neq p(y|s_i)$. Az X_i jegy gyengén releváns, ha nem erősen releváns, és van egy olyan S_i' részhalmaza az S_i jegyeknek, amelyekre létezik egy olyan x_i, y és s_i', hogy $p(x_i, s_i') > 0$ és $p(y|x_i, s_i') \neq p(y|s_i')$. Egy jegy releváns, ha gyengén vagy erősen releváns; amúgy irreleváns [9].

A Bayes-hálók sokoldalúsága rengeteg lehetőséget kínál a relevancia reprezentálására [14]. A következő tétel egy elégséges feltételt ad a releváns jegyek Bayes-hálós reprezentálására.

9.1. Tétel. *Egy (G, θ) Bayes-háló által definiált p eloszlás esetében a* $\mathrm{bd}(Y, G)$ *változók Y Markov-takarója, ahol* $\mathrm{bd}(Y, G)$ *Y szüleinek, gyerekeinek és gyerekei egyéb szüleinek halmaza [14]. Ha a p eloszlás stabil és G perfekt térképe, akkor* $\mathrm{bd}(Y, G)$ *az egyértelmű és minimális Markov-takarója Y-nak (*$\mathrm{MBS}_p(Y)$*), továbbá, $X_i \in \mathrm{MBS}_p(Y)$ ha X_i erősen releváns [16].*

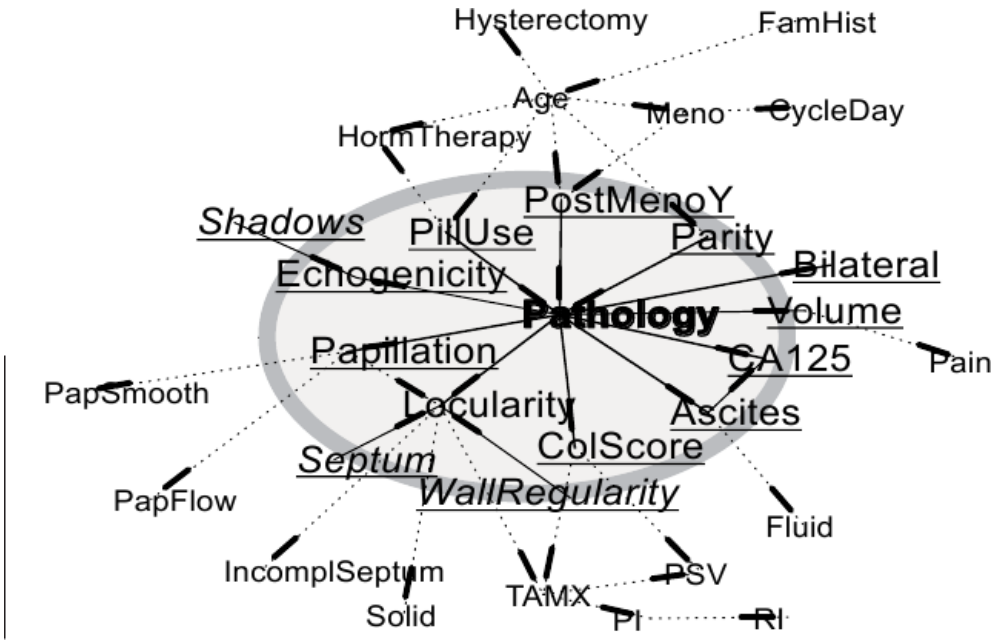
A továbbiakban $\mathrm{bd}(Y, G)$-re mint Y G-beli Markov-takarójára hivatkozunk $\mathrm{MBS}(Y, G)$ jelöléssel, azzal az implicit feltevéssel, hogy p Markov-kompatibilis G-vel[†]. Hasonlóan, a származtatott (szimmetrikus) páronkénti relációt is

$$\mathrm{MBM}(Y, X_j, G) \Leftrightarrow X_j \in \mathrm{bd}(Y, G) \tag{9.2}$$

Markov-takaróbeliségnek hívjuk.

A Markov-takaró jelentőségét az adja, hogy egy olyan minimális változóhalmazt azonosít, amely szükséges és elégséges egy változóhalmaz esetén. A 9.2. ábra egy valós orvosi diagnosztikai modell Markov-takaróját mutatja.

[†]Egy általános Bayes-i formalizációban (például Dirichlet-eloszlások alkalmazásával $p(\theta|G)$ paraméter prioroknál), a $\mathrm{bd}(Y, G)$ szomszédok 1 valószínűséggel alkotnak Markov-határt [13].



9.2. ábra. Egy preoperatív petefészekrák diagnosztikai modell [3]. A Pathology célváltozót félkövér kiemelés jelzi, Markov-takaróját szürke keret.

9.3. Bayes-i többszintű relevancia-elemzés

Korábbi relevancia-elemzési módszerek, amelyek Bayes-hálókat használtak: a Markov-takaró Közelítő Algoritmus [11], a kiterjesztései [18], illetve az IAMB algoritmus és variánsai [2, 16, 17]. Az optimalizációs alapú, maximum likelihood vagy maximum a posteriori (MAP) identifikációs módszerek sztochasztikus és Bayes-i kiterjesztései is megjelentek (egy randomizált módszert lásd [15]). A számításigényesebb, Bayes-i megközelítésben az adott Y célváltozóra vonatkozó relevanciák különböző reprezentációinak az a posteriori valószínűségi eloszlását szeretnénk megismerni. Korábbi munkákban a cél a tárgyterület átfogó jellemzése volt MBM poszteriorokkal [8, 10, 12].

Az FSS problémát könnyedén ki lehet terjeszteni, hogy tartalmazza a releváns változók interakciós struktúráját is, nevezetesen a Markov-takaró gráf mint strukturális modelltulajdonság vezethető be (osztályozási algráfként is gyakran hivatkozott [1, 2].

9.3. Definíció (Markov-takaró gráf)**.** A G Bayes-háló-struktúra Markov-takaró részgráfja vagy határoló mechanizmusok modellje $\mathrm{MBG}(Y, G)$ az Y változóra, ha tartalmazza a $\mathrm{bd}(Y, G)$ Markov-takarót és az Y-ba és gyerekeibe befutó éleket.

Az MBG-knek létezik valószínűségi és kauzális értelmezése. Minderről, valamint a megfigyelés ekvivalens MBG-kről, a számosságukra adható korlátról és a predikcióban való használatukról bővebb információ a [1, 2] irodalmakban található. Az MBG-k egy fontos tulajdonsága, hogy teljes adathalmaz esetén az MBG ismerete elégséges feltétel a releváns változók meghatározásához. Sajnos az MBG poszterior számítása exponenciális

komplexitású, azonban egy változósorrendre alapozott sorrend feltételes poszterior polinom időben számítható, ami kihasználható sorrendi MCMC-módszerekkel [2]. Az MBM és az MBS (vagy MBG) elemek a Bayes-hálós jegyeken alapuló modellezés két különböző megközelítése. Az előbbi esetében a jegyek és lehetséges értékeik száma könnyen kezelhető (a változók függvényében lineáris vagy kvadratikus). Ekkor az egyes MBM-jegyek a teljes modell egy kis részét reprezentálják, és ezek integrálásával jutunk a teljes modellt leíró képhez. Ilyen jegyek a páronkénti élek, a kényszerített élek és a Markov-takaróbeliség (MBM). Egy lehetséges másik megközelítésben egy komplex jegy szolgál átfogó képpel a teljes modellről. Ilyen jegyek lehetnek statisztikailag szignifikáns algráfok, mint például Markov-takaró gráfok (MBG-k). A többszintű Bayes-i relevanciaanalízis annyival nyújt többet, hogy mindkét megközelítést magába foglalja, ezáltal még teljesebb képet ad a teljes modellről. Lehetővé teszi továbbá az egyes egyszerű jegyek (MBM), jegyek halmazai (MBS), illetve a jegyalgráfok (MBG) a posteriori valószínűségeinek számítását és összekapcsolását. További szintek is lehetségesek tárgyterület-specifikus tudás felhasználásával, mellyel a változók típus szerinti csoportosítása válik lehetővé. Továbbá lehetséges az MBG-k által kifeszített térnél „szűkebb" CRPDAG-ok által kifeszített teret használni. A skálázhatóság megértéséhez vegyük észre, hogy az MBM, MBS, és MBG jegyek egyre növekvő komplexitású szinteket definiálnak ($|MBM| \ll |\,\mathrm{MBS}\,| < |\mathrm{MBG}| < |BN|$).

9.4. Többváltozós skálázhatóság: a k-MBS jegy

A többszintű Bayes-i relevanciaanalízis- (BMLA-) módszer a különböző absztrakciós szintek alkalmazásával széleskörű elemzést tesz lehetővé. Az MBS és az MBG jegyek sokkal kifejezőbbek az MBM jegyeknél, ám kardinalitásuk exponenciális, illetve szuperexponenciális, míg az MBM esetén ez lineáris a változók számának függvényében. Ennek megfelelően előfordulhat, hogy az MBS és az MBG a posteriori valószínűségek „laposak", mikor MBM poszteriorok már rég„csúcsosak" (azaz 0-hoz vagy 1-hez vannak közel). A „lapos" poszterior azt jelenti, hogy számos, akár száz jegy rendelkezik közepesen magas valószínűséggel, és nincs igazán közöttük legjobb. A „csúcsos" poszterior ezzel szemben azt jelenti, hogy a jegyek sokasága közül van egy-két olyan, amelyik markánsan nagyobb valószínűséggel rendelkezik a többinél. Tipikusan - még lapos poszteriorok esetén is a legvalószínűbb MBS és MBG jegyek rendelkeznek közös részekkel. Ennek kezelésére vezethetőek be a k-MBS és k-MBG jegyek, melyek a „k" paraméter segítségével skálázható komplexitásúak.

9.4. Definíció (k-MBS)**.** Egy $p(\underline{V})$ eloszlás esetén ($|\underline{V}| = n$), ha minden $X_i \in \underline{s}$ változó, ahol $\underline{s} \subseteq \underline{V}$, Markov-határbeliek mbs és $|\underline{s}| = k$, akkor $\underline{s}$ egy k-s Markov-határ subset[‡] $\text{k-MBS}_p(\underline{s}, Y) \Leftrightarrow (\exists \text{mbs} : \text{MBS}_p(\text{mbs}, Y), \underline{s} \subseteq \text{mbs}$.

A k-MBS_p fogalom gráf-alapú meghatározása a következő.

[‡] Mivel p 1 valószínűséggel stabil Dirichlet-paramétereloszlások esetén [13], szintén használjuk az indikátorfüggvényt $\text{k-MBS}(s, Y, G)$ feltéve, hogy p kompatibilis G-vel. Azonban a nem-stabil esetek miatt, ezeket a halmazokat k-s Markov-takaró részhalmazoknak is nevezik.

9.1. Propozíció. *Egy stabil p eloszlás esetén, amit (G, θ) Bayes-háló definiál, s egy k-s Markov-határ* k-MBS$_p(s, Y)$*, ha* $s \subseteq \mathrm{bd}(Y, G)$ *és* $|s| = k$.

A k-MBS jegyek előnye, hogy skálázhatóak, kardinalitásuk polinomiális $\mathcal{O}(n^k)$, éppen ezért jól alkalmazhatóak a relevanciaanalízis során. A gyakorlatban ez azt jelenti, hogy megvizsgálhatjuk a legvalószínűbb k-MBS(Y) jegyeket a k paraméter egy elég széles tartományában. További előnyük, hogy a k-MBS és k-MBG poszteriorok offline számíthatók a MBS és MBG poszteriorok közelítő értékéből. A legnagyobb k érték, amelynél az egyes modell-tulajdonságok (egyes strukturális jegyek) nagy valószínűséggel megjelennek, problémafüggő. Megfelelő k érték választásához bottum-up vagy top-down megközelítést kell alkalmazni, azaz értelemszerűen az előbbi esetben a vizsgált k paraméter kezdeti értéke $k = 1$, míg az utóbbinál $k = |V|$.

Szimmetria-okok miatt adódik a következő általánosítása a k-as Markov-takaró határ egy k prediktorra korlátozott fogalmának [5].

9.5. Definíció. Legyen az mbs változó halmaz egy Markov-takaró a $p(\underline{V})$ eloszlás esetén. Egy $\underline{s}$ változóhalmazt relevánsbelinek és k-as Markov-takaró-részhalmaznak (k-subMBS) nevezünk, ha $|\underline{s}| = k$ és $\underline{s} \subseteq$ mbs. Egy $\underline{s}$ változóhalmazt részben relevánsnak és k-as Markov-takaró-fedőhalmaznak nevezzük (k-supMBS), ha $|\underline{s}| = k$ és mbs $\subseteq \underline{s}$.

A k-subMBS és k-supMBS fogalmak a releváns változók jelenlétét és hiányát hivatottak kifejezni. Egy s_{sub} k-subMBS halmaz azokat a változókat tartalmazza, amelyek biztosan (szükségszerűen) erősen relevánsak. Egy s^c_{sup} k-supMBS halmazban nem szereplő változók a biztosan nem erősen releváns változókat tartalmazza (azaz egy k-supMBS részben releváns halmaz egy elégséges változóhalmazt tartalmaz). Vegyük észre, hogy a k-subMBS és k-supMBS fogalmak egy k-ban indexelt hierarchikusan kapcsolódó, átlapolódó hipotézishalmazt jelölnek. Valójában a k-subMBS-ek és k-supMBS-ek k-ban polinomiális számossága az MBM jegyek lineáris számosságát és az MBS-ek exponenciális számosságát hidalja át: $\mathcal{O}(n) < \mathcal{O}(\binom{n}{k}) < \mathcal{O}(2^n)$, ahol n jelöli változók számát. Mivel az MBG-k és DAG-ok számossága még ennél is magasabb [6], az MBM-ek, k-subMBS-ek/k-supMBS-ek, MBS-ek, MBG-k, esszenciális gráfok és DAG-ok egy egymásba ágyazott, egyre komplexebb hipotézisosztályt alkotnak a relevanciával kapcsolatban. Ennek megfelelően ezek a hierarchia-szintek természetes módon használhatóak fel egy többszintű relevancia elemzésben, amelyben a k-MBS-ek változó k-ra egy skálázhatóan többváltozós relevancia-elemzést tesznek lehetővé.

A Bayes-i megközelítésben egy s halmaz relevánsbeliségének poszteriorja:

$$\underline{p}(s|D_N) = p(\mathrm{MBS}(Y, G) = s|D_N) + \sum_{s' : s \subset s'} p(\mathrm{MBS}(Y, G) = s'|D_N), \tag{9.3}$$

Analóg módon, egy s halmaz részbenreleváns voltának poszteriorja:

$$\bar{p}(s|D_N) = p(\mathrm{MBS}(Y, G) = s|D_N) + \sum_{s' : s \supset s'} p(\mathrm{MBS}(Y, G) = s'|D_N). \tag{9.4}$$

9.5. Többcélváltozós relevancia

Egy összetett vizsgálatnál előfordulhat, hogy egyszerre több célváltozót kell együttesen megvizsgálni. Ilyen esetben a célváltozók Y halmazához keressük a releváns változókat, és a célváltozók közötti kapcsolat nem játszik szerepet. Tekinthetünk erre úgy is, mint egyfajta aggregálásra, ami hasonlít a korábban bemutatott jegyek aggregálására, csak ezúttal a célváltozókon elvégezve. Szerencsére a relevancia alapvető összefüggései egyszerűen kiterjeszthetőek célváltozó halmazokra.

9.6. Definíció (Multi-target relevance)**.** Egy jegy (véletlen változó) X_i erősen (gyengén) releváns $\underline{Y}$ célváltozókra, ha erősen (gyengén) releváns bármely $Y_i \in \underline{Y}$ elemre.

A Markov-takaró részgráf több célváltozóra való kiterjesztése hasonlóképp történik. A több célváltozóra számított MBG szintén meghatározza a szükséges és elégséges függőségi struktúrát és célváltozók predikciójához szükséges paramétereket.

9.7. Definíció. Egy Bayes-háló G részgráfját Y célváltozóhalmaz Markov-takaró részgráfjának nevezzük (MBG($\underline{Y}, G$)), ha az tartalmazza az Y célváltozóhalmaz Markov-takarójának csomópontjait és célváltozókba valamint azok gyermekeibe futó éleket.

9.6. Poszterior-dekomponáláson alapuló interakció és redundancia

A relevancia-analízis során a hangsúly jellemzően a nagy a posteriori valószínűségű jegyek elemzésére kerül, habár az alacsony valószínűség is jelezhet fontos összefüggéseket. Többek közt létrehozhatók olyan mértékek, melyek révén magasszintű szemantikus jellemzők mérhetőek. Ilyen az általunk létrehozott interakció és redundancia felfedését elősegítő mérték (score). Ennek számításához az egzakt k-MBS poszterior és annak MBM alapú approximációja szükséges. Az approximáció a k-MBS-beli változók (egy adott Y központi változóra vonatkozó) MBM valószínűségeinek szorzataként áll elő az alábbiak szerint:

$$p(\text{k-MBS}(\mathbf{X}', Y, G)|D_n) \approx \prod_{X_i \in \mathbf{X}'} p(\text{MBM}(Y, X_i, G)|D_n), \tag{9.5}$$

Ez a közelítő számítás alapvetően a struktúra poszterior dekomponálhatóságához kötődik és egy közvetlen Bayes-i megközelítést tesz lehetővé a redundancia és az interakció tulajdonságok vizsgálatára. Ugyanis ha egy magasabb rendű k-MBS poszterior nagyobb, mint egy approximált alacsonyabb rendű k-MBS poszterior, az azt jelenti, hogy a releváns változók halmazában vannak interakciós tagok. Az ellenkező eset – vagyis ha az approximált poszterior a nagyobb, mint a közvetlenül számított – pedig redundáns változók jelenlétét jelzi. Ez azzal magyarázható, hogy az approximált k-MBS poszterior számítása úgy történik, mintha a k-MBS független változókból állna, viszont a számított k-MBS poszterior a változók együttes hatásáról ad képet. Mindez a következő definícióval formalizálható:

9.8. Definíció (Interaction and redundancy)**.** Az $\mathbf{X}' = \{X_{i_1}, \ldots, X_{i_k}\}$ jegyek 1,k-szorzat interakciósak (redundánsak), ha a poszterior $p(\text{k-MBS}(\mathbf{X}', Y, G)|D_N)$ nagyobb (kisebb) mint $\prod_j p(\text{MBM}(X_{i_j}, Y, G)|D_N)$.

Megjegyezzük, hogy ez a definíció általánosítható magasabb rendű k-ra (azaz $k > 1$), illetve több célváltozóra. A redundás jegyek feltárására lehet úgy is tekinteni, mint a stabil jegyek feltárásának komplementerére, vagyis legegyszerűbb esetben olyan jegyeket kereshetünk, melyek a stabil jegyek mellett tűnnek fel. A k-(sub)MBS poszterior a statisztikai interakció új, rendszerszintű jellemzését teszi lehetővé, amely a valódi poszterior és alacsonyabb rendű k-subMBS poszteriorokon alapuló közelítés különbségén alapul, és a változók modellen keresztüli kölcsönös információtartalmával függ össze.

9.7. MBS poszteriorok utófeldolgozása és megjelenítése

Az MBS poszterior utófeldolgozásában és megjelenítésében a következő fogalmak és módszerek kiemelkedő fontosságúak (részletes bemutatásuk a Bioinformatika jegyzetben található).

1. *Feltételes MBS poszteriorok megjelenítése a modell struktúrára vetítve:* A Bayes-háló-struktúra felhasználható az MBSs és az MBM marginális poszteriorok megjelenítésére, amely akár a következő feltételes formában is megkonstruálható: $p(\text{mbs}|D_N, \alpha(\text{mbs}))$, ahol $\alpha(\text{mbs})$ egy tetszőleges logikai kifejezés a prediktorok MBS státuszáról.
2. *MBS és k-MBS poszteriorok megjelenítése részhalmaz hálón:* Mind a megjelenítés, mind az utófeldolgozás kihasználhatja a részhalmazok azon tulajdonságát, hogy a metszet és unió műveletekkel egy hálót alkotnak, ahol a minimális és maximális elemek az üres és a teljes halmazok. A megjelenítésben a háló tranzitív redukált térképe (TRM) használható, ahol a csomópontok a k. oszlopban a k méretű részhalmazokhoz tartoznak. A TRM egy DAG-ként is ábrázolható, ahol az élek a „part of" relációt jelölik.
3. *A relevancia-fa:* A relevancia-fa a relevanciabeliség poszteriorja szerint mutatja a prediktorok halmazait. A prediktorok részhalmazai méret szerint rendezve jelenik meg, mivel egy halmaz megjelenítésének vízszintes pozíciója, színe, mérete a halmaz relevanciabeliségétől függ (ami értelemszerűen monoton változik a mérettel).
4. A relevancia-interakció: A páronkénti, relevancia alapú statisztikai interakció egy hierarchikus interakciós diagramon ábrázolható. Ezen az egyes prediktorok (például SNP-k) erős relevanciáját egy oszlop jelzi a belső körön, a belső gyűrű egy magasabb aggregációs szintnek felel meg (például géneknek), a külső rész reprezentálja a legmagasabb szintű entitások relevanciáját (például nagyobb kromoszomális régiók). Az élek vastagsága arányos az interakciók erősségével, illetve piros jelzi az interakciót és kék a redundanciát.

9.8. Tudás alapú utóaggregálás

A relevancia Bayes-i megközelítésének az az előnye, hogy a modell poszterior elméleti megkötések nélkül transzformálható és értelmezhető. Jelen esetben a Bayes-háló-struktúrák terét alkalmazva ez azt jelenti, hogy a poszterior aggregálható a G modellstruktúrák felett, ahol minden particionálás egy potenciálisan új értelmezést tesz lehetővé. Jellemzően kevés partíció rendelkezik általános vagy tárgyterület-specifikus értelmezéssel. A nem-informatív modellaggregálás mellett lehetséges informatív aggregálás is az a priori tárgyterületi tudás felhasználásával. Mindkét esetben az aggregálás (1) lehetővé teszi a tárgyterületi relevancia-relációk általános leírását, valamint (2) magasabb konfidencia-szintű numerikus eredményeket eredményez. Például egynukleotidos polimorfizmusok (SNP-k) esetén a génszintre aggregálás egy természetes lépés, mivel számos SNP kötődik egy adott génhez. Az aggregálás révén a gének szintjén is számítható a Markov-takaróba tartozás (MBM) és a Markov-takaró halmaz (MBS) relációk. A számítás módja levezethető a megfelelő SNP szintű számításokból. Az alábbiakban erre látható egy példa, amely egy adott génhez tartozó SNP-k Y változó Markov-takarójába tartozásának valószínűségét adja meg:

$$p(MBM(Y,g|D)) = \sum_{G:\exists s:onGene(g,s))\wedge MBM(Y,s,G)} p(G|D). \tag{9.6}$$

9.9. Összefoglaló

A Bayes-hálón alapuló többszintű Bayes-i metodológia egy igen részletes relevancia-elemzést tesz lehetővé, amely révén többek között képet kapunk a mintaszám elégséges voltáról is. Továbbá lehetőséget nyújt széleskörű tárgyterületi a priori tudás felhasználására, és kiválóan alkalmazható kis mintaméret esetén is. Az interakciók MBG jegy alapú egzakt modellezése lehetővé teszi a releváns jegyek és a köztük lévő interakciók tanulási bizonytalanságának számszerűsítését. Az MBS és MBG komplex modelltulajdonságok célváltozó-fókuszáltak, de rendszerszemléletűek, skálázhatóak, polinom komplexitással. Több célváltozó (célváltozóhalmaz) együttes vizsgálatát is lehetővé teszi, illetve interakció és redundancia feltárására is alkalmas, ami alapvetően a struktúra poszterior dekomponálhatóságán alapszik.

Irodalomjegyzék

[1] S. Acid, L. M. de Campos, and J. G. Castellano, Learning Bayesian network classifiers: searching in a space of partially directed acyclic graphs. *Machine Learning*, 59:213–235, 2005.

[2] C.F. Aliferis, I. Tsamardinos, and A. Statnikov, Large-scale feature selection using Markov blanket induction for the prediction of protein-drug binding, 2003.

[3] P. Antal, G. Fannes, Y. Moreau, D. Timmerman, and B. De Moor, Using literature and data to learn Bayesian networks as clinical models of ovarian tumors. *Artificial Intelligence in Medicine*, 30:257–281, 2004.

[4] P. Antal, G. Hullám, A. Gézsi, and A. Millinghoffer, Learning complex Bayesian network features for classification. In *Proc. of third European Workshop on Probabilistic Graphical Models*, pages 9–16, 2006.

[5] P. Antal, A. Millinghoffer, G. Hullám, Cs. Szalai, and A. Falus, A Bayesian view of challenges in feature selection: Feature aggregation, multiple targets, redundancy and interaction. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 4:74–89, 2008.

[6] G. F. Cooper and E. Herskovits, A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[7] N. Friedman and D. Koller, Being Bayesian about network structure. In *Proc. of the 16th Conf. on Uncertainty in Artificial Intelligence(UAI-2000)*, pages 201–211. Morgan Kaufmann, 2000.

[8] N. Friedman and D. Koller, Being Bayesian about network structure. *Machine Learning*, 50:95–125, 2003.

[9] R. Kohavi and G. H. John, Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324, 1997.

[10] M. Koivisto and K. Sood, Exact Bayesian structure discovery in Bayesian networks. *Journal of Machine Learning Research*, 5:549–573, 2004.

[11] D. Koller and M. Sahami, Toward optimal feature selection. In *International Conference on Machine Learning*, pages 284–292, 1996.

[12] D. Madigan, S. A. Andersson, M. Perlman, and C. T. Volinsky, Bayesian model averaging and model selection for Markov equivalence classes of acyclic digraphs. *Comm.Statist. Theory Methods*, 25:2493–2520, 1996.

[13] C. Meek, Causal inference and causal explanation with background knowledge. In *Proc. of the 11th Conf. on Uncertainty in Artificial Intelligence (UAI-1995)*, pages 403–410. Morgan Kaufmann, 1995.

[14] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, CA, 1988.

[15] J.M. Pena, R. Nilsson, J. Bjorkegren, and J. Tegnér, Towards scalable and data efficient learning of Markov boundaries. *International Journal of Approximate Reasoning*, 45:211–232, 2007.

[16] I. Tsamardinos and C. Aliferis, Towards principled feature selection: Relevancy, filters, and wrappers. In *Proc. of the Artificial Intelligence and Statistics*, pages 334–342, 2003.

[17] I. Tsamardinos, C. F. Aliferis, and A. Statnikov, Algorithms for large-scale local causal discovery and feature selection in the presence of limited sample or large causal neighbourhoods. In *The 16th International FLAIRS Conference*, 2003.

[18] Lei Yu and Huan Liu, Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5:1205–1224, 2004.

10. fejezet

Hálózatbiológia

10.1. Bevezetés

A XXI. század első évtizedében új korszak köszöntött be az orvosbiológiai kutatások történetében. Ezen – gyakran „poszt-genomikus" névvel illetett – korszak sajátossága a különböző sejtszintű komponensek holisztikus, rendszerszintű szemlélete; egyes entitások (pl. gének, fehérjék) vizsgálata helyett komplex kapcsolatok és interakciós mintázatok leírása. A számítástechnika és méréstechnikák fejlődése hatalmas ugráshoz vezetett a heterogén, különböző omikai szinteken létező biológiai adatok mennyiségében, új kihívásokat teremtve napjaink tudósainak. A rendszerbiológia célja, hogy újszerű betekintést nyújtson, illetve több sejtbiológiai szinten egyszerre operáló eszközökkel támogassa a kutatókat ezen erőpróba során.

Nem kell sokáig keresgélnünk, ha a rendszerbiológiai szemléletet matematikai keretbe próbáljuk foglalni: a hálózatelmélet az egyik kézenfekvő választásként adódik. A gráfelmélet ezen alterületének meglátása szerint az egész több, mint a részek összessége, így az érdeklődés középpontjában diszkrét entitások közötti kapcsolatok, mintázatok, illetve a hálózatok emergens tulajdonságai állnak. A „hálózat" kifejezés azonban kissé pongyola, számos különböző fogalmat jelölhet, amelyek gyakorlati haszna eltérő lehet. Tisztázzuk tehát, hogy mit is érthetünk „hálózat" alatt az alábbi négy fogalmi szint elkülönítésével:

1. **Hasonlósági hálózatok**, pl. szekvencia hasonlósági hálózatok egyszerűen generálhatók tetszőleges hasonlósági mátrixokból. Bár számos alkalmazás során igen hasznosnak bizonyultak, jóval kevésbé kifinomultak, mint a 3. és 4. pont kvantitatív modelljei.

2. **Leíró gráfok**, pl. a fehérje–fehérje interakciós hálózatok a hálózatbiológia főáramát képviselik; számos kutató ezt a szintet tartja „A" rendszerbiológia szintjének.

3. **Függetlenségi térképek és oksági diagramok**, pl. a Bayes-hálók nagy népszerűségnek örvendenek a bioinformatika területén, bár hagyományosan inkább tartják egyfajta statisztikai megközelítésnek, mint a hálózatelmélet és hálózatbiológia részének.

4. **Kvantitatív szabályozási hálózatok** kifinomult matematikai modelljei különböző sejtszintű folyamatoknak és funkcióknak; gyakran közönséges és parciális differenciálegyenletek segítségével modelleznek biokémiai reakciókat.

Ebben a fejezetben bevezetjük az olvasót a leíró hálózatelmélet alapfogalmaiba, amelyek elsősorban az első két kategóriához tartoznak, és nem feltétlenül rendelkeznek generatív kvalitásokkal. Egyes érvelések szerint az „igazi" rendszerbiológia éppen az utolsó két kategóriában található; egyelőre azonban nincs egyetértés abban, hogy egyes tudósok mit is értenek rendszer- (hálózat-) biológia alatt. Végül megjegyezzük, hogy egy kimerítő összefoglalás messze meghaladná e tankönyv kereteit, így teljesebb áttekintésekért más szerzők műveire hivatkozunk [1, 2].

10.2. Biológiai hálózatok

A biológiai hálózatok – a legegyszerűbb sejtektől teljes ökoszisztémákig – közös jellemzői az összetett interakciók az egyes komponensek között. Számos példája ismert az ezek leírására törekvő biológiai hálózatoknak, ezek közül álljon itt néhány ismertebb:

- **Szekvencia/szerkezeti hasonlósági hálózatok**at entitáspárokra értelmezett hasonlóságmérték meghatározásával tudunk származtatni. Entitás alatt leggyakrabban géneket, fehérjéket, kismolekulákat (pl. gyógyszereket), vagy – a struktúrán és szekvencián túl – elvontabb objektumokat értünk (pl. betegségek, génexpressziós profilok). A hálózatok ezen válfaja széles alkalmazási területe folytán meglehetősen népszerű (pl. funkció és interakciók predikciója [3, 4], gyógyszerkutatás [5]).

- **Fehérje-fehérje interakciós hálózatok (PPI, PIN)** építése fizikai fehérjekötődési adatok alapján történik, rendszerint nagy áteresztőképességű eszközök felhasználásával. Elsődleges alkalmazási területük a fehérjék funkciójának meghatározása interakcióik elemzésével. Néhány publikus adatbázis: DIP [6], MINT [7].

- **Metabolikus hálózatok**at élő szervezetek metabolikus útvonalainak vizsgálatára használunk. Építőelemei között találunk enzimeket, ezek szubsztrátjait és termékeit (metabolitok), valamint a katalizált reakciók reprezentációit. A legszélesebb körben elterjedt nyílt adatbázisok pl. a KEGG [8] és a BioCyc [9].

- **Szignál transzdukciós hálózatok** a szignálok továbbítását, releváns molekuláris útvonalakat és a cross-talk mechanizmusokat helyezik a középpontba. Előbbire példa a MiST [10] és TRANSPATH [11] adatbázisok; kifejezetten cross-talk mechanizmusok elemzésére szolgál a SignaLink [12].

- **Szabályozási hálózatok (GRN)** a génexpresszió szabályozását vizsgálják, ideértve a szabályozási régiókat, transzkripciós faktorokat, RNS interferenciát, poszttranszlációs módosításokat és más faktorokkal történő interakciókat. Két publikus adatbázis a JASPAR [13] a TRANSFAC [14].

- **Egyéb integrált hálózatok** hozhatók létre több heterogén információforrás kombinálásával, így az entitásokat egységes nézőpontból vizsgálhatjuk. Ilyenek például a többrétegű szabályozási hálózatok, gyógyszer–betegség–gén hálózatok és számos más publikus eszköz – itt említhető a Connectivity Map, amely betegségeket, kismolekulákat és génexpressziós adatokat integrál [7].

10.3. Gráfelméleti alapok

Ebben a fejezetben néhány gráfelméleti alapfogalommal ismerkedünk meg. A *gráf* egy *csúcs*okból és *él*ekből álló gyűjtemény, amelyet a $G = (V, E)$ rendezett párral jelölünk, ahol V a csúcsok (vagy *csomópontok*) halmaza, míg E az élek (vagy *kapcsolatok*) halmaza. Minden él megfeleltethető egy V-beli csúcspárnak – egy él mindig két, *szomszédos*nak nevezett csúcsot köt össze (ám ez a kettő lehet ugyanaz a csúcs). Számos esetben szükségessé válik az élek *irányítása* – képzeljünk csak el egy családfát, amely így az *irányított gráfok* csoportját gazdagítja. E gráfokban az élek rendezett csúcspárokként reprezentálhatók; más esetekben a kapcsolatok szimmetrikus volta ezt nem követeli meg (*irányítatlan gráfok*). Az irányított gráfok speciális esetei az *irányított körmentes gráfok* (DAG), amelyek, ahogy azt nevük is sugallja, nem tartalmazhatnak kört – e tulajdonság számos alkalmazásban nagyon fontosnak bizonyul. Néhány esetben hasznos, ha az élekhez számszerű értékeket rendelünk. Ezeket *súlyozott él*nek nevezzük, a gráfot pedig *súlyozott gráf*nak.

Egy adott csúcsra illeszkedő (kapcsolódó) élek számát nevezzük a csúcs *fokszámának*. A *szabályos gráfok*ban minden csúcs fokszáma megegyezik. A *teljes gráf* az előbbinek speciális esete, ahol bármely két csúcsra illeszkedik él. Értelemszerűen nem minden gráf teljes, sőt, még csak nem is feltétlenül *összefüggő*. Összefüggőnek nevezzük a gráfot, ha bármely két csúcsa között létezik út – ellenkező esetben a gráf *nem összefüggő*. Egy gráf *részgráf*ja az eredeti gráf kiválasztott csúcsaiból és éleiből áll, ahol a kiválasztott élek kiválasztott csúcsokra illeszkednek. A maximális (lehető legnagyobb) összefüggő részgráfokat *komponens*eknek nevezzük, azaz egy nem összefüggő gráf több komponens tartalmaz, míg egy összefüggő gráf pontosan egyet. Egy gráf teljes részgráfjait *klikk*eknek nevezzük, a lehető legnagyobb klikkeket pedig *maximális klikk*eknek. A gráfok egy speciális fajtája a *páros gráf*, ahol a csúcsok két diszjunkt halmazt alkotnak, ahol azonos halmazbeli csúcsokra nem illeszkedik él – képzeljünk el egy sakktáblát, ahol minden fekete mező csak fehérrel szomszédos, és fordítva. Végül, egy *klaszter* a csúcshalmaz egy olyan részhalmaza, amelyben a csúcsok „sokkal erősebben” kapcsolódnak egymáshoz, mint a gráf többi részéhez.

A gráf klasztereződésének méréséhez a *klasztereződési együttható* különböző definícióit lehet igénybe venni. További fontos mértékek például a *legrövidebb út*, az *átlagos úthossz*, a *hálózati centralizáció*, csomóponti *centralitás*ok (pl. *fokszám-*, *közelségi*, *sajátvektor-* stb. centralitás). Ezek tárgyalása túlmutat a könyv keretein, így a részletekért más művekre hivatkozunk [1, 2].

10.4. Hálózatelemzés

A hálózatelemzés a hálózat kvalitatív és kvantitatív tulajdonságait vizsgálja, ideértve a mögöttes strukturális alapelveket, funkcionális szerveződést, lokális mintázatokat, emergens tulajdonságokat és dinamikus viselkedést. Interdiszciplináris területről lévén szó, alkalmazási területe nem korlátozódik a hálózatbiológiára; hasonló eszközöket használnak a telekommunikációban, szociális hálózatok elemzésében és számos egyéb területen.

10.4.1. Hálózati topológia

A hálózati topológia a csomópontok és kapcsolataik elrendeződését jellemzi, azaz leírja, hogyan kapcsolódnak, „kommunikálnak" egymással az egyes csomópontok. Ahogy a 10.3. alfejezetben láthattuk, a gráfok gyakran rendelkeznek jól meghatározott strukturális elemekkel (pl. klikkek, klaszterek); ebben az alfejezetben hasonló, hálózatelemzésben gyakran vizsgált elemekkel ismerkedünk meg, amelyek jelentősen befolyásolják a hálózat viselkedését.

Az átlagosnál sokkal több kapcsolattal rendelkező csomópontokat *hub*-oknak nevezzük. A hubok bizonyos értelemben a hálózat kulcsszereplői – törlésük rendszerint a hálózat gyors degradációjához, izolált klaszterekre való széteséséhez vezet. Ez a jelenség PPI hálózatok esetén „centralitási–letalitási szabály" néven ismert, mivel a hub-ok gyakran nélkülözhetetlen fehérjéknek felelnek meg. Lokális topológiai struktúrák még a *motívumok* (szignifikánsan felülreprezentált irányított részgráfok) és *graphletek* (az előbbiek irányítatlan megfelelői).

A hálózatbiológia nevezéktanában a *modul* többé-kevésbé a gráfelméleti klaszternek felel meg. Gyakran funkcionális alrendszereket reprezentálnak, pl. bizonyos sejtszintű folyamatokat vagy funkciókat. Összetett rendszerekben több típusú interakció is elképzelhető az egyes modulok között, például *átlapolódás*on vagy *hidak*on (modulokat összekötő csomópontokon) keresztül. Ha egy híd az egyetlen összekötő elem két modul között, *bottleneck*-nek nevezzük. A modulok hierarchikus elrendeződést is mutathatnak; kisebb, interakcióban lévő modulok nagyobb, lazább modulok alkotóiként szerepelhetnek. A hálózatok klaszterezése intenzíven kutatott terület, amely a modulok azonosítását célozza. Széles eszköztárában megtalálhatók gráfelméleti, statisztikai és gépi tanulási eljárások egyaránt.

A csomóponti *centralitás* általánosságban „befolyásos" csomópontok jelenlétére utal; ha léteznek a hálózat egyfajta globális „koordinátoraként" viselkedő csomópontok, ezek magas centralitással bírnak. Néhány centralitási mértéket említettünk az előző alfejezetben. Idevágó fogalom a hálózati *centralizáció*, amely a csomóponti centralitások eloszlását veszi figyelembe, tehát a hálózat egészére vonatkozik – erősen centralizált hálózatok gyakran csillagszerű topológiát mutatnak, a skála másik végén egyenletesebb eloszlással találkozhatunk. A magas centralitású csomópontokból álló alhálózatot *csontváz*nak nevezzük.

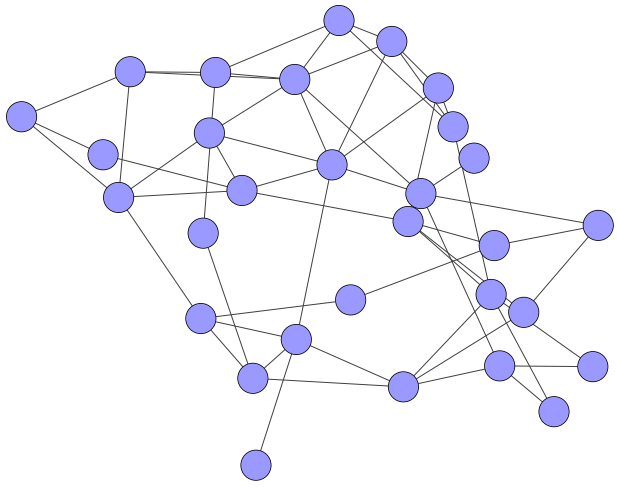
A valós hálózatok egyik lenyűgöző tulajdonsága a meglepően alacsony átlagos úthossz, a hálózat esetenként hatalmas mérete ellenére. Ezt a jelenséget gyakran *kisvilág*-tulajdonságnak nevezik. A kifejezés a társadalomtudományból és Stanley Milgram kutatásaiból származik, bár elsőként Karinthy Frigyes vetette fel; példájában kifejti, hogy

bármely személy a földön elérhető személyes ismeretségek útján legfeljebb öt lépésben (később: „six degrees of separation").

10.4.2. Hálózati modellek és dinamika

A valóságban sok hálózat – különösen a biológiai rendszereket modellezők – időben folyamatosan változik és fejlődik. A hálózati dinamika rohamosan gyarapodó területe ezeket a temporális aspektusokat hivatott vizsgálni. A komplex hálózatok tulajdonságainak megértéséhez célszerű megfigyelni azok kialakulását és fejlődését, felfedezni a mögöttes szerveződési alapelveket. Ezek a modellek lényegében „prototípusai" a valóságban fellelhető hálózatoknak, céljuk pedig betekintést nyújtani abba, hogyan következnek az emergens tulajdonságok kis számú egyszerű konstrukciós szabályból. Az elmúlt ötven évben számos modellt alkottak, amelyek közül a leghíresebbek az Erdős–Rényi-modell [16], a Watts–Strogatz-modell [17] és a Barabási–Albert-modell [18].

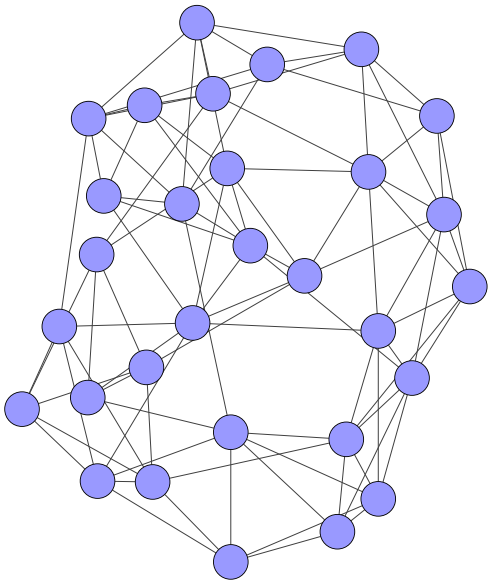
Az Erdős–Rényi-modell az egyik legegyszerűbb modell véletlen gráfok leírására. A konstrukció N csomóponttal indul, majd véletlenszerűen húz be éleket az $N(N-1)/2$ lehetőségből. E modell példányai rendelkeznek a kisvilág-tulajdonsággal, ám a fokszámok között csak kis variancia tapasztalható, azaz nem képesek megmagyarázni a valós hálózatok klasztereződési tendenciáját (pl. hubok formálódását).



10.1. ábra. Az Erdős–Rényi-modell egy példánya, 30 csomóponttal és $p = 0.1$ valószínűségi paraméterrel

A Watts–Strogatz-modell mind a kisvilág-tulajdonságot, mind a lokális klasztereződést reprodukálja. Kezdetben az N darab csomópont egy körben van elrendezve, továbbá minden csomópont össze van kötve $k/2$ legközelebbi szomszédjával. Ezután minden él egy kis p valószínűséggel „áthuzalozódik", azaz egyik vége egy véletlenszerűen kiválasztott csomóponthoz csatlakozik – ennek köszönhető a kisvilág-tulajdonság. Ha p-t megfelelően,

de nem extrém módon kicsire választjuk, elfogadható mértékű lokális klaszterező́dés marad a hálózatban; $p = 1$-re az Erdős–Rényi-modellt kapjuk vissza.

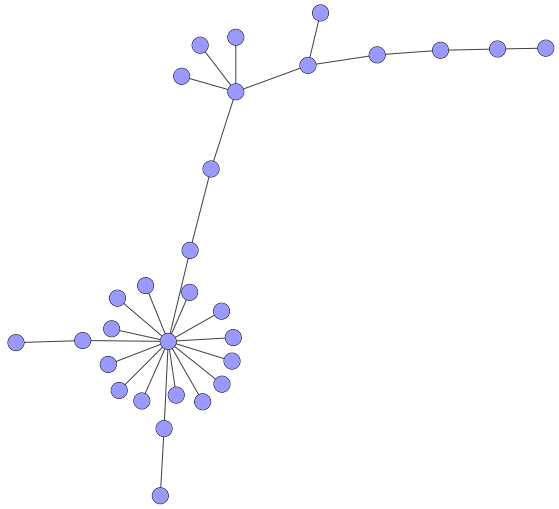


10.2. ábra. A Watts–Strogatz-modell egy példánya, 30 csomóponttal és $p = 0.1, k = 3$ paraméterezéssel

A Barabási–Albert-modell nemcsak a fenti tulajdonságokat mutatja, hanem skálafüggetlen fokszámeloszlást is, amely gyakran megfigyelhető valós hálózatokban, például a biológia területén vagy az Interneten (lásd a következő alfejezetet). A modell alapötlete a *növekedés* és *preferenciális kapcsolódás* alkalmazása. A hálózat ismételten új csomópontokkal egészül ki (növekedés), ezek kapcsolatai pedig valószínűségi alapon, a többi csomópont aktuális fokszámát figyelembe véve alakulnak ki; más szavakkal, az új csomópont a már eddig is sok kapcsolattal rendelkezőket preferálja a kapcsolódás során (preferenciális kapcsolódás, „a gazdag még gazdagabbá válik”). A preferenciális kapcsolódás hűen modellezi számos valós (pl. szociális) hálózat formálódási szabályait; meggyőző magyarázatok állnak rendelkezésre arról is, hogy sejtszintű hálózatok miért követik szintén ezt a sémát és rendelkeznek skálafüggetlen topológiával [19].

10.4.3. Asszortativitás, fokszámeloszlás és skálafüggetlen hálózatok

Az *asszortativitás* a csomópontok „hasonló” csomópontokhoz történő preferenciális kapcsolódását írja le; „hasonló” alatt rendszerint hasonló fokszámot értünk. Asszortatív hálózatokban a sok kapcsolattal rendelkező csomópontok más, sok kapcsolattal rendelkező csomópontokat preferálnak; a biológiai hálózatok rendszerint *diszasszortatívek*, azaz magas fokszámú csomópontok alacsony fokszámúakhoz kapcsolódnak [20].

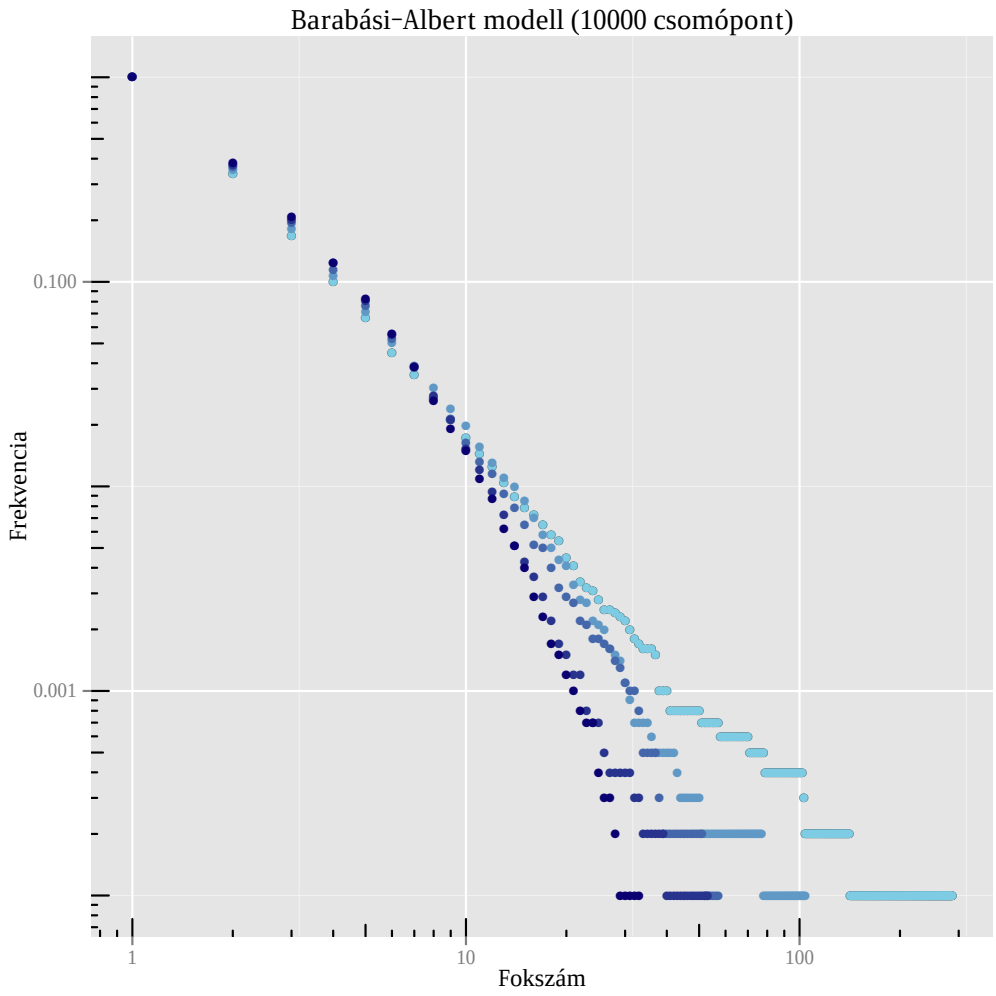


10.3. ábra. A Barabási–Albert-modell egy példánya, 30 csomóponttal és $\gamma = 2$ fokszám-kitevővel

A biológiai hálózatok további kulcsfontosságú tulajdonsága, hogy a fokszámeloszlás hatványfüggvényt követ, ún. *skálafüggetlen* hálózatot eredményezve. A fokszámeloszlás ($p(k)$) annak valószínűségét adja meg, hogy egy csomópont fokszáma pontosan k. Az Erdős–Rényi-modellben a fokszámeloszlás binomiális, ami nagy hálózatokban Poisson-eloszlással becsülhető, tehát az átlagos fokszámnál erősen csúcsosodik (az átlagostól nagyon eltérő fokszámú csomópontok extrém ritkák). A skálafüggetlen hálózatok $p(k) \sim k^{-\gamma}$ alakú fokszámeloszlást követnek, így néhány magas fokszámú csomópontra (hubok) sok alacsony fokszámú jut (10.4. ábra). A γ *fokszámkitevő* alapvetően meghatározza a hálózat viselkedését. Minél magasabb az értéke, a $p(k)$ függvény annál meredekebb lesz, így $\gamma > 3$ értékeknél nagy hubok már csak elvétve fordulnak elő és nem játszanak lényeges szerepet; fordítva pedig, γ alacsonyabb értékeinél a hubok jelenléte kifejezett. A legtöbb biológiai hálózat fokszámkitevője 2 és 3 között van. Mint kiderült, ezek a hálózatok ráadásul „ultra-kicsik" abban az értelemben, hogy az átlagos úthossz jelentősen rövidebb, mint véletlen hálózatok esetében. További részletekért ajánljuk Barabási és munkatársainak közleményeit [18, 19].

10.4.4. Feladatok és kihívások

A gyakorlatban a biológiai rendszerekről rendelkezésre álló tudásunk sosem teljes. Ennek számos oka lehet – elméleti tudatlanság, gyakorlati korlátok, eredendő bizonytalanságok, hibák, lustaság, csak hogy néhányat említsünk. Ebből következik, hogy a legjobb módszertannal és végrehajtással is csak tökéletlen modellekhez juthatunk. Bár a tökéletesség elérése a gyakorlatban kivitelezhetetlen, a modellek jelentősen javíthatók az adatokba ágyazott „rejtett" struktúrák és kapcsolatok kihasználásával, ezzel eddig ismeretlen információt hozva felszínre. Ez a hálózatbiológia kontextusában hálózatelemzési problémák



10.4. ábra. Fokszámeloszlások különböző fokszámkitevőkkel

megoldását jelenti, amelyeknek számos válfaja ismert:

- **Csomópontok és kapcsolatok jóslása** az egyik legkézenfekvőbb feladat. Csomópontok és kapcsolatok jósolhatók például hasonlóságok, topológiai vagy temporális tulajdonságok, vagy hálózati összehasonlítás felhasználásával [3].
- **Klaszteranalízis** használható funkcionális modulok felismerésére és interakcióik elemzésére biológiai rendszerekben.
- **Klasszifikáció, regresszió és rangsorolás** a gépi tanulás területéről származó általános fogalmak. A hálózatelemzési problémák széles körében alkalmazhatók, pl. csomópontok vagy kapcsolatok jóslására, tulajdonságaik felderítésére stb.

- **Centralitás-elemzés, útkeresés, robosztusság elemzése** használható a hálózat szerveződésének megértésére és a csomópontok „kommunikációjának" leírására. Egy nyilvánvaló alkalmazás gyógyszercélpontok azonosítása, azaz annak eldöntése, hogy milyen csomópontokat vagy éleket érdemes megtámadni a betegség hatásainak kiküszöbölése érdekében, a legkevesebb mellékhatás elérése mellett – vagy éppen hogyan lehet a sejtet minél hatékonyabban elpusztítani (antibiotikumok, rákellenes szerek).

- **Gráf-izomorfizmus és hálózatillesztés** a hálózatintegrációval rokon új keletű feladatok. Egy kutatásban például több faj PPI hálózatait illesztették és fehérjék funkcionális ortológiájára következtettek [22].

- **Gráf motívumkeresés**, amely az előbbihez hasonló, sikeresen alkalmazták például metabolikus hálózatokra a szerkezetük és építőelemeik mélyebb megértéséhez [23].

