Magyar Tudomány 179(2018)5, 630–635
DOI: 10.1556/2065.179.2018.5.5

OKOS HÁLÓZATOK, AVAGY HOGYAN TEGYÜK AZ RNS-SZEKVENÁLÁST RELEVÁNSABBÁ

SMART GRAPHS: TURNING RNA-SEQ DATA INTO MEANINGFUL NETWORK

Makai Szabolcs

PhD, bioinformatikus, tudományos munkatárs, MTA Agrártudományi Kutatóközpont[1]
bionformatika@thegreatmagic.com

ÖSSZEFOGLALÓ

Az újgenerációs transzkriptom szekvenálás egy robusztus eszköz, amelyet gyakran használnak genetikai funkciók és szabályozó körök feltárására. Ugyanakkor nem minden koexpressziós esemény hordoz biológiailag releváns információt. Sőt, szigorúan véve az RNS-szekvenálás elsősorban egy adott gén aktivitását mutatja, és csak indirekt módon utal a gén termékének aktivitására. Ezért az RNS-szekvenálásból nyert információk elsősorban egy adott gén szabályozásáról, pontosabban szabályozó régiójáról adnak felvilágosítást. A transzkriptom adatok társítása a szabályozó régió elemeinek adataival egy ígéretes eszköz lehet a szabályozó körök feltárására.

ABSTRACT

Next generation sequencing of the whole transcriptome is a powerful tool to describe genetic functions and to detect regulatory circuits. However, not all events of co-expression have biological relevance. Speaking of protein coding genes, gene expression is primarily a measure of the gene's activity and only gives vague indications with regards of the activity of its product. Thus, the results of the whole transcriptome analysis give us information on the regulation of genes, and most importantly on its regulatory region. Coupling up expression data with regulatory region data is a promising tool to detect small regulatory circuits.

Kulcsszavak: hálózatelemzés, génexpresszió

Keywords: network analysis, gene expression

[1] A cikk írásakor.

A genomi éra tizenhetedik évében a funkcionális genomika legszélesebb körben alkalmazott technológiája az RNS-szekvenálás. Az újgenerációs, nagy áteresztőképességű technológiának hála ma olcsó lett, és kellően megbízható az élő szövetekben aktívan átíródó genomi DNS detektálása. A legújabb divat szerint ezt hívjuk transzkriptomnak, mert egyszerre az összes átírt RNS-szekvenciát képesek vagyunk meghatározni, avagy olvasni.

Röviden, a mérés során kinyerik a minta teljes RNS állományát, amelyet átírnak cDNS-sé, majd fragmentálnak. Aztán meghatározzák e fragmentumok nukleotid sorrendjét (szekvenciáját) egy bizonyos hosszon. Az így kapott, nemritkán milliós nagyságrendű olvasatokat a már ismert genomhoz illesztik, és az illeszkedés alapján meghatározzák a gént, amelyet a legvalószínűbben reprezentál. A génekhez illesztett fragmentumok száma lineárisan arányos a gén expressziós szintjével. Szokták ezt a számot a gén hosszával korrigálni, mert hiszen a hosszabb gén statisztikailag több fragmentumot adhat, és aktívabbnak tűnhetne. A gének expressziója azt mutatja, hogy az adott génből mennyi átirat, transzkript volt jelen a mintavétel pillanatában a szövetben.

A transzkriptom elemzés ezeknek a génhez rendelt számoknak az elemzése, mely során arra törekszünk, hogy meghatározzuk azon géneket vagy génegyütteseket, amelyek az adott kísérletben a legmeghatározóbbak lehetnek (például nagyon eltér a kontrollbeállításoknál mért számoktól). Másként mondva, felelős lehet egyfajta fejlődési állapotért vagy stresszválaszért. Köznapi hasonlattal élve olyan ez, mint amikor egy teremben sok ember gyűlik össze, sokan egyszerre beszélnek, és talán még több hallgat. Az első kérdésünk az lehet, hogyan azonosítsuk a beszélőket? Ezt tesszük a fragmentumok szekvenálásával, majd az illeszkedésen alapuló génazonosítással. A legáltalánosabb kísérleti beállítás szerint mindig egy kontrollbeállítást vetünk össze egy „megzavart" beállítással. Ilyen, amikor a teremben lekapcsolják a világítást, vagy például egy növényt, a szálkaperjét *(Brachypodium distachyon)* szárazságnak teszünk ki.

Ha váratlanul eloltják a lámpákat, a teremben sokan talán egyszerre felhördülnek, megijednek, majd elhallgatnak. Aztán lassan egy-két ember elkezd beszélgetni, nemritkán halkan, feltehetőleg azonos témát feszegetve: Mi történt? Hogyan állítsuk helyre a világítást? Lennének, akik kacagnának, élveznék a váratlan helyzetet. A szárazságnak kitett növény esetén is valami hasonló eseménysorozat játszódhat le, ám mindebből sokkal kevesebbet vagyunk képesek kimérni, és még kevesebbet értelmezni.

A koexpressziós hálózatok abban segítenek, hogy meghatározhassuk (azon túl, hogy ki beszél) azt is, hogy kik beszélnek egyszerre, avagy mely gének azok, amelyek egyszerre íródnak át. Azon egyszerű, ám korántsem helytálló feltételezés által, hogy akik egyszerre beszélnek, nagy eséllyel egymással is beszélnek. Habár a teremben és feltételezhetően a sejtmagban közelről sem ez a helyzet,

mégis valamilyen módon tájékoztatást kapunk a kezelés vagy zavarás által megszólaltatott, illetve elhallgattatott gének mibenlétéről.

Javítja a helyzetet, ha valamilyen módon meghatározzuk, hogy vajon ki miről beszélhet. A témák alapján pontosabban következtethetünk arra, hogy aki egyszerre, hasonló témáról társalog, az jó eséllyel azonos körhöz is tartozik. A címben említett okos hálózatok ezt teszik. Azaz funkcionális címkék (annotációk) és/vagy szabályozó motívumok alapján az okos hálózat fókuszba hoz egy-egy feltételezett génkört (vagy társaságot a makrovilágbéli hasonlatunkban), amelyek az adott kísérletben a leginkább jellemzőek. Szerencsére a gének funkciója nem annyira változatos, amennyire egy ember témaválasztása lehetséges egy társasági csevegésben. Legtöbbször a gének funkciója egzakt módon ismert, vagy legalább sejthető.

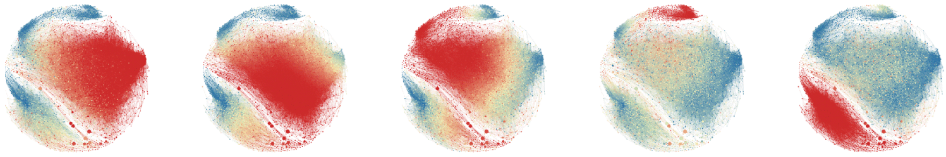
A szálkaperje szárazságra adott válaszát azért vizsgáltuk, mert könnyen hozzáférhető, karcsú genomjával ígéretes modellnövénye lehet a gazdaságilag fontos búzának, amelynek még az emberénél is ötször nagyobb a genomja, és kevésbé ismert. A kezelt és kontrollnövény kalászaiból vett idősoros mintákból meghatároztuk, hogy mely gének, mikor, milyen erősen fejeződtek ki. A koexpressziós hálózatot ezen expressziós értékek korrelációja alapján szerkesztettük meg.

A koexpressziós hálózatoknak nagyon széles irodalma van, amelyben a módszertani cikkek jelentős részt foglalnak el. Nem is meglepő, mert koexpresziót számolni sokféleképpen lehet, és mindegyik számítás egy kicsit másként releváns. A leggyakrabban használt számítás a Pearson-féle korrelációs együttható, a Spearman-rangkorreláció, az euklideszi távolság és legújabban a kölcsönös információn alapuló távolság. Ritkábban, de előfordul a dinamikus *time warping* algoritmus, ami idősoros adatoknál lehet hasznos.

Mindegyiknek megvan a maga előnye és hátránya, mindegyik másra jó. A Pearson-féle együttható lineáris összefüggést feltételez, és érzékeny a zajra, cserébe független a mértéktől. Az euklideszi nem független a mértéktől, kevésbé érzékeny a zajra, de nem képes negatív korrelációt mérni. A kölcsönös információ nem feltételez linearitást, skálaérzékeny, érzékeli a kiugró értékeket, amely tulajdonságok a biológiai rendszerekről is elmondhatóak. Két gén közötti kölcsönhatás lehet negatív, pozitív és mindezekkel együtt erősítő/csökkentő. Ezen erősítő hatás sajnos láthatatlan marad a kölcsönös információn alapuló számításoknál, de a Pearson-korrelációval társítva már eredményesen pontosítható a hálózat. Ugyanakkor bárhogyan is számoljuk a korrelációt, attól mi még csak arra vagyunk képesek, hogy megmondjuk, egy teremben kik beszélnek egyszerre, és azt még nem tudjuk megmondani, hogy kik beszélnek egymással. Fontos tisztázni, hogy RNS-szekvenálásra alapozva ezt nem is fogjuk tudni megmondani.

Első kérdésünk a brachypodiumos kísérletben az volt, hogy a kontroll- és a szárazságkezelt növények között mely gének expresszáltak eltérően. Sajnos a felsorolt módszerek ebben az esetben nem segítettek. Az történt ugyanis, hogy a növény a

szemfejlődést a nem várt szárazság hatására felgyorsította. Pearson-korrelációval ezt nem tudtuk kiszűrni, mert csak annyi történt, hogy máshol lettek a maximumok és a minimumok, de a szórások közel hasonlóan alakultak, másként mondva az értékek eltolódtak. Itt bizonyult nagyon szerencsésnek a nehezen lefordítható dinamikus time warping (időtorzításos) algoritmus, amelyet leggyakrabban a beszédfelismerésnél szoktak alkalmazni. (Hiszen ugyanazt a szót lehet gyorsabban és lassabban is kiejteni, attól még ugyanaz marad.) Azt néztük meg, hogy a gének önmagukhoz képest mennyire tolódtak el. Ezzel az elemzéssel rögtön azok a gének kerültek a fókuszpontba, amelyekről ismert, hogy szárazság, vízmegvonás hatására kapcsolnak be. Képletes példánkban valami olyan történhetett, hogy a teremben bejelentették, hogy elfogyott az ital. Ezért mindenki szép lassan befejezte a mondatát, elindult kifelé, és bizony akadtak beszélgetések, amelyek módosított helyszínen, például a ruhatárban vagy már az utcán fejeződtek be.

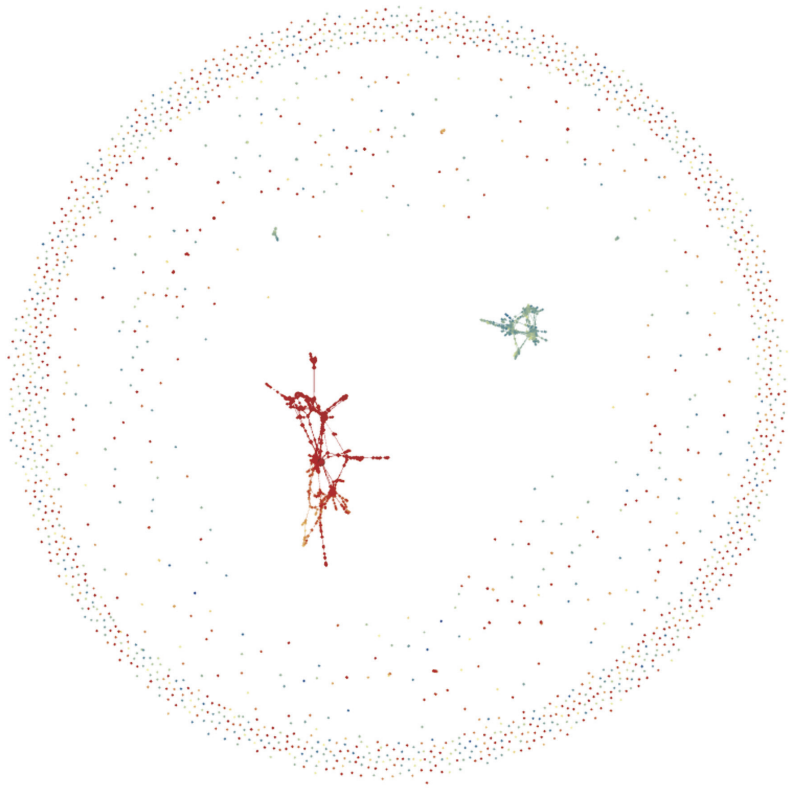


1. ábra. Pearson-korreláció-alapú koexpressziós hálózat. Balról jobbra a Brachypodium distachyon kalászának transzkriptomja a virágzást követően öt időpontban. Látványos, ám a túl sok korreláció elfedi a lényeget.

Mindezek mellett a tisztán expresszióra alapozott hálózatok sűrűsége viszonylag magas, azaz sok él köt össze nem kevés csomópontot (nódust) *(1. ábra).* A nódusok kapcsoltsági fokának eloszlása ritkán skálafüggetlen, ami viszont az egyik legárulkodóbb jele annak, ha egy biológiailag releváns hálózatra akadtunk. A kérdésünk az, hogy milyen módszerrel lehet kiszűrni a kusza és sűrű hálózatokból a releváns összefüggéseket, avagy fókuszálttá, divatos szóval okossá tenni azokat.

Az egyik, még mindig általános kérdéseket megválaszoló, de hatékony módszer a funkciók hasonlóságát, összefüggéseit veszi figyelembe. Amennyiben ismerjük a gének funkcióját, akkor egy képzeletbeli szemantikai térben meghatározhatóak a gének „szemantikai" távolságai. Az így kapott értékkel súlyozott korrelációs együtthatók egy fókuszáltabb gráfot eredményeznek. A gének funkcionális osztályozására a legjobb módszer a gének ontológiai besorolása. A génontológia maga is gráfszerkezetű, ám a csomópontokban funkciók és kategóriák állnak, és az élek a besorolás viszonyát jelölik. A szálkaperje kísérlet ily módon megszerkesztett fókuszált hálózatát mutatja be a *2. ábra.*

Amennyiben nem feltétlenül egy általános eredményre vagyunk kíváncsiak, hanem valamilyen speciális kérdésre várunk választ, akkor lehetőségeink

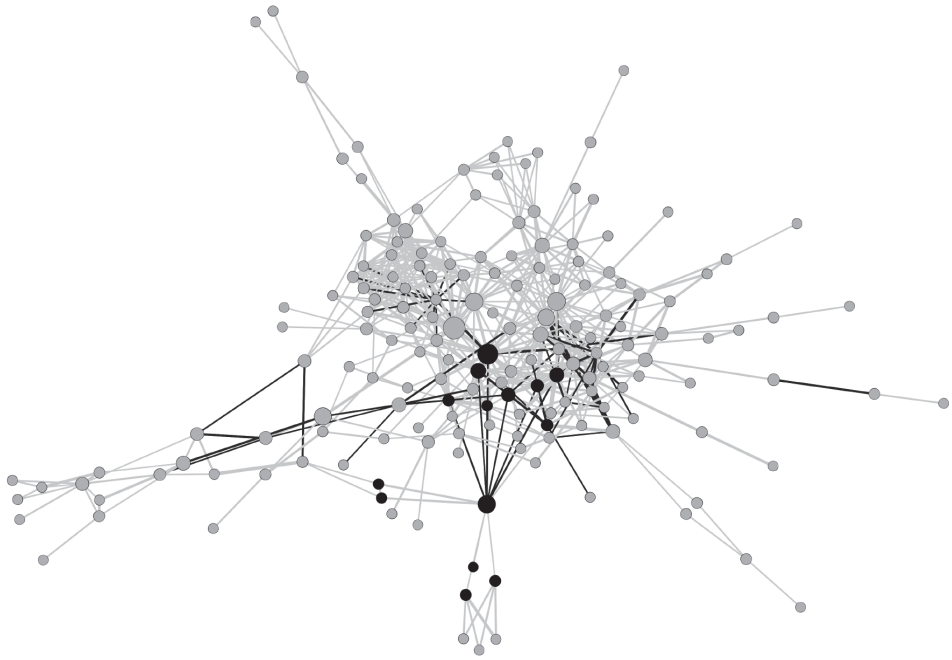


2. ábra. Génontolóia (GO) alapján fókuszált koexpressziós hálózat.
A fókuszálás eredményeként kirajzolódott egy skálafüggetlen topológia.

a kérdés jellegétől függően bővülnek. Ilyen speciális kérdés lehet, hogy mely transzkripciós faktorok azok, melyek egy bizonyos gén aktivitását szabályozzák. A szálkaperje esetén arra voltunk kíváncsiak, hogy a növény prolamin és glutenin jellegű tartalékfehérje-génjeit mely transzkripciós faktorok szabályozzák. Ezen tartalékfehérjék búzában előforduló rokonai a legfontosabb összetevői a lisztnek.

A gének szabályozásáért részben transzkripciós faktorok felelősek, amelyek egy célgén szabályozó szakaszához kötődve fejtik ki serkentő vagy gátló hatásukat. Ezen faktorok legtöbbször csoportokban dolgoznak, és kötőhelyeik jellemzőek rájuk. Azt feltételezzük, ha két kötőhely egymás közelében helyezkedik el a célgének szabályozó régióiban, akkor az ahhoz kötődő faktorok jó eséllyel együtt fejtik ki szabályozó hatásukat, avagy kölcsönhatásban állnak egymással. Köznapi példánkban tehát a teremben két ember nemcsak egyszerre beszél, hanem egymás közelében is állnak, és ez így már kielégítően erős indikáció arra nézve, hogy egymással beszélgetnek.

Tehát amennyiben a két gén korrelációja magas, illetve egyenként a célgénnel is megfigyelhető negatív vagy pozitív korreláció, akkor a célgén kötőhelyszerkezete alapján megszerkeszthető egy olyan fókuszált hálózat, amely már a biológiailag releváns összefüggéseket fogja megjeleníteni. Ez történt a szálkaperje esetén is, ahol a fent részletezett fókuszálást követően sikerült azonosítani egy fontos, szálkaperjében eddig nem ismert szabályozó gént *(3. ábra).*



3. ábra. Prolamin- és glutenin-gének szabályozó hálózata. A gráf a célgének (target) és a kölcsönható transzkripciós faktorok feltételezett kapcsolatát mutatja. Világos élek pozitív, sötét élek negatív korrelációt jeleznek. Az élek súlyai a kölcsönös információ alapján számolt távolság, amelyet a szabályozó régiók motívumszerkezete alapján fókuszáltunk.

A jó hír, hogy ahogy egyre bővülnek a biológiai ismereteink, egyre többféleképpen leszünk képesek a hálózatainkat speciális kérdésekre fókuszálni, és ily módon speciális válaszokat kihámozni az amúgy kusza és zajos koexpressziós gráfokból.

A munkában részt vettek: Pólya Sára (Eötvös Loránd Tudományegyetem), Gell Gyönyvér (MTA Agrártudományi Kutatóközpont), Juhász Angéla (MTA Agrártudományi Kutatóközpont), Gáti Zsófia (Eötvös Loránd Tudományegyetem), Jäger Katalin (MTA Agrártudományi Kutatóközpont), Fábián Attila (MTA Agrártudományi Kutatóközpont).