Instrumente de clasificare pentru reţele multi-relaţionale şi serii de timp multidimensionale cu aplicaţii în biomedicină

- Rezumat -

UNIVERSITATEA BABEŞ-BOLYAI
Facultatea de Matematică şi Informatică

Student Doctorand: Ramona-Carmen Stoica
Coordonator Ştiinţific: Bazil Pârv
Facultatea de Matematică şi Informatică
Universitatea Babeş-Bolyai

A thesis submitted for the degree of
PhD

2017

1

Cuprinsul tezei de doctorat

Contents

2

Introducere

2.1 Motivaţia cercetării

În zilele noastre, datele experimentale disponibile cercetătorilor şi oamenilor de ştiinţă cresc exponenţial. Reducerea costurilor pentru secvenţializarea datelor din biologie şi medicină oferă o oportunitate imensă pentru analiza, modelarea şi simularea datelor. Bioinformatica, sistemele, biologia şi ingineria bio-modelelor sunt considerate discipline promiţătoare care vizează studiul şi înţelegerea sistemelor biologice. Ştiinţele exacte, cum ar fi fizica sau chimia, precum şi ingineria, utilizează modelarea matematică şi computaţională pentru a înţelege şi analiza mai bine sistemele existente ale lumii reale, prin crearea de modele mai complexe şi prin utilizarea simulărilor pentru a reproduce comportamentul lor dinamic.

Cu toate acestea, majoritatea sistemelor sunt atât de complexe încât toate ipotezele existente pot explica doar parţial comportamentul lor. Prin urmare, o întrebare firească este: cât de multe informaţii despre un sistem sunt necesare pentru a formula un model matematic care să îl explice pe deplin?

Motivaţia acestei teze provine dintr-o nevoie tot mai mare de a dezvolta modele mai bune pentru sisteme complexe, de mari dimensiuni, cu scopul de a obţine noi perspective asupra comportamentului acestor sisteme. Dezvoltarea de noi metode pentru studierea interacţiunilor care apar în reţele complexe are un impact mare asupra multor discipline, de la ştiinţele sociale până la ştiinţele biologice. Cu toate acestea, această activitate se concentrează în principal pe interacţiunile bio-entitate în reţelele biomedicale. Provocările acestei teze includ dezvoltarea aplicaţiilor cât mai generale astfel încât să

poată fi aplicate pe tipuri diferite de reţele biologice, precum şi instrumentele practice care oferă perspective noi asupra comportamentului acestor reţele în diverse condiţii (cum ar fi, de exemplu, mutaţiile).

2.2 Obiectivele cercetării

Obiectivele principale ale acestei teze pot fi împarţite în două categorii de probleme din bioinformatică.

Prima categorie la care se referă această teză sunt reţelele multi-relaţionale (majoritatea reţelelor biologice sunt reţele multi-relaţionale). Al doilea aspect al tezei se referă la clustering-ul multidimensional al seriilor de timp (mai multe seturi de date biologice şi experimente adună informaţii în timp).

Scopul acestei lucrări se concentrează în jurul a două tipuri de probleme şi obiectivele propuse sunt dezvoltarea şi analiza modelelor de reţele multi-relaţionale, clustering-ul multidimensional al seriilor de timp şi evidenţierea unor direcţii generale pentru extinderea modelelor dezvoltate în această teză, care pot evidenţia o perspectivă suplimentară asupra problemelor abordate. Această teză prezintă două idei principale de abordare a problemelor biologice. Cu toate acestea, abordările noastre nu sunt limitate doar la domeniul biologic sau cel medical ci pot fi aplicate cu succes şi în alte domenii.

Primul subiect se referă la reţele multi-relaţionale (reţele cu mai multe legături de diferite tipuri între două noduri). Sunt abordate trei aspecte ale acestor reţele:

- Compararea reţelelor multi-relaţionale,
- Identificarea de graphlets şi motive în reţele multi-relaţionale, precum şi
- Identificarea comunităţilor în reţele multi-relaţionale.

Provocările includ rezolvarea problemelor NP-complete (cum ar fi izomorfismul în reţea), identificarea algoritmilor eficienţi pentru grafuri şi motive precum şi identificarea modalităţii corecte de reprezentare a reţelelor şi a interacţiunilor dintre ele (avem mai multe reţele, unele din tipurile de muchii pot să lipsească într-una sau mai multe reţele, etc).

Al doilea obiectiv al acestei teze este abordarea datelor pentru seriile de timp multidimensionale precum şi clustering-ul acestor entităţi. Cele mai multe date biologice sunt organizate ca serii de timp.

De multe ori, experimentele biologice sunt efectuate în condiţii sau situaţii paralele, sau în medii diferite. Cu toate acestea, scopul final al experimentelor este acela de-a identifica entităţile (gene, proteine, enzime etc.) care se comportă similiar în acelaşi organism (sau organisme similare) în condiţii diferite. Aceasta implică identificarea mai multor serii de timp care caracterizează o entitate şi clustering-ul acestor entităţi in funcţie de aceste serii de timp. Problema este interesantă şi este utilă abordarea acesteia într-un mod diferit.

Un alt obiectiv al acestei teze a fost acela de a dezvolta aplicaţii practice care să contribuie la validarea soluţiilor propuse pentru problemele de cercetare abordate în această teză.

2.3 Contribuţiile tezei

Contribuţiile acestei teze pot fi grupate în două categorii: teoretice / conceptuale şi aplicative / practice. În prima categorie menţionăm algoritmul MultiNetCom (descris în Capitolul 7), conceput pentru detectarea comunităţilor în reţele multi-relaţionale.

Unele îmbunătăţiri ale clustering-ului pentru seriile de timp multidimensionale sunt prezentate în Capitolul 4.

A doua categorie de contribuţii conţine patru aplicaţii software dezvoltate pentru analiza reţelelor multi-relaţionale: (a) o aplicaţie care compară reţele multi-relaţionale, NCTool; (b) o aplicaţie, MultiMot pentru identificarea motivelor; (c) o aplicaţie MultiNetCom pentru detectarea comunităţilor şi (d) MDTSC, o aplicaţie pentru clustering-ul multidimensională a seriilor de timp.

Datele de intrare pentru aplicaţia NCTool constau în două sau mai multe reţele multi-relaţionale. Pe aceste date de intrare se aplică câteva operaţii de bază, cum ar fi reuniunea dintre două reţele, intersecţia dintre două reţele, subgrafurile, gradul de distribuţie al nodurilor. Aceste aplicaţii sunt implementate pentru reţele de interacţiune genă-genă şi reţele de interacţiune proteină-proteină, dar pot fi extinse şi adaptate cu uşurinţă şi altor tipuri de reţele. Utilizatorul are opţiunea de a selecta unul, două, mai multe sau toate tipurile de legături dintr-o reţea şi de-a rula operaţiile pentru tipurile de legături selectate.

Aplicaţia MultiMot identifică motive de o anumită dimensiune (sunt implementate dimensiunile 3, 4 şi 5 noduri), având posibilitatea de a detecta motive customizabile.

Pot fi aplicate atât pentru grafurile directe cât şi pentru cele indirecte.

Aplicaţia MultiNetCom implementează o extensie a algoritmului lui Fruchterman. Algoritmul propus de noi (conceput iniţial pentru reţele cu o singură relaţie) este folosit pentru a rezolva problema detectării de comunităţi în reţele multi-relaţionale.

Aplicaţia MDTSC a fost implementată pentru a fi folosită la clustering-ul entităţilor compuse din mai multe serii de timp. Aplicaţia are o interfaţă uşor de utilizat.

Ţinând cont de domeniile de aplicabilitate, teza abordează aplicaţii din domeniul computaţional care sunt utilizate în cea mai mare parte pentru a rezolva problemele din biologie şi medicină, dar sunt destul de generale pentru a fi aplicate şi în alte domenii similare.

Ţinând cont de principalele domenii abordate, putem grupa contribuţiile în două categorii: *reţele* (sau *grafuri*) şi *clustering*. În cadrul acestei teze ne-am focalizat pe aspectele specifice ale acestor topice şi anume reţele multi-relaţionale (care acceptă mai multe muchii de diferite tipuri între două noduri). De asemenea, am abordat conceptul de clustering al seriilor de timp multidimensionale (entităţi compuse din mai multe serii de timp).

Contribuţiile principale ale acestei teze sunt evidenţiate mai jos:

- Definirea conceptelor necesare pentru înţelegerea reţelelor multi-relaţionale şi extinderea de la caracteristicile specifice reţelelor cu o singură relaţie la caracteristicile specifice reţelelor multi-relaţionale.
- Proiectarea, implementarea şi testarea algoritmilor pentru compararea reţelelor biologice multi-relaţionale (47).
- Proiectarea, implementarea şi testarea algoritmilor pentru detectarea motivelor şi graphlets în reţele multi-relaţionale. Acestă aplicaţie are două părţi:
  - motivele de o anumită dimensiune (date de numărul de noduri) pot fi găsite folosind un algoritm de "sampling" (dimensiunile 3, 4 şi 5 sunt implementate, dar motivele sunt uşor de adaptat la dimensiuni mai mari) şi
  - motive customizabile care pot fi detectate folosind un algoritm de backtracking. Pot fi căutate simultan mai multe motive.

- Proiectarea, implementarea şi testarea unui nou algoritm pentru detectarea comunităţilor în reţele multi-relaţionale (46).

3

Reţele complexe

Sistemele complexe sunt adesea analizate ca şi reţele din cauza topologiei lor comune. Acest capitol defineşte conceptul de reţea şi noţiunile specifice într-o abordare ştiinţifică, şi descrie, de asemenea, clasificarea şi reprezentarea acestora. Noţiunile de bază despre reţele includ: definiţiile reţelelor, clasificări, structură şi reprezentare (ca matrice, listă de adiacenţă şi listă de incidenţă). Sunt descrise unele măsurători utilizate în reţele (gradul de centralitate, distribuţia gradului, coeficientul de clustering, centralitatea). Modelele de reţele, cum ar fi modelul Erdos-Renyi (2), modelul Watts-Strogatz (34) şi modelul Barabasi-Albert (2), necesare în capitolele următoare sunt de asemenea menţionate şi descrise pe scurt. Ultima secţiune descrie câteva proprietăţi ale reţelelor: motive în reţea şi grafuri.

3.1 Definiţiile reţelelor

O mare varietate de sisteme naturale şi sociale poate fi descrise de reţele cu o topologie complexă. De obicei, reprezentate ca şi grafuri aleatoare, este unanim recunoscut faptul că evoluţia şi topologia lor în lumea reală sunt conduse de reguli şi principii puternice (1). Această subsecţiune introduce conceptele de bază ale reţelelor complexe şi oferă o imagine de ansamblu asupra principalelor modele de reţele întâlnite în lumea reală, cum ar fi reţelele de dimensiuni reduse precum şi cele aleatorii (34).

Din cauza asemănării lor cu grafurile, reţelele sunt considerate ca fiind grafuri mult mai mari şi sunt definite după aceleaşi structuri şi relaţii:

- Fie $G = (V, E)$ reprezentarea unui graf, unde V reprezintă setul de noduri (sau vârfuri) şi E reprezintă setul de muchii (sau conexiuni, legături între noduri) (8).
- Fie $|V|$ numărul nodurilor (cardinal de V) din graf. Fie $|E|$ numărul de muchii din graf.
- Fie $Deg(v), v \in V$, numărul de legături (muchii) conectate la v.
- Un drum între nodurile $s, t \in V$ este definit de o secvenţă alternantă de noduri şi legături, începând cu s şi terminând cu t. Fiecare legătură conectează nodurile adiacente (8). Lungimea drumului este determinată de suma greutăţilor aflate pe muchiile sale tinând cont de funcţia de greutate (8).

Reţelele pot fi clasificate în funcţie de cel puţin trei criterii: zona în care se aplică, structura şi modelul acestora:

1. În funcţie de aplicabilitatea lor avem următoarele tipuri de reţele:
   - Reţele de calculatoare
   - Reţele de transport
   - Reţele de conexiuni sinaptice
   - Reţele funcţionale ale creierului
   - Reţele de boli (gene bolnave)
   - Reţele ecologice care sintetizează ecosistemele
   - Reţele de telecomunicaţii
   - Reţele World Wide Web
   - Reţele de socializare
   - Reţele biologice, etc.
2. După structura lor avem următoarele tipuri de reţele:
   - Considerăm $\omega(e), e \in E$, funcţia de greutate reprezentând legăturile sale.
   - O reţea neponderată şi simplă are $\omega(e) = 0$ pentru toate legăturile şi doar o legătură e $\in E$ între două noduri $u, v \in E$ (8).

- O reţea ponderată are aceeaşi structură pentru legături ca şi cea neponderată, dar cu $\omega(e) \in R$ (51).
- O reţea multiplă permite mai multe legături între aceeaşi pereche de vârfuri şi poate fi clasificată şi după greutăţile muchiilor: aceasta poate fi ponderată sau neponderată (51).
- O reţea direcţionată este descrisă de setul de muchii care au o direcţie asociată cu acestea; poate fi şi ponderată sau neponderată (51).

3. După nodurile şi gradul de distribuţie care reprezintă modelele de reţele:
   - O reţea de dimensiuni mici are o distribuţie specială a vârfurilor: distanţa D între două noduri creşte proporţional cu logaritmul numărului de noduri $|V|$ din reţea (51).
   - O reţea scale-free (majoritatea reţelelor biologice sunt scale-free) respectă o distribuţie asemănătoare muchiilor în care gradul fiecărui nod respectă formula pentru putere (51).
   - O reţea aleatoare nu are nici o restricţie de grad: nu se respectă legea gradelor nodurilor (51).

Cu siguranţă există o clasificare mai detaliată a reţelelor, cu mai multe informaţii, însă aceste clasificări sunt generale şi acoperă cele mai populare tipuri de reţele.

3.2 Reprezentarea reţelelor

Aşa cum am definit mai devreme, o reţea este definită de $G(V, E)$ având setul de vârfuri V şi setul de muchii E. Reţelele pot fi reprezentate în mai multe moduri, în funcţie de conexiunea existentă între noduri:

- reprezentarea ca şi matrice de adiacenţă,
- reprezentarea ca şi listă de adiacenţă,
- reprezentarea ca şi matrice de incidenţă,
- eprezentarea ca şi listă de adiacenţă.

4

Reţele biologice

Acest capitol descrie diferite tipuri de reţele biologice şi prezintă câteva aspecte comune între reţelele biologice şi cele reale.

Există mai multe tipuri de reţele biologice însă nu toate fac obiectul de studiu al acestei lucrări. Acest capitol insistă asupra unor tipuri de reţele care vor fi utilizate în experimente sau reţele similare care pot fi abordate ca şi extensii ulterioare ale aplicaţiilor implementate în această teză. Unele dintre reţelele studiate includ interacţiunile dintre gene, interacţiunile dintre proteine, interacţiunile dintre elementele biochimice (reacţiile chimice) şi reţelele metabolice.

