Learning by Fusing Heterogeneous Data

A dissertation presented
by

Marinka Žitnik

to
The Faculty of Computer and Information Science
in partial fulfilment of the requirements for the degree of
Doctor of Philosophy
in the subject of
Computer and Information Science

Ljubljana, 2015

APPROVAL

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgement has been made in the text.

— Marinka Žitnik —
June 2015

The submission has been approved by

dr. Igor Kononenko
Professor of Computer and Information Science
examiner

dr. Peter Šemrl
Professor of Mathematics
examiner

dr. Florian Markowetz
Senior group leader at Cancer Research UK Cambridge Institute
external examiner
University of Cambridge

PREVIOUS PUBLICATIONS

I hereby declare that the research reported herein was previously published in peer reviewed journals or publicly presented at the following occasions:

[1] M. Žitnik and B. Zupan. NIMFA: A Python library for nonnegative matrix factorization. *Journal of Machine Learning Research*, 13, 849–853, 2012.

[2] M. Žitnik, V. Janjić, C. Larminie, B. Zupan and N. Pržulj. Discovering disease-disease associations by fusing systems-level molecular data. *Scientific Reports*, 3, 3202, 2013. doi: 10.1038/srep03202

[3] M. Žitnik and B. Zupan. Matrix factorization-based data fusion for drug-induced liver injury prediction. *Systems Biomedicine*, 2, 1: 16–22, 2014. doi: 10.4161/sysb.29072

[4] M. Žitnik and B. Zupan. Gene network inference by probabilistic scoring of relationships from a factorized model of interactions. *Bioinformatics*, 30, 12: 246–254, 2014. doi: 10.1093/bioinformatics/btu287

[5] M. Žitnik and B. Zupan. Matrix factorization-based data fusion for gene function prediction in baker's yeast and slime mold. In R. B. Altman, A. K. Dunker, L. Hunter, T. E. Klein and M. D. Ritchi, editors, *Proc. of the Pacific Symposium on Biocomputing*, 19, 400–411, HI, USA, 2014. World Scientific. doi: 10.1142/9789814583220_0038

[6] M. Žitnik and B. Zupan. Data fusion by matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37, 1: 41–53, 2015. doi: 10.1109/TPAMI.2014.2343973

[7] M. Žitnik and B. Zupan. Data imputation in epistatic MAPs by network-guided matrix completion. *Journal of Computational Biology*, 22, 6: 595–608, 2015. doi: 10.1089/cmb.2014.0158

[8] M. Žitnik and B. Zupan. Survival regression by data fusion. *Systems Biomedicine*, 2, 3: 47–53, 2015. doi: 10.1080/21628130.2015.1016702

[9] M. Žitnik and B. Zupan. Gene network inference by fusing data from diverse distributions. *Bioinformatics*, 31, 12: 230–239, 2015. doi: 10.1093/bioinformatics/btv258

[10] M. Žitnik, E. A. Nam, C. Dinh, A. Kuspa, G. Shaulsky and B. Zupan. Gene prioritization by compressive data fusion and chaining. *PLoS Computational Biology*, 11, 10: e1004552, 2015. doi: 10.1371/journal.pcbi.1004552

I certify that I have obtained a written permission from the copyright owners to include the above published materials in my thesis. I certify that the above materials describe work completed during my graduate study at the University of Ljubljana.

Univerza *v Ljubljani*
Fakulteta *za računalništvo in informatiko*

Marinka Žitnik

Učenje z zlivanjem heterogenih podatkov

POVZETEK

Podatkovno-intenzivni postopki v tehnologiji in znanosti nam v zadnjih letih omogočajo zajem velike količine heterogenih podatkov, ki opisujejo sisteme na različnih nivojih granularnosti in z različnih zornih kotov. Zbrani podatki so pogosto predstavljeni v *povsem različnih podatkovnih domenah*, kar predstavlja izziv za algoritme, ki gradijo napovedne modele z zlivanjem podatkov. Naše raziskave temeljijo na premisi, da je heterogene podatke mogoče "organizirati," tako da vzpostavimo ustrezne preslikave med posameznimi dimenzijami vhodnih podatkovnih domen. Ozko grlo, ki nas loči od boljšega razumevanja podatkovne domene in s tem tudi od bolj učinkovite gradnje napovednih modelov z zlitjem velikih heterogenih podatkov, je prepoznava vrste informacije, ki jo je možno prenesti med povezanimi podatkovnimi nabori, objekti različnih tipov in napovednimi nalogami. V disertaciji predlagamo več zanimivih in zmogljivih napovednih modelov za učenje iz heterogenih podatkov. Ti pristopi so splošni, dosegajo visoko napovedno točnost in so enostavni za uporabo: v veliki meri se izognejo dolgotrajnim in zahtevnim predobdelavam podatkov, na katere se zanašajo trenutni modeli, ki heterogene podatke najpogosteje poskušajo preslikati v enovit podatkovni prostor. Razviti algoritmi so se izkazali za obetavne na večih področjih človekovega delovanja, a smo se v tem delu osredotočili na reševanje aktualnih problemov v molekularni in sistemski biologiji. Ti med drugim vključujejo napovedovanje genskih funkcij in farmakoloških akcij, rangiranje obetavnih genov za nadaljnje biološke raziskave, odkrivanje vzorcev povezav med boleznimi, odkrivanje toksičnosti zdravil in analizo umrljivosti.

Pomemben vidik naših raziskav predstavlja študij *latentnih faktorskih modelov*. Razvijemo več latentnih modelov s faktoriziranimi parametri, ki lahko sočasno naslavljajo več vrst podatkovne heterogenosti; to je, raznolikosti, ki zaobsega heterogene podatkovne domene, več tipov entitet in različne napovedne naloge. Prednost naših algoritmov pred uveljavljenimi pristopi je sposobnost ohranitve strukture odvisnosti med podatki tekom gradnje napovednih modelov, kar smo empirično preverili v večih študijah. Naše nedavno delo na tem področju obsega *pristope za gradnjo mrež* z analizo podatkov iz večih morebitno različnih podatkovnih porazdelitev, ki smo jih uporabili za avtomatično gradnjo genskih regulatornih mrež pri bolezni raka. Modelirali smo tudi epistazo, ki predstavlja pomemben koncept v genetiki. V ta namen smo predlagali učinkovite algoritme za določitev vrstnega reda delovanja genov v genskih poteh, ki porabijo nekajkrat manj računskih virov od znanih tehnik.

Ena izmed osrednjih tem doktorske disertacije je analiza *velikih podatkovnih zbirk*. V empiričnih študijah smo namreč opazili, da je za zanesljive napovedi v bioinformatiki, zaželjene na primer pri odkrivanju odvisnosti med boleznimi in ocenjevanju vpletenosti genov v razne fenotipe, pogosto koristno sklepati na osnovi meritev, ki izhajajo iz različnih eksperimentalnih ali predhodnih računskih postopkov. Med drugim v delu analiziramo 30 heterogenih podatkovnih zbirk, ki nam služijo za ocenjevanje toksičnosti zdravil, in več kot 40 zbirk o odvisnostih med geni v človeku. Slednje predstavlja analizo najobsežnejše zbirke podatkov v dosedanjih študijah latentnih faktorskih modelov. Tolikšna razsežnost podatkov poraja nova vprašanja o izbiri ustreznih podatkovnih virov za zlivanje, za kar predlagamo splošni pristop ocenjevanja občutljivosti med viri.

Ključne besede: napoved genskih funkcij, genska prioritizacija, gradnja mrež, sočasna matrična faktorizacija, matrično dopolnjevanje, faktorski modeli, zlivanje podatkov, bioinformatika, statistično relacijsko učenje, strojno učenje

University *of Ljubljana*
Faculty *of Computer and Information Science*

Marinka Žitnik
Learning by Fusing Heterogeneous Data

ABSTRACT

It has become increasingly common in science and technology to gather data about systems at different levels of granularity or from different perspectives. This often gives rise to data that are represented in *totally* different input spaces. A basic premise behind the study of learning from heterogeneous data is that in many such cases, there exists some correspondence among certain input dimensions of different input spaces. In our work we found that a key bottleneck that prevents us from better understanding and truly fusing heterogeneous data at large scales is identifying the kind of knowledge that can be transferred between related data views, entities and tasks. We develop interesting and accurate data fusion methods for predictive modeling, which reduce or entirely eliminate some of the basic feature engineering steps that were needed in the past when inferring prediction models from disparate data. In addition, our work has a wide range of applications of which we focus on those from molecular and systems biology: it can help us predict gene functions, forecast pharmacological actions of small chemicals, prioritize genes for further studies, mine disease associations, detect drug toxicity and regress cancer patient survival data.

Another important aspect of our research is the study of *latent factor models*. We aim to design latent models with factorized parameters that simultaneously tackle multiple types of data heterogeneity, where data diversity spans across heterogeneous input spaces, multiple types of features, and a variety of related prediction tasks. Our algorithms are capable of retaining the relational structure of a data system during model inference, which turns out to be vital for good performance of data fusion in certain applications. Our recent work included the study of *network inference* from many potentially nonidentical data distributions and its application to cancer genomic data. We also model the epistasis, an important concept from genetics, and propose algorithms to efficiently find the ordering of genes in cellular pathways.

A central topic of our Thesis is also the analysis of *large data compendia* as predictions about certain phenomena, such as associations between diseases and involvement of genes in a certain phenotype, are only possible when dealing with lots of data. Among others, we analyze 30 heterogeneous data sets to assess drug toxicity and over 40 human gene association data collections, the largest number of data sets considered by a collective latent factor model up to date. We also make interesting observations about deciding which data should be considered for fusion and develop a generic approach that can estimate the sensitivities between different data sets.

Key words: gene function prediction, gene prioritization, network inference, collective matrix factorization, matrix completion, factor models, data fusion, bioinformatics, statistical relational learning, machine learning

ACKNOWLEDGEMENTS

First, I would like to thank my advisor Blaž Zupan for his advice and immense support that I got from him during almost three years of my doctoral study. This Thesis would not have been possible without him guiding me through my early research steps when I was still an undergraduate student in mathematics and computer science about five years ago. I have benefited over and over again from long meetings with Blaž, who showed me to think differently about the problems and often enlightened me to ask different questions. His wisdom, flexibility, insightful criticisms, and the generous amounts of red ink Blaž left on my papers have helped me to better understand research and academic life and gain deep technical knowledge about the topics presented in this Thesis. I am especially grateful to Blaž for the many opportunities to go on research visits and for the research freedom he has given me over the last years I spent at University of Ljubljana.

I have been very lucky to have had a wonderful group of collaborators and coauthors. Each of them deserves my appreciation: Charles Boone, Brenda Andrews, Uroš Petrovič, Mojca Ušaj, Matej Ušaj, Petra Kaferle, Nataša Pržulj, Vuk Janjić, Chris Larminie, Balaji Santhanam, Mariko Katoh, Amber Miller, Rafe Rosengarten, Eddie Nam, Chris Dinh, Adam Kuspa, Gad Shaulsky, Jordi Puigvert, Rok Sosič and Jure Leskovec. Thank you for the unselfish help, feedback and for making the research a fun collaborative effort.

I have spent great months at University of Toronto, Imperial College London, Baylor College of Medicine and Stanford University. Many chapters of this Thesis got done during this time. I would like to thank my hosts for openly accepting me into their groups and for the trust in studying the genomes of yeast, amoeba, human disease and society.

I also wish to thank my Thesis committee, Igor Kononenko, Peter Šemrl and Florian Markowetz for their advice and comments.

Of course, my research journey so far would not be the same without the fellow Biolab data miners, and Leskovec, Shaulsky, Boone and Pržulj groups for making vibrant and friendly research environments. I will never forget our discussions about the biology of yeast, genomics of slime mold, my attempts of learning to pipette and the wonders I have seen under your microscopes. I also thank the fellow colleagues at the ACM and the women squad at Google. Thank you Bruna, Lara and Nikola for making my time in Houston fun. Thank you Alice and Norm for the wonderful time in California.

Finally, my parents and my brother Slavko—I will thank you in person.

— Marinka Žitnik, Ljubljana, June 2015.

CONTENTS

Introduction

The main interest of our research presented in this Thesis has been in understanding the different types of heterogeneity in predictive modeling and in developing computational approaches for learning in such settings. Which are efficient and effective ways of considering circumstantial evidence during model inference? How to include drug side effects into a model that predicts associations between diseases? Or, how can we take into account different types of movie roles actors have played when recommending which movie a user should see next? How to map the heterogeneous input spaces to a common space and construct a single prediction model with good generalization performance? Which data sets are complementary when making predictions? How to detect problematic data sets from a collection of tens or even hundreds of data sets? Answers to such questions are important for most problems in science and engineering where we can obtain data sets that describe the observed system from various perspectives and record the behavior of its individual components.

These settings open many new applications, yet they pose new challenges from the algorithm perspective. How can we link seemingly uncorrelated prediction tasks to mutually boost their learning performance? Different tasks might seem to be *totally* uncorrelated with each other if their examples are in different data spaces. For example, in cross-lingual classification, the first task might be classifying a set of English documents whose input space consists of English vocabulary, and the second task might be classifying a set of Slovenian documents whose input space consists of Slovenian vocabulary. Another example is simultaneous classification of documents and images. Here, the first task might be document classification where input space consists of the document vocabulary, and the second task is image classification, whose input space consists of the image vocabulary, such as features extracted from different image regions. Yet another example is gene function prediction in bioinformatics, where different tasks correspond to the different functional roles that genes might have in the cell, and relevant vocabularies span the space of related cellular pathways that genes belong to and diseases that genes are associated with. The possibility of jointly learning multiple *arbitrary* tasks described with heterogeneous input spaces so that they could benefit from each other is vital to a range of application areas from the cross-lingual classification, movie recommendation, the identification of disease-disease associations, drug toxicity detection and gene function prediction to experimental design in biology.

A basic premise behind the study of learning from heterogeneous data is that in many real applications, there might exist some correspondence among certain input dimensions of different tasks. In the example of cross-lingual classification, there is a natural correspondence between the words from two different languages (e.g., a “boat” in English means “čoln” in Slovenian); in the example of document-image classification, some words can be naturally translated into some image regions; in the example of gene function prediction, genes are linked via the protein-protein interactions or comorbidity of diseases that these genes cause. The correspondence across different input spaces hence provides an important connection among different tasks.

The goal of learning by fusing heterogeneous data is to leverage different types of data heterogeneity to improve the performance of predictive modeling. We study three types of data heterogeneity and also their combinations, where multiple types of data heterogeneity can interleave and lead to increasingly more challenging tasks in predictive modeling:

- *Relation heterogeneity:* Compared with traditional task heterogeneity, where the input space is *homogeneous* across different tasks, learning via data fusion is able to borrow consistent patterns across *many* potentially *heterogeneous* input spaces.
- *Object type heterogeneity:* Compared with traditional relation heterogeneity, where the examples are described by features of a *single* type across different data sets, data fusion in a multi-object type setting is able to leverage *heterogeneous* types of features to improve the learning performance in each task.
- *Task heterogeneity:* Compared with traditional collective study of multiple object types, where each prediction task is modeled *independently* across different types of objects, data fusion exploits *related* prediction tasks to transfer knowledge between data views.

1.1 Motivation and applications

Traditionally prediction models were constructed by utilizing a single data source where practitioners typically aimed at encoding it into an example-by-feature data table. For example, when classifying tissue samples into cancerous and non-cancerous, one might describe each sample with a profile (a vector) containing levels of gene expression in that sample and a binary value indicating whether the sample was cancerous or not.

A plethora of machine learning and data mining models have been developed in recent decades to tackle such data representations and predict quantities of interest, e.g. whether a patient suffers from a particular disease or not. Though such off-the-shelf models are very expressive, they often fail to scale to diverse data representations that potentially come from heterogeneous input spaces. Moreover, many times we need to work with tens or even hundreds of diverse data sets to reliably estimate a quantity of interest; thus, the focus moves to the study of heterogeneous data collection as a whole.

Today with the ubiquity of high-throughput technologies across engineering, natural and life sciences, there are several opportunities to study phenomena and systems at large scale and from different perspectives that were not possible before. This can be summarized by the following three points:

- Observations of natural phenomena (e.g., human genome) and technological systems (e.g., web) have detailed data that describe complex relationships between many entities of different types (e.g., genes, RNA molecules and cellular pathways in the case of the human genome; users, events and groups in the case of the web), where many of the entities are circumstantially and in a priori unknown way related to the problem of interest (e.g., the relevance of environmental exposure to a particular genetic disease; the influence of user's online social circle on her movie preferences).
- "Big data" generated by such experiments can be seen as "a large collection of small or medium-sized data sets" opposed to "a single homogeneous big data set" (Zoubin Ghahramani, personal communication).
- Such rich data come with various levels of uncertainty in their measurements and in diverse data representations, such as feature-based data tables, associations, networks and ontologies.

For example, ENCODE (Consortium et al., 2012) is an encyclopedia of DNA elements that aims to identify all functional elements in the human genome sequence. Teams of computational and laboratory-based scientists have worked to apply high-throughput biotechnological approaches to detect sequence elements, which might carry biological functions. This new and varied content has in addition to the human genome (Venter et al., 2001) and numerous studies in molecular biology and functional genomics led to a flurry of research activity in *systems biology* that aims to mine the diverse content and infer useful data from it (e.g. detection of disease causing variants and stratification of cancer patients). Other such examples from different data domains include: data generated by ATLAS experiment at the CERN (Toor et al., 2012) that searches for new particles using head-on collisions of protons at high energy to detect diverse types of events; online recommendation systems (Feuerverger et al., 2012), which are capable of considering movie preferences, demographics data and movie, actor and genre information to support thousands of users selecting which movie to see next; the fusion of multiple global navigation satellite systems (Li et al., 2015) to improve the reliability of positioning and optimize the spatial geometry; or for example, online social networks (Szell et al., 2010; Mucha et al., 2010) that capture various complex communication patterns, such as "likes," "upvotes," and sharing of posts between either individuals or online communities.

Ubiquity of high-throughput data presents many unique opportunities and challenges. A key bottleneck that prevents us from better understanding and truly fusing heterogeneous data at large scales is defining the kind of knowledge that can be transferred between related data views, entities and tasks. Throughout this Thesis our algorithms rely on one of the following three assumptions:

- *Relation transfer:* We build the relational map called *a data fusion graph* of all the relations considered in data fusion and relax the assumptions about independently and identically distributed relations.

- *Object type transfer:* We assume that there exists a common feature space shared by the input spaces, which can be used as a bridge to transfer knowledge.

- *Parameter transfer:* We make use of latent model parameterization and assume that heterogeneous input spaces have shared latent parameters and hyperparameters.

The approaches to sharing of information between related views are aligned with the types of heterogeneity considered in this Thesis. To model individual heterogeneity types and their blend we follow the following three steps in this Thesis:

- *STAGE 1 – Exploration:* We ask a question, which is motivated by a current challenge in molecular and systems biology, do background research and construct a hypothesis. We gather data from public biological data repositories and in-house data from our collaborating institutions, perform measurements and identify one or more types of heterogeneity, which need to be considered during model inference.

- *STAGE 2 – Modeling:* Given observations about different types of data heterogeneity, we design models that give predictions and probabilistic estimates about a problem of interest. We test our hypothesis by doing experiments and further analyze our data.
- *STAGE 3 – Algorithms:* Finally, we present new *generic* algorithms for data fusion and empirically evaluate their effectiveness and prediction power against state-of-the-art systems. Depending on a question asked, our predictions are further validated by biologists in the wet laboratory.

We study six cases where we show that principled approaches of learning by fusing heterogeneous data can improve the quality and performance of inferred prediction models. The six cases are reflected in the map of this Thesis shown in Fig. 1.1.

Thus, the Thesis naturally breaks into six pieces as also shown in Table 1.1: the rows correspond to the research challenges and the columns correspond to previously described types of heterogeneity that are modeled by the respective parts. Next we give the motivation for each of the six parts, following by the summary of our contributions.

Table 1.1

Structure of this Thesis with references to the parts.

Thesis part		Types of data heterogeneity		
		Relation	Object type	Task
Part I	Network side information	✓		✓
Part II	Network inference	✓		
Part III	Compressive data fusion	✓	✓	✓
Part IV	Latent chaining and profiling		✓	
Part V	Regression by data fusion		✓	✓
Part VI	Large-scale data fusion selection	Exploring types of heterogeneity		

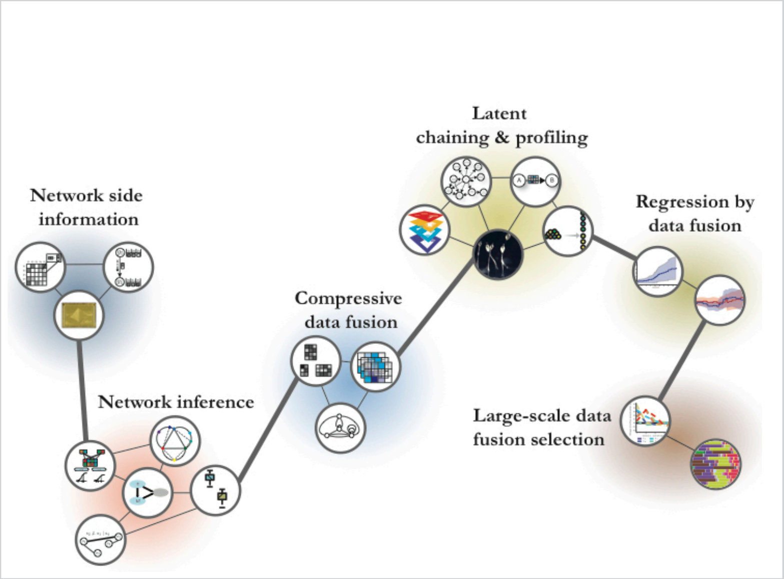


Figure 1.1

The map of this Thesis.

1.1.1 Relation heterogeneity

In many data analysis tasks there exist several different ways to describe the same set of objects. This can lead to multiple distinct representations, "relations" or "views," that encode patterns relevant to the problem of interest. How can we integrate these representations in a way that allows us to effectively identify these patterns? In some applications we may have access to a set of relations that are entirely *consistent* – the same patterns occur across all relations. The problem then becomes estimation of a single consensus model summarizing the patterns common to all relations. However, in some cases, substantial discord may exist between the data in different relations. An effective data fusion procedure must detect common patterns, reconcile the disagreements while also preserving patterns that are specific to each relation.

Applications. The predictive modeling of multi-relation domains has found the following applications in this Thesis:

- *Modeling background knowledge represented with networks:* Rich relational data can naturally be modeled and encoded with networks. Our methods form means of including network side information within the inference of a latent data model to improve prediction of genetic interactions. Our methods are useful when making predictions for objects that are entirely missing from a certain relation, i.e. addressing the cold-start problem.

- *Epistasis analysis:* One of the cornerstone questions in genetics is concerned with the estimation of how mutation in a gene affects the activity of genes that act downstream in a certain cellular pathway. How does a cell orchestrate complex relationships when pathways contain many genes? Until now, it was computationally very hard to infer gene networks based on epistasis analysis. Our results help form a promising basis for inference of pathways from genome scale data that can readily be investigated by biologists.

- *Network inference:* When data in multiple relations come from potentially *non-identical* data distributions, powerful data fusion algorithms should be capable of jointly modeling the data while accounting for statistics of each distribution.

1.1.2 Object type heterogeneity

It has become increasingly common to gather data about a system at different levels of granularity or from different perspectives. This can give rise to data that are represented in *totally* different input spaces. Consider, for example, a typical online book recommendation service, which aims to recommend books that would be of interest to a user. A primary data set for such recommendation engine might be user's reading history, i.e. a user-by-book data table. However, one can easily envision the potential of considering authors' biographies, i.e. an author-by-biography data table, during model inference.

The challenges arising in multi-object type domains are typically resolved in a labor-intensive way through feature engineering that transforms data into profiles describing objects of a single type, e.g. users. Data preprocessing is neither standardized nor trivial and may lead to loss of information. Can we design algorithms that can model multi-object type data without them necessitating substantial feature engineering?

Applications. Working on heterogeneous object types has led us to develop novel methodology to study:

- *Object profiling in the latent space:* Analyzing heterogeneous object types within a single prediction model gives us means to *chain* latent spaces of individual object types. This allows for easy profiling of one object type in the latent space of another object type. Profiles constructed by chaining are useful as input to general machine learning algorithms.

- *Gene prioritization:* The identification of genes involved in a certain disease often requires time consuming and expensive examination of many candidate genes, since genome-wide techniques such as association studies and linkage analysis frequently select many hundreds of candidates. Using many heterogeneous data sets we can more accurately prioritize genes at scales that were not possible before. Our work allows us, for example, to identify which genes are most likely involved in mechanisms of bacterial resistance in *Dictyostelium*.

- *Mining disease associations:* To find relationships between diseases one needs to shift away from linking diseases simply based on their shared genes towards evidence from fusing all available molecular interaction and ontology data. Our work highlights the importance in the paradigm shift towards systems-level data fusion.

- *Drug toxicity detection:* Development of tools for early identification of adverse effects in drugs is a major challenge within pharmaceutical industry and clinical medicine. Our large-scale efforts to forecast drug-induced liver injuries suggest that joint analysis of toxicogenomic data together with circumstantial data sets allows prediction of liver injuries in humans from animal data. The ability to model objects of different types, e.g. genes, drugs, samples and biological processes, is important by itself as it allows us to discover patterns not found in data sets that are limited to a single object type.

1.1.3 Task heterogeneity

The third type of heterogeneity addressed in this Thesis deals with the analysis of data from two or more tasks. Whereas single-task learning solves each task in *isolation* and ignores potential relations between the tasks, multi-task learning solves the tasks jointly. Such analysis exploits the relations between the tasks to reduce the hypothesis space and improve generalization. The advantage of learning multiple tasks across heterogeneous input spaces manifests when the tasks are truly related and the transfer of information between related tasks is properly employed. To take a recent example from the pharmacology domain, prediction of aspirin's pharmacological actions benefited largely from joint modeling of aspirin as an "inhibitor of platelet aggregation" as well as an "cyclooxygenase inhibitor," rather than independent analysis of the two chemical actions. However, directly modeling many tasks on a large scale proved difficult.

Applications. In this Thesis, different tasks are permitted to have different input spaces. Our models assume that each task has its own features but might also share features with other tasks:

- *Gene function prediction:* Development of effective methods that can predict gene functions in an easily extensible way is an important goal in computational biology. The data fusion in our work is achieved by simultaneous analysis of data and sharing of latent data structure between data sets and tasks. This allows, for example, prediction of ontological annotations in slime mold *D. discoideum* and recognition of proteins in baker's yeast *S. cerevisiae* that participate in the ribosome or are located in the cell membrane.

- *Mining pharmacological data:* Integrative analysis of gene ontological annotations is related to the prediction of pharmacological actions for small chemicals. It forces us to develop *general* algorithms and tools that scale to large heterogeneous data collections.

1.1.4 *Dual and triple data heterogeneity*

Ultimately we aim at designing methods capable of addressing problems with multiple types of data heterogeneity. In several chapters of the Thesis (Table 1.1) we break the limit of a single type of data heterogeneity in an attempt of extending our methods to a wider range of applications. Our algorithms are thus designed to take advantage of the consistency across many data relations, the ease of modeling heterogeneous object types, and the relatedness of multiple tasks.

Applications. Our focus on analyzing and modeling dual and triple data heterogeneity is useful, for example, when trying to understand the complexity of cellular machinery or to predict cancer progression in patients:

- *Model selection in data fusion:* When tackling several *tens of genome-wide data sets*, which is a common theme of the Thesis, one becomes interested in how changes of one relation (data set) affect the latent model representation of another relation in the context of a given data fusion algorithm. How, for example, would a change of casting affect user's preferences in a user-movie recommendation system? Our results help identify surprising data sets and problematic data sets that contain potential experimental errors.
- *Survival regression:* Cancer subtype classification is a prominent problem in cancer genome studies, whereby a heterogeneous population of tumor samples is broken into clinically meaningful subtypes. Stratification of tumors typically relies on the similarity of molecular profiles. It aims at predicting important clinical properties including patient survival time and response to chemotherapy. Although individual data sets have long been used to stratify patients, stratification based on multiple types of data, such as expression, methylation and somatic mutation profiles, has been more challenging. These data sets are fundamentally different from each other, both in type and in structure. Our work in this area demonstrates that problems originating from data diversity can be largely surmounted by data fusion, which provides gains in accuracy through data integration.

1.2 Thesis overview and contributions

The Thesis addresses a number of important questions regarding the inference in settings where plenty of heterogeneous data are available. It investigates how to organize diverse data sets such that predictive modeling can benefit from transferring information between related data views, types of objects and prediction tasks.

The work presented here focuses on modeling heterogeneous data with latent factor models where we achieve the transfer of information via sharing of latent parameters or by a factorized representation of model parameters (Žitnik and Zupan, 2012).

Overall, the Thesis aims to show that factorized parameterization and sharing are two powerful techniques that can transform traditional models, which typically learn from homogeneous data, like Markov networks or matrix factorization, into general data fusion methods. Our Thesis has a "six-by-three" structure: it analyzes six problem domains where each of them is examined from the perspective of at least one out of three types of data heterogeneity: Relation heterogeneity, Object type heterogeneity and Task heterogeneity. Most Thesis parts investigate dual or even triple data heterogeneity. Table 1.1 gives the overall structure of our research with the mapping to the parts of this Thesis.

In what follows we describe the main questions this Thesis asks and answers. We break each of them into three steps that follow the above mentioned stages: Exploration, Modeling and Algorithms, which are consistent with the scientific method.

1.2.1 Part I – Network side information

To develop an integrative approach to data analysis one needs to include additional information into the model inference itself. A celebrated model that might benefit from inclusion of side information is matrix completion, which estimates a factorized latent model from a relational data table that contain many unobserved entries. A common assumption employed by matrix completion algorithms is that observed data has been generated by an unknown (i.e., hidden or latent) process with substantially fewer degrees of freedom (i.e., dimensions) than the dimensionality of the original data.

The first part of the Thesis presents our results on a Bayesian view of matrix completion, which readily allows us to couple the inference of with the network-based side information in order to improve the quality of a latent data model itself.

Stage 1 – Exploration: What is the role of network side information in various prediction settings?

First we present a study, which evaluates the significance of side information for four distinct patterns of unknown entries that might appear in a data matrix (Žitnik and Zupan, 2014d). The simplest pattern has unknown elements distributed independently and uniformly at random. While this scenario is most often empirically evaluated, it is less relevant in real world applications where unknown entries typically have a certain structure. In more realistic situations all matrix entries corresponding to a subset of objects might be unknown, or all interactions between two disjoint subsets of objects might be missing. These scenarios occur, for example, in genetic interaction studies, where interactions within a group of essential genes typically cannot be measured, or when two genetic interaction data sets that share a subset of genes are combined into one large data set. Another example of the latter setting are patient data from studies that used various experimental platforms to collect the same type of measurements for different patient subgroups. The fourth prediction setting, which represents the hardest challenge from the learning perspective, hides all values from a subset of rows or columns of a data matrix. It is known as a *cold start* setting and arises in interaction studies when complete genetic interaction profiles are missing. We explored several genetic interaction data sets and observed that inclusion of additional genomic data is crucial when our goal is to predict interactions that follow a structured pattern (Žitnik and Zupan, 2014d). These findings are important for recommendation systems in collaborative filtering and interaction studies in genetics.

Stage 2 – Modeling: How can we model network side information?

We examine network side information by studying a Bayesian matrix completion model and prior knowledge presented with potentially many weighted networks. We formulate a probabilistic model, in which distribution of a latent feature vector depends on the latent vectors of its direct neighbors in the provided networks (Žitnik and Zupan, 2015d). It is the individual latent vectors that collectively give rise to the propagation of their influence over the network. A latent vector of a given object “A” in our model should thus be “close” to the latent vectors of objects located in the network neighborhood of “A”. In fact, our network-guided matrix completion is capable of transferring information across all available measurements and network neighborhoods, which can lead to more accurate inference than simply estimating a particular target measurement independently of any additional knowledge.

Stage 3 – Algorithms: How do we effectively infer latent models using circumstantial network data?

Last we examine a question of how one can effectively learn the latent vectors and estimate network weights. We used the *maximum a posteriori* principle to develop an algorithm that maximizes the posterior probability over the latent vectors. In contrast to previous models, network-guided matrix completion includes side information encode in connectivity of the networks, which allows it to predict matrix rows and columns of objects for which none of the entries is observed, i.e., a cold start setting, while still being mathematically tractable. We show how network side information can be used to predict genetic interactions in epistatic miniarray profile (E-MAP) data assays. In a validation study with several large-scale interaction data sets we were able to demonstrate superior performance of network-guided matrix completion over competing *local* models, which rely on neighbor-based methods and local least squares, and *global* models, which assume a global covariance structure between all genes in the E-MAP data set (Žitnik and Zupan, 2015d). We found that global methods perform poorly when groups of genes predominantly have distinct local similarity patterns and that local methods achieve solid performance across data sets of various size. Moreover, we empirically studied model generalization in various prediction settings. We showed that distribution and the abundance of unmeasured genetic interactions have a significant impact on predictive performance and can limit direct application of non-integrative prediction methods to E-MAPs.

Contributions and impact:

- We developed the network-guided matrix completion, which is a generic and mathematically tractable probabilistic matrix completion model. Moreover, network-guided matrix completion is unique in *fusing relational data* with *network side information* through inference of a single prediction model. We targeted gene interaction data sets and showed that our approach achieved very good generalization.
- Our work on analyzing genetic interaction data has high practical value for the prediction of *entire gene interaction profiles* for genes whose interactions otherwise

cannot be measured directly due to limits of biotechnology.

1.2.2 Part II – Network inference

In the first part of the Thesis our focus is on employing network data as background knowledge to improve completion of partially observed data matrices. The second part of the Thesis investigates the inverse question—it presents our results on statistical network inference, where our goal is to estimate the network edges between objects for which experimental data are available. In bioinformatics, developing insights into complex associations in high-dimensional data sets is important for inference of gene regulatory networks, automatic reconstruction of gene pathways, suggesting promising drug targets and finding potential disease causing genes, among others.

The most straightforward approach to network inference is to observe correlations between data profiles, which typically infers dependencies that are circumstantial rather than *causative*. Whereas *direct network inference* provides useful knowledge about, for example, genes that participate in a common biological process or share a cellular function, we turned our attention to *model-based network inference*, which, for example, carries the potential to identify, which gene acts upstream of another gene in a cellular pathway. We investigated two cases of such inference via undirected graphical models and probabilistic scoring of epistatic relationships and we were able to estimate networks through integrative analysis on a large scale. We found that network edges estimated by our models can be related to causal inference and reveal complex dependencies that cannot be uncovered otherwise using either techniques of direct network inference or a single data source.

Stage 1 – Exploration: What are patterns of gene-gene relationships and probability distributions describing them?

Here we examine how different types of data follow distinct data distributions. Our basic premise is that disregarding information about data distribution can adversely affect performance of prediction models. This work had influence on developing a Markov network model for inference from multiple related but *non-identical data distributions* (Žitnik and Zupan, 2015b). For example, to date, bioinformatic studies commonly assumed that data follow a Gaussian distribution. While this assumption holds for log ratio gene expression values generated by the microarrays, we found that increasing quantity of high-throughput omics data, such as that from the next-generation sequencing, come from skewed distributions. For example, gene expression levels generated by the RNA-sequencing technology count how many times a transcript maps to a specific genomic location and as such these data would be more appropriately modeled with the Poisson or the negative binomial distribution rather than a Gaussian distribution. Another example is data on different types of mutation and copy number variation that follow a categoric data distribution. We showed that such data can be effectively modeled if one considers a broad class of *exponential family* distributions. Surprisingly, despite the growing body of non-Gaussian data and our ability to collect them, computational approaches to support non-Gaussian distributed data are at best scarce. Moreover, there is only a handful of techniques that support *epistasis analysis*, an important concept from classical genetics capable of estimating the order-of-action in gene pathways from mutant-based phenotypes.

Stage 2 – Modeling: How can we jointly model multiple non-identical data distributions?

We present two models that utilize heterogeneous data for network inference. First, Réd estimates gene networks that are *consistent* with observed gene-gene epistatic relationships (Žitnik and Zupan, 2014c). This means that Réd aims to minimize the number of edges that violate the rules defined by the epistasis analysis. The model relies on quantitative but potentially noisy and missing mutant phenotype data. Réd defines a probabilistic latent model for the *entire set of pairwise gene relationships*. In contrast to previous small-scale models that perform local structural changes to the evolving network, Réd uses *global* latent data representation to account for noise and data sparsity. We show that accurate scores indicating preference for different types of gene-gene relationships, i.e. epistasis, partial interdependence and parallelism, can be derived from Réd's latent data model. Whereas Réd is addressing the scarcity of computational methods for epistasis analysis in genetic interaction studies, it lacks a broad appeal of a general network inference method. To this end, we develop FuseNet, which is a generic approach for network assembly from data arising from potentially *many nonidentical exponentially family distributions* (Žitnik and Zupan, 2015b). The principal innovation of this work is a latent parameterization of a Markov network model such that latent parameters are shared between models for different exponential family distributions. We show that FuseNet's power of generalization comes from its two key components: the ability to model non-Gaussian distributions and the fusion of data through reuse of latent model parameterization.

Stage 3 – Algorithms: How can we predict interdependence of genes at large scales?

We developed algorithms for Réd and FuseNet that handle genome-scale genetic interaction data sets and large-scale heterogeneous cancer genomic data of The International Cancer Genome Consortium and The Cancer Genome Atlas. Réd shows promising performance in accuracy and speed relative to the competing techniques. Using the latent model of Réd we efficiently search the space of all possible networks and estimate model parameters for networks with thousands of genes in a matter of minutes, while alternative approaches use ensembling and sampling with a runtime of several days (Žitnik and Zupan, 2014c). On a related note, FuseNet couples model parameters of different data distributions and thus cannot directly utilize existing optimization algorithms for undirected graphical model selection. To this end, we propose to fit FuseNet's models through a cyclical coordinate descent along the entire path of regularization parameters (Žitnik and Zupan, 2015b).

Contributions and impact:

- We developed FuseNet, an *off-the-shelf network inference framework* for mixed data arising from any combination of exponential family distributions. Furthermore, FuseNet is the first model that is able to combine the theory of Markov network inference and data fusion.

- We analyzed heterogeneous data from the International Cancer Genome Consortium and found that joint network inference by FuseNet from multiple related data sets, i.e. RNA-sequencing and somatic mutation data, showed greater functional enrichment than networks learned from any data type alone.

- We developed Réd, an approach to epistasis-based gene network inference that is able to reconstruct known cellular pathways more accurately than competing methods.

- Réd allowed us to infer networks consistent with the theory of epistasis analysis by considering hundreds of thousands of genetic interaction measurements, the largest data compendium considered for epistasis analysis up to date.
- Réd has been harnessed by the molecular biology community, e.g. by Uroš Petrovič at Institut Jožef Stefan and the group of Thomas Helleday from SciLifeLab at Karolinska Institutet.

1.2.3 Part III – Compressive data fusion

The third part of the Thesis presents our work on triple data heterogeneity, namely the development of models that address relation, object type and task heterogeneity (Table 1.1). Recently, a variety of real applications emerged, which exhibit triple heterogeneity. We show how modeling of multiple types of data heterogeneity gives as opportunities to predict biological functions of genes and pharmacological actions of small chemicals using large amounts of diverse data that were previously practically impossible to consider.

Stage 1 – Exploration: What is a relational map of heterogeneous data?

We present *fusion graph*, a relational map of heterogeneous data compendium that is considered for data fusion (Žitnik and Zupan, 2015a). We view each data set as a dyadic relation that encodes relationships between objects of two types. This abstraction is of sufficient generality to apply to many data-rich problem domains, e.g., functional genomics, pharmacology, social networks and recommendation systems, that contain tens or even hundreds of data sets, each potentially relating different object types.

Stage 2 – Modeling: How can we model triple data heterogeneity?

We present our work on multiplex, multiscale and multi-object type matrix factorization models. Researchers in data mining and machine learning have long been excited about "matrix decomposition," where the intuition is to approximately decompose a *large data matrix* into a "useful" product of several *much smaller* data matrices typically referred to as latent matrices or latent factors. We develop DFMF, a penalized matrix tri-factorization model that collectively tri-factorizes many data matrices such that each data matrix is decomposed into a product of three latent matrices. The essence of the model comes from its design, which *reuses the latent matrices* when co-factorizing *related data matrices* (Žitnik and Zupan, 2014a). This formalization has a wide range of applications in the area of relational learning.

DFMF modifies standard factorization formulation such that it can consider multi-relational and multi object-type data without necessitating substantial data transformation. In this way it breaks through conventional feature-based data types and factorization of a single dyadic relation. Few existing matrix factorization approaches for data integration (see Žitnik and Zupan (2015a) and references therein) can model multiple relations between the same two sets of objects, e.g., genes and drugs, or can vary object types along one dimension of data matrices. They would often require full set of pairwise relations between all pairs of object types. On the contrary, DFMF can model multiple relations between multiple object types without imposing any assumptions about structural properties of the matrices.

Stage 3 – Algorithms: How can we predict gene functions and pharmacological actions via collective latent modeling?

The collective penalized matrix tri-factorization model has also led us to efficient and theoretically sound algorithms for collective matrix factorization. Our approach *provably* finds *a quality estimate* of the latent matrices (Žitnik and Zupan, 2015a). We utilized the approach for gene function prediction in yeast and amoeba, where the task was to predict ontological annotations of genes derived from the Gene Ontology (Žitnik and Zupan, 2015a, 2014a). The approach was flexible and, in contrast to state-of-the-art kernel-based methods required minimal preprocessing of the input data. The whole-genome gene function prediction on compendia with tens of data sets required minutes of computation time compared to hours required by competing algorithms. We showed that inclusion of circumstantial evidence improved the accuracy of prediction models. Beyond the task at hand, we showed that the same algorithm can be used to decide the pharmacological actions of small chemicals.

Contributions and impact:

- We developed DFMF, a model for collective matrix factorization and its variant for collective matrix completion. We proved that latent matrices found by our algorithm for the estimation of DFMF model locally minimize the total reconstruction error of a data system presented with the data fusion graph.
- We found that latent matrices estimated by the DFMF algorithm have high predictive power relative to the performance of methods that transform data into a single feature-based data table, i.e. *early integration*, and methods that explicitly address the multiplicity of data via multiple kernel learning, i.e. *intermediate integration*.
- Our approach is general and flexible. We successfully used it for prediction of gene annotations in amoeba, identification of chemical actions and for recognizing yeast proteins that participate in the ribosome or are located in the cell membrane. Follow-up works with collaborators at Baylor College of Medicine and Karolinska Institutet showed promising performance of our approach on human cancer data sets, mouse data related to the development of retinal diseases, data from fruit fly model organism and on large-scale data from amoeba organism.

1.2.4 Part IV – Latent chaining and profiling

To use latent models in various prediction settings one also needs to understand the different roles that latent data matrices returned by a particular decomposition algorithm might have. The fourth part of the Thesis presents our work on utilizing latent models of data systems with tens of data sets for clustering, association mining and gene ranking. We show how integrative analysis allows us to recognize patterns that were practically invisible in small-scale studies (Žitnik et al., 2015b).

Stage 1 – Exploration: How does systems-level view complement disease-disease, gene-phenotype and gene-drug associations?

Here we examine how the advent of genome-scale genetic and genomic studies enables new insights into identification of genes involved in the onset of a phenotype, discovery of disease-disease associations and into prediction of drug toxicity levels. We build on a recent shift from relating human diseases simply based on pathological analysis and clinical symptoms towards systems-level integration of molecular data. By fusing *11 genome-scale human data sets* we identify several *disease-disease associations that were not present in Disease Ontology* for which we find strong support in the literature and significant comorbidity effects in associated diseases (Žitnik et al., 2013). We show that even sparse data sets with only a few data points might be important for effective integration. Surprisingly, we found that genetic interaction data were the most predictive underlying factor of disease-disease associations despite their current small size. Another evidence in support of a systems-level view is our study on *29 toxicogenomic data sets*, where we find that drug-induced liver injuries in humans can be predicted from animal data and circumstantial data sets (Žitnik and Zupan, 2014b). Furthermore, starting from *14 whole-genome data sets from amoeba* and only four genes relevant to bacterial response in *Dictyostelium*, we were able to recommend *eight candidate genes that were readily validated* as necessary for the response of *Dictyostelium* to Gram-negative bacteria (Žitnik et al., 2015a).

Stage 2 – Modeling: How can we reduce degrees of data heterogeneity?

We also study the utility of inferred latent factors for prediction. We analyzed the reconstruction quality of data matrices, each of which may have only a few percentage of observed cells. We further co-clustered objects of various types, e.g., drugs, diseases and genes, based on their latent profiles obtained via co-factorization of a data system. To revert triple heterogeneity exhibited by our applications in molecular biology domain to problems with dual heterogeneity, i.e. relation heterogeneity and task heterogeneity, we introduce Collage, a model that *chains latent matrices along paths defined by the fusion graph* (Žitnik et al., 2015a). These findings are important for construction of a feature-based data tables that can subsequently be used as an input to an established machine learning algorithm.

Stage 3 – Algorithms: How do we prioritize genes, disease and drugs relative to the reference knowledge?

Last, we present way of how latent space of a data system induced by a collective factor model can be used to profile objects in the input space of any other object type based on the connectivity in the data fusion graph. We show that *latent matrix chaining* is an effective technique for construction of dense profiles that include the most informative features obtained by collectively compressing a data system via matrix factorization. Our approach (Žitnik and Zupan, 2014b) ranked first in the "Critical Assessment of Massive Data Analysis" competition, where the task was to predict drug adverse effects from in vivo and in vitro animal toxicogenomic data, hematological and clinical chemistry data, and human gene expression data. Beyond prediction of drug toxicity, we used the same algorithm to successfully mine relationships between human diseases.

Contributions and impact:

- We developed Collage, an approach to gene prioritization. Given a handful of *seed genes* important for a biological function of interest, Collage aims to identify the most promising candidate genes for further studies. In contrast to *gene-centric prioritization algorithms*, Collage represents an advancement in the breadth of data it can incorporate, the ease of data integration without complex feature engineering, and the ability to retain the relational data structure during model inference.

- Our formalization of gene prioritization and models for detection of drug toxicity and discovery of disease-disease associations have had a wide range of implications for researchers in the life sciences. For example, the identification and characterization of four seed genes for the bacterial resistance study was a laborious task that required several months of laboratory work per gene. Collage *substantially simplified this task by suggesting eight genes that were successfully validated in the wet lab.*

- Our approach for drug toxicity detection in toxicogenomic studies received the best analysis award at ISMB CAMDA 2013 conference (Žitnik and Zupan, 2014b).

- Follow-up works with collaborators from Baylor College of Medicine later confirmed the potential of our gene prioritization approach to identify promising genes involved in human retinal diseases.

1.2.5 Part V – Regression by data fusion

Whereas methodology presented in the third part and the fourth part of the Thesis focuses on finding a compressed, i.e. latent, data representation of a heterogeneous data system, the fifth part of the Thesis presents our results on simultaneous estimation of a latent data model and regression against a target data variable. Our results in survival analysis highlight the benefits of data fusion for inference of survival models that are predictive of clinical outcomes.

Stage 1 – Exploration: Which are insightful data types in cancer genome studies?

Individual cancer data sets have long been used to partition a population of tumor samples into clinically meaningful subtypes. We analyzed one of the largest available collections of cancer data from The International Cancer Genome Consortium, trying to find how well different types of data, such as levels of protein expression or somatic mutation data, predict the clinical outcomes of patients. We observed substantial differences in predictive performance when estimating survival models using different data types (Žitnik and Zupan, 2015e). We also found that transformation of high-dimensional cancer genomic data to a low-dimensional space was vital for modeling patient survival time.

Stage 2 – Modeling: How can survival models consider circumstantial evidence?

Stratification of patients based on multiple types of data, such as expression, methylation and somatic mutation profiles, is an important challenge in cancer bioinformatics. The challenge stems from fundamentally different data sets, both in type and in structure. For example, somatic mutation profiles are sparse and discrete since typically only a small fraction of genes are mutated and patients diagnosed with the same cancer type share few mutations. On the other hand, methylation data are typically dense and real-valued. We developed a DFMF-SR model that couples the additive survival regression model with collective matrix factorization into a joint inference procedure (Žitnik and Zupan, 2015e). In contrast to existing survival regression models, DFMF-SR allows *simultaneous* modeling of patient latent data profiles and estimation of the influence of latent factors on survival time.

Stage 3 – Algorithms: How do we predict patient survival times?

Last, we developed an efficient algorithm for DFMF-SR model that is based on computing a solution to the Sylvester matrix equation, a well-characterized type of linear matrix equation. Our approach (Žitnik and Zupan, 2015e) ranked first in the "Critical Assessment of Massive Data Analysis" competition, where the question was whether the integration of comprehensive cancer data consisting of gene expression, microRNA expression, protein expression profiles, somatic mutations and methylation can identify disease causal changes. We also showed that DFMF-SR gave information about the time-varying effects of latent factors on patient survival time. We found that the most informative factors are related to known cancer processes. Beyond the task at hand, our findings point to a potential utility of the proposed approach for uncovering critical factors and their changing influence on cancer progression.

Contributions and impact:

- We developed DFMF-SR, a data fusion approach to survival regression, and an efficient algorithm for the estimation of model parameters. We showed for selected cancer data from The International Cancer Genome Consortium that DFMF-SR performs well relative to a popular approach that first transforms data into the latent space and then does survival regression independently of data transformation. Moreover, DFMF-SR is the first approach that is able to infer a latent data model and regression coefficients of a survival model *at the same time*.
- Our approach for survival regression via data fusion received the best analysis award at ISMB CAMDA 2014 conference (Žitnik and Zupan, 2015e).

1.2.6 Part VI – Large-scale data fusion selection

The work presented in the first five parts of the Thesis generally consider many data sets for each prediction task at hand. In the sixth part we take a step back and ask an important question of how changes in one data set affect the latent representation of another data set in the context of a given collective latent factor model. Answers to this question are vital if we would like to select how many and which data sets should be considered for data fusion. In the sixth part we aim to understand the sensitivity of one data set to perturbations in another data set when both data sets are modeled by a collective matrix factorization.

Stage 1 – Exploration: How changes in one relation affect the latent representation of another relation?

For example, in a user-movie recommendation system, how would a change of animation technology affect users's preferences? We study a data system of 13 data sets from molecular biology and another system of 40 experimental protein physical interaction data sets, *the largest data compendium considered by a collective latent factor model to date* (Žitnik and Zupan, 2015c). We investigate how additions or removals of data sets change the quality of the fitted latent data models and find that there does not exist a simple relationship between, for example, the number of data points in a data set and its affect on the latent representation of other modeled data sets.

Stage 2 – Modeling: What is sensitivity between different relations of a latent model?

Motivated by our observations, we present our work on modeling the inter-relation sensitivity in collective matrix factorization. We use concepts from *matrix algebra* and the *Fréchet derivation* to develop Forensic, a generic approach to sensitivity estimation that is readily applicable to many different collective factorization models (Žitnik and Zupan, 2015c). In fact, our work is the first to directly quantify the amount of sensitivity between relations in large data collections analyzed with latent models. We arrive at a simple yet surprisingly accurate scoring technique with high levels of agreement when applied to different factorization models and scores that report sensitivities, which are intrinsic properties of a relational data structure rather than a confound of a given factorization model.

Stage 3 – Algorithms: How can we select data for fusion and identify surprising data sets?

Last, we develop an algorithm that uses the LAPACK norm estimator to efficiently estimate Forensic's scores. Forensic is capable of estimating sensitivity for *any pair of modeled relations* for which it needs *a one-time-only inference of a factorized model.* In this way, Forensic avoids computational burdens associated with the repeated runs of a factorization algorithm. We found that Forensic opens many new applications that were previously not possible. When analyzing large compendiums of data sets their sheer number increases the likelihood that there will be an outlier data set of lower quality. We showed how Forensic can be used to detect low-quality experimental data sets. Forensic also provides recommendations as to which data sets should be used for integrative analysis and offers insights into "surprising," i.e. potentially problematic, data sets.

Contributions and impact:

- We developed Forensic, a *general* and *computationally efficient* approach to inter-relation sensitivity estimation in collective latent factor models. Moreover, Forensic is the first principled model offering such functionality for collective latent factor models. Forensic shows promising results to be used as a scoring technique for selection of data sets for fusion.
- We analyzed a compendium of 40 experimental protein physical interaction data sets, which is to the best of our knowledge the *largest collection of data sets examined with a collective latent factor model* reported in the literature up to date. We demonstrated that Forensic can correctly identify data sets that contain experimental errors.

Next, we present basic concepts and preliminaries, introduce the notation and briefly survey the related work. We then proceed with each of the six main parts of the Thesis.

2

Overview and survey

In this chapter we review the basic concepts and terminology used in this Thesis and introduce the most important notation. Next, we survey the work on machine learning methods that learn from heterogeneous data, as well as latent factor models and methods that estimate their parameters.

2.1 Basic concepts and definitions

Next, we briefly define concepts and terminology that is used throughout the Thesis. We introduce factorization of a single data matrix and review fundamental concepts from molecular biology needed for fully understanding the problems addressed in the Thesis.

2.1.1 Single matrix factorization for data analysis

Let us consider tabulated data, organized in the observed matrix $\boldsymbol{X} \in \mathbb{R}^{n \times m}$, which we would like to approximate by a product of two matrices $\boldsymbol{U}\boldsymbol{V}^T$, where $\boldsymbol{U} \in \mathbb{R}^{n \times k}$ and $\boldsymbol{V} \in \mathbb{R}^{m \times k}$. If we view the rows of $\boldsymbol{X}$ as data vectors $\boldsymbol{X}_i$, then each such data vector is approximated by a linear combination $\boldsymbol{U}_i\boldsymbol{V}^T$ of the rows of $\boldsymbol{V}^T$. We think of the rows of $\boldsymbol{V}^T$ as *latent factors* and the entries of $\boldsymbol{U}$ as *coefficients* of the linear combinations. In a geometrical setting, the data vectors $\boldsymbol{U}_i \in \mathbb{R}^m$ are approximated by a k-dimensional linear subspace spanned by the rows in $\boldsymbol{V}^T$. The converse also holds: the columns of $\boldsymbol{X}$ can be viewed as linear combinations of the columns of $\boldsymbol{U}$. We refer to $\boldsymbol{U}$ and $\boldsymbol{V}$ as *latent matrices* or *latent factor matrices*.

If we do not impose additional constraints on matrices $\boldsymbol{U}$ and $\boldsymbol{V}$, then the matrices, which can be *exactly* factored as $\hat{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{V}^T$ are those matrices of rank at most k. That is, approximating a matrix $\boldsymbol{X}$ by an unconstrained factorization is equivalent to approximating it by a rank-k matrix.

We were ambiguous about the notion of "approximating" the data matrix. In what sense do we desire to approximate the data? And, what is the measure of discrepancy between the data $\boldsymbol{X}$ and the model $\hat{\boldsymbol{X}}$ that we would like to optimize for? Can we see the "approximation" as fitting a suitable probabilistic model?

Unconstrained factorizations. The most common measure of discrepancy between the data $\boldsymbol{X}$ and the model $\widehat{\boldsymbol{X}}$ is the sum-squared error, or the Frobenius norm of the difference between $\boldsymbol{X}$ and $\widehat{\boldsymbol{X}}$:

$$\|\boldsymbol{X} - \widehat{\boldsymbol{X}}\|_{\text{Fro}}^2 = \sum_{i,j} (\boldsymbol{X}_{ij} - \widehat{\boldsymbol{X}}_{ij})^2. \tag{2.1}$$

The popularity of the Frobenius low-rank approximation is due to the simplicity of computing the factorization. It is a standard result that the k-rank matrix $\widehat{\boldsymbol{X}}$, which minimizes the Frobenius distance to $\boldsymbol{X}$, is given by the k leading components of the singular value decomposition of $\boldsymbol{X}$ (Jolliffe, 2002).

Constrained factorizations. So far we referred to *unconstrained* matrix factorization, where $\boldsymbol{U}$ and $\boldsymbol{V}$ can vary over all matrices in the space $\mathbb{R}^{n\times k}$ and $\mathbb{R}^{m\times k}$, respectively. This means that $\widehat{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{V}^T$ is limited only by its rank. In data analysis it is often appropriate to additionally constrain the factor matrices, i.e. introduce additional regularization terms in the objective function. This can alleviate the interpretation of the factor matrices, or in order to reduce the complexity of the model, allow identification of more factors. Imposing constraints on the factor matrices removes the degrees of freedom on the factorization $\boldsymbol{U}\boldsymbol{V}^T$ of a reconstructed $\widehat{\boldsymbol{X}}$. Lee and Seung (2000) studied various constraints on the factor matrices, including a very popular constraint about non-negativity of the latent matrices. We refer the reader to Žitnik and Zupan (2012); Wang and Zhang (2013) for a comprehensive review of different types of regularization, such as nonnegativity, orthogonality, stochasticity, sparseness and preservation of local topological properties, and the relationships between them.

A unified view of matrix factorization. Recently, Singh and Gordon (2008b) presented a unified view of matrix factorization that frames the differences among popular methods, such as non-negative matrix factorization (Lee and Seung, 2000), weighted singular value decomposition (Srebro et al., 2003), exponential principal component analysis (Collins et al., 2001), maximum margin matrix factorization (Srebro et al., 2004), probabilistic latent semantic indexing (Hofmann, 1999), Bregman co-clustering (Gordon, 2002), and many others in terms of a small number of modeling choices.

Definition 1: A matrix factorization can be defined by the following choices, which are sufficient to include many "popular approaches" (Fig. 2.1):

1. Data weights $\boldsymbol{W} \in \mathbb{R}_{+}^{m \times n}$.
2. Prediction link $f : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{m \times n}$.
3. Hard constraints on factors, $\boldsymbol{U}, \boldsymbol{V} \in \mathscr{C}$.
4. Weighted loss between $\boldsymbol{X}$ and $\hat{\boldsymbol{X}} = f(\boldsymbol{U}\boldsymbol{V}^T)$, $\mathscr{D}(\boldsymbol{X} \| \hat{\boldsymbol{X}}, \boldsymbol{W}) \geq 0$.
5. Regularization penalty, $\mathscr{R}(\boldsymbol{U}, \boldsymbol{V}) \geq 0$.

Given these choices, the optimization for the model $\boldsymbol{X} \approx f(\boldsymbol{U}\boldsymbol{V}^T)$ is:

$$\underset{\boldsymbol{U}, \boldsymbol{V} \in \mathscr{C}}{\arg\min}\, \mathscr{D}(\boldsymbol{X} \| f(\boldsymbol{U}\boldsymbol{V}^T), \boldsymbol{W}) + \mathscr{R}(\boldsymbol{U}, \boldsymbol{V}). \tag{2.2}$$

Here, prediction link f allows nonlinear relationships between $\boldsymbol{U}\boldsymbol{V}^T$ and the data $\boldsymbol{X}$ (Singh and Gordon, 2008b).

A concept very closely related to matrix factorization is that of *matrix completion*. The aim of matrix completion is to recover an unknown matrix from a subset of its entries (Todeschini et al., 2013; Lee and Shraibman, 2013). The problem has received prominent attention in the context of recommendation systems, cf. e.g., Shi et al. (2012a). A central approach to this problem is to generate a matrix of the lowest possible complexity that agrees with the partially observed matrix. Here, complexity is typically measured using rank or trace norms. The performance of this approach has been well studied under the assumption that observed matrix entries are sampled uniformly at random (Candès and Recht, 2009; Candès and Tao, 2010).

Factor models in data analysis. Matrix factorization has turned out to be very good at discovering intricate structures in high-dimensional data and is therefore applicable to many domains of business, science and government. In addition to beating records in collaborative filtering and recommendation systems (Bell and Koren, 2007), it has had many successes in dimensionality reduction (Jolliffe, 2002; Li et al., 2009c; Maurus and Plant, 2014), clustering (Hochreiter et al., 2010; Arora et al., 2013) and low-rank

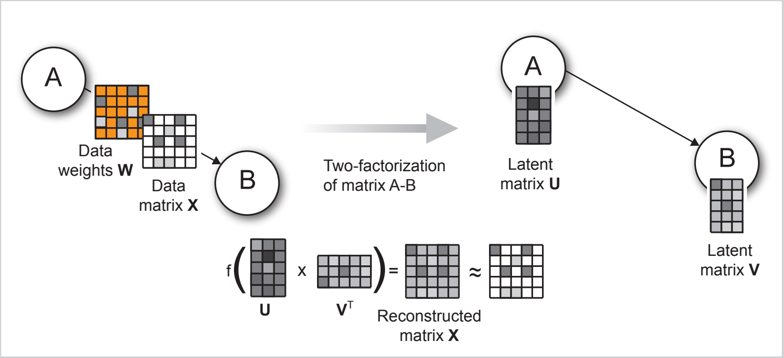


Figure 2.1

Matrix two-factorization. Given a data matrix X, prediction link f and data weights W, a model of matrix two-factorization optimizes for latent matrices U and V that satisfy potential constraints and minimize a measure of discrepancy between input data matrix and its reconstruction.

approximation (Matsushita and Tanaka, 2013), among others.

One way to measure the fit of a learned factor model is to use metric such as root mean squared error. This metric was adopted in the Netflix Prize Contest (`http://www.netflixprize.com`) as the evaluation metric. However, it is now recognized that approaches optimized to minimize the error rate can achieve poor performance on classification and ranking tasks (Rendle, 2010; Rendle et al., 2010). In collaborative filtering, users focus their attention on only a small number of recommendations, effectively ignoring all but a short list of recommended items. For this reason, the ultimate goal of factor models in collaborative filtering is to *generate a top-N list of relevant items to individual users.* Generation of recommendation lists is a ranking task, i.e. ranking items according to their relevance to the user. Consequently, new learning algorithms and factor models that are being developed optimize for a variety of metrics used for ranking, classification and regression (Rendle et al., 2009; Rendle, 2010; Shi et al., 2012b, 2013). Factor models can thus be applied not only to regression, where the estimated latent matrices can be used directly as predictors and the optimization criterion is e.g., the minimal least squared error, but also for binary classification, where parameters are optimized for hinge loss or logit loss (Rendle, 2010), and for ranking, where optimization is done with pairwise or listwise classification loss functions (Shi et al., 2013). This posits that factor models are general predictors working with any data matrix representation (Rendle, 2010). They model interactions between variables using factorized parameters and are capable of estimating interactions between variables even in problems with huge sparsity, such as recommendation systems, where other methods fail (Rendle, 2013).

2.1.2 *Important concepts from molecular biology*

> *"Computers are to biology what mathematics is to physics."*
>
> — Harold J. Morowitz

Next, we attempt to provide enough background for a computer scientist to be able to appreciate the relevance of biological applications studied in the Thesis. This section provides a very brief overview, interested reader is referred to Hunter (1993); Alberts et al. (2007) for a better understanding of cell biology.

Inherited characteristics of an organism are contained in a single molecule: deoxyribonucleic acid, or DNA. These characteristics are encoded in a simple, linear, four-element code, which is known as organism's *genotype*. The resulting physical properties of an organism are called its *phenotype*.

The composition of cells. Organisms can either be single-cellular or multi-cellular. The main advantage of multi-cellular organisms is *specialization*. This means that not every cell in a multi-cellular organism needs to be able to protect itself, extract nutrients, sense the environment, reproduced itself, etc. These complex tasks are typically divided so that many different classes of cells work together and accomplish tasks that single cells cannot. Groups of cells specialized for a particular function are *tissues*. We say that cells in a tissue have *differentiated.* When a cell differentiates, it typically cannot change from one type to another. Despite all of the variation, all cells in a multicellular organism have exactly the same genetic code. These differences can be explained by differences in *gene expression*, that is, whether or not the product that a gene codes for is produced, and how much of the product is produced. Genes code for products that turn on and off other genes, which in turn regulate other genes, and so on (Hunter, 1993). One of the key research areas in biology is development: how the interrelated genetic regulatory processes are managed, and how cells "know" what to differentiate into, and when and where they do it (Alberts et al., 2007).

Despite the many differences, most cells have a great deal in common with each other: they contain cytoplasm and genetic material, are enclosed by a membrane and have the basic mechanisms for *translating genetic material* into the main type of biological molecule, the *protein*.

Genetic material codes for all other parts of the cell. This information is typically stored in long strands of DNA. While proteins are the workhorses of the biochemical world, *nucleic acids*, e.g., DNA, are the drivers; they control the action. Besides DNA, another very important polymer is ribonucleic acid or RNA, which directs the synthesis of proteins. Both types of nucleic acids are polymers of four simple units called *nucleotides*. There are four nucleotides found in DNA: *adenine* (A), *guanine* (G), *cytosine* (C) and *thymine* (T). Nucleotides are sometimes called *bases*, and, since DNA consists of two complementary strands bonded together, these units are often called *base pairs*. In RNA, *uracil* (U) takes the place of thymine.

Proteins are the molecules that accomplish most of the functions of the living cell. The number of different functions and structures that proteins take on in a single organism is staggering. They make possible all of the chemical reactions in the cell by acting as *enzymes* that promote chemical reactions, which would otherwise occur slowly. Proteins also provide structural support and are vital for the immune system to distinguish itself from the invaders. They provide the means for acquiring and transforming energy, as well as the transmission of information. All proteins are constructed from linear sequences of smaller molecules known as *amino acids*. There are twenty naturally occurring amino acids. Long proteins may contain as many as 4,500 amino acids. Hence, the space of possible proteins is very large: 20^{4500} or 10^{5850}. Additionally, proteins fold up to form three dimensional shape, which give them their specific chemical functionality.

The defining part of eukaryotic cells are the *nuclei*. The nucleus contains the genetic material of the cell in the form of *chromatin*, i.e. long stretches of DNA in a variety of conformations.

Other important parts of cells include membranes, cytoplasm, ribosomes, mitochondria and chroloplasts, endoplasmic reticulum, Golgi appratus and lysosomes.

Genes, the genome and the genetic code. The genetic information of an organism can be stored in one or more distinct DNA molecules; each is called a *chromosome*. In some organisms, called *diploids*, each chromosome contains two similar DNA molecules that physically bound together, one from each parent. Human beings are diploid with 23 pairs of chromosomes. All of the genetic information of an organism is referred to as its *genome*. The primary role of nucleic acids is to carry the encoding of the primary structure of proteins. Each non-overlapping triple of nucleotides is called a *codon* and corresponds to a particular amino acid. Four nucleotides can form $4^3 = 64$ possible triplets, which is more than the 200 triplets that are needed to code for each amino acid. Most amino acids are encoded by more than one codon. For example, alanine is represented in DNA by the codons GCT, GCC, GCA and GCG.

The basic process of synthesizing proteins involves mapping a sequence of codons, i.e. a *gene*, to a sequence of amino acids, i.e. a protein. However, there are many important complications. The structure of a gene typically consists of many elements of which the actual protein coding sequence may be only a small part. The non-coding sequences are called *introns* and are *spliced out* before the sequence is mapped into amino acids. The segments of DNA that actually end up coding for a protein, i.e. segments that get *expressed*, are called *exons*. DNA contains a large amount of information in addition to the coding sequences of proteins (Hunter, 1993; Alberts et al., 2007).

Transcription and translation. The process of mapping a DNA sequence to a folded protein in eukaryotes involves many steps. The most important steps are: (1) *transcription,* which transforms a portion of DNA into an RNA molecule called a messenger RNA (mRNA); (2) *intron splicing,* which splices the exons together; (3) *translation,* which uses mRNA as a blueprint for the production of a protein at the ribosome; and (4) *protein folding* and *post-translational modifications*. Once the protein has folded, other transformations can occur. Various chemicals can be bound to different places on the proteins, which can change the shape of the protein, and may be necessary to make the protein active, or may keep it from having an effect before it is needed.

Model organisms. Model organisms are a vital source of biological knowledge. The investigation of even a single organism can take many scientists many careers worth of time. Moreover, biological experimentation is often complex, time consuming and difficult. Some of the most valuable biological methods are invasive, or require organisms to be sacrificed, or require many generations of observation, or observations on large populations (Hunter, 1993). Such work is impractical or unethical to carry out on humans. Hence, biologists have selected a variety of *model organisms for experimentation*. These creatures have qualities that make possible controlled laboratory experiments at reasonable cost and difficulty with results that can often be *translated to people.* The main models used in molecular biology include: bacterium *Escherichia coli*, brewer's yeast *Saccharomyces cervesiae*, common weed *Arabidopsis thaliana*, common fruit fly *Drosophila melanogaster*, mouse *Mus musculus*, nematode worm *Caenorhabditis elegans*, and amoeba *Dictyostelium discoideum*.

2.2 *Machine learning approaches to data integration*

Computational methods for integrative data analysis are capable of analyzing heterogeneous data. These methods combine data arising from diverse background distributions, relations, dimensions and formats to enhance the statistical significance and to obtain more accurate predictive models (Boström et al., 2007). *Data fusion,* a term borrowed from engineering (Hall and Llinas, 1997), has in recent years emerged in various areas of predictive modeling to reflect combining distinct heterogeneous data sources, even when they differ in their conceptual, contextual and typographical representations (Aerts et al., 2006).

Data heterogeneity may arise due to various reasons. It may be due to differences in data extraction methods or different perspectives/scales at which the problem of interest is being studied. Furthermore, there might be heterogeneity at the measurement scale, data dimensionality or the types of features. For example, data representations range from high-resolution images, text documents, feature-based data tables to structured data, such as networks, hierarchies of associations and ontologies. Different data types naturally use different formats and can be nominal, ordinal, represented with intervals, ratios, etc. Some of the successful applications include integrative methods for gene prioritization (Aerts et al., 2006; Sifrim et al., 2013), gene and protein function prediction (Savage et al., 2010; Saddiki et al., 2014; Klein et al., 2014), signal processing (Subrahmanya and Shin, 2010), visual object recognition (Bucak et al., 2014), information retrieval (Dwork et al., 2001; Zhu et al., 2013), network analysis (Tang

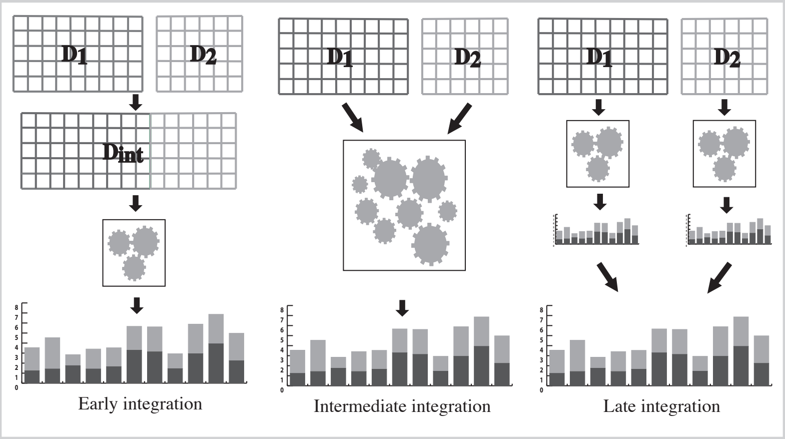


Figure 2.2

Data integration strategies. Early integration transforms all data sets into a single, feature-based table and treats this as a single data set that can be explored by any of the well-established attribute-based machine learning algorithms. It relies on procedures for feature construction and often neglects possible relational structure of the data. In late integration, each data set gives rise to a separate model. Prior to model inference, it is necessary to transform each data set to encode relations to the target concept (non-trivial). Intermediate integration actively includes additional information in the method itself.

et al., 2012), and text processing (Lin and Kolcz, 2012; Rebholz-Schuhmann et al., 2012).

According to Pavlidis et al. (2002); Schölkopf et al. (2004); Maragos et al. (2008); Greene and Cunningham (2009), data fusion approaches can be classified into three main categories depending on the modeling stage at which fusion takes place (Fig. 2.2). In *early integration*, features from different sources are concatenated and fed to a single learner. *Late integration* involves feeding different features to different classifiers whose decisions are then combined by a fixed or trained combiner. The youngest branch of data fusion algorithms is *intermediate integration*. Intermediate integration does not merge the input data, nor does it develop separate models for each data source. It instead retains the structure of the data sources by incorporating it within the structure of predictive model. Algorithms in this category explicitly address the multiplicity of data and fuse them through inference of a single joint model by actively including additional information by the fusion algorithm itself. This particular approach is often preferred because of its superior predictive accuracy (Pavlidis et al., 2002; Lanckriet et al., 2004c; Gevaert et al., 2006; Tang et al., 2009; van Vliet et al., 2012), but for a given model type, it requires the development of a new inference algorithm.

Although there are many applications, which attempt to estimate prediction models from heterogeneous data, most often these are heuristic approaches that depend heavily on specific problems that are being studied. Such methods might be difficult to generalize. On the other hand, *kernel-based methods* and *graphical models* are two general approaches with many successful applications in learning from multiple data sets. This Thesis is premised on the notion that *collective latent factor models* represent another group of general machine learning predictors that are appropriate for data fusion.

In what follows we briefly overview each of the three classes of methods for integrative data analysis. Further related work focused on a specific field of study, e.g., network inference or gene function prediction, is provided in the respective chapters of the Thesis.

2.2.1 *Graphical model-based methods*

Bayesian modeling has been widely used in multi-task learning and multi-view learning over the last decade, where the goal has been to harness multiple data views, i.e. data sets, describing a given set of objects and to leverage related tasks to improve the learning performance in each task. Research dedicated to Bayesian hierarchical modeling has demonstrated effectiveness and improvement in predictive performance (Bakker and Heskes, 2003; Guo et al., 2011; Han et al., 2012). These methods have been successfully applied to areas, such as information retrieval (Blei et al., 2004) and computer vision (Luo et al., 2013; Ding et al., 2012). Typical approaches to transfer information among multiple views and tasks include: sharing hidden nodes in neural networks, placing a common prior in hierarchical models (He and Lawrence, 2011; Zhang et al., 2011a; Yang and He, 2014), sharing a common structure on the predictor space (Yu et al., 2005; He et al., 2014), and structured regularization in kernel methods (He and Lawrence, 2011), among others.

2.2.2 *Multiple kernel-based methods*

Kernel methods are nonparametric learning methods that use kernel functions (Shawe-Taylor and Cristianini, 2004) to implicitly define the similarity of a pair of data points according to the features describing them. There are several advantages to the use of kernel methods for data fusion. Due to nonparametric characteristic of the kernels, one does not need to make prior assumptions about data distributions. Furthermore, kernel functions can effectively model nonlinear relationships between data features. Also, since the size of the kernel matrices depend only on the number of data points and not on the number of features, kernel-based methods are suitable for high-dimensional data with many features. Most popular kernel functions include linear, polynomial and Gaussian, although many other forms, e.g., diffusion, string and tree kernel functions (Lodhi et al., 2002; Zhu et al., 2004; Da San Martino et al., 2012), have been successfully employed.

Multiple kernel learning. In recent years, several methods have been proposed to combine multiple kernels instead of using a single one (Gönen and Alpaydın, 2011). Multiple kernels are useful for modeling either a single homogeneous data set or many heterogeneous data sets. In the first setting, one can vary kernel functions and their parameters to construct multiple kernel matrices over a given data set. For the purpose of integrative analysis on heterogeneous data, a separate kernel matrix can be created for each data set.

In contrast to Bayesian modeling, multiple kernel learning typically does not need to model prior data distributions and relationships between different types of features. However, selecting appropriate kernel function and its parameters in an important issue in multiple kernel learning methods. Typically, a cross-validation procedure is used to choose the best performing kernel function among a set of kernel functions on a separate validation set different from the training set.

A common implementation of multiple kernel learning can be seen as a technique, that optimizes the parameters used to combine a set of predefined kernels, i.e. we assume that kernel functions and the corresponding kernel parameters are known before training (Qiu and Lane, 2009; Cortes et al., 2010). It is also possible to enhance multiple kernel learning, such that parameters integrated into the kernel functions are optimized during training (Yang et al., 2009; Gönen and Alpaydin, 2008). Most of the existing algorithms fall into the first category and *try to combine predefined kernels in an optimal way.*

The reasoning behind combining many kernels is similar to combining different classifiers. Instead of choosing a single kernel function, it is better to have a set promising kernel functions and let an algorithm do the selection of a kernel or their combination. There can be two uses multiple kernel learning:

- First, different kernels correspond to different notions of similarity and instead of trying to find, which works best, a learning method does the picking for us, or may use a combination of them. Using a specific kernel may be a source of bias, and in allowing a learner to choose among a set of kernels, a better solution can be found.

- Second, different kernels may be using inputs coming from different representations, from different sources or modalities. Since these are different representations, they have different measures of similarity corresponding to different kernels. In such a case, combining kernels is one possible way to combine multiple information sources in a sense typical of intermediate data integration.

There are different ways in which kernel combination can be done. The most popular are methods that combine kernels via an unweighted sum, i.e. using the sum or the mean of the kernels as the combined kernel, or through a weighted linear combination (Lanckriet et al., 2004b). Other multiple kernel learning algorithms use nonlinear functions of kernels, e.g., multiplication, power, exponentiation (Varma and Babu, 2009), or use specific kernel weights for each data point determined in a data-driven way (Yang et al., 2010).

2.2.3 *Collective latent factor models and* parameter sharing

A collective latent factor model typically factors each data matrix using a generalized-linear link function, but *whenever an object type is involved in more than one data relation, it ties the factors of respective relations together.*

Multi-object type latent factor models. Wang et al. (2008) and Wang et al. (2011a) proposed tri-SPMF and S-NMTF, respectively, a collective clustering of multi-type relational data via symmetric nonnegative matrix tri-factorization. These two methods consider both *inter-type* data relations, i.e. relationships between objects of *different types*, and *intra-type* data relations, i.e. relationships between objects of the *same type*.

Data sets are viewed as dyadic relations and are encoded in relation and constraint matrices. A relation matrix $\boldsymbol{R}_{ij}$ is a $n_i \times n_j$ real-valued matrix, in which rows correspond to objects of type i, columns to objects of type j and the element $\boldsymbol{R}_{ij}(k,l)$ represents the relationship between objects k and l. A constraint matrix $\boldsymbol{\Theta}_i$ is a $n_i \times n_i$ real-valued matrix that relates objects of type i to themselves. It contains pairwise constraints indicating dissimilarity/similarity between objects. The objective function of a latent factor model is such that latent matrices minimizing it achieve good reconstruction of observed elements in the relation matrices and adhere to the constraints (Fig. 2.3). For example, in matrix tri-factorization models, $\boldsymbol{R}_{ij}$ is approximated by three latent matrices such that $\boldsymbol{R}_{ij} \approx \boldsymbol{F}_{ij}\boldsymbol{S}_{ij}\boldsymbol{G}_{ij}^T$, where $\boldsymbol{F}_{ij} \in \mathbb{R}^{n_i \times k_{ij}}$, $\boldsymbol{S}_{ij} \in \mathbb{R}^{k_{ij} \times c_{ij}}$ and $\boldsymbol{G}_{ij} \in \mathbb{R}^{n_j \times c_{ij}}$. Here, k_{ij} and c_{ij} are factorization ranks, which typically in predictive modeling are substantially smaller than the original data dimensionality, $k_{ij} \ll n_i$, $c_{ij} \ll n_j$. Since profiles, i.e. row vectors in $\boldsymbol{R}_{ij}$, of many objects are represented by relatively few vectors from $\boldsymbol{S}_{ij}$ and low dimensional vectors in $\boldsymbol{G}_i$ and $\boldsymbol{G}_j$, a good approximation can only be achieved if these vectors span a space that reveals some latent structure present in the original data (Fig. 2.3). *Collective factor models* of Wang et al. (2008) and Wang et al. (2011a) are able to tri-factorize a *collection of relation and constraint matrices* but require that relations between any two modeled object types are available. This requirement is rarely satisfied in real-world data fusion settings, where we do not have access to relation matrices between all possible pairs of object types. While these models require little engineering by hand and can take advantage of increases in the amount of available data, new *generic* and *effective* learning algorithms that are currently being developed for collective data analysis will extend their applicability to various data domains and prediction tasks.

In the context of text processing, matrix tri-factorization can be interpreted as follows (Li et al., 2009b). Given a term-by-document matrix $\boldsymbol{R}_{ij}$, latent matrices $\boldsymbol{F}_{ij}$ and $\boldsymbol{G}_{ij}$ specify soft membership of terms and documents in one of k_{ij} and c_{ij} classes, respectively, where $\boldsymbol{F}_{ij}$ represents knowledge in the word space, i.e. i-th row of $\boldsymbol{F}_{ij}$ represent the posterior probability of word i belonging to the k_{ij} classes, and $\boldsymbol{G}_{ij}$ represents knowledge in document space, i.e. the i-th row of $\boldsymbol{G}_{ij}$ represents the posterior probability of document i belonging to the c_{ij} classes. A third latent factor, $\boldsymbol{S}_{ij}$, provides a condensed view of $\boldsymbol{R}_{ij}$. When performing collective matrix decomposition, the strategy of sharing latent factors between relation matrices depends on a particular design of collective latent model (Wang et al., 2008, 2011a; Žitnik and Zupan, 2015a).

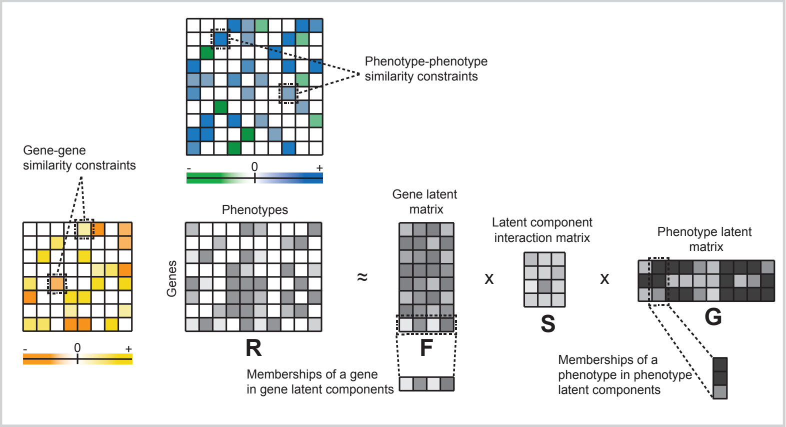

Phenotype-phenotype similarity constraints
Gene-gene similarity constraints
-
0
+
Phenotypes
Genes
R
≈
Gene latent matrix
F
x
Latent component interaction matrix
S
x
Phenotype latent matrix
G
Memberships of a gene in gene latent components
Memberships of a phenotype in phenotype latent components

Figure 2.3

Constrained matrix tri-factorization of an exemplar gene-by-phenotype matrix. The matrix is accompanied by two constraint matrices expressing similarity between genes (matrix in yellow and orange) and between phenotypes (matrix in blue and green). Elements in constraint matrices that have greater negative values are *must-link constraints*, i.e. the corresponding genes (phenotypes) should have more similar latent profiles. Positive values in constraint matrices represent *cannot-link constraints* and penalize the latent data model if the corresponding genes (phenotypes) have similar latent profiles.

Multiple data features of different types can also be modeled with tensor decompositions. However, in present tensor decompositions (Kolda and Bader, 2009; Sutskever, 2009; Rendle et al., 2011; Rettinger et al., 2012; Xu et al., 2014), tensors become increasingly sparse and computationally intractable for higher dimensions.

Multi-relational latent factor models. Zhang et al. (2012) proposed a collective matrix factorization to decompose a number of data matrices $\boldsymbol{R}_i$ into a common latent matrix $\boldsymbol{W}$ and different coefficient matrices $\boldsymbol{H}_i$, such that $\boldsymbol{R}_i \approx \boldsymbol{W}\boldsymbol{H}_i$ by minimizing $\sum_i ||\boldsymbol{R}_i - \boldsymbol{W}\boldsymbol{H}_i||^2_{\text{Fro}}$. This is an intermediate integration approach but it can only describe relations that involve fixed objects across data matrices. Similar two-factor approaches but with various regularization types were also proposed (Li and Yeung, 2007; Zhang et al., 2011b; Singh and Gordon, 2008a, 2010).

There is an abundance of work on factorized models that consider a *single data matrix* or *multiple data matrices over the two types of objects* (Wang et al., 2008; Sutskever, 2009; Li et al., 2009a; Wang et al., 2012). For example, Nickel et al. (2011) proposed a tri-factorization model for multiple dyadic relations, which factorizes every $\boldsymbol{R}_i$ as $\boldsymbol{R}_i \approx \boldsymbol{A}\boldsymbol{S}_i\boldsymbol{A}^T$, where latent matrix $\boldsymbol{A}$ is shared between all data relations.

Part I

Network side information

3

Prior knowledge presented with networks

Matrix completion is among the most popular techniques in relational learning, where one of its most celebrated application areas include collaborative filtering. One challenge of matrix completion is how to utilize available auxiliary information to improve prediction accuracy.

In this chapter we study the problem of including side information as an additional feature of matrix completion. We incorporate the *mechanism of information propagation over the networks* into the factorized model in a principled way. To inject network influence in our model we make latent features of every object dependent on the latent features of its direct neighbors in the network. Using this idea, latent features of objects indirectly connected in the network become dependent and hence information gets propagated.

Cold start objects, e.g., genes for which no measurements are available, are an important challenge in matrix completion models. Since cold start objects rely more on the auxiliary information compared to the objects with many measurements, the effect of using the principle of information propagation is vital for poorly characterized objects. Moreover, in many genomic data sets a very large portion of genes might not be considered in any of the experiments for various reasons, such as gene essentiality, but these genes appear in the background knowledge represented here in the form of gene networks. Hence, using only observed measurements would not allow to learn the latent features for such genes. The model presented in this chapter forces gene feature vectors to be close to those of their neighbors. As such, the model is capable of learning the latent features for genes with no or very few measurements.

We have conducted experiments on several large-scale genetic interaction data sets. Our experiments demonstrate that modeling propagation of information over the networks while inferring a latent factor model leads to a substantial increase in prediction accuracy, in particular for cold start genes.

3.1 Background

The epistatic miniarray profile (E-MAP) technology (Schuldiner et al., 2005; Collins et al., 2006; Roguev et al., 2008; Wilmes et al., 2008; Surma et al., 2013) is based on a synthetic genetic array (SGA) approach (Tong et al., 2001, 2004) and generates quantitative measurements of both positive and negative genetic interactions (GIs) between genes. E-MAP was developed to study the phenomenon of epistasis, wherein the presence of one mutation modulates the effect of another mutation. The power of epistasis analysis is greatly enhanced by quantitative measurements of interactions (Collins et al., 2006). E-MAP has provided high-throughput measurements of hundreds of thousands of GIs in yeast (Schuldiner et al., 2005; Collins et al., 2007; Wilmes et al., 2008) and has been shown to significantly improve gene function prediction (Collins et al., 2007). However, E-MAP data suffer from a large number of missing values that can be as high as ∼40% for a given assay (see also Table 3.1). Missing values correspond to pairs of genes for which the strength of the interaction could not be measured during the experimental procedure or that were subsequently removed due to low reliability. A high proportion of missing values can adversely affect analysis algorithms or even prevent their use (Nanni et al., 2012). Missing data can introduce instability in clustering results (de Brevern et al., 2004) or bias the inference of prediction models (Liew et al., 2011). Accurate imputation of quantitative GIs is therefore an appealing option to improve downstream data analysis and correspondence between genetic and functional similarity (Collins et al., 2007; Pu et al., 2008; Bandyopadhyay et al., 2008; Ulitsky et al., 2008; Järvinen et al., 2008).

The missing value problem in E-MAPs resembles that from gene expression data where imputation has been studied well (Troyanskaya et al., 2001; Brock et al., 2008; Liew et al., 2011). The objective of both tasks is to estimate the values of missing elements in the given incomplete data matrix. Both types of data may exhibit correlation between mutant and gene profiles that is indicative of pathway membership in the case of E-MAP data (Ryan et al., 2010) and co-regulation in the case of gene expression data. E-MAP data sets are therefore often investigated with tools originally developed for gene expression data analysis (Zheng et al., 2010). However, there are important differences between E-MAP and gene expression data that limit direct application of gene expression imputation techniques to E-MAPs (Ryan et al., 2010). E-MAP matrices report on pairwise relations between genes and have substantially different dimensionality than gene expression data sets. They often contain substantially more missing values than gene expression data sets with the latter having up to 5% missing data rate (Bø et al., 2004; Liew et al., 2011). These differences coupled with the biological significance of E-MAP studies have spurred the development of specialized computational techniques for recovery of missing interaction measurements in E-MAP-like data sets (Ryan et al., 2010).

In this chapter we present network-guided matrix completion (NG-MC), a *hybrid* and *knowledge-assisted* method for imputing missing values in E-MAP-like data sets. NG-MC builds upon two concepts: probabilistic matrix completion and propagation of NG-MC-inferred latent gene interaction profiles. Matrix completion uses information on global correlation of elements in the E-MAP score matrix. Propagation of latent profiles exploits the local similarity of genes as specified by the gene networks. The use of prior knowledge in the form of gene networks gives NG-MC the potential to improve imputation accuracy beyond purely data-driven approaches. This could be especially important for data sets with small number of genes and high missing data rate such as E-MAPs. In what follows we present mathematical formulation of the proposed approach and in a comparative study that includes several state-of-the-art imputation techniques demonstrate its accuracy across several E-MAP data sets.

3.2 Related work on data imputation

Imputation algorithms for gene expression data sets are reviewed in Liew et al. (2011) where they are categorized into four classes based on how they utilize or combine local and global information from within the data (*local, global* and *hybrid* algorithms) and their use of prior knowledge in imputation (*knowledge-assisted* algorithms). Local methods based on k-nearest neighbors that include KNNimpute (Troyanskaya et al., 2001), local least squares (LLS) (Kim et al., 2005) and adaptive least squares (LSimpute) (Bø et al., 2004) rely on local similarity of genes to recover missing values. Global methods decompose data matrices using variations of singular value decomposition (SVDimpute) (Troyanskaya et al., 2001), singular value thresholding algorithm for matrix completion (SVT) (Cai et al., 2010) and Bayesian principal component analysis (BPCA) (Oba et al., 2003). Hybrid imputation approaches for gene expression data make predictions by combining estimates from both local and global imputation methods (Jörnsten et al., 2005).

Only a handful of missing data imputation algorithms directly address E-MAP-like data sets. Ulitsky et al. (2009) experimented with a variety of genomic features, such as the existence of physical interaction or co-expression between genes, that were used as input to a classification algorithm. The NG-MC differs from this approach as it directly uses the matrix of measured GI scores and does not require data-specific feature engineering. Ryan et al. (2010, 2011) considered four general strategies for imputing missing values—three local methods and one global method—and adapted these strategies for E-MAPs. They modified unweighted and weighted k-nearest neighbors imputation methods (uKNN and wNN, respectively) ans adapted LLS and BPCA algorithms to handle symmetric E-MAP data. We refer the reader to Ryan et al. (2010) for details on the algorithm modifications. We compare their imputation approaches with the NG-MC (Sec. 3.4). Pan et al. (2011) proposed an ensemble approach to combine the outputs of two global and four local imputation methods based on diversity of estimates of individual algorithms. In this chapter, we focus on the development of a single algorithm that, if necessary, could be used in an ensemble, and therefore compare it with ensemble-free algorithms.

Another venue of research focuses on predicting qualitative, *i.e.* binary instead of quantitative interactions. Here, predictions estimate the presence or absence of certain types of interactions rather than their strength (Wong et al., 2004; Kelley and Ideker, 2005; Qi et al., 2008; Pandey et al., 2010). A major distinction between these techniques and the method presented in this chapter is that we aim to accurately impute quantitative genetic interactions using the scale of GI scores. Individual GI may by itself already provide valuable biological insight as each interaction attests to a functional relationship of a pair of genes. Prediction of synthetic sick and lethal interaction types in *S. cerevisiae* was pioneered by Wong et al. (2004), who applied probabilistic decision trees to diverse genomic data. Wong et al. introduced *2-hop features* to capture the relationship between a pair of genes and a third gene. For example, if protein g physically interacts with protein h, and gene w is synthetic lethal with the encoding gene of h, then this observation increases the likelihood of a synthetic lethal interaction between the encoding gene of g and gene w. Two-hop features were shown to be crucial when predicting GIs (Wong et al., 2004; Bandyopadhyay et al., 2008; Ulitsky et al., 2009) and are the rationale behind our concept of propagating latent profiles over gene networks.

3.3 Network-guided matrix completion

We start by presenting a probabilistic model of matrix completion for missing value imputation in E-MAP-like data sets in which the prediction of missing interaction measurement depends only on the E-MAP score matrix. We then develop an efficient model fitting approach called network-guided matrix completion (NG-MC), which can additionally consider the prior knowledge in the form of any number of gene networks. NG-MC uses information on topology of gene networks to propagate latent gene interaction profiles among neighboring genes. It exploits the transitivity of interactions, that is, the property of the relationship between a gene pair and a third gene (Sec. 3.2). As such, NG-MC predicts missing values by integrating E-MAP data with available network data. Any type of knowledge that can be expressed in the form of gene networks can be passed to NG-MC. In our experiments we consider Gene Ontology (Ashburner et al., 2000) semantic similarity network and protein-protein interaction network.

3.3.1 Problem definition

In the E-MAP study we have a set of n genes, $\{g_1, g_2, \dots, g_n\}$. Genetic interaction of two genes is scored according to the fitness of the corresponding double mutant and reported with an S-score, which reflects both the magnitude and the sign of observed interaction measurement (Collins et al., 2006). Scored GIs are reported in partially observed matrix $\boldsymbol{G} \in \mathbb{R}^{n \times n}$. In this matrix, the element $\boldsymbol{G}_{ij}$ contains measurement of GI between g_i and g_j. We assume that $\boldsymbol{G}$ is symmetric, $\boldsymbol{G}_{ij} = \boldsymbol{G}_{ji}$, and has its values scaled to $[0, 1]$-interval. Genetic interactions are mapped to $[0, 1]$-interval by normalizing $\boldsymbol{G}$ before data imputation is performed.

Network-guided matrix completion can simultaneously consider multiple gene networks. Given a weighted adjacency matrix $\boldsymbol{P} \in \mathbb{R}^{n \times n}$ of a gene network from a collection of networks $\mathscr{P}$, N_g^P denotes a set of direct neighbors of g in $\boldsymbol{P}$, where for $h \in N_g^P$ the value $\boldsymbol{P}_{gh}$ ($\boldsymbol{P}_{gh} \neq 0$) represents the strength of association of gene g with gene h. Prior to the inference of factorized model we normalize each row of $\boldsymbol{P}$ by the sum of the weights of incident edges such that $\sum_{j=1}^{n} \boldsymbol{P}_{ij} = 1$ for all i. A non-zero entry $\boldsymbol{P}_{gh}$ denotes the dependence of g-th latent feature vector on h-th latent feature vector.

Using this idea, latent features of genes that are indirectly connected in the network $\boldsymbol{P}$ become dependent after a certain number of algorithm iterations, the number of steps being determined by the distance between genes in the network. Hence, information about gene latent representation propagates through network $\boldsymbol{P}$.

The model inference task is defined as follows: given a pair of genes, g_i and g_j, for which $\boldsymbol{G}_{ij}$ (and $\boldsymbol{G}_{ji}$) is unknown, predict quantitative GI between g_i and g_j using $\boldsymbol{G}$ and $\mathscr{P}$. Let $\boldsymbol{F} \in \mathbb{R}^{k \times n}$ and $\boldsymbol{H} \in \mathbb{R}^{k \times n}$ be gene latent feature matrices with column vectors $\boldsymbol{F}_i$ and $\boldsymbol{H}_j$ representing k-dimensional gene-specific latent feature vectors of g_i and g_j, respectively. Let $\boldsymbol{W} \in \mathbb{R}^{n \times |\mathscr{P}|}$ be the networks weighting matrix where $\boldsymbol{W}_{ip}$ represents the influence of g_i's neighborhood in $\boldsymbol{P} \in \mathscr{P}$ on the latent feature vector of g_i. Network-guided matrix completion infers gene latent feature matrices and network weighting matrix and utilizes them for missing value imputation in E-MAP-like data sets.

