Bayesian Markov random field analysis for integrated network-based protein function prediction

Yiannis A.I. Kourmpetis
Thesis Committee

Thesis Supervisor
Prof. dr. Cajo J.F. ter Braak
Personal chair at Biometris
Wageningen University and Research Centre

Thesis Co-supervisor
Dr. Roeland C.H.J. van Ham
Vice-President Bioinformatics and Modeling
Keygene N.V., Wageningen

Other members
Prof. dr. J. Molenaar, Wageningen University
Prof. dr. M.A.M Groenen, Wageningen University
Prof. dr. M.J.T. Reinders, Delft University of Technology
Prof. dr. D.R. Westhead, University of Leeds, UK

This research was conducted under the auspices of the Graduate School of Experimental Plant Sciences
Bayesian Markov random field analysis for integrated network-based protein function prediction

Yiannis A.I. Kourmpetis
Yiannis A.I. Kourmpetis
Bayesian Markov random field analysis for integrated network-based protein function prediction

Thesis, Wageningen University, Wageningen, NL (2011)
With references, with summaries in Dutch and English
To my lovely son Yiankos
Contents

1 General Introduction 5
  1.1 Computational protein function prediction ................. 6
  1.2 Gene Ontology ..................................... 10
  1.3 Objectives and scope .................................. 11
  1.4 Outline of the thesis .................................. 11

2 Gene function prediction and Gene Ontology 13
  2.1 Abstract .............................................. 13
  2.2 Introduction .......................................... 14
  2.3 Materials and Methods .................................. 15
  2.4 Results and discussion ................................ 17
  2.5 Conclusions ........................................... 20
  2.6 Acknowledgements ...................................... 21

3 Network based protein function prediction 23
  3.1 Abstract .............................................. 23
  3.2 Introduction .......................................... 24
  3.3 Methods .............................................. 27
    3.3.1 Markov Random Fields ........................... 27
    3.3.2 MRF-Deng ........................................ 29
    3.3.3 BMRF ............................................ 29
    3.3.4 Datasets ......................................... 30

4 Arabidopsis protein function prediction 43
  4.1 Abstract .............................................. 43
  4.2 Introduction .......................................... 43
  4.3 Materials and Methods .................................. 46
    4.3.1 Protein-Protein Interaction (PPI) network .......... 46
    4.3.2 Coexpression (COEX) Network ..................... 47
    4.3.3 Functional Domains .............................. 47
    4.3.4 Gene Ontology Annotations ....................... 47
## Contents

4.3.5 Protein function prediction ........................................... 48  
4.3.6 Performance Evaluation ........................................... 50  
4.3.7 Construction and evaluation of list with novel predictions 51  
4.3.8 Performance comparison of BMRF with other prediction  
methods ................................................................. 52  
4.4 Results ........................................................................ 53  
4.4.1 Model selection .......................................................... 53  
4.4.2 Protein function prediction for Arabidopsis .................... 54  
4.4.3 Flowering and floral organ development in Arabidopsis . 56  
4.5 Discussion .................................................................. 59

5 Gene Ontology consistent protein function prediction 65

6 General discussion .......................................................... 79  
6.1 The usage of Gene Ontology in automated protein function  
prediction ................................................................. 79  
6.2 Probabilistic protein function prediction by BMRF ............ 81  
6.3 Integration of diverse data sources .................................. 85  
6.4 BMRF and Gene Ontology consistent protein function prediction 86

Summary ........................................................................ 87

Samenvatting .................................................................. 91

References ........................................................................ 93

Acknowledgments ................................................................ 107

List of publications .......................................................... 109

Curriculum vitae ................................................................ 111
Chapter 1

General Introduction

The Food and Agriculture Organization of the United Nations estimates that one billion people are undernourished in the world. Malnutrition causes around 36 million deaths per year, either directly or indirectly through diseases. The quality of life for a large part of this underprivileged population can be critically improved through the development of new crop varieties that are productive in soil of poor quality, thrive under circumstances of drought and that are resistant to pathogens. In addition to challenges posed to world food production, many social and environmental challenges arise due to emerging climate change. The understanding of how life works at the molecular level has a key role in finding sustainable solutions for global chronic hunger and to mitigate the effects of climate change. Genes, gene products (in particular their translated proteins) and their interactions constitute the foundations of living cells. Knowledge of the biological functions of genes and their products is crucial for the efficacy of breeding programs of agricultural crops and animals and in rational design of new drugs, medical treatments and diagnostics.

Protein functions can be inferred through thorough genetic and biochemical experiments, but as targeted approach, such experiments poorly scale to large sets of genes. On the other hand, modern high-throughput experimental methods provide a less precise yet genome-wide view to different aspects of protein function. Genome sequencing, measurements of gene expression levels and identification of protein interaction maps are some examples. The availability of such diverse functional genomics datasets for species of agricultural or biomedical interest is now growing rapidly. The challenge addressed in this thesis concerns the integration of this information in order to scale up protein function inference, either solely by computational means or by providing directions for the design of targeted experiments that need to be carried out to analyze protein function.
Chapter 1. General Introduction

1.1 Computational protein function prediction

There are two main types of approaches to computational protein function prediction, referred to as the feature based and the network based approach. The first is to determine by some statistical or machine learning method common properties of proteins that perform the same function. The second type of approach is to propagate the functional information through a network of functionally associated proteins.

For the feature based approach, the features can be of diverse nature and related to protein sequence and structure, statistics derived from these or they can be derived from experimental measurements of gene expression in different conditions, cellular location, etc. Feature based approaches are divided into two types. In the first, a set of proteins performing the same function is used to fit a statistical model of the features. Any query protein can be tested whether its features fit to that model. Hidden Markov Model (HMM) (Eddy, 1998) based protein function prediction (or family classification) is a popular methodology that follows this approach. In this case the model is built using the multiple alignment of the proteins as data matrix. An HMM builds a probabilistic model that describes the transition between features of a sequence in a multiple alignment, and further the more complex dependencies between them. From the biological point of view this approach is based on the assumption that the function of a protein is determined by its secondary and tertiary structure, which in turn are both determined by the primary sequence. Proteins with the same or similar function are therefore expected to exhibit similar sequence patterns. Proteins are classified in families according to these sequence patterns (for example PFAM (Finn et al., 2010)) or domains identified by HMM. InterPro (Mulder et al., 2005a) is a resource that aggregates several databases with known feature profiles identified from protein sequences and structures. Newly identified proteins can be used to search these databases to assess whether they fit to any of the known profiles. Although profile-based approaches are powerful, they require high-quality multiple alignments of representative samples from available protein families and this is not achievable for all families. Profile collections cover only a small part of the entire functional spectrum and are mainly concerned with the enzymatic (or molecular) function of a protein. Profiles that relate sequences with biological processes are still limited, most likely because the underlying patterns cannot be identified using the current methods. In addition to profile methods, another common method of this type is the analysis of differentially expressed genes using data from microarrays or RNA-seq. Gene sets that are active under particular biological conditions (for example under drought or heat stress, or after infection by a pathogen) are likely to be involved in related biological pathways and processes.

In the second type of feature based approaches, protein features are associated with functions by developing discriminative models that separate two
1.1. Computational protein function prediction

sets of proteins through a function of their features. This involves analysis in which the protein features serve as predictors and the presence/absence of a function as the response variable. Because the response variable is binary, this task of protein function prediction can be identified as logistic regression or as classification. Artificial Neural Networks or Support Vector Machines have been used for this purpose (Barutcuoglu et al., 2006) because of their ability to deal with high dimensional problems. Microarray and RNA-seq data analysis involve building discriminative models that separate the expression profiles from genes that perform a particular function from the other genes in the experiment. Those experimental methodologies involve large sets of genes and generate statistical challenges for developing methodologies that accurately determine the association of a gene with a particular condition of the organism. Here the main challenges from the statistical point of view are feature selection for high dimensional input spaces, dealing with missing data and fitting regression models with more parameters than observations.

The network based approach to computational protein function prediction involves propagation of available functional information through a network of physically or functionally associated proteins. In such networks, nodes denote proteins and edges denote associations between them and may be based on different types of data. For instance, proteins with similar primary, secondary and tertiary structures may perform similar functions. Also, proteins that interact physically in molecular complexes, that are co-localized in cellular compartments or that exhibit correlated patterns of expression are all likely to be involved in the same biological process. The addition of diverse types of associations, defines the structure of the functional association network containing intra- and inter-species edges between proteins.

The most common method of this type is to transfer function information between proteins on the basis of similarity at the sequence or structural level. The underlying assumption is that proteins that show relatively high similarity are evolutionary related (homologs or orthologs), and are therefore likely to have retained the same or similar functions. Pairwise (Smith and Waterman, 1981; Needleman and Wunsch, 1970) or multiple (Thompson et al., 2002) dynamic programming algorithms facilitate the task of aligning and identifying similarities between protein sequences. BLAST (Altschul et al., 1997) is a fast pairwise alignment algorithm that allows the identification of similar sequences in large databases and provides the most widely used means for computational protein function prediction. When one or more such sequences are identified, the function information is manually inspected or automatically analyzed in order to select the pieces of information that should be transferred to the query sequence. Computational tools such as Blast2GO (Conesa et al., 2005) automate this procedure. A major source of errors in homology-based transfer of function information stems from paralogous relationships between
proteins. Paralogs may have high sequence similarity but their functions are different. Fully automated transfer of annotations between proteins bears the risk of propagating wrong pieces of information and percolating those errors in the sequence databases (Gilks et al., 2002). Identifying and removing those mis-annotations is a time-consuming operation. Although there are several computational methods aiming to discriminate orthology from paralogy relationships, this task is fraught with difficulties and currently an active field of research. Further, the protein universe contains several millions of proteins, and novel proteins are being identified rapidly but the fraction of these for which experimentally verified functional annotations are available is extremely small. Even for model species this fraction is small (Sharan et al., 2007). It is therefore very common that none of the homologs of a newly identified protein is functionally annotated.

Another type of evidence indicating that pairs of proteins are functionally associated is their physical interaction, their concerted expression or their interaction at the genetic level. High-throughput identification of such types of interactions can be obtained by microarray analysis, RNA sequencing, perturbation screens, affinity purification experiments and yeast two hybrid screens. The networks of associations inferred from such data can be analyzed in a similar manner as done for social or computer networks. The first approach in this case is to identify highly connected clusters in those protein networks using network clustering algorithms. Those clusters are called functional modules. Once they are identified, the function information can be transferred between the proteins belonging to the same cluster. Statistical methodology for finding overrepresented functions in the clusters has been developed for this task (Subramanian et al., 2005). This approach is a fast way to assign functions to proteins but given that a protein may be involved in one or more clusters, assigning each protein to exactly one cluster may be problematic. Fuzzy clustering (ter Braak et al., 2009) or fuzzy community detection algorithms may provide valuable tools to overcome this problem.

Alternative approaches are available that do not require the clustering step to pass function information through the network topology. A primitive method of this type is guilt by association, in which the annotation of a protein is based solely on its direct neighbors in the network. The weak point of this approach is that it is not possible to directly determine which piece of functional information should be transferred between a pair of proteins. There are several factors that introduce uncertainty. In particular, proteins may be associated only in a specific functional context (see Fig 1.1A for an illustration). It is common for most proteins to have multiple distinct roles in the cell with each one of these being performed as transient interactions with different sets of proteins. Hence, association between a pair of proteins does not automatically imply that they share all of their biological roles. Protein interactions maybe stable or tran-
1.1. Computational protein function prediction

Figure 1.1: Examples of questions that arise in protein function annotation. A. AGL15, AGL1 and AGL5 are *A. thaliana* proteins. AGL1 is known to be functionally involved in the processes of fruit dihescence and carpel development. AGL5 is known to be involved in carpel development and in the regulation of transcription. What can be inferred about the function of AGL15, given that it is associated with the other two proteins by protein interaction? B. Can we infer the function of AGL71, given that we know the function of AGL2?

sient. In the first case the proteins may have different molecular functions but they are all involved in the same set of biological processes. However, many of the protein interactions in the cell are transient. Such pairs of proteins perform a particular biological task together, but they have additional distinct biological roles. This property of protein interactions puts the task of the propagation of functional information in a probabilistic context for which proper statistical models need to be developed. Another major challenge in function prediction through network propagation concerns the sparsity of the available functional information. Even for the best-studied model species, 10-75% of the proteins lacks functional annotation (Sharan et al., 2007). Therefore, it is very likely that some or even all of the network neighbors of an uncharacterized protein, will be uncharacterized as well (Fig 1.1B).

Statistical methods offer ground to deal with the probabilistic nature of the above mentioned challenges. Here, an edge between two proteins in a network represents a probability that they share a function. Statistical models that quantify this type of relationships have been developed, in particular methods for spatial analysis developed for computer vision (or image analysis) or for modeling of geospatial processes. Markov Random Fields (MRF) provide a generic probabilistic framework to model the joint distribution of dependent
random variables. They are used in computer vision, such as the restoration of missing parts of images where the color information is propagated through the grid of the neighboring pixels in the image (Geman and Geman, 1984). In analogy, for protein function prediction, the function information is transferred through the network which can be seen as a binary image. The main qualitative difference is the complex structure of the function association network, compared to the regular square or rectangular grid in the image in computer vision. Protein networks are very similar to those of social or computer networks, having statistical properties such as the presence of hubs, small world properties and power-law distributions of the connectivity degrees, while in image analysis, the nodes represent pixels with constant number of neighbors (for pixels that do not lie in the boundary of the image). The framework of using binary MRF analysis for protein function prediction was originally proposed by Deng et al. (2004). MRF methods allow the prediction of functions for proteins that have neighbors that are not functionally characterized. In this approach the probability that a protein performs a particular function depends on two parameters, i.e: i) the number of direct neighbors in the network that perform the function, and; ii) the number of direct neighbors that do not perform the function. The parameters of this conditional distribution are estimated from the set of proteins with known function by fitting a logistic regression model. Function prediction of the unannotated proteins is subsequently done by posterior simulation i.e by Gibbs sampling, that is sampling the state of each uncharacterized protein sequentially given the state of all the rest (or its neighbors due to the Markov property). The posterior probabilities are then interpreted as level of confidence for each predicted function annotation. The advances in high throughput experimental methods allow the generation of different data types for large numbers of proteins that can be used either to connect the protein directly to a function or to relate it to other proteins that are functionally associated. Usage of those two approaches in a unified framework is a key challenge for protein function prediction since it enables the effective usage of the available data so as to improve of function prediction accuracy.

1.2 Gene Ontology

Predicting protein functions by computational methods assumes that those functions can be defined. The Gene Ontology (GO) (Ashburner et al., 2000) is a widely used controlled vocabulary for the description of functions of genes, gene products and their locations. GO terms are separated in three distinct branches: Molecular Function (MF), Biological Process (BP) and Cellular Component (CC). In each class the terms are arranged in a Directed Acyclic Graph. Due to this hierarchical structure, a protein that is assigned to a par-
1.3 Objectives and scope

The aim of this thesis is to develop computational methodology for protein function prediction aimed at practical application by the biological community. The methodology should provide accurate predictions, should be able to integrate heterogeneous information and should be applicable to large datasets with manageable computational effort. The key questions addressed in this thesis are:

- Q1: How is the Gene Ontology used in practice in the annotation of different (model) species?
- Q2: What types of statistical methodologies are most appropriate for dealing with the protein function prediction challenge?
- Q3: How can we deal with the function annotation of proteins that are linked to proteins that are themselves functionally uncharacterized?
- Q4: Which is an appropriate model for the integration of different types of data?
- Q5: How can we deal with uncertainty in the model parameters?
- Q6: How can we make a method that provides biologically interpretable predictions?

1.4 Outline of the thesis

The key questions are addressed in the different chapters of this thesis. Chapter two studies the GO annotations for the proteomes of seven model species (S. cerevisiae, C. elegans, Drosophila melanogaster, H. sapiens, M. musculus, A. thaliana and O. sativa) (Q1). In most genomes large parts of the proteins are multi-functional and are annotated using multiple and hierarchically independent GO terms. Therefore a protein function prediction method should be able to predict multiple GO terms (Q2). Also, it was shown that the annotation strategy used for the different model species varies and that this affects the resulting annotation. In Chapter three an MRF-based method is developed.
that solves the problem of extending the propagation of function information beyond directly associated proteins (Q3) and estimates the model parameters accurately by performing simultaneous estimation of the parameters and protein function prediction (Q5). The method is based on an adaptive Markov Chain Monte Carlo algorithm, named Differential Evolution Markov Chain (Ter Braak, 2006; Ter Braak and Vrugt, 2008). The methodology is tested against the original MRF and two alternative methods using yeast data. The results of the comparison showed that the performance of our method is competitive in the current state of art, and importantly, that it is applicable to larger datasets than alternative methods. Novel functions for a significant number of yeast proteins were predicted. Chapter four studies approaches for the integration of diverse data sources (Q4). The method is applied to sequence, protein-protein interaction, and co-expression data from A. thaliana and provided predictions for a large fraction of proteins that hitherto lacked functional characterization. A number of predictions are validated by experimental studies (Q4). Chapter five develops a method that takes as input the probabilistic outputs per GO term and then imposes the GO DAG consistency to the final function inference. This method makes predictions more interpretable, but also improves the performance of the method in terms of precision and recall (Q6). Chapter six summarizes and discusses the results from the studies described in this thesis and provides an outlook for further research.
Chapter 2

Gene function prediction and genome annotation using the Gene Ontology\textsuperscript{1}

2.1 Abstract

The Gene Ontology (GO) is a widely used controlled vocabulary for the description of gene function. In this study we quantify the usage of multiple and hierarchically independent GO terms in the curated genome annotations of seven well-studied species. In most genomes, significant proportions (6 - 60\%) of genes have been annotated with multiple and hierarchically independent terms. This may be necessary to attain adequate specificity of description. One noticeable exception is Arabidopsis thaliana, in which genes are much less frequently annotated with multiple terms (6 - 14\%). In contrast, an analysis of the occurrence of InterPro hits in the proteomes of the seven species, followed by a mapping of the hits to GO terms, did not reveal an aberrant pattern for the A. thaliana genome.

This study shows the widespread usage of multiple hierarchically independent GO terms in the functional annotation of genes. By consequence, probabilistic methods that aim to predict gene function automatically through integration of diverse genomic datasets, and that employ the GO, must be able to predict such multiple terms.

We attribute the low frequency with which multiple GO terms are used

\textsuperscript{1}Published as: Kourmpetis YAI., van der Burgt A, Bink MCAM, ter Braak, C.J.F. and van Ham, R.C.H.J. The use of multiple hierarchically independent Gene Ontology terms in gene function prediction and genome annotation. In silico biology 2007 7(6): 575-582
in Arabidopsis to deviating practices in the genome annotation and curation process between communities of annotators. This may bias genome-scale comparisons of gene function between different species. GO term assignment should therefore be performed according to strictly similar rules and standards.

2.2 Introduction

The Gene Ontology (GO) (Ashburner et al., 2000) provides a controlled vocabulary for the description of gene and gene product attributes in any species. It uses three key domains that provide descriptions of molecular function, biological process and cellular component. Common applications of GO include the functional annotation of genes predicted from whole genome sequences, the functional annotation and comparison of genes assayed in microarray experiments and the analysis of cellular pathways. Our main interest in GO lies in its application as a classification scheme in automated, probabilistic methods for gene function prediction (Barutcuoglu et al., 2006; Eisner et al., 2005; Pavlidis et al., 2002; Troyanskaya et al., 2003). Such methods usually employ powerful statistical methods, but they have one built-in restriction: they perform classifications in which terms from at most one functional class of the GO hierarchy are predicted. Recently, the subject of multi-label classification in gene function prediction was addressed in a number of studies. In some of them (Clare and King, 2003; Vens et al., 2008; Roth and Fischer, 2007; Zhang and Zhou, 2007), a single-label classification algorithm was extended for multi-label purposes, but it did not employ the GO. A second study (Barutcuoglu et al., 2006) aimed to improve the performance and consistency of gene function prediction by ensuring the True Path Rule, but it could not predict classes that are hierarchically independent. This strongly contrasts with the observation made in this note that genes in the genomes of well-studied model species have often been community-annotated using multiple independent GO terms. The following example illustrates why in principle, it is most relevant for methods of automated gene function prediction to be able to predict multiple independent GO terms.

In the GO datasets 2, the genes YDL029W and YHR107C from *Saccharomyces cerevisiae* are both annotated with three hierarchically independent terms in the molecular function domain: YDL029W is annotated with “ATP binding” (GO: 0005524), “actin binding” (GO: 0003779) and “structural constituent cytoskeleton” (GO: 0005200) and gene YHR107C is annotated with the terms “GTPase activity” (GO:0003924), “phosphatidylinositol binding” (GO:0005545) and “structural constituent cytoskeleton” (GO: 0005200). Both genes contribute to the structural integrity of the cytoskeleton, yet they have

---

Table 2.1: Species and number of annotated genes analyzed. Names of the species and number of annotated genes per branch of the Gene Ontology that were used for study. The last column refers to the revision number of the respective gene.association.species files provided by the Gene ontology project. BP, MF and CC denote Biological Process, Molecular Function and Cellular Component respectively.