4.1 Reţele metabolice

Activitatea unei celule este stabilită în reţele complexe de reacţii chimice care produc diferite funcţii celulare. Procesul care ne ajută să identificăm reacţiile dintr-o reţea se numeşte reconstrucţia reţelei. Suma proceselor fizice şi metabolice care determină caracteristicile celulelor, biochimice şi fiziologice, formează o reţea metabolică. Prin însăşi natura sa, aceste reţele compun reacţiile chimice ale metabolismului, căile metabolice şi, în acelaşi timp, interacţiunile de reglementare care conduc aceste reacţii (18, 22, 25, 39, 43).

Metabolismul intermediar poate fi considerat un instrument chimic care transformă materialele de bază în energie, precum şi blocurile necesare pentru a produce structuri biologice, pentru a susţine celulele şi pentru a menţine diferitele funcţii ale celulelor. Acest instrument chimic se schimbă extrem de mult, acceptă legile chimiei şi ale fizicii

şi, din acest motiv, este limitat de diferite constrângeri fizico-chimice. În acelaşi timp, are o structură de reglementare complicată, care îi permite să răspundă la o combinaţie de perturbaţii externe. Dezechilibrul metabolic este rădăcina principalelor boli întâlnite la oameni, cum ar fi bolile de inimă, cancerul, diabetul şi obezitatea. Metabolismul este compus din două tipuri diferite de transformări chimice, cum ar fi: căile catabolice care descompun substraturile în metaboliţi simpli şi căile anabolice care sintetizează aminoacizii, acizii nucleici, acizii graşi şi alte blocuri necesare.

Pe parcursul acestor procese, există un schimb complex între diferite grupuri chimice şi potenţialele de reducere ale oxidării (redox) care apar într-un un set de molecule. Aceste molecule de transport şi proprietăţile pe care le transferă astfel leagă împreună reţeaua metabolică.

Ierarhia în reţele metabolice Reţelele metabolice sunt greu de înţeles pentru mintea umană deoarece reconstrucţiile pe scară genomică ale reţelelor metabolice sunt compuse din procese mari care constau în multitudinea de metaboliţi şi din când în când conţin peste o mie de reacţii. Prin urmare, avem nevoie de modele matematice pentru a studia proprietăţile lor şi a le simula funcţiile. Cu toate acestea, putem vedea proprietăţile unei reţele într-un mod ierarhic pentru a simplifica abordarea (teoria) funcţiilor reţelei.

Mai jos sunt prezentate cele patru niveluri de descompunere funcţională ale metabolismului:

- Nivelul 1: întreaga celulă
- Nivelul 2: sectoarele metabolice
- Nivelul 3 3: căile
- Nivelul 4: reacţiile individuale

Există două metode de reconstrucţie prezentate mai jos:

- Reconstrucţia metabolică a genomului
- Multiple reţele genomice

În reţelele metabolice avem metaboliţi şi căi metabolice. Metaboliţii sunt particule microscopice cum ar fi glucoza şi aminoacizii, sau macromolecule precum polizaharidele, şi carbohidraţii.

Căile metabolice sunt secvenţe de reacţii biochimice consecutive pentru o anumită funcţie metabolică, de exemplu glicoliza sau sinteza penicilinei, care transformă un metabolit într-altul. Enzimele sunt proteine care catalizează (accelerează) reacţiile chimice. Astfel, într-o cale metabolică, metaboliţii şi enzimele sunt corespondenţii nodurilor, iar reacţiile metabolice corespund muchiilor. Abordările mai simple indică faptul că nodurile reprezintă metaboliţi şi muchiile direcţionate sunt reacţiile care transformă un metabolit într-altul. Exemple ale unei părţi dintr-o cale de glicoliză, o reprezentare a metabolitului, reacţii, i metaboliţii sunt prezentate în figurile 4.1, 4.2, şi 4.3:

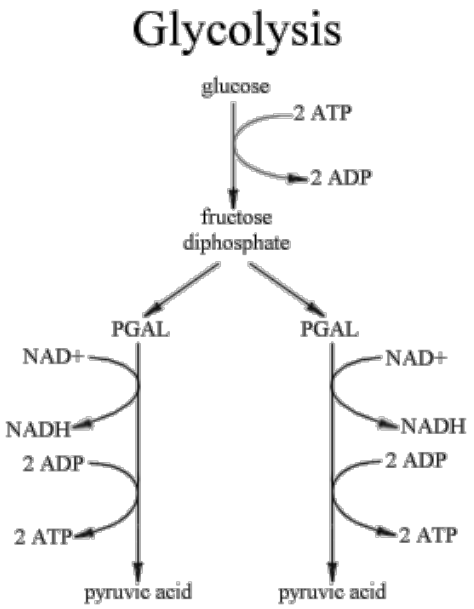


Figure 4.1: Parte a căii de glicoliză.
source: `http://library.thinkquest.org/27819/ch4_4.shtml`

Toate căile metabolice ale unei celule formează o reţea metabolică. O astfel de reţea este compusă din vizualizarea completă a metabolismului celular şi a fluxului de material. Celulele se bazează pe această reţea pentru a digera substraturi din mediul înconjurător, pentru a genera energie şi pentru a sintetiza componentele necesare din mediul înconjurător pentru creşterea şi supravieţuirea lor. Aceste reţele sunt, de asemenea, utile pentru tratarea bolilor metabolice umane printr-o mai bună înţelegere a mecan-

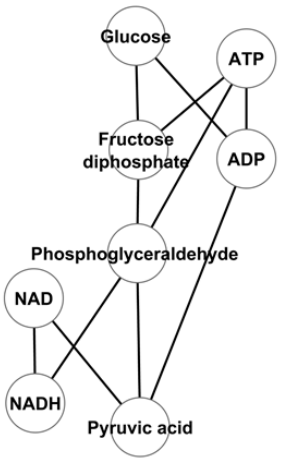


Figure 4.2: Reprezentarea unui metabolit

ismelor metabolice sau prin controlul infecţiilor patogene prin înţelegerea diferenţelor metabolice dintre om şi agenţii patogeni (49).

Căile metabolice sunt construite parţial experimental şi parţial din secvenţa genomului (omologie). Aceste reţele sunt generale şi sunt folosite pentru multe organisme, de la bacterii până la oameni (26). Proiectul Enciclopedia de Gene şi Genomul (KEGG) este o colecţie vastă de baze de date online care au legătură cu genomul, căile enzimatice şi substanţele chimice biologice.

KEGG este o bază de date creată cu scopul de a înţelege funcţiile şi utilităţile sistemelor biologice. Ajută la învăţarea sistemelor biologice cum ar fi organismul, celula şi ecosistemele din punct de vedere molecular şi genomic. Se compune din următoarele baze de date: informaţiile chimice, informaţiile reţelei şi informaţiile genomice (27).

Sistemul biologic este descris în figura 4.4. Reprezentarea include blocuri moleculare ale genelor şi proteinelor (de exemplu, informaţii genomice) şi substanţe chimice (informaţii chimice) care sunt combinate cu cunoştinţele de diagrame de interacţiune, reacţii şi relaţii între reţele (informaţii ale sistemului) (26).

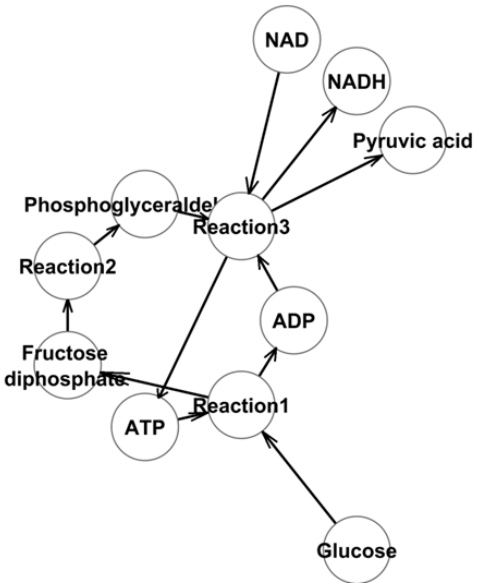


Figure 4.3: Reacţii şi metaboliţi.

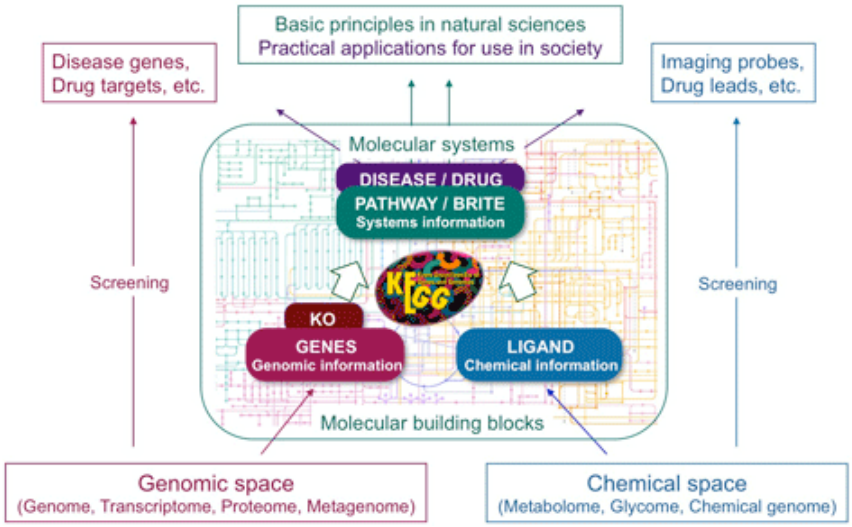


Figure 4.4: Reprezentarea digitală a sistemului biologic.
source: `http://www.genome.jp/kegg/kegg1a.html` (26)

4.2 Reţele de interacţiune genă-genă

Reţelele de interacţiune dintre gene (GGI) sunt reţele în care nodurile reprezintă genele (de la un anumit organism) şi legăturile sunt muchiile dintre ele. Aceste conexiuni sunt obţinute din diferite studii, din experimente şi din numeroasele baze de date existente. Există câteva instrumente GGI, dar una dintre cele mai comune şi mai flexibile este Genemania (33, 37). Interacţiunile dintre genele incluse în Genemania sunt următoarele (50):

- **Co-expresie.** Două gene sunt de tipul co-expresie dacă nivelurile de expresie sunt similare în funcţie de anumite condiţii într-un studiu de exprimare al genei.
- **Interacţiunea fizică.** Două gene interacţionează fizic dacă ele interacţionează într-un studiu de interacţiune între proteine sau într-o bază de date.
- **Interacţiune genetică.** Două gene au o interacţiune genetică între ele dacă schimbările unei gene au un efect asupra modificărilor celeilalte gene.
- **Suprafeţe comune de proteine.** Două gene (de exemplu proteine) au acest tip de interacţiune daca au acelaşi domeniu proteic.
- **Co-localizare.** Două gene sunt co-localizate dacă ambele sunt exprimate în acelaşi ţesut sau dacă produsele lor genetice sunt ambele identificate în aceeaşi locaţie celulară.
- **Pathway.** Două gene (de exemplu proteine) sunt conectate daca sunt parte a aceleaşi reacţie într-o cale.
- **Predictibile.** Două gene au o relaţie funcţională predictibilă dacă există relaţii funcţionale cunoscute de la un alt organism prin ortologie.
- **Alte interacţiuni.** Toate celelalte relaţii: date privind genomica chimică, corelaţii fenotipice de la Ensembl, informaţii despre boală de la OMIM.

5

Clustering în reţele complexe

Acest capitol prezintă noţiunile de bază de clustering, metricele de similitudine şi cîţiva algoritmi pentru clustering. De asemenea, capitolul descrie conceptul de clustering al seriilor de timp, extensia acestor concepte de clustering de la serii de timp unidimensionale la serii de timp multidimensionale şi importanţa practică a acestor algoritmi. Un algoritm nou de clustering al seriilor de timp multidimensionale este propus şi implementat, iar implementarea este testată pe cîteva seturi de date reale.

5.1 Noţiuni de bază

Clustering-ul e procesul prin care obiectele într-o mulţime sunt divizate în submulţimi distincte, astfel încît obiectele similare aparţin la aceeaşi submulţime. Clustering-ul face parte din clasa algoritmilor de învăţare nesupervizată, deoarece nu există nici o proprietate care ar indica cum ar trebui grupate obiectele.

Un *cluster* conţine obiecte care satisfac următoarele proprietăţi:

- obiectele în fiecare cluster sunt foarte similare
- obiectele în clustere diferite sunt foarte diferite

Există multe definiţii de similitudine in literatura ştiinţifică (13). Aceste definiţii se bazează pe:

- tipul datelor luate în considerare
- tipul similitudinii dorite

De obicei, metricele de similitudine sunt bazate pe o funcţie de distanţă $d(x, y)$. Ideal, funcţia aleasă ar trebui să satisfacă următoarele restricţii (17, 54):

- $d(x, y) \geqslant 0$
- $d(x, y) = 0 \iff x = y$
- $d(x, y) = d(y, x)$
- $d(x, z) \leqslant d(x, y) + d(y, z)$

5.2 Clustering pe serii de timp

Seriile de timp sunt secvenţe de numere reale care corespund valorilor unor parametri observate la intervale egale de timp.

Seriile de timp pot fi continue (variabila e definită pentru toate punctele în timp) sau discrete. Seriile de timp folosite în clustering sunt de obicei discrete şi sunt o combinaţie de mai multe componente (6):

1. un trend
2. fluctuaţii de o anumită regularitate
3. un component sezonier
4. un efect aleator

Clustering-ul seriilor de timp are multe aplicaţii reale în diverse domenii ştiinţifice. Seriile de timp de obicei au cîteva valori extreme, mai ales în mulţimile de date foarte mari. Elementele seriilor de timp au o ordine temporala şi operaţiile pe astfel de date sunt des întîlnite in domeniul extragerii datelor. Algoritmii de clustering ai seriilor de timp au fost studiaţi extensiv şi eficienţa acestor algoritmi a fost testată în numeroase aplicaţii reale (3, 5, 11, 19, 21, 30, 41, 48).

