EXPLOITING NOISY AND INCOMPLETE BIOLOGICAL DATA FOR PREDICTION AND KNOWLEDGE DISCOVERY

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus Prof.ir. K.C.A.M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op donderdag 7 oktober 2010 om 12.30 uur
door

Yunlei LI

elektrotechnisch ingenieur
geboren te Sichuan, China

Dit proefschrift is goedgekeurd door de promotor:
Prof. dr. ir. M.J.T. Reinders

Copromotor:
Dr. ir. D. de Ridder

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. ir. M.J.T. Reinders,	Technische Universiteit Delft, promotor
Dr. ir. D. de Ridder,	Technische Universiteit Delft, copromotor
Prof. dr. ir. J.J. Heijnen,	Technische Universiteit Delft
Prof. dr. A. van Kampen,	Universiteit van Amsterdam
Prof. dr. M.A. Huynen,	Radboud Universiteit Nijmegen
Prof. dr. J. Heringa,	Vrije Universiteit Amsterdam
Prof. dr. A.P.J.M. Siebes,	Universiteit Utrecht
Prof. dr. ir. A.P. de Vries,	Technische Universiteit Delft, reserve lid

nbic Netherlands
Bioinformatics
Centre

This work was part of the BioRange program of the Netherlands Bioinformatics Center (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).

ISBN 978-94-90818-03-6
Published by TU Delft Mediamatica
Printed by Universal Press - NL

Copyright © 2010 by Yunlei Li
All rights reserved. No part of this thesis may be reproduced, transmitted, or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without prior written permission of the copyright owner.

To my family

CONTENTS

1. Introduction . . . 1
- 1.1 Background . . . 1
  - 1.1.1 Computational methods to handle noisy data . . . 2
  - 1.1.2 Computational methods to handle incomplete data . . . 2
- 1.2 Scope . . . 3
- 1.3 Outline . . . 6

2. Classification in the presence of class noise . . . 9
- 2.1 Introduction . . . 10
- 2.2 The Lawrence and Schölkopf model . . . 11
  - 2.2.1 Modifications to the Lawrence and Schölkopf model . . . 12
- 2.3 Probabilistic Kernel Fisher and Probabilistic Fisher . . . 13
  - 2.3.1 Kernel Fisher discriminant . . . 14
  - 2.3.2 Probabilistic Kernel Fisher . . . 15
  - 2.3.3 Probabilistic Fisher . . . 17
- 2.4 Experiments . . . 17
  - 2.4.1 Simulated data sets . . . 17
  - 2.4.2 CGH data set . . . 18
  - 2.4.3 Results on the simulated data sets . . . 19
  - 2.4.4 Results on the CGH data set . . . 20
- 2.5 Conclusions . . . 20

3. Classification using prior knowledge on measurement noise . . . 23
- 3.1 Introduction . . . 24
- 3.2 Integrative Kernel Method . . . 25
  - 3.2.1 Theoretical basis . . . 25
  - 3.2.2 Transformation model . . . 27
- 3.3 Elliptical kernel methods . . . 29
- 3.4 Data . . . 30
  - 3.4.1 Protein complex prediction . . . 30
  - 3.4.2 Data set . . . 31
  - 3.4.3 Prior knowledge on measurement noise . . . 31
- 3.5 Results . . . 32
- 3.6 Conclusions . . . 37

4. M-Pal: Aligning metabolic pathways between species 39
4.1 Introduction 40
4.2 Method 42
4.2.1 Reaction retrieval 42
4.2.2 Building block alignment 43
4.2.3 Pathway assembly 44
4.2.4 Scoring function 46
4.3 Results and discussion 47
4.4 Conclusions 51

5. M-PAS: Measuring metabolic pathway similarities 53
5.1 Introduction 54
5.2 Method 56
5.2.1 Reaction representation 56
5.2.2 Reaction alignment 56
5.2.3 Scoring function 58
5.2.4 Pathway construction 61
5.3 Results and discussion 62
5.3.1 The scoring function can address different biological questions 62
5.3.2 Comparing results of different queries can help infer additional details 64
5.3.3 Combining the component scores makes sense 65
5.3.4 The conserved part of two aligned networks is scale-free 66
5.3.5 Short pathways lead to interpretable results 66
5.3.6 M-PAS reveals pathway diversity and alternatives 67
5.3.7 New links between different parts of metabolism are found 70
5.3.8 Primary metabolism is highly conserved 70
5.4 Conclusions 71

6. RM-PAS: Aligning regulatory-metabolic pathways 73
6.1 Introduction 74
6.2 Method 75
6.2.1 Regulatory-metabolic building blocks 75
6.2.2 Pathway assembly 77
6.2.3 Scoring function 77
6.3 Data 79
6.4 Experiments and results 79
6.4.1 Identifying conserved regulatory-metabolic network elements 80
6.4.2 Using one level to infer missing information at another level 82
6.4.3 Revealing the differences between two levels 83
6.5 Conclusions 84
6.6 Appendix A: Permutation test 84

6.7 Appendix B: Commentary . . . 85
  6.7.1 Experiments & results . . . 85
  6.7.2 Method . . . 87

7. Discussion . . . 89

Bibliography . . . 93

Summary . . . 103

Samenvatting . . . 105

Acknowledgements . . . 107

Curriculum vitae . . . 109

1

INTRODUCTION

1.1 Background

Molecular biology aims to globally and systematically characterize the basic properties of gene products and their interactions under certain biological conditions, so that we can gain a comprehensive understanding of the basic mechanisms underlying cellular behavior. New measurement techniques have revolutionized molecular biology over the last decade. Various high-throughput (HT) technologies simultaneously quantify thousands of interdependent biological variables of complex biomolecular systems at various levels, from the level of a cell to that of the whole organism [104, 130]. The available techniques enable us, among others, to measure genome-wide transcription levels [155], protein abundance [41], protein-protein interactions [145], protein-DNA interactions [121], subcellular localizations of proteins [79], gene deletion phenotypes [154], DNA copy number variations [76], etc. However, experimental data are often *noisy* and *incomplete*, which poses problems for data interpretation, model construction and prediction generation.

By *noise* we mean any deviation of the measurements from the true values. These measures can be about class labels or attribute features, and can be discrete or continuous. Noise generates *variation* in the measurement values. When the measurement is corrupted by systematic noise, a measurement *bias* is introduced. Each technology used has a different type and degree of noise [59]. For example, gene expression data obtained from microarrays suffer from noise related to hybridization and readout steps [144], while protein-protein interaction and protein-DNA interaction screens contain a large number of false positives and false negatives [21, 149].

Incompleteness is caused by the fact that not everything is or can be measured. Although these new technologies provide massive amounts of data, there may be levels of cellular activities that remain unknown. Even for the known levels, we still miss a lot of information. For example, many metabolites and enzymes have not been identified [73, 104]. As another example, von Mering *et al.* estimate that less than one-third of the complete set of protein-protein interactions has been discovered for *Saccharomyces cerevisiae* [149]. Even sequence homology-based methods fail to assign functions to a considerable fraction (30-80%) of genes in completely sequenced genomes [61], and have been known to produce incomplete or imprecise annotations [23, 62].

The tremendous amount of noisy and incomplete biological data formulates a formidable task for bioinformatics to develop computational strategies and frameworks that filter, organize, and interpret data into models which describe cellular functions. Such models can be used to systematically generate hypotheses [75, 110] and direct biological knowledge discovery. Many of these computational techniques utilize methods developed in statistical learning, data mining, and artificial intelligence [63].

1.1.1 Computational methods to handle noisy data

When multiple data sets contaminated by noise are available, simply taking their average (for continuous values) or intersection (for categorical values) does not remove random noise completely, and in the mean time sacrifices either sensitivity or specificity [102, 149]. Therefore, many studies try to integrate these data probabilistically by giving them different weights depending on their noise levels, using Bayesian approaches [65], kernel-based methods [69, 82], and other integration methods such as Fisher's χ^2 [59].

1.1.2 Computational methods to handle incomplete data

Given a variety of data sources, a properly designed integration approach is expected to not only reduce noise, but also reduce incompleteness of the combined data [4, 59, 64]. This is because different experiments provide complementary perspectives of the biological system. Thus, their integration offers a more detailed and comprehensive picture. In particular, two techniques have proved to be successful in integrating incomplete data, namely an *evolution-based* approach and a *network-based* approach.

Evolution-based approach

Based on similarities across species, we can generate hypotheses to recover missing information in relatively poorly characterized species [130]. This approach has been used to identify candidate enzymes for certain functions [47], predict transcription factor binding sites [80], and infer metabolic pathways [18, 68, 129]. Evolutionary conservation can also serve as a powerful criterion to distinguish signals from random noise, because noise is unlikely to reproduce in multiple species. For instance, observing coregulation of a pair of genes over large evolutionary distances implies that it confers a selective advantage during evolution and that the genes therefore are likely functionally related [137].

Network-based approach

Cells function due to the interactions between the myriad of biomolecules they produce. To understand any biological function, we therefore not only need the information of individual molecules, but also the information about their relationships [5, 106]. This information can be gathered from various experiments, and be represented by networks, e.g. protein-protein interaction, regulatory and metabolic networks. Computational methods using networks place the individual molecule in a context (i.e. relationship with other molecules), and as such enable improved functional annotation [10, 73], visualization, systematic analysis of the network properties [43], operon prediction [143, 162, 163], and phenotype prediction [111, 132]. Furthermore, comparison and integration can be done at the network level, for different species, conditions, time points [130], or interaction types [27, 71, 159].

Evolution & network-based approach

Often, the evolution-based approach can also be applied at a network level. That is, networks can be compared by aligning them across multiple species, so as to evaluate various hypotheses concerning evolution, and to predict unknown functions or interactions from the results [131]. Numerous studies exploit network similarities between different species for detecting conservation [116], drug target identification [48, 134], predicting novel networks or parts of networks [34, 83, 96, 115], biotechnological application design [19], and phylogenetic tree reconstruction [13, 32, 33, 53].

1.2 Scope

In this thesis, we try to exploit noisy and incomplete biological data to improve classification (prediction) and knowledge discovery via different computational approaches. The thesis is divided into two parts. Part I focuses on the noise problem in a classification setting, in order to improve the class label prediction of the biological instances themselves or of the relationships between them. In Part II, we study the noisy and incomplete data of biomolecules and their interactions within a network and evolutionary context, to facilitate knowledge discovery. Fig. 1.1 depicts the two situations.

Part I

In the first part we pursue a theoretical investigation to build noise-tolerant two-class classifiers. Noise is normally categorized into two types, namely

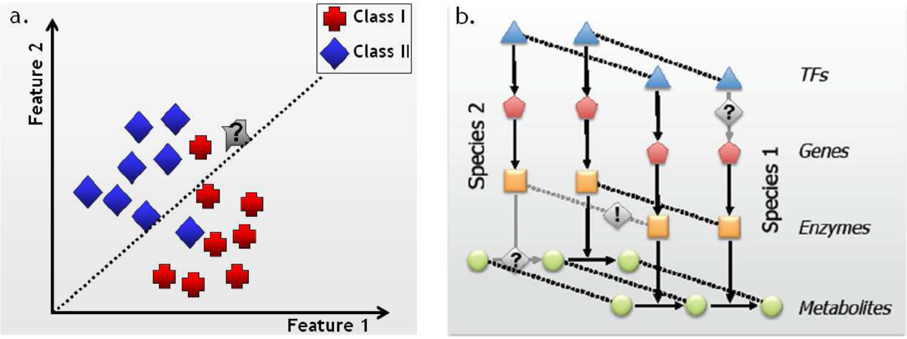

Figure 1.1: Illustration of the computational approaches in this thesis. **a)** Part I. Here nodes represent tumor samples (in the class noise problem) or protein pairs (in the measurement noise problem). A classifier is inferred given the training data to predict the class label of the test data, indicated by a question mark. **b)** Part II. There are four types of nodes, representing transcription factors, enzyme-coding genes, enzymes, and metabolites. Solid lines connecting nodes represent interactions/relationships, e.g. transcriptional regulation or metabolic reaction. Dotted lines represent similarities between nodes. By placing the nodes in a network context across species, we aim to predict the information indicated by a question mark, and discover knowledge about the similarities highlighted by an exclamation mark.

class noise and *attribute noise* [156]. Class noise occurs when the training samples are incorrectly labelled. If attribute noise is introduced by inaccurate measurements, we call it *measurement noise*, which can be further divided into *systematic* and *random* measurement noise [103]. We address both random class noise and random measurement noise in this part.

Noise-contaminated training data undoubtedly deteriorates the accuracy of the classifiers built upon. Techniques handling noisy data can be grouped into data cleaning (i.e. detect and eliminate noisy training samples prior to classifier construction) [49, 117, 167] and noise-tolerant classifier construction [81, 100, 148]. Our research falls into the latter category, which avoids the potential risk of removing precious good samples while keeping the bad ones.

We aim to build classifiers that are robust when the number of samples is small compared to the number of features. This is often the case in high-throughput experiments, which offer genome-wide measurements as features, but for which samples are scarce or expensive to acquire (e.g. tumor samples). This so called *small sample size* problem forms a major challenge [25, 119]. Since only a few samples are given, it becomes impossible for the classifiers to estimate a large number of parameters characterizing the high-dimensional data distribution, introducing over-fitting to the training data and resulting in poor generalization performance.

For the class noise problem, we adopt the model of Lawrence and Schölkopf [81], which casts the class noise generation process probabilistically. That means, whether each sample belongs to a certain class is expressed by a probability. The goal is to calculate these probabilities for all samples, so that we can recover the underlying real distribution of each class and build a more accurate classifier.

We contribute by extending their model in three ways. First, the distribution assumption previously made, i.e. that class conditional variances should be equal, is relaxed. Second, we present a novel incorporation of the noise model in the Kernel Fisher discriminant and standard Fisher discriminant, which offer improved performance in some scenarios. Third, our algorithms achieve a large performance gain on non-Gaussian data sets and data sets with relatively large numbers of features compared to their sample sizes.

In the measurement noise problem, the measured (observed) value randomly deviates from the true value. It may be due to the sample being measured (variability of the sample), the type of measurement technique (precision of the technique), or the measured feature value itself (e.g. high protein abundance is generally more precisely measured [112]). Using training data corrupted by measurement noise directly to build classifier is inaccurate, because the data distribution is simply distorted. To address this problem, most current statistical pattern recognition methods assume independent and identically distributed measurement noise, treating all samples, features, and feature values equally. Only a few studies address the reliabilities of different experimental techniques [59]. Therefore, our goal was to design a classifier which can incorporate the diverse noise levels for different instances.

Our main contribution to this problem is that we have explicitly specified the manner of incorporating prior knowledge on measurement noise, for individual samples, features, and feature values, in kernel-density based classifiers. We chose this type of classifier because of its interpretability and ability to incorporate the noise level easily. We show that including prior knowledge is especially beneficial in a relatively under-sampled data set when compared to the number of features.

Part II

In this part, we pursue a practical study to integrate noisy and incomplete data of biomolecules and their interactions within a network and evolutionary context. We are interested in metabolic reactions (consisting of metabolites and enzymes) and the underlying transcriptional regulation of these enzymes. These interactions are not isolated, but are intertwined with others to form pathways (series of connected interactions) and networks (collections of pathways), such as regulatory networks and metabolic networks. Our knowledge about these networks is still developing, as some information is still missing or unreliable, e.g. information about the regulatory binding and presence of enzymes or reactions.

Our goal is to exploit the available partial information, using an evolutionary network approach. That is, we integrate networks at different levels (transcriptional regulatory and metabolic) and compare the integrated networks across species. Figure 1.1b gives an illustration. By doing so, we aim to not only provide a more comprehensive view of the cellular system, but also to generate more reliable information and hypotheses [55, 130]. This is because when an interaction or a series of interactions are observed at multiple levels and/or across multiple species, we can be more confident about the reliability of the observations.

Our approach is one of the first attempts to conduct a systematic alignment of the full metabolic networks of multiple species, rather than parts of conventional networks (e.g. KEGG pathways). The core contributions lie in our novel alignment and scoring frameworks. That is, we align all reactions in entire metabolic networks of two species and assemble them into pathways, taking mismatches (different reactions with similar or dissimilar enzymes), gaps (different numbers of reactions in two species) and crossovers (different sequential order of the transformations) into account. To prioritize the resulting pathways, we have developed a comprehensive and flexible scoring function for pathway similarity that combines all relevant and uncorrelated information sources. Together, this allows us to make predictions although the information is only partially available.

1.3 Outline

Part I

Chapter 2 presents three noise-tolerant classifiers for training data contaminated with class noise, i.e. Probabilistic Kernel Fisher (PKF), Probabilistic Fisher (PF), and Component-based Probabilistic Algorithm (CPA). They are based on a probabilistic model proposed by Lawrence and Schölkopf [81], in which class labels are represented by probabilities and optimized. We apply this general idea to the Bayes classifier, Fisher discriminant, and Kernel Fisher discriminant. We test the algorithms on several simulated noisy data sets with different distributions, sizes and noise levels, and on a comparative genomic hybridization (CGH) data set. The results show that the proposed approaches substantially improve standard classifiers in noisy data sets, and achieve larger performance gain in non-Gaussian data sets and small sample size data sets.

This chapter was published in *Pattern Recognition*, 2007 [88].

In Chapter 3, we investigate the benefit of incorporating prior knowledge about measurement noise into classifier construction. A new kernel density based classifier, called the Integrative Kernel Method, is proposed. Instead of using an identical spherical kernel for each sample, we use the prior knowledge to set a distinct kernel and weight for each sample, distinguishing between different levels of measurement precision and sample importance. The integration procedure is straightforward and easy to interpret. We show how to estimate the diverse measurement noise levels in a protein complex prediction data set. Compared to standard methods, the new kernel density classifier can yield a significantly better classification performance, particularly for data sets suffering from the small sample size problem.

This chapter was published in *Pattern Recognition*, 2008 [87].

Part II

Chapter 4 presents a comparative analysis of metabolic reaction networks between different species. Our method, Metabolic Pathway ALignment (M-Pal), systematically investigates full metabolic networks of *S. cerevisiae* and *E. coli* at the same time, with the goal of identifying highly similar yet non-identical pathways which perform the same metabolic function, i.e. the transformation of a specific substrate into a certain end product via similar reactions. To this end, we first align two to four similar reactions in two species into so called building blocks, and then assemble these into pathways of a desired length. In each building block, a specific substrate is transformed into a specific product via similar but not necessarily identical reactions in the two species. We also propose a scoring scheme which prioritizes the results according to functional and sequence similarity of the enzymes involved. The analysis helps to gain insight in the biological differences between species and provides comprehensive information on diversity in pathways between species and alternative pathways within species, which is potentially useful for pharmaceutical and industrial bioengineering targets.

This chapter was published in the *Series on Advances in Bioinformatics & Computational Biology*, 2008 [86].

In Chapter 5, Metabolic Pathway Alignment and Scoring (M-PAS) extends our work in Chapter 4. We propose a novel scoring method to quantify the level of conservation in a comprehensive and flexible manner, such that we can focus on different pathways given different biological motivations. This similarity measure compares all components of two pathways by measuring similarities between substrate sets, product sets, enzyme functions, enzyme sequences, and alignment topology. These individual similarity measures are integrated into a single score. It has a hierarchical and generic form, and is capable of measuring pathway similarity given different biological emphases.

This chapter was published in *BMC Systems Biology*, 2008 [85].

In Chapter 6, the alignment framework introduced in Chapter 4 and the scoring function proposed in Chapter 5 are further enhanced. Here we present a more comprehensive method, RM-PAS, that searches for network elements that are conserved in evolution at both the regulatory and metabolic level, and measures the extent of this conservation. RM-PAS extends the building block construction and the scoring function to include transcriptional regulation information. That is, for each enzyme in a reaction, we add the transcription factors that regulate the enzyme-coding genes. We demonstrate how RM-PAS can be applied to identify conserved regulatory-metabolic network elements, infer missing reactions, prioritize and corroborate TF-gene binding hypotheses, and reveal diverse regulation in pathways that are conserved at metabolic level.

This chapter was published in *Proceedings of the* 8^{th} *Annual International Conference on Computational Systems Bioinformatics*, Stanford, USA, 2009 [84].

2

CLASSIFICATION IN THE PRESENCE OF CLASS NOISE

In machine learning, class noise (i.e. noise in the labelling of objects) occurs frequently and deteriorates classifiers derived from noisy data sets. This chapter presents three promising classifiers for this problem based on a probabilistic model proposed by Lawrence and Schölkopf [81]. The proposed algorithms are able to tolerate class noise, and extend the earlier work of Lawrence and Schölkopf. First, we present a novel incorporation of their probabilistic noise model in the standard Fisher discriminant and the Kernel Fisher discriminant. Second, the distribution assumption previously made is relaxed in our work. The methods were evaluated on diverse simulated noisy data sets and a real world comparative genomic hybridization (CGH) data set. The results show that the proposed approaches substantially improve performance of standard classifiers on noisy data sets, and achieve larger performance gain on non-Gaussian data sets and small size data sets.

This chapter was published in *Pattern Recognition*, 2007 [88].

2.1 Introduction

In inductive machine learning, it is quite frequent for noise to be introduced into a data set. Due to the fact that the noise is unlikely to be completely excluded, the inferences derived from the data set may become less reliable. The study of effective noise handling is therefore of great importance. Generally, there are two types of noise, namely attribute noise and class noise [156]. Class noise usually means the erroneous labelling of the training examples. As summarized by Brodley [8], class noise can occur for several reasons including subjectivity, data-entry error, or inadequacy of the information used to label each object. This chapter focuses on the class noise problem.

There are a large number of possible solutions to deal with the existence of class noise. Data cleaning, detection, and elimination of mislabelled training examples prior to classifier induction may increase the classification accuracy [117]. The early approaches tried to remove the instances misclassified by some form of nearest neighbor algorithm [20, 38, 152]. Brodley and Freidl [8] cast this problem into a filtering framework and employed an ensemble of classifiers that served as both filter and final classifier. Different criteria were proposed to identify the mislabelled samples. For example, Guyon *et al.* [49] removed the noisy instances with high information gain and checked them further by a human expert. While the saturation filter [36] assessed the CLCH (complexity of the least complex correct hypothesis) value reduction. In other methods, potential noisy data were pruned either by C4.5 [67, 167] or by neural network [161].

Due to the potential risk of data cleaning when noisy examples are retained while good examples are removed, in which cases the reduced training set can be much less accurate than the full training set, efforts have been taken to construct noise tolerant classifiers directly. Mingers [100] used rule truncation and tree pruning to reduce the chance of over-fitting to noise. A boosting algorithm [118, 125] avoided the noise influence on constructing the classifier via combining a set of classifiers' predictions by voting. To improve the decision-tree approach, some noise-tolerant Occam algorithms were applied [123]. Later on, the decision tree was enhanced to process the training sets with labels specified by belief functions [148].

Among the numerous solutions, the algorithm introduced by Lawrence and Schölkopf [81] has a sound theoretical foundation and elegantly includes the class noise in a generative model. However, it remains unclear how to apply the proposed probabilistic model in more complicated data sets, which cannot be characterized by one Gaussian distribution for each class. Furthermore, their method constrains the class conditional variances to be equal. Here we propose a new method, probabilistic Kernel Fisher (PKF), which extends the previous work to non- Gaussian data sets by providing an explicit implementation of the probabilistic model to Kernel Fisher discriminant (KFD) in a projected space. In addition, we present a simpler version of PKF, probabilistic Fisher (PF), which enables the standard Fisher discriminant to tolerate class noise in linearly separable data sets. We evaluate all these approaches on diverse data sets of different distributions and sizes.

The remainder of this chapter is organized as follows. In section 2.2, we briefly review the probabilistic model proposed by Lawrence and Schölkopf [81]. Then we introduce a modified method, CPA. In section 2.3, we describe the new methods, PKF and PF, in detail. The experimental evaluation is carried out in section 2.4. Finally, section 2.5

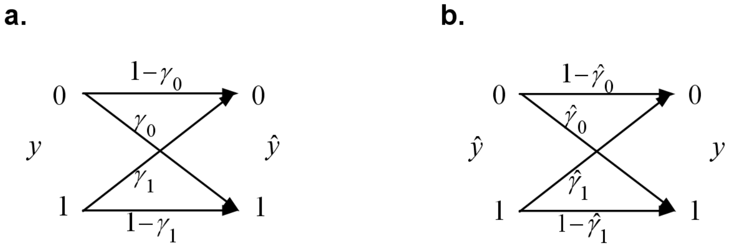

Figure 2.1: Illustration of the *classification noise process*. **a)** Conditional distribution $P(\hat{y}|y)$ with flipping rates γ_0 and γ_1. **b)** Conditional distribution $P(y|\hat{y})$ with flipping rates $\hat{\gamma}_0$ and $\hat{\gamma}_1$.

concludes the results with a discussion.

2.2 The Lawrence and Schölkopf model

Following Lawrence and Schölkopf [81], we now describe their method briefly. The class noise is assumed to have been generated by a *classification noise process* [3]. In this kind of noise process, the input feature distribution remains the same but their labels are independently and randomly reversed with probabilities γ_0 and γ_1, the flipping rates for the two classes, respectively. Let $\mathbf{x}$ denote the input feature vector, $y \in \{0, 1\}$ be the corresponding true class label, and $\hat{y}$ be the observed noisy class label. The noise introducing process can then be specified as:

$$P(\hat{y} = 1|y = 0) = \gamma_0, \; P(\hat{y} = 0|y = 1) = \gamma_1. \tag{2.1}$$

In practice, however, we only have access to the observed noisy class labels. Therefore it is necessary to express:

$$P(y = 1|\hat{y} = 0) = \hat{\gamma}_0, \; P(y = 0|\hat{y} = 1) = \hat{\gamma}_1. \tag{2.2}$$

Eqs. 2.1 and 2.2 are illustrated in Fig. 2.1.

Lawrence and Schölkopf [81] provide a general form that describes the data-generative process probabilistically. In this model, a data point is represented by the joint distribution of its feature vector $\mathbf{x}$, true class label y, and noisy observed class label $\hat{y}$. The joint distribution may be factorized into a class conditional distribution and a probability of the label being flipped:

$$P(\mathbf{x}, y, \hat{y}) = P(y|\hat{y})p(\mathbf{x}|y)P(\hat{y}). \tag{2.3}$$

Now if we can determine the three terms in Eq. 2.3, the true underlying distribution can be obtained to infer an appropriate classifier. Firstly, $P(\hat{y})$ is typically estimated as the proportion of the data in each class. Secondly, making use of the noise model Eq. 2.2, $P(y|\hat{y})$ can be expressed by the flipping rates $\hat{\gamma}_0$ and $\hat{\gamma}_1$. The last term to be estimated is the class conditional distribution. To simplify the model at this moment, it can be assumed to have a multidimensional Gaussian distribution of random vector $\mathbf{x}$ with mean $\mathbf{m}_y$ and covariance matrix $\mathbf{\Sigma}_y$. That is, $p(\mathbf{x}|y) = N(\mathbf{x}|\mathbf{m}_y, \mathbf{\Sigma}_y)$.

Summarizing, four parameters have to be estimated: the flipping rates $\hat{\gamma}_0$, $\hat{\gamma}_1$ and the Gaussian parameters $\mathbf{m}_y$, $\mathbf{\Sigma}_y$. Lawrence and Schölkopf showed that these parameters can be computed by optimizing a modified form of the log-likelihood via an EM algorithm. More precisely, in the "Expectation-step", the posterior distribution of the true class label y is computed as follows:

$$P(y|\mathbf{x},\hat{y},\Theta) = \frac{p(\mathbf{x},y|\hat{y},\Theta)}{p(\mathbf{x}|\hat{y},\Theta)}, \text{ where } \Theta = \{\mathbf{m}_y, \mathbf{\Sigma}_y\} \tag{2.4}$$

In the "Maximization-step", the optimization of the modified log-likelihood is achieved by the following update equations:

$$\mathbf{m}_y = \frac{1}{\nu_y}\sum_{n=1}^{N} P(y|\mathbf{x}_n,\hat{y}_n,\Theta)\mathbf{x}_n, \tag{2.5}$$

$$\mathbf{\Sigma}_y = \frac{1}{\nu_y}\sum_{n=1}^{N} P(y|\mathbf{x}_n,\hat{y}_n,\Theta)(\mathbf{x}_n - \mathbf{m}_y)(\mathbf{x}_n - \mathbf{m}_y)^{\mathrm{T}}, \tag{2.6}$$

$$\hat{\gamma}_0 = \frac{1}{\nu_y}\sum_{n=1}^{N} P(y|\mathbf{x}_n,\hat{y}_n,\Theta)(1-y)\hat{y}_n, \tag{2.7}$$

$$\hat{\gamma}_1 = \frac{1}{\nu_y}\sum_{n=1}^{N} P(y|\mathbf{x}_n,\hat{y}_n,\Theta)(1-\hat{y}_n)y, \tag{2.8}$$

where N is the total number of samples in the data set, and $\nu_y = \sum_{n=1}^{N} P(y|\mathbf{x}_n,\hat{y}_n,\Theta)$ is the expected number of samples in class y. To implement the above EM algorithm, Θ can be initialized as the means and covariances of the observed data set, and $\hat{\gamma}_0$ and $\hat{\gamma}_1$ may be set to any positive number. After the convergence of the EM steps, the underlying distribution may be recovered, and the priors $P(y)$ can also be calculated to derive a Bayes classifier.

2.2.1 Modifications to the Lawrence and Schölkopf model

We made two modifications to the model of Lawrence and Schölkopf [81]. The first concerns the update equations for $\hat{\gamma}_0$ and $\hat{\gamma}_1$, since the original equations 2.7 and 2.8 do not comply with the definition in Eq. 2.2. The new update equations are:

$$\hat{\gamma}_0 = \frac{1}{N_{\hat{y}}}\sum_{n=1}^{N} P(y|\mathbf{x}_n,\hat{y}_n,\Theta)(1-\hat{y}_n)y, \tag{2.9}$$

$$\hat{\gamma}_1 = \frac{1}{N_{\hat{y}}}\sum_{n=1}^{N} P(y|\mathbf{x}_n,\hat{y}_n,\Theta)(1-y)\hat{y}_n, \tag{2.10}$$

where $N_{\hat{y}}$ is the number of samples in class $\hat{y}$.

Secondly, we introduced a mixture-of-Gaussians model – instead of the single Gaussian in the Lawrence and Schölkopf model – to be flexible with respect to non-Gaussian class conditional densities. First, a clustering algorithm is applied to find the mixture components in the overall data set. This is possible because the generative class noise model does not alter the distribution of features. More specifically, the optimal number of components K is found when the clustering captures the data structure with the highest total log-likelihood. Since the likelihood of the complete data set will undoubtedly rise with increasing K, the data set is first divided into a mutually exclusive training and test data set. For each K, we cluster the training set using a mixture-of-Gaussians and assess the corresponding total log-likelihood on the test set. By doing so, we aim to avoid overfitting the training data. The optimal number of mixtures is determined as that value of K, which produces the highest total log-likelihood on the test set. Again, to avoid sampling effects, the likelihood is calculated by averaging the values over a number of different training and test-set samplings.

After associating each component with the observed noisy class label, the class conditional mixture density serves as the initialization of the probabilistic model. Each of the components is then optimized via the EM scheme by applying Eqs. 2.4 - 2.6, 2.9, and 2.10. By doing so, the underlying true class conditional distribution may be recovered as labelled mixture components. We denote this modified Lawrence and Schölkopf model the component-based probabilistic algorithm (CPA). The sequential steps of CPA are summarized below:

1. Estimate the number of mixture components K and the mixture density parameters.
2. Map mixture components to classes.
3. Apply Eqs. 2.4 - 2.6, 2.9, and 2.10 to optimize the mixture density parameters.
4. Map updated mixture components to classes.
5. Create a Bayes classifier.

2.3 Probabilistic Kernel Fisher and Probabilistic Fisher

In this section, the generalization of the class conditional distribution is handled by kernel-based methods to achieve a nonlinear classifier in the input feature space without modelling the input distribution explicitly. Here we first nonlinearly transform $\mathbf{x}$ into a higher-dimensional space F by Φ, then seek to construct a Fisher discriminant [30] in F. In this case, the discriminant function has the form:

$$g(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}) + \omega_0, \tag{2.11}$$

where $\mathbf{w}$ is the weight vector in F, $\Phi(\mathbf{x})$ is the mapping of the input vector $\mathbf{x}$ in F, and ω_0 is a constant threshold. The decision rule assigns $\mathbf{x}$ to $y = 0$ if $g(\mathbf{x}) > 0$, and $y = 1$ otherwise.

In the following the KFD proposed by Mika *et al.* [99] is introduced first. Then, we show how the KFD can be extended to the PKF method to tolerate class noise.

2.3.1 Kernel Fisher discriminant

After the nonlinear mapping Φ to the new feature space F, we seek to find a direction $\mathbf{w} \in F$ maximizing the Rayleigh quotient:

$$J(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}} S_B^{\Phi} \mathbf{w}}{\mathbf{w}^{\mathrm{T}} S_W^{\Phi} \mathbf{w}}, \tag{2.12}$$

with

$$S_B^{\Phi} = (\mathbf{m}_0^{\Phi} - \mathbf{m}_1^{\Phi})(\mathbf{m}_0^{\Phi} - \mathbf{m}_1^{\Phi})^{\mathrm{T}}, \tag{2.13}$$

$$S_W^{\Phi} = \sum_{y=0,1} \sum_{\mathbf{x} \in \mathbf{X}_y} (\Phi(\mathbf{x}) - \mathbf{m}_y^{\Phi})(\Phi(\mathbf{x}) - \mathbf{m}_y^{\Phi})^{\mathrm{T}}, \tag{2.14}$$

$$\mathbf{m}_y^{\Phi} = \frac{1}{N_y} \sum_{\mathbf{x} \in \mathbf{X}_y} \Phi(\mathbf{x}), \tag{2.15}$$

where $\mathbf{X}_y = \{\mathbf{x}_1, ..., \mathbf{x}_{N_y}\}$ denotes the samples from class y.

The solution relies only on scalar products between the transformed feature vectors, which can be replaced by some kernel function:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \Phi^{\mathrm{T}}(\mathbf{x}_i)\Phi(\mathbf{x}_j), \text{ with } i, j \in N \tag{2.16}$$

provided that the kernel can be written as an inner product, which means it must satisfy Mercer's condition [16]. By using the kernel, we are able to compute the Fisher discriminant $g(\mathbf{x})$ (Eq. 2.11) in F efficiently without mapping to F explicitly. This advantage could be remarkable when the dimension of the feature space F is high or even infinite. Some commonly used kernels include Gaussian kernels $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-|\mathbf{x}_i - \mathbf{x}_j|^2/c^2)$ and polynomial kernels $k(\mathbf{x}_i, \mathbf{x}_j) = (1 + \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j)^d$ for some positive constants c and d, respectively.