3.3.2 Preliminaries

We begin with a probabilistic view of matrix completion for missing value imputation that does not consider prior biological knowledge. This approach builds upon probabilistic matrix factorization of Mnih and Salakhutdinov (2007) and Salakhutdinov and Mnih (2008) and we refer to it as MC. Genome-scale genetic interaction mapping (Costanzo et al., 2010) has suggested the existence of coherent groups of genes participating in related biological processes. Hence, a desirable computational model of interactions should model interactions not only in terms of pairwise measurements, but also in terms of how these measurements relate to each other. Matrix completion models this intuition by assuming E-MAP score matrix $\boldsymbol{G}$ has low rank and factorizes observed values in $\boldsymbol{G}$ into a product of two low-dimensional latent feature matrices, $\boldsymbol{F}$ and $\boldsymbol{H}$. In order to learn gene latent feature matrices MC formulates the conditional probability of observed interactions as:

$$p(\boldsymbol{G}|\boldsymbol{F},\boldsymbol{H},\sigma_G^2) = \prod_{i=1}^{n}\prod_{j=1}^{n} \mathcal{N}(\boldsymbol{G}_{ij}|g(\boldsymbol{F}_i^T \boldsymbol{H}_j), \sigma_G^2)^{I_{ij}^G}, \tag{3.1}$$

where $\mathcal{N}(x|\mu,\sigma^2)$ is Gaussian distribution with mean μ and variance σ^2 and $\boldsymbol{I}_{ij}^G$ is an indicator function that is equal to 1 if the interaction measurement of g_i and g_j is available and is equal to 0 otherwise. As such, the conditional probability of interaction data regards only observed entries in matrix $\boldsymbol{G}$. It should be noted that predictions of matrix completion are not biased by a priori setting the missing entries in $\boldsymbol{G}$ to some fixed value selected in an ad hoc manner, which is otherwise common in matrix factorization algorithms (Lee and Seung, 2000; Lee et al., 2012; Wang et al., 2013). Another appealing property of matrix completion is sharing of gene latent feature vectors between all estimates of interaction measurements that involve a certain gene. In particular, latent feature vector $\boldsymbol{F}_i$ is used in estimations of interaction measurements $\boldsymbol{G}_{ij}$ for all j. Similar factor sharing is used in estimations of $\boldsymbol{H}$. The function g is a logistic function, $g(x) = 1/(1 + e^{-0.5x})$, which bounds the range of $g(\boldsymbol{F}_i^T \boldsymbol{H}_j)$ within interval $(0, 1)$. Our assumption of Gaussian distribution in Eq. (3.1) is justified by the scoring scheme of genetic interactions in E-MAP technology that uses a modified t-value score, called S-score (Collins et al., 2006). We further assume a zero-mean Gaussian prior for gene latent feature vectors in $\boldsymbol{F}$ given by $p(\boldsymbol{F}|\sigma_F^2) = \prod_{i=1}^n \mathcal{N}(\boldsymbol{F}_i|0, \sigma_F^2 \boldsymbol{I})$ and similarly, endow $\boldsymbol{H}$ with Gaussian prior distribution, $p(\boldsymbol{H}|\sigma_H^2) = \prod_{i=1}^n \mathcal{N}(\boldsymbol{H}_i|0, \sigma_H^2 \boldsymbol{I})$, parameterized by σ_F^2 and σ_H^2, respectively.

Through Bayesian inference we obtain the log-posterior probability of latent feature matrices given the interaction measurements, $p(\boldsymbol{F}, \boldsymbol{H}|\boldsymbol{G}, \sigma_G^2, \sigma_F^2, \sigma_H^2)$. We then select the factorized model consisting of $\boldsymbol{F}$ and $\boldsymbol{H}$ by finding maximum *a posteriori* estimate with gradient descent technique while keeping the observation noise variance σ_G^2 and prior variance σ_F^2 and σ_H^2 fixed.

3.3.3 Network-guided matrix completion

Network-guided matrix completion (NG-MC) extends matrix completion model (MC) from the previous section by borrowing latent feature information from neighboring genes in networks $\mathscr{P}$.

An illustration of NG-MC algorithm with prior knowledge in the form of a gene network is shown in Fig. 3.1. The figure shows a hypothetical E-MAP data set with five genes given, $\{g_1, \ldots, g_5\}$. Prior knowledge is presented through a gene network $\boldsymbol{P}$. Gene interaction profiles are listed next to corresponding nodes in gene network $\boldsymbol{P}$ (left in Fig. 3.1) and are shown in the sparse and symmetric matrix $\boldsymbol{G}$ (right in

Fig. 3.1).Different shades of grey quantify interaction strength while white elements in $\boldsymbol{G}$ denote missing values. Matrices $\boldsymbol{F}$ and $\boldsymbol{H}$ are gene latent feature matrices. Gene latent feature vector $\boldsymbol{F}_{g_i}$ depends in each iteration of the NG-MC on the latent feature vectors of g_i's direct neighbors in $\boldsymbol{P}$. For instance, the latent vector of gene g_1 in $\boldsymbol{F}$ depends in the first iteration of the NG-MC algorithm on latent vectors of its neighbors g_4 and g_5 ($\boldsymbol{F}_{g_4}$ and $\boldsymbol{F}_{g_5}$ are shown on input edges of g_1) whose degrees of influence are determined by $\boldsymbol{P}_{14}$ and $\boldsymbol{P}_{15}$, respectively. In the second iteration, the update of $\boldsymbol{F}_{g_1}$ depends also on the latent vector of g_1's 2-hop neighbor, g_2, hence the influence of gene latent feature vectors propagates through $\boldsymbol{P}$. Gene latent feature matrix $\boldsymbol{H}$ is not influenced by gene neighborhoods in $\boldsymbol{P}$.

The biological motivation for the propagation of interactions stems from the transitive relationship between a gene pair and a third gene (see Sec. 3.2) and indicates that the behavior of a gene is affected by its direct and indirect neighbors in the underlying gene networks $\mathscr{P}$. In other words, the latent feature vector of gene g, $\boldsymbol{F}_g$, is in each iteration of NG-MC algorithm dependent on the latent feature vectors of its direct neighbors $h \in N_g$ in networks $\mathscr{P}$. The influence is formulated as $\widehat{\boldsymbol{F}}_g = \sum_{P\in\mathscr{P}} \boldsymbol{W}_{gp} \sum_{h\in N_g} \boldsymbol{P}_{gh}\boldsymbol{F}_h$, where $\widehat{\boldsymbol{F}}_g$ is the estimated latent feature vector of g given feature vectors of its direct neighbors and $\boldsymbol{W}_{gp}$ is the weight of g in network $\boldsymbol{P}$ as inferred by NG-MC. Thus, latent feature vectors in $\boldsymbol{F}$ of genes that are indirectly connected in networks $\mathscr{P}$ are dependent and hence information about their latent representation propagates according to the connectivity of gene networks as the NG-MC algorithm progresses.

Suppose that for a given i and j, the observation in $\boldsymbol{G}_{ij}$ comes from distribution:

$$\mathcal{N}(\boldsymbol{G}_{ij}|g(\boldsymbol{F}_i^T\boldsymbol{H}_j),\sigma_G^2). \tag{3.2}$$

Considering that interaction measurements are generated independently, we model partially observed matrix $\boldsymbol{G}$ as:

$$p(\boldsymbol{G}|\boldsymbol{F},\boldsymbol{H},\sigma_G^2) = \prod_{i=1}^{n}\prod_{j=1}^{n} \mathcal{N}(\boldsymbol{G}_{ij}|g(\boldsymbol{F}_i^T\boldsymbol{H}_j),\sigma_G^2)^{I_{ij}^G}. \tag{3.3}$$

We achieve the coupling of interaction measurements by sharing latent gene profiles among all measurements of a certain gene. Note that incorporating prior knowledge

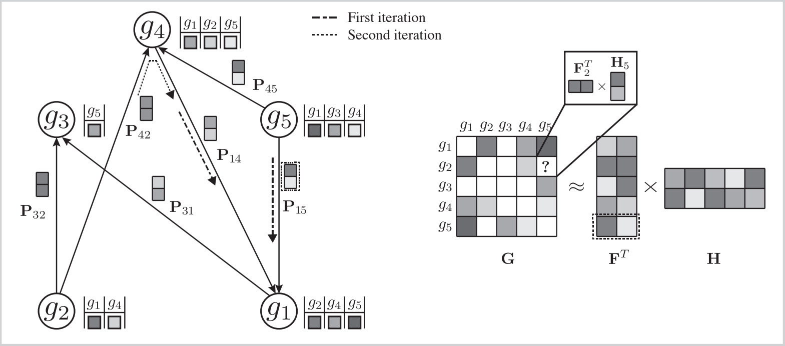


Figure 3.1

A toy application of network-guided matrix completion (NG-MC) algorithm. Given is a hypothetical E-MAP data set with five genes, $\{g_1, \dots, g_5\}$. Prior knowledge is presented through a gene network $\boldsymbol{P}$. Gene interaction profiles are listed next to corresponding nodes in gene network $\boldsymbol{P}$ (left) and are shown in the sparse and symmetric matrix $\boldsymbol{G}$ (right). See Sec. 3.3.3 for further explanation.

in the form of gene networks $\mathscr{P}$ does not change our probabilistic model of observed interaction measurements from Eq. (3.1). Instead, it only affects the formulation of gene latent feature vectors in $\boldsymbol{F}$. We describe them with two factors: a zero-mean Gaussian prior to avoid overfitting and a conditional distribution of gene latent feature vectors given the latent feature vectors of their direct neighbors:

$$\begin{aligned} p(\boldsymbol{F}|\mathscr{P}, \boldsymbol{W}, \sigma_F^2, \sigma_{\mathscr{P}}^2) \quad &\propto \quad \prod_{i=1}^{n} \mathcal{N}(\boldsymbol{F}_i|0, \sigma_F^2 \boldsymbol{I}) \times \\ &\quad \prod_{i=1}^{n} \mathcal{N}(\boldsymbol{F}_i| \sum_{P\in\mathscr{P}} \boldsymbol{W}_{ip} \sum_{j\in N_i^P} \boldsymbol{P}_{ij} \boldsymbol{F}_j, \sigma_{\mathscr{P}}^2 \boldsymbol{I}). \qquad (3.4) \end{aligned}$$

Such formulation of gene latent matrix keeps gene feature vectors in $\boldsymbol{F}$ both small and close to the latent feature vectors of their direct neighbors. Because NG-MC borrows its strength across all available observations and gene neighborhoods in estimating each $\boldsymbol{G}_{ij}$, it can lead to more accurate inference than simply learning $\boldsymbol{G}_{ij}$ independently of any additional domain knowledge. In a Bayesian estimation setting of our NG-MC model, one is interested in the behavior of the posterior distribution of gene latent feature matrices $\boldsymbol{F}$ and $\boldsymbol{H}$ given the observed genetic interaction scores $\boldsymbol{G}$ and gene networks $\mathscr{P}$. It follows that the posterior $p(\boldsymbol{F}, \boldsymbol{H}|\boldsymbol{G}, \mathscr{P}, \boldsymbol{W}, \sigma_G^2, \sigma_{\mathscr{P}}^2, \sigma_F^2, \sigma_H^2)$ is proportional to the following expression:

$$\prod_{i=1}^{n}\prod_{j=1}^{n}\mathcal{N}(\boldsymbol{G}_{ij}|g(\boldsymbol{F}_i^T\boldsymbol{H}_j),\sigma_G^2)^{\boldsymbol{I}_{ij}^G}\times\prod_{i=1}^{n}\mathcal{N}(\boldsymbol{F}_i|\sum_{P\in\mathscr{P}}\boldsymbol{W}_{ip}\sum_{j\in N_i^P}\boldsymbol{P}_{ij}\boldsymbol{F}_j,\sigma_{\mathscr{P}}^2\boldsymbol{I})\times \\ \times\prod_{i=1}^{n}\mathcal{N}(\boldsymbol{F}_i|0,\sigma_F^2\boldsymbol{I})\times\prod_{j=1}^{n}\mathcal{N}(\boldsymbol{H}_j|0,\sigma_H^2\boldsymbol{I}). \quad (3.5)$$

We then compute the log-posterior probability $\ln p(\boldsymbol{F},\boldsymbol{H}|\boldsymbol{G},\mathscr{P},\boldsymbol{W},\sigma_G^2,\sigma_{\mathscr{P}}^2,\sigma_F^2,\sigma_H^2)$ to obtain the expression:

$$-\frac{1}{2\sigma_G^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{I}_{ij}^G(\boldsymbol{G}_{ij}-g(\boldsymbol{F}_i^T\boldsymbol{H}_j))^2-\frac{1}{2\sigma_F^2}\sum_{i=1}^{n}\boldsymbol{F}_i^T\boldsymbol{F}_i-\frac{1}{2\sigma_H^2}\sum_{j=1}^{n}\boldsymbol{H}_j^T\boldsymbol{H}_j- \\ -\frac{1}{2\sigma_P^2}\sum_{i=1}^{n}((\boldsymbol{F}_i-\sum_{P\in\mathscr{P}}\boldsymbol{W}_{ip}\sum_{j\in N_i^P}\boldsymbol{P}_{ij}\boldsymbol{F}_j)^T(\boldsymbol{F}_i-\sum_{P\in\mathscr{P}}\boldsymbol{W}_{ip}\sum_{j\in N_i^P}\boldsymbol{P}_{ij}\boldsymbol{F}_j))- \\ -\frac{1}{2}nk(\ln\sigma_F^2+\ln\sigma_H^2+\ln\sigma_{\mathscr{P}}^2)-\frac{1}{2}(\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{I}_{ij}^G)\ln\sigma_G^2+\mathscr{C}. \quad (3.6)$$

Our goal is to learn $\boldsymbol{F}$, $\boldsymbol{H}$ and $\boldsymbol{W}$ that maximize the conditional posterior probability over gene latent feature vectors. To do so, we formulate a minimization problem that is equivalent to maximization of the log-posterior probability in Eq. (3.6) and employ gradient descent technique on $\boldsymbol{F}$, $\boldsymbol{H}$ and $\boldsymbol{W}$ to solve it. In particular, we minimize the objective function:

$$\begin{aligned}\mathscr{L}(\boldsymbol{G},\mathscr{P},\boldsymbol{W},\boldsymbol{F},\boldsymbol{H}) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} I_{ij}^{G}(\boldsymbol{G}_{ij} - g(\boldsymbol{F}_i^T\boldsymbol{H}_j))^2 + \\ + \frac{\lambda_{\mathscr{P}}}{2}\sum_{i=1}^{n}((\boldsymbol{F}_i - \sum_{P\in\mathscr{P}} \boldsymbol{W}_{ip}\sum_{j\in N_i^P}\boldsymbol{P}_{ij}\boldsymbol{F}_j)^T(\boldsymbol{F}_i - \sum_{P\in\mathscr{P}} \boldsymbol{W}_{ip}\sum_{j\in N_i^P}\boldsymbol{P}_{ij}\boldsymbol{F}_j)) \\ + \frac{\lambda_F}{2}\sum_{i=1}^{N}\boldsymbol{F}_i^T\boldsymbol{F}_i + \frac{\lambda_H}{2}\sum_{j=1}^{N}\boldsymbol{H}_j^T\boldsymbol{H}_j, \quad (3.7)\end{aligned}$$

where $\lambda_F = \sigma_G^2/\sigma_F^2$, $\lambda_H = \sigma_G^2/\sigma_H^2$ and $\lambda_{\mathscr{P}} = \sigma_G^2/\sigma_{\mathscr{P}}^2$. We normalize interaction measurements in $\boldsymbol{G}$ before performing numerical optimization such that the elements of $\boldsymbol{G}$ are in [0,1] interval. Normalization is due to estimates in $\widehat{\boldsymbol{G}} = g(\boldsymbol{F}^T\boldsymbol{H})$ being bounded by the logistic function g. We keep the observation noise variance σ_G^2 and prior variances σ_F^2, σ_H^2 and $\sigma_{\mathscr{P}}^2$ fixed and use gradient descent algorithm to find the local minimum of $\mathscr{L}(\boldsymbol{G},\mathscr{P},\boldsymbol{W},\boldsymbol{F},\boldsymbol{H})$ and estimate gene latent feature matrices. The parameters λ_F and λ_H serve as to regularize latent gene profiles and the presence of $\lambda_{\mathscr{P}}$ trades off the sole reliance on observed measurements against the inclusion of domain knowledge.

NG-MC algorithm (Algorithm 1) iteratively updates gene latent feature vectors $\boldsymbol{F}_i$ and $\boldsymbol{H}_j$ for each i and j based on the latent feature vectors from the previous iteration and gene neighbors in networks $\mathscr{P}$. In each iteration, NG-MC also refines weights of genes in considered gene networks given in $\boldsymbol{W}$ in order to account for the contribution of genes to current latent feature vectors of their neighbors. Successive updates of $\boldsymbol{F}_i$ and $\boldsymbol{H}_j$ converge to a maximum *a posteriori* estimate of the posterior probability formulated in Eq. (3.5). In practice, the algorithm stops iterating once the reconstruction error over observed interaction measurements does not decrease after the update of $\boldsymbol{F}$, $\boldsymbol{H}$ and $\boldsymbol{W}$. We observed that parameter values $\lambda_H = \lambda_F = 0.01$ and learning rates $\alpha = 0.1$ and $\alpha_{\mathscr{P}} = 0.001$ gave accurate results across a number of different data sets. Parameter $\lambda_{\mathscr{P}}$, which controls the influence of gene networks $\mathscr{P}$ on gene latent feature vectors in $\boldsymbol{F}$, depended on data set complexity (Brock et al., 2008).

Algorithm 1

NG-MC, the proposed approach for matrix completion prior knowledge presented in the form of networks. Source code is available at http://github.com/marinkaz/ngmc.

```
Input:
    ■ Sparse matrix $\boldsymbol{G} \in \mathbb{R}^{n \times n}$ containing interaction measurements,
    ■ gene networks $\mathscr{P} = \{\boldsymbol{P} \in \mathbb{R}^{n \times n}\}$,
    ■ parameters $\lambda_F = \lambda_H$, $\lambda_{\mathscr{P}}$,
    ■ rank $k$,
    ■ learning rates $\alpha$ and $\alpha_{\mathscr{P}}$.
Output:
    ■ Data matrix $\widehat{\boldsymbol{G}}$,
    ■ latent matrices $\boldsymbol{F}$ and $\boldsymbol{H}$
    ■ gene networks weights $\boldsymbol{W}$.

1. Normalize each row of $\boldsymbol{P} \in \mathscr{P}$ such that $\sum_{j=1}^{n} \boldsymbol{P}_{ij} = 1$.
2. Sample $\boldsymbol{F} \sim \mathscr{U}[0,1]^{k \times n}$ and $\boldsymbol{H} \sim \mathscr{U}[0,1]^{k \times n}$ and set $\boldsymbol{W} = [\frac{1}{|\mathscr{P}|}]^{n \times |\mathscr{P}|}$.
3. Repeat until convergence:
    a. For $i, j = 1, 2, \dots, n$:
        $\begin{aligned} \frac{\partial \mathscr{L}}{\partial \boldsymbol{F}_i} &= \sum_{j=1}^{n} I_{ij}^{G} \boldsymbol{H}_j g'(\boldsymbol{F}_i^T \boldsymbol{H}_j)(g(\boldsymbol{F}_i^T \boldsymbol{H}_j) - \boldsymbol{G}_{ij}) + \lambda_F \boldsymbol{F}_i + \\ &+ \lambda_{\mathscr{P}}(\boldsymbol{F}_i - \sum_{P \in \mathscr{P}} \boldsymbol{W}_{ip} \sum_{j \in N_i^P} \boldsymbol{P}_{ij} \boldsymbol{F}_j) - \\ &- \lambda_{\mathscr{P}} \sum_{P \in \mathscr{P}} \sum_{\{j | i \in N_j^P\}} \boldsymbol{W}_{jp} \boldsymbol{P}_{ji} (\boldsymbol{F}_j - \sum_{R \in \mathscr{P}} \boldsymbol{W}_{jr} \sum_{l \in N_j^R} \boldsymbol{R}_{jl} \boldsymbol{F}_l), \end{aligned}$
        $\frac{\partial \mathscr{L}}{\partial \boldsymbol{H}_j} = \sum_{i=1}^{n} I_{ij}^{G} \boldsymbol{F}_i g'(\boldsymbol{F}_i^T \boldsymbol{H}_j)(g(\boldsymbol{F}_i^T \boldsymbol{H}_j) - \boldsymbol{G}_{ij}) + \lambda_H \boldsymbol{H}_j.$
    b. For $i = 1, 2, \dots, n$ and $p = 1, 2, \dots, |\mathscr{P}|$:
        $\begin{aligned} \frac{\partial \mathscr{L}}{\partial \boldsymbol{W}_{ip}} &= -\lambda_{\mathscr{P}} \boldsymbol{F}_i^T \sum_{j \in N_i^P} \boldsymbol{P}_{ij} \boldsymbol{F}_j + \lambda_{\mathscr{P}} \boldsymbol{W}_{ip} \sum_{j \in N_i^P} \boldsymbol{P}_{ij} \boldsymbol{F}_j^T \sum_{k \in N_i^P} \boldsymbol{P}_{ik} \boldsymbol{F}_k + \\ &+ \frac{\lambda_{\mathscr{P}}}{2} \sum_{j \in N_i^P} \boldsymbol{P}_{ij} \boldsymbol{F}_j^T \sum_{\substack{\bar{P} \in \mathscr{P} \\ \bar{p} \neq p}} \boldsymbol{W}_{i\bar{p}} \sum_{j \in N_i^{\bar{P}}} \bar{\boldsymbol{P}}_{ij} \boldsymbol{F}_j. \end{aligned}$
    c. Set $\boldsymbol{F}_i \leftarrow \boldsymbol{F}_i - \alpha \frac{\partial \mathscr{L}}{\partial F_i}$ for $i = 1, 2, \dots, n$.
    d. Set $\boldsymbol{H}_j \leftarrow \boldsymbol{H}_j - \alpha \frac{\partial \mathscr{L}}{\partial H_j}$ for $j = 1, 2, \dots, n$.
    e. Set $\boldsymbol{W}_{ip} \leftarrow \boldsymbol{W}_{ip} - \alpha_{\mathscr{P}} \frac{\partial \mathscr{L}}{\partial W_{ip}}$ for $i = 1, 2, \dots, n$ and $p = 1, 2, \dots |\mathscr{P}|$.
4. Compute $\widehat{\boldsymbol{G}} = g(\boldsymbol{F}^T \boldsymbol{H})$. Predict interaction of $g_i$ and $g_j$ as $(\widehat{\boldsymbol{G}}_{ij} + \widehat{\boldsymbol{G}}_{ji})/2$.
```

3.4 A case study: imputation of genetic interactions

Next, we evaluate the performance of network-guided matrix completion against several alternative approaches for prediction of genetic interactions in yeast *S. cerevisiae*. We also study how the amount and distribution of missing values affect predictive performance and whether performance can be improved through inclusion of side information.

3.4.1 Experimental setup

In the experiments we consider an existing incomplete E-MAP matrix from each of the E-MAP studies and artificially introduce an additional 1% of missing values for a set of randomly selected gene pairs representing unmeasured interactions (Ryan et al., 2010; Pan et al., 2011). These gene pairs and their data constitute a test set on which we evaluate performance of imputation algorithms. Because of E-MAP symmetry, for a given test gene pair and its corresponding entry $\boldsymbol{G}_{ij}$, we also hide the value of $\boldsymbol{G}_{ji}$. We repeat this process 30 times and report on the averaged imputation performance.

It may be noted that established performance evaluation procedure of missing value imputation methods for gene expression data is not directly applicable to E-MAPs for several reasons discussed in Ryan et al. (2010). That procedure first constructs a complete data matrix by removing genes with missing values and then artificially introduces missing values for evaluation. Gene expression data contain substantially lower fraction of missing data than E-MAPs (Table 3.1) and removing a small number of genes and experimental conditions does not significantly reduce the size of gene expression data sets, whereas this does not hold for E-MAP data sets.

We select the latent dimensionality k and regularization parameters λ_F and $\lambda_{\mathscr{P}}$ of the NG-MC with the following procedure. For each data set and before the performance evaluation, we leave out 1% of randomly selected known values and attempt to impute them with varying values of parameters in grid search fashion. Parameter values that result in the best estimation of the left-out values are then used in all experiments involving the data set. Notice that the left-out values are determined before performance evaluation and are therefore not included in the test data set. We set the parameters of competitive methods to values recommended by Ryan et al. (2010) (for wNN, LLS and BPCA) or optimize parameter selection through grid search (for SVT, MC and NG-MC).

We consider two measures of imputation accuracy. These are the Pearson correlation (CC) between the imputed and the true values, and the normalized root mean square error (NRMSE) (Oba et al., 2003) given as NRMSE $= \sqrt{E((\hat{\boldsymbol{y}} - \boldsymbol{y})^2)/Var(\boldsymbol{y})}$, where $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ denote vectors of true and imputed values, respectively. More accurate imputations give a higher correlation score and a lower NRMSE.

To test if the differences in performance of imputation methods are significant, we use the Wilcoxon signed-rank test, a non-parametric equivalent of a paired t-test. Its advantage is that it does not require Gaussian distribution or homogeneity of variance, but it has less statistical power, so there is a risk that some differences are not recognized as significant.

3.4.2 Data

We consider four E-MAP data sets in a comparative evaluation of NG-MC with five state-of-the-art methods for missing value imputation. The evaluated data sets are from budding yeast *S. cerevisiae*; they include S-scores of interaction measurements, but differ in the subset of studied genes and the proportion of missing values (Table 3.1):

- Chromosome Biology (Collins et al., 2007) is the largest data set considered, encompassing interactions between 743 genes involved in various aspects of chromosome biology, such as chromatid segregation, DNA replication and transcriptional regulation.
- RNA processing (denoted by RNA) (Wilmes et al., 2008) focused on the relationships between and within RNA processing pathways involving 552 mutants, 166 of which were hypomorphic alleles of essential genes.
- The Early Secretory Pathway (denoted by ESP) (Schuldiner et al., 2005) generated genetic interaction maps of genes acting in the yeast early secretory pathway to identify pathway organization and components of physical complexes.
- Lipid E-MAP (Surma et al., 2013) focused on lipid metabolism, sorting, trafficking and various aspects of lipid biology, and its data were indicative of a

dedicated bilayer stress response for membrane homeostasis.

Table 3.1

Overview of the E-MAPs considered.

Data set	Genes	Missing Interactions	Measured Interactions
Chromosome Biology	743	34.0%	187,000
Lipid	741	9.2%	250,000
RNA	552	29.6%	107,000
Early Secretory Pathway	424	7.5%	83,000

We have considered two data sources for gene network construction. The first network is constructed based on Gene Ontology (Ashburner et al., 2000) (GO) annotation data. It is a weighted network of genes included in the E-MAP study whose edge weights correspond to the number of shared GO terms between connected genes, excluding annotations inferred from GI studies (*i.e.* those with the IGI evidence code). The second network represents physical interaction data from BioGRID 3.2 (Stark et al., 2006). The physical interaction network is a binary network in which two genes are connected if their gene products physically interact. Depending on the considered network, we denote their corresponding NG-MC models by NG-MC-GO and NG-MC-PPI, respectively.

3.4.3 *Imputation performance*

Table 3.2 shows the CC and NRMSE scores of imputation algorithms along with the baseline method of filling-in with zeros. NG-MC-PPI and NG-MC-GO achieved highest accuracies on all considered data sets. We compared their scores with the performance of the second-best method (*i.e.* LLS on Chromosome Biology data set, SVT on ESP data set and MC on RNA data set) and found that improvements were significant in all data sets.

We did not observe any apparent connection between the proportion of missing values in a data set and the performance of any of the imputation methods. The performance was better on smaller ESP and RNA data sets, although differences were small and further investigation appears to be worthwhile.

The baseline method of filling-in with zeros had the worst performance on all data sets. While this approach seems naïve, it is justified by the expectation that most genes do not interact. We observed that BPCA failed to match the performance of weighted neighbor-based and local least squares methods, wNN and LLS, respectively, on all three evaluated E-MAP data sets. Local imputation methods, wNN and LLS, demonstrated good performance across all three data sets. Solid performance of neighbor-based methods on larger data sets could be explained by a larger number of neighbors to choose from when imputing missing values, which resulted in more reliable missing value estimates.

Global methods, BPCA, SVT and MC, performed well on the ESP data set but poorly on a much larger Chromosome Biology data set. These methods assume the existence of a global covariance structure between all genes in the E-MAP score matrix. When this assumption is not appropriate, *i.e.* when genes predominantly exhibit local similarity substructure, the imputation becomes less accurate. The comparable performance of SVT and MC across data sets was expected. Both methods solve related optimization problems and operate under the assumption that the E-MAP score matrix has low rank.

The superior performance of NG-MC models over other imputation methods can be explained by their ability to introduce circumstantial evidence into model inference. As a hybrid imputation approach, NG-MC can benefit from both global information present in the E-MAP data and local similarity structure between genes. One could vary the level of influence of global and local patterns on the imputation through $\lambda_{\mathscr{P}}$ parameter of the NG-MC model, where a higher value of $\lambda_{\mathscr{P}}$ indicates more emphasis on locality. In this way, our approach can adequately address data of varying underlying complexity (Brock et al., 2008), where data complexity indicates the difficulty of mapping the E-MAP score matrix to a low-dimensional space. To quantify the complexity of gene expression matrices, Brock et al. (2008) devised an entropy-based imputation algorithm selection scheme that was based on observation that global imputation methods performed better on gene expression data with lower data complexity and local methods performed better on data with higher complexity. Their selection

Table 3.2

Accuracy as measured by the Pearson correlation coefficient (CC) and normalized root mean squared error (NRMSE) across three E-MAP data sets and eight imputation methods. MC denotes matrix completion model (Sec. 3.3.2). The NG-MC-GO and NG-MC-PPI are network-guided matrix completion models (Sec. 3.3.3) that utilize Gene Ontology annotation and physical interaction data, respectively. For descriptions of other methods see Sec. 3.2. Highlighted results are significantly better than the best non-NG-MC method according to the Wilcoxon signed-rank test at 0.05 significance level.

Approach	Chromosome Biology		ESP		RNA	
	CC	NRMSE	CC	NRMSE	CC	NRMSE
Filling with zeros	0.000	1.021	0.000	1.011	0.000	1.000
BPCA ($k = 300$)	0.539	0.834	0.619	0.796	0.589	0.804
wNN ($k = 50$)	0.657	0.744	0.625	0.776	0.626	0.787
LLS ($k = 20$)	0.678	0.736	0.626	0.764	0.626	0.776
SVT ($k = 40$)	0.631	0.753	0.672	0.719	0.649	0.765
MC ($k = 40$)	0.641	0.742	0.653	0.722	0.651	0.760
NG-MC-GO ($k = 60$)	0.691	0.693	*0.732*	*0.648*	*0.727*	*0.641*
NG-MC-PPI ($k = 60$)	*0.722*	*0.668*	*0.742*	0.667	0.701	*0.652*

scheme could be adapted to work with E-MAP-like data sets and be used to set $\lambda_{\mathscr{P}}$ in an informed way.

We studied the sensitivity of NG-MC to variations in algorithm parameters. In particular, we investigated how NG-MC imputation performance was affected as a function of parameters values. The parameters of NG-MC algorithm are the latent dimensionality of the factorized model (k), the degree of regularization of latent matrices (λ_F) and the impact of network neighborhood ($\lambda_{\mathscr{P}}$). In additional experiments performed on ESP data set (Fig. 3.2) and with NG-MC-GO model we found that performance of our NG-MC approach is robust for a broad range of parameters values.

3.4.4 Missing value abundance and distribution

Ulitsky et al. (2009) described three different scenarios of missing values in E-MAP experiments (Fig. 3.3). The simplest and the most studied scenario is the *Random* model for which we assume that missing measurements are generated independently and uniformly by a random process. The *Submatrix* model corresponds to the case

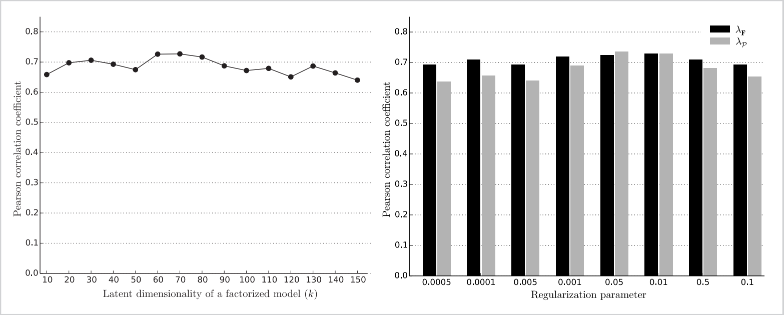


Figure 3.2

Sensitivity of network-guided matrix completion to selection of latent dimensionality (left) and regularization (right). When studying the latent dimensionality we set the regularization to $\lambda_F = 0.01$ and $\lambda_{\mathcal{P}} = 0.01$, and when investigating the influence of regularization we set the latent dimensionality to $k = 60$ and the remaining regularization parameter to 0.01. Results are for the early secretory pathway and the network derived from Gene Ontology. Similar behavior was observed with other E-MAP data sets.

where all interactions within a subset of genes (*e.g.* essential genes) are missing. The *Cross* model arises when interactions between two disjoint subsets of genes are missing. This model concurs with the situation when two E-MAP data sets that share a subset of genes are combined into a single large data set. We identified the fourth missing value configuration, which we call the *Prediction* scenario (Fig. 3.3). It occurs when complete GI profiles are missing. Learning in such setting is substantially harder than learning with other missing value arrangements as genes with missing values in the Prediction scenario do not have any associated interaction measurements. In the previous section, we compared the imputation methods using the Random configuration and we study other configurations in this section. We were here interested in the effect that various missing data configurations have on NG-MC and we compared the NG-MC algorithm to its variant, which does not use domain knowledge (MC).

Fig. 3.4 reports on the predictive performance of our matrix completion approach obtained by varying the fraction of missing values in the four missing data scenarios presented in Fig. 3.3. For $x = 5, 10, 20, \ldots, 90$ we hid x% of E-MAP measurements in the ESP data and inferred prediction model. Our results are reasonably accurate (CC > 0.4) when up to 60% of the E-MAP values were hidden in the Random and Submatrix models. It should be noted that when we hide 60% of the ESP E-MAP measurements, the E-MAP scores are present in less than 40% of the matrix because the original ESP data set already contains ~8% missing values (Table 3.1). When more than 80%

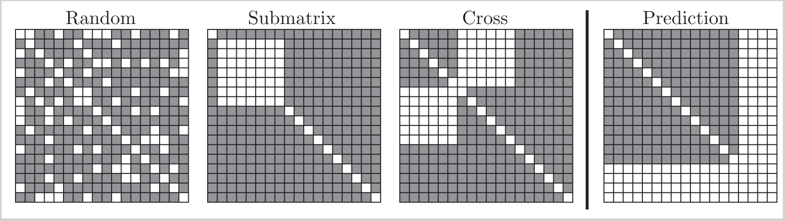


Figure 3.3

The four patterns of missing values. *Random* configuration has hidden genetic interactions selected uniformly at random. *Submatrix* and *Cross* configurations have hidden all interactions within a random set of genes or between two random disjoint sets of genes, respectively. In the *Prediction* scenario, complete genetic interaction profiles of genes are removed.

of the data were removed, the three considered prediction models still achieved higher accuracy (CC ≈ 0.2) than filling-in with zeros. As expected, predictions were more accurate for the Random model than for the Submatrix model for almost all fractions of hidden data (cf. Fig. 3.4). However, the difference in performance between the Random and the Submatrix models tended to be small when less than 30% or more than 70% of the measurements were hidden. From this experiment we conclude that inclusion of additional genomic data is more useful in structured missing value scenarios, *i.e.* the Submatrix and the Cross model (Fig. 3.4), demonstrating that individual gene networks provide complementary information.

Imputation accuracy has improved (Fig. 3.4) when E-MAP data were combined with gene annotation (NG-MC-GO) or protein-protein interaction (NG-MC-PPI) networks. These results support findings from experimental studies (Tong et al., 2004; Collins et al., 2007; Costanzo et al., 2010) that showed that if two proteins act together to carry out a common function, deletions of their corresponding encoding genes may have similar GI profiles. Furthermore, Gene Ontology annotations and synthetic lethality are correlated with ∼12% and ∼27% of genes that genetically interact having either identical or highly similar Gene Ontology annotations, respectively (Tong et al., 2004; Michaut and Bader, 2012). Our NG-MC-GO and NG-MC-PPI models could exploit these strong links between functionally similar genes, physically interacting proteins and GIs. Performance of integrated models in Fig. 3.4 suggests the importance of combining interaction and functional networks for prediction of missing values in E-MAP data sets.

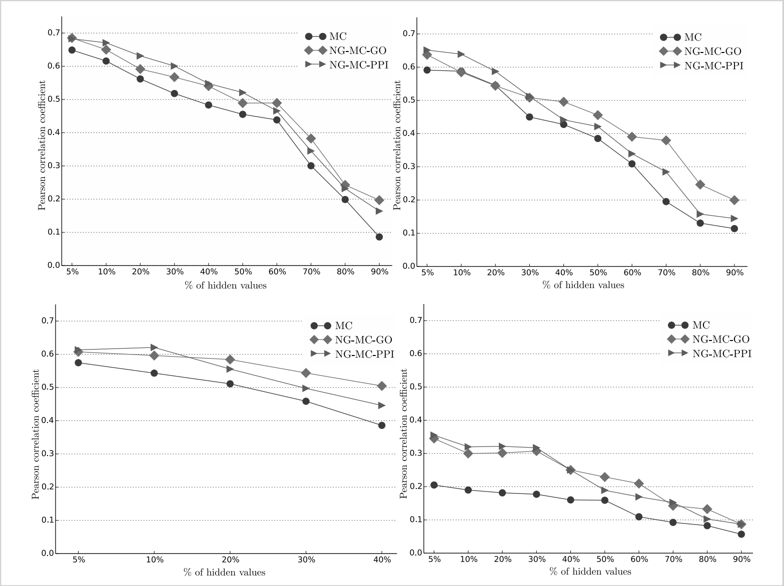


Figure 3.4

Performance of imputation methods (Pearson correlation coefficient) proposed in this chapter for different missing data rates and missing value configurations (first row: *Random* and *Submatrix* scenarios, second row: *Cross* and *Prediction* scenarios). Refer to the main text and Fig. 3.3 for description of the missing value scenarios. MC denotes matrix completion approach (Sec. 3.3.2). Network-guided matrix completion (Sec. 3.3.3) is represented by NG-MC-GO and NG-MC-PPI. Performance was assessed for the early secretory pathway E-MAP data set because it contains the least missing values. The Cross configuration is not applicable when more than 50% of the values are missing.

We observed deterioration of imputation accuracy when complete genetic interaction profiles were removed and NG-MC could only utilize circumstantial evidence (Fig. 3.4, second row, right). Decreased prediction performance suggests that measured gene interactions are the best source of information for predicting missing values in the E-MAP data. However, when the proportion of missing interactions was increased, the inclusion of additional genomic data was more helpful. With the exception of the Prediction model for which the opposite behavior was observed, the performance difference between MC and NG-MC was small (∼10%) as long as <50% of the data were removed, but rose to above 20% when ≥60% of the data were removed (Fig. 3.4).

3.4.5 Data imputation by integration of gene networks

We studied imputation performance of our proposed approach on the recent lipid E-MAP data set by Surma et al. (2013). Fig. 3.5 shows the Pearson correlation between the imputed and true interaction measurements when different types of circumstantial evidence were considered and various amounts and distributions of genetic interactions were excluded from the training set. Similarly as in experiments with the ESP data set (Fig. 3.4), prediction models inferred from the lipid E-MAP data that included prior knowledge performed better than models, which considered only interaction measurements. Fig. 3.5 also reveals that best performance was attained when our NG-MC approach collectively considered both protein-protein interaction network and network derived from gene functional annotation data (NG-MC-GO-PPI). The NG-MC can simultaneously consult multiple gene networks during model inference and modify gene weights in each of the networks to achieve better prediction accuracy. As such, it does not require substantial network preprocessing prior model inference and is able to adjust for network influence by taking into account entire collection of considered networks. Fig. 3.5 also conveys that the inclusion of additional knowledge into prediction models is more pronounced in scenarios with high missing data rates and non-trivial structure of missing measurements. Good performance of our approach in such scenarios is an appealing property and hence, NG-MC seems to be an attractive data imputation approach.

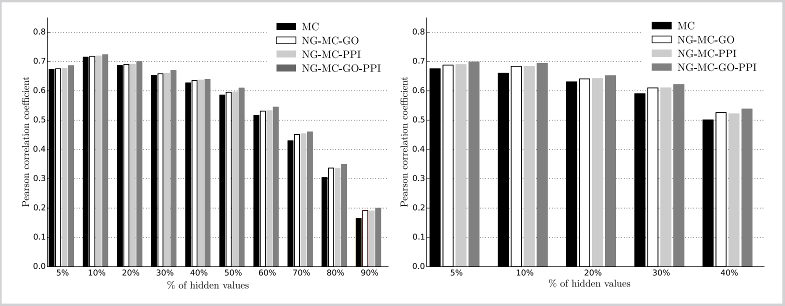


Figure 3.5

Imputation performance of network-guided matrix completion (NG-MC) for different fractions of missing values in the lipid E-MAP data set and for various sources of biological network information. Shown are results for the *Random* (left) and *Cross* (right) scenarios. Prior knowledge is included in the form of protein-protein interaction network (PPI), a network derived from Gene Ontology annotation data (GO) and collective consideration of both PPI and GO.

3.5 Conclusion

We have proposed a new missing value imputation method called network-guided matrix completion (NG-MC) that targets gene interaction data sets. The approach is unique in combining gene interaction and network data through inference of a single probabilistic model. Experiments with epistatic MAP interaction data sets show that inclusion of prior knowledge is crucial and helps NG-MC to perform better than a number of state-of-the-art algorithms we have included in our study. The results are encouraging and have potentially high practical value for prediction of genetic interactions that are otherwise unavailable to existing interaction measurements.

Part II

Network inference

4

Epistasis-based network inference

Epistasis analysis is an essential tool of classical genetics for inferring the order of function of genes in a common pathway. Typically, it considers single and double mutant phenotypes and for a pair of genes observes if a change in the first gene masks the effects of the mutation in the second gene. Despite the recent emergence of biotechnology techniques that can provide gene interaction data on a large, possibly genomic scale, very few methods are available for quantitative epistasis analysis and epistasis-based network reconstruction.

In this chapter we describe a conceptually new probabilistic approach to gene network inference from quantitative interaction data. The approach is founded on epistasis analysis. Its features are joint treatment of the mutant phenotype data with a factorized model and probabilistic scoring of pairwise gene relationships that are inferred from the latent gene representation. The resulting gene network is assembled from scored pairwise relationships. In an experimental study, we show that the proposed approach can accurately reconstruct several known pathways and that it surpasses the accuracy of current approaches.

4.1 Background

Epistasis analysis is a tool of classical genetics for inferring the order of genes in pathways from mutant-based phenotypes (Botstein and Maurer, 1982; Avery and Wasserman, 1992). Epistasis asserts that two genes interact if the mutation in one gene masks the effects of perturbations in the other gene. Then, assuming a common pathway, the first, masking gene would be downstream, and the products of the second gene would regulate the expression of the first one (Avery and Wasserman, 1992; Huang and Sternberg, 1995; Roth et al., 2009; Cordell, 2002). Epistasis analysis uncovers the relationship between a pair of genes. Its logic can be further extended to uncover parallelism, where both genes have an effect on the phenotype but where there is no epistasis (Zupan et al., 2003; Battle et al., 2010). Uncovered pairwise relationships in a group of genes can give rise to a reconstruction of more complex multi-gene networks. An enlightening demonstration of the power of epistasis for assembly of gene networks is for instance a reconstruction of a four-gene cell death pathway in *C. elegans* (Metzstein et al., 1998).

Fig. 4.1 shows a toy example of epistasis analysis with three genes, u, v and w. The phenotype a double or single knockout mutants are denoted with $\boldsymbol{R}$. For example, $\boldsymbol{R}(u\Delta v\Delta)$ and $\boldsymbol{R}(v\Delta))$ correspond to the quantified phenotypes of a double knockout mutant $u\Delta v\Delta$ and single knockout mutant $v\Delta$, respectively. Expected double mutant phenotypes, which assume no interaction between genes, are denoted with $\boldsymbol{E}$ (*e.g.* $\boldsymbol{E}(u\Delta v\Delta)$). Three types of pairwise gene relationships are typically considered in epistasis analysis:

Fig. 4.1a: A double mutant $u\Delta v\Delta$ has a phenotype similar to that of a single mutant $v\Delta$, which indicates that v is epistatic to u.

Fig. 4.1b: From the activity of genes v and w we conjecture that gene v partially depends on gene w, *i.e.,* v also acts through a separate pathway because their double mutant $v\Delta w\Delta$ has a phenotype that is equally similar to the single knockout $\boldsymbol{R}(w\Delta)$ and the expected phenotype $\boldsymbol{E}(v\Delta w\Delta)$.

Fig. 4.1c: The phenotype of double knockout $u\Delta w\Delta$ is close to the expected phenotype of $u\Delta w\Delta$, $\boldsymbol{E}(u\Delta w\Delta)$, which may be explained by u and w acting independently in parallel pathways.

Given gene-gene relationships that are concordant with the phenotypic measurements, the goal of epistasis-based gene network inference is to estimate a joint network, which is consistent with observations and scored gene-gene relationships. The multi-gene network in Fig. 4.1d represents such a candidate pathway on genes u, v and w.

Emergent technologies from molecular biology that record phenotypes of single and double mutants at a large, possibly genomic scale, prompt for the development of systematic approaches for epistasis analysis and pose the need to devise computational tools that support gene network inference. Approaches of mutagenesis by homologous recombination (Tong et al., 2004; Collins et al., 2006) or RNA interference can yield phenotype observations for thousands or even millions of mutants (Costanzo et al., 2010). Several past studies considered mutant assays with qualitative phenotypes (Zupan et al., 2003), quantitative fitness scores (Drees et al., 2005; St Onge et al., 2007; Beerenwinkel et al., 2007; Battle et al., 2010; Phenix et al., 2011, 2013) or even whole-genome transcriptional profiles (Van Driessche et al., 2005; Hughes, 2005). Majority of these studies present gene networks as collections of directly observed pairwise interactions (*e.g.*, St Onge et al. (2007); Phenix et al. (2013)) and do not propose a generally

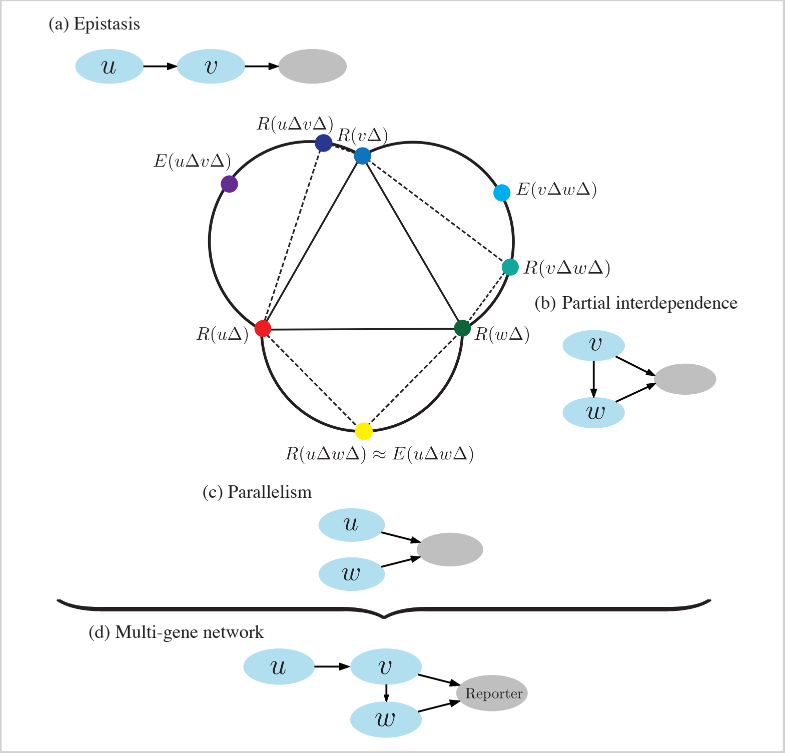


Figure 4.1

A hypothetical example of epistasis analysis with three genes, u, v and w. Nodes in the central graph represent mutant phenotypes. The phenotypic difference between a double knockout (e.g. $R(u\Delta v\Delta)$) and a single knockout mutant (e.g. $R(v\Delta)$) is given by the length of the corresponding dotted edge. Expected double mutant phenotypes, which assume no interaction between genes, are denoted with E (e.g. $E(u\Delta v\Delta)$). See Sec. 4.1 for further explanation.

applicable formalism to model the data. Only few general purpose algorithms for inference of epistatic networks have been proposed. Zupan et al. (2003) introduced formal rules and inference algorithm to infer different types of relationships between genes, but could treat only qualitative phenotypes and could not handle noise. These limitations were elegantly bypassed by a Bayesian approach of Battle et al. (2010) that can handle larger data sets with few hundred genes. This algorithm is to our knowledge also the only modern approach to inference of epistasis networks.

Gene epistasis analysis infers interactions that stem directly from mutant phenotypes. Its causative reasoning is different from other network reconstruction tools that observe correlations between gene profiles (*e.g.* Ahn et al. (2011); Mohammadi et al. (2012))

and infer relationships that are circumstantial (Hughes et al., 2000). Despite the growing body of quantitative genetic interaction data and our ability to collect such data computational approaches and tools to support epistasis are at best scarce (Battle et al., 2010; Jaimovich and Friedman, 2011; Zhang and Zhao, 2013). Devising methods for inference of gene pathways from mutant-based phenotypes and developing related software tools remains a major challenge of computational systems biology.

We here present a new epistasis analysis-inspired computational approach to infer gene networks from a collection of quantitative mutant phenotypes. We refer to our method as Réd (pronounced as *réd*, meaning "order" in Slovene). Our work was motivated by the Bayesian learning method of Battle et al. (2010), henceforth denoted by APN (activity pathway network), that starts from a random network and then iteratively refines it to best match data-inferred relationships. The model refinement in APN is carried out through a succession of local structural changes of the evolving network. This procedure may substantially depend on (arbitrary) initialization of network structure, and hence requires ensembling across many runs of the algorithm to raise accuracy of the final network.

Our approach is conceptually different from APN. We first simultaneously infer a probabilistic model for the entire set of pairwise relationships. Relationship probabilities serve as preferences for different types of pairwise relationships (*e.g.* epistasis, parallelism and partial interdependence) used in a single-step construction of a gene network. In contrast to APN's local network changes, Réd applies a global procedure to infer the relationships between genes and does not require ensembling. The probabilistic model of Réd uses matrix completion-derived latent data representation to account for noise and sparsity. Inference of factorized model also includes construction of a gene-specific data transformation to account for the differences in single mutant backgrounds, which may affect the phenotype of double mutants. In an experimental study, we show that both components are necessary for inferring gene networks of high accuracy.

4.2 Probabilistic view of epistatic relationships

Réd, the proposed gene network reconstruction algorithm (Algorithm 2), considers quantitative phenotype measurements over a set of single and double mutants, pro-

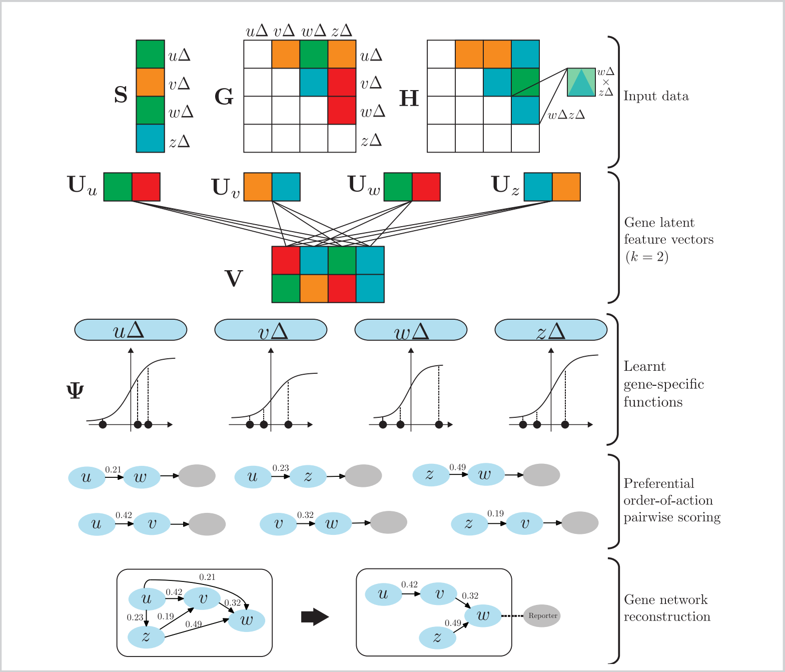


Figure 4.2

An overview of Réd, a novel approach for automatic gene network inference from mutant data. Inputs to the preferential order-of-action factorized algorithm of Réd include a matrix of double knockout phenotypes ($\boldsymbol{G}$), a vector of single knockout phenotypes ($\boldsymbol{S}$) and a matrix of expected phenotypes corresponding to the assumption of absent interactions between genes ($\boldsymbol{H}$). Réd estimates a factorized model from $\boldsymbol{G}$, whose gene latent feature vectors capture the global structure of the phenotype landscape, and learns a parametrized logistic map Ψ, which is a gene-dependent nonlinear mapping from latent to phenotype space. A scoring scheme is then applied to the inferred model to estimate the probabilities of pairwise gene relationships of different types. Finally, a multi-gene network is reconstructed, which aims to minimize the number of violating and redundant edges.

vides preferential order-of-action scores of possible pairwise relationships, and assembles them in a joint gene network. The essential steps of the algorithm are overviewed in Fig. 4.2 and are described in detail below.

4.2.1 Problem definition

In quantitative analysis of genetic interactions we typically observe pairwise interactions between n genes and measure mutant phenotypes, such as the fitness of an organism or expression of a reporter gene (*Reporter*). Measurements over a set of double knockout mutants are given in a sparse matrix $\boldsymbol{G} \in \mathbb{R}^{n \times n}$ and those of single knockout mutants in a vector $\boldsymbol{S} \in \mathbb{R}^n$. In these matrices, $\boldsymbol{G}_{u,v}$ quantifies a phenotype of double mutant $u\Delta v\Delta$ and $\boldsymbol{S}_u$ denotes a phenotype of single mutant $u\Delta$. The expected mutant phenotypes, which represent phenotypes of double mutants in the absence of genetic interactions, are given by a matrix $\boldsymbol{H}$.

We aim to reconstruct a gene network that is consistent with pairwise gene relationships inferred from $\boldsymbol{G}$, $\boldsymbol{H}$ and $\boldsymbol{S}$. Inputs to network reconstruction are preferential scores for all four modeled gene relationships that include epistasis $u \rightarrow v$, epistasis $u \leftarrow v$, parallelism $v||u$, and partial interdependence $v\triangle u$ (Table 4.1). Réd represents the scores as $\boldsymbol{P} = (\boldsymbol{P}^{\rightarrow}, \boldsymbol{P}^{\leftarrow}, \boldsymbol{P}^{||}, \boldsymbol{P}^{\triangle})$ and computes them from the latent gene representation, which is obtained in the inference of a factorized model.

4.2.2 Factorized model of interactions

To deal with noise and address possibly incomplete input data, Réd estimates probabilities of gene relationships through a factorized model. We utilize a Bayesian inference approach and formulate the conditional probability of observed double mutant phenotype data, given their latent representation, as:

$$p(\boldsymbol{G}|\boldsymbol{U},\boldsymbol{V},\boldsymbol{\Psi},\sigma_G^2) = \prod_{u=1}^{n}\prod_{v=1}^{n}(\mathcal{N}(\boldsymbol{G}_{u,v}|g(\boldsymbol{U}_u^T\boldsymbol{V}_v;\boldsymbol{\Psi}_{u,v}),\sigma_G^2))^{I_{u,v}^G},$$

where $\mathcal{N}(x|\mu,\sigma^2)$ is a normal distribution with mean μ and variance σ^2, and $I_{u,v}^G$ indicates if the phenotypic measurement of $u\Delta v\Delta$ is available.

We assume that the observed phenotype of $u\Delta v\Delta$ is governed by the latent features associated with both genes u and v. In order to learn the latent features of u and v, we factorize double mutant phenotype data ($\boldsymbol{G}$) into a product of two low-dimensional latent matrix factors $\boldsymbol{U}^{k\times n}$ and $\boldsymbol{V}^{k\times n}$. Their column vectors, $\boldsymbol{U}_u$ and $\boldsymbol{V}_v$, represent k-dimensional u-specific and v-specific gene latent feature vectors, respectively. Instead of using linear latent Gaussian model of gene interactions, we pass the dot product $\boldsymbol{U}_u^T\boldsymbol{V}_v$ through a parameterized logistic function g. Thus, the model of interaction between genes u and v is represented by the factorized parameter $g(\boldsymbol{U}_u^T\boldsymbol{V}_v;\boldsymbol{\Psi}_{u,v})$. In the factorization, gene interactions depend on each other as they overlap and share parameters. For instance, given genes u, v, and w, their factorized parameters $g(\boldsymbol{U}_u^T\boldsymbol{V}_v;\boldsymbol{\Psi}_{u,v})$ and $g(\boldsymbol{U}_u^T\boldsymbol{V}_w;\boldsymbol{\Psi}_{u,w})$ share a common gene latent feature vector $\boldsymbol{U}_u$.

Parametrized logistic function g is given by:

$$g(x;\psi^{(1)},\psi^{(2)},\psi^{(3)}) = \frac{\psi^{(3)}}{1+\psi^{(1)}\exp(-\psi^{(2)}x)}$$

and bounds the range of factorized parameters by modeling saturation of the *Reporter*. Here, parameter $\psi^{(3)}$ represents the limiting value of the output past which g cannot grow and $\psi^{(1)}$ represents the number of times that $\boldsymbol{U}_u^T\boldsymbol{V}_v$ must grow to reach the value of $\psi^{(3)}$. If $\psi^{(2)}$ is positive, g is increasing in x, otherwise g is a decreasing function. Notice that $g(x;1,1,1)$ corresponds to the well-known sigmoid function. For every double mutant $u\Delta v\Delta$ we represent its logistic function parameters in a triple $\boldsymbol{\Psi}_{u,v} = (\boldsymbol{\Psi}_{u,v}^{(1)}, \boldsymbol{\Psi}_{u,v}^{(2)}, \boldsymbol{\Psi}_{u,v}^{(3)})$ and define $\boldsymbol{\Psi}$ to hold the parameterized logistic function representation over all possible double mutants: $\boldsymbol{\Psi} = (\boldsymbol{\Psi}^{(1)}, \boldsymbol{\Psi}^{(2)}, \boldsymbol{\Psi}^{(3)})$. We reduce the complexity of this factorized model in Sec. 4.2.3 by replacing dense parameterization of $\boldsymbol{\Psi}$ (one parameter set for every factorized parameter, $|\boldsymbol{\Psi}| = 3n^2$) with gene-dependent parameterization (one parameter set for every gene, $|\boldsymbol{\Psi}| = 3n$).

We employ a Gaussian prior centered at 1 for logistic function parametrization $\boldsymbol{\Psi}$ over given phenotypic measurements:

$$p(\boldsymbol{\Psi}|\sigma_\Psi^2) = \prod_{i=1}^{3}\prod_{u=1}^{n}\prod_{v=1}^{n}(\mathcal{N}(\boldsymbol{\Psi}_{u,v}^{(i)}|1,\sigma_\Psi^2\boldsymbol{I}))^{I_{u,v}^G}.$$

For gene latent feature vectors in $\boldsymbol{U}$ and $\boldsymbol{V}$ we assume zero-mean Gaussian priors to avoid overfitting:

$$p(\boldsymbol{U}|\sigma_U^2) = \prod_{u=1}^{n}\mathcal{N}(\boldsymbol{U}_u|\boldsymbol{0},\sigma_U^2\boldsymbol{I}),\;\; p(\boldsymbol{V}|\sigma_V^2) = \prod_{v=1}^{n}\mathcal{N}(\boldsymbol{V}_v|\boldsymbol{0},\sigma_V^2\boldsymbol{I}).$$

Through Bayesian inference we derive the posterior probability of gene latent vectors and logistic function parametrization given the available double mutants phenotypes:

$$\begin{aligned} p(\boldsymbol{U},\boldsymbol{V},\boldsymbol{\Psi}|\boldsymbol{G},\sigma_G^2,\sigma_U^2,\sigma_V^2,\sigma_\Psi^2) \quad &\propto \quad p(\boldsymbol{G}|\boldsymbol{U},\boldsymbol{V},\boldsymbol{\Psi},\sigma_G^2)p(\boldsymbol{U}|\sigma_U^2) \\ &\quad p(\boldsymbol{V}|\sigma_V^2)p(\boldsymbol{\Psi}|\sigma_\Psi^2). \end{aligned} \tag{4.1}$$

We select the factorized model according to the maximum *a posteriori* (MAP) estimation by maximizing the log-posterior of Eq. (4.1) over latent feature matrices and logistic function parametrization. The measurement noise variance (σ_G^2) and prior variances (σ_U^2, σ_V^2 and σ_Ψ^2) are kept fixed. Finding maximum *a posteriori* is equivalent to minimizing the following objective function, which is a sum of squared errors with quadratic regularization terms:

$$\begin{aligned}\mathscr{L}(\boldsymbol{G},\boldsymbol{U},\boldsymbol{V},\boldsymbol{\Psi}) \quad = \quad & \frac{1}{2}\sum_{u=1}^{n}\sum_{v=1}^{n} I_{u,v}^{G}(\boldsymbol{G}_{u,v} - g(\boldsymbol{U}_u^T\boldsymbol{V}_v;\boldsymbol{\Psi}_{u,v}))^2 \\ & +\frac{\lambda_U}{2}\sum_{u=1}^{n}\boldsymbol{U}_u^T\boldsymbol{U}_u + \frac{\lambda_V}{2}\sum_{v=1}^{n}\boldsymbol{V}_v^T\boldsymbol{V}_v \\ & +\frac{\lambda_\Psi}{2}\sum_{i=1}^{3}\sum_{u=1}^{n}\sum_{v=1}^{n} I_{u,v}^{G}(\boldsymbol{\Psi}_{u,v}^{(i)} - 1)^2, \end{aligned} \tag{4.2}$$

where $\lambda_U = \sigma_G^2/\sigma_U^2$, $\lambda_V = \sigma_G^2/\sigma_V^2$ and $\lambda_\Psi = \sigma_G^2/\sigma_\Psi^2$.

Here, $\boldsymbol{\Psi}$, $\boldsymbol{U}$ and $\boldsymbol{V}$ are unknown, and unfortunately the function $\mathscr{L}$ is not convex in all unknowns. In particular, $\mathscr{L}$ is convex in either $\boldsymbol{U}$ or $\boldsymbol{V}$ but not in both factors together, which is a known result from matrix factorization studies (Lee and Seung, 2000; Koren et al., 2009). In our study, $\mathscr{L}$ is further coupled by the parametrization of $\boldsymbol{\Psi}$. Thus, it is unrealistic to expect an algorithm to solve the optimization problem defined by $\mathscr{L}$ in the sense of finding global minimum. We thus estimate latent features and logistic function parameters by finding a local minimum of the objective function $\mathscr{L}$ through application of gradient descent. Derivatives of $\mathscr{L}$ with respect to gene latent features and logistic parameters are given by:

$$\frac{\partial \mathscr{L}}{\partial \boldsymbol{U}_u} = \sum_{v=1}^{n} h(u,v)\boldsymbol{V}_v g'(\boldsymbol{U}_u^T \boldsymbol{V}_v; \boldsymbol{\Psi}_{u,v}) + \lambda_U \boldsymbol{U}_u, \tag{4.3}$$

$$\frac{\partial \mathscr{L}}{\partial \boldsymbol{V}_v} = \sum_{u=1}^{n} h(u,v)\boldsymbol{U}_u g'(\boldsymbol{U}_u^T \boldsymbol{V}_v; \boldsymbol{\Psi}_{u,v}) + \lambda_V \boldsymbol{V}_v, \tag{4.4}$$

$$\frac{\partial \mathscr{L}}{\partial \boldsymbol{\Psi}_{u,v}^{(1)}} = -\frac{h(u,v)\boldsymbol{\Psi}_{u,v}^{(3)} \exp(\boldsymbol{\Psi}_{u,v}^{(2)} \boldsymbol{U}_u^T \boldsymbol{V}_v)}{(\exp(\boldsymbol{\Psi}_{u,v}^{(2)} \boldsymbol{U}_u^T \boldsymbol{V}_v) + \boldsymbol{\Psi}_{u,v}^{(1)})^2} + t(u,v,1), \tag{4.5}$$

$$\frac{\partial \mathscr{L}}{\partial \boldsymbol{\Psi}_{u,v}^{(2)}} = \frac{h(u,v)\boldsymbol{\Psi}_{u,v}^{(1)}\boldsymbol{\Psi}_{u,v}^{(3)} \boldsymbol{U}_u^T \boldsymbol{V}_v \exp(\boldsymbol{\Psi}_{u,v}^{(2)} \boldsymbol{U}_u^T \boldsymbol{V}_v)}{(\exp(\boldsymbol{\Psi}_{u,v}^{(2)} \boldsymbol{U}_u^T \boldsymbol{V}_v) + \boldsymbol{\Psi}_{u,v}^{(1)})^2} + t(u,v,2), \tag{4.6}$$

$$\frac{\partial \mathscr{L}}{\partial \boldsymbol{\Psi}_{u,v}^{(3)}} = \frac{h(u,v)}{1 + \boldsymbol{\Psi}_{u,v}^{(1)} \exp(-\boldsymbol{\Psi}_{u,v}^{(2)} \boldsymbol{U}_u^T \boldsymbol{V}_v)} + t(u,v,3), \tag{4.7}$$

where for convenience of notation $h(u,v)$ is substituted for:

$$h(u,v) = \boldsymbol{I}_{u,v}^{G}(g(\boldsymbol{U}_u^T \boldsymbol{V}_v; \boldsymbol{\Psi}_{u,v}) - \boldsymbol{G}_{u,v}),$$

penalty term $t(u,v,i)$ stands for $t(u,v,i) = \lambda_\Psi \boldsymbol{I}_{u,v}^{G}(\boldsymbol{\Psi}_{u,v}^{(i)} - 1)$, and $g'(x; \boldsymbol{\Psi}_{u,v})$ is logistic function derivative with respect to x. Efficiency in training Réd model comes from finding point estimates of model unknowns instead of inferring the full posterior distribution over them.

4.2.3 Gene-dependent weighting

We further reduce complexity of the model described in the previous section by combining evidence from multiple phenotypic measurements through their latent representation. We replace entrywise (double-mutant-phenotype-dependent) logistic function parametrization $\boldsymbol{\Psi}$ with gene-dependent parametrization that is given by $\boldsymbol{\Psi}_{u,v}^{(i)} \leftarrow \frac{1}{n-1}\sum_w \boldsymbol{\Psi}_{u,w}^{(i)}$ for $i = 1,2,3$. This reduces the number of parameters in $\boldsymbol{\Psi}$ that have to be learned from $3n^2$ to $3n$. Intuitively, measurements that involve gene u are not independent from each other but are rather governed by the gene pathways in which u participates. Gene-dependent parametrization of $\boldsymbol{\Psi}$ represents a method of regularization allowing us to remove penalty terms in Eqs. (4.5)–(4.7).

Derivatives of $\boldsymbol{\Psi}$ utilize only available phenotypic measurements due to the application of an indicator function (cf. Eqs. (4.5)–(4.7)). We relax this limitation by considering current estimates of $\boldsymbol{G}$ when computing the derivatives of $\boldsymbol{\Psi}$. These estimates are given by $\widehat{\boldsymbol{G}}_{u,v} = g(\boldsymbol{U}_u^T \boldsymbol{V}_v; \boldsymbol{\Psi}_{u,v})$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are latent matrix factors from the previous iteration of gradient descent (step 3c in Algorithm 2).

4.2.4 *Preferential order-of-action scoring of gene pairs*

Probabilities of gene-gene relationships in $\boldsymbol{P}$ are computed from the inferred phenotypes given by $\widehat{\boldsymbol{G}} = g(\boldsymbol{U}^T \boldsymbol{V}; \boldsymbol{\Psi})$ with the rules outlined in Table 4.1. Estimated probabilities in $\boldsymbol{P}$ approach 1 when inferred phenotypic values in $\widehat{\boldsymbol{G}}$ are close to the phenotypes, which would be expected if a certain network structure ($\rightarrow$, $\leftarrow$, $||$, $\triangle$) existed between genes, and they slowly vanish when the inferred values deviate from the values expected by a certain type of relationship.

For instance, an epistatic genetic interaction $u \leftarrow v$ is inferred when the trait $\widehat{\boldsymbol{G}}_{u,v}$ of the double mutant $u\Delta v\Delta$ is very similar to the single mutant $u\Delta$ phenotype $\boldsymbol{S}_u$ and the two single mutant phenotypes are different ($\boldsymbol{S}_u \not\approx \boldsymbol{S}_v$). This brings $|\widehat{\boldsymbol{G}}_{u,v} - \boldsymbol{S}_u|$ close to 0 and, consequently, $\boldsymbol{P}^{\leftarrow}_{u,v}$ close to 1. With different single mutant phenotypes, the expected phenotype $\boldsymbol{H}_{u,v}$ of the double mutant that assumes no genetic interaction is different from both single mutant phenotypes ($\boldsymbol{S}_u \not\approx \boldsymbol{S}_v \Rightarrow \boldsymbol{S}_v \not\approx \boldsymbol{H}_{u,v} \wedge \boldsymbol{S}_u \not\approx \boldsymbol{H}_{u,v}$), bringing $\boldsymbol{P}^{||}_{u,v}$ and $\boldsymbol{P}^{\triangle}_{u,v}$ close to 0. Likewise, the phenotype of $v\Delta$ would be different from the phenotype of the double mutant, bringing $\boldsymbol{P}^{\rightarrow}_{u,v}$ close to 0.

Cases with less pronounced differences between phenotypes would lead to smaller differences in relationship probabilities. Preferential order-of-action scores generalize the epistasis analysis framework by Avery and Wasserman (1992), wherein the signal and the genes under study were strictly on or off with no intermediate levels of activity. An appealing feature of scores in $\boldsymbol{P}$ is that they have a direct probabilistic interpretation.

4.2.5 *Multi-gene network inference*

Given probabilistic scores of gene-gene network structures in $\boldsymbol{P}$ from Sec. 4.2.4, we reconstruct a detailed multi-gene network that is consistent with the inferred relationship probabilities and contains a minimum number of *violating* and *redundant* edges.

Algorithm 2

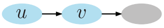
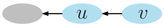
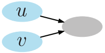
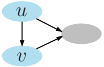
Réd, the proposed approach for gene network inference by scoring relationships from a factorized model of interactions. Source code is available at http://github.com/biolab/red.

```
Input:
    ▪ sparse matrix of double mutant phenotypes $\boldsymbol{G} \in \mathbb{R}^{n \times n}$,
    ▪ typical interaction values $\boldsymbol{H} \in \mathbb{R}^{n \times n}$,
    ▪ measured phenotypes of single mutants $\boldsymbol{S} \in \mathbb{R}^{n}$,
    ▪ parameters $\lambda_U$, $\lambda_V$, rates $\alpha$ and $\beta$, and rank $k$.
Output:
    ▪ preferential order-of-action score matrices $\boldsymbol{P}$,
    ▪ completed matrix $\widehat{\boldsymbol{G}}$,
    ▪ gene-dependent logistic function parametrization $\boldsymbol{\Psi}$,
    ▪ inferred gene network for a gene subset of interest.

1. Initialize $\boldsymbol{U} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})^{k \times n}$ and $\boldsymbol{V} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})^{k \times n}$.
2. Initialize $\boldsymbol{\Psi}^{(i)}$ as $\boldsymbol{1}_{n \times n}$ for $i = 1, 2, 3$.
3. Repeat until convergence:
    a. Compute $\frac{\partial \mathcal{L}}{\partial U}$ and $\frac{\partial \mathcal{L}}{\partial V}$ with Eq. (4.3) and Eq. (4.4), respectively.
    b. Update $\boldsymbol{U} \leftarrow \boldsymbol{U} - \alpha \frac{\partial \mathcal{L}}{\partial U}$ and $\boldsymbol{V} \leftarrow \boldsymbol{V} - \alpha \frac{\partial \mathcal{L}}{\partial V}$.
    c. Compute $\frac{\partial \mathcal{L}}{\partial \Psi^{(i)}}$ for $i = 1, 2, 3$ using Eqs. (4.5)–(4.7),
       respectively. Substitute $h(u, v)$ therein with $h(u, v) = g(\boldsymbol{U}_u^T \boldsymbol{V}_v; \boldsymbol{\Psi}_{u,v}) - \boldsymbol{X}_{u,v}$,
       where $\boldsymbol{X}_{u,v} = \boldsymbol{G}_{u,v}$ if $\boldsymbol{I}^G_{u,v} = 1$ and $\boldsymbol{X}_{u,v} = \widehat{\boldsymbol{G}}_{u,v}$ otherwise. Here, $\widehat{\boldsymbol{G}}_{u,v}$ is
       computed using the latent matrix factors from the previous iteration.
    d. Update $\boldsymbol{\Psi}^{(i)} \leftarrow \boldsymbol{\Psi}^{(i)} - \beta \frac{\partial \mathcal{L}}{\partial \Psi^{(i)}}$ for $i = 1, 2, 3$.
    e. Set gene-dependent weights $\boldsymbol{\Psi}^{(i)}_{u,v} \leftarrow \frac{1}{n-1} \sum_w \boldsymbol{\Psi}^{(i)}_{u,w}$ for $i = 1, 2, 3$ and $\forall u, v$.
4. Compute preferential order-of-action scores $\boldsymbol{P}^i_{u,v}$ for $i \in \{\rightarrow, \leftarrow, ||, \triangle\}$ and $\forall u, v$ using Eqs. from Table 4.1.
5. Normalize $\boldsymbol{P}^i_{u,v} \leftarrow \boldsymbol{P}^i_{u,v} / \sum_j \boldsymbol{P}^j_{u,v}$ for $i \in \{\rightarrow, \leftarrow, ||, \triangle\}$ and $\forall u, v$.
6. Compute $\widehat{\boldsymbol{G}}_{u,v} = g(\boldsymbol{U}_u^T \boldsymbol{V}_v; \boldsymbol{\Psi}_{u,v})$.
7. Given a gene subset of interest, infer a network (Sec. 4.2.5).
```

Table 4.1

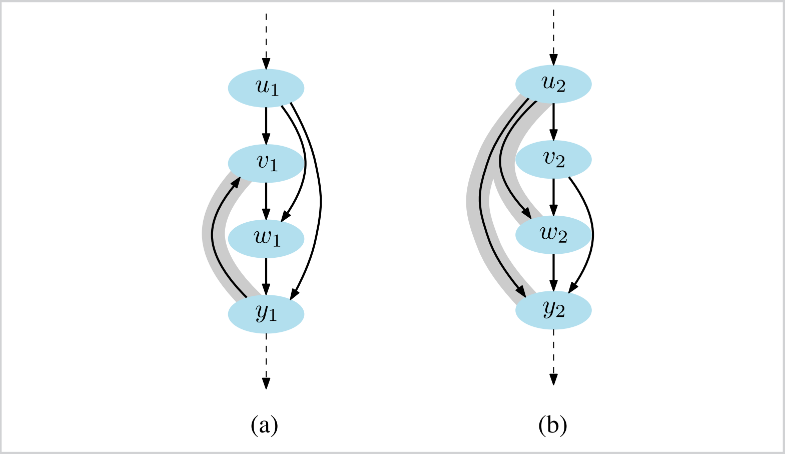
Probabilistic scoring of gene-gene relationships. Given genes u and v, the table shows all four pairwise relationships and their corresponding network structures. These relationships have already been considered by Battle et al. (2010) but are here studied with probabilistic scoring functions. See main text for explanation of preferential order-of-action scores.

Gene-gene relationship	Network structure	Preferential order-of-action score
u and v in a linear pathway, v downstream, gene v is *epistatic* to gene u		$P^{\rightarrow}_{u,v} = \frac{2}{1+\exp(\lvert\widehat{G}_{u,v}-S_v\rvert)}$
u and v in a linear pathway, u downstream, gene u is *epistatic* to gene v		$P^{\leftarrow}_{u,v} = \frac{2}{1+\exp(\lvert\widehat{G}_{u,v}-S_u\rvert)}$
u and v affect the reporter separately		$P^{\vert\vert}_{u,v} = \frac{2}{1+\exp(\lvert\widehat{G}_{u,v}-H_{u,v}\rvert)}$
u and v are partially interdependent, each has also a path to the reporter that is independent of the other		$P^{\triangle}_{u,v} = \frac{2}{1+\exp(\lvert\widehat{G}_{u,v}-\frac{1}{2}(H_{u,v}+\max(S_u,S_v))\rvert)}$

Examples of inferred networks are given in Figs. 4.4–4.7. A network is a weighted directed graph with genes as vertices and directed edges that determine the order of action. A designated vertex represents the observed quantitative trait. A directed edge from u to v is *violating* (Fig. 4.3a) if there is evidence in $\boldsymbol{P}$ for both $u \rightarrow v$ and $u \leftarrow v$ (*e.g.* $\boldsymbol{P}^{\rightarrow}_{u,v} \approx \boldsymbol{P}^{\leftarrow}_{u,v}$). A directed edge from u to v is *redundant* (Fig. 4.3b) if there is evidence in $\boldsymbol{P}$ that some intermediate gene exists between u and v. That is, u and v are not adjacent in a genetic network but rather u indirectly affects v, *i.e.,* $\boldsymbol{P}^{\rightarrow}_{u,v}$ captures the extent to which strict weak ordering of u and v holds.

Network inference procedure assigns a level to every gene in a manner that if there is strong evidence in $\boldsymbol{P}$ that gene u is placed upstream of gene v, that is, if v is epistatic to u, then level(u) > level(v). In the case of stronger evidence of parallelism or partial interdependence between u and v the level(u) ≈ level(v). Several genes can be assigned the same level, but a designated vertex corresponding to a phenotype of interest is the only vertex placed on the lowest level.

Inference of a genetic network involves two phases. In the first phase we perform an approximate topological sort through construction of a directed weighted graph. Given genes u and v and the inferred epistasis relationships between them, the direction and weight of a between-level edge are determined by the maximum of the values $\boldsymbol{P}^{\rightarrow}_{u,v}$ (edge $u \rightarrow v$) and $\boldsymbol{P}^{\leftarrow}_{u,v}$ (edge $u \leftarrow v$). Given a parallelism or partial interdependence relationship between u and v, a within-level edge is determined by the maximum of



(a) (b)

Figure 4.3

Illustration of violating (a) and redundant (b) edges (in grey) in a pathway with four genes. Edge $y_1 \rightarrow v_1$ is violating because there is evidence that v_1 is placed upstream of y_1 ($v_1 \rightarrow w_1$ and $w_1 \rightarrow y_1$) but also that y_1 is upstream of v_1 ($y_1 \rightarrow v_1$). Edge $u_2 \rightarrow w_2$ is redundant because there is evidence of an intermediate gene v_2. Similarly, edge $u_2 \rightarrow y_2$ is redundant because of two intervening genes, v_2 and w_2.

the values $\boldsymbol{P}^{||}_{u,v}$ (no edge between u and v) and $\boldsymbol{P}^{\triangle}_{u,v}$ (edge $u \rightarrow v$). This graph may contain directed cycles and finding an exact topological ordering of its vertices with the minimal set of violating edges is a known NP-hard problem (Eades et al., 1993; Charbit et al., 2007). Thus, we proceed in the following way. We select a vertex with no incoming between-level edges, assign that vertex to the currently top-most level and recurse on the graph with that vertex removed. We also look for vertices with no outgoing between-level edges and assign them to the currently lowest level. If in some step multiple vertices have no incoming or outgoing between-level edges, they are assigned the same level. It can happen that all vertices have incoming and outgoing between-level edges. In this case, we select the vertex with the highest differential between weighted incoming between-level degree and weighted outgoing between-level degree.

In the second phase of gene network inference we retain within-level edges and those edges that link adjacent levels and are directed downwards. The latter procedure eliminates violating edges. As a final step, we remove redundant edges according to their definition above.

4.3 Present mutant phenotype data and known gene pathways

We assess the accuracy of Réd by applying our inference approach to the data sets of Jonikas et al. (2009) and Surma et al. (2013) and compare results to known or partially known networks. Experiments that use data from Jonikas et al. closely follow the setup by Battle et al. and use the same data sets and reference pathways.

Mutant phenotype data

Jonikas et al. (2009) measured unfolded protein response (UPR) levels in single and double mutants to systematically characterize functional interdependence of yeast genes with roles in endoplasmic reticulum (ER) folding. The data set contains 444 genes that caused high UPR reporter inductions. The interaction data include phenotypes of 42,240 distinct double mutants (matrix $\boldsymbol{G}$) corresponding to 43% of all possible double mutants. Jonikas et al. also computed typical (*i.e.* expected) values of genetic interactions for every double mutant (matrix $\boldsymbol{H}$). They considered multiplicative neutrality function (Mani et al., 2008) and computed it using reporter levels of pairs of single mutants, modified by a Hill function to account for the saturation of the reporter signal.

Surma et al. (2013) considered 741 genes and observed the growth phenotype (colony size) for all pairs of double mutants. In total, after filtering out unreliable measurements, their data set comprises 251,383 double mutant fitness scores. We computed single mutant scores by averaging across all scores of double mutants that included mutations of the corresponding genes. We considered multiplicative model to calculate the expected fitness of a double mutant in the absence of a genetic interaction.

Gene pathways

We compare gene networks inferred by Réd to a number of known or partially known cellular pathways that include genes whose perturbations are measured by Jonikas et al.:

- the N-linked glycosylation pathway consisting of 10 genes whose true ordering is known (Helenius and Aebi, 2004),

- the ER-associated degradation (ERAD) pathway for which many functional interdependencies between its member genes are known,
- tail-anchored protein biogenesis machinery consisting of tail-anchored (TA) proteins important for transmembrane trafficking and the recently discovered GET pathway (Stefanovic and Hegde, 2007; Schuldiner et al., 2008; Bozkurt et al., 2009).

We also compare Réd's networks to well-characterized cellular pathways of phospholipid biosynthesis whose gene mutants are measured by Surma et al. and that include:

- the Kennedy pathway involved in the synthesis of phosphatidylethanolamine and phosphatidylcholine (PC), and
- the phosphatidylserine to PC conversion pathway.

Experimental setup

In the first part of the experiments, we use mutant phenotype data to qualitatively evaluate the reconstruction of five gene pathways from Sec. 4.3. In the second part of the experiments, we evaluate the accuracy of gene ordering through three different setups. In the first two setups, the data-inferred gene ordering was compared to the known pathways. In the third setup, we use cross-validation to estimate the accuracy of prediction of gene interaction scores with the following experiments:

1. Battle et al. provided 168 test gene pairs (v, u) from common KEGG pathways (Kanehisa et al., 2008). For 21 gene pairs v is known to be upstream of u, and for 147 gene pairs v is not known to be upstream of u. Given a gene pair, Réd predicted the probability of epistasis as $\boldsymbol{P}^{\rightarrow}_{u,v}/(\boldsymbol{P}^{\rightarrow}_{u,v} + \boldsymbol{P}^{\leftarrow}_{u,v})$, and the accuracy of predictions on entire set of 168 gene pairs.
2. Using the setup from Battle et al. we evaluate the accuracy of prediction of direct edges $u \rightarrow v$ in the N-linked glycosylation pathway (Fig. 4.4) based on the model-estimated probability of epistasis $\boldsymbol{P}^{\rightarrow}_{u,v}$.
3. We estimate the accuracy when predicting that two genes are in epistasis, that is $u \rightarrow v$ or $v \rightarrow u$. Notice that in the literature this relationship is also referred to

as an *alleviating interaction*, where the phenotype of a double mutant is less severe than expected from the phenotypes of the corresponding single mutants (Mani et al., 2008; Jonikas et al., 2009). For the data from Jonikas et al. this means that the double mutant cell responds to ER stress surprisingly better than how the ER stress would typically be mitigated. The data for this experiment was preprocessed according to the procedure described by Battle et al.. A positive set included gene pairs (u, v) with significant alleviating genetic interactions, for which the observed phenotype (interaction score) was negative with a magnitude greater than $|\boldsymbol{G}_{u,v} - \max(\boldsymbol{S}_u, \boldsymbol{S}_v)|$ (see St Onge et al. (2007)). It was further required that the double mutant phenotype data contained a sufficient number of observations that included $u\Delta$ or $v\Delta$, such that the geometric mean of such measurements for u and for v was at least 180. There are 2723 gene pairs in the data of Jonikas et al. that match these criteria. In each test run, we form a test set with a random selection of 5% of the positive gene pairs and a negative set of equal size of gene pairs that fail to satisfy the selection criteria. We remove the test data from the interaction score matrix $\boldsymbol{G}$, and predict whether a test gene pair is alleviating using the probability that u and v occur together in a linear pathway, *i.e.* $\boldsymbol{P}^{\rightarrow}_{u,v} + \boldsymbol{P}^{\leftarrow}_{u,v}$. We report an averaged accuracy across ten different test runs.

We characterize the accuracy of predictions through the area under the ROC curve (AUC), with a baseline of 0.5 (random networks) and a perfect score of 1.0 (inferred networks that are identical to gold standard – known networks).

We compare Réd, our network inference approach, to a recently published Bayesian approach by Battle et al.. They developed preference scoring functions over all possible pairwise gene relationships and applied annealed importance sampling to reconstruct high scoring multi-gene networks. Their method (referred here as APN) was shown to be superior to a number of other approaches that can infer networks from gene interaction data by Jonikas et al.. These other approaches include baseline techniques such as Pearson correlation of genetic interaction profiles and raw interaction values as well as more sophisticated techniques such as Gaussian process regression (GP; Williams and Rasmussen (1996)), a method that uses the correlation of observed interaction profiles, the diffusion kernel method (DK; Qi et al. (2008)) and GenePath (Zupan et al., 2003). For brevity, we therefore focus on comparing our method with APN, which was run with default parameters as chosen by Battle et al. for the data set of Jonikas et al., but we also report the accuracies achieved by GP and DK.

Two essential components of Réd are latent representation of gene interactions and their transformation through the logistic function. To test the extent to which the performance of Réd depends on these two components we also run experiments where the algorithm infers probabilities and makes predictions from raw (not factorized) phenotypes, and where the latent representation is used without logistic transformation. We refer to these two approaches as RAW and MF, respectively.

In all experiments with data from Jonikas et al., the parameters of Réd are set as: $\lambda_U = \lambda_V = 1 \times 10^{-4}$, $\beta = 0.1$, $\alpha = 0.1$, $k = 100$. The same parameters are used on data from Surma et al. with the exception of $\alpha = 1 \times 10^{-3}$ and $k = 50$, that were selected to minimize the normalized root mean square error of $\widehat{\boldsymbol{G}}$. This choice of regularization parameters and learning rates is common (cf. Min and Lee (2005); Pedregosa et al. (2011)). We also show (see Sec. 4.4.6) that the performance of Réd does not critically depend on the rank of factorization k. Réd's optimization by gradient descent is terminated when the Frobenius distance between $\boldsymbol{G}$ and $\widehat{\boldsymbol{G}}$ over known values fails to decrease between the two consecutive iterations of optimization.

4.4 A case study: reconstruction of known gene pathways

4.4.1 Reconstruction of a gene pathway from data by Jonikas et al.

We analyzed the ability of Réd to reconstruct the known N-linked glycosylation pathway. Fig. 4.4 shows the inferred network next to the known pathway as reported by Helenius and Aebi (2004). Genes *CWH41, DIE2* and *ALG8* are correctly placed such that they are dependent on the other genes. Also, *ALG12* is placed upstream of *ALG9*, which is also upstream of *ALG3*. *OST3* is correctly placed downstream, but *OST5* is incorrectly placed, likely because double mutant data with the other ALG genes were not available. Surprisingly, Réd correctly placed *CWH41*, a gene which encodes glucosidase I, an integral membrane protein of the ER involved in sensing ER stress (Romero et al., 1997), at the beginning of the pathway despite mild downstream effects observed in *CWH41* mutants. Notice that the interaction profile of *CWH41* is only moderately correlated with the those of ALG genes, thus *CWH41* was not clustered together with them (Jonikas et al., 2009). We hence conclude that Réd inference of the N-linked glycans synthesis pathway was successful with a network that closely resembles that reported in the literature.

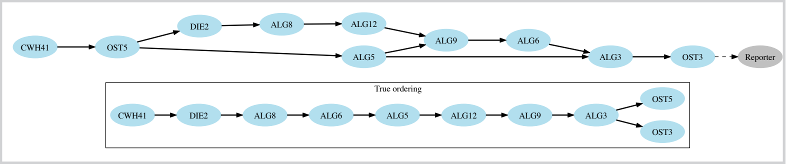


Figure 4.4

Gene network of the N-linked glycosylation pathway inferred by Réd. For reference, we show the true ordering of this pathway (Helenius and Aebi, 2004) as adapted from Battle et al. (2010). The inferred gene network reflects many correct gene placements.

4.4.2 Reconstruction of gene pathways from data by Surma et al.

We applied Réd to mutant data by Surma et al. to reconstruct two thoroughly studied pathways of phospholipid biosynthesis. Réd's ordering of genes in the phosphatidylserine to phosphatidylcholine conversion pathway is fully consistent with the reference pathway (Fig. 4.5a). In the Kennedy pathway, Réd correctly placed *PCT1* upstream of *CPT1* and *CKI1* upstream of *CPT1* with high confidence (Fig. 4.5b), but it misplaced gene pair *PCT1* and *CKI1* likely due to the ambiguity in the data. However, as Réd performs global reasoning by combining evidence from all measurements, it handled the data uncertainty by assigning PCT1 → CKI1 structure the lowest score in the reconstruction of the Kennedy pathway.

4.4.3 Reconstruction of partially known gene pathways

Jonikas et al. (2009) identified several pathways that are important for ER protein folding. Of these, the pathways for ER-associated degradation and tail-anchored protein insertion were considered in Battle et al. (2010). Réd-inferred networks for these two pathways are shown in Figs. 4.6–4.7. The solid edges in these figures are those inferred by our algorithm, while the dotted edges indicate gene interactions reported in the literature (Jonikas et al., 2009; Battle et al., 2010; Kim et al., 2005; Carvalho et al., 2006; Nakatsukasa and Brodsky, 2008; Clerc et al., 2009).

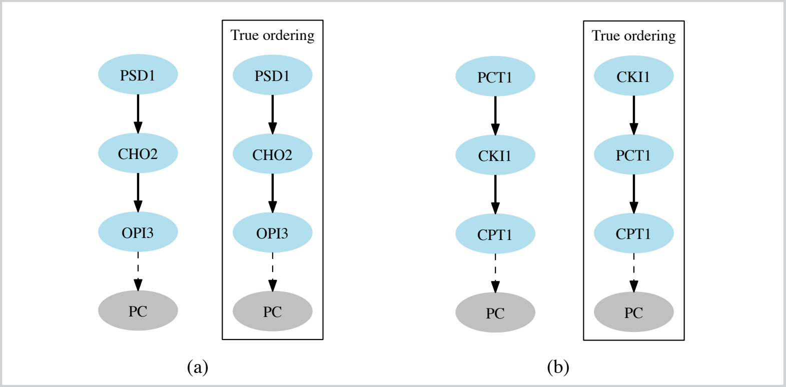


Figure 4.5

Gene networks of the phosphatidylserine to phosphatidylcholine (PC) conversion pathway (a) and the Kennedy pathway (b) as inferred by Réd. For reference, we show the true orderings in both pathways adapted from Surma et al. (2013). Réd correctly and with high confidence ($P > 0.80$) inferred all three pairwise gene relationships of the PC conversion pathway. It also correctly predicted two out of three gene relationships of the Kennedy pathway with the wrong prediction (PCT1 → CKI1) being assigned a low confidence ($P = 0.25$).

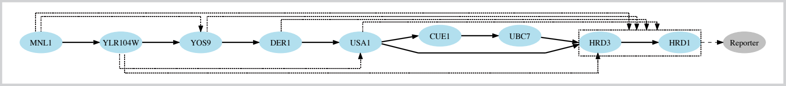


Figure 4.6

The ER-associated degradation (ERAD) pathway predicted by Réd is shown by solid edges. Placement of genes in the inferred network is very consistent with known interdependencies (dotted edges).

The ordering of inferred networks is entirely consistent with the partially known gene pathways. In the network for the ER-associated degradation pathway (Fig. 4.6), the upstream placement of *MNL1* to *YOS9* is consistent with existing data showing that *MNL1* generates the sugar species recognized by *YOS9* (Clerc et al., 2009). Also, *MNL1*, *YOS9*, *DER1* and *USA1* are placed upstream of *HRD3* and *HRD1*, which is compatible with data showing that degradation of certain substrates requires all six components (Kim et al., 2005; Carvalho et al., 2006; Nakatsukasa and Brodsky, 2008). For the tail-anchored protein insertion pathway Réd inferred a network (Fig. 4.7) that placed the poorly characterized protein *SGT2* upstream of the tail-anchored protein biogenesis machinery components according to its function in the insertion of tail-anchored proteins into membranes (Battle et al., 2010).

Similarly, positive results of network inference are also reported in (Battle et al., 2010).

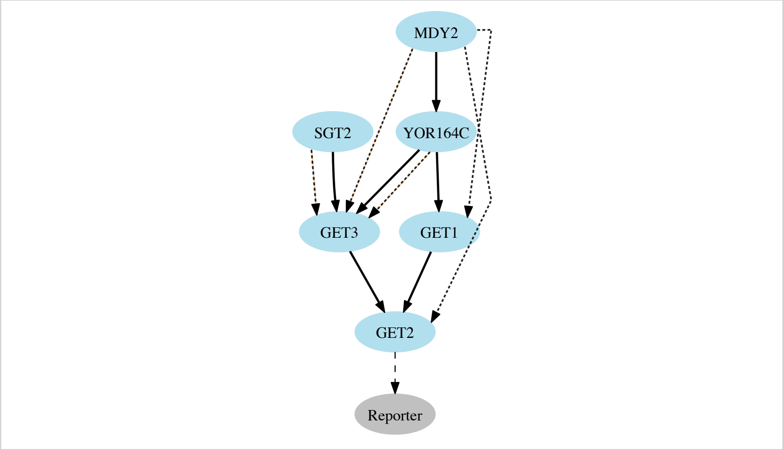


Figure 4.7

Gene network inferred by Réd that represents the likely ordering of genes belonging to the tail-anchored protein biogenesis machinery (solid edges). Known relationships between genes are denoted by dotted edges. Notice that the predicted ordering strongly reflects known interdependencies between genes.

Their method inferred a number of candidate networks of which the best-scored were shown to be partially consistent with known gene interdependencies. In contrast, for each pathway, Réd inferred a single network that is entirely consistent with known gene relationships.

4.4.4 Quantitative analysis of gene ordering

Table 4.2 reports the accuracies of gene ordering prediction obtained by four different algorithms, Réd, APN, and two simplified variants of Réd. In comparison to APN, Réd performs substantially better in predicting the edges of the KEGG pathways and slightly better in predicting the edges of the N-linked glycosylation pathway (Fig. 4.8).

The poor performance of the simplified variants of Réd (RAW and MF) indicates that Réd's latent representation inferred from the factorized model, the nonlinear logistic map and gene-dependent weighting are the essential components of Réd. Without any of these, Réd would not be able to achieve the resulting accuracy.

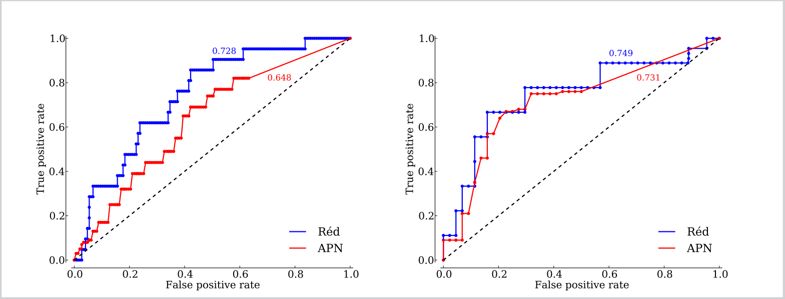


Figure 4.8

The ROC curves for the prediction of gene ordering in KEGG pathways (left) and the N-linked glycosylation pathway (right) by Réd, our proposed approach, and a Bayesian learning method APN (Battle et al., 2010)). Each curve is annotated with its corresponding area under the curve (AUC).

Table 4.2

The predictive accuracy (AUC) of gene ordering by a Bayesian learning method (APN; Battle et al. (2010)), Réd, our proposed approach, and its simplified variants: without factorization (RAW) and with factorization but in the absence of transformation by logistic function (MF).