- **Hálózatok becslése** vagy „visszafejtése" (reverse engineering) alatt a hálózat struktúrájának adatokból történő meghatározását értjük. Fontos megjegyezni, hogy az így meghatározott szerkezet nagyban függ az alkalmazott módszertől, ezért egyre inkább több becslés integrációjára és együttes felhasználására kerül a hangsúly.

- **Hálózat-integráció**, célja több hálózat kombinációja, amellyel a tudásfúzió területére jutunk.

- **Hálózat-vizualizáció** a legegyszerűbb, mégis a legfontosabb feladatok egyike. A Cytoscape rendszer valószínűleg a legnépszerűbb eszköz biológiai hálózatok vizualizációjára; emellett rendkívüli segítséget jelenthet a hálózatelemzési problémák széles skáláján.

10.5. Néhány alkalmazás

A gyógyszerkutatás és -fejlesztés hagyományosan elsősorban olyan molekulák tervezését tűzte ki célul, amelyek egyetlen, legfeljebb néhány célponthoz kötnek maximális szelektivitással. Bár régóta ismert, hogy számos sikeres gyógyszer kifejezetten sok célponton hat egyszerre, a hálózati biológia és a gyógyszerkutatás csak az utóbbi néhány évben kezdtek egymásra találni (*network pharmacy*). Ez az egyesülés új, hatásosabb és alacsonyabb toxicitású gyógyszerek ígéretét rejti magában. A hálózati megközelítés az ún. gyógyszer-újrapozicionálás szempontjából is vonzó. Mivel a gyógyszeripar új molekula-kibocsátása évről évre csökken, a már forgalomban lévő gyógyszerek „újrahasznosítása" más indikációkban ésszerű stratégiát képvisel.

E szakterület fiatal kora ellenére jónéhány közlemény született, amelyek a hálózatelemzés módszereit kísérelték meg a gyógyszerfejlesztés és gyógyszer-újrapozícionálás területén kamatoztatni. Ezek közül számos próbálkozás gyógyszercélpontok azonosítására törekedett az előző alfejezetben ismertetettekhez hasonló eljárásokkal; mások a hasonlósági megközelítést követve több információs szintet hoztak létre (pl. gyógyszer–gyógyszer és betegség–betegség hasonlósági hálók), majd *ad hoc* módon kombinálták ezen szinteket.

A Lamb és mtsai által fejlesztett Connectivity Map a génexpressziós változások nyelvét használta fel a gyógyszerek, betegségek és gének szintjeinek egyesítésére [7]. A génexpressziós profilok változásait experimentálisan határozták meg számos gyógyszer és betegség esetében; a gyógyszer–betegség kapcsolatokat a profilok ellentétes irányú változásai alapján állapították meg. A PREDICT rendszer [24] nagyszámú hasonlóságot definiál gyógyszerek között (kémiai leírások, mellékhatások, szekvencia, PPI-hálózatbeli közelség és funkcionális annotáció alapján), valamint betegségek között (pl. fenotípusos és genetikai jellemzők alapján). Ezután egy gépi tanulási megközelítést használva gyógyszer–betegség párokat azonosítanak ismert párokhoz való hasonlítás alapján.

Minden $drug_i$–$disease_j$ párhoz jellemzők számolhatók az alábbi pontozófüggvénnyel:

$$score(drug_i, disease_j) = \max_{k \neq i, l \neq j} \sqrt{sim(drug_i, drug_k) \cdot sim(disease_j, disease_l)},$$

ami lényegében a legközelebbi ismert gyógyszer–betegség párhoz való hasonlóságot számítja ki minden sim hasonlóságmértékre. Ezeket jellemzőkként használva az ismeretlen párok klasszifikálhatók logisztikus regresszió útján, amely egyben a jellemzők súlyozását is elvégzi, és egy végső klasszifikációs pontszámot ad.

Irodalomjegyzék

[1] G. A. Pavlopoulos, M. Secrier, C. N. Moschopoulos, T. G. Soldatos, S. Kossida, J. Aerts, R. Schneider, and P. G. Bagos, Using graph theory to analyze biological networks. *BioData Min*, 4:10, 2011.

[2] Björn H. Junker and Falk Schreiber, *Analysis of Biological Networks*. Wiley Series in Bioinformatics, Wiley-Interscience, 2008.

[3] T. Phuong and N. Nhung, Predicting gene function using similarity learning. *BMC Genomics*, 14 Suppl 4:S4, Oct. 2013.

[4] Q. Chen, W. Lan, and J. Wang, Mining featured patterns of MiRNA interaction based on sequence and structure similarity. *IEEE/ACM Trans Comput Biol Bioinform*, 10(2):415–422, 2013.

[5] P. Csermely, T. Korcsmaros, H. J. Kiss, G. London, and R. Nussinov, Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol. Ther.*, 138(3):333–408, June 2013.

[6] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg, DIP: the database of interacting proteins. *Nucleic Acids Res.*, 28(1):289–291, Jan. 2000.

[7] A. Chatr-aryamontri, A. Ceol, L. M. Palazzi, G. Nardelli, M. V. Schneider, L. Castagnoli, and G. Cesareni, MINT: the Molecular INTeraction database. *Nucleic Acids Res.*, 35 (Database issue):D572–574, Jan. 2007.

[8] M. Kanehisa and S. Goto, KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28(1):27–30, Jan. 2000.

[9] R. Caspi, T. Altman, R. Billington, K. Dreher, H. Foerster, C. A. Fulcher, T. A. Holland, I. M. Keseler, A. Kothari, A. Kubo, M. Krummenacker, M. Latendresse, L. A. Mueller, Q Ong, S. Paley, P. Subhraveti, D. S. Weaver, D. Weerasinghe, P. Zhang, and P. D. Karp, The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*, 42(1):D459–471, Jan. 2014.

[10] L. E. Ulrich and I. B. Zhulin, MiST: a microbial signal transduction database. *Nucleic Acids Res.*, 35 (Database issue):D386–390, Jan. 2007.

[11] F. Schacherer, C. Choi, U. Gotze, M. Krull, S. Pistor, and E. Wingender, The TRANSPATH signal transduction database: a knowledge base on signal transduction networks. *Bioinformatics*, 17(11):1053–1057, Nov. 2001.

[12] D. Fazekas, M. Koltai, D Turei, D. Modos, M. Palfy, Z. Dul, L. Zsakai, M. Szalay-Bekő, K. Lenti, I. J. Farkas, T. Vellai, P. Csermely, and T. Korcsmaros, SignaLink 2 - a signaling pathway resource with multi-layered regulatory networks. *BMC Syst Biol*, 7:7, 2013.

[13] A. Sandelin, W. Alkema, P. Engstrom, W. W. Wasserman, and B. Lenhard, JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, 32 (Database issue):D91–94, Jan. 2004.

[14] E. Wingender, X. Chen, R. Hehl, H. Karas, I. Liebich, V. Matys, T. Meinhardt, M. Pruss, I. Reuter, and F. Schacherer, TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.*, 28(1):316–319, Jan. 2000.

[15] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J. P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wei, S. A. Armstrong, S. J. Haggarty, P. A. Clemons, R. Wei, S. A. Carr, E. S. Lander, and T. R. Golub, The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–1935, Sep. 2006.

[16] P. Erdős and A. Rényi, On the evolution of random graphs. In: *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, pages 17–61, 1960.

[17] M. E. Newman, S. H. Strogatz, and D. J. Watts, Random graphs with arbitrary degree distributions and their applications. *Phys Rev E Stat Nonlin Soft Matter Phys*, 64(2 Pt 2):026118, Aug. 2001.

[18] A. L. Barabasi and R. Albert, Emergence of scaling in random networks. *Science*, 286(5439):509–512, Oct. 1999.

[19] A. L. Barabasi and Z. N. Oltvai, Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, 5(2):101–113, Feb. 2004.

[20] M. E. Newman, Assortative mixing in networks. *Phys. Rev. Lett.*, 89(20):208701, Nov. 2002.

[21] Linyuan Lü and Tao Zhou, Link prediction in complex networks: A survey. *Physica A*, 390(6):11501170, 2011.

[22] R. Singh, J. Xu, and B. Berger, Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proc. Natl. Acad. Sci. U.S.A.*, 105(35):12763–12768, Sep. 2008.

[23] V. Lacroix, C. G. Fernandes, and M. F. Sagot, Motif search in graphs: application to metabolic networks. *IEEE/ACM Trans Comput Biol Bioinform*, 3(4):360–368, 2006.

[24] A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan, PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.* 7:496, 2011.

11. fejezet

Dinamikus modellezés a sejtbiológiában

A kísérleti biológia nagy áteresztőképességű módszereinek köszönhetően mára hatalmas mennyiségű adat van a birtokunkban. Ahogy az adatgyűjtés egyszerűvé vált, úgy válik az értelmezés egyre inkább kihívássá. A modellezés a tudás formális specifikációba rendezésének eszköze, ezt felhasználva egy iteratív folyamatban felépíthetünk egy biológiai tudásbázist. A mérések alapján az elméleti biológusok pontosabb modelleket specifikálhatnak és szimulációs módszerekkel a rendszer várható viselkedése jósolható. Ezek a szimulációk úgy tekinthetők mint virtuális mérések és összehasonlíthatók a kísérleti adatokkal, majd a modell vagy megerősítést nyer, vagy elvetésre kerül. Egy közvetlenebb megközelítés a biológiai kísérletek modell alapú tervezése azzal a céllal, hogy maximalizáljuk az eredményekből nyerhető információk mennyiségét. Úgy tekinthetünk a modellekre mint közös nyelvre a kísérleti és az elméleti kutatók között, mely lehetővé teszi a biológiai adat és az elmélet közvetlen kapcsolatát [1].

Első lépésként egy formális modellt alkotunk meg a biológiai tudás alapján. A modell egzakt módon specifikálja a biológiai rendszerről meglévő hipotéziseinket, és csak biológiai feltételezéseket tartalmaz. Ez a modellezési szint ideális a tudományos társadalmon belüli, valamint eltérő módszerekre építő szoftverek közötti tudáscserére. Hogy szimulációkat végezhessünk finomítanunk kell a modellünket a számítási kerettől függő feltételezésekkel. Néhány esetben ez a finomítás automatizálható, de a feltételezések elfogadása minden esetben modellezési döntés eredménye kell, hogy legyen. Például, ha folytonos változóként kezelünk koncentrációkat, az eredményünk helyes lesz abban az esetben, ha egy nagy térfogatban lejátszódó reakciót szimulálunk, de helytelen eredményre vezet extrém kis térfogatok esetében, például egy mitokondrium esetében, ahol a reagáló részecskék diszkrét volta nagy jelentőséget kap.

11.1. Biokémiai fogalmak, ezek számításos reprezentációi

A biokémiai modellek alapvető építőelemei a reakciók. A reakciókat szubsztrátjaikkal, termékeikkel, sztöchiometrikus tényezőikkel és sebességi állandóikkal specifikálhatjuk, pl.:

$$n_a A + n_b B \rightarrow n_c C + n_d D : k.$$

A sztöchiometrikus tényező (n_x) megadja a reaktáns vagy termék relatív mennyiségét, tehát definiálja a reakció struktúráját. A sebességi állandó k azt a gyakoriságot fejezi ki, amivel a reaktáns molekulák – n_a db A és n_b db B – kellő energiával összeütköznek, hogy a termékek képződhessenek. A reakció aktuális sebessége – a fluxus – a reaktánsok koncentrációinak szorzatával arányos, figyelembe véve a sztöchiometriai konstansokat is:

$$J([A],[B]) = k[A]^{n_a}[B]^{n_b},$$

ahol $[A]$ jelöli A koncentrációját, általában mol/L egységben.

Szigorúan véve minden reakció visszafordítható, és az alábbiak szerint írható le mint két irreverzibilis reakció eredője:

$$n_aA + n_bB \to n_cC + n_dD : k_1,$$
$$n_cC + n_dD \to n_aA + n_bB : k_2,$$

egyszerűbb alakban

$$n_aA + n_bB \leftrightarrow n_cC + n_dD : k_1, k_2.$$

Amikor a két fluxus megegyezik:

$$k_1[A]^{n_a}[B]^{n_b} = k_2[C]^{n_c}[D]^{n_d},$$

a rendszer egyensúlyban van, és a koncentrációkat meghatározhatjuk a fenti algebrai egyenlet átrendezésével:

$$\frac{[C]^{n_c}[D]^{n_d}}{[A]^{n_a}[B]^{n_b}} = \frac{k_1}{k_2} = K_{eq}.$$

Ha $k_2 \to 0$ – a reakció irreverzibilis –, az egyensúlyt akkor érjük el, ha a kiindulási anyagok elfogynak.

Mikroszkopikus szinten részecske-számokat használunk moláris koncentrációk helyett, és a reakció sebességét hazárd függvények formájában fejezzük ki: $h_i(x)$, ahol x a rendszer állapotát jelöli — a részecskeszám-vektor. Annak a valószínűsége, hogy az i-edik reakció megtörténik dt időintervallum alatt: $h_i(x)dt$. Ha az A vegyület koncentrációja $[A]$ egy V térfogatú kompartmentben, a részecskeszám $n_A[A]V$, ahol n_A az Avogadro-féle szám.

Ha az i-edik reakció elsőrendű kinetikát követ és a j-edik vegyület a reakció szubsztrátja, a hazárd függvény az alábbi alakú:

$$h_i(x) = x_j c.$$

Bimolekuláris reakció esetén a hazárd függvény alakja

$$h_i(x) = \begin{cases} x_j x_k c_i, & \text{ha } j \neq k, \\ \frac{x_j(x_j-1)}{2} c_i & \text{egyébként.} \end{cases}$$

Könnyen látható, hogy a k makroszkopikus sebességi állandó és a c sztochasztikus sebességi állandó közötti konverzió függ a konkrét reakció rendűségétől [2, 3].

Természetes módon specifikálhatunk például egy konstans befelé irányuló fluxust a rendszerbe a mikroszkopikus szintű modellben, de a koncentráció változás mértéke a kompartment térfogatától függ, tehát a folytonos modellben k térfogatfüggő:

$$k = \frac{c}{n_A V}.$$

Elsőrendű reakciók esetén k és c mindig egyenlő, mivel cdt dimenzió nélküli mennyiség: azon szubsztrát relatív mennyisége, amely átalakul dt idő alatt.

Magasabb rendű reakciók esetén c fordítottan arányos V-vel, mert az intermolekuláris ütközés valószínűsége koncentrációfüggő. Például egy másodrendű reakció sebességi állandóira igaz, ha a két szubsztrát eltérő, hogy:

$$c = \frac{k}{n_A V},$$

és ha csak egy szubsztrát van, melyhez tartozó sztöchiometrikus konstans 2:

$$c = 2\frac{k}{n_A V}.$$

Továbbá érdemes még megemlíteni, hogy a fentiekből következően az alábbi két rendszer nem egyenértékű kinetikai értelemben:

$$A + B + C \to ABC : k,$$
$$A + B \to AB : k_1; AB + C \to ABC : k_2.$$

Csatolt biokémiai reakciók rendszerei általában komplex hálózatos struktúrával rendelkeznek, és természetes megközelítés, hogy gráfokként ábrázoljuk őket. Nincs felső korlátja azon reakciók számának, amelyekben egy konkrét vegyület részt vehet, tehát a vegyületeket csomópontokként kell formalizálnunk. Ugyanakkor egy kémiai reakciónak több mint egy kiindulási anyaga és/vagy terméke lehet, tehát a hálózat hiperéleket tartalmaz. Másik lehetőség, hogy a reakciókat is csomópontokként formalizáljuk, és definiálunk egy $V(S, R, E)$ címkézett irányított páros gráfot, ahol egy irányított él fut $s \in S$ vegyület-csomópontból $r \in R$ reakció-csomópontba akkor és csak akkor, ha s szubsztrátja r-nek, vagy egy irányított él fut r-ből s-be akkor és csak akkor, ha s terméke r-nek. Minden élre egy $E \to \mathbb{N}$ címkézés sztöchiometrikus konstansokat definiál az adott reakcióban. Ez a gráf formalizálja a rendszer kvalitatív struktúráját. Egy címkézés szintén definiált a vegyület-csomópontokon – melyet marking-nak (jelölés) nevezünk –, és a vegyületek részecskeszámait definiálják. Ezt a fajta páros gráfot Petri-hálónak nevezzük, és részletes elmélete van. A Petri-hálók esetében használt terminológia S-et a hely-halmaznak nevezi (és P-vel jelöli), R-et az átmenetek halmazának (és T-vel jelöli).

Most már definiálhatunk egy S ún. sztöchiometrikus mátrixot, ahol s_{ij} a részecskeszámváltozás i vegyület esetében, amikor a j-edik reakció megtörténik, tehát a mátrix elemei a reakció előjeles sztöchiometriás konstansai: ha i kiindulási anyag, az előjel negatív, ha i termék, az előjel pozitív. A reakció megtörténtét az átmenet tüzelésének hívjuk a Petri-hálók terminológiájában.

Legyen M_0 kezdeti állapot, és r a megtörtént reakciók vektora, ekkor a rendszer új állapota

$$M' = M_0 + Sr.$$

Az S mátrix vizsgálata érdekes információkkal szolgálhat a rendszer struktúrájáról. Vizsgáljuk meg S mátrix magterét, azon x vektorok által kifeszített teret, melyek megoldásai az alábbi egyenletnek:

$$Sx = 0.$$

Vagy intuitív definícióval keressük az összes olyan reakció-szekvenciát, amely visszaviszi a rendszert eredeti állapotába. Ha x egy megoldása a fenti egyenletnek, akkor T-invariánsa a Petri-hálónak, azaz elemi módusa a biokémiai útvonalnak.

Most vizsgáljuk meg S transzponáltjának, $S^\top$ mátrixnak a magterét:

$$S^\top y = yS = 0.$$

A fenti egyenlet megoldásait P-invariánsnak nevezzük, ezek alkotják a rendszer megmaradási törvényeit.

11.2. Modellezés differenciálegyenletekkel

A koncentráció-változás dt idő alatt Jdt, tehát egy egyváltozós differenciálegyenlet írható fel minden vegyület koncentrációjára (állapotváltozók):

$$\frac{d[A]}{dt} = \frac{d[B]}{dt} = -k_1[A]^{n_A}[B]^{n_B} + k_2[C]^{n_C}[D]^{n_D},$$
$$\frac{d[C]}{dt} = \frac{d[D]}{dt} = k_1[A]^{n_A}[B]^{n_B} - k_2[C]^{n_C}[D]^{n_D}.$$

A fenti differenciálegyenlet-rendszer egyszerűen megoldható, és a rendszer dinamikus viselkedése vizsgálható. Az egyensúlyi állapot meghatározásához egy algebrai egyenletet kell megoldanunk, ahol minden derivált nulla:

$$0 = -k_1[A]^{n_A}[B]^{n_B} + k_2[C]^{n_C}[D]^{n_D},$$

amely ugyanannak a dinamikus egyensúlynak felel meg, amelyet az egyensúlyi konstans származtatásánál már tárgyaltunk.

A fenti differenciálegyenlet-rendszer az alábbi általános vektoriális alakban írható:

$$\frac{dv}{dt} = f(v),$$

ahol v az állapotváltozók vektora, jelen esetben a koncentrációké. Vagy az alábbi alakban:

$$\frac{dv}{dt} = SJ(v),$$

ahol S a sztöchiometrikus mátrix, és J a reakció fluxusok vektora.

A módszer mögött az az implicit feltételezés áll, hogy a koncentrációkat folytonos változóként kezelhetjük.

11.3. Sztochasztikus modellezés

A sejtszintű folyamatokban néha igen kis anyagmennyiségek vesznek részt, ezért az ezekben rejlő alapvető kvantáltság relevánssá válik. Ilyen például, ha a rendszerben lévő molekulák száma néhány száznál kevesebb.

Ebben az esetben a rendszer állapotváltozóit egész értékű részecskeszámokkal szimuláljuk koncentrációk helyett. A reakciót egy valószínűségi eseményként definiáljuk, ahol a molekuláris ütközés valószínűsége $h_i(x, c_i)dt$ arányos a kiindulási anyagok részecskeszámainak szorzatával. Ez a fajta modell Monte Carlo-módszerekkel szimulálható. A legnyilvánvalóbb módja a rendszer szimulációjának, ha diszkrét időléptékeket használunk, és egy generált véletlen szám alapján döntünk, hogy ütközés történt-e vagy sem. Amennyiben történt, módosítjuk az állapotváltozókat a reakciónak megfelelően.

Ez az eljárás számításintenzív, és csak közelítése a folytonos idejű Markov-láncnak. Ha egzakt módon szeretnénk eljárni és olyan alacsony időléptéket választunk, hogy minden lépésben maximum egy reakció történhessen, az algoritmus pazarló lesz, mivel számos időlépést szimulálunk, amikor semmi sem történik.

Megmutatható, hogy egy adott időintervallumban történő reakciók száma Poisson eloszlást követ, és két esemény közötti idő eloszlásfüggvénye is analitikus alakban írható: az időkülönbségek exponenciális eloszlást követnek. Ez adja az alapötletét a Gillespie-algoritmusnak: ahelyett, hogy számos diszkrét időlépésben kiszámítanánk a rendszer állapotát, kiszámíthatjuk a következő reakció időpontját, majd szimuláljuk azt [2].

1. Inicializálás: $t = 0$; $n = 0$; $x = x_0$.
2. Számítsuk ki: $h_i(x, c_i)\, i = 1..M$; $h_0 = \sum_{i=1}^{M} h_i(x, c_i)$.
3. Véletlen számot generálunk: $r_1, r_2 \sim U(0, 1)$.
4. Számítsuk ki: $\tau = \frac{1}{h_0} \ln \frac{1}{r_1}$.
5. Határozzuk meg μ-t amelyre $\sum_{v=1}^{\mu-1} h_i < r_2 h_0 \leq \sum_{v=1}^{\mu} h_i$.
6. Alkalmazzuk R_μ reakciós szabályt; $n = n + 1$; $t = t + \tau$.
7. Ha $t < T_{max}$: vissza 2-re.

Az inicializálást követően a hazárdokat kiszámítjuk a rendszer jelenlegi állapota alapján. Ezután mintavételezzük a következő reakció időpontját, és annak típusát az inverz eloszlások módszere szerint (3.-5. lépés) A 6. lépésben a megfelelő reakciós szabályt alkalmazzuk, tehát a megfelelő számú reaktánst eltávolítjuk, és a terméket hozzáadjuk az állapotvektorhoz.

Ahelyett, hogy a következő reakció bekövetkezésének idejét mintavételeznénk, meghatározhatjuk minden reakcióra a következő bekövetkezés időpontját a rendszer jelenlegi állapota mellett, majd a legközelebbit választjuk ki. Első ránézésre ez a módszer kevésbé hatékony, mert minden lépésben, minden reakcióhoz egy külön véletlen szám generálását

igényli. A gyakorlatban két esetben is gyorsítást érhetünk el. Ha a reakció hazárdja nem változott az előző lépés óta, a reakció következő bekövetkezési ideje továbbra is érvényes. Ha a hazard a korábbi h_i értékről h_i'-re változott, az előzőleg mintavételezett bekövetkezési időig hátralévő intervallum újraskálázható:

$$\Delta t_i' = \frac{h_i}{h_i'} \Delta t_i.$$

Ez az alapötlete a Gibson–Bruck-algoritmusnak, mely egy hatékony alternatívája a Gillespie-eljárásnak.

11.4. Hibrid módszerek

Számos közbenső lehetőség létezik a módszerkiválasztás megkönnyítésére. Egy rendszerben, ahol a reaktánsok mennyisége alacsony, a kompartmentek kicsik, a reakciók sztochasztikus természetét kezelni kell a szimulációban. Ugyanakkor a sztochasztikus szimuláció, még egy szofisztikált algoritmus használata esetén is sokkal erőforrás-igényesebb, mint egy differenciálegyenletek megoldására építő módszer. Kompromisszumot kell kötnünk tehát a pontosság és a kezelhető modell maximális komplexitása között. Egy átmeneti vagy hibrid módszer segíthet, hogy jó kompromisszumot köthessünk.

A matematikában, fizikában és közgazdaságtanban széles körben használt klasszikus módszerek használhatók a problémák sztochasztikus, de folytonos közelítésére. Intuitív származtatásukhoz használjuk fel, hogy:

$$\lim_{\lambda \to \infty} Po(\lambda) \sim N(\lambda, \lambda),$$

tehát diszkrét sztochasztikus szimuláció helyett megoldhatunk egy sztochasztikus differenciálegyenlet (SDE) formájában felírt folytonos közelítést, a folyamat Langevin-egyenletét.

$$\frac{dv}{dt} = f(v) + n(t),$$

ahol n egy zajtag, az egyenletet sztenderd technikákkal megoldhatjuk. Általános vektoriális alakjában egy SDE az alábbiak szerint írható:

$$\frac{dX}{dt} = \mu(X) + \sigma(X) dW,$$

ahol W a Wiener-folyamatot jelöli, melynek definíciója:

$$W(0) = 0, \qquad W(t+\tau) - W(t) \sim N(0, \tau),$$

és minden nem átfedő inkremens egymástól független véletlen változó.

A legegyszerűbb numerikus eljárás SDE-k megoldására az Euler-módszer általánosításának tekinthető Euler–Maruyama-módszer:

$$X_{n+1} = X_n + \mu(X_n)\Delta t + \sigma(X_n)\Delta W_n, \quad \text{ahol} \quad \Delta W_n \sim N(0, \Delta t).$$

Egy másik lehetőség, hogy kiszámítjuk a valószínűségi sűrűségfüggvény időbeni viselkedését oly módon, hogy származtatjuk a fenti Langevin-egyenlethez tartozó „Kolmogorov's forward" egyenletet:

$$\frac{\partial}{\partial t}p(x,t) = -\sum_{i=1}^{k}\frac{\partial}{\partial x_i}\{\mu_i(x)p(x,t)\} + \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k}\frac{\partial^2}{\partial x_i \partial x_j}\{\beta_{i,j}(x)p(x,t)\}.$$

Ezt Fokker–Planck-egyenletnek nevezzük.

Egy további lehetőség hibrid eljárások származtatására, ha a rendszer változóinak egy halmazát diszkrétként kezeljük, a többit folytonosként. Ebben az esetben kezelnünk kell a rendszer állapotának folytonos változását két szimulációs lépés között, tehát a Poisson folyamatunk inhomogén lesz.

11.5. Reakció–diffúzió-rendszerek

Minden eddig tárgyalt megközelítés feltételezi, hogy a vizsgált rendszer jól keveredő, a vegyületek koncentrációi és ütközési valószínűségeik azonosak a rendszer minden részében. Ha ezek a feltételezések legalább közelítőleg helytállóak, minden reakciót úgy kezelhetünk, mintha a tér azonos pontján játszódnának le. Egy sejtben azonban a reakciók jól lokalizáltak, és ez a lokalizáció elengedhetetlen a komplex szabályozási mechanizmusok működéséhez. Ebben az esetben tehát az idő mellett a térbeli koordinátákat is be kell vezetni mint változókat. A térbeli transzport-folyamat formalizálása immár elengedhetetlen, és a legegyszerűbb ilyen folyamat a diffúzió. A diffúzió egy statisztikai természetű spontán folyamat. A részecskék Brown mozgása folyamatos keveredést vált ki a rendszerben. Az egyedi molekulák szintjéről nézve egy részecske véletlen bolyongást végez a térben. Egy részecske távolsága a kiindulási helyétől várható értékben $\sqrt{N}\lambda$ ahol N az ütközések száma és λ az átlagos szabad úthossz.

Populációs szinten egy kicsi i-edik dx térrészben n_i a részecskék száma. Egy rövid időszelet alatt annak a valószínűsége, hogy a részecske átlép egy térrész-határt: p, tehát ha az i-edik térrészben a lokális koncentráció nagyobb, mint a szomszédos térrészekben, a térrészből kilépő részecskék várható száma nagyobb, mint az oda belépők várható száma.

A lineáris két dimenziós esetet tekintve annak a valószínűsége, hogy a részecske átlép egy konkrét határt, 0,5 tehát

$$\Delta n_i^k = \frac{1}{2}n_{i+1}^k - n_i^k + \frac{1}{2}n_{i-1}^k.$$

Véve a térrész méretének határértékét nullában, az alábbi differenciálegyenlethez jutunk, melyet diffúziós egyenletnek nevezünk:

$$\frac{\partial C(x,t)}{\partial t} = D\frac{\partial^2 C(x,t)}{\partial^2 x},$$

ahol D a diffúziós konstans [4]. A molekuláris fluxus arányos a koncentráció gradiensével:

$$J_F(x,t) = -D\frac{\partial C(x,t)}{\partial x}.$$

A fenti két egyenletből a makroszkopikus Fick-egyenlethez jutunk:

$$\frac{\partial C(x,t)}{\partial t} = \frac{\partial J_F(x,t)}{\partial x}.$$

A fenti egyenletek egy dimenzióban vannak megadva, de egyszerűen származtathatók háromdimenziós megfelelőik is. A reakciók által alkotott differenciálegyenlet-rendszerrel kombinálva megkapjuk a reakció–diffúzió-rendszert reprezentáló parciális differenciálegyenlet-rendszert:

$$\frac{\partial C_i(r,t)}{\partial t} = f(C_1, C_2, .., C_N) + D_i \nabla^2 C_i(r,t).$$

Mikor megoldjuk ezeket az egyenleteket, a peremfeltételeknek, úgy-mint a sejtek térbeli alakjának nagy hatása van a megoldás alakjára. A reakciókinetika és a diffúzió összjátéka kifejezetten komplex mintázatokat hozhat létre, ha a két folyamat hasonló időskálán játszódik le. Ezeket gyakran Turing-mintázatoknak nevezzük, mert Alan Turing „The Chemical Basis of Morphogenesis" című híressé vált publikációjában tárgyalja a jelenséget [5]. A cikkben reakció–diffúzió-egyenleteket alkalmazott modell-rendszereken, és a megoldások tulajdonságait vizsgálta.

Az élővilágban számos példa található olyan motívumokra, melyek erősen emlékeztetnek a Turing-mintázatokra. Láthatóak például állatok szőrzetén, mint például a cirmos macskák csíkjai vagy a leopárd foltjai.

11.6. Modell-illesztés

Az alapvető kapcsolatot a modell és a kísérlet között az adat testesíti meg. A modell paraméterei a kísérleti adatok segítségével határozhatók meg, a modell-illesztésre gépi tanulási módszereket használunk. A differenciálegyenletes módszer esetében az $f(v)$ függvény meghatározása a modell-illesztés célja. Erre a célra tetszőleges regressziós módszert használhatunk.

Sztochasztikus szimuláció esetén a modell-illesztés sokkal nehezebb feladat és jelenleg is aktív kutatás tárgyát képezi. Az a feltételezés, hogy minden reakció bekövetkezésének pontos időpontjával rendelkezünk, irreális, tehát a sztochasztikus modell-tanulás kontextusában a hiányos adat kezelésének problematikájával találjuk magunkat szembe. Úgynevezett Markov-lánc Monte Carlo-módszereket használhatunk a sztochasztikus modellek Bayes-i paraméterbecslésére [6]. Egy adat-imputációt tartalmazó mintavételezési sémát használhatunk, hogy meghatározzuk a modell-paraméterek *a posteriori* eloszlását a hiányos megfigyelések ismeretében.

Egy alternatív megközelítés, hogy a paramétertanulást a sztochasztikus modell egy folytonos normális eloszlású közelítésén hajtjuk végre. Ez a modell szintén igényel imputációt, mivel általában nem áll rendelkezésünkre elég sűrűn minta, hogy közvetlenül alkalmazhassuk a sztochasztikus differenciálegyenlet Euler–Moruyama-közelítését [7].

11.7. Teljes-sejt-szimuláció

Egy olyan komplex biológiai rendszernek, mint egy teljes sejtnek a megértése több szinten történik. Amikor egy organizmus teljes genomját szekvenálják, egyértelmű, hogy a rejtélyek nagy része még megoldatlan. Mikor minden gént annotálnak, a géntermékeket azonosítják, a szerkezetüket meghatározzák, még mindig számos nyitott kérdés marad. A tudás egy következő szintjét a géntermékek funkciója és a közöttük lévő komplex kölcsönhatások képezik. Továbbá fennállnak kölcsönhatások a géntermékek és a kromatin-struktúra között is. A kölcsönhatás lehet közvetlen vagy közvetett, melyet közös metabolitok rendeznek biokémiai útvonalakba. Ha meg tudjuk rajzolni ezt a térképet, és az organizmus teljes metabolomját ismerjük, még mindig van a tudásnak egy fennmaradó szintje: a sejt dinamikus viselkedése [8]. Ezt a szintet tekinthetjük az organizmus legmagasabb szintű fenotípusának, ha figyelmen kívül hagyjuk a környezetet. Az egyetlen megvalósítható módja, hogy a sejt dinamikus viselkedését tanulmányozzuk, az *in silico* szimuláció.

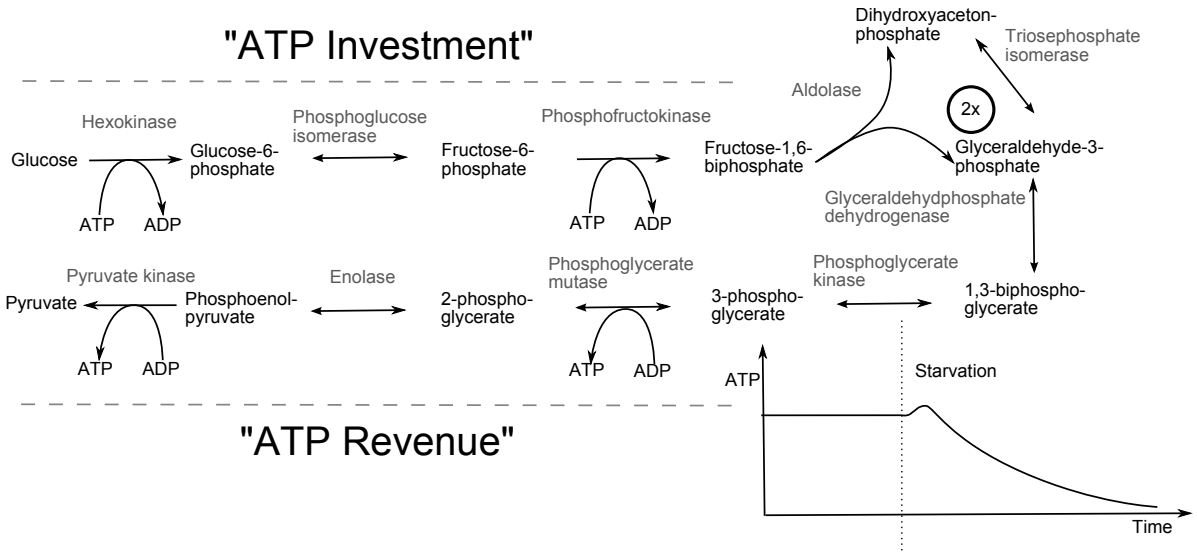
Az elvárásunk egy modelltől valamiféle alapvetően új előrejelzés. Ezeknek az előrejelzéseknek két eltérő nézőpontját nevezték találóan Freddolino és munkatársai a fizikus nézőpontjának és a mérnök nézőpontjának [9]. Az első típus egy széles körben alkalmazható rendezőelv, amely segítheti a rendszerről való tudományos gondolkodást, a második típus egy praktikusabb, általában kvantitatív becslés, mely valamely mérnöki feladatban lehet hasznos, például hatóanyag szűrésben.

A *Mycoplasma genitalium* nevű patogén mikroba rendelkezik a legkisebb genommal minden ismert organizmus között: 525 azonosított génje és 580kb hosszú genomja van. Nem meglepő tehát, hogy a teljes sejt szimulációra tett első kísérletek az *M. genitalium*ot használták modellorganizmusként. Mivel még ez az organizmus is relatíve nagy számú génnel rendelkezik, valamint a génkiütéses vizsgálatok megmutatták, hogy nem minden gén esszenciális a mikroorganizmus túléléséhez, lehetséges egy minimális génhalmaz – egy minimális genom – kiválasztása. Azt a mesterséges sejtet, mely ezt a genomot tartalmazza, minimális önfenntartó sejtnek (angolul self-surviving cell, SSC) nevezzük.

Az E-CELL modell (127 gén, 495 reakciós szabály) glukózt fogyaszt a környezetéből és laktátot termel mint anyacseréjének végtermékét [10]. Ez a triviális viselkedés *in silico* szimuláció nélkül is megjósolható, de ez az egyszerű modell is képes néhány érdekes jelenség előrejelzésére.

Ha a környezeti glukóz-szint eléri a nullát, a sejt éhezni kezd. Paradox módon a modellek azt jósolják, hogy az éhezés nagyon korai szakaszában az ATP-szint ideiglenesen emelkedik, majd később esni kezd mindaddig, míg az ATP-készletek kimerülnek (11.1. ábra) [8, 9].

Ez a fajta szimuláció hatékonyan használható fel patológiás állapotok vagy egyéni különbségek modellezésére, hogy személyre szabott beavatkozásokat választhassunk ki. Egy teljes értékű humán sejt modellezése még nem elérhető, de humán eritrocita modellek már léteznek. Ezek a modellek lehetővé teszik bizonyos fajta örökletes anémiák vizsgálatát [8].



11.1. ábra. A glikolízis első felében két ATP/glukóz-molekula befektetésre van szükség, a második felében pedig 2×2 ATP nyereség realizálható, tehát a nettó nyereség 2 ATP/glukóz. A második részben a fluxus kétszerese az elsőének (lásd a „2×" jelet az összefutó reakcióknál)

11.8. Áttekintés

Ebben a fejezetben bemutattuk a dinamikus modellezés fontosságát, és áttekintettünk néhány számítási eljárást ennek végrehajtásához. Ezek az eljárások leginkább a vizsgált rendszerre vonatkozó alapvető feltevéseikben különböznek. A tárgyalt keretrendszerek csoportosításához lásd a 11.1. táblázatot. A reakció–diffúzió-rendszerek sztochasztikus kezelésének lehetőségével jelen fejezetben nem foglalkoztunk.

11.1. táblázat. Keretrendszerek kulcsszavakban

	Determinisztikus	Sztochasztikus	
	Folytonos	Diszkrét	Folytonos
Homogén	Differenciálegyenletek	Poisson folyamatok, Gillespie algoritmus, Gibson–Bruck-algoritmus	SDE, Langevin-egyenlet, Fokker–Planck-egyenlet
Heterogén	Parciális differenciál-egyenletek	*nem tárgyaltuk*	*nem tárgyaltuk*

Irodalomjegyzék

[1] J. M. Bower and H. Bolouri, *Computational Modeling of Genetic and Biochemical Networks*. Bradford Books, MIT Press, 2001.

[2] D. T. Gillespie, Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[3] D. J. Wilkinson, *Stochastic modelling for systems biology*, Chapter Chemical and biochemical kinetics. Chapman and Hall/CRC mathematical and computational biology series, [11], Chapman & Hall/CRC, Boca Raton, Fla., 2006.

[4] G. Bormann, F. Brosens, and E. De Schutter, *Computational Modeling of Genetic and Biochemical Networks*, Chapter Diffusion. Bradford Books, MIT Press, [1], 2001.

[5] A. M. Turing, The Chemical Basis of Morphogenesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 237(641):37–72, Aug. 1952.

[6] R. J. Boys, D. J. Wilkinson, and T. B. L. Kirkwood, Bayesian inference for a discretely observed stochastic kinetic model. *Statistics and Computing*, 18(2):125–135, 2008.

[7] Andrew Golightly and Darren J. Wilkinson, Bayesian sequential inference for stochastic kinetic biochemical network models. *Journal of Computational Biology*, 13(3):838–851, 2006.

[8] M. Tomita, Whole-cell simulation: a grand challenge of the 21st century. *TRENDS in Biotechnology*, 19(6):205–210, 2001.

[9] P. L. Freddolino and S. Tavazoie, The dawn of virtual cell biology. *Cell*, 150(2):248–250, July 2012.

[10] M. Tomita, K. Hashimoto, K. Takahashi, T. S. Shimizu, Y. Matsuzaki, F. Miyoshi, K. Saito, S. Tanida, K. Yugi, J. C. Venter, and C. A. Hutchison, E-CELL: software environment for whole-cell simulation. *Bioinformatics*, 15(1):72–84, 1999.

[11] D. J. Wilkinson, *Stochastic modelling for systems biology*. Chapman and Hall/CRC mathematical and computational biology series, Chapman & Hall/CRC, Boca Raton, Fla., 2006.

12. fejezet

Oksági következtetések az orvosbiológiában

Ebben a fejezetben összefoglaljuk az elméleti hátterét és megközelítési módját olyan induktív következtetési eljárásoknak, amelyek egy tárgyterület összes vagy egy célváltozót közvetlenül érintő oksági relációjának a feltérképezését segítik. A megközelítés alapja a relációk létezésének jellemzése, amire a Bayes-statisztikai keretrendszer felhasználását mutatjuk be. Bemutatjuk a posztgenomikai korszak azon változásait is, amelyek indokolják ezt a megközelítést, és bemutatjuk a módszer jelenlegi határait, nyitott kérdéseit.

Jelölések*

x,$\underline{x}$,$\underline{\underline{x}}$	skalár, (oszlop)vektor vagy halmaz, mátrix
$X, x, p(X)$	véletlen változó X, érték x, valószínűségi tömegfüggvény/sűrűségfüggvény X
$\mathrm{E}_{X,p(X)}[f(X)]$	$f(X)$ várható értéke $p(X)$ szerint
$\mathrm{var}_{p(X)}[f(X)]$	$f(X)$ varianciája $p(X)$ szerint
$I_p(\underline{X}\|\underline{Z}\|\underline{Y})$	$\underline{X}$ és $\underline{Y}$ megfigyelési függetlensége $\underline{Z}$ feltétellel p esetében
$(X \perp\!\!\!\perp Y \| Z)_p$	$I_p(\underline{X}\|\underline{Z}\|\underline{Y})$
$(X \not\perp\!\!\!\perp Y \| Z)_p)$	$\neg I_p(\underline{X}\|\underline{Z}\|\underline{Y})$
$CI_p(\underline{X};\underline{Y}\|\underline{Z})$	$\underline{X}$ és $\underline{Y}$ beavatkozási függetlensége $\underline{Z}$ feltétellel p esetében
$\prec$	(részleges) sorrendezés
$\prec^c$	a változók egy teljes sorrendezése
$\prec^G$	adott G irányított körmentes gráffal kompatibilis sorrendek halmaza
$\prec(n)$	n objektum sorrendjeinek (permutációinak) a halmaza
$G, \underline{\theta}$	Bayes-háló struktúrája és paraméterei
$G^{\sim}$	G irányított körmentes gráf esszenciális gráfja
$\mathcal{G}(n)/\mathcal{G}^k(n)$	n csomópontú maximum k szülőjű DAG-ok halmaza
$\mathcal{G}^{\prec}$	adott $\prec$ sorrenddel kompatibilis DAG-ok halmaza

*További konvenciók az egyes fejezetekben jelöltek.

$\mathcal{G}^G$	adott G DAG-gal megfigyelési ekvivalens DAG-ok halmaza
$\sim$	kompatibilitási reláció
$\mathrm{pa}(X_i, G) \sim \prec$	$\mathrm{pa}(X_i, G)$ szülői halmaz kompatibilis $\prec$ sorrendezéssel
$\mathrm{MB}_p(X_i)$	Markov-takarója X_i-nek p-ben
$\mathrm{pa}, \mathrm{pa}(X_i, G)$	szülői változók halmaza, X_i szüleinek halmaza G-ben
pa_{ij}	a j. konfigurációja a szülői értékeknek egy sorrendben
$\mathrm{bd}(X_i, G)$	X_i szüleinek, gyerekeinek és gyerekei egyéb szüleinek halmaza G-ben
$\mathrm{MBG}(X_i, G)$	a Markov-takaró algráfja X_i-nek G-ben
$\mathrm{MBM}(X_i, X_j, G)$	a Markov-takaróbeliség relációja
n	valószínűségi változók száma
k	maximális szülőszám DAG-okban
N	mintaszám
V	összes valószínűségi változók száma
Y	válasz, kimeneteli, függő változó
$N_+/N_{\ldots,+,\ldots}$	$N_i/N_{\ldots,i,\ldots}$ megfelelő összegei
$D\|X$	X változóhalmazra szűkített adathalmaz
$\|\|$	kardinalitás
$1()$	indikátorfüggvény
f', f''	f függvény első és második deriváltjai
A^T	A mátrix transzponáltja
$\underline{x} \cdot \underline{y}$	$\underline{x}$ és $\underline{y}$ vektorok skalárszorzata
ξ^+/ξ^-	informatív/nem informatív információs kontextus
$\neg, \wedge, \vee, \neq, \rightarrow$	standard logikai operátorok
$\cap, \cup, \setminus, \Delta$	standard halmazműveletek
$KB \vdash_i \alpha$	α bizonyíthatósága KB-ből
Γ	a Gamma függvény
$\mathrm{Beta}(x\|\alpha, \beta)$	a Béta eloszlás sűrűségfüggvénye (pdf)
$\mathrm{Dir}(x\|\underline{\alpha})$	a Dirichlet-eloszlás sűrűségfüggvénye
$\mathrm{N}(x\|\mu, \sigma)$	az egyváltozós normál eloszlás sűrűségfüggvénye
$\mathrm{N}(x\|\underline{\mu}, \underline{\underline{\Sigma}})$	a többváltozós normál eloszlás sűrűségfüggvénye
BD,BD_e	Bayesian Dirichlet-prior, megfigyelési ekvivalens BD-prior
BD_{CH}	Bayesian Dirichlet (BD) prior 1 hiperparaméterekkel
$\mathrm{BD}_{\mathrm{eu}}$	megfigyelési ekvivalens és uniform BD prior
$L(\underline{\theta}; D_N)$	$p(D_N\|\underline{\theta})$ likelihood függvénye
$H(X, Y)$	X és Y entrópiája
$I(X; Y)$	X és Y kölcsönös információja
$\mathrm{KL}(X\|Y)$	X és Y Kullback–Leibler-divergenciája
$H(X\|Y)$	X és Y keresztentrópiája
$L_1(,)$, $L_2(,)$	az abszolútértékbeli (Manhattan) négyzetes (euklidészi) távolságok
$L_0(,)$	0-1 veszteség
$\mathcal{O}()/\Theta()$	aszimptotikus, nagyságrendi felső és alsó határ

Rövidítések

ROC	Receiver Operating Characteristic (ROC) görbe
AUC	ROC-görbe alatti terület
BMA	Bayes-i modell átlagolás
BN	Bayes-háló
DAG	irányított körmentes gráf
FSS	jegykiválasztási probléma
MAP	maximum a posteriori
MI	kölcsönös információ
ML	maximum likelihood
MBG	Markov-határ gráf
MB	Markov-takaró
MBM	Markov-takaróbeliség
(MC)MC	(Markov-láncos) Monte Carlo
NBN	naiv Bayes-háló

12.1. Bevezető

Az omikai mérési technikák elterjedése lehetővé tették a hipotézismentes orvosbiológiai kutatásokat. Az omikai adatok nagy változószáma és az ehhez képesti alacsony mintaszáma egyszerű (kevés statisztikai) teszten alapuló statisztikai elemzéseket indokol, amelyek azonban a reméltől elmaradó eredményeket hoztak például a biomarker-kutatások, új gyógyszercélpontok és új klinikai végpontok felfedezésének területén is. A komplexebb modellek alkalmazására a Bayes-statisztikai keretrendszer kínál egy konzisztens, önkorrigáló lehetőséget, különösen az azon belüli Monte Carlo alapú következtetések utóbbi negyedszázadban bekövetkezett fejlődése. Ennek részben oka a számítástechnika fejlődése, illetve az ezredfordulótól megfigyelhető trendfordulása is, ami a párhuzamos számítási erőforrások fejlődését jelenti: az általános célú grafikus kártyák, elosztott „grid" rendszerek és a felhő alapú számítási közmű elterjedését. Ezen tényezők eredményeként átfogó, oksági modellek induktív strukturális vizsgálata is lehetővé vált. Az oksági kutatásoknak ez az ága különösen relevánssá vált az omikai megközelítés miatt, amely vizsgálatot követhetnek más típusú oksági következtetések, mint például az adott oksági modellen belüli hatáserősség identifikálásának és becslésének a kérdései, illetve funkcionális oksági modelleken alapuló vizsgálatai kontrafaktuális jellegű következtetéseknek. Az oksági relációk rendszerszintű vizsgálatát a Bayes-statisztikai keretben mutatjuk be, amelyhez elsőként összefoglaljuk a passzív megfigyelésekből történő tanulás elméleti korlátait, és bemutatunk olyan idealisztikus tanulási algoritmusokat, amelyek aszimptotikus mennyiségű adatot tételeznek fel. Ezt követően bemutatjuk egy elterjedt poszterior származtatását az oksági modellekhez, amely képes oksági priorokat és oksági (beavatkozásokat is tartalmazó) adatokat is integrálni. Végül bemutatunk olyan strukturális modelltulajdonságok feletti Bayes-következtetést, amely modelltulajdonságok sokrétű oksági értelmezéssel bírnak.

Az oksági relációk tanulásával kapcsolatos kihívások illusztrálására érdemes felidézni, hogy egy okozati reláció

1. inkább a determinisztikus és nem bizonytalan világképhez tartozik,
2. aszimmetrikus, szemben az információs, asszociációs bizonytalansággal,
3. aktív cselekvések, beavatkozások következményeihez kapcsolódik, és nem passzív megfigyelésekhez,
4. mechanizmusokhoz kapcsolódik, amelyek autonómok, modulárisak az őket terhelő zajok és a beavatkozások viszonylatában,
5. időaspektussal is rendelkezik.

A bizonytalanság modellezésében az asszociációs relációk és az oksági relációk megkülönböztetésére több szempontrendszert is megfogalmaztak, ilyen például az orvosbiológiai kutatásokból származó következő lista, mely az oksági relációkkal szemben támasztott követelményeket sorolja fel [21]:

1. Erő. Erős statisztikai asszociáció.
2. Konzisztencia, specifikusság, koherencia. Például az ok megszüntetésével a hatás is szűnjön meg (szükségesség), és az ok bekövetkeztével a hatás is erősödjön (elégségesség).
3. Gradiens. Legyen a következmény arányos a hatással (dózis–hatás elv).
4. Temporalitás. X időben előzze meg Y-t.
5. Plauzibilitás és analógia. Létezzen magyarázat, és ne legyenek alternatív, zavaró tényezőre is építő alternatív magyarázatok.
6. Kísérleti adatok léte.

12.2. Függetlenségi és oksági relációk reprezentálása Bayes-hálókkal

A feltételes függetlenség fogalma központi szerepet játszik az oksági relációk tanulásának tisztázásában. Követve a Dawid [7] által bevezetett jelölést a feltételes függetlenség a következőképpen definiálható.

12.1. Definíció. Legyen $p(\underline{V})$ együttes eloszlás esetén $\underline{X}, \underline{Y}, \underline{Z} \subseteq \underline{V}$ diszjunkt részhalmazok. Jelölje $\underline{X}$ és $\underline{Y}$ $\underline{Z}$ feltétel melletti függetlenségét $I_p(\underline{X}|\underline{Z}|\underline{Y})$, azaz

$$(X \perp\!\!\!\perp Y|Z)_p \text{ } \textit{iff} \text{ } (\forall \underline{x}, \underline{y}, \underline{z} \; p(\underline{x}, \underline{y}|\underline{z}) = p(\underline{x}|\underline{z})p(\underline{y}|\underline{z}) \text{ } \textit{ha} \text{ } p(\underline{z}) > 0). \tag{12.1}$$

Az $(X \perp\!\!\!\perp Y|Z)_p$ feltételes függetlenségre egy másik jelölés az $I_p(X|Z|Y)$ és az $I_p(X;Y|Z)$. Egyértelműség esetén az alsóindexet és a feltételt elhagyjuk. A függetlenség hiányát, azaz a függést $(X \not\perp\!\!\!\perp Y|Z)_p$ jelöli.

Egy eloszlásban fennálló függetlenségek teljes rendszerét reprezentálja a következő

12.2. Definíció. Egy $P(X_1, \ldots, X_n)$ eloszlás M_P függetlenségi modellje pontosan a P-ben érvényes $I_P(X, Y|Y)$ függetlenségi állításokat tartalmazza.

Az oksági kutatásban központi szerepet játszó Bayes-hálók valószínűségi definíciójához szükséges a következő két fogalom.

12.3. Definíció. Egy G irányított, körmentes gráfban az $X, Y, Z \subseteq V$ diszjunkt csomópont halmazok esetében jelölje $I_G(X|Z|Y)$, illetve $I_G(X;Y|Z)$, ha X és Y *d-elválasztottak* Z által, azaz ha minden p út X és Y között blokkolt Z által a következőképpen:

1,2 a p út tartalmaz egy Z-beli n csomópontot nem összetartó élekkel (azaz így $\rightarrow n \rightarrow$ vagy így $\leftarrow n \rightarrow$),

3 a p út tartalmaz egy nem Z-beli n csomópontot összetartó élekkel (azaz így $\rightarrow n \leftarrow$), amelynek nincs leszármazottja Z-ben.

12.4. Definíció. A $p(X_1, \ldots, X_n)$ eloszlásra teljesül a globális Markov-feltétel G szerint, ha

$$\forall\, X, Y, Z \subseteq V :\ I_G(X;Y|Z)_G \Rightarrow (X \perp\!\!\!\perp Y|Z)_p. \tag{12.2}$$

Ekkor a Bayes-háló egy lehetséges definíciója a következő.

12.5. Definíció. A G irányított körmentes gráf a $P(V)$ eloszlás Bayes-hálója, ha minden változót a gráf egy csomópontja reprezentál, a gráfra teljesül valamelyik (és így az összes) Markov-feltétel, és a gráf minimális (azaz bármely él elhagyásával a Markov-feltétel már nem teljesül).

