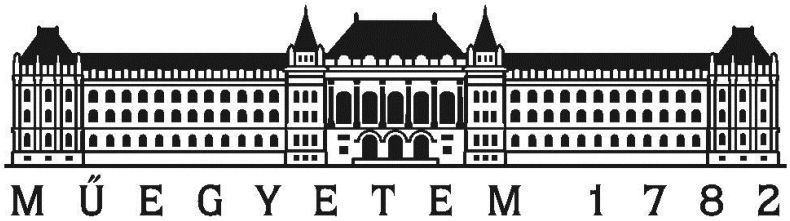


BUDAPESTI MŰSZAKI ÉS GAZDASÁGTUDOMÁNYI EGYETEM
VILLAMOSMÉRNÖKI ÉS INFORMATIKAI KAR
MÉRÉSTECHNIKA ÉS INFORMÁCIÓS RENDSZEREK TANSZÉK

Orvosi és genetikai információk fuzionálása döntéstámogató modellekben örökletes mellrák területén

TDK dolgozat

Készítette:
Gável Dorina
BSc. IV. évfolyam

Konzulens:
Dr. Antal Péter
docens

2014.10.22.

Tartalomjegyzék

1. Bevezető

Napjaink orvostudományának egyik legnagyobb kihívása a rákos megbetegedések gyógyítása. Más és más terápiás kezelés szükséges jóindulatú daganatos megbetegedés és rosszindulatú tumor esetén. A gyógyulás valószínűsége függ a beteg általános kockázati tényezőitől, ilyen például a nem, a kor és az etnikum, továbbá jelentős a genetikai kockázat és a családtól öröklött hajlam szerepe is.

A hatékony gyógyítás érdekében manapság az egyedi terápiás kezeléseket részesítik előnyben, figyelembe véve a beteg genetikai hátterét, ugyanis így célzott területen tudják a megfelelő gyógymódot alkalmazni. Az orvosok számára azonban nem egyszerű feladat az adott genetikai háttérhez megfelelő terápia kiválasztása. Ezen probléma megoldására számos kockázati modellt hoztak létre, amelyek döntéstámogatást is nyújtanak.

Önálló laboratóriumom és szakmai gyakorlatom során napjaink egyik legjelentősebb rákos betegségével, a mellrákkal foglalkoztam, amely világszerte a nők 12,3%-át érinti [1]. Korai diagnózis felállítással megfelelő terápia alkalmazható, melynek következtében hatékonyan kezelhető a mellrákos megbetegedés. A mellrák kialakulásának tényezőinek vizsgálata során azt tapasztaltam, hogy habár sok gén és variánsa jelentősen növelheti a mellrák kockázatát, a legismertebb kockázati modellekben csak a legnagyobb rizikójú géneket vizsgálták. Ebből a gondolatból kiindulva olyan modellek létrehozásának módszereit kerestem, amelyek az aktuális kutatási annotációkat feldolgozva feltérképezhetik az egyes gének és variánsok közötti kapcsolatokat. Ezen új és a korábbi kockázati modellek összehasonlítása gyakorlati szempontból is hasznos eredményeket hozhat.

Kockázati modellek létrehozására, vizualizálására és vizsgálatára szolgáló hatékony eszközök a valószínűségi hálók. Bayes-hálók segítségével különböző ok-okozati összefüggéseket vizsgálhatunk, és döntési fákkal segíthetjük a megfelelő terápia kiválasztását. A gének és variánsaik közötti kapcsolatok felderítése azonban nem egyszerű feladat és felveti azt a kérdést, hogy az általam létrehozott valószínűségi hálókban tényleg minden él releváns-e, nem pedig false-pozitív vagy false-negatív eredmény. A kutatást kiegészítettem egy harmadik tényezővel, a génmutációk fehérjeútvonalakra gyakorolt hatásának vizsgálatával, amely segítheti az orvosokat a megfelelő terápia kiválasztásában.

Jelen dolgozatomban felvázolom a fent ismertetett, vizsgálandó hierarchiát, amely alapján új valószínűségi hálókat hozok létre. Ezen modell újszerű, a genetikai háttér mélyebb felderítésére és eddig nem ismert összefüggések kutatására szolgál. A Bayes-hálókat az általam fejlesztett szoftverrel, a PAGEVA-val hoztam létre. Statisztikai alapon, az annotációk feldolgozásával felderítem a gráf lehetséges csomópontjait, melyek között az éleket a csomópontok adott annotációban való együttes előfordulásának valószínűségével súlyozom. A szoftver bemenetére több annotációs forrást is

adhatunk, amelyekből az élek és csomópontok megkeresése után a szoftver megvizsgálja az élek valószínűségi értékeit és kiszűri a kevésbé relevánsakat.

A PAGEVA szoftver működését és eredményeit a mellrákkal kapcsolatos Pubmed absztrakciókra alkalmazva mutatom be. A vizsgálat során olyan cikkek absztraktjait használtam fel, amelyek 2010 és 2014 között jelentek meg, továbbá tartalmazzák a „breast cancer" kulcsszót. Ezen feltételek mellett 78209 darab absztrakciót vizsgáltam meg.

A dolgozatomban bemutatott szoftveremmel elvégezhetjük több statisztikai forrás együttes feldolgozását, azaz több valószínűségi háló összekapcsolását és együttes értékelését. A PAGEVA felépítése lehetővé teszi más betegségek genetikai modellezését is, ugyanis a valószínűségi hálóban csak a kialakítandó hierarchia kötött, a vizsgálandó gének, variánsok és útvonalak, azaz a csomópontok a felhasznált annotációktól függenek.

A dolgozat vázlatos felépítése:

- *2. fejezet:* Bayes-hálók rövid ismertetése
- *3. fejezet:* a mellrák kialakulásának kockázati tényezőinek és modelleknek rövid áttekintése. Az új, vizsgálandó hierarchia felvázolása, összehasonlítása az eddigi kockázati modellekkel. A kutatáshoz használt variánsok, gének rövid ismertetése.
- *4. fejezet:* a PAGEVA szoftver szerkezetének és működésének bemutatása.
- *5. fejezet:* a PAGEVA szoftverrel létrehozott hálók kiértékelése.
- *6. fejezet:* kitekintés, a szoftver bővítési lehetőségeinek ismertetése, további területeken történő alkalmazása.

2. Valószínűségi hálók

A valószínűségi hálók – más néven Bayes-hálók – a dolgozatom és kutatásom alapjai. A Bayes-hálók olyan adatstruktúrák, melyek lehetőséget biztosítanak valószínűségi változók közötti függőségek leírására és bármely együttes valószínűség-eloszlás függvény tömör megadására. A valószínűségi hálók olyan irányított, körmentes gráfok, melyekre az alábbi feltételeknek kell teljesülniük:

- *csomópontok:* valószínűségi változók egy halmaza, melyek lehetnek diszkrétek és folytonosak is
- *irányított élek:* összeköt csomópontokat, amely szülő-gyermek kapcsolatot reprezentál
- *csomópontokhoz tartozik:* $P(X_i|Szülők(X_i))$ feltételes valószínűség-eloszlás, ami számszerűen megadja a szülők hatását a csomóponti változóra.

Az egyes változók eloszlásait feltételes valószínűségi táblákkal (FVT) reprezentálhatjuk, melyekben minden sor az egyes csomóponti értékek feltételes valószínűségét írja le az adott sorhoz tartozó szülői feltétel esetén (elsősorban diszkrét eloszlásokhoz használjuk). A szülői feltétel a szülői csomópontok egy lehetséges kombinációja, amely egy elemi eseményt testesít meg [2].

A valószínűségi hálók egy együttes valószínűségi eloszlásfüggvény leírása. Ahhoz, hogy gráfként kezelhessük, szükség van az eloszlásfüggvény dekompozíciójára, amelyet az alábbi szorzatalak biztosít:

$$\mathrm{P}(x_1 \dots x_n) = \prod_{i=1}^{N} P(x_i|szülők(X_i))$$

ahol a szülők(x_i) a Szülők(X_i)-ben szereplő változók adott értékeinek együttesét jelöli. Így az adott változó feltételes valószínűségi táblázatának bejegyzései a szülő csomópontok feltételes valószínűségi táblák megfelelő elemeinek egy szorzata [2].

A dekomponálást – más néven faktorizálást - minden egyes együttes valószínűségre alkalmazva, megfelelő sorszámozás esetén az alábbi összefüggést kapjuk, amelyet láncszabálynak neveznek:

$$\mathrm{P}(x_1 \dots x_n) = \prod_{i=1}^{N} P(x_i|x_{i-1}, \dots x_1)$$

A láncszabályban szereplő szorzótényezők alapján már építhető egy Bayes-háló, ugyanis ismertek a feltételes valószínűségi táblákhoz szükséges valószínűség-eloszlások és paraméterek. A láncszabály azonban csak abban az esetben egy helyes reprezentáció, ha minden csomópont feltételesen független a csomópontot sorrendezésben őt megelőzőktől [2].

A Bayes háló értelmezhető egy függetlenségi modellként is, amely tartalmazza a feltételes függőségeket, azaz

$$M_p = \{I_1(X_1|Y_1|Z_1), I_2(X_2|Y_2|Z_2)\ldots\},$$

ahol $I_p(X|Y|Z)$ jelöli, hogy X és Y függetlenek Z feltétellel p eloszlásban. Ezzel a megközelítéssel azonban nem garantált a gráffal való reprezentálhatóság. A Markov-feltételek teljesülése azonban biztosítja a Bayes háló létrehozhatóságát:

- *sorrendi Markov-feltétel:*

$$\forall i = 1..n : I\left(X_{\pi(i)} \middle| Pa(X_{\pi(i)}) | \{X_{\pi(j)} | j < i\} \backslash Pa(X_{\pi(i)})\right)$$

  π a struktúra egy topologikus rendezése, I(X|Y|Z) reláció X feltételes függetlensége Z-től Y feltétellel.

- *lokális Markov-feltétel* G szerint: bármely változó független nem-leszármazottaitól, feltéve szüleit.

- *globális Markov-feltétel* G szerint:

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

  azaz ha x független y-tól feltéve z-t, vagyis z d-szeparálja x-et y-tól a G gráfban.

A Markov feltételek és a faktorizálás követelménye a Bayes háló szempontjából ekvivalens, azaz bármely teljesülése biztosítja a valószínűségi háló létrehozhatóságát [3].

A valószínűségi háló oksági értelmezést is kap, ugyanis minden irányított él reprezentálható úgy, mint ok-okozati összefüggés, azaz a szülő csomópont a gyerek csomópont eseményét okozza, a szülőnek közvetlen befolyása van a gyerekre. Ez a szemlélet segíti a következtetési eljárások tervezését [2].