Now we show how the Rayleigh quotient (Eq. 2.12) can be expressed in terms of scalar products in F, so as to be replaced by some form of kernel. Since $\mathbf{w}$ lies in the space determined by the mapping of all training samples, it can be expanded as

$$\mathbf{w} = \sum_{n=1}^{N} \alpha_n \Phi(\mathbf{x}_n). \tag{2.17}$$

From this expansion and the definition in Eq. 2.15, we obtain

$$\mathbf{w}^{\mathrm{T}}\mathbf{m}_y^{\Phi} = \frac{1}{N_y} \sum_{n=1}^{N} \sum_{j=1}^{N_y} \alpha_n \Phi^{\mathrm{T}}(\mathbf{x}_n)\Phi(\mathbf{x}_j) = \frac{1}{N_y} \sum_{n=1}^{N} \sum_{j=1}^{N_y} \alpha_n k(\mathbf{x}_n, \mathbf{x}_j) = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{M}_y \tag{2.18}$$

with

$$(\mathbf{M}_y)_n = \frac{1}{N_y} \sum_{j=1}^{N_y} k(\mathbf{x}_n, \mathbf{x}_j). \tag{2.19}$$

Let $\mathbf{M} = (\mathbf{M}_0 - \mathbf{M}_1)(\mathbf{M}_0 - \mathbf{M}_1)^{\mathrm{T}}$, so the numerator of Eq. 2.12 becomes

$$\mathbf{w}^{\mathrm{T}} S_B^{\Phi} \mathbf{w} = \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{M} \boldsymbol{\alpha}. \tag{2.20}$$

Let $\mathbf{K}_y$ denote the kernel matrix for class y. It is an $N \times N_y$ matrix with the $(n,j)^{th}$ entry: $k(\mathbf{x}_n, \mathbf{x}_j)$. Then the denominator of Eq. 2.12 can be expressed as

$$\mathbf{w}^{\mathrm{T}} S_W^{\Phi} \mathbf{w} = \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{N} \boldsymbol{\alpha}, \tag{2.21}$$

where

$$\mathbf{N} = \sum_{y=0,1} \mathbf{K}_y (\mathbf{I} - \mathbf{1}_{N_y}) \mathbf{K}_y^{\mathrm{T}}. \tag{2.22}$$

Here $\mathbf{I}$ is the identity matrix, and $\mathbf{1}_{N_y}$ is the matrix with all entries $1/N_y$. Hence the Rayleigh quotient becomes

$$J(\mathbf{w}) = \frac{\boldsymbol{\alpha}^{\mathrm{T}} \mathbf{M} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^{\mathrm{T}} \mathbf{N} \boldsymbol{\alpha}}. \tag{2.23}$$

The maximum of J can be found for $\boldsymbol{\alpha} = \mathbf{N}^{-1}(\mathbf{M}_0 - \mathbf{M}_1)$. Consequently the projection of the input $\mathbf{x}$ onto $\mathbf{w}$ is

$$z = \mathbf{w}^{\mathrm{T}} \Phi(\mathbf{x}) = \sum_{n=1}^{N} \alpha_n k(\mathbf{x}_n, \mathbf{x}). \tag{2.24}$$

In this way, we obtain the projected data z in a one-dimensional space directly without mapping to F. The final discriminant Eq. 2.11 can be constructed based on z by determining ω_0 in a similar way as in the Fisher discriminant.

2.3.2 Probabilistic Kernel Fisher

In the presence of class noise, Lawrence and Schölkopf [81] proposed that by assuming the class conditional densities to be Gaussian distributions with equal covariances in the mapped space, a Fisher discriminant may be computed through some EM procedure incorporating the posterior probability $P(y|\mathbf{x}, \hat{y})$. However, the paper did not shed light on the detailed implementation, including how to compute the posterior probability and how to optimize the Fisher discriminant in the mapped space, which is only implicitly defined by kernels. Here we answer the above questions, and provide an actual realization of incorporating the probabilistic model into KFD.

First, we show how the projection direction $\mathbf{w}$ can be optimized in KFD when the data contains class noise. From Eqs. 2.13 - 2.15, it can be seen that to estimate S_B^{Φ} and S_W^{Φ} correctly, the key issue is to assign the data sample $\mathbf{x}_n$ to the correct class y. In the presence of class noise, the noisy labels are certainly not reliable enough to be used directly. Instead, we can use some posterior probability $P(y_n|\mathbf{x}_n, \hat{y}_n)$ to attach a class membership weight to each sample. The probabilistically weighted form of Eqs. 2.14 and 2.15 then becomes

$$S_W^{\Phi} = \sum_{y=0,1} \sum_{n=1}^{N} P(y_n|\mathbf{x}_n, \hat{y}_n)(\Phi(\mathbf{x}_n) - \mathbf{m}_y^{\Phi})(\Phi(\mathbf{x}_n) - \mathbf{m}_y^{\Phi})^{\mathrm{T}}, \tag{2.25}$$

$$\mathbf{m}_y^{\Phi} = \frac{1}{\nu_y} \sum_{n=1}^{N} P(y_n|\mathbf{x}_n, \hat{y}_n) \Phi(\mathbf{x}_n) \tag{2.26}$$

with ν denoting the expected number of samples in class y: $\nu_y = \sum_{n=1}^{N} P(y_n|\mathbf{x}_n, \hat{y}_n)$.

In KFD, the mapping is done implicitly by using kernels, in which case we do not compute $\Phi(\mathbf{x})$ explicitly. Still, the posterior weighting can be performed by applying 'the kernel trick'. That is, the kernel matrix $\mathbf{K}_y$ becomes an $N \times N$ matrix with the $(n, j)^{th}$ entry

$$(\mathbf{K}_y)_{n,j} = P(y_j|\mathbf{x}_j, \hat{y}_j)k(\mathbf{x}_n, \mathbf{x}_j), \tag{2.27}$$

and Eqs. 2.19 and 2.22 become

$$(\mathbf{M}_y)_n = \frac{1}{\nu_y} \sum_{j=1}^{N} P(y_j|\mathbf{x}_j, \hat{y}_j)k(\mathbf{x}_n, \mathbf{x}_j), \tag{2.28}$$

$$\mathbf{N} = \sum_{y=0,1} \mathbf{K}_y(\mathbf{I} - \mathbf{1}_{\nu_y})\mathbf{K}_y^{\mathrm{T}}. \tag{2.29}$$

Thus we achieve, given $P(y_n|\mathbf{x}_n, \hat{y}_n)$, the probabilistic expression for the kernelized Rayleigh quotient in Eq. 2.23, from which we can optimize $\mathbf{w}$ and compute the projection on $\mathbf{w}$ as in Eq. 2.24.

Now consider how to obtain the required posterior probability $P(y_n|\mathbf{x}_n, \hat{y}_n)$. Recall that the EM algorithm in Lawrence and Schölkopf [81] estimates the posterior probability in the input space by assuming Gaussian class conditional distributions. In KFD, this is not possible, since we do not have access to the mapped space, F. However, if we assume that the noisy labels are correct for the moment, we can apply KFD to compute a projection direction, $\mathbf{w}$, and project the data to z. Then we can model z with Gaussian distributions, and apply the EM algorithm on z to obtain the posterior label probabilities in the one-dimensional projected space. It is assumed that the distribution in the final projected space approximates the structure in the mapped space, and the posterior probability which maximizes the likelihood[1] of z also maximizes the likelihood of $\Phi(\mathbf{x})$, and eventually the likelihood of $\mathbf{x}$. That is, if the two classes are well separated on $\mathbf{w}$, they are also well separated in F and the original space.

Unlike Lawrence and Schölkopf [81], here no distribution assumption is made in F. After estimating $P(y_n|\mathbf{x}_n, \hat{y}_n)$, it is incorporated in the process of finding $\mathbf{w}$ to maximize the kernelized Rayleigh quotient as shown above in Eqs. 2.27 - 2.29. These two steps are repeated so that the $\mathbf{w}$ is adjusted to best separate the two classes. After that, the threshold ω_0 is computed by a Bayes classifier using the optimized distribution parameters in the projected space. Finally, the discriminant is determined according to Eqs. 2.11 and 2.24. The complete procedure of PKF is summarized below:

1. Initialization: apply KFD on the noisy data set directly to compute $\mathbf{w}$ and the projection z.
2. Apply Eqs. 2.4 - 2.6, 2.9, and 2.10 on z to estimate the posterior probability $P(y_n|\mathbf{x}_n, \hat{y}_n)$, log-likelihood L, and the flipping rates $\hat{\gamma}_0$ and $\hat{\gamma}_1$.
3. Incorporate the probabilities Eqs. 2.27 - 2.29, find the new $\mathbf{w}$ and project again.
4. Repeat Step 2 and Step 3 until convergence in L, $\hat{\gamma}_0$ and $\hat{\gamma}_1$.
5. Apply Eqs. 2.4 - 2.6, 2.9, and 2.10 on the final z to estimate the density parameters and priors for each class in the projected space.
6. Determine the threshold ω_0.
7. Create the discriminant function.

---

[1] The modified form of log-likelihood as in Lawrence and Schölkopf [81].

2.3.3 Probabilistic Fisher

Similarly, the procedure of PKF above can be applied to Fisher's Linear discriminant to tolerate class noise. In this case, the posterior probability is estimated by employing the EM algorithm Eqs. 2.4 - 2.6, 2.9, and 2.10 in the projected space $z = \mathbf{w}^T\mathbf{x}$, which is a linear transformation of the input space according to the Fisher's criterion. The computation of the projection direction $\mathbf{w}$ incorporates that posterior probability as stated in Eqs. 2.25 and 2.26 with $\Phi(\mathbf{x})$ replaced by $\mathbf{x}$, and is iteratively optimized. We call it the PF algorithm.

2.4 Experiments

Extensive experiments have been carried out to evaluate the classification performance of the PKF method using a Gaussian kernel [2] and the PF method in noisy data sets, compared with the modified model of Lawrence and Schölkopf (CPA) and the corresponding standard classifiers which do not model the class noise explicitly. The comparison is performed for different types of data sets, and for different noise levels. This section first describes the experiment using simulated data sets, followed by the introduction of a real world medical data set used for the evaluation. Finally, we report and discuss the experimental results.

2.4.1 Simulated data sets

Three simulated data sets were created to represent different types of distributions. The analysis of the performances of the classifiers in these data sets will shed light on their properties in the presence of noise. The noiseless data sets are illustrated in Fig. 2.2. In particular, G-ellipse is analogous to the toy data set in Lawrence and Schölkopf [81]. The class labels were randomly flipped at certain rates [3]: (1) $\gamma_0 = \gamma_1 = 0.25$; (2) $\gamma_0 = \gamma_1 = 0.45$; (3) $\gamma_0 = 0.25, \gamma_1 = 0$; (4) $\gamma_0 = 0.45, \gamma_1 = 0$; (5) $\gamma_0 = 0.75, \gamma_1 = 0$. Classifiers were trained on the same noisy data sets and the classifiers' accuracies were tested on independent large noiseless data sets with the same distributions as the corresponding training sets. By comparing the error rates of classifiers with and without modelling noise, the capability of tolerating class noise can be assessed for the proposed approaches.

Four standard classifiers which do not explicitly model the class noise were tested on the simulated data sets: quadratic Bayes normal classifier (QDC), *k*-nearest neighbor classifier (*k*-NN), Fisher's linear discriminant (Fisher), and the Kernel Fisher discriminant using a Gaussian kernel (KFD, [99]). These classifiers were chosen as representatives of linear and nonlinear, parametric and non-parametric approaches. Since PKF and PF are essentially noise-tolerant extensions of KFD and Fisher, respectively, the comparison between their performances may provide evidence on how the noise-modelled classifiers correct for the noisy labelling effect.

---

[2] Polynomial kernels with degree two were also investigated, which performed similar to or worse than the Gaussian kernel in the simulated data sets.

[3] A noise rate of $x\%$ means that for a randomly chosen subset of $x\%$ of the training samples, the class labels are flipped to the other class.

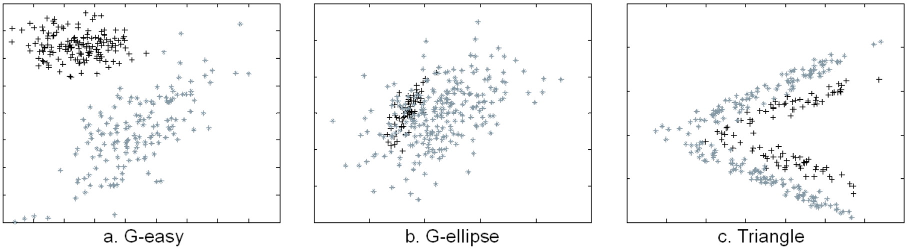

Figure 2.2: Illustration of the noiseless simulated data sets: **a)** G-easy; **b)** G-ellipse; **c)** Triangle. The black "+" sign represents a positive class sample, while the grey "*" sign represents a negative class sample.

As to the data set size, the training set size was set to be 100 × 2 (100 samples and 2 features) and 300 × 2 for all data sets, and 100 × 10 for Triangle data set. The independent test sets were produced five times as large as their training counterparts.

The EM algorithm of the probabilistic model was initialized with flipping rates $\hat{\gamma}_0 = \hat{\gamma}_1 = 0.3$ (as chosen in [81]). Actually, the model is insensitive to this initial setting. The optimization was considered to have converged when both the log-likelihood and the parameters changed by less than 10^{-2}. In k-NN, the number of neighbors, k, was chosen automatically by optimizing the leave-one-out error in the training set. For CPA, we estimate the number of mixture components K as follows. We split the data 50 times randomly into a training set (80%) and a test set (20%). For each split and each $K = 2, \ldots, 10$, we apply the EM algorithm and calculate the maximum log-likelihood over 100 random initializations, to avoid local maxima. We select the K which gives the maximum average log-likelihood over these 50 splits. Moreover, the Gaussian kernel width in KFD and PKF, c, was set as the average distance between the training samples. This is done not only for comparison purposes, but also because c cannot be optimized using noisy labels. Finally, each experiment was repeated 30 times.

2.4.2 CGH data set

*BRCA*1 mutation carriers usually have a high risk of developing breast cancer. Recently, an approach has been developed to identify potential *BRCA*1 mutation carriers within a group of sporadic breast carcinomas based on comparative genomic hybridization (CGH) profiles [146, 151]. The data set, collected at the Netherlands Cancer Institute, includes 34 proven *BRCA*1 mutation carriers (class 'B1') and 32 sporadic tumors (class 'C') [4]. This data set is a special case because we have prior knowledge that only the positive class ('B1') can be flipped into the negative class ('C'). That is, a sample is labelled as negative unless there is definitive proof that it is a positive one. PKF is assessed on both a hold-out test and the complete data set. In the former case, the test samples were excluded in the training procedure, and their estimated labels were verified with the true labels. In the latter case, the classifiers were trained and tested on the entire data set, and each sample's estimated posterior probability of class membership $P(y_n|\mathbf{x}_n, \hat{y}_n)$ was analyzed.

---

[4] See Wessels *et al.* [151] for detailed data preparation and description.

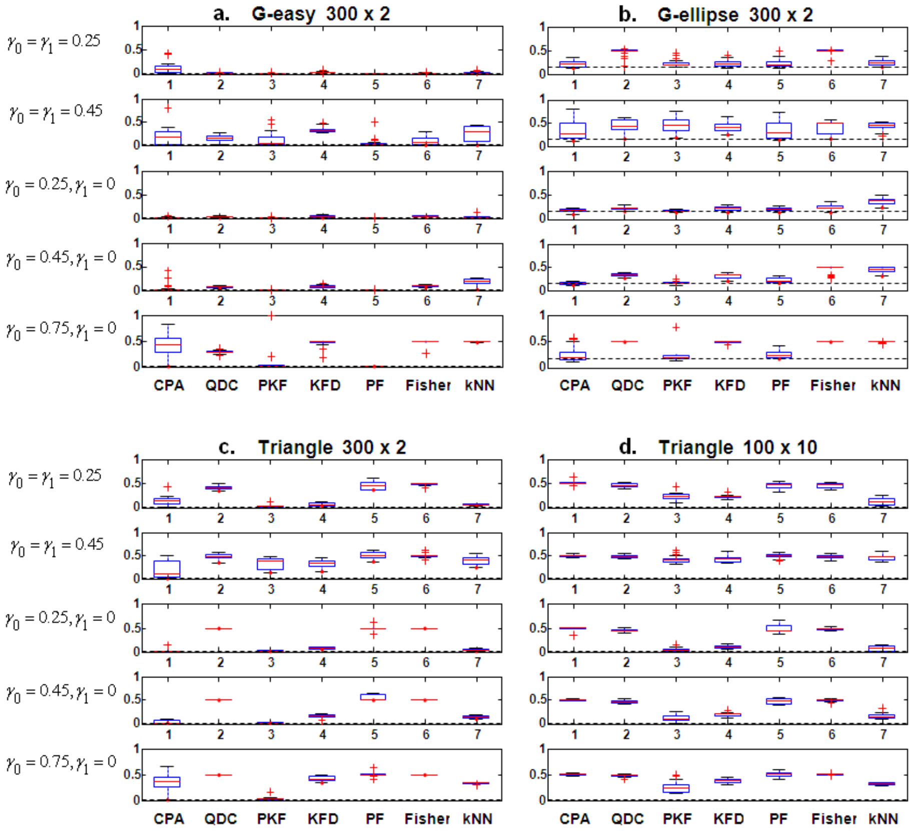

Figure 2.3: Average error rates of the seven classifiers. **a)** G-easy 300 × 2. The baseline error (indicated by a dashed line) is taken to be QDC's error rate in the noiseless data set. **b)** G-ellipse 300 × 2. The baseline error is taken to be QDC's error rate in the noiseless data set. **c)** Triangle 300 × 2. The baseline error is taken to be KFD's error rate in the noiseless data set. **d)** Triangle 100 × 10. The baseline error is taken to be KFD's error rate in the noiseless data set.

In particular, two samples (namely '1A' and '1B') that were initially labelled as sporadic were later confirmed to be *BRCA*1 carriers. This real situation offers an opportunity to test our noise-model algorithms on these two samples. The same two genomic features as in Wessels *et al.* [151] were employed in our experiment. In addition, the EM algorithm was initialized as $\hat{\gamma}_0 = 0, \hat{\gamma}_1 = 0.3$ to exploit the prior knowledge about possible flipping rates.

2.4.3 Results on the simulated data sets

Fig. 2.3 presents the average error rates of the seven classifiers in the simulated data sets as box plots. The baseline error is also plotted as a dashed line in each figure. As expected, the noise-model classifiers generally perform better when the ratio of sample size to dimensionality increases, i.e. 300 × 2 data sets in our experiment. In this more favorable scenario, we can better investigate their capabilities of tolerating the class noise.

The results indicate that the noise models (i.e. CPA, PKF, and PF) improve on the standard classifiers (i.e. QDC, KFD, and Fisher) in most cases. The exceptions occur when the classifier itself is not suitable for the data set. For example, Fisher is not suitable for the linearly inseparable Triangle data set, so neither is PF. Not surprisingly, higher noise levels can be tolerated when only one class is flipped.

In particular, PKF distinctly outperforms the others in the non-Gaussian data set (Triangle 300 × 2), and reaches the baseline error when $\gamma_0 = \gamma_1 = 0.25$ and $\gamma_0 = 0.75, \gamma_1 = 0$. Compared to the application of the modified Lawrence and Schölkopf's model in the original feature space (i.e. CPA), PKF exemplifies the advantage of the kernel-based method in classifying nonlinearly separable data sets with more complicated distributions, as well as the capability of PKF to address class noise. The improvement of PKF over CPA becomes more obvious when the number of features increases compared to the number of samples, as we can see from the Triangle data set of size 100 × 10 (Fig. 2.3d). Furthermore, PKF estimates the flipping rates very well (results not shown). When only the positive class was flipped into the negative class, $\gamma_1 = 0$ was always correctly estimated regardless of the initialization. In addition, the non-parametric classifier, k-NN, shows performance similar to the more complicated classifier KFD.

2.4.4 Results on the CGH data set

The two wrongly labelled *BRCA*1 samples have been successfully detected. In the hold-out test, PKF was trained on the data set excluding '1A' and '1B', and then the two samples were classified. As shown in Fig. 2.4a, PKF assigned both '1A' and '1B' to *BRCA*1.

In the other experiment, PKF was trained on the entire noisy CGH data set. Fig. 2.4b depicts the difference of the posterior probabilities, in which we see '1A' and '1B' are far more likely to be class 'B1' instead of their current labels 'C'. Interestingly, from the result we also found some other sporadic samples that seem to be potential *BRCA*1 tumors. In-depth evaluation on these samples is suggested.

2.5 Conclusions

In this chapter we analyzed the class noise problem, presented and investigated two noise-tolerant classifiers, which are applications of the probabilistic noise model proposed by Lawrence and Schölkopf [81]. More specifically, PF and PKF aim to optimize the projection direction in noisy data, yielding linear and non-linear classifiers in the original space, respectively. Explicit distribution assumptions in the input space are circumvented. Furthermore, we modified the probabilistic model of Lawrence and Schökopf in the original feature space, and extended it to a component-based probabilistic algorithm (CPA) to handle non-Gaussian data sets.

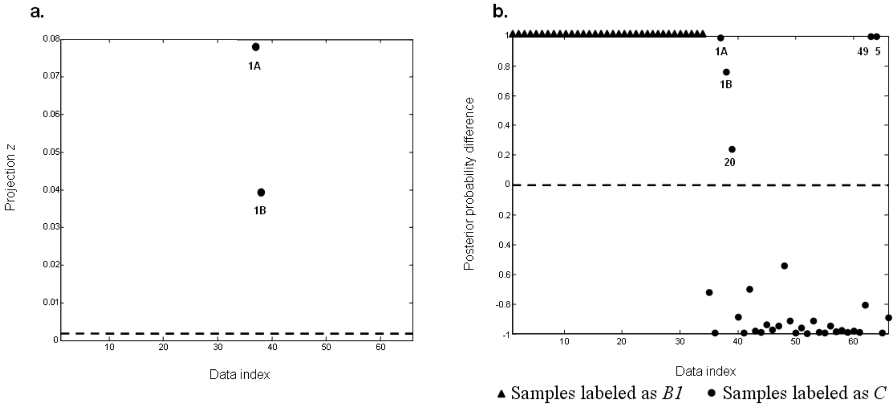

Figure 2.4: Classification results of PKF in the CGH experiment. **a)** Hold-out test. Projected test data values and the threshold ω_0 computed by PKF. The data points above the decision threshold (horizontal dashed line) indicate they are classified as *BRCA*1. **b)** Posterior probability difference $P(B1|\mathbf{x}_n, \hat{y}_n) - P(C|\mathbf{x}_n, \hat{y}_n)$ estimated by PKF. The data points above the threshold (horizontal dashed line) indicate possible *BRCA*1 carriers.

The experimental results are promising. On the whole, the proposed noise models improve the standard classifiers when properly applied. PKF exhibited significant advantages on non-Gaussian data sets and on data sets with relatively large numbers of features compared to the sample size. When applied to the *BRCA*1 data set, PKF correctly detected the wrongly labelled samples.

Having been pointed out by Lawrence and Schölkopf [81], the computational problem in large data sets is a major handicap of kernel-based methods. More specifically, the kernelization algorithm increases the complexity to $O(N^3)$. For this reason, we did not implement the experiment on larger data set with more iterations.

It should be noted that this study has addressed only random class noise in a two-class problem. The remaining issues to be studied include how to handle other types of class noise (e.g. not random or not independent) as well as multiclass problems. How to distinguish noisy labelled samples from outliers is still a challenging subject [135]. In addition, when only one class is flipped, it resembles the situation of the one-class classification problem [140]. Then the one-class classifier techniques may be adopted to handle class noise.

3

CLASSIFICATION USING PRIOR KNOWLEDGE ON MEASUREMENT NOISE

Samples can be measured with different precisions and reliabilities in different experiments, or even within the same experiment. These varying levels of measurement noise may deteriorate the performance of a pattern recognition system, if not treated with care. Here we investigate the benefit of using prior knowledge about measurement noise during system construction. We propose a kernel density classifier which integrates such prior knowledge. Instead of using an identical kernel for each sample, we transform the prior knowledge into a distinct kernel for each sample. The integration procedure is straightforward and easy to interpret. In addition, we show how to estimate the diverse measurement noise levels in a real world data set. Compared to basic methods, the new kernel density classifier can give a significantly better classification performance. As expected, this improvement is more obvious for data sets with small sample sizes and large numbers of features.

This chapter was published in *Pattern Recognition*, 2008 [87].

3.1 Introduction

In practice, prior knowledge about the problem domain at hand is usually beneficial or even essential for solving a pattern recognition problem. Prior knowledge can take various forms, ranging from knowledge about the importance of a class, the informativeness of features, the quality of samples, to the dependency of variables. If properly exploited, prior knowledge can substantially improve a pattern recognition system's performance at all stages, including domain understanding, data preparation, data selection, feature selection, model design, result interpretation, and performance evaluation. As the simplest example, Bayesian approaches use prior probabilities of class occurrences as one form of encoding prior knowledge [9]. In image classification, transformation invariance and locality information can be incorporated in designing Support Vector Kernels [126]. In other cases, expert knowledge can be utilized to specify the topology of a Bayesian network, which circumvents learning the structure from possibly insufficient data [128].

Despite the broad application of prior knowledge, the use of knowledge about the measurement devices used to measure features, such as their noise levels, is hardly addressed. By measurement noise we mean the deviation of the measured (observed) value from the true value. It can depend on the type of feature (measurement technique), the object being measured, or the measured feature value itself. Unfortunately, most current statistical pattern recognition methods assume independent, identically distributed measurement noise, which might result in less reliable models. For instance, the noisy training samples, especially those close to the decision boundary, will distort the boundary if they are treated equally. Although some work addresses the reliabilities of different experimental technologies [59], the diverse measurement noise levels of different samples in the same experiment are ignored.

To address this problem, this chapter proposes a new methodology of utilizing prior knowledge about measurement noise to construct a kernel density classifier. The kernel density classifier is well-studied and has been successfully used in many applications [31, 113, 122]. We now interpret the concept of the kernel from the view of measurement noise. When measurements are corrupted by random variations, each observed sample can be characterized by a kernel centered at its true measurement value. In this sense, a kernel actually indicates how precisely the sample is measured, and how trustworthy the sample is as a representation of the true value. Therefore, to approximate the original data distribution from the observed samples, we can sum these characteristic kernels. This idea is used in the kernel density classifier.

There are three aspects that may hinder the basic kernel method's performance. First, all training samples in a class have the same kernel with the same shape and weight, which implies they are equally well measured and trusted. Second, common solutions to estimate the kernel are solely based on the measured data itself, without any knowledge of measurement noise. Finally, as a density-based classifier, the kernel method needs many samples to achieve a reasonably good performance.

The new approach presented here, the Integrative Kernel Method, aims to incorporate the prior knowledge about measurement noise and consequently improve the performance of the kernel density classifier. The main contribution is that we explicitly specify the manner of transforming the measurement noise knowledge into a distinct kernel for each sample. Moreover, we clearly quantify the prior knowledge about measurement noise in a biological data set, and successfully demonstrate the benefit of our proposed method. Experimental results show classification improvements of the new method over three more basic kernel methods at various sample sizes and feature sizes. This means we need fewer training samples to reach the same performance after the integration of prior knowledge, and eventually the time and the cost for collecting samples can be reduced.

Section 3.2 first introduces the new Integrative Kernel Method, which transforms prior knowledge into distinct kernels. Besides the basic kernel method using spherical kernels, two other kernel methods using elliptical kernels are briefly mentioned in section 3.3. Section 3.4 describes a yeast co-expression data set and how to represent the prior knowledge in this data set. Section 3.5 presents the experimental results comparing the new method and the basic methods, followed by conclusions in section 3.6.

3.2 Integrative Kernel Method

As explained in the introduction, each sample is measured with different uncertainty, and this information can be integrated into the kernel density classifier by applying different kernels. Although some research has been done to estimate the kernel width sample-wise [7, 90], those solutions were proposed to better estimate the density according to the various sparseness of the data. The novelty of our method is that instead of estimating the kernel only from data (the measured feature values), we intend to utilize our additional knowledge to improve classification by explicitly transforming it into the parameters of the kernel.

In this section, we first show how the basic kernel method can be modified theoretically. Then, we propose a simple model to transform noise level knowledge into kernel parameters.

3.2.1 Theoretical basis

In the basic kernel method, a spherical unimodal kernel is used, and all samples have the same weight. That is, given a training set $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ of n samples represented by p features, the density of an object $\mathbf{z}$ is estimated as

$$\hat{p}(\mathbf{z}) = \frac{1}{n}\sum_{i=1}^{n} K(\mathbf{z} - \mathbf{x}_i, h_\mathrm{s}), \tag{3.1}$$

where the kernel width h_s is a scalar and can be calculated per class using for example leave-one-out maximum likelihood estimation [25].

K is a multivariate kernel that has an area of one. Here we choose the most widely used spherical Gaussian kernel:

$$K(\mathbf{z} - \mathbf{x}_i, h_\mathrm{s}) = \frac{1}{(\sqrt{2\pi}h_\mathrm{s})^p}\exp\Big(- \frac{||\mathbf{z} - \mathbf{x}_i||^2}{2h_\mathrm{s}^2}\Big). \tag{3.2}$$

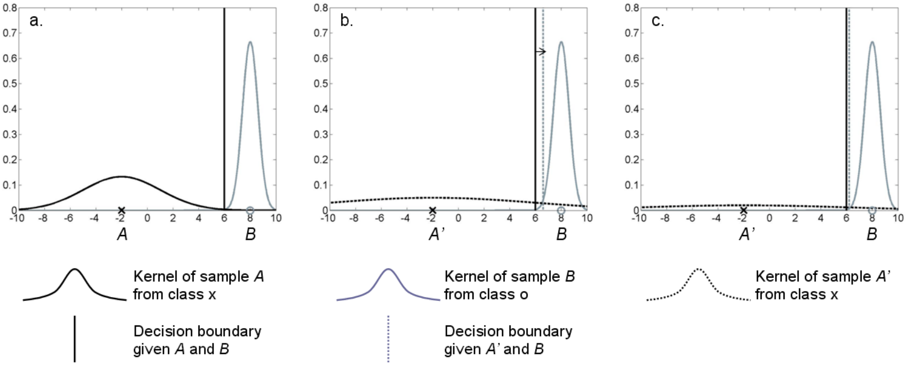

Figure 3.1: A demonstration of the use of kernel weights. **a)** The decision boundary given A and B. **b)** If a less precisely measured sample A' is only assigned a wider kernel with the same weight as A, the decision boundary may shift *away* from A', which is counter-intuitive. **c)** By downweighting A', its contribution to the overall density estimate and classification is reduced, resolving the problem.

So Eq. 3.1 can be written as

$$\hat{p}(\mathbf{z}) = \frac{1}{n}\sum_{i=1}^{n}\Big\{\prod_{d=1}^{p}\frac{1}{\sqrt{2\pi}h_{\mathrm{s}}}\exp\Big(-\frac{(z_d - x_{i,d})^2}{2h_{\mathrm{s}}^2}\Big)\Big\}. \tag{3.3}$$

Note that Eq. 3.3 is written as a sum of product kernels in p features, which is the most widely used form for multivariate density estimation. It assumes local independence, without implying global independence of features.

Since samples can be measured with different noise levels in different experiments, it is necessary to use a different kernel for each sample. More specifically, the prior knowledge on measurement noise can be used to construct a distinct kernel through two parameters. On the one hand, through the kernel *width*. For imprecise measurement, it is more likely the measured value is far away from the actual value than for a precise measurement. Hence, its contribution to the density estimate should be "spread out" more. On the other hand, through the kernel *weight*. Not all samples are equally important due to their various measurement noise levels, and should not influence the system construction equally. An imprecise measurement should get less weight in the overall density estimation and classification. For an illustration, see Fig. 3.1.

This makes it reasonable to replace the identical spherical kernel by a distinct elliptical kernel for each sample. That is, the scalar kernel width h_{s} in Eq. 3.3 is replaced by $h_{i,d}$, indicating the kernel width in feature d for sample i. Similarly, we can assign a distinct weight w_i to each sample instead of an identical weight. This weight can be derived from the sample's weight at each feature independently, i.e. $w_i = \prod_{d=1}^{p} w_{i,d}$. Then Eq. 3.3 becomes:

$$\hat{p}(\mathbf{z}) = \frac{1}{\eta}\sum_{i=1}^{n}\Big\{w_i\prod_{d=1}^{p}\frac{1}{\sqrt{2\pi}h_{i,d}}\exp\Big(-\frac{(z_d - x_{i,d})^2}{2h_{i,d}^2}\Big)\Big\}, \tag{3.4}$$

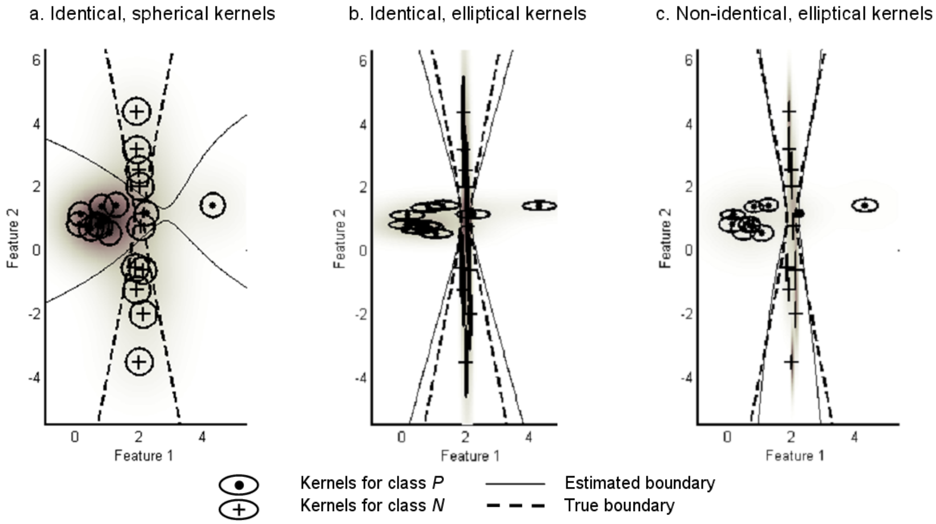

Figure 3.2: The effect of using different kernels on the Highleyman data set. The estimated density is represented by the grey-value shadow. Both estimated boundaries (solid line) and the true boundary (dashed line) are shown. The true boundary is computed based on the known underlying data distribution. Different kernel methods are used to compute the kernels. **a)** The basic kernel method using identical spherical kernels. **b)** The kernel method using identical elliptical kernels (see Eq. 3.8 in section 3.3). **c)** The Integrative Kernel Method using non-identical elliptical kernels (Eq. 3.4), where the measurement noise is simulated for illustration.

where the normalization factor $\eta = \sum_{i=1}^{n} w_i$.

The effect of using different kernels is illustrated on the Highleyman data set in Fig. 3.2. The data set consists of two overlapping Gaussian classes with different covariance matrices according to the Highleyman distribution [26, 54].

3.2.2 Transformation model

Now we introduce a simple model to transform measurement noise knowledge into a distinct kernel width and weight for each sample. Estimation of the measurement noise depends on the application, which will be elaborated on in section 3.4. For the moment, suppose the measurement noise of sample i in feature d is known, and is represented by its certainty level $r_{i,d}$. The larger $r_{i,d}$, the more certain the measurement, hence a smaller width $h_{i,d}$ and a larger weight $w_{i,d}$ should be assigned to this sample at this feature. To keep it simple, we use linear functions to model the negative dependency between the kernel variance $h^2_{i,d}$ and $r_{i,d}$, and the positive dependency between $w_{i,d}$ and $r_{i,d}$.

The dependencies are robustly estimated piece-wise linear functions as shown in Figs. 3.3 and 3.4. The slope of $h^2_{i,d} \sim r_{i,d}$ is estimated on values between the 25th

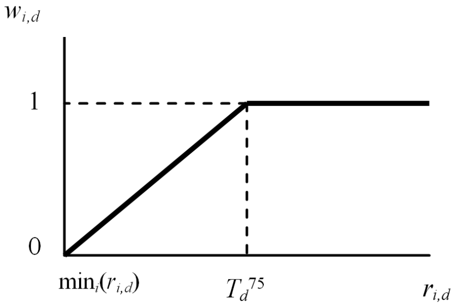

Figure 3.3: The dependency function between the certainty level $r_{i,d}$ and $w_{i,d}$, i.e. the weight of sample i at feature d.

