1. Orvosi bioinformatika

1.1. Az orvosbiológia helyzete az adatgazdag tudománytörténeti korszakban

Az utóbbi évtized egy technológiai és tudománytörténeti korszakhatárnak is tekinthető, amikor a XX. század második felére jellemző számítás-intenzív, szimulációs korszakot egy adat-intenzív, adatelemző korszak váltotta fel. Az adatok mennyiségének exponenciális gyarapodása több tudományágban is megfigyelhető jelenség, ami a részecskefizikában és asztrofizikában már korábban elkezdődött, de a molekuláris biológia és az egészségügy területein az utóbbi évtizedben gyorsult fel. Ezt alapvetően új biotechnológia eljárások, nevezetesen speciális méréstechnikák robbanásszerű fejlődése tette lehetővé. A grandiózus tudományos projekteknek mint például a Human Genome Project-nek és a kapcsolódó technológiai áttöréseknek köszönhetően a biológia az 1990-as években egy rendkívül gyors, régóta várt transzformáción ment keresztül:

1. A biológia „adat-gazdag" tudománnyá lett, tekintve az adatok mennyiségét (szekvencia adatok), az adatok sokféleségét (genomikai, fehérjék, térszerkezetek, molekuláris interakciók, mutációk, fenotipikus, klinikai adatok) és az adatok dimenzióját (átfogó génaktivitás és fehérje interakció vizsgálatok gén- vagy fehérje-chipekkel).
2. Gyakorlati közelségbe kerültek a biológia „nagy egyesített elméletei". Ilyenfajta egyesítést kínál az emberi genom ismerete, ami egy olyan központi szervezési pontot biztosít, ahova a makromolekuláris részletektől a klinikai szintig felfűzhetők az ismeretek.
3. A biológia a közvélemény fókuszába került, például a betegségek korai diagnosztizálásának, személyre szabott kezelésének az ígéretével, genetikailag módosított élelmiszerek kérdésével, a klónozással vagy a genetikai információk etikai, üzleti kérdésével.

Az előző jegyeknek is köszönhetően aposztrofálták a biológiát mint, a „XXI. század tudományát" (v.ö. fizika a XX. század tudománya). Ennek a megalapozottságát az is jelzi, hogy az 1990 évektől a biológia újfajta kutatási módszereket, tudományfilozófiát, publikálási metódust, számítási potenciált, tudományos-ipari együttműködést és új iparágakat indukált. További indokok a következők:

4. Felhasznált biológiai számítási kapacitás. Jelenleg a molekuláris biológia és biokémia átveszi a vezetést a felhasznált számítási kapacitást tekintve olyan klasszikus fizikai számításoktól, mint az áramlástan, időjárás és klíma változás előrejelzés vagy nukleáris szimulációk. Egy ezt jelző példa volt a Celera által használt szuperszámítógép az emberi genom összeállításához. Jelenleg azonban már újabb szuperszámítógép generációk üzemelnek a Sandia National Lab-ban, amely az U.S. egyik elsőszámú védekelmi kutatásokra specializált állami laboratóriuma. Paul Robinson-t, a Sandia elnökét idézve: " We in the nuclear weapons community felt for many years nothing could be more complex than nuclear physics, but I'm now convinced nothing beats the complexity of biological science."
5. Biológiai adatok mennyisége, sokfélesége és dimenzója. A technológia újításoknak köszönhetően a génszekvenciák meghatározása exponenciálisan gyorsuló ütemben történhetett meg, ami azt jelentette, hogy a generált információ mennyiségének duplázódása 15-18 hónaponta történik meg.

6. Tudásbázisok sokasága és sokfélesége. Az előbb említett „adatok” általában tudáselemek is egyben. Például egy nukleinsav szekvencia részlethez általában az „annotálása” megadhatja az általa tartalmazott gént, evolúció szempontjából más faj releváns génjeit, a gén átírását szabályozó rész-szekvenciákat, a kódolt fehérje funkcióit, annak térbeli atomi szerkezetét, a releváns metabolikus hálózatokat, a génhez kapcsolódó lehetséges mutációkat és azokhoz tartozó betegségeket, azok szimptomáit, stb. A különböző vonatkozású adatot/tudást különböző tudásbázisok tartalmazzák, ezek között megtalálhatók klasszikus formális logikai tudásbázisok (EcoCyc), ontológiák (Gene Ontology), tezauruszok (UMLS), szemi-struktúrált és természetes nyelvi szövegeket tartalmazó adatbázisok, szakmai publikácók (PubMed), illetve pszeudo-adatbázisok is, amelyek procedurális alapú szolgáltatásokként képesek entitás relációkat megállapítani, mint például szekvencia hasonlóság, fehérje struktúra hasonlóság, fehérje családba tartozás.
7. Adatok elosztottsága. A biológiai adatok és információk egyik jellegzetessége, hogy autonóm közösségek által fenntartott, interneten keresztül elérhető adatbázisok sokaságában tárolódnak.
8. Redukcionista versus holisztikus biológia. A „post-genom” biológia egyik nagy ígérete a genetikai ismereteken alapuló hatékonyabb gyógyszer „tervezés” („pharmacogenomics”), illetve a személyre szabott orvosságok és kezelés alkalmazása. Ez például azt jelentheti, hogy egy adott adott klinikai szituációban lévő, adott genetikai típusba tartozó beteg, egy adott szövetében, adott mutációtípussal kialakuló rákos sejtjeit célzottan megpróbáljuk megsemmisíteni egy alkalmas hatóanyaggal, a sejt genetikai, fehérje és metabolit hálózatának megbénításával. Egy ilyen hatóanyag megtervezéséhez, pontosabban potenciális hatóanyagok generálásához biológiai, orvosi és klinikai szintek és aspektusok sokaságát kell együttesen figyelembe venni.
9. Publikálás és kutatás. Az adatok, tudás és metodológiák mennyisége és sokfélesége a „post-genom” biológiában is fokozattan megkövetelte az együttműködödést, az eredmények összekapcsolhatóságát (v.ö. a biológiai tudásban meghatározó esetlegességeket szemben a fizika axiomatikus/redukcionista származtathatóságával). Az eredmények könnyű, lehetőleg automatikus összekapcsolása, felhasználása így rövid idő alatt átalakította a tudományos publikálást, megkövetelve az adatok, eredmények és a publikáció olyan „szemantikus” közzétételét, ami által az összekapcsolás automatikusan, szoftvereszközökkel is elvégezhető.

Megállapítható, hogy a „post-genom” biológia által előzőekben vázolt tulajdonságai, úgymint a magas dimenziójú, nagy mennyiségű, elosztott adat és tudás sok szempontú elemzése nem egy ideiglenes állapot, hanem egy általános és hosszú távú tudománytörténeti trendnek a jelzése.

A molekuláris biológiai méréstechnika fejlődését jól jellemzi, hogy az általa generált adatok mennyiségére szintén megfogalmazható a számítástechnikából jól ismert Moore törvény, amely szerint a molekuláris biológiai adatok mennyisége 18 havonta megduplázódik[1, 2].

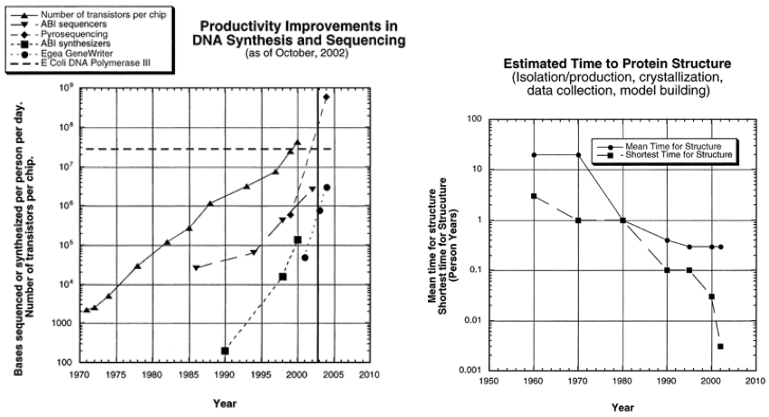


Figure 1 Bal: DNS szintézis és szekvenálás produktivitásának növekedése Moore törvényéhez viszonyítva (felfele háromszögek). Az egyes pontok a feldolgozható DNS mennyiségét jelzik, amit egy személy több mérőeszköz felhasználásával nyolc órás napi munkában elérhet. Jobb: Fehérjetérszerkezetek meghatározásának idejének változása[2].

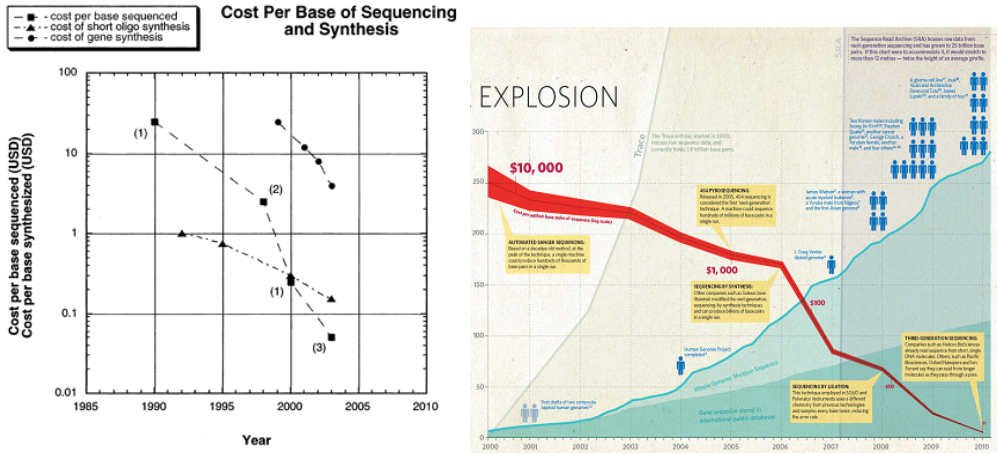


Figure 2 Bal: DNS szekvenálás és szintézi becsült költségeinek változása. Jobb: Szekvencia adatok mennyiségének változása [1].

Hasonló növekedési jelleget mutat az orvosbiológiai szakcikkek változása.

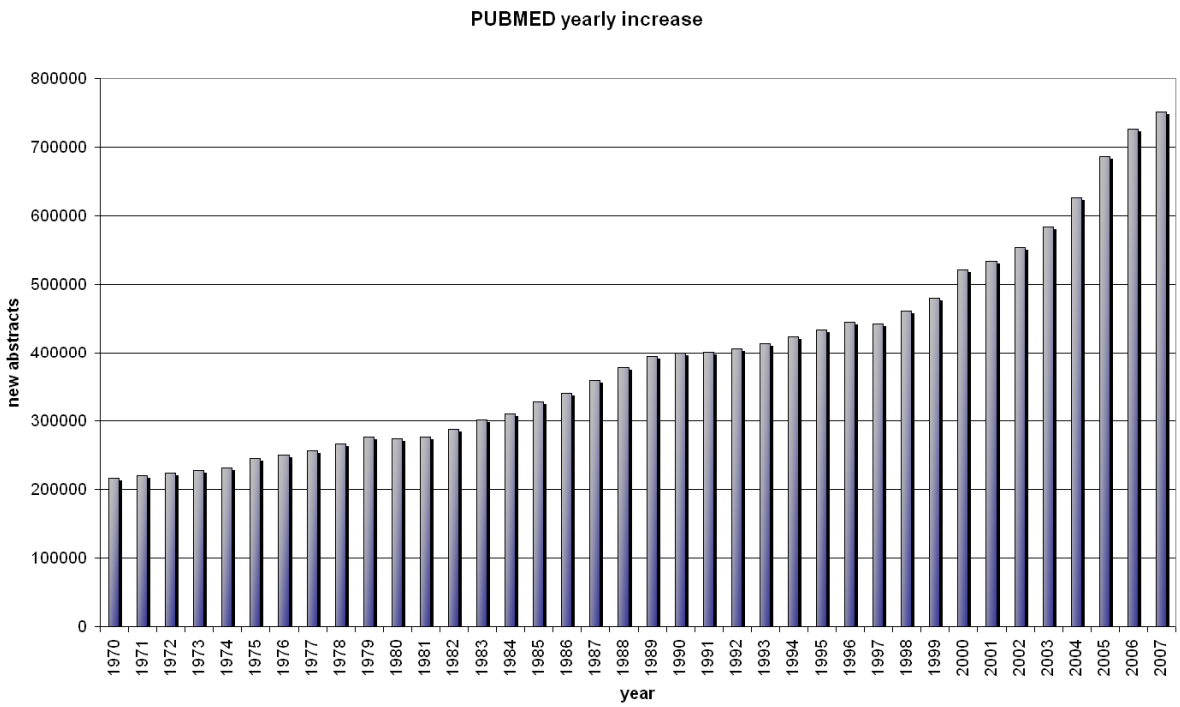


Figure 3 Orvosbiológiai szakcikkek száma a PubMed adatbázisban.

Ez a korábban elképzelhetetlen adatmennyiség a biotechnológiai és orvosbiológiai tudományterületeket alapvetően meg is változtatta: a korábbi tudásgazdag és hipotézisvezérelt jelleg helyett egy merőben új, adatgazdag-hipotézismentes megközelítés vált lehetségessé, amelyben az adatok elemzése, és az elemzési eredmények és ismeretek fúziója kritikus fontosságú.

Azonban nem csak a biotechnológia területeit változtatta meg az új megfigyelések gyorsulva gyarapodó mennyisége, hanem ez kölcsönösen is igaz: a számítástudománynak, a statisztikának, és az informatikának is a molekuláris biotechnológiai/orvosbiológiai területek a legfőbb fogyasztója és így húzóágazata, trend teremtője lett. Ez különösen igaz az adatelemzési eredményeket és háttértudást integráló módszerek kutatására és fejlesztésére, mivel ez tipikusan az internet segítségével megy végbe, elosztott adat- és tudásbázisok, hálózati szolgáltatások ezreinek a segítségével.

Az orvosbiológiai kutatások és az informatika kölcsönös fontosságát jelzi a bioinformatikának nevezett tudományterület kialakulása is, amely informatikai eszközöket és módszereket alkalmaz magasabb szintű jelenségek molekuláris biológiai szintű hátterének modellezésére, elemzésére, és megismerésére. Az egészségügy területén különösen nagy elvárások fogalmazódtak meg a poszt-genomikai korszakban, amelyeknek valóra váltásában új matematikai és informatikai megoldásokra is szükség van, illetve a napi klinikai gyakorlatba beépülő mérnöki konstrukciókra is szükség van, hogy a

biotechnológiai-bioinformatikai tudománytörténeti fordulópont egy civilizációs-kultúrtörténeti fordulópont is legyen, hasonlóan a kommunikációs technológiák utóbbi évszázadban megtörtént kibontakozásához.

A bioinformatika gyors megjelenése és robbanásszerű fejlődése miatt a terület átfogó tárgyalása helyett a magyar oktatásban eddig nem lefedett, ám központi fontosságú fogalmakat tekintjük át.

Az ezredforduló során született áttekintés a (1) szekvenciaelemzések (illesztés, motívumkeresés, filogenetika), és (2) szerkezeti modellezés két fő témakörre bontotta, amit adatbáziselmélet és biostatisztikai területek is kiegészítettek. Az orvosi-gyógyszerészeti vonal felerősödése miatt (3) a statisztika és (4) az orvosi döntéstámogatás, a genetikai variabilitás központba kerülése miatt (5) a genetika, a rendszerszemlélet miatt a (6) rendszerbiológia és (7) hálózatelmélet, a tudás és publikációk felhalmozódása miatt pedig (8) a tudásbázisok, ontológiák, és (9) szövegbányászat vált hasonlóan fontos területté. Az induktív következtetéssel, tudáskezeléssel, és rendszerszemléleten alapuló publikációk száma, különösen az orvosi alkalmazói cikkeket is figyelembe véve, mára már meg is haladja a bioinformatika klasszikus ágaival foglalkozó cikkek számát. A magyarországi bioinformatika képzés követte is ezt a trendet, és szekvenciaelemzésre, strukturális modellezésre, és genomikai méréstechnikákra készültek egyetemi jegyzetek. Azonban az orvosbiológia kutatásokban és alkalmazásokban, különös tekintettel a genetikai, genomikai és általában az -omikai kutatásokban mára már kritikus szerepet játszanak a bioinformatika következő területei

1. kísérlettervezés és adatgyűjtés,
2. kis-mintás statisztikai adatelemzések,
3. oksági modellek és valószínűségi gráfos modellek,
4. relevancia alapú biomarker elemzés,
5. tudásfúzió az elemzések értelmezésére,
6. orvosi döntéstámogatás.

A kísérlettervezés a pre-omikai, klasszikus keretben a hipotézis igazolásához szükségesnek vélt változók és mintaméret meghatározását jelentette. A post-genomikai korszak hipotézismentes volta és –omikai méréstechnikái látszólag a változók meghatározását szükségtelenné teszi . Azonban egy adott omikai szinten belül is különböző fedésű és pontosságú mérések érhetők el más és más költséggel. Egy adott költségkeretnél így az omikai méréstechnika megválasztása a minőségi paraméterek és az elérhető mintaszám együttes megfontolását igényli. Ez több omikai méréstechnika

figyelembevételénél többszörös kombinációt kínál. Önálló megfontolást igényel egy speciális –omikai szint a fenome, azaz a fenotípusok leírásának szintje, aminek standardizálása az egyik jelenlegi legnagyobb kihívás. Gyakorlati szempontból pedig nagyon fontos kérdés a biológiai minták tárolása biobankokban, az adatok tárolása betegregiszterekben, ahol megfelelő minőségbiztosítás és adatintegrálás érhető el.

Az induktív következtetés, benne a statisztikai adatelemzéssel szintén egy korszakba ért az omikai méréstechnikák biztosította rendkívül nagy változószám („dimenzió") és az Moore törvény szerint a 2000-s évekig gyorsuló, majd onnantól fogva párhuzamosítás révén növekvő számítási kapacitások miatt. Azonban az új méréstechnikai eljárások költségeinek csökkenése ellenére a közeljövőben, azaz 2-3, de várhatóan akár 5-10 éves perióduson belül is, a mintaszám viszonylagosan alacsony fog maradni a változószámhoz és az illeszteni kívánt modellek komplexitásához képest. Ennek oka egyrészt az entitások rendkívül nagy száma, környezeti feltételek nagy száma, de további oka lehet a betegségek ritka vagy gyorsan változó volta, illetve oka lehet a használni kívánt modellek egyre komplexebb volta is. Ennek egy megjelenési formája a többszörös hipotézistesztelés problémája is, a megismételhetetlenség. A kismintás-számításintenzív induktív módszerek egyik egyre népszerűbb formája a bayes statisztika.

Az omikai megközelítés nagy változószámai az egyváltozós asszociációs hipotézisek esetén komoly kihívást jelentenek statisztikai, azaz mintaméret vonatkozásában. A hipotézistesztelési keretben ez a hamis felfedezés (elsőfajú hiba, Type I.) és az elmulasztott felfedezés (másodfajú hiba, Type II.) közötti helyes arány megtalálását is jelenti. A többváltozós megközelítés a hipotézisek számát nagyságrendekkel tovább növeli, amely megközelítés azonban a gyakori betegségek multifaktoriális jellege miatt szükségszerű. Az asszociációs, prediktív alapú hipotézisekkel szemben az oksági, generatív, autonóm mechanizmus alapú hipotézisek mindig is elsőbbséget élveztek, bár elérhetőségük bizonyos értelemben statisztikán túlinak, tárgyterület specifikusnak volt elkönyvelve.Az utóbbi három évtized matematikai kutatásai az oksági relációk induktív következtetésénél új eredményeket, egy új tudományág megszületését hozta. Az oksági hálózatok kutatása a hipotézisek számának csökkentését, zavaró tényezők hatékonyabb kezelését, a hatáserősség jobb megbecslését is biztosítja, akár a kísérlettervezés adaptív voltával (ld. aktív tanulás).

Az asszociáció vagy prediktív alapú statisztikai megközelítés a biomarkerek tekintetében is a fenti problémákhoz vezet, azaz mind egy, mind többváltozós esetben a hipotéziszámhoz a szükséges mintaszámok nem elérhetőek. Az oksági kutatások egyik mellékhatása, hogy megszülettek olyan matematikai fogalmak is, mint például a valószínűségi gráfos modellek és az erős relevancia, amelyek oksági értelmezés nélkül is lehetővé teszik a kisebb hipotézisszám elérését, a robusztusabb hipotézisek használatát. A valószínűségi gráfos modellek így mintegy köztes lépcsőfokot jelentenek az asszociációs hálózati és az oksági hálózatok között, amely a biomarker kutatásban jelent új lehetőségeket.

Az adat- és tudásfúzió kérdése a helyes következtetés, és azon belül az indukció évezredes problémaköreihez sorolható. A Jacob Bernoulli-tól, Thomas Bayes-től eredeztett szubjektív értelmezésű valószínűségi megközelítés az a priori hiedelmek és a megfigyelések normatív kombinálására adott megoldást, azonban az elméleti egyszerűség ellenére ez nagy számításigényú szimulációkat igényelnek. Ennek megfelelően a bayesi tudásintegrációs megközelítések csupán az utóbbi két évtizedben lettek egyre népszerűbbek. Az adat- és tudásfúziós kutatásokban egy követlező korszakot egy új feladat, a prioritizálás megjelenése hozta, ami a korábbi hálózati és klaszterezési fúziós megközelítéseknek egy jelentős egyszerűsítése. A prioritizálás, például génprioritizálás egy sorrendi tanulási feladat, amit nagyban motivált a szabadszöveges keresőgépek működésének egyszerűsége (pl. PageRank). Leegyszerűsítve a feladat adott ``pozitív'' mintákhoz megkeresni a leginkább kapcsolódókat a felcímkézetlenek közül, koherensen felhasználva az összes lehetséges adatforrást.

A személyreszabott medicina egyik ígérete a költséghatékony egészségügy, amikor személyreszabott gyógyszerek betegek egy alcsoportján eredményesebben alkalmazható. Ekkor megfelelő diagnosztika segítségével egy többlépéses, szekvenciális, azaz időben elhúzódó kezelés optimális megtervezése szükséges, amelyet döntéstámogató rendszerek segíthetnek. Adott információs állapotban megbecsülhetik egy adott beszerezhető információ hasznosságát, vagy akár irreleváns voltát, és akár, a természettel sakkozás analógiájára, több lépést előre szimulálva megbecsülhetik egy terápia felhasználásának közvetlen, de akár távoli következményeinek hasznosságát is.

Ezek a területek egy egységes adatközpontú munkafolyamat egyes fázisaiként is szemlélhetőek ld. 1. Ábra.

