Bayesi módszerek a releváns változók kiválasztásának problémájára és alkalmazásuk az orvosbiológiában 2008 okt. 1. – 2012 április 30.

Bayesian methods for the generalized feature subset selection problem and their biomedical applications

Bevezető

A nagy áteresztőképességű -genetikai, genomikai, proteomikai, metabolikai mérések új lehetőségeket nyitottak az orvosbiológiában, a személyre szabott megelőzés, diagnózis, hatóanyagok és kezelés reményével. Azonban az utóbbi évek kutatásainak egyik legfőbb eredménye annak megértése, hogy a komplex, gyakori betegségek genetikai hátterében rendkívül sok genetikai útvonal, gén és genetikai variáns érintett, amelyek egy adott életstílus és környezeti hatások mellett különbözőképpen nyilvánulhatnak meg. Szerepük megértéséhez a gyakori genetikai variánsokon és eset-kontroll megközelítésen alapuló vizsgálatokat jelentősen ki kell terjeszteni a ritka genetikai variánsokra, epigenetikai változásokra, illetve a fenotípus, az életmód és a környezet részletes leírásaira.

A megfigyeléseink „teljessé" („-omikaivá") válása, azaz a változók nagy száma azonban egyszerre jelent unikális lehetőséget a valódi okozati tényezők azonosítására, és komoly kihívást is a valós összefüggések statisztikailag megbízható felismerésére. **A projekt folyamán egy adatelemzési módszertant fejlesztettünk tovább, a Bayes háló alapú bayesi többszintű relevancia elemzést (Bayesian network based Bayesian Multilevel Analysis of relevance, BN-BMLA), amely rendszerelméleti/hálózati megközelítésen alapulva komplex fenotípusok esetén is képes feltárni a releváns változók és interakcióik részletes szerepét, azokon alapuló hipotézisek hiearchiáit [1,2,3,9,10,12,13,14].** A BN-BMLA módszertan komplex modellek felett átlagolva származtat a változók és azok egyre magasabb szintű interakcióinak a relevanciájára a posteriori valószínűségeket. Komplex fenotípusok esetén a releváns változók szerepéhez is származtatunk a posteriori valószínűséget, nevezetesen, hogy a változó gyengén (más változó által közvetítetten) vagy erősen (közvetlenül) releváns, esetleg több ponton is releváns a betegség előrehaladásában. A modellek feletti átlagolást az alkalmazott Monte Carlo módszerek párhuzamosításával oldottuk meg. A BMLA módszertan további előnye, hogy eredményei egy valószínűségi tudásbázisként is használhatóak és későbbi kísérletek optimális megtervezését is segítik, amelyeket szintén vizsgáltunk.

Grafikus valószínűségi modellek alkalmazása genetikai asszociációs vizsgálatokban

A grafikus valószínűségi modellek használata genetikai asszociációs vizsgálatokban a családfa elemzésekhez kapcsolódott, majd a genetikai variánsok kapcsoltsága miatt a tagSNP-k és a haplotípusok kezelésénél jelent meg [52-56]. A genetikai interakciók, komplex fenotípusok és életmódbeli, környezeti módosító hatások figyelembevétele miatt az utóbbi években a grafikus valószínűségi modellek, különösen az oksági kapcsolatok modellezésére alkalmas bayes hálózatok használata genetikai asszociációs vizsgálatokban egyre elterjedtebbé váltak. A kutatási projektünkben kidolgozott módszertan szisztematikus lehetőséget kínál az asszociációs és többváltozós, vagy akár oksági kapcsolatok részletes jellemzésére is.

A bayes statisztika alkalmazása genetikai asszociációs vizsgálatokban

A bayes statisztika orvosbiológiai alkalmazását kezdetben olyan általános tulajdonságok motiválták, mint a statisztikai értelemben vett kismintás esetekben történő felhasználás, és az a priori ismeretek koherens beléptetése a statisztikai következtetésbe. Az omikai vizsgálatok ezt a két irányt felerősítették, de projektünkben két további előnyét is vizsgáltuk és kihasználtuk a bayes statisztikai megközelítésnek, nevezetesen a hipotézismentes modelltulajdonság felfedezést és a bayesi adatelemzés eredményeinek a rugalmasabb utófeldolgozását és kombinálását. A statisztikai értelemben vett kismintás eset az orvosbiológiai kontextusban rendkívül nagyra növekedett változószám miatt lép fel, ami legegyszerűbb esetben a genetikai asszociációs vizsgálatokban a többszörös hipotézistesztelés problémájaként aposztrofálódik. Az a priori ismeretek felhasználásának fontossága a viszonylagosan alacsony mintaszám és komplex modellek miatt, illetve az orvosbiológiai háttértudás sokrétűsége és gazdagsága miatt fontos. A nagy áteresztőképességű, omikai mérések miatt lehetségessé vált hipotézismentes kutatás azonban a bayesi megközelítés azon előnyét is fontossá tette, hogy komplex modellek tulajdonságai kikövetkeztethetők lehetnek, annak ellenére, hogy a modellek között nincsenek dominánsak, sem nagy a posteriori valószínűségi régiók kis kiterjedéssel. Ekkor az adott adat mint feltétel meghatározza az adott modellosztályt használó konkrét elemzés során fennálló statisztikai bizonytalanságot, és az érdekes, megerősített modelltulajdonságok felismerése egy sokrétű feladatként jelenik meg. Valójában a hipotézismentes kutatási paradigma térnyerésének egy lassan felismert következménye, hogy az értelmezéssel eltöltött idő rendkívül hosszúvá válik, és a kutatásoknak és elemzéseknek az értelmezés jelenti a szűk keresztmetszetet. Emiatt projektünkben ezt a problémát több szinten is vizsgáltuk:

- a nagy változószámú bayesi relevancia elemzések eredményeinek szintaktikai aggregációit és azok vizualizációját [1,2,3,9,10,12,13,14],
- a bayesi adatelemzés eredményeinek szemantikai aggregációját egy valószínűségi tudásbázis elemeiként, amely kodifikált orvosbiológiai taxonómiákat és logikai tudásbázisokat is tartalmaz [1,2,19,39],
- a bayesi adatelemzés eredményeinek döntéselméleti elemzését, és annak felhasználását további kísérletek megtervezésében (adaptív avagy szekvenciális kísérlettervezésben) [1,2,17,21,43].

A bayesi döntéselmélet alkalmazása genetikai asszociációs vizsgálatokban

A hipotézismentes, nagy változószámú omikai vizsgálatok döntéselméleti tervezése a génexpressziós vizsgálatok esetében jelent meg elsőként. A genetikai asszociációs vizsgálatok kiértékelésének nehézségei és szekvenciális, adaptív jellege azonban felvetette a döntéselméleti keret alkalmazását a kiértékelésben is. A kutatási projektünkben mind a kísérlettervezési, mind az értelmezési fázisban megvizsgáltuk a döntéselméleti keret alkalmazását, az első esetben a kísérlet következő fázisában potenciálisan mérésre kerülő változók prioritizálása miatt, a második esetben pedig a részletesebb vizsgálatra, vagy publikálásra érdemes változók prioritizálása miatt [1,2,17,21,43].

A Bayesi többszintű relevancia elemzés és fejlődésének fázisai

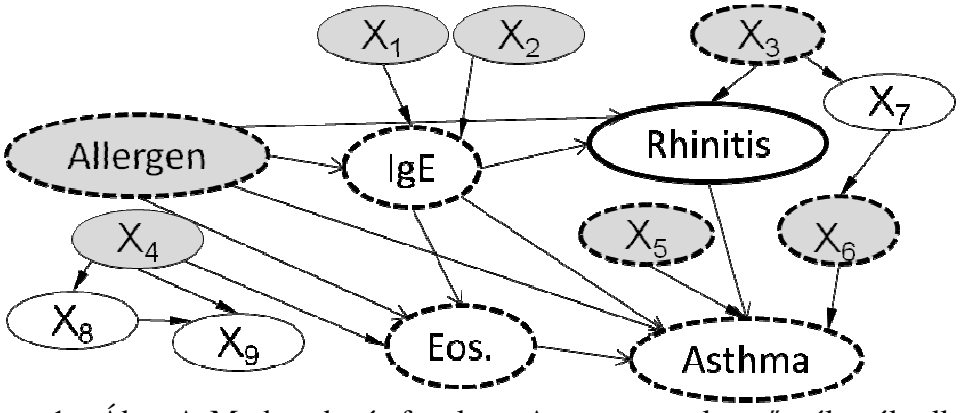
A Bayes hálókon alapuló bayesi többszintű relevancia elemzés (BN-BMLA) kulcsgondolata azon a felismerésen alapult, hogy

- a Bayes hálózatok unikális lehetőséget kínálnak függések és oksági kapcsolatok hierarchikus, egyre részletesebb leírására,
- a Bayes statisztikai keret pedig lehetővé teszi a posteriori értékek számítását ezen egymásba ágyazott hipotézisekhez.

