Izvirni znanstveni članek ■

Odkrivanje pravil uravnavanja izražanja genov z razvrščanjem na podlagi pravil

Tomaž Curk, Blaž Zupan, Uroš Petrovič, Gad Shaulsky

Izvleček. Zapis in struktura regulatornih regij genov pogojujeta program izražanja genov v odzivu celice na notranje in zunanje dražljaje. Ključen predpogoj za uspešno eksperimentalno potrditev in razumevanje programov uravnavanja izražanja genov je računalniško odkrivanje relacij oziroma pravil, ki opisujejo povezavo med izmerjenim izražanjem in strukturo regulatorne regije gena. Težava pri tovrstnih analizah je njihova izjemna kombinatorična kompleksnost; možnih pravil, ki jih je potrebno pri analizi preveriti, je zelo veliko. Navadno imamo namreč na razpolago mnogo potencialnih vezavnih mest transkripcijskih faktorjev, s katerimi je možno opisati strukturo regulatornih regij. V članku opisujemo metodo, ki z uporabo hevrističnih pristopov gradi omenjena pravila in predlagamo različne načine predstavitve rezultatov tovrstne analize.

Rule-based clustering for discovery of patterns in gene expression regulation

Abstract. The genetic response programs of cells to their internal state and outside environment are predominately determined by the sequence and structure of gene regulatory regions. Computational discovery of relations between gene expression, sequence and structure of regulatory regions is a prerequisite for experimental validation and a successful understanding of such programs. Given a large base of regulatory elements (*i.e.* putative or known transcription factor binding sites) which can be used to infer rules, the main obstacle posed is the high combinatorial explosion of possible rules which need to be tested. The rule-based clustering method we developed, combined with an effective presentation of discovered rules, can successfully handle this combinatorial problem.

■ **Infor Med Slov**: 2006; 11(1): 52-59

Institucije avtorjev: Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, Slovenija (TC, BZ), Baylor College of Medicine, Houston, USA (BZ, GS), Institut Jožef Stefan, Ljubljana, Slovenija (UP).

Kontaktna oseba: Tomaž Curk, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, Tržaška 25, SI-1001, Ljubljana. email: tomaz.curk@fri.uni-lj.si.

Uvod

Temeljni korak na poti določitve in razumevanja mehanizmov ter programov uravnavanja genov je analiza regulatornih regij genov.[1] Regulatorne regije so deli zapisa DNA, na katere se vežejo posebni proteini, imenovani transkripcijski faktorji, ki vzbujajo ali zavirajo transkripcijo gena v bližini regulatorne regije. Vezava transkripcijskih faktorjev je le eden izmed načinov uravnavanja izražanja genov in posledično tvorbe proteinov. Na izražanje namreč vplivajo tudi uravnavanje na nivoju strukture in oblike kromatina, epigenetski učinki, post-transkripcijsko uravnavanje, translacijsko in post-translacijsko uravnavanje ter drugi nivoji uravnavanja.[1] Ker je podatkov o slednjih zelo malo, se večina študij osredotoča na postopke iskanja povezav med vsebino regulatornega zaporedja DNA in izmerjenim izražanjem gena. Tehnologija DNA mikromrež omogoča vzporedno merjenje izražanja genov celotnega genoma, vendar so tako dobljene meritve dokazano nezanesljive.[2] Zato je nujno potrebna previdnost pri interpretaciji dobljenih rezultatov. Pomembnejše rezultate lahko na primer še dodatno preverimo z uporabo zanesljivejših metod (kot je na primer metoda PCR), ki pa običajno ne dovoljujejo hkratnega merjenja velikega števila genov.[3]

Prvi korak analize odnosov med genskim zaporedjem in njegovim izrazom je določitev regulatorne regije in vezavnih mest. Regulatorna regija se navadno nahaja v neposredni bližini kodirajočega dela gena ali pa se celo z njim prepleta. Področji se ločita v pogostosti posameznih nukleotidov ter pogostosti zaporednih trojk nukleotidov (kodonov), ki v kodirajočem delu gena določajo zaporedje aminokislin končnega proteina. Te in podobne lastnosti regulatornih in kodirajočih regij s pridom uporabljajo algoritmi za napovedovanje regulatornih regij.[4] Ker se regulatorna področja navadno ne prepisujejo v mRNA, je drugi, posredni in eksperimentalni način določanja regulatornih regij odkrivanje delov zaporedja DNA, ki se sploh kdaj prepišejo v mRNA (ang. *EST – expressed sequence tags*). Področja v bližini genov, ki se nikoli ne prepišejo, so kandidati za regulatorne regije.