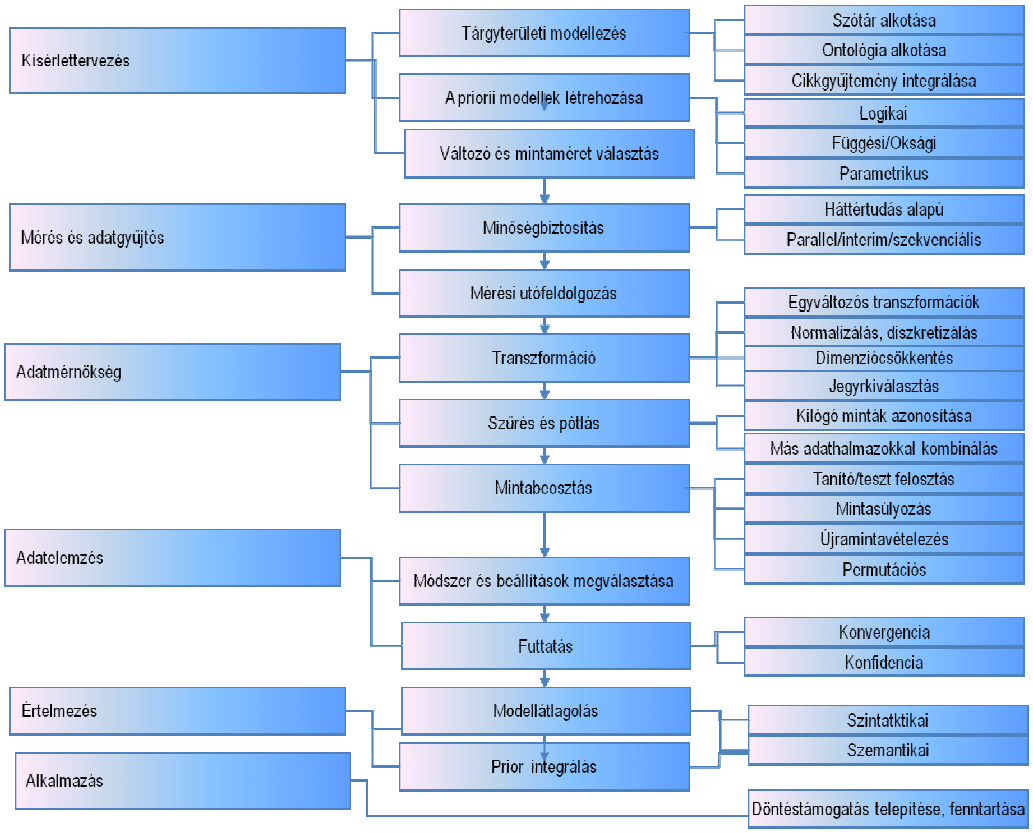


Figure 4 Genomikai kísérlettervezés és kiértékelés általános sémája és fázisai.

A 2.1 Ábra fázisai a következők alfeladatokhoz kapcsolódnak:

1. Általános a priori információ források és adatok megszerzése és konverziója, amely általános, sok esetben akadémiai felhasználásnál ingyenesen vagy kedvezményesen elérhető tudásbázisokra támaszkodik, mint a PubMed, Ingenuity, Ariadne, KEGG, Gene Ontology, HAPMAP, dbSNP, PreMod, TRANSFAC, miRDB, vagy OMIM.

2. Saját probléma-specifikus a priori ismeretek létrehozása, amely magába foglalja szakcikkgyűjtemények meghatározását, általános szótárakból, ontológiákból releváns részek kivonatolását, „szövegbányászatot”, tipikusan szakirodalomelemzést, benne entitások előfordulási statisztikáinak elemzésével, relációk kivonatolását, tudásbázis építését, korábbi kapcsolódó kísérletek adatainak beszerzését, transzformálását, környezeti és klinikai változók oksági vagy függési modelljének létrehozását, kapcsolódó útvonal diagrammok exportálását és klinikai szinthez kapcsolását.

3. Modell absztrakciós szintjének és határainak meghatározása.

4. Kísérlettervezés, benne a potenciális cél relációk vagy cél entitások meghatározásával, kísérlet típusának és a megfigyelni kívánt változóknak a meghatározása, adatgyűjtési protokoll kidolgozása, különös tekintettel az adatminőségbiztosítási protokoll kidolgozására, a potenciális szisztematikus hiányzás kezelésére és az együttműködés okozta torzítás kezelésére.

5. Mérés és adatgyűjtés, amelyben fontos szerephez jutnak a biobankok a biológiai minták megbízható tárolása miatt, a laboratóriumi információs rendszerek, amelyek a mérés folyamatának adatait is rögzítik, az adat megbízhatóságának segítésére, a betegregiszterek és adatelemzési tudásbázisok, amelyek a nyers adatokat, az elemzésre előkészített adatokat, és akár későbbi meta-elemzésekbe kiadott megosztott adatokat és részleges statisztikákat is tartalmazzák

6. Adatmérnökség, amely mind változók és mind minták szinjén transzformációkat vagy például kiválasztást jelent. Az elsőre az SNP-haplotípus transzformáció vagy akár egy adott útvonal menti genetikai variánsokból létrehozott aggregált, episztatikus interakciókat leíró változó a példa. A minta szintű beosztásra pedig a tesztadathalmaz létrehozása egy gyakori példa.

7. Az adatelemzés során tipikusan több statisztikai keretrendszert és több modellosztályt is meg kell vizsgálni. Ilyen keretrendszerek a a hipotézistesztéshez kötődő (klasszikus) „frekventista", hatáserősségvizsgálatban használt „bootstrap", vagy a bayes statisztikai keretrendszer. Modellosztályok esetén az egyváltozós és többváltozós vizsgálat megszokott, ahol a többváltozós vizsgálaton belül az úgynevezett feltételes modellek népszerűbbek, ahol a prediktor változók közti függéseket nem modellezük, azonban a számítási kapacitás növekedésével ez már egyre inkább lehetséges és sok esetben segítséget jelent. A legnagyobb kihívást ekkor az jelenti, hogy a különböző keretrendszerek és modellek együttes alkalmazása konzisztens és megbízható legyen, amit egy átfogó induktív döntéselméleti keret tud garantálni.

8. Az adatelemzési eredmények értelmezése kapcsán nagy kihívást jelent az orvosbiológiai terület azon sajátossága, hogy autonóm, sokszor csak gyengén kapcsolt szintek sokasága létezik, saját –omikai ontológiával, adat- és tudásbázisokkal, mint például a genetikai (DNS), genomikai (mRNS), proteomikai, lipidomikai, metabolomikiai, immunomikai, sintek vagy a „drugome", „diseasome", „phenome" szintjei. Ezeknek az integrált felhasználása egy adott értelmezés során egy formálódó hipotézis megvizsgálására az egyik leggyorsabban fejlődő kutatási terület.

9. A transzlációs kutatások egyik végcélja a klinikai gyakorlatban hasznosuló új diagnosztikai és terápiás eljárások, eszközök megalkotása („theranostics"). A személyreszabott medicina várható trendjében az új diagnosztikai és terápiás eljárások komplexitása növekedni fog, és várhatóan a változásuk sebessége is az orvosi képzési folyamatokhoz képest igen gyors lesz. Emiatt az informatikai döntéstámogató eszközök szerepe növekedni fog. Ezt a folyamatot az egyre kiterjedtebb elektronikus betegadatregiszterek is támogatják, ami akár az otthoni időskori monitorozó rendszerekből is kaphat adatokat.

1.2. Kísérlettervezés és adatgyűjtés

A kísérlettervezés a pre-omikai, klasszikus keretben a hipotézis igazolásához szükségesnek vélt változók és mintaméret meghatározását jelenti. A post-genomikai korszak hipotézismentes volta és –omikai méréstechnikái látszólag a változók meghatározását szükségtelenné teszi . Azonban egy adott omikai szinten belül is különböző fedésű és pontosságú mérések érhetők el más és más költséggel. Egy adott költségkeretnél így az omikai méréstechnika megválasztása a minőségi paraméterek és az elérhető mintaszám együttes megfontolását igényli. A 3.1 Ábra a genetikai adatok egy hierachiáját mutatja, amely teljes genom szekvenálástól, részleges szekvenálási információkon és teljes genom asszociációs adatokon át a részleges genomszűrési és jelölt gén asszociációs adatokig tart. A 2011-ben tipikus viszonylagos költség a $x10^6$ HUF, $x10^5$ HUF, 10^4 HUF, a változók száma pedig $x10^9$, $x10^6$, $x10^3$.

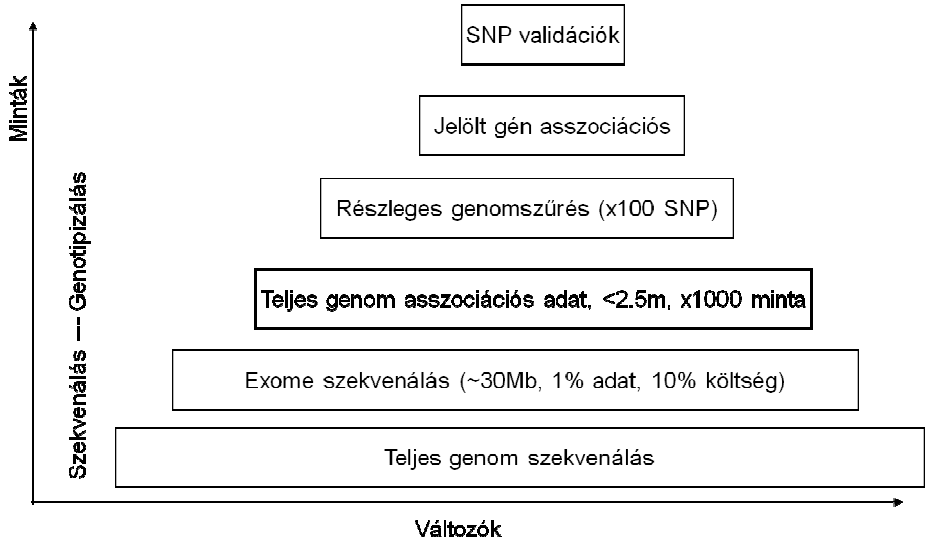


Figure 5 Genetikai adatok hierarchiája.

Több omikai méréstechnika figyelembevételénél ez többszörös kombinációt kínál, mivel kísérlettervezés során az egyes omikai szintek adatai integráltan is felhasználhatóak, egymást normatívan erősítve. Erre mutat példát a 3.2 Ábra.

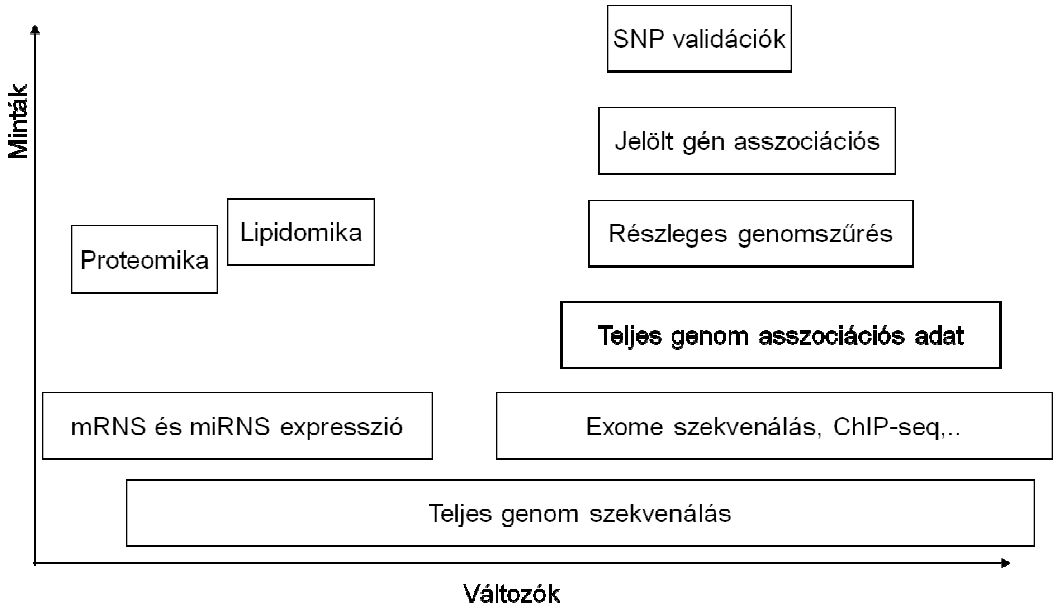


Figure 6 Omikai adatok hierarchiája.

A bioinformatika számára többek között új kihívást jelentettek a következők.

- A nagy mennyiségű omikai adatnak a szabványos tárolása, különös tekintettel a fenotípusok leírásának a standardizálására.
- Adatgyűjtés folyamatának támogatása, amely a standardizáláson túl adatvédelmi és adatbiztonsági szempontokat is kielégít, minőségbiztosítási szempontokat, és elosztott hozzáférést és integrációt is biztosít.
- Prioritizálási tehcnikák, amely a heterogén információforrások felhasználásával, benne a szakirodalom szövegbányászati feldolgozásával, adnak támogatást régiók, gének, vagy variánsok megerősítéses vizsgálatára.
- Szekvenciális kísérlettervezés, amely egy integrált kísérlettervezési-adatelemzési folyamatot tételez fel.

A következőkben ezen új területeket tekintjük át röviden, mivel a klasszikus kísérlettervezés adattípus, adatgyűjtéstípus, mintaméret, együttműködés, szisztematikus hiányzás, zavaró tényezők, modelltípusok, és kérdések tekintetében több epidemiológiai munka is elérhető.

1.2.1. Adatszabványok, ontológiák

A nagy mennyiségű omikai adatnak a szabványos tárolásának kérdése elsőként az időrendben elsőként populárissá váló omikai adatok, a génexpreszziós adatok kapcsán került elő. Megoldására született az MGED és MIAME szabványok kidolgozása, illetve a GEO adatbázis létrehozása. Genetikai adatok terén hasonló adatbázisok a dbGAP és az EGA adatbázisok.

Önálló megfontolást igényel egy speciális –omikai szint a fenome, azaz a fenotípusok leírásának a standardizálása. Ez sokáig lehetetlen vállalkozásnak tűnt, bár több kísérlet is volt rá. Az egyik az UMLS, amely egy meta-ontológia és ezzel integrálja a benne szereplő számos betegség ontológiáját, taxonomiáját, és szakszótárát. A fogalmak száma milliós nagyságrendű, azonban az egyes betegségspecifikus ontológiák eltérő kidolgozottsági szintje gyakorlatban nem tette sem népszerűvé, de különösen nem de facto szabvánnyá. Egy másik lehetséges vonal az egészségügyi finanszírozásban is használt IDC10 kódrendszer lehetne, azonban kutatási célokra ezen a granularitási szinten ez nem használható. Kényszerűségből így az egyes betegségek esetén önálló kódrendszert/szakszótárt szokás használni, ami azonban a kutatások és adatok integrálását teszi nehézkessé.

Ennek a problémának több módon is várható a megoldása. A teljes genomszélességű asszociációs vizsgálatok eredményessége elmaradt a remélt szinttől, ami a hiányzó örökletesség problémájának megfogalmazásához is vezetett [missingherit]. Ennek a feloldására egyebek mellett a részletesebb fenotípus használatát tartja egy lehetséges megoldásnak a tudományos közvélemény. Ennek a szemléletnek egy továbbvitele a gyakori betegségek olyan részletes szisztematikus leírását kívánja megvalósítani, amely az egyes gyakori betegség altípusokat a ritka szintig bontja le. A Human Phenotype Ontology (HPO) távlatilag olyan de facto szabványnak tekinthető ontológiát, benne pedig kódrendszert kíván adni a fenotípusok leírására, mint Gene Ontology (GO) a genomikai területen [3]. Ehhez a trendhez igazodika ritka betegségek vizsgálatának a kiemelt fontossága is, amelynek megfelelően az IDC11 már tartalmazni fogja a ritka betegségek jelentős részének kódolását.

A fenotípus leírás szabványosításával ideális esetben lehetségessé válnaa jelenlegi kétszeres, vagy sokszor akár még többszörös redundáns adatgyűjtés egyszerűsítése is, ami a finanszírozás, intézményi menedzsment, és kutatási projektek miatt most sokszor jelen van. Továbbá a fenotípus szint egységesítése valóban utat nyitna a hatékony meta-elemzések előtt is, illetve segítene megváltoztatni az adatok efemerális jellegét, és hosszabb időtávon is elérhetőséget és felhasználást tenne lehetővé.

1.2.2. Adatgyűjtés folyamatának támogatása

A genetikai és fenotípusos adatok standardizált reprezentálásának kidolgozása csak az egyik eleme az adatelemzés számára használható adatok biztosításának.

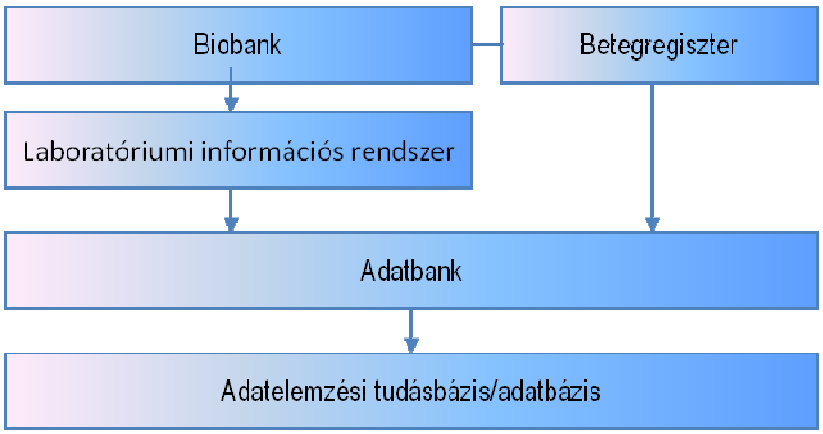


Figure 7 Az adatgyűjtés folyamata és moduljai.

Ideális esetben az adatok egy laboratóriumi információs rendszerből érkeznek, amely egy elektronikusan integrált biobankra támaszkodik. Ebben az esetben a biológiai minták követése, a mérés folyamata, a nyers és utófeldolgozott mérési adatok, automatizáltan kerülnek be nemzetközi integrációt lehetővé tévő adatbankokba. A fenotípusos adatok egy elosztott, internetes hozzáférésen keresztül a genetikai adatokkal integráltan kerülnek be betegregiszterekbe. Az adatelemzés eredményei pedig egy adatelemzési tudásbázisba kerülnek, ami komplex lekérdezést, illetve más eredményekkel való szemantikus integrálást tesz lehetővé.

Ennek a folyamatnak a támogatására, de különös tekintettel a betegregiszterek működésére és az adatok statisztikai minőségének biztosítására több ajánlást is megfogalmaztak, továbbá például adatvédelmi, adatbiztonsági szempontokat, hatókörökre bontott és elosztott hozzáférést, illetve az adatintegráció milyenségét. Az adatok minőségbiztosításának fontosságát jelzi, hogy sok esetben az adatelemzésre fordított idő szinte kizárólagosan adattisztításra és transzformációra fordítódik, illetve, hogy az adatok döntő része sok esetben nem is használható fel az elemzésben.

A WHO általános meghatározása szerint egy betegregiszter dokumentumok olyan gyűjteménye, mely egységes információt tartalmaz egyénekről, amelyeket szisztematikus és átfogó módon gyűjtöttek előre meghatározott tudományos, klinikai vagy szakmapolitikai célból. Létrehozása egy komplex folyamat, számos technikai és szervezési feladat megoldást igényel(a genetikai aspektusok miatt a jelenlegi kutatásokban egyre fontosabbá váló ritka betegségekre vonatkozó ajánlásokat lásd [4, 5]):

1.Tervezés

- cél meghatározása (prevalencia - incidencia monitorozás, betegség lefolyása – befolyásoló tényezői, biztonság monitorozás – beavatkozás, klinikai hatékonyság mérés, minőségmérés, etiológiai kutatás, betegnyilvántartás)
- adatkör kiválasztása, figyelembe véve, hogy ezen adatokat más helyen gyűjtötték-e (összehasonlíthatóság)

- vizsgálati populáció meghatározása (geográfiai, időbeni kiterjedés)

2. Implementációs terv, amely ideálisan előír egy pilot vizsgálatot az adatgyűjtés használhatóságáról, a rendszer érthetőségéről. Részletes dokumentációkat tartalmaz a minőség és hatékonyság növelésére a következő aspektusokról:

- Működtető megnevezése (szakmai, informatikai)
- Vizsgálati populáció, kiválasztási, kizárási kritériumok
- Adatok definiálása (kódszótár)
- Adatgyűjtés módszere
- Adatbevitel, adatbevivő képzése, motivációja, elkötelezettsége
- Minőségbiztosítás, folyamatos monitorozása és fejlesztése
- Adatfeldolgozás
- Hardver és szoftver leírás
- Elemzés leírása
- Adattárolás
- Adatvédelmi leírások
- Hozzáférési szabályok

5. Flexibilitás: az adatkör változtathatósága a tudomány állásának megfelelően

6. Technológia: az adatbevitel korszerű, online, felhasználó barát felületen keresztül.

7. Az adatbázis összekapcsolhatósága egyéb adatbázisokkal, nemzetközi összehasonlíthatóság.

8. Adatlekérdezés szabályozása, amely hatékony, időben elfogadható, flexibilis.

9. Az adatok és eredmények terjesztése, elosztott elemzés és értelmezés támogatása.

10. Etikai és jogi megfontolások

- Engedélyek
- Beleegyező nyilatkozat
- Személyi adatok védelme
    - Anonim
    - anonimizált (irreverzibilis)
    - közvetett módon azonosítható (kód)

- közvetlenül azonosítható

11. Költségvetés – finanszírozás, fenntartás kérdése.

A kísérlettervezés egy ideális és egyre gyakrabban elvárt eleme egy formális, szemantikus protokoll, amely egyrészt a szakértők számára is érthetően, másrészt az adatgyűjtő informatikai rendszer számára is direkt módon elérhető. Ez a klinikai aspektusok mellett, amelyek súgási funkciókkal elérhetőek, tartalmazza a gyűjtött adatok statisztikai típusát, kényszereket, összefüggéseket, strukturált kérdőív esetén a feltételeket.

1.2.3. Prioritizálási technikák, szövegbányászat

Az a priori ismeretek mennyiségének gyors növekedése, különösen az elektronikusan is elérhető a priori információk mennyiségének gyors növekedése a tudásmérnökség, a gépi tanulás és a számítógépes nyelvészet számára is újfajta elméleti és gyakorlati kihívást jelent. A jelenlegi szemantikai web-hez kapcsolódó erőfeszítések ellenére is, az a priori ismeretek jelentős része természetes nyelvű cikkkekben testesül meg, bár a szemantikusan strukturált tudományos cikkek terjedése vagy a formális tudásbázisok fejlődése mellett szerepük valószínűleg csökkeni fog. Ezekkel kapcsolatban két fontos módszercsalád a következő.

- **Információ keresés**, ami egy adott probléma, adott előéletű felhasználó, éppen aktuális információs igénye alapján próbál meg releváns cikket és/vagy esetleg bennük is releváns információt keresni.
- **Információ kivonatolás**, ami egy szűkebb értelmezésben egy szabatosan megfogalmazott típusbeli információt például adott típusú entitásokat és relációkat próbál meg kivonatolni, egy tágabb értelmezésben viszont célja inkább egy átfogó tudásbázis építés vagy a cikkek tömör összefoglalása.

