Univerza v Ljubljani
Fakulteta za računalništvo in informatiko
Fakulteta za matematiko in fiziko

Pristop matrične faktorizacije za gradnjo napovednih modelov iz heterogenih podatkovnih virov

Marinka Žitnik

Ljubljana, 2012

Univerza v Ljubljani
Fakulteta za računalništvo in informatiko
Fakulteta za matematiko in fiziko

Pristop matrične faktorizacije za gradnjo napovednih modelov iz heterogenih podatkovnih virov

Marinka Žitnik

Delo je pripravljeno v skladu s Pravilnikom o podeljevanju Prešernovih nagrad študentom, pod mentorstvom prof. dr. Blaža Zupana.

Ljubljana, 2012

Povzetek

Na mnogih področjih človekovega udejstvovanja je v zadnjem času značilna hitra rast podatkov tako v količini kot tudi raznolikosti. Skupna obravnava heterogenih virov informacij je zato velik izziv. V delu predlagamo novo računsko metodo za gradnjo napovednih modelov iz heterogenih podatkovnih virov. Ta uporablja simetrično kazensko matrično tri-faktorizacijo in prioritizira napovedi z ocenjevanjem verjetnosti iz razcepnih matričnih faktorjev. Metoda predstavlja nov koncept vmesne strategije združevanja podatkov, ki ni le splošno uporaben, temveč daje tudi zanesljive in zelo dobre rezultate. Glavne prednosti predlaganega pristopa so elegantna matematična formulacija problema, zmožnost integracije vseh vrst podatkov, ki se lahko izrazijo v matrični obliki, in visoka napovedna točnost.

Uspešnost metode smo v našem delu eksperimentalno preverili z napovedovanjem genskih določitev amebe *D. discoideum*, pri čemer smo združili podatke o genskih izrazih, omrežja proteinskih interakcij in znane genske pripise. Model, zgrajen s predlagano metodo, dosega višjo točnost od strategij zgodnje in pozne integracije, ki združujejo vhodne podatke ali napovedi, in so se v preteklosti izkazale za koristne pri povečanju točnosti napovednih modelov.

S predlaganim pristopom smo napovedali tudi nekaj genov *D. discoideum*, ki imajo lahko pomembno vlogo v bakterijski rezistenci in pred tem niso bili povezani s to funkcijo. Ameba je zelo pomemben modelni organizem, znana tudi kot plenilec bakterij, tudi takih, ki so človeku nevarne in so v zadnjem času vse bolj odporne na razvite antibiotike. Do sedaj je bila znana le peščica genov, vpletenih v poti amebine bakterijske rezistence. Naše napovedi petih novih kandidatnih genov so bile eksperimentalno potrjene na sodelujoči instituciji (Baylor College of Medicine, Houston, ZDA). Razširitev seznama teh genov je ključna v študijah mehanizmov bakterijske rezistence in lahko prispeva pri snovanju alternativnih metod klasičnega antibakterijskega zdravljenja.

Ključne besede

matrična faktorizacija, heterogeni podatkovni viri, združevanje podatkov, napovedni model, genska določitev, bakterijska rezistenca

Abstract

Today we are witnessing rapid growth of data both in quantity and variety in all areas of human endeavour. Integrative treatment of these sources of information is a major challenge. We propose a new computation method for inference of prediction models. The method uses symmetric penalized matrix tri-factorization and prioritizes predictions by estimating probabilities from matrix factors. The approach represents a new concept of data integration by intermediate strategy which is both generally applicable as well as highly effective and reliable. Major advantages of the approach are an elegant mathematical formulation of the problem, ability to integrate any kind of data that can be expressed in matrix form, and high predictive accuracy.

We tested the effectiveness of the proposed method on predicting gene annotations of social amoebae *D. discoideum*. The developed model integrates gene expressions, protein-protein interactions and known gene annotations. Model, inferred by proposed method, achieves higher accuracy than standard techniques of early and late integration, which combine inputs and predictions, respectively, and have in the past been favourably reported for their accuracy.

With the proposed approach we have also predicted that there are a few genes of *D. discoideum* that may have a role in bacterial resistance and which were previously not associated with this function. Amoebae is an important model organism, also known for its predation of bacteria, among which are some dangerous to humans and have recently been increasingly resistant to developed antibiotics. Until now, only a handful of genes were known to participate in related bacterial recognition pathways of amoebae. Our predictions of five new genes were experimentally confirmed in wet-lab experiments at the collaborating institution (Baylor College of Medicine, Houston, USA). Expanding the list of such genes is crucial in the studies of mechanisms for bacterial resistance and can contribute to the research in development of alternative antibacterial therapy.

Keywords

matrix factorization, heterogeneous data sources, data fusion, prediction model, gene annotation, bacterial resistance

Kazalo

"What we know is a drop, what we don't know is an ocean."

Sir Isaac Newton (1642–1727)

Seznam uporabljenih kratic in simbolov

BLAST	osnovni algoritem lokalne poravnave zaporedij (angl. *basic local alignment search tool*)
EC	označevalni sistem encimov (angl. *enzyme commission*)
MF	matrična faktorizacija (angl. *matrix factorization*)
miRNA	mikro RNA (angl. *micro RNA*)
mRNA	sporočilna RNA (angl. *messenger RNA*)
NNDSVD	nenegativni dvojni singularni razcep (angl. *nonnegative double singular value decomposition*)
omrežje **PPI**	omrežje proteinskih interakcij (angl. *protein-protein interaction network*)
PCA	analiza glavnih komponent (angl. *principal component analysis*)
PMF	kazenska matrična faktorizacija (angl. *penalized matrix factorization*)
pogoji **KKT**	pogoji Karush-Kuhn-Tucker v teoriji optimizacije
SDP	semidefinitno programiranje (angl. *semidefinite programming*)
SNMNMF	večrazcepna hkratna redka matrična faktorizacija z mrežno regularizacijo (angl. *sparse network-regularized multiple nonnegative matrix factorization*)
SVD	singularni razcep (angl. *singular value decomposition*)
SVM	metoda podpornih vektorjev (angl. *support vector machines*)
tri-NMF	nenegativna matrična tri-faktorizacija (angl. *nonnegative matrix tri-factorization*)
tri-PMF	kazenska matrična tri-faktorizacija (angl. *penalized matrix tri-factorization*)
tri-SPMF	simetrična kazenska matrična tri-faktorizacija (angl. *symmetric penalized matrix tri-factorization*)
učenje **PU**	učenje iz pozitivnih in neoznačenih primerov (angl. *learning from positive and unlabeled examples*)
WT	sev divjega tipa, nemutirani protein (angl. *wild-type*)

Poglavje 1

Uvod

V mnogih nalogah podatkovne analize obstaja več naravnih predstavitev oziroma pogledov, ki z različnih zornih kotov osvetljujejo isti nabor entitet. Posamezne predstavitve so pogosto nepopolne, pomanjkljive in neskladne. Združevanje različnih podatkovnih virov je eden izmed ključnih izzivov v analizi podatkov in pomembno vpliva na kakovost odkrivanja struktur v domeni [14]. V primeru skladnih podatkovnih virov – enaki vzorci so razpoznani v vseh dostopnih predstavitvah entitet – se problem združevanja virov poenostavi v določanje konsenznega modela, ki opisuje vzorce, skupne vsem predstavitvam. Če je med podatkovnimi viri razdor, je potreben učinkovit postopek združevanja predstavitev, ki odpravi neskladja, razpozna skupne vzorce in hkrati ohranje tiste, ki so edinstveni za vsako predstavitev.

Široko razlikovanje med tehnikami združevanja podatkovnih virov opredeljuje tri splošne strategije [49], (i) *zgodnja integracija* zajema neposredno združevanje podatkov iz več pogledov v eno samo učno množico pred začetkom učnega procesa, (ii) *vmesna integracija* zajema izračune relacij med podatkovnimi viri in proizvede kombinirane poglede, ki so nato dani učnemu algoritmu, in (iii) *pozna integracija*, ki uporabi učni algoritem ločeno v vsaki predstavitvi in nato združuje rezultate. Tako zgodnja kot tudi pozna strategija, ki združujeta vhodne podatke ali napovedi, sta se izkazali za koristni pri povečevanju točnosti napovednih modelov [46, 54, 33]. Vendar vmesni

pristop omogoča ocenjevanje prispevkov podatkovnih virov in vključevanje znanih relacij med viri v učni proces, zato je ta pristop zaželjen. Vpeljavo vmesne integracije v gradnjo napovednih modelov ovirajo učni algoritmi. Ti na področju strojnega učenja najpogosteje niso prilagojeni za sočasno obravnavo večih predstavitev, njihove razširitve niso enostavne ali zahtevajo prilagoditev, uporabno le za izbrano nalogo [69, 44].

V delu predlagamo novo računsko metodo za gradnjo napovednih modelov iz heterogenih podatkovnih virov, ki uporablja matrično faktorizacijo in ocenjevanje verjetnosti napovedi iz razcepnih matričnih faktorjev ter temelji na vmesni integraciji podatkov.

Uporaba matrične faktorizacije za vmesno združevanje heterogenih virov je nov pristop, obetaven na raznih področjih. Eno izmed teh je bioinformatika, kjer hitra rast količine in raznolikosti javno dostopnih bioloških podatkov zahteva učinkovite računske metode za združevanje heterogenih predstavitev [31, 43]. Matrične faktorizacije so dobro znane v bioinformatični skupnosti, posebno nenegativne faktorizacije so uspešne pri odkrivanju skritih vzorcev v genskih izrazih iz mikromrež [10], odkrivanju bioloških procesov, razvrščanju in gručenju genskih izrazov [21], razpoznavanju vzorcev zvijanja proteinov, napovedovanju genskih funkcij [62], proteinskih interakcij [7] in koreguliranih genskih ter miRNA modulov [69]. Večina metod uporablja en sam vir informacij, čeprav je v splošnem za vse opisane naloge na voljo več virov. Za napovedovanje genskih funkcij so med njimi zapisi genskih izrazov, znani genski pripisi, proteinske interakcije, biološke informacije, pridobljene iz literature in drugi viri [6].

V pričujočem delu predstavimo matrične faktorizacije in njihovo uporabo v bioinformatiki ter raziščemo strategije združevanja heterogenih virov pri gradnji napovednih modelov (poglavje 2). Predlagano ogrodje in matrično faktorizacijo uvedemo v poglavju 3 in napovedovanje iz razcepnih matričnih faktorjev v poglavju 4. Uspešnost ogrodja za združevanje heterogenih virov ponazorimo na dveh aktualnih problemih v bioinformatiki. To sta napovedovanje genskih določitev amebe *D. discoideum*, predstavljeno v poglavju 5,

in napovedovanje bakterijske rezistence te amebe v poglavju 6. Rezultati slednje uporabe so izrednega pomena, saj je trenutno znanih le nekaj genov in poti v *D. discoideum*, ki so odgovorni za bakterijsko razpoznavo in rezistenco. Z novo metodo smo napovedali več genov te amebe, ki so bili do sedaj neznani v analizi bakterijske rezistence. Geni so bili potrjeni z biološkimi eksperimenti, ki so jih na osnovi naših napovedi izvedli v laboratorijih prof. dr. Gada Shaulskyja in prof. dr. Adama Kuspe na Baylor College of Medicine v Houstonu, ZDA. Odzivi ameb na bakterije v okolju so pomembni za okužbe pri ljudeh, ker so ti verjetno razvili odzive iz poti, ki so jih uporabljali primitivni evkarionti za obrambo pred bakterijami. Velika razširjenost bakterij, odpornih na antibiotike, zmanjšuje učinkovitost klasičnih antibakterijskih zdravil, zato je določitev novih genov in genskih poti pomembna za odkrivanje alternativnih metod antibakterijskega zdravljenja pri človeku.

Alternativa predlaganemu pristopu z matrično faktorizacijo so dobro znane tehnike nadzorovanega učenja. Spodbudni rezultati računske analize in opravljenih eksperimentov v biološkem laboratoriju potrjujejo prednosti predlagane metode vmesne integracije pred tehnikami nadzorovanega učenja z zgodnjim ali poznim združevanjem. Hkrati je predlagana metoda zaradi univerzalnega pristopa, ki seveda ni omejen le na področje bioinformatike, uporabna tudi na drugih področjih, kjer se srečujemo z mnogoterimi viri podatkov.

Poglavje 2

Pregled sorodnega dela

2.1 Postopki matrične faktorizacije

Faktorizacijski modeli so pogosto naravni pri analizi različnih vrst tabelaričnih podatkov. Ti med drugim vključujejo uporabniške izbire iz seznama predmetov, genske izraze in zbirke dokumentov ali slik. Osnovna predpostavka teh modelov je, da je pomembne vidike podatkov možno zajeti z zmanjšanjem razsežnosti predstavitve, ki ustreza določenim omejitvam.

Ponazorimo predpostavko na preprostem primeru, ki opisuje uporabniške ocene filmov. Nabor je predstavljen z matriko, pri čemer uporabniki ustrezajo vrsticam in filmi stolpcem matrike. Koncept faktorizacijskega modela v tem primeru predpostavlja, da so uporabnikove želje določene z majhnim številom dejavnikov in uporabniški zapis predstavljen s pripadnostmi uporabnika posameznim dejavnikom. V linearnem faktorizacijskem modelu je vsak dejavnik vektor in uporabniški zapis ustreza linearni kombinaciji dejavnikov. Koeficienti linearne kombinacije predstavljajo uporabnika v nizko razsežnem prostoru. Učenje dejavnikov ustreza *faktorizaciji* matrike ocen v več (pogosto dve) manjših matrik ali iskanju faktorizacije, ki dobro ustreza podatkovni matriki.

Matrična faktorizacija je lahko koristna na več načinov:

Rekonstrukcija signala. Predstavitev v nizko razsežnem prostoru ustreza

skritim signalom ali procesom, ki so bili opazovani posredno [37, 70, 36]. Sodobno uporabo je mogoče najti v analizi genskih izrazov, ki je namenjena zaznavi celičnih procesov in stanj na osnovi opazovanih ravni izražanja genov.

Stiskanje podatkov z izgubami. Tradicionalno analiza glavnih komponent uporablja nizko dimenzionalni prikaz kot kompaktno predstavitev, ki še vedno vsebuje večino pomembnih informacij prvotno visoko razsežne vhodne predstavitve [22, 10]. Delo z zmanjšano predstavitvijo zmanjša pomnilniške zahteve in računske stroške.

Razumevanje struktur v podatkih. Matrična faktorizacija se pogosto uporablja v nenadzorovanem učenju za razpoznavanje strukture domene, kot je zbirka dokumentov ali slik [13, 1]. Vsak element (dokument ali slika) ustreza vrstici v matriki in stolpci ustrezajo postavkam (besede ali barvni nivoji). S faktorizacijo zgrajene matrike lažje razumemo odnose med elementi in glavne načine variacij v podatkih.

Napovedovanje. Če so elementi v podatkovni matriki le delno opazovani (n.pr. uporabniki niso videli in ocenili vseh filmov), se lahko matrična faktorizacija uporablja za napovedovanje neznanih vrednosti (n.pr. ocen filmov) [32, 11, 40].

Raznovrstne uporabe matrične faktorizacije se razlikujejo po omejitvah, ki so včasih naložene na faktorizacijo in v meritvah odstopanj med matričnimi razcepnimi faktorji in dejanskimi podatki. Če matrični faktorji nimajo omejitev, so matrike, ki jih je možno razcepiti v dve manjši matriki, natanko matrike z rangom, ki je omejen s številom dejavnikov. To z drugimi besedami pomeni, da je aproksimacija matrike s faktorizacijo brez omejitev enakovredna aproksimaciji z matriko nizkega matričnega ranga.

Najpogostejša oblika matrične faktorizacije išče dobro matrično aproksimacijo nizkega ranga polne podatkovne matrike, tako da minimizira vsoto kvadratov razlik med matrično aproksimacijo in vhodno podatkovno matriko.

Če imajo stolpci vhodne matrike ničelne srednje vrednosti, je ta postopek bolje znan kot analiza glavnih komponent (PCA), saj razcepni faktorji predstavljajo glavne smeri sprememb v podatkih. Takšna matrična aproksimacija je dana v zaprti obliki s singularnim razcepom matrike (SVD). Singularni razcep predstavlja lastne vrednosti in lastne vektorje kovariančne matrike vrstic in stolpcev podatkovne matrike.

V mnogih primerih je primerno razmisliti o drugih funkcijah izgube v aproksimaciji (n.pr. ko so ciljni elementi nenumerični ali ustrezajo verjetnostnim modelom) in omejitvah, naloženih na matrično faktorizacijo, primera slednjih sta nenegativnost matričnih faktorjev [36, 10] ali redkost matričnih faktorjev [22, 69]. Z omejitvami se je možno naučiti več razcepnih faktorjev in jih razlikovati glede na njihovo kakovost. Pogost zaplet v izvedbi matrične faktorizacije je, da je podatkovna matrika le delno opazovana (redka), kar zahteva posebne napore za ustrezno upoštevanje neznanih vrednosti v procesu faktorizacije.

Za raziskovalne namene smo razvili programsko knjižnico v jeziku Python, imenovano Nimfa[1] [71], ki vsebuje več modelov matričnih faktorizacij (n.pr. nenegativni, standardni, gladki, večrazcepni), konkretnih konstrukcij faktorizacij z omejitvami in brez omejitev ter mere za ocenjevanje kakovosti razcepnih faktorjev. Nekatere elemente te programske knjižnice opišemo tudi v besedilu, ki sledi.

2.1.1 Matrična aproksimacija s faktorizacijo

Naj bodo vhodni podatki predstavljeni v podatkovni matriki $\mathbf{Y} \in \mathbb{R}^{n \times m}$, ki jo želimo predstaviti s produktom dveh matrik $\mathbf{U}\mathbf{V}^T$, pri čemer $\mathbf{U} \in \mathbb{R}^{n \times k}$ in $\mathbf{V} \in \mathbb{R}^{m \times k}$. Označimo vrstice matrike $\mathbf{Y}$ z vektorji podatkov Y_i. Vsak tak vektor Y_i je aproksimiran z linearno kombinacijo $U_i\mathbf{V}^T$ vrstic matrike $\mathbf{V}^T$ – te imenujemo *faktorje* in vnosi v matriki $\mathbf{U}$ so *koeficienti* linearnih kombinacij. V geometrijski predstavitvi ta formalizacija pomeni: vektorji $Y_i \in \mathbb{R}^m$ so aproksimirani s k-razsežnim linearnim podprostorom, ki ga razpenjajo

[1] Programska knjižnica je dostopna na `http://nimfa.biolab.si`.

vrstični vektorji matrike $\mathbf{V}^T$. Geometrijski pogled je seveda simetričen, pri čemer so stolpci matrike $\mathbf{Y}$ izraženi z linearno kombinacijo stolpcev matrike $\mathbf{U}$. Matriki $\mathbf{U}$ in $\mathbf{V}$ bomo imenovali *razcepna matrična faktorja*.

Če razcepna matrična faktorja nista izpostavljena drugim omejitvam, lahko matriko $\mathbf{Y}$, katere rang je navzgor omejen s številom faktorjev k, faktoriziramo brez izgub v $\mathbf{X} = \mathbf{U}\mathbf{V}^T$. V zgornji razpravi ostaja nedoločen pojem aproksimacije podatkovne matrike. V kakšnem smislu želimo aproksimirati vhodno matriko? Kaj je merilo razhajanja med podatki $\mathbf{Y}$ in modelom $\mathbf{X}$, ki ga minimiziramo? Ali je možno matrično aproksimacijo razumeti kot gradnjo verjetnostnega modela?

2.1.2 Frobeniusova razdalja matrične aproksimacije

Osnovno merilo kakovosti, ki je pogosto sestavni del kompleksnih kriterijskih funkcij pri matričnih faktorizacijah z omejitvami, je razlika kvadratov napak ali Frobeniusova razdalja. To je Frobeniusova norma razdalje med podatki $\mathbf{Y}$ in modelom $\mathbf{X}$

$$||\mathbf{Y} - \mathbf{X}||^2_{\text{Fro}} = \sum_{ia} (\mathbf{Y}_{ia} - \mathbf{X}_{ia})^2. \tag{2.1}$$

V metodi analize glavnih komponent je dovoljen dodatni aditivni člen srednjih vrednosti. To je, vhodna matrika $\mathbf{Y} \in \mathbb{R}^{n \times m}$ je aproksimirana z matriko $\mathbf{X} \in \mathbb{R}^{n \times m}$ matričnega ranga k in z vrstičnim vektorjem $\mu \in \mathbb{R}^m$ tako, da je Frobeniusova razdalja

$$\sum_{ia} (\mathbf{Y}_{ia} - (\mathbf{X}_{ia} + \mu_a))^2 \tag{2.2}$$

minimalna. Matrika $\mathbf{X}$ zajame glavne smeri sprememb vrstic v matriki $\mathbf{Y}$ glede na vektor μ. Z dodanim vektorjem srednjih vrednosti μ problem matrične faktorizacije ni več simetričen, saj so vrstice in stolpci obravnavani različno.

Minimizacijo Frobeniusove razdalje je možno razumeti kot ocenjevanje

maksimalnega verjetja v prisotnosti aditivnega neodvisnega in enako porazdeljenega (i. i. d.) Gaussovega šuma s konstantno varianco [60]. Predpostavimo, da opazimo naključno matriko generirano na način

$$\widetilde{\mathbf{Y}} = \mathbf{X} + \widetilde{\mathbf{Z}}, \tag{2.3}$$

kjer je $\mathbf{X}$ matrika ranga k in $\widetilde{\mathbf{Z}}$ matrika i. i. d. normalnih slučajnih spremenljivk z ničelnimi srednjimi vrednostmi in konstantno varianco σ^2. Potem je verjetje matrike $\mathbf{X}$ ob dani opazovani matriki $\mathbf{Y}$ enako

$$\begin{aligned} \log P(\widetilde{\mathbf{Y}} = \mathbf{Y}|\mathbf{X}) &= -\frac{nm}{2}\ln 2\pi\sigma^2 - \sum_{ia}\frac{(\mathbf{Y}_{ia} - \mathbf{X}_{ia})^2}{2\sigma^2} \\ &= -\frac{1}{2\sigma^2}||\mathbf{Y} - \mathbf{X}||_{\text{Fro}} + \text{Const.} \end{aligned} \tag{2.4}$$

Enačba (2.4) kaže, da je iskanje maksimalnega verjetja matrike $\mathbf{X}$ ekvivalentno minimizaciji Frobeniusove razdalje.

Frobeniusova aproksimacija nizkega ranga je razširjena predvsem zaradi enostavnosti izračuna. Znano je, da je ta aproksimacija dana s k glavnimi komponentami singularnega razcepa matrike $\mathbf{Y}$. Matrične faktorizacije, ki minimizirajo druge mere kakovosti aproksimacije, pogosto niso enostavne. Kljub temu je potrebno obravnavati druge pristope, saj model Gaussovega šuma ni vedno primeren.

2.1.3 Ne-Gaussovi pogojni verjetnostni modeli

Minimizacija Frobeniusove razdalje med matriko $\mathbf{X}$ nizkega ranga in podatkovno matriko $\mathbf{Y}$ ustreza verjetnostnemu modelu, v katerem je vsak element $\mathbf{Y}_{ia}$ realizacija slučajne spremenljivke $\widetilde{\mathbf{Y}}_{ia} = \mathbf{X}_{ia} + \widetilde{\mathbf{Z}}_{ia}$. Pri tem je $\widetilde{\mathbf{Z}}_{ia} \sim \mathcal{N}(0, \sigma^2)$ Gaussova napaka z ničelno srednjo vrednostjo in varianco σ^2. Ta verjetnostni model lahko razumemo kot določitev pogojne porazdelitve $\widetilde{\mathbf{Y}}|\mathbf{X}$, pri čemer je slučajna spremenljivka $\widetilde{\mathbf{Y}}_{ia}|\mathbf{X}_{ia}$ normalno porazdeljena s srednjo vrednostjo $\mathbf{X}_{ia}$ in varianco σ^2 neodvisno od vnosa (i, a) v slučajni

matriki $\widetilde{\mathbf{Y}}$.

Pogosto so namesto normalne porazdeltve primerni drugi modeli verjetnostnih porazdelitev za $\widetilde{\mathbf{Y}}_{ia}|\mathbf{X}_{ia}$ [24, 57]. Ti so določeni z enoparametrično družino porazdelitev $p(y; x)$.

Poseben razred pogojnih porazdelitev izhaja iz aditivnih, ne nujno Gaussovih modelov, to je $\widetilde{\mathbf{Y}}_{ia} = \mathbf{X}_{ia} + \widetilde{\mathbf{Z}}_{ia}$, kjer so $\widetilde{\mathbf{Z}}_{ia}$ neodvisne slučajne spremenljivke s fiksno porazdelitvijo. Tovrstne verjetnostne modele imenujemo *aditivni šumni modeli*. Včasih aditivni model šuma $\widetilde{\mathbf{Y}} = \mathbf{X} + \widetilde{\mathbf{Z}}$, v katerem je slučajna matrika $\widetilde{\mathbf{Z}}$ neodvisna od $\mathbf{X}$, opustimo. To se zgodi, ko je šum multiplikativen ali je $\mathbf{Y}$ matrika z diskretnimi vrednostmi.

Oglejmo si modeliranje klasifikacijske matrike $\mathbf{Y}$ z binarnimi oznakami. Z vložitvijo oznak v množico realnih števil (n.pr. nič–ena ali ± 1) lahko uporabimo postopke matrične faktorizacije, ki minimizirajo srednjo kvadratno napako. Toda domneva Gaussovega verjetnostnega modela ni primerna. Naravni verjetnostni model je logistični model, parametriziran z matriko $\mathbf{X} \in \mathbb{R}^{n \times m}$ nizkega ranga, tako da je verjetnost $P(\widetilde{\mathbf{Y}}_{ia} = +1|\mathbf{X}_{ia}) = g(\mathbf{X}_{ia})$ neodvisna za vsak ia in je g logistična funkcija $g(x) = \frac{1}{1+e^{-x}}$. Model $\mathbf{X}$ poiščemo z iskanjem največjega verjetja $P(\widetilde{\mathbf{Y}} = \mathbf{Y}|\mathbf{X})$.

Logistični model je le primer splošnega pristopa [15], v katerem pogojne porazdelitve $\widetilde{\mathbf{Y}}_{ia} = \mathbf{X}_{ia} + \widetilde{\mathbf{Z}}_{ia}$ pripadajo družini eksponencialnih porazdelitev in so vrednosti $\mathbf{X}_{ia}$ naravni parametri.