<table>
<thead>
<tr>
<th>Species name</th>
<th>BP</th>
<th>MF</th>
<th>CC</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>6473</td>
<td>6473</td>
<td>6473</td>
<td>1.337</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>9555</td>
<td>10996</td>
<td>6256</td>
<td>1.83</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>10334</td>
<td>10469</td>
<td>7739</td>
<td>1.95</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>24940</td>
<td>29639</td>
<td>22823</td>
<td>1.48</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>15845</td>
<td>17341</td>
<td>16059</td>
<td>1.664</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>27937</td>
<td>30577</td>
<td>28869</td>
<td>1.1156</td>
</tr>
<tr>
<td>Oryza sativa</td>
<td>13994</td>
<td>12669</td>
<td>46942</td>
<td>1.33</td>
</tr>
</tbody>
</table>

different molecular functions. It is evident that a single GO term, irrespective of its specificity, will often not suffice to describe the function of a protein in a complete and unique way.

In this study, we investigate and quantify the incidence with which genes from well-studied species (Table 2.1) have been annotated using multiple independent GO terms. We emphasize the importance, not only of developing new methods for automated gene function prediction based on multi-label classification techniques, but also of establishing transparent and standard procedures for GO-term assignment in the curation of gene function annotations.

## 2.3 Materials and Methods

We studied patterns of GO term assignment in the community-based, manually curated annotations of seven well-studied species (Table 2.1) by analyzing the frequency of multiple term usage in the annotation files provided by the GO project (gene.association.species files; \(^3\)). Hierarchical relationships between the terms were checked using the version 1.12 of the GO DAG \(^4\). The primary data files were divided according to the three ontology domains. Multiple annotations of a gene with the same term but with use of different evidence codes were counted as single-term annotations. We counted all genes that were annotated using multiple, hierarchically independent GO terms. This was done by examining the hierarchical relationships between all pairs of annotations.

---

\(^3\)http://www.geneontology.org/GO.current.annotations.shtml

\(^4\)http://www.geneontology.org/GO.downloads.ontology.shtml
for a gene. In the case two terms were hierarchically related (parent-child relationship), the annotation was counted as single-term at the deeper node in the GO DAG (see Figure 2.1 for an illustration).

Figure 2.1: **Example of a gene with multiple annotations.** The HSL1 gene from Saccharomyces is annotated using five terms from the Biological Process Ontology. Node GO:000074 is the parent of nodes GO:000078 and GO:000135. Thus, four of the five GO terms from the root are hierarchically unrelated, i.e. GO:000078, GO:000135, GO:000086, GO:0006468.

We compared the curated GO project annotations with those that can be automatically derived from the Integr8 database (Kersey et al., 2005). The Proteome Analysis section of this database provides protein annotations based on InterPro hits to protein families, domains and functional sites. The precomputed files for the species listed in Table 1 were downloaded from Integr8 and InterPro hits were mapped to GO terms using the interpro2go map 5, available from the GO website. *Oryza sativa* was excluded from this study because it was not available at Integr8 at the time of analysis. Since InterPro entries may correspond to terms from different domains of the GO, the mapped files were again divided according to the ontology. The counting of multiple GO term usage based on InterPro domains was subsequently performed as described for the GO project annotations.

We also investigated the frequency distributions of evidence codes used in GO term assignment for each of the species. The differences in the use of evidence codes between species and between single and multiple annotated genes were analyzed by log-ratio Principal Component Analysis for compositional data (Aitchison and Greenacre, 2002). In this analysis we excluded the ND evidence code (designation for “No biological data available”), as it cannot, by definition, result in multiple annotations for a gene.

5 Mappings of External Classification Systems to GO: http://www.geneontology.org/GO.indices.shtml
2.4 Results and discussion

Analysis of the community-based, curated gene annotations of seven well-studied species, as provided by the GO project, show that significant propor-

Figure 2.2: Multiple annotated genes in seven species. Percentages of genes with multiple hierarchically independent GO terms calculated: (A) using the annotation files provided by the GO project, including all evidence codes; (B) using the same annotation files, but excluding the ND evidence code (“no biological data available”), and; (C) using the InterPro hits as provided by the Integr8 database. Integr8 did not contain data for *O. sativa* at the time of study. The percentage for the Biological Process Ontology for *M. musculus* is missing due to a technical problem. Species designations are as in Table 1.
tions of the genes are annotated with multiple GO terms (Fig. 2.2A). Proportions range between 6% and 60% among the species and the different domains of the ontology. A large proportion of genes are annotated with three to six independent GO terms and for a considerable number of genes this number exceeds ten terms (Fig. 2.3). An extreme example is the gene Notch from *Drosophila melanogaster* (Flybase ID: Fbg0004647) which is annotated with 52 independent terms from the Biological Process Ontology.

It is commonly accepted that human-curated annotations using the GO provide the as yet most reliable and standardized functional descriptions of genes. For many genes, this involves the use of multiple, hierarchically independent GO terms. Multiple annotations may be required either to describe a single function, process or cellular component as completely and uniquely as possible, or to describe the multiple functions, processes or cellular components in which a single gene product can take part. Most classification methods currently used in automated gene function prediction are not designed to assign multiple GO terms. This implies that such methods will produce incomplete and low-quality functional annotations for many genes.

A wholly unexpected and surprising observation in the comparison of species is the much lower proportion of genes with multiple GO terms in *Arabidopsis thaliana*. Further analysis showed that a relatively large number of genes in the *Arabidopsis* annotation have been assigned the term “unknown” (36.1%, 42% and 35.8% of the annotations in the molecular function, biological process, and cellular component domains, respectively). The term “unknown” is used at the root node of each domain of the GO. In our analysis these nodes were treated as single term annotations. Because this might have biased the comparison, we rescaled the proportions for all the species, excluding genes with the ND evidence code for “No biological data available”. The resulting histograms are given in Figure 2.2B and persistently show lower proportions of multiple annotations in *Arabidopsis*. To exclude further the possibility that the deviant pattern in *Arabidopsis* is an artifact resulting from usage of erroneous files in the GO repository, we analyzed two older versions of *Arabidopsis* annotation files (versions 1.949 and 1.959 from March and April 2006, respectively). Both analyzes showed similar and much lower percentages of multiple annotated genes in Arabidopsis compared to the other species (data not shown).

To investigate whether the aberrant proportions for *Arabidopsis* can be explained biologically, we performed an independent re-annotation of the gene sets from each species on the basis of their hit lists to InterPro accessions (Mulder et al., 2005b). If the *Arabidopsis* proteome would be biologically different from the other species, as suggested by the GO project annotations (Fig. 2.2A), an InterPro-based GO annotation can be expected to reveal a similar aberration in the proportion of genes with multiple functional signatures. The
2.4. Results and discussion

Figure 2.3: Proportions of genes annotated with one or more hierarchically independent GO terms used. Each histogram represents the relative composition of genes annotated with the numbers of GO terms indicated.

results presented in Figure 2.2C do not reveal such a difference. It seems obvious therefore that the aberrant proportion of multiple annotations in the GO project must be explained by the different usage of rules for GO term assignment in the annotation process for Arabidopsis.

Because a formal comparison between the annotation strategies for the different species could not be conducted directly, we compared their distributions of usage of evidence codes for both single and multiple annotated genes by log-ratio principal component analysis. The resulting biplot (Fig. 2.4) shows a widely deviating pattern of evidence code usage. The differences between species appear to be much larger than the differences between single annotated and multiple annotated genes within species. Arabidopsis does not show any particular deviating pattern relative to the other species and usage of evidence codes appears to be independent of whether genes are annotated with single or multiple hierarchically independent GO terms. Saccharomyces appears to be most different in this analysis, which can be attributed to a higher frequency of codes reserved for experimental evidence (IGI, IPI and IMP, see legend for abbreviations). These results further illustrates that systematic differences underlie practices of functional annotation by the annotators and curators of the various species.
Figure 2.4: Biplot of evidence code distributions used for single and multiple annotated genes. Each point lies originally in a thirteen dimensional space where each coordinate corresponds to the frequency of use of one evidence code. This two-dimensional representation was achieved by performing a special form of PCA for compositional data [11]. A polygon is drawn to cluster the points for each species. Evidence codes: IC: Inferred by Curator; IDA: Inferred from Direct Assay; IEA: Inferred from Electronic Annotation; IEP: Inferred from Expression Pattern; IGC: Inferred from Genomic Context; IGI: Inferred from Genetic Interaction; IMP: Inferred from Mutant Phenotype; IPI: Inferred from Physical Interaction; ISS: Inferred from Sequence or Structural Similarity; NAS: Non-traceable Author Statement; RCA: inferred from Reviewed Computational Analysis; TAS: Traceable Author Statement; NR: Not Recorded.

2.5 Conclusions

Assigning function to a gene is an important but also complex operation. Controlled vocabularies like GO provide an excellent infrastructure for the functional description of genes and gene products. As part of our effort to develop statistical methodology for gene function prediction based on data integration and multi-label classification, we have studied GO term usage in genome annotations. We find that the availability of vocabularies alone does not guarantee
that annotators will employ these consistently and in exactly the same manner for each species. This study reveals that the practice of GO term assignment differs considerably between communities of annotators. This will bias genome-scale comparisons of gene function between different species. The quality and comparability of functional annotations will therefore benefit, not only from controlled vocabularies such as the GO, but also from strict application of formal rules for the assignment of GO terms. First and foremost such rules must specify what criteria must be met in order for a GO term to be included in an annotation, irrespective of what evidence code is used for that GO term assignment. A proposal for such rules has been put forward by Clare et al. (2006). Finally, the observation of extensive usage of multiple, independent terms to describe a gene function underlines the importance of using multi-label classification methods in the development and application of methods for automated gene function prediction.

2.6 Acknowledgements

This work is part of the BioRange program of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI).
Chapter 3

Network based protein function prediction

3.1 Abstract

Inference of protein functions is one of the most important aims of modern biology. To fully exploit the large volumes of genomic data typically produced in modern-day genomic experiments, automated computational methods for protein function prediction are urgently needed. Established methods use sequence or structure similarity to infer functions but those types of data do not suffice to determine the biological context in which proteins act. Current high-throughput biological experiments produce large amounts of data on the interactions between proteins. Such data can be used to infer interaction networks and to predict the biological process that the protein is involved in.

Here, we develop a probabilistic approach for protein function prediction using network data, such as protein-protein interaction measurements. We use an adaptive Markov Chain Monte Carlo algorithm that leads to more accurate parameter estimates and consequently to improved prediction performance compared to the standard Markov Random Fields method. We tested our method using a high quality *S. cerevisiae* validation network with 1622 proteins against 90 Gene Ontology terms of different levels of abstraction. Compared to three other protein function prediction methods, our approach shows very good prediction performance.

\[1\] Published as: Kourmpetis YAI, van Dijk ADJ, Bink MCAM, van Ham RCHJ, ter Braak CJF (2010) Bayesian Markov Random Field Analysis for Protein Function Prediction Based on Network Data. PLoS ONE 5(2): e9293

23
Our method can be directly applied to protein-protein interaction or coexpression networks, but also can be extended to use multiple data sources. We apply our method to physical protein interaction data from *S. cerevisiae* and provide novel predictions, using 340 Gene Ontology terms, for 1170 unannotated proteins and we evaluate the predictions using the available literature.

### 3.2 Introduction

Functional annotation of proteins is an important goal in post-genomics research. However, despite the many recent technological advances that have allowed the production of various types of molecular data at a genome-wide scale, the function of large numbers of proteins in fully sequenced genomes still remains unknown. This is true even for six of the most-studied model species, in which the proportion of unannotated proteins varies between 10% and 75% (Sharan et al., 2007). The general problem is that on the one hand, large-scale experimental approaches give only indirect information about the function of proteins, whereas on the other hand small-scale experiments provide more direct evidence but are labor intensive. The development of accurate computational methods for protein function prediction can therefore aid in reducing the gap between the speed of whole-genome sequencing and the functional annotation of their encoded proteomes.

The most common approach in computational prediction of protein function is to use sequence or structure similarity to transfer functional information between proteins (Punta and Ofran, 2008). Blast (Altschul et al., 1997) and InterPro (Mulder et al., 2005a) searches are popular methods for such predictions. However, sequence similarity does not necessarily imply functional equivalence and thus Blast based annotation transfers can be erroneous e.g. proteins from gene duplication may have high sequence similarity but different functions. Also, homology based annotation transfers lead to the percolation of misannotations in databases. Furthermore, sequence data do not provide information on the biological context of protein functions, *e.g.* the metabolic pathway or biological process that the protein is involved in. Such contextual information can be derived from large-scale data on interactions (*i.e.* physical, genetic, co-expression) between genes or gene-products, such as proteins. These data are commonly represented as networks, with nodes representing proteins and edges representing the detected interactions (Figure 3.1).

In a review of the existing computational methods that exploit network data for function prediction, Sharan et al. (2007) distinguished direct and indirect methods. Direct methods predict the function of a protein from the known functions of its neighbors (the proteins it interacts with) (Letovsky and Kasif, 2003; Vazquez et al., 2003; Deng et al., 2003; Karaoz et al., 2004; Mostafavi et al., 2008). Indirect methods first identify functional modules in the network
and subsequently assign overrepresented (enriched) functions in the module to their unannotated components (Enright et al., 2002; Bader and Hogue, 2003; Ulitsky et al., 2008). Sharan et al. (2007) judged the direct methods as slightly superior to the indirect ones.

A pioneering direct method is the binary Markov Random Fields (MRF) method proposed by Deng et al. (2003) (hereafter referred to as “MRF-Deng”). In MRF-Deng, the probability that a protein performs a particular function depends on two numbers, namely the number of its direct neighbors in the network that perform the function and the number of those that do not. The parameters of this relationship are learned from a training set by logistic regression (McCullagh and Nelder, 1989) using these numbers as predictors. Then, Gibbs sampling is employed for functional inference of the proteins with unknown function (“unannotated proteins”). Letovsky and Kasif (LK) (Letovsky and Kasif, 2003) developed an approach that is similar to MRF-Deng, but with another parameter estimation method and with Gibbs sampling replaced by belief propagation for the prediction step. GeneMania (Mostafavi et al., 2008) is based on a Gaussian (instead of a binary) MRF and leads to a relatively easy to solve quadratic program for making predictions.

Lanckriet et al. (2004) proposed an approach based on Support Vector Machines (SVM). In this approach, a similarity kernel between the proteins is computed and then a classifier is built by maximizing the margin between the proteins that perform a particular function and those that do not. The authors showed that the SVM approach leads to improved performance compared to MRF-Deng. One extension of this method is the Multi-Label Hierarchical Classification method (MLHC) (Barutcuoglu et al., 2006; Guan et al., 2008) where predictions are first made by SVM, independently per Gene Ontology (GO) (Ashburner et al., 2000) term, which are then made consistent with the GO hierarchy by using a Bayesian Network.

Lee et al. (2006) combined the appealing properties of MRF and SVM methods into Kernel Logistic Regression (KLR). Whereas the predictors in MRF-Deng are derived from the adjacency matrix that represents the network, they are derived from a similarity kernel in KLR. Parameter estimation and predictions are made by logistic regression instead of by SVM, because logistic regression is much faster. Lee et al. used a diffusion kernel (Kondor and Lafferty, 2002), whereby the protein neighborhoods are expanded or pruned depending on the diffusion parameter, and showed that diffusion based KLR outperforms MRF-Deng and performs comparably to diffusion kernel based SVM. In the recent experiment of (Peña Castillo et al., 2008), several state of art methods were assessed using Mus musculus genomic datasets (Llewellyn and Eisenberg, 2008) leading to the conclusion that GeneMania, MLHC and KLR showed appealing performance.

The application of diffusion kernel based KLR or SVM to large networks
is difficult or even impossible because of the huge computational cost of the required matrix exponentiation. In this paper we therefore try to improve the original MRF-Deng method without introduction of diffusion kernels.

We discovered an important potential problem with MRF-Deng. The parameter estimation step of MRF-Deng is problematic in that proteins with known function (“annotated proteins”) have unannotated proteins as neighbors so that the predictors used in the logistic regression carry uncertainty due to the unannotated proteins (Figure 3.1). This problem increases with increasing numbers of unannotated proteins. MRF-Deng neglects this problem by disregarding the unannotated proteins in the first step. By this strategy, the neighborhood counts of a large number of proteins are reduced and therefore the parameter estimates tend to take larger absolute values (McCullagh and Nelder, 1989). During the Gibbs sampling, the unannotated proteins are taken into account, but the model parameters are those estimated from the pruned neighborhoods.

Here we amend the MRF-Deng method, by performing joint parameter estimation and prediction (Figure 3.1) as suggested by (Lee et al., 2006; Wei and Li, 2007) i.e. in a way that the computational cost is still modest compared to diffusion kernel based KLR. Joint analysis is a standard approach to deal with missing data in the context of semi-supervised learning and can be performed by iteratively estimating the parameters by maximizing the PseudoLikelihood Function (PLF) using logistic regression as a first step and estimating the unknown function by optimizing the objective function of the MRF in the second step, till convergence is met (Besag, 1986). If there are many unannotated proteins in a given dataset then there are so many unknowns (in the second step), that optimizing them leads to a loss of statistical consistency in parameter estimation. In such cases it is much better to allow for the uncertainty therein and “average across” the unknowns (MacKay, 2002). We do so by taking a Bayesian approach. We model the joint posterior distribution of the model parameters and the functional states of the unannotated proteins and sample from this joint distribution by a Markov Chain Monte Carlo (MCMC) algorithm (Figure 3.1). We name the new method Bayesian Markov Random Field analysis (BMRF) and evaluate its performance under severe conditions, i.e. when half of the proteins in a network is unannotated. We show that BMRF outperforms MRF-Deng, and is competitive to diffusion KLR. Using a high quality protein-protein interaction data set of (Collins et al., 2007) we provide functional predictions for 1170 unannotated S. cerevisiae proteins in terms of 340 nodes (“GO terms”) of the biological process ontology of The Gene Ontology Consortium (Ashburner et al., 2000) and we evaluate a subset of these predictions using available literature.
3.3 Methods

3.3.1 Markov Random Fields

MRF methods provide the framework for probabilistic modeling of dependent random variables. They are widely applied to a variety of problems with spatial dependencies, such as image analysis (Geman and Geman, 1984), where a picture is considered as a square grid of pixels (i.e. an undirected graph) and each pixel corresponds to a variable whose value (i.e. color) depends on the values of its neighborhood pixels. In image restoration problems, MRF methods are used to restore the missing parts of the images. The most probable coloring configurations of the missing pixels can be inferred from the full joint probability distribution. The colors of the missing pixels thereby are predicted simultaneously, allowing prediction in cases where the entire neighborhoods of pixels have to be predicted. MRF is thus particularly suited for a guilt-by-association approach.