A természetes nyelvű cikkkek különböző szintű elemzése így kitüntetett az a priori információk felhasználásában, amely elemzések a nyelvi strukturális („deep linguistic”) elemzéstől a szövegek egyszerű statisztikai („shallow statistical”) elemzéséig terjednek. Az elemzések egyik fontos célkitűzése egy adott szakterület entitásainak és relációiknak az automatikus kivonatolása, például fehérje-fehérje relációknak vagy kémiai anyagok-szimptómák-betegségek relációinak a kivonatolása. Az entitás relációk automatikus kivonatolásában a nyelvi strukturális elemzés mellett az egyszerű statisztikai elemzések is hasznosnak bizonyultak. Ezek a módszerek a nyelv struktúráját figyelmen kívül hagyják, azonban egyszerűségük miatt a módszerek számítási komplexitása nagyságrendekkel alacsonyabb a nyelvi elemzésekkel szemben, illetve egyszerűségük miatt más pontosság-megbízhatóság (accuracy-recall, specificity-sensitivity) karakterisztikával rendelkeznek mint a strukturális nyelvi elemzők. Ez a két tulajdonság együttesen nagy méretű

dokumentumgyűjteményeken való alkalmazhatóságot, illetve a lehetséges relációkat nagy eséllyel, de alacsonyabb pontossággal való detektálását jelenti, amely kombináció a gyakorlatban igen hasznosnak bizonyult egy adott szakterület áttekintésére, potenciális relációk feltérképezésére. Az entitásrelációk kivonatolásának szövegek egyszerű statisztikai elemzésén alapuló módszerei két tág, egymással átlapoló csoportba sorolhatók. Az egyik az entitások együttes előfordulásán alapul, azonos dokumentumban, bekezdésben, mondatban, mondatrészben, amely esetleg az entitások szótávolságát is figyelembe veszi. A másik módszer az entitások leírásának vagy idézettségüknek a hasonlóságán alapul, azaz például közös kulcsszavakon, közös hivatkozásokon. Ezek a megközelítések a következő módon csoportosíthatóak.

- Teljes (deep) nyelvészeti elemzés.
- Felszíni (shallow) nyelvészeti elemzés, ami például „part-of-speech” elemzést takar.
- Statisztikai elemzés, ami a fogalmak egyes cikkbeli előfordulási mintázatának a valószínűségi elemzése.
- Oksági elemzés, ami a fogalmak egyes cikkbeli előfordulási mintázatának a globális vagy lokális oksági elemzése.

Az együttes előfordulás legegyszerűbb esete az entitások neveinek együttes előfordulása, aminek orvosi, biológiai hasznosságát Swanson 1980-as évektől folytatódó vizsgálatai jelezték [6]. Genetikai területen is folyamatosan vizsgálták az ilyen relációk biológia adekvátságát, gyakorlati hasznosságát, illetve a gyakorlati alkalmazhatóság okait[7-10]. A szakértői, kézi elemzésük konklúziójaként megállapították, hogy a relációk pontosság-megbízhatóság karakterisztikája a gyakorlatban elfogadható biológiailag releváns relációk kivonatolására. Genetikai területen az első problémát az entitások nevek szinonímjainak a kezelése jelenti. A kivonatolt "szakirodalmi hálózat" összehasonlítása génaktivitási adatok elemzéséből kapott hierarchikus csoportosítással szintén megerősítette ezt a következtetést. Végeredményben a módszert összességében igen alkalmasnak minősítették nagy mennyiségű szakirodalom szempont mentes áttekintésére, kis szakértői és számítási időigény mellett.

Az entitások leírásának hasonlósága, különösen a strukturált annotálások hasonlósága egy kézenfekvő módszer entitás relációk származtatására (például gének, fehérjék esetében azonos biológiai folyamatban való részvétel, amit azonos kulcsszavak jelölnek). Egy indirekt, a szakirodalmat is felhasználó módszer, ha nem a leírások egymáshoz való statisztikai hasonlóságát vizsgáljuk, hanem a leírások szakcikkekhez való hasonlóságának korreláltságát, azaz a a szakcikkekhez való hasonlósági relációk hasonlóságát.

A felsorolt információ kivonatoló módszerek gyakorlati alkalmazhatósága a kiértékelések alapján megalapozott, azaz a kivonatolt relációk képesek egy intelligens áttekintést adni egy szakterületről, lehetséges relációkat beazonosítani, a relációkhoz kvantitatív, kvalitatív és szöveges jellemzőket is

kivonatolni (például név együttelőfordulási gyakoriságot, annak szignifikancia szintjét, illetve a közös előfordulást tartalmazó dokumentumok kulcsszavait).

Egy gyakori megközelítésben feltesszük, hogy a tárgyterület változóihoz tartozó kontrollált szótár, szinonima szótár, kifejezés szótár és elhagyandó szavak listája, illetve egy általános szóelemző alapján rendelkezésre állnak, ahol az egyes szavak információs tartalmát súlyok, például a TFIDF súlyok tükrözik [11].

Jelölje az X_j egy változó nevének (illetve szinonimjainak) a jelenlétét a d_i dokumentumban a bináris p^N_{ij} érték. Így $(p_{ij}^N)^N_{j=1}$ egy bináris, N dimenziójú vektor, ami az i dokument tartalmát fejezi ki az változók nevének jelenlétével. Hasonlóan $(p_{ij}^N)^L_{i=1}$ egy L dimenziós bináris vektor, ami az X_j változót jellemzi kapcsolódó dokumentumokkal. $\underline{\mathbf{P^N}}$ jelöli a teljes mátrixot a változók egy adott halmaza és egy adott dokumentumgyűjtemény esetén.

A változó leírások hasonlóságán és együttes relevanciáján alapuló módszerekhez a dokumentumok egy hasonlósági mértékét szokás például használni, ami az információ keresés esetén széles körben elterjedt. A hasonlósági mérték a dokumentumok vektor reprezentációján alapul, többféle szósúlyozás esetén, például az általunk is használt TFIDF szósúlyok esetén. A D_X, D_Y dokumentum párok szemantikus és információ elméleti kapcsolódását a normalizált tf-idf vektor reprezentációk W_X, W_Y szorzata definiálja.

$sim(D_X,D_Y)=cos(W_X,W_Y)$

Ekkor egyszerűen definiálhatók a kivonatolni kívánt relációk, amelyeket a feltevésünknek megfelelően változó párokra és feltételeket nem tartalmazó formában adunk meg. Jelöljön X és Y tárgyterületi változókat. Ekkor COOC(X;Y) jelöli az együttes név előfordulás relációt. A P^N_i és P^K_i valószínűségi változókat felhasználva a definíciók a következők (az X változó nevének jelenlétéhez és leírásának relevanciájához tartozó bináris valószínűségi változó jelölése P^N_X, P^K_X, amelyek értékeit p_x jelöli):

$COOC(X;Y) = P(P^N_X=1,P^N_Y=1|(P^N_X=1)\vee(P^N_Y=1))$

Becslésük a „szakirodalmi” adat alapján például $COOC(X_i;X_j) \approx n_{i,j}/ n_i+n_j-n_{i,j}$), ahol n_i, n_j és $n_{i,j}$ az X_i és X_j változók nevei előfordulásának a számai. Hasonló szövegalapú relációkat együttesen a továbbiakban mint „szöveges”, „annotációs” vagy „szakirodalmi” relációkra hivatkozunk, az TXT(X;Y) jelöléssel. Feltételezéseink szerint a relációk kétváltozósak, azaz egy élsúly mátrixal reprezentálhatók. A gráf alapú vizualizálást az élsúly-élvastagság összefüggés beállíthatóságával, illetve az élsúlyok megjelenítésére vonatkozó küszöbérték beállításával segíti. További lehetőség több reláció együttes megjelenítése a feltételek független beállításával. A relációtípusonkénti külön-külön feltételekkel

történő vizualizáció esetében az összehasonlítás túlnyomórészt a szakértő feladata. A relációk logikai és numerikus, iteratív kombinálása egy jövőbeni célkitűzés.

A szakirodalom alapú a priori ismeretek felhasználása mellett a kísérlettervezésben számos egyéb heterogén információforrás is rendelkezésre áll, amibe a korábbi adatok, és adatelemzések is beleértendők. Ez utóbbiak kezelése felveti, hogy a kísérlettervezés ideálisan egy integrált kísérlettervezést és adatelemzést jelent, azaz egy szekvenciális, adaptív kísérlettervezést (lásd Szekvenciális kísérlettervezés és Fúzió alfejezeteket). Az egylépéses kísérlettervezésre fókuszálva elsőként azt a tágabb kontextust illusztráljuk, amivel egy fúziós kísérlettervezés szembesül. A heterogén információforrások információt szolgáltatnak gének kromoszomális pozíciójáról, saját és más faj evolúcionisztikasan kapcsolódó „hasonló" génjeiről, átírási faktoraikról, esetleges alternatív átírási módjukról, lehetséges mutációikról. Adott kísérleti körülmények, adott sejttípus és mutáció esetén a gének átírási aktivitás mintázatáról (relatív mRNS gyakoriságokról) a gén chipek szolgáltatnak információt. A lehetséges gén produktumok, a fehérjék esetében az átírás utáni módosítás, funkció, struktúra, interakciókra vonatkozó információk mellett, tesztelhetők és megkereshetők a bizonyos önállósággal bíró egyszerűbb vagy bonyolultabb szekvencia részek. A megismert gén-fehérje átírási kapcsolatok, annak szabályozási faktorai, az empirikusan megerősített fehérje-fehérje interakciók, fehérjekomplexek, átírás utáni fehérje módositások, metabolikus, enzimatikus reakciók alapján végül komplex hálózatokat reprezentáló tudásbázisok építhetők. Ezek a tudásbázisok kapcsolódhatnak a terület általános logikai modelljét leíró ontológiákhoz, illetve egy tudásbázis lefedhet több fajt is, ahol a tudásreprezentáció a fajok taxonómiai elrendeződésének megfelelően hierarchikusan kódolja a gén-fehérje-metabolit hálózatokat, azaz egy entitást vagy relációt mindig a lehető legáltalánosabb, az evolúciós fa lehető legősibb csomópontjának szintjén reprezentál.

Szekvencia adatbázisok: GenBank, EMBL, ExProt, SWISS-PROT/TrEMBL, PIR,TRANSFAC

Fehérje motif adatbázisok: Blocks, InterPro, Pfam, PRINTS, SUPFAM, PROSITE

Fehérje struktúra adatbázisok: PDB, MMDB

Génaktivitás-mintázatok adatbázisai: GEO, YMGV

Molekuláris kölcsönhatások adatbázisai: BIND, DIP, BRENDA

Metabolikus hálózat adatbázisok: EcoCyc, MetaCyc, GeneNet

Mutációs adatbázisok: OMIM

Ontológiák, tezauruszok: Go, UMLS, MeSH, Galen

Publikációk: PubMed

A biológiai információk sokfélesége és mennyisége külön-külön is új kihívásokat jelentett több tudományág számára. A nukleotid és fehérje szekvenciák visszakeresése a „hasonlóság" és illesztés új statisztikai és algoritmikus módszereket igényelt. A szekvenciák átírás szempontjából kitüntetett részeinek felismerése, az autonóm, moduláris elemeinek (motívumainak) automatikus felismerése új módszerek sokaságát indukálta. Természetesen a fehérjék térszerkezetének, interakcióiknak, komplexeiknek a megjósolása továbbra is központi jelentőségű maradt. A nem struktúrált szövegek visszakeresése új szótárak, tezauruszok és ontológiák létrehozását indukálta, amelyek mérete, komplexitása és sokasága is kihívást jelentett. A természetes nyelvi, esetleg részben strukturált publikációk természetes nyelvi elemzése, statisztikai elemzése, (gén, fehérje, gyógyszer) nevek és entitás relációk felismerése és automatikus kivonatolása egy új, komplex és gyakorlati alkalmazást jelent a nyelvi elemzés számára. A tudásbázisok, a tudásmérnökség számára is egy precedens nélküli kihívást jelentett és jelent az ilyen nagy tömegű és ilyen sokféle információ reprezentálása, automatikus vagy kézi létrehozása, karbantartatása, validálása.

Az igazi kihívást azonban a sokféle információ együttes kezelése és elemzése jelenti, ahol illusztrációként tekintsük azt az egyszerűsítést, amikor az a priori információkat páronkénti relációkként formalizáljuk. Ekkor a következő entitás-entitás relációk fogalmazhatóak meg

1. Gén-gén többváltozós relációk
    - génszekvenciák (evolúcionisztikus) hasonlósága, illeszthetősége
    - közös átírási faktorok
    - kromoszomális közelség, azonos génklaszterekben szereplés
    - Különböző genomokbani való jelenlétből generált mintázatok (filogenetikai mintázat) hasonlósága
    - Korábbi kísérletekben mért génaktivitás mintázatok hasonlósága, vagy azok klaszterezése alapján generált klaszterekben való együttes előfordulás
    - Ismert közös átírási szabályozás
    - Szakirodalmi relációk, például gyakori együttes előfordulás publikációkban (lásd még külön is)
2. Fehérje-fehérje többváltozós relációk
    - Fehérje szekvenciák hasonlósága, illeszthetősége
    - Fehérje struktúrák térbeli hasonlósága, illeszthetősége
    - Interakció, közös komplexumban való részvétel kísérleti megállapítása
    - Genomok összehasonlításából származó összekapcsolódásra utaló reláció (domain fusion)
    - Ismert közös metabolikus hálózatban való részvétel
    - Szakirodalmi relációk, például gyakori együttes előfordulás publikációkban (lásd még külön is)
3. Gén-fehérje több-több változós relációk (v.ö. az egy gén-egy fehérje egyszerűsítéssel)
    - Egy gén aktiváláshoz kapcsolódó egyéb gének és fehérjék
    - Egy fehérje, fehérje komplexum kialakulásához releváns gének és fehérjék

- Szakirodalmi relációk, például gyakori együttes előfordulás publikációkban (lásd még külön is)

4. Általános relációk, főként szakértői annotációkból, leírásokból, vagy a szakirodalomból származó relációk
    - entitások (gének, fehérjék, gyógyszerek, betegségek) nevének (és azok szinonimjainak) együttes előfordulási gyakorisága publikációkban vagy az ebből generált klaszterezés adta közös csoporttagság
    - entitások kulcsszavas, strukturált vagy természetes nyelvi leírásának a statisztikai hasonlósága vagy vagy az ebből generált klaszterezés adta közös csoporttagság
    - típusos entitások típusos relációinak, mint „szabályoz", „kötődik", „okoz", „hat" a kivonatolása a természetes nyelvi publikációkból

Ezekben az esetekben az entitások hasonlósága egy skaláris mennyiséggel reprezentálható, ami például gének szintjén az egyes információforrások esetében egy-egy hasonlósági mátrixot jelent. A kísérlettervezés egy gyakori, megerősítési fázisban fontossá váló kérdése ekkor a következő problémára vezet. Tegyük fel, hogy omikai vizsgálatokból egy génhalmaz adott, mint szignifikánsan releváns hipotézis. Ekkor a megerősítésükre egy olyan kísérlet tervezhető, amely már főként ezekere a génekre, és hozzájuk kapcsolódó génekre irányul. Ennek a kiterjesztett génhalmaznak a megkeresése egy "google"-szerű megközelítésben is elképzelhető, amelyben a kiindulási génhalmazunkat mint kérdés tesszük fel, és a hasonlósági mércék alapján keresünk olyan más géneket, amelyek átlagos hasonlósága a kérdében szereplő génekhez nagy. Természetesen génszint és gének helyett más szint is analog módon használható, például variánsok, ennek megfelelően prioritizáló rendszerek mind a gén vagy régió és mind a variáns prioritizálásban is használtak. A részleges GA kísérletek tervezésénél az első kérdés a lefedni kívánt régió vagy gének meghatározása. Ez szaktudáson, szakirodalmon és szövegbányászati eszközökön, útvonal adatbázisokon, tudásbázisokon, már gyanús géneken, korábbi expressziós és protemikai adatokon alapulhat. A régió vagy gének meghatározása után válik lehetővé az SNP-k meghatározása, amit a tervezett funkcionális elemzés szempontjai (pl. kódolás, szabályozás) és a primertervezés is befolyásolnak. Ekkor az a feladat, hogy egy adott orvosbiológiai fogalomhoz vagy a priori génhalmazhoz keressünk releváns géneket felhasználva a szakirodalmat, korábbi adatokat, és meglévő tudásbázisokat, mára már egy önálló problémakörré vált, ez az úgynevezett génprioritizálási feladat. A génprioritizáló rendszerek egy sikeres csoportja a szakirodalomból épített útvonal tudásbázisok és az adatokból épített korrelációs mátrixokból fúziós technikákkal javasol egy a priori génhalmazhoz további releváns géneket. Jelentősebb génprioritizáló rendszereket a lenti táblázat sorol fel.

Table 1 Génprioritizáló rendszerek.

Tool	Website
SUSPECT	http://www.genetics.med.ed.ac.uk/suspects/
ToppGene	http://toppgene.cchmc.org/
PolySearch	http://wishart.biology.ualberta.ca/polysearch/index.htm
Prioritizer	http://www.prioritizer.nl

CAESAR	http://visionlab.bio.unc.edu/caesar/
MimMiner GeneSeeker	http://www.cmbi.ru.nl/MimMiner/cgi-bin/main.pl
PhenoPred	http://www.phenopred.org
PGMapper	http://www.genediscovery.org/pgmapper/index.jsp
Endevour	http://www.esat.kuleuven.be/endeavour
G2D	http://www.ogic.ca/projects/g2d_2/
TOM	http://www-micrel.deis.unibo.it/~tom/
SNPs3D	http://www.SNPs3D.org
GenTrepid	http://www.gentrepid.org/
GFSST	http://gfsst.nci.nih.gov/
GeneWanderer	http://compbio.charite.de/genewanderer
Bitola	http://www.mf.uni-lj.si/bitola/
CANDID	https://dsgweb.wustl.edu/hutz/candid.html
aGeneApart	http://www.esat.kuleuven.be/ageneapart
GenePredictor	http://www.hugenavigator.net/HuGENavigator/geneProspectorStartPage.do

SNP-prioritizáló rendszereket a lenti táblázat mutat be.

Table 2 SNP-prioritizáló rendszerek.

Alkalmazás	Adatbázis	Döntéstámogatás	Keresés	Pontozás	Halmazkiválasztás	Többrétegű	Tárgyterület
Genomizer	Több	Nincs	Van	Van	Nincs	Nem	SNP
QuickSNP	Egy	Nincs	Van	Nincs	Nincs	Nem	SNP
GoldSurfer	Egy	Nincs	Van	Nincs	Nincs	Nem	SNP
HAPLOT	Egy	Nincs	Van	Nincs	Nincs	Nem	SNP
SNPHunter	Egy	Nincs	Van	Nincs	Van	Nem	SNP
SNPSelector	Egy	Nincs	Van	Van	Nincs	Nem	SNP
OSIRIS	Több	Van	Van	Nincs	Nincs	Nem	Gén
SNPbrowser	Több	Nincs	Van	Nincs	Nincs	Nem	SNP
SNPs3D	Több	Van	Van	Nincs	Nincs	Nem	Gén/SNP

SNPStats	Egy	Van	Nincs	Nincs	Nincs	Nem	SNP

- *Adatbázis – az alkalmazás egy, vagy több orvosbiológiai adatbázisra támaszkodik*
- *Döntéstámogatás – az alkalmazás döntéstámogatást nyújt*
- *Keresés – az alkalmazás lehetőséget ad az elemek szűrésére, keresésére*
- *Pontozás - az alkalmazás használ pontozásos rendszert az elemek sorrendezéséhez, a döntéstámogatáshoz*
- *Halmazkiválasztás - az alkalmazás alkalmaz halmazkiválasztási algoritmust*
- *Többrétegű - az alkalmazás többrétegű modellt alkalmaz*
- *Tárgyterület - az alkalmazás által kezelt elemek típusa*

A prioritizáló rendszerek szekvenciális felhasználása, azaz szekvenciális kísérlettevezésre való felhasználása egy ígéretes lehetőség, amelyhez áttekintjük az adaptív kísérlettervezés lényegi elemeit.

1.2.4. Szekvenciális kísérlettervezés

A korábbi kísérletekből származó adatok elérése egyesíti a kísérlettervezést és adatelemzést, mivel a megfigyelni vagy módosítani kívánt változó kiválasztása, a gyűjteni kívánt adat mennyisége egyre inkább direkt módon függ korábbi adatoktól, az azokból levont konklúziók és az a priori szaktudás mellett. A szekvenciális kísérlettervezés természetesen a metaelemzések elméletére is támaszkodik, ami akorábbi kísérletekből származó adatok újrafelhasználásának kérdésköre, magába foglalva az eltérő protokoll szerint gyűjtött adatoknak az együttes felhasználásáának és a hiányzós megfigyelések kezelésének problémáit. Az adaptív statisztikai kíséletervezés gyakran alkalmazott eljárás a klinikai kísérletek során is, különös tekintettel a farmakológiára. A mintagyûjtés gazdasági, orvosi és egyéb költségei miatt az adaptív kísérlettervezés hagyományosan a mintaméret adaptív megválasztására fókuszál, a változók egy fix méretû halmazát, és adott statisztikai tesztet feltételezve.

Az egyre népszerűbb , általánosabb keretet kínáló, és didaktikailag is egyszerűbb Bayesi megközelítés a tudásgazdag orvosbiológiai területhez megfelelõ eszközt nyújt, amelyet a hagyományos, kisváltozószámú szekvenciális kísérlettervezés (SSD) területén már alkalmaztak. A genetikai célterületen, például a parciális szekvenciális asszociációs kísérlettervezés során, ekkor megpróbáljuk az egymást követõ, egymásraépülõ kísérletek során kapott adatokra alapozva a lépések közötti elemzéssel, illetve meta-analízis segítségével kiválasztani azokat a változókat, amelyek relevánsak az adott betegségre nézve.

A szekvenciális kísérlettervezés fázisai

Ez illeszkedik a teljes (GWA) és részleges (PGA) genetikai asszociációs kutatások egymásra épüléséhez is, amit a lenti ábra illusztrál.

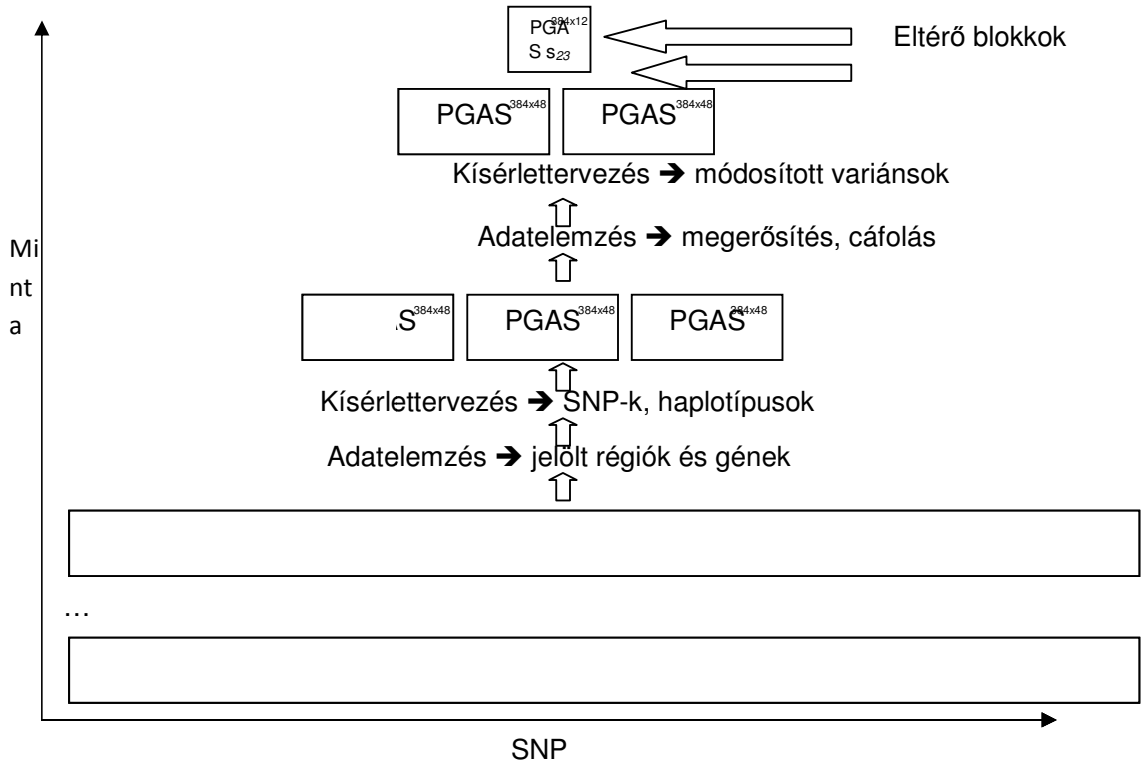


Figure 8 ***A teljes (GWA) és részleges (PGA) genetikai asszociációs kutatások egymásra épülése.***

Az adaptív kísérlettervezés kérdése még inkább felvetődik, ha a genotipizáló készülék átbocsátóképessége limitált, azaz ha olyan elemzésekben is használják, ahol a változó és mintaszám a készülék többszöri felhasználásával biztosítható. Ekkor a mérésekkel párhuzamosan zajló elemzéssel válik lehetségessé az adaptív kísérlettervezés.