2.1.4 Druge ocenitvene funkcije

Pogosto se želimo izogniti ocenjevanju največjega verjetja in pripadajočih mer kakovosti aproksimacije, ki ustrezajo logaritmu verjetij elementov v matriki $\mathbf{X}$ do homogenosti in aditivnosti natančno

$$\begin{aligned} \mathcal{D}(\mathbf{X}; \mathbf{Y}) &= \sum_{ia} \text{loss}(\mathbf{X}_{ia}; \mathbf{Y}_{ia}), \\ \text{loss}(x; y) &= -\log P(y|x). \end{aligned} \tag{2.5}$$

Zaželena je neposredna obravnava mer kakovosti aproksimacije, ne da bi bilo potrebno izpeljati verjetnostni model. Če so v matriki $\mathbf{Y}$ binarne razredne oznake, so namesto logističnega ali drugega verjetnostnega modela primernejše standardne ocenitvene funkcije za klasifikacijske probleme. Te vključujejo 0/1-predznačeno izgubo, to je ujemanje pozitivnih oznak s pozitivnimi elementi v matriki $\mathbf{X}$

$$\text{loss}(x; y) = \begin{cases} 0 & \text{če } xy > 0 \\ 1 & \text{sicer,} \end{cases} \tag{2.6}$$

in razne konveksne ocenitvene funkcije, kot je Hingeova ocenitvena funkcija

$$\text{loss}(x; y) = \begin{cases} 0 & \text{če } xy > 1 \\ 1 - xy & \text{sicer.} \end{cases} \tag{2.7}$$

2.1.5 Matrična faktorizacija z omejitvami

Do sedaj smo si ogledali le pristope matričnih faktorizacij *brez omejitev*, to so postopki, kjer sta matriki $\mathbf{U}$ in $\mathbf{V}$ poljubni realni matriki in je model $\mathbf{X} = \mathbf{U}\mathbf{V}^T$ omejen le z matričnim rangom. Pogosto je primerno omejiti prostor razcepnih matričnih faktorjev. To je potrebno zaradi

- skladnosti interpretacije matričnih faktorjev (n.pr. določanje verjetnostnih porazdelitev),
- zmanjševanja kompleksnosti iskanja modela, in
- določitve večih različnih naborov razcepnih faktorjev.

Omejitve razcepnih matričnih faktorjev običajno zmanjšajo število stopenj prostosti faktorizacije $\mathbf{U}\mathbf{V}^T$ v rekonstrukciji $\mathbf{Y}$ in olajšajo interpretacijo.

Lee in Seung sta raziskovala različne omejitve razcepnih faktorjev, med drugim nenegativnost [36] in stohastične omejitve [35]. Razvoj metod matričnih faktorizacij z omejitvami in razširitve na večrazcepne faktorizacijske

modele je aktivno raziskovalno področje. Naj omenimo le nekaj razširitev, ki smo jih uporabili v programski knjižnici Nimfa:

- Bayesovska nenegativna matrična faktorizacija z Gibbsovim vzorčenjem [55],
- binarna matrična faktorizacija [70],
- nenegativna faktorizacija z interativnimi pogojnimi načini [55],
- nenegativna faktorizacija za izločanje lokalnih značilk [37, 65],
- nenegativna faktorizacija z metodo alternirajočih najmanjših kvadratov in projeciranim gradientom [39],
- gladka matrična faktorizacija [48],
- nenegativna matrična faktorizacija z matriko povezanosti [10] in divergenco Kullback-Leibler [36],
- verjetnostna faktorizacija [34],
- verjetnostna redka faktorizacija [22],
- večrazcepna hkratna redka faktorizacija z mrežno regularizacijo [69],
- kazenska faktorizacija za gručenje z omejitvami [64].

2.2 Pristopi združevanja heterogenih podatkovnih virov v napovednih modelih

Izkoriščanje vseh dostopnih podatkovnih virov pri gradnji napovednih modelov postaja s hitro rastjo količine podatkov izredno pomemben izziv v podatkovni analizi [43]. Motivacija za integracijo podatkovnih virov je utemeljena na vseh področjih bioznanosti, še posebej v fiziki delcev, vesoljskih raziskavah in biologiji. V pričujoči nalogi nas predvsem zanima vključevanje heterogenih bioloških virov v učni proces gradnje napovednega modela.

Tehnike odkrivanja znanj iz podatkov (angl. *data mining*) nam lahko pomagajo pri tvorjenju novih hipotez, ki izhajajo iz bioloških podatkov. Pri

tem pogosto želimo uporabiti predznanje, ki je na področju bioznanosti mnogokrat podano v obliki ontologij.

Čeprav v biologiji obstajajo številni tipi podatkov, tabela 2.1 povzema le prevladujoče kategorije. Napovedni model, ki napoveduje na osnovi različnih virov, mora zagotoviti bolj natančne napovedi, kot jih je mogoče doseči z uporabo katerega koli posameznega vira podatkov, da je njegova uporaba smiselna. Vendar raznolika narava virov predstavlja izziv pri gradnji enotnega modela.

Tip	Primer	Opis
Gensko zaporedje	AACTAG	Kategorični tip, oznake za štiri nukleinske kisline
Proteinsko zaporedje	ARNDCEQ	Kategorični tip, oznake za dvajset aminokislin
Genske izrazi	(abc3, 9h) = 0.246	Številski tip, zvezne vrednosti o količini mRNA
Proteinske interakcije	(brca1, esr1, 0.999)	Številski tip, zvezne vrednosti o medsebojnem vplivu proteinov
Proteinske strukture	$(0.24, 4.12, 82.1)$ $\phi = 0.23$	Številski tip, koordinate v prostoru in eksperimentalni rezultati
Masna spektrometrija	123.45	Številski tip, zvezne vrednosti predstavljajo razmerje mase in električnega naboja v peptidu
Filogenetski podatki	Phy0010CMS–Phy000CVNF	Številski in kategorični tip, morfološki in molekularni podatki v rodoslovnih drevesih
Metabolični podatki	ko04110-Cell cycle[2]	Številski in kategorični tip, omrežja presnovnih poti

Tabela 2.1: Prevladujoči formati bioloških podatkov.

Različne metode za združevanje bioloških podatkov (angl. *biological data fusion*) lahko razdelimo v pet kategorij [46]:

[2]Primer znane podatkovne baze metaboličnih poti je baza KEGG. Dostopna je na naslovu `http://www.genome.jp/kegg/pathway.html`.

Vektorsko združevanje. Primer podatkovnega nabora je opisan z vektorjem značilk, izločenih iz različnih podatkovnih virov. Poljubna standardna napovedna metoda se nato uporabi nad atributnim zapisom učnega nabora. Gre za obliko zgodnjega združevanja.

Združevanje napovedi. Napovedni model je zgrajen za vsak podatkovni vir neodvisno, nato se združi njihove napovedi v končno napoved. Združevanje napovedi je primer poznega združevanja in temelji na principu večkratne razlage. Ta trdi, da moramo za optimalno rešitev problema upoštevati vse hipoteze, ki so skladne z vhodnimi podatki.

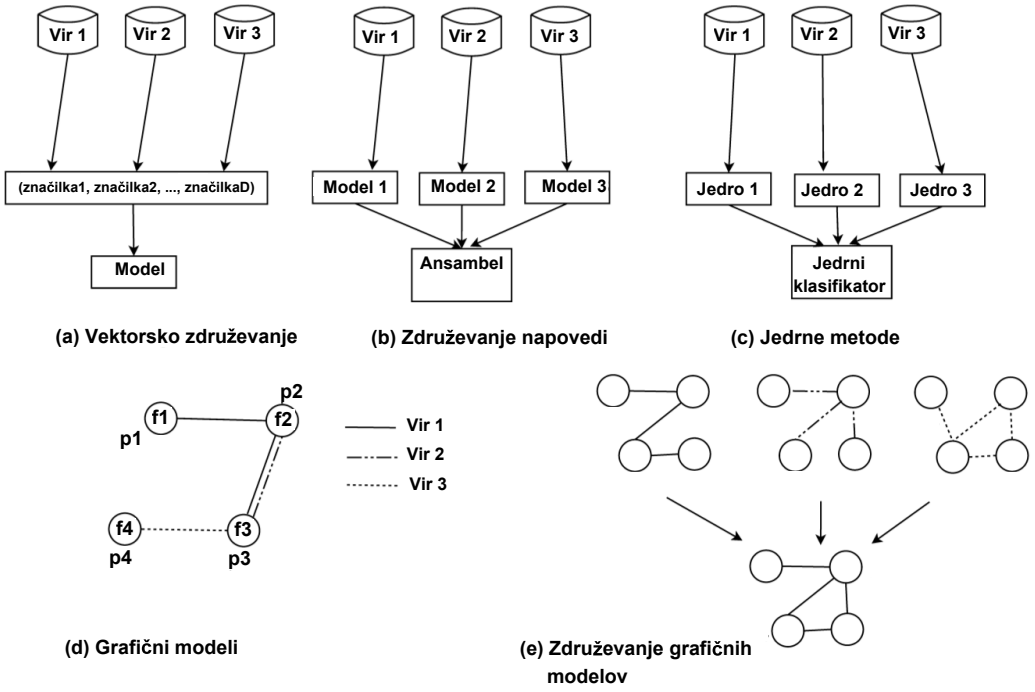
Jedrne metode. Te metode zagotavljajo jasen okvir za povezovanje različnih podatkovnih virov, ki se uporabljajo tudi, če podatki niso predstavljeni z vektorji. Več virov podatkov tvori mrežo, ki je informativna glede odnosov med elementi. Primer take mreže je omrežje proteinskih interakcij. Proteini, ki pogosto interagirajo, verjetno sodelujejo v isti (n.pr. presnovni ali genski) poti. Zato je omrežje proteinskih interakcij v celici informativno za napovedovanje proteinskih funkcij.

Grafični modeli. Usmerjeni in neusmerjeni grafični modeli nudijo verjetnostni okvir za združevanje heterogenih podatkovnih virov. Modeliranje se doseže s predstavitvijo lokalnih verjetnostnih odvisnosti. Struktura teh modelov je naravna izbira za zajemanje mreže funkcionalnih razmerij.

Združevanje grafičnih modelov. Večkrat je koristno povezovanje večih omrežij funkcionalnih odnosov, kot so različne oblike interakcij, skupnega izražanja in regulacije.

Shematski prikaz omenjenih pristopov združevanja heterogenih virov je prikazan na sliki 2.1. Pristopa z jedrnimi metodami in grafičnimi modeli ter vektorsko združevanje kombinirajo vhodne predstavitve podatkov iz različnih virov in uporabljajo zgodnjo strategijo združevanja. Združevanje napovedi, ki so pripravljene neodvisno za vsak vir, z ansambelskim pristopom sodi med

strategije pozne integracije. Grafični modeli so obetavno področje za razvoj novih pristopov vmesne integracije. Guo in sod. (2010) [26] so predlagali napovedno metodo za združevanje Gaussovih grafičnih modelov iz heterogenih podatkovnih virov. Uporaba enega grafičnega modela nad heterogenimi viri zasenči podatke, ki so prisotni le v nekaterih predstavitvah, zato pristop Guoja in sod. upošteva več modelov hkrati z uvedbo hierarhičnih kazenskih členov. Ti kazenski členi usmerjajo učenje k ohranjanju opisov, ki so skupni vsem predstavitvam podatkov, a hkrati dopuščajo, da del opisa podatkov izhaja le iz neke predstavitve.



Slika 2.1: Shematski prikaz pogostih pristopov združevanja heterogenih bioloških podatkovnih virov.

2.2.1 Vektorsko združevanje

Morda najpreprostejša oblika povezovanja podatkov je povzemanje ustreznih značilk različnih podatkovnih vrst v vektorje fiksne dolžine in uporaba klasi-

fikacijske ali regresijske metode. To je primer zgodnje strategije združevanja. Ta pristop je preprost, vendar enaka obravnava različnih podatkovnih vrst ne dovoljuje vključevanja domenskega znanja v gradnjo modela.

Zgodnji primer takšnega pristopa pri napovedovanju proteinskih funkcij je opisal Jardins s sod. (1997) [17]. To delo predstavlja omejeno obliko združevanja podatkov; uporablja namreč veliko različnih vrst proteinskih lastnosti, vendar je večina značilk izpeljanih iz proteinskih zaporedij. Te vključujejo dolžino proteinskih zaporedij, molekulsko maso, naboj, aminokislinsko sestavo (t.j. frekvence ostankov) in izoelektrično točko. Za učne primere z dostopno informacijo o tri-dimenzionalni strukturi je bilo vključenih več značilk o sekundarni strukturi beljakovin. Avtorji uporabljajo dobro znane algoritme strojnega učenja, kot so odločitvena drevesa, metoda najbližjih sosedov in naivni Bayesov klasifikator, ter primerjajo njihovo uspešnost z algoritmom BLAST za poravnavo aminokislinskih zaporedij.

Podobno so Jensen in sod. (2002) [30] z značilkami, izpeljanimi iz aminokislinskih zaporedij (n.pr. povprečna hidrofobičnost in število negativno nabitih ostankov) gradili nevronske mreže za napovedovanje različnih vrst sprememb v prevajanju zaporedij in regij spremenjene kompleksnosti.

Beljakovinske strukture se pogosto ohranjajo tudi, ko ni mogoče zaznati značilnih ohranitev v aminokislinskih zaporedjih. To pomeni, da je včasih primerno namesto neposredne primerjave zaporedij sestaviti vektorje značilk na osnovi sekundarne in terciarne strukture beljakovin. Dobson in Doig (2005) [21] sta uporabila slednje z metodo podpornih vektorjev za napovedovanje oznak EC encimov z znano strukturo.

Yan in sod. (2010) [68] so z naključnimi gozdovi gradili modele za napovedovanje genskih določitev mušice *D. melanogaster*. Značilke so izračunali iz zapisov genskih izrazov, ohranjenih proteinskih domen, proteinskih interakcij, podobnosti genskih zaporedij in fizikalnih lastnosti. Oznake razredov so predstavljale kategorije konceptov iz ontologije GO Ontology[3] [4].

Eden izmed izzivov v vektorskem združevanju je določitev prispevka zna-

[3]Genska ontologija je dostopna na `http://www.geneontology.org`.

čilke k točnosti napovednega modela in iskanje majhnih podmnožic značilk, ki ohranjajo ali izboljšajo točnost modela. Ta naloga je znana kot *izbor značilk* (angl. *feature selection*) in je aktivno področje raziskav v strojnem učenju. Bralec bo več o metodah izbora značilk našel v [27]. Najenostavnejši pristop izbiranja značilk je metoda filtriranja, ki za vsako značilko izračuna statistiko, ki odraža njeno informativnost. Primeri tovrstnih statistik so Pearsonov korelacijski koeficient, površino pod krivuljo ROC, Fisherjevo kriterijsko funkcijo in druge.

Neodvisno ocenjevanje vsake značilke ne upošteva redundance, ki je značilna za visoko razsežne podatke o genskih izrazih, in ne upošteva lastnosti napovednega modela. Te ovire odpravljamo z *metodami ovojnic* (angl. *wrapper methods*) in vgrajenimi pristopi (angl. *embedded methods*). Metode ovijanja z napovedno metodo ocenijo prispevek množice značilk in se lahko kombinirajo s skoraj vsemi napovednimi metodami. V vgrajeni tehniki je napovedni model del algoritma za izbor značilk in slednji upošteva lastnosti modela pri izboru informativnih značilk. Primer preproste vgrajene metode je rekurzivno odstranjevanje značilk (angl. *recursive feature elimination*), ki v linearnih klasifikacijskih modelih iterativno odstranjuje značilke z najmanjšo pripadajočo utežjo klasifikacijskega vektorja.

Očitna slabost vektorskega združevanja je modeliranje vseh značilk na enak način. Eden izmed načinov za odpravljanje te slabosti je učenje različnih modelov za vsak podatkovni vir in nato kombiniranje napovedi – združevanje s pozno strategijo. Tak pristop imenujemo združevanje napovedi in je opisan v poglavju 2.2.2. Jedrne metode, predstavljene v 2.2.3, gradijo le en model, a ta omogoča večjo fleksibilnost pri združevanju podatkovnih virov z določitvijo mer podobnosti za vsak podatkovni vir. Poleg tega se jedrne metode uporabljajo za modeliranje virov, kot so proteinska zaporedja, kjer predstavitev učnih primerov z vektorji ni enostavna.

2.2.2 Združevanje napovedi

Drugi pristop k izgradnji enotnega napovednega modela sestavi več modelov in nato združuje njihove napovedi. Najpogosteje so napovedni modeli dobljeni z algoritmi razvrščanja. Področje kombiniranja klasifikatorjev in metaučenja je v raziskavah strojnega učenja deležno veliko pozornosti, saj so združene napovedi pogosto bolj točne od posameznih napovedi.

Iskanje genov v zaporedjih DNA je znan bioinformatični problem, kjer združevanje večih klasifikacijskih metod zagotavlja natančnejše napovedi [52]. Sarac in sod. (2010) [54] so s pozno strategijo kombiniranja napovedi klasifikatorjev BLAST-k-najbližjh sosedov, podpornimi vektorji in vložitvijo proteinskih zaporedij v nizko razsežni prostor z algoritmom SPMap izboljšali točnost napovedovanja molekularnih funkcij, izpeljanih iz genske ontologije GO Ontology.

Metode za kombiniranje lahko razdelimo v več skupin:

(i) združevanje modelov različnih metod, zgrajenih na enakih podatkovnih naborih;

(ii) kombiniranje večih modelov ene metode, zgrajenih na podatkovnih podmnožicah ali s podmnožicami značilk – *ansambel klasifikatorjev*;

(iii) združevanje večih modelov, ki so zgrajeni iz različnih podatkovnih virov.

Na področju gradnje ansamblov je bilo največ raziskav opravljenih z uporabo enega samega učnega algoritma. Različnost modelov dosežemo z vključitvijo naključnosti v učni algoritem (n.pr. naključni gozdovi), spreminjanjem sestave učne množice (n.pr. metodi boosting in bagging) ali s spreminjanjem značilk množice. Napovedi nato združujemo z navadnim ali uteženim glasovanjem.

V primeru več različnih algoritmov na eni sami učni množici lahko uporabimo metodo skladanja klasifikatorjev. S tehnikami metaučenja točnost napovedi še izboljšamo tako, da se učimo izbirati: (i) primerno pristranskost

učnih algoritmov, (ii) ustrezni učni algoritem in (iii) kombiniramo osnovne napovedi klasifikatorjev.

2.2.3 Jedrne metode

Jedrne metode so dobro znana tehnika v strojnem učenju in ta priljubljen pristop je prisoten tudi v računski biologiji [8, 7].

Jedro je preslikava, ki določa podobnosti med pari objektov. Jedrna metoda je algoritem, ki dostopa do podatkov le preko jedra, definiranega nad podatki. Natančneje, jedro je mera podobnosti, ki zadošča zahtevam skalarnega produkta v nekem vektorskem prostoru – jedro $K(x, y)$ se izraža kot $\langle \phi(x), \phi(y) \rangle$, kjer je ϕ nelinearna preslikava. Ta tehnika je znana že desetletja, a je pridobila na veljavi v zvezi s posebno močnim algoritmom razvrščanja, znanim kot metoda podpornih vektorjev (SVM) [56]. Znani trik z jedri – preslikava podatkov v več razsežni prostor z vnaprej določeno jedrno funkcijo – pogosto pretvori osnovni problem v nalogo z več dimenzijami kot primeri. SVM se izjemno dobro spopada s takimi primeri in je učinkovit pri zmanjšanju prekletstva dimenzionalnosti (angl. *curse of dimensionality*). Jedrne metode so uporabne za klasifikacijske in regresijske naloge, uvrščanje, analizo glavnih komponent in drugo.

Jedrne metode zagotavljajo jasen okvir za združevanje podatkovnih virov, saj jedrna funkcija zagotavlja obliko, v kateri se predstavlja veliko različnih vrst podatkov, vključno z vektorji, matrikami, nizi, drevesi in grafi. Praviloma metoda dostopa do podatkov preko jedrne funkcije, zato je mogoče nabor n elementov povzeti kot *jedrno matriko* velikosti $n \times n$. To lastnost izkoriščamo tudi v predlaganem matričnem pristopu združevanja heterogenih podatkovnih virov. Jedrna matrika je zadostna predstavitev. Ko se izračuna, se izvorne podatke lahko zavrže in jedrna metoda še vedno opravlja svojo nalogo. Poleg tega jedrne matrike iz različnih virov podatkov lahko povezujemo v preprosto jedrno algebro, ki definira operacije seštevanja, množenja in konvolucije [28].

Najenostavnejši način združevanja jeder je tvorjenje aditivnih kombina-

cij – seštevanje slik jedrnih funkcij je enakovredno konkatenaciji vektorskih predstavitev. Prostor značilk definiran z množenjem jeder je produkt prostora značilk posamenih jeder. Ta pristop je bil uporabljen pri napovedovanju proteinskih interakcij [7].

Možno je opravljati vektorsko združevanje z jedrnimi metodami. Koristno je normalizirati vrednosti značilk, da zavzemajo podoben obseg, saj so jedrne metode občutljive na različno skalirane vrednosti. Alternativno lahko normaliziramo jedrno funkcijo namesto predstavitve podatkov tako, da uporabimo jedro $K'(x, y) = K(x, y)/\sqrt{K(x, x)K(y, y)}$, kar ustreza projeciranju predstavitve podatkov na enotsko sfero.

Pavlidis in sod. (2001) [50] so uporabili jedrno metodo združevanja za napovedovanje proteinskih funkcij. Avtorji so kombinirali genske izrazne in filogenetske zapise v eno jedro in z metodo SVM uvrščali gene kvasovke v funkcijske kategorije. V primerjavi z navadnim vektorskim združevanjem, ki zgolj konkatenira dve vektorski predstavitvi podatkov, in združevanjem napovedi iz dveh ločeno zgrajenih modelov SVM, je ta pristop dosegal višjo točnost. Ključna razlika med tem pristopom in vektorskim združevanjem je uporaba polinomskega jedra tretje stopnje na obeh podatkovnih množicah pred združitvijo. Polinomsko jedro preslika podatke v večrazsežni prostor, v katerem so značilke monomi prvotnih značilk s stopnjo manjšo ali enako tri. Preslikava vsake množice podatkov posebej namesto konkateniranih vektorskih predstavitev vključuje predhodno znanje – odvisnosti med značilkami enega tipa so bolj verjetne kot odvisnosti med značilkami različnih tipov.

Namesto enostavnega seštevanja vrednosti jedrnih funkcij lahko sestavimo linearno kombinacijo jeder, ki upošteva informativost preslikav. Če vemo, da je nabor A koristnejši (n.pr. bolj informativen in z manj šuma) od nabora B, lahko pripadajoči preslikavi kombiniramo $K_{AB} = \lambda K_A + K_B$. Utežni faktor λ pri združevanju dveh jeder lahko določimo z notranjim prečnim preverjanjem.

Lanckriet in sod. (2004) [33] so predstavili statistično ogrodje za združevanje podatkov z jedri, pri čemer so vsak podatkovni vir utežili z utežnimi faktorji. Namesto zahteve po vnaprejšnji določitvi uteži so avtorji uporabili

metodo SVM za hkratno učenje utežnih faktorjev s semidefinitnim programom (SDP). Ta pristop so primerjali z združevanjem z Markovskimi naključnimi polji. Avtorji so razvrščali gene kvasovke v trinajst širših funkcijskih skupin baze MIPS FunCat[4] in uporabljali pet podatkovnih virov: (i) struktura domene proteinov, (ii) znane proteinske interakcije, (iii) genske interakcije, (iv) proteinske ko-komplekse, in (v) genske izraze. Zmogljivost modela so ocenjevali s krivuljo ROC. Pristop z metodo podpornih vektorjev in semidefinitnim programiranjem je bil boljši od vektorskega združevanja in Markovskih polj.

Zanimiv pristop k napovedovanju proteinskih funkcij so predlagali Borgwardt in sod. (2005) [8]. Strukturo beljakovin so predstavili z grafom, čigar vozlišča so sekundarni strukturni elementi in povezave predstavljajo bližino aminokislinskih zaporedij in prostorske strukture proteinov. Avtorji so podobnost proteinov opredelili z jedrom z naključnim sprehodom. Predlagano jedro je združevalo lokalne lastnosti proteinov in globalno prostorsko strukturo proteinov.

2.2.4 Grafični modeli

Veliko podatkov s področja bioinformatike je podanih v obliki mreže ali jih je mogoče pretvoriti v omrežno strukturo. Primer takšnega omrežja so proteinske interakcije – beljakovine, ki interagirajo, so pogosto udeležene v istem biološkem procesu, imajo podoben vzorec v lokalizaciji in v manjši meri tudi podobno funkcijo. Druge vire podatkov, ki niso podani v neposredni obliki omrežij, lahko vanje pretvorimo. Genske izrazne zapise predstavimo z grafom, čigar vozlišča so geni, povezave pa vzpostavljene le, če so si genski izrazni zapisi med seboj povezanih genov dovolj podobni. V primeru zaporedij DNA uteži povezav zavisijo od podobnosti med zaporedji nukleinskih kislin – podobnost lahko izračunamo z algoritmi za poravnavo, kot sta Smith-Waterman

[4]Podatkovna baza MIPS FunCat s funkcijskimi oznakami za sistematično razvrščanje proteinov je dostopna na `http://mips.helmholtz-muenchen.de/proj/funcatDB/search_main_frame.html`

ali algoritem PSI-BLAST.

Pomembna naloga je poenotenje številnih danih omrežij med pari podatkovnih tipov v enotno omrežje. Opišimo enostaven pristop za združevanje treh virov: (i) korelirana evolucija iz filogenetskih zapisov, (ii) izraze mRNA, in (iii) vzorci bioloških domen. Napovedi, ki jih podpira večina zgrajenih modelov, so razumljene kot zanesljive, zato zanesljive napovedi razširimo skozi omrežje, da označimo nove primere podatkov. Zgolj upoštevanje napovedi, ki so enake v več modelih, ima pomanjkljivosti, posebno, če se napovedi modelov razlikujejo v zanesljivosti. Združevanje je uspešnejše, če upoštevamo ocene zanesljivosti napovedi iz različnih omrežij. Povezavi L med proteinoma v omrežju E in kontekstu konkretne poti ali določitve priredimo verjetje LLS (angl. *log-likelihood score*)

$$\mathrm{LLS}(L|E) = \log \frac{P(L|E)/P(\widetilde{L}|E)}{P(L)/P(\widetilde{L})}, \tag{2.8}$$

kjer $P(L|E)$ označuje pogostost povezave L v podatkih in $\widetilde{L}$ primere, v katerih povezava ni realizirana.

Podoben problem naslavljata sistema MAGIC [62] in bioPixie [45] za združevanje podatkovnih virov pri napovedovanju genskih funkcij in odkrivanju poti v kvasovki *S. cerevisiae*. V sistemu MAGIC se ocenjuje verjetnost, da sta proteina i in j funkcijsko povezana. Obstoj povezanosti se modelira z več relacijami med proteini, kot so skupna izražava, fizikalne interakcije, genske interakcije in sodelovanje v kompleksu. Nato se uporabi Bayesovska mreža za ocenjevanje verjetnosti. Bayesovska mreža je verjetnostni model, ki predstavi verjetnostne odvisnosti v podatkih v obliki usmerjenega grafa tako, da je mogoč učinkovit izračun verjetnostnih porazdelitev.

Za ponazoritev pristopa si oglejmo preprost model. Naj bo R slučajna spremenljivka, ki označuje obstoj funkcijske povezanosti proteinov in X_1, $X_2, \ldots, X_d$ slučajne spremenljivke, ki pomenijo znane relacije med proteini, to je, dane podatke. Zanima nas verjetnost $P(R|X_1, X_2, \ldots, X_d)$. Verjetnost izrazimo z Bayesovim pravilom

$$P(R|X_1, X_2, \ldots, X_d) = \frac{P(X_1, X_2, \ldots, X_d|R)P(R)}{P(X_1, X_2, \ldots, X_d)}. \tag{2.9}$$

Naivni Bayesov model predpostavlja, da je vsak vir pogojno neodvisen od drugih virov, torej $P(X_i|R, X_j) = P(X_i|R)$ za $i \neq j$. Zato lahko izraz (2.9) pišemo

$$\begin{aligned} P(X_1, X_1, \ldots, X_d|R) &= P(X_1|R)P(X_2, \ldots, X_d|R, X_1) \\ &= P(X_1|R)P(X_2, \ldots, X_d|R) \\ &= P(X_1|R)P(X_2|R)P(X_3, \ldots, X_d|R) \\ \ldots &= \prod_{i=1}^{d} P(X_i|R). \end{aligned} \tag{2.10}$$

Končno verjetnost ocenimo z

$$P(R|X_1, X_2, \ldots, X_d) = \frac{\prod_i P(X_i|R)P(R)}{\prod_i P(X_i)}. \tag{2.11}$$

Ta preprost grafični model interpretiramo tako, da so ob dani funkcijski povezanosti proteinov druge relacije, kot so proteinske interakcije in skupni izrazi, neodvisne in so posledica funkcijske povezanosti. Bayesovsko omrežje, ki so ga predlagali Troyanskaya in sod. (2003, 2005) [62, 45] uvaja nekatere odvisnosti med različnimi viri podatkov, a splošna struktura odvisnosti je podobna zgoraj predstavljeni. Avtorji so pogojne verjetnostne porazdelitve v vozliščih določili s strokovnjaki iz področja molekularne biologije kvasovke in teh niso avtomatsko izločili iz podatkov.