A fentiek alapján a valószínűségi hálók egy duális struktúrát reprezentálnak, melynek részei:

- függetlenségi térkép/oksági modell, amely a DAG szerkezetből adódik - az egyes csomópontok kapcsolatainak meghatározásával és korlátozásával foglalkozik.
- paraméteres szemlélet, amely az együttes feltételes eloszlásból következik - a feltételes valószínűségi táblák paramétereinek megadásával.

Jelen dolgozatomban mindkét szemlélet megjelenik:

- a PAGEVA szoftverrel az oksági modellt megalkotása, a háló gráf alapú struktúráját létrehozva
- a korábban vizsgált kockázati modellek és szakirodalmak alapján a háló csomópontjainak megválasztása

3. Problématerület

Ebben a fejezetben ismertetem a mellrák legfontosabb kockázati és genetikai tényezőit, melyek kapcsán felmerült egy új, hierarchikus modell létrehozása és kutatása. Az új modell jelentőségét néhány korábbi kockázati modellel összehasonlítva szemléltetem.

3.1 A mellrák kockázati és genetikai tényezői

A mellrák a prosztatarák után a második leggyakrabban előforduló rákos megbetegedés [1]. Elsősorban a nőket érinti, azonban napjainkban egyre több férfinál diagnosztizálják. A betegség kialakulásának tényezői a két nem esetén eltérő, melyek közül jelen dolgozatomban csak a nők mellrák kockázati tényezőit és modelljeit ismertetem.

A kockázati modellek két nagyobb csoportba oszthatóak: általános kockázati modellekre és BRCA1/2 génmutációkat felhasználó modellekre. Az általános kockázati modellek a páciens családi kórtörténetét és általános jellemzőit (pl. kor, etnikum stb.) használják fel, míg a genetikai tényezőket tartalmazó modellek a mellrák kockázatát jelentősen növelő BRCA1 és BRCA2 génmutációk bekövetkezésének tényezőit vizsgálja.

3.1.1. Általános kockázati modellek

a.) Gail modell

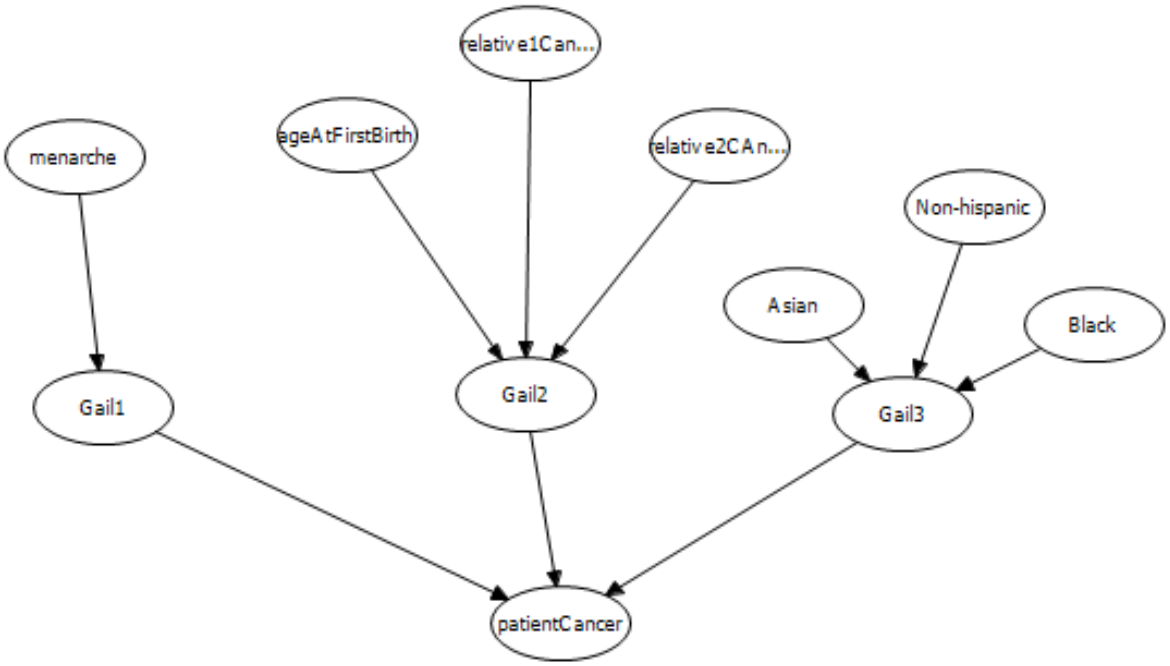
Az általános kockázati modellek közül a legismertebb, széles körben alkalmazott modell a Gail-modell, amely Dr. Mitchell Gail professzorról kapta a nevét. A modell a Breast Cancer Detection Demonstration Project (BCDDP) eredménye, amely a Breast Cancer Risk Assessment Tool szoftver alapjául is szolgál. A Gail-modell kezdetben csak a beteg korát, saját és elsőfokú rokonainak kórtörténetét használta fel, azaz hogy volt-e a betegnek vagy a családtagjának valaha mellrákos megbetegedése [4,5]. A modellt később kiegészítették további tényezőkkel is, mint például a beteg első menstruációjának időpontja, a páciens kora az első élő gyermek születésekor, a korábbi mell biopsziák száma. A kutatást először csak angol fehér nőkön végezték, később kibővítették afrikai, kaukázusi, ázsiai és fekete nőkre is, azaz fontos tényezővé vált a beteg etnikuma is. A Gail-modell szerint minél idősebb a beteg, annál nagyobb a mellrák kialakulásának kockázata [6].

Önálló laboratóriumom során ehhez a modellhez három statisztikai eredményt használtam fel, amelyek az alábbi kockázati tényezőket vizsgálták:

- Páciens életkora
- Életkor a menstruáció kezdetekor
- Életkor az első gyermek élveszülésekor

- Mellrákkal érintett elsőfokú rokonok száma
- Rassz/etnikum
- A páciens rokonainak mellrákkal rendelkezésének valószínűsége (0, 1 vagy legalább 2 rokon rendelkezik-e mellrákkal)

A fenti tényezőkből építettem egy valószínűségi hálót, amely reprezentálja a statisztikákban vizsgált összefüggéseket.



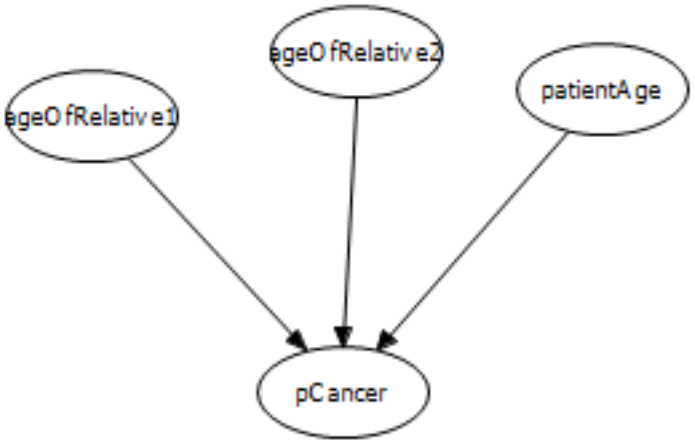
1. ábra: Gail-modell statisztikai eredményei alapján épített valószínűségi háló

Az 1. ábrán megfigyelhető a felhasznált statisztikai eredmények egy modellezése. A gráf baloldalán a páciens első menstruációjának életkorából számítja ki a mellrák kialakulásának valószínűségét (Gail1). A középső ág a beteg korát az első élő gyermek szülésekor és a páciens rokonainak mellrákkal rendelkezésének valószínűségét használja fel a mellrák kialakulás valószínűségének kiszámításához (Gail2) [4]. A harmadik ága a hálónak az etnikum függvényében vizsgálja a beteg mellrák kockázatát (Gail3) [6]. A három különböző statisztikából kiszámolt eredményt a patientCancer változóban egyesítettem, mely során noisy-OR kapuval súlyoztam az egyes alhálók eredményeinek relevanciáját a célcsomópontra.

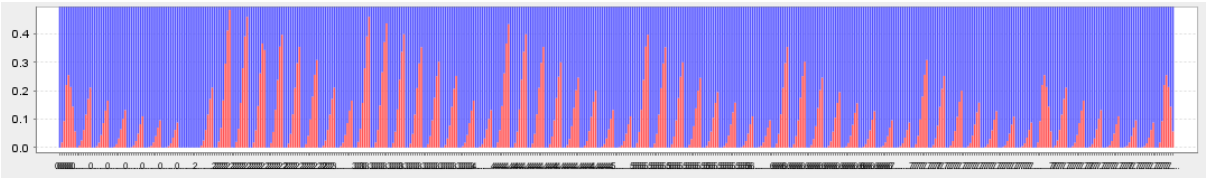
b.) Claus modell

Ezen modell kapcsán a kutatást olyan személyeken végezték, akiknél legalább az egyik rokonnál már diagnosztizáltak korábban mellrákot. A vizsgálat során felhasznált kockázati tényezők:

- a páciens kora (patientAge változó)
- mellrákban érintett rokon kora a betegségének diagnosztizálásakor (ageOfRelative változók)[7]



2. ábra: Claus modell statisztikáját reprezentáló valószínűségi háló. A pCancer változó adja meg a páciens mellrák kialakulásának valószínűségét.



3. ábra: Az általam létrehozott és felparaméterezett hálóból P(patientCancer) változó eloszlása a három paraméter függvényében: Egy „periódus” az egyik elsőfokú rokon életkorához tartozik, egy „tüske” a másik elsőfokú rokon életkorához és 1 oszlop a páciens életkorához kapcsolódik. A rokonok életkora balról jobbra növekszik 10 éves időintervallumokkal, 20 és 80 éves kor között. Az első periódus a 20 évnél fiatalabb rokonokra vonatkozik, azonban abban a korban még nem diagnosztizáltak még mellrákot, így abban az esetben a mellrák általános kockázatának adatait használtam fel.

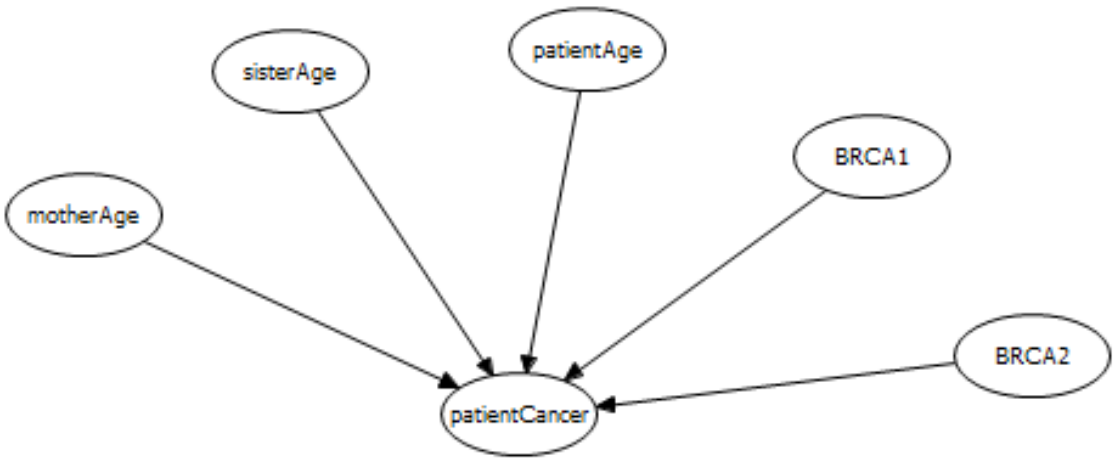
A 3. ábrából is megfigyelhető az a megállapítás, miszerint minél később diagnosztizálják az egyes családtagoknál a mellrákot, annál kisebb a valószínűsége, hogy a páciensnél is kialakul a mellrákos megbetegedés.