Míg ez a definíció egyértelműen a valószínűségi függetlenségek rendszerének reprezentációjaként tekint a Bayes-hálóra, addig a mérnöki gyakorlatban közkedvelt az alábbi, praktikus meghatározás.

12.6. Definíció. A V valószínűségi változók Bayes-hálója a (G, θ) páros, ha G egy irányított körmentes gráf, amelyben a csomópontok jelképezik V elemeit, θ pedig a csomópontokhoz tartozó $P(X_i|Pa(X_i))$ feltételes eloszlásokat leíró numerikus paraméterek összessége.

A Markov-feltétel teljesülése biztosítja, hogy minden gráfból kiolvasott függetlenség teljesüljön az eloszlásban, azonban a másik irányhoz, ahhoz tehát, hogy minden függetlenség kiolvasható is legyen a gráfból, annak stabilnak is kell lennie.

12.7. Definíció. Egy P(U) eloszlás stabil, ha létezik olyan G DAG, hogy P(U)-ban pontosan a G-ből d-szeparációval kiolvasható függések és függetlenségek teljesülnek benne (azaz G perfekt térkép).

A DAG-reprezentáció korlátját alapvetően az jelenti, hogy numerikusan a struktúra szerint nem szükségszerű függetlenségek is lekódolhatóak. A triviális redundanciákon túl ezek rejtett formákban is megjelenhetnek, például nem tranzitív függések képében vagy

alacsonyabb rendű függetlenségek képében (például egy Markov-láncban megfelelő paraméterezés mellett előfordulhat, hogy a függések nem tranzitívak).

Az eloszlás stabilitásának és szigorú pozitivitásának feltevése sem zárja ki, hogy az eloszlás függetlenségi modelljének több DAG is perfekt térképe legyen. Viszont éppen ez a DAG-okból d-szeparációval indukált közös függetlenségi modellek teszik lehetővé egy DAG-ok feletti ekvivalencia-reláció bevezetését [14, 20, 13].

12.8. Definíció. Két DAG G_1, G_2 *megfigyelési ekvivalens*, ha pontosan ugyanazokat a d-szeparációs relációkat definiálják, azaz $((X \perp\!\!\!\perp Y|Z)_{G_1}) \Leftrightarrow (X \perp\!\!\!\perp Y|Z)_{G_2}$.

Az azonos ekvivalencia-osztályba tartozó DAG-ok tulajdonságainak megértése több szempontból is fontos. Egyrészt szükséges tisztázni a DAG-ok szándékolt, intuitív oksági szemantikájának fenntarthatóságát, nevezetesen azt, hogy milyen korlátok között maradhatna érvényes ez az oksági értelmezés. Másrészt azonos megfigyelési ekvivalencia-osztályba tartozó DAG-ok Bayes-hálóit azonos módon kellene felparaméterezni, ami akauzális megközelítésben is fontos következményekhez fog vezetni. Az azonos ekvivalencia-osztályba tartozó DAG-ok jellemzése két észrevételen nyugszik. Az első, hogy az azonos megfigyelési ekvivalencia-osztályba tartozó DAG-ok irányítatlan váza azonos, mivel a DAG-ban egy él egy közvetlen függést reprezentál, amelynek minden Markov-kompatibilis DAG-ban meg kell jelennie [14]. A második észrevétel, hogy ha X, Y és Y, Z közötti közvetlen függések léteznek, úgy, hogy nincs közvetlen függés X, Z között és nincs olyan függetlenség, hogy $(X \perp\!\!\!\perp Z|\{Y, S\})$, azt mindenképpen egy összetartó élpárral kell jelezni $X \to Y \leftarrow Z$, egy úgynevezett *v-struktúrát* létrehozva. Az azonos ekvivalencia-osztályba tartozó DAG-ok jellemzését a következő tétel biztosítja.

12.1. Tétel ([14, 4])**.** *Két DAG G_1, G_2 pontosan akkor* megfigyelési ekvivalens*, ha az irányítatlan vázuk megegyezik és ugyanazon* v-struktúrákat *tartalmazzák (azaz konvergáló éleket, amelyek talpánál nincs él) [14]. Ha a Bayes-hálók $(G_1, \underline{\theta}_1)$ és $(G_2, \underline{\theta}_2)$ diszkrét változókat tartalmaznak és lokális modelljeik multinomiális eloszlások, akkor G_1, G_2 megfigyelési ekvivalenciája egyenlő dimenzionalitást és bijektív leképezhetőséget jelent a $\underline{\theta}_1$ és $\underline{\theta}_2$ paraméterezések között, amit* eloszlásbeli ekvivalenciának *neveznek [4].*

Mint látható, ha elfogadjuk az Ockham-elv által diktált modellminimalitás elvét, és egy eloszlásmodellezésnél (az egyszerűség kedvéért stabil eloszlást feltételezve) a függetlenségi modelljét minimális módon reprezentáló DAG-okat tekintjük, akkor bizonyos élek irányítása önkényes, így oksági értelmezése, a priori információk hiányában értelmetlen. Azonban a 12.1. Tételben szereplő v-struktúráknál több élre jelenthet megkötést a megfigyelési osztályba tartozás, hiszen bizonyos élek irányítása azért lehet egyértelmű, mert amúgy v-struktúrát hoznának létre (ami kivezetne az ekvivalencia-osztályból). Ez a következő definícióhoz vezet el.

12.9. Definíció. Az *esszenciális gráf* a megfigyelési ekvivalens DAG-ok halmazát reprezentálja egy részlegesen irányított DAG-gal (PDAG), amely gráfban csak azok az úgynevezett kényszerített élek irányítottak, amelyek az ekvivalenciaosztálybeli DAG-okban azonosan irányítottak. A többi él irányítatlansága az (élszintű) eldönthetetlenséget jelzi.

Az esszenciális gráf meghatározására hatékony algoritmust közölt Meek [13].

A klasszikus kérdés, hogy hogyan lehet megkülönböztetni az oksági kapcsolatokat a függésektől („korreláció versus kauzalitás"), azaz, hogy hogyan lehetne meghatározni az oksági státuszát passzívan megfigyelt X és Y közötti statisztikai függésnek, az felbontható a valószínűségi Bayes-hálós reprezentációkhoz tartozó fogalmakkal, mint stabilitás és az esszenciális gráf. Elsőként megfontolandó, hogy vajon az összes közvetlen függés oksági-e. Ez erősen vitatható feltevés volna, amelyre hosszabban kitérünk. Másodsorban a stabilitás feltevése is megfontolható, hiszen annak hiányában (a Bayes-hálós reprezentáció definíciója szerint) nem fennálló függéseket is implikálni fog a struktúra. Harmadsorban, meg lehet fontolni, hogy az esszenciális gráf és a kényszerített élek definiálásánál használt „Boolean" Ockham-elv (amely szerint csak a minimális, konzisztens modelleket vettük figyelembe) a Bayes-i kontextusban nem terjeszthető-e ki?

Ezen kérdések megfontolásához vezessük be az oksági modell fogalmát, amely a korábbi, Bayes-hálókon alapuló intuíciót formalizálja.

12.10. Definíció. Egy DAG-ot *oksági struktúrának* nevezünk változók V halmaza felett, ha minden csomópont egy változót reprezentál, az élek pedig közvetlen ráhatást szimbolizálnak. Egy *oksági modell* olyan oksági struktúra *lokális valószínűségi modellekkel* $p(X_i|\,\mathrm{pa}(X_i))$ minden egyes csomóponthoz, amely leírja az adott X_i csomópont sztochasztikus függését a $\mathrm{pa}(X_i)$ szüleitől. Mivel a feltételes modellek gyakran parametrikus modellcsaládból származnak, az X_i-hez tartozó feltételes modell paramétereit $\underline{\theta}_i$ jelöli, és $\underline{\theta}$ jelöli a teljes modell paraméterezését.

A stabilitás feltevésével az esszenciális gráf egzakt módon reprezentálja a függetlenségi relációkat, és a Boolean Ockham-elv szerinti modellminimalitásnak megfelelően maximális mértékben jelzi a potenciális oksági relációkat, így elfogadásával az oksági relációk rendszer alapú kikövetkeztetésére láthatnánk példát. A feltevések jogosságának vizsgálatához vezessük be az alábbi formális feltételt, amely egy oksági struktúra validitását és elégségességét biztosítja.

12.11. Definíció. Egy G oksági struktúra és p eloszlás teljesíti az *oksági Markov-feltételt (CMA*, ha p-ben teljesül a G szerinti lokális Markov-feltétel.

Az oksági Markov-feltétel Reichenbach „közös ok elv"-én alapul, amely szerint X és Y események közötti függés azért áll fenn, mert vagy X okozza Y-t, vagy Y okozza X-et, vagy közös ok befolyásolja X-et és Y-t is [16, 10]. Ennek megfelelően az oksági Markov-feltétel akkor áll fenn (p, G) párra, ha a V változóhalmaz *okságilag elégséges*, azaz nincs rejtett, nem V-beli, közös ok (vagy másképpen fogalmazva: minden közös ok $X, Y \in V$ párokra V-beli). Ez természetesen nem azt jelenti, hogy nem lehetnek rejtett változók, hiszen ez egy adott absztrakciós szinten elkerülhetetlen, de csak azon változóknak szükséges V-ben szerepelni, amelyek két vagy több változót is közvetlenül befolyásolnak.

Az oksági Markov-feltétel összekapcsolja az oksági relációkat és a függéseket, és az oksági modell (modellezés) elégségességét követeli meg a megfigyelt függésekhez (mondhatni úgy is, hogy az élek elégségesek). Érdemes észrevenni, hogy a stabilitás feltevése éppen az élek szükségességét jelenti (mondhatni úgy is, hogy nincsen felesleges él). Ez a két

feltevés biztosíthatja, hogy a Bayes-háló által implikált függetlenségek valóban fennállnak és a függések is egzakt módon reprezentáltak az oksági modellben [9].

Az oksági következtetések valószínűségi megközelítéséhez vezessük be a beavatkozás $do()$ műveletét a „*manipulációs tétel*" ([19]) és „gráf csonkolás" ([16]) szerint.

12.12. Definíció. Egy $G, \underline{\theta}$ oksági modell esetén $p(Y|z, do(X = x))$ jelölje azt az eloszlást, amelyet úgy kapunk, hogy a (perfekt) beavatkozáshoz tartozó X változó(k) bemenő éleit töröljük és ezeket a változókat az előírt értékre beállítjuk (azaz a faktorizációban a beállított változókhoz tartozó faktorok nem szerepelnek) [15].

A beavatkozás fogalmára támaszkodva egy ahhoz kapcsolódó függetlenség is bevezethető.

12.13. Definíció. Jelölje $p(.|do(.))$ a megfelelő beavatkozási eloszlásokat, és legyenek $\underline{X}, \underline{Y}, \underline{Z} \subseteq \underline{V}$ diszjunkt részhalmazok. Ekkor a $\underline{X}$ és $\underline{Y}$ oksági függetlensége (irrelevanciája) $\underline{Z}$ esetében $CI_p(\underline{X}; \underline{Y}|\underline{Z})$ akkor áll fenn, ha

$$CI_p(\underline{X}; \underline{Y}|\underline{Z}) \text{ } \textit{iff} \text{ } (\forall \underline{x}, \underline{y}, \underline{z} \; p(\underline{y}|do(\underline{z}), do(\underline{x})) = p(\underline{y}|do(\underline{z}))). \tag{12.3}$$

Ezen oksági függetlenséghez is tartozik gráf alapú reprezentáció.

12.2. Tétel. *Egy (G, θ) Bayes-hálóval definiált p stabil eloszlásban az irányított útlefogás egzakt módon reprezentálja az oksági irrelevanciát, azaz $int(X \perp\!\!\!\perp Y|Z)_G \Leftrightarrow (X \perp\!\!\!\perp Y|Z)_p$, $\forall \; X, Y, Z \subseteq V$), ahol $int(X \perp\!\!\!\perp Y|Z)_G$ jelöli, hogy Z minden irányított utat lefog X-ből Y-ba, azaz minden X-ből Y-ba vezető s út tartalmaz egy csomópontot Z-ben.*

12.3. Oksági relációk kényszer alapú tanulása

A kényszer alapú struktúra-tanulási algoritmusok lehetőség szerint minimális számú függetlenségi tesztet végrehajtva próbálnak olyan Bayes-háló-struktúrát találni, amely az adatokban megjelenő függetlenségi viszonyokat hűen reprezentálja [16, 10, 19] (minimális függetlenségi térkép, lásd Valószínűségi gráfos modellek fejezet). Ezekre az algoritmusokra példa az „Inductive Causation" (IC) algoritmus, amely egy stabil eloszlást tételez fel és ekkor helyes megoldást ad:

1. *Váz:* Konstruáljuk meg az irányítatlan gráfot (vázat) úgy, hogy $X, Y \in \underline{V}$ akkor legyen összekötve, ha $\forall S (X \perp\!\!\!\perp Y|S)_P$, ahol $S \subseteq \underline{V} \setminus \{X, Y\}$.
2. *v-struktúrák:* Irányítsuk $X \rightarrow Z \leftarrow Y$, ha X, Y nem szomszédosak, Z egy közös szomszéd és $\neg \exists S$ úgy, hogy $(X \perp\!\!\!\perp Y|S)_P$, ahol $S \subseteq \underline{V} \setminus \{X, Y\}$ és $Z \in S$.
3. *propagation:* Irányítsuk a maradék irányítatlan éleket úgy, hogy nem hozunk létre új v-struktúrát, sem irányított kört.

12.3. Tétel. *A következő szabályok szükségesek és elégségesek.*

R_1 *Ha* $(a \not\to c) \wedge (a \to b) \wedge (b - c)$, *akkor* $b \to c$.

R_2 *Ha* $(a \to c \to b) \wedge (a - b)$, *akkor* $a \to b$.

R_3 *Ha* $(a - b) \wedge (a - c \to b) \wedge (a - d \to b) \wedge (c \not\to d)$, *akkor* $a \to b$.

R_4 *Ha* $(a - b) \wedge (a - c \to d) \wedge (c \to d \to b) \wedge (c \not\to b) \wedge (a - d)$, *akkor* $a \to b$.

Bár stabil eloszlás esetében a módszerek aszimptotikus adatmennyiségnél azonosan viselkednek, véges adatmennyiségnél nincsen gyakorlati tanács a szignifikancia-szintek kezelésére, sem a globálisan kiadódó modell átfogó szignifikancia szintjére. Azonban alacsony számítási igénye miatt és rejtett változókat is kezelő kiterjesztései miatt ez a megközelítés lokális oksági részstruktúrák kikövetkeztetésére egy vonzó lehetőség. Elsőnek vizsgáljuk meg azt az esetet, hogy nem lehetnek zavaró tényezők [5, 17].

12.1. Példa. Az oksági Markov-feltétel garantálja, hogy három változó esetén már oksági relációkat tudunk kikövetkeztetni passzív megfigyelésekből is. Ekkor azon függetlenségi modell, amely tartalmazza $X, Y\ Y, Z$ közötti közvetlen függéseket, X, Z függetlenségét és $(X \not\perp\!\!\!\perp Z|\{Y\})$ feltételes függését, csak az úgynevezett v-struktúrát mutató $X \to Y \leftarrow Z$ DAG-gal reprezentálható.

Érdekes módon oksági relációk bizonyos esetekben zavaró tényezők potenciális jelenlétében is kikövetkeztethetőek, azaz amikor az oksági Markov-feltétel nem teljesül (lokális oksági felfedező algoritmusokért lásd [5, 17, 12].

12.2. Példa. Ha potenciális zavaró tényezők nincsenek a priori kizárva, akkor az előző példát folytatva még egy változót meg kell figyelni ennek kizárásához (oksági sorrend a priori feltevése esetén ismét elég három változó). Az előző példában szereplő függetlenségi modellt folytatva tételezzük fel, hogy megfigyelünk egy további W változót, Y, W direkt függéssel és feltételes függetlenséggel $(W \perp\!\!\!\perp \{X, Z\}|Y)$ (a stabilitás feltevése miatt W függ X-től és Z-től is). Mivel az Y feltétel függetlenséget jelent, a globális d-elválasztásos reprezentáció megköveteli, hogy legyen $Y \to W$ él, hiszen egy közvetítő $*$ zavaró tényező $* \ Y \to * \to W$ élekkel nem lenne lefogva Y által.

12.4. Teljes oksági modellek Bayes-i tanulása

Az oksági modellek kényszer alapú tanulásával szemben a pontszám alapú módszerekben egy globális pontszám a teljes modellnek az adathoz és az a priori ismeretekhez való illeszkedését jelzi. A pontszámokra egy természetes választás a modellek a posteriori valószínűsége a D_N adat feltételében. Egy Bayes-háló-struktúra poszteriorja a struktúrapriornak és a modell-likelihood-nak a szorzata:

$$p(G|D_N) \propto p(G) \int p(D_N|\underline{\theta}, G) p(\underline{\theta}|G)\, \mathrm{d}\underline{\theta} = p(G) p(D_N|G). \tag{12.4}$$

A likelihood tényezőre egy hatékonyan számolható képlet vezethető le (lásd [6, 18, 11]):

$$p(G, D_N) = p(G) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij+})}{\Gamma(\alpha_{ij+} + n_{ij+})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk+n_{ijk}})}{\Gamma(\alpha_{ijk})}, \tag{12.5}$$

Ezt *Bayesian Dirichlet*-poszteriornak nevezik, és ha az $\underline{\alpha}$ kezdeti hiperparaméterek kielégítik azt a feltételt, hogy a likelihood egy megfigyelési ekvivalenciaosztályon belül azonos értéked ad, akkor BD_e jelöli [11]. Ha a kezdeti $\underline{\alpha}$ hiperparaméterek konstans 1 értékűek, akkor $\mathrm{BD}_{\mathrm{CH}}$ jelöli [6]. Ha a kezdeti hiperparaméterek a lokális multinomiális modell paraméterei számának reciproka, akkor jele $\mathrm{BD}_{\mathrm{eu}}$ [3, 11].

Beavatkozásos adatoknál az Oksági modellek fejezetben bevezetett „do" szemantika szerint annyit változik ez a pontszám, hogy a beállított változókhoz tartozó szorzatok nem jelennek meg [10].

12.5. Oksági jegyek következtetése Bayes-hálók feletti átlagolással

A grafikus valószínűségi modellek használata genetikai asszociációs vizsgálatokban a családfa-elemzésekhez kapcsolódott, majd a genetikai variánsok kapcsoltsága miatt a tagSNP-k és a haplotípusok kezelésénél jelent meg. A genetikai interakciók, komplex fenotípusok és életmódbeli, környezeti módosító hatások figyelembevétele miatt az utóbbi években a grafikus valószínűségi modellek, különösen az oksági kapcsolatok modellezésére alkalmas Bayes-hálózatok használata genetikai asszociációs vizsgálatokban egyre elterjedtebbé váltak.

Az ismertetett módszertan másik eleme a Bayes-statisztika. Mielőtt megvizsgálnánk az oksági Bayes-hálók felhasználását ezen keretrendszerben, összefoglaljuk a Bayes-statisztikai keret általános sémáját. Ebben a statisztikai megközelítésben, parametrikus modelleket feltételezve, egy adott információs ellátottságú ξ szituációban a megfigyelések feletti $p(x|\xi)$ bizonytalan elvárásokat úgy állítjuk elő, hogy első lépésként meghatározzuk a releváns, θ paraméterezésű $p(x|\theta)$ modelleket, majd ezen θ paraméterezés felett egy $p(\theta|\xi)$ valószínűség eloszlást (az x_i mennyiségek a megfigyelhető, a θ paraméter a tipikusan nem megfigyelhető kategóriába esnek). A ξ információs kontextus és a valószínűségek feltételeiben való szerepeltetése a valószínűségek szubjektív értelmezését hivatott hangsúlyozni. Gyakran használt jelölés a ξ^+ és ξ^-, amelyek a neminformatív és informatív szituációkat jelölik. A $p(x, \theta|\xi)$ együttes eloszlás megkonstruálása után a valószínűségszámítás szabályai szerint tetszőleges következtetések lehetségesek uniform módon használva a megfigyelhető x_i mennyiségeket és a nem megfigyelhető θ paramétereket. A gyakorlatban elterjedt megközelítés szerint a hierarchikus specifikációban a releváns $\mathcal{M}^i$ modellosztályok specifikációjával, majd az azokon belüli $\mathcal{S}_k^i$ vagy M_k^i modell-struktúrák specifikációjával, és végül a modell-struktúrákhoz tartozó θ_k^i paraméterek specifikációjával történik. Ennek megfelelően egy adott i modellosztálybeli k struktúra θ_k^i paraméterezéséhez tartozó *a priori* bizonytalan elvárás egy szorzatként fejezhető ki:

$$p(\theta_k^i, M_k^i, \mathcal{M}^i) = p(\mathcal{M}^i)p(M_k^i|\mathcal{M}^i)p(\theta_k^i|M_k^i). \tag{12.6}$$

A modellek eloszlásainak specifikációját a megfigyelhető mennyiségekre vonatkozó $p(x|\theta, \phi)$ avagy $p(x|\theta_k^i, M_k^i)$ feltételes eloszlás egészíti ki a Bayes-statisztikai megközelítéshez tartozó teljes együttes eloszlássá.

A Bayes-statisztikai orvosbiológiai alkalmazását kezdetben olyan általános tulajdonságok motiválták, mint a statisztikai értelemben vett kismintás esetekben történő felhasználás, és az a priori ismeretek koherens beléptetése a statisztikai következtetésbe. Az omikai vizsgálatok ezt a két irányt felerősítették, mivel a statisztikai értelemben vett kismintás eset az orvosbiológiai kontextusban rendkívül nagyra növekedett változószám miatt lép fel. Ez a probléma a legegyszerűbb egyváltozós statisztikai vizsgálatokban, például genetikai asszociációs vizsgálatokban a többszörös hipotézistesztelés problémájaként aposztrofálódik. A Bayes-statisztika egyik előnye az a priori ismeretek felhasználásának fontossága a viszonylagosan alacsony mintaszám és komplex modellek miatt, illetve az orvosbiológiai háttértudás sokrétűsége és gazdagsága miatt fontos. A nagy áteresztőképességű, omikai mérések miatt lehetségessé vált hipotézismentes kutatás azonban a Bayes-i megközelítés másik előnyét is fontossá tette, hogy komplex modellek tulajdonságai kikövetkeztethetők lehetnek, annak ellenére, hogy a modellek között nincsenek dominánsak, sem nagy a posteriori valószínűségi régiók kis kiterjedéssel. Ekkor az adott adat mint feltétel meghatározza az adott modellosztályt használó konkrét elemzés során fennálló statisztikai bizonytalanságot, és az érdekes, megerősített modelltulajdonságok utólagos, adatelemzési eredményekből történő felismerése egy sokrétű feladatként jelenik meg.

Elsőként is vegyük észre, hogy a Bayes-i modellátlagolás a DAG-ok felett, nem csak a modellstruktúrák tulajdonságainak Bayes-következtetésében, hanem több feladatban is megjelenik (θ paraméterek feletti átlagolást analitikusan oldja meg a (12.5) képlet). Megjelenik a kérdéses modelltulajdonságot jelző F_c indikátorfüggvény valószínűségének becslésében, egy adott modell (vagy akár tulajdonság) várható veszteségének becslésében és megjelenik az úgynevezett teljes Bayes-i következtetésben is:

$$p(F_c = f_c|D_N) = \sum_G 1(F_c(G) = f_c)p(G|D_N), \tag{12.7}$$

$$L_{\hat{G}|D_N} = \mathrm{E}_{p(G|D_N)}[L(G,\hat{G})] = \sum_G L(G,\hat{G})p(G|D_N), \tag{12.8}$$

$$p(\underline{y}|\underline{x}, D_N) = \mathrm{E}_{p(G|D_N)}[\mathrm{E}_{p(\underline{\Theta}|G,D_N)}[p(\underline{y}|\underline{x},\underline{\Theta},G)]]. \tag{12.9}$$

Az oksági Bayes-hálók strukturális részét reprezentáló DAG-ok, közvetlenül és közvetve is, számos oksági értelmezéssel bíró modelltulajdonság definiálását teszik lehetővé, mint például az élek, irányítatlan élek, kényszerített élek, irányított utak; páronkénti, részleges és teljes változósorrendek, szülői halmazok és Markov-takaró gráfok.

12.5.1. Élek: közvetlen páronkénti függések

Az oksági Markov-feltétel mellett a legnyilvánvalóbb oksági Bayes-háló jegy az irányított él, amely egy „közvetlen” (nem mediált és feltétlen) páronkénti relációt reprezentál (a „közvetlenség” az oksági Markov-feltétel szerint értendő, tehát a modellezett szint alatt természetesen létezhetnek közvetítő változók, ám azok nem befolyásolnak más modellbeli változókat 12.11). Ha a hipotézisosztályok a Bayes-hálók megfigyelési ekvivalencia osztályai, akkor az ezeket reprezentáló esszenciális gráfokbeli kényszerített élek jelölnek egy potenciálisan oksági értelmezéssel felruházható relációt (a stabilitás feltevése és oksági

Markov-feltétel mellett). A megfelelő poszteriorok a következőek:

$$p(X_i \to_G X_j|D_N) = \sum_G 1(X_i \to_G X_j)p(G|D_N) \tag{12.10}$$

$$p(CompE(X_i, X_j|G)|D_N) = \sum_G CompE(X_i, X_j|G)p(G|D_N). \tag{12.11}$$

12.5.2. Áttételes páronkénti oksági relációk

A kényszerített él közvetlen volta ellenére egy teljes modelltől függő, globális aspektusokat is mutató páronkénti reláció. Összetett, azaz áttételes kapcsolatokat is megengedve számos további páronkénti oksági reláció definiálható, amelyek hasonlóan a teljes modelltől függnek. A 12.1. táblázat összefoglaló jelleggel mutat asszociációs, relevancia és oksági relációkat.

12.1. táblázat. Asszociációs, relevancia- és oksági relációk definíciói gráfos valószínűségi modellek felhasználásával

Reláció	Rövidítés	Gráfbeli definíció
Direkt oksági relevancia	DCR(X,Y)	Létezik él X és Y között
Tranzitív oksági relevancia	TCR(X,Y)	Létezik irányított út X és Y között
Oksági relevancia	CR	DCR vagy TCR
Zavart relevancia	ConfR(X,Y)	X-nek és Y-nak van közös őse
Asszociáció	A	DCR vagy TCR vagy ConfR
Tisztán (főhatás nélküli) interakciós relevancia	PIR(X,Y)	X-nek és Y-nak van közös gyermeke
Erős relevancia	SR(X,Y)	PIR vagy DCR

Több célváltozó esetén a következő komplex relációk is hasznosak lehetnek, amelyeket az a 12.2. táblázat foglal össze.

Egy $R(X, Y)$ páronkénti reláció poszteriorja a következőképpen adódik:

$$p(R(X, Y)|D_N) = \sum_G 1(R(X, Y); \text{holds}; \text{in}; G)p(G|D_N). \tag{12.12}$$

12.5.3. Markov-takaró (al)gráf

A diagnosztikai biomarkereknél központi szerepet betöltő Markov-takaró halmazt általánosítani lehet oly módon, hogy a releváns változók interakcióját (vagy annak hiányát) explicit módon reprezentáljuk.

12.14. Definíció (Markov-takaró gráf)**.** A G Bayes-háló-struktúra Markov-takaró részgráfja vagy határoló mechanizmusok modellje $\text{MBG}(Y, G)$ az Y változóra tartalmazza a $\text{bd}(Y, G)$ Markov-takarót és az Y-ba és gyerekeibe befutó éleket.

12.2. táblázat. Relevancia több célváltozó esetén

Reláció	Rövidítés	Gráfbeli definíció
Közvetlen relevancia egy vagy több célhoz	EdgeToAny(X,Y)	Létezik él X és valamely Y között.
Egyszeres közvetlen relevancia	EdgeToExactlyOne(X,Y-,Y')	Pontosan egy olyan Y van, amelyhez létezik él X-ből.
Többszörös közvetlen relevancia.	MultipleEdges(X,Y)	Több olyan Y van, amelyhez létezik él X-ből.
Közvetlen relevancia más célhoz	EdgeToSomewhereElse(X,Y)	Létezik él X és valamely nem Y-beli elem között'.

Ezzel az MBG(Y, G) Markov-takaró gráf, mint strukturális modelltulajdonság vezethető be (osztályozási algráfként is gyakran hivatkozott [1, 2].

Az oksági értelmezés szempontjából az MBG-knek egy fontos tulajdonsága, hogy az Y-ra vonatkozó autonóm mechanizmusok rendszerszintű kapcsolódásáról hordoz együttes, de mégis koncentrált információt. Sajnos az MBG poszterior számítása exponenciális komplexitású, azonban egy változósorrenddel vett feltételes poszterior polinom időben számítható [2]. A kapcsolódó MBG poszterior a következőképpen definiált:

$$p(\mathrm{MBG}(Y, G) = \mathrm{mbg}\,|D_N) = \sum_G 1(\mathrm{MBG}(Y, G) = \mathrm{mbg})p(G|D_N). \tag{12.13}$$

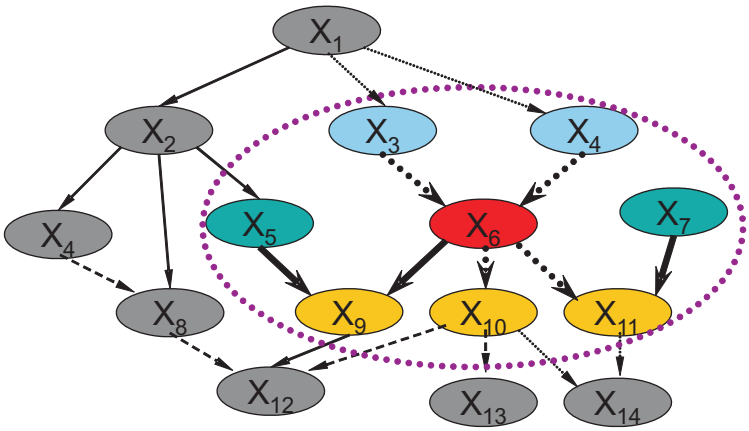
12.5.4. Hatásmódosítók

Az interakciók központi szerepe ellenére genetikai asszociációs, gén-környezet és farmakogenomikai kutatásokban az interakciók típusai jelenleg még nincsenek kidolgozva. A fejezetben tárgyalt rendszerszintű megközelítés lehetővé teszi altípusok definiálását, mint például a 12.1. ábrán látható asszociációs típusok: pontozott vonal jelzi az asszociált változókat X_6-tal, a szaggatott útvonal X_4-től X_{13}-ba jelzi azokat a változókat, amelyek potenciálisan befolyásoltak vagy relevánsak az X_4, X_{13} relációra, illetve a pontozott útvonal X_1-től X_{14}-ig jelzi azokat a változókat, amelyek potenciálisan asszociáltak vagy relevánsak az X_1, X_{14} oksági relációra.

Az oksági Bayes-hálók felhasználásának illusztrálására fontoljuk meg a következő kérdést:

Oksági relevancia hatáserősség-módosítója. Mi az a minimális halmaz, amely elszigeteli az X változón történő beavatkozás Y-ra gyakorolt hatását a többi változótól?

Adott feltételek mellett erre a válasz az X-ből Y-ba vezető utakon lévő csomópontjainak a szüleinek a halmaza, amelyhez a Bayes-i modellátlagolásos keretben szintén becsülhető poszterior.



12.1. ábra. Hatásmódosítók információs és oksági relevancia esetében

12.5.5. Változók sorrendje

Bár a változók teljes sorrendje ritkán jelenik meg önálló célként, implicit módon a DAG reprezentációban és így bármely tanulási eljárásban jelen van. Az oksági értelmezésben egy adott DAG-gal kompatibilis (topológiai) sorrendek oksági értelmezése az eredmények értelmezése szempontjából is alapvető fontosságú lehet. A sorrendek ezen technikai és oksági szerepe miatt is figyelemreméltó eredmény, hogy maximált szülőszám mellett egy adott változósorrend poszteriorja polinom időben kiszámítható [8]. A változók teljes sorrendjére (permutációira) is származtatott poszterior:

$$p(\prec |D_N) \quad = \quad \sum_G 1(G \in \mathcal{G}^{\prec}) p(G|D_N). \tag{12.14}$$

Irodalomjegyzék

[1] S. Acid, L. M. de Campos, and J. G. Castellano, Learning Bayesian network classifiers: searching in a space of partially directed acyclic graphs. *Machine Learning*, 59:213–235, 2005.

[2] P. Antal, G. Hullám, A. Gézsi, and A. Millinghoffer, Learning complex Bayesian network features for classification. In *Proc. of third European Workshop on Probabilistic Graphical Models*, pages 9–16, 2006.

[3] W. L. Buntine, Theory refinement of Bayesian networks. In *Proc. of the 7th Conf. on Uncertainty in Artificial Intelligence (UAI-1991)*, pages 52–60. Morgan Kaufmann, 1991.

[4] D. M. Chickering, A transformational characterization of equivalent Bayesian network structures. In *Proc. of 11th Conference on Uncertainty in Artificial Intelligence (UAI-1995)*, pages 87–98. Morgan Kaufmann, 1995.

[5] G. Cooper, A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery*, 2:203–224, 1997.

[6] G. F. Cooper and E. Herskovits, A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[7] A. P. Dawid, Conditional independence in statistitical theory. *J. of the Royal Statistical Soc. Ser.B*, 41:1–31, 1979.

[8] N. Friedman and D. Koller, Being Bayesian about network structure. In *Proc. of the 16th Conf. on Uncertainty in Artificial Intelligence(UAI-2000)*, pages 201–211. Morgan Kaufmann, 2000.

[9] D. Galles and J. Pearl, Axioms of causal relevance. *Artificial Intelligence*, 97(1-2):9–43, 1997.

[10] C. Glymour and G. F. Cooper, *Computation, Causation, and Discovery*. AAAI Press, 1999.

[11] D. Heckerman, D. Geiger, and D. Chickering, Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

[12] Subramani Mani and Gregory F. Cooper, A simulation study of three related causal data mining algorithms. In *International Workshop on Artificial Intelligence and Statistics*, pages 73–80. Morgan Kaufmann, San Francisco, CA, 2001.

[13] C. Meek, Causal inference and causal explanation with background knowledge. In *Proc. of the 11th Conf. on Uncertainty in Artificial Intelligence (UAI-1995)*, pages 403–410. Morgan Kaufmann, 1995.

[14] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, CA, 1988.

[15] J. Pearl, Causal diagrams for empirical research. *Biometrika*, 82(4):669–710, 1995.

[16] J. Pearl, *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[17] C. Silverstein, S. Brin, R. Motwani, and J. D. Ullman, Scalable techniques for mining causal structures. *Data Mining and Knowledge Discovery*, 4(2/3):163–192, 2000.

[18] D. J. Spiegelhalter, A. Dawid, S. Lauritzen, and R. Cowell, Bayesian analysis in expert systems. *Statistical Science*, 8(3):219–283, 1993.

[19] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*. MIT Press, 2001.

[20] T. Verma and J. Pearl, *Equivalence and synthesis of causal models*, volume 6, pages 255–68. Elsevier, 1990.

[21] M. Woodward, *Epidemiology: Study design and data analysis*. Chapman&Hall, 1999.

13. fejezet

Szövegbányászati módszerek a bioinformatikában

13.1. Bevezetés

Az emberiség egészen a digitális korszak kezdete óta számítógépet használt tudásának tökéletesítésére, tárolására és megosztására. Napjainkban évente több millió publikáció születik; e hatalmas mennyiségű kollektív tudással lépést tartani a kutatók számára reménytelen vállalkozás, még saját szakterületükön is. A szövegbányászat rohamosan fejlődő tudománya ezt a nehézséget hivatott orvosolni; pontosabban szólva, a szövegbányászat célja rejtett tudás felfedése nagy mennyiségű szöveges adat feldolgozásával. Orvosbiológiai kontextusban ez rendszerint cikkek tízezreinek vagy akár millióinak elemzését jelenti, amely lehetővé teszi eddig ismeretlen kapcsolatok felderítését és új hipotézisek generálását. A szövegbányászatra tekinthetünk az adatbányászat vadhajtásaként, amelyet először a 80-as években kezdtek alkalmazni, de a kutatás főáramába csak a XX. század végén került be. Az orvosbiológiai szövegbányászat azóta hatalmas fejlődésen ment át, részben a számítástechnika, részben más kapcsolódó területek (adatbányászat, gépi tanulás, statisztika, számítógépes lingvisztika) párhuzamos fejlődésének köszönhetően. E fejezetben alapfogalmakat és gyakran alkalmazott technikákat tekintünk át.

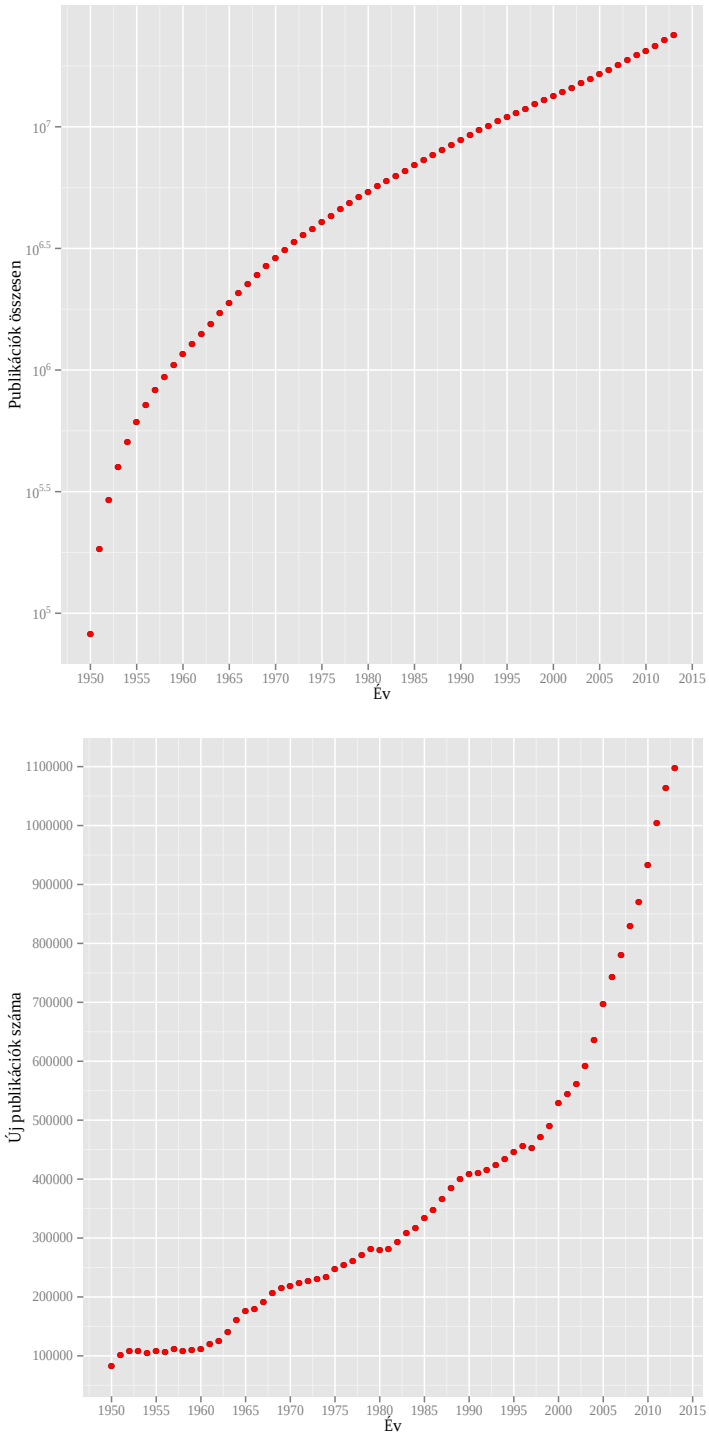
13.2. Orvosbiológiai szövegbányászat

Általánosságban – ám nem mindig – igaz, hogy az orvosbiológiai szövegbányászat a felhalmozott tudással tudományos közlemények formájában találkozik; egyéb források lehetnek például jelentések, szabadalmak, gyógyszer-tájékoztatók, blogbejegyzések stb. A folyamat bemeneteként a *korpusz* (dokumentumgyűjtemény) szolgál, amelyet gyakran kísér a kifejezések egy kontrollált *szótára* és a háttértudás egyéb forrásai. Kimenetként strukturált adatot kapunk, amelyet – hasonlóan a kutatás során felmerülő egyéb adatbázisokhoz – tárolni és rendszerezni kell, és akár nagyobb tudásbázisokba beépíthető. Egy általános munkafolyamat a következőképpen nézhet ki:

1. **Feladatleírás, eszközök megválasztása.** Az első lépések közé tartozik a problématerület meghatározása és a feladat leírása – mi a célunk, mit remélünk elérni a szövegbányászat alkalmazásával. Fontos a megfelelő eszközök megválasztása ezen célok eléréséhez; e fejezet többek között ebben kíván segítséget nyújtani.
2. **Korpuszépítés.** A korpusz a szövegbányászati folyamat bemenetéül szolgáló dokumentumok gyűjteménye. A korpuszépítés során nagy mennyiségű szöveges adat letöltésére, szűrésére kerül sor; szükséges lehet több feladatspecifikus korpusz létrehozása is.
3. **Korpusz feldolgozása.** A feldolgozás során az adatok könnyebben kezelhető formátumba kerülnek, így további műveletek végezhetők rajtuk. Az ebben a fázisban végezhető néhány transzformáció (pl. szótövezés) leírása a 13.2.1. alfejezetben található.
4. **Szótárépítés (opcionális).** Bemenetként számos eljárás igényli a vizsgálandó kifejezések kontrollált listáját. Megjegyezzük, hogy az ilyen szótárak építése esetenként bonyolultságuk miatt igen fárasztó és időigényes munka lehet (13.2.2. alfejezet).
5. **Jegykivonatolás (opcionális).** A gépi tanulási algoritmusok jellegzetessége, hogy az adatokat kivonatolt *jegyek* (feature) formájában várják – ezek tulajdonképpen az adatok kompakt, lényegre törő reprezentációi. A jegykivonatolás célja alkalmas jegyek számítása, amelyek hatékonyan kezelhetők és nagy mennyiségű információt hordoznak.
6. **Elemzés.** Rengeteg módszer létezik, kezdve az egyszerű előfordulás-alapú statisztikáktól a természetes nyelvi feldolgozáson (NLP) át a gépi tanulásig és egyéb kifinomult módszerekig; a fejezet további részében számos példát láthatunk.
7. **Adatszervezés, integráció, további lépések.** A kimenetként kapott strukturált adat más forrásokból származó adatokkal integrálható, így szélesebb tudásbázishoz juthatunk, amely számtalan módon felhasználható: pl. keresés, következtetés, válaszkeresés stb.

13.2.1. Korpuszépítés

A fellelhető biomedikális szövegek egésze – más néven a *bibliom* – felfogható a korpuszépítési folyamat bemeneteként. Az orvosbiológiai szövegbányászati alkalmazások hagyományosan a bibliom egy kitüntetett részét, a tudományos közlemények absztraktjait helyezték előtérbe; ennek legfőbb okai a kompakt, lényegre törő írásmód és a nyílt hozzáférés voltak. Napjainkra a hangsúly egyéb dokumentumtípusok (pl. szabadalmak, teljes cikkek) felé tolódott; ezek elérhetősége a szabad hozzáférés elvének köszönhetően folyamatosan növekszik. A dokumentumok közös jellemzője, hogy nem-strukturált adatot tartalmaznak, azaz a strukturált adatokkal szemben semmilyen előre meghatározott szerkezetet vagy modellt nem követnek, ami egy adatbázis esetén elvárható lenne. Nem-strukturált adatot hordoznak például a videók, képek és a szabadszöveges leírások. A bibliom egy kis része félig strukturált dokumentumokból áll, például XML fájlok formájában, amelyek így átmenetet képeznek az adatbázisok és a nem-strukturált adatok között.



13.1. ábra. Az összes és új publikációk száma az egyes években a PubMed adatbázisban

A korpuszépítés lépései közé tartozik a bibliom lekérdezése nyílt eszközökkel, például PubMed, Google vagy más keresőszolgáltatások segítségével. A feladattól függően szükséges lehet az eredmények szűrése, a különböző zavaró tényezők (lásd 13.3.6. alfejezet) kiküszöbölése és tárgyterület-specifikus korpuszok gyártása érdekében. A szűrés rengeteg szempont alapján végezhető, pl. publikációs dátum, cikktípus, kulcsszavak, MeSH termek, folyóiratok stb. szerint. Amennyiben a gyűjteményt többé-kevésbé teljesnek ítéljük, sor kerülhet a feldolgozására és eltárolására egy erre alkalmas formátumban.

A feldolgozás fogalma alatt temérdek eljárást érthetünk, pl. szótövezés, lemmatizálás (szótári alakra történő redukció), stopszó-szűrés (nemkívánatos vagy zavaró szavak, pl. kötőszavak) vagy tokenizáció (kisebb egységekre, pl. mondatokra történő szegmentáció). A feldolgozási eljárások egy speciális példája a korpusz *annotáció*ja, amelynek során nemszöveges információt csatolunk a dokumentum egyes elemeihez. Ez a biomedikális területen rendszerint szemantikus annotációt jelent, azaz egyes elemeket, pl. a gének vagy a fehérjék neveit megjelöljük egy előre meghatározott ontológia alapján. Ilyen annotált korpusz például a GENIA [1].

13.2.2. Szótárépítés

Szótár alatt a vizsgálandó kifejezések egy listáját értjük, amely a szótáralapú szövegbányászati módszerek elengedhetetlen bemenete. E módszerek rendszerint a megadott kifejezések keresésén alapulnak, és olyan feladatokat hajtanak végre, mint például az entitásfelismerés, együtt-előfordulási elemzés, szemantikus annotáció, szövegklasszifikáció stb. A szótáraknak sok formáját ismerjük:

- **Kontrollált szótár**ak általános értelemben különböző tudásforrások alapján építhetők, a legfontosabbak ezek közül a szakértői tudás és az online adatbázisok. A kifejezések kivonatolása és szűrése történhet félig vagy teljesen automatizált módon, számos online adatbázis nyújt ilyen szolgáltatásokat (UMLS, HUGO, OMIM stb.). A kifejezések szabadszöveges írásokból is kivonhatók, ezzel újabb szövegbányászati területekre jutunk (pl. ontológiák készítése [2]).
- **Taxonómiák** alatt hierarchikus struktúrával rendelkező kontrollált szótárakat értünk; a kifejezés hagyományosan az élőlények rendszertanát jelölte. Néhány említésre méltó példa: a Betegségek Nemzetközi Osztályozása (BNO), a gyógyszerek ATC-klasszifikációja, valamint egy sereg szakterület-specifikus taxonómia.
- **Tezauruszok** az előbbiektől eltérően nem csak hierarchikus kapcsolatokat engednek meg a kifejezések között. Az UMLS Metathesaurus például orvosbiológiai és egészségügyi kifejezések millióit, ezek szinonimáit és kapcsolatait tartalmazza.
- **Ontológiák**ról szigorú értelemben formális, számítógép által is olvasható reprezentációs nyelven leírt szótárak esetén beszélünk; a gyakorlatban azonban a fenti kategóriák mindegyikére használják az „ontológia” kifejezést. Az Open Biological and Biomedical Ontologies (OBO) Foundry a szakterületek széles skáláján elhelyezkedő ontológiákat tart fent.

13.2.3. Szövegbányászati feladatok

Még ha csak az orvosbiológiai kutatásra szorítkozunk is, a szövegbányászat igen széles alkalmazási területtel bír. Gyakran felmerülő feladatok:

- **Információ-visszakeresés** során releváns entitásokat adunk vissza a felhasználó által meghatározott kritériumok (lekérdezés) alapján. Az információ-visszakereső rendszereket gyakran keresőmotoroknak is nevezik. Erre mutat példát a PubMed, az egyik legszélesebb körben használt keresőmotor [3].
- **Entitásfelismerés.** Célja a szövegben egyedi „dolgokat" képviselő kifejezések megtalálása és megjelenítése – ilyenek például a gének vagy fehérjék szimbólumai, betegségek vagy más, névvel ellátható entitások. A következő lépésben, az ún. *normalizáció* során ezen találatokat külső adatbázisok azonosítóihoz rendeljük. A következő fejezetben részletesebben is megismerkedünk az entitásfelismerés elterjedt módszereivel.
- **Reláció-kivonatolás.** Az ide tartozó eljárások célja az entitások közötti kapcsolatok azonosítása; gyakran követi az entitásfelismerés lépését. Bár az entitásfelismerést sokan megoldottnak tartják, a reláció-kivonatolás sokkal összetettebb probléma, amely a jelentős erőfeszítések ellenére máig sem megoldott; néhány megközelítést szintén leírunk a következő fejezetben 13.3.4.
- **Hipotézis-generálás.** A kivonatolt relációk és statisztikai asszociációk rendszerét elemezve rejtett információk kerülhetnek felszínre, amelyek új hipotézisek alapjául szolgálhatnak.
- **Klasszifikáció és klaszterezés.** Mindkét kifejezés az entitások egyfajta „csoportosítására" utal, előbbi esetben előre ismert, utóbbiban ismeretlen kategóriákba. Ezen entitások lehetnek a korábban említett, névvel ellátott entitások vagy magasabb szintű objektumok, például dokumentumok vagy témák. A gépi tanulás területén a klasszifikáció és a klaszterezés jól ismert feladatok, leírásuk számos tankönyvben megtalálható.
- **Összefoglalás.** Az eljárás során egy kompakt összefoglalás keletkezik a dokumentumról a magas információtartalom megőrzése mellett. Rendszerint magában foglalja az egyes mondatok pontozását (többféle szempont, pl. pozíció vagy kulcsszavak alapján), majd a leginformatívabbnak ítélt mondatok kivonását. Egy másik lehetséges módszer az absztrakció: a szöveg egy szemantikus reprezentációját felhasználva természetes nyelvű összefoglalás generálható. Sajnos a természetes nyelvi generálás még mindig gyerekcipőben jár.
- **Ontológiakészítés.** Röviden említettük az előző alfejezetben. További részletekért lásd pl. [2].
- **Válaszkeresés.** A válaszkereső rendszerek felfoghatók speciális információ-visszakereső rendszerekként, amelyek természetes nyelvi interfésszel rendelkeznek. Az ilyen rendszerek szintaktikai és szemantikai elemzésnek vetik alá a lekérdezést. A következő lépésben az informatív szövegrészletek kivonására, szűrésére és pontozására

kerül sor; a feladatra sok megközelítés alkalmas, pl. következtetés, gépi tanulás vagy információ-visszakeresési technikák.

13.3. Alapvető szövegbányászati technikák

Ebben az alfejezetben egyszerű eljárásokat, majd néhány kifinomultabb megközelítést mutatunk be, amelyeket gyakran alkalmaznak az orvosbiológiai szövegbányászatban. A leírt technikák vagy az általános szövegbányászat mélyebb részletei iránt érdeklődők további információt a [4] és [5] tankönyvekben találhatnak.

13.3.1. Mintaillesztés

A mintaillesztés során előre meghatározott „mintákat" keresünk a szövegben; ez egyben a legtöbb szövegbányászati technika alapját is képezi. A minták lehetnek egyszerű sztringek (karaktersorozatok) vagy reguláris kifejezések (követelményeket reprezentáló speciális kifejezések, amelyek többféle sztringhez is illeszkedhetnek). A XX. század második felében mindkét célra rengeteg algoritmust terveztek. Előbbire példa a Boyer–Moore algoritmus [6]; a reguláris kifejezések és véges állapotú automaták részleteiért Cox összefoglaló művére hivatkozunk [7].

Az ún. „fuzzy" mintaillesztéssel (más néven hibatűrő mintaillesztés) adott távolságmérték alapján mért „hozzávetőleges" egyezések is megtalálhatók. Ezen módszerek nemcsak a szövegbányászatban, hanem a szekvenciaillesztésben is hasznosak. Néhány gyakran használt távolságmérték:

- **Hamming-távolság:** egyforma hosszúságú sztringekben azon pozíciók száma, ahol a karakterek eltérnek.
- **Levenshtein-távolság:** inzerciók, deléciók és szubsztitúciók száma, esetleg valamilyen súlyozási sémával.
- **Manhattan-távolság:** vektortér-reprezentációban a koordináták abszolút különbségeinek összege.
- **Biológia által inspirált távolságok:** Needleman–Wunsch, Smith–Waterman távolság; eredetileg szekvenciaillesztésben alkalmazták.

13.3.2. Dokumentumok reprezentációja

A szabadszöveges leírások számítógépes elemzéséhez elengedhetetlen a dokumentumok reprezentációja valamely jól definiált, gép által is olvasható módon – más szóval, strukturált adatként. A feladattól függően több lehetőség közül választhatunk; leggyakrabban a vektortér-modellt és a valószínűségi megközelítéseket használják.