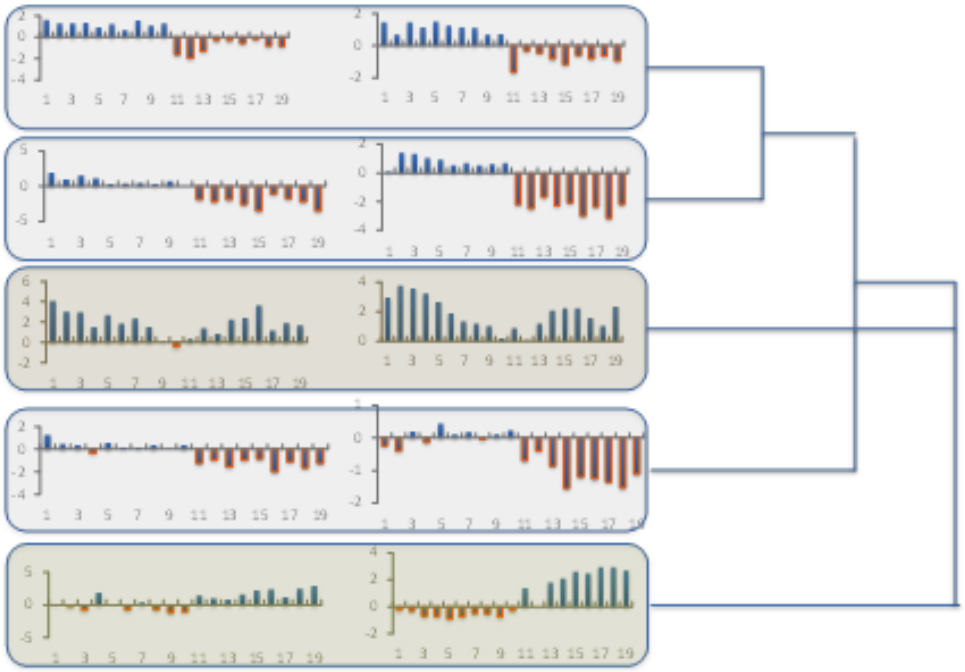


Figure 5.1: Serii de timp bi-dimensionale: exemplu de clustering ierarhică

Clustering pe serii de timp multidimensionale O serie de timp e definită printr-un şir $X = (x_1, x_2, \ldots, x_n)$ de măsurări în timp a unui anumit parametru.

O *serie de timp multidimensională* e reprezentată prin:

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{pmatrix}$$

unde fiecare $X_i, 1 \leq i \leq N$, e o serie de timp separată. Seriile de timp individuale pot avea mărimi diferite.

Clustering-ul seriilor de timp multidimensionale presupune gruparea entităţilor de forma X. Figura 5.1 prezintă cinci entităţi pe care s-a făcut clustering în mod ierarhic ca serii de timp bi-dimensionale.

Clustering-ul seriilor de timp unidimensionale necesită gruparea mai multor puncte de date temporale: de exemplu, clustering-ul unor oraşe în raport cu temperaturile raportate in fiecare oraş timp de doi ani. Această serie e formată din 730 puncte bi-dimensionale (2 ani cîte 365 zile). În cazul seriilor de timp multidimensionale, fiecare

punct de date e format din mai multe serii de timp. De exemplu, ar fi util de aplicat clustering-ul pe oraşele pe baza vitezei vîntului, presiunii, volumului precipitaţiilor, etc. şi nu doar pe baza temperaturilor. Figurile de mai jos prezintă exemple de entităţi cu două serii de timp. Unele entităţi sunt similare comform primei serii de timp, în timp ce altele sunt similare comform celei de-a doua serii de timp. În cazul seriilor de timp multidimensionale, clustering-ul trebuie să grupeze entităţile pe baza similitudinii ambelor serii de timp. Acest exemplu e ilustrat în Figura 5.2.

Abordarea comună în cazul multidimensional este de a concatena toate seriile de timp într-una singură şi de a transforma problema într-o problemă cu o singură dimensiune a seriei de timp. Dar acest lucru poate duce la pierderea aspectelor generale ale problemei. Prin urmare, este avantajos să se trateze seriile temporale multidimensionale fără a le transforma: pe de o parte, ele oferă un punct de vedere global şi prezintă unele patologii critice care decurg din discrepanţe evidente şi, pe de altă parte, permit integrarea informaţiilor conţinute în fiecare serie de timp unidimensională a X şi, prin urmare, este util atunci când fiecare matrice este redusă (15).

5.3 O variantă a clustering-ului pe serii de timp multidimensionale

O măsură de similitudine este de obicei utilizată pentru a calcula asemănarea dintre două serii de timp. În acest capitol, tratăm diferenţa dintre fiecare serie temporală a unei instanţe temporale multidimensionale ca o funcţie obiectivă care trebuie minimizată.

5.3.1 Măsuri de similitudine

Pentru a compara similitudinea între două obiecte X şi Y, unde:

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{pmatrix}, Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix}$$

definim o funcţie N-dimensională $F = (f_1, f_2, \ldots, f_N)$:

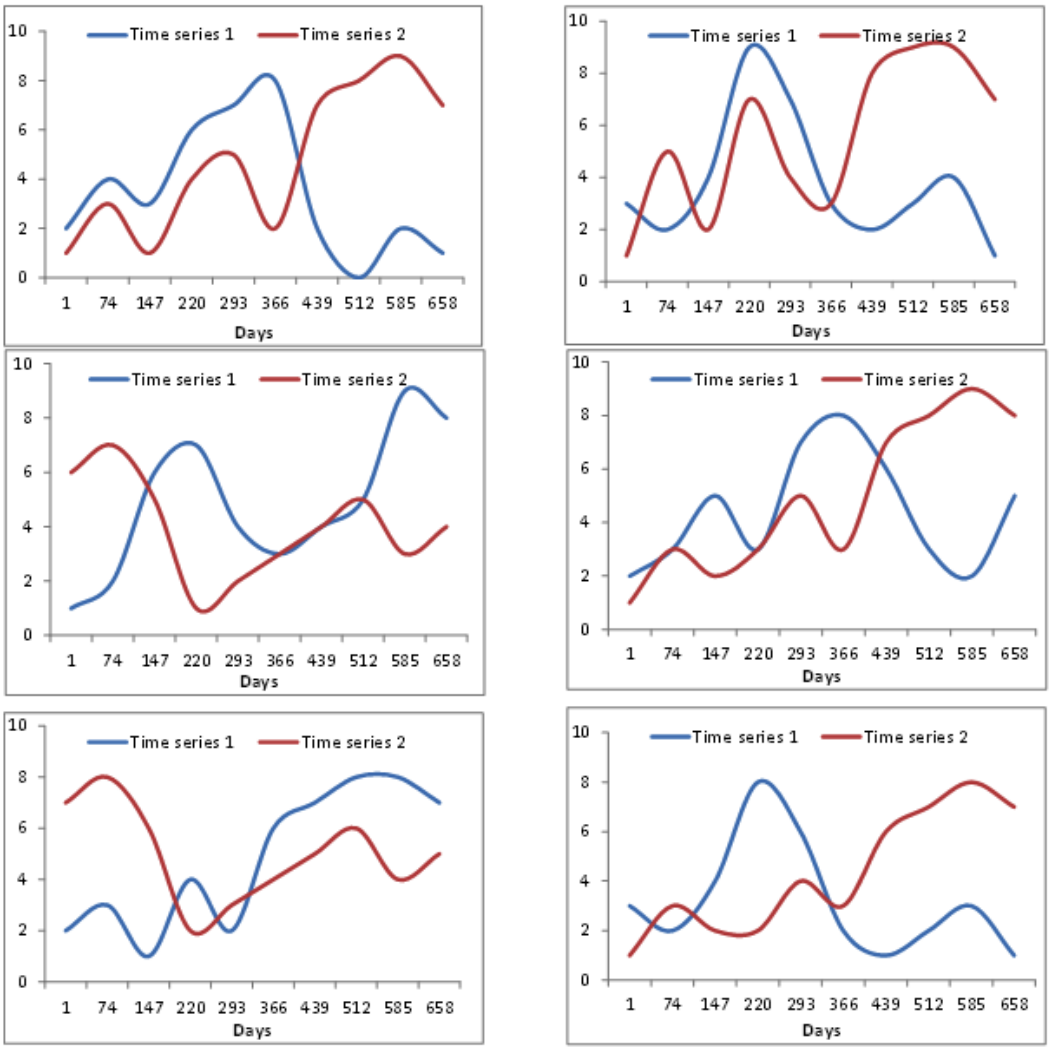


Figure 5.2: Seturile de măsurători pe parcursul celor doi ani

$$F = \begin{pmatrix} f_1 = d(X_1, Y_1) \\ f_2 = d(X_2, Y_2) \\ \vdots \\ f_N = d(X_N, Y_N) \end{pmatrix}$$

unde $d(\cdot)$ defineşte mărimea similitudinii.

Clasificăm seria temporală multidimensională folosind algoritmul de grupare k-means. Utilizăm funcţia F pentru a determina grupul la care fiecare element trebuie atribuit. Calculăm valoarea de similitudine ca o combinaţie liniară a lui N şi vom desemna rezultatul cu d_{sim}:

$$d_{sim} = \sum_{i=1}^{N} w_i f_i$$

unde w_i sunt greutăţile care determină importanţa fiecarei serii de timp în clustering. În experimentele noastre, toate seriile de timp au importanţa egală ($w_i = 1, 1 \leq i \leq N$) şi distanţele implementate sunt:

- Distanţa euclidiană
- Distanţa Manhattan
- Distanţa maximă
- Distanţa medie

5.3.2 Alegerea valorii pentru parametrul k

Utilizatorul poate selecta una din aceste patru măsuri din meniul principal, precum şi procentul maxim de distanţe (setat în mod implicit la 0,6 în experimentele noastre). Valoarea iniţială a k este setată la numărul total de elemente din setul de date. După fiecare repetare, seminţele care nu au nici un punct de date alocat acestora sunt eliminate. Algoritmul trece prin numărul necesar de iteraţii pentru ca grupurile să se stabilizeze (distanţe mici între punctele din acelaşi grup şi distanţe mari între punctele care aparţin la clustere diferite).

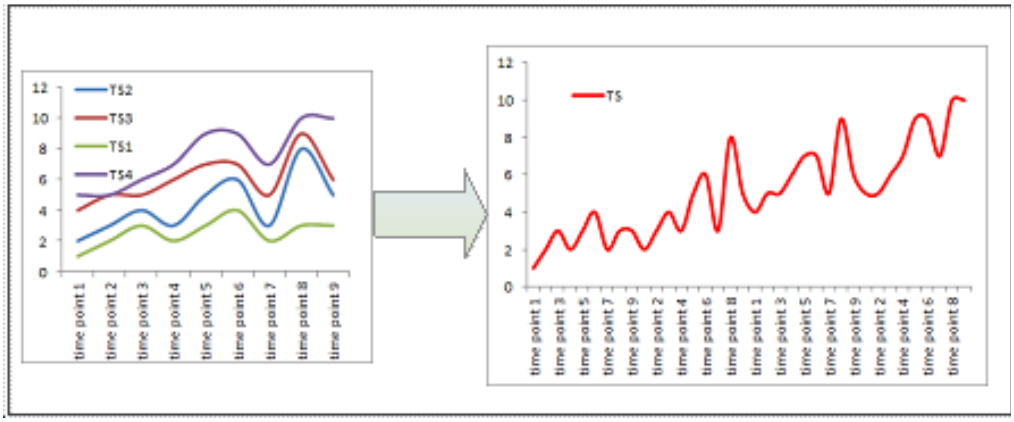


Figure 5.3: Comparaţia cu metodele tradiţionale.

5.4 Compararea cu metodele tradiţionale

Metodele clasice clasifică de obicei datele: deduc un singur parametru care caracterizează toate seriile de timp (16). Deci, în mod esenţial, o serie temporală multidimensională este convertită într-o singură dimensiune, aşa cum arată Figura 5.3. Am implementat şi această tehnică, am comparat rezultatele cu cele obţinute de abordarea noastră şi am constatat că metoda tradiţională tinde de obicei să împartă datele în mai multe clustere decât se aştepta. Acest lucru a fost deosebit de evident atunci când se utilizează valorile de similitudine medie şi Manhattan pe primul set de date sau când se utilizează măsura medie la cel de-al doilea set de date.

5.5 Rezumat

Seriile temporale multidimensionale sunt o generalizare a seriei temporale unidimensionale. Ele au mai mulţi parametri care sunt utilizaţi pentru a descrie fiecare instanţă de date şi, prin urmare, gruparea acestora este mai dificilă decât gruparea seriilor temporale unidimensionale. În acest capitol, am abordat problema de grupare a seriilor de timp multidimensionale ca o problemă multi-obiectiv şi am implementat mai multe măsuri de distanţă geometrică în algoritmul de grupare k-means pentru a testa rezultatul măsurii de similitudine împreună cu procesul de grupare. Am validat rezultatele pe trei seturi de

date şi am ajuns la concluzia că distanţa cea mai frecvent utilizată pentru gruparea seriei temporale unidimensionale (distanţa euclidiană) ar putea să nu fie cea mai potrivită măsură pentru gruparea seriei temporale multidimensionale. Aceste rezultate au fost publicate în cadrul conferinţei Intelligent System Design and Applications (ISDA 2013) (45) şi în Studia Universitatis Babeş Bolyai în 2013 (44).

6

Compararea reţelelor multi-relaţionale

6.1 Introducere

În acest capitol sunt introduse definiţiile şi conceptele legate de reţele multi-relaţionale prin comparaţie cu cele uni-relaţionale. Unele proprietăţi utile ale acestor reţele sunt prezentate şi este descrisă o aplicaţie implementată pentru compararea reţelelor, împreună cu principalele operaţii asupra reţelei. Ultima secţiune analizează experimentele referitoare la compararea reţelelor de interacţiune dintre gene pentru diferite tipuri de cancer. Aceste experimente sunt efectuate pentru a valida ideile şi conceptele introduse în acest capitol.

6.2 Definiţii

Termenul de reţea poate fi interpretat ca un grup de persoane, organizaţii, locuri etc. care sunt conectate sau lucrează împreună (7). În domeniul informaticii, o reţea este descrisă ca un graf mai mare care are un set de vârfuri conectate prin muchii. Prima diferenţă între grafuri simple şi multigrafuri este dată de numărul de interacţiuni dintre noduri: grafurile simple au cel mult o interacţiune (muchie), în timp ce în multigrafuri

pot exista interacţiuni multiple între oricare două noduri. Un multigraf are un set de vârfuri şi o mulţime de perechi neordonate de vârfuri care definesc muchiile. Un multigraf are toate muchiile de acelaşi tip. În cazul reţelelor (biologice), pot exista mai multe tipuri de interacţiuni între două noduri, astfel încât o reţea biologică este o generalizare a multigrafurilor.

Combinând aceste definiţii în bioinformatică, putem spune că o reţea biologică este formată dintr-un set de vârfuri (gene, proteine, metaboliţi etc.) şi un set de muchii reprezentând interacţiunile (de diferite tipuri) între vârfuri.