3.1.2. BRCA1 és BRCA2 génmutációkat vizsgáló kockázati modellek

A modellek másik nagy csoportja a BRCA1 és BRCA2 gének mutációját figyelembe vevő modellek, ugyanis számos kutatás összefüggést talált aBRCA1/2 gének mutációja és a mellrák kialakulása között. Napjaink orvoslása a személyre szabott kezelések irányába tart, továbbá a genetikai vizsgálatok költsége egyre kisebb, ezért a kutatások a genetikai hatások felmérése felé fordultak. Fontos továbbá, hogy a gének mutációja általában több betegség kialakulásának kockázatát is növeli, pl. a BRCA1 és BRCA2 gén mutációja jelentősen növeli a petefészekrák és a prosztatarák kialakulásának valószínűségét. Jelen dolgozatomban ezen modellek közül a BOADICEA modellt és a Myriad laboratórium eredményeit foglalom össze.

a.) BOADICEA modell

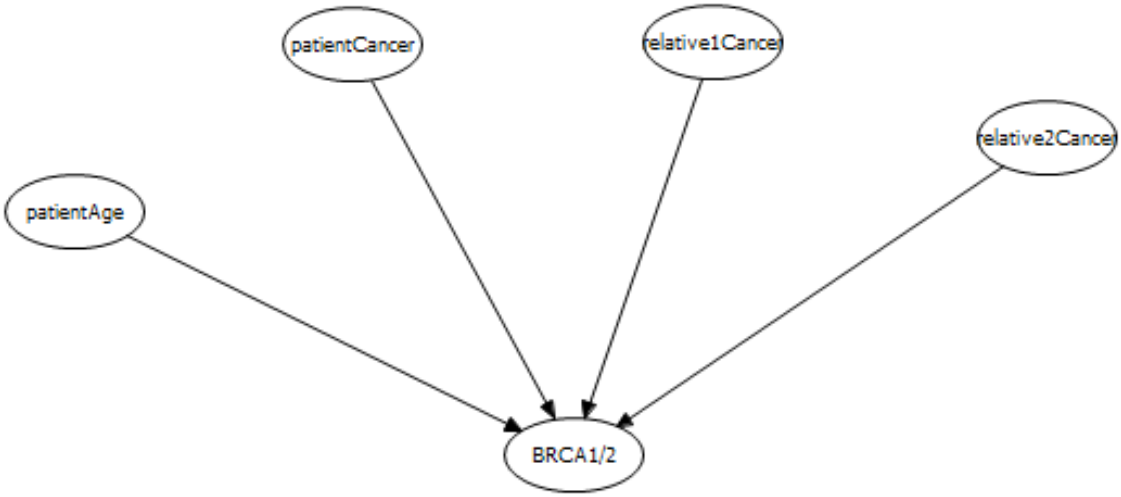
A BOADICEA modell (Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm) a BRCA1 és BRCA2 gének mutációjának következtében vizsgálj a mellrák kialakulásának valószínűségét a páciens korának és családi kórtörténete függvényében. A mutáció hordozásának valószínűsége annál nagyobb, minél fiatalabb a beteg, továbbá ha az édesanyánál és/vagy lánytestvérnél már diagnosztizáltak korábban mellrákot. A rokonok mellrák diagnosztizálásának éve is jelentős, ugyanis minél később alakul ki valamely családtagnál a mellrák, annál kisebb a génmutáció jelenlétének valószínűsége [8].



4. ábra: a BOADICEA modell eredményeit felhasználó valószínűségi háló. Ennél a modellnél a BRCA1 és BRCA2 génmutáció jelenléte feltételezett, továbbá a páciens édesanyjának és lánytestvérének kórtörténetét is felhasználták a kutatáshoz [8].

b.), Myriad

A Myriad Genetikai Laboratórium kutatási eredménye a BRCA1/2 gének mutációjának valószínűségét ábrázolja a család kórtörténete és a páciens kora alapján, ugyanakkor itt megjelennek azok az eshetőségek is, melyek szerint nem volt mellrákos megbetegedés a családban. Ebbe a kutatásba továbbá bevették a petefészekrák vizsgálatát is, amely a BRCA1 és a BRCA2 gének mutációját szintén előidézheti. A valószínűségi hálómba nem vettem fel, de végeztek kutatásokat a férfi mellrák kialakulására is a fent ismertetett tényezőket figyelembe véve [9].



5. ábra: a Myriad laboratórium kutatási eredményei alapján létrehozott valószínűségi háló.

A fenti ábrán az egyes valószínűségi változók jelentése:

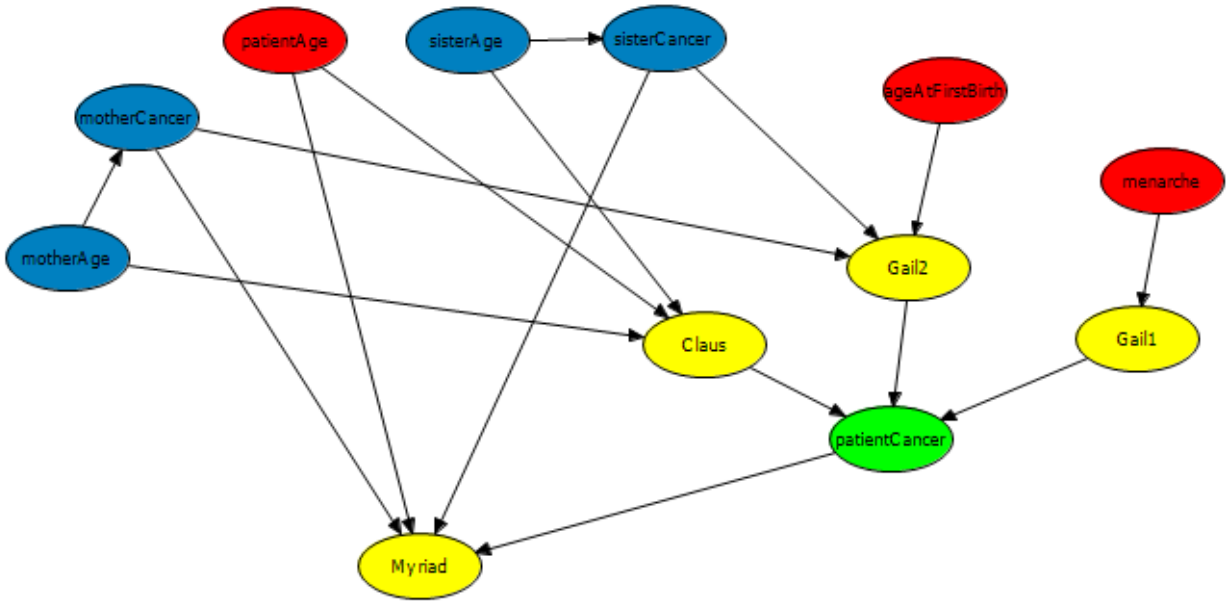
- *patientAge:* a páciens kora a mellrák diagnosztizálásakor
- *patientCancer:* a páciens mellrákkal rendelkezésének valószínűsége
- *relativeCancer:* a beteg elsőfokú rokonainak mellrákkal rendelkezésének valószínűsége
- *BRCA1/2:* a BRCA1 és BRCA2 gének együttes, törlődéses mutációjának valószínűsége

A fenti változók alapján megfigyelhető, hogy a Myriad laboratórium fordítottan végzett kutatásokat, azaz nem azt vizsgálta, hogy a BRCA1 és BRCA2 gének mutációja mekkora valószínűséggel növeli a mellrák kialakulásának kockázatát, hanem a mellrák megléte előidézhet-e génmutációt.

3.1.3. Összetett, fuzionált modell

Az 3.1.1. és a 3.1.2. fejezetekben ismertetett modellekből Önálló laboratóriumom során létrehoztam parametrikus módszerrel egy összetett, fuzionált modellt. A modelleknél ismertetett kockázati tényezők között kerestem azonosakat, melyek mentén összekapcsolhattam a különböző statisztikai forrásokból származó eredményeket. Az összetett modellhez felhasznált modellek:

- Claus modell
- Myriad
- Gail-modell



6. ábra: az általam létrehozott összetett, fuzionált modell. A patientCancer változó reprezentálja a mellrák kialakulásának kockázatát, melybe noisy-OR segítségével súlyoztam és bekötöttem a Claus és a Gail modell által kapott eredményeket. A Myriad változó megadja a BRCA1 és a BRCA2 gének mutációjának valószínűségét. A piros valószínűségi változók a beteg általános kockázati tényezői, míg a kék változók a beteg rokonaira vonatkozó kockázati tényezők.

A BOADICEA modellhez tartozó statisztikai adatok a páciens lánytestvérére és anyjára vonatkozóan készültek, a többi modellben azonban a rokonokra csak elsőfokú rokonsági megkötés van. Ennek következtében a fuzionált valószínűségi hálómban az édesanya és a lánytestvér állapotát vizsgálom, lehetőséget biztosítva a BOADICEA modell eredményeinek felhasználását a további kutatásokhoz.

A motherAge és sisterAge valószínűségi változók megadják a páciens rokonainak korát, azaz mikor diagnosztizáltak náluk mellrákot. Ebből következően függővé tettem a motherCancer, illetve a sisterCancer változókat a megfelelő korváltozóktól, amelyekre az a megkötés érvényes, miszerint ha nincs rák diagnosztizálva, akkor általános, a populációra jellemző mellrák kockázattal számolok.

A modellek fuzionálásához zajos VAGY kaput használtam, melyet az alábbi módon paramétereztem: Claus modell esetén r = 0.5 értéket választottam, Gail1 modellnél r = 0.05, míg Gail2 modellnél r = 0.75 valószínűségeket adtam meg, saját hiedelem alapján. A legszélesebb körben alkalmazott modell a Gail modell, továbbá jelentős a genetikai öröklés mellráknál, ezért a Gail2 modellt veszem figyelembe a legnagyobb mértékben. A Gail1 modell csak kis mértékben befolyásolja a modellt, mivel az a tény, miszerint csak az első menstruáció alapján diagnosztizáljunk mellrákot, önmagában nem releváns, azonban más modellel együtt alkalmazva érdemleges információt szolgáltathat. A noisy-OR kapuhoz 0.1 nagyságú leaky-node-ot állítottam be.