Učenje presnovnih poti je primer učenja funkcijske povezanosti. Pri tem se učimo omrežij, katerih vozlišča so encimi in povezave označujejo, da dana encima sodelujeta pri katalizaciji sosednih reakcij na presnovni poti. Yamanishi in sod. (2005) [67] so predlagali dva pristopa za učenje presnovih poti. Prvi pristop je direkten, tako da se zgradi klasifikator, ki čim bolje loči med pozitivnimi primeri – znani pari encimov iz iste presnovne poti – in

negativnimi primeri. V drugem pristopu so Yamanishi in sod. pred uporabo klasifikacijskega algoritma izračunali nizko razsežno predstavitev podatkov. Ta je izpostavila bližino med encimi na isti presnovni poti. Njihovi rezultati so pokazali, da zmanjšanje dimenzij vhodne predstavitve pred razvrščanjem značilno izboljša točnost napovedi glede na pristop brez redukcije dimenzij.

2.2.5 Kontekst v grafičnih modelih in združevanje

Ko v omrežje predstavimo znana razmerja med pari vozlišč, lahko neznane oznake izpeljemo s pravilom "krivde zaradi povezanosti" (angl. *guilty by association*) tako, da obravnavamo soseščine znanih vozlišč – v preprostem pristopu le z večinskim glasovanjem med oznakami sosednjih vozlišč. Ta način delovanja je poenostavitev problema, saj ne upošteva širšega konteksta, v katerem se pojavlja vozlišče. Alternativni pristop zgolj zanašanja na lokalne informacije je povezovanje podatkov preko omrežij. GeneMANIA[5] je orodje, ki išče gene, povezane z nizom vhodnih genov in uporablja zelo veliko naborov s podatki o funkcijski povezanosti genov [66]. Ti vključujejo proteinske in genske interakcije, znane genske poti in skupne genske izraze, fizikalne interakcije in skupne lokalizacije. Orodje je bilo nedavno razširjeno z atributnimi omrežji, ki omogočajo enostavno upoštevanje informacij iz atributnih predstavitev.

V tem razdelku opišemo nekaj pristopov, ki upoštevajo celotno omrežje pri pripravi napovedi. Vsi napovedujejo razredne oznake elementov (n.pr. izbrano funkcijo proteinov) z uporabo omrežij, v katerih je vsak element (n.pr. protein) vozlišče. Oznaka vozlišča lahko zavzame vrednosti nič in ena glede na pripadnost razredu ter posebno vrednost, če pripadnost ni znana. Povezavam so dodane uteži o povezanosti med vozlišči. Možna je uporaba več omrežij z njihovim združevanjem ali deljenjem naključnih spremenljivk oznak med omrežji.

Mnogi podatkovni viri izražajo simetričnost – proteinske interakcije in skupni genski izrazi – in jim ne moremo določiti usmerjenosti. Bayesovske

[5]Orodje GeneMANIA je dostopno na strani `http://genemania.org`.

mreže, ki se zanašajo na usmerjene grafe za modeliranje odvisnosti med spremenljivkami, za te probleme niso primerne. Uporabimo Markovska polja, grafične modele, ki predstavijo verjetnostne odvisnosti z neusmerjenimi grafi. Deng in sod. (2003) [16] so predlagali model Markovskih polj za obravnavo večih omrežij, pri čemer vsako omrežje neodvisno modelira en razred (n.pr. proteinsko funkcijo). Gre za pristop poznega združevanja z Markovskimi polji. Vozlišču i v omrežju razreda c je dodeljena binarna slučajna spremenljivka X_i, ki meri pripadnost vozlišča i razredu c. Realizacijo slučajne spremenljivke X_i označimo z x_i. Realizacijo slučajnega vektorja $\mathbf{X} = (X_1, X_2, \ldots, X_N)$ označimo z $\mathbf{x}$ in njena verjetnost je v modelu Deng in sod. določena z

$$P(\mathbf{x}) = e^{-U(\mathbf{x})} \frac{1}{Z(\theta)}, \tag{2.12}$$

pri čemer je $Z(\theta)$ normalizacijski faktor, ki zavisi od parametrov modela. Vrednost $U(\mathbf{x})$ se izračuna kot

$$\begin{aligned} U(\mathbf{x}) &= -\alpha \sum_{i=1}^{N} x_i - \beta \sum_{(i,j)\in S} [(1-x_i)x_j + x_i(1-x_j)] \\ &\quad -\gamma \sum_{(i,j)\in S} x_i x_j - \kappa \sum_{(i,j)\in S} (1-x_i)(1-x_j), \end{aligned} \tag{2.13}$$

kjer S pomeni množico vseh povezav v grafu, $\alpha = \log \frac{\pi}{1-\pi}$ in π je apriorna verjetnost razreda c. Prvi člen v (2.13) predstavlja apriorno verjetnost konfiguracije $\mathbf{x}$. Ostali členi zajamejo interakcije med sosedi v omrežju – po vrsti, število sosedov z različnimi oznakami razreda; število sosedov, ki pripadajo razredu c; število sosedov, ki so negativni primeri.

Deng in sod. so s predstavljenim modelom Markovskega polja napovedali proteinske funkcije. Učni nabor so tvorili proteini, katerih funkcije so znane. Verjetnosti genskih funkcij testnih proteinov so ocenili s pogojevanjem na stanja učnih proteinov. Verjetnost, da ima testni protein dano funkcijo, je enaka vsoti preko vseh konfiguracij drugih testnih proteinov. To verjetnost so avtorji ocenili z Gibbsovim vzorčenjem.

Do sedaj smo predstavili Markovsko polje, ki uporablja le eno omrežje. Če je dostopnih več omrežij, je verjetnost enaka produktu členov oblike (2.13) z deljenjem vektorja $\mathbf{X}$. Deng in sod. so temu dodali še komponento, ki upošteva sestavo domene proteinov. Posplošitev predstavljenega pristopa je iterativna uporaba pravila "krivde zaradi povezanosti", dokler mreža ne doseže stanja, ki je maksimalno skladno z opazovanimi podatki.

Namesto da poskušamo oceniti verjetnostni model omrežja, se lahko poslužimo neposredne gradnje napovedi. Tak pristop so opisali Tsuda in sod. (2005) [63]. Avtorji sledijo znani paradigmi v strojnem učenju in optimizirajo dvočleno kriterijsko funkcijo. Ta sestoji iz člena napake, ki meri, kako dobro se napovedi ujemajo z opazovanimi podatki, in člena za regularizacijo. Z ozirom učenja iz omrežij regularizacija pomeni, da sosednjim vozliščem ustrezajo podobne napovedi. Tsuda in sod. so definirali graf podobno kot Deng in sod., vozlišča imajo binarne oznake glede na pripadnost razredu. Naj bo n število vozlišč in naj bodo dane oznake za prvih p vozlišč v binarnem vektorju $\mathbf{y}$ z elementi ± 1. Vozlišča z neznano oznako imajo vrednosti $y_i = 0$. Nadaljnje z f_i označimo napoved vozlišča i. Vektor napovedanih oznak $\mathbf{f} = (f_1, f_2, \ldots, f_n)$ se določi z minimizacijo funkcije

$$\sum_{i=1}^{p}(f_i - y_i)^2 + \mu \sum_{p+1}^{n} f_i^2 + \sum_{i,j} w_{ij}(f_i - f_j)^2, \tag{2.14}$$

pri čemer je w_{ij} utež povezave med vozliščema i in j. Prvi člen v (2.14) meri napako, druga sta regularizacijska člena – po vrsti, omejitev vrednosti f_i za neoznačena vozlišča in sosednja vozlišča imajo podobno napoved. Če postavimo $\mu = 1$, se (2.14) poenostavi v

$$\sum_{i=1}^{n}(f_i - y_i)^2 + \sum_{i,j} w_{ij}(f_i - f_j)^2, \tag{2.15}$$

saj je zadnjih $n - p$ vrednosti v vektorju $\mathbf{y}$ ničelnih. Ekvivalenten optimizacijski problem predstavimo z opisom

$$\min_{\mathbf{f},\gamma} \sum_{i=1}^{n} (f_i - y_i)^2 + r\gamma, \quad \mathbf{f}^T \mathbf{L} \mathbf{f} \leq \gamma, \tag{2.16}$$

pri čemer je $\mathbf{L}$ Laplaceova matrika, podana z $\mathbf{L} = \mathbf{D} - \mathbf{W}$, $\mathbf{W}$ matrika uteži in $\mathbf{D} = \mathrm{diag}(d_i)$, $d_i = \sum_j w_{ij}$.

Opis iz (2.16) uporabimo za združevanje večih omrežij na način

$$\min_{\mathbf{f},\gamma} \sum_{i=1}^{n} (f_i - y_i)^2 + r\gamma, \quad \mathbf{f}^T \mathbf{L}_k \mathbf{f} \leq \gamma, \tag{2.17}$$

kjer je $\mathbf{L}_k$ Laplaceova matrika za omrežje k. Dualna predstavitev te naloge daje učinkovit algoritem za reševanje, ki je uporaben tudi za velika omrežja. Tsuda in sod. so z opisanim pristopom napovedali funkcijske skupine proteinov kvasovke *S. cerevisiae* izvedene iz baze MIPS FunCat. Zgradili so omrežja iz proteinskih kompleksov, genskih interakcij, fizikalnih interakcij MIPS in podobnosti v sestavi domen iz baze Pfam[6]. Uspešnost pristopa je primerljiva z metodo podpornih vektorjev in semidefinitnim programiranjem, opisanem v razdelku 2.2.3.

2.3 Delno nadzorovano učenje z matrično faktorizacijo

V zadnjih letih narašča zanimanje za delno nadzorovane metode učenja, katerih cij je razvrščanje podatkov, upoštevajoč dodatne informacije, ki usmerjajo učenje. Nadzorne informacije so običajno podane z omejitvami, ki odražajo podobnost med pari elementov. Poleg pristopov matrične faktorizacije k delno nadzorovanim metodam so uspešne tudi tehnike, osnovane na teoriji grafov, kot sta bipartitna spektralna particija grafov [19] in relacijska povzemalna omrežja [42]. Kljub uspešnim empiričnim rezultatom in strogi teoretični analizi ti algoritmi izkoriščajo le informacije o odnosih med različnimi

[6]Podatkovna baza Pfam hrani oznake proteinskih družin in poravnave njihovih zaporedij. Dostopna je na `http://pfam.sanger.ac.uk`.

podatkovnimi tipi, najpogosteje le med dvema podatkovnima tipoma. Toda pri reševanju številnih problemov imamo na voljo ne le znanje o relacijah med različnimi podatkovnimi tipi, temveč tudi o povezanosti elementov znotraj enega podatkovnega tipa.

V tem delu predlagamo nov pristop za gradnjo napovednih modelov z matrično faktorizacijo iz heterogenih podatkovnih virov, s katerim naslavljamo obe omenjeni slabosti. Delno nadzorovano učenje z matrično faktorizacijo [64] nam služi kot osnova za izpeljavo tega pristopa, zato si zasluži posebno pozornost. V tabeli 2.2 povzemamo najpomembnejše uporabljene oznake. V razdelku 2.3.1 in 2.3.2 formalno uvedemo kazensko matrično faktorizacijo.

Oznaka	Opis
n	št. primerov v podatkovnem viru (n_i v primeru več virov)
c	faktorizacijski rang (c_i v primeru več virov)
$\mathbf{x}_i$	i-ti primer (stolpec), $x_i \in \mathbb{R}^d$
$\mathbf{X}$	podatkovna matrika velikosti $d \times n$
$\mathbf{f}_i$	i-ti faktor (stolpec) matrike $\mathbf{F}$
$\mathbf{F}$	razcepni matrični faktor velikosti $d \times c$
$\mathbf{G}$	razcepni matrični faktor velikosti $n \times c$
$\mathbf{\Theta}$	omejitvena matrika velikosti $n \times n$
$\mathbf{\Theta}_i$	omejitvena matrika za vir i velikosti $n_i \times n_i$
$\mathbf{R}$	bločna matrika relacij velikosti $\sum_i n_i \times \sum_i n_i$
$\mathbf{R}_{ij}$	matrika relacij med virom i in j

Tabela 2.2: Uporabljene matematične oznake in njihov pomen.

2.3.1 Kazenska matrična faktorizacija

Naj bo dana matrika $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$. Iščemo matrična razcepna faktorja $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_c] \in \mathbb{R}^{d \times c}$ in $\mathbf{G} \in \mathbb{R}^{n \times c}$, ki dobro aproksimirata matriko $\mathbf{X}$. Taki matriki torej rešujeta minimizacijsko nalogo s kriterijsko funkcijo

$$J = ||\mathbf{X} - \mathbf{F}\mathbf{G}^T||^2_{\text{Fro}}. \tag{2.18}$$

Če interpretiramo razcepni faktor $\mathbf{G}$ kot binarno matriko pripadnosti elementov c skupinam in stolpce v razcepnem faktorju $\mathbf{F}$ kot predstavnike skupin, je ta optimizacijska naloga ekvivalentna razvrščanju s c voditelji. Z uvedbo relaksacij v matriki $\mathbf{G}$ – matrika $\mathbf{G}$ vsebuje realne vrednosti z intervala $[0, 1]$ – lahko elemente razvrstimo v c skupin $\pi = \{\pi_1, \pi_2, \ldots, \pi_c\}$ z matrično faktorizacijo. Za določanje pripadnosti elementa skupini π_i z danima faktorjema $\mathbf{F}$ in $\mathbf{G}$ obstaja mnogo pristopov.

Pogosto znanje o pripadnosti elementov nekemu razredu predstavimo z množico omejitev med pari primerov $\mathcal{M}$ (angl. *must-link constraints*). Množica $\mathcal{M}$ vsebuje kazni $\theta_{ij} \geq 0$, ki se dodelijo, če primera $\mathbf{x}_i$ in $\mathbf{x}_j$ nimata podobnih zapisov v matričnem razcepnem faktorju. Znanje o ne-pripadnosti opišemo z množico omejitev med pari primerov $\mathcal{C}$ (angl. *cannot-link constraints*), omejitve $\tilde{\theta}_{ij} \geq 0$ izražajo kazni, če $\mathbf{x}_i$ in $\mathbf{x}_j$ pripadata podobna zapisa v razcepu. Podobnost zapisov pomeni, da sta zapisa uvrščena v isto skupino v razvrstitvi π.

Kriterijska funkcija kazenske matrične faktorizacije (2.19) vsebuje tri člene. Prvi člen zahteva, da je predstavniški vektor $\mathbf{f}_c$ skupine c v prostoru $\mathbf{X}$ blizu primerom v skupini c. Kazen, če primera iz $\mathcal{M}$ nimata podobnih zapisov, se zamenja z nagrado, če imata primera podoben zapis (od tod sprememba predznaka v izrazu (2.20)). Tretji člen kaznuje podobne zapise primerov iz $\mathcal{C}$. Kriterijsko funkcijo zapišemo

$$\begin{aligned} J(\pi) &= \sum_c \sum_{\mathbf{x}_i \in \pi_c} ||\mathbf{x}_i - \mathbf{f}_c||^2 - \sum_{\substack{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{M} \\ \pi(\mathbf{x}_i) = \pi(\mathbf{x}_j)}} \theta_{ij} + \sum_{\substack{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{C} \\ \pi(\mathbf{x}_i) = \pi(\mathbf{x}_j)}} \tilde{\theta}_{ij} \\ &= \sum_c \sum_{\mathbf{x}_i} g_{ic} ||\mathbf{x}_i - \mathbf{f}_c||^2 + \sum_c \sum_{i,j} g_{ic} g_{jc} \mathbf{\Theta}_{ij}, \end{aligned} \tag{2.19}$$

kjer g_{ij} določajo pripadnosti

$$g_{ij} = \begin{cases} 1 & \text{če } \mathbf{x}_i \in \pi_j \\ 0 & \text{sicer,} \end{cases}$$

in *omejitvena matrika* $\mathbf{\Theta}$ določa omejitve med vektorji v $\mathbf{X}$

$$\mathbf{\Theta}_{ij} = \begin{cases} \tilde{\theta}_{ij} & \text{če } (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{C} \\ -\theta_{ij} & \text{če } (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{M} \\ 0 & \text{sicer.} \end{cases} \tag{2.20}$$

Omejitvena matrika hrani naše apriorno znanje o podatkovnem viru.

Pogosto želimo za učinkovito računanje kriterijsko funkcijo zapisati v matrični obliki. Zapis (2.19) prepišemo v obliko $J(\pi) = ||\mathbf{X} - \mathbf{F}\mathbf{G}^T||^2_{\text{Fro}} + tr(\mathbf{G}^T\mathbf{\Theta}\mathbf{G})$. Zopet opustimo zahtevo o binarni vsebini razcepnega faktorja $\mathbf{G}$, saj ni znan učinkovit polinomski algoritem za optimizacijsko nalogo s to zahtevo. Sedaj lahko definiramo optimizacijsko nalogo, ki jo rešuje kazenska matrična faktorizacija (PMF)

$$\begin{aligned} \min_{\mathbf{F},\mathbf{G}} \quad & J(\pi) = ||\mathbf{X} - \mathbf{F}\mathbf{G}^T||^2_{\text{Fro}} + tr(\mathbf{G}^T\mathbf{\Theta}\mathbf{G}) \\ \text{p.p.} \quad & \mathbf{G} \geq 0. \end{aligned} \tag{2.21}$$

Wang in sod. (2008) [64] so za reševanje optimizacijske naloge (2.21) predlagali preprost iterativni algoritem, ki ga povzema algoritem 1. Dokaz pravilnosti in konvergence algoritma podajamo v dodatku A.

Vhod: podatkovna matrika $\mathbf{X}$, omejitvena matrika $\mathbf{\Theta}$, rang c.
Izhod: razcepna faktorja $\mathbf{F}$ in $\mathbf{G}$.
Naključno enakomerno inicializiraj $\mathbf{G}$ z vrednostmi z intervala $[0, 1]$;
while *konvergenca ni dosežena* **do**
  Za dani $\mathbf{G}$ posodobi $\mathbf{F} = \mathbf{X}\mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}$;
  $\mathbf{\Theta} = \mathbf{\Theta}^+ - \mathbf{\Theta}^-$;
  $\mathbf{F}^T\mathbf{F} = (\mathbf{F}^T\mathbf{F})^+ - (\mathbf{F}^T\mathbf{F})^-$;
  $\mathbf{X}^T\mathbf{F} = (\mathbf{X}^T\mathbf{F})^+ - (\mathbf{X}^T\mathbf{F})^-$;
  Za dani $\mathbf{F}$ posodobi $\mathbf{G}$ s pravilom
  $\mathbf{G}_{ij} \leftarrow \mathbf{G}_{ij}\sqrt{\frac{(\mathbf{X}^T\mathbf{F})^+_{ij} + [\mathbf{G}(\mathbf{F}^T\mathbf{F})^-]_{ij} + (\mathbf{\Theta}^-\mathbf{G})_{ij}}{(\mathbf{X}^T\mathbf{F})^-_{ij} + [\mathbf{G}(\mathbf{F}^T\mathbf{F})^+]_{ij} + (\mathbf{\Theta}^+\mathbf{G})_{ij}}}$;
end

Algoritem 1: Kazenska matrična faktorizacija (PMF).

2.3.2 Kazenska matrična tri-faktorizacija

Ena izmed omejitev kazenske matrične faktorizacije je, da lahko z njo rešujemo le optimizacijske naloge z enim samim podatkovnim tipom. To pomeni, da je PMF primerna za obdelavo homogenih podatkov. Kljub temu se vsakodnevno srečujemo s problemi v podatkovni analizi, ki zahtevajo obravnavo heterogenih podatkov. Zato razširimo PMF s kazensko matrično tri-faktorizacijo (tri-PMF) [64], ki se lahko spopada z diadičnimi podatkovnimi nabori, ki vključujejo dva podatkovna tipa.

Pojem *tri-faktorizacije* se nanaša na število razcepnih matričnih faktorjev, s katerimi aproksimiramo vhodno podatkovno matriko. Poleg števila razcepnih matrik je za opis matrične faktorizacije pomembno še število vhodnih matrik. Najbolj preprosti matrični modeli razstavijo eno vhodno matriko v produkt dveh (običajno manjših) matrik – tak je tudi model PMF. Namesto tega je možna hkratna obravnava večih vhodnih matrik, ki bodisi predstavljajo dodatne regularizacijske člene bodisi so predmet razcepa. Model tri-PMF obravnava več vhodnih matrik, a je le ena predmet dekompozicije,

ostali dve sta matriki omejitev med elementi podatkovnih tipov. Pristop za gradnjo napovednih modelov, ki ga predlagamo v tem delu, uporablja več vhodnih matrik, namenjenih faktorizaciji in več matrik za regularizacijo. Nedavno so Zhang in sod. (2011) [69] predlagali večrazcepno hkratno redko faktorizacijo z mrežno regularizacijo (SNMNMF)[7], ki vsako izmed dveh vhodnih matrik aproksimira s tremi matričnimi faktorji in uporablja še dve matriki za predstavitev apriornega znanja ter usmerjanje faktorizacije.

V diadičnem podatkovnem naboru so prisotni elementi dveh podatkovnih tipov, $\mathcal{X}_1$ velikosti n_1 in $\mathcal{X}_2$ velikosti n_2. Pogosto lahko iz $\mathcal{X}_1$ in $\mathcal{X}_2$ zgradimo *relacijsko matriko* $\mathbf{R}_{12} \in \mathbb{R}^{n_1 \times n_2}$ – n.pr. z jedrnimi funkcijami. Element $\mathbf{R}_{12}(ij)$ hrani oceno moči relacije med i-tim elementom $\mathcal{X}_1$ in j-tim elementom $\mathcal{X}_2$. Relacijske matrike največkrat opisujejo simetrične relacije. Naš cilj je dobra aproksimacija relacijske matrike. Dobljene razcepne matrične faktorje je možno uporabiti za hkratno razvrščanje elementov $\mathcal{X}_1$ in $\mathcal{X}_2$ v skupine.

Na relacijski matriki $\mathbf{R}_{12}$ lahko uporabimo nenegativno matrično tri-faktorizacijo (tri-NMF). Ding in sod. (2006) [20] so pokazali, da rešitev tri-NMF ustreza sproščeni rešitvi razvrščanja stolpcev in vrstic relacijske matrike. Točneje, tri-NMF rešuje sledečo optimizacijsko nalogo

$$\min_{\mathbf{G}_1 \geq 0, \mathbf{G}_2 \geq 0, \mathbf{S} \geq 0} ||\mathbf{R}_{12} - \mathbf{G}_1 \mathbf{S} \mathbf{G}_2^T||^2_{\text{Fro}}, \tag{2.22}$$

pri čemer $\mathbf{G}_1$ in $\mathbf{G}_2$ predstavita elemente tipa $\mathcal{X}_1$ v c_1 razsežnem prostoru in tipa $\mathcal{X}_2$ v c_2 razsežnem prostoru. Če odstranimo zahtevo nenegativnosti razcepnega faktorja $\mathbf{S}$, rešujemo semi-nenegativno matrično tri-faktorizacijo.

O podatkih $\mathcal{X}_1$ in $\mathcal{X}_2$ imamo lahko še dodatno znanje, ki ga predstavimo z omejitvami med pari elementov istega tipa. Namesto reševanja naloge (2.22) zato rešujemo optimizacijsko nalogo

$$\min_{\mathbf{G}_1 \geq 0, \mathbf{G}_2 \geq 0} ||\mathbf{R}_{12} - \mathbf{G}_1 \mathbf{S} \mathbf{G}_2^T||^2_{\text{Fro}} + P(\mathcal{X}_1) + P(\mathcal{X}_2), \tag{2.23}$$

kjer člena $P(\mathcal{X}_1)$ in $P(\mathcal{X}_2)$ označujeta kazni za kršenje omejitev. Ta predsta-

[7]Algoritem modela SNMNMF je implementiran v programski knjižnici Nimfa in je dostopen na `http://nimfa.biolab.si`.

vimo s kvadratnima formama $P(\mathcal{X}_1) = tr(\mathbf{G}_1^T \mathbf{\Theta}_1 \mathbf{G}_1)$ in $P(\mathcal{X}_2) = tr(\mathbf{G}_2^T \mathbf{\Theta}_2 \mathbf{G}_2)$, pri čemer sta $\mathbf{\Theta}_1 \in \mathbb{R}^{n_1 \times n_1}$ in $\mathbf{\Theta}_2 \in \mathbb{R}^{n_2 \times n_2}$ omejitveni matriki za $\mathcal{X}_1$ in $\mathcal{X}_2$. Vrednost $\mathbf{\Theta}_1(i,j)$ ($\mathbf{\Theta}_2(i,j)$) pomeni kazen za kršitev omejitev med i-tim in j-tim elementov $\mathcal{X}_1$ ($\mathcal{X}_2$).

Sedaj definirajmo optimizacijski problem, ki ga rešuje kazenska matrična tri-faktorizacija, kot minimizacijsko nalogo

$$\min_{\mathbf{G}_1 \geq 0, \mathbf{G}_2 \geq 0} J = ||\mathbf{R}_{12} - \mathbf{G}_1 \mathbf{S} \mathbf{G}_2^T||_{\text{Fro}}^2 + tr(\mathbf{G}_1^T \mathbf{\Theta}_1 \mathbf{G}_1) + tr(\mathbf{G}_2^T \mathbf{\Theta}_2 \mathbf{G}_2). \quad (2.24)$$

Wang in sod. (2008) [64] so predstavili iterativni algoritem za reševanje naloge (2.24), imenovan kazenska matrična tri-faktorizacija (tri-PMF), ki je priložen v algoritmu 2. Tri-PMF je razširitev modela PMF, dokaz pravilnosti in konvergence je podoben kot pri modelu PMF. Bralec bo dokaza pravilnosti in konvergence za algoritem tri-PMF našel v [64]. Opazimo, da relacijsko matriko $\mathbf{R}_{12}$ aproksimiramo s tremi razcepnimi matričnimi faktorji, zato je tako poimenovanje faktorizacije primerno.

Aproksimacija matrike $\mathbf{R}_{12}$ je dana z $\mathbf{R}_{12} \approx \mathbf{G}_1 \mathbf{S} \mathbf{G}_2^T$.

Vhod: relacijska matrika $\mathbf{R}_{12}$, omejitvena matrika $\mathbf{\Theta}_1$, rang c_1, omejitvena matrika $\mathbf{\Theta}_2$, rang c_2.

Izhod: razcepni faktorji $\mathbf{G}_1$, $\mathbf{S}$ in $\mathbf{G}_2$.