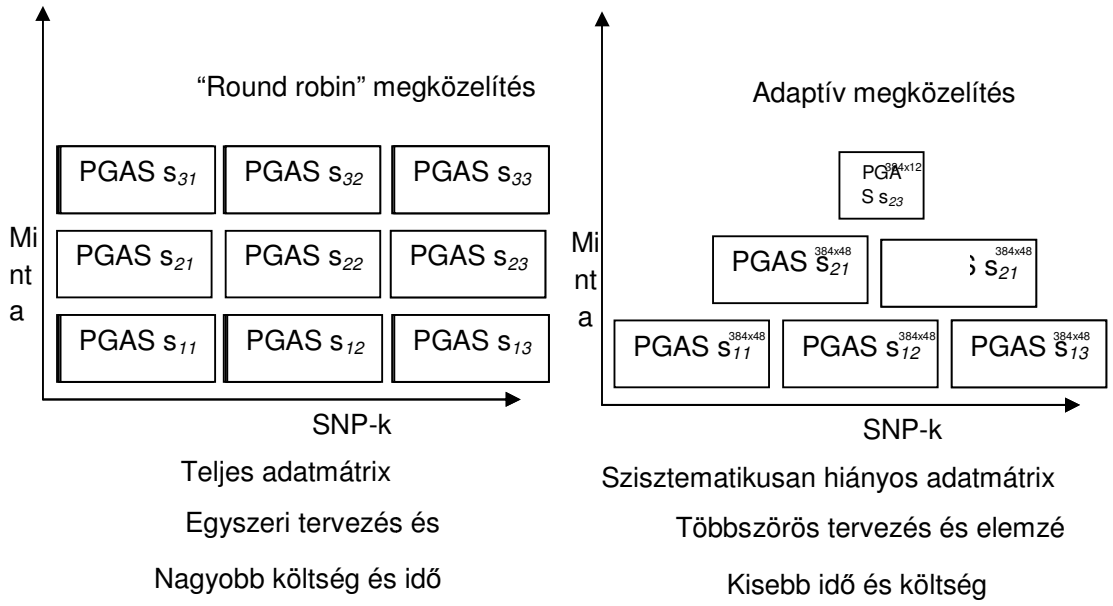


Figure 9 A változószám csökkentés az adaptív kísérlettervezésben.

A változószám csökkentésére több elméleti keret is lehetséges. Egy flexibilis elméleti keretet a Bayesi döntéselmélet szolgáltat, amit az 10. ábra illusztrál.

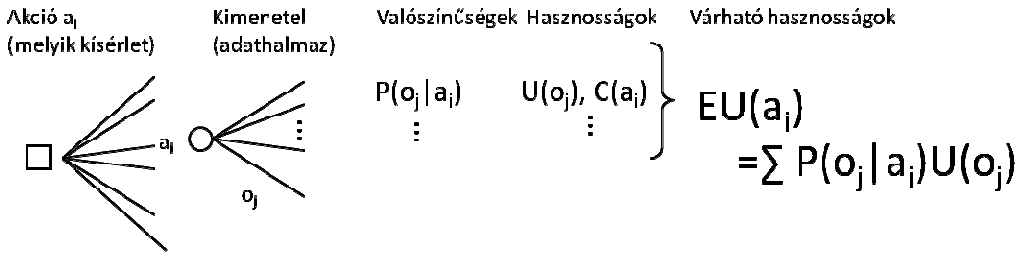


Figure 10 A kísérlettervezés bayesi döntéselméleti kerete.

A bayesi döntéselméleti keretben a kísérlettervezés tipikusan a gyűjtött változók halmazának és a minták számának a megválasztását jelenti. Ezt az adatgyűjtés után adott hipotézisek, vagy akár felfedezések megkonstruálását jelenti különböző valószínűséggel és hasznossággal, amelyekből az adott kísérlet várható hasznossága kiszámítható. Az egyes kísérlettevezéshez kiszámított várható hasznosságok alapján kvantitatív módon lehet dönteni az optimális kísérlettervezésről. Korábbi munkákban már voltak javaslatok Bayesi eljárások felhasználására gén szabályozási hálózatok felderítéséhez microarray kísérletek alkalmazásával. Ezek már mind a változók, mind a minták dimenziójában bemutatnak beavatkozásos, illetve megfigyeléses módszert is. Az eljárás lokális oksági kapcsolatok feltérképezésére mutatott példát.

További lehetőségként annak az elméleti megfontolása is felvetődik, hogy vegyük figyelembe a minták szekvenciális érkezését, amelyek adott gyakorisággal, költséggel, és hasznossággal érkeznek, amelyek kontextuális és megfigyelendő változóinak van egy-egy adott költsége. Ekkor a mintákról való döntési probléma az aktív tanulási keretben fogalmazható meg.

1.3. Adatelemzés

A monogénes, főként családfa elemzésekre támaszkodó korszakot az 1990-es évektől egyre inkább felváltotta a a gyakori betegségek-gyakori variánsok hipotézis és az asszociációs vizsgálatok. Ezek teljesfokú megvalósítói a teljes genomszélességű asszociációs elemzések, amelyek akár SNP, CNV, vagy akár epigenetikai szinteken is lehetségesek. A reméltnél kevesebb eredmény miatt került megfogalmazásra a hiányzó örökletesség problémája, amely a korábban már említett részletes fenotípus igény mellett a teljes genomszélességű asszociációs vizsgálatok másik elemét is górcső alá veszi, és a gyakori betegségek-gyakori variánsok hipotézis helyett a ritka genetikai variánsokat helyezi előtérbe [12]. Az összekapcsolni kívánt két szint mellett az összekapcsolásra szolgáló statisztikai modelleket és módszereket is komoly kritika érte. B.Maher az alábbi pontokban összegezte a „hiányzó örökletesség” jelensége mögött lehetséges főbb magyarázatokat[13].

- „Right under everyone’s noses”: ritka variánsok szerepe,
- „Out of sight”: nagy számú, gyenge hatás szerepe,
- „In the architecture”: strukturális, rendszerszrintű jelenségek szerepe,
- „In underground networks”: episztatikus interakciók szerepe,
- „The great beyond”: epigenetika szerepe,
- „Lost in diagnosis”: részletesebb fenotípus szerepe.

Ez egyrészt további biológiai szintek fontosságára, azaz azok méréseire hívta fel a figyelmet, mint ritka variánsok és epigenetikai módosulások szintjeire, másrészt a már tárgyalt részletesebb fenotípusos leírásra (lásd 11. ábra).

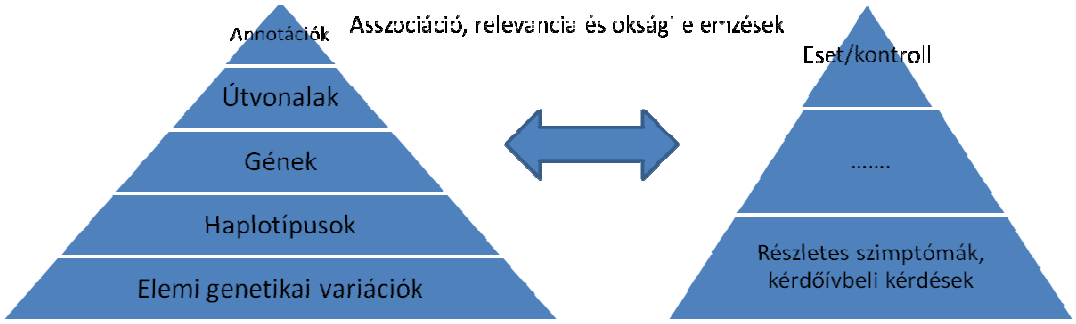


Figure 11 Asszociációk és oksági kapcsolatok vizsgálata hierarchikus prediktor és kimeneti változók esetén.

Továbbá azonban olyan statisztikai, modellezési technikák szükségességére, amelyek a gyenge hatáserősségeket, a többváltozós relációkat, a statisztikai és biológiai interakciókat, a környezet és életstílus módosító hatásait, nagyszámú zavaró tényezőt is képesek kezelni, és oksági kapcsolatok felderítését és hatáserősségek megbecslését is lehetővé teszik. Ezek a kérdések matematikai szempontból sok tekintetben összekapcsolódnak és a komplex modellek stasztikai használatának kérdéskörében egységesen tárgyalhatók. Az 1980-as évektől kialakuló „gráfos valószínűségi modellek" („probabilistic graphical models") kutatása egy új szakterületet hozott létre, amely a statisztikai modellek felhasználásában, kommunikálásában is új lehetőségeket nyitott. A gráfos valószínűségi modellek osztályától elméletileg független az alkalmazására szolgáló statisztikai keretrendszer megválasztása. Azonban ezen modellosztály használatához szükséges statisztikai minták nagy mennyisége és a nagy számítási komplexitás a bayes statisztikai keretrendszert helyezte előtérbe, amit a számítástechnikai hardver fejlődése is támogatott.

A genetikai vizsgálatokban az oksági kapcsolatok felismerésének nehezítő körülményeit illusztrálja az 12. ábra. Zavaró tényező van jelen, akkor ha a genotípus és a fenotípus is asszociált egy alpopulációs markerrel (etnikum), helyszínnel (éghajlat, környezet), életkorral (bevándorlás), akkor az egy nem oki statisztikai asszociációt hoz létre a kérdéses genotípus és fenotípus között. További nehézséget jelent a nem-funkcionális, kapcsolt genetikai variánsok jelenléte.

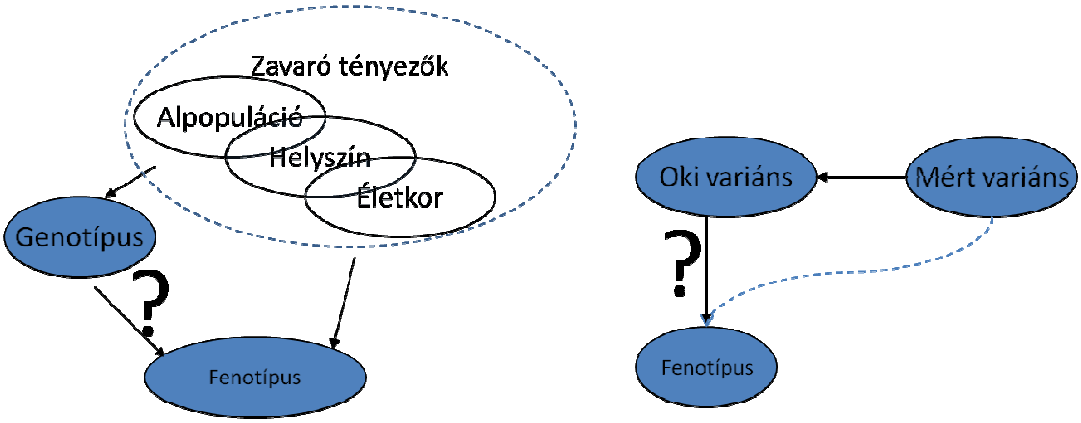


Figure 12 Oksági kapcsolatok felismerésének nehezítő körülményei: zavaró tényezők és nem-funkcionális, kapcsolt variánsok.

Az oksági kapcsolatok megállapításához szükséges követelményeknek egy gyakran hivatkozott listája az alábbi (X➔Y viszonylatában):

- Valószínűségi (statisztikai) asszociáció.
- Időbeli asszimmetria.
- Dózis-hatás (beavatkozási) reláció
    - szükségesség: X megszüntetésének, csökkentésének hatása,
    - elégségesség: X fennálllásának, növelésének hatása.
- Kontrafaktuális (be nem bekövetkezett/nem fennálló tényállásokon alapuló) reláció
    - szükségesség: Ha X nem lett volna, Y sem,
    - elégségesség: Ha X lett volna, Y is.
- Korlátos környezeti függés: az elkerülhetetlen környezeti feltételek elég gyakoriak, hogy a reláció releváns legyen.
- Elfogadható mechanisztikus magyarázat létezése, amely a zavaró tényezők hatását eliminálja.

A fenti követelményeknek eleget tevő relációk elvártan egy autonóm, mechanizmust azonosítanak be, amely a környezeti feltételek mellett stabil, hordozható, és felhasználható több szinten is következtetésekre:

- megfigyeléseken,
- beavatkozásokon,
- be nem bekövetkezett/nem fennálló tényállásokon eseményeken

alapúló következtetések elvégzésére. Ezek a következtetések a gyógyszerkutatás, kísérlettervezés, megelőzés, diagnózis, terápiaválasztás esetén is megjelennek, illetve a monitorozás, kiértékelés fázisában is, továbbá akár olyan akár olyan újszerű kérdésekben, amikor egy korábbi beteganyagon szeretnénk egy valódítól eltérő terápia hatásosságát megbecsülni

A következő alfejezetben részletesen tárgyaljuk a gráfos valószínűségi modellosztályt, ami a fenti oksági következtetéseket is támogatja [14, 15]. Az általános adatelemzési alfejezet zárásaként a bayes statisztikai keretrendszer főbb jellemzőit foglaljuk még össze [16].

A bayesi megközelítés alapját és értelmezésének egyszerűségét a bayes szabály biztosítja, amely a feltételes valószínűségek definiciójából

$$p(X \mid Y) = \frac{p(Y, X)}{p(Y)} \; p(Y \mid X) = \frac{p(Y, X)}{p(X)}$$

egy egyből következő állítás

$$p(X \mid Y) = \frac{p(Y \mid X)\, p(X)}{p(Y)} \propto p(Y \mid X)\, p(X)\,.$$

Formálisan egy inverziót tesz lehetővé, ami az a posteriori valószínűség (poszterior) p(X|Y) mennyiség a prior valószínűségből (priorból) való származtatását teszi lehetővé (a konstans erejéig egyenlő $\propto$ jel felhasználásával ez még egyszerűbben jelezhető). Diagnosztikai felhasználásaa szakértői tudásként gyakran elérhető p(Következmény|Ok) okozati iránnyal megegyező feltételes valószínűségek „átfordítása” diagnosztikai irányba:

$$p(Ok \mid Következmény) \propto p(Következmény \mid Ok) \times p(Ok)\,.$$

Gyakori statisztikai megfogalmazása pedig a következő

$$p(Model \mid Data) \propto p(Data \mid Model)\, p(Model)\,,$$

amely direkt valószínűségi állítást (a poszteriort) eredményez egy modellre az adatok tükrében, amely a prior és a „likelihood” p(Data|Modell) faktorok egyensúlyából származik.

A modellekre vonatkozó direkt valószínűségi állítások eltérnek a hipotézistesztelésbeli megismétlésekre vonatkozó cáfolati valószínűségi értékétől (p-értéktől). A modellekre vonatkozó a posteriori értékek követik a bayesi megközelítésnél használt szubjektív értelmezését a valószínűségnek, amely természetesen nem érinti a valószínűségszámítás axiómáit. A szubjektív értelmezés szerint a valószínűségek egy koherens preferencia relációt reprezentálnak események bekövetkeztére vonatkozós elvárásokkal kapcsolatban, azaz egy adott probléma kapcsán a valószínűségi értékek szubjektívek, információfüggők, ugyanannak a ténynek a valószínűsége más és más különböző rendszerekben. Ez az értelmezés eltér a mérnöki gyakorlatban inkább elterjedt objektív vagy frekventista értelmezésektő, amelyek szerint bizonyos jelenségeknél egy rögzített eloszlás jelenléte inherens tulajdonság („fizikai” következmény), illetve, hogy tapasztalati törvény, hogy bizonyos jelenségek sorozatos megfigyelésénél a sorozatok adott konvergencia tulajdonságokat mutatnak (például a gyakoriság mint határérték létezik és ez a határérték minden effektíve kiválasztható részsorozatra ugyanannyi). Azonban bármelyik értelmezésnél a valószínűségszámításnak ugyanaz a formális rendszere adódik. Továbbá a valószínűségek szubjektív értelmezésénél sem helytálló az a kijelentés, hogy ekkor ezek önkényesen, a „valóságtól” függetlenül választhatók meg. Egyrészt az a priori valószínűségek is valóságos tapasztalatokon nyugvó preferencia relációhoz tartoznak, másrészt mint azt a Bayes szabály kapcsán láttuk, az a priori értékeket a megfigyelések úgy módosíthatják, hogy a megfigyelések súlya egyre növekszik és meghatározóvá válik.

A bayesi megközelítés leegyszerűsítve a következőket javasolja.

1. Az a priori tudás alapján konstruáljunk egy $Q(\boldsymbol{\Theta})$ a priori eloszlást, ahol a $\boldsymbol{\theta}$ paraméterhez/címkéhez tartozó modellek elképzelhetőnek tartott valószínűségi eloszlását definiálják a lehetséges kimeneteleknek (likelihood modellek). A felhasznált modellek alkothatnak hierarchiát vagy tartozhatnak különböző modelcsaládokba is.
2. Az $x_1,..,x_m$ megfigyelések a Bayes szabály alapján meghatázorrák a $Q(\boldsymbol{\Theta}|x_1,..,x_m)$ a posteriori eloszlást.

   $Q(\boldsymbol{\theta}|x_1,..,x_m) \propto Q(\boldsymbol{\theta})L(x_1,..,x_m; \boldsymbol{\theta})$, ahol a megfigyeléseken alapuló $L(x_1,..,x_m; \boldsymbol{\theta})$ likelihood

   $L(x_1,..,x_m; \boldsymbol{\theta}) =_{def} p(x_1,..,x_m|\boldsymbol{\theta})$, amely függetlenség feltételezése esetén $\Pi_{i=1..m}\, p(x_i|\boldsymbol{\theta})$.

A következtetés során a feladat, hogy megbecsüljük egy adott esemény, vagy egy modell feltételes valószínűségét az alapismereteink és a megfigyelési adatok szerint. Az első esetben *prediktív,* a másodikban *parametrikus* becslésről (a posteriori eloszlás számításáról) beszélünk. Mindkét esetben a Bayes-tételből indulunk ki, amelynek segítségével események feltételes valószínűségét számíthatjuk (a továbbiakban D az adatokat, G egy struktúrát, θ pedig egy paraméterezést jelöl):

(1) $$P(G,\theta \mid D) = \frac{P(D \mid G,\theta)P(G,\theta)}{P(D)}$$

Az a posteriori eloszlást (röviden *posterior*t) az előzőleg említett modelltér egy struktúrájának paraméterezésére, vagy magára a struktúrák terére is kiszámíthatjuk. A paraméterek esetén a képlet formailag megegyezik (1)-gyel, a struktúrákra vonatkozó pedig a paraméterek kiintegrálása után adódik:

(2) $$P(G \mid D) = \frac{\int P(D \mid G,\theta)P(G,\theta)d\theta}{P(D)}$$

Predikció esetén még egy lépést teszünk: a keresett valószínűséget kiszámítjuk minden létező modellre, és ezeknek a modellek a posteriori valószínűségével súlyozott átlagát vesszük:

(3) $$p(x \mid D) = \sum_k p(G_k \mid D) \int p(x \mid \theta_k) p(\theta_k \mid G_k, D) d\theta_k$$

A fenti posteriorok gyakran nem mintavételezhetők, ezért Monte Carlo módszereket kell alkalmaznunk, például a fontossági mintavételezést vagy a Markov-lánc Monte Carlo (MCMC) módszerek egyikét [17]. A posterior vizsgálata helyett a feladat gyakran egy vagy több $\overline{f} = E_{p(x|D)}[f(x)]$ alakú várható érték becslésére egyszerűsödik. Ez megtehető a következő lépésekkel, melyek helyességét az MCMC módszereknél igazolt nagy számok törvénye (2. pont) és centrális határeloszlás tétel (3. pont) biztosítja:

1. a $\{x_i\}_{i=1}^N$ minta vételezése
2. $\overline{f}$ becslése az $\hat{f} = \frac{1}{N}\sum_{i=1}^N f(x_i)$ képlet alapján
3. konfidenciabecslés az $\left|\overline{f} - \hat{f}\right|$ eltérésre

A Monte Carlo mintavételezés mellett még gyakori a meghatározott számú, legnagyobb a posteriori valószínűségű modellstruktúra alapján történő kiszámítás, amely legegyszerűbb esetben egyetlen, az ún. MAP (maximum a posteriori) modell használatát jelenti.

A következő listában röviden összefoglaljuk a fentebb ismertetett bayesi módszertannak a klasszikus statisztikával szembeni előnyeit :

- A paraméterek bizonytalanságát a felettük definiált eloszlással jellemezzük, így minden statisztikai következtetés egy direkt valószínűségi állítás, ami az automatizált többlépéses tanulási rendszereknél és tudásbázisok generálásánál igen előnyös.
- A paraméterbecslés egy inverziós feladatként fogható fel, hisz itt kizárólag az adatból következtetünk arra a paraméterre, amely annak generálását meghatározta. A Bayes-tétel pontosan ezt az inverziót formalizálja, így a következtetést a hipotetikus viselkedés figyelmen kívül hagyásával végzi, szemben a klasszikus statisztika egyes módszereivel.

- Az a priori eloszlások (röviden *priorok*) használata alkalmas az előismeretek összegzésére vagy akár a teljes ismerethiány kifejezésére is.
- A priorok – mivel leggyakrabban korábbi megfigyeléseken vagy vizsgálatokon alapulnak – az ismeretszerzési folyamat egyes fázisainak tekinthetők, hisz új tudásunkat (az a posteriori eloszlást) ez alapján szerezzük.
- A bayesi következtetés a Bayes-tétel segítségével egyenrangú módon, normatívan kombinálja az előismeretekben és az adatokban rejlő információkat. Így a Bayes-tétel használata az adatok és az előismeretek egyfajta súlyozását valósítja meg: az adatok mennyiségének növekedtével azok befolyása is nő az a posteriori eloszlásra.
- Az a posteriori eloszlás használata pontbecslés helyett a predikció során nem csak a legvalószínűbb konfiguráció alapján számol, hanem figyelembe veszi a kevésbé valószínű eseteket is, ami a modell komplexitásához képest kis mennyiségű megfigyelés esetén fontos.

A főbb különbségeket a hipotézistesztelés és bayesi megközelítés között az 3. táblázat tartalmazza.

Table 3 A hipotézistesztelés és bayesi megközelítés főbb tulajdonságai.