(Claus, Gail1, Gail2)	true	false
(true, true, true)	0.98313	0.01687
(true, true, false)	0.9775	0.0225
(true, false, true)	0.6625	0.3375
(true, false, false)	0.55	0.45
(false, true, true)	0.96625	0.03375
(false, true, false)	0.955	0.045
(false, false, true)	0.325	0.675
(false, false, false)	0.1	0.9

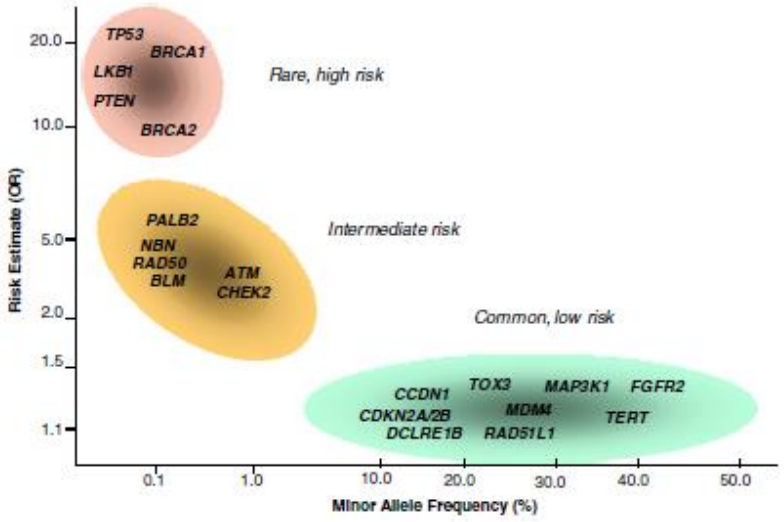
7. ábra: A P(patientCancer) feltételes valószínűségi táblája a fent ismertetett noisy-OR értékek esetén

A felparaméterezett modell alapján az általános kockázat mellrák kialakulására **P(patientCancer)=0.13594**, amely jól közelíti a statisztikailag meghatározott általános kockázatot (P(motherCancer=P(sisterCancer))=0.123).

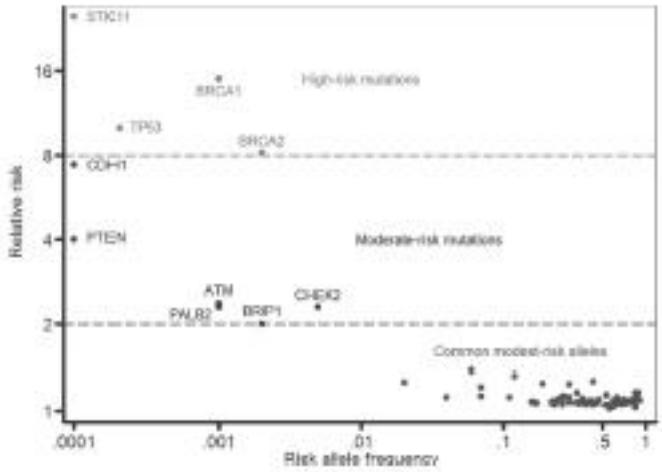
Az összetett, fuzionált modellt parametrikus módszerrel hoztam létre, már meglévő modelleket, statisztikai eredményeket egyesítettem a közös kockázati tényezőkön keresztül, azaz együttes eloszlásfüggvények alapján. Ez a módszer azonban jelentős korlátokat szab, amely jelentkezett a BOADICEA modell felhasználhatóságánál is: a rokonoknál vizsgált tényezőkre szűkebb megkötéseket tettek, így nem lehetett egyértelműen összeegyeztetni a többi modellel. Ebből következően a bővítés komoly akadályokba ütközik. Új valószínűségi változók felvétele is sokszor problémát okoz, ugyanis a fenti modellek nem szolgáltatnak elég információt az integráláshoz, például ha a PALB2 gén mutációját is fel szeretnénk használni, akkor azt egy másik statisztikai eredmény mentén kell megtennünk, mert a korábban ismertetett modellek nem vizsgálják. Számos más kockázati modellt is kidolgoztak, de a gyakorlatban alkalmazott modellek genetikai oldalról csak a BRCA1 és BRCA2 gének mutációjának vizsgálatával foglalkoztak. A korábbi modellek és a parametrikus hálóépítés korlátai keltette fel érdeklődésemet a strukturális modellépítés és a genetika részletesebb megismerésére, a mellrák kialakulását növelő gének feltérképezésére.

3.2 Vizsgálandó gének és variánsaik

A vonatkozó szakirodalom feldolgozása során megállapítottam, hogy számos más gén mutációja is jelentősen növelheti a mellrák kialakulásának kockázatát.



8. ábra: A mellrák kockázatát növelő gének csoportosítása a Hereditary Cancer in Clinical Practice szerint [10]



9. ábra: Csoportosítás a The American Journal of Pathology kutatásai alapján [11].

Az 8. és 9. ábrán megfigyelhető, hogy a géneket három nagy csoportba lehet sorolni előfordulási valószínűségük és kockázatnövelő hatásuk szerint:

a.) magas rizikójú, ritka előfordulású gének

b.) közepes rizikójú, közepes előfordulású gének

c.) közepes vagy általános rizikójú, gyakori előfordulású gének

Szoftverem tesztelésére néhány magas rizikójú és közepes rizikójú gént használtam fel, amelyek a DNS-hibajavításban játszanak fontos szerepet [9, 10]. Ennek megfelelően a kutatást a BRCA1 és BRCA2 gének mellett kiegészítettem még a PALB2, TP53, ATM, CHEK2 és PTEN gének vizsgálatával is.

Korábban a PALB2 gént közepes kockázatú génként tartották számon (1-2. ábra). Azonban több kutatóintézet is kimutatta, hogy kockázatnövelő hatása megközelíti a BRCA2 génét. A hasonlóság abból is ered, hogy a PALB2 gén a BRCA2 génnel együttműködve végzi a DNS javítását. A PALB2 gén mutációja közel 35%-kal növeli a mellrák kialakulásának kockázatát, amely tovább növekedhet családi halmozódás esetén [12, 13].

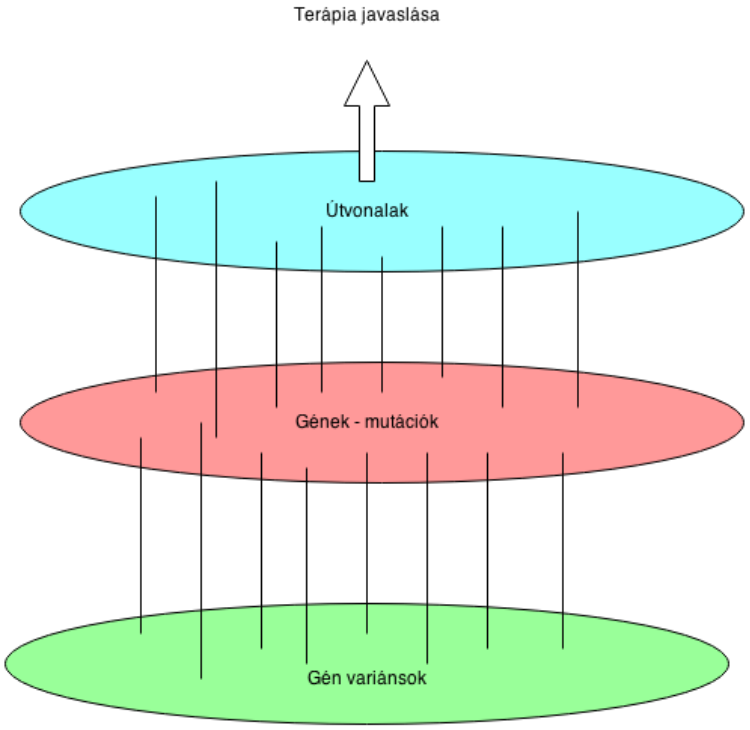
A kutatást kibővítettem néhány génvariáns vizsgálatával is, ugyanis bizonyos variánsok jelenléte a fent listázott génekben megnöveli a mellrák kialakulásának kockázatát. Egy génvariáns több génben is előfordulhat, így ezen kapcsolatok feltérképezése új eredményeket hozhatnak. A vonatkozó szakirodalom feldolgozása alapján a vizsgáltra kiválasztott génvariánsokat az 1. táblázatban foglaltam össze [14].

1. táblázat: A vizsgálatban szereplő gének variánsai

Gén	*BRCA1*	*BRCA2*	*ATM*	*CHEK2*	*TP53*
Kiválasztott génvariáns	Q 356 R rs1799950	N 289 H rs766173	V 182 L rs3218707	I 157 T rs17879961	P 72 R rs1042522
	D 693 N rs4986850	N 372 H rs144848	L 546 V rs4987945		
	S 1140 G rs2227945	T 1915 M rs4987117	S 707 P rs4986761		
	K 1183 R rs16942	R 2034 C rs1799954	D 814 E rs3218695		
	S 1613 G rs1799966	S 2835 P rs11571746	F 858 L rs1800056		
		E 2856 A rs11571747	P 1054 R rs1800057		
		I 2944 F rs4987047	H 1380 Y rs3092856		
		K 3326 stop rs11571833	L 1420 F rs1800058		
		I 3412 V rs180142	D 1853 V rs1801673		

3.3 Vizsgált hierarchia bemutatása

A korábbi kockázati modellektől eltérően a változók egy olyan hierarchiáját ismertetem, mellyel megvizsgálom alaposabban a beteg genetikai hátterét, azaz elvégzem a génvariánsok és az útvonalak vizsgálatát is.



10. ábra: a gén variánsok, gének és útvonalak hierarchiája. Az egyes szinteken belül is futhatnak kapcsolatok, amely összetettebb ok-okozati vizsgálatot tesz lehetővé.

A fenti ábrán megfigyelhető a kialakítandó hierarchia (valószínűségi háló), melynek szintjei a következőket reprezentálják:

- Alsó (zöld) szint: a génvariánsok csomópontjai, melyek a valószínűségi hálóban azt a valószínűséget fejezik ki, hogy a betegnél megfigyelhető-e az adott génvariáns.
- Középső (rózsaszín) szint: gének csomópontjai. Az egyes csomópontok megadják, mekkora valószínűséggel következik be mutáció az adott génnél.
- Felső (kék) szint: útvonalak (fehérje, gén stb.). Ezek a valószínűségi változók megadják, mekkora valószínűséggel módosulhat az adott útvonal.

A felső szint alapján már javasolható személyre szabott terápiás kezelés. Jelen dolgozatomban az alsó két szintet vizsgálom, továbbá a szoftver szerkezete lehetővé teszi a későbbiekben a felső szint vizsgálatát is.