6.3 Reţele multi-relaţionale

Analiza şi studierea reţelelor a devenit tot mai importantă într-o varietate de domenii, cum ar fi biologia, informatica şi sociologia. În acest capitol ne concentrăm asupra reţelelor biologice. Prin progresele recente în biologie, o cantitate mare de date a devenit disponibilă. Cu toate acestea, datorită dimensiunii acestor date, provocarea constă în a studia aceste date. Trebuie găsită o modalitate de a transforma aceste date în modele mai uşor de înneles. Un astfel de model care s-a dovedit a fi foarte eficient în descrierea relaţiilor complexe şi a interacţiunilor în sistemele biologice este reţeaua biologică.

O reţea este descrisă ca un set de noduri (vârfuri) şi un set de conexiuni (muchii) care au diferite tipuri de relaţii între noduri. În biologie, reţelele sunt folosite pentru a descrie atât structura, cât şi dinamica unui sistem biologic, prin urmare, identificarea reţelelor biologice este esenţială în biologia sistemelor.

Reţelele biologice sunt compuse din entităţi biologice, iar interacţiunile dintre ele, pot fi folosite pentru a descrie multe tipuri diferite de relaţii. Exemple comune de reţele biologice sunt: reţelele de reglementare transcripională (GRN), reţelele metabolice şi biochimice şi reţelele de interacţiune dintre proteine (PPI). Aceste reţele biologice sunt, de obicei, multi-relaţionale: două noduri pot avea mai multe tipuri de interacţiuni între ele, iar aceste interacţiuni sunt reprezentate prin legăturile din reţea. De exemplu, căile metabolice pot fi descrise ca reţele multi-relaţionale. Moleculele sunt nodurile, iar

activităţile enzimei, transducţia semnalului sau reacţiile chimice sunt muchiile (35). Spre deosebire de reţelele uni-relaţionale care permit un singur tip de muchii între vârfuri, reţelele multi-relaţionale sunt mult mai potrivite pentru a reprezenta interacţiunile din lumea reală, dar sunt mai greu de analizat.

6.4 Experimente şi rezultate

6.4.1 Seturi de date pentru compararea reţelelor multi-relaţionale

Aplicaţia nostră a fost testată pe trei seturi de gene care sunt reprezentative pentru cancerul endometrial, ovarian şi cel mamar. Am folosit aplicaţia Genemania pentru a crea reţelele şi a le salva în fişiere pe care le-am transformat în formatul acceptat de aplicaţia noastră: fiecare muchie este reprezentată sub forma $(gena1, gena2, relaţie\ tip)$; unde *gena1* şi *gena2* reprezintă nodurile iar *tipul de relaţie* defineşte tipul de legătură care există între cele două noduri.

6.4.2 Rezultate

Am efectuat experimente pe trei tipuri de cancer: sân, ovarian şi endometrial. Datele de intrare au fost preluate din baza de date pusă la dispoziţie de Institutul Sanger al Institutului de tiin Weizmann (38) şi de alte câteva studii. Interacţiunile pentru cele mai semnificative gene asociate fiecărui tip de cancer au fost extrase din Genemania (4, 35) şi sunt prezentate în figura 6.1 (vizualizarea foloseşte un aspect de cerc sortat gradat şi a fost desenată folosind aplicaţia Cytoscape (38, 42**?**)).

Detaliile pentru fiecare din cele trei reţele sunt prezentate în Table 6.1.

6.5 Rezumat

În acest capitol am prezentat toate definiţiile şi noţiunile de bază necesare pentru extinderea de la o reţea uni-relaţională la o reţea multi-relaţională. Reţelele multi-relaţionale pot fi aplicate pe probleme din mai multe domenii diferite, iar reţelele bio-

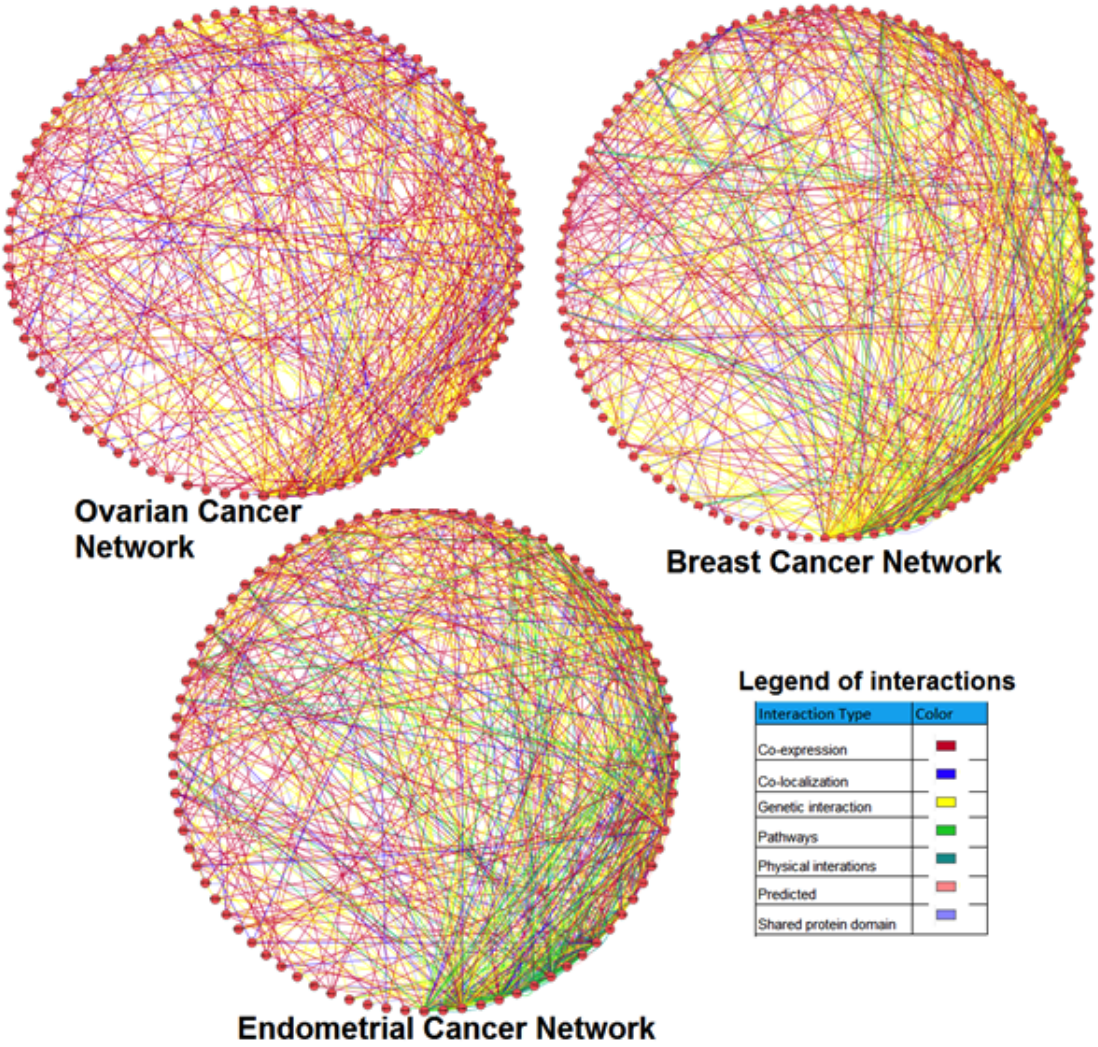


Figure 6.1: Reprezentarea reţelelor de interacţiune dintre gene pentru cele trei tipuri de cancer: ovarian, mamar şi endometrial (aplicaţia Cytoscape (9, 38, 42) a fost folosită pentru a desena reţelele).

	Vancer ovarian	Cancer mamar	Cancer endometrial
Noduri	83	101	83
Muchii	590	817	801
Noduri cu mai multe perechi de muchii	47	74	108
Noduri izolate	0	0	0
Densitatea reţelei	0.157	0.14	0.19
Heterogenitatea reţelei	0.39	0.48	0.41

Table 6.1: Descrierea reţelelor de interactiune dintre genele implicate în cancerul ovarian, mamar şi endometrial.

logice sunt una dintre cele mai comune aplicaţii ale acestora. Capitolul a prezentat, de asemenea, o aplicaţie care este concepută pentru a comparara reţele multi-relaţionale. Au fost efectuate şi descrise experimente privind datele reale legate de diferite tipuri de cancer feminin. Experimentele au fost efectuate pentru trei tipuri de reţele de gene implicate în trei tipuri de cancer similare (ovarian, mamar şi endometrial) şi rezultatele au fost prezentate la Conferinţa Internaională de Design şi Aplicatii a Sistemelor Inteligente (ISDA) (47).

7

Motive în reţele multi-relaţionale

7.1 Introducere

Acest capitol prezintă motive în reţea şi grafuri şi este structurat în trei secţiuni. În prima secţiune sunt prezentate noţiunile de bază despre motive şi grafuri. De asemenea, descriem semnificaţia motivelor şi a grafturilor şi importanţa acestora în biologie. O extindere a motivelor şi al modului de detectare al acestora, în cazul reţelelor multi-relaţonale, este prezentată împreună cu operaţiile de bază asupra motivelor din reţele. Este descrisă o aplicaţie care detectează motive în reţele multi-relaţionale, care este o extensie a unei aplicaţii celebre dezvoltate în laboratorul lui Uri Alon la Institutul Weizmann, descrise în a doua parte a capitolului. Extensia nu se referă doar la reţele multi-relaţionale (extinse din reţelele uni-relaţionale), ci şi la o aplicaţie de căutare a motivelor de către utilozator. Utilizatorul are posibilitatea de a-şi defini propriul motiv şi de a-l căuta în reţea. În ultima parte a acestui capitol sunt efectuate experimente, inclusiv date biologice reale, pentru a valida eficienţa acestei aplicaţii.

7.2 Motive în reţele multi-relaţionale

Cantitatea de informaţie care a devenit disponibilă în ultimii ani a permis studiul unor reţele foarte mari: World Wide Web, reţele sociale şi, în special, reţele biologice,

printe care se numără reţelele de interacţiune dintre proteine (PPI) (20, 23, 24), reţelele metabolice (25) şi reţelele de interacţiune dintre gene (GGI) (31, 32). Aceste reţele aduce o multitudine de date, dar extragerea de informaţii semnificative din aceste date are provocările sale. Trebuie dezvoltaţi algoritmi care sunt eficienţi şi în acelaşi timp robuşti din punct de vedere al erorilor întâlnite în datele analizate.

Din nefericire, datorită complexităţii lor, unele reţele nu pot fi analizate complet. Genomul uman este un exemplu clasic al unei astfel de reţele. Aceste reţele sunt, de obicei, extrem de importante. Prin urmare, obţinerea oricărei noi înţelegeri în comportamentul lor este foarte important. O modalitate de a realiza acest lucru este prin abordarea paralelă cu reţele mai simple, care sunt mai uşor de analizat. Odată ce un subsistem dintr-un organism mai simplu este bine înţeles, această experienţă poate fi aplicată organismului mai complex. Această abordare permite, în esenţă, izolarea şi analizarea în mod independent a subsistemelor mici în organismele mai complexe.

O clasă de astfel de algoritmi este *detectarea motivelor* (32). Aceşti algoritmi urmăresc să detecteze modelele de conectivitate care apar mai des într-o reţea şi să ofere, de asemenea, o imagine asupra modularităţii şi structurii unei reţele (20, 31, 36, 53). S-a demonstrat că aceleaşi motive se găsesc în multe organisme diferite, care le transformă într-un instrument esenţial pentru transferul de cunoştinţe de la organisme simple la subsisteme mai complexe.

Pe lângă aplicaţiile lor în reţelele biologice, motivele au fost aplicate cu succes şi în alte zone.

Reţelele biologice prezintă un volum mare de date, modelate ca reţele de interacţiune, căi metabolice şi de semnalizare, reţele de reglementare etc. Grafurile multi-relaţionale sunt cele mai utilizate pentru a arăta complexitatea şi amploarea lor. Caracteristicile acestui tip de reţele sunt importante în analiza interacţiunilor dintre entităţile biologice. Ele sunt numite motive în grafuri.

În general, reţelele biologice (şi motivele dintr-o reţea) sunt multi-relaţionale. Acest lucru complică procesul de identificare a motivelor.

Există mai multe aplicaţii (29, 52) care, în grafuri simple unidirecţionle şi multidirecţionale,

identifică motive de dimensiuni diferite dintr-o reţea. Aplicaţia propusă de noi detectează motivele din reţelele multi-relaţionale.

Oferim două moduri de identificare a motivelor. În primul rând, *template matching*, utilizatorul poate specfica unul sau mai multe template-uri de grafuri pe care le caută.

În al doilea rând, abordarea de eşantionare a subgrafelor este o variaţie a algoritmului pentru grafurile multi-relaţionale (28). Estimează frecvenţa subgrafurilor de dimensiune N folosind subgrafurile k obţinute în mod aleator din reţea.

7.3 Experimente şi rezultate

7.3.1 Seturi de date pentru detectarea motivelor în reţele multi-relaţionale

Considerăm reţeaua definită astfel:

$node_a$ $node_b$ $edge_1$

$node_a$ $node_d$ $edge_3$

$node_e$ $node_a$ $edge_1$

$node_b$ $node_e$ $edge_2$

$node_d$ $node_a$ $edge_1$

$node_e$ $node_f$ $edge_1$

unde $node_x$ reprezintıa un nod (x este eticheta sa) si $edge_i$ reprezintıa o relaţie între două noduri (i este tipul relaţiei).

7.3.2 Rezultate

Pentru a ilustra modul în care funcţionează aplicaţia MultiMot, am selectat un exemplu de reţea multi-relaţională de interacţiune dintre gene . Am considerat prmiele 50 de gene neregulate ale unui ţesut cardiac de la două specii de şoareci. Datele sunt preluate din baza de date Gene Expression Omnibus (12). Reţelele de interacţiune ale genelor (descrise în figura 7.1) au fost generate utilizând Genemania.