Prediction	AUC			
	RAW	MF	APN	Réd
KEGG pathway ordering	0.563	0.583	0.648	0.728
N-linked glycosylation pathway	0.591	0.638	0.731	0.749

4.4.5 *Prediction of alleviating genetic interactions*

Given the training and separate test data sets, we predict whether an interaction is alleviating (see Sec 4.3). Table 4.3 shows that Réd performs substantially better than APN (p-value < 0.001). Réd also outperforms standard two-factor matrix factorization (MF) by a large margin, which is an indicator that transformation via a logistic map is essential to the performance of our algorithm. We compare these results with those obtained by Gaussian process regression (GP) (Williams and Rasmussen, 1996) using squared exponential autocorrelation model constructed from the genetic interaction profiles, and with the interactions predicted with the diffusion kernel method (DK) (Qi et al., 2008). Réd achieves significantly higher accuracy than GP (p-value < 0.01) and DK (p-value < 0.001), although the difference with GP is small and may be worth of further study. Notice that RAW, a Réd variant without factorization, is not applicable for this experiment as it does not generalize across gene interaction scores.

Table 4.3

Prediction of unknown alleviating genetic interactions. We report the accuracy of predicted interactions based on the diffusion kernel method (DK; Qi et al.), predictions based on latent representation obtained with standard two-factor matrix factorization (MF), APNs learned through a Bayesian method by Battle et al., predicted genetic interaction values from Gaussian process regression (GP; Williams and Rasmussen) that uses the correlation of observed interaction profiles, and Réd, our proposed approach.

Prediction	AUC				
	MF	DK	APN	GP	Réd
Alleviating interactions	0.723	0.759	0.783	0.862	0.906

We have observed that the probabilities of alleviating gene pairs predicted by Réd are well correlated to the strength of alleviating interactions (Spearman $r = -0.704$, p-value $< 1 \times 10^{-100}$; Fig. S3). Réd scores gene pairs with stronger alleviating effects (negative interaction values with greater magnitude) higher than those that interact moderately.

4.4.6 Sensitivity and repeatability analysis

We analyze the sensitivity of Réd to reduced measurement precision by introducing increasing levels of random noise to the data set of Jonikas et al. (2009) and, for each noise level, re-running inference by Réd with a fixed initialization of matrix factors. For every measurement of a single and double mutant in the data set we sample the noise component from a Gaussian distribution with zero mean and standard deviation s, and add this value to the original measurement. For each run, using a specific value for s, we compare all estimates in $\boldsymbol{P}$ to its original, noise-free estimates. Fig. 4.9 shows the correlation between the original estimates and estimates inferred from the noisy data set. The results suggest that good probability estimates of network relationships between genes are possible even in settings with increased noise. Thus, Réd could also infer accurate networks from data that includes more noise than otherwise present in the data set by Jonikas et al. (2009).

For twenty runs of Réd learning with different initializations of matrix factors $\boldsymbol{U}$ and $\boldsymbol{V}$, we estimate $\boldsymbol{P}$ for the edges potentially connecting each pair of genes. For every run we compare all probability estimates in $\boldsymbol{P}$ to the corresponding estimates from every other run. The maximum difference for any two runs and for any pair of genes is less than 1×10^{-8}, demonstrating that Réd estimates are highly repeatable and that the performance of Réd does not substantially vary with initialization of the latent factors.

Similarly, we run Réd several times for different values of the latent dimension k ($k \in \{40, 60, 80, 100, 120\}$). We compare the corresponding probability estimates in $\boldsymbol{P}$ from every two runs. The mean difference for any two runs and for any edge is less than 1×10^{-3} and the standard deviation is less than 1×10^{-2}. Thus, Réd is robust and performs well on the data by Jonikas et al. (2009) for a broad range of sensible values for the latent dimension.

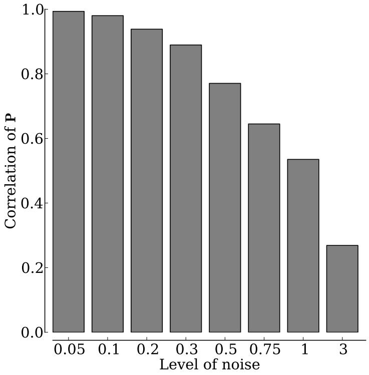


Figure 4.9

Sensitivity of $\boldsymbol{P}$ to measurement noise. We vary the level of Gaussian noise introduced into phenotypic measurements of single and double mutants for the Jonikas et al. data and compute the correlation between $\boldsymbol{P}$ as estimated from the original or noise induced data.

4.5 Conclusion

Réd is a conceptually new approach for inference of gene networks from quantitative genetic interaction data. It implements a probabilistic epistasis analysis and assembles pairwise relationships into gene networks. In our experiments, Réd was able to reconstruct several known and partially known pathways with accuracy above that of the state-of-the-art approaches. Réd outperforms APN, the state-of-the-art method by Battle et al. (2010), both in accuracy and speed, with CPU runtime of only a few minutes compared to APN's 30 minutes for an inference of a single full network in an ensemble of 500 networks. We also show that Réd's power of generalization comes from its two key components, a factorized model with latent representation of gene interactions and a gene-dependent logistic map of interaction scores.

Our evaluation in this chapter was computational and thus limited to data sets for which several gene pathways or at least partial gene orderings were available (Jonikas et al., 2009; Battle et al., 2010). Réd can efficiently handle similar data sets as well as much larger ones, such as that from the recent yeast experiments by Costanzo et al. (2010). These are also the data sets for which we foresee future applications of Réd and which will require subsequent verification of inferred networks in the wet lab.

5

Collective network inference

Markov networks are undirected graphical models that are widely used to infer relations between genes from experimental data. Their state-of-the-art inference procedures assume the data arise from a Gaussian distribution. High-throughput omics data, such as that from next generation sequencing, often violates this assumption. Furthermore, when collected data arise from *multiple* related but otherwise *nonidentical distributions*, their underlying networks are likely to have common features. New principled statistical approaches are needed that can deal with different data distributions and jointly consider collections of data sets.

In this chapter we describe FuseNet, a Markov network formulation that infers networks from a collection of nonidentically distributed data sets. Our approach is computationally efficient and general: given any number of distributions from an exponential family, FuseNet represents model parameters through shared latent factors that define neighborhoods of network nodes. In a simulation study we demonstrate good predictive performance of FuseNet in comparison to several popular graphical models. We show its effectiveness in an application to breast cancer RNA-sequencing and somatic mutation data, a novel application of graphical models. Fusion of data sets offers substantial gains relative to inference of separate networks for each data set. Our results demonstrate that network inference methods for non-Gaussian data can help in accurate modeling of the data generated by emergent high-throughput technologies.

5.1 Background

Undirected graphical models or Markov networks are a popular class of statistical tools for probabilistic description of complex associations in high-dimensional data (cf. Rue and Held, 2005). Biological processes in a cell involve complex interactions between genes and it is important to understand, which genes conditionally depend on each other. These dependencies can be inferred from the experimental data and represented in a gene network. As a popular approach to network modeling, Markov networks are particularly appealing because they focus on finding such conditional dependence relationships. Intuitively, the existence of a link between genes A and B in a Markov network indicates that the behavior of gene A is still predictive of gene B given all available measurements about gene A and its immediate neighbors in a network. Hence, Markov networks can help us to find a rich set of direct dependencies between genes that are stronger than gene correlations (Allen and Liu, 2013).

Markov networks have been well studied in bioinformatics and numerous applications are concerned with inferring the network structure primarily from microarray and next generation sequencing gene expression data (Segal et al., 2003; Kotera et al., 2012; Gallopin et al., 2013). They are complementary but not superior to other gene network inference approaches (Marbach et al., 2012). However, the increasing variety of data generating technologies and heterogeneity of resulting data draw attention to two challenges in the context of Markov network inference: inference from non-Gaussian distributed data, and simultaneous inference from many data sets.

In bioinformatics, many data sets are high dimensional, contain a limited number of samples with a large number of zeros, and come from skewed distributions. Most existing methods assume that data follow a Gaussian distribution. While this assumption holds for typical log ratio expression values from microarray data, it is violated for measurements obtained from sequencing technologies. For example, gene expression levels from RNA-sequencing count how many times a transcript maps to a specific genomic location (Wang et al., 2009) and as such these data are not Gaussian (Allen and Liu, 2013). The Gaussian assumption is also violated for categorical data sets, such as data on mutation types and copy number variation data (Hudson et al., 2010). While it would be possible to design a network inference for each specific data type, we could benefit from a procedure that can treat a wide class of distributions and can jointly consider all available data during network inference (Žitnik and Zupan, 2015a).

We have developed a novel approach, called FuseNet, for inference of undirected networks from a number of high-dimensional data sets (Fig. 5.1). Our approach builds upon recent theoretical results about Markov networks (Yang et al., 2012, 2013) and, unlike the previous works in Markov modeling, can be applied to settings where data arise from *multiple* related but otherwise *nonidentical distributions*. To achieve this level of modeling flexibility, we represent model parameters with *latent factors*. FuseNet implements data fusion through sharing of latent factors that are common to all data sets and distributions, and handles data diversity through inference of factors specific to a particular data set.

In simulation studies FuseNet recovers the true networks underlying the observed data more accurately than several alternative approaches. The improved performance

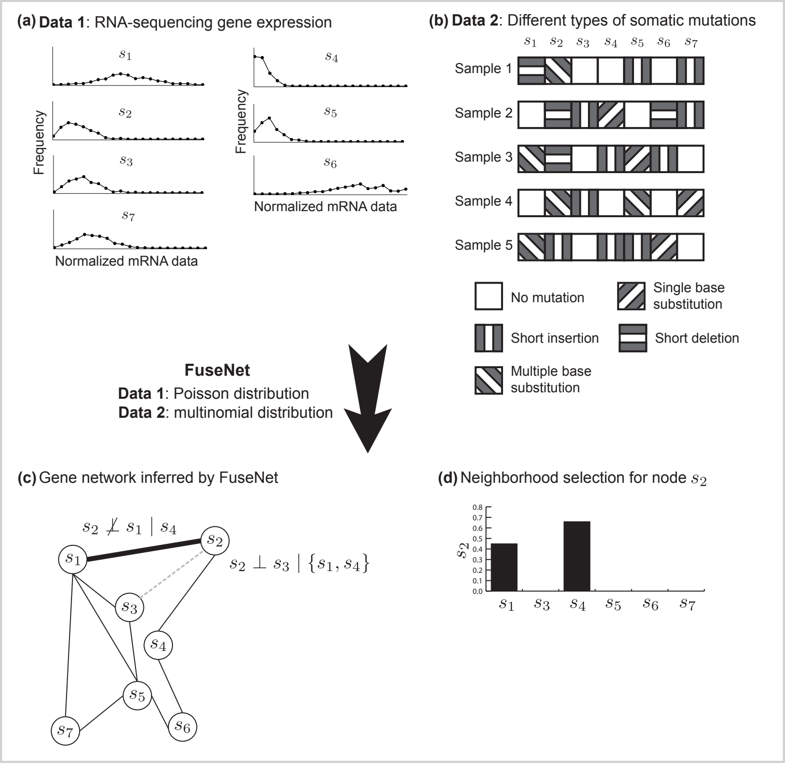

(a) Data 1: RNA-sequencing gene expression
s_1
s_2
s_3
s_7
s_4
s_5
s_6
Frequency
Frequency
Normalized mRNA data
Normalized mRNA data
(b) Data 2: Different types of somatic mutations
s_1 s_2 s_3 s_4 s_5 s_6 s_7
Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
No mutation
Single base substitution
Short insertion
Short deletion
Multiple base substitution
FuseNet
Data 1: Poisson distribution
Data 2: multinomial distribution
(c) Gene network inferred by FuseNet
$s_2 \not\perp s_1 \mid s_4$
$s_2 \perp s_3 \mid \{s_1, s_4\}$
s_1
s_2
s_3
s_4
s_5
s_6
s_7
(d) Neighborhood selection for node s_2
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
s_2
s_1 s_3 s_4 s_5 s_6 s_7

Figure 5.1

An overview of FuseNet in a toy application to network inference. FuseNet's input is a collection of data sets that can follow different exponential family distributions. The example from the figure uses two data sets: (a) gene expressions from next-generation sequencing follow the Poisson distribution, and (b) somatic mutation data follow the multinomial distribution. (c) FuseNet infers a network by collectively modeling dependencies between any two genes conditioned on the rest of the genes. The absence of an edge between s_2 and s_3 (dotted line in grey) implies that s_2 acts independently of s_3 given s_1 and s_4, the neighbors of s_2. The $\perp$ symbol stands for conditional independence. Genes s_1 and s_2 are linked because data profiles of s_2 in (a-b) are still predictive of the profile values of s_1 given s_4, the neighbor of s_2. (d) Shown are FuseNet-inferred coefficients that relate s_2 to all other genes. Non-zero values indicate gene dependency. In the resulting network, gene s_2 has two neighbors, s_1 and s_4.

demonstrates that FuseNet can find conditional dependencies between genes that could not be reconstructed with Gaussian-based approaches. In a case study with breast cancer RNA-sequencing expression values and somatic mutation data, we demonstrate the benefits of joint network inference from multiple related data sets. The networks inferred collectively from both types of data show greater functional enrichment than networks learned from any data type alone.

5.2 Related work on gene network inference

The most straightforward approach to network inference is a similarity-based approach, which assumes that functionally related genes are likely to share high similarity with respect to a given data set. A well known network obtained with this approach is the *S. cerevisiae* genetic interaction network by Costanzo et al. (2010). Whenever the similarity value between two genes is above a threshold they are linked by an edge, which is referred to as a *direct network inference* approach (Kotera et al., 2012). In contrast to direct network inference, *model-based network inference* via graphical models focuses on local dependencies between genes, where each gene is directly affected by a relatively small number of genes. Edges estimated by a graphical model can be related to causal inference (Pearl and Verma, 1991).

The problem of learning a network structure associated with an undirected graphical model has seen a wide range of applications ranging from social networks and image and speech processing (Metzler and Croft, 2005; Wang et al., 2013) to genomics. Applications in bioinformatics include estimation of molecular pathways from protein interaction and gene expression data (Segal et al., 2003; Stingo and Vannucci, 2011), reconstruction of gene regulatory networks from microarray data (Marbach et al., 2012), inference of a cancer signaling network from proteomic data (Mukherjee and Speed, 2008) and reconstruction of genetic interaction networks from integrated experimental data (Isci et al., 2014). Methods applied to these problems and many other recent gene network inference algorithms (Schäfer and Strimmer, 2005; Meinshausen and Bühlmann, 2006; Friedman et al., 2008; Anjum et al., 2009; Ravikumar et al., 2010) estimate Gaussian or binary Markov networks, *i.e.*, they assume that data follow an approximately Gaussian distribution.

Although non-Gaussian data are becoming increasingly common in biology, until now, very few network inference algorithms have been proposed for their treatment. When dealing with non-Gaussian data, some authors simply use methods that are based on a Gaussian assumption (Cai et al., 2012). We show in experiments that this decision may result in poor predictive performance. Recently, various extensions of Gaussian Markov networks have been proposed that first Gaussianize the data, using for example a copula transform (Liu et al., 2009, 2012; Murray et al., 2013) or a log transform, and then apply algorithms that rely on an assumption of normality. While these approaches perform better than naïve application of Gaussian-based methods to untransformed data, they are ill-suited to data generated by next generation sequencing technologies (Allen and Liu, 2013). A handful of recent algorithms (Allen and Liu, 2013; Gallopin et al., 2013) have considered Markov networks for non-Gaussian data, using for example the Poisson distribution for RNA-sequencing read counts. In contrast to our FuseNet, these methods can not integrate data sets across different data types, thereby limiting their ability to fuse information from many data sets.

5.3 Gene network inference by fusing data from diverse distributions

FuseNet takes as its input a collection of data sets where each data set consists of a set of gene profiles (Fig. 5.1). Gene profiles can be heterogeneous and belong to different data types, *e.g.*, data can be continuous, discrete or categorical. For example, measurements from RNA-sequencing represent the numbers of fragments that were mapped to a specific genomic location (Wang et al., 2009). The RNA-sequencing expression values are then non-negative and integer valued and, hence, are not approximately Gaussian, but rather follow the Poisson or negative binomial distribution. This is in contrast to copy number variation data and mutation data, *i.e.*, single base substitutions, short indels, or multiple base substitutions, that might be modeled better with multinomial or categorical distributions. On the other end of spectrum are microarray gene expression data, which are approximately Gaussian distributed.

The crucial feature of FuseNet is the *representation of model parameters via latent factors*. This feature, together with the *sharing of latent factors between data sets*, allows us to infer a network by simultaneously considering many data sets that each can arise from a different exponential family distribution (Sec. 5.3.7).

We exemplify FuseNet by deriving Markov network models for two distributions from an exponential family, the Poisson distribution (Sec. 5.3.3) and the multinomial distribution (Sec. 5.3.5). Since the exponential family includes not only Gaussian but also binomial, multinomial, Poisson, gamma distributions and others, FuseNet can achieve great flexibility in estimating gene networks from diverse data (Sec. 5.3.6) and also comes with an efficient algorithm for network structure estimation (Sec. 5.4).

Our work provides two novel contributions over current approaches to gene network inference discussed in Related work:

- FuseNet *simultaneously* infers networks from data sets that may be generated by nonidentical distributions, and
- FuseNet estimates large-scale genomic networks from increasingly common *non-Gaussian distributed* data.

5.3.1 Preliminaries

Markov networks

A Markov network specifies conditional dependence relationships between genes. In particular, if there is no edge between genes s and t then this implies that the behavior of s is independent of t given the set of immediate neighbors of s. From this local property (Murphy, 2012), one can easily see that two genes (nodes) are conditionally independent given the rest of the genes iff there is no direct edge between them. The conditional independence (Markov) properties permit a rich set of dependencies among the nodes and hence, the connectivity of a Markov network can reveal complex relationships between its nodes (Jalali et al., 2011; Allen and Liu, 2013).

Exponential family

The probability distributions that we study in this chapter are specific examples of a broad class of distributions called the exponential family (Duda and Hart, 1973). Members of the exponential family have many important properties in common. Given parameters θ, the exponential family of distributions over X is defined to be the set of distributions of the form:

$$P(X) = \exp(\theta B(X) + C(X) - D(\theta)), \tag{5.1}$$

where $B(X)$ are sufficient statistics, $C(X)$ is a base measure and $D(\theta)$ is a log-normalization constant (Murphy, 2012). The exponential family includes many widely used distributions, such as Bernoulli, binomial, Poisson, gamma, multinomial and Gaussian distributions.

Parameterization of Markov networks

Let $\boldsymbol{X} = (X_1, X_2, \dots, X_p)$ be a random vector with X_i being a random variable. Suppose $\boldsymbol{G} = (V, E)$ is an undirected graph with p nodes representing p variables in $\boldsymbol{X}$, $|V| = p$. Then the corresponding undirected graphical model is any distribution defined on $\boldsymbol{X}$ that satisfies Markov independence assumptions with respect to graph $\boldsymbol{G}$ (Murphy, 2012). By the Hammersley-Clifford theorem (Murphy, 2012), any such distribution of $\boldsymbol{X}$ decomposes according to graph $\boldsymbol{G}$ in the following way. Let $\mathscr{C}$ be a set of maximal cliques (fully-connected subgraphs) in graph $\boldsymbol{G}$ and let $\{\phi_c(X_c), c \in \mathscr{C}\}$ be "clique potential" functions. By the Hammersley-Clifford theorem, any distribution of X within the graphical model family defined by $\boldsymbol{G}$ can be represented as an exponential of a weighted sum of potential functions over the maximal cliques $\mathscr{C}$:

$$P(X) \propto \exp(\sum_{c \in \mathscr{C}} \theta_c \phi_c(X_c)), \tag{5.2}$$

where $\{\theta_c, c \in \mathscr{C}\}$ are the weights of potential functions.

An important question is how one would select potential functions $\{\phi_c, c \in \mathscr{C}\}$ to obtain various multivariate extensions of univariate distributions. Recently, Yang et al. (2012) showed that if a node-conditional univariate distribution, *i.e.*, distribution of a random variable conditioned on all other variables, belongs to an *exponential family*, it *necessarily* follows that the joint distribution of $\boldsymbol{X}$ has the form:

$$P(\boldsymbol{X}) \propto \exp(\sum_{s \in V} \theta_s B(X_s) + \sum_{s \in V} \sum_{t \in \mathscr{N}(s)} \theta_{st} B(X_s) B(X_t) + \tag{5.3}$$

$$\sum_{s \in V} \sum_{t_2, \dots, t_k \in \mathscr{N}(s)} \theta_{s, t_2, \dots, t_k} B(X_s) \prod_{j=2}^{k} B(X_{t_j}) + \sum_{s \in V} C(X_s)),$$

where the cliques are of size at most k, $\mathscr{N}(s)$ are neighbors of node s, $\boldsymbol{B}$ represent sufficient statistics and $\boldsymbol{C}$ is the base measure of the a given exponential family distribution (cf. Proposition 1 and Proposition 2 in Yang et al. (2012)). These results tell us that the joint distribution specified in Eq. (5.3) has *the most general form under the assumption of exponential family node-conditional distributions*. Hence, learning a graphical model from the data can be reduced to learning weights $\{\theta_s\} \cup \{\theta_{st}\} \cup \dots \cup \{\theta_{s,t_2,\dots,t_k}\}$ of distribution-specific sufficient statistics.

5.3.2 *Problem definition*

Suppose we are given a collection $\mathscr{D}$ of n observations, $\mathscr{D} = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(n)}\}$, where $\boldsymbol{x}^{(i)}$ is a p-dimensional vector drawn i.i.d. from a specific distribution of the form in Eq. (5.3). This distribution has parameters $\{\theta_c^*, c \in \mathscr{C}\}$ and is associated with a graph $\boldsymbol{G} = (V, \boldsymbol{E}^*)$ on p nodes. Graph $\boldsymbol{G}$ encodes Markov independence properties between the respective variables. The goal of learning the structure of $\boldsymbol{G}$ is to infer an edge set $\boldsymbol{E}^*$ that corresponds to distribution, which generated observations in $\mathscr{D}$. We can express $\boldsymbol{E}^*$ as a function of parameters $\{\theta_c^*, c \in \mathscr{C}\}$ and write it as:

$$E^* = \{(s,t) \in V \times V \,:\, \exists \text{ clique } c \in \mathscr{C} \,:\, \{s,t\} \subseteq c \wedge \theta_c^* \neq 0\}.$$

Hence, learning the network structure reduces to the problem of estimating weights $\{\hat{\theta}_c, c \in \mathscr{C}\}$ that should be as close as possible to the true but otherwise unknown parameters $\{\theta_c^*, c \in \mathscr{C}\}$.

In this chapter, we focus largely on a special case of *pairwise* Markov networks, where the joint distribution has cliques of size at most two:

$$P(\boldsymbol{X}) \propto \exp(\underbrace{\sum_{s \in V} \theta_s^* B(X_s)}_{\text{set of nodes}} + \underbrace{\sum_{(s,t) \in V \times V} \theta_{st}^* B(X_s) B(X_t)}_{\text{set of edges}} + \sum_{s \in V} C(X_s)) \tag{5.4}$$

with entries $\theta_{st}^* \neq 0$ if $t \in \mathscr{N}(s)$ and $\theta_{st}^* = 0$ if $t \notin \mathscr{N}(s)$. Following the work of Ravikumar et al. (2010), Jalali et al. (2011) and Allen and Liu (2013) we approach the problem of Markov network structure learning via neighborhood estimation, where we obtain the global network estimate $\hat{\boldsymbol{E}}$ by stitching together the estimated neighborhoods of the nodes. The overall network structure is then:

$$\hat{E} = \bigcup_{s \in V, t \in \widehat{\mathscr{N}}(s)} \{(s,t)\}, \tag{5.5}$$

where (s,t) denotes an edge between s and t and $\widehat{\mathscr{N}}(s) = \{t \in V \setminus \{s\} \,:\, \hat{\theta}_{st} \neq 0\}$ is the estimated neighborhood of node s.

In the remainder of this section we formulate two pairwise Markov networks, which assume either Poisson or multinomial data distribution. These two exponential family models are taken as an example through which we specify a general scheme for network inference from multiple potentially non-identical data distributions.

5.3.3 Poisson model specification

Following the work of Yang et al. (2012) and Allen and Liu (2013) we define a Poisson Markov network model by specifying a distribution where all node-conditional distributions follow a univariate Poisson distribution. Our Poisson Markov network model is then a series of locally defined models, one for every variable (node). A local model for s is given by a distribution of X_s conditioned on all other variables:

$$P(X_s|X_{V\setminus s}) \sim \text{Poisson}(\exp\{\boldsymbol{u}_s + \sum_{t\in V\setminus\{s\}} \boldsymbol{u}_s^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{u}_t X_t\}), \tag{5.6}$$

where $X_{V\setminus s} = \{X_t | t \in V \setminus \{s\}\}$ denotes the rest of the variables, and $\boldsymbol{u}_s \in \mathbb{R}^r$ and $\boldsymbol{W} \in \mathbb{R}^{r\times r}$ are model parameters. An $\boldsymbol{r}$-dimensional vector $\boldsymbol{u}_s$ is a latent factor for node s that consists of $\boldsymbol{r}$ latent components. For now, we assume that the number of latent components $\boldsymbol{r}$ is given; we will later discuss how to automatically determine $\boldsymbol{r}$. Notice that the latent factor of node s, $\boldsymbol{u}_s$, represents the strength of membership of node s to $\boldsymbol{r}$ latent components and $\boldsymbol{W}$ models the interactions between all combinations of $\boldsymbol{r}$ latent components. The formulation of the Poisson conditional distribution in Eq. (5.6) ensures that node pair-wise weights are symmetric, which is an appealing property when studying undirected graphical models. In particular, the contribution of X_t towards $P(X_s|X_{V\setminus s})$ is the same as is the contribution of X_s towards $P(X_t|X_{V\setminus t})$.

We refer to our model as a model *parameterized via latent factorization*, since model parameters $\boldsymbol{u}_s$, $\boldsymbol{u}_t$ and $\boldsymbol{W}$ form a factorization of the edge weight θ_{st}, which is specified by a Markov network model in Eq. (5.4). The importance of latent factor parameterization will be obvious later in Sec. 5.3.7 when we discuss collective network inference from many data sets.

Recall the univariate Poisson distribution is given by the mass function $P(X = x) = \lambda^x \exp(-\lambda)/x!$, where λ is the shape parameter. Our model extends the univariate Poisson in a natural and strict sense to the multivariate graphical model setting. The latter can be obtained from the univariate Poisson by setting the shape parameter to $\lambda = \exp(\boldsymbol{u}_s + \sum_{t\in V\setminus s} \boldsymbol{u}_s^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{u}_t X_t)$. We then write the expression in Eq. (5.6) as:

$$\begin{aligned} P(X_s|X_{V\setminus s}) = \exp\{\boldsymbol{u}_s X_s - \log(X_s!) + \sum_{t\in V\setminus\{s\}} (\boldsymbol{u}_s^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{u}_t X_s X_t \\ - \exp(\boldsymbol{u}_s + \boldsymbol{u}_s^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{u}_t X_t))\} \end{aligned} \tag{5.7}$$

Intuitively, variable X_s in Eq. (5.7) can be viewed as the response variable in a latent factor Poisson regression in which the other variables $X_{V\setminus s}$ play the role of the predictors. Variables with strong relationships with gene s will have non-zero regression coefficients, and these will be connected to node s in the inferred graph.

5.3.4 Optimization of the Poisson model

The node-conditional distributions specified in Eq. (5.7) define a global distribution that factors according to the cliques of the underlying graph $\boldsymbol{G}$ that we would like to estimate. We obtain edge set $\widehat{\boldsymbol{E}}$ by stitching node neighborhoods as prescribed by Eq. (5.5), where we define the neighborhood of node s as $\widehat{\mathcal{N}}(s) = \{t \in V \setminus \{s\} : \boldsymbol{u}_s^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{u}_t \neq 0\}$. This means that edge (s,t) is included in the network if the estimated product of respective latent factors of variables X_s and X_t is non-zero.

To estimate edge set $\widehat{\boldsymbol{E}}$ we have to determine the node neighborhoods of all nodes in V. To achieve this goal, we solve a sparsity constrained conditional maximum likelihood estimation problem:

$$\min_{U,W} \sum_{s\in V} \ell_s(\boldsymbol{U}, \boldsymbol{W}; \mathcal{D}) + \alpha(\mathrm{Reg}(\boldsymbol{U}) + \mathrm{Reg}(\boldsymbol{W})). \tag{5.8}$$

Here, $\boldsymbol{U}$ is a matrix with node latent factors placed in the columns, $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_n]$.

Eq. (5.8) consists of two parts, which we discuss next. Terms involving Reg represent the elastic net penalties (Zou and Hastie, 2005). The penalty is defined for $\boldsymbol{U}$ as $\mathrm{Reg}(\boldsymbol{U}) = (1-\lambda)\frac{1}{2}\|\boldsymbol{U}\|_{2,1}^2 + \lambda\|\boldsymbol{U}\|_{1,1}$, where $\lambda \geq 0$ is a regularization parameter controlling the amount of sparsity in the node neighborhood. The definition of the penalty term for $\boldsymbol{W}$ is analogous. Notice that the $L_{2,1}$ norm is the sum of 2-norms of the columns, $\|\boldsymbol{U}\|_{2,1} = \sum_{s=1}^{p} \|\boldsymbol{u}_s\|_2^2$, and the $L_{1,1}$ norm is the sum of 1-norms of the columns, $\|\boldsymbol{U}\|_{1,1} = \sum_{s=1}^{p} \|\boldsymbol{u}_s\|_1$. Since latent factors are affected by the strength of regularization, the choice of parameter λ is important. Procedure for selection of λ is described in Sec. 5.4.

The crucial part of Eq. (5.8) is, however, the sum of the node-wise Poisson likelihood functions. Given node s and n realizations of the associated random variable X_s, the

Poisson likelihood function ℓ_s follows directly from Eq. (5.7) and can be written as:

$$\begin{aligned}\ell_s(\boldsymbol{U}, \boldsymbol{W}; \mathscr{D}) &= -\frac{1}{n} \log \prod_{i=1}^{n} P(X_s = \boldsymbol{x}_s^{(i)} | X_{V \setminus s} = \boldsymbol{X}_{\setminus s}^{(i)}) \\ &= -\frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_s^{(i)} \boldsymbol{X}_{\setminus s}^{(i)} \boldsymbol{U}_{\setminus s}^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{u}_s \\ &\quad - \exp(\boldsymbol{X}_{\setminus s}^{(i)} \boldsymbol{U}_{\setminus s}^T \boldsymbol{W}^T \boldsymbol{W})),\end{aligned} \tag{5.9}$$

where $\boldsymbol{x}_s^{(i)}$ is the i-th realization of X_s in data $\mathscr{D}$, $\boldsymbol{X}_{\setminus s}^{(i)}$ denotes the i-th realization of the rest of the variables $X_{V \setminus s}$, and $\boldsymbol{U}$ and $\boldsymbol{W}$ are matrix unknowns. Notice that node-wise terms are ignored here for simplicity.

5.3.5 *Multinomial model specification and optimization*

We now develop a multinomial Markov network model that relies on latent factor parameterization of the model parameters and follows the same paradigm as our Poisson model described in the previous section. The multinomial model presented here is a natural extension of the multinomial graphical model described by Jalali et al. (2011).

We start with the neighborhood recovery of one fixed node s and then combine the neighborhood sets across nodes to estimate the network. The multinomial model assumes that each variable X_i from a random vector $\boldsymbol{X}$ follows a multinomial distribution with potentially different parameters. This means that X_i can take any value from a small discrete set $\{1, 2, \dots, m\}$ of cardinality m. Probabilities of different values are not independent so that, given any $m - 1$ of the probabilities, the probability of the remaining value is fixed. It is convenient to express the distribution in terms of only $m - 1$ values, thereby leaving $m - 1$ probability parameters that need to be estimated.

The distribution of X_s conditioned on other variables $X_{V \setminus s} = \{X_t : t \in V \setminus \{s\}\}$ is given by:

$$P(X_s = j | X_{V \setminus s}) = \frac{\exp(\theta_{sj} + \sum_{t \in V \setminus \{s\}} \sum_k \theta_{st;jk} \mathscr{I}_k(X_t))}{1 + \sum_l \exp(\theta_{sl} + \sum_{t \in V \setminus \{s\}} \sum_k \theta_{st;lk} \mathscr{I}_k(X_t))} \tag{5.10}$$

for all $j \in \{1, 2, \dots m - 1\}$. Here, θ_{sj} represents a node-wise term that models the probability of variable X_s taking value j. The other model parameter is $\theta_{st;jk}$, which models dependency between variable X_s and variable X_t when they take values j and k, respectively. We can view Eq. (5.10) as a multiclass logistic (softmax) regression, where X_s is the response variable and indicator functions associated with other variables:

$$\{\mathcal{I}_k(X_t), t \in V \setminus \{s\}, k \in \{1, 2, \dots, m-1\}\},$$

where $\mathcal{I}_k(X_t) = 1$ if $X_t = k$ else 0, are the predictors.

We now proceed by writing model parameters θ_{sj} and $\theta_{st;jk}$ in the form of a product of latent factors. We gather node-wise terms θ_{sj} into a matrix $\boldsymbol{Q} \in \boldsymbol{R}^{p \times (m-1)}$. We factorize $\theta_{st;jk}$ as $\theta_{st;jk} = \boldsymbol{u}_s^T \boldsymbol{Q}_{sj} \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{Q}_{tk} \boldsymbol{u}_t$. Here, $\boldsymbol{u}_s$ and $\boldsymbol{u}_t$ are $\boldsymbol{r}$-dimensional latent factors and $\boldsymbol{W} \in \mathbb{R}^{r \times r}$ encodes interactions between latent components in the same way as is described in Sec. 5.3.3.

To estimate the latent factors and node-wise terms from the data we solve the following convex optimization program:

$$\min_{U,Q,W} \sum_{s \in V} \ell_s(\boldsymbol{U}, \boldsymbol{Q}, \boldsymbol{W}; \mathcal{D}) + \alpha(\mathrm{Reg}(\boldsymbol{U}) + \mathrm{Reg}(\boldsymbol{Q}) + \mathrm{Reg}(\boldsymbol{W})), \tag{5.11}$$

where definitions of $\boldsymbol{U}$, $\boldsymbol{W}$ and Reg are the same is in the previous section. Here, the node-wise multinomial likelihood function ℓ_s for node s follows from Eq. (5.10) and can be written as:

$$\begin{aligned} \ell_s(\boldsymbol{U},\boldsymbol{Q},\boldsymbol{W};\mathcal{D}) = -\frac{1}{n} \log \prod_{i=1}^{n} P(X_s = \boldsymbol{x}_s^{(i)} | X_{V \setminus s} = \boldsymbol{X}_{\setminus s}^{(i)}) = \\ -\frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{Q}_{sx_s^{(i)}} + \sum_{t \in V \setminus \{s\}} \sum_{k} \boldsymbol{u}_s^T \boldsymbol{Q}_{sx_s^{(i)}} \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{Q}_{tk} \boldsymbol{u}_t \mathcal{I}_k(\boldsymbol{x}_t^{(i)}) - \\ \log(1 + \sum_{l} \exp(\boldsymbol{Q}_{sl} + \sum_{t \in V \setminus \{s\}} \sum_{k} \boldsymbol{u}_s^T \boldsymbol{Q}_{sl} \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{Q}_{tk} \boldsymbol{u}_t \mathcal{I}_k(\boldsymbol{x}_t^{(i)})))), \end{aligned} \tag{5.12}$$

where $\boldsymbol{x}_s^{(i)} \in \{1, 2, \dots, m-1\}$ is the i-th realization of X_s in data $\mathcal{D}$, $\boldsymbol{X}_{\setminus s}^{(i)}$ denotes the i-th realization of the rest of the variables $X_{V \setminus s}$, and $\boldsymbol{U}$, $\boldsymbol{Q}$ and $\boldsymbol{W}$ are matrix unknowns. Given latent factor estimates $\boldsymbol{U}$ and $\boldsymbol{W}$, and the estimate of node-wise terms $\boldsymbol{Q}$, we determine the neighborhood for node s as $\widehat{\mathcal{N}}(s) = \{t \in V \setminus \{s\} : \sum_{j,k} \boldsymbol{u}_s^T \boldsymbol{Q}_{sj} \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{Q}_{tk} \boldsymbol{u}_t \neq 0\}$. This means that edge (s, t) is included in the network if product $\boldsymbol{u}_s^T \boldsymbol{Q}_{sj} \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{Q}_{tk} \boldsymbol{u}_t$ does not vanish over at least one choice of categories j and k.

5.3.6 Other exponential family distributions

So far, we described in Sec. 5.3.3–5.3.5 the Poisson model and the multinomial model that are suitable for separately inferring the edge set of a Poisson or a multinomial Markov network. In this section we would like to allude to the fact that a procedure with derivations very similar to those in the above sections can be applied to any exponential family distribution.

From Eq. (5.1) we see that the unnormalized probability of an exponential family distribution can be expressed as an exponential of a weighted linear combination of sufficient statistics. These sufficient statistics correspond to clique potential functions (see Sec. 5.3.1). Under the assumption of joint distribution having cliques of size at most two, node-conditional distributions take the form:

$$P(X_s|X_{V\setminus s}) \propto \exp(\theta_s B(X_s) + \sum_{t\in\mathcal{N}(s)} \theta_{st} B(X_s)B(X_t) + C(X_s))$$

where $\{\theta_s, s \in V\}$ and $\{\theta_{st}, s, t \in V\}$ are parameters that shall be estimated from the data.

FuseNet yields a general framework for including data from any exponential family distribution, such as Gaussian, binomial, Poisson or multinomial distributions, in its predictive model by simply expressing weights $\{\theta_s, s \in V\}$ and $\{\theta_{st}, s, t \in V\}$ of a given distribution as products of *appropriately* selected latent factors. Here, factorization of the weights is *appropriate* if it allows fusion of data from diverse distributions, such that factorization consists of both latent factors that are shared between different distributions and factors that are specific to a particular distribution (Žitnik and Zupan, 2015a), a property that we describe in the following section.

5.3.7 Collective inference of a gene network

We proceed by formulating a collective network inference model, wherein a network is jointly estimated from multiple nonidentical data distributions.

Let $\mathcal{D}_x = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(n_x)}\}$ be a set of n_x observations of a random vector X, where each p-dimensional vector $\boldsymbol{x}^{(i)}$ is drawn from a distribution P_x of the form of

Eq. (5.4) and let $\mathscr{D}_y = \{\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}, \ldots, \boldsymbol{y}^{(n_y)}\}$ be a set of n_y observations where each p-dimensional vector $\boldsymbol{y}^{(i)}$ is drawn from distribution $\boldsymbol{P}_y$ of the form of Eq. (5.4). Importantly, distributions $\boldsymbol{P}_x$ and $\boldsymbol{P}_y$ are not necessarily identical in terms of their parameters or distribution type. For example, $\boldsymbol{P}_x$ might denote the Poisson distribution and $\boldsymbol{P}_y$ might be the multinomial distribution or they could both describe multinomial distributions that have different parameters. For simplicity of notation we provide here the formulation for the case with only two data sets, $\mathscr{D}_x$ and $\mathscr{D}_y$, but notice that our analysis generalizes to any number of data sets.

In collective network inference we solve for:

$$\begin{aligned} \min_{\substack{U, Q_x, Q_y,\\ W_x, W_y}} \sum_{s \in V} (&\ell_{s;P_x}(\boldsymbol{U}, \boldsymbol{Q}_x, \boldsymbol{W}_x; \mathscr{D}_x) \\ &+ \ell_{s;P_y}(\boldsymbol{U}, \boldsymbol{Q}_y, \boldsymbol{W}_y; \mathscr{D}_y)) + \text{reg. param.}, \end{aligned} \tag{5.13}$$

where regularization parameters depend on the form of data distributions. In a specific scenario in which $\boldsymbol{P}_x$ and $\boldsymbol{P}_y$ are the Poisson and the multinomial distributions, respectively, we set $\boldsymbol{Q}_x = \boldsymbol{I}$. We specify the regularization according to the Poisson model in Eq. (5.8) and the multinomial model in Eq. (5.11) as:

$$\lambda(\text{Reg}(\boldsymbol{U}) + \text{Reg}(\boldsymbol{W}_x) + \text{Reg}(\boldsymbol{Q}_y) + \text{Reg}(\boldsymbol{W}_y)),$$

where Reg is the elastic net penalty defined in Sec. 5.3.3. The estimated neighborhood of node s, which corresponds to a random variable $X_s \in X$, are then nodes whose behavior depends on behavior of s according to any of considered data distributions, $\widehat{\mathcal{N}} = \{t \in V \setminus \{s\} : \hat{\theta}_{st;P_x} \neq 0 \vee \hat{\theta}_{st;P_y} \neq 0\}$. In our specific scenario, parameters $\hat{\theta}_{st;P_x}$ and $\hat{\theta}_{st;P_y}$ would be given by $\hat{\theta}_{st;P_x} = \boldsymbol{u}_s^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{u}_t$ and $\hat{\theta}_{st;P_y} = \sum_{j,k} \boldsymbol{u}_s^T Q_{sj} \boldsymbol{W}^T \boldsymbol{W} Q_{tk} \boldsymbol{u}_t$.

It is important to notice the coupling of the parameters in FuseNet through which data fusion is achieved (Žitnik and Zupan, 2015a). As is evident from Eq. (5.13), the latent factor of node s, $\boldsymbol{u}_s$, participates both in terms associated with $\boldsymbol{P}_x$ and terms related to $\boldsymbol{P}_y$. Hence, a good estimate of $\boldsymbol{u}_s$ should simultaneously minimize both $\ell_{s;P_x}$ and $\ell_{s;P_y}$, but should do so in a way that statistics internal to both data distributions are considered. To account for the fact that data sets may disagree and differ in how accurately they capture biological signals, FuseNet has parameters that are specific to every distribution. In particular, we allow that interactions between latent components in $\mathscr{D}_x$ are different from those in $\mathscr{D}_y$ and hence, the model has one latent matrix $\boldsymbol{W}$ for each distribution. An additional parameter $\boldsymbol{Q}$ captures the characteristics of a particular exponential family distribution, *e.g.*, the bias associated with m categories in the multinomial distribution.

5.4 Learning the models in practice

Now that we defined the FuseNet model, we explain how to solve related optimization problems. Notice that exact optimization problem one needs to solve depends on a particular data setting, *i.e.*, the particular combination of exponential family distributions that generated a collection of data sets.

There has been a strong line of work on developing fast algorithms to solve sparse regression problems that are similar to Eq. (8) and Eq. (11) including the work by Krishnapuram et al. (2005), Meier et al. (2008), Jalali et al. (2011) and Allen and Liu (2013). Existing algorithms for undirected graphical model selection assume that model parameters are independent of each other. This, however, is not true in FuseNet due to reasons discussed in Sec. 5.3.7, which ensure data fusion. Consequently, this also means that we cannot use off-the-shelf optimization solvers.

5.4.1 Node neighborhood selection

We propose to fit our FuseNet by computing cyclical coordinate descent along the path of regularization parameter λ. Taking derivatives of Eq. (13) and with optimization techniques by Friedman et al. (2007a); Yuan (2008); Friedman et al. (2010) we can obtain solutions over a range of values for regularization parameter with approximately the same speed as fitting a model at a single value of λ. The technique uses current parameter estimates as warm restarts.

FuseNet employs elastic net penalties (Zou and Hastie, 2005) in their models. Elastic net is a compromise between the ridge penalty ($\lambda = 0$) and the lasso penalty ($\lambda = 1$) and is useful in situations where $p \gg n$ or when many variables are correlated. As λ increases from 0 to 1, for a given α the sparsity of the solution (*i.e.* the number of latent components equal to zero) increases monotonically from 0 to the sparsity of the lasso solution. In each iteration of the coordinate descent we apply soft thresholding to the current FuseNet estimates to care of the lasso contribution to the penalty, and then apply a proportional shrinkage for the ridge penalty (Meinshausen and Bühlmann, 2006; Friedman et al., 2007a; Simon et al., 2013).

5.4.2 Selecting regularization parameters

The choice of λ is critical since different λ's can lead to different network sparsity patterns, *i.e.* the number and position of edges in the inferred network. We estimate λ in data-dependent way via stability selection (Meinshausen and Bühlmann, 2010), a technique which was shown to lead to better results for the network inference than other parameter selection methods including cross validation, Akaike's information criterion and Bayesian information criterion (Liu et al., 2010; Yu et al., 2012).

For now, we assume that the number of latent components r is given. Here, we choose λ so as to use the least amount of regularization that simultaneously makes the network sparse and stable, *i.e.*, replicable under random sampling. FuseNet employs recently proposed stability selection technique called StARS (Liu et al., 2010). Briefly, StARS repeatedly sub-samples data $\mathscr{D}$ to obtain many data samples $\mathscr{D}_s$. Here, $\mathscr{D}_s$ denotes s-th data sample. It then estimates a separate network $\widehat{\boldsymbol{E}}_s(\lambda, r)$ for each $\mathscr{D}_s$ and each λ from a vector of regularization parameters λ; the latter being possible due to coordinate descent computed along a regularization path. Selected value for regularization controls the average variance over the edges of the networks inferred from sub-sampled data:

$$\lambda^{(r)}_{\text{opt}} = \arg\min_{\rho}\{\min_{0\leq\lambda\leq\rho}(\sum_{j<k} 2\bar{\boldsymbol{A}}_{jk}(\lambda, r)(1-\bar{\boldsymbol{A}}_{jk}(\lambda, r))/\binom{p}{2}) \leq \beta\}$$

where $\bar{\boldsymbol{A}}_{jk}(\lambda, r) = \frac{1}{S}\sum_{s=1}^{S}\mathscr{I}((j,k) \in \widehat{\boldsymbol{E}}_s(\lambda, r))$. We set β and the size of data samples $\mathscr{D}_s$ to the values recommended in Allen and Liu (2013). We note that we obtain different optimal values of $\lambda^{(r)}_{\text{opt}}$ for different choices of r. Next, we describe how we select r, which in effect determines the exact value of regularization.

5.4.3 Selecting the number of latent components

Our FuseNet has another parameter, the number of latent components r, which otherwise does not appear in current Markov models. The latent dimensionality is selected from a set of predefined candidate values $\{0.05n, 0.1n \dots, 0.5n\}$, where n is the mean number of observations across all considered data sets. We seek to use the fewest number of latent components that produce stable and sparse network:

$$r_{\text{opt}} = \arg\min_{\tau} \lambda_{\text{opt}}^{(\tau)}.$$

As a consequence, the optimal regularization value is $\lambda_{\text{opt}} = \lambda_{\text{opt}}^{(r_{\text{opt}})}$. Notice that the entire set of computations including path-wise coordinate descent and selection of regularization via stability selection can be performed in parallel for each candidate value of r.

Source code of FuseNet is available at `http://github.com/marinkaz/fusenet`.

5.5 Evaluating the quality of network inference

We compare the performance of FuseNet to several state-of-the-art Markov network models in estimating the true underlying network structure. The success of network recovery is evaluated by comparison to the gold standard networks, when they are available, and by functional enrichment of the inferred networks.

Assessing the accuracy of network recovery

Simulated data come with complete and unambiguous true underlying networks, hence we can assess the performance of the algorithms as follows. We report receiver operator curves (ROC) computed by varying the regularization parameter λ, precision-recall (PR) curves, and true and false positive rates for fixed λ as estimated via stability selection. The true positive rate is estimated as proportion of the edges found by a network inference algorithm that are also in the true network. The false positive rate represents proportion of the edges in the inferred network that are not present in the true network. An algorithm with a perfect performance achieves an area under the ROC curve of 1, precision of 1 and recall of 1, a true positive rate of 1 and a false positive rate of 0.

Quantifying the functional content of inferred networks

We employ two approaches to evaluate "functional correctness" of the networks inferred from cancer data.

First, we use SANTA (Cornish and Markowetz, 2014) to quantify the strength of association between sets of functionally related genes and the inferred network. The input to SANTA are a gene network and a gene set and the output is a score representing statistical significance of their association. We obtain gene sets from the Gene Ontology (GO) (Ashburner et al., 2000) and test only GO terms associated with between 20 and 100 network genes to ensure that the functional sets are not too thinly or thickly spread.

Second, we overlay the inferred network with gene information from the GO and for every GO term assess how community-like a subnetwork of genes that belong to a particular GO term is. Four different structural notions of network communities exist in networks and we report the values of their representative scoring functions (Yang and Leskovec, 2012). Given is the inferred network $G(V, \hat{E})$, where $p = |V|$. Let $T \subseteq V$ be genes that belong to a specific GO term and let p_T be their number, $p_T = |T|$. We also need m_T, which is the number of edges in G whose both endpoints are annotated with a given GO term, $m_T = |\{(s,t) \in \hat{E} : s \in T, t \in T\}|$, and c_T, which counts how many edges are on the boundary of set T, $c_T = |\{(s,t) \in \hat{E} : s \in T, t \notin T\}|$. We denote degree of gene s with $d(s)$. Scoring functions build on the intuition that communities are sets of genes with many connections between the members and few connections to the rest of the network. We consider the following four scoring functions:

- *triangle participation ratio (TPR)* is the fraction of genes in T that belong to a triad, $|\{s : s \in T, \{(t,u) : t,u \in T, (s,t) \in \hat{E}, (s,u) \in \hat{E}, (t,u) \in \hat{E}\} \neq \emptyset\}|/p_T$;
- *cut ratio* is the fraction of all possible edges in T that connect T to the remainder of the network, $\frac{c_T}{p_T(p-p_T)}$;

- *conductance* is the fraction of total edge volume that points outside the GO term T, $\frac{c_T}{2m_T+c_T}$;
- *flake-over-median-degree (flake-ODF)* is the fraction of genes in T with fewer edges linking inside than outside of T, $|\{s : s \in T, |\{(s,t) \in \widehat{E} : t \in T\}| < d(s)/2\}|/p_T$.

The functions take values from [0, 1] interval. To make the higher the better, we report $(1 - \text{Conductance})$, $(1 - \text{Cut ratio})$ and $(1 - \text{flake-ODF})$ for conductance, cut ratio and flake-ODF, respectively.

Considered gene network inference algorithms

We experiment with the Poisson FuseNet (Sec. 5.3.3), the multinomial FuseNet (Sec. 5.3.5) and FuseNet with fusion of Poisson and multinomial data distributions (Sec. 5.3.7). We compare our models to the Graphical Lasso (GLASSO) (Friedman et al., 2007b), which is a widely used Markov network model based on a Gaussian assumption. To see how FuseNet relates to techniques that perform data preprocessing we consider the GLASSO after applying a log transform to the data plus one (Gallopin et al., 2013) and the GLASSO with the nonparanormal Gaussian copula transformation (NPN-Copula) (Liu et al., 2009). We also compare FuseNet with two Markov network models that are designed for non-Gaussian distributed data: the Local Poisson Graphical Model (LPGM) (Allen and Liu, 2013), and the Multinomial Markov Network Model (Mult-GM) (Jalali et al., 2011). The crucial parameter of these methods is degree of regularization, which controls sparsity of the networks. We select the value for regularization via stability selection (see Sec. 5.4).

5.6 *Simulated multivariate and real genomic data*

Network inference algorithms are evaluated based on simulated data and large-scale cancer genomic data sets.

Multivariate data simulation

Four network structures are simulated: (1) the Erdős Rényi random network, where an edge between each pair of nodes is set with equal probability and independently of other edges; (2) a hub network, where each node is connected to one of three hub nodes; (3) a scale-free network, in which node degree distribution follows a power-law; and (4) a small-world network, in which most nodes are not neighbors of each other but most nodes can be reached from every other by a small number of hops.

In simulations involving the Poisson model we closely follow the approach described by Karlis (2003) and Allen and Liu (2013). We generate n independent observations with p nodes, $\mathscr{D} = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(n)}\}$, where $\boldsymbol{x}^{(i)}$ is a p-dimensional count data vector, $\boldsymbol{x}^{(i)} \in \{0, 1, \dots, \infty\}^p$. A matrix of observations $\boldsymbol{X} = [\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(n)}]^T$ is obtained from the model $\boldsymbol{X} = \boldsymbol{Y}\boldsymbol{B} + \boldsymbol{E}$. Here, $\boldsymbol{Y}$ is a $n \times (p + p(p-1)/2)$ matrix with each entry $\boldsymbol{Y}_{ij} \overset{iid}{\sim} \text{Poisson}(\lambda_{\text{true}})$ and $\boldsymbol{E}$ is a $n \times p$ matrix with $\boldsymbol{E} \overset{iid}{\sim} \text{Poisson}(\lambda_{\text{noise}})$. Let $\boldsymbol{A}^*$ denote the adjacency matrix of a given true network structure E^*. The adjacency matrix is encoded by matrix $\boldsymbol{B}$ as $\boldsymbol{B} = [\boldsymbol{I}_p; \boldsymbol{P} \odot (\mathbf{1}_p \text{tri}(\boldsymbol{A}^*)^T)]^T$. Here, $\boldsymbol{P}$ is a $p \times (p(p-1)/2)$ permutation matrix, $\odot$ represents the entry-wise product and $\text{tri}(\boldsymbol{A}^*)$ is the $(p(p-1)/2) \times 1$ vectorized upper triangular part of $\boldsymbol{A}^*$. As done by Allen and Liu (2013) we simulate data at two signal-to-noise ratio (SNR) levels. We set $\lambda_{\text{true}} = 1$ with $\lambda_{\text{noise}} = 0.5$ for the high SNR level and $\lambda_{\text{noise}} = 5$ for the low SNR level.

In simulations involving the multinomial model we fix the alphabet size to $m = 3$. For a given true network structure E^*, we pick the parameter set $\theta_{st;jk} \in \{\theta_{st;jk} : s,t \in V; (s,t) \in E^*; j,k \in \{1,2\}\}$ as follows. If $(s,t) \in E^*$ then each nonzero entry $\theta_{st;jk}$ for $j,k \in \{1,2\}$ is set to $\theta_{st;jk} \in [-0.5, 0.5]$ uniformly at random; there are $4 = (3-1)^2$ such entries. We then generate n observations to construct a data set according to the probability distribution corresponding to $\theta_{st;jk}$. We solve the problem in Eq. (12) and compare the inferred network $\widehat{E}$ with the true network E^*.

Cancer genomic data

We apply network inference algorithms to two examples of non-Gaussian high-throughput genomic data to learn (1) an mRNA expression network, (2) a somatic mutation network and (3) a collectively inferred gene network from both data types.

We download breast cancer (BRCA-US) gene expression data measured by next generation sequencing and breast cancer (BRCA-US) simple somatic mutation data from the International Cancer Genome Consortium (ICGC) (Hudson et al., 2010) portal (release 17). We follow the steps in Allen and Liu (2013) and process the data to be approximately Poisson as is shown in Fig. 5.2. Genes with little variation across samples, the bottom 50%, are filtered out, and the data is adjusted for possible overdispersion by transforming them via a power $\alpha \in (0, 1]$ where α is chosen to yield approximately Poisson data as assessed via Kolmogorov-Smirnov tests (Li et al., 2011). The power transformation has another advantage. When neighboring genes have extremely large counts, the exponential in Eq. (6) causes the conditional Poisson mean to become large. The transformation limits the extreme counts and subsequently improves the fit of the model. Data preprocessing results in a matrix with rows as the subjects ($n_{\text{exp}} = 1{,}012$) and columns as genes ($p_{\text{exp}} = 657$). These genes form the nodes of our Poisson breast cancer mRNA network.

Breast cancer simple somatic mutation data from the ICGC portal include single base substitutions, multiple base substitutions and short indels. Mutation data are converted into a matrix with rows as subjects ($n_{\text{mut}} = 954$) and columns as genes containing mutations or variations (25,834 genes). Each matrix entry is categorized into one of three groups based on the type of mutation: no mutation, single base substitution, insertion/deletion of < 200 base pairs. Differentially mutated genes, *i.e.* genes containing mutations relative to the corresponding normal sample data, are ordered by their percentage of mutations across all samples and the top $p = 500$ genes were used in our analysis. These genes form the nodes of our multinomial breast cancer somatic mutation network.

For the collectively inferred network, we consider both gene expression profiles and somatic mutation data provided by the ICGC assuming the Poisson model for the RNA-seq data and the multinomial model for the mutation data. The genes that form the nodes of this network are taken as the union of sets of genes from the respective gene expression and somatic mutation matrices ($p = |V_{\text{exp}} \cup V_{\text{mut}}|$). Mutational and expression profiles from both matrices are matched by the subjects.

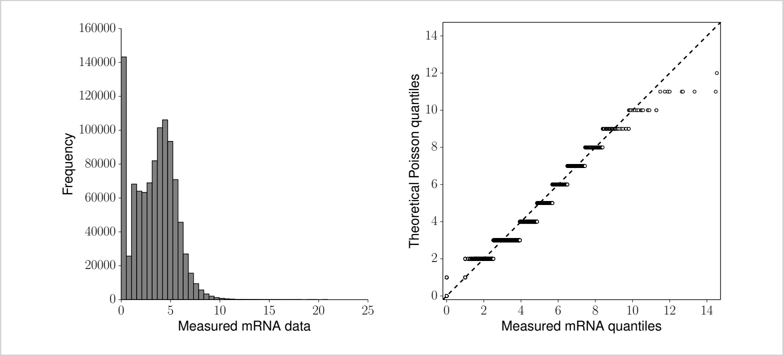


Figure 5.2

A histogram of the overall breast cancer RNA-seq data from the ICGC (Hudson et al., 2010) (left) and a comparison of these data to the quantiles of the Poisson distribution via a q-q plot (right). A q-q plot shows that breast cancer RNA-seq data approximately follow the Poisson distribution. The multivariate count data arising from the measurements of gene expression with the next generation sequencing technology is only an example of recent high-throughput technologies that produce non-Gaussian distributed data.

5.7 A case study: simulated and cancer genomic networks

Next, we evaluate the ability of FuseNet to recover networks from simulated data following various exponential family distributions. We also compare FuseNet to several gene network inference methods on cancer genomic data.

5.7.1 Network recovery with simulated data

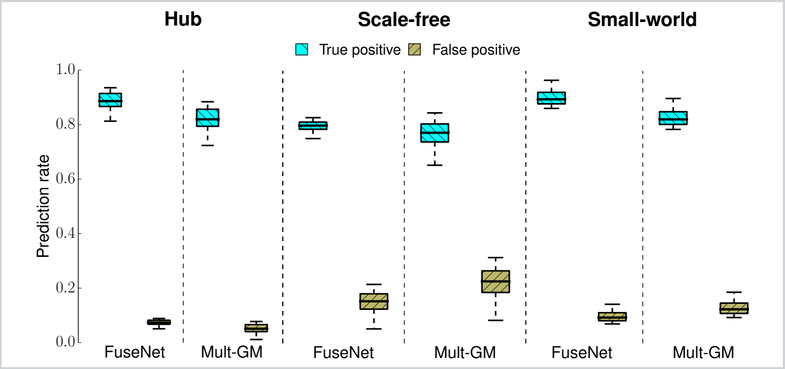
In every simulation, we generated a data set of observations based on a simulated network and then applied different network inference algorithms to determine whether the algorithms successfully recovered complex relationships between data variables.

We simulated four network types, which are known to resemble the structure of real biological networks (Costanzo et al., 2010; Allen and Liu, 2013). We report receiver operator curves computed by varying the regularization parameter λ in Fig. 5.4, boxplots of true and false positive rates for fixed λ as determined by stability selection in Fig. 5.4 and Fig. 5.3.

Experimental evidence indicates that FuseNet outperforms Gaussian-based competitors (GLASSO, Log-GLASSO and NPN-Copula) as well as existing methods that are designed specifically for the Poisson and the multinomial data (LPGM in Fig. 5.3 and

Figure 5.3

Application of gene network inference algorithms to multinomial-distributed simulated data. Simulation studies on three network types were performed: hub, scale-free and small-world. For each graph type, we generated $n = 300$ observations at a high signal-to-noise ratio (SNR) with $p = 50$ variables (nodes) taking values from an alphabet of size $m = 3$. Boxplots are shown for multinomial FuseNet (proposed here) and the multinomial graphical model (Mult-GM) (Jalali et al., 2011).



Mult-GM in Fig. 5.4). The overall good performance of FuseNet is consistent across the four types of network structure and the two data distributions that we considered in experiments.

The improved statistical power of FuseNet and LPGM over methods that during network inference rely heavily on the assumption of normality is particularly impressive. Results in Fig. 5.4 suggest that in situations where this assumption is not satisfied, we can expect reduced prediction performance if we naively apply Gaussian-based methods, (GLASSO) or if we perform insufficient data preprocessing (Log-GLASSO). However, we note that sophisticated techniques that replace Gaussian distributed data by the transformed data obtained, for example, through a semiparametric Gaussian copula (NPN-Copula; Liu et al. (2009)), can give substantial gains in accuracy over the naive analysis. These observations are not surprising as disregarding information about data distribution can adversely affect performance of prediction models. Our results demonstrate that employing the "correct" statistical model, in this case FuseNet or LPGM, can lead to more accurate network inference.

Next, we try to understand which algorithmic component of FuseNet contributes most to its good performance relative to existing algorithms for network structure learning. The primary difference between FuseNet and non-Gaussian-based methods considered here, LPGM and Mult-GM, is representation of model parameters

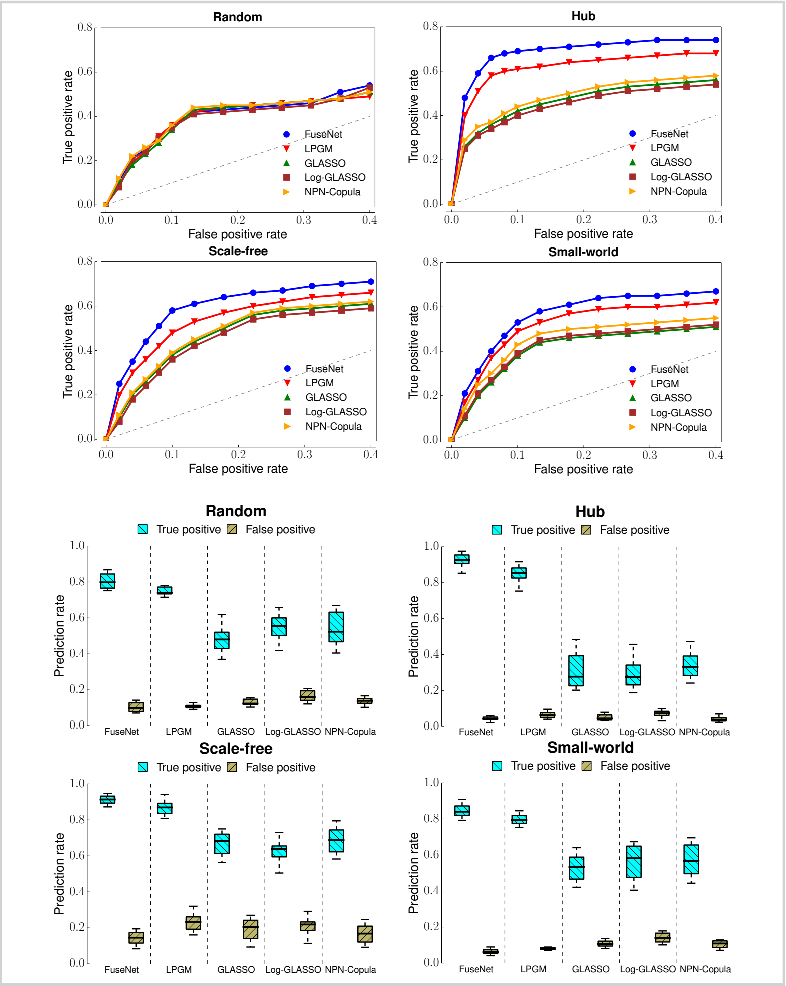


Figure 5.4

Application of gene network inference algorithms to Poisson-distributed simulated data. Simulation studies on four network types were performed: random, hub, scale-free and small-world. These graph structures appear in many real biological networks. For each graph type, we generated data with $n = 200$ observations with $p = 100$ variables (nodes) at a low (first row) and high (second row) signal-to-noise ratio (SNR). Receiver operating curves and boxplots are shown for Poisson FuseNet (proposed here), the Local Poisson Graphical Model (LPGM) (Allen and Liu, 2013), the Graphical Lasso (GLASSO) (Friedman et al., 2007b), the GLASSO on log-transformed data (Log-GLASSO) (*e.g.* cf. Gallopin et al., 2013) and the GLASSO on data transformed through nonparanormal Gaussian copula (NPN-Copula) (Liu et al., 2009) with products of latent factors. In LPGM and similarly in Mult-GM, a prediction model is fitted locally by an algorithm, which performs a series of independent penalized regressions. This is in contrast with FuseNet, where different model parameters are not entirely independent of each other but rather rely on borrowing strength from each other via factorization. Our results on simulated data suggest that representation of model parameters through the use of latent factors is beneficial. Furthermore, latent parameterization can improve performance of network recovery beyond what is possible with models that do not use latent factors. On the downside, we note that due to coupling of model parameters, FuseNet is not trivially parallelizable, which is otherwise true for LPGM and Mult-GM.

Results shown in Fig. 5.3 and Fig. 5.4 are reported for data sets with a few hundred observations (n) and a few tens of variables (p; see figure captions). We note that reported results are consistent with experiments done in various high-dimensional scenarios even when the number of variables is greater than the number of observations ($p > n$). Results therein reveal the same trend, namely, the overall strong performance of FuseNet in recovering true networks from non-Gaussian data.

5.7.2 Functional content of genomic networks

An important challenge in cancer systems biology is to uncover complex dependencies between genes implicated in cancer. Since our knowledge about genome-scale gene networks is incomplete and only a few functional modules are known for higher organisms (Rolland et al., 2014), our aim is to quantify associations between the inferred gene networks and known cellular functions and phenotypes, and to assess the significance of these associations.

Comparison of FuseNet *variants with existing methods*

To characterize how functionally informative the inferred networks are we employ four structural definitions of network communities (Figs. 5.5–5.7). Inferred networks were overlaid with GO terms and subnetworks induced by each GO term were assessed for how well they corresponded to network communities. Four different scoring functions are used to quantify the presence of different structural notions of communities

(Sec. 5.5) that appear in biological networks. These represent four possible notions of association between a given GO term and the inferred network (Yang and Leskovec, 2012). The triangle participation ratio quantifies how well genes that are members of a given GO term are linked to each other in the inferred network. The cut ratio captures the abundance of external connectivity, *i.e.*, edges between genes of a GO term and the rest of the network, whereas conductance and flake-ODF consider both internal and external network connectivity. Through these four measures we are able to estimate the overall concordance of inferred gene networks and known functional annotation of genes. For these reasons, networks that score higher on many measures should be considered more informative across a wider spectrum of cellular functions.

Fig. 5.5 shows that gene network inferred by FuseNet through fusion of breast cancer RNA-sequencing data and somatic mutation data is more concordant with functional annotation data in the GO than are networks inferred by FuseNet from either RNA-sequencing or somatic mutation data alone. We note that we used Poisson FuseNet to infer network from RNA-sequencing data, multinomial FuseNet to infer network from somatic mutation data and collective FuseNet for joint network inference from RNA-sequencing and mutation data. These results demonstrate that combining data through the use of latent factors can perform better than independent modeling of each data set alone.

For each of the four community scoring measures in Fig. 5.5, we compared score distributions of GO terms across three networks inferred by FuseNet using Kolmogorov-Smirnov tests. We concluded that the network inferred by FuseNet through fusion of RNA-sequencing and mutation data associates with GO significantly more strongly than the other two networks (p-value $< 1 \times 10^{-5}$ on all four measures from Fig. 5.5). This experiment shows how cancer genomic data provide different levels of information about cellular machinery, highlighting that it is possible to infer a network that better explains the mechanisms of cancer by combining multiple data sets in a principled statistical way.

We further compared FuseNet to existing network inference methods on cancer data. The comparison was made only with LPGM, as this was the best performing method in our study on simulated data (Sec. 5.7.1) and in the cancer-data study of Allen and Liu (2013). Fig. 5.6 shows the functional content of the networks inferred from RNA-

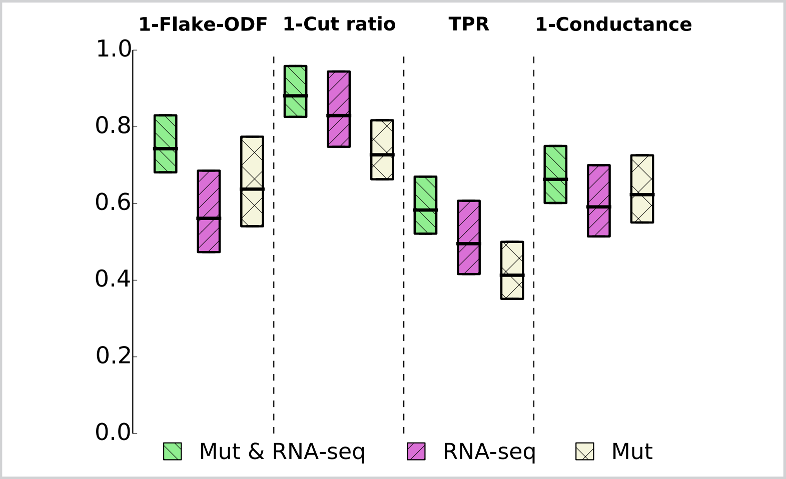


Figure 5.5

The strength of association between gene sets from the Gene Ontology (GO) and networks inferred with FuseNet. Considering breast cancer RNA-sequencing (RNA-seq) and somatic mutation data (Mut), these boxplots show the gains that fusion of data from different distributions (Mut & RNA-seq) can offer over network inference from any data set alone, either RNA-seq or Mut. Poisson FuseNet was used with RNA-sequencing data, multinomial FuseNet with somatic mutation data and fully-specified FuseNet for joint consideration of RNA-sequencing and mutation data. Flake-ODF, flake-over-median-degree; TPR, triangle participation ratio.

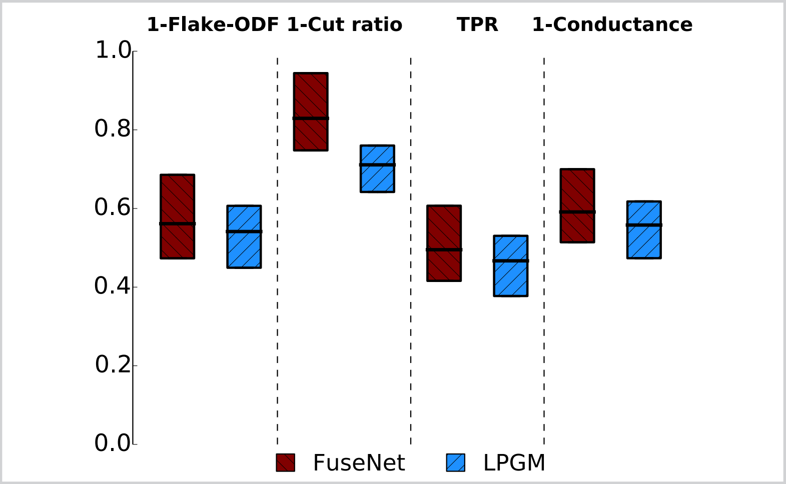


Figure 5.6

The strength of association between gene sets from the Gene Ontology (GO) and networks inferred with Poisson FuseNet (proposed here) and LPGM (Allen and Liu, 2013). Results are shown for breast cancer RNA-sequencing data because LPGM method was designed for Poisson distributed data. Flake-ODF, flake-over-median-degree; TPR, triangle participation ratio.

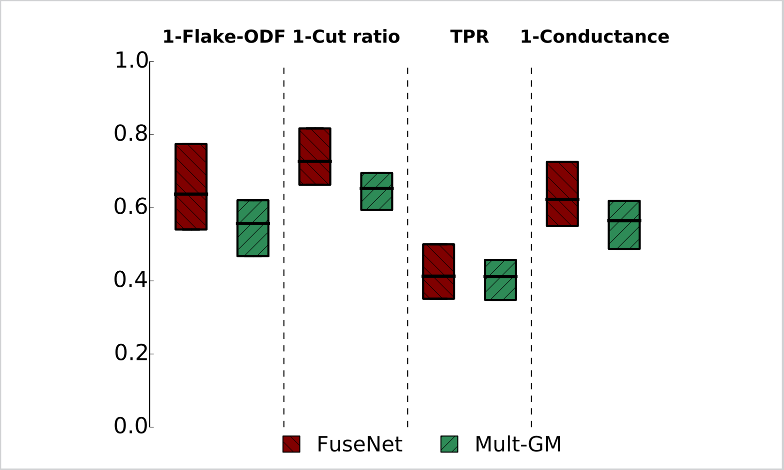


Figure 5.7

The strength of association between gene sets from the Gene Ontology (GO) and networks inferred with multinomial FuseNet (proposed here) and multinomial graphical model (Mult-GM) (Jalali et al., 2011). Results are shown for breast cancer somatic mutation data because Mult-GM method was designed for multinomial distributed data. Flake-ODF, flake-over-median-degree; TPR, triangle participation ratio.

sequencing data by either Poisson FuseNet or LPGM. On a related note, Fig. 5.7 shows enrichment of the networks inferred from somatic mutation data by either multinomial FuseNet or Mult-GM. Notice that LPGM and Mult-GM were designed for data that are approximately Poisson distributed, such as measurements from RNA-sequencing, and multinomially distributed, such as various types of gene variations, respectively. These results demonstrate that networks inferred by FuseNet can better capture known GO annotations than networks obtained by methods such as LPGM and Mult-GM, whose prediction models do not have factorized representation. These observations are consistent across four complementary structural definitions of GO terms, where every GO term is viewed as a network community defined by its member genes.

Networks via breast cancer data

We employ SANTA (Cornish and Markowetz, 2014) to quantify the functional content of gene networks. SANTA extends the concept of gene set enrichment analysis to networks. We observed that GO terms indeed cluster more strongly on Poisson FuseNet's networks than on networks inferred by GLASSO and Log-GLASSO (p-value $< 1\times10^{-6}$, RNA-seq network), NPN-Copula (p-value $< 1\times10^{-5}$, RNA-seq network) and LPGM (p-value $< 1 \times 10^{-4}$, RNA-seq network). These results suggest that network edges inferred by FuseNet might represent more accurate indication of shared cellular functions than edges inferred by other considered methods. This effect was independent of the GO term size and was strongest for specific cellular functions such as "centrosome cycle" (p-value $< 1 \times 10^{-9}$), "cellular response to DNA damage stimulus" (p-value $< 1 \times 10^{-9}$), "apoptotic process" (p-value $< 1 \times 10^{-9}$) and "regulation of cytokinesis" (p-value $< 1 \times 10^{-8}$). We observed similar results when inferring networks from somatic mutation data. Gene network inferred by multinomial FuseNet was functionally richer than network inferred by Mult-GM. Here, the functional content of a network was quantified with SANTA as proportion of evaluated GO terms whose association strength with the network had p-value $< 1 \times 10^{-5}$.

Interactions that are captured by fusing both cancer related data sets recovered many gene-gene associations that have been previously linked to increased breast cancer predisposition and metastasis. For example, FuseNet revealed a hypothesized transcriptional regulatory *GATA3* module (Wang et al., 2014) consisting of fully connected *GATA3*, *PTCH1*, *NFIB* and *PPARA*. *GATA3* is an important transcriptional regulator in breast cancer (Theodorou et al., 2013), and low expression levels of *GATA3* are associated with a poor prognosis (Albergaria et al., 2009). It has been shown by Wang et al. (2014) that *PTCH1*, *PPARA* and *NFIB* exhibit epistatic interactions with *GATA3*, have negatively correlated expression levels with *GATA3* and that *GATA3* binds to gene regions near *NFIB*, *PTCH1* and *PPARA* in breast epithelial tumor cell line.

Other interactions identified in our network include *ATM* and *BRCA1*, *ATM* and *BRCA2*, and *CHEK2* and *BRCA2*, which are known gene-gene interactions whose mutations affect breast cancer susceptibility (Turnbull et al., 2012).

Another transcriptional module that was found by FuseNet consists of *FLI1*, *JAK2* and *CCND2*. This module has been only recently associated with breast cancer patient outcome (Wang et al., 2014). Interestingly, *FLI1* module has been captured by FuseNet when fusing RNA-sequencing and mutation data but has been missed when using FuseNet with any of the two cancer data sets in isolation, as well as by any other inference algorithm considered in this study. One possible explanation for the latter result might be observations made by Wang et al. (2014). Wang et al. examined The Cancer Genome Atlas breast cancer patient survival data and found that low expression *or* mutation in one or more members of the *FLI1* module is associated with reduced overall survival time in all patients. The illustrative example of *FLI1* module highlights an advantage of FuseNet over methods considering a single data set during network inference.

5.8 Conclusion

FuseNet is an approach for automatic inference of gene networks from data arising from potentially many nonidentical distributions. It is based on the theory of Markov networks, where the inferred network edges denote a type of direct dependence that is stronger than merely correlated measurements. An appealing property of FuseNet is its ability to estimate network edges by *fusing potentially many data sets*. In the case studies FuseNet's models outperform several state-of-the-art undirected graphical models. We show that FuseNet's high performance is attributed to the ability to model non-Gaussian distributions and fusion of data through sharing of latent representations. Our work here has broadened the class of off-the-shelf network inference algorithms for simultaneously considering a wide range of parametric distributions and has combined Markov network inference with data fusion.

Part III

Compressive data fusion

6

Factorial multi-relation and multi-object type model

For most problems in science and engineering we can obtain data sets that describe the observed system from various perspectives and record the behavior of its individual components. Data fusion can focus on a specific target relation and exploit directly associated data together with contextual data and data about system's constraints.

In the chapter we describe a data fusion approach with collective penalized matrix tri-factorization (DFMF) that simultaneously factorizes data matrices to reveal hidden associations. The approach can directly consider any data that can be expressed in a matrix, including those from feature-based representations, ontologies, associations and networks.

In the following chapters we demonstrate the utility of DFMF for gene function prediction task with eleven different data sources and for prediction of pharmacologic actions by fusing six data sources. Our data fusion algorithm compares favorably to alternative data integration approaches and achieves higher accuracy than can be obtained from any single data source alone.

6.1 Background

Data abound in all areas of human endeavor. We may gather various data sets that are directly related to the problem, or data sets that are loosely related to our study but could be useful when combined with other data sets. Consider, for example, the exposome (Rappaport and Smith, 2010) that encompasses the totality of human endeavor in the study of disease. Let us say that we examine susceptibility to a particular disease and have access to the patients' clinical data together with data on their demographics, habits, living environments, friends, relatives, movie-watching habits, and movie genre ontology. Mining such a diverse data collection may reveal interesting patterns that would remain hidden if we would analyze only directly related, clinical data. What if the disease was less common in living areas with more open spaces or in environments where people need to walk instead of drive to the nearest grocery? Is the disease less common among those that watch comedies and ignore politics and news?

Methods for data fusion collectively treat data sets and combine diverse data sources even when they differ in their conceptual, contextual and typographical representation (Aerts et al., 2006; Boström et al., 2007). Individual data sets may be incomplete, yet because of their diversity and complementarity, fusion can improve the robustness and predictive performance of the resulting models (Greene and Cunningham, 2009; Lanckriet et al., 2004c).

According to Pavlidis et al. (2002), data fusion approaches can be classified into three main categories depending on the modeling stage at which fusion takes place. *Early (or full) integration* transforms all data sources into a single feature-based table and treats this as a single data set that can be explored by any of the well-established feature-based machine learning algorithms. The inferred models can in principle include any type of relationships between the features from within and between the data sources. Early integration relies on procedures for feature construction. For our exposome example, patient-specific data would need to include both clinical data and information from the movie genre ontologies. The former may be trivial as this data is already related to each specific patient, while the latter requires more complex feature engineering. Early integration also neglects the modular structure of the data.

In *late (decision) integration*, each data source gives rise to a separate model. Predictions of these models are fused by model weighting. Again, prior to model inference, it is necessary to transform each data set to encode relations to the target concept. For our example, information on the movie preferences of friends and relatives would need to be mapped to disease associations. Such transformations may not be trivial and would need to be crafted independently for every data source.

The youngest branch of data fusion algorithms is *intermediate (partial) integration*. Algorithms in this category explicitly address the multiplicity of data and fuse them through inference of a single joint model. Intermediate integration does not merge the input data, nor does it develop separate models for each data source. It instead retains the structure of the data sources by incorporating it within the structure of predictive model. This particular approach is often preferred because of its superior predictive accuracy (Pavlidis et al., 2002; Lanckriet et al., 2004c; Gevaert et al., 2006; Tang et al., 2009; van Vliet et al., 2012), but for a given model type, it requires the development of a new inference algorithm.

We here report on the development of a new method for intermediate data fusion based on constrained matrix factorization. Our aim was to construct an algorithm that requires no or only minimal transformation of input data and can fuse feature-based representations, ontologies, associations and networks. We focus on the challenge of dealing with collections of heterogeneous data sources, and while showing that our method can be used on sizable problems from current research, scaling is not the focus of the present chapter. We first present our data fusion algorithm, henceforth DFMF (Sec. 6.2), and then place it within the related work of relational learning approaches (Sec. 6.3). We also refer to related data integration approaches, specifically to methods of kernel-based data fusion (Sec. 6.3). We then examine the utility of DFMF and experimentally compare it with intermediate integration by multiple kernel learning, early integration with random forests, and tri-SPMF (Wang et al., 2008), previously proposed matrix tri-factorization approach (Sec. 3.4).

6.2 Data fusion by collective matrix factorization

The DFMF considers r object types $\mathscr{E}_1, \dots, \mathscr{E}_r$ and a collection of data sources, each relating a pair of object types $(\mathscr{E}_i, \mathscr{E}_j)$. In our introductory example of the exposome, object types could be a patient, a disease or a living environment, among others. If there are n_i objects of type $\mathscr{E}_i$ (o_p^i is p-th object of type $\mathscr{E}_i$) and n_j objects of type $\mathscr{E}_j$, we represent the observations from the data source that relates $(\mathscr{E}_i, \mathscr{E}_j)$ for $i \neq j$ in a sparse matrix $\boldsymbol{R}_{ij} \in \mathbb{R}^{n_i \times n_j}$. An example of such a matrix would relate patients and drugs by reporting on patient's current drug prescriptions. Notice that matrices $\boldsymbol{R}_{ij}$ and $\boldsymbol{R}_{ji}$ are in general asymmetric. A data source that provides relations between objects of the same type $\mathscr{E}_i$ is represented by a constraint matrix $\boldsymbol{\Theta}_i \in \mathbb{R}^{n_i \times n_i}$. Examples of such constraints are social networks and drug interactions.

In real-world scenarios we might not have access to relations between all pairs of object types. Our data fusion algorithm still integrates all available data if the underlying graph of relations between object types is connected. In that case, low-dimensional representations of objects of certain type borrow information from related objects of the different type. Fig. 6.1 shows an example of an underlying graph of relations and a block configuration of the fusion system with four object types.

To retain the block structure of our fusion system and hence model distinct relations between object types, we propose the simultaneous factorization of all relation matrices $\boldsymbol{R}_{ij}$ constrained by $\boldsymbol{\Theta}_i$. The resulting system contains factors that are specific to each

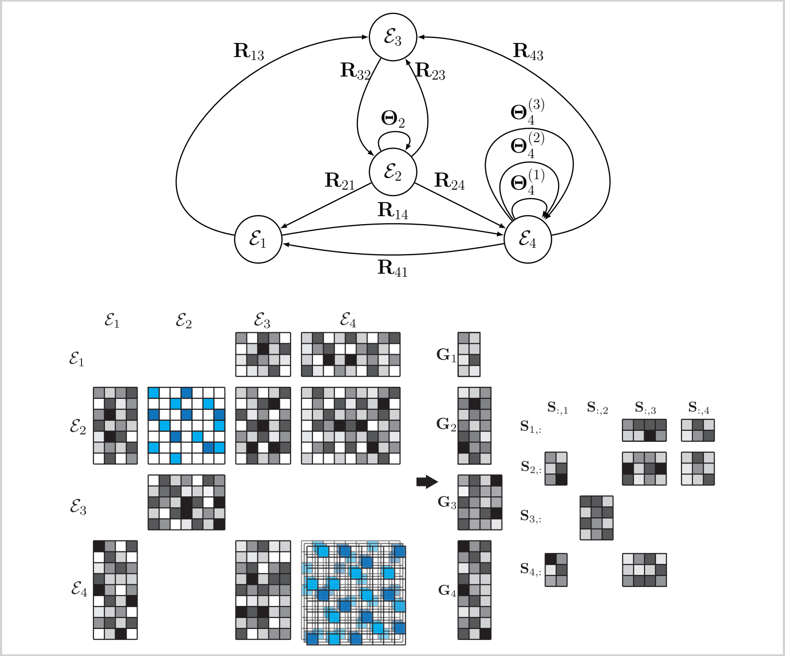


Figure 6.1

Conceptual fusion configuration for four object types, $\mathscr{E}_1$, $\mathscr{E}_2$, $\mathscr{E}_3$ and $\mathscr{E}_4$, equivalently represented by the graph of relations between object types (top) and the block-based matrix structure (bottom). Every data source relates a pair of object types and is denoted by an arc in the graph (top) and given in a matrix with shades of gray (bottom). For example, data matrix $\boldsymbol{R}_{23}$ relates object types $\mathscr{E}_2$ and $\mathscr{E}_3$. Some relations are entirely missing. For instance, there is no data source relating objects from $\mathscr{E}_3$ and $\mathscr{E}_1$, as there is no arc linking nodes $\mathscr{E}_3$ and $\mathscr{E}_1$ in (top) or equivalently, a matrix $\boldsymbol{R}_{31}$ is missing in (bottom). Relations can be asymmetric, such that $\boldsymbol{R}_{23} \neq \boldsymbol{R}_{32}^T$. Constraints denoted by loops in (top) or matrices with blue entries in (bottom) relate objects of the same type. In our example configuration, constraints are provided for object types $\mathscr{E}_2$ (one constraint matrix) and $\mathscr{E}_4$ (three constraint matrices).

data source and factors that are specific to each object type. Through factor sharing we fuse the data but also identify source-specific patterns.

We have developed a variant of three-factor penalized matrix factorization that simultaneously decomposes all available relation matrices $\boldsymbol{R}_{ij}$ into $\boldsymbol{G}_i \in \mathbb{R}^{n_i \times k_i}$, $\boldsymbol{G}_j \in \mathbb{R}^{n_j \times k_j}$ and $\boldsymbol{S} \in \mathbb{R}^{k_i \times k_j}$, and regularizes their approximation through constraint matrices $\boldsymbol{\Theta}_i$ and $\boldsymbol{\Theta}_j$ such that $\boldsymbol{R}_{ij} \approx \boldsymbol{G}_i \boldsymbol{S}_{ij} \boldsymbol{G}_j^T$. Approximation can be rewritten such that entry $\boldsymbol{R}_{ij}(p, q)$ is approximated by an inner product of the p-th row of matrix $\boldsymbol{G}_i$ and a linear combination of the columns of matrix $\boldsymbol{S}_{ij}$, weighted by the q-th column of $\boldsymbol{G}_j$. The matrix $\boldsymbol{S}_{ij}$, which has relatively few vectors compared to $\boldsymbol{R}_{ij}$ ($k_i \ll n_i$, $k_j \ll n_j$), is used to represent many data vectors, and a good approximation can only be achieved in the presence of the latent structure in the original data.

The proposed fusion approach is different from treating an entire system (e.g., from Fig. 6.1) as a large single matrix. Factorization of such a matrix would yield factors that are not object type-specific and would thus disregard the structure of the system. We also show (Sec. 7.8) that such an approach is inferior in terms of predictive performance.

In comparison with existing multi-type relational data factorization approaches (see Sec. 6.3) the following characterizes our DFMF data fusion method:

i DFMF can model *multiple* relations between *multiple* object types.

ii Relations between some object types can be completely missing (see Fig. 6.1).

iii Every object type can be associated with multiple constraint matrices.

iv The algorithm makes no assumptions about structural properties of relations (*e.g.* symmetry of relations).

In order to be applicable to general real-world fusion problems, data fusion algorithm would need to jointly address all of these characteristics. Besides DFMF proposed in this manuscript, we are not aware of any other approach that would do so. Most real-world data integration problems would usually consider a larger number of object types, but with growing number of object types, it is likely that data relating a pair of object types is either not available nor meaningful. On the other side, there may be various data sources available on interactions between objects of the same type that also require appropriate treatment. For example of this type of data, consider abundance of data bases on drug or disease interactions.

In the case study presented in this chapter we apply data fusion to infer relations between two target object types, $\mathcal{E}_i$ and $\mathcal{E}_j$ (Sec. 6.2.6 and Sec. 6.2.7). This relation, encoded in a target matrix $\boldsymbol{R}_{ij}$, will be observed in the context of all other data sources (Sec. 6.2.1). We assume that our target $\boldsymbol{R}_{ij}$ is a $[0, 1]$-matrix that is only partially observed. Its entries indicate a degree of relation, 0 denoting no relation and 1 denoting the strongest relation. We aim to predict unobserved entries in $\boldsymbol{R}_{ij}$ by reconstructing them through matrix factorization. Such treatment in general applies to multi-class or multi-label classification tasks, which are conveniently addressed by multiple kernel fusion (Yu et al., 2010), with which we compare our performance in this chapter.

In the following, we present the factorization model, objective function, derive the updating rules for optimization, and describe the procedure for prediction of relations from matrix factors. In the optimization part, we closely follow (Wang et al., 2008) in notation, mathematical derivation and proof technique.

6.2.1 Multi-relation and multi-object type factorial model

An input to DFMF is a relation block matrix $\boldsymbol{R}$ that conceptually represents all relation matrices:

$$\boldsymbol{R} = \begin{bmatrix} * & \boldsymbol{R}_{12} & \cdots & \boldsymbol{R}_{1r} \\ \boldsymbol{R}_{21} & * & \cdots & \boldsymbol{R}_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{R}_{r1} & \boldsymbol{R}_{r2} & \cdots & * \end{bmatrix}. \tag{6.1}$$

Here, an asterisk ("*") denotes the relation between the same type of objects that DMFM does not model. Notice that our method does not require the presence of all relation matrices in Eq. (6.1). Depending on a particular data setup, any subset of relation matrices might be missing and thus, not considered in the analysis. A block in the i-th row and j-th column ($\boldsymbol{R}_{ij}$) of matrix $\boldsymbol{R}$ represents the relationship between object type $\mathscr{E}_i$ and $\mathscr{E}_j$. The p-th object of type $\mathscr{E}_i$ (i.e. o_p^i) and q-th object of type $\mathscr{E}_j$ (i.e. o_q^j) are related by $\boldsymbol{R}_{ij}(p,q)$. An important aspect of Eq. (6.1) for data fusion and what distinguishes DMFM from other conceptually related matrix factorization models such as S-NMTF (Wang et al., 2011a) or even tri-SPMF (Wang et al., 2008) is that it is designed for multi-object type and multi-relational data where the relations can be asymmetric, $\boldsymbol{R}_{ji} \neq \boldsymbol{R}_{ij}^T$, and some can be completely missing (unknown $\boldsymbol{R}_{ij}$) (Sec. 6.2.3).

We additionally consider constraints relating objects of the same type. Several data sources of this kind may be available for each object type. For instance, personal relations may be observed from a social network or a family tree. Assume there are $t_i \geq 0$ data sources for object type $\mathscr{E}_i$ represented by a set of constraint matrices $\boldsymbol{\Theta}_i^{(t)}$ for $t \in \{1, 2, \ldots, t_i\}$. Constraints are collectively encoded in a set of constraint block diagonal matrices $\boldsymbol{\Theta}^{(t)}$ for $t \in \{1, 2, \ldots, \max_i t_i\}$:

$$\boldsymbol{\Theta}^{(t)} = Diag(\boldsymbol{\Theta}_1^{(t)}, \boldsymbol{\Theta}_2^{(t)}, \ldots, \boldsymbol{\Theta}_r^{(t)}) \tag{6.2}$$

The i-th block along the main diagonal of $\boldsymbol{\Theta}^{(t)}$ is zero if $t > t_i$. Entries in constraint matrices are positive for objects that are not similar and negative for objects that are similar. The former are known as *cannot-link constraints* because they impose penalties on the current approximation of the matrix factors, and the latter are *must-link constraints*, which are rewards that reduce the value of the cost function during optimization. Must-link constraint expresses the notion that a pair of objects of the same type should be close in their latent component space. An example of must-link constraints are, for instance, drug-drug interactions, and example of cannot-link constraints the matrix of adversaries. Typically, data sources with must-link constraints are more abundant.

The block matrix $\boldsymbol{R}$ is tri-factorized into block matrix factors $\boldsymbol{G}$ and $\boldsymbol{S}$:

$$\boldsymbol{G} = Diag(\boldsymbol{G}_1^{n_1 \times k_1}, \boldsymbol{G}_2^{n_2 \times k_2}, \dots, \boldsymbol{G}_r^{n_r \times k_r}),$$

$$\boldsymbol{S} = \begin{bmatrix} * & \boldsymbol{S}_{12}^{k_1 \times k_2} & \cdots & \boldsymbol{S}_{1r}^{k_1 \times k_r} \\ \boldsymbol{S}_{21}^{k_2 \times k_1} & * & \cdots & \boldsymbol{S}_{2r}^{k_2 \times k_r} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{S}_{r1}^{k_r \times k_1} & \boldsymbol{S}_{r2}^{k_r \times k_2} & \cdots & * \end{bmatrix}. \tag{6.3}$$

Matrix $\boldsymbol{S}$ in Eq. (6.3) has the same block structure as $\boldsymbol{R}$ in Eq. (6.1). It is in general asymmetric (*i.e.* $\boldsymbol{S}_{ji} \neq \boldsymbol{S}_{ij}^T$) and if a relation matrix is missing in $\boldsymbol{R}$ then also its corresponding matrix factor in $\boldsymbol{S}$ will be missing. These two properties of $\boldsymbol{S}$ stem from our decision to model relation matrices without assuming their structural properties or their availability for every possible combination of object types.

A factorization rank k_i is assigned to $\mathscr{E}_i$ during inference of the factorized system. Factor $\boldsymbol{S}_{ij}$ defines the latent relation between object types $\mathscr{E}_i$ and $\mathscr{E}_j$, while factor $\boldsymbol{G}_i$ is specific to objects of type $\mathscr{E}_i$ and is used in the reconstruction of every relation with this object type. In this way, each relation matrix $\boldsymbol{R}_{ij}$ obtains its own factorization $\boldsymbol{G}_i\boldsymbol{S}_{ij}\boldsymbol{G}_j^T$ with factor $\boldsymbol{G}_i$ ($\boldsymbol{G}_j$) that is shared across all relations which involve object types $\mathscr{E}_i$ ($\mathscr{E}_j$). This can also be observed from the block structure of the reconstructed system $\boldsymbol{G}\boldsymbol{S}\boldsymbol{G}^T$:

$$\begin{bmatrix} * & \boldsymbol{G}_1\boldsymbol{S}_{12}\boldsymbol{G}_2^T & \cdots & \boldsymbol{G}_1\boldsymbol{S}_{1r}\boldsymbol{G}_r^T \\ \boldsymbol{G}_2\boldsymbol{S}_{21}\boldsymbol{G}_1^T & * & \cdots & \boldsymbol{G}_2\boldsymbol{S}_{2r}\boldsymbol{G}_r^T \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{G}_r\boldsymbol{S}_{r1}\boldsymbol{G}_1^T & \boldsymbol{G}_r\boldsymbol{S}_{r2}\boldsymbol{G}_2^T & \cdots & * \end{bmatrix}. \tag{6.4}$$

Here, the p-th row in factor $\boldsymbol{G}_i$ holds the latent component representation of object o_p^i. By holding $\boldsymbol{G}_j$ and $\boldsymbol{S}_{ij}$ fixed, it is clear that latent component representation of o_p^i depends on $\boldsymbol{G}_j$ as well as on the existence of relation $\boldsymbol{R}_{ij}$. Consequently, all direct and indirect relations have a determining influence on the calculation of o_p^i-th latent representation. Just as the objects of type $\mathscr{E}_i$ are represented by $\boldsymbol{G}_i$, each relation is represented by factor $\boldsymbol{S}_{ij}$, which models how the latent components interact in the respective relation. The asymmetry of $\boldsymbol{S}_{ij}$ takes into account whether a latent component occurs as a subject or an object of corresponding relation $\boldsymbol{R}_{ij}$.