A BN-BMLA doktori kutatásomban használt szintjei a következőek voltak [57]:

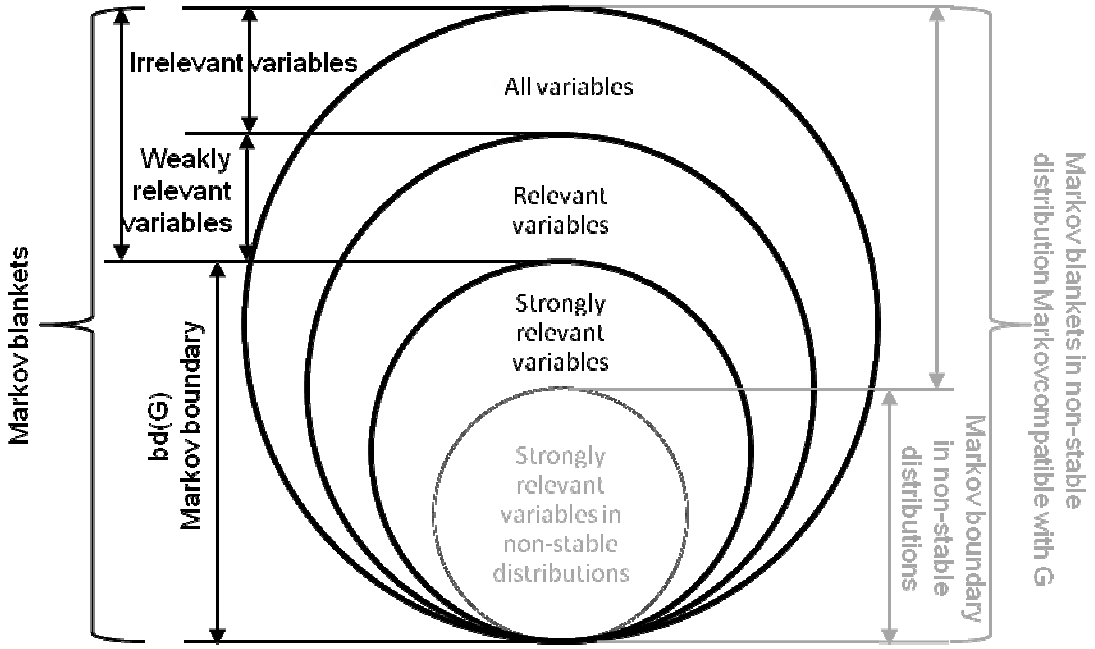
- a Markov határbeliség, ami az egyes változókhoz tartozó jellemző (egyváltozós szint),
- a Markov határhalmaz, ami több változóhoz tartozó jellemző (többváltozós szint),
- a Markov határhalmazt kifeszítő algráf, ami több változóhoz és interakcióikhoz tartozó jellemző (interakciós szint).

A BN-BMLA módszer projetkbeli kiterjesztése során nagy hangsúlyt kapott az orvosbiológiai kutatásokban 2000-től egyre inkább előtérbe kerülő genetikai variánsok hatásának vizsgálata. Ennek keretében megismerkedtünk a genetikai asszociációs kísérlettervezési és elemzési módszerekkel, különös tekintettel az egynukleotidos polimorfizmusok (SNP-k) redundanciájának következményeivel, valamint a haplotípusok szintjével. Ennek során valós alkalmazások keretein belül mértük fel a megszokott asszociációs statisztikai eszköztár és Bayes hálók biztosította gazdagabb fogalomtár különbözőségét, mint például a páronkénti asszociációs kapcsolat és a Markov határbeliség és erős relevancia viszonya. A Markov határ fogalmát Bayes hálózatokban az 1. Ábra illusztrálja (részletes leírásért ld. [3]).

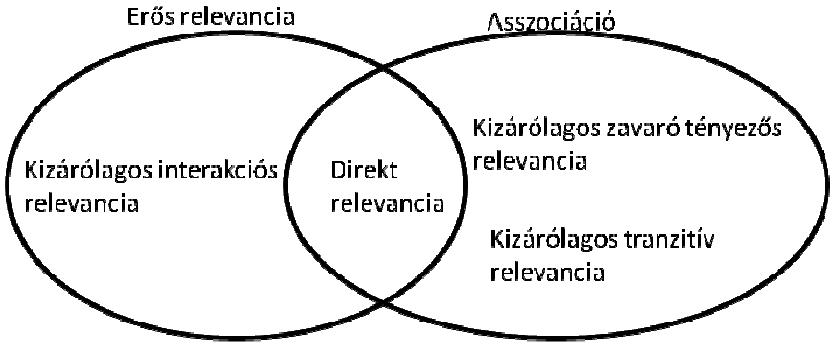


1. Ábra A Markov határ fogalma. A szaggatott keretű változók alkotják a Rhinitis változó Markov határát (részletes leírásért ld. [10]).

A Markov határbeliség az erős relevancia fogalmához kötődik, amely a Bayes hálók reprezentációs szemantikájához jól illeszkedik. Az erős és gyenge relevancia, illetve az asszociáció kapcsolatát az 2. Ábra és 3. Ábra illusztrálja (részletes leírásért ld. [3]).

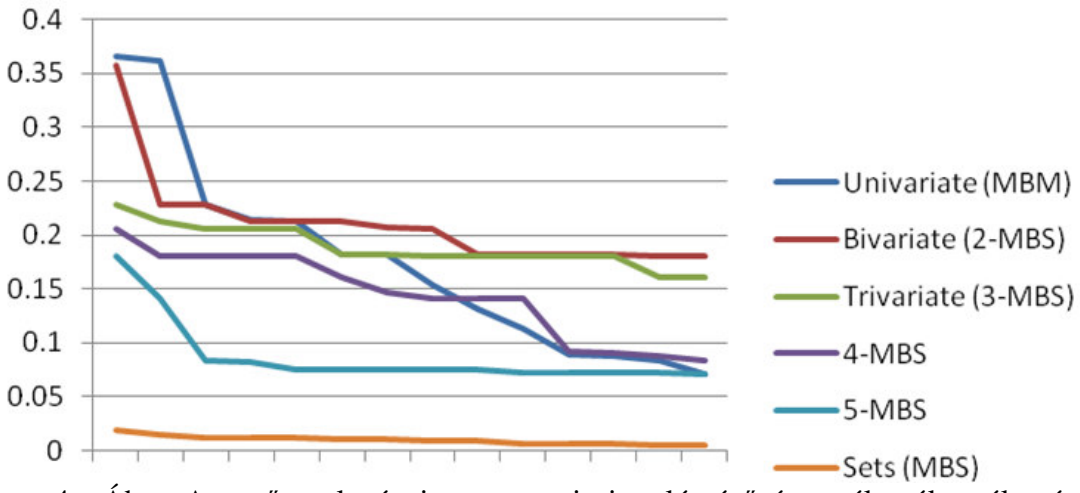


2. Ábra Változók lehetséges relevancia típusai egy adott G struktúrájú Bayes hálóval kompatibilis eloszlásokban, fekete a stabil, szürke a nem stabil eloszlások esetére vonatkozik (részletes leírásért ld. [3]).



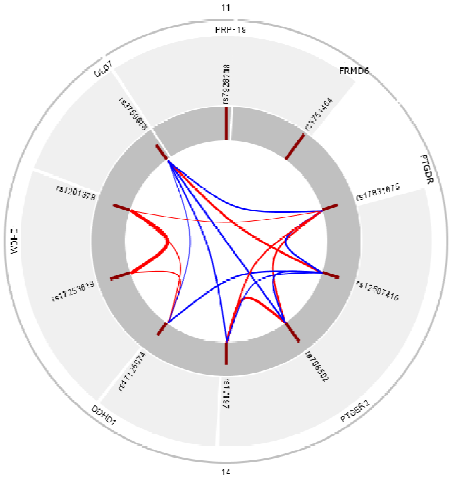
3. Ábra A páronkénti relevancia típusainak viszonya (részletes leírásért ld. [3,10]).

Az egyváltozós és többváltozós szintek közötti fokozatos átmenet támogatására bevezettük a k-subMBS tulajdonságot, ami egy k változót tartalmazó halmaznál akkor áll fenn, ha része a Markov határnak [3,9,10]. Mivel a k-subMBS számossága polinomiális, ez egy fokozatos, adatméret és változószámhoz igazított átmenetet tesz lehetővé. Nevezetesen kiszámítható az a posteriori valószínűsége, hogy változó párok, változó hármasok, és rendre változók k-sai, erősen relevánsak, lásd 4. Ábra (részletes leírásért ld. [3,9,10]).



4. Ábra Az erős relevancia a posteriori valószínűsége változók, változó párok, változó hármasok, változók k-sai esetén (részletes leírásért ld. [3,9,10]).

A többváltozós elemzések kiegészítéséhez bevezettünk egy interakció és redundancia pontszámot a k-subMBS a posteriori valószínűségei alapján, ami az interakciót strukturális dekomponálhatóság (modellbeli közös előfordulás) alapján jellemzi. A kiszámolt interakciókat és redundanciákat az 5. Ábra illusztrálja (részletes leírásért ld. [3]).