Reţelele de interacţiune dintre gene ale ţesutului cardiac exprimate diferit sunt reprezentate în figura 7.1. Tipurile de legături reprezentate cu culori diferite sunt prezen-

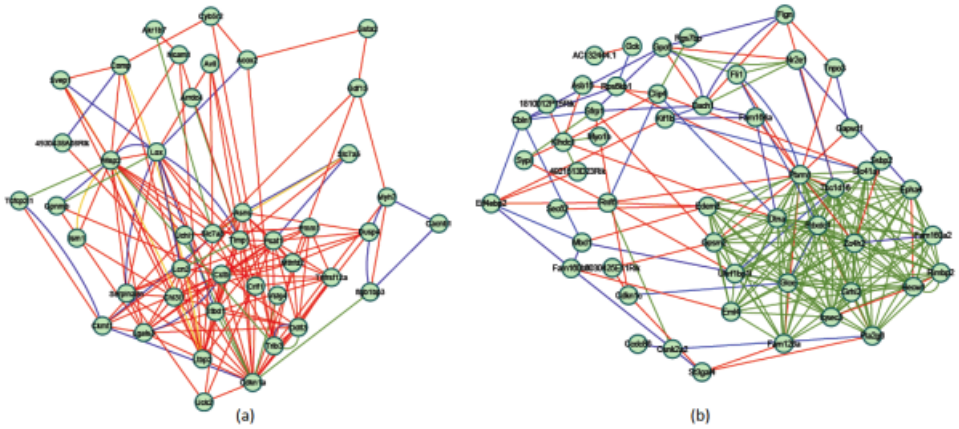


Figure 7.1: Reprezentarea reţelelor de interacţiune dintre gene ale a unui ţesut cardiac exprimat diferenţial.

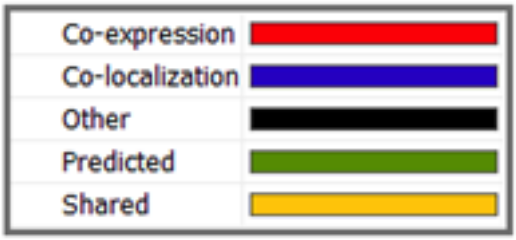


Figure 7.2: Reprezentarea legendei reţelelor de interacţiune dintre gene ale unui ţesut cardiac exprimată diferenţiat.

tate în figura 7.2.

Dacă metoda trebuie să găsească toate motivele de dimensiune 3 în fiecare dintre reţele, atunci rezultatul este cel din Figura 7.3 respectiv Figura 7.4.

Legenda tipurilor de legături sunt reprezentate în Figura 7.2.

Motivele de dimeniune 3 găsite de MultiMot în reţele, Figura 7.2 (a) şi Figura 7.2 (b) sunt reprezentate în Figura 7.3 şi Figura 7.4.

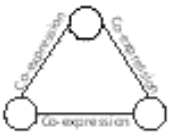
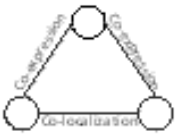
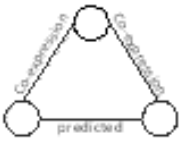
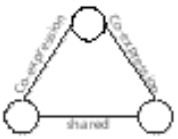
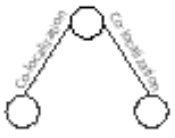
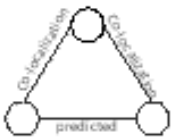
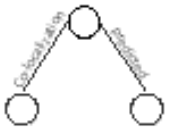
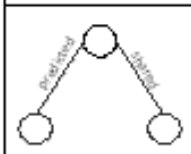
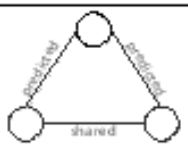
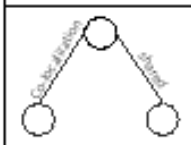
Motif	Network (a)	Network (b)	Motif	Network (a)	Network (b)
	223	22		63	5
	48	69		6	1
	60	19		2	4
	17			3	
	1	7		1	1
	4	19		4	4
	1			1	
	1			1	1

Figure 7.3: Frecvenţa motivelor de dimensiunea 3 găsite de MultiMot în reţelele din Figura 7.1 (a) şi Figura 7.1 (b).

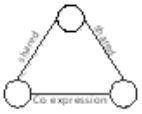
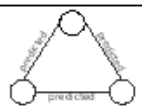
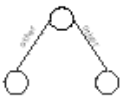
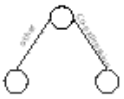
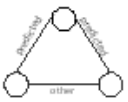
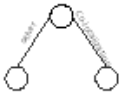
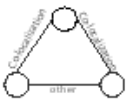
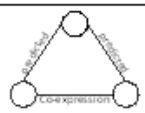
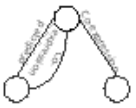
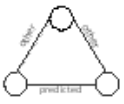
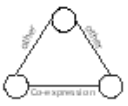
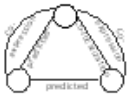
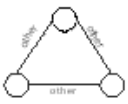
	2	219		1	
	31			0	20
	4			0	6
	9			0	1
	6			0	8
				0	10
	0	1		0	1
	3	0		0	1
	0	1		0	1

Figure 7.4: Frecvenţa motivelor de dimensiunea 3 găsite de MultiMot în reţelele din Figura 7.1 (a) şi Figure 7.1 (b) (continuare).

7.4 Rezumat

Aceast capitol prezintă un algoritm nou de detectare al motivelor în reţele multi-relaţionale.

Identificarea unor modele similare şi frecvent întâlnite (motive în reţea) în reţele oferă informaţii importante pentru o mai bună înţelegere a bolilor şi are un imens impact în studiile biologice. Această lucrare prezintă un algoritm nou de identificare a motivelor, care este o extindere a algoritmului folosit pentru detectarea motivelor în reţele uni-relaţionale şi se ocupă de grafuri eterogene multi-relaionale. Nu numai că găseşte motive standard de o anumită dimensiune, ci caută şi motive customizate (personalizate) (care pot avea un grad mai mare de complexitate). Acesta consideră că atât în grafurile direcţionate, cât şi în cele nedirecţionate, pot fi detectate motive pe un subset de muchii şi pot fi folosite metacaractere pentru etichetarea muchiilor, dacă este necesar, ceea ce la creşterea generalităţii.

8

Comunităţi in reţele multi-relaţionale

8.1 Introducere

Acest capitol oferă o prezentare generală a comunităţilor din reţele multi-relaţionale şi reţele uni-relaţionale şi mai mulţi algoritmi pentru identificarea comunităţilor în reţele uni-relaţionale, respectiv reţele multi-relaţionale. De asemenea, propune un algoritm nou pentru detectarea comunităţilor în reţele multi-relaţionale. În continuare, în acest capitol prezentăm o aplicaţie implementată de noi care are la bază algoritmul pe care l-am propus, precum şi rezultatele obţinute prin rularea algoritmului pe mai multe seturi de date.

8.2 Comunităţi în reţele multi-relaţionale

Reţelele complexe sunt modele matematice care sunt folosite pentru a reprezenta într-o formă citibilă de maşini multe fenomene de interacţiune care au loc în lumea reală. Următoarele reţele sunt exemple de reţele mari:

- World Wide Web: nodurile reprezintă paginile web, iar legăturile reprezintă hyperlink-uri.

- Reţele metabolice: nodurile reprezintă enzime si legăturile reprezintă reacţiile între ele.

- Facebook: nodurile reprezintă utilizatorii si legăturile reprezintă relaţiile de prietenie.

Comunităţile pot fi definite ca grupuri discontinue de entităţi într-un graf, astfel încât fiecare entitate să fie "mai aproape" de toate celelalte entităţi din acelaşi grup decât de entităţile din afara acesteia. Detectarea comunităţilor într-o reţea reprezintă un pas important în identificarea modelelor din acea reţea (10).

Detectarea comunităţilor este foarte importantă în domeniul informaticii şi al biologiei, unde informaţiile sunt reprezentate ca reţele. Cele mai multe reţele în natură sunt multi-relaţionale. Cu toate acestea, detectarea comunităţilor în reţele multi-relaţionale este dificilă, deoarece complexitatea creşte odată cu dimensiunea reţelei şi cu numărul de tipuri de conexiuni. Prin urmare, atunci când căutăm comunităţi, este adesea mai uşor să reducem reţeaua multi-relaţională la o reţea uni-rela]ctională. Acest lucru duce în mod inevitabil la pierderea informaţiilor, iar rezultatele nu sunt întotdeauna exacte.

8.3 Experimente şi rezultate

8.3.1 Seturi de date pentru detectarea comunităţilor în reţele biologice multi-relaţionale - aplicaţia MultiNetCom

Considerăm 11 seturi de date luate din (14) şi din (40). Acestea conţin gene implicate în 11 tipuri de cancer cum ar fi: leucemia limfoblastică acută (ALL), leucemia mieloidă acută (AML), carcinomul mamar (BC), adenocarciomul colorectal (CA), limfomul non-Hodgkin (NHL) Carcinomul spastic (ST), cancerul gastric (GC), cancerul hepatic (LIC), cancerul pulmonar (LUC), cancerul sistemului nervos (NSC). Apoi, utilizăm Genemania (50), o platformă software care generează reţele de interacţiune gene-genă din milioane de publicaţii disponibile în diferite baze de date medicale şi biologice. Obţinem o reţea multi-relaţională pentru fiecare tip de cancer. Genele reprezintă nodurile reţelei

şi interacţiunile dintre gene reprezintă legăturile sau muchiile. În reţelele pe care le-am obţinut, s-au generat următoarele tipuri de interacţiuni: interacţiuni prezise, co-localizare, coexpresie, domenii de proteine partajate, interacţiuni fizice, interacţiuni genetice, interacţiuni pe căi.

Tabelul 8.1 conţine detaliile pentru fiecare reţea: numărul nodurilor, numărul muchiilor şi numărul tipurilor de muchii.

Datele de cancer			
Tip de cancer	Nr noduri	Nr muchii	Tipul muchiilor
ALL	47	260	6
AML	92	822	7
BC	26	163	7
CA	39	411	7
NHL	19	49	6
NSCLC	35	166	6
ST	21	52	5
GC	511	28148	7
LIC	322	9694	7
LUC	74	687	7
NSC	47	20276	7

Table 8.1: Descrierea seturilor de date folosite in experimente.

8.3.2 Rezultate

Realizăm două teste pentru fiecare reţea, separat: pentru prima, reducem reţeaua multi-relaţională la o reţea uni-relaţională şi apoi aplicăm algoritmul iniţial al lui Fruchterman pentru a obţine comunităţile. Pentru al doilea test, considerăm reţeaua multi-relaţională, şi apoi aplicăm versiunea nouă a algoritmului lui Fruchterman propusă de noi în această lucrare, pentru a obţine comunităţile. Figura 8.1 arată numărul de comunităţi şi numărul de gene din fiecare comunitate obţinute de fiecare algoritm pentru cele 11 seturi de date.

Din rezultatele prezentate aici, este evident faptul că folosind cele două abordări obţinem comunităţi diferite. Numărul comunităţilor obţinute pentru reţeaua uni-relaţională este aproape întotdeauna diferit de numărul comunităţilor obţinute de reţeaua multi-

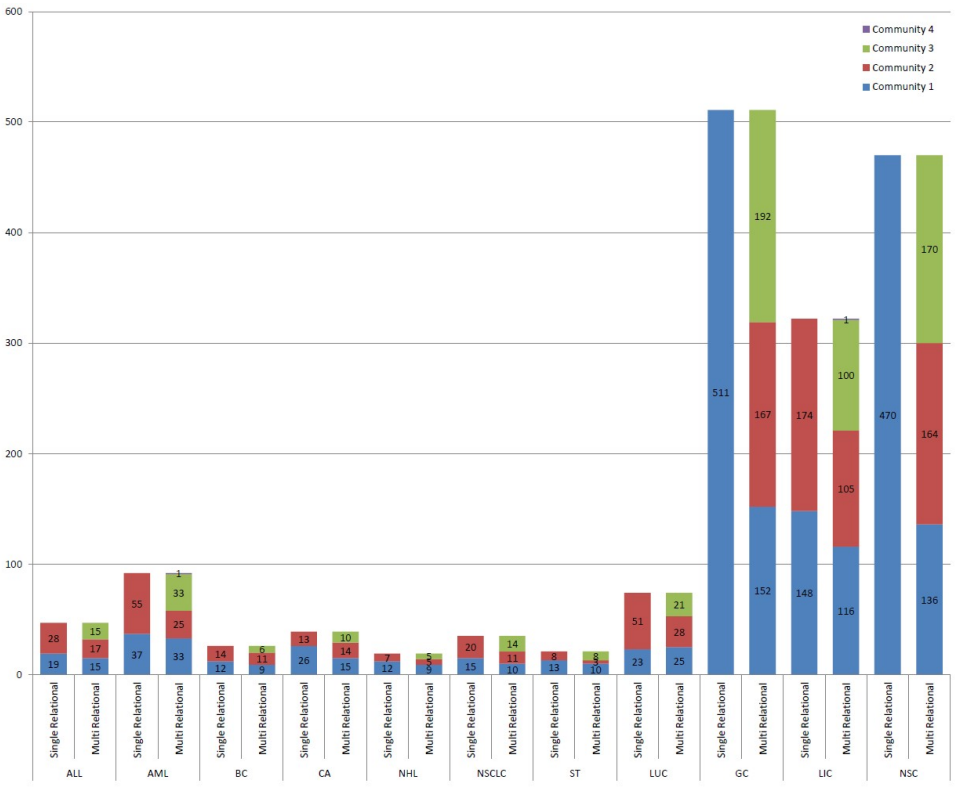


Figure 8.1: Numărul comunităţilor şi genelor în fiecare comunitate obţinută prin una din două metode: reţa uni-relaţională transformată şi reţeaua multi-relaţională originală.

relaţională, aşa cum se poate vedea în figura 8.1. Mai mult decât atât, raportul de suprapunere între oricare două comunităţi este foarte scăzut pentru anumite reţele, fără gene comune (ca în cazul reţelelor ALL, AML, NHL, ST şi LIC) sau doar o genă în comun (ca în cazul Din reţelele AML, BC, CA şi LIC). În unele cazuri, unele comunităţi obţinute în reţeaua multi-relaţională sunt incluse în comunităţile obţinute în reţeaua uni-relaţională. Merită menţionat faptul că testele multi-relaţionale generează întotdeauna mai multe comunităţi decât cele unice.

8.4 Rezumat

Lucrarea prezentată aici are scopul de a demonstra că reţelele multi-relaţionale şi operaţiile computaţionale aplicate acestora trebuie să păstreze aspectul multi-relaţional şi nu pot fi reduse la reţele uni-relaţionale. Transformarea unei reţele multi-relaţionale într-o reţea uni-relaţională poate reduce în mod semnificativ costul calculului, dar duce, de obicei, la pierderea informaţiilor. În această lucrare propunem un algoritm pentru detectarea comunităţilor în reţele multi-relaţionale, dezvoltat iniţial pentru reţele uni-relaţionale. Algoritmul a fost extins la reţele multi-relaţionale, iar comparaţiile dintre cei doi algoritmi au fost efectuate pe 11 seturi de date luate din date implicate în 11 tipuri de cancer care reprezintă reţele genetice de interacţiune (interacţiune genă-gena). Rezultatele arată diferenţe semnificative între cele două abordări care ar putea fi luate în considerare de către cercetătorii care lucrează în domeniul biomedical. Algoritmul pe care l-am propus este unul general, astfel încât orice algoritm (nu doar algoritmul lui Fruchterman prezentat aici) poate fi extins cu uşurinţă de la reţelele uni-relaţionale la cele multi-relaţionale.