The framework for protein function prediction based on MRF was originally proposed by (Deng et al., 2003). Given a set of N proteins and a set E of pair-wise interactions, we construct a network where nodes represent proteins and edges represent the interactions between them. Next each node is colored depending on whether the corresponding protein performs or does not perform a particular function (e.g. one GO term), where the coloring nodes of unannotated proteins remains unknown (Figure 3.1). The coloring is encoded in an N-dimensional binary vector x, i.e. \( x_i = 1 \) if the \( i^{th} \) protein performs a particular function, \( x_i = 0 \), if it does not. Our aim is to assign each unannotated protein to one of the two possible states. In fact, this problem is similar to the image restoration problem described above. The MRF model entails that the probability of state x of the network given a vector \( \theta \) of model parameters (discussed below) is

\[
P(x|\theta) = \frac{1}{Z(\theta)} \exp(U(x, \theta)),
\]

where \(-U\) is known as the energy function and \(Z(\theta)\) is a normalizing constant that depends on \(\theta\). In a homogeneous second order MRF, \(U\) can be written as (Besag, 1986; Sharan et al., 2007)

\[
U(x, \theta) = \sum_{i=1}^{N} G_1(x_i) + \sum_{i=1}^{N} \sum_{j=i+1}^{N} G_2(x_i, x_j),
\]

where \(G_1\) and \(G_2\) are problem-dependent functions. \(G_1\) takes one value per state, without considering the interactions of the protein, i.e. \(G_1(1) = \alpha\) and \(G_1(0) = 0\). The function \(G_2\) is equal to zero if proteins \(i\) and \(j\) do not interact. For interacting proteins Deng et. al. (2003) used three classes of
interactions. If both of the interacting proteins perform the function of interest then \( G_2(1,1) = \beta^{11} \). If only one of them performs the function then then \( G_2(1,0) = G_2(0,1) = \beta^{10} \), and when none of them performs the function \( G_2(0,0) = \beta^{00} \). We denote the number of protein pairs in these three classes by \( N_{11}, N_{10} \) and \( N_{00} \), respectively. The energy function of this MRF is then

\[
\alpha \sum_{i=1}^{N} x_i + \beta^{11} N_{11} + \beta^{10} N_{00} + \beta^{00} N_{00},
\]

which can be rewritten in terms of the elements of \( x \) as

\[
U(x, \theta) = \alpha \sum_{i=1}^{N} x_i + \beta^{11} \sum_{(i,j) \in E} x_i x_j + \beta^{10} \sum_{(i,j) \in E} [x_i(1 - x_j)] + \beta^{00} \sum_{(i,j) \in E} (1 - x_i)(1 - x_j),
\]

with \( \theta = (\alpha, \beta^{11}, \beta^{10}, \beta^{00}) \). We now compare two ways of coloring the network that differ only in the value of the \( i^{th} \) protein. By inserting equation (3.2) in (3.1) and setting \( \beta^1 = (\beta^{11} - \beta^{10}) \) and \( \beta^0 = (\beta^{10} - \beta^{00}) \), the log-odds (the logarithm of ratio of their probabilities) can be shown to be:

\[
\log \frac{P(x_i = 1 | x_{-i}, \alpha, \beta^1, \beta^0)}{P(x_i = 0 | x_{-i}, \alpha, \beta^1, \beta^0)} = \alpha + \beta^1 \sum_{j \in S_i} x_j + \beta^0 \sum_{j \in S_i} (1 - x_j)
\]

where \( x_{-i} \) denotes \( x \) without the \( i^{th} \) element and \( S_i \) the set of proteins that interact with protein \( i \). This equation is known from logistic regression. It has two predictors \( M_{i1} \) and \( M_{i0} \) counting the number of neighboring proteins of protein \( i \) that do and do not perform the function, respectively, and three unknown parameters, whereas the function \( U \) had four parameters. This is no surprise when noting that one parameter in \( U \) is redundant, because the sum of \( N_{11}, N_{10} \) and \( N_{00} \) is a constant that is independent of \( x \). When the right-hand side of the logistic equation is a known value \( v_i \), the conditional probability that unannotated protein \( i \) performs the function is given by the logistic function \((1 + \exp(-v_i))^{-1}\). In this way we can sample the state of each unannotated protein when we know the parameters and the states of its neighbors. The problem that some or all neighbors have an unknown state can be circumvented by repeated sampling of states, starting from an initial configuration, until convergence. This process is called Gibbs sampling (Geman and Geman, 1984) and is performed across all unannotated proteins. Finally, the PseudoLikelihood Function (PLF) is the product of the conditional probabilities across nodes (Li, 1995).
3.3. Methods

\[ PLF(x|\alpha, \beta^1, \beta^0) = \prod_{i=1}^{N} P(x_i|x_{-i}, \alpha, \beta^1, \beta^0). \]

3.3.2 MRF-Deng

MRF-Deng (Deng et al., 2003) consists of two tasks. In the first task, the parameters are estimated by maximizing the PLF (Li, 1995). This can be achieved by logistic regression, in which each protein is a statistical unit, the response variable is the value of \( x_i \) and two predictors are the numbers of neighbors of protein that do and do not perform the function. Unannotated proteins give rise to units with missing response (which are simply deleted from the regression) and to uncertain values of predictors for neighboring units (Figure 3.1). Thus, the two predictors cannot be precisely calculated when the neighborhood of a protein contains unannotated proteins. Consequently, the logistic regression can no longer be carried out. The authors overcame this problem by simply ignoring the unannotated proteins. In the second task, MRF-Deng makes functional inferences by Gibbs sampling across all unannotated proteins, as described above.

In summary, MRF-Deng disregards the neighborhood uncertainty in the parameter estimation step, but takes it into account during the labeling step. By disregarding unannotated proteins in the first task, neighborhoods are pruned compared to the full network. We expected that this strategy will work worse as the proportion of unannotated proteins in the network is large.

3.3.3 BMRF

In this study we develop a Bayesian strategy and draw from the joint posterior density of \( x, \alpha, \beta^0, \beta^1 \) using an MCMC algorithm and starting from an initial configuration. As in (Deng et al., 2003), we will use the PLF rather than the full likelihood, as the latter has an intractable normalizing constant. A uniform prior is used as a joint prior distribution of the model parameters. The outline of our method is given in Figure 3.1. It is Gibbs sampling in which, at iteration, \( t \), the elements of \( x^{(t)} \) corresponding to unannotated proteins are updated conditionally on the values of the parameters \( \alpha, \beta^0, \beta^1 \), as described above, and the parameters are updated conditionally on \( x^{(t)} \). The parameter update uses the adaptive MCMC algorithm called the Differential Evolution Markov Chain (DEMC) (Ter Braak and Vrugt, 2008) as follows. A candidate point \( \theta^* = (\alpha^*, \beta^{0*}, \beta^{1*}) \) is obtained using the equation:

\[ \theta^* = \theta + \gamma(Z_{R1} - Z_{R2}) + e, \]

where \( \theta \) denotes the current state of the parameter vector, \( \gamma \sim U(\gamma^*/2, \gamma^*) \) is the scaling parameter and \( \gamma^* = \frac{2.38}{\sqrt{d}} \) is the optimal step size (Ter Braak, 2006),
where \( d \) is the parameter dimension. In our problem, \( d = 3 \) and therefore \( \gamma^* = 0.97 \). \( Z_{R1}, Z_{R2} \) are uniformly selected from past samples of the Markov Chain as stored in a matrix \( Z \) and \( e \sim MVN(0, 10^{-4}) \). \( \theta^* \) is accepted using a Metropolis step, with probability:

\[
r = \min \left( 1, \frac{PLF(x^{(t)} | \theta^*)}{PLF(x^{(t)} | \theta)} \right).
\]

The labelling vector \( x \) is initialized using the output of the MRF-Deng. The \( Z \) matrix is initialized in the following way. First, the Maximum Penalized Pseudolikelihood Estimates of \( \theta, \hat{\mu} \) and \( \hat{\Sigma} \) are obtained by logistic regression. We used the penalization to reduce the bias of the parameter estimates due to the small number of positive examples in the specific GO terms. Those parameter estimates were obtained using the brglm R package (Kosmidis, 2007). Then \( m = 10d \) parameter values are sampled from \( N(\hat{\mu}, \hat{\Sigma}) \) and stored in \( Z \), where \( d \) is the dimension of the parameter vector (eq 3.3). During the simulation, the state of \( \theta \) is appended to \( Z \) in every iteration (Ter Braak, 2006). DEMC gave near optimal acceptance rates (0.23). Convergence was tested by performing multiple independent runs from dispersed starting points. We found, by visual comparison of the posterior means of multiple runs that 2,000 iterations were sufficient to achieve convergence. The time needed for each run was around 20 seconds. The posterior probability that a protein performed the function under study was calculated by averaging the conditional probabilities that the protein performed the function, \( (1+exp(-v_i))^{-1} \), across iterations. Note that \( v_i \) varies across iterations because parameter values and states of neighboring unannotated proteins may vary across iterations. Receiving Operating Characteristic (ROC) curves were constructed from the resulting posterior probabilities. The prediction performance was measured using the Area Under the ROC Curve (AUC) (Hanley and McNeil, 1982a). The R code of BMRF is freely available at the website: https://gforge.nbic.nl/projects/bmrf/.

### 3.3.4 Datasets

We constructed a \textit{S. cerevisiae} interaction network using the physical protein-protein interaction dataset of (Collins et al., 2007). They used a scoring system called purification enrichment (PE) to evaluate each interaction. According to their study, selecting the interactions with PE score larger than 3.19 leads to a high quality network. This network contains 1,622 proteins (from which 84 are unannotated, corresponding to 5% of the total) and 9,074 interactions (Figure 3.2). We used this set of proteins and this topology as validation network for evaluating the performance of our method. Since the network provides information on the cellular process of the proteins, we used the set of GO terms that belong to the Biological Process (BP) ontology.
3.3. Methods

Figure 3.1: Bayesian Markov Random Fields analysis (BMRF) for protein function prediction in a nutshell. A. The topology of the interaction network is given. B. Functional annotations of proteins using a set of Gene Ontology terms. C. A partially annotated network. D-E. BMRF analysis.

Performance evaluation

To evaluate the prediction performance of our method, we selected by stratified sampling 800 out of 1622 proteins and treated them as unannotated. This masks the annotation of about half of the proteins in the network. Such a proportion of unannotated proteins is common even for the most well studied species (Sharan et al., 2007). The originally unannotated proteins were excluded from masking, but were kept in the network. MRF-Deng and BMRF were applied to the obtained data (i.e. a partially labelled network, containing the masked, the unmasked proteins and unannotated proteins), resulting in posterior probabilities for each protein and for each method. The masked proteins constituted the test set and their corresponding probabilities were used to construct ROC curves and to calculate the AUC score (Figure 3.5). We performed “out-of-bag” evaluation on 90 GO terms (Figure 3.4), selected by stratified sampling across different levels of abstraction of the GO Directed Acyclic Graph. The most sparse GO term contained 21 annotated proteins, while the most general 789. We considered the parameter values as estimated from the data prior to masking as the “true ones” (Figure 3.6).
Figure 3.2: **Number of unannotated proteins and number of interactions against Purification Enrichment (PE) score.** The numbers are divided by their values for PE = 0 (*i.e.* the network without any cutoff that contains the full set of proteins and edges). The validation network was constructed using PE = 3.19 as suggested by (Collins et al., 2007).
3.3. Methods

Function predictions for unannotated proteins

For actual prediction purposes we constructed an expanded network using the Collins et. al. (Collins et al., 2007) dataset. Figure 3.2, shows that for PE threshold of 0.65, most of the low confidence edges of the network are excluded while the majority of the proteins with unknown functions are included. We considered this network as suitable for protein function prediction purposes. It contained 5,419 proteins (1,170 of which were unannotated) and 89,685 interactions. The proteins assigned to the GO term “biological process unknown” were treated as unannotated. We applied our method to 340 GO terms from the BP ontology.

Comparison with other methods

Besides MRF-Deng, we compared the performance of BMRF with two other methods for protein function prediction i.e. diffusion based KLR (Lee et al., 2006) and the method proposed by Letovsky and Kasif (2003). KLR performs logistic regression on the diffusion kernel of the protein interaction network. First the diffusion kernel $K = e^{\tau L}$ is computed, where $\tau$ is the diffusion constant and $L$ is the opposite Laplacian of the adjacency matrix of the protein interaction network. We computed $K$ using the “expm” function of the “Matrix” R package that uses the squaring and scaling with Padé approximation. Predictions are made from the model of eq (3.3) using the diffusion matrix $K$ (instead of the original adjacency matrix) to define protein neighborhoods and the annotated proteins only, that is, KLR uses:

$$M_{i1} = \sum_{j \in S_i'} K(i, j)x_j$$

$$M_{i0} = \sum_{j \in S_i'} K(i, j)(1 - x_j)$$

in eq (3.3), where $S_i'$ denotes the set of neighbors of protein $i$ that have known function. Therefore, KLR ignores the neighborhood uncertainty in both parameter estimation and prediction, and also involves one more parameter, $\tau$. As in (Lee et al., 2006), we used a range of values for $\tau = (0.1, 0.5, 1.0, 3.0)$ and found that the best performance was achieved for $\tau = 0.1$ and therefore performed further computations using this value. Parameters were estimated by logistic regression. The motivation behind LK is that the number neighbors of protein $i$ that are in state 1 is binomially distributed, conditioned on the state of the protein $x_i$. The derived model can be expressed in similar manner as eq (3.3). In LK inferences for the unannotated proteins of the network are made by a heuristic algorithm based on belief propagation.
Chapter 3. Network based protein function prediction

Function predictions for unannotated proteins

For actual prediction purposes we constructed an expanded network using the Collins dataset (Collins et al., 2007). Figure 3.2, shows that for PE threshold of 0.65, most of the low confidence edges of the network are excluded while the majority of the proteins with unknown functions are included. We considered this network as suitable for protein function prediction purposes. It contained 5,419 proteins (1,170 of which were unannotated) and 89,685 interactions. The proteins assigned to the GO term “biological process unknown” were treated as unannotated. We applied our method to 340 GO terms from the BP ontology.

Results

Performance evaluation

We compared the prediction performance of BMRF with three other protein function prediction methods, i.e. MRF-Deng, LK (Letovsky and Kasif, 2003) and KLR on 90 GO terms (Figure 3.4), by treating 800 randomly chosen proteins (out of 1622) as unannotated and using the AUC score as an indicator of the prediction performance. The AUC score denotes the probability that a randomly chosen protein that performs the function is given a higher posterior mean by the predictor than a randomly chosen protein that does not (Hanley and McNeil, 1982a). The mean AUC values for the 90 GO terms were: 0.8195 for KLR, 0.8137 for the BMRF, 0.7867 for LK and 0.7578 for MRF-Deng. BMRF performed better than LK and MRF-Deng, that served as its basis, but slightly underperformed compared to KLR (Figure 3.5A). The improvement of BMRF over MRF-Deng is due to the fact that BMRF estimated the interaction parameters much better. Figure 3.6 illustrates the parameter values based on the simulation for GO term GO:0042592 (homeostatic process). Both methods estimate the intercept parameter reasonably well (Figure 3.6C) but the interaction parameters ($\beta^0$ and $\beta^1$) as estimated in MRF-Deng deviate far more from the true values than those of BMRF (Figure 3.6 AB). This led to the improvement in the prediction performance (Figure 3.6D). A further explanation is that the neighborhood counts of a large number of proteins are reduced in the MRF-Deng method because it disregards interactions with unannotated proteins and therefore the parameter estimates take larger absolute values. During the Gibbs sampling, the unannotated proteins are taken into account, but the model parameters are estimated from the pruned neighborhoods. This discrepancy explains the reduced performance of MRF-Deng compared to BMRF. This trend was observed for the majority of GO terms that we tested. The maximum improvement in the AUC score was 0.31 while the maximum deterioration was 0.1. We further calculated the precision when
the recall is set to 20% (PR20R). The mean PR20R across all the GO terms was 0.70 for KLR, 0.62 for BMRF, 0.54 for LK and 0.31 for MRF-Deng.

Figure 3.3: Running times for KLR and BMRF. The horizontal axis represents the size of the network and the vertical the time (in seconds) needed by each method. The computations were performed using the same hardware i.e. a Pentium 4 with dual core processor with 4GB of RAM and Linux operating system. The crosses denote the network size where the running times were evaluated. For BMRF the running time grows linearly with the network size while for KLR it grows polynomially.

Another important aspect of our comparison is the computational cost of the methods. BMRF has by definition larger computational cost than MRF-Deng, since it uses MRF-Deng for labelling initialization and also involves the additional parameter updating step, but the improvement in prediction performance compensates this increased cost. We did not compare with LK because our R implementation of this method was not sufficiently optimized for the speed. We compared KLR and BMRF in five networks of different sizes, constructed from the Collins et. al. data (Collins et al., 2007) by setting different PE score cut-offs (PE = 0.65, 1.29, 1.92, 2.55, 3.19). BMRF shows much better scaling properties and therefore is more suitable for large networks (Figure 3.3). The dominant factor of the computational cost of KLR is the computation of the diffusion kernel. In our implementation of KLR the diffusion kernel is obtained by scaling and squaring method with Padé approximation.
which is considered to be one of most competitive method currently (Moler and Loan, 2003). Still, matrix exponentiation is an active field of research in Numerical Analysis and therefore faster methods or implementations may exist (i.e. the power iteration method)

**Novel predictions for unannotated proteins**

We applied the BMRF method for 340 GO terms, aiming to predict the functions of 1170 unannotated *S. cerevisiae* proteins. Lists of protein names, GO terms probabilities and ranks per GO term are provided as supplementary material (Table S1). We checked for further information concerning the unannotated proteins in the literature and in the Saccharomyces Genome Database (SGD, accessed during December 2008). When functional information was found, we compared it with our predictions. In the majority of cases, existing information was in accordance with our predictions (Table 3.1). Below we give a number of examples of these predictions and evaluations.

YNR024W is involved in the degradation of “cryptic” non-coding RNA (Milligan et al., 2008), on the basis of which it is now annotated in SGD with a number of GO terms, including the term "nuclear-transcribed mRNA catabolic process". In our prediction, YNR024W is indeed predicted top ranking (1st) for GO term “mRNA catabolic process” (GO:0006402) which is the parent term of the previously assigned GO term.

There is evidence that protein YDL176W is involved in glycolysis and glucoleogenesis (Ulitsky et al., 2008; Ferre and King, 2006). We predict this protein as top ranking (1st) in the GO term “Glucose metabolic process” (GO:0006006), which is in agreement with the existing information.

YMR233W is a Small Ubiquitin-like Modifier (SUMO) substrate (Chen et al., 2007) and in mammals is involved pre-mRNA 3'-end processing (Vethantham et al., 2007). We predict the protein YMR233W to be top ranking (1st) for the GO term “RNA 3'-end processing“ (GO:0031123). Targeted experiments are needed to provide more direct evidence for the role of YMR233W in mRNA processing in yeast.

YOR093C is related to increased stress levels caused by the accumulation of unfolded proteins in the endoplasmic reticulum (Chen et al., 2005). YOR093C ranked first in “protein folding” (GO:0006457) in our predictions.

Information from SGD, based on the work of (Cheeseman et al., 2002), reveals that YLR315W and YDR383C are non-essential subunits of the Ctf19 central kinetochore complex. The kinetochore complex is known to have a central role in chromosome segregation. In our predictions YLR315W and YDR383C ranked 1st and 2nd respectively for the term “chromosome segregation” (GO:0007059) which is in accordance with the experimental evidence.

Proteins YGL128C (1st), YBL104C (2nd), YHR156C (3rd), were co-predicted to four hierarchically dependent GO terms concerning the nuclear
3.3. Methods
spliceosome mRNA splicing. They interact with proteins related to mRNA splicing in a very dense neighborhood of the protein interaction network. Information from SGD suggests that YGL156C is located in the snRNP U5 compartment and probably linked to mRNA splicing. YGL128C is annotated in SGD as "putatively involved in pre-mRNA splicing", while there is an IEA annotation (Inferred from Electronic Annotation) to the RNA splicing GO term. This is a parent node of our prediction and thus we provide a more detailed prediction. Also, this protein is located in the spliceosome and therefore in principle associated with the splicing processes. SGD does not provide information on the protein YBL104C. However, using BLAST we found the protein YPR178W (e-value = 0.043) to be a distant homologue. This protein is assigned to the GO term "nuclear mRNA splicing, via spliceosome" and contains a "splicing factor motif" in its sequence. The region of similarity with YBL104C is however located outside of this motif.

YOR227W is involved in the organization of the endoplasmic reticulum (Federovitch et al., 2008), on the basis of which it is now annotated in SGD with the GO term "endoplasmic reticulum organization". This protein ranked 4th for the GO term "organelle organization" (GO:0006996) which is the parent of the GO term assigned by SGD. According to SGD, YKR021W is proposed to regulate the endocytosis of the plasma membrane. This protein is top ranking for the GO term "Cellular localization", which is related to the proposed function.

SGD states that YBR227C is possibly a mitochondrial chaperone with non-proteolytic function while our predictions place this protein as first ranking for cation transport. This mismatch does not necessarily imply that our prediction is false, since functional evidence from SGD can be still weak and also it is rather common that proteins have multiple functions.

**Discussion**

Development of computational methods for protein function prediction based on interaction data is a challenging problem in bioinformatics. Here, we present a method to tackle this problem based on MRF. We followed the seminal work by Deng et al. (2003) in formulating the problem but we solved it in a significantly improved way. Our MCMC algorithm samples the MRF parameter values jointly with functional inference, whereas these are estimated in a single, questionable, training step in the work of (Deng et al., 2003). Our method outperforms Deng’s MRF method in efficiency of both parameter estimation and prediction performance. Also, we showed that our method performs better than the method proposed by Letovsky and Kasif (Letovsky and Kasif, 2003). The Kernel Logistic Regression (KLR) method (Lee et al., 2006) performed
3.3. Methods

Table 3.1: Manually evaluated predictions of protein functions.

<table>
<thead>
<tr>
<th>ORF</th>
<th>Protein function [reference]</th>
<th>Predicted GO term definition</th>
<th>RP Score</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>YNR024W</td>
<td>Nuclear transcribed mRNA catabolic process (Milligan et al., 2008)</td>
<td>mRNA catabolic process</td>
<td>56.87</td>
<td>1</td>
</tr>
<tr>
<td>YDL176W</td>
<td>Glycolysis and gluconeogenesis (Ulitsky et al., 2008; Ferre and King, 2006)</td>
<td>Glucose metabolic process</td>
<td>22.91</td>
<td>1</td>
</tr>
<tr>
<td>YMR233W</td>
<td>pre-mRNA 3'-end processing (Vethantham et al., 2007; Chen et al., 2007)</td>
<td>RNA 3'-end processing</td>
<td>31.01</td>
<td>1</td>
</tr>
<tr>
<td>YOR093C</td>
<td>Increased levels of unfolded proteins (Chen et al., 2005)</td>
<td>Protein folding</td>
<td>28.71</td>
<td>1</td>
</tr>
<tr>
<td>YLR315W</td>
<td>Ctf19 central kinetochore complex (Cheeseman et al., 2002)</td>
<td>Chromosome segregation</td>
<td>32.78</td>
<td>1</td>
</tr>
<tr>
<td>YDR383C</td>
<td>Ctf19 central kinetochore complex (Cheeseman et al., 2002)</td>
<td>Chromosome segregation</td>
<td>31.68</td>
<td>2</td>
</tr>
<tr>
<td>YGL128C</td>
<td>putatively involved in pre-mRNA splicing (SGD)</td>
<td>Nuclear mRNA splicing, via spliceosome</td>
<td>43.47</td>
<td>1</td>
</tr>
<tr>
<td>YBL104C</td>
<td>nuclear mRNA splicing, via spliceosome (Blast hit)</td>
<td>Nuclear mRNA splicing, via spliceosome</td>
<td>42.68</td>
<td>2</td>
</tr>
<tr>
<td>YHR156C</td>
<td>putatively involved in pre-mRNA splicing (SGD)</td>
<td>Nuclear mRNA splicing, via spliceosome</td>
<td>41.15</td>
<td>3</td>
</tr>
<tr>
<td>YOR227W</td>
<td>endoplasmic reticulum (Federovitch et al., 2008)</td>
<td>Organelle organization</td>
<td>1.63</td>
<td>4</td>
</tr>
<tr>
<td>YFR003C</td>
<td>Transporter activity (SGD)</td>
<td>Ion transport</td>
<td>6.53</td>
<td>8</td>
</tr>
<tr>
<td>YKR021W</td>
<td>Ubiquitin-mediated endocytosis (SGD)</td>
<td>Cellular localization</td>
<td>3.65</td>
<td>3</td>
</tr>
<tr>
<td>YBR227C</td>
<td>possibly a mitochondrial chaperone (SGD)</td>
<td>Cation transport</td>
<td>8.86</td>
<td>1</td>
</tr>
</tbody>
</table>
Chapter 3. Network based protein function prediction

Figure 3.5: **Performance comparison for 90 GO terms, using the Area Under the ROC Curve (AUC).** The points above the diagonal denote improved performance of BMRF against A. MRF-Deng B. LK C. KLR. BMRF performs better for the majority of the tests compared to MRF-Deng and LK. KLR performs slightly better, but it is difficult to be applied in large datasets.

slightly better than BMRF, but this method involves an expensive matrix exponentiation operation, that is needed to compute the diffusion kernel. This makes KLR impractical for large networks.