Jelölje $t_k, \quad k = 1, 2, ..., m$ a kifejezéseket, valamint $d_i, \quad i = 1, 2, ..., n$ a dokumentumokat. Legyen D egy $m \times n$ mátrix (kifejezés–dokumentum mátrix), amelyre $D_{ki} = 1$

ha a d_i dokumentum tartalmazza a t_k kifejezést. Így a t_k kifejezéseknek D_k sorai felelnek meg, továbbá az egyes sorokra gondolhatunk egy n-dimenziós vektortér elemeiként – innen a modell neve. Hasonlóképp, a D^i oszlopok dokumentumokat képviselnek, és egy m-dimenziós vektortér elemeit adják. Látható, hogy ez a modellcsalád nem veszi figyelembe a kifejezések dokumentumbeli sorrendjét, gyakran hívják ezért „szózsák" (bag of words) modellnek is. A kifinomultabb változatok n_{ki}-t, a t_k kifejezés d_i dokumentumbeli frekvenciáját használják bináris előfordulás helyett, vagy más összetett súlyozási sémát használnak. Igen elterjedt séma a tf–idf (kifejezésfrekvencia–inverz dokumentumfrekvencia), amely a következőképpen számolható:

$$D_{ki} = tf(t_k, d_i) \cdot idf(t_k, D) = \frac{n_{ki}}{|D^i|} \cdot \log\left(\frac{n}{n_k}\right),$$

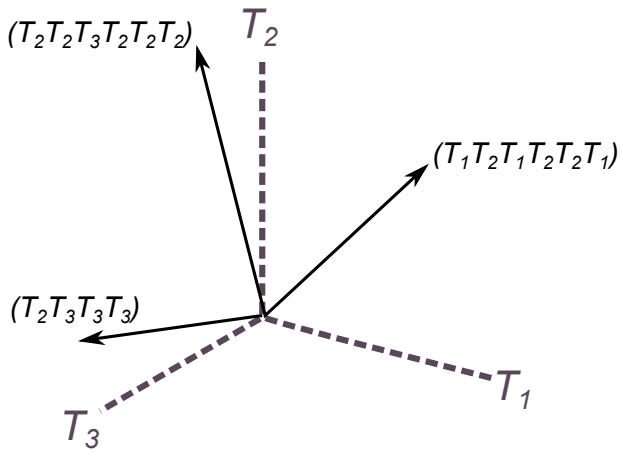
ahol $tf(t_k, d_i)$ a t_k kifejezés d_i dokumentumbeli relatív frekvenciája, n_k azon dokumentumok száma, amelyekben a t_k kifejezés előfordul, valamint $idf(t_k, D)$ jelöli a t_k kifejezés inverz dokumentumfrekvenciáját (megállapodás szerint logaritmust alkalmazva). A vektortér-modell figyelemre méltó előnye, hogy különösen egyszerűvé teszi dokumentum–dokumentum és kifejezés–kifejezés hasonlóságok kiszámítását, ami igen jól jön egyes feladatoknál (klasszifikáció, klaszterezés). Rengeteg hasonlóságmérték közül válogathatunk, az egyszerű koszinusz-hasonlóságtól egészen komplex, kifinomult hasonlóságmértékekig.

Nyilvánvaló, hogy a vektortér-reprezrezentációk általában rendkívül magas dimenziójúak és igen ritkák. Gyakorlati problémák esetén a dimenzionalitás redukciójára algoritmusok széles körét javasolták. Az alábbi listán néhány példát láthatunk:

- **Lingvisztikai megközelítések:** szótövezés, lemmatizáció, stopszó-szűrés.
- **Mátrix-dekompozíciók:** szinguláris értékek szerinti felbontás (SVD, ebben a kontextusban még: látens szemantikus indexelés, LSI), CUR dekompozíció, más alacsony rangú approximációk.
- **Gépi tanulási eljárások:** jegykiválasztás/kivonatolás, főkomponens-analízis (principle component analysis, PCA), multidimenzionális skálázás (multidimensional scaling, MDS), önszerveződő térképek (self-organizing maps, SOM).

A reprezentáció kapcsán gyakran esik a választás a valószínűségszámításra és valószínűségi modellekre. E megközelítéseket elsőként információ-visszakereső rendszerekben és levélszemét-szűrőkben alkalmazták. Mivel számos feladatban felülmúlják a többi modellt, ráadásul kitűnően alkalmazhatók orvosbiológiai kontextusban, mára a szövegbányászati eszköztár nélkülözhetetlen elemeivé váltak. Részletes tárgyalásuk sajnálatos módon messze túlmutat e tankönyv keretein, így csupán néhány bevált technikát sorolunk fel, a valószínűségi modellek további részleteiért más művekre hivatkozunk [8].

- Markov véletlen mezők (Markov Random Field, MRF), feltételes véletlen mezők (Conditional Random Field, CRF)
- Rejtett Markov-modellek (Hidden Markov Model, HMM)
- Bayes-i modellek



13.2. ábra. A vektortérmodell sematikus ábrázolása. T_1, T_2 és T_3 kifejezéseket jelölnek, a nyilak pedig az ezekből álló „dokumentumokat”.

- Bayes-hálók (Bayesian Network)
- Valószínűségi környezetfüggetlen nyelvtanok (Probabilistic Context-Free Grammar, PCFG és LPCFG)

13.3.3. Az entitásfelismerés módszerei

Az entitásfelismerés (named entity recognition, NER) egyedi, „nevesített” entitások felismerését és megjelölését jelenti. Négy fő megközelítést ismerünk:

- **Szótáralapú módszerek**, amelyek rendszerint egzakt vagy hibatűrő mintaillesztést használnak az entitások azonosítására.
- **Szabályalapú módszerek** alatt különböző empirikus szabályokkal operáló rendszereket értünk. Ismert, hogy már néhány intuitív szabály is elfogadható teljesítményhez vezet: figyelembe vehetők például a nagybetűk, kontextuális jegyek (idézőjelek, zárójelek), pozíció a szövegtörzsben vagy a címben, frekvencia, szakterület-specifikus jegyek stb. Hasonló szabályok akár tanulhatók is gépi tanulási technikákkal.
- **Gépi tanulási eljárások** szintén sikerrel alkalmazhatók. A klasszifikáció-alapú megközelítések a gépi tanulásban leírt klasszifikációs algoritmusok széles tárházából válogatnak; ezek előzetesen annotált korpuszon történő tanítást igényelnek. A különböző szekvencia-alapú eljárások – néhányat már láttunk a valószínűségi modellek leírásánál – ún. „tag”-ekkel felcímkézett korpuszok felhasználásával parametrizálhatók; működésük során a legvalószínűbb címkéket jósolják az egyes szavakra.
- **Hibrid megoldások** ötvözhetik az előzőeket.

További részletek és nyílt eszközök leírása megtalálható a hivatkozott irodalomban [9]. A következő lépés rendszerint a normalizáció, azaz a felismert entitások hozzákötése különbö-

ző adatbázisok azonosítóihoz – könnyű feladat szótáralapú megoldásoknál, míg a többinél munkaigényessé válhat.

13.3.4. A relációkivonatolás módszerei

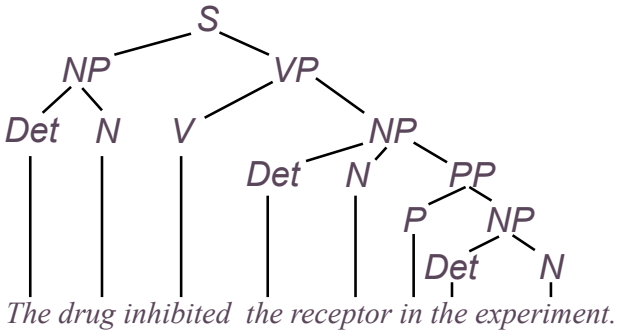
A relációkivonatolás entitások között fennálló különböző típusú relációk felismerését jelenti. Helyesen használva rendkívül hatékony eszköze lehet a hipotézis-generálásnak, mivel az adatokba ágyazott, emberi léptékben láthatatlan kapcsolatokra deríthet fényt. A relációkivonatolás azonban összehasonlíthatatlanul nehezebb feladat, mint az entitásfelismerés, mivel a relációkat meghatározó kifejezések gyakran elszórva helyezkednek el a mondatokban és bekezdésekben. Az előző részben leírt megközelítések a relációkivonatolásban is használhatók, azaz léteznek szótáralapú, szabályalapú és gépi tanulási rendszerek. A kivonatolt relációk a következőképpen oszthatók fel:

- **Statisztikai relációk** detektálása a legegyszerűbb feladat. A szótáralapú entitásfelismerő eljárások jól használhatók kifejezés-előfordulások megszámlálására, melyeket együtt-előfordulási statisztikák kiszámítására lehet felhasználni. Az igen/nem együtt-előforduláson és frekvencia-alapú modelleken túl meghatározhatunk kifinomultabb mértékeket is, pl. kölcsönös információ (mutual information). Az elképzelés súlyos hátulütője, hogy nem veszi figyelembe a kontextust: a csak felvetett, gyanított, sőt, egyenesen tagadott állítások ugyanúgy valid relációkként fognak megjelenni.
- **Szemantikai relációk**at rendszerint természetes nyelvi feldolgozás (Natural Language Processing, 13.3.5. alfejezet) útján azonosíthatunk. E rendszerek a mondatok szintaktikai szerkezetét tükröző elemzési fát (parse tree) építenek, majd ezekben különböző szerkezeteket azonosítanak a relációk felismerése érdekében. Ilyen szerkezetek az RDF adatmodell által is használt tárgy–predikátum–objektum hármasok: a „cAMP inhibits Ras” fordulat például ilyen struktúrára fordítható.
- **Szintaktikai relációk**, amelyek mostanában kerültek a kutatás középpontjába, és erősen kapcsolódnak a kernel-alapú relációs tanuláshoz. Az ötlet lényege, hogy a relációkra szintaktikai struktúraként (elemzési fa vagy függőségi gráf) gondolunk, majd ismert relációkat tanítómintaként használva gépi tanulás útján próbálunk további hasonló relációkat találni. A módszer jó teljesítményt mutatott a gyakorlatban [10].

Korábban már hangsúlyoztuk a relációkivonatolás hasznát a hipotézis-generálásban. A legelső modellt, amely ezt a megközelítést alkalmazta, Swanson javasolta 1986-ban [11]. A „felfedezés ABC-modellje” néven híressé vált elgondolás a szakirodalom két elszigetelt régiójából indul ki (azaz a külön csoportba tartozó szerzőknek nincs közös cikke, nem idézik egymást és nem idézik őket együtt). Ekkor ha az A és B entitások közötti relációt leírják az egyik csoportban, valamint a B és a C közötti relációt a másikban, akkor egy eddig ismeretlen, A és C között fennálló relációra következtethetünk. Az Arrowsmith-eszköz együtt-előfordulási statisztikákkal kombinálta a megközelítést, és sikeresen használta fel kifejezések közötti relációk indukciójára. Szakirodalom alapú felfedezést szolgáló rendszerek és leírásuk megtalálhatók a hivatkozott irodalomban [12].

13.3.5. Lexikalizált valószínűségi környezetfüggetlen nyelvtanok

A formális nyelvek elmélete a matematikai logika, számítógépes nyelvészet és a számítástudományok határán helyezkedik el. Bár a terület évszázadok óta ismert, még ma is születnek új alkalmazásai. A lexikalizált valószínűségi környezetfüggetlen nyelvtanok (LPCFG, SLCFG) a természetes nyelvi elemzés különösen hatékony eszközei, amelyeket a legkorszerűbb elemzők implementálnak (pl. a Stanford Parser [13]). Az orvosbiológiai szövegbányászatban ezeket az eszközöket a tudományos publikációkat alkotó mondatok elemzési fáinak építésére használhatjuk fel, messze meghaladva a hagyományos együttelőfordulási és szabályalapú modelleket.



13.3. ábra. Egy egyszerű mondat elemzési fája

Környezetfüggetlen nyelvtanok (CFG) alatt a $G = (N, \Sigma, R, S)$ négyest értjük, ahol

- N a nem-terminális szimbólumok véges halmaza, pl. `S` (mondat), `VP` (igei kifejezés), `NP` (főnévi kifejezés), `NN` (főnév), `Vi`/`Vt` (intranzitív/tranzitív ige).
- Σ a terminális szimbólumok véges halmaza, pl. `cAMP`, `Ras`, `inhibit`.
- R az átírási szabályok véges halmaza, amelyek a következő formában írhatók: $X \to Y_1 Y_2 ... Y_n$, ahol X egyetlen nem-terminális szimbólum, Y_i pedig bármilyen szimbólum; pl. `S` $\to$ `NP VP`, `NN` $\to$ `cAMP`.
- $S \in N$ a start szimbólum, amely az elemzési fa gyökerét képezi (`S`).

Az átírási szabályok használatával minden nyelvtanilag helyes mondathoz egy vagy több elemzési fa építhető. A valószínűségi CFG az előbbi triviális kiterjesztése. A kétértelműség feloldása érdekében minden átírási szabályhoz valószínűséget rendelünk:

$$\begin{aligned} P(\texttt{S} \to \texttt{NP VP}) &= 1.0, \\ P(\texttt{VP} \to \texttt{Vi}) &= 0.6, \\ P(\texttt{VP} \to \texttt{Vt NP}) &= 0.4, \\ P(\texttt{NN} \to \texttt{cAMP}) &= 0.001. \end{aligned}$$

A lehetséges elemzési fáknál a valószínűségeket összeszorozva kiválaszthatjuk a mondatot legnagyobb valószínűséggel jellemző elemzési fát (13.3. ábra). A lexikalizált PCFG-k egy

további lépést jelentenek, ahol az átírási szabályokban konkrét szimbólumok kerülnek a feltételekhez:

$$\begin{aligned} P(\texttt{VP} \rightarrow \texttt{Vt NP} | \texttt{Vt} = \texttt{inhibit}, Head(\texttt{NP}) = \texttt{Ras}) &= 0.1, \\ P(\texttt{VP} \rightarrow \texttt{Vt NP} | \texttt{Vt} = \texttt{inhibit}, Head(\texttt{NP}) = \texttt{spaceship}) &= 0.00001. \end{aligned}$$

13.3.6. Az orvosbiológiai szövegbányászat kihívásai

Jelentős erőfeszítéseink ellenére az orvosbiológiai szövegbányászat eredendő buktatóinak megkerülése igen nehéz feladatnak bizonyul:

- **Rokonértelműség** (szinonímia) elsősorban a szótáralapú entitásfelismerést érinti. A kielégítően pontos felismeréshez és normalizációhoz elkerülhetetlen a szinonimák figyelembe vétele; ez hatalmas ugrást eredményez a kifejezések számában, amely viszont a teljesítmény csökkenéséhez vezet.
- **Azonosalakúság** (homonímia) alatt azonosan írt, de teljesen más jelentésű kifejezéseket értünk, amely értelemszerűen az entitásfelismerő rendszerek pontosságát is befolyásolja.
- **Visszautalások** (anafora) alatt egy korábbi szövegrészre utaló nyelvtani elemet (pl. mutatószók, névmások) értünk. A visszautalások automatikus feloldása ma is erősen kutatott terület [14].
- **Morfológiai variánsok** gyakran fordulnak elő az orvosbiológiai szakirodalomban; rendszerint szinonimaként hozzáadva vagy hibatűrő mintaillesztéssel kezelik.
- **Betűhibák** szintén elkerülhetetlenek nagy terjedelmű szabad szöveg elemzésénél. Hibatűrő mintaillesztés használható az elírt entitások felismeréséhez.
- **Rövidítések** rendkívül gyakoriak a biomedikális közleményekben, ami komoly kihívást jelent az entitásfelismerő rendszereknek; mi több, a rövidítések körében azonosalakúság sem ritka, amely a normalizációt is megnehezíti (pl. egy génszimbólum több, teljesen független génre is vonatkozhat). Végül pedig számos rövidítés alakra teljesen azonos egyéb rövid szavakkal, amely szintén rontja a tisztán szótáralapú eszközök teljesítményét (némiképp kiküszöbölhető szabályalapú kiegészítések beépítésével).
- **Kifejezés-határok** megállapítása nem egyértelmű az esetenkénti átlapolódás vagy kontextusfüggőség miatt. Számos rendszer szabályalapú megközelítést vagy szintaktikai elemzést használ.
- **A szótárak elavulása** a tudomány fejlődésével viszonylag gyorsan bekövetkezik; fenntartásuk jelentős munkát igényel.
- **A normalizáció referencia-adatbázisai** rendszerint hiányosak. A kapcsolódás és az adatbázisok közötti leképezések igazi kihívásnak bizonyulhatnak.

Az orvosbiológiai szövegbányászat kapcsán a szisztematikus hiba (bias) lehetősége is felmerül:

- **Publikációs bias.** „Pozitív” eredményeket sokkal nagyobb valószínűséggel publikálnak, mint „negatívakat”; a probléma megkerüléséhez sok hatóság és folyóirat megköveteli a tanulmány az indítás előtti regisztrációját. Ennek ellenére 2009-ben a regisztrált klinikai kísérletek kevesebb, mint feléről publikáltak eredményeket [15].
- **Szelekciós bias.** Mivel nem minden publikáció szabadon hozzáférhető, a nagyléptékű szövegbányászati kutatások rendszerint absztraktokra szorítkoznak, amelyek viszont csak részleges információt tartalmaznak. A nyílt hozzáférés (Open Access) egyre növekvő elfogadottsága lehetővé teheti e hiba elkerülését.
- **Mintavételezési bias.** Az orvosbiológiai kutatásokban gyakran tanulmányozott entitások iránti preferencia szintén torzíthatja a levont következtetéseket.

13.4. Szövegbányászat és tudásszervezés

E fejezetben beszéltünk a nem-strukturált (szabad) szöveg elemzéséről és strukturált adattá történő konverziójáról. Ez az átmeneti reprezentáció számos formát ölthet; láttuk a szózsákmodellt, valószínűségi modelleket, elemzési fákat vagy függőségi/fogalmi gráfokat stb. A reprezentációk között alapvető különbség a szemantika „mennyisége”: míg a szózsákmodell csak előfordulásokat jellemző adatvektorokká redukálja a szöveget, a természetes nyelvi feldolgozás során adódó reprezentációk sokat megőriznek az eredeti gazdag szemantikából.

Rengeteg szövegbányászati algoritmus induktív következtetést alkalmaz az átmeneti strukturált adaton, más szóval általános szabályokat azonosít a modellben hordozott konkrét megfigyelések, pl. együtt-előfordulások alapján. Bár a következtetés ezen formája adatbányászati területen és szövegbányászatban egyaránt remekül működik, nem használja ki a természetes nyelv gazdag kifejezőerejét. Sokkal „természetesebb” megközelítés volna abduktív vagy deduktív következtetés útján új tudást felfedni a szöveg szemantikai tartalmának alkalmas reprezentációjából.

A megközelítés tovább erősíthető a szemantikus publikáció elveinek követésével. A fogalom a tudományos közlemények szemantikai információval való feldúsítását jelenti, lényegében egy formális tudásreprezentációs réteg létrehozásával, amely az információvisszakeresést és tudásfelfedezést támogathatná, valamint a teljes szakirodalom egységes szemléletét tehetné lehetővé. Bár számos útmutató, szemantikus nyelv és fogalom (pl. „strukturált digitális absztrakt”) született, a tudományos publikáció ezen új korszaka még várat magára.

Irodalomjegyzék

[1] J. D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, GENIA corpus–semantically annotated corpus for bio-textmining. *Bioinformatics*, 19 Suppl 1:i180–182, 2003.

[2] Philipp Cimiano, *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[3] Z. Lu, PubMed and beyond: a survey of web tools for searching biomedical literature. *Database (Oxford)*, 2011:baq036, 2011.

[4] Matthew S. Simpson and Dina Demner-Fushman, Biomedical Text Mining: A Survey of Recent Progress. In: Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 465–517. Springer, 2012.

[5] Sholom M. Weiss, Nitin Indurkhya, and T. Zhang, *Text Mining. Predictive Methods for Analyzing Unstructured Information*. Springer, Berlin, 1st. ed. 2004.

[6] Robert S. Boyer and J. Strother Moore, A Fast String Searching Algorithm. *Commun. ACM* 20(10):762–772, October 1977.

[7] Russ Cox, Regular expression matching can be simple and fast, 1 2007.

[8] Yizhou Sun, Hongbo Deng, and Jiawei Han, Probabilistic Models for Text Mining. In: Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 259–295. Springer, 2012.

[9] U. Leser and J. Hakenberg, What makes a gene name? Named entity recognition in the biomedical literature. *Brief Bioinform*, 6(4):357–369, December 2005.

[10] Chad M. Cumby and Dan Roth, On Kernel Methods for Relational Learning. In: T. Fawcett and N. Mishra, editors, *Proceedings of the 20th International Conference on Machine Learning (ICML 2003)*, pages 107–114, Washington, DC, USA, August 2003. AAAI Press.

[11] D. R. Swanson, Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect. Biol. Med.*, 30(1):7 18, 1986.

[12] M. Yetisgen-Yildiz and W. Pratt, Evaluation of Literature-Based Discovery Systems. *Literature-based Discovery*, pages 101–113. 2008.

[13] Dan Klein and Christopher D. Manning, Accurate Unlexicalized Parsing. In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, Vol. 1, ACL '03, pages 423–430, Association for Computational Linguistics, Stroudsburg, PA, USA, 2003.

[14] Jennifer D'Souza and Vincent Ng, Anaphora Resolution in Biomedical Literature: A Hybrid Approach. In: *Proceedings of the 3rd ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pages 113–122, 2012.

[15] S. Mathieu, I. Boutron, D. Moher, D. G. Altman, and P. Ravaud, Comparison of registered and published primary outcomes in randomized controlled trials. *JAMA*, 302(9):977–984, Sep. 2009.

14. fejezet

Kísérlettervezés: az alapoktól a tudásgazdag és aktív tanulásos kiterjesztésekig

14.1. Bevezetés

A kísérletezés az emberiség egyik leghatékonyabb eszköze a körülötte lévő világ felfedezésére; bármiféle tudományos (vagy akár filozófiai!) előrehaladás elképzelhetetlen volna gondosan megtervezett kísérletek nélkül. Nem meglepő, hogy – a fejlődéslélektan legtöbb képviselője szerint – a kísérletezés az emberi kognitív fejlődésben is központi szerepet tölt be. Jean Piaget a 12–18 hónapos gyermekeket egyenesen „fiatal tudósoknak" tartotta, akik a világot kísérletek tervezésén és kivitelezésén keresztül fedezik fel.

Mindennek dacára a matematikusok érdeklődését csak a XX. században kezdte felkelteni a kérdés. Amióta Ronald Fisher, az egyik legnevesebb statisztikus (egyben elismert evolúcióbiológus és genetikus) megírta „The Design of Experiments" c. művét (1935), a kísérlettervezés a matematikai statisztika jelentős alterületévé nőtte ki magát. Ebben a fejezetben áttekintjük a kísérlettervezés folyamatát a biológus és a statisztikus nézőpontjából egyaránt.

14.2. A kísérlettervezés alapjai

A kísérlettervezés (KT; angolul Design of Experiments, DOE) célja, hogy egy kísérlet valamilyen értelemben vett *optimális* voltát biztosítsa. Ez rendszerint azt jelenti, hogy a lehető legtöbb információt akarjuk kinyerni a lehető legkisebb torzítás, hiba, idő és költségek mellett. Szintén elsődleges cél helyes kérdések feltétele, valamint helyes következtetések levonásának lehetősége; az értelmetlen kérdések és a tervezés hibáiból fakadó félremagyarázások az egész kutatás sorsát megbélyegezhetik, függetlenül a minták minőségétől és a mérések kivitelezésétől. Az orvosbiológiai KT magában foglal olyan gyakorlati feladatokat is, mint például a mintagyűjtés és mintatárolás megszervezése, a felszerelés

használatának és személyzeti kérdések menedzselése, stb. Bár az orvosbiológiai KT jelentős mértékben támaszkodik az epidemiológiai tanulmányok tervezésére, erre ebben a fejezetben nincs módunk kitérni; további információért lásd pl. [1].

14.2.1. Az orvosbiológiai kísérlettervezés lépései

Az orvosbiológiai KT a következő főbb lépésekre bontható:

1. **Tárgyterület modellezése.** Rendszerint magában foglalja a szakirodalom alapos átkutatását; leggyakrabban maguk a tudósok végzik, változó mértékű bioinformatikai támogatással. Ennek egyik végleteként gondolhatunk egy kutatóra, aki különböző keresőkkel (pl. PubMed) publikációkat gyűjt és olvas; a másik véglet lehet egy teljesen integrált adat- és szövegbányász rendszer, amely emberi beavatkozás nélkül végrehajtja a szakirodalomban fellelhető tudás kivonatolását, modellezését és vizualizációját.

2. **Célok kitűzése.** Ez a lépés szoros kapcsolatban áll a *hipotézisek* felállításával. Egyrészt, a kísérletek általában a versengő lehetséges magyarázatok közötti döntés megkönnyítését szolgálják. Másrészt viszont – legalábbis a biológia területén – az előre felállított hipotézisek immár nem szükségesek: a poszt-genomikus korszak számos nagy áteresztőképességű méréstechnikát kínál, amelyek nem igénylik hipotézisek felállítását, sőt, akár hipotézisek *generálására* is felhasználhatók.

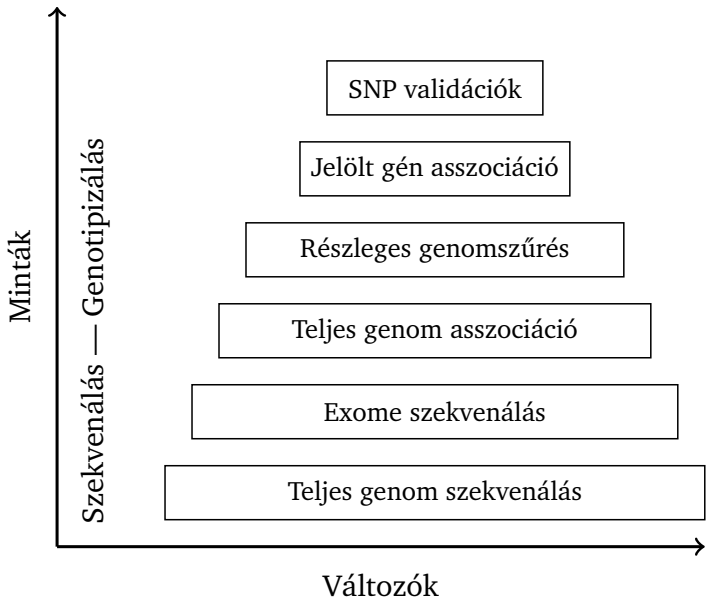
3. **Mintaszám és célváltozók meghatározása.** A célváltozók lényegében a kísérlet kimeneti változói: egy kísérletben különböző bemeneti paramétereket vagy *faktorokat* beállítva azt vizsgáljuk, hogy ezek milyen hatással vannak a kimenetre (*célváltozókra*). A jó kérdésfeltevés gyakran a sikeres kísérlet kulcsa, ennek pedig központi eleme a megfelelő mintaszám és célváltozó-halmaz meghatározása.

4. **Technikai részletek finomítása.** Ebben a lépésben technikai részletek kerülnek kidolgozásra, mint például az adat- vagy mintagyűjtési protokoll, tárolás, hiányos adatok kezelése, előfeldolgozás, technológia és felszerelés megválasztása (valamint ehhez kapcsolódó egyéb tevékenységek, pl. assay-tervezés), etikai és jogi kérdések, stb. Számos feladat ezek közül szintén jelentős bioinformatikai támogatást igényel.

14.2.2. A biológiai kísérletek fajtái

A kísérletek felosztása számos szempontnak megfelelően történhet. A feladat matematikai-statisztikai természete alapján például a következő kategóriákat állíthatjuk fel:

- **Asszociációk felderítése.** Asszociációról beszélhetünk akkor, ha egy entitás (pl. génvariáns) szignifikánsan gyakrabban fordul elő egy adott betegségben szenvedő emberekben; nem feltétlenül jelent azonban ok-okozati kapcsolatot vagy kóroki tényezőt.



14.1. ábra. Az egyes kísérlettípusok során felmerülő nagyságrendek

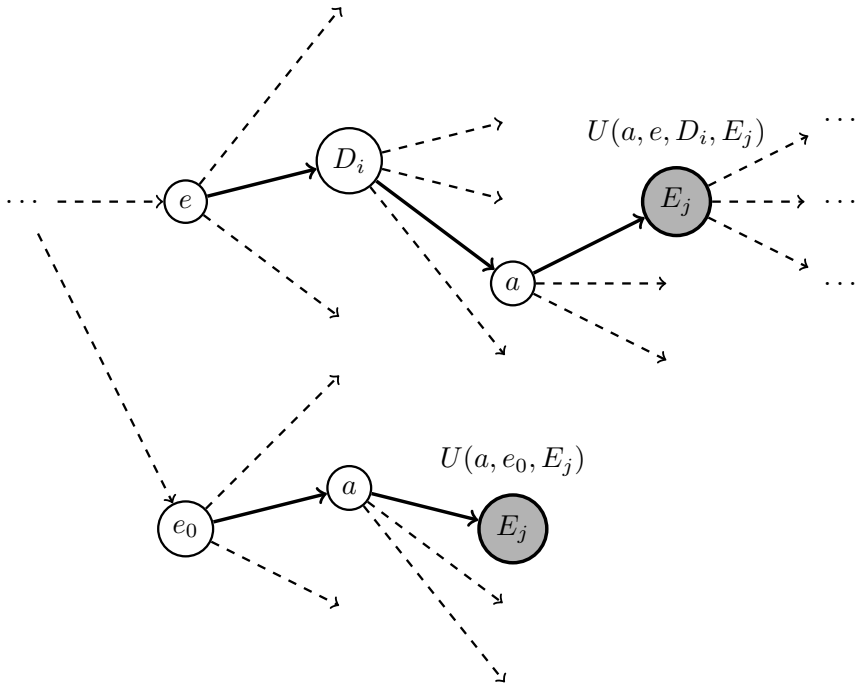
- **Klasszifikáció.** A klasszifikáció vagy osztályozás során adott mintákat próbálunk előre meghatározott osztályokba sorolni. Gondolhatunk például a kötelező szűrővizsgálatokra, ahol ezek az osztályok értelemszerűen a „beteg" és a „nem beteg".

- **Klaszterezés.** A klaszterezés annyiban különbözik az előbbitől, hogy nem állnak rendelkezésre előre meghatározott osztályok, a célunk mégis a minták csoportosítása. Gyakran használjuk génexpressziós adatok elemzésénél (pl. microarray adatok biklaszterezése).

- **Regresszió.** A regresszió során számszerű értékeket próbálunk jósolni az egyes mintákhoz, illetve meghatározni a célváltozóra legerősebb hatást gyakorló faktorokat; felhasználható például betegségek kimenetelének jóslására.

- **Összehasonlítás.** Az összehasonlítás a hipotézisek felállításának egyik legegyszerűbb és leghatékonyabb módja.

- **Modellezés/hipotézisgenerálás.** A modellezés során a valós világban megtalálható bonyolult kapcsolatrendszert képezzük le egy egyszerűbb matematikai konstrukcióra. Ez a folyamatot nevezik absztrakciónak is, amelynek során tehát „lényeges" és „lényegtelen" tulajdonságokat próbálunk elkülöníteni. A lényeg kivonásával és hatékony reprezentációjával lehetőség nyílik eddig rejtett információk felderítésére, döntéstámogatásra, hipotézisek felállítására, vagy akár szisztematikus generálására.



14.2. ábra. A munkafolyamatot szemléltető valószínűségi gráf [2]. A kísérleteket e, az adatokat D_i, a cselekvéseket a, az eseményeket E_j jelöli. Az e_0 csomópont nem elvégzett kísérletet jelöl; $U(\cdot)$ a hasznosságfüggvény.

14.3. A kísérlettervezés döntéselméleti megközelítése

14.3.1. A kísérlet várható értéke

A kísérlettervezés statisztikai megközelítésének megértéséhez először meg kell ismerkednünk a hasznosságelmélet alapfogalmaival. Képzeljünk el egy munkafolyamatot, ahol minden *kísérlet* fogad egy bemeneti adathalmazt és paramétereket, majd kimeneteként *adat* keletkezik. Ezt az adatot figyelembe véve különböző *cselekvések* közül választhatunk, amelyek *eseményekhez* vezetnek. Egy-egy ilyen esemény alapján további kísérletek elvégzése mellett dönthetünk. A rendszer felírható például a 14.2. ábrán látható valószínűségi gráf formájában. A munkafolyamat során a kutató a fa éleit követve mozog. Minden kimenetel egyfajta „értéket" képvisel a számunkra, ezt nevezzük *hasznosságnak*. Egy ésszerű stratégia mindig azon kísérlet elvégzése, amely a várható hasznosságot maximalizálja. Ez a gondolatmenet részletesebben kifejtve megtalálható Bernardo és Smith eredeti művében [2].

Jelölje e a kísérleteket, D_i az adatot, a a cselekvéseket, E_j az eseményeket, valamint U a hasznosságfüggvényt. Az átmenetek valószínűségi természetét figyelembe véve, egy

cselekvés várható hasznossága az $E_j \in \mathcal{E}$ események kiátlagolásával

$$EU(a, e, D_i) = \sum_j U(a, e, D_i, E_j) p(E_j | a, e, D_i).$$

Az a_i^* cselekvés, amely maximalizálja a várható hasznosságot

$$a_i^*\big|_{E_j \in \mathcal{E}} = \arg\max_{a \in \mathcal{A}} EU(a, e, D_i)$$

egyben az optimális döntés az $\mathcal{A}$ rendelkezésre álló cselekvésekre nézve minden (e, D_i)-re. Ekkor

$$EU(e, D_i) = EU(a_i^*, e, D_i).$$

Ismét hátralépve, a kísérlet várható hasznossága D_i kiátlagolásával

$$EU(e) = \sum_i EU(e, D_i) p(D_i | e),$$

ahol az utolsó tag a D_i adat likelihoodját jelöli adott kísérlet mellett.

Ezen a ponton új problémába ütközünk. Mikor hagyjuk abba a kísérletezést, és érjük be az eddig összegyűjtött tudással? Az orvosi etika egyik alapelve például kimondja, hogy csak olyan vizsgálatot szabad elvégezni, amelynek eredménye befolyásolja a beteg kezelését. A probléma akkor lenne megoldva, ha valamiképpen meg tudnánk mérni a jövőben összegyűjthető adat „befolyását". Pontosan erre ad lehetőséget az *adat várható értéke (EVD)* és a *kísérlet várható értéke (EVE)*.

Az e_0 nem elvégzett kísérlet várható hasznossága értelemszerűen

$$EU(e_0) = EU(a_0^*, e_0) = \max_{a \in \mathcal{A}} \sum_{E_j \in \mathcal{E}} EU(a, e_0, E_j) p(E_j | a, e_0).$$

Így a D_i, azaz a jövőben e kísérlettel megszerezhető adat várható értéke kiszámolható az e kísérlet elvégzéséből és nem-elvégzéséből fakadó várható hasznosságok különbségeként:

$$EVD(e, D_i) = EU(e, D_i) - EU(e_0).$$

Ez a mennyiség az **adat várható értéke** (Expected Value of the Data). A D_i adat kiátlagolásával megkapjuk a **kísérlet várható értékét** (Expected Value of the Experiment):

$$EVE(e) = \sum_{D_i \in \mathcal{D}} EVD(e, D_i) p(D_i | e).$$

14.3.2. Adaptív kísérlettervezés és költségkorlátozott tanulás

A valós kutatási folyamatokat szinte minden esetben megkötések terhelik – ez jelenthet financiális, időbeli, felszerelést illető, stb. kényszereket. A legtöbb esetben a cél a lehető legnagyobb mennyiségű információ megszerzése a költségvetés kimerüléséig. A *költségkorlátozott tanulás* (budgeted learning) és az *adaptív tanulás* szorosan kapcsolódó fogalmak, amelyeket kezdetben elsősorban a farmakológia és klinikai kísérletek területén használtak, és hagyományosan a mintaméret adaptív megválasztására törekedtek.

A 70-es évek vége óta egyre nagyobb hangsúlyt kapott a rögzített mintaszámú kísérletek kiváltásának lehetősége; e tanulmányok központi hibája ugyanis, hogy a rögzített mintaszámtól való eltérés nem lehetséges, így az adatok nem is hozzáférhetők egészen a kísérlet végéig. A gazdasági hátrányokon (pl. a feleslegesen nagy mintaszámból fakadó költségeken) túl etikai és adminisztratív hátulütők is megjelennek. Számos megközelítést javasoltak ezen hátulütők orvoslására (lásd pl. [3]):

1. **Csoport-szekvenciális módszer.** A csoport-szekvenciális (group-sequential) módszer az adatok fix időközönként történő megtekintését teszi lehetővé. Ha egy ponton a kísérlet sikeresnek bizonyul (megfelelő szignifikanciaszint elérésével), akkor a mintagyűjtés és egyúttal a kísérlet véget ér. Mivel azonban a szignifikáns eltérés *legalább egy csoportban* sokkal magasabb együttes elsőfajú hibát eredményezne, a nominális szignifikanciaszinteket minden megtekintésnél megfelelően korrigálni kell. Egy áttekintés a korrekció lehetőségeiről megtalálható a hivatkozott irodalomban [4].

2. **Alfa-költő megközelítések.** Az alfa-költő (alpha-spending) megközelítés az előző módszer kiterjesztésének tekinthető, amely megengedi az adatok irreguláris időközönkénti megtekintését is (azaz a csoportméretek eltérhetnek). Ebben a megközelítésben a megkövetelt együttes elsőfajú hiba mértéke előre rögzített, és az akkumulálódó elsőfajú hibát követjük (matematikailag: definiálhatunk egy „hibaköltő" függvényt, melyre $f(0) = 0$ és $f(t) = \alpha$) minden $t \geq 1$-re; minden megtekintésnél a nominális szignifikanciaszint e függvény alapján számítható).

3. **Whitehead trianguláris módszere.** Másképpen a **határ-módszer**, az előzőektől eltérően az adatok folyamatos megfigyelését igényli. Minden megtekintésnél két statisztika számolható; az egyik az aktív és a kontroll csoportok közötti különbséget, a másik ennek varianciáját mutatja. Ezeket egy 2D koordináta-rendszer tengelyeiként használva a felhalmozódó adat ábrázolható. A „sikert" és a „kudarcot" jelképező elméleti határok az előbbi koordináta-rendszerben egyenes vonalakként ábrázolódnak. Amennyiben az akkumulálódó adat metszi a felső határt, a kísérlet sikeres, ennek ellenkezője érvényes az alsó határra. A kísérlet addig folytatódik, amíg az adat a határok által bezárt folytatási régióba esik (amely tipikusan háromszög alakú, innen az eljárás neve).

4. **Sztochasztikus kizárás**. Ez a megközelítés a kísérlet várható kimenetelét becsüli. Amennyiben az elvárt szignifikanciaszint a jövőben beérkező mintáktól függetlenül elérhető, vagy épp ellenkezőleg, ennek valószínűsége kicsi, a kísérlet megállítható.

14.3. ábra. A Bayes-i szekvenciális kísérlettervezés munkafolyamata

A fenti módszerek közös előnye, hogy jobban illeszkednek a valós kísérletekhez (pl. rendszeres monitorozás), kényelmesebben alkalmazhatók, valamint lehetővé teszik a korai leállítást, ami alacsonyabb mintaszámhoz és rövidebb tanulmányokhoz vezet.

14.3.3. Szekvenciális döntési folyamatok Bayes-i keretben

A Bayes-i statisztikai eszköztár és a Bayes-hálók igen jól használhatók szekvenciális döntési folyamatok modellezésére. Az elmúlt években sok kutatás célozta meg a Bayes-i keretrendszer további kiterjesztését, például informatív priorok és hasznosságfüggvények konstruálásával, párhuzamos számításokkal, illetve egyéb, korábban nem kapcsolódó eljárások (pl. génprioritizálás) integrálásával. Ebben az alfejezetben bemutatunk egy adaptív technikát, amely alkalmas kísérletsorozatok tervezésére – ehhez minden lépésben a legígéretesebb változókat (pl. SNP-ket) választja ki, így viszonylag nagy mintaszámot biztosít adott költségvetés esetén. A leírás során felhasználjuk az előző alfejezetekben bemutatott eszköztárat és a Bayes-i megközelítést. A módszert először az asthma genetikai hátterének felderítésére használták PGAS adatokon [5].

Az alapötlet *relevancia-analízisek* (olyan változók azonosítása, amelyek szorosan kapcsolódnak a kísérlet tárgyához, pl. egy fenotípushoz) és *változó-kizárások* (variable pruning) iteratív alkalmazása. A munkafolyamatot a 14.3. ábra mutatja. Először egy kezdeti jelölt változóhalmaz kerül kiválasztásra a mérési adatok és szakértői tudás alapján (illetve egyéb eszközök alkalmazásával, pl. keresők, prioritizálók, szövegbányász eszközök, stb.). A jelölt változóhalmaz ezután egy ciklusba kerül, ahol kísérleteken, relevancia-analíziseken és változó-kizárásokon megy át; ennek során az algoritmus a legnagyobb várható hasznosságú változókat tartja meg. Minden iteráció után egy döntés történik a kísérletek folyatására vagy leállítására; utóbbi során megtörténik a relevánsnak ítélt változóhalmazok „jelentése”.

Az alábbi leírás a módszert kifejlesztő csoport közleményét követi [5]. Tekintsük az

$f \in \mathcal{F}$ strukturális jegyeket és a posteriort az $\mathcal{F}$ jegytér felett, az i. lépésben meglévő $D_{<i}$ tudásunk mellett. Az f^* optimális jelentett jegy megállapítható az egyes $\hat{f}$ jegyek jelentésének várható hasznossága feletti maximalizálásával:

$$f^* = \arg\max_{\hat{f}} E_{p(f|D_{<i})}\left[U(\hat{f}|f)\right].$$

Minden lépésben dönteni kell a kísérletek folytatásáról vagy leállításáról. Utóbbi esetben az eddigi lépések hasznossága, $U(D_{<i})$, megegyezik az optimális jelentés hasznosságával; folytatás esetén $U(D_{<i})$ a várható adat hasznosságaként határozható meg.

$$U(D_{<i}) = \max(U^R(D_{<i}), U^C(D_{<i})) = \max\left(E_{p(f|D_{<i})}\left[U(f^*|f)\right], E_{p(D_i|D_{<i})}\left[U(D_{\leq i})\right]\right).$$

Megjegyzendő, hogy $U(D_{\leq i})$ becsülhető a $U^R(D_{\leq i})$ jelentés hasznosságával. Ezek után az egyetlen hiányzó elem maga a hasznosságfüggvény. Ahogy azt a fenti egyenlet rekurzív definíciója jelzi, előbb-utóbb egy direkt pontozófüggvényre lesz szükségünk. Legyenek tehát az f strukturális jegyek változóhalmazok és jelölje $\mathcal{S}$ a változóhalmazok halmazát. A direkt pontozófüggvény $U^D : \mathcal{S} \to \mathbb{R}$ a következőképpen definiálható:

$$U^D(s) = \sum_{v \in s} \lambda_V V(v) + \lambda_S S(s) + \lambda_G G(s),$$

ahol $V(v)$ az s változóhalmaz egy elemének MBM-pontszáma, $S(s)$ a halmaz MBS-pontszáma és $G(s)$ a halmaz MBG-pontszáma (ezek definíciója és további tudnivalók a Bayes-i többszintű elemzésről megtalálható a hivatkozott irodalomban [6]).

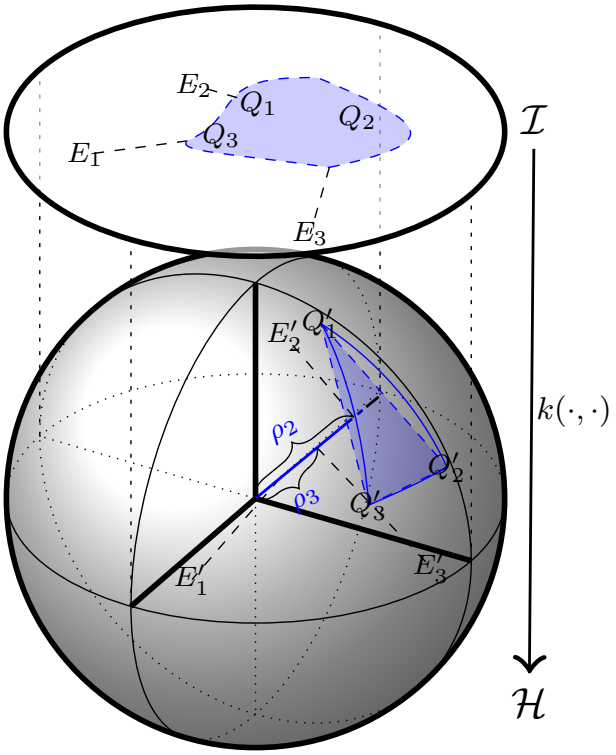
14.4. A célváltozók kiválasztását szolgáló módszerek

14.4.1. Génprioritizálás

A génprioritizálás sorrendi tanulási feladat, amelynek célja egy adott lekérdezéshez legrelevánsabb entitások megtalálása. Gondolhatunk rá egyfajta „orvosbiológiai Google"-ként, ahol a lekérdezés állhat betegségekből, betegség-génekből, kulcsszavakból, stb. A prioritizáló rendszer kimenete a gének egy relevancia szerint rendezett sorrendje. Ahogy a heterogén „omikai" információforrások integrációja egyre inkább bekerült a köztudatba, szoros kapcsolatok alakultak ki a génprioritizálás és az adatfúzió területei között is.

Bár a legtöbb prioritizáló rendszer páronkénti hasonlóságokat, illetve hálózat-alapú megközelítéseket használ, más módszerek is napvilágot láttak, pl. sorrendi statisztikai [7] illetve Bayes-háló alapú megközelítések [8]. Számos rendszer leírása megtalálható egy 2011-es összefoglaló közleményben [9]. Ebben a fejezetben bemutatunk egy hasonlósági génprioritizáló rendszert, amely ún. szupport-vektor gépekre (SVM) épül.

A könnyebb érthetőség érdekében a génprioritizálást egy gyakorlati példán keresztül vizsgáljuk meg. Tegyük fel, hogy olyan géneket keresünk, amelyek valamilyen szerepet töltenek be a sejtciklus szabályozásában. Ehhez rendelkezésre állnak génexpressziós profilok



14.4. ábra. SVM-alapú prioritizálás. A lekérdezést Q, a többi entitást E jelöli. Az entitások a szaggatott kék vonallal jelölt felülettől való távolságuk alapján vannak sorrendezve. A felületet a hasonlóságok által meghatározott transzformált térben számítjuk ki.

microarray-vizsgálatokból, valamint ismerünk proto-onkogéneket (lekérdezés). Feltesszük továbbá, hogy „hasonló" expressziós profillal rendelkező gének többé-kevésbé azonos funkciót látnak el. Ezen a ponton meg kell határoznunk a „hasonlóság" fogalmát, amelyhez számtalan hasonlóságmérték közül választhatunk – ez a választás egyben a szakértői tudás bevitelének egyik módja is. Az ún. egyosztályos szupportvektor-gép a hasonlóságok által meghatározott matematikai térben egy olyan felületet számít ki, amely a lehető legnagyobb margóval elválasztja a lekérdezést a többi géntől. A következő lépésben a gének sorrendezhetők a felülettől való távolságuk alapján; minél kisebb a távolság, annál valószínűbb, hogy a gén szerepet játszik a sejtciklusban (14.4. ábra).

További részletek az egyosztályos és ν-SVM-ről megtalálhatók az eredeti közleményben [10]. Az egyosztályos SVM primál feladata a következőképp írható:

$$\begin{aligned} \min_{\boldsymbol{w},\boldsymbol{\xi},\rho} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} - \rho + \frac{1}{\nu l}\sum_i \xi_i \\ s.t. \quad & \boldsymbol{w}^T\phi(\boldsymbol{x_i}) \geq \rho - \xi_i \\ & \boldsymbol{\xi} \geq 0, \quad i = 1, 2, ..., l \end{aligned}$$

ahol a célfüggvény első tagja a modell simaságát biztosítja, ρ jelöli a margót, ν szabályozza a komplexitást és ξ_i a soft-margin formalizációhoz szükséges slack változók. $\phi(\cdot)$ képezi le a mintákat a $\mathcal{H}$ reprodukáló kernel Hilbert-térbe, azaz $K(\boldsymbol{x_i}, \boldsymbol{x_j}) = \langle \phi(\boldsymbol{x_i}), \phi(\boldsymbol{x_j}) \rangle_{\mathcal{H}}$. A duál

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \mathcal{D}(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}^T K \boldsymbol{\alpha} \\ s.t. \quad & 0 \leq \boldsymbol{\alpha} \leq 1, \quad \mathbf{1}^T \boldsymbol{\alpha} = \nu l. \end{aligned}$$

A prioritizáció során az origótól számított hipersíkra ortogonális távolság:

$$f(\boldsymbol{x}) = \frac{\sum_i \alpha_i K(\boldsymbol{x_i}, \boldsymbol{x})}{\sqrt{\boldsymbol{\alpha}^T K \boldsymbol{\alpha}}}$$

ahol a nevező a normalizációért felel, a konstans ρ paramétert pedig elhagyjuk.

14.4.2. Aktív tanulás

Tekintsük a fenti keretet egy apró módosítással. Tegyük fel, hogy rendelkezésünkre állnak a gének és az expressziós profilok, de semmit nem tudunk a funkciókról, így annak felfedéséhez, hogy egy adott gén rendelkezik-e az általunk vizsgált funkcióval, külön kísérlet szükséges. A célunk, hogy a funkcióval rendelkező géneket találjunk megfelelő pontossággal és relatíve kisszámú kísérlettel. Ez a feladat a gyógyszerkutatási folyamatra emlékeztet, ahol a cél aktív vegyületek felfedezése hatalmas molekuláris könyvtárakban. 2003-ban Warmuth egy elegáns keretrendszert javasolt ilyen problémák kezelésére, amely az aktív tanulás fogalmán alapult. Az aktív tanulás egy iteratív folyamat, amely a következő lépésekkel írható le:

1. Modellépítés egy kezdeti mintahalmaz alapján (a mérettel megegyező számú kísérlet elvégzése szükséges).
2. Eddig ismeretlen minták kiválasztása valamilyen kritérium alapján, majd címkéjének felfedése (ismét egy kísérlettel).
3. A modell finomítása az eredmény alapján.
4. A 2-3. lépések ismétlése konvergenciáig.

Esetünkben két ésszerű kiválasztási stratégia lehet a felülethez legközelebbi, vagy éppen a legtávolabbi gén kiválasztása („belül", azaz a felület „pozitív" oldalán!). Az előbbi választás az ún. Minimum Marginal Hyperplane eljárások alapja, amely végeredményben azokat a mintákat válogatja be, amelyekben a modellünk a leginkább bizonytalan, majd az ilyen „határesetek" megvizsgálásával javít a modellen. Az utóbbi stratégia (Maximum Marginal Hyperplane) a biztosnak ítélt predikciók felülvizsgálatán alapul. Egyéb kiválasztási stratégiákat és ezek viselkedését Warmuth eredeti közleményében láthatunk [11]. Az „aktív" kifejezés az adatok aktív felfedezését jelenti, szemben az előző algoritmusokkal, amelyek egy statikus tanítóhalmazt használtak ismert címkékkel. Szintén vegyük észre az algoritmus szekvenciális természetét, ami más fogalmakhoz, például a szekvenciális kísérlettervezéshez vagy az adaptív kísérlettervezéshez való kapcsolatot sugall.

14.5. Egyéb, a gyakorlatban felmerülő bioinformatikai feladatok

A korszerű kísérlettervezés elképzelhetetlen lenne bioinformatikai támogatás nélkül. A legfontosabb, bioinformatikára erősen támaszkodó lépések a következők:

- **Irodalomkutatás.** A szakirodalom feldolgozása és a releváns ismeretek kinyerése ma már jelentős mértékű bioinformatikai támogatással történik. A legelterjedtebb keresőmotorok (pl. PubMed) rengeteg szolgáltatást nyújtanak, ideértve a szűrési és rendszerezési eljárásokat, idézési segédeszközöket, alkalmazási programozási felületeket (API), stb. Emellett több, félig vagy teljesen automatizált szövegbányász rendszer is a kutató rendelkezésére áll.

- **Minta- és adatgyűjtés.** A kísérletben résztvevők kérdőíveinek elkészítése, kiküldése, begyűjtése és feldolgozása (esetleg elektronikus felület biztosítása), valamint a mintaazonosítás és -szállítás mind-mind erős informatikai hátteret követelnek meg.

- **Tárolási feladatok.** A fizikai mintatárolás rendszerint elektronikus készletnyilvántartó rendszerekkel egészül ki. Hasonlóképpen, az adattárolás, pl. bemeneti és mérési adatok szabványos tárolása és elérése is korszerű adatbázis-rendszerekkel valósítható meg.

- **Biztonság.** Az adatbiztonság mind jogi, mind etikai szempontból kritikus fontosságú. Kapcsolódó fogalom a megosztott hozzáférés, amely a kísérletet végző, különböző feladatokkal megbízott személyek tevékenységének összehangolását teszi egyszerűbbé. Szintén a biztonsághoz kapcsolódik a minőségbiztosítás kérdésköre.

Irodalomjegyzék

[1] W. Ahrens and I. Pigeot, *Handbook of Epidemiology*. Springer, 2007.

[2] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*. Wiley Series in Probability and Statistics, John Wiley & Sons Canada, Ltd., 2007.

[3] S. Senn, *Statistical issues in drug development*. Wiley-Interscience, 2007.

[4] C. Jennison and B. W. Turnbull, *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC Interdisciplinary Statistics, Taylor & Francis, 1999.

[5] P. Antal, G. Hajós, A. Millinghoffer, G. Hullám, Cs. Szalai, and A. Falus, Variable pruning in Bayesian sequential study design. *Machine Learning in Systems Biology*, page 141, 2009.

[6] Péter Antal, András Gézsi, Gábor Hullám, and András Millinghoffer, Learning complex bayesian network features for classification. In: *Proc. of third European Workshop on Probabilistic Graphical Models*, pages 9–16, 2006.

[7] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L. C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, P. Carmeliet, and Y. Moreau, Gene prioritization through genomic data fusion. *Nat. Biotechnol.*, 24:537–544, May 2006.