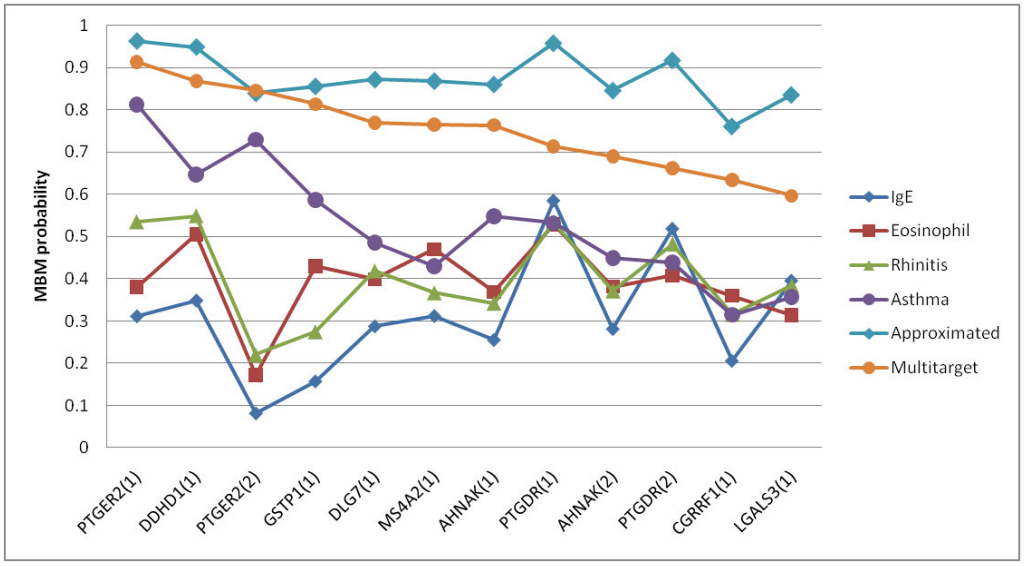
5. Ábra. Az egyes változók (SNP-k) erős relevanciáját a belső körre illesztett oszlopok jelzik, a belső szegmensek a gének és külsők a kromoszómák szerint vannak címkézve. Az átkötések vastagsága az interakció (piros) és redundancia (kék) pontszámával arányos (részletes leírásért ld. [3,12]).

A BN-BMLA elemzésből származó a posteriori valószínűségek , amelyek azt jelentik, hogy pontosan egy adott változóhalmaz az erősen releváns halmaz, ezek felhasználható az eredmények szemantikai aggregálására is. Nevezetesen például az egyes genetikai variánsokra

(egynukleotidos polimorfizmusokra, SNP-kre) vonatkozó eredményeket génekre, majd Gene Ontology kifejezésekre tudjuk aggregálni, azaz direkt módon kiszámítható, hogy milyen a posteriori valószínűséggel kapcsolódik az adott kifejezés olyan génhez vagy régióhoz, amelyben erősen releváns biomarker, például genetikai variáns fordul elő, lásd 6. Ábra (részletes leírásért ld. [1,3]).

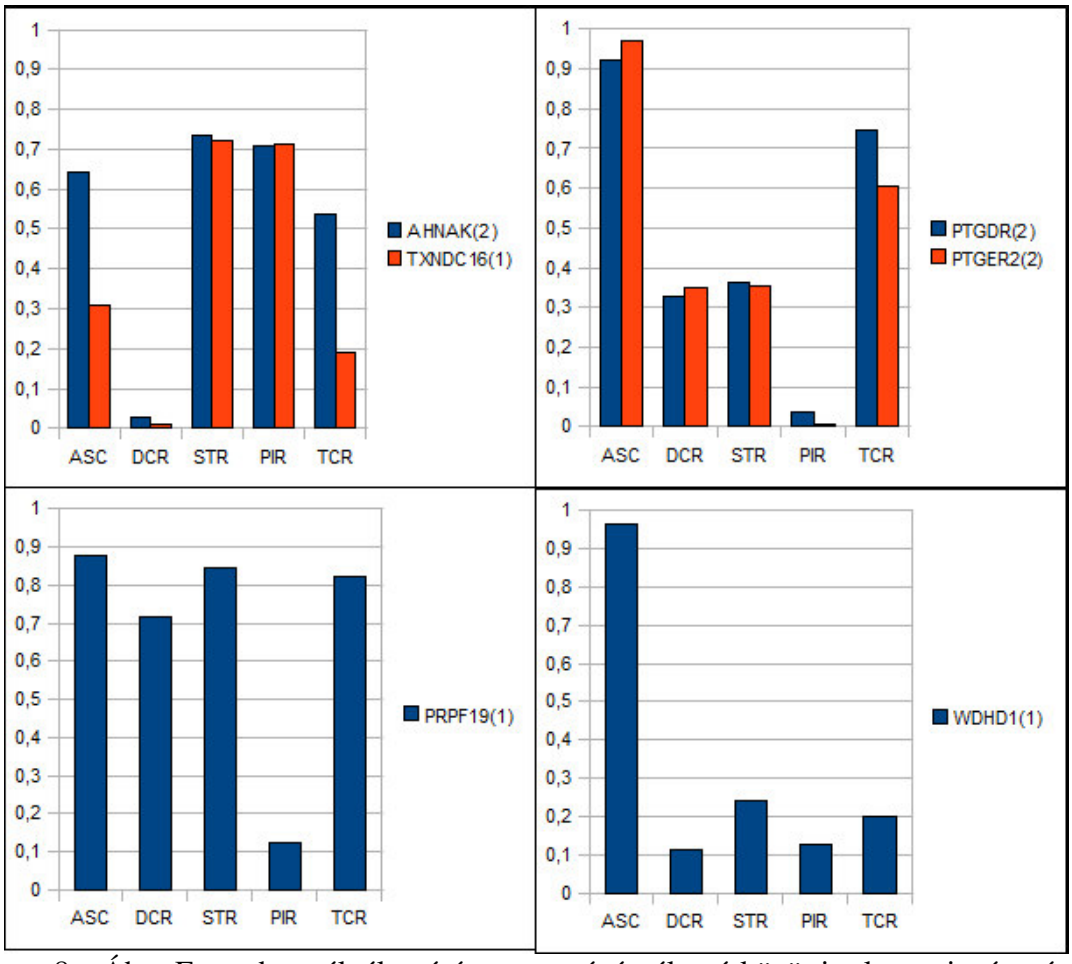
6. Ábra. Genetikai variánsok (SNP-k) relevanciájának a BN-BMLA elemzésének az eredménye Gene Ontology taxonómiára terjesztve (részletes leírásért ld. [1,3]).

Több célváltozó esetén az egyes magyarázó változók szerepe többféle is lehet, mint például „bármelyikhez”, „pontosan egyhez”, „pontosan egyhez nem”, „több mint egyhez” való erős relevancia. Ezen relációk a Bayes hálózatok reprezentációs szemantikájával hatékonyan kifejezhetők, és a Bayes statisztikai keretben ezekhez a posteriori valószínűség becsülhető, ld. 7. Ábra (részletes leírásért ld. [1,3,9,10]).



7. Ábra Az x tengelyen szereplő egyes faktorok erős relevanciájának a posteriori valószínűsége különböző célváltozókra, azok együttesére, és egyes célváltozókra vonatkozó értékeken alapuló approximáció (részletes leírásért ld. [3,10]).

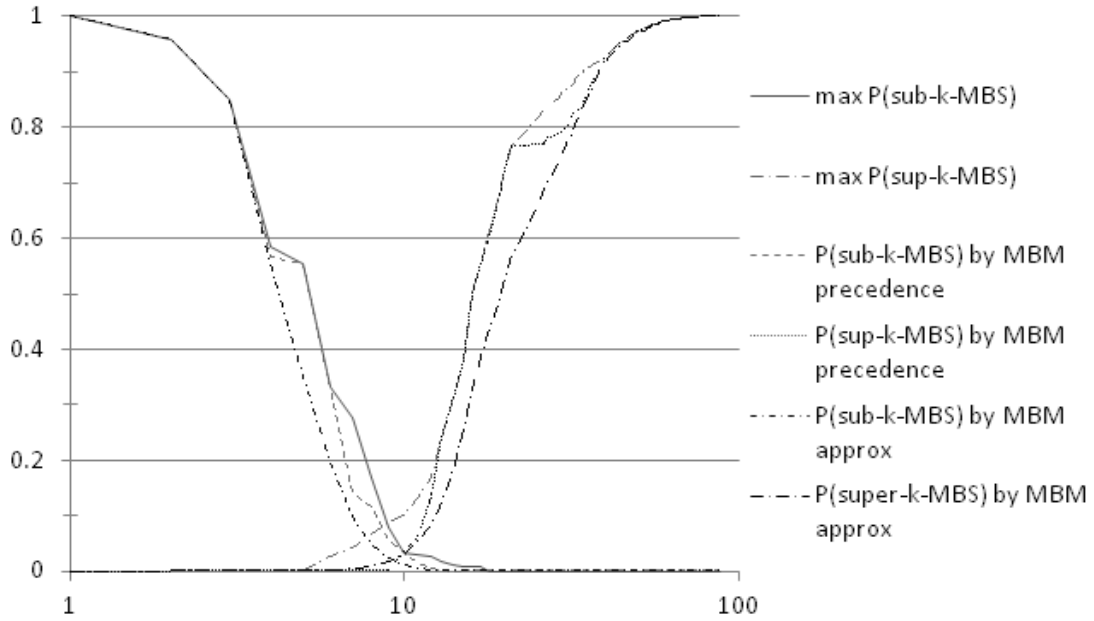
Hasonlóan, az egyes magyarázó változók szerepe is többféle lehet akár egyetlen célváltozó esetében is, ahogyan azt a 3. Ábra is illusztrálja. Ezek a relációk szintén hatékonyan kifejezhetők Bayes hálókkal, amelyek felhasználása több célváltozó esetén, különösen komplex, oksági relációkban álló fenotípusok esetén lehet hasznos (ld. 8. Ábra).