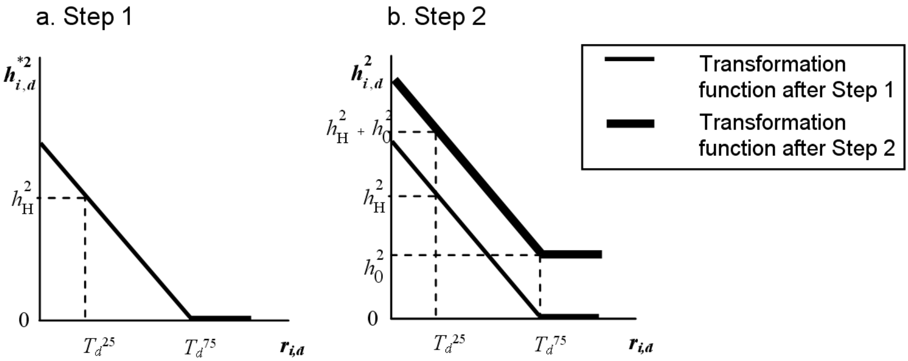

Figure 3.4: The two steps of specifying the dependency function between the certainty level $r_{i,d}$ and the kernel variance $h^2_{i,d}$. **a)** The $h^{*2}_{i,d}$ function after Step 1. **b)** The final $h^2_{i,d}$ function after Step 2.

(T_d^{25}) and 75th (T_d^{75}) percentiles of $\{r_{1,d}, ..., r_{n,d}\}$ to avoid undue influence of outlying certainty levels. We consider samples having $r_{i,d} > T_d^{75}$ as equally trustworthy, which makes the function robust against samples with extraordinary large $r_{i,d}$.

Now we describe the two dependency functions in detail. The dependency function between $r_{i,d}$ and $w_{i,d}$ is straightforward: the sample which has the minimum $r_{i,d}$ will have a weight $w_{i,d} = 0$, and the samples which are larger than T_d^{75} will have $w_{i,d} = 1$. The weights for the remaining samples can be found using the linear function:

$$w_{i,d} = \frac{r_{i,d} - \min_i(r_{i,d})}{T_d^{75} - \min_i(r_{i,d})}.$$

The linear dependency function between $r_{i,d}$ and $h^2_{i,d}$ is specified in two steps as follows.

Step 1: Slope. Initially, $h^{*2}_{i,d}$ at T_d^{75} is set to 0. Furthermore, $h^{*2}_{i,d}$ at T_d^{25} is set to h^2_{H}, which is derived from the optimal spherical kernel width h_{S}: $h^2_{\mathrm{H}} = (C \cdot h_{\mathrm{S}})^2$, where C is a constant amplification coefficient. Thus the slope is specified robustly based on the middle quartiles of $\{r_{1,d}, ..., r_{n,d}\}$, and its scale is adjusted accordingly to different sample sizes and feature sizes.

Step 2: Offset. An offset h_0^2 is added to the $h_{i,d}^{*2}$ function in order to regularize it. h_0 is found by maximizing the leave-one-out total likelihood of the training set as follows:

$$\begin{aligned} LL &= \sum_{k=1}^{n} \log(\hat{p}(\mathbf{x}_k)) \\ &= \sum_{k=1}^{n} \log\Big\{ \sum_{i=1, i\neq k}^{n} \Big\{ w_i \prod_{d=1}^{p} \frac{1}{\sqrt{2\pi(h_{i,d}^{*2} + h_0^2)}} \exp\Big(- \frac{(x_{k,d} - x_{i,d})^2}{2(h_{i,d}^{*2} + h_0^2)} \Big) \Big\} \Big\} - D, \end{aligned} \tag{3.5}$$

where D is a constant: $\sum_{k=1}^{n} \log(\sum_{i=1, i\neq k}^{n} w_i)$.

To implement this optimization, h_0 is initialized as h_s, the optimal spherical kernel width estimated by the basic kernel method, and a hill-climbing searching method is applied to maximize LL.

After the two-step transformation, the prior knowledge of the noise in terms of a certainty level $r_{i,d}$ is mapped to a kernel width within a proper range. The final dependency function is depicted in Fig. 3.4b, and the density estimate in the classification stage is computed by Eq. 3.4, with $h_{i,d} = \sqrt{h_{i,d}^{*2} + h_0^2}$.

3.3 Elliptical kernel methods

The basic kernel method, which uses an identical spherical kernel for all samples, has been introduced in section 3.2.1. Nevertheless, there are some arguments that using a different width in each dimension may generate a better approximation of the data density. Let h_d denote the width at dimension d, then the same trick can be applied as was used to arrive at Eq. 3.4, but here the density is computed using an identical elliptical kernel for all samples:

$$\hat{p}(\mathbf{z}) = \frac{1}{n} \sum_{i=1}^{n} \prod_{d=1}^{p} \frac{1}{\sqrt{2\pi} h_d} \exp\Big(- \frac{(z_d - x_{i,d})^2}{2h_d^2} \Big). \tag{3.6}$$

Regarding the estimation of h_d , there are two commonly used solutions. In the simplest approach, which we refer to as the *Scaled Elliptical Kernel Method*, the variance of the data in each dimension is taken into account to scale the kernel width accordingly. More specifically, the h_d per class is estimated by the following three steps:

Step 1: Normalize each feature to a unit variance by dividing all feature values by the standard deviation σ_d of that feature.

Step 2: Find the optimal spherical kernel width h_s in this normalized space using the basic kernel method (see section 3.2.1).

Step 3: For each dimension, inverse-scale the found kernel width h_s to the original feature space according to $h_d = h_s \cdot \sigma_d$.

Another approach to estimate h_d is to maximize the leave-one- out total likelihood. We refer to it as the *ML Elliptical Kernel Method*. An Expectation-Maximization algorithm is proposed [109] and briefly described here for completeness:

E-step: Evaluate the conditional probability of each sample given every other sample.

$$\begin{aligned} p(\mathbf{x}_i|\mathbf{x}_k) &= \frac{p(\mathbf{x}_k|\mathbf{x}_i)}{\sum_{l=1,l\neq k}^{n} p(\mathbf{x}_k|\mathbf{x}_l)} \\ &= \frac{\prod_{d=1}^{p} \frac{1}{\sqrt{2\pi}h_d} \exp(-(x_{k,d}-x_{i,d})^2/2h_d^2)}{\sum_{l=1,l\neq k}^{n} \prod_{d=1}^{p} \frac{1}{\sqrt{2\pi}h_d} \exp(-(x_{k,d}-x_{l,d})^2/2h_d^2)}. \end{aligned} \tag{3.7}$$

M-step: Determine the new h_d by maximizing the expectation of the leave-one-out total likelihood, using the conditional probabilities computed in the E-step.

$$h_d^2 = \frac{1}{n} \sum_{k=1}^{n} \sum_{i=1,i\neq k}^{n} p(\mathbf{x}_i|\mathbf{x}_k)(x_{k,d}-x_{i,d})^2. \tag{3.8}$$

3.4 Data

This section first introduces a biological classification problem, followed by a discussion of the data set we use and the accompanying knowledge about the measurement noise in this data set.

3.4.1 Protein complex prediction

The genome-wide discovery of protein complexes is crucial to elucidate the biological system's behavior [39, 131]. In pattern recognition terms, this problem can be cast as a classification problem [149, 160], in which the goal is to predict whether two proteins belong to the same complex or not.

Proteins in the same complex are often co-expressed, which means the corresponding genes that code for the proteins have similar activities in terms of mRNA expression levels under various conditions, e.g. different environments for the cell. Therefore, the mRNA co-expression coefficient is an important characteristic feature and has been broadly used [40, 64, 65, 91]. More specifically, for a pair of genes with expression vectors $\mathbf{x}$ and $\mathbf{y}$, their co-expression level is represented by their Pearson correlation coefficient ρ:

$$\rho(\mathbf{x},\mathbf{y}) = \frac{\mathrm{cov}(\mathbf{x},\mathbf{y})}{\sqrt{\mathrm{var}(\mathbf{x})\mathrm{var}(\mathbf{y})}}. \tag{3.9}$$

As we will discuss later, the measurement noise of this mRNA co-expression coefficient can be approximated from the expression vectors $\mathbf{x}$ and $\mathbf{y}$. Fig. 3.5 schematically shows the mRNA expression levels of two genes under different conditions.

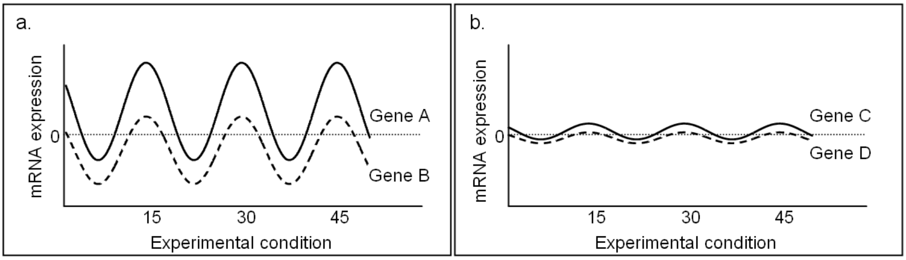

Figure 3.5: Schematic illustration of the co-expression of a pair of genes in different experimental conditions. Two situations are shown both with $\rho(\mathbf{x}, \mathbf{y}) = 1$. **a)** The two genes A and B are very active in the experiments. **b)** The two genes C and D are inactive in the experiments. The profiles are simplified as sinusoid curves purely to illustrate the linear relationship between them, while emphasizing their amplitudes. If the experiments concern different time points, the x-axis is an ordered axis, i.e. the time series. In other types of experiments, the axis is merely an index.

To investigate the influence of the dimensionality on the classification performance, we use expression data sets from different labs to obtain various mRNA co-expression coefficients for the same protein pair. Due to the different experimental setup, techniques, conditions, and many other reasons, the measurements for the same protein pair can contain diverse noise levels in different experiments. This provides an opportunity to study the new method's power of integrating diverse measurement noise both within an experiment and across multiple experiments.

3.4.2 Data set

A data set is constructed containing $\sim$ 2 million protein pairs of the model organism *Saccharomyces cerevisiae*, or baker's yeast. Each protein pair is a sample in the data set. There are 11 features that are mRNA co-expression coefficients computed from the expression profiles measured from 11 labs independently. The data come from the Stanford Microarray Database [44], Tai *et al.* [138], and Hughes *et al.*[58]. All data used in this work can be found at http://ict.ewi.tudelft.nl/$\sim$yunlei/measurementnoise.

In accordance with previous research [65, 91], we define the true class label of each sample based on the MIPS complexes catalog [98]. This results in 7929 samples in the positive class 'P' (the two proteins belong to the same complex) and 2 129 049 samples in the negative class 'N'.

3.4.3 Prior knowledge on measurement noise

Let S denote the divisor in Eq. 3.9, to which we refer as *Product of Standard-deviations*:

$$S(\mathbf{x}, \mathbf{y}) = \sqrt{\mathrm{var}(\mathbf{x})\mathrm{var}(\mathbf{y})}. \tag{3.10}$$

When S is relatively small, it means the activity of at least one gene in the pair does not change significantly across these experimental conditions, which suggests that the obtained expression vector is likely to be only noise instead of a real pattern of activity. Therefore, the expression vector is not informative and the Pearson correlation coefficient ρ thus computed may not reflect the true linear relationship between the two genes reliably. That is, a high ρ of a pair with a small S cannot guarantee the genes are truly co-expressed, as low variance here indicates lack of information rather than a precise measurement. On the contrary, we are more confident about whether ρ reflects the co-expression level well when S is larger. These two situations are illustrated in Fig. 3.5.

This suggests that we can utilize the Product of Standard-deviations S to estimate the certainty level $r_{i,d}$ of each sample in this data set, and have it serve as the input of the transformation function in the Integrative Kernel Method (section 3.2.2). After this step, the dependency functions $h_{i,d}^2 \sim r_{i,d}$ and $w_{i,d} \sim r_{i,d}$ are fully specified. The dependency $h_{i,d}^2 \sim r_{i,d}$ indicates that the variance of the kernel is proportional to the combined variance (Product of Standard-deviations) of the gene pair's mRNA expression levels, with both sides of the equation having the same metric.

The certainty level $r_{i,d}$ can be estimated either directly or indirectly from S. In the simplest case, $r_{i,d}$ is computed based on the experimental expression profile of the pair of genes directly, according to Eq. 3.10. However, we observed a significant dependency between the co-expression coefficient ρ and the Product of Standard-deviations S. Fig. 3.6 displays the regression result of S on ρ obtained by non-parametric Locally Weighted Regression [14]. It clearly shows that S is noticeably dependent on ρ: when a gene pair is more positively (or negatively) co-expressed, the Product of Standard-deviations S tends to be larger, and therefore the measured feature value is more reliable. This result is not surprising, because the noisy expression profiles produced by inactive protein pairs are less likely to have a linear relationship. Consequently, the certainty level $r_{i,d}$ can also be estimated indirectly from the regression of S on ρ. This essentially means that our knowledge of measurement noise now comes from two sources, namely observed Product of Standard-deviations and measured co-expression coefficient, considering the individual activity changes of two genes (S) as well as their joint change (ρ), and providing a more consistent and informative measure of certainty levels.

3.5 Results

To demonstrate the added value of prior knowledge about measurement noise, we investigate the classification performances of the proposed method and basic kernel methods on the yeast co-expression data set. The classification performance is measured by the AUC (Area Under ROC Curve) [51], which considers the overall performance of a classifier under all possible class priors (relative class sizes) and can be averaged over multiple experiments. This measure is general and not specific to the yeast co-expression data set.

The experiments are carried out as follows:

Step 1: Generate a training set and a test set by sampling randomly from the entire data

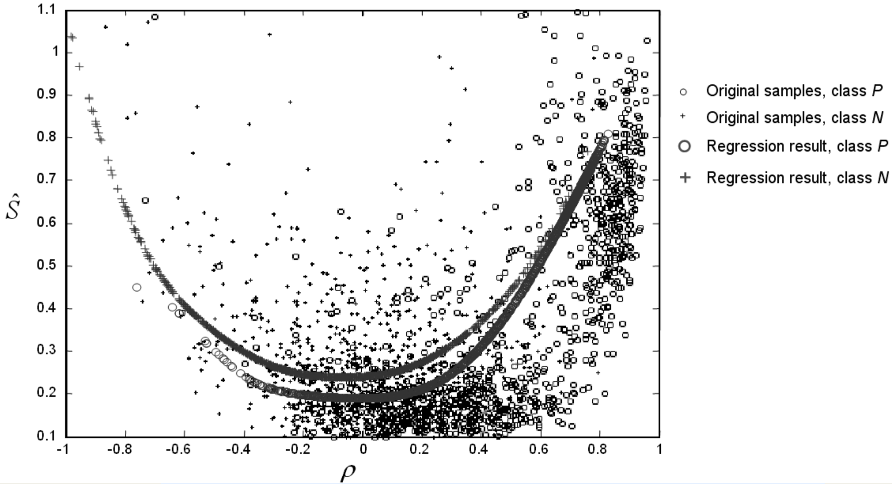

Figure 3.6: The regression result of the Product of Standard-deviations S on the co-expression coefficient ρ obtained by Local Weighted Regression on 'Spellman' data set of Stanford Microarray Database.

set. The training set contains 10, 50, and 300 samples per class. The test set has 2000 independent samples per class.

Step 2: On the training set, rank the 11 features by the Mann-Whitney U test [94].

Step 3: On the training set with incrementally added features (best 1, best 2, ...; according to the U test), apply all kernel methods to compute the kernel width (and weight) for each training sample. Note that all parameters, including $r_{i,d}$, are estimated on the training set only.

Step 4: Estimate the posterior probability of the test samples, i.e. P(Class 'P' | test sample) and P(Class 'N' | test sample).

Step 5: Compute the AUC for each method.

Step 6: Repeat Steps 1-5 200 times.

The results are presented in Fig. 3.7, which shows the average AUC over the 200 iterations for increasing feature set size. When the ratio of the feature set size divided by the sample size is large (e.g. 10 samples per class with 11 features), the Integrative Kernel Method outperforms all other kernel methods. The pairwise t-test indicates that the performance difference between the Integrative Kernel Method and the best basic method (i.e. the Scaled Elliptical Kernel Method) is statistically significant: p-value = 10^{-6}, 2×10^{-29}, and 7×10^{-15} for 10, 50, and 300 samples with 11 features, respectively.

This improvement demonstrates the benefit of using prior knowledge, in this case the measurement noise knowledge, when there are a limited number of training samples with high dimensionality. On the other hand, it means fewer samples are needed for the Integrative Kernel Methods to reach a certain performance, compared to the other methods. In Fig. 3.7 for example, to obtain the same performance in 11-D, the Inte-

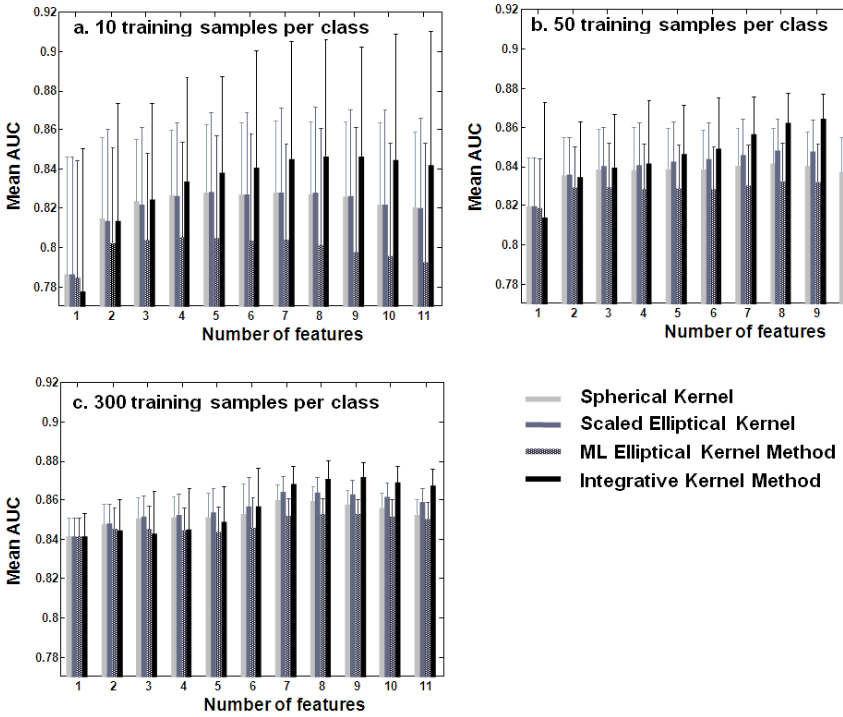

Figure 3.7: Classification performances in terms of average AUC. The average AUCs of each method at different feature sizes are shown as bars. The standard deviation of each method at a certain feature size is indicated by a whisker extending from the average AUC: **a)** 10 samples per class in the training set. **b)** 50 samples per class in the training set. **c)** 300 samples per class in the training set.

grative Kernel Method only needs about 20% of the training samples that the Scaled Elliptical Kernel Method needs.

For the previous results, the amplification coefficient C (section 3.2.2, Step 1) is taken to be 2.5. The classification performance, however, is quite robust w.r.t. the choice of this parameter. This is shown in Fig. 3.8, which presents the results for C = 1.5, 2, 2.5, 3, and 3.5.

We also investigated the performance of the Integrative Kernel Method using the measured Product of Standard-deviations S directly to estimate the certainty level $r_{i,d}$. The performances of the Integrative Kernel Method using the two different ways to estimate the certainty level are shown in Fig. 3.9. We can clearly observe a difference between the two approaches, which becomes more apparent when the sample size increases. In Fig. 3.9a, when the training sample size increases to 300 per class, the performance of the direct estimation approach degrades. This is because some features (experiments) contain many measurements that are inconsistent with the regression relationship between the Product of Standard-deviations S and the co-expression coefficient ρ, i.e. the experimental results are noisy and far away from the regression curve,

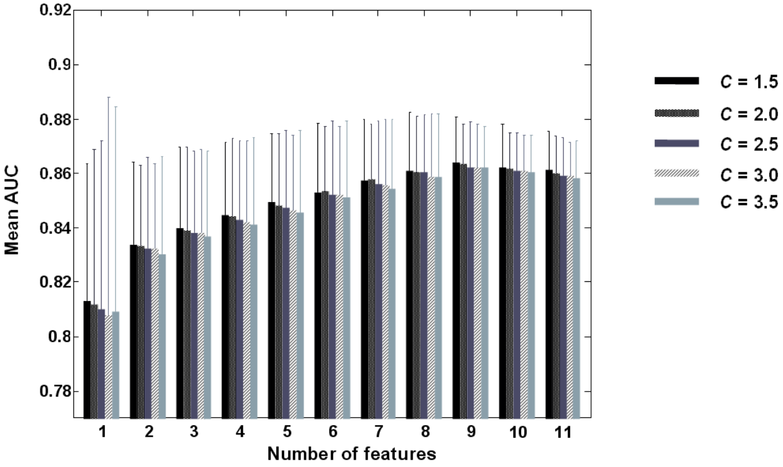

Figure 3.8: The average AUC of the Integrative Kernel Method for different C. The average AUCs for each C value at different feature set sizes are shown as bars. The standard deviation is indicated by a whisker extending from the average AUC. The similar performance results indicate that the method is insensitive to the value of C. The training set size is 50 samples per class. The results for other training set sizes are similar (results not shown).

in spite of the high ranks of the features in the Mann-Whitney U test. Therefore, direct application of the measured S results in inconsistent kernel widths and weights. Consequently, the performance of the Integrative Kernel Method degrades when such noisy features are encountered, and becomes worse given more noisy samples.

As we can see from Fig. 3.9b, this drawback of direct application of the measured Product of Standard-deviations S is overcome by the regression procedure as we proposed. For these noisy data, the corresponding part of the regression curve tends to be flat, which means all samples will have similar certainty levels and consequently similar kernel widths and weights. This actually alleviates the erroneous estimates made by using the measured S directly.

Nevertheless, the performance of the Integrative Kernel Method is still deteriorated by those noisy features. In Fig. 3.7c for example, the Integrative Kernel Method is inferior to the basic methods for the first four or first five features. This brings up an important issue of classifier construction, namely feature selection. When prior knowledge is incorporated in the classifier construction, only those features that contain consistent and accurate information about measurement noise should be used.

We proved the advantage of feature selection by a followup experiment, in which three noisy features were removed. These features were found to have the largest average MSE in 200 iterations of the leave-one-out Locally Weighted Regression on normalized data [1]. The performance improvements of the Integrative Kernel Method over the

---

[1] The data set has 10 samples per class. Each feature is normalized to zero mean and unit variance prior to the regression, so that the MSE results of different features can be compared with each other.

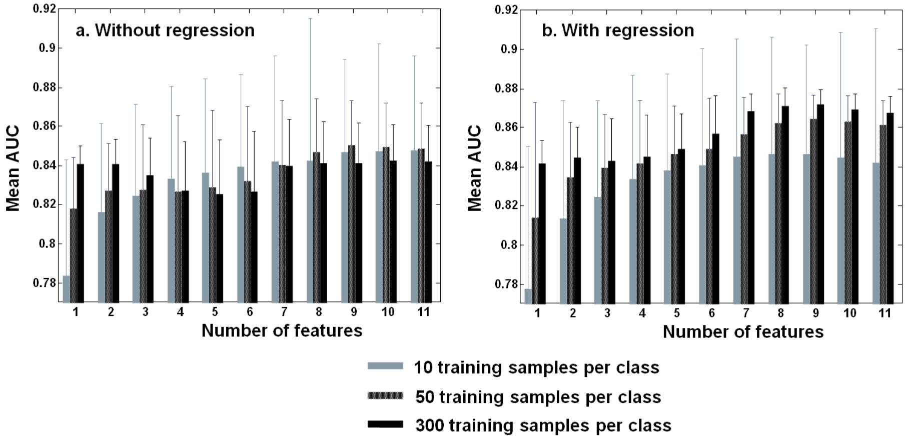

Figure 3.9: The average AUC of the Integrative Kernel Method without and with regression. The average AUCs of each training set size are shown as bars. The standard deviation is indicated by a whisker extending from the average AUC. **a)** The measured Product of Standard-deviations S is taken to be the certainty level $r_{i,d}$ directly. **b)** The regressed Product of Standard-deviations $\hat{S}$ is taken to be the certainty level $r_{i,d}$, which improves the classification performance of the Integrative Kernel Method, especially for larger sample sizes (same results as in Fig. 3.7).

Scaled Elliptical Kernel Method are shown in Fig. 3.10, from which we can see the effect of feature selection. The performance difference between the Integrative Kernel Method and the Scaled Elliptical Kernel Method is increased substantially after the feature selection as the pairwise t-test indicates. For 10, 50, and 300 samples with seven features, for example, the p-values are $3 \times 10^{-6}, 6 \times 10^{-11}$, and 2×10^{-5}, respectively, before feature selection; after feature selection, the p-values become $3 \times 10^{-21}, 7 \times 10^{-32}$, and 9×10^{-15}.

Regarding the other kernel methods, it is shown in Fig. 3.7 that the Scaled Elliptical Kernel Method is slightly better than the Spherical Kernel Method given a reasonable number of samples, e.g. more than $\sim$ 50 per class. Interestingly, the ML Elliptical Kernel Method behaves worst in almost all cases. In-depth investigation indicates that this result stems from the fact that the objectives and the corresponding criteria are different between the parameter estimation stage and the final application stage. That is, the kernel width is estimated to optimize the density estimation in terms of total likelihood, not the classification performance in terms of AUC. Therefore, when applied to the yeast co-expression data set and the Highleyman data set, it is not surprising that this approach outperforms the Spherical Kernel Method in density estimation. However, where classification performance is concerned, this approach suffers from overfitting in the more complex data set, i.e. the yeast co-expression data set, especially when the number of samples is limited.

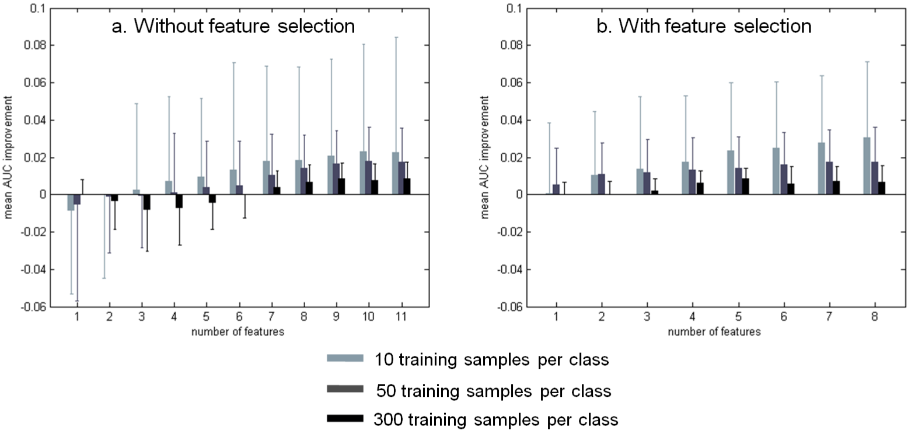

Figure 3.10: The average AUC improvement of the Integrative Kernel Method over the Scale Elliptical Kernel Method without and with feature selection. **a)** All 11 features are used (same results are used as in Fig. 3.7). **b)** Three features are removed in the feature selection.

3.6 Conclusions

This work exploits the prior knowledge about measurement noise in constructing kernel density based classifiers. A new kernel density based classifier which transforms this knowledge into kernel widths and weights is investigated, and compared to three basic kernel methods.

Our methodology can be used for any specific application –

Step 1: Estimate the certainty level r carefully, so that it represents the measurement noise level precisely and consistently. As we saw from the comparison of the two ways to estimate the certainty levels in the yeast co-expression data set, the regression procedure greatly improves the certainty level estimation, which eventually results in a better classification performance. A related important issue is feature selection, which means removing features for which it is difficult to estimate r well.

Step 2: Transform r into the kernel width h and weight w (see section 3.2.2), and perform classification using Eq. 3.4. This transformation model is general and not dependent on the application. Of course, the parameters should be estimated on the data set at hand.

Classification performance on the yeast co-expression data set demonstrates the power of utilizing the prior knowledge, in which the proposed Integrative Kernel Method significantly outperforms the basic methods. In particular, the integration of prior knowledge about measurement noise is especially beneficial in a relatively under-sampled data set when compared to the number of features. When the number of training samples approaches infinity, the Integrative Kernel Method will converge to the basic kernel methods. This is what we expect, since the characteristic of measurement noise can be fully represented by the infinite number of observed samples, and the benefit of prior knowledge will therefore disappear.

The methodology is applied on the kernel classifier in this work because of its interpretability and ability to incorporate the noise level easily. The concept can be applied to other pattern recognition techniques and problems which require data integration by transforming the noise knowledge of the problem into an embedded variable of an appropriate classifier, e.g. a sample-wise parameter for misclassification cost (C) in the support vector machine [6, 15].

Our work can be thought to fall into the category of robust statistics [50, 56, 95], as we seek to build a robust classifier based on noisy data which deviate from the true values. Moreover, the dependency functions $h^2_{i,d} \sim r_{i,d}$ and $w_{i,d} \sim r_{i,d}$ we propose resemble some Ψ functions of an M-estimator [50]. In general, most robust methods aim to cope with data sets contaminated by gross error or outliers, rounding and grouping errors, missing data, departure from an assumed sample distribution, and the random error which models the deviations occurring if multiple independent measurements were to be taken for the same quantity.

Here we have no assumption on the sample distribution, and we do not have multiple independent measurements for each feature of each sample. Our problem and methodology are different from the standard robust setting, in that we have only one observation (measurement) per sample per feature, but we do possess additional information about the observation quality. Assuming the random error of the measurement for each feature of each sample has a normal distribution, we use the additional information to estimate its width. Moreover, we also use the same prior knowledge to compute sample weights (i.e. influences), unlike standard robust methods. As a result, our method can greatly improve the efficiency of classifiers, since fewer measurements are required.

4

M-PAL: ALIGNING METABOLIC PATHWAYS BETWEEN SPECIES

Comparative analysis of metabolic networks in multiple species yields important information on their evolution, and has great practical value in metabolic engineering, human disease analysis, drug design etc. In this work and the next two chapters, we aim to systematically search for pathways conserved in two species, to quantify their similarities, and to focus on the variations between them.

Our method systematically investigates full metabolic networks of two species by exploring reaction arrangement possibilities, with the goal of identifying highly similar yet non-identical pathways which perform the same metabolic function. We present a clear framework for matching metabolic pathways, taking mismatches, gaps and crossovers into account. We also propose a scoring scheme which combines enzyme functional similarity with protein sequence similarity. This analysis helps to gain insight in the biological differences between species and provides comprehensive information on diversity in pathways between species and alternative pathways within species, which is useful for pharmaceutical and industrial bioengineering targets. The results also generate hypotheses for improving current metabolic networks or constructing such networks for currently unannotated species.

This chapter was published in the *Series on Advances in Bioinformatics & Computational Biology*, 2008 [86].

4.1 Introduction

The metabolic network of a species represents all known chemical reactions of metabolism within a cell. A single, relatively isolated cascade of such reactions is normally called a metabolic pathway. Most metabolic reactions are catalyzed by specific groups of enzymes. These enzymes are annotated by EC numbers [105], hierarchically organized numbers indicating the type(s) of reaction they catalyze. Studying the metabolic network is a powerful tool to elucidate the cellular machinery. Therefore, it has been an active research field for the last decade [12, 13, 19, 33, 92, 107, 142, 165].

Comparing pathways between multiple species provides valuable information to understand evolutionary conservation and variation. Kelley *et al.* [70] align protein interaction networks and predict protein function and interaction using conserved pathways. We extend their alignment concept to the metabolic level, to discover conserved metabolic pathways. Such a pathway transforms a specific substrate into a specific end product via very similar reactions in multiple species. These reactions are similar since they have common substrates and common products. However, they may have different co-substrates or co-products, be catalyzed by different enzymes, need different numbers of reactions to complete the transformation, or reactions may occur in a different order.

Although many comparative analyses at the metabolic level have been performed, little work focuses explicitly on the discrete differences between conserved pathways, and to our knowledge no global search has been carried out yet. For example, Forst *et al.* [33] perform a phylogenetic analysis on four pre-chosen pathways by combining the sequence information of a set of enzymes and gene-coded metabolites in a pathway. Dandekar *et al.* [19] also limit their study, to the glycolysis pathway. As for the similarity measure for matching pathways, Tohsato *et al.* [142] align pathways based on enzyme EC number similarity, discarding information on the involved metabolites. In Clemente *et al.*[12, 13], sets of reactions in multiple pathways are compared, omitting connectivity between the reactions.

Inspired by the PathBLAST algorithm of Kelley *et al.* [70], we propose a novel approach to align metabolic pathways. Our method, Metabolic Pathway ALignment (M-Pal), aligns entire metabolic networks of different species in order to explore highly conserved pathways. In the resulting aligned pathways, most reactions are identical; the remaining reactions are not identical, yet similar (see Fig. 4.1 for illustration). These conserved pathways are very likely to be essential or efficient pathways. More importantly, our method sheds light on differences between species in the use of non-identical but similar reactions, revealing between-species diversity and within-species alternatives. We introduce *diversity* in a pathway as a term indicating that each species has its own unique mechanism to allow a certain biochemical transformation to take place. If both species share a common reaction, but one of the species has a second, unique reaction to perform the same transformation, then this last transformation forms part of a unique *alternative* pathway. Fig. 4.2 gives a schematic explanation of these two terms, in which different types of arrows are used to indicate unique reactions of one species.

Diversity and alternatives across species give insight into biological differences between species, provide potential candidate enzymes for bioengineering, and generate hypotheses on missing enzymes or incorrect annotations in current metabolic net-

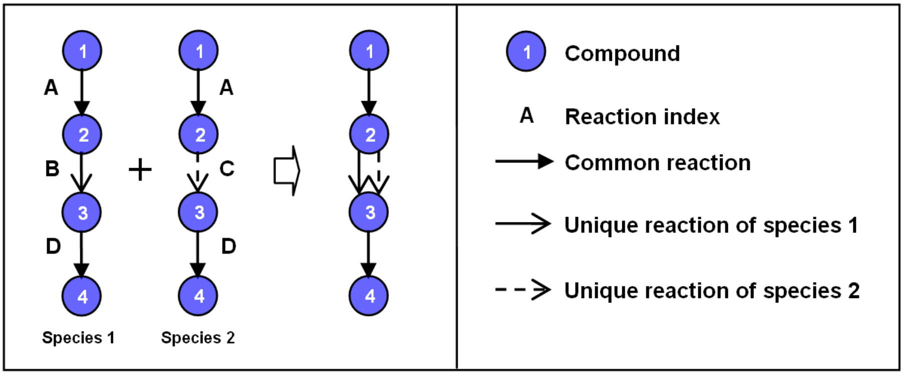

Figure 4.1: Illustration of our searching target. The pathways in two species share common reactions (A and D), but also have variations (B and C).