6.2.2 Objective function

The objective function minimized by DFMF aims at good approximation of the input data and adherence to must-link and cannot-link constraints:

$$\begin{aligned} \min_{G\geq 0} J(\boldsymbol{G}; \boldsymbol{S}) \quad &= \quad \sum_{R_{ij}\in\mathscr{R}} ||\boldsymbol{R}_{ij} - \boldsymbol{G}_i\boldsymbol{S}_{ij}\boldsymbol{G}_j^T||^2 + \\ &+ \sum_{t=1}^{\max_i t_i} \mathrm{tr}(\boldsymbol{G}^T\boldsymbol{\Theta}^{(t)}\boldsymbol{G}), \end{aligned} \tag{6.5}$$

Here, $||\cdot||$ and $\mathrm{tr}(\cdot)$ denote the Frobenius norm and trace, respectively, and $\mathscr{R}$ is the set of all relations included in our model. Our objective function explicitly allows that relations between some object types are entirely missing.

Notice that in Eq. (6.5) we do not approximate input data by $||\boldsymbol{R} - \boldsymbol{G}\boldsymbol{S}\boldsymbol{G}^T||^2$ as was proposed in related approaches of S-NMTF (Wang et al., 2011a) and tri-SPMF (Wang et al., 2008). To model the data system such as that from Fig. 6.1, one could be tempted to replace the missing relation matrices with zero matrices. This would enable the optimization to further reduce the value of objective function, but would also introduce relations in factorized system that were intentionally not present in the input data. Their inclusion in the model would distort inferred relations between other object types (see Sec. 7.4).

6.2.3 *Computing the factorization*

The DFMF algorithm for solving the minimization problem specified in Eq. (6.5) is shown in Algorithm 3. The algorithm first initializes matrix factors (Sec. 6.2.8) and then iteratively refines them by alternating between fixing $\boldsymbol{G}$ and updating $\boldsymbol{S}$, and then fixing $\boldsymbol{S}$ and updating $\boldsymbol{G}$, until convergence. Successive updates of $\boldsymbol{G}_i$ and $\boldsymbol{S}_{ij}$ converge to a *local minimum of the problem given in Eq. (6.5)*.

We derive multiplicative updating rules for regularized decomposition of relation matrices by fixing one matrix factor (e.g., $\boldsymbol{G}$) and considering the roots of the partial derivative with respect to the other matrix factor (e.g., $\boldsymbol{S}$, and vice-versa) of the Lagrangian function. The latter is constructed from the objective function (Eq. 6.5):

$$\begin{aligned} J(\boldsymbol{G};\boldsymbol{S}) \quad = \quad & \sum_{R_{ij}\in\mathcal{R}} \mathrm{tr}(\boldsymbol{R}_{ij}^T\boldsymbol{R}_{ij} - 2\boldsymbol{G}_j^T\boldsymbol{R}_{ij}^T\boldsymbol{G}_i\boldsymbol{S}_{ij} + \\ & + \boldsymbol{G}_i^T\boldsymbol{G}_i\boldsymbol{S}_{ij}\boldsymbol{G}_j^T\boldsymbol{G}_j\boldsymbol{S}_{ij}^T) + \\ & + \sum_{t=1}^{\max_i t_i}\sum_{i=1}^{r} \mathrm{tr}(\boldsymbol{G}_i^T\boldsymbol{\Theta}_i^{(t)}\boldsymbol{G}_i). \end{aligned} \tag{6.6}$$

Regarding the correctness and convergence of the algorithm in Algorithm 3 we have the following two theorems.

Theorem 1: (Correctness of DFMF algorithm). If the update rules for matrix factors $\boldsymbol{G}$ and $\boldsymbol{S}$ from Algorithm 3 converge, then the final solution satisfies the Karuch-Kuhn-Tucker (KKT) conditions (Kuhn and Tucker, 1951) of optimality.

Proof 1: We introduce the Lagrangian multipliers $\lambda_1, \lambda_2, \dots, \lambda_r$ and construct the Lagrange function:

$$L = J(\boldsymbol{G};\boldsymbol{S}) - \sum_{i=1}^{r}\mathrm{tr}(\lambda_i\mathbf{1}_{n_i\times k_i}\boldsymbol{G}_i^T). \tag{6.7}$$

Then for i, j, such that $\boldsymbol{R}_{ij} \in \mathscr{R}$:

$$\frac{\partial L}{\partial \boldsymbol{S}_{ij}} = -2\boldsymbol{G}_i^T \boldsymbol{R}_{ij} \boldsymbol{G}_j + 2\boldsymbol{G}_i^T \boldsymbol{G}_i \boldsymbol{S}_{ij} \boldsymbol{G}_j^T \boldsymbol{G}_j,$$

and for $i = 1, 2, \ldots, r$:

$$\begin{aligned} \frac{\partial L}{\partial \boldsymbol{G}_i} &= \sum_{j:R_{ij} \in \mathscr{R}} (-2\boldsymbol{R}_{ij} \boldsymbol{G}_j \boldsymbol{S}_{ij}^T + 2\boldsymbol{G}_i \boldsymbol{S}_{ij} \boldsymbol{G}_j^T \boldsymbol{G}_j \boldsymbol{S}_{ij}^T) + \\ &+ \sum_{j:R_{ji} \in \mathscr{R}} (-2\boldsymbol{R}_{ji}^T \boldsymbol{G}_j \boldsymbol{S}_{ji} + 2\boldsymbol{G}_i \boldsymbol{S}_{ji}^T \boldsymbol{G}_j^T \boldsymbol{G}_j \boldsymbol{S}_{ji}) + \\ &+ \sum_{t=1}^{\max_i t_i} 2\boldsymbol{\Theta}_i^{(t)} \boldsymbol{G}_i - \lambda_i \mathbf{1}_{n_i \times k_i}. \end{aligned} \tag{6.8}$$

Fixing $\boldsymbol{G}_1, \boldsymbol{G}_2, \ldots, \boldsymbol{G}_r$ and letting $\frac{\partial L}{\partial S_{ij}} = 0$ for all $i, j = 1, 2, \ldots, r$, we obtain:

$$\boldsymbol{S} = (\boldsymbol{G}^T \boldsymbol{G})^{-1} \boldsymbol{G}^T \boldsymbol{R} \boldsymbol{G} (\boldsymbol{G}^T \boldsymbol{G})^{-1}.$$

We then fix $\boldsymbol{S}$ and let $\frac{\partial L}{\partial G_i} = 0$ for $i = 1, 2, \ldots, r$. We get an expression for the KKT multiplier λ_i from Eq. (6.8). Then the KKT complementary condition for the nonnegativity of $\boldsymbol{G}_i$ is:

$$\begin{aligned} \mathbf{0} &= \lambda_i \mathbf{1}_{n_i \times k_i} \circ \boldsymbol{G}_i = \\ &= \Bigg[\sum_{j:R_{ij} \in \mathscr{R}} (-2\boldsymbol{R}_{ij} \boldsymbol{G}_j \boldsymbol{S}_{ij}^T + 2\boldsymbol{G}_i \boldsymbol{S}_{ij} \boldsymbol{G}_j^T \boldsymbol{G}_j \boldsymbol{S}_{ij}^T) + \\ &+ \sum_{j:R_{ji} \in \mathscr{R}} (-2\boldsymbol{R}_{ji}^T \boldsymbol{G}_j \boldsymbol{S}_{ji} + 2\boldsymbol{G}_i \boldsymbol{S}_{ji}^T \boldsymbol{G}_j^T \boldsymbol{G}_j \boldsymbol{S}_{ji}) + \\ &+ \sum_{t=1}^{\max_i t_i} 2\boldsymbol{\Theta}_i^{(t)} \boldsymbol{G}_i \Bigg] \circ \boldsymbol{G}_i. \end{aligned} \tag{6.9}$$

Here, $\circ$ denotes the Hadamard product. Let us here introduce variables $\boldsymbol{\Gamma}_i$ to denote $\boldsymbol{\Gamma}_i = \lambda_i \circ \boldsymbol{G}_i$. Eq. (6.9) is a fixed point equation and the solution must satisfy it at convergence. We let:

$$\begin{aligned}
\boldsymbol{\Theta}_i^{(t)} &= [\boldsymbol{\Theta}_i^{(t)}]^+ - [\boldsymbol{\Theta}_i^{(t)}]^- \\
\boldsymbol{R}_{ij}\boldsymbol{G}_j\boldsymbol{S}_{ij}^T &= (\boldsymbol{R}_{ij}\boldsymbol{G}_j\boldsymbol{S}_{ij}^T)^+ - (\boldsymbol{R}_{ij}\boldsymbol{G}_j\boldsymbol{S}_{ij}^T)^- \\
\boldsymbol{S}_{ij}\boldsymbol{G}_j^T\boldsymbol{G}_j\boldsymbol{S}_{ij}^T &= (\boldsymbol{S}_{ij}\boldsymbol{G}_j^T\boldsymbol{G}_j\boldsymbol{S}_{ij}^T)^+ - (\boldsymbol{S}_{ij}\boldsymbol{G}_j^T\boldsymbol{G}_j\boldsymbol{S}_{ij}^T)^- \\
\boldsymbol{R}_{ji}^T\boldsymbol{G}_j\boldsymbol{S}_{ji} &= (\boldsymbol{R}_{ji}^T\boldsymbol{G}_j\boldsymbol{S}_{ji})^+ - (\boldsymbol{R}_{ji}^T\boldsymbol{G}_j\boldsymbol{S}_{ji})^- \\
\boldsymbol{S}_{ji}^T\boldsymbol{G}_j^T\boldsymbol{G}_j\boldsymbol{S}_{ji} &= (\boldsymbol{S}_{ji}^T\boldsymbol{G}_j^T\boldsymbol{G}_j\boldsymbol{S}_{ji})^+ - (\boldsymbol{S}_{ji}^T\boldsymbol{G}_j^T\boldsymbol{G}_j\boldsymbol{S}_{ji})^-
\end{aligned}$$

where all matrices on right-hand sides are nonnegative. Then, given an initial guess of $\boldsymbol{G}_i$, the successive updates of $\boldsymbol{G}_i$ using Eq. (6.13)–(6.15) converge to a local minimum of the problem in Eq. (6.5). It can be easily seen that using such a rule, at convergence, $\boldsymbol{G}_i$ satisfies $\boldsymbol{\Gamma}_i \circ \boldsymbol{G}_i = \boldsymbol{0}$, which is equivalent to $\boldsymbol{\Gamma}_i = 0$ (Eq. (6.9)) due to nonnegativity of $\boldsymbol{G}_i$. □

Theorem 2: (Convergence of DFMF algorithm). The objective function $\boldsymbol{J}(\boldsymbol{G};\boldsymbol{S})$ given by Eq. (6.5) is nonincreasing under the updating rules for matrix factors $\boldsymbol{G}$ and $\boldsymbol{S}$ in Algorithm 3.

Proof 2: Our proof follows the concept of *auxiliary functions* often used in convergence proofs of approximate matrix factorization algorithms (Lee and Seung, 2000). The proof is performed by introducing an appropriate function $F(\boldsymbol{G},\boldsymbol{G}')$, which is an auxiliary function of the objective $\boldsymbol{J}(\boldsymbol{G};\boldsymbol{S})$ that satisfies:

$$\begin{aligned}
F(\boldsymbol{G}',\boldsymbol{G}') &= \boldsymbol{J}(\boldsymbol{G}';\boldsymbol{S}), \\
F(\boldsymbol{G},\boldsymbol{G}') &\geq \boldsymbol{J}(\boldsymbol{G};\boldsymbol{S}).
\end{aligned}$$

If such an auxiliary function F can be found and if $\boldsymbol{G}$ is updated in $(m+1)$-th iteration as the minimizer of the auxiliary function F, i.e.:

$$\boldsymbol{G}^{(m+1)} = \arg\min_{\boldsymbol{G}} F(\boldsymbol{G},\boldsymbol{G}^{(m)}), \tag{6.10}$$

then the following holds:

$$\begin{aligned} J(\boldsymbol{G}^{(m+1)}; \boldsymbol{S}) &\leq F(\boldsymbol{G}^{(m+1)}, \boldsymbol{G}^{(m)}) \leq \\ &\leq F(\boldsymbol{G}^{(m)}, \boldsymbol{G}^{(m)}) = \\ &= J(\boldsymbol{G}^{(m)}; \boldsymbol{S}). \end{aligned} \tag{6.11}$$

That is, if F is an auxiliary function of $J(\boldsymbol{G}; \boldsymbol{S})$, then $J(\boldsymbol{G}; \boldsymbol{S})$ is nonincreasing under the update Eq. (6.10). In the proof we show that the update step for $\boldsymbol{G}$ in Eq. (6.15) is exactly the update in Eq. (6.10) with a proper auxiliary function. For that we make use of an auxiliary function specified by Wang et al. (2008) (cf. Appendix II in Wang et al. (2008)). Wang et al. (2008) constructed a function $F_{\text{Wang}}(\boldsymbol{A}, \boldsymbol{A}'; \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D})$ and showed that it satisfied the conditions of auxiliary functions for functions of the form $J(\boldsymbol{A}; \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}) = tr(-2\boldsymbol{A}^T\boldsymbol{B} + \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}^T) + tr(\boldsymbol{A}^T\boldsymbol{C}\boldsymbol{A})$, where $\boldsymbol{C}$ and $\boldsymbol{D}$ are symmetric, and $\boldsymbol{A}$ is nonnegative. To prove the convergence of our algorithm, we show that the objective function from Eq. (6.5) is a special case of $J(\boldsymbol{A}; \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D})$.

First, we view $J(\boldsymbol{G}; \boldsymbol{S})$ in Eq. (6.6) as a function of $\boldsymbol{G}_1$ and construct the auxiliary function $F_{\text{Wang}}(\boldsymbol{A}, \boldsymbol{A}'; \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D})$ such that:

$$\begin{aligned} \boldsymbol{A} &= \boldsymbol{G}_1, \\ \boldsymbol{B} &= \sum_{j:R_{1j}\in\mathcal{R}} \boldsymbol{R}_{1j}\boldsymbol{G}_j\boldsymbol{S}_{1j}^T + \sum_{i:R_{i1}\in\mathcal{R}} \boldsymbol{R}_{i1}^T\boldsymbol{G}_i\boldsymbol{S}_{i1}, \\ \boldsymbol{C} &= \sum_{t=1}^{\max_i t_i} \boldsymbol{\Theta}_1^{(t)}, \\ \boldsymbol{D} &= \sum_{j:R_{1j}\in\mathcal{R}} \boldsymbol{S}_{1j}\boldsymbol{G}_j^T\boldsymbol{G}_j\boldsymbol{S}_{1j}^T + \sum_{i:R_{i1}\in\mathcal{R}} \boldsymbol{S}_{i1}^T\boldsymbol{G}_i^T\boldsymbol{G}_i\boldsymbol{S}_{i1}. \end{aligned} \tag{6.12}$$

With these values for $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ and $\boldsymbol{D}$, the auxiliary function F_{Wang} is convex in $\boldsymbol{G}_1$. Notice that each of the two summation terms in the right-hand side expression for $\boldsymbol{D}$ represents the sum of the symmetric matrices of the form $(\boldsymbol{G}_j\boldsymbol{S}_{1j}^T)^T(\boldsymbol{G}_j\boldsymbol{S}_{1j}^T)$ and $(\boldsymbol{G}_i\boldsymbol{S}_{i1})^T(\boldsymbol{G}_i\boldsymbol{S}_{i1})$, respectively. Thus, $\boldsymbol{D}$ is symmetric. The global minimum

(Eq. (6.10)) of $F_{\text{Wang}}(\boldsymbol{A}, \boldsymbol{A}'; \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D})$ is exactly the update rule for $\boldsymbol{G}_1$ in Eq. (6.13)–(6.15).

We repeat this process by constructing the remaining $r-1$ auxiliary functions by separately considering $\boldsymbol{J}(\boldsymbol{G}; \boldsymbol{S})$ as a function of matrix factors $\boldsymbol{G}_2 \dots, \boldsymbol{G}_r$. From the theory of auxiliary functions it then follows that $\boldsymbol{J}$ is nonincreasing under the update rules for each of $\boldsymbol{G}_1, \boldsymbol{G}_2, \dots, \boldsymbol{G}_r$. Letting $\boldsymbol{J}(\boldsymbol{G}_1, \boldsymbol{G}_2, \dots, \boldsymbol{G}_r, \boldsymbol{S}) = \boldsymbol{J}(\boldsymbol{G}; \boldsymbol{S})$, we have:

$$\begin{aligned} \boldsymbol{J}(\boldsymbol{G}_1^0, \boldsymbol{G}_2^0, \dots, \boldsymbol{G}_r^0, \boldsymbol{S}) &\geq \boldsymbol{J}(\boldsymbol{G}_1^1, \boldsymbol{G}_2^0, \dots, \boldsymbol{G}_r^0, \boldsymbol{S}) \geq \\ &\geq \cdots \\ &\geq \boldsymbol{J}(\boldsymbol{G}_1^1, \boldsymbol{G}_2^1, \dots, \boldsymbol{G}_r^1, \boldsymbol{S}). \end{aligned}$$

Since $\boldsymbol{J}(\boldsymbol{G}; \boldsymbol{S})$ is certainly bounded from below by zero, we proved the theorem. □

6.2.4 *Stopping criterion*

Recall that the optimization task from Eq. (6.5) is nonconvex and it thus has multiple local minima, each with different depths, for which the optimum is called the global minimum. The global minimum of our multi-relational system remains elusive and is impossible to determine in practice, where large dimensional data are common. However, we were still able to prove in the previous section that DFMF algorithm given in Algorithm 3 converges to a local minimum of Eq. (6.5).

Next, we would like to apply data fusion to infer relations between two target object types, $\mathcal{E}_i$ and $\mathcal{E}_j$. We hence define the stopping criterion that observes convergence in approximation of only the target matrix $\boldsymbol{R}_{ij}$. Our convergence criterion is $||\boldsymbol{R}_{ij} - \boldsymbol{G}_i \boldsymbol{S}_{ij} \boldsymbol{G}_j^T||^2 < \epsilon$, where ϵ is a user-defined parameter, possibly refined through observing log entries of the target matrix approximation error for several runs of the factorization algorithm. In our experiments ϵ was set to 10^{-5}. To reduce the computational load, the convergence criterion was assessed only every fifth iteration.

Algorithm 3

DFMF, data fusion by collective matrix factorization. Source code of DFMF and of its extensions to collective matrix completion and treatment of multiple relations over an object type pair is available at http://github.com/marinkaz/scikit-fusion.

```
Input:
  ▪ A set $\mathscr{R}$ of relation matrices $\boldsymbol{R}_{ij}$,
  ▪ constraint matrices $\boldsymbol{\Theta}^{(t)}$ for $t \in \{1, 2, \dots, \max_i t_i\}$
  ▪ factorization ranks $k_1, k_2, \dots, k_r$ $(i, j \in [r])$.
Output:
  ▪ Matrix factors $\boldsymbol{S}$ and $\boldsymbol{G}$.
1. Initialize $\boldsymbol{G}_i$ for $i = 1, 2, \dots, r$.
2. Repeat until convergence (Sec. 6.2.4) or a time limit is reached:
   a. Construct $\boldsymbol{R}$ and $\boldsymbol{G}$ using their definitions in Eq. (6.1) and Eq. (6.3).
   b. Update $\boldsymbol{S}$ using:
        $\boldsymbol{S} \leftarrow (\boldsymbol{G}^T\boldsymbol{G})^{-1}\boldsymbol{G}^T\boldsymbol{R}\boldsymbol{G}(\boldsymbol{G}^T\boldsymbol{G})^{-1}.$
   c. Set $\boldsymbol{G}_i^{(e)} \leftarrow \boldsymbol{0}$ for $i = 1, 2, \dots, r$.
   d. Set $\boldsymbol{G}_i^{(d)} \leftarrow \boldsymbol{0}$ for $i = 1, 2, \dots, r$.
   e. For $\boldsymbol{R}_{ij} \in \mathscr{R}$:
        $\boldsymbol{G}_i^{(e)} \mathrel{+}= (\boldsymbol{R}_{ij}\boldsymbol{G}_j\boldsymbol{S}_{ij}^T)^+ + \boldsymbol{G}_i(\boldsymbol{S}_{ij}\boldsymbol{G}_j^T\boldsymbol{G}_j\boldsymbol{S}_{ij}^T)^-$
        $\boldsymbol{G}_i^{(d)} \mathrel{+}= (\boldsymbol{R}_{ij}\boldsymbol{G}_j\boldsymbol{S}_{ij}^T)^- + \boldsymbol{G}_i(\boldsymbol{S}_{ij}\boldsymbol{G}_j^T\boldsymbol{G}_j\boldsymbol{S}_{ij}^T)^+$
        $\boldsymbol{G}_j^{(e)} \mathrel{+}= (\boldsymbol{R}_{ij}^T\boldsymbol{G}_i\boldsymbol{S}_{ij})^+ + \boldsymbol{G}_j(\boldsymbol{S}_{ij}^T\boldsymbol{G}_i^T\boldsymbol{G}_i\boldsymbol{S}_{ij})^-$
        $\boldsymbol{G}_j^{(d)} \mathrel{+}= (\boldsymbol{R}_{ij}^T\boldsymbol{G}_i\boldsymbol{S}_{ij})^- + \boldsymbol{G}_j(\boldsymbol{S}_{ij}^T\boldsymbol{G}_i^T\boldsymbol{G}_i\boldsymbol{S}_{ij})^+ \qquad (6.13)$
   f. For $t = 1, 2, \dots, \max_i t_i$:
        $\boldsymbol{G}_i^{(e)} \mathrel{+}= [\boldsymbol{\Theta}_i^{(t)}]^-\boldsymbol{G}_i \quad \text{for } i = 1, 2, \dots, r$
        $\boldsymbol{G}_i^{(d)} \mathrel{+}= [\boldsymbol{\Theta}_i^{(t)}]^+\boldsymbol{G}_i \quad \text{for } i = 1, 2, \dots, r \qquad (6.14)$
   g. Construct $\boldsymbol{G}$ as:
        $\boldsymbol{G} \leftarrow \boldsymbol{G} \circ Diag(\sqrt{\frac{\boldsymbol{G}_1^{(e)}}{\boldsymbol{G}_1^{(d)}}}, \sqrt{\frac{\boldsymbol{G}_2^{(e)}}{\boldsymbol{G}_2^{(d)}}}, \dots, \sqrt{\frac{\boldsymbol{G}_r^{(e)}}{\boldsymbol{G}_r^{(d)}}}), \qquad (6.15)$
      where $\circ$ denotes the Hadamard product. The $\sqrt{\cdot}$ and $\frac{\cdot}{\cdot}$ are entry-wise operations.
```

6.2.5 Parameter estimation

Parameters to DFMF algorithm are factorization ranks, $k_1, k_2, \ldots, k_r$. These are chosen from a predefined interval of possible rank values such that their choice maximizes the estimated quality of the model. To reduce the number of required factorization runs we mimic the bisection method by first testing rank values at the midpoint and borders of specified ranges and then for each rank value selecting the subinterval for which the resulting model was of higher quality. We evaluate the models through the explained variance, the residual sum of squares (RSS) and a measure based on the cophenetic correlation coefficient ρ (Brunet et al., 2004). We compute these measures for the target relation matrix. The RSS is computed over observed associations (o_p^i, o_q^j) in $\boldsymbol{R}_{ij}$ as $\mathrm{RSS}(\boldsymbol{R}_{ij}) = \sum \left[(\boldsymbol{R}_{ij} - \boldsymbol{G}_i \boldsymbol{S}_{ij} \boldsymbol{G}_j^T)(p, q)\right]^2$. Similarly, explained variance is $R^2(\boldsymbol{R}_{ij}) = 1 - \mathrm{RSS}(\boldsymbol{R}_{ij}) / \sum [\boldsymbol{R}_{ij}(p, q)]^2$.

We assess the three quality scores through internal cross-validation and observe how $R^2(\boldsymbol{R}_{ij})$, $\mathrm{RSS}(\boldsymbol{R}_{ij})$ and $\rho(\boldsymbol{R}_{ij})$ vary with changes of factorization ranks. We select ranks $k_1, k_2, \ldots, k_r$ where the cophenetic coefficient begins to fall, the explained variance is high and the RSS curve shows an inflection point (Hutchins et al., 2008).

6.2.6 Prediction from matrix factors

The approximate relation matrix $\widehat{\boldsymbol{R}}_{ij}$ for the target pair of object types $\mathcal{E}_i$ and $\mathcal{E}_j$ is reconstructed as $\widehat{\boldsymbol{R}}_{ij} = \boldsymbol{G}_i \boldsymbol{S}_{ij} \boldsymbol{G}_j^T$. When the model is requested to propose relations for a new object $o_{n_i+1}^i$ of type $\mathcal{E}_i$ that was not included in the training data, we need to estimate its factorized representation and use the resulting factors for prediction. We formulate a non-negative linear least-squares and solve it with an efficient interior point Newton-like method (Van Benthem and Keenan, 2004) for $\min_{x_l \geq 0} ||(\boldsymbol{G}_l \boldsymbol{S}_{li} + \boldsymbol{G}_l \boldsymbol{S}_{il}^T)\boldsymbol{x}_l - \boldsymbol{o}_{n_i+1}^{i,l}||_2^2$, where $\boldsymbol{o}_{n_i+1}^{i,l} \in \mathbb{R}^{n_l}$ is the original description of object $o_{n_i+1}^i$ (if available) and $\boldsymbol{x}_l \in \mathbb{R}^{k_i}$ is its factorized representation (for $l = 1, 2, \ldots, r$ and $l \neq i$). A solution vector given by $\sum_l \boldsymbol{x}_l^{*T}$ is added to $\boldsymbol{G}_i$ and a new $\widehat{\boldsymbol{R}}_{ij} \in \mathbb{R}^{(n_i+1) \times n_j}$ is computed.

We would like to identify object pairs (o_p^i, o_q^j) for which the predicted degree of relation $\widehat{\boldsymbol{R}}_{ij}(p, q)$ is unusually high. We are interested in candidate pairs (o_p^i, o_q^j) for which the estimated association score $\hat{\boldsymbol{R}}_{ij}(p,q)$ is greater than the mean estimated score of all known relations of o_p^i:

$$\hat{\boldsymbol{R}}_{ij}(p,q) > \frac{1}{|\mathscr{A}(o_p^i, \mathscr{E}_j)|} \sum_{o_m^j \in \mathscr{A}(o_p^i, \mathscr{E}_j)} \hat{\boldsymbol{R}}_{ij}(p,m), \tag{6.16}$$

where $\mathscr{A}(o_p^i, \mathscr{E}_j)$ is the set of all objects of $\mathscr{E}_j$ related to o_p^i. Notice that this rule is row-centric, that is, given an object of type $\mathscr{E}_i$, it searches for objects of the other type ($\mathscr{E}_j$) that it could be related to. We can modify the rule to become column-centric, or even combine the two rules.

For example, let us consider that we are studying disease predispositions for a set of patients. Let the patients be objects of type $\mathscr{E}_i$ and diseases objects of type $\mathscr{E}_j$. A patient-centric rule would consider a patient and his medical history and through Eq. (6.16) propose a set of new disease associations. A disease-centric rule would instead consider all patients already associated with a specific disease and identify other patients with a sufficiently high association score.

We can combine row-centric and column-centric approaches. For example, we can first apply a row-centric approach to identify candidates of type $\mathscr{E}_i$ and then estimate the strength of association to a specific object o_q^j by reporting an inverse percentile of association score in the distribution of scores for all true associations of o_q^j, that is, by considering the scores in the q-ed column of $\hat{\boldsymbol{R}}_{ij}$. In our gene function prediction study, we use row-centric approach for candidate identification and column-centric approach for association scoring, and in the experiment from cheminformatics we apply row-centric approach to both tasks.

6.2.7 An ensemble approach to prediction

Different initializations of $\boldsymbol{G}_i$ may in practice give rise to different factorizations of the fusion system. To leverage this effect we construct an ensemble of factorization models. The resulting matrix factors in each model may also be different due to small random perturbations of selected factorization ranks. We use each factorization system for inference of associations (Sec. 6.2.6) and then select the candidate pair through a majority vote. That is, the rule from Eq. (6.16) must apply in more than one half of factorized systems of the ensemble. Ensembles improved the predictive accuracy and stability of the factorized system and the robustness of the results. In our experiments the ensembles combined 15 factorization models.

6.2.8 Matrix factor initialization

The inference of the factorized system in Sec. 6.2.1 is sensitive to the initialization of factor $\boldsymbol{G}$. Proper initialization sidesteps the issue of local convergence and reduces the number of iterations needed to obtain matrix factors of equal quality. We initialize $\boldsymbol{G}$ by separately initializing each $\boldsymbol{G}_i$, using algorithms for single-matrix factorization. Factors $\boldsymbol{S}$ are computed from $\boldsymbol{G}$ (Algorithm 3) and do not require initialization.

Wang et al. (2008) and several other authors (Lee and Seung, 2000) use simple random initialization. Other more informed initialization algorithms include random C (Albright et al., 2006), random Acol (Albright et al., 2006), non-negative double SVD and its variants (Boutsidis and Gallopoulos, 2008), and k-means clustering or relaxed SVD-centroid initialization (Albright et al., 2006). We show that the latter approaches are indeed better over a random initialization (Sec. 7.7). We use random Acol in our case study. Random Acol computes each column of $\boldsymbol{G}_i$ as an element-wise average of a random subset of columns in $\boldsymbol{R}_{ij}$.

6.3 Related work on data integration and latent factor models

Approximate matrix factorization estimates a data matrix $\boldsymbol{R}$ as a product of low-rank matrix factors that are found by solving an optimization problem. In two-factor decomposition, $\boldsymbol{R} \in \mathbb{R}^{n\times m}$ is decomposed to a product $\boldsymbol{W}\boldsymbol{H}$, where $\boldsymbol{W} \in \mathbb{R}^{n\times k}$, $\boldsymbol{H} \in \mathbb{R}^{k\times m}$ and $k \ll \min(n, m)$. A large class of matrix factorization algorithms minimize discrepancy between the observed matrix and its low-rank approximation, such that $\boldsymbol{R} \approx \boldsymbol{W}\boldsymbol{H}$. For instance, SVD, non-negative matrix factorization and exponential family PCA all minimize Bregman divergence (Singh and Gordon, 2008b).

Although often used in data analysis for dimensionality reduction, clustering or low-rank approximation, there have been only a few applications of matrix factorization in data fusion. Lange and Buhmann (2005) proposed an integration by non-negative matrix factorization of a target matrix, which was a convex combination of similarity matrices obtained from multiple information sources. Their work is similar to that of Wang et al. (2012), who applied non-negative matrix tri-factorization with input matrix completion. Note that both approaches implement early integration and can model only multiple dyadic relations. Their approaches cannot be used to model relations between more than two object types, which is a major distinction with the algorithm proposed in this chapter.

Zhang et al. (2012) proposed a joint matrix factorization to decompose a number of data matrices $\boldsymbol{R}_i$ into a common basis matrix $\boldsymbol{W}$ and different coefficient matrices $\boldsymbol{H}_i$, such that $\boldsymbol{R}_i \approx \boldsymbol{W}\boldsymbol{H}_i$ by minimizing $\sum_i ||\boldsymbol{R}_i - \boldsymbol{W}\boldsymbol{H}_i||^2_{\text{Fro}}$. This is an intermediate integration approach with different data sources but it can describe only relations whose objects (*i.e.* rows in $\boldsymbol{R}_i$) are fixed across relation matrices. Similar approaches but with various regularization types were also proposed, such as network- or relation-regularized constraints (Li and Yeung, 2007; Zhang et al., 2011b) and hierarchical priors (Singh and Gordon, 2008a, 2010). Our work generalizes these approaches by simultaneously dealing with objects of different types, where we can vary object types along both dimensions of relation matrices, $\boldsymbol{R}_{ij}$) and can constrain objects of every type.

There is an abundance of work on matrix factorization models that consider a single dyadic relation matrix or multiple relation matrices between the same two types of objects (Wang et al., 2008; Sutskever, 2009; Li et al., 2009a; Singh and Gordon, 2010; Zhang et al., 2011b; Wang et al., 2012) that are subsumed in our approach. For instance, Nickel et al. (2011) proposed a tri-factorization model for multiple dyadic relations that factorized every $\boldsymbol{R}_i$ as $\boldsymbol{R}_i \approx \boldsymbol{A}\boldsymbol{S}_i\boldsymbol{A}^T$. Although their model is appropriate for certain tasks of collective learning, all $\boldsymbol{R}_i$ describe relations between the same two sets of objects, whereas our approach models multi-relational and multi-object type data.

Rettinger et al. (2012) proposed context-aware tensor decomposition for relation prediction in social networks, CARTD. They decompose a tensor into additive factorized matrices using two-factor decomposition. They assume that input data is provided together with the contextual information that describes one specific relation, the recommendation. The drawback of their and similar approaches (Kolda and Bader, 2009; Sutskever, 2009; Rendle et al., 2011) for r-ary tensors is that in higher dimensions

($r > 3$) the tensors become increasingly sparse and the computational requirements become infeasible. Notice that here r corresponds to number of different object types in DFMF. In comparison, the approach proposed in this chapter can handle tens of different object types.

Wang et al. (2008) and Wang et al. (2011a) proposed tri-SPMF and S-NMTF, respectively, a simultaneous clustering of multi-type relational data via symmetric non-negative matrix tri-factorization. These two methods are conceptually similar to our approach and use both inter-type and intra-type relations, but they require a full set of symmetric relation matrices, $\boldsymbol{R}_{ij} = \boldsymbol{R}_{ji}^T$. These assumptions of tri-SPMF and S-NMTF are rarely met in real-world fusion scenarios (see, for example, a fusion configuration from Fig. 7.1, which is not a 6-clique), where we do not have access to relation matrices between all possible pairs of object types (*i.e.* $\boldsymbol{R}_{ij}$ for $1 \leq i < j \leq r$). The tri-SPMF and S-NMTF algorithms do not converge to a local minimum if described relations are asymmetric ($\boldsymbol{R}_{ij} \neq \boldsymbol{R}_{ji}^T$).

We are currently witnessing increasing interest in the joint treatment of heterogeneous data sets and the emergence of approaches specifically designed for data fusion. Besides matrix factorization-based methods as reviewed above, these approaches include canonical correlation analysis (Chaudhuri et al., 2009), combining many interaction networks into a composite network (Mostafavi and Morris, 2012), multiple graph clustering with linked matrix factorization (Tang et al., 2009), a mixture of Markov chains associated with different graphs (Zhou and Burges, 2007), dependency-seeking clustering algorithms with variational Bayes (Klami and Kaski, 2008), latent factor analysis (Lopes et al., 2011; Luttinen and Ilin, 2009), nonparametric Bayes ensemble learning (Xing and Dunson, 2011), approaches based on Bayesian theory (Zhang and Ji, 2006; Alexeyenko and Sonnhammer, 2009; Huttenhower et al., 2009), neural networks (Carpenter et al., 2005), and module guided random forests (Chen and Zhang, 2013).

Data integration approaches from the previous paragraph either fuse input data (early integration) or predictions (late integration) and do not directly combine heterogeneous representation of objects of different types. A state-of-the-art approach that can address such data through intermediate integration is kernel-based learning. Multiple kernel learning (MKL) has been pioneered by (Lanckriet et al., 2004a) and (Bach et al., 2004) and is an additive extension of single kernel SVM to incorporate multiple kernels in classification, regression and clustering. The MKL assumes that $\mathscr{E}_1, \dots, \mathscr{E}_r$ are r different representations of the same set of n objects. Extension from single to multiple data sources is achieved by additive combination of kernel matrices, given by $\Omega = \left\{ \sum_{i=1}^{r} \theta_i \boldsymbol{K}_i \middle| \forall i : \theta_i \geq 0, \sum_{i=1}^{r} \theta_i^{\delta} = 1, \boldsymbol{K}_i \succeq 0 \right\}$, where θ_i are weights of the kernel matrices, δ is a parameter determining the norm of constraint posed on coefficients (for L_2, L_p-norm MKL, see (Kloft et al., 2009, 2011; Yu et al., 2010, 2012)) and $\boldsymbol{K}_i$ are normalized kernel matrices centered in the Hilbert space. Among other improvements, (Yu et al., 2010) extended the framework of the MKL in (Lanckriet et al., 2004a) by optimizing various norms in the dual problem of SVMs that allows non-sparse optimal kernel coefficients θ_i^*. (Gönen and Alpaydın, 2011) recently reviewed several MKL algorithms and concluded that, in general, using multiple kernels instead of a single one is useful. The heterogeneity of data sources in the MKL is resolved by transforming different object types and data structures (e.g., strings, vectors, graphs) into kernel matrices. These transformations depend on the choice of the kernels, which in turn affects the method's performance (Debnath and Takahashi, 2004).

7

Case study: functional genomics and pharmacology

We present two case studies from bioinformatics and cheminformatics, where recent technological advancements have allowed researchers to collect large and diverse experimental data sets (Parikh and Polikar, 2007; Pandey et al., 2010; Savage et al., 2010; Xing and Dunson, 2011). From bioinformatics, we study prediction of gene function, where the target relation is given by a binary matrix representing relationships between genes of the amoeba *Dictyostelium discoideum* and their associated functions or processes (Gene Ontology (GO) terms, $\boldsymbol{R}_{12}$). In the cheminformatics study, the binary target matrix encodes the pharmacologic actions of a subset of chemicals from PubChem database. We apply DFMF to fuse eleven data matrices for gene function prediction and six data matrices for the prediction of pharmacologic actions. During testing, we estimate the relation for a previously-unseen pair (Gene, GO Term) or (Chemical, Pharmacologic Action).

We compare our collective matrix factorization model DFMF to an early integration by random forests (Breiman, 2001; Boulesteix et al., 2008), intermediate integration by multiple kernel learning (MKL) (Yu et al., 2010) and relational learning by matrix factorization (tri-SPMF) (Wang et al., 2008). Kernel-based fusion used a multi-class L_2 norm MKL with Vapnik's SVM (Ye et al., 2008). The MKL was formulated as a second order cone program (SOCP) and its dual problem was solved by the conic optimization solver SeDuMi. Random forests from the Orange data mining suite were used with default parameters. Relational learning by tri-SPMF used the matrix factorization algorithm from Wang et al. (2008) and a procedure described in Sec. 6.2.6 for predicting associations.

7.1 Gene function prediction task

Various classification schemes were developed to standardize the association of genes to its function. Of these, Gene Ontology (GO) (Ashburner et al., 2000) is adopted widely and is thus suitable for computational studies (Mostafavi and Morris, 2012; Radivojac et al., 2013). In our study, given a gene, we aimed to predict a set of its associated GO terms along with the confidence of the association.

7.1.1 Data

We observed six object types (Fig. 7.1): genes (type 1), ontology terms (type 2), experimental conditions (type 3), publications from the PubMed database (PMID) (type 4), Medical Subject Headings (MeSH) descriptors (type 5), and KEGG pathways (Kanehisa et al., 2014) (type 6). The data included gene expression measured during different time-points of a 24-hour development cycle (Parikh et al., 2010) ($\boldsymbol{R}_{13}$, 14 experimental conditions), gene annotations with experimental evidence code to 148 generic slim terms from the GO ($\boldsymbol{R}_{12}$), PMIDs and their associated *D. discoideum* genes from dictyBase ($\boldsymbol{R}_{14}$), genes participating in KEGG pathways ($\boldsymbol{R}_{16}$), assignments of MeSH descriptors to publications from PubMed ($\boldsymbol{R}_{45}$), references to published work on associations between a specific GO term and gene product ($\boldsymbol{R}_{42}$), and associations of enzymes involved in KEGG pathways and related to GO terms ($\boldsymbol{R}_{62}$).

To balance $\boldsymbol{R}_{12}$, our target relation matrix, we added an equal number of non-associations for which there is no evidence of any type in the GO. We constrained our system by considering gene interaction scores from STRING v9.0 ($\boldsymbol{\Theta}_1$) and slim term similarity scores ($\boldsymbol{\Theta}_2$) computed as -0.2^{hops}, where $hops$ was the length of the shortest path between two terms in the GO graph. Similarly, MeSH descriptors were constrained with the average number of hops in the MeSH hierarchy between each pair of descriptors ($\boldsymbol{\Theta}_5$). Constraints between KEGG pathways corresponded to the number of common ortholog groups ($\boldsymbol{\Theta}_6$). The slim subset of GO terms was used to limit the optimization complexity of the MKL and the number of variables in the SOCP, and to ease the computational burden of early integration by random forests, which inferred a separate model for each of the terms.

We conducted three experiments in which we considered either 100 or 1000 most GO-annotated genes or the whole *D. discoideum* genome ($\sim$12,000 genes). We also examined the predictions of gene associations with any of nine GO terms that are of specific relevance to the current research in the *Dictyostelium* community (upon consultations with Gad Shaulsky, Baylor College of Medicine, Houston, TX; see Table 7.2). Instead of using a generic slim subset of terms, we examined the predictions in the context of a complete set of GO terms. This resulted in a data set with $\sim 2,000$ terms, each term having $\sim$10 direct gene annotations.

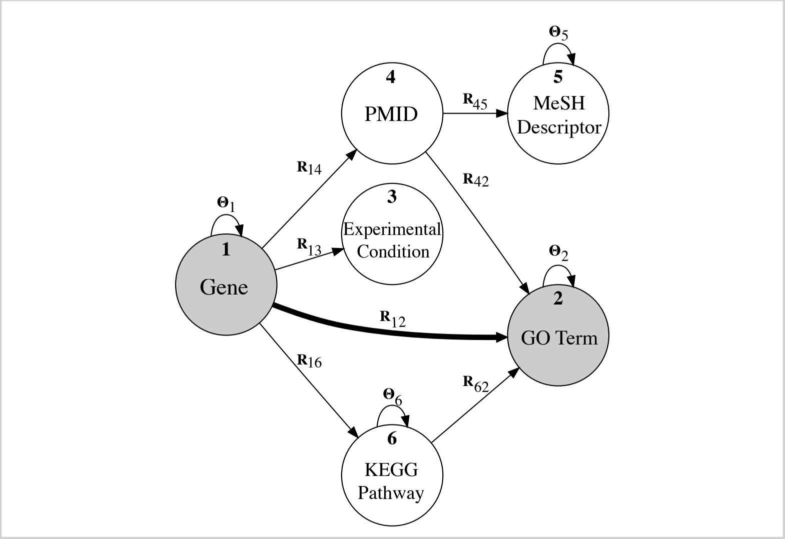


Figure 7.1

The fusion configuration for gene function prediction task in *D. discoideum*. Some relations are entirely missing, for instance $\boldsymbol{R}_{23}$. Nodes represent object types used in our study. Edges correspond to relation and constraint matrices. The arc that represents the target matrix $\boldsymbol{R}_{12}$ and its object types are highlighted.

7.1.2 Preprocessing for kernel-based fusion

We generated an RBF kernel for gene expression measurements from $\boldsymbol{R}_{13}$ with the RBF function $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-||\boldsymbol{x}_i - \boldsymbol{x}_j||^2/2\sigma^2)$, and a linear kernel for $[0,1]$-protein-interaction matrix from $\boldsymbol{\Theta}_1$. This particular choice of kernels was motivated by the experimental study and kernel comparison in (Lanckriet et al., 2004c). Kernels were applied to data matrices. We used a linear kernel to generate a kernel matrix from *D. discoideum* specific genes that participate in pathways ($\boldsymbol{R}_{16}$), and a kernel matrix from PMIDs and their associated genes ($\boldsymbol{R}_{14}$). Several data sources describe relations between object types other than genes. For kernel-based fusion we had to transform them to explicitly relate to genes. For instance, to relate genes and MeSH descriptors, we counted the number of publications that were associated with a specific gene ($\boldsymbol{R}_{14}$) and were assigned a specific MeSH descriptor ($\boldsymbol{R}_{45}$, see also Fig. 7.1). A linear kernel was applied to the resulting matrix. Kernel matrices that incorporated relations between KEGG pathways and GO terms ($\boldsymbol{R}_{62}$), and publications and GO terms were obtained in similar fashion.

To represent the hierarchical structure of MeSH descriptors ($\boldsymbol{\Theta}_5$), the semantic structure of the GO graph ($\boldsymbol{\Theta}_2$) and ortholog groups that correspond to KEGG pathways ($\boldsymbol{\Theta}_6$), we considered the genes as nodes in three distinct large weighted graphs. In the graph for $\boldsymbol{\Theta}_5$, the link between two genes was weighted by the similarity of their associated sets of MeSH descriptors using information from $\boldsymbol{R}_{14}$ and $\boldsymbol{R}_{45}$. We considered the MeSH hierarchy to measure these similarities. Similarly, for the graph for $\boldsymbol{\Theta}_2$ we considered the GO semantic structure in computing similarities of sets of GO terms associated with genes. In the graph for $\boldsymbol{\Theta}_6$, the gene edges were weighted by the number of common KEGG ortholog groups. Kernel matrices were constructed with a diffusion kernel (Kondor and Lafferty, 2002).

The resulting kernel matrices $\boldsymbol{K} \in \mathbb{R}^{n \times n}$ were centered as $\boldsymbol{K}^c(i,j) = \boldsymbol{K}(i,j) - 1/n \sum_i \boldsymbol{K}(i,j) - 1/n \sum_j \boldsymbol{K}(i,j) + 1/n^2 \sum_{ij} \boldsymbol{K}(i,j)$ and normalized using the formula $\boldsymbol{K}^n(i,j) = \boldsymbol{K}^c(i,j)/\sqrt{\boldsymbol{K}^c(i,i)\boldsymbol{K}^c(j,j)}$. The parameters for all kernels were selected through internal cross-validation. In cross-validation, only the training part of the matrices was optimized for learning, while centering and normalization were performed on the entire data set. The prediction task was defined through the classification matrix of genes and their associated GO slim terms from $\boldsymbol{R}_{12}$.

7.1.3 Preprocessing for early integration

The gene-related data matrices prepared for kernel-based fusion were also used for early integration and were concatenated into a single data table. Each row in the table represented a gene profile obtained from all available data sources. For our case study, each gene was characterized by a fixed 9,362-dimensional feature vector. For each GO slim term, we then separately developed a classifier with a random forest of classification trees and reported cross-validated results.

7.1.4 Preprocessing for tri-SPMF learning

Relation and constraint matrices prepared for DFMF were also used for tri-SPMF factorization algorithm. Tri-SPMF requires a full set of relation matrices between all pairs of object types. Thus, we used zero matrices for non-existing relations from Fig. 7.1. For instance, $\boldsymbol{R}_{63}$ and $\boldsymbol{\Theta}_4$ were represented by zero matrices of proper dimensions. Because tri-SPMF requires that relations are symmetric, we set $\boldsymbol{R}_{ji} = \boldsymbol{R}_{ij}^T$ for all available relation matrices.

7.2 Pharmacologic action prediction task

Identification of the mechanisms of action of chemical compounds is a crucial task in drug discovery (Paolini et al., 2006; Iorio et al., 2010). Here, our aim was to computationally predict pharmacologic actions of chemical compounds as defined in the PubChem database (Wang et al., 2009).

7.2.1 Data

We considered six object types (Fig. 7.2): chemicals (type 1), PubChem's (Wang et al., 2009) pharmacologic actions (type 2), publications from the PubMed database (PMID) (type 3), depositors of chemical substances (type 4) and their categorization (type 6), and PubChem substructure fingerprints (type 5).

The data included 1,260 chemicals extracted from the complete DrugBank (Law et al., 2014) database (accessed in Feb. 2014) that were identified with at least one pharmacologic action in the PubChem Compound database. In that way, every chemical (drug) was assigned one or more MeSH headings that described its pharmacologic actions and corresponded to D27.505 tree of the 2014 MeSH Tree Structure (target relation $\boldsymbol{R}_{12}$). For example, established pharmacologic actions for Aspirin include "Anti-Inflammatory Agents, Non-Steroidal", "Fibrinolytic Agents" and "Antipyretics." To increase the number of chemicals assigned to a particular pharmacologic action, the actions of the chemical also included those from its direct parents in the D27.505 tree.

Other data considered were publications from the PubMed database ($\boldsymbol{R}_{13}$), data on depositors who submitted substances of the chemicals present in PubChem Compound records ($\boldsymbol{R}_{14}$), categories of data depositors ($\boldsymbol{R}_{46}$) and PubChem substructure fingerprints ($\boldsymbol{R}_{15}$). These fingerprints consist of a series of 881 binary indicators, each denoting the presence or absence of a particular substructure in a molecule. Collectively, these binary keys provide a "fingerprint" of a particular chemical structure form. Chemicals are constrained by a matrix of substructure-based Tanimoto 2D similarity ($\boldsymbol{\Theta}_1$) obtained through PubChem Score Matrix Service.

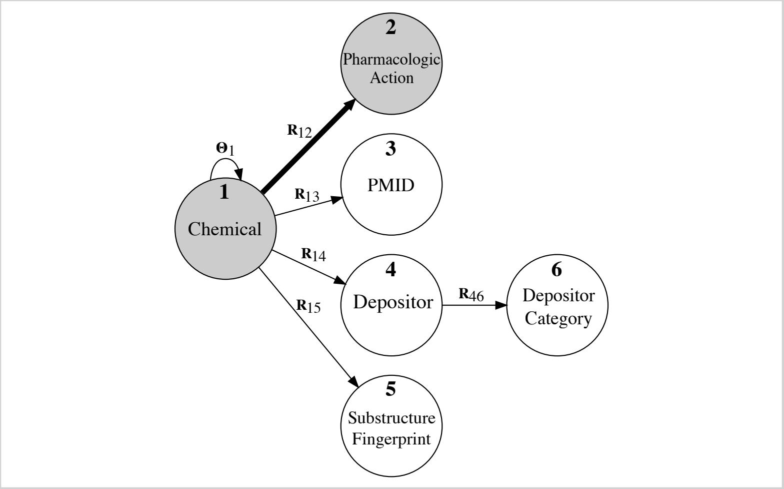


Figure 7.2

The fusion configuration for the prediction of pharmacologic actions of chemicals, with object types denoted with nodes and relations between them with edges. The edge representing the target relation and its corresponding data matrix R_{12} is highlighted.

7.2.2 Preprocessing for alternative learning methods

For the kernel-based fusion, we generated the kernel matrices for chemicals from $\boldsymbol{R}_{13}$, $\boldsymbol{R}_{14}$, $\boldsymbol{R}_{15}$ and $\boldsymbol{\Theta}_1$ (Fig. 7.2) using the polynomial kernel of degree 2. We included data on depositors ($\boldsymbol{R}_{46}$) by applying a polynomial kernel to $\boldsymbol{R}_{14}\boldsymbol{R}_{46}$. The resulting kernel matrices were centered and normalized, and the kernel parameters were selected in internal cross-validation (see Sec. 7.1.2 for details). Preprocessing for early integration by random forests and tri-SPMF learning followed the same procedures as described in Sec. 7.1.3 and Sec. 7.1.4, respectively. The prediction task was defined by the associations of chemicals to pharmacologic actions given by $\boldsymbol{R}_{12}$ (Fig. 7.2).

7.3 Scoring

We estimated the quality of inferred models by ten-fold cross-validation. In each iteration, we split the set of genes (chemicals) to a train and test set. The corresponding data on genes (chemicals) from the test set was entirely omitted from the training data. We developed prediction models from the training data and tested them on the genes (chemicals) from the test set. The performance was evaluated using an F_1 score, a harmonic mean of precision and recall, and area under ROC curve (AUC). Both scores were averaged across cross-validation runs.

7.4 Predictive performance

Table 7.1 presents the cross-validated F_1 and AUC scores for both gene function prediction (data set of slim GO terms) and prediction of pharmacologic actions. The accuracy of DFMF is at least comparable to MKL and substantially higher than that of early integration by random forests and relational learning by tri-SPMF. When more genes and hence more data were considered for the gene function prediction the performance of all four fusion approaches improved.

Poorer performance of tri-SPMF was most probably due to required introduction of relations into factorized system that were not present in the input data. Consequently, the ability of tri-SPMF to infer relations of interest between other object types deteriorated considerably. Notice also that tri-SPMF could not be applied if fusion schemes in Figs. 7.1 or 7.2 would contain asymmetric or one-way relations, such as those from the analysis of signed networks (Leskovec et al., 2010) and computational biology (Notebaart et al., 2009), among others. We also observed numerical instability with tri-SPMF, which was exhibited as an increase in the value of objective function between successive iterations. In contrast, DFMF exhibited numerical stability in all experiments (results not shown).

The accuracy for nine GO terms selected by domain expert is given in Table 7.2. The DFMF performs consistently better than the other three approaches. Again, the early integration by random forests is inferior to all three intermediate integration methods. Notice that, with only a few exceptions, both F_1 and AUC scores of DFMF are high. This is important, as all nine gene processes and functions observed are relevant for current research of *D. discoideum* where the methods for data fusion can yield new candidate genes for focused experimental studies.

Our fusion approach is faster than multiple kernel learning. DFMF required 18 minutes of runtime on a standard desktop computer compared to 77 minutes for MKL to finish one iteration of cross-validation of the whole-genome variant of gene function prediction task. The factorization algorithm of DFMF also took less time to execute

Table 7.1

Cross-validated F_1 and AUC accuracy scores for fusion by matrix factorization (DFMF), kernel-based method (MKL), random forests (RF) and relational learning-based matrix factorization (tri-SPMF).

Prediction task	DFMF		MKL		RF		tri-SPMF	
	F_1	AUC	F_1	AUC	F_1	AUC	F_1	AUC
100 *D. discoideum* genes	0.799	0.801	0.781	0.788	0.761	0.785	0.731	0.724
1000 *D. discoideum* genes	0.826	0.823	0.787	0.798	0.767	0.788	0.756	0.741
D. discoideum genome	0.831	0.849	0.800	0.821	0.782	0.801	0.778	0.787
Pharmacologic actions	0.663	0.834	0.639	0.811	0.643	0.819	0.641	0.810

Table 7.2

Gene Ontology term-specific cross-validated F_1 and AUC accuracy scores for fusion by matrix factorization (DFMF), kernel-based method (MKL), random forests (RF) and relational learning-based matrix factorization (tri-SPMF). Seq.-spec. DNA TFA, sequence-specific transcription factor activity; Activation of ACA, activation of adenylate cyclase activity.

GO term name	Term identifier	Size	DFMF		MKL		RF		tri-SPMF	
			F_1	AUC	F_1	AUC	F_1	AUC	F_1	AUC
Activation of ACA	0007190	11	0.834	0.844	0.770	0.781	0.758	0.601	0.729	0.731
Chemotaxis	0006935	58	0.981	0.980	0.794	0.786	0.538	0.724	0.804	0.810
Chemotaxis to cAMP	0043327	21	0.922	0.910	0.835	0.862	0.798	0.767	0.838	0.815
Phagocytosis	0006909	33	0.956	0.932	0.892	0.901	0.789	0.619	0.836	0.810
Response to bacterium	0009617	51	0.899	0.870	0.788	0.761	0.785	0.761	0.817	0.831
Cell-cell adhesion	0016337	14	0.883	0.861	0.867	0.856	0.728	0.725	0.799	0.834
Actin binding	0003779	43	0.676	0.781	0.664	0.658	0.642	0.737	0.671	0.682
Lysozyme activity	0003796	4	0.782	0.750	0.774	0.750	0.754	0.625	0.747	0.625
Seq.-spec. DNA TFA	0003700	79	0.956	0.948	0.894	0.901	0.732	0.759	0.892	0.852

than tri-SPMF due to redundant representation of fusion system required by tri-SPMF.

7.5 Sensitivity to inclusion of data sources

Inclusion of additional data sources improves the accuracy of prediction models. We illustrate this for gene function prediction in Fig. 7.3, where we started with only the target data source $\boldsymbol{R}_{12}$ and then added either $\boldsymbol{R}_{13}$ or $\boldsymbol{\Theta}_1$ or both. Similar effects were observed when we studied other combinations of data sources (not shown here for brevity). Notice also that due to ensembling the cross-validated variance of F_1 is small.

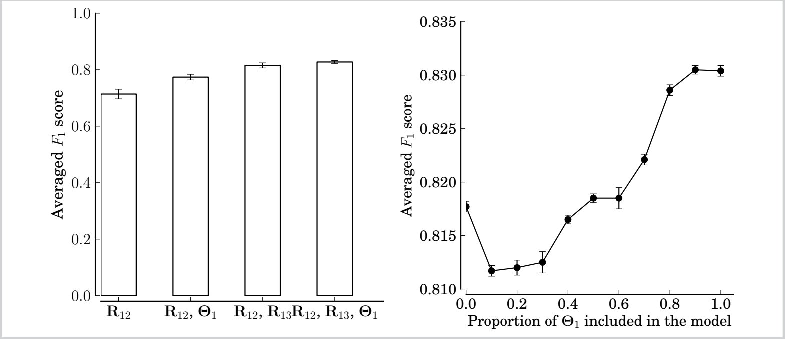


Figure 7.3

Adding new data sources (left) or incorporating more object-type-specific constraints in Θ_1 (right) both increase the accuracy of matrix factorization-based models for gene function prediction task.

7.6 Sensitivity to inclusion of constraints

We varied the sparseness of gene constraint matrix $\mathbf{\Theta}_1$ by holding out a random subset of protein-protein interactions. We set the entries of $\mathbf{\Theta}_1$ that corresponded to held-out constraints to zero so that they did not affect the cost function during optimization. Fig. 7.3 shows that including additional information on genes in the form of constraints improves the predictive performance of DFMF for gene function prediction.

7.7 Matrix factor initialization study

We studied the effect of matrix factor initialization on DFMF by observing the reconstruction error after one and after twenty iterations of optimization procedure, the latter being about one fourth of the iterations required for the optimization algorithm to converge when predicting gene functions. We estimated the error relative to the optimal $(k_1, k_2, \dots, k_6)$-rank approximation given by the SVD. For iteration v and matrix $\boldsymbol{R}_{ij}$ the error was computed by:

$$\text{Err}_{ij}(v) = \frac{||\boldsymbol{R}_{ij} - \boldsymbol{G}_i^{(v)} \boldsymbol{S}_{ij}^{(v)} (\boldsymbol{G}_j^T)^{(v)}||^2 - d_F(\boldsymbol{R}_{ij}, [\boldsymbol{R}_{ij}]_k)}{d_F(\boldsymbol{R}_{ij}, [\boldsymbol{R}_{ij}]_k)}, \qquad (7.1)$$

where $\boldsymbol{G}_i^{(v)}$, $\boldsymbol{G}_j^{(v)}$ and $\boldsymbol{S}_{ij}^{(v)}$ were matrix factors obtained after v iterations of factorization algorithm. In Eq. (7.1), $d_F(\boldsymbol{R}_{ij}, [\boldsymbol{R}_{ij}]_k) = ||\boldsymbol{R}_{ij} - \boldsymbol{U}_k \boldsymbol{\Sigma}_k \boldsymbol{V}_k^T||^2$ denotes the Frobenius distance between $\boldsymbol{R}_{ij}$ and its k-rank approximation given by the SVD, where

Table 7.3

Effect of initialization algorithm on reconstruction error of DFMF's factorization model.

Method	Time $\boldsymbol{G}^{(0)}$	Storage $\boldsymbol{G}^{(0)}$	$\mathrm{Err}_{12}(1)$	$\mathrm{Err}_{12}(20)$
Rand.	0.0011 s	618K	5.11	3.61
Rand. C	0.1027 s	553K	2.97	1.67
Rand. Acol	0.0654 s	505K	1.59	1.30
K-means	0.4029 s	562K	2.47	2.20
NNDSVDa	0.1193 s	562K	3.50	2.01

$k = \max(k_i, k_j)$ is the approximation rank. $\mathrm{Err}_{ij}(v)$ is a pessimistic measure of quantitative accuracy because of the choice of k. This error measure is similar to the error of the two-factor non-negative matrix factorization from (Albright et al., 2006).

Table 7.3 shows the results for the experiment with 1000 most GO-annotated *D. discoideum* genes and selected factorization ranks $k_i < 65$, $i \in [6]$. The informed initialization algorithms surpass the random initialization. Of these, the random Acol algorithm performs best in terms of accuracy and is also one of the simplest.

7.8 Early integration by matrix factorization

Our data fusion approach simultaneously factorizes individual blocks of data in $\boldsymbol{R}$. Alternatively, we could also disregard the data structure, and treat $\boldsymbol{R}$ as a single data matrix. Such data treatment would transform our data fusion approach to that of early integration and lose the benefits of structured system and source-specific factorization. To prove this experimentally, we considered the 1,000 most GO-annotated *D. discoideum* genes. The resulting cross-validated F_1 score for factorization-based early integration was 0.576, compared to 0.826 obtained with our proposed data fusion algorithm. This result is not surprising as neglecting the structure of the system also causes the loss of the structure in matrix factors and the loss of zero blocks in factors $\boldsymbol{S}$ and $\boldsymbol{G}$ from Eq. (6.3). Clearly, data structure carries substantial information and should be retained in the model.

Conclusion

8

We have described a new matrix factorization-based data fusion algorithm called DFMF. The approach is flexible and, in contrast to state-of-the-art kernel-based methods, requires minimal, if any, preprocessing of input data. This latter feature, the ability to model multi-relational and multi-object type data, and DFMF's excellent accuracy and time response, are the principal advantages of our new algorithm.

DFMF can model any collection of data sets, each of which can be expressed as a matrix. Tasks from bioinformatics and cheminformatics considered here that were traditionally regarded as classification problems exemplify just one type of data mining problems that can be addressed with our method. We anticipate the utility of factorization-based data fusion in multi-task learning, association mining, clustering, link prediction or structured output prediction.

Part IV

Latent chaining and profiling

Gene prioritization

9

In everyday life, we make decisions by considering all the available information, and often find that inclusion of even seemingly circumstantial evidence provides an advantage. Our new computational method Collage prioritizes genes from a large collection of heterogeneous data. In a case study on social amoeba *Dictyostelium*, we started from four bacterial response genes and 14 different data sets ranging from gene expression to pathway and literature information. Collage proposed eight candidate genes that were tested in the wet lab. Mutations in all eight candidates reduced the ability of the amoebae to grow on Gram-negative bacteria. This is a remarkably accurate result since only about a hundred of the 12,000 *Dictyostelium* genes are estimated to be responsible for bacterial response.

Data integration procedures combine heterogeneous data sets into predictive models, but they are limited to data explicitly related to the target object types, such as genes. Collage is a new data fusion approach to gene prioritization. It considers data sets of various association levels with the prediction task, utilizes collective matrix factorization to compress the data, and chaining to relate different object types in the data. Collage prioritizes genes based on their similarity to several seed genes. We tested Collage by prioritizing bacterial-response genes in *Dictyostelium* as a novel model system for prokaryote-eukaryote interactions. Using 4 seed genes and 14 data sets, only one of which was directly related to bacterial responses, Collage proposed 8 candidate genes that were readily validated as necessary for the response of *Dictyostelium* to Gram-negative bacteria. These findings establish Collage as a method for inferring biological knowledge from the integration of heterogeneous and coarsely related data sets.

9.1 Background

In the natural sciences, incorporating all the data, especially circumstantial information, can be conceptually and computationally challenging. The difficulty stems from the heterogeneity and abundance of data sets. Consider a typical data analysis task in molecular biology: besides experimental data, such as levels of gene expression, there are plenty of other data sets at our disposal, such as protein-protein binding sites, genetic and metabolic pathways, functional annotations, genetic interactions, phenotype ontologies, diseases, drugs and their side effects. Intuitively, collective mining of all available information sources should improve accuracy of predictive modeling. However, the challenges are to integrate seemingly unrelated concepts from heterogeneous data sets (Ormrod, 2011) and fuse various data sets into a single predictive model.

Here we present a method called Collage that can consider a large number of potentially indirectly related data sets and use them for gene prioritization. Computational prediction of gene function is a formidable challenge. Given a small set of seed genes that are known to be responsible for a particular function, gene prioritization (Moreau and Tranchevent, 2012) aims to identify the most promising candidates for further studies. Present data integration approaches for gene prioritization can be divided into four groups: methods that consecutively filter one data set at a time (Franke et al., 2004); methods that stitch together gene profiles from different data sources and then treat the stitched parts equally (Sifrim et al., 2013); methods that use each data set separately to estimate the similarity of candidates to the seed genes and then fuse similarity scores through weighting (Lanckriet et al., 2004b; Aerts et al., 2006; De Bie et al., 2007; Sun et al., 2009; Chen et al., 2009; Yu et al., 2010; Fontaine et al., 2011; Schlicker et al., 2010); and methods that construct gene correlation networks independently from each data set and find genes that are similar to the seed genes in the composite network (Sharma et al., 2010; Köhler et al., 2008; Mostafavi et al., 2008; Mostafavi and Morris, 2012; Wang et al., 2014).

These approaches are limited to data that *explicitly* refer to genes. They cannot readily treat data that are relevant for gene prioritization but are provided in a non-gene data space, such as disease ontologies, phenotype classifications, drug interactions and annotations of small chemicals. A labor-intensive approach to consider data from non-gene space is feature engineering, which transforms circumstantial data into gene profiles. However, feature engineering is neither standardized nor effortless and is a bottleneck that prevents the implementation of truly large-scale data fusion for gene prioritization. As an alternative to gene-centric approaches, Collage represents a major advancement in (i) the breadth of data it can incorporate, (ii) the ease of data integration without complex feature engineering, (iii) the high prediction accuracy, (iv) the ability to retain the relational structure both within and between data sets during model inference and (v) the capacity to incorporate knowledge of data structure in model design.

We used Collage to solve a problem in an exciting and relatively new field of interest – the use of *Dictyostelium* as a model system to explore the interaction between eukaryotes and prokaryotes. *D. discoideum* is a free-living soil amoeba that feeds on bacteria. The amoebae eat both Gram-negative and Gram-positive bacteria, but they respond differently to bacteria from these two groups. Early studies have shown that mutations can impair the ability of the amoebae to grow on either Gram-positive or on Gram-negative bacteria (Newell et al., 1977). Other studies have shown that the amoebae can serve as a model for the interaction between eukaryotes and prokaryotes, including pathogenesis (Bozzaro and Eichinger, 2011; Lima et al., 2011; Steinert, 2011). This system is an important addition to the field because *Dictyostelium* is a very convenient model organism that offers a variety of experimental tools, including classical genetics and modern genomic approaches.

The interaction between *D. discoideum* and several Gram-positive and Gram-negative bacteria has recently been explored with genetic and genomic methods (Nasser et al., 2013). These studies revealed transcriptome-level responses to the two bacterial groups and discovered a handful of genes that are essential for growth of amoebae on bacteria. The genetic analysis suggested that one in a hundred of the 12,000 genes in the *D. discoideum* genome is required for bacterial discrimination. Identifying and characterizing these genes is a laborious task that requires several months of work per gene. We hypothesized that Collage could simplify this task by prioritizing genes and suggesting which ones should be tested by direct experiments.

9.2 Gene prioritization by compressive data fusion and chaining

Next, we overview the Collage gene prioritization algorithm. The fundamental building block of Collage is matrix tri-factorization of a single relation matrix (Fig. 9.1). To model a particular relation, tri-factorization decomposes the data matrix into three smaller, low-dimensional latent matrices, whose product should well reconstruct the original matrix. Two latent *recipe matrices* map objects A and B into the latent space, and the remaining *backbone matrix* describes the relations in the latent space. In essence, the backbone matrix is a compressed version of the original data matrix.

We proceed by providing a more detailed overview of gene prioritization algorithm. The entire operation of Collage can be decomposed into four major parts:

Step I: Compressive data fusion – Collage collectively models many data matrices that share object types. We organize the matrices in a data fusion graph. Object types are denoted as nodes (A to G in Fig. 9.2), which may correspond to genes, ontology terms, diseases and patients, etc. Instead of separately tri-factorizing each data matrix, Collage collectively factorizes all the matrices to a set of backbone matrices (edges, matrices in blue, one for each original data matrix) and recipe matrices (nodes, one for each object type), where the recipe matrices are shared across data sets that report on a common object type.

Step II: Chaining of latent matrices – Collage chains latent matrices of the resulting factorized model to profile target objects, e.g., genes, in the latent space of any other object type. For example, Fig. 9.2c shows the profiling of objects A in the latent space C. Object profiles are constructed by chaining that starts at node A and traverses the graph to node C through D and F. Chaining multiplies the recipe matrix A by the backbone matrices along the traversed path. The A-to-C path in Fig. 9.2c is one of nine chains through which we can profile objects A in our exemplar data fusion graph. The nine chains of latent matrices for the exemplar fusion graph from Fig. 9.2a are shown in Fig. 9.2d.

Step III: Similarity estimation – Collage uses the profile matrices obtained by chaining in Step II to estimate similarity between target objects (object type A, genes, in Fig. 9.2) and seed objects. The number of profile vectors for each object of type A corresponds to the number of chains. Collage compares the profiles of candidate genes to the profiles of the seed genes. Given a candidate gene, Collage records its rank correlation-based similarities in a similarity score matrix with seed genes in the columns and chained profiles in the rows (Fig. 9.2e). The final score estimates the similarity of a candidate to a set of seed genes and is obtained by summarizing the similarity score matrix with a single value (green circle) computed by a median-based L-estimator.

Step IV: Gene ranking – The similarity score of a gene is a proxy for its degree of involvement in the phenotype characterized by the set of seed genes. Hence the prioritization is defined by ranking the candidates according to their seed-similarity scores (Fig. 9.2f).

In the remainder of this section we provide a more detailed overview of each of the four components of Collage.

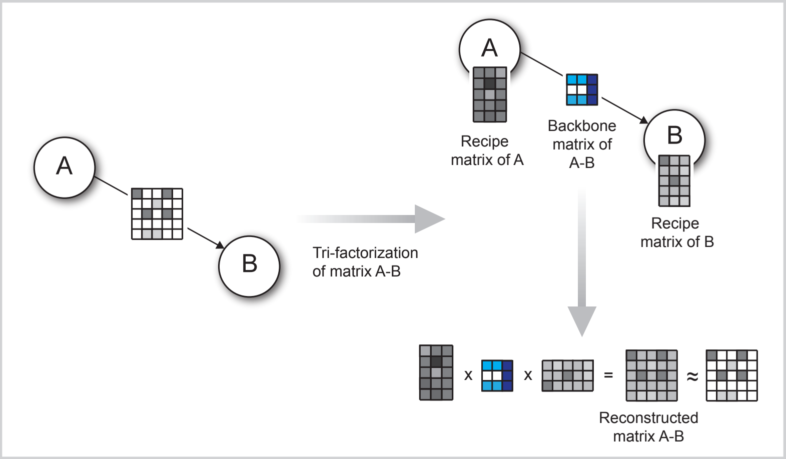


Figure 9.1

Each data matrix in Collage relates two object types. We graphically represent this relation such that nodes A and B represent object types and the directed edge A-B connects the two nodes with an associated data matrix. The matrix has objects of type A, e.g., genes, in the rows and objects of type B, e.g., experimental conditions, in the columns as indicated by the edge directionality. Grey cells in the matrix represent quantitative measurements, e.g., mRNA transcript abundance, or binary memberships that relate objects in rows to objects in columns. Empty cells denote missing values. See Sec. 9.2 for further explanation.

9.2.1 Step I: Compressive data fusion

Collage starts with a collection of data sets and can consider any kind of information (data tables, ontologies, associations, networks) that can be encoded in a matrix (Fig. 9.3). Each data set is viewed as a relation between two object types. For example, gene expression data relate gene names (columns) to experimental conditions (rows), where the entries represent transcript abundance. Literature annotation data relate research papers and their contents to annotation terms, where the entries are Boolean. Such data sets are abundant in the field of molecular biology and they report on dyadic relations that can be encoded in matrices. Matrix data representation is suitable for a wide range of data types, including tables, associations, ontologies and networks. Whenever data sets share object types, we can connect them in a data fusion graph with object types as nodes and data matrices as edges. In the simplest data fusion graph shown in Fig. 9.1, node A may represent known genes in a certain genome and node B may denote various experimental conditions. A gene from A could be related to an experimental condition in B through a level of its mRNA abundance. Relationships between all genes and experimental conditions are represented in a data matrix that is placed on the edge A-B.

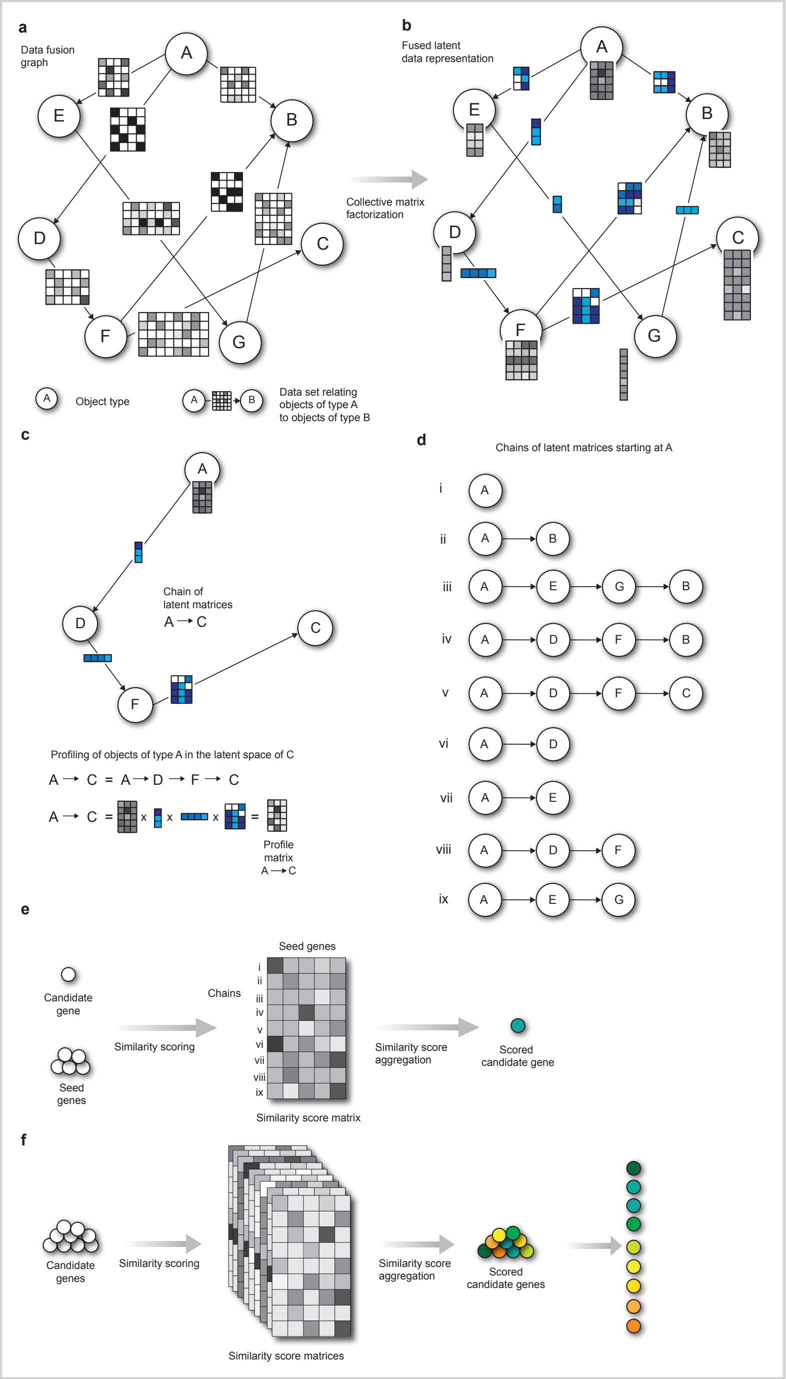


Figure 9.2

Overview of the Collage prioritization algorithm. Collage algorithm comprises four main steps: (a, b) compressive data fusion, (c, d) object profiling by chaining of latent matrices, (e) similarity estimation, and (f) gene ranking. See Sec. 9.2 for further explanation.

We model the system of data sets (Fig. 9.2a) through data fusion by collective matrix factorization (Žitnik and Zupan, 2015a). Matrix factorization compresses the data matrices to a latent space and infers recipes to convert the latent representation back to the original data domain. Each data matrix is decomposed into a product of three low-dimensional latent matrices: a *backbone matrix* encodes the relations between the latent components, and two *recipe matrices* transform the backbone matrix to the original space of the object types (Fig. 9.1). Data sets that are directly related and share a node in the fusion graph report on a common object type and hence use a common recipe matrix in their decomposition. Importantly, decomposition of any data set in the system depends on all other data sets according to a design of the fusion graph (Fig. 9.2b). Sharing of recipe matrices ensures data fusion and allows Collage to incorporate knowledge about the relations between data sets.

9.2.2 Step II: Chaining of latent matrices

Collage profiles objects in the latent space of any other object type based on the connectivity in the data fusion graph. In the simplest scenario, where object types are adjacent, such as A and D in Fig. 9.2b, Collage profiles objects of type A in the latent space of D by multiplying the recipe matrix of A by the backbone matrix A-D. The resulting profile matrix has objects of type A in rows and the latent components of type D in columns. The advantage of Collage over other gene prioritization tools is its ability to profile objects whose types are not direct neighbors in the fusion graph, such as A and C in Fig. 9.2b. To profile objects of A in the latent space of C Collage starts with the recipe matrix of A and multiplies it by backbone matrices A-D, D-F and F-C on the path from A to C (Fig. 9.2c). If A represented genes, D literature, F literature annotations and C chemical compounds, this procedure would yield profiles of genes in the latent space of chemical compounds. We refer to this technique as latent matrix chaining. It constructs dense profiles that include the most informative features obtained by collectively compressing data via matrix factorization. Intuitively, chaining is able to establish links between genes and chemical compounds even though relationships between these object types are not available in input data in Fig. 9.2a.

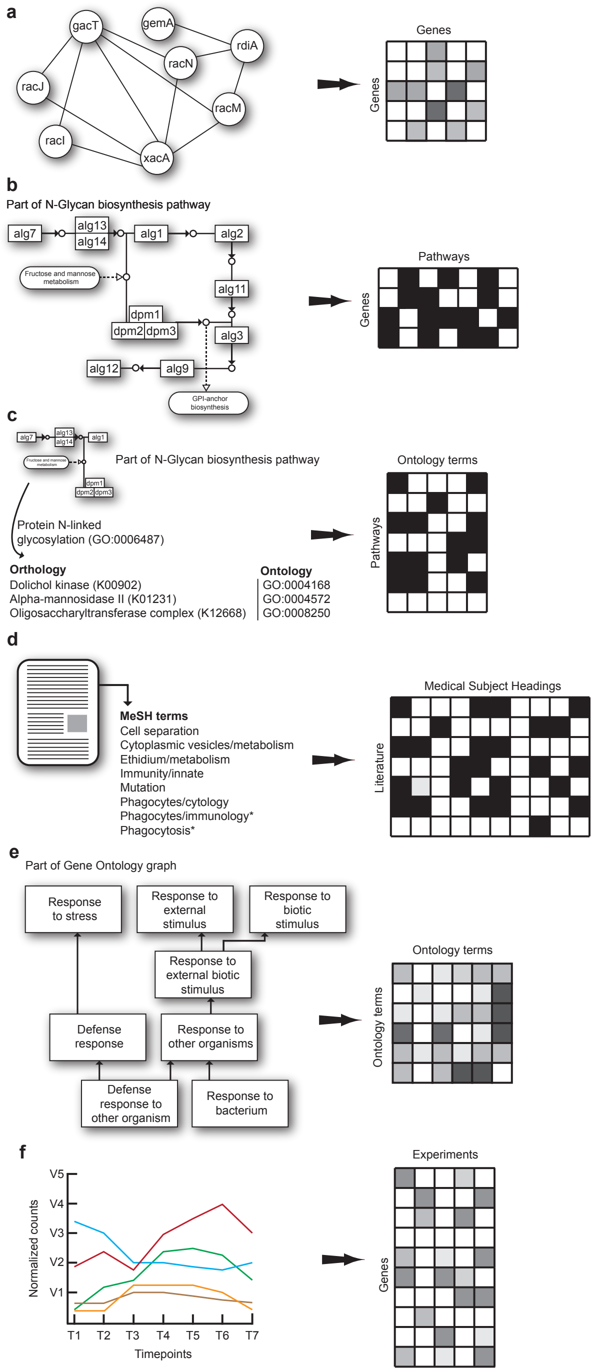


Figure 9.3

Representation of information sources with data matrices. Matrices in Collage describe relationships between objects of two types. Matrix rows correspond to objects of one type, columns correspond to objects of the other type and matrix elements express the degree of a relationship between the corresponding objects. The figure illustrates matrix representation of six distinct data sets. (a) Degrees of protein-protein interactions from the STRING database are represented in a gene-to-gene matrix. (b) Membership of genes in pathways are represented in binary matrices, one column for each pathway. Binary matrices are also used to associate (c) pathways with gene ontology terms and (d) research articles with Medical Subject Headings. (e) The structure of Gene Ontology can be represented with a real-valued matrix, whose elements report on distance or semantic similarity between the corresponding ontological terms. (f) Levels of gene expression, an experimental data set, are represented by a matrix of stacked gene expression profile vectors.

9.2.3 Steps III and IV: Gene prioritization

Collage prioritizes objects of the target object type, e.g., genes, node A in Fig. 9.2, based on a small set of seed objects (previously characterized genes). For each target object, it constructs a set of profile matrices by considering all possible chains of latent matrices that start in the target node and end in any node that is reachable in data fusion graph (Fig. 9.2d). A profile matrix corresponds to a particular latent matrix chain and encodes the latent space of the chain's last node. Each profile matrix is used to estimate the similarity between any two targets (genes) by comparing their respective profiles. Collages estimates the overall similarity between a candidate gene and the seed genes by aggregating similarity scores of the candidate gene across all profile matrices (Fig. 9.2e). As a final step, Collage ranks all the genes based on their overall similarity with the seed genes (Fig. 9.2f).

9.3 *A case study: bacterial response gene prioritization in* Dictyostelium

Collage is agnostic to data types it can consider and can be applied to any collection of data sets and any phenotype of interest. We used Collage to find genes that affect *D. discoideum* growth on the Gram-negative bacteria *Klebsiella pneumoniae*. We started with four seed genes that have been previously identified in a genetic screen for *D. discoideum* mutants that fail to grow on Gram-negative bacteria (Table 9.1). We fused 14 publicly available data sets that were considered relevant to the problem. Collectively, these data sets describe relations between 10 object types (see data fusion graph in Fig. 9.4). Our prioritization task was particularly challenging since there is not a lot of information about *Dictyostelium* in the literature and in public databases and only one of the data sets (Fig. 9.4, Bacterial RNA-seq, node 9) was directly related to the task. Collage ranked ∼12,000 genes from the *Dictyostelium* genome. The prioritized gene list was then filtered by the reported availability of *D. discoideum* gene knockout strains in the Dicty Stock Center (http://dictybase.org/StockCenter/StockCenter.html). We selected eight genes listed in Table 9.4 from the 30 top-ranked candidates in Table 9.3 (left column) for direct testing.

Table 9.1

Seed *D. discoideum* genes used for Gram-negative bacterial response gene prioritization. Seed genes used for prioritization by Collage were selected based on the experiments published in (Nasser et al., 2013).

Gene	DictyBase ID	Description
nip7	DDBG0295477	Ortholog of the conserved NIP7 nucleolar protein that is required for 60S ribosome subunit biogenesis; contains a PUA domain.
clkB	DDBG0278487	Similar to the cell division cycle 2-related protein kinase 7 (CRK7) and other cell division cycle 2-like protein kinases; belongs to the CMGC group of protein kinases.
spc3	DDBG0290851	Ortholog of the conserved microsomal signal peptidase 23 kDa subunit; the signal peptidase complex is a membrane-bound endoproteinase that removes signal peptides from nascent proteins as they are translocated into the lumen of the endoplasmic reticulum; contains a putative signal peptide.
alyL	DDBG0286229	Amoeba lysozyme family protein (aly), but divergent compared to *alyA-D*.

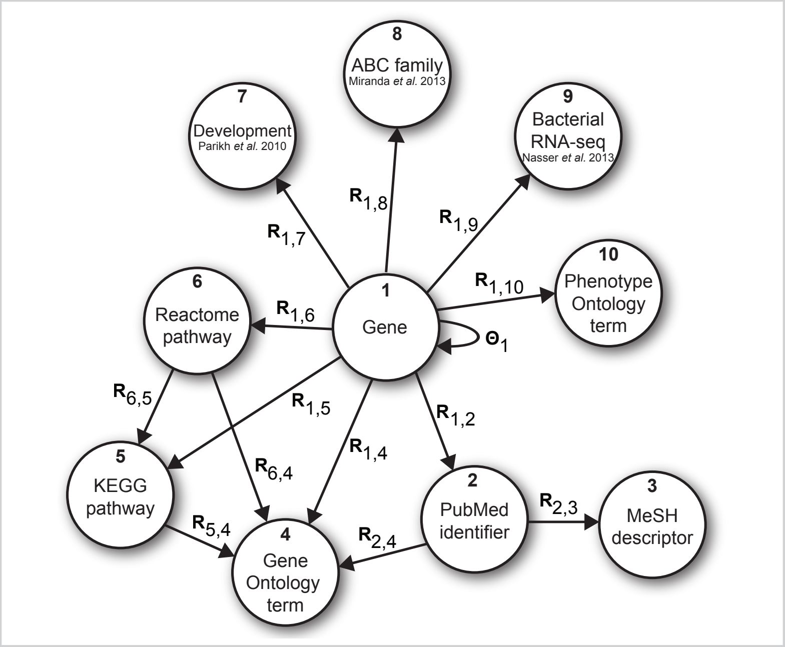


Figure 9.4

A data fusion graph for bacterial response gene prioritization in *Dictyostelium*. Collage considered 14 data sets (edges, represented by arrows) in this study describing the relations between 10 object types (nodes, represented by circles). The data sets included three whole-genome *D. discoideum* RNA-seq experiments ($R_{1,7}$, $R_{1,8}$, $R_{1,9}$), protein-protein interactions from the STRING database (Θ_1), gene mentions in research articles ($R_{1,2}$) and their Medical Subject Headings (MeSH) annotations ($R_{2,3}$), pathway memberships from the Kyoto Encyclopedia of Genes and Genomes (KEGG) and Reactome databases ($R_{1,6}$, $R_{1,5}$, $R_{6,5}$), associations of genes to phenotypes from Phenotype Ontology ($R_{1,10}$), gene functions in Gene Ontology ($R_{1,4}$) and interrelatedness of Reactome and KEGG pathways and research literature with Gene Ontology terms ($R_{6,4}$, $R_{5,4}$, $R_{2,4}$).

9.3.1 Considered data sets

A total of 14 data sets and 10 object types were considered for Gram-negative bacterial response gene prioritization. Data sets were organized in a data fusion graph (Fig. 9.4).

We used RPKM-normalized RNA-seq transcriptional profiles of 35 abc-transporter mutant strains and wild-type AX4 strain in two biological replicates and at four different time points during development (Miranda et al., 2013) ($\boldsymbol{R}_{1,8}$), normalized gene expression profiles analyzed by RNA-seq and measured at 4-hour intervals during the 24-hour development of *D. discoideum* in two biological replicates (Parikh et al., 2010) ($\boldsymbol{R}_{1,7}$), and normalized abundances of gene transcripts in two replicates and four different bacterial growth conditions analyzed with RNA-seq (Nasser et al., 2013) ($\boldsymbol{R}_{1,9}$).

We also included the following publicly available data sets: Phenotype Ontology (Fey et al., 2009) annotations ($\boldsymbol{R}_{1,10}$) downloaded from the DictyBase data portal in March 2014, protein-protein interactions from the STRING v.9 database (Franceschini et al., 2013) ($\boldsymbol{\Theta}_1$), membership of *D. discoideum* genes in pathways from the Reactome database (Croft et al., 2014) ($\boldsymbol{R}_{1,6}$) downloaded in August 2013, Kyoto Encyclopedia of genes and genomes (KEGG) pathway memberships (Kanehisa et al., 2014) ($\boldsymbol{R}_{1,5}$), and annotations of genes in Gene Ontology (Ashburner et al., 2000) ($\boldsymbol{R}_{1,4}$). Additionally, we cross-referenced Reactome and KEGG pathways ($\boldsymbol{R}_{6,5}$), Gene Ontology terms and Reactome pathways ($\boldsymbol{R}_{6,4}$), and KEGG orthology groups and Gene Ontology terms ($\boldsymbol{R}_{5,4}$). Literature data included associations of genes to research articles from PubMed ($\boldsymbol{R}_{1,2}$) accessed in August 2013 through DictyBase, mapping of research articles to Gene Ontology terms ($\boldsymbol{R}_{2,4}$) and their Medical Subject Headings (MeSH) ($\boldsymbol{R}_{2,3}$). Table 9.2 summarizes the number of objects of each type and the data sets considered in our analysis.

9.3.2 Inference of a joint latent factor model

A total of 14 data sets and 10 object types were considered for Gram-negative bacterial response gene prioritization (Fig. 9.4). Data sets are viewed as dyadic relations and are encoded in relation and constraint matrices. Given a collection of relation matrices $\mathscr{R}$ ($\boldsymbol{R}_{ij}$ for different choices of i and j) and a collection of constraint matrices $\mathscr{C}$ ($\boldsymbol{\Theta}_i^{(l)}$ for different choices of i, where l enumerates constraint matrices available for object type i), collective matrix factorization simultaneously decomposes all the relation matrices in $\mathscr{R}$ while regularizing the inferred latent model with the constraints in $\mathscr{C}$ using the algorithm described in Chapter 6.

The inferred low-dimensional matrix factors $\boldsymbol{G}_i$, $\boldsymbol{G}_j$ and $\boldsymbol{S}_{ij}$ decompose the associated relation matrix such that $\boldsymbol{R}_{ij} \approx \boldsymbol{G}_i \boldsymbol{S}_{ij} \boldsymbol{G}_j^T$. We call a $n_i \times c_i$ nonnegative latent matrix $\boldsymbol{G}_i$ a *recipe matrix*. It contains the latent profiles of objects of type i in the rows. Another recipe matrix is a $n_j \times c_j$ nonnegative latent matrix $\boldsymbol{G}_j$ with profiles of objects of type j in the rows. We refer to a $c_i \times c_j$ latent matrix $\boldsymbol{S}_{ij}$ as a *backbone matrix*. The backbone matrix $\boldsymbol{S}_{ij}$ models interactions between latent components in the (i, j)-th data set. Latent profile of an object of type i is given by its corresponding row vector in $\boldsymbol{G}_i$ and encodes membership of the object to c_i latent components.

Table 9.2

Summary of data sets considered for bacterial response gene prioritization in *D. discoideum*. The notation of the data sets ("Data matrix" column) is the same as in the data fusion graph (Fig. 9.4). All relation data matrices were normalized before data analysis such that the Frobenius norm of every row profile was equal to 1.

Data matrix	Matrix size	Description
$\boldsymbol{R}_{1,4}$	12,873 × 3,083	Gene annotations from the Gene Ontology (Ashburner et al., 2000).
$\boldsymbol{R}_{1,8}$	12,873 × 282	RPKM-normalized RNA-seq transcriptional profiles of 35 abc-transporter mutant strains and wild-type AX4 strain in two replicates and at four different time points during development (Miranda et al., 2013).
$\boldsymbol{R}_{1,2}$	12,873 × 3,424	Associations of *D. discoideum* genes to research articles from PubMed accessed in August 2013.
$\boldsymbol{R}_{1,5}$	12,873 × 99	Memberships of *D. discoideum* genes in the KEGG pathways (Kanehisa et al., 2014).
$\boldsymbol{R}_{1,6}$	12,873 × 92	*D. discoideum* pathways from the Reactome database (Croft et al., 2014) in August 2013.
$\boldsymbol{R}_{1,7}$	12,873 × 14	Normalized gene expression profiles analyzed with RNA-seq and measured at 4-hour intervals during 24-hour *D. discoideum* development in two replicates (Parikh et al., 2010).
$\boldsymbol{R}_{1,9}$	12,873 × 8	Normalized abundances of gene transcripts in two replicates and four different bacterial growth conditions analyzed with RNA-seq (Nasser et al., 2013).
$\boldsymbol{R}_{1,10}$	12,873 × 503	Gene annotations from the DictyBase Phenotype Ontology in March 2014.
$\boldsymbol{R}_{2,4}$	3,424 × 3,083	Cross-references of research articles from the PubMed and Gene Ontology terms. We counted the words from the Gene Ontology term names that occurred in the abstracts of articles from the PubMed database.
$\boldsymbol{R}_{2,3}$	3,424 × 2,804	Assignments of Medical Subject Headings (MeSH) to research articles from the PubMed.
$\boldsymbol{R}_{5,4}$	99 × 3,083	Cross-references of the KEGG orthology groups and Gene Ontology terms. We mapped KEGG pathways to KEGG orthology groups and used the mapping between ortholog groups and Gene Ontology terms as specified by the KEGG pathway browser.
$\boldsymbol{R}_{6,4}$	92 × 3,083	Cross-references of the Reactome pathways and Gene Ontology terms available in the generic Gene Ontology Slim subset.
$\boldsymbol{R}_{6,5}$	92 × 99	Cross-references of the Reactome and KEGG pathways by semantic similarity of KEGG pathway names and Reactome pathways display names.
Θ_1	12,873 × 12,873	Protein-protein interaction data from the STRING v.9 database (Franceschini et al., 2013) in April 2014. Ortholog mapping of *Dictyostelium* genes onto interactions from other organisms is performed with the Clusters of Orthologous Group (COGs).

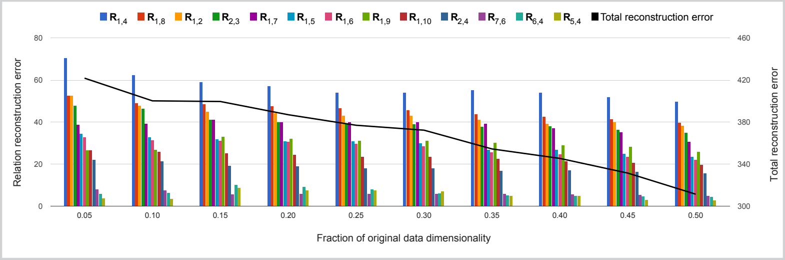


Figure 9.5

Reconstruction error as a function of factorization rank. Ranks were set using a parameter k as kn_i for object type i with n_i objects. The value of k was selected by observing the change of the "total reconstruction error" (black line), when varying k between 0.05 and 0.5 (x-axis, "fraction of original data dimensionality"). The bars show reconstruction errors of individual data matrices ("relation reconstruction error").

The algorithm for inference of the fused latent model given in Chapter 6 is an iterative algorithm that starts by randomly initializing latent matrices $\boldsymbol{G}_i$ and then alternates between updating matrices $\boldsymbol{G}_i$ and $\boldsymbol{S}_{ij}$ until convergence. To ensure robust prioritization, the algorithm was run 20 times with different initializations of latent matrices. The algorithm was run for a maximum of 200 iterations or was terminated early if the total reconstruction error between consecutive iterations changed by less than 0.01.

Parameters of the algorithm are factorization ranks, c_i, for every object type i in the data fusion system. Our prioritization of *D. discoideum* genes included 10 types of objects; we have selected latent dimensionality of object types through a single parameter representing the fraction of the original data dimensionality such that $(c_1, c_2, \ldots, c_{10}) = (kn_1, kn_2, \ldots, kn_{10})$. The value of k was obtained by observing kinks in a diagram of total reconstruction error, $\sum_{R_{ij}\in\mathcal{R}} \|\boldsymbol{R}_{ij} - \widehat{\boldsymbol{R}}_{ij}\|_{\text{Fro}}$, when varying k from 0.05 to 0.5 (Fig. 9.5). The reconstruction error was estimated by 50 repetitions of collective matrix factorization, where each repetition was run with a different random initialization of latent matrices. We selected $k = 0.1$ where a maximum kink was attained. This choice resulted in latent data dimensionality $(c_1, c_2, \ldots, c_{10}) = (1287, 342, 280, 308, 9, 9, 5, 28, 5, 50)$ with a limitation on minimum factorization rank set to 5.

9.3.3 Gene profiling

We profiled genes by considering latent data representation inferred by data fusion. Each gene was characterized through a collection of profiles determined by the topology of the data fusion graph. We obtained gene profiles by starting at a gene node and its corresponding recipe matrix ($\boldsymbol{G}_1$), and traversing along the edges of the data fusion graph, multiplying the edge-associated backbone latent matrices.