8. Ábra Egy adott célváltozó és magyarázó változó közötti relevancia típusának az a posteriori valószínűsége (Asc: páronként asszociált, DCR: direkt/feltétlen erős relevancia, PIR: feltételes erős relevancia, STR: erős relevancia, TCR: oksági tranzitív relevancia). Részletes leírásért ld. [10].

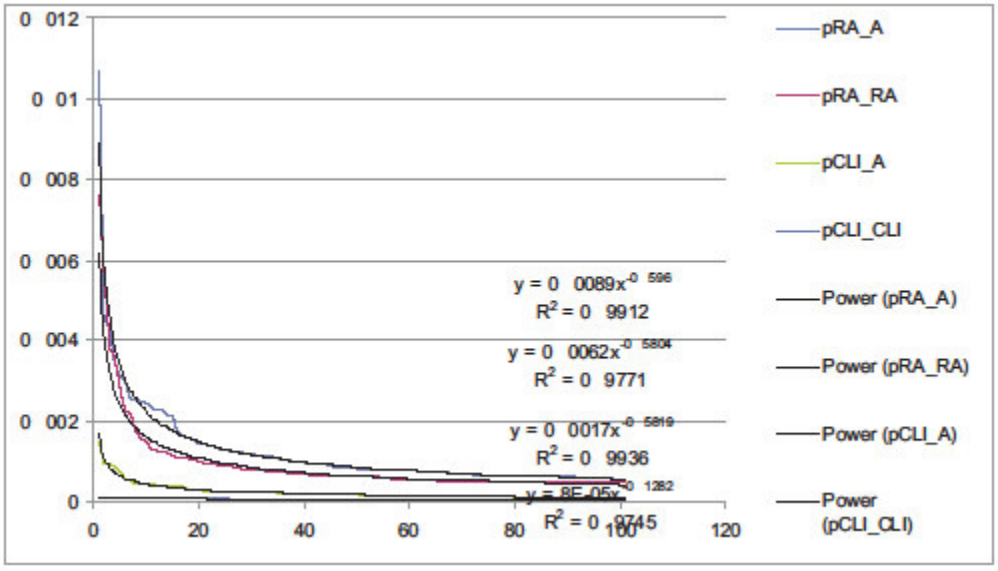
A k-subMBS mintájára bevezettük a k-supMBS fogalmát, amely akkor áll fenn k változóra, ha a kérdéses k változó tartalmazza az erősen releváns változók halmazát [14]. A k-subMBS fogalom lehetővé tette, hogy erősen releváns változók interakcióban lévő, együttes hatásuk miatt összetartozó csoportjait fokozatosan azonosítsuk be, míg a k-supMBS fogalom lehetővé tette, hogy a nem erősen releváns változókat adattól függő megerősítésük szerint fokozatosan zárjuk ki.

Máshogy fogalmazva e két fogalom lehetővé teszi, hogy fokozatosan és konzisztensen kiegyensúlyozva a bevételt és kizárást, többváltozós módon az interakciókat figyelembe egyrészt meghatározzuk az adott megbízhatósági szinten erősen releváns változókat (szükségesek) és kizárjuk a nem erősen relevánsakat (megtartva az elégségeseket). A szükséges (k-subMBS) változók és az elégséges (k-supMBS) változók mintegy közrefogják és egy eldöntetlen státuszba sorolják a maradék változókat, ld. 9. Ábra (részletes leírásért ld. [2,3,14]).



9. Ábra Az erős relevancia a posteriori valószínűségének változása. Baloldal: ahogyan egyre több változót nyilvánítunk erősen relevánsnak (k-subMBS egyre nagyobb). Jobboldal: ahogyan egyre több változót zárunk ki a potenciálisan erősen releváns változók halmazából (k-supMBS) (részletes leírásért ld. [2,3,14]).

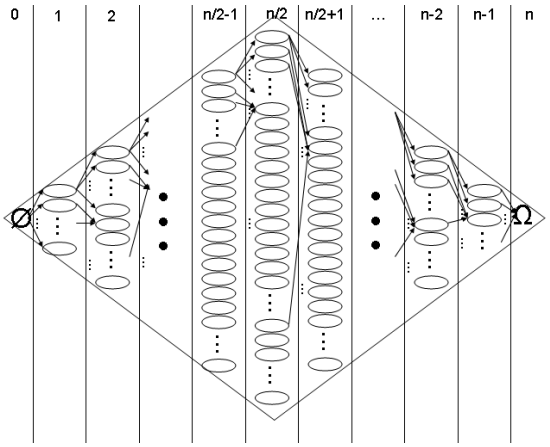
Amint a 9. Ábra is illusztrálja, az a posteriori valószínűsége az erős relevanciának, illetve annak kizárhatóságának tipikusan gyorsan csökken és nem kínál természetes küszöbértéket. A k halmazméret függvényében a maximális a posteriori értékek, a k-subMBS és k-supMBS tulajdonságokra egy olyan szinten metszik egymást, ami igazolhatóan az egyes változóhalmazok erős relevanciájának az a posteriori értékei felett van, azaz a „maximum a posteriori változóhalmaz" erős relevanciájának az a posteriori értéke felett van. Ez azonban tipikusan egy igen alacsony érték, amint azt a 10. Ábra is illusztrálja.



10. Ábra. Az erős relevencia pontos fennállásának a posteriori valószínűségei a magyarázó váltózók egyes halmazaira, különböző célváltozók és mintahalmazok esetén (részletes leírásért ld. [3,10]).

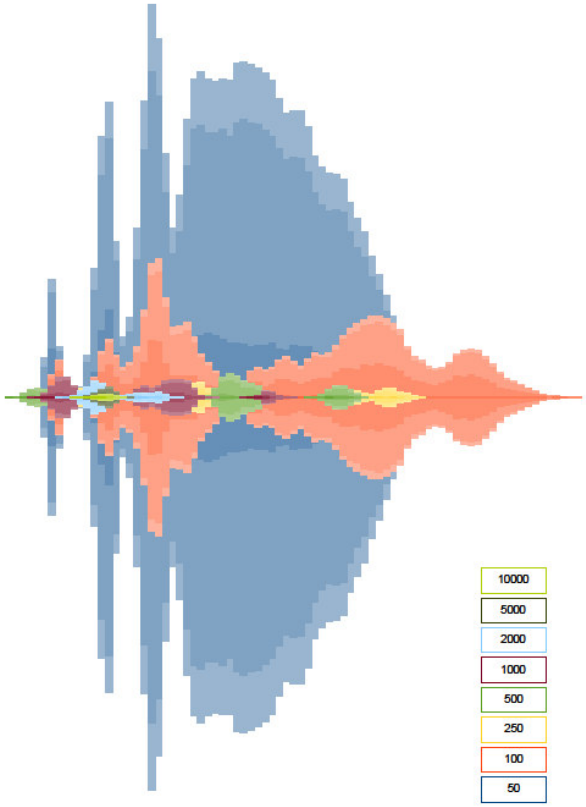
Az a posteriori értékek „elkentsége” ellenére a k-subMBS és k-supMBS fogalmak támogatást adnak erősen releváns és nem erősen releváns változóhalmazok kiválasztására, egy adott küszöbérték mellett. Az adott küszöbérték mellett érvényes erősen releváns és nem erősen releváns változóhalmazok száma azonban igen nagy lehet, n^k nagyságrendű, ami felvetette a változóhalmazok nem redundáns jelentésének kérdését, vizualizálásának, és várható hibáinak kérdését.

A változóhalmazok erős relevanciájának az a posteriori valószínűségeinek, és k-subMBS/ k-supMBS tulajdonságuk a posteriori valószínűségeinek a vizualizációjára és a változóhalmazok halmazainak a hatékony leírására vezettük be a részhalmaz hálón alapuló részhalmaz térképet (ld. 11. ábra).