Az egyes szintek között keressük a kapcsolatokat, azaz egy adott gén variánsának jelenléte hogyan befolyásolja a gén mutációjának valószínűségét, illetve egy gén mutációja milyen hatást fejt ki az adott fehérjeútvonalra. A középső szinten belül is értelmezhetők kapcsolatok, ugyanis egy gén mutációja előidézheti más gén mutációját is, mivel egyes gének bizonyos feladatokat együttműködve látnak el.

3.4. Vizsgálat során felhasznált módszerek és tényezők

A hierarchiát egy valószínűségi hálóval vizsgálom, amelyet egy adott annotációs valószínűségi hálóból származtatok le. Az annotációs valószínűségi háló abban különbözik a hagyományos Bayes-hálótól, hogy a gráf élei az annotációk alapján súlyozottak valamilyen statisztikai módszer által, ezzel elősegítve az élek fontosságának eldöntését. Az annotációs valószínűségi háló éleit többféle módszerrel is minősíthetjük, például együttes névelőfordulás alapján vagy kernel alapú hasonlóság módszerrel is [15]. Az élek minősítésének elvégzése után szűrést végzünk: megnézzük, mely élek valószínűségi értékei esnek egy adott küszöbérték felé. Ezen élek megtartása után már nincs szükség az élek valószínűségi értékeire, így megkapjuk a várt Bayes-hálót, amelyet felhasználhatunk további kutatásokhoz.

A valószínűségi hálót a fentieknek megfelelően strukturálisan hozom létre, melynek során az adott annotációk, szakirodalmak feldolgozása szolgáltatja és súlyozza az éleket a gráfban. A szakirodalomból tanulás nagyobb szabadságot biztosít a hálók építésében és bővítésében is, mint a parametrikus hálóépítés. Ezen feladat megoldására a BMC Informatics együttes névelőfordulás módszerét használtam. [15,16].

Az együttes névelőfordulás valószínűségének kiszámolása

A BMC Informatics az együttes névelőfordulást a gének kapcsolatának minősítésére használták fel. Ennél a módszernél a PubMed és a Gene Ontology annotációk szolgáltatták a szükséges információkat. A módszer az együttes névelőfordulást vizsgáló módszerhez hasonlít, ugyanis itt is azt vizsgáljuk, mekkora valószínűséggel fordul elő az adott absztraktban a két gén megnevezése együttesen. Az egyes génpárok együttes előfordulását az adott annotációban az alábbi valószínűséggel adható meg:

$p_{link}(\#absztrakciók\ száma \geq k\ |n,m,N) = 1 - \sum_{i=0}^{k-1} p(i|n,m,N)$, ahol

$p(i|n,m,N) = \frac{n!(N-n)!m!(N-m)!}{(n-i)!i!(m-i)!(N-n-m+i)!N!}$

Jelölések:

- n,m: azon absztrakciók száma, melyben az n illetve m gén megjelenik
- i: azon absztrakciók száma, melyben mindkét gén megtalálható
- N: a vizsgált absztrakciók száma

A fenti képlettel a két valószínűség egyezőségét, az annotációkban való megfigyelések számának egyezőségét (azaz együtt fordulnak elő az absztrakciókban) [16].

Az élek minősítésére az élsúly kiszámolása:

$N(Edge_{ij}) = Ceil(10 * p(i,j))$,

azaz az élhez kiszámolt valószínűség 10-szeresének felfelé kerekített egész értékét veszi [16].

A fenti módszeren alapszik a Matlab egyik eszközkészlete, a Kevin Murphy's BNT package, amely a kivonatolt adatokból kitanul egy valószínűségi hálót [16].

A vizsgálatban keresendő kifejezések

A PAGEVA szoftver számára a kernel hasonlóság módszerben alkalmazott kernelekhez hasonlóan kell megadni a keresett változókhoz tartozó kifejezéseket. A mátrix minden egyes sora egy valószínűségi változót reprezentál, melyeknél felsoroljuk a változóhoz tartozó szinonimákat, kifejezéseket a keresés végrehajtásához [15]. A vizsgálathoz a 3.1.3. fejezetben ismertetett fuzionált modell valószínűségi változóit használtam fel, mint általános kockázati tényezőket, továbbá a 3.2. fejezetben ismertetett géneket és variánsaikat.

A fent leírtak alapján a vizsgálathoz felhasznált kernel mátrixot a következő táblázatban ismertetem:

2. táblázat: a vizsgálat kernel mátrixa

Valószínűségi változó típusa	Valószínűségi változó neve	Kifejezések, szinonimák
Gén	BRCA1	BRCA1, BRCA1/2
Gén	BRCA2	BRCA2, BRCA1/2
Gén	PALB2	PALB2
Gén	TP53	TP53
Gén	ATM	ATM
Gén	CHEK2	CHEK2
Gén	PTEN	PTEN
Variáns	Q356R	Q356R, Q 356 R, rs1799950
Variáns	D693N	D693N, D 693 N, rs4986850
Variáns	S1140G	S1140G, S 1140 G, rs2227945
Variáns	K1183R	K1183R, K 1183 R, rs16942
Variáns	S1613G	S1613G, S 1613 G, rs1799966
Variáns	N289H	N289H, N 289 H, rs766173
Variáns	N372H	N372H, N 372 H, rs144848
Variáns	T1915M	T1915M, T 1915 M, rs4987117
Variáns	R2034C	R2034C, R 2034 C, rs1799954
Variáns	S2835P	S2835P, S 2835 P, rs11571746
Variáns	E2856A	E2856A, E 2856 A, rs11571747
Variáns	I2944F	I2944F, I 2944 F, rs4987047
Variáns	K3326STOP	K3326stop, K 3326 stop, rs11571833
Variáns	I3412V	I3412V, I 3412 V, rs180142, rs3218707
Variáns	L546V	L546V, L 546 V, rs4987945
Variáns	V182L	V182L, V 182 L, rs3218707
Variáns	S709P	S709P, S 709 P, rs4986761
Variáns	F858L	F858L, F 858 L, rs1800056
Variáns	P1054R	P1054R, P 1054 R, rs1800057
Variáns	H1380Y	H1380Y, H 1380 Y, rs3092856
Variáns	L1420F	L1420F, L 1420 F, rs1800058
Variáns	D1853V	D1853V, D 1853 V, rs1801673

Variáns	I157T	I157T, I 157 T, rs17879961
Variáns	P72R	P72R, P 72 R, rs1042522
Általános tényező	mother_age	mother's age
Általános tényező	sister_age	sister's age
Általános tényező	patient_age	patient's age
Általános tényező	breast cancer	heritable breast cancer
Általános tényező	ethnic	ethnic
Általános tényező	menarche	menarche
Általános tényező	age_at_first_birth	age at firth live birth

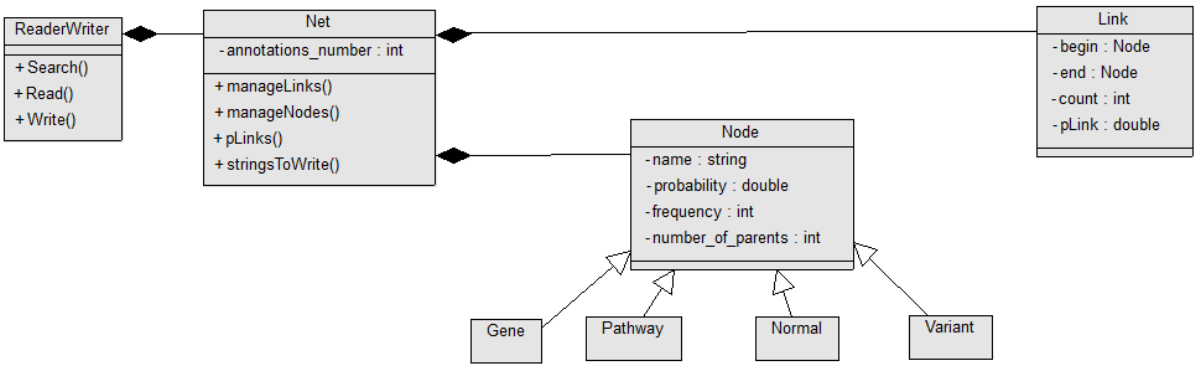
A 2. táblázatban egyetlen olyan változó van, amelyet még nem részleteztem: a breast cancer (heritable breast cancer) változó segítségével külön vizsgálatot végzek az örökletes mellrák kockázati tényezőinek felderítésére. Mivel a PubMed absztrakciókat elsősorban „breast cancer" kulcsszó alapján gyűjtöttem össze, ezért megjelenik a szomatikus mellrákot és az örökletes mellrákot befolyásoló tényezők is. A 2.1.2-ben ismertetett modellek szintén az örökletes mellrák kockázati tényezőit vizsgálja, így szükséges volt egy új változó bevezetése.

4. A PAGEVA szoftver ismertetése

A PAGEVA szoftver - melynek neve a „Pathway”, „Gene” és ” Variant” szavakból alkotott mozaikszó – 2014-ben, C# nyelven fejlesztett, általam készített program. A szoftver célja a 3.3. fejezetben ismertetett hierarchia vizsgálatához szükséges annotációk, absztraktak feldolgozása és a keresési eredmény alapján egy súlyozott, annotált valószínűségi háló létrehozása a hierarchiának megfelelően.

A szoftver szerkezete:

A PAGEVA szoftver szerkezetét úgy terveztem meg, hogy könnyedén lehessen bővíteni a genetikai vizsgálatokhoz. A 3.3. fejezetben ismertetett hierarchia szintjei egy-egy valószínűségi változó csoportot reprezentálnak, így külön osztályokat kaptak. Az egyes csoportoknak számos egyedi tulajdonság megadható a későbbi kutatások érdekében, pl. a változók értékkészlete, színe a BayesCube-ban.



11. ábra: a PAGEVA szoftver szerkezetét összefoglaló, de nem teljes UML osztálydiagram. A Net osztály egy adott valószínűségi hálót reprezentál, mely éleket (Link) és különböző csomópontokat (Node) tartalmaz.

A szoftver egyes osztályainak ismertetése:

ReaderWriter:

Input-output kezelésért felelős, azaz a szövegfájlok beolvasásáért és az adatok kiíratásáért egy adott valószínűségi háló esetén. Először beolvassa a kernel mátrixhoz szükséges adatokat, amely alapján felveszi a háló csomópontjait. Ezután feldolgozza az annotációkat a kernel mátrixban megadott kifejezések alapján, miközben tárolja az egyes éleket és előfordulásaiknak gyakoriságát, továbbá a csomópontok gyakoriságát is. A feldolgozás végén előállíttatja a hálóval az egyes élek valószínűségeit, végül kiíratja különböző fájlokba az eredményeket.