[8] A. Parikh, E. Huang, C. Dinh, B. Zupan, A. Kuspa, D. Subramanian, and G. Shaulsky, New components of the Dictyostelium PKA pathway revealed by Bayesian analysis of expression data. *BMC Bioinformatics*, 11:163, 2010.

[9] L. C. Tranchevent, F. B. Capdevila, D. Nitsch, B. De Moor, P. De Causmaecker, and Y. Moreau, A guide to web tools to prioritize candidate genes. *Brief. Bioinformatics*, 12:22–32, Jan 2011.

[10] Bernhard Schölkopf, John C. Platt, John C. Shawe-Taylor, Alex J. Smola, and Robert C. Williamson, Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13:1443–1471, July 2001.

[11] M. K. Warmuth, J. Liao, G. Ratsch, M. Mathieson, S. Putta, and C. Lemmen, Active learning with support vector machines in the drug discovery process. *J Chem Inf Comput Sci*, 43(2):667–673, 2003.

15. fejezet

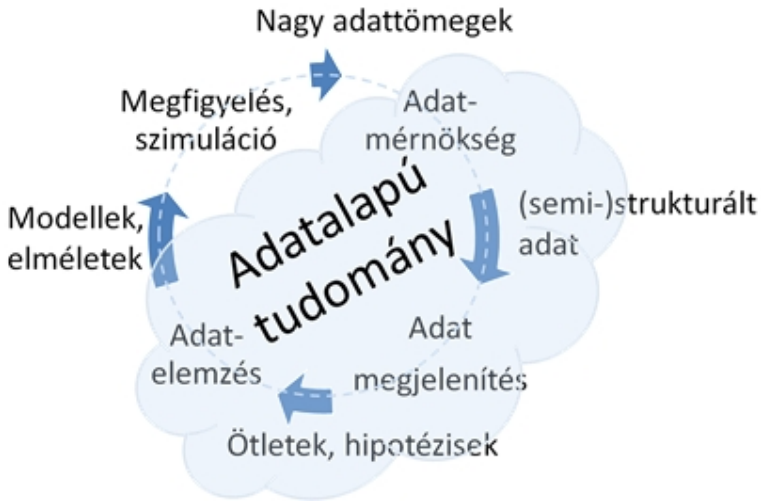
Nagy adattömegek az orvosbiológiában

Amelyben áttekintjük a biológiában megjelenő nagy adattömegek első hullámába tartozó szekvencia, strukturális és expressziós adatokat, majd összefoglaljuk ezek egyre heterogénabb, ám még mindig akadémiai forrású második hullámát. Ezt követően áttekintjük a jelenleg formálódó a mindennapi életből származó nagy adattömegek forrásait az internettől a hordható elektronikai eszközökön át az otthoni egészségmonitorozó rendszerekig. Megvizsgáljuk ezek orvosbiológiai relevanciáját, illetve fordítva is a nagy adattömegekre kifejlesztett módszerek orvosbiológiai adatokhoz való adekvátságát. Végezetül megvizsgáljuk, hogy ezen adatok alapján milyen betegoldali és orvosoldali adatelemzési igény és döntéstámogatás is várható.

15.1. Bevezető

Az 1965-ben G. Moore által megfogalmazott törvény a tranzisztorok sűrűségéről az elektronika egy általános törvényévé vált, amely az eredeti fizikai alapoktól elválva a számítási teljesítmény és adattárolás sokféle vonatkozásában is helyénvalónak bizonyult. Az adattárolás fejlődésével párhuzamosan a méréstechnika is exponenciális fejlődési szakaszon ment keresztül pl. a csillagászat, meteorológia, részecskefizika, kémia, molekuláris és neurobiológia területén. A felhalmozódó adatok miatt az ezredforduló tudománytörténeti korszakhatárnak is tekinthető, amikor a XX. század második felére jellemző számításintenzív, szimulációs korszakot egy adatintenzív, adatelemző korszak váltotta fel. Tudománytörténészek egy új kutatási paradigma, az e-science megjelenését is vizionálták, amelynek központi eleme ezen nagy adattömegek léte, hatékony begyűjtése, tárolása, elemzése és modellalkotásban, kísérlettervezésben való felhasználása. Fontos felismerni azonban, hogy az adatgazdagság nem kizárója, hanem csak megelőző fázisa a számításintenzív szimulációknak, így ezek megjelenése egyre komplexebb területeken várható.

Az e-science paradigma elméleti és gyakorlati háttere több tudományterületen is elosztva fejlődik, amely a következő kulcsszavak köré szerveződik: (1) a számításintenzív szimuláció, (2) a nagy adattömegek, a „Big Data”, (3) közösségi kutatás, (4) a nyílt elérés, hatékony kombinálhatóság, újrafelhasználhatóság.



15.1. ábra. A kutatási ciklus nagy adattömegeknél

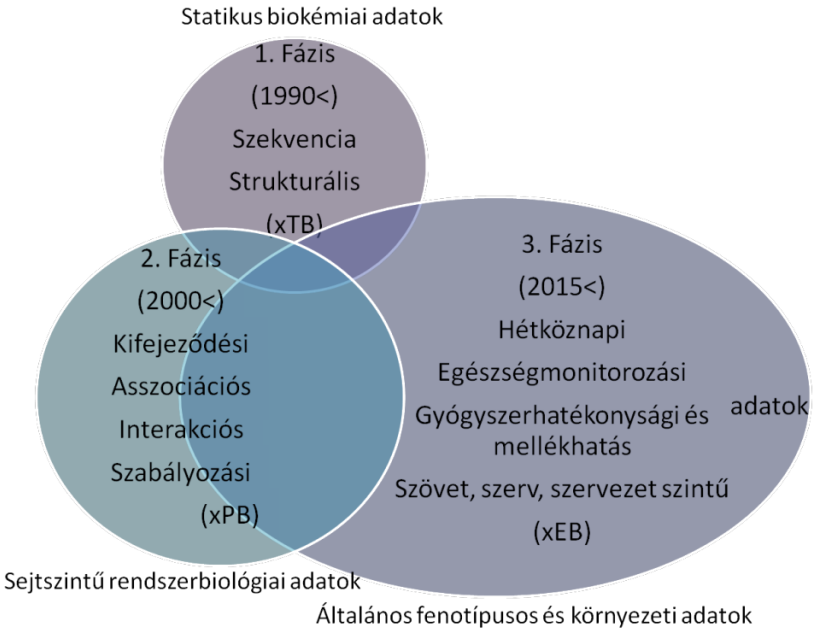
A nagy „adattömeg/adatbőség/adattenger/adatlavina" („Big Data") meghatározó elem az e-science vonatkozásában, és a (Big) „Data Science" kifejezést így az e-science szinonimájaként is használt. Az orvosbiológiában jelenlévő, megjelenő és várhatóan megjelenő nagy adattömegek azonban speciális sajátosságokkal bírnak a nagy adattömeg („Big Data") megszokott definícióihoz képest, bár a fenotípusos adatok fontosságának előtérbe kerülésével a hétköznapi nagy adattömegek orvosbiológiai felhasználása is egyre fontosabb. A fejezetben ezt a kérdést vizsgáljuk meg több szempontból is.

Érdemes észrevenni, hogy más tudományterületeken, mint a fizika, csillagászat vagy klímakutatás területén a harmadik, mindennapokból származó nagy adattömegek bekapcsolódása nem indokolt, így ez unikális az orvosbiológiára.

15.2. Az orvosbiológia klasszikus nagy adattömegei

A biológiai, biokémiai adatok évtizedeken át meghatározó forrása a fehérjetérszerkezet-adatok voltak, azonban a Humán Genom Program indulásával a genetikai szekvenciaadatok mennyisége vált meghatározóvá. A génexpressziós adatok ezredfordulón bekövetkezett méréstechnikai fejlődésével a biológiai adatok három fő területe kialakult, amelyek a strukturális, a szekvenciákra vonatkozó, és a kifejeződésekre vonatkozó adatok. Érdekes, hogy a molekuláris biológiai méréstechnika fejlődése is jellemezhető a számítástechnikából jól ismert Moore-törvény szerint, amely alapján a molekuláris biológiai adatok mennyisége éves nagyságrendben megduplázódik [1, 2]. Ezen Carlson-törvények szerint a DNS szintézis és szekvenálás produktivitásának növekedése is jellegében a Moore-törvényéhez viszonyítva változik, illetve a fehérje-térszerkezetek meghatározási idejének változása is.

Az autonóm omikai szinteknek megfelelően a génexpressziós kifejeződési szinttel analóg módon megjelentek a (kvantitatív) transzkripciós, proteomikai, lipidomikai, metabolomikai szintek is, önálló ontológiákkal és adattárházakkal. A gyógyszerkutatásban betöltött



15.2. ábra. Az orvosbiológiai nagy adattömegek három hulláma

szerepe miatt önálló, a bioinformatikai kutatásoktól kissé elváló utat járt be a hatóanyagok és gyógyszerek reprezentálásának és adatbázisainak fejlődése.

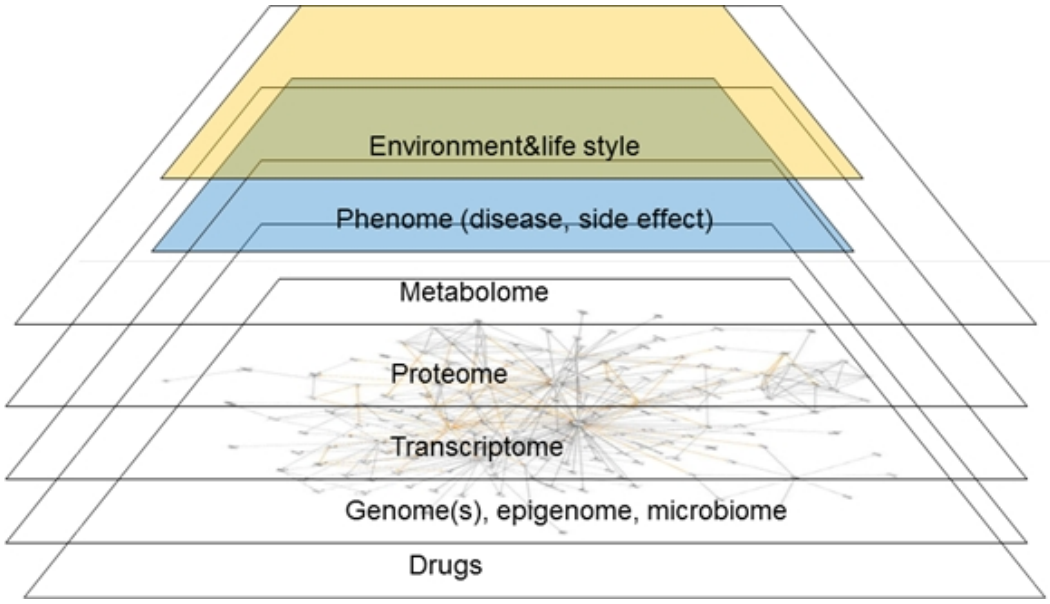
Gyors, bár ehhez nem fogható növekedési jelleget mutat az orvosbiológiai szakcikkek számának gyarapodása is.

A molekuláris entitásokról szóló adatok mellett a „páronkénti" adatok, mind a génszabályozási vonalán, mind a fehérje-fehérje interakciók kapcsán, illetve a genetikai variánsok és betegségek kapcsán a genetikai asszociációs adatbázisok. A hatóanyagok és gyógyszerek adatbázisainak fejlődése jól tükrözte a kemoinformatika önálló fejlődését, hogy az orvosbiológiai nagy adattömegek megjelenésének első hullámában, az ezredfordulóig, a gyógyszer-célpont adatbázisokon túl, a gyógyszer-betegség relációban nem jöttek még létre nagy mennyiségben adatok.

15.3. Posztgenomikai nagy adattömegek az orvosbiológiában

A Humán Genom Program lezárulása után, amely egy többé-kevésbé lezárt referenciaszekvenciát eredményezett, a genetikai variánsok feltérképezésére helyeződött a hangsúly. A genotipizálás és génszekvenálás elérhetősége folyamatosan javult a Carlson-törvényeknek megfelelően, amelyeknek érvényessége csupán napjainkban, 2013-ban látszott sérülni. A létrejött új generációs szekvenálási módszerek felhasználásával új programok indultak, amelyek több ezer teljes emberi genomot határoztak meg.

A nagy adattömegek elérhetősége ellenére azonban mind a diagnosztikai biomarkerek



15.3. ábra. Omikai szintek

felfedezésében, mind a gyógyszerkutatásban az elért eredmények elmaradtak az ezredfordulón még fenntartott várakozásoktól. Az elmaradt eredmények magyarázatára több javaslat is megjelent, amelyek egy része az ezredforduló után felismert jelentőségű új omikai szintekhez, mint például a microRNS-ek szintjéhez vagy az epigenetikai módosulások szintjeihez kapcsolódott.

Új leíró szintekre példa, amit az új generációs szekvenálási eljárások tesznek lehetővé, a mikrobiális vizsgálatokat segítő metagenomikai vizsgálatok, amelyek akár az emberi szervezet egy bakteriális ökoszisztémával kialakított szimbiózisát is képesek vizsgálni. Ennek jelentőségét az adja, hogy az emberi szervezetben 10^{14} nagyságrendű baktérium él, meghaladva az emberi sejtek számát is [3, 4].

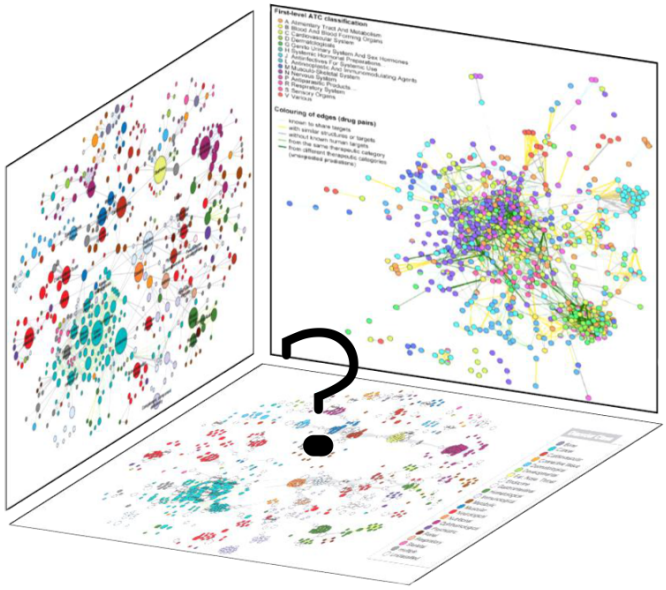
Egy másik, szintén az új generációs szekvenálási eljárások által lehetővé vált módszer az immunrendszer karakterizálását végzi el a T és/vagy B sejtek repertoárjának felmérésére, a sejtek immunológiailag releváns szekvenciarégióinak feltérképezésével. Ezek számossága szintén 109-es nagyságrendet meghaladó lehet, amelyek követése autoimmunbetegségekben rendkívül ígéretes.

Más magyarázatok szerint a relációk és mechanizmusok többváltozós és kontextuális jellege nehezíti a felfedezést. Érdekes módon a magyarázatok egy része magát az omikai megközelítést, a hipotézismentes kutatási paradigmát is támadta. Ezeknek az alapja a többszörös hipotézistesztelési problematika, amely szerint még egyváltozós statisztikai asszociációs elemzésekben is a változók, pontosabban a független statisztikai tesztek szerint a hibás felfedezés kontrollálása miatt a statisztikai tesztek érvényességének az elfogadását egyre szigorúbb kritériumokhoz kell kötni. Többváltozós, akár interakciót is megengedő modelleknél a lehetséges tesztek száma a változók számában akár egy igen gyorsan növő függvény is lehet, amely a többszörös hipotézistesztelés problemáját még inkább súlyos-

bítja. Bár kezelésére több statisztikai módszertan is megjelent, az alapvető problémát az adatok viszonylagos, a modellek sokaságához, komplexitásához viszonyított volta jelenti. Ennek megfelelően az adatok és a meglévő a priori tudás fúziója került az előtérbe, nevezetesen a heterogén omikai szintek kapcsolódására vonatkozó és egyéb háttérinformációk integrálása. Ennek egy olvasata, hogy az orvosbiológia adatgazdagsága viszonylagos, és a nagy mennyiségű háttérismeret felhasználása elengedhetetlen, amelynek rendszer alapú elemzése jelenthet segítséget a statisztikai aluldeterminálttsággal szemben. A rendszer alapú megközelítés sok tekintetben kötődik a beavatkozásokhoz, autonóm mechanizmusokhoz, oksági modellezéshez, amelynek matematikai alapjainak fejlődése az utóbbi negyedszázadban rohamos fejlődésen ment át [5].

Az orvosbiológiában megjelenő posztgenomikai nagy adattömegek második hulláma ezen rendszerszintű megközelítéssel is jellemezhető, azaz olyan szisztematikus vizsgálatok, amelyek beavatkozásokhoz, szabályozások, autonóm mechanizmusok feltértékepzéséhez kapcsolódnak.

A hatóanyag/gyógyszer-génexpresszió-betegség/genetikai profil/szövet hármas együttes megközelítése miatt a Connectivity MAP volt annak első példája, amely egy molekulakönyvtárat különböző sejtvonalakon alkalmazva azok transzkripciós, illetve egyéb omikai profiljait vizsgálta. Egy specifikusabb követője ennek a hatóanyag -expresszió-sejttípus hármasok szisztematikus szűrési paradigmának a Genomics of Drug Sensitivity in Cancer.



15.4. ábra. A hatóanyag -expresszió-sejttípus integrált adatok felhasználásának problémája: gyógyszer-gyógyszer, gén-gén, betegség-betegség kapcsolati hálók

Teljessége miatt szintén kiemelkedik az ENCODE projekt, amely különböző transzkripciós faktorok kötőhelyeit térképezi fel szisztematikusan, epigenetikai térképeket is alkotva, szövetspecifikusan.

Az orvosbiológiai nagy adattömegek egy speciális szegmensét alkotják a most bein-

duló agykutatási programok, amelyek jelentősége a remények szerint a Humán Genom Projekthez hasonló lesz, és többléptékű adatok sokaságát fogja eredményezni: az idegsejt membránpotenciáljától az agyi képalkotó eljárások kimeneteléig. Ennek kapcsolódása a genomikai kutatásokhoz több ponton is várható, különösen a következőkben tárgyalt komplex fenotípus kapcsán.

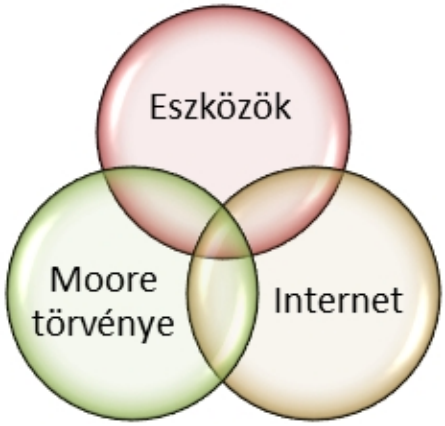
Végezetül a számításintenzív szimulációkból származó adatokat említjük meg, amely forrás szerepeltetése meglepőnek tűnhet, különösen a „4. tudományos kutatási paradigma" korszakában [6]. A 2. paradigmának nevezett analitikus egyenletrendszerek, majd a 3. korszaknak nevezett számításintenzív szimulációk után a jelenlegi korszak adatvezérelt, amelyben az adat azonban valóban a *lingua franca*, amely származhat mind valós megfigyelésekből, mind adott valósághűségű számításintenzív szimulációkból. Fontos felismerni, hogy a molekuláris biológiai, biokémiai méréstechnika fejlődése mellett az ismeretek gyarapodása és számítási kapacitások bővülése is olyan mértékű, hogy sok esetben alternatívaként jelenik meg az adott pontosságú, költségű és infrastrukturális igényű valós mérés és számításintenzív szimuláció. Ez különösen igaz az általános célú grafikus kártyák (GPU) fejlődésével és a számítási közmű/felhő egyre általánosabb elérhetőségével. A sejt, szerv, szervrendszer, teljes szervezet modellezésének az ismeretek további gyarapodása és a számítási erőforrások további növekedése mellett a többszintű szimulációs eszközök fejlődése adhatna újabb lendületet.

A nagyméretű, kvantitatív modellek szimulációja, különösen ezen modellek nagyszámú, populációszintű futtatása rendkívül nagy számításigényt jelenthet, viszont a beavatkozás lehetősége miatt ez unikális, a valóságban nem kivitelezhető megfigyeléseket biztosít. Az így keletkezett adatok adott pontosságú tárolása ugyanúgy kérdés, mint a valós adat esetén, hiszen ez az adat is az előállítási költsége mellett a kiszámításához szükséges időt is jelenti, akár valós vagy szimulált környezettel, és egyfajta prekompilált, disztillált tudás is sok esetben.

15.4. Hétköznapokból származó nagy adattömegek

A tárgyalt tudománytörténeti váltásnak megfelelően a nagy adattömegek megjelentek a nukleáris fizikában, majd a molekuláris biológiában, csillagászatban, klímakutatásban, a most induló agykutatási programokban is. Az akadémiai megjelenés mellett a nagy adattömegek a kereskedelemben, iparban és a mindennapokban is megjelentek természetesen. Kezdve a banki tranzakciókkal, majd az elektronizáció és internet terjedése, illetve a beágyazott elektronikai eszközök miatt megjelentek a mobiltelefon-adatok, felhasználói adatok (klikkek sorozatától a feltöltött fényképekig és videókig), email-adatok, blogok, internetkeresési adatok, társasági hálózati adatok. Emellett az idősek és betegek otthoni életvitelét támogató rendszerek, az egészségmonitorozó rendszerek, a viselhető elektronikai rendszerek, a kiberfizikai rendszerek, intelligens otthonok, szenzorhálózatok is egyre nagyobb tömegű adatot szolgáltatnak. Ezen hétköznapi nagy adattömegek megjelenését a Moore-törvény, az elektronikus eszközök és az internet hármasa biztosította, és meghatározó sajátossága az egybemosódó fizikai-informatikai világ (*E. Dumbill: Making sense*

of big data, Big Data, vol. 1, no. 1, 2013).



15.5. ábra. Hétköznapi nagy adattömegek prekurzorai

A felhasználók számának növekedésével ezek összességükben az akadémiai nagy adattömegek mennyiségét messze felülmúló értéket képesek generálni, amely azonban, mint látni fogjuk, összekapcsolható akár kutatási céllal is az akadémiai adattömegekkel. Ezen hétköznapi „nagyon nagy" adattömeg megjelenéséhez valószínűleg a számítási/adattárolási közmű szolgáltatásának fejlődése is szükséges, azonban ennek tárgyalása kívül esik a jegyzet keretein, így a jelenlegi szintű hétköznapi adattömegeket tételezzük fel. A hétköznapi adattömeg/„Big Data" megjelenése például a következő területeken már megszokott:

1. Pénzügyi és tőzsdei tranzakciók (előrejelzés, visszaélés-felderítés)
2. Telefon (híváshálló elemzése célzott reklámhoz, visszaélés-felderítés)
3. Szoftverhasználat (használhatóság elemzése szoftverhasználati jogok alapján)
4. Webes keresés (hivatkozásstruktúra elemzése)
5. Webhasználat (weboldal felépítésének optimalizálása)
6. Járműforgalom-elemzés (GPS-ek alapján, terhelésoptimalizálás, dugóelkerülés)
7. Villamosenergia-hálózat mérése (predikció)
8. Növénytermesztés (visszaélés-felderítés műholdkép-elemzéssel)

Annak megértéséhez, hogy az orvosbiológiai nagy adattömegek és az ipari, kereskedelmi, hétköznapi nagy adattömegek miben is hasonlóak és eltérőek, és így a rájuk kifejlesztett eszközök miben is mások, vizsgáljuk meg a nagy adattömegek meghatározását. A „Big data" kifejezés első használata a megszokot értelmű, akkori informatikai kereteket meghaladó adatra vonatkozott [7], ami 2001-ben egy igen állandósult 3xV váltott fel: **volume, variety, and velocity** (2001). A rengeteg definíció között egy orvosbiológiai szempontból releváns a következő:

> „[big data] ... **represents the totality or the universe of observations. That is what qualifies as big data. You do not have to have a hypothesis in advance before you collect your data.** You have collected all there is–**all the data there is about a phenomenon."**
> (*E.Dumbill: Making sense of big data, Big Data, vol. 1, no. 1, 2013*)

amely a megszokott omikai definíció. Egy gyakran előforduló megkülönböztetés az ipari, kereskedelmi és mindennapi életből származó „big data" és az akadémiai, speciálisan a bioinformatikai, kemoinformatikai nagy adattömegek között az előbbiek időbeliségén, pontosabban a felhasználásuk időbeliségén alapszik. Amire egy példa azon feladat, hogy küldjünk egy olyan célzott elektronikus üzenetet azoknak a felhasználóknak, akik egy bizonyos helyszínen tartózkodnak, adott termékről tudnak és kommunikációs/kapcsolati hálóikban ezzel kapcsolatos aktivitásuk valamilyen értelemben központi szerepet tölt be.

A gyors reakciójú felhasználás alapján történő megkülönböztetés ellenére a hétköznapi nagy adattömegek az élet egyre kiterjedtebb részét fedik le, információtartalmuk egyre nő, így az orvosbiológiai kutatások egyre inkább relevánsak, akár orvosbiológiai nagy adattömegekként is tekinthetők. A hétköznapi nagy adattömegek orvosbiológiai relevanciáját orvosbiológiai és gyógyszerkutatási oldalon bekövetkező változások is segítik, amit a következőkben tekintünk át.

15.5. A hétköznapi nagy adattömegek az orvosbiológiában

A már tárgyalt reménttől elmaradó sikerességére a genetikai asszociációs kutatásoknak több magyarázat is az asszociáció leírásának elégtelen voltát emelte ki. Magának a fenotípusnak a leírása is kritika tárgya, például a sok betegségben megszokott eset-kontroll bináris felbontást elégtelennek, finomabb felbontást viszont már szakmailag szubjektívnek tartanak, molekuláris biológia végpontokkal történő karakterizálás pedig legtöbbször csak kutatási célként létezik. Hasonlóan biomarkereknél a kontextus részletesebb leírása is fontos volna, azaz a potenciálisan módosító tevékenységek és a környezet leírása. Kapcsolódó metodológiai változás, hogy a célzott eset-kontroll elemzések helyett a nagy kohorsz-vizsgálatok lesznek preferálva, aminek statisztikai mintaszám okai is vannak.

A fenotípusadatokon és hétköznapi nagy adattömegeken belül különös fontosságra tettek szert a gyógyszerfogyasztással és gyógyhatású készítmények használatával kapcsolatos információk. Ezek az alap-orvosbiológiai kutatások mellett akadémiai gyógyszerkutatások, gyógyszeripari kutatások, népegészségügyi kutatások, de egészségbiztosítási vizsgálatok szempontjából is vitális információkat hordoznak eredményességről és hatékonyságról, illetve mellékhatásokról. A mellékhatások szisztematikus és átfogó követésére több európai program is indult, amely a gyógyszerkutatásokban egy új korszakot nyithat. Ettől független, de megjegyzendő, hogy a gyógyszeripar stratégiai megváltozására az is példa, hogy több gyártó a molekulakönyvtárának és azokon végzett kutatásainak bizonyos fokú

kinyitására készül, illetve, hogy az engedélyeztetési eljárásban keletkező adatok nyers formájukban is elérhetőek lesznek. Hasonló érdeklődés az élelmiszerbiztonság és a kémiai biztonság irányából is várható.

Végezetül a legalapvetőbb tényező a hétköznapi nagy adattömegek orvosi felhasználása mellett maga az egyének önmegismerő és egészségmegőrző törekvése. Bár a hordható elektronikai eszközök az ezredfordulótól folyamatosan a tömeggyártás és tömeges elterjedés határán vannak, a társadalmi szintű fogékonyság és ipari felkészültség több felmérés szerint is most fog egy kritikus szintet elérni.

A viselhető („wearable") számítástechnika, a beágyazott, transzparens számítástechnika („ambient assisted living") miatt várhatóan további új adatforrások is megjelennek a közeljövőben (1–5 év), mint például a következők:

1. Testszenzorok, okosóra: folyamatos orvosi alapadatok és hanginformációk teljes körű potenciális rögzítése.
2. Okos szemüveg: vizuális információk teljes körű potenciális rögzítése.
3. Gyógyszerhatékonyság és mellékhatás-információk jobb követése.
4. Beágyazott számítástechnika, okos otthon, idős- és beteggondozás: mindennapi tevékenység teljes körű potenciális rögzítése.
5. Elektronikus tárgykövetés: mindennapi használati tárgyak helyzetének teljes körű potenciális rögzítése.

A hétköznapi nagy adattömegek megjelenésénél említett egybemosódó fizikai-informatikai világban a mindennapi élet egyre nagyobb részéhez tartozik egy elektronikai-informatikai vetület is, amelyben modellek „követik" a tevékenységeket és direkt vagy indirekt módon hatnak vissza a valós világra. Egy leegyszerűsített kép szerint ebbe a virtuális térbe a következők kerülhetnek be:

1. a fizikai tárgyak hely- és állapotjellemzői (egy elektronikai követőrendszeren keresztül)
2. személyek fiziológiai állapota (különböző passzív mérőrendszereken keresztül)
3. személyek kognitív leírói (aktív közreműködéssel határidőnaplók, teendők listájának a használatával vagy passzív modellezés útján).

A teljesség igénye nélkül az ebben megjelenő információk a következőek lehetnek.

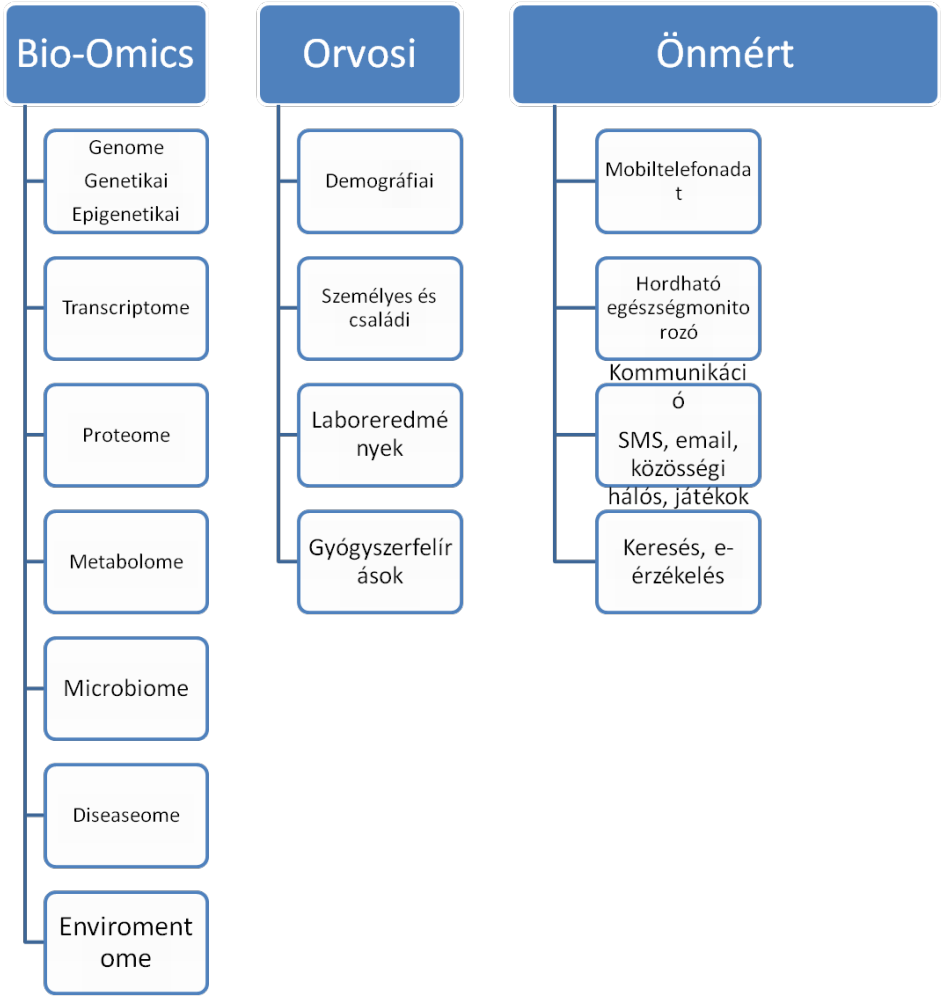
1. Általános fiziológiás állapot követése
   - Testhőmérséklet
   - Pulzus, EKG; származtatott mutatók
   - Légzés; kapacitás, gyakoriság
   - Vérnyomás
   - Bőrellenállás

- Súly
- Kalóriabevitel
- Vércukorszint
- Testmozgás

2. Kommunikáció
   - Telefon
   - Elektronikus üzenetek
   - Közösségi hálókon és számítógépes játékokban való részvétel
3. Otthoni környezet
   - Háztartási gépek aktivitása, használata
   - Általános állapotleírók
4. Közlekedés
   - Útvonal
   - Eltöltött idő
5. Betegségspecifikus állapot követése
   - Elektronikus kórtörténet, leletek
   - Gyógyszerhasználat
   - Egészségmegőrző aktivitiás
   - Patologikus mozgás
     - Tüsszentés
     - Köhögés
     - Remegés
     - Elesés

Ezen adatoknak fontos sajátossága a többszintű, több idői lépték mentén elhelyezkedő adatok, amit az alábbi példával illusztrálunk, bemutatva az allergiás állapot többszintű követésének adatait:

- Szakorvosi adatok: szezon szerinti és éves vizitek, eseti megkeresések.
- Laboradatok: szezonszerinti és éves viziteken mért immunológiai profilok.
- Tünetek: szervrendszer és klinikai végpontok szerinti pontszámok helyszínnel órás, napi, heti, szezonális és évi bontásban.
- Gyógyszerelés: napi, heti, szezonális és évi bontásban.
- Mellékhatások: napi, heti, szezonális és évi bontásban.
- Meteorológiai adatok: helyszínnel órás, napi, heti, szezonális és évi bontásban.



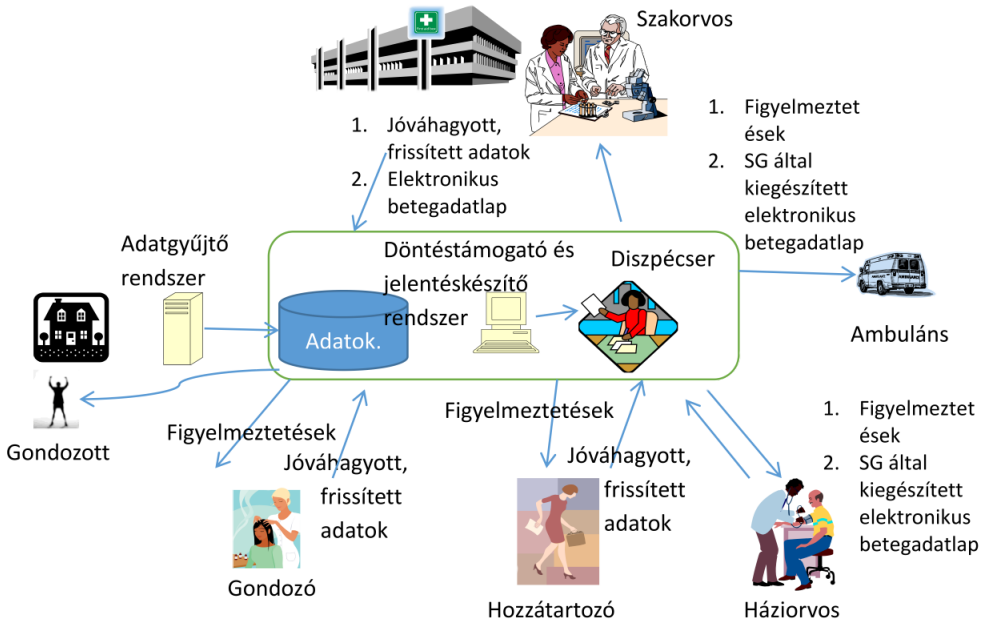
15.6. ábra. Orvosbiológiai nagy adattömegek biológiai, orvosi és hétköznapi típusai

- Légszennyezettségi adatok: helyszínnel órás, napi, heti, szezonális és évi bontásban.
- Pollenadatok: helyszínnel órás, napi, heti, szezonális és évi bontásban.
- Beteg genetikai adatai.
- Beteg életviteli adatai: fizikai aktivitás, környezeti kitettség, táplálkozási napló.

15.6. A hétköznapi nagy adattömegek bioinformatikai kihívásai

A hétköznapi nagy adattömegeket az orvosbiológiai nagy adattömegek harmadik hullámának tekinthetjük, amelyek merőben új lehetőségeket kínálnak, mint a leíró jellegű első hullámba tartozók, és az oksági/mechnizmus-orientált második hullámbeli adattömegek. Míg az első két korszakba tartozó alapkutatás-orientált volt, addig a harmadik korszakbeli

adatok alapvetően transzlációs orientációjúak, az egyének motiváltságán alapulnak, és sok esetben a „big data” kereskedelmi, ipari megközelítésének megfelelően azonnali feldolgozást, döntéseket és cselekvéseket igényelnek. Erre példák a hasonló betegek keresése, az interneten keresztüli orvosi tanácsadás, illetve akár az időskori otthoni gondozás feladata, amely az idősődő populáció miatt egy egyre fontosabb. Ennek egy keretét a 3. ábra mutatja.



15.7. ábra. A hétköznapi nagy adattömegekre épülő orvosi döntéstámogató rendszerek lehetséges szerepei

Ebben a szereplők az idős korú vagy otthon lábadozó egyén maga, hozzátartozók, gondozói ellátás, háziorvosi ellátás, szakorvosi ellátás, diszpécserközpont. Az itt keletkező adatok lehetővé teszik például a következőket:

1. A háziorvos, a szakorvos, a gondozó és a hozzátartozók elérhetik
   (a) az elektronikus nyers adatokat,
   (b) azok automatizált korrigáltját,
   (c) a kézzel történő jóváhagyását (ez a kézi megerősítés például otthoni gyógyszeradagolás (bevétel) esetén lehet fontos).
2. Az adatok, annak statisztikai leírói, és az adaptív modellek mind részévé válhatnak az általánosan elérhető elektronikus betegadatlapnak.
3. A háziorvos és a szakorvos az adatok és a követő modellek egyedi és csoportos elemzésével pontosabb, személyre szabottabb

(a) megelőzést,

(b) diagnózist és

(c) kezelést érhet el.

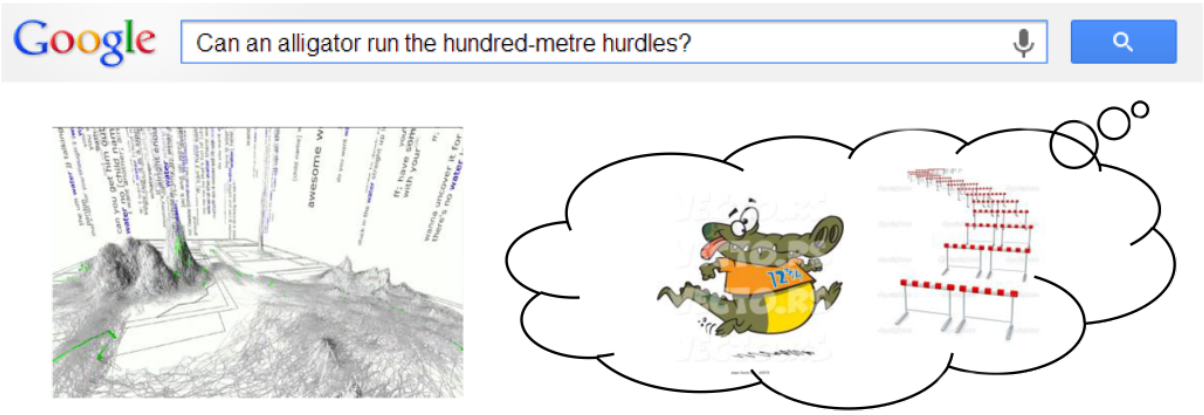
(d) Elektronikus betegadatok (kórtörténet, laboreredmények), gyógyszerezés, általános és betegségspecifikus fiziológiás adatok alapján figyelmeztetést kérhet gyógyszerbeállításra.

4. Anomáliákra, potenciális veszélyhelyzetekre való figyelmeztetést és magyarázatot kaphatnak logikai és bizonytalanságot is kezelő modellek felhasználásával,

   (a) a modellek mind egyetlen, mind több személy adatai alapján adaptívak lehetnek,

   (b) speciális, személyre szabott követési vagy figyelmeztetési modelleket hozhatnak létre, amelyek a gondozói és hozzátartozói kapcsolatban jelenthetnek nagy segítséget.

5. A gondozott egyén maga is átfogó rátekintést kaphat az állapotáról, amit felhasználhat a gyógymódjának segítésében, illetve ebből akár személyes profilt is kialakíthat, amit valós vagy akár virtuális közösségekben is felhasználhat a gyógyulás elősegítésére.



15.8. ábra. Hétköznapi logika és fogalmak kapcsolata a nagy adattömegek elemzésével: a „víz” fogalmának megjelenése egy ház térképén gyermeki szemmel és a a számítógépes jelentés és szemantika kérdése

A döntéstámogatás során a Bayes-i döntéselméleti keret és a döntési hálók egy általános keretet biztosítanak, amely az általános bioinformatikai nagy adattömegekkel való integrálást is biztosítja.

A mindennapi életből származó nagy adattömegek, azok nyílt, szabályozatlan, gyakran természetes nyelvi, sőt várhatóan audiovizuális reprezentációja felveti annak kérdését,

hogy lehetséges-e „józan ész” (common sense) nélkül ezeket az adatok elemezni. A kérdés tárgyalása meghaladja a jegyzet kereteit, de valójában éppen ez a mindennapi életből származó nagy adattömeg biztosíthatja a magasabb absztrakciós szinten lévő bioinformatikai és kemoinformatikai adatok értelmezését, hatékony kihasználását.

Erre várhatóan első példákat a hétköznapi nagy adattömegek azon felhasználása fog eredményezni posztgenomikai kutatásokban és gyógyszerhatékonysági, mellékhatás-követési vizsgálatokban, amikor ezen adatok mint egy részletes környezeti leírás és a lehető legteljesebb szervezet/egyén szintű fenotípus-leírás kerülenk felhasználásra, új végpontokat biztosítva (vö. a génexpresszió mint „ultimate” sejt szintű fenotípus [8]–[11]).

Irodalomjegyzék

[1] [Anonymous], THE SEQUENCE EXPLOSION. *Nature*, 464(7289):670–670, 2010.

[2] Carlson R, The Pace and Proliferation of Biological Technologies. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science* 2004, 1(3).

[3] Wooley J, Godzik A, Friedberg I, A Primer on Metagenomics. *Plos Computational Biology*, 6(2) 2010.

[4] Wooley J, Ye Y, Metagenomics: Facts and Artifacts, and Computational Challenges. *Journal of Computer Science and Technology* 25(1):71–81, 2010.

[5] Pearl J, Causality: models, reasoning, and inference. Cambridge University Press, Cambridge, U.K.; New York, 2000.

[6] Bell G, Hey T, Szalay A, Beyond the Data Deluge. *Science*, 323(5919):1297–1298, 2009.

[7] Bryson S, Kenwright D, Cox M, Ellsworth D, Haimes A, Visually exploring gigabyte data sets in real time. *Communications of the Acm*, 42(8):82–90, 1999.

[8] Schadt E, Monks S, Drake T, Lusis A, Che N, Colinayo V, Ruff T, Milligan S, Lamb J, Cavet G et al., Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422(6929):297–302, 2003.

[9] Schadt E, Monks S, Friend S, A new paradigm for drug discovery: integrating clinical, genetic, genomic and molecular phenotype data to identify drug targets. *Biochemical Society Transactions*, 31:437–443, 2003.

[10] Schadt E, Lamb J, Yang X, Zhu J, Edwards S, GuhaThakurta D, Sieberts S, Monks S, Reitman M, Zhang C et al., An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, 37(7):710–717, 2005.

[11] Emilsson V, Thorleifsson G, Zhang B, Leonardson A, Zink F, Zhu J, Carlson S, Helgason A, Walters G, Gunnarsdottir S et al., Genetics of gene expression and its effect on disease. *Nature*, 452(7186):423–U422, 2008.

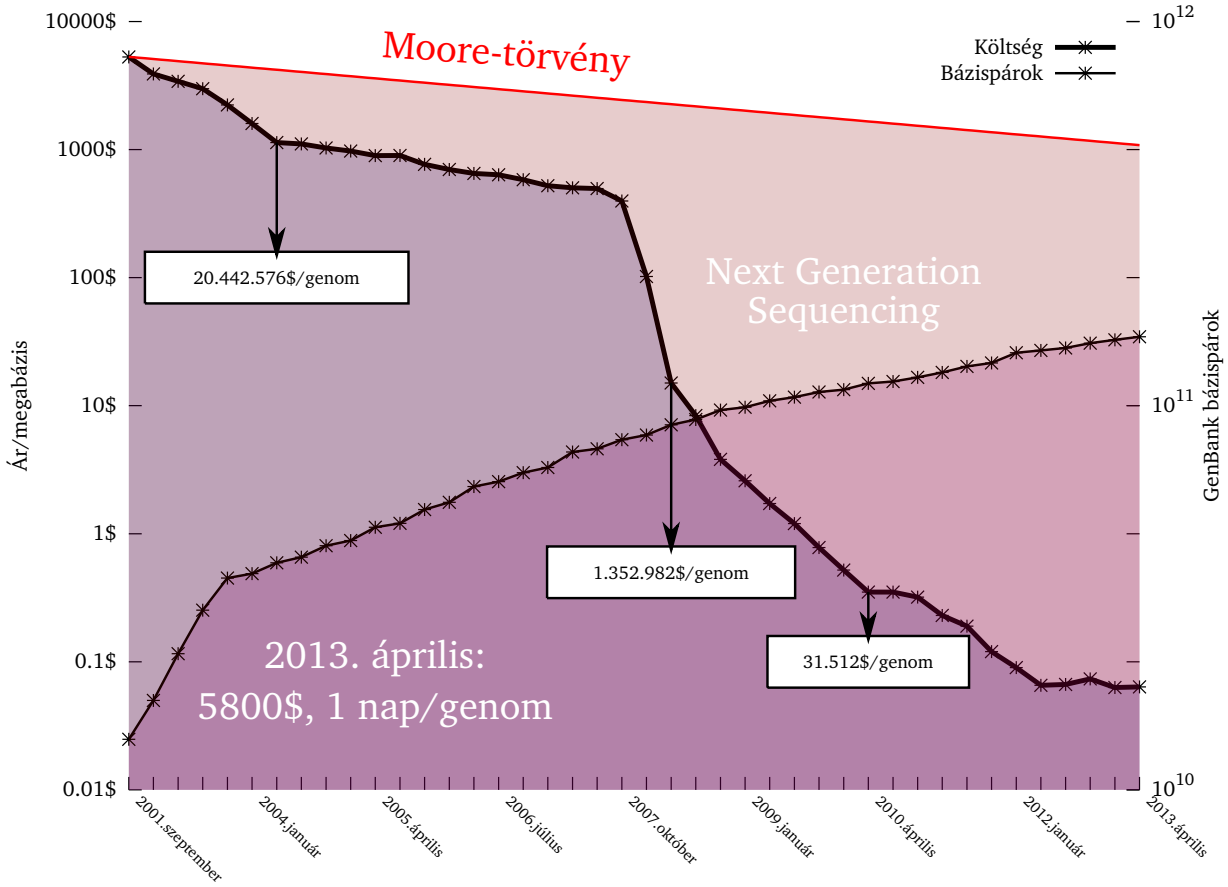
16. fejezet

Heterogén biológiai adatok fúziós elemzése

16.1. Bevezetés

A modern orvosbiológiai, bioinformatikai kutatások egyik legfőbb mozgatórugója az a technológiai forradalom, amely a XX. század, a „fizika évszázadának" második felében kezdődött és mind a mai napig tart. A számítási teljesítmény növekedésének, a csíkszélesség csökkenésének ütemét megfogalmazó Moore-törvényhez hasonlóan más tudományterületeken is hasonló észrevételek születtek, amelyek a méréstechnikák exponenciális fejlődését jósolták (pl. Carlson-törvények [1]). Ennek megfelelően a XXI. században – amelyet sokan a „biológia évszázadának" tartanak – rengeteg nagy áteresztőképességű biológiai módszer látott napvilágot, és hatalmas mennyiségű, heterogén mérési adat született, amelynek „fejben" történő szintetizálása és elemzése reménytelen vállalkozás. A biológiai és számítástudományi fejlődés, valamint ezzel párhuzamosan a mérési módszerek és számítások árának csökkenése együttesen új kutatási megközelítések kialakulásához vezetett. Ezek közé tartozik a hipotézismentes kutatási paradigma („génhalászat"), illetve a kapcsolt omikai (genomikai, proteomikai stb.) szintek együttes vizsgálatának ötlete. Az új évezred elejétől a modern biológiai alapkutatás az entitásszintű szemléletet maga mögött hagyva egyre inkább a rendszerszintű elemzések felé mozdult el (systems biology). A növekvő adatmennyiséggel párhuzamosan az orvosbiológiai adatbázisok száma is emelkedett, amelyek a következőképpen oszthatók fel (a teljesség igénye nélkül):

- Szekvencia: GenBank, EMBL, ExProt, SWISS-PROT/TrEMBL, PIR
- Útvonal: KEGG, Reactome
- Reguláció: miRBase, TRANSFAC, TRANSPATH
- Epigenetika: PubMeth
- Fehérje motif: Blocks, InterPro, Pfam, PRINTS, SUPFAM, PROSITE
- Fehérjestruktúra: PDB, MMDB



16.1. ábra. A szekvenálás költségeinek alakulása a 2000-es évek elejétől. A számítástudomány területén megfogalmazott Moore-törvényhez hasonlóan a méréstechnikák is hasonló, sőt, gyorsabb ütemben fejlődtek, a szekvenálás költségei exponenciálisan csökkentek.

- Gén–betegség asszociációk: HuGENet, PharmGKB, GenAtlas
- Farmakológia, farmakogenomika: DrugBank, SIDER, PharmGKB, PubChem
- Génexpresszió: GEO, YMGV
- Molekuláris kölcsönhatások: BIND, DIP, BRENDA, BioGRID
- Metabolikus hálózat: EcoCyc, MetaCyc, GeneNet
- Mutációk, variációk: OMIM, dbSNP, HGMD
- Ontológiák, tezauruszok: Go, UMLS, MeSH, Galen
- Publikációk: PubMed

16.2. Tudásfúzió és adatfúzió

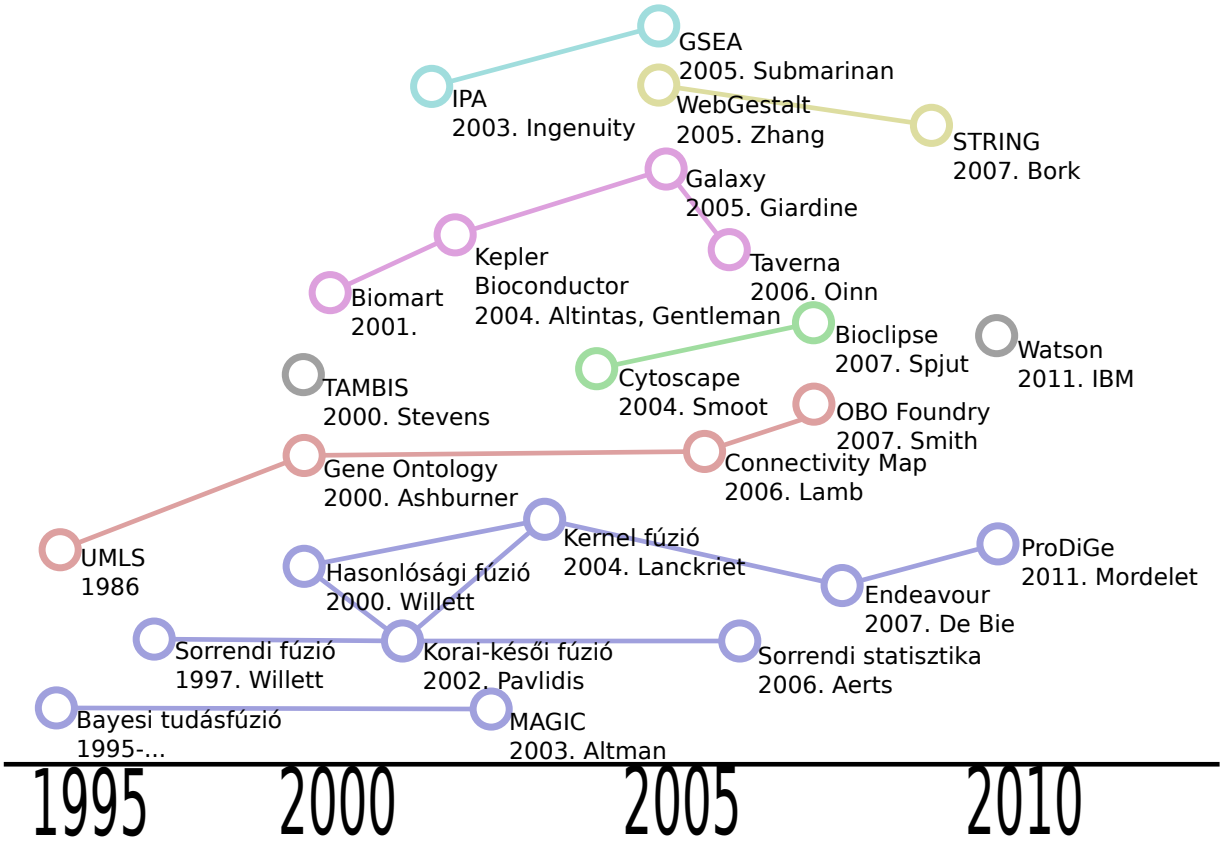
A heterogén biológiai ismeretanyag fúziója során elkülöníthetjük az tudásfúziót, illetve ennek egy szűkebb értelmezését, az adatfúziót. A tudásfúzió lényege a kutatás támogatása a különböző forrásokból származó tudás együttes, koherens felhasználásával; az adatfúzió eszköztára a nyers biológiai adatok kombinálására szorítkozik (pl. szekvenciák, expressziós mérések eredményei), gyakran numerikus módszerek alkalmazásával. A fúziós paradigma központi kérdéseihez tartozik a mérési adatok és a háttértudás egyesítése, amely így átmenetet képez az adat- és a tudásfúzió területe között. Mindegyik megközelítés az adatelemzést és -értelmezést, a kísérlettervezést és a döntéstámogatást szolgálja. A fúziós rendszerek Synnergren felosztása szerint az alábbi kategóriákba sorolhatók [2]:

- Tudáskivonatoló rendszerek
- Tudásintegrációs rendszerek
- Tudásfúziós rendszerek

A tudáskivonatolás a lekérdezéshez kapcsolható információk automatizált kinyerését jelenti a különböző biológiai tudásbázisokból, leggyakrabban adat- és szövegbányászati technikákra támaszkodva. Lehetőséget biztosítanak a kinyert tudás vizualizálására, rendszerezésére és böngészésére. Ide sorolható a legtöbb automatizált adatbányász rendszer (DAVID [3], WebGestalt [4]).