In the bacterial response gene prioritization study there were 15 such paths of latent matrices (Fig. 9.4), and correspondingly 15 different profile matrices with gene profiles for every candidate gene: $\boldsymbol{G}_1$, $\boldsymbol{G}_1\boldsymbol{S}_{1,7}$, $\boldsymbol{G}_1\boldsymbol{S}_{1,8}$, $\boldsymbol{G}_1\boldsymbol{S}_{1,9}$, $\boldsymbol{G}_1\boldsymbol{S}_{1,10}$, $\boldsymbol{G}_1\boldsymbol{S}_{1,2}$, $\boldsymbol{G}_1\boldsymbol{S}_{1,6}$, $\boldsymbol{G}_1\boldsymbol{S}_{1,5}$, $\boldsymbol{G}_1\boldsymbol{S}_{1,4}$, $\boldsymbol{G}_1\boldsymbol{S}_{1,2}\boldsymbol{S}_{2,3}$, $\boldsymbol{G}_1\boldsymbol{S}_{1,6}\boldsymbol{S}_{6,5}$, $\boldsymbol{G}_1\boldsymbol{S}_{1,6}\boldsymbol{S}_{6,4}$, $\boldsymbol{G}_1\boldsymbol{S}_{1,2}\boldsymbol{S}_{2,4}$, $\boldsymbol{G}_1\boldsymbol{S}_{1,5}\boldsymbol{S}_{5,4}$ and $\boldsymbol{G}_1\boldsymbol{S}_{1,6}\boldsymbol{S}_{6,5}\boldsymbol{S}_{5,4}$.

9.3.4 Gene prioritization

The inputs to gene prioritization were candidate genes, seed genes and the set of profile matrices. We aimed to find genes whose profiles are similar to the profiles of seed genes. We estimated the similarities independently for each profile matrix, and then aggregated the resulting scores to obtain the final prioritization. Each row in a profile matrix corresponds to a profile of a gene. We assessed similarity between a candidate gene and a seed gene by computing Spearman rank correlation of two respective row vectors. This procedure yielded a $15 \times |\text{seed genes}|$ similarity score matrix of rank correlations for each candidate gene. Similarity score matrices were aggregated in a two-step median value computation along score matrix dimensions to produce a single rank value per gene. We obtained empirical P-values by randomizing seed set of genes. Randomization of seed genes was repeated 50 times. Empirical P-value of a candidate rank was estimated as the fraction of randomizations with higher aggregated score than the score obtained from the original seed set.

As a gene profile similarity measure, Spearman rank correlation was chosen for its correspondence to similarity of gene assignments to latent components. A promising candidate gene should have a latent profile similar to the profile of a seed gene. Given a profile matrix $\boldsymbol{X}$, candidate gene g and seed gene s, gene g is considered promising if its latent component with the largest membership is the same as that of seed gene s. We formalize this intuition by measuring whether $\arg\max_j \boldsymbol{X}(g, j) = \arg\max_j \boldsymbol{X}(s, j)$. The same should hold for the latent component of the second largest, third largest, and all remaining value-ordered gene memberships. Quantitatively, the described procedure corresponds to rank correlations between candidate and seed genes.

9.3.5 Sensitivity of gene prioritization to the inclusion of data sets

To study the sensitivity of gene prioritization to the number of data sets in the data fusion graph, we observed how the rankings of the validated candidate genes changed when the overall prioritization was obtained by fusing different subsets of data sets from our initial collection. Four independent gene prioritization predictive models were inferred in addition to the original model that contained 14 data sets (Table 9.3). The scenarios considered seven, four, three and two data sets, where each scenario considered a different subset of the data sets (Fig. 9.6). The selection of data sets was in part determined by the data fusion graph. In particular, for data fusion to take place, the associated graph has to be connected such that information can be shared between data matrices.

9.3.6 Validation of top ranked candidate genes in the wet laboratory

To validate the selected candidate genes, we assessed growth of the *D. discoideum* knockout strains by making serial dilutions of the amoebae and co-culturing the cells with *K. pneumoniae* bacteria on nutrient agar. We observed a significant difference in the growth of all the mutants compared to the wild type AX4 (Fig. 9.7). In this system, the bacteria grow faster than the amoebae so the first observation is the appearance of a thick opaque lawn of bacteria on the surface of the agar plate within 24 hours (not shown). Later on, as the amoebae eat the bacteria, they clear parts or all of the bacterial lawn, depending on their density and growth rate. When there are numerous, fast growing amoebae, we observe a cleared lawn, e.g., Fig. 9.7, AX4, 10^4 cells, Day 2. When there are very few amoebae, we observe distinct plaques that appear as darker spots in the bacterial lawn, e.g., Fig. 9.7, AX4 Day 3, 10^2 cells. When the bacteria are consumed, the amoebae starve, aggregate, and form developmental structures (Fig. 9.7, AX4 Day 3, 10^4 cells). Growth of the Collage-predicted knockout strains was compared to the wild type (AX4, top row in Fig. 9.7) and to the most severe mu-

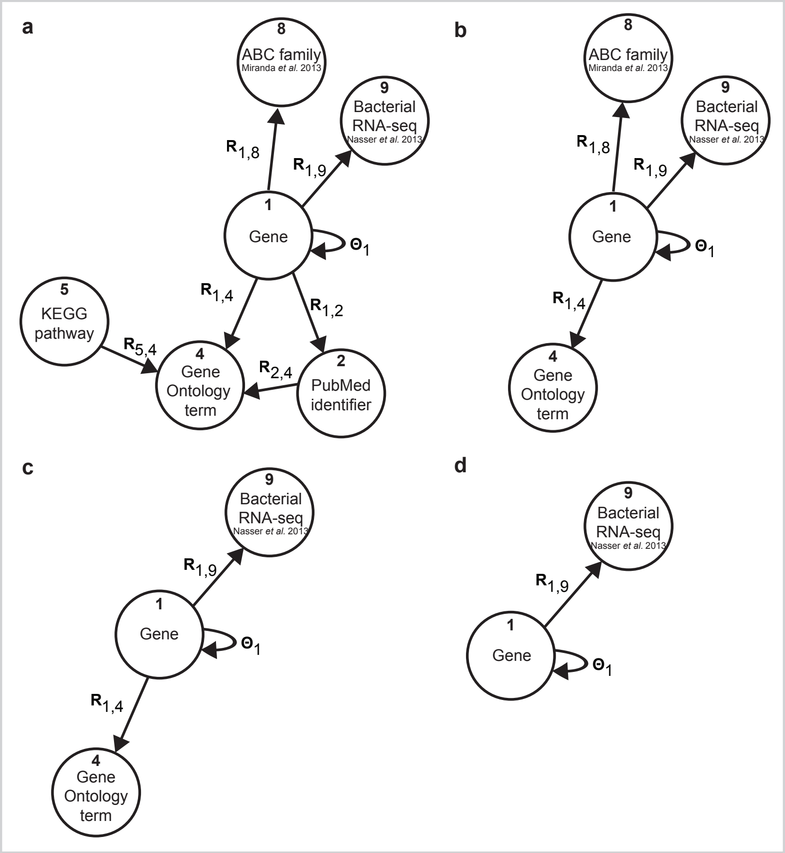


Figure 9.6

Data fusion graphs for the study of sensitivity to data set selection. Besides the full collection of data sets (data fusion graph in Figure 2), we have considered data collections with a smaller number of data matrices and studied the impact of this reduction on gene prioritization (Table 9.3). We ran gene prioritization analyses by considering subsets of (a) seven, (b) four, (c) three and (d) two data sets that were included in our original study.

Table 9.3

The impact of modeling circumstantial data on the overall *D. discoideum* bacterial response gene prioritization. The table lists the top-30 candidate genes obtained by prioritization by data fusion of 14, 7, 4, 3 and 2 data sets from the data fusion graphs in Fig. 9.6. Genes in red are the ones selected for the experimental study.

14 data sets	7 data sets	4 data sets	3 data sets	2 data sets
cf50-1	shkA	rbsk	DDB_G0271348	arpE
smlA	DDB_G0288519	DDB_G0272614	DDB_G0268872	DDB_G0278663
acbA	pten	DDB_G0278163	DDB_G0287153	DDB_G0281091
pirA	cf50-1	qtrt1	yelA	DDB_G0267742
rps10	acbA	DDB_G0279263	sibD	pten
abpC	smlA	DDB_G0286079	DDB_G0272380	DDB_G0277937
tirA	DDB_G0288947	adprh	DDB_G0288519	DDB_G0271120
DDB_G0272184	DDB_G0275057	DDB_G0279939	dnaja1	yipf1
pikB	tra2	DDB_G0272382	rabT2	DDB_G0267494
vps46	sibC	gdt6	DDB_G0292920	DDB_G0272016
pikA	rbsk	ku80	sibB	eif2b1
swp1	DDB_G0281967	arpF	DDB_G0278163	empB
ggtA	pikA	cofD-1	adprh	DDB_G0291926
DDB_G0288519	DDB_G0272614	DDB_G0288551	lvsG	vps13l
pten	DG1112	empB	DDB_G0285403	cenB
DDB_G0288551	adprh	gacV	tpsB	ku80
tra2	DDB_G0288551	DDB_G0294629	ndm	DDB_G0288161
DDB_G0286429	DD_G0283989	swp1	DDB_G0281559	DDB_G0268232
dscA-1	dscA-1	gbqA	DDB_G0275671	rbsk
cinC	gdt6	DDB_G0291926	DDB_G0288963	atg12
udpB	piaA	DDB_G0273031	gbqA	vps46
sfbA	DDB_G0279145	DDB_G0287643	uduA1	DDB_G0290575
modA	DDB_G0290575	DDB_G0268876	acrA	DDB_G0267958
DDB_G0287399	abcA1	abkD	arpE	DDB_G0287153
prmt5	DDB_G0272380	DDB_G0268206	uduC	gacV
dpoA	DDB_G0272801	DDB_G0279145	DG1098	DDB_G0276509
DDB_G0278663	lipA	DDB_G0272380	DDB_G0273451	DDB_G0279971
psiP	cepG	plbG	adprt3	usp39
sibC	lvsG	cct3	DDB_G0288031	DDB_G0280477
DDB_G0291926	uduA1	cct3	yipf1	DDB_G0292098

tant available (*tirA*–, bottom row in Fig. 9.7). Cells that carry an inactivating mutation in the *tirA* gene (*tirA*$^-$ cells) exhibit impaired growth on *K. pneumoneae* (Chen et al., 2007). We used these cells as a control in our assay and indeed they exhibited no clearing of the bacterial lawn when plated at the same initial density as the wild type cells (Fig. 9.7, AX4 vs. *tirA*$^-$, Day 2, 10^4 cells). We note that *tirA*$^-$ cells can grow to some extent on *K. pneumoniae* bacteria under certain conditions, indicating that the growth phenotype is continuous even though many researchers tend to describe it as Boolean.

We tested the predictions made by Collage on eight genes—*acbA*, *smlA*, *pikA*, *pikB*, *pten*, *abpC*, *modA* and *cf50-1* (Table 9.4). In the case of *pikA* and *pikB* we used a double knockout strain because of previously reported overlap in the functions of these two genes (Zhou et al., 1995). Strikingly, when we assessed the ability of the mutant cells to grow on bacteria, they all exhibited varying degrees of growth defects compared to the equivalent wild type (AX4) control (Fig. 9.7). Comparing only one condition, disruption of *acbA*, *abpC* and *modA* resulted in small individual plaques in the bacterial lawn but not complete clearing as observed in AX4 (Fig. 9.7, black box, Day 2, 10^4 cells). In contrast, mutations in *smlA*, *pikA/pikB*, *pten*, and *cf50-1* caused phenotypes as severe as the loss of *tirA* with no clearing on Day 2 (Fig. 9.7, black box, Day 2, 10^4 cells). Further distinction in the ability to grow on bacteria was revealed when the mutant cells were observed for an additional day. For example on Day 2, *pikA*$^-$/*pikB*$^-$ and *pten*$^-$ cells exhibited similar growth defects, but by Day 3, the loss of *pten* did not hinder growth on bacteria as much as the loss of *pikA* and *pikB* (Fig. 9.7).

Details on the experimental analysis of Dictyostelium *mutants*

D. discoideum strains were obtained from the Dicty Stock Center and grown axenically in HL-5 at 22^{o}C (Nasser et al., 2013). *K. pneumoniae* was maintained in SM broth at 22^{o}C. To assess the ability of *D. discoideum* to grow on bacteria, *D. discoideum* cells were collected from axenic cultures during logarithmic growth and washed once with Sorensen's buffer (Nasser et al., 2013). *D. discoideum* cells were serially diluted with bacteria (OD_{600} = 1.0) and spotted onto SM agar plates. The plates were incubated in a humid chamber at 22°C, and images of plates were taken every 24 hours. Images were taken at 2 and 3 days after plating to show the progression of amoebae growth in time. Each experiment was performed in duplicate. Representative images of three

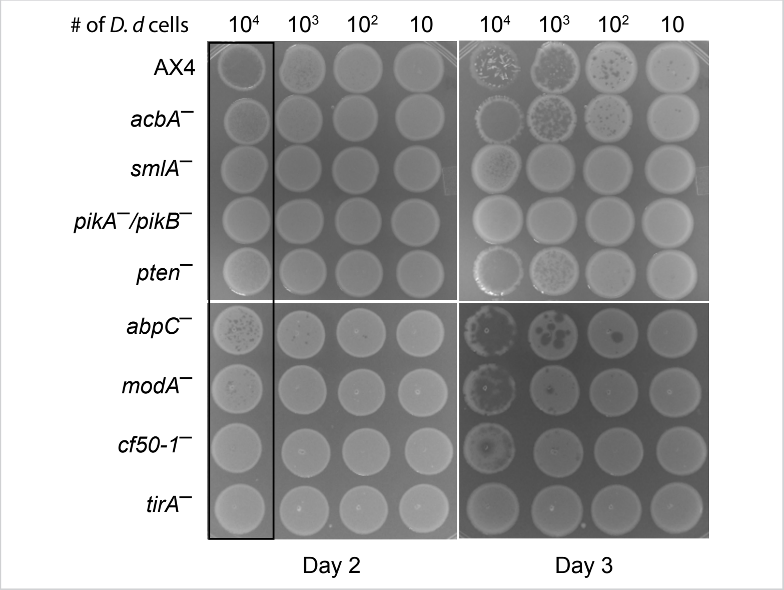


Figure 9.7

Experimental validation of top ranked candidate genes. Co-cultures of *D. discoideum* (*D.d.*) and bacteria were generated by serial dilutions of axenically grown *D. discoideum* amoebae with a large excess of *K. pneumoniae* bacteria such that the number of amoebae plated in each spot was between 10 and 10^4 as indicated above each column. The relevant genotypes of the amoebae strains are indicated on the left of each row. The larger white opaque spots are lawns of the *K. pneumoniae* bacteria. Growth of the *D. discoideum* amoebae results in the formation of plaques within the opaque spots in cases of low amoebae cell density or clearing of much or all of the opaque spots in cases of high amoebae cell density.

independent experiments are shown in Fig. 9.7.

9.4 *Discussion and conclusion*

The results indicate that Collage is capable of prioritizing genes in a reliable manner and identifying genes with various effects on the tested phenotype. This allows the analysis of a broad spectrum of genes in a given biological pathway. Application of the method to this specific question required only a few days of computational work and the validation step required a few more days of work. Considering the low yield of standard genetic screens, it would have taken about a year to identify 8 new genes in the bacterial response pathway.

Six of the eight validated bacterial growth genes—*cf50-1*, *abpC*, *smlA*, *pten*, *pikA* and *pikB*, are involved in actin polymerization and cell motility (Gao et al., 2007; Dormann et al., 2004; Cox et al., 1996; Brock et al., 2002). One explanation for the enrichment

Table 9.4

Top-ranked candidate *D. discoideum* genes tested for Gram-negative bacterial response. The name of the candidate gene, DictyBase ID and description from DictyBase are shown, together with the rank (out of all *D. discoideum* gene knockout strains available in the Dicty Stock Center) at which the candidate was prioritized by Collage using the data sets from the fusion graph in Fig. 9.4.

Gene	DictyBase ID	Description	Rank position
cf50-1	DDBG0273175	Component of the counting factor complex, which includes CF60, CF50, CF45-1, and CtnA (countin).	1
smlA	DDBG0287587	Cytosolic protein present in vegetative and developing cells.	2
acbA	DDBG0270658	Precursor of SDF-2; similar to diazepam binding inhibitor; enriched in prespore cells.	3
abpC	DDBG0269100	120 kDa F-actin binding protein also often called filamin; involved in actin cytoskeleton organization, motility, sand development; enriched in prestalk cells.	6
pikB	DDBG0283081	Phosphatidylinositol kinase.	9
pikA	DDBG0278727	Phosphatidylinositol kinase.	11
pten	DDBG0286557	Phosphatase and tensin homolog.	15
modA	DDBG0269154	Protein post-translational modification mutant.	23

of these genes is that the availability of preexisting knockout strains may be enriched with cell motility genes. This is because *D. discoideum* has been used extensively as a model system for chemotaxis, and many genes involved in cell motility have been disrupted and made available to the community. Nonetheless, the importance of actin in the consumption of bacteria may have been previously oversimplified, and the enrichment of these genes could be due to an essential role for actin in bacterial consumption. Proper regulation of actin is required for cell motility, phagocytosis and intracellular trafficking of phagosomes to lysosomes (Gao et al., 2007; Dormann et al., 2004; Cox et al., 1996; Brock et al., 2002). Each of these processes could be important in hunting, consuming and digesting bacteria.

We identified the sugar modifying alpha-glucosidase II enzyme, ModA (Ebert DL and JA, 1989). Complex sugar modifications are important for biogenesis and intracellular trafficking of proteins. Others have shown that disruption of *modA* results in a lack of anionic N-glycan, which is associated with lysosomal enzymes (Hykollari et al., 2014). While it may not be surprising to identify genes that regulate actin and lysosomes in a direct genetic screen, it is important to see that Collage did so too.

We also identified one gene, *acbA*, with a less salient relationship to bacterial consumption. Gene *acbA* encodes an Acyl-CoA Binding protein, which is similar to the mammalian diazepam binding inhibitor. Acyl-CoA Binding protein is secreted during *D. discoideum* development and cleaved to form the SDF-2 peptide (Spore Differentiation Factor-2) (Cabral et al., 2006, 2010). The role of Acyl-CoA Binding protein and SDF-2 in growth on bacteria is unclear. It is unlikely to be due to disruption of a general cellular growth pathway, since *acbA*$^{-}$ cells grow normally in axenic medium and it is unclear whether the SDF-2 peptide is secreted during growth because the system that produces it is developmentally-regulated. The identification of *acbA* suggests that novel gene functions can be discovered with our gene prioritization method.

The ranking of candidate genes depends on the particular collection of data sets we consider for gene prioritization. Removal of data sets from the data fusion graph (Fig. 9.6) changes the prioritization. When fewer data sets are considered, the validated genes from our study become ranked lower, below the top 30 (Table 9.3). This is an intuitive dependence – less information should result in reduced accuracy, and it is also validated by simulations. Our computational studies in data fusion with collective matrix factorization show that exclusion of data sets gradually reduces the quality of the predictions, e.g., see Chapters 7, 10 and 11. We can attribute our success in identification of genes that participate in Gram-negative response pathways to the proposed approach and the appropriate choice of 14 relevant data sets. In the absence of a much larger set of known genes for this pathway, we cannot claim that this particular selection of data sets is optimal.

Collage builds upon our data fusion method by collective matrix factorization introduced in Chapter 6, which can achieve high predictive accuracy and enables effortless integration of a range of very diverse data sets. Collective learning hence provides means for Collage to constitute a useful complement to large-scale ranking of genes in various organisms and to ranking of other objects contained in the fusion graph, such as drugs, diseases and pathways.

Disease-disease association prediction

10

The advent of genome-scale genetic and genomic studies allows new insight into disease classification. Recently, a shift was made from linking diseases simply based on their shared genes towards systems-level integration of molecular data. Here, we aim to find relationships between diseases based on evidence from fusing all available molecular interaction and ontology data.

We describe in this chapter a multi-level hierarchy of disease classes that significantly overlaps with existing disease classification. In it, we find 14 disease-disease associations currently not present in Disease Ontology and provide evidence for their relationships through comorbidity data and literature curation. Interestingly, even though the number of known human genetic interactions is currently very small, we find they are the most important predictor of a link between diseases. Finally, we show that omission of any one of the included data sources reduces prediction quality, further highlighting the importance in the paradigm shift towards systems-level data fusion.

10.1 Background

Disease Ontology (DO) (Schriml et al., 2012) is a well established classification and ontology of human diseases. It integrates disease nomenclature through inclusion and cross mapping of disease-specific terms and identifiers from Medical Subject Headings (MeSH) (Nelson et al., 2004), World Health Organization (WHO) International Classification of Diseases (ICD) (Aymé et al., 2010), Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) (Cornet and De Keizer, 2008), National Cancer Institute (NCI) thesaurus (Sioutos et al., 2007) and Online Mendelian Inheritance in Man (OMIM) (Amberger et al., 2011). It relates and classifies human diseases based on pathological analysis and clinical symptoms. However, the growing number of heterogeneous genomic, proteomic, transcriptomic and metabolic data currently does not contribute to this classification. Understanding of even the most straightforward monogenic classic Mendelian disorders is limited without considering interactions between mutations and biochemical and physiological characteristics. Hence, redefining human disease classification to include evidence from heterogeneous data is expected to improve prognosis and response to therapy (Loscalzo et al., 2007). In this chapter we examine whether inclusion of modern molecular level data can improve disease classification.

Several studies have reported on efforts and benefits of relating human diseases through their molecular causes. Loscalzo et al. (2007) cataloged diseases through a network-based analysis of associations among genes, proteins, metabolites, intermediate phenotype and environmental factors that influence pathophenotype. Gulbahce et al. (2012) constructed a "viral disease network" of disease associations to decipher the interplay between viruses and disease phenotypes. They uncover several diseases that have not previously been associated with infection by the corresponding viruses. A similar approach was used by Lee et al. (2008) to gain insights into disease relationships through a network derived from metabolic data instead of virological implications. They demonstrated that known metabolic coupling between enzyme-associated diseases reveal comorbidity patterns between diseases in patients. Goh et al. (2007) studied the position of disease genes within the human interactome in order to predict new cancer-related genes. Conversely, a gene-centric approach to disease association discovery was used by Linghu et al. (2009): they took 110 diseases for which a set of disease genes are known, and compared gene sets and their positions within the gene network to infer associations of related diseases. More details can be found in two recent surveys of current network analysis methods aimed at giving insights into human disease (Janjić and Pržulj, 2012; Emmert-Streib et al., 2013), as well as in a review of different data sources that can provide complementary disease-relevant information (Piro and Di Cunto, 2012).

A challenge in relating diseases and molecular data is in the multitude of information sources. Disease profiling may include data from genetics, genomics, transcriptomics, metabolomics or any other omics, all potentially related to susceptibility, progress and manifestation of disease. Such data may be related on their own: for example, information on transcription factor binding sites, gene and protein interactions, drug-target associations, various ontologies and other less-structured knowledge bases, such as literature repositories, are all inter-dependent and it is not trivial to integrate them in a way that will yield new information about diseases. This stresses the need for an integrated approach of current models to exploit all these heterogeneous data simultaneously when inferring new associations between diseases (Emmert-Streib et al., 2013).

Data from heterogeneous sources of information can be integrated by *data fusion* (Yu et al., 2010). Common fusion approaches follow early or late integration strategies, combining inputs (Mostafavi and Morris, 2012) or predictions (Pandey et al., 2010), respectively. Another and often preferred approach is an intermediate integration, which preserves the structure of the data while inferring a single model (Lanckriet et al., 2004b; Gevaert et al., 2006; van Vliet et al., 2012). An excellent example of intermediate integration is multiple kernel learning that convexly combines several kernel matrices constructed from available data sources (De Bie et al., 2007; Yu et al., 2010). Data fusion has been successfully applied for tasks such as gene prioritization (Aerts et al., 2006; De Bie et al., 2007; Yu et al., 2010), or gene network reconstruction and function prediction (Mostafavi and Morris, 2012; Chen and Zhang, 2013). To our knowledge, we present the first application of data fusion to disease association mining.

We choose the intermediate data fusion approach for its accuracy of inferring prediction models (i.e. how well a model can learn to predict disease-disease associations) and the ability to explicitly measure the contribution of each data set to the extracted knowledge (Lanckriet et al., 2004b; Gevaert et al., 2006). Kernel-based fusion can only use data sources expressed in the "disease space", i.e. all data sources have to be expressed as kernel matrices encoding relationships between diseases, which may incur loss of information when transforming circumstantial data sources into appropriate feature space. In our study, most of the data sources are only indirectly related to diseases, hence we employ an alternative and recently proposed intermediate data fusion algorithm by matrix factorization (Žitnik and Zupan, 2015a), which has an accuracy comparable to kernel-based fusion approaches, but can treat all data sources directly (i.e. no transformation of data into "disease space" is necessary). The key idea of our data fusion approach lies in sharing of low-rank matrix factors between data sources that describe biological data of the same type. For instance, genes are one data type which can be linked to other data types such as Gene Ontology (GO) terms or diseases through two distinct data sources, namely GO annotations and disease-gene mapping. The fused factorized system contains matrix factors that are specific to every molecular data type, as well as matrix factors that are specific to every data source. Thus, low-rank matrix factors can simultaneously capture both source- and object type-specific patterns.

We report on the ability of our recently developed data fusion approach to mine human disease-disease associations. Starting from Disease Ontology, we revise the links between diseases using related systems-level data, including protein-protein and genetic

Table 10.1

Data sets used for our disease association study. Relation matrices $\boldsymbol{R}_{ij}$ relate objects of two different types and their numbers are reported separately (delimited by a forward slash).

Matrix	Name	Nodes	Edges	Density	Availability
$\Theta_1^{(1)}$	Protein interactions	10,360	55,787	0.00104	BioGRID Rel 3.1.94 (Stark et al., 2011)
$\Theta_1^{(2)}$	Gene co-expression	539	869	0.00600	Prieto et al. (2008)
$\Theta_1^{(3)}$	Cell signaling data	1,217	7,517	0.01016	KEGG (Kanehisa et al., 2012)
$\Theta_1^{(4)}$	Genetic interactions	542	511	0.00349	BioGRID Rel 3.1.94 (Stark et al., 2011)
$\Theta_1^{(5)}$	Metabolic network	5,908	1,505,831	0.08630	KEGG (Kanehisa et al., 2012)
Θ_4	Drug interaction data	4,477	21,821	0.00218	DrugBank v3.0 (Knox et al., 2011)
Θ_3	GO semantic structure	11,853	43,924	0.00063	Gene Ontology (Ashburner et al., 2000)
Θ_2	DO semantic structure	1,536	1,098	0.00093	Disease Ontology (Schriml et al., 2012)
$\boldsymbol{R}_{13}$	Gene annotations	17,428/11,853	100,685	0.00049	Gene Ontology (Ashburner et al., 2000)
$\boldsymbol{R}_{14}$	Drug-target relations	1,978/4,477	7,977	0.00009	DrugBank v3.0 (Knox et al., 2011)
$\boldsymbol{R}_{12}$	Gene-disease relations	5,267/1,536	22,084	0.00273	Mapped GeneRIF (Osborne et al., 2009)

interactions, gene co-expressions, metabolic data, drug-target relations, and other (see Sec. 10.2). By fusing these data we identify several disease-disease associations that were not present in Disease Ontology and validate their existence by finding strong support in the literature and significant comorbidity effects in associated diseases. We also quantify the contribution of each molecular data source to the integrated disease-disease association model.

10.2 Data sets

In this study, we integrate biological data on objects of four different types (nodes in Fig. 10.1): genes, diseases (Disease Ontology terms), drugs, and Gene Ontology (GO) terms. We observe them through 11 sources of information (edges in Fig. 10.1). Every source of information is represented by a distinct data matrix that either relates objects of two different types (such as drugs and their associated target proteins) or objects of the same type (such as genetic interactions between genes): relations between objects of types i and j are represented by a *relation matrix*, $\boldsymbol{R}_{ij}$, and relations between objects of the same type i are represented by a *constraint matrix*, $\boldsymbol{\Theta}_i$. Table 10.1 summarizes all 11 data sets.

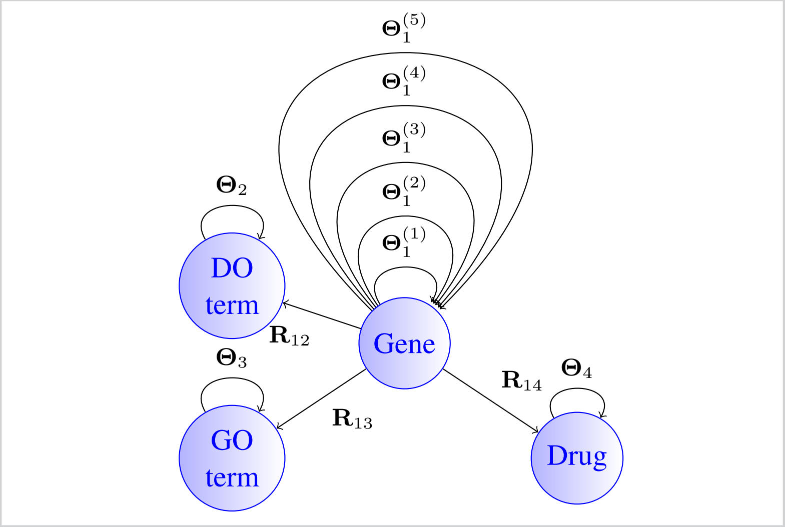


Figure 10.1

System-level data fusion approach to disease reclassification. The figure shows the relationships between data sources: nodes represent four types of objects, i.e. genes, GO terms, DO terms and drugs; arcs denote data sources that relate objects of different types (relation matrices, $\boldsymbol{R}_{ij}, i \neq j$), or objects of the same type (constraints, Θ_i).

10.2.1 Disease data

The principal source of information on human disease associations is Disease Ontology (DO) (Schriml et al., 2012). DO semantically combines medical and disease vocabularies and addresses the complexity of disease nomenclature through extensive cross-mapping of DO terms to standard clinical and medical terminologies of MeSH, ICD, NCI's thesaurus, SNOMED and OMIM. It is designed to reflect the current knowledge of human diseases and their associations with phenotype, environment and genetics. We extract $1,536$ DO terms from the latest version of the disease ontology hosted by the OBO Foundry (`http://www.obofoundry.org`) and construct a binary matrix $\boldsymbol{R}_{12}$ from $22,084$ associations between genes and diseases. DO leverages the semantic richness through linking terms by computable relationships in the hierarchy (e.g. mediastinum ganglioneuroblastoma is_a peripheral nervous system ganglioneuroblastoma, which is_a ganglioneuroblastoma and then in turn is_a neuroblastoma) first by etiology and then by the affected body system. We use the semantic structure of DO to reason over is_a relations. Since entries in the constraint matrices are positive for objects that are not similar and negative for objects that are similar, the constraint between two DO terms in $\mathbf{\Theta}_2$ is set to $-0.8^{\texttt{hops}}$, where `hops` is the length of the path between corresponding terms in DO graph. We empirically chose 0.8 from $[0, 1]$ range — 0 meaning that no two terms in the DO graph are related, and 1 meaning that two DO terms are always related (regardless of the path distance between them in the DO graph) — by performing standardized internal cross-validation using values between 0 and 1 with a 0.1 step (i.e. $0, 0.1, 0.2, \dots, 1$). Scores of multiple parentage (multiple is_a relationships) are summed to produce the final value of semantic association. Throughout the chapter, we use *disease* and *DO term* interchangeably, which both refer to a unique DO identifier (DOID).

10.2.2 Gene Ontology data

We use relations between 11,853 distinct genes and 100,685 gene annotations that are given by Gene Ontology (GO) (Ashburner et al., 2000) to construct a binary matrix of direct annotations $\boldsymbol{R}_{13}$. Topology of the GO graph is included by reasoning over is_a, $part\_of$ and *has_part* relations between GO terms to populate $\mathbf{\Theta}_3$ in the same way as $\mathbf{\Theta}_2$ with the constraint between two GO terms set to $-0.9^{\texttt{hops}}$.

10.2.3 Drug data

We obtain drug data from DrugCard entries in the DrugBank (http://www.drugbank.ca) database that contains chemical, pharmacological and pharmaceutical drug information with comprehensive drug target details. Our model contains 4,477 distinct drugs, each identified by a DrugBank accession number. Drugs are related to their target proteins in $\boldsymbol{R}_{14}$, which is populated by 7,977 binary drug-target relationships from DrugBank. We use reported side-effects of drug combinations form DrugBank as 21,821 binary indicators of interactions between drugs in $\mathbf{\Theta}_4$.

10.2.4 Gene interaction data

We obtain the relationships between genes from five sources of interaction data (top five rows in Table 10.1). Genes are identified by their NCBI gene IDs. We first map the approved gene symbols and Uniprot IDs to Entrez gene IDs using the index files from HGNC database (Seal et al., 2011), downloaded in November 2012. This is done to convert all gene annotations, drug-target, and co-expression data into NCBI IDs. To increase coverage of gene and protein interaction data, we include all genes (or equivalently, proteins) for which at least two supporting pieces of information were available in any of the data sources listed in Table 10.1. In total, these sources include: 55,787 protein-protein interactions (PPIs) between 10,360 proteins ($\mathbf{\Theta}_1^{(1)}$), 869 pairs of co-expressed genes ($\mathbf{\Theta}_1^{(2)}$), 7,517 cell signaling interactions ($\mathbf{\Theta}_1^{(3)}$), 511 human and inter-species genetic interactions ($\mathbf{\Theta}_1^{(4)}$), and 1,505,831 pairs of genes involved in metabolic pathways ($\mathbf{\Theta}_1^{(5)}$).

10.3 Inference of a joint prediction model

We infer human disease-disease associations by integrating a multitude of relevant molecular data sources. We use a data mining approach based on matrix representation of these molecular data, which works by simultaneous matrix tri-factorization and is presented in Chapter 6.

Data fusion for disease-disease association prediction consists of three main steps illustrated in Fig. 10.2 and in Algorithm 4:

- First, we construct relation and constraint matrices from all the available data (Fig. 10.1). Recall that a relation matrix encodes relations between objects of two different types (e.g. gene to Gene Ontology term annotation) and a constraint matrix describes relations between objects of the same type (e.g. protein-protein interactions). The molecular data encoded in these matrices are sparse, incomplete and noisy and some matrices are completely missing because associated data sources are not available (e.g. no link between GO terms and drugs in Fig. 10.2).
- We then simultaneously factorize all the relation matrices under given constraints (Algorithm 3).
- Finally we score statistically significant associations in the matrix decomposition and identify disease classes (Algorithm 4).

The objective function minimized by matrix factorization algorithm (Algorithm 3) enforces good approximation of the input matrices and is regularized by using available

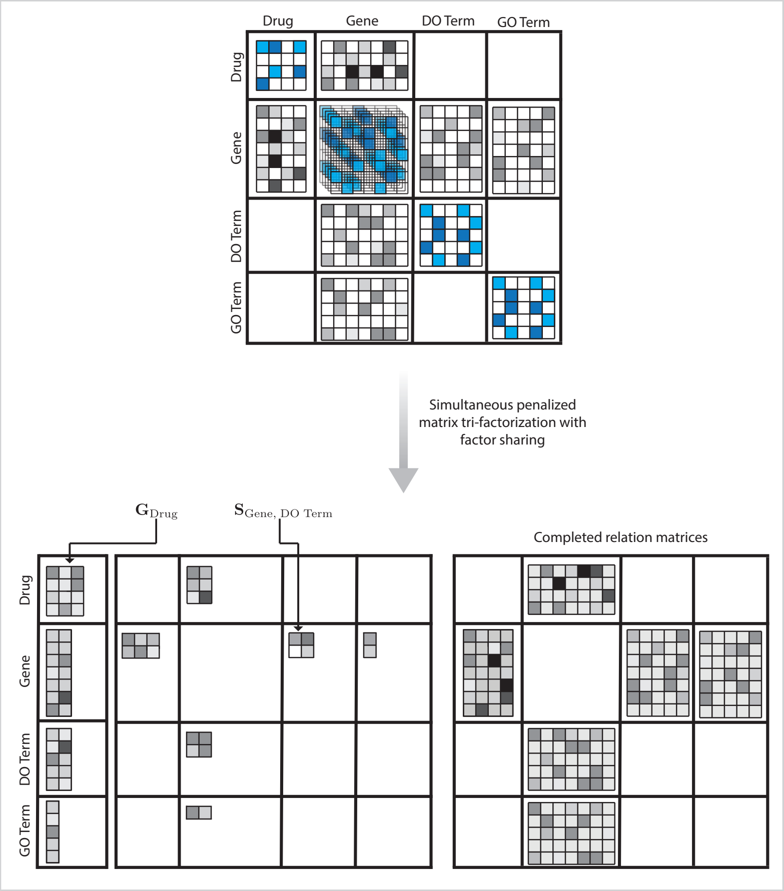

Drug
Gene
DO Term
GO Term
Drug
Gene
DO Term
GO Term
Simultaneous penalized matrix tri-factorization with factor sharing
$\mathbf{G}_{\text{Drug}}$
$\mathbf{S}_{\text{Gene, DO Term}}$
Completed relation matrices
Drug
Gene
DO Term
GO Term

Figure 10.2

A graphical representation of our data fusion by matrix factorization approach to discovering disease-disease associations. (top) The block-based matrix representation exactly corresponds to the data fusion graph in Fig. 10.1. We combine 11 data sources on four different types of objects (see Sec. 10.2): drugs, genes, Disease Ontology (DO) terms and Gene Ontology (GO) terms. These data are encoded in two types of matrices: constraint matrices, which relate objects of the same type (such as drugs if they have common adverse effects) and are placed on the main diagonal illustrated by matrices with blue entries; and relation matrices, which relate objects of different types and are placed off the main diagonal illustrated by matrices with grey entries. The molecular data encoded in these matrices are sparse, incomplete and noisy (depicted by different shades of blue and grey). (bottom, left) The resulting factorized system contains matrix factors that are specific to every type of objects (four matrices in left part; e.g. G_{Drug}), and matrix factors that are specific to every data set (six matrix factors in right part; e.g. $S_{Gene,\ DO\ Term}$). (bottom, right) We use matrix factors to reconstruct relation matrices and complete their unobserved entries.

constraint matrices presented in $\mathbf{\Theta}^{(t)}$. For prediction of disease associations, the input to the data fusion algorithm consists of five constraint block matrices $\mathbf{\Theta}_1^{(t)}, 1 \leq t \leq 5$ due to five sources of interaction data that represent relations between genes, three constraint matrices corresponding to relationships between drugs, diseases and Gene Ontology terms, respectively, and three relation matrices that connect different genes, diseases, drugs and Gene Ontology terms. Recall that Algorithm 3 estimates latent matrices $\boldsymbol{G}_i$ and $\boldsymbol{S}_{ij}$, which we utilize for the identification of disease classes.

Parameters of the fusion algorithm are factorization ranks, k_i, which determine the degree of dimension reduction for four object types in our fusion graph. These factorization ranks are selected from a predefined set of possible values to optimize the quality of the model in its ability to reconstruct the input data from gene-disease relation matrix $\boldsymbol{R}_{12}$. For example, gene-disease profiles of length ≈1,500 in the original space are reduced to profiles with ≈70 factors in data fusion space. We find this approach to be robust and small variations in initial parameter tuning do not impede the overall final quality of the fused system (data not shown). In our study, factorization ranks of 50 to 80 yield models of similar quality. In general, we find that if the data contain meaningful information (as opposed to randomized input), the optimized factorization ranks are much smaller than input dimensions because these data can be effectively compressed, and low-dimensional representation will provide a good estimate of the target relation matrix. Conversely, this would not hold true if we were to predict arbitrarily assigned labels. In that case factorization ranks would have to be substantially larger in order to produce somewhat comparable models.

10.3.1 Disease class assignment

Each factorization run produces a set of matrix factors that reconstruct the three relation matrices in our model. For disease association discovery, we are interested in approximating $\boldsymbol{R}_{12} \approx \boldsymbol{G}_1 \boldsymbol{S}_{12} \boldsymbol{G}_2^T$, specifically factor $\boldsymbol{G}_2$ that contains meta profiles of DO terms and is used to identify classes of diseases. Class membership of a disease is determined by maximum column-coefficient in the corresponding row of $\boldsymbol{G}_2$. This is a well-known approach in applications of non-negative matrix factorization (Brunet et al., 2004; Kim and Tidor, 2003). A binary connectivity matrix $\boldsymbol{C}$ is then obtained from class assignments with $\boldsymbol{C}_{ij}$ set to 1 if disease i and disease j belong to the same

Algorithm 4

Disease class assignment.

```
Input:
    ▪ Latent matrices obtained by $r$ repetitions of factorization given in Algorithm 3,
      $\boldsymbol{G}_2^{(i)}$ for $1 \leq i \leq r$.
Output:
    ▪ A consensus matrix $\bar{\boldsymbol{C}}$,
    ▪ a set $\mathscr{D}$ of disease classes, $\boldsymbol{D}$.

1. Repeat the following for each matrix factor $\boldsymbol{G}_2^{(i)}$ for $1 \leq i \leq r$.
    a. For each disease $j$ compute its class as $\arg\max_m \boldsymbol{G}_2^{(i)}(m, j)$.
    b. Compute connectivity matrix $\boldsymbol{C}^{(i)}$ from class assignment such that $\boldsymbol{C}^{(i)}(r, s)$
       is set to 1 if disease $r$ and $s$ were assigned the same class in step a.
2. Compute consensus matrix as $\bar{\boldsymbol{C}} = \frac{1}{r} \sum_i \boldsymbol{C}^{(i)}$.
3. Extract new disease classes, $\mathscr{D} = \{\boldsymbol{D} \mid \forall i, j \in \boldsymbol{D} \wedge i \neq j \,:\, \bar{\boldsymbol{C}}(i, j) = 1\}$.
```

class (see Algorithm 4). Repeating factorization process 15 times with different initial random conditions and factorization ranks gives a collection of connectivity matrices, $\boldsymbol{C}^{(i)}, i \in 1, 2, \ldots, 15$. These are averaged to obtain the consensus matrix $\bar{\boldsymbol{C}}$ that is then used to assess reliability and robustness of disease associations. The entries in the consensus matrix range from 0 to 1 and indicate the probability that diseases i and j cluster together. If the assignment of diseases into classes is stable, we would expect that the connectivity matrix does not vary among runs and that the entries in the consensus matrix tend to be close to 0 (no association) or to 1 (full consensus for association). To recover informative and relevant disease associations, we are interested in diseases with high values in the consensus matrix. The process is outlined in Algorithm 4.

10.3.2 Disease association scoring

Disease associations are scored by permuting the entries in gene-disease relation matrix $\boldsymbol{R}_{12}$ and inferring the prediction model from the permuted matrix. Matrix $\boldsymbol{R}_{12}$ encodes relations between genes and diseases, and via genes to the rest of the fusion model, so permuting its entries is sufficient for a complete rewiring of associations. To compute the p-values for the disease associations observed in our inferred model, we generate 70 consensus matrices (each one is averaged over 15 permutations of a disease-gene connectivity matrix, giving $70 \times 15 = 1{,}050$ unique matrices) and express the p-value of a particular disease association as the fraction of factorization runs in which it was observed.

10.4 Discovering disease-disease associations by fusing systems-level molecular data

We fuse systems-level molecular data by using our recently developed matrix factorization approach to gain new insight into the current state-of-the-art human disease classification. This large-scale data integration results in 108 highly reliable disease classes (each corresponding to a clique in the consensus matrix, $\bar{\boldsymbol{C}}$; see Algorithm 4). Size distribution of the 108 disease classes is as follows: 60 disease classes contain 2 diseases; 31 disease classes contain 3 or 4 diseases; 9 disease classes contain 5, 6 or 7 diseases; 5 disease classes contain 8, 9 or 10 diseases; 2 disease classes contain 11 or 17 diseases; and 1 disease class contains 146 disease. For each class we examine the associations between its member diseases to inspect how the obtained classes align with currently accepted disease classification.

Using Disease Ontology (DO) and literature curation, we find that the 107 smaller classes successfully capture closely-related diseases that are also placed near each other in DO (see below for details). Also, we find that in the largest identified disease class (i.e. the one containing 146 diseases), the most represented major disease is cancer (31.5%), followed by nervous system diseases (14.4%), inherited metabolic disorders (9.6%) and immune system diseases (5.5%). This class primarily contains diseases of anatomical entity (45.2%), cellular proliferation (25.4%) and metabolic diseases (14.3%), with other major concepts of DO being rarely represented. The large size of this class may reflect the following underlying biases in various data sources — its constituents represent either larger majority groups in DO, or minority groups at a lower level of ontology:

- diseases of anatomical entity, because diseases are often described based on tissue/organ;
- cellular proliferation, because of the heavy enrichment of cancers and the sub-

classification of these into many variant diseases, also possibly driven by rich gene and pathway annotation around cell cycle and proliferation;

- metabolic diseases, because of significant representation of metabolic diseases and significant understanding of metabolic pathways. Metabolic disease is a primary focus for systems modeling and simulation, as much is known from pathways and a wealth of omics data available.

Since the obtained distribution appears unbalanced due to one large class containing 146 disease, we further decompose that class by repeating data fusion analysis on its disease members. This effectively gives us a multi-layer hierarchical breakdown of disease classes (see Fig. 10.3). The large class is broken down into 10 classes (only those observed in all 15 inferred models are taken into account; see Sec. 10.3.2). The distribution of disease class sizes is: 9 disease classes with 2 or 3 diseases, and 1 disease class with 51 diseases. The diseases captured by the 9 smaller classes are: two classes consist of cancer diseases, three consist of inherited metabolic disorders, one contains nervous system diseases, two contain respiratory system diseases, and the last one has cardiovascular system diseases. The largest disease class (containing 51 disease members) is further decomposed into 8 disease classes. The distribution of disease class sizes at this level of hierarchy is: 7 disease classes with 2 or 3 diseases, and 1 disease class with 18 diseases. The diseases captured by the 7 smaller classes are: two classes with immune system diseases, one class with cognitive disorders, one class with acquired metabolic diseases, one with cancer, and the last three were split between cognitive disorders and metabolic diseases. The largest class (containing 18 disease members; again, under the most stringent agreement threshold; see Sec. 10.3.1) is finally decomposed into six conserved diseases (the remaining 12 diseases grouped less reliably under our stringent threshold): lung metastasis, dysgerminoma, serous cystadenoma (cellular proliferation and cancer), abetalipoproteinemia (metabolic disorder), related factor XIII deficiency and plasmodium falciparum malaria.

10.4.1 Significant comorbidity of diseases in captured classes

A comorbidity relationship exists between diseases whenever they affect the same individual substantially more than expected by chance. We want to know whether diseases assigned to the same disease class by our data fusion method exhibit higher comorbidity

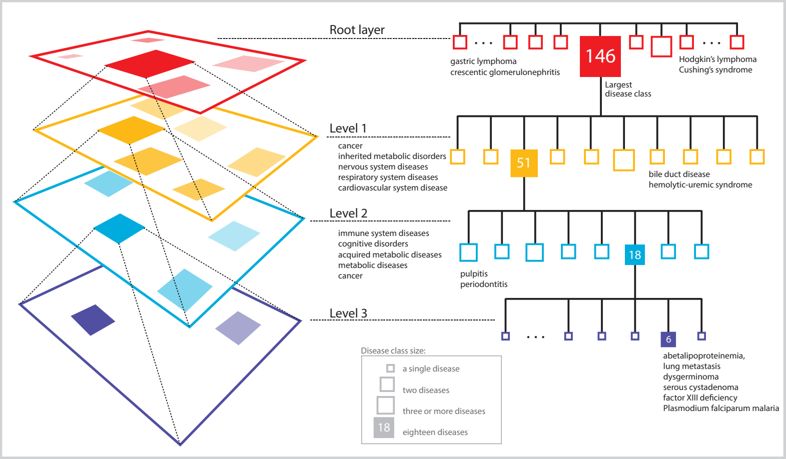


Figure 10.3

Multi-layered hierarchical decomposition of disease classes. Our analysis yields 108 disease classes using the most stringent threshold for predicting disease-disease associations. Identified classes are rather small and each class contains at most 17 diseases with the exception of the largest disease class that consists of 146 diseases (at root layer). We further decompose the largest class by re-running the data fusion on the set of its diseases to identify its fine-grained structure (level one). We repeat data fusion analysis using this top-down strategy two more times (levels two and three) to obtain a hierarchical decomposition of disease classes.

than diseases assigned to different classes. Hidalgo et al. (Hidalgo et al., 2009) proposed two comorbidity measures (`http://barabasilab.neu.edu/projects/hudine`) to quantify the distance between two diseases: a relative risk (defined below) and Pearson's correlation between prevalences of two diseases (ϕ). A *relative risk* (RR) of two diseases is defined as the fraction between the number of patients diagnosed with both diseases and random expectation based on disease prevalence. Expressing the strength of comorbidity is difficult because different statistical distance measures are biased to under- or over-estimating the relationships between rare and prevalent diseases. The RR overestimates associations between rare diseases and underestimates associations involving highly prevalent diseases, whereas ϕ has low values for diseases with extremely different prevalence, but is good at recognizing comorbidities between disease pairs of similar prevalence.

We find that 66 (out of 107) disease classes have a significantly higher comorbidity than what would be expected by chance (p-value < 0.001 with Bonferroni multiple comparison correction applied to all p-values). We assess the statistical significance by randomly sampling disease sets of the same size as the disease class in question, and computing the comorbidity enrichment scores of the sampled sets according to the two comorbidity measures, RR and ϕ, as proposed by Hidalgo et al (Hidalgo et al., 2009). The enrichment score is then computed as the mean of comorbidity values between all disease pairs in a disease class. For subsequent layers of hierarchical decomposition of the largest disease class (i.e. the one containing 146 diseases), we find that: 7 out of 10 first level disease classes have a significantly higher comorbidity (measured by both RR and ϕ) than what would be expected by chance; comorbidity data was available for only 3 out of 8 second-level disease classes, and 2 of them exhibited significantly higher comorbidity than what would be expected by chance.

10.4.2 Evaluating disease classes through Disease Ontology

To see how well our fusion approach captures disease-disease associations already present in the semantic structure of DO, we look at the overlap between 107 disease classes (again, we perform enrichment analysis of the largest above-described class separately, see below) and find that 79 classes have at least 80% of disease members directly connected in DO via is_a relationship; an example of one such disease class is given in Fig. 10.4. We assess the statistical significance of such a high number of classes being enriched in known relations from DO by computing the p-value as follows. First, we remove all DO-related information (i.e. we remove the constraint matrix $\mathbf{\Theta}_2$; see Sec. 10.2) and then we perform the data fusion again without any prior information on relationships between diseases. We find that such a high number of classes is unlikely to be enriched in known relations from DO by chance (p-value < 0.001).

This result is very interesting as it indicates that DO could, in principle, be reconstructed from molecular data only. Our findings suggest that disease classification derived from pathological analysis and clinical symptoms (DO) can be largely reproduced by considering *only* molecular data. In other words, data fusion of different types of evidence could be used to infer a hierarchy of disease relations whose coverage and power might be very similar to those of the manually curated DO.

The decomposition of the largest disease class yields similar results: 5 out of 9 first-level classes have their members directly linked in DO via is_a relationships; 4 out of 7 second-level disease classes have their members directly linked in DO via is_a relationships; the third-level class of size six does not significantly overlap with the DO graph, but is partially supported by literature (Holst et al., 1999).

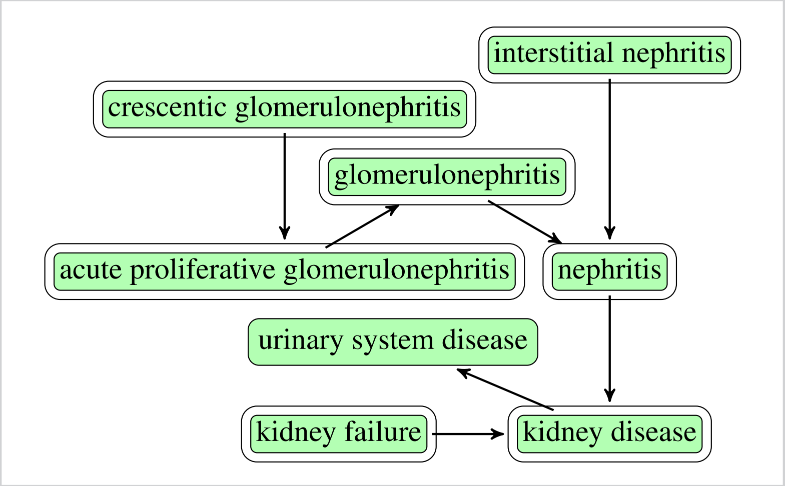


Figure 10.4

An example of disease class predicted by data fusion overlaid with a DO graph. Members of the disease class are outlined. This illustrates the ability of data fusion to successfully capture real disease classes: diseases associated with crescentic glomerulonephritis are presented.

10.4.3 Finding new links between diseases

In addition to examining classes of multiple diseases, we can use our fused model to rank individual disease-disease associations based on supporting molecular evidence, and make novel predictions linking previously seemingly unrelated diseases. Among all the highest-ranked disease-disease associations in the fused model (i.e. disease pairs from the most stable classes — obtained in step 3 of Algorithm 4 — with less than 6 disease members), we find 14 associations not recorded in Disease Ontology. We perform literature curation and find evidence for *all* 14 of the predicted disease associations (Table 10.2). Such high accuracy is due to our choice to take a highly stringent approach that requests the association to be observed in all 15 of the inferred models (see Sec. 10.3.2 for details). Comorbidity data were available for 4 out of 14 predicted disease associations and all 4 of these disease-disease associations were found to have significantly high comorbidity: (DOID:11198, DOID:12336), (DOID:12252, DOID:8543), (DOID:423, DOID:13166), and (DOID:11202, DOID:11335).

Table 10.2

14 predicted disease-disease associations currently not captured by the semantic structure of Disease Ontology. Literature support for them is listed under the column denoted by "References". Reported *p*-values measure how likely it would be for a disease association to emerge if gene-disease relation matrix was permuted, as described in Sec. 10.3.2.

Disease pair	Literature evidence (quoted verbatim from the referenced source)	P-value
vitamin B deficiency (DOID:8449), endogenous depression (DOID:1595)	"Vitamin B complex deficiency causes the psychiatric symptoms of atypical endogenous depression. Dementia and depression have been association with this deficiency possibly from under production of methionine." (Keuter, 1958; Carney et al., 1990)	< 0.001
crescentic glomerulonephritis (DOID:13139), gastric lymphoma (DOID:10540)	"Mixed cryoglobulinemia-associated membranoproliferative glomerulonephritis disclosed gastric MALT lymphoma. Glomerulonephritis and lymphoma tend to co-exist in the same patients (relative risk 34.0; $P < 0.0001$)." (Buob and Copin, 2006; Skopouli et al., 2000; Von Vietinghoff et al., 2006)	< 0.001
thyroid medullary carcinoma (DOID:3973), cholestasis (DOID:13580)	"Paraneoplastic cholestasis and hypercoagulability associated with medullary thyroid carcinoma. Cholestasis is likely a paraneoplastic effect of thyroid medullary carcinoma." (Tiede et al., 1994)	0.001
crescentic glomerulonephritis (DOID:13139), miliary tuberculosis (DOID:9861)	"Complex-mediated diffuse proliferative glomerulonephritis with crescentic formation is associated with miliary tuberculosis. Antituberculous agents successfully treat miliary tuberculosis and recovered renal function." (Kohler et al., 1994; Wen and Chen, 2009)	0.001
thyroid adenoma (DOID:2891), thymoma (DOID:3275)	"Coexistence of bilateral paraganglioma of the A. carotis, thymoma and thyroid adenoma. A common neuroectodermal origin is proposed as an explanation for the coexistence of the carotid body tumor and multiple endocrine tumors." (Refior and Mees, 2000)	0.001
early myoclonic encephalopathy (DOID:308), Angelman syndrome (DOID:1932)	"Angelman syndromes share a range of clinical characteristics, including intellectual disability with or without regression and infantile encephalopathy. It presents in infancy with nonspecific features, such as psychomotor delay and seizures. This can lead to the descriptive labels of cerebral palsy or static encephalopathy." (Willemsen et al., 2012; Dagli et al., 2012)	< 0.001
autoimmune polyendocrine syndrome (DOID:14040), myositis (DOID:633)	"Autoimmune polyendocrine syndrome type 2 (known as Schmidt's syndrome) can be associated with interstitial myositis, an inflammatory myopathy which can be pathologically distinguished from idiopathic polymyositis and inclusion body myositis." (Heuss et al., 1995)	< 0.001
primary hyperparathyroidism (DOID:11202), sarcoidosis (DOID:11335)	"Primary hyperparathyroidism simulates sarcoidosis. Coexisting primary hyperparathyroidism and sarcoidosis cause increased Angiotensin-converting enzyme and decreased parathyroid hormone and phosphate levels." (Lim and Clarke, 2013)	< 0.001
cerebrotendinous xanthomatosis (DOID:4810), viral hepatitis (DOID:1884)	"Mutations in the sterol 27-hydroxylase gene (CYP27A) cause hepatitis of infancy as well as cerebrotendinous xanthomatosis. Accumulation of cholesterol and cholestanol can lead to the xanthomata, neurodegeneration, cataracts and atherosclerosis that are typical of cerebrotendinous xanthomatosis." (Clayton et al., 2002)	< 0.001
lepromatous leprosy (DOID:10887), mental depression (DOID:1596)	"The precipitating causes of relapse in leprosy include mental depression which downgrades immunity. The prevalence of dementia and depression in older leprosy patients is high." (Su et al., 2012)	0.001
male infertility (DOID:12336), DiGeorge syndrome (DOID:11198)	"Complex chromosome rearrangements (CCR) are rare structural chromosome aberrations that can be found in patients with phenotypic abnormalities or in phenotypically normal patients presenting infertility. The malsegregation of CCR can lead to partial 10p12.3 to 10p14 deletion, associated with the DiGeorge like phenotype." (Karmous-Benailly et al., 2006; Christopoulou et al., 2013)	0.001
Cushing's syndrome (DOID:12252), Hodgkin's lymphoma (DOID:8543)	"Hodgkin's lymphoma is highly responsive to steroids and Cushing's syndrome results from over exposure to corticosteroids, so it could be considered a treatment side effect. However, the co-existence in one patient of Cushing's disease (caused by a tumour in the pituitary) that suppressed the Hodgkin's lymphoma has been reported." (Howell et al., 2004)	< 0.001
crescentic glomerulonephritis (DOID:13139), prostate cancer (DOID:10283)	"There can be two potential causes for the association: 1) that the drugs and treatment regimen that cancer patients are on causes the glomerulonephritis, or 2) that features of the cancer may cause the glomerulonephritis with ANCA being associated in both cases." (Von Vietinghoff et al., 2006)	< 0.001
allergic bronchopulmonary aspergillosis (DOID:13166), myopathy (DOID:423)	"Allergic Bronchopulmonary aspergillosis is caused by a fungal disease. Fungal diseases are often treated with triazoles. Drug-induced myopathies are well recognized with triazole class of drugs. The association between these two may therefore be based on the treatment and risk it carries, rather than a common mechanism." (Valiyil and Christopher-Stine, 2010)	< 0.001

10.4.4 Contribution of each data set to the fused model

We have seen that data fusion can successfully retrieve existing and uncover new associations between diseases. Now we examine the contribution of each individual data set to the final disease-disease association model. We estimate the relative importance of each of the fused data sources in predicting disease associations by comparing the quality of the inferred model that includes the data source, to the quality of the model that excludes it. The measured quality is represented by a tuple of residual sum of squares (RSS; lower values are better) and explained variance (Evar; higher values are better; see Žitnik and Zupan (2015a) for details) of gene-disease relationship matrix $\boldsymbol{R}_{12}$ (see Sec. 10.2). So an increase in RSS and a decrease in Evar hinder the quality of the inferred model, and conversely, a decrease in RSS and an increase in Evar improve the quality of the inferred model. We find that omission of each of the five data sources that specify interactions between genes ($\boldsymbol{\Theta}_1^{(1)}, \ldots, \boldsymbol{\Theta}_1^{(5)}$) reduces the overall quality of the model. Surprisingly, the largest model degradation is observed in the absence of genetic interactions when Evar drops by 9.5% and RSS increases by 13.3%. This result is unexpected, because the number of available genetic interactions is small (511). This may confirm the proposed importance of genetic interactions and functional buffering as being critical for understanding disease evolution and for design of new therapeutic approaches (Ashworth et al., 2011). Although the dataset of genetic interactions is currently small, the observed interactions are more likely to be causative as opposed to correlative and may therefore have less noise associated, hence they appear to be more informative and have a larger importance on relationships between diseases than other data sources. Exclusion of other sources results in a smaller decrease in quality (Table 10.3), but nevertheless, these results confirm that all of the fused data sources contribute to the quality of the model.

10.5 Discussion and conclusion

We integrate a wide range of modern systems-level molecular interaction and ontology data using our recently proposed data-fusion approach, and apply it to finding relationships between diseases previously unrecorded in DO. We validate our findings through comorbidity data and literature curation to demonstrate that such a systems-level integration can recover known and successfully identify currently unrecorded relationships

Table 10.3

Relative contribution of each data set to the fused model. Starting from the configuration given in Fig. 10.1, we remove individual data sources, re-run the data fusion algorithm, and compute residual sum of squares (RSS) and explained variance (Evar) changes for the resulting model. For example, if we remove protein-protein interaction data (column labeled "Θ_4"), the quality of the resulting fused model drops by 2.2% (i.e. RSS increases by 2.2% and Evar decreases by 1.3%). The column labeled "$\Theta_4 + R_{14}$" corresponds to the configuration in which we remove all drug-related information from the system, while the one labeled "Θ_4" indicates that only drug side-effects information was removed.

Data set	$\Theta_1^{(4)}$	$\Theta_1^{(2)}$	$\Theta_1^{(3)}$	$\Theta_1^{(5)}$	$\Theta_1^{(1)}$	Θ_4	$\Theta_4 + R_{14}$	Θ_3	$\Theta_3 + R_{13}$
RSS increase (↑)	13.3%	6.3%	2.0%	2.0%	2.0%	2.2%	3.8%	1.0%	1.9%
Evar decrease (↓)	9.5%	4.5%	2.5%	2.0%	2.0%	1.3%	4.6%	1.8%	3.2%

between diseases.

When searching for disease-disease associations not present in DO, we considered only those associations that are present in all of the inferred models. This conservative approach gave us 14 disease-disease association predictions which we validated through literature and comorbidity data. Relaxing the threshold of association to be predicted, i.e. requiring a disease-disease association to be present in 95%, 90%, 85% or fewer of inferred models yields a higher number of predicted disease associations. For instance, we found 89 associations unrecorded by DO when requiring them to be present in at least 80% of the models. Exploring the effects of lowering this threshold remains a subject of future research, as we were able to demonstrate our goal to find potentially useful associations using the most stringent threshold. Specifically, two of the fourteen predicted disease-disease associations — between gastric lymphoma and crescentic glomerulonephritis, and between Cushing's syndrome and Hodgkin's lymphoma — demonstrate the ability of the approach to find interesting novel links, but also highlight the fact that it is not possible to determine causal from correlative relationships (which, indeed, in many cases may not be known), given our current scientific understanding.

Perhaps even more interesting is the fact that the newly identified relations between diseases could, in principle, be used to systematically update and extend DO, or even develop a parallel data-driven hierarchy of disease relations. Utilizing data fusion for disease re-classification, as well as linking these results with genome-wide association studies (GWAS) is a subject open to future research.

We show that all available molecular data — regardless of their sparseness — are important for effective integration. Surprisingly, we find that genetic interaction data are the most predictive underlying factor of disease-disease associations despite their current small size. The flexibility of our data fusion approach allows us to extend the model with new data sources or omit some sources of information to study their effects on predictive performance. We only require that the underlying graph of data fusion graph (Fig. 10.1) be connected. This gives our data fusion algorithm the power to share latent representations of object types between different data sources. For instance, we cannot omit data on drug targets ($\boldsymbol{R}_{14}$ in Fig. 10.1 without also removing data on adverse side-effects of drug combinations ($\boldsymbol{\Theta}_4$). Thus, we report in Results on the quality of all models that exclude any reasonable first-order combination of data sources and use these data to estimate contributions of data sources to the quality of the fused model.

Since our data fusion approach is a semi-supervised learning method, it is less prone to over-fitting than supervised methods, i.e. ones that make distinctions between objects on the basis of predefined class label information. Additionally, in order to avoid over-fitting, we selected data fusion parameters through internal cross-validation and used constraint matrices — which express the notion that a pair of similar objects of the same type, such as a pair of drugs or a pair of diseases, should be close in their latent component space — to impose penalties on matrix factors. Thus, the observed reduction in model quality when any one of the included data sets is omitted is caused by the exclusion of complementary information provided by the data set rather than by the lack of robustness of the model.

We have seen the role of data fusion in successful retrieval of existing and uncovering of novel links between diseases. Future improvements of such a comprehensive integration of molecular data would allow better understanding of underlying mechanisms that drive diseases and would, in turn, improve choice of medical therapy.

Drug toxicity prediction

11

Traditional studies of liver toxicity involve screening compounds through in vivo and in vitro tests. They need to distinguish between compounds that represent little or no health concern and those with the greatest likelihood to cause adverse effects in humans. High-throughput and toxicogenomic screening methods coupled with a plethora of circumstantial evidence provide a challenge for improved toxicity prediction and require appropriate computational methods that integrate various biological, chemical and toxicological data.

We report in this chapter on a data fusion approach for prediction of drug-induced liver injury potential in humans using microarray data from the Japanese Toxicogenomics Project (TGP) as provided for the contest by Critical Assessment of Massive Data Analysis (CAMDA) 2013 Conference. Our aim was to investigate if the data from different TGP studies could be fused together to boost prediction accuracy. We were also interested if in vitro studies provided sufficient information to refrain from studies in animals. We show that our recently proposed matrix factorization-based data fusion provides an elegant computational framework for integration of the TGP and related data sets, twenty-nine data sets in total. Fusion yields a high cross-validated accuracy (AUC of 0.819 for in vivo assays), which is above the accuracy of the established machine learning procedure of stacked classification with feature selection. Our data analysis shows that animal studies may be replaced with in vitro assays (AUC = 0.799) and that liver injury in humans can be predicted from animal data (AUC = 0.811). Our principal contribution is a demonstration that analysis of toxicogenomic data can substantially benefit from data fusion with directly and circumstantially related data sets.

11.1 Background

Drug-induced liver injury (DILI) is the most frequent reason for drug withdrawal during early development and clinical trials as well as after drugs are approved for the marketplace (Lee, 2003). Some drugs are more likely to cause hepatic adverse events than others, and some may even lead to severe liver injuries. Development of tools for early detection of adverse effects and identification of a drug's toxic potential is a major challenge within the pharmaceutical industry and clinical medicine (Chen et al., 2011; Ju and Reilly, 2012; Kaplowitz, 2013). The toxicology and drug safety evaluation communities have made great efforts in developing methodologies to assess drug toxicity risks (Dix et al., 2007; Yang et al., 2008; Shukla et al., 2010). These large-scale efforts also intend to elucidate whether animal studies can be replaced with in vitro assays and if liver injuries in humans can be predicted using toxicogenomic data from animals. Critical Assessment of Massive Data Analysis (CAMDA) (Tilstone, 2003) created a challenge in 2013 to assess the performance of different analytic methods to predict the human hepatotoxic potential of drugs using the Japanese Toxicogenomics Project (TGP) (Uehara et al., 2010) data set. The challenge aimed to foster the development of computational approaches and to promote these within the scope of tools for drug toxicity estimation.

Molecular biology abounds with data from sequencing, expression studies, function annotations, and studies of interactions between genes, proteins and drugs. These data sets are related, and analysis of one data set could benefit from the inclusion of information from others. We proposed in Chapter 6 a data fusion approach that can elegantly integrate heterogeneous data sets, representing each data set in a matrix and fusing the data sets by simultaneous matrix factorization. We focus in this chapter on the fusion of 29 data sets from the TGP and related data repositories to predict DILI risk. We assess the value of combining conventional toxicogenomic data sets with circumstantial evidence for more informed prediction of adverse drug reactions and hepatotoxicity. We compare the accuracy of data fusion to that of a standard multi-classifier approach where we stack four state-of-the-art classification algorithms. We additionally investigate feature subset selection by CUR matrix decomposition applied before combining classifiers with stacking.

11.2 A data collection of 29 data sets

We performed two computational experiments, one with a multi-classifier and the other with a data fusion approach. The multi-classifier approach considered gene expression data sets provided by the Japanese Toxicogenomics Project (TGP), which consisted of two in vivo studies (performed on rats) and two in vitro studies (one performed on rat and one on human cell lines). In addition to gene expression data, the data fusion approach also included data on drugs available from DrugBank (`http://www.drugbank.ca`), gene annotations from Gene Ontology (`http://www.geneontology.`

org), protein-protein interactions from STRING (http://string-db.org), and hematological and clinical chemistry data for each animal and sample metadata information.

Data fusion considered 14 types of objects (nodes in Fig. 11.1, e.g. genes, GO terms, or drugs) and a collection of 29 data sets, each relating a pair of object types (arcs in Fig. 11.1, e.g. gene annotations that relate genes and GO terms). We represent the observations from a data source that relates two distinct object types i and j in a sparse relation matrix $\boldsymbol{R}_{ij}$. For example, the matrix $\boldsymbol{R}_{1,13}$ encodes the annotations of genes from the rat in vivo single dose study. A data source that provides relations between objects of the same type i is represented by a constraint matrix $\boldsymbol{\Theta}_{ii}$ (e.g., $\boldsymbol{\Theta}_{10,10}$ for DrugBank's drug interactions).

Fused data sets in Fig. 11.1 include gene annotations that are encoded in $\{0, 1\}$-matrices $\boldsymbol{R}_{1,13}$, $\boldsymbol{R}_{2,13}$, $\boldsymbol{R}_{3,13}$ and $\boldsymbol{R}_{4,13}$; expression profiles ($\boldsymbol{R}_{1,5}$, $\boldsymbol{R}_{2,6}$, $\boldsymbol{R}_{3,7}$, $\boldsymbol{R}_{4,8}$); hematology, body weight and clinical chemistry data for each rat ($\boldsymbol{R}_{5,12}$, $\boldsymbol{R}_{6,12}$, $\boldsymbol{R}_{12,5} = \boldsymbol{R}^T_{5,12}$, $\boldsymbol{R}_{12,6} = \boldsymbol{R}^T_{6,12}$); array metadata information such as dose level, dosage time and sacrifice time ($\boldsymbol{R}_{5,9}$, $\boldsymbol{R}_{6,9}$, $\boldsymbol{R}_{7,9}$, $\boldsymbol{R}_{8,9}$, $\boldsymbol{R}_{9,5} = \boldsymbol{R}^T_{5,9}$, $\boldsymbol{R}_{9,6} = \boldsymbol{R}^T_{6,9}$, $\boldsymbol{R}_{9,7} = \boldsymbol{R}^T_{7,9}$, $\boldsymbol{R}_{9,8} = \boldsymbol{R}^T_{8,9}$); drug targets ($\boldsymbol{R}_{1,10}$, $\boldsymbol{R}_{2,10}$, $\boldsymbol{R}_{3,10}$, $\boldsymbol{R}_{4,10}$); indication of medical drugs tested with samples ($\boldsymbol{R}_{5,10}$, $\boldsymbol{R}_{6,10}$, $\boldsymbol{R}_{7,10}$, $\boldsymbol{R}_{8,10}$) and structure and categorization of drugs ($\boldsymbol{R}_{10,11}$, $\boldsymbol{R}_{11,10} = \boldsymbol{R}^T_{10,11}$). Constraint matrices encode protein-protein interactions ($\boldsymbol{\Theta}_{1,1}$, $\boldsymbol{\Theta}_{2,2}$, $\boldsymbol{\Theta}_{3,3}$, $\boldsymbol{\Theta}_{4,4}$), drug interactions ($\boldsymbol{\Theta}_{10,10}$) and the semantic structure of the Gene Ontology graph ($\boldsymbol{\Theta}_{13,13}$).

11.2.1 Gene expression data and sample metadata

The TGP (Uehara et al., 2010) created a gene expression database using the Affymetrix GeneChip array to measure the effects of 131 chemicals, mainly medical drugs, on the liver. Approximately 20,000 samples (tissue/drug combinations) were studied both in vivo and in vitro. The in vivo study used the rat as the species of analysis and considered two experimental designs: a single dose study, consisting of multiple time points with multiple dose levels and a repeated dose study, consisting of multiple dose periods with multiple dose levels. The probe level intensity ratios were quantile normalized, corrected for chemical batch effects and summarized using FARMS technique (Hochreiter et al., 2006) to obtain expression values per genes (Clevert et al., 2012). Replicate

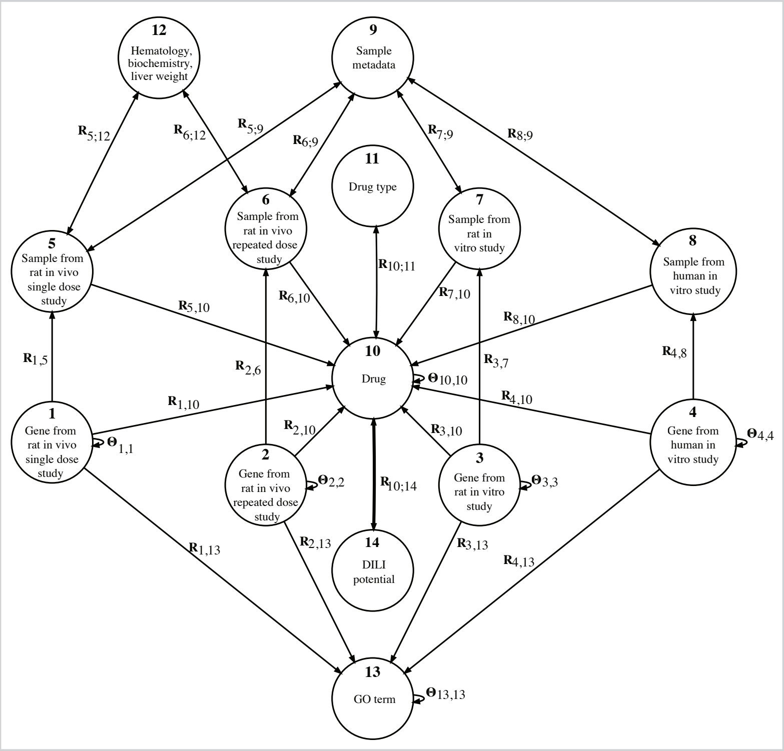

12
Hematology, biochemistry, liver weight
9
Sample metadata
R5;12
R6;12
R5;9
R6;9
R7;9
R8;9
11
Drug type
6
Sample from rat in vivo repeated dose study
7
Sample from rat in vitro study
5
Sample from rat in vivo single dose study
8
Sample from human in vitro study
R10;11
R6,10
R7,10
R5,10
R8,10
R1,5
R2,6
R3,7
R4,8
10
Drug
Θ10,10
R1,10
R4,10
R2,10
R3,10
1
Gene from rat in vivo single dose study
Θ1,1
2
Gene from rat in vivo repeated dose study
Θ2,2
R10;14
3
Gene from rat in vitro study
Θ3,3
4
Gene from human in vitro study
Θ4,4
R1,13
R2,13
R3,13
R4,13
14
DILI potential
13
GO term
Θ13,13

Figure 11.1

Data fusion graph. Nodes represent 14 object types. Arcs denote data sets that relate objects of different type (relation matrices, R_{ij}) or objects of the same type (constraints, Θ_{ii}) for a total of 29 matrices-data sets. The bold arc ($R_{10,14}$, $R_{14,10} = R^T_{10,14}$) represents relation between drugs and DILI potential that we try to augment. Sec Sec. 11.2 for further explanation of the relations.

measurements were collapsed to one measurement per gene, which resulted in 12,088 rat genes and 18,988 human genes. We removed samples whose corresponding chemicals were not annotated with human DILI potential and retained 4,824 samples from the rat in vivo single dose study ($\boldsymbol{R}_{1,5}$), 4,827 samples from the rat in vivo repeated dose study ($\boldsymbol{R}_{2,6}$), 2,424 samples from the rat in vitro study ($\boldsymbol{R}_{3,7}$) and 1,116 samples from the human in vitro study ($\boldsymbol{R}_{4,8}$). For each sample we considered seven metadata features ($\boldsymbol{R}_{5,9}$, $\boldsymbol{R}_{6,9}$, $\boldsymbol{R}_{7,9}$, $\boldsymbol{R}_{8,9}$), including animal sacrifice period, dose and dose level, animal age in weeks and sex type.

11.2.2 Histological and clinical chemistry data

Data obtained from each animal in single dose and repeated dose TGP studies included histopathology, animal weight, food consumption, hematology and blood chemistry. For each animal sample we included 41 attributes ($\boldsymbol{R}_{5,12}$, $\boldsymbol{R}_{6,12}$) describing hematology, such as the levels of monocytes and lymphocytes, biochemistry, such as the concentration of albumin (RALB), direct bilirubin (DBIL) and total bilirubin (TBIL), and body and liver weight.

11.2.3 Drug data

We obtained drug information from the DrugBank (Knox et al., 2011) database. We related drugs to their gene targets (binary matrices $\boldsymbol{R}_{1,10}$, $\boldsymbol{R}_{2,10}$, $\boldsymbol{R}_{3,10}$, $\boldsymbol{R}_{4,10}$) and assigned structural groups (binary matrix $\boldsymbol{R}_{10,11}$). We considered joint adverse effects of drug pairs and DILI risk class co-membership of drugs and included them in the training set ($\boldsymbol{\Theta}_{10,10}$). A constraint between a pair of drugs was set to $(-1)^c k/10^{-3}$, where k was the number of joint adverse effects of a drug pair and c indicated if the two drugs belonged to the same class of DILI risk. The DILI severity in humans was determined for 101 out of 131 drugs based on FDA-approved drug labeling (Chen et al., 2011). Each drug was assigned to one of three categories resulting in 41 drugs of severe DILI concern, 51 drugs of moderate DILI concern and 8 drugs of mild or no DILI concern.

11.2.4 Protein-protein interaction data

We included protein-protein interactions from the STRING (Franceschini et al., 2013) database as constraints between corresponding genes. Degrees of interaction were represented with STRING confidence scores and used to populate constraint matrices, $\boldsymbol{\Theta}_{1,1}$, $\boldsymbol{\Theta}_{2,2}$, $\boldsymbol{\Theta}_{3,3}$, $\boldsymbol{\Theta}_{4,4}$.

11.2.5 Gene Ontology data

We considered gene annotations from Gene Ontology (GO) (Ashburner et al., 2000). We extracted 7,056 GO terms to populate binary relation matrices $\boldsymbol{R}_{1,13}$, $\boldsymbol{R}_{2,13}$ and $\boldsymbol{R}_{3,13}$ with 169,816 rat gene annotations and 288,764 human gene annotations to construct matrix $\boldsymbol{R}_{4,13}$. The hierarchical structure of GO ($\boldsymbol{\Theta}_{13,13}$) was included by reasoning over *has_part*, *part_of* and *is_a* relations in the GO graph. A constraint between a pair of GO terms was set to -0.2^{hops}, where $hops$ was the length of the shortest path between the two GO terms.

11.3 A factorial data fusion approach

We applied data fusion to infer relations between drugs and DILI potential. This relation, encoded in target matrix $\boldsymbol{R}_{10,14}$, was observed in the context of all other data sets. Matrix $\boldsymbol{R}_{10,14} \in \mathbb{R}^{131\times 3}$ was a $[0,1]$-matrix that was only partially observed. Its entries indicated the degree of membership of drugs to the three DILI severity classes.

Our approach involves three main steps:

1. First, data are encoded in constraint and relations matrices as specified by the data fusion graph in Fig. 11.1.

2. In the second step, relation matrices $\boldsymbol{R}_{ij}$ are simultaneously factorized under constraints given by $\boldsymbol{\Theta}_i$. Recall that every relation matrix is decomposed into a product of three low-rank matrix factors, such that a relation matrix $\boldsymbol{R}_{ij}$ is approximated as $\widehat{\boldsymbol{R}}_{ij} = \boldsymbol{G}_i \boldsymbol{S}_{ij} \boldsymbol{G}_j^T$ using the collective matrix factorization presented in Chapter 6. Constraint matrices serve to regularize the low-rank approximations of relation matrices. The key idea of the data fusion approach is

sharing low-rank matrix factors between relation matrices that describe objects of common type. For instance, the latent matrix factor of drugs, $\boldsymbol{G}_{10}$, is shared between decompositions of all relation matrices in Fig. 11.1 whose arcs point to the drug node but the matrix factor $\boldsymbol{S}_{7,10}$ is used only in reconstruction of the corresponding relation matrix between in vitro samples performed on rat cell lines and drugs. The resulting fused system contains latent matrices $\boldsymbol{S}_{ij}$ that are specific to every relation matrix (data source) and latent matrices $\boldsymbol{G}_i$ that are specific to every object type.

3. Finally, we use matrix factors to complete unobserved entries in relation matrices and to transform new objects to the fused latent space. In this study, we aim to predict the unobserved entries in $\boldsymbol{R}_{10,14}$. The DILI severity of d-th drug is determined as $\arg\max_i \widehat{\boldsymbol{R}}_{10,14}(d,i)$. Predictions for d-th drug in the binary classification problem of severe DILI risk against moderate or mild DILI risks are estimated by $\widehat{\boldsymbol{R}}_{10,14}(d,2)/\widehat{\boldsymbol{R}}_{10,14}(d,:)$.

11.4 A multi-classifier approach and CUR matrix decomposition

We employed CUR matrix decomposition (Mahoney and Drineas, 2009) to identify a small set of information carrying genes. CUR matrix decomposition approximates target matrix $\boldsymbol{A}$ in an unsupervised manner as $\boldsymbol{A} \approx \boldsymbol{C}\boldsymbol{U}\boldsymbol{R}$, where $\boldsymbol{C}$ and $\boldsymbol{R}$ are low-dimensional matrix factors that contain a subset of columns and rows from $\boldsymbol{A}$, respectively. The advantage of CUR decomposition over some well known low-rank matrix decompositions such as principal component analysis (PCA) or singular value decomposition (SVD) is its explicit representation in terms of a small number of actual columns and rows of target data matrix. The CUR decomposition-selected features corresponded to original gene expression profiles instead of their linear combinations as with PCA and SVD. We then applied several state-of-the-art classifiers to predict the DILI concern in humans from the matrix factor $\boldsymbol{C}$ obtained for each toxicogenomic study separately. We used gradient tree boosting (Friedman, 2002), random forests (Breiman, 2001) and a support vector machine with polynomial kernel to predict drug-induced toxicity. Output class probabilities generated by the classifiers were combined through stacking to compensate for classifier biases (Wolpert, 1992). Stacking took as input predicted class probabilities and generalized over them with logistic regression, which increased the accuracy of the best of the individual classifiers, reduced the variance and prevented overfitting. It was shown that relatively simple combiners that can avoid overfitting on highly correlated input models often produce most accurate results (Džeroski and Ženko, 2004; Reid and Grudic, 2009).

11.5 Experimental setup

The performance of above described modeling techniques and fusion scenarios was assessed through 10-fold cross-validation and evaluated with the area under the receiver operating characteristic curve (AUC). The AUC score represents the probability that, given a pair of randomly drawn drugs from the positive and negative classes, a predictor predicts higher probability for the positive drug than for the negative drug. The AUC is robust to class imbalance and is not biased against minority class (Guo et al., 2008).

In the multi-classifier approach, we considered the problem of predicting drug-induced toxicity as a binary classification of severe DILI concern against moderate or mild DILI potential. In order to compare the performance of data fusion to multi-classifier approach we casted predictions made by fusion into a binary problem as was done for the multi-classifier experiments. Feature subset selection for the multi-classifier approach was performed within cross-validation on a training data set. Parameters of the classification algorithms, such as the number of iterations and the sizes of the constituent trees in gradient boosting trees, the penalty parameter in support vector machine and the regularization term in logistic regression, were estimated through internal cross-validation on the training data.

The matrix decomposition algorithm used in data fusion required a 14-tuple of factorization ranks, one value per object type, which were selected from a predefined set of values by estimating the quality of low-rank fit of target matrix $\widehat{\boldsymbol{R}}_{10,14}$ using explained variance (Evar) and residual sum of squares (RSS). Initial values of matrix factors were set uniformly at random. The algorithm terminated when the improvement in convergence of target matrix approximation between consecutive iterations measured as the Frobenius distance was below $1{\times}10^{-5}$.

Table 11.1

Predictive performance of the multi-classifier approach for DILI potential prediction with and without CUR dimensionality reduction. Ten-fold cross-validated AUC scores are reported. Acronyms: RF - random forests (Breiman, 2001), GBT - gradient boosting trees (Friedman, 2002), LR - logistic regression, SVM - support vector machine (polynomial third degree kernel).

Stacking using LR		human	rat	rat	rat
Base predictors	Projection	in vitro	in vitro	in vivo single	in vivo repeated
RF, GBT, LR, SVM	PCA	0.741	0.765	0.748	0.761
RF, GBT, LR, SVM	CUR	0.758	0.755	0.764	0.778

11.6 Drug-induced liver injury prediction

Next, we evaluate the predictive performance of the factorial data fusion approach against an established multi-classifier approach based on stacked generalization. We then examine different low-dimensional data projections, which serve to reduce the data dimensionality and to select informative gene profiles. We conclude by investigating the effects of individual data sets on the overall predictive power.

11.6.1 Comparison to a multi-classifier approach with feature selection

Our first experiment focused on a multi-classifier approach to predict DILI risk from the preprocessed TGP microarray data. In particular, we used stacked generalization (Wolpert, 1992) to combine predictions of random forests (Breiman, 2001), gradient boosting trees (Friedman, 2002), logistic regression and support vector machines (Cortes and Vapnik, 1995) (Table 11.1).

We applied gene filtering to perform feature selection and to identify genes with high statistical leverage. We applied the CUR matrix decomposition (Mahoney and Drineas, 2009) of the TGP microarray data sets for gene subset selection. CUR decomposition computes leverage scores for matrix columns (i.e. genes) and uses them for weighted column sampling, preferring those columns with a larger score and assembling a lower-dimensionality matrix. Statistical leverage scores capture the influence of genes on the best low-rank fit of gene expression matrix. Table 11.2 shows the top ten genes with highest normalized statistical leverage as computed separately from animal in vitro and in vivo data.

Table 11.2

Genes with the most influence on the fit of low-rank CUR decomposition of rat in vitro and rat in vivo expression data. Higher values indicate the higher statistical leverage of a gene.

Rat in vitro		Rat in vivo, single dose	
Gene symbol	Leverage score	Gene symbol	Leverage score
Cyp1a1	0.671	*Fam111a*	0.972
Angptl4	0.121	*RGD1309362*	0.953
Cyp4a3	0.119	*Aldh1a7*	0.919
Gdf15	0.086	*Ephx2*	0.906
Chac1	0.086	*Ubd*	0.873
Ctgf	0.084	*Ilf3*	0.735
Acta1	0.080	*Ifit1*	0.714
Hmgcs2	0.079	*Hamp*	0.664
G0s2	0.075	*Akr1c12*	0.565
Ccl20	0.074	*RT1-Bb*	0.492

11.6.2 A data fusion-based approach

We used data fusion by matrix factorization (Sec. 11.3) to integrate various data sets. Data sets are represented as matrices, each relating objects of two types. We considered objects such as genes, gene ontology (GO) terms, drugs, and tissue samples. For instance, genes and tissue samples from rat in vivo single study are related through corresponding gene expression data. Genes and drugs are related through a matrix of drug targets. All together, we consider 29 data sets that provide relations between 14 object types (Fig. 11.1). Data fusion simultaneously considers all data sets (relations) in the factorization schema and factorizes them into substantially smaller relation matrices.

Our target relation in this system was a drug's DILI potential, which describes various degrees of drug toxicity. Toxicity was provided for 101 drugs and expressed as severe, moderate or mild. In a cross-validation study, a subset of considered drugs was excluded to serve for testing of predictions of the data fusion model developed from remaining drugs and all other data sets in the factorization schema. In particular, given latent matrix factors inferred from the training data and a new drug, we

Table 11.3

Predictive performance of fusing various subsets of assays for DILI potential prediction. Ten-fold cross-validated AUC scores are reported.

Fused studies	AUC
In vivo studies	0.819
In vitro studies	0.790
Human in vitro study	0.793
Animal in vitro study	0.799
Animal studies	0.811
Human studies	0.792
All studies	0.810

estimated drug's latent profile by transforming available relations about it to inferred latent space and then used the estimated profile to predict the target relation, namely drug's DILI potential. In that way, we avoided the unwanted information flow between the training and test sets. Table 11.3 shows the 10-fold cross-validated accuracy for seven data fusion scenarios that considered various data sets of the complete fusion model from Fig. 11.1. The model inferred from all four TGP studies used all available data sets. Other models considered only selected toxicogenomic studies and associated non-expression data. For instance, fusion of in vivo assays omitted all data sets from in vitro studies (object types 3, 4, 7, and 8).

11.6.3 Effect of circumstantial data on latent model quality

We estimated the effect of circumstantial data (gene annotations, drug structural information, hematology data, sample metadata) on the quality of the fused factorized model. We observed the reconstruction quality of the target data set, which related drugs to DILI risk, through explained variance (Evar) and residual sum of squares (RSS). Better models have high Evar and low RSS. The influence of the data set was determined by observing the change in reconstruction quality when this data set was excluded from training. Reconstruction of DILI potential when considering the entire collection of data sets achieved Evar of 0.911 and RSS of 8.779 in 10-fold cross-validated study when the entire collection of data sets was considered. The reconstruction quality decreased by 1.0% in Evar and 11.7% in RSS when omitting the data on hematology, biochemistry and liver weight (type 12; Fig. 11.1) from the entire collection of data sets. In contrast, we observed a 9.6% decrease in Evar and a 12.8% increase in RSS when excluding array metadata (type 9; Fig. 11.1) from the collection, and a 0.7% decrease in Evar and 9.4% increase in RSS without considering related drug data (type 11; Fig. 11.1). Exclusion of gene annotations (type 13; Fig. 11.1) slightly worsened the model with respect to Evar (a decrease of 0.2%) but improved RSS by 0.3%.

11.7 Discussion

From a computational perspective, our contributions are two-fold. First, we evaluated the performance of unsupervised matrix decomposition to select genes that exhibit high statistical leverage and employed a reduced data set using well-established classification ensemble methods. Second, we pursued a novel data fusion approach based on matrix factorization to assess the hepatotoxic risk associated with individual drugs by fusing gene expression profiles with a plethora of related and heterogeneous data sets.

In our first experiment we considered the DILI prediction problem for each study separately and pursued a multi-classifier approach (Table 11.1). The training data consisted of microarray profiles (independent variables) and associated drug with a given DILI potential (dependent variable). Feature subset selection by CUR matrix decomposition substantially reduced the number of input features. For instance, and as averaged across iterations of cross-validation, a subset of only about 300 genes were used for training the prediction models in the human in vitro study instead of the original 18,988 genes included by FARMS summarization. The solid performance of multi-classifier approach was not surprising (Džeroski and Ženko, 2004; Pandey et al., 2010) as several previous studies (Pessiot et al., 2013; Bowles and Shigeta, 2013) on this data have already reported good results with single classification algorithms such as support vector machines or gradient boosting. In our case the performance was boosted by both feature selection and classifier ensembling. Also of note is the comparable performance of data preprocessing by CUR factorization and principal component analysis (PCA). As CUR performs feature selection rather than feature transformation, it could be a preferable procedure to identify gene biomarkers (Table 11.2).

Results in Table 11.1 show that using repeated dose studies (rat, in vivo repeated) when forecasting the toxic potency of compounds in humans yielded better results than employing single dose animal studies (rat, in vivo single). According to Greim et al. (2006) and Blaauboer and Andersen (2007) repeated dose studies in animals represent critical data for hazard identification and risk assessment in humans. They claimed that the 28-day toxicity study, which was also used by the TGP, is the minimum requirement to evaluate the organ specific effects of compounds. Our results of the multi-classifier approach show that in the absence of such information, the assessment of continuous human exposure to hazardous compounds is incomplete.

For an integrative approach that simultaneously considers all available experimental and circumstantial data, we use data fusion by matrix factorization (Algorithm 3), an intermediate data integration approach that is able to fuse heterogeneous data sets. Intermediate integration is often the preferred integration strategy (van Vliet et al., 2012; Gevaert et al., 2006; Lanckriet et al., 2004b) as it embeds the structure of the data into a predictive model and thus often achieves higher accuracy. Data fusion surpassed the accuracy of the multi-classifier approach for predicting DILI potential in humans (Table 11.3). The most accurate model was inferred by fusing in vivo assays, which scored an AUC of 0.819. It is surprising that in vivo assays, which relied on an animal model, performed better than human assays, given the aim was to predict DILI potential in humans. However, Pessiot et al. (2013) similarly observed that using in vivo animal data was more informative than using in vitro data from humans. Their AUC scores obtained by a linear support vector machine classifier and inferred from separate toxicogenomic studies were surpassed by those reported by our fusion-based approach.

The fusion-based model inferred from animal assays (two in vivo studies and one in vitro study) outperformed the model obtained by fusing human assays only (one human in vitro study), with the first achieving an AUC of 0.811 and the latter an AUC of 0.792. One might expect that the administration of drugs to animal models would fail to identify the risk of liver injury for drugs prescribed to humans due to differences in metabolic pathways and the current lack of suitable animal models that reproduce human risk factors (Kaplowitz, 2013). Our results do not confirm this hypothesis;

however, differences in performance are small and further investigations seem worthwhile.

The study of influence of data sets on the reconstruction quality of target relation between drugs and DILI risk (see Sec. 11.6.3) showed that, though some data sets were small in their size, they substantially affected reconstruction of target relation. For example, sample metadata included only seven features, such as information about animal sacrifice period and dose level, yet its exclusion from data fusion resulted in a near 10% decrease in reconstruction quality of target relation. In contrast, we observed only a slight reduction in model quality when gene annotation data were omitted from the fused model despite annotation data recording associations to more than 7,000 GO terms.

11.8 Conclusion

Although gene expression profiling is an accepted approach for identifying drugs with potential safety problems (Uehara et al., 2010), our results suggest that integrating expression profiles with circumstantial data on drugs, arrays and genes can further improve predictive performance of analytic approaches and pinpoint the mechanisms that underlie drug toxicity. Our data fusion approach should be applicable to other toxicity endpoints, such as neurotoxicity, or mechanisms of action, such as regenerative hyperplasia. We anticipate that efforts in data analysis hold the promise to replace animal studies with in vitro assays and predict the outcome of liver injuries in humans using in vitro animal toxicogenomic data.

Part V

Regression by data fusion

Factorial survival regression

12

Any knowledge discovery could in principal benefit from the fusion of directly or even indirectly related data sources. In this chapter we explore whether data fusion by simultaneous matrix factorization could be adapted for survival regression. We propose a new method that jointly infers latent data factors from a number of heterogeneous data sets and estimates regression coefficients of a survival model. We have applied the method to CAMDA 2014 large-scale Cancer Genomes Challenge and modeled survival time as a function of gene, protein and miRNA expression data, and data on methylated and mutated regions. We find that both joint inference of data factors and regression coefficients and the data fusion procedure are crucial for performance. Our approach is substantially more accurate than the baseline Aalen's additive model. Latent factors inferred by our approach could be mined further; for CAMDA challenge, we found that the most informative factors are related to known cancer processes.

In Chapter 6 we described a data fusion approach called DFMF ("data fusion by matrix factorization") that jointly factorizes possibly many data matrices into products of low-dimensional matrix factors in a way that latent matrices are shared between factorizations of related data matrices. So far, we reported the utility of DFMF in functional genomics (Chapter 7), gene prioritization (Chapter 9), inference of new diseases associations (Chapter 10), and drug-induced liver injury prediction (Chapter 11). Next, we extend DFMF in a supervised manner to perform survival regression.

12.1 Background

Identification of driving events and their hazard rates for cancer progression remains a major challenge in cancer studies (Garraway and Lander, 2013). Recently, initiatives such as The Cancer Genome Atlas (TCGA) (Collins et al., 2007) and International Cancer Genome Consortium (ICGC) (Hudson et al., 2010) were launched to coordinate large-scale cancer genome studies across different cancer types and subtypes of clinical importance. They collect data that span patients, cancer types and diverse biological data types to address the richness of genomic and molecular mechanisms that play critical roles during cancer development. Importantly, these include data from matched tumor and non-tumor tissues (Pleasance et al., 2009). Rich, diverse, large and complex data sets generated within cancer genome projects now require computational methods that can collectively address them, provide interpretations on the genome-scale, and further integrate them with other genomic, clinical and functional information.