Net:

Adott valószínűségi hálót kezel, azaz csomópontokat és éleket, illetve előállítja a háló paraméterei alapján a kiíratáshoz szükséges stringeket.

Változók:

- *links*: éllista
- *nodes*: csomópontok listája
- *annotations_number*: felhasznált annotációk száma – az élsúlyok kiszámításához szükséges

Metódusok:

- *manageLinks(), manageNodes()*: a háló objektumainak kezeléséért felelős függvénycsoportok.
- *pLinks()*: élek súlyozását végzi el az annotációk feldolgozása után. Ebben a metódusban implementáltam a 3.4. fejezetben ismertetett együttes névelőfordulás egyenletét.
- *stringsToWrite()*: a fájlba kiíratáshoz szükséges stringeket állítja elő (pl. XML fájlhoz a szükséges paramétereket)

Link:

A valószínűségi háló egyes éleihez biztosítja az adatok eltárolását.

Változók:

- *begin, end*: az él szülő-, illetve gyermek csomópontját reprezentáló Node objektumok.
- *count*: az élek előfordulásának gyakoriságát tároló változó, azaz hányszor fordult elő mindkét kifejezés 1 annotáción belül.
- *pLink*: adott élhez kiszámított élsúly.

Metódusok: csak a változók értékeinek beállítására és lekérdezésére szolgáló függvényeket tartalmazza.

Node:

A Bayes háló csomópontjait reprezentáló osztály. Ez egy ősosztály a különböző csomópontok együttes kezelésére, tárolására és közös változók, metódusok előírására.

Változók:

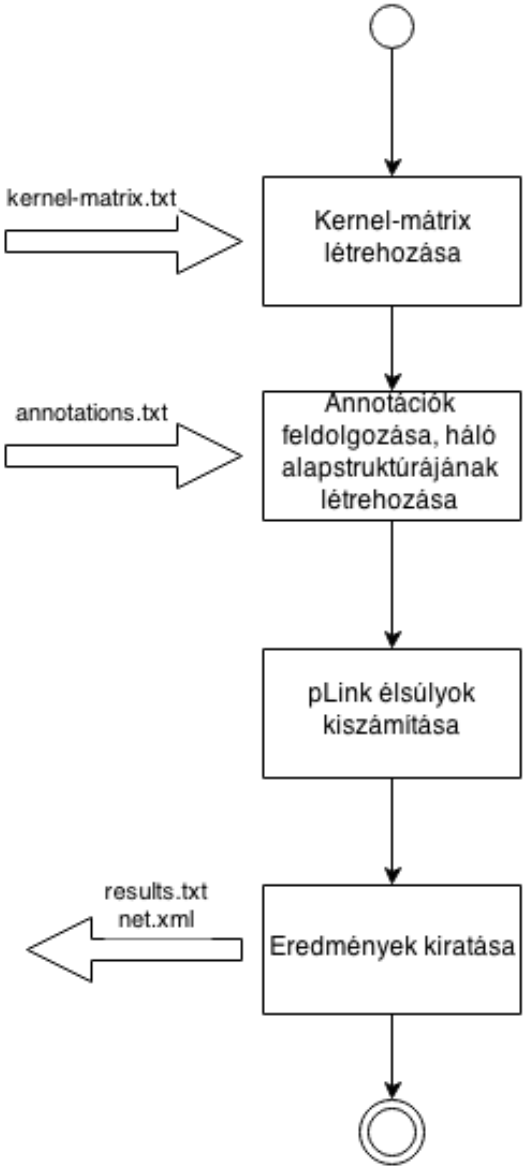
- *name:* csomópont neve
- *probability:* a csomópont valószínűségi - most még nem használt, a szoftver későbbi bővíthetőségéért lett megadva. Később átalakítom feltételes valószínűségi táblák megadására.
- *number_of_parents:* a csomópont szülő csomópontjainak száma – az XML formátumú fájlba kiíráshoz szükséges.
- *frequency:* a csomópont előfordulásának száma az adott annotációs forrásban.

Metódusok: ebben az osztályban ismét csak a változók értékeinek beállítására és lekérdezésére szolgáló függvények szerepelnek.

Gene, Pathway, Variant, Normal node: leszármaztatott osztályok, amelyek a 3.3 fejezetben ismertetett hierarchia egy-egy absztrakciós szintjének felelnek meg.

A szoftver bemenete és kimenete:

A szoftver bemenetére először egy olyan szöveges, txt formátumú fájlt kell megadni, amely tartalmazza a kernel-mátrix elemeit a következő formátumban: minden kifejezést vesszővel kell elválasztani, a sor első eleme a változó típusát tartalmazza, a második eleme a változó nevét és további 8 kifejezést lehet megadni a kereséshez. Miután a PAGEVA szoftver feldolgozta és létrehozta a kereséshez szükséges kernel-mátrixot, egy újabb txt fájlban megadjuk az annotációkat, absztrakciókat tartalmazó fájlt. Ezen fájl alapján létrehozza a valószínűségi hálót élsúlyokkal együtt. Ezután egy-egy txt fájlba kiíratja a csomópontokra és élekre vonatkozó információkat a debuggolás és validálás segítéséhez. A legvégén a program előállít egy XML fájlt, amely egy BayesCube-os modellt reprezentál és importálhatóságot biztosít a BayesCube-ba. A szoftver alapműködését a 12. ábra foglalja össze.



12. ábra: a PAGEVA szoftver alapműködését bemutató folyamatábra

A szoftver jelenleg PubMed absztakciók feldolgozására íródott, azonban a ReaderWriter osztály segítségével könnyedén implementálhatók más annotációk, absztrakciók feldolgozása is (pl. Gene Ontology annotációi).

5. Eredmények kiértékelése

Az új háló létrehozásához 78209 PubMed cikk absztraktját dolgoztam fel, amelyek cikkeit legkorábban 2010-ben publikáltak. Az egyes változók előfordulásait az 1. diagramon összegeztem.

1. diagram: a források feldolgozásakor kapott eredmények. A különböző színezések az egyes absztrakciós szinteket reprezentálják – bordó a géneké, halványsárga az általános kockázati tényezőké és zöld a variánsoké.

Az 1. diagramon csak azon változókat jelenítettem meg, amely legalább egyszer előfordult valamely annotációban. Ennek következtében 24 változó előfordulása olvasható le a 38-ból. A BRCA1 és BRCA2 gének előfordulása a legnagyobb, amely igazolja, mennyire is aktuális ezen gének kutatása napjainkban. A többi gén is jelentős számban fordult elő a forrásban, bizonyítva a génkutatások aktivitását.

A 3. fejezetben ismertetett kockázati modellektől eltérően a páciens és rokonainak kora háttérbe szorult, helyettük az etnikum és a menstruáció kezdete került előtérbe, amelyek csak a Gail-modellben szerepeltek. Meglepő eredmény továbbá, hogy lánytestvérre vonatkozó cikk nem fordult elő a forrásban. A vizsgálatba bevont 24 darab génvariáns közül csak 10-et lehetett megtalálni az absztrakciókban, amely arra enged következtetni, hogy habár ezen a területen is folynak már kutatások, nem mindegyik variánst vizsgálták részletesen a mellrákos megbetegedés szempontjából.

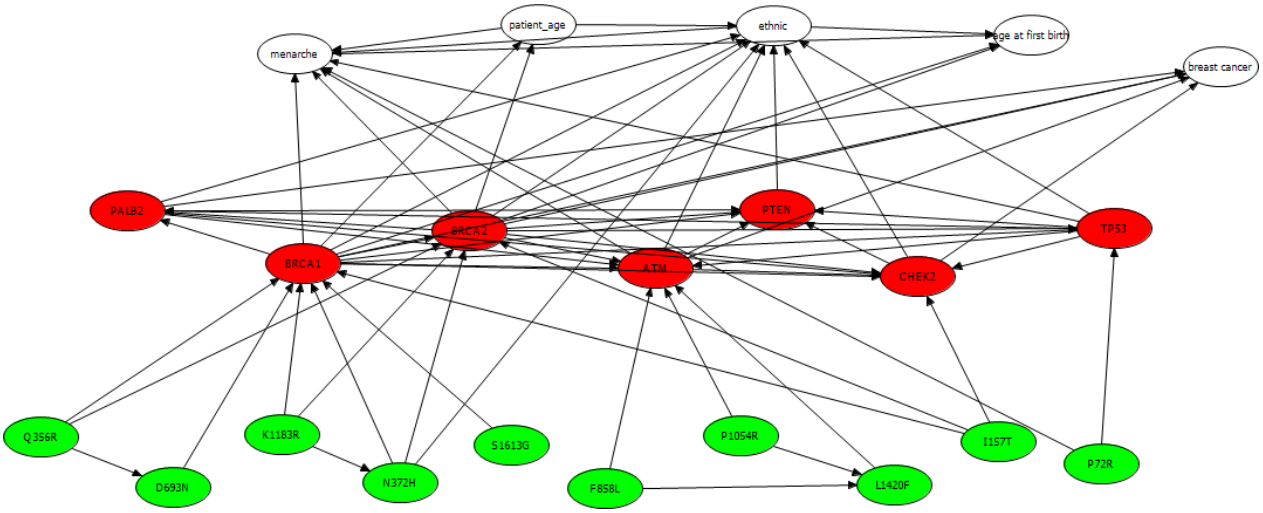
Az absztrakciókból 68 kapcsolatot sikerült felderítenem, melynek eredményeit megjelenítettem a 3. táblázatban. A 12. ábrán megfigyelhető valószínűségi hálónak az egyes csomópontok által képviselt absztrakciós szintjei, továbbá a gének közötti nagyszámú kapcsolat.