Naključno enakomerno inicializiraj $\mathbf{G}_1$ z vrednostmi z intervala $[0, 1]$;

Naključno enakomerno inicializiraj $\mathbf{G}_2$ z vrednostmi z intervala $[0, 1]$;

while *konvergenca ni dosežena* **do**

Za dana $\mathbf{G}_1, \mathbf{G}_2$ posodobi $\mathbf{S} = (\mathbf{G}_1^T\mathbf{G}_1)^{-1}\mathbf{G}_1^T\mathbf{R}_{12}\mathbf{G}_2(\mathbf{G}_2^T\mathbf{G}_2)^{-1}$;

$\mathbf{\Theta}_1 = \mathbf{\Theta}_1^+ - \mathbf{\Theta}_1^-$;

$\mathbf{\Theta}_2 = \mathbf{\Theta}_2^+ - \mathbf{\Theta}_2^-$;

$\mathbf{R}_{12}\mathbf{G}_2\mathbf{S}^T = (\mathbf{R}_{12}\mathbf{G}_2\mathbf{S}^T)^+ - (\mathbf{R}_{12}\mathbf{G}_2\mathbf{S}^T)^-$;

$\mathbf{R}_{12}^T\mathbf{G}_1\mathbf{S} = (\mathbf{R}_{12}^T\mathbf{G}_1\mathbf{S})^+ - (\mathbf{R}_{12}^T\mathbf{G}_1\mathbf{S})^-$;

$\mathbf{S}\mathbf{G}_1^T\mathbf{G}_1\mathbf{S}^T = (\mathbf{S}\mathbf{G}_1^T\mathbf{G}_1\mathbf{S}^T)^+ - (\mathbf{S}\mathbf{G}_1^T\mathbf{G}_1\mathbf{S}^T)^-$;

$\mathbf{S}^T\mathbf{G}_2^T\mathbf{G}_2\mathbf{S} = (\mathbf{S}^T\mathbf{G}_2^T\mathbf{G}_2\mathbf{S})^+ - (\mathbf{S}^T\mathbf{G}_2^T\mathbf{G}_2\mathbf{S})^-$;

Za dana $\mathbf{S}, \mathbf{G_2}$ posodobi $\mathbf{G}_1$ s pravilom

$$\mathbf{G}_{1ij} \leftarrow \mathbf{G}_{1ij}\sqrt{\frac{(\mathbf{R}_{12}\mathbf{G}_2\mathbf{S}^T)^+_{ij}[\mathbf{G}_1(\mathbf{S}^T\mathbf{G}_2^T\mathbf{G}_2\mathbf{S})^-]_{ij} + (\mathbf{\Theta}_1^-\mathbf{G}_1)_{ij}}{\mathbf{R}_{12}\mathbf{G}_2\mathbf{S}^T)^-_{ij} + [\mathbf{G}_1(\mathbf{S}^T\mathbf{G}_2^T\mathbf{G}_2\mathbf{S})^+]_{ij} + (\mathbf{\Theta}_1^+\mathbf{G}_1)_{ij}}} \ ;$$

Za dana $\mathbf{S}, \mathbf{G_1}$ posodobi $\mathbf{G}_2$ s pravilom

$$\mathbf{G}_{2ij} \leftarrow \mathbf{G}_{2ij}\sqrt{\frac{(\mathbf{R}_{12}^T\mathbf{G}_1\mathbf{S})^+_{ij}[\mathbf{G}_2(\mathbf{S}\mathbf{G}_1^T\mathbf{G}_1\mathbf{S}^T)^-]_{ij} + (\mathbf{\Theta}_2^-\mathbf{G}_2)_{ij}}{(\mathbf{R}_{12}^T\mathbf{G}_1\mathbf{S})^-_{ij} + [\mathbf{G}_2(\mathbf{S}\mathbf{G}_1^T\mathbf{G}_1\mathbf{S}^T)^+]_{ij} + (\mathbf{\Theta}_2^+\mathbf{G}_2)_{ij}}} \ ;$$

end

Algoritem 2: Kazenska matrična tri-faktorizacija (tri-PMF).

Poglavje 3

Združevanje heterogenih podatkovnih virov s simetrično kazensko matrično tri-faktorizacijo

Metode matrične faktorizacije, ki gradijo razcepne matrične faktorje iz več podatkovnih tipov, so bile predstavljene nedavno [13, 69, 64] in se šele uveljavljajo na področju odkrivanja znanj iz podatkov. Najpogosteje se namreč uporabljajo tehnike zgodnjega združevanja. Te vključujejo dodatno znanje z vektorskim združevanjem, tako da dodajajo nove stolpce ali vrstice v vhodno podatkovno matriko. Takšno združevanje zahteva normalizacijo podatkovne matrike pred izvajanjem matrične faktorizacije in lahko vnese v postopek nestabilnost zaradi velikih razsežnosti podatkovne matrike. Drugi način zgodnjega združevanja z matrično faktorizacijo namesto razširjanja podatkovne matrike spreminja njeno vsebino, tako da nad različnimi viri uporabi preslikave, ki ocenjujejo prispevke virov. Izkaže se [69], da standardne faktorizacije z dvema razcepnima matričnima faktorjema ne morejo učinkovito uporabljati matrik, ki opisujejo več različnih vrst spremenljivk. Dvorazcepne matrične faktorizacije ne omogočajo vmesnega združevanja večih matrik z različnimi podatkovnimi tipi skupaj s predhodnim znanjem [69], kot so omrežja, ki predstavljajo odnose med spremenljivkami istega tipa ali različnih tipov.

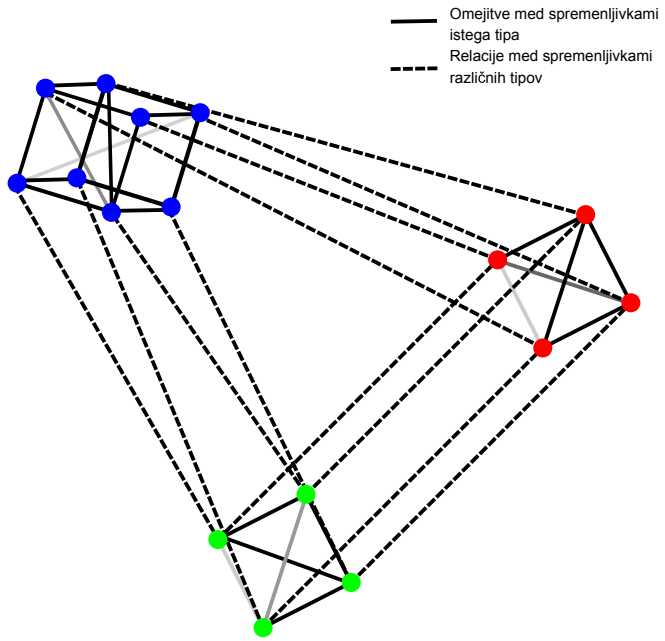
V želji po vključevanju heterogenih podatkovnih virov in predhodnega znanja v gradnjo napovednih modelov z matrično faktorizacijo predstavimo pristop, ki temelji na delno nadzorovani matrični tri-faktorizaciji iz razdelka 2.3.2.

3.1 Matematični zapis problema

Naj bo danih K podatkovnih virov, $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_K\}$, pri čemer i-ti vir vsebuje n_i primerov $\mathcal{X}_i = \{\mathbf{x}_1^i, \mathbf{x}_2^i, \ldots, \mathbf{x}_{n_i}^i\}$. Uporabljene oznake sledijo tistim, predstavljenim v razdelku o kazenski matrični faktorizaciji v tabeli 2.2. Poudarimo, da so viri $\mathcal{X}$ lahko bodisi različne predstavitve (pogledi) entitet istega tipa bodisi predstavljajo različne tipe entitet, med katerimi so znane relacije. Dodajanje podatkovnih virov je konceptualno prikazano na sliki 3.1 s tremi podatkovnimi viri. Sprva viru $\mathcal{X}_1$ (označenem z zeleno barvo) dodamo vir $\mathcal{X}_2$ (označen z rdečo barvo), tako da opredelimo relacijo med dodanim virom $\mathcal{X}_2$ in prvotnim $\mathcal{X}_1$. Ob dodajanju tretjega vira $\mathcal{X}_3$ določimo njegove povezave z obema prej dodanima viroma. Tako dodajanje r-tega vira spominja na razširitev grafa hiperkocke Q_{r+1} na Q_{r+2}.

Naj bo dana množica *relacijskih matrik* $\{\mathbf{R}_{ij} \in \mathbb{R}^{n_i \times n_j}\}$, ki podajajo relacije med vsemi pari podatkovnih virov $\mathcal{X}$. To pomeni, matrika $\mathbf{R}_{ij}$ hrani preslikavo med primeri vira $\mathcal{X}_i$ in $\mathcal{X}_j$. Element $\mathbf{R}_{ij}(k, l)$ veže primera $\mathbf{x}_k^i$ in $\mathbf{x}_l^j$. Predpostavimo še, da relacijske matrike opisujejo simetrične relacije, in torej velja $\mathbf{R}_{ji} = \mathbf{R}_{ij}^T$. Število potrebnih relacijskih matrik za predstavitev K podatkovnih virov se zmanjša za polovico.

Z dostopnim dodatnim znanjem za podatkovi vir $\mathcal{X}_i$ zgradimo *omejitveno matriko* $\mathbf{\Theta}_i$. Če za nek vir predhodno znanje ni na voljo, je pripadajoča omejitvena matrika ničelna. Omejitvena matrika $\mathbf{\Theta}_i$ je le posebna relacijska matrika, v kateri element $\mathbf{\Theta}_i(k, l)$ veže primera $\mathbf{x}_k^i$ in $\mathbf{x}_l^i$. V nasprotju z relacijskimi matrikami omejitvene matrike le usmerjajo faktorizacijski proces in jih ne razcepimo na matrične faktorje. Omejitve so dveh vrst. Pozitivne omejitve izražajo nagrade, če imata ustrezna primera podoben zapis v ma-



Slika 3.1: Konceptualna predstavitev večih podatkovnih virov z relacijami med spremenljivkami različnih tipov (prekinjene črte) in omejitvami med spremenljivkami istega tipa. Pozitivne omejitve so označeno s sivimi neprekinjenimi črtami, negativne omejitve s črnimi neprekinjenimi črtami. Prikazan je primer za tri podatkovne vire $\mathcal{X}_1, \mathcal{X}_2$ in $\mathcal{X}_3$, ki so označeni z različnimi barvami (rdeča, modra, zelena). Za vključitev novega podatkovnega vira je potrebno določiti relacije med novim in obstoječimi viri ter morebitne lastne omejitve vira.

tričnih razcepnih faktorjih. *Zapis* ali *profil* primera $\mathbf{x}_k^i$ je vektor dolžine c_i v k-ti vrstici razcepnega matričnega faktorja za i-ti podatkovni vir. Negativne omejitve so kazni, ki usmerjajo pravila za posodabljanje razcepnih faktorjev, če imata pripadajoča primera različna zapisa. Nagradam v matriki ustrezajo negativne vrednosti, kaznim pa pozitivne vrednosti. Rešujemo namreč minimizacijsko optimizacijsko nalogo in nagrade zmanjšujejo vrednost kriterijske funkcije (nagrajujejo trenutno aproksimacijo), kazni pa vrednost kriterijske funkcije večajo. Omejitvene matrike lahko vsebujejo ničelne elemente, če omejitve niso znane. Kompakten zapis omejitvenih matrik je določen natanko tako kot v primeru kazenske matrične faktorizacije z enačbo (2.20) v

razdelku 2.3.1.

Simetrična kazenska matrična tri-faktorizacija (tri-SPMF) je razširitev metode tri-PMF [64]. Gre za tri-razcepno faktorizacijo, kar pomeni, da vsako relacijsko matriko predstavimo s produktom treh manjših razcepnih faktorjev. Optimizacijska naloga, ki jo rešujemo, posplošuje nalogo tri-PMF

$$\min_{\mathbf{G}_1 \geq 0, \ldots, \mathbf{G}_K \geq 0} \sum_{0<i<j\leq K} ||\mathbf{R}_{ij} - \mathbf{G}_i \mathbf{S}_{ij} \mathbf{G}_j^T|| + \sum_i tr(\mathbf{G}_i^T \mathbf{\Theta}_i \mathbf{G}_i), \tag{3.1}$$

kjer so matrike $\mathbf{G}_i \in \mathbb{R}^{n_i \times c_i}$ in $\mathbf{S}_{ij} \in \mathbb{R}^{c_i \times c_j}$ razcepni matrični faktorji metode tri-SPMF.

Množico relacijskih in omejitvenih matrik želimo zapisati v kompaktni obliki za učinkovito računanje in predstavitev. Wang in sod. (2008) [64] so pokazali v lemi 4.1 v [64], da je reševanje optimizacijske naloge metode tri-PMF ekvivalentno reševanju optimizacijske naloge, v kateri so relacijske in omejitvene matrike ter razcepni matrični faktorji le bloki večjih relacijskih, omejitvenih in razcepnih bločnih matrik. To je zelo dober rezultat, ki ga uporabimo za metodo tri-SPMF. Nove bločne matrike iz množice relacijskih in omejitvenih matrik in razcepnih faktorjev predstavimo v obliki

$$
\begin{aligned}
\mathbf{R} &= \begin{bmatrix} \mathbf{0} & \mathbf{R}_{12}^{n_1\times n_2} & \cdots & \mathbf{R}_{1K}^{n_1\times n_K} \\ \mathbf{R}_{21}^{n_2\times n_1} & \mathbf{0} & \cdots & \mathbf{R}_{2K}^{n_2\times n_K} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_{K1}^{n_K\times n_1} & \mathbf{R}_{K2}^{n_K\times n_2} & \cdots & \mathbf{0} \end{bmatrix}, \\
\mathbf{\Theta} &= \begin{bmatrix} \mathbf{\Theta}_1^{n_1\times n_2} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Theta}_2^{n_2\times n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{\Theta}_K^{n_K\times n_K} \end{bmatrix}, \\
\mathbf{G} &= \begin{bmatrix} \mathbf{G}_1^{n_1\times c_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_2^{n_2\times c_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{G}_K^{n_K\times c_K} \end{bmatrix}, \\
\mathbf{S} &= \begin{bmatrix} \mathbf{0} & \mathbf{S}_{12}^{c_1\times c_2} & \cdots & \mathbf{S}_{1K}^{c_1\times c_K} \\ \mathbf{S}_{21}^{c_2\times c_1} & \mathbf{0} & \cdots & \mathbf{S}_{2K}^{c_2\times c_K} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_{K1}^{c_K\times c_1} & \mathbf{S}_{K2}^{c_K\times c_2} & \cdots & \mathbf{0} \end{bmatrix}.
\end{aligned}
\tag{3.2}
$$

Matrika $\mathbf{R}$ iz (3.2) je simetrična bločna matrika z ničelnimi bloki na glavni diagonali ter relacijskimi matrikami ob straneh. V (3.2) je $\mathbf{\Theta}$ bločna diagonalna matrika z omejitvenimi matrikami na diagonali. Matriki $\mathbf{G}$ in $\mathbf{S}$ hranita razcepne faktorje, $\mathbf{G}$ je bločna matrika z neničelnimi bloki na glavni diagonali in $\mathbf{S}$ je simetrična bločna matrika z ničelnimi bloki na glavni diagonali.

Optimizacijsko nalogo (3.1) tri-SPMF zapišemo v novi obliki

$$
\min_{\mathbf{G}\geq 0} ||\mathbf{R} - \mathbf{G}\mathbf{S}\mathbf{G}^T|| + tr(\mathbf{G}^T\mathbf{\Theta}\mathbf{G}). \tag{3.3}
$$

Nova predstavitev ni le kompaktni zapis. Sedaj lahko na nalogo (3.3) gledamo kot na faktorizacijo simetrične matrike $\mathbf{R}$ v razcepna faktorja $\mathbf{G}$ in $\mathbf{S}$. Zaradi simetričnosti matrike $\mathbf{R}$ in kazenskih členov v $\mathbf{\Theta}$ nalogo imenujemo

simetrična kazenska matrična tri-faktorizacija. Za reševanje optimizacijske naloge lahko uporabimo algoritem 2 za tri-PMF. Tega z upoštevanjem nove predstavitve navajamo v algoritmu 3. Pravilnost in konvergenca tri-SPMF sledita iz lastnosti tri-PMF. Relacijsko matriko $\mathbf{R}_{ij}$ rekonstruiramo na način

$$\mathbf{R}_{ij} \approx \mathbf{G}_i \mathbf{S}_{ij} \mathbf{G}_j^T \tag{3.4}$$

iz pripadajočih razcepnih matričnih faktorjev $\mathbf{G}_i$, $\mathbf{G}_j$ in $\mathbf{S}_{ij}$.

Vhod: relacijske matrike $\{\mathbf{R}_{ij}\}_{1 \le i < j \le K}$, omejitvene matrike $\mathbf{\Theta}_1, \mathbf{\Theta}_2, \ldots \mathbf{\Theta}_K$, rangi $c_1, c_2, \ldots, c_K$.
Izhod: razcepna faktorja $\mathbf{G}$ in $\mathbf{S}$.
Naključno enakomerno inicializiraj $\mathbf{G}_1, \mathbf{G}_2, \ldots \mathbf{G}_K$ z vrednostmi z intervala $[0, 1]$;
Sestavi matrike $\mathbf{R}$, $\mathbf{G}$ in $\mathbf{\Theta}$ kot v (3.2);
while *konvergenca ni dosežena* **do**
  Za dani $\mathbf{G}$ posodobi $\mathbf{S} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{R}\mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}$;
  $\mathbf{\Theta} = \mathbf{\Theta}^+ - \mathbf{\Theta}^-$;
  $\mathbf{RGS} = (\mathbf{RGS})^+ - (\mathbf{RGS})^-$;
  $\mathbf{SG}^T\mathbf{GS} = (\mathbf{SG}^T\mathbf{GS})^+ - (\mathbf{SG}^T\mathbf{GS})^-$;
  Za dani $\mathbf{S}$ posodobi $\mathbf{G}$ s pravilom
  $\mathbf{G}_{ij} \leftarrow \mathbf{G}_{ij} \sqrt{\frac{(\mathbf{RGS})^+_{ij} [\mathbf{G}(\mathbf{SG}^T\mathbf{GS})^-]_{ij} + (\mathbf{\Theta}^- \mathbf{G})_{ij}}{(\mathbf{RGS})^-_{ij} + [\mathbf{G}(\mathbf{SG}^T\mathbf{GS})^+]_{ij} + (\mathbf{\Theta}^+ \mathbf{G})_{ij}}}$;
end

Algoritem 3: Simetrična kazenska matrična tri-faktorizacija (tri-SPMF).

3.2 Gradnja omejitvenih in relacijskih matrik

Vsebina omejitvenih in relacijskih matrik zavisi od konkretnega problema, ki ga rešujemo. Konvergenco izboljšamo, če relacijske matrike pred faktorizacijo normaliziramo – običajno normaliziramo vsak stolpec (vrstico) relacijske matrike posebej. Višje absolutne vrednosti omejitvenih matrik pomenijo večje

kaznovanje (nagrajevanje) podobnih zapisov primerov. Zato je primerno, da nagrade in kazni zavzemajo primerljiv številski obseg.

Pri gradnji omejitvenih in relacijskih matrik je pomembna obravnava neznanih vrednosti. Ta problem je znan kot *matrično dopolnjevanje* (angl. *matrix completion*). Enostavna rešitev je, da na mestu, kjer omejitev ali povezanost nista znani, vstavimo ničelno vrednost. Algoritem tri-SPMF, ki smo ga predstavili, ne obravnava posebej ničelnih vrednosti. Prav tako tudi ne upošteva ocene zanesljivosti o vrednostih v matrikah.

Če bomo matrično faktorizacijo uporabili za gradnjo napovednega modela, nastavljanje ničelnih vrednosti vsekakor ni najboljši pristop, saj uvaja pristranskost v napovedni model. Ta pristop je na področju strategij filtriranja medsebojnih povezanosti (angl. *collaborative filtering*) poimenovan *vsi manjkajoči kot negativni* (angl. *all-missing-as-negative*). Nasprotje te metode je tehnika, imenovana *vsi manjkajoči kot neznani* (angl. *all-missing-as-unknown*), ki mestom v matriki z neznanimi vrednostmi dodeli neko vnaprej določeno oznako [47].

Candes in Recht (2012) [11] sta predlagala dopolnjevanje matrik z reševanjem konveksnega optimizacijskega programa in pokazala tesno spodnjo mejo za število potrebnih primerov v vzorcu za pravilno rekonstrukcijo matrike z veliko verjetnostjo. Druge strategije dopolnjevanja temeljijo na ideji vzorčenja in uteževanja manjkajočih vrednosti. Prva strategija predpostavlja, da je verjetnost relacije v negativnem smislu na mestu manjkajoče vrednosti enaka za vse primere v matriki. Zato se manjkajoči vrednosti naključno enakomerno dodeli utež $\delta \in [0, 1]$. Naslednja strategija predpostavlja, če je primer pogosto povezan z drugimi primeri v pozitivnem smislu, potem je verjetnost negativne povezave na mestu manjkajoče vrednosti visoka. Pogostost povezanosti z drugimi primeri merimo z razmerjem med vsoto ocen pozitivnih povezav in vsoto ocen negativnih povezav. Ideji vzorčenja in uteževanja manjkajočih vrednosti sta med seboj primerljivi in delno odpravljata pristranskost pristopa, ki obravnava manjkajoče vrednosti kot negativne. Reševanje konveksnega optimizacijskega programa ali predstavitev vhodnih podatkov v nizko

razsežnem prostoru [47] sta trenutno najboljši znani metodi za dopolnjevanje matrik.

3.3 Inicializacija matričnih razcepnih faktorjev

Algoritmi za PMF, tri-PMF in tri-SPMF najdejo lokalne rešitve optimizacijskih nalog. Njihovo izvajanje je deterministično z izjemo inicializacije matričnih faktorjev. Zato so inicializacijski postopki izredno pomembni v matričnih faktorizacijah, saj ne vplivajo le na kvaliteto končne rešitve, temveč tudi na hitrost konvergence. Večina faktorizacijskih algoritmov, ki se uporabljajo v podatkovni analizi, namreč sledi standardni shemi iterativnega izboljševanja matričnih faktorjev, dokler niso izpolnjeni pogoji konvergence.

Testi, ki smo jih opravili na umetno generiranih relacijskih matrikah, kažejo, da dober inicializacijski algoritem pri enakem številu iteracij v faktorizaciji izboljša kvaliteto končne rešitve (v smislu Frobeniusove razdalje) tudi za 10 %. To pomeni, da je potrebno izvesti manj iteracij faktorizacijskega algoritma, da dobimo zadovoljivo rešitev. V tabeli 3.1 navajamo srednjo Frobeniusovo razdaljo med relacijskimi matrikami in njihovimi aproksimacijami z razcepnimi faktorji v algoritmu tri-SPMF. Relacijske matrike smo zgradili z enakomerno naključno izbranimi elementi z intervala $[0, 1]$, omejitvene matrike smo predstavili z ničelnimi matrikami, saj gre za umetno zgrajeni podatkovni nabor brez znanih omejitev. Število podatkovnih virov v $\mathcal{X}$ in matrični rangi $c_1, c_2, \ldots c_K$ niso korelirani z izbiro inicializacijskega algoritma. Čeprav je naključna inicializacija najenostavnejša, predlagamo za tri-SPMF uporabo katerega izmed inicializacijskih postopkov, navedenih v tabeli 3.1[1].

Naprednejši inicializacijski algoritmi so časovno zahtevnejši, a primerni zaradi hitrejše konvergence faktorizacije in boljše končne rešitve. Algoritem

[1]Inicializacijski algoritmi iz tabele 3.1 so uporabnikom na voljo v programski knjižnici Nimfa, dostopni na `http://nimfa.biolab.si`.

Inic. algoritem \ Opis	$(100, 200, 300)$ $(20, 20, 30)$	$(500, 500, 500)$ $(80, 120, 65)$	$(600, 500, 700)$ $(45, 85, 150)$
NNDSVD	49.662	123.317	145.846
Random C	47.211	118.199	136.762
Random Vcol	47.212	117.401	135.281
Naključno z $[0, 1]$	51.011	127.775	151.210

Tabela 3.1: Srednja Frobeniusova razdalja med relacijskimi matrikami in njihovimi aproksimacijami v algoritmu tri-SPMF za različne inicializacijske algoritme. Opis povzema nastavitve tri-SPMF, in sicer vhodne dimenzije matrik (n_1, n_2, n_3) ter faktorizacijske range (c_1, c_2, c_3). Vsaka izvedba je obsegala 100 iteracij.

random C je nezahtevna oblika inicializacije, ki temelji na ideji dekompozicije CUR – ta dekompozicija predstavi podatkovno matriko z majhnim številom njenih stolpcev in (ali) vrstic. Algoritem random C naključno izbere p največjih stolpcev podatkovne matrike glede na njihovo 2-normo. Vrednost p je vnaprej določeni parameter. Stolpec v razcepnem matričnem faktorju se inicializira s srednjim vektorjem p izbranih stolpcev. Postopek se ponavlja, dokler ni razcepni faktor v celotni inicializiran.

Algoritem *random Vcol* [1] inicializira stolpce (vrstice) razcepnega faktorja s povprečenjem p naključno izbranih stolpcev (vrstic) podatkovne matrike. Ta algoritem je hitrejši od algoritma random C, a manj učinkovit. Njegova zmogljivost je med naključno inicializacijo in *inicializacijo s centroidi*. Slednja inicializira razcepne faktorje na osnovi rešitve centroidne dekompozicije.

Algoritma random C in random Vcol vključujeta naključnost, zato lahko ob njuni uporabi faktorizacijo večkrat ponovimo in upoštevamo najboljši rezultat. Oba algoritma značilno izboljšata kakovost matričnih razcepnih faktorjev glede na povsem naključno inicializacijo, pri čemer je njuno izvajanje hitro. Algoritem dvojnega nenegativnega singularnega razcepa (NNDSVD) [9] je deterministično in temelji na računanju dveh singularnih razcepov. Prvi razcep deluje na podatkovni matriki, drugi pa na matriki, sestavljeni iz pozitivnih elementov razcepnih faktorjev prvega razcepa. Zaradi

računanja dveh singularnih razcepov je NNDSVD časovno zahtevnejši od večine drugih naključnostnih algoritmov.

3.4 Optimizacija matričnih izračunov

Za učinkovito izvedbo rešitve tri-SPMF moramo upoštevati lastnosti matrik, ki jih konstruira algoritem. Relacijska matrika $\mathbf{R}$, omejitvena matrika $\mathbf{\Theta}$ in razcepna faktorja so bločne matrike.

V pravilu za posodabljanje faktorja $\mathbf{S}$ algoritem 3 računa le z matrikama $\mathbf{G}$ in $\mathbf{R}$. Ob računanju člena $\mathbf{X} = \mathbf{G}^T\mathbf{G}$ množimo med seboj soležne bločne matrike na diagonali, teh je K. Matrika $\mathbf{G}$ ni bločno diagonalna, ker njeni bloki niso kvadratne matrike. Toda računati je potrebno inverz člena $\mathbf{X}$, ki je bločno diagonalna matrika. Računanje pospešimo, če upoštevamo, da je inverz bločno diagonalne matrike zopet bločno diagonalen, sestavljen iz inverzov posameznih blokov. Če matrika $\mathbf{X}$ ni obrnljiva, njen inverz ne obstaja. V takih slučajih namesto inverza računamo Moore-Penroseov psevdoinverz $\mathbf{X}^+$. Psevdoinverza se poslužujemo tudi, če je matrika $\mathbf{X}$ slabo pogojena, torej ima veliko število občutljivosti. Pri dobro pogojenih matrikah je ostanek med prvotno matriko in njenim približkom dobra ocena za napako v rešitvi, pri slabo pogojenih matrikah pa iz majhnega ostanka ne moremo sklepati, da je tudi napaka rešitve majhna.