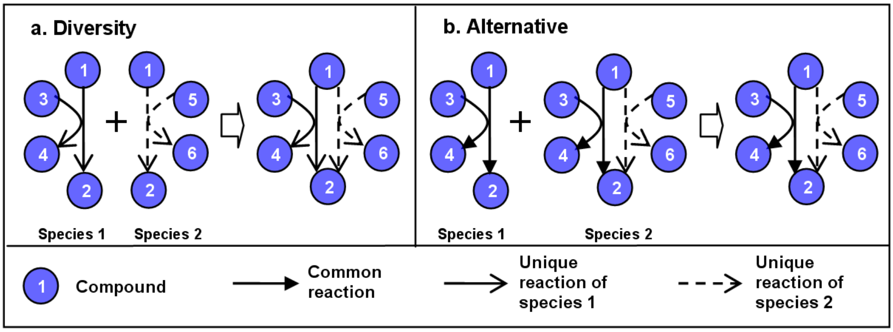

Figure 4.2: Illustration of diversity and alternative pathway. In each case, the reactions in both species are combined into a unified representation for conciseness.

works. Moreover, the resulting pathways give more options in pathway engineering and constructing metabolic networks for unannotated species. Finally, this method unites reactions in isolated metabolisms into a large network, relating reactions with upstream substrates and downstream products which might be elusive if we only look at a subset of the network.

We apply M-Pal to *Saccharomyces cerevisiae* and *Escherichia coli*, and find 2518 short conserved pathways. In each conserved pathway, 4-5 reactions from one species are aligned with similar reactions from another species. Among the results, $\sim$ 1500 pathways are diverse or contain unique alternative enzyme activities. We categorize the differences between pathways and refine the search result by scoring each pathway according to functional and sequence similarity of the enzymes involved. This scoring scheme enables us to focus on highly conserved pathways with similar enzymes. We show that a number of metabolic annotations can be attached to each of the resulting pathways, demonstrating the strength of our systematic search in unearthing novel cross-links in metabolic networks.

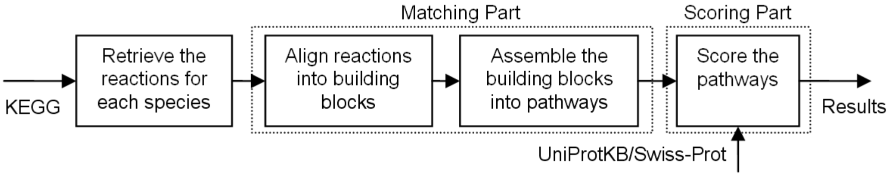
Matching Part
Scoring Part
KEGG
Retrieve the reactions for each species
Align reactions into building blocks
Assemble the building blocks into pathways
Score the pathways
Results
UniProtKB/Swiss-Prot

Figure 4.3: M-Pal flow chart.

We describe M-Pal in detail in section 4.2. The results are presented and discussed in section 4.3. Section 4.4 ends with some conclusions and an outlook to further work.

4.2 Method

Since we seek to investigate diversity and alternatives in highly conserved metabolic pathways, we align the pathways from two species into a conserved pathway in a rather strict way. That is, we align two pathways only if most of the involved reactions in these two species use similar enzymes to catalyze common substrates into common products, introducing only a limited amount of freedom into the alignment. More specifically, let P_1 and P_2 denote two metabolic pathways in two species containing reactions $[R_{11}, R_{12}, ..., R_{1L}]$ and $[R_{21}, R_{22}, ..., R_{2L}]$, respectively. P_1 and P_2 can be aligned into a conserved pathway only if the individual reactions are aligned in the right order. That is, R_{11} is aligned with R_{21}, R_{12} is aligned with R_{22} *etc.*, until R_{1L} is aligned with R_{2L}. We call each pair of matching reactions, e.g. R_{11} and R_{21}, a *building block*.

Given the restrictions mentioned above, we propose an efficient matching mechanism which constructs all building blocks first, and then assembles them into pathways of a desired length, taking reaction directions into account. After the aligned pathways are obtained, we compute an enzyme similarity score for each aligned pathway. In this way, we eventually get a list of conserved pathways, ordered by this score.

This sequential procedure of matching and scoring (see Fig. 4.3) ensures the search for all matching pathways is complete and allows for a flexible scoring function. The exhaustive search results can be pre-computed and, as scoring is performed separately, no potential match will be missed because of prematurely discarding a pathway in the search. Our method is explained in detail in the remainder of this section.

4.2.1 Reaction retrieval

We obtained the general reaction definitions from Release 42.0 of the KEGG LIGAND composite database [45], updated on May 14, 2007. For each species, we acquired the subset of reactions present in that species, together with the EC numbers and ORF names of the enzymes which catalyze each reaction, from the KEGG/XML and KEGG/PATHWAY databases.

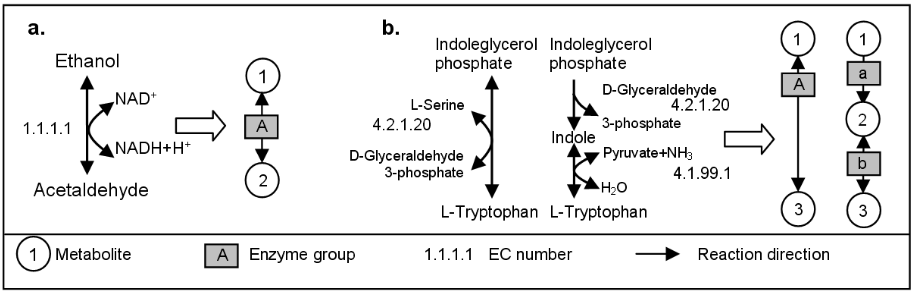

Figure 4.4: Reaction representation. **a)** Illustration of two representations of reactions in our method. **b)** One reaction from *S. cerevisiae* (on the left) and two reactions from *E. coli* (on the right) share a common substrate (Indoleglycerol phosphate) and product (L-Tryptophan). This situation forms one "gap", i.e. the difference in the number of reactions to transform Indoleglycerol phosphate into L-Tryptophan is one.

In M-Pal, reactions are represented as a combination of the classic "enzyme-centric" and "compound-centric" representations. Thus, a reaction is represented by all elements involved: metabolites, (a group of) enzyme(s), and its direction. Fig. 4.4a gives an example. To allow us to compare reactions from different species, we plot them next to each other, with the matching substrate or product in the same row. Sometimes, a single reaction and a series of reactions connected in tandem may share common substrates and products. This introduces "*gaps*", indicating that the number of reactions to transform the specific substrates into the specific products differs between species. Fig. 4.4b illustrates this: one reaction from *S. cerevisiae* and two reactions from *E. coli* form a "gap".

4.2.2 Building block alignment

Two reactions R_{1l} and R_{2l} can be aligned to form a building block when they have a common substrate and a common product, and at least one pair of enzymes (one from each species) share functional similarity such that the first two digits of their EC numbers are the same. Note that a reaction can be catalyzed by a group of enzymes, which may have multiple EC numbers. By allowing some variation, we introduce a number of *building block types* (see Fig. 4.5). If R_{1l} and R_{2l} are identical, i.e. the same reaction is present in both species, the resulting building block is called "*identical*" (*i*). If R_{1l} and R_{2l} are different reactions, because of different co-substrates or co-products according to the definition in section 4.2.1, they form a "*direct*" building block (*d*). To incorporate alternative pathways, evolutional diversity and annotation errors, we also allow one "mismatch" or one "gap" in a building block. Thus, in an "*enzyme mismatch*" building block (*em*), the first two digits of the EC numbers of the enzymes involved are not the same. The building blocks containing one "gap" are "*direct-gap*" (*dg*) and "*enzyme mismatch-gap*" (*eg*). Furthermore, we include "*enzyme crossover match*" building blocks (*ec*) to accommodate possible variation in the order of the catalyses: there are two reactions in each species sharing common substrates and end products with the EC numbers of the first and second reaction in one species being similar to those of the

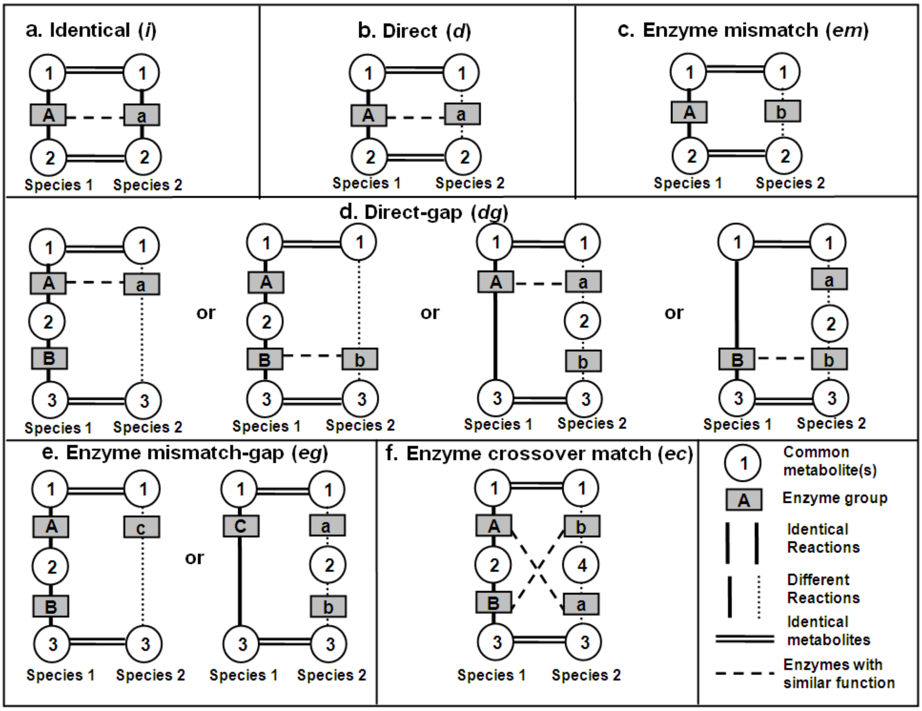

Figure 4.5: Illustration of the six types of building blocks. The reaction directions are omitted in the figure for simplicity. A dashed link is drawn between two groups of enzymes if they share the same first two digits of their EC numbers.

second and first reaction in the other species, respectively.

To summarize, the reaction alignment method described above results in six types of building blocks, each containing one or two reactions from each species. Note that 26 "current metabolites" [92, 165], listed below [1], were excluded from consideration as common substrate or product to avoid finding large numbers of trivial conserved pathways.

4.2.3 Pathway assembly

Next, we focus on finding conserved short acyclic pathways. We only assemble four building blocks into a pathway, ensuring that one reaction does not appear more than once in a pathway. Moreover, we demand that out of these four building blocks, at least three must be of type "identical" or "direct", representing the conserved part of the pathway. Only a single building block of type "enzyme mismatch", "direct-gap",

---

[1] ATP, ADP, UTP, UDP, GTP, GDP, AMP, UMP, GMP, NAD, NADH,NADP, NADPH, Acetyl-CoA, CoA, Propanoyl-CoA, L-Glutamine, L-Glutamate, 2-Oxoglutarate, CTP, CDP, CMP, H_2O, CO_2, NH_2, Phosphate.

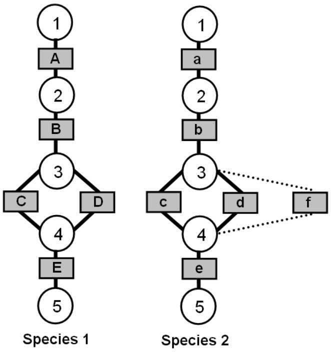

Figure 4.6: Illustration of the removal of redundant pathways. See Fig. 4.5 for legends. Six possible pathway alignments can be induced in this example (each reaction is represented by the corresponding enzyme groups): **(1)** Reactions A-B-C-E of species 1 with a-b-c-e of species 2, obtaining an *"i-i-i-i"* alignment. **(2)** Reactions A-B-D-E of species 1 with a-b-d-e of species 2, obtaining an *"i-i-i-i"* alignment. **(3)** Reactions A-B-C-E of species 1 with a-b-d-e of species 2, obtaining an *"i-i-x-i"* alignment, where x indicates one of the five non-"identical" building block types. This alignment is redundant with (1) and (2). **(4)** Reactions A-B-D-E of species 1 with a-b-c-e of species 2, which is also redundant with (1) and (2). **(5)** Reactions A-B-C-E of species 1 with a-b-f-e of species 2, obtaining an *"i-i-x-i"* alignment. This is a novel alternative pathway, since reaction f is unique in species 2, hence *"i-i-i-i"* alignment is impossible. **(6)** Reactions A-B-D-E of species 1 with a-b-f-e of species 2 also is a novel pathway. In the end, four aligned pathways are obtained: (1), (2), (5) and (6).

"enzyme mismatch-gap" or "enzyme crossover match" is allowed in a pathway. Abbreviations are used to denote the pathway composition of building blocks regardless of the order, e.g. *"i-i-i-d"* indicates a pathway with three reactions of type "identical" and one of type "direct", in any order. In total, there are 21 such compositions possible for pathway alignment. These are used as 21 *pathway categories* in the discussion of our results.

To enhance the informativeness of our resulting set of pathways, we remove some redundant pathways. First, building blocks whose substrate and product are identical in one species (after removing current metabolites) will not be selected to construct a pathway. Furthermore, we reduce the redundancy in the result by enforcing uniqueness in choosing the building blocks of the five types other than "identical", see Fig. 4.6. A non-"identical" building block can be chosen only if it contains at least one reaction absent in one of the species. This is because if all reactions in the building block are present in both species, two building blocks of type "identical" will already be constructed. Consequently, any other combinations of these reactions are redundant. Conversely, a reaction unique to one species provides an interesting novel alternative pathway.

4.2.4 Scoring function

Two factors indicate the extent to which an aligned pathway is conserved. One is the pathway category, i.e. the building block composition. For instance, we consider an *"i-i-i-d"* pathway to be more conserved than an *"i-i-i-dg"* pathway. The other factor is enzyme similarity, which we evaluate here based on functional similarity (EC numbers) and sequence similarity. Since they are not fully correlated, we integrate them to introduce a more informative measure of true orthology. In the following, we explain how to calculate functional similarity and sequence similarity of a building block, followed by their integration.

Given a building block containing one reaction from each species, enzyme functional similarity E_f is taken to be the maximum number of digits of EC numbers that the two groups of enzymes share. This is a simple and straightforward manner to measure enzyme functional similarity [53, 89], since EC numbers form a functional hierarchy. Although more complex methods exist [13, 142], their validity is still under research. Let the EC numbers in the reaction for species 1 be EC_{11}, EC_{12}, ..., EC_{1m}, and for species 2 EC_{21}, EC_{22}, ..., EC_{2n}, we count the number of shared digits for each possible pair of EC numbers, and use the maximum as the functional similarity E_f for this building block. For "direct-gap" and "enzyme mismatch-gap" building blocks, for which one group of enzymes should be compared to two groups of enzymes, we compute E_f for both pairs of groups, and choose the larger E_f. For "enzyme crossover match" building blocks, E_f is taken to be the averaged value of the crossover enzyme group comparisons.

For the sequence similarity E_s between two reactions, we take the minimum BLAST E-value between all possible enzyme pairs. For "direct-gap" and "enzyme mismatch-gap" building blocks, E_s is computed between the two groups of enzymes which have the larger E_f. For "enzyme crossover match" building blocks, E_s is averaged. BLAST (version 2.2.15) is performed with $e = 100$ on the protein sequences in UniProtKB / Swiss-Prot Release 51.6.

After computing the E_f and E_s scores for all building blocks in a pathway, we sum all E_fs in a pathway and transform the result into a score $S_f \in [0, 1]$; likewise for all E_ss in the pathway to obtain $S_s \in [0, 1]$. Tables 4.1 and 4.2 detail these transformations. Since the original values of E_f and E_s have very different ranges, this transformation step actually scales these two measures into the same range in a sensible way, so that they are comparable and easy to combine. The intervals in the transformation tables are chosen to reflect our objective in finding conserved pathways with similar enzymes: high functional similarity values are examined in more detail in the score. For sequence similarity, we focus on the traditional cutoff value 10^{-2} for weak sequence similarity [70], thus the intervals around 10^{-2} are smaller than those for high sequence similarities. We do not restrict ourselves to highly similar sequences because our main interest is to reveal the alternatives and diversities in the pathways. Since the maximum value for E_s is 100 (due to the parameter setting used for BLAST), the intervals for $S_s \geq 0.8$ indicate the number of building blocks with very dissimilar enzyme sequences.

Finally, the two scores are summed so as to combine the functional and sequence similarity:

$\sum_{b\in P} E_f(b)$	16	15.5	15	14.5	14	13.5	13	12	11	10	8
S_f	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1

Table 4.1: Transformation of the total functional similarity $\sum_{b\in P} E_f(b)$ into the score S_f.

$\sum_{b\in P} E_s(b)$	$(0, 10^{-80})$	$[10^{-80}, 10^{-60})$	$[10^{-60}, 10^{-40})$	$[10^{-40}, 10^{-20})$
S_s	0	0.1	0.2	0.3
$\sum_{b\in P} E_s(b)$	$[10^{-20}, 10^{-10})$	$[10^{-10}, 10^{-6})$	$[10^{-6}, 10^{-2})$	$[10^{-2}, 100)$
S_s	0.4	0.5	0.6	0.7
$\sum_{b\in P} E_s(b)$	$[100, 200)$	$[200, 300)$	$[300, \infty)$	
S_s	0.8	0.9	1	

Table 4.2: Transformation of the total sequence similarity $\sum_{b\in P} E_s(b)$ into the score S_s.

$$S(P) = S_f\left(\sum_{b\in P} E_f(b)\right) + S_s\left(\sum_{b\in P} E_s(b)\right) \tag{4.1}$$

in which b denotes a building block and P denotes an aligned pathway. The lower this score, the more similar the enzymes in P are.

4.3 Results and discussion

Of 881 enzymatic reactions in *S. cerevisiae* and 1106 in *E. coli*, 588 reactions are present in both species (Fig. 4.7a). Based on the total of 1399 unique reactions, six types of building blocks are assembled into 2518 unique pathways of length 4. Fig. 4.7b shows the number of reaction involved in the resulting pathways. Table 4.3 summarizes the number of building blocks of each type found. These results indicate that the reactions and building blocks in the resulting pathways reasonably cover all available reactions and building blocks, demonstrating the strength of our systematic search.

For each pathway category containing a specific composition of building blocks, the total number of resulting pathways is shown in Fig. 4.8a, and their average functional similarity score S_f and sequence similarity score S_s are shown in Fig. 4.8b. As shown in Fig. 4.8a, $\sim$ 1000 completely conserved pathways of type *"i-i-i-i"* are found. Not surprisingly, their enzyme sequences are highly similar, with BLAST E-values ranging from 10^{-10} to 10^{-6} on average. The pathway with the best score, 0, is depicted in Fig. 4.9a. However, the variance of the sequence similarity score is also large, indicating that some reactions in these pathways do not have enzymes with similar sequences.

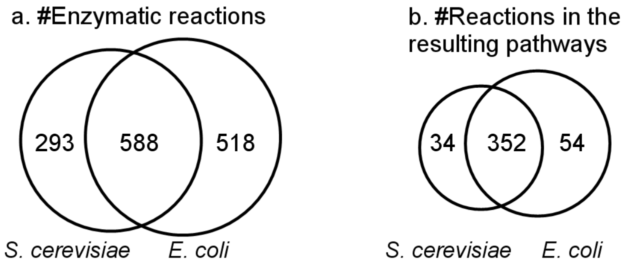

Figure 4.7: Venn diagrams showing **a)** the total number of enzymatic reactions in the two species and **b)** the number of reactions involved in the results.

Type	Identical (*i*)	Direct (*d*)	Direct -gap (*dg*)	Enzyme mismatch (*em*)	Enzyme crossover match (*ec*)	Enzyme mismatch-gap (*eg*)
♯Building blocks	516	116	108	27	40	52
♯Building blocks in the resulting pathways	352	67	64	11	12	29

Table 4.3: The number of each of the six types of building blocks.

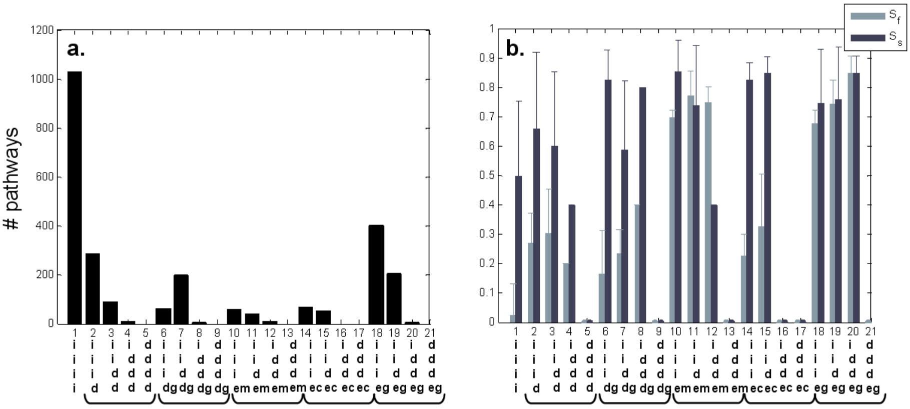

Figure 4.8: **a)** Total number of pathways in 21 pathway categories. Note that long conserved pathways may result in multiple short overlapping pathways. **b)** The average enzyme functional similarity score and sequence similarity score of each pathway category. Whiskers indicate standard deviations.

This might arise because of different specificity, horizontal gene transfer, gene fusions, or the fact that only subunits of the enzymes are the same.

We also found ∼ 1500 highly conserved pathways which contain some diversity be-

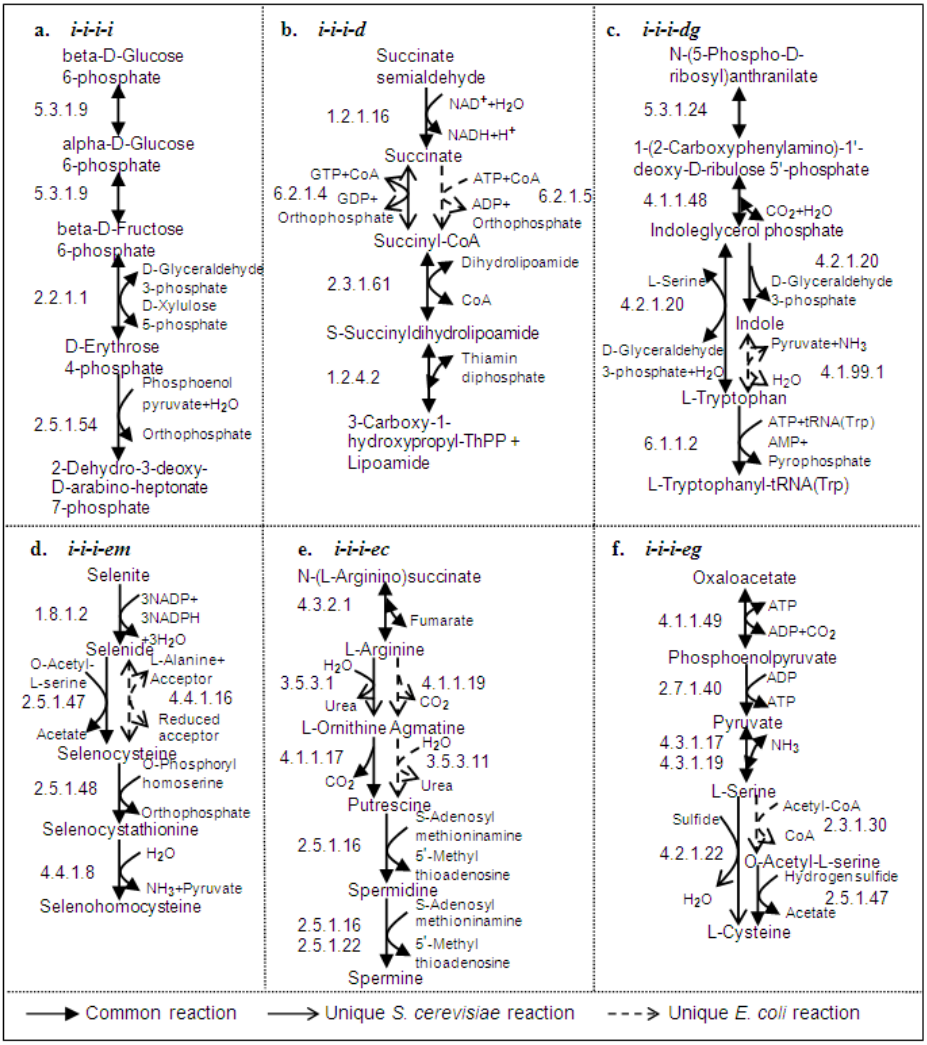

Figure 4.9: The pathways with the best scores in categories 1, 2, 6, 10, 14, and 18 of Fig. 4.8. **a)** The pathway with the best score ($S = 0$) in the results. It has an *"i-i-i-i"* alignment. **b)** One of the pathways with the best score ($S_f = 0.2, S_s = 0.1$) within category *"i-i-i-d"*. **c)** The pathways with the best score ($S_f = 0, S_s = 0.5$) within category *"i-i-i-dg"*. **d)** One of the pathways with the best score ($S_f = 0.7, S_s = 0.6$) within category *"i-i-i-em"*. **e)** The pathways with the best score ($S_f = 0.2, S_s = 0.7$) within category *"i-i-i-ec"*. **f)** One of the pathways with the best score ($S_f = 0.7, S_s = 0.3$) within category *"i-i-i-eg"*.

tween both species or unique alternatives within one species. Each of these pathways has a building block of type "direct", "direct-gap", "enzyme mismatch", "enzyme mismatch-gap", or "enzyme crossover match". Examples are given in Fig. 4.9b-4.9f. These pathways are of great interest in bioengineering as they manifest the hidden information about pathway diversity and alternatives, which will not be found if we only look at a subset of the metabolic network in one species.

The results are useful in many applications. First, some resulting pathways suggest a more exact EC number annotation of their enzymes is possible and call for detailed comparison of the enzymes. For example, the enzymes in the pathways of type "*i-d-d-em*" in Fig. 4.8b have dissimilar EC numbers, but their sequences are actually very similar (low S_s and high S_f). They might be incorrectly annotated, since they both transform a common substrate into a common product. Another example is given in Fig. 4.9c, in which the enzymes with EC number 4.2.1.20 in *E. coli* (trpA and trpB) could also be annotated as 4.1.2.8, which is the α-subunit of 4.2.1.20. Comparing the enzymes in alternative pathways in different species can also be beneficial to understand their structural difference and relationship. In Fig. 4.9c for instance, the two enzymes in *E. coli*, 4.2.1.20 and 4.1.99.1, might be different subunits of the enzyme 4.2.1.20 in *S. cerevisiae*. The same can be observed in Fig. 4.8b, where the sequence similarity in the pathways with "*dg*" is generally worse than in those with "*d*" only, implying that the enzymes in "*dg*" are only subunits of the corresponding enzymes in "*d*".

Second, the results can help to understand diversity in metabolism and evolution. Reactions which are unique to one species are highlighted in Fig. 4.9. Investigation of the biological difference between the two species is expected to explain their uniqueness. Further, we can project the knowledge to a new species. For instance, if the new species has the enzymes which catalyze a unique reaction of *S. cerevisiae*, then probably they are very closely related in the phylogenetic tree, and therefore share more common properties. Nevertheless, the revealed diversity might be an artifact of current metabolic network databases. Therefore it is recommended to examine whether the other species also has this unique enzyme, or whether some enzymes (and reactions) are missing in the pathways with "gaps". Another interesting result which might be worthy of further research is shown in Fig. 4.8b, for the group containing enzyme crossover match building blocks (*ec*). Although the crossover enzymes have similar functions, their sequences are very dissimilar. Possible reasons could be that the enzymes have different substrate specificities, or the intermediate substrates are very different. They could also have been isoenzymes in parallel pathways, having become specialized to one species in evolution.

Third, the unique alternative pathways revealed by M-Pal provide potential candidate enzymes for bioengineering. Certain natural enzymes can be removed or changed so that we can choose between different alternative pathways, or enforce the reaction direction to produce the product of our interest. In the pathway shown in Fig. 4.9c, *E. coli* has two alternative pathways to transform Indoleglycerol phosphate into L-tryptophan, one being reversible (catalyzed by 4.2.1.20) and the other one reported to be irreversible (catalyzed by 4.2.1.20 and 4.1.99.1). If the enzymes of 4.2.1.20 in the irreversible pathway are indeed also possibly annotated as 4.1.2.8, we can remove the 4.2.1.20 enzyme activity to enforce the direction towards producing tryptophan, which is an essential amino acid in human nutrition [157].

Finally, our results provide additional opportunities to construct the metabolic networks for currently unannotated species. As discussed above, our method points out possible missing enzymes and suggests related enzymes in well-studied species. The alternative pathways also provide more possibilities for optimizing the network to fit the found enzymes and reactions better.

4.4 Conclusions

The systematic search of M-Pal associates different parts of metabolic networks with each other and combines information from multiple species to discover diversity and alternatives in highly conserved pathways. The results shed light on the small differences found in the conserved pathways and provide useful information for many applications. Gene knock-out experiments can be performed to test our hypotheses, and the essentiality of the resulting pathways should be examined.

Our research is still at an early stage, and can be refined in a number of ways. Possible extensions include increasing the freedom in the alignment, e.g. allowing for more gaps or mismatches, further separated crossover matches, and longer pathways. This implies the search algorithm will have to become more sophisticated, as exhaustive enumeration will become infeasible. Next, the scoring function can be modified to prefer certain types of alignment. Non-identical metabolites could be included in the matching, implying a need for a compound similarity measure to be added to the scoring function. The enzyme sequence similarity measure could also be refined using protein domain information. The current scoring mechanism assumes functional and sequence similarity is equally important. Weights could be added to model a trade-off between the two [13]. The scoring function itself could be enhanced by using a probabilistic framework such as in Kelley *et al.* [70], allowing us to look for relatively rather than absolutely conserved pathways and to attach a p-value to the pathways found. Other possible enhancements to the score are to take reversibility of reactions and the presence of isoenzymes into account.

Currently, this method is performed on two species only and is expected to give more informative results if applied on species not closely related. An extension could be to apply M-Pal on multiple species, at different evolutionary distances. We expect that larger differences will be found as evolutionary distance increases. The results will give insight to understand evolution and specialization, provide new building blocks and alternatives for pathway engineering, and be of great value for prediction of unannotated genes.

5

M-PAS: MEASURING METABOLIC PATHWAY SIMILARITIES

In the previous chapter we presented a pathway alignment framework to align metabolic pathways (M-Pal). In this chapter, we extend this framework with a scoring scheme which is able to quantify the level of conservation of aligned pathways in a comprehensive and flexible manner. The scoring function compares all components of two pathways by measuring similarities between substrate sets, product sets, enzyme functions, enzyme sequences, and alignment topology. These individual similarity measures are then integrated into a single score in a hierarchical way, which enables us to weight the individual similarity measures in order to express different biological emphases.

Using M-PAS, we detected 2597 length-four conserved pathways between *Saccharomyces cerevisiae* and *Escherichia coli*. The proposed scoring function ranks these pathways given five biological motivations and reveals the diverse similarity fingerprint of each type of alignment. Not surprisingly, parts of primary metabolism are found to be abundant in our top-scoring pathways.

This chapter was published in *BMC Systems Biology*, 2008 [85].

5.1 Introduction

Comparative analysis of metabolic networks in different species yields information important for both biology (understanding evolution/speciation, annotating new genomes etc.) and life science applications (e.g. in biotechnology, pharmacology). Therefore, it has been an active research field for the last decade. For example, Dandekar *et al.* [19] combined biochemical data analysis, elementary flux mode analysis and comparative genome analysis to compare glycolytic pathways in 17 species. Jeong *et al.* [66] and Ravasz *et al.* [120] studied the global topological properties of the metabolic networks in 43 species. In addition, Küffner *et al.* [78] used Petri nets to compare database contents and define differential metabolic displays (DMDs), which allow to compare metabolic networks by identifying intersection and difference sets of reactions. As one of the applications, Heymans *et al.* [53] derived phylogenetic trees based on metabolic pathway comparison. Guimerà *et al.* [48] analyzed the modularity of the metabolic networks of 18 organisms, and classified metabolites and enzymes based on their roles in connecting different functional modules. Díaz-Mejía *et al.* [24] investigated the relation of network modularity and distance between reactions with the retention of gene duplicates in various species and databases. More generally, a review on biological network comparison problems, techniques and applications is given by Sharan *et al.* [130].

In studies up till now, however, only little work focused explicitly on the variations between species in conserved pathways, and to our knowledge no alignment of entire networks, exploiting all reaction arrangement possibilities, has been carried out yet. Moreover, the similarity measures used to align metabolic pathways is often not comprehensive, as compounds or network structure are neglected. For example, Tohsato *et al.* [142] align pathways based on enzyme EC number similarity only, discarding information on the compounds involved. Yang *et al.* [158] perform path matching and graph matching to query certain metabolic pathways or subgraphs in a predefined graph, but also use a similarity measure based on EC numbers only. Although Forst *et al.* [33] define the distance between pathways as a combination of distances between compounds and distances between enzymes, they only consider sequence similarity, and the compounds are limited to amino acids. In [13], sets of reactions in multiple pathways are compared, omitting the connectivity between the reactions. Finally, the pathway similarity score in [53, 116, 164] combines EC number similarity and network topology, but does not include compounds, and alignments are between predefined sub-networks only. Therefore, the comparison is limited to conventional pathways, and different parts of the cellular metabolism are not associated with each other.

In this work, we align entire metabolic networks of two species and quantify their similarities comprehensively, to identify highly conserved pathways. We particularly focus on the variations in these pathways, as illustrated in Fig. 4.1. Here a pathway is defined as a series of chemical reactions of metabolism within a cell (see also section 4.1). Therefore they are not necessarily routes through the network from uptake to secretion, as represented by many conventional pathway representations.

A naive approach to find conservation and variations between metabolic networks would be to search for common reactions and reaction pairs, using different cofactors or enzymes in the two species. Besides being inefficient, this approach isolates reactions from their upstream and downstream processes. Instead, we search for conserved *pathways*, rather than single *reactions*. In this way, we place the reactions in their

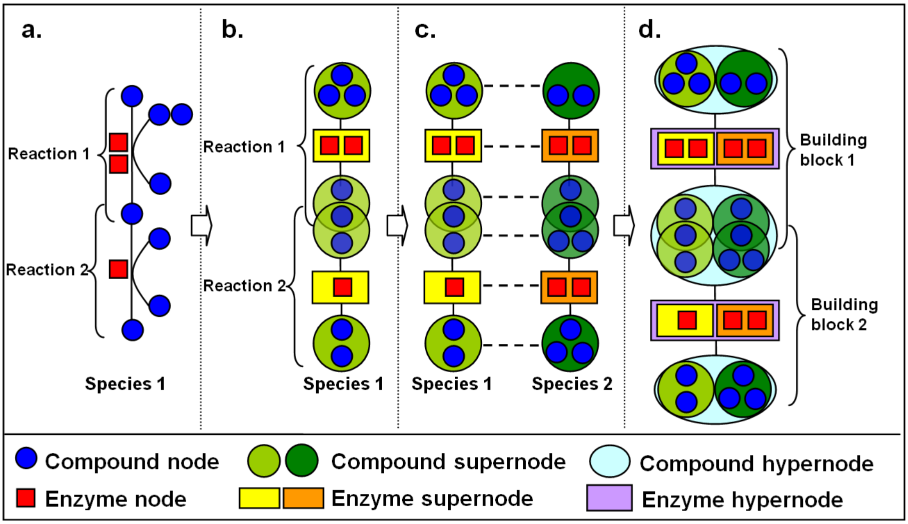

Figure 5.1: Overview of the alignment method. First, compound nodes and enzyme nodes (**a**) are generalized into compound supernodes and enzyme supernodes (**b**). Two reactions of species 1 are aligned with two reactions of species 2 (**c**), by pairing the supernodes into compound hypernodes and enzyme hypernodes (**d**). Each pair of aligned reactions forms a building block, from which an aligned pathway can be assembled. The reaction directions are omitted in the figure for simplicity.

metabolic functional context, which helps to 1) filter out isolated reactions not involved in pathways, 2) provide more evidence to claim part of a pathway is conserved, given that neighboring reactions are conserved, 3) interpret the resulting pathways.