In this study we focused on the methodological aspect and limit our experiments to a single data source. In this way, we could clearly show that our method is more powerful than its predecessor. Our method can handle multiple data sources such as expression correlation datasets, co-occurrence of protein names in literature obtained via text-mining, or cross-species sequence comparisons (e.g. orthology networks (Kuzniar et al., 2008; Gabaldon et al., 2009)). The datasets can then either be merged into a single network (e.g. (Nariai et al., 2007)), or used separately, leading to additional terms in the energy function and additional parameters (Deng et al., 2004) which can then be treated in the Bayesian way as proposed here. Also, protein networks for most of the species are far from complete and therefore dealing with the uncertainty of the network topology is another direction for future research.

Importantly, we showed that our approach is suitable for networks in which a large proportion of the proteins is unannotated. Our method can be applied for protein function prediction in species for which large-scale interaction datasets are available. We provided Gene Ontology predictions for 1,170 unannotated yeast proteins and for many high-ranking predictions we found supporting information in the literature.
Figure 3.6: Comparison of parameter estimation and prediction performance between BMRF and MRF-Deng for the GO term “homeostatic process”. A-B. In BMRF the parameters $\beta^0$ and $\beta^1$ are sampled closeby to the true parameter values, in contrast to MRF-Deng where the parameters are estimated using only the annotated part of the network and lead to overestimated values. C. Both methods estimate the intercept reasonably well. D. ROC curves for the prediction performance of the two methods. The AUC value for BMRF is 0.79 and for MRF-Deng is 0.71.
Acknowledgments

We thank Jeroen Engelberts from LifeScience Grid project for his support during our computations and Ioannis Stergiopoulos for his assistance on preparing the figures. We thank the three reviewers for their helpful comments.
Chapter 4

Protein function prediction by integrating multiple data sources

1

4.1 Abstract

Although Arabidopsis thaliana is the best studied plant species, the biological role of one third of its proteins is still unknown. We developed a probabilistic protein function prediction method that integrates information from sequences, protein-protein interactions and gene expression. The method was applied to proteins from Arabidopsis thaliana. Evaluation of prediction performance showed that our method has improved performance compared to single source-based prediction approaches and two existing integration approaches. An innovative feature of our method is that enables transfer of functional information between proteins that are not directly associated with each other. We provide novel function predictions for 5,807 proteins. Recent experimental studies confirmed several of the predictions. We highlight these in detail for proteins predicted to be involved in flowering and floral organ development.

4.2 Introduction

Arabidopsis thaliana is the most widely used model organism in plant research. Unraveling the biological processes in this species is therefore essential for the

1Published as: Kourmpetis YAI, van Dijk ADJ, van Ham RCHJ, ter Braak CJF (2011) Genome-Wide Computational Function Prediction of Arabidopsis Proteins by Integration of Multiple Data Sources. Plant Physiology 155(1): 271-281
understanding of plant biology in general and for the transfer of this knowledge to other species. Fundamental to this goal is the functional annotation of Arabidopsis proteins. While the aim of the National Science Foundation 2010 initiative on Arabidopsis was to reveal the function of each of its proteins by 2010 (Berardini et al., 2004), currently one third of the proteins still lack a functional annotation in terms of their biological role. It is unlikely that these missing annotations can be generated solely by large-scale experimental analysis, as these often provide only a general view on the functions of proteins. At the same time, targeted experiments remain time-consuming and often require a specific prior hypothesis of function, which for many proteins can not be formulated. As a complementary approach, computational methods are therefore needed that can accurately predict protein functions on a large scale or provide leads for hypotheses of function and the design of targeted experiments. Methods like BLAST (Altschul et al., 1997) and InterProScan (Zdobnov and Apweiler, 2001; Mulder et al., 2005b) infer the functions of proteins by identifying their homologs in sequence databases and by the presence of domains that are associated with particular functions, respectively. This homology-based transfer is a powerful approach for functional annotation of novel proteins, but also one that can lead to erroneous inferences because similarity at the sequence level does not necessarily imply that proteins carry out the same function. Also, for lineage-specific or highly divergent proteins the probability of identifying a functionally characterized homolog is small. Finally, homology transfer cannot deal with sub- and neo-functionalization of recent paralogs. Besides plain sequences, other types of information therefore need to be integrated to maximize the coverage and accuracy of function prediction (Forslund and Sonnhammer, 2008).

Proteins that participate in the same biological process often interact physically or exhibit correlations in their expression patterns. High-throughput experiments provide genome-wide information on such associations. In addition, protein-protein interactions can be predicted from sequences (Marcotte et al., 1999; Itzhaki et al., 2006; Van Dijk et al., 2008). The associations can be viewed as a network with nodes representing proteins and edges representing the interactions between them. Computational methods can employ those networks for function predictions by analyzing the topology of the network to identify sets of proteins with dense interactions between them (Enright et al., 2002), or to analyze their direct relationships (Vazquez et al., 2003; Letovsky and Kasif, 2003) using the guilt by association principle. Such methods employ statistical models, the performance of which relies on appropriate selection of the parameters. Recently, we (Kourmpetis et al., 2010a) developed a method that accurately estimated these model parameters using a Bayesian approach and which outperformed other related methods.

Integrated approaches for protein function prediction make use of diverse
types of data. Peña Castillo et al. (2008) evaluated such methods using multiple genomic datasets from *Mus musculus* and concluded that different data sources provide complementary pieces of information on protein function.

For *Arabidopsis thaliana* various types of genomic datasets are available. The genome sequence is completed (Arabidopsis, 2000), gene coexpression levels have been calculated using expression values from a wide variety of conditions (Obayashi et al., 2009) and physical protein-protein interactions have been identified through experiments or predicted through homology (Geisler-Lee et al., 2007).

Despite the availability of these data, only a limited number of studies has focused on function prediction of Arabidopsis proteins through the integration of data. Clare et al. (2006) predicted the molecular functions of proteins integrating sequence features with expression experiments. The authors used the decision tree algorithm Q4.5 (Quinlan, 1993) to predict function terms of the controlled vocabularies of Gene Ontology (GO) (Ashburner et al., 2000) and MIPS (Frisman et al., 2001). Their algorithm was developed for predictions that required functional classes to be ordered in a hierarchical tree structure. GO has a Directed Acyclic Graph (DAG) structure which is not a tree. Clare et al. (2006) therefore restricted their predictions to the GO terms that are related to Molecular Functions and further to the most general terms that have a tree structure. Lan et al. (2007) predicted the functions of Arabidopsis proteins that are involved in plant response to abiotic stress by combining different gene expression experiments. Horan et al. (2008) grouped Arabidopsis proteins with similar expression patterns by cluster analysis and predicted functions based on overrepresented GO terms in each identified cluster. The agglomerative clustering algorithm assigned each protein to exactly one cluster, while the complex nature of the biological processes leads to the expectation that proteins will belong to multiple clusters. Further, the cluster analysis did not provide information on the uncertainty of each prediction. GeneMania (Mostafavi et al., 2008) is a Gaussian Markov Random Fields-based method for protein function prediction that combines multiple networks. In the evaluation experiment of (Peña Castillo et al., 2008) GeneMania was shown to be one of most accurate methods and besides predictions for the *Mus musculus* proteins it was further applied to several species including Arabidopsis. Bradford et al. (2010) combined sequence data, gene location in the chromosome, phylogenetic profiles, physical protein-protein interactions and expression levels to predict functions of proteins in Arabidopsis. Using a two-step approach the authors first constructed ranked lists of proteins that are functionally associated with each query protein. Functions were then inferred by Gene Set Enrichment Analysis (Subramanian et al., 2005) of these lists. Since a large fraction of Arabidopsis proteins lack functional annotations, the ranked lists may contain no or only few proteins with GO terms assigned to them. The analysis has thus
difficulty identifying infrequent GO terms. Lee et al. (2010) derived a composite functional linkage network (Karaoz et al., 2004) for the Arabidopsis proteins by integrating data from sequences, coexpressions and physical interactions from Arabidopsis and from other species. As in the previous approach, functional inference was only possible when at least one direct neighbor of the query protein had a known function. From the total set of 7,465 Arabidopsis proteins without functional annotation, 2,986 (40%) were not linked to any protein with known function and therefore function predictions for them was not possible. VirtualPlant (Katari et al., 2010) is visualization software that integrates different sources of data for Arabidopsis including GO function information on the proteins. VirtualPlant is valuable for bridging the gap between biologists and bioinformaticians by providing an intuitive way to integrate and mine diverse data sources but does not perform de novo function prediction. In this study we performed genome wide function prediction for Arabidopsis proteins by integrating protein sequences, gene expression data and experimentally derived or predicted protein-protein interactions. We applied Bayesian Markov Random Fields (BMRF) (Kourmpetis et al., 2010a), a probabilistic method shown to be suitable when the function of a large number of proteins has to be predicted, such as in the case of Arabidopsis. A powerful feature of BMRF is that it can transfer functional information beyond direct interactions and so can provide function predictions for proteins linked with other proteins of unknown function. In the studies of Bradford et al. (2010) and Lee et al. (2010) such predictions were not possible. We extended the original BMRF to multiple data sources using the framework of Deng et al. (2004) and, in a one-step approach, we optimize data source integration for function prediction. Our analysis resulted in 64,721 novel protein function predictions for 5,807 proteins in 867 Gene Ontology terms that provide detailed functional descriptions. We provide the predictions in the website (http://www.ab.wur.nl/bmrf/). After demonstrating the performance of our method using cross-validation as a validation step, we investigated recent experimental evidence for our predictions. As an example of the usefulness of our predictions, we evaluated our predictions on proteins involved in the flowering process in Arabidopsis.

4.3 Materials and Methods

4.3.1 Protein-Protein Interaction (PPI) network

Physical interactions between proteins provide valuable information for their functions. Proteins that interact are members of the same complex and involved in the same biological process or pathway. There are around 3,000 experimentally identified interactions between Arabidopsis proteins. In addition, interactions can be predicted by detecting interacting orthologs. Such
predicted interactions are called interologs. Geisler-Lee et al. (2007) used the orthology detection algorithm INPARANOID (Remm et al., 2001) to identify Arabidopsis interologs from several well studied species, including *Saccharomyces cerevisiae, Caenorhabditis elegans, Drosophila melanogaster* and *Homo sapiens*. We downloaded the experimentally derived interactions and the interologs from the TAIR website and constructed a protein-protein interaction (PPI) network that contains 7,177 proteins with 72,266 interactions.

### 4.3.2 Coexpression (COEX) Network

Besides direct physical interactions, proteins involved in the same biological process present correlations in their gene expression. Genome wide expression experiments therefore provide an important data source for protein function prediction. In the recent study of Lee et al. (2010), gene expression was found to be the most informative data source for protein function prediction. The ATTED-II database (Obayashi et al., 2009) aggregates 58 experiments and 1,388 microarray slides in total. We downloaded the coexpression data from the ATTED-II website on the 21st of July 2009. For each coexpression, a confidence value is provided, which is defined by the mutual rank of the coexpression of two proteins. The authors calculate the Pearson Correlation Coefficient (PCC) between all pairs of proteins. Then for each protein they ranked the PCC. As confidence value for an interaction they calculate the mutual rank, i.e. the square root of the product for the rankings in both directions of the interaction. This coexpression measure is useful for protein function prediction (Obayashi and Kinoshita, 2009). Our coexpression network (referred to as COEX) was constructed by setting the maximum mutual rank to 60. COEX contains 22,133 proteins with 358,540 interactions.

### 4.3.3 Functional Domains

Sequence signatures are an important source of information concerning the function of a protein. InterPro (Mulder et al., 2005b) aggregates the most important tools and databases that are used to identify such sequence patterns and to link them to particular functions, primarily at the molecular level. We used the identified InterPro functional signatures for the proteome of Arabidopsis. This dataset was downloaded directly from the TAIR website in October 2009.

### 4.3.4 Gene Ontology Annotations

We downloaded the annotation file for Arabidopsis from the Gene Ontology, with version 1.1271 deposited on 13 October 2009. This file contains in total 14,038 Biological Process (BP) annotations. Annotations with evidence code
“Inferred from Electronic Annotation” (IEA) were removed from the dataset because these are derived from InterPro hits, which we used independently in our study for function prediction. All remaining annotations were up-propagated to their more general terms using the GO DAG structure. In total there were 2,894 GO terms appearing at least once, but many of them were extremely sparse, i.e. containing less than 10 proteins assigned to them (after up-propagation). Our final set contained 1,024 GO terms and 8,247 proteins lacking BP annotations in the Arabidopsis thaliana proteome.

4.3.5 Protein function prediction

BMRF for single network

We investigated different approaches for protein function prediction. Our starting point was the Bayesian Markov Random Fields (BMRF) method for protein function prediction based on a single network described in an earlier study (Kourmpetis et al., 2010a). In particular, given a network that contains \( N \) proteins and \( S \) edges (indicating interactions between proteins), a particular function of interest (i.e. a GO term) is represented by a \( N \)-dimensional binary vector \( X \) with element \( x_i = 1 \) if protein \( i \) is annotated as performing the function and \( x_i = 0 \) otherwise. The elements of \( X \) that correspond to unannotated proteins are unknown. The objective of BMRF is to infer the unknowns given the observed part of \( X \) using the edges of the protein network. The log odds of the probability that an unannotated protein \( x_i \) performs the function of interest, given the annotations for all other proteins, denoted by \( X_{-i} \), depends on the number of its direct neighbors performing the function and the number of them that do not perform the function:

\[
\log \frac{P(x_i = 1 \mid X_{-i})}{P(x_i = 0 \mid X_{-i})} = \alpha + \beta_1 \sum_{j \in S_i} x_j + \beta_0 \sum_{j \in S_i} (1 - x_j) = \alpha + \beta_1 M_{i1} + \beta_0 M_{i0}, \quad (4.1)
\]

where \( \alpha \) denotes the intercept, \( \beta_1 \), \( \beta_0 \) are interaction parameters and \( S_i \) denotes the set of proteins that interact with protein \( i \), so that \( M_{i1} \) denotes the number of proteins that interact with protein \( i \) and perform the function, while \( M_{i0} \) denotes those that interact with protein \( i \) but do not perform the function. Inference for the unannotated part of \( X \) can be made using a Markov Chain Monte Carlo approach (Kourmpetis et al., 2010a). We refer to BMRF-PPI when this method is applied to the PPI network and to BMRF-COEX when it is applied to the COEX network. In both cases the predictions are limited to the proteins that appear in the network that BMRF is applied to. For example it is not possible to make predictions for the proteins appearing only in the COEX network, by applying BMRF to the PPI network.
4.3. Materials and Methods

BMRF for multiple networks

A natural way to integrate multiple networks through BMRF is by using a set of interaction parameters per network. This approach was originally proposed by Deng et al. (2004):

\[
\log \frac{P(x_i = 1 | X_{-i})}{P(x_i = 0 | X_{-i})} = \alpha + \beta_1^{PPI} M_{i1}^{PPI} + \beta_0^{PPI} M_{i0}^{PPI} + \beta_1^{COEX} M_{i1}^{COEX} + \beta_0^{COEX} M_{i0}^{COEX}.
\]

We refer to this model as BMRF-MULTI. A special case of this model is obtained by constraining the interaction parameters between the networks to be equal, i.e. \(\beta_1^{PPI} = \beta_1^{COEX}\) and \(\beta_0^{PPI} = \beta_0^{COEX}\). This approach is equivalent to applying BMRF to the single network that is the union of the PPI and COEX networks. The union network has an edge if an edge appears in at least one of the networks. We refer to the latter model as BMRF-UNION.

Elastic Nets for functional domains

Given the set of \(M\) available InterPro domains, we constructed the \(N \times M\) binary matrix \(D\), where the element \(d_{nm}\) is equal to 1 if protein \(n\) contains the InterPro domain \(m\) and 0 otherwise. The probability that a protein performs the function of interest depends on the presence/absence profile of domains. We write this relationship as a logistic regression with binary variables:

\[
\log \frac{P_1(d_i | \beta^D)}{P_0(d_i | \beta^D)} = \beta_0^D + \sum_{m=1}^{M} \beta_m^D d_{im}
\] (4.2)

The parameter vector \(\beta^D\) contains the regression coefficients for the domains and can be estimated using the proteins with known functional annotation. A particular GO term is expected to be related to only a small subset of the domains and those domains usually act in a concerted way. Therefore, we aim to perform variable selection while keeping highly correlated variables into the model. A suitable method for this purpose is the Elastic Net (Zou and Hastie, 2005) version for logistic regression (Park and Hastie, 2007). EN combines Lasso (Tibshirani, 1996) and Ridge regression (Hoerl and Kennard, 1970). In Lasso regression, the sum of the absolute values of the regression coefficients is penalized, while in Ridge regression, their sum of squares is penalized. EN combines both regularization methods using a convex parameter for which on one extreme the model becomes equivalent to Lasso and on the other extreme to Ridge regression. In cases with highly correlated variables, Lasso tends to include only one of those variables in the model. In our application, we aim to obtain a sparse model that includes the set of domains that are related to the
function. For this reason, we selected EN as the most appropriate method for this application. EN has two parameters to be set prior to model selection. The first is the convex parameter (taking values between 0 and 1) and the second is the penalty parameter. Usually those parameters are estimated through cross validation. We adopted a simple approach by fixing the convex parameter to 0.5 (that gives equal weight to both methods) and by selecting from a series of penalty parameters the one that leads to the largest model containing no more than ten variables (domains). All computations involving EN were made using the GLMNET R package (Friedman et al., 2010). We refer to this function prediction model as EN-DOMAINS.

Integration of networks and functional domains

Let $P_d$ denote the output (on logit scale) from EN-DOMAINS. We insert $P_d$ into the BMRF model, also adding one more parameter $\beta_d$:

$$
\log \frac{P(x_i = 1 \mid X_i)}{P(x_i = 0 \mid X_i)} = \alpha + \beta_d P_{di} + \beta_{PPI} M_{i1}^{PPI} + \beta_{PPI} M_{i0}^{PPI} + \beta_{COEX} M_{i1}^{COEX} + \beta_{COEX} M_{i0}^{COEX}
$$

Function prediction is further performed by BMRF and updating $\beta_d$ in the same way with the other parameters in the model. We remark that all the quantities in this model are updated during BMRF, while $P_{di}$ remains constant.

4.3.6 Performance Evaluation

We estimated the performance of each protein function prediction model by constructing 100 benchmarking datasets, one for each of 100 GO terms randomly selected from different levels of generality. For each GO term, we selected 300 proteins with known function (i.e. known whether it is assigned or not to this particular GO term) to be treated as unknowns. The selection of these “masked” proteins was done using the following procedure. First the proteins with known functions were classified in three sets: those that appear only in the PPI network, those that appear only in the COEX network and finally those that appear in both networks. One hundred proteins were randomly selected from each set to be treated as unannotated. Consequently, 200 proteins appear in the PPI network and 200 in the COEX network while 100 appear in both. For the very sparse GO terms (i.e. those with less than 20 proteins assigned to them) we randomly selected exactly half of the proteins that belong to the GO term to be masked. All the protein function prediction models were applied to the hundred benchmarking datasets so obtained. For the evaluation of prediction performance we used the Area Under the Receiver Operating Characteristic Curve (AUC) (Hanley and McNeil, 1982b; Fawcett, 2006), Precision
(Prec), Recall (Rec) and F-score (F-score = 2*Prec*Rec/(Prec+Rec), i.e. the harmonic mean of Recall and Precision). Recall and Precision are defined as the fraction of proteins correctly predicted of having the function out of the total number of proteins having the function and the fraction of proteins correctly predicted of having the function out of the total number of proteins predicted having the function, respectively. High Recall and high Precision are conflicting aims and the F-score is a compromise between them that is often used in information retrieval. All performance metrics were computed using the ROCR R package (Sing et al., 2005).

