Ph.D. értekezés tézisei

Bioinformatikai analízis és digitális jelfeldolgozás génexpressziós adatokon

Készítette: Kelemen János-Zsigmond

Témavezető: Dr. Puskás László

MTA Szegedi Biológiai Központ

Funkcionális Genomika Laboratórium

Szegedi Tudományegyetem

2007

Bevezetés

Az utóbbi években egyre gyarapodó biológiai adatbázisok – mint például a DNS szekvencia, génexpressziós mintázat, fehérje-fehérje kölcsönhatás adattárak – a rendszer biológia dinamikus fejlődését eredményezték. A ma hozzáférhető magas adatátvitelű gén-expressziós technológiák hasonlóan hozzájárultak a rendszer biológia tudományterület fejlődéséhez. A rendszer biológia lehetővé teszi az élettani folyamatok jobb megértését és az orvosi biológia területén megbízhatóbb diagnosztikát és orvosi kezelést ígér. Ez azáltal válik elérhetővé, hogy törekszik matematikailag modellezni és szimulálni a sejtben zajló komplex biológiai folyamatokat. Ismeretes, hogy a rákos megbetegedések altípusai eltérően válaszolhatnak az eltérő kezelésekre. Ezért is indokolt a kezelést megelőző pontos diagnózis. A gén-expressziós mintázata alapján, a rákos sejt egy olyan több-állapotos rendszerként fogható fel, ahol az egyes állapotok a rák altípusainak feleltethetők meg. Ez az elképzelés vezetett el a rákos megbetegedés gén-

expresszión alapuló molekuláris klasszifikációjához – ami nem más, mint matematikai módszereken alapuló diagnózis. Az utóbbi években kitűntetett tudományos érdeklődésnek tartanak számot az elsősorban DNS microarray alapú ide sorolható módszerek. Nagyszámú mesterséges intelligencián alapuló algoritmusok, mint amilyenek a *support vector machine, mesterséges neuron hálók, nearest neighbor osztályozó,* vagy a *random forests*, azzal a céllal kerültek alkalmazásra, hogy pontosabb és megbízhatóbb diagnosztikát tegyenek lehetővé.

Sajnos a meglévő gén-expressziós adatokon (QRT-PCR, DNS microarray), a kísérleti körülményekből adódó hiba (zaj) és a biológiai eredetű variancia együttesen megfigyelhető. Ezért indokolt egy több lépésből álló adat elő-feldolgozás és további módszertani fejlesztések is.

Célkitűzés

Célkitűzéseink a gén-expressziós adatfeldolgozással és módszertani fejlesztéssel kapcsolatosak. Nevezetesen:

- A standard gén-expressziós adatfeldolgozási módszerek alkalmazása QRT-PCR és microarray adatokon.
- Statisztikailag megvizsgálni, hogy a laboratóriumban használt protokollok alapján hogyan befolyásolhatók az egyes gén-expressziós változások.
- Klaszterezés és marker gének azonosítása szkizofréniás betegek gén-expressziós mintázatában.
- Új normalizációs és zajszűrési (Kálmán Szűrő) módszerek fejlesztése és alkalmazása a molekuláris szintű rák diagnosztikában.

A gén-expressziós kovariancia, mely a gének közti funkcionális kapcsolatot is mutatja, fontos szereppel bír a betegségek molekuláris osztályozásában. A Kálmán Szűrő figyelembe veszi a gén-expressziós kovarianciát. Célunk, hogy a Kálmán Szűrő segítségével kiszűrjük a kísérleti zajt és megbecsüljük a minták biológiai állapotát. Nem utolsó sorban szándékunkban állt megvizsgálni a Kálmán Szűrővel kezelt adatok osztályozhatóságát, osztályozó algoritmusok segítségével.

Eredmények

A disszertációban közölt eredmények bioinformatikai módszerek alkalmazását mutatják be. A kötelező gén-expressziós adat elő-feldolgozási lépések, nevezetesen a minőség ellenőrzés és a LOWESS normalizáció illetve a *t*-próba, mely a gén-expressziós eltéréseket tárja föl, a társszerzős publikációk eredményeiben kerültek bemutatásra. A fenti módszerek alkalmazásának részletes leírása került bemutatásra, abban a publikációban, mely az *ntrR* által szabályozott géneket azonosítja *Sinorhizobium meliloti* modell organizmusban. Egy DNS microarray kísérletben az *S. meliloti* egy *ntrR* funkcióvesztéses mutánsát hasonlítottuk össze a vad típussal aerob és mikroaerob körülmények között.

Egerek makrofág sejtjein végzett lipopoliszacharidos kezelés egy olyan kísérletnek szolgált alapul, melyben a DNS amplifikációnak a gén-expressziós változásra mért hatását vizsgáltuk. A cDNS amplifikációt két protokoll - exponenciális fázisban megállított amplifikáció illetve szaturációs amplifikáció - alapján végeztük el és az

eredményezett gén-expressziós változást mutató adatokon $\chi 2$ próbát hajtotunk végre. A kísérlet kontrolljaként egy non-amplifikációs protokoll szolgált. A statisztikai eredmények azt igazolták, hogy az exponenciális fázisban megállított amplifikáció megbízhatóbb, szemben a szaturációs amplifikációval, a microarray kísérletek reprodukálhatósága szempontjából.