În acest capitol se propune un nou algoritm de detectare a comunităţilor pentru reţele multi-relaţionale (46). Am aplicat acest algoritm în reţele biologice multi-relaţionale şi am comparat rezultatele cu cele produse de un algoritm de detectare a comunităţilor similare pentru reţelele uni-relaţionale. Am arătat că algoritmul nostru este capabil să detecteze mai multe comunităţi, în comparaţie cu un algoritmul implementat pentru

reţele uni-relaţionale, deoarece informaţiile se pierd atunci când tratăm reţeaua iniţială multi-relaţională ca fiind o reţea uni-relaţională. Rezultatele prezentate în acest capitol sunt incluse într-un articol trimis la revista PlosOne.

9

Concluzii

Ultimul capitol analizează succint eforturile de dezvoltare şi analiză a modelelor de reţele multi-relaţionale şi a clustering-ului pe serii de timp multidimensionale şi identifică câteva direcţii generale pentru extinderea modelelor dezvoltate în această teză, care pot da perspective suplimentare. În continuare sunt descrise mai multe probleme deschise în analiza reţelelor multi-relaţionale şi gruparea multidimensională a seriilor de timp.

Această teză a prezentat două idei principale care sunt utilizate aici pentru abordarea problemelor biologice. Cu toate acestea, ele nu sunt construite special pentru biologie sau medicină şi pot fi aplicate şi în alte domenii.

Primul subiect se referă la reţele multi-relaţionale (reţele în care există mai multe linkuri şi mai mult de un tip de muchii între două noduri). Ea abordează trei aspecte ale acestor reţele:

- Comparaţia reţelelor multi-dimensionale,
- Căutarea de graphlets şi motive în reţele multi-dimensionale,
- Detectarea comunităţilor în reţele multi-relaţionale.

Provocările întâlnite în această parte a tezei au fost:

- Abordarea unor problem NP-complete (cum ar fi izomorfismul reţelelor),

- Depistarea algoritmilor eficienţi pentru graphlets şi motive,
- Găsirea modului corect de a reprezenta reţelele şi interacţiunile acestora,
- Implementarea măsurilor potrivite pentru compararea reţelelor,
- Detectarea comunităţilor în reţele multi-relaţionale.

Totuşi, încă mai trebuie depuse multe eforturi pentru o comparaţie cuprinzătoare a reţelelor şi pentru implementarea unor algoritmi şi mai eficienţi pentru găsirea motivelor şi detectarea comunităţilor. Am realizat până în prezent, următoarele:

- Proiectarea şi implementarea unei aplicaţii de comparaţie a reţelelor:
  - Este nevoie de două sau mai multe reţele
  - Este conceput pentru reţele de interacţiune genetică şi reţele de interacţiune proteină-proteină, dar este uşor de extensibil şi adaptabil
  - Implementează câteva operaţii de bază, cum ar fi uniunea, intersecţia, subgrafurile comune, distribuţia gradelor
  - Permite selectarea a unul, două, mai multe sau toate tipurile de link-uri existente şi efectuează operaţiuni pentru tipurile de linkuri selectate
- Proiectarea şi implementarea unei aplicaţii pentru detecarea de graphlets şi a motive:
  - Proiectat pentru reţele multi-relaţionale
  - Găseşte motive de o anumită dimensiune dată (sunt implementate mărimi de 3, 4 şi 5 noduri)
  - Găseşte motive customized
  - Lucrează cu grafuri direcţionate şi nedirecţionate (reţele)
- Experimente şi aplicaţii pentru problemele biologice:

  - Pentru reţelele de gene implicate în trei tipuri similare de cancer feminin (ovarian, endometrial şi sân)
  - Compararea reţelelor de interacţiune genetică bazate pe motivele conţinute (adică blocurile de bază)

- Proiectarea şi implementarea unei aplicaţii pentru clustering-ul multidimensional:
  - Entităţile formate din mai multe serii de timp sunt grupate
  - Este efectuată o comparaţie a unei similitudini bazate pe distanţă geometrică

10

Cuvinte cheie

- bioinformatică
- reţele complexe
- reţele cu o singură relaţie
- reţele multi-relaţionale
- reţele biologice multi-relaţionale
- clusters
- clastering multidimensional al seriilor de timp
- motive în graf
- motive în reţele multi-relaţionale
- detectarea comunităţilor
- comunităţi în reţele multi-relaţionale
- reţele de interacţiune genetică
- analiza reţelelor
- aplicaţii soft

11

Referinţele tezei

[1] AARTS/KORST. Simulated annealing and boltzmann machines. A stochastic approach to combinatorial opti- mization and neural computing. John Wiley., 1990.

[2] Gaurav Agarwal and David Kempe. Modularitymaximizing graph communities via mathematical programming. The European Physical Journal B, 66(3):409-418, 2008.

[3] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. Journal of Machine Learning Research, 9(Sep):1981-2014, 2008.

[4] Reka Albert and Albert-Laszlo Barabasi. Statistical mechanics of complex networks. Reviews of modern physics, 74(1):47, 2002.

[5] Reka Albert, Hawoong Jeong, and Albert-Laszlo Barabasi. Error and attack tolerance of complex networks. arXiv preprint cond-mat/0008064, 2000.

[6] Charles J Alpert and So-Zen Yao. Spectral partitioning: the more eigenvectors, the better. In Proceedings of the 32nd annual ACM/IEEE Design Automation Conference, pages 195-200. ACM, 1995.

[7] Khaled Alsabti, Sanjay Ranka, and Vineet Singh. An eficient k-means clustering

algorithm. 1997.

[8] Michael R Anderberg. Cluster analysis for applications.monographs and textbooks on probability and mathematical statistics, 1973.

[9] E Michael Azoff. Neural network time series forecasting of financial markets. John Wiley and Sons, Inc., 1994.

[10] Gary D Bader, Doron Betel, and Christopher WV Hogue. Bind: the biomolecular interaction network database. Nucleic acids research, 31(1):248-250, 2003.

[11] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. Nature Reviews Genetics, 5(2):101-113, 2004.

[12] Alain Barrat, Marc Barthelemy, Romualdo Pastor- Satorras, and Alessandro Vespignani. The architecture of complex weighted networks. Proceedings of the Na- tional Academy of Sciences of the United States of America, 101(11):3747-3752, 2004.

[13] Federico Battiston, Jacopo Iacovacci, Vincenzo Nicosia, Ginestra Bianconi, and Vito Latora. Emergence of multiplex communities in collaboration networks. PloS one, 11(1):e0147451, 2016.

[14] Anais Baudot, Gonzalo Gomez-Lopez, and Alfonso Valencia. Translational disease interpretation with molecular networks. Genome biology, 10(6):221, 2009.

[15] Laura Bennett, Aristotelis Kittas, Gareth Muirhead, Lazaros G Papageorgiou, and Sophia Tsoka. Detection of composite communities in multiplex biological networks. Scientific reports, 5, 2015.

[16] Laura Bennett, Songsong Liu, Lazaros G Papageorgiou, and Sophia Tsoka. Detection of disjoint and overlapping modules in weighted complex networks. Advances in Complex Systems, 15(05):1150023, 2012.

[17] Michele Berlingerio, Michele Coscia, Fosca Giannotti, Anna Monreale, and Dino

Pedreschi. Foundations of multidimensional network analysis. In Advances in So- cial Networks Analysis and Mining (ASONAM), 2011 In- ternational Conference on, pages 485-489. IEEE, 2011.

[18] Bela Bollobas, Gautam Das, Dimitrios Gunopulos, and Heikki Mannila. Time-series similarity problems and well-separated geometric sets. In Proceedings of the thirteenth annual symposium on Computational geometry, pages 454-456. ACM, 1997.

[19] David E Booth. Time series, 1992.

[20] Stefan Bornholdt and Heinz Georg Schuster. Handbook of graphs and networks: from the genome to the internet. John Wiley and Sons, 2006.

[21] Laurie A Boyer, Tong Ihn Lee, Megan F Cole, Sarah E Johnstone, Stuart S Levine, Jacob P Zucker, Matthew G Guenther, Roshan M Kumar, Heather L Murray, Richard G Jenner, et al. Core transcriptional regulatory circuitry in human embryonic stem cells. cell, 122(6):947-956, 2005.

[22] Ulrik Brandes. A faster algorithm for betweenness centrality*. Journal of Mathematical Sociology, 25(2):163- 177, 2001.

[23] Rainer Breitling, David Gilbert, Monika Heiner, and Richard Orton. A structured approach for the engineering of biochemical network models, illustrated for signalling pathways. Briefings in Bioinformatics, 9(5):404-421, 2008.

[24] Karin Breuer, Amir K Foroushani, Matthew R Laird, Carol Chen, Anastasia Sribnaia, Raymond Lo, Geoffrey L Winsor, Robert EW Hancock, Fiona SL Brinkman, and David J Lynn. Innatedb: systems biology of innate immunity and beyondrecent updates and continuing curation. Nucleic acids research, 41(D1):D1228-D1233, 2013.

[25] Piotr Brodka, Tomasz Filipowski, and Przemys law Kazienko. An introduction to community detection in multi-layered social network. In World Summit on Knowledge Society, pages 185-190. Springer, 2011.

[26] Deng Cai, Zheng Shao, Xiaofei He, Xifeng Yan, and Jiawei Han. Community mining from multi-relational networks. In European Conference on Principles of Data Mining and Knowledge Discovery, pages 445-452. Springer, 2005.

[27] Muffy Calder, Stephen Gilmore, and Jane Hillston. Modelling the influence of rkip on the erk signalling pathway using the stochastic process algebra pepa. In Transactions on computational systems biology VII, pages 445-452. Springer, 2006.

[28] Melissa S Cline, Michael Smoot, Ethan Cerami, Allan Kuchinsky, Nerius Landys, Chris Workman, Rowan Christmas, Iliana Avila-Campilo, Michael Creech, Benjamin Gross, et al. Integration of biological networks and gene expression data using cytoscape. Nature pro- tocols, 2(10):2366-2382, 2007.

[29] Michele Coscia, Fosca Giannotti, and Dino Pedreschi. A classification for community discovery methods in complex networks. Statistical Analysis and Data Mining, 4(5):512-546, 2011.

[30] Nello Cristianini, John Shawe-Taylor, and Jaz S Kandola. Spectral kernel methods for clustering. In Advances in neural information processing systems, pages 649-655, 2002.

[31] Gautam Das, Dimitrios Gunopulos, and Heikki Mannila. Finding similar time series. In European Symposium on Principles of Data Mining and Knowledge Discovery, pages 88-100. Springer, 1997.

[32] Darcy A Davis and Nitesh V Chawla. Exploring and exploiting disease interactions from multi-relational gene and phenotype networks. PloS one, 6(7):e22670, 2011.

[33] Kara Dolinski, Andrew Chatr-aryamontri, and Mike Tyers. Systematic curation of protein and genetic interaction data for computable biology. BMC biology, 11(1):43, 2013.

[34] Benjamin S Duran and Patrick L Odell. Cluster analysis: a survey. Springer-Verlag

New York, 1974.

[35] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. Nucleic acids research, 30(1):207-210, 2002.

[36] William H Elliott, Daphne C Elliott, John R Jefferson, and John Wheldrake. Biochemistry and molecular biology. Oxford University Press Oxford, 1997.

[37] Graphical Enumeration. F. harary and em palmer, 1973.

[38] Paul Erdos and Alfred Renyi. On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci, 5(1):17- 60, 1960.

[39] Brian S Everitt. Unresolved problems in cluster analysis. Biometrics, pages 169-181, 1979.

[40] http://cancer.sanger.ac.uk/census/.

[41] Santo Fortunato. Community detection in graphs. Physics reports, 486(3):75-174, 2010.

[42] Andrea Franceschini, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer Bork, Christian von Mering, et al. String v9. 1: protein-protein interaction networks, with increased coverage and integration. Nucleic acids research, 41(D1):D808-D815, 2013.

[43] Marco Franciosi and Giulia Menconi. Multi-dimensional sparse time series: feature extraction. arXiv preprint arXiv:0803.0405, 2008.

[44] Thomas MJ Fruchterman and Edward M Reingold. Graph drawing by force-directed placement. Software: Practice and experience, 21(11):1129-1164, 1991.

[45] Sylvia Fruhwirth-Schnatter, Christoph Pamminger, Rudolf Winter-Ebmer, and Andrea Weber. Model-based clustering of categorical time series with multinomial logit

classification. In AIP Conference Proceedings, volume 1281, pages 1897-1900. AIP, 2010.

[46] Guojun Gan, Chaoqun Ma, and JianhongWu. Data clus- tering: theory, algorithms, and applications, volume 20. Siam, 2007.

[47] Michael R Garey and David S Johnson. Crossing number is np-complete. SIAM Journal on Algebraic Discrete Methods, 4(3):312-316, 1983.

[48] Claude Gerard, Didier Gonze, and Albert Goldbeter. Effect of positive feedback loops on the robustness of oscillations in the network of cyclin-dependent kinases driving the mammalian cell cycle. FEBS Journal, 279(18):3411-3431, 2012.

[49] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. Proceedings of the national academy of sciences, 99(12):7821-7826, 2002.

[50] Roger Guimera and Luis A Nunes Amaral. Functional cartography of complex metabolic networks. Nature, 433(7028):895-900, 2005.

[51] Roger Guimera, Marta Sales-Pardo, and Luis A Nunes Amaral. A network-based method for target selection in metabolic networks. Bioinformatics, 23(13):1616-1622, 2007.

[52] Dimitrios Gunopulos and Gautam Das. Time series similarity measures (tutorial pm-2). In Tutorial notes of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 243-307. ACM, 2000.

[53] Jing-Dong J Han, Nicolas Bertin, Hao Tong, Debra S Goldberg, et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature, 430(6995):88, 2004.

[54] Frank Harary and Edgar M Palmer. Graphical enumeration. Technical report,

DTIC Document, 1973.

[55] John A Hartigan. Clustering algorithms. John Wiley and Sons, Inc., 1975.

[56] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1):100-108, 1979.

[57] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. Social networks, 5(2):109-137, 1983.

[58] William M Holmes. Time series: Sir maurice kendall and j. keith ord, (edward arnold, great britain, 1990) pp. 296, 1992.

[59] Thomas C Ings, Jose M Montoya, Jordi Bascompte, Nico Bluthgen, Lee Brown, Carsten F Dormann, Francois Edwards, David Figueroa, Ute Jacob, J Iwan Jones, et al. Review: Ecological networks-beyond food webs. Journal of Animal Ecology, 78(1):253-269, 2009.