Opozoriti velja, da je reševanje ustreznega sistema linearnih enačb praviloma učinkoviteje od računanja inverzne matrike. Inverz $\mathbf{X}^{-1}$ lahko zapišemo kot rešitev matrične enačbe $\mathbf{X}\mathbf{X}^{-1} = \mathbf{I}$, ki jo zapišemo po stolpcih kot zaporedje linearnih sistemov z isto matriko $\mathbf{X}$

$$\mathbf{X}\mathbf{x}^{(j)} = \mathbf{e}^{(j)}\ j = 1, 2, \ldots, n,$$

pri čemer so $\mathbf{x}^{(j)}$ stolpci inverzne matrike $\mathbf{X}^{-1}$, $\mathbf{e}^{(j)}$ pa ustrezni stolpci enotske matrike $\mathbf{I}$.

Če je narava problema taka, da relacijske in omejitvene matrike vsebujejo veliko ničelnih elementov, lahko privarčujemo na prostoru, če matrike predstavimo v enem izmed redkih formatov[2]. Pogosti formati za predstavitev redkih matrik so (i) predstavitev s slovarjem, (ii) povezani seznami, (iii) redki vrstični format in (iv) redki stolpični formati.

[2]Programska knjižnica NIMFA podpira delo z redkimi matriki v formatih `scipy.sparse`.

Poglavje 4

Napovedovanje iz razcepnih matričnih faktorjev

Nastavitev relacijskih in omejitvenih matrik ter izvedba algoritma tri-SPMF je le prvi korak pri gradnji napovednih modelov z matrično faktorizacijo. Faktorizaciji namreč sledi uporaba razcepnih matričnih faktorjev. Predpostavljamo, da nam je na voljo model, ki ga vrne algoritem tri-SPMF – to sta bločna razcepna matrična faktorja $\mathbf{G}$ in $\mathbf{S}$. Uporaba modela zavisi od interpretacije razcepnih faktorjev. V tem poglavju predstavimo nov pristop uporabe razcepnih faktorjev za uvrščanje v skupine in ocenjevanje verjetnosti napovedi.

Razcepni matrični faktorji se v splošnem uporabljajo za (naloge so padajoče razvrščene po pogostosti uporabe):

Zmanjševanje dimenzionalnosti podatkov (angl. *dimensionality reduction*) [36, 1, 29]. Navadno pričakujemo, da podatkovni nabor z več značilkami vsebuje več informacij in bodo izpeljani napovedni modeli točnejši. Toda odkrivanje informacij iz večjega števila značilk je zahtevnejše, znano kot prekletstvo dimenzionalnosti. V analizi dokumentov, genskih podatkov in slik sta zato tehniki *izločanja značilk* (angl. *feature extraction*) in izbiranja značilk (angl. *feature selection*) zelo pomembni. Dekompozicije CUR podatkovno matriko izrazijo s kombinacijo pod-

množice njenih stolpcev in vrstic, zato so primerne za izbor značilk. Ob izločanju značilk iščemo preslikave iz večrazsežnega prostora v prostor z manj dimenzijami s čim več informacije. Običajno so nove značilke zapisi primerov v matričnih razcepnih faktorjih - z omejitvami matričnih faktorjev imajo značilke posebne lastnosti (n.pr. nenegativnost in redkost).

Razvrščanje v skupine (angl. *clustering*) [10, 25]. Primere v podatkovni matriki razvrstimo v c skupin, kjer je c matrični rang faktorizacije. Kriteriji uvrščanja primera v skupino (več skupin) temeljijo na maksimalnih elementih v zapisih primerov.

Aproksimacija nizkega ranga (angl. *low-rank approximation*). Zelo pomembna uporaba v numerični linearni algebri za naloge kot so reševanje linearnih sistemov enačb, računanje približnih matričnih inverzov in iskanje rešitev enačb z več spremenljivkami.

Regresijske in klasifikacijske naloge [32, 38]. Matrične faktorizacije so zelo uspešne na področju strategij filtriranja medsebojnih povezanosti v priporočilnih sistemih.

4.1 Izbor kandidatov

Naj v množici heterogenih podatkovnih virov $\mathcal{X}$ obstajata vira $\mathcal{X}_i$ in $\mathcal{X}_j$, pri čemer $\mathcal{X}_i = \{\mathbf{x}_1^i, \mathbf{x}_2^i, \ldots, \mathbf{x}_{n_i}^i\}$ vsebuje primere in $\mathcal{X}_j = \{\mathbf{x}_1^j, \mathbf{x}_2^j, \ldots, \mathbf{x}_{n_j}^j\}$ opise oznak. Znanje o njuni neposredni povezanosti je predstavljeno v matriki $\mathbf{R}_{ij}$. V modelu tri-SPMF so pripadajoči razcepni faktorji $\mathbf{G}_i \in \mathbb{R}^{n_i \times c_i}$, $\mathbf{G}_j \in \mathbb{R}^{n_j \times c_j}$ in $\mathbf{S}_{ij} \in \mathbb{R}^{c_i \times c_j}$, tako da velja $\widehat{\mathbf{R}}_{ij} = \mathbf{G}_i \mathbf{S}_{ij} \mathbf{G}_j^T$. Kandidati so dvojice $(\mathbf{x}_k^i, \mathbf{x}_l^j)$, sestavljene iz primera $\mathbf{x}_k^i$ in oznake $\mathbf{x}_l^j$. Seveda nas pri napovedovanju zanimajo le povezave z neznanimi vrednostmi na mestih $\mathbf{R}_{ij}(k, l)$. V pričujočem delu se ukvarjamo s klasifikacijskim problemom, zato bodisi $\mathbf{x}_k^i \in \mathbf{x}_l^j$ bodisi $\mathbf{x}_k^i \notin \mathbf{x}_l^j$.

Za določitev kandidatov $(\mathbf{x}_k^i, \mathbf{x}_l^j)$ uporabimo matriko $\widehat{\mathbf{R}}_{ij}$. Nekatere uporabe matričnih faktorizacij za določitev pripadnosti primerov razredom iščejo največje koeficiente v vrsticah (stolpcih) matričnih razcepnih faktorjev [31, 10] ali uporabijo pragovno funkcijo na izračunani standardni z-oceni za vsako mesto v razcepnem faktorju [69].

Tako določevanje ni ustrezno, če je primeru $\mathbf{x}_k^i$ lahko dodeljenih več oznak iz $\mathcal{X}_j$. S tem se srečujemo pri napovedovanju genskih funkcij v poglavju 5. Zato za vsak primer v $\mathcal{X}_i$ izračunamo srednjo vrednost njegovega zapisa v matriki $\widehat{\mathbf{R}}_{ij}$ preko vseh oznak. Par $(\mathbf{x}_k^i, \mathbf{x}_l^j)$ je dodan v seznam kandidatov, če je vrednost $\mathbf{R}_{ij}(k,l)$ večja od srednje vrednosti zapisa

$$candidate_{\widehat{\mathbf{R}}}(k,l) \iff s(k,l) = \widehat{\mathbf{R}}_{ij}(k,l) > \frac{1}{n_j}\sum_{m=1}^{n_j} \widehat{\mathbf{R}}_{ij}(k,m). \tag{4.1}$$

Ob izbiranju kandidatov ne upoštevamo zapisov oznak v matričnih razcepnih faktorjih. Ocene oznak so potrebne v naslednji fazi, ko ocenjujemo verjetnosti kandidatov, to je verjetnost, da primer $\mathbf{x}_k^i$ označimo z oznako $\mathbf{x}_l^j$.

V rezultatih, predstavljenih v poglavjih 5 in 6, uporabljamo ansambelski pristop v matrični faktorizaciji. Zaradi naključnostnih algoritmov za inicializacijo razcepnih faktorjev model konvergira različnim lokalno optimalnim rešitvam. Izkaže se, da za značilno izboljšanje napovedi zadostujejo že tri do devet ponovitev. V ansambelskem pristopu par dodamo v seznam kandidatov po pravilu glasovanja, če ima primer visoko srednjo vrednost zapisa v večini ponovitev.

Izbor kandidatov, določen z (4.1), je pri napovedovanju genskih funkcij in rezistence amebe *D. discoideum* najuspešnejši. Predlagamo še dva alternativni načina za izbor kandidatov. Prvi način računa srednjo vrednost zapisa oznake $\mathbf{x}_l^j$ in uporablja zapis primera za ocenjevanje verjetnosti

$$candidate_{\widehat{\mathbf{R}}}(k,l) \iff s(k,l) = \widehat{\mathbf{R}}_{ij}(k,l) > \frac{1}{n_i}\sum_{m=1}^{n_i} \widehat{\mathbf{R}}_{ij}(m,l). \tag{4.2}$$

Izbor (4.2) je simetričen pravilu (4.1). V drugem načinu izračunamo standardno z-oceno za par $(\mathbf{x}_k^i, \mathbf{x}_l^j)$ iz vrstic v razcepnih faktorjev $\mathbf{G}_i$ in $\mathbf{G}_j$

$$candidate_{\widehat{\mathbf{R}}}(k,l) \iff z(k,l) = \frac{\widehat{\mathbf{R}}_{ij}(k,l) - \mu_k}{\sigma_k} > T_1, \tag{4.3}$$

pri čemer je μ_k srednja vrednost zapisa $\mathbf{x}_k^i$ (k-ta vrstica) v faktorju $\mathbf{G}_i$, in σ_k standardni odklon. Vrednost T_1 je nastavitveni parameter. V matričnih faktorizacijah je simetrija zelo pogosta lastnost. Zapis v (4.3) se osredotoča na primer $\mathbf{x}_k^i$, z zamenjavo izračunov srednjih vrednosti in standardnega odklona lahko dobimo razredno orientirano pravilo za razred (oznako) $\mathbf{x}_l^j$, tako da velja

$$candidate_{\widehat{\mathbf{R}}}(k,l) \iff z(k,l) = \frac{\widehat{\mathbf{R}}_{ij}(k,l) - \mu_l}{\sigma_l} > T_2. \tag{4.4}$$

Ideja za pravili (4.3) in (4.4) izhaja iz dela Zhang in sod. (2011), v katerem so z dvorazcepno faktorizacijo SNMNMF razpoznavali komodule genov in miRNA.

Iz pravil (4.3) in (4.4) lahko sestavimo linearno kombinacijo, v kateri upoštevamo zapise primera in oznake v razcepnih faktorjih

$$candidate_{\widehat{\mathbf{R}}}(k,l) \iff \alpha \cdot \frac{\widehat{\mathbf{R}}_{ij}(k,l) - \mu_k}{\sigma_k} + \beta \cdot \frac{\widehat{\mathbf{R}}_{ij}(k,l) - \mu_l}{\sigma_l} > T_3. \tag{4.5}$$

Parametra α in β lahko nastavimo z notranjim prečnim preverjanjem.

4.2 Prioritizacija napovedi z ocenjevanjem verjetnosti

Po zaključenem izboru kandidatov nam je na voljo seznam parov z obetavnimi prireditvami oznak $\mathbf{x}_l^j$ primerom $\mathbf{x}_k^i$. Pogosto je naša delovna hipoteza, da je manjše število kandidatov v seznamu zares pomembnih. Kandidate želimo *prioritizirati*, prednostno razvrstiti od najbolj do najmanj verjetnih.

V biologiji to lahko pomeni, da le eden ali nekaj genov neposredno povzroča redko bolezen, ki jo preučujemo – to so bolezenski geni (angl. *disease genes*). Prepoznavanje najbolj obetavnih genov v seznamu kandidatov je zamudno in zahteva drage laboratorijske raziskave. Značilno je, da morajo biologi ročno pregledati sezname kandidatov, preveriti, kaj je trenutno znanega o vsakem genu, in oceniti, ali gre za obetavnega kandidata. Bioinformatična skupnost je zato uvedla koncept *prioritizacije genov* (angl. *gene prioritization*), ki izkorišča napredek v veliki količini javnih genetskih podatkov. Podobnost med vsemi strategijami prioritizacije genov je uporaba "krivde zaradi povezanosti" – najbolj obetavni geni so podobni genom, za katere je že znano, da so povezani s preučevanim biološkim procesom. Tranchevent in sod. (2010) [61] so pregledali rešitve za prioritizacijo genov, dostopne na spletu.

Problem razvrstitve kandidatov je v strojnem učenju znan kot primer *učenja iz pozitivnih in neoznačenih primerov* (učenje PU). Pri učenju PU sta na voljo dve skupini primerov; skupina pozitivnih primerov $\mathcal{P}$ in skupina mešanih primerov $\mathcal{U}$. Skupina $\mathcal{U}$ vsebuje tako pozitivne kot tudi negativne primere, a primeri niso označeni. V običajnem nadzorovanem učenju nam je na voljo učna množica z označenimi primeri obeh razredov. Učenje PU so prvi predstavili Liu in sod. (2002) [40] in je pomembno v klasifikaciji dokumentov. Večina pristopov v učenju PU temelji na reševanju optimizacijskega naloge. Sprva iz neoznačene skupine $\mathcal{U}$ izberemo množico primerov $\mathcal{N}$, ki predstavljajo negativne primere v učenju. V drugem koraku gradimo binarne klasifikatorje. Oba koraka skupaj je mogoče razumeti kot iterativno tehniko povečevanja množice $\mathcal{N}$, pri čemer morajo pozitivni primeri ostati pravilno klasificirani.

Tukaj predlagamo nov način prioritizacije kandidatov, ki temelji na "krivdi zaradi povezanosti." Ta koristi za statistično mero in oceno verjetnosti, da primer $\mathbf{x}_k^i$ pripada razredu $\mathbf{x}_l^j$. Po razvrstitvi kandidatov s pragovno funkcijo določimo, kateri so najbolj obetavni.

Statistično značilnost kandidata $(\mathbf{x}_k^i, \mathbf{x}_l^j)$ ocenimo z uvrstitvijo njegove

ocene $\widehat{\mathbf{R}}_{ij}(k,l)$ v porazdelitev ocen pozitivnih parov za razred $\mathbf{x}_l^j$. Množica ocen pozitivnih parov za razred $\mathbf{x}_l^j$ je $\mathcal{R} = \{\widehat{\mathbf{R}}_{ij}(m,l) \mid \mathbf{x}_m^i \in \mathbf{x}_l^j, m = 1,2,\ldots,n_i\}$. Pričakujemo, da velja: če primer dejansko pripada razredu, bo njegova ocena na mestu v matriki $\widehat{\mathbf{R}}_{ij}$ podobna oceni drugih pozitivnih primerov za ta razred. Ta pristop je razredno osredotočen, ker upošteva porazdelitev ocen za kandidatov razred.

Namesto opazovanja ocen za kandidatov razred $\mathbf{x}_l^j$ se lahko opazuje porazdelitev ocen za kandidatov primer $\mathbf{x}_k^i$. Tedaj se kandidatova ocena uvrsti v porazdelitev ocen $\mathcal{R} = \{\widehat{\mathbf{R}}_{ij}(k,m) \mid \mathbf{x}_k^i \in \mathbf{x}_m^j, m = 1,2,\ldots,n_j\}$. Ideja je podobna kot zgoraj z zamenjano vlogo primera in razreda. Če primer dejansko pripada razredu, bo njegova ocena v $\widehat{\mathbf{R}}_{ij}$ podobna ocenam drugih razredov, ki jim ta primer pripada. Od tod takšna konstrukcija množice $\mathcal{R}$. Ta pristop se osredotoča na primer in ne na razred.

Katerega izmed dveh opisanih načinov izberemo, zavisi od konkretne naloge. Če imamo veliko razredov in vsak razred vsebuje le malo primerov, je smiselen pristop z osredotočenjem na primere. Če o primerih še ni veliko znanega in so razredi veliki, bo bolje deloval razredno osredotočen pristop. Sicer lahko oba pristopa kombiniramo.

4.3 Vrednotenje napovednega modela

Uspešnost napovednega modela zgrajenega s simetrično matrično tri-faktorizacijo iz heterogenih podatkovnih virov ocenjujemo s prečnim preverjanjem. Predpostavimo, da modela ne uporabljamo za izločanje značilk ali za zmanjševanje dimenzionalnosti podatkov, temveč za razvrščanje. To pomeni, da v modelu obstaja relacijska matrika $\mathbf{R}_{ij}$ z racepnimi matričnimi faktorji $\mathbf{G}_i, \mathbf{G}_j$ in $\mathbf{S}_{ij}$, pri čemer vir $\mathcal{X}_i = \{\mathbf{x}_1^i, \mathbf{x}_2^i, \ldots, \mathbf{x}_{n_i}^i\}$ interpretiramo kot množico primerov in vir $\mathcal{X}_j = \{\mathbf{x}_1^j, \mathbf{x}_2^j, \ldots, \mathbf{x}_{n_j}^j\}$ kot množico razrednih oznak. Naš cilj je čim bolje napovedati razrede, katerim pripada dani primer, in oceniti zanesljivost napovedi. Naj par $(\mathbf{x}_k^i, \mathbf{x}_l^j)$ označuje nalogo, v kateri bodisi $\mathbf{x}_k^i \in \mathbf{x}_l^j$ bodisi $\mathbf{x}_k^i \notin \mathbf{x}_l^j$.

Vse znane pare $(\mathbf{x}_k^i, \mathbf{x}_l^j)$ razvrstimo v C enako velikih množic. Nato C-krat zgradimo model in napovemo oznake, vsakič je ena množica testna, ostale so učne. Če je par $(\mathbf{x}_k^i, \mathbf{x}_l^j)$ v testni množici, pomeni le, da je na mestu $\mathbf{R}_{ij}(k,l)$ neznana vrednost. Torej je v vseh C ponovitvah učenja in preverjanja matrika $\mathbf{R}_{ij}$ enake velikosti, to je $\mathbf{R}_{ij} \in \mathbb{R}^{n_i \times n_j}$, spreminja se le njena polnost.

Korektno testiranje zahteva še eno pozornost. Nastavitev mesta $\mathbf{R}_{ij}(k,l)$ na neznano vrednost ne zadostuje, poskrbeti moramo, da para $(\mathbf{x}_k^i, \mathbf{x}_l^j)$ ne upoštevamo v drugih relacijskih in omejitvenih matrikah.

Zaradi jasnosti si oglejmo primer. Dani so trije podatkovni tipi $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3\}$, $\mathcal{X}_1$ vsebuje primere, $\mathcal{X}_2$ oznake. Naj bo par $(\mathbf{x}_k^1, \mathbf{x}_l^2)$ v testni množici in mesto $\mathbf{R}_{12}(k,l)$ nastavimo na neznano vrednost. Recimo, da $(\mathbf{x}_k^1, \mathbf{x}_l^2)$ uporabljamo pri gradnji matrik $\mathbf{R}_{13}$ in $\mathbf{R}_{23}$, na primer za računanje $\mathbf{R}_{13}(k,m)$ in $\mathbf{R}_{23}(m,l)$. Testiranje ni korektno, ker posredno uporabljamo testni primer za gradnjo napovednega modela. Če imamo več podatkovnih virov, je tako iskanje ciklov zamudno. Iskanje ciklov ni potrebno, če v vsaki izmed C ponovitev ponovno zgradimo vse relacijske in omejitvene matrike.

Uspešnost napovedovanja ocenjujemo z uteženo oceno F_1. Ocena F_1 je definirana s priklicem r in točnostjo p

$$F_1 = 2 \cdot \frac{p \cdot r}{p + r}, \quad p = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \quad r = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{4.6}$$

Priklic podaja, kolikšen delež primerov ciljnega razreda je klasifikator pravilno napovedal. Točnost nam pove, kolikšen del napovedi dejansko pripada ciljnemu razredu. Relativni prispevek točnosti in priklica oceni F_1 je enak.

V učni množici se lahko pojavlja največ n_j različnih oznak. Zato uspešnost ocenjujemo z uteženo oceno F_1. Za vsak razred i posebej izračunamo oceno F_1 in ocene utežimo s pogostostjo razreda $w(i)$

$$F_{1\ \mathrm{fin}} = \sum_{i=i}^{n_j} w(i) \cdot F_1(i). \tag{4.7}$$

Poglavje 5

Napovedovanje genskih določitev amebe *D. discoideum*

Genom in genskim produktom organizma so pogosto pripisani koncepti genske ontologije. *Genski pripis* ali *določitev* za nas pomeni opis dejavnosti in lokalizacije genskih produktov ter opredelitev, kakšna vrsta dokaza potrjuje določitev. Utemeljitve se razlikujejo po stopnji zanesljivosti. Najbolj verodostojne so eksperimentalne potrditve, manj utemeljitve kuratorjev ontologije ali izpeljave iz računskih analiz.

Med najbolj znanimi ontologijami v bioinformatičnimi skupnostmi je projekt GO Ontology[1][4]. Ta zagotavlja ontologije opredeljenih konceptov, ki opisujejo lastnosti genskih produktov. Pokriva tri velika področja: (i) celične komponente, dele celice in zunajcelično okolje; (ii) molekularne funkcije, elementarne dejavnosti genov na nivoju molekul, kot so vezava ali kataliza; in (iii) biološke procese, sklopi molekularnih dogodkov z opredeljenim začetkom in koncem v celicah, tkivih, organih in organizmih. Ontologija je strukturirana kot usmerjen aciklični graf, vsak koncept ima opredeljene povezave do

[1]Genska ontologija GO Ontology je javno dostopna na spletu na naslovu `http://www.geneontology.org`.

enega ali več drugih konceptov. Ontološki slovar je zasnovan tako, da je vsebina neodvisna od vrste organizma, vključeni koncepti veljajo za evkarionte, prokarionte, eno- in večcelične organizme.

Omejena različica genske ontologije se imenuje GO Slim. Ta vsebuje podmnožico konceptov iz celotne ontologije in daje širok pregled nad vsebino brez natančno opredeljenih drobnozrnatih konceptov. Omejene različice ustvarjajo raziskovalne skupine glede na svoje potrebe in so lahko specifične za vrsto organizma ali področje ontologije.

Naš cilj je zgraditi napovedni model iz razcepnih faktorjev simetrične kazenske matrične tri-faktorizacije, ki bo za gene amebe *D. discoideum* čim bolje napovedal neznane koncepte genske ontologije. Omenimo naj, da je ta sicer izjemno zanimiva socialna ameba še relativno slabo raziskana in da je eksperimentalno potrjenih genskih določitev iz njenega genoma le malo. Zato je razvoj orodij, ki bodo predlagala dobre hipoteze na tem področju, zelo zanimiva in aktualna naloga. Ustrezna orodja, ki bi služila napovedovanju genskih določitev *D. discoideum* in pri tem uporabila različne podatkovne vire, do sedaj še niso bila predlagana.

Pri gradnji modela za amebo *D. discoideum* združujemo tri podatkovne vire:

- **znane genske pripise** – Za 13295 amebinih genov imamo na voljo 19810 neposrednih pripisov o 2046 konceptih iz ontologije. Eksperimentalno potrjenih določitev je 4326, pripisov iz računskih analiz je 9316 in pregledanih pripisov kuratorjev je 6168. Le 5348 genov ima znane določitve, izmed teh ima vsak gen povprečno dve določitvi. Iz omejene splošne različice ontologije GO Slim izpeljemo 83 konceptov, ki pripadajo področjema celičnih komponent in bioloških procesov.

- **proteinske interakcije** – Interakcije so podane s trojicami (p_1, p_2, s), pri čemer $s \in [0, 1]$ pomeni oceno zanesljivosti interakcije med proteinoma p_1 in p_2. Le 3982 genov, ki kodirajo proteine, ima znane proteinske interakcije. Interakcije so javno objavljene in dostopne v bazi

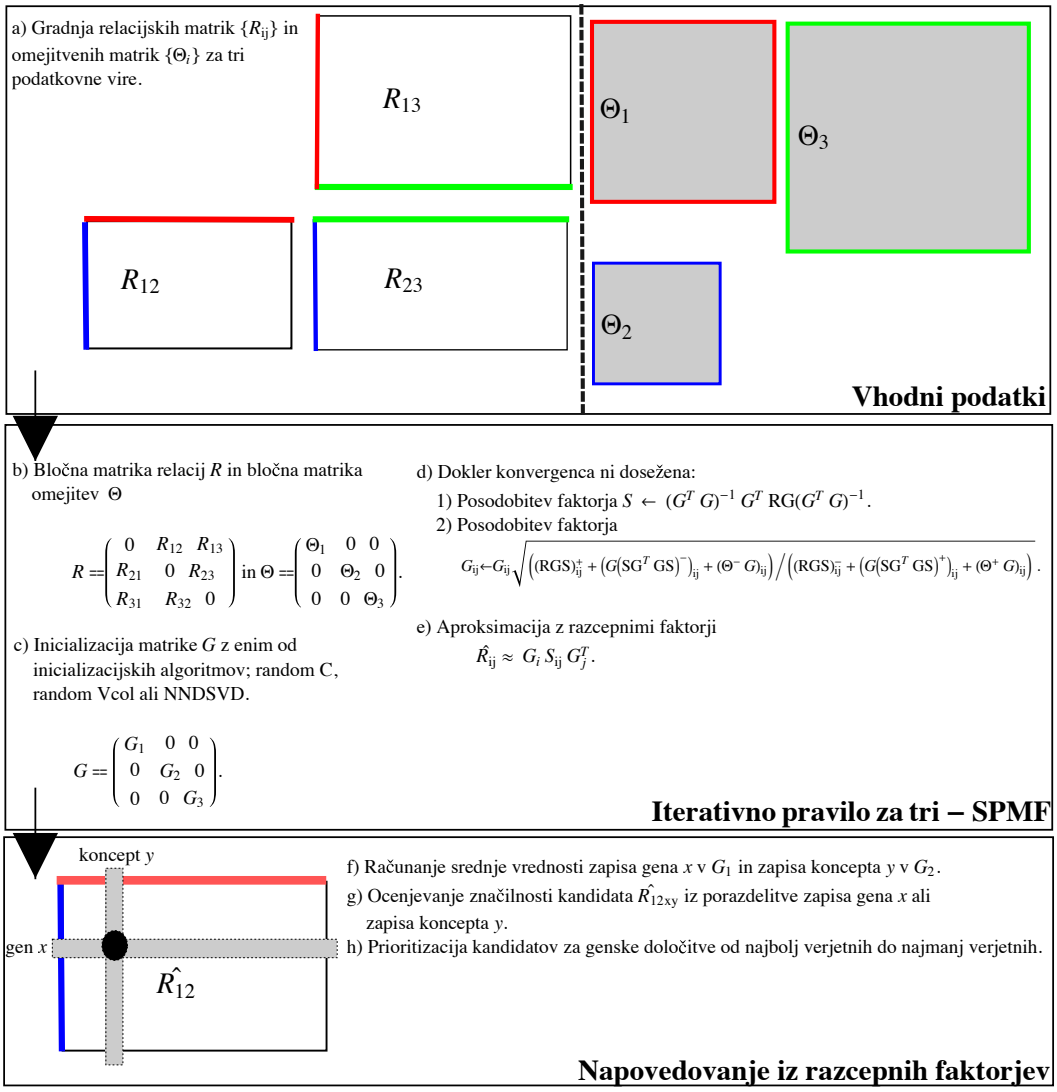
znanih in napovedanih proteinskih interakcij STRING[2].