Naslednji korak analize je določitev potencialnih vezavnih mest transkripcijskih faktorjev v odkritih regulatornih regijah. Vezavno mesto je navadno krajše zaporedje (4 do 20 nukleotidov),[1] ki je v manjših variacijah posameznih nukleotidov ohranjeno v regulatornih regijah reguliranih genov. Za računalniško obdelavo je najbolj primerna predstavitev vezavnega mesta v obliki matrike pogostosti nukleotidov na posamezni poziciji zaporedja, kar lahko predstavimo tudi grafično, v obliki tako imenovanih "logo-v" (tabela 1 in slika 1). Tovrstna, eksperimentalno potrjena zaporedja, najdemo tudi v javnih bazah podatkov, katere primer je baza TRANSFAC.[5] V primeru, da analiziramo gene s (še) neznanimi regulatorji oziroma neznanimi vezavnimi mesti, je možno uporabiti programska orodja, ki z lokalno ali globalno poravnavo zaporedij poiščejo pogosta, krajša podzaporedja,[6] ki jih v nadaljnji analizi lahko obravnavamo kot potencialna vezavna mesta. Podroben opis in primerjavo tovrstnih orodij podaja pregledni članek Tompa in sodelavcev.[7]

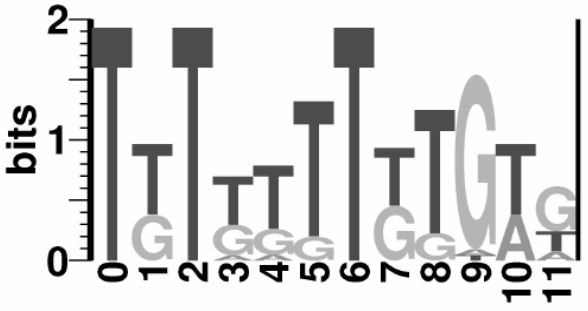
Tabela 1 Matrična predstavitev vezavnega mesta.

	0	1	2	3	4	5	6	7	8	9	10	11
A	0	0	0	0.2	0.1	0	0	0	0	0.1	0.6	0.1
C	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0.5	0	0.4	0.4	0.1	0	0.8	0.4	0.8	0	0.6
T	1.0	0.5	1.0	0.4	0.5	0.9	1.0	0.2	0.6	0.1	0.4	0.3

Pojasnilo: Prikazane so frekvence štirih nukleotidov A, C, G in T na posamezni poziciji (od 0 do 11) potencialnega vezavnega mesta.

Večina postopkov namenjenih preučevanju povezave med prisotnostjo potencialnih vezavnih mest transkripcijskih faktorjev v regulatornih regijah in izmerjenim izražanjem genov temelji na začetnem razvrščanju genov v skupine (ang. *clustering*) na podlagi izražanja, funkcije ali drugega kriterija. Razvrščanju sledi iskanje vezavnih mest značilnih za posamezno skupino genov.[8,9] Uspeh

teh pristopov je močno odvisen od števila skupin, kar je navadno parameter metode razvrščanja, ki ga mora uporabnik podati vnaprej. Rahlo spremenjeni začetni pogoji ali drugače izbran prag pri razvrščanju v skupine lahko privede do popolnoma različnih skupin in posledično do popolnoma drugačnega nabora na ta način odkritih značilnih vezavnih mest v posamezni skupini. Druga, večja pomanjkljivost teh pristopov je, da se osredotočajo samo na izključujoče se podskupine genov, čeprav je znano, da se isti gen lahko odziva na več načinov oziroma opravlja več funkcij. Primer takšnega gena podajamo v razdelku z eksperimentalnimi rezultati.



Slika 1 Enostaven in razumljiv grafični prikaz vezavnega mesta določenega v tabeli 1. Prikazana je ohranjenost nukleotidov na posamezni poziciji vezavnega mesta.

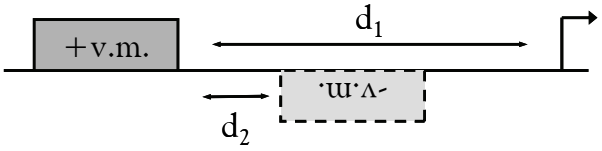
Komplementaren pristop zgoraj opisanemu prične s podatki o vezavnih mestih, ter nato išče takšne opise regulatornih regij, ki so skupne samo skupinam podobno izraženih genov.[10] Primer tovrstnega pristopa je delo Chianga in sodelavcev,[11] kjer za vsako vezavno mesto izračunajo povprečno izražanje skupine vseh genov, ki to mesto vsebujejo. Povprečno izražanje posamezne skupine nato primerjajo z izražanjem naključne, enako velike skupine genov in izračunajo statistično značilnost prisotnosti vezavnega mesta in koherence izražanja skupine genov. Takšno neodvisno obravnavanje posameznih vezavnih mest zanemarja dejansko kombinatorično naravo uravnavanja izražanja genov, ki jo določa vezava skupine transkripcijskih faktorjev na različna vezavna gena. Bolj napredne metode zato preverjajo koherenco izražanja skupin genov, katerih regulatorne regije vsebujejo kombinacijo več potencialnih vezavnih mest.[10,12] Ozko grlo tovrstnih pristopov je izčrpno, kombinatorično iskanje, ki ga te metode navadno uporabljajo in zato tipično preiskujejo samo kombinacije dveh ali največ treh vezavnih mest. Enostaven izračun pove, da se že pri preverjanju tisoč vezavnih mest število možnih dvojk in trojk hitro povzpne v več sto milijonov. Preiskovanje postane še toliko bolj zahtevno oziroma skoraj neizvedljivo, če želimo v analizo vključiti tudi razdalje med vezavnimi mesti, razdalje med vezavnimi mesti in določenimi pomembnimi mejniki v strukturi genov (na primer mesto začetka transkripcije ali translacije, itd.), njihovo orientacijo, število pojavitev posameznega vezavnega mesta, itd. Zato se večina postopkov omejuje na kombinatorično iskanje opisov z največ tremi ali izjemoma štirimi elementi.[12]

Da bi presegli zgoraj omenjene kombinatorične omejitve smo razvili hevristično metodo preiskovanja prostora opisov oziroma pravil, ki opisujejo kompleksne strukture regulatornih regij. Naša metoda razvrščanja na podlagi pravil uporablja informacijo o podobnosti izražanja genov in tako usmerjeno preiskuje le najbolj perspektivne (in koherentne) podskupine genov, ki imajo tudi podobno regulatorno strukturo.

Opisni jezik in iskanje pravil

Cilj našega postopka je poiskati pravila, ki opišejo skupno regulatorno strukturo genov, katerih izražanje je med seboj čimbolj podobno. V našem postopku podobnost med genskimi izrazi določamo na podlagi Pearsonove korelacije.

Za opis strukturnih lastnosti regulatornih regij smo uporabili bogat opisni jezik, s katerim skušamo zajeti razdaljo vezavnih mest od položaja začetka transkripcije ali translacije, medsebojno razdaljo različnih vezavnih mest ter njihovo relativno in absolutno orientacijo glede na neki referenčni položaj (slika 2).



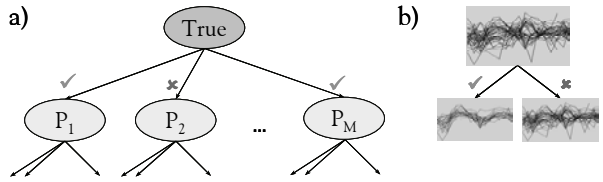
Slika 2 Elementi uporabljenega jezika hipotez, ki omogoča opis medsebojne razdalje vezavnih mest (v.m., razdalja d_2), razdalje od ATG (to je, mesta začetka translacije, razdalja d_1) ter orientacijo vezavnih mest v smeri branja DNA (+) ali v nasprotni smeri (-).

Pri tako bogatem opisnem jeziku je možnih pravil izredno veliko. Posledica je izjemno velik preiskovalni prostor in velika nevarnost, da se metode, ki v tem prostoru iščejo napovedna pravila, preveč prilagodijo podatkom (ang. *overfitting*). Da bi ta problem omilili, smo se pri razvoju metode hevrističnega iskanja zgledovali po metodi gradnje dreves za razvrščanje kot so jo predlagali Blockeel in sodelavci,[13] in tako razvili bolj splošno metodo iskanja pravil, ki poskuša preiskati le najbolj obetavne dele prostora. Vsak nadaljnji korak iskanja novih podskupin je v našem postopku ocenjen in izbran na podlagi korelacije izražanja genov v trenutno odkritih skupinah. Za nadaljnjo izostritev izberemo pravila, ki opisujejo le najbolj obetavne podskupine.

Postopek zahteva neko "ciljno" množico genov za katere želimo poiskati pravila in jih tako razvrstiti v skupine. Ta množica je lahko celoten genom, lahko pa je rezultat predhodne analize podatkov meritev DNA mikromrež, kjer na primer izberemo samo gene s statistično značilno spremembo izražanja v nekem ali več preučevanih pogojih (mutacija gena, zunanji, kemijski, mehanični, temperaturni ali kakšni drugi vplivi). Postopek prične z množico vseh genov, kar predstavimo s pravilom oziroma njegovim pogojem za proženje *TRUE*. To je tudi edino pravilo v začetni množici odkritih pravil. Sklepni del pravila je povprečno izražanje genov, ki jih opisuje pogojni del. Postopek iskanja poskuša izostriti trenutno odkrita pravila z dodajanjem novih pogojev. Na primer, pogojni del pravila, ki ga zapišemo z "M_1" in ki zahteva prisotnost vezavnega mesta M_1, lahko izostrimo z dodatnim pogojem, da je to vezavno mesto orientirano v pozitivno smer, kar zapišemo z "M_1+." Prvotni pogoj "M_1" lahko izostrimo tudi z zahtevo, da se vezavno mesto M_1 pojavlja na razdalji od -100 do -80 nukleotidov relativno glede na začetek translacije (ATG), kar zapišemo kot "M_1 @-100..-80(ref:ATG)." Za referenco lahko uporabimo neko drugo vezavno mesto, na primer M_2, kar zapišemo z "M_1 @ -100..-80(ref:M_2)." Prvotno pravilo "M_1" lahko izostrimo tudi z dodatno zahtevo o prisotnosti vezavnega mesta M_2, kar zapišemo z "M_1 in M_2."

Vsako izostreno pravilo pokrije manj genov od prvotnega pravila, vendar pa pri postopku zahtevamo, da pravila pokrijejo vsaj N ciljnih genov. N je parameter algoritma, ki ga določi uporabnik. Hkrati pa se mora podobnost med geni, ki ustrezajo izostrenemu pogoju, značilno povečati glede na medsebojno podobnost izražanja genov, ki ustrezajo prvotnemu pravilu. Če so ti pogoji izpolnjeni, nov pogoj oz. pravilo dodamo v množico pravil za nadaljnjo izostritev. V nasprotnem primeru pa ga le ocenimo in obdržimo, če se uvrsti med K najboljših pravil (uporabnik določi vrednost parametra K). Da bi še dodatno omejili preiskani prostor je velikost množice pravil za nadaljnjo izostritev omejena na največ L najboljših pravil (L je parameter algoritma, ki ga določi uporabnik). Kvaliteto pravila merimo s povprečno razdaljo izražanja genov v skupini, ki jo pravilo opisuje. Značilnost povečanja podobnosti genov med prvotnim in izostrenim pravilom merimo z uporabo F-testa, kjer v osnovi testiramo zmanjšanje variance v razdaljah med geni znotraj prvotne in izostrene skupine. Razdalja v izražanju genov je v našem postopku definirana s formulo: 1.0 – Pearsonova korelacija. Pravilo sprejmemo, če pokrije vsaj N ciljnih genov in opisuje skupino genov, katerih povprečna medsebojna razdalja je manjša od parametra D, ki ga določi uporabnik. Za osnovni korak preiskovalnega algoritma glej sliko 3, za preiskovalni algoritem pa sliko 4.

Velja poudariti, da lahko sprejeta pravila pokrijejo tudi gene izven ciljne množice. Metodo lahko zato uporabimo za iskanje genov, ki sicer niso bili vključeni v ciljno skupino, čeprav bi jih na podlagi izražanja in strukture regulatorne regije morali obravnavati skupaj z ostalimi geni v ciljni skupini.



Slika 3 a) Osnovni korak preiskovanja je izostritev posameznega pravila v trenutni množici odkritih pravil. b) Opazovana podobnost genov v skupini, ki jo opisuje izostreno pravilo, se mora značilno povečati v primerjavi s skupino prvotnega pravila. Dve različni izostritvi istega prvotnega pravila lahko privedeta do dveh različno homogenih podskupin (kljukica predstavlja značilno, križec pa neznačilno povečanje koherence izražanja genov dveh izostrenih pravil).

množica L najboljših pravil za izostritev B = {True}
množica K najboljših odkritih pravil R = {}
WHILE B ≠ {}
  iz B vzemi najboljše pravilo Pb
  FOR EACH k IN 1..število vseh vezavnim mest
    Pravilo Pb izostri z vezavnim mestom Mk in tako tvori pravilo Pn.
    IF pravilo Pn sprejemljivo AND značilno povečanje v podobnosti med pokritimi geni v Pb in Pn THEN v B dodaj novo pravilo Pn in v B ohrani le L najboljših pravil.
  Pravilo Pb dodaj v R, če je med K najboljšimi.
vrni množico K najboljših odkritih pravil R.

Slika 4 Preiskovalni algoritem. Izostritev pravila Pb z vezavnim mestom Mk se dejansko opravi na več načinov, z dodajanjem pogojev o prisotnosti, orientaciji in razdalji vezavnega mesta Mk glede na že prisotne člene v pravilu.

Predlagana metoda se razlikuje od klasičnih prekrivnih algoritmov za iskanje pravil, kot je na primer algoritem CN2,[14] saj omogoča odkrivanje prekrivajočih skupin genov. Osnovna verzija algoritma CN2 namreč iterativno odstranjuje primere (v našem primeru gene), ki jih v dani iteraciji opiše najboljše odkrito pravilo, ter nato ponovi iskanje na tako zmanjšani množici. To ponavlja dokler ne pokrije vseh primerov.

Predlagana metoda pa odkriva nova pravila vse dokler je pravila možno izostriti in se pri tem ne ozira na dejansko pokritost skupine genov s pravili.

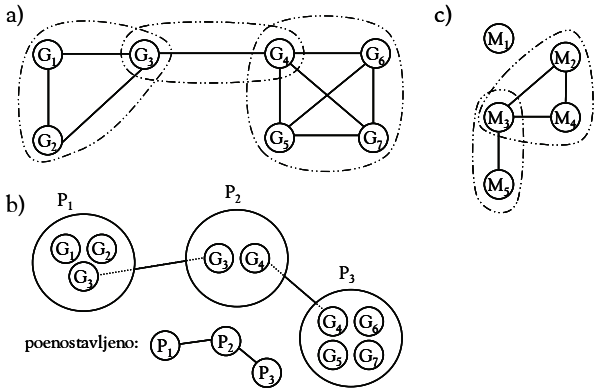
Izčrpno preiskovanje prostora relativno enostavnih (kratkih) pravil hitro preraste v neobvladljiv problem zaradi prej omenjene kombinatorične eksplozije. Izrazita prednost tu opisanega hevrističnega pristopa je zmožnost učinkovitega opisovanja uravnavanja izražanja genov, kjer lahko za osnovo obravnavamo več tisoč vezavnih mest in iz njih tvorimo kompleksnejše opise. V razdelku z eksperimentalnimi rezultati navajamo število hevristično preiskanih pravil in to primerjamo s številom pravil, ki bi jih sicer morali preveriti z izčrpnim preiskovanjem.

Prikaz odkritih pravil

Zaradi bogatega opisnega jezika in predvsem zaradi zmožnosti odkrivanja prekrivajočih skupin je število odkritih pravil oziroma skupin lahko zelo veliko. Da bi uporabnik lažje analiziral odkrita pravila, jih je smiselno prikazati z uporabo grafov, ki omogočajo boljši vpogled v skupne značilnosti in strukturo odkritih pravil in podskupin.

Kot osnovni prikaz odkritih pravil smo uporabili graf, kjer med seboj povežemo vse gene, ki jih opisuje posamezno pravilo (slika 5a). Pri velikem številu genov in odkritih pravil lahko tovrstni izris hitro postane zasičen in nepregleden. Naslednji višji nivo abstrakcije, s katerim lahko prikažemo iste rezultate vendar na dosti manj zasičen način, je graf pravil (slika 5b). Tu vsako vozlišče predstavlja eno pravilo. Dve pravili povežemo, če opisujeta vsaj en ali poljubno izbrano število skupnih genov. Z višanjem praga dobimo vse manj povezan graf, ohranjene povezave pa lahko kažejo na veliko strukturno in izrazno podobnost vpletenih genov. Zadnji nivo abstrakcije (slika 5c) opisuje podobnosti med pravili na podlagi skupnih členov, ki nastopajo v pravilih. Vozlišča so v tem primeru členi pravil. Povežemo jih, če nastopajo v (vsaj enem) istem pravilu. Podobno lahko tudi tu spreminjamo prag zahtevane prisotnosti člena v

različnem številu pravil in tako opazujemo kateri člani so centralni, torej nastopajo v mnogo pravilih in so zato morda vezavna mesta nekih splošnih regulatorjev. Opazujemo lahko tudi kateri člani postanejo hitro nepovezani in lahko predstavljajo zato vezavna mesta, ki določajo specifično izražanje podskupine genov.



Slika 5 a) Mreža genov. b) Prikaz povezanosti pravil oziroma skupin genov, ki jih pravila določajo. c) Prikaz podobnosti pravil, kjer povežemo pravila s skupnimi člani. Pravilo P_1 ("M_1")testira prisotnost vezavnega mesta M_1, pravilo P_2 ("M_2 in M_3 in M_4") prisotnost vezavnih mest M_2, M_3 in M_4, pravilo P_3 ("M_3 in M_5") pa prisotnost vezavni mest M_3 in M_5 v regulatornih regijah genov.

Eksperimentalni rezultati

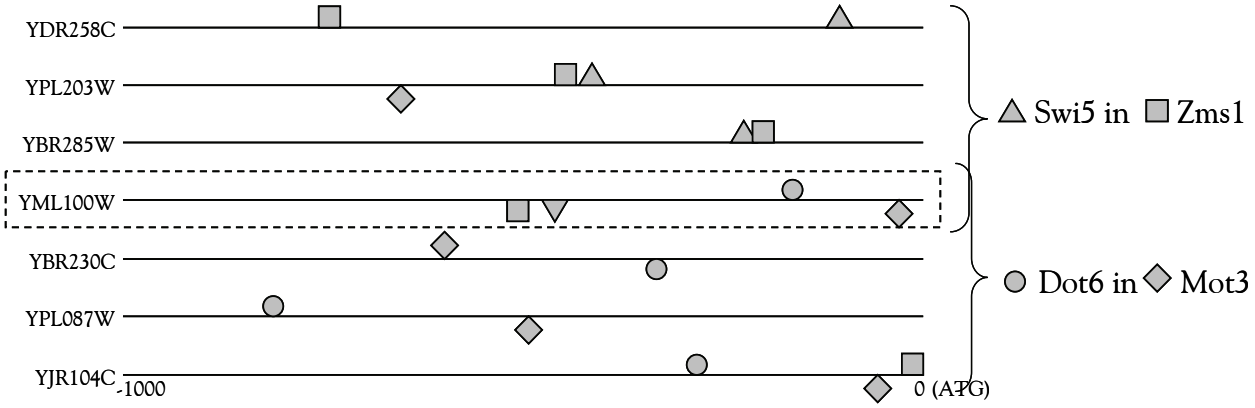
Z metodo razvrščanja na podlagi pravil smo analizirali regulatorne regije kvasnih genov, katerim so Gasch in sodelavci[15] izmerili in določili značilno spremembo izražanja v različnih stresnih pogojih. Za ciljno množico smo izbrali 281 genov s povečanim izražanjem v stresnih pogojih. Vzeli smo regulatorne regije dolžine 1000 nukleotidov od mesta začetka translacije (ATG, na poziciji 0). Za podatke o vezavnih mestih smo uporabili bazo znanih in eksperimentalno potrjenih vezavnih mest,[16] ter jim dodali vezavna mesta, ki smo jih odkrili s programom za lokalno poravnavo zaporedij MEME[6]. Pri gradnji pravil smo tako upoštevali približno 3100 vezavnih mest. Iskali smo skupine z najmanj štirimi ciljnimi geni (N=4) in povprečno Pearsonovo korelacijo genov v skupini nad 0.45 (D=1.0 – 0.45=0.55). Velikost množice pravil za nadaljnjo izostritev je bila omejena na L=1000 pravil. Glede na dano dolžino regulatorne regije, so relativne razdalje med elementi v razponu od –1000 do 1000 nukleotidov. Razdalje med elementi smo zaokroževali na 40 nukleotidov natančno, kar pomeni, da smo obravnavali 50 (=2000/40) možnih različnih razdalj med elementi. Večina odkritih pravil je sestavljenih iz dveh členov. Najdaljše odkrito pravilo vsebuje štiri člene in opisuje razdalje med štirimi vezavnimi mesti ter njihovo orientacijo. Število vseh možnih pravil, ki opisujejo samo prisotnost in orientacijo štirih vezavnih mest je ogromno (vezavnih mest je 3100, ki so lahko s pozitivno, negativno ali brez določene orientacije, faktor 3):

$$\binom{3100 \cdot 3}{4} \approx 3.11 \cdot 10^{14}$$

Če bi želeli izčrpno preiskati tudi pravila, ki opisujejo razdalje med vezavnimi mesti, bi se zgornje število možnih pravil povečalo za faktor $50^4 = 6.25 \cdot 10^6$. Naša metoda hevrističnega iskanja je na tej domeni pregledala $3.3 \cdot 10^9$ pravil oziroma manj kot 0.0011% prostora vseh možnih pravil s štirimi členi, kar se je na delovni postaji s procesno enoto Pentium 4, 3.4 GHz izvajalo približno eno uro in dvajset minut.

S predstavljeno metodo smo uspešno odkrili in opisali regulatorne regije skupine genov, za katere je bilo predhodno že pokazano, da so povezane s transkripcijskimi faktorji Msn2p, Msn4p in Yap1p. Prav tako smo odkrili druga, domnevna regulatorna vezavna mesta, ki nastopajo v kombinaciji z že znanimi mesti, kar nakazuje možnost novih mehanizmov uravnavanja. S pregledom grafa pravil (ni prikazan) smo hitro odkrili dve prekrivajoči skupini, prikazani na sliki 6. Pravili opisujeta regulatorne regije dveh različnih skupin genov s skupnim genom YML100W. Pregled znanih določitev funkcij (ang. *annotation*) v genski ontologiji[17] (ang. *Gene Ontology*) nam pove, da je značilni biološki proces

zgornjih treh genov na sliki 6 *celični metabolizem proteinov* (ang. *cellular protein metabolism*), značilni biološki proces spodnjih treh genov pa *odziv na stres* (ang. *response to stress*). Pregled določitve funkcije gena YML100W (imenovanega tudi TSL1) nam pokaže, da sta genu eksperimentalno določeni obe funkciji. Ta primer posredno potrjuje zmožnost metode za odkrivanje funkcijsko smiselno se prekrivajočih skupin genov.



Slika 6 Primer dveh enostavnih odkritih pravil, ki zahtevata le prisotnost posameznih vezavnih mest, in prikaz regulatorne regije genov, ki jih pravili opisujeta. Posamezni simbol ponazarja vezavno mesto. Obe pravili ("Swi5 in Zms1" in "Dot6 in Mot3") opisujeta regulatorno regijo gena YML100W. Zaporedja so poravnana glede na začetek translacije (ATG) na skrajni desni. Prikazana je regulatorna regija dolžine tisoč nukleotidov.

Zaključek

Dobljeni eksperimentalni rezultati kažejo, da je možno dokaj učinkovito in relativno hitro poiskati kompleksne opise regulatornih regij skupin med seboj podobno izraženih genov. Različni prikazi dobljenih rezultatov še dodatno pripomorejo k boljšemu razumevanju in biološki interpretaciji. Glavno uporabnost predstavljene metode vidimo v luči iskanja dodatnih dokazov, da so geni v neki teoretično ali pa eksperimentalno določeni skupini dejansko tudi regulatorno medsebojno povezani oziroma imajo skupne regulatorje. Metoda lahko na ta način postavi izbrane hipoteze, ki jih je moč preveriti z naknadnimi eksperimenti. Implementacija tu predstavljene metode v obliki spletne aplikacije, ki bo omogočila biologom enostavno uporabo opisanega orodja, je v delu.

Literatura

1. Wasserman WW, Sandelin A: Applied bioinformatics for the identification of regulatory elements. Nat Reviews Genet 2004; 5: 276-87.
2. Kothapalli R, Yoder SJ, Mane S, et al.: Microarray results: how accurate are they? BMC Bioinformatics 2002; 3:22.
3. Chuanqui RF, Bonner RF, Best CJM, et al.: Post-analysis follow-up and validation of microarray experiments. Nature Genetics Supplement 2002; 32:509-514.
4. Bajic VB, Tan SL, Suzuki Y, et al.: Promoter prediction analysis on the whole human genome, Nature Biotechnology 2004; 22:1467-1473.
5. Wingender E, Dietze P, Karas H, et al.: TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res 1996; 24(1): 238-41.
6. Bailey TL, Elkan C: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 1994; 2: 28-36.
7. Tompa M, Li N, Bailey TL, et al.: Assessing computational tools for the discovery of

transcription factor binding sites, Nature Biotechnology 2005; 23:137-144.
8. Pennacchio LA, Rubin EM: Genomic strategies to identify mammalian regulatory sequences. Nat Reviews Genet 2001; 2: 100-9.
9. Conlon EM, Liu XS, Lieb JD, et al.: Integrating regulatory motif discovery and genome-wide expression analysis. Proc Natl Acad Sci USA 2003; 100(6): 3339-44.
10. Pilpel Y, Sudarsanam P, Church GM: Identifying regulatory networks by combinatorial analysis of promoter elements. Nat Genet 2001; 29(2): 153-9.
11. Chiang DY, Brown PO, Eisen MB: Visualizing associations between genome sequences and gene expression data using genome-mean expression profiles. Bioinformatics 2001; 17(supp. 1):S49-S55.
12. Beer MA, Tavazoie S: Predicting gene expression from sequence, Cell 2004; 117:185-198.
13. Blockeel H, De Raedt L, Ramon J: Top-down induction of clustering trees. Machine Learning, Proceedings of the 15th International Conference 1998; Morgan Kaufmann.
14. Clark P, Niblett T: The CN2 induction algorithm. Machine Learning 1989; 3(4): 261-283.
15. Gasch AP, Spellman PT, Kao CM, et al.: Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 2000; 11(12): 4241-57.
16. Lee TI, Rinaldi NJ, Robert F, et al.: Transcriptional regulatory networks in Saccharomyces cerevisiae. Science 2002; 298:799-804.
17. The Gene Ontology Consortium: Gene Ontology: tool for unification of biology. Nature Genetics 2000; 25:25-29.