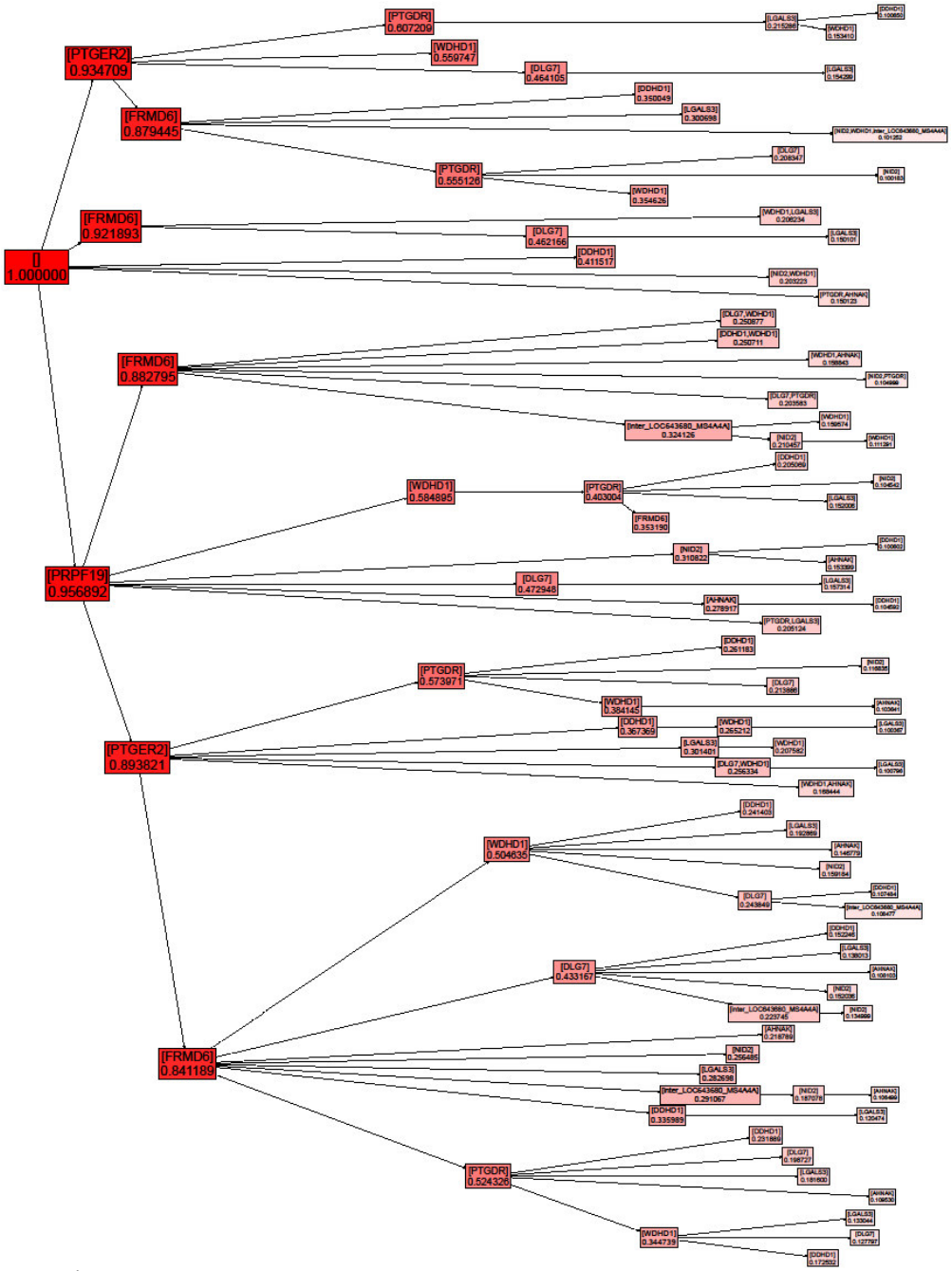
11. Ábra. A részhalmaz hálón alapuló részhalmaz térkép: a k. oszlop a k elemszámú halmazokat tartalmazza, él pedig az egy elemben (változóban) különböző halmazokat köti össze (részletes leírásért ld. [3,14]).

A részhalmaz térkép felhasználását egy változóhalmazok feletti eloszlás, az erős relevancia a posteriori eloszlásának a vizualizálására a 12. Ábra illusztrálja.



12. Ábra A változóhalmazok feletti eloszlások vizualizálása a részhalmaz térkép felhasználásával. A részhalmaz térkép egyes pontja egy adott részhalmazhoz tartoznak, azonban ez különböző elemzésekben változhat. Az egyes színek különböző adatmérethez tartozó a posteriori eloszlásokhoz tartoznak egy genetikai asszociációs elemzésből, az egyes színek sötétebb árnyalata nagyobb poszteriorokat jelent (részletes leírásért ld. [3,14]).

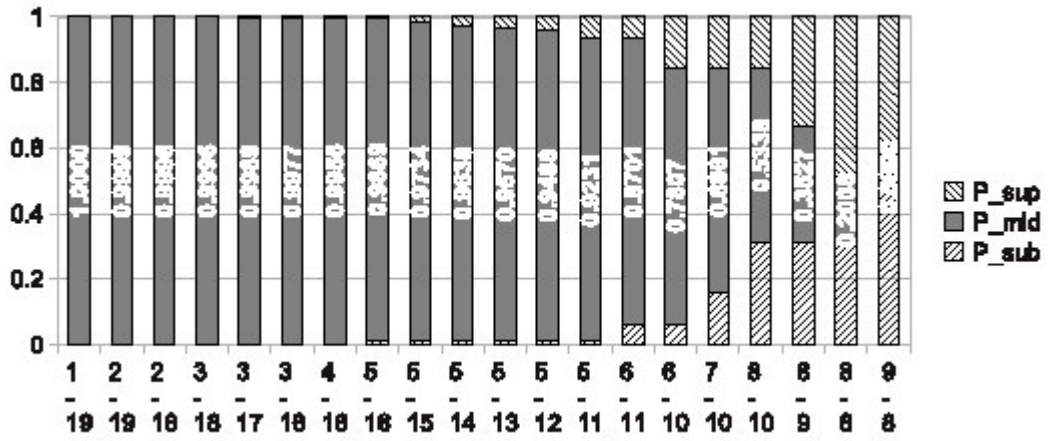
A részhalmaz háló topológiája a vizualizáláson túl unikális lehetőséget kínál azon változóhalmazok karakterizációjára, amelyekben a változók együttes erős relevanciájának az a posterior valószínűsége egy adott küszöbértéket meghalad, mivel definiálható egy olyan részhalmazokat tartalmazó (sub-relevancia) határ, amely ezeket a részhalmazokat tartalmazza. A 13. Ábra bemutat egy adott küszöbértékhez tartozó részhalmazokat tartalmazó fát, amelynek a levelei egy ilyen (sub-relevancia) határhoz tartoznak.



13. Ábra A részhalmazok k-subMBS státuszának az a posteriori valószínűségeinek az ábrázolása egy (részhalmaz relevancia) dendrogram felhasználásával. A fa csomópontjai részhalmazokat reprezentálnak, a levél csomópontjai egy adott küszöbértékhez tartozó sub-relevancia határt jelölnek ki (részletes leírásért ld. [3,14]).

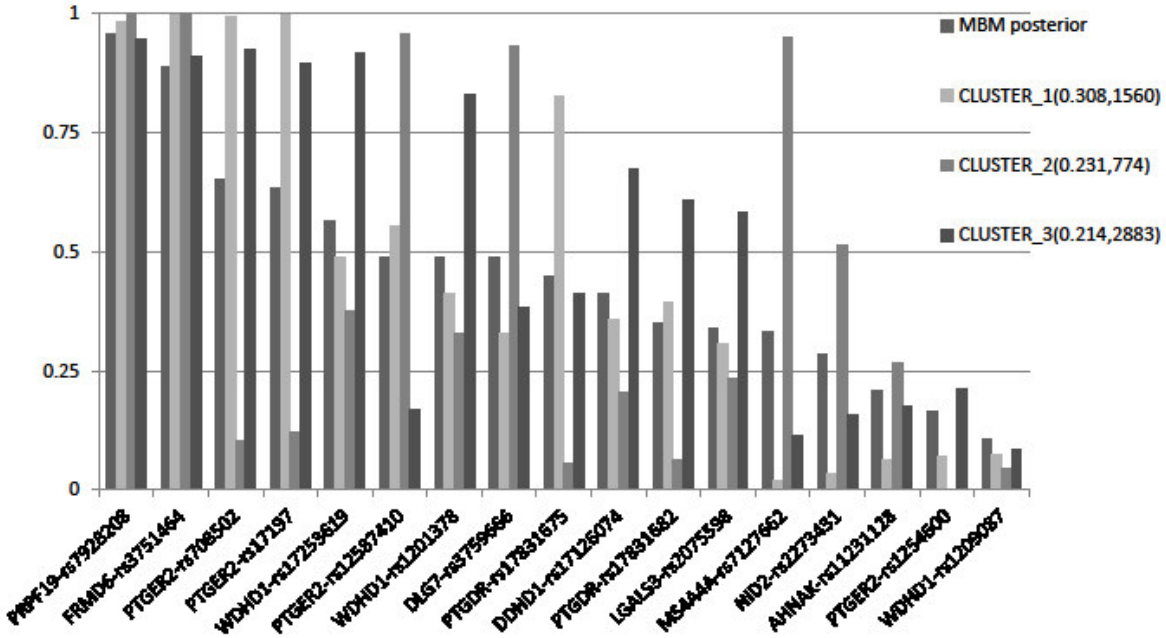
A részhalmaz háló topológiáján alapuló sub-relevancia határ analógiájára bevezettük a sup-relevancia határt is, amely a kizárhatósághoz tartozik. A sub-relevancia és a sup-relevancia határok viszont amellett, hogy a nagy a posteriori valószínűséggel erősen releváns és nem erősen

releváns változók halmazai határolják le, közrefogják a legvalószínűbb változóhalmazokat is. A részhalmaz térkép ezen felhasználása a verziótér valószínűségi általánosításának fogható fel (ld. 14. ábra).



14. Ábra A sub-relevancia és sup-relevancia határok által közrefogott változóhalmazok összesített a posteriori értéke, amikor a határokat kiegyensúlyozottan és egyesével változtatjuk a változók egy rögzített sorrendje esetén (részletes leírásért ld. [3,14]).