3. táblázat: a valószínűségi változók között felderített kapcsolatok és előfordulásaik gyakorisága

Szülő csomópont	Gyermek csomópont	Előfordulások száma	pLink	Minősítés
BRCA1	BRCA2	1389	0,999999	10
BRCA1	PTEN	54	0,999999	10
BRCA1	ethnic	77	0,999999	10
BRCA2	ethnic	70	0,999999	10
BRCA1	TP53	88	0,999998	10
BRCA2	PTEN	36	0,999998	10
PTEN	ethnic	5	0,999981	10
BRCA2	TP53	70	0,999931	10
BRCA1	ATM	93	0,999848	10
BRCA1	menarche	14	0,999804	10
TP53	ethnic	7	0,999613	10
BRCA2	ATM	61	0,998276	10
BRCA2	menarche	13	0,997917	10
ATM	ethnic	6	0,994209	10
ethnic	menarche	22	0,993218	10
TP53	PTEN	46	0,985412	10
BRCA1	PALB2	85	0,980787	10
BRCA1	CHEK2	73	0,96551	10
BRCA2	PALB2	107	0,939042	10
ATM	PTEN	21	0,92546	10
BRCA2	CHEK2	70	0,907072	10
BRCA1	patient_age	1	0,894774	9
PALB2	ethnic	6	0,891606	9
TP53	ATM	29	0,890193	9
TP53	menarche	1	0,880859	9
CHEK2	ethnic	6	0,847707	9
ATM	menarche	1	0,82655	9
BRCA2	patient_age	1	0,794121	8

Szülő csomópont	Gyermek csomópont	Előfordulások száma	pLink	Minősítés
patient_age	ethnic	2	0,711713	8
PALB2	PTEN	11	0,634741	7
CHEK2	PTEN	13	0,568198	6
TP53	PALB2	17	0,561026	6
BRCA1	age_at_first_birth	1	0,496157	5
TP53	CHEK2	29	0,494289	5
PALB2	ATM	30	0,472672	5
ATM	CHEK2	37	0,407996	5
BRCA2	age_at_first_birth	1	0,378657	4
ethnic	age_at_first_birth	2	0,310654	4
PALB2	CHEK2	31	0,279506	3
I157T	BRCA1	3	0,278392	3
patient_age	menarche	2	0,270819	3
I157T	BRCA2	1	0,201562	3
BRCA1	heritable breast cancer	3	0,137663	2
N372H	BRCA1	3	0,112088	2
BRCA2	heritable breast cancer	3	0,096901	1
Q356R	BRCA1	3	0,083726	1
K1183R	BRCA1	3	0,083726	1
menarche	age_at_first_birth	16	0,079835	1
N372H	BRCA2	4	0,078254	1
N372H	ethnic	2	0,061715	1
K1183R	BRCA2	3	0,059247	1
Q356R	BRCA2	1	0,059247	1
P72R	TP53	8	0,043902	1
P72R	menarche	1	0,030352	1
S1613G	BRCA1	1	0,0287	1
D693N	BRCA1	1	0,0287	1
ATM	breast	1	0,017195	1
I157T	CHEK2	11	0,01664	1
PALB2	heritable breast cancer	1	0,008349	1
CHEK2	heritable breast cancer	1	0,007206	1
F858L	ATM	1	0,003465	1
P1054R	ATM	1	0,003465	1
L1420F	ATM	1	0,003465	1
K1183R	N372H	1	0,000228	1
Q356R	D693N	1	3,90E-05	1
F858L	P1054R	1	1,30E-05	1
F858L	L1420F	1	1,30E-05	1
P1054R	L1420F	1	1,30E-05	1

A 3. táblázat eredményei azt mutatják, hogy a leggyakrabban előforduló változókhoz tartoznak a legerősebb kapcsolatok (még ha az együttes előfordulások néhol nem magas értékűek), hiszen minél többször fordul elő egy adott változó a forrásban, annál nagyobb a valószínűsége, hogy egyszerre jelen van egy másik változó ugyanabban az annotációban. Az egyes absztrakciós szinteken belül is felfedezhetünk kapcsolatokat, amelyek a gének esetében erősebbek, míg a variánsoknál gyengébbek. Az összes kapcsolatot megjelenítettem a 13. ábrában az egyes absztrakciós szintek szerint rendezve. A körmentesség biztosítására az egyes szintekből csak felfelé vezethetnek élek, kiegészítve a szinteken belüli kapcsolatokkal, melyek irányát néhány esetben meg kellett fordítani. A szinteken belüli élek megfordítása nem probléma, ugyanis az ok-okozati összefüggés csupán a kernel-mátrixban felírt változók sorrendjéből erednek.



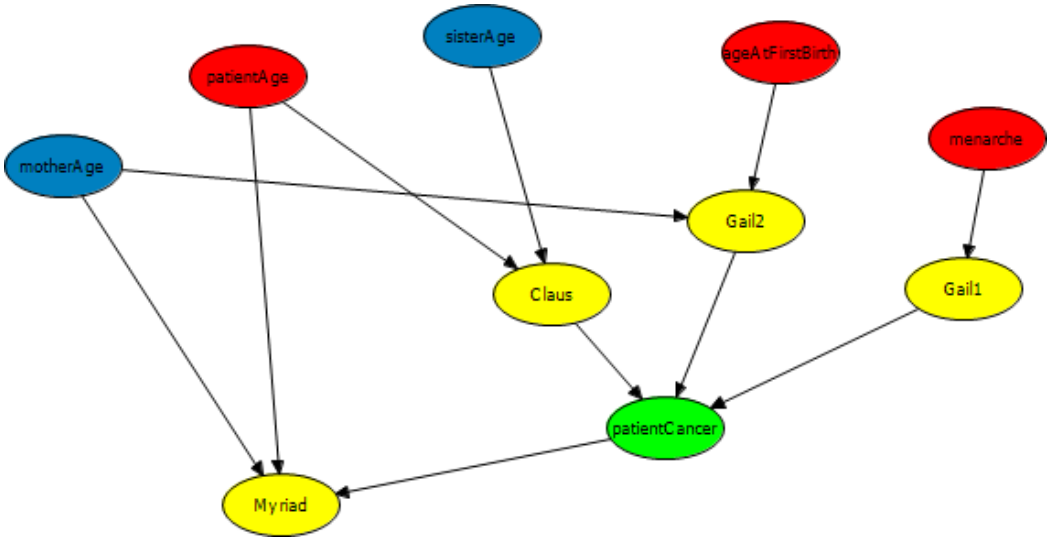
13. ábra: az annotációk alapján megalkotott háló az összes változóval és kapcsolattal reprezentálva. Az ábrát és a modellt BayesCube szoftverrel hoztam létre, amelyet elsősorban strukturális vizsgálatokhoz és következtetésekhez használnak. A hálót beimportáltam a BayesCube-ba a PAGEVA szoftverem által létrehozott XML fájlon keresztül.

A 13. ábrán megfigyelhető, hogy habár sok variánsról nem találtam információt (nem szerepelt egyetlen absztraktban sem), számos variáns megjelenik több génben (főként a BRCA1 és BRCA2 gének esetén feltűnő). Továbbá a gén variánsok absztrakciós szintjén belül is összekapcsolhattam néhány variánst egymással.

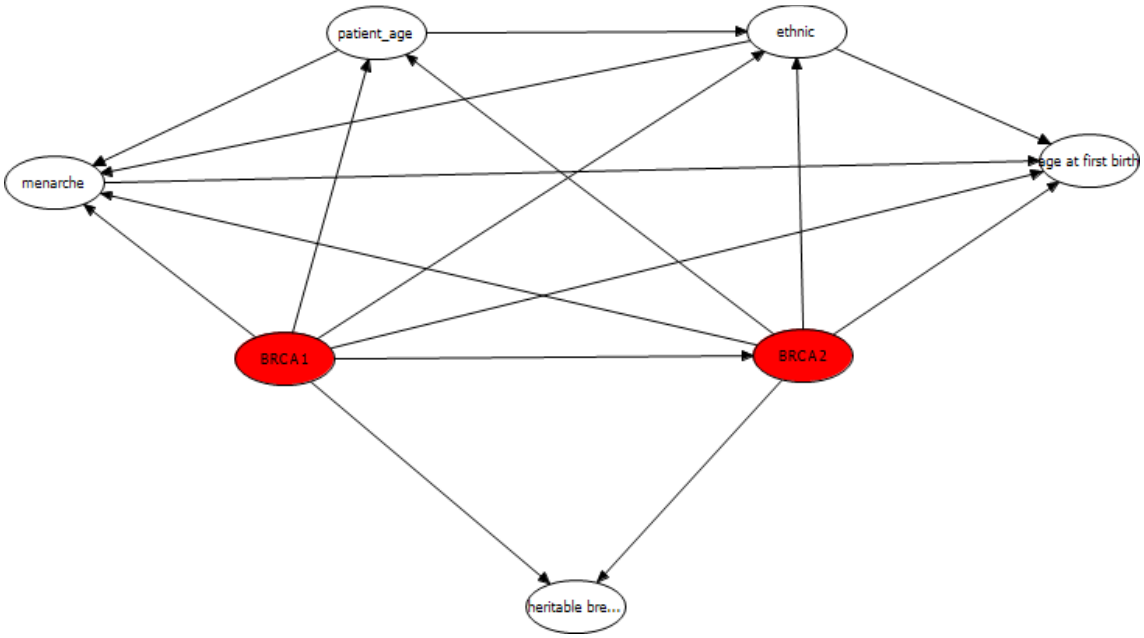
Az örökletes mellrák változóhoz más általános változó nem kapcsolódott, azonban a gének közül a PTEN és TP53 kivételével mindegyike hatással van a breast cancer változóra.

Az összetett, fuzionált modellemmel való összehasonlíthatóság érdekében az új háló csak azon változóit hagytam meg, melyekkel elvégezhető az összevetés, továbbá a fuzionált modellből kivettem a rokonokhoz kapcsolódó mellrák valószínűségi változókat, ugyanis azokra nem végeztem keresést.

Az eredményeket a 14. és 15. ábrán jelenítettem meg.

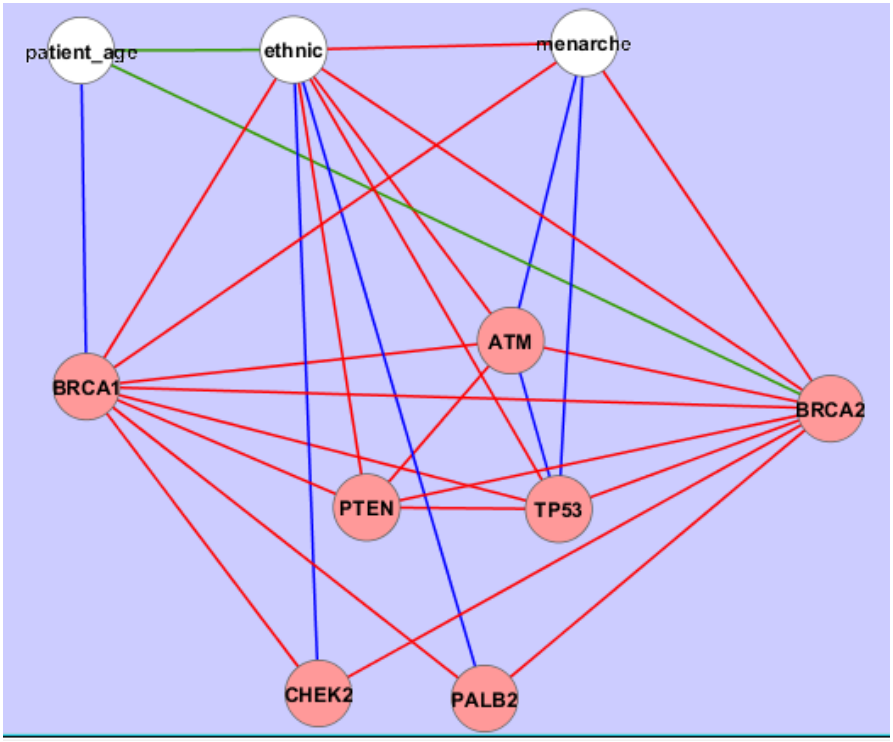


14. ábra: a fuzionált modellem, melynél az egyes kockázati tényezők között nincs kapcsolat.