4.3.7 Construction and evaluation of list with novel predictions

BMRF computes for each protein the probability of membership to each GO term, except for the most general ones (with more than 3,000 proteins annotated to them). From these memberships probabilities, we constructed a list of novel predictions by selecting the cutoff per GO term that maximized the F-score in the set of proteins with known annotations. We tested this method for cutoff selection on the benchmarking datasets. We first obtained the optimal cutoff using the proteins with known function and then applied this cutoff to the predicted part. Both sets of proteins estimate closely the Recall values (Figure 4.1). The Precision when estimated from the set of proteins with known function is a conservative estimation of the one obtained from the set of masked proteins.

The novel predictions were compared with the new annotations in the annotation file of April 18, 2010. Because, the predictions may be related to the correct annotations by being more general or more specific, we used the GO DAG structure to up-propagate the predicted and the “true” annotations per protein. We define as True Positives the set of GO terms that appear in both lists, False Positives those that appears only in the predicted one and False Negatives those that appear only in the “true” list. From the measurements we calculated Precision, Recall and the F-score. The F-score was tested for statistical significance, for which we used a Monte Carlo permutation test, in which the prediction lists of proteins were randomly permuted (shuffled) among proteins that were common to our predictions and the new annotations. After each shuffle, the F-score was calculated. The p-value of the test is the rank of the F-score in the data among all the F-scores calculated from the shuffled data, divided by the number of permutations. With 100,000 permutations the lowest obtainable p-value is thus 0.00001.
Figure 4.1: Recall and precision scores estimated from the held out set (proteins with “masked” annotation) versus those from the training set (proteins with known annotation)(x-axis). In supervised learning, performance estimates are based on the held out set. A The Recall rates of the training and the held out set are in accordance. B Precision estimated from the training set provides a conservative estimate of the true precision estimated by the held out set.

4.3.8 Performance comparison of BMRF with other prediction methods

We used the list of new annotations as validation set to compare the performance of BMRF with two state of art methods that provide function predictions for Arabidopsis proteins, i.e. Aranet (Lee et al., 2010) and GO-AT (Bradford et al., 2010). We obtained function prediction lists with confidence scores by querying the web-servers of these two methods. Precision and Recall values were calculated in the full range of scores per method, by applying cutoffs and up-propagating the resulting lists. The posterior probabilities from BMRF are uncalibrated in the sense that it is not useful to apply a single cutoff for all the GO terms. For this comparison, we calibrated the probabilities using the function:

$$p_{ng}^{cal} = \frac{1}{1 + exp(-U(p_{ng}, p_g))}$$  \hspace{1cm} (4.3)
4.4 Results

with

\[ U(p_{ng}, p_g) = \alpha \log \frac{p_{ng}}{1 - p_{ng}} + (1 - \alpha) \frac{p_g}{1 - p_g} \] (4.4)

and \( p_{ng} \) is the BMRF posterior probability for protein \( n \) at GO term \( g \) and \( p_g \) the prior probability of membership for GO term \( g \) (that is the proportion of the proteins in our data set that are assigned to this term). After some experimenting using yeast data Kourmpetis et al. (2010a), the parameter \( \alpha \) was set to 2, which gives, for sparse GO terms, calibrated probabilities that are the product of \( p_{ng} \) and \( p_{ng} / p_g \). The calibrated probabilities are available from http://www.ab.wur.nl/bmrf/.

4.4 Results

4.4.1 Model selection

We extended our BMRF function prediction approach (Kourmpetis et al., 2010a) to deal with multiple and diverse datasets, and applied it to sequence data, protein-protein interaction data and coexpression data available for *Arabidopsis thaliana*. The first and most crucial step in our study was to identify the best performing function prediction model. We investigated six models of different levels of complexity. Three of those used only one type of data (sequence, protein-interaction or coexpression). The other three used different ways to integrate the various datasources. For benchmarking, we masked the annotation of a set of proteins with known annotations. This set was divided in three strata: network-specific proteins that appear in either 1) the protein-protein interaction network or 2) the coexpression network, and 3) those that appear in both. We randomly selected 100 proteins per stratum and predicted their functions. We evaluated the performance by constructing 100 such benchmarking datasets, one for each of 100 GO terms, and used the Area Under the Receiver Operating Characteristic Curve (AUC) as performance measure.

Table 4.4.1 shows the mean AUC scores for the candidate models in four different evaluation settings (higher value means better performance). Overall, the best performing model is the one that integrates networks and sequence information (BMRF-UNION-DOMAINS). In general, the predictions based on the integrated network outperformed those from single networks (Figure 4.2A-B). The prediction performance improved not only for proteins that appear in both networks but also for network-specific proteins (Figure 4.2A-B, Table 4.4.1). The latter was unexpected because the neighborhood of a network-specific protein does not change after the integration of the networks. This performance improvement therefore reveals an appealing property of BMRF, namely the propagation of information over long ranges across the network. A
Table 4.1: Mean AUC scores for the evaluation datasets. BMRF-PPI (BMRF-COEX) denotes the application of BMRF to the protein-protein interaction (coexpression) network, EN-Domains denotes application of Elastic Net to the Domain information. BMRF-MULTI denotes the integration of the PPI and COEX networks internally by BMRF. BMRF-UNION denotes the application of BMRF to the union of the PPI and COEX networks, whereas BMRF-UNION-DOMAINS also adds the Domain information. Columns: “PPI (COEX) only” evaluates performance for the masked proteins that appear only in the PPI (COEX) network, “Intersection” for the masked proteins that appear in both networks and “All” for all masked proteins. NA denotes Not Available. The best performing score per category is shown in bold.

<table>
<thead>
<tr>
<th>Model,Protein sets</th>
<th>PPI only</th>
<th>COEX only</th>
<th>Intersection</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMRF-PPI</td>
<td>0.67</td>
<td>NA</td>
<td>0.68</td>
<td>NA</td>
</tr>
<tr>
<td>BMRF-COEX</td>
<td>NA</td>
<td>0.66</td>
<td>0.67</td>
<td>NA</td>
</tr>
<tr>
<td>EN-DOMAINS</td>
<td>0.61</td>
<td>0.63</td>
<td>0.61</td>
<td>0.62</td>
</tr>
<tr>
<td>BMRF-MULTI</td>
<td>0.71</td>
<td>0.70</td>
<td>0.74</td>
<td>0.70</td>
</tr>
<tr>
<td>BMRF-UNION</td>
<td>0.70</td>
<td>0.70</td>
<td>0.74</td>
<td>0.68</td>
</tr>
<tr>
<td>BMRF-UNION-DOMAINS</td>
<td>0.76</td>
<td>0.77</td>
<td>0.79</td>
<td>0.75</td>
</tr>
</tbody>
</table>

4.4.2 Protein function prediction for Arabidopsis

We applied our probabilistic method integrating protein-protein interactions, gene coexpression, and functional domains to predict functions for 8,247 Arabidopsis proteins with unknown biological role. Our procedure computes a posterior probability for each protein against and for each GO term, which makes the interpretation of the predictions somewhat difficult. To overcome this problem, we constructed a list with positive predictions after obtaining the optimal F-score cutoff on the posterior probability from the set of annotated proteins and applying it to the set of proteins with unknown function. This resulted in a list of 64,721 predictions for 5,807 proteins against 867 GO terms (the list is available from http://www.ab.wur.nl/bmrf/). For each prediction we calculated the Precision and Recall at the given probability cutoff in order to facilitate further use (biological interpretation) of the list. Both metrics are high in the list of predictions (Figure 4.3). The density of Recall rates shows that an appreciable fraction of proteins received a prediction, while the Pre-
cision rates, which are even higher than the Recall rates, shows that the list contains a large fraction of correctly predicted proteins.

For validation, we investigated whether there was recent experimental evidence in the literature supporting our predictions and that was not available at the time of our computations. For this purpose, we downloaded the annotation file for Arabidopsis on the 18th of April 2010 from Gene Ontology and identified the proteins that were annotated after the 13th of October 2009 (the date on which we downloaded the annotation file used in our predictions). There were 194 new annotations with GO terms from the Biological Process branch for 103 proteins that were included in our prediction list. In 14 cases, we predicted the exact GO term (Table 4.2) or a more detailed one according to the GO DAG. For 109 new annotations we predicted one or more GO terms that are more general but related. Hence, in total we predicted a GO DAG related function (either more general, exact or more specific) for 123 out of the 194 (63%). This level of performance is highly significant (p-value < 0.00001) as judged by a permutation test.

Table 4.2: Experimentally verified predictions where BMRF predicted the exact GO term or a more detailed one. Rel denotes the relation between the GO terms of the new annotation and the BMRF prediction (E: exact prediction, D: the GO term predicted by BMRF is a successor of the annotation according to the GO Directed Acyclic Graph).

<table>
<thead>
<tr>
<th>Proteins</th>
<th>Annotation</th>
<th>Reference</th>
<th>BMRF prediction</th>
<th>Rel</th>
</tr>
</thead>
<tbody>
<tr>
<td>YUC8</td>
<td>Auxin biosynthetic process</td>
<td>Rawat et al., (2009)</td>
<td>Auxin biosynthetic process</td>
<td>E</td>
</tr>
<tr>
<td>BIR1</td>
<td>Negative regulation of defense response</td>
<td>Gao et al., (2009)</td>
<td>Negative regulation of defense response</td>
<td>E</td>
</tr>
<tr>
<td>AFH3</td>
<td>Actin nucleation</td>
<td>Ye et al., (2009)</td>
<td>Actin nucleation</td>
<td>E</td>
</tr>
<tr>
<td>AT1G18370</td>
<td>Cytokinesis</td>
<td>Oh et al., (2008)</td>
<td>Cytokinesis</td>
<td>E</td>
</tr>
<tr>
<td>AT3G10570</td>
<td>Flower development</td>
<td>Li-Beisson et al., (2009)</td>
<td>Flower development</td>
<td>E</td>
</tr>
<tr>
<td>AT3G23070</td>
<td>Seed development</td>
<td>Asakura et al., (2008)</td>
<td>Embryonic development ending in seed dormancy</td>
<td>D,E</td>
</tr>
<tr>
<td>AT5G64580</td>
<td>Embryonic development</td>
<td>Mutwil et al., (2010)</td>
<td>Embryonic development ending in seed dormancy</td>
<td>D</td>
</tr>
<tr>
<td>AT3G08710</td>
<td>Cell Communication</td>
<td>Meng et al., (2010)</td>
<td>Intracellular signaling cascade</td>
<td>D</td>
</tr>
</tbody>
</table>

We also compared the prediction performance of BMRF with two recently
published integration methods, i.e. Aranet (Lee et al., 2010) and GO-AT (Bradford et al., 2010), using the list of new annotations as validation dataset. Each method provides scored predictions from which we calculated Precision and Recall at a series of cutoffs. The precision of BMRF was higher than the other methods at any given recall rate (Figure 4.4).

For the newly annotated proteins we also make 718 predictions that were not inferred by the newly obtained experimental data. We expect that at least some of our novel predictions will be confirmed in future experiments. Below, we further comment on some of the supported predictions.

Monaghan and coworkers (Monaghan et al., 2009) performed double mutant analysis on the Arabidopsis proteins MAC3A and MAC3B and showed that they are involved in defense response against plant pathogens. InterPro searches did not return information related to the function of those proteins, while the BLAST2GO tool predicted the more general term “defense response”. On the other hand, BMRF predicted the GO terms: “defense response to bacterium” (GO:0042742) for both proteins which is in complete agreement with the aforementioned experimental study (Figure 4.5A). Also, BMRF predicted the involvement of MAC3B in “activation of innate immune response” (GO:0002218), which is also a defense related process. The same gene was identified to have a ubiquitin-protein ligase molecular function (Wiborg et al., 2008). BMRF predicted that MAC3B is involved in the biological process “protein ubiquitination” (GO:0016567), which is in accordance with that study. Furthermore, Borges et al. (2008) performed a genome wide transcriptome analysis identifying MAC3B to be involved in “embryonic sac development” (GO:0009553), a function that was also predicted by BMRF. Other examples in which BMRF accurately predicted protein functions include YUC8 that was recently identified by Rawat et al. (2009) to be involved in the “auxin biosynthetic process” (Figure 4.5B) and BIR1 that was identified to be involved in “negative regulation of defense response” (Gao et al., 2009) (Figure 4.5C). BMRF predicted that AT3G8710 is involved in “intracellular signaling cascade” (GO:0023034). On the basis of results described by Meng et al. (2010), this protein is newly assigned to the more general term “cell communication” (GO:0007154).

4.4.3 Flowering and floral organ development in Arabidopsis

As a specific example of the usefulness of our method, we here focus on the evaluation of predictions for flowering and floral organ development. Obviously, this is a biological process for which annotation transfer between species is only possible within the plant kingdom. Given the current sparsity of annotation for plants, homology based methods have limited scope for annotation transfer
from other species and our network based approach is therefore in principle better suited.

GO terms were selected which describe processes related to flowering and floral organ development (Suppl Table 4.3). We first discuss a few groups of proteins, including transcription factors, that are predicted for the selected GO terms, and then focus on two particular terms for “floral transition” (GO:0010228) and “corolla development” (GO:0048465).

One important class of transcription factors (TFs) with known roles in the regulation of floral transition and in flower development are the MADS domain proteins (Coen and Meyerowitz, 1991; Ferrario et al., 2004; Ng and Yanofsky, 2001). For several members of this family, BMRF predicted additional functions which are consistent with those known functions. For example, for the MADS domain protein AGL6, the GO term “floral organ development” was predicted, as well as the more detailed term for “carpel, gynoecium and ovule development”. Although in Arabidopsis the function of AGL6 has remained elusive so far (due to the lack of a single loss-of-function mutant exhibiting a clear phenotype), recently it was shown that an AGL6 homolog is involved in petal and anther development in petunia (Rijpkema et al., 2009). Hence, our prediction for floral organ development is supported by independent evidence.

A second MADS domain protein predicted for carpel/gynoecium development was AGL15. This protein has a known function in the floral transition process (Adamczyk et al., 2007), but our prediction suggested that it has a broader function in the development of floral organs.

Several other MADS domain proteins with unknown functions to date were predicted to function in flower development, including AGL13, AGL14, AGL71, AGL72 and AGL79. Note that several of these proteins arose through lineage- or species-specific duplications that occurred in the MADS domain protein family. Such duplications render annotation transfer based on orthology inadequate because it cannot deal with sub- or neo-functionalization, while our network-based method can in principle deal with these cases. Two additional predictions for carpel development, the MADS domain proteins AP3 and PI, seem incorrect in light of existing knowledge that these proteins are only involved in development of petals and stamens, although it is known that PI is temporarily expressed in the fourth whorl where carpel formation takes place (Goto and Meyerowitz, 1994).

Regulation of transcription via MADS domain TFs involves histone modification proteins (Hill et al., 2008; Ng et al., 2009). Similarly, chromatin modifications are important in the regulation of the floral transition (for a recent review see (He, 2009)). An interesting aspect of our predictions is that several proteins related to chromatin modifications are predicted to be involved in flower development, including histone H3, and SPT16 and SSRP1, which are part of a chromatin remodelling complex. These predictions do not necessarily
Chapter 4. Arabidopsis protein function prediction

imply that those proteins have a very specific function in flower development, as it could well be the case that many different TFs which are involved in various biological processes fulfill their functions via such proteins. Indeed, histone H3 is predicted to be involved in some other developmental processes as well (e.g. leaf morphogenesis).

In addition, two methyltransferases are predicted to be involved in the floral transition, one of which (PRMT6) is closely related to a histone-regulating methyltransferase (PRMT10) with a known function in regulating the floral transition (Niu et al., 2007). Note that for the floral transition in particular, an epigenetic mechanism is biologically meaningful as a way to bridge the temporal separation between induction of a flowering-competent state by e.g. vernalization and initiation of flowering in spring (Jung and Muller, 2009). TFs are important for regulating biological processes in general and flower development in particular. At a lower level, however, target genes with more specific molecular functions are obviously involved in those processes. One particular set of proteins predicted by our method are hydrolases (more specific, hydrolases that hydrolyze glycosyl compounds). For one of those (AT3G56310, a putative alpha-galactosidase), which is assigned to the process ‘positive regulation of flower development’, there is indeed supporting literature evidence (Rojo et al., 2003; Van Doorn and Woltering, 2008). Other predicted hydrolases include an alpha-galactosidase (positive regulation of flower development) and AT3G48700 which is expressed during petal differentiation and expansion stage according to TAIR. For the GO term 0048573 (photoperiodic control of flowering time) among others the hydrolases AtXTH17,18 and 19 are predicted. However, according to Osato et al. (2006) these are preferentially expressed in the roots, and there is evidence for a principal role for the AtXTH18 gene in primary root elongation. Hence this prediction seems unlikely. We now briefly discuss our predictions for two particular processes, the floral transition (GO: 0010228), and corolla development (GO:0048465). The floral transition refers to the transition from the vegetative to reproductive phase. The methyltransferase PRMT6, a methyltransferase-related protein (AT5G53920, ribosomal protein L11 methyltransferase-related) and protein AtBAG2 were all predicted for floral transition. For PRMT6, the closely related protein PRMT10 is indeed known to be involved in this process (reference). AtBAG2 is one of the BAG (Bcl-2-associated athanogene) proteins, which are plant homologs of mammalian regulators of apoptosis. These proteins regulate apoptotic-like processes associated with e.g. pathogen attack, abiotic stress, or plant development. For the two BAG family members AtBAG2 and AtBAG6, knock-outs have been shown to give early flowering (Doukhanina et al., 2006), which provides strong support for our predictions.

Another interesting prediction for floral transition is AT1G10320, which is an U2 snRNP auxiliary factor-related protein (it is predicted as well for
photoperiodic control of flowering time). This protein is involved in splicing regulation (Lorković et al., 2000) and evidence is mounting for a role of alternative splicing (AS) in the floral transition (Terzi and Simpson, 2008). In particular, some MADS domain proteins have AS variants with a putative role in this process. One example is FLM which via exon-skipping can form two different variants (Severing et al, manuscript in preparation). Corolla development refers to the development of the petals of a flower. Here, several members of the glutaredoxin family were predicted. As there are indeed indications for the involvement of glutaredoxin in petal development (Xing et al., 2005) this prediction seems reasonable.

Table 4.3: Selected GO terms for flowering and floral organ development

<table>
<thead>
<tr>
<th>GO term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0010080</td>
<td>regulation of floral meristem growth</td>
</tr>
<tr>
<td>GO:0060860</td>
<td>regulation of floral organ abscission</td>
</tr>
<tr>
<td>GO:0048460</td>
<td>flower formation</td>
</tr>
<tr>
<td>GO:0048439</td>
<td>flower morphogenesis</td>
</tr>
<tr>
<td>GO:0009908</td>
<td>flower development</td>
</tr>
<tr>
<td>GO:0010582</td>
<td>floral meristem determinacy</td>
</tr>
<tr>
<td>GO:0010451</td>
<td>floral meristem growth</td>
</tr>
<tr>
<td>GO:0048437</td>
<td>floral organ development</td>
</tr>
<tr>
<td>GO:0048438</td>
<td>floral whorl development</td>
</tr>
<tr>
<td>GO:0009910</td>
<td>negative regulation of flower development</td>
</tr>
<tr>
<td>GO:0009911</td>
<td>positive regulation of flower development</td>
</tr>
<tr>
<td>GO:0009909</td>
<td>regulation of flower development</td>
</tr>
<tr>
<td>GO:0010228</td>
<td>vegetative to reproductive phase transition</td>
</tr>
<tr>
<td>GO:0048573</td>
<td>photoperiodism, flowering</td>
</tr>
<tr>
<td>GO:0048574</td>
<td>long-day photoperiodism, flowering</td>
</tr>
<tr>
<td>GO:0048438</td>
<td>floral whorl development</td>
</tr>
<tr>
<td>GO:0048465</td>
<td>corolla development</td>
</tr>
<tr>
<td>GO:0048467</td>
<td>gynoecium development</td>
</tr>
<tr>
<td>GO:0048440</td>
<td>carpel development</td>
</tr>
<tr>
<td>GO:0048445</td>
<td>carpel morphogenesis</td>
</tr>
<tr>
<td>GO:0048481</td>
<td>ovule development</td>
</tr>
</tbody>
</table>

4.5 Discussion

In this study we apply BMRF, a computational method for protein function prediction, to the proteome of Arabidopsis thaliana. By integrating diverse
data sources (experimentally identified protein interactions, expression levels and sequence derived features) we predict 64,721 novel GO terms for 5,807 Arabidopsis proteins. Performance metrics such as Precision and Recall are estimated for each prediction. We show that our predictions are of high precision and may provide leads for the design of new hypothesis-driven experiments.

It is well-known that high throughput datasets such as those used in our study contain measurement errors. Taking this error into account, by incorporating the edge confidence values in the BMRF model, may lead to further improvement in the prediction performance. The coexpression values that we used in this study capture correlations between expression levels show in a wide range of biological conditions. However, some proteins may interact only in particular circumstances and therefore have correlated expressions only under those conditions. We plan to work on using such data more efficiently in the BMRF model. Further, the Gene Ontology annotation files we used do not contain all the available information concerning the functions of Arabidopsis proteins. Integration with additional sources of function information (e.g. by literature mining) may improve the prediction performance of BMRF.

We statistically evaluated the prediction performance of BMRF and compared our predictions with recent new annotations deposited in the Gene Ontology. From the total of 194 such new annotations, BMRF provided exact or more detailed function predictions for 14 cases and more general but related GO terms for 109 cases. Thus, for 63% of the novel annotations, BMRF was able to predict a relevant function, which is a highly significant result as judged on the basis on a permutation test. BMRF gave better predictions than two recently proposed integrative approaches as judged on the basis of the precision-recall curves for the new annotations. We further studied the predictions related to flowering processes and found several cases where our predictions are supported by the literature and may therefore provide information for further experimental validation.

BMRF is a computational method for function prediction that integrates large-scale datasets and transfers functional information between proteins that interact indirectly. These two properties make BMRF a very useful method for protein function prediction in the genomic era as shown here by the application to the Arabidopsis proteome. Based on the results presented here we expect that our method will also show its value for other plant and animal species.
Figure 4.2: Scatterplots showing the relative performance (AUC score) of different protein function prediction models. The performance was evaluated for 100 Gene Ontology terms using four sets of masked proteins: those that appear only in the PPI network (light blue) or only in the COEX network (dark blue), proteins that appear in both (red) and the full set of proteins (yellow). A-B. Integrated approach BMRF-Union against BMRF-PPI (A) and BMRF-COEX (B). For the majority of the cases the integrated approach performs better, not only for proteins that appear in both networks, but also for the network-specific proteins (light and dark blue). C. Comparison between the two network integration methods BMRF-UNION and BMRF-MULTI shows little difference in performance. D. Comparison of BMRF-UNION-DOMAINS with the BMRF-UNION and EN-DOMAINS. The performance of the fully integrated approach is significantly better compared to the other methods.
Chapter 4. Arabidopsis protein function prediction

Figure 4.3: Densities for the Precision (light blue) and Recall (red) from the list of predictions. The mass for the Recall lies in the region of 0.2 and larger. The precision is high, with its mode at 0.8.

Figure 4.4: Precision versus Recall curves for BMRF (blue), GO-AT (red, dashed) and Aranet (light green) using as validation set the newly annotated proteins (those deposited in GO database after 18th April 2010). The Precision level at any Recall is higher for BMRF than for the other two methods.
4.5. Discussion

Figure 4.5: Illustrations of three experimentally verified cases, where BMRF successfully recovered the exact function of proteins.  

A. MAC3A and MAC3B were predicted to be involved in “defense response to bacterium”. MAC3A has an experimentally determined interaction with CDC5 as well as one predicted by the interolog from Homo sapiens with PRL1. PRL1 is also predicted to interact with MAC3B by a Saccharomyces cerevisiae interolog. Both CDC5 and PRL1 are involved in “defense response to bacterium”.  

B. YUC8 was successfully predicted to be involved in “auxin biosynthetic process”. This protein does not interact with any proteins involved in this process. Still the prediction was based on the presence of two InterPro domains in its sequence.  

C. BIR1 is involved in “negative regulation of defense response” which is correctly predicted by BMRF through its coexpressions with four other proteins known to be involved in this process.
Chapter 4. Arabidopsis protein function prediction
Chapter 5

Evolutionary discrete optimization for Gene Ontology consistent protein function prediction: the FALCON algorithm

Abstract

Gene Ontology is a hierarchical vocabulary for the description of biological functions and locations, often employed by computational methods for protein function prediction. Due to the structure of the Gene Ontology, function predictions can be self contradictory. For example, a protein may be predicted to belong to a detailed functional class, but not in a broader class that, due to the vocabulary structure, includes the predicted one.

We present a novel discrete optimization algorithm that resolves such contradictions. The Gene Ontology is modeled as a discrete Bayesian Network. For any given input of GO term membership probabilities, the algorithm returns the most probable GO term assignments that are in accordance with the Gene Ontology structure. The optimization is done using the Differential Evolution algorithm. Performance is evaluated on simulated and real data from Arabidopsis showing improvement compared to other related approaches.

\[\text{Submitted for publication}\]
Introduction

Central aim of computational protein function prediction methods is to provide reliable and interpretable results, in order to be useful for the biological community. For this reason, prediction methods often make use of the Gene Ontology (GO) controlled vocabulary (Ashburner et al., 2000) to describe functional properties of proteins. GO terms are organized in three separate domains that describe different aspects of gene and protein function: Molecular Function (MF), Biological Process (BP) and Cellular Component (CC). Within each domain the terms are arranged in a Directed Acyclic Graph (DAG). Due to the hierarchical structure of the GO-DAG, a protein that is assigned to a particular term is by definition assigned to all of its predecessors, which are more general GO terms. On the other hand, if a protein does not perform a particular function, it is not assigned to the corresponding GO term, nor to any of the successors (more detailed terms) of that term. This constraint of the GO-DAG is referred to as the True Path Rule (TPR) and provides a framework to ensure that functional descriptions of proteins are not self-contradictory. Computational methods often neglect the TPR in their predictions, making their interpretations problematic. Taking the GO DAG (and thus TPR) into account in protein function prediction may lead to improvement of the performance and interpretation.

Violation of TPR can be described in a continuous or in a discrete manner. In the former, the probability (or confidence) of membership to a GO term does not decrease monotonically when moving from more general GO terms to the more detailed ones. Therefore the space of probability vectors (where a vector denotes the joint set of per-GO term probabilities of memberships) can be divided in two sets: one set \( C \) that contains the probability vectors that satisfy the monotonicity constraint and another set \( V \) that contains those that violate the constraint. The challenge from the continuous point of view is, given a vector \( V \) to find an optimal corresponding vector in \( C \), according to a criterion.

Obozinski et al. (2008) developed different ”reconciliation” approaches to infer consistent probability vectors from Support Vector Machines (SVM) outputs transformed to probabilities. Performance comparisons between methods based on Belief Propagation, Kullback-Leibner minimization and Isotonic Regression (IR), showed that the last outperformed the rest. In IR (Burdakov et al., 2006a,b) predictors are the ranks in the ordering of terms in the GO-DAG from general to detailed and the responses are the membership probabilities. The aim is to identify the probability vector that minimizes the squared error with the original input vector and that is monotonic for the predictors and thus belongs to \( C \).

In the discrete case, the interest is shifted from the probabilities of membership to the memberships themselves. The TPR violation can be evaluated by
checking whether all dependencies are satisfied or not. Given an inconsistent probability vector, the aim is to find the most probable set of GO assignments that do not violate TPR. The task of inferring the most probable latent binary vector given the input probabilities is a decoding problem, which is well-studied in information theory when the underlying structure of constraints has a tree-like structure (including chains). The Viterbi algorithm (Viterbi et al., 1967) (also called min-sum (MacKay, 2002)) performs such exact inference in tree-like structures. Standard hierarchical classification is not a suitable approach to this problem due to the the DAG structure of GO and multi-functionality of proteins (Kourmpetis et al., 2007). For instance, applying hierarchical classification to the DAG depicted in Figure 5.1A, one starts from the root ($x_1$) and moves to either $x_2$ or $x_3$. Regardless the outcome of this classification, it is not possible to give a positive prediction for $x_4$ without violating the TPR (since exactly one of its parents will not be predicted). However, Vens et al. (2008) proposed an hierarchical classification methodology adapted for the GO vocabulary. Exact inference in DAG structures is an NP-hard problem (Jiang et al., 2008) that can be performed by the Junction Tree algorithm (Lauritzen and Spiegelhalter, 1988) but the computational cost is intractable for the size of graphs such as the GO. Barutcuoglu et al. (2006) modeled the GO-DAG as a Bayesian Network and they combined SVM outputs per GO term in order to obtain GO assignments. In their case, exact inference was feasible because of the small size of the GO-DAG part used in the study (105 terms). Another related approach was developed by Sokolov and Ben-Hur (2010) where SVM classifiers for structured spaces, such as the Gene Ontology, were developed. Valentini (2010) developed an ensemble algorithm that transfers the decisions between base (GO term) classifiers according to the GO DAG structure. Jiang et al. (2008) first converted the GO DAG to a tree structure and then applied exact inference.

Here, we take a discrete approach to the problem of TPR violations and we develop an algorithm for the inference of most probable TPR consistent assignments using per-GO term probabilities as input. Our approach is based on the global optimization method of Differential Evolution (Storn and Price, 1997), which is adapted discrete space. We test our algorithm on small graphs of size 6 and 15 nodes, for which we can perform exact inference. We show that our algorithm consistently finds the correct optimal configuration. Further we evaluate the performance of the algorithm on probabilistic outputs of Bayesian Markov Random Fields (BMRF) (Kourmpetis et al., 2010a) as applied previously in Arabidopsis thaliana protein function prediction (Kourmpetis et al., 2010b). Our algorithm is applied to a graph that contains 1024 GO terms. We show that our algorithm improves the performance of the predictions compared to a supervised method used in a previous study.
Table 5.1: Parent-child relationship in a GO-DAG. C denotes consistent configuration while V denotes violating configuration.

<table>
<thead>
<tr>
<th>Parent</th>
<th>Child 1</th>
<th>Child 0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
<td>V</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>C</td>
</tr>
</tbody>
</table>

**Materials and Methods**

The GO is a vocabulary that describes the functions and locations of genes and its terms are arranged in a DAG structure, i.e. every node has zero, one or more parents and children. A protein can be assigned to one or multiple terms from each domain of GO (Kourmpetis et al., 2007). The TPR of the GO-DAG implies that when a protein is known to be assigned to a particular GO term, it should also be assigned to all ancestor terms. In contrast, when a protein is known not to be a member of a GO term, it should not be a member of any of all the successors of that term. By GO-DAG consistency we denote satisfaction of the TPR (also see Table 1). In terms of prediction, given a probably inconsistent vector of input probabilities, one has to find the most probable multiple and consistent GO-DAG paths that the protein has to be annotated to.

Naturally, methods that treat GO terms independently and neglect the DAG structure of the GO can make predictions that are inconsistent. In particular for probabilistic methods those inconsistencies may appear in the form of \( p_i > p_j \) in which the term \( j \) is an ancestor of term \( i \), and thus more general. In this study, we aim to find the most probable consistent GO term assignments, using such probability vectors as input. We first describe the general probabilistic setting, then derive two likelihood based objective functions and finally an evolutionary algorithm for the optimization.

**Bayesian Network modelling of GO**

Consider a Directed Acyclic Graph (DAG) \( G = (V, E) \) with nodes \( V \) (denoting the set of GO terms) and \( E \) directed edges (the set of parent-child relationships). Vector \( \theta \) denotes the input probability vector which is \( |V| \) - dimensional and \( x \) is the corresponding binary labeling, where \( x_g = 1 \) denotes membership for a particular protein to the \( g \)-th GO term in \( V \) GO term.

We model the GO-DAG as a Bayesian Network, with density for \( x \):\[ p(x \mid G, \theta) = \prod_{g=1}^{\left|V\right|} p(x_g \mid x_{pa(g)}) \]
Table 5.2: Conditional Probability Tables, under the DAG constraints.

<table>
<thead>
<tr>
<th>( x_g )</th>
<th>( \prod x_{pa(g)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 ( \theta_g )</td>
</tr>
<tr>
<td>0 ( 1 - \theta_g )</td>
<td>1</td>
</tr>
</tbody>
</table>

where \( pa(g) \) denotes the parent set of node \( g \) and \( x_{pa(g)} \) the set of labels that correspond to those parents.

The probability \( p(x_g \mid x_{pa(g)}) \), under the DAG constraints, is given using the Conditional Probability Table (CPT) of Table 2. The table shows that when \( \min(x_{pa(g)}) = 0 \) (i.e at least one of the parents has label 0) then \( x_g = 0 \) with probability 1. Otherwise \( x_g = 1 \) with probability \( \theta_g \) and \( x_g = 0 \) with probability \( 1 - \theta_g \). Note that all inconsistent labelings have zero probability.

Given equation 5.1 and conditional probability tables with parameters \( = (\theta_1, \cdots, \theta_{|V|}) \), one wishes to identify the most probable labelings vector \( x \). There are two challenges in this. The first is how to choose the parameter vector \( \theta \), discussed in this section, and the second is how to search for the most probable labelings vector \( x \), which is discussed in the next section.

Many function assignment methods act per GO term and disregard the GO hierarchy. Such methods - and BMRF is one of them- do not return the conditional probabilities in the sense of equation 5.1. The memberships probabilities that they return are perhaps best viewed as marginal probabilities, i.e. summed over all configurations for GO terms other than the specific term \( g \). We might have tried to retrieve \( \theta \) from the relation between marginal and conditional probabilities, but this is certainly not an easy way. We attempted other ways.
Methods such as BMRF return low probabilities for detailed GO terms and high ones for general terms. Prioritization of the proteins in a particular GO term can then be achieved by simply sorting them. By contrast, prioritization of GO terms for a particular protein (a more important task) is not simple as the sets of probabilities for different GO terms are not directly comparable. To make them comparable, the probabilities need to be calibrated. We derive two approaches.

The first, called DeltaL is based on the maximization of the difference of the likelihood and prior probability of the labelings as they are defined in eq (5.1). The second, called LogitR, is based on explicit calibration of the input probability vector.

For DeltaL, we modify the objective function of equation (5.1) by incorporating the prior probabilities of membership $\theta^*$. We use the log ratio between the input probability and the prior, $\log(p(x_g | \theta_g)/p(x_g | \theta^*_g))$ as score function for the labeling of the $g$-th GO term. For $x_g = 1$ the score is equal to $\log(\theta_g/\theta^*_g)$, while for $x_g = 0$ it takes the value of $\log((1-\theta_g)/(1-\theta^*_g))$. When $\theta_g > \theta^*_g$ then $x_g = 1$ maximizes the function. In the opposite case $x_g = 0$ gives the maximum. The extended function is given by the difference of the log likelihoods:

$$\Delta L(X; \theta, \theta^*) = \sum_{g=1}^{V} \log \frac{p(x_g | \theta)}{p(x_g | \theta^*)},$$ (5.2)

giving

$$\Delta L(X; \theta, \theta^*) = \log(p(X | \theta)) - \log(p(X | \theta^*)).$$ (5.3)

Note that when the input probabilities are very close to the priors, the objective function of DeltaL becomes multimodal.

In LogitR optimization of eq 5.1 is performed on a calibrated input probability vector. The calibration is done as follows:

$$\logit(\theta_{cg}) = \logit(\theta^*_g) + \alpha(\logit(\theta_g) - \logit(\theta^*_g))$$ (5.4)

where $\theta_{cg}$ is the calibrated probability for node $g$ and can be calculated using the inverse of the logistic transformation, $\theta^*_g$ is the prior probability of membership for node $g$ and $\alpha$ a slope parameter. In this objective function, when the posterior probability $\theta = \theta^*$ then the probability of membership is equal to $\theta^*$ (Fig 5.2A). As $\theta$ deviates from the prior, the calibrated probability $\theta_{cg}$ changes according to the logistic function given $\theta_g$ and $\alpha$ (Fig 5.2B). The $\alpha$ parameter was tuned using Saccharomyces cerevisiae data. In particular, for a range of $\alpha = 1, 1.5, 2.0, 2.5, 3.0$ the logitR algorithm was applied taking as input BMRF based predictions obtained from a previous study (Kourmpetis et al., 2010a) before March 2010. The evaluation set consisted of 327 proteins that were annotated after March 2010, according to the GO annotation file of
Figure 5.2: Calibration of posterior probabilities using $\alpha = 2$. A. Calibrated probabilities (y-axis) against the posterior probabilities (x-axis) when the prior is equal to 0.2. B. Image plot, for the entire range of prior and posterior probabilities. The colors denote the calibrated probabilities.

July 2011. The relevant part of GO DAG contained 423 terms from Biological Process. For each value of $\alpha$ the prediction performance was measured using the F-score, which is the harmonic mean of precision and recall. The largest F-score was obtained for $\alpha = 2$ and therefore we fixed $\alpha$ to that value.

**Optimization by Differential Evolution: The FALCON algorithm**

The DeltaL in eq(5.3) and LogitR in eq(5.4) approaches do not involve directly the TPR constraints. We develop an optimization algorithm inspired from Differential Evolution (DE) (Storn and Price, 1997) that by construction is restricted to the sub space of consistent labelings. We call our algorithm Functional Annotation with Labeling CONsistency (FALCON). In general, DE works by evolving a population of candidate solutions to explore the search space and retrieve the maximum. Because DE is derivative free, it has appealing global optimization properties. Also, it is suitable for optimization in discrete spaces (like the labelings space in our problem).

Our optimization algorithm is based on the generation and evolution of a population of consistent labelings. The graph representation of the labelings is helpful to explain how the algorithm works. Given the graph $G$ and its corresponding labeling $X$, we define a reduced graph $R = (V_R, E_R)$ which
Figure 5.3: Examples of graph (upper row) and logical (lower row) operations, using the DAG structure of Figure 1A.

Table 5.3: Logical operations OR and AND for all the combinations of labels. In this example X1 and X2 are univariate.

<table>
<thead>
<tr>
<th>x1</th>
<th>x2</th>
<th>OR (x1 ∨ x2)</th>
<th>AND (x1 ∧ x2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

contains the nodes with corresponding labels x = 1. If X is consistent, in the TPR sense, R will be a connected sub-graph of G and maintaining the original structure for the VR nodes. Consider two labelings L1, L2 and their graphs R1, R2 respectively is given in figure 5.3. Graph union R1 ∪ R2 gives the expanded graph R\text{union} = (V_{R1} \cup V_{R2}, E_{R1} \cup E_{R2}), while graph intersection R1 ∩ R2, gives the contracted one R\text{int} = (V_{R1} \cap V_{R2}, E_{R1} \cap E_{R2}). The nodes that will be included in the resulting graph are given by set operations (i.e (V_{R1} \cup V_{R2}) and \(V_{R1} \cap V_{R2}\) respectively), but also equivalently by performing logical OR (for union), \(X_1 ∨ X_2\), and logical AND (for intersection), \(X_1 ∧ X_2\) operations on the labelings directly. Table 3 and Figure 5.3 illustrate those operations.

Those operations between consistent graphs (labelings) result in consistent graphs (labelings) as well, because the edge set of the last is the union or the intersection of the operands and therefore a particular edge has to pre-exist in at least one of the operands without violating the TPR. This property can be seen as follows: For any parent-child pair of nodes there are three types of
configurations that are consistent (Table 1). Graph union and intersection between any combination of those pairs leads to locally consistent labeling. This holds for all the parent-child pairs, so it holds for the full labeling. Therefore the outcome of graph union and intersection will be consistent as well. Further, operations between more than two labelings will be consistent as well due to the associativity property. In our algorithm, a population is initialized with consistent labelings and evolved exploiting the graph-union and graph-intersection operations between individuals. Through the generations, all the constructed labelings will be consistent due to the abovementioned property. In our optimization problem we used four strategies to propose a new candidate solution (labelings) for the $i$-th graph $R_i$:

- **S1:** Global Union $R_{Cand} = R_1 \lor R_2 \lor e$
- **S2:** Global Intersection $R_{Cand} = R_1 \land R_2 \lor e$
- **S3:** Local Union $R_{Cand} = R_i \lor R_1 \lor R_2 \lor e$
- **S4:** Local Intersection $R_{Cand} = R_i \land R_1 \lor R_2 \lor e$

The first two types are called global because they do not involve $R_i$ while the latter two are local moves. Graph $e$ is a random sub graph of the original full graph (i.e. GO-DAG). It is constructed by picking a random node and the DAG and assigning that node and its ancestors to state 1. $e$ ensures that all consistent configurations can be eventually proposed and reached.

The scheme of the FALCON algorithm is as follows:

**Initialize** Population $R$ of size $2 | V |$ by picking random consistent vectors (see below):

while Convergence or Maximum generations reached do
  for $i = 1$ to PopSize do
    Sample two labelings from the population $R_1, R_2 \neq R_i$
    Construct $R_{Cand}$ using the a randomly picked strategy $S1, S2, S3, S4$
    if $f(R_{Cand}) > f(R_i)$ then, $R_i := R_{Cand}$
  end for
end while

Initialization of the population for DeltaL is done by sampling GO terms to be in state 1 according to their individual score (log ratio of the input and prior probability), while LogitR by sampling from the binomial distribution with probability equal to the calibrated one. In both cases the nodes were up-propagated in order to construct a consistent labeling. The computation was terminated after 10,000 generations or after reaching a plateau (i.e there is no improvement in the objective function for 100 generations). Finally we point that a valid Markov Chain Monte Carlo algorithm cannot be derived using those proposal strategies because they do not represent reversible moves. The
Chapter 5. Gene Ontology consistent protein function prediction

bitwise exclusive OR move proposed by Sterns in (Strens, 2003) is reversible but does not lead to consistent labelings. Implementation of the algorithm was done in R language for Statistical Computing and using the igraph R package (Csardi and Nepusz, 2006).

Performance Evaluation

We evaluated the performance of the FALCON algorithm on the DeltaL and LogitR objective function using Precision, Recall and F-score. Precision is defined as the percentage of correct GO terms in the list of the GO predictions. Recall is equal to the percentage of the GO assignments that were identified and F-score is the geometric mean of the Precision and Recall.

Simulated data

First, we tested the capability of FALCON to retrieve the most probable graph using the full graphs in Figure 5.1 with hundred simulated probability vectors. The first contains six nodes and the second fifteen. Because of the small size of the graphs exhaustive search of the most probable labeling was computationally tractable. We generated a hundred random probability vectors, by sampling probabilities for each node from the uniform distribution. Then we identified the most probable labeling for each simulated probability vector and the one returned by FALCON using equation (5.1) as objective function. Performance measures were calculated by comparing the vectors obtained by FALCON with the most probable ones as calculated from the exhaustive search.

Real data

The performance of FALCON was further evaluated using as input the GO membership probabilities of the Arabidopsis proteins as computed by BMRF in (Kourmpetis et al., 2010b). This method provides membership probabilities per GO term independently. We constructed two evaluation datasets from those data. First, we randomly picked 100 Arabidopsis proteins that were already annotated at the time of computing the BMRF posterior probabilities. One constraint was that they should have at least fifty annotations (after up-propagating their original annotations). In this way we ensured that they were annotated in rather detailed GO terms, and therefore the attempt to get GO-DAG consistent predictions would be sensible. Although these proteins had a fixed labeling in the computations, BMRF can calculate membership probabilities for them, by reconstitution, i.e as if they were unknown. The second dataset consisted of 387 proteins that were annotated later than the date of the BMRF computations. Thus, at the time of the computation the proteins were not annotated. We used this second set of proteins to evaluate
the performance of FALCON in realistic conditions. In addition, we obtained a further list of predictions using the supervised approach proposed in (Kourmpetis et al., 2010b). In this approach, from the posterior probabilities of the annotated proteins, an F-score based optimal threshold was calculated per GO term. Using this approach, called maxF, we derived a set of predictions for each evaluation dataset. Note that those lists are not guaranteed to be GO-DAG consistent.

Results and Discussion

Performance of FALCON on the simulated dataset

We initially evaluated the performance of FALCON in the two small graphs of Figure 5.1. For each graph we simulated 100 probability vectors by drawing from the uniform. Because the graphs are small we could identify the most probable labeling by exhaustive searching. Using equation 5.1 as objective function and without using prior probabilities LogitR retrieved the 98/100 of the labelings for the 6-node graph and 92/100 of the labelings for the 15-node graph. The DeltaL approach also retrieved 98/100 labelings for the small graph (using priors = 0.5 for all the nodes).

Performance using the two evaluation datasets

The true labelings of the proteins included in the evaluation datasets were known, so we were able to calculate performance metrics. Table 4 shows mean performance measures per protein and per GO term. The LogitR approach leads to the highest F-score, while maxF comes second and DeltaL comes last. We see that all three of them follow the precision-recall trade off (i.e for larger precision there is lower recall and vice versa) with maxF being more precise but with reduced recall and the opposite for DeltaL. LogitR stays in the middle. In Figure 4 performance measures are shown in relation to the GO term level of detail and to the number of GO assignments per protein. Using the F-score to summarize the performance (Fig 5.4A) we see that for the GO terms that are rather general DeltaL (yellow) performs well, but for the more detailed ones its performance deteriorates. On the other hand LogitR and maxF perform well in detailed GO terms. In terms of Precision (Fig 5.4B) and Recall (Fig 5.4C), the latter methods have similar performance but LogitR performs slightly better. On the other hand DeltaL predicts large numbers of terms and therefore shows high recall but low precision, in particular for the detailed GO terms that are of real interest. Comparing the performance of predicting the assignments per protein (Fig 5.4D-F), the LogitR approach performs consistently better than the others in terms of the proteins that need or small either large number of
GO terms to be functionally described.

We further evaluated the performance of our approaches using a set of proteins that were annotated after obtaining the BMRF predictions (Table 5). From the total of 387 newly annotated proteins, maxF returned predictions for 84 of them, DeltaL for 328 proteins and LogitR for 147 proteins. Again, maxF and DeltaL show comparable performance while logitL returned an improved list in terms of F score. Further, the higher recall rates of DeltaL tend to give longer lists of predictions. Importantly however, DeltaL and LogitR return predictions that are consistent with GO-DAG and are therefore preferred because such predictions are biologically interpretable.

**Table 5.4: Mean Performance measures for the evaluation dataset consisting of 100 Arabidopsis proteins**

<table>
<thead>
<tr>
<th></th>
<th>LogitR</th>
<th>DeltaL</th>
<th>maxF</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Per Protein</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>0.79</td>
<td>0.27</td>
<td>0.85</td>
</tr>
<tr>
<td>Recall</td>
<td>0.55</td>
<td>0.90</td>
<td>0.46</td>
</tr>
<tr>
<td>F-score</td>
<td>0.63</td>
<td>0.41</td>
<td>0.56</td>
</tr>
<tr>
<td><strong>Per GO term</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>0.70</td>
<td>0.25</td>
<td>0.81</td>
</tr>
<tr>
<td>Recall</td>
<td>0.50</td>
<td>0.80</td>
<td>0.40</td>
</tr>
<tr>
<td>F-score</td>
<td>0.70</td>
<td>0.44</td>
<td>0.66</td>
</tr>
</tbody>
</table>

**Table 5.5: Mean Performance measures for the newly annotated proteins. Every method predicted different numbers of proteins. The number of proteins returned (out of the total 387) are given in the last column of the table.**

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
<th>proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>maxF</td>
<td>0.34</td>
<td>0.35</td>
<td>0.23</td>
<td>84</td>
</tr>
<tr>
<td>DeltaL</td>
<td>0.08</td>
<td>0.58</td>
<td>0.19</td>
<td>328</td>
</tr>
<tr>
<td>LogitR</td>
<td>0.26</td>
<td>0.50</td>
<td>0.27</td>
<td>147</td>
</tr>
</tbody>
</table>

The FALCON algorithm shows appealing performance properties for inference of GO terms in large graphs. To the best of our knowledge there are no other algorithms to perform this task at such scale. Using the LogitR as objective function, FALCON outperforms maxF. The predictions of FALCON are GO-DAG consistent and therefore biologically much easier interpreted by the curators of protein function annotations. Further, FALCON is unsupervised i.e
Figure 5.4: Performance on the evaluation dataset for the methods LogitR (red), DeltaL (blue), maxF (yellow). **ABC.** F-score, Precision and Recall scores for different size of GO terms. **DEF.** The same scores against the number of annotations per protein. Smoothed splines in each subplot show fitted generalized additive models and using the R function smooth.spline. Because a large number of points in the scatterplot coincided, we performed jittering by adding a small error term to each value $e \sim N(0, 10^{-4})$, in order to make the maximum number of points visible.

it can infer the optimal GO term assignment using only the probability vectors and the prior probabilities per GO term, which can be easily computed from a set of predictions or using external Gene Ontology information. In contrast, maxF needs a training set in order to obtain the optimal cutoffs. In this study both approaches were applicable but the FALCON algorithm is expected to have broader applicability.
Chapter 6

General discussion

6.1 The usage of Gene Ontology in automated protein function prediction

In this thesis different aspects of computational protein function prediction were approached from a classification perspective. The central aim was to develop a computational method for protein function prediction that integrates diverse data sources. Each function was represented by a class and for this reason the use of a rich set of terms to describe function was necessary. The controlled vocabulary provided by the Gene Ontology (GO) was found to be suitable for this purpose, since it is actively maintained and covers a wide range of functional descriptions. As of January 2011, the GO contained almost 44,000 terms divided in three branches that describe aspects of function as “Molecular Functions”, “Biological Processes” and “Cellular Compartment”. In Chapter 2 of this thesis, it was shown that it is common for proteins to perform multiple distinct molecular functions, to be involved in different cellular processes and to localize to different compartments of the cell. Therefore the full description of the function of a protein often requires multiple terms from each category, either by selecting them manually or by predicting them using a computational classification method. Further, the GO terms in each category are structured in a hierarchy, allowing the level of detail for the description to be adjusted to the evidence available for the protein of interest. For example, a protein may be known to be involved in a certain “metabolic process”, but in case of collection of further evidence the function description could be made more precise, e.g. by describing it with a GO term for “hormone catabolic process”. The Directed Acyclic Graph structure of the GO implies the True Path Rule (TPR): if a protein is assigned to a particular GO term, then it should also be assigned to all of the ancestors of that GO term (less detailed terms). The implication
of this property for the functional annotation of a protein is that sub-graphs of the GO hierarchy have to be used to consistently describe the function and multiple sub-graphs need to be used to fully describe the protein function. Given that the TPR is satisfied, usage of the deeper (more detailed) GO terms suffices for the function description. Statistical prediction methodology for protein function prediction needs to account for the correct usage of the GO vocabulary in order to make the predictions useful and interpretable by the biological community. For example, a method may predict membership to a particular GO term but in addition the non-membership for one or more of the ancestors of this term. GO DAG consistent predictions are more easily interpreted by the biological community and therefore more useful.

It is also important for a computational method to allow for the prediction of multiple functions. For example, a protein that is involved in a given disease pathway may have additional and unrelated biological roles. A drug that inhibits the function of that protein, besides interfering in the disease, is likely to cause multiple side-effects. Prediction of all different functions of the protein is therefore necessary for the development of an effective treatment. A prediction method that cannot handle this multiplicity will not provide the relevant leads necessary for the design of drugs for such a treatment.

The study described in Chapter 2 showed that the fraction of proteins assigned to multiple distinct functions is much smaller in *A. thaliana* than in other well studied model species. Comparison of the domain composition of the proteome of Arabidopsis to those of the other species showed that there is no biological factor that can account for this difference. Further, it was shown by log-ratio Principal Components Analysis that the main part of the annotations for *A. thaliana*, unlike the other species, were assigned by “Reviewed Computational Analysis”, which refers to human-reviewed computational procedures for inferring the protein functions. We concluded from this that systematic differences in the annotation strategies used for the various model species and by different research communities provides the most probable explanation for the different annotation patterns. This conclusion points out the importance of standardization in the practice of function annotation and in the development of effective automated procedures. Computational methods for protein function prediction can consistently apply standardized annotation strategies. Even if the annotation model changes, the annotations can be easily updated or reconsidered. Development of standards and usage of proper computational models will therefore help the community to resolve discrepancies in their annotation strategies.
6.2 Probabilistic protein function prediction by BMRF

Protein annotation, performed manually or computationally, involves two main practices. The first considers features of the proteins themselves, such as sequences, domains or expression levels in particular conditions. The second is based on measuring various sorts of associations with other proteins that are already functionally characterized, including similarity in pairwise or multiple alignments, indications for physical or genetic interactions, co-localization and others. Computational methodology for protein function prediction can be made most powerful in selecting and combining evidence from diverse sources of information (sequence and domain databases, high throughput experiments, etc.) and in automating the procedure of function annotation. In Chapters 3 and 4 of the thesis, Markov Random Field (MRF) based methodology was developed for prediction of protein function from single and multiple data sources. The MRF framework is suitable for the modeling of inter-dependent variables, with variables represented as nodes of undirected graphs or networks. For function prediction with BMRF each variable (node) is binary and represents the state of a protein annotation i.e. whether or not a protein performs a given function. The undirected graph structure is defined by the set of interactions between the proteins. There are three types of interactions that reflect the states of the linked proteins: i) those in which both proteins perform the function, ii) those in which one of them performs the function and iii) those in which none of them performs the function. One parameter is assigned to each type of interaction which is determined by the states of the participating proteins. The sum of all the terms across all the interactions gives the objective function or what is called the “energy” of the MRF. This energy can also be written as a Gibbs density corresponding to the joint probability distribution of the annotations in the protein network. Due to the Markov property the conditional distribution of each annotation depends on the annotations of the direct neighbors in the network. Besag (Besag, 1986) and (Geman and Geman, 1984) showed that sampling from the joint distribution is feasible through Gibbs sampling. Since some of the proteins have known annotations (states), the sampling is done conditional on those states. Furthermore, in the original MRF-based method for protein function prediction developed by Deng et al. (2004), the parameters of conditional distributions are estimated by logistic regression, with the response variable being the state of a protein and predictors representing the states of the proteins in the neighborhood. The parameters are estimated by maximizing the Pseudo-Likelihood assuming that the observations are independent, while they are not due to the complex inter-dependencies between the variables (protein states) which serve both as predictors and response variables (auto-regression model). The proteins lacking annotations are
not taken into account in the parameter estimation step. In the study reported in Chapter 3, it was found that this parameter estimation strategy is not accurate when a large part of the proteins is unannotated, because those proteins introduce uncertainty to the data matrix that is used for the fitting. This issue is highly relevant for the available genome-wide annotations of most species, for which it is common that a large part of their proteins lack an annotation. The solution for sparsely annotated networks proposed in Chapter 3 makes inferences by sampling from the joint posterior density of the protein annotations and model parameters. This step was done by employing an adaptive Markov Chain Monte Carlo algorithm, the Differential Evolution Markov Chain (DEMC) (Ter Braak, 2006; Ter Braak and Vrugt, 2008). Our proposed approach, called “Bayesian MRF” (BMRF) (Kourmpetis et al., 2010a) leads to improvement of model parameter estimates compared to the original method of (Deng et al., 2004). Consequently, the prediction performance, as measured by the Area Under the Curve, was significantly improved compared to the original method and as evaluated by cross validation using \textit{S. cerevisiae} protein interaction data. BMRF was indeed shown to be especially suitable for datasets where a large fraction of the proteins is unannotated. The standard approach to tackle the missing data problem, as in sparsely annotated networks, would involve the Expectation-Maximization (EM) framework. However, in this case computing the expectation of the joint density (E-step) is intractable. Therefore one would have to rely on a stochastic (or Monte-Carlo) EM algorithm, but again numerical computation of the E-step in each iteration would not be practical, since several iterations will be required for converge for each EM iteration. Besag (Besag, 1986) proposed the idea of a two step iterative estimation of the parameters by maximizing the Pseudo-likelihood and maximizing the MRF objective function for finding the annotation states and this approach could also be extended to a kind of stochastic EM Algorithm, i.e by iteratively drawing a sample from the annotations density and then updating the model parameters by fitting a logistic regression model, given each sample. However, given the large fraction of unannotated proteins, sampling from the full joint density of the parameters to make inferences is the preferable approach. In the BMRF model, the annotation states are sampled by Gibbs sampling, with the model parameters by the DEMC. The latter has been shown to be superior compared to other Markov Chain Monte Carlo settings attempted during the thesis study such as sampling the parameter vectors from the posterior Normal density with mean equal to the Maximum Pseudo-likelihood estimates and variance equal to the inverse Fischer information matrix. Those two parameters were obtained by fitting the logistic regression model after the states for all the unannotated proteins were drawn by sequential sampling of each state conditioned on the rest (i.e Gibbs sampling). Updating the posterior mean and variance involves the use of a fitting algorithm that is normally used for logis-
6.2. Probabilistic protein function prediction by BMRF

tic regression, such as the Iterative Reweighted Least Squares method. The convergence of such derivative based algorithm can be problematic for datasets that are extremely unbalanced, i.e those that include only a limited number of proteins performing the function. MCMC sampling therefore becomes computationally unstable unless there are further adjustments through the proposal distribution. Tuning the parameters for the proposal distribution (either local, global or independence sampler) makes the generalization of the method difficult, since the algorithm was performing well for some datasets but for others numerical difficulties were encountered. In contrast, DEMC adapts the mean and variance of the proposal density making the BMRF methodology easier to apply to a variety of datasets. This illustrates the importance of use of adaptive MCMC algorithms in complex problems.

There are other popular methods that deal with network data. One of the most popular is based on a two-step procedure in which the network first is clustered, followed by a one-tailed Fisher’s exact test for GO term enrichment in the clusters. If overrepresented GO terms are found then all the proteins in the cluster are assigned to these. This procedure is straightforward to apply. However, there are two points to be considered. First, most network clustering algorithms assign each protein to exactly one cluster. As discussed above, it is well known that proteins often have multiple diverse biological roles and this multiplicity should be represented by membership of multiple clusters. This is a limitation of the approaches based on network clustering, although fuzzy community detection in complex networks is an active field of research (ter Braak et al., 2009). In contrast, the BMRF type of approach does not rely on clustering to make predictions for proteins. Second, hypothesis testing for GO term enrichment is actually done in the presence of missing data. In particular for the Fisher exact test not all the data are observed since some of the proteins are still unannotated. The methods that use this approach neglect this fact and apply the test treating the data as fully observable. The effect on the statistical conclusions (i.e. the outcome of the test) depends on the fraction of the unannotated proteins in the network and also on their underlying model (or pattern) of missing values. One procedure to deal with missing data is by data imputation. Here, the variables (i.e. the annotations of the proteins) are highly interdependent due to the network structure and therefore single imputation is not suitable. Instead, multiple imputation that is related to data augmentation is a better alternative. From a Bayesian perspective, multiple imputation is done by sampling from the joint posterior distribution of the model parameters and the data and finally drawing conclusions from those samples. Although BMRF does not pose the protein function prediction problem as a hypothesis testing problem, it employs a model that can be seen as multiple imputation since it samples from the full joint distribution. In Chapter 4, the prediction performance of BMRF was compared to two other methods that make use
Chapter 6. General discussion

of the GO Enrichment approaches. BMRF showed to perform better than the other two methods. One explanation for this outcome is the imputation strategy that BMRF employs.

In Chapter 3 the performance of BMRF was compared with another protein function prediction method called Kernel Logistic Regression (KLR). This method propagates the function information from the annotated proteins through the network via process that is similar to the diffusion of heat sources in a network. The diffusion is computed by exponentiation of the heat matrix i.e., a matrix that contains the initial heats in each node. For protein function prediction the probability that a protein performs a function depends on its proximity to the heat sources which represent the proteins performing the function. In KLR, the first step is to compute the diffused matrix by exponentiation of the original network adjacency matrix. In the diffused matrix the neighborhoods of the proteins are expanded and therefore information from proteins not directly linked is used for the function prediction step. Diffusion processes are related to random walks in graphs, which give a Markov process view on this operation. The diffusion matrix can also be interpreted as the average occupancy for each position during a random walk in the network. The Gibbs sampling from the MRF can also be seen as a random walk in the network. Therefore given this underlying relationship between the posterior simulation and estimation by exponentiation, the similarity in prediction performance between BMRF and KLR is not too surprising. However, an important difference between the two approaches is that the diffusion operation in KLR can account for weighted edges in the network, while the current version of BMRF uses a binary representation. Incorporation of weighted edges in the BMRF model is a topic for further investigation and could be achieved, for example, by selecting a proper weight for the edges or by sampling network topologies during the BMRF run. The KLR method, on the other hand, has a serious drawback in that fast exponentiation of large matrices, corresponding to large networks, is difficult and requires special implementation and computational capacity. The Gibbs sampling employed in BMRF uses a more transparent probabilistic formulation that does not need the diffusion kernel parameter selection and it has a computational complexity that grows only linearly with the number of proteins in the network (Chapter 3). This makes it more easy to implement and apply to network sizes that are relevant in the field of protein function prediction. A remaining problem with the BMRF method is that the posterior simulation from the MRF-based probability density using the Gibbs sampler may cause difficulties with convergence, although in the studies using yeast (Chapter 3) and Arabidopsis (Chapter 4) data, such difficulties were not encountered.
6.3 Integration of diverse data sources

There is no doubt that the best view on the function of a protein can be obtained only by looking at various types of data. The first step in this procedure is to identify the pieces of information from various sources that are related with the function of interest and then to weight each piece of evidence. In Chapter 4, the BMRF model was extended to integrate multiple types of information, i.e., data based on sequence domains, protein-protein interactions, and co-expression networks. These three types of information are important for integration since the sequences contain mainly information on the Molecular Function of the protein, while the interactions between them contain information on the Biological Processes. The protein sequences were first transformed to binary domain matrices, indicating which InterPro signatures each protein contains. In order to identify a set of signatures related to each function, the annotation vector was regressed to the domain matrix using penalized regression. We used Elastic Nets which provides a penalty that is a linear combination of the Lasso and Ridge regression penalties. After fitting that model, the probability of each protein to perform a function was estimated from its domain composition. The BMRF model remained the same for the network data, but assigned a different set of parameters to each network. In this particular study, it was shown that the more parsimonious model of using the same parameters for the different networks did not lead to deterioration of the performance. The probability that a protein performs a function was estimated through the linear combination of the EN and network-based information. The model parameters that involve updating of the relative weight of each data type were automatically estimated in the BMRF execution. In the study described in Chapter 4, it was shown that integration of different data types leads to significant improvement of the prediction performance.

In future studies, larger numbers of data sources may be needed to be combined. The parameters per data sources do not need to be either equal or completely independent. In Bayesian analysis they may be related through a prior on their distance. Such an extension of BMRF would be of interest since it may have impact on the prediction performance and also would minimize the need for tuning of parameters.

Next generation genomics provides a means to build a multi-level perspective on biological functions. Examples of datasets that Next Generation Sequencing (NGS) technologies provide are sequences, sequence variation (i.e. DNA seq), transcriptomics (RNA-seq), epigenetic information (DNA methylation), and DNA-protein interactions (ChIP-seq). Integrated approaches for analyzing those types of data are necessary and investigation on how to adjust BMRF to use such NGS data is a promising direction of research.
6.4 BMRF and Gene Ontology consistent protein function prediction

The BMRF method provides probability outputs denoting the level of confidence for a protein to be assigned to a particular GO term. This operation is done for each GO term independently. Predictions from BMRF solely may be inconsistent in the sense of the GO DAG structure. We developed a post-processing method, called FALCON, that ensures that the predictions are consistent. This method identifies the most probable consistent assignment of GO terms given the per GO term membership probabilities. An evolutionary optimization method (Differential Evolution) was used to tackle this problem. Tests with simulation and real data showed that this algorithm identifies the most probable consistent labeling and further it improves the prediction performance (in terms of precision and recall) compared to predictions that do not take into account the GO DAG structure. This provides interpretable and more accurate predictions. It is still an open question whether this method can be incorporated into BMRF, so that predictions are made consistent directly.

BMRF was shown to be an effective method for protein function prediction by integrating heterogeneous information. Combined with FALCON, it will provide informative and accurate predictions using the GO vocabulary.

The challenge for computational protein function prediction methods is to make accurate predictions of detailed function descriptions. The BMRF-based predictions made in this study use GO terms that have at least 20 proteins annotated to them. Although those GO terms are reasonably detailed, there is a significant number of GO term that are even more detailed. For more sparse terms, sampling of the BMRF model parameters becomes problematic because the dataset becomes extremely unbalanced with regard to the composition of annotated and unannotated proteins in the term. One way to overcome this difficulty is to use the GO DAG hierarchy inside the BMRF model, so that the more detailed terms use the information from their ancestors. In this case the model parameters across different GO terms are not treated independently but they would be connected through a prior on their difference. Also, the True Path Rule may be implied directly during the MCMC sampling so that posterior probabilities are consistent with the GO DAG, in the sense that they are monotonically decreasing along a path from the most general to the most detailed GO term. This provides an alternative approach to the FALCON algorithm which is now used as a post processing step and also provides consistent but binary predictions.

Already in their present form, BMRF and FALCON are practically useful for protein function prediction, able to handle multiple, large, functional genomics datasets and performing at least as well as other state of the art methods in this field.
Summary

Accurate, comprehensive annotation of genes and their products is essential for the understanding of biological processes. Experimental annotation procedures are impractical since they are laborious, expensive and time consuming. This thesis deals with the development and the application of a computational protein function prediction method that integrates heterogeneous data sources.

In chapter 1 the two main approaches for protein function prediction are introduced. The first is based on identifying the set of features that proteins performing a particular function have in common. Such features can be based for instance on their sequences or the level of expression of their related genes in particular conditions. The second type of approaches is based on the propagation on the function information between proteins. BLAST based annotations or guilt by association based methods belong to this category. Chapter 1 further introduces the Gene Ontology (GO) that is a widely used controlled vocabulary for protein function annotation. GO is divided in three non overlapping branches i.e Molecular Function, Biological Process and Cellular Component. This vocabulary is extended enough to allow the accurate description of the enzymatic function of a protein, the process that it is involved in and the cellular compartment that the protein is located in.

In chapter 2 the usage of GO is studied using data from seven well studied species. Obtaining insight on how this vocabulary is used in practice is very useful for the development of proper computational function prediction methods. Chapter 2 shows that it is very common that proteins have multiple distinct roles in the cell and in different places. This fact implies that prediction methods should be able to assign multiple independent functions to each protein. Further, it was shown that annotation strategies differ considerably between species resulting in different annotation patterns. That was indicated by the fact that proteins with multiple annotation in *Arabidopsis thaliana* are much less common than in the other species of the study and without any underlying biological reason. Formal rules for the assignment of GO terms would make the function annotation of different species more comparable.

Chapter 3 deals with the development of a probabilistic method for pro-
tein function prediction. This method is based on Markov Random Fields that allows the modeling of inter-dependent variables. When those variables are represented as nodes and their dependencies as edges, an undirected graph (or network) is formed. Protein association networks are constructed based on protein similarity, their interactions or the co-regulated expression of their relevant genes. In such networks, a node represents a protein while an edge denotes pairs of proteins that are functionally associated. Whether or not a protein is annotated with the function of interest, defines the state of the relevant network node. The Bayesian Markov Random Fields (BMRF) method developed in chapter 3 uses the structure of the network to propagate the functional annotations from proteins that are characterized to those that are not. This is done by sampling from the joint probability distribution of protein annotations augmented with the model parameters. Sampling of the unknown protein annotations is done by Gibbs sampling, while the model parameters are sampled by Differential Evolution Markov Chain. The sampling scheme has shown to be suitable for sparsely annotated networks. The BMRF method was further applied in *Saccharomyces cerevisiae* delivering new GO term annotations for 1170 proteins.

In chapter 4, an extended BMRF model is presented for the integration of multiple data sources and is applied to GO term prediction in *Arabidopsis thaliana*. Networks constructed from protein-protein interactions and co-expressions are combined with protein sequences. Two models for network integration were evaluated, one that the BMRF model parameters are allowed to be sampled independently for each network and another that the parameters are fixed to be equal between the networks. For the Arabidopsis data that were used in the study the second model was found to be more suitable. For the protein sequences, a binary domain matrix was constructed indicating the InterPro signatures that each protein contains. A regression model was fitted based on Elastic Nets using as predictors the domain matrix and response the annotation vector. The resulted integrated prediction model was based on the linear combination of the BMRF and the Elastic Nets probabilities.

In chapters 3 & 4 BMRF based predictions are made in a GO term basis without considering whether or not those predictions are consistent with the True Path Rule (TPR) or not. TPR entails that if a protein is assigned to a GO term then it should be assigned to the GO terms that are ancestors of that term, according to the GO Directed Acyclic Graph. Chapter 5 deals with the development of an optimization algorithm that takes as input the BMRF posterior probabilities that were computed per GO term and returns the most probable GO term assignment that is consistent with the TPR. The algorithm called Functional Annotation with Labelling Consistency (FALCON) is based on discrete evolutionary optimization. In particular, a population of TPR consistent assignments is maintained and evolved using graph operations.
In this chapter it was shown that the FALCON algorithm consistently finds the optimal set of consistent annotations, improving the quality of annotations as compared to when they are done on a per GO term basis. In chapter 6 the findings of the previous chapters are discussed and directions for future research are proposed.
Samenvatting

Nauwkeurige en uitvoerige annotatie van genen en hun producten is essentiële voor het begrijpen van biologische processen. De op lab experimenten gebaseerde annotatie procedures zijn niet praktisch omdat ze omslachtig, duur en tijdrovend zijn. In dit proefschrift worden de ontwikkeling en toepassing gepresenteerd van een in silico eiwitfunctie voorspellings methode die heterogene databronnen geïntegreerd behandelt.

In hoofdstuk 1 worden de twee belangrijkste aanpakken voor eiwitfunctie predictie geïntroduceerd. De eerste is gebaseerd op het identificeren van eigenschappen die eiwitten met een bepaalde functie gemeen hebben. Dergelijke eigenschappen kunnen bijvoorbeeld sequentie elementen zijn of het expressie niveau van de bijbehorende genen onder bepaalde omstandigheden. Het tweede aanpak is gebaseerd op het overbrengen van functie gerelateerde informatie van het ene eiwit op een ander. De op BLAST of “schuld door associatie” gebaseerde methoden behoren tot deze categorie. In hoofdstuk 1 wordt ook de Gene Ontology (GO), een veel gebruikte woordenschat voor eiwitfunctie annotaties, geïntroduceerd. GO termen vallen in de volgende drie categorieën: moleculaire-functies, biologische processen en cellulair componenten. Deze woordenschat is uitgebreid genoeg om de enzymatische functie van een eiwit, het proces waarbij het betrokken is en het cellulaire compartiment waar het voorkomt nauwkeurig te kunnen beschrijven.

In hoofdstuk 2 wordt het gebruik van GO met behulp van data van zeven goed bestudeerde soorten bestudeerd. Het verkrijgen van inzicht over hoe deze woordenschat in de praktijk toegepast wordt, is van belang voor de ontwikkeling van degelijke in silico functie predictie methoden. Hoofdstuk 2 laat zien dat het heel gewoon is voor eiwitten om meerdere rollen te vervullen, en wel op verschillende plaatsen in de cel. Dit impliceert dat de predictie methoden in staat moeten zijn om aan elk eiwit meerdere onafhankelijke functies toe te wijzen. Verder werd aangetoond dat de annotatie strategieën aanzienlijk verschillen per soort. Dit werd benadrukt door de feit dat er veel minder eiwitten met meerdere annotaties in Arabidopsis thaliana zijn dan in de andere soorten. Er was geen enkele biologische verklaring voor dit verschijnsel. Het
opstellen van formele regels voor het toewijzen van GO termen zal de functie annotatie van verschillende soorten beter vergelijkbaar maken.

**Hoofdstuk 3** behandelt de ontwikkeling van een probabilistische methode voor eiwitfunctie voorspelling. Deze methode is gebaseerd op Markov Random Fields die het modelleren van onderling afhankelijke variabelen mogelijk maakt. Wanneer deze variabelen als knooppunten worden weergegeven en hun onderlinge afhankelijkheden als verbindingenlijnen, ontstaat er een ongerichte graaf (of netwerk). Voor eiwit associatie netwerken wordt er gebruik gemaakt bijvoorbeeld de gelijkenis tussen de eiwitten, hun interacties of de co-regulatie van de expressie van bijbehorende genen. In dergelijke netwerken corresponderen de knooppunten met eiwitten en de verbindingen lijnen met eiwit-paren die gerelateerd zijn qua functie. Het wel of niet geannoteerd zijn met de bestudeerde functie bepaalt de toestand van een knooppunt in het netwerk. De Bayesian Markov Random Fields (BMRF) methode die in **hoofdstuk 3** wordt ontwikkeld maakt gebruik van de structuur van het netwerk om de functie van geannoteerde eiwitten over te brengen naar eiwitten zonder bekende annotatie. Dit wordt gedaan door steekproeven te trekken uit de gezamenlijke kansverdeling van de onbekende eiwit annotaties en de model parameters. Steekproeven van de onbekende eiwit annotaties werd getrokken met de procedure die met Gibbs sampling wordt aangeduid. De steekproeven van de model parameters werden verkregen met de methodie die met Differential Evolution Markov Chain wordt aangeduid. In **hoofdstuk 3** wordt aangetoond dat dit steekproefschema werkt voor schaars geannoteerde netwerken. De BMRF methode werd toegepast in Saccharomyces cerevisiae en heeft nieuwe GO term annotaties voor 1170 eiwitten opgeleverd.

In **hoofdstuk 4** wordt het BMRF model uitgebreid voor de integratie van meerdere gegevensbronnen en toegepast om GO termen te voorspellen voor Arabidopsis thaliana. Eiwit-eiwit interactie- en co-expressie netwerken werden gecombineerd met eiwit-sequenties. Er werden twee wijzen voor de integratie van de netwerken geëvalueerd. Op de eerste wijze worden de BMRF modellparameters voor elk netwerk onafhankelijk bemonsterd en op de tweede wijze hebben de parameters van de netwerken gelijke waarden. De tweede wijze bleek geschikter te zijn voor de gegevens van Arabidopsis die in de studie werden gebruikt. Voor de eiwit-sequenties werd er een binaire domein-matrix geconstrueerd die de Interpro-domein annotaties van elk eiwit voorstelt. Een logistisch regressie model werd aangepast op basis van de methode die met Elastische Net wordt aangeduid, waarin de domein-matrix als voorspeller en de annotatie vector als afhankelijke variabele werden gebruikt. Het resulterende geïntegreerde voorspellingsmodel was gebaseerd op de lineaire combinatie van de lineaire predictor van het Elastic Net en die van BMRF met alleen netwerk informatie.

In **hoofdstukken 3 en 4** deed BMRF voorspellingen over welke eiwitten
tot verschillende GO termen behoren zonder dat er rekening werd gehouden of deze voorspellingen overeenstemmen met wat wordt aangeduid als de True Path Rule (TPR). De TPR is gebaseerd op de gerichte acyclische graaf van GO termen en houdt in dat, als een bepaalde GO term aan een eiwit is toegewezen, ook alle voorouders van die term in de GO graaf aan het eiwit moeten worden toegewezen. **Hoofdstuk 5** behandelt de ontwikkeling van een optimalisatie algoritme dat de a posteriori waarschijnlijkheden die BMRF per GO term berekend heeft, als invoer neemt en vervolgens de meest waarschijnlijke GO annotatie teruggeeft die voldoet aan de TPR. Het algoritme heet Functional Annotation with Labeling Consistency (FALCON) en is gebaseerd op een discrete evolutionaire optimalisatie methode. Er wordt namelijk een populatie van TPR consistente toewijzingen onderhouden die door het toepassen van graafoperaties evolueert naar de beste toewijzing. Met kleine voorbeelden wordt getoond dat het FALCON-algoritme de optimale set van consistente annotaties vindt. Toepassing van FALCON verhoogt van de kwaliteit van de annotaties. **In hoofdstuk 6** worden de bevindingen van de eerdere hoofdstukken besproken en worden mogelijke richtingen voor vervolg onderzoek voorgesteld.
References


References


References


Acknowledgments

On late December 2005, I moved from Komotini (my hometown in North Greece) to Wageningen to start a Bioinformatics PhD in Biometris. I could never imagine what would happen in the coming years. Many wonderful (and a couple of difficult) moments sum up to the most exciting period of my life.

Several people were directly or indirectly involved in making this thesis possible. First of all, I want to express my deepest gratitude to my supervisor Cajo ter Braak for the constant guidance and unlimited support on all the aspects on the thesis. It was a great privilege to work under the supervision of Cajo all those years and I was greatly benefited from this interaction. Further, Cajo showed me invaluable understanding and support on issues related to my personal life and I am indebted to him for all this support. I am also indebted to Roeland van Ham for the excellent supervision all those years. Roeland helped me very much in keeping a biological focus on my project. I was always enjoying the discussions with Roeland finding them very motivational. Besides the PhD project, I want to thank Roeland for believing in me in undertaking a project on Next Generation Sequencing which showed to be of great interest.

I want to thank Aalt-Jan van Dijk and Marco Bink for providing feedback and support and spending a lot of their times in our meetings, commenting and correcting manuscripts. I am very lucky that I had the chance to work with those people the previous years.

I am grateful to my friends and colleagues in the Biometris and Bioinformatics groups for their support and tolerance all this time.

Edouard, Arnold, Patrick and Jacques it was great sharing the workspace with you. Thank you for your tolerance. Edouard, thanks for all your help in finalizing the last bits of the thesis.

Wies and Vaia, thank you for being there for me in the good and in the dif-
Acknowledgments

Difficult times. Without you, I would never manage to finish the PhD. Patricia, thank you for all your support. Santosh and Martin, thanks for the nice and relaxing discussions during our billiard sessions. Ioanni “Post” thank you for your friendship and for being a good example on how to stay strict on ethical matters, whatever it takes. Stelio and Vasili, my friends from early childhood, thank you for your support during all those years.

I want to thank my former supervisors Filippos Aravanopoulos and Giorgios Rodakis for believing in me and for their support during my BSc and MSc studies.

Special thanks to the system administrators Bas te Lintel Hekkert and Jan van Haarst for their support in using the computer systems.

My family made several sacrifices in order to support me to pursue my dreams. Mother, Father and my brother Dimitri, I am deeply grateful to you for all your love. Thank you for your support during all this period. Thank you for believing in me, tolerating me, for your invaluable help in resolving difficult and complicated situations. Dimitri, thank you for taking me from airports at 2 o’clock at the morning or getting me there early mornings while being exhausted. Thanks for sharing those great future plans with me. Thanks for keeping me motivated. Thanks for being a great brother. Thanks to uncle Tassos, aunt Mairi and my cousins Yanni and Dimitri for your constant support. Uncle “De l’art” Tassos thank you for being a source of inspiration for me. Also, special thanks to Nick and Tom for feeding my soul during the difficult times.

Finally, I want to express my gratitude to Vana for teaching me some important lessons about life, but mainly for showing much love and care to our most important person in our lives, Yiankos.
List of publications

Kourmpetis YAI and ter Braak CJF. Evolutionary discrete optimization for Gene Ontology consistent protein function prediction: the FALCON algorithm. Submitted for publication.

Bousios A, Kourmpetis Y, Pavlidis P, Minga E, Tsaftaris A, Darzentas N. The turbulent life of Sirevirus retrotransposons in the maize B73 genome: more than ten thousand intact elements tell the story. Submitted for publication.


Kourmpetis YAI, A van der Burgt, MCAM Bink, CJF ter Braak, and RCHJ. van Ham. 2007. The use of multiple hierarchically independent gene ontology
Curriculum vitae

Yiannis Kourmpetis was born on 30th of October 1978 in Komotini, Greece. In 2003 Yiannis received a BSc in Agricultural Sciences from the Aristotle University of Thessaloniki. Later in the same year, Yiannis moved to Athens to study Bioinformatics at the National University of Athens and on 2005 he received a MSc degree in Bioinformatics. During his Master research project working on Computational Phylogenetics he got introduced to the basic ideas of Bayesian Phylogenetics. On January 2006, he started his PhD in Bioinformatics in Wageningen University on the topic of Bayesian gene function prediction. Currently, Yiannis works at Wageningen University in a research project on Next Generation Sequencing data analysis. On December 2007, his son Yiankos was born in Ede.
**Education Statement of the Graduate School**

**Experimental Plant Sciences**

Issued to: Yiannis Kournpitis  
Date: 04 October 2011  
Group: Biometris, Wageningen University & Research Centre

### 1) Start-up phase
- First presentation of your project  
  Oral presentation on Bayesian gene function prediction  
  Jul 13, 2006  
- Writing or rewriting a project proposal  
  Bayesian gene function prediction  
  May 2006  
- Writing a review or book chapter  
- MSc coursework  
- Laboratory use of isotopes

**Subtotal Start-up Phase** 7.5 credits

### 2) Scientific Exposure
- EPS PhD student days  
  EPS Student Day 2006, Wageningen University  
  Sep 19, 2006  
  EPS Student Day 2007, Wageningen University  
  Sep 13, 2007  
- EPS theme symposia  
  Theme 4 symposium 'Genome Plasticity', Wageningen University  
  Dec 12, 2008  
- NWO Lunteren days and other National Platforms  
  Biorange Meeting 2006, Lunteren  
  Mar 05-06, 2008  
  NBIC conference 2009, Lunteren  
  Mar 17-18, 2009

**Seminars (series), workshops and symposia**  
- Capita Selecta 'Structure and dynamics of biomolecules in vitro and in vivo'  
  Oct-Dec 2006  
- Systems Biology for plant design  
  Jul 08-11, 2010  
- Statistical Challenges on the 1000euro genome sequences in plants, Toulouse  
  Apr 14-15, 2011  
- Next Generation Sequencing Symposium, Cologne  
  Nov 13-15, 2010  
- Integrative Bioinformatics Symposium, Wageningen  
  Mar 21-23, 2011

**Presentations**  
- Diverse data combination for gene function prediction by evolutionary ensembles (ISMB 2007, Vienna)  
  Jul 19-20, 2007  
- Integrated prediction of protein functions in Arabidopsis thaliana by Bayesian Markov Random Fields (poster)  
  Mar 21-23, 2011  
- Next Generation Sequencing for genome variation in Arabidopsis natural populations (Statfis 2011, Toulouse)  
  Apr 14-15, 2011  
- CAF/A11: Network based protein function prediction using Bayesian Markov Random Fields (ISMB 2011, oral)  
  Jul 15-19, 2011  
- IAB interview  
  Dec 09, 2008

**Excursions**

**Subtotal Scientific Exposure** 15.2 credits

### 3) In-Depth Studies
- EPS courses or other PhD courses  
  Introduction to Stochastic Processes (LNB) Utrecht  
  Sep 04.05-11.12, 2006  
  Simulation (LNB) Utrecht  
  Sep 28-Dec 04, 2006  
  Continuous Optimization (LNB) Utrecht  
  Sep 28-Dec 04, 2006  
  Multivariate Analysis (PERC)  
  Apr 18-26, 2007  
  Applied Bioinformatics in Plant Sciences (EU-Cost Action), Athens  
  Dec 13-17, 2010

**Journal club**  
- Joint meetings Applied Bioinformatics and Bioinformatics groups  
  2006-2011

**Individual research training**  
- EU-COST Short Term Scientific Mission (Uppsala university)  
  Mar 28-Apr 13, 2011

**Subtotal In-Depth Studies** 38.4 credits

### 4) Personal development
- Skill training courses  
  PhD Competence Assessment  
  Sep 12, 2006  
  PhD Career Assessment  
  Nov 19, 2008

**Organisation of PhD students day, course or conference**  
- Statistics for –omics data analysis  
  Dec 11-13, 2006  
- EPS PhD course Molecular Phylogenetics  
  Oct 16-19, 2010  
- Mic course in Advanced Bioinformatics  
  Feb 14, 2011

**Membership of Board, Committee or PhD council**

**Subtotal Personal Development** 3.6 credits

**TOTAL NUMBER OF CREDIT POINTS** 43

*The Graduate School declares that the PhD candidate has completed the educational requirements set by the Educational Committee of EPS which comprises of a minimum total of 30 ECTS credits.

* A credit represents a normative study load of 28 hours of study.