Hipotézistesztelés („frekventista”)	Bayesi
-	A priori valószínűségek
Null hipotézis	-
Indirekt: cáfolati bizonyítás	Direkt állítás
Modell kiválasztás	Modell átlagolás
p-érték	-!
-!	A posteriori valószínűségek
Konfidencia intervallum	„Credible” régió
Szignifikancia szint	Költségmátrix.
Korrigálás többszörös hipotézistesztelésre	Hipotézistér együttes figyelembevétele

1.4. Gráfos valószínűségi modellek és oksági modellek

A hiányzó örökletesség problémájánál előlegeztük meg, hogy a gráfos valószínűségi modellek több javasolt irányt is lefednek, például a többszörös hipotézistesztelés kezelését, interakciók kezelését, a zavaró tényezők kezelését, és oksági kapcsolatok felderítését. Alapvető és viszonylagosan új sajátossága ennek a modellosztálynak a tárgyterületi jellege, azaz a prediktorok és kimeneteli változók együttes, sőt uniform kezelése. Ez eltér a feltételes modellektől, mint például a regressziós modellektől, beleértve az általánosított és logisztikus regressziót is, neurális hálózatokat, kernel gépeket, illetve bizonyos értelemben a klaszterezési módszereket is.

A továbbiakban feltételezzük, hogy a döntési helyzethez tartozó változók diszkrétek, létezik elfogadott a priori diszkretizálás vagy a diszkretizálás statisztikai megfontolások szerint történik, így adottak az $\mathbf{x}_i$ elemi események, a változók értékeinek együttes konfigurációi $\mathbf{x}_i=[x_{i_1},..,x_{i_n}]$. Jelölje az elemi események számát N. Így a cél az $\mathbf{X}=[X_1,..,X_n]$diszkrét valószínűségi változók $P(\mathbf{X})$ eloszlásának, azaz pontosabban a $P(\mathbf{X}|\boldsymbol{\theta})$ parametrikus eloszlásoknak és egy kapcsolódó $Q(\boldsymbol{\Theta})$ a priori eloszlásnak a kezelése. Feltesszük, hogy létezik egy X_j változó, továbbiakban Y-nal jelöljük, ami a döntési helyzet szempontjából egy „kimeneteli" változónak tekinthető (később több kimeneteli változót is megengedünk).

Az $\mathbf{X}=[X_1,..,X_n]$diszkrét valószínűségi változók ismeretében egy kézenfekvő lehetőségnek tűnhet, ha az $\mathbf{X}$ együttes valószínűségi változót egyetlen változónak tekintjük és eloszlását a Dirichlet eloszlás felhasználásával definiáljuk.

2.1 Definíció. Dirichlet eloszlás. Egy folytonos $\boldsymbol{\Theta}=[\Theta_1,..,\Theta_k]$ valószínűségi vektor változó k-dimenziós, $\boldsymbol{\alpha}=[\alpha_1,..,\alpha_{k+1}]$ paraméterű $Di_k(\boldsymbol{\Theta}|\boldsymbol{\alpha})$ Dirichlet eloszlást követ ($\alpha_i>0$, i=1,..,k+1), ha $0<\theta_i<1$ és $\theta_1+..+\theta_k<1$, továbbá sűrűségfüggvénye

$$Di_k(\boldsymbol{\theta}|\boldsymbol{\alpha}) \propto \theta^{\alpha_1-1}... \theta^{\alpha_k-1}(1-\Sigma_{i=1..k}\theta_i)^{\alpha_{k+1}-1}.$$

A Dirichlet eloszlásban szereplő Θ_i várható értéke $E[\Theta_i]=\alpha_i/(\alpha_1,.., \alpha_{k+1})$. Ez beleilleszkedik abba az elterjedt értelmezésbe, hogy Z diszkrét, k értékű valószínűségi változó eloszlása definiálható $\boldsymbol{\theta}$-val, $P(Z=i)=\theta_i$ megfeleltetéssel és $Di_k(\boldsymbol{\Theta}|\boldsymbol{\alpha})$ feltételezésével, ahol az α_i paraméterek felfoghatók a „i" érték „számlálójának", abszolút gyakoriságának. Mivel ekkor a θ_i valószínűségi változók, mint paraméterek a P(Z=i) valószínűségeket definiálják, az $\boldsymbol{\theta}$ eloszlását definiáló α_i paramétereket hiperparamétereknek nevezzük.

Visszatérve a példánkra, ekkor az $\mathbf{x}$ elemi események eloszlása is ennek megfelelően definiálható egy $P(\mathbf{X}=i)=\theta_i$ megfeleltetéssel és az α_i paraméterek megadásával i=1.. N esetén, továbbá a $Di_N(\boldsymbol{\theta}|\boldsymbol{\alpha})$ eloszlás feltételezésével (ekkor α_i hiperparaméter az adott elemi esemény korábbi előfordulásainak

száma). Ezzel $P(\mathbf{X}|\boldsymbol{\theta})$ parametrikus eloszlást és a kapcsolódó $Q(\boldsymbol{\Theta})$ eloszlást elméletileg definiálhattuk. A gyakorlatban azonban az elemi események nagy száma, amely a változók számában exponenciális, két okból is problémát jelent:

1. A szakértők nem ismerik az elemi konfigurációk valószínűségét. Gondoljuk meg egy orvosi döntési helyzet esetén egy konfigurációban az összes döntési helyzetben releváns betegség és tünet szerepel. A nagy számú konfigurációk valószínűségét automatikusan becsülni is nehéz, mivel igen sok esetre/mintára lenne szükség.
2. Az érvelés számítási komplexitása. Az elemi konfigurációk száma az értékészletük (ri) szorzata, így a változók számában exponenciális. A legegyszerűbb kérdés is, például "mennyi P(Xj=x)", az összes kompatibilis elemi konfiguráció összegzését jelenti.

A módszer naív felhasználása tehát nem lehetséges, keressünk ezért olyan tulajdonságokat, amik kezelhetővé teszik a feladatot. Vizsgáljuk meg a következő alaphelyzetet.

1.4.1. A naív Bayes model

Tételezzük fel, hogy a változók halmaza tartalmaz egy hipotézis (Y) változót és megfigyeléseket ($X_1,..X_n$). Elterjedt más elnevezések az ok, modell, diagnózis, illetve okozat, bizonyíték, megfigyelés, tünet, szimptóma. A megfigyelések csoportot tekinthetjük különböző típusú megfigyeléseknek (például tüneteknek egy orvosi problémában) vagy azonos típusú, de eltérő idejű megfigyelések szekvenciáinak.

Ideiglenesen feltesszük, hogy célunk a $P(Y, X_1,..X_n)$ eloszlás modellezése, azaz a bayesi paradigmától eltérve egyetlen valószínűségi modellt szeretnénk konstruálni. Feltételezzük továbbá a $P(Y, X_1,..X_n)$ eloszlásról, hogy az X_i megfigyelések teljesen függetlenek ha ismert az Y hipotézis, azaz

$$P(X_j| X_{i_1},..,X_{i_m},Y)=P(X_j|Y) \text{ bármely } X_{i_1},..,X_{i_m} \text{ részhalmazra.}$$

Ha ekkor a $P(X_1|Y),...,P(X_n|Y)$ feltételes eloszlások, lokális valószínűségi függések ismertek, akkor a hipotézis eloszlását adott $X_{i_1},..,X_{i_m}$ megfigyelések esetén a következőképpen adódik. A Bayes szabály alkalmazásával

$$P(Y| X_{i_1},..,X_{i_m}) = P(X_{i_1},..,X_{i_m} |Y)\, P(Y) / P(X_{i_1},..,X_{i_m}).$$

Mivel a feltételezett függetlenség alapján fenáll, hogy

$$P(X_{i_1},..,X_{i_m} |Y) = P(X_{i_1} |Y)... P(X_{i_m} |Y).$$

A hipotézis feltételes eloszlása az $X_{i_1},..,X_{i_m}$ megfigyelések feltételekkel a következőképpen egyszerűsíthető

$$P(Y| X_{i_1},..,X_{i_m}) = P(X_{i_1} |Y)... P(X_{i_m} |Y)\, P(Y) / P(X_{i_1},..,X_{i_m}).$$

Ez tovább alakítható, ha figyelembe vesszük, hogy a hipotézisek kölcsönösen kizáróak és teljesek, azaz

$$P(X_{i_1},..,X_{i_m}) = \Sigma_{y\in \text{Range}(Y)}\, P(X_{i_1},..,X_{i_m} |y)\, P(y)$$

Végül mivel egy adott $x_{i_1},..,x_{i_m}$ megfigyelés esetén $P(x_{i_1},..,x_{i_m})$ csupán egy normalizációs konstans, írhatjuk, hogy

$$P(Y|\ X_{i_1},..,X_{i_m}) \propto P(X_{i_1}\ |Y)...\ P(X_{i_m}\ |Y)\ P(Y)$$

azaz egy hipotézis valószínűsége az a priori valószínűségének és a szimptómák feltéles valószínűségének a szorzata. Ez az eredmény azt mutatja, hogy ha feltétevésünk a döntesi helyzetre, szakterületre vonatkozó feltételes függetlenségekről helyes és ismerjük a $P(X_i|Y)$ eloszlásokat, illetve a P(Y) a priori eloszlást, akkor a modell mérete és a számítás a változók számában nem exponenciális, hanem lineáris. Az algebrai eredmény egy intuitív vizuális kifejezését a 3. Ábra mutatja.

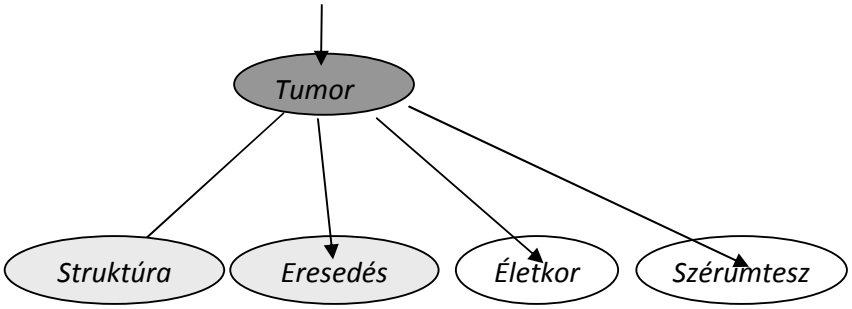


Figure 13 Egy "Naív" bayes háló.

Az ábra intuitíve jelzi, hogy egyedül a hipotézis változó befolyásolja direkt módon a megfigyelhető szimptómákat, azok között direkt interakció nem lehetséges, továbbá hogy a nyílak által jelzett függéseket kell valószínűségek megadásával definiálni. Mint látni fogjuk ennek az értelmezésnek a formalizálása és általánosítása eredményezi majd a valószínűségi hálózatok elméletet.

1.4.2. Függetlenségek reprezentálása és felhasználása

Az előző speciális példa jól demonstrálja a függetlenségek központi szerepét egy valószínűségi, bayesi modell dekomponálásában. A dekomponálás pedig a modellépítés, statisztikai adaptáció és a következtetés hatékonyságát is meghatározzák, ami a gyakorlatban sokszor a megvalósíthatóságot jelenti. A dekomponálás nyújtotta előnyőkről az általános esetekben sem kell lemondani. Tegyük fel, hogy megkonstruáltunk egy $P(X_1,..X_n)$ valószínűségi modellt és továbbra is eltekintünk a a bayesi paradigmához tartozó „több modell-a priori eloszlás" elvtől. Válasszunk egy önkényes sorrendezést a valószínűségi változók számára és tekintsük a következő, minden esetben lehetséges dekompozíciót:

$$P(X_1,..,X_n) = P(X_1|\ X_2..,X_n)P(X_2|\ X_3..,X_n)...P(X_n)=\Pi_{i=1..n}\ P(X_i|\ X_{i+1}..,X_n)$$

Valós modelleknél a feltételes függetlenségek általában gyakoriak, így a $P(X_i|\ X_{i+1}..,X_n)$ feltételes valószínűségek $P(X_i|X_{i_1},..,X_{i_m})$-re cserélhetők, ahol $[X_{i_1},..,X_{i_m}]\subseteq Pa(X_i)=_{def}[\ X_{i+1}..,X_n\ ]$. Mivel a $Pa(X_i)$,i=1..n „szülői" halmazok mérete gyakran egy adott korlát alatt marad, ez a dekomponálás a gyakorlatban igen hasznosnak bizonyulhat. Például, ha egy adott sorrendezésnél $\#(\ Pa(X_i)) <k$,i=1..n, azaz a változók csak maximum k számú másiktól függnek, akkor az eloszlás reprezentálása, a táblák mérete, nagyságrendekkel kisebb lehet. Bináris változók esetén ez 2^n, illetve $n*2^k$ nagyságrendű model méretet jelent, ez n=20 és k=5 esetén 1 Gigabyte nagyságrendet, illetve 1 Kilobyte nagyságrendet

jelent. Ez a feltételes függetlenségeken alapuló dekompozíció hatékonyabb számítást, például marginalizációt is lehetővé tesz.

A feltételes függetlenségek hatékony felhasználása azonban csak akkor lehetséges, ha hatékonyan reprezentáljuk és kezeljük őket, hiszen hiába adottak implicite a definiált eloszlással, a definíción alapuló ellenőrzés nem praktikus. Explicit reprezentációjuk és modellezésük szükséges. Vezessük be ezért a következő jelölést:

2.2 Definíció. Feltételes függetlenség. I(**V,Z,W)** jelölje azt, hogy **Z**-t ismerve **V** és **W** független, ahol **V,Y,W**$\subseteq$ **X**.

2.3 Definíció. Függőségi modellnek nevezzük ilyen állítások együttesét: $M = [I_1,...,I_L]$.

Érdekes, hogy „függetlenség” nemcsak a valószínűségszámítás szerinti értelmezéssel, hanem például adatbázisok esetén is megfogalmazható: I(**V,Z,W**): **Z**-t ismerve **V** nem jelent korlátozást **W** értékeire.

A bevezetett jelöléssel a feltételes függetlenségeket, mint logikai állításokat kezelhetjük. Ennek bevezetése és hasznosságának felismerése – a dekompozícióban és vizualizációban játszott központi szerepének a felismerése - a valószínűségszámítás szempontjából, döntő fontosságú lépés volt, ami lehetővé tette a bizonytalanság hatékony kezelését, valószínűségi alapú szakértői és döntéstámogató rendszerek széles körű alkalmazhatóságát. A célul kitűzött valószínűségi modell hatékony megkonstruálásához ezért egy M függőségi modellt konstruálnunk meg, ami majd remélhetőleg olyan dekomponálást definiál, ami lehetővé teszi a modell hatékony felparaméterezését, kvantifikációt. A feltételes függetlenségek logikai modelljének további előnye hogy logikailag is lehet rajta érvelni, azaz definiálható egy érvényes és hatékony következtetési módszer. Egy másik előny, hogy a feltételes függetlenségek önálló kezelését az emberi szakértőhöz való igazodás is indokolja. Egyrészt az emberi szakértő ezt a kvalitatív struktúrát akkor is ismeri és használja, hogyha a kvantitatív függésekre nicsenek becslései. Ez általában azt is jelenti, hogy egy szakértő által adott kvalitatív struktúra megbízhatóbb, mint az általa adott valószínűségek.

Az M függőségi modellre természetesen kényszereket jelent, ha egy adott döntési helyzethez tartozó P valószínűségi eloszlás modellezésére hozták létre. Azaz az I állítások rendszere nem tetszőleges, például

$$I(\mathbf{V,Z,W}) \in M \Leftrightarrow I(\mathbf{W,Z,V}) \in M.$$

Az, hogy M egy eloszlásra vonatkozik, olyan szigorú megszorításokat jelent, hogy M igen sok esetben "vizuálisan", gráfokkal is reprezentálható. Az összefüggések és függetlenségek gráffal történő reprezentálhatóságának formális bizonyítása egy igen régi elvárást erősít meg az emberi gondolkodás természetével kapcsolatban. Az asszociációkon, "gráf-struktúrákban terjedő aktivációkon" alapuló pszichológiai modellek elfogadottak és így alátámasztják azt az elvárást, hogy az emberi szakértők könnyen tudnak ehhez a formalizmushoz illeszkedni. Továbbá érv a gráfok alkalmazása mellett, hogy vizuálisak. Bemutatjuk továbbá, hogy a gráfok a függetlenségek reprezentálása mellett , hogyan

használhatók fel a valószínűségi eloszlás kvantitatív leírásához mind a nem bayesi és bayesi paradigmában (egyetlen **θ** felparaméterezés, illetve a $Q(\boldsymbol{\Theta})$a priori eloszlás definiálásához). A valószínűségi eloszlásoknak, függési modelleknek, az eloszlások dekomponálhatóságának, illetve a függési modellek gráfokkal történő reprezentálásának kölcsönös összefüggését a 4. Ábra mutatja.

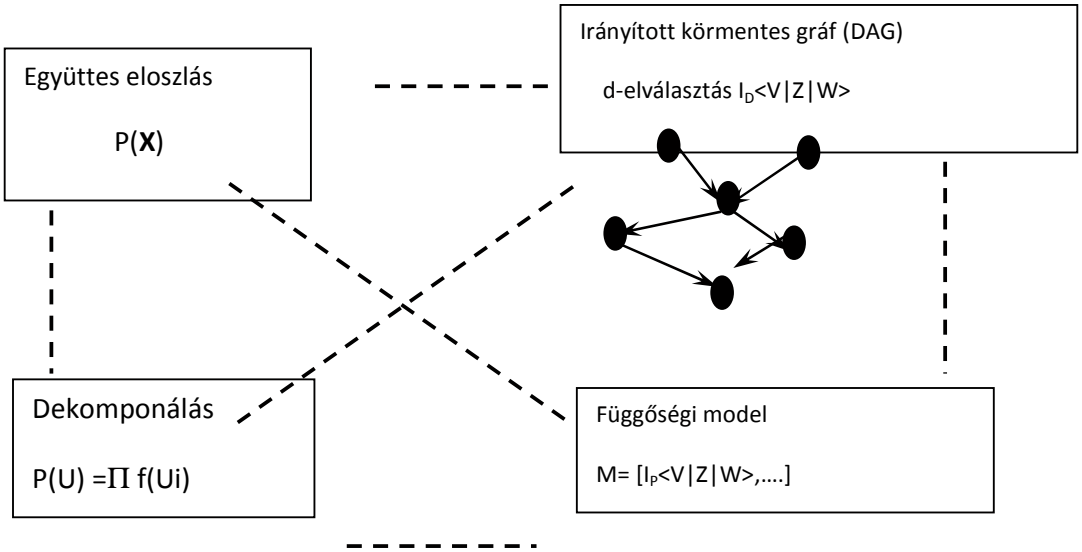


Figure 14 Eloszlások, gráfok, dekomponálások és függési modellek kapcsolata.

1.4.3. Markov hálók

Adott függőségi modell M=[I_1,...,I_L] reprezentálására egy G=(**P**,**E**) irányítatlan gráfot használunk, ahol **P** a (csomó)pontok halmaza és **E** **P**-be tartozó pontpárok, élek halmaza. Foglaljuk össze a definíciókat.

2.3 Definíció. Jelölje <**V**|**Z**|**W**>$_G$, ha minden **V** és **W** közötti úton létezik egy csomópont ami eleme **Z**-nek. Ezzel a feltételes függetlenséget próbáljuk majd reprezentálni és megfelel annak az intuíciónak, hogy például a tünetek függetlenné válnak, ha közös okuk egy betegség ismert.

2.4 Definíció. G-t M-hez tartozó *függetlenségi-diagrammnak* (*I-diagrammnak*) nevezzük, ha <**V**|**Z**|**W**>$_G$⇒I(**V,Z,W**). Ekkor a gráfban megállapítható függetlenség biztosan fenáll, de nem biztos, hogy reprezentálva van. Ez azt jelenti, hogy nem minden függetlenség van reprezentálva. Egy teljes gráf természetesen I-diagramm, mivel nem mond ki függetlenséget.

2.4 Definíció. G-t M-hez tartozó *perfekt-diagrammnak* nevezzük, ha tökéletesen reprezentálja tudásunkat a függetlenségekről, azaz <**V**|**Z**|**W**>$_G$⇔I(**V,Z,W**). Eloszlásokra vonatkozó M-eknél ilyen nem mindig lehetséges. Gondoljunk arra, hogy egy P eloszlásnál lehetséges, hogy I(**V,Z,W**) és **¬**I(**V,Z',Y**)}, ahol **Z⊂Z'**, azaz további információk addig feltételesen független változókat függővé tesznek. Például, ha két érmével függetlenül dobunk és XOR-olt eredményüket tekintjük egy harmadik valószínűségi váltózónak. Ekkor ha az eredményről nem tudunk semmit, a két változó feltételesen független. Ha azonban a harmadik változó ismert, akkor egy dobás függvénye a másik dobásnak. Azonban formalizálható, hogy milyen megkötéseknek eleget tevő M-hez létezik tökéletes reprezentáció.

2.5 Definíció. G-t M-hez tartozó *Markov-hálónak* nevezzük, ha I-diagramm és bármely élének törlésére már nem lenne az.

2.6 Definíció. G egy p∈ P csomópontjához tartozó **B**(p) *Markov-határnak* nevezzük azt a minimális elemszámú **B**⊆**X**, v∉ **B** részhalmazt, ha I(p**,B,V-B-**p), azaz ha **B** mintegy „elválasztja” p-t a hálózat többi részétől.

Ezekután egy tétel formájában is kimondható hogyan konstruáljuk meg egy szigoróan pozitív P eloszlás Markov-hálóját[Pearl]:

2.1 Propozíció. Ha G(P,E) esetén [p_i, p_j]∈ E⇔p_j∈ **B**(p_i), ahol B(p) egy szigoróan pozitív P eloszlás által definiált, akkor G a P eloszlás által definiált M függési modellhez tartozó Markov-háló.

Egy adott P eloszláshoz tartozó gráfreprezentációhoz a következő tétel alapján lehet hozzárendelni az eloszlás kvantitatív jellemzőit, a valószínűségeket (a bayesi kiterjesztést a következő, jobban elterjedt gráf reprezentáció kapcsán ismertetjük majd).

2.2 Propozíció. P valószínűségi eloszlás *dekomponálható* G gráf alapján, ha G I-diagrammja P-nek és G "háromszögesített", azaz minden négynél hosszabb köre tartalmaz egy nem szomszédos csomópontot összekötő élt. Ebben az esetben P eloszlás előáll G gráf klikkjeihez tartozó eloszlások

szorzataként osztva a klikkek metszeteihez tartozó eloszlásokkal (Klikknek nevezzük egy gráf csomópontjainak egy részhalmazát, ami teljesen összekötött és maximális).

Az, hogy a gráf részenként reprezentálja a teljes P eloszlást két szempont miatt lényeges:

- a szakértőtől és adatokból a részeit elég megbecsülni,
- hatékony számítást tesz lehetővé, hiszen P-vel kapcsolatos számítások is dekomponálódnak a klikkekre, így a számítás, például marginalizáció, csak a klikkméretekben lehet exponenciális.

1.4.4. Bayes hálók: reprezentáció

A valószínűségi megközelítésben bizonytalan tudásunkat sztochasztikus változók együttes eloszlásával reprezentáljuk. A szisztematikus struktúrával nem rendelkező tárgyterületek esetén (szemben pl. a kép- és hangfeldolgozással) az ilyen eloszlások modellezésére használt elsődleges eszközt ma a Bayes-hálók jelentik. Ezekben egy irányított körmentes gráfban (DAG – directed acylic graph) reprezentálják a változókat és a köztük lévő összefüggéseket: minden csomópont egy-egy változót jelöl, és minden csomóponthoz tartozik egy lokális feltételes valószínűségi modell, amely leírja a változó függését a szüleitől (a pontos definíciót a következő fejezet tartalmazza).