A részhalmaz térképbeli sub-relevancia határ kizárja a túlságosan egyszerű, a sup-relevancia határ pedig kizárja a túlságosan komplex változóhalmazokat (modelleket). A két határ közrefogja az adattal kompatibilis vagy máshogyan fogalmazva az adat által adott küszöbnél jobban megerősített változóhalmazokat (modelleket). A részhalmaz térkép szerinti hasonlóságban/távolságban azonban gyakran több lokális maximuma is van az erős relevancia a posteriori valószínűségeinek, amelyek akár alternatív vagy parallel mechanizmusokhoz, útvonalakhoz is tartozhatnak, más és más erősen releváns változóhalmazzal. Az értelmezésük támogatására bevezettük az erős relevancia feltételes a posteriori valószínűségének a fogalmát, amely esetben a feltétel részhalmazok szemantikailag összetartozó halmaza (klasztere) vagy a részhalmazok felett definiált hasonlóság/távolság szerint összetartozó halmaza (klasztere). Az egyes változók erős relevanciájának és feltételes relevanciájának az a posteriori értékeit és azok viszonyát a 15. Ábra illusztrálja.



15. Ábra Az erős relevancia és feltételes relevanciák a posteriori értékei, ahol a feltételben szereplő klasztereket a változóhalmazok közötti távolság és a posteriori értékeik szerint határoztuk meg (részletes leírásért ld. [3,14]).

A nagy számú változó és nagyságrendekkel nagyobb számosságú részhalmazaik és interakciós modelljeik a frekventista/hipotézis tesztelési statisztikai keretben a többszörös hipotézistesztelési problémához vezetnek. A bayes statisztikai keret normatív megoldást kínál erre a problémára, amely komplex, teljes tárgyterületi modellek használatára és azok a posteriori eloszlásának marginalizációjára épül. Ezen kettős a BN-BMLA módszertannak is alapja. Természetesen a statisztikai alapprobléma ebben a bayesi keretben is jelen van: a nagy változószám, modell komplexitás, és viszonylagosan kevés adat egy „lapos" poszteriorban fog megjelenni. A tanulás sebességét a különböző komplexitású relevancia szinteken szisztematikusan vizsgálatuk. A többszörös hipotézistesztelési probléma azonban több más módon is megjelenik, amelyeket a kutatásunkban szintén vizsgáltunk [2,14]. (1) Egy nem kellően felismert, nyílt kutatási kérdés a Monte Carlo módszerek használata nagy számú jegy a posteriori valószínűségének becslésére. Ekkor a céltól függetlenül, ami lehet hipotézistesztelés, például, hogy a posterior egy adott küszöbértéknél nagyobb/kisebb–, vagy konfidencia intervallum számítás, a becsült értékek nagy számát figyelembe kell venni, ami természetesen független hatás a változók nagy számának a hatásától. Ennek kezelésére klasszikus és „false discovery rate" (FDR) alapú korrekciós módszereket vizsgáltunk. (2) A Monte Carlo becslések többszörös hipotézistesztelési problémájának a viszonylagosan kevéssé kutatott státuszával szemben az a priori valószínűségek megválasztásának problematikája egy központi kérdés a bayes statisztika támogatói és kritikusai számára is. A kutatásunkban ezt a kérdést a többcélváltozókra vonatkozó relevancia altípusok esetén vizsgáltuk, nevezetesen véletlen gráf modellekben szimulációkkal vizsgáltuk, hogy különböző a priori és közelített a posteriori él és algráf valószínűségek esetén, milyen valószínűséggel jelennek meg többcélváltozókra vonatkozó relevancia típusok. (3) Továbbá

kidolgoztunk egy többváltozós, bayesi döntéselmélet alapú megközelítést a felfedezések hibáinak, benne az FDR, kontrollálására, és egy hozzátartozó „d-értéket", ami a „q-érték" egy ezen keretbeli analógja [2,14].

A többváltozós, bayesi döntéselméleti megközelítés nem csupán az eredmények értelmezésében és konklúziók, például tudományos közlések, elérésében használható fel, hanem követő kísérletek megtervezésében is. A priori ismeretek, korábbi adatok önálló és együttes felhasználását is vizsgáltuk a kísérlettervezés területén, különösen az adaptív, szekvenciális kísérlettervezés tekintetében. [1,2,17,21,43].

A kifejlesztett módszerek párhuzamosított implementációja gyakran több alternatív módon is, kódszinten és alkalmazásszinten is elkészült, amelyek az alkalmazási lehetőségeket nagyban segítették. A projekt folyamán a következő területeken kezdtük el alkalmazni a kifejlesztett metodológiát.

1. Asztma és allergia genetikai háttere (Szalai Csaba, Falus András, SE Genetikai, Sejt- és Immunbiológiai Intézet) [1,10].
2. Leukémia genetikai háttere (Szalai Csaba, Erdélyi Dániel, Falus András, SE Genetikai, Sejt- és Immunbiológiai Intézet) [5,12].
3. Autoimmun betegségek genetikai háttere és egyéb immunomikai alkalmazások (Buzás Edit, Falus András, Edit, SE Genetikai, Sejt- és Immunbiológiai Intézet) [4,6,11,15].
4. Függőségek és kognitív döntési mechanizmusok genetikai háttere (Sasvári Mária, SE Orvosi Vegytani, Molekuláris Biológiai és Patobiokémiai Intézet) [8,9].
5. Impulzivitás pszichogenetikája (Székely-Veres Anna, ELTE, Pszichológiai Intézet) [9].
6. Preoperatív petefészekrák diagnosztika (Dirk Timmerman, K.U.Leuven).
7. Gyógyszermellékhatások és felhasználásuk adatfúzión alapuló gyógyszerkutatásokban (Mátyus Péter, SE Gyógyszerkutatási és Gyógyszerbiztonsági Centrum) [13].
8. Vesedialízissel kapcsolatos genetikai faktorok (Kiss István, Dél-Budai Nephrológiai központ).
9. Koraszülött halálozást és szövődményeket befolyásoló faktorok (Szabó Miklós, SE I. Gyermekklinika).
10. A depresszió és szorongás genetikai háttere (Juhász Gabriella, Baghdy György, SE Gyógyszerhatástani Intézet).

A Bayes háló alapú bayesi többszintű elemzés fogalomkészletét a projektbeli kutatásunkban az alábbiakkal bővítettük egy teljesebb, gyakorlatban is használható metodológiává:

- a részleges (k-subMBS) és fedőleges (k-supMBS) erős relevancia fogalmai [3,14],
- a strukturális posztererior alapú interakciós és redundancia pontszámok [1,2,3,12],
- az erős relevancia relációt tartalmazó tudásbázisok valószínűségi szemantikája („adatelemzési tudásbázisok") [1,2,3,19],
- specifikus erős relevancia típusai több célváltozó esetén [1,2,3,10],
- globális oksági, standard páronkénti asszociációs és erős relevancia altípusai [1,2,3,10],
- a részhalmaz háló alapú részhalmaz térkép, határhalmazok, valószínűségi verziótér a részhalmazok relevanciájára, és részhalmazok klaszterei és feltételes relevancia fogalma [13],

- döntéselméleti támogatás az elemzések kiértékelésére, különös tekintettel a nagy számú változó esetén fellépő hibás felfedezések kontrolljára [2,13],
- komplex, különösen több célváltozóhoz tartozó relevancia viszonyok előfordulási valószínűségének vizsgálata véletlen gráf alapú szimulációkkal,
- nagy számú hipotézis a posteriori valószínűségének Monta Carlo becslése esetén szükséges (többszörös hipotézistesztelési) korrekció,
- döntéselméleti támogatás szekvenciális kísérletek tervezésére [1,2,17,21,43],
- kódszintű és alkalmazásszintű párhuzamosítások, amelyek a projektben beszerzett tesztelésre alkalmas infrastuktúra mellett a Genagrid projekt nagyszámítógépes infrastruktúráján is elérhetőek. A következő címen tölthető le egy keretszoftver http://redmine.genagrid.eu/projects/bayeseyedownload/wiki.

A módszertan publikálása még jelenleg is tart, 4 cikk és könyvfejezet elbírálás alatt van, és több kézirat is beadás, vagy ismételt beadás előtt áll. Következő lépésként az adatfúzió, különös tekintettel a sok tekintetben komplementer kernel fúziós módszerek integrálását tervezzük, amely megoldást kínál majd több kemoinformatikai és gyógyszerkutatási kérdésre is.