One of the fundamental goals of bioinformatics approaches in cancer studies is cancer subtype classification (Yuan et al., 2011; Network et al., 2011; Hofree et al., 2013; Pal et al., 2014), whereby a heterogeneous population of tumor samples is partitioned into biologically and clinically meaningful subtypes. Stratification of tumors is typically determined by the similarity of molecular profiles and correlated with clinical phenotypes including patient survival time and response to chemotherapy. Most current attempts to stratify tumors have used a single source of biological information and have derived molecular profiles from mRNA expression data (Reis-Filho and Pusztai, 2011; Pal et al., 2014), somatic mutations (Greenman et al., 2007; Alexandrov et al., 2013) or methylation data (Gifford et al., 2004). They have discovered informative subtypes in diseases such as breast cancer and glioblastoma but have also reported a lack of correlation between derived profiles and clinical phenotypes in certain cancer types, including colorectal and lung tumors (Network et al., 2011, 2012). These shortcomings might be due to data incompleteness, noise inherent to biological measurements and limitations of data analysis methods.

Although individual data sets have long been used to stratify patients, stratification based on multiple types of data, such as expression, methylation and somatic mutation profiles, has been more challenging. These data sets are fundamentally different from each other, both in type and in structure. Somatic mutation profiles are extremely sparse and dispersed since typically only a small fraction of genes are mutated and patients diagnosed with the same cancer type share few, if any, mutations (Lawrence et al., 2013). On the other hand, methylation, miRNA expression and gene expression measurements assign quantitative values to nearly all markers, miRNAs and genes, respectively, in every patient. These data also naturally come at different levels of granularity and describe distinct biological data types, such as genes, proteins, miRNAs and methylation markers, among others. Heterogeneity of data generated by an increasing number of cancer studies hence limits the usage of naive computational approaches that either cannot be applied to such data or have to discard potentially beneficial biological information.

Here we report that the problems that stem from data diversity can be largely surmounted by data fusion, which can collectively consider a plethora of data sets coming from both directly and indirectly related data domains and can provide gains in accuracy through data integration. We focus on the prediction of patient survival time and the identification of crucial clinical and molecular features. We propose a new machine learning approach that can consider a potentially large number of heterogeneous data sets to infer latent factors for a survival regression model. Its principal innovation is simultaneous inference of patient profiles and estimation of the influence of latent factors on patient survival time. Below we describe the key concepts behind the proposed approach and demonstrate its high predictive accuracy in three ICGC cancer studies.

12.2 Overview of survival regression by data fusion

We introduce in this chapter a method called DFMF-SR ("data fusion by matrix factorization for survival regression") that couples Aalen's additive model for survival regression and matrix factorization-based data fusion into a joint inference procedure. The principal novelty of the approach is the establishment of interdependence between Aalen's time-varying regression coefficients and fused latent matrix factors during model inference. Intuitively, in each iteration of the algorithm, current estimates of patients' survival time influence the optimization of latent matrix factors and vice-versa.

Fig. 12.1 shows an exemplar data fusion graph of eight data sets together with patient survival data and their corresponding latent matrix factors as inferred by DFMF-SR. We summarize relationships present in every data set ($\boldsymbol{R}_{ij}$) with a mapping from objects, i.e. the units of analysis, to sets of objects called latent factors (columns in $\boldsymbol{G}_i$ and $\boldsymbol{G}_j$) and pairwise relations between latent factors themselves ($\boldsymbol{S}_{ij}$). The inference process aims at identifying objects that are similar to each other in terms of their affiliation with latent factors. Similar objects are mapped to the same latent factor. Individual objects are allowed to instantiate similarity patterns with multiple latent factors.

Overall, the goal of analysis with DFMF-SR is to identify the mapping of objects to a fixed number of latent factors, the pairwise relations among the factors, and regression coefficients of the survival model. The latter are optimized against good prediction of hazard rates using the mapping of individuals to latent factors. It should be noted that latent factors are inferred simultaneously for all objects and every object type in the sys-

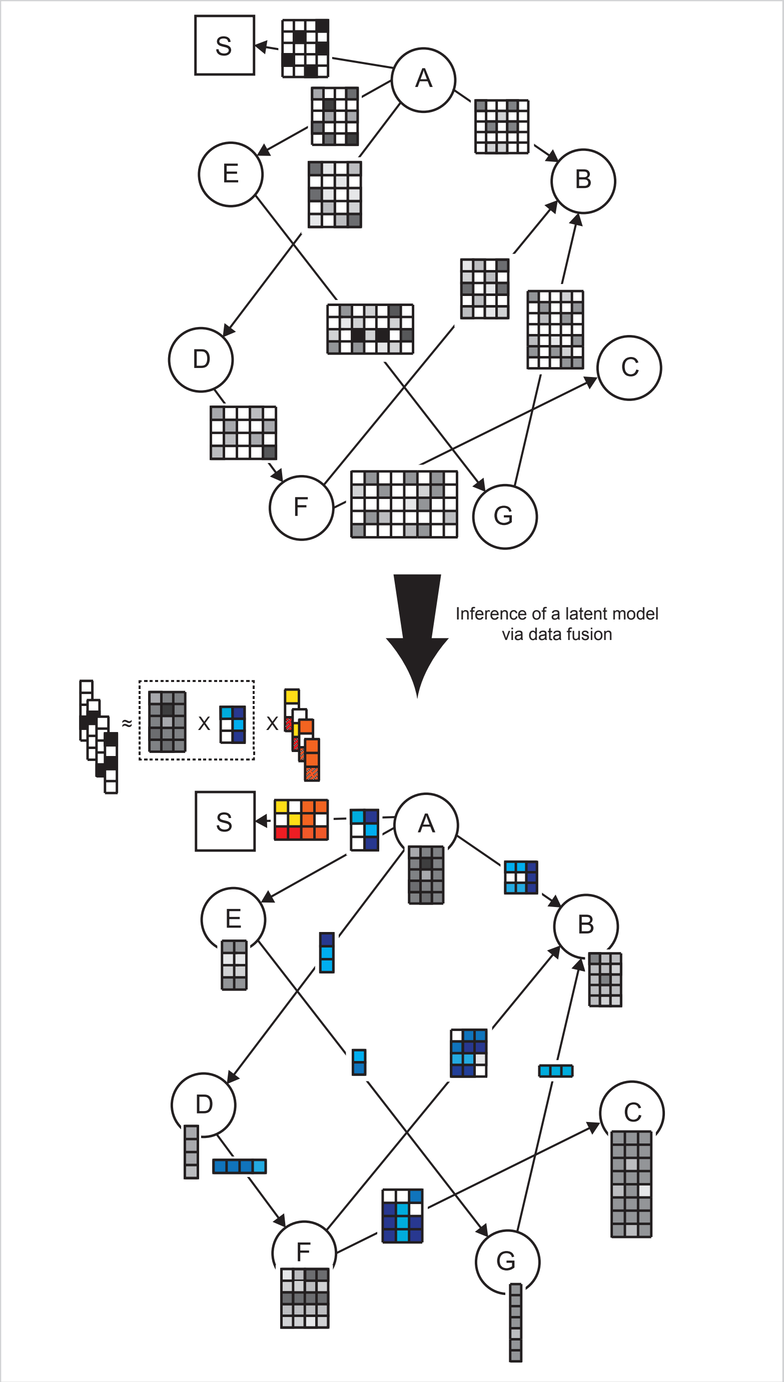


Figure 12.1

Example illustrating survival regression by data fusion (DFMF-SR). The top panel shows the data fusion graph. Nodes in the fusion graph correspond to different types of objects considered by the system. Edges represent data matrices that describe relationships between objects of different types. For example, rows of matrix ("A", "E") correspond to objects "A" and columns agree with objects of type "E". A designated node "S" in the square box serves for the times of the events. Matrix ("A", "S") contains patient survival data. It is a binary matrix indicating the times when the respective objects of type "A" experienced the event. Type "A" most often corresponds to patients or tumor samples and hence ("A", "S") encodes the amount of time that has passed from primary diagnosis until patient's death. DFMF-SR naturally interleaves collective matrix factorization with estimation of survival regression coefficients. The bottom panel shows a latent data model inferred by DFMF-SR. Let us assume data matrix ("A", "E") was selected as a data set whose latent factors are used in the survival model. In each iteration of DFMF-SR, the current tri-factorization of ("A", "E") is updated towards both better reconstruction of the matrix ("A", "E") and improved accuracy of the survival model. Parameterization of the survival model is given by vectors with red and orange entries. The number of vectors corresponds to the number of time points in the survival data. Each vector holds information about the importance of all latent factors on survival up to the respective time point. The dimensionality of each vector corresponds to the number of latent factors in ("A", "E"), i.e. the number of columns in the matrix with blue entries, plus one. An additional entry in each vector is reserved for the time-varying baseline hazard for survival.

tem as shown in Fig. 12.1. DFMF-SR couples latent factors with survival coefficients, which are estimated by regressing latent factors against patient survival data. Selection of a data set whose latent factors are used in survival model estimation is done prior to model inference. However, DFMF-SR is flexible in the sense that it allows one to consider for survival analysis the latent representation of any data set included in the system.

Next, we briefly describe the Aalen's additive model for survival analysis and a recent approach to collective matrix factorization, which form the foundation of our work in this chapter. We then present our survival regression model that uses data fusion and latent factor parametrization.

12.3 Preliminaries

Survival analysis studies the relationship between risk factors and a patient's time to the event, e.g., death, cancer relapse. The patient is referred to as right-censored if the event has not yet occurred by the end of the study. Traditional statistical techniques usually cannot be applied because of the skewness of the distribution of patient lifetime data, time-dependent features and data censoring. The survival probability until at least some time point is most often estimated with Kaplan-Meier statistics. When additional patient data are available, such as clinical covariates or information about somatic mutations that are present in the tumor, we can model time to the event through survival regression.

12.3.1 Aalen's additive model of survival regression

Aalen's additive model is an alternative to Cox's proportional hazards model (Aalen, 1989, 1993; Abadi et al., 2011). It has time-varying regression coefficients, poses no assumptions about their parametric form and can provide information about the changing effects of data features on survival. Let $\lambda(t)$ denote a vector of hazard rates for n individuals where $\lambda_i(t)$ denotes the hazard rate of individual i. The additive model is given by $\lambda(t) = \boldsymbol{X}(t)\beta(t)$, where vector $\beta(t) \in \mathbb{R}^{m+1}$ holds the baseline hazard and m regression coefficients that measure the influence of the respective features in $\boldsymbol{X}(t) \in \mathbb{R}^{n\times(m+1)}$. The matrix $\boldsymbol{X}(t)$ is constructed as follows. If the i-th individual is at risk at time t (the event has not yet occurred), then the corresponding row of $\boldsymbol{X}(t)$ contains the individual's feature profile, otherwise it is replaced with an all-zeros row. Aalen's model estimates cumulative regression coefficients defined by $\boldsymbol{B}_i(t) = \int_0^t \beta_i(s)\mathrm{d}s$, $i \in [m+1]$. This is done by finding $\boldsymbol{B}^*(t) = \sum_{t_k<t} \boldsymbol{V}(t_k)\boldsymbol{I}_k$, where t_k are ordered times of events and $\boldsymbol{I}_k$ is a binary vector indicating an individual who experiences the event at time t_k. The matrix $\boldsymbol{V}(t)$ is computed by the least squares formula from $\boldsymbol{X}(t)$.

12.4 Factorized data fusion model for survival regression

Let i and j denote two types of objects, such as genes and Gene Ontology terms, and let there be n_i objects of type i and similarly n_j objects of type j. DFMF-SR considers a collection $\mathscr{R}$ of relation matrices $\boldsymbol{R}_{ij} \in \mathbb{R}^{n_i \times n_j}$, where $\boldsymbol{R}_{ij}$ encodes relations between objects of types i and j, and a collection $\mathscr{C}$ of constraint matrices $\boldsymbol{\Theta}_i^{(l)}$ for $l \in [l_i]$, where $\boldsymbol{\Theta}_i^{(l)}$ is l-th constraint matrix for objects of type i. Similarly to DFMF, DFMF-SR organizes data sets in a data fusion graph, an example of which is shown in Fig. 12.2. DFMF-SR infers latent matrix factors $\boldsymbol{G}_i$ ($\boldsymbol{G}_i \geq \boldsymbol{0}$) and $\boldsymbol{S}_{ij}$ for all i and j, and cumulative regression coefficients $\boldsymbol{B}(t)$ for all time points of the events, $t_1 < t_2 < \cdots < t_n$, by minimizing the following objective function:

$$\sum_{R_{ij}\in\mathscr{R}} \|\boldsymbol{R}_{ij} - \boldsymbol{G}_i \boldsymbol{S}_{ij} \boldsymbol{G}_j^T\|_{\mathrm{Fro}}^2 + \sum_{\Theta_i\in\mathscr{C}} \sum_{l=1}^{l_i} \mathrm{tr}(\boldsymbol{G}_i^T \boldsymbol{\Theta}_i^{(l)} \boldsymbol{G}_i) + \sum_{t_k<t_n} \|\boldsymbol{I}_k - \boldsymbol{G}_p \boldsymbol{S}_{pr}(t_k)\beta(t_k)\|_{\mathrm{Fro}}^2. \quad (12.1)$$

Here, p and r are object types and specify data set whose fused latent representation we use to regress against survival data. The example in Fig. 12.1 uses data set ("A", "E") to regress against survival data ("A", "S"), hence in that example p corresponds to "A" and r to "E" (see also Fig. 12.1). The times t_k in Eq. (12.1) are ordered times of the events and $\boldsymbol{I}_k \in \mathbb{R}^{n_p}$ is a binary vector consisting of zeros except for a one in the position corresponding to an individual who experiences the event at time t_k. In our analysis, p refers to samples and r to features, e.g., protein expression profiles or mutated chromosomal regions.

We expand the objective function in Eq. (12.1) using a trace operator similar to our work in Chapter 6 and derive the iterative multiplicative update rules for the unknowns from the associated Lagrangian L. Derivatives of L with respect to $\boldsymbol{G}_i$ for $i \neq p$ remain the same as in Chapter 6 and thus, their update rules are unchanged. The multiplicative update of latent matrix factor $\boldsymbol{G}_p$ (not shown here) follows from the following expression after some algebraic manipulation:

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{G}_p} \quad = \quad & 2\sum_{j:R_{pj}\in\mathscr{R}}(-\boldsymbol{R}_{pj}\boldsymbol{G}_j\boldsymbol{S}_{pj}^T+\boldsymbol{G}_p\boldsymbol{S}_{pj}\boldsymbol{G}_j^T\boldsymbol{G}_j\boldsymbol{S}_{pj}^T)+ \\ & 2\sum_{j:R_{jp}\in\mathscr{R}}(-\boldsymbol{R}_{jp}^T\boldsymbol{G}_j\boldsymbol{S}_{jp}+\boldsymbol{G}_p\boldsymbol{S}_{jp}^T\boldsymbol{G}_j^T\boldsymbol{G}_j\boldsymbol{S}_{jp})+2\sum_{l=1}^{l_p}\boldsymbol{\Theta}_p^{(l)}\boldsymbol{G}_p+ \\ & 2\sum_{t_k<t_n}(-\boldsymbol{I}_k\beta(t_k)\boldsymbol{S}_{pr}^T+\boldsymbol{G}_p(t_k)\boldsymbol{S}_{pr}\beta(t_k)^T\beta(t_k)\boldsymbol{S}_{pr}^T)-\boldsymbol{C}_p\mathbf{1}_{n_p\times k_p}.\end{aligned} \tag{12.2}$$

Similarly, update rules of latent matrix factors $\boldsymbol{S}_{ij}$ for $i, j \neq p, r$ are the same as those reported in Chapter 6. The rule for $\boldsymbol{S}_{pr}$ is obtained from the associated partial derivative of the Lagrangian L given by:

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{S}_{pr}} \quad = \quad & -2\boldsymbol{G}_p^T\boldsymbol{R}_{pr}\boldsymbol{G}_r+2\boldsymbol{G}_p^T\boldsymbol{G}_p\boldsymbol{S}_{pr}\boldsymbol{G}_r^T\boldsymbol{G}_r-2\sum_{t_k<t_n}\boldsymbol{G}_p(t_k)^T\boldsymbol{I}_k\beta(t_k)+ \\ & 2\sum_{t_k<t_n}\boldsymbol{G}_p(t_k)^T\boldsymbol{G}_p(t_k)\boldsymbol{S}_{pr}\beta(t_k)^T\beta(t_k).\end{aligned} \tag{12.3}$$

To properly formulate the multiplicative update rule of $\boldsymbol{S}_{pr}$, one would need to solve a generalized linear matrix equation (Horn and Johnson, 1991; Bhatia and Rosenthal, 1997; Horn and Johnson, 2012). Such equations are difficult to analyze in their full generality, and necessary and sufficient conditions for the existence of their solutions are not known (Simoncini, 2014). Also, current numerical techniques for solving generalized linear matrix equations are lacking or are not robust in large-scale settings (Simoncini, 2014). We tackle this problem by randomly selecting a particular t_k in each iteration of the DFMF-SR algorithm and its associated term from the last component of the right side of Eq. (12.3). Based on this reduction we update $\boldsymbol{S}_{pr}$ by solving a Sylvester equation, a well-characterized type of linear matrix equation in which the coefficient matrices occur on both sides of the unknown matrix $\boldsymbol{S}_{pr}$.

Finally, Aalen's time-varying coefficients are computed in each iteration of DFMF-SR by regressing current estimates of $\boldsymbol{G}_p\boldsymbol{S}_{pr}(t_k)$ for all t_k against lifetimes ordered by the times of the events with regularized least squares formulation. The parameter selection and stopping criteria of the DFMF-SR algorithm are similar to those of the base DFMF algorithm (Chapter 6).

12.5 Determining assignment of objects to latent factors

DFMF-SR regresses against latent factors in $\boldsymbol{G}_p\boldsymbol{S}_{pr}$. Latent factor in $\boldsymbol{G}_i$, i.e. a column in $\boldsymbol{G}_i$, corresponds to a group of objects of type i. Since a latent factor does not directly represent any individual object, it is not readily interpretable in a biologically meaningful manner. To decipher the meaning of any latent factor, we wish to identify objects that are associated with it. By definition, the elements in $\boldsymbol{G}_i$ can only take nonnegative values and represent object membership strengths to latent factors. Membership strengths are real-valued due to the relaxation of orthogonality constraints on $\boldsymbol{G}_i$ in DFMF. Therefore, from the values in $\boldsymbol{G}_i$ for a given latent factor c we can determine, which objects are most important and have the greatest membership to factor c. Specifically, object x of type i belongs to a factor c if $c = \arg\max_{\tilde{c}} \boldsymbol{G}_i(x, \tilde{c})$.

12.6 Data and experimental setup

We consider large-scale cancer studies of three cancer types selected for the CAMDA 2014 Challenge in the 15.1 release of the International Cancer Genome Consortium (ICGC; `http://dcc.icgc.org`) (Hudson et al., 2010). These are head and neck squamous cell carcinoma (HNSC; 368 donors), kidney renal clear cell carcinoma (KIRC; 505 donors) and lung adenocarcinoma (LUAD; 461 donors). The ICGC provides data from matched tumor and non-tumor tissues. For each cancer type, data include protein, miRNA and normalized gene expression values, genome-wide information on the state of methylated fragments, somatic mutations and clinical annotation. We consider these data sets alongside Gene Ontology annotations, amounting to a total of ten data sources (Fig. 12.2) for each cancer study. The base object type (p) is given by tumor samples that are associated with survival data based on the donor's last known vital status ("donor's vital status") and the interval from primary diagnosis to the last follow-up date in months ("donor's interval of last follow-up").

We evaluate the performance of survival models by leave-one-out cross-validation of tumor samples and score the models based on predicted survival times. We report

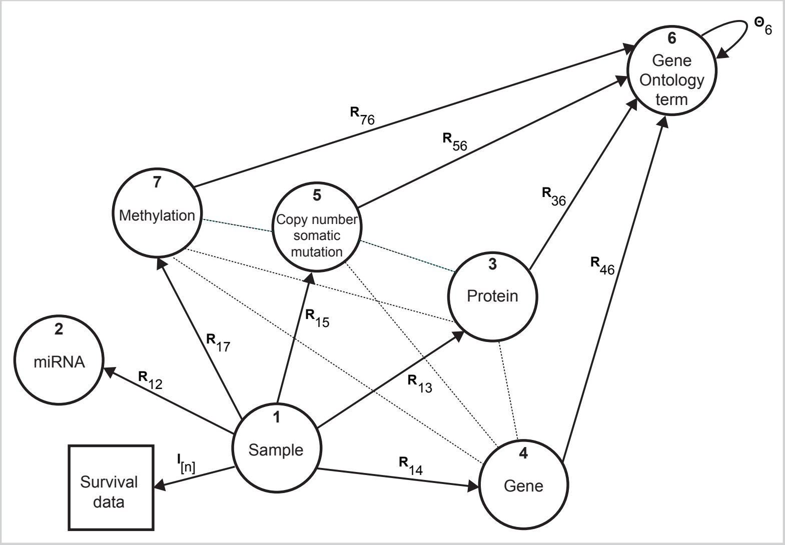


Figure 12.2

Data sources and their relations. Nodes in the graph correspond to different types of objects and edges denote data matrices R_{ij} or constraint matrices Θ_i. For example, matrix R_{13} contains protein expression values, R_{15} relates tissue samples to mutated genes in the tumor, and DNA methylation matrix R_{17} reports on gene-based methylation Beta values of interrogated sites. Gene annotations from Gene Ontology are given in matrices R_{x6}, $x \in \{3, 4, 5, 7\}$. Constraint matrix Θ_6 encodes the semantic similarity of Gene Ontology terms as defined by the directed acyclic graph of the ontology.

transformed absolute error loss of survival time defined by $l(y, \hat{\mathbf{y}}) = |\log(y) - \log(\hat{\mathbf{y}}_m)|$, where $\hat{\mathbf{y}}_m$ is the predicted median of survival time y. The median is the optimal predictor of the absolute error loss and is less affected by the long tails of survival distributions than the squared error loss. Log transformation addresses the concern that the absolute difference between predicted and actual survival time at a distant time point should result in smaller error than the same absolute difference achieved at a nearer time point (Lawless and Yuan, 2010).

12.7 Prediction of patient survival time

Table 12.1 reports the errors of predicting survival time for lung, kidney and head/neck cancer studies. We use protein expression and somatic mutation (p = sample, r = protein or r = copy number somatic mutation; see Sec. 12.4) data to regress against survival data. Our DFMF-SR approach (last row in the Table) outperforms an alternative approach that does sequential survival regression by first transforming data into the latent space and then inferring a survival model independently of data transforma-

Table 12.1

Cross-validated error of predicted survival time. Latent data representations of protein expression values or somatic mutation data are regressed against patient survival data for three different cancer studies. We compare our approach (DFMF-SR) to a procedure which first infers predictive factors by data fusion (DFMF in Step I) or principal component analysis (PCA in Step I) and then learns a regression model (Aalen in Step II). Aalen's regression modeling could be in principal applied to raw data (first row without feature construction in Step I), but fails due to high dimensionality of data sets.

Approach		Protein expression			Somatic mutation		
Step I	Step II	HNSC	KIRC	LUAD	HNSC	KIRC	LUAD
n. a.	Aalen	0.83	0.89	0.80	0.95	0.91	0.99
PCA	Aalen	0.73	0.70	0.69	0.71	0.73	0.72
DFMF	Aalen	0.67	0.65	0.66	0.61	0.68	0.61
DFMF-SR		0.56	0.62	0.59	0.54	0.58	0.53

tion (second and third row in the Table). Similar gains in accuracy of DFMF-SR are observed for other choices of $\boldsymbol{r}$ but are omitted here for brevity.

Models inferred by DFMF-SR are also substantially better than Aalen's regression from the raw data (first line in Table 12.1). The less well-studied cancer data sets in CAMDA 2014 are challenging to analyze due to noisy measurements, missing data and high right censorship (given the available data). For example, 30% of tumor samples from the HNSC study do not have information about donors' last known vital status or time intervals since their primary diagnoses. Of the remaining samples, 86% belong to censored individuals. We observed that model performance crucially depends on the ability to infer latent space and reduce data dimensionality, and survival regression analysis fails to detect predictive signals if applied to high-dimensional untransformed data sets in the original data domain.

The additive regression model benefits from incorporating time into estimation of regression coefficients and can give information about effects of data features on patient survival time by plotting components of cumulative regression coefficients $\boldsymbol{B}^*(t_k)$ against time. Fig. 12.3 shows cumulative regression functions for two somatic mutation latent factors and the baseline regression coefficient in the HNSC cancer study. The baseline coefficient starts off small in the first ten months after primary diagnosis and then increases (Fig. 12.3, right panel). Notice the different dynamics of regression coefficients for the two latent factors (Fig. 12.3, left panel). Gene sets belonging to

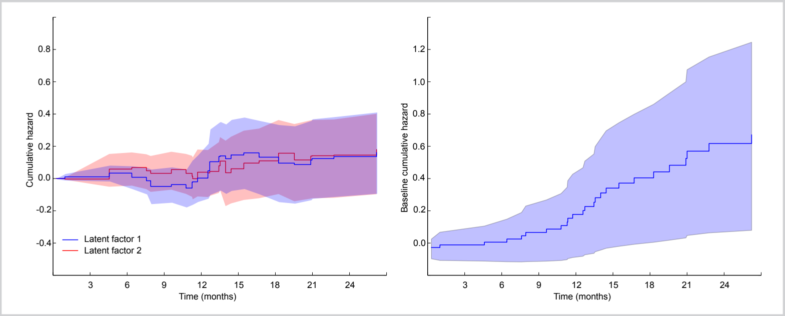


Figure 12.3

Cumulative hazard plots produced by DFMF-SR showing the cumulative hazards of selected somatic mutation latent factors (left; i.e. $B_i^*(t_k)$ of latent factor i for the times of the events t_k) and the baseline hazard (right) in the HNSC cancer study. Notice that regression coefficients are the derivatives of the cumulative hazards and so it is the slopes of the plots that are informative.

these latent factors are enriched in biological processes known to play a role in the development of cancer (Garraway and Lander, 2013), such as regulation of nitric-oxide synthase activity, monooxygenase and oxidoreductase activity, nitric oxide processes, and cyclase activity (FDR $< 4 \times 10^{-4}$). This finding points to a possible utility of the proposed approach for uncovering critical factors and their changing influence across different stages of cancer progression.

12.8 Conclusion

We here introduced data fusion for survival regression, a method for predicting patient survival time from a collection of heterogeneous data sets. Our approach builds upon recently proposed collective matrix factorization and a well-known Aalen's additive model for survival regression. Unlike existing methods for survival time prediction, we formulated a joint inference procedure that allows us to simultaneously infer model parameters of collective matrix factorization and regression coefficients of Aalen's model. We demonstrated improved performance of our method over several baselines in case studies involving three cancer types from the International Cancer Genome Consortium and diverse data sets, such as gene and miRNA expression profiles, somatic mutation data, methylation and gene annotations from the Gene Ontology. We showed that both latent data representation and joint inference, the two features of our approach, contribute substantially to accurate prediction of survival time. The work here alludes to the potential benefits of data fusion for inference of prediction models that are predictive of clinical outcomes.

Part VI

Data set selection for large-scale data fusion

Inter-relation sensitivity in collective matrix factorization

Most branches of science and technology are data rich, both in volume and heterogeneity of available data sets. We can view data sets as relation matrices, and represent the entire data domain as a relation graph. This representation has recently been explored in fusion by collective matrix factorization to jointly infer predictive models with very high accuracy.

We are interested in this chapter in how changes in one relation (data set) affect the latent representation of another relation in the context of a given collective matrix factorization model. For example, in a user-movie recommendation system, how would a change of casting affect users's preferences? We present Forensic, an approach for inter-relation sensitivity estimation in collective matrix factorization. Forensic derives from theory of Fréchet derivation and condition number estimation. It can estimate sensitivity for all pairs of relations within a single run of inference algorithm and can be applied to any collective matrix factorization.

We investigate the properties of Forensic in a study consisting of 13 data sets from molecular biology. Furthermore, we demonstrate its utility on a collection of 40 experimental protein physical interaction data sets, where Forensic is able to correctly identify surprising data sets and data sets containing experimental errors. *To our best knowledge, the latter study involves the largest number of data sets to date that were considered by any collective matrix factorization model.*

Results show that estimated sensitivity highly correlates with the changes of target relation reconstruction error when effect relation is removed. Forensic exhibits a surprisingly high level of agreement when applied to different factorization models and hence reports sensitivities that are properties of a relational data structure rather than a confound of a given factorization model. Experiments provide evidence that Forensic could be used as a scoring technique in data set selection for data fusion.

13.1 Background

Many applications of machine learning in social networks, e-commerce and molecular biology involve heterogeneous data that describe multiple relations between multiple types of objects (Pavlidis et al., 2001; Sutskever, 2009; Kim and Leskovec, 2013; Wang et al., 2014). For example, a biological domain with genes, phenotypes, cellular pathways and experiments might have four relations, as shown by a subgraph of relation graph in Fig. 13.1: a real matrix representing expression values of genes at different time points ($\boldsymbol{R}_{17}$), a matrix representing the phenotypes exhibited by the mutants ($\boldsymbol{R}_{12}$) and two binary matrices indicating the pathways each gene belongs to ($\boldsymbol{R}_{13}$) and whether genes interact physically or not ($\boldsymbol{\Theta}_1$) (Hofree et al., 2013). In such multi-relational domains, one could fit each relation matrix separately but this approach would not take advantage of any correlations between relations (Greene and Cunningham, 2009). If genes from a particular pathway are those whose mutants have similar phenotypes, one would like to exploit this correlation to improve the prediction (Stingo and Vannucci, 2011).

One class of inference algorithms that can treat multiple, in principle tens or hundreds of relations, are techniques of collective matrix factorization (Singh and Gordon, 2008c). Many collective factorization models were proposed recently (Long et al., 2006; Banerjee et al., 2007; Singh and Gordon, 2008a; Wang et al., 2008, 2011a; Nickel et al., 2011; Žitnik and Zupan, 2015a; London et al., 2013). Given the increased interest in considering a plethora of data sets during model inference such factorization models are expected to become even more abundant in the future (Žitnik and Zupan, 2014b).

Collective matrix factorization aims at improving predictive accuracy by exploiting information from one relation while predicting another (Singh and Gordon, 2008a). Typically, one of the modeled relations represents target, such as users' ratings matrix in a recommendation task or genes' functional annotation matrix in gene function prediction. Methods of collective matrix factorization have to address the issues of incomplete relations, missing patterns and possible disagreements between relations that arise due to integrative nature of the analysis (Greene and Cunningham, 2009). Therefore, understanding dependencies between relations seems essential both for exploratory analysis and for performing various predictive modeling tasks (Tang et al., 2013). One would like to consider in a factorized model only relations that positively affect completion of target relation and often for reasons of computational efficiency remove relations with insignificant influence on target relation. On the other side, a relation to which target relation is highly sensitive might be of analyst's interest by itself as the sensitivity can arise due to the unique characteristics of problem domain or noise within a relation (Greene and Cunningham, 2009; Xing and Dunson, 2011).

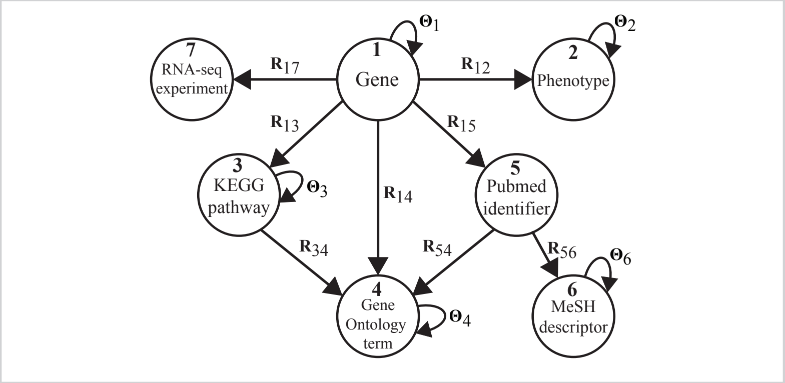


Figure 13.1

Relation graph of a biological domain consisting of seven object types (nodes) and thirteen relations (edges). For example, relation $\boldsymbol{R}_{12}$ is a matrix with genes in the rows and phenotypes in the columns, and element $R_{12}(i, j)$ indicates whether gene mutant i exhibits phenotype j, relation Θ_1 encodes protein physical interaction network, and R_{17} contains gene expression profiles.

In this chapter we aim to understand the sensitivity of one relation to perturbations in another relation when both relations are modeled by a collective matrix factorization. We approach this challenge by providing a formal definition of inter-relation sensitivity in collective matrix factorization and show how to estimate it efficiently. We propose Forensic, a new method of inter-relation sensitivity estimation for collective matrix factorization. To best of our knowledge, we are not aware of any existing approach that would provide such functionality. Our formulation derives from matrix algebra and Fréchet derivation (Higham, 2008) and provides consistent sensitivity estimates across many collective factorization models. An appealing feature of Forensic is its ability to estimate sensitivity for any pair of modeled relations for which it needs a one-time-only inference of a factorized model. As such, Forensic avoids computational burdens of controlling for latent factor initialization and additional parameter setting of factorization algorithm. Further innovation of our approach is that we can estimate sensitivity between relations coming from different data domains if they are related in a relation graph. In the example from Fig. 13.1 we can relate phenotypic annotation of genes ($\boldsymbol{R}_{12}$) to Medical Subject Heading (MeSH) description of research papers ($\boldsymbol{R}_{56}$) through relation that records gene occurrences in research literature ($\boldsymbol{R}_{15}$).

Moreover, the use of the Fréchet derivation in latent factor models opens many new applications that were previously not possible. Forensic can be used to detect low-quality experimental data sets. In biology, an often underappreciated issue is that even when an experimental readout is mapped in a sample, it is usually done with few, if any, replicates owing to cost, time or sample material availability (Ernst and Kellis, 2015). As a result, experimental variability can confound biological comparisons. This situation is exacerbated when analyzing large compendiums of data sets whose sheer number increases the likelihood that there will be outlier data sets of lower quality (Ernst and Kellis, 2015). This observation, together with the increasingly popular joint analysis of large data collections using collective latent factor models calls for an efficient and principled approach for estimation of sensitivities between data sets (relations). As such, Forensic can provide recommendations as to which data sets to integrate and can offer insights about "surprising," i.e. potentially problematic, data sets. While identification of problematic data sets is related to outlier mining in high-dimensional data (Angiulli and Pizzuti, 2005), the most important distinctions between that body of work and ours center on: (1) the estimation of sensitivity between *relations* rather than differences between *individual objects* coming from a single data set, and (2) the computational mechanisms that make Forensic readily applicable to any present collective latent factor model.

Here, we first provide the background in collective matrix factorization and the Fréchet derivation and then present our approach to inter-relation sensitivity estimation (Sec. 13.3). We investigate properties of the proposed approach in a domain with thirteen relations from molecular biology and several collective factorization models (Fig. 13.1). In a domain with forty protein interaction data sets we demonstrate the utility of Forensic for investigation of influences between relations and identification of relations to which a given target relation is most or least sensitive (Sec. 13.5).

13.2 Preliminaries

13.2.1 Collective matrix factorizations

Low-rank matrix factorization have been widely used for pattern recognition in the fields of data mining, signal processing, computer vision, bioinformatics, finance and economics, among others (Wang et al., 2013). Existing algorithmic variants include factorizations that impose the nonnegativity constraints on matrix factors or constraints such as sparsity, locality and orthogonality through regularization. Recent algorithms bySingh and Gordon (2008a); Sutskever (2009); Banerjee et al. (2007); Wang et al.

(2008, 2011a); Nickel et al. (2011); Žitnik and Zupan (2015a); London et al. (2013) modify standard factorization formulations to break through conventional data types or factorization modes.

Multi-relational factorization simultaneously factorizes many data matrices and shares latent factors between relations that have object types in common. Data Fusion by Matrix Factorization (DFMF) (Žitnik and Zupan, 2015a) takes a system of relation matrices and collectively factorizes them. Given a relation $\boldsymbol{R}_{ij} \in \mathbb{R}^{n_i \times n_j}$ between n_i objects of type i and n_j objects of type j, DFMF tri-factorizes it into a product of three low-dimensional matrices in the following way. DFMF find a rank-c_i,c_j factorization of $\boldsymbol{R}_{ij}$ as $\boldsymbol{R}_{ij} \approx \boldsymbol{G}_i \boldsymbol{S}_{ij} \boldsymbol{G}^T$, where an $c_i \times c_j$ matrix factor $\boldsymbol{S}_{ij}$ represents a relation-specific factor, and an $n_i \times c_i$ matrix factor $\boldsymbol{G}_i$ and an $n_j \times c_j$ matrix factor $\boldsymbol{G}_j$ are object type-specific matrix factors. The latter two matrix factors are shared among decompositions of relations that describe objects of type i and j, respectively. Related models of simultaneous matrix decomposition include Symmetric Penalized Matrix Tri-Factorization (tri-SPMF) (Wang et al., 2008) and Symmetric Nonnegative Matrix Factorization (S-NMTF) (Wang et al., 2011a). They differ from DFMF by incorporating graph regularization, requiring full set of relation matrices between all pairs of object types and the symmetry of relations. Another model, RESCAL (Nickel et al., 2011), employs a tensor factorization to take the structure of relational data into account. Given a collection of relation matrices of the same dimensions, RESCAL finds a rank-c factorization of k-th relation $\boldsymbol{R}_k \in \mathbb{R}^{n \times n}$ as $\boldsymbol{R}_k \approx \boldsymbol{A} \boldsymbol{S}_k \boldsymbol{A}^T$, where an $n \times c$ matrix factor $\boldsymbol{A}$ contains global latent components and $\boldsymbol{S}_k$ is a $c \times c$ asymmetric matrix that models participation of the latent components in the k-th relation. Typically, learning of these models is iterative by nature and alternates between updates of local latent factors until convergence criteria are satisfied. Although the formed factorization models are conceptually different, we demonstrate in the experiments that Forensic can be applied to them.

13.2.2 Condition numbers and Fréchet derivation

Ideally, a factorization algorithm returns not only an approximate solution but also an interpretable estimate for the error in that solution. Producing an a priori error bound for an algorithm can be very difficult (Higham, 2008), as it involves analysis of trunctation errors and rounding errors and, int the case of collective factorization algorithm, their propagation across data sets. A separate question, usually easier to answer, is how sensitive is the solution of the problems to perturbations in the data. Knowledge of problem sensitivity can be crucial in applications, where it gives insight into whether the problem of collective factorization has been well formulated, allows prediction of the effects of data inaccuracies, and indicates the best reconstruction error that any algorithm can be expected to provide (Higham, 2008).

Sensitivity is determined by the derivative of the function that maps the input data to the latent model. For a matrix function view of the latent factor models the appropriate derivative is the Fréchet derivative and its norm can determine the condition number for the problem as explained in the following sections. Thus every latent factor model gives rise to the related problem of characterizing, computing and estimating the associated Fréchet derivative and its norm.

Condition numbers

Sensitivity is measured by condition numbers (Higham, 2008). We start by recalling how condition numbers are defined for scalar functions $f(x)$. The standard definition of *relative condition number* is:

$$\mathrm{cond}_{\mathrm{rel}}(f,x) = \lim_{\epsilon\to 0} \sup_{|\Delta x|\leq\epsilon|x|} \left| \frac{f(x+\Delta x)-f(x)}{\epsilon f(x)} \right|. \tag{13.1}$$

This number measures by how much, at most, small changes in the data x can be magnified by the function f, when both changes are measured in a relative sense. If f is continuously differentiable, $f(x)\neq 0$ and $x\neq 0$ then it follows that the condition number of function f at point x is:

$$\mathrm{cond}_{\mathrm{rel}}(f,x) = \left| \frac{xf'(x)}{f(x)} \right|. \tag{13.2}$$

This definition of relative condition number extends readily to arbitrary matrix functions $\boldsymbol{F} : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$. Higham (2008) defined the relative condition number of matrix function $\boldsymbol{F}$ at value $\boldsymbol{X}$ by

$$\mathrm{cond}_{\mathrm{rel}}(\boldsymbol{F},\boldsymbol{X}) = \lim_{\epsilon\to 0} \sup_{\|E\|\leq\epsilon\|X\|} \frac{\|\boldsymbol{F}(\boldsymbol{X}+\boldsymbol{E})-\boldsymbol{F}(\boldsymbol{X})\|}{\epsilon\|\boldsymbol{F}(\boldsymbol{X})\|}, \tag{13.3}$$

where $\boldsymbol{E}$ is a perturbation matrix and the norm is any matrix norm. Some care is needed in interpreting Eq. (13.3) for matrix functions not defined throughout $\mathbb{C}^{n\times n}$. There exists a corresponding *absolute condition number*, which measures the change in the data and the function in an absolute sense (Higham, 2008).

The Fréchet derivative

To obtain an explicit expression analogous to Eq. (13.2) we need an appropriate notion of derivative for matrix function. The Fréchet derivative (Higham, 2008) of a matrix function $F : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ at point $\boldsymbol{X} \in \mathbb{C}^{n\times n}$ is a linear mapping:

$$\begin{tikzcd} \mathbb{C}^{n\times n} \arrow[r, "L_F"] & \mathbb{C}^{n\times n} \\ \boldsymbol{E} \arrow[r, mapsto] & L_F(\boldsymbol{X}, \boldsymbol{E}) \end{tikzcd} \qquad (13.4)$$

such that for all perturbation matrices $\boldsymbol{E} \in \mathbb{C}^{n\times n}$ the following holds:

$$F(\boldsymbol{X}+\boldsymbol{E}) - F(\boldsymbol{X}) - L_F(\boldsymbol{X}, \boldsymbol{E}) = o(\|\boldsymbol{E}\|) \qquad (13.5)$$

The value $L_F(\boldsymbol{X}, \boldsymbol{E})$ is referred to as the Fréchet derivative of F at $\boldsymbol{X}$ in direction $\boldsymbol{E}$. The notation $L_F(\boldsymbol{X}, \boldsymbol{E})$ can be read as "the Fréchet derivative of F at $\boldsymbol{X}$ in the direction $\boldsymbol{E}$" or "the Fréchet derivative of F at $\boldsymbol{X}$ applied to the matrix $\boldsymbol{E}$" (Higham, 2008).

The relative condition number can be expressed in terms of the norm of $L_F(\boldsymbol{X})$, which is defined by:

$$\|L_F(\boldsymbol{X})\| = \max_{Z\neq 0} \frac{\|L_F(\boldsymbol{X}, \boldsymbol{Z})\|}{\|\boldsymbol{Z}\|}. \qquad (13.6)$$

The relative condition number is then given by (cf. Theorem 3.1 in Higham (2008)):

$$\mathrm{cond}_{\mathrm{rel}}(F, \boldsymbol{X}) = \frac{\|L_F(\boldsymbol{X})\|\,\|\boldsymbol{X}\|}{\|F(\boldsymbol{X})\|}. \qquad (13.7)$$

Approximation to the Fréchet derivative of a matrix function

It is usually not straightforward to obtain an explicit formula or representation for the Fréchet derivative. In order to estimate $\mathrm{cond}_{\mathrm{abs}}(F, \boldsymbol{X})$ efficiently we have to be able to evaluate $\boldsymbol{L_F}(\boldsymbol{X}, \boldsymbol{E})$ for many directions $\boldsymbol{E}$. Al-Mohy and Higham (2010) proposed a complex step approximation to the Fréchet derivative of the matrix function $\boldsymbol{F}$, which is the idea that we employ in the estimation inter-relation sensitivity defined in Sec. 13.3.2. They approximate the Fréchet derivative by evaluating a real-valued matrix function $\boldsymbol{F}$ at a complex argument as:

$$\boldsymbol{L_F}(\boldsymbol{X}, \boldsymbol{E}) = \mathrm{Im}\ F(\boldsymbol{X} + ih\boldsymbol{E})/h + O(h^2), \tag{13.8}$$

where $i = \sqrt{-1}$ is unit imaginary number. The complex step approximation is known in the scalar case, where it can be derived from the Taylor series expansion. The use of complex arithmetic is appealing because of two reasons. First, unlike in the finite difference approximation to the Fréchet derivative, subtractive cancellation is not intrinsic in the expression Im $F(\boldsymbol{X} + ih\boldsymbol{E})/h$. This means that the complex step approximation offers the promise of selecting $\boldsymbol{h}$ based solely on the need to make the truncation error sufficiently small. Practical experience with the scalar-based complex step approximation, e.g., Cox and Harris (2004), has indeed demonstrated the ability of Eq. (13.8) to produce accurate approximations even in scenarios with $\boldsymbol{h}$ as small as 10^{-100}. Second, the complex-valued approximation produces an estimate of the Fréchet derivative with an order of convergence more than the real-valued approximation. The last attractive property holds due to the cancellation of imaginary unit in the even-powered terms of the Taylor series expansion.

Although complex step approximation is known in the scalar case, it is new in terms of matrix functions and can produce estimates that are more reliable than those obtained by techniques that use finite differences (Al-Mohy and Higham, 2010).

13.3 Inter-relation sensitivity estimation in collective matrix factorization

Suppose we have a collective matrix factorization model $\boldsymbol{F_\theta}$ and we denote the setting of its latent parametrization collectively by θ. We view $\boldsymbol{F_\theta}$ as a matrix function (Higham, 2008) that takes as it input a collection of relations $\mathscr{C}$, infers latent representation θ and specifies $\boldsymbol{F_\theta}(\boldsymbol{R})$ to be a relation of the same dimensions as $\boldsymbol{R} \in \mathscr{C}$; it does so in a way that provides a useful decomposition of $\boldsymbol{R}$ into latent components, most often found by minimizing a reconstruction loss with additional constraints. Suppose that $\mathscr{C}$ contains two designated relations, $\boldsymbol{R}^{(t)} \in \mathbb{R}^{n_{t_1} \times n_{t_2}}$ and $\boldsymbol{R}^{(e)} \in \mathbb{R}^{n_{e_1} \times n_{e_2}}$, which we refer to as target and effect relations, respectively. Selection of target and effect relations depends on a predictive modeling task. It is unrelated to concepts of supervised learning since matrix factorization typically implements unsupervised or semi-supervised model.

Forensic can estimate inter-relation sensitivity when $\boldsymbol{R}^{(t)}$ and $\boldsymbol{R}^{(e)}$ share either both dimensions ($n_{t_1} = n_{e_1} \wedge n_{t_2} = n_{e_2}$), one dimension ($n_{t_1} = n_{e_1} \veebar n_{t_2} = n_{e_2}$) or neither of them ($n_{t_1} \neq n_{e_1} \wedge n_{t_2} \neq n_{e_2}$). This characteristic permits Forensic the analysis of any two relations included in $\mathscr{C}$. Given F_θ, relations $\boldsymbol{R}^{(t)}$ and $\boldsymbol{R}^{(e)}$, our goal is to quantify the effects that relation $\boldsymbol{R}^{(e)}$ has on target relation $\boldsymbol{R}^{(t)}$ in the context of F_θ. We aim to do so efficiently without necessitating rerun of factorization inference algorithm.

13.3.1 *Definition of* Forensic *inter-relation sensitivity score*

We begin with definition of Forensic and appropriate condition numbers. We define Forensic's ϕ score of inter-relation sensitivity as an estimate that quantifies the effects of changes of effect relation $\boldsymbol{R}^{(e)}$ on target relation $\boldsymbol{R}^{(t)}$ in the context of a collective matrix factorization F_θ:

$$\phi(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)}; F_\theta) = \frac{\|L_{F_\theta}(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)})\| \, \|\boldsymbol{R}^{(t)}\|}{\|F_\theta(\boldsymbol{R}^{(t)})\|}, \qquad (13.9)$$

where L_{F_θ} is Fréchet derivative of F_θ, $F_\theta(\boldsymbol{R}^{(t)})$ is the estimate of target relation provided by the parameterized latent factor model F_θ, and the norm can be any matrix norm (we focus on matrix 1-norm in the next sections).

The Fréchet derivative $L_{F_\theta}(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)})$ in Eq. (13.9) evaluates sensitivity of target relation $\boldsymbol{R}^{(t)}$ to small perturbations in $\boldsymbol{R}^{(e)}$, where perturbations are determined by model F_θ. It is the essential part of the above formula and represents the mapping instead of its value in any particular direction. We refer to it as Fréchet derivative because of its conceptual analogy with Fréchet derivatives (Higham, 2008). We define Fréchet derivative of F_θ at $\boldsymbol{R}^{(t)}$ with respect to perturbation $\boldsymbol{E}$ applied to $\boldsymbol{R}^{(e)}$ as a linear mapping $L_{F_\theta}(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)}; \boldsymbol{E}) \in \mathbb{R}^{n_{t_1} \times n_{t_2}}$ such that:

$$F_{\bar\theta}(\boldsymbol{R}^{(t)} | \bar\theta = \theta_{e^-} \cup \{\boldsymbol{R}^{(e)} \boxplus \boldsymbol{E}\}) - F(\boldsymbol{R}^{(t)}) - L_{F_\theta}(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)}; \boldsymbol{E}) = o(\|\boldsymbol{E}\|) \quad (13.10)$$

for any perturbation matrix $\boldsymbol{E}$. Detailed definition and estimation of $F_{\bar{\theta}}$ and $\boxplus$ operator will become clear in the next section. Intuitively, $F_{\bar{\theta}}$ evaluates target relation when small perturbations of respect relation are performed and the $\boxplus$ operator transforms effect relation and perturbation to the same data domain as specified by a factorization scheme. To estimate Fréchet derivative that does not depend on perturbation direction we have to estimate matrix norm of the Fréchet derivative:

$$\|L_{F_\theta}(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)})\| = \max_{\|E\|=1} \|L_{F_\theta}(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)}; \boldsymbol{E})\| \tag{13.11}$$

Forensic measures by how much, at most, small changes in the data can be magnified by F_θ, when both changes are measured in a relative sense. Recall that sensitivity is measured by condition numbers and Forensic generalizes the relative condition number of a scalar function f at point x defined in Eq. (13.2). More explicitly, $\phi(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)}; F_\theta)$ represents relative condition number of factorization F_θ at $\boldsymbol{R}^{(t)}$ for changes made in $\boldsymbol{R}^{(e)}$:

$$\phi(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)}; F_\theta) = \lim_{\epsilon \to 0} \sup_{\|E\| < \epsilon \|R^{(t)}\|} \frac{\|F_{\bar{\theta}}(\boldsymbol{R}^{(t)} | \bar{\theta} = \theta_{e^-} \cup \{\boldsymbol{R}^{(e)} \boxplus \boldsymbol{E}\}) - F_\theta(\boldsymbol{R}^{(t)})\|}{\epsilon \|F_\theta(\boldsymbol{R}^{(t)})\|}. \tag{13.12}$$

This alternative but equivalent view of Eq. (13.9) implies that:

$$\frac{\|F_{\bar{\theta}}(\boldsymbol{R}^{(t)} | \bar{\theta} = \theta_{e^-} \cup \{\boldsymbol{R}^{(e)} \boxplus \boldsymbol{E}\}) - F_\theta(\boldsymbol{R}^{(t)})\|}{\|F_\theta(\boldsymbol{R}^{(t)})\|} \leq \phi(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)}; F_\theta) \frac{\|\boldsymbol{E}\|}{\|\boldsymbol{R}^{(t)}\|} + o(\|\boldsymbol{E}\|)$$

and provides an approximate perturbation bound for small perturbations $\boldsymbol{E}$. We next outline the procedure for estimation of Forensic's ϕ score.

13.3.2 Estimation of Forensic *score*

In this section we focus on estimating the essential part of Forensic's ϕ score, the matrix norm of Fréchet derivative $\|L_{F_\theta}(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)})\|$. Estimation of matrix norm requires evaluating Fréchet derivative for certain $\boldsymbol{E}$, a task on which we focus first. Matrix norms of target relation and its reconstruction by F_θ, which are also needed in Forensic's ϕ score, can be estimated with standard matrix norm estimators (cf. matrix 1-norm estimation in Higham and Tisseur (2000)).

Estimation of $L_{F_\theta}(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)}; \boldsymbol{E})$

Forensic estimates the Fréchet derivative L_{F_θ} via complex step approximation:

$$L_{F_\theta}(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)}; \boldsymbol{E}) \approx \operatorname{Im} \frac{F_{\bar{\theta}}(\boldsymbol{R}^{(t)} | \bar{\theta} = \theta_{e^-} \cup \{\boldsymbol{R}^{(e)} \boxplus ih\boldsymbol{E}\})}{h}, \tag{13.13}$$

where $L_{F_\theta}(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)}; \boldsymbol{E})$ is evaluated at complex argument $ih\boldsymbol{E}$. This expression approximates Fréchet derivative of F_θ at $\boldsymbol{R}^{(t)}$ with respect to change of $\boldsymbol{R}^{(e)}$ in the direction $ih\boldsymbol{E}$ for suitably small h. In the scalar case, complex step approximation is derived from the Taylor series expansion. Forensic generalizes it to matrix factorizations over real numbers. Replacing $\boldsymbol{E}$ by $ih\boldsymbol{E}$ in Eq. (13.10), where $\boldsymbol{E}$ is independent of h, and using the linearity of L_{F_θ}, we obtain:

$$F_{\bar{\theta}}(\boldsymbol{R}^{(t)} | \bar{\theta} = \theta_{e^-} \cup \{\boldsymbol{R}^{(e)} \boxplus ih\boldsymbol{E}\}) - F(\boldsymbol{R}^{(t)}) - ihL_{F_\theta}(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)}) = o(h). \tag{13.14}$$

Thus, if F operates over non-complex relations and $\boldsymbol{R}^{(t)}$, $\boldsymbol{R}^{(e)}$ and $\boldsymbol{E}$ are real matrices then:

$$L_{F_\theta}(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)}; \boldsymbol{E}) = \lim_{h \to 0} \operatorname{Im} \frac{F_{\bar{\theta}}(\boldsymbol{R}^{(t)} | \bar{\theta} = \theta_{e^-} \cup \{\boldsymbol{R}^{(e)} \boxplus ih\boldsymbol{E}\})}{h}, \tag{13.15}$$

which justifies complex step approximation. To be able to determine the rate of convergence of the approximation as $h \to 0$, we need stronger assumptions about F. In particular, if the operation of F can be described by an analytic matrix function then we rely on theory of matrix functions: the analyticity of that function is sufficient but not necessary condition to ensure a second order approximation of Fréchet derivative (cf. Lai and Crassidis (2008) and Theorem 1 in Al-Mohy and Higham (2010)). In Sec. 13.4 we show the efficacy of L_{F_θ} in predictive modeling.

We select h in Eq. (13.13) such that $h \leq \sqrt{u} \|\boldsymbol{R}^{(t)}\|_{\text{Fro}} / \|\boldsymbol{E}\|_{\text{Fro}}$ and u is the unit round-off. Complex step approximation is attractive because it can be implemented as long as the update rules of F can be evaluated at a complex argument, which is true for existing collective matrix factorization algorithms. Perturbation matrix $\boldsymbol{E}$ has same dimensions as target relation, $\boldsymbol{E} \in \mathbb{R}^{n_{t_1} \times n_{t_2}}$. Hence, the operator $\boxplus$ is concerned with transforming perturbation matrix to latent space of effect relation, that is, $\boxplus$ specifies a mapping from $\mathbb{R}^{n_{t_1}\times n_{t_2}}$ to $\mathbb{R}^{c_{e_1}\times c_{e_2}}$. This procedure depends on the algorithm of collective matrix factorization but typically involves a few multiplications of perturbation matrix with inferred latent matrices that are part of $\boldsymbol{F}_\theta$.

Latent parametrization $\bar{\theta}$ in Eqs. (13.13–13.15) is obtained from θ by replacing latent parameters specific to effect relation with their perturbed version obtained by the $\boxplus$ operation. We next exemplify evaluation of $\boldsymbol{F}_{\bar{\theta}}(\boldsymbol{R}^{(t)}|\bar{\theta} = \theta_{e^-} \cup \{\boldsymbol{R}^{(e)} \boxplus ih\boldsymbol{E}\})$ in Eq. (13.13) for two different collective factorized models $\boldsymbol{F}$. In general, $\boldsymbol{F}_{\bar{\theta}}$ applies the update rule specified by $\boldsymbol{F}$ to latent factors that are shared between target relation and other relations modeled by $\boldsymbol{F}_\theta$.

Example 1 – Evaluation of $\boldsymbol{F}_{\bar{\theta}}(\boldsymbol{R}^{(t)}|\bar{\theta} = \theta_{e^-} \cup \{\boldsymbol{R}^{(e)} \boxplus ih\boldsymbol{E}\})$ in Nickel et al. (2011) model. Factorized model RESCAL (Nickel et al., 2011) factorizes k-th relation as $\boldsymbol{R}_k \approx \boldsymbol{A}\boldsymbol{S}_k\boldsymbol{A}^T$, where $\theta = \{\boldsymbol{A}\} \cup \{\boldsymbol{S}_k\}$. FORENSIC evaluates operator $\boxplus$ as $\boldsymbol{R}^{(e)} \boxplus ih\boldsymbol{E} = \boldsymbol{S}_e + ih\boldsymbol{A}^T\boldsymbol{E}\boldsymbol{A}$. Latent parametrization $\bar{\theta}$ is then formed from θ by replacing $\boldsymbol{S}_e$ with $\boldsymbol{S}_e + ih\boldsymbol{A}^T\boldsymbol{E}\boldsymbol{A}$. As such, data perturbation affects only $\boldsymbol{S}_e$, which is the latent factor in RESCAL specific to effect relation $\boldsymbol{R}^{(e)}$. Given $\bar{\theta}$, $\boldsymbol{F}_{\bar{\theta}}$ performs an update of latent matrix $\boldsymbol{A}$ as defined by RESCAL factorization scheme and returns reconstructed target relation $\boldsymbol{R}^{(t)}$ as $\bar{\boldsymbol{A}}\boldsymbol{S}_t\bar{\boldsymbol{A}}$.

Example 2 – Evaluation of $\boldsymbol{F}_{\bar{\theta}}(\boldsymbol{R}^{(t)}|\bar{\theta} = \theta_{e^-} \cup \{\boldsymbol{R}^{(e)} \boxplus ih\boldsymbol{E}\})$ in Žitnik and Zupan (2015a) model. DFMF (Žitnik and Zupan, 2015a) is a factorized model that models multiple relations between different types of objects, whereas RESCAL models multiple relations between two types of objects. DFMF factorizes relation between object types j and l ($\boldsymbol{R}_{jl} \in \boldsymbol{R}^{n_j\times n_l}$) into a product of three low-rank matrices $\boldsymbol{R}_{jl} \approx \boldsymbol{G}_j\boldsymbol{S}_{jl}\boldsymbol{G}_l^T$, where $\boldsymbol{G}_j \in \mathbb{R}^{n_j\times c_j}$, $\boldsymbol{G}_l \in \mathbb{R}^{n_l\times c_l}$ and $\boldsymbol{S}_{jl} \in \mathbb{R}^{c_j\times c_l}$ are latent matrices and c_j, c_l represent model dimensionality. Latent parametrization of DFMF is given by $\theta = \{\boldsymbol{G}_j\} \cup \{\boldsymbol{S}_{jl}\}$. For target relation $\boldsymbol{R}^{(t_1t_2)}$ and effect relation $\boldsymbol{R}^{(e_1e_2)}$, both included in DFMF model $\boldsymbol{F}_\theta$, we evaluate $\boldsymbol{R}^{(e)} \boxplus ih\boldsymbol{E}$ as follows:

$$\boldsymbol{R}^{(e)} \boxplus ih\boldsymbol{E} = \begin{cases} \boldsymbol{S}_{e_1e_2} + ih\boldsymbol{G}_{t_1}^T\boldsymbol{E}\boldsymbol{G}_{t_2}\boldsymbol{S}_{t_1t_2}^T\boldsymbol{S}_{e_1e_2} & \text{if } t_1 = e_1 \\ \boldsymbol{S}_{e_1e_2} + ih\boldsymbol{S}_{e_1e_2}\boldsymbol{S}_{t_1t_2}^T\boldsymbol{G}_{t_1}^T\boldsymbol{E}\boldsymbol{G}_{t_2} & \text{if } t_2 = e_2 \\ \boldsymbol{S}_{e_1e_2} + ih\boldsymbol{S}_{e_1e_2}\boldsymbol{G}_{e_1}^T\boldsymbol{E}\boldsymbol{G}_{t_2}\boldsymbol{S}_{t_1t_2}^T & \text{if } t_1 = e_2 \\ \boldsymbol{S}_{e_1e_2} + ih\boldsymbol{S}_{t_1t_2}^T\boldsymbol{G}_{t_1}^T\boldsymbol{E}\boldsymbol{G}_{t_2}\boldsymbol{S}_{e_1e_2} & \text{if } t_2 = e_1 \end{cases} \qquad (13.16)$$

Similarly as in RESCAL, $\boldsymbol{F}_{\bar{\theta}}$ performs an update of latent factors that are shared between target relation and other relations in the model. In particular, it applies an update rule of DFMF to latent matrix factors $\boldsymbol{G}_{t_1}$ and $\boldsymbol{G}_{t_2}$. The return value of $\boldsymbol{F}_{\bar{\theta}}$ is reconstruction of target relation $\boldsymbol{R}^{(t)}$ computed as $\bar{\boldsymbol{G}}_{t_1} \boldsymbol{S}_{t_1 t_2} \bar{\boldsymbol{G}}_{t_2}^T$.

Estimation of matrix 1-norm of $L_{F_\theta}(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)})$

Next, we explain how to estimate the matrix 1-norm of the Fréchet derivative. Since the Fréchet derivative $L_{F_\theta}(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)})$ is a linear operator:

$$\mathrm{vec}(L_{F_\theta}(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)}; \boldsymbol{E})) = K_{F_\theta}(\boldsymbol{R}^{(t)}; \boldsymbol{R}^{(e)})\mathrm{vec}(\boldsymbol{E}) \tag{13.17}$$

for some $K_{F_\theta}(\boldsymbol{R}^{(t)}; \boldsymbol{R}^{(e)}) \in \boldsymbol{R}^{n_{t_1}^2 \times n_{t_2}^2}$ that is independent of $\boldsymbol{E}$. We refer to matrix $K_{F_\theta}(\boldsymbol{R}^{(t)}; \boldsymbol{R}^{(e)})$ as the Kronecker form of the Fréchet derivative. This form is attractive because it explicitly captures the linearity of the Fréchet derivative and permits Forensic to exploit standard linear algebra techniques to estimate $\|L_{F_\theta}(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)})\|$.

For large n, explicitly forming $K_{F_\theta}(\boldsymbol{R}^{(t)}; \boldsymbol{R}^{(e)})$ is prohibitively expensive and so the Forensic's ϕ score must be estimated rather than computed exactly. In practice, what is needed is an estimate that is of the correct order of magnitude — more than one correct significant digit is not needed.

In Algorithm 5 we give a matrix 1-norm estimator for $L_{F_\theta}(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)})$. The idea of applying a matrix norm estimator to the Kronecker form of the Fréchet derivative has been used before for estimating the condition numbers of matrix exponential and matrix logarithm (Al-Mohy and Higham, 2009; Al-Mohy et al., 2013), where it was shown to produce more reliable matrix norm estimates than approaches based on finite differences. Algorithm 5 is Lanczos-based and requires two evaluations of the Fréchet derivative per iteration (steps 7 and 12 in Algorithm 5). Forensic computes them via a complex step approximation given by Eq. (13.13).

Algorithm 5 does not require a starting matrix. In contrast to the power method, which is an alternative matrix norm estimation technique, Algorithm 5 has a "built-in" starting matrix (step 2 in Algorithm 5). Another advantage of Algorithm 5 is that it has a natural formulation of the convergence test. It also has a more predictable

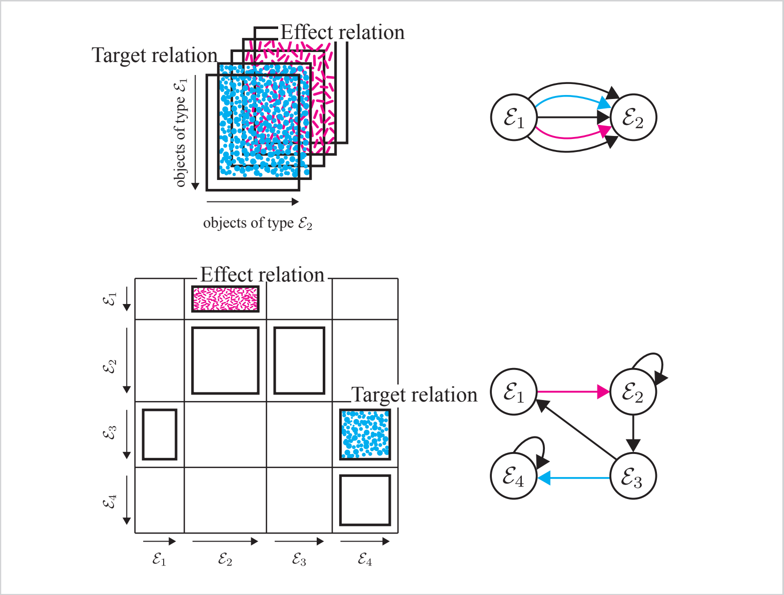


Figure 13.2

Collective matrix factorization models can be divided into two groups: (top) models for two-object type and multi-relation data and (bottom) models for multi-object type and multi-relation data. Shown are representative data settings for the two groups. In a two-object type setting, we are given a collection of relation matrices between two types of objects ($\mathscr{E}_1$ and $\mathscr{E}_2$), whereas in a multi-object type setting we are presented with a collection of relation matrices that potentially describe relationships between different object types ($\mathscr{E}_1$ to $\mathscr{E}_4$ in the bottom pane). Next to the relation matrices are shown relation graphs with nodes corresponding to different objects types and edges denoting distinct relations. Target and relation matrices might not share an object type.

number of iterations. In all our experiments it needed less than ten Fréchet derivative evaluations for convergence.

Estimation of $L_{F_\theta}(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)})$ when $t \cap e = \varnothing$

Some care is needed in interpreting Eq. (13.16) for multi-object type factorizations such as previously mentioned DFMF (Fig. 13.2, bottom), where it can occur that target and effect relations do not describe a common object type. For example, target relation matrix in the bottom pane of Fig. 13.2 contains relationships between objects of type $\mathscr{E}_3$ and type $\mathscr{E}_4$, whereas effect relation matrix relates objects of type $\mathscr{E}_1$ to objects of type $\mathscr{E}_2$. This means that the particular choice of target and effect relations in Fig. 13.2 is represented in a relation graph by non-neighboring edges.

In scenarios, where target and relation matrices do not share an object type, Forensic defines the Fréchet derivative through a relation chain that starts at the effect relation

Algorithm 5

LAPACK Matrix 1-norm estimator on the Fréchet derivative in factorization F_θ; $\|L_{F_\theta}(R^{(t)}, R^{(e)})\|_1$. Given the latent factor model F_θ, target relation $R^{(t)}$ and effect relation $R^{(e)}$, this algorithm uses the LAPACK norm estimator to produce an estimate of $\|L_{F_\theta}(R^{(t)}, R^{(e)})\|_1$, given the ability to compute $L_{F_\theta}(R^{(t)}, R^{(e)}; E)$ for any E.

Input:

- matrix factorization F_θ with latent parameters θ,
- target relation $\boldsymbol{R}^{(t)}$,
- effect relation $\boldsymbol{R}^{(e)}$.

Output:

- an estimate $\gamma = \|L_{F_\theta}(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)})\|_1$.

```
1:  $\boldsymbol{v} = \text{vec}(L_{F_\theta}(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)}; (n_{t_1} n_{t_2})^{-1}[\mathbf{1}]_{n_{t_1} \times n_{t_2}}))$
2:  $\gamma = \|\boldsymbol{v}\|_1$
3:  $\xi = \text{sign}(\boldsymbol{v})$
4:  $\boldsymbol{x} = \text{vec}(L_{F_\theta}(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)}; [\xi^T]_{n_{t_1} \times n_{t_2}}))$
5:  *repeat*
6:      $j = \min\{i : |\boldsymbol{x}_i| = \|\boldsymbol{x}\|_\infty\}$
7:      $\boldsymbol{v} = \text{vec}(L_{F_\theta}(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)}; [\boldsymbol{e}_j]_{n_{t_1} \times n_{t_2}}))$, where $\boldsymbol{e}_j \in \{0,1\}^{n_{t_1} n_{t_2}}$ is a standard unit vector
8:      $\bar{\gamma} = \gamma$
9:      $\gamma = \|\boldsymbol{v}\|_1$
10:     *if* $\text{sign}(\boldsymbol{v}) = \pm\xi$ or $\gamma \leq \bar{\gamma}$ *then goto* line 14
11:     $\xi = \text{sign}(\boldsymbol{v})$
12:     $\boldsymbol{x} = \text{vec}(L_{F_\theta}(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)}; [\xi^T]_{n_{t_1} \times n_{t_1}}))$
13: *until* $\|\boldsymbol{x}\|_\infty = \boldsymbol{x}_j$
14: $\boldsymbol{x}_i = (-1)^{i+1}(1 + \frac{i-1}{n_{t_1} n_{t_2} - 1}),\ i = 1 : (n_{t_1} n_{t_2})$
15: $\boldsymbol{x} = \text{vec}(L_{F_\theta}(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)}; [\boldsymbol{x}]_{n_{t_1} \times n_{t_2}}))$
16: *if* $2\|\boldsymbol{x}\|_1/(3n_{t_1} n_{t_2}) > \gamma$ *then* $\gamma = 2\|\boldsymbol{x}\|_1/(3n_{t_1} n_{t_2})$
```

Here, sign is element-wise sign function, vec is an operator that stacks columns of a matrix into one long vector, and $[x]_{n_1 \times n_2}$ denotes dematricization operation, which reshapes the vectorized version of $\boldsymbol{x}$ back to its full matrix form.

and end at the target relation:

$$\|L_{F_\theta}(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)})\| = \sum_{\mathscr{C}} \min_{1\leq i<|\mathscr{C}|} \|L_{F_\theta}(\boldsymbol{R}^{(\mathscr{C}_{i+1})}, \boldsymbol{R}^{(\mathscr{C}_i)})\|. \tag{13.18}$$

Here, $\mathscr{C}$ is a relation chain between the target and the effect relations and is determined by the connectivity of a particular relation graph. Formally, a relation chain is given by a sequence of relations:

$$\mathscr{C} = [\boldsymbol{R}^{(1)}, \boldsymbol{R}^{(2)}, \ldots, \boldsymbol{R}^{(k)}], \tag{13.19}$$

where $\boldsymbol{R}^{(1)} = \boldsymbol{R}^{(e)}$ and $\boldsymbol{R}^{(k)} = \boldsymbol{R}^{(t)}$. The ordering of relations in $\mathscr{C}$ is such that $\boldsymbol{R}^{(\mathscr{C}_{i+1})}$ and $\boldsymbol{R}^{(\mathscr{C}_i)}$ share an object type. Notice that in a well formulated collective latent factor model there always exists a relation chain between any target and effect relations.

An example of a relation chain. A data domain with users, movies, genres, actors, directors and user demographic profiles might have five relations shown in Fig. 13.3 representing: users' ratings of movies, users' demographic profiles, the genres each movie belongs to, the movies each actor appeared in, and the movies delivered by each director. To estimate sensitivity of $\boldsymbol{R}_{\text{User–Demographics}}$ to perturbations of $\boldsymbol{R}_{\text{Genre–Movie}}$, FORENSIC considers a relation chain $\mathscr{C} = [\boldsymbol{R}_{\text{Genre–Movie}}, \boldsymbol{R}_{\text{Movie–User}}, \boldsymbol{R}_{\text{User–Demographics}}]$. Intuitively, Eq. (13.18) accounts for possible ways in which perturbations of $\boldsymbol{R}_{\text{Genre–Movie}}$ can propagate through relation graph to reach $\boldsymbol{R}_{\text{User–Demographics}}$.

13.3.3 Normalization of FORENSIC score

So far, we defined FORENSIC's ϕ score as a value that measures how perturbation of one relation affects another relation if both relations are collectively modeled by a latent factor model. We defined appropriate condition numbers and described computational steps needed to efficiently estimate ϕ values. Next, we further normalize ϕ values to ensure that we can compare the values when different effect and target relation pairs are considered.

We define the normalized ϕ_n score as:

$$\phi_n(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)}; F_\theta) = \frac{\phi(\boldsymbol{R}^{(t)}, \boldsymbol{R}^{(e)}; F_\theta)}{\|F_\theta(\boldsymbol{R}^{(t)})\|\|F_\theta(\boldsymbol{R}^{(e)})\|}, \tag{13.20}$$

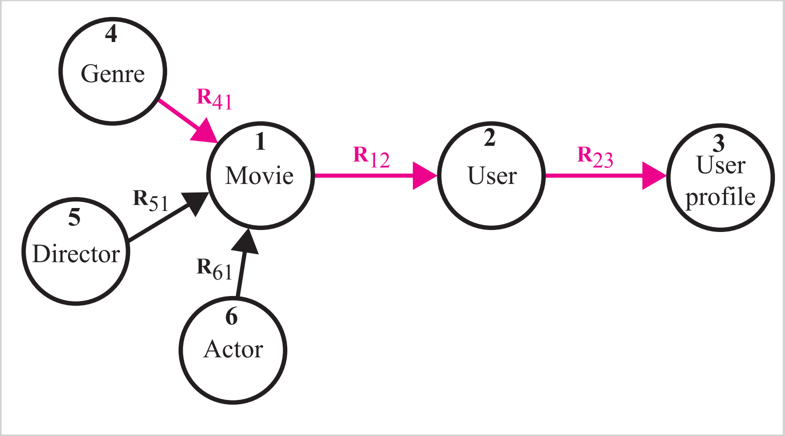


Figure 13.3

An example of a data domain with six object types (nodes) that are related by five relations (edges). Multi-object type data domains can be analyzed with collective matrix factorization such as DFMF (Žitnik and Zupan, 2015a) and S-NMTF (Wang et al., 2011a). To estimate the sensitivity of user's demographics profile (R_{23}) to perturbations of the genres each movie belongs to (R_{41}), FORENSIC considers sensitivity scores along a highlighted chain of relations ($R_{41} \rightarrow R_{12} \rightarrow R_{23}$).

where $\|F_\theta(\boldsymbol{R}^{(t)})\|$ and $\|F_\theta(\boldsymbol{R}^{(e)})\|$ are matrix norms of the estimated target and effect relations, respectively. The estimates are obtained by reconstructing target and effect relations from the latent factors provided by $\boldsymbol{F}$.

The reasoning behind the normalization terms introduced into the ϕ score is as follows. We would like to have a score that would use an equal *relative amount of perturbation* for various effect relations. When potential effect relations are of very different size then the amount of perturbation, $\boldsymbol{E}$, that is induced into the latent factor model also varies. This situation is exacerbated when analyzing large collections of data sets (see Sec. 13.5) whose sheer difference in the number of data points per data set increases the likelihood that a given target relation will be more sensitive to a larger data set. We thus divide the ϕ score by the norm of the estimated effect relation.

The second normalization term in Eq. (13.20) has a related role. A given effect relation should have an approximately equal *relative opportunity* to affect various target relations, which in practice might contain substantially different number of data points. We hence further divide the ϕ score by the norm of the estimated target relation to obtain the final normalized ϕ score. We note that we employ matrix 1-norm in all our experiments.

13.4 A case study with 13 data sets from molecular biology

In this section, we show that Forensic has excellent inter-relation sensitivity estimation power by applying it to four recent collective matrix factorizations, DFMF (Žitnik and Zupan, 2015a), tri-SPMF (Wang et al., 2008), S-NMTF (Wang et al., 2011a) and RESCAL (Nickel et al., 2011) (overviewed in Sec. 13.2.1), and evaluating its performance in two case studies. First, we consider thirteen relations from molecular biology that describe relationships between seven object types including genes, phenotypes and cellular pathways. In what follows, we describe the data sets, report performance of Forensic and its utility for data analysis.

13.4.1 Experimental setup

Accurate identification of genes whose mutations are implicated in a certain phenotype is a major challenge in biology (Radivojac et al., 2013; Sifrim et al., 2013). We had gene expression profiles of *D. discoideum* development cycle by RNA-seq (Parikh et al., 2010) ($\boldsymbol{R}_{17}$), gene annotations from Gene Ontology (Ashburner et al., 2000) ($\boldsymbol{R}_{14}$) and Phenotype Ontology downloaded from dictyBase (http://dictybase.org) ($\boldsymbol{R}_{12}$), mentions of genes in research articles from PubMed database ($\boldsymbol{R}_{15}$, $\boldsymbol{R}_{54}$), categorization of research articles by Medical Subject Headings (MeSH) ($\boldsymbol{R}_{56}$), and cellular pathway information ($\boldsymbol{R}_{13}$, $\boldsymbol{R}_{34}$). Additionally, we include information about the structure of gene and phenotype ontologies ($\boldsymbol{\Theta}_4$, $\boldsymbol{\Theta}_2$), MeSH hierarchy ($\boldsymbol{\Theta}_6$), pathway organization ($\boldsymbol{\Theta}_3$) and protein physical interaction network ($\boldsymbol{\Theta}_1$).

Fig. 13.1 shows the relation graph considered in this study. We applied DFMF model to thirteen relations for the prediction of gene-phenotype associations in *D. discoideum*, where target gene-phenotype relation matrix is denoted by $\boldsymbol{R}_{12}$. Two other considered factorization algorithms, tri-SPMF and S-NMTF, require as their input a collection of symmetric relation matrices between every possible pair of object types. We thus set $\boldsymbol{R}_{ij} = \boldsymbol{R}_{ji}^T$ for the choice of i and j that correspond to the relations shown in Fig. 13.1 and set $\boldsymbol{R}_{ij} = \boldsymbol{0}$ otherwise.

For the RESCAL model, which can consider data describing a one object type, we preprocessed the relations from Fig. 13.1 to obtain several data matrices that described relationships between genes. We constructed eight relation matrices:

$$\begin{aligned}
\boldsymbol{R}_1 &= \boldsymbol{R}_{17}\boldsymbol{R}_{17}^T\\
\boldsymbol{R}_2 &= \boldsymbol{R}_{16}\boldsymbol{R}_{16}^T\\
\boldsymbol{R}_3 &= \boldsymbol{R}_{14}\boldsymbol{R}_{14}^T\\
\boldsymbol{R}_4 &= \boldsymbol{R}_{13}\boldsymbol{R}_{13}^T\\
\boldsymbol{R}_5 &= \boldsymbol{R}_{12}\boldsymbol{R}_{12}^T\\
\boldsymbol{R}_6 &= \boldsymbol{R}_{16}\boldsymbol{R}_{62}\boldsymbol{R}_{62}^T\boldsymbol{R}_{16}^T\\
\boldsymbol{R}_7 &= \boldsymbol{R}_{14}\boldsymbol{R}_{42}\boldsymbol{R}_{42}^T\boldsymbol{R}_{14}^T\\
\boldsymbol{R}_8 &= \boldsymbol{R}_{14}\boldsymbol{R}_{45}\boldsymbol{R}_{45}^T\boldsymbol{R}_{14}^T
\end{aligned}$$

In this way, RESCAL factorized relation matrices that were originally given as multi-object type data. Intuitively, $\boldsymbol{R}_3 = \boldsymbol{R}_{14}\boldsymbol{R}_{14}^T$ counts the number of common gene annotations between any two considered genes, and $\boldsymbol{R}_4 = \boldsymbol{R}_{13}\boldsymbol{R}_{13}^T$ counts the number of molecular pathways to which any two genes simultaneously belong. The reasoning behind the construction of other six relations considered by RESCAL follows similar principles.

When excluding a certain relation from a factorized model, we either removed it (in DFMF and RESCAL) or replaced the corresponding two symmetric relation matrices with zero matrices (in tri-SPMF and S-NMTF).

13.4.2 Data set selection with Forensic

Table 13.1 shows the changes of relative reconstruction error of target relation $\boldsymbol{R}_{12}$:

$$(\|\boldsymbol{R}_{12} - \widehat{\boldsymbol{R}}_{12}\|_{\text{Fro}} - \|\boldsymbol{R}_{12} - \widehat{\boldsymbol{R}}_{12}^{\text{excl.}}\|_{\text{Fro}})/\|\boldsymbol{R}_{12}\|_{\text{Fro}}, \tag{13.21}$$

when either a high-ranked or a low-ranked effect relation was excluded from inference of a factorized model. The 'excl.' in Eq. (13.21) indicates the estimate $\widehat{\boldsymbol{R}}_{12}$ when a reduced data collection was used as the input to the factorization algorithm.

The message of the results in Table 13.1 is two-fold. First, we can see that sensitivity estimates are well aligned with the changes of target reconstruction error. For example, omitting relation with the greatest effect on target relation in terms of Forensic's ϕ

Table 13.1

The changes of relative reconstruction error (%) of target relation R_{12} (or the corresponding matrix in the RESCAL model) when effect relations with the largest (1^{-}), the second-largest (2^{-}), the next-to-smallest (8^{-}) and the smallest (9^{-}) normalized Forensic's ϕ_n scores were excluded from a factorized model.

Model F_θ	$\Delta\mathrm{Err}(\boldsymbol{R}_{12}; F_{\theta,1^-})$	$\Delta\mathrm{Err}(\boldsymbol{R}_{12}; F_{\theta,2^-})$	$\Delta\mathrm{Err}(\boldsymbol{R}_{12}; F_{\theta,8^-})$	$\Delta\mathrm{Err}(\boldsymbol{R}_{12}; F_{\theta,9^-})$
DFMF	2.040	2.001	< 0.001	< 0.001
S-NMTF	4.179	3.084	< 0.001	< 0.001
tri-SPMF	3.427	3.153	< 0.001	< 0.001
RESCAL	13.832	4.640	0.655	0.637

score corresponded to the largest improvement of target approximation quality; similar observations held for smaller effect sizes. A surprising aspect of Forensic approach is its ability to estimate inter-relation sensitivity without the need to rerun inference algorithm. Thus, efficient estimation and its superiority over baseline approach that omits a relation and reruns the algorithm make Forensic an attractive option for understanding inter-relation structure when applying methods of collective matrix factorization to tens of different relations. Another message from Table 13.1 is that Forensic performs well for many different models of collective matrix factorization including multi-object type and multi-relation models.

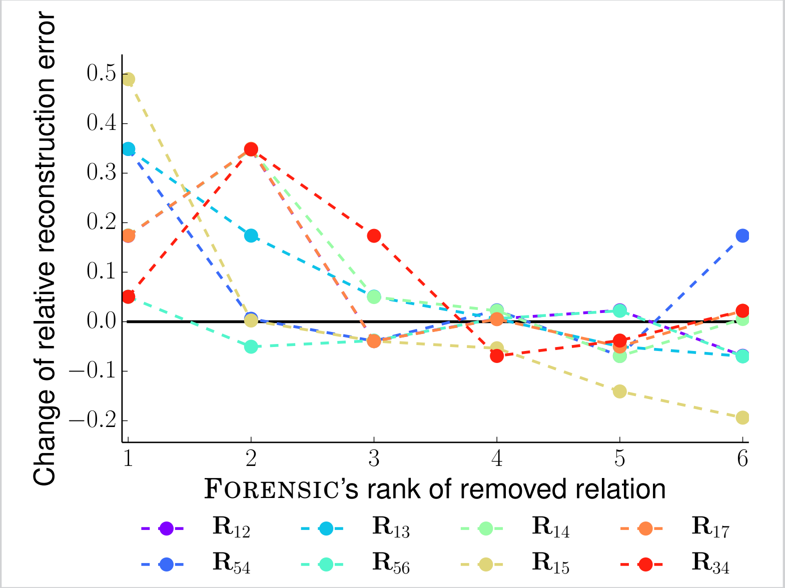
On a related note, we also observed the changes of relative reconstruction error of the entire data system:

$$\frac{\sum_{i,j} \|\boldsymbol{R}_{ij} - \widehat{\boldsymbol{R}}_{ij}\|_{\mathrm{Fro}}}{\sum_{i,j} \|\boldsymbol{R}_{ij}\|_{\mathrm{Fro}}} - \frac{\sum_{i,j} \|\boldsymbol{R}_{ij} - \widehat{\boldsymbol{R}}_{ij}^{\mathrm{excl.}}\|_{\mathrm{Fro}}}{\sum_{i,j;R_{ij}\ \text{not excl.}} \|\boldsymbol{R}_{ij}\|_{\mathrm{Fro}}}, \tag{13.22}$$

when, for a given target relation, either a high-ranked or a low-ranked effect relation was removed from the data collection considered by a factorization algorithm. Similarly as above, the 'excl.' in Eq. (13.22) indicates the estimate $\widehat{\boldsymbol{R}}_{ij}$ when a reduced data collection was used as the input to the factorization algorithm. Fig. 13.4 shows that for any choice of target relation from the relation graph in Fig. 13.1 the reconstruction error of the entire system reduces *more* when the effect relation with the *larger* Forensic's ϕ score is excluded from model inference. This can be seen from the decreasing trajectories of changes of reconstruction error in Fig. 13.4.

Figure 13.4

Exclusion of relations with high normalized Forensic's ϕ_n scores improves the reconstruction quality. We evaluated each of the relations from Fig. 13.1 in turn, and observed its forensic's ϕ_n scores when any of the remaining relations had the role of the effect relation. We then ordered the effect relations by the decreasing Forensic's ϕ_n scores and evaluated the relative reconstruction error of the entire data system when high- or low-ranked relations were excluded from inference of a DFMF factorized model. For a given target relation, its corresponding dashed line shows the relative reconstruction error defined in Eq. (13.22) when the effect relation with the largest ϕ_n score (i.e., rank-'1') up to the smallest ϕ_n score (i.e., rank-'6') was excluded from inference of a collective factor model.



Sometimes, exclusion of a relation can worsen model inference. Indeed, negative values in Fig. 13.4 indicate that exclusion of an effect relation reduced the quality of the inferred model. The good result is that Forensic is capable of detecting such relations and can do so without the need to re-run factorization algorithm on the reduced data collection. The results suggest that Forensic could be used for selection of data sets considered by a collective latent factor model. In particular, relations that score low by Forensic are those whose removal would in general worsen model quality. By contrast, relations that score high by Forensic are those whose removal would improve model quality. One could follow this guiding principle to select the appropriate data sets within a *single* run of factorization algorithm. Results in Fig. 13.4 are shown for the DFMF model, however, we note that similar behavior was observed for other factorization models as well.

Next, we investigated whether sensitivity estimates returned by Forensic depend on the parameters of model inference. In particular, we were interested if the number of iterations of a factorization algorithm and the selected latent dimensionality correlated with Forensic's ϕ scores. A high correlation between the scores and model parameters would be an undesired effect. Forensic's ϕ scores exhibited no dependence on the number of performed algorithm iterations during factorized model inference ($\tau > 0.9$, p-value $< 1 \times 10^{-3}$; all evaluated factorization models) or on reasonable selections of factorized model dimensionality ($\tau > 0.9$, p-value $< 1 \times 10^{-4}$; all evaluated factorization models). Note that we reported Kendall's tau coefficient of the scores obtained by varying the respective type of model parameter.

13.4.3 *Concordance of* Forensic *scores across factor models*

Table 13.2 shows concordance of Forensic's scores in different collective matrix factorizations when applied to the exact same data. Results indicate strong agreement of estimates across models and suggest that Forensic provides sensitivities that are properties of inter-relation structure rather than individual factorization algorithms. Paired with the relations are any number of algorithms that can be used to factorize them, that is, to find their latent representation. This appealing property of Forensic follows from the definition of ϕ score and the theoretical underpinnings of Forensic (cf. Eq. (13.12)).

13.4.4 *Discussion*

Results estimated by Forensic are consistent with genomics literature (Sifrim et al., 2013). Fig. 13.5 shows that, for example, target gene-phenotype relation ($\boldsymbol{R}_{12}$) was most sensitive to literature data ($\boldsymbol{R}_{15}$). Although mining the literature is a powerful way of identifying new associations, it tends to over optimistically identify straightforward candidates for which abundant knowledge is already available. It is likely that many known gene-phenotype associations are explicitly stated in the literature and thus, their latent factors relied strongly on the literature-derived relationship, hence high sensitivity. On the other side, relations about gene pathways ($\boldsymbol{R}_{13}$, $\boldsymbol{R}_{34}$) had mild effects on $\boldsymbol{R}_{12}$ and we observed insignificant reduction of prediction performance when these relations were excluded from learning.

Table 13.2

Kendall's tau (τ) coefficients of the correspondence of sensitivity estimates between different factorized models. For a given factorized model and a target relation we estimated the sensitivity of target relation to each of the remaining seven relations (effect relations) and ranked the relations by their normalized Forensic's ϕ_n scores. Correlation coefficients for a given target relation are either in the upper triangle (target is in the first row of a block) or in the lower triangle of each block (target is in the third row of a block).

	Target	DFMF	S-NMTF	tri-SPMF	Target	DFMF	S-NMTF	tri-SPMF
DFMF	$\boldsymbol{R}_{12}$	—	0.944	0.944	$\boldsymbol{R}_{25}$	—	0.833	0.833
S-NMTF		1.000	—	0.944		0.944	—	0.833
tri-SPMF	$\boldsymbol{R}_{13}$	1.000	0.833	—	$\boldsymbol{R}_{34}$	0.777	0.833	—
DFMF	$\boldsymbol{R}_{14}$	—	0.722	0.944	$\boldsymbol{R}_{35}$	—	0.944	0.777
S-NMTF		0.888	—	0.888		0.888	—	0.722
tri-SPMF	$\boldsymbol{R}_{15}$	0.833	1.000	—	$\boldsymbol{R}_{45}$	1.000	0.944	—
DFMF	$\boldsymbol{R}_{23}$	—	0.944	1.000				
S-NMTF		1.000	—	1.000				
tri-SPMF	$\boldsymbol{R}_{24}$	1.000	1.000	—				

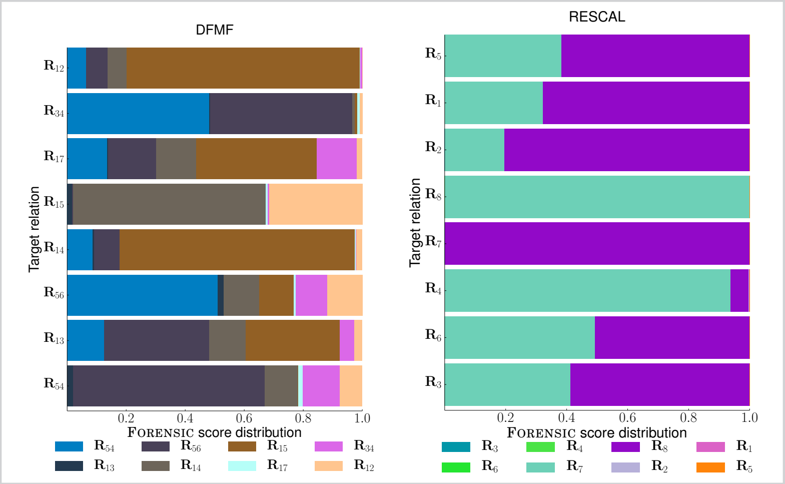


Figure 13.5

Normalized Forensic's ϕ_n scores representing sensitivity between any two relation matrices considered in the molecular biology study. For description of individual relations see Fig. 13.1. (left) In DFMF model, many relations were most sensitive to literature data, as can be seen from the width of the bands that correspond to R_{15}. (right) Similarly, in RESCAL model, all data sets exhibited the largest sensitivity to relations that were derived from literature data as shown by the width of the bands for matrices R_7 and R_8.

13.5 A case study with 40 human protein interaction data sets

Next we consider data sets from 40 human physical protein interaction studies. In what follows, we describe the data sets, report performance of Forensic and its utility for identification of potentially problematic data sets.

13.5.1 Experimental setup

We obtained human gene association network data from the GeneMANIA data archive (`http://genemania.org/data`) (Mostafavi et al., 2008). Forty protein physical interaction data sets listed in Table 13.3 were considered in the experiments. We represented each network with a symmetric network adjacency matrix, wherein protein-protein interactions were represented with nonnegative weights that corresponded to interaction strengths and were provided with the data.

Our case study was a multi-relation and two object-type task. This means that we had many relation matrices, $\boldsymbol{R}_{12}^{(1)}$ to $\boldsymbol{R}_{12}^{(40)}$, and they all encoded relationships between proteins. As can be seen from Table 13.3, data sets differ substantially in the number of interactions they contain. For example, human protein interaction network from the BioGRID data source (Chatr-aryamontri et al., 2014) contained more than 100,000 protein interactions curated from the primary biomedical literature, whereas many experimentally derived interaction networks were much smaller in size and each of them contained a few hundreds interactions (Table 13.3).

We evaluated an extension of recent DFMF algorithm (Žitnik and Zupan, 2015a) that performs matrix completion instead of matrix factorization. This means that a collective latent factor model is optimized over protein interactions that have been observed in model organisms so far. Matrix completion is a more realistic approach for our case study than matrix factorization. The reason is that interactions between proteins that have yet to be reported by the biologists are not viewed by a matrix completion algorithm as they would not exist (Rolland et al., 2014). In contrast, matrix factorization algorithms require dense matrices at their input and often make unrealistic assumptions about unknown values, such as substituting them with zero values.

We computed Forensic sensitivity scores for all target-effect relation pairs of data matrices $\boldsymbol{R}_{12}^{(1)}$ to $\boldsymbol{R}_{12}^{(40)}$ that were modeled by a matrix completion-based extension of DFMF algorithm. We aimed to investigate whether Forensic can detect potential experimental errors in the data, such as interaction mix-ups (Westra et al., 2011). We simulated interaction mix-ups in a given effect relation and deliberately introduced protein mix-ups by swapping entire interaction profiles, i.e. respective rows and columns in a given relation matrix. We then compared Forensic's scores of all relations before and after a selected effect relation was mixed-up.

13.5.2 Detection of surprising or problematic data sets

Experimental procedures in genomics and molecular biology typically involve many steps before actual analysis of the data, during each of which samples could be accidentally swapped. Sample mix-ups, cross-reacting antibodies or other experimental errors can arise during sample collection, handling, genotyping or data management (Ernst and Kellis, 2015). Since many studies are pushing toward larger sample-sizes in order to be able to get a more detailed view of cellular machinery, the presence of sample mix-ups becomes almost unavoidable (Westra et al., 2011).

For quality control in particular, we show in Fig. 13.6 and Fig. 13.7 that Forensic's scores are informative of potential errors present in the data. To simulate experimental errors in a data set we randomly swapped entire interaction profiles of $\boldsymbol{R}_{12}^{(31)}$ (Behrends et al., 2010) and obtained a permuted version of the original relation matrix, $\widetilde{\boldsymbol{R}}_{12}^{(31)}$. We then observed changes of Forensic's ϕ_n scores when $\widetilde{\boldsymbol{R}}_{12}^{(31)}$ was either a target or an effect relation. Compared with Fig. 13.6 results in Fig. 13.7 show that Forensic's ϕ_n scores of $\widetilde{\boldsymbol{R}}_{12}^{(31)}$ increased substantially relative to the scores of $\boldsymbol{R}_{12}^{(31)}$. These observations were consistent for different choices of target relations and when $\widetilde{\boldsymbol{R}}_{12}^{(31)}$ had the role of an effect relation. We note that we observed similar trends when simulating experimental errors in other data sets. These results suggest that Forensic can reveal problematic data sets that would otherwise unintentionally be included in a collective latent factor model.

Table 13.3

Human gene protein interaction networks.