Mint reprezentációs eszköz, egy Bayes-háló háromféle értelmezést kaphat, ezek a felsorolás sorrendjében egyre erősebb modellezési, értelmezési lehetőséggel bírnak:

- Tekinthető egyszerűen az együttes eloszlás egy hatékony ábrázolásának, hisz a csomópontonkénti feltételes valószínűségi modellekre való faktorizálással a felhasznált paraméterek száma jelentősen csökken.

- Egy adott struktúra meghatározza, hogy az ábrázolt eloszlásban milyen feltételes függések és függetlenségek lehetnek, azaz az élek tekinthetők a közvetlen valószínűségi összefüggések reprezentációjának, míg a teljes gráf a reprezentált eloszlás függési térképének.

- Az előzőnél is erősebb a kauzális értelmezés, amelyben minden élt az érintett két csomópont közötti ok-okozati összefüggésként értelmezzük.

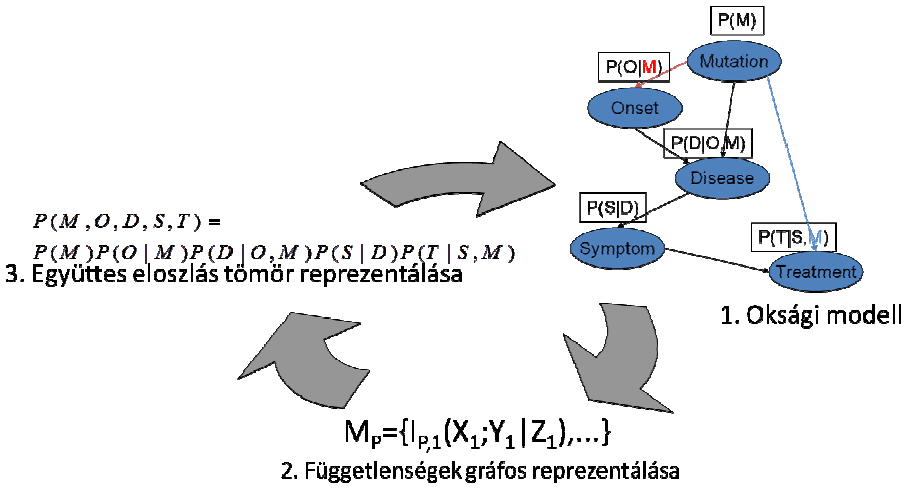


Figure 15 A Bayes háló értélmezésének három szintje.

1.4.5. A valószínűségi definíció: szintaxis és szemantika

Egy Bayes-háló struktúrája és a reprezentálni kívánt eloszlás közti kapcsolatot az alábbi négy feltételre alapozhatjuk, melyekről belátható (Cowel1999), hogy ekvivalensek.

- A $P(X_1, ..X_n)$ eloszlás *faktorizálható* a G DAG szerint, ha:

$$P(X_1,...X_n) = \prod P(X_i \mid Pa(X_i)),$$

  ahol $Pa(X_i)$ az X_i csomópont szülői halmaza.

- A $P(X_1, ..X_n)$ eloszlásra teljesül a *sorrendi Markov-feltétel* G szerint, ha

$$\forall i = 1..n : I(X_{\pi(i)} \mid Pa(X_{\pi(i)}) \mid \{X_{\pi(j)} \mid j < i\} \setminus Pa(X_{\pi(i)}))_P,$$

  ahol az I(X|Y|Z) reláció az X feltételes függetlenségét jelenti a Z-től Y feltétellel, π pedig a struktúra egy topologikus rendezése

- A $P(X_1, ..X_n)$ eloszlásra teljesül a *lokális (szülői) Markov-feltétel* G szerint, ha bármely változó független nem-leszármazottaitól, feltéve szüleit.

- A $P(X_1, ..X_n)$ eloszlásra teljesül a *globális Markov-feltétel* G szerint, ha

$$\forall x, y, z \subseteq \{X_i\} : I(x \mid z \mid y)_G \Rightarrow I(x \mid z \mid y)_P,$$

azaz ha z d-szeparálja x-et y-tól a G gráfban, akkor x független y-tól. Itt $I()_G$ a d-szeparálást jelöli, hogy 'z' d-szeparálja 'x'-et és 'y'-t a 'G' gráfban ($x.y.z \subseteq V(G)$), ha minden 'x' és 'y' között menő irányítatlan 'p' utat blokkol, azaz, ha (1) 'p' tartalmazza 'z' egy elemét nem összefutó élekkel, vagy (2) 'p' tartalmaz egy 'n'csomópontot összefutó élekkel, hogy 'z' nem tartalmazza sem 'n'-t, sem valamelyik leszármazottját

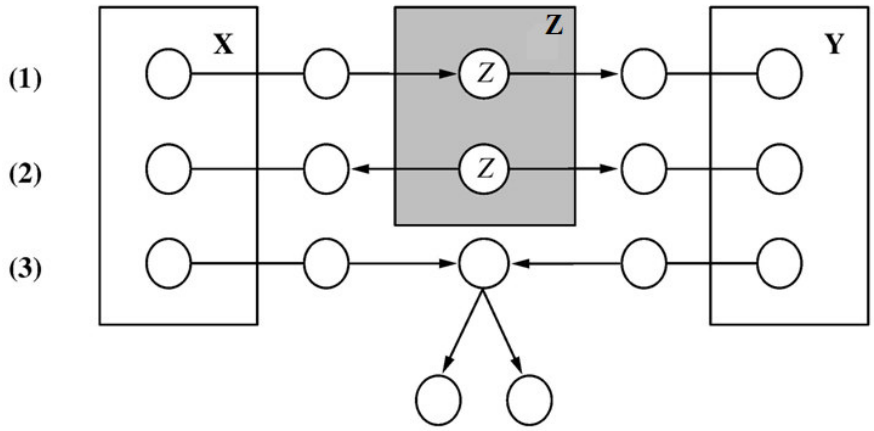


Figure 16 A d-szeparáció (irányított elválasztás) esetei.

Egy elfogadott definíció a Markov-feltételek által adott függőségi rendszer tulajdonságaira épít [15]:

A 'G' irányított körmentes gráf a 'P(U)' eloszlás Bayes-hálója *(U az összes változó halmaza), akkor és csak akkor, ha minden* $u \in U$ *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 teljesülne).*

Míg ez a definíció egyértelműen a valószínűségi függetlenségek rendszerének reprezentációjaként tekint a Bayes-hálóra, addig a mérnöki gyakorlatban közkedvelt az alábbi, praktikus meghatározás:

Az 'U' valószínűségiváltozó-halmaz Bayes-hálója *a (G, ϑ) páros,ha 'G' irányított körmentes gráf, amelyben a csomópontok jelképezik U elemeit, ϑ pedig csomópontokhoz tartozó 'P(X|Pa(X))' feltételes eloszlásokat leíró numerikus paraméterek összessége.*

Fontos megjegyezni, hogy a definiált modellosztályban a lehetséges struktúrák száma a csomópontok számában szuperexponenciális, ez pedig pl. a később tárgyalandó tanulás komplexitását is befolyásolja.

Bár egy Bayes-háló egyaránt tartalmazhat diszkrét és folytonos változókat is, mi a továbbiakban kizárólag diszkrét, véges változókkal foglalkozunk, feltéve továbbá, hogy minden lokális feltételes valószínűségi modell a multinomiális eloszlásokhoz tartozik, így a paraméterek ún. feltételes valószínűségi táblák (FVT-k) elemei .

Egy adott Bayes-háló struktúrája meghatározza, hogy az milyen függéseket írhat le (pl. külön komponensekben lévő változók közt nem lehet függés), azonban különböző struktúrákhoz is tartozhat azonos implikált függési rendszer. Ha két struktúrából ugyanazok a feltételes függetlenségek olvashatók ki, a két gráfot *megfigyelés-ekvivalensnek* mondjuk. Belátható [14, 15], hogy két gráf akkor és csak akkor megfigyelés-ekvivalens, ha irányítás nélküli vázuk, illetve v-struktúráik (az A→B←C típusú részgráfok, úgy, hogy A és C közt nincs él) megegyeznek.

A megfigyelési ekvivalencia segítségével a struktúrákat diszjunkt osztályokba sorolhatjuk. Minden ilyen ekvivalencia osztályt egy ún. esszenciális gráffal reprezentálhatunk. Az esszenciális gráf váza

megegyezik az osztályba tartozó gráfokéval, és csak azok az élei irányítottak, amelyek iránya mindegyik gráfban megegyezik (ezek az ún. kényszerített –compelled– élek).

1.4.6. Kauzális definíció

Az előző, tisztán valószínűségi definíciók bevezetése után formálisan könnyen áttérhetünk a Bayes-hálók kauzális értelmezésére: *egy (G, ϑ) páros* kauzális Bayes-hálója *a P(U) eloszlásnak, ha egyrészt a tárgyterület valószínűségi modellje az előző értelmezések szerint, továbbá minden él közvetlen ok-okozati viszonyt jelképez.*

Hasonlóan, itt is létezik egy Markov-feltétel: *egy P(U) eloszlás és egy kauzális relációkat leíró G gráf teljesíti a* kauzális Markov-feltételt*, ha G és P(U) teljesíti a lokális Markov-feltételt.*

A Markov-feltétel teljesülése biztosítja, hogy minden (kazuális) függés kiolvasható a gráfból, a másik irányhoz, ahhoz tehát, hogy minden a gráfból kiolvasott függés teljesüljön az eloszlásban, annak stabilnak kell lennie. *Egy P(U) eloszlás* stabil*, ha létezik olyan G gráf, hogy P(U)-ban pontosan a G-ből d-szeparációval kiolvasható függések és függetlenségek teljesülnek benne* (pl. megfelelő paraméterezés mellett előfordulhat, hogy egy A→B→C struktúrában A és C függetlenek).

A fenti kauzális definíció a modell és a tárgyterület összefüggéseinek értelmezését illetően igen erős, a megfigyelési adatok statisztika elemzésének kereteit meghaladó eszközt szolgáltat. Alkalmazásakor figyelembe kell vennünk, milyen nem kauzális kapcsolatok okozhatnak valószínűségi összefüggést két változó között, azaz milyen korlátai vannak a kauzális értelmezésnek. Ilyenek lehetnek pl.:

- Zavaró változók: a két változó közti függést okozhatja egy közös ős (az ún. zavaró változó) is.
- Kiválasztási bias: a változók közti függés lehet az adatgyűjtési mód következménye is (pl. ha egy orvosi adatbázisba csak a komolyabb megfázással kezelt betegek kerülnek be, akkor a láz és torokfájás között direkt függést figyelhetünk meg).
- Az ős-ok, leszármazott-okozat megfeleltetés és a DAG gráfstruktúra kizárja a mechanizmusokban lévő visszacsatolások (ciklikusságok), illetve az oda-vissza ható okozatiság lehetőségét.
- A modelltér maga (azaz, hogy milyen változók szerepelnek, illetve azok milyen értékkészlettel rendelkeznek) szintén befolyásolja, hogy milyen direkt függések jelennek meg (azaz a gráf struktúrát).

1.4.7. Bayes-hálók és a tudásmérnökség

A föntebb definiált Bayes-háló a tudásmérnökség eszközeként jelent meg a 80-as években, konstruálása jellemzően a szakértőktől származó adatokból történt manuálisan. A kézi konstruálás még napjainkban is jelentős súlyt képvisel a Bayes-hálók alkalmazásában, másrészt ahol az adathoz viszonyítva jelentős a priori tudás áll rendelkezésre, ott a Bayes-hálók tudásmérnöki alkalmazása a bayesi keretrendszer alkalmazásának egy kezdeti fázisát jeleni, nevezetesen a prior konstruálást.

A tudásmérnökség metodikájára nagy hatással volt a nagy mennyiségű elektronikus tárgyterületi információ megjelenése, a megfelelő mennyiségű statisztikai adat elérhetősége, valamint a Bayes-

statisztikai alapú gépi tanulási módszerek elterjedése. A felépített tudásbázissal szemben követelményként jelent meg a bayesi módszerek alkalmazásakor, hogy támogassa a priorok konstruálását, hiszen a valószínűségekkel leírt a priori tudás és a rendelkezésre álló adatok bayesi frissítéssel történő kombinációja szolgáltatja a végső tudásmodellt. Mindemellett fontos, hogy a tudásbázis segítse komplex, akár szabad szöveges háttérismereteket is tartalmazó valószínűségi állítások megfogalmazását, valamint tegye lehetővé a szakértőktől származó szubjektív információ tárolását, mely releváns a bayesi a priori tudásmodell megalkotásánál.

Egy tudásbázis megépítéséhez olyan környezetben, ahol rendelkezésre áll elektronikus tárgyterületi tudás, elegendő statisztikai adat, valamint a megfelelő bayesi módszerek, az alábbi lépések szükségesek (amelyekből a specifikusokat részletezzük):

1. *Célok, alkalmazási terület és modellezési szintek identifikációja*
   T*erminológia és ontológia elfogadása.*

2. *Nem rendszerezett tudás begyűjtése*
   Ehhez a lépéshez tartozik az összes releváns elektronikus és egyéb szövegalapú információforrás feldolgozása, ami magába foglalja az a priori információ kinyerését különféle szövegbányászati metódusok alkalmazásával, mint például az általunk kifejlesztett módszer, amit a 4.8. fejezetben mutatunk be.

3. *Struktúra kinyerése*
   A G DAG struktúrák feletti p(G) priorok konstruálása, melyek egyesítik a szakértők által megadott információkat az elektronikus forrásokból kinyert információkkal. A p(G) a priori eloszlást többnyire normalizálatlan formában lehet előállítani: pl. egy adott referencia struktúrától való eltérés alapján

   $$P(G|\xi) \propto \kappa^{\delta};\ 0<\kappa<1,$$

   ahol δ a referenciától való tetszőlegesen definiált strukturális tulajdonságokbeli eltéréseknek a száma.

4. *Paraméter és hiperparaméter kinyerése*
   A valószínűségi paraméterek számos módon nyerhetők: adatbázisok, szakirodalom vagy szakértők szubjektív véleménye alapján. A $p(\theta|G)$ paraméter prior specifikációja az általunk vizsgált diszkrét, véges esetben egy egyszerű módszerrel megtehető, ha feltehetjük az egyes változókhoz és szülői értékkonfigurációkhoz tartozó paraméterek függetlenségét

   $$P(\theta|\ G_0,\xi) \propto \prod_{i=1\ldots n}\prod_{j=1\ldots qi} P(\theta_{i,j}|\ G_0,\xi).$$

   Egy szinte kizárólagosan használt eloszláscsalád az adott változó, adott szülői értékkonfigurációjához tartozó $P(\theta_{i,j}|\ G_0,\xi)$ megadására a Dirichlet eloszlás $Dir(\theta_{i,j}|\ \alpha_{ij,i}\ \xi)$, ahol az $\alpha_{ij,i}$ hiperparaméter jelentése a paraméterhez tartozó szülői értékkonfiguráció korábban megfigyelt eseteinek számait jelenti (Cowel1999). Megmutatható, hogy a Dirichlet család az egyetlen lehetséges választás, ha az ugyanazon megfigyelési ekvivalencia osztályba tartozó G struktúrákhoz ekvivalens priorokat szeretnénk megadni, ami kauzális modellezésnél nem szükségszerű (Heckerman1995a). További feltevések mellett az is bizonyitható, hogy az összes struktúrához konzisztens $p(\theta|G)$ definíciója ekvivalens egy teljes modellhez tartozó

pontparametrizációnak és egyetlen korábban megfigyelt összesetszámot jelentő hiperparaméternek a megadásával. E kettő együtt valójában egy a priori adathalmazt definiál, ami korábban megfigyelt eseteket tartalmazza, így az összesetszámot virtuális vagy a priori mintaszámnak nevezünk.

5. *Érzékenységi analízis, verifikáció és validáció*
   A modellek posteriorjának vizsgálata magába foglalja egyrészt az a priori eloszlásokra való érzékenység vizsgálatát (ami különösen fontos a több szakértőt és tudásbázist is magában foglaló automatizáltan származtatott prioroknál), másrészt referencia priorokkal való összehasonlítást. Mindkét esetben gyakran szükséges a modellosztály komplexitása miatt, egyrészt hogy modell jegyeket használjunk, másrészt hogy MAP modellre alapozzuk a vizsgálatot.

Mint ahogy az látható, a tudásbázis építése a bayesi modellkiértékeléssel zárul. A kiértékelés tartalmazza az adat és a modell kompatibilitásának vizsgálatát és az a posteriori valószínűségek vizsgálatát, azaz a tudásmérnöki folyamat lényege az a priori modell konstruálása a későbbi tanulási folyamat számára.

1.4.8. Reprezentációk teljessége, hűsége, esetlegessége

Mindegyik módszer esetén más és más függetlenséget nem lehet reprezentálni, ám ez végülis csak hatékonyság vesztés. A gráf reprezentációk egyik nyilvánvaló hiányossága, hogy nem képesek ábrázolni azt a jelenséget, hogy a valószínűségszámítás szerint a változók páronkénti függetlenségéből nem következik a változók halmazainak függetlensége.

Korábban említettük, hogy az emberi szakértők a függőséget intuitíve képesek kezelni. Több alternatíva esetén használhatunk akár több modellcsaládot is, bevezetve egy újabb $P(M_i)$ szintet a bayesi paradigma szerint, ami az apriori véleményt fejezi ki egyes M_i modellcsaládokra nézve. Egy egyszerűbb módszer, ha továbbra is egyetlen modellcsaláddal dolgozunk, azonban több függést veszünk be, úgy gondolva hogy, legfeljebb kevésbé lesz hatékony a rendszer. Sajnos ez a módszer is alapos mérlegelést igényel, hiszen a következtetési algoritmusok komplexitása erősen függ a gráf típusától, így egyetlen él is döntő lehet a számítási komplexitásra nézve. Másik probléma, hogy nem tudjuk becsülni az óvatosságból betett éleket, így lehet, hogy megéri függetlennek feltételezni, mert bár ebben az esetben egy összefügést elhanyagolunk, de a kisebb függést viszont már jobban tudjuk becsülni.

A bayesi paradigmától eltérve vizsgáljuk meg most a Bayes hálók, mint reprezentációs technika viszonyát egy P(X) eloszlással.

A fentebbi definiciókból az alábbi algoritmus vezethető le egy P eloszláshoz tartozó Bayes-háló megkonstruálásához (Verma):

2.3 Propozíció. Válasszunk önkényesen egy $X_1..,X_n$ sorrendet az eloszlást alkotó $X_i \in \mathbf{X}$ változókhoz, majd legyenek G gráfban X_i-hez tartozó csomópont szülei azon minimális méretű $Pa(X_i) \subseteq \mathbf{X}$ részhalmaz elemei, amelyekhez tartozó valószínűségi változókra fennáll, hogy

$P(X_i|X_1,...,X_{i-1}) = P(X_i|Pa(X_i))$, ahol $Pa(i) \subseteq X_1,...,X_{i-1}$. Ekkor G gráf egy P eloszláshoz tartozó Bayes háló.

A fentebb formálisan ismertetett módszer szerint, informálisan fogalmazva

- a csomópontok változóknak felelnek meg,
- az élek "direkt függést" jelölnek,
- minden csomóponthoz tartozik egy feltételes valószínűségi tábla, a szülőkkel mint feltételekkel.

Ezzel a módszerrel minden eloszlás reprezentálható, lásd 2.3 Propozíció, illetve ha adott egy Bayes háló, akkor az egy eloszlást definiál:

- ha irányított körmentes gráf, akkor létezik a csomópontoknak egy topológikus sorrendje:$(i<j) \Rightarrow$ (X_j nem szerepel X_i feltételeként a feltételes valószínűségeken alapuló dekompozícióban),
- a sorrend alapján lehetséges feltételes valószínűségek szorzatára való bontás, ahol a feltételes valószínűségi táblák ismertek.

További előny, hogy mivel nincs redundancia, nem lehet ellentmondást létrehozni. A struktúra és a paraméterek egy módon azonban eltérhetnek, mivel a paraméterek (eloszlás) szerinti függetlenséget nem feltétlenül reprezentálja a struktúra (gráf).

A módszer egyik kritikus pontja a változók sorrendjének megválasztása, ami a változók között egyfajta ok-okozat viszonyt, kauzalitást definiál, a feltételes valószínűségek „irányát". Továbbá a sorrend befolyásolja a reprezentációban szereplő feltételes valószínűségek megválasztását is, ami azt is jelentheti, hogy ugyanazon eloszlás esetén bizonyos sorrendnél egy igen ritka gráf, más sorrendnél igen sok élű gráf jöhet ki. Gondoljuk például végig a naív Bayes model példáját $(Y,X_1,..,X_n)$ és $(X_1,..,X_n ,Y)$ sorrend esetén. Ez a tulajdonsága a reprezentációnak furcsának tűnhet, hiszen a Bayes hálók egy másik neve is - "Causality network" - az okok és okozatok kapcsolatának leírására utal. és a megszokott emberi okoskodás szerint az okok és okozatok rendszere objektív szükségszerűség, ami az idő, entrópia és kapcsolódó fogalmakon alapu. A kauzalitás vonatkozó tárgyalására megtalálható például J. Pearl műveiben [14, 15], itt csupán két álláspontot említünk.

- Objektív szemlélet. Egy adott területhez tartozó kauzalitási modell a terület objektív és szükségszerű következményel.
- Szubjektív szemlélet. Egy adott területhez tartozó kauzalitási modellt egyrészt a terület objektív tulajdonságai, másrészt számítási kényszerek, szubjektív preferenciák esetlegességek alakítanak ki. A hatékony számításhoz, érveléshez aktívan keresni kell a dekomponáló

változókat, a hatékony sorrendezést, hogy a kiadódó reprezentáció hatékonyan írhassa le az eloszlás objektíve létező feltételes függetlenségit.

1.4.9. Bayes-hálók struktúra tanulása

Mivel a teljes bayesi következtetés annak komplexitása miatt csak különleges esetekben hajtható végre, gyakran a teljes modelltér helyett csak egyetlen modellt használunk. Ha elegendő statisztikai adat áll rendelkezésre, a fent bemutatott manuális konstruálás mellett szerepet kaphat az optimális modell keresése, a *tanulás*, mely végezhető a szakértői modellből kiindulva, annak finomításával, vagy tabula rasa alapon is. A tanulás, mint az optimális modell keresése, a parametrikus következtetés alkalmazásának tekinthető, és megmutatható, hogy NP-teljes bonyolultságú.

A tanulás két szinten lehetséges: kereshetjük adott struktúra mellett az optimális paraméterezést (paramétertanulás), vagy a legjobb struktúrát és annak paraméterezését (struktúratanulás). Az optimalitás valamilyen mérték szerint értendő, ez legegyszerűbb esetben a modell a posteriori valószínűsége.

A MAP modell keresése mellett elképzelhető más kritériumfüggvény is, amely leggyakrabban az a posteriori valószínűség egyenletes priorral, kiegészítve valamilyen, a struktúra bonyolultságát büntető taggal. Az ilyen büntetés alkalmazása felfogható a prior módosításának: minél erősebb a büntetés, annál kisebb a bonyolult struktúrák valószínűsége. A leggyakoribb ilyen minősítési függvény a bayesi információ-kritérium függvény (BIC – Bayesian information criterion), ennek képlete:

$$BIC(G,D) = \log P(D \mid G) - \frac{\log N}{2}|\theta|, \tag{4}$$

ahol 'N' a tanító minták, |θ| pedig a háló paramétereinek száma. A *log*N-nel arányos mellett még elképzelhető N-ben lineáris vagy polinomiális büntetés is.