Az integrációt szolgáló eszközök célja a tudás reprezentációja és vizualizációja egy egységes felületen (STRING [5]); rendszerint tartalmaznak kivonatoló és komplex lekérdező alrendszert is (pl. természetes nyelvi lekérdező szolgáltatások), illetve kapcsolatot biztosítanak a releváns publikációkhoz és elemzésekhez. A tudásbázis-integráció egy korai példája a TAMBIS [6], amelynek kifinomult lekérdező rendszere lefordítja a kérést a heterogén adat- és tudásbázisok, szolgáltatások számára, majd a válaszokat integrálja és egy egységes felületen jeleníti meg.

E két megközelítés során a tényleges fúziót maga a kutató végzi a megjelenített információ felhasználásával, támaszkodva saját szakértelmére is. A szűkebb értelemben vett tudásfúziós rendszerek lényege a heterogén adatok transzformációja egy egységes reprezentációt képviselő szintre. A közös nyelv bevezetését célzó korai kutatások központi eleme a szemantikai integráció volt. A fogalmak egységesítését szolgáló tezauruszokon, fordítókon, szótárakon (pl. UMLS, UniGene) túl ide sorolhatók a relációk szintjén történő egységesítésre vonatkozó törekvések (pl. Gene Ontology). Egy újabb megközelítés a Connectivity Map [7], ahol a közös nyelv szerepét a különböző betegségek, gyógyszerek és egyéb molekulák hatására bekövetkező génexpressziós változások töltik be; ezek korrelációjából lehet következtetni a heterogén entitások között fennálló kapcsolatra. A korszerű technikák közé tartoznak még a gráfos megközelítések (valószínűségi gráfos modellek, pl. MAGIC [8]), a formális logikai leíró nyelvek és sztochasztikus induktív logikai programozás, hasonlóság alapú fúzió (kernel módszerek, pl. Endeavour [9]), illetve a különböző burkoló környezetek (pl. Bioclipse [10], Cytoscape [11], munkafolyamat-rendszerek), amelyek számos egységes reprezentációt és elemző algoritmust biztosítanak, rendszerint kibővíthető, moduláris (plugin) felépítéssel.



16.2. ábra. A tudásfúziós technikák időbeli alakulása. Kék színnel jelöltük az adatfúziós, pirossal a szemantikai integrációs rendszereket, zölddel a programozási, lilával a workflow környezeteket, sárgával az adatbányász, világoskékkel az útvonal-elemző eszközöket. A rendszerek leírása és további információ elérhető a hivatkozott irodalomban [2].

16.3. Az adatfúzió módszereinek felosztása

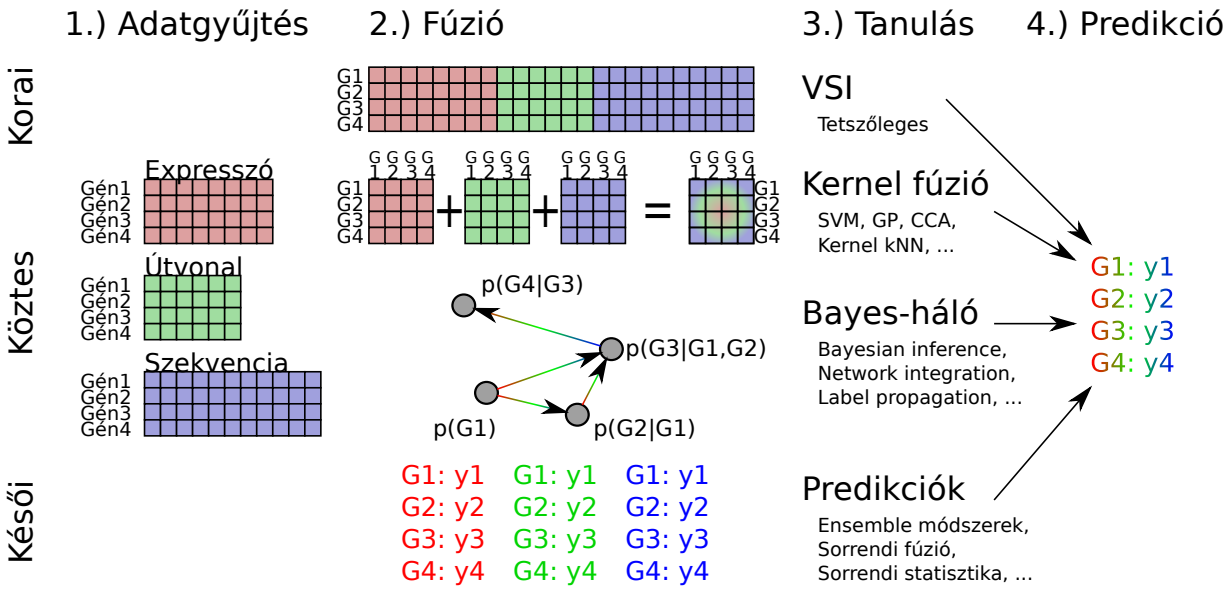
Az új paradigmák egyik központi kérdésévé a heterogén adatok fúziója vált. Az adatfúziós eljárásokkal kapcsolatban jogos elvárásként fogalmazódhatnak meg az alábbiak:

- az eltérő aspektusok figyelembevételével javuljon az eredmények minősége
- legyen lehetőség szakértői tudás integrálására
- legyen automatizált
- legyen könnyen használható, felhasználóbarát
- legyen használható különböző formátumú adatok esetén (pl. nem-vektoriális adatok)
- rendelkezzen stabil matematikai alapokkal
- legyen hatékonyan számítható
- jól skálázódjon az adatforrások számával és méretével

- kezelje a hiányos adatokat, legyen zajtűrő

A különböző technikák hagyományosan három csoportba sorolhatók (16.3. ábra) [12]:

- Korai/alacsony szintű fúzió
- Köztes/középszintű fúzió
- Késői/magas szintű fúzió



16.3. ábra. Fúziós megközelítések. A korai megközelítés az a priori *tudás adatszintű integrációja után elemez, míg a köztes lényege az adatok egy átmeneti reprezentációja pl. kernelek (hasonlósági mátrixok) vagy valószínűségi gráfos modellek (PGM) formájában. A késői módszer a forrásokon külön-külön végzett elemzésének eredményeit kombinálja. A tanulás során használható fontosabb algoritmusokat is feltüntettük.*

16.3.1. Korai fúzió

A korai fúzió (másképpen: adatintegráció) lényege az entitásokhoz tartozó különböző leírások adatszintű kombinálása. Ennek legegyszerűbb és leggyakrabban használt módszere az adatok konkatenálása (VSI, vektortér-integráció), majd az így kombinált adatokon az elemző algoritmus futtatása. Egyszerűsége mellett előnye, hogy hatékonyan számítható (az elemzést csak egyszer kell futtatni), illetve az algoritmus megkap minden információt minden forrásból, azaz az entitások leírásai között fennálló korrelációkból közvetlenül, a forrásoktól függetlenül profitál. Hátrányai közé tartozik a többi megközelítéshez képest a viszonylagos rugalmatlanság, a reprezentáció nehézségei pl. nem-vektoriális adatoknál, valamint a tárgyterületre vonatkozó *a priori* tudás (háttértudás) bevitelének problémái.

16.3.2. Köztes fúzió

A köztes módszer az adatok egy köztes reprezentációja alapján fuzionál. A két legelterjedtebb technika a kernel módszerek családja (pl. szupportvektor-gépek, Gauss-folyamatok) és a gráf alapú megközelítések (kiemelten a valószínűségi gráfos módszerek). Előbbinél az átmeneti reprezentáció az entitások páronkénti hasonlóságait tartalmazó mátrixok (kernelek), utóbbinál leggyakrabban a Bayes-háló. A köztes megközelítés ötvözi a korai fúzió hatékonyságát a késői fúzió rugalmasságával, így a gyakorlatban rendkívül elterjedtté vált.

A kernel technikák stabil matematikai alapokkal bírnak, bármilyen formátumú adatoknál használhatók (amennyiben tudunk hasonlóságokat származtatni az entitások között), rendkívül hatékonyan számíthatók, a hasonlóságmértékek szabad megválasztásával és speciális kernelek tervezésével részben lehetőség nyílik *a priori* tudás integrálására is; ugyanakkor gyakran nehézkes a használható algoritmusok és a kernelek megfelelő paraméterezésének megtalálása. A Bayes-hálókban a háttértudás *a priori* modellek feletti valószínűségi eloszlásokban tárolódik, amelyeket az adatokkal *a posteriori* eloszlások konstruálására használnak fel, így végeredményként könnyen értelmezhető valószínűségi állításokat kapunk. Előnye továbbá a bizonytalanság és a hiányos adatok kezelése, viszont nehéz az *a priori* ismeretek lefordítása a valószínűségek nyelvére, valamint hátránya a nagy számításigény.

16.3.3. Késői fúzió

A késői fúzió (másképpen: döntés-szintű fúzió) során az elemző algoritmust minden adatforrásra külön-külön futtatják, és az így nyert eredményeket kombinálják. Egyik legnagyobb előnye a nagyfokú rugalmasság: gyakorlatilag bármilyen jellegű adat kombinálható, és lehetőség van forrásonként eltérő elemző algoritmusok használatára is; ezek közül a problémának leginkább megfelelők kiválasztása egyben a szakértői tudás bevitelének egyik lehetősége is. Mivel a kimenetek már rendszerint azonos formátumúak, a fúzió könnyen elvégezhető. Hátrányként említhető a nagy számításigény (forrásonkénti elemzés, majd az eredmények kombinálása), illetve a döntési szinten megjelenő jelentős dimenzió-redukció: a fúziónál maguk az adatok már nem látszanak, csak az elemzések kimenetei. Emiatt a késői módszer kevésbé érzékeny az adatok közti korrelációkra, mint a korai.

Az egyszerűbb módszerek közt tartják számon a kimenetek algebrai kombinációját (pl. összegzés, súlyozott átlag, medián stb.), míg a kifinomultabb technikák közé tartoznak az ensemble-módszerek (Mixture of Experts, bagging, boosting, stacking), illetve a sorrendi fúzió különböző formái (sorrendi statisztika, Borda ranking, parallel selection, Pareto-ranking stb.). Számos sorrendi fúziós módszer leírása és teljesítményük összehasonlítása megtalálható Svensson közleményében [13].

- **Sum rank:** adott entitás összes sorrendezésben elért pozícióit összeadjuk, a végső sorrend az így nyert pozíciók alapján alakul.
- **Sum score:** adott entitás összes sorrendezésben elért pontszámait elosztjuk az adott

sorrendben megtalálható legmagasabb pontszámmal, majd az így nyert értékeket összeadjuk. A végső sorrend ezen relatív pontszámok alapján alakul.

- **Pareto ranking:** adott entitás végső sorrendben elfoglalt pozíciója attól függ, hogy hány entitás ér el nála magasabb rangot a sorrendekben. A döntetlenek a sum rank módszerrel dőlnek el.
- **Rank vote:** minden sorrend „szavaz" az első n elemére, az entitások végső sorrendje a kapott szavazatok alapján alakul. A döntetlenek a sum score módszerrel dőlnek el.
- **Parallel selection:** minden sorrendből párhuzamosan kiválasztjuk a legjobb entitást. Ha olyan jönne, amely egy másik sorrendből már bekerült, akkor helyette a következőt választjuk, majd ismételjük az eljárást.

16.4. Hasonlóság alapú adatfúzió

Az entitások páronként és forrásonként vett hasonlóságain alapuló fúzió elsőként a génexpressziós adatok klaszterezésénél jelent meg a 2000-es évek elején, azonban csak Lanckriet meghatározó 2004-es közleménye után terjedt el széles körben [14]. Itt a fúzió során a hasonlósági mátrixok (kernelek) súlyozott összegét használták, a tanulási fázist pedig ún. szupportvektor-géppel (SVM) végezték (ami egyben a források optimális súlyozását is megtalálta). Az SVM fontosabb előnyei az automatikus súlyozás mellett a gyorsaság, a jelenleg egyik legjobbnak tartott általánosító képesség és pontosság, valamint a jó skálázódás nagy méretű adatokra is.

Minden szimmetrikus pozitív szemidefinit hasonlósági mátrix (kernel) meghatároz egy Hilbert-teret, amelyet Reproducing Kernel Hilbert-térnek (RKHS) nevezünk. Legyen adott $k : \mathbb{R}^l \times \mathbb{R}^l \to \mathbb{R}$ kernelfüggvény (hasonlóságmérték), ahol például

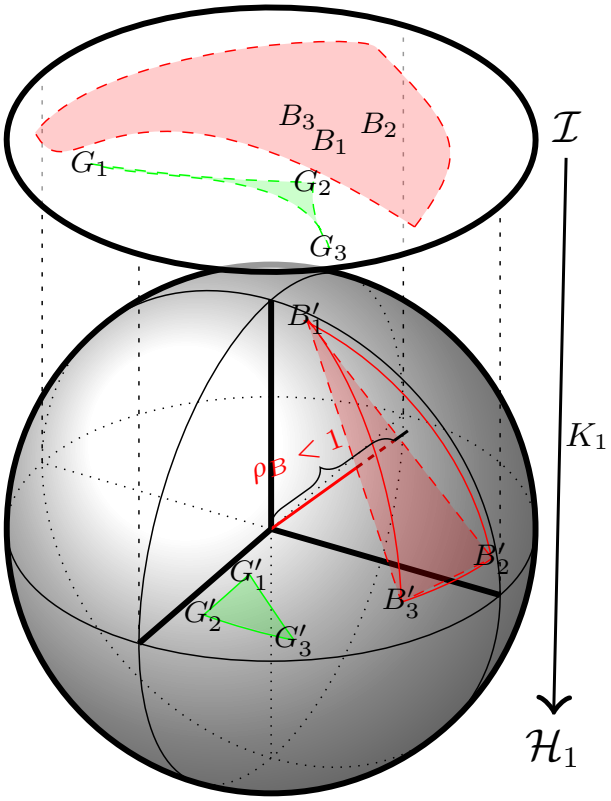
$$k(\boldsymbol{x_i}, \boldsymbol{x_j}) = e^{-\gamma \left\| \boldsymbol{x_i} - \boldsymbol{x_j} \right\|^2},$$

a kernelmátrix ezeket az értékeket tartalmazza. Ehhez létezik a $\mathcal{H}$ Hilbert-tér, ahol

$$k(\boldsymbol{x_i}, \boldsymbol{x_j}) = \langle \phi(\boldsymbol{x_i}), \phi(\boldsymbol{x_j}) \rangle_{\mathcal{H}},$$

ahol $\phi(\cdot)$ végzi az adatok vetítését a RKHS-be, az SVM pedig ebben a térben végzi a tanulást. A fenti példában definiált k függvényt nevezzük Gauss-féle radiális bázisfüggvénynek (RBF). Megmutatható, hogy az ehhez tartozó tér például végtelen dimenziós.

Több információforrás integrálására nyújt lehetőséget a kernelfúzió (Multiple Kernel Learning), amely kezdetben a mátrixok egyszerű összegét, vagy súlyozott átlagát vette alapul [12, 14]. Itt lehet kihasználni a tényt, hogy a források optimális súlyozása megkapható, ha a súlytényezőket sikerül beépíteni az optimalizációs feladatba, amire több formalizáció



16.4. ábra. A kernelsúlyozási technika szemléltetése. Az adatokat a bemeneti térből a $\mathcal{H}_1$ térbe transzformáljuk, ahol az SVM a tanulást végzi. A B-vel jelölt, valójában egymáshoz kevéssé hasonló entitások nagy térrészt feszítenek ki, így ehhez az információforráshoz alacsony súlyt rendelünk. Ha lekérdezésként G-t adjuk meg, a kifeszített térrész kisebb, azaz az információforrás jól jellemzi a lekérdezést (és így magasabb súlyt kap).

is született. Ugyancsak ezen a ponton jelent meg a súlyok regularizációjának kérdése, ahol az ún. $L2$-normalizáció vált be a sparse ($p < 2$) módszerekkel szemben [15].

A kernelsúlyok optimalizációs feladatba való integrálására számos megközelítés született [16, 17, 18]. Egy 2010. végi formalizációval a probléma differenciálható duál célfüggvényre vezethető, amely lehetővé teszi a hagyományos SVM-nél igen jól bevált SMO algoritmus alkalmazását [19]. Ha a tanulási fázisban ún. egyosztályos SVM-et alkalmazunk, a primál probléma így írható:

$$\begin{aligned} \min_{\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{d},\rho} \quad & \frac{1}{2}\sum_k \frac{\boldsymbol{w_k}^T \boldsymbol{w_k}}{d_k} - \rho + \frac{1}{\nu l}\sum_i \xi_i + \frac{\lambda}{2}\left(\sum_k d_k^p\right)^{\frac{2}{p}} \\ s.t. \quad & \sum_k \boldsymbol{w_k}^T \phi_k(\boldsymbol{x_i}) \geq \rho - \xi_i \\ & \boldsymbol{\xi} \geq 0, \boldsymbol{d} > 0, \quad i = 1, 2, ..., l, \end{aligned}$$

ahol λ szabályozza a d_k kernelsúlyokra vonatkozó Lp regularizációt. A duál:

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \mathcal{D}(\boldsymbol{\alpha}) = -\frac{1}{8\lambda}\left(\sum_k \left(\boldsymbol{\alpha}^T K_k \boldsymbol{\alpha}\right)^q\right)^{\frac{2}{q}} \\ s.t. \quad & 0 \leq \boldsymbol{\alpha} \leq 1, \quad \mathbf{1}^T\boldsymbol{\alpha} = \nu l, \quad \frac{1}{p}+\frac{1}{q}=1. \end{aligned}$$

Ha az eszközt prioritizálásra akarjuk használni, az origótól számított, hipersíkra merőleges távolságot megkaphatjuk a

$$f(\boldsymbol{x}) = \frac{\sum_i \alpha_i \sum_k d_k K_k(\boldsymbol{x_i}, \boldsymbol{x})}{\sqrt{\sum_k d_k \boldsymbol{\alpha}^T K_k \boldsymbol{\alpha}}}$$

formulával, ahol a nevező a normalizációt szolgálja, a konstans ρ tagot pedig elhagyjuk.

Láttuk, hogy a kernelfúziós keretrendszer is alkalmas sorrendezések elvégzésére. Erre mutat példát a Leuveni Katolikus Egyetemen kifejlesztett Endeavour rendszer [9], vagy az ennek továbbfejlesztett változata, a ProDiGe [20]. E megközelítés több tekintetben meghaladja a hagyományos, globális hasonlóságokra támaszkodó technikákat. A források automatikus súlyozásával a módszer kontextusfüggővé válik, azaz a fúziót a lekérdezés információtartalmára is támaszkodva végzi el. További előnyt jelent, hogy így a lekérdezés elemeinek akár ismeretlen összetartozására is fény derülhet: ha például tudtunkon kívül azonos biológiai útvonalon fekvő géneket adunk meg lekérdezésként, és van útvonal alapú információforrásunk, az magas súlyt fog kapni. Az egyosztályos SVM másik kedvező tulajdonsága és egyben hagyományos alkalmazási területe az ún. outlier detekció: ha kiugró elemeket tartalmazó, inhomogén lekérdezést adunk meg, az algoritmus ezt detektálja. Hátrány, hogy ekkor egyúttal a sorrend is értelmetlenné válhat, szélsőséges esetben a lekérdezés akár az utolsó helyekre is szorulhat. További hátrány a módszer viszonylagos érzékenysége a zajos kernelekre, így az információforrások helyes megválasztása kritikus fontosságú.

Irodalomjegyzék

[1] R. Carlson, The pace and proliferation of biological technologies. *Biosecur Bioterror*, 1:203–214, 2003.

[2] J. Synnergren, B. Olsson, and J. Gamalielsson, Classification of information fusion methods in systems biology. *In Silico Biol. (Gedrukt)*, 9:65–76, 2009.

[3] d. a. W. Huang, B. T. Sherman, Q. Tan, J. Kir, D. Liu, D. Bryant, Y. Guo, R. Stephens, M. W. Baseler, H. C. Lane, and R. A. Lempicki, DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res.*, 35:W169–175, July 2007.

[4] B.Zhang, S. Kirov, and J. Snoddy, WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res.*, 33:W741–748, July 2005.

[5] C. von Mering, L. J. Jensen, M. Kuhn, S. Chaffron, T. Doerks, B. Kruger, B. Snel, and P. Bork, STRING 7 – recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.*, 35:D358–362, Jan. 2007.

[6] P. G. Baker, A. Brass, S. Bechhofer, C. Goble, N. Paton, and R. Stevens, TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. An Overview. In: *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology (ISMB'98)*, pages 25–34, Menlow Park, California, June 28–July 1 1998. AAAI Press.

[7] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J. P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wei, S. A. Armstrong, S. J. Haggarty, P. A. Clemons, R. Wei, S. A. Carr, E. S. Lander, and T. R. Golub, The Connectivity Map: using gene-expression signatures to connect small molecules, genes and disease. *Science*, 313(5795):1929–1935, Sep. 2006.

[8] O. G. Troyanskaya, K. Dolinski, A. B. Owen, R. B. Altman, and D. Botstein, A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proc. Natl. Acad. Sci. U.S.A.*, 100:8348–8353, July 2003.

[9] T. De Bie, L. C. Tranchevent, L. M. van Oeffelen, and Y. Moreau, Kernel-based data fusion for gene prioritization. *Bioinformatics*, 23:i125–132, July 2007.

[10] O. Spjuth, T. Helmus, E. L. Willighagen, S. Kuhn, M. Eklund, J. Wagener, P. Murray-Rust, C. Steinbeck, and J. E. Wikberg, Bioclipse: an open source workbench for chemo- and bioinformatics. *BMC Bioinformatics*, 8:59, 2007.

[11] M. E. Smoot, K. Ono, J. Ruscheinski, P. L. Wang, and T. Ideker, Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27:431–432, Feb. 2011.

[12] P. Pavlidis, J. Weston, J. Cai, and W. S. Noble, Learning gene functional classifications from multiple data types. *J. Comput. Biol.*, 9:401–411, 2002.

[13] F. Svensson, A. Karlen, and C. Skold, Virtual screening data fusion using both structure- and ligand-based methods. *J Chem Inf Model*, 52(1):225–232, Jan. 2012.

[14] G. R. G. Lanckriet, M. Deng, N. Cristianini, M. I. Jordan, and W. S. Noble, Kernel-based data fusion and its application to protein function prediction in yeast. In: *Proceedings of the Pacific Symposium on Biocomputing*, 2004.

[15] S. Yu, T. Falck, A. Daemen, L. C. Tranchevent, J. A. Suykens, B. De Moor, and Y. Moreau, L2-norm multiple kernel learning and its application to biomedical data fusion. *BMC Bioinformatics*, 11:309, 2010.

[16] Alain Rakotomamonjy, Francis R. Bach, Stephane Canu, and Yves Grandvalet, SimpleMKL. *Journal of Machine Learning Research*, 9:2491–2521, November 2008.

[17] Marius Kloft, Ulf Brefeld, Soeren Sonnenburg, Pavel Laskov, Klaus-Robert Müller, and Alexander Zien, Efficient and Accurate Lp-Norm Multiple Kernel Learning. In: Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 997–1005, 2009.

[18] Francis R. Bach, Gert R. G. Lanckriet, and Michael I. Jordan, Multiple kernel learning, conic duality, and the SMO algorithm. In: *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, pages 6–, ACM, New York, NY, USA, 2004.

[19] S. V. N. Vishwanathan, Z. Sun, N. Theera-Ampornpunt, and M. Varma, Multiple Kernel Learning and the SMO Algorithm. In: *Advances in Neural Information Processing Systems*, December 2010.

[20] F. Mordelet and J. P. Vert, ProDiGe: Prioritization Of Disease Genes with multitask machine learning from positive and unlabeled examples. *BMC Bioinformatics*, 12:389, 2011.

17. fejezet

A Bayes-i enciklopédia

Ebben a fejezetben áttekintjük az orvosbiológiai adatok, tudományos eredmények és számítási modellek egységes reprezentálásának trendjeit és lehetőségeit. Elsőként áttekintjük az adatok, a szakirodalom és számítási modellek gyors bővülését, amelyet az adatok gyors felhalmozódása indított el. Az adatok hatékony, nyilvános megosztása érdekében ontológiák és annotált adattárházak jöttek létre, a szakcikkek adatbázisaihoz hasonlóan, azonban a kettő között csak minimális kapcsolat jött létre, alapvetően a természetes nyelvű közlemény egészéhez kapcsolt nyilvánosan elérhető adathalmaz formájában. A genetikai asszociációs kutatások kapcsán bemutatjuk a hátrányait ennek a jelenlegi publikciós gyakorlatnak, amelynek tünetei (1) a téves, statisztikailag megalapozatlan állítások magas aránya, (2) a kísérletek megismételhetetlensége, (3) a statisztikailag gyenge eredmények publikálhatatlansága és ezért teljes elvesztése, illetve (4) a gyorsan elavuló, önkényes határvonalú, szakértők által konstruált tudásbázisok. Ígéretes megoldásként áttekintjük a szemantikus publikálás helyzetét, az adatelemzési tudásbázisokat, illetve a dekomponált modellek és modellkönyvtárak trendjeit és a modell alapú számítások fejlődési irányait. Végezetül bemutatjuk az adatokat, adatelemzési eredményeket és modellrészeket egységesen kezelő valószínűségi adatbázisokat és a Bayes-logikai megközelítést, amelyek az egységes valószínűségi reprezentáción túl egységes következtetésre is lehetőséget adnak.

17.1. Bevezető

Az emberiség tudásanyagának megosztásában az információtechnológiai fejlődés alapvető változásokat idézett elő: jelentősen leegyszerűsödött és felgyorsult a tudás közzététele. Ennek következményeként évente kb. egymillió tudományos közlemény jelenik meg csak orvosbiológiai témakörökben, de ez egy szűkebb területen is ezres nagyságrendet jelent. Ekkora számosságú cikk követése meghaladja az emberi kogníció határait, pedig a heterogén ismeretek integrálása, jelentőségének felismerése a tudományos haladás egyik záloga. A hatékony információelérést lehetővé tevő szemantikus technológiák már korábban megjelentek, azonban felhasználásuk számos megoldatlan probléma miatt csak korlátozottan jellemző. A szemantikus web és szemantikus technológiák az internet gyors elterjedésével

az ezredfordulón nagy elvárásokkal szembesültek, amelyek részben nagyméretű, szabadszöveges, közösségi szerkesztésű, informális adat- és tudásbázisok révén, részben a kódrendszerek, taxonómiák és ontológiák fejlődésével teljesültek. Ez utóbbiak különösen gyors fejlődésen mentek át az orvosbiológiában, a kémiai szinttől a molekuláris biológiai szinten át a sejtfolyamatok leírásáig. Az ontológiák megjelenése sokrétűen forradalmasította az orvosbiológiai kutatásokat, az egységes annotáció mellett lehetővé tette új statisztikai eljárások megjelenését. Azonban a szabadszövegekből automatikusan kivonatoló eszközök teljesítménye az entitás felismerésen túl a relációk azonosításában már nehezen fokozható a természetes nyelv gazdagsága miatt. A posztgenomikai korszak egyre gyarapodó, klinikai validitással is rendelkező genetikai asszociációs és farmakogenomikai eredményeinek halmozódása egyre inkább előtérbe helyezi a szövegbányászati módszerekkel és szakértők segítségével költségesen kialakított tudásbázisok felváltását vagy kiegészítését a szerzők által létrehozott strukturált digitális kivonatokkal és szemantikus közlemények egy rétegével. Ezt a lehetőséget erősítik a területen megjelenő szabványok, amelyek a mérés folyamatának, eredményeinek és a létrehozott prediktív modelleknek a közlését is szabályoznák.

A tudományos eredmények, az azokat alátámasztó empirikus adatok és az azok származtatását leíró számítási modellek együttes leírása szinte az írásbeliséggel egyidős enciklopédista hagyományokig visszavezethető. Ennek modern kori háttere a pozitivizmus, majd a Bécsi Kör gondolatvilága, illetve a logikai pozitivizmus, valamint H.G. Wells „World Brain" víziója és E. Garfield „Informatorium" elképzelése is [17, 18]. Napjainkban ennek az irányzatnak az átfogó képviselője a Wikipedia, amely emberi felhasználásra szánt, bár szemantikus technológiákkal kiegészített verziói egyes szakterületekre elérhetőek [5]. Az egységes reprezentálás előzményének tekinthető a Cyc projekt, bár annak eredeti 1990-es évekbeli célja a hétköznapi tudás (a józan ész, „common sense") reprezentálása volt [28]. Az egységes leírás, a köztes nyelv megteremtése szempontjából pedig meghatározó jelentőségű volt az ontológiák fejlődése, mint például orvosbiológiában a Unified Medical Language Systems (UMLS) vagy a Gene Ontology (GO) [32, 8]. Az egységes leírás gondolata megjelenik a „4. tudományos kutatási paradigma" és az „e-science" meghatározásában is [1, 23, 22].

Az egységes reprezentáció eléréséhez számos megoldásra váró problémára kell megoldást találni, amelyek az egyes alterületeken belül, illetve azokon átívelő módon is jelen vannak. Egy általános kihívás a minden területen jelenlévő bizonytalanság. A bizonytalanság kezelésére a valószínűségszámítás általános keretrendszert kínál, amelynek szubjektív értelmezése a tudásintegrációra egy koherens, sőt normatív rendszert kínál (származtatását a Valószínűségi döntéstámogatás című jegyzetben tárgyaljuk). Ennek megoldására informatikai oldalról több szinten is új elméleti eredmények, szabványok és rendszerek is jelentek meg, mint például a valószínűségi adatbázisok és valószínűségi logikák területén.

Az adatelemzés kapcsán megoldatlan feladat a többlépéses, megerősítő méréseket is tartalmazó vagy jelentős utófeldolgozást igénylő molekuláris biológiai mérések reprezentálása, mint az új generációs genetikai szekvenálási adatok (next-generation sequencing, NGS) vagy áramlási citométer (Fluorescence-activated cell sorting, FACS) adatok esetén.

Megoldatlan a részletes fenotípus-információk szabványos leírása, különös tekintettel a mindennapi életben keletkező nagy adattömegekre, amelyek az elektronikus kommu-

nikáció különböző formáiból, a hordható elektronikus eszközökből, az intelligens otthon eszközeiből származnak.

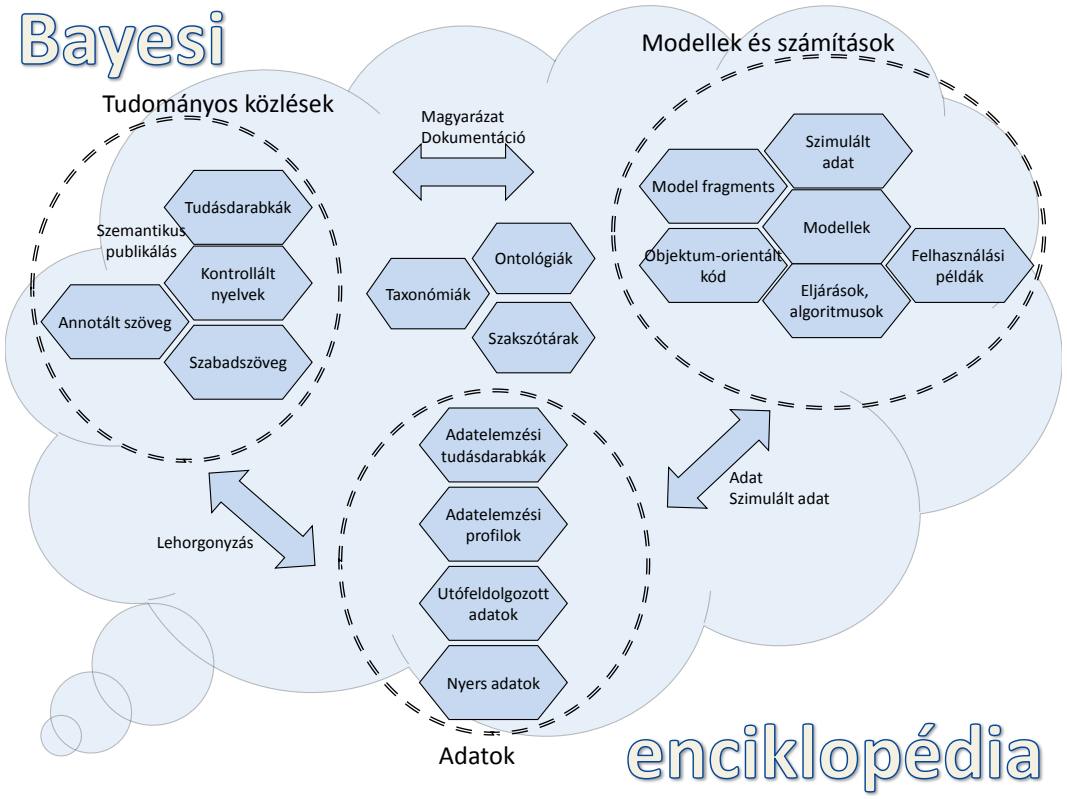
Az adatelemzés eredményeinek közlése, reprezentálása is megoldatlan, különösen a többváltozós, kontextuális, bizonytalan információk reprezentálása. Ez már a tudományos eredmények közlésének problematikájaként is felfogható a kapcsolódó értelmezések miatt.

A szemantikus publikálás területén általában hiányoznak a szemantikus publikáláshoz szükséges, széles körben elfogadott fogalmi rendszerek (ontológiák), a dokumentumok szabványosított felépítése sem alakult ki, és a szükséges szerkesztő eszközök sem terjedtek el széles körben. Nevezetesen, megoldatlan problémák a következőek: 1) egy adott tárgyterület heterogén ontológiáinak, szabványainak konzisztens együttes használata, 2) a szabadszöveges publikációkhoz és empirikus eredményekhez való kapcsolat, 3) a tudományos információközlésbe való beilleszkedés, 4) informális és formális következtetésekben való felhasználás, 5) számítási szempontból hatékony következtetés. Különösen fontos kérdés volna a genomikai szabványok érvényesítése a mérés folyamatának, a mérés eredményeiből származtatott genomikai asszociációknak és prediktív modelleknek a leírásában. Megoldatlan kérdés a szemantikus publikálás kapcsolódása a ma elterjedt szövegbányászati módszerekhez. Elméleti és gyakorlati oldalról központi kérdés a bizonytalan tudás reprezentációja, aminek része az említett statisztikai adatelemzési eredményeknek a szemantikus publikálása. Az információközlés folyamatában tisztázásra vár a szerzők segítésének módja, a kiadók egységes szabványosítása, amely az alkalmazott webtechnológiák szabványosítását is jelenti. A személyre szabott medicina, de különösen a rákbetegségek területén kulcskérdés volna az alapkutatási és a klinikai hasznossággal bíró eredmények gyors és megbízható megjelenése a klinikai gyakorlatban. A szemantikus publikálás révén potenciálisan létrejövő, adatelemzési eredményeket integráló, valószínűségi tudásbázis akár egy szűkebb tárgyterület kapcsán is közlemények tízezreit tartalmazhatja. Az ebben való logikai következtetés sikere azonban alapvető módon függ a számítási hatékonyságtól.

Megoldatlan kérdés a szakértői, „kézi' összeállítású tudásbázisok együttes használata, elsődlegesen emberi felhasználásra szánt kapcsolódáson túli betagozodásuk egy egységes tudásbázisba, hatékony fenntartásuk, határaik, megbízhatóságuk explicit reprezentálása. Önmagában is megoldatlan probléma az adatelemzéshez való részletes kapcsolat reprezentálása, a szövegbányászati eszközök hatékony használata, különösen a fentebb említett kontextuális, bizonytalan, többváltozós eredmények kivonatolása. Ezen eredményeknél a fenntartás és aktualizálás különösen fontos, mivel általánosságban minél bizonytalabb és komplexebb egy tudáselem, annál kézimunka igényesebb, és aktualitása, fennállása is annál gyorsabban változhat. A szakértői tudásbázisok egységes rendszerben való felhasználását tovább nehezíti gyakori kereskedelmi voltuk, illetve szabadalmi védettségük is.

Végezetül a modellek és számítási eljárások dokumentálása, az adat és eredmények közti útvonal formális reprezentálása is megemlítendő mint jelenleg megoldatlan feladat. Egyrészt a modellek, moduláris modellrészek reprezentálása megoldatlan, másrészt kombinálásuk, transzformálásuk és felhasználásuk módjának leírása, azaz a felparaméterezésük standardizált leírása is megoldatlan, ami a replikálhatóság miatt kap egyre nagyobb hangúlyt az adat utófeldolgozása és elemzése határvonalán (például az új generációs szekvenálási adatok utófeldolgozása és elemzése kapcsán).

Az idealizált egységes tudásbázis részeit és az egészükben lévő főbb kapcsolatokat – egy Bayes-i enciklopédia keretében – a 17.1. ábra mutatja.

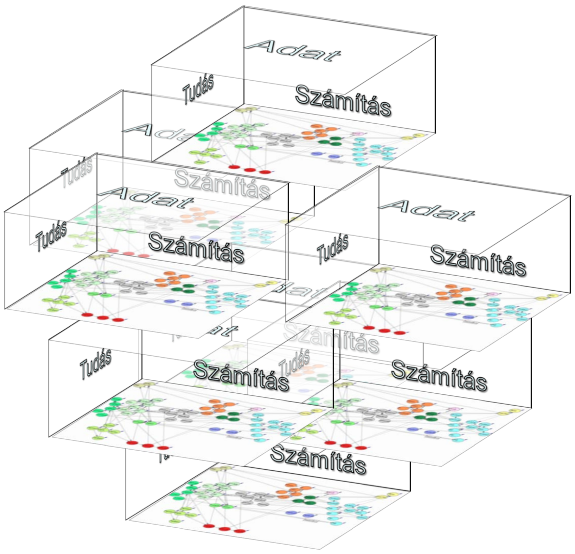


17.1. ábra. Egy Bayes-i enciklopédia összetevői és kerete

Egy adat, tudás, számítási modellek egységén alapuló tudásbázis még egy szűk szakterületen, mint például a genetikai asszociációs területen belül – akár csak egyetlen (útvonal) betegséghez kapcsolódó farmakogenomikai szakterületen belül is – nagy kihívás a klinikai felhasználás miatt. Ekkor a diagnosztikai mérések, a lelet előállítása, a lelet értelmezése és a terápiás döntések is mind kapcsolódnak egy ilyen tudásbázishoz. Fontos megjegyezni, hogy ezen komplex, egységes tudásbázisok célja nem a betegek közvetlen tájékoztatása leegyszerűsítő vagy szakorvoshoz orientáló módon. Hasonlóan, a legátfogóbb tudásbázis létrehozása sem pótolja a felhasználásra vonatkozó szakértő tudást, és ezen tudásbázisok nem a kreativitás és emberi, klinikai relevancia felismerésének kiváltását, hanem éppen annak kiegészítését szolgálhatják.

17.2. Az adat, tudás, számítás hármasának modern kori megjelenései

Az 1990-es évektől induló, majd egyre gyorsuló ütemben halmozódó orvosbiológiai nagy adattömegek egyedi helyzetet teremtettek a tudásgazdag, autonóm szintekkel rendelkező orvosbiológiában. A nagy adattömegek korábbi megjelenése a nukleáris fizikai vagy űrkutatási területeken a redukcionista megközelítés szolgálatában történt, mégha azok akár a jelenleg két végpontnak tekinthető elmélethez is igazodtak, mint a részecskefizika vagy éppen a gravitációs kutatások. Az orvosbiológiában ezzel szemben a nagy adattömegek újabb és újabb autonóm, gyengén kapcsolt szinteknek a megjelenését is elősegítették, mint például a genetikai variánsok, epigenetika, mikro-RNS-k szintje vagy a mikrobiome. Így a nagy adattömegek egyelőre inkább leíró jelleggel egyre nagyobb mennyiségű tudáselemet generálnak a szintek szabályozási, számítási modelljeivel együtt, mintsem általános, több szint jelenségeit prediktáló elméletet eredményeztek volna a redukcionista megközelítés szerint. Az egyes alterületeken halmozódó adat, tudás és számítási modellek így komoly kihívást jelentenek és várhatóan nem egy múló tranzienst, ami némiképpen eltér a 4. paradigmának nevezett adatvezérelt kutatási paradigmától és közelebb esik az e-science „kiberfizikai” világképéhez [27]. Hasonló helyzet várható a most induló agykutatási programok területén is, amelyen belül az ioncsatorna-modellezéstől a sejtmodelleken át a klinikai képalkotásig várhatóan szintek és megközelítések sokasága fogja az adat, tudás és számítási modellek hármasait létrehozni.



17.2. ábra. Az adat, tudás és számítási modellek hármasainak építőkockái

17.3. Az adat, tudás, számítás hármasa a genetikai asszociációs kutatásokban

Az adat, tudás, számítási modellek hármasával kapcsolatos trendek áttekintéséhez elsőként is tekintsük át a jelenlegi gyakorlat főbb vonásait Az 1990-től egyre gyarapodó orvosbiológiai nagy adattömegek első hulláma a fajszintű genetikai szekvencia-adatokat és fehérjékre vonatkozó strukturális adatokat tartalmazott. Az ezredfordulótól kiteljesedő sejtszintű kifejeződési adatok, mint a génkifejeződési, proteomikai, metabolomikai adatok már egyed-, betegség- és szövetspecifikusak voltak, hasonlóan a genetikai variációkra vonatkozó adatokhoz. Az orvosbiológiában megjelenő nagy adattömegek harmadik hulláma az egyedszintű fenotípus- és környezeti adatok, mint például a klinikai adatok, a mindennapi kommunikációs adatok és a viselhető elektronikai eszközökből, egészségmonitorozó eszközökből származó adatok. Az adatok megosztásának igénye önmagában is, de a tudományos közlések rendszerének átalakulása miatt is fontos szempont volt a megismételhetőség, eltérő elemzés és metaelemzés miatt is. Ennek eredményeképpen jelentek meg a Microarray Gene Expression Data (MGED) standard és a Minimum Information About a Microarray Experiment (MIAME) standard, illetve olyan adatbázisok, mint a Gene Expression Omnibus (GEO) [13, 12]. A később induló genetikai variácók feltérképezésével analóg módon jelent meg a Minimum Information about a Genotyping Experiment (MIGEN) [24] és olyan adatbázisok, mint például a European Genotyping Archive, amely főként teljes genomszélességű adatok tárolására jött léttre (genome-wide association studies, GWASs).

Párhuzamosan az adattárolási szabványok kialakulásával az adatok tudáselemekkel történő összekapcsolának legelemibb formáját, az annotációkat is törekedtek szabványosítani, amire példa a Gene Ontology (GO) és a Unified Medical Language System (UMLS) megjelenése, bár az utóbbi inkább különböző minőségű ontologiák és szakszótárak együttese csupán. Az orvosbiológiai és kémiai publikáció tárolására átfogó megoldást kínáltak a PubMed és MedChem adatbázisok, amelyek kulcsszavait a Medical Subject Headings (MeSH) adja. A genetikai asszociációs kísérletek kivitelezésének és közlésének az egységes színvonalának a biztosítására ajánlások sorozata született:

1. STREGA: STrengthening the REporting of Genetic Associations [29],
2. STROBE: STrengthening the Reporting of OBservational studies in Epidemiology [45],
3. STROBE-ME: STrengthening the Reporting of OBservational studies in Epidemiology: Molecular Epidemiology [16],
4. GRIPS: Strengthening the reporting of genetic risk prediction studies: the GRIPS statement [25].

A szakcikkek mellett főként állami támogatással nyilvános, átfogó tudásbázisok is létrejöttek, mint például az NCBI tudásbázisai, amelyeket főként szakértői böngészésre szántak. Strukturáltabb és részben kereskedelmi termékek sokasága is létrejött, mint például

17.1. táblázat. Adatbázisok

SNP database	Availability
HuGe	http://www.hugenavigator.net/
OMIM	http://www.nslij-genetics.org/search_omim.html
S-SNPs	http://pga.gs.washington.edu
HGVBaseG2P	http://www.hgvbaseg2p.org/index
dbGAP	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap http://geneticassociationdb.nih.gov
LOVD	http://www.lovd.nl/2.0/
PharmKB	http://www.pharmgkb.org/
SNPedia	http://www.snpedia.com/index.php/SNPedia
GAD, Genecards, GoDisease, IPA, Ariadne, Alamut, GODisease and Knome	

az Online Inheritence In Man (OMIM), GeneCards, PharmGKB, IPA, Ariadne, Alamut, GODisease és a Knome.

Bár kvantitív modellek leírására jelentek meg ajánlások, mint például a GRIPS ajánlás és a Predictive Model Markup Language (PMML), de átfogó megoldások nem jelentek még meg.

Mint látható, a jelenlegi gyakorlatban az adatvilág és a tudásvilág integrálása szakcikkek egészének a szintjén, illetve szakértői értelmezés támogatására történik. A széttagoltság megértésére érdemes áttekinteni a teljes „kutatói labortól a betegágyig" folyamatot a személyre szabott medicina keretében (lásd 17.3. ábra).

A gépi feldolgozás számára a szeparáltság következményei például az alábbiak:

1. Kísérlettervezés. A szakirodalom és korábbi adatok integrálásának nehézsége a gén- és variánsprioritizáló rendszerekbe.

2. Az adatgyűjtési protokoll ad hoc jelleggel használt az adatelemzési fázisban.

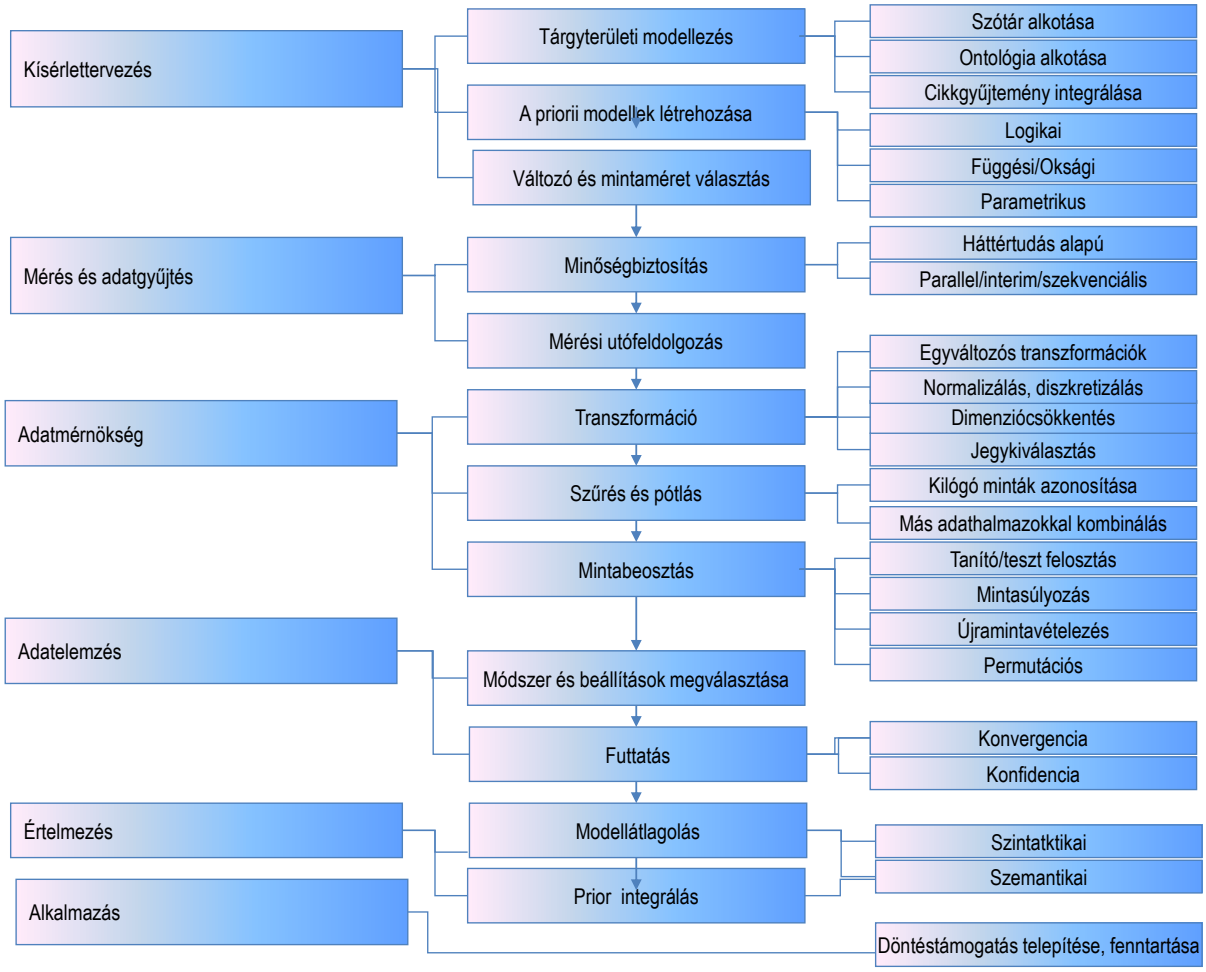
3. A szakirodalomból ad hoc módszerekkel származtatható a priori tudás az adatelemzés támogatására.

4. Az adatelemzési eredmények értelmezése az egyik legmeghatározóbb szűk keresztmetszetté vált, mivel a szakirodalmi ismeretek nehezen integrálhatóak.

5. Az elemzés során az egyes adatverziók, elemzési verziók és értelmezések sokasága ad hoc módon kezelt.

6. A gyenge megerősítésű statisztikai adatok nem publikálhatóak, így elvesznek.

7. Kvantitatív modellek, modellrészek nem kerülnek publikálásra.



17.3. ábra. Az adat, tudás, modellek hármasán átívelő munkafolyamat a kísérlettervezéstől az adatelemzésen át a tudományos eredmények közléséig, majd döntéstámogatásig

8. Klinikai gyakorlatban a szakirodalom ad hoc módon használható leletannotálásra és döntési modellek ajánlásának magyarázatgenerálására.

Lehetséges megoldásokat az alábbiakban összegzünk. Az adatbázisok és a bibliomikai adatbázisok közötti egyre halványuló határt cikkek sorozatában tárgyalták [14, 20, 19, 38, 21, 39, 4, 41, 42].

17.4. Trendek az adatvilágban

Az adatok tárolása kapcsán akut problémát jelent a komplex adatfeldolgozási lánc dokumentálása, a gazdag fenotípusos adatok standardizálása, illetve a mesterséges adatok generálásának és tárolásának a helyes egyensúlya.

17.4.1. Új generációs szekvenálási adatok feldolgozásának dokumentálása

A modern új generációs szekvenálási (NGS) méréstechnikák egy faj genomjának a költséghatékony és gyors meghatározásán túl már felhasználhatóak akár egy egyeden belüli sejtpopuláció genomjainak az átfogó vizsgálatára, mint például egy daganat vagy az immunrendszer esetében, felhasználhatóak egy ökoszisztéma genomiális vizsgálatára, például egy élelmiszerbiztonsági vagy környezetszennyezési kérdésben, de felhasználhatóak a genomok epigenetikai módosulásainak vizsgálatában és a genomok működésének kvantitatív vizsgálatában is. Az NGS méréstechnikák ezen robbanásszerű fejlődése a mérési folyamat egyszerűsödésével és standardizálásával is jár, ami a klinikai, mezőgazdasági vagy ipari rutin-felhasználáshoz szükséges volna. Azonban jelenleg még a kísérlet- és méréstervezés, mérés adatainak előfeldolgozása, elemzése, majd értelmezése nem csupán egy szakmai specializációknak megfelelően összeállított szoftverfolyamat-rendszert igényel, hanem az automatizált mérésből származó nyers mérési adatok szakértői előfeldolgozását, majd legtöbb esetben statisztikai elemzéseket, diagnosztikai következtetéseket, majd azok értelmezését, és végül optimális döntésekben való felhasználását. Ez a komplex munkafolyamat méréstechnikai, adatmérnökségi, statisztikai adatelemzési, szakterület-specifikus értelmezési és döntéselméleti fázisokat is tartalmaz. Az egyes fázisokhoz tartozó zárt gyártói vagy nyílt akadémiai eszközök tartoznak, amelyek az adott problémára specifikusan összeállított rendszerét vagy ad hoc módon hozzák létre vagy, egy munkafolyamat keretrendszerében. Azonban mindkét esetben jellemző az elemzési folyamat iteratív, többszöri részleges megismétlése, a konkrét adatokhoz legjobban illeszkedő paraméterbeállítások időrabló megkeresése, majd az elemzési lánc ismételt megismétlése. Különösen jelentős kihívás, hogy a munkafolyamat végén lévő eredmények értelmezése orvosbiológiai szakterületi tudást igényel, így gyakran derül ki, hogy egy bizonytalan eredmény értelmezése volna a szakterület szempontjából a legérdekesebb, amely a munkafolyamat méréstechnikai, adatmérnöki, majd statiszikai újrafeldolgozását és megismétlését igényli a bizonytalan eredmény pontosabbá tétele, robusztusságának vizsgálata miatt. Ennek formális dokumentálása és az elemzésben, az eredmények értelmezésében történő automatizált felhasználása fontos feladat.