Matrix	Data set	Genes	Interactions	Matrix	Data set	Genes	Interactions
$R_{12}^{(2)}$	Kneissl et al. (2003)	162	81	$R_{12}^{(27)}$	Lehner and Sanderson (2004)	854	427
$R_{12}^{(4)}$	Ouyang et al. (2009)	210	105	$R_{12}^{(15)}$	Shi et al. (2011)	538	269
$R_{12}^{(5)}$	Napolitano et al. (2011)	162	81	$R_{12}^{(18)}$	Bennett et al. (2010)	8,772	4,386
$R_{12}^{(13)}$	Wagner et al. (2011)	9,180	4,590	$R_{12}^{(23)}$	Wang et al. (2011b)	6,798	3,399
$R_{12}^{(8)}$	Jager et al. (2012)	70	35	$R_{12}^{(12)}$	Havugimana et al. (2012)	27,420	13,710
$R_{12}^{(21)}$	Varjosalo et al. (2013)	612	306	$R_{12}^{(7)}$	Reinke et al. (2013)	286	143
$R_{12}^{(9)}$	Zanon et al. (2013)	388	194	$R_{12}^{(28)}$	Abu-Odeh et al. (2014)	432	216
$R_{12}^{(24)}$	Nakayama et al. (2002)	252	126	$R_{12}^{(29)}$	de Hoog et al. (2004)	450	225
$R_{12}^{(38)}$	Jin et al. (2004)	466	233	$R_{12}^{(22)}$	Jones et al. (2006)	344	172
$R_{12}^{(3)}$	Cannavo et al. (2007)	202	101	$R_{12}^{(14)}$	Behzadnia et al. (2007)	224	112
$R_{12}^{(36)}$	Alexandru et al. (2008)	228	114	$R_{12}^{(6)}$	Barr et al. (2009)	330	165
$R_{12}^{(35)}$	Brehme et al. (2009)	1,158	579	$R_{12}^{(40)}$	van Wijk et al. (2009)	612	306
$R_{12}^{(20)}$	Ravasi et al. (2010)	1,270	635	$R_{12}^{(31)}$	Behrends et al. (2010)	1,532	766
$R_{12}^{(39)}$	Kahle et al. (2011)	264	132	$R_{12}^{(26)}$	Tarallo et al. (2011)	478	239
$R_{12}^{(19)}$	Rowbotham et al. (2011)	228	114	$R_{12}^{(11)}$	Pichlmair et al. (2011)	200	100
$R_{12}^{(17)}$	Gao et al. (2012)	324	162	$R_{12}^{(1)}$	Wong et al. (2012)	230	115
$R_{12}^{(32)}$	Woods et al. (2012)	1854	927	$R_{12}^{(34)}$	Blandin et al. (2013)	1,316	658
$R_{12}^{(37)}$	Roy et al. (2013)	306	153	$R_{12}^{(10)}$	Foerster et al. (2013)	322	161
$R_{12}^{(30)}$	BIND	14,156	7,078	$R_{12}^{(16)}$	BioGRID	254,414	127,207
$R_{12}^{(33)}$	InnateDB	7,740	3,870	$R_{12}^{(25)}$	OPHID	88,984	43,492

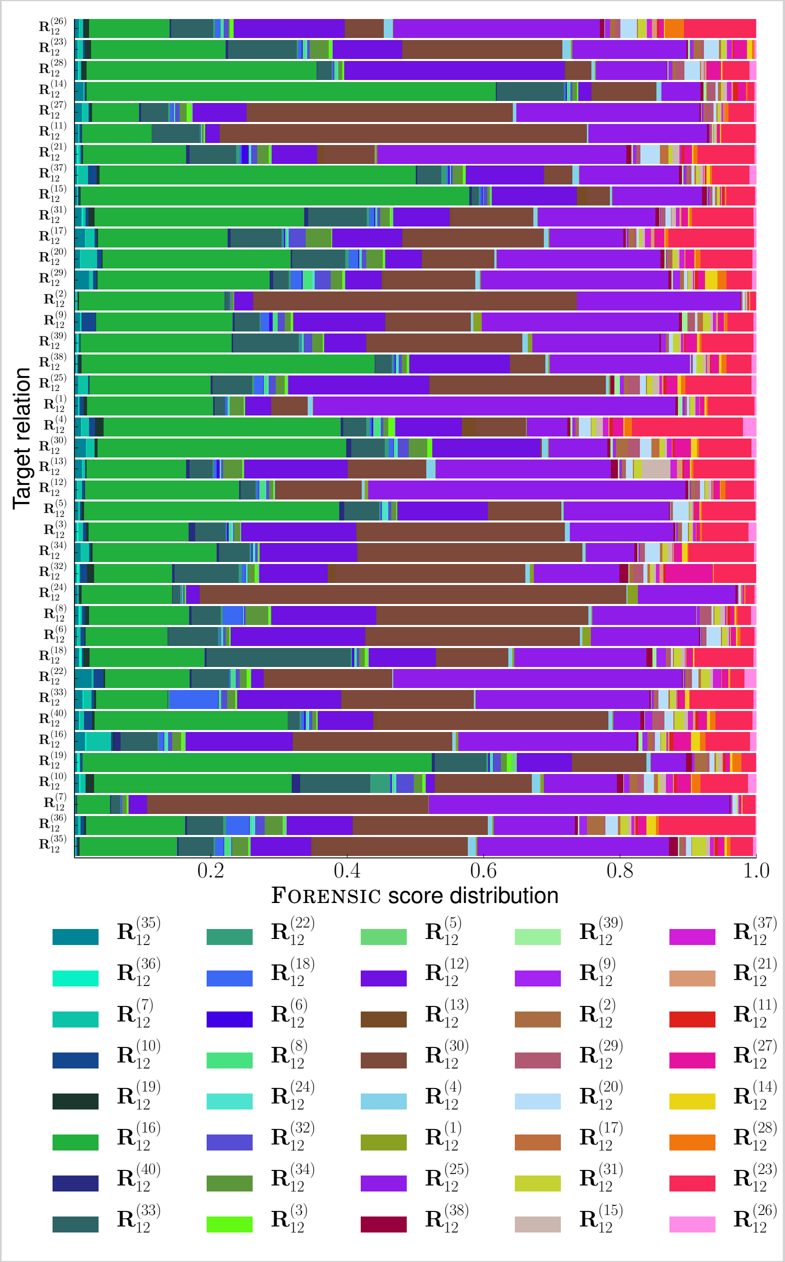


Figure 13.6

Normalized FORENSIC's ϕ_n scores representing sensitivity between any two relation matrices considered in the human gene association study. For description of individual relations see Sec. 13.5.1. For a given target relation shown in the left, a horizontal bar chart represents a distribution of FORENSIC's ϕ_n scores when each of the remaining thirty-nine relations has, in turn, the role of an effect relation.

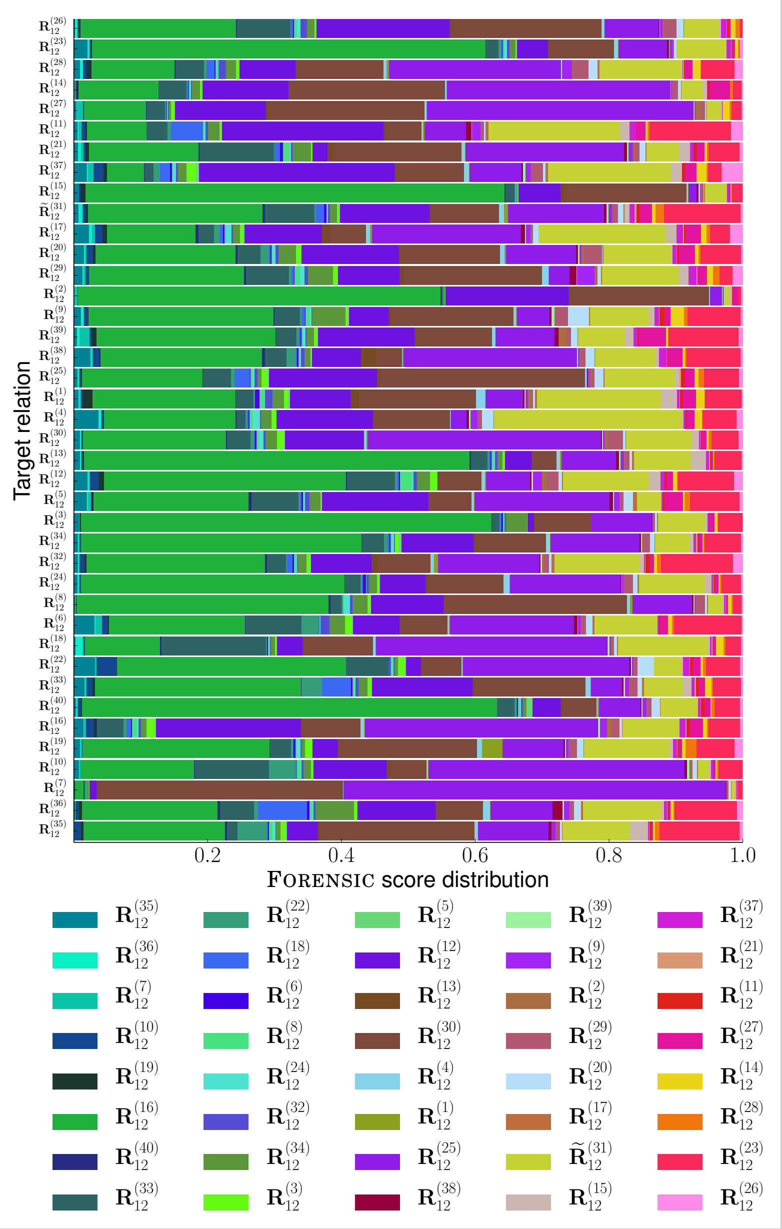


Figure 13.7

Normalized FORENSIC's ϕ_n scores representing sensitivity between any two relation matrices considered in the human gene association study when the protein interactions in relation $R_{12}^{(31)}$ (Behrends et al., 2010) were randomly permuted. Permuted relation matrix is denoted by $\widetilde{R}_{12}^{(31)}$. For description of individual relations see Sec. 13.5.1. For a given target relation shown in the left, a horizontal bar chart represents a distribution of FORENSIC's ϕ_n scores when each of the remaining thirty-nine relations takes, in turn, the role of an effect relation. Compared with results for $R_{12}^{(31)}$ in Fig. 13.6, this figure shows a substantial increase in FORENSIC's ϕ_n scores of $\widetilde{R}_{12}^{(31)}$ for most choices of target relations.

13.6 Conclusion

The estimation of inter-relation sensitivity in collective latent factor models opens many new applications that were previously not possible. We demonstrated two such applications of Forensic, our new approach to sensitivity estimation. In a study with data sets from molecular biology, we used Forensic's scores to identify data sets that worsened latent model quality and to select complementary data sets, which improved the quality of relations modeled by a collective matrix factorization algorithm. In another study we considered forty human protein association data sets, the largest number of data sets analyzed by a collective latent factor model to date. We used global discrepancies between Forensic's sensitivity scores as a quality control metric. By simulating experimental errors, we were able to detect surprising and potentially problematic data sets.

We believe that *post hoc* methods, such as Forensic, which provide insights into estimated latent factor models, represent a structured approach when handling multiple large-scale and sparse data sets with latent factor models. Methods offering such functionality are currently scarce, however, we expect that they will quickly become a valuable tool and a necessary step in doing state-of-the-art data fusion.

Part VII

Conclusions and future directions

The abundance and the ubiquity of complex data and rich computing applications in the life sciences provide computer science with a unique opportunity to design and build computing systems and applications capable of handling large volumes of heterogeneous data. Indeed, producing large quantities of genomic data is now relatively easy, but analyzing these data is not (Vihinen, 2015). For example, working out whether a particular genetic variation of an individual is important relative to the reference genome, and understanding the roles of these variants in disease, is a complex and time-consuming quest. To fulfill the potentials of incentives, such as recently unveiled "The Precision Medicine Initiative" (Collins and Varmus, 2015), we need to develop *scalable, reliable* and *integrative* data analysis tools that can draw the connections between genetic variation and disease, which are then further analyzed by domain experts.

Our Thesis presents a combination of (i) empirical work and experiments, (ii) design and analysis of prediction models, and (iii) development of machine learning algorithms and data mining tools. The research focus of this Thesis is to analyze and model large heterogeneous data compendia via methods of data fusion. Our contributions so far are the following. We introduced the probabilistic matrix completion model that can consider side information presented with networks. We also developed two approaches to network inference: the epistasis-based probabilistic model for gene network inference, and the general network inference model that can fuse data from many potentially nonidentical data distributions. We designed algorithms for efficiently estimating its parameters. In addition to methods that model single and dual heterogeneity, we also introduced the collective matrix factorization model for emerging applications that exhibit triple data heterogeneity. We also introduced the technique that reduces a triple data heterogeneity problem to a problem with dual data heterogeneity. Furthermore, we developed the algorithm in which our collective matrix factorization model walks hand in hand with the regression-based survival model (Fig. VII.1).

On the application side, we presented analyses of the roles genes have in cells, associations between diseases and drug toxicity levels. We showed that networks inferred from cancer genomic data and prediction of cancer patient survival time benefit from inclusion of circumstantial evidence. We also showed that the ability to accurately predict genetic interactions does not simply increase monotonically with the number of available interaction measurements, but rather reflects more subtle features of genetic interaction landscape. Finally, in a fruitful collaboration with biologists from Baylor

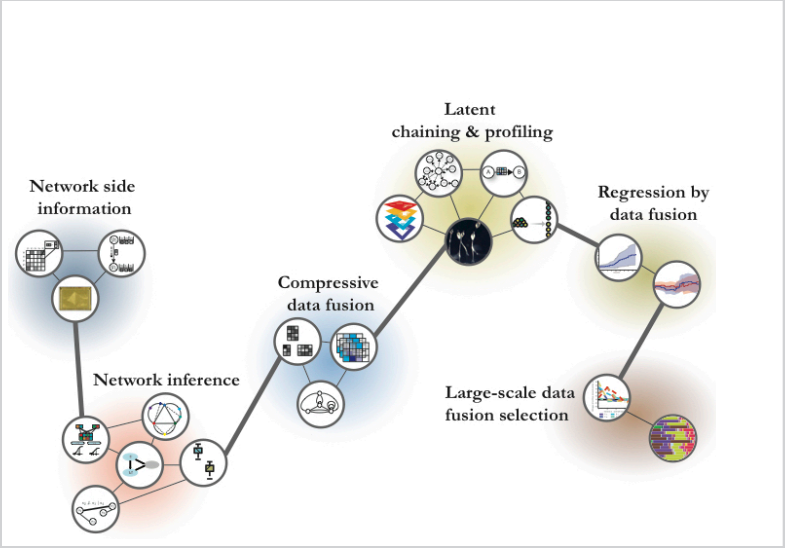

Network side information
Latent chaining & profiling
Regression by data fusion
Compressive data fusion
Network inference
Large-scale data fusion selection

Figure VII.1

The map of this Thesis.

College of Medicine we were able to successfully validate eight genes, which were predicted by our gene prioritization method to have a role in the bacterial resistance of *Dictyostelium*.

Last, we also showed how working with large data compendia gives us opportunities to arrive at observations that are practically invisible at small scales. We demonstrated this by analyzing the largest number of data sets with a collective latent factor model so far, and made novel observations about selection of data sets from which data fusion might benefit.

In the long run, outside the scope of this Thesis, we would like to build tools for modeling heterogeneous data both in the life sciences and also in other data domains, such as in the social sciences. We want to study how complementary different data perspectives are and how to harness data diversity to improve prediction modeling. Ideally, we would like to marry these two views, so that we can detect "surprising" patterns that bridge across the disciplines.

Next, we give a summary of contributions and our vision for future work.

Summary of contributions

We summarize our contributions by grouping them by the columns defined by the Thesis structure in Table 14.1. The Thesis adheres to the following three types of data heterogeneity and their combinations. (1) In relation data heterogeneity, learning by fusing heterogeneous data aims to harness heterogeneous input data spaces. (2) In object type heterogeneity, approaches to data fusion leverage heterogeneous types of features. (3) In task heterogeneity, data fusion exploits related prediction tasks to transfer knowledge between data views.

Table 14.1

Structure of this Thesis with references to the parts.

Thesis		Types of data heterogeneity		
part		Relation	Object type	Task
Part I	Network side information	✓		✓
Part II	Network inference	✓		
Part III	Compressive data fusion	✓	✓	✓
Part IV	Latent chaining and profiling		✓	
Part V	Regression by data fusion		✓	✓
Part VI	Large-scale data fusion selection	Exploring types of heterogeneity		

Relation heterogeneity (Part II):

- We developed FuseNet, an *off-the-shelf network inference framework* for mixed data arising from any combination of exponential family distributions. Moreover, FuseNet is the first model that is able to combine the theory of Markov network inference, latent factor models and data fusion.
- We developed Réd, an approach to epistasis-based gene network inference that is able to reconstruct known cellular pathways more accurately than present state-of-the-art methods.
- Using Réd we were able to infer networks consistent with the theory of epistasis analysis by considering hundreds of thousands of genetic interaction measurements, the largest data compendium considered for epistasis analysis up to date.

- We analyzed heterogeneous data from the International Cancer Genome Consortium and found that joint network inference by FuseNet from multiple related data sets, i.e. RNA-sequencing and somatic mutation data, showed greater functional enrichment than networks learned from any data type alone.

Object type heterogeneity (Part IV):

- We developed Collage, an approach to gene prioritization. Given a handful of *seed genes* important for a biological function of interest, Collage aims to identify the most promising candidate genes for further studies. Collage represents a major advancement relative to *gene-centric prioritization algorithms* in the breadth of data it can incorporate, the ease of data integration without complex feature engineering, and the ability to retain the relational data structure during model inference.
- We provided a new formalization of gene prioritization and designed models for assessment of drug toxicity and discovery of disease-disease associations that have had a wide range of implications for researchers in the life sciences. For example, the identification and characterization of four seed genes for the bacterial resistance study in *Dictyostelium* was a laborious task that required several months of laboratory work per gene. Collage *has substantially simplified this task by suggesting eight genes that have been successfully validated in the wet lab.*

Dual data heterogeneity (Parts I and V):

- We developed the network-guided matrix completion, which is mathematically tractable and general probabilistic matrix completion model. Network-guided matrix completion is unique in *fusing relational data* with *network side information* by inferring a single predictive model. It achieves better generalization than competing approaches in predicting genetic interactions.
- We showed that our work on analyzing genetic interaction data has high practical value for the prediction of *entire gene interaction profiles* for genes whose interactions otherwise cannot be measured directly due to limits of biotechnology.

- We developed DFMF-SR, a data fusion model of survival regression and an efficient algorithm for the estimation of its parameters. We analyzed cancer data from the International Cancer Genome Consortium and showed that DFMF-SR performs well relative to a popular approach that first transforms data into the latent space and then does survival regression independently of data transformation. Moreover, DFMF-SR is the first approach that is able to infer a latent data model and regression coefficients of a survival model *at the same time*.

Triple data heterogeneity (Part III):

- We developed DFMF, an algorithm for collective matrix factorization and its variant for collective matrix completion. We proved that latent matrices found by our algorithm for the DFMF model locally minimize the total reconstruction error of a data system presented with the data fusion graph.
- We found that latent matrices estimated by the DFMF algorithm have high predictive power and compare favorably to techniques that transform data into a single feature-based data table, i.e. *early integration*, and to techniques that explicitly address the multiplicity of data via multiple kernel learning, i.e. *intermediate integration*.

Exploring types of data heterogeneity (Part VI):

- We developed Forensic, a *general* and *computationally efficient* approach to inter-relation sensitivity estimation in collective latent factor models. Furthermore, Forensic is the first principled model offering such functionality for collective latent factor models and shows a potential to be used as a scoring technique for selection of data sets for data fusion.
- We analyzed a compendium of 40 experimental protein physical interaction data sets, which is to the best of our knowledge the *largest collection of data sets examined with a collective latent factor model* up to date. We demonstrated that Forensic can correctly pinpoint corrupted data.

15

Goals for the future

Our long-term research goal is to tackle large data compendia to understand, model, predict, and finally, enhance biological, technological and social systems. We would like to create accurate and explanatory predictive models of relationships and roles of large groups of biological entities, e.g., genes, drugs and diseases; societal entities, e.g., people, communities, social events; and technological systems, e.g., the web. Many times, complementary data descriptions of various entities are available and integrative methods of machine learning and statistics can be applied to heterogeneous data, which yield effective models with boosted prediction ability. Based on our research experience and recent results, we believe that the study of latent and factor models is one of the promising ways to develop such understandings, as these models can naturally share information between related data views, different types of objects and various predictive tasks.

In our Thesis research, we made several steps towards this long-term goal. We now better understand mixed, multiscale, multiplex and multislice data and models that connect the different types of data heterogeneity. Moreover, we can efficiently fit the latent models to the data and make predictions about the systems. We also have a clearer view of how inclusion of circumstantial evidence affects the model performance, what the relational structure of data systems are, and how to select relevant data for fusion.

On the way towards the long-term goal, our research will center on three dimensions: (1) addressing problems with multiple types of data heterogeneity and designing powerful models to encompass rich and complex data, (2) scaling up the analysis to huge and massive data collections, and (3) developing explanatory system-view models.

15.1 Medium-term aims

We first allude to the medium-term future aims that build on the work presented in the Thesis.

Time-dependent and locality-aware analysis of multiscale and multiplex data

Many experiments in the biological and technological sciences generate series of measurements that are snapshots of different states that a particular system might be in.

The series of measurements might be taken at different scales and positions within the system, or recorded at different stages of the system's operation. For example, an RNA-seq experiment measures RNA-content within a cell population and produces data in the form of millions of short nucleotide sequences that are informative of the activity of genes in a particular environment and developmental stage of the organism. Profiling gene expression over time then provides information about the dynamical behavior of genes. Moreover, the precise roles of genes frequently depend on their tissue context and cell-type identity. We want to understand how, for example, in the biological domain, genes that participate in distinct cellular processes according to developmental and anatomical context, rewire in different tissues to associate with different functional partners, and, more abstractly, how pathways rewire, arise and decay in different contexts. We would like to design and explore *integrative methods* that can answer questions that are *specific* to individual system's components, e.g., a single gene in a single tissue, which is important, for e.g., human diseases, where tissue and cell-type specific factors combine in the context of a whole organism. We believe that the key here is to relate data on a microscopic scale with a macroscopic view, connect local to global, and complement data about comprehensive system's operation with the assays on specific features of interest.

Data fusion selection

Beyond simply including more data into the analysis, one can try to understand levels of *consistency* across different data views and degrees of *relatedness* of various predictive tasks. In practice, data fusion often encounters one or more of the following issues. Data relations are typically incomplete, where each relation contains a subset of the total set of objects in the domain. Furthermore, patterns that are present in one data set, can be largely or entirely absent from another data set. Such disagreements can be the result of unique properties of the problem domain, or can simply arise due to noise and experimental errors. Explorations in this realm could lead to data fusion strategies that would help to identify the most informative data sets for a given predictive task or data subparts of questionable quality. A natural next step is then to suggest experiments that one could perform to collect data, which would maximally boost predictive performance of data fusion methods. In this context, we plan to continue with our work on modeling sensitivity and interdependence of data sets in large data compendia. Furthermore, better understanding of the collected data could help us decide the type of integrative data analysis that is most appropriate for the predictive task at hand.

Huge data and scalability

Another important aspect of our research work focuses on large-scale data and analysis architectures for manipulation of large data collections. To handle such *data compendia with hundreds of data sets and billions of data points*, scalability becomes an issue. Alternating least squares and stochastic gradient descent type algorithms are two popular approaches that were employed in several parallelizations of the latent factor algorithms. However, alternating least squares type techniques are not scalable to large-scale data due to their cubic time complexity in the dimensionality of the latent model, i.e. the factorization rank. On the other hand, the updates of stochastic gradient descent are efficient but usually have slow convergence. The question here is what kind of matrix update sequences can be easily parallelized on multi-core and distributed systems to scale to thousands of machines. Here, additional challenges arise when we want to *jointly co-factorize multiple matrices,* one of our primary research interests, for which parallel and distributed algorithms that can exploit thread-level parallelism, in-memory processing and asynchronous communication have yet to be developed. We plan to extend our software library to parallel architectures and batch and streaming scenarios, and explore *coordinate descent* type optimization techniques to scale to many skinny and wide data matrices with billions of data points.

15.2 Long-term vision

Last, we present the long-term goals of our research.

Universal data fusion

The scope of the Thesis is centered on current challenges in bioinformatics and systems biology. We tackle these challenges by developing mathematically sound and computational data fusion methods, which capture interesting patterns and relationships. However, we believe there is still a long way to a truly *comprehensive data fusion of everything*. For this reason it is important to study how our results translate to other data domains, e.g., the social sciences and the fields of engineering and technology. Human activities leave digital traces in various data systems, which collectively capture our "social genome," the footprints of our society. On the other hand, for example, experiments in physics and engineering have already generated massive heterogeneous data discerning the "technological genomes," the blueprints of natural phenomena seen through the lenses of technology. Like the human genome, the "social genome" data and the "technological genomes" data have much buried in the massive almost chaotic data compendia. Here, we plan to continue our preliminary work in collective data analysis models for predictive tasks beyond those presented in the Thesis, e.g., for classification and ranking (Žitnik and Zupan, 2016). This line of research will allow us to respond to the specific requirements of science, society and technology, and make predictions with levels of reliability that cannot be achieved by considering a single data perspective.

Explanatory multi-modeling with dynamic feedback

Ultimately, understanding a phenomenon entails development of both *accurate* as well as *explanatory* models that can *continuously change* as new evidence arrives to add circumstantial support. To fulfill the vaunted promise of precision medicine, i.e. prevention and treatment strategies that take individual variability into account (Collins and Varmus, 2015), substantial gains still need to be made in computational methods for data analysis, integration and interpretation. In precision medicine, for example, incorporating the variety of information about environmental exposures, genetic exposures, and prior clinical courses may *better explain* the disease burden. We plan to focus on machine learning methods for determining causality and evidence evaluation in incremental and online learning settings, which will allow for more explanatory prediction models.

Some further last words

Data become most powerful when integrated. Fragmented efforts to make prediction from a single data source or of a single data type are neither effective nor efficient. Furthermore, current integrative approaches are often inaccessible to domain experts while making data less useful. On a long term we envision facilitating the massive amounts of data by building a transparent infrastructure of tools and machine learning algorithms that will make our decision making more informative.

> *What if knowing the daily habits of a patient's Facebook friends could enhance predicting patient's clinical outcome of a selective drug therapy (Christakis and Fowler, 2014)? Consider the totality of information! Whether you are mapping pathways in cancer, matching genetics to phenotype, modeling the electrical behavior of neurons, or recommending which product to buy next, data fusion will make it easy to build artificial intelligence into the "fuseome" — the connections among all of the information-rich resources across disciplines, scales and data types. Soon you will add your data to the fuseome, choose the type of prediction you want to get returned via an "app", and you will be alerted when the results are in.*

If properly designed and interpreted, this "fuseome" could one day offer insights into many of the most challenging problems facing our society.

Del A

Razširjeni povzetek

A.1 Uvod

V naših raziskavah poskušamo razumeti različne vrste heterogenosti podatkov, s katerimi se soočamo pri gradnji napovednih modelov. To znanje nato uporabimo za razvoj *učinkovitih* in *zmogljivih* algoritmov za učenje v heterogenih podatkovnih okoljih. Razviti algoritmi pri gradnji napovednih modelov uporabljajo predznanje, ki je lahko podano v različnih podatkovnih formatih, kot so tabele značilk, ontologije in mreže, in opisano z značilkami različnih tipov. Cilj sočasne obravnave večih podatkovnih virov tekom gradnje napovednega modela je izboljšanje kakovosti modela, ki jo ocenimo z raznimi merami za oceno točnosti napovedi, zmožnostjo razlage in razumljivosti napovedi ter delom problemskega področja, ki ga je model sposoben obravnavati. Nekaj vprašanj, ki si jih zastavimo v disertaciji in nanje tudi odgovorimo, je sledečih. Kateri so učinkoviti in uspešni pristopi za vključitev dodatnih podatkovnih virov v učenje? Kako znanje o stranskih učinkih zdravil vključiti v model, ki napoveduje povezave med boleznimi? Ali, kako upoštevati različne vloge filmskih igralcev, ko želimo uporabniku predlagati film, ki bi ga utegnili zanimati? Kako lahko zlijemo heterogene podatkovne prostore in zgradimo enotni napovedni model z odlično napovedno uspešnostjo? Kateri podatkovni viri so komplementarni pri gradnji napovedi? Kako dovolj zgodaj zaznati problematične podatkovne nabore z napačnimi meritvami, ko sočasno obravnavamo več deset ali celo več sto podatkovnih naborov? Odgovori na tovrstna vprašanja so pomembna v številnih sodobnih izzivih znanosti, tehnologije in družbe, kjer lahko zberemo veliko podatkov, ki sisteme opisujejo z različnih zornih kotov in beležijo delovanje njihovih sestavnih delov.

Vseprisotnost domen z veliko raznovrstnih podatkov ponuja priložnosti za uporabo metod, ki gradijo modele z zlivanjem podatkov, a hkrati predstavlja številne algoritmične izzive. Kako lahko povežemo na prvi pogled neodvisne napovedne naloge, da izboljšamo napovedno uspešnost? Včasih se zdi, da med različnimi nalogami ne moremo vzpostaviti povezovalnih elementov, če so podatkovni nabori, ki tem nalogam pripadajo v *povsem* neprekrivajočih se podatkovnih prostorih. Na primer, v večjezikovnem uvrščanju je lahko prva naloga uvrščanje zbirke angleških dokumentov, katerih podatkovni prostor sestoji iz angleškega besednjaka, druga naloga je lahko uvrščanje zbirke slovenskih dokumentov, katerih vhodni prostor je sestavljen iz slovenskega besednjaka. S podobno heterogenimi podatkovnimi viri se srečamo ob hkratnem razvrščanju dokumentov in slik. Tu je lahko prva naloga gručenje dokumentov, opisanih z besedilnimi značilkami, in druga naloga razvrščanje slik, opisanih z značilkami izračunanimi nad različnimi slikovnimi območji. Razvoj novih tehnologij je omogočil zbiranje veliko raznovrstnih podatkov ne le na področju analize besedil in slik, temveč tudi v vedah o življenju, kot je biologija. Pri napovedovanju genskih funkcij tako napovednim nalogam ustrezajo različni biološki procesi in vloge, ki jih imajo geni v celici, vsak proces pa lahko opišemo z relevantnimi genskimi potmi in boleznimi, povezanimi z okvarami teh poti. Možnost skupnega učenja večih nalog, podanih v heterogenih podatkovnih prostorih, tako da učenje ene naloge izboljša učenje povezanih nalog, je ključnega pomena na številnih področjih, med katerimi so večjezikovno razvrščanje besedil, gradnja priporočilnih sistemov, odkrivanje povezav med boleznimi, napovedovanje toksičnosti zdravil in genskih funkcij ter načrtovanje eksperimentov v biologiji.

Študije, ki gradijo napovedne modele z zlivanjem heterogenih podatkov, tipično predpostavljajo, da so podatkovni prostori različnih napovednih nalog vsaj posredno povezani. V primeru večjezikovnega uvrščanja se zdi naravna povezava med besedami iz dveh različnih jezikih (na primer, ”boat” v angleščini pomeni ”čoln” v slovenščini); v primeru sočasnega razvrščanje dokumentov in slik lahko besede prevedemo v slikovna področja; v primeru napovedovanje genskih funkcij lahko vzpostavimo odnose med geni preko interakcij proteinov, ki jih ti geni kodirajo, ali preko komorbidnosti bolezni, ki jih ti geni povzročajo. Odvisnosti med različnimi podatkovnimi prostori tako vzpostavljajo pomembne povezave med različnimi napovednimi nalogami.

Cilj učenja z zlivanjem heterogenih podatkov je izkoriščanje različnih podatkovne heterogenosti za izboljšanje učinkovitosti napovednega modeliranja. V pričujoči disertaciji preučujemo tri vrste podatkovne heterogenosti in njihove kombinacije, ki prepletajo več vrst heterogenosti in vodijo v vse bolj računsko in algoritmično zahtevne izzive:

- *Heterogenost podatkovnih relacij*: V primerjavi s tradicionalnim razumevanjem heterogenosti napovednih nalog, kjer so različne naloge opisane s *homogenimi* podatkovnimi prostori, lahko učenje z zlivanjem podatkov izmenjuje vzorce med *številnimi* potencialno *heterogenimi* vhodnimi prostori.

- *Heterogenost tipov objektov*: V primerjavi s tradicionalnim razumevanjem heterogenosti podatkovnih relacij, kjer so učni primeri opisani z značilkami *enega tipa* v različnih podatkovnih relacijah, zlivanje podatkov obravnava *heterogene* *vrste značilk*, da se izboljša učinkovitost učenja posameznih napovednih nalog.

- *Heterogenost napovednih nalog*: V primerjavi z uveljavljenimi pristopi za analizo značilk, ki opisujejo več tipov objektov, in napovedne naloge obravnavajo *ločeno* preko različnih tipov objektov, zlivanje podatkov izkorišča *morebitno povezanost med nalogami*, s čimer se prenaša znanje med različnimi podatkovnimi relacijami.

Motivacija

Uveljavljeni pristopi pri gradnji napovednih modelov tipično uporabljajo en podatkovni nabor in učne primere predstavijo z vektorji značilk. Na primer, pri razpoznavanju rakavega tkiva lahko vsak vzorec tkiva opišemo z vektorjem (profilom) genskih izrazov v danem tkivu in z binarno spremenljivko, ki kaže, ali je vzorec rakavega izvora ali ne. Številni pristopi strojnega učenja in tehnike odkrivanja znanj iz podatkov razviti v zadnjih desetletjih obravnavajo tabelarične podatkovne predstavitve in gradijo modele za napovedovanje ciljnih spremenljivk, kot so verjetnost razvoja bolezni v posamezniku. Čeprav so ti modeli zmogljivi in imajo veliko izrazno moč, pogosto ne morejo obravnavati različnih podatkovnih predstavitev, ki izhajajo iz heterogenih podatkovnih prostorov. Poleg tega moramo pogosto analizirati več deset ali celo več sto podatkovnih naborov, da lahko zanesljivo ocenimo vrednost ciljne spremenljivke; torej, naše raziskovanje se osredotoči na sočasno računsko analizo velike heterogene zbirke podatkov.

Vseprisotnost visoko prepustnih tehnologij v naravoslovnih, humanističnih in tehnoloških vedah poraja veliko možnosti za študij pojavov in sistemov v velikem obsegu in iz različnih perspektiv, kar še pred kratkim ni bilo mogoče. To je mogoče povzeti z naslednjimi tremi točkami:

- Meritve naravnih sistemov (na primer človeškega genoma) in tehnoloških sistemov (na primer splet) vsebujejo podrobne podatke, ki opisujejo kompleksne odnose med številnimi objekti različnih tipov (kot so geni, molekule RNK in celične poti v primeru človeškega genoma; uporabniki, dogodki in skupnosti v primeru spleta), kjer so objekti posredno in na vnaprej neznani način povezani s ciljno spremenljivko (na primer pomen posameznikovega okolja na razvoj genetske bolezni; vpliv prijateljev s spletnega omrežja na posameznikovo naklonjenost izbranemu filmu).

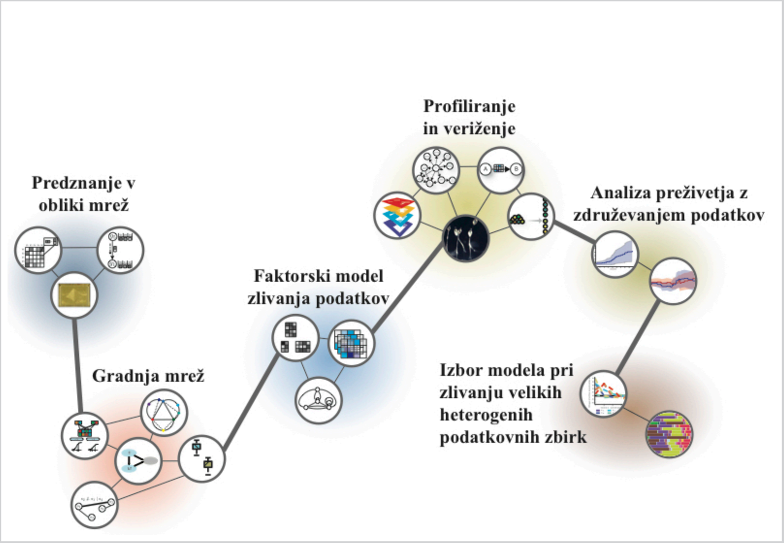
- "Velike podatke" zbrane s tovrstnimi meritvami je mogoče razumeti kot "veliko zbirko manjših ali srednje velikih podatkovnih naborov" v primerjavi z alternativnim pogledom "ene velike podatkovne tabele" (Zoubin Ghahramani, osebna komunikacija).

- Tako bogati podatki so podani z različnimi stopnjami negotovosti in opisani z raznimi podatkovnimi predstavitvami, kot so tabele, povezave, omrežja in ontologije.

Na primer, projekt ENCODE (Consortium et al., 2012) je enciklopedija elementov DNK, ki si prizadeva določiti vse funkcionalne elemente v človeškem genomu. Računski strokovnjaki in biologi uporabljajo visoko prepustne biotehnološke pristope za določanje zaporedij DNK, ki imajo biološke funkcije. Ta nedavni vir informacij je poleg človeškega genoma (Venter et al., 2001) in številnih meritev v molekularni biologiji in funkcijski genomiki privedel do razvoja *sistemske biologije*, ki si prizadeva celostno analizirati biološke sisteme (na primer, razpoznava sekvenčnih variant, ki povzročajo bolezni in stratifikacija bolnikov z rakom). Primeri drugih podatkovno intenzivnih področij vključujejo: podatke, ki jih proizvede eksperiment ATLAS v projektu CERN (Toor et al., 2012), ki išče nove delce s trkanjem protonov pri visoki energiji in odkriva različne vrste dogodkov; spletni priporočilo sistemi (Feuerverger et al., 2012), ki upoštevajo podatke o preteklih ogledih filmov, demografske profile uporabnikov in informacije o filmih, igralcih ter žanrih, da nudijo podporo več sto tisočim uporabnikov pri izbiri filmov; globalni satelitski navigacijski sistemi, ki zlivajo podatke za izboljšanje zanesljivosti pozicioniranja in optimizacije prostorske geometrije (Li et al., 2015); ali na primer spletna družbena omrežja (Szell et al., 2010; Mucha et al., 2010), ki zajemajo kompleksne komunikacijske vzorce, kot so "všečki," "glasovi," in kaskade razširjanja objav med posamezniki oziroma spletnimi skupnostmi.

Analiza podatkov, ki se beležijo v tovrstnih sistemih, predstavlja številne edinstvene priložnosti in izzive. Ozko grlo, ki nam preprečuje, da bi bolje razumeli in resnično zlili raznovrstne podatke v velikem obsegu predstavlja opredelitev znanja, ki se lahko prenaša med podatkovni nabori, tipi značilk in napovednimi nalogami. V disertaciji predlagamo algoritme, ki za vzpostavitev povezav med heterogenimi podatkovnimi viri uporabljajo eno ali več izmed naslednjih predpostavk:

- *Prenos podatkovnih relacij/pogledov:* Gradimo karto podatkovnih relacij — *graf zlivanja*, ki opisuje odnose med heterogenimi podatkovnimi viri. Za učinkovito sočasno obravnavo heterogenih podatkovnih virov se pogosto ne moremo zanašati na predpostavko o neodvisno in enako porazdeljenih podatkovnih virih.
- *Prenos tipov objektov:* Tu predpostavljamo, da obstajajo značilke v podatkih, ki so skupne različnim podatkovnim prostorom. Te značilke izkoristimo za prenos znanja med heterogenimi podatkovnimi domenami.
- *Izmenjava učnih parametrov:* Tu se zanašamo na parametrizacijo latentnega podatkovnega modela in predpostavljamo, da so nekateri parametri in hiperparametri souporabljeni v modelih različnih podatkovnih virov.


Profiliranje in veriženje
Predznanje v obliki mrež
Analiza preživetja z združevanjem podatkov
Faktorski model zlivanja podatkov
Gradnja mrež
Izbor modela pri zlivanju velikih heterogenih podatkovnih zbirk

Slika A.1

Organizacija doktorske disertacije.

Pristopi k prenosu informacij med sorodnimi podatkovnimi pogledi, povezanimi tipi objektov in učnimi parametri so usklajeni z vrstami podatkovne heterogenosti, ki jih obravnavamo v disertaciji. Pri modeliranju posameznih heterogenosti sledimo naslednjih trem korakom:

- *FAZA 1 - Raziskava:* Tu zastavimo vprašanje, ki se osredotoča na trenutne izzive v sistemski in molekularni biologiji, raziščemo uveljavljene in sorodne pristope ter oblikujemo delovno hipotezo. Zberemo podatke iz podatkovnih baz, ki hranijo meritve bioloških eksperimentov, in podatke s sodelujočih ustanov ter določimo eno ali več vrst podatkovne heterogenosti, ki jih želimo obravnavati tekom gradnje podatkovnih modelov.

- *FAZA 2 - Modeliranje:* Tu razvijemo računske modele, ki nam služijo za gradnjo napovedi in ocenjevanje verjetnosti povezano z zastavljenim vprašanjem. Preizkusimo našo hipotezo in opravimo dodatne analize podatkov.

- *FAZA 3 - Algoritmi:* Predstavimo *splošne* algoritme za zlivanje podatkov, empirično ovrednotimo njihovo zmogljivost in učinkovitost ter jih primerjamo z uveljavljenimi pristopi. Naše napovedi partnerji s sodelujočih ustanov preverijo z biološkimi eksperimenti, če to dopušča zastavljeno vprašanje.

V disertaciji preučujemo šest smeri, kjer pokažemo, da lahko z načelnimi pristopi za zlivanje podatkov izboljšamo kakovost zgrajenih napovednih modelov. Obravnavane smeri našega dela so prikazane na karti doktorske disertacije na Sliki A.1.

Disertacija se tako naravno organizira v šest delov, kot je prikazano v Tabeli 1: vrstice ustrezajo zastavljenim raziskovalnim vprašanjem in stolpci predstavljajo prej opisane vrste podatkovne heterogenosti, ki so obravnavane v pripadajočih delih disertacije.

Tabela 1

Različne vrste podatkovne heterogenosti in deli doktorske disertacije, ki jih naslavljajo.

Del	disertacije	Vrste heterogenosti		
		Relacija	Tip objektov	Naloga
Del I	Predznanje v obliki mrež	✓		✓
Del II	Gradnja mrež	✓		
Del III	Faktorski model zlivanja podatkov	✓	✓	✓
Del IV	Profiliranje in veriženje		✓	
Del V	Analiza preživetja z združevanjem podatkov		✓	✓
Del VI	Izbor modela pri zlivanju velikih zbirk	Odvisnosti med heterogenostmi		

A.2 Latentni faktorski modeli

V tem razdelku orišemo latentne faktorske modele, ki so namenjeni obravnavi *posameznih podatkovnih matrik.* Ti modeli so temeljni gradnik pristopov zlivanja podatkov, ki jih obravnava pričujoča disertacija.

Naj bo dana podatkovna tabela predstavljena z matriko $\boldsymbol{X} \in \mathbb{R}^{n\times m}$, ki jo želimo aproksimirati s produktom dveh matrik $\boldsymbol{U}\boldsymbol{V}^T$, kjer $\boldsymbol{U} \in \mathbb{R}^{n\times k}$ in $\boldsymbol{V} \in \mathbb{R}^{m\times k}$. Če na vrstice matrike $\boldsymbol{X}$ gledamo kot na podatkovne vektorje $\boldsymbol{X}_i$, potem vsak tak vektor predstavimo z linearno kombinacijo $\boldsymbol{U}_i\boldsymbol{V}^T$ vrstic v matriki $\boldsymbol{V}^T$. O vrsticah v $\boldsymbol{V}^T$ lahko razmišljamo kot o *latentnih faktorjih* in o elementih v $\boldsymbol{U}$ kot *utežeh* te linearne kombinacije. Z geometrijskega vidika so vektorji $\boldsymbol{U}_i \in \mathbb{R}^m$ predstavljeni s k-razsežnim linearnim podprostorom, ki ga razpenjajo vrstice $\boldsymbol{V}^T$. Velja tudi obratno: stolpce v matriki $\boldsymbol{X}$ je možno razumeti kot linearne kombinacije stolpcev matrike $\boldsymbol{U}$. Matriki $\boldsymbol{U}$ in $\boldsymbol{V}$ sta pogosto označeni kot *latentni matriki* ali *matriki latentnih faktorjev*.

V kolikor ne zahtevamo, da matriki $\boldsymbol{U}$ in $\boldsymbol{V}$ zadoščata dodatnim omejitvam, torej sta lahko poljubni realni matriki ustreznih razsežnosti, potem so matrike, ki jih je možno zapisati s produktom $\widehat{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{V}^T$, natanko tiste, katerih rang je omejen s k. To pomeni, da je faktorizacija matrike $\boldsymbol{X}$ brez dodatnih omejitev enakovredna k-razsežni aproksimaciji matrike $\boldsymbol{X}$.

Zgornji opis se namenoma ne ukvarja z razumevanjem pojma "aproksimacije" podatkovne matrike. V kakšnem smislu želimo aproksimirati podatke? Nadaljnje, kako merimo neskladje med podatki $\boldsymbol{X}$ in modelom $\widehat{\boldsymbol{X}}$, ki ga želimo zgraditi? Ali lahko na "aproksimacijo" gledamo kot na gradnjo primernega verjetnostnega modela?

Faktorizacija brez omejitev. Najpogosteje uporabljeno merilo neskladja med podatki $\boldsymbol{X}$ in modelom $\widehat{\boldsymbol{X}}$ je Frobeniusova razdalja med $\boldsymbol{X}$ in $\widehat{\boldsymbol{X}}$:

$$\|\boldsymbol{X} - \widehat{\boldsymbol{X}}\|^2_{\text{Fro}} = \sum_{i,j}(\boldsymbol{X}_{ij} - \widehat{\boldsymbol{X}}_{ij})^2. \tag{1}$$

Matrične faktorizacije, ki izhajajo iz optimizacije Frobeniusove razdalje, se pogosto uporabljajo zaradi enostavnosti njihovega izračuna. Izkaže se namreč, da je matrika $\widehat{\boldsymbol{X}}$ ranga k, ki minimizira vsoto razlike kvadratov do matrike $\boldsymbol{X}$, dana s k glavnimi komponentami v singularnem razcepu matrike $\boldsymbol{X}$ (Jolliffe, 2002).

Faktorizacija z omejitvami. V dosedanjem opisu smo se osredotočili na faktorizacije brez dodatnih omejitev, kjer lahko matriki $\boldsymbol{U}$ in $\boldsymbol{V}$ zavzameta kateri koli matriki iz prostorov $\mathbb{R}^{n\times k}$ oziroma $\mathbb{R}^{m\times k}$. To pomeni, da je model $\hat{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{V}^T$ omejen le z matričnim rangom. V analizi podatkov pogosto želimo, da latentne matrike zadoščajo dodatnim omejitvam, kar dosežemo z uvedbo regularizacije v cenitveni funkciji. Regularizacija latentnega modela omogoča lažjo interpretacijo latentnih faktorjev, zmanjša prostor vseh možnih rešitev in dovoljuje obstoj večih latentnih matrik, ki so z vidika cenitvene funkcije enake kakovosti. V splošnem vpeljava omejitev zmanjša število prostostnih stopenj razcepa $\boldsymbol{U}\boldsymbol{V}^T$. Lee and Seung (2000) sta raziskovala različne omejitve faktorskih matrik, vključno z zelo razširjeno omejitvijo, ki določa, da morata latentni matriki vsebovati le nenegativne vrednosti. Celovit pregled različnih vrst regularizacije, kot so nenegativnost, ortogonalnost, stohastičnost, redkost in ohranitev topoloških lastnosti med različnimi omejitvami, so podani v Žitnik and Zupan (2012); Wang and Zhang (2013).

Enotni pogled na matrično faktorizacijo. Singh and Gordon (2008b) sta nedavno predstavila formalno ogrodje za matrično faktorizacijo, ki omogoča opis zelo različnih vrst faktorizacije s spreminjanjem majhnega števila modelnih parametrov. To ogrodje obsega razširjene metode, kot so nenegativna matrična faktorizacija (Lee and Seung, 2000), uteženi singularni razcep (Srebro et al., 2003), eksponentna analiza glavnih komponent (Collins et al., 2001), matrična faktorizacija z velikim robom (Srebro et al., 2004), verjetnostni model latentnega semantičnega indeksiranja (Hofmann, 1999), Bregmanovo sočasno razvrščanje (Gordon, 2002) in številne druge.

Definicija 1: Matrično faktorizacijo lahko opredelimo z izbiro naslednjih možnosti, ki so dovolj splošne, da obsegajo številne pogosto rabljene matrične razcepe v analizi podatkov:

1. Podatkovne uteži $\boldsymbol{W} \in \mathbb{R}_+^{m\times n}$.
2. Funkcija preslikave $f : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$.
3. Omejitve latentnih matrik, $\boldsymbol{U}, \boldsymbol{V} \in \mathscr{C}$.
4. Utežena mera napake med $\boldsymbol{X}$ in $\hat{\boldsymbol{X}} = f(\boldsymbol{U}\boldsymbol{V}^T)$, $\mathscr{D}(\boldsymbol{X}\|\hat{\boldsymbol{X}}, \boldsymbol{W}) \geq 0$.

5. Parametri regularizacije, $\mathscr{R}(\boldsymbol{U}, \boldsymbol{V}) \geq 0$.

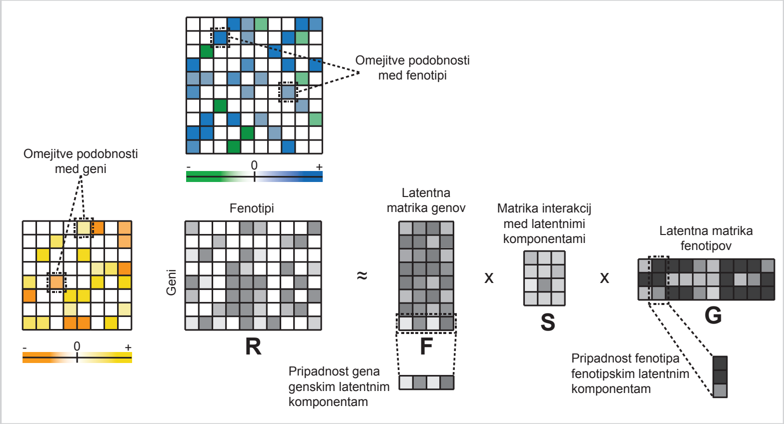
Z danimi izbirami zgornjih možnosti poiščemo latentni model $\boldsymbol{X} \approx f(\boldsymbol{U}\boldsymbol{V}^T)$ z reševanjem ustrezne optimizacijske naloge:

$$\underset{U,V \in \mathscr{C}}{\arg\min}\, \mathscr{D}(\boldsymbol{X} \| f(\boldsymbol{U}\boldsymbol{V}^T), \boldsymbol{W}) + \mathscr{R}(\boldsymbol{U}, \boldsymbol{V}). \qquad (2)$$

Funkcija preslikave f omogoča modeliranje nelinearnih odvisnosti med modelom $\boldsymbol{U}\boldsymbol{V}^T$ in podatki $\boldsymbol{X}$ (Singh and Gordon, 2008b).

Matrično dopolnjevanje je koncept, ki je tesno povezan z matrično faktorizacijo. Cilj matričnega dopolnjevanja je rekonstrukcija podatkovne matrike, pri čemer imamo na vhodu na voljo le podmnožico njenih elementov (Todeschini et al., 2013; Lee and Shraibman, 2013). Problem matričnega dopolnjevanja se pojavi v priporočilnih sistemih, cf. Shi et al. (2012a). Pristopi, ki ga rešujejo, najpogosteje poiščejo matriko čim manjše kompleksnosti, ki se ujema s podatki na vhodu. Tu se kompleksnost matrike običajno meri z velikostjo ranga ali ocenjevanjem norme sledi matrik. Ti računski pristopi so dobro raziskani ob predpostavki, da je množica matričnih vrednosti na vhodu vzorčena naključno enakomerno (Candès and Recht, 2009; Candès and Tao, 2010).

Matrična tri-faktorizacija z omejitvami. Mnogo podatkovnih naborov ustreza diadičnim relacijam, ki opisujejo odnose med objeki dveh tipov. Tovrstne nabore algoritmi matrične faktorizacije predstavijo z relacijskimi oziroma omejitvenimi matrika. Relacijska matrika $\boldsymbol{R}_{ij}$ je realna matrika razsežnosti $n_i \times n_j$, v kateri vrstice ustrezajo objektom tipa i, stolpci predstavljajo objekte tipa j in element $\boldsymbol{R}_{ij}(k, l)$ opisuje razmerje med objektom k in l. Omejitvena matrika $\boldsymbol{\Theta}_i$ je realna matrika dimenzij $n_i \times n_i$, ki opisuje razmerja med objekti tipa i. Njene vrednosti kodirajo podobnosti/razlike med objekti. Cenitvena funkcija latentnega faktorskega modela je taka, da bolje ceni latentne matrike, ki zadoščajo omejitvam in dobro rekonstruirajo elemente matrik na vhodu (Slika A.2). Na primer, modeli matrične tri-faktorizacije razcepijo relacijsko matriko $\boldsymbol{R}_{ij}$ na tri latentne matrike, tako da $\boldsymbol{R}_{ij} \approx \boldsymbol{F}_{ij}\boldsymbol{S}_{ij}\boldsymbol{G}_{ij}^T$, kjer $\boldsymbol{F}_{ij} \in \mathbb{R}^{n_i \times k_{ij}}$, $\boldsymbol{S}_{ij} \in \mathbb{R}^{k_{ij} \times c_{ij}}$ in $\boldsymbol{G}_{ij} \in \mathbb{R}^{n_j \times c_{ij}}$. Parametra k_{ij} in c_{ij} predstavljata rang matrične faktorizacije in sta v analizi podatkov običajno bistveno manjša od razsežnosti vhodnih matrik, $k_{ij} \ll n_i$, $c_{ij} \ll n_j$. Matrična tri-faktorizacija predstavi profile, t.j. vrstične vektorje v $\boldsymbol{R}_{ij}$, z bistveno manj vektorji v $\boldsymbol{S}_{ij}$ in z nizko-razsežnimi vektorji $\boldsymbol{G}_i$ in $\boldsymbol{G}_j$. To pomeni, da je dobro rekonstrukcijo možno doseči le, če ti vektorji razpenjajo prostor, ki razkriva strukturo, ki je skrita, a lastna vhodnim podatkom (Fig. A.2). Slednja lastnost je ključna predpostavka, na katero se zanašajo pristopi latentnih faktorskih modelov v strojnem učenju in odkrivanju znanj iz podatkov.



Slika A.2

Matrična tri-faktorizacija z omejitvami. Poleg podatkovne matrike sta na voljo omejitveni matriki, ki izražata stopnjo podobnosti med geni (matrika z rumenimi in oranžnimi celicami) oziroma med fenotipi (matrika z modrimi in zelenimi celicami). Negativne vrednosti v omejitvenih matrikah nagrajujejo latentne matrike, v katerih imajo pripadajoči geni oziroma fenotipi podobne profile. Velja tudi obratno; večje pozitivne vrednosti kaznujejo latentne modele, v katerih imajo pripadajoči geni oziroma fenotipi podobno predstavitev.

Latentni faktorski modeli v analizi podatkov. Latentni faktorski modeli so se izkazali za zelo uspešne pri odkrivanju zapletenih struktur v visoko-dimenzionalnih podatkih in se zato uporabljajo na številnih domenah in raznih poslovnih, tehnoloških, znanstvenih in raziskovalnih področjih. Poleg izjemnega uspeha, ki so ga algoritmi matrične faktorizacija dosegli v priporočilnih sistemih (Bell and Koren, 2007), se te tehnike med drugim uspešno uporabljajo v pristopih za zmanjšanje dimenzij podatkov (Jolliffe, 2002; Li et al., 2009c; Maurus and Plant, 2014), razvrščanje (Hochreiter et al., 2010; Arora et al., 2013) in nizko-razsežno aproksimacijo podatkov (Matsushita and Tanaka, 2013).

Eden izmed načinov za merjenje prileganja faktorskega modela učnim podatkom je uporaba raznih metrik, kot je kvadratni koren povprečne kvadratne napake med vhodnimi meritvami in napovedanimi vrednostmi modela. Slednja metrika se je uporabljala za vrednotenje rešitev v izzivu Netflix Prize Contest (`http://www.netflixprize.`

com), ki je eden najpomembnejših katalizatorjev uporabe matričnih metod v zadnjem desetletju strojnega učenja. V zadnjem času se veliko matričnih pristopov osredotoča na naloge uvrščanja in rangiranja, kjer naivna uporaba metrik, kot je povprečna kvadratna napaka, ne vrača zadovoljivega rezultata (Rendle, 2010; Rendle et al., 2010). Na primer, v nalogah skupinskega filtriranja se uporabniki osredotočijo le na nekaj najbolj obetavnih priporočil. To pomeni, da mora biti razvoj računskih pristopov usmerjen na generiranje kakovostnega a kratkega seznama priporočenih izdelkov. Primerne metrike za optimizacijo in vrednotenje prileganja faktorskih modelov za danega uporabnika v tovrstnih primerih ocenjujejo *relevantnost seznama prvih-N izdelkov.* Mnogi nedavno predlagani algoritmi in faktorski modeli tako merijo neskladje med vhodnimi in napovedanimi vrednostmi z metrikami za vrednotenje rangiranja, uvrščanja in regresije (Rendle et al., 2009; Rendle, 2010; Shi et al., 2012b, 2013). Faktorski modeli se zato lahko uporabljajo ne le za regresijske naloge, kjer se primerni matrični produkt latentnih matrik neposredno uporablja za gradnjo napovedi in je merilo optimizacije kvadratna napaka, temveč tudi v dvorazrednem uvrščanju, kjer se parametri določijo z optimizacijo funkcij napak, kot sta hinge in logit (Rendle, 2010), in v nalogah rangiranja, kjer je optmizacija usmerjena v iskanje relevantnih seznamov objektov (Shi et al., 2013). Nedavni razvoj tovrstnih pristopov torej postavlja latentne faktorske modele v *skupino splošnih napovednih modelov*, ki lahko obravnavajo raznovrstne matrične predstavitve podatkov (Rendle, 2010). Ti modeli obravnavajo interakcije med spremenljivkami s pomočjo faktoriziranih parametrov in so sposobni robustnega ocenjevanja interakcij tudi v primerih, kjer so podatki zelo redki in so uveljavljene tehnike strojnega učenja manj uspešne (Rendle, 2013).

A.3 Predznanje podano z omrežji

Problem matričnega dopolnjevanja se pojavi v mnogih nalogah podatkovnega rudarjenja in je v zadnjih letih deležen izjemne pozornosti na področju gradnje priporočilnih sistemov, kjer so tovrstni algoritmi med najuspešnejšimi (Shi et al., 2012b). Ti algoritmi redko matriko na vhodu dopolnijo na način, da je kompleksnost dopolnjene matrike čim manjša in da se dopolnjena matrika dobro ujema z znanimi vrednostmi.

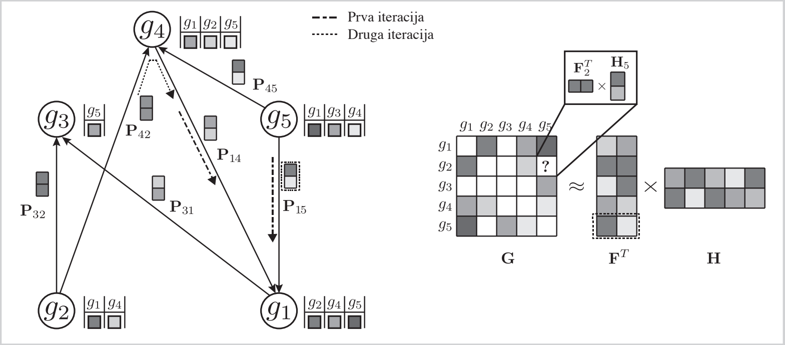
Razvili smo pristop matričnega dopolnjevanja na osnovi izmenjave podatkov med viri, ki upošteva predznanje podano v obliki omrežij (Slika A.3). Iterativni algoritem gradi verjetnostni latentni model in vključuje predznanje po principu tranzitivnosti. To pomeni, da so latentni profili objektov v vsaki iteraciji algoritma odvisni od profilov njihovih neposrednih sosedov v danih omrežjih. Iterativna narava algoritma omogoča širjenje vpliva med latentnimi profili objektov glede na njihove lokalne okolice v omrežjih.

Primer delovanja algoritma na meritvah petih objektov prikazuje Slika A.3. Naj bodo dani objekti $g_1 \dots g_5$. Znane vrednosti so prikazane ob pripadajočih vozliščih omrežja $\boldsymbol{P}$ in v matriki $\boldsymbol{G}$. Matriki $\boldsymbol{F}$ in $\boldsymbol{H}$ sta latentni matriki, ki ju želimo določiti. Profili objektov v $\boldsymbol{F}$ so v vsaki iteraciji algoritma odvisni od profilov njihovih sosedov v omrežju $\boldsymbol{P}$. Na primer, algoritem v prvi iteraciji pri posodobitvi vektorja $\boldsymbol{F}_{g_1}$ upošteva predstavitev njegovih sosedov g_4 in g_5 (profila $\boldsymbol{F}_{g_4}$ in $\boldsymbol{F}_{g_5}$ sta prikazana na vhodnih povezavah objekta g_1 na Sliki A.3), pri čemer je stopnja vpliva določena s $\boldsymbol{P}_{14}$ in $\boldsymbol{P}_{15}$. V drugi iteraciji se pri posodobitvi $\boldsymbol{F}_{g_1}$ upošteva profil objekta g_2 (Slika A.3).

Matrično dopolnjevanje, ki v gradnjo modela vključi predznanje, nam omogoča, da napovemo vrednosti za vrstice oziroma stolpce vhodne matrike, za katere sicer ni na voljo nobenih meritev. Ta zanimiv izziv je v literaturi priporočilnih sistemov znan kot *problem hladnega zagona* in ga je brez vključitve predznanja zelo težko ustrezno nasloviti.

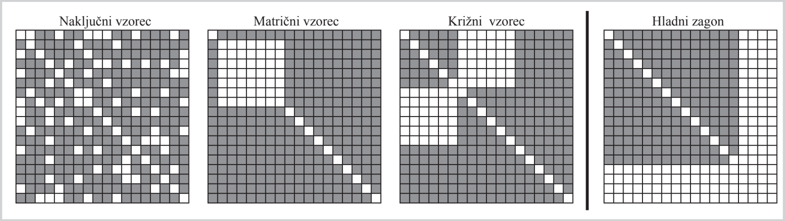
Slika A.3

Matrično dopolnjevanje s predznanjem podanim z omrežji. Dani so podatki E-MAP s petimi geni, $\{g_1, \dots, g_5\}$. Predznanje je podano z mrežo $\boldsymbol{P}$. Interakcijski profili genov so podani v matriki $\boldsymbol{G}$ (desno) in zraven pripadajočih vozlišč genov (levo). Algoritem faktorizira vhodno matriko $\boldsymbol{G}$ z latentnima faktorjema $\boldsymbol{F}$ in $\boldsymbol{H}$. Struktura njunih latentnih komponent je skladna z oddaljenostjo genov v dani mreži.



Vrednotenje razvitih metod. V disertaciji poročamo o več empiričnih eksperimentih, v katerih gradimo napovedne modele genskih interakcij v študijah epistatičnih mikromrež (E-MAP) (Wilmes et al., 2008). Predlagani algoritem primerjamo z večimi obstoječimi metodami za napovedovanje genskih interakcij in ga ovrednotimo na večih podatkovnih naborih in z različnimi omrežji, kot sta omrežje proteinskih interakcij in omrežje genskih pripisov. Izkaže se, da je matrično dopolnjevanje učinkovit pristop, ki se pri napovedovanju genskih interakcij obnese bolje od alternativnih tehnik. Zelo dobro obnašanje modela je mogoče razložiti z njegovo zmožnostjo vključitve dodatnih virov informacij in z naravo algoritma, ki upošteva tako *globalno kovariančno strukturo* kot tudi *lokalno izmenjavo latentnih profilov med sosednimi geni v omrežju.*

Preučili smo vpliv velikosti učne množice meritev in porazdelitve znanih vrednosti na napovedno točnost zgrajenih modelov. V realnih situacijah meritve pogosto ne vzorčijo domenskega prostora enakomerno naključno, kar pomeni, da manjkajoče vrednosti sledijo vzorcu, ki je posledica tehnoloških ali domenskih omejitev (Slika A.4). Rezultati empiričnih raziskav kažejo, da je obravnava večih virov informacij vselej koristna. Še posebej dobro se modeli s predznanjem obnesejo v situacijah, v katerih manjkajoče meritve sledijo netrivialnim vzorcem prikazanim na Sliki A.4.



Slika A.4

Porazdelitev manjkajočih vrednosti v podatkih. Manjkajoče meritve so lahko porazdeljene enakomerno naključno (*Naključni vzorec*). Alternativno lahko manjkajo vse meritve interakcij znotraj množice objektov (*Matrični vzorec*), med dvema disjunktnima množicama objektov (*Križni vzorec*) ali celotni meritveni profili (*Hladni zagon*).

A.4 Sočasna gradnja mrež iz večih virov

V disertaciji smo razvili dva pristopa za gradnjo mrež, ki temeljita na izmenjavi latentnih informacij med podatkovnimi viri.

Gradnja genskih mrež na osnovi analize epistaze

Pristopi h gradnji genskih mrež nas zanimajo v smislu napovedovanja vrstnega reda delovanja genov, to je, njihove urejenosti v bioloških poteh, o čemer lahko sklepamo iz fenotipskih podatkov enojnih in dvojnih mutant. *Analiza epistaze* je princip znan v klasični genetiki, ki ocenjuje vpliv in urejenost dveh genov na osnovi meritev njunih fenotipov. Fenotip je najpogosteje podan z oceno fitnesa, to je, sposobnost organizma, da se razvija in raste, ali z oceno o izraženosti izbranega gena. Analiza epistaze primerja fenotip dvojne mutante s fenotipom ustreznih enojnih mutant in oceni, kateri izmed pripadajočih genov deluje v genski poti bližje izhodnemu signalu (Roth et al., 2009). Epistaza ne omogoča le sklepanja o linearni urejenosti genov in o soodvisnosti njihovih vlog v celici, ampak je koristna tudi za odkrivanje delnih odvisnosti in razkrivanje vzporednih bioloških poti. Razkrivanje neposrednih funkcijskih odvisnosti med geni in razlaga vzročno-posledičnih razmerij sta ključni lastnosti, v katerih se analiza epistaze razlikuje od drugih pristopov h gradnji genskih mrež, ki temeljijo na računanju podobnosti med profili genskih interakcij in lahko o odvisnostih med geni sklepajo le posredno (Costanzo et al., 2010; Mostafavi and Morris, 2012).

Razvili smo algoritem Réd za gradnjo velikih genskih mrež, ki so skladne z epistatičnimi razmerji genov. Algoritem gradi napovedni model na osnovi fenotipskih meritev enojnih in dvojnih mutant in je zaradi faktorizacije modelnih parametrov primeren za šumne in redke podatke. Pristop se konceptualno razlikuje od obstoječih tehnik, saj sočasno gradi nelinearni verjetnostni model za vse pare genov in vse možne funkcijske odvisnosti, to je, linearno, vzporedno ali vzporedno-odvisno delovanje genov. Latentni podatkovni model služi za izračun ocen verjetnosti različnih odvisnosti med geni in za izgradnjo genske mreže.

Vrednotenje razvitih metod. Zmogljivost predlaganega algoritma Réd vrednotimo na večih podatkovnih naborih (Jonikas et al., 2009; Costanzo et al., 2010; Surma et al., 2013), tako da merimo ploščino pod krivuljo ROC in rekonstrukcijsko napako med napovedanimi genskimi mrežami ter referenčnimi mrežami oziroma znanimi odvisnostmi med geni. Na področju gradnje genskih mrež z analizo epistaze obstaja le nekaj algoritmov (Battle et al., 2010), ki lahko obravnavajo nabore z meritvami nekaj sto mutant. Réd je *računsko učinkovit* in za izgradnjo napovednega modela vseh mutant v kvasovki z več tisoč geni in sto tisoči meritev potrebuje le nekaj minut na osebnem računalniku. Znani pristopi niso primerni za analizo podatkov takih razsežnosti. Poleg računske učinkovitosti Réd rekonstruira genske mreže s presenetljivo visoko preciznostjo in dosega točnost, ki je vsaj primerljiva, a večinoma boljša od uveljavljenih tehnik.

Gradnja mrež iz večih virov in raznovrstnih podatkovnih porazdelitev

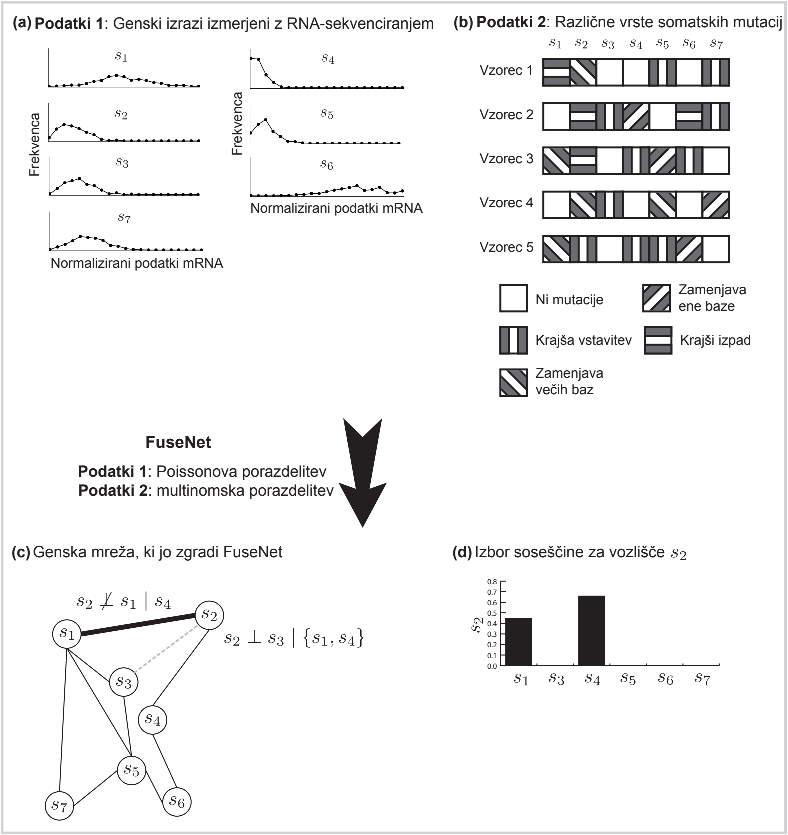
Markovska mreže so neusmerjena grafični modeli, ki se pogosto uporabljajo za odkrivanje kompleksnih odnosov med objekti iz meritev o njihovem delovanju (Rue and Held, 2005). Uveljavljeni postopki za gradnjo markovskih mrež tipično temeljijo na analizi podatkov, ki sledijo Gaussovi porazdelitvi (Friedman et al., 2008; Ravikumar et al., 2010). Obsežni podatki pridobljeni z visoko-prepustnimi tehnologijami, kot so tehnike sekvenciranja RNA v molekularni biologiji, sledijo raznovrstnim podatkovnim porazdelitvam in pogosto kršijo predpostavke Gaussove porazdelitve. Ko je za dane objekte na voljo več virov podatkov, ki izhajajo iz različnih porazdelitev, je verjetno, da imajo markovske mreže, zgrajene nad različnimi viri, določene skupne strukturne lastnosti.

V disertaciji se ukvarjamo z novim statističnim pristopom, ki lahko sočasno obravnava heterogene zbirke podatkov, kjer heterogenost izhaja iz raznolikosti podatkovnih porazdelitev. V ta namen smo razvili algoritem FuseNet, ki gradi markovske mreže iz večih morebitno različno porazdeljenih podatkovnih virov. FuseNet je *računsko učinkovit* in *splošen* pristop, ki lahko sočasno obravnava več virov opisanih z različnimi porazdelitvami iz *eksponentne družine*. FuseNet parametrizira napovedni model s faktoriziranimi parametri, ki souporabljajo latentne faktorje, le ti pa opredeljujejo soseščino vozlišč v zgrajeni mreži (Slika A.5).

Vrednotenje razvitih metod. V empiričnih študijah pokažemo dobro napovedno točnost pristopa FuseNet v primerjavi z več uveljavljenimi neusmerjenimi grafičnimi modeli (Allen and Liu, 2013; Friedman et al., 2007a; Gallopin et al., 2013; Liu et al., 2009). Učinkovitost pristopa raziščemo z analizo podatkov RNA-sekvenciranja in podatkov o somatskih mutacijah vzorcev rakavega tkiva v International Cancer Genome Consortium, kar je nova uporaba neusmerjenih grafičnih modelov. Zlivanje virov znatno izboljša točnost zgrajenih mrež in stopnjo njihove biološke obogatenosti v

Slika A.5

Metoda FuseNet za gradnjo genskih mrež. FuseNet avtomatično zgradi markovsko mrežo s hkratno obravnavo večih podatkovnih naborov, pri čemer lahko nabori sledijo različnim porazdelitvam iz eksponentne družine. Prikazan je primer z dvema naboroma: (a) meritve genskih izrazov, ki sledijo Poissonovi porazdelitvi in (b) podatki somatskih mutacij, ki jih je možno modelirati z multinomsko porazdelitvijo. (c) FuseNet zgradi gensko mrežo s hkratnim učenjem odvisnosti med geni in v kontekstu vseh meritev genske izraženosti in mutacijskih profilov. Odsotnost povezave med s_2 and s_3 (prekinjena črta v sivem) nakazuje, da gen s_2 deluje neodvisno od gena s_3 pri znanem delovanju genov s_1 in s_4, ki sta neposredna soseda gena s_2 v mreži. Simbol ⊥ predstavlja pogojno neodvisnost. Gena s_1 in s_2 sta povezana, ker genski profili s_2 v (a-b) nosijo informacijo o delovanju gena s_1 pri znanem s_4, neposrednim sosedom s_2. (d) Prikazani so koeficienti odvisnosti med s_2 in vsemi ostalimi geni v sistemu. Neničelne vrednosti nakazujejo odvisnost v delovanju pripadajočih genov. Iz slike (d) izhaja, da ima gen s_2 dva soseda v mreži, s_1 in s_4.


(a) Podatki 1: Genski izrazi izmerjeni z RNA-sekvenciranjem
s_1
s_2
s_3
s_7
s_4
s_5
s_6
Frekvenca
Frekvenca
Normalizirani podatki mRNA
Normalizirani podatki mRNA
(b) Podatki 2: Različne vrste somatskih mutacij
s_1 s_2 s_3 s_4 s_5 s_6 s_7
Vzorec 1
Vzorec 2
Vzorec 3
Vzorec 4
Vzorec 5
Ni mutacije
Zamenjava ene baze
Krajša vstavitev
Krajši izpad
Zamenjava večih baz
FuseNet
Podatki 1: Poissonova porazdelitev
Podatki 2: multinomska porazdelitev
(c) Genska mreža, ki jo zgradi FuseNet
$s_2 \not\perp s_1 \mid s_4$
$s_2 \perp s_3 \mid \{s_1, s_4\}$
s_1
s_2
s_3
s_4
s_5
s_6
s_7
(d) Izbor soseščine za vozlišče s_2
s_2
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
s_1 s_3 s_4 s_5 s_6 s_7

primerjavi z mrežami, ki so zgrajene ločeno in nedvisno nad posameznim virom podatkov. Naši rezultati tudi kažejo, da lahko metode za gradnjo markovskih mrež iz ne-Gaussovih porazdelitev izboljšajo modeliranje podatkov, pridobljenih z nastajajočimi visoko-prepustnimi tehnologijami v sistemski biologiji.

A.5 Večrelacijski in večtipni faktorski model zlivanja podatkov

Algoritmi matrične faktorizacije razcepijo podatkovno matriko v več latentnih matričnih faktorjev nižjega ranga, ki jih poiščemo z reševanjem ustrezne optimizacijske naloge. Čeprav se algoritmi, ki temeljijo na latentnih modelih s faktoriziranimi parametri, uspešno uporabljajo v podatkovni analizi za raznovrstne naloge, kot so manjšanje dimenzij v visoko-dimenzionalnih podatkih, gručenje in kompaktna predstavitev matrik, so pristopi zlivanja virov, ki temeljijo na latentnih faktorskih modelih, maloštevilni.

Predlagali smo algoritem DFMF za sočasno matrično tri-faktorizacijo z omejitvami, ki omogoča hkratni razcep načeloma poljubnega števila podatkovnih matrik v produkte treh razcepnih matričnih faktorjev. Prednost pristopa je, da lahko obravnava matrike, ki opisujejo različne tipe objektov (na primer gene, bolezni, zdravila in kemikalije) Pomembno je, da so latentni faktorji *deljeni med razcepi matrik*, ki opisujejo objekte istega tipa (Slika A.6), kar omogoča sočasno obravnavo večih podatkovnih virov. Algoritem predstavi vsak podatkovni nabor z matriko, pri čemer razlikuje med omejitvenimi matrikami, ki opisujejo relacije med objekti istega tipa, na primer interakcije med geni, in relacijskimi matrikami, ki opisujejo razmerja med objekti različnih tipov, na primer pripisi konceptov hierarhije MeSH (angl. Medical Subject Headings) znanstvenim člankom. Pristop sestoji iz treh glavnih korakov:

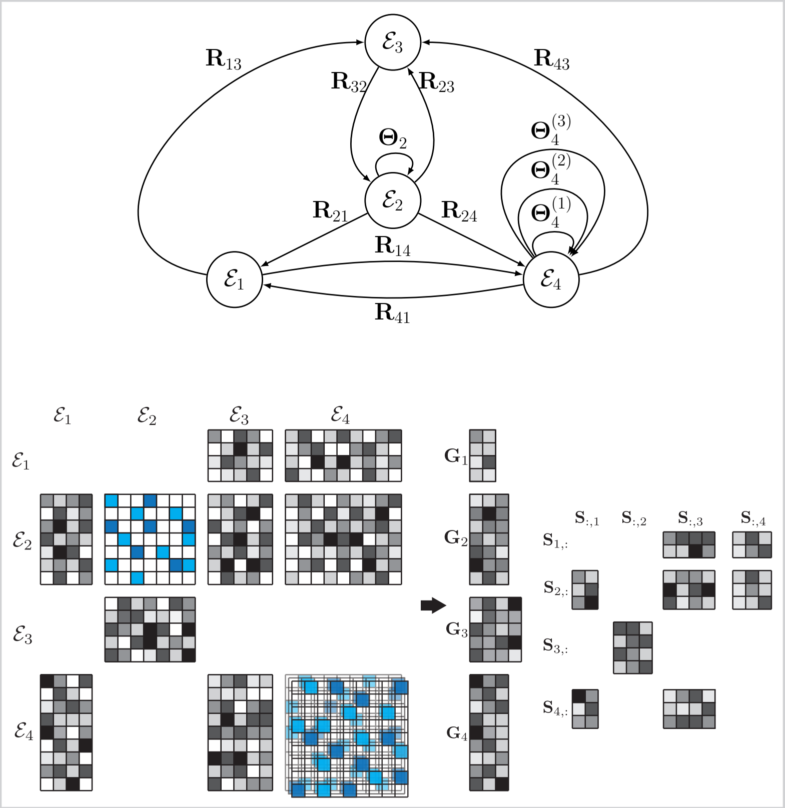
1. Določitev podatkovnih virov za zlivanje in njihova organizacija v *graf zlivanja*. Graf zlivanja na Sliki A.6 prikazuje shemo enajstih podatkovnih virov med štirimi tipi objektov.

2. Sočasna matrična tri-faktorizacija vseh relacijskih matrik, pri čemer omejitvene matrike služijo za regularizacijo razcepnih matričnih faktorjev. Ključni korak zlivanja je *souporaba* latentnih faktorjev v razcepih sorodnih relacijskih matrik (Slika A.6, spodaj).

3. Uporaba zgrajenega latentnega modela za gradnjo napovedi v nalogah, kot so:

Slika A.6

Delovanje faktorskega modela za zlivanje podatkov na primeru štirih tipov objektov, $\mathscr{E}_1, \mathscr{E}_2, \mathscr{E}_3$ in $\mathscr{E}_4$. Podatkovne nabore, ki opisujejo različne tipe podatkov, prikažemo z grafom relacij med tipi objektov (zgoraj) ali z enakovredno bločno matrično predstavitvijo (spodaj). Faktorski model za sočasno učenje predpostavlja, da dani podatkovni nabor opisuje odnose med dvema objektnima tipoma. Viri so prikazani s povezavami v grafu (zgoraj) oziroma s sivinskimi matrikami v spodnji bločni predstavitvi. Na primer, podatkovna matrika $\boldsymbol{R}_{23}$ opisuje odnose med objekti $\mathscr{E}_2$ in $\mathscr{E}_3$. Nekatere relacije so lahko odsotne. Na primer, v dani shemi ne obstaja nabor podatkov, ki bi vzpostavil odnose med objekti $\mathscr{E}_3$ in $\mathscr{E}_1$, zaradi česar v grafu ni usmerjene povezave med $\mathscr{E}_3$ in $\mathscr{E}_1$ oziroma enakovredno, matrika $\boldsymbol{R}_{31}$ ni na voljo. Omejitvene matrike so prikazane z zankami (zgoraj) oziroma matrikami z modrimi celicami (spodaj). Omejitvene matrike predstavljajo podatkovne nabore, ki opisujejo odnose med objekti istega tipa. Dani primer vsebuje omejitve za $\mathscr{E}_2$ (ena omejitvena matrika) in $\mathscr{E}_4$ (tri omejitvene matrike).



- rekonstrukcija relacijskih matrik z namenom dopolnitve njihovih manjkajočih vrednosti,
- veriženje razcepnih matričnih faktorjev vzdolž poti v grafu zlivanja,
- sočasno razvrščanje objektov različnih tipov v skupine na osnovi njihove pripadnosti latentnim komponentam.

Vrednotenje razvitih metod. Zmogljivost razvitega pristopa primerjamo s *pristopi zgodnjega združevanja virov*, kot so naključni gozdovi (Breiman, 2001; Chen and Zhang, 2013), *metodami poznega združevanja*, kot je zlaganje napovedi obstoječih učnih algoritmov (Pandey et al., 2010), in s *tehnikami vmesnega združevanja*, kot so metode učenja z večimi jedri (Gönen and Alpaydın, 2011; Yu et al., 2012). Naš pristop primerjamo z večimi algoritmi matrične faktorizacije v smislu njihove napovedne moči, časovne zahtevnosti in uporabnosti. Ti pristopi vključujejo enostavne dvo-razcepne matrične faktorizacije (Zhang et al., 2011b), ki obravnavajo diadične relacije, in tri-razcepne matrične faktorizacije za obravnavo večih diadičnih relacij (Wang et al., 2008, 2011a).

Metode ovrednotimo, tako da zlivamo več deset podatkovnih naborov iz molekularne biologije, kot so genske interakcije, genski pripisi, podatki o izrazih mRNA, metilacijski in mutacijski profili, metabolična omrežja, genske poti in podatki o celičnem signaliziranju. Pri tem povezujemo objekte različnih tipov, kot so geni, zdravila, bolezni, fenotipi, pacienti.

Zanimajo nas aktualni problemi v molekularni in sistemski biologiji:

- Napovedovanje funkcij genov in proteinov v večih modelnih organizmih z zlivanjem več deset podatkovnih virov, med drugim genske izraze, omrežja proteinskih interakcij, znane genske pripise, podatke o vključenosti genov v presnovne poti ter izvlečke iz znanstvenih objav. Genske funkcije so opredeljene z ontološkimi koncepti v Gene Ontology (Ashburner et al., 2000), *Dictyostelium* Phenotype Ontology (Fey et al., 2009), Disease Ontology (Schriml et al., 2012) in Yeast Genome Database (Güldener et al., 2005).
- Napovedovanje farmakoloških akcij kemikalij, pri čemer farmakološke akcije ustrezajo konceptom ustrezne MeSH podhierarhije.

- Odkrivanje povezav med boleznimi z zlivanjem več kot desetih molekularnih podatkovnih virov.
- Napovedovanje toksičnosti zdravil z namenom zgodnjega odkrivanja stranskih učinkov zdravil na delovanje jeter s sočasno obravnavo skoraj trideset podatkovnih virov. Ta problem ni zanimiv le z raziskovalnega vidika, temveč je tudi izrednega pomena za zmanjšanje nezaželenih učinkov zdravil in stroškov razvoja, ki je posledica pozno ugotovljene toksičnosti zdravil.
- Rangiranje (prioritizacija) genov glede na oceno verjetnosti njihove vpletenosti v izbrani biološki proces. Biološki procesi med drugim vključujejo raziskavo bakterijske rezistence v amebi *Dictyostelium* in bolezni mrežnice pri človeku.

Rezultati empiričnih raziskav kažejo, da predlagani pristop DFMF dosega primerljive ali višjo točnost od uveljavljenih pristopov, ki gradijo napovedne modele z združevanjem podatkovnih virov. Prav tako pristop v večih empiričnih študijah napovedovanja genskih funkcij, farmakoloških akcij in toksičnosti zdravil, bistevno izboljša zmogljivost modelov, zgrajenih nad enim samim podatkovnim virom. To spoznanje je pomembno, saj kaže na prednosti, ki jih ima učenje z zlivanjem podatkov pred metodami za ločeno analizo posameznih podatkovnih naborov.

Poleg tega ima predlagani pristop nekaj zaželenih lastnosti, zaradi katerih je *uporaben v raznovrstnih napovednih nalogah*, dosega večjo *fleksibilnost* kot znane tehnike in je enostaven za uporabo. Algoritem DFMF za sočasno matrično faktorizacijo namreč ohranja *relacijsko strukturo podatkov* in lahko obravnava *heterogene podatkovne predstavitve brez njihove predhodne transformacije v enotni podatkovni prostor*. Ta zaželena lastnost omogoča nadaljnjo analizo objektov katerega koli tipa vključenega v zlivanje, pri čemer se izkorišča bogata latentna predstavitev celotne zbirke virov.

A.6 Profiliranje in veriženje

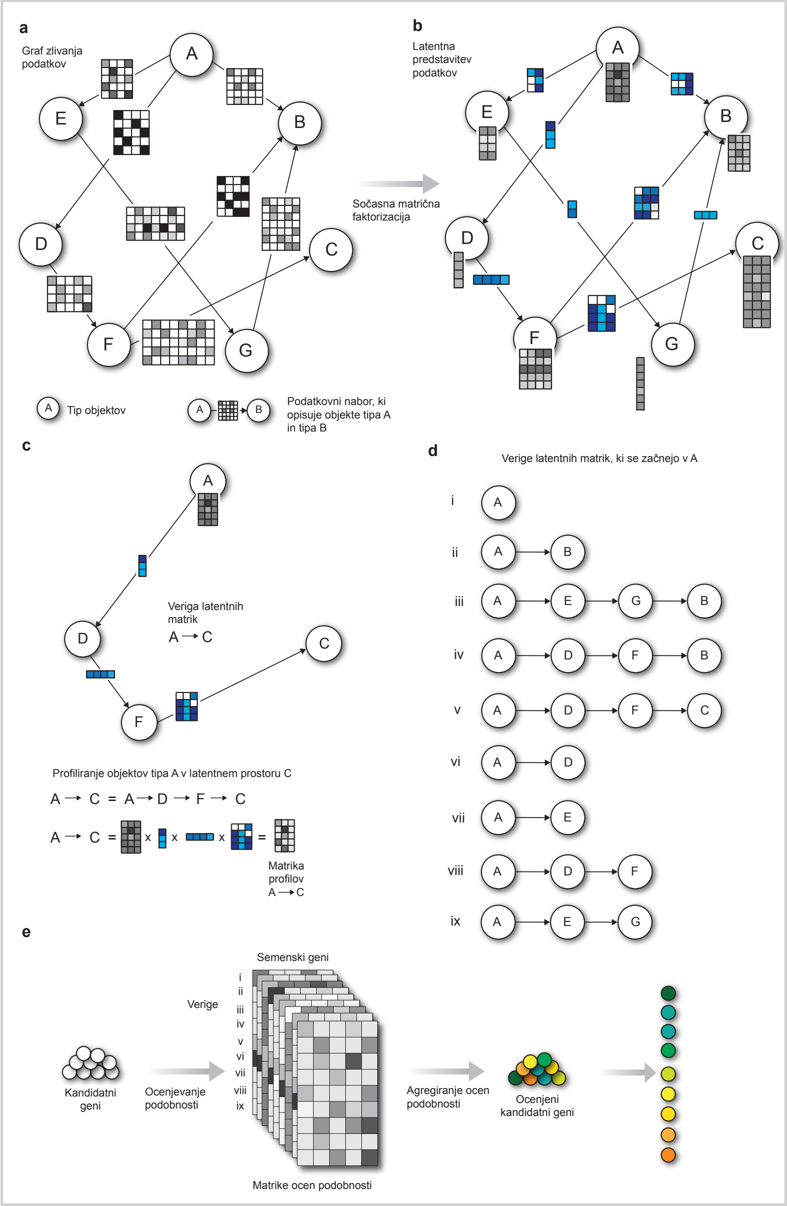
Algoritmi zlivanja virov s pristopi matrične faktorizacije opisani v prejšnjem razdelku poiščejo latentno podatkovno predstavitev celotne zbirke podatkov. Ta latentni prostor ohranja bogato relacijsko strukturo raznih virov in tipov objektov, ki jih definira graf zlivanja. Latentna predstavitev podatkov ponuja veliko priložnosti za gradnjo napovedi. Verjetno najbolj naravna in pogosta raba je dopolnjevanje relacijskih matrik, ki jo dosežemo z matričnim množenjem primernih latentnih matrik. V disertaciji poleg omenjenega dopolnjevanja raziščemo še nekaj možnosti, kot so uporaba latentnih matrik za razvrščanje objektov izbranega tipa, na primer genov, in sočasno razvrščanje objektov večih tipov, na primer hkratno gručenje genov in bolezni.

Da bi izkoristili relacijsko strukturo latentnega prostora (Slika A.6, spodaj desno), v disertaciji predlagamo nov način profiliranja objektov, imenovan *veriženje.* Veriženje latentnih matrik poteka vzdolž poti v grafu zlivanja in služi izpeljavi vektorjev značilk, ki so primerni za nadaljnjo analizo z uveljavljenimi algoritmi strojnega učenja.

Vrednotenje razvitih metod. Veriženje je osrednji sestavni element pristopa Collage, ki je predstavljen v disertaciji in služi prioritizaciji genov (Slika A.7). S predlaganim pristopom smo napovedali nekaj genov amebe *D. discoideum*, ki imajo lahko pomembno vlogo v bakterijski rezistenci in pred tem niso bili povezani s to funkcijo. Ameba je pomemben modelni organizem, ki se hrani z bakterijami, a je pogosto tudi njihova žrtev. Boljše razumevanje amebinega odziva v okolju z raznovrstnimi bakterijami, tudi takimi, ki so človeku nevarne in postajajo vse bolj odporne na razvite antibiotike, je pomembno za okužbe pri ljudeh. Do sedaj je bila znana le peščica genov, vpletenih v poti amebine bakterijske rezistence, ki so v naši študiji imeli vlogo *semenskih genov* zoper katerih smo ocenjevali *obetavnost* kandidatnih genov. Obetavnost izbranega kandidata smo merili z ocenjevanjem podobnosti genskih profilov izvedenih s postopkom veriženja (Slika A.7, cde). Naše napovedi osmih novih kandidatnih genov so bile eksperimentalno potrjene na sodelujoči instituciji (Baylor College of Medicine, Houston, ZDA). Razširitev seznama genov genov povezanih z razpoznavo bakterij ni le ključna v raziskavah mehanizmov bakterijske rezistence, temveč lahko prispeva pri snovanju alternativnih metod antibakterijskega zdravljenja.

A.7 Analiza preživetja z združevanjem podatkov

V mnogih pristopih analize podatkov je možno izboljšati kakovost zgrajenih modelov z združevanjem neposredno ali posredno povezanih virov. V disertaciji predlagamo razširitev algoritma za sočasno matrično tri-faktorizacijo DFMF, tako da lahko gradimo latentni model z zlivanjem podatkov in hkrati ocenjujemo parametre regresijskega modela za analizo preživetja. Novi pristop, imenovan DFMF-SR, je latentni faktorski



Slika A.7

Collage, predlagana metoda za gensko prioritizacijo. Algoritem vsebuje štiri pomembne korake: (a,b) izračun latentne predstavitve podatkov s sočasno matrično faktorizacijo, (c,d) profiliranje genov z veriženjem latentnih matrik, (e) ocenjevanje podobnosti in rangiranje genov na osnovi agregiranih vrednosti.

model, ki princip izmenjave latentnih matrik združi z Aalenovim aditivnim modelom analize preživetja (Aalen, 1989; Abadi et al., 2011).

Vrednotenje razvitih metod. Predlagani pristop ovrednotimo na heterogenih meritvah vzorcev rakavega in zdravega tkiva v zbirki podatkov iz International Cancer Genome Consortium. Obdobje preživetja med ugotovljeno diagnozo in dogodkom modeliramo kot funkcijo genskih izrazov, izraženosti proteinov in molekul miRNA, podatkov o metiliranih regijah na genomu ter profilov somatskih mutacij. Empirične študije kažejo, da sta tako izmenjava latentnih matrik in analiza večih virov kot tudi sočasno ocenjevanje regresijskega modela preživetja ključnega pomena za gradnjo zmogljivih napovednih modelov. Pristop DFMF-SR gradi bistveno bolj točne napovedi kot izhodiščni Aalenov aditivni model. Izkaže se tudi, da so najbolj informativni latentni faktorji statistično značilno povezani z biološkimi procesi povezanimi z razvojem rakavih obolenj.

A.8 Izbor modela pri zlivanju velikih heterogenih podatkovnih zbirk

Z združevanjem več deset podatkovnih virov nastopijo novi izzivi, eden izmed njih je problem izbire podatkovnih virov, ki naj bodo vključeni v zliti latentni model. Gre za posplošitev znanega problema izbora informativnih značilk v danem podatkovnem naboru, pri čemer nas pri zlivanju podatkov zanima, kateri so informativni podatkovni viri v dani zbirki virov. V ta namen smo se poslužili ocenjevanja občutljivosti latentnih matrik na vključitev novega vira v obstoječo zbirko virov.

Predlagali smo *računsko učinkovit* algoritem Forensic, ki temelji na tehnikah numerične linearne algebre. Forensic za dani latentni faktorski model definira Fréchetov odvod ciljne matrike pri izbrani vhodni matriki kot spremembo latentne predstavitve ciljne matrike pri majhni perturbaciji latentnega modela vhodne matrike. Forensic za ocenjevanje občutljivosti ciljne matrike na spremembe vhodne matrike uporablja inducirane matrične norme in ocenjevanje pogojenostnih števil matrik. Privlačna lastnost pristopa je njegova sposobnost, da oceni prispevke posameznih virov na zgrajeni latentni model, ne da bi zahteval večkratno gradnjo latentnega modela na manjši zbirki virov.

Vrednotenje razvitih metod. V empiričnih raziskavah nas je še posebej zanimalo, ali je Forensic možno uporabiti za odkrivanje "presenetljivih" oziroma problematičnih podatkovnih naborov, ki vsebujejo eksperimentalne napake. V ta namen smo zgradili latentni faktorski model za zbirko 40 naborov genskih interakcij s sočasno matrično tri-faktorizacijo. Opazovali smo, kako se ocene, ki jih izračuna Forensic, spreminjajo, ko v posameznih naborih simuliramo napake, na primer zamenjave bioloških vzorcev. Ugotovili smo, da Forensic uspešno odkrije problematične podatkovne vire. Prav tako smo pri zlivanju velike zbirke molekularnih virov uspeli izboljšati kakovost zgrajenega latentnega modela, tako da smo izključili vire z visoko občutljivostjo. Izbor relevantnih podatkovnih virov je vsekakor zanimiv problem, ki v integrativnih latentnih faktorskih modelih še ni naslovljen, saj se faktorski modeli za sočasno analizo velikih zbirk virov šele razvijajo. Verjamemo, da to področje ponuja veliko možnosti za nadaljnje delo.

A.9 Zaključki in prihodnje delo

V pričujoči doktorski disertaciji so podani naslednji izvirni prispevki k znanosti.

Heterogenost podatkovnih relacij (Del II):

- Razvili smo FuseNet, *splošni* in *učinkovit* pristop za sočasno gradnjo mrež iz raznovrstnih podatkov, ki sledijo različnim porazdelitvam iz eksponentne družine. FuseNet je prvi računski model, ki temelji na teoriji markovskih mrež in izkorišča lastnosti latentnih faktorskih modelov za zlivanje podatkov.

- Razvili smo Réd, računski pristop za gradnjo genskih mrež, skladnih s teorijo epistaze. Pokazali smo, da Réd lahko rekonstruira genske poti iz podatkov o fenotipu enojnih in dvojnih mutant, ki so bolj točne od mrež, zgrajenih z uveljavljenimi pristopi.

- Algoritem metode Réd je *računsko učinkovit*. Tako smo lahko gradili mreže na osnovi več sto tisoč meritev genskih interakcij, kar je največja tovrstna analiza epistaze do sedaj.