Továbbá, egy szkizofréniás betegekből álló populációt használtunk fel arra, hogy megbízható marker géneket keressünk a kór molekuláris diagnosztizálásához. A *DRD2* és a *Kir2.3* bizonyultak marker génnek. Annak ellenőrzésére, hogy a fenti gének esetében valóban a betegség marker génjeivel állunk szemben, hierarchikus klaszterezést hajtottunk végre, beteg és egészséges személyektől származó adatokon. A klaszterező eljárás látványosan kimutatta, hogy a szkizofrén minták elkülönültek a normál mintáktól a fenti gének tekintetében.

A továbbiakban a gén-expressziós adatok klasszifikációja állt érdeklődésünk középpontjában. A klasszifikáció hatékonyságának javítása érdekében a Kálmán Szűrőt

vezettük be. Munkánk szempontjából a legfontosabb tulajdonsága ennek a matematikai módszernek, hogy elkülőníti a biológiailag értelmezhető varianciát a mérési zajtól. A microarray kísérletben biológiai állapotnak tekintjük a gének valós expressziós szintjét. Az osztályozási felállásban ez az állapot az egyes alosztályoknak megfelelően változik. Ezt az esetet stochasztikusan modelleztük. A mérési zaj szintén stochaszikusan volt megjeleníthető. A Kálmán Szűrő a stochaszikus modellek mellett fölhasznál még egy a microarray folyamatnak megfelelő determinisztikus modellt. Ezek segítségével vált felbecsülhetővé a gének valós expressziós szintje azaz a biológiai állapot.

A fenti módszert 7 különböző publikus, tumoros eredetű adatsoron alkalmaztuk. A leghasználatosabb klasszifikációs módszereket szűrt és nem szűrt adatokon egyaránt teszteltük. Statisztikailag igazoltuk, hogy a Kálmán Szűrő szignifikánsan javítja az osztályozhatóságot. Három különböző grafikai ábrázolást alkalmaztunk, annak demonstrálására, hogy az egyes osztályok szemmel láthatóan elkülönülnek egymástól. Új

markerek azonosítását is tárgyaljuk, annak bizonyítására, hogy a szűrt expressziós adatok, már kis számú gén esetében is predikciós portenciállal bírnak az osztályozásra nézve.

Közlemények jegyzéke

A disszertációhoz kapcsolódó közlemények:

Kelemen JZ, Kertesz-Farkas A, Kocsor A, Puskas LG. Kalman filtering for disease-state estimation from microarray data. Bioinformatics. 2006 Dec 15;22(24):3047-53. Epub 2006 Oct 25.

Zvara A, Szekeres G, Janka Z, **Kelemen JZ**, Cimmer C, Santha M, Puskas LG. Over-expression of dopamine D2 receptor and inwardly rectifying potassium channel genes in drug-naive schizophrenic peripheral blood lymphocytes as potential diagnostic markers. Dis Markers. 2005;21(2):61-9.

Nagy ZB, **Kelemen JZ**, Feher LZ, Zvara A, Juhasz K, Puskas LG. Real-time polymerase chain reaction-based exponential sample amplification for microarray gene expression profiling. Anal Biochem. 2005 Feb 1;337(1):76-83.

Puskas LG, Nagy ZB, **Kelemen JZ**, Ruberg S, Bodogai M, Becker A, Dusha I. Wide-range transcriptional modulating effect of ntrR under microaerobiosis in Sinorhizobium meliloti. Mol Genet Genomics. 2004 Oct;272(3):275-89. Epub 2004 Sep 9.

Egyéb közlemények:

Feher LZ, Balazs M, **Kelemen JZ**, Zvara A, Nemeth I, Varga-Orvos Z, Puskas LG. Improved DOP-PCR-based representational whole-genome amplification using quantitative real-time PCR. Diagn Mol Pathol. 2006 Mar;15(1):43-8. Erratum in: Diagn Mol Pathol. 2006 Jun;15(2):123.

Nyilatkozat

Kelemen János-Zsigmond a „Over-expression of dopamine D2 receptor and inwardly rectifying potassium channel genes in drug-naive schizophrenic peripheral blood lymphocytes as potential diagnostic markers”[3] (Dis Markers 21(2):61-9), a “Wide-range transcriptional modulating effect of *ntrR* under microaerobiosis in *Sinorhizobium meliloti”*[2] (Mol Genet Genomics 272(3):275-89) és a “Real-time polymerase chain reaction-based exponential sample amplification for microarray gene expression profiling”[1] (Anal Biochem 337(1):76-83) című közleményeiben ismertetett eredményeivel jelentős mértékben hozzájárult az említett publikációk létrejöttéhez. Ezért, mint a fenti cikkek felelős szerzője, támogatom azt, hogy e publikációkat doktori fokozatszerzéséhez felhasználja.

Szeged,
2007, április 23

[1]**Dr. Puskás László**
tudományos tanácsadó

……..

[2]**Dr. Dusha Ilona**
tudományos tanácsadó

……..

[3]**Dr. Zvara Ágnes**
tudományos munkatárs

……..

Poszterek

Péter Hankovszky, Liliána Z. Fehér, **János Zsigmond Kelemen**, Sándor Sonkodi, László G. Puskás; Vasoactive effects of the long-term administration of erythropoietin and global gene-expression analysis in the rat mesenteric artery, 2005. Istanbul ERA-EDTA XLII Congress - poster

Bodogai M., Puskás L.G., Nagy Zs.B., **Kelemen J.Zs.**, Rüberg S., Becker A., Dusha I., Wide-range transcriptional modulating effect of ntrR under microaerobiosis in Sinorhizobium, The 6th European Conference on Nitrogen Fixation, 2004 - poster

Előadások

Kelemen JZ. "Methods in Microarray Analysis – Classification", Second International School on Biology, Computation and Information (BCI 2005), Dobbiaco (BZ), Italy, September 12-16, 2005