Számításigényét tekintve a tiszta a posteriori kritériumfüggvény, és a teljes, független, azonos eloszlású minták alapján végzett tanulás a legegyszerűbb. Ekkor, Dirichlet eloszlású paraméterpriort feltéve, adott struktúra a posteriori valószínűsége egyszerű, zárt formában számítható [18-20]:

$$P(G \mid D) \propto P(G,D) = P(G)\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{(N'_{ij}+r_i-1)!}{(N_{ij}+N'_{ij}+r_i-1)!}\prod_{k=1}^{r_i}\frac{(N_{ijk}+N'_{ijk})!}{N'_{ijk}!}, \tag{5}$$

ahol az N_{ijk} az i. változó j. szülői konfigurációjának és k. értékének az előfordulását, q_i az i. változó szülői konfigurációinak a számát és r_i az értékeinek számát jelenti (N_{ij} a megfelelő marginális). Az N'_{ijk} a megfelelő virtuális mintaszámokat jelöli (ezek előismeretek hiányában 1-nek választhatók).

Paramétertanulás esetén az optimális paraméterezés az FVT-k külön-külön, relatív gyakoriságokkal való kitöltésével elérhető, struktúratanulás esetén pedig minden csomóponthoz külön megkereshető az optimális szülői halmaz, feltéve hogy ismert a csomópontok egy kauzális rendezése[1]. Ha ilyen

[1]Egy kauzális rendezésben a csomópontok szülei csak az őket megelőző változók közül kerülhetnek ki. Egy kauzális rendezés a reprezentáns DAG csúcsainak egy topologikus rendezése.

információ nem áll rendelkezésre, ügyelni kell, hogy a DAG tulajdonság ne sérüljön, pl. úgy, hogy minden lehetséges sorrendet külön megvizsgálunk.

1.4.10. Bayes-hálók tanulása hiányos adatok alapján

Amennyiben a tanító adatok hiányosak, azaz bizonyos változók értéke nem minden esetben ismert a tanulás feladata jóval nehezebbé válik. Ilyenkor a paramétertanulásban iteratív eljárások használhatók, a legismertebbek ezek közül a gradiens alapú közelítő eljárások vagy ezek robosztusabb változatai, a konjugált gradiens és a skálázott konjugált gradiens algoritmusok, vagy az expectation maximization algoritmus [21, 22].

Struktúratanulás esetén, mivel a szülői halmazok nem tanulhatók külön még adott sorrendnél sem, a teljes struktúrateret bejáró keresésre van szükség. Mivel a lehetséges struktúrák száma a csomópontok számával szuperexponenciálisan nő, a gyakorlatban nem teljes keresési eljárásokat kell alkalmazni, pl. mohó keresést vagy szimulált lehűtést (ekkor az elemi lépés pl. egy él törlése, beszúrása, megfordítása lehet).

Ezek az eljárások is csak akkor működnek azonban, ha az adatokra teljesül a véletlenszerű eltűnés (MAR – missing at random) feltétele, azaz ha a bejegyzések eltűnése nem függ az eltűnt értéktől [16].

1.5. Biomarker elemzés

Az orvosbiológiai kutatások egyik alapkérdése egy vagy több kimeneteli változó esetén azon változók beazonosítása, amelyek prediktív (diagnosztikai) vagy beavatkozási (terápiás) lehetőségeket kínálnak [23, 24].

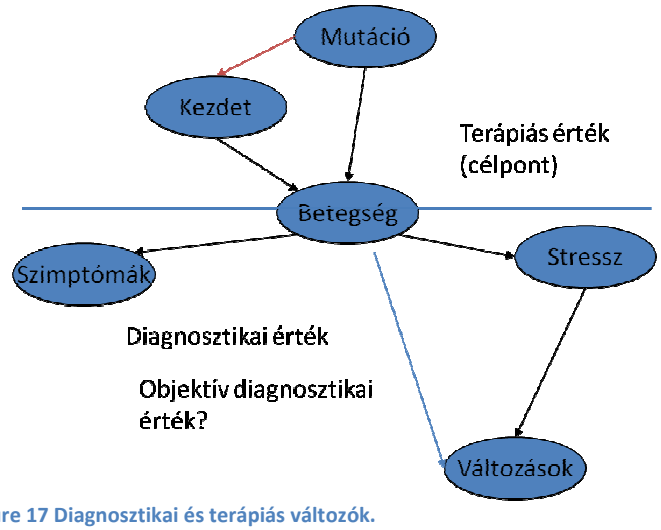


Figure 17 Diagnosztikai és terápiás változók.

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 kanonikus volta is jelezhet. Oksági aspektusban szintén nyilvánvaló

követelmény a hatáserősség, illetve itt is a rendszerszintű kanonikus volta változóknak. Mindkét esetben közös szempont lehet az elérhetőség és a költség aspektusa. A formalizálás kidolgozásához tekintsük a következő fogalmakat.

Egy Xi sztochasztikus változó (jegy) erősen releváns Y és az összes változó U vonatkozásában, ha I(Xi | U\ [Xi, Y] | Y). Gyengén releváns, ha létezik olyan U' részhalmaza U-nak, amelyre I(Xi | U'\[Xi , Y] | Y). Irreleváns, ha sem erősen, sem gyengén nem releváns. Egy sztochasztikus változó (jegy) erősen (gyengén) releváns **Y** célváltozó halmazra vonatkoztatva, ha erősen (gyengén) releváns valamely **Y**-beli Y esetén.

Sztochasztikus változók (jegyek) egy S részhalmaza Markov-takaró Y célváltozó és az összes változó U vonatkozásában, ha I(U\[S,Y] | S| Y). U részhalmaza S Markov-határ Y célváltozó és az összes változó U vonatkozásában, ha I(U\[S,Y] | S| Y) és S minimális.

Tétel formájában kimondható, hogy ha az U feletti P eloszlást G Bayes-háló definiálja, akkor a bayes statisztikai keretben - matematikai értelemben 1 valószínűséggel - Y változó Markov-határa a Y változót reprezentáló csomópont szülei, gyermekei, és gyermekei egyéb szülei, lásd 18. és 19. ábra. További tétel mondja ki, hogy a Markov-határ pontosan az erősen releváns változókat tartalmazza.

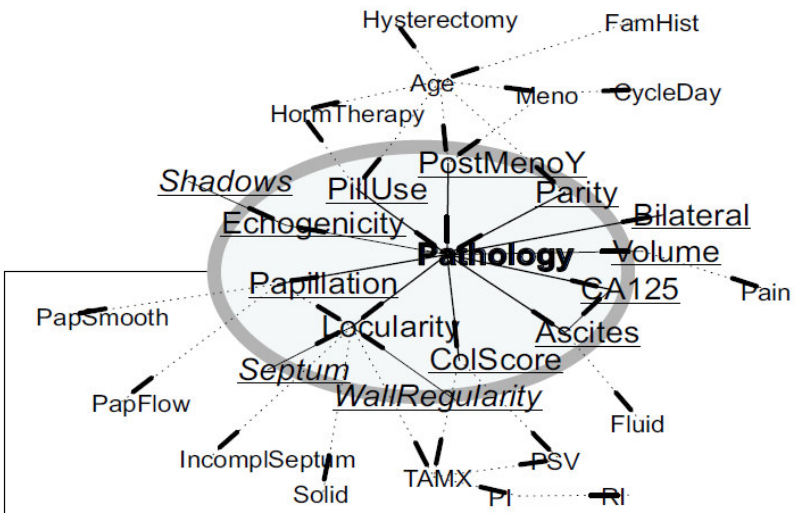


Figure 18 Egy preoperatív petefészekrák diagnosztikai modell. A Pathology célváltozót félkövér kiemelés jelzi, Markov-takaróját szürke keret.

A Markov-takaró jelentőségét az adja, hogy egy olyan minimális változóhalmazt azonosít, amely szükséges és elégséges U változóhalmaz esetén [15].

Több célváltozó esetén a Markov-határ fogalmát az 19. ábra illusztrálja.

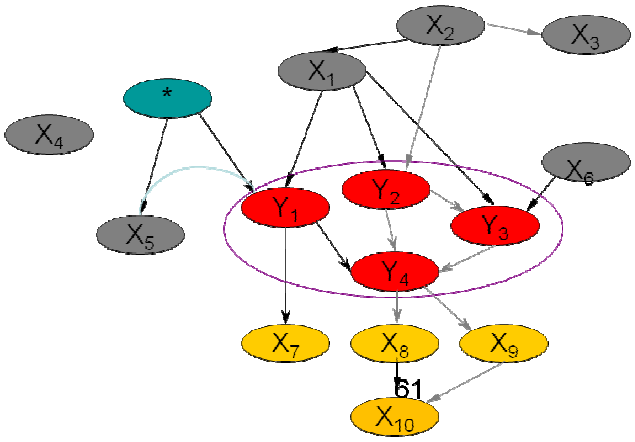


Figure 19 Markov-takaró több célváltozó esetén. Piros változók jelzik a célváltozókat, X1, X2, X6, X7, X8, X9 változók alkotják a Markov takaró, X4 változó irreleváns, a többi változó gyengén releváns.

A Markov-határ fogalma lehetővé teszi egy szimmetrikus, páronkénti reláció definiálását az Markov-határbeliséget, az egymás Markov-határában való előfordulást (MBM(X,Y) – Markov boundary membership), amely tehát általában egybeesik az erős relevancia fogalmával. Az erős relevancia fogalma azonban érdekes módon eltér a ma domináns asszociáció fogalmától. Közös találkozási pontjuk a direkt relevancia. Azonban csak az erős relevanciába tartozik bele a „kizárólagos interakciós relevancia", amely például episztatikus, főhatás nélküli, azaz csak hatásmódosító változók relevanciáját jelöli. Ezzel szemben csak az asszociációba tartozik bele a „kizárólagos zavaró tényezős relevancia", amely csak zavaró tényezők által létrehozott statisztikai asszociációt jelöl,illetve a „kizárólagos tranzitív relevancia", amely csak más kapcsolt változók által mediált asszociáció (lásd 20. ábra).

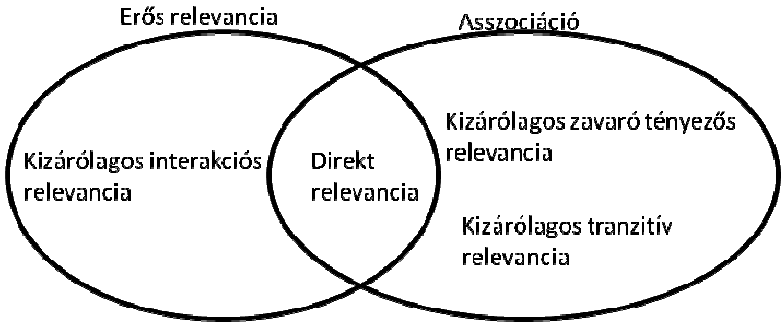


Figure 20 Az erős relevancia és asszociáció kapcsolata.

A gráfos valószínűségi modellek, benne a Bayes hálók egy könnyen értelmezhető reprezentációt jelentenek a relevancia vizualizálására, illetve egy hatékony matematikai reprezentációt is kínálnak. Azonban a Markov-határ kardinalitása exponenciális, ennek megfelelően a teljes többváltozós megközelítés nem lehetséges, a hipotézistesztelési keretben ez gyenge szignifikanciát, a bayesi keretben pedig „lapos" poszteriort jelent, azaz amikor nincs domináns modell, sem modelltulajdonság, lásd 21. ábra.

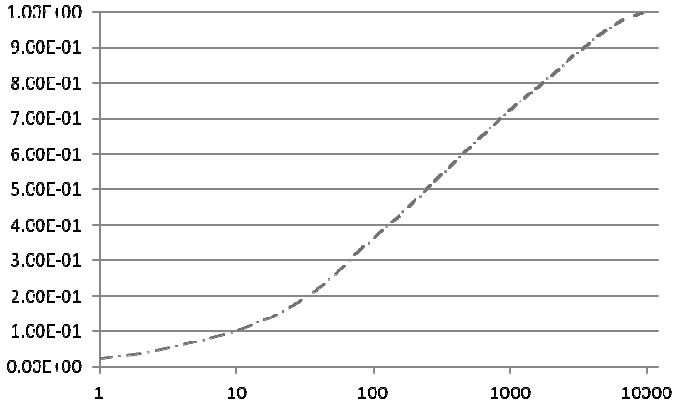


Figure 21 A legvalószínűbb Markov-takarók a posteriori valószínűségének kumulatív értéke.

Mivel tipikusan - még lapos posteriorok esetén is a legvalószínűbb MBS jegyek rendelkeznek közös részekkel, érdemes használni a k-MBS jegyeket [25].

[*k-MBS*] Egy adott $p(\mathbf{V})$ eloszlás esetén, ($|\mathbf{V}|=n$), ha minden $X_i \in \mathbf{s}$, $\mathbf{s} \subseteq \mathbf{V}$, változóra igaz, hogy tagja az mbs Markov-határ halmaznak és $|\mathbf{s}|=k$, akkor $\mathbf{s}$ egy k-adrendű Markov-határ halmaz $k-MBS_p(\mathbf{s},Y) \Leftrightarrow (\exists mbs : MBS_p(mbs,Y), \mathbf{s} \subseteq mbs$.

A $k-MBS$ jegyek előnye hogy skálázhatóak, kardinalitásuk polinomiális $O(n^k)$, éppen ezért jól alkalmazhatóak a relevanciaanalízis során. A gyakorlatban ez azt jelenti, hogy megvizsgálhatjuk a legvalószínűbb $k-MBS(Y)$ a k paraméter egy elég széles tartományában. További előnyük, hogy a k-MBS posteriorok offline számíthatók a MBS posteriorok közelítő értékéből. A legnagyobb *k* érték, amelynél az egyes modell tulajdonságok (egyes strukturális jegyek) nagy valószínűséggel megjelennek, problémafüggő, de tipikus értékei a 2-5 tartomány.

1.6. Tudásfúzió: tudásbázisok és elemzések integrálása

Az adat- és tudásfúzió kérdése az orvosbiológia kutatásokban több tényezőnek köszönhetően vált központivá.

1. A gyengén kapcsolt, autonóm szintek sokasága miatt, amelyek önállóan is nagy méretűek, omikai voltukban.

2. Mint tudás-gazdag, adatszegény tudomány a felhalmozódó megfigyelések újrafelhasználása, meta-elemzése egy fontos jellemvonása a modern orvosbiológiának.

3. A számítási kapacitás exponenciális felfutása, ami a 2000-es évtől már nem gyorsaságban, hanem párhuzamos elérhetőségben, azaz fajlagos költségben jelenik meg a számításintenzív technikákat egyre inkább lehetővé tették.

4. A szemantikus világháló megjelenése a kutatások elosztott voltát továbberősítette, ami már az 1990-es években is sikeres fúziókat tett lehetővé (ld. R.Altman: Riboweb).

5. A szemantikus világháló megjelenése az elektronikus tudásábrázolást is forradalmasította. Az ontológiák megjelenése lehetővé tette az adatok, modellek, és kodifikált tudás egységes leírását.

6. A nagy-áteresztő képességű orvosbiológiai mérések megjelenése az orvosbiológiában hirtelen jött adatgazdagságot eredményezett, ami azonban a változók nagy száma, a modellek komplexitása, és a viszonylagosan kis mintaszám mellett nem vezetett, nem vezethetett egyértelmű hipotézisekhez.

A felhalmozódó milliós nagyságrendű orvosbiológiai fogalom és reláció, az felhalmozódó milliárdos nagyságrendű adatok, és az ezer-milliárdos nagyságrendű számítási kapacitás idealizált metszetében helyezhető el egy tudás- és számításintenzív induktív következtetési irány (lásd 22. ábra). Hagyományait tekintve ez akár az enciklopédista, majd a pozitivista, bécsi körhöz is kapcsolható, hiszen szándéka a megfigyelésekig „lehorgonyzott", visszavezetett tudásreprezentáció és az egységes nyelven, automatizáltan lekövethető induktív következtetés.

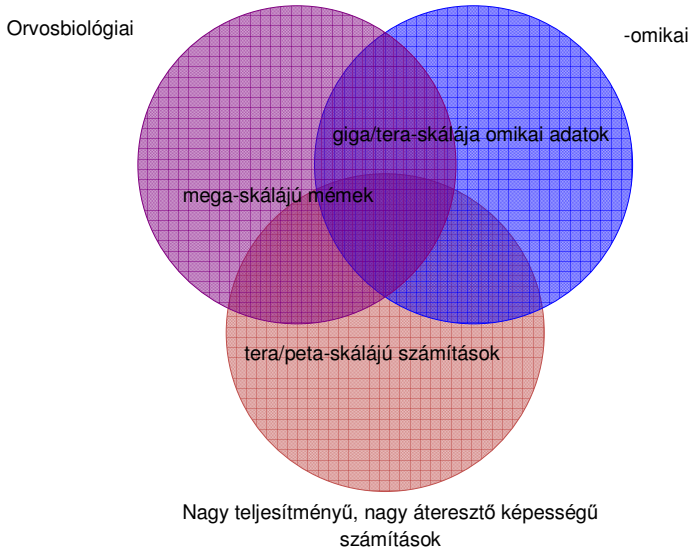


Figure 22 Tudás, adat, és számítási kapacitás integrálása.

Az indokok és lehetőségek sokféleségének köszönhetően a fúziós kutatások is sokfélék, kapcsolódó szűkebb szakterületek a következők:

1. természetes nyelvi elemzések, orvosbiológiai szövegbányászat, tudáskivonatolás;
2. adattárházak, adatbázis integráció;
3. logikai integráció, logikai következtetés;
4. szemantikus web, szemantikus integráció;
5. orvosbiológiai tudásbázisok és ontológiák;
6. szemantikus modell-leíró nyelvek;
7. munkafolyamat rendszerek;
8. integrált, orvosbiológia modellek;
9. fúziós elvek és keretrendszerek;
    a. adatelemzési eredmények és logikai tudásbázisok integrálása
    b. kernel-fúziós keretek [26],
    c. adatelemzési tudásbázisok, off-line meta-elemzések,
    d. valószínűségi logikai keretrendszerek.

Az egyik legnagyobb kihívást jelenleg az adatelemzéshez kapcsolódó fúzió jelenti, különös tekintettel a heterogén adatok egymással történő fúziójára, és az adatok és a háttértudás fúziójára. Erre a következő megközelítések jelentek meg.

1. Metaanalízis, amelyben a korábbi adatok egységesítésével és átfogó modell megalkotásával lehetségessé válik az adat-szintű fúzió, a meglévő háttértudás pedig opcionálisan a modellalkotásban, a priori eloszlásokban, vagy az értelmezésben kerül felhasználásra.
2. Bayes szekvenciális elemzés, amelyben a korábbi adatok elemzéséből származó eredményekből és a meglévő háttértudásból „posterior prior”-okat konstruálunk, amelyek a korábbi adatok, ismeretek, és invesztált számítások összegzései.
3. [Sztochasztikus] Induktív logikai programozás, amely a háttértudás és az adatok logikai reprezentációja esetén megpróbál konzisztens hipotéziseket kikövetkeztetni. Valószínűségi kiterjesztése is megjelent már.

4. Kernel-alapú fúzió, ami korábbi adatok és akár a szakirodalom egységes statisztikai jellemzésén alapul speciális hasonlósági relációk felhasználásával.

5. Adat-világ kontra tudás-világ fúzió, amelyben az adatok tipikusan szeparált elemzésének eredményét az aktuális adatelemzéstől tipikusan független tudásbázissal vetjük össze, az eredményeket abban karakterizáljuk.

6. Programozási (burkoló) környezetek, amelyek támogatást adnak a tetszőleges szintű fuzionáltatásra, mint például a Bioclipse, Cytoscape, Bioconductor.

7. Munkafolyamatos (burkoló) környezetek, amelyekben grafikus felületen tervezhető meg az összetett elemzés Biopipe, biomart, stb.

8. Nyelvi lekérdező (burkoló) környezetek, amelyekben egy kontrollált természetes nyelvi felületen lehet összetett elemzéseket végrehajtatni.

Egy integrált informatikai rendszer célja lehet, hogy a sokféle forrás adat- és tudásbázis és szolgáltatás különbözőségét elfedje, egységesítve jelenítse meg őket. Ez egy komplex lekérdezést biztosító bioinformatikai lekérdező, amely felhasználva egy biológiai ontológiát és az adat- és tudásbázisok és szolgáltatások modelljeit, a komplex kérést lefordítja a heterogén forrásokra, ütemezi azokat, begyűjti, integrálja, majd a komplex választ egy egységes felületen megjeleníti. A források integrálása azonban rendkívül idő és munkaigényes, lekérdező nyelv komplexitásának léteznek korlátai („mi kérdezhető"), a felhasznált források lehet nem az adott kontextushoz legjobban illőek vagy nem a legfrissebbek, ami az ilyen típusú rendszerek elkerülhetetlen negatív tulajdonsága.

Jelenleg a leginkább elterjedt fúziós módszer a duális, globális, környezetfüggetlen „adat-világ kontra tudás-világ" fúzió. Tipikus példái a következőek, amelyek a gén expressziós adatok elemzési igénye miatt 2000-től jelentek meg először kutatási eszközökben, ma pedig már kereskedelmi eszközökben is.

- Adatelemzési eredmények, például releváns változók halmazának vagy egy adott klaszternek a szöveges karakterizációja, tipikusan ontológiai leírók vektor alapú profilja, ami az adott halmazbeli entitások szöveges vektor reprezentációjában szignifikánsan eltérően reprezentált,

- Duális klaszterezés, ami az entitások adatalapú és szakirodalom alapú klaszterezését is jelenti, majd ezek összevetését.

- Duális hálózatelemzés, ami az entitások interakcióinak, függési rendszerének, vagy oksági viszonyainak az adatalapú és szakirodalom alapú létrehozását és összevetését jelenti.

Table 4 Adat- és tudásfúziós módszerek.

Módszer	Adatelemzési	Háttértudás bemenet	Kimenet

	bemenet		
Ingenuity – IPA	Omikai (génszintű) pontszám	Diagrammatikus, logikai tudásbázis	Pontszámokkal annotált útvonalak
GSEA	Omikai (génszintű) sorrend	Halmazok (annotáció, ontológia, vagy útvonal alapon)	Felülreprezentált elemek
Hasonlósági/kernel-fúzió	Hasonlósági/kernel mátrixok	Hasonlósági/kernel mátrixok	Hasonlósági/kernel mátrixok
Prioritizálás	Halmaz	Hasonlósági/kernel mátrixok	Sorrendezés

Az egyik legsikeresebb fúziós rendszer az Ingenuity az egyváltozós adatelemzési eredményeket integrálja a háttértudásbázisával, ami vizualizálást, szó-profilozást, és egyéb tudásmenedzsmentbeli támogatást jelent (lásd 5. ábra). Tulajdonságait és hiányosságait a 2. táblázat összegzi.

Table 5 az Ingenuity-IPA adat- és tudásfúziójának tulajdonságai.