- Z metodo FuseNet smo analizirali heterogene podatke v International Cancer Genome Consortium. Ugotovili smo, da mreže, zgrajene z združevanjem genskih izraznih in mutacijskih profilov, izražajo večjo funkcijo obogatenost kot mreže, zgrajene iz le enega podatkovnega vira.

Heterogenost tipov objektov (Del IV):

- Razvili smo Collage, računski pristop h genskih prioritizaciji. Collage na osnovi peščice semenskih genov, relevantnih za izbrani biološki proces ali funkcijo, predlaga najbolj obetavne gene za nadaljnje biološke študije. Collage predstavlja velik napredek v razvoju algoritmov genske prioritizacije, saj omogoča sočasno obravnavo velikih podatkovnih zbirk brez kompleksnih predobdelav podatkov in ohranja relacijsko strukturo podatkov tekom gradnje napovednega modela.

- Predlagali smo novo formalizacijo genske prioritizacije in postavili modele za napovedovanje toksičnosti zdravil in odkrivanje povezanosti med boleznimi, ki imajo veliko možnosti uporabe na področju raziskav ved o življenju. Na primer, odkritje in raziskava štirih semenskih genov, povezanih z bakterijsko rezistenco v amebi *Dictyostelium*, je bilo zahtevno opravilo, ki je zahtevalo več mesecev laboratorijskega dela za vsak gen. *Z uporabo metode Collage smo predlagali osem genov vpletenih v poti bakterijske razpoznave, ki so bili potrjeni z biološkimi raziskavami.* Na ta način je Collage znatno poenostavil in skrajšal čas, potreben za iskanje genov, ki so relevantni za dani biološki proces.

Dvojna heterogenost podatkov (Del I in Del V):

- Razvili smo verjetnostno metodo matričnega dopolnjevanja, ki obravnava predznanje podano z mrežami. Razviti algoritem je učinkovit in dosega boljšo točnost pri napovedovanju genskih interakcij kot uveljavljeni pristopi.

- Pristop matričnega dopolnjevanja zgradi en napovedni model, ki sočasno zliva relacijske podatke s predznanjem. Pokazali smo, da je ta lastnost izrednega pomena za učinkovito napovedovanje *polnih interakcijskih profilov* genov, katerih interakcij sicer ni možno izmeriti zaradi biotehnoloških omejitev. Vključeno predznanje naslavlja problem hladnega zagona, ki se pojavlja v številnih domenah.

- Razvili smo metodo DFMF-SR, računski pristop za analizo preživetja z združevanjem podatkovnih virov. Analizirali smo podatke o raku v International Cancer Genome Consortium, kjer smo pokazali, da DFMF-SR deluje bolje od uveljavljenih pristopov, ki sprva transformirajo podatke v latentni prostor in nato neodvisno od transformacije izvedejo analizo preživetja. DFMF-SR je prvi pristop, ki lahko *sočasno* gradi latentno podatkovno predstavitev in ocenjuje regresijske koeficiente modela za analizo preživetja.

Trojna heterogenost podatkov (Del III):

- Razvili smo algoritem DFMF za sočasno matrično faktorizacijo in ga razširili z metodo za matrično dopolnjevanje. Matematični model smo podkrepili z dokazi pravilnosti in konvergence algoritma za iskanje latentnih matrik. Latentna predstavitev, ki jo zgradi DFMF, minimizira rekonstrukcijsko napako celotnega sistema podatkov v grafu zlivanja.
- V empiričnih študijah smo ugotovili, da lahko z uporabo latentne predstavitve podatkov gradimo napovedi, ki so bolj točne od tistih, dobljenih z uspešnimi metodami za *zgodnjo integracijo* podatkov, kot so naključni gozdovi, in napovednih modelov za *vmesno integracijo*, kot so večjedrne metode.

Odvisnosti med podatkovnimi heterogenostmi (Del VI)

- Razvili smo Forensic, *splošni* in *računsko učinkovit* pristop za ocenjevanje občutljivosti podatkovnih naborov, ki so vključeni v večrelacijski faktorski model. Forensic je prvi pristop, s katerim lahko ocenimo vpliv izbranega vira podatkov na preostale vire v večrelacijskih faktorskih modelih, in se lahko uporabi za izbor virov, ki naj se vključijo v napovedni model zlivanja podatkov.
- Analizirali smo 40 podatkovnih naborov z meritvami fizičnih interakcij med proteini, ki je največja zbirka virov analizirana z večrelacijskih faktorskim modelom do sedaj. S predlagano metodo smo pravilno odkrili vire z neskladnimi podatki in vire, ki vsebujejo eksperimentalne napake.

Prihodnje delo

Naš dolgoročni cilj je analiza velikih in heterogenih podatkovnih zbirk, da bi bolje razumeli, modelirali in napovedali obnašanje ter izboljšali delovanje bioloških, tehnoloških in družbenih sistemov. Želeli bi razviti zmogljive napovedne modele, ki bi lahko razložili odnose in vloge različnih bioloških entitet, kot so geni, zdravila in bolezni; družbenih enot, kot so posamezniki, skupnosti in dogodki; in tehnoloških sistemov, kot je splet. V pričujoči disertaciji smo zastavili nekaj možnih poti v smeri našega dolgoročnega cilja. Sedaj bolje razumemo mešane, večrelacijske, večtipne podatke in napovedne modele ki obravnavajo različne vrste podatkovne heterogenosti. Prav tako lahko učinkovito gradimo latentne modele, ki jih nato uporabljamo za napovedovanje delovanja sistemov na različnih ravneh, nivoju posameznih entitet ali skupnosti. Nadaljnje, dosedanja analiza ponuja zanimive poglede na odnose med predznanjem in zmogljivostjo napovednih modelov, relacijsko strukturo podatkovnih zbirk in izbor relevantnih podatkovnih naborov za dano napovedno nalogo.

V prihodnje se bomo osredotočili na sledeče vidike našega dela, ki ponujajo veliko priložnosti za izboljšave:

- Razvoj učinkovitih napovednih modelov zlivanja podatkov za obravnavo prostorske in časovne lokalnosti ter analizo na različnih stopnjah podatkovne granularnosti.
- Razvoj metod za izbor relevantnih podatkovnih naborov, ki izboljšajo kakovost integrativnih modelov. Zanimiv je tudi razvoj pristopov za analizo skladnosti vzorcev preko podatkovnih naborov in iskanje naborov vprašljive kakovosti.
- Razvoj in uporaba naprednih tehnik za porazdeljeno in vzporedno matrično algebro v smeri podpore interaktivne analize in analize ogromnih podatkovnih zbirk z več sto podatkovnimi nabori in več milijardami podatkovnih točk.

Naša vizija za prihodnost obsega vzpostavitev učinkovitega ogrodja algoritmičnih pristopov, s katerim bomo gradili *raznovrstne napovedne modele* za naloge uvrščanja, razvrščanja in rangiranja v *velikem in heterogenem podatkovju,* ki lahko opisuje različne tipe objektov, uporablja raznolike semantične predstavitve in sledi raznovrstnim podatkovnim porazdelitvam.

BIBLIOGRAPHY

Odd O Aalen. A linear regression model for the analysis of life times. *Statistics in Medicine*, 8(8):907–925, 1989.

Odd O Aalen. Further results on the non-parametric linear regression model in survival analysis. *Statistics in Medicine*, 12(17):1569–1588, 1993.

Alireza Abadi, Saeed Saadat, Parvin Yavari, Chris Bajdik, and Parvin Jalili. Comparison of Aalen's additive and Cox proportional hazards models for breast cancer survival: analysis of population-based data from british columbia, canada. *Asian Pacific Journal of Cancer Prevention*, 12: 3113–3116, 2011.

M. Abu-Odeh, T. Bar-Mag, H. Huang, T. Kim, Z. Salah, S. K. Abdeen, M. Sudol, D. Reichmann, S. Sidhu, P. M. Kim, and R. I. Aqeilan. Characterizing WW domain interactions of tumor suppressor WWOX reveals its association with multiprotein networks. *J. Biol. Chem.*, 289 (13):8865–8880, Mar 2014.

Stein Aerts, Diether Lambrechts, Sunit Maity, Peter Van Loo, Bert Coessens, Frederik De Smet, Leon-Charles Tranchevent, Bart De Moor, Peter Marynen, Bassem Hassan, Peter Carmeliet, and Yves Moreau. Gene prioritization through genomic data fusion. *Nature Biotechnology*, 24(5):537–544, 2006. ISSN 1087-0156. doi: 10.1038/nbt1203.

Jaegyoon Ahn, Youngmi Yoon, Chihyun Park, Eunji Shin, and Sanghyun Park. Integrative gene network construction for predicting a set of complementary prostate cancer genes. *Bioinformatics*, 27(13):1846–1853, 2011.

Awad H Al-Mohy and Nicholas J Higham. A new scaling and squaring algorithm for the matrix exponential. *SIAM Journal on Matrix Analysis and Applications*, 31(3): 970–989, 2009.

Awad H Al-Mohy and Nicholas J Higham. The complex step approximation to the fréchet derivative of a matrix function. *Numerical Algorithms*, 53(1):133–148, 2010.

Awad H Al-Mohy, Nicholas J Higham, and Samuel D Relton. Computing the fréchet derivative of the matrix logarithm and estimating the condition number. *SIAM Journal on Scientific Computing*, 35(4):C394–C410, 2013.

André Albergaria et al. Expression of FOXA1 and GATA3 in breast cancer: the prognostic significance in hormone receptor-negative tumours. *Breast Cancer Research*, 11(3): R40, 2009.

Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, Peter Walter, et al. Molecular biology of the cell. *Garland Science, New York*, 5, 2007.

R. Albright, J. Cox, D. Duling, A. N. Langville, and C. D Meyer. Algorithms, initializations, and convergence for the nonnegative matrix factorization. Technical report, Department of Mathematics, North Carolina State University, 2006.

Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel AJR Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolò Bolli, Ake Borg, Anne-Lise Børresen-Dale, et al. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, 2013.

G. Alexandru, J. Graumann, G. T. Smith, N. J. Kolawa, R. Fang, and R. J. Deshaies. UBXD7 binds multiple ubiquitin ligases and implicates p97 in HIF1alpha turnover. *Cell*, 134(5):804–816, Sep 2008.

Andrey Alexeyenko and Erik L L Sonnhammer. Global networks of functional coupling in eukaryotes from comprehensive data integration. *Genome Research*, 19(6):1107–16, 2009. ISSN 1088-9051. doi: 10.1101/gr.087528.108.

Genevera I Allen and Zhandong Liu. A local poisson graphical model for inferring networks from sequencing data. *IEEE Transactions on NanoBioscience*, 12(3):189–198, 2013.

J. Amberger, C. Bocchini, and A. Hamosh. A new face and new challenges for online mendelian inheritance in man (OMIM). *Human Mutation*, 32(5):564–567, 2011.

Fabrizio Angiulli and Clara Pizzuti. Outlier mining in large high-dimensional data sets. *IEEE Transactions on Knowledge and Data Engineering*, 17(2):203–215, 2005.

Shahzia Anjum, Arnaud Doucet, and Chris C Holmes. A boosting approach to structure learning of graphs with and without prior knowledge. *Bioinformatics*, 25(22): 2929–2936, 2009.

Raman Arora, Maya R Gupta, Amol Kapila, and Maryam Fazel. Similarity-based clustering by left-stochastic matrix factorization. *The Journal of Machine Learning Research*, 14(1):1715–1746, 2013.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene Ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000. doi: 10.1038/75556.

Alan Ashworth, Christopher J Lord, and Jorge S Reis-Filho. Genetic interactions in cancer progression and treatment. *Cell*, 145(1):30–8, April 2011. ISSN 1097-4172.

Leon Avery and Steven Wasserman. Ordering gene function: the interpretation of epistasis in regulatory hierarchies. *Trends in Genetics*, 8(9):312–316, 1992.

S. Aymé, A. Rath, and B. Bellet. WHO international classification of diseases (ICD) revision process: incorporating rare diseases into the classification scheme: state of art. *Orphanet Journal of Rare Diseases*, 5(Suppl 1):P1, 2010.

Francis R. Bach, Gert R. G. Lanckriet, and Michael I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *Proceedings of the 21st International Conference on Machine Learning*, pages 6–14, 2004. ISBN 1-58113-838-5. doi: 10.1145/1015330.1015424.

Bart Bakker and Tom Heskes. Task clustering and gating for bayesian multitask learning. *The Journal of Machine Learning Research*, 4:83–99, 2003.

Sourav Bandyopadhyay et al. Functional maps of protein complexes from quantitative genetic interaction data. *PLoS Computational Biology*, 4(4):e1000065, 2008.

Arindam Banerjee, Sugato Basu, and Srujana Merugu. Multiway clustering on relation graphs. In *SDM*, 2007.

A. J. Barr, E. Ugochukwu, W. H. Lee, O. N. King, P. Filippakopoulos, I. Alfano, P. Savitsky, N. A. Burgess-Brown, S. Muller, and S. Knapp. Large-scale structural analysis of the classical human protein tyrosine phosphatome. *Cell*, 136(2):352–363, Jan 2009.

Alexis Battle, Martin C Jonikas, Peter Walter, Jonathan S Weissman, and Daphne Koller. Automated identification of pathways from quantitative genetic interaction data. *Molecular Systems Biology*, 6, 2010.

Niko Beerenwinkel, Lior Pachter, Bernd Sturmfels, Santiago F Elena, and Richard E Lenski. Analysis of epistatic interactions and fitness landscapes using a new geometric approach. *BMC Evolutionary Biology*, 7, 2007.

Christian Behrends, Mathew E Sowa, Steven P Gygi, and J Wade Harper. Network organization of the human autophagy system. *Nature*, 466(7302):68–76, 2010.

N. Behzadnia, M. M. Golas, K. Hartmuth, B. Sander, B. Kastner, J. Deckert, P. Dube, C. L. Will, H. Urlaub, H. Stark, and R. Luhrmann. Composition and three-dimensional EM structure of double affinity-purified, human prespliceosomal A complexes. *EMBO J.*, 26(6): 1737–1748, Mar 2007.

Robert M Bell and Yehuda Koren. Lessons from the Netflix prize challenge. *ACM SIGKDD Explorations Newsletter*, 9 (2):75–79, 2007.

E. J. Bennett, J. Rush, S. P. Gygi, and J. W. Harper. Dynamics of cullin-RING ubiquitin ligase network revealed by systematic quantitative proteomics. *Cell*, 143(6): 951–965, Dec 2010.

Rajendra Bhatia and Peter Rosenthal. How and why to solve the operator equation AX-XB=Y. *Bulletin of the London Mathematical Society*, 29(1):1–21, 1997.

Bas J Blaauboer and Melvin E Andersen. The need for a new toxicity testing and risk analysis paradigm to implement REACH or any other large scale testing initiative. *Archives of Toxicology*, 81(5):385–387, 2007.

G. Blandin, S. Marchand, K. Charton, N. Daniele, E. Gicquel, J. B. Boucheteil, A. Bentaib, L. Barrault, D. Stockholm, M. Bartoli, and I. Richard. A human skeletal muscle interactome centered on proteins involved in muscular dystrophies: LGMD interactome. *Skelet Muscle*, 3(1):3, 2013.

David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua J. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. *Advances in Neural Information Processing Systems*, 16:17, 2004.

Trond Hellem Bø et al. LSimpute: accurate estimation of missing values in microarray data with least squares methods. *Nucleic Acids Research*, 32(3):e34, 2004.

Henrik Boström, Sten F. Andler, Marcus Brohede, Ronnie Johansson, Alexander Karlsson, Joeri vanLaere, Lars Niklasson, Maria Nilsson, Anne Persson, and Tom Ziemke. On the definition of information fusion as a field of research. Technical report, University of Skovde, School of Humanities and Informatics, Skovde, Sweden, 2007.

David Botstein and Russell Maurer. Genetic approaches to the analysis of microbial development. *Annual Review of Genetics*, 16(1):61–83, 1982.

Anne-Laure Boulesteix, Christine Porzelius, and Martin Daumer. Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value. *Bioinformatics*, 24(15):1698–1706, 2008.

Christos Boutsidis and Efstratios Gallopoulos. SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4):1350–1362, 2008. ISSN 0031-3203. doi: 10.1016/j.patcog.2007.09.010.

Mike Bowles and Ron Shigeta. Statistical models for predicting liver toxicity from genomic data. *Systems Biomedicine*, 1(3):1–6, 2013.

Gunes Bozkurt, Goran Stjepanovic, Fabio Vilardi, Stefan Amlacher, Klemens Wild, Gert Bange, Vincenzo Favaloro, Karsten Rippe, Ed Hurt, Bernhard Dobberstein, et al. Structural insights into tail-anchored protein binding and membrane insertion by Get3. *Proceedings of the National Academy of Sciences*, 106(50):21131–21136, 2009.

Salvatore Bozzaro and Ludwig Eichinger. The professional phagocyte Dictyostelium discoideum as a model host for bacterial pathogens. *Current Drug Targets*, 12(7):942, 2011.

M. Brehme, O. Hantschel, J. Colinge, I. Kaupe, M. Planyavsky, T. Kocher, K. Mechtler, K. L. Bennett, and G. Superti-Furga. Charting the molecular network of the drug target Bcr-Abl. *Proc. Natl. Acad. Sci. U.S.A.*, 106(18):7414–7419, May 2009.

Leo Breiman. Random forests. *Machine Learning*, 45(1): 5–32, 2001.

Debra A Brock, R Diane Hatton, Dan-Victor Giurgiutiu, Brenton Scott, Robin Ammann, and Richard H Gomer. The different components of a multisubunit cell number-counting factor have both unique and overlapping functions. *Development*, 129(15):3657–3668, 2002.

Guy N Brock et al. Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes. *BMC Bioinformatics*, 9(1):12, 2008.

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

Serhat S Bucak, Rong Jin, and Anil K Jain. Multiple kernel learning for visual object recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1354–1369, 2014.

D. Buob and M. C. Copin. [Mixed cryoglobulinemia-associated membranoproliferative glomerulonephritis, disclosing gastric MALT lymphoma]. *Annales de Pathologie*, 26(4):267–270, Sep 2006.

Matthew Cabral, Christophe Anjard, William F Loomis, and Adam Kuspa. Genetic evidence that the acyl coenzyme A binding protein AcbA and the serine protease/ABC transporter TagA function together in dictyostelium discoideum cell differentiation. *Eukaryot Cell*, 5(12): 2024–2032, 2006.

Matthew Cabral, Christophe Anjard, Vivek Malhotra, William F Loomis, and Adam Kuspa. Unconventional secretion of AcbA in Dictyostelium discoideum through a vesicular intermediate. *Eukaryot Cell*, 9(7):1009–1017, 2010.

Jian-Feng Cai et al. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

Ying Cai, Bernard Fendler, and Gurinder S Atwal. Utilizing RNA-seq data for cancer network inference. In *IEEE GENSIPS*, pages 46–49, 2012.

Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

E. Cannavo, B. Gerrits, G. Marra, R. Schlapbach, and J. Jiricny. Characterization of the interactome of the human MutL homologues MLH1, PMS1, and PMS2. *J. Biol. Chem.*, 282(5):2976–2986, Feb 2007.

M.W.P. Carney, T.K.N. Chary, M. Laundy, T. Bottiglieri, I. Chanarin, E.H. Reynolds, and B. Toone. Red cell folate concentrations in psychiatric patients. *Journal of Affective Disorders*, 19(3):207 – 213, 1990. ISSN 0165-0327.

Gail A. Carpenter, Siegfried Martens, and Ogi J. Ogas. Self-organizing information fusion and hierarchical knowledge discovery: a new framework using ARTMAP neural networks. *Neural Netw.*, 18(3):287–295, 2005. ISSN 0893-6080. doi: 10.1016/j.neunet.2004.12.003.

Pedro Carvalho, Veit Goder, and Tom A Rapoport. Distinct ubiquitin-ligase complexes define convergent pathways for the degradation of ER proteins. *Cell*, 126(2):361–373, 2006.

Pierre Charbit, Stéphan Thomassé, Anders Yeo, et al. The minimum feedback arc set problem is NP-hard for tournaments. *Combinatorics, Probability and Computing*, 16: 1–4, 2007.

Andrew Chatr-aryamontri, Bobby-Joe Breitkreutz, Rose Oughtred, Lorrie Boucher, Sven Heinicke, Daici Chen, Chris Stark, Ashton Breitkreutz, Nadine Kolas, Lara O'Donnell, et al. The BioGRID interaction database: 2015 update. *Nucleic Acids Research*, page gku1204, 2014.

Kamalika Chaudhuri, Sham M. Kakade, Karen Livescu, and Karthik Sridharan. Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th International Conference on Machine Learning*, pages 129–136, 2009. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553391.

Guokai Chen, Olga Zhuchenko, and Adam Kuspa. Immune-like phagocyte activity in the social amoeba. *Science*, 317(5838):678–681, August 2007. doi: 10.1126/science.1143991.

Jing Chen, Eric E Bardes, Bruce J Aronow, and Anil G Jegga. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res*, 37(suppl 2):W305–W311, 2009.

M. Chen, V. Vijay, Q. Shi, Z. Liu, H. Fang, and W. Tong. FDA-approved drug labeling for the study of drug-induced liver injury. *Drug Discov. Today*, 16(15-16): 697–703, 2011.

Zheng Chen and Weixiong Zhang. Integrative analysis using module-guided random forests reveals correlated genetic factors related to mouse weight. *PLoS Computational Biology*, 9(3):e1002956, 2013. doi: 10.1371/journal.pcbi.1002956.

Nicholas A Christakis and James H Fowler. Friendship and natural selection. *Proceedings of the National Academy of Sciences*, 111(3):10796–10801, 2014.

G. Christopoulou, C. Sismani, M. Sakellariou, M. Saklamaki, V. Athanassiou, and V. Velissariou. Clinical and molecular description of the prenatal diagnosis of a fetus with a maternally inherited microduplication 22q11.2 of 2.5Mb. *Gene*, 527(2):694–697, 2013.

P. T. Clayton, A. Verrips, E. Sistermans, A. Mann, G. Mieli-Vergani, and R. Wevers. Mutations in the sterol 27-hydroxylase gene (CYP27A) cause hepatitis of infancy as well as cerebrotendinous xanthomatosis. *Journal of Inherited Metabolic Disease*, 25(6):501–513, Oct 2002.

Simone Clerc, Christian Hirsch, Daniela Maria Oggier, Paola Deprez, Claude Jakob, Thomas Sommer, and Markus Aebi. Htm1 protein generates the N-glycan signal for glycoprotein degradation in the endoplasmic reticulum. *The Journal of Cell Biology*, 184(1):159–172, 2009.

Djork-Arne Clevert, Martin Heusel, Andreas Mitterecker, Willem Talloen, Hinrich Göhlmann, Jörg Wegner, Andreas Mayr, Günter Klambauer, and Sepp Hochreiter. Exploiting the Japanese toxicogenomics project for predictive modelling of drug toxicity. *CAMDA 2012*, pages 26–29, 2012.

Francis S Collins and Harold Varmus. A new initiative on precision medicine. *New England Journal of Medicine*, 372(9):793–795, 2015.

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

Sean R Collins et al. A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. *Genome Biology*, 7:R63, 2006.

Sean R Collins et al. Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature*, 446(7137):806–810, 2007.

ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.

Heather J Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20):2463–2468, 2002.

R. Cornet and N. De Keizer. Forty years of SNOMED: a literature review. *BMC Medical Informatics and Decision Making*, 8(Suppl 1):S2, 2008.

Alex J Cornish and Florian Markowetz. SANTA: quantifying the functional content of molecular networks. *PLoS Computational Biology*, 10(9):e1003808, 2014.

C Cortes and V Vapnik. Support vector machine. *Machine Learning*, 20(3):273–297, 1995.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Two-stage learning kernel algorithms. In *Proceedings of the 27th International Conference on Machine Learning*, pages 239–246, 2010.

Michael Costanzo, Anastasia Baryshnikova, et al. The genetic landscape of a cell. *Science*, 327(5964):425–431, 2010.

D Cox, D Wessels, DR Soll, J Hartwig, and J Condeelis. Re-expression of ABP-120 rescues cytoskeletal, motility, and phagocytosis defects of ABP-120-Dictyostelium mutants. *Molecular Biology of the Cell*, 7(5):803–823, 1996.

MG Cox and PM Harris. Numerical analysis for algorithm design in metrology. Software support for metrology best practice guide. Technical report, National Physical Laboratory, Teddington, 2004.

David Croft, Antonio Fabregat Mundo, Robin Haw, Marija Milacic, Joel Weiser, Guanming Wu, Michael Caudy, Phani Garapati, Marc Gillespie, Maulik R Kamdar, et al. The Reactome pathway knowledgebase. *Nucleic Acids Res*, 42(D1):D472–D477, 2014.

Giovanni Da San Martino, Nicolo Navarin, and Alessandro Sperduti. A tree-based kernel for graphs. In *SDM*, pages 975–986, 2012.

A. Dagli, K. Buiting, and C. A. Williams. Molecular and clinical aspects of Angelman syndrome. *Molecular Syndromology*, 2(3-5):100–112, Apr 2012.

Tijl De Bie, Léon-Charles Tranchevent, Liesbeth M M van Oeffelen, and Yves Moreau. Kernel-based data fusion for gene prioritization. *Bioinformatics*, 23(13):i125–32, 2007. ISSN 1367-4811. doi: 10.1093/bioinformatics/btm187.

Alexandre G de Brevern et al. Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. *BMC Bioinformatics*, 5(1):114, 2004.

C. L. de Hoog, L. J. Foster, and M. Mann. RNA and RNA binding proteins participate in early stages of cell spreading through spreading initiation centers. *Cell*, 117(5): 649–662, May 2004.

Rameswar Debnath and Haruhisa Takahashi. Kernel selection for the support vector machine. *IEICE Transactions*, 87-D(12):2903–2904, 2004.

Lei Ding, Alper Yilmaz, and Rong Yan. Interactive image segmentation using Dirichlet process multiple-view learning. *IEEE Transactions on Image Processing*, 21(4): 2119–2129, 2012.

David J Dix, Keith A Houck, Matthew T Martin, Ann M Richard, R Woodrow Setzer, and Robert J Kavlock. The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicological Sciences*, 95(1):5–12, 2007.

Dirk Dormann, Gerti Weijer, Simon Dowler, and Cornelis J Weijer. In vivo analysis of 3-phosphoinositide dynamics during Dictyostelium phagocytosis and chemotaxis. *Journal of Cell Science*, 117(26):6497–6509, 2004.

Becky L Drees, Vesteinn Thorsson, Gregory W Carter, Alexander W Rives, Marisa Z Raymond, Iliana Avila-Campillo, Paul Shannon, and Timothy Galitski. Derivation of genetic interaction networks from quantitative phenotype data. *Genome Biology*, 6(4):R38, 2005.

Richard O Duda and Peter E Hart. *Pattern classification and scene analysis*. Wiley, 1973.

Cynthia Dwork, Ravi Kumar, Moni Naor, and Dandapani Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th International Conference on World Wide Web*, pages 613–622. ACM, 2001.

Saso Džeroski and Bernard Ženko. Is combining classifiers with stacking better than selecting the best one? *Machine learning*, 54(3):255–273, 2004.

Peter Eades, Xuemin Lin, and William F Smyth. A fast and effective heuristic for the feedback arc set problem. *Information Processing Letters*, 47(6):319–323, 1993.

Dimond RL Ebert DL, Bush JM and Cardelli JA. Biogenesis of lysosomal enzymes in the alpha-glucosidase II-deficient modA mutant of Dictyostelium discoideum: retention of alpha-1,3-linked glucose on N-linked oligosaccharides delays intracellular transport but does not alter sorting of alpha-mannosidase or beta-glucosidase. *Arch Biochem Biophys*, 273:479–490, 1989.

F Emmert-Streib, S Tripathi, R Simoes, A Hawwa, and M Dehmer. The human disease network: opportunities for classification, diagnosis and prediction of disorders and disease genes. *Systems Biomedicine*, 1:15–22, 2013.

Jason Ernst and Manolis Kellis. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nature Biotechnology*, 34(224):203–215, 2015.

Andrey Feuerverger, Yu He, Shashi Khatri, et al. Statistical significance of the Netflix challenge. *Statistical Science*, 27 (2):202–231, 2012.

Petra Fey, Pascale Gaudet, Tomaž Curk, Blaž Zupan, Eric M Just, Siddhartha Basu, Sohel N Merchant, Yulia A Bushmanova, Gad Shaulsky, Warren A Kibbe, et al. dictyBase – a Dictyostelium bioinformatics resource update. *Nucleic Acids Res*, 37(Suppl 1):D515–D519, 2009.

S. Foerster, T. Kacprowski, V. M. Dhople, E. Hammer, S. Herzog, H. Saafan, S. Bien-Moller, M. Albrecht, U. Volker, and C. A. Ritter. Characterization of the EGFR interactome reveals associated protein complex networks and intracellular receptor dynamics. *Proteomics*, 13(21):3131–3144, Nov 2013.

Jean-Fred Fontaine, Florian Priller, Adriano Barbosa-Silva, and Miguel A Andrade-Navarro. Génie: literature-based gene prioritization at multi genomic scale. *Nucleic Acids Res*, 39(suppl 2):W455–W461, 2011.

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 Res*, 41(D1):D808–D815, 2013.

Lude Franke, Harm Van Bakel, Begoña Diosdado, Martine Van Belzen, Martin Wapenaar, and Cisca Wijmenga. TEAM: a tool for the integration of expression, and linkage and association maps. *European Journal of Human Genetics*, 12(8):633–638, 2004.

Jerome Friedman, Trevor Hastie, Holger Höfling, Robert Tibshirani, et al. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007a.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the lasso. *Biostatistics*, 9:432–441, 2007b.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

Jerome H Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002.

Mélina Gallopin, Andrea Rau, and Florence Jaffrézic. A hierarchical Poisson log-normal model for network inference from RNA sequencing data. *PLoS One*, 8(10):e77503, 2013.

Tong Gao, Celine Roisin-Bouffay, R Diane Hatton, Lei Tang, Debra A Brock, Tiffany DeShazo, Laura Olson, Wan-Pyo Hong, Wonhee Jang, Elvia Canseco, et al. A cell number-counting factor regulates levels of a novel protein, SslA, as part of a group size regulation mechanism in Dictyostelium. *Eukaryot Cell*, 6(9):1538–1551, 2007.

Z. Gao, J. Zhang, R. Bonasio, F. Strino, A. Sawai, F. Parisi, Y. Kluger, and D. Reinberg. PCGF homologs, CBX proteins, and RYBP define functionally distinct PRC1 family complexes. *Mol. Cell*, 45(3):344–356, Feb 2012.

Levi A Garraway and Eric S Lander. Lessons from the cancer genome. *Cell*, 153(1):17–37, 2013.

Olivier Gevaert, Frank De Smet, Dirk Timmerman, Yves Moreau, and Bart De Moor. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics*, 22(14):e184–90, 2006. ISSN 1367-4811. doi: 10.1093/bioinformatics/btl230.

Gillian Gifford, Jim Paul, Paul A Vasey, Stanley B Kaye, and Robert Brown. The acquisition of hMLH1 methylation in plasma DNA after chemotherapy predicts poor survival for ovarian cancer patients. *Clinical Cancer Research*, 10(13):4420–4426, 2004.

Kwang-il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-Laszlo Barabasi. The human disease network. *PNAS*, 104(21):8685–8690, 2007.

Mehmet Gönen and Ethem Alpaydin. Localized multiple kernel learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 352–359. ACM, 2008.

Mehmet Gönen and Ethem Alpaydın. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12: 2211–2268, 2011. ISSN 1532-4435.

Geoffrey J. Gordon. Generalized^2 linear^2 models. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *NIPS*, pages 577–584. MIT Press, 2002. ISBN 0-262-02550-7.

Derek Greene and Pádraig Cunningham. A matrix factorization approach for integrating multiple data views. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 423–438, Berlin, Heidelberg, 2009. Springer-Verlag. ISBN 978-3-642-04179-2. doi: 10.1007/978-3-642-04180-8_45.

Christopher Greenman, Philip Stephens, Raffaella Smith, Gillian L Dalgliesh, Christopher Hunter, Graham Bignell, Helen Davies, Jon Teague, Adam Butler, Claire Stevens, et al. Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132):153–158, 2007.

Helmut Greim, Michael Arand, Herman Autrup, Hermann M Bolt, James Bridges, Erik Dybing, Remi Glomot, Vito Foa, and Rolf Schulte-Hermann. Toxicological comments to the discussion about REACH. *Archives of Toxicology*, 80(3):121–124, 2006.

Natali Gulbahce, Han Yan, Amélie Dricot, Megha Padi, Danielle Byrdsong, Rachel Franchi, Deok-Sun Lee, Orit Rozenblatt-Rosen, Jessica C Mar, Michael Calderwood, Amy Baldwin, Bo Zhao, Balaji Santhanam, Pascal Braun, Nicolas Simonis, Kyung-Won Huh, Karin Hellner, Miranda Grace, Alyce Chen, Renee Rubio, Jarrod A Marto, Nicholas A Christakis, Elliott Kieff, Frederick P Roth, Jennifer Roecklein-Canfield, James a Decaprio, Michael E Cusick, John Quackenbush, David E Hill, Karl Münger, Marc Vidal, and Albert-László Barabási. Viral perturbations of host networks reflect disease etiology. *PLoS Computational Biology*, 8(6):e1002531, 2012.

Ulrich Güldener, Martin Münsterkötter, Gabi Kastenmüller, Normann Strack, Jacques van Helden, Christian Lemer, J Richelles, Shoshana J Wodak, José García-Martínez, José Enrique Pérez-Ortín, et al. CYGD: the comprehensive yeast genome database. *Nucleic Acids Research*, 33 (suppl 1):D364–D368, 2005.

Shengbo Guo, Onno Zoeter, and Cédric Archambeau. Sparse bayesian multi-task learning. In *Advances in Neural Information Processing Systems*, pages 1755–1763, 2011.

Xinjian Guo, Yilong Yin, Cailing Dong, Goping Yang, and Guangtong Zhou. On the class imbalance problem. In *Fourth International Conference on Natural Computation*, volume 4, pages 192–201. IEEE, 2008.

David L Hall and James Llinas. An introduction to multisensor data fusion. *Proceedings of the IEEE*, 85(1):6–23, 1997.

Shaobo Han, Xuejun Liao, and Lawrence Carin. Cross-domain multitask learning with latent probit models. In *Proceedings of the 19th International Conference on Machine Learning*, 2012.

P. C. Havugimana, G. T. Hart, T. Nepusz, H. Yang, A. L. Turinsky, Z. Li, P. I. Wang, D. R. Boutz, V. Fong, S. Phanse, M. Babu, S. A. Craig, P. Hu, C. Wan, J. Vlasblom, V. U. Dar, A. Bezginov, G. W. Clark, G. C. Wu, S. J. Wodak, E. R. Tillier, A. Paccanaro, E. M. Marcotte, and A. Emili. A census of human soluble protein complexes. *Cell*, 150(5):1068–1081, Aug 2012.

Jingrui He and Rick Lawrence. A graph-based framework for multi-task multi-view learning. In *Proceedings of the 28th International Conference on Machine Learning*, pages 25–32, 2011.

Jingrui He, Yan Liu, and Qiang Yang. Linking heterogeneous input spaces with pivots for multi-task learning. In *Proceedings of the SIAM International Conference on Data Mining*, pages 181–189. SIAM, 2014.

Ari Helenius and Markus Aebi. Roles of N-linked glycans in the endoplasmic reticulum. *Annual Review of Biochemistry*, 73(1):1019–1049, 2004.

D. Heuss, A. Engelhardt, H. Gobel, and B. Neundorfer. Myopathological findings in interstitial myositis in type II polyendocrine autoimmune syndrome (Schmidt's syndrome). *Neurological Research*, 17(3):233–237, Jun 1995.

César a Hidalgo, Nicholas Blumm, Albert-László Barabási, and Nicholas A. Christakis. A dynamic network approach for the study of human phenotypes. *PLoS Computational Biology*, 5(4):e1000353, April 2009. ISSN 1553-7358.

Nicholas J Higham. *Functions of matrices: theory and computation*. SIAM, Philadelphia, PA, USA, 2008.

Nicholas J Higham and Françoise Tisseur. A block algorithm for matrix 1-norm estimation, with an application to 1-norm pseudospectra. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1185–1201, 2000.

Sepp Hochreiter, Djork-Arné Clevert, and Klaus Obermayer. A new summarization method for affymetrix probe level data. *Bioinformatics*, 22(8):943–949, 2006. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl033.

Sepp Hochreiter, Ulrich Bodenhofer, Martin Heusel, Andreas Mayr, Andreas Mitterecker, Adetayo Kasim, Tatsiana Khamiakova, Suzy Van Sanden, Dan Lin, Willem Talloen, et al. Fabia: factor analysis for bicluster acquisition. *Bioinformatics*, 26(12):1520–1527, 2010.

Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57. ACM, 1999.

Matan Hofree, John P Shen, Hannah Carter, Andrew Gross, and Trey Ideker. Network-based stratification of tumor mutations. *Nature Methods*, 10:1108–1115, 2013.

Friedrich Georg Ernst Holst, Christoph Josef Hemmer, Christian Foth, Rainer Seitz, Rudolf Egbring, and Manfred Dietrich. Low levels of fibrin-stabilizing factor (factor XIII) in human Plasmodium falciparum malaria: correlation with clinical severity. *American Journal of Tropical Medicine and Hygiene*, 60:99–104, 1999.

Roger A Horn and Charles R Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.

Roger A Horn and Charles R Johnson. *Matrix Analysis*. Cambridge University Press, 2012.

D.L. Howell, J. Bergsagel, R. Chu, and L. Meacham. Suppression of Hodgkin's disease in a patient with Cushing's syndrome. *Journal of Pediatric Hematology/Oncology*, 26 (5):301–303, 2004.

Linda S Huang and Paul W Sternberg. Genetic dissection of developmental pathways. *Methods in Cell Biology*, 48: 97–122, 1995.

Thomas J Hudson, Warwick Anderson, Axel Aretz, Anna D Barker, Cindy Bell, Rosa R Bernabé, MK Bhan, Fabien Calvo, Iiro Eerola, Daniela S Gerhard, et al. International network of cancer genome projects. *Nature*, 464(7291): 993–998, 2010.

Timothy R Hughes. Universal epistasis analysis. *Nature Genetics*, 37(5):457–457, 2005.

Timothy R Hughes, Matthew J Marton, Allan R Jones, Christopher J Roberts, Roland Stoughton, Christopher D Armour, Holly A Bennett, Ernest Coffey, Hongyue Dai, Yudong D He, et al. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–126, 2000.

Lawrence Hunter. *Artificial intelligence and molecular biology*. Granite Hill Publishers, 1993.

Lucie N. Hutchins, Sean M. Murphy, Priyam Singh, and Joel H. Graber. Position-dependent motif characterization using non-negative matrix factorization. *Bioinformatics*, 24(23):2684–2690, 2008. ISSN 1367-4803. doi: 10.1093/bioinformatics/btn526.

Curtis Huttenhower, K. Tsheko Mutungu, Natasha Indik, Woongcheol Yang, Mark Schroeder, Joshua J. Forman, Olga G. Troyanskaya, and Hilary A. Coller. Detailing regulatory networks through large scale data integration. *Bioinformatics*, 25(24):3267–3274, 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp588.

Alba Hykollari, Martin Dragosits, Dubravko Rendić, Iain BH Wilson, and Katharina Paschinger. N-glycomic profiling of a glucosidase ii mutant of Dictyostelium discoideum by "off-line" liquid chromatography and mass spectrometry. *Electrophoresis*, 35(15):2116–2129, 2014.

Francesco Iorio, Roberta Bosotti, Emanuela Scacheri, Vincenzo Belcastro, Pratibha Mithbaokar, Rosa Ferriero, Loredana Murino, Roberto Tagliaferri, Nicola Brunetti-Pierri, Antonella Isacchi, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proceedings of the National Academy of Sciences*, 107(33):14621–14626, 2010.

Senol Isci et al. Bayesian network prior: network analysis of biological data using external knowledge. *Bioinformatics*, 30(6):860–867, 2014.

S. Jager, P. Cimermancic, N. Gulbahce, J. R. Johnson, K. E. McGovern, S. C. Clarke, M. Shales, G. Mercenne, L. Pache, K. Li, H. Hernandez, G. M. Jang, S. L. Roth, E. Akiva, J. Marlett, M. Stephens, I. D'Orso, J. Fernandes, M. Fahey, C. Mahon, A. J. O'Donoghue, A. Todorovic, J. H. Morris, D. A. Maltby, T. Alber, G. Cagney, F. D. Bushman, J. A. Young, S. K. Chanda, W. I. Sundquist, T. Kortemme, R. D. Hernandez, C. S. Craik, A. Burlingame, A. Sali, A. D. Frankel, and N. J. Krogan. Global landscape of HIV-human protein complexes. *Nature*, 481(7381):365–370, Jan 2012.

Ariel Jaimovich and Nir Friedman. From large-scale assays to mechanistic insights: computational analysis of interactions. *Current Opinion in Biotechnology*, 22(1):87–93, 2011.

Ali Jalali, Pradeep D Ravikumar, Vishvas Vasuki, and Sujay Sanghavi. On learning discrete graphical models using group-sparse regularization. In *AISTATS*, pages 378–387, 2011.

Vuk Janjić and Nataša Pržulj. Biological function through network topology: a survey of the human diseasome. *Briefings in Functional Genomics*, September 2012. ISSN 2041-2657.

Aki P Järvinen et al. Predicting quantitative genetic interactions by means of sequential matrix approximation. *PLoS One*, 3(9):e3284, 2008.

J. Jin, F. D. Smith, C. Stark, C. D. Wells, J. P. Fawcett, S. Kulkarni, P. Metalnikov, P. O'Donnell, P. Taylor, L. Taylor, A. Zougman, J. R. Woodgett, L. K. Langeberg, J. D. Scott, and T. Pawson. Proteomic, functional, and domain-based analysis of in vivo 14-3-3 binding proteins involved in cytoskeletal regulation and cellular organization. *Curr. Biol.*, 14(16):1436–1450, Aug 2004.

Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

R. B. Jones, A. Gordus, J. A. Krall, and G. MacBeath. A quantitative protein interaction network for the ErbB receptors using protein microarrays. *Nature*, 439(7073): 168–174, Jan 2006.

Martin C Jonikas, Sean R Collins, Vladimir Denic, Eugene Oh, Erin M Quan, Volker Schmid, Jimena Weibezahn, Blanche Schwappach, Peter Walter, Jonathan S Weissman, et al. Comprehensive characterization of genes required for protein folding in the endoplasmic reticulum. *Science*, 323(5922):1693–1697, 2009.

Rebecka Jörnsten et al. DNA microarray data imputation and significance analysis of differential expression. *Bioinformatics*, 21(22):4155–4161, 2005.

Cynthia Ju and Timothy Reilly. Role of immune reactions in drug-induced liver injury (DILI). *Drug Metabolism Reviews*, 44(1):107–115, 2012. doi: 10.3109/03602532.2011.645579.

J. J. Kahle, N. Gulbahce, C. A. Shaw, J. Lim, D. E. Hill, A. L. Barabasi, and H. Y. Zoghbi. Comparison of an expanded ataxia interactome with patient medical records reveals a relationship between macular degeneration and ataxia. *Hum. Mol. Genet.*, 20(3):510–527, Feb 2011.

M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(D1): D109–D114, 2012.

Minoru Kanehisa, Michihiro Araki, Susumu Goto, Masahiro Hattori, Mika Hirakawa, Masumi Itoh, Toshiaki Katayama, Shuichi Kawashima, Shujiro Okuda, Toshiaki Tokimatsu, et al. KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, 36(suppl 1): D480–D484, 2008.

Minoru Kanehisa, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Data, information, knowledge and principle: back to metabolism in kegg. *Nucleic Acids Research*, 42(D1): D199–D205, 2014.

Neil Kaplowitz. Avoiding idiosyncratic DILI: Two is better than one. *Hepatology*, 2013. ISSN 1527-3350. doi: 10.1002/hep.26295.

Dimitris Karlis. An EM algorithm for multivariate Poisson distribution and related models. *Journal of Applied Statistics*, 30(1):63–77, 2003.

H. Karmous-Benailly, F. Giuliano, C. Massol, C. Bloch, D. De Ricaud, J. C. Lambert, and S. Perelman. Unbalanced inherited complex chromosome rearrangement involving chromosome 8, 10, 11 and 16 in a patient with congenital malformations and delayed development. *European Journal of Medical Genetics*, 49(5):431–438, 2006.

Ryan Kelley and Trey Ideker. Systematic interpretation of genetic interactions using protein networks. *Nature Biotechnology*, 23(5):561–566, 2005.

E. J. Keuter. [Vitamin B complex deficiency causing the psychiatric symptoms of atypical endogenous depression]. *Ned Tijdschr Geneeskd*, 102(31):1501–1503, Aug 1958.

Hyunsoo Kim et al. Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics*, 21(2):187–198, 2005.

Myunghwan Kim and Jure Leskovec. Nonparametric multi-group membership model for dynamic networks. In *NIPS*, 2013.

Philip M. Kim and Bruce Tidor. Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Research*, 13:1706–1718, 2003. doi: 10.1101/gr.903503.fully.

Arto Klami and Samuel Kaski. Probabilistic approach to detecting dependencies between data sets. *Neurocomput.*, 72(1-3):39–46, 2008. ISSN 0925-2312. doi: 10.1016/j.neucom.2007.12.044.

Hans-Ulrich Klein, Martin Schäfer, Bo T Porse, Marie S Hasemann, Katja Ickstadt, and Martin Dugas. Integrative analysis of histone chip-seq and transcription data using bayesian mixture models. *Bioinformatics*, 30(8): 1154–1162, 2014.

Marius Kloft, Ulf Brefeld, Soren Sonnenburg, Pavel Laskov, Klaus-Robert Muller, and Alexander Zien. Efficient and accurate L_p-norm multiple kernel learning. *Advances in Neural Information Processing Systems*, 21:997–1005, 2009.

Marius Kloft, Ulf Brefeld, Sören Sonnenburg, and Alexander Zien. L_p-norm multiple kernel learning. *J. Mach. Learn. Res.*, 12:953–997, 2011. ISSN 1532-4435.

M. Kneissl, V. Putter, A. A. Szalay, and F. Grummt. Interaction and assembly of murine pre-replicative complex proteins in yeast and mouse cells. *J. Mol. Biol.*, 327(1): 111–128, Mar 2003.

Craig Knox, Vivian Law, Timothy Jewison, Philip Liu, Son Ly, Alex Frolkis, Allison Pon, Kelly Banco, Christine Mak, Vanessa Neveu, Yannick Djoumbou, Roman Eisner, Anchi Guo, and David S. Wishart. Drugbank 3.0: a comprehensive resource for 'Omics' research on drugs. *Nucleic Acids Research*, 39(Database-Issue):1035–1041, 2011.

L. J. Kohler, A. F. Gohara, R. W. Hamilton, and R. S. Reeves. Crescentic fibrillary glomerulonephritis associated with intermittent rifampin therapy for pulmonary tuberculosis. *Clinical Nephrology*, 42(4):263–265, Oct 1994.

Sebastian Köhler, Sebastian Bauer, Denise Horn, and Peter N Robinson. Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet*, 82(4): 949–958, 2008.

Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

Risi Imre Kondor and John D. Lafferty. Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of the 19th International Conference on Machine Learning*, pages 315–322, 2002. ISBN 1-55860-873-7.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8):30–37, 2009. ISSN 0018-9162. doi: 10.1109/MC.2009.263.

Masaaki Kotera et al. GENIES: gene network inference engine based on supervised analysis. *Nucleic Acids Research*, 2012.

Balaji Krishnapuram et al. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 27(6):957–968, 2005.

H. W. Kuhn and A. W. Tucker. Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 481–492. University of California Press, 1951.

K-L Lai and JL Crassidis. Extensions of the first and second complex-step derivative approximations. *Journal of Computational and Applied Mathematics*, 219(1):276–293, 2008.

Gert R. G. Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I. Jordan. Learning the kernel matrix with semidefinite programming. *J. Mach. Learn. Res.*, 5:27–72, 2004a. ISSN 1532-4435.

Gert R. G. Lanckriet, Tijl De Bie, Nello Cristianini, Michael I. Jordan, and William Stafford Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635, 2004b. ISSN 1367-4803.

Gert R. G. Lanckriet, Tijl De Bie, Nello Cristianini, Michael I. Jordan, and William Stafford Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635, 2004c. ISSN 1367-4803. doi: 10.1093/bioinformatics/bth294.

Tilman Lange and Joachim M. Buhmann. Fusion of similarity data in clustering. In *Advances in Neural Information Processing Systems*, pages 723–730, 2005.

Vivian Law, Craig Knox, Yannick Djoumbou, Tim Jewison, An Chi Guo, Yifeng Liu, Adam Maciejewski, David Arndt, Michael Wilson, Vanessa Neveu, et al. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Research*, 42(D1):D1091–D1097, 2014.

Jerald F Lawless and Yan Yuan. Estimation of prediction error for survival models. *Statistics in Medicine*, 29(2): 262–274, 2010.

Michael S Lawrence, Petar Stojanov, Paz Polak, Gregory V Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L Carter, Chip Stewart, Craig H Mermel, Steven A Roberts, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218, 2013.

Clare M Lee, Manikhandan AV Mudaliar, DR Haggart, C Roland Wolf, Gino Miele, J Keith Vass, Desmond J Higham, and Daniel Crowther. Simultaneous non-negative matrix factorization for multiple large scale gene expression datasets in toxicology. *PLoS One*, 7(12): e48238, 2012.

D-S Lee, J Park, K a Kay, N A Christakis, Z N Oltvai, and Albert-László Barabási. The implications of human metabolic network topology for disease comorbidity. *PNAS*, 105(29):9880–5, July 2008. ISSN 1091-6490.

Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *Advances in Neural Information Processing Systems*, pages 556–562. MIT Press, 2000.

Troy Lee and Adi Shraibman. Matrix completion from any given set of observations. In *Advances in Neural Information Processing Systems*, volume 26, pages 1781–1787, 2013.

W. M. Lee. Drug-induced hepatotoxicity. *N. Engl. J. Med.*, 349(5):474–485, 2003.

B. Lehner and C. M. Sanderson. A protein interaction framework for human mRNA degradation. *Genome Res.*, 14 (7):1315–1323, Jul 2004.

Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Signed networks in social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1361–1370, 2010.

Jun Li et al. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*, pages 1–16, 2011.

Tao Li, Yi Zhang, and Vikas Sindhwani. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 244–252, 2009a. ISBN 978-1-932432-45-9.

Tao Li, Yi Zhang, and Vikas Sindhwani. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing*, pages 244–252. Association for Computational Linguistics, 2009b.

Wu-jun Li and Dit-yan Yeung. Relation regularized matrix factorization. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1126–1131, 2007.

Wu-Jun Li, Dit-Yan Yeung, and Zhihua Zhang. Probabilistic relational PCA. In *Advances in Neural Information Processing Systems*, pages 1123–1131, 2009c.

Xingxing Li, Xiaohong Zhang, Xiaodong Ren, Mathias Fritsche, Jens Wickert, and Harald Schuh. Precise positioning with current multi-constellation global navigation satellite systems: Gps, glonass, galileo and beidou. *Scientific Reports*, 5:8328, 2015.

Alan Wee-Chung Liew et al. Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Briefings in Bioinformatics*, 12(5):498–513, 2011.

V. Lim and B. L. Clarke. Coexisting primary hyperparathyroidism and sarcoidosis cause increased Angiotensin-converting enzyme and decreased parathyroid hormone and phosphate levels. *Journal of Clinical Endocrinology and Metabolism*, 98(5):1939–1945, May 2013.

Wanessa C Lima, Emmanuelle Lelong, and Pierre Cosson. What can Dictyostelium bring to the study of Pseudomonas infections? In *Seminars in Cell & Developmental Biology*, volume 22, pages 77–81. Elsevier, 2011.

Jimmy Lin and Alek Kolcz. Large-scale machine learning at Twitter. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 793–804. ACM, 2012.

Bolan Linghu, Evan S Snitkin, Zhenjun Hu, Yu Xia, and Charles Delisi. Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. *Genome Biology*, 10(9):R91, January 2009. ISSN 1465-6914.

Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *The Journal of Machine Learning Research*, 10:2295–2328, 2009.

Han Liu, Kathryn Roeder, and Larry Wasserman. Stability approach to regularization selection (StARS) for high dimensional graphical models. In *NIPS*, pages 1432–1440, 2010.

Han Liu, Fang Han, Ming Yuan, John Lafferty, Larry Wasserman, et al. High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, 2012.

Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. Text classification using string kernels. *The Journal of Machine Learning Research*, 2:419–444, 2002.

Ben London, Theodoros Rekatsinas, Bert Huang, and Lise Getoor. Multi-relational learning using weighted tensor decomposition with modular loss. *arXiv preprint arXiv:1303.1733*, 2013.

Bo Long, Xiaoyun Wu, Zhongfei (Mark) Zhang, and Philip S. Yu. Unsupervised learning on k-partite graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 317–326, New York, NY, USA, 2006. ISBN 1-59593-339-5. doi: 10.1145/1150402.1150439.

Hedibert Freitas Lopes, Dani Gamerman, and Esther Salazar. Generalized spatial dynamic factor models. *Computational Statistics & Data Analysis*, 55(3):1319–1330, 2011.

Joseph Loscalzo, Isaac Kohane, and Albert-László Barabási. Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. *Molecular Systems Biology*, 3(124):124, January 2007. ISSN 1744-4292.

Yong Luo, Dacheng Tao, Bo Geng, Chao Xu, and Stephen J Maybank. Manifold regularized multitask learning for semi-supervised multilabel image classification. *IEEE Transactions on Image Processing*, 22(2):523–536, 2013.

Jaakko Luttinen and Alexander Ilin. Variational Gaussian-process factor analysis for modeling spatio-temporal data. In *Advances in Neural Information Processing Systems*, pages 1177–1185. Curran Associates, Inc., 2009. ISBN 9781615679119.

Michael W. Mahoney and Petros Drineas. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702, 2009. doi: 10.1073/pnas.0803205106.

Ramamurthy Mani, Robert P St.Onge, John L Hartman, Guri Giaever, and Frederick P Roth. Defining genetic interaction. *Proceedings of the National Academy of Sciences*, 105(9):3461–3466, 2008.

Petros Maragos, Patrick Gros, Athanassios Katsamanis, and George Papandreou. Cross-modal integration for performance improving in multimedia: A review. In Petros Maragos, Alexandros Potamianos, and Patrick Gros, editors, *Multimodal Processing and Interaction*, volume 33 of *Multimedia Systems and Applications*, pages 1–46. Springer US, 2008. ISBN 978-0-387-76315-6. doi: 10.1007/978-0-387-76316-3_1.

Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Manolis Kellis, James J Collins, Gustavo Stolovitzky, et al. Wisdom of crowds for robust gene network inference. *Nature Methods*, 9(8):796–804, 2012.

Ryosuke Matsushita and Toshiyuki Tanaka. Low-rank matrix reconstruction and clustering via approximate message passing. In *Advances in Neural Information Processing Systems*, pages 917–925, 2013.

Samuel Maurus and Claudia Plant. Ternary matrix factorization. In *Proceedings of the 14th IEEE International Conference on Data Mining*, pages 400–409. IEEE, 2014.

Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1): 53–71, 2008.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462, 2006.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Donald Metzler and W Bruce Croft. A Markov random field model for term dependencies. In *ACM SIGIR*, pages 472–479, 2005.

Mark M Metzstein, Gillian M Stanfield, and H Robert Horvitz. Genetics of programmed cell death in *C. elegans*: past, present and future. *Trends in Genetics*, 14(10): 410–416, 1998.

Magali Michaut and Gary D Bader. Multiple genetic interaction experiments provide complementary information useful for gene function prediction. *PLoS Computational Biology*, 8(6):e1002559, 2012.

Jae H Min and Young-Chan Lee. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications*, 28 (4):603–614, 2005.

Edward Roshan Miranda, Olga Zhuchenko, Marko Toplak, Balaji Santhanam, Blaž Zupan, Adam Kuspa, and Gad Shaulsky. ABC transporters in D*ictyostelium discoideum* development. *PLoS One*, 8(8):e70040, 2013.

Andriy Mnih and Ruslan Salakhutdinov. Probabilistic matrix factorization. In *NIPS*, pages 1257–1264, 2007.

Shahin Mohammadi, Giorgos Kollias, and Ananth Grama. Role of synthetic genetic interactions in understanding functional interactions among pathways. In *Pacific Symposium on Biocomputing*, pages 43–54, 2012.

Yves Moreau and Léon-Charles Tranchevent. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nature Reviews Genetics*, 13(8):523–536, 2012. ISSN 1471-0064. doi: 10.1038/nrg3253.

Sara Mostafavi and Quaid Morris. Combining many interaction networks to predict gene function and analyze gene lists. *Proteomics*, 12(10):1687–96, 2012. ISSN 1615-9861.

Sara Mostafavi, Debajyoti Ray, David Warde-Farley, Chris Grouios, and Quaid Morris. GeneMANIA: A real-time multiple association network integration algorithm for predicting gene function. *Genome Biol.*, 9 (Suppl 1), S4, 2008.

Peter J Mucha, Thomas Richardson, Kevin Macon, Mason A Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–878, 2010.

Sach Mukherjee and Terence P Speed. Network inference using informative priors. *Proceedings of the National Academy of Sciences*, 105(38):14313–14318, 2008.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT Press, 2012.

Jared S Murray, David B Dunson, Lawrence Carin, and Joseph E Lucas. Bayesian Gaussian copula factor models for mixed data. *Journal of the American Statistical Association*, 108(502):656–665, 2013.

Kunio Nakatsukasa and Jeffrey L Brodsky. The recognition and retrotranslocation of misfolded proteins from the endoplasmic reticulum. *Traffic*, 9(6):861–870, 2008.

M. Nakayama, R. Kikuno, and O. Ohara. Protein-protein interactions between large proteins: two-hybrid screening using a functionally classified library composed of long cDNAs. *Genome Res.*, 12(11):1773–1784, Nov 2002.

Loris Nanni, Alessandra Lumini, and Sheryl Brahnam. A classifier ensemble approach for the missing feature problem. *Artificial Intelligence in Medicine*, 55(1):37–50, 2012.

L. M. Napolitano, E. G. Jaffray, R. T. Hay, and G. Meroni. Functional interactions between ubiquitin E2 enzymes and TRIM proteins. *Biochem. J.*, 434(2):309–319, Mar 2011.

Waleed Nasser, Balaji Santhanam, Edward Roshan Miranda, Anup Parikh, Kavina Juneja, Gregor Rot, Chris Dinh, Rui Chen, Blaž Zupan, Gad Shaulsky, et al. Bacterial discrimination by dictyostelid amoebae reveals the complexity of ancient interspecies interactions. *Current Biology*, 23(10):862–872, 2013.

S.J. Nelson, M. Schopen, A.G. Savage, J.L. Schulman, and N. Arluk. The MeSH translation maintenance system: structure, interface design, and implementation. *Medinfo*, 11(Pt 1):67–69, 2004.

Cancer Genome Atlas Network et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337, 2012.

Cancer Genome Atlas Research Network et al. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474 (7353):609–615, 2011.

PC Newell, RF Henderson, D Mosses, and DI Ratner. Sensitivity to Bacillus subtilis: a novel system for selection of heterozygous diploids of Dictyostelium discoideum. *Journal of General Microbiology*, 100(1):207–211, 1977.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on Machine Learning*, pages 809–816, 2011.

Richard A Notebaart, Philip R Kensche, Martijn A Huynen, Bas E Dutilh, et al. Asymmetric relationships between proteins shape genome evolution. *Genome Biology*, 10(2): R19, 2009.

Shigeyuki Oba et al. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16):2088–2096, 2003.

Jeanne Ellis Ormrod. *Human learning*. Pearson, Upper Saddle River, New Jersey, USA, 2011.

John D Osborne, Jared Flatow, Michelle Holko, Simon M Lin, Warren a Kibbe, Lihua Julie Zhu, Maria I Danila, Gang Feng, and Rex L Chisholm. Annotating the human genome with Disease Ontology. *BMC Genomics*, 10:S6, January 2009. ISSN 1471-2164.

J. Ouyang, Y. Shi, A. Valin, Y. Xuan, and G. Gill. Direct binding of CoREST1 to SUMO-2/3 contributes to gene-specific repression by the LSD1/CoREST1/HDAC complex. *Mol. Cell*, 34(2):145–154, Apr 2009.

Sharmistha Pal, Yingtao Bi, Luke Macyszyn, Louise C Showe, Donald M O'Rourke, and Ramana V Davuluri. Isoform-level gene signature improves prognostic stratification and accurately classifies glioblastoma subtypes. *Nucleic Acids Research*, 42(8):e64, 2014.

Xiao-Yong Pan, Ye Tian, Yan Huang, and Hong-Bin Shen. Towards better accuracy for missing value estimation of epistatic miniarray profiling data by a novel ensemble approach. *Genomics*, 97(5):257–264, 2011.

Gaurav Pandey, Bin Zhang, Aaron N Chang, Chad L Myers, Jun Zhu, Vipin Kumar, and Eric E Schadt. An integrative multi-network and multi-classifier approach to predict genetic interactions. *PLoS Computational Biology*, 6(9), 2010. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000928.

Gaia V Paolini, Richard HB Shapland, Willem P van Hoorn, Jonathan S Mason, and Andrew L Hopkins. Global mapping of pharmacological space. *Nature Biotechnology*, 24(7):805–815, 2006.

Anup Parikh, Edward Roshan Miranda, Mariko Katoh-Kurasawa, Danny Fuller, Gregor Rot, Lan Zagar, Tomaz Curk, Richard Sucgang, Rui Chen, Blaž Zupan, William F Loomis, Adam Kuspa, and Gad Shaulsky. Conserved developmental transcriptomes in evolutionarily divergent species. *Genome Biology*, 11(3):R35, 2010. ISSN 1465-6914. doi: 10.1186/gb-2010-11-3-r35.

Devi Parikh and Robi Polikar. An ensemble-based incremental learning approach to data fusion. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 37(2):437–450, 2007.

Paul Pavlidis, Jason Weston, Jinsong Cai, and William Noble Grundy. Gene functional classification from heterogeneous data. In *Proceedings of the 5th International Conference on Computational Biology*, pages 249–255, New York, NY, USA, 2001. ISBN 1-58113-353-7. doi: 10.1145/369133.369228.

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

J. Pearl and T. Verma. A theory of inferred causation. In *Conference on the Principles of Knowledge Representation and Reasoning*, pages 441–452, 1991.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Jean-François Pessiot, Pui Shan Wong, Toru Maruyama, Ryoko Morioka, Sachiyo Aburatani, Michihiro Tanaka, and Wataru Fujibuchi. The impact of collapsing data on microarray analysis and DILI prediction. *Systems Biomedicine*, 1(3):1–7, 2013.

Hilary Phenix, Katy Morin, Cory Batenchuk, Jacob Parker, Vida Abedi, Liu Yang, Lioudmila Tepliakova, Theodore J Perkins, and Mads Kærn. Quantitative epistasis analysis and pathway inference from genetic interaction data. *PLoS Computational Biology*, 7(5):e1002048, 2011.

Hilary Phenix, Theodore Perkins, and Mads Kærn. Identifiability and inference of pathway motifs by epistasis analysis. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 23(2):025103–025103, 2013.

A. Pichlmair, C. Lassnig, C. A. Eberle, M. W. Gorna, C. L. Baumann, T. R. Burkard, T. Burckstummer, A. Stefanovic, S. Krieger, K. L. Bennett, T. Rulicke, F. Weber, J. Colinge, M. Muller, and G. Superti-Furga. IFIT1 is an antiviral protein that recognizes 5'-triphosphate RNA. *Nat. Immunol.*, 12(7):624–630, Jul 2011.

Rosario M Piro and Ferdinando Di Cunto. Computational approaches to disease-gene prediction: rationale, classification and successes. *The FEBS Journal*, 279(5):678–96, March 2012. ISSN 1742-4658.

Erin D Pleasance, R Keira Cheetham, Philip J Stephens, David J McBride, Sean J Humphray, Chris D Greenman, Ignacio Varela, Meng-Lay Lin, Gonzalo R Ordóñez, Graham R Bignell, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 463(7278):191–196, 2009.

C. Prieto, A. Risueño, C. Fontanillo, and J. De Las Rivas. Human gene coexpression landscape: confident network derived from tissue transcriptomic profiles. *PLoS One*, 3 (12):e3911, 2008.

Shuye Pu et al. Local coherence in genetic interaction patterns reveals prevalent functional versatility. *Bioinformatics*, 24(20):2376–2383, 2008.

Yan Qi et al. Finding friends and enemies in an enemies-only network: a graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions. *Genome Research*, 18(12): 1991–2004, 2008.

Shibin Qiu and Terran Lane. A framework for multiple kernel support vector regression and its applications to sirna efficacy prediction. *IEEE Transactions on Computational Biology and Bioinformatics*, 6(2):190–199, 2009.

Predrag Radivojac, Wyatt T Clark, Tal Ronnen Oron, Alexandra M Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, Christopher Funk, Karin Verspoor, Asa Ben-Hur, et al. A large-scale evaluation of computational protein function prediction. *Nature Methods*, 10: 221–227, 2013.

S. M. Rappaport and M. T. Smith. Environment and disease risks. *Science*, 330(6003):460–461, 2010.

T. Ravasi, H. Suzuki, C. V. Cannistraci, S. Katayama, V. B. Bajic, K. Tan, A. Akalin, S. Schmeier, M. Kanamori-Katayama, N. Bertin, P. Carninci, C. O. Daub, A. R. Forrest, J. Gough, S. Grimmond, J. H. Han, T. Hashimoto, W. Hide, O. Hofmann, A. Kamburov, M. Kaur, H. Kawaji, A. Kubosaki, T. Lassmann, E. van Nimwegen, C. R. MacPherson, C. Ogawa, A. Radovanovic, A. Schwartz, R. D. Teasdale, J. Tegner, B. Lenhard, S. A. Teichmann, T. Arakawa, N. Ninomiya, K. Murakami, M. Tagami, S. Fukuda, K. Imamura, C. Kai, R. Ishihara, Y. Kitazume, J. Kawai, D. A. Hume, T. Ideker, and Y. Hayashizaki. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, 140 (5):744–752, Mar 2010.

Pradeep Ravikumar et al. High-dimensional ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

Dietrich Rebholz-Schuhmann, Anika Oellrich, and Robert Hoehndorf. Text-mining solutions for biomedical research: enabling integrative biology. *Nature Reviews Genetics*, 13(12):829–839, 2012.

M. Refior and K. Mees. Coexistence of bilateral paraganglioma of the A. carotis, thymoma and thyroid adenoma: a chance finding? *Laryngorhinootologie*, 79(6):337–340, Jun 2000.

Sam Reid and Greg Grudic. Regularized linear models in stacked generalization. In *Multiple Classifier Systems*, pages 112–121. Springer, 2009.

A. W. Reinke, J. Baek, O. Ashenberg, and A. E. Keating. Networks of bZIP protein-protein interactions diversified over a billion years of evolution. *Science*, 340(6133): 730–734, May 2013.

Jorge S Reis-Filho and Lajos Pusztai. Gene expression profiling in breast cancer: classification, prognostication, and prediction. *The Lancet*, 378(9805):1812–1823, 2011.

Steffen Rendle. Factorization machines. In *2010 Proceedings of the 10th International Conference on Data Mining*, pages 995–1000. IEEE, 2010.

Steffen Rendle. Scaling factorization machines to relational data. In *Proceedings of the VLDB*, volume 6, pages 337–348. VLDB Endowment, 2013.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pages 452–461. AUAI Press, 2009.

Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 811–820. ACM, 2010.

Steffen Rendle, Zeno Gantner, Christoph Freudenthaler, and Lars Schmidt-Thieme. Fast context-aware recommendations with factorization machines. In *Proceedings of the 34th international ACM Conference on Research and Development in Information Retrieval*, pages 635–644. ACM, 2011.

Achim Rettinger, Hendrik Wermser, Yi Huang, and Volker Tresp. Context-aware tensor decomposition for relation prediction in social networks. *Social Network Analysis and Mining*, 2(4):373–385, 2012.

Assen Roguev et al. Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *Science*, 322(5900):405–410, 2008.

Thomas Rolland et al. A proteome-scale map of the human interactome network. *Cell*, 159(5):1212–1226, 2014.

Pedro A Romero, GJ Peter Dijkgraaf, Serge Shahinian, Annette Herscovics, and Howard Bussey. The yeast CWH41 gene encodes glucosidase I. *Glycobiology*, 7(7):997–1004, 1997.

Frederick P Roth, Howard D Lipshitz, and Brenda J Andrews. Q&A: Epistasis. *Journal of Biology*, 8:35, 2009.

S. P. Rowbotham, L. Barki, A. Neves-Costa, F. Santos, W. Dean, N. Hawkes, P. Choudhary, W. R. Will, J. Webster, D. Oxley, C. M. Green, P. Varga-Weisz, and J. E. Mermoud. Maintenance of silent chromatin through replication requires SWI/SNF-like chromatin remodeler SMARCAD1. *Mol. Cell*, 42(3):285–296, May 2011.

S. J. Roy, I. Glazkova, L. Frechette, C. Iorio-Morin, C. Binda, D. Petrin, P. Trieu, M. Robitaille, S. Angers, T. E. Hebert, and J. L. Parent. Novel, gel-free proteomics approach identifies RNF5 and JAMP as modulators of GPCR stability. *Mol. Endocrinol.*, 27(8):1245–1266, Aug 2013.

Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. CRC Press, 2005.

Colm Ryan et al. Missing value imputation for epistatic MAPs. *BMC Bioinformatics*, 11(1):197, 2010.

Colm Ryan et al. Imputing and predicting quantitative genetic interactions in epistatic MAPs. In *Network Biology*, pages 353–361. Humana Press, 2011.

Hachem Saddiki, Jon McAuliffe, and Patrick Flaherty. GLAD: a mixed-membership model for heterogeneous tumor subtype classification. *Bioinformatics*, 31(2): 225–232, 2014.

Ruslan Salakhutdinov and Andriy Mnih. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *ICML*, pages 880–887, 2008.

Richard S. Savage, Zoubin Ghahramani, Jim E. Griffin, Bernard J. de la Cruz, and David L. Wild. Discovering transcriptional modules by Bayesian data integration. *Bioinformatics*, 26(12):i158–i167, 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq210.

Juliane Schäfer and Korbinian Strimmer. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764, 2005.

Andreas Schlicker, Thomas Lengauer, and Mario Albrecht. Improving disease gene prioritization using the semantic similarity of Gene Ontology terms. *Bioinformatics*, 26 (18):i561–i567, 2010.

B Schölkopf, K Tsuda, and J-P Vert. *Kernel Methods in Computational Biology*. Computational Molecular Biology. MIT Press, Cambridge, MA, USA, 8 2004.

L.M. Schriml, C. Arze, S. Nadendla, Y.W.W. Chang, M. Mazaitis, V. Felix, G. Feng, and W.A. Kibbe. Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Research*, 40(D1):D940–D946, 2012.

Maya Schuldiner, Jutta Metz, Volker Schmid, Vladimir Denic, Magdalena Rakwalska, Hans Dieter Schmitt, Blanche Schwappach, and Jonathan S Weissman. The GET complex mediates insertion of tail-anchored proteins into the ER membrane. *Cell*, 134(4):634–645, 2008.

Maya Schuldiner et al. Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell*, 123(3):507–519, 2005.

Ruth L. Seal, Susan M. Gordon, Michael J. Lush, Mathew W. Wright, and Elspeth A. Bruford. genenames.org: the HGNC resources in 2011. *Nucleic Acids Research*, 39(Database-Issue):514–519, 2011.

Eran Segal, Haidong Wang, and Daphne Koller. Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, 19(suppl 1):i264–i272, 2003.

Amitabh Sharma, Sreenivas Chavali, Rubina Tabassum, Nikhil Tandon, and Dwaipayan Bharadwaj. Gene prioritization in Type 2 Diabetes using domain interactions and network analysis. *BMC Genomics*, 11(1):84, 2010.

John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA, 2004. ISBN 0521813972.

Y. Shi, D. W. Chan, S. Y. Jung, A. Malovannaya, Y. Wang, and J. Qin. A data set of human endogenous protein ubiquitination sites. *Mol. Cell Proteomics*, 10(5): M110.002089, May 2011.

Yue Shi, Alexandros Karatzoglou, Linas Baltrunas, Martha Larson, Alan Hanjalic, and Nuria Oliver. TFMAP: Optimizing MAP for top-n context-aware recommendation. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 155–164. ACM, 2012a.

Yue Shi, Alexandros Karatzoglou, Linas Baltrunas, Martha Larson, Nuria Oliver, and Alan Hanjalic. CLiMF: learning to maximize reciprocal rank with collaborative less-is-more filtering. In *Proceedings of the 6th ACM International Conference on Recommender Systems*, pages 139–146. ACM, 2012b.

Yue Shi, Alexandros Karatzoglou, Linas Baltrunas, Martha Larson, and Alan Hanjalic. GAPfm: Optimal top-n recommendations for graded relevance domains. In *Proceedings of the 22nd ACM International Conference Information & Knowledge Management*, pages 2261–2266. ACM, 2013.

Sunita J Shukla, Ruili Huang, Christopher P Austin, and Menghang Xia. The future of toxicity testing: a focus on *in vitro* methods using a quantitative high-throughput screening platform. *Drug Discovery Today*, 15(23): 997–1007, 2010.

Alejandro Sifrim, Dusan Popovic, Leon-Charles Tranchevent, Amin Ardeshirdavani, Ryo Sakai, Peter Konings, Joris R Vermeesch, Jan Aerts, Bart De Moor, and Yves Moreau. eXtasy: variant prioritization by genomic data fusion. *Nature Methods*, 10(11):1083–1084, 2013.

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

V Simoncini. Computational methods for linear matrix equations. Technical report, Department of Mathematics, University of Bologna, Piazza di Porta San Donato 5, I-40127, January 2014.

Ajit P Singh and Geoffrey J Gordon. Relational learning via collective matrix factorization. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 650–658, 2008a.

Ajit P. Singh and Geoffrey J. Gordon. A unified view of matrix factorization models. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 358–373, 2008b. ISBN 978-3-540-87480-5. doi: 10.1007/978-3-540-87481-2_24.

Ajit P. Singh and Geoffrey J. Gordon. Relational learning via collective matrix factorization. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 650–658, New York, NY, USA, 2008c. ACM. ISBN 978-1-60558-193-4. doi: 10.1145/1401890.1401969.

Ajit P Singh and Geoffrey J Gordon. A Bayesian matrix factorization model for relational data. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, pages 556–563, 2010.

N. Sioutos, S. Coronado, M.W. Haber, F.W. Hartel, W.L. Shaiu, and L.W. Wright. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics*, 40(1): 30–43, 2007.

Fotini N. Skopouli, Urania Dafni, John P.A. Ioannidis, and Haralampos M. Moutsopoulos. Clinical evolution, and morbidity and mortality of primary Sjögren's syndrome. *Seminars in Arthritis and Rheumatism*, 29(5):296 – 304, 2000. ISSN 0049-0172.

Nathan Srebro, Tommi Jaakkola, et al. Weighted low-rank approximations. In *ICML*, volume 3, pages 720–727, 2003.

Nathan Srebro, Jason Rennie, and Tommi S Jaakkola. Maximum-margin matrix factorization. In *Advances in neural information processing systems*, pages 1329–1336, 2004.

Robert P St Onge, Ramamurthy Mani, Julia Oh, Michael Proctor, Eula Fung, Ronald W Davis, Corey Nislow, Frederick P Roth, and Guri Giaever. Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. *Nature Genetics*, 39(2):199–206, 2007.

Chris Stark, Bobby-Joe Breitkreutz, Andrew Chatraryamontri, Lorrie Boucher, Rose Oughtred, Michael S. Livstone, Julie Nixon, Kimberly Van Auken, Xiaodong Wang, Xiaoqi Shi, Teresa Reguly, Jennifer M. Rust, Andrew G. Winter, Kara Dolinski, and Mike Tyers. The BioGRID interaction database: 2011 update. *Nucleic Acids Research*, 39(Database-Issue):698–704, 2011.

Chris Stark et al. BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, 34(suppl 1): D535–D539, 2006.

Sandra Stefanovic and Ramanujan S Hegde. Identification of a targeting factor for posttranslational membrane protein insertion into the ER. *Cell*, 128(6):1147–1159, 2007.

Michael Steinert. Pathogen–host interactions in Dictyostelium, Legionella, Mycobacterium and other pathogens. In *Seminars in Cell & Developmental Biology*, volume 22, pages 70–76. Elsevier, 2011.

Francesco C Stingo and Marina Vannucci. Variable selection for discriminant analysis with Markov random field priors for the analysis of microarray data. *Bioinformatics*, 27 (4):495–501, 2011.

T. W. Su, L. L. Wu, and C. P. Lin. The prevalence of dementia and depression in Taiwanese institutionalized leprosy patients, and the effectiveness evaluation of reminiscence therapy–a longitudinal, single-blind, randomized control study. *International Journal of Geriatric Psychiatry*, 27(2): 187–196, Feb 2012.

Niranjan Subrahmanya and Yung C Shin. Sparse multiple kernel learning for signal processing applications. *IEEE Transcations on Pattern Analysis and Machine Intelligence*, 32(5):788–798, 2010.

Jingchun Sun, Peilin Jia, Ayman H Fanous, Bradley T Webb, Edwin JCG Van den Oord, Xiangning Chen, Jozsef Bukszar, Kenneth S Kendler, and Zhongming Zhao. A multi-dimensional evidence-based candidate gene prioritization approach for complex diseases–schizophrenia as a case. *Bioinformatics*, 25(19):2595–6602, 2009.

Michal A Surma, Christian Klose, Debby Peng, Michael Shales, Caroline Mrejen, Adam Stefanko, Hannes Braberg, David E Gordon, Daniela Vorkel, Christer S Ejsing, et al. A lipid E-MAP identifies Ubx2 as a critical regulator of lipid saturation and lipid bilayer stress. *Molecular Cell*, 51(4):519–530, 2013.

Ilya Sutskever. Modelling relational data using bayesian clustered tensor factorization. In *Advances in Neural Information Processing Systems*, pages 1821–1828, 2009.

Michael Szell, Renaud Lambiotte, and Stefan Thurner. Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences*, 107(31):13636–13641, 2010.

Lei Tang, Xufei Wang, and Huan Liu. Community detection via heterogeneous interaction analysis. *Data Mining and Knowledge Discovery*, 25(1):1–33, 2012.

Wei Tang, Zhengdong Lu, and Inderjit S. Dhillon. Clustering with multiple graphs. In *Proceedings of the 9th IEEE International Conference on Data Mining*, pages 1016–1021, 2009. ISBN 978-0-7695-3895-2. doi: 10.1109/ICDM.2009.125.

Yichuan Tang, Ruslan Salakhutdinov, and Geoffrey Hinton. Tensor analyzers. In *ICML*, 2013.

R. Tarallo, A. Bamundo, G. Nassa, E. Nola, O. Paris, C. Ambrosino, A. Facchiano, M. Baumann, T. A. Nyman, and A. Weisz. Identification of proteins associated with ligand-activated estrogen receptor α in human breast cancer cell nuclei by tandem affinity purification and nano LC-MS/MS. *Proteomics*, 11(1):172–179, Jan 2011.

Vasiliki Theodorou et al. GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility. *Genome Research*, 23(1):12–22, 2013.

D. J. Tiede, A. Tefferi, R. Kochhar, G. B. Thompson, and I. D. Hay. Paraneoplastic cholestasis and hypercoagulability associated with medullary thyroid carcinoma. Resolution with tumor debulking. *Cancer*, 73(3):702–705, Feb 1994.

Claire Tilstone. DNA microarrays: vital statistics. *Nature*, 424(6949):610–612, 2003.

Adrien Todeschini, François Caron, and Marie Chavent. Probabilistic low-rank matrix completion with adaptive spectral regularization algorithms. In *Advances in Neural Information Processing Systems*, volume 26, pages 845–853, 2013.

Amy Hin Yan Tong et al. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, 294 (5550):2364–2368, 2001.

Amy Hin Yan Tong et al. Global mapping of the yeast genetic interaction network. *Science*, 303(5659):808–813, 2004.

Salman Toor, Rainer Toebbicke, Maitane Zotes Resines, and Sverker Holmgren. Investigating an open source cloud storage infrastructure for CERN-specific data analysis. In *IEEE 7th International Conference on Networking, Architecture and Storage*, pages 84–88. IEEE, 2012.

Olga Troyanskaya et al. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001.

Clare Turnbull et al. Gene-gene interactions in breast cancer susceptibility. *Human Molecular Genetics*, 21(4):958–962, 2012.

Takeki Uehara, Atsushi Ono, Toshiyuki Maruyama, Ikuo Kato, Hiroshi Yamada, Yasuo Ohno, and Tetsuro Urushidani. The Japanese toxicogenomics project: Application of toxicogenomics. *Molecular Nutrition and Food Research*, 54:218–227, 2010. doi: 10.1002/mnfr.200900169.

Igor Ulitsky et al. From E-MAPs to module maps: dissecting quantitative genetic interactions using physical interactions. *Molecular Systems Biology*, 4(1), 2008.

Igor Ulitsky et al. Towards accurate imputation of quantitative genetic interactions. *Genome Biology*, 10(12):R140, 2009.

R. Valiyil and L. Christopher-Stine. Drug-related myopathies of which the clinician should be aware. *Current Rheumatology Reports*, 12(3):213–220, 2010.

Mark H. Van Benthem and Michael R. Keenan. Fast algorithm for the solution of large-scale non-negativity-constrained least squares problems. *Journal of Chemometrics*, 18(10):441–450, 2004. ISSN 1099-128X. doi: 10.1002/cem.889.

Nancy Van Driessche, Janez Demšar, Ezgi O Booth, Paul Hill, Peter Juvan, Blaž Zupan, Adam Kuspa, and Gad Shaulsky. Epistasis analysis with global transcriptional phenotypes. *Nature Genetics*, 37(5):471–477, 2005.

Martin H van Vliet, Hugo M Horlings, Marc J van de Vijver, Marcel J T Reinders, and Lodewyk F A Wessels. Integration of clinical and gene expression data has a synergetic effect on predicting breast cancer outcome. *PLoS One*, 7 (7):e40358, 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0040358.

S. J. van Wijk, S. J. de Vries, P. Kemmeren, A. Huang, R. Boelens, A. M. Bonvin, and H. T. Timmers. A comprehensive framework of E2-RING E3 interactions of the human ubiquitin-proteasome system. *Mol. Syst. Biol.*, 5: 295, 2009.

M. Varjosalo, S. Keskitalo, A. Van Drogen, H. Nurkkala, A. Vichalkovski, R. Aebersold, and M. Gstaiger. The protein interaction landscape of the human CMGC kinase group. *Cell Rep*, 3(4):1306–1320, Apr 2013.

Manik Varma and Bodla Rakesh Babu. More generality in efficient multiple kernel learning. In *Proceedings of the 26th International Conference on Machine Learning*, pages 1065–1072. ACM, 2009.

J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *Science*, 291 (5507):1304–1351, 2001.

Mauno Vihinen. No more hidden solutions in bioinformatics. *Nature*, 521(7552):261, 2015.

S. Von Vietinghoff, W. Schneider, F.C. Luft, and R. Kettritz. Crescentic glomerulonephritis and malignancy guilty or guilt by association? *Nephrology Dialysis Transplantation*, 21(11):3324–3326, 2006.

S. A. Wagner, P. Beli, B. T. Weinert, M. L. Nielsen, J. Cox, M. Mann, and C. Choudhary. A proteome-wide, quantitative survey of in vivo ubiquitylation sites reveals widespread regulatory roles. *Mol. Cell Proteomics*, 10(10): M111.013284, Oct 2011.

Chaohui Wang, Nikos Komodakis, and Nikos Paragios. Markov random field modeling, inference & learning in computer vision & image understanding: A survey. *Computer Vision and Image Understanding*, 117(11): 1610–1627, 2013.

Fei Wang, Tao Li, and Changshui Zhang. Semi-supervised clustering via matrix factorization. In *Proceedings of the SIAM International Conference on Data Mining*, pages 1–12, 2008.

Hua Wang, Heng Huang, and Chris H. Q. Ding. Simultaneous clustering of multi-type relational data via symmetric nonnegative matrix tri-factorization. In *Proceedings of the 20th ACM CIKM International Conference on Information and Knowledge Management*, pages 279–284, 2011a. ISBN 978-1-4503-0717-8. doi: 10.1145/2063576.2063621.

Hua Wang, Heng Huang, Chris H. Q. Ding, and Feiping Nie. Predicting protein-protein interactions from multimodal biological data sources via nonnegative matrix tri-factorization. In *Research in Computational Molecular Biology*, volume 7262, pages 314–325. Springer, 2012. ISBN 978-3-642-29626-0.

J. Wang, K. Huo, L. Ma, L. Tang, D. Li, X. Huang, Y. Yuan, C. Li, W. Wang, W. Guan, H. Chen, C. Jin, J. Wei, W. Zhang, Y. Yang, Q. Liu, Y. Zhou, C. Zhang, Z. Wu, W. Xu, Y. Zhang, T. Liu, D. Yu, Y. Zhang, L. Chen, D. Zhu, X. Zhong, L. Kang, X. Gan, X. Yu, Q. Ma, J. Yan, L. Zhou, Z. Liu, Y. Zhu, T. Zhou, F. He, and X. Yang. Toward an understanding of the protein interaction network of the human liver. *Mol. Syst. Biol.*, 7:536, 2011b.

Xiaoyue Wang, Audrey Q Fu, Megan E McNerney, and Kevin P White. Widespread genetic epistasis among cancer genes. *Nature Communications*, 5:4828, 2014.

Yanli Wang, Jewen Xiao, Tugba O Suzek, Jian Zhang, Jiyao Wang, and Stephen H Bryant. Pubchem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Research*, 37(suppl 2):W623–W633, 2009.

Yu-Xiong Wang and Yu-Jin Zhang. Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on Knowledge and Data Engineering*, 25(6):1336–1353, 2013.

Y. K. Wen and M. L. Chen. Crescentic glomerulonephritis associated with miliary tuberculosis. *Clinical Nephrology*, 71(3):310–313, Mar 2009.

Harm-Jan Westra, Ritsert C Jansen, Rudolf SN Fehrmann, Gerard J te Meerman, David Van Heel, Cisca Wijmenga, and Lude Franke. MixupMapper: correcting sample mixups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics*, 27(15):2104–2111, 2011.

M. H. Willemsen, J. H. Rensen, H. M. van Schrojenstein-Lantman de Valk, B. C. Hamel, and T. Kleefstra. Adult phenotypes in Angelman- and Rett-Like syndromes. *Molecular Syndromology*, 2(3-5):217–234, Apr 2012.

Christopher K.I. Williams and Carl Edward Rasmussen. Gaussian processes for regression. In *Advances in Neural Information Processing Systems*, pages 514–520, 1996.

Gwendolyn M Wilmes et al. A genetic interaction map of RNA-processing factors reveals links between Sem1/Dss1-containing complexes and mRNA export and splicing. *Molecular Cell*, 32(5):735–746, 2008.

David H Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

K. A. Wong, J. Wilson, A. Russo, L. Wang, M. N. Okur, X. Wang, N. P. Martin, E. Scappini, G. K. Carnegie, and J. P. O'Bryan. Intersectin (ITSN) family of scaffolds function as molecular hubs in protein interaction networks. *PLoS ONE*, 7(4):e36023, 2012.

Sharyl L Wong et al. Combining biological networks to predict genetic interactions. *PNAS*, 101(44):15682–15687, 2004.

N. T. Woods, R. D. Mesquita, M. Sweet, M. A. Carvalho, X. Li, Y. Liu, H. Nguyen, C. E. Thomas, E. S. Iversen, S. Marsillac, R. Karchin, J. Koomen, and A. N. Monteiro. Charting the landscape of tandem BRCT domain-mediated protein interactions. *Sci Signal*, 5(242):rs6, 2012.

Chuanhua Xing and David B Dunson. Bayesian inference for genomic data integration reduces misclassification rate in predicting protein-protein interactions. *PLoS Computational Biology*, 7(7):e1002110, 2011. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002110.

Zenglin Xu, Feng Yan, and Yuan Qi. Bayesian nonparametric models for multiway data analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99:1–14, 2014.

C Yang, CH Hasselgren, S Boyer, K Arvidson, S Aveston, P Dierkes, R Benigni, RD Benz, J Contrera, NL Kruhlak, et al. Understanding genetic toxicity through data mining: the process of building knowledge by integrating multiple genetic toxicity databases. *Toxicology Mechanisms and Methods*, 18(2-3):277–295, 2008.

Eunho Yang, Genevera Allen, Zhandong Liu, and Pradeep K Ravikumar. Graphical models via generalized linear models. In *NIPS*, pages 1358–1366, 2012.

Eunho Yang, Pradeep Ravikumar, Genevera I Allen, and Zhandong Liu. On Poisson graphical models. In *NIPS*, pages 1718–1726, 2013.

Hongxia Yang and Jingrui He. Learning with dual heterogeneity: a nonparametric bayes model. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 582–590. ACM, 2014.

Jaewon Yang and Jure Leskovec. Defining and evaluating network communities based on ground-truth. In *ACM MDS*, 2012.

Jingjing Yang, Yuanning Li, Yonghong Tian, Lingyu Duan, and Wen Gao. Group-sensitive multiple kernel learning for object categorization. In *Proceedings of the 12th IEEE International Conference on Computer Vision*, pages 436–443. IEEE, 2009.

Jingjing Yang, Yuanning Li, Yonghong Tian, Ling-Yu Duan, and Wen Gao. Per-sample multiple kernel approach for visual concept learning. *Journal on Image and Video Processing*, 2010:2, 2010.

Jieping Ye, Shuiwang Ji, and Jianhui Chen. Multi-class discriminant kernel learning via convex programming. *J. Mach. Learn. Res.*, 9:719–758, 2008. ISSN 1532-4435.

Kai Yu, Volker Tresp, and Anton Schwaighofer. Learning gaussian processes from multiple tasks. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 1012–1019. ACM, 2005.

Shi Yu, Tillmann Falck, Anneleen Daemen, Leon-Charles Tranchevent, Johan Ak Suykens, Bart De Moor, and Yves Moreau. L_2-norm multiple kernel learning and its application to biomedical data fusion. *BMC Bioinformatics*, 11:309, 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-309.

Shi Yu, Leon Tranchevent, Xinhai Liu, Wolfgang Glanzel, Johan A. K. Suykens, Bart De Moor, and Yves Moreau. Optimized data fusion for kernel k-means clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(5):1031–1039, 2012. ISSN 0162-8828. doi: 10.1109/TPAMI.2011.255.

Ming Yuan. Efficient computation of ℓ_1 regularized estimates in Gaussian graphical models. *Journal of Computational and Graphical Statistics*, 17(4):809–826, 2008.

Yinyin Yuan, Richard S Savage, and Florian Markowetz. Patient-specific data fusion defines prognostic cancer subtypes. *PLoS Computational Biology*, 7(10):e1002227, 2011.

A. Zanon, A. Rakovic, H. Blankenburg, N. T. Doncheva, C. Schwienbacher, A. Serafin, A. Alexa, C. X. Weichenberger, M. Albrecht, C. Klein, A. A. Hicks, P. P. Pramstaller, F. S. Domingues, and I. Pichler. Profiling of Parkin-binding partners using tandem affinity purification. *PLoS ONE*, 8(11):e78648, 2013.

Dan Zhang, Jingrui He, Yan Liu, Luo Si, and Richard Lawrence. Multi-view transfer learning with a large margin approach. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1208–1216. ACM, 2011a.

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

Shihua Zhang, Chun-Chi Liu, Wenyuan Li, Hui Shen, Peter W. Laird, and Xianghong Jasmine Zhou. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Research*, 40(19): 9379–9391, 2012. doi: 10.1093/nar/gks725.

Xiao-Dong Zhang and Xing-Ming Zhao. Computational approaches for identifying signaling pathways from molecular interaction networks. *Current Bioinformatics*, 8 (1):56–62, 2013.

Yongmian Zhang and Qiang Ji. Active and dynamic information fusion for multisensor systems with dynamic Bayesian networks. *Trans. Sys. Man Cyber. Part B*, 36(2):467–472, 2006. ISSN 1083-4419. doi: 10.1109/TSMCB.2005.859081.

Jiashun Zheng et al. Epistatic relationships reveal the functional organization of yeast transcription factors. *Molecular Systems Biology*, 6(1), 2010.

Dengyong Zhou and Christopher J. C. Burges. Spectral clustering and transductive learning with multiple views. In *Proceedings of the 24th International Conference on Machine Learning*, pages 1159–1166, 2007. ISBN 978-1-59593-793-3. doi: 10.1145/1273496.1273642.

Kemin Zhou, Kaoru Takegawa, Scott D Emr, and Richard A Firtel. A phosphatidylinositol (PI) kinase gene family in Dictyostelium discoideum: biological roles of putative mammalian p110 and yeast Vps34p PI 3-kinase homologs during growth and development. *Mol Cell Biol*, 15(10):5645–5656, 1995.

Qiusha Zhu, Zhao Li, Haohong Wang, Yimin Yang, and Mei-Ling Shyu. Multimodal sparse linear integration for content-based item recommendation. In *IEEE International Symposium on Multimedia*, pages 187–194. IEEE, 2013.

Xiaojin Zhu, Jaz Kandola, Zoubin Ghahramani, and John D Lafferty. Nonparametric transforms of graph kernels for semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 1641–1648, 2004.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

Blaž Zupan, Janez Demšar, Ivan Bratko, Peter Juvan, John A Halter, Adam Kuspa, and Gad Shaulsky. GenePath: a system for automated construction of genetic networks from mutant data. *Bioinformatics*, 19(3):383–389, 2003.

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

Marinka Žitnik and Blaž Zupan. Matrix factorization-based data fusion for gene function prediction in baker's yeast and slime mold. In *Pacific Symposium on Biocomputing*, volume 19, pages 400–412, 2014a. doi: 10.1142/9789814583220_0038.

Marinka Žitnik and Blaž Zupan. Data fusion by matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):41–53, 2015a. doi: 10.1109/TPAMI.2014.2343973.

Marinka Žitnik and Blaž Zupan. Collective pairwise classification for multi-way analysis of disease and drug data. In *Pacific Symposium on Biocomputing*, volume 21, 2016.

Marinka Žitnik and Blaž Zupan. Matrix factorization-based data fusion for drug-induced liver injury prediction. *Systems Biomedicine*, 2(1):16–22, 2014b. doi: 10.4161/sysb.29072.

Marinka Žitnik and Blaž Zupan. Gene network inference by probabilistic scoring of relationships from a factorized model of interactions. *Bioinformatics*, 30(12):i246–i254, 2014c. doi: 10.1093/bioinformatics/btu287.

Marinka Žitnik and Blaž Zupan. Imputation of quantitative genetic interactions in epistatic maps by interaction propagation matrix completion. In *Research in Computational Molecular Biology*, pages 448–462. Springer, 2014d.

Marinka Žitnik and Blaž Zupan. Gene network inference by fusing data from diverse distributions. *Bioinformatics*, 31(12):230–239, 2015b. doi: 10.1093/bioinformatics/btv258.

Marinka Žitnik and Blaž Zupan. Inter-relation sensitivity estimation in collective matrix factorization. *Journal of Machine Learning Research*, In submission, 2015c.

Marinka Žitnik and Blaž Zupan. Data imputation in epistatic maps by network-guided matrix completion. *Journal of Computational Biology*, 2015d. doi: 10.1089/cmb.2014.0158.

Marinka Žitnik and Blaž Zupan. Survival regression by data fusion. *Systems Biomedicine*, 2(3):49–55, 2015e. doi: 10.1080/21628130.2015.1016702.

Marinka Žitnik, Vuk Janjić, Chris Larminie, Blaž Zupan, and Nataša Pržulj. Discovering disease-disease associations by fusing systems-level molecular data. *Scientific Reports*, 3: e3202, 2013. doi: 10.1038/srep03202.

Marinka Žitnik, Edward Nam A., Christopher Dinh, Adam Kuspa, Gad Shaulsky, and Blaž Zupan. Gene prioritization by compressive data fusion and chaining. *PLoS Computational Biology*, 11(10):e1004552, 2015a. doi: 10.1371/journal.pcbi.1004552.

Marinka Žitnik, Rok Sosič, and Jure Leskovec. Ranking network communities by relevancy. *In review*, 2015b.