- **genske izraze** – Celotni genomski transkripcijski zapisi (RNA-seq) amebe *D. discoideum* seva AX4 v 24-urnem razvojnem ciklu, ki je enakomerno vzorčen vsake štiri ure. Razpoložljivi so podatki celic divjega tipa in približno 50 mutant iz družine vezav ATP. Zapisi celic divjega tipa so prosto dostopni v orodju PIPA[3].

Genske določitve amebe *D. discoideum* bomo napovedovali z novim pristopom gradnje napovednih modelov z matrično faktorizacijo iz heterogenih virov. Delo lahko grobo razdelimo v tri faze: (1) gradnja relacijskih in omejitvenih matrik iz heterogenih virov, (2) izvedba simetrične kazenske matrične tri-faktorizacije in (3) napovedovanje iz razcepnih matričnih faktorjev s prioritizacijo napovedi. Opisani postopek je shematsko prikazan na sliki 5.1.

[2]Baza STRING je dostopna na naslovu `http://string-db.org`.

[3]Orodje PIPA za analizo in upravljanje podatkov naslednje generacije sekvenciranja je dostopno na `http://pipa.biolab.si`.



Slika 5.1: Napovedovanje genskih določitev *D. discoideum* z matrično faktorizacijo iz heterogenih virov. Tri podatkovne vire predstavimo s tremi simetričnimi relacijskimi matrikami. Vsakemu viru pripada še omejitvena matrika za vključitev predznanja o viru. Rob matrike je označen z barvo vira, ki je v njej opisan.

5.1 Stanje raziskav v svetu

Število razpoznanih genov zaradi napredka tehnik sekvenciranja v zadnjem desetletju zelo hitro narašča. Določanje genskih funkcij je postal eden izmed osrednjih problemov v molekularni biologiji [23]. Ročno določanje je zaradi velike količine podatkov nemogoče, zato postajajo vse pomembnejše računske metode, ki pomagajo biologom pri načrtovanju bioloških eksperimentov z izborom najbolj obetavnih kandidatov.

Poskusi avtomatskega določanja sledijo dvema glavnima pristopoma. V prvem pristopu primerjamo gen, čigar določitve napovedujemo, z javnimi podatkovnimi bazami znanih določitev. Določitve z najboljšo oceno glede na definicijo podobnosti se prenesejo na ciljni gen. To je *pristop s prenosom* (angl. *transfer approach*) [59]. Kljub znanim slabostim, kot so pretirano upoštevanje tranzitivnosti, nizka senzitivnost in specifičnost ter propagiranje napak v zbirkah podatkov, so tovrstne tehnike danes najbolj razširjene.

Pristop s prenosom je bil zgodovinsko prvi uspešni način avtomatskega določanja [18], razvit še v času, ko so podatkovne zbirke vsebovale bistveno manj genskih zaporedij. Drugi pristop oblikuje problem napovedovanja določitev kot klasifikacijsko nalogo, v kateri ontološki koncepti predstavljajo razrede in geni primere za razvrščanje. *Klasifikacijski pristop* [54] temelji na znanih klasifikacijskih algoritmih, kot so metode podpornih vektorjev in umetne nevronske mreže. Te metode se učijo tako iz pozitivnih kot tudi negativnih primerov, zato so pogosto bolj točne od tehnik s prenosom. Ovira nastopi pri izboru negativnih primerov za učno množico, zato se uporabljajo principi učenja iz pozitivnih in neoznačenih primerov. Genske določitve lahko namreč interpretiramo na več načinov, pri čemer natančno določevanje zavisi od konteksta uporabe gena. Podobne genske funkcije predstavljajo ontološki koncepti z različnimi stopnjami specifičnosti. Za uspešno učenje so potrebni pozitivni in negativni učni primeri za vsak koncept. Pripravo podatkov otežujejo biološke podobnosti med dejavnostmi, ki jih opisujejo koncepti, in genska pripadnost večim konceptom. Metode razdelimo v tri skupine na

osnovi uporabljenih podatkovnih virov:

- analiza homoloških značilnosti [41],
- analiza podzaporedij [53],
- atributno učenje [54].

Tehnike z analizo homoloških značilnosti primerjajo podobnost ciljnega genskega zaporedja z zaporedji pozitivnih in negativnih primerov v učni množici, da napovedo ontološke koncepte. Znano je, da je visoka stopnja identičnosti zaporedij močan pokazatelj funkcijske homologije. Homologija je kvalitativna oznaka, ki poudarja evolucijsko povezanost dveh zaporedij – njuno podobnost po izvoru. Najbolj uveljavljene metode za iskanje podobnosti med genskimi zaporedji so algoritmi lokalne poravnave zaporedij BLAST [2] in PSI-BLAST.

Analiza podzaporedij se osredotoča na dobro ohranjene regije motivov in domen, ki so ključnega pomena za proteine, da opravljajo določene naloge. Te metode so še posebej učinkovite, ko je genska funkcija, opisana z ontološkim konceptom, neposredno povezana z obstojem domene ali motiva. Uporabne so tudi pri sklepanju v oddaljenih homoloških razmerah.

Pri atributnem učenju se sprva določi biološko pomembne lastnosti iz primarnih zaporedij. Te so frekvence ostankov, molekulska masa, sekundarna struktura, koeficienti izumrtja in druge fizikalno-kemijske lastnosti. Njihove vrednosti se predstavi z vektorskim zapisom in nad njimi uporabi kateri koli znan klasifikacijski algoritem. Koncepti s področja bioloških procesov ali celičnih komponent vključujejo gene z zelo različnimi lastnostmi, ki pripadajo istemu razredu, in gene s podobnimi lastnostmi, ki pripadajo različnim razredom. To predstavlja problem za klasifikacijski algoritem z diskriminativnim pristopom optimizacije mej med razredi. Homološki pristopi so nanj manj občutljivi, saj odločitev o razredu upošteva le nekaj najbolje ocenjenih zadetkov.

5.2 Podatkovni viri

Napovedni model s simetrično kazensko matrično tri-faktorizacijo gradimo iz treh podatkovnih virov. To pomeni, da vire predstavimo s šestimi matrikami; s tremi relacijskimi matrikami $\mathbf{R}_{12}$, $\mathbf{R}_{13}$ in $\mathbf{R}_{23}$ za opis relacij med vsemi kombinacijami dveh virov in s tremi omejitvenimi matrikami $\mathbf{\Theta}_1$, $\mathbf{\Theta}_2$ in $\mathbf{\Theta}_3$ za opis omejitev znotraj podatkovnih virov. Tabela 5.1 povzema njihovo vsebino.

Matrika	Vsebina
$\mathbf{R}_{12}$	Genske določitve, potrjene z biološkimi eksperimenti
$\mathbf{R}_{23}$	Srednji transkripcijski zapisi mutant glede na pripadnost konceptom – izračun podan s formulo (5.5)
$\mathbf{R}_{13}$	Genomski transkripcijski zapisi mutant
$\mathbf{\Theta}_1$	Proteinske interakcije z ocenami zanesljivosti
$\mathbf{\Theta}_2$	Predhodno znanje o konceptih, izpeljano iz strukture acikličnega grafa, relacije med koncepti *is-a, part-of, regulates, positively-regulates, negatively-regulates*
$\mathbf{\Theta}_3$	Predhodno znanje o transkripcijskih zapisih v 24-urnem ciklu

Tabela 5.1: Vsebina relacijskih in omejitvenih matrik pri napovedovanju genskih določitev v modelnem organizmu *D. discoideum*.

5.2.1 Omejitvene matrike

Ocene zanesljivosti proteinskih interakcij so v normalizirani obliki že shranjene v bazi proteinskih interakcij STRING. Omejitveno matriko $\mathbf{\Theta}_1$ nastavimo z

$$\mathbf{\Theta}_1(k,l) = \mathbf{\Theta}_1(l,k) = \begin{cases} -s & \text{če } (k,l,s) \in \text{PPI} \\ 0 & \text{sicer.} \end{cases} \tag{5.1}$$

Proteinske interakcije izražajo omejitve pozitivnega tipa. Naša hipoteza je namreč, da imata gena z visoko oceno interakcije podobno neznano predstavitev, ki jo izražata s podobnima zapisoma v razcepnih matričnih faktor-

jih. Pozitivne omejitve so v algoritmu nagrade, utežene z razcepnimi faktorji podobnosti, zato jih moramo zapisati z negativnimi števili. Prispevajo namreč k boljši rešitvi in zmanjšujejo rešitveno vrednost kriterijske funkcije. Zaradi predstavitev interakcij z neusmerjenimi grafi je matrika $\mathbf{\Theta}_1$ simetrična.

Genska ontologija je strukturirana v obliki grafa, v katerem so koncepti predstavljeni z vozlišči grafa in relacije med kocepti s povezavami. Relacije med koncepti so natančno opredeljene in kategorizirane. Te v ontologiji GO Ontology obsegajo skupine (i) podtip *is-a*, (ii) del *part-of*, (iii) reguliranje *regulates*, (iv) pozitivno reguliranje *positively-regulates* in (v) negativno reguliranje *negatively-regulates*. Ontologija GO Ontology uporablja le nekaj vrst relacij izmed možnih, ki jih opredeljuje ontologija relacij OBO Relations Ontology. Relacije v GO Ontology upoštevajo format OBO (angl. *open biomedical ontologies format*). Smith in sod. (2005) [58] so predstavili metodologijo za zagotavljanje doslednih in nedvoumnih formalnih definicij relacijskih izrazov, ki se uporabljajo v takih ontologijah. Ta metodologija omogoča povezovanje ontologij iz različnih bioloških domen in definira logično sklepanje o sestavljenih relacijah.

Znanje o strukturi genske ontologije in relacijah med koncepti predstavimo v omejitveni matriki $\mathbf{\Theta}_2$. Vsaki skupini relacij i pripišemo utež pomembnosti α_i. Te uteži določimo z notranjim prečnim preverjanjem ali jih določimo vnaprej. S pravili logičnega sklepanja izračunamo ocene kompozicij večih relacij. Ponazorimo s primerom. Naj velja `A regulates B` in `B part-of C`. Sklepamo na `A regulates C`.

Za dana koncepta k_1 in k_2 nato seštejemo prispevke vseh kompozicij, ki jih utežimo s faktorjem znižanja (angl. *discount factor*) $\gamma \leq 1$

$$g(k_1, k_2) = \sum_{\substack{l=1 \\ \text{is-a}}}^{d} \gamma^l \cdot \alpha_{\text{is-a}} + \sum_{\substack{l=1 \\ \text{part-of}}}^{d} \gamma^l \cdot \alpha_{\text{part-of}} + \sum_{\substack{l=1 \\ \text{regulates}}}^{d} \gamma^l \cdot \alpha_{\text{regulates}}, \qquad (5.2)$$

pri čemer l označuje dolžino poti in d maksimalno dolžino poti, ki jo še upoštevamo. Uporabimo $\gamma = 0.5$ in maksimalno dolžino $d = 4$.

Omejitvena matrika $\mathbf{\Theta}_2$ je določena z

$$\mathbf{\Theta}_2(k_1, k_2) = \begin{cases} -g(k_1, k_2) & \text{če } \exists \text{ pot med } k_1 \text{ in } k_2 \\ 0 & \text{sicer.} \end{cases} \tag{5.3}$$

Graf ontologije je usmerjen, zato matrika $\mathbf{\Theta}_2$ ni simetrična. Kot v primeru proteinskih interakcij so omejitve pozitivnega tipa, od tod negativne vrednosti v (5.3).

Predhodnega znanja o transkripcijskih zapisih $\mathbf{\Theta}_3$ ne uporabljamo. Formalno matriki $\mathbf{\Theta}_3$ ustreza ničelna matrika.

5.2.2 Relacijske matrike

Relacijska matrika genskih določitev $\mathbf{R}_{12}$ je v najbolj preprosti predstavitvi binarna matrika

$$\mathbf{R}_{12}(g, k) = \begin{cases} 1 & \text{če } g \text{ pripada konceptu } k \\ 0 & \text{sicer.} \end{cases} \tag{5.4}$$

V poglavju 3.2 smo predstavili slabosti take predstavitve in predlagali strategije z vzorčenjem znanih genskih določitev. Matrika $\mathbf{R}_{13}$ vsebuje eksperimentalne podatke genskih izrazov. Transkripcijske zapise pred faktorizacijo normaliziramo, tako da vsak zapis delimo z njegovo drugo vektorsko normo.

Za matriko $\mathbf{R}_{23}$ izračunamo srednje transkripcijske zapise genov glede na njihovo pripadnost konceptom

$$\mathbf{R}_{23}(k, :) = \frac{1}{|\mathcal{P}(k)|} \sum_{g \in \mathcal{P}(k)} \text{norm-izr-zapis}(g), \tag{5.5}$$

pri čemer je $\mathcal{P}(k)$ množica vseh genov, ki imajo pripisan koncept k in norm-izr-zapis(g) označuje transkripcijski zapis genskih izrazov gena g.

5.3 Rezultati

Razvili smo pristop združevanja z matrično faktorizacijo, opisan v poglavju 3, in napovedovanje iz matričnih razcepnih faktorjev, opisano v poglavju 4, ter primerjali uspešnost z naključnimi gozdovi z zgodnjo integracijo. Kolikor nam je znano, drugi pristopi za napovedovanje genskih določitev iz heterogenih virov z vmesno integracijo ne obstajajo. Metoda, predlagana v pričujočem delu, je uspešnejša od strategij zgodnje integracije. Ničelno hipotezo o enaki uspešnosti zgodnje integracije z naključnimi gozdovi in faktorizacijskega pristopa zavrnemo pri stopnji značilnosti $\alpha = 0.05$. Vsi predstavljeni rezultati temeljijo na 10-kratnem prečnem preverjanju in ocenjevanju z uteženo F_1 oceno, oboje opisano v razdelku 4.3.

Parametre pristopa z matrično faktorizacijo smo nastavili z notranjim prečnim preverjanjem. Najbolj pomembni so faktorizacijski rangi relacijskih matrik in število ponovitev faktorizacije v ansambelskem pristopu – v razdelku o izboru kandidatov 4.1. Poskusi na umetnih množicah, napovedovanju genskih določitev in analizi bakterijske rezistence kažejo, da zadostuje že tri do devet ponovitev simetrične kazenske matrične tri-faktorizacije z naključnostnimi algoritmi za inicializacijo za značilno povečanje uspešnosti napovedovanja glede na F_1 oceno. Faktorizacijski rang v splošnem ni mogoče vnaprej določiti. Pri simetrični matrični tri-faktorizaciji je število parametrov za faktorizacijski rang enako številu relacijskih matrik, saj vsako relacisko matriko faktoriziramo v produkt treh razcepnih matričnih faktorjev nižjih dimenzij. Za napovedovanje genskih določitev so primerne vrednosti rangov $c_1 \in [38, 43], c_2 \in [35, 45]$ in $c_3 \in [55, 64]$.

5.3.1 Primerjava matričnega pristopa z naključnimi gozdovi

Vsak gen ima lahko več genskih določitev, a vsi ontološki koncepti iz ontologije GO Ontology nimajo pripisanih genov. Razumna predpostavka je, da

za razrede pri učenju upoštevamo le ontološke koncepte s pripisanimi geni. To ni potrebno pri pristopu z matrično faktorizacijo, saj pri napovedovanju genskih določitev iz razcepnih faktorjev sodelujejo tudi drugi podatkovni viri, v katerih morda obstajajo podatki o konceptih brez pripisanh genov. Ovira nastopi, če želimo oceniti značilnost izbranih kandidatov in verjetnost genskih določitev, saj na voljo ni znanih pripisov. Naključni gozdovi za učenje določitev konceptov brez pripisanih genov niso mogoči – namesto naključnih gozdov je potrebno uporabiti nenadzorovane tehnike.

Z namenom zagotovitve enakih pogojev za obe primerjani metodi izpeljemo razrede iz omejene različice genske ontologije GO Slim generic[4]. Ta vsebuje 148 ontoloških konceptov, ki pripadajo področjema bioloških procesov in celičnih komponent. Genske določitve, ki so potrjene z biološkimi eksperimenti, sestavljajo učno množico pozitivnih primerov. Kandidati za negativne primere so tiste ne-določitve med genskimi produkti in koncepti, ki niso podprte z nobenih dokazom – eksperimentalnim, računskim ali pregledom biološke literature. Določitev množice negativnih primerov je odvisna od tehnike učenja PU, opisanega v razdelku 4.2.

Podatkovne vire – genske izraze, proteinske interakcije in znane genske pripise – za naključne gozdove združimo v vektorsko atributno predstavitev. Nato naključne gozdove z 200 odločitvenimi drevesi, omejitvijo vsaj petih primerov v vozliščih in 10-odstotno verjetnostjo preskoka atributa uporabimo v obliki binarnih klasifikatorjev za učenje vsakega razreda (ontološkega koncepta) posebej. Število zgrajenih modelov naključnih gozdov raste linearno z množico ontoloških konceptov. Tako učenje je časovno zahtevno že za omejeno različico genske ontologije in se zdi neizvedljivo za večje število konceptov. V predlaganem pristopu z matrično faktorizacijo učenje za vsak razred posebej ni potrebno.

Napovedovanje s simetrično matrično tri-faktorizacijo z vmesno integracijo dosega značilno boljše rezultate glede na uteženo F_1 oceno. V tabeli 5.2 povzemamo rezultate obeh metod pri napovedovanju genskih določitev za

[4]Omejena različica genske ontologije je dostopna na `http://www.geneontology.org/GO.slims.shtml`.

izbrano množico genov. Izbrali smo gene s največ pripisi, upoštevajoč vse utemeljitve pripisov iz omejene različice genske ontologije.

Metoda	ocena F_1 za 100 genov	ocena F_1 za 1000 genov
Napovedovanje s tri-SPMF	0.7782	0.8009
Naključni gozdovi	0.7431	0.7527

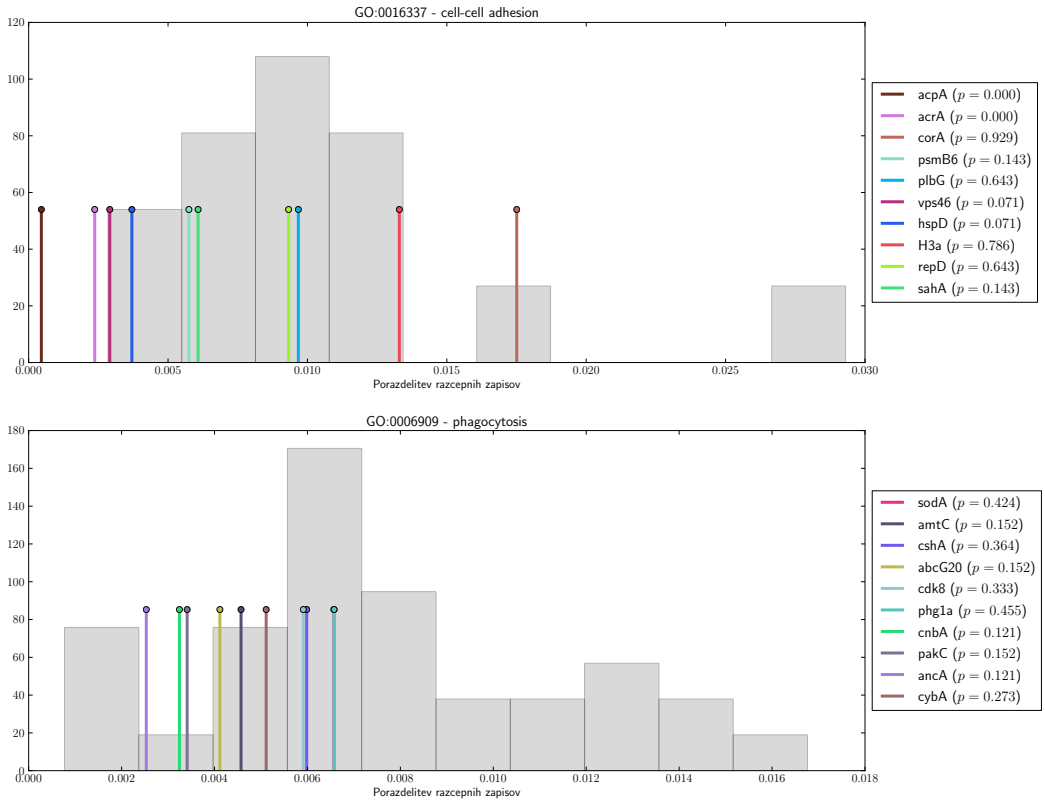
Tabela 5.2: Rezultati 10-kratnega prečnega preverjanja pristopa z matrično faktorizacijo in naključnih gozdov pri napovedovanju genskih določitev za 100 in 1000 genov *D. discoideum* z največ genskimi določitvami. Podane so F_1 ocene, utežene s pogostostjo razredov.

Biologi s sodelujoče institucije so pripravili seznam konceptov, ki so za modelni organizem *D. discoideum* v trenutnih raziskavah najbolj pomembni. Ocene F_1 po konceptih so prikazane v tabeli 5.3. Matrični pristop se izkaže za uspešnejšega pri skoraj vseh konceptih.

Po izboru kandidatov je smiselno oceniti verjetnosti genskih pripisov iz razcepnih matričnih faktorjev, kot je opisano v razdelku 4.2. Značilnost napovedanega pripisa ocenimo z uvrstitvijo njegove ocene v matriki $\widehat{\mathbf{R}}_{12}$ v porazdelitev ocen znanih genskih pripisov za dani koncept. Na sliki 5.2 so prikazane uvrstitve desetih izbranih genov za koncepta celične adhezije GO:0016337 in fagocitoze GO:0006909.

Koncept	Opis koncepta	P	D	$F_1^{(MF)}$	$F_1^{(RF)}$
GO:0007190	activation of adenylate cyclase activity	11	BP	0.8300	0.7762
GO:0006935	chemotaxis	58	BP	0.9825	0.8562
GO:0043327	chemotaxis to cAMP	21	BP	0.9756	0.9210
GO:0006909	phagocytosis	33	BP	0.9524	0.8813
GO:0003700	DNA binding transcription factor activity	79	MF	0.9467	0.7918
GO:0009617	response to bacterium	51	BP	0.8913	0.8641
GO:0016337	cell-cell adhesion	14	BP	0.8800	0.7978
GO:0003779	actin binding	43	MF	0.6562	0.6610
GO:0003796	lysozyme activity	4	MF	0.7726	0.7533

Tabela 5.3: Rezultati 10-kratnega prečnega preverjanja pristopa z matrično faktorizacijo in naključnih gozdov pri napovedovanju genskih določitev *D. discoideum* za izbrane ontološke koncepte. Podpora v stolpcu P označuje število znanih pripisov genov danemu konceptu. Področje v stolpcu D se nanaša na izpeljane koncepte; BP - biološki procesi, MF - molekularne funkcije. Ocena F_1 iz naključnih gozdov je označena s $F_1^{(RF)}$, iz matričnega pristopa s tri-SPMF pa s $F_1^{(MF)}$.

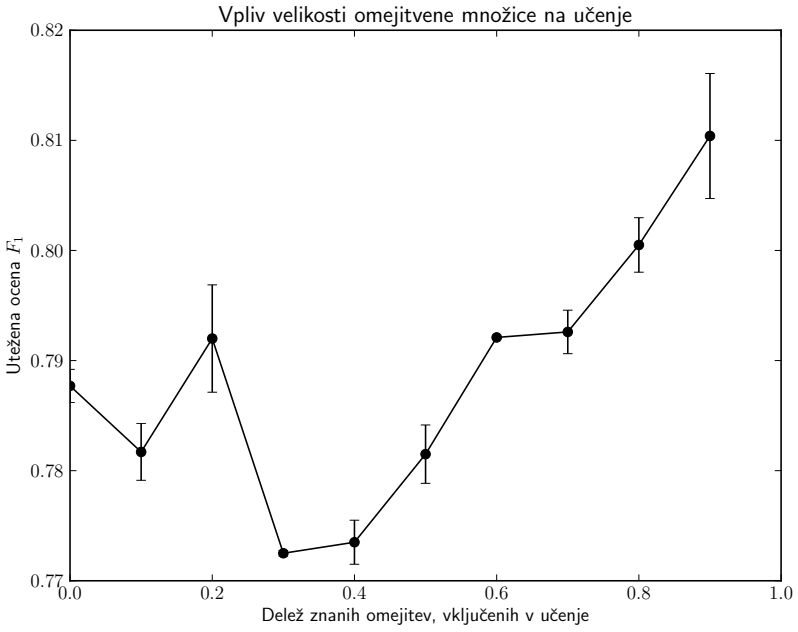


Slika 5.2: Ocenjevanje značilnosti kandidatnih genov za pripise iz razcepnih zapisov za koncepta celične adhezije GO:0016337 in fagocitoze GO:0006909. Oba koncepta pripadata področju bioloških procesov. Kandidatovo oceno v matriki $\widehat{\mathbf{R}}_{12}$ (obarvane navpične črte) uvrstimo v porazdelitev (sivi stolpci) ocen znanih določitev za obravnavani koncept.

5.3.2 Vpliv omejitev na napovedi matričnega pristopa

Pri učenju s simetrično kazensko matrično tri-faktorizacijo smo zgradili omejitveni matriki $\mathbf{\Theta}_1$ in $\mathbf{\Theta}_2$, upoštevajoč omejitve med pari genov in pari konceptov. Uporaba omejitev je smiselna le, če te prispevajo k povečanju točnosti napovednega modela, sicer kršimo princip najkrajšega opisa (angl. *minimum description length principle*). Ta pravi, da je najpreprostejša razlaga tudi najbolj zanesljiva.

Korist omejitev ocenimo tako, da gradnjo napovednega modela večkrat ponovimo in spreminjamo delež uporabljenih omejitev. Na mesto odstranjenih omejitev postavimo ničelne vrednosti, kar ustreza interpretaciji omejitvenih matrik – ničelne omejive ne vplivajo na vrednost kriterijske funkcije ali usmerjanje posodabljanja razcepnih faktorjev. Spreminjanje ocene F_1 z dodajanjem omejitev pri napovedovanju genskih določitev na sliki 5.3 kaže, da te prispevajo k povečevanju točnosti napovednega modela.

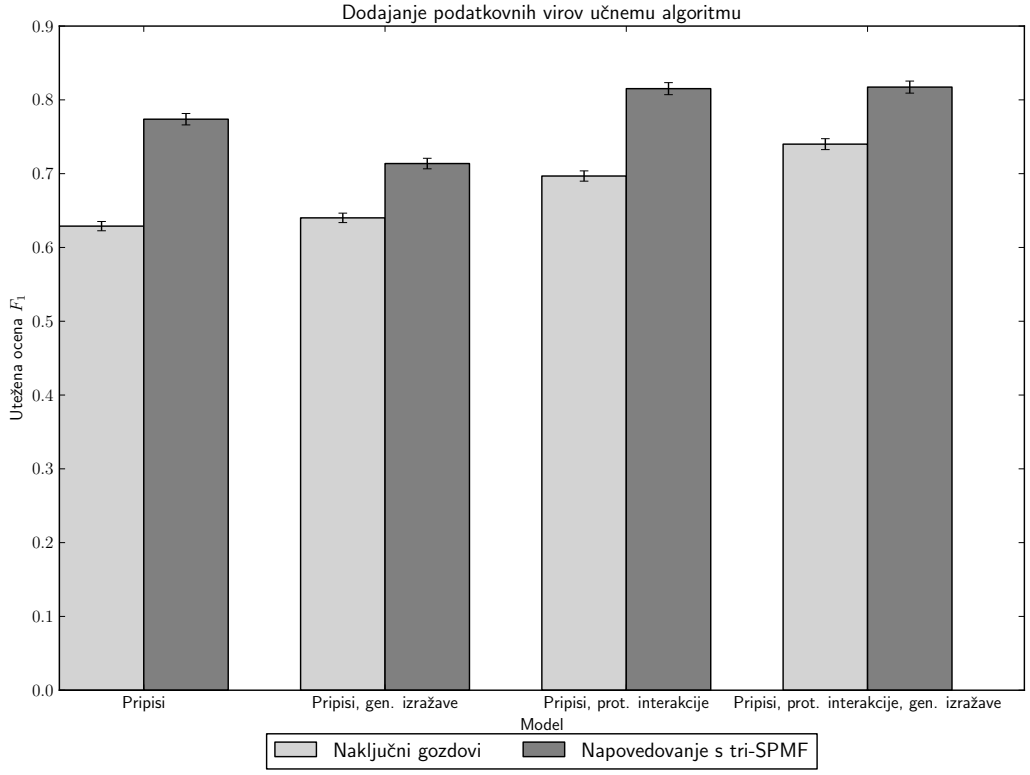


Slika 5.3: Vpliv spreminjanja velikosti omejitvene množice na uspešnost učenja z matričnim pristopom.