Könyvfejezetek

1. P. Antal, A. Millinghoffer, G. Hullám, G. Hajós, Cs. Szalai, A. Falus: **A bioinformatic platform for a Bayesian, multiphased, multilevel analysis in immunogenomics**, in Bioinformatics for Immunomics, Ed.: M.N.Davies, S.Ranganathan, D.R.Flower, Springer, 2010, 3:157-185
2. P. Antal, András Millinghoffer, Gábor Hullám, Gergely Hajós, Csaba Szalai, András Falus: **Bayesian, systems-based, multilevel analysis of associations for complex phenotypes: from interpretation to decisions**, in Probabilistic graphical models in genetics, genomics, postgenomics, eds.: Christine Sinoquet, Raphael Mourad (submitted)

Folyóiratcikkek

3. P. Antal, A. Millinghoffer, G. Hullám, Cs. Szalai, A. Falus: **A Bayesian View of Challenges in Feature Selection: Feature Aggregation**, Multiple Targets, Redundancy and Interaction, JMLR: Workshop and Conference Proceedings 4, 74-89
4. T. G Szabó, R. Palotai, P. Antal, I. Tokatly, L. Tóthfalusi, O. Lund, Gy. Nagy, A. Falus, E. I. Buzás: **Critical role of glycosylation in determining the length and structure of T cell epitopes- As suggested by a combined in silico systems biology approach**, Immunome Res. 2009 Sep 24;5:4
5. Semsei A.F, Antal P, Szalai Cs.: **Strengths and weaknesses of gene association studies in childhood acute lymphoblastic leukemia,** Leuk Res. 2010 Mar;34(3):269-71
6. S. Srivastava, P. Antal, J. Gál, G. Hullám, A.F. Semsei, G. Nagy, A. Falus, E. I. Buzás: **Lack of evidence for association of two functional SNPs of CHI3L1 gene(HC-gp39) with rheumatoid arthritis,** Rheumatol Int. 2011 Aug;31(8):1003-7
7. G. Hullám, P. Antal, Cs. Szalai, A. Falus: **Evaluation of a Bayesian model-based approach in GA studies**, JMLR Workshop and Conference Proceedings, 8:30-43, 2010.
8. Marx Péter, Arany Ádám, Rónai Zsolt, Antal Péter, Sasvári-Székely Mária: **Az oxitocinreceptor genetikai variabilitásának in silico elemzése**, Neuropsychopharmacologia Hungarica, 2011. XIII . évf. 3. Szám, 139-144
9. Gabor Varga, Anna Szekely, Peter Antal, Peter Sarkozy, Zsofia Nemoda, Zsolt Demetrovics, Maria Sasvari-Szekely: **Independent effects of serotonergic and dopaminergic polymorphisms on trait impulsivity**, Am J Med Genet B Neuropsychiatr Genet. 2012 Apr;159B(3):281-8
10. Ildikó Ungvári, Gábor Hullám, Péter Antal, Petra Sz. Kiszel, András Gézsi, Éva Hadadi, Viktor Virág, Gergely Hajós, András Millinghoffer, Adrienne Nagy, András Kiss, Ágnes F. Semsei, Gergely Temesi, Béla Melegh, Péter Kisfali, Márta Széll, András Bikov, Gabriella Gálffy, Lilla Tamási, András

Falus,Csaba Szalai**: Evaluation of a partial genome screening of two asthma susceptibility regions using Bayesian network based Bayesian multilevel analysis of relevance,** PLoS One. 2012;7(3):e33573

11. Zsuzsanna Pál; Péter Antal; Sanjeev K Srivastava; Gábor Hullám; Ágnes Félné Semsei; János Gál; Mihály Svébis; Györgyi Soós; Csaba Szalai; Sabine André; Elena Gordeeva; György Nagy; Herbert Kaltner; Nicolai V Bovin; Mária J Molnár; András Falus; Hans-Joachim Gabius; Edit Irén Buzás: **Non-synonymous single nucleotide polymorphisms in genes for immunoregulatory galectins: association of galectin-8 (F19Y) occurrence with autoimmune diseases in a Caucasian population**, Biochimica et Biophysica Acta-General Subjects, 2012, in press
12. Orsolya Lautner-Csorba, András Gézsi, Ágnes F. Semsei1, Dániel J. Erdélyi, Péter Antal, Géza Schermann, Nóra Kutszegi, Katalin Csordás, Márta Hegyi, Gábor Kovács, András Falus, Csaba Szalai: **Candidate gene association study in pediatric acute lymphoblastic leukemia evaluated by Bayesian network based Bayesian multilevel analysis of relevance**, (submitted to BMC Medical Genomics)
13. Ádám Arany, Peter Antal, Bence Bolgár, Balázs Balogh, Péter Mátyus**: A New Strategy for Repositioning: Drug Prioritization by Adaptive Fusion of Medicinal Chemical, Target and Side-Effect-Related Information**, (submitted to Current Medicinal Chemistry)
14. Peter Antal, Andras Millinghoffer, Gabor Hullam, Gergely Hajos, Andras Gezsi, Peter Sarkozy, Yves Moreau, Csaba Szalai, Andras Falus: **Deep Bayesian characterization of relevant factors and interactions in allergy and asthma using subset map and multivariate FDR control**, (under revision)
15. Sanjeev Srivastava, Péter Antal, Mercédesz Mazán, Mária Pásztói, Ilona Újfalussy, Bernadett Rojkovich, Judit Kelemen, Ildikó Ungvári, Csaba Szalai, Tamás Gáti, Gábor Hullám, György Nagy, András Falus, Edit I Buzás: **Combined analysis of two single nucleotide polymorphisms of the glucuronidase gene shows strong association with rheumatoid arthritis,** The Journal of Rheumatology (under revision)

Konferenciacikkek

16. P. Antal, A. Millinghoffer, G. Hullám, Cs. Szalai, A. Falus: **A Bayesian View of Challenges in Feature Selection: Feature Aggregation, Multiple Targets, Redundancy and Interaction**, ECML/PKDD, Workshop on New challenges for feature selection in data mining and knowledge discovery 2008 (FSDM08), Antwerp, JMLR: Workshop and Conference Proceedings 4, 74-89
17. P Antal, G Hajós, P Sárközy: **Bayesian network based analysis in sequential partial genome screening studies**, MODGRAPH, June 8., 2009, Nantes, France
18. G. Hullám, P. Antal, Cs. Szalai, A. Falus: **Evaluation of methods in GA studies: yet another case for Bayesian network**, Machine Learning in System Biology 2009 (MLSB09), Sept 5-6, Ljubljana, Slovenia, Proc. of the Third International Workshop, 35-44
19. P. Sarkozy, P. Marx, A. Millinghoffer, G. Varga, A. Szekely, Zs. Nemoda, Zs. Demetrovics, M. Sasvari-Szekely, P. Antal: **Bayesian data analytic knowledge-bases for genetic association studies**, in Working Notes of the Workshop on Probabilistic Problem Solving in BioMedicine (ProBioMed'11), the European Conference on Artificial Intelligence in Medicine (AIME'11), 13th European Conference on Artificial Intelligence in Medicine, AIME'11, 2011, pp 55-67

Előadások (konferenciakiadványban összefoglalóval)

20. P.Antal, G. Hajós, G.Hullám, A.Millinghoffer Cs.Szalai and A. Falus: **A bioinformatic platform for a model-based, knowledge-rich study design and Bayesian analysis of partial genome screening studies**, Magyar Biokémiai Egyesület 2008. évi Vándorgyűlése (Szeged, 2008. augusztus 31-szeptember 03.)
21. P.Antal, G. Hajós, G.Hullám, A.Millinghoffer Cs.Szalai and A. Falus: **Adaptive Sequential Partial Genome Screening Studies: a Case Study in Asthma**, Human Genome Variation Society, Human Variome Project, Towards Establishing Standards, 22nd May 2009, Vienna, Austria
22. Csaba Szalai, Ágnes F. Semsei, Ildikó Ungvári, Petra Kiszel, Péter Antal, András Falus: **Investigation of the genomic background of obesity using single nucleotide polymorphism analysis in candidate genes**, 2nd Central European Congress on Obesity, October 1-3, 2009, Budapest, Hungary
23. Peter Antal, G. Hullam, Cs. Szalai, A. Falus: **A "Bayesian Tour" from Study Design through Data Analysis to Clinical Decision Support**, 41st ANNUAL SCIENTIFIC MEETING OF THE HUNGARIAN MEDICAL ASSOCIATION OF AMERICA, Sarasota, Florida, October 25 - October 30, 2009