[60] Kuhn Ip, Caroline Colijn, and Desmond S Lun. Analysis of complex metabolic behavior through pathway decomposition. BMC systems biology, 5(1):91, 2011.

[61] Negin Iranfar, Danny Fuller, and William F Loomis. Transcriptional regulation of post-aggregation genes in dictyostelium by a feed-forward loop involving gbf and lagc. Developmental biology, 290(2):460-469, 2006.

[62] Shalev Itzkovitz, Reuven Levitt, Nadav Kashtan, Ron Milo, Michael Itzkovitz, and Uri Alon. Coarse-graining and self-dissimilarity of complex networks. Physical Re- view E, 71(1):016127, 2005.

[63] Ariel Jaimovich, Gal Elidan, Hanah Margalit, and Nir Friedman. Towards an integrated protein-protein interaction network: A relational markov network approach. Journal of Computational Biology, 13(2):145-164, 2006.

[64] Anil K Jain and Richard C Dubes. Algorithms for clus- tering data. Prentice-Hall, Inc., 1988.

[65] E Ferrell James Jr. Feedback loops and reciprocal regulation: recurring motifs in the systems biology of the cell cycle. Current Opinion in Cell Biology, 25(6):676- 686, 2013.

[66] Hawoong Jeong, Sean P Mason, Albert-Laszlo Barabasi, and Zoltan N Oltvai. Lethality and centrality in protein networks. arXiv preprint cond-mat/0105306, 2001.

[67] Hawoong Jeong, Balint Tombor, Reka Albert, Zoltan N Oltvai, and A-L Barabasi. The large-scale organization of metabolic networks. Nature, 407(6804):651-654, 2000.

[68] Stephen C Johnson. Hierarchical clustering schemes. Psychometrika, 32(3):241-254, 1967.

[69] Bjorn H Junker and Falk Schreiber. Analysis of biological networks, volume 2. John Wiley and Sons, 2008.

[70] Minoru Kanehisa. A database for post-genome analysis. Trends in genetics: TIG, 13(9):375, 1997.

[71] Minoru Kanehisa. The kegg database. In Novartis Found Symp, volume 247, pages 91-101, 2002.

[72] Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, Yasushi Okuno, and Masahiro Hattori. The kegg resource for deciphering the genome. Nucleic acids re- search, 32(suppl 1):D277-D280, 2004. 23

[73] Ravi Kannan, Santosh Vempala, and Adrian Vetta. On clusterings: Good, bad and spectral. Journal of the ACM (JACM), 51(3):497-515, 2004.

[74] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. An eficient k-means clustering algorithm: Analysis

and implementation. IEEE transactions on pattern analysis and machine intelligence, 24(7):881-892, 2002.

[75] Nadav Kashtan, Shalev Itzkovitz, Ron Milo, and Uri Alon. Mfinder tool guide. Department of Molecular Cell Biology and Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot Israel, Tech. Rep, 2002.

[76] Nadav Kashtan, Shalev Itzkovitz, Ron Milo, and Uri Alon. Eficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. Bioinformatics, 20(11):1746-1758, 2004.

[77] Leonard Kaufman and Peter J Rousseeuw. Finding groups in data: an introduction to cluster analysis, volume 344. John Wiley and Sons, 2009.

[78] Jinki Kim and Gwan-Su Yi. Rmod: A tool for regulatory motif detection in signaling network. PloS one, 8(7):e68407, 2013.

[79] Mikko Kivela, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. Journal of complex networks, 2(3):203-271, 2014.

[80] Mark Kozdoba and Shie Mannor. Community detection via measure space embedding. In Advances in Neural Information Processing Systems, pages 2890-2898, 2015.

[81] Cho Kwang-Hyun, Shin Sung-Young, Kim Hyun-Woo, Olaf Wolkenhauer, Brian McFerran, and Walter Kolch. Mathematical modeling of the influence of rkip on the erk signaling pathway. In Computational methods in sys- tems biology, pages 127-141. Springer, 2003.

[82] Godfrey N Lance and William Thomas Williams. A general theory of classi
catory sorting strategies 1. hierarchical systems. The computer journal, 9(4):373-380, 1967.

[83] Tong Ihn Lee, Nicola J Rinaldi, Francois Robert, Duncan T Odom, Ziv Bar-Joseph,

Georg K Gerber, Nancy M Hannett, Christopher T Harbison, Craig M Thompson, Itamar Simon, et al. Transcriptional regulatory networks in saccharomyces cerevisiae. science, 298(5594):799-804, 2002.

[84] Xutao Li, Michael K Ng, and Yunming Ye. Multicomm: Finding community structure in multi-dimensional networks. IEEE Transactions on Knowledge and Data Engineering, 26(4):929-941, 2014.

[85] T Warren Liao. Clustering of time series dataa survey. Pattern recognition, 38(11):1857-1874, 2005.

[86] Luana Licata, Leonardo Briganti, Daniele Peluso, Livia Perfetto, Marta Iannuccelli, Eugenia Galeota, Francesca Sacco, Anita Palma, Aurelio Pio Nardozza, Elena Santonico, et al. Mint, the molecular interaction database: 2012 update. Nucleic acids research, 40(D1):D857-D861, 2012.

[87] M. Lichman. UCI machine learning repository, 2013.

[88] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. The global k-means clustering algorithm. Pattern recognition, 36(2):451-461, 2003.

[89] Azi Lipshtat, Sudarshan P Purushothaman, Ravi Iyengar, and Avi Maayan. Functions of bifans in context of multiple regulatory motifs in signaling networks. Biophysical journal, 94(7):2566-2579, 2008.

[90] James MacQueen et al. Some methods for classi

cation and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, volume 1, pages 281-297. Oakland, CA, USA., 1967.

[91] Shmoolik Mangan and Uri Alon. Structure and function of the feed-forward loop network motif. Proceedings of the National Academy of Sciences, 100(21):11980-11985, 2003.

[92] Sergei Maslov, Kim Sneppen, and Uri Alon. Correlation profiles and motifs in complex networks. Handbook of Graphs and Networks: From the Genome to the Internet, pages 168-198, 2003.

[93] Louis L McQuitty. Elementary linkage analysis for isolating orthogonal and oblique types and typal relevancies. Educational and Psychological Measurement, 1957.

[94] Marina Meila and Jianbo Shi. Learning segmentation by random walks. In NIPS, volume 14, 2000.

[95] Manuel Middendorf, Etay Ziv, and Chris H Wiggins. Inferring network mechanisms: the drosophila melanogaster protein interaction network. Proceedings of the National Academy of Sciences of the United States of America, 102(9):3192-3197, 2005.

[96] Ron Milo, Shalev Itzkovitz, Nadav Kashtan, Reuven Levitt, Shai Shen-Orr, Inbal Ayzenshtat, Michal Sheffer, and Uri Alon. Superfamilies of evolved and designed networks. Science, 303(5663):1538-1542, 2004.

[97] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. Science, 298(5594):824-827, 2002.

[98] Alexander Y Mitrophanov and Eduardo A Groisman. Positive feedback in cellular control systems. Bioessays, 30(6):542-555, 2008.

[99] Sara Mostafavi, Debajyoti Ray, David Warde-Farley, Chris Grouios, Quaid Morris, et al. Genemania: a realtime multiple association network integration algorithm for predicting gene function. Genome Biol, 9(Suppl 1):S4, 2008.

[100] Panagiotis Moulos, Julie Klein, Simon Jupp, Robert Stevens, Jean-Loup Bascands, and Joost P Schanstra. The kupnetviz: a biological network viewer for multipleomics datasets in kidney diseases. BMC bioinformatics, 14(1):235, 2013.

[101] KA Abdul Nazeer and MP Sebastian. Improving the accuracy and efficiency of the

k-means clustering algorithm. In Proceedings of the World Congress on Engineering, volume 1, pages 1-3, 2009. 109 [102] Mark EJ Newman. Models of the small world. Journal of Statistical Physics, 101(3):819-841, 2000.

[103] Mark EJ Newman. Clustering and preferential attachment in growing networks. Physical review E, 64(2):025102, 2001.

[104] Mark EJ Newman. The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences, 98(2):404-409, 2001.

[105] Mark EJ Newman. The structure and function of complex networks. SIAM review, 45(2):167-256, 2003.

[106] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. Physical review E, 69(2):026113, 2004.

[107] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In Ad- vances in neural information processing systems, pages 849-856, 2002 [108] Michaek Kwok-Po Ng, Xutao Li, and Yunming Ye. Multirank: co-ranking for objects and relations in multirelational data. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1217-1225. ACM, 2011.

[109] BO Palsson. Properties of reconstructed networks. Cambridge: Systems Biology, 2006.

[110] Tony Pawson and Rune Linding. Network medicine. FEBS letters, 582(8):1266-1270, 2008.

[111] Matteo Pellegrini, Edward M Marcotte, Michael J Thompson, David Eisenberg, and Todd O Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proceedings of the Na- tional Academy of Sciences, 96(8):4285-4288, 1999.

[112] Steven J Phillips. Acceleration of k-means and related clustering algorithms. In Algorithm Engineering and Experiments, pages 166-177. Springer, 2002.

[113] V Pihur, Somnath Datta, and Susmita Datta. Finding common genes in multiple cancer types through meta- analysis of microarray experiments: A rank aggregation approach. Genomics, 92(6):400-403, 2008.

[114] Ivan Plavec, Oksana Sirenko, Sylvie Privat, Yuker Wang, Maya Dajee, Jennifer Melrose, Brian Nakao, Evangelos Hytopoulos, Ellen L Berg, and Eugene C Butcher. Method for analyzing signaling networks in complex cellular systems. Proceedings of the National Academy of Sciences of the United States of America, 101(5):1223-1228, 2004.

[115] Teresa Przytycka. An important connection between network motifs and parsimony models. In Research in Computational Molecular Biology, pages 321-335. Springer, 2006.

[116] Marcos G Quiles, Elbert EN Macau, and Nicolas Rubido.Dynamical detection of network communities. Scientific reports, 6, 2016.

[117] Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi. Defining and identifying communities in networks. Proceedings of the National Academy of Sciences of the United States of America, 101(9):2658-2663, 2004.

[118] Silvia Rausanu, Crina Grosan, Zujian Wu, Ovidiu Parvu, Ramona Stoica, and David Gilbert. Computational models for inferring biochemical networks. Neural Computing and Applications, 26(2):299-311, 2015.

[119] Erzsebet Ravasz, Anna Lisa Somera, Dale A Mongru, Zoltan N Oltvai, and A-L Barabasi. Hierarchical organization of modularity in metabolic networks. science, 297(5586):1551-1555, 2002.

[120] F James Rohlf. 12 single-link clustering algorithms. Handbook of Statistics, 2:267-284, 1982.

[121] Michal Ronen, Revital Rosenberg, Boris I Shraiman, and Uri Alon. Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. Proceedings of the national academy of sciences, 99(16):10555-10560, 2002.

[122] Jill A Rosenfeld, Dina Amrom, Eva Andermann, Frederick Andermann, Martin Veilleux, Cynthia Curry, Jamie Fisher, Stephen Deputy, Arthur S Aylsworth, Cynthia M Powell, et al. Genotype-phenotype correlation in interstitial 6q deletions: a report of 12 new cases. neurogenetics, 13(1):31-47, 2012.

[123] Martin Rosvall and Carl T Bergstrom. Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences, 105(4):1118-1123, 2008.

[124] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20:53-65, 1987.

[125] Louis A Saddic, Barbel Huvermann, Staver Bezhani, Yanhui Su, Cara M Winter, Chang Seob Kwon, Richard P Collum, and Doris Wagner. The leafy target lmi1 is a meristem identity regulator and acts together with leafy to regulate expression of cauliflower. Development, 133(9):1673-1682, 2006.

[126] Rintaro Saito, Michael E Smoot, Keiichiro Ono, Johannes Ruscheinski, Peng-Liang Wang, Samad Lotia, Alexander R Pico, Gary D Bader, and Trey Ideker. A travel guide to cytoscape plugins. Nature methods, 9(11):1069-1076, 2012.

[127] Regina Samaga and Steffen Klamt. Modeling approaches for qualitative and semi-quantitative analysis of cellular signaling networks. Cell communication and signaling, 11(1):43, 2013.

[128] S Schuster, DA Fell, and T Dandekar. A general definition of metabolic pathways

useful for systematic organization and analysis of complex metabolic networks. Nature Biotechnology, 18(3):326-332, 2000.

[129] http://ccgd-starrlab.oit.umn.edu/search.php.

[130] Jimmy Shadbolt and John Gerald Taylor. Neural Net- works and the Financial Markets: Predicting, Combining, and Portfolio Optimisation. Springer, 2002.

[131] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research, 13(11):2498-2504, 2003.

[132] Junming Shao, Zhichao Han, and Qinli Yang. Community detection via local dynamic interaction. arXiv preprint arXiv:1409.7978, 2014.

[133] Peter HA Sneath. The application of computers to taxonomy. Journal of general microbiology, 17(1):201-226, 1957.

[134] Daniel A Spielmat and Shang-Hua Teng. Spectral partitioning works: Planar graphs and

nite element meshes. In Foundations of Computer Science, 1996. Proceedings., 37th Annual Symposium on, pages 96-105. IEEE, 1996.

[135] Jorg Stelling, Steffen Klamt, Katja Bettenbrock, Stefan Schuster, and Ernst Dieter Gilles. Metabolic network structure determines key aspects of functionality and regulation. Nature, 420(6912):190-193, 2002.

[136] RAMONA STOICA. Multiobjective approach of multidimensional time series clustering. Studia Universitatis Babes-Bolyai, Informatica, 59(1), 2014.

[137] Ramona Stoica, Mihaela Ola, and Mihai Paraschivescu. Multidimensional temporal clustering: geometrical similarity measures analysis in k-means. Intelligent Systems

Design and Applications, 2013.

[138] Ramona Stoica, Bazil Parv, and Crina Grosan. Communities detection in multi-relational networks. PlosOne, 26(2):299-311, 2017.

[139] Ramona Stoica and Liviu Stirb. A tool for comparing multirelational networks from biology. In Intelligent Sys- tems Design and Applications (ISDA), 2013 13th International Conference on, pages 242-246. IEEE, 2013.

[140] Gemma Swiers, Roger Patient, and Matthew Loose. Genetic regulatory networks programming hematopoietic stem cells and erythroid lineage specification. Developmental biology, 294(2):525-540, 2006.

[141] Damian Szklarczyk, Andrea Franceschini, Michael Kuhn, Milan Simonovic, Alexander Roth, Pablo Minguez, Tobias Doerks, Manuel Stark, Jean Muller, Peer Bork, et al. The string database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic acids research, 39(suppl 1):D561-D568, 2011.