15. ábra: az új háló struktúrája. A sisterAge változóra nem volt találat, helyette felvettem az ethnic változót, amely a heritable breast cancer kivételével teljes gráfot alkot a többi valószínűségi változóval. A két háló között további különbség, hogy míg a Myriad laboratórium a mellrákos megbetegedésből következtetett a BRCA1 és BRCA2 gének mutációjára, addig az új hálóban a BRCA1/2 gének vizsgálatából következtetünk az örökletes mellrák kialakulására.

Készítettem a létrehozott hálóhoz még egy ábrát, amelyen csak a legfontosabbnak minősített éleket jelenítettem meg. Az egyes minősítésű éleknek más és más színt adtam meg az átláthatóságért, az egyes absztrakciós szinteket a korábbiakban használt színekkel jelöltem, azaz pirossal a géneket és fehérrel az általános tényezőket.



16. ábra: a legjobb minősítésű éleket és csomópontjaikat ábrázoló háló. A zöld színű élek 8-as minősítésűek, a kékek 9-esek, míg a pirosak 10-esek. A piros változók a géneket reprezentálják, míg a fehérek az általános tényezőket. Az ábrát a Cytoscape szoftver segítségével hoztam létre, amelyet elsősorban valószínűségi hálók vizualizálására használnak.

A 16. ábrán megfigyelhető a legfontosabb éleket tartalmazó háló. Ebben a Bayes hálóban észrevehetjük, hogy az első 10 leggyakrabban előforduló változó jelent meg, továbbá az első 3 leggyakoribb változó (BRCA1, BRCA2 és ethnic) minden más változóra hatással van. A 4. leggyakoribb, a PTEN változó azonban nem minden változóval van kapcsolatban. Az ATM változó azonban több más változóval össze lett kötve, mint a PTEN. A kockázati modellekben felhasznált változók közül csak a heritable breast cancer – „örökletes mellrák” és az age at first birth – „kor az első élő gyermek születésekor” valószínűségi változók nem jelentek meg ebben a hálókban, ugyanis 5-ös és 3-as szintűek a kapcsolataik.

A fenti eredmények alapján érdemes lenne az etnikum és a menstruáció hatását alaposabban megvizsgálni a kockázati modellekben, továbbá néhány génnel kiegészíteni a kutatást (pl. PALB2), parametrikus úton megvizsgálva. A variánsok felparaméterezése után, illetve az útvonalak vizsgálatával kiegészítve a kutatást az egyes változók egymásra fejtett hatását jelentős lehet felmérni.

A létrehozott hálók változóihoz nem rendeltem eloszlást, ugyanis a PAGEVA szoftver feladata az volt, hogy egy struktúrát alkosson a feldolgozott absztraktak alapján.

6. Kitekintés

A 4. fejezetben bemutatott PAGEVA szoftver 1.0 verziójú, nem rendelkezik felhasználóbarát kezelőfelülettel. A jövőben szeretnék hozzá készíteni egy ablakos felületet, amelyen keresztül megadhatjuk a kernel mátrixot, a felhasználni kívánt annotációkat és absztraktakat. További fejlesztéssel megtekinthetnénk az egyes előfordulási értékeket, kiválasztva a kiimportálandó változókat és éleiket. Valószínűségi hálók létrehozására további más módszer is létezik, amellyel érdemes lenne kibővíteni a programot és összehasonlítani az egyes módszerek eredményeit

A szoftvert a Gene Ontology annotációkra is lehetne alkalmazni, a különböző útvonalak vizsgálatával kiegészítve. Mivel egyes betegségek bizonyos gének mutációival kapcsolatban állnak (egy gén mutációját több betegség is előidézheti), ezért a mellrák kutatásába be lehetne vonni más betegségeket is, pl. petefészekrák és prosztatarák vizsgálatát, kiegészítve a nem vizsgálatának fontosságával. További általános kockázati tényezők is befolyásolják a mellrák kialakulásának kockázatát, amely újabb bővítési lehetőséget biztosít, pl. a dohányzás és az alkoholfogyasztás. Döntési fákkal kiegészítve a hálókat terápiás kezelések javaslására is lehetne használni az eredményeket.

Köszönetnyilvánítás

Szeretném megköszönni konzulensemnek, Dr. Antal Péternek a problématerület megismerésében és a dolgozat írásában nyújtott segítségét, az elmúlt egy év szakmai támogatását.

Irodalomjegyzék

[1]: National Cancer Institute – Mellrák statisztikai adatok [Internet]: http://seer.cancer.gov/statfacts/html/breast.html

[2]: Stuart Russell, Peter Norvig – Mesterséges Intelligencia modern megközelítésben, 2. kiadás, 2005 Hungarian Translation Panem Könyvkiadó, Budapest

[3]: Millinghoffer András, Hullám Gábor, Antal Péter - Statisztikai adat- és szövegelemzés Bayes-hálókkal: a valószínőségektől a függetlenségi és oksági viszonyokig http://home.mit.bme.hu/~milli/docs/millinghoffer_06_AdatEsSzovegelemzes.pdf

[4]: Joseph P. Costantino, Mitchell H. Gail, David Pee, Stewart Anderson, Carol K. Redmond, Jacques Benichou, H. Samuel Wieand - Validation Studies for Models Projecting the Risk of Invasive and Total Breast Cancer Incidence (szept. 1999.), JNCI J Natl Cancer Inst (1999) 91(18): 1541-1548.

[5]: National Cancer Institute – Breast Cancer Risk Assessment Tool [Internet]: http://www.cancer.gov/bcrisktool/breast-cancer-risk.aspx

[6]: Tanja Hoegg - Statistical Modelling of Breast Cancer Risk for British Columbian Women The University of British Columbia (2013)

[7]: American Cancer Society kutatási eredményei: Ian D. Young – Introduction to risk calculation in genetic counseling

[8]: AC Antoniou, PPD Pharoah, P Smith and DF Easton - The BOADICEA model of genetic susceptibility to breast and ovarian cancer, British Journal of Cancer (2004), 91, 1580–1590. doi:10.1038/sj.bjc.6602175

[9]: Thomas S. Frank, Amie M. Deffenbaugh, Julia E. Reid, Mark Hulick, Brian E. Ward, Beth Lingenfelter, Kathi L. Gumpper, Thomas Scholl, Sean V. Tavtigian, Dmitry R. Pruss, and Gregory C. Critchfield - Clinical Characteristics of Individuals With Germline Mutations in BRCA1 and BRCA2: Analysis of 10,000 Individuals, 2002, J Clin Oncol 20:1480-1490.

[10]: Natalia Bogdanova, Sonja Helbig and Thilo Dörk - Hereditary breast cancer: ever more pieces to the polygenic puzzle, Bogdanova et al. Hereditary Cancer in Clinical Practice 2013, 11:12

[11]: Maya Ghoussaini, Paul D.P. Pharoah, and Douglas F. Easton: Inherited Genetic Susceptibility to Breast Cancer, Am J Pathol 2013, 183: 1038e1051

[12]: Antonis C. Antoniou, Ph.D., Silvia Casadei, Ph.D., Tuomas Heikkinen, Ph.D., Daniel Barrowdale, B.Sc., Katri Pylkäs, Ph.D., Jonathan Roberts, B.Sc., Andrew Lee, Ph.D., Deepak Subramanian, M.B., B.Chir., Kim De Leeneer, Ph.D., Florentia Fostira, Ph.D., Eva Tomiak, M.D., Susan L. Neuhausen, Ph.D., Zhi L. Teo, Ph.D., Sofia Khan, Ph.D., Kristiina Aittomäki, M.D., Ph.D., Jukka S. Moilanen, M.D., Ph.D., Clare Turnbull, M.D., Ph.D., Sheila Seal, M.I.Biol., Arto Mannermaa, Ph.D., Anne Kallioniemi, M.D., Ph.D., Geoffrey J. Lindeman, F.R.A.C.P., Ph.D., Saundra S. Buys, M.D., Irene L. Andrulis, Ph.D., Paolo Radice, Ph.D.,

Carlo Tondini, M.D., Siranoush Manoukian, M.D., Amanda E. Toland, Ph.D., Penelope Miron, Ph.D., Jeffrey N. Weitzel, M.D., Susan M. Domchek, M.D., Bruce Poppe, M.D., Ph.D., Kathleen B.M. Claes, Ph.D., Drakoulis Yannoukakos, Ph.D., Patrick Concannon, Ph.D., Jonine L. Bernstein, Ph.D., Paul A. James, M.B., Ch.B., Ph.D., Douglas F. Easton, Ph.D., David E. Goldgar, Ph.D., John L. Hopper, Ph.D., Nazneen Rahman, M.D., Ph.D., Paolo Peterlongo, Ph.D., Heli Nevanlinna, Ph.D., Mary-Claire King, Ph.D., Fergus J. Couch, Ph.D., Melissa C. Southey, Ph.D., Robert Winqvist, Ph.D., William D. Foulkes, M.B., B.S., Ph.D., and Marc Tischkowitz, M.D., Ph.D.- Breast-Cancer Risk in Families with Mutations in *PALB2,* N Engl J Med 2014; 371:497-506 August 7, 2014 DOI: 10.1056/NEJMoa1400382

[13]: Melissa C Southey, Zhi L Teo, James G Dowty, Fabrice A Odefrey, Daniel J Park, Marc Tischkowitz, Nelly Sabbaghian, Carmel Apicella, Graham B Byrnes, Ingrid Winship, Laura Baglietto, Graham G Giles,David E Goldgar, William D Foulkes, John L Hopper, kConFab, for the Breast Cancer Family Registry - A PALB2 mutation associated with high risk of breast cancer, Southey et al. Breast Cancer Research 2010, 12:R109

[14]: Nichola Johnson, Olivia Fletcher, Claire Palles, Matthew Rudd, Emily Webb, Gabrielle Sellick, Isabel dos Santos Silva, Valerie McCormack4, Lorna Gibson, Agnes Fraser, Angela Leonard, Clare Gilham, Sean V. Tavtigian, Alan Ashworth, Richard Houlston and Julian Peto - Counting potentially functional variants in BRCA1, BRCA2 and ATM predicts breast cancer susceptibility, Human Molecular Genetics, 2007, Vol. 16, No. 9 1051 1057, doi:10.1093/hmg/ddm050

[15]: Peter Antal, Geert Fannes, Dirk Timmerman, Yves Moreau, Bart De Moor - Using literature and data to learn Bayesian networks as clinical models of ovarian tumors, Artificial Intelligence in Medicine 30 (2004) 257–281

[16]: Shouguo Gao, Xujing Wang - Quantitative utilization of prior biological knowledge in the Bayesian network modeling of gene expression data, Gao and Wang BMC Bioinformatics 2011, 12:359