24. Ágnes F. Semsei, Dániel Erdelyi, Ildikó Ungvári, Edit Cságoly, Márta Z. Hegyi, András Millinghoffer, Gábor Hullám, Péter Antal, András Falus, Gábor Kovács, Csaba Szalai: **The role of ATP-binding cassette transporter polymorphisms on antracycline induced cardiotoxicity in childhood acute lymphoblastic leukemia**, Semmelweis Egyetem PhD Tudományos Napok 2009, 2009. március 30-31. Budapest SE NET
25. Petra Sz. Kiszel, Ágnes F. Semsei, Ildikó Ungvári, Adrienne Nagy, Márta Széll, Béla Melegh, Péter Kisfali, Péter Antal, Gábor Hullám, András Falus, Csaba Szalai: **Screening for susceptibility genes of asthma on chromosome 11 and 14** , Allergy & Asthma Symposium: Bridging Innate and Adaptive Immunity, May 28-29, 2009 Bruges, Belgium
26. Ágnes F. Semsei, Dániel Erdélyi, Ildikó Ungvári, Edit Cságoly, Márta Z. Hegyi, András Millinghoffer, Gábor Hullám, Péter Antal, András Falus, Gábor Kovács, Csaba Szalai: **The role of ATP-binding cassette transporter polymorphisms on antracycline induced cardiotoxicity in childhood acute lymphoblastic leukemia**, PHARMACOGENOMICS & PERSONALIZED MEDICINE, September 12 - 15, 2009, Wellcome Trust Conference Centre, Hinxton, UK
27. Peter Antal, G. Hullam, Cs. Szalai, A. Falus: **A "Bayesian Tour" in Omics,** '3rd HUNGARIAN-SINGAPOREAN WORKSHOP on SYSTEMS BIOLOGY and COMMUNICATION SYSTEMS, Budapest, March 29-30, 2010
28. Temesi Gergely, Antal Péter, Szalai Csaba, Falus András, Hajós Gergely, Gézsi András, Sárközy Péter, Marx Péter: **Több forrásból származó omikai adatok és információk kiértékelése hibrid tudásbázis technológiával**, Magyar Biokémiai Egyesület Budapesten 2010. évi Vándorgyűlése, Magyar Biokémiai Egyesület 2010. évi Vándorgyűlése, Budapest, 2010. augusztus 25-28
29. Pál Zsuzsanna, Antal Péter, Millinghoffer András, Hullám Gábor, Pálóczi Krisztina, Tóth Sára, Hans-Joachim Gabius, Falus András, Buzas Edit, Molnár Mária Judit: **A galectin-1 és interleukin-2 receptor β új haplotípusának asszociációja autoimmun myasthenia gravis-szal**, MAGYAR HUMÁNGENETIKAI TÁRSASÁG VIII. KONGRESSZUSA, 2010. szeptember 2-4., Debrecen
30. Antal Péter, Sárközy P., Hajós G., Szalai Cs.: **Polimorfizmusok kontra ritka variánsok – hipotézismentes omikától a tudásgazdag adatelemzésig**, MAGYAR HUMÁNGENETIKAI TÁRSASÁG VIII. KONGRESSZUSA, 2010. szeptember 2-4., Debrecen
31. Antal P: **Génprioritizálástól az oksági következtetésig**, XIII. Magyar Neuropszichofarmakológiai Kongresszus, Neuropsychopharmacologia Hungarica 2010. XII. évf. Suppl. 1., 2010, október 9-13, Tihany
32. Antal Péter(1), Szalai Csaba (2), Falus András(3): **Gyakori és ritka variánsok a gyakori betegségekben – helyzetkép és konklúziók statisztikai szemmel**, A Magyar Személyre Szabott Medicina Társaság 1. Konferenciája, 2010 október 28, Budapest
33. Antal, P. Marx, A. Millinghoffer, G. Hullam, I. Ungvary, Cs. Szalai, A. Falus: **Bayesian fusion of heterogeneous signs for biomarker and pathway discovery**, Capita Selecta in Complex Disease Analysis (CSCDA 2010), Leuven (Belgium), 25-27 August 2010
34. Arany Ádám, Bolgár Bence, Balogh Balázs, Antal Péter, Mátyus Péter: **Hatóanyagok újrapozícionálása in silico információforrások fúzionáltatásával**, Magyar Kémikusok Egyesülete MKE 1. Nemzeti Konferencia, Sopron, 2011
35. Antal Péter, Arany Ádám, Bolgár Bence, Balogh Balázs, Mátyus Péter: **Bioinformatics and Medicinal Chemistry: Drug repositioning**, 3rd Balaton Course on Medicinal Chemistry, Balatonszemes, Sept. 15-17, 2011
36. Antal P., Hajós G., Millinghoffer A., Hullám G., Marx P., Sárközy P., Gézsi A., Temesi G., Szalai Cs., Falus A: **Oksági biomarkerek kutatásának fogalmai és keretei: asszociációtól a relevanciáig, családfától az oksági hálózatokig**, A Magyar Személyre Szabott Medicina Társaság 2. Konferenciája, 2011, szeptember, Eger
37. Arany Ádám, Bolgár Bence, Hajós Gergely, Balogh Balázs, Mátyus Péter, Antal Péter: **Gyógyszer újrapozícionálás a személyre szabott medicina korszakában: a mellékhatások kihasználásától a mellékhatások kontrollálásáig**, A Magyar Személyre Szabott Medicina Társaság 2. Konferenciája, 2011, szeptember, Eger
38. Sárközy Péter, Antal Péter, Rónai Zsolt: **Az OpenArray rendszer alkalmazása a pszichogenetikában. Bioinformatikai analízis**, XIV. Magyar Neuropszichofarmakológiai Kongresszus, Neuropsychopharmacologia Hungarica, 2011, október, Tihany

Poszterek (konferenciakiadványban összefoglalóval)

39. P. Antal, A. Millinghoffer, G. Hullám, Cs. Szalai, A. Falus: **BysCyc: A Bayesian Logic for the Integrative Analysis of Knowledge and Data in Genetic Association Studies**, The Second International Workshop on Machine Learning in System Biology (MLSB08), September 13-14 2008, Brussels, Belgium
40. G. Hajós, P. Antal, Y. Moreau, Cs. Szalai, A. Falus: **Model-based SNP set selection in study design using a multilevel, sequential, Bayesian analysis of earlier data sets**, The Second International Workshop on Machine Learning in System Biology (MLSB08), September 13-14 2008, Brussels, Belgium
41. Petra Sz. Kiszel, Ágnes F. Semsei, Ildikó Ungvári, Adrienne Nagy, Márta Széll, Béla Melegh, Péter Kisfali, Péter Antal, Gábor Hullám, András Falus, Csaba Szalai: **Screening for susceptibility genes of asthma on chromosome 11 and 14**, Allergy & Asthma Symposium: Bridging Innate and Adaptive Immunity, May 28-29, 2009 Bruges, Belgium
42. P. Antal, A. Millinghoffer, Cs. Szalai, A. Falus: **On the Bayesian applicability of graphical models in genome-wide association studies**, Machine Learning in System Biology 2009 (MLSB09), Sept 5-6, Ljubljana, Slovenia
43. G. Hajós, P. Antal, Y. Moreau, Cs. Szalai, A. Falus: **Variable Pruning in Bayesian Sequential Study Design**, Machine Learning in System Biology 2009 (MLSB09), Sept 5-6, Ljubljana, Slovenia
44. P. Antal, P. Sárközy, Z.Balázs, P. Kiszel, A. Semsei, Cs. Szalai, A. Falus: **Averaging over measurement and haplotype uncertainty using probabilistic genotype data**, Machine Learning in System Biology 2009 (MLSB09), Sept 5-6, Ljubljana, Slovenia
45. Gézsi A., Antal P., Hajós G., Millinghoffer A., Szalai Cs., Falus A.: **Bayesi módszerek felskálázásának vizsgálata teljes genom asszociációs elemzésekben**, Magyar Biokémiai Egyesület 2010. évi Vándorgyűlése, Budapest, 2010. augusztus 25-28.
46. Antal P., Hajós G., Millinghoffer A., Hullám G., Balázs Z., Gézsi A., Temesi G., Sárközy P., Félné Semsei Á, Ungváry I., Németh Zs., Virág V., Hadadi É., Debreceni G., Kormos K., Lévai P, Szalai Cs., Falus A.: **A Genagrid bioinformatikai szolgáltatás genetikai asszociációs elemzések támogatására** (Genagrid bioinformatic services for the support of genetic association studies), Magyar Biokémiai Egyesület 2010. évi Vándorgyűlése, Budapest, 2010. augusztus 25-28.
47. Hajós Gergely, Antal Péter, Szalai Csaba, Falus András: **Bayes-háló alapú adaptív kísérlettervezés parciális genomszűrési kísérletekhez** (Bayes-network based adaptive study design for partial genome association studies), Magyar Biokémiai Egyesület 2010. évi Vándorgyűlése, Budapest, 2010. augusztus 25-28.
48. Marx, P. Antal, G. Hullám, A. Millinghoffer, Cs. Szalai, A. Falus: **Sorrendi fúziós algoritmusok alkalmazása klinikai genomikában**, Magyar Biokémiai Egyesület 2010. évi Vándorgyűlése, Budapest, 2010. augusztus 25-28.
49. Sárközy P., Antal P., Balázs Z., Sasvári M., Szalai Cs., Falus A.: **A bizonytalan genotípusos és in silico rekonstruált haplotípusos adatok elemzésének lehetőségei modell átlagolással**, Magyar Biokémiai Egyesület 2010. évi Vándorgyűlése, Budapest, 2010. augusztus 25-28.
50. Antal P., Sárközy P., Balázs Z., Sasvári M., Szalai Cs., Falus A.: **Haplotype- and Pathway-based Aggregations for the Bayesian Analysis of Rare Variants**, Machine Learning in System Biology 2010 (MLSB10), Oct. 15-16
51. Antal P., Gézsi A., Hajós G., Millinghoffer A., Szalai Cs., Falus A.: **On the applicability of Bayesian univariate methods as filters in complex GWAS analysis**, Machine Learning in System Biology 2010 (MLSB10), Oct. 15-16

Külső hivatkozások

52. Mourad R, Sinoquet C, Leray P.: **Probabilistic graphical models for genetic association studies**, Brief Bioinform. 2012 Jan;13(1):20-33.
53. Pearl J: **Causality : models, reasoning, and inference**. Cambridge, U.K. ; New York: Cambridge University Press; 2000.
54. Koller D, Friedman N: **Probabilistic graphical models : principles and techniques**. Cambridge, Mass.: MIT Press; 2009.
55. Friedman N: **Inferring cellular networks using probabilistic graphical models**. *Science* 2004, **303**(5659):799-805.
56. D. J. Balding: Handbook of Statistical Genetics, Wiley&Sons

57. P. Antal: Integrative Analysis of Data, Literature, and Expert Knowledge, Ph.D. dissertation, K.U.Leuven, D/2007/7515/99, 2007