5.3.3 Vpliv podatkovnih virov na napovedi matričnega pristopa

Podobno kot smo si ogledali, kako velikost omejitvene množice vpliva na učenje, nas zanima, ali tudi dodajanje podatkovnih virov v resnici izboljša napovedni model. V splošnem je odgovor pritrdilen. V številnih primerih [54, 68, 33] zgodnje in pozne integracije se je združevanje podatkovnih virov izkazalo za koristno. Pričakovali bi, da vmesna integracija s hkratno obravnavo podatkovnih virov dosega boljše rezultate od zgodnje in pozne integracije.

Matrični pristop, ki temelji le na znanih genskih določitvah, namesto simetrične kazenske tri-faktorizacije tri-SPMF uporablja kazensko matrično faktorizacijo PMF. Če znanim genskim določitvam pridružimo še en vir, bodisi genske izraze bodisi proteinske interakcije, razcepne faktorje pridobimo iz kazenske tri-faktorizacije tri-PMF. Za večje število podatkovnih virov se vsakokrat uporabi simetrična kazenska tri-faktorizacija tri-SPMF. V naključnih gozdovih dodajanje virov le spreminja dolžino atributne predstavitve genov. Na sliki 5.4 vidimo, da matrični pristop dosega boljšo oceno od naključnih gozdov ne glede na nabor virov, kar izhaja iz prednosti vmesne strategije združevanja. Nadaljnje, dodajanje podatkovnih virov izboljša napovedi obeh pristopov, a je v matričnem modelu izboljšava večja.



Slika 5.4: Vpliv dodajanja podatkovnih virov na učenje z matričnim pristopom.

Poglavje 6

Prioritizacija genov bakterijske rezistence amebe *D. discoideum*

Klasičnim antibakterijskih zdravilom se zmanjšuje učinkovitost zaradi vse večje razširjenosti bakterij, odpornih na antibiotike. Potreben je razvoj novih metod zdravljenja bakterijskih okužb. Razpoznava bakterijske rezistence genov in genskih poti *D. discoideum* je začetek raziskav na tem področju. Ameba *D. discoideum* je namreč bakterijski plenilec in včasih tudi žrtev bakterijskih okužb, zato je razpoznava genov in genskih poti, ki sodelujejo v njeni rezistenci, zelo pomembna.

Odziv amebe v okolju z bakterijami je pomemben za okužbe pri ljudeh, ker se je le ta verjetno razvil iz poti, ki so jih uporabljali primitivni evkarionti za obrambo pred bakterijami. Nekaj eksperimentalnih testov je že potrdilo povezanost imunskega odziva *D. discoideum* s tistim pri ljudeh, kar obeta možen razvoj novih antibiotičnih zdravil, a so raziskave še v povojih.

Trenutno je znanih le nekaj genov in genskih poti, ki so v *D. discoideum* odgovorni za bakterijsko razpoznavo in rezistenco. Nedavno so biologi z genskim pregledom mutant, ki rastejo bodisi v okolju Gram pozitivnih bodisi v okolju Gram negativnih bakterij, a ne preživijo v obeh okoljih, določili

prve kandidatne gene, teh je zgolj sedem. Naš cilj je napoved novih kandidatov z uporabo računskega pristopa, ki ga predlagamo v tem delu. Pri tem združujemo heterogene podatkovne vire, ki so znani za *D. discoideum* – eksperimentalni podatki o genskih izrazih, genske določitve in proteinske interakcije.

Podatki, s katerimi se ukvarjamo, so precej redki. *D. discoideum* je modelni organizem, v katerem ima le približno tretjina genov pripisane določitve, ki so potrjene z eksperimenti. Večina informacij o proteinskih interakcijah izhaja iz homolognih genov v drugih organizmih. Dodatna ovira je majhnost množice mutant z znanim odzivom preživetja v Gram pozitivnem in Gram negativnem okolju. Ta je navedena v dodatku B in vsebuje 45 mutant, izmed teh 28 mutant pripada razredu, ki počasi raste v Gram pozitivnem in Gram negativnem okolju. Vse mutante v učnem naboru so razvrščene v šest razredov. Za analizo bakterijske rezistence sta najbolj pomembna razreda mutant, občutljivih na eno od obeh Gram okolij – v strojnem učenju je problem znan kot cenovno občutljiva klasifikacija. Iz primerov mutant, občutljivih na eno izmed obeh Gram okolij lahko domnevamo, da so spremenjeni geni mutant vpleteni v bakterijsko rezistenco *D. discoideum*. Žal je teh mutant le sedem, tri so občutljive na Gram pozitivno okolje in štiri na Gram negativno okolje, kot je prikazano v dodatku B. Izredno malo znanih mutant z odzivom, občutljivim na bodisi Gram pozitivno bodisi Gram negativno okolje, je največja ovira pri izgradnji napovednega modela.

6.1 Napovedovanje iz razcepnih matričnih faktorjev

Na voljo imamo enake podatkovne vire kot v gradnji modela genskih določitev (poglavje 5), zato uporabimo enako matrično predstavitev relacijskih in omejitvenih matrik. Spremeni se le napovedovanje iz razcepnih faktorjev, ki jih vrne algoritem tri-SPMF. Uvrstitev primerov v razrede rasti mutant *D. discoideum* ni neposredno vsebovana v matrični predstavitvi. Razrede rasti zato

računamo iz razcepnih zapisov genskih izrazov v $\mathbf{G}_3$ in aproksimirane relacijske matrike $\widehat{\mathbf{R}}_{13}$ s pravilom "krivde zaradi poveznosti," tako da določimo minimalno razdaljo razcepnih zapisov med testnim primerom t in srednjimi zapisi predstavnikov razredov rasti

$$\text{class}(t) = \underset{C \in \{\text{SG, FG, NG, NGA, GPD, GND}\}}{\arg\min} \frac{1}{|C|} \sum_{x \in C} d(\mathbf{G}_3(t,:), \mathbf{G}_3(x,:)). \quad (6.1)$$

Značilnost dobljenih razdalij ocenimo s permutacijskim testom, v katerem permutiramo oznake razredov.

6.2 Rezultati in razprava

Alternativna metoda predlaganemu pristopu z matrično faktorizacijo je nadzorovano učenje z zgodnjo integracijo. Podatkovne vire predstavimo z značilkami in vsak učni primer opišemo z vektorjem vrednosti izpeljanih značilk kot so transkripcijski zapisi genskih izrazov mutant v raznih okoljih, genske določitve, proteinske interakcije in druge informativne lastnosti primerov. Tak način kodiranja učnih podatkov je v bioinformatiki zelo pogost. Atributna predstavitev ima slabosti, saj je izrazno šibkejša od predstavitev, ki kodirajo strukture, in je običajno shranjena v obliki velikih in redkih atributnih matrik.

Napovedovanje s simetrično kazensko matrično tri-faktorizacijo smo primerjali z naključnimi gozdovi s številom odločitvenih dreves na intervalu [100, 500], zahtevo po vsaj petih primerih v vozliščih za delitev vozlišča in verjetnostjo preskočitve atributa 0.1. Oba modela sta imela na voljo popolnoma enak nabor vhodnih podatkov. Tega smo za naključne gozdove predstavili v atributni obliki. Učna množica znanih odzivov mutant *D. discoideum* na okolje Gram pozitivnih in Gram negativnih bakterij vsebuje le 45 označenih primerov. Izmed teh spadajo trije primeri v razred občutljivosti na Gram pozitivno okolje in štirje primeri v razred občutljivosti na Gram negativno

okolje. Zaradi neuravnoteženosti učnega nabora, večinski razred zavzema 62 % učne množice, smo primere manjšinskih razredov vzorčili navzgor. Za vsak manjšinski razred smo enakomerno naključno in z vračanjem izbrali pripadajoče primere ter jih dodali v učno množico, da je ta postala povsem uravnotežena. To pomeni, da je učni nabor po vzorčenju navzgor vseboval ponovitve nekaterih primerov. Pri testiranju smo vsakokrat izločili en primer v testno množico in zgradili napovedna modela s simetrično kazensko matrično-tri faktorizacijo ter naključnimi gozdovi na preostalem delu množice.

Model s simetrično kazensko matrično tri-faktorizacijo tri-SPMF se je izkazal za uspešnejšega. Metodo naključnih gozdov smo razširili z izborom in gradnjo novih značilk z notranjim prečnim preverjanjem. Značilke gradimo na dva načina. Prvi vključuje računanje medsebojnega prispevka in sinergije [3] $Syn(A_1, A_2; C) = I(A_1, A_2; C) - I(A_1, C) - I(A_2, C)$ med pari atributov A_1 in A_2 ter razredom C. Medsebojno informacijo $I(X; Y)$ izračunamo z entropijo $I(X; Y) = H(X) - H(X|Y)$. V konstruktivni indukciji so zelo zanimive značilke, ki same po sebi niso informativne, a dosegajo v kombinaciji z drugimi značilkami visoko sinergijo.

Drugi način gradnje značilk izkorišča uporabo matričnih faktorizacij za zmanjšanje dimenzionalnosti podatkov. Zapis učnega primera iz razcepnih faktorjev simetrične kazenske matrične tri-faktorizacije smo vključili v atributni vektor in nad novo predstavitvijo uporabili naključne gozdove. Napovedovanji z metodo tri-SPMF in naključnimi gozdovi, katerim sta dodana gradnja in izbor značilk, dosegata primerljive rezultate v smislu dosežene utežene ocene AUC in klasifikacijske točnosti, prikazane v tabeli 6.1. Pomembna je razpoznava genov, ki sodelujejo v bakterijski rezistenci amebe *D. discoideum*. To so geni, katerih mutante rastejo bodisi v okolju Gram pozitivnih bodisi okolju Gram negativnih bakterij.

Pri analizi bakterijske rezistence *D. discoideum* nas zanima ne le čim točnejša uvrstitev primera v razred, temveč tudi izbor najbolj obetavnih kandidatov, ki bi bili primerni za načrtovanje dragih bioloških eksperimen-

Metoda	ocena AUC za Gram neg.	ocena AUC za Gram poz.
Večinski razred	0.5000	0.5000
Naključni gozdovi	0.6665	0.8312
Naključni gozdovi s konstruktivno indukcijo	0.7506	0.9043
Napovedovanje s tri-SPMF	0.7599	0.9643

Tabela 6.1: Ocena uspešnosti napovednih modelov matričnega pristopa in naključnih gozodov za oba razreda Gram okolij z oceno AUC. Napovedovanje preživetja mutant *D. discoideum* v Gram okoljih je pomembno za analizo bakterijske rezistence.

tov. V poglavju 4 o napovedovanju iz matričnih razcepnih faktorjev smo predlagali več računskih pravil za izbor kandidatov (v 4.1) in prioritizacijo njihovih ocen (v 4.2). V naključnih gozdovih smo kandidate razvrstili glede na verjetnost pripadnosti razredu. Napovedovanji iz razcepnih faktorjev in naključnih gozdov z gradnjo ter izborom značilk proizvedeta podobni množici desetih najboljših kandidatov za razreda občutljivosti v Gram pozitivnem in Gram negativnem okolju, v katerih se ujema 87.2 % genov. Seznama za oba Gram razreda sta izpisana v tabeli 6.2. Delež ujemanja med seznamoma zgornjih tridesetih kandidatov je 0.736. Opravljeni so bili biološki eksperimenti, v katerih se je opazovalo preživetje mutant *D. discoideum* izbranih genov iz seznama 6.2 v obeh Gram okoljih. Trenutni rezultati bioloških raziskav so zelo spodbudni, saj vse testirane mutante v laboratoriju kažejo spremembe rasti bodisi v okolju Gram pozitivnih bakterij bodisi v okolju Gram negativnih bakterij kot viru hrane za njihovo preživetje. Težavnost klasifikacijskega problema bakterijske rezistence *D. discoideum* – za več kot 13000 testnih genov *D. discoideum* so dani le trije učni primeri Gram pozitivne in štirje učni primeri Gram negativne občutljivosti – in biološki testi kažejo na obetavnost predlagane metode gradnje napovednih modelov z matrično faktorizacijo iz heterogenih podatkovnih virov.

[1]Oznake mutant so iz podatkovne baze DictyBase, dostopne na naslovu `http://dictybase.org`.

Tabela 6.2: Seznama desetih najbolje ocenjenih genov *D. discoideum* za razreda občutljivosti v Gram pozitivnem in Gram negativnem okolju, kot ju predlaga napovedovanje iz razcepnih faktorjev simetrične kazenske matrične tri-faktorizacije. Množici desetih najbolje ocenjenih genov za oba razreda se v pristopu s faktorizacijo in naključnih gozdovih z gradnjo ter izborom značilk ujemata v 87.2 %.

(a) Občutljivost v okolju Gram pozitivnem okolju

Mutanta[1]	Ime gena
DDB_G0286117	*rps10*
DDB_G0273063	*dscA-1*
DDB_G0291568	*dnaja1*
DDB_G0284613	*D7*
DDB_G0291123	*glpD*
DDB_G0287399	
DDB_G0281861	
DDB_G0293194	*abcD2*
DDB_G0273149	
DDB_G0279291	

(b) Občutljivost v okolju Gram negativnem okolju

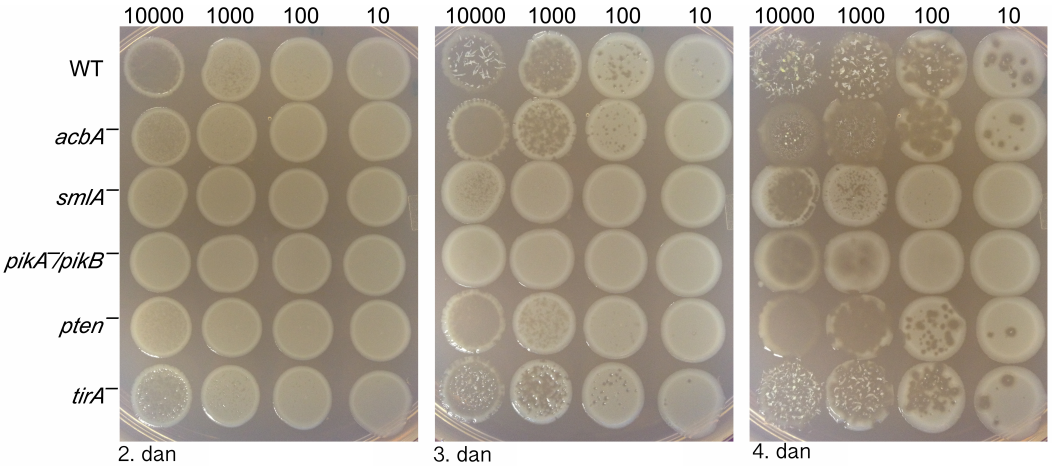
Mutanta[1]	Ime gena
DDB_G0269100	*abpC*
DDB_G0277355	
DDB_G0273105	
DDB_G0273175	*cf50-1*
DDB_G0287585	*rgaA*
DDB_G0273095	
DDB_G0279291	
DDB_G0275057	
DDB_G0290575	
DDB_G0291123	*phdA*

6.3 Biološki eksperimenti

Z novo metodo smo napovedali več genov amebe *D. discoideum*, ki so bili do sedaj neznani v analizi bakterijske rezistence, in jih prednostno razvrstili od najbolj do najmanj verjetnih z ocenjevanjem značilnosti njihovih ocen. Ta postopek je predlagan in opisan v poglavju 3 in 4.

Na osnovi naših napovedi najbolje ocenjenih genov so v laboratorijih prof. dr. Gada Shaulskyja in prof. dr. Adama Kuspe na Baylor College of Medicine v Houstonu v ZDA opravili biološke eksperimente, s katerimi so potrdili napovedane gene. Razpoznava genov z doslej neznano vlogo v bakterijski rezistenci je izredno pomemben prispevek pri snovanju alternativnih metod klasičnega antibakterijskega zdravljenja.

Z eksperimenti so ocenili rast sevov z izničenimi geni *D. discoideum* v okolju Gram negativne bakterije *K. pneumoniae*, tako da so vzgojili ko-kulturo celic amebe *D. discoideum* in bakterije *K. pneumoniae*. Celice amebe in bakterije so rastle na agarnih ploščah v vlažni komori pri 22 °C, pri čemer so bile amebe nacepljene na gojišče in serijsko razredčene z bakterijo *K. pneumoniae*. Rezultati bakterijske rasti so vidni v pojavu debele neprozorne plasti na vseh sevih 2. dne z 10 amebami na posamezni točki. Amebe divjega tipa porabijo bakterije na način, ki je uvodoma viden kot majhne plake (sev WT v drugem dnevu s 1000 celicami amebe na sliki 6.1), nato napreduje v očiščeno plast (sev WT v tretjem dnevu s 1000 celicami amebe na sliki 6.1) ter se konča z razvojem večceličnih struktur (sev WT v četrtem dnevu s 1000 in 10000 amebami na sliki 6.1). Za primerjavo, celice mutante $tirA^-$ kažejo zmanjšano rast v okolju *K. pneumoniae*, kot so poročali Chen in sod. (2007) v [12]. Sevi z motnjami v genih, katere smo napovedali z novo metodo, so pokazali napake v rasti. Mutante $smlA^-$, $pikA^-/\,pikB^-$ in $pten^-$ izražajo močno zmanjšano rast na sliki 6.1. Celice mutante $acbA^-$ prav tako kažejo prizadeto rast v primerjavi z divjim tipom, vendar v manjši meri kot v primeru drugih mutant.



Slika 6.1: Ko-kulture celic *D. discoideum* in bakterij so bile narejene z mešanjem aksenično vzgojenih celic *D. discoideum* z bakterijami *K. pneumoniae* (OD_{600} = 1.0). Celice amebe so bile razredčene z bakterijami in nacepljene na gojišča z 10000, 1000, 100 in 10 amebami na posamezni točki. Agarne plošče so bile inkubirane v vlažni komori s posnetimi slikami v drugem, tretjem in četrtem dnevu. Imena sevov so navedena na levi, pri čemer WT označuje sev divjega tipa. Bele neprozorne lise označujejo rast bakterij *K. pneumoniae*. Rast ameb se kaže v tvorjenju čistih plak znotraj neprozornih delov. Po izčrpanju bakterijske kulture se pojavi proces stradanja celic *D. discoideum* in razvoj večceličnih struktur (tretji in četrti dan).

6.3.1 Materiali in metode

Sevi *D. discoideum* so bili pridobljeni v centru DictyBase Stock in vzdrževani aksenično v HL-5 pri temperaturi 22 °C. Bakterija *K. pneumoniae* je bila shranjena v mešanici SM pri 22 °C. Za oceno zmožnosti rasti celic *D. discoideum* v bakterijskem okolju so bile amebe izbrane iz akseničnih kultur med logaritemsko rastjo in izprane enkrat s Sorensenovim pufrom. Celice *D. discoideum* so bile serijsko razredčene z bakterijskimi celicami (OD_{600} = 1.0) in nacepljene na agarne plošče SM. Plošče so bile inkubirane v vlažni komori pri 22 °C in slike plošč so bile posnete vsakih 24 ur.

Poglavje 7

Sklep

Predstavili smo učinkovit pristop za gradnjo napovednih modelov iz heterogenih virov. Hitro večanje količine in raznolikosti podatkov postavlja izziv za razvoj metod, ki vire inteligentno združujejo. Zgodnja in pozna strategija združevanja podatkov, ki kombinirata vhodne podatke ali napovedi, sta se v sorodnih, že objavljenih študijah, izkazali za koristni pri izboljšanju točnosti napovednih modelov. Toda vmesna integracija, ki sočasno obravnava vire v učnem algoritmu, omogoča ocenjevanje prispevkov različnih virov in vključevanje znanih relacij v učni proces, zato je ta pristop skoraj vedno zaželjen, a je težji za izvedbo. Naša predpostavka je, da je pravi vir informacij dejansko kombinacija podatkov iz različnih podatkovnih zbirk in da je zanesljive ter robustne napovedi mogoče doseči le z integracijo heterogenih virov.

Predstavljeni pristop temelji na simetrični kazenski matrični tri-faktorizaciji. Podatkovne vire predstavimo z dvema skupinama matrik. Relacijske matrike so simetrične in hranijo preslikavo med poljubnima viroma. Vsak vir ima poleg opisov relacij z drugimi viri pridruženo omejitveno matriko, ki predstavlja znane nagrade in kazni pri razvrščanju zapisov primerov v matričnih razcepnih faktorjih. Omejitve izražajo naše apriorno znanje o podatkovnem viru.

V splošnem so te matrike zelo velike, zato jih organiziramo v matematični

konstrukt s ciljem zmanjšanja dimenzionalnosti, tako da najdemo dobro nadomestno predstavitev z manjšim številom matričnih faktorjev. Ti razkrivajo skrite vzorce v podatkih, odstranjujejo šum in manj pomembne variacijske načine podatkov. Iz matričnih faktorjev, ki jih vrne faktorizacijski algoritem, zgradimo seznam kandidatov, ki predstavljajo našo hipotezo o neznanih relacijah med viri. Ročno pregledovanje in iskanje najbolj obetavnih kandidatov v velikih seznamih je naporno in zamudno delo. Zato za zanimiv problem prioritizacije in rangiranja kandidatov v delu predstavimo več računskih pravil, ki upoštevajo zapise v razcepnih faktorjih. Kakovost kandidatov ocenimo z uvrstitivijo njihovih razcepnih zapisov v porazdelitev zapisov znanih relacij med viri in z izračunom značilnosti uvrstitve.

Standardni postopki matričnih faktorizacij so uporabni za izražanje odnosov le med dvema podatkovnima tipoma. To ne velja za pristop, predstavljen v tem delu, ki lahko vključi vse vrste podatkov preko omejitev in relacij znotraj skupnega matematičnega okvira. Glavna novost našega predloga je celostno reševanje naloge vmesne integracije podatkovnih virov v gradnji napovednih modelov.

Poskusi na umetnih in realnih podatkovnih virih dajejo zelo spodbudne rezultate. Metodo smo uporabili za napovedovanje genskih določitev amebe *D. discoideum*. Simetrična kazenska matrična tri-faktorizacija hkrati obravnava poljubno število podatkovnih virov. V napovedovanje genskih določitev smo vključili zapise genskih izrazov, omrežja proteinskih interakcij in znane genske pripise. V vseh testih je predlagana tehnika za gradnjo napovednih modelov uspešnejša od učnih algoritmov z zgodnjim in poznim združevanjem. Primerljivi pristopi h gradnji napovednih modelov iz heterogenih virov z vmesno strategijo združevanja nam niso poznani. Uporaba matričnih faktorizacij v funkcijski genomiki je uspešna, a redka. Faktorizacije, ki kombinirajo dodatno znanje, namreč z nekaj izjemami uporabe v zelo specifičnih okoljih ne obstajajo.

Prispevek pričujočega dela pa ni samo metodološki. V študiji smo namreč na problemih iz molekularne biologije pokazali tudi na njegovo uporabno

vrednost. Tehniko s simetrično kazensko matrično tri-faktorizacijo smo zelo uspešno preizkusili v napovedovanju genov, ki sodelujejo pri bakterijski rezistenci *D. discoideum*. Biološki eksperimenti, opravljeni v laboratoriju priznane institucije Baylor College of Medicine iz Houstona, ZDA, so potrdili občutljivost napovedanih genov na bakterije v okolju. Ker je bila do sedaj znana le peščica genov, pomembnih za bakterijsko rezistenco te amebe, so naše potrjene računsko pridobljene hipoteze pomemben prispevek k napredku bioloških znanosti. Odziv amebe na bakterije je pomemben za raziskovanje okužb pri ljudeh, saj eksperimentalni testi kažejo obstoj povezav med amebinim imunskim sistemom in imunskim sistemom ljudi, vendar so raziskave na tem področju še v povojih. Razširjenost bakterij, odpornih na antibiotike, zmanjšuje učinkovitost klasičnih antibakterijskih zdravil, zato je določitev novih genov ter genskih poti, vpletenih v amebino bakterijsko rezistenco, prispevek k razvoju novih metod zdravljenja.

Podatkovni viri, s katerimi se ukvarjamo, so pogosto precej redki. Za učinkovito računanje morajo biti temu prilagojeni razviti algoritimi. Razvili smo javno dostopno programsko knjižnico, imenovano Nimfa[1] [71]. Ta podpira računanje z redkimi in gostimi matrikami ter vsebuje implementacije številnih tehnik matrične faktorizacije, mere za ocenjevanje kakovosti matričnih faktorjev, primere uporabe in je dobro dokumentirana.

Predlagani pristop odpira številne možnosti za razširitve in izboljšave na področju izgradnje modela ter uporabe:

- **Nadzorovana matrična faktorizacija.** Trenutno znane faktorizacijske tehnike sodijo v nenadzorovano ali delno nadzorovano učenje. Razvoj takih metod zahteva povsem nov pogled na faktorizacijske modele [5, 51].

- **Rangiranje seznama kandidatov.** To je precej raziskano področje, a v bioinformatični skupnosti iskanje postopka, ki ne bi bil vezan na konkretno uporabo, še vedno predstavlja izziv.

[1]Programska knjižnica Nimfa je dostopna na naslovu `http://nimfa.biolab.si`.

- **Matrično dopolnjevanje** išče ustrezne predstavitve za mesta z neznanimi vrednostmi v vhodni podatkovni matriki. Hitri hevristični pristopi in vzorčenje domene pogosto ne odpravljajo problema. Nedavno je bil predstavljen nov pristop za neznane vrednosti, ki za matrično dopolnjevanje rešuje konveksni optimizacijski program [11].

- **Izgradnja omrežij iz razcepnih faktorjev.** Odkriti skušamo organizacijo povezanosti najbolj obetavnih kandidatov, s čimer bi v bioinformatični uporabi sklepali o genskih poteh. Genska regulatorna omrežja lahko predstavimo z nizom omejitev preko genskih interakcij. Zato bi bili zanimivo proučiti obnašanje predlaganega pristopa v analizi genskih interakcij.

Te razširitve bomo raziskali v nadaljnjem raziskovalnem delu. V času pisanja pričujočega poročila pa že potekajo priprave na nadaljnje študije uporabe predlaganih metod na področju sistemske biologije kvasovke in študije epistaze oziroma rekonstrukcije regulatornih mrež tega organizma.

Literatura

[1] R. Albright, J. Cox, D. Duling, A. N. Langville, in C. D Meyer. Algorithms, initializations, and convergence for the nonnegative matrix factorization. Tehnično poročilo, Department of Mathematics, North Carolina State University, 2006.

[2] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, in David J. Lipman. A basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[3] Dimitris Anastassiou. Computational analysis of the synergy among multiple interacting genes. *Molecular Systems biology*, 3(83), 2007.

[4] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, in G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29, 2000.

[5] Elnaz Barshan, Ali Ghodsi, Zohreh Azimifar, in Mansoor Zolghadri Jahromi. Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recogn.*, 44(7):1357–1371, Julij 2011.

[6] Riccardo Bellazzi in Blaž Zupan. Towards knowledge-based gene expression data mining. *Journal of Biomedical Informatics*, 40(6):787–802, 2007.

[7] Asa Ben-Hur in William Stafford Noble. Kernel methods for predicting protein–protein interactions. *Bioinformatics*, 21(1):38–46, 2005.

[8] Karsten M. Borgwardt, Cheng Soon Ong, Stefan Schönauer, S. V. N. Vishwanathan, Alex J. Smola, in Hans-Peter Kriegel. Protein function prediction via graph kernels. *Bioinformatics*, 21(1):47–56, 2005.

[9] Christos Boutsidis in Efstratios Gallopoulos. SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recogn.*, 41(4):1350–1362, 2008.

[10] Jean-Philippe Brunet, Pablo Tamayo, Todd R. Golub, in Jill P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *PNAS*, 101(12):4164–4169, 2004.

[11] Emmanuel Candès in Benjamin Recht. Exact matrix completion via convex optimization. *Commun. ACM*, 55(6):111–119, 2012.

[12] Guokai Chen, Olga Zhuchenko, in Adam Kuspa. Immune-like phagocyte activity in the social amoeba. *Science*, 317(5838):678–81, 2007.

[13] Yanhua Chen, Lijun Wang, in Ming Dong. Non-negative matrix factorization for semisupervised heterogeneous data coclustering. *IEEE Trans. Knowl. Data Eng.*, 22(10):1459–1474, 2010.

[14] Vladimir Cherkassky, Feng Cai, in Lichen Liang. Predictive learning with sparse heterogeneous data. V *Proceedings of the 2009 international joint conference on Neural Networks*, IJCNN'09, strani 3155–3162, Piscataway, NJ, USA, 2009. IEEE Press.

[15] Michael Collins, Sanjoy Dasgupta, in Robert E. Schapire. A generalization of principal component analysis to the exponential family. V *Advances in Neural Information Processing Systems*. MIT Press, 2001.

[16] Minghua Deng, Ting Chen, in Fengzhu Sun. An integrated probabilistic model for functional prediction of proteins. V *Proceedings of the seventh annual international conference on Research in computational molecular biology*, RECOMB '03, strani 95–103, New York, NY, USA, 2003.

[17] Marie desJardins, Peter D. Karp, Markus Krummenacker, Thomas J. Lee, in Christos A. Ouzounis. Prediction of enzyme classification from protein sequence without the use of sequence similarity. V *Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology*, strani 92–99. AAAI Press, 1997.

[18] Damien Devos in Alfonso Valencia. Practical limits of function prediction. *Proteins*, 41:98–107, 2000.

[19] Inderjit S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. V *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01, strani 269–274, New York, NY, USA, 2001.

[20] Chris Ding, Tao Li, Wei Peng, in Haesun Park. Orthogonal nonnegative matrix t-factorizations for clustering. V *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, strani 126–135, New York, NY, USA, 2006.

[21] Paul Dobson in Andrew Doig. Predicting enzyme class from protein structure without alignments. *J. Mol. Biol.*, 345(1):187–199, 2005.

[22] Delbert Dueck, Quaid D. Morris, in Brendan J. Frey. Multi-way clustering of microarray data using probabilistic sparse matrix factorization. *Bioinformatics*, 21(1):144–151, 2005.

[23] Iddo Friedberg. Automated protein function prediction - the genomic challenge. *Briefings in Bioinformatics*, 7(3):225–242, 2006.

[24] Geoffrey J. Gordon. Generalized^2 linear^2 models. V Suzanna Becker, Sebastian Thrun, in Klaus Obermayer, editors, *NIPS*, strani 577–584. MIT Press, 2002.

[25] Derek Greene in Pádraig Cunningham. A matrix factorization approach for integrating multiple data views. V *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part I*, ECML PKDD '09, strani 423–438, Berlin, Heidelberg, 2009. Springer-Verlag.

[26] Jian Guo, Elizaveta Levina, George Michailidis, in Ji Zhu. Joint estimation of multiple graphical models. *Biometrika*, 2010.

[27] Isabelle Guyon, Steve Gunn, Masoud Nikravesh, in Lotfi A. Zadeh. *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[28] David Haussler. Convolution kernels on discrete structures. Tehnično poročilo, UC Santa Cruz, 1999.

[29] Patrik O. Hoyer. Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res.*, 5:1457–1469, 2004.

[30] L. J. Jensen, R. Gupta, N. Blom, D. Devos, J. Tamames, C. Kesmir, H. Nielsen, H. H. Stærfeldt, K. Rapacki, C. Workman, C. A. F. Andersen, S. Knudsen, A. Krogh, A. Valencia, in S. Brunak. Prediction of human protein function from post-translational modifications and localization features. *J Mol Biol*, 319:1257–1265, 2002.

[31] Philip M. Kim in Bruce Tidor. Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Res.*, 13:1706–1718, 2003.

[32] Yehuda Koren, Robert Bell, in Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[33] Gert R. G. Lanckriet, Tijl De Bie, Nello Cristianini, Michael I. Jordan, in William Stafford Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635, 2004.

[34] Hans Laurberg, Mads G. Christensen, Mark D. Plumbley, Lars K. Hansen, in Soren H. Jensen. Theorems on positive data: on the uniqueness of NMF. *Computational Intelligence and Neuroscience*, 2008.

[35] Daniel D. Lee in H. Sebastian Seung. Unsupervised learning by convex and conic coding. V Michael Mozer, Michael I. Jordan, in Thomas Petsche, editors, *NIPS*, strani 515–521. MIT Press, 1996.

[36] Daniel D. Lee in H. Sebastian Seung. Algorithms for non-negative matrix factorization. V Todd K. Leen, Thomas G. Dietterich, in Volker Tresp, editors, *NIPS*, strani 556–562. MIT Press, 2000.

[37] Stan Z. Li, XinWen Hou, HongJiang Zhang, in QianSheng Cheng. Learning spatially localized, parts-based representation. V *CVPR (1)*, strani 207–212. IEEE Computer Society, 2001.

[38] Tao Li, Yi Zhang, in Vikas Sindhwani. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. V *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, strani 244–252, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[39] Chih-Jen Lin. Projected gradient methods for nonnegative matrix factorization. *Neural Comput.*, 19(10):2756–2779, 2007.

[40] Bing Liu, Wee Sun Lee, Philip S. Yu, in Xiaoli Li. Partially supervised classification of text documents. V *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, strani 387–394, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[41] Yaniv Loewenstein, Domenico Raimondo, Oliver C Redfern, James Watson, Dmitrij Frishman, Michal Linial, Christine Orengo, Janet Thornton, in Anna Tramontano. Protein function annotation by homology-based inference. *Genome Biology*, 10(2), 2009.

[42] Bo Long, Xiaoyun Wu, Zhongfei (Mark) Zhang, in Philip S. Yu. Unsupervised learning on k-partite graphs. V *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, strani 317–326, New York, NY, USA, 2006.

[43] Kenneth McGarry, Sheila Garfield, in Nick Morris. Recent trends in knowledge and data integration for the life sciences. *Expert Systems*, 23(5):330–341, 2006.

[44] Sara Mostafavi in Quaid Morris. Fast integration of heterogeneous data sources for predicting gene function with limited annotation. *Bioinformatics*, 26(14):1759–1765, 2010.

[45] Chad Myers, Drew Robson, Adam Wible, Matthew Hibbs, Camelia Chiriac, Chandra Theesfeld, Kara Dolinski, in Olga Troyanskaya. Discovery of biological networks from diverse functional genomic data. *Genome Biology*, 6(13):R114, 2005.

[46] William Stafford Noble in Asa Ben-Hur. Integrating information for protein function prediction. V *Bioinformatics–From Genomes to Therapies*, strani 1297–1314. Wiley VCH Verlag GmbH & Co KGaA, Weinheim, Germany, 2007.

[47] Rong Pan, Yunhong Zhou, Bin Cao, Nathan N. Liu, Rajan Lukose, Martin Scholz, in Qiang Yang. One-class collaborative filtering. V *Pro-*

ceedings of the 2008 Eighth IEEE International Conference on Data Mining, ICDM '08, strani 502–511, Washington, DC, USA, 2008. IEEE Computer Society.

[48] Alberto Pascual-Montano, J. M. Carazo, Kieko Kochi, Dietrich Lehmann, in Roberto D. Pascual-Marqui. Nonsmooth nonnegative matrix factorization (nsNMF). *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(3):403–415, 2006.

[49] Paul Pavlidis, Jinsong Cai, Jason Weston, in William Stafford Noble. Learning gene functional classifications from multiple data types. *Journal of Computational Biology*, 9:401–411, 2002.

[50] Paul Pavlidis, Jason Weston, Jinsong Cai, in William Noble Grundy. Gene functional classification from heterogeneous data. V *Proceedings of the fifth annual international conference on Computational biology*, RECOMB '01, strani 249–255, New York, NY, USA, 2001.

[51] Irina Rish, Genady Grabarnik, Guillermo Cecchi, Francisco Pereira, in Geoffrey J. Gordon. Closed-form supervised dimensionality reduction with generalized linear models. V *Proceedings of the 25th international conference on Machine learning*, ICML '08, strani 832–839, New York, NY, USA, 2008. ACM.

[52] Sanja Rogic, B. F. Francis Ouellette, in Alan K. Mackworth. Improving gene recognition accuracy by combining predictions from two gene-finding programs. *Bioinformatics*, 18(8):1034–1045, 2002.

[53] Omer Sinan Sarac, Özge Gürsoy-Yüzügüllü, Rengul Cetin-Atalay, in Volkan Atalay. Subsequence-based feature map for protein function classification. *Comput. Biol. Chem.*, 32(2):122–130, 2008.

[54] Ömer Sinan Saraç, Volkan Atalay, in Rengul Cetin-Atalay. GOPred: GO Molecular Function Prediction by Combined Classifiers. *PLoS ONE*, 5(8):e12382, 2010.

[55] Mikkel N. Schmidt, Ole Winther, in Lars Kai Hansen. Bayesian non-negative matrix factorization. V *Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation*, ICA '09, strani 540–547, Berlin, Heidelberg, 2009. Springer-Verlag.

[56] John Shawe-Taylor in Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA, 2004.

[57] Ajit P. Singh in Geoffrey J. Gordon. A unified view of matrix factorization models. V *Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases - Part II*, ECML PKDD '08, strani 358–373, Berlin, Heidelberg, 2008. Springer-Verlag.

[58] Barry Smith, Werner Ceusters, Bert Klagges, Jacob Köhler, Anand Kumar, Jane Lomax, Chris Mungall, Alan L. Rector Fabian Neuhaus and, in Cornelius Rosse. Relations in biomedical ontologies. *Genome Biology*, 6, 2005.

[59] Artem Sokolov in Asa Ben-hur. A structured-outputs method for prediction of protein function. V *In Proceedings of the 3rd International Workshop on Machine Learning in Systems Biology*, strani 49–58, 2008.

[60] Nathan Srebro in Tommi Jaakkola. Linear dependent dimensionality reduction. V *In Advances in Neural Information Processing Systems 16*. MIT Press, 2003.

[61] Léon-Charles Tranchevent, Francisco Bonachela Capdevila, Daniela Nitsch, Bart De Moor, Patrick De Causmaecker, in Yves Moreau. A guide to web tools to prioritize candidate genes. *Briefings in Bioinformatics*, 12(1):22–32, 2010.

[62] Olga G. Troyanskaya, Kara Dolinski, Art B. Owen, Russ B. Altman, in David Botstein. A bayesian framework for combining heterogeneous data sources for gene function prediction (in S. cerevisiae). *Proc Natl Acad Sci USA*, 100(14):8348–53, 2003.

[63] Koji Tsuda, Hyunjung Shin, in Bernhard Schölkopf. Fast protein classification with multiple networks. *Bioinformatics*, 21(2):59–65, 2005.

[64] Fei Wang, Tao Li, in Changshui Zhang. Semi-supervised clustering via matrix factorization. V *SDM*, strani 1–12. SIAM, 2008.

[65] Yuan Wang in Yunde Jia. Fisher non-negative matrix factorization for learning local features. V *In Proc. Asian Conf. on Comp. Vision*, strani 27–30, 2004.

[66] David Warde-Farley, Sylva L. Donaldson, Ovi Comes, Khalid Zuberi, Rashad Badrawi, Pauline Chao, Max Franz, Chris Grouios, Farzana Kazi, Christian Tannus Lopes, Anson Maitland, Sara Mostafavi, Jason Montojo, Quentin Shao, George Wright, Gary D. Bader, in Quaid Morris. The genemania prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*, 38(Web-Server-Issue):214–220, 2010.

[67] Yoshihiro Yamanishi, Jean-Philippe Vert, in Minoru Kanehisa. Supervised enzyme network inference from the integration of genomic data and chemical information. V *ISMB (Supplement of Bioinformatics)*, strani 468–477, 2005.

[68] Han Yan, Kavitha Venkatesan, John E. Beaver, Niels Klitgord, Muhammed A. Yildirim, Tong Hao, David E. Hill, Michael E. Cusick, Norbert Perrimon, Frederick P. Roth, in Marc Vidal. A genome-wide gene function prediction resource for Drosophila melanogaster. *PLoS ONE*, 5(8):e12139, 2010.

[69] Shi-Hua Zhang, Qingjiao Li, Juan Liu, in Xianghong Jasmine Zhou. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. *Bioinformatics*, 27(13):401–409, 2011.

[70] Zhongyuan Zhang, Tao Li, Chris Ding, in Xiangsun Zhang. Binary matrix factorization with applications. V *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, ICDM '07, strani 391–400, Washington, DC, USA, 2007. IEEE Computer Society.

[71] Marinka Žitnik in Blaž Zupan. NIMFA: A Python library for nonnegative matrix factorization. *Journal of Machine Learning Research*, 13:849–853, 2012.

Dodatek A

Pravilnost in konvergenca kazenske matrične faktorizacije

Dokaz o pravilnosti in konvergenci kazenske matrične faktorizacije povzemamo po objavi dela Wang in sod. (2008) [64].

Izrek A.1 *(***Pravilnost algoritma PMF***) Če pravili za posodabljanje matričnih razcepnih faktorjev* $\mathbf{F}$ *in* $\mathbf{G}$ *v modelu PMF, danim z algoritmom 1, konvergirata, potem končna rešitev zadošča Karush-Kuhn-Tuckerjevim (KKT) pogojem optimalnosti.*

Dokaz.

Kriterijsko funkcijo modela PMF iz (2.21) prepišemo v kompaktno matrično obliko, da upoštevamo lastnosti Frobeniusove norme in matričnega operatorja sledi. Kriterijska funkcija se sedaj izraža

$$J(\pi) = tr(\mathbf{X}^T\mathbf{X} - 2\mathbf{F}^T\mathbf{X}\mathbf{G} + \mathbf{G}\mathbf{F}^T\mathbf{F}\mathbf{G}^T + \mathbf{G}^T\mathbf{\Theta}\mathbf{G}). \tag{A.1}$$

Sledimo standardni teoriji iskanja vezanih ekstremov, tako da uvedemo matriko Lagrangeovih multiplikatorjev β in definiramo Lagrangeovo funkcijo

$$L = J(\pi) - tr(\beta\mathbf{G}^T). \tag{A.2}$$

Sedaj združimo izraza (A.1) in (A.2), da izpeljemo parcialni odvod kriterijske funkcije J glede na matrični razcepni faktor

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{F}} &= -2\mathbf{X}\mathbf{G} + 2\mathbf{F}\mathbf{G}^T\mathbf{G}, && \text{(A.3)} \\ \frac{\partial L}{\partial \mathbf{G}} &= -2\mathbf{X}^T\mathbf{F} + 2\mathbf{G}\mathbf{F}^T\mathbf{F} + 2\boldsymbol{\Theta}\mathbf{G} - \beta. && \text{(A.4)} \end{aligned}$$

Za dani $\mathbf{G}$ zahtevamo, da je parcialni odvod po $\mathbf{F}$ ničelen, $\frac{\partial L}{\partial \mathbf{F}} = 0$. Zato lahko iz (A.3) izrazimo pravilo za posodabljanje matrike $\mathbf{F}$, ki se glasi $\mathbf{F} = \mathbf{X}\mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}$. Prav tako za dani $\mathbf{F}$ zahtevamo, da je parcialni odvod po $\mathbf{G}$ ničelen, $\frac{\partial L}{\partial \mathbf{G}} = 0$. Zato lahko iz (A.4) izrazimo matriko Lagrangeovih multiplikatorjev $\beta = -2\mathbf{X}^T\mathbf{F} + 2\mathbf{G}\mathbf{F}^T\mathbf{F} + 2\boldsymbol{\Theta}\mathbf{G}$.

Sedaj upoštevajmo komplementarni pogoj KKT na nenegativnost matrike $\mathbf{G}$ iz optimizacijske naloge (2.21), da dobimo pogoj

$$(-2\mathbf{X}^T\mathbf{F} + 2\mathbf{G}\mathbf{F}^T\mathbf{F} + 2\boldsymbol{\Theta}\mathbf{G})_{ij}\mathbf{G}_{ij} = \beta_{ij}\mathbf{G}_{ij} = 0. \qquad \text{(A.5)}$$

Izraz (A.5) je potrebni pogoj, kateremu mora zadoščati rešitev, ki konvergira. Poudarimo, da smo v izreku predpostavili konvergenco pravil za posodabljanje matričnih razcepnih faktorjev.

Zapišimo sledeče matrike v obliki razlike med njihovo pozitivno in negativno vsebino

$$\begin{aligned} \boldsymbol{\Theta} &= \boldsymbol{\Theta}^+ - \boldsymbol{\Theta}^-, \\ \mathbf{F}^T\mathbf{F} &= (\mathbf{F}^T\mathbf{F})^+ - (\mathbf{F}^T\mathbf{F})^-, \\ \mathbf{X}^T\mathbf{F} &= (\mathbf{X}^T\mathbf{F})^+ - (\mathbf{X}^T\mathbf{F})^-. && \text{(A.6)} \end{aligned}$$

Vseh šest matrik na desnih straneh v (A.6) je nenegativnih. Potem z danim začetnim približkom za razcepni matrični faktor $\mathbf{G}$ posodabljanje matrike $\mathbf{G}$ s pravilom

$$\mathbf{G}_{ij} \leftarrow \mathbf{G}_{ij}\sqrt{\frac{(\mathbf{X}^T\mathbf{F})^+_{ij} + [\mathbf{G}(\mathbf{F}^T\mathbf{F})^-]_{ij} + (\mathbf{\Theta}^-\mathbf{G})_{ij}}{(\mathbf{X}^T\mathbf{F})^-_{ij} + [\mathbf{G}(\mathbf{F}^T\mathbf{F})^+]_{ij} + (\mathbf{\Theta}^+\mathbf{G})_{ij}}} \tag{A.7}$$

konvergira proti lokalnemu minimumu optimizacijske naloge PMF. V primeru konvergence namreč velja $\mathbf{G}^{(\infty)} = \mathbf{G}^{(t+1)} = \mathbf{G}^{(t)} = \mathbf{G}$, torej sledi enakost

$$\mathbf{G}_{ij} = \mathbf{G}_{ij}\sqrt{\frac{(\mathbf{X}^T\mathbf{F})^+_{ij} + [\mathbf{G}(\mathbf{F}^T\mathbf{F})^-]_{ij} + (\mathbf{\Theta}^-\mathbf{G})_{ij}}{(\mathbf{X}^T\mathbf{F})^-_{ij} + [\mathbf{G}(\mathbf{F}^T\mathbf{F})^+]_{ij} + (\mathbf{\Theta}^+\mathbf{G})_{ij}}}, \tag{A.8}$$

ki je ekvivalentna izrazu $(-2\mathbf{X}^T\mathbf{F} + 2\mathbf{G}\mathbf{F}^T\mathbf{F} + 2\mathbf{\Theta}\mathbf{G})_{ij}\mathbf{G}^2_{ij} = 0$. Ta pa je ekvivalenten potrebnemu pogoju (A.5), ki mu zadošča rešitev optimizacijske naloge PMF.

□

Definicija A.1 *(***Pomožna funkcija***) Funkcija $Z(\mathbf{G}, \mathbf{G}')$ je pomožna funkcija funkcije $L(\mathbf{G})$, če za vsak par $\mathbf{G}, \mathbf{G}'$ velja*

$$Z(\mathbf{G}, \mathbf{G}') \geq L(\mathbf{G}) \;\; in \;\; Z(\mathbf{G}, \mathbf{G}) = L(\mathbf{G})$$

Izrek A.2 *(***Konvergenca algoritma PMF***) Posodabljanje matričnih razcepnih faktorjev $\mathbf{F}$ in $\mathbf{G}$, kot je določeno z modelom PMF v algoritmu 1, zagotavlja konvergenco algoritma.*

Dokaz.

Naj bo $\{\mathbf{G}^{(t)}\}$ zaporedje približkov matričnih razcepnih faktorjev $\mathbf{G}$, dobljenih v modelu PMF z algoritmom 1, pri čemer (t) označuje zaporedno številko iteracije v algoritmu. Sedaj definirajmo $\mathbf{G}^{(t+1)} = \arg\min_{\mathbf{G}} Z(\mathbf{G}, \mathbf{G}^{(t)})$, pri čemer je $Z(\mathbf{G}, \mathbf{G}')$ pomožna funkcija za $L(\mathbf{G}) = J(\pi)$, določena z definicijo (A.1). Iz definicije pomožne funkcije sledi, da je $L(\mathbf{G}^{(t)})$ monotono padajoče zaporedje približkov, saj velja

$$L(\mathbf{G}^{(t)}) = Z(\mathbf{G}^{(t)}, \mathbf{G}^{(t)}) \geq Z(\mathbf{G}^{(t+1)}, \mathbf{G}^{(t)}) \geq L(\mathbf{G}^{(t+1)}). \tag{A.9}$$

Torej moramo za dokaz izreka poiskati ustrezno funkcijo $Z(\mathbf{G}, \mathbf{G}')$, ki bo izpolnjevala pogoje pomožne funkcije za kriterijsko funkcijo $J(\pi) = J(\mathbf{F}, \mathbf{G})$ optimizacijske naloge (2.21) modela PMF.

Izkaže se, da znamo poiskati funkcijo $Z(\mathbf{G}, \mathbf{G}')$, ki je pomožna funkcija za $J(\pi)$. Še več, $Z(\mathbf{G}, \mathbf{G}')$ je tudi konveksna funcija matrične spremenljivke $\mathbf{G}$ in doseže globalni minimum pri

$$\mathbf{G}_{ij} = \mathbf{G}_{ij} \sqrt{\frac{(\mathbf{X}^T\mathbf{F})^+_{ij} + [\mathbf{G}(\mathbf{F}^T\mathbf{F})^-]_{ij} + (\mathbf{\Theta}^-\mathbf{G})_{ij}}{(\mathbf{X}^T\mathbf{F})^-_{ij} + [\mathbf{G}(\mathbf{F}^T\mathbf{F})^+]_{ij} + (\mathbf{\Theta}^+\mathbf{G})_{ij}}}. \tag{A.10}$$

Bralec bo izpis pomožne funkcije $Z(\mathbf{G}, \mathbf{G}')$ in dokaz navedenih lastnosti našel v izreku 6.1 v delu Wang in sod. (2008) [64]. Od tod sledi, da za kriterijsko funkcijo $J(\pi) = J(\mathbf{F}, \mathbf{G})$ velja

$$J(\mathbf{F}^0, \mathbf{G}^0) \geq J(\mathbf{F}^1, \mathbf{G}^0) \geq J(\mathbf{F}^1, \mathbf{G}^1) \geq \cdots \geq L(\mathbf{G}), \tag{A.11}$$

kar pomeni, da kriterijska funkcija optimizacijske naloge PMF monotono pada. Izraz $J(\mathbf{F}, \mathbf{G})$ je navzdol omejen, zato je izrek dokazan.

□

Postopek dokazovanja je pomemben zaradi uporabljene tehnike. Pri dokazovanju konvergence matričnih faktorizacij je princip pomožnih funkcij pogosto uporabljen – za dokazovanje sta jih uporabila začetnika nenegativne matrične faktorizacije Lee in Seung (2000) [36].

Dodatek B

Podatki mutant o bakterijski rezistenci *D. discoideum*

Tabela B.1: Rast mutant *D. discoideum* v okolju Gram pozitivne bakterije *S. aureus* in Gram negativne bakterije *K. pneumoniae*. Naslednje oznake se nanašajo na vedenje mutant v obeh okoljih; SG – počasna rast, FG – hitra rast, NG – nespremenjena rast, NGA – brez rasti. Za analizo bakterijske rezistence so zanimive mutante, ki ne rastejo bodisi v okolju Gram pozitivnih bodisi Gram negativnih bakterij; GPD – brez rasti v Gram pozitivnem okolju, GND – brez rasti v Gram negativnem okolju.

Mutanta	Ime gena	Rast
DDB_G0286229	*alyL*	GND
DDB_G0295477	*nip7*	GND
DDB_G0278487		GND
DDB_G0290851	*spc3*	GND
DDB_G0286195	*nagB1*	GPD
DDB_G0283673	*gpi*	GPD
DDB_G0289479	*swp1*	GPD
DDB_G0276559		SG
DDB_G0276459	*pakB*	SG
DDB_G0292196		SG
DDB_G0270300		SG

Mutanta	Ime gena	Rast
DDB_G0284227		SG
DDB_G0277667		SG
DDB_G0280447	*pelo*	SG
DDB_G0293440		SG
DDB_G0281967		SG
DDB_G0288459		SG
DDB_G0277169		SG
DDB_G0283267	*shkA*	SG
DDB_G0287007	*arpF*	SG
DDB_G0281447		SG
DDB_G0285803		SG
DDB_G0291279		SG
DDB_G0276355	*dhcA*	SG
DDB_G0273915		SG
DDB_G0269608		SG
DDB_G0276277		SG
DDB_G0272382		SG
DDB_G0292054		SG
DDB_G0293476	*fncl*	SG
DDB_G0291802	*qtrt1*	SG
DDB_G0293540	*empB*	SG
DDB_G0278783	*elp4*	SG
DDB_G0281741	*cct3*	SG
DDB_G0279939		SG
DDB_G0278929	*usp39*	NG
DDB_G0283883	*mlh3*	NG
DDB_G0285747		NG
DDB_G0278663		NG
DDB_G0276263		NG
DDB_G0272999		NG
DDB_G0274115	*abcG12*	NG
DDB_G0288161		FG
DDB_G0281785		FG
DDB_G0280893	*rbsk*	NGA

Zahvala

Profesorju dr. Blažu Zupanu se zahvaljujem za izzive, ki so me razveseljevali skozi študij, priložnosti za nove izkušnje ter za svetovanje in spodbudo pri nastajanju pričujočega dela. Njegovo podporo iskreno cenim.

Zahvaljujem se prof. dr. Gadu Shaulskyju in prof. dr. Adamu Kuspi z ustanove Baylor College of Medicine v Houstonu, ZDA, in njunim laboratorijem za opravljene biološke eksperimente.

Članom Laboratorija za bioinformatiko na Fakulteti za računalništvo in informatiko v Ljubljani se želim zahvaliti za nasvete, ki so jih delili z mano v preteklih letih.

Iskrena hvala tudi staršem in bratu Slavku za podporo, spodbudo in potrpežljivost.