Ingenuity alapú fúzió tulajdonságai	Hiányzó/komplementer tulajdonság
Egyváltozós statisztikai elemzések	Többváltozós statisztikai elemzések
Bizonytalanság heterogén kezelése (eltérő jelentésű pontszámok és szignifikancia szintek)	Egységes modell tulajdonságok feletti poszteriorok
Rögzített adatelemzési kimenetelek	Szabad, szemantikai adatelemzési kimenetelek
Propozicionális logika (diagrammatikus reprezentáció)	Leíró logika

A duális fúzióra egy másik példa a GSEA rendszer, amely egy sorrendet fogad az adatelemzésből és referencia halmazokat az MSigDB forrásból, illetve tetszőleges halmazokat. Az MSigDB génhalmazok öt fő csoportba oszthatóak: C1: „positional” (326), C2: „curated” (3272), C3: „motif” (836),C4: „computational” (88), C5:”GO” (1454), amelyek szekvenciális, közös motívumok, vagy akár a Gene Ontology alapján közös folyamatok, útvonalak alapján vannak meghatározva. A GSEA egy olyan számítógépes módszer, amely megmutatja, hogy az adataink alapján létrehozott sorrendben fel van-e dúsulva vajon egy előzetesen meghatározott génkészlet (úgynevezett gene set). Ezáltal a módszer a teljes rangsort, azaz bár a sorrend egyváltozós, de mégis több változó pozíciója alapján határozza meg egy halmaz (azaz

egy annotáció) feldúsulását. Az elemzés elsődleges eredménye az úgynevezett feldúsulási érték (Enrichment Score, ES), amely annak a fokát mutatja meg, hogy egy bizonyos génkészlet (például valamilyen jelátviteli útvonal) génjei egyenlően oszlanak-e el a rangsorolt listán, vagy pedig a rangsorolt lista tetején vagy az alján halmozódnak fel. Az algoritmus lefelé haladva pásztázza a rangsorolt listát, és növeli a futó-összegző statisztikát, ha egy gén eleme egy génkészletnek, és csökkenti, ha nem tartozik bele. Pozitív ES a rangsorolt lista tetején, míg a negatív ES a rangsorolt lista alján jelez génkészlet feldúsulást. Ha több génkészletet tesztelünk egyszerre, szükséges az ES normalizálása, melyet a normalizált feldúsulási érték (Normalized Enrichment Score, NES) fejez ki. Azt, hogy egy adott génkészlet feldúsulása mennyire szignifikáns, a nominális p érték szimbolizálja. Azonban, ha több génkészlet feldúsulását vizsgáljuk egyszerre, mindenképpen szükséges annak a valószínűségnek a becslése, hogy egy adott NES érték mekkora eséllyel hamis pozitív, tehát csak a véletlen műve, hogy a génkészlet szerepel az eredmények között. Ennek a kifejezésére való az FDR-q érték (False Discovery Rate), amely egy többszörös hipotézisteszteléses, permutációs teszt alapú általánosítása a p-értéknek.

A duális fúzionálásra egy másik példa az Endeavour rendszer, ami génprioritizálás esetén használja fel az asszociációs relációkat, akár a hasonlósági mátrixok lineáris vagy nem-lineáris kombinációjára, vagy akár egy kernel alapú fúziónak a felhasználásával, hozva létre egy lekérdezés függő fúzionált, hasonlósági mátrixot (lásd X ábra).

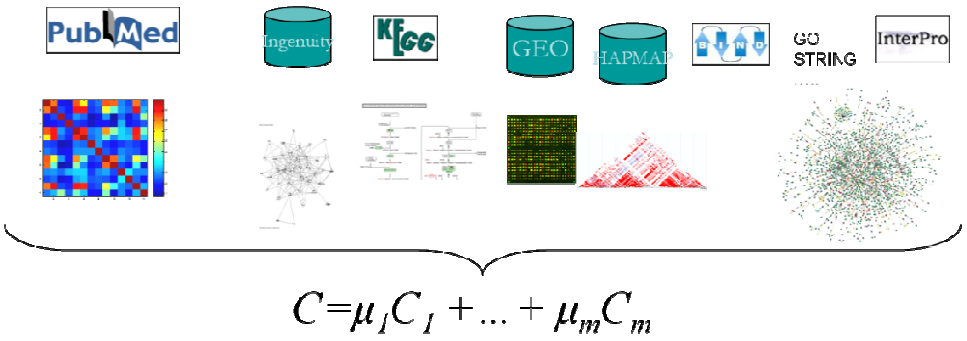


Figure 23 Hasonlósági mátrixok lineáris fúziója.

Ennek a megközelítésnek a meglévő és leginkább hiányzó tulajdonságait a 3. táblázat mutatja be.

Table 6 A kernelfúziós módszerek tulajdonságai.

Kernel alapú fúzió	Hiányzó tulajdonság
A hasonlóságok heterogénitása	Egységes ontológia
Páronkénti	Többváltozós

Szimmetrikus	Aszimmetrikus
Tranzitíve (indirekt asszociációs relációk)	intranzitivitás (csak direct relációk)
Negálás és hiányzás	Negálás
Környezet független	Tárgyterület/probléma figyelembe vétele
Sorrendezés és vizualizáció	Szimbolikus lekérdezés
Numerikus (algebrai megközelítés)	Szimbolikus (logikai megközelítés)
Passzív megfigyelési	Beavatkozás és okozatiság
Hasznosság	Valószínűségek
Egyváltozós sorrendezés/pontozás	Többváltozós tetszőleges szemantikai állítás
Hiányzó vagy ad hoc következtetés	Modell elméleti szemantikával rendelkező kalkulus
Atemporális	Temporalitás
Amodális	Modalitás (például okozatiság)
Hatékony számíthatóság	Nagy számításigény
kvantifikálásra nincs lehetőség	Kvantifikált állítások
Logikai tudás transzformációja nem szemantikai	Logikai tudás megőrzése eredi gazdagságában
Modellmentes	Gazdag modellezési lehetőség

Az egyre népszerűbb bayesi elemzések lehetőséget adnak arra, hogy a heterogén szinteken megjelenő direkt valószínűségi állításokat egy központi, például génszinten összesítsük, majd terjesszük magasabb, például útvonalak szintjére (lásd 24. ábra).

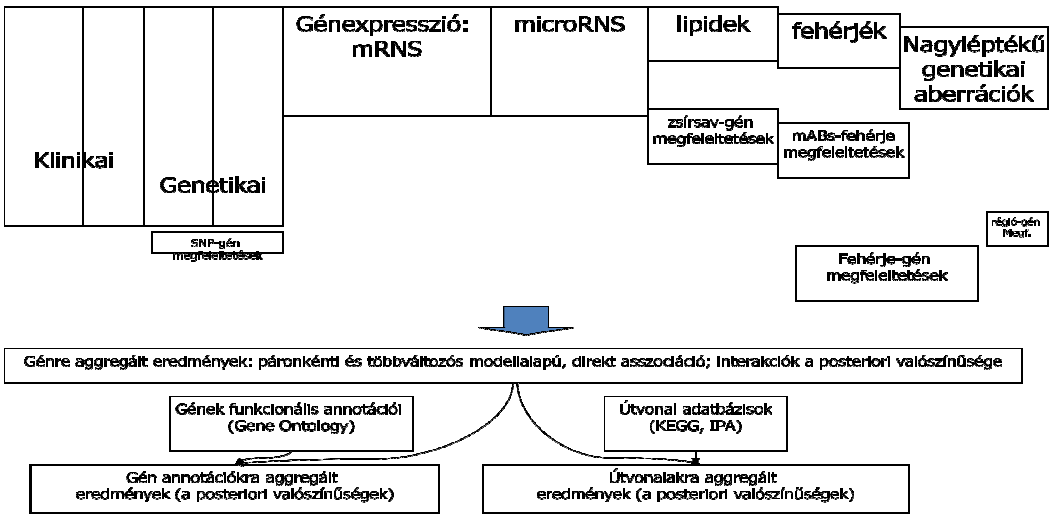
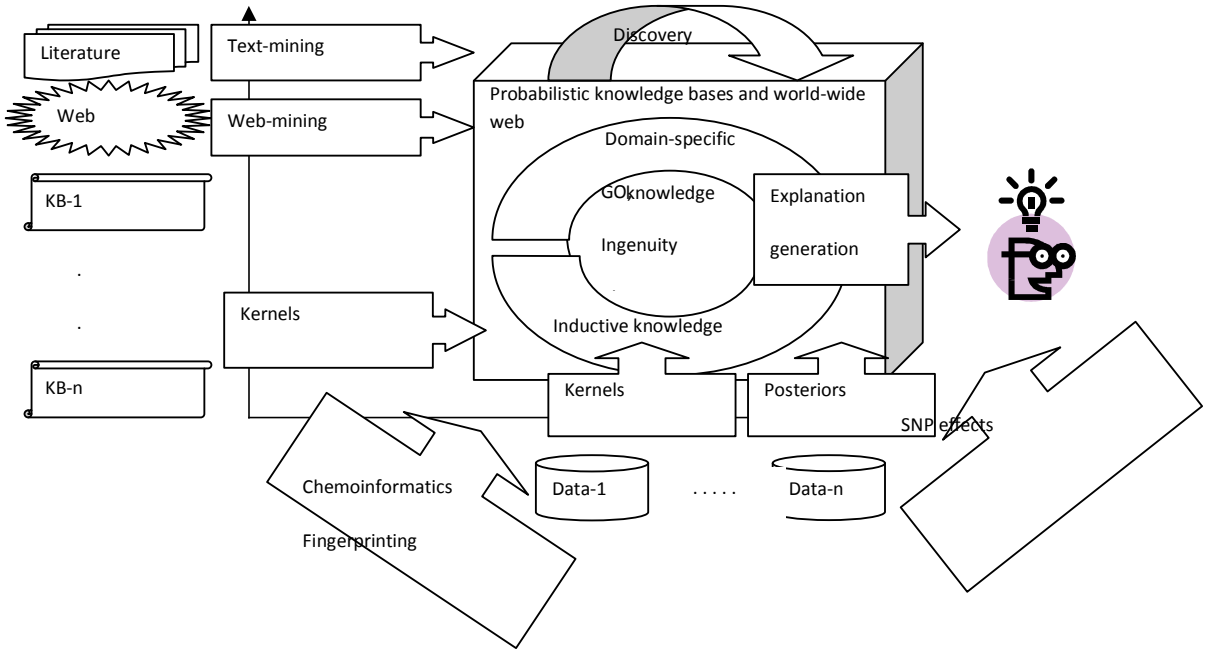


Figure 24 Omikai szinteken és absztrakciós szinteken átívelő bayesi tudásfúzió.

Ez a valószínűségi-szemantikus fúziós keret jelenleg azonban nagyon korlátos: logikai értelemben az ontológiák leírására optimalizált egyszerű leíró logikai szintre képes indukálni az adatelemzésből származó poszteriorokat, másrészt jelenleg nincsenek integrálva a klaszterezési/hasonlósági/kernel-fúziós technikák. Mindazonáltal szemantikai jellege miatt ez tűnik alapvetőnek ezen irányok integrálására. Az 1. ábra illusztrál egy jövőben várható elképzelt keretrendszert. A központi részt általános orvosbiológiai tudásbázisok alkotják, mint a Gene Ontology, Human Phenome Ontology, OMIM, dbSNP, vagy akár az Ingeniuty. Az ezek alkotta magot automatikusan kivonatolt, formálisan reprezentált betegség-specifikus információk egészítik ki, bizonytalansági és környezeti információkkal. A másik lényegi kiegészítője a rendszernek az induktív következtetések eredményeinek formális reprezentációja, különös tekintettel a Bayes adatelemzésre és az asszociációs relációkra, és a kernel-alapú fúziós technikák eredményeire. További modulok a tudásbázishoz viszonyított szerepük szerint csoportosíthatók, úgymint konstruktorok, gazdagítók, lekérdezők, következtetők, és magyarázat generálók.



$$P(\phi \mid KB) = \sum_{\text{pKB}} P(\phi \text{ is provable} \mid pKB, lKB, kKB) p(pKB)$$

Figure 25 Valószínűségi-szemantikus adat- és tudásfúziós keret.

1.7. Orvosi döntéstámogatás

A klinikai adatok egyre teljesebb, részletesebb elektronikus elérhetősége kiegészül a már tárgyalt poszt-genomikai adatokkal, de akár a mindennapi életvitelből is származó adatokkal (például otthoni, időskori monitoring mellett). A személyreszabott medicina eszköztára is paradigmatikus változások előtt áll, amelyben várhatóan jelen lesznek az egyes genetikai variánsok megléte alapján adható orvosságok, akár a személyes érzékenység miatt a mellékhatásokra, vagy akár a betegség genetikai alábontása miatt („útvonal-gyógyszerek”). Az orvosi munkát várhatóan egyre nagyobb mértékben egészíti majd ki orvosi döntéstámogató és szakértői rendszerek . Az időskorú populáció miatt egy egyre fontosabb scenáriót,az időskori otthoni gondozás egy keretét az 26. ábra mutatja.

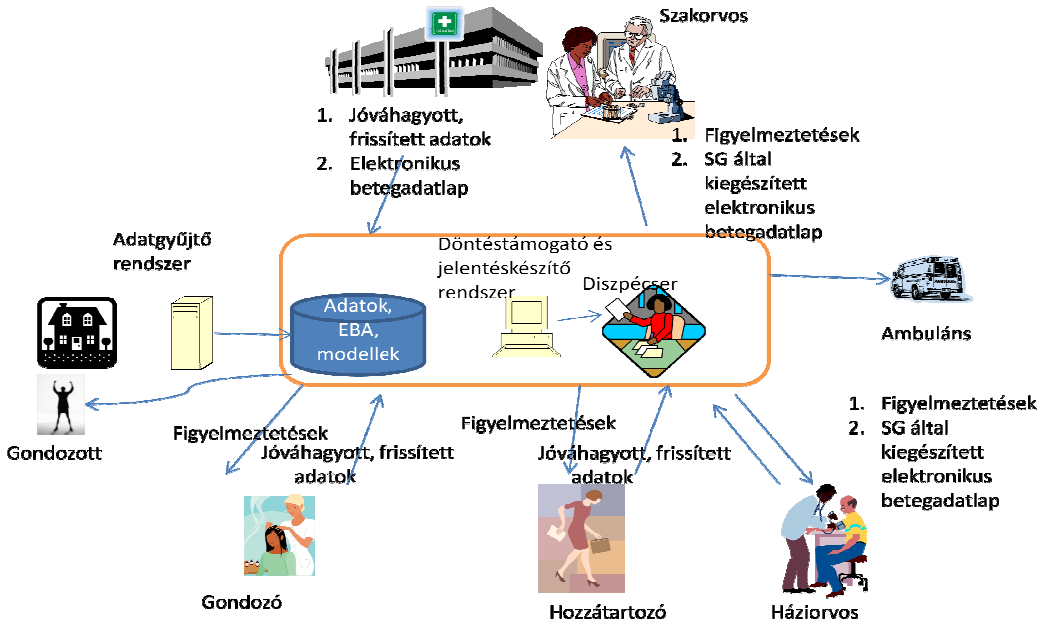


Figure 26 Orvosi döntéstámogató rendszerek lehetséges szerepei.

A korábban tárgyalt gráfos valószínűségi modellek mellett szükséges a lehetséges döntések formális reprezentálása, illetve a hasznosságok és veszteségek reprezentálása, lásd 27. és 28. ábra.

- Döntési szituáció:
  - Akciók a_i
  - Kimenetelek o_j
  - Kimenetelek valószínűségei $p(o_j \mid a_i)$
  - Kimenetelek hasznosságai/veszteségei $U(o_j \mid a_i)$
    - QALY, micromort
  - Maximális várható hasznosság elve $EU(a_i) = \sum_j U(o_j \mid a_i) p(o_j \mid a_i)$
    - Legjobb választás a maximális hasznosságú opció $a^* = \arg\max_i EU(a_i)$

Figure 27 Egy döntési helyzet elemei.

A döntési helyzet elemeit az akciókat, és hasznosságokat egy bemenetek (szülők) nélküli „cselekedet-csomópont” jelöli egy téglalapként és egy determinisztikus (függvény-) csomópont, amit egy rombusz jelöl [27]. Természetesen, egy valóságos modellben sokféle cselekedet van, amelyeket a számítási igény csökkentése miatt akár bizonyos más változókhoz mint feltételekhez is lehet kötni, amelynek az lehet a felhasználása, hogy bizonyos feltételek teljesülése esetén számolja csak ki adott cselekedet

várható hasznosságát. Az 28. ábra egy olyan döntési hálót ábrázol, amelyben egy teszt elvégzésének a költség-haszon elemzése formalizálható.

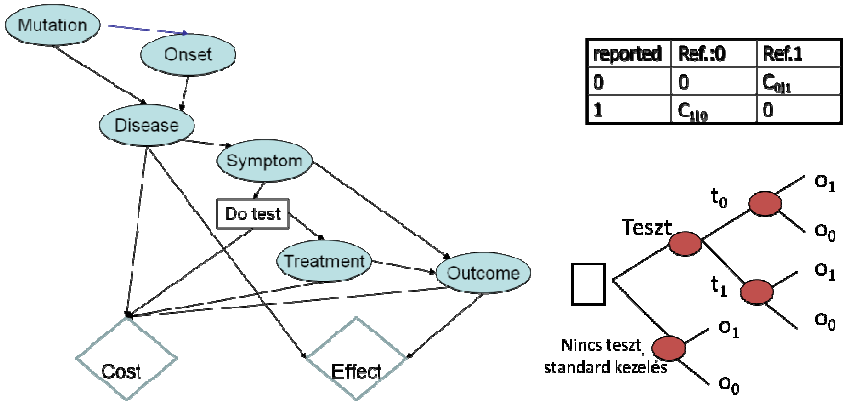


Figure 28 Egy döntési háló, veszteség mátrix, és döntési fa.

Ekkor ha feltesszük, hogy a t (genetikai) tesztnek

- C_t a költsége,
- két kimenetele van t_0, t_1 , $p(t_1)=q$ valószínűséggel,
- segíti a kezelés megválasztását x kontextusban: $p(Y=1|X=x, t_i)=p_i$

Akkor a teszt várható értéke: $EL - ((1-q)EL_0 + qEL_1)$, mivel

- A teszt nélkül várható veszteség: $EL=\min(pC_{0|1},(1-p)\ C_{1|0})$
- A teszttel a várható veszteség $(1-q)EL_0 + qEL_1$
  - t_0: $EL_0=\min(p_0C_{0|1},(1-p_0)\ C_{1|0})$
  - t_1: $EL_1=\min(p_1C_{0|1},(1-p_1)\ C_{1|0})$

Ha feltesszük, hogy ha a teszt negatív, akkor nem nyerünk információt, azaz $EL_0 \approx EL$, akkor $(1-q)EL_0 + qEL_1 - EL \approx q(EL_1 - EL)$, azaz $q(p-p_1)C_{0|1}$, ami azt az intuitive is elvárt eredményt formalizálja, hogy minél gyakoribb profil esetén használható a teszt, és ott minél nagyobb mértékben növeli meg az érzékenységet, annál hasznosabb.

1.8. Irodalom

1. **THE SEQUENCE EXPLOSION**. *Nature* 2010, **464**(7289):670-670.

2. Carlson R: **The Pace and Proliferation of Biological Technologies**. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science* 2004, **1**(3).
3. Robinson P, Mundlos S: **The Human Phenotype Ontology**. *Clinical Genetics* 2010, **77**(6):525-534.
4. Richard Gliklich M: **Patient Registries and Rare Diseases - Applied Clinical Trials**. *Applied Clinical Trials* 2012.
5. Rubinstein YR, Groft SC, Bartek R, Brown K, Christensen RA, Collier E, Farber A, Farmer J, Ferguson JH, Forrest CB *et al*: **Creating a global rare disease patient registry linked to a rare diseases biorepository database: Rare Disease-HUB (RD-HUB)**. *Contemp Clin Trials* 2010, **31**(5):394-404.
6. Swanson D, Smalheiser N: **An interactive system for finding complementary literatures: A stimulus to scientific discovery**. *Artificial Intelligence* 1997, **91**(2):183-203.
7. Thompson P, McNaught J, Montemagni S, Calzolari N, del Gratta R, Lee V, Marchi S, Monachini M, Pezik P, Quochi V *et al*: **The BioLexicon: a large-scale terminological resource for biomedical text mining**. *Bmc Bioinformatics* 2011, **12**.
8. Andronis C, Sharma A, Virvilis V, Deftereos S, Persidis A: **Literature mining, ontologies and information visualization for drug repurposing**. *Briefings in Bioinformatics* 2011, **12**(4):357-368.
9. Glenisson P, Coessens B, Van Vooren S, Mathys J, Moreau Y, De Moor B: **TXTGate: profiling gene groups with text-based information**. *Genome Biology* 2004, **5**(6).
10. Glenisson P, Antal P, Mathys J, Moreau Y, De Moor B: **Evaluation of the vector space representation in text-based gene clustering**. In: *Pacific Symposium on Biocomputing (PSB03)*. 2003: 391-402.
11. Baeza-Yates R, Ribeiro-Neto B: **Modern Information Retrieval**: ACM PresS; 1999.
12. McClellan J, King M: **Genetic Heterogeneity in Human Disease**. *Cell* 2010, **141**(2):210-217.
13. Maher B: **Personal genomes: The case of the missing heritability**. *Nature* 2008, **456**(7218):18-21.
14. Pearl J: **Causality : models, reasoning, and inference**. Cambridge, U.K. ; New York: Cambridge University Press; 2000.
15. Pearl J: **Probabilistic reasoning in intelligent systems : networks of plausible inference**. San Mateo, Calif.: Morgan Kaufmann Publishers; 1988.
16. Gelman A: **Bayesian data analysis**, 1st edn. London ; New York: Chapman & Hall; 1995.
17. Gilks WR, Richardson S, Spiegelhalter DJ: **Markov chain Monte Carlo in practice**. London: Chapman & Hall; 1996.
18. COOPER G, HERSKOVITS E: **A BAYESIAN METHOD FOR THE INDUCTION OF PROBABILISTIC NETWORKS FROM DATA**. *Machine Learning* 1992, **9**(4):309-347.
19. HECKERMAN D, GEIGER D, CHICKERING D: **LEARNING BAYESIAN NETWORKS - THE COMBINATION OF KNOWLEDGE AND STATISTICAL-DATA**. *Machine Learning* 1995, **20**(3):197-243.
20. Koller D, Friedman N: **Probabilistic graphical models : principles and techniques**. Cambridge, Mass.: MIT Press; 2009.
21. Bishop CM: **Pattern recognition and machine learning**. New York: Springer; 2006.
22. Bishop CM: **Neural networks for pattern recognition**. Oxford,New York: Clarendon Press ;Oxford University Press; 1995.
23. Buchen L: **MISSING THE MARK**. *Nature* 2011, **471**(7339):428-432.
24. Gewin V: **Missing the mark**. *Nature* 2007, **449**(7164):770-771.
25. Antal P, Millinghoffer A, Hullám G, Falus CSaA: **A Bayesian View of Challenges in Feature Selection: Feature Aggregation, Multiple Targets, Redundancy and Interaction**. In: *JMLR Proceeding.* vol. 4; 2008: 74-89.
26. De Bie T, Tranchevent L, Van Oeffelen L, Moreau Y: **Kernel-based data fusion for gene prioritization**. *Bioinformatics* 2007, **23**(13):I125-I132.
27. Russell SJ, Norvig P: **Artificial intelligence : a modern approach**, 3rd edn. Upper Saddle River, N.J.: Prentice Hall; 2010.