17.4.2. Gazdag fenotípusos adatok

A fenotípusos adatok leírásának standardizálása olyan megoldatlan probléma, amely a genetikai asszociációs kutatások haladásának is záloga (a hiányzó örökletességgel kapcsolatos szerepét lásd [30]; a „deep phenotyping" szerepéről a pszichogenetikában , lásd [26]). Fenotípusos adatok skálája a biológiai, sejtszintű oldalon a kifejeződési adatokkal mint „végső" fenotípusokkal kezdődik [31, 7, 37, 9, 11]. Az általánosan elfogadott szint a demográfiai adatok és klinikai adatok, azonban ezek leírása is megoldatlan, amit a tumorpatológiák többféle, alternatív leírása is jól illusztrál. Sajnos a klinikai gyakorlatban használt IDC10 és IDC11 granularitása kutatási célokra általában nem elegendő. Egy ígéretes kísérlet a Human Phenotype Ontology (HPO) [34], illetve egy sikeres példa a Medical Diction-

ary for Regulatory Activities (MedDRA), amely gyógyszer-mellékhatások és -hatékonyság követését támogatja.

17.5. Trendek a tudásvilágban: szemantikus publikálás és adatelemzési tudásbázisok

17.5.1. Szemantikus publikálás

Az automatizált szövegbányászati módszerek és kereskedelmi, szakirodalom alapú bibliomikai adatbázisok mellett a szemantikus publikálás egy ígéretes jelölt. A szemantikus publikálás a szabadszöveges közlemények kibővítése formális tudásreprezentációs rétegekkel, mint például a következőek:

1. annotálás
   (a) nyelvtani annotálás, például part-of-speech jelölés,
   (b) szakszótárakból történő tartalmi annotálás,
   (c) adatokra történő mikrohivatkozás (azaz részletes adatelemzési eredményekre vonatkozó hivatkozás),
   (d) más cikkekre történő mikrohivatkozás (azaz részletes, valamely közlemény valamely pontos állítására történő hivatkozás),
2. kivonatolás
   (a) automatizált kivonatolás,
   (b) kontrollált nyelvi átírás,
3. logikai tudásreprezentáció

A szemantikus publikálás hátterét a szemantikus technológiák, a szemantikus web teremtették meg [44, 2, 3, 10, 40, 33]. A szemantikus publikálás fejlődésének illusztratív mérföldkövei a következőek:

1. Jelölő (mark-up) nyelvek használata a strukturális kémiában, majd más területeken is [44, 36].
2. Az adatbázisok és szabadszöveges közlemények közti határ elmosódásáról szóló cikksorozat [14, 20, 19, 38, 4, 41].
3. Egy példapublikáció [42].
4. A „Structured Digital Abstract” javaslat, amely egy strukturált XML összefoglalót javasolt tenni a közlemények mellé [21].

5. A FEBS javaslata a digitális összefoglalókra [39].

6. Az Elsevier Initiatives In Bioinformatics And Semantic Enrichment állásfoglalása.

7. Szövegbányaszati módszerek vizsgálata a szemantikus publikálás támogatására [21, 39].

A szemantikus publikálás általános elterjedése nem következett be a több évtizedes rutinhasználata ellenére sem bizonyos területeken. Ennek oka egyrészt az ontológiák hiányai, másrészt nagyban felelős lehet a szerzők motiválatlansága. Ez utóbbi változhatna (1) szemantikus publikáláson alapuló hasznos kutatási eszközök megjelenésével, (2) a szemantikus publikálás kötelezővé tételével, amit akár a szerzők, akár osztottan az egyenletes színvonal miatt a kiadó is végezhetne a a kulcsszavakhoz hasonlóan (3) a tudományos hozzájárulás új rendszerének kialakításával, amely az adatokra, adatelemzési eredményekre és más közleményekbeli részletes állításokra való hivatkozásokon alapulna. Végül fontos volna olyan szövegbányászati eszközökkel támogatott beviteli rendszerek fejlesztése, amelyek hatékonyak, akár a cikk főbb üzenetének a jobb kiemelését is támogatják. Ebben triviális volna szakterületi ajánlások formalizálása, mint például a genomikai területen a STREGA, STROBE, GRIPS ajánlások. Hasonlóan fontos kérdés volna a genomikai szabványok érvényesítése a mérés folyamatának, a mérés eredményeiből származtatott genomikai asszociációknak és prediktív modelleknek a leírásában. Megoldatlan kérdés a szemantikus publikálás kapcsolódása a ma elterjedt szövegbányászati módszerekhez. Az információ közlés folyamatában tisztázásra vár a szerzők segítésének módja, a kiadók egységes szabványosítása, amely az alkalmazott webtechnológiák szabványosítását is jelenti. A személyre szabott medicina, de különösen a rákbetegségek területén kulcskérdés volna az alapkutatási és a klinikai hasznossággal bíró eredmények gyors és megbízható megjelenése a klinikai gyakorlatban. A szemantikus publikálás hordozza ennek lehetőségét, de e cél elérésének módja egyelőre kutatásra vár. A szemantikus publikálás révén potenciálisan létrejövő, adatelemzési eredményeket integráló, valószínűségi tudásbázis akár egy szűkebb tárgyterület kapcsán is közlemények tízezreit tartalmazhatja. Az ebben való logikai következtetés sikere azonban alapvető módon függ a számítási hatékonyságtól.

17.5.2. Adatelemzési tudásbázisok

A nagy teljesítményű mérési módszerek megjelenésével az adatok (adatvilág) és a faktuális hipotézisek (faktuális tudásvilág) között rendkívül nagy szerephez kezdenek jutni az adott megbízhatóságú adatelemzésből származó tudáselemek, például Bayes-statisztikai adatelemzésből származó modelltulajdonságok. Az egyes modellekre vonatkozó bizonytalan tudással kapcsolatban több aspektus is egyszerűen nem létezik még jelenleg, mint például a következők.

1. Szemantikai nyelvek és ontológiák adatelemzésből származó bizonytalansága. Érdekes módon szinte minden bioinformatikai adatra léteznek szemantikai nyelvek és ontológiák, mint például a MIAME-MGED szabvány expressziós adatra, illetve teljes

modellekre is léteznek ilyenek, mint például az XBN Bayes-hálókra, vagy Predictive Model Markup Language, illetve orvosbiológiai tudásbázisokra is szemantikai nyelvek és ontológiák sokasága létezik, azonban jelenleg nincsenek bizonytalan modell tulajdonságokat leíró információkra vonatkozó szabványosítások, szemantikai nyelvek és ontológiák. A bizonytalan információk internetes megjelenésének szabványosítását megcélzó W3 csoport 2008-ban alakult meg.

2. Bizonytalansági információk tudományos közlése. A statisztikai információk közlése egy hagyományosan nehéz feladat, amit várhatóan mind szabványosítással, mind tudománypolitikai eszközökkel is támogatni fognak.

3. A faktuális tudás és az adatelemzésből származó bizonytalan tudás fúziója. A faktuális tudás felhasználása az adatelemzésben jelenleg strukturális kényszer alapú vagy kvantitatív a priori eloszlásokkal történik. A modelltulajdonságok szisztematikus leírásával, egy úgynevezett adatelemzési tudásbázissal azonban a bizonytalan tudásvilág explicit bevezetésével egy újfajta fúzió is lehetséges, amelyben a felhasznált számítást már megőrizve, de az eredményeket a lehető legérintetlenebb formájukban hagyjuk meg későbbi utófeldolgozások, értelmezések és metaelemzések számára.

Az adatelemzési tudásbázisok kapcsán cél lehet a többváltozós megközelítés, a bizonytalanság kezelése, a kontextualitás, a direkt, lehetőleg oksági relációk használata, szemantikai megközelítés (negálás, szimbolikus lekérdezés), beavatkozás és okozatiság kezelése, valószínűségi szemantika használata, logikai tudás megőrzése eredeti gazdagságában, modellek explicit kezelése. Viszont ezen adott megbízhatóságú tudáselemek tudományos kommunikációja, szabványosított felhasználása, szemantikus reprezentálása, adatbázisbeli reprezentálása, illetve fúziós módszerekbeli felhasználása még nem megoldott.

17.6. Trendek a modellvilágban

Az adatok és tudáselemek, közlemények világához képest legkevésbé kidolgozott a modellek leírása. Korai kísérletek megjelentek a modellek adatokkal, adatgyűjtési protokollal történő összekapcsolására, az adatok esetalapú értelmezésére, illetve a modellek szakcikkekkel történő összekapcsolására, mind a modellkonstruálás, modelltanulás és információkeresés, mind a magyarázatgenerálás támogatására. Jelenlegi próbálkozásként a szintetikus biológiában megjelenő BioBricks rendszer említhető [35, 6, 15], illetve a hálózat leíró rendszerek említhetőek [43].

Irodalomjegyzék

[1] A. Szalay, G. Bell, and T. Hey, Beyond the data deluge. *Science*, 323(5919):1297–1298, 2009.

[2] T. Berners-Lee and J. Hendler, Publishing on the semantic web. *Nature*, 410:1023–1024, 2001.

[3] T. Berners-Lee, J. Hendler, and O. Lassila, The semantic web. *Scientific American*, May:29–37, 2001.

[4] P. Bourne, Will a biological database be different from a biological journal? *Plos Computational Biology*, 1(3):179–181, 2005.

[5] S. Brohee, R. Barriot, and Y. Moreau, Biological knowledge bases using wikis: combining the flexibility of wikis with the structure of databases. *Bioinformatics*, 26(17):2210–2211, 2010.

[6] Y. Cai, M. L. Wilson, and J. Peccoud, Genocad for igem: a grammatical approach to the design of standard-compliant constructs. *Nucleic Acids Res.*, 38(8):2637–44, 2010.

[7] V. G. Cheung and R. S. Spielman, Genetics of human gene expression: mapping dna variants that influence gene expression. *Nat. Rev. Genet.*, 10(9):595–604, 2009.

[8] The Gene Ontology Consortium, Gene ontology: tool for the unification of biology. *Nature Genetics*, pages 25–29, 2000.

[9] A. Darvasi, Genomics: Gene expression meets genetics. *Nature*, 20(422(6929)):269–70, 2003.

[10] S. Decker, P. Mitra, and Sergey Melnik, Framework for the semantic web: an rdf tutorial. *IEEE Internet Computing*, 410:68–73, Nov.-Dec. 2000.

[11] E. T. Dermitzakis, From gene expression to disease risk. *Nat. Genet.*, 40(5):492–3, 2008.

[12] Ron Edgar, Michael Domrachev, and Alex E. Lash, Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acid Research*, 30(1):207–210, 2002.

[13] A. Brazma et al., Minimum information about a microarray experiment (miame) – toward standards for microarray data. *Nature genetics*, 29:365–371, 2001.

[14] R. J. Roberts et al., Building a 'genbank' of the published literature. *Science*, 291:2318–2319, 2001.

[15] P. Fu, A perspective of synthetic biology: assembling building blocks for novel functions. *Biotechnol J.*, 1(6):690–9, 2006.

[16] V. Gallo et al., Strengthening the reporting of observational studies in epidemiology - molecular epidemiology (strobe-me): An extension of the strobe statement. *Preventive Medicine*, 53(6):377–387, 2011.

[17] E. Garfield, *Essays of an Information Scientist*, chapter Towards the World Brain. ISI Press, Cambridge, MA, 1977.

[18] Eugene Garfield, From the world brain to the informatorium. *Information Services & Use*, 19:99–105, 1999.

[19] M. Gerstein, E-publishing on the web: Promises, pitfalls, and payoffs for bioinformatics. *Bioinformatics*, 15(6):429–431, 1999.

[20] M. Gerstein and J. Junker, Blurring the boundaries between scientific 'papers' and biological databases, 2001. *Nature* (web debate, on-line 7 May 2001).

[21] M. Gerstein, M. Seringhaus, and S. Fields, Structured digital abstract makes text mining easy. *Nature*, 447(7141):142–142, 2007.

[22] David Heckerman, *The Fourth Paradigm in Practice*. Creative Commons, 2012.

[23] Tony Hey, *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009.

[24] J. Huang et al., Minimum information about a genotyping experiment (migen). *Standards in Genomic Sciences*, 5(2):224–229, 2011.

[25] A. Janssens et al., Strengthening the reporting of genetic risk prediction studies: The grips statement. *Genetics in Medicine*, 13(5):453–456, 2011.

[26] R. Joober, The 1000 genomes project: deep genomic sequencing waiting for deep psychiatric phenotyping. *J Psychiatry Neurosci*, 36(3):147–9, 2011.

[27] L. Z. Karvalics, *Information Society Policies*, Chapter Science at the crossroads, pages 64–73. A. Rab UNESCO IFAP, 2011.

[28] Douglas Lenat and R. V. Guha, *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*. Addison-Wesley, 1990.

[29] J. Little et al., Strengthening the reporting of genetic association studies (strega): an extension of the strobe statement. *Human Genetics*, 125(9):131–151, 20091.

[30] B. Maher, Personal genomes: The case of the missing heritability. *Nature*, 456(7218):18–21, 2008.

[31] O. Nachtomy, A. Shavit, and Z. Yakhini, Gene expression and the concept of the phenotype. *Stud. Hist. Phil. Biol. & Biomed. Sci.*, 38:238–254, 2007.

[32] S. J. Nelson, T. Powell, and B. L. Humphreys, The unified medical language system (umls) project, 2001. `http://www.nlm.nih.gov`.

[33] H. Pearson, The future of the electronic scientific literature. *Nature*, 413:1–3, 2001.

[34] P. N. Robinson and S. Mundlos, The human phenotype ontology. *Clin Genet*, 77:525–534, 2010.

[35] G. Rokke, E. Korvald, J. Pahr, O. Oyas, and R Lale, Biobrick assembly standards and techniques and associated software tools. *Methods Mol Biol.*, 1116:1–24, 2014.

[36] H. Rzepa and P. Murray-Rust, A new publishing paradigm: Stm articles as part of the semantic web. *Learned Publishing*, 14(3):177–182, 2001.

[37] E. E. Schadt, S. A. Monks, T. A. Drake, A. J. Lusis, N. Che, V. Colinayo, T. G. Ruff, S. B. Milligan, J. R. Lamb, G. Cavet, P. S. Linsley, M. Mao, R. B. Stoughton, and S. H. Friend, Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 20(422(6929)):297–302, 2003.

[38] M. Seringhaus and M. Gerstein, Publishing perishing? Towards tomorrow's information architecture. *BMC Bioinformatics*, 8, 2007.

[39] M. Seringhaus and M. Gerstein, Manually structured digital abstracts: A scaffold for automatic text mining. *Febs Letters*, 582(8):1170–1170, 2008.

[40] N. Shadbolt, What does the science in e-science, *IEEE Intelligent Systems*, 17(May/June):2–3, 2002.

[41] D. Shotton, Semantic publishing: the coming revolution in scientific journal publishing. *Learned Publishing*, 22(2):85–94, 2009.

[42] D. Shotton, K. Portwin, G. Klyne, and A. Miles, Adventures in semantic publishing: Exemplar semantic enhancements of a research article. *Plos Computational Biology*, 5(4):179–181, 2009.

[43] T. Slater, Recent advances in modeling languages for pathway maps and computable biological networks. *Drug Discov Today*, 19(2):193–198, 2014.

[44] Vanessa Speding, Xml to take science by storm. *Scientific Computing World*, Supplement (Autumn):15–18, 2001.

[45] J. Vandenbroucke et al., Strengthening the reporting of observational studies in epidemiology (strobe): Explanation and elaboration. *Plos Medicine*, 4(10):1628–1654, 2007.

18. fejezet

Bioinformatikai munkafolyamat-rendszerek — esettanulmány

A bioinformatika, mint interdiszciplináris tudományág a számítógépes számítási kapacitás nagyságának és elérhetőségének növekedésével született. A szuperszámítógépek és elosztott számítási rendszerek megjelenése utat nyitott számos olyan eljárás előtt, amely annak számításigényes volta miatt korábban nem volt praktikusan alkalmazható.

A megnövekedett számítási kapacitás kihasználása azonban nem csak új lehetőségeket, hanem új feladatokat is hozott magával: egy szuperszámítógép vagy egy elosztott számítási rendszer hatékony kiaknázása komoly informatikai feladatot jelent. Ebben a fejezetben egy ilyen rendszer esettanulmány jellegű áttekintését tesszük meg, aminek a segítségével jobb rálátást kaphatunk az ilyen rendszerek megvalósításakor felmerülő problémákra és azok megoldási lehetőségeikre.

A fejezet további részeiben a következőkkel foglalkozunk: a 18.1. szakasz egy általános áttekintést ad a vizsgált rendszerről, a 18.2. szakaszban az alkalmazott adatmodellt ismertetjük. A 18.3. szakaszban a rendszer magasabb szintű felhasználói eseteivel és a megvalósítás architekturális elemeivel foglalkozunk, míg a 18.4. szakasz a szerveroldali megvalósítás részleteit tárgyalja. A 18.5. szakasz foglalkozik a munkafolyamat-rendszer záró elemével, az utófeldolgozással.

18.1. A feladat áttekintése

A vizsgált munkafolyamat-rendszer alapját a BMLA-analízisek adják, amelyek elsődleges feladata, hogy MCMC-szimulációk eredményeinek felhasználásával, Bayes-hálós modellek strukturális jegyeinek segítségével vizsgálják egy adott tárgyterület összefüggéseit.

Mivel az ilyen MCMC-szimulációk számítási igénye meglehetősen nagy, valamint egy-egy BMLA-analízis lefuttatásához több MCMC-futtatásra is szükség van, a megvalósítandó munkafolyamat-rendszernek rendelkeznie kell a következő tulajdonságokkal:

- Össze kell tudnia fogni az egy BMLA-analízishez tartozó MCMC-futásokat, az általuk felhasznált bemeneti adatokat és a létrejövő eredményeket.
- A rendszernek (a hosszú futási idők miatt) számon kell tudni tartania az egyes felhasználók által indított analíziseket, anélkül, hogy ez egy állandó kapcsolat fenntartását igényelné a felhasználótól.
- A rendelkezésre álló erőforrások felhasználásáról automatizáltan kell tudni gondoskodnia.

A fenti követelmények egy többszintű kliens-szerver architektúra irányába mutatnak, amelyben a kliens (a felhasználó) összeállíthat és feltölthet (elindíthat) BMLA-elemzéseket a szerveren, amelyek állapotát, eredményét később lekérdezheti.

18.2. Adatmodell és -reprezentáció

A BMLA-elemzések alapjául tehát a Bayes-hálós modellek és a hozzájuk tartozó megfigyelési adatok szolgálnak. A *BayesCube* program ezek szerkesztéséhez és kezeléséhez teljeskörű eszköztárat biztosít, így ezek a bemeneti adatok a vizsgált munkafolyamat-rendszer szempontjából adottnak tekinthetők*.

A megfigyelési adatok és a hozzájuk tartozó modell mellett még specifikálni kell a BMLA-elemzés során végrehajtandó MCMC-futások további paramétereit is. Ezek a futásokat meghatározó információk a következők:

Célváltozók halmaza. E változókkal kapcsolatban fog történni a statisztikák gyűjtése. A célváltozók halmazának szűkítésére (vagyis az exploratív, minden változóra kiterjedő statisztikagyűjtés elhagyására) a gyűjtött minták nagy (adott esetben akár GB-os nagyságrendű) mérete miatt van szükség.

Célváltozók kezelése az MCMC-futások során. Több célváltozó esetén lehetőség van arra, hogy pl. az MBS tulajdonságot az összes célváltozóra együttesen vonatkozóan vagy külön-külön gyűjtsük. Egy harmadik lehetőség, ha minden célváltozóhoz egy olyan modellt hozunk létre, amelyből a többi célváltozót elhagyjuk.

Vizsgált tulajdonságok halmaza. A legtipikusabbak az MBM, MBS és MBG tulajdonságok, illetve a változópárok egymáshoz való strukturális viszonyát (gyermek–szülő, leszármazott–ős, közös őssel rendelkező pár, stb.) leíró ún. oksági reláció.

Magasabb szintű vizsgálatok. Az egyedi MCMC-futások szintje felett is lehetséges magasabb szintű vizsgálatokat, teszteket végezni: ilyen lehet pl. a *permutációs teszt* vagy a *bootstrap* alkalmazása. A statisztikai megbízhatóság illetve konvergencia- és konfidencia-tesztek végzéséhez hasznos lehet ugyanazon futtatás többszörös végrehajtása is.

*A Bayes-hálókhoz és a megfigyelési adatokhoz kapcsolódó *BayesCube* szerkesztési funkciókkal itt nem foglalkozunk, mivel azok egy másik fejezetben már részletesen tárgyalva voltak.

MCMC-futások paraméterei. A MCMC-szimulációkat végrehajtó programnak magának is számos lehetséges futtatási paramétere van. Ezek értékeit, értékkombinációit is itt tudjuk megadni.

Mivel a *BayesCube* a fenti BMLA-konfigurációk szerkesztését is támogatja, a kliens oldalán ezzel előállt a teljes bemeneti adathalmaz. Ennek ismeretében már áttekinthetjük, hogy a megvalósítandó munkafolyamat-rendszernek milyen funkciókat kell támogatni a kliens felé, illetve, hogy ez milyen architekturális felépítést igényel a részéről.

18.3. Felhasználói esetek és architektúra

Az alapreprezentáció megismerése után áttekinthetjük a legfontosabb felhasználói eseteket, amelyek alapján már megtervezhető a munkafolyamat-rendszer architektúrája.

Az alapvető *use-case*-ek listája a munkafolyamat-rendszer használatában a következő:

Bemeneti adatok összeállítása. Ez a *BayesCube* szoftver által kezelt lépés tekinthető az előkészítési fázisnak: a felhasználó összeállítja a megfigyelési adatok halmazát és a hozzájuk tartozó modellt, valamint meghatározza a végrehajtandó MCMC futások halmazát a 18.2. szakaszban bemutatott konfigurációs fájl összeállításával. Ebben a szakaszban még nem történik interakció a munkafolyamat-rendszerrel.

BMLA-analízis végrehajtásának indítása. Az előző pontban összeállított adathalmazt a felhasználó feltölti a BMLA-szerverre, ahol egyrészt eltárolódnak az alapadatok, kiegészítve a feltöltő azonosítójával, másrészt elkezdődnek a végrehajtandó programfuttatások.

Analízis állapotának lekérése. Mivel a teljes analízis lefutása akár több napig is tarthat, illetve a valós számítások megkezdését más futó analízisek is késleltethetik, fontos, hogy a felhasználó az előrehaladottság állapotát igény szerint monitorozni tudja.

Eredmények lekérése. Az utolsó lépés természetesen a lefutott analízis eredményeinek lekérése a szerverről a lokális kliens-gépre, amelyen a *BayesCube* segítségével a nyers eredmények további utófeldolgozási és elemzési lépései megtehetők.

A fentiek megvalósítására szolgáló teljes rendszer architektúrája a következő elemekből épülhet fel:

Kliens-oldali interfész-függvénykönyvtár. A modularitás és újrafelhasználhatóság érdekében a fenti felhasználói esetek elindításáért felelős funkcionalitásokat egy függvénykönyvtárba ágyazva implementáljuk, amely a lehető legegyszerűbb módon valósítja meg a szerverrel történő kommunikációt. Minden elemi felhasználói eset egy-egy függvényhívás lesz a függvénykönyvtár által szolgáltatott interfészen, amely így

könnyen beépíthető bármilyen szoftvereszközbe, amely a BMLA-analízisek kezelésével foglalkozik (mint pl. a *BayesCube*).

Ennek a modulnak a fő célja tehát a munkafolyamat-rendszer belső részleteinek elfedése, az általa nyújtott szolgáltatások absztrakciója.

Webservice alkalmazás. A kliens-oldali függvénykönyvtár szerveroldali megfelelője: minden elemi szerverszolgáltatáshoz egy webservice-en keresztül elérhető függvényt rendel, így az előző modullal együtt tekinthető a valós megvalósítás és a felhasználók közötti webes kapcsolatot elfedő absztrakciós réteg részének.

Az ebben a modulban megvalósított függvények már közvetlenül érik el az architektúra további elemeit, azokon szükség szerint végrehajtva a megfelelő műveleteket.

Háttér adatbázis. Adminisztratív funkciókat lát el: a felhasználói azonosítók mellett minden feltöltött BMLA-analízishez tárolja az alapadatokat (megfigyelési adatok és modell, valamint a konfigurációs fájl és a feltöltés ideje), valamint az adott analízisre vonatkozó utolsó állapotlekérés eredményét.

Szerveroldali szoftvereszközök. A központi webserver alkalmazás által meghívott eszközök végzik el a következő alapvető elemi műveleteket: (1) futtatandó számítások halmazának összeállítása, (2) a számítások elindítása, (3) a számítások állapotának lekérdezése, esetleges leállítása, (4) az eredmények összeállítása (és a kliens számára történő elérhetővé tétele).

Feladatütemező rendszer. A teljes futtatási rendszerben számos egyedi programvégrehajtást kell koordinálni, hisz egyszerre több BMLA-elemzés is futhat párhuzamosan, illetve egyetlen BMLA-elemzés is több egyedi futtatásból áll. Emellett több különálló számítógép is rendelkezésre állhat a számítások végrehajtására. Ez a két tényező már egyértelműen egy feladatütemező rendszer alkalmazásának igényét veti fel, egy olyan rendszerét, amely képes több programfuttatási feladatnak egy elosztott rendszeren belüli párhuzamos lefuttatásának koordinálására.

A BMLA-munkafolyamat-rendszeren belül erre a célra a *HTCondor* rendszert alkalmazzuk, vagyis minden egyes végrehajtandó programfuttatáshoz egy-egy *HTCondor* feladatot (*job*ot) hozunk létre, amelynek végrehajtásáról és ütemezéséről a *HTCondor* rendszer gondoskodik majd.

Számítási csomópontok (ún. *node*-ok). Mint látható, a *HTCondor* rendszer egy újabb absztrakciós réteget hoz létre, amely a BMLA-eszközök elől fedi el a futtatáshoz használt hardverelemeket. A rendszer szoftverelemeinek megvalósítása szempontjából tehát a végrehajtáshoz használt számítógépek figyelmen kívül hagyhatók, azokkal kapcsolatban csak azt kell biztosítani, hogy (1) rajtuk telepítve legyenek a *HTCondor* rendszerhez való csatlakozáshoz szükséges eszközök, illetve (2) képesek legyenek az MCMC-szimulációkat kivitelező programok futtatására.

18.4. A szerver működési részletei

Ebben a szakaszban azokat a szerveroldali alprogramokat tekintjük át, amelyek a rendszer alapvető működését biztosítják a fő szerveralkalmazás koordinációja alapján.

HTCondor. Mint az előző szakaszban láttuk, a *HTCondor* általános feladatütemező rendszer feladata, hogy a BMLA munkafolyamatok elől elrejtse a futtatáshoz használt számítógéppark részleteit. A *HTCondor* rendszer a következő, a mi szempontunkból fontos fő tulajdonságokkal és szolgáltatásokkal rendelkezik:

- A végrehajtandó feladat egy ún. *job* formájában írható le, amely a futtatandó állomány mellett megadja az annak átadandó paramétereknek és az általa felhasznált bemeneti fájloknak a listáját. Minden *job* rendelkezhet egy részletes erőforrásigény-leírással is, a BMLA rendszerben azonban ilyen szempontból nem teszünk különbséget az egyes *job*ok között.
- A számításokat végrehajtó számítógépek (*node*-ok) mint erőforrások jelennek meg, a *HTCondor* rendszer pedig folyamatosan monitorozza a szabad erőforrások halmazát, és annak elemeihez (alapértelmezés szerint érkezési sorrendben) hozzárendeli a még ki nem osztott *job*okat. Az egyes *job*ok futási állapotának figyelése mellett a rendszer gondoskodik arról, hogy a lefutott *job*ok által előállított eredmények az eredeti (a szerveren lévő) futtatási könyvtárba kerüljenek.
- Az egyes *job*ok között lehetőség van egy elsőbbségi sorrend (precedencia) meghatározására, amely segítségével biztosítható, hogy a más *job*ok kimenetét felhasználó feladatok (pl. az eredmények aggregálását végző program) csak akkor fussanak le, amikor már az összes általuk igényelt bemeneti állomány létrejött.

soapbmla.cmd.GenerateCondorJobs.class Ennek az eszköznek a feladata, hogy a BMLA konfigurációs fájl alapján előállítsa a végrehajtandó MCMC-futások listáját. Mint azt már korábban láttuk, a konfigurációs fájlok által tartalmazott paraméterek két csoportba oszthatók: (1) a közvetlenül az MCMC-futás végző programnak átadandókéba, illetve (2) a magasabb szintűekébe, amelyek pl. a többszörös futtatások számáról vagy a permutációs tesztekről rendelkeznek. Ennek megfelelően a *HTCondor* rendszerbe feltöltendő *submit* fájlok listájának előállítása az alábbi lépésekben történik:

(1) A legtöbb magasabb szintű teszt és eljárás az adat és/vagy a modell valamilyen átalakítást is igényli[†]; ha van előírva ilyen, akkor megtörténik a segéd adat- és modellfájlok előállítása.

(2) A fentiek és a megadott MCMC-paraméter-kombinációk alapján előáll az összes különböző paraméterezésű futtatási kombináció.

[†]Például egy permutációs teszt végzése a célváltozóra vonatkozó megfigyelési adatok randomizálását, egy bootstrap-módszer alkalmazása pedig az eredeti adatfájl újramintavételezését igényli.

(3) Ha szükséges (meg van adva a `number-of-runs` paraméter), a teljes *submit*-fájl halmaz többszörözve lesz.

(4) A teljes futtatás-halmazhoz tartozik még az eredmények összesítését végző program (`mergeResults.exe`) futtatása.

Az összes fenti futtatás egy *HTCondor dagman*[‡] leíróban lesz összefogva, amelynek segítségével a teljes halmaz futtatása egyetlen *job* feltöltésével elindítható.

bn-MCMC.exe Ez a program végzi az MCMC-futások végrehajtását, bemenete az adat- és a modellfájl, illetve a parancssori argumentumokként átadott MCMC-paraméterek halmaza, kimenete az MCMC által gyűjtött statisztikákat tartalmazó fájlok halmaza. A `bn-MCMC.exe` példányainak futtatása a *HTCondor* rendszerben történik az annak átadott *submit* fájlok alapján.

mergeResults.exe A `bn-MCMC.exe` által előállított nyers eredmények összegzését végzi. Az MCMC-futások után automatikusan végrehajtódik, hogy az eredmények lekérése hatékonyabban történhessen (adott esetben több száz fájlból hoz létre néhány jóval tömörebbet), de adott esetben „kézileg" is futtatható (az ezzel kapcsolatos lehetőségekről a 18.5. szakaszban lesz szó).

18.5. Utófeldolgozási lépések

A számítások sikeres lefutása után az eredmények a kliens-oldalra kerülnek, ahol megtörténhet annak szakértői feldolgozása, értelmezése. Ezekhez a műveletekhez a *BayesCube* szoftver szolgáltat eszközöket; ezek azonban nem tartoznak szorosan magához a BMLA-munkafolyamathoz.

Az utófeldolgozás során használható másik eszköz a *mergeResults.exe* program, amely a nyers MCMC-eredmények összefésülését és aggregálását végzi. Mivel egy tipikus BMLA-elemzés számos különálló MCMC-futásból áll össze az effajta adatintegrálási lépés jelentős haszonnal járhat mind praktikus (tárhelyigény csökkentése, eredmények áttekinthetőségének növelése), mind elméleti (alapvető statisztikák, egyszerűbb konvergencia- és konfidencia-mutatók számítása) szempontból.

Maga a `mergeResults.exe` program a következők szerint működik:

- Bemeneteként az egyes MCMC-futások nyers eredményei, illetve az MCMC-paramétereket tartalmazó naplófájlok szolgálnak.
- Az eredmények feldolgozása során az ekvivalens paraméterezésű futások eredményeit a program egybefésüli.

[‡]Ez az eszköz használható az összetartozó *job*ok közti precedencia megadására.

- Az előző lépésben összefésült eredményekre kiszámol néhány alapvető statisztikát, ilyenek pl. az átlag, szórás, minimum és maximum.
- Az összefésült eredmények kerülnek a programfutás kimenetébe, igény szerint megadhatóan adott paraméterek értékei szerint külön állományokba csoportosítva.

A fenti lépések során egy fontos kérdés még, hogy mely MCMC-paraméterezések tekinthetők ekvivalensnek. Alapértelmezés szerint csak azok, amelyek minden paramétere pontosan egyezik, adott esetben azonban lehetőség van bizonyos paraméterek „kiaggregálására". Egy (vagy több) paraméter „kiaggregálása" egyszerűen annyit jelent, hogy azokat az MCMC-futásokat, amelyek paraméterezése csak a vonatkozó paraméter(ek)ben térnek el egymástól, ekvivalenseknek tekintjük, és a számítandó statisztikákat ezek halmaza felett értékeljük ki.

A fejezetben áttekintett BMLA-munkafolyamat tehát a fenti utófeldolgozási lépésekkel zárul, amelyek végrehajtása után adott esetben azok interpretációja, értelmezése, vagy egy a tapasztalatok alapján átkonfigurált, újabb BMLA-elemzés következhet.

19. fejezet

A gyógyszeripari kutatás informatikai aspektusai

19.1. A fejlesztési folyamat áttekintése

Jelen fejezet célja, hogy rövid bevezetésként szolgáljon a kismolekulás hatóanyag-tervezés modern technikáinak megismeréséhez, különösképpen az informatika, matematika és a szerves kémia határterületéről, valamint kiindulópontként szolgáljon az érdeklődő olvasónak. A tárgyalt témában számos könyv és folyamatosan növekvő számú tudományos közlemény érhető el.

Egy fejlesztési terv alapvető eleme a cél definíció, legyen az egy elérendő hatás, vagy egy jól definiált molekuláris célpont. Molekuláris célpontnak általában egy makromolekulát nevezünk a vizsgált organizmusban ami hatóanyaggal modulálható. Hatóanyag lehet kismolekula és makromolekula is – például antitestek, rövid peptidek – de jelen fejezetben csak kismolekulás gyógyszerek fejlesztésével foglalkozunk. Molekuláris célpont kiválasztásra kerülhet a betegségről rendelkezésünkre álló biológiai vagy orvosi háttértudás, vagy már ismert gyógyszer ismert hatásmechanizmusa alapján.

Ha a célpontot meghatároztuk, biztató vegyületek egy halmaza kiválasztható *in silico* szűréssel vagy *in vitro* nagy áteresztőképességű szűréssel (HTS). Első lépésként nagy számú vegyületet – egy molekuláris könyvtárat – szűrünk át találatok után kutatva. Egy könyvtár alatt érthetjük valódi vegyületek gyűjteményét, de egy virtuális vegyületkönyvtárat is. Ezután különböző tulajdonságok alapján a találatokból egy kisebb molekulahalmazt válogatunk ki. A vezérmolekulákat és analógjaikat ezután optimalizáljuk és preklinikai kísérletekben vizsgáljuk.

A preklinikai fázis kettős szerepet tölt be: az *in vitro* és állatkísérletek minimalizálják a toxicitásból adódó kockázatokat az emberi alanyokon végzett klinikai vizsgálatok megkezdése előtt, másrészt csökkenti az esélyét, hogy sikertelen klinikai vizsgálatot kezdjünk, hatalmas anyagi veszteséget szenvedve el ezzel. Az analógok tesztelésekor szerzett adatokat továbbá felhasználjuk arra, hogy a struktúra–hatás összefüggéseket modellezzük a vezérmolekula körüli kémiai térben.

A preklinikai kiértékelést követően önkéntesek részvételével sor kerül a klinikai vizsgá-

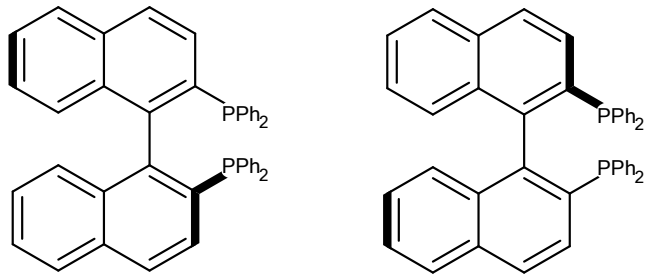
latra, hogy meghatározzák a gyógyszer biztonságossági profilját és hatásosságát. A klinikai vizsgálat folyamata három hagyományos (I, II, III) és egy további posztmarketing (IV) fázisra oszlik. Ez alatt a biztonságos humán dózisok meghatározásra kerülnek (I. fázis) és az adott egészségügyi állapotra vonatkozó hatásosság placebo-kontrollált körülmények között kerül vizsgálatra több lépésben, növekvő mintaméret mellett (II. és III. fázis). A mellékhatások gyűjtése folyamatos az I. fázistól kezdve a posztmarketing fázisig, mikor a gyógyszer már a piacon van. A klinikai vizsgálat teljes folyamatát statisztikai monitorozásnak vetik alá – úgynevezett interim analízis zajlik –, amely lehetővé teszi, hogy a folyamatot leállítsák etikai vagy gazdasági okokból.

19.2. Kemoinformatikai háttér

Ahhoz, hogy egy megfelelő tulajdonságokkal rendelkező új, farmakológiailag aktív vegyületre bukkanjunk, néha több mint egy-millió vegyületet kell megvizsgálnunk. Egy ilyen hatalmas adatbázis vegyületeit nem lehet gazdaságosan megszintetizálni, első lépésként tehát gyakran egy virtuális könyvtáron végezzük el a szűrést: nagy számú, a kereskedelemben elérhető, vagy adott esetben csak sejthetően szintetizálható vegyülethalmazon, melyek között lehetnek olyanok, amiket még soha sem szintetizáltak. A virtuális könyvtárat reprezentáló adatbázis a vegyületek szerkezete mellett tartalmazhat számos számított tulajdonságot. Általánosságban véve egy kémiai szerkezet definiálható az atomok címkézett szomszédossági mátrixával (gráf reprezentáció), kiegészítve további információval a részstruktúrák térbeli relatív helyzetéről.

Egy adott atom-atom kapcsolódási hálózat számos háromdimenziós szerkezetet reprezentálhat. Ha a háromdimenziós szerkezetek egy halmaza szobahőmérsékleten a termikus mozgás révén szabadon egymásba alakulhat, akkor a struktúrákat azonos vegyületnek tekintjük, az egyes szerkezetek a vegyület **konformerei**. Tehát az energiagát két konformer között olyan alacsony, hogy a gyakorlatban nem izolálhatók, minden konformer megtalálható egyazon mintában a Boltzmann eloszlásnak megfelelő valószínűséggel. Ha relatíve nagy energiagát van 3D szerkezetek két halmaza között, a két halmaz két elkülöníthető vegyületet reprezentál, melyek **izomerek**. Ennek egy speciális esete, ha az atomok kapcsolódása azonos, csak a háromdimenziós szerkezet tér el két vegyület között: ezeket **sztereoizomereknek** hívjuk. A fogalom **kiralitás**ként (A görög kéz szóból, jelentése „kézszerű") is ismert. Egy királis objektum meghatározó tulajdonsága, hogy nem hozható fedésbe tükörképével.

Hogy kódolhassuk a két sztereoizomer közötti különbséget, ki kell egészítenünk a molekulagráfot további információkkal. Például olyan négy vegyértékű szén esetén, melynek mind a négy szubsztituense eltérő, két eltérő kapcsolódási sorrendet különböztethetünk meg. Az ilyen atomok – ún. kiralitás centrumok – és más királis elemek címkézésére egy konvenciót, a Cahn–Ingold–Prelog-szabályt (CIP-szabály) alkalmazzák. A lehetséges címkék: S (*Sinister*, latinul bal) és R (*Rectus*, latinul jobb). A CIP-konvenció alapötlete, hogy felcímkézünk minden szubsztituenst a centrumhoz közvetlen kapcsolódó atom rendszáma szerinti sorrendben iteratívan, majd a molekulát úgy helyezzük el a térben, hogy



19.1. ábra. Példa az axiális kiralitásra. A BINAP (2,2'-bisz(difenilfoszfino)-1,1'-binaftil) két izomerének kétdimenziós ábrája. A Ph fenil csoportot jelöl, a kivastagított vonalak azon kötéseket melyek a kép síkja fölött vannak.

a legkisebb számmal jelzett szubsztituens a papír síkja alatt helyezkedjen el. Ekkor a másik három szubsztituens vagy az óra járásának megfelelő, vagy azzal ellentétes módon számozódik. A pontos szabály megtalálható bármely szerves kémia tankönyvben vagy az IUPAC vonatkozó ajánlásában [1, 2].

Vannak a kiralitásnak speciálisabb esetei, úgymint az axiális kiralitás (lásd a 19.1. és 19.2. ábrát). Vegyületek egy csoportja, a helicének, melyek összekapcsolt aromás gyűrűkből állnak, háromdimenziós spirált alkotnak. A helicénekben nem található kiralitáscentrum, mégis két formájuk létezik: egy az óramutató járásának megfelelő és egy azzal ellentétes csavarmenettel.

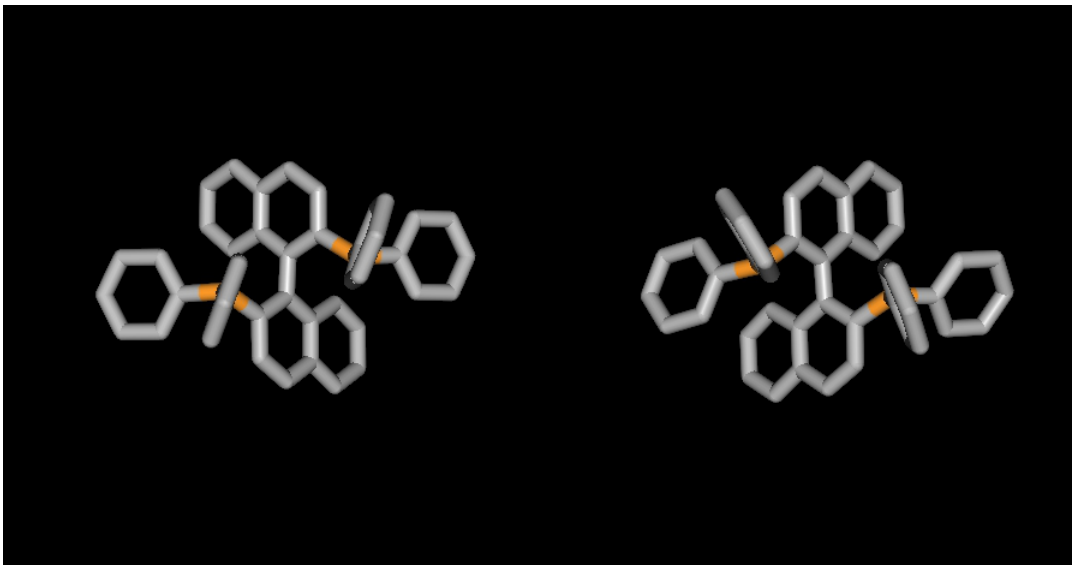
Biológiai rendszerekben az eltérő sztereoizomereknek jelentősen eltérő hatásuk lehet, mivel a molekuláris célpont és a hatóanyag geometriai illeszkedése elengedhetetlen. Egy kiroszelektív rendszerben az illeszkedési pontok minimális száma három. További feltétel, hogy ezen interakciók hozzájárulása a kötési energiához közel azonos legyen, ellenkező esetben kevesebb, mint három interakció dominálja a kötődést, és az izomerek affinitásában csak csekély különbség lép fel. Például a talidomid nevű szedatív szer (S) sztereoizomere teratogén. Ezt a szert eredetileg terhes anyák reggeli rosszulléteinek kezelésére fejlesztették és Contergan márkanéven volt forgalomban. A talidomid jó példa egy másik jelenségre is, melyet racemizációnak nevezünk: vannak vegyületek, melyek izomerjei átalakulhatnak egymásba biológiai rendszerekben jelen lévő enzimek segítségével. Ebből következően a tiszta (R)-talidomid szintén teratogén tulajdonságokat mutat. Ahogy még a fejezet későbbi részében látni fogjuk, még ez a veszélyes vegyület is használható számos új indikációban, ahol a terhesség kizárható.

Egy molekula adott célpontra mutatott affinitása egy disszociációs állandóval definiálható, melyet általában K_d jelöl. Adott az alábbi reakció:

$$T + L \leftrightarrow TL,$$

ahol T a ligandum mentes célpontot, L a szabad ligandumot és TL a komplexet jelöli. K_d dimenziója moláris koncentráció, és definíciója

$$K_d = \frac{[T][L]}{[TL]},$$



19.2. ábra. A BINAP izomerek háromdimenziós szerkezete. Nincs aszimmetrikus szénatom a molekulákban.

ahol a kapcsos zárójelek egyensúlyi moláris koncentrációkat jelölnek [3].

Minél kisebb a K_d, annál aktívabb a vegyület. Az 1uM affinitás azt jelenti, hogy a célpontok fele komplex formájában van jelen a modulátor 1uM/l koncentrációjú oldatában, mivel ha $[L] = K_d$, akkor

$$K_d = \frac{[T]}{[TL]} K_d,$$

tehát

$$\frac{[T]}{[TL]} = 1.$$

A kölcsönhatás erősségét a Gibbs-szabadentalpia segítségével fejezhetjük ki. A két mennyiség közötti kapcsolat:

$$\ln K_d = \frac{\Delta G}{RT},$$

ahol T a rendszer hőmérséklete és R az egyetemes gázállandó.

19.3. Szűrési kritériumok

A farmakológiai tulajdonságok két fő csoportra oszthatók: farmakodinámiás (PD) és farmakokinetikai (PK) tulajdonságokra. A farmakodinámia általában azt írja le: „Hogyan hat a gyógyszer a biológiai rendszerre?”, úgymint mi a célpont, mennyire potens a gyógyszer, mennyire szelektív a ligandum és hasonlók. A farmakokinetika arra kérdez rá: „Hogyan hat a biológiai rendszer a gyógyszerünkre?”, úgymint: hogyan történik a vegyület szállítása, elosztása, átalakítása a szervezetben.

Egy gyógyszerfejlesztési folyamatban a várható biológiai aktivitás csak egy a számos teljesítendő kritérium közül. További nagyon fontos kritériumok egy csoportjára utal az angol ADMET betűszó: Absorption, Distribution, Metabolism, Excretion and Toxicity, azaz Felvétel, Eloszlás, Metabolizmus, Kiválasztás és Toxicitás.

A legegyszerűbb mód a kinetika leírására, ha a molekulákat fizikokémiai tulajdonságaik segítségével írjuk le, úgymint oldhatóság, poláris felszín, lipofilicitás, molekulatömeg stb., melyek alacsony átlagos hibával becsülhetők tisztán számításos úton. Egy klasszikus kísérlet a nem gyógyszerszerű vegyületek kiszűrésére a Lipinski-féle ötös szabály alkalmazása. Ez a szabály orálisan aktív gyógyszerek esetén maximálja a hidrogénkötés-donorok számát 5-ben, az akceptorokét 10-ben, a molekulatömeget 500-ban, és az oktanol-víz megoszlási hányadost (lásd az alábbi keretes részt) 5-ben [4]. Ezek alól a szabályok alól természetesen vannak kivételek. Egy másik hasonló szabály a szigorúbb „Hármas szabály" a fragmens alapú tervezés területén (nem azonos a Jörgensens-féle hármas szabállyal), mely a hidrogénkötés-donorok és akceptorok számát 3-3-ban, a molekulatömeget 300-ban, az oktanol–víz megoszlási hányadost pedig 3-ban maximálja [5]. Ezek a tulajdonságok nem csak jól becsülhetők, de relatíve könnyen hangolhatók is a vezérmolekula kémiai módosításával.

Oktanol–víz megoszlási hányados (LogP)

A megoszlási hányadost két határfelületükön egymással egyensúlyban lévő nem elegyedő oldatban mért koncentráció arányával definiáljuk.

$$\log P = \log \frac{[L]_{octanol}}{[L]_{water}},$$

ahol L a vegyület nem ionizált formája. A logP a lipofilicitás mértékének tekinthető. Ha a logP alacsony, a vegyületet hidrofilnak, ha magas, lipofilnak nevezzük.

Egy koncepcionálisan eltérő farmakokinetikai terület a metabolizmus, melynek becslése jóval nehezebb. A lehetséges metabolikus reakciók általában megjósolhatók azáltal, hogy reakciós mintákat illesztünk a vizsgált vegyületekre, de számos erősen aspecifikus enzim kötődés-profilját kell számításba venni, hogy a valóban releváns metabolikus útvonalat azonosítani lehessen. A metabolizmus célja, hogy az idegen anyagot vízoldhatóbbá tegye és elősegítse a kiválasztását. A folyamat két fő részre osztható: A Fázis I. metabolikus reakciók általában oxidatívak, míg a Fázis II. metabolikus folyamatokban endogén vegyületek konjugálódnak az idegen anyagra. Például egy nagy oxidáz családnak, a Citokróm P450 családnak – általános rövidítésük CYP –, kiemelkedő szerepe van számos gyógyszer hepatikus metabolizmusában.

A metabolizmus ugyanakkor a farmakogenomika egyik első területe is, és ezeknek az enzimeknek számos polimorfizmusát azonosították gyógyszerek személyenként eltérő hatásával kapcsolatban. Néhány esetben, mint a warfarin és a CYP2C9 egyes polimorfizmusai, az asszociációt feltüntetik a gyógyszer betegtájékoztatóján is, és a genotipizálást a klinikai gyakorlatban is alkalmazzák – segítve ezzel a dózis beállítását [6]. Számos más specifikus kölcsönhatás húzódhat még meg a gyógyszerek farmakokinetikai tulajdonságai mögött,

mint transzporterekhez és szövetspecifikus enzimekhez való kötődés, tehát a PK probléma egyszerű fizikokémiai alapú kezelésének lehetőségei korlátozottak.

A farmakodinámiás tulajdonságok becslésének problémája természeténél fogva komplexebb. Általában feltételezzük, hogy a gyógyszer hatását egy vagy több, a kismolekula és egy molekuláris célpont között létrejött specifikus kötődési kölcsönhatás közvetíti. Ugyanakkor a célpontok száma nagy lehet az ún. piszkos vegyületek esetében, illetve aspecifikus vagy ellentmondásos lehet mint például az etanol és a lipid membránok kölcsönhatásai.

Miután néhány kedvező tulajdonságokkal rendelkező találatot kiválasztottunk, a következő lépés az optimalizáció. Ebben a lépésben a jelölt számos analógját szintetizáljuk és szűrjük azzal a céllal, hogy jobb jelölteket találjunk. A kiválasztási kritériumok között ebben a fázisban nem csak az aktivitás, de a fent említett további fontos tulajdonságok is szerepelnek. Egy ún. QSAR (Quantitative Structure-Activity Relationship) modellt illeszthetünk az analógszűrés eredményeire, hogy aztán egy iteratív folyamatban valószínűleg jobb tulajdonságokkal rendelkező vegyületeket tervezhessünk. Ennek során a jelölt molekulatömege és lipofilicitása tipikusan növekszik. A növekvő méret problematikus lehet, tekintettel az ADME tulajdonságokra, lásd például a Lipinski szabályokat, ezért az egyensúly megtartása fontos. Egy mérőszám az ún. ligandum-hatékonyság széles körben használatos, amivel figyelembe vehető a méret és aktivitás egymással ellentétes hatása:

$$LE = \frac{\Delta G}{N_{hv}},$$

ahol N_{hv} a nem-hidrogénatomok száma, az ún. nehézatom-szám. Állandó hőmérsékletet feltételezve ΔG és $\log K_d$ felcserélhetőek egymással. Az alábbi metrikák definiálásához ΔG-t fogjuk használni, de számos más aktivitás vagy affinitás jellegű mennyiség használható a gyakorlatban, például a pK_d vagy a pIC_{50}*. A mérőszám egy módosított verzióját is javasolták, hogy korrigálják a molekulaméret – átlagos aktivitás összefüggés nemlinearitását. Ezt a mutatót SILE-nek (size-independent ligand efficiency) nevezik:

$$SILE = \frac{\Delta G}{{N_{hv}}^{0.3}}.$$

A definíciós formula alakja azzal magyarázható, hogy az energia-hozzájárulás részben a molekula-térfogattal, részben az oldószer által elérhető molekulafelszínnel arányos [7].

Egy további hatékonysági mérték az LLE (lipophilic ligand efficiency) az alacsony lipofilicitás és a nagy affinitás közti egyensúly elérését segíti:

$$LLE = \Delta G - logP;$$

illetve egy általános mérőszám mindkettőre az LELP (ligand efficiency-dependent lipophilicity):

$$LELP = \frac{logP}{LE}.$$

*Az IC_{50} az az inhibitor-koncentráció, amely mellett a vizsgált enzim aktivitása fele az inhibitor nélkül mérhetőnek [3].