Our method is designed to conduct this process efficiently and comprehensively. More specifically, our pairwise pathway alignment is based on a mechanism we proposed earlier [86], which is inspired by the alignment concept of [70]. It first aligns two to four similar reactions in two species into building blocks, and then assembles these into pathways of a desired length (Fig. 5.1). In each building block, a specific substrate is transformed into a specific product via similar but not necessarily identical reactions in two species. That is, they may have different co-substrates or co-products, be catalyzed by different enzymes, need different numbers of reactions to complete the transformation, or reactions may occur in a different order. In other words, our method enables to explore topological arrangement possibilities of reactions both between species (by building block assembly) and within species (by pathway assembly).

Further, we rank the aligned pathways according to their similarities (i.e. level of conservation), which prioritizes them for further investigation. To this end, a novel scoring function is proposed, which forms the core contribution of this chapter. It compares all components of two pathways by measuring similarities between substrate sets, product sets, enzyme functions, enzyme sequences, and alignment topology. The resulting individual similarity measures are then integrated into a single score. This scoring function has a generic form and is flexible enough to address various biological questions, by selecting different parameter settings.

5.2 Method

We align the pathways from two species in a strict way, in order to investigate highly conserved metabolic pathways, i.e. pathways with very similar structure and limited variation between species. More specifically, two metabolic pathways can be aligned into a conserved pathway only if their individual reactions transform common substrates into common products in each step. We call such a pair of matching reactions a building block (BB). Next, these building blocks are assembled into pathways of a specified length, taking reaction directions into account. Finally, we compute the similarity score for each aligned pathway, and obtain interesting pathways as those pathways that have high similarity scores.

5.2.1 Reaction representation

In M-PAS, reactions are represented at three levels of generalization: nodes, supernodes and hypernodes, respectively (see Fig. 5.1). The low-level representation gives the finest details of reactions, in which each compound and each enzyme constitutes a node (Fig. 5.1a). The medium-level representation generalizes reactions, so that all substrates and products of a reaction compose two compound supernodes, and all enzymes in that reaction form an enzyme supernode (Fig. 5.1b). Such a generalized representation is useful due to the multiple-to-multiple property of metabolic reactions, i.e. multiple substrates can be catalyzed by multiple enzymes into multiple products [77, 130]. Finally, at the high-level representation, the corresponding compound supernodes and enzyme supernodes from two aligned reactions are combined into compound hypernodes and enzyme hypernodes, respectively (Fig. 5.1c-d).

These different levels of representation enable the comparison of reactions in a detailed yet flexible manner. Thus, a particular compound node can be part of various compound supernodes given different co-factors in different reactions, and further can be part of various compound hypernodes due to different alignments with other compound supernodes. The same holds for enzyme nodes. This flexible representation not only reflects the versatility of the metabolic network conveniently, but is also necessary in order to express and quantify the similarity of reactions, which will be explained in section 5.2.3.

5.2.2 Reaction alignment

The reaction alignment part is proposed in our previous work [86] and is briefly explained here for comprehensibility and completeness of our methodology. Two reactions can be aligned to form a building block when they have at least one common substrate node and one common product node (Fig. 5.1d). To allow for some variation, we introduce six types of building blocks (see Fig. 5.2, which uses different legends than Fig. 4.5). If the same reaction is present in both species, the resulting building block is called "identical" (*i*). If the two reactions are different, but the first two digits

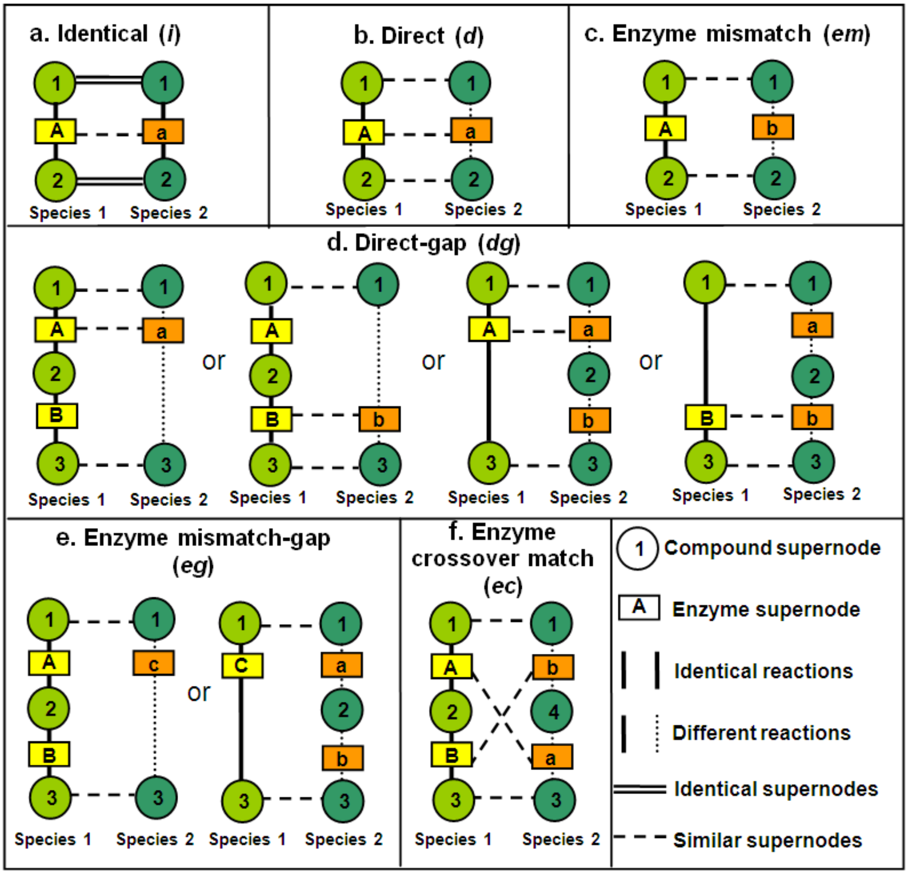

Figure 5.2: Illustration of the six types of building blocks. The reaction directions are omitted in the figure for simplicity. Two compound supernodes are considered similar if they share at least one common compound node. Two enzyme supernodes are considered similar if there exists a pair of enzymes which share the same first two digits in their EC numbers.

of the EC numbers of their enzymes are the same, they form a "direct" building block (*d*).

We allow for up to one mismatch or one gap in a building block, in order to incorporate alternative pathways, evolutionary diversity and annotation errors. That is, in an "enzyme mismatch" building block (*em*), the first two digits of the EC numbers of their enzymes are not the same. Gaps occur when a single reaction and a series of reactions connected in tandem share common substrates and products, indicating that the number of reactions to transform the specific substrates into the specific products differs between species. The building blocks containing one gap are "direct-gap" (*dg*) and "enzyme mismatch-gap" (*eg*). Finally, we include "enzyme crossover match" building blocks (*ec*) to accommodate possible variations in the order of the catalysis. That is, apart from sharing common substrates and end products in two reactions in each species, the first two EC number digits of the first and second reaction in one species are the same as those of the second and first reaction in the other species, respectively.

To enhance the informativeness of these resulting pathways, we add a constraint to avoid redundant building blocks. That is, a non-identical building block can be constructed only if it contains at least one unique reaction in one of the species, which is absent in the other species. This is explained in Fig. 4.6.

5.2.3 Scoring function

We set out by specifying a number of criteria for the design of the scoring function. First, similarities of all reaction components should be considered: substrate sets, product sets, enzyme functions and enzyme sequences, respectively. Second, the scoring function should be flexible and adaptable according to the user's biological interests. For example, the user might want to find pathways containing a particular structure (e.g. with a gap); or focus on enzymes only, but not on compounds; or seek to find a completely alternative pathway in which the enzymes are very dissimilar between two species. Third, since we aim to investigate many aspects of an aligned pathway and obtain multiple similarity scores, a reasonable way of integrating these is required. Finally, we should consider specificity in computing similarities, since both distributions of compound connectivity and enzyme EC number hierarchy show large variation [18, 142], i.e. some compounds and EC subclasses appear more often than the others in the background.

1) **Total score** According to the criteria above, we first compute similarity scores independently for all compound hypernodes and enzyme hypernodes in an aligned pathway, taking all aspects into account. These are then converted into z-scores before integration to account for their diverse distributions.

Let $Z(x)$ denote the z-score of x. Then $Z(P)$ is the total z-score for an aligned pathway P, a weighted sum of the scores of N building blocks B in P:

$$\begin{aligned} Z(P) &= \frac{1}{\sqrt{2N}} \sum_{\forall B \in P} [Z_0(B) + Z(B)] \\ &= \frac{1}{\sqrt{2N}} \sum_{\forall B \in P} \left[Z_0(B) + \frac{1}{\sqrt{\omega_c^2 + \omega_e^2}} \left(\omega_c Z(C_B) + \omega_e Z(E_B) \right) \right] \end{aligned} \tag{5.1}$$

$Z(B)$ is the z-score for a building block B. Let c and e denote a compound hypernode and an enzyme hypernode respectively, and denote the set of all c's and e's in a building block B by C_B and E_B, respectively. Users can define a preferred building block structure by assigning different biases ($Z_0(B)$) to different building block types. For example, if building blocks with gaps are preferred in a query, then these types of building block can be assigned a large positive bias. Weights $\omega_c, \omega_e \in [0, 1]$ can be used to assign different relative importance to compound similarity and enzyme similarity (resembling the α parameter in [13]).

Note that the z-scores are hierarchically combined using Liptak-Stouffer's method [52, 59]. In the following we explain how to compute $Z(C_B)$ and $Z(E_B)$ in detail.

2) **Compound similarity** $Z(C_B)$ is composed of compound similarities $Z(c)$ of the two compound hypernodes in the building block (i.e. the substrate hypernode and product hypernode). We express $Z(c)$ in two terms:

$$Z(C_B) = \frac{1}{\sqrt{2}} \sum_{\forall c \in C_B} Z(c) = \frac{1}{\sqrt{2}} \sum_{\forall c \in C_B} \frac{1}{\sqrt{2}} \left[Z_A(c) + Z_S(c) \right] \tag{5.2}$$

The *agreement* $Z_A(c)$ is the extent of the overlap in number of compounds between the two aligned compound supernodes. This is computed as the probability of observing the amount of overlap between the two compound supernodes by chance, according to a hypergeometric distribution [57]:

$$P_A(c) = \frac{\binom{|c_1|}{|c_1 \cap c_2|}\binom{|c_1 \cup c_2| - |c_1|}{|c_2| - |c_1 \cap c_2|}}{\binom{|c_1 \cup c_2|}{|c_2|}} = \frac{\binom{|c_1|}{|c_1 \cap c_2|}}{\binom{|c_1 \cup c_2|}{|c_2|}}, \tag{5.3}$$

where c_1 and c_2 denote the compound supernodes that form c, and $|x|$ denotes the number of compound nodes in x.

Next, this probability is transformed to a z-score:

$$Z_A(c) = \frac{P_A(c) - \mu_{AC}}{\sigma_{AC}}, \tag{5.4}$$

where μ_{AC} and σ_{AC} are the mean and standard-deviation of $P_A(c)$ over all possible compound supernode pairs, which represent the expected amount of overlap when the pairing would be random.

The other term is $Z_S(c)$, the *specificity* of the overlap when compared to all possible supernode pairs. That is, if two compound supernodes have overlapping compounds, we take into account the frequency of obtaining this particular overlap at random. We consider two sets of substances to be more similar if the overlapping part is more specific, i.e. not observed frequently by chance. Moreover, considering specificity of compounds may result in more biologically meaningful pathways, since metabolic pathways seem to represent paths through the least "promiscuous" compounds [18].

Suppose there are in total m compound supernodes in species 1 and n in species 2. Then we have:

$$P_S(c) = 1 - \frac{\#\text{observed } (c_1 \cap c_2) \text{ in the intersection}}{mn}, \tag{5.5}$$

$$Z_S(c) = \frac{P_S(c) - \mu_{SC}}{\sigma_{SC}}, \tag{5.6}$$

where μ_{SC} and σ_{SC} are the mean and standard-deviation of $P_S(c)$ computed over all $m \times n$ compound supernode pairs. The numerator in Eq. 5.5 is the number of times the specific overlap in compound node in c, i.e. $(c_1 \cap c_2)$, is observed in the intersections of all possible compound supernode pairs.

3) **Enzyme similarity** The enzyme hypernode similarity score, $Z(E_B)$, is defined by a functional similarity score $Z_F(e)$ and a sequence similarity score $Z_Q(e)$. In addition, users can specify weights $\omega_f, \omega_q \in [-1, 1]$ for the functional and sequence similarity scores to indicate their relative importance. Setting these weights to negative values actually enables us to search for dissimilar enzymes, which associates reactions with different mechanisms and provides more possibilities to annotate new species. For generality, suppose there are k enzyme hypernodes in building block B ($k = 2$ for "enzyme crossover match" building blocks, $k = 1$ for others). The enzyme similarity is then given by:

$$\begin{aligned} Z(E_B) &= \frac{1}{\sqrt{k}} \sum_{\forall e \in E_B} Z(e) \\ &= \frac{1}{\sqrt{k}} \sum_{\forall e \in E_B} \frac{1}{\sqrt{\omega_f^2 + \omega_q^2}} \left[\omega_f Z_F(e) + \omega_q Z_Q(e) \right]. \end{aligned} \tag{5.7}$$

$Z_F(e)$ is computed similar to Eq. 5.2-Eq. 5.6, containing agreement and specificity of the EC number overlap:

$$Z_F(e) = \frac{1}{\sqrt{2}} \left[Z_A(e) + Z_S(e) \right]. \tag{5.8}$$

The enzyme functional agreement score $Z_A(e)$ is derived from $P_A(e)$, the probability of obtaining by chance the number of common subclasses between the EC numbers of e_1 and e_2, the two enzyme supernodes that form hypernode e. Let $\mathcal{T}$ denote the set of all subclasses, and $\mathcal{M}$ be the overlapping subclasses. For instance, for e_1 = 1.2.3.4 and e_2 = 1.2.4.4, $\mathcal{T} = \{1, 1.2, 1.2.3, 1.2.4, 1.2.3.4, 1.2.4.4\}$, and $\mathcal{M} = \{1, 1.2\}$. These sets are then used to assess the extent of overlap between two EC numbers, analogous to Eq. 5.3:

$$P_A(e) = \frac{\binom{4}{|\mathcal{M}|}\binom{|\mathcal{T}|-4}{4-|\mathcal{M}|}}{\binom{|\mathcal{T}|}{4}} = \frac{\binom{4}{|\mathcal{M}|}}{\binom{|\mathcal{T}|}{4}} \tag{5.9}$$

$$Z_A(e) = \frac{P_A(e) - \mu_{AE}}{\sigma_{AE}}, \tag{5.10}$$

where μ_{AE} and σ_{AE} are computed from $P_A(e)$ over all possible enzyme supernode pairs.

To address the specificity of the observed $\mathcal{M}$, we also count the number of times the common EC number subclasses of two enzyme supernodes contains this $\mathcal{M}$, and compute $P_S(e)$, μ_{SE}, σ_{SE} and $Z_S(e)$, analogous to Eq. 5.5-Eq. 5.6:

$$P_S(e) = 1 - \frac{\#\text{observed } \mathcal{M} \text{ in the overlapping subclasses}}{uv}, \tag{5.11}$$

$$Z_S(e) = \frac{P_S(e) - \mu_{SE}}{\sigma_{SE}}, \tag{5.12}$$

with u and v the total numbers of enzyme supernodes in the two species.

Finally, the sequence similarity score $Z_Q(e)$ is derived from the BLAST E-value $L(e)$:

$$Q(e) = -log_{10}L(e), \qquad Z_Q(e) = \frac{Q(e) - \mu_q}{\sigma_q}, \tag{5.13}$$

where μ_q and σ_q are the mean and standard-deviation of $Q(e)$ over all possible enzyme supernode pairs.

Note that there might exist multiple EC numbers and multiple sequences in each enzyme supernode, as illustrated in Fig. 5.1. So we first compute all $Z(e)$ given all possible combinations of EC numbers and corresponding sequences in enzyme hypernode e. Since we aim to find the conserved part between pathways, the highest $Z(e)$ is taken to be the enzyme similarity score for this pair of supernodes, indicating the similarity of the most conserved part between them.

Moreover, when gaps are present, we align two enzyme supernodes in one species with one enzyme supernode in another species separately, obtaining two $Z(e)$. Again, the higher one is selected for this building block to represent the similarity of the most conserved part.

5.2.4 Pathway construction

Reaction definitions were obtained from Release 42.0 of the KEGG LIGAND composite database [45], updated on May 14, 2007. The species-specific reactions and enzyme lists were retrieved from KEGG/XML and KEGG/PATHWAY. Protein sequences were downloaded from UniProtKB/Swiss-Prot [97]. Discrepancies and missing information (e.g. gene names and EC numbers) were resolved manually. Twenty-six currency metabolites[1] are excluded from consideration during pathway construction to avoid finding large numbers of pathway shortcuts [28, 92, 120]. Note that the reactions containing these metabolites are still included in the algorithm. Currency metabolites are only excluded in aligning reactions into building blocks and assembling pathways, i.e. we do not match or connect two reactions if they only share the same currency metabolites.

Based on 881 enzymatic reactions in *S. cerevisiae* (with 1762 compound supernodes and 881 enzyme supernodes) and 1106 enzymatic reactions in *E. coli* (with 2212 compound supernodes and 1106 enzyme supernodes), 640 building blocks are constructed. These are further concatenated into pathways using a backtracking search, starting from a certain substrate. Each pathway contains four different building blocks, and is constrained so that one reaction cannot appear more than once in one species, and one compound (excluding the currency metabolites) cannot be traversed more than once in one species, e.g. a compound can not be both the substrate and product of a reaction, or be the product of more than one reaction in the pathway. Using 69% of all available building blocks, 2597 length-four pathways are assembled, starting from 245 substrates. These substrates are not restricted to external metabolites, since our pathways are not necessarily routes from uptake to secretion.

---

[1] See section 4.2.2.

	ω_c	ω_e	ω_f	ω_q	Z_0	Emphasis
Query 1	0.5	0.5	0.5	0.5	0 for all	overall
Query 2	0	1	0.5	0.5	0 for all	enzyme
Query 3	1	0	0	0	0 for all	compound
Query 4	0.5	0.5	0.5	0.5	100 for *"dg"*and *"eg"*, 0 otherwise	gap
Query 5	0	1	1	0	0 for all	enzyme function

Table 5.1: The parameter settings and biological emphases in the five queries.

5.3 Results and discussion

We conducted five queries using different settings for the parameters as described in section 5.2.3, corresponding to five different interests. Table 5.1 summarizes the parameters used.

In each query, the similarity scores of all 2597 length-four pathways found are computed using Eq. 5.1 and the highest-scoring pathway(s) of a certain substrate is referred as the *best pathway* for that substrate.

It is useful to investigate the building block types as they reflect the differences between species in terms of reactions use, which is not reflected in the scores. Therefore, we categorize the pathways w.r.t. their configurations of building blocks, in order to gain insight in the impact of the parameter settings on the resulting pathway properties. Abbreviations are used to denote the six categories: *"i-i-i-i"* indicates a pathway consists of four "identical" building blocks; *"d"* indicates that the pathway has *at least one* "direct" building block; *"em"*, *"dg"*, *"eg"* and *"ec"* are defined likewise.

Of all 2597 length-four pathways, 1198 have *"i-i-i-i"* configuration, and 1399 differ between the species, starting from 160 substrates. Among these 426 contain *"d"*, 192 *"em"*, 199 *"dg"*, 709 *"eg"* and 194 *"ec"*. For each type of configuration, Fig. 5.3a gives the percentage of best pathways found in all pathways with a particular configuration. Fig. 5.3b corrects the percentages shown in Fig. 5.3a by comparing the number of best pathways with the baseline number of best pathways, which is the maximum possible number of best pathways with that configuration. Therefore Fig. 5.3b actually presents the extent to which a query succeeds in finding a certain type of pathway when only best pathways are concerned.

5.3.1 The scoring function can address different biological questions

Using our scoring function, different parameter settings result in different best pathways, highlighting different aspects of the pathway features.

Table 5.1 and Fig. 5.3 can be used as a guide to design a query for a specific purpose. For example, Query 1 finds generally similar pathways in two species. Query 2 only considers enzyme similarity, therefore more best pathways containing *"dg"* and *"ec"* are found (Fig. 5.3b). Query 5 is a special case of Query 2, looking for conserved pathways with similar enzyme functions. Compound and enzyme sequence similarities

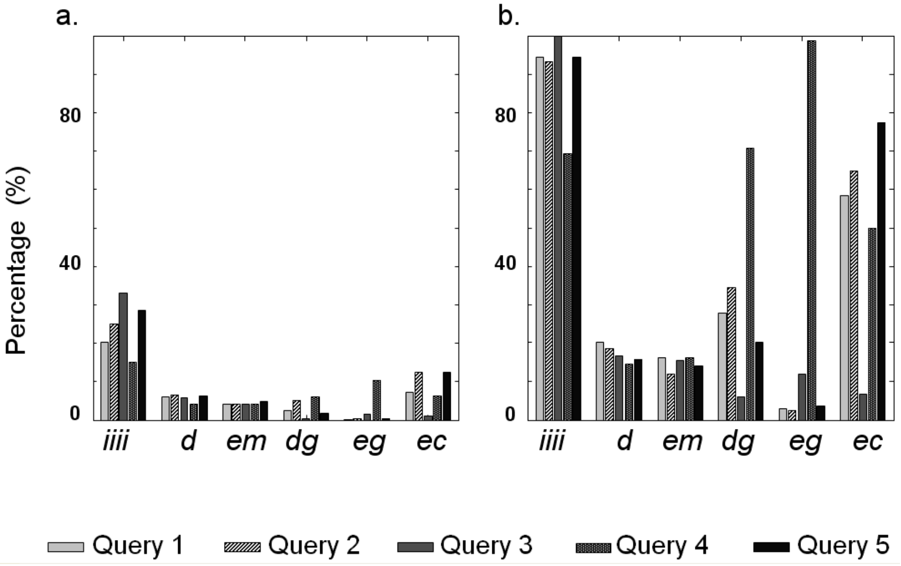

Figure 5.3: The percentages of found best pathways in the five queries, with particular pathway configurations. **a)** Percentage in all pathways with this configuration. **b)** Percentage in all possible best pathways with this configuration. See text for details.

are neglected, thus providing more possibilities for predicting the functions of unannotated genes.

Query 3, on the other hand, considers compound similarity only. If two reactions have the same compounds, they are identical reactions. So all best pathways with "*i-i-i-i*" configuration are found in Query 3 (Fig. 5.3b). Identical reactions are highly conserved in the metabolism of different species, and can be used as a measure of phylogenetic distance. Furthermore, those very specific processes containing the most unique compounds will score the highest (see Eq. 5.2). Fig. 5.4a shows an example, in which the non-currency compounds are only present in the shown pathway, which is specific to biotin metabolism.

Gaps are preferred in Query 4. Indeed, we can see a large increase in best pathways with "*dg*" and "*eg*" in Fig. 5.3b. Moreover, in-depth analysis shows that the numbers of "*dg*" and "*eg*" building blocks in the pathways have also increased four to seven times, demonstrating that the increase of found best pathways with "*dg*" and "*eg*" is not because a limited number of building blocks are used repeatedly. The results may hint at additional intriguing evolutionary phenomena: if one enzyme in species 1 is comparable to the combined functionality of two enzymes in species 2, it may be caused by gene fusion in species 1, or gene duplication in species 2 [116].

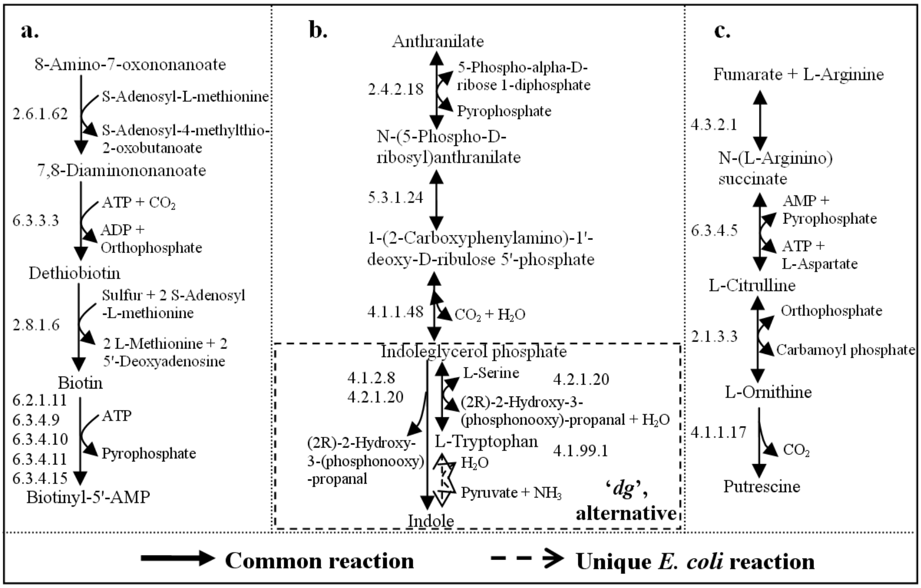

Figure 5.4: Examples of the highest-scoring pathways. **a)** One of the highest-scoring pathways in Query 3, which is involved in biotin metabolism. **b)** One of the highest-scoring pathways in Query 2, but not in Query 1 or Query 4. The last building block is a *"dg"*, which contains one unique reaction in *E. coli*, and constitutes an alternative pathway (see text). Involved KEGG maps include: phenylalanine, tyrosine and tryptophan biosynthesis; benzoxazinone biosynthesis; tryptophan metabolism; nitrogen metabolism. **c)** One of the highest-scoring pathways in Query 5, but not in Query 2. Involved KEGG maps include: urea cycle and metabolism of amino groups; alanine and aspartate metabolism; arginine and proline metabolism.

5.3.2 Comparing results of different queries can help infer additional details

It can be instructive to investigate the differences in the results between various queries. For instance, the best pathways of a certain substrate in Query 2 and not found in the best pathways of the same substrate in Query 1 have similar enzymes but use different cofactors or less specific substrates. They are well-conserved, a-specific enzymes. Many pathways containing *"dg"* are found in Query 2 for this reason, as we can see from Fig. 5.3b. Fig. 5.4b shows an example, which is found in Query 2 due to its high enzyme similarity, but not in Query 1 or Query 4 for the same substrate due to its low compound similarity. In another example (not shown), a best pathway in Query 2 producing pyruvate is filtered out in Query 1 because pyruvate is less specific, as it is present in 147 reactions [18].

In addition, the best pathways of a certain substrate in Query 5 and not found in the best pathways of the same substrate in Query 2 have similar enzyme functions but dissimilar enzyme sequences. These enzymes might be non-homologous but evolved

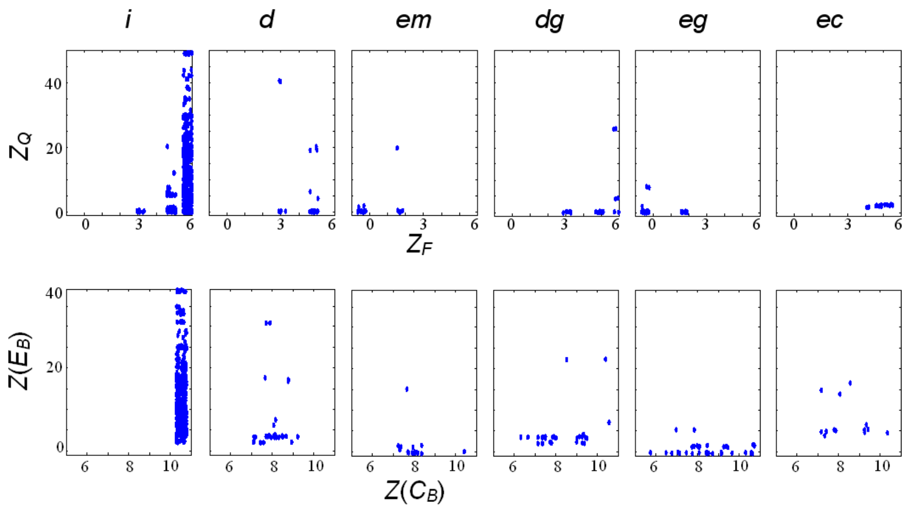

Figure 5.5: The distributions of the four component scores for each type of building block. Z_F and Z_Q are computed as in Eq. 5.8 and Eq. 5.13. $Z(C_B)$, $Z(E_B)$ are computed as in Eq. 5.2 and Eq. 5.7 with the parameter settings of Query 1 (see Table 5.1).

into the same function, or the functions have been maintained although their sequences have been changed. An example is given in Fig. 5.4c. The enzymes in the fourth building block, spe1 from *S. cerevisiae* and speC, speF from *E. coli*, have very dissimilar sequences (E-value > 100). Although spe1, speC and speF are non-homologous, lysA (EC: 4.1.1.20) in *E. coli* has a sequence similar to that of spe1 (E-value $= 2.5 \times 10^{-7}$). According to Sandmeier *et al.* [124], speC and speF belong to group III decarboxylases, and spe1 and lysA belong to group IV decarboxylases. Although the homology among the enzymes within each group is established, no evidence has been obtained that the sequences of these two groups are related. Therefore, they seem to have different evolutionary origin. This result demonstrates that enzyme function and sequence do not always correlate with each other. In addition, more "*ec*" are found in Query 5 (see Fig. 5.3b) exactly because on average "*ec*" has high enzyme functional similarity but low sequence similarity, as shown in Fig. 5.5.

5.3.3 Combining the component scores makes sense

Fig. 5.5 presents the component scores of each type of building block and shows that the various information sources are not correlated (see also [142]), making it worthwhile to combine them. In addition, Fig. 5.5 reveals the diverse similarity fingerprint of each type of building block, which calls for further research. For example, the variance of the sequence similarity score in "*i*" is large, which might arise because of different specificity, horizontal gene transfer, gene fusions, or the fact that only subunits of the enzymes are the same. As to "*ec*", their sequences are very dissimilar in spite of their similar functions. Possible reasons could be that the enzymes have different substrate specificities, or that intermediate substrates are very different. They could also

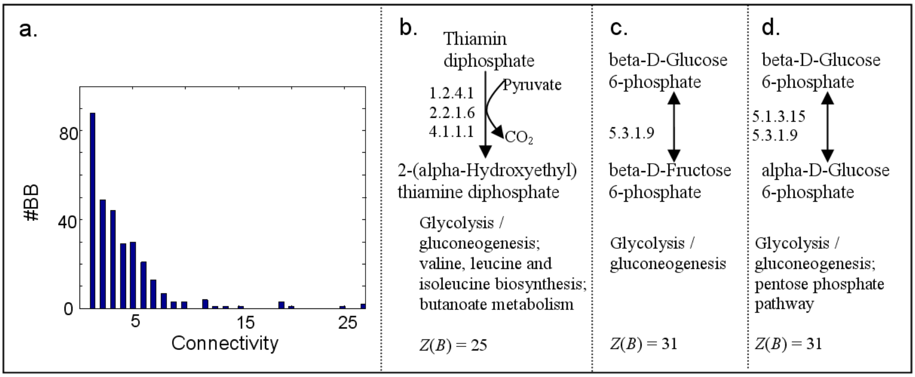

Figure 5.6: The building block connectivity. **a)** Histogram of the number of best pathways in which a building block is involved in Query 1. **b) - d)** Three building blocks which are involved in 27, 27 and 25 best pathways in Query 1, respectively. Scores and involved KEGG maps are given underneath the building blocks.

have been isoenzymes in parallel pathways, having become specialized to one species during evolution.

5.3.4 The conserved part of two aligned networks is scale-free

We inspected the connectivity of each building block in Query 1, i.e. the number of best pathways in which a building block is involved. Fig. 5.6a shows that building block connectivity follows a power-law distribution. It has already been pointed out that metabolic networks as a whole are scale-free networks [66]; but our finding provides evidence from a new perspective, indicating that the conserved part of aligned networks, composed of the building blocks in the best pathways, is also scale-free. Fig. 5.6b-d shows the three building blocks with the highest connectivity to be involved in primary metabolism glycolysis/gluconeogenesis, which is known to be highly conserved and plays a role in many processes.

5.3.5 Short pathways lead to interpretable results

Our methodology has no inherent limit on the pathway lengths. That is, it can construct and score pathways consisting of any number of building blocks. To find longer pathways, one can simply extend the pathway length in the search step. Actually, we conducted experiments without a length limit, which resulted in aligned pathways up to a length of 42 building blocks. Another solution would be to assemble the current length-four short pathways into longer pathways.

However, not all pathway lengths give meaningful results. When the length becomes too short, the method starts to compare individual reactions and loses the power of

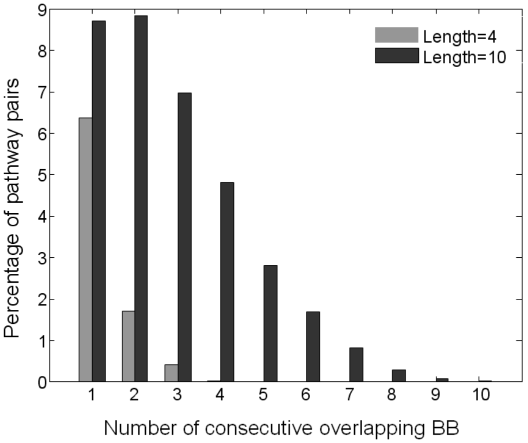

Figure 5.7: The impact of pathway length on the resulting overlap. A frequency graph of the number of consecutive overlapping building blocks in all pairs of pathways of the same length found in Query 1. When pathway length is increased, the overlap between resulting pathways increases significantly, hampering interpretation.

metabolic functional context, as stated in section 5.1. As a result, some isolated reactions are also included in the results. For example, 31% of building blocks (i.e. length-one pathways) contain isolated reactions, which are not included in any length-four pathway.

When the pathway length becomes too large, the method produces many highly overlapping results. For example, when running M-PAS with the pathway length set to ten, the number of found pathways increases to 15939 (as compared to the 2597 found pathways when this length is set to four). However, Fig. 5.7 shows that the average overlap between any two pathways also increases significantly. This makes it more difficult to interpret the results. Moreover, longer pathway lengths stress pathway conservation more, and will inevitably miss some interesting short pathways. For example, 128 building blocks (20%) which are present in the results of length-four are not found in the set of length-ten pathways. Therefore, although limiting the pathway length to four might not be the optimal choice, it is within a reasonable range which produces meaningful results.

5.3.6 M-PAS reveals pathway diversity and alternatives

As mentioned above, we found that 54% of the length-four pathways are not "*i-i-i-i*", which occur in 65% of the substrates. Interestingly, 17 start substrates do not have any "*i-i-i-i*" pathways, which means the length-four pathways starting with these substrates *always* differ in these two species. When only best pathways are concerned, we found 16% of these are not "*i-i-i-i*", starting from 13% of the substrates. Fig. 5.8 displays two best pathways in Query 1, which contain unique reactions in both species.

These pathways are highly conserved, yet exhibit differences between the two species.

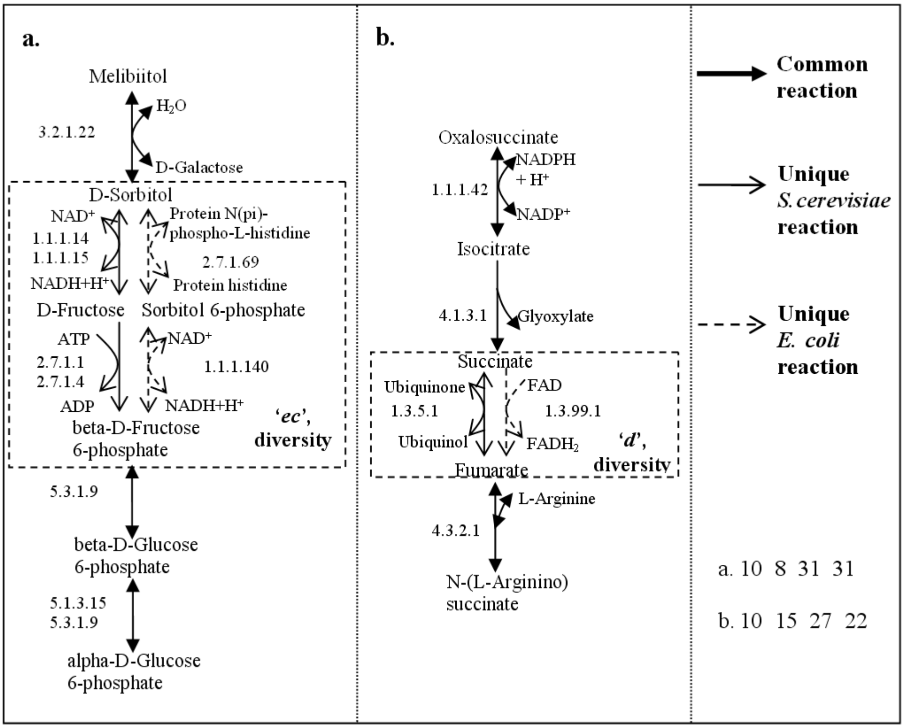

Figure 5.8: Two examples of non-"*i-i-i-i*" best pathways in Query 1. The non-identical building blocks are highlighted, which exhibit diversities. Scores of all building blocks are shown at the bottom right. The involved KEGG maps are: **a)** galactose metabolism; fructose and mannose metabolism; glycolysis / gluconeogenesis; pentose phosphate pathway. **b)** citrate cycle (TCA cycle); glyoxylate and dicarboxylate metabolism; urea cycle and metabolism of amino groups; alanine and aspartate metabolism; arginine and proline metabolism; butanoate metabolism (only for *E. coli*); reductive carboxylate cycle (CO_2 fixation) (only for *E. coli*).

Note that M-PAS goes beyond simple reaction comparison and always places these differences in metabolic functional context. In this way, our method sheds light on variations between species in the use of non-identical but similar reactions in pathways, revealing between-species diversity and within-species alternatives. When both species have their own unique reactions to transform a particular substrate into a particular product, we call this *diversity*. If only one of the species has a unique reaction, which performs the same transformation as another common reaction does in both species, then this unique transformation forms part of an *alternative* pathway. Fig. 4.2 gives a schematic explanation of these two terms.

Recall the constraint in section 5.2.2 which enforces uniqueness in constructing a non-identical building block. Consequently, these non-identical building blocks contain unique reactions in either one or both species, introducing diversity or alternatives in the assembled pathways. In other words, all resulting pathways which do not have an

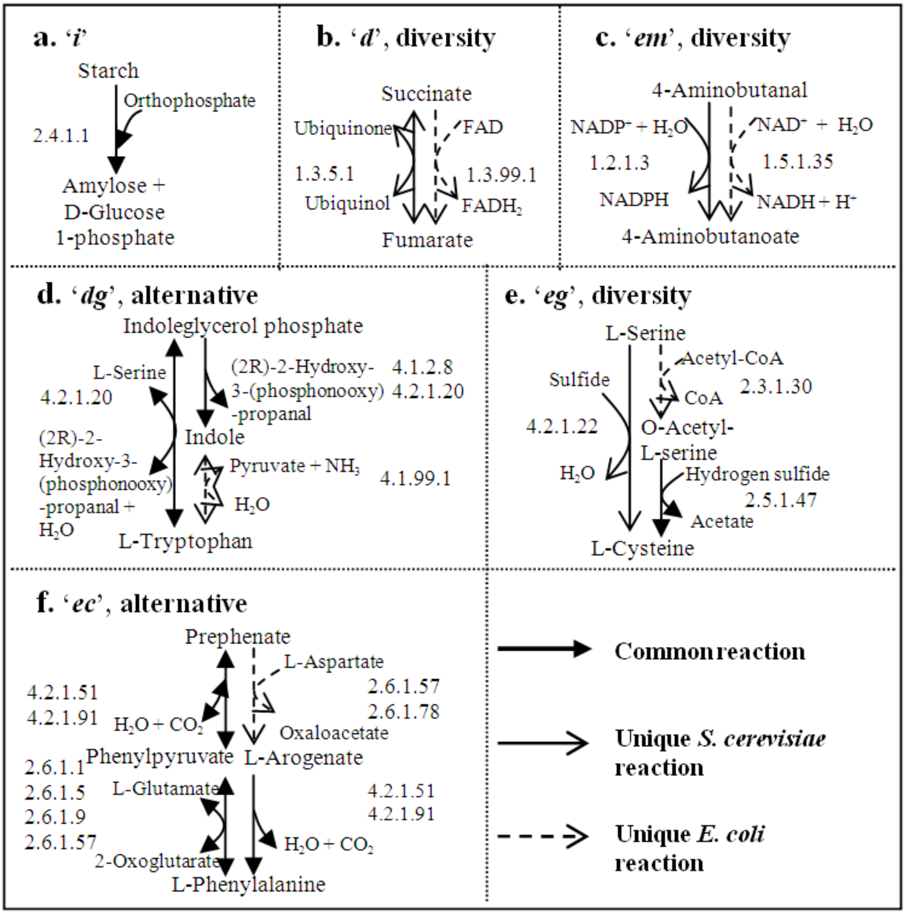

Figure 5.9: High-scoring building blocks in Query 1. **a)** One of the highest-scoring "identical" building blocks ($Z(B) = 34$). **b)** One of the highest-scoring "direct" building blocks ($Z(B) = 27$). **c)** The highest-scoring "enzyme mismatch" building block ($Z(B) = 16$). **d)** The highest-scoring "direct-gap" building block ($Z(B) = 23$). **e)** The highest-scoring "enzyme mismatch-gap" building block ($Z(B) = 9$). **f)** The highest-scoring "enzyme crossover match" building block ($Z(B) = 18$).

"*i-i-i-i*" configuration contain diversity or alternatives. For example, the fourth building block in Fig. 5.4b contains a reaction unique to *E. coli*, constituting a unique alternative pathway. On the other hand, the second building block in Fig. 5.8a and the third building block in Fig. 5.8b contain unique reactions in both species, therefore they show diversity in the pathways. More examples are given in Fig. 5.9, which displays the most similar building blocks of each type in Query 1.

These results demonstrate the value of including non-identical building blocks, as otherwise these strongly conserved pathways would have been overlooked. In particular, building blocks with gaps or crossovers would be hard to detect manually, e.g. Fig. 5.8a and Figs. 5.9d-f. Take Fig. 5.9d as an example. By comparing reactions in two species, normally we can only find a reversible reaction present in both species which catalyzes indoleglycerol phosphate into L-tryptophan. However, considering gaps allows us to find two consecutive reactions in one of the species which perform the same transformation in two steps. In the end, our algorithm found a unique alternative pathway in *E. coli* which transforms indoleglycerol phosphate to indole first by an irreversible reaction, followed by a unique reaction transforming indole to L-tryptophan.

5.3.7 New links between different parts of metabolism are found

Our method is global, starting from constructing building blocks to the assembly of pathways. Therefore, the resulting pathways have a reasonable coverage of the network, and explicitly include links between different parts of metabolism, which are displayed in 202 pathway maps of metabolism in KEGG [45]. For example, Fig. 5.8 shows four to seven such maps are linked together in each aligned pathway (see caption).

Since our alignment method operates on individual reactions, independent of the existing pathways as given in current databases, we not only reconstruct known pathways (as presented by KEGG, e.g. Figs. 5.4, 5.8a, 5.10b-c), but also discover new pathway possibilities with the component reactions annotated in different maps and not linked with each other in the original database, e.g. Figs. 5.8b and 5.10a. These pathways will not be found if we only look at the pathways shown in the maps and the links between maps.

Moreover, M-PAS not only links different parts of metabolism within one species, but also associates diverse parts in two species with each other, offering potential interesting targets for bioengineering. For instance, in Fig. 5.9e, the unique reaction of *S. cerevisiae* is found in glycine, serine and threonine metabolism, while the unique reaction of *E. coli* is found in cysteine metabolism. Therefore it will not be found if we only look at one map or one species at a time.

5.3.8 Primary metabolism is highly conserved

Three pathways with the highest scores in Query 1 are shown in Fig. 5.10. They represent the most conserved part of the metabolic network in the two species and are therefore expected to be important. Not surprisingly, the three pathways are all involved in primary metabolism. Moreover, they all have "*i-i-i-i*" configuration, meaning all reactions in the pathways are conserved across species. Clement *et al.* [12] also pointed out that "vital biological processes in a group of related species should be conserved and expressed by a significant number of reactions in all the organisms of the group".

We can also observe this in Fig. 5.9, where the involved parts of metabolism in the highest-scoring building blocks are rather central processes, e.g. starch and sucrose metabolism, citrate cycle (TCA cycle), CO_2 fixation and other important amino acid metabolisms.

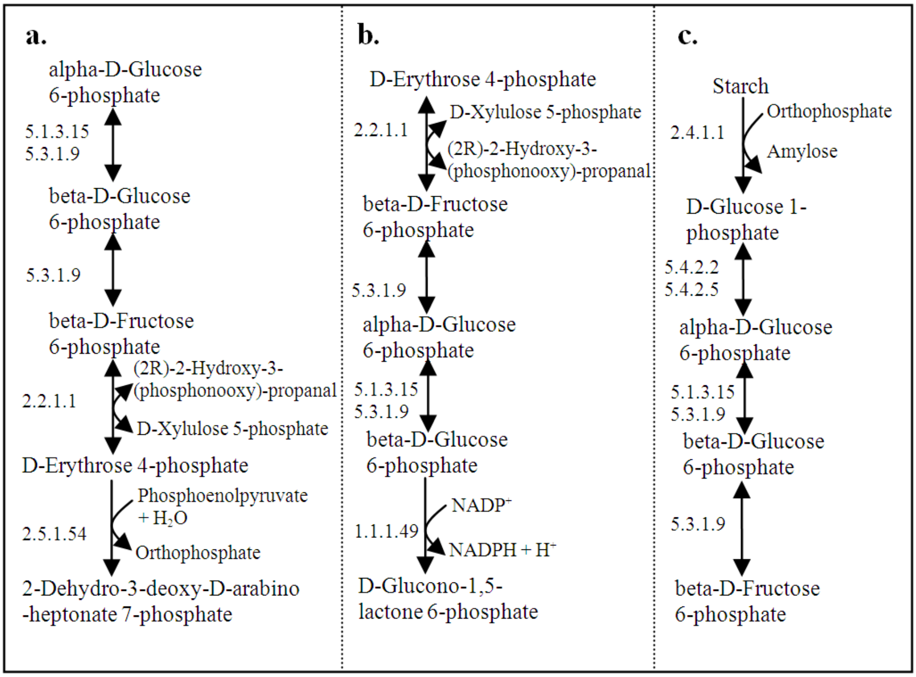
a.
alpha-D-Glucose 6-phosphate
5.1.3.15
5.3.1.9
beta-D-Glucose 6-phosphate
5.3.1.9
beta-D-Fructose 6-phosphate
(2R)-2-Hydroxy-3-(phosphonooxy)-propanal
2.2.1.1
D-Xylulose 5-phosphate
D-Erythrose 4-phosphate
Phosphoenolpyruvate + H2O
2.5.1.54
Orthophosphate
2-Dehydro-3-deoxy-D-arabino-heptonate 7-phosphate
b.
D-Erythrose 4-phosphate
D-Xylulose 5-phosphate
2.2.1.1
(2R)-2-Hydroxy-3-(phosphonooxy)-propanal
beta-D-Fructose 6-phosphate
5.3.1.9
alpha-D-Glucose 6-phosphate
5.1.3.15
5.3.1.9
beta-D-Glucose 6-phosphate
NADP+
1.1.1.49
NADPH + H+
D-Glucono-1,5-lactone 6-phosphate
c.
Starch
Orthophosphate
2.4.1.1
Amylose
D-Glucose 1-phosphate
5.4.2.2
5.4.2.5
alpha-D-Glucose 6-phosphate
5.1.3.15
5.3.1.9
beta-D-Glucose 6-phosphate
5.3.1.9
beta-D-Fructose 6-phosphate

Figure 5.10: Three pathways with the highest scores in Query 1. The solid-headed arrow indicates the reactions exist in both species, constituting an "*i-i-i-i*" pathway. **a)** $Z(P) = 41$. Involved KEGG maps include: glycolysis/gluconeogenesis; pentose phosphate pathway; starch and sucrose metabolism; phenylalanine, tyrosine and tryptophan biosynthesis. **b)** $Z(P) = 40$. Involved KEGG maps include: pentose phosphate pathway; glycolysis/gluconeogenesis; starch and sucrose metabolism; glutathione metabolism. **c)** $Z(P) = 39$. Involved KEGG maps include: starch and sucrose metabolism; glycolysis/gluconeogenesis; galactose metabolism; streptomycin biosynthesis; pentose phosphate pathway.

5.4 Conclusions

In this work, we extend our former alignment framework and propose a novel scoring method to identify conserved metabolic pathways and quantify the level of conservation in an efficient and comprehensive manner. Based on the six types of building blocks, a systematic search is conducted in the network. We find and rank conserved pathways given certain substrates, and shed light on the variations between species within a metabolic functional context. This is not possible by simple comparison of reaction lists or enzyme lists.

Our method combines individual reactions, so that we can find conserved pathways that are not represented in conventional databases. Since the alignment and search are conducted in the whole network, M-PAS unites reactions in different KEGG maps, revealing links and relating reactions with common upstream substrates and downstream products which might be elusive if we only look at subsets of the network.

Our similarity measure combines uncorrelated information sources, including similarities of substrate sets, product sets, enzyme functions, enzyme sequences and alignment configurations. The function has a generic form and is capable of measuring pathway similarity given different biological emphases. Due to its hierarchical integration structure, it is readily extensible to include other relevant similarity measures if available (e.g. enzyme affinities), or to modify a component score (e.g. using compound molecular similarity scores). Moreover, the proposed function is plausible since parts of primary metabolism, which are known to be well conserved, are found to be abundant in our top-scoring pathways and building blocks.

M-PAS reveals highly conserved pathways containing diversity or alternatives, which yields important information for biology and life sciences. First, the results give insight into the evolutionary differences between species. For instance, the two species apparently diverged to process 17 substrates differently, so that no "*i-i-i-i*" pathways are found starting from them. This divergence calls for special treatment of these substrates per species in analysis and applications. Second, the diversity and alternatives in conserved pathways also provide additional ways to construct metabolic networks for currently unannotated species. Third, our analysis lists potential candidate enzymes for bioengineering, i.e. certain natural enzymes can be removed, introduced, or changed so that we can select a favorable pathway to enforce production of a metabolite of interest, or block pathways leading to certain unfavorable products. In particular, alternative pathways have to be considered in drug design, because blocking central enzymes might not be effective when alternative pathways provide other routes, and cause drug resistance in the pathogen population [19].

M-PAS is currently constrained to finding linear pathways which are strictly similar. Although further processing these linear pathways, e.g. combining them, could reconstruct some tree-like subnets and cycles, not all network structures can be captured. M-PAS could be extended to construct and score more complex pathway topologies that capture more variation. First, to capture more variation, one may extend the building block definition to include larger differences, e.g. a "*dg*" with two gaps, or to allow compound mismatch. But care needs to be taken to keep the computational load acceptable and to avoid linking unrelated pathways. Alternatively, one may reduce the pathway length, e.g. to assemble two building blocks into a pathway to capture diverse pathways with short overlaps. However, as discussed earlier, when the pathway length becomes too short, the method starts to compare individual reactions. To find more complex pathway topologies, a more complex search algorithm is required. An alternative would be to expand our building block definition to incorporate more types of network motifs. But again, the computational load will increase significantly.

The complementary reaction information of multiple well-studied model species provides more confidence and more possibilities to transfer this information to a new species. Although M-PAS currently only performs pairwise alignment on two species, we expect even more informative results when it is applied on multiple species, and larger differences will be found as the phylogenetic distance increases. Finally, by relating different sets of enzymes in different species to a common metabolic function, this work provides an infrastructure in which regulatory factors can be incorporated, and functional hypotheses can be generated.

6

RM-PAS: ALIGNING REGULATORY-METABOLIC PATHWAYS

Integrating different types of biological networks and aligning networks across species are two useful but challenging comparative methods in systems biology nowadays. By combining these in one framework, we can expect to generate more reliable information and hypotheses. In this study, we systematically integrate the transcriptional regulation network of enzyme-coding genes and the corresponding metabolic network, and align these integrated networks between two species. By applying a scoring function to measure the alignment similarity, our method can be used to identify conserved elements (allowing for small variations) of evolution at both the regulatory and metabolic level, to reveal the interrelation and divergence between species and to use information at one level to predict missing information at the other level.

This chapter was published in *Proceedings of the* 8^{th} *Annual International Conference on Computational Systems Bioinformatics*, Stanford, USA, 2009 [84].

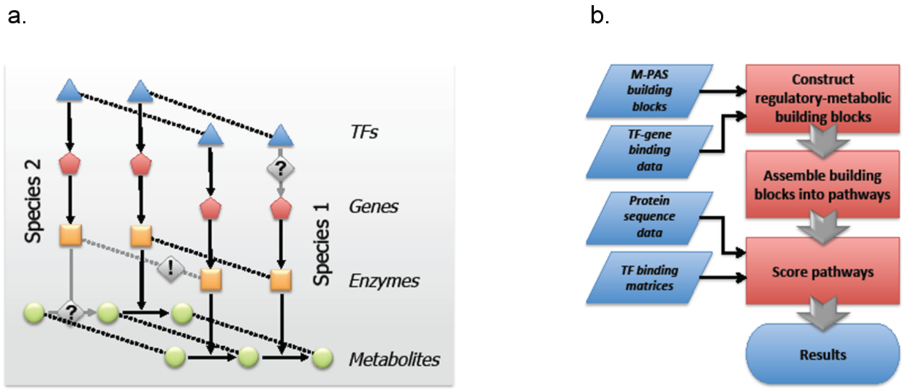

Figure 6.1: Method overview. **a)** The goal of our method is to align metabolic pathways and their regulation between two species, using suitably defined similarity measures between compounds, enzymes and transcription factors (illustrated by the dotted lines), in order to find conserved elements and learn about differences between species (illustrated by the exclamation mark). From missing links in an otherwise conserved context, we can infer missing reactions or regulation within species (illustrated by the question marks). **b)** The RM-PAS flowchart.

6.1 Introduction

Most metabolic reactions in cells are catalyzed by enzymes, and the genes which code for these enzymes are regulated by transcription factors (TFs). That is, TFs can bind to the promoter sequence of genes and subsequently activate or repress the transcription of these genes. This information flow from the regulatory level to the metabolic level is illustrated in Fig. 6.1a. At each level, these interactions form a network, i.e. a transcriptional regulatory network and a metabolic network, respectively.

Comparing networks between species at each level individually can help to filter noise, and produce insights into the principles governing evolution. For example, Gasch *et al.* [37] found that many of the known cis-regulatory systems in *Saccharomyces cerevisiae* (yeast) have been conserved in 13 ancient fungi species. Tanay *et al.* [139] studied the promoter evolution of co-regulated genes in 17 yeast species, and suggested an intermediate redundant regulatory program underling the evolvability and increased redundancy of transcriptional regulation in higher organisms. Alkema *et al.* [2] improved the prediction of co-regulated genes based on the conservation of protein sequences and regulatory mechanisms. At the metabolic level, Jeong *et al.* [66] and Ravasz *et al.* [120] studied the global topological properties of the metabolic networks in 43 species. Heymans *et al.* [53] derived phylogenetic trees based on metabolic pathway comparison.

Comparing networks at different levels simultaneously can be even more informative. Since different types of network present different perspectives on the biological system, integrating them may offer a more comprehensive picture. Particularly when elements are conserved at multiple levels, we can be more confident about the reliability of the observed conservation. This allows us to make predictions, using information at one level to infer information at another level, or using information of one species to infer information for another species.

Although integrating different types of network within one species has received quite some attention [17, 60, 114, 130], little advances have been made on the alignment of regulatory and metabolic networks across species. Here we present a method that searches for network elements that are conserved in evolution at both the regulatory and metabolic level, and measures the extent of this conservation. A schematic overview of our goal is given in Fig. 6.1a.

Previously we developed M-PAS [85], a framework for metabolic pathway alignment and scoring based on the notion of building blocks (see Fig. 5.2), to align the metabolic networks of *Saccharomyces cerevisiae* and *Escherichia coli*. In the current work, we integrate TF-gene interaction and TF binding site (TFBS) information into M-PAS, and form a more comprehensive method, RM-PAS. We applied RM-PAS to *S. cerevisiae* and *E. coli*, two of the best-annotated model organisms, with relatively much TF binding and TFBS data available. Since these species are not closely related, many differences are expected, and the resulting conservation is expected to be quite informative.

6.2 Method

The building block method used in M-PAS has shown to be an appropriate approach to align metabolic pathways [85, 86]. First, it is able to explore topological arrangement possibilities of reactions both between species (by building block construction) and within species (by pathway assembly). Second, by defining building blocks, we can focus on conserved pathways while allowing small variations. Third, the method is adaptable and can easily be extended to include more information.

Here, we extend the building block construction and the scoring function to include transcriptional regulation information. That is, for every enzyme in a reaction, we add the transcription factors that regulate the enzyme-coding genes. In the end, we consider the building blocks be the *conserved elements* that we are interested in. The flowchart is given in Fig. 6.1b and will be explained in the remaining of this section. Note that given curated databases (see section 6.3) and user-defined parameters as input, each step in the flowchart is automated.

6.2.1 Regulatory-metabolic building blocks

We add transcriptional regulation to the metabolic building blocks in M-PAS, to construct regulatory-metabolic (RM) building blocks. That is, we add a link between a transcription factor and the enzyme in the reaction. This is only done when there is experimental evidence showing that the transcription factor indeed regulates the gene coding for the enzyme.

Like in the metabolic building block approach, we also categorize the RM building blocks with different TF regulation scenarios in the two species, as well as different TF similarity scenarios, i.e. (1) whether the TFs which bind to the enzyme-coding genes are similar (”direct TF”) or dissimilar (”mismatch TF”), and (2) whether there exist

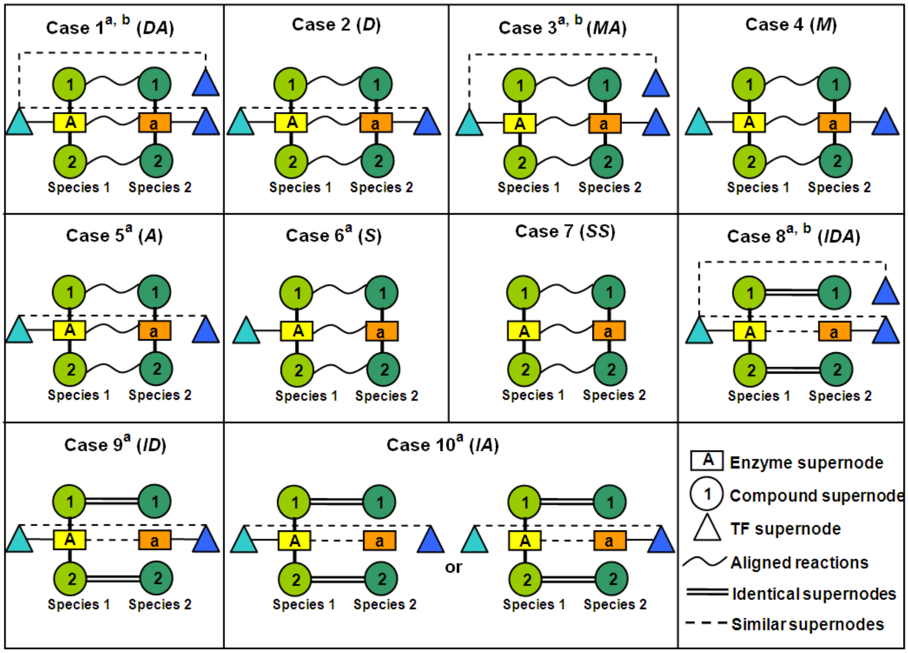

Figure 6.2: Illustration of the ten cases of RM building blocks. ***D***: direct TF. ***A***: alternative TF. ***M***: mismatch TF. ***S***: absent TF. ***I***: missing reaction. [a] The labels “species 1” and “species 2” can be exchanged. [b] The alternative TF can be present in one or both species. Aligned reactions denote any of the six types of metabolic building blocks (see Fig. 5.2). A *TF supernode* is the set of TFs which bind to the enzyme-coding genes in a reaction. Two TF supernodes are considered similar if their TFBS are more similar than average, i.e. for “direct TF”: $Z_{TB}(B) > 0$ (Eq. 6.6), and for “alternative TF”: $Z_{TU}(B) > 0$ (Eq. 6.8). In cases 8-10, the two enzyme supernodes have the same EC number.

additional TFs (“alternative TF”) in one species which are similar to the TFs in another species, but are not found to bind to the genes in that reaction. When one species has neither bound TFs nor alternative TFs, we call the RM building block has “absent TF” in that species. The seven possible cases where TFs are added to the metabolic building blocks are shown in Fig. 6.2, cases 1-7.

In addition to the reactions present in the database, we also look for possible reactions which are currently missing in one of the species (“missing”). In this scenario, one reaction is present in only one species, but the other species does contain the reaction’s compounds and enzymes with identical function in terms of EC number. An RM building block is then constructed when there is evidence from the transcriptional regulation control indicating that the missing reaction might be present. That is, when there exist “direct” and/or “alternative” TFs, we hypothesize the reaction might exist in both species. These three cases are shown in Fig. 6.2, cases 8-10.

6.2.2 Pathway assembly

After building blocks are constructed, they are concatenated into pathways, if the product of the upstream building block is the substrate of the immediate downstream building block. Since we are interested in small differences (as illustrated in Fig. 5.2 and Fig. 6.2), instead of generating a few highly conserved longer pathways, we generate a ranked list of short pathways with the same length. Because the amount of overlap between pathways increases substantially when pathway length increases, we limit each pathway to contain four building blocks.

To implement an exhaustive search for all length-four pathways, we start a backtracking search from each substrate. During the search, all building blocks in a pathway should be different, and one reaction cannot appear more than once in one species. Note that twenty-six currency metabolites[1] are excluded from consideration during pathway assembly to avoid finding large numbers of pathway shortcuts.

6.2.3 Scoring function

We rank the aligned pathways according to the extent of conservation, in order to prioritize the interesting pathways for further investigation. The M-PAS scoring function [85] integrates multiple similarity scores of all reaction components. It has a generic form and is capable of measuring pathway similarity given different biological emphases. This allows user to specifically look for certain characteristic differences between species in otherwise highly conserved pathways: by setting the appropriate parameters, differences will be allowed between enzymes, compounds and/or TFs. Due to its hierarchical integration structure, it is readily extensible to include other relevant similarity measures. In this study, the M-PAS scoring function [85] is adapted such that transcriptional regulation similarities are included.

Total score

Our goal is to reflect all aspects of an aligned pathway in the total similarity score. These include similarities at the regulatory level and the metabolic level, i.e. similarities between transcription factors, substrate sets, product sets, enzyme functions, enzyme sequences and alignment topology, respectively.

To account for their diverse distributions of similarities, we first compute similarity scores independently for each aspect, and then convert the raw scores into z-scores before integration. The integration of multiple z-scores is done hierarchically using Liptak-Stouffer's method [52]. In this way, we obtain a decomposable score for a pathway:

$$\begin{aligned} Z(P) &= \frac{1}{\sqrt{N}} \sum_{\forall B \in P} Z(B) \qquad (6.1) \\ &= \frac{1}{\sqrt{N}} \sum_{\forall B \in P} \frac{1}{\sqrt{3}} \left[Z_0(B) + Z_R(B) + Z_T(B) \right], \end{aligned}$$

---

[1] See section 4.2.2.

where $Z(P)$ denotes the total z-score of an aligned pathway P, which contains N building blocks B. $Z_0(B)$ is the user-specified bias for the building block alignment type. For example, if the user is interested in building blocks with gaps, then the building blocks with gaps, i.e. *"direct-gap"* and *"enzyme mismatch-gap"* in Figs. 5.2d-e, can be assigned a large positive bias. $Z_R(B)$ and $Z_T(B)$ denote the reaction and transcription factor similarity z-scores in B. $Z_R(B)$ is discussed in detail in Chapter 5 and [85], and will be briefly described below. Here we mainly focus on the TF similarity score.

Reaction score

The reaction similarity score $Z_R(B)$ is a weighted sum of its compound score $Z(C_B)$ and enzyme score $Z(E_B)$:

$$Z_R(B) = \frac{1}{\sqrt{\omega_c^2 + \omega_e^2}} \left[\omega_c Z(C_B) + \omega_e Z(E_B)\right]. \tag{6.2}$$

Compound weight ω_c and enzyme weight ω_e can be used to assign different relative importance to compound similarity and enzyme similarity. The compound score $Z(C_B)$ combines the similarities between the substrate sets and between the product sets in a building block B, considering the amount and specificity of the overlapping compounds. The enzyme score $Z(E_B)$ is a weighted sum of a functional similarity score (with weight ω_f) and a sequence similarity score (with weight ω_q).

Transcription factor score

We measure TF similarity to see whether regulation is conserved in the two species, and whether we can find possible alternative TFs. Therefore, the TF score contains two parts: (1) the similarity between the bound TFs in two species (Z_{TB}), and (2) the similarity between the bound TFs in one species and TFs that are not found to bind in the other species (Z_{TU}). Weights are given to these two parts for finding different cases in Fig. 6.2. Thus the TF score can be written as an integrated z-score:

$$Z_T(B) = \frac{1}{\sqrt{\omega_{tb}^2 + \omega_{tu}^2}} \left[\omega_{tb} Z_{TB}(B) + \omega_{tu} Z_{TU}(B)\right]. \tag{6.3}$$

First, we need to compute the raw similarity scores between TFs. A TF is characterized by its corresponding transcription factor binding site (TFBS), which can be quantitatively described by position weight matrices (PWM) or position frequency matrices (PFM) [150]. We take the standard approach of comparing PFM profiles [74, 127] to measure the similarities between different TFs in an RM building block. More specifically, we applied MatCompare [127] to calculate the Kullback-Leibler divergence [133] between the PFM matrices. This measures the information divergence between the matrix entries. If matrices m and m' have w columns, indicating the length of the TFBS sequence, the divergence between them is:

$$D(m, m') = \sum_{i=1}^{w} \sum_{j \in \{A,C,G,T\}} (m_{ij} - m'_{ij}) \log(m_{ij}/m'_{ij}). \tag{6.4}$$

If one of the two matrices has fewer columns, that matrix is compared to all possible starting columns in the other matrix to find the best match.

For a building block B, there might be multiple TFs, each of which might have multiple PFM matrices. Let M_{B1} and M_{B2} denote the complete set of PFM matrices of all bound TFs involved in B in the two species, respectively. Then the raw TF similarity between bound TFs is the best match in all pairs of bound TF PFM matrices:

$$S_{TB}(B) = \max_{m \in M_{B1}, m' \in M_{B2}} -D(m, m'). \tag{6.5}$$

This similarity is further transformed into a z-score:

$$Z_{TB}(B) = \frac{S_{TB}(B) - \mu_{TB}}{\sigma_{TB}}, \tag{6.6}$$

where μ_{TB} and σ_{TB} are the average and standard-deviation of S_{TB} over all possible permuted pairs of M_{B1} and M_{B2}.

Similarly, we compute the raw similarity score between bound TFs in one species and the alternative TFs in the other species, which is the best match in all pairs between bound TF PFM matrices in one species and the alternative TF PFM matrices in the other species:

$$S_{TU}(B) = \max\{\max_{m \in M_{B1}, m' \notin M_{B2}} -D(m, m'), \max_{m \notin M_{B1}, m' \in M_{B2}} -D(m, m')\}, \tag{6.7}$$

$$Z_{TU}(B) = \frac{S_{TU}(B) - \mu_{TU}}{\sigma_{TU}}, \tag{6.8}$$

where μ_{TU} and σ_{TU} are the average and standard-deviation of S_{TU} over all possible permuted pairs of M_{B1} and M_{B2}.

6.3 Data

Reaction definitions were obtained from Release 42.0 of the KEGG LIGAND composite database [46], updated on Aug. 18, 2008. The species-specific reactions and enzyme lists were retrieved from KEGG/XML and KEGG/PATHWAY. Protein sequences were downloaded from UniProtKB/SwissProt [97] Release 56.0, updated on July 22, 2008.

For *S. cerevisiae*, the experimentally verified TF-gene binding data is collected from TRANSFAC [153] Release 11.4 and Yeastract [141] version 2008515. The PFM matrices are obtained from TRANSFAC, Yeastract, SwissRegulon [108], IMD [11], and ooTFD [42].

For *E. coli*, the experimentally verified TF-gene binding data is collected from EcoCyc [72] Release 11.6 and RegulonDB [35] Release 6.0. The TFBS matrices are obtained from RegulonDB and SwissRegulon.

6.4 Experiments and results

Based on 957 enzymatic reactions in yeast and 1175 enzymatic reactions in *E. coli*, we constructed 697 RM building blocks, including 5 of cases 8-10 in Fig. 6.2. They are assembled into 8397 length-four pathways, starting from 259 substrates.

Query	ω_c	ω_e	ω_f	ω_q	ω_{tb}	ω_{tu}	Z_0
1	1	1	1	1	1	0	0 for all
2	1	1	1	1	0	1	0 for all
3a	1	1	1	1	-1	1	100 for non-"*i*"
3b	1	1	1	1	-1	1	100 for "*i*"

Query	**Target**	**Case**
1	Full conservation	1,2,8,9
2	Missing TF-gene bindings	1,3,5,8,10
3a	Differences between two levels	3
3b	Differences between two levels	3

Table 6.1: The parameter settings in the three queries. "*i*" refers to the identical metabolic building block type in Fig. 5.2. The cases refer to those in Fig. 6.2, illustrating the scenarios for each query.

Here we demonstrate our method using three example queries, to find fully conserved pathways, missing TF-gene bindings, and differences between the regulatory and metabolic level. Each query uses a different parameter setting, including the building block type bias Z_0, four reaction score weights (i.e. ω_c, ω_e, ω_f and ω_q), and two TF score weights (i.e. ω_{tb} and ω_{tu}). In each query, the similarity scores of all pathways found are computed using Eq. 6.1, and the highest-scoring pathway(s) of a certain substrate is referred as the *best pathway* for that substrate.

Table 6.1 lists the parameter settings in the queries. The motivations for, and results of the queries are discussed in the following.

6.4.1 Identifying conserved regulatory-metabolic network elements

In Query 1, all aspects of known information at both the regulatory and the metabolic level are considered. Therefore, the resulting pathways represent elements fully conserved at both levels. Fig. 6.3a gives an example, which is involved in the citrate cycle (TCA cycle) and the biosynthesis of several essential amino acids, i.e. valine, leucine and isoleucine.

The addition of TF similarity helps to refine the results of Query 1 in M-PAS, which only uses reaction similarity. Consequently, the ranks of found length-four pathways in RM-PAS might be different than those in M-PAS, revealing that regulatory mechanisms are not uniformly conserved in metabolic pathways.

For the 2427 pathways common in the results of RM-PAS and M-PAS, we calculated the rank of each pathway among the group of pathways which share the same starting substrate, using both scoring methods. This rank was then normalized by dividing by the size of the group to obtain a normalized rank in the range of [0,1], i.e. the most conserved pathway in a group ranks 1. In the end, 52% of pathways have normalized

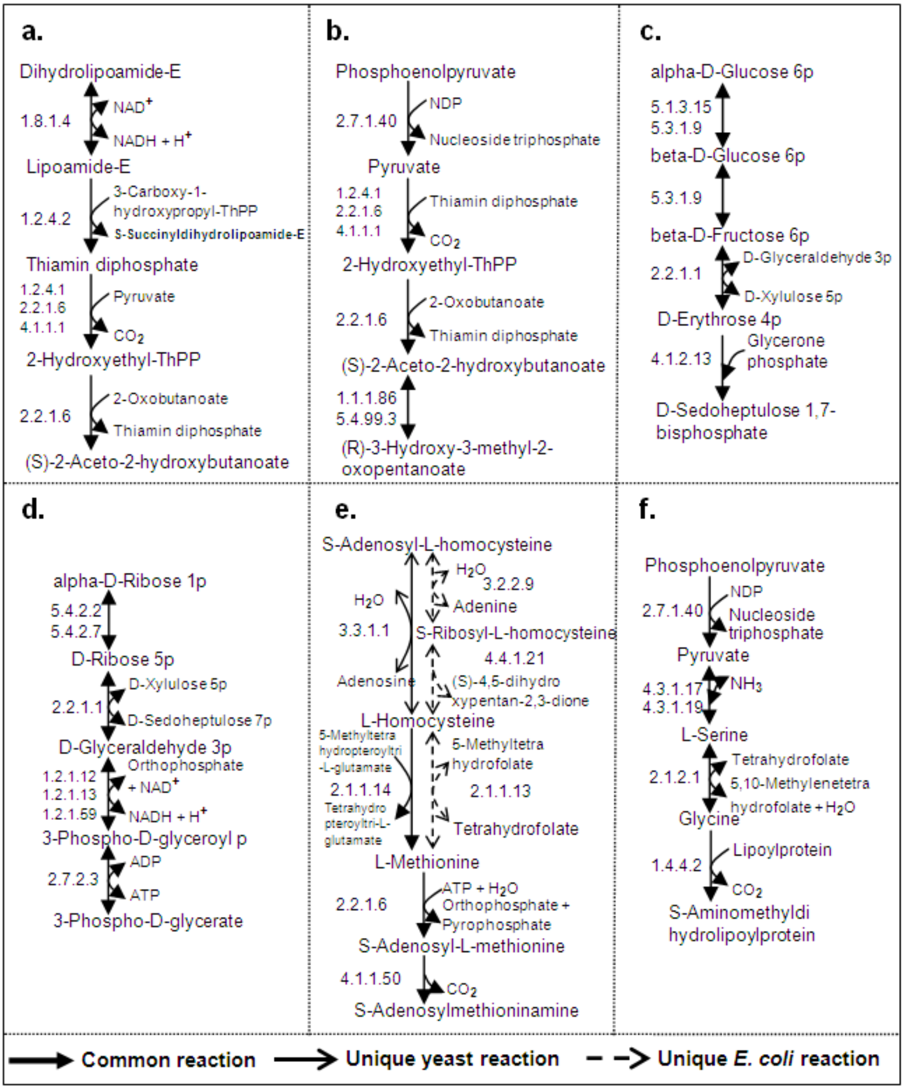

Figure 6.3: Examples in the three queries. **a)** An example best pathway in Query 1. **b)** One pathway which ranks differently in Query 1 of RM-PAS and M-PAS. **c-d)** Example best pathways in Query 2. **e-f)** Example best pathways in Query 3. See text for details.

ranks higher in RM-PAS than in M-PAS, while 28% have lower ranks. Note that only 16% of the changes in the ranking is caused solely by changes in the group size.

In-depth analysis shows the TFs are indeed different in the pathways whose ranks are lower in RM-PAS. For instance, the pathway in Fig. 6.3b has the highest score in M-PAS, but its RM-PAS score is the 30th highest. This is because the TFs in the first

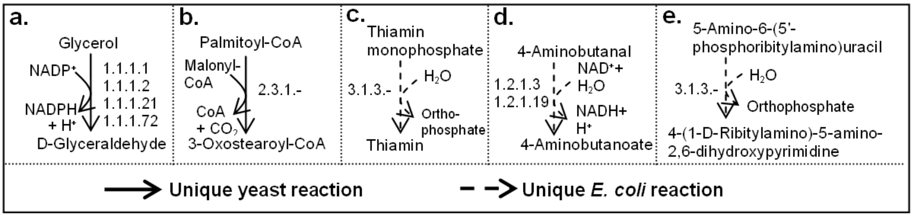

Figure 6.4: Five building blocks belonging to cases 8-10 in Fig. 6.2.

building block are quite different: not only in TFBS matrices, but also in their functional annotations, binding domains, and protein sequences. In fact, the binding domain of the *E. coli* TF fruR is only present in bacteria.

6.4.2 Using one level to infer missing information at another level

Inferring missing reactions Here we use conservation at the regulatory level to infer missing reactions at the metabolic level. Based on the data collected, we constructed five building blocks corresponding to cases 8-10 (see Fig. 6.2), which are shown in Fig. 6.4. In particular, Fig. 6.4d is found in six length-four pathways. In each example, although the reaction is not found in the database for one of the species, we hypothesize that it is actually present. The evidence comes from both metabolic and regulatory levels: all involved compounds and enzymes with the required function are present in the species, and they are also regulated similarly.

Inferring missing TF-gene bindings In Query 2, we try to use conservation at the metabolic level to prioritize a list of hypothetical TF-gene bindings with higher confidence. Overall, the predictions on yeast TF-gene bindings by RM-PAS are significantly better than random predictions. This is validated by a permutation test (see Appendix A), which shows that the TFs predicted by RM-PAS are more likely to bind to the respective genes than random predictions for 50% of the genes.

Here, we give two examples. Fig. 6.3c shows the highest-scoring pathway, involved in glycolysis/gluconeogenesis, pentose phosphate pathway, and carbon fixation. In the fourth building block, we find the bound yeast TF GCR1 is similar to an alternative *E. coli* TF cueR, with MatCompare score $= 0.3$ (the original paper defines two TFs are similar when this score is ≤ 1). It suggests cueR might bind to the *E. coli* enzyme fbaA.

We applied Regulatory Sequence Analysis Tools (RSAT [1]) to see whether the upstream region of fbaA contains the TFBS of cueR. RSAT scans the upstream coding sequence of fbaA for the TFBS matrices of cueR. It outputs a segment score for each sequence segment, which is calculated as the log-ratio between the probability to generate the sequence segment given the TFBS matrix, and the probability to generate the sequence segment given the first-order Markov chain-based background model. The result shows not only that there exists one matching site at -141bp to -120bp, but also that it has a higher segment score than all TFBS of the bound TFs (i.e. fruR and crp) with site-wise p-value $= 0.0005$.

Another example is shown in Fig. 6.3d. In the first building block, we find the bound *E. coli* TF Fis is similar to an alternative yeast TF WAR1, with MatCompare score = 0.5. It suggests WAR1 might bind to the yeast enzyme PGM2. RSAT shows that the TFBS matrix of WAR1 has a higher segment score than 20 (83%) bound TFs, with site-wise p-value $= 0.00002$. In addition, WAR1 shares the same domain "Zn clus" with six bound TFs, according to Pfam [29].

We applied co-expression analysis to investigate the likelihood of this latter TF-gene binding. Our reasoning is that if a particular gene g is regulated by a particular TF T, then g should be more similar than random genes r to other genes g' also regulated by T, in terms of correlation of mRNA expression. This means the average co-expression coefficient between g and g' should be significantly larger than that between r and g'. We used an mRNA microarray data set described earlier [87]. The result shows that the average co-expression coefficient between PGM2 and the set of genes known to be regulated by WAR1 is significantly higher than the co-expression between a randomly drawn gene and the same gene set ($p = 0.001$).

6.4.3 Revealing the differences between two levels

The target pathways in Query 3 are conserved at the metabolic level, yet differ at the regulatory level. As depicted in case 3 in Fig. 6.2, the bound TFs are a "mismatch", even though there exist "alternative" TFs. We further refine our investigation by looking at two types of conservation at metabolic level.

Query 3a looks into the diverse regulation in non-"identical" metabolic building blocks, which contain unique reactions with different cofactors in two species. Therefore, the query actually is designed to find cofactor-specific TFs. Since the enzymes catalyze different reactions in two species, we hypothesize that the different cofactors might have induced different TFs to bind the enzyme-coding genes. These enzyme products in turn enable the same transformation of a particular substrate to a particular product, when different cofactors are available.

Another possible explanation is that different species have evolved separately to produce different cofactors, e.g. ATP, which are actually the main products in some pathways. Several studies show that mutations in active-site residues produce new catalytic properties for enzymes, which enable the formation of new pathways [101]. In our results, we find examples of different TF binding domains that have evolved in different species. For instance, the first building block in Fig. 6.3e contains unique reactions in both species, and the yeast TFs have a bHLH domain present in eukaryotes, and a Zn(2)-C6 fungal-type domain only present in fungal TFs. The second building block contains a unique reaction in *E. coli*, and its enzyme metR has a HTH lysR-type DNA-binding domain unique to bacteria.

Query 3b finds divergent TFBS in the most conserved pathways at metabolic level, with identical reactions in both species. This might indicate the evolution of TFBS [93], and the mutational robustness during the evolution.

Although binding sites are subject to random mutations, evolution has naturally driven TFBS to be unspecific so that the functional phenotype is somewhat insensitive to mutations [147]. Previous research also shows that orthologous transcription factors may regulate orthologous genes through divergent TFBS in distantly related species [2]. This is reflected in our results. For example, the TFBS in the first building block in Fig. 6.3f are very dissimilar in two species with MatCompare score = 2.1, although the enzymes share similar sequences with BLAST E-value $= 4 \times 10^{-68}$.

6.5 Conclusions

RM-PAS combines biological knowledge across species, and across levels of cellular organization. By setting different weight parameters in the scoring function, we showed how RM-PAS can be applied to identify conserved regulatory-metabolic network elements, infer missing reactions, prioritize and corroborate TF-gene binding hypotheses, and reveal diverse regulation in pathways that are conserved at metabolic level.

Our findings may be further exploited to analyze the integrated and aligned network properties, study evolutionary processes in multiple species, seek metabolic engineering targets, predict operons, and provide more possibilities to construct such a multilevel network for a new genome.

6.6 Appendix A: Permutation test

This appendix details one of the validation methods in section 6.4.2. To validate whether the TFs predicted by RM-PAS are more likely to bind to a particular gene than random predictions, we applied the following procedure:

1. Generate the RM-PAS prediction data set. This data set contains the TF-gene pairs predicted by RM-PAS in Query 2. In particular, the genes are the enzyme-coding genes in the best pathways of Query 2, with $Z_{TU} > 0$ (Eq. 6.8).
2. Generate the permuted data set. For each TF-gene pair in the prediction data set, fix the gene and pair it with 10 random TFs that have matrices and that are not known/predicted to bind to this gene.
3. Run RSAT on both the prediction data set and the permuted data set, to obtain a segment score for each TF-gene pair.
4. For each gene, test whether the segment scores of predicted TFs are significantly higher than those of random TFs in permuted data set. This is a one-tailed t-test, assuming that two sets of scores come from normal distributions with unknown and possibly unequal variances. If $p < 0.05$, RM-PAS "wins" this gene test.
5. Perform (4) for all genes in the prediction data set, and obtain the percentage of genes for which RM-PAS wins.

Results: Given 40 genes in total in the prediction data set, RM-PAS is significantly better than random in predicting TFs for 20 genes. Out of the other 20 genes where $p > 0.05$, 19 genes only have 2 or 3 predicted TFs, indicating that small sample size is a major cause of lack of significance.

6.7 Appendix B: Commentary

This appendix presents a follow-up experiment to investigate the added-value of RM-PAS in TF-gene binding prediction, which is addressed in section 6.4.2.

6.7.1 Experiments & results

Our goal is to globally quantify the prediction performance of RM-PAS. To this end, we design a hold-out test in which all experimentally verified yeast TF-gene binding information is left out in the prediction phase, and is used in the validation phase to calculate the true-positive and false-positive prediction rates by RM-PAS. We use the performance of RSAT [1] and random prediction as references. The experimental procedure is as follows (see section 6.7.2 for details):

(A) Generate the RM-PAS prediction data set, containing the yeast TF-gene pairs with their TF similarity scores and reaction similarity scores given by RM-PAS.

(B) Generate the RSAT prediction data set, containing the yeast TF-gene pairs with their RSAT segment scores given by RSAT.

(C) Calculate the overall true positive rate (TPR) and false positive rate (FPR) for RM-PAS. In total, 15 different TF similarity thresholds (t_T) are tested.

(D) Calculate the overall TPR and FPR for RSAT as in (C). In total, 15 different segment score thresholds (t_S) are tested.

The results are shown as ROC curves in Fig. 6.5a. We can see that RM-PAS performs better than random prediction, but not as good as RSAT. We then set out to test two hypotheses as to why this may be the case. Hypothesis I is that uninformative PFMs (position frequency matrices) may lead to large numbers of false positives. To test this hypothesis, we attempt to include the information content of the TF matrices in calculating the TF similarity score in Step (A). More specifically, Kullback-Leibler divergence [133] (see Eq. 6.4) is used to calculate TF similarity. It quantifies the divergence of two matrices, but does not take into account the information content of the matrices. Therefore, two identical and informative matrices have a divergence of 0, but also two identical yet *un*informative matrices have a divergence of 0. We include the information content in the experiments as described in section 6.7.2 below, and refer to this modified version *RM-PASi*.

Hypothesis II is that false positive TF-gene interactions may occur mainly in low-confidence RM-PAS predictions, i.e. in reactions with low similarity score. From Fig. 6.5a, we already see that RM-PAS helps to improve the prediction, using the conservation in the “identical” building blocks (see section 6.7.2 for RM-PAS prediction data set). To test this hypothesis, we add another threshold (t_R) regarding the reaction similarity into RM-PASi, so that the TF-gene interactions are predicted only for building blocks whose reaction similarities exceed t_R. Four thresholds are used. We call this second modification *RM-PASir*.

The performances of the two modified versions of RM-PAS are shown in Fig. 6.5b, using the same way to calculate TPR and FPR as in Steps (C) and (D). Comparable results are observed for a larger RM-PAS prediction data set using all building blocks in the

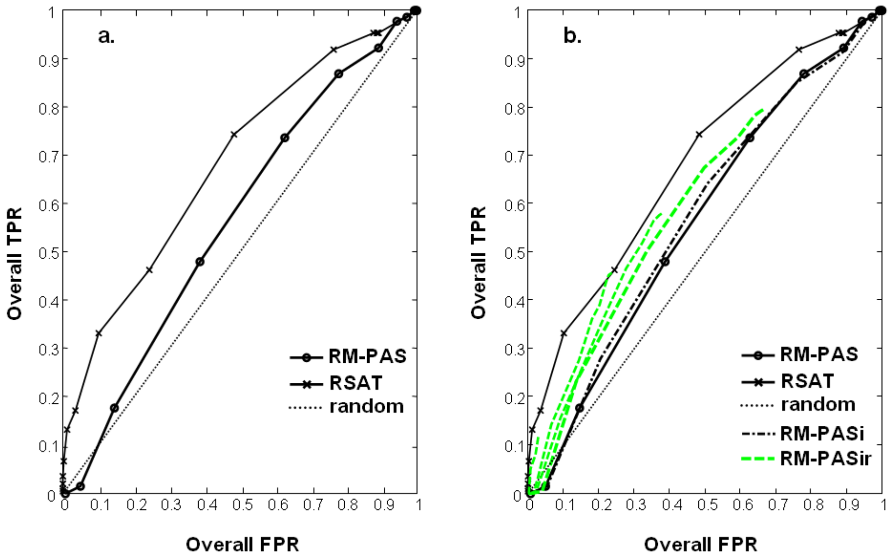

Figure 6.5: a) Global quantification of TF-gene binding prediction performances of RM-PAS and RSAT in a hold-out test. **b)** The prediction performances of RM-PASi and RM-PASir in the same experiment. The ROC curves of RM-PAS and RSAT are the same as in a). The reaction similarity thresholds t_R for the four RM-PASir curves (clockwise) are: 30, 25, 20, 15.

best pathways in Query 2 (see section 6.4) and for all involving genes. We can see from Fig. 6.5b that using information content only marginally improves RM-PAS' performance, indicating the PFM quality in terms of informativeness in the current database is acceptable. Using different reaction similarity thresholds, on the other hand, significantly improves RM-PAS' performance presented by the ROC curves. In particular, given a certain TPR, predictions from more similar reactions (e.g. $t_R = 30$) have a lower FPR than those from less similar reactions (e.g. $t_R = 15$). This indicates that conservation at the metabolic level is indeed correlated with the prediction accuracy at the regulatory level. Therefore, the reaction similarity score can be used to prioritize TF-gene binding hypotheses, although further investigation is needed to learn how to best apply this to obtain more accurate predictions. Moreover, the results also indicate that our scoring function is designed properly to quantify reaction similarity, otherwise we would not be able to obtain this correlation.

In summary, both hypotheses are true, albeit to different extents. However, they still cannot fully explain why RM-PAS is not as good as RSAT. Therefore, we investigated the experiment design and discovered that RSAT has two advantages in getting a better performance. One advantage is that RSAT matches gene sequences with the TFs' PFMs, which is a more direct form of evidence for binding. The other advantage is that RSAT uses the same information source in the training and testing phases. More specifically, a TF is represented by its PFMs, which are mostly obtained from its known bindings to certain genes in experiments. Therefore, if TF a is not represented by PFM m (because this is not found by experiments), scanning any gene sequence for m will not find a in the prediction for any threshold. In case TF a actually binds to certain genes through m, these bindings will not be predicted by RSAT, and are counted as true negatives exactly because they are not experimentally known. In other words, using known TFs to measure RSAT's performance may be biased, and we might have found true bindings by RM-PAS that cannot be validated by RSAT or known bindings.

6.7.2 Method

RM-PAS prediction data set
The data set is generated from all "*identical*" building blocks. The following steps are implemented on each of these building blocks:

1) Calculate its reaction similarity score according to Eq. 6.2, using the parameters of Query 2.

2) Calculate its TF similarity, which is expressed by Kullback-Leibler divergence [133] (see Eq. 6.4). This is calculated for each pair of yeast TF and *E. coli* TF using MatCompare [127]. In such a TF pair y_{TF}-e_{TF}, the yeast TF y_{TF} is one of all yeast TFs which have PFM information, and the *E. coli* TF e_{TF} is one of the TFs that are experimentally known to bind to the enzyme-coding genes in the *E. coli* reaction of that building block. See section 6.3 for the data sources of the PFM information and known bindings.

3) Pair each yeast TF y_{TF} with the yeast enzyme-coding genes in the building block. In case there exist multiple yeast enzymes with different EC numbers, y_{TF} is paired to the genes which have the most similar EC number to that of their *E. coli* counterparts bound by e_{TF}.

In the end, we obtain a data set with yeast TF-gene pairs. Each pair is associated with a reaction similarity score and a K-L divergence score.

RSAT prediction data set
Regulatory Sequence Analysis Tools (RSAT) is a web server which computes the binding likelihood for a TF-gene pair. It scans the upstream region of the gene for the PFM of the TF, and outputs a segment score for each matching sequence segment. To reduce the web query load for RSAT, 30 yeast enzyme-coding genes were randomly chosen, and their upstream regions were scanned for all yeast TFs by RSAT. In the end, we obtain a data set with yeast TF-gene pairs. Each pair has a segment score which is the highest in all segments of that pair.

TPR & FPR calculation
For each method and each threshold, overall TPR and FPR are calculated for a set of genes G by comparing the predicted TFs for G with the ground truth. G includes 28 yeast enzyme-coding genes which overlap between the prediction sets in Steps (A) and (B), and have known yeast TFs (i.e. the bindings are verified by experiments) with PFM information available. The ground truth for the TFs which bind to G comes from the known bindings. See section 6.3 for the data sources of the PFM information and known bindings. The detailed procedure is as follows for each method and each threshold:

1) Select the TF-gene pairs containing genes in G from the prediction set in Step (A) or (B).

2) Obtain the prediction: It is the subset of the pairs in the previous step, whose similarity score exceeds the threshold. For RM-PAS, the K-L divergence score should be lower than t_T. For RSAT, the segment score should be larger than t_S.

3) Obtain the ground truth for known TFs.

4) For each gene in G, compare its predicted TFs with its known TFs, obtaining the numbers in the confusion matrix [2]. This is done for all 28 genes, resulting in an overall TPR and an FPR for each threshold.

RM-PASi

The information content of the PFMs are included in RM-PASi as follows:

1) Calculate the information content I for each matrix m [136].

$$I(m) = \sum_{i=1}^{w} \sum_{j\in\{A,C,G,T\}} m_{ij} \log_2 \frac{m_{ij}}{P_j}, \tag{6.9}$$

where $m_{ij} = \max(m_{ij}, 10^{-10})$, the frequency of observing base j at position i. P_j is the frequency of base j in the whole genome of that species.

2) Transform each I into a scaling factor F through a linear function:

$$F(I) = H - (H-1) \times \frac{I - I_{min}}{I_{max} - I_{min}}, \tag{6.10}$$

where I_{min} and I_{max} are the minimum and maximum of I. The transformation maps I into $F \in [1:H]$ [3].

3) Scale the original K-L divergence D (Eq. 6.4) into D' for a TF pair *a*-*b*:

$$D'(F_a, F_b, D) = \max(F_a, F_b) \times D. \tag{6.11}$$

This means if one of the two matrices under comparison has low information content, their K-L divergence becomes larger (i.e. less similar) after scaling. In the end, D' is used to express TF similarity instead of D to calculate TPR and FPR.

---

[2] The TFs are counted at the complex level, because the TFs in a complex often share the same PFM in the database. Therefore, only one TP is counted even if multiple TFs in a complex are correctly predicted. Similarly, only one FP is counted even if multiple TFs in a complex are wrongly predicted.

[3] Different H are tested. The best performance of RM-PASi is achieved when H is 10.

7

DISCUSSION

From this thesis, we have learned that although various ways of measuring biological data using current technology are unavoidably noisy and incomplete, proper computational methods can help to integrate them to improve predictive power and to aid in knowledge discovery. In Chapter 2 and Chapter 3, it was shown that both class noise and measurement noise significantly impact classification accuracy. Therefore, we should not use the measured data directly without dealing with the noise explicitly. Moreover, both studies pinpoint the importance of understanding the data for noise modelling and classifier construction. In Chapters 4-6, we discovered knowledge and generated hypotheses from incomplete data by data integration. We learned that comparing data from different sources (e.g. species) and types (e.g. metabolic and regulatory) may compensate for incompleteness of the individual data sources. In these studies, we also showed that systematic comparison between complete metabolic networks is necessary to fully explore possible links between reactions, and a refined yet flexible quantification of similarity is valuable.

In the future, developments in the following fields are required to decrease the problems related to noise and data incompleteness. First, measurement techniques should be more efficient and cheaper, so that we can measure more replicates to reduce noise and have a broader coverage. Moreover, attaching a reliability value to individual measurements will be useful for subsequent processing. Second, public databases are to be unified and more comprehensive. This can greatly improve the efficiency and capability of data integration, and also prevent finding artifacts in individual databases. Third, computationally efficient methods are needed to enable more sophisticated noise-handling algorithms and alignment/query algorithms. Fourth, even more species are to be studied, which calls for integration and alignment algorithms for multiple species, where evolutionary distance should be taken into account.

In particular, the following challenges are important:

Understand the cause of the noise. As mentioned in the respective chapters, noise can be caused by many factors including human error and measurement techniques. Understanding the various causes can greatly improve the system's performance, because solutions can be designed to address the specific issues involved. Furthermore, this understanding can also help to identify the steps in the techniques that need to be modified to reduce noise. We already exploited the causes in Chapter 3, for instance, to estimate the noise level per object (i.e. protein pair) and per measurement technique (i.e. experiment). The various noise sources are then individually modelled during classifier construction. As another example, to measure gene expressions from microarray data, Tu *et al.* [144] design sets of replicate experiments to separate the noise introduced in different experimental processes, such as sample preparation and hybridization. This enables them to process the measured data with quantitative characterization of different noise sources.

Reduce the noise before classifier construction. Proper data pre-processing can greatly enhance classification accuracy (see [166] for a review). In particular, data cleansing and feature selection are two useful tools when noise is present. It has been proved that appropriate data cleansing, such as eliminating noisy instances, predicting unknown (or missing) attribute values, or correcting noisy values, can improve classification accuracy. However, this procedure is not trivial, as erroneous cleansing will actually deteriorate a system's performance. Next, feature selection is even more important to build classifiers given noisy data, than given noise-free data. Three factors have to be considered in the selection. The first factor is the quality of the feature. Of course we prefer less noisy features, which can be quantified using some quality analysis methods to estimate their reliability [21, 22]. The second factor is the knowledge about the noise in the feature. As shown in Chapter 3, if we can measure or estimate the noise accurately, the knowledge can be used during the classification. The last factor is the importance of the feature. It is found that noise in different features gives different impact on a system's performance [166]. The higher the correlation between the feature and the class label, the more impact the feature has when a certain amount of noise is introduced. Therefore, to improve classification performance efficiently, we may focus on the important features.

Address the small sample size problem. The small sample size problem is common in computational biology. It means a classifier has to be built upon a few samples with a large number of features, introducing over-fitting to the training samples. The problem becomes more severe when the training data is noisy, because more training samples are required to compensate the distorting effect of the noise, and to recover the true data distribution. Therefore, designing classifiers which can handle the small sample size problem should have our special attention. We have investigated this matter in Chapters 2 & 3, and have shown two useful approaches. One is using classifiers which project data onto a one-dimension space (i.e. PKF in Chapter 2). The other approach includes prior knowledge about the noise in the classifier (Chapter 3).

Validate predictions by experiments. The results of our integrated network comparison approaches are generated by exploiting two complementary types of information, namely similarity and difference. That is, we discover conservation by looking for the similar parts of the networks in both species and at both levels. From the remaining differences, we then learn about the diversity between species and the divergence between levels. Predictions can be made using these differences to infer missing information, e.g. missing reactions and TF-gene bindings, assuming the "different" parts are actually the same. However, the question is: How do we distinguish between real differences and missing information? To consider this, we have to validate our predictions carefully. Given the incomplete data we currently have, ideally we would integrate all available information into the prediction model, leaving no independent data to test the prediction. Therefore, experimental validation is an irreplaceable way to test the findings.

Refine the input networks. Currently, the networks we use only indicate presence/absence of edges (i.e. metabolic reactions or regulatory interactions), in all conditions and time points. That is, they are static and qualitative topologies. However, the actual interactions occur only at certain moments, and under certain environmental and contextual conditions (e.g. combinatorial TF regulations). Moreover, the intensity of the interactions can have a more complex distribution than simply presence/absence. Therefore, a more refined alignment approach ought to describe the edges by their dynamics and quantitative intensities [5, 63]. In the mean time, this poses a daunting task to develop a suitable computational method for comparative network analysis.

BIBLIOGRAPHY

[1] *http://rsat.ulb.ac.be/rsat/*. Pages: 82, 85

[2] W. Alkema, B. Lenhard, and W. Wasserman. Regulog analysis: detection of conserved regulatory networks across bacteria: application on *Staphylococcus aureus*. *Genome Res.*, 14:1362–1373, 2004. Pages: 74, 84

[3] D. Angluin and P. Laird. Learning from noisy examples. *Mach. Learn.*, 2:343–370, 1988. Pages: 11

[4] Z. Bar-Joseph, G. Gerber, and T. Lee *et al.* Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.*, 21:1337–1342, 2003. Pages: 2

[5] A. Barabási and Z. Oltvai. Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet.*, 5:101–113, 2004. Pages: 3, 91

[6] A. Ben-Hur and W. Noble. Kernel methods for predicting protein-protein interactions. *Bioinfo.*, 21:i38–i46, 2005. Pages: 38

[7] L. Breiman, W. Meisel, and E. Purcell. Variable kernel estimates of multivariate densities. *Technometrics*, 19:135–144, 1977. Pages: 25

[8] C. Brodley and M. Freidl. Identifying mislabeled training data. *J. Artif. Intell. Res.*, 11:131–167, 1999. Pages: 10

[9] P. Cheeseman. On finding the most probable model. In J. Shrager and P. Langley, editors, *Computational Models of Scientific Discovery and Theory Formation*, pages 73–95. Morgan Kaufmann Publishers, Inc, San Mateo, CA, 1990. Pages: 24

[10] L. Chen and D. Vitkup. Predicting genes for orphan metabolic activities using phylogenetic profiles. *Genome Biol.*, 7:R17, 2006. Pages: 3

[11] Q. Chen, G. Hertz, and G. Stormo. Matrix search 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Comput. Appl. Biosci.*, 11:563–566, 1995. Pages: 79

[12] J. Clemente, K. Satou, and G. Valiente. Finding conserved and non-conserved reactions using a metabolic pathway alignment algorithm. *Genome Info.*, 17:46–56, 2006. Pages: 40, 70

[13] J. Clemente, K. Satou, and G. Valiente. Phylogenetic reconstruction from non-genomic data. *Bioinfo.*, 23:e110–e115, 2006. Pages: 3, 40, 46, 51, 54, 58

[14] W. Cleveland. Robust locally weighted regression and smoothing scatterplots. *J. Amer. Stat. Assoc.*, 74:829–836, 1979. Pages: 32

[15] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995. Pages: 38

[16] R. Courant and D. Hilbert. *Methods of Mathematical Physics*. Wiley, New York, NY, 1959. Pages: 14

[17] M. Covert, E. Knight, J. Reed, M. Herrgard, and B. Palsson. Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, 429: 92–96, 2004. Pages: 75

[18] D. Croes, F. Couche, S. Wodak, and J. van Helden. Inferring meaningful path-

wyas in weighted metabolic networks. *J. Mol. Biol.*, 356:222–236, 2006. Pages: 2, 58, 59, 64

[19] T. Dandekar, S. Schuster, B. Snel, M. Huynen, and P. Bork. Pathway alignment: Application to the comparative analysis of glycolytic enzymes. *Biochem. J.*, 343: 115–124, 1999. Pages: 3, 40, 54, 72

[20] B. Dasarathy. Nosing around the neighborhood: a new system structure and classification rule for recognition in partial exposed environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2:67–71, 1980. Pages: 10

[21] C. Deane, L. Salwinski, I. Xenarios, and D. Eisenberg. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Proteomics*, 1:349–356, 2002. Pages: 1, 90

[22] M. Deng, F. Sun, and T. Chen. Assessment of the reliability of protein-protein interactions and protein function prediction. *PSB*, 8:140–151, 2003. Pages: 90

[23] D. Devos and A. Valencia. Intrinsic errors in genome annotation. *Trends Genet.*, 17:429–431, 2001. Pages: 2

[24] J. Díaz-Mejía, E. Pérez-Rueda, and L. Segovia. A network perspective on the evolution of metabolism by gene duplication. *Genome Biol.*, 8:R26, 2007. Pages: 54

[25] R. Duin. On the choice of smoothing parameters for parzen estimators of probability density functions. *IEEE Trans. Comput.*, 25:1175–1179, 1976. Pages: 4, 25

[26] R. Duin. *PRTools Version 3.0, A Matlab Toolbox for Pattern Recognition*. Pattern Recognition Group, Delft University of Technology, The Netherlands, 2000. Pages: 27

[27] P. Durek and D. Walther. The integrated analysis of metabolic and protein interaction networks reveals novel molecular organizing principles. *BMC Sys. Biol.*, 2:100, 2008. Pages: 3

[28] D. Fell and A. Wagner. The small world of metabolism. *Nat. Biotech.*, 18:1121–1122, 2000. Pages: 61

[29] R. Finn, J. Tate, and J. Mistry *et al.* The Pfam protein families database. *Nucl. Acids Res.*, 36:D281–D288, 2008. Pages: 83

[30] R. Fisher. The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, 7:179–188, 1936. Pages: 13

[31] E. Fix and J. Hodges. Discriminatory analysis, nonparametric discrimination: consistency properties. In *Report no. 4, Project no. 21-49-004*. USAF School of Aviation Medicine, Randolph Field, TX, 1951. Pages: 24

[32] C. Forst and K. Schulten. Evolution of metabolisms: A new method for the comparison of metaboic pathways using genomic information. *J. Comput. Biol.*, 6:343–360, 1999. Pages: 3

[33] C. Forst and K. Schulten. Phylogenetic analysis of metabolic pathways. *J. Mol. Evol.*, 52:471–489, 2001. Pages: 3, 40, 54

[34] C. Francke, R. Siezen, and B. Teusink. Reconstructing the metabolic network of a bacterium from its genome. *Trends in Microbiology*, 13:550–558, 2005. Pages: 3

[35] S. Gama-Castro, V. Jacinto, and M. Peralta-Gil *et al.* RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucl. Acids Res.*, 36:D120–D124, 2008. Pages: 79

[36] D. Gamberger, N. Lavrac, and C. Groselj. Experiments with noise filtering in a medical domain. In *Proceedings of the 16th International Conference on Machine Learning*, pages 143–151, 1999. Pages: 10

[37] A. Gasch, A. Moses, D. Chiang, H. Fraser, M. Berardini, and M. Eisen. Conser-

vation and evolution of cis-regulatory systems in ascomycete fungi. *PLoS Biol.*, 2:e398, 2004. Pages: 74

[38] G. Gates. The reduced nearest neighbor rule. *IEEE Trans. Inf. Theory*, 18:431–433, 1972. Pages: 10

[39] A. Gavin, M. Bösche, and R. Krause *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415:141–147, 2002. Pages: 30

[40] H. Ge, Z. Liu, G. Church, and M. Vidal. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat. Genet.*, 29:482–486, 2001. Pages: 30

[41] S. Ghaemmaghami, W. Huh, and K. Bower *et al.* Global analysis of protein expression in yeast. *Nature*, 425:737–741, 2003. Pages: 1

[42] D. Ghosh. OOTFD (object-oriented transcription factors database): an object-oriented successor to TFD. *Nucl. Acids Res.*, 26:360–361, 1998. Pages: 79

[43] D. Goldberg and F. Roth. Assessing experimentally derived interactions in a small world. *PNAS*, 100:4372–4376, 2003. Pages: 3

[44] J. Gollub, C. Ball, and G. Binkley *et al.* The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res.*, 31:9496, 2003. Pages: 31

[45] S. Goto, T. Nishioka, and M. Kanehisa. Ligand: chemical database for enzyme reactions. *Bioinfo.*, 14:591–599, 1998. Pages: 42, 61, 70

[46] S. Goto, T. Nishioka, and M. Kanehisa. LIGAND: chemical database for enzyme reactions. *Bioinfo.*, 14:591–599, 1998. Pages: 79

[47] M. Green and P. Karp. A bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinfo.*, 5:76, 2004. Pages: 2

[48] R. Guimerà, M. Sales-Pardo, and L. Amaral. A network-based method for target selection in metabolic networks. *Bioinfo.*, 23:1616–1622, 2007. Pages: 3, 54

[49] I. Guyon, N. Matic, and V. Vapnik. Discovering informative patterns and data cleaning. In *Advances in Knowledge Discovery and Data Mining*, pages 181–203. AAAI/MIT Press, Cambridge, MA, 1996. Pages: 4, 10

[50] F. Hampel, E. Ronchetti, P. Rousseeuw, and W. Stahel. *Robust StatisticsThe Approach Based on Influence Functions*. Wiley, New York, NY, 1986. Pages: 38

[51] J. Hanley and B. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:2936, 1982. Pages: 32

[52] L. Hedges and I. Olkin. *Statistical methods for meta-analysis*. Academic Press, Orlando, FL, 1985. Pages: 58, 77

[53] M. Heymans and A. Singh. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. *Bioinfo.*, 19:i138–i146, 2003. Pages: 3, 46, 54, 74

[54] W. Highleyman. Linear decision functions, with application to pattern recognition. In *Proceedings of the 50th IRE*, pages 1501–1514, 1962. Pages: 27

[55] L. Hood and D. Galas. The digital code of DNA. *Nature*, 421:444–448, 2003. Pages: 6

[56] P. Huber. *Robust Statistics*. Wiley, New York, NY, 1981. Pages: 38

[57] J. Hughes, P. Estep, S. Tavazoie, and G. Church. Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, 296:1205–1214, 2000. Pages: 59

[58] T. Hughes, M. Marton, and A. Jones *et al.* Functional discovery via a compendium of expression profiles. *Cell*, 102:109–126, 2000. Pages: 31

[59] D. Hwang, A. Rust, and S. Ramsey *et al.* A data integration methodology for systems biology. *PNAS*, 102:17296–17301, 2005. Pages: 1, 2, 5, 24, 58

[60] J. Ihmels, R. Levy, and N. Barkai. Principles of transcriptional control in the

metabolic network of *Saccharomyces cerevisiae*. *Nature Biotech.*, 22:86–92, 2004. Pages: 75

[61] I. Iliopoulos, S. Tsoka, and M. Andrade *et al.* Genome sequences and great expectations. *Genome Biol.*, 2:i0001.1–i0001.3, 2001. Pages: 2

[62] I. Iliopoulos, S. Tsoka, and M. Andrade *et al.* Evaluation of annotation strategies using an entire genome sequence. *Bioinfo.*, 19:717–726, 2003. Pages: 2

[63] R. Jansen and M. Gerstein. Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Curr. Opin. Microbiol.*, 7:535–545, 2004. Pages: 2, 91

[64] R. Jansen, N. Lan, J. Qian, and M. Gerstein. Integration of genomic datasets to predict protein complexes in yeast. *J. Struct. Func. Genomics*, 2:71–81, 2002. Pages: 2, 30

[65] R. Jansen, H. Yu, and D. Greenbaum *et al.* A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302:449–453, 2003. Pages: 2, 30, 31

[66] H. Jeong, B. Tombor, R. Albert, Z. Oltvai, and A. Barabási. The large-scale organization of metabolic networks. *Nature*, 406:651–654, 2000. Pages: 54, 66, 74

[67] G. John. Robust decision tree: removing outliers from data. In *Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining*, pages 174–179, 1995. Pages: 10

[68] P. Karp, C. Ouzounis, and S. Paley. HinCyc: a knowledge base of the complete genome and metabolic pathways of H. influenzae. In *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, pages 116–124, 1996. Pages: 2

[69] T. Kato, K. Tsuda, and K. Asai. Selective integration of multiple biological data for supervised network inference. *Bioinfo.*, 21:2488–2495, 2005. Pages: 2

[70] B. Kelley, R. Sharan, R. Karp, T. Sittler, D. Root, B. Stockwell, and T. Ideker. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *PNAS*, 100:11394–11399, 2003. Pages: 40, 46, 51, 55

[71] R. Kelley and T. Ideker. Systematic interpretation of genetic interactions using protein networks. *Nat. Biotech.*, 23:561–566, 2005. Pages: 3

[72] I. Keseler, J. Collado-Vides, and S. Gama-Castro *et al.* Ecocyc: a comprehensive database resource for *Escherichia coli*. *Nucl. Acids Res.*, 33:D334–D337, 2005. Pages: 79

[73] P. Kharchenko, D. Vitkup, and G. Church. Filling gaps in a metabolic network using expression information. *Bioinfo.*, 20:i178–i185, 2004. Pages: 2, 3

[74] S. Kielbasa, D. Gonze, and H. Herzel. Measuring similarities between transcription factor binding sites. *BMC Bioinfo.*, 6:237, 2005. Pages: 78

[75] H. Kitano. Systems biology: A brief overview. *Science*, 295:1662–1664, 2002. Pages: 2

[76] S. Knuutila, A. Björkqvist, and K. Autio *et al.* DNA copy number amplifications in human neoplasms - review of comparative genomic hybridization studies. *Ame. J. Pathol.*, 152:1107–1123, 1998. Pages: 1

[77] L. Krishnamurthy, J. Nadeau, G. Ozsoyoglu, M. Ozsoyoglu, G. Schaeffer, M. Tasan, and W. Xu. Pathways database system: An integrated system for biological pathways. *Bioinfo.*, 19:930–937, 2003. Pages: 56

[78] R. Küffner, R. Zimmer, and T. Lengauer. Pathway analysis in metabolic databases via differential metabolic display (DMD). *Bioinfo.*, 16:825–836, 2000. Pages: 54

[79] A. Kumar, S. Agarwal, and J. Heyman *et al.* Subcellular localization of the yeast

proteome. *Genes Dev.*, 16:707–719, 2002. Pages: 1

[80] H. Lähdesmäki, A. Rust, and I. Shmulevich. Probabilistic inference of transcription factor binding from multiple data sources. *PLoS ONE*, 3:e1820, 2008. Pages: 2

[81] N. Lawrence and B. Schölkopf. Estimating a kernel fisher discriminant in the presence of label noise. In *Proceedings of the 18^{th} International Conference on Machine Learning*, page 306313, 2001. Pages: 4, 5, 6, 9, 10, 11, 12, 15, 16, 17, 18, 20, 21

[82] D. Lewis, T. Jebara, and W. Noble. Support vector machine learning from heterogeneous data: an empirical analysis using protein sequence and structure. *Bioinfo.*, 22:2753–2760, 2006. Pages: 2

[83] S. Li, C. Armstrong, and N. Bertin *et al.* A map of the interactome network of the metazoan *C. elegans*. *Science*, 303:540–543, 2004. Pages: 3

[84] Y. Li, J. Bot, M. Reinders, and D. de Ridder. A pathway approach to align regulatory-metabolic networks. In *Proceedings of the 8^{th} Annual International Conference on Computational Systems Bioinformatics*, 2009. Pages: 8, 73

[85] Y. Li, D. de Ridder, M. de Groot, and M. Reinders. Metabolic pathway alignment between species using a comprehensive and flexible similarity measure. *BMC Systems Biology*, 2:111, 2008. Pages: 7, 53, 75, 77, 78

[86] Y. Li, D. de Ridder, M. de Groot, and M. Reinders. Metabolic pathway alignment (M-Pal) reveals diversity and alternatives in conserved networks. In A. Brazma, S. Miyano, and T. Akutsu, editors, *Advances in Bioinformatics & Computational Biology*, volume 6, pages 273–285. Imperial College Press, London, 2008. Pages: 7, 39, 55, 56, 75

[87] Y. Li, D. de Ridder, R. Duin, and M. Reinders. Integration of prior knowledge of measurement noise in kernel density classification. *Pattern Recognition.*, 41: 320–330, 2008. Pages: 7, 23, 83

[88] Y. Li, L. Wessels, D. de Ridder, and M. Reinders. Classification in the presence of class noise using a Probabilistic Kernel Fisher method. *Pattern Recognition*, 40: 3349–3357, 2007. Pages: 6, 9

[89] Z. Li, S. Zhang, Y. Wang, X. Zhang, and L. Chen. Alignment of molecular networks by integer quadratic programming. *Bioinfo.*, 23:1631–1639, 2007. Pages: 46

[90] D. Loftsgaarden and C. Quesenberry. A nonparametric estimate of a multivariate density function. *Ann. Math. Stat.*, 36:1049–1051, 1965. Pages: 25

[91] L. Lu, Y. Xia, A. Paccanaro, H. Yu, and M. Gerstein. Assessing the limits of genomic data integration for predicting protein networks. *Genome Res.*, 15:945–953, 2005. Pages: 30, 31

[92] H. Ma and A. Zeng. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinfo.*, 19:270–277, 2003. Pages: 40, 44, 61

[93] M. Madan Babu and S. Teichmann. Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. *Nucleic Acids Res.*, 31:1234–1244, 2003. Pages: 83

[94] H. Mann and D. Whitney. On a test of whether one of 2 random variables is stochastically larger than the other. *Ann. Math. Stat.*, 18:5060, 1947. Pages: 33

[95] R. Maronna, D. Martin, and V. Yohai. *Robust Statistics: Theory and Methods*. Wiley, New York, NY, 2006. Pages: 38

[96] L. Matthews, P. Vaglio, J. Reboul, H. Ge, B. Davis, J. Garrels, S. Vincent, and M. Vidal. Identification of potential interaction networks using sequence-based

searches for conserved protein-protein interactions or "interologs". *Genome Res.*, 11:2120–2126, 2001. Pages: 3

[97] The UniProt Consortium. The universal protein resource (UniProt). *Nucleic Acids Res.*, 36:D190–D195, 2008. Pages: 61, 79

[98] H. Mewes, C. Amid, and R. Arnold *et al.* MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.*, 32:D41–D44, 2004. Pages: 31

[99] S. Mika, G. Ratsch, J. Weston, B. Schölkopf, and K. Muller. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX, Proceedings of the 1999 IEEE Signal Processing Society Workshop*, page 4148, 1999. Pages: 13, 17

[100] J. Mingers. An empirical comparison of pruning methods for decision tree induction. *Mach. Learn.*, 4:227–243, 1989. Pages: 4, 10

[101] A. Murzin. Can homologous proteins evolve different enzymatic activities? *Trends. Biochem. Sci.*, 18:403–405, 1993. Pages: 83

[102] C. Myers, D. Robson, A. Wible, M. Hibbs, C. Chiriac, C. Theesfeld, K. Dolinski, and O. Troyanskaya. Discovery of biological networks from diverse functional genomic data. *Genome Biol.*, 6:R114, 2005. Pages: 2

[103] R. Nadon and J. Shoemaker. Statistical issues with microarrays: processing and analysis. *Trends Genet.*, 18:265–271, 2002. Pages: 4

[104] J. Nicholson and I. Wilson. Understanding "global" systems biology: metabonomics and the continuum of metabolism. *Nat. Reviews Drug Disc.*, 2:668–676, 2003. Pages: 1, 2

[105] Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NCIUBMB). *http://www.chem.qmul.ac.uk/iubmb/enzyme/*. Pages: 40

[106] H. Ogata, W. Fujibuchi, S. Goto, and M. Kanehisa. A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucl. Acids Res.*, 28:4021–4028, 2000. Pages: 3

[107] R. Overbeek, N. Larsen, W. Smith, N. Maltsev, and E. Selkov. Representation of function: the next step. *Gene*, 191:GC1–GC9, 1997. Pages: 40

[108] M. Pachkov, I. Erb, N. Molina, and E. van Nimwegen. SwissRegulon: a database of genome-wide annotations of regulatory sites. *Nucl. Acids Res.*, 35:D127–D131, 2007. Pages: 79

[109] P. Paclík, J. Novovicv, P. Pudil, and P. Somol. Road sign classification using Laplace kernel classifier. *Pattern Recognition Lett.*, 21:1165–1173, 2000. Pages: 30

[110] B. Palsson. The challenges of in silico biology. *Nat. Biotech.*, 18:1147–1150, 2000. Pages: 2

[111] B. Palsson. In silico biotechnology. era of reconstruction and interrogation. *Curr. Opin. Biotech.*, 15:50–51, 2004. Pages: 3

[112] C. Pan, G. Kora, and W. McDonald *et al.* ProRata: A quantitative proteomics program for accurate protein abundance ratio estimation with confidence interval evaluation. *Anal. Chem.*, 78:7121–7131, 2006. Pages: 5

[113] E. Parzen. On estimation of a probability density function and mode. *Ann. Math. Stat.*, 33:1065–1076, 1962. Pages: 24

[114] K. Patil and J. Nielsen. Uncovering transcriptional regulaton of metabolism by using metabolic network topology. *PNAS*, 102:2685–2689, 2005. Pages: 75

[115] M. Pellegrini, E. Marcotte, M. Thompson, D. Eisenberg, and T. Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *PNAS*, 96:4285–4288, 1999. Pages: 3

[116] R. Pinter, O. Rokhlenko, E. Yeger-Lotem, and M. Ziv-Ukelson. Alignment of metabolic pathways. *Bioinfo.*, 21:3401–3408, 2005. Pages: 3, 54, 63

[117] J. Quinlan. Induction of decision trees. *Mach. Learn.*, 1:81–106, 1986. Pages: 4, 10

[118] J. Quinlan. Bagging, boosting and C4.5. In *Proceedings of the 13th National Conference on Artificial Intelligence*, pages 725–730, 1996. Pages: 10

[119] S. Raudys and A. Jain. Small sample size effects in statistical pattern recognition: recommendations for practitioners. *IEEE Trans. Pattern Analy. Mach. Intel.*, 13: 252–264, 1991. Pages: 4

[120] E. Ravasz, A. Somera, D. Mongru, Z. Oltvai, and A. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551–1555, 2002. Pages: 54, 61, 74

[121] B. Ren, F. Robert, and J. Wyrick *et al.* Genome-wide location and function of DNA binding proteins. *Science*, 290:2306–2309, 2000. Pages: 1

[122] M. Rosenblatt. Remarks on some non-parametric estimates of a density function. *Ann. Math. Stat.*, 27:832–837, 1956. Pages: 24

[123] Y. Sakakibara. Noise-tolerant occam algorithms and their applications to learning decision trees. *Mach. Learn.*, 11:37–62, 1993. Pages: 10

[124] E. Sandmeier, T. Hale, and P. Christen. Multiple evolutionary origin of pyridoxal-5′-phosphate-dependent amino acid decarboxylases. *European J. Biochem.*, 221:997–1002, 1994. Pages: 65

[125] R. Schapire. The strength of weak learnability. *Mach. Learn.*, 5:197–227, 1990. Pages: 10

[126] B. Schölkopf, P. Simard, V. Vapnik, and A. Smola. Prior knowledge in support vector kernels. *Adv. Neural Inf. Process Syst.*, 10:640–646, 1998. Pages: 24

[127] D. Schones, P. Sumazin, and M. Zhang. Similarity of position frequency matrices for transcription factor binding sites. *Bioinfo.*, 21:307–313, 2005. Pages: 78, 87

[128] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition specific regulators from gene expression data. *Nat. Genet.*, 34:166–176, 2003. Pages: 24

[129] E. Selkov, N. Maltsev, G. Olsen, R. Overbeek, and W. Whitman. A reconstruction of the metabolism of *Methanococcus jannaschii* from sequence data. *Gene*, 197: GC11–GC26, 1997. Pages: 2

[130] R. Sharan and T. Ideker. Modeling cellular machinery through biological network comparison. *Nat. Biotech.*, 24:427–433, 2006. Pages: 1, 2, 3, 6, 54, 56, 75

[131] R. Sharan, S. Suthram, R. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R. Karp, and T. Ideker. Conserved patterns of protein interaction in multiple species. *PNAS*, 102:1974–1979, 2005. Pages: 3, 30

[132] E. Smid, D. Molenaar, J. Hugenholtz, W. de Vos, and B. Teusink. Functionaly ingredient production: application of global metabolic models. *Curr. Opin. Biotech.*, 16:190–197, 2005. Pages: 3

[133] A. Smith, P. Sumazin, and M. Zhang. Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *PNAS*, 102:1560–1565, 2005. Pages: 78, 85, 87

[134] P. Sridhar, T. Kahveci, and S. Ranka. An iterative algorithm for metabolic network-based drug target identification. *PSB*, 12:88–99, 2007. Pages: 3

[135] A. Srinivasan, S. Muggleton, and M. Bain. Distinguishing exception from noise in non-monotonic learning. In *Proceedings of the 2nd ILP Workshop*, pages 97–107, 1992. Pages: 21

[136] G. Stormo. DNA binding sites: representation and discovery. *Bioinfo.*, 16:16–23, 2000. Pages: 88

[137] J. Stuart, E. Segal, D. Koller, and S. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302:249–255, 2003. Pages: 3

[138] S. Tai, V. Boer, P. Daran-Lapujade, M. Walsh, J. deWinde, J. Daran, and J. Pronk.

Two-dimensional transcriptome analysis in chemostat cultures: combinatorial effects of oxygen availability and macronutrient limitation in *Saccharomyces cerevisiae*. *J. Biol. Chem.*, 280:437–447, 2005. Pages: 31

[139] A. Tanay, A. Regev, and R. Shamir. Conservation and evolvability in regulatory networks: The evolution of ribosomal regulation in yeast. *PNAS*, 102:7203–7208, 2005. Pages: 74

[140] D. Tax and R. Duin. Support vector data description. *Mach. Learn.*, 54:45–66, 2004. Pages: 21

[141] M. Teixeira, P. Monteiro, and P. Jain *et al.* The YEASTRACT database: a tool for the analysis of transcription regulatory associations in *Saccharomyces cerevisiae*. *Nucl. Acids Res.*, 34:D446–D451, 2006. Pages: 79

[142] Y. Tohsato, H. Matsuda, and A. Hashimoto. A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*, pages 376–383, 2000. Pages: 40, 46, 54, 58, 65

[143] T. Tran, P. Dam, Z. Su, F. Poole II, M. Adams, G. Zhou, and Y. Xu. Operon prediction in *Pyrococcus furiosus*. *Nucl. Acids Res.*, 35:11–20, 2007. Pages: 3

[144] Y. Tu, G. Stolovitzky, and U. Klein. Quantitative noise analysis for gene expression microarray experiments. *PNAS*, 99:14031–14036, 2002. Pages: 1, 90

[145] P. Uetz, L. Giot, and G. Cagney *et al.* A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403:623–627, 2000. Pages: 1

[146] E. van Beers, T. vanWelsem, L. Wessels, Y. Li, R. Oldenburg, P. Devilee, C. Cornelisse, S. Verhoef, F. Hogervorst, L. van't Veer, and P. Nederlof. Comparative genomic hybridization profiles in human BRCA1 and BRCA2 breast tumors highlight differential sets of genomic aberrations. *Cancer Res.*, 65:822–827, 2005. Pages: 18

[147] E. van Nimwegen, J. Crutchfield, and M. Huynen. Neutral evolution of mutational robustness. *PNAS*, 96:9716–9720, 1999. Pages: 83

[148] P. Vannoorenberghe and T. Denoeux. Handling uncertain labels in multiclass problems using belief decision trees. In *Proceedings of the International Conference on Processing and Management of Uncertainty*, pages 1916–1926, 2002. Pages: 4, 10

[149] C. von Mering, R. Krause, B. Snel, M. Cornell, G. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417:399–403, 2002. Pages: 1, 2, 30

[150] W. Wasserman and A. Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, 5:276–287, 2004. Pages: 78

[151] L. Wessels, T. van Welsem, A. Hart, L. van't Veer, M. Reinders, and P. Nederlof. Molecular classification of breast carcinomas by comparative genomic hybridization: a specific somatic genetic profile for BRCA1 tumors. *Cancer Res.*, 62: 7110–7117, 2002. Pages: 18, 19

[152] D. Wilson and T. Martinez. Instance pruning techniques. In *Proceedings of the 14th International Conference on Machine Learning*, pages 404–411, 1997. Pages: 10

[153] E. Wingender, X. Chen, and E. Fricke *et al.* The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.*, 29:281–283, 2001. Pages: 79

[154] E. Winzeler, D. Shoemaker, and A. Astromoff *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science*, 285: 901–906, 1999. Pages: 1

[155] L. Wu, T. Hughes, A. Davierwala, M. Robinson, R. Stoughton, and S. Altschuler. Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nat. Genet.*, 31:255–265, 2002. Pages: 1

[156] X. Wu. *Knowledge aquisition from database*. Ablex Publishing Corp., Norwood, NJ, USA, 1995. Pages: 4, 10

[157] R. Wurtman, W. Shoemaker, and F. Larin. Mechanism of the daily rhythm in hepatic tyrosine transaminase activity: role of dietary tryptophan. *PNAS*, 59: 800–807, 1968. Pages: 50

[158] Q. Yang and S. Sze. Path matching and graph matching in biological networks. *J. Comput. Biol.*, 14:56–67, 2007. Pages: 54

[159] C. Yeang and M. Vingron. A joint model of regulatory and metabolic networks. *BMC Bioinfo.*, 7:332, 2006. Pages: 3

[160] H. Yu, N. Luscombe, and X. Lu *et al.* Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res.*, 14:1107–1118, 2004. Pages: 30

[161] X. Zeng and T. Martinez. A noise filtering method using neural networks. In *IEEE International Workshop on Soft Computing Techniques in Instrumentation, Measurement and Related Applications*, pages 26–31, 2003. Pages: 10

[162] G. Zhang, Z. Cao, Q. Luo, Y. Cai, and Y. Li. Operon prediction based on SVM. *Comput. Biol. Chem.*, 30:233–240, 2006. Pages: 3

[163] Y. Zheng, J. Szustakowski, L. Fortnow, R. Roberts, and S. Kasif. Computational identification of operons in microbial genomes. *Genome Res.*, 12:1221–1230, 2002. Pages: 3

[164] L. Zhenping, S. Zhang, Y. Wang, X. Zhang, and L. Chen. Alignment of molecular networks by integer quadratic programming. *Bioinfo.*, 23:1631–1639, 2007. Pages: 54

[165] D. Zhu and Z. Qin. Structural comparison of metabolic networks in selected single cell organisms. *BMC Bioinfo.*, pages 6–8, 2005. Pages: 40, 44

[166] X. Zhu and X. Wu. Class noise vs. attribute noise: A quantitative study of their impacts. *A.I. Rev.*, 22:177–210, 2004. Pages: 90

[167] X. Zhu, X. Wu, and Q. Chen. Eliminating class noise in large datasets. In *Proceedings of the 20^{th} International Conference on Machine Learning*, pages 920–927, 2003. Pages: 4, 10

SUMMARY

In modern molecular biology, the vast amount of experimental data enables us to obtain more comprehensive understanding of cellular activities, from transcription to metabolism. However, due to the inherent complexity of the cell and the various limitations of the measuring techniques, these data are often noisy and incomplete. Therefore, conclusions and hypotheses generated from these data are unreliable and remain partial. This poses a major challenge in molecular biology.

This thesis contributes to this matter by proposing several approaches to handle noisy and incomplete biological data, in order to improve prediction accuracy and ease knowledge discovery. It is divided into two parts which address different problems. Part I is dedicated to the theoretical study of building noise-tolerant classifiers in the presence of class noise and measurement noise, i.e. when class labels or measured attribute values of biological instances are erroneous. For the class noise problem, we present three classifiers using probabilistic models to recover the true distribution of each class. In particular, our novel incorporation of the noise model in the Kernel Fisher discriminant offers improved prediction performance, especially on non-Gaussian data sets and data sets with relatively large numbers of features compared to their sample sizes. For measurement noise, we propose to integrate prior knowledge of the noise into kernel density based classifiers, using distinct kernels for individual samples, features, and feature values. The inclusion of prior knowledge is also shown to be especially beneficial in relatively under-sampled data sets.

In Part II, we exploit the incomplete metabolic reaction and transcriptional regulation data, using both a network-centric and evolution-based approach. That is, we integrate metabolic networks and regulatory networks within species, and compare the integrated networks across different species. This integrated evolutionary network method not only provides a more comprehensive view of the cellular system, but also helps to generate more reliable information and hypotheses. Our alignment framework allows to automatically align the full metabolic networks of two species, taking into account all reaction arrangement possibilities and allowing small differences in otherwise similar reactions. We present a scoring function which measures pathway similarity in a comprehensive and flexible manner, hierarchically integrating all relevant and uncorrelated information sources. Using this method, we have identified fully conserved pathways and their variations at regulatory and metabolic level, discovered new pathway possibilities which are not represented in conventional databases, and generated hypotheses on the missing information using the information of its counterpart at another level and/or another species.

SAMENVATTING

De moderne moleculaire biologie beschikt over een enorme hoeveelheid data, die ons in staat stelt om een beter begrip te krijgen van de cellulaire activiteiten van transcriptie tot metabolisme. Door de inherente complexiteit van de cel en de beperkingen van de meettechnieken is de data echter vaak incompleet en bevat ruis. Daarom zijn de conclusies en hypotheses die op deze data gebaseerd zijn incompleet en onbetrouwbaar. Dit vormt een grote uitdaging voor de moleculaire biologie.

Dit proefschrift draagt bij aan deze materie door verschillende methoden voor te stellen om met incomplete data en data met ruis om te gaan, ten einde de accuraatheid van voorspellingen te verbeteren en het ontdekken van kennis te vergemakkelijken. Het bevat twee delen die verschillende problemen addresseren. Het eerste deel bevat een theoretische studie betreffende het ontwikkelen van ruistolerante klassificatoren in de aanwezigheid van: 1) fouten in de toegewezen klassen van voorbeelden in de leerset (klassenruis) en 2) meetruis in de kenmerken. Voor het probleem van de klassenruis, presenteren we drie klassificatoren gebaseerd op waarschijnlijkheidsmodellen om de werkelijke distributies van de klassen te ontdekken. Vooral onze toevoeging van een ruismodel aan de *Kernel Fisher discriminant* biedt betere performance, in het bijzonder op niet-Gaussische datasets en datasets met een relatief groot aantal kenmerken ten opzichte van het aantal voorbeelden in de leerset. Om beter om te gaan met meetruis in de kenmerken stellen we voor om voorkennis over de ruis te integrereren in *kernel-density* klassificatoren, door specifieke *kernels* te gebruiken voor individuele objecten, kenmerken en meetwaarden. De toevoeging van voorkennis blijkt vooral heilzaam voor kleine leersets.

In het tweede deel gebruiken we incomplete data over metabolische reacties en transcriptie regulatie met behulp van zowel een netwerk- als een evolutie-gebaseerde benadering. Dat betekent dat we geïntegreerde metabole netwerken en regulatie netwerken binnen één soort met elkaar te vergelijken en om deze geïntegreerde netwerken tussen verschillende soorten met elkaar te vergelijken. Door op netwerknivo te vergelijken krijgen we niet alleen een uitgebreider inzicht in het cellulaire systeem, maar kunnen we ook betrouwbaardere hypotheses genereren op basis van de initieel incomplete data. Deze geïntegreerde netwerkmethode is gebasseerd op een nieuwe uitlijningsmethodiek die het mogelijk maakt om automatisch de volledige metabole netwerken van twee soorten uit te lijnen, rekening houdend met alle mogelijke combinaties en met kleine verschillen in vergelijkbare reacties. Daarnaast presenteren we een uitgebreide en flexibele score functie voor de vergelijking van de geïntegreerde netwerken die gebruik maakt van alle relevante en ongecorreleerde informatiebronnen. Als resultaat hebben we volledig geconserveerde netwerken en hun variaties op regulatie en metabolisch niveau geïdentificieerd, nieuwe mogelijke netwerken ontdekt die in conventionele databanken niet voorkomen en hypotheses gegenereerd over ontbrekende data, gebaseerd op tegenhangers op een andere niveau en/of in een andere soort.

ACKNOWLEDGEMENTS

Six years after I graduated from university, I came back to academia. It wasn't easy, plus the huge differences between western and Chinese education styles. Many people have helped me along the way. Even though I'd like to keep my acknowledgements brief, I hope everyone who has helped me will receive my sincere gratitude.

First of all, I would like to thank three supervisors during my M.Sc. and Ph.D. study: Marcel Reinders, Dick de Ridder, and Lodewyk Wessels. Marcel supervised me for five years. His passion for science and dedication to education set a good example and often reminded me of how lucky I am. Above all, I'm grateful for his trust. It's him who always sets a high standard and pushes me to reach higher. I also learned a lot from Dick, from open-minded thinking to down-to-earth attitude. He gave me the utmost support to the utmost detail, and always provided a frank communication platform on any subject. Certainly I should also thank him for making an impressive Peking duck, which will keep shining in my memory. My interest in bioinformatics started from my M.Sc. thesis project, many thanks to Lodewyk. His collaboration with The Netherlands Cancer Institute showed me the connection between research and application. I enjoyed and learned from the discussions with these supervisors. Each debate was challenging and intense, like playing table tennis with three or sometimes four people!

Secondly, I am very lucky to have Bob Duin as my counselor in the pattern recognition part of my research. Discussions with him are always insightful and inspiring. His endeavor to understand human cognition also brought me, among others, to many interesting discussions where we can share our consciousness freely.

Next, I want to thank Marco de Groot and Jan Bot for their contribution in my network alignment papers. The brainstorm discussions with Marco were fruitful and informative. He patiently answered thousands of questions from me, and asked millions of questions back. Jan helped me to run web queries on RSAT, day and night. Somehow he has a magic stick to make things work, within deadlines! Meanwhile, I like to thank Wouter van Winden and Christof Franke for the constructive discussions and their earnest suggestions.

After almost eight years, I still like to take this chance to thank Inald Lagendijk, for his enthusiastic lectures and the small group class especially for Chinese students. The ICT group has an enjoyable learning atmosphere. You can basically knock at someone's door and get invaluable advice back. David Tax is a general practitioner, who is interested in all sorts of problems and loves to contribute. Marco de Groot and Domenico Bellomo taught me a lot about biochemistry. Carmen Lai always brings me the bright side of the world, among which is her warm hug before my M.Sc. defense. Pavel Paclík, Serguei Verzakov, Jun Wang, Bangjun Lei, and Piotr Juszczak have shared interesting thoughts with me. Thank you all! My sincere gratitude also goes to my fellows: Theo Knijnenburg, Rogier van Berlo, Marc Hulsman, Martin van Vliet, Wouter Meuleman, Christiaan Klijn, Erik van den Akker, Jeroen de Ridder, Eugéne van Someren, Peter van Nes and Fengyuan Hu. They not only helped me with my research, but also gave me wonderful accompany during miscellaneous events. To name a few, I won't forget playing table tennis with Theo, Rogier, and Domenico. Neither will I forget the raft binding and rowing competition, where Jeroen dragged Marcel down into the water...

For sure, I wouldn't have completed my study smoothly without the great support from Robbert Eggermont, Anja van den Berg, Saskia Peters, and Ben van den Boom. I also wouldn't have had so much laughters without Emile Hendriks (although I missed the opportunities to visit his enormous farm), Alan Hanjalic, Bart Kroon + Stefan Borchert + Robbert Eggermont (Christmas Game on!), Bartek Gedrojc, Wan-Jui Lee (and her delicious Chinese "zongzi"), Maarten Clements, Mark van Staalduinen, Ronald Westerlaken, Stevan Rudinac, Vikas Gupta (and his encouraging Everest climber), Berend Berendsen (with his colorful expeditions), Yan Li + Tianmu (special thanks for recording my laughters!), Feifei Huo, Zekeriya Erkin, Gjenna Stippel, Umut Naci, Hasan Celik, Richard Hendriks, Peter-Jan Doets, and Alessandro Ibba (for his brilliant sales advice)!

It's the end of an era, as Gineke ten Holt put it when we hugged goodbye. Indeed, together with Jeroen Lichtenauer, we had a memorable time during our Ph.D. studies at HB.10.300. They were always there to share my highs and lows, give me help and encouragement. I learned that Chinese and Dutch can become close friends, and cultural diversity makes our life more interesting. Especially, I would like to express my deep thanks to Jeroen's generous help on many many occasions.

Finally, I owe a lot to my dear family. My parents gave me the best guidance throughout my 21 years' education. I would not have reached this far without their encouragement and my little sister's support. Xin is a truly lucky star, who brings happiness to everyone around her. I have enjoyed very much studying with her, and the cozy family moments with her, together with Jason and later two angels, Serena and Rebecca. To Mike, I'd like to thank him by just being myself, since that is what he appreciates.

CURRICULUM VITAE

Yunlei Li was born on July 8, 1975 in Emei Shan, Sichuan, China. In 1996 she obtained her B.Sc. diploma at Zhejiang University, China. Afterwards she worked for the China National Petroleum Corporation as an instrument engineer. In 2002 she started studying Electrical Engineering at Delft University of Technology, The Netherlands. In 2004 she received her M.Sc. degree with honors, for her graduation thesis work conducted in the Information & Communication Theory Group, in collaboration with The Netherlands Cancer Institute (NKI). After graduation, she worked for AsiaInfo Holdings Inc. in China as a business intelligence engineer.

In 2005 she came to The Netherlands again and started her Ph.D. research in the same group, under supervision of Marcel Reinders and Dick de Ridder. The project was part of the BioRange programme of the Netherlands Bioinformatics Center (NBIC). The results of the research are presented in this dissertation. Since September 2009, she has been working as a business specialist at Ortec Finance.

List of publications

The following publications have resulted from Yunlei Li's Ph.D. studies:

Y. Li, J.J. Bot, M.J.T. Reinders, and D. de Ridder. A pathway approach to align regulatory-metabolic networks, *The 8th Annual International Conference on Computational Systems Bioinformatics*, Stanford University, USA, August 2009.

Y. Li, D. de Ridder, R.P.W. Duin, and M.J.T. Reinders. Integration of prior knowledge of measurement noise in kernel density classification, *Pattern Recognition*, vol. 41(1), pp. 320-330, 2008.

Y. Li, D. de Ridder, M.J.L. de Groot, and M.J.T. Reinders. Metabolic Pathway Alignment (M-Pal) reveals diversity and alternatives in conserved networks, *Series on Advances in Bioinformatics Computational Biology*, vol.6, pp. 273-285, Imperial College Press, 2008.

Y. Li, D. de Ridder, M.J.L. de Groot, and M.J.T. Reinders. M-PAS: a comprehensive and flexible metabolic pathway alignment and scoring method, The 14^{th} Annual Conference of the Advanced School for Computing and Imaging, pp. 121-128, June

2008.

Y. Li, D. de Ridder, M.J.L. de Groot, and M.J.T. Reinders. Metabolic pathway alignment between species using a comprehensive and flexible similarity measure, *BMC Systems Biology*, vol. 2, pp. 111, 2008.

Y. Li, D. de Ridder, R.P.W. Duin, and M.J.T. Reinders. An integrative kernel method dealing with diverse measurement noise in classification, The 13^{th} Annual Conference of the Advanced School for Computing and Imaging, pp. 127-134, June 2007.

Y. Li, L.F.A. Wessels, D. de Ridder, and M.J.T. Reinders. Classification in the presence of class noise using a Probabilistic Kernel Fisher method, *Pattern Recognition*, vol. 40(12), pp. 3349-3357, 2007.

Y. Li, L.F.A. Wessels, and M.J.T. Reinders. Class-noise tolerant classification based on a probabilistic noise model, The 12^{th} Annual Conference of the Advanced School for Computing and Imaging, 2006.

E.H. van Beers, T. van Welsem, L.F.A. Wessels, Y. Li, R.A. Oldenburg, P. Devilee, C.J. Cornelisse, S. Verhoef, F.B.L. Hogervorst, L.J. van't Veer, and P.M. Nederlof. Comparative genomic hybridization profiles in human BRCA1 and BRCA2 breast tumors highlight differential sets of genomic aberrations, *Cancer Research*, vol. 65, pp. 822-827, 2005.