[142] Bosiljka Tadic, Miroslav Andjelkovic, Biljana Mileva Boshkoska, and Zoran Levnajic. Algebraic topology of multi-brain connectivity networks reveals dissimilarity in functional patterns during spoken communications. PloS one, 11(11):e0166787, 2016.

[143] Bing Tian Dai, Freddy Chong Tat Chua, and Ee-Peng Lim. Structural analysis in multi-relational social networks.In Proceedings of the 2012 SIAM International Conference on Data Mining, pages 451-462. SIAM, 2012.

[144] Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. Random walk with restart: fast solutions and applications. Knowledge and Information Systems, 14(3):327-346, 2008.

[145] Ruey S Tsay. Analysis of

nancial time series. Financial econometrics, a wiley-interscience publication, 2002.

[146] Joshua R Tyler, Dennis M Wilkinson, and Bernardo A Huberman. E-mail as

spectroscopy: Automated discovery of community structure within organizations. The Information Society, 21(2):143-153, 2005.

[147] Ikuo Uchiyama. Mbgd: microbial genome database for comparative analysis. Nucleic acids research, 31(1):58- 62, 2003.

[148] Olga Vechtomova. Introduction to information retrieval christopher d. manning, prabhakar raghavan, and hinrich schutze (stanford university, yahoo! research, and university of stuttgart) cambridge: Cambridge university press, 2008, xxi+ 482 pp; hardbound, isbn 978-0- 521-86571-5, 2009.

[149] David Warde-Farley, Sylva L Donaldson, Ovi Comes, Khalid Zuberi, Rashad Badrawi, Pauline Chao, Max Franz, Chris Grouios, Farzana Kazi, Christian Tannus Lopes, et al. The genemania prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic acids research, 38(suppl 2):W214-W220, 2010.

[150] Duncan J Watts and Steven H Strogatz. Collective dynamics of small-worldnetworks. nature, 393(6684):440- 442, 1998.

[151] Yen-Chuen Wei and Chung-Kuan Cheng. Towards efficient hierarchical designs by ratio cut partitioning. In Computer-Aided Design, 1989. ICCAD-89. Digest of Technical Papers., 1989 IEEE International Conference on, pages 298-301. IEEE, 1989.

[152] Sebastian Wernicke. Efficient detection of network motifs. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 3(4), 2006.

[153] Dennis M Wilkinson and Bernardo A Huberman. A method for Finding communities of related genes. pro- ceedings of the national Academy of sciences, 101(suppl 1):5241-5248, 2004.

[154] WT Williams and JM t Lambert. Multivariate methods in plant ecology: V. similarity analyses and information-analysis. The Journal of Ecology, pages 427-445, 1966.

[155] Elisabeth Wong, Brittany Baur, Saad Quader, and Chun-Hsi Huang. Biological network motif detection: principles and practice. Briefings in bioinformatics, 13(2):202-215, 2011.

[156] Elisabeth A Wong and Brittany Baur. On network tools for network motif nding: a survey study, 2010.

[157] Zhiang Wu, Zhan Bu, Jie Cao, and Yi Zhuang. Discovering communities in multi-relational networks. In User Community Discovery, pages 75-95. Springer, 2015.

[158] Stefan Wuchty. Scale-free behavior in protein domain networks. Molecular biology and evolution, 18(9):1694-1702, 2001.

[159] Ioannis Xenarios, Danny W Rice, Lukasz Salwinski, Marisa K Baron, Edward M Marcotte, and David Eisenberg. Dip: the database of interacting proteins. Nucleic acids research, 28(1):289-291, 2000.

[160] Gang Xu, Laura Bennett, Lazaros G Papageorgiou, and Sophia Tsoka. Module detection in complex networks using integer optimisation. Algorithms for Molecular Biology, 5(1):36, 2010.

[161] Yoshihiro Yamanishi, Michihiro Araki, Alex Gutteridge, Wataru Honda, and Minoru Kanehisa. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics, 24(13):i232-i240, 2008.

[162] Muhammed A Yildirim, Kwang-Il Goh, Michael E Cusick, Albert-Laszlo Barabasi, and Marc Vidal. Drug- target network. Nature biotechnology, 25(10):1119, 2007.

[163] Chang Hun You, Lawrence B Holder, and Diane J Cook. Application of graph-based data mining to metabolic pathways. In Data Mining Workshops, 2006. ICDM Workshops 2006. Sixth IEEE International Conference on, pages 169-173. IEEE, 2006.

[164] Alon Zaslaver, Avi E Mayo, Revital Rosenberg, Pnina Bashkin, Hila Sberro,

Miri Tsalyuk, Michael G Surette, and Uri Alon. Just-in-time transcription program in metabolic pathways. Nature genetics, 36(5):486, 2004.

[165] Bin Zhang and Sargur N Srihari. Properties of binary vector dissimilarity measures. In Proc. JCIS Intl Conf. Computer Vision, Pattern Recognition, and Image Processing, volume 1, 2003.

[166] Bin Zhang and Sargur N Srihari. Fast k-nearest neighbor classification using cluster-based trees. IEEE Transactions on Pattern analysis and machine intelligence, 26(4):525-528, 2004.

[167] Yan Zhang, Zhifeng Yang, and Xiangyi Yu. Ecological network and emergy analysis of urban metabolic systems: model development, and a case study of four chinese cities. Ecological Modelling, 220(11):1431-1442, 2009.

[168] Chunxiang Zheng, Chad R Weisbrod, Juan D Chavez, Jimmy K Eng, Vagisha Sharma, Xia Wu, and James E Bruce. Xlink-db: Database and software tools for storing and visualizing protein interaction topology data. Journal of proteome research, 12(4):1989-1995, 2013.

[169] Xiaowei Zhu, Mark Gerstein, and Michael Snyder. Getting connected: analysis and principles of biological networks. Genes and development, 21(9):1010-1024, 2007.

References

[1] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002. 12

[2] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *arXiv preprint cond-mat/0008064*, 2000. 12

[3] E Michael Azoff. *Neural network time series forecasting of financial markets*. John Wiley & Sons, Inc., 1994. 22

[4] Albert-László Barabási and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004. 31

[5] Béla Bollobás, Gautam Das, Dimitrios Gunopulos, and Heikki Mannila. Time-series similarity problems and well-separated geometric sets. In *Proceedings of the thirteenth annual symposium on Computational geometry*, pages 454–456. ACM, 1997. 22

[6] David E Booth. Time series, 1992. 22

[7] Stefan Bornholdt and Heinz Georg Schuster. *Handbook of graphs and networks: from the genome to the internet*. John Wiley & Sons, 2006. 29

[8] Ulrik Brandes. A faster algorithm for betweenness centrality*. *Journal of Mathematical Sociology*, 25(2):163–177, 2001. 13

[9] Melissa S Cline, Michael Smoot, Ethan Cerami, Allan Kuchinsky, Nerius Landys, Chris Workman, Rowan Christmas, Iliana Avila-Campilo, Michael Creech, Benjamin Gross, et al. Integration of biological networks and gene expression data using cytoscape. *Nature protocols*, 2(10):2366–2382, 2007. 32

[10] Michele Coscia, Fosca Giannotti, and Dino Pedreschi. A classification for community discovery methods in complex networks. *Statistical Analysis and Data Mining*, 4(5):512–546, 2011. 42

[11] Gautam Das, Dimitrios Gunopulos, and Heikki Mannila. Finding similar time series. In *European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 88–100. Springer, 1997. 22

[12] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002. 36

[13] Brian S Everitt. Unresolved problems in cluster analysis. *Biometrics*, pages 169–181, 1979. 21

[14] `http://cancer.sanger.ac.uk/census/`. 42

[15] Marco Franciosi and Giulia Menconi. Multi-dimensional sparse time series: feature extraction. *arXiv preprint arXiv:0803.0405*, 2008. 24

[16] Sylvia Frühwirth-Schnatter, Christoph Pamminger, Rudolf Winter-Ebmer, and Andrea Weber. Model-based clustering of categorical time series with multinomial logit classification. In *AIP Conference Proceedings*, volume 1281, pages 1897–1900. AIP, 2010. 27

[17] Guojun Gan, Chaoqun Ma, and Jianhong Wu. *Data clustering: theory, algorithms, and applications*, volume 20. Siam, 2007. 22

[18] Roger Guimera and Luis A Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, 2005. 15

[19] Dimitrios Gunopulos and Gautam Das. Time series similarity measures (tutorial pm-2). In *Tutorial notes of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 243–307. ACM, 2000. 22

[20] Jing-Dong J Han, Nicolas Bertin, Hao Tong, Debra S Goldberg, et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430(6995):88, 2004. 35

[21] William M Holmes. Time series: Sir maurice kendall and j. keith ord, (edward arnold, great britain, 1990) pp. 296, 1992. 22

[22] Kuhn Ip, Caroline Colijn, and Desmond S Lun. Analysis of complex metabolic behavior through pathway decomposition. *BMC systems biology*, 5(1):91, 2011. 15

[23] Ariel Jaimovich, Gal Elidan, Hanah Margalit, and Nir Friedman. Towards an integrated protein–protein interaction network: A relational markov network approach. *Journal of Computational Biology*, 13(2):145–164, 2006. 35

[24] Hawoong Jeong, Sean P Mason, Albert-Laszlo Barabasi, and Zoltan N Oltvai. Lethality and centrality in protein networks. *arXiv preprint cond-mat/0105306*, 2001. 35

[25] Hawoong Jeong, Bálint Tombor, Réka Albert, Zoltan N Oltvai, and A-L Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000. 15, 35

[26] Minoru Kanehisa. The kegg database. In *Novartis Found Symp*, volume 247, pages 91–101, 2002. 18, 19

[27] Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, Yasushi Okuno, and Masahiro Hattori. The kegg resource for deciphering the genome. *Nucleic acids research*, 32(suppl 1):D277–D280, 2004. 18

[28] Nadav Kashtan, Shalev Itzkovitz, Ron Milo, and Uri Alon. Mfinder tool guide. *Department of Molecular Cell Biology and Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot Israel, Tech. Rep*, 2002. 36

[29] Jinki Kim and Gwan-Su Yi. Rmod: A tool for regulatory motif detection in signaling network. *PloS one*, 8(7):e68407, 2013. 35

[30] T Warren Liao. Clustering of time series dataa survey. *Pattern recognition*, 38(11):1857–1874, 2005. 22

[31] Shmoolik Mangan and Uri Alon. Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences*, 100(21):11980–11985, 2003. 35

[32] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002. 35

[33] Sara Mostafavi, Debajyoti Ray, David Warde-Farley, Chris Grouios, Quaid Morris, et al. Genemania: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol*, 9(Suppl 1):S4, 2008. 20

[34] Mark EJ Newman. Models of the small world. *Journal of Statistical Physics*, 101(3):819–841, 2000. 12

[35] V Pihur, Somnath Datta, and Susmita Datta. Finding common genes in multiple cancer types through meta–analysis of microarray experiments: A rank aggregation approach. *Genomics*, 92(6):400–403, 2008. 31

[36] Michal Ronen, Revital Rosenberg, Boris I Shraiman, and Uri Alon. Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proceedings of the national academy of sciences*, 99(16):10555–10560, 2002. 35

[37] Jill A Rosenfeld, Dina Amrom, Eva Andermann, Frederick Andermann, Martin Veilleux, Cynthia Curry, Jamie Fisher, Stephen Deputy, Arthur S Aylsworth, Cynthia M Powell, et al. Genotype–phenotype correlation in interstitial 6q deletions: a report of 12 new cases. *neurogenetics*, 13(1):31–47, 2012. 20

[38] Rintaro Saito, Michael E Smoot, Keiichiro Ono, Johannes Ruscheinski, Peng-Liang Wang, Samad Lotia, Alexander R Pico, Gary D Bader, and Trey Ideker. A travel guide to cytoscape plugins. *Nature methods*, 9(11):1069–1076, 2012. 31, 32

[39] S Schuster, DA Fell, and T Dandekar. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnology*, 18(3):326–332, 2000. 15

[40] `http://ccgd-starrlab.oit.umn.edu/search.php`. 42

[41] Jimmy Shadbolt and John Gerald Taylor. *Neural Networks and the Financial Markets: Predicting, Combining, and Portfolio Optimisation*. Springer, 2002. 22

[42] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003. 31, 32

[43] Jörg Stelling, Steffen Klamt, Katja Bettenbrock, Stefan Schuster, and Ernst Dieter Gilles. Metabolic network structure determines key aspects of functionality and regulation. *Nature*, 420(6912):190–193, 2002. 15

[44] RAMONA STOICA. Multiobjective approach of multidimensional time series clustering. *Studia Universitatis Babes-Bolyai, Informatica*, 59(1), 2014. 28

[45] Ramona Stoica, Mihaela Ola, and Mihai Paraschivescu. Multidimensional temporal clustering: geometrical similarity measures analysis in k-means. *Intelligent Systems Design and Applications*, 2013. 28

[46] Ramona Stoica, Bazil Parv, and Crina Grosan. Communities detection in multi-relational networks. *PlosOne*, 26(2):299–311, 2017. 11, 45

[47] Ramona Stoica and Liviu Stirb. A tool for comparing multirelational networks from biology. In *Intelligent Systems Design and Applications (ISDA), 2013 13th International Conference on*, pages 242–246. IEEE, 2013. 10, 33

[48] Ruey S Tsay. Analysis of financial time series. financial econometrics, a wiley-interscience publication, 2002. 22

[49] Ikuo Uchiyama. Mbgd: microbial genome database for comparative analysis. *Nucleic acids research*, 31(1):58–62, 2003. 18

[50] David Warde-Farley, Sylva L Donaldson, Ovi Comes, Khalid Zuberi, Rashad Badrawi, Pauline Chao, Max Franz, Chris Grouios, Farzana Kazi, Christian Tannus Lopes, et al. The genemania prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic acids research*, 38(suppl_2):W214–W220, 2010. 20, 42

[51] Duncan J Watts and Steven H Strogatz. Collective dynamics of small-worldnetworks. *nature*, 393(6684):440–442, 1998. 14

[52] Elisabeth A Wong and Brittany Baur. On network tools for network motif finding: a survey study, 2010. 35

[53] Alon Zaslaver, Avi E Mayo, Revital Rosenberg, Pnina Bashkin, Hila Sberro, Miri Tsalyuk, Michael G Surette, and Uri Alon. Just-in-time transcription program in metabolic pathways. *Nature genetics*, 36(5):486, 2004. 35

[54] Bin Zhang and Sargur N Srihari. Properties of binary vector dissimilarity measures. In *Proc. JCIS Intl Conf. Computer Vision, Pattern Recognition, and Image Processing*, volume 1, 2003. 22