Ezt a mérőszámot optimalizálás során minimalizáljuk, ellentétben a korábban tárgyaltakkal. A megalkotóik szavaival élve: azt az árat fejezi ki, amit lipofilicitásban fizetnünk kell egy egységnyi ligandum hatékonyságért [8].

Mélyebb elméleti nézőpontból tekintve a molekulaméret és lipofilicitás növekedése az entrópia-vezérelt optimalizációs stratégiának tulajdonítható. Hogy áttekintést kaphassunk az entrópia- és az entalpia-vezérelt optimalizáció természetéről, vessünk egy pillantást a Gibbs-szabadentalpia definíciójára:

$$\Delta G = \Delta H - T\Delta S,$$

ahol ΔH a nettó entalpiaváltozás és ΔS a nettó entrópia változás a kötődési folyamat alatt. A Gibbs-szabadentalpia optimalizálható ΔH minimalizálásával – entalpia-vezérelt stratégia –, vagy ΔS maximalizálásával – entrópia-vezérelt stratégia.

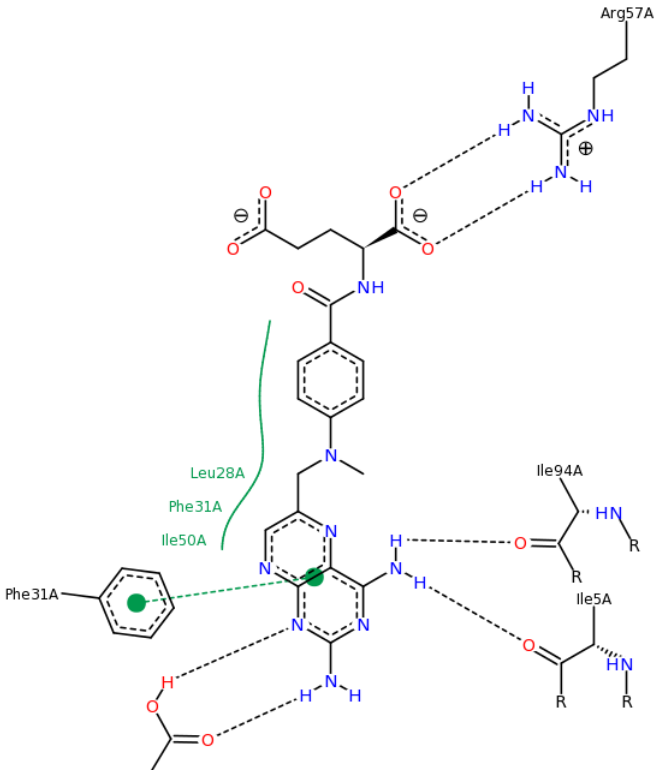
A gyakorlatban nagyon nehéz pusztán az egyik tag optimalizálása anélkül, hogy jelentős kompenzáció lépne fel a másikban. Például ha egy erős kölcsönhatást tervezünk a ligandum és a célpont közé, ez korlátozni fogja a ligandum konformációs flexibilitását és entrópiabüntetést eredményez [9].

Az entalpia-tag fő komponenseit a célpont és a ligandum közötti poláris kölcsönhatások – például hidrogénkötés – (kedvező) és a víz, valamint a ligandum/kötőhely poláris csoportjainak kölcsönhatása (kedvezőtlen) adják. Az entrópia-tag komponensei a szolvatációs entrópia és a konformációs entrópia. A szolvatációs entrópiaváltozás kedvező, azt a taszító kölcsönhatást reprezentálja, mely a lipofil csoportok és a víz között lép fel, de ez a kötődési folyamat egy nyilvánvalóan nem szelektív komponense. A konformációs entrópiaváltozás kedvezőtlen, melyet a konformációs tér szűkülése okoz a kötődés során. A fentiekből nyilvánvalóan látszik, hogy egy nagy lipofil molekulának nagy affinitása lehet. Tudjuk azonban, hogy az affinitás csak egy a paraméterek közül amit optimalizálni szeretnénk.

19.4. Módszerek

Ha a molekuláris célpont ismert, az aktív modulátorok keresését a szerkezetre vonatkozó információk segítségével végezhetjük, esetlegesen ismerve az ismert modulátorokkal – mind endogén, mind exogén – történő kölcsönhatásokat. Azokat a módszereket, melyek feltételezik, hogy a célpont szerkezete ismert, szerkezet alapú módszereknek nevezzük. A módszerek másik csoportja – az ún. ligandum alapú módszerek – csak az ismert aktív vegyületek struktúrájára épít és olyan modellek építését célozza, melyekkel azonosíthatók a közös strukturális jegyek vagy a szerkezet–hatás összefüggések.

A célpont–ligandum kölcsönhatás legegyszerűbb modellje a kulcs-zár modell. Ebben feltételezzük, hogy a célpont rendelkezik egy specifikus, relatíve merev felszínű régióval – a kötőhellyel – és a ligandum valamely konformációja tökéletesen beleillik ebbe a zsebbe. A geometria mellett más tulajdonságok egyezésére is szükség van, amit a töltések, a hidrogénkötések és hidrofób helyek határoznak meg (lásd a 19.3. ábrát). A kölcsönhatás egy összetettebb modellje az indukált illeszkedés modellje. Ebben nem csak a ligandumot



19.3. ábra. A Methotrexate (MTX) kötött állapotban célpontjának, a Dihidrofolát reduktáz enzimnek az aktív helyén. (Forrás: RCSB PDB, 1RG7) Célpont–ligandum kölcsönhatások: π-π stacking a pteridin gyűrű és a fenilalanin aromás oldallánca között (zöld pötty), egy leucin, egy izoleucin és egy fenilalanin oldallánc hidrofób kölcsönhatása az MTX középső régiójával, ionos kölcsönhatás az MTX egyik karboxil-csoportja és egy pozitívan töltött arginin oldallánc között, valamint három hidrogén-kötés az MTX egy aromás nitrogénje továbbá két amin-csoportja részvételével.

tekintjük flexibilisnek, hanem a célpontot is. Ahogy a ligandum a kötőhelyhez közeledik, kölcsönös erők ébrednek a ligandum és a célpont között, melyek konformációs változásokat indukálnak a kölcsönható felekben.

A szerkezet alapú módszerek egy példája a molekuláris dokkolás, amely egy geometria alapú módszer és segítségével megbecsülhető a molekulák komplexének szerkezete és a kölcsönhatás erőssége. A dokkolási eljárás egy állapotteres keresési algoritmus az alábbi optimalizációs probléma megoldására: meg kell találni a ligandum optimális orientációját a célponthoz viszonyítva, és ki kell értékelni a kölcsönhatások erősségét egy klasszikus fizikai tényezőket tartalmazó közelítő pontozófüggvény segítségével. A dokkolást merev testek segítségével is végre lehet hajtani, illetve köztes esetnek tekinthetjük, ha a receptor merev, de a ligandum flexibilis. A dokkolás egy sokkal számításintenzívebb verziója az indukált illeszkedést is számításba veszi.

Az optimalitási kritérium a dokkolás során lehet egy empirikus pontozófüggvény, vagy a komplex becsült potenciális energiája, amely egy erőtérrel: heurisztikusan meghatáro-

zott függvénnyel és annak paramétereivel van definiálva. Általánosságban az energiát egy összeg formájában írják fel, mint például:

$$E = E_{bond} + E_{angle} + E_{dihedral} + E_{VdW} + E_{Coulomb}.$$

A használt erőtértől függően a hozzájárulások alakja és a paraméterek eltérnek. A paramétereket empirikusan hangolják be kísérletes eredmények és nagy pontosságú kvantumkémiai számítások segítségével.

Például a kötéshosszra vonatkozó potenciál lehet egyszerű harmonikus, vagy lehet Morse-potenciál:

$$V_M = D_e \left(1 - e^{-a(r-r_e)}\right)^2,$$

ahol D_e a disszociációs energia, r_e az egyensúlyi kötéshossz és a a szélesség paraméter.

A Van der Waals-potenciál Lennard–Jones-potenciállal közelíthető:

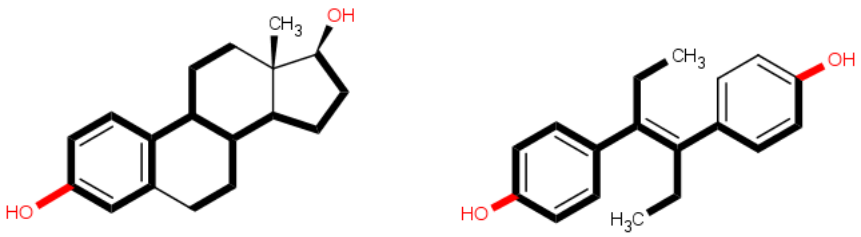
$$V_{LJ} = 3\epsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right],$$

ahol ϵ a potenciálárok mélysége és σ az a távolság, ahol a potenciál nulla. Számos más alakú függvényt is használnak a fent említett példákon túl. Dokkolás esetén a modellezett folyamat víz jelenlétében játszódik le, tehát gyakran vezetnek be további tagot a szolvatáció implicit modellezésére.

A ligandum alapú QSAR és QSPR (Quantitative Structure-Property Relationship) széles körben elfogadott és népszerű eljárások a gyógyszertervezésben. Ezeket a kifejezéseket összefoglalóan használjuk minden statisztikai modellre, ami kapcsolatokat ír le valamely tulajdonság (mint aktivitás QSAR esetén, vagy valamely fizikokémiai tulajdonság QSPR esetén) és a kémiai struktúra között. Ezek a modellek általában a kémiai tér valamely korlátozott tartományában érvényesek: az analógok egy halmazán. Számos statisztikai módszer alkalmas QSAR modellek építésére, pl.: regressziós módszerek (általában dimenziócsökkentéssel, mint a PLS), neurális hálózatok, SVMek.

Ha a molekuláris célpont nem ismert, számos hasonlóság alapú keresési módszer használható. Ezeknek a módszereknek számos közös tulajdonsága van a QSAR modellezéssel. Mindkét esetben az első lépés a vegyületek reprezentációját szemantikailag értelmezhető formára transzformálni. Egy lehetséges megoldás az ujjlenyomatok készítése. Ebben az esetben a szerkezeteket szekvenciális adattá, általában bináris sztringgé vagy számok sorozatává alakítjuk. Minden szám egy elemi tulajdonságnak felel meg, mint például egy szerkezeti elem megléte. A strukturális kulcsok a gráfreprezentáción vagy akár a háromdimenziós szerkezeten is kiértékelésre kerülhetnek. A 3D ujjlenyomatok egy speciális esete a farmakofór ujjlenyomatoké. A farmakofór jelentése gyógyszer- (pharmacon) tulajdonságok hordozója (phoros); strukturális elemek egy halmaza és ezek relatív orientációja melyet a célpont felismer. Normális esetben sokkal több elkülöníthető tulajdonság létezik, mint ahány bitünk egy molekula reprezentációjára rendelkezésre áll, ezért egy alacsony ütközési valószínűséggel rendelkező hash-függvényt használunk, hogy tömörítsük az ujjlenyomatot.

A fent említett ligandum alapú módszerek nyilvánvaló összhangban vannak a hasonló tulajdonságok elvével: ha két molekula nagyon hasonló, a tulajdonságaik is valószínűleg



19.4. ábra. Ösztradiol (balra) és Dietilstilbösztrol (jobbra). A vastagított kötések a közös részstruktúrát mutatják. A jobb oldali molekulában két kötés hiányzik a váz B és C gyűrűjéből, lehetővé téve ezzel a molekula egyes részeinek szabad rotációját. Ugyanakkor egy bevezetett kettős kötés valamelyest korlátozza a konformációs tér méretét. Részletes konformációanalízisért lásd Wiese és munkatársai munkáját [10].

hasonlók. A klasszikus módszerek fő hátránya, hogy a kiindulási pont szűk környezetében keresik az új vegyületeket. Egy hasonló farmakológiai tulajdonságokkal, de eltérő alapvázzal rendelkező molekula hasznos lehet egyes esetekben, például nagyon gyenge ADME tulajdonságok esetén, vagy ha szabadalmi probléma merül fel. Ez a szükséglet látszólag ellentmondásban van a hasonló tulajdonságok elvével. A konfliktus megoldását az alapváz ugrás (scaffold hopping, core hopping) módszere nyújthatja. Ahelyett, hogy az oldalláncokat módosítjuk, a molekula alapvázát transzformáljuk szisztematikusan, vagy teljesen lecseréljük úgy, hogy a szerkezet lényegi elemei ne változzanak meg. Többé-kevésbé folyamatos a spektrum az egy-atom helyettesítéses módszerektől az új alapváz tervezéséig. Jó példát szolgáltatnak a köztes módszerekre a gyűrűmanipulációk. Farmakodinámiás értelemben egy merev molekula magas összekötöttséggel előnyös, mert a merev struktúrának kevesebb konformere van, tehát a célponthoz való kötődés energetikailag kedvezőbb: a rendszer entrópiavesztesége mérsékeltebb. Ha van egy flexibilis molekulánk és ismerjük ennek aktív konformációját, rögzíthetjük a molekulát ebben a konformációban egy gyűrűzáró kötés bevezetésével. További előnyös tulajdonsága egy merev molekulának a magasabb szelektivitás. A sok előnynek ugyanakkor ára is van. Egy merev rendszer számos gyűrűvel általában rosszabb oldhatósággal rendelkezik és ADME tulajdonságai rosszabbak. Néha fel kell nyitnunk gyűrűket, hogy kedvezőbb ADME tulajdonságokkal rendelkező rendszereket hozzunk létre, vagy szándékosan csökkentsük a vegyület hatását egy adott célponton.

A dietilstilbösztrol, egy a 40-es évektől a 70-es évekig széles körben használ szintetikus ösztrogén, nagyon hasonló az ösztradiol egy gyűrű-felnyitott analógjához (lásd a 19.4. ábrát).

19.5. Fragmens alapú tervezés

Egy biztató, a nagy áteresztőképességű módszereket kiegészítő megközelítés a fragmens alapú tervezés. Ebben a megközelítésben jelentősen kevesebb vegyületet szűrünk le a molekuláris célponton. Ez a kisebb könyvtár kicsi molekulákat tartalmaz, és a cél olyan

kis kölcsönhatások detektálása, amik felhasználhatók egy nagy affinitású jelölt fragmensekből történő felépítésére. Ez a nagy érzékenységet követel meg, mely arra készteti a vegyészeket, hogy nagy információ tartalmú kísérleti módszereket, például NMR spektroszkópiát alkalmazzanak *in silico* módszerek helyett. Ez a módszer kísérletektől való függőségéhez vezet, habár újabban egyre többen tesznek kísérletet fragmensek azonosítására számításos módszerekkel is. Egy erre alkalmas módszer lehet a dokkolás [11]. Az affinitás meghatározására használt módszer – legyen bár kísérleti vagy *in silico* – strukturális információkkal szolgálhat a gyenge kölcsönhatásokról, lehetővé téve, hogy olyan egymással nem átfedő fragmensekből, melyek közeli kötőhelyeken kötődnek, ligandumot építsünk fel. Egy megfelelő *in silico* eljárás lehet erre a dokkolás. Ha a nem átfedő fragmenseket azonosítottuk, megfelelő linkerek tervezhetők közéjük. Átfedő fragmensek esetén összeolvasztásos stratégia használható. Ez a fajta „oszd meg és uralkodj" stratégia nagy kémiai tér bejárását teszi lehetővé exponenciális mértékű erőforrás-megtakarítás mellett. Egy minden lehetséges gyógyszerszerű vegyületet reprezentáló halmazzal történő szűrés lehetetlen a kémiai tér méretei miatt, de a kis méretű fragmensek terében ez egy realisztikus cél lehet. A molekuláris célpont karakterizálható egy fragmens-szűrés segítségével, így a célpont „gyógyszerelhetősége" megbecsülhető. A fragmens alapú megközelítés segíteni tudja a vezérmolekula-optimalizálás fázisát is, mivel a fragmenseket valamely ligandumhatékonyság alapú kritérium segítségével választhatjuk ki, tehát a molekulatömeg és a lipofilicitás kontrollálható.

19.6. Gyógyszer-újrapozicionálás

A gyógyszer-újrapozicionálás (drug repositioning) egy kifejezés, arra a gyakorlatra utal mikor egy már elfogadott hatóanyagot újrahasználunk egy új terápiás indikációban. Ez a koncepció népszerűségét annak köszönheti, hogy költséghatékony: a biztonságossági és toxicitásvizsgálatok már egyszer lezajlottak, és az eredményeik – vagy azok egy része – újra felhasználható. Az újrapozicionálás kontextusában sokkal gazdagabb információforrások állnak rendelkezésre, úgymint már ismert mellékhatások, indikációk, már ismert molekuláris célpontok és hasonlók. A gyógyszerkutatás történetében számos véletlenszerű újrapozicionálás történt. Egy jól ismert példa a sildenafil esete, melyet eredetileg kardiológiai indikációkra fejlesztettek ki (angina pectoris, magas vérnyomás) majd később Viagra márkanéven került forgalomba mint erektilis diszfunkció kezelésére szolgáló gyógyszer. A két indikáció közös tulajdonságát a gyógyszer értágító hatása célozza meg, melyet annak egy foszfodiészteráz altípuson a PDE5-ön mutatott gátló hatása közvetít.

A gyógyszer-újrapozicionálás hatékony eszköze a ritka betegségek elleni gyógyszerfejlesztésnek is. A ritka betegség és a hozzá társuló „orphan drug" számos országban jogi kategória, intuitíven úgy definiálható, mint egy olyan betegség (és a kezelésére szolgáló gyógyszer), mely olyan ritka, hogy a gyógyszerfejlesztés klasszikus megközelítései nehezen kivitelezhetők és nagyon gazdaságtalanok. Például a korábban említett teratogén gyógyszer, a talidomid újrapozicionálható néhány lepra-típus és daganatos megbetegedés ellen, továbbá immunszupresszáns tulajdonságokkal is rendelkezik. Nincs éles határ az „orphan

drug" koncepció és a „valódi" személyre szabott medicina között, mivel számos ritka betegséget ritka genetikai mutációk okoznak, és extrém esetben a betegség kezelése erősen betegspecifikus kell, hogy legyen.

A gyógyszer-újrapozicionálás kontextusában az adatfúziós technikák (melyeket a „Heterogén biológiai adatok fúziós elemzése" című fejezetben tárgyalunk) különösen hasznosak lehetnek [12]. Számos különböző típusú információforrással rendelkezünk, úgymint a kémiai szerkezet, a mellékhatások, genetikai faktorok, a molekuláris célpontok, érintett biokémiai útvonalak stb. A hasonlóság alapú megközelítés kiterjeszthető ezekre az adatforrásokra is. Igen gazdag adatbázis – például számos fenotípusos információ – nyerhető korábbi vizsgálatokból és a posztmarketing információkból. A fenotípus – a fogalom tradicionális értelmezésében – statikus tulajdonság, az organizmus egy megfigyelhető jellegzetessége. Gyógyszerhatóanyagok esetén a „kémiailag gerjesztett" biológiai rendszer néhány tulajdonságát vizsgáljuk, mint a biokémiai változásokat, hatásokat, mellékhatásokat. A mellékhatás alapú hasonlósági mértéket például Campillos és munkatársai javasolták 2008-ban [13]. A hipotézis az alábbi volt: ha két gyógyszernek számos mellékhatása közös, feltehetően van közös molekuláris célpontjuk, vagy legalább vannak olyan célpontjaik, melyek egyazon biokémiai útvonalon helyezkednek el.

A gyógyszer-újrapozicionálás területén elérhető információk gazdagsága ideális határterületté teheti azt a gyógyszerkémia, biológia és a „big data" kutatások számára.

Irodalomjegyzék

[1] Lajos Novák and József Nyitrai, *Szerves kémia*. 2001.

[2] International Union of Pure and Applied Chemistry. Commission on the Nomenclature of Organic Chemistry, R. Panico, W. H. Powell, and J. C. Richer, *A Guide to IUPAC Nomenclature of Organic Compounds: Recommendations 1993*. IUPAC chemical data series. Blackwell Scientific Publications, 1993.

[3] Kenneth A. Krohn and Jeanne M. Link, Interpreting enzyme and receptor kinetics: keeping it simple, but not too simple. *Nuclear Medicine and Biology*, 30(8):819–826, 2003. Workshop on Receptor-Binding Radiotracers 2003.

[4] Christopher A. Lipinski, Franco Lombardo, Beryl W. Dominy, and Paul J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 23(1–3):3–25, 1997.

[5] Miles Congreve, Robin Carr, Chris Murray, and Harren Jhoti, A 'Rule of Three' for fragment-based lead discovery? *Drug Discovery Today*, 8(19):876–877, 2003.

[6] Guruprasad P. Aithal, Christopher P. Day, Patrick J. L. Kesteven, and Ann K. Daly, Association of polymorphisms in the cytochrome P450 CYP2C9 with warfarin dose requirement and risk of bleeding complications. *The Lancet*, 353(9154):717–719, 1999.

[7] J. Willem M. Nissink, Simple size-independent measure of ligand efficiency. *Journal of Chemical Information and Modeling*, 49(6):1617–1622, 2009. PMID:19438171.

[8] György G. Ferenczy and György M. Keserű, Thermodynamics guided lead discovery and optimization. *Drug Discovery Today*, 15(21–22):919–932, 2010.

[9] Adam J. Ruben, Yoshiaki Kiso, and Ernesto Freire, Overcoming roadblocks in lead optimization: A thermodynamic perspective. *Chemical Biology & Drug Design*, 67(1):2–4, 2006.

[10] T. E. Wiese, D. Dukes, and S. C. Brooks, A molecular modeling analysis of diethylstilbestrol conformations and their similarity to estradiol-17 beta. *Steroids*, 60(12):802–808, 1995.

[11] Huameng Li and Chenglong Li, Multiple ligand simultaneous docking: Orchestrated dancing of ligands in binding sites of protein. *Journal of Computational Chemistry*, 31(10):2014–2011, 2010.

[12] A. Arany, B. Bolgar, B. Balogh, P. Antal, and P. Matyus, Multi-aspect candidates for repositioning: Data fusion methods using heterogeneous information sources. *Current Medicinal Chemistry*, 20(1):95–107, 2013-01-01T00:00:00.

[13] Monica Campillos, Michael Kuhn, Anne-Claude Gavin, Lars Juhl Jensen, and Peer Bork, Drug target identification using side-effect similarity. *Science*, 321(5886):263–266, 2008.

20. fejezet

Metagenomika

20.1. Bevezetés

A mikrobák mindenütt ott vannak. Az $5 * 10^{30}$-ra becsült bakteriális és archaea sejt (azaz a prokarióták) az alapvető tápanyagok (szén, nitrogén, foszfor) legnagyobb raktárai a Földön, és egyes becslések szerint a biomassza legnagyobb részét is ezek alkotják [1]. Bolygónkon rengeteg olyan extrém környezet található, ahol csak a prokarióták képesek a túlélésre, legyen az rendkívül meleg, hideg, savas vagy sós hely. Léteznek mikrobák, amelyek képesek a természetben előforduló toxinok vagy az emberi tevékenységek melléktermékeként keletkező mesterséges toxinok (pl. olajfoltok) lebontására. Bár többnyire szabad szemmel nem láthatók, a mikrobák valójában létfontosságúak a Földön élő minden életforma, köztük az ember számára is [2]. A mikrobák alakítják vissza az élettelen anyagot abba a formába, amelyet már minden más élőlény közvetlenül fel tud használni. Majdnem minden többsejtű eukarióta élőlény szoros szimbiózisban él olyan mikrobiális közösségekkel, amelyek létfontosságú tápanyagokat és vitaminokat állítanak elő a gazdaszervezet számára. Az emésztőrendszerünkben és szánkban élő mikroorganizmusok teszik lehetővé, hogy kinyerjük az energiát azokból az ételekből, amelyek egyébként emészthetetlenek lennének. A bennünk és rajtunk élő komplex mikrobiális közösségek aktívan részt vesznek a betegséget okozó ágensek elleni védelemben. Valójában az emberi test egyfajta szuperorganizmusnak is tekinthető, hiszen a saját kb. 10^{13} darab sejtünknél mintegy 10-szer több, 10^{14} baktérium él a szervezetünkben [1, 2].

Az 1995-ben végzett első bakteriális teljes genom projekt óta [3] a mai napig ezernél is több baktérium genomi szekvenciája vált ismertté. Ezek a vizsgálatok és az általuk szerzett nagy mennyiségű adat és tudás nagyban elősegítették a komparatív genomika és a rendszerbiológia tudományának fejlődését. Mindazonáltal – az így szerzett hatalmas mennyiségű adat és tudás ellenére – az egyetlen organizmuson végzett kutatásoknak szükségszerű korlátai vannak: Először is, annak érdekében, hogy egy mikroba teljes genomját meg lehessen szekvenálni, a jelenlegi technológiai elvárások szerint az adott organizmust először ki kell tenyészteni. Ez pedig nagyon ritkán sikerül, ugyanis a természetben élő mikróbáknak csak nagyon kis százalékát lehet laboratóriumi körülmények között felszaporítani. Másodszor, a mikróbák rendszerint bonyolult közösségekben élnek, amelyekben az egyes fajok kölcsönhatásban állnak egymással és a környezetükkel. Emiatt a kitenyésztett

organizmusok vizsgálata nem képes valós képet nyújtani az egyes élőlények kölcsönhatásairól, a funkcionális képességeiről vagy a populációban megfigyelhető genomi változatosságáról.

Az új generációs szekvenálási technológiák megjelenése nagyban megkönnyítette a mikróbák vizsgálatát a fent említett korlátozások kiküszöbölésével. A környezeti mintavételezés lehetővé teszi, hogy közvetlenül a mikrobiális közösségek természetes élőhelyéről szerezzük be a genomi információt. Néhány faj egyedenkénti vizsgálata helyett az új technológia képessé tesz minket arra, hogy a közösséget mint egészt vizsgáljuk. Ezek nyomán új tudományág született: a metagenomika – a közvetlenül a környezetből származó genomi szekvenciák (azaz a metagenom) vizsgálata.

Mindazonáltal a környezeti szekvenálásnak is megvannak a maga korlátai. Egy egyedi organizmust vizsgáló genom projekt során majdnem teljes képet kaphatunk a mikroba genomjáról: a rövid genomi szekvenciák összeilleszthetők, annotálhatók, a gének és operonok helye kikövetkeztethető. Ezzel szemben a környezeti mintavételezés nem ilyen egyszerű. Minden egyes szekvenciatöredék különböző fajhoz tartozó élőlényekből is származhat, és sok különböző faj is előfordulhat a mintában. Emiatt a teljes genomok összeillesztése csak speciális környezetek esetén lehetséges, amelyben például egyetlen faj dominálja a mintát, és még ebben az esetben is csak a domináns faj genomja határozható meg. A természetben előforduló környezetek legnagyobb részében rengeteg különböző faj található, így a genomok összeillesztése nem lehetséges. Ezekben az esetekben a rövid szekvenciákból összeillesztett kontigok mérete általában nem haladja meg az 5000 bázispárt. Következésképpen a szekvenciák annotációja csak részben lehetséges, így mindössze vázlatos képet kaphatunk a mikrobiális közösség felépítéséről.

Ebben a fejezetben áttekintjük a metagenomok elemzésének fő megközelítéseit, majd végigkövetjük egy tipikus metagenomikai projekt munkafolyamatát.

20.2. A metagenom elemzése

Ebben az alfejezetben röviden áttekintjük a metagenomok elemzésének fő megközelítéseit.

20.2.1. A közösséget alkotó fajok beazonosítása

Előfordulhat, hogy csak arra vagyunk kíváncsiak, hogy milyen fajokból áll a vizsgált környezet („Kik vannak ott?”). Ebben az esetben a teljes genomi szekvenálás helyett marker gének szekvenálása is elegendő lehet univerzális primerek segítségével, ami egy relatíve gyors és költséghatékony módja a bakteriális diverzitás megbecslésének. Emellett ezt a módszert gyakran használják nagyobb metagenomikai vizsgálatok előzetes lépéseként is a környezet kezdeti felmérésére [4], illetve a bakteriális közösség összetételének időbeli és térbeli változásának monitorozása céljából [5].

A leggyakrabban használt marker gén a 16S rRNS a prokarióták, illetve a 18S rRNS az eukarióták vizsgálatára. A riboszomális RNS (rRNS) a fehérjeszintézisben szerepet játszó riboszómák elengedhetetlenül fontos alkotórésze, amely az evolúció során erősen konzerválódott, ugyanakkor elegendő mértékben változatos is ahhoz, hogy az evolúciós

távolság egy jó markere lehessen. A széleskörű használatát a hatalmas rRNS génszekvencia adatbázisok is elősegítik [6, 7].

A 16S rRNS használatának egyik hátránya, hogy a különféle bakteriális fajokban eltérő számú másolattal rendelkezik, amely erősen befolyásolja a közösség összetételének becslési pontosságát. Ennek a hátránynak a kiküszöbölésére más, egyetlen kópiában meglévő géneket (pl. *RpoB*) is alkalmaztak hasonló célokból. Ezek ugyanis lehetővé teszik a közösségi összetétel pontosabb becslését, szemben a változó számú kópiával rendelkező 16S rRNS használatával [8]. Mindazonáltal a létező bakteriális szekvencia-adatbázisok lényegesen kevesebb – ilyen génekből származó – szekvenciát tartalmaznak.

A marker gének használatának másik hátránya az, hogy a gén szekvenciájának meghatározásához mindenképpen valamilyen módon primereket kell választani. Annak ellenére, hogy ezek a gének evolúciósan konzerválódtak, mindig megvan az esélye, hogy a kiválasztott primerek nem illeszkednek (teljesen) a mintában található egyes fajok DNS-szekvenciájára, ami ezen fajok azonosítását erősen megnehezíti.

Virális közösségek beazonosítása még ennél is nehezebb, ugyanis nem létezik univerzálisan konzerválódott marker gén a vírusok esetén. Ebben az esetben a shotgun-szekvenálás az egyetlen lehetőség.

20.2.2. Funkcionális metagenomika

A közösséget alkotó fajok beazonosítása mellet arra is kíváncsiak lehetünk, hogy a vizsgálandó metagenom funkcionálisan mire képes („Vajon mit csinálhatnak?”). Ebben az esetben nem feltétlenül szükséges tudnunk, hogy melyik gén melyik szervezetből származik; ugyanannak a génnek a terméke ugyanazt (vagy nagyon hasonló) szerepet tölt be attól függetlenül, hogy melyik fajból származik eredetileg. Ezen általános feltevésnek megfelelően a funkcionális metagenomikai megközelítésben a különböző fajok helyett a közösség egészének génkészletére fókuszálunk.

Ebben az esetben a környezetből nagy mennyiségű DNS-t mintavételezünk, majd hagyományos Sanger-módszerrel vagy új generációs szekvenálási technológiával meghatározzuk a szekvenciák bázissorrendjét. Ezután a leolvasott szekvenciákat a lehetőségekhez mérten összeillesztjük, meghatározzuk a lehetséges nyitott leolvasási kereteket (open reading frame, ORF), majd meghatározzuk ezek biológiai funkcióit. Ezt *funkcionális annotálás*nak nevezzük. Az így meghatározott biológiai funkciókat és géneket ezután azonosítjuk meglévő biológiai hálózatokban, például metabolikai útvonalakban. Az alul-, illetve felülreprezentált biológiai funkciók és útvonalak a bakteriális közösség funkcionális képességeiről árulkodnak.

Természetesen ennek a módszernek is megvannak a maga korlátai. A legtöbb esetben a közösség túlságosan bonyolult ahhoz, hogy teljes vagy akár csak majdnem teljes genomösszerekást lehessen végezni, így csak a nyitott leolvasási kereteknek csak részeit lehet azonosítani. Ezek homológ szekvenciáit meg lehet keresni létező adatbázisokban ahhoz, hogy a jósolt kódolt fehérje funkcióját meghatározzuk, de ezt szükségszerűen korlátozza az adatbázisokban rendelkezésre álló információ mennyisége. A nyitott leolvasási keretekben lehet motívumokat vagy más szekvenciamintázatokat is keresni, amelyek a kódolt fehérje funkciójára utalhatnak („Mire képes a jósolt fehérje?”), de ebbe a folyamatba sok

hiba csúszhat a nyitott leolvasási keretek töredékes volta miatt vagy a motívumkereső algoritmusok és a tudásunk hiányosságai miatt [1].

A közösség funkcionális képességeinek meghatározása mellett a véletlen shotgun-szekvenálás akár több információt is tud nyújtani a közösség diverzitásával, taxonómiai összetételével kapcsolatban, mint a marker géneken alapuló módszerek, ugyanis ezt nem korlátozzák a primer szekvenciák használatával összefüggő problémák. Ebből eredően ezen módszerrel képesek vagyunk bakteriofágok és egyéb vírusok azonosítására is a prokarióták és eukarióták mellett. Sőt, új fajok detektálására is, amelyeket a nem túlságosan „univerzális" primerek használatával nem találtunk volna meg.

20.3. Metagenomika lépésről lépésre

Ebben az alfejezetben röviden demonstráljuk egy tipikus véletlen shotgun-szekvenálás alapú metagenom projekt elemzésének tipikus lépéseit.

20.3.1. Mintavételezés

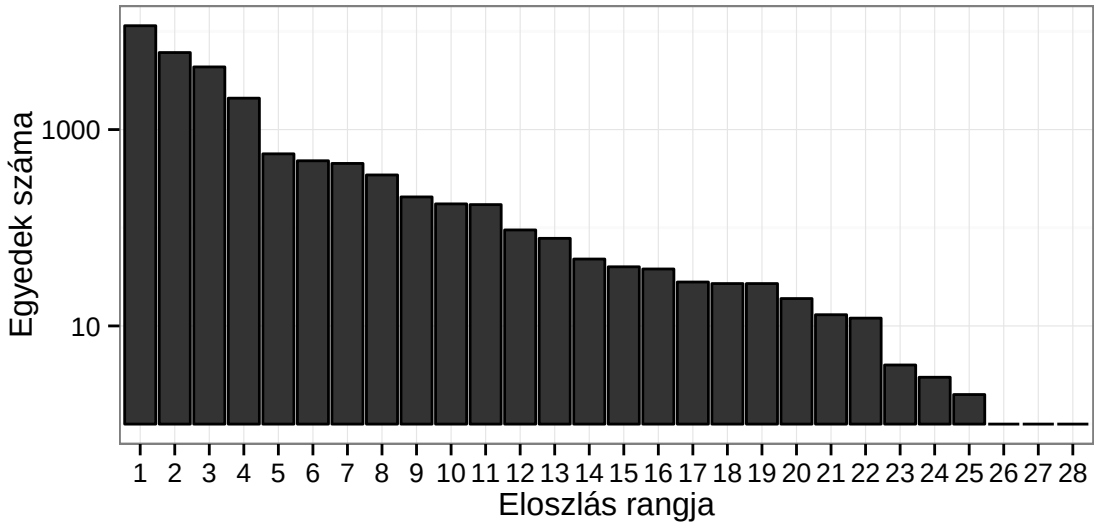
Mintaméret-megfontolások a fajok diverzitásának tükrében

Egy metagenomikai projekt a környezetből való mintavételezéssel kezdődik. A fő kérdés ezzel kapcsolatban az, hogy honnan tudjuk, hogy elegendő mintát gyűjtöttünk, ha nem látjuk azokat az organizmusokat, amelyeket össze szeretnénk gyűjteni?

Emellett vajon hány szekvencia lesz elég? Ez egyrészt a bakteriális közösség struktúráján (biodiverzitásán), másrészt pedig a vizsgálatunk céljától függ. A továbbiakban ezeket a szempontokat fogjuk részletezni.

A közösség struktúrájának komplexitása az azt alkotó különböző fajok számától (*richness*, gazdagság) és azok relatív gyakoriságától (*evenness*, egyenletesség) függ. A legtöbb, természetben előforduló környezetben a fajok relatív gyakorisága nem egyenletes. A legegyszerűbb módszer ennek az egyenetlenségnek az ábrázolására az ún. rang-gyakoriság görbe, amelyben minden egyes taxonómiai egységet egy – a gyakoriságával arányos nagyságú – oszlop reprezentál a leggyakoribb fajtól a legritkábbig (lásd az 20.1. ábrát). Egy kiegyensúlyozott populációban a rang-gyakoriság görbe egyenletes lenne.

Hogyan kapcsolódik mindez a szekvenáláshoz? Ha egy szekvenálási platform képes lenne egyetlen sejt teljes genomjának a pontos szekvenálására, akkor sejtenként egyetlen szekvencia elegendő lenne ahhoz, hogy meglehetősen jó képet kapjunk egy egyetlen fajhoz tartozó egyetlen egyedről. Ugyanakkor a jelenlegi technikai feltételek mindössze 50–700 bázispár hosszúságú leolvasásokat engednek meg, és a rövid fragmenseket a leolvasásokban szereplő átfedő részek alapján kell összerakni. Az egy nukleotidra jutó átlagos leolvasások számát *lefedettség*nek nevezzük. Tételezzük fel, hogy a környezetben található domináns faj genomjának mérete 3 Mbp (pl. a *S. pneumoniae* genomjának mérete kb. 2.2 Mbp), a relatív gyakorisága a populációban legyen 10%. Tegyük fel, hogy a szekvenálás során 700 Mbp-nyi szekvenciát olvastunk le (egy futás során a Roche GS FLX Titanium XL+ rendszerének tipikus teljesítménye). Ebben az esetben a domináns fajt körülbelül 70 Mbp



20.1. ábra. Rang-gyakorisági görbe

szekvencia reprezentálja, ami megközelítőleg 23.3X lefedettséget eredményez. Ugyanakkor egy alacsony gyakoriságú faj esetén (legyen például 0.1% a populációban) az átlagos lefedettség 0.23X lesz.

Ahogyan azt korábban említettük, a vizsgálat céljai szintén befolyásolják, hogy mennyit szükséges szekvenálni. Több mint 20-szoros lefedettség szükséges ahhoz, hogyha a populációban jelenlévő genetikai variációt (pl. egypontos nukleotid polimorfizmusokat) is meg szeretnénk figyelni. Az előző példában említett domináns faj esetén kiszámított lefedettség ehhez elegendő. Körülbelül 6-szoros lefedettség szükséges egy vázlatos genomösszerakáshoz. Mindazonáltal sokkal kevesebb szekvencia is elegendő lehet ahhoz, hogy a közösségben mint egészben felülreprezentált géneket azonosítani lehessen [9].

Metaadatok

A környezeti mintavételezés mellett a metaadatok pontos rögzítése elengedhetetlen: hol, mikor és milyen körülmények között vettük a mintákat. A metaadatok köre környezetenként változó: egy talajból vagy természetes vízből származó minta esetén szükséges rögzíteni biokémiai adatokat (pl. pH-érték, oxigéntartalom stb.), földrajzi adatokat (pl. GPS-koordináták), a minták kezelésére vonatkozó adatokat (dátum és időpont, DNS-kivonatolási eljárás stb.). Emberi mikrobiális mérések esetén fontos rögzíteni az orvosi, kezelésre vonatkozó adatokat (patológia, kórtörténet stb.); a mintakezelésre vonatkozó adatokat (mintavételezési dátum és időpont, a pontos hely és szövet, ahonnan a minta származik stb.) [9, 1, 2].

20.3.2. Szekvenálás

Az új generációs szekvenálási platformok (next generation sequencing, NGS) megjelenése nagyban lecsökkentette a környezeti mintákból származó DNS szekvenálásának költségeit és idejét a korábbi technológiákhoz képest. Mindazonáltal a Sanger-szekvenálás a hosszú leolvasási hossz (>700 bp) és az alacsony szekvenálási hibaarány miatt továbbra is alternatívát jelenhet [10].

Két NGS technológiát használtak eddig jellemzően metagenomikai kutatásokban: a 454/Roche és az Illumina/Solexa platformokat, amelyek közül most röviden bemutatjuk a Roche technológiáját. A GS FLX+ rendszer egy futása során a munkafolyamat három fő lépésből áll: a DNS-könyvtár előkészítése, emulziós PCR és a szekvenálás. A DNS-könyvtár előkészítése során rövid, univerzális adaptereket adnak hozzá a véletlenszerűen feldarabolt DNS fragmensek mindkét végéhez. Ezeket az adaptorokat a további amplifikációs és szekvenálási lépések során használják. A DNS darabkákat ezután mikroszkopikus gyöngyökhöz kapcsolják, és beleöntik egy víz-az-olajban emulziós keverékbe (egy fragmens egy gyöngyön, egy vízcseppben). Az emulziós PCR során a gyöngyön található egyetlen templát DNS molekulát felsokszorozzák, míg végül néhány millió másolata fog a gyöngyhöz kapcsolódni. A gyöngyöket egy speciális plate (PicoTiterPlate$^{\text{TM}}$, PTP) apró üregeibe töltik a piroszekvenálási reakcióhoz szükséges enzimekkel együtt. A szekvenálási lépés során nukleotidokat áramoltatnak keresztül a PTP-en egymást követő turnusokban, és a templát szálakkal komplementer nukleotidok beépülnek DNS polimeráz közreműködésével, ami a beépült nukleotidok számával arányos erősségű fénykibocsátással jár. A kibocsátott fotonokat egy CCD kamera rögzíti és konvertálja bázissorrenddé [11]. Ez a folyamat masszívan párhuzamosan történik, amely ~ 1 millió leolvasást (rövid szekvenciát) eredményez futásonként. Kevesebb mint egy nap alatt összesen 700 Mbp hosszúságú szekvencia keletkezik; a leolvasások hosszának mediánja körülbelül 700 bázispár [12]. Multiplexelés használatával pedig egyetlen futás során akár 132 minta szekvenálására is lehetőség van.

20.3.3. Genomösszerakás

A *genomösszerakás* folyamata során a leolvasásokat összeillesztjük az átfedő részszekvenciák alapján nagyobb, összefüggő DNS szakaszokká, ún. *kontig*okká. A kontig konszenzusos szekvenciáját ezután általában az adott pozícióban leggyakoribb nukleotid alapján állítjuk elő.

Egyetlen organizmus teljes genomjának összerakása is problémás lehet a genomjában szereplő repetitív régiók miatt. Ugyanakkor a metagenom összerakása általában még bonyolultabb. A szekvenciák különböző organizmusokból származnak, és ezen szekvenciák összeillesztése téves eredményre, ún. kimérák keletkezéséhez vezet. Ez a jelenség még gyakrabban fordul elő közeli rokonságban álló organizmusok esetén. A szekvenálási erőfeszítéseinktől függően az alacsony gyakoriságú fajokról esetleg csak néhány szekvenciát sikerül leolvasni, ami elméletileg is lehetetlenné teszi a genomjuk összerakását.

Ezekből következően egy tipikus metagenomikai vizsgálatban az összeillesztett kontigok mérete általában nem haladja meg a néhány ezer bázispárt. Ennek súlyos következményei vannak a további elemzési lépések szempontjából, ugyanis ez a mérettartomány csak

a rövid géneket és fehérjedomaineket fedi le – hosszabb funkcionális egységeket, például operonokat, hosszabb géneket vagy teljes kromoszómákat nem fogunk tudni összeilleszteni [1].

A leolvasott szekvenciák összeillesztése megfogalmazható úgy, mint egy útkeresési algoritmus a szekvenciákat reprezentáló gráfban. Minden egyes leolvasott szekvencia megfeleltethető a gráf egy csomópontjának, és két csomópont között akkor fut él, ha az adott szekvenciák átfednek. Ebben az esetben a genom összerakása megfelel egy Hamilton-kör keresési problémának, amelyben minden csomópontot pontosan egyszer látogatunk meg. Ez azonban metagenomikai vizsgálatok során nem alkalmazható a feladat NP-teljes számítási komplexitása miatt, a Hamilton-kör megtalálásához szükséges idő ugyanis exponenciális mértékben nő a leolvasások számának növekedésével. Ezt a megoldást általában csak kisebb genomok összerakására és hosszabb (tipikusan Sanger) szekvenciák leolvasása esetén szokták alkalmazni.

Egy másik megközelítésben a gráf csomópontjai k-méretű szavakat jelentenek, és a leolvasott szekvenciák azoknak az éleknek feleltethetők meg, amelyek a megfelelő csomópontokat (rész-szavakat) összekötik. Ennél fogva a csomópontok száma független a leolvasott szekvenciák számától. A genom összerakása ekkor egy Euler-kör keresési problémaként fogalmazható meg, amelyben minden élet pontosan egyszer látogatunk meg. Erre létezik lineáris idejű algoritmus, ami ezáltal lehetővé teszi a genom összerakását metagenomikai alkalmazások esetén is (természetesen a korábban megfogalmazott korlátozásokkal). Több, szabadon hozzáférhető eszköz is létezik, amely ezt az algoritmust valósítja meg, mint például az EULER [13], a Velvet [14] vagy a MetaVelvet [15].

20.3.4. Besorolás

A megagenom összerakása során egybefüggő kontigokat és egyedüli (singleton) leolvasásokat kapunk eredményül. Azt a folyamatot, amikor ezeket összerendeljük azokkal az organizmusokkal (vagy magasabb taxonómiai egységekkel), amelyekből származnak, *besorolás*nak (*binning*) nevezzük. Ebben az alfejezetben két besorolási eljárást mutatunk be: a szekvencia alapú és a tartalom alapú besorolást.

Szekvencia alapú besorolás

Az egyik leggyakrabban használt besorolás eljárás azon alapul, hogy egy adott szekvenciához hasonló szekvenciákat keresünk egy annotált referencia-adatbázisban lokális szekvenciaillesztéssel, például a Basic Local Alignment Search Tool (BLAST) [16] felhasználásával. Ez a módszer akkor vezet jó eredményre, ha a legtöbb szekvenciához találunk szignifikánsan hasonló referenciaszekvenciákat, amelyek ismert organizmusokból származnak. Ugyanakkor a nem teljes vagy pontatlan adatbázisok használata erősen befolyásolja a kapott eredmények megbízhatóságát.

Tartalom alapú besorolás

Egy másik besorolási módszer a szekvenciák nukleotidkompozícióján alapul. Jól ismert tény például, hogy a DNS GC tartalma erősen variábilis és jó ismertetőjegye a különbö-

ző fajoknak. Szofisztikáltabb módszerek oligonukleotidok (k méretű szavak) gyakoriságán vagy kodonhasználati jellemzők vizsgálatán alapulnak, amelyek szintén különböznek az eltérő fajok genomjai között [17]. Oligonukleotidok használata esetén a szavak mérete különböző lehet, 1-től kezdve (GC tartalom) 4-en keresztül (tetranukleotid, pl. TETRA [18]) 8-ig (pl. RDP osztályozó riboszomális RNS-re [7]).

Azonban az olyan rövid szekvenciák besorolása, amelyek nem illeszthetők nagyobb kontigokba, problémás lehet, ugyanis ezek kevesebb szót tartalmaznak, ami miatt a besorolás bizonytalanná válik. Ezekben az esetekben a szekvencia alapú besorolási módszer használható.

20.3.5. Génfelismerés és funkcionális annotáció

A genom alapvető funkcionális egységei a gének. A minta DNS-ből származó génszekvenciák azonosítását *génfelismerés*nek (*gene calling*) nevezzük. A génfelismerés metagenomikai minták esetén különösen nagy kihívást jelent a környezeti DNS töredékes természete és hiányos összerakása miatt.

A génfelismerés alapvető módszere szerint az összerakott kontigokhoz hasonló géneket vagy fehérjéket keresünk a BLAST segítségével létező adatbázisokban. A szekvenálási hibák vagy az összeillesztett kontigok rövidsége azonban megnehezíti és néhány esetben lehetetlenné teszik a homológ szekvenciák azonosítását. Emellett a BLAST nem használható új gének megtalálására sem, hiszen ezeknek nincs ismert homológjuk a létező adatbázisokban. Így, a homológiakeresés során az új géneket teljesen figyelmen kívül hagyjuk [1].

Egy másik megközelítésben „ab initio" génfelismerést is használhatunk akkor, amikor a homológiakeresés nem vezet kellő eredményre. A létező eszközök statisztikai mintázatfelismerést valósítanak meg, azaz a DNS szekvenciák azon belső jellemzőit ismerik fel, amelyek a kódoló és nem kódoló szakaszokat megkülönböztetik. Teljes genomok esetén az „ab initio" génfelismerés általában könnyebb, ugyanis a modellek az adott genom alapján betaníthatók és a működésük finomhangolható. Metagenomikai minták esetén azonban csak a domináns egyedek vizsgálhatók ilyen módon. Ezek szekvenciáit ugyanis elválaszthatjuk a minta többi részétől (besorolási eljárással). Az alacsony gyakoriságú minták esetén azonban csak általános modelleket használhatunk. Például a MetaGene [19] szoftver két általános modellt használ: egyet archaea-ra és egyet baktériumok esetén.

A génfelismerés végrehajtása után általában arra keressük a választ, hogy a mikrobiális közösség vajon milyen potenciális funkciót tölt be („Mire képesek közösségként?"). A származtatott génlistákat össze lehet hasonlítani például metabolikus útvonal-adatbázisokkal (mint amilyen a Kyoto Encyclopedia of Genes and Genomes (KEGG) [20]), amely a géneket hozzárendeli azokhoz a biológiai funkciókhoz, amelyekben azok részt vesznek. Az alul-, illetve felülreprezentált útvonalak és biológiai folyamatok a közösség funkcionális képességeiről árulkodnak.

Emellett adott gének jelenlétének vagy hiányának megállapítása is felfedhet fontos funkcionális jellemzőket. Például antibiotikum-rezisztencia gének jelenléte alapján megjósolható az antibiotikus kezelés hatásossága, illetve esetleges káros következményei [2].

Irodalomjegyzék

[1] John C. Wooley, Adam Godzik, and Iddo Friedberg, A primer on metagenomics. *PLoS Computational Biology*, 6(2), February 2010. PMID: 20195499 PMCID: PMC2829047.

[2] George M. Weinstock, Genomic approaches to studying the human microbiota. *Nature*, 489(7415):250–256, September 2012. PMID: 22972298.

[3] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science (New York, N.Y.)*, 269(5223):496–512, July 1995. PMID: 7542800.

[4] Peter J. Turnbaugh, Micah Hamady, Tanya Yatsunenko, Brandi L. Cantarel, Alexis Duncan, Ruth E. Ley, Mitchell L. Sogin, William J. Jones, Bruce A. Roe, Jason P. Affourtit, Michael Egholm, Bernard Henrissat, Andrew C. Heath, Rob Knight, and Jeffrey I. Gordon, A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480–484, January 2009.

[5] J. Gregory Caporaso, Christian L. Lauber, Elizabeth K. Costello, Donna Berg-Lyons, Antonio Gonzalez, Jesse Stombaugh, Dan Knights, Pawel Gajer, Jacques Ravel, Noah Fierer, Jeffrey I. Gordon, and Rob Knight, Moving pictures of the human microbiome. *Genome Biology*, 12(5):R50, 2011. PMID: 21624126 PMCID: PMC3271711.

[6] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glockner, The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1):D590–D596, November 2012.

[7] J. R. Cole, Q. Wang, E. Cardenas, J.Fish, B. Chai, R. J. Farris, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, T. Marsh, G. M. Garrity, and J. M. Tiedje, The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Research*, 37(suppl 1):D141–D145, January 2009.

[8] Rebecca J. Case, Yan Boucher, Ingela Dahllöf, Carola Holmström, W. Ford Doolittle, and Staffan Kjelleberg, Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Applied and environmental microbiology*, 73(1):278–288, January 2007. PMID: 17071787.

[9] Victor Kunin, Alex Copeland, Alla Lapidus, Konstantinos Mavromatis, and Philip Hugenholtz, A bioinformatician's guide to metagenomics. *Microbiology and molecular biology reviews: MMBR*, 72(4):557–578, December 2008. PMID: 19052320.

[10] Torsten Thomas, Jack Gilbert, and Folker Meyer, Metagenomics – a guide from sampling to data analysis. *Microbial informatics and experimentation*, 2(1):3, 2012.

[11] Michal Janitz, editor, *Next-Generation Genome Sequencing: Towards Personalized Medicine*. Wiley-Blackwell, 1. ed., October 2008.

[12] Products – GS FLX+ System: 454 Life Sciences, a Roche Company. `http://454.com/products/gs-flx-system/`

[13] Mark J. Chaisson and Pavel A. Pevzner, Short read fragment assembly of bacterial genomes. *Genome research*, 18(2):324–330, February 2008. PMID: 18083777.

[14] Daniel R. Zerbino and Ewan Birney, Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, 18(5):821–829, May 2008. PMID: 18349386.

[15] Toshiaki Namiki, Tsuyoshi Hachiya, Hideaki Tanaka, and Yasubumi Sakakibara, MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic acids research*, 40(20):e155, November 2012. PMID:22821567.

[16] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, October 1990. PMID: 2231712.

[17] S. Karlin, J. Mrázek, and A. M. Campbell, Compositional biases of bacterial genomes and evolutionary implications. *Journal of bacteriology*, 179(12):3899–3913, June 1997. PMID 9190805.

[18] Hanno Teeling, Jost Waldmann, Thierry Lombardot, Margarete Bauer, and Frank Oliver Glöckner, TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences *BMC bioinformatics*, 5:163, October 2004. PMID: 15507136.

[19] Hideki Noguchi, Jungho Park, and Toshihisa Takagi, MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic acids research*, 34(19):5623–5630, 2006. PMID: 17028096.

[20] Minoru Kanehisa, Michihiro Araki, Susumu Goto, Masahiro Hattori, Mika Hirakawa, Masumi Itoh, Toshiaki Katayama, Shuichi Kawashima, Shujiro Okuda, Toshiaki Tokimatsu, and Yoshihiro Yamanishi, KEGG for linking genomes to life and the environment. *Nucleic acids research*, 36(Database issue):D480–484, January 2008. PMID: 18077471.