Analysis of Protein-Protein Interaction Networks

by Means of Annotated Graph Mining Algorithms

Hossein Rahmani

Analysis of Protein-Protein Interaction Networks

by Means of Annotated Graph Mining Algorithms

Proefschrift

ter verkrijging van

de graad van Doctor aan de Universiteit Leiden,

op gezag van Rector Magnificus prof.mr. P.F. van der Heijden,

volgens besluit van het College voor Promoties

te verdedigen op dinsdag 30 oktober 2012

klokke 15.00 uur

door

Hossein Rahmani

geboren te Tehran, Iran
in 1983

PhD committee

Promotor:	Prof. dr. J.N. Kok	Leiden University
Co-promotor:	Dr. H. Blockeel	Leiden University
Other members:		
	Prof. dr. T. Baeck	Leiden University
	Dr. Ir. F.J. Verbeek	Leiden University
	Prof. dr. K. Marchal	Universiteit Gent

The work reported in this thesis has been carried out at the Leiden Institute of Advanced Computer Science at Leiden University, under the auspices of the research school IPA (Institute for Programming research and Algorithmics). This research is supported by the Dutch Science Foundation (NWO) through a VIDI grant.

The cover illustration depicts a part of Human Disease Network discussed in this thesis.
Cover design by: Hossein Rahmani.
Published by: Uitgeverij BOXPress, 's-Hertogenbosch
ISBN: 978-90-8891-488-1

Contents

Chapter 1

Introduction

1.1 General Introduction

This thesis is a collection of 5 papers discussing solutions to several open problems in Protein-Protein Interaction (PPI) networks with the aid of Knowledge Discovery. PPI networks are usually represented as undirected graphs, with nodes corresponding to proteins and edges representing interactions among protein pairs. A Large amount of available PPI data and noise within it has made the knowledge discovery process a necessary central part for the network analysis. We define Knowledge Discovery as a process of extracting informative knowledge from the huge amount of data. Much success has been achieved when the input data is represented as a set of independent instances and their attributes. But, in the context of PPI networks, there is interesting knowledge to be mined from the relationships between instances (proteins). The resulting research area is called "Graph Mining". Here, the input data is modeled as a graph and the output could be any type of knowledge. In this thesis, we propose several graph mining algorithms to examine structural characteristics of PPI networks and link them to the information useful for biologists, such as function or disease.

This chapter consists of two main sections. In the first section, we discuss the knowledge discovery process and its high-level subprocesses. In the second section, we discuss the area of PPI networks, its open problems and our proposed methods for solving them.

1.2 Knowledge Discovery Process

As a quote from John Naisbitt: "We are drowning in information but starved for knowledge" indicates, the amount of data available in different aspects of life increases every second and the task to mine data and extract useful knowledge becomes more and more challenging. The main goal of the knowledge discovery process is to extract informative knowledge from a large amount of data in a human understandable structure. Considering the whole knowledge discovery process as a system which takes a certain type of data as input and produces informative knowledge as output, Figure 1.1 shows three main subprocesses of the whole system. The first subprocess is called "Data Pre-Processing" which takes raw input data and outputs the cleaned version of the data. The second subprocess is called "Machine Learning" and its main task is to extract potential informative patterns from the cleaned data. The last subprocess is called "Data Post-Processing"; it validates and evaluates the extracted patterns. We will discuss each of these subprocesses in more details in the following sections.

1.2.1 Input Data

Before discussing the details of the different knowledge discovery subprocesses, we briefly introduce some basic concepts about the *Input Data* we deal with in this thesis. We model the PPI network as a graph $G(V, E)$, where V is a set of nodes (proteins in our context) and E is a set of edges (interactions in our context) connecting pairs

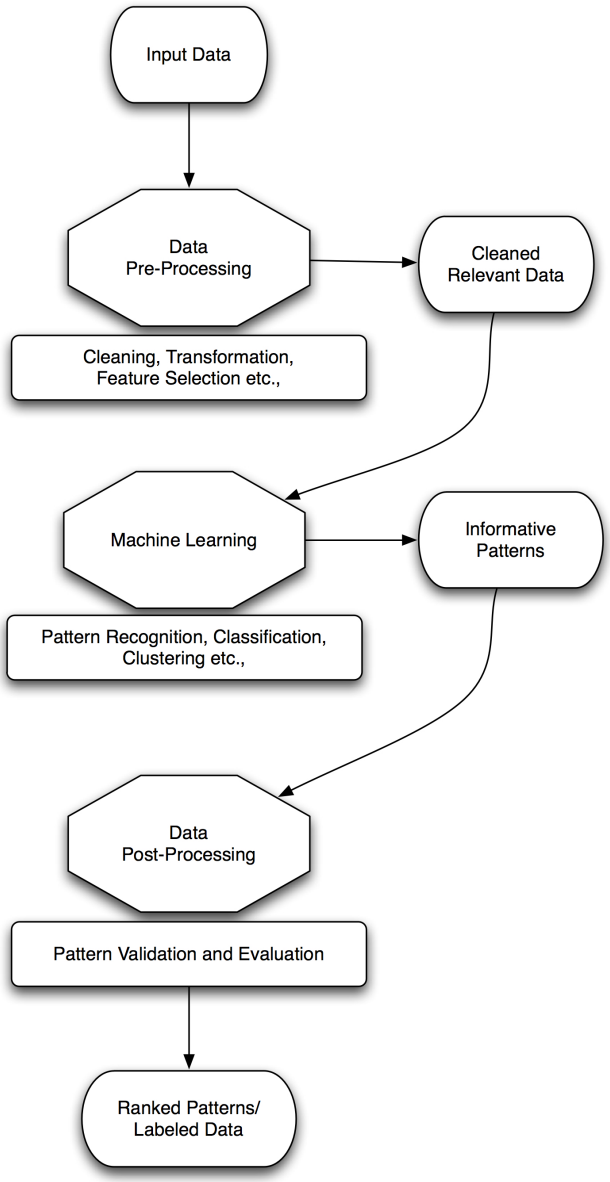
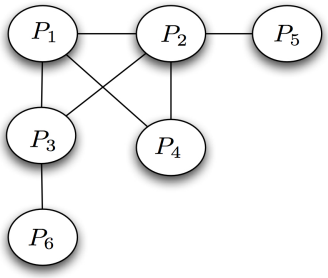


Figure 1.1: Three main subprocesses of the whole knowledge discovery process. The first subprocess is called "Data Pre-Processing" which takes raw input data and output the cleaned version of the data. The second subprocess is called "Machine Learning" and its main task is to extract potential informative patterns from the cleaned data. The last subprocess is called "Data Post-Processing" and validates and evaluates the extracted patterns.



$$Degree(P_1) = 3$$
$$Path(P_5 \rightarrow P_6) = P_5, P_2, P_1, P_3, P_6$$
$$Path(P_5 \rightarrow P_6) = P_5, P_2, P_3, P_6$$
$$Length(P_5, P_2, P_1, P_3, P_6) = 4$$
$$Length(P_5, P_2, P_3, P_6) = 3$$
$$Distance(P_5, P_6) = 3$$

Figure 1.2: A simple graph $G(6, 7)$ and some of its graph-based features. We use these features to describe graph nodes.

of nodes. Assuming a and b are two arbitrary nodes in the graph $G(V, E)$, we define the following graph-based features:

- $Degree(a)$: Number of edges a is connected with.
- $Path(a \rightarrow b)$: Sequence of nodes starting with node a and ending with node b, such that there is an edge between each two subsequent nodes of sequence.
- $Length(Path(a \rightarrow b))$: Number of edges in the $Path(a \rightarrow b)$.
- $Distance(a, b)$: Length of the shortest path between a and b.

In the following parts, we use these features to describe nodes. Figure 1.2 shows a simple graph $G(6, 7)$ in addition to some of its graph-based features.

1.2.2 Data Pre-Processing

In the knowledge discovery process, input data may contain noisy and irrelevant data that should be cleaned before further analysis of the data (garbage in, garbage out). The main goal of Data Pre-Processing is to prepare a final training set for the machine learning algorithms and that may include cleaning, transformation, feature selection etc. Number of the features equals to the size of the graph ($|V|$). For example, in feature $Path(a \rightarrow b)$, a is a node to describe and then, for each different node b we have different feature $Path(a \rightarrow b)$. So, we need feature selection algorithms to select useful and informative node b and then, describe node a based on them.

Due to the large and noisy nature of the PPI network, a natural way to reduce the dimensionality is using a feature selection method to filter out the least interesting features. Next, we will discuss two feature selection methods Chi-square and Anova (Analysis of variance) for this purpose.

	$F=0$	$F=1$	$Total$
$C=0$	a	b	$a+b$
$C=1$	c	d	$c+d$
	$a+c$	$b+d$	$a+b+c+d$

Table 1.1: The contingency table of a binary feature F w.r.t. a binary class variable C. a, b, c, and d count the number of times F and C have the corresponding value. The χ^2 value of F w.r.t. C is derived from this.

Chi-Square Feature Selection

An often used measure for determining the relevance of a binary feature F for a class variable C is the χ^2 score which is defined by Liu et al. [69] as follows:

$$\chi^2 = \frac{(ad-bc)^2 * (a+b+c+d)}{(a+b)(c+d)(b+d)(a+c)}, \tag{1.1}$$

where a, b, c and d are defined by the contingency table in Table 1.1.

Anova-based Feature Selection

The Anova-inspired selection measure (briefly, Anova) is defined as follows. Let P^+ be the set of input cases labeled as class C, and P^- the set of input cases not labeled as such. For each input case q, we introduce a feature d_q; In the context of PPI networks, $d_q(p)$ denotes the shortest-path distance between p and q (viewed here as a feature of p). We consider for each q the mean and variance of $d_q(p)$, taken over all C-related and non-C-related p:

$$m_q^+ = \frac{\sum_{p\in P^+} d_q(p)}{|P^+|} \tag{1.2}$$

$$m_q^- = \frac{\sum_{p\in P^-} d_q(p)}{|P^-|} \tag{1.3}$$

$$var_q^+ = \frac{\sum_{p\in P^+} (d_q(p)-m_q^+)^2}{|P^+|-1} \tag{1.4}$$

$$var_q^- = \frac{\sum_{p\in P^-} (d_q(p)-m_q^-)^2}{|P^-|-1} \tag{1.5}$$

Seeing P^+ and P^- as two groups of input cases, the following formula compares the variance between groups to the variance within groups (as it is used for relative ranking only, constant factors are dropped):

$$A_q = \frac{(m_q^+ - m_q^-)^2}{var_q^+ + var_q^-} \tag{1.6}$$

A high A_q means that d_q varies little within groups and/or much between groups, which indicates that d_q has high predictive power for the group. Features d_q can be ranked according to A_q, and the top-k features selected as actual features to be included in the description of all proteins.

1.2.3 Machine Learning

Machine Learning is a subfield of Artificial Intelligence in which the main goal is to learn knowledge through experience. Tom Mitchel in his book [79] defines the "ability to learn" as follows: A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Based on the problem definition and the type of training data (whether it is labeled or unlabeled), we focus on two high level main machine learning tasks: *Supervised learning* and *Unsupervised learning*. An output of supervised learner is a classifier that has the ability to predict the correct label for any valid input data while an unsupervised learner tries to infer hidden structure among unlabeled data. In this thesis, we deal with different annotating problems of PPI networks and accordingly, our final goal is to propose a supervised learner. Next, we discuss briefly about some classifiers that we use in this thesis.

Naive Bayes Classifier

In the category of probabilistic classifiers, the naive Bayes classifier is the simple classifier which applies 'Bayes Theorem' by assuming independency among features. The probability model for this classifier is a conditional model $P(C|F_1, \ldots, F_n)$ over a dependent class variable C, conditional on several feature variables F_1 through F_n. If the number of features n becomes large or feature F_i can take large number of values, then basing this model on probability tables may not be feasible. We reformulate the model using Bayes' Theorem shown in Formula 1.7.

$$p(C|F_1, \ldots, F_n) = \frac{p(C)p(F_1, \ldots, F_n|C)}{p(F_1, \ldots, F_n)} \tag{1.7}$$

In Formula 1.7, only the numerator of fraction is dependent to class variable C and denominator is practically constant. Now, by applying the conditional independency assumption among different features shown in Formula 1.8 we can express the model as Formula 1.9.

$$p(F_i|C, F_j) = p(F_i|C) \tag{1.8}$$

$$\begin{aligned} p(C|F_1, \ldots, F_n) &= \frac{1}{H}p(C)p(F_1|C)p(F_2|C, F_1)\ldots p(F_n|C, F_1, F_2, F_3, ...F_{n-1}) \\ &= \frac{1}{H}p(C)\prod_{i=1}^{n} p(F_i|C) \end{aligned} \tag{1.9}$$

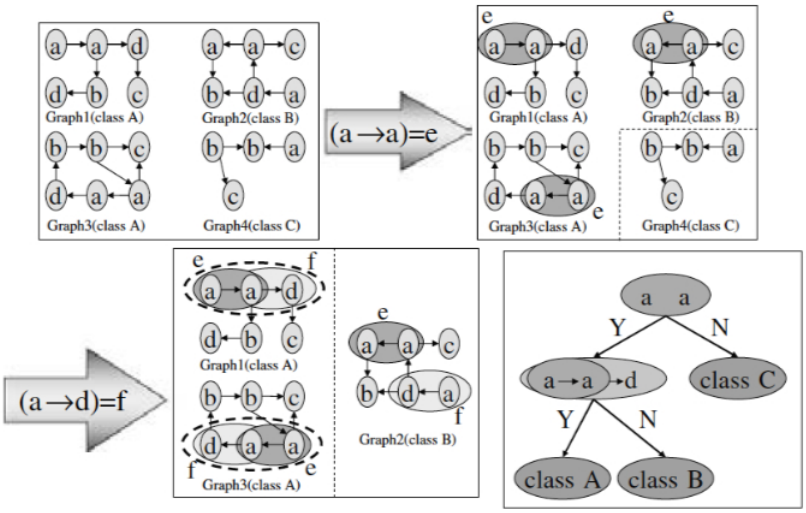


Figure 1.3: The process of building the simple decision tree classifier based on 4 input graph data annotated with three class labels: class A, class B and class C. Considering each graph $G(V, E)$ with node set V and edge set E, the features of the decision tree classifier would be whether an edge $v_i \rightarrow v_j \in E$ or not. The Figure is taken from [20].

where H is a scaling factor dependent only on $F_1, \ldots, F_n$.

Decision Tree Classifier

This classifier uses a tree-like structure to predict the label of the data. In the tree-like structure, leaves are class labels and branches represent conjunctions of features that lead to those class labels. A decision tree can be constructed by recursively splitting the training set into subsets based on the feature value. At each step, the feature that most reduces the uncertainty about the class in each partition, is selected and is used as a split. The recursion is completed when all elements of the subset at a node have the same label value, or when splitting no longer adds value to the predictions. Figure 1.3 shows the process of building the simple decision tree classifier based on 4 input graph data categorized with three class labels: class A, class B and class C. Considering each graph $G(V, E)$ shown in Figure 1.3 with node set V and edge set E, the features of the decision tree classifier would be whether an edge $v_i \rightarrow v_j \in E$ or not. As Figure 1.3 shows, the most discriminative edge is $a \rightarrow a$ which classifies class C from classes A and B.

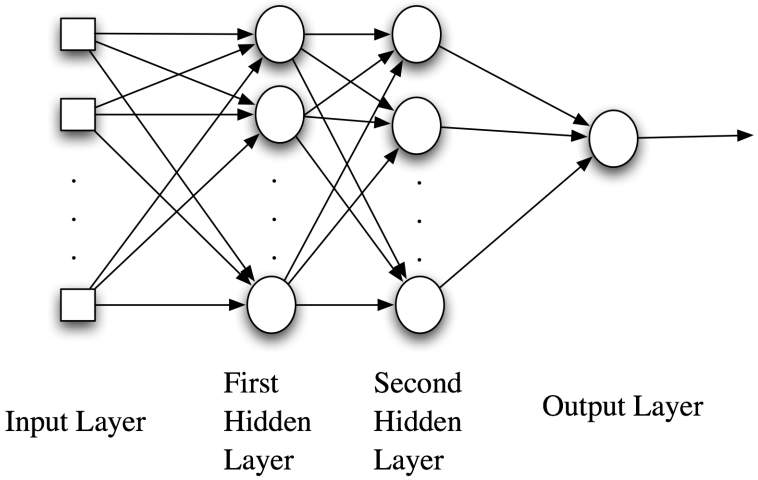


Figure 1.4: A simple feedforward Artificial Neural Network with two hidden layers.

Artificial Neural Networks

A simple and still efficient way of solving a complex problem is through using the *divide and conquer* strategy which solves a problem by breaking the complex problems into smaller (and still the same type as the original problem) subproblems. Then, recursively solve the simple subproblems and integrate the solutions. Networks can be used for this strategy where each node acts as a computational unit (i.e., receive input data, process it and generate output data) and the network connections show the information flow and the way different computational units integrate their outputs.

One type of network models the nodes based on the structural and functional aspects of biological neurons. They are called Artificial Neural Networks (ANNs). ANNs can be used for both supervised and unsupervised learning. The performance of ANN mainly depends on the following parameters:

- Network Connectivity: How different nodes interconnect with each other.
- Learning Process: How to update the weights of the interconnections.
- Activation Function: How to convert a neuron's weighted input to its output activation.

Figure 1.4 shows a simple neural network with two hidden layers. Networks such as the one shown in Figure 1.4 are commonly called *feedforward* network, because their graph is a directed acyclic graph. Networks with cycles are commonly called *recurrent*.

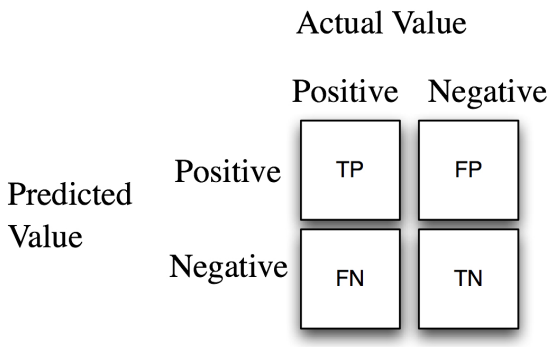


Figure 1.5: Definition of True Positive (TP), False Positive (FP), False Negative (FN) and True Negative (TN) in a binary classification.

1.2.4 Data Post-Processing

The extracted informative patterns could be further processed. We could evaluate the patterns, simplify, visualize, interpret and incorporate them into an existing system. In this section, we discuss different evaluation measures/techniques that we use for evaluating our methods.

Precision, Recall and Fmeasure

In this thesis, we evaluate our predictions according to Precision, Recall and Fmeasure as follows:

$$Precision = \frac{tp}{tp + fp} \tag{1.10}$$

$$Recall = \frac{tp}{tp + fn} \tag{1.11}$$

$$Fmeasure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{1.12}$$

where tp, fp and fn denote the number of true positives, false positives, and false negatives, respectively and are defined in Figure 1.5.

Different Cross Validation Techniques

We need a technique to show how well the learned model from the training data will perform on future independent data. In *k-fold cross validation*, we partition the input data into k folds and then, use one fold for validating the model and the remaining $k - 1$ folds for training the algorithm. We repeat this process k times, with each of

the k folds used exactly once as the validation data. Finally, we average the k results from the folds to produce a single estimation. In this thesis, we mostly assume $k = 10$ for evaluating our methods.

One particular case of cross validation techniques is *leave-one-out cross validation (LOOCV)*, where we consider a single instance from the input data as a validation data, and the remaining instances as the training data. We repeat this process for each instance in the input data and we average the N (= number of the examined instances) results to produce a single estimation.

One special case of LOOCV is to find a rank of some previously selected instances relative to 99 randomly selected instances as follows:

1. We select 99 instances randomly from the input data ($randSet$).
2. For each previously selected instance psi
   (a) We build the $trainSet$ by excluding the $\{psi \cup randSet\}$.
   (b) We train the prediction method M with $trainSet$ and then, we apply M to rank psi relative to the 99 randomly selected isntances ($rank(psi)$). M should return small rank values for more relevant input instances.
3. We repeat steps 1 to 2b, 10 times and we calculate the average rank of each psi over different iterations ($avg(rank(psi))$).

1.2.5 Knowledge Discovery Tool: WEKA

For each knowledge discovery subprocess *Data Pre-Processing*, *Machine Learning* and *Data Post-Processing* shown in Figure 1.1, there are hundreds of possible methods and algorithms available in the literature. Instead of implementing those techniques from scratch, we can benefit from the use of free, Java-based open source, off-the-shelf tool WEKA [127] (Waikato Environment for Knowledge Analysis). WEKA contains a collection of state-of-the-art algorithms and tools for each high-level subprocess of knowledge discovery shown in Figure 1.1, in addition to an easy graphical user interface for those functionalities. Table 1.2 shows a brief list of WEKA's capabilities for each subprocess.

1.3 PPI Network and Its Open Problems

In recent years, much effort has been invested in the construction of protein-protein interaction (PPI) networks [118]. Much can be learned from the analysis of such networks with respect to the metabolic and signalling processes present in an organism, and the knowledge gained can also be prospectively employed e.g. to the task of protein function prediction [78, 98, 18, 111, 121, 119, 57, 13], identification of functional modules [71], interaction prediction [48, 129], identification of disease candidate genes [27, 109, 26, 58, 106, 37, 87, 130, 132] and drug targets [104, 81], according to an analysis of the resulting network [72].

Pre-Processing	Change data formats (e.g., From nominal to binary etc), Feature selection (e.g., Chi-square, Principle Component Analysis (PCA), Information gain etc), Discretize data etc.
Machine Learning	Naive Bayes, Support Vector Machine (SVM), K Nearest Neighbors (KNN), Tree-based classifiers (ID3, J48 and Random Forest), EM Clustering, Apriori, etc.
Post-Processing	k-fold cross validation, Cost sensitive evaluation, Visualization and etc.

Table 1.2: Very brief list of WEKA's capabilities in each knowledge discovery sub-process shown in Figure 1.1.

In the following sections, first, we introduce three types of datasets that we use in this thesis. Second, we describe some open problems of PPI networks and finally, we discuss our proposed methods for each open problem.

1.3.1 Data Set

We consider a PPI network as an undirected annotated graph $(P, E, \lambda_F, \lambda_D)$ where P is a set of proteins, $E \subseteq P \times P$ is a set of interactions between these proteins, and λ_F and λ_D are so-called annotation functions; for each p, λ_F and λ_D denote the additional information we have about p. In this work, we assume that $\lambda_F(p)$ simply lists all the biological functions that are associated with p; we call it the function set (or function vector) of p, and denote it $FS(p)$. $\lambda_D(p)$ lists all the diseases/cancers that protein p is involved in; we call it the disease list of p and denote it $dizList(p)$. According to these annotation information, we consider three categories of datasets: PPI datasets, Function datasets and Cancer/Disease datasets.

PPI Datasets

This category of datasets describes the proteins and the way different proteins interact with each other. We use two types of PPI networks: *S. cerevisiae* datasets and Human datasets. Tables 1.3 shows the number of proteins and number of interactions for each PPI network used in this thesis.

The Milenkovic et al. [77] data set is the union of three human PPI datasets: HPRD [91], BIOGRID [116] and the dataset used by Radivojac et al. [97]. When we say "union", we mean that the new network contains all the nodes and edges (proteins and interactions) found in either of these networks. The aim of merging these three datasets was to obtain as complete a human PPI network as possible, i.e., a network that covers with its edges as many proteins in the human proteome as possible. Milenkovic et al. [77] provide details on the construction of the union network.

Dataset Type	Name	Proteins Count	Interactions Count
S. cerevisiae	DIP-Core [25]	2,388	4,400
S. cerevisiae	VonMering [123]	2,708	22,000
S. cerevisiae	Krogan [59]	2,708	14,246
S. cerevisiae	MIPS [76]	7,928	44,514
Human	Milenkovic et al. [77]	10,282	47,303

Table 1.3: PPI networks used in this thesis.

Function Datasets

This category of datasets describes the functional annotation of each protein in the PPI network. We use different function datasets for *S. cerevisiae* and Human datasets.

The protein function annotation for *S. cerevisiae* PPI networks are obtained from the Yeast Genome Repository [39]. In this dataset, functions can be described in different levels of detail. For example, two functions 11.02.01 (rRNA synthesis) and 11.02.03 (mRNA synthesis) are considered the same up to the second function level (i.e., 11.02 = RNA synthesis), but not on deeper levels. Figure 1.6 shows high level categories of this dataset.

Functional Category
01 METABOLISM
02 ENERGY
10 CELL CYCLE AND DNA PROCESSING
11 TRANSCRIPTION
12 PROTEIN SYNTHESIS
14 PROTEIN FATE (folding, modification, destination)
16 PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic)
18 REGULATION OF METABOLISM AND PROTEIN FUNCTION
20 CELLULAR TRANSPORT, TRANSPORT FACILITIES AND TRANSPORT ROUTES
30 CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM
32 CELL RESCUE, DEFENSE AND VIRULENCE
34 INTERACTION WITH THE ENVIRONMENT
38 TRANSPOSABLE ELEMENTS, VIRAL AND PLASMID PROTEINS
40 CELL FATE
41 DEVELOPMENT (Systemic)
42 BIOGENESIS OF CELLULAR COMPONENTS
43 CELL TYPE DIFFERENTIATION

Figure 1.6: MIPS high level function categories.

We annotate the proteins in the Human PPI datasets based on Gene Ontology (GO) [36]. GO unifies the representation of gene and gene product attributes by

introducing three ontology domains: Cellular Components, Molecular Functions and Biological Process. Figure 1.7 shows part of the Gene Ontology domains.

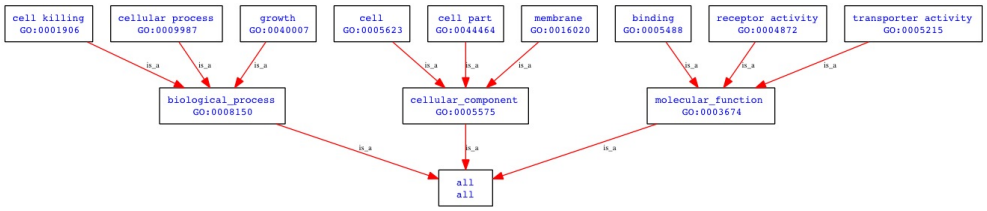


Figure 1.7: Part of the Gene Ontology domains.

Cancer/Disease Datasets

We denote as "known cancer proteins" the set of proteins implicated in cancer that is available from the following databases: Cancer Gene Database [23], Cancer Genome Project-the Cancer Gene Census [95], GeneCards [32] and Kyoto Encyclopedia of Genes and Genomes [83]. Similarly, proteins in Online Mendelian Inheritance in Man [47] are annotated as "disease-related proteins".

1.3.2 Predicting the Functions of Proteins in PPI Networks

Some of the proteins in PPI networks are annotated with biological functions. The task of function prediction in a PPI network is trying to predict the functions of unannotated proteins based on the information available in the PPI network. In Figure 1.8, the color of each protein shows the protein's functional annotations and the main target is to predict the colors of white proteins in this network. As we discussed in section 1.3.1 , we use four PPI datasets: DIP-Core, VonMering, Krogan and MIPS and the functional dataset Yeast Genome Repository [39] for this task. We discuss our proposed methods for this problem in chapters 2 and 3 of this thesis.

In chapter 2, we predict functions from relative position of proteins in the PPI network.

- First, we classify the previous methods on protein's function prediction into *Inductive* and *Transductive* methods. We define inductive approaches as model-based approaches which construct a model (a mathematical function) to map a description of a protein onto its function. On the other hand, we define transductive approaches as non-model based methods which immediately make predictions for the proteins in the PPI network without going through the intermediate stage of constructing a model. We compare characteristics of inductive and transductive methods and we conclude that using inductive methods have more advantages comparing to using transductive methods.

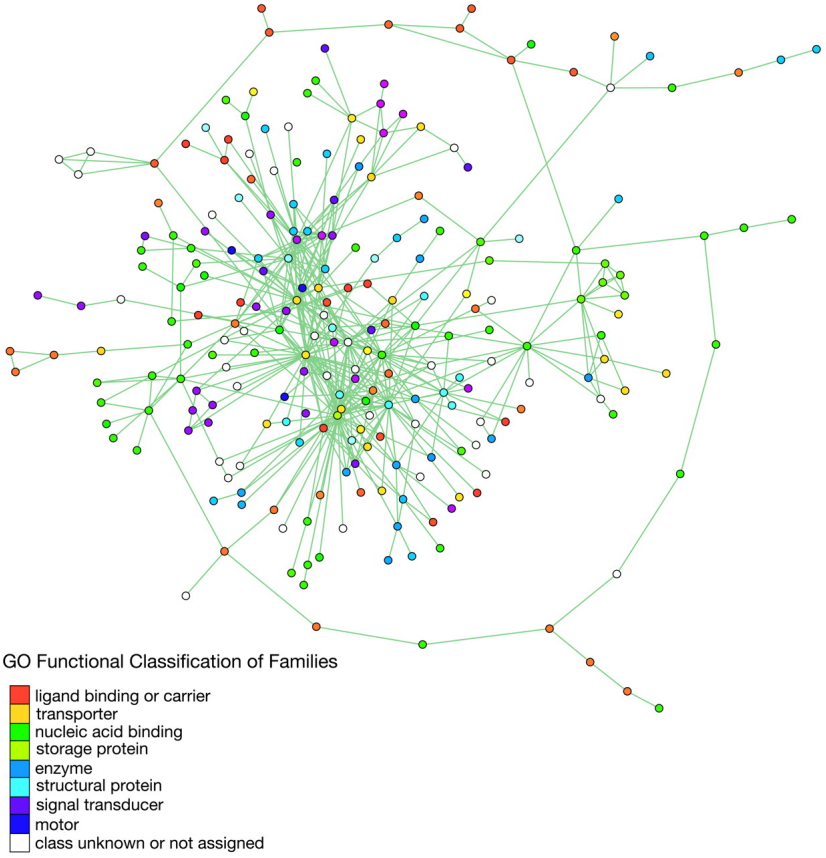


Figure 1.8: Predicting the functions of proteins in PPI network. The color of each protein shows protein's functional annotations and the main target is to predict the colors of white proteins based on the information available in the network.

- Second, we introduce an inductive approach that uses a global protein description for the task of function prediction in PPI networks as follows: Assume that there are n nodes in the network, identified through numbers 1 to n. Each node is then described by an n-dimensional vector. The i'th component in the vector of a node v gives the length of the shortest path in the graph between v and node i. A potential disadvantage of this method is that in large graphs, one gets very high-dimensional descriptions, and not all learners handle learning from high-dimensional spaces well. It is possible, however, to reduce the dimensionality of the vector by only retaining the shortest-path distance to a few "important" nodes. This essentially represents a feature selection problem. A node i is important if the shortest-path distance of some node v to i is likely

to be relevant for v's classification. We use the Anova measure (discussed in section 1.2.2) for selecting the "important" nodes in the PPI network.

- Third, given the input data and a particular function to predict, any standard machine learning tool can be used to build a model that predicts from a node's description, whether the node has a particular function or not. We compare several methods, as available in the WEKA data mining toolbox [127], namely decision trees (J48), Random Forests (RF), an instance based learner (IBk), Naive Bayes, radial basis function networks, Support Vector Machine (libSVM), Classification via Regression (CVR) and Voting Feature Intervals (VFI), with each other and we observe that Random Forests are our best candidate for learning from the given type of data, and we use this method in the remaining experiments.

- Fourth, we compare the performance of this system with that of Majority Rule (MR) [111], a transductive learner. MR simply assigns to a protein the k functions that occur most frequently among its neighbors (with k a parameter). We see that, over the four datasets, RF has higher precision (11% higher in average) but smaller Recall (10% smaller in average). RF and MR perform almost similarly with respect to Fmeasure. RF tends to have higher scores (+6%) with respect to AUC.

- Fifth, we investigate the effect of the Anova-based node selection criterion on predictive performance: Does a reduction of the number of important nodes increase or decrease the predictive performance, and is there a clear optimum with respect to the number of important nodes that should be selected? We notice that for most of the functions, selecting 50-70 important proteins is enough to obtain good classification results. Beyond this area, there is usually no major improvement in performance.

- Sixth, we investigate whether our method also classifies proteins accurately on more detailed MIPS function levels. We examine up to five different function levels and for each level compare our method with Majority Rule. The highest improvement is observed for function level 2, when our method has more than 8% higher Fmeasure value, on average, for the DIP-Core and VonMering datasets. The difference is smallest for very general (level 1) or very specific (level 5) function prediction.

In chapter 3, we predict functions from information about the functions of proteins it interacts with.

- First, we categorize the previous methods into structural based and non-structural based methods. Structural based methods rely on the local or global structure of the PPI network and do not use information about the functions of other nodes to predict the functions of a particular protein. Methods that do use such information form a non-structural category. A prototypical example is the

Majority Rule (MR) approach [111]. A common drawback of the second category of approaches is that they rely solely on the assumption that neighboring proteins tend to have the same functions. It is not unreasonable to assume that proteins with one particular function tend to interact with proteins with specific other functions. We call such functions "collaborative" functions. We assume that a biological process is a complex aggregation of many individual protein functions, in which topologically close proteins have collaborative, but not necessarily the same, functions. We define collaborative functions as pairs of functions that frequently interface with each other in different interacting proteins.

- Second, we propose a Reinforcement Based Collaborative Function Prediction (RBCFP) that increases the collaboration value of two functions if they interface with each other in two sides of one interaction and decreases the collaboration value if just one of the functions occurs on either side of an interaction. After calculating the collaboration value for each pairs of functions in the PPI network, at prediction time, this method ranks candidate functions based on how well they collaborate with the neighborhood of unclassified protein.

- Third, we propose a Self Organizing Map (SOM) based collaborative function prediction that has a one-layered network with as many inputs as there are functions in the PPI network, and equally many output neurons. Each input is connected to each output. After training the SOM, the network takes as input the functions occurring in a protein's neighborhood, and outputs information about the protein's functions.

- Fourth, we compare our collaboration-based methods (RBCFP and SOM) with similarity-based methods using leave-one-out cross validation (discussed in section 1.2.4) in five different function levels. We observe that collaboration based methods predict functions more accurately than similarity based methods. As we consider more detailed function levels, the difference between their performance increases.

1.3.3 Predicting Cancer-related Proteins in PPI networks

Some of the proteins in PPI networks are annotated as being involved in cancer (cancer-related proteins). The task of predicting cancer-related proteins is trying to predict which other proteins in the PPI network are most likely involved in cancer.

As we discussed in section 1.3.1 , we use Human PPI and Gene Ontology datasets for this task. We discuss our proposed methods for this problem in chapters 4 and 5 of this thesis.

In chapter 4, we predict cancer-related proteins from functional and structural information in the PPI network.

- First, we discuss two types of previous methods: guilt-by-proximity and feature-based methods. Methods classified in "guilt-by-proximity" category are based on

the assumption that genes that directly interact, or, more generally, lie close to each other in the network, are more likely to be involved in the same diseases (as argued by, e.g., Gandhi et al. [31]). The methods vary based on how they define proximity. In "feature-based" methods, each individual protein is described by means of a fixed set of features such as protein degree, protein length, protein GO annotations etc. Next, using machine learning methods, a model is learned that links some of these features to cancer-relatedness. Figure 1.9 shows the general procedure of feature-based methods. We compare characteristics of guilt-by-proximity and feature-based categories and we conclude that feature-based approaches have a number of advantages over proximity-based approaches with respect to flexibility and data integration.

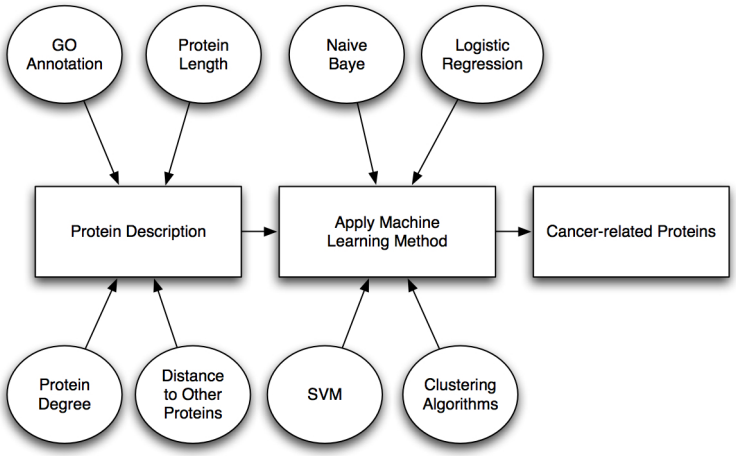


Figure 1.9: The general procedure of feature-based approaches for predicting cancer-related proteins. In the first step, we describe each proteins based on some features (e.g., protein degree, protein length, protein GO annotations and its shortest path distance to some other proteins in the network). In the second step, we apply a machine learning method (e.g., Naive Bayes, logistic regression, support vector machine and classification by clustering) to the proteins descriptions. In the last step, we evaluate the newly predicted cancer-related proteins.

- Second, we assume that GO annotations of proteins are often incomplete, and by collecting GO information from the neighbors of a protein p, we may get more information about p itself. This argument is backed up by the fact that GO annotations of proteins can often be predicted well from the GO annotations of their neighbors; see, e.g., [111, 99]. However, this is not the only effect; there is also a direct relationship between a protein's involvement in cancer and the GO annotations of the proteins it interacts with. We propose a new type of

functional feature that considers the functions of proteins interacting with the target protein (rather than the protein itself).

- Third, we propose a new type of structural features which considers the relative position of the target protein with respect to specific other proteins selected according to the Anova (discussed in section 1.2.2) based measure.

- Fourth, by applying literature mining to the most discriminative functional and structural features, we succeed to find the biological relevance for all the proposed features.

- Fifth, we describe the proteins in PPI network based on network contextual information (Functional and Structural features) and then, we apply the naive Bayes classifier for the prediction task. We observe that a simple and efficient machine learning method (here Naive Bayes) that uses a combination of functional information about the neighbors and shortest-path distance to specific proteins, predicts cancer-related proteins with higher accuracy than any previous PPI-based methods.

- Sixth, we analyze a list of 20 genes predicted to be involved in cancer by our method, but not annotated in this manner in our training dataset, and we find that virtually all of them (at least 18 out of 20) could be linked to cancer in scientific publications. So, not only our classification results improve upon previous methods, but that also our 'false' positive predictions could in many cases be verified to be linked to cancer in more recent literature.

In chapter 5, we predict cancer-related proteins by combining Gene Ontology annotations with information contained in the topology of a PPI network.

- First, we discuss standard machine learning approaches [77, 29, 66, 30] for the task of predicting cancer-related proteins in PPI networks. We observe that these methods typically use a feature-based representation of proteins as input, and their success depends strongly on the relevance of the selected features. Figure 1.9 shows the general procedure of feature-based methods .In earlier work it has been shown that the Gene Ontology (GO) annotations of a protein have high relevance. Accordingly, several authors [29, 66] propose to use the χ^2-based feature selection method discussed in section 1.2.2 to select the most relevant GO terms.

- Second, we observe that selecting individual discriminative functions based on the original χ^2 formula does not consider the network topology and the way different functions interact with each other in the network. For example, independent of how four proteins shown in Figure 1.10 interact with each other, the χ^2 value of each function is same. We believe that for the task of predicting cancer-related proteins, it is possible that a function f_i does not correlate itself with cancer-involvement, but when a protein with function f_i interacts with a

protein with function f_j, this interaction may be an indication of the former protein being involved in a cancer. So, we propose a interaction-based feature selection for predicting cancer-related proteins in PPI networks.

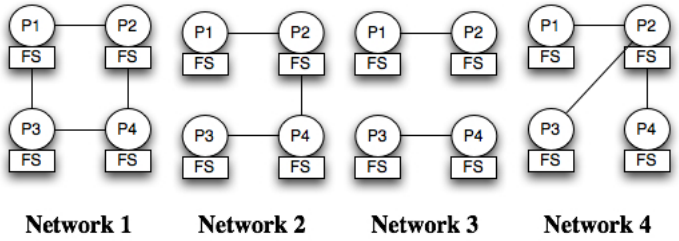


Figure 1.10: Independent of how four proteins P_1, P_2, P_3 and P_4 interact with each other, the χ^2 value of each function $f_i \in FS$ is same.

- Third, we compare our proposed interaction based feature selection with individual-based feature selection with respect to Femasure and we observe that interaction based feature selection outperforms the individual-based method with 7.8%, on average, with respect to Fmeasure.

1.3.4 Predicting Disease-related Proteins in PPI network

In chapter 6, we predict disease-related proteins from information in the relationships among different diseases.

- First, we notice that in almost all the previous methods, prediction accuracy depends directly on the initial disease-related proteins, which we refer to as *seed proteins*. As the initial seed proteins of each disease suffer from several 'False Negative' cases (i.e., disease-related proteins which are not annotated as being involved in disease), dependency of previous methods to the incomplete seed proteins is the main drawback of these methods.
- Second, we propose an informative Human Disease Network (HDN) in which each node is a disease and each weighted edge shows a relationship between two diseases. Each directed edge $d_i \rightarrow d_j$ between two diseases d_i and d_j in the HDN, shows how much seed proteins of disease d_j are predictable based on the information in the seed proteins of disease d_i using a given prediction method M. Although our proposed approach for building the HDN is very general and any prediction method M could be used, the quality of the resulting HDN still depends on the prediction method M. We will discuss some recommended prediction methods in the next step.
- Third, we analyze different *Structural* (using Anova measure and Random Walk) and *Functional* (using χ^2 and interactive-χ^2 as in chapter 5) prediction methods

and we conclude that a hybrid method which considers both structural and functional information in the PPI network is the best method for building the HDN.

- Fourth, we build the HDN based on 20 diseases (Alzheimer ,Amyotrophic, Anemia, Breast cancer, Cataract, Charcot-marie-tooth, Colorectal cancer, Deafness, Diabets, Dystonia, Ehlers-danlos, Epilepsy, Emolytic-anemia, Long QT Syndrome, Lymphoma, Mental-retardation, Parkinson, Usher-syndrome, Xeroderma, Zellweger) and we show that the resulting HDN is biologically meaningful. There are 380 (20×19) possible edges in the original HDN. We prune the HDN by sorting the edges based on their weight descendingly and then, keeping the 38 (10% of the original HDN) highest-weighted edges. Figure 1.11 shows the pruned HDN. For each edge $(d_i) \xrightarrow{rank} (d_j)$, Figure 1.11 shows the rank of the relationship between two diseases d_i and d_j among all the 380 disease pairs. The highest-ranking found relationship is $(deafness) \xrightarrow{1} (usher\ syndrome)$.

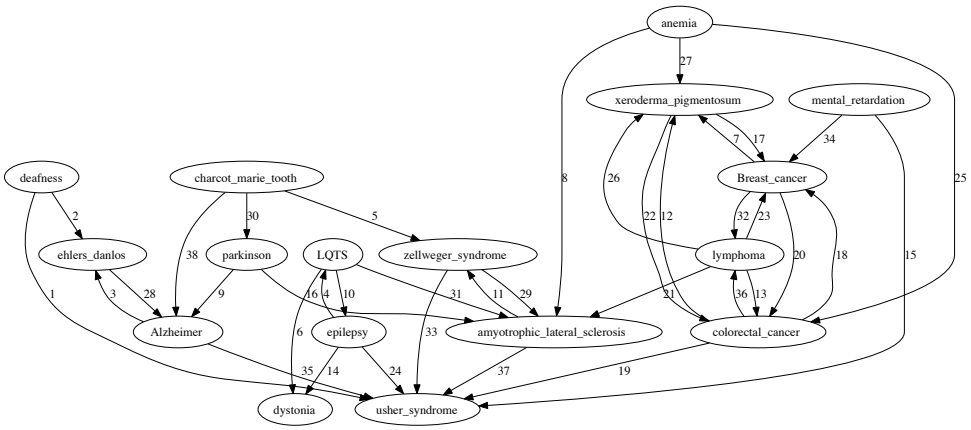


Figure 1.11: Pruned Human Disease Network by keeping only 38 (10% of the original HDN) high-ranked relationships among different diseases. The best found relationship is $(deafness) \xrightarrow{1} (usher\ syndrome)$.

- Fifth, we cluster the HDN and we augment the seed proteins of diseases based on the cluster they belong to. Finally, we predict disease-related proteins based on the augmented version of seed proteins. Literature mining of the newly found disease-related proteins proved the usefulness of using our proposed HDN for predicting disease-related proteins.

1.3.5 Conclusions

In the context of the PPI networks, noisy nature of the networks, high-deminsionality and incompleteness of initial annotation information are potential problems for any knowledges discovery method. To overcome the first two problems, we proposed different feature selection methods using Anova (Analysis of variance) and interaction-based chi-square. To conquer the problem of incompleteness of initial annotation information we proposed a new type of network called Human Disease Network (HDN). We also converted the "Homophily" assumption behind the function prediction methods to "Selective Heterophily" assumption by introducing collaborative functions. There is a "Conclusions" chapter in this thesis which discusses our main contributions in more details.

Chapter 2

Predicting Proteins Functions Using Network Global Information

Based on

Hossein Rahmani, Hendrik Blockeel and Andreas Bender, "Predicting the functions of proteins in PPI networks from global information", JMLR: Workshop and Conference Proceedings, International Workshop on Machine Learning in Systems Biology, Ljubljana, Slovenia, 5-6 September 2009, volume 8, pages 82-97, 2010.

2.1 abstract

In this work we present a novel approach to predict the function of proteins in protein-protein interaction (PPI) networks. We classify existing approaches into inductive and transductive approaches, and into local and global approaches. As of yet, among the group of inductive approaches, only local ones have been proposed for protein function prediction. We here introduce a protein description formalism that also includes global information, namely information that locates a protein relative to specific important proteins in the network. We analyze the effect on function prediction accuracy of selecting a different number of important proteins. With around 70 important proteins, even in large graphs, our method makes good and stable predictions. Furthermore, we investigate whether our method also classifies proteins accurately on more detailed function levels. We examined up to five different function levels. The method is benchmarked on four datasets where we found classification performance according to F-measure values indeed improves by 9 percent over the benchmark methods employed.

2.2 Introduction

In recent years, much effort has been invested in the construction of protein-protein interaction (PPI) networks [118]. Much can be learned from the analysis of such networks with respect to the metabolic and signalling processes present in an organism, and the knowledge gained can also be prospectively employed e.g. to predict which proteins are suitable drug targets, according to an analysis of the resulting network [72]. One particular machine learning task that has been considered is predicting the functions of proteins in the network.

A variety of methods have been proposed for predicting the classes of proteins. On a high level we can distinguish two types of approaches, namely inductive and transductive ones. Inductive learning approaches, also called model-based approaches, construct a model (a mathematical function) that maps a description of a protein onto its functions. Transductive approaches, on the other hand, immediately make predictions for the proteins in the network, without going through the intermediate stage of constructing a model that can be used afterwards for making predictions. The difference between these two will be described more formally in the next section.

Transductive approaches are often "global": information on the whole network is taken into account when making predictions. The inductive approaches that have been used until now are typically local, in the sense that the description of a protein (from which its labels are to be predicted) contains information about the local neighborhood of the protein, not about the network as a whole. This is not an inherent property of inductive approaches, though; one might just as well try to construct a description that contains global information. Accordingly, in this paper we explore the usefulness of one particular kind of global information for the task of protein function prediction, namely the relative position of a protein with respect to specific

other proteins.

This paper is structured as follows. In Section 2 we define the learning problem formally. In Section 3 we briefly review approaches that have been proposed before to solve this problem. In Section 4 we present a new inductive learning approach; we do not present any new learning algorithms but a new description format of proteins, which contains global rather than local information. In Section 5 we empirically evaluate the performance of several learning algorithms when using this format, and, as a control experiment, compare this performance to that of a previously proposed approach. We present our conclusions in Section 6.

2.3 Problem Statement

Mathematically, PPI networks can be represented as graphs, and the problem we consider is that of predicting the labels of nodes in this graph.

Consider an undirected graph G with node set V and edge set E, where each node $v \in V$ is annotated with a description $d(v) \in D$ and, optionally, a label $l(v) \in L$. We assume that there exists a "true" labelling function λ from which l is a sample, that is, $l(v) = \lambda(v)$ where $l(v)$ is defined.

In **transductive** learning, the task is to predict the label of all the nodes. That is, given the graph $G = (V, E, d, l)$, with l a partial function, the task is to construct a completed version $G' = (V, E, d, l')$ with l' a complete function that is consistent with l where $l(v)$ is defined.

In practice, there is an additional constraint that l' should approximate λ. This is imposed by some optimization criterion o, the exact form of which expresses assumptions about λ. For instance, o could express that nodes that are directly connected to each other tend to have similar labels by stating that the number of $\{v_1, v_2\}$ edges where $l'(v_1) \neq l'(v_2)$ should be minimal. The assumptions made about λ are called the bias of the transductive learner.

In **inductive learning**, the task is to learn a function $f : D \to L$ that maps a node description $d(v)$ onto its label $l(v)$. That is, given $G = (V, E, d, l)$, we need to construct $f : D \to L$ such that $f(d(v)) = l(v)$ when $l(v)$ is defined, and f is defined for all elements of D. Note that f differs from l in that it maps D, not V, onto L. This implies, for instance, that it can also make predictions for a node v that was not in the original network, as long as $d(v)$ is known.

Besides the bias expressed by the optimization criterion o (which may still be present), there is now also a bias imposed by the choice of D: whenever two different nodes have the same description, they are assumed to have the same labels: $d(v_1) = d(v_2) \Rightarrow \lambda(v_1) = \lambda(v_2)$. Additionally, the learning algorithm used to learn f has its own inductive bias [79]: given exactly the same inputs, two different learning algorithms may learn different functions f, according to assumptions they make about the likely shape of f.

Thus we have three types of bias. Transductive learners have a transductive bias, which is implied by the choice of the optimization criterion o. Inductive learners have

a description bias, imposed by the choice of d, as well as an inductive bias, imposed by the choice of the learning algorithm that is used to learn f from $(d(v), l(v))$ pairs. In this paper we will explore for one particular description function d wether it represents a suitable description bias.

In the context of protein function prediction in PPI networks, the nodes v are proteins; the descriptions $d(v)$ can be any description of the protein that can be derived from the network structure (where no additional information, such as the protein structure, is assumed to be available; we assume we learn from the network structure only); the labels $l(v)$ are sets of protein functions.

Note that many proteins have more than one function [102]; this is why a node label can be any set of functions. Most off-the-shelf machine learning techniques can only learn classifiers that predict a single value, not a set of values. The fact that node labels are sets may seem to form a problem in this respect. To remedy this situation, if we have n possible functions, the task of predicting a subset of these functions can easily be transformed into n single-function prediction tasks: for each possible function a binary classification task is then constructed where nodes are to be assigned the class true or false depending on whether the protein has that function or not. This is the setting we will focus on in this paper.

2.4 Related work

Among transductive approaches to the protein function prediction problem, the Majority Rule approach has a prominent role [111]. This method assigns to a protein those functions that occur most frequently among its neighbors (typically a fixed number of functions is predicted, for instance, the three most frequently occurring functions in the neighborhood). One problem with this approach is that it only considers neighbors of which the function is already known, ignoring all others. To address this problem, global optimization-based function prediction methods have been proposed. Any probable function assignment to the whole set of unclassified proteins is given a score, counting the number of interacting pairs of nodes with no common function; the function assignment with the lowest value will be the best assignment [121, 119].

Another improvement over the original implementation was made by observing higher-level interactions [18]. Level k interaction between two proteins means that there is a path of length k between them in the network. Proteins that have both a direct interaction and shared level-2 interaction partners have turned out to be more similar to each other (i.e. having same functions). Taking this further, one can make the assumption that in dense regions (subgraphs with many edges, relative to the number of nodes) most nodes have similar functions. This has led to clustering approaches which first cluster the networks (with clusters corresponding to dense regions), and subsequently predict the function of unclassified proteins based on the cluster they belong to [57, 13].

Among the inductive approaches, Milenkovic and Przulj's (2008) graphlet-based approach has been used in the area of protein function predictions. The node descrip-

tion $d(v)$ that is built here, in their terminology the "signature vector", describes the local neighborhood of the node in terms of so-called graphlets, small graph structures as a part of which each node occurs. Most other inductive approaches use similar signatures. Typical for them is that they describe only the local structure of the network near the node to be predicted, however remote changes in the network do not influence the signature at all.

2.5 A global description of proteins

In this work we will now introduce an inductive approach that uses global node descriptions to the area of protein-protein interactions; that is, any change (e.g., addition or removal of an edge) in the network, wherever it occurs, may influence a node's description. Our hypothesis is that the inclusion of additional information will improve the function prediction of unknown nodes which will be investigated in the following in detail.

We describe a node as follows. Assume that there are n nodes in the network, identified through numbers 1 to n. Each node is then described by an n-dimensional vector. The i'th component in the vector of a node v gives the length of the shortest path in the graph between v and node i.

It has been hypothesized before that shortest-path distances are relevant in PPI network analysis; for instance, [103] cluster nodes based on shortest-path distance profiles. As of yet, however, such shortest-path distances have not been considered in the context of inductive learning of protein function predictors which is the rationale behind the current work.

A potential disadvantage of this method is that in large graphs, one gets very high-dimensional descriptions, and not all learners handle learning from high-dimensional spaces well. It is possible, however, to reduce the dimensionality of the vector by only retaining the shortest-path distance to a few "important" nodes. This essentially represents a feature selection problem. A node i is important if the shortest-path distance of some node v to i is likely to be relevant for v's classification. If the feature f_i denotes the shortest path distance to node i, one possible measure of the relevance of f_i for the label of a node (which is a set of functions) is the following.

For each function j, let G_j be the set of all proteins that have that function j. Let $Mean_{k \in G_j}(f_{ik})$ be the average f_i value take over all proteins k in G_j, and $Var_{k \in G_j}(f_{ik})$ the variance of the f_i value take over all proteins k in G_j. The following formula, inspired by ANOVA (analysis of variance), gives an indication of how relevant f_i is for the function set as a whole:

$$\forall p_i \in P; A_i = \frac{Var_j[Mean_{k \in G_j}(f_{ik})]}{Mean_j[Var_{k \in G_j}(f_{ik})]} \tag{2.1}$$

where P is the set of all proteins in the network and F contains all possible functions. Var_j and $Mean_j$ denote the Variance and Mean operators taken over all values of j. A high A_i denotes a high relevance of feature f_i. Figure 2.1 shows the

intuitive representation of formula 2.1. Imagine there are three functions F_1, F_2 and F_3 in the network. First, we put the proteins having the same function in the same group and forming the G_1, G_2 and G_3 groups. Second, in order to calculate the ANOVA value of each protein P_i in the network, we find the shortest path distance of the protein P_i to all the members of each group (i.e., f_{ij}). Finally, we calculate the average and variance of different f_{ij} in each function group G_j.

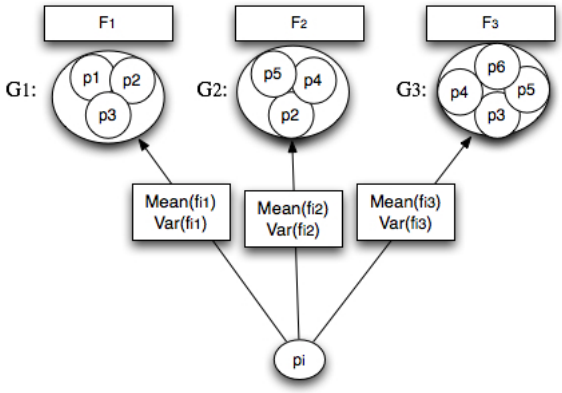


Figure 2.1: Intuitive representation of formula 2.1.

To illustrate this measure, figure 2.2 shows two different scenarios. In the first scenario, all three averages and all three variances are equal. If the X axis shows the value of shortest path distance to protein P_i, then we can not predict one specific function based on the shortest path distance to the protein P_i . So, protein P_i does not discriminate different functions in this scenario and is not an “important" protein. In the second scenario, the values of variances are equal but average values are different. In this scenario, if the shortest path distance of one protein to protein P_i is smaller than μ_1 or bigger than μ_3 then we predict function F_1 or F_3 for that protein respectively. If the shortest path distance is between μ_1 and μ_3 then we predict the function F_2 for that protein. In the second scenario, protein P_i discriminates different functions so we could use it as an important protein.

In the following, we will empirically determine whether the shortest-path distances to all, or a few particular, nodes are indeed informative with respect to a protein's functions by evaluating the performance of the method on a benchmark dataset.

2.6 Experiments

We performed four consecutive experiments.

1. We evaluated the potential of the proposed protein description for protein function prediction by assessing multiple learning systems and finding the learning

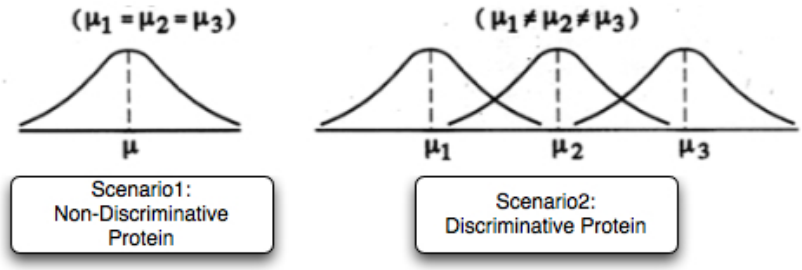


Figure 2.2: Discriminative protein versus non-discriminative protein.

system whose inductive bias best fits our dataset. This step was made to alleviate the risk of concluding that the description is unsuitable, when the cause for bad results is in fact a poor choice of learner.

2. We compared the performance of this system with that of Majority Rule [111], a transductive learner.[1]

3. We investigated the effect of the ANOVA-based node selection criterion on predictive performance: Does a reduction of the number of important nodes increase or decrease the predictive performance, and is there a clear optimum with respect to the number of important nodes that should be selected?

4. While these experiments focused on prediction of functions on the highest level of the functional hierarchy, we check whether our method also yields good predictive accuracy at lower levels.

We evaluate predictive performance using the following measures: area under the ROC curve (AUC) [96], precision, recall, and F-measure. We do not include predictive accuracy (percentage of predictions that are correct) because for several function prediction tasks, the class distribution is highly skewed (e.g., 1% of the protein has that function, 99% does not), and in such cases predictive accuracy (the percentage of predictions that is correct) does not carry much information. AUC and precision/recall are much more robust to skewed class distributions.

2.6.1 Datasets

We apply our method to four *S. cerevisiae* PPI networks: DIP-Core [25], VonMering [123], Krogan [59] and MIPS [76], which contain 4400, 22000, 14246 and 44514 interactions among 2388, 1401, 2708 and 7928 proteins respectively. We consider 17 high level functions for evaluating our function predictors. Figure 2.3 shows high level MIPS function categories with their corresponding function number.

[1]Majority Rule was selected for its ease of implementation, and because it is still a regularly used reference method.

Functional Category
01 METABOLISM
02 ENERGY
10 CELL CYCLE AND DNA PROCESSING
11 TRANSCRIPTION
12 PROTEIN SYNTHESIS
14 PROTEIN FATE (folding, modification, destination)
16 PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic)
18 REGULATION OF METABOLISM AND PROTEIN FUNCTION
20 CELLULAR TRANSPORT, TRANSPORT FACILITIES AND TRANSPORT ROUTES
30 CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM
32 CELL RESCUE, DEFENSE AND VIRULENCE
34 INTERACTION WITH THE ENVIRONMENT
38 TRANSPOSABLE ELEMENTS, VIRAL AND PLASMID PROTEINS
40 CELL FATE
41 DEVELOPMENT (Systemic)
42 BIOGENESIS OF CELLULAR COMPONENTS
43 CELL TYPE DIFFERENTIATION

Figure 2.3: MIPS high level function categories.

2.6.2 Comparison of Learners

Given the input data and a particular function to predict, any standard machine learning tool can be used to build a model that predicts from a node's description, whether the node has a particular function or not. We have experimented with several methods, as available in the Weka data mining toolbox [127], namely decision trees (J48), random forests, an instance based learner (IBk), Naive Bayes, radial basis function networks, Support Vector Machine (libSVM), Classification via Regression (CVR) and Voting Feature Intervals (VFI). We examined three kernel functions namely polynomial, radial basis and sigmoid kernels in the libSVM method, and select the kernel which gives the highest AUC value among the three types of kernel function. These methods were chosen to be representative for a broad range of machine learning methods. This comparative evaluation was made on the DIP-Core data set. The results are shown in Figure 2.4. Looking at average AUC over the functions to be predicted, we see that Random Forests score best (AUC = 0.7), with IBk a close second (0.67). These averages may seem close, but when we look at individual labels, we see that there is only one win and one draw for IBk, and 15 losses, compared to RF. This shows that the difference, while small, is significant.

It is interesting to see that RF performs best among all learners in 13 out of 17 cases, and the 4 cases where it does not are all characterized by a high class skew. (Figure 2.5 visualises this.) This is, in hindsight, not surprising: Random Forests

are ensembles of decision trees, and these are known to perform less well on highly skewed class distributions. In our case, however, while most datasets have a strong class skew, for the large majority of them this is not problematic.

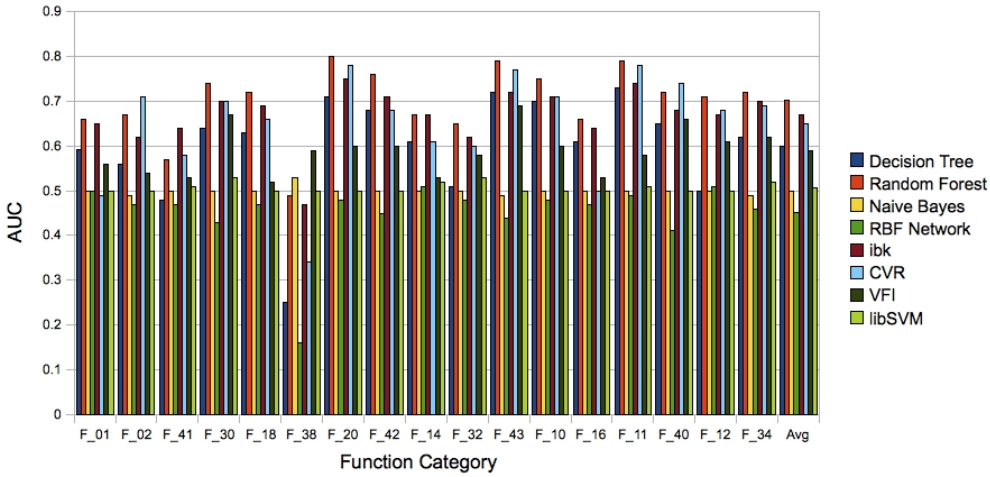


Figure 2.4: Comparison of different machine learning methods on the DIP-Core Dataset.

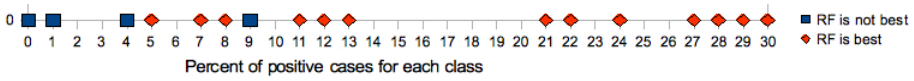


Figure 2.5: Percentage of positive instances in cases where RF does / does not perform best among all learners. The graph shows that whether RF performs best is strongly related to class skew.

We have concluded from the above results that Random Forests are our best candidate for learning from the given type of data, and we have used this method in the remaining experiments.

2.6.3 Comparison with a transductive method

We next compared Random Forests and Majority Rule in predicting the proteins' functions in four datasets DIP-Core, VonMering, Krogan and MIPS. Firstly, we selected 700 nodes based on the ANOVA Measure. Then, we found the shortest path of each protein to those selected proteins. We used this information as the input for Weka and calculated the average Precision, Recall, F-measure and AUC for each function class in a 10-fold cross validation. Figure 2.6 compares the average precision,

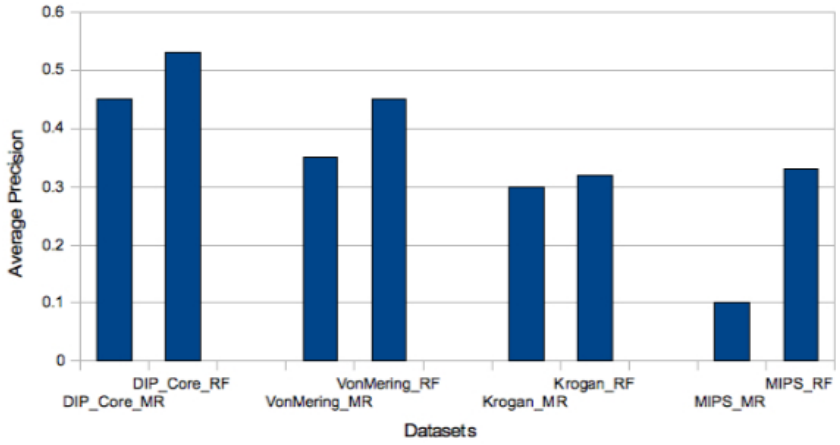


Figure 2.6: Average precision of MR and RF in four datasets.

over all classes, of Majority Rule (MR) and Random Forests (RF). Figure 2.7 similarly compares the recall of MR and RF, and Figure 2.8 the F-measures. We see that, over the four datasets, RF has higher precision (11% higher in average) but smaller Recall (10% smaller in average). RF and MR perform almost similarly with respect to F-measure. The AUCs are compared in Figure 2.9 ; again, RF tends to have higher scores (+6%).

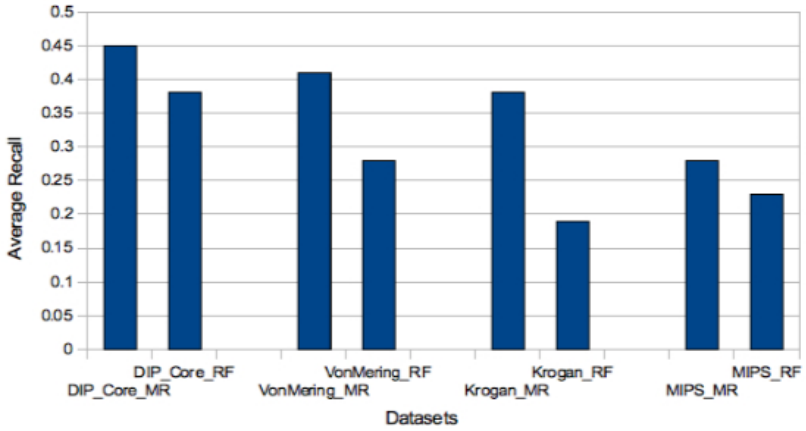


Figure 2.7: Average recall of MR and RF in four datasets.

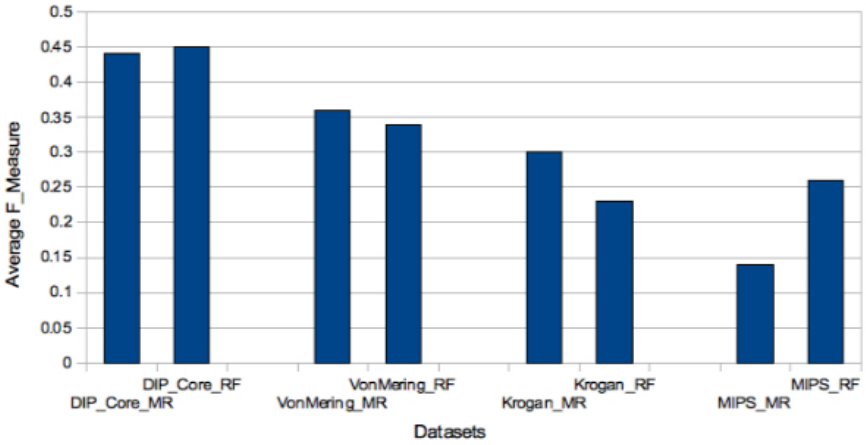


Figure 2.8: Average F-measure of MR and RF in four datasets.

2.6.4 Different Number of Important Proteins

Furthermore, we investigated the effect of selecting a different number of important proteins on the classification metrics. We selected the 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200 and 300 most important proteins according to Equation 2.1. For each number n, we created a dataset where each protein in the network is described using its distance to the n most important proteins. We trained a random forest from that dataset and recorded its precision, recall, F-measure and AUC; these numbers are finally plotted against n. For each combination of function and dataset this gives a separate curve.

The shape of the curves is qualitatively very similar in all cases. Figures 2.10–2.13 show a few representative cases. In general, we observe the following behavior:

- When the number of important proteins is limited to less than 10, this typically yields bad predictive performance. Even though in a few cases the curves already start increasing in this area, they do not reach their maximum.
- In the area of 10-50 important proteins, there is usually a major improvement in all four metrics.
- For most of the functions, selecting 50-70 important proteins is enough to obtain good classification results. Beyond this area, there is usually no major improvement in performance. Nevertheless, in a small number of cases, as visible in Figures 2.10 and 2.13, the performance keeps increasing significantly when the number of important proteins is raised to 300.

We did not systematically increase the number of important proteins beyond 300 because of computational complexity reasons. (Weka's random forest learning method

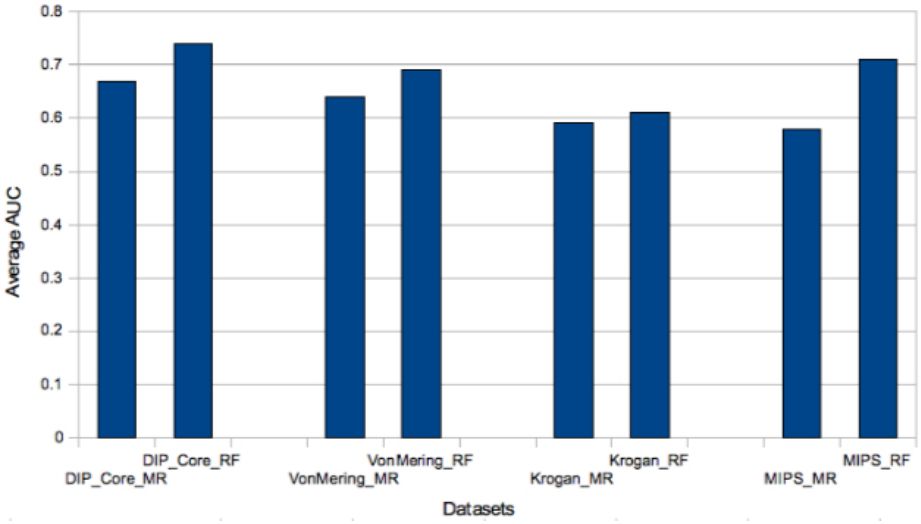


Figure 2.9: Average AUC of MR and RF in four datasets.

is relatively slow for datasets of this size; the fact that it needs to be run for each separate function and dataset, and that each time a ten fold cross-validation is performed, makes it necessary to limit the number of datasets on which it is run.) Nevertheless, our results show that in the large majority of cases a reduction of the number of important proteins to a relatively small number (50-70) is possible without predictive performance suffering too much from this. To really maximize predictive performance, however, experimenting with a larger number of important proteins may be useful.

Looking at the Figures 2.10–2.13, we further notice that the effect of the number of important proteins is much more pronounced for the F-measure than for the AUC. In fact, in several cases there is no clearly perceivable trend in the AUC metric: even though in general AUC tends to go up with increasing F-measure, in some cases it remains relatively constant, and random variations are relatively large compared to the systematic variation. Figure 2.14 illustrates this clearly. Thus, from the point of view of comparing the quality of different models, the F-measure seems a more dependable metric than AUC.

Comparison with Random Selection

The above experiments show that it is possible to reduce the number of important nodes significantly without predictive performance suffering too much, but they do not answer the question whether this is because our ANOVA-based selection method performs well, or because any small number of "important" nodes would simply give us enough information, no matter how we define "important". To answer this question,

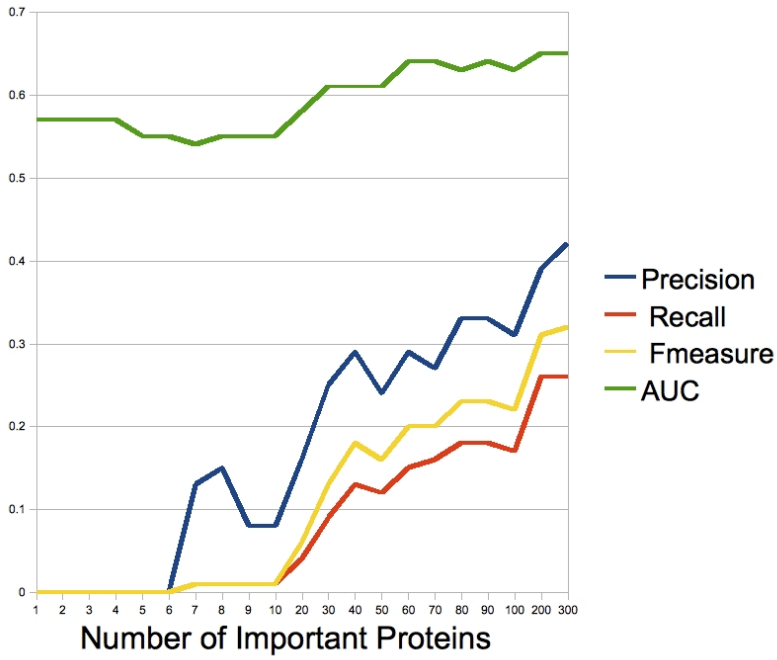


Figure 2.10: Function F-10 in VonMering dataset. Selecting less than 10 important proteins is not enough for discriminating proteins with different function.

we compared our ANOVA-based selection method to random selection (i.e. simply choosing n proteins at random and describing other proteins by their distances to these proteins).

We compared the F-measures obtained when using the ANOVA-based and random selection criteria for eleven different numbers of important proteins: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 300. Since random selection yields different results depending on the random choices made, and this influenced the F-measure quite a bit, for each number of important proteins we ran the random selection based method 20 times and reported average F-measure.

Figure 2.15 compares the ANOVA based selection method with random selection, evaluated according to the F-measure metric, in the DIP-Core dataset. If the number of important proteins is less than 20 or over 100, then there is no big difference between these two methods, however between 20 and 100 selected proteins the ANOVA-based selection method clearly improves upon random selection. This result suggests that when the protein description is very detailed (distances to many other proteins are given), it does not matter what these other proteins are, and when it is very coarse (distances to very few other proteins are given), there is not enough information in

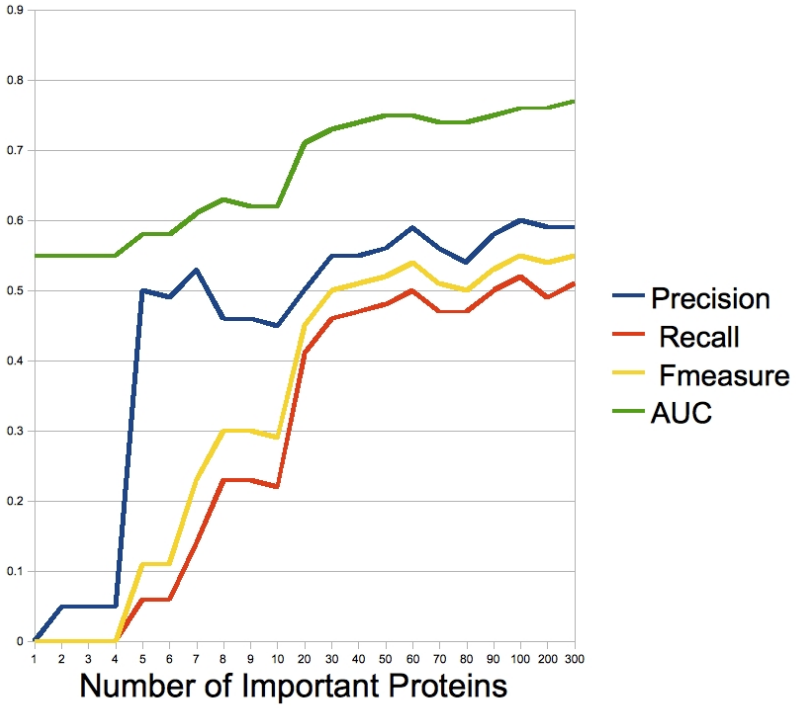


Figure 2.11: Function F-10 in DIP-Core dataset. Major improvement happens when the number of important proteins is between 10 to 50 proteins.

these distances, regardless of whether the important proteins are selected at random or according to the ANOVA criterion. However, in between these extremes, our ANOVA criterion clearly selects proteins with a higher information content than randomly selected proteins.

2.6.5 Function Levels

Proteins' functions have hierarchical structures. As we discussed in the section 2.6.1, we only consider high level MIPS functions. For example, two functions 11.02.01 (rRNA synthesis) and 11.02.03 (mRNA synthesis) are considered similar up to the second function level (i.e. 11.02 =RNA synthesis), but not on deeper levels. In this section, we investigate whether our method also classifies proteins accurately on more detailed levels. We examine up to five different function levels and for each level compare our method with Majority Rule.

Figure 2.16 compares the F-measure obtained by Majority Rule and our method on the DIP-Core and VonMering datasets, for five different function levels 1–5, where

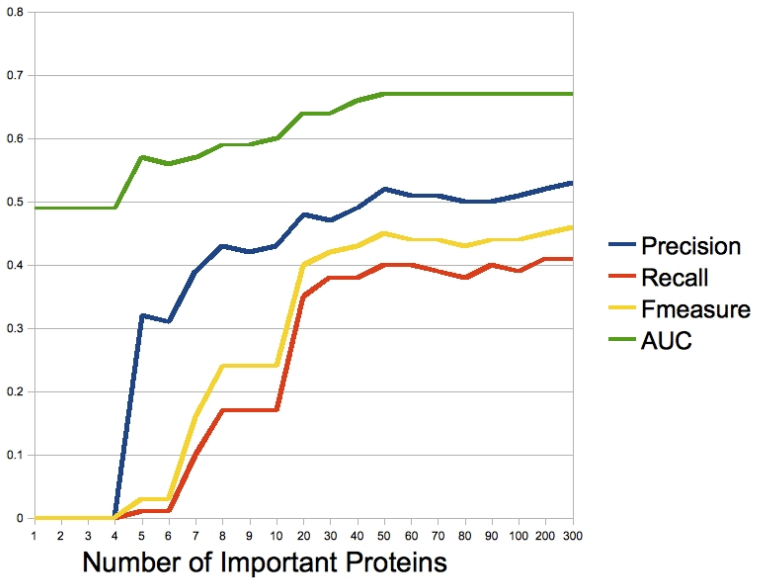


Figure 2.12: Function F-14 in DIP-Core dataset. Uniform metrics' values after selecting 50 important proteins.

1 is the level we used in our earlier experiments. Figure 2.17 visualizes the difference in F-measures between both approaches. The highest improvement is observed for function level 2, when our method has more than 8% higher F-measure value, on average, for the DIP-Core and VonMering datasets. The difference is smallest for very general (level 1) or very specific (level 5) function prediction.

2.7 Conclusions

To summarize, we have firstly classified existing methods for the prediction of node properties in a network into transductive and inductive methods. This distinction provides insight into potential strengths and weaknesses of the methods, particularly in terms of learning bias. Inductive learning methods make different assumptions about the true labeling function than transductive methods, which guided our choice of algorithm employed in this work. Secondly, we observed that existing inductive learning methods for predicting protein functions in PPI networks use local information, while the use of global information for such methods has as of yet remained unexplored. Accordingly, we have, thirdly, introduced a node description formalism that has not been used previously for protein function prediction and which takes global information into account. Together with this node description formalism we have introduced and evaluated a method for reducing the number of features needed

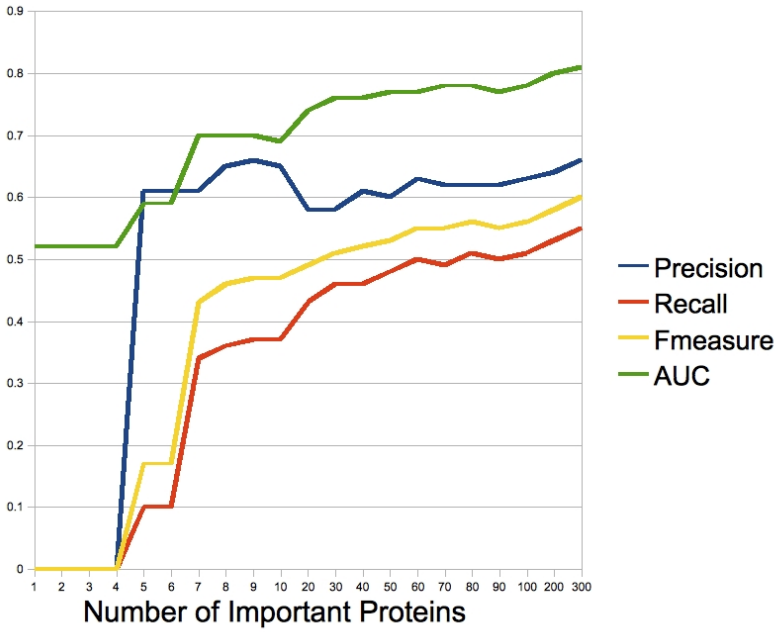


Figure 2.13: Function F-20 in DIP-Core dataset. This is an example of a case where classification accuracy may keep improving when increasing the number of important proteins beyond 300.

for the description. We analyzed the effect of selecting a different number of important proteins on the classification metrics. We found that, for most of the functions, selecting 50–70 important proteins is enough to obtain good classification results. Beyond this area, there is usually no major improvement in performance. Furthermore, we investigate whether our method also classifies proteins accurately on more detailed levels. We examine up to five different function levels. On four benchmark datasets, DIP-Core, VonMering, Krogan and MIPS, we have shown that a standard learner using this formalism outperforms the benchmark Majority Rule approach according to Precision, F-measure and AUC and, hence, that our description formalism is informative with respect to the prediction of a protein's functions from its location in the PPI network.

In the future, a more extensive comparison with other learners would be warranted. It would also be interesting to determine to what extent the information in our global protein description is complementary to that used in other (local inductive, or transductive) approaches. The reason is that when several predictors exploit different information when making their predictions, they can typically be combined into a single composite predictor that performs better than each individual one. Finally,

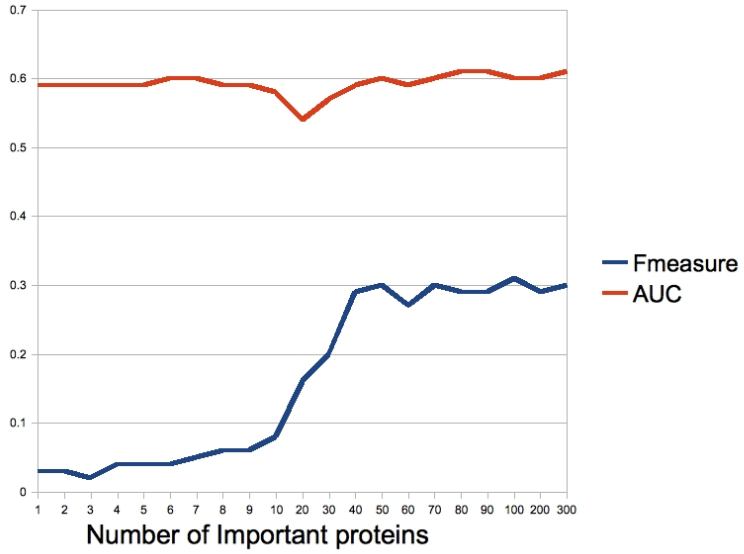


Figure 2.14: Function F-16 in the Krogan dataset. The AUC is more or less constant when varying the number of important proteins, whereas the F-measure shows a clear increase.

while we have focused here on models that predict a single class at a time, there exist a few methods that predict multiple classes simultaneously [9]. Hence, it would be useful to investigate to what extent these classifiers yield better predictions than the single-label prediction approach presented here.

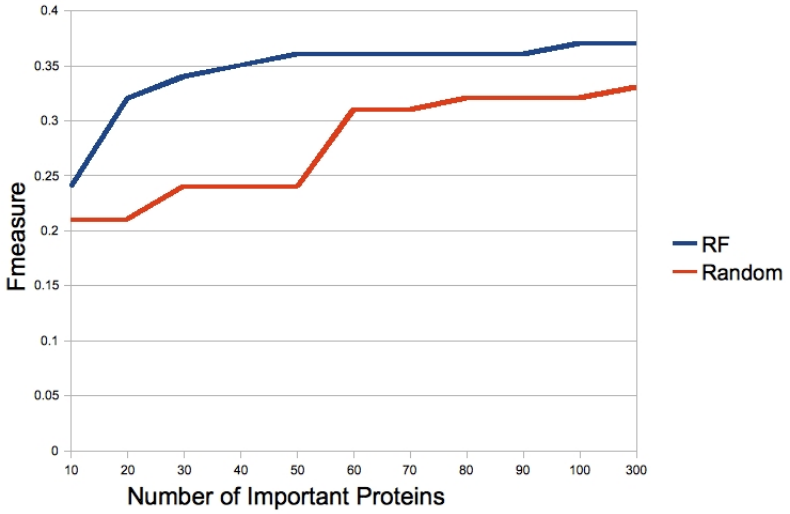
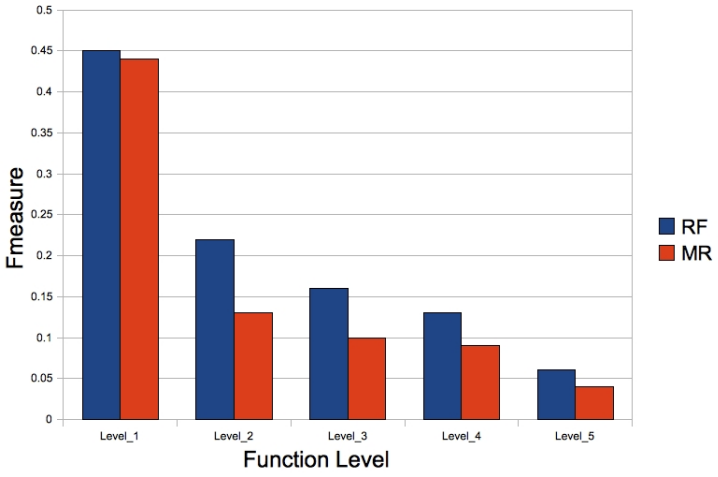
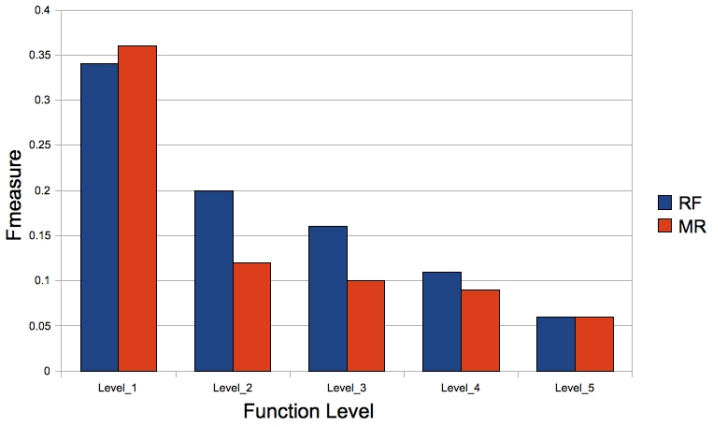


Figure 2.15: Average F-measure obtained with Random Forests using the ANOVA-based feature selection, versus using random selection, in the DIP-Core dataset.

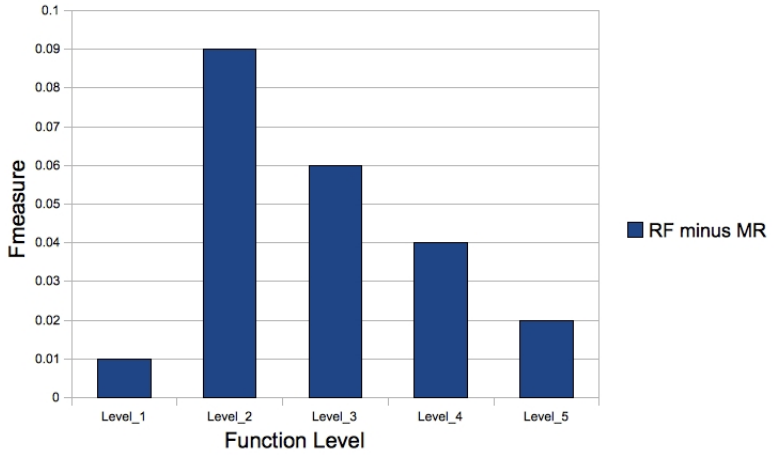


(a) DIP-Core

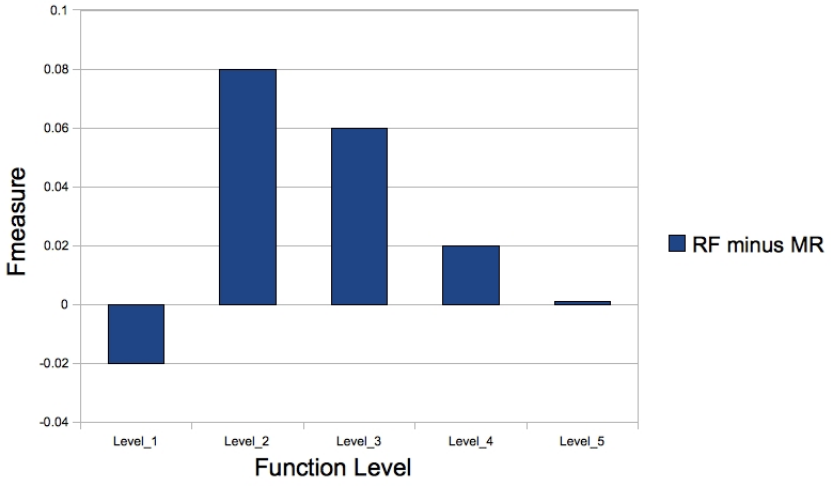


(b) VonMering

Figure 2.16: Average F-measure of MR and RF for different function levels in DIP-Core and VonMering datasets.



(a) DIP-Core



(b) VonMering

Figure 2.17: Difference of RF and MR based on F-measure in five function levels in DIP-Core and VonMering datasets.

Chapter 3

Predicting Proteins Functions Using Collaborative Functions

Based on

Hossein Rahmani, Hendrik Blockeel and Andreas Bender, "Collaboration-based function prediction in protein-proein interaction networks", In: Proceedings of the 10th International Symposium on Advances in Intelligent Data Analysis: 318-327, Lecture Notes in Computer Science 7014 Springer (2011)

3.1 abstract

The cellular metabolism of a living organism is among the most complex systems that man is currently trying to understand. Part of it is described by so-called protein-protein interaction (PPI) networks, and much effort is spent on analyzing these networks. In particular, there has been much interest in predicting certain properties of nodes in the network (in this case, proteins) from the other information in the network. In this paper, we are concerned with predicting a protein's functions. Many approaches to this problem exist. Among the approaches that predict a protein's functions purely from its environment in the network, many are based on the assumption that neighboring proteins tend to have the same functions. In this work we generalize this assumption: we assume that certain neighboring proteins tend to have "collaborative", but not necessarily the same, functions. We propose a few methods that work under this new assumption. These methods yield better results than those previously considered, with improvements in F-measure ranging from 3% to 17%. This shows that the commonly made assumption of homophily in the network (or "guilt by association"), while useful, is not necessarily the best one can make. The assumption of collaborativeness is a useful generalization of it; it is operational (one can easily define methods that rely on it) and can lead to better results.

3.2 Introduction

In recent years, much effort has been invested in the construction of protein-protein interaction (PPI) networks [118]. Much can be learned from the analysis of such networks with respect to the metabolic and signalling processes present in an organism, and the knowledge gained can also be prospectively employed e.g. to predict which proteins are suitable drug targets, according to an analysis of the resulting network [72]. One particular machine learning task that has been considered is predicting the functions of proteins in the network.

A variety of methods have been proposed for predicting the functions of proteins. A large class of them relies on the assumption that interacting proteins tend to have the same functions (this is sometimes called "guilt by association"; it is also related to the notion of homophily, often used in other areas). In this paper we investigate a generalized version of this notion. We rely on the fact that topologically close proteins tend to have *collaborative* functions, not necessarily the same functions. We define collaborative functions as pairs of functions that frequently interface with each other in different interacting proteins. In this way, the assumption becomes somewhat tautological (this definition of collaborative functions implies that the assumption cannot be wrong), but the question remains whether one can, through analysis of PPI networks, correctly identify collaborative functions, and how much gain in predictive accuracy can be obtained by this.

We propose two methods that predict protein functions based on function collaboration. The first method calculates the collaboration value of two functions using an

iterative reinforcement strategy; the second method adopts an artificial neural network for this purpose. We perform a comprehensive set of experiments that reveal a significant improvement of F-measure values compared to existing methods.

The rest of paper is organized as follows. Section 2 briefly reviews approaches that have been proposed before to solve this problem. We present the proposed collaboration-based methods in Section 3, and evaluate them in Section 4. Section 5 contains our conclusions.

3.3 Related work

Various approaches have been proposed for determining the protein functions in PPI networks. A first category contains what we could call structure-based methods. These rely on the local or global structure of the PPI network. For instance, Milenkovic et al. [78] describe the local structure around a node by listing for a fixed set of small graph structures ("graphlets") whether the node is part of such a graphlet or not. Rahmani et al. [98] describe nodes by indicating their position in the network relative to specific important proteins in the network, thus introducing information about the global graph structure.

The above methods do not use information about the functions of other nodes to predict the functions of a particular protein. Methods that do use such information form a second category. A prototypical example is the Majority Rule approach [111]. This method simply assigns to a protein the k functions that occur most frequently among its neighbors (with k a parameter). One problem with this approach is that it only considers neighbors of which the function is already known, ignoring all others. This has been alleviated by introducing global optimization-based methods; these try to find global function assignments such that the number of interacting pairs of nodes without any function in common is minimal [121, 119]. Another improvement over the original Majority Rule method consists of taking a wider neighborhood into account [18]. Level k interaction between two proteins means that there is a path of length k between them in the network. Proteins that have both a direct interaction and shared level-2 interaction partners have been found more likely to share functions [18]. Taking this further, one can make the assumption that in dense regions (subgraphs with many edges, relative to the number of nodes) most nodes have similar functions. This has led to Functional Clustering approaches, which cluster the network (with clusters corresponding to dense regions), and subsequently predict the functions of unclassified proteins based on the cluster they belong to [57, 13].

A common drawback of the second category of approaches is that they rely solely on the assumption that neighboring proteins tend to have the same functions. It is not unreasonable to assume that proteins with one particular function tend to interact with proteins with specific other functions. We call such functions "collaborative" functions. Pertinent questions are: can we discover such collaborative functions, and once we know which functions tend to collaborate, can we use this information to obtain better predictions? The methods we propose next, try to do exactly this.

3.4 Two collaboration-based methods

We propose two different methods. Each of them relies on the assumption that interacting proteins tend to have collaborative functions. They try to estimate from the network which functions often collaborate and, next, try to predict unknown functions of proteins using this information.

In the first method, first we extract the collaborative function pairs from the whole network. Then, in order to make prediction for an unclassified protein, we extract the candidate functions based on position of the protein in the network. Finally, we calculate the score of each candidate function. High score candidate functions are those which collaborate more with the neighborhood of unclassified protein. The second method adopts a neural network for modeling the function collaboration in PPI networks.

We use the following notation and terminology. The PPI network is represented by protein set P and interaction set E. Each $e_{pq} \in E$ shows an interaction between two proteins $p \in P$ and $q \in P$. Let F be the set of all the functions that occur in the PPI network. Each classified protein $p \in P$ is annotated with an $|F|$-dimensional vector FS_p that indicates the functions of this protein: $FS_p(f_i)$ is 1 if $f_i \in F$ is a function of protein p, and 0 otherwise. FS_p can also be seen as the set of all functions f_i for which $FS_p(f_i) = 1$. Similarly, the $|F|$-dimensional vector NB_p describes how often each function occurs in the neighborhood of protein p. $NB_p(f_i) = n$ means that among all the proteins in the neighborhood of p, n have function f_i. The neighborhood of p is defined as all proteins that interact with p.

3.4.1 A Reinforcement Based Function Predictor

Consider the Majority Rule method. This method considers as *candidate functions* (functions that might be assigned to a protein) all the functions that occur in its neighborhood, and *ranks* them according to their frequency in that neighborhood (the most frequent ones will eventually be assigned).

Our method differs in two ways. First, we consider extensions of Majority Rule's candidate functions strategy. Instead of only considering functions in the direct neighborhood as candidates, we can also consider functions that occur at a distance at most k from the protein. We consider $k = 1, 2, 3, 4$ and call these strategies First-FL (First function level, this is Majority Rule's original candidate strategy), Second-FL, Third-FL and Fourth-FL. Finally, the All-FL strategy considers all functions as candidate functions.

The second difference is that our method ranks functions according a "function collaboration strength" value, which is computed through iterative reinforcement, as follows. Let $FuncColVal(f_i, f_j)$ denote the strength of collaboration between functions f_i and f_j. We consider each classified protein $p \in P$ in turn. If function f_j occurs in the neighborhood of protein p (i.e., $NB_p(f_j) > 0$) then we increase the

collaboration value between f_j and all the functions in FS_p:

$$\forall f_i \in FS_p : FuncColVal(f_i, f_j) += \frac{NB_p(f_j) * R}{support(f_j)}$$

If $NB_p(f_j) = 0$, we decrease the collaboration value between function f_j and all the functions belonging to FS_p:

$$\forall f_i \in FS_p : FuncColVal(f_i, f_j) -= \frac{P}{support(f_j)}$$

$support(f_j)$ is the total number of times that function f_j appears on the side of an edge e_{pq} in the network. R and P are "Reward" and "Punish" coefficients determined by the user.

Formula (3.1) assigns a collaboration score to each candidate function f_c:

$$Score(f_c) = \sum_{\forall f_j \in F} NB_p(f_j) * FuncColVal(f_j, f_c) \tag{3.1}$$

High score candidate function(s) collaborate better with the functions observed in the neighborhood of p and are more likely to be predicted as p's functions.

We call the above method "Reinforcement based collaborative function prediction" (RBCFP), as it is based on reinforcing collaboration values between functions as they are observed.

3.4.2 SOM Based Function Predictor

The second approach presented in this work employs an artificial neural network, and is inspired by self-organizing maps (SOMs). From the PPI network, a SOM is constructed as follows. We make a one-layered network with as many inputs as there are functions in the PPI network, and equally many output neurons. Each input is connected to each output. The network is trained as follows. All weights are initialized to zero. Next, the training procedure iterates multiple times over all proteins in the PPI network. Given a protein p with function vector FS_p and neighborhood vector NB_p, the network's input vector is set to NB_p, and for each j for which $FS_p(f_j) = 1$, the weights of the j'th output neuron are adapted as follows:

$$W_{ij,New} = W_{i,j,Current} + LR * (NB_p(j) - W_{i,j,Current}) \tag{3.2}$$

where W_{ij} is the weight of the connection from input i to output j, and LR (learning rate) is a parameter. Intuitively, this update rule makes the weight vector $W_{.j}$ of output j gradually converge to a vector that is representative for the NB vectors of all proteins that have f_j as one of their functions. Once the network has been trained, predictions will be made by comparing the NB vector of a new protein q to the weight vectors of the outputs corresponding to candidate functions, and predicting the k

functions for which the weight vector is closest to NB_q (using Euclidean distance), with k a parameter determined by the user.

Normally, in a SOM, the weights of the winner neurons (the output neurons whose weights are closest to the input) and that of neurons close to them in the SOM lattice are adjusted towards the input vector. The difference with our method is that our learning method is supervised: we consider as "winner neurons" all output neurons corresponding to the functions of the protein. As usual in SOMs, the magnitude of the weight update decreases with time and with distance from the winner neuron. Here, we take some new parameters into consideration which are $LearningRate(LR)$, $DecreasingLearningRate(DecLR)$ and $TerminateCriteria(TC)$ parameters. LR determines how strongly the weights are pulled toward the input vector, and $DecLR$ determines how much LR decreases with each iteration. TC determines when the training phase of SOM terminates: it indicates the minimum amount of change required in one iteration; when there is less change, the training procedure stops.

Algorithm 3.4.1 summarizes the Training phase of the SOM method.

Algorithm 3.4.1: SOM TRAINING PHASE$(LR, DecLR, TC)$

procedure SOM-TRAINING$(LR, DecLR, TC)$
$maxChangeInNetworkWeights \leftarrow 0$;
repeat
  for each $classified\ protein\ p \in P$
  do
    $build\ \ NB_p$
    for each $f_i \in F$
      do $inputNeuron(i) = NB_p(f_i)$
    for each $f_j \in FS_p$
      do
        $apply\ Formula$ (3.2).
        $update\ \ maxChangeInNetworkWeights$
  $LR \leftarrow LR * DecLR$
until $(maxChangeInNetworkWeights < TC)$

3.5 Experiments

3.5.1 Dataset and Annotation Data

We apply our method to three *S. cerevisiae* PPI networks: DIP-Core [25], VonMering [123] and Krogan [59] which contain 4400, 22000 and 14246 interactions among 2388, 1401 and 2708 proteins respectively. The protein function annotation for *S. cerevisiae* PPI networks were obtained from the Yeast Genome Repository [39]. Functions can be described in different levels of detail. For example, two functions 11.02.01 (rRNA synthesis) and 11.02.03 (mRNA synthesis) are considered the same up to the second

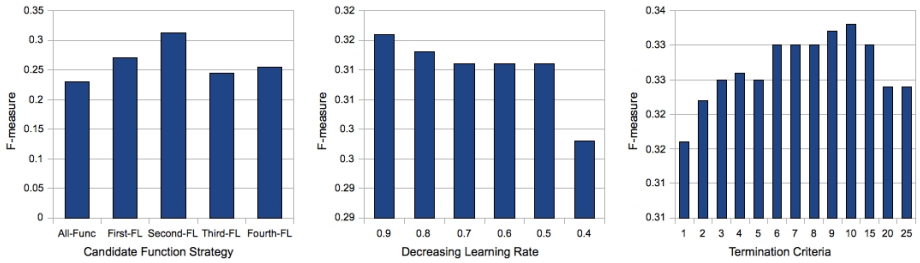


Figure 3.1: Effects of tuning the "Candidate function strategy" parameter, the "Decreasing Learning Rate", and the "Termination Criteria" of SOM network in Krogan Dataset. Second-FL with DecLR=0.9 and TC=10 produces the best result on Krogan.

function level (i.e., 11.02 = RNA synthesis), but not on deeper levels. The function hierarchy we use contains five different levels, which we will refer to as F-L-i. Thus, for each dataset, five different versions can be produced, one for each function level.

3.5.2 Parameter Tuning

Our methods have parameters for which good values need to be found. Parameters can be tuned by trying out different values, and keeping the one that performed best. Obviously, such tuning carries a risk of overestimating the performance of the tuned method, when it is evaluated on the same dataset for which it was tuned. To counter this effect, we tuned our methods on the Krogan dataset labeled with F-L-1 functions (the most general level of functions in the function hierarchy), and evaluated them with the same parameter values on the other datasets; results for DIP-Core and Von Mering are therefore unbiased. Conversely, for the Krogan dataset, we used parameter settings tuned on DIP-Core. This way, all the results are unbiased.

We tuned the parameters manually, using the following simple and non-exhaustive strategy. Parameters were tuned one at a time. After finding the best value for one parameter, it was fixed and other parameters were tuned using that value. For parameters not yet fixed when tuning a parameter p, we tried multiple settings and chose a value for p that appeared to work well on average. With this approach, the order in which the parameters are tuned can in principle influence the outcome, but we found this influence to be very small in practice.

Fig. 3.1 shows the effects of the consecutive tuning of the different parameters. The best value for "Candidate function strategy" parameter is Second-FL; using this value we found a best DecLR value of 0.9, and using these two values we found an optimum for TC at 10. For LR the default setting of 1 was used.

"Candidate function strategy" = Second-FL, $R = 1$ and $P = 2$ turns out to be the best parameter setting for the RBCFP method.

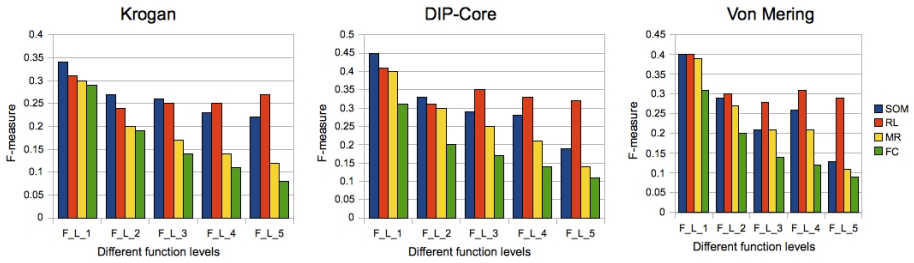


Figure 3.2: F-measures obtained by the new collaboration-based methods (SOM and RBCFP), compared to existing similarity-based methods (MR and FC) at five different function levels, for the Krogan, DIP-Core, and VonMering datasets. On all levels, collaboration based methods predict functions more accurately than similarity based methods.

3.5.3 Comparison to previous methods

In this section, we compare our collaboration-based methods (RBCFP and SOM) with similarity-based methods (Majority Rule and Functional Clustering) on the Krogan, VonMering and DIP-Core datasets, using average F-measure as the evaluation criterion. We perform a leave-on-out cross-validation, leaving out one protein at a time and predicting its functions from the remaining data. For each protein, we predict a fixed number of functions, namely three; this is exactly what was done in the Majority Rule approach we compare to, so our results are maximally comparable. In the proposed methods, we use the parameter values tuned in the previous section. Majority Rule (MR) selects the three most frequently occurring functions in the neighborhood of the protein in the network. Functional clustering (FC) methods differ mainly in their cluster detection technique. Once a cluster is obtained simple methods are usually used for function prediction within the cluster. In our evaluation, we use the clusters from [39] (which were manually constructed by human experts).

Fig. 3.2 compares collaboration-based and similarity-based methods on the Krogan, DIP-Core and VonMering datasets respectively. F-L-i refers to the i'th function level in the function hierarchy. We compare the methods on five different function levels.

In all three datasets, collaboration based methods predict functions more accurately than similarity based methods. As we consider more detailed function levels, the difference between their performance increases. In order to have a general idea about the performance of two method types in different function levels, we take the average of F-measure difference between collaboration based methods and similarity based methods in three datasets. Fig. 3.3 shows the average F-measure difference between two method types. For general function descriptions (first and second function levels), collaboration based methods outperform the similarity based methods with some 4 percent. For more specific function descriptions, for example function level 5,

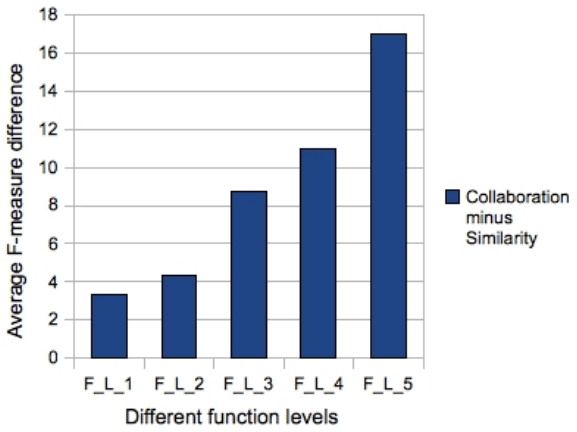


Figure 3.3: Difference in F-measure between the best collaboration-based and the best similarity-based method, averaged over three datasets, for five function levels. The difference increases as we consider more detailed function levels.

the performance difference between two methods increases up to 17 percent.

3.5.4 Extending Majority Rule

We identified the notion of collaboration-based prediction (as supposed to similarity-based prediction) as the main difference between our new methods and the ones we compare with. However, in the comparison with Majority Rule, there is another difference: while Majority Rule assigns only functions from the direct neighborhood to a protein, we found that using candidate functions from a wider neighborhood (including neighbors of neighbors) was advantageous. This raises the question whether majority rule can also be improved by making it consider a wider neighborhood.

We tested this by extending the Majority Rule so that it can consider not only direct neighbors, but also neighbors at distance 2 or 3. We refer to these versions as MR(NB-Li). Fig. 3.4 shows the effect of considering a wider neighborhood in Majority Rule in the three datasets Krogan, VonMering and DIP-Core. There is no improvement in the Krogan and VonMering datasets, and only a small improvement (1%) in DIP-Core, for MR(NB-L2). This confirms that the improved predictions of our methods are due to using the new collaboration-based scores, and not simply to considering functions from a wider neighborhood.

3.6 Conclusion

To our knowledge, this is the first study that considers function collaboration for the task of function prediction in PPI networks. The underlying assumption behind our

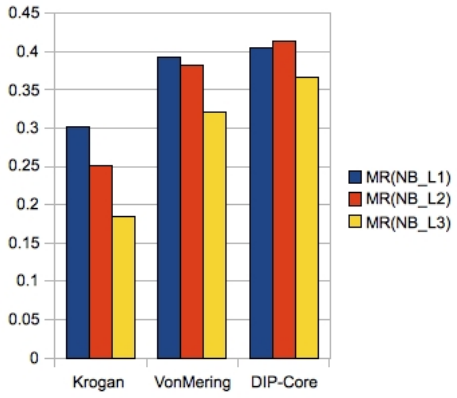


Figure 3.4: Extending Majority Rule by considering other function neighborhood levels. NB-Li represents the 1-, 2- or 3-neighborhood of the protein.

approach is that a biological process is a complex aggregation of many individual protein functions, in which topologically close proteins have collaborative, but not necessarily the same, functions. We define collaborative functions as pairs of functions that frequently interface with each other in different interacting proteins.

We have proposed two methods based on this assumption. The first method rewards the collaboration value of two functions if they interface with each other in two sides of one interaction and punishes the collaboration value if just one of the functions occurs on either side of an interaction. At prediction time, this method ranks candidate functions based on how well they collaborate with the neighborhood of unclassified protein. The second method uses a neural network based method for the task of function prediction. The network takes as input the functions occurring in a protein's neighborhood, and outputs information about the protein's functions.

We selected two methods, Majority Rule and Functional Clustering, as representatives of the similarity based approaches. We compared our collaboration based methods with them on three interaction datasets: Krogan, DIP-Core and VonMering. We examined up to five different function levels and we found classification performance according to F-measure values indeed improved, sometimes by up to 17 percent, over the benchmark methods employed. Regarding the relative performance of the proposed methods, their classification performances are similar in the high level function levels but the RBCFP method outperforms the SOM method in more detailed function levels.

Our results confirm that the notion of collaborativeness of functions, rather than similarity, is useful for the task of predicting the functions of proteins. The information about which functions collaborate, can be extracted easily from a PPI network, and using that information leads to improved predictive accuracy.

These results may well apply in other domains, outside PPI networks. The notion

of homophily is well-known in network analysis; it states that similar nodes are more likely to be linked together. The notion of collaborativeness, in this context, could also be described as "selective heterophily". It remains to be seen to what extent this notion may lead to better predictive results in other types of networks.

Acknowledgements

This research is funded by the Dutch Science Foundation (NWO) through a VIDI grant. At the time this research was performed, Andreas Bender was funded by the Dutch Top Institute Pharma, project number: D1-105.

Chapter 4

Predicting Cancer-Related Proteins Using Network Contextual Information

Based on

Hossein Rahmani, Hendrik Blockeel and Andreas Bender, "Predicting Genes Involved in Human Cancer Using Network Contextual Information" Journal of Integrative Bioinformatics, 9(1):210, 2012.

4.1 Abstract

Protein-Protein Interaction (PPI) networks have been widely used for the task of predicting proteins involved in cancer. Previous research has shown that functional information about the protein for which a prediction is made, proximity to specific other proteins in the PPI network, as well as local network structure are informative features in this respect. In this work, we introduce two new types of input features, reflecting additional information: (1) Functional Context: the functions of proteins interacting with the target protein (rather than the protein itself); and (2) Structural Context: the relative position of the target protein with respect to specific other proteins selected according to a novel ANOVA (analysis of variance) based measure. We also introduce a selection strategy to pinpoint the most informative features. Results show that the proposed feature types and feature selection strategy yield informative features. A standard machine learning method (Naive Bayes) that uses the features proposed here outperforms the current state-of-the-art methods by more than 5% with respect to F-measure. In addition, manual inspection confirms the biological relevance of the top-ranked features.

4.2 Introduction

In recent years, much effort has been invested in the construction of protein-protein interaction (PPI) networks [118]. Much can be learned from the analysis of such networks with respect to the metabolic and signalling processes present in an organism, and the knowledge gained can also be prospectively employed e.g. to the task of protein function prediction [78, 98, 111, 121, 119, 57, 13, 18], identification of functional modules [71], interaction prediction [48, 129], identification of disease candidate genes [106, 37, 132, 87] and drug targets [104, 81], according to an analysis of the resulting network [72].

Wu et al. [130] present an excellent overview of multiple methods for detecting proteins involved in cancer or disease. Among the different methods discussed in [130], "guilt-by-proximity" methods are well known. Methods classified in this category are based on the assumption that genes that directly interact, or, more generally, lie close to each other in the network, are more likely to be involved in the same diseases (as argued by, e.g., Gandhi et al. [31]). The methods vary based on how they define proximity: Some methods consider only direct neighbors to be in the proximity (e.g., [87, 3]), some quantify proximity of two proteins using the length of the shortest-path between them, some compute a "Global Distance Measure" that also takes into account how many paths there are between the two proteins, and how long these are; an example is the approach by Chen et al. [16], who use a PageRank based model for this.

While the basic guilt-by-proximity methods require that certain nodes in the network are already known to be involved in the disease under study, Wu et al. also discuss methods that rely on proximity to nodes known to be involved in other, simi-

lar diseases. Wu et al. define *de novo* methods as methods that can predict nodes to be involved in a particular disease even if no other nodes in the network are known to be involved in it.

The methods discussed by Wu et al. mostly rely on notions of proximity (to genes known to be disease-related) from the area of graph analysis. An entirely different type of approaches are those that rely on feature-based descriptions. There, each individual protein is described by means of a fixed set of features. Next, using machine learning methods, a model is learned that links some of these features to disease-relatedness. In the context of predicting involvement in cancer, examples of feature-based methods include Milenkovic et al.[77], Furney et al. [29] and Li et al. [66]. Milenkovic et al. [77] characterize a protein using a "signature vector" that describes the local network structure around the node in terms of so-called graphlets, small fixed graph structures in which the node occurs. By applying a series of clustering methods, they show that protein that are involved in cancer have similar "topological signatures", which distinguish them from other proteins, and these nodes need not be close to each other in the network. Furney et al. [29] use the Gene Ontology annotations of a protein as features, as well as a number of other properties; they use a chi-square-based selection criterion to select the likely most relevant features, then apply Naive Bayes. Li et al. [66] compare three classifiers: SVM, Naive Bayes and logistic regression and they find that the SVM classifier on average performed slightly better than the Naive Bayes and logistic regression methods, and that among SVMs using different types of features individually, including GO annotations as features gives the best performance, while sequence and conservation features have relatively weak predictive power.

When learning from PPI networks, feature-based approaches have a number of advantages over proximity-based approaches. First, defining the problem in a machine learning setting gives access to a wide range of machine learning techniques, making this type of approaches very flexible. Second, data integration is more easily achieved: one can easily define additional features for proteins, possibly using background information (i.e., information external to the PPI network) for this. Third, these method are inductive: they do not yield predictions, but a model for making predictions. This is interesting in terms of Wu et al.'s definition of *de novo* methods. Information about disease genes is needed when *constructing* the model, but not when *applying* it, so the model can be applied to other PPI networks, or in other areas of the same PPI network. Finally, inductive methods can yield interpretable models, which may by themselves yield new insights.

A difficulty with feature-based methods, however, is that the quality of the learned model depends on the features used. When the input data is a PPI network, the main challenge is to find features with good predictive power that can be computed from this network. The approaches mentioned above all do this in some way. In this work, we propose two new types of input features, reflecting additional information that can be extracted from a PPI network: (1) Functional Context: the functions of proteins interacting with the target protein (rather than the protein itself); (2) Structural Context: the relative position of the target protein with respect to specific other proteins selected according to a novel ANOVA (analysis of variance) based

measure. We show that these features have good predictive power. It is not our goal to compare different machine learning algorithms; we restrict ourselves to the Naive Bayes classifier. The performance of the method might be optimized by using another learning method, but we expect the difference to be small (see also Li et al. [66]). Our main claim lies in the usefulness of the new features.

4.3 Methods

4.3.1 Formal Definition

We consider a PPI network as an undirected annotated graph (P, E, λ) where P is a set of proteins, $E \subseteq P \times P$ is a set of interactions between these proteins, and λ is a so-called annotation function; for each p, $\lambda(p)$ denotes the additional information we have about p (for instance, its GO annotations). In this work, we assume that $\lambda(p)$ simply lists all the GO functions that are associated with p; we call it the function set (or function vector) of p, and denote it $FS(p)$. If $F = \{f_1, f_2, \ldots, f_n\}$ is the set of all the functions in the network, then $FS(p)$ is an $|F|$-dimensional binary vector; the i'th component of $FS(p)$, denoted $FS_i(p)$, is 1 if function f_i is associated with p, and 0 otherwise. We will also write $f_i \in FS(p)$ to denote $FS_i(p) = 1$.

4.3.2 Protein Description Based on Functional Context

Given a protein p, we define the interactor set of p, denoted $IS(p)$, as the set of proteins it interacts with, i.e., $IS(p) = \{q|(p, q) \in E\}$. Besides the function vector of p itself, we also define the "interacting function counts" vector $IFC(p)$ as the number of interacting proteins that are annotated with that function.

$$IFC(p) = \sum_{q \in IS(p)} FS(q) \tag{4.1}$$

Note that, while methods for predicting involvement in cancer have considered GO annotations of proteins as predictive features (e.g., [29, 66]), no methods up till now have considered GO annotations of the neighbors of those proteins at the same time. That is, for predicting involvement in cancer of a protein p, the $FS(p)$ vector has been considered as a predictive feature, but the vector $IFC(p)$ has not. One may wonder what the advantage is of using GO annotations of related proteins, rather than the protein itself. One argument is that GO annotations are often incomplete, and by collecting GO information from the neighbors of a protein p, we may get more information about p itself. This argument is backed up by the fact that GO annotations of proteins can often be predicted well from the GO annotations of their neighbors; see, e.g., [111, 99]. However, as we will show, this is not the only effect; there is also a direct relationship between a protein's involvement in cancer and the GO annotations of the proteins it interacts with.

We will refer to the information in $FS(p)$ and $IFC(p)$ as the *functional context* of p. We experimentally compare two different versions of this functional context: using $FS(p)$ only as input vector (i.e., ignoring the information in the neighborhood of p), and using the sum of $FS(p)$ and $IFC(p)$ as input vector (thus taking into account functional information about the neighborhood of p, including p itself). We call these two approaches FS and $FS + IFC$, respectively.

As defined above, the $FS(p)$ and $IFC(p)$ vectors have high dimensionality; the number of components equals the number of functions in the Gene Ontology. A natural way to reduce this dimensionality is using a feature selection method to filter out the least interesting features (functions, in this case). An often used measure for determining the relevance of a binary feature F for a class variable C is the χ^2 score, defined as follows:

$$\chi^2 = \frac{(ad - bc)^2 * (a + b + c + d)}{(a + b)(c + d)(b + d)(a + c)} \tag{4.2}$$

where a, b, c and d are defined by the contingency table in Table 4.1.

Table 4.1: **The contingency table of a binary feature F w.r.t. a binary class variable C**

	$F = 0$	$F = 1$	total
C=0	a	b	a+b
C=1	c	d	c+d
	a+c	b+d	a+b+c+d

a, b, c, and d count the number of times F and C have the corresponding value. The χ^2 value of F w.r.t. C is derived from this.

In our case, the class variable C indicates whether a protein p is involved in cancer or not, and the binary feature F indicates whether a particular component of $FS(p)$ or $FS(p) + IFC(p)$ is zero ($F = 0$) or not ($F = 1$).

Apart from allowing us to reduce the dimensionality of the vectors describing a protein p, the χ^2 measure also ranks functions from highly relevant (for predicting involvement in cancer) to less relevant.

4.3.3 Protein Description Based on Structural Context

Besides the functional context of a protein, defined before, we will also consider its so-called *structural context*. This structural context relates to the relative position of p in the network.

Several methods discussed in Wu et al. [130] describe each protein p based on the shortest-path distance of p to some previously known cancer/disease proteins. We refer to this category of methods as "distanceToCancer" methods (DisToCancer).

Alternatively, we can describe a protein's position relative to other proteins than only cancer-related ones. Rahmani et al. [98] proposed a relevance measure for

proteins that is inspired by statistical ANOVA (analysis of variance), and showed that shortest-path distance to a relatively small number of proteins (selected according to the ANOVA-based measure) is informative for the task of function prediction in the PPI networks. Since the ANOVA method works well for function prediction, it is natural to check whether it also gives good results for the task of predicting cancer-related proteins, and this is one of the purposes of the current study. We therefore propose the use of similar features for predicting proteins involved in cancer.

The ANOVA-inspired selection measure (briefly, ANOVA) is defined as follows. Let P^+ be the set of proteins labeled as being involved in cancer, and P^- the set of proteins not labeled as such. For each protein q, we introduce a feature d_q; $d_q(p)$ denotes the shortest-path distance between p and q (viewed here as a feature of p). We consider for each q the mean and variance of $d_q(p)$, taken over all cancer-related and non-cancer-related p:

$$m_q^+ = \frac{\sum_{p \in P^+} d_q(p)}{|P^+|} \tag{4.3}$$

$$m_q^- = \frac{\sum_{p \in P^-} d_q(p)}{|P^-|} \tag{4.4}$$

$$var_q^+ = \frac{\sum_{p \in P^+} (d_q(p) - m_q^+)^2}{|P^+| - 1} \tag{4.5}$$

$$var_q^- = \frac{\sum_{p \in P^-} (d_q(p) - m_q^-)^2}{|P^-| - 1} \tag{4.6}$$

Seeing P^+ and P^- as two groups of proteins, the following formula compares the variance between groups to the variance within groups (as it is used for relative ranking only, constant factors are dropped):

$$A_q = \frac{(m_q^+ - m_q^-)^2}{var_q^+ + var_q^-} \tag{4.7}$$

A high A_q means that d_q varies little within groups and/or much between groups, which indicates that d_q has high predictive power for the group. Features d_q can be ranked according to A_q, and the top-k features selected as actual features to be included in the description of all proteins. We will call the category of methods that use these features DisToAnova methods, or DisToAnova(k) when referring to a particular setting for the parameter k.

Finally, we can combine the information in the DisToCancer and DisToAnova descriptors; we do this by first filtering the proteins, retaining only those known to be involved in cancer, and ranking these according to the Anova criterion. This combined version is referred to as DisToCancerAnova.

4.3.4 Protein Description Based on Functional and Structural Context

This refers to protein descriptions that include both information from functional and structural context. The input consists of the $FS + IFC$ vector concatenated with the DisTo(Cancer/Anova/CancerAnova) vector.

4.4 Results

4.4.1 Dataset

We evaluate our methods on the dataset used by Milenkovic et al. [77]. This dataset is the union of three human PPI datasets: HPRD [91], BIOGRID [116] and the dataset used by Radivojac et al. [97]. When we say "union", we mean that the new network contains all the nodes and edges (proteins and interactions) found in either of these networks. The aim of merging these three datasets was to obtain as complete a human PPI network as possible, i.e., a network that covers with its edges as many proteins in the human proteome as possible. We denote as "known cancer genes" the set of genes implicated in cancer that is available from the following databases: Cancer Gene Database [23], Cancer Genome Project-the Cancer Gene Census [95], GeneCards [32] Kyoto Encyclopedia of Genes and Genomes [83] and Online Mendelian Inheritance in Man [47]. Some statistical information is shown in Table 4.2. We have chosen to evaluate our methods on this dataset to make a precise quantitative comparison to their graphlet-based method possible.

Table 4.2: Statistical information of union of three human PPI datasets: HPRD [91], BIOGRID [116] and Radivojac et al. [97]

Number of Proteins	10,282
Average Degree	9.201
Min Degree	1
Max Degree	272
Number of Cancer Genes	939

While the dataset employed here is of high quality, as it is based on large and widely employed datasets, it should be kept in mind that it is not trivial (or in the narrow sense probably even impossible) to define it in a flawless fashion. One of the limitations lies in the role of 'genes involved in cancer' - cancers are different, so while a gene may play a role in one cancer, it might play no role at all in another one. Also, there are spatial and temporal conditions involved in the annotation we do not include here. On the other hand, a limitation lies in the construction of the interactome we define in our dataset. Again, temporal conditions are excluded, and likely many interactions have not been identified in experiment yet; hence our dataset

likely contains a substantial number of missing annotations (while likely also false positive interactions are included due to experimental noise and errors). Nonetheless, the dataset employed here is as good as we can do currently both in size and quality; and, in particular, it has been employed in related studies before, which enables us to perform benchmark experiments in a comparative manner by utilizing it.

This dataset determines uniquely the network structure, and therefore the values of all features, used in our experiments. The actual datasets we use differ only with respect to what features are included.

4.4.2 Biological Interpretation of the Most Relevant Functions

The number of different functions occurring in our human dataset is 9833; this is also the dimensionality of FS and IFC if no dimensionality reduction is used. As mentioned before, we can use a χ^2-based feature selection method to reduce this number; at the same time, this technique ranks functions according to how relevant they are for prediction of cancer involvement.

Most Relevant Functions in FS

Tables 4.3, 4.4 and 4.5 show the 20 highest ranked functions. As the Gene Ontology actually uses three domains (biological function, molecular function, cellular component), we have separated the functions according to their domain.

Searching the most relevant functions in the cancer literature proved the usefulness of chi-square for detecting these functions. For example, based on cancer literature, function "GO:0008284" is involved in various cancers: "Breast Cancer", "Prostate Cancer" and "Lung Cancer". Besides using the statistic to select a limited number of features, we can also use it to inspect the top-ranked functions, which can be used both as a soundness check (are the functions that we expect to be relevant indeed highly ranked?) and as a method for discovering potentially new information (when there are unexpected functions among the top-ranked ones).

Many of the biological functions contained in Table 4.3 are obviously related to cancer, such as GO:0008284, the Positive regulation of cell proliferation, which is a synonym for uncontrolled cell growth, as are positions 3, 5 and 10 in the list (GO:0045944 Positive regulation of transcription from RNA polymerase, GO:0006355 Regulation of transcription, DNA-dependent and GO:0045941 Positive regulation of transcription). Similarly, position 4 (GO:0008285 Negative regulation of cell proliferation) has an obvious connection to cancer; where positive stimulation of cell growth can stimulate tumor growth, an inhibition of the negative regulatory elements will have the very same effect. Fibroblasts are involved in wound healing, a process not taking place properly in cancerous settings [50]. We can also find biological processes linked to small molecules in the list, at positions 6 and 7, namely GO:0014070 Response to organic cyclic substance and GO:0042493 Response to drug. It is known that many carcinogenic substances such as benzo[a]pyren, or even smaller molecules

Table 4.3: **Most discriminative functions from Biological Process based on *FS* method**

Index	Function	Short Info	chi-square	p-value
1	GO:0008284	Positive regulation of cell proliferation	163.02	<0.0001
2	GO:0008543	Fibroblast growth factor receptor signaling pathway	105.80	<0.0001
3	GO:0045944	Positive regulation of transcription from RNA polymerase II promote	99.65	<0.0001
4	GO:0008285	Negative regulation of cell proliferation	71.40	<0.0001
5	GO:0006355	Regulation of transcription, DNA-dependent	69.84	<0.0001
6	GO:0014070	Response to organic cyclic compound	69.76	<0.0001
7	GO:0042493	Response to drug	69.18	<0.0001
8	GO:0043434	Response to peptide hormone stimulus	67.09	<0.0001
9	GO:0001658	Branching involved in ureteric bud morphogenesis	64.91	<0.0001
10	GO:0045941	Positive regulation of transcription	64.89	<0.0001
11	GO:0007050	Cell cycle arrest	62.73	<0.0001
12	GO:0001656	Metanephros development	62.07	<0.0001
13	GO:0032355	Response to estradiol stimulus	59.99	<0.0001

Table 4.4: **Most discriminative functions from Molecular Function based on *FS* method**

Index	Function	Short Info	chi-square	p-value
1	GO:0016563	Transcription activator activity	88.85	<0.0001
2	GO:0004713	Protein tyrosine kinase activity	84.60	<0.0001
3	GO:0003700	Sequence-specific DNA binding transcription factor activity	83.11	<0.0001
4	GO:0005515	Protein binding	82.88	<0.0001
5	GO:0004716	Receptor signaling protein tyrosine kinase activity	68.76	<0.0001
6	GO:0043565	Sequence-specific DNA binding	67.69	<0.0001

such as benzene, are linked to cancer risk. Unfortunately, one of the limitations of the GO terms is their low selectivity; hence the term 'response to drug' remains rather vague. Positions 9 and 12, GO:0001658 Branching involved in ureteric bud morphogenesis and GO:0001656 Metanephros development, are both linked to growth factors, and hence in turn to the development of cancers.

Molecular functions returned as significantly enriched among cancer genes, listed

Table 4.5: **Most discriminative function from Cellular Component based on FS method**

Index	Function	Short Info	chi-square	p-value
1	GO:0005634	A membrane-bounded organelle of eukaryotic cells in which chromosomesare housed and replicated.	99.76	<0.0001

in Table 4.4, frequently refer to transcription factor (position 1) and kinase activity (positions 2 and 5). On the other hand, the cellular component category was less revealing, only listing one significantly enriched category related to cancer genes - the nucleus (where increased transcription takes place, leading to uncontrolled cell growth). Unfortunately, the GO term employed is too general to draw more detailed conclusions from this analysis.

Most Relevant Functions in $FS + IFC$

18 out of 20 functions with the highest χ^2, calculated based on the $FS+IFC$ method, belong to the Biological Process ontology and are listed in Table 4.6. As is apparent from Table 4.6 (when compared to Table 4.3, which results from the use of the FS method), very different discriminative GO terms from the Biological Process ontology are retrieved. Many biological processes retrieved by this method seem to be more specific, such as GO:0043491 at position 1, naming the protein kinase B signaling cascade as involved in cancerogenesis (which is known from literature [12]). It is interesting that now also secondary processes known to be relevant for cancerogenesis and, in particular, cancer growth and the formation of metastases, are captured (which was not the case by purely applying the FS method), such as at position 6 (GO:0001525) for the formation of blood vessels essential for the rapid growth of cancerogenous tissue, and at position 12 (GO:0030335) with respect to cell migration, important for the formation of metastases. Also novel in the list are biological processes related to insulin and the insulin-like growth factor receptor (IGFR), at positions 2 (GO:0048009) and 5 (GO:0032869). This is supported by literature, as insulin has been linked to pancreatic cancer development [28], while the literature regarding insulin-like growth factor receptor is still inconclusive [19, 88]. Still, due to their apparent role in cell proliferation, it is certainly a possibilty that IGFRs play a role in the development of at least some cancer subtypes.

As shown in Table 4.6, chi-square values when calculated based on $FS + IFC$ are greater than the chi-square values when we use the FS method for the calculation, illustrating how our additional annotations add information to the feature selection step; P-values of all the highly ranked functions are < 0.0001 which is very significant.

Overall, from the discussion above, it becomes apparent that the $FS + IFC$ method, as proposed in this work, is able to retrieve significantly different biologi-

cal processes, compared to using the FS method; thus it adds to the information that can be gained from the same set of data. Hence, we suggest it to be a worthwhile method to be employed in the analysis of signaling networks, as shown in this particular case study.

Table 4.6: **Most discriminative functions from Biological Process based on $FS + IFC$ method**

Index	Function	Short Info	chi-square	p-value
1	GO:0043491	Protein kinase B signaling cascade	280.70	<0.0001
2	GO:0048009	Insulin-like growth factor receptor signaling pathway	231.72	<0.0001
3	GO:0008284	Positive regulation of cell proliferation	223.62	<0.0001
4	GO:0034097	Response to cytokine stimulus	223.05	<0.0001
5	GO:0032869	Cellular response to insulin stimulus	218.56	<0.0001
6	GO:0001525	Angiogenesis	213.14	<0.0001
7	GO:0043066	Negative regulation of apoptosis	211.77	<0.0001
8	GO:0001701	In utero embryonic development	208.366	<0.0001
9	GO:0009887	Organ morphogenesis	207.71	<0.0001
10	GO:0042493	Response to drug	205.81	<0.0001
11	GO:0030097	Hemopoiesis	202.45	<0.0001
12	GO:0030335	Positive regulation of cell migration	202.38	<0.0001
13	GO:0051091	Positive regulation of sequence-specific DNA binding transcription factor activity	194.37	<0.0001
14	GO:0046326	Positive regulation of glucose import	194.13	<0.0001
15	GO:0043627	Response to estrogen stimulus	192.34	<0.0001
16	GO:0044419	Interspecies interaction between organisms	191.29	<0.0001
17	GO:0014070	Response to organic cyclic compound	189.94	<0.0001
18	GO:0045944	Positive regulation of transcription from RNA polymerase II promoter	189.32	<0.0001

4.4.3 Biological Interpretation of the Most Discriminative Proteins

Our dataset contains 10,282 proteins. The DisToAnova method uses the ANOVA measure to select the most relevant among these. More detailed information could be obtained from an ANOVA analysis of the most relevant proteins among the full set of 10,282 proteins. Table 4.7 shows the 10 proteins with the highest ANOVA measure.

Zinc finger protein (ZNF467) is known to be upregulated in a variety of breast cancers; however usually its close link with BRCA1 has been seen as the reason for its causal relation with cancers [56]. STATIP1 is involved in histone H3 and H4 acetylation and its interactions with STAT3 and JAK1/2 - which are all involved in cell growth and differentiation processes - have been documented in literature [22]. JUNB

has been documented as a proto-oncogene and IL22 (along with its subunit IL22RA2) is involved in Stat3 phosphorylation [15]. FGFR4 (fibroblast growth factor receptor 4) is associated with cancer nearly by definition (and in alignment with fibroblasts being identified earlier in the context of biological functions). The chemokine ligand 1 receptor, CCL1, has been implicated in cancer before and also it has been suggested as a therapeutic target in this context [45]. Platelet derived growth factor C (PDGFC) is part of the PDGFR-alpha signalling pathway and the influence of PDGFR expression on metastatic behaviour has been well documented [126]. STAT1 is involved in cell growth processes [133], hence its appearance in this list is reasonable. C20ORF185 is an interesting case in that it is annotated as possibly being involved in recognizing/binding specific classes of odorants or serving as a defence mechanism by removing pathogenic microorganisms from the mucosa [24]. On the other hand, its recommended name is the "Long palate, lung and nasal epithelium carcinoma-associated protein 3 precursor", rendering its inclusion in the list of proteins most involved in cancer reasonable.

Table 4.7: **Most discriminative proteins based on ANOVA measure**

Index	Protein Name	Official Full Name
1	ZNF467	Zinc finger protein 467
2	STATIP1	Elongator complex protein 2
3	JUNB	Transcription factor jun-B
4	IL22RA2	Interleukin-22 receptor subunit alpha-2
5	FGFR4	Fibroblast growth factor receptor 4
6	CCL1	Cyclin associated with protein kinase Kin28p
7	PDGFC	Platelet-derived growth factor C
8	IL22	Interleukin 22
9	STAT1	Signal transducer and activator of transcription 1-alpha/beta
10	C20ORF185	Long palate, lung and nasal epithelium carcinoma-associated protein 3

4.4.4 Comparing Different Contextual Methods

We divided the dataset into a training set containing 90%, and a test set containing the remaining 10%, of the proteins for the selection of contextual method and tuning. For the final evaluation we use 10-fold cross validation. Features were selected according to the above-mentioned χ^2 and ANOVA methods; in both cases only the training set was used to rank features according to relevance. We have varied the number of features (functions for functional context methods, proteins for structural context methods) from low to high, in order to investigate the effect of this parameter on predictive performance.

With each method, we predict the cancer-relatedness of the nodes in the test set using our various methods, and evaluate the predictions according to Recall, Precision and F-measure:

$$Precision = \frac{tp}{tp + fp} \tag{4.8}$$

$$Recall = \frac{tp}{tp + fn} \tag{4.9}$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4.10}$$

where proteins involved in cancer are considered as the positive class, and tp, fp and fn denote the number of true positives, false positives, and false negatives, respectively.

Figure 4.1 shows the evaluation metrics for two functional context methods FS and $FS + IFC$ in different function counts: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500 with respect to F-measure, Precision and Recall. Independent from the function count, the $FS + IFC$ method always outperforms the FS method with respect to F-measure and this proves our assumption about the usefulness of considering the whole functional context of proteins (not just the functions of the protein itself but also those of its neighbors) for predicting the proteins involved in cancer. The best obtained F-measure with FS is 29% while the best obtained F-measure for $FS + IFC$ is 37% in one case and 35% in three cases.

These results show that considering the functional annotation of the neighbors allows for more accurate prediction of which genes are involved in cancer. Since it was already known that the functional annotation of a protein's neighbors can be used to predict the protein's own functions [111, 99], and that the protein's own functions are relevant for its involvement in cancer [29, 66], one might wonder to what extent our results are simply a consequence of these two facts. We can test this by enriching proteins in the PPI network with predicted GO annotations (predicted from the GO annotations of their neighbors), and next applying the FS method. We tested this by using a Majority Rule method [111] for enriching the GO annotations of the proteins, in two different ways. In the first approach, we perform function prediction for each protein p which $|FS(p)| = 0$ (reasoning that if a protein is not annotated with any functions, it is likely that its functions are simply not known), while in the second, less conservative, approach, we extend the function set of each protein p with $|FS(p)| < 10$ to a total of ten functions. In the notation employed here, the $||$ operator returns the size of the function set of protein p, with 10 being the average function count of proteins in our dataset before applying the Majority Rule method. We call the enriching approaches "Unclassified" and "Extended", respectively. Figure 4.2 compares the FS and $FS + IFC$ methods with their "Unclassified" and "Extended" versions, and we can see that there is no major difference between the original methods employed, compared to their respective functionally enriched versions. This confirms that the functions of neighboring proteins directly influence disease-relatedness; the influence cannot be explained by the relationship between the functions of neighboring

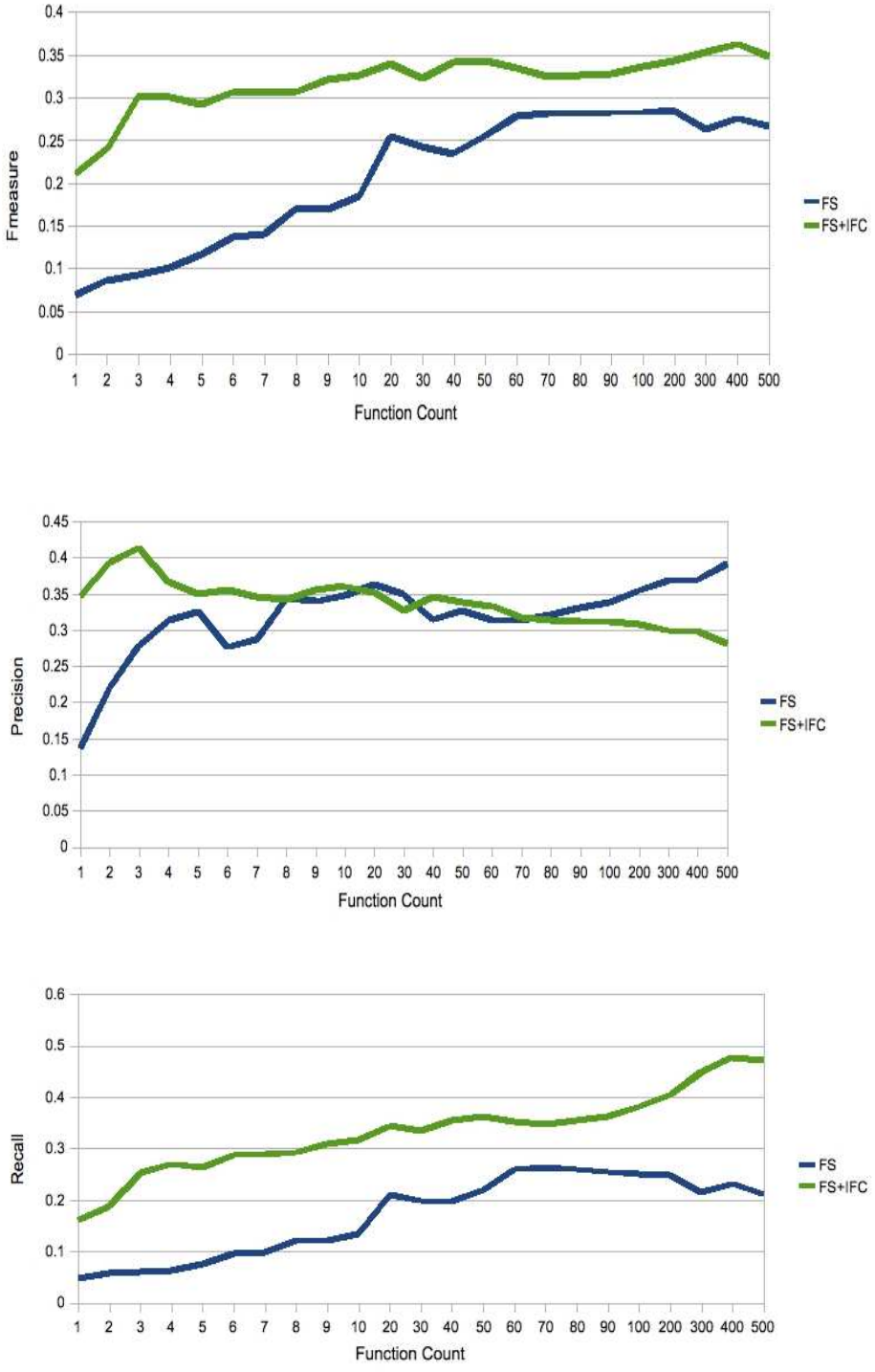


Figure 4.1: Comparing different Functional Context methods. The $FS+IFC$ method always outperforms the FS method with respect to F-measure.

proteins on the one hand, and between a protein's own functions and involvement in disease on the other hand.

Figure 4.3 compares three structural context methods disToCancer, disToAnova and disToCancerAnova in different protein counts: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500 with respect to F-measure, Precision and Recall. (As disToCancer has no natural criterion for selecting a subset of cancer-related proteins, proteins were selected randomly in this case, to arrive at comparable counts.) It turns out that, in order to get reasonable F-measure results, selecting less than 30 proteins is enough in the structural context methods. With respect to F-measure, methods using ANOVA for selecting the important proteins almost always outperform the method that selects previously known cancer-related proteins.

In Figure 4.4, we show the result of integrating the functional context method $FS+IFC$ with any one of the three structural context methods, disToAnova, disToCancer and disToCancerAnova. We vary the number of analyzed functions from 5 to 30, and the number of analyzed proteins from 1 to 40. The integration of $FS + IFC$ with disToAnova slightly outperforms the other two integrated methods. Although it may seem that applying the ANOVA method results in only small numerical improvements, Figure 4.4 shows that its integration with the functional annotation of the proteins consistently results in improved results with respect to F-measure values. Compared to functional and structural context methods, the integrated method gives rise to more cases (17 out of 52 in $(FS + IFC)$-DisToCancerAnova, as opposed to 0 out of 52 in $FS + IFC$) with F-measure over 35% (and up to 39% in one case).

4.4.5 Comparing with Previous Methods

Milenkovic et al. [77] have evaluated their method using a leave-one-out cross validation and report an F-measure of 24%. They compare this result to that of Aragues et al. [3], who use information from heterogeneous data sources: (i) Protein Protein Interaction networks, (ii) differential expression data, (iii) structural and functional properties of cancer genes; Aragues et al. report an F-measure of 18.15% for their method. Further, we will compare our results to the method of Furney et al. [29]. As Furney et al. reported results on another dataset, to obtain more comparable results we have implemented their method by selecting 100 functions based on the chi-square value, describing each protein based on those selected functions, and using the Weka machine learning system to apply Naive-Bayes for predicting the proteins involved in cancer.

Our method uses as parameters the number of functions and proteins to be selected by the feature selection method. To optimize these parameters, we divided the human dataset into three parts: 80% for training the model with a particular parameter setting; 10% for tuning the different parameters (that is, models trained with particular parameter values are tested on this 10% and the parameter settings that perform optimally here will be used for the final evaluation), and 10% for evaluating the model; note that this last 10% was not involved in the training in any way. Table 4.8 shows the optimal parameter settings for each method.

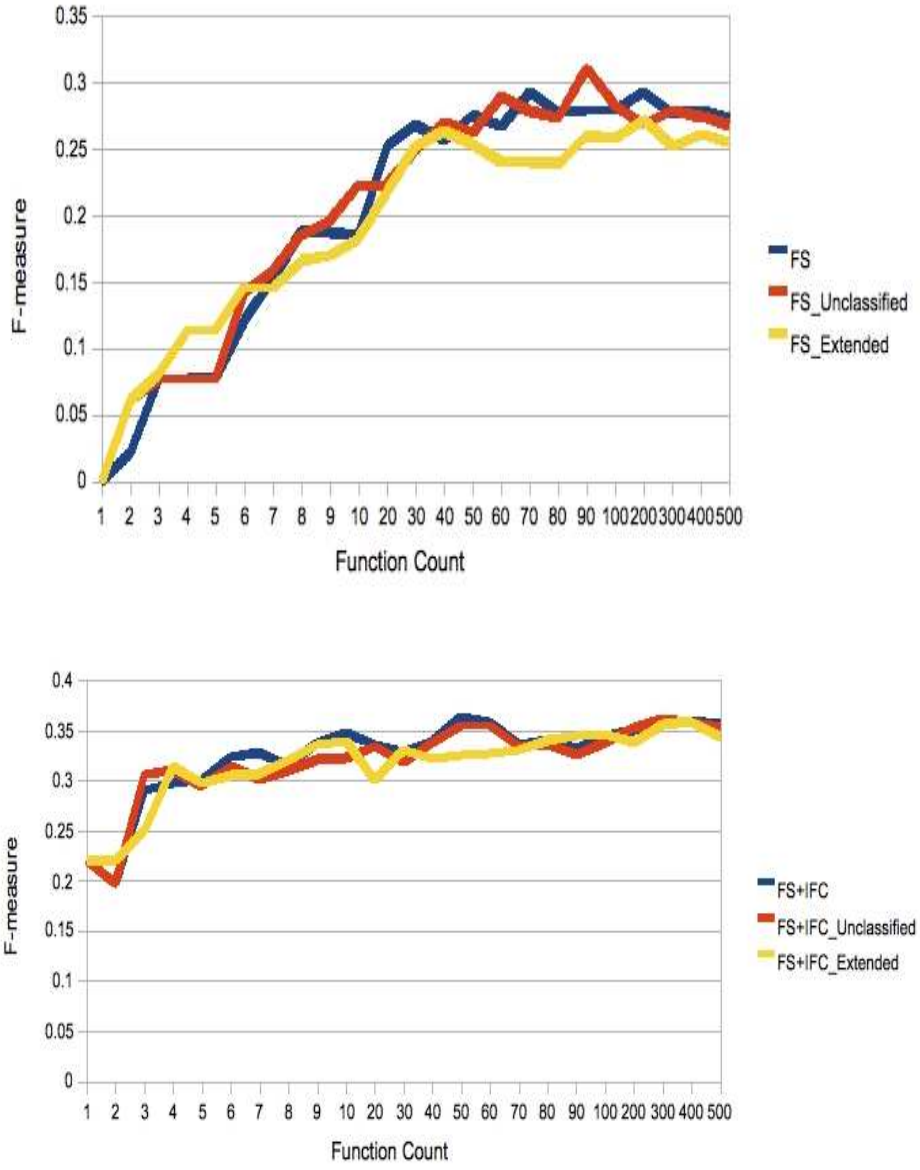


Figure 4.2: Comparing different Functional Context methods with their enriched functional versions. There is no major difference between the original methods employed, compared to their respective functionally enriched versions.

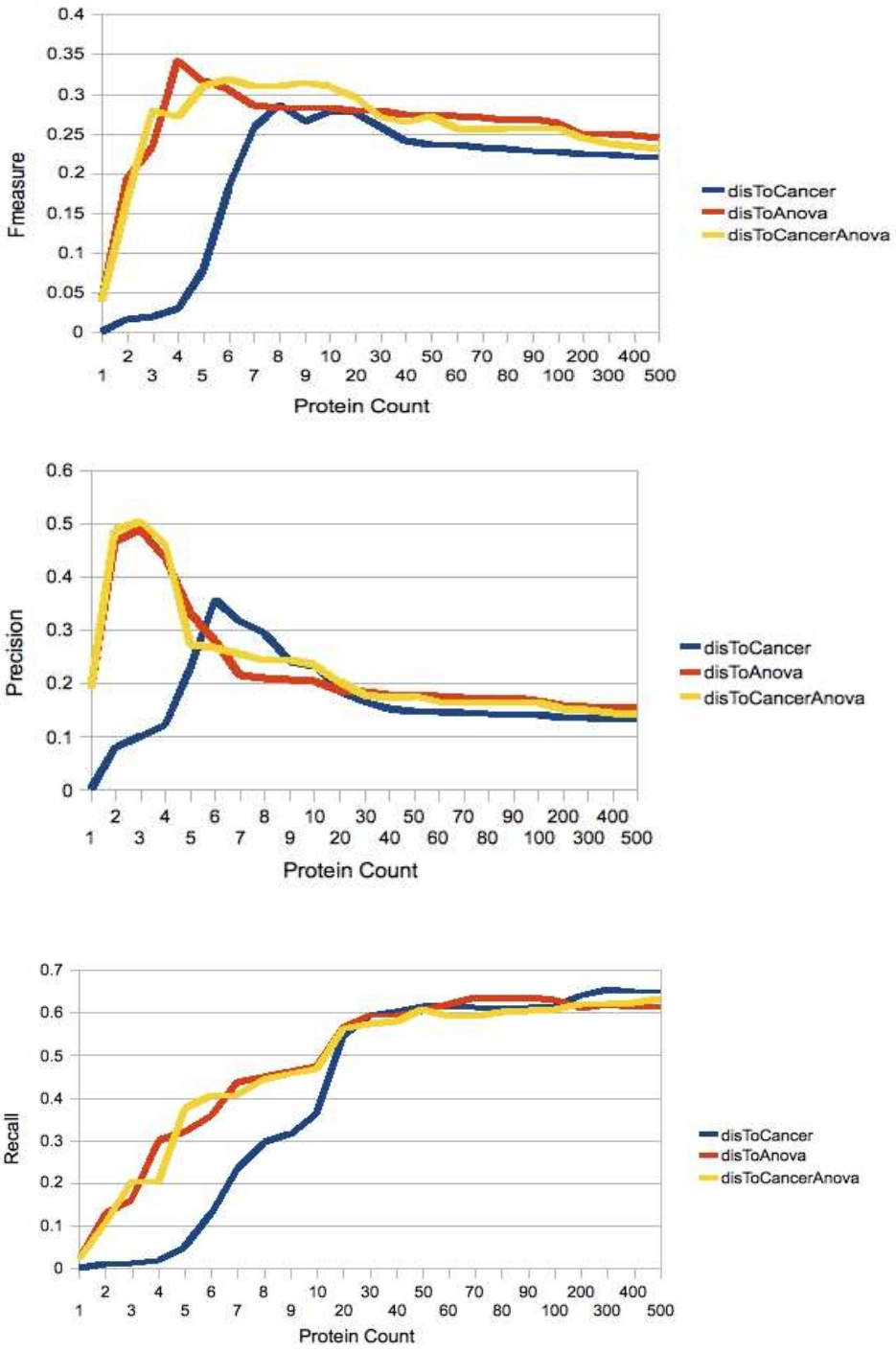


Figure 4.3: Comparing different Structural Context methods. With respect to F-measure, methods using ANOVA for selecting the important proteins almost always outperform the method which selects the previously known cancer-related proteins.

Figure 4.5 compares all the proposed methods with each other using 10-fold cross validation. The method (FS+IFC)-DisToCancerAnova which considers network contextual information outperforms all other proposed methods.

Table 4.8: **Tuning result of each method**

Method Name	Best Feature Count	F-measure in Test Set
FS	90 Functions	28
$FS + IFC$	100 Functions	34
disToAnova	10 Proteins	28
disToCancer	10 Proteins	29
disToCancerAnova	10 Proteins	30
$(FS + IFC)$-DisToCancer	10 Functions and 4 Proteins	35
$(FS + IFC)$-DisToAnova	10 Functions and 5 Proteins	37
$(FS + IFC)$-DisToCancerAnova	10 Functions and 9 Proteins	37

Figure 4.6 compares our best proposed method with previous methods using 10-fold cross validation. The method (FS+IFC)-DisToCancerAnova which considers network contextual information outperforms the previous methods (Furney et al. [29], Aragues et al. [3] and Milenkovic et al. [77]), by 5%, 13% and 8%, respectively, with respect to F-measure.

4.4.6 Random Feature Selection

We showed that the ANOVA method for selection of proteins and the chi-square based feature selection work well. A conclusion might be that the feature-selection methods work, and that it is indeed the case that some functions, or some proteins, have a higher predictive power than others. To test whether this is really the case, we compare these feature selection methods with random selection of proteins or functions. In order to evaluate the *Random* versions of the proposed methods, we did the following:

1. We chose K features randomly.
2. We used the selected method M to describe each protein in the test set based on the randomly selected features. We called the new method: M-Random.
3. We applied the naive Bayes classifier to calculate the F-measure values.
4. We repeated the steps 1 to 3 fifty times, and report the average of the F-measure values.

We assign $K = 100$ and $K = 10$ for the functional and the structural context methods, respectively. Figure 4.7 compares our proposed methods with their corresponding

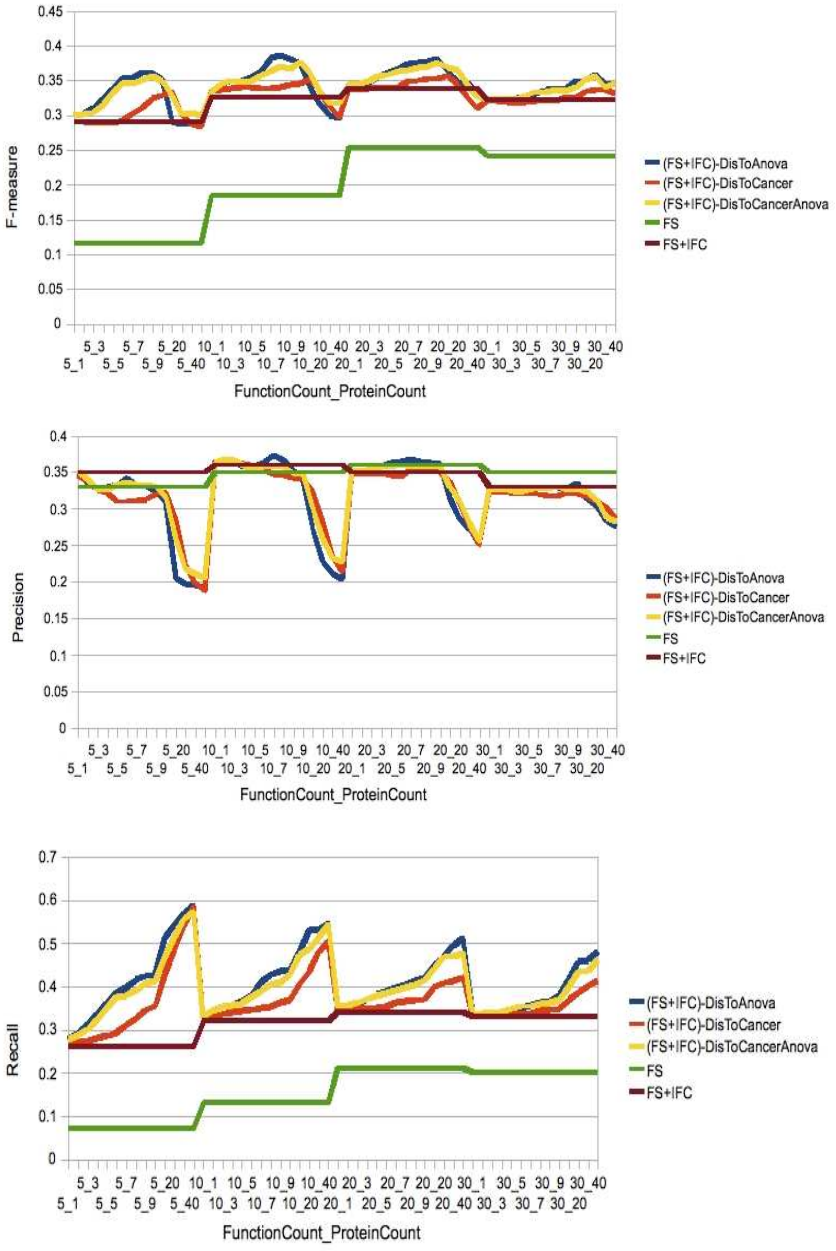


Figure 4.4: Comparing different Integrated methods. Comparing to functional and structural context methods, the integrated method gives rise to more cases (17 out of 52 in $(FS + IFC)$-DisToCancerAnova, as opposed to 0 out of 52 in $FS + IFC$) with F-measure over 35% (and up to 39% in one case).

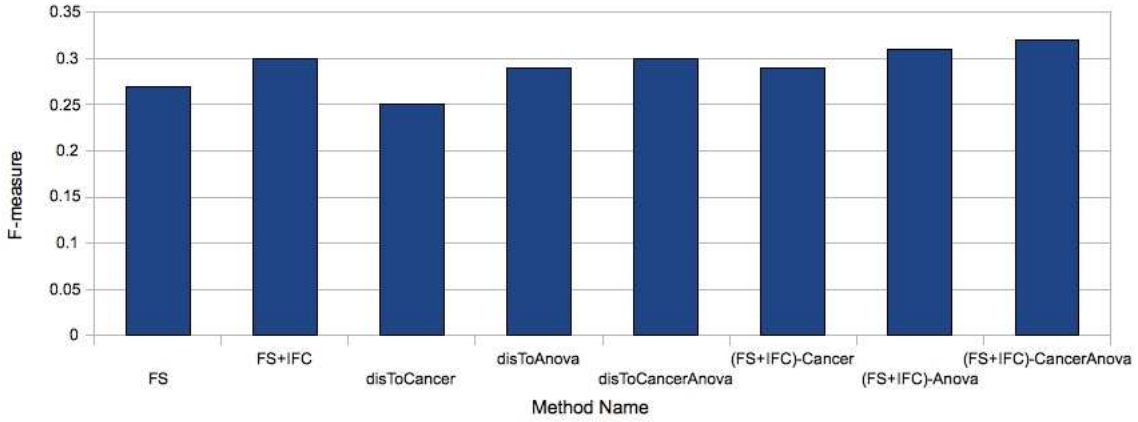


Figure 4.5: Comparison of all the proposed methods with each other. The method (FS+IFC)-DisToCancerAnova which considers network contextual information outperforms all other proposed methods.

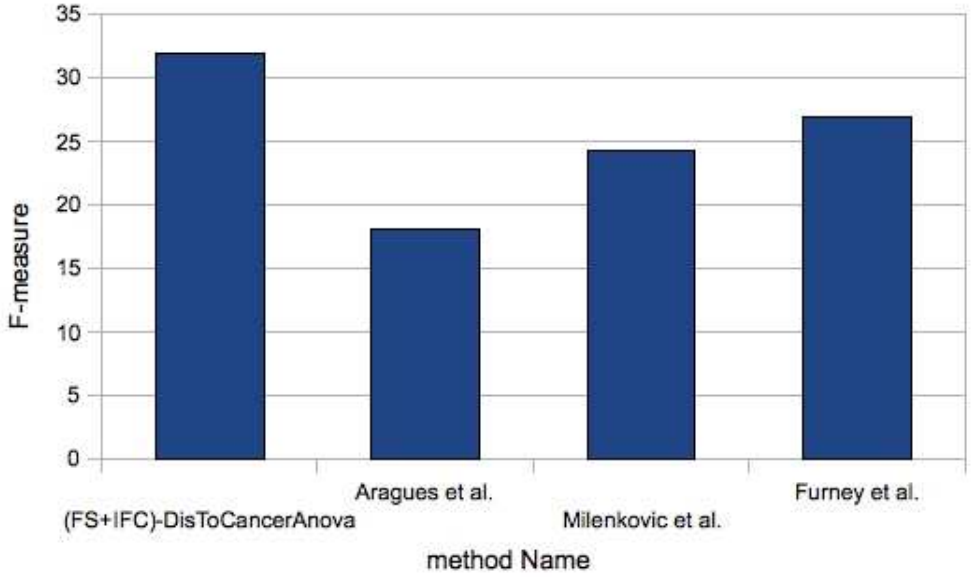


Figure 4.6: Comparing with previous methods. The method (FS+IFC)-DisToCancerAnova which considers network contextual information outperforms the previous methods (Furney et al. [29], Aragues et al. [3] and Milenkovic et al. [77]), by 5%, 13% and 8%, respectively, with respect to F-measure.

Random versions. It turns out that the feature selection algorithms outperform random selection in all cases, with F-measure improvements from 5% (for disToCancer, which also selects randomly but only among proteins known to be cancer-related) up to 26% (for FS).

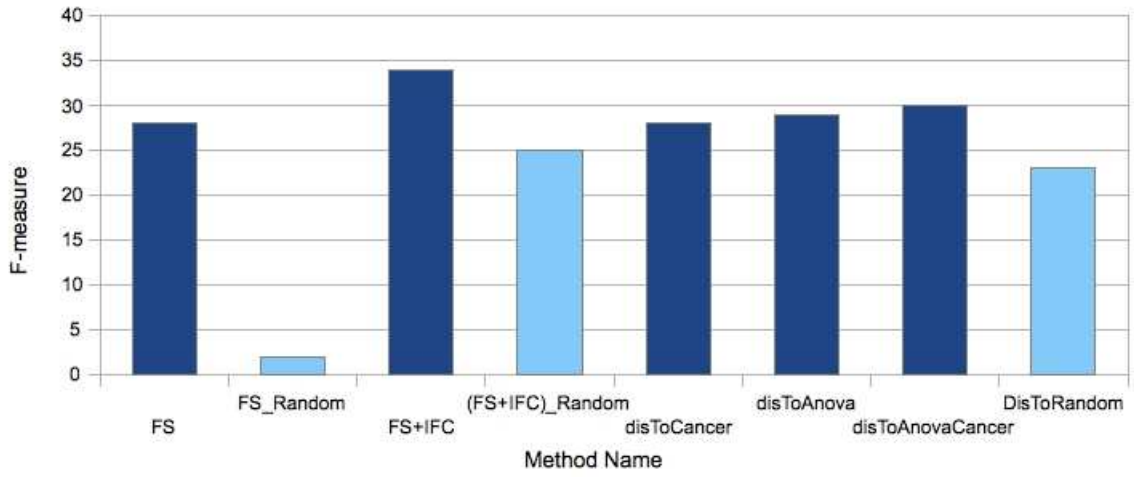


Figure 4.7: Comparing different proposed methods with their corresponding *Random* versions. The feature selection algorithm does not matter very much for the disToCancer method, with 5% improvement in F-measure over the Random version, but it matters a lot for the FS method, with 26% improvement in F-measure comparing to the Random version.

4.4.7 Capacity Identification of New Cancer-Related Proteins

The following steps were performed for predicting new cancer-related proteins:

1. A new training set was built containing all the proteins annotated as being involved in cancer (positive set) in addition to 500 randomly selected proteins (negative set).
2. A test set was built containing all the remaining proteins in the network.
3. 100 functional features were selected based on the $FS + IFC$ method.
4. 10 structural features were selected based on the ANOVA method.
5. Train set and test set were described based on the selected features.
6. The naive Bayes classifier was applied for ranking the proteins in the test set.

7. CiteXplore[68] was used to search for the high-ranked candidate proteins in the literature.

Table 4.9 lists the highly ranked newly identified cancer-related proteins. Given that, since the compilation of our dataset, novel literature linking genes to diseases (such as, in this case, cancer) have been identified, we attempted to find literature evidence for our novel gene-cancer links. As evident from Table 4.9, we found such evidence in a surprising number of at least 18 of the 20 highest-ranking genes that were not annotated in this way in the training dataset.

As can be seen, the majority of genes is now associated with breast cancer (14 out of 20), which is likely due to the fact that genotyping is currently routinely performed in this cancer type due to the different personalized treatment options available. The three genes with the least current literature information linking them to cancer are CORO2A (coronin, actin binding protein, 2A), DAZ1 (deleted in azoospermia 1) and CRSP7 (cofactor required for Sp1 transcriptional activation, subunit 7, 70kDa; now MED26, mediator complex subunit 26). However, CORO2A is involved in cell cycle progression which makes its link to cancer at least plausible. DAZ1 is involved in spermatogenesis, and it is hypothesized to bind to the 3'UTR of mRNAs to regulate their translation. While involvement of this gene in adult cancers is probably not the case, a link to regulation and cell cycle progression is also given here. Likewise, CRSP7/MED26 is a cofactor required for transcriptional activation of RNA polymerase-II dependent genes - hence, while unspecific, the link of the highest ranked genes with respect to their involvement in cancer gives a consistent link to transcriptional and, more general, cell cycle regulation events.

Overall, we were able to find literature evidence for most genes predicted to be involved in cancer, but not annotated in this manner in our training dataset. This underlines the quickly-evolving knowledge in the molecular biology field, but it also gives us more confidence that we are prospectively able to identify cancer-related genes with the approach described in the current work.

4.5 Discussion

Previous work on predicting disease-related proteins based on PPI networks has mostly focused on the functional information about the protein for which a prediction is made, or proximity of known disease-related genes in the PPI network. Several methods have been described that take into account more general features related to the graph structure, with good results. In this article, we introduce two new types of features, reflecting additional information: (1) the functions of proteins interacting with the target protein; (2) the relative position of the target protein with respect to specific other proteins, as measured by shortest-path distance. Our results indicate that:

1. Functions of proteins interacting with the target protein are informative: they offer additional information over the functions of the target protein itself. This

Table 4.9: **Capacity Identification of New Cancer-Related Proteins**

Index	Protein	Cancer Types Identified in CiteXplore	References
1	ITGAV	Breast Cancer	[11, 42]
2	CTNND2	Cervical, Prostate, Urinary Bladder	[70, 125, 46]
3	CORO2A	—	—
4	SMAD1	Breast, Colon, Lung, Prostate, Rectal, Renal cell	[62, 86, 67]
5	RPS6KB1	Breast, Colon or Rectal ovarian	[114, 90, 64]
6	VIL2	Breast and Prostate	[93, 110]
7	FST	Breast, Gastric, Lung, Prostate, Stomach, Thyroid	[85, 10, 108]
8	HSP90AA1	Gastric, Lung	[14, 115]
9	PPP2CA	Breast, Colon, Lung, Prostate	[7, 4, 128]
10	SUMO1	Breast, Lung, Prostate	[41, 80, 55]
11	SKP1A	Esophageal	[84]
12	EIF4EBP1	Breast, Colon, Head, Neck, Ovarian, Prostate	[135, 107, 5]
13	DAZ1	—	—
14	CRSP7	—	—
15	TGFB3	Breast,Colon, Prostate, Pancreatic	[112, 34, 63]
16	FHL2	Breast, Colon, Gastrointestinal, Liver, Prostate	[40, 82, 131]
17	TLN1	Breast, Prostate	[61, 94, 105]
18	GFI1B	Breast, Gastric, leukemia, Ovarian	[53, 134, 74]
19	IGFBP7	Breast, Cervical, Colorectal, leukemia, Liver, Lung, Neck, Thyroid carcinogenesis	[17, 44, 33]
20	COL4A2	Breast, Gastric, Lung, Pancreatic	[113, 8, 43]

is visible both in the expert interpretation of the results and in the predictive accuracy of the method.

2. A relatively small number of GO functions suffices to obtain maximal predictive accuracy.

3. Shortest-path distances to selected fixed proteins in the network are relevant, even more relevant than shortest-path distances to other disease-related proteins;

4. A small number of such fixed proteins (10, in our experiments) is sufficient to obtain good predictive power;

5. The χ^2 and ANOVA measures for selecting relevant functions, respectively proteins, yield interpretable results.

6. A simple and efficient machine learning method (here Naive Bayes) that uses a combination of functional information about the neighbors and shortest-path

distance to specific proteins, predicts cancer-related proteins with higher accuracy than any previous PPI-based methods.

7. What is particularly remarkable of the current work is that not only our classification results improve upon previous methods, but that also our 'false' positive predictions could in many cases be verified to be linked to cancer in more recent literature. Namely, we analyzed a list of 20 newly found cancer-related proteins that were identified by our method, and we find that virtually all of them (at least 18 out of 20) could be linked to cancer in scientific publications.

We concluded from this that the proposed features are informative for predicting cancer-related proteins as they increase the accuracy of predictive models and have a biological interpretation.

4.6 Acknowledgments

This research is funded by the Dutch Science Foundation (NWO) through VIDI grant 639.022.605. The authors thank Tijana Milenkovic for her cooperation.

Chapter 5

Predicting Cancer-Related Proteins using Interaction-based Features

Based on

Hossein Rahmani, Hendrik Blockeel and Andreas Bender: Interaction-based feature selection for predicting cancer-related proteins in protein-protein interaction networks. In: Proceedings Fifth International Workshop on Machine Learning in System Biology (2011)

5.1 Introduction

The task of predicting in a protein-protein-interaction (PPI) network which proteins are involved in certain diseases, such as cancer, has received a significant amount of attention in the literature [29, 66]. Multiple approaches haven been proposed, some based on graph algorithms, some on standard machine learning approaches. Machine learning approaches such as Milenkovic et al.[77], Furney et al. [29], Li et al. [66], Furney et al. [30] and Kar et al. [52] typically use a feature-based representation of proteins as input, and their success depends strongly on the relevance of the selected features. In earlier work it has been shown that the Gene Ontology (GO) annotations of a protein have high relevance. For instance, Li et al. [66] found predictive performance to depend only slightly on the chosen machine learning method, but strongly on the chosen features, and among many features considered, GO annotations turned out to be particularly important.

In previous work, when a protein p is to be classified as disease-related or not, the GO annotations used for that prediction are usually those of p itself. In this paper, we present a new type of GO-based features. These features are based not on the GO annotation ("function") of a single protein, but on pairs of functions that occur on both sides of an edge in the PPI network. We call them *interaction-based features*.

5.2 Interaction-based feature selection

A PPI network is a graph where nodes are proteins and an edge between two nodes indicates that those two proteins are known to interact. In our application, proteins in the training set are also labeled as cancer-related or not (supervised learning). Additionally, each protein p is annotated with a vector $FS(p)$ that indicates the functions that p has according to the Gene Ontology. Let $F = \{f_1, \ldots, f_{|F|}\}$ be the set of all functions in GO. $FS(p)$ is then an $|F|$-dimensional vector with $FS_i(p) = 1$ if protein p has function f_i, and $FS_i(p) = 0$ otherwise.

Several authors [29, 66] propose to use a χ^2-based feature selection method to select the most relevant GO terms. Let C and $\bar{C}$ be the set of proteins that are cancer-related (C) or not ($\bar{C}$), and let, for each f_i, P_i be the set of proteins annotated with f_i and $\bar{P}_i$ the set of proteins not annotated with it. With $a = |C \cap P_i|$, $b = |C \cap \bar{P}_i|$, $c = |\bar{C} \cap P_i|$ and $d = |\bar{C} \cap \bar{P}_i|$, we have

$$\chi^2(f_i) = \frac{(ad - bc)^2 * (a + b + c + d)}{(a + b)(c + d)(b + d)(a + c)} \tag{5.1}$$

Selecting individual discriminative functions based on equation 5.1 does not consider the network topology and the way different functions interact with each other in the network. Recent approach by Rahmani et al. [99] showed that considering Collaborative Functions: Pairs of functions that frequently interface with each other in different interacting proteins, improves the prediction of proteins functions. For the task of predicting cancer-related proteins, it is not impossible that a function f_i

does not correlate itself with cancer-involvement, but when a protein with function f_i interacts with a protein with function f_j, this interaction may be an indication of the former protein being involved in a cancer.

To be able to take into account the information in the interactions, we here define new features f_{ij}. These do not describe nodes, but directed edges between nodes. Although edges in a PPI network are undirected, we can see them as pairs of directed edges. A directed edge $p \rightarrow q$ is considered positive if p is a cancer-related protein, and negative otherwise. By definition, $f_{ij}(p \rightarrow q) = 1$ if $FS_i(p) = 1$ and $FS_j(q) = 1$, and 0 otherwise. If C is the set of positive edges, $\bar{C}$ the set of negative edges, and for each feature f_{ij}, P_{ij} is the set of edges for which $f_{ij} = 1$ and $\bar{P}_{ij}$ is the set of edges for which $f_{ij} = 0$, then the χ^2 value of f_{ij} can be defined exactly as above (substituting f_{ij} and P_{ij} for f_i and P_i in the formulas for a, b, c, d and χ^2). Intuitively, an f_{ij} with high χ^2-value is relevant for the class of the protein on the i-side.

The f_{ij} features describe edges, but we need instead features that describe proteins. Therefore, we define features F_{ij} as follows: $F_{ij}(p) = \sum_q f_{ij}(p \rightarrow q)$ if $FS_i(p) = 1$, and $F_{ij}(p) = -1$ otherwise. Note that by introducing -1 as a separate value indicating that $FS_i(p) = 0$, each F_{ij} encodes implicitly the corresponding f_i feature.

In this work we compare how well cancer-involvement can be predicted from: (1) a limited number of f_i features, when those features are selected according to their χ^2 value as defined above, and (2) the same number of F_{ij} features, when those features are selected according to the following score, which combines the overall relevance of f_i, f_j, and f_{ij}:

$$score(F_{ij}) = \chi^2(f_i) + \chi^2(f_j) + \chi^2(f_{ij}).$$

In the following we will call the f_i individual-based features, and the F_{ij} interaction-based features.

5.3 Results

We evaluate our methods on the dataset used by Milenkovic et al. [77]. This dataset is the union of three human PPI datasets: HPRD [91], BIOGRID [116] and the dataset used by Radivojac et al. [97]. Milenkovic et al. provide details on the construction of the integrated network; some statistical information is shown in Table 5.1.

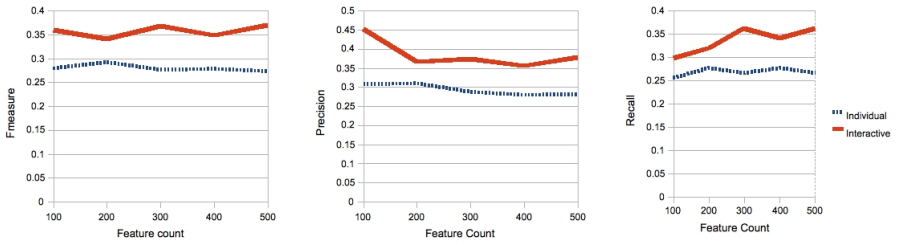
We divided the dataset into a training set containing 90%, and a test set containing the remaining 10%, of the proteins. We used information in the train set to select the $K(= 100, 200, 300, 400, 500)$ highest scoring individual-based, respectively interaction-based, features. Then, we described each protein in the test set based on the selected features and finally, we applied the Naive Bayes classifier for predicting cancer-related proteins.

Figure 5.1 compares our interaction-based features with the individual-based features with respect to the Fmeasure, Precision and Recall metrics. Our proposed method outperforms the individual-based method with 7.8%, on average, with respect to Fmeasure. This confirms our assumption about the usefulness of considering

Number of proteins	10,282
Average Degree	9.201
Min Degree	1
Max Degree	272
Number of Cancer Genes	939

Table 5.1: Statistical information of union of three human PPI datasets: HPRD [91], BIOGRID [116] and Radivojac et al. [97].

network interactions in feature selection. Table 5.2 lists five high-ranked function pairs; it shows that the functions in these pairs are not necessarily among the highest ranking functions with respect to their own χ^2.



(a) Fmeasure (b) Precision (c) Recall

Figure 5.1: Comparing interaction-based feature selection with protein-based feature selection with respect to the Fmeasure, Precision and Recall metrics. Interaction-based feature selection outperforms the protein-based method with 7.8%, on average, with respect to Fmeasure.

What is interesting about Table 5.2 is that terms from two of the ontologies used, namely 'Molecular Function' as well as 'Biological Process', are selected using our feature selection method. This is the case both for pairs of terms from the same ontology, as well as for pairs of terms taken from both ontologies. More explicitly, GO terms 5515 and 3700 relate to 'protein amino acid binding' and 'DNA binding transcription factor activity', and are hence related to cellular replication (first entry in Table 5.2). Subsequent entries have slightly different character though, such as relating protein binding (GO term 5515) to events such as signal transduction (GO term 7165), and they are hence alerting to the particular kinds of proteins that are often involved in cancer, namely kinases (such as EGFR) involved in a large number of signaling processes in the cell. It is interesting that GO terms 60571, and also 1823 and 1656 are returned by our analysis, the former relating to 'morphogenesis of an epithelial fold', and the latter two to different stages of kidney development. Hence, some of the terms returned can also be seen as tissue-specific as well as organ-specific, and in this way a more subtle differentiation of ontology annotations can be achieved

f_i	f_j	$Rank(\chi^2(f_i))$	$Rank(\chi^2(f_j))$	$Rank(score(f_i, f_j))$
GO-0005515	GO-0003700	5	6	1
GO-0005515	GO-0007165	5	46	2
GO-0060571	GO-0001656	175	17	3
GO-0060571	GO-0001823	175	105	4
GO-0060571	GO-0050768	175	170	5

Table 5.2: Five high-score interactive function pairs. Function members of interactive pairs are not necessarily among the functions with high chi-score value.

than by using single terms alone.

5.4 Conclusions

Earlier work showed that Gene Ontology annotations of a protein are relevant for predicting whether it is involved in cancer. In this work we have shown that predictive accuracy can be improved significantly by combining this information with the information contained in the topology of a PPI network. Although the combination of GO-based features and features based on network topology has been considered before, the idea of attributing GO-based features to edges, rather than nodes, is novel, and is shown here to substantially improve predictive accuracy, and to identify functional interactions for which the involved functions would not normally be found relevant by themselves.

Acknowledgements

This research was funded by the Netherlands Organisation for Scientific Research (NWO) through a Vidi grant.

Chapter 6

Predicting Disease-Related Proteins Using Human Disease Network

Based on

Hossein Rahmani, Hendrik Blockeel and Andreas Bender, "Predicting Disease-Related Proteins using Informative Human Disease Network" submitted to IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB).

6.1 abstract

Identification of novel proteins likely involved in diseases is an important issue in the area of computational biology. Protein-Protein Interaction (PPI) networks have been widely used for the task of predicting proteins involved in diseases. Previous methods assume to have a set of proteins which are previously known to be involved in disease (i.e., seed proteins) and then, they try to extend the seed proteins by predicting new disease-related proteins. While the initial seed proteins of each disease is incomplete and suffers from 'False Negative' cases, dependency of previous methods to the incomplete seed proteins is the main drawback of these methods. In this paper, we reduce the number of False Negative cases in the initial seed proteins of 20 analyzed diseases by proposing an informative Human Disease Network (HDN) considering both functional and structural information in the PPI network. After building a biologically meaningful HDN, we cluster the HDN nodes based on the connectivity and then, we augment the seed proteins of each disease based on the cluster it belongs to. Finally, we predict new disease-related proteins based on augmented seed proteins. Literature mining of newly predicted proteins proved the usefulness of the proposed HDN.

6.2 Introduction

In recent years, much effort has been invested in the construction of protein-protein interaction (PPI) networks [118]. Much can be learned from the analysis of such networks with respect to the metabolic and signalling processes present in an organism, and the knowledge gained can also be prospectively employed e.g., for the task of protein function prediction [78, 98, 18], identification of functional modules [71], interaction prediction [48, 129], identification of disease candidate genes [27, 109, 26, 58, 106, 37, 87, 130, 132] and drug targets [104, 81], according to an analysis of the resulting network [72].

Wu et al. [130] present an excellent overview of multiple methods for detecting proteins involved in disease or cancer. Among the different methods discussed in [130], "guilt-by-proximity" methods are well known. Methods classified in this category are based on the assumption that genes that directly interact, or, more generally, lie close to each other in the network, are more likely to be involved in the same diseases (as argued by, e.g., Gandhi et al. [31]). The methods vary based on how they define proximity: Some methods consider only direct neighbors to be in the proximity (e.g., [87, 3]), some quantify proximity of two proteins using the length of the shortest-path between them, some compute a "Global Distance Measure" that also takes into account how many paths there are between the two proteins, and how long these are; an example is the approach by Chen et al. [16], who use a PageRank based model for this.

The methods discussed by Wu et al. [130] mostly rely on notions of proximity (to genes known to be disease-related) from the area of graph analysis. An entirely

different type of approaches are those that rely on feature-based descriptions [132, 77, 29, 66]. There, each individual protein is described by means of a fixed set of features. Next, using machine learning methods, a model is learned that links some of these features to disease-relatedness.

In almost all the discussed methods, prediction accuracy depends directly on the initial disease-related proteins, which we refer to as seed proteins. While the initial seed proteins of each disease suffers from several 'False Negative' cases (i.e., disease-related proteins which are not annotated as being involved in disease), dependency of previous methods to the incomplete seed proteins is the main drawback of these methods. In this paper, first, we propose an informative Human Disease Network (HDN) considering both functional and structural information in the PPI network. Second, we cluster the HDN nodes based on connectivity. Third, we augment the seed proteins of each disease based on the cluster it belongs to. Fourth, we predict new disease-related proteins based on augmented seed proteins. Finally, we analyze the literature to prove the usefulness of the proposed HDN.

6.3 Methods

6.3.1 Formal Definition

We consider a PPI network as an undirected annotated graph $(P, E, \lambda_F, \lambda_D)$ where P is a set of proteins, $E \subseteq P \times P$ is a set of interactions between these proteins, and λ_F and λ_D are so-called annotation functions; for each p, λ_F and λ_D denote the additional information we have about p. In this work, we assume that $\lambda_F(p)$ simply lists all the GO functions that are associated with p; we call it the function set (or function vector) of p, and denote it $FS(p)$. $\lambda_D(p)$ lists all the diseases that protein p is involved in; we call it the disease list of p and denote it $dizList(p)$. If $D = \{diz_1, diz_2, \ldots, diz_m\}$ is the list of m analyzed diseases in our paper, then $dizList_i(p) = 1$ if p is involved in diz_i and 0 otherwise. We also define seed proteins $SP(diz_i)$ as the set of proteins involved in disease diz_i ($diz_i \in dizList(p) \Leftrightarrow p \in SP(diz_i)$).

6.3.2 Human Disease Network

We consider a Human Disease Network (HDN) as a directed graph $HDN(D, R)$ where D is a set of diseases and $R \subseteq D \times D$ is a set of directed relationships between these diseases. We build our proposed HDN as follows: For each disease $d_i \in D$:

1. We build $testSet$ as a union of the seed proteins of each disease $d_k \in D$ where $k \neq i$.
   $testSet = \bigcup_{d_k \in D \text{ and } k \neq i} SP(d_k)$

2. We consider the remaining proteins in P as $trainSet$. We assume the seed proteins $SP(d_i)$ as positive cases and the remaining proteins in the $trainSet$ as negative cases.

3. We choose a prediction method M, we train M with $trainSet$ and then, we use method M to calculate the prediction-value $PV(p)$ for each protein $p \in testSet$. M will return high PV values for more relevant disease-related proteins.
4. We repeat step 3 , 10 times and we calculate the average prediction-value of each protein $p \in testSet$ (APV(p)).
5. For each disease $d_j \in D(j \neq i)$, we add a directed edge $d_i \rightarrow d_j$ in HDN based on Formula 6.1.

$$weight(d_i \rightarrow d_j) = \frac{\sum_{\forall p \in SP(d_j)} APV(p)}{|SP(d_j)|} \tag{6.1}$$

In Formula 6.1, the $||$ operator returns the number of seed proteins of disease d_j.

The resulting HDN is the directed fully-connected network in which each node is a disease and each weighted edge shows a relationship between two diseases. In order to focus on the most important relationships in HDN, we prune the network by keeping only the highest-ranked edges.

Although our proposed approach for building HDN is very general and any prediction method M could be used in step 3 of building HDN, the quality of the resulted HDN still depends on the prediction method M. We next discuss some recommended prediction methods.

6.3.3 Recommended Prediction Methods

In this section, we discuss about three categories of methods used for predicting proteins involved in diseases. *Structural* methods predict proteins involved in diseases based on the topological location of the proteins in the PPI network while *functional* methods use the functional annotation of the proteins for the prediction. *Hybrid* methods take both structural and functional information into account.

Structural Category: Random Walk based Method (*ST-RW*)

Berger et al. [6] assume that disease-related proteins fall closer on average to the seed proteins than they do on average to the rest of the network. They calculate the score of each protein p_j in the network based on Formula 6.2 and then, select high-scoring proteins as disease-related proteins.

$$score_s(p_j) = \frac{\frac{\sum_{i \in C'} T_{ij}}{|C'|} - \frac{\sum_{i \in C} T_{ij}}{|C|}}{\frac{\sum_i T_{ij}}{|C|+|C'|}} \tag{6.2}$$

In Formula 6.2, T_{ij} is the average number of steps a random walker takes to walk from a specified node i to another specified node j, C is the set of seed proteins and C' is the set of all other proteins in the network. In the rest of this paper, we refer to this method as *ST-RW*.

Structural Category: ANOVA based Method (*ST-Anova*)

Rahmani et al. [98] proposed a relevance measure for proteins that is inspired by statistical ANOVA (analysis of variance), and showed that shortest-path distance to a relatively small number of proteins (selected according to the ANOVA-based measure) is informative for the task of function prediction in the PPI network. Since the ANOVA method works well for function prediction, it is natural to check whether it also gives good results for the task of predicting disease-related proteins. We therefore propose the use of similar features for predicting proteins involved in disease.

The ANOVA-inspired selection measure (briefly, ANOVA) is defined as follows. Let P^+ be the set of proteins labeled as being involved in disease *diz*, and P^- the set of proteins not labeled as such. For each protein q, we introduce a feature d_q; $d_q(p)$ denotes the shortest-path distance between p and q (viewed here as a feature of p). We consider for each q the mean and variance of $d_q(p)$, taken over all *diz*-related (m_q^+ and var_q^+ respectively) and non-*diz*-related p (m_q^- and var_q^- respectively).

$$m_q^+ = \frac{\sum_{p \in P^+} d_q(p)}{|P^+|} \tag{6.3}$$

$$m_q^- = \frac{\sum_{p \in P^-} d_q(p)}{|P^-|} \tag{6.4}$$

$$var_q^+ = \frac{\sum_{p \in P^+} (d_q(p) - m_q^+)^2}{|P^+| - 1} \tag{6.5}$$

$$var_q^- = \frac{\sum_{p \in P^-} (d_q(p) - m_q^-)^2}{|P^-| - 1} \tag{6.6}$$

Seeing P^+ and P^- as two groups of proteins, the following formula compares the variance between groups to the variance within groups (as it is used for relative ranking only, constant factors are dropped):

$$A_q = \frac{(m_q^+ - m_q^-)^2}{var_q^+ + var_q^-} \tag{6.7}$$

A high A_q means that d_q varies little within groups and/or much between groups, which indicates that d_q has high predictive power for the group. Features d_q can be ranked according to A_q, and the top-k features selected as actual features to be included in the description of all proteins. In the end, we apply the naive Bayes classifier to the proteins descriptions for predicting *diz*-related proteins. In the rest of this paper, we refer to this method as *ST-Anova*.

Functional Category: Individual based Method (*Func-Indiv*)

In this method, first, we use a χ^2-based feature selection method to select the most relevant individual functions. Let D and $\bar{D}$ be the set of proteins that are disease-related (D) or not ($\bar{D}$), and let, for each function f_i, P_i be the set of proteins annotated

Table 6.1: List of the 4 different hybrid methods considering structural and functional information in the network.

Structural Method	Functional Method	Hybrid Method
ST-RW	Func-Indiv	RW-Indiv
ST-RW	Func-Collab	RW-Collab
ST-Anova	Func-Indiv	Anova-Indiv
ST-Anova	Func-Collab	Anova-collab

with f_i and $\bar{P}_i$ the set of proteins not annotated with it. With $a = |D \cap P_i|$, $b = |D \cap \bar{P}_i|$, $c = |\bar{D} \cap P_i|$ and $d = |\bar{D} \cap \bar{P}_i|$, we have

$$\chi^2(f_i) = \frac{(ad - bc)^2 * (a + b + c + d)}{(a + b)(c + d)(b + d)(a + c)} \tag{6.8}$$

We calculate the chi-square of each individual function f_i in the network. Then, we describe each protein p_j in the network based on the high-scored individual functions. In the end, we apply the naive Bayes classifier for predicting disease-related proteins. In the rest of this paper, we refer to this method as *Func-Indiv*.

Functional Category: Collaboration based Method (*Func-Collab*)

Selecting individual discriminative functions based on $\chi^2(f_i)$ does not consider the network topology and the way different functions interact with each other in the network. Rahmani et al. [100] showed that for the task of predicting cancer-related proteins, it is possible that a function f_i does not correlate itself with cancer-involvement, but interaction of the same function with some function f_j does correlate with the former protein being involved in a cancer. Rahmani et al. [100] proposed a new way of calculating the χ^2 of the function pairs in the PPI network. They select high-ranked collaborative function pairs and then, they describe the proteins based on the high-ranked function pairs. In the end, they applied the naive Bayes classifier for predicting the proteins involved in cancer. In the rest of this paper, we refer to this method as *Func-Collab*.

Hybrid Category: Integrating Functional and Structural Information

Structural-based and functional-based methods can be combined into hybrid methods as shown in Table 6.1. The hybrid method is calculated as follows:

$$score_h(p) = norm(score_s(p)) + norm(score_f(p)) \tag{6.9}$$

In Formula 6.9, $score_s(p)$ and $score_f(p)$ show disease-relatedness score of p using *Structural* (*ST-RW* and *ST-Anova*) and *Functional* (*Func-Indiv* and *Func-Collab*) methods, respectively. In order to avoid a bias toward either of these categories, we use Formula 6.10 to normalize the disease-relatedness scores. In Formula 6.10,

$min(x)$ and $max(x)$ return minimum and maximum values taken over all values of x, respectively.

$$norm(x_i) = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{6.10}$$

6.4 Empirical Results

6.4.1 Dataset

We applied our method for building HDN to the PPI network used by Milenkovic et al. [77]. This dataset is the union of three human PPI datasets: HPRD [91], BIOGRID [116] and the dataset used by Radivojac et al. [97] and contains 47,303 physical interactions among 10,282 proteins. When we say "union", we mean that the new network contains all the nodes and edges (proteins and interactions) found in either of these networks. The aim of merging these three datasets was to obtain as complete a human PPI network as possible, i.e., a network that covers with its edges as many proteins in the human proteome as possible. Milenkovic et al. [77] provide details on the construction of the integrated network.

Table 6.2 shows the list of 20 different diseases analyzed in this paper in addition to the number of proteins involved in each disease (seed count).

6.4.2 Comparing Recommended Prediction Methods

In this section, we use the following leave-one-out cross validation to compare the different prediction methods discussed in section 6.3.3:

For each disease $d_i \in D$:

1. We select 99 proteins randomly from the PPI network ($randSet$).
2. For each seed proteins $p_i \in SP(d_i)$
   (a) We build the $trainSet$ by excluding the $\{p_i \cup randSet\}$.
   (b) We apply different prediction methods M to rank p_i relative to the 99 randomly selected proteins ($rank(p_i)$). M should return small rank values for more relevant disease-related proteins.
3. We repeat steps 1 to 2b, 10 times and we calculate the average rank of each $p_i \in SP(d_i)$ over different iterations ($avg(rank(p_i))$).

Figure 6.1 compares the discussed prediction methods for the 20 different diseases shown in Table 6.2 with respect to overall rank of seed proteins among 99 randomly selected proteins (Formula 6.11). *RW-Indiv* achieves the best overall performance, compared to the other methods, and is therefore a good candidate method for building HDN.

Table 6.2: List of the 20 different diseases analyzed in this paper.

Disease ID	Disease Name	Seed Count
D-1	Alzheimer	7
D-2	Amyotrophic	4
D-3	Anemia	36
D-4	Breast Cancer	21
D-5	Cataract	14
D-6	Charcot-marie-tooth	11
D-7	Colorectal-cancer	20
D-8	Deafness	28
D-9	Diabets	23
D-10	Dystonia	5
D-11	Ehlers-danlos	7
D-12	Emolytic-anemia	11
D-13	Epilepsy	11
D-14	Long QT Syndrome	13
D-15	Lymphoma	27
D-16	Mental-retardation	19
D-17	Parkinson	8
D-18	Usher-syndrome	5
D-19	Xeroderma	10
D-20	Zellweger	8

$$overallRank(d_i) = \frac{\sum_{p_i \in SP(d_i)} avg(rank(p_i))}{|SP(d_i)|} \tag{6.11}$$

For each discussed method M, Table 6.3 shows the set of diseases for which M produces the best result. It is clear that *Func-Indiv* and *RW-Indiv* are overall the best performing methods.

Figure 6.2 compares the three best methods, *ST-RW*, *Func-Indiv* and *RW-Indiv*, to each other for each disease. The figure shows that, for those diseases where *Func-Indiv* scores best (e.g., D-6, D-11 and D-19), it is only slightly better than the second-best method, whereas in those cases where it is not best, the difference with the best method can be large (e.g., D-3, D-7, D-15, D-16). *RW-Indiv*, on the other hand, never differs much with the best method, making it it the most stable method for predicting disease-related proteins in PPI networks.

Figure 6.3 compares all methods on six different diseases where neither *Func-Indiv* nor *RW-Indiv* achieves the best overall performance. Although *RW-Indiv* is not the best method for any of these, Figure 6.3 shows that on average it ranks second, with a very small difference compared to the best method.

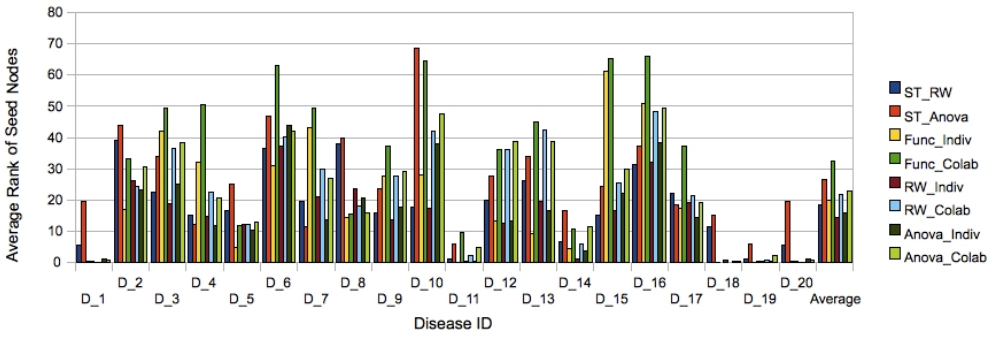


Figure 6.1: Average rank of seed proteins in 20 different diseases shown in Table 6.2. *RW-Indiv* achieves the best overall performance comparing to the other methods and is therefore a good candidate method for building HDN.

Table 6.3: Set of diseases in which each method produces the best result.

Method M	Set of diseases which M produces the best results	Count
ST-RW	D15, D16	2
ST-Anova	D7	1
Func-Indiv	D2, D5, D6, D8, D11, D13, D19	7
Func-Collab	D18	1
RW-Indiv	D1, D3, D9, D10, D12, D14, D20	7
RW-Collab	–	0
Anova-Indiv	D4, D17	2
Anova-Collab	–	0

6.4.3 Informative Human Disease Network

We choose the *RW-Indiv* prediction method to build our proposed HDN for 20 different diseases shown in Table 6.2. There are 380(20×19) possible edges in the original HDN. We prune HDN by sorting the edges based on their weight descendingly and then, keeping the 38 (10% of original HDN) highest-weighted edges. Figure 6.4 shows the pruned HDN. For each edge $(d_i) \xrightarrow{rank} (d_j)$, Figure 6.4 shows the rank of the relationship between two diseases d_i and d_j among all the 380 disease pairs. The highest-ranking found relationship is $(deafness) \xrightarrow{1} (usher\ syndrome)$. Analyzing the literature, we found biological evidence for most of the relationships shown in Figure 6.4.

Goh et al. [38] propose a simple method for building undirected Human Disease Network. They connect two diseases d_i and d_j in the network if there is at least one gene that implicated in both. We applied the Goh et al. [38] to our disease dataset and

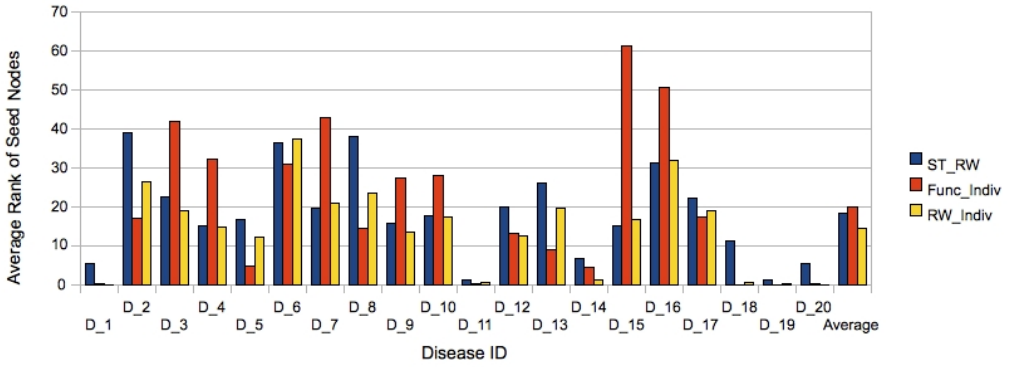


Figure 6.2: Comparing *ST-RW*, *Func-Indiv* and *RW-Indiv* methods with each other with respect to average rank of seed proteins in 20 different diseases shown in table 6.2.

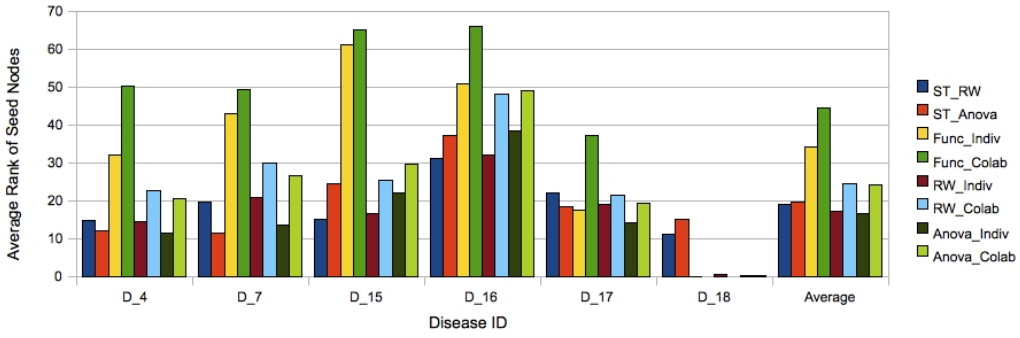


Figure 6.3: Comparing different prediction methods in 6 different diseases in which neither *Func-Indiv* nor *RW-Indiv* achieves the best overall performance. According to the average rank column , *RW-Indiv* is the second best method for these diseases with a very small difference with the best method.

the resulted HDN is shown on Figure 6.5. For each edge $(d_i) \leftrightarrow (d_j)$, Figure 6.5 shows the number of proteins involved in both diseases d_i and d_j ($|SP(d_i) \cap SP(d_j)|$). The best found relationship is $(anemia) \leftrightarrow (emolytic\ anemia)$. Comparing our proposed HDN (Figure 6.4) with the disease network discussed by Goh et al. [38] (Figure 6.5), we observe that our HDN is more informative than the network proposed by Goh et al. [38].

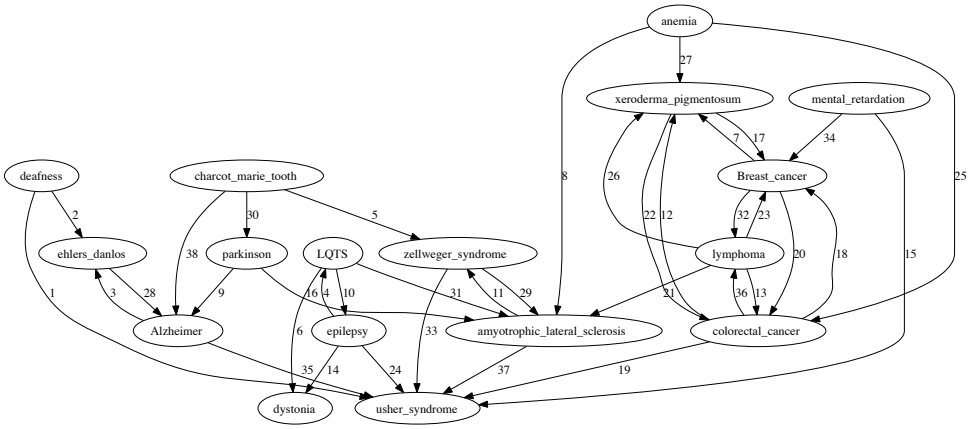


Figure 6.4: Pruned Human Disease Network by keeping only 38 (10% of original HDN) high-ranked relationships among different diseases. The best found relationship is $(deafness) \xrightarrow{1} (usher\ syndrome)$.

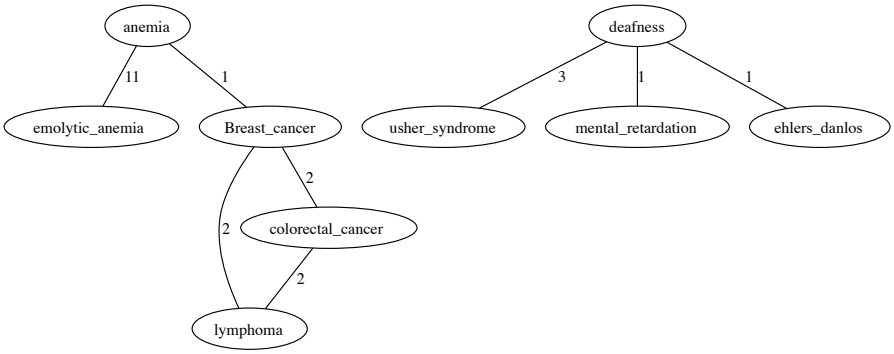


Figure 6.5: Human Disease Network based on the common proteins (Proposed by Goh et al. [38]). Edge's weight shows the number of common proteins between two related diseases.

6.4.4 Biological Interpretation of the Pruned HDN

We will now briefly discuss the biological significance of the observed findings. The highest ranked connection (1.) between deafness and Usher's Syndrome is appar-

ent, given the latter is an inherited form of deafness. The link between Deafness and Ehlers-Danlos syndrome however may be attributed also to misdiagnosis of joint laxity, given that the combination of this observation with deafness is more likely to be correctly classified as Stickler syndrome [65]. Epilepsy and Dystonia are both characterized by seizures, and given the proximity of both terms in the Figure also a mechanistic connection between both disorders can be elucidated. Interesting is the relationship of Long QT Syndrome (LQTS) to both Dystonia and Epilepsy, which hints at the importance of ion channels being important in all of those cases. On the other hand, Anyotrophic Lateral Sclerosis (ALS) is more related to Parkinson's disease (but neither epilepsy nor dystonia), hinting at fundamentally different mechanisms behind those, on the surface similar, disorders characterized by seizures. Apart from this seizure-cluster, also various cancer variations are found to be closely related, namely Xeroderma Pigmentosum (leading to sensitivity to UV light), breast cancer, lymphomas and colorectal cancer. What is interesting is the close link of ALS with the cluster of cancers, since indeed it is assumed that ALS, as a motor neuron disease, may represent a particular case of paraneoplastic encephalomyelitis [122].

6.4.5 Predicting Disease-Related Proteins using the Pruned HDN

In the context of involvement in disease, one main drawback of previous methods is their dependency on a list of seed proteins which is likely incomplete. In this section, we use our proposed HDN for augmenting the seed proteins of different diseases as follows: First, we cluster the pruned HDN into n clusters $C_1 \dots C_n$ based on the network connectivity. Second, we augment the seed proteins of each disease member d_i of cluster C_j by unioning the seed proteins of all the disease members of cluster C_j ($d_i \in C_j \Rightarrow Aug(SP(d_i)) = \cup_{d_k \in C_j} SP(d_k)$). $Aug(SP(d_i))$ is the augmented list of seed proteins of disease d_i. Third, we use a hybrid prediction method *RW-Indiv* for predicting new proteins involved in the disease. Table 6.4, Table 6.5, Table 6.6 and Table 6.7 show the four clusters extracted from the pruned HDN shown in Figure 6.4 in addition to the 10 highest-ranked proteins predicted for each cluster. The first cluster, covering Alzheimer and Ehler-Danlos syndrome, covers both known and potential novel protein targets to treat those diseases. In case of Alzheimer's, BACE2, HSD17B10 and TM2D1 have been implicated in literature before, while COL5A3, which encodes one of the fibrillar collagens, has been established to be involved in Ehler-Danlos syndrome. On the other hand, genes (and proteins) not explicitly associated with those diseases are TGBF2, THBS1 and SPON1, all of which are known to be involved in cell-to-cell interactions, cell-to-matrix interactions, and cell adhesion, respectively. In particular SPON1 can readily be understood to be of importance, given its involvement of attachment of neuron cells and neurite outgrowth.

Similar results covering both established and novel genes are observed for the second cluster, with LQTS, Epilepsy and Dystonia. Dopamine levels and epilepsy have been linked for a long time (DRD1, DRD3 and DRD4; [117] Dystonia) and they are of practical relevance for treatment. The KCNQ4 ion channel on the other hand

Table 6.4: 10 highest-ranked proteins predicted for cluster 1 = {Alzheimer, ehler-danlos}.

Index	Protein Symbol	Full Protein Name
1	COL5A3	Collagen, type V, alpha 3
2	THBS1	Thrombospondin 1
3	TGFB2	Transforming growth factor, beta 2
4	COL5A2	Collagen, type V, alpha 2
5	PDGFA	Platelet-derived growth factor alpha polypeptide
6	SPON1	Spondin 1, extracellular matrix protein
7	HSD17B10	Hydroxysteroid (17-beta) dehydrogenase 10
8	HADH2	Hydroxysteroid (17-beta) dehydrogenase 10
9	BACE2	Beta-site APP-cleaving enzyme 2
10	TM2D1	TM2 domain containing 1

has been previously linked with Long QT Syndrom (LQTS). What is interesting, with potential practial implications, is the importance of ALG10 in this cluster, which gates rat ether-a-go-go (the human homolog of the hERG channel involved in LQTS) and which might hence also play an important role in human. No explicit involvement of the EPM2AIP1 gene, encoding laforin, has been described in literature yet; however, our analysis makes a rather strong disease implication for the three diseases present in this cluster.

The third cluster of neoplastic diseases, covering Xeroderma pigmentosum, breast cancer, lymphoma and colorectal cancer gives relatively little surprises, with agreement on MSH3 and MSH6 which are both involved in DNA repair, on the APC tumor suppressor protein, and the RELA oncogene (which bings to the NF kappa b transcription factor with known involvement in cancerogenesis).

The fourth and final disease cluster, of Zellweger syndrome, ALS, and Usher's Syndrome, involves the myosins MYO6, MYO3A and MYO15A which are all known to be involved either in hearing loss or, in the latter case, the actin organization in the hair cells of the cochlea. What is apparent is the link of this set of disorders to the peroxisome, which has been established for this disease cluster before (the involvement of PEX7 and PEX12 which are involved in the assembly of peroxisomes is characteristic, but also ABCD1 is involved in fatty acid transport into the peroxisome, and PXMP3 is involved in its biogenesis). The potentially most surprising gene located in this disease cluster is SIRT3, which is known to be involved in epigenetic silencing and which has been characterized as a potential antineoplastic target - given its prominent role in this analysis, it might hence also play a role for drug treatments of this set of diseases in the future.

Table 6.5: 10 highest-ranked proteins predicted for cluster 2 = {LQTS, Epilepsy, Dystonia}.

Index	Protein Symbol	Full protein Name
1	DRD4	Dopamine receptor D4
2	DRD3	Dopamine receptor D3
3	DRD1	Dopamine receptor D1
4	ALG10B	Asparagine-linked glycosylation 10, alpha-1,2-glucosyltransferase homolog B (yeast)
5	KCR1	A membrane Protein That Facilitates Functional Expression of Non-inactivating K+ Currents Associates with Rat EAG Voltage-dependent K+Channels
6	EPM2AIP1	EPM2A (laforin) interacting protein 1
7	KCNQ4	Potassium voltage-gated channel, KQT-like subfamily, member 4
8	TOR1B	Torsin family 1, member B (torsin B)
9	HSPC163	–
10	GCHFR	GTP cyclohydrolase I feedback regulator

Table 6.6: 10 highest-ranked proteins predicted for cluster 3 = {xeroderma-pigmentosum, breast-cancer-leon, lymphoma, colorectal-cancer}.

Index	Protein Symbol	Protein Full Name
1	MSH6	MutS homolog 6 (E. coli)
2	MSH3	MutS homolog 3 (E. coli)
3	APC	Adenomatous polyposis coli
4	RELA	V-rel reticuloendotheliosis viral oