Enumerációs-alapú diád predikciós algoritmus alkalmazása növényi promóterek analízisében

Ph.D. értekezés tézisei

Cserháti Mátyás

Témavezető: Dr. Pongor Sándor és Dr. Györgyey János

MTA Szegedi Biológiai Központ, Növénybiológiai Intézet
Biológia Doktori Iskola
SZTE TTIK

Szeged
2011

1. Bevezető

Az abiotikus stresszhatásokra adott növényi válaszokban a gének nagy számban játszanak szerepet, mivel ez egy olyan komplex alkalmazkodási folyamatok, amelyek kiterjedten érintik a növény teljes szervezetét. A növények általában kétféleképpen válaszolnak; vagy megpróbálják fenntartani az eredeti fiziológiai állapotukat, vagy alkalmazkodnak a megváltozott körülményekhez. Abiotikus stresszhatások körébe sok olyan környezeti tényező tartozik, mely a növény vízháztartását érinti: a szárazság, a talajban előforduló sók, ozmotikus hatású anyagok, sőt bizonyos mértékig a hideg is ilyen stressz-tényező. A stressz-szignált általában a membránban található receptorok közvetítik (sokszor hormonok hatására, pl. ABA, citokininek, vagy etilén), az érzékelt jelet másodlagos hírvivők továbbítják a citoplazmán belül (pl. ROS, Ca^{2+} vagy IP molekulák). A sejten, illetve a sejtmagon belül pedig több transzkripciós faktor kölcsönhatása révén alakul ki a stresszre adott génexpressziós válasz.

A vízhiánnyal kapcsolatos abiotikus stresszválaszok szabályozási útjait általában két csoportra különítjük el, az egyikbe az ABA-függő, a másikba az ABA-független utakat soroljuk. Eközött a két szabályozási hálózat között azonban jelentős átfedések, illetve kölcsönhatások vannak, ez közösen szabályozott transzkripciós faktorokban, illetve mindkettőhöz tartozó transzkripciós faktor kötőhelyekben is megnyilvánul.

2. Célok

Mivel az abiotikus stresszválaszban szereplő gének sok esetben hasonló szabályozás alatt állhatnak, így feltételezhetjük azt, hogy ennek alapját legalább részben a hasonló szabályozó elemeket is tartalmazó a promóter régióik adhatják. Mivel összetett molekuláris genetikai folyamatról van szó, így az is feltételezhető, hogy számos gén, transzkripciós faktor, illetve szabályozó motívum együttműködéséről van szó.

Eddig számos motívumkereső algoritmust fejlesztettek ki DNS szekvenciák analízisére és motívumok előrejelzésére. Ezek leginkább rövid oligonukleotid motívumokat képesek felfedezni. Tompa és kollégái vizsgálatai alapján több jól ismert motívumkereső program vagy algoritmus érzékenysége alacsony, így jelentősen javítani lehet ezeket az algoritmusokat. Mivel olyan algoritmus-összeállítás eddig nem volt, amely együtt szabályozott gének promóter régióiban szabályozó elem párokat, azaz diádokat keressen, azt tűztük ki célul, hogy ilyen algoritmust kifejlesszünk. Az algoritmust a jövőben más, megismert genomszekvenciájú eukarióta élőlényre, így többek között különböző fontos gazdasági növény (árpa, búza, rizs, stb.) promótereinek a vizsgálatára lehet majd alkalmazni.

Az algoritmus egyik fontos eredménye az, hogy bemeneti promóter szettekben, azokra jellemző feltételezett diádokat képes optimalizáltan megtalálni. Az így kapott diádokkal pedig a vizsgált élőlény promóteromját lehet tovább elemezni, és megkeresni a többi hasonló diádot tartalmazó promótert, és ezáltal a promóterek génjeiről előrejelzést lehet vele tenni arra vonatkozóan, hogy állhatnak hasonló transzkripciós szabályozás alatt, így jó eséllyel hasonló folyamatokban is vesz részt, mint a kiindulási promóterek génjei.

3. Az algoritmus leírása

Az algoritmus DNS szekvencia diádokat keres, amelyeket az $M_1N_nM_2$ formula határoz meg, ahol M_1 a feji, illetve M_2 a farki motívum. Köztük pedig egy jellemző hosszúságú, szekvencia-specificitás nélküli spacer régió van, ami n bázis hosszú, kevés lötyögéssel megengedve. A fej és farok motívumok ugyanolyan hosszúak, és a spacer régió pedig 0-52 bp hosszú lehet.

Az elemzés több fázisban zajlik. Az első fázis abból áll, hogy a megfelelően együtt szabályozott géneket kiválasztjuk, és a hozzájuk tartozó promóter szekvenciákat egy adatsorba gyűjtjük, majd ezeket különböző szettekbe osztjuk be. A promótereket általában 2 Kbp hosszúságig vizsgáltuk, illetve akkor tekintettük rövidebbnek, ha közelebb már másik gént találtunk. A promótereket pozitív illetve negatív tanuló szettekbe, valamint pozitív illetve negatív teszt szettekbe kell beosztani.

A következő fázis a tanuló fázis, amikor is az algoritmus összeszámolja az összes lehetséges diád előfordulását a pozitív és negatív tanuló promóter szettekben. Ezek után a diádokat súlyozza, kiszámolja az ún, cdr értékét, és a diádokat eszerint rangsorolja. Egy diád cdr értékét így lehet kiszámolni:

$$cdr = \frac{N_{positive} - N_{negative}}{N_{positive}}$$

Itt a $N_{positive}$ azon promóterek száma a pozitív tanuló promóter szettben, amelyben a diád előfordul, míg $N_{negative}$ azon promóterek száma a negatív tanuló szettben, amelyben a diád szintén előfordul. A cdr értéke $-\infty$-től 1-ig terjedhet (csak azokat az eseteket vettük figyelembe ahol $N_{positive}$ nagyobb volt, mint nulla). Minél magasabb egy diád cdr értéke, annál jelentősebb szerepe lehet abban a folyamatban, amit vizsgálunk. Esetünkben ez a vízhiánnyal kapcsolatos abiotikus stressz volt. A kapott diád adatbázis további szűrés, homopolimerek, repetitív szekvenciák eltávolítása után lehet tovább használni.

A tanuló fázist követi a teszt fázis, amelyben különböző paraméterek szerint vizsgáljuk a diádokat, hogy a pozitív és negatív promóter szettek elkülönítésére

legoptimálisabb diádokat kiválasszuk. Ezek a paraméterek a következők: küszöbérték a pozitív tanuló szettben való előfordulás gyakoriságára, a spacer régió „lötyögésének" mértéke ±5bp-ig, illetve küszöbérték a diádok cdr értékére. Minden diád szett esetében visszakeressük a találatokat a pozitív és a negatív promóterek teszt szettjeiben. ROC analízis során azt az optimális diád szettet választjuk ki, amelyet majd a promóterom szűréséhez lehet használni. A promóterom szűrése alapján az egyes promótereket a bennük előforduló diádok alapján osztályozzuk és rangsoroljuk.

4. Eredmények

Az algoritmusunkat először lúdfű (*Arabidopsis thaliana)* esetében alkalmaztuk, és igazoltuk használhatóságát, majd rizs *(Oryza sativa)* esetében is alkalmaztuk; hogy a kétszikű modellnövény mellett egy egyszikű növény promóteromjának vizsgálatában is teszteljük. A lúdfű esetében az abiotikus stresszhez kapcsolódó és a kontroll (nem stressz) tanuló promóter szettekbe 125-125 promóter került, míg a megfelelő teszt szettekbe 44-44 darab. Rizs esetében 87-87 promóter került a két tanuló szettbe, míg 42 illetve 56 promóter került bele a pozitív illetve negatív teszt promóter szettbe. A pozitív szettekben a promóterek olyan génekhez tartoztak, amelyek annotációjuk alapján, egy microarray expressziós adatbázis (Genevestigator) adatai alapján, vagy saját kísérleti eredményeink alapján indukálhatóak voltak abiotikus stressz hatására.

Lúdfű esetében az optimalizálással 81 feltételezhető diádot sikerült találnunk. Ebben az esetben a minimum 14-szeri előfordulás a pozitív tanuló promóter szettben, minimum 0.9-es cdr értékkel, ±2bp lötyögéssel adta a pozitív és negatív szettek szétválasztását. Rizs esetében 38 diáddal kaptuk a legjobb eredményt, minimum 9-szer előfordulva a pozitív tanuló promóter szettben, minimum 0.89-es cdr értékkel, és lötyögés megengedése nélkül. A 81 lúdfű diád közül 38-at 11 csoportba klasztertük az alapján, hogy mennyire volt hasonló két adott diád egymáshoz képest. A hasonlóság egy Hamming-távolság volt, ahol a maximális hasonlóság 10 volt (mivel pentamer párokat vizsgáltunk). A Hamming távolság küszöbértéke 3 volt.

A promóterom szűrés alapján a lúdfű esetén a legjobb cdr értéket adó 3100 promótert vizsgáltuk, rizs esetében a legjobb 4600-t. Úgy húztuk meg az ezekhez a számokhoz tartozó határértékeket, hogy ezek között csak minimális arányban forduljanak elő fals pozitívok, azaz a program tanításához használt kontroll (nem stressz) promóterek. Lúdfű esetében a promóterom szűréssel talált génekről 78.6% esetében igazolta vissza rendelkezésre álló expressziós adat, hogy stressz indukálható. Rizs esetében ez 98.7% volt. A predikcióval 1273 (49+1224) hipotetikus vagy újonnan stressz indukálhatónak mondható gént jelzett előre az algoritmus. Rizs esetében ez 2682 gén volt (1437 új gén, illetve 1245 hipotetikus gén).

Az lúdfűben megtalált diádok, illetve a belőlük képzett klaszterek előfordulását vizsgáltuk ismert géncsaládokban. A 11-ből 7 cluster, valamint 5 egyedi diád kulcsfontosságú szerepet játszanak 5 *cor*, illetve 4 *erd* gén hálózat-szerű szabályozásában. Ezen túl 1224 feltételezhető SZEP (szabályozó elem pár) (amelyeknek a módosított cdr értékük 0.5 fölött volt) jelezhető előre, hogy valamilyen módon szerepet játszhat az abiotikus stressz válaszban.

Lúdfűben promóterome keresést végeztünk azért, hogy megnézzük, milyen SZEP-ek fordulnak elő benne. Kiszámoltuk a Jacquard-együtthatót minden promóter és minden *cor* gén promóter között. Ez alapján a SZEP tartalom különbségét számoltuk ki minden egyes promóter párra. A legjobb 25 promótert választottuk ki, amelyeknek a SZEP tartalom különbsége 0.5 alatt volt. Ezek közül 1 hipotetikus és 5 új ismeretlen funkciójú gént jelöltünk meg, amelyről az elemzés alapján feltételezhetjük, hogy hasonló funkcióval rendelkezhet, mint a *cor* gének.

Harminc rizs aldo-keto reduktáz gén vizsgálata során 28 jellemző, feltételezhető diád elemet fedeztünk fel, amelyek legalább 7 bemeneti promóterben előfordulnak, és amelynek legalább 0.9 volt a cdr értéke. A 30 AKR gén közül három expresszióját vizsgálták meg kísérletesen. Ezek közül az AKR1 (Os01g0847600) több jellegzetes diádot tartalmazott, mint az AKR2 és AKR3. Ez egybecseng a génexpressziós kísérleti eredményekkel, melyek szerint az AKR1 erősen, a másik két gén kevésbé indukálhatóak

ozmotikus stressz által. Ezzel egy, a kiindulási génszettől független esetben is igazoltuk az algoritmusunk használhatóságát.

Hat rizs géncsaládhoz (glükanázok, kitinázok, PR1, PR4, PT5, és PR9) tartozó 91 promóterrel is végeztünk elemzéseket. Ezen promótereket tartalmazó gének olyan búza génekkel homológok, amelyek szerepet játszanak a biotikus stressz válaszban. Bennük sok olyan motívumot megtaláltunk, amelyeket már ismerünk, és előfordulnak a növényi cisz-regulátor elemeket tartalmazó PlantCARE adatbázisban is (pl. a W1-box, az EIRE, és a WUN-motif). Így azt tételezzük fel, hogy a szekvencia-hasonlóságok révén ezeket, vagy ezekhez hasonló motívumokat meg lehet találni az ortológ búza gének promótereiben is. Ugyanezeket a géneket később megvizsgáltuk búzában is, és azt találtuk, hogy a prediktált diádok fele kis módosulással ugyan, de egyezik a megfelelő rizs diáddal.

Az algoritmusunkat lefuttattuk erre a 91 rizs biotikus stresszhez kapcsolódó promóterre. Közülük 13 szerepelt a pozitív tanuló szettben (mivel ezek voltak a feltételezhető ortológok), míg az összes többi a negatív tanuló promóter szettet alkották. Mivel a vizsgált promóterek száma csekély volt, tetrád diádokat kerestünk. Összesen 263 diádot talált az algoritmus, amelyeknek a cdr értéke legalább 0.9 volt. 28 olyan diád volt közöttük, amely vagy a 6 vizsgált géncsalád közül legalább 4-ben előfordult, és így statisztikailag jelentős volt, vagy megtalálhatóak voltak a PLACE adatbázisban.

Az algoritmust két jól ismert motívumkereső algoritmussal hasonlítottuk össze, a YMF-fel és a dyad-analysis-szel. Ehhez mindkét programmal vizsgáltuk a lúdfű promóteromot, kiindulásként használva ugyanazt a 125 stressz tanuló promótert, amit a saját fejlesztésű módszerrel is használtunk. Az YMF programmal 283 promótert találtunk meg, amely jelentős számú szabályozó elemet tartalmazott. Ezek közül mindössze 3 tartozott az eredeti 125-ös tanuló promóter szetthez, és csak 3.1%-uk volt stressz-indukálható a Genevestigator adatai alapján. A dyad-analysis program 149 promótert talált meg, amely jelentős számú feltételezhető szabályozó szekvencia elemet tartalmazott. Ezek közül azonban csak 1 tartozott az eredeti 125-höz. Ezeknek a 3.6%-a volt indukálható abiotikus stressz által a Genevestigator adatai alapján. Ezek az

eredmények azt mutatják, hogy a mi algoritmusunk ehhez a két programhoz képest sokkal hatékonyabban tud új, feltételezhetően stresszben szerepet játszó szabályozó elemeket előre jelezni, valamint az eredetileg vizsgált, együtt szabályozódó génekhez olyan újabb tagokat találni, amelyek sokat tartalmaznak az előre jelzett motívumokból.

Az algoritmust 64 bites IRIX64 programozási környezetben valósítottuk meg awk (GNU Awk 3.1.5.), C shell, és C (GCC 3.4.6.) szkriptek és programok kombinációival. Az algoritmusnak saját weboldala is van rövid leírással, és egy letölthető, PC-n futtatható önálló programmal: http://bhd.szbk.u-szeged.hu/dyadscan/. Bemeneti paraméterként meg kell adni a pozitív és negatív promóter szetteket, a keresett motívumok hosszát, a spacer régió maximális hosszát, a minimális előfordulást a pozitív tanuló promóter szettben, illetve a diádok minimális cdr értékét. A program eredménye egy olyan feltételezhető diád lista, ami megfelel a bemeneti kritériumoknak. A kimenetben láthatók a diádok szekvenciái, a pozitív és negatív tanuló szettekben való előfordulásuk, illetve a cdr értékeik.

A disszertáció alapját képező közlemények:

Turóczy, Z., Kis, P., Török, K., **Cserháti, M.**, Lendvai, Á., Dudits, D., and Horváth, G: Overproduction of a rice aldo-keto reductase increases oxidative and heat stress tolerance by malondialdehyde and methylglyoxal detoxification, *Plant Molecular Biology*, 2011. (kiadás alatt)

Cserháti M., Turóczy, Z., Zombori, Z., Cserző, M., Dudits, D., Pongor, S., Györgyey, J.: Prediction of new abiotic stress genes in Arabidopsis thaliana and Oryza sativa according to enumeration-based statistical analysis, *Molecular Genetics and Genomics*, 2011 (kiadás alatt).

Cserháti, M., Pongor, S. and Györgyey, J: Statistical methods for finding biologically relevant motifs in promóter regions and a few of its implementations, In: 5th International Conference of PhD Students, University of Miskolc, Hungary, 14-20 August 2005, (Eds L. Lehoczky and L. Kalmár) Published by University of Miskolc, Innovation and Technology Transfer Centre, pp. 41-46, 2005

Cserháti, M., Pongor, S., Dudits, D., and Györgyey, J: (2006). „Enumerációs módszereken alapuló algoritmusok használata promóter motívumok keresésére." Tavaszi Szél 2006 conference. Kaposvár. ISBN 963 229 773 3

Cserháti M.: Usage of enumeration method based algorithms for finding promóter motifs in plant genomes. Acta Biol Szeged 2006, 50(3-4):145.

Veronika Pós, Klára Manninger, Krisztián Halász, Éva Hunyadi-Gulyás, Emília Szájli, **Mátyás Cserháti**, Huijun Duan, Katalin Medzihradszky, János Györgyey, Noémi Lukács: Proteomic changes of the wheat apoplast associated with resistance against leaf rust. 15th International Congress of the Hungarian Society for Microbiology: July 18-20, 2007, Eötvös Loránd University (Budapest, Hungary)

Pós Veronika, **Cserháti Mátyás**, Hunyadi-Gulyás Éva, Manninger Sándorné, Györgyey János, Medzihradszky Katalin, Lukács Noémi: KÖZÖS CISZ-REGULÁLÓ elemek LEVÉLROZSDA FERTŐZÉSSEL ASSZOCIÁLT BÚZA APOPLASZTFEHÉRJÉK GÉNEXPRESSZIÓJÁBAN. A Magyar Biokémiai Egyesület 2007. évi Vándorgyûlése 2007. augusztus 26-29. Debreceni Egyetem (Debrecen, Magyarország)

További közlemények:

Cserháti, M. and Györgyey J. 2006. „Génkutatás i*n silico*", könyvfejezet: „Korszakváltás a molekuláris biológiában" c. könyvben. Szerkesztő: Dudits Dénes.

Dudits, D., **Cserháti, M.**, Miskolczi, P., Horváth, G. The growing family of plant cyclin-dependant kinases with multiple functions in cellular and developmental regulation. 2006. Cell cycle control and plant development. Editor Dirk Inzé. Blackwell Publishing, Oxford.

Cserháti, M., Turóczy, Z., Dudits, D., Horváth, G., and Györgyey, J: Bioinformatic analysis of heptamer palindromes in rice stress promóters. 3rd EPSO vonference, Visegrád, poster.

Turóczy, Z., Kis, P., **Cserháti, M.** Dare to bet? –from the *in silico* predictions to the demonstration of stress induced gene expression. 7th Biologist Days, Cluj Napoca, Romania

Turóczy, Z., Kis, P., **Cserháti, M.**, Dudits, D., Horváth, G. Response of rice AKR genes to abiotic stresses: expression profiling and enzyme activity characterization. 3rd EPSO conference, Visegrád, poster.

Dénes, D., **Cserháti, M.**, Miskolczi, P., Fehér, A., Ayaydin, F. and Horváth, G. V.: Use of Alfalfa In Vitro Cultures in Studies on Regulation of Cyclin-Dependent Kinase (CDK) Functions. 2006. Proceedings of the 11th IAPTC&B Congress, Beijing. Editors: Z. Xu, J. Li, I.K. Vasil, Y. Xue and W. Yang.

Cserháti, M., Turóczy, Z., Sečenji, M., Pongor, S., Cserző, M., Dudits, D., Horváth V., G., Györgyey, J. Növényi promóterek analízise abiotikus stressz folyamatok megértésében. 2006. Straub napok előadás, November 15-17.

András Cseri, András Palágyi, **Mátyás Cserháti**, János Pauk, Dénes Dudits, Ottó Törjék: EcoTILLING analysis of drought related candidate genes in barley. Plant Abiotic Stress - from signaling to development, 2nd meeting of INPAS(Interrnational Network of Plant Abiotic Stress), 14-17 May 2009, Tartu, Estonia