Integracija strukturnih omejitev pri izpeljavi gensko regulatornih omrežij

Žiga Pušnik, Miha Moškon
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Večna pot 113, Ljubljana
ziga.pusnik@fri.uni-lj.si, miha.moskon@fri.uni-lj.si

Izvleček

Inferenca gensko regulatornih omrežij (GRO) iz ekspresijskih podatkov je še vedno težak problem. Število genov je velikokrat večje od števila poskusov, kjer gensko izražanje še dodatno spremlja določena mera šuma. Zato predlagamo uporabo strukturnih omejitev pri izpeljavi GRO na podlagi predhodnega znanja v obliki referenčnih omrežij. Naša ideja izvira iz dejstva, da vsebujejo GRO vzorce, tako imenovane motive, ki se pojavljajo bistveno pogosteje, kot bi to pričakovali v naključno generiranih omrežjih. Uporabo predhodnega znanja dosežemo s spreminjanjem uteži posameznega gena v cenovni funkciji linearne regresije. Uteži spreminjamo iterativno z gradientnim spustom. Naš pristop temelji na že uveljavljeni parcialno korelacijski metodi SPACE. S spreminjanjem uteži na podlagi prisotnosti motivov, porazdelitve stopenj genov in pričakovanega števila regulatornih genov za odtenek izboljšamo točnost, natančnost, priklic in F 1 oceno omrežij izpeljanih iz GRO bakterije E. coli.

Ključne besede: Gensko regulatorna omrežja, Inferenca omrežij, Integracija strukturnih podatkov, Referenčna omrežja

Abstract

The inference of gene regulatory networks (GRNs) from the gene expression data remains a challenging task. The number of genes is significantly larger than the number of experiments, where each experiment contains a noise component. We impose structural constraints on the inferred gene regulatory network based on the structure of reference GRNs. Our idea is motivated by the fact that GRNs contain a vast number of patterns, i.e. motifs, that are significantly more common than in randomized networks. We impose these constraints by modifying the weights of genes contributing to the joint loss function in the regression problem. We modify weights iteratively with gradient descent. Our approach is based on the already established partial correlation method dubbed SPACE. By extracting the expected number of regulatory genes, gene degree distribution and motifs from the reference network, we have improved by a small margin the inference accuracy, precision, recall and F1 score in the inference of GRNs derived from the GRN of the E. coli bacteria.

Keywords: Gene regulatory network, integrative data, network inference, reference network

1 UVOD

Izpeljavo gensko regulatornih omrežij (GRO) iz ekspresijskih podatkov lahko umestimo v širši kontekst računske biologije kot ključen korak za odkrivanje zapletenih bioloških procesov. Metode za inferenco omrežij v splošnem delimo na (1) verjetnostne pristope, (2) korelacijske in parcialno korelacijske pristope, ter (3) pristope na podlagi teorije informacij [Allen et al., 2012]. Rezultat teh metod je usmerjen graf, neusmerjen graf, Bayesovska mreža ali Boolova mreža.

Težava, s katero se srečujemo pri izpeljavi GRO iz ekspresijskih podatkov, je nizko število eksperimentov. Tipično je število eksperimentov veliko manjše od števila vozlišč (genov) v omrežju. Slednje močno vpliva na pravilnost izpeljanega omrežja. Da bi zmanjšali velikost prostora rešitev, nekateri pristopi najprej združijo gene s podobno dinamiko v tako imenovane meta-gene [Martin et al., 2007].

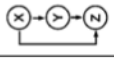
Nekoliko naprednejše metode, kot je na primer metoda SPACE [Peng et al., 2009], ohranijo število genov in so zato primernejše za inferenco omrežij z večjo biološko relevanco. SPACE določi povezave GRO na podlagi ocene parcialnih korelacij, pridobljenih z minimizacijo regularizirane cenovne funkcije.

Problem obstoječih metod za inferenco omrežij je neupoštevanje predhodnega znanja o splošni strukturi podobnih omrežij. Zaradi majhnega števila eksperimentov in prisotnosti šuma so lahko izpeljana omrežja nerealna, oziroma se prekomerno prilagodijo podatkom. V primerjavi z naključno generiranimi omrežji so GRO redko povezana in vsebujejo gene z velikim številom povezav. Ti imajo ključno vlogo pri genski regulaciji, saj nadzirajo gene s skupno globalno funkcijo, kot je na primer celični odziv. Poleg tega GRO vsebujejo pogosto ponavljajoče se vzorce vozlišč oziroma fragmente, ki se pojavljajo pogosteje, kot bi pričakovali v naključnih omrežjih. Take fragmente imenujemo motivi. Zaradi evolucijskih prednosti, ki izhajajo iz načina izvedbe različnih funkcij, so se motivi razvili neodvisno v različnih organizmih. Predhodno poznavanje prisotnosti motivov in njihove strukture je moč uporabiti za izboljšanje rezultatov inference.

Nekatere metode predhodno znanje v omejenem obsegu že upoštevajo. Parcialno korelacijski pristop ESPACE [Yu et al., 2017], ki izhaja iz metode SPACE, uvede dodatno kazen. Ta je nižja za gosto povezane gene z več kot sedmimi povezavami [Yu et al., 2017]. ESPACE se na realnih omrežjih odreže bolje kot SPACE, če predhodno vemo, katera vozlišča so gosto povezana. Če to znanje ni na voljo, je ESPACE enaka metodi SPACE. Poleg tega metoda ESPACE ne upošteva dodatnega predhodnega znanja o omrežju, kot so motivi, povezanost omrežja in splošna struktura GRO.

Naš pristop temelji na metodi SPACE, je kontekstno odvisen in vključuje integracijo predhodnega znanja, ki ga podamo v obliki referenčnih omrežij. Cilj takšnega pristopa je izpeljati GRO iz transkriptomskih podatkov in v omrežju ohraniti strukturne lastnosti referenčnih omrežij, kot je povezanost omrežja in prisotnost motivov. Na tak način želimo zagotoviti izpeljavo pravilnih in biološko verodostojnih GRO. Integracijo predhodnega znanja dosežemo z iterativnim spreminjanjem uteži cenovne funkcije. Uteži spreminjamo glede na pričakovano število povezav, porazdelitev stopnje vozlišč in prisotnostjo fragmentov s tremi vozlišči. Tak pristop lahko uporabimo v različnih aplikacijah. V kontekstu sistemske biologije in medicine pogosto poznamo strukturo referenčnega omrežja (npr. zdrava celica), spremembe, ki pripeljejo do določene bolezni (npr. mutacije) pa so neznane ali poznane le delno. V kontekstu sintezne biologije lahko rešujemo podoben problem. Želen odziv sintetičnega sistema je poznan, pri čemer so spremembe referenčnega omrežja, na primer omrežja gostiteljske celice, potrebne za pridobitev takega odziva, poznane le delno. Uporabo predlaganega pristopa v nadaljevanju ponazorimo na podlagi desetih referenčnih omrežji in štirih testnih omrežij. Omrežja smo z orodjem GeneNetWeaver [Schaffter et al., 2011] izpeljali na podlagi GRO *E. coli*.

Tabela 1: **Primeri motivov organizma E. coli. Negativna avtoregulacija pripomore k povečanju robustnosti sistema. Naprej usmerjena zanka generira in zakasni pulz. Geni motiva bi-fan imajo skupno globalno funkcijo, kot je na primer odziv sistema na zunanje dražljaje. Vsebina tabele je povzeta po [Alon, 2006, Alon, 2007].**

Motiv	Struktura	*Z*-vrednost
Negativna avtoregulacija		32
Naprej usmerjena zanka		10
Bi-fan		13

2 INTEGRACIJA STRUKTURNIH OMEJITEV

Naj ima naključno generirano omrežje N vozlišč in E povezav. V naključnem Erdos-Renyi omrežju [Erdös and Rényi, 1959] je verjetnost povezave enaka $p = E/N^2$. Če primerjamo vzorce, ki se pojavljajo v bioloških GRO, lahko ugotovimo, da je pojavnost določenih fragmentov bistveno večja, kot bi to pričakovali v naključno generiranih omrežjih [Alon, 2007]. Tabela 1 prikazuje različne vrste motivov GRO

E. coli in njihovo *Z*-vrednost [Alon, 2006]. V tem primeru predstavlja *Z*-vrednost število standardnih od- klonov pojavnosti motivov v bioloških GRO v primerjavi z naključnimi omrežji z enakim številom vozlišč in povezav[Alon, 2006].

Splošno strukturo GRO narekujejo motivi in regulatorni geni z velikim številom povezav. Podatke o motivih GRO lahko poleg same strukture omrežja v proces inference vključimo posredno preko referenčnega omrežja. Referenčno omrežje se lahko v tem primeru nanaša na omrežje sorodnega organizma, za katerega strukturo že poznamo, ali pa na omrežje pred izvedbo določene perturbacije (npr. pred pojavom bolezni).

Izhajamo iz parcialno korelacijske metode SPACE [Peng et al., 2009]. Parcialna korelacija je mera line-

arne povezanosti dveh spremenljivk, pri čemer omilimo vpliv vseh ostalih spremenljivk. Gena v GRO sta povezana, če je njuna parcialna korelacija različna od nič. Parcialno korelacijo lahko izrazimo tudi preko inverza kovariančne matrike oziroma matrike natančnosti Σ^{-1} z enačbo

$$\rho^{ij} = \frac{-\sigma^{ij}}{\sqrt{\sigma^{ii}\,\sigma^{jj}}}, \tag{1}$$

kjer je σ^{ij} element i-te vrstice j-tega stolpca matrike natančnosti. Peng in sodelavci reformulirajo problem iskanja povezav v omrežju v regresijski problem iskanja neničelnih parcialnih korelacij [Peng et al., 2009]. Zato predlagajo sledečo cenovno funkcijo

$$L(\Theta, \sigma, Y) = \frac{1}{2}\left(\sum_{i=1}^{m} w^i \left|\left| Y_i - \sum_{i \neq 1} \rho^{ij} \sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}}\, Yj \right|\right|^2\right) + \lambda \sum_{ij} \left|\rho^{ij}\right|, \tag{2}$$

kjer je $\Theta = (\rho^{12}, \ldots, \rho^{(p-1)p})$, $\sigma = diag(\Sigma^{-1})$ in λ regularizacijski parameter. Utež w_i predstavlja prispevek napake gena Y_i k skupni napaki. Regularizacija $L1$ dodatno spodbuja k inferenci redko povezanih omrežij, ki so v naravi pogostejša. Metoda SPACE deluje po korakih. V prvem koraku parameter σ le ocenimo $1/\hat{\sigma}^{ii} \approx var(y_i)$ in ga v sledečih iteracijah posodabljamo. Metodo ustavimo ob konvergenci ali po preteku maksimalnega števila iteracij. Neničelne parcialne korelacije predstavljajo povezave izpeljanega omrežja [Peng et al., 2009, Yu et al., 2017]. Zaradi slednjega metoda SPACE izpelje neusmerjen graf.

Topološke značilnosti referenčnega omrežja v pristop vključimo s prilagajanjem uteži cenovne funkcije. Vhodni argumenti našega pristopa so podatki genske ekspresije Y ter matrike povezanosti referenčnih omrežij. Iz referenčnih omrežij na začetku izluščimo pričakovano število povezav p_r, porazdelitev vozlišč d_r in prisotnost fragmentov s tremi vozlišči f_r [Gal et al., 2020]. Pričakovano število povezav je skalar. Pri porazdelitvi vozlišč opazujemo vozlišča, ki imajo od 0 do vključno 10 povezav, pri čemer agregiramo vsa vozlišča z več kot desetimi povezavami v skupen razred. Ker se ukvarjamo z neusmerjenimi grafi, sta možna samo dva različna fragmentka s tremi vozlišči, to sta trikotnik (angl. *triangle*) in pot (angl. *path*). Želimo si izpeljati omrežje, ki je topološko podobno referenčnemu omrežju.

Naš pristop deluje iterativno. V prvi iteraciji topološke lastnosti izpeljanega omrežja še niso na voljo, zato utežimo vsak gen z enako utežjo $w_i = 1$. Nato apliciramo metodo SPACE. Na podlagi pričakovanega števila povezav pi, porazdelitve stopenj vozlišč di in prisotnosti fragmentov f_i izpeljanega omrežja minimiziramo cenovno funkcijo

$$C(d_i, f_i, p_i) = aS\left(\left|\left|d_i - d_r\right|\right|_2\right) + bS\left(\left|\left|f_i - f_r\right|\right|_2\right) + cS(p_i - p_r). \tag{3}$$

Zaradi različnih magnitud pri posameznih topoloških lastnostih cenovno funkcijo še dodatno normaliziramo s sigmoidno funkcijo S. Koeficiente a, b in c smo določili eksperimentalno, in sicer $a = 0,6$, $b = 0,2$, $c = 0,2$. Uteži iterativno v stotih iteracijah prilagajamo z gradientnim spustom in stopnjo učenja $\alpha = 0,1$

Naš pristop poleg iterativnega popravljanja uteži vnaša dodatno računsko zahtevnost z numeričnim odvajanjem.

$$w_i = w_i - \alpha \frac{\delta C}{\delta w_i}. \tag{4}$$

Tabela 2: **Primerjava rezultatov inference omrežij pridobljenimi z osnovno metodo SPACE in z metodo SPACE s prilagojenimi utežmi. Omrežja s štiridesetimi vozlišči smo iz GRO *E. coli* pridobili z orodjem GeneNetWeaver.**

	Osnoven				Prilagojene uteži			
Omrežje	**Točnost**	**Natančnost**	**Priklic**	**F1**	**Točnost**	**Natančnost**	**Priklic**	**F1**
1	0,823	0,252	0,514	0,338	0,833	0,266	0,514	0,351
2	0,68	0,122	0,39	0,186	0,684	0,124	0,39	0,188
3	0,663	0,166	0,776	0,274	0,668	0,168	0,776	0,277
4	0,811	0,22	0,465	0,299	0,82	0,231	0,465	0,308

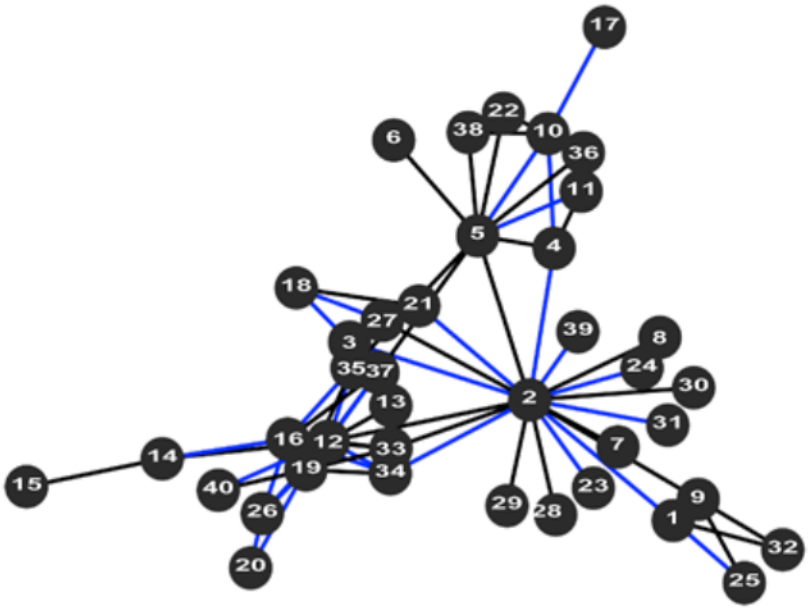
3 DISKUSIJA IN REZULTATI

Pristop smo ovrednotili na sintetičnih podatkih štirih GRO, ki smo jih izpeljali iz večjega GRO *E. coli*. Takšna omrežja so za naš pristop še vedno obvladljiva, hkrati pa izhajajo iz realnega omrežja, ki vsebuje že omenjene lastnosti GRO, kot je prisotnost motivov. Izpeljana omrežja vsebujejo 40 vozlišč in imajo vsaj 20 regulatornih genov. Podatke genske ekspresije smo z orodjem GeneNetWeaver generirali na enak način kot v izzivu DREAM4[Marbach et al., 2009, Marbach et al., 2010, Prill et al., 2010] s stohastičnimi diferencialnimi enačbami. Za inferenco omrežij smo uporabili časovno serijo genskega izražanja in izražanje z utišanimi geni (t.i. knockout in knockdown geni).

Tabela 2 prikazuje rezultate izpeljave inference omrežij z metodo SPACE brez modifikacij in z modificiranimi utežmi. Kljub nizki natančnosti in F1 oceni naš pristop v določeni meri izboljša rezultate inference. Omrežja izpeljana s prilagajanjem uteži imajo manj lažno pozitivnih povezav, kar je razvidno iz ocene natančnosti. Slika 1 prikazuje graf omrežja 1, izpeljanega z modifikacijami uteži. Zaradi pregle-dnosti lažno pozitivnih povezav ne prikazujemo. Že na manjšem grafu, kot je ta, so razvidne topološke lastnosti GRO. Opazimo lahko prisotnost gosto po-vezanih vozlišč, manjše število trikotnikov in večje število vozlišč stopnje 1.

4 ZAKLJUČEK

Predlagali smo pristop integracije strukturnih omejitev pri izpeljavi GRO s spremembami uteži, preko



Slika 1: **Rezultati inference omrežja 1, pridobljeni z uporabo metode SPACE s prilagojenimi utežmi. Črne povezave so pravilno izpeljane, modre povezave so prisotne a niso izpeljane. Napačno izpeljanih povezav zaradi večje preglednosti ne prikazujemo.**

katerih smo posredno opisali pričakovane lastnosti referenčnih omrežij. Te smo določili na podlagi pričakovanega števila povezav, porazdelitve stopenj vozlišč in prisotnosti motivov s tremi vozlišči. Te informacije pristop avtomatsko pridobi iz podanih referenčnih omrežij. Predlagani pristop temelji na parcialno korelacijski metodi SPACE [Peng et al., 2009]. Pokazali smo, da takšna integracija predhodnega znanja lahko izboljša točnost izpeljanih omrežij, četudi le za odtenek.

Naša ideja je izhajala iz dejstva, da struktura GRO sledi določenim pravilom. Taka omrežja so redko povezana ter vsebujejo motive in regulatorne gene z velikim številom povezav [Alon, 2006]. Naš pristop zato lahko uporabimo v primeru izpeljave GRO na podlagi poznanega referenčnega omrežja sorodnega organizma ali poznanega omrežja pred izvedbo sekvence določenih perturbacij omrežja. Pomankljivost našega pristopa je višja časovna zahtevnost. Ker naš pristop prilagaja uteži iterativno, s tem vpeljemo dodaten računski čas. Naša druga skrb je prekomerno prilagajanje strukture izpeljanega GRO omrežja k referenčnemu omrežju. Slednji problem bi lahko naslovili z vpeljavo dodatnega skalirnega faktorja. Če bi nastavili vrednost skalirnega faktorja na 0, bi izpeljali omrežje brez omejitev referenčnega omrežja. V nasprotnem primeru bi upoštevali topološke značilnosti referenčnega omrežja sorazmerno z velikostjo skalirnega faktorja. V nadaljnjem delu bomo pristop uporabili na realnih omrežjih genske regulacije. Pri tem bomo kot referenčna omrežja preizkusili omrežja sorodnih organizmov. Zanimalo nas bo predvsem, kako se pristop obnese pri večjih omrežjih ter kakšna je občutljivost metode na perturbacije referenčnega omrežja. Zaradi nekoliko slabših rezultatov inferenčne metode SPACE se bomo osredotočili na drugačne pristope, kot so hevristični algoritmi in multikriterijska optimizacija.

LITERATURA

[1] [Allen et al., 2012] Allen, J. D., Xie, Y., Chen, M., Girard, L., and Xiao, G. (2012). Comparing statistical methods for constructing large scale gene networks. *PloS one*, 7(1):e29348.

[2] [Alon, 2006] Alon, U. (2006). *An introduction to systems biology: design principles of biological circuits*. Chapman and Hall/CRC.

[3] [Alon, 2007] Alon, U. (2007). Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450.

[4] [Erdös and Rényi, 1959] Erdös, P. and Rényi, A. (1959). On random graphs, i. *Publicationes Mathematicae (Debrecen)*, 6:290–297.

[5] [Gal et al., 2020] Gal, E., Perin, R., Markram, H., London, M., and Segev, I. (2020). Neuron geometry underlies universal network features in cortical microcircuits. *bioRxiv*, page 656058.

[6] [Marbach et al., 2010] Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D., and Stolovitzky, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the national academy of sciences,* 107(14):6286–6291.

[7] [Marbach et al., 2009] Marbach, D., Schaffter, T., Mattiussi, C., and Floreano, D. (2009). Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *Journal of computational biology*, 16(2):229–239.

[8] [Martin et al., 2007] Martin, S., Zhang, Z., Martino, A., and Faulon, J.-L. (2007). Boolean dynamics of genetic regulatory networks inferred from microarray time series data. *Bioinformatics*, 23(7):866–874.

[9] [Peng et al., 2009] Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746.

[10] [Prill et al., 2010] Prill, R. J., Marbach, D., Saez-Rodriguez, J., Sorger, P. K., Alexopoulos, L. G., Xue, X., Clarke, N. D., Altan-Bonnet, G., and Stolovitzky, G. (2010). Towards a rigorous assessment of systems biology models: the dream3 challenges. *PloS one*, 5(2):e9202.

[11] [Schaffter et al., 2011] Schaffter, T., Marbach, D., and Floreano, D. (2011). GeneNetWeaver: in silico bench- mark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16):2263– 2270.

[12] [Yu et al., 2017] Yu, D., Lim, J., Wang, X., Liang, F., and Xiao, G. (2017). Enhanced construction of gene regulatory networks using hub gene information. *BMC bioinformatics*, 18(1):186.

Žiga Pušnik je asistent in doktorski študent na Fakulteti za računalništvo in informatiko Univerze v Ljubljani. Njegovi raziskovalni interesi so računska in sintezna biologija, strojno učenje in inferenca omrežij. Trenutno poučuje pri predmetih Osnove digitalnih vezij, Računalniška arhitektura ter Brezžična in mobilna omrežja.

Miha Moškon je izredni profesor na Fakulteti za računalništvo in informatiko Univerze v Ljubljani. Raziskovalno se ukvarja z vzpostavitvijo in uporabo metod za računsko-podprto modeliranje in analizo bioloških sistemov na področju sistemske biologije in medicine ter za računsko-podprto snovanje bioloških sistemov na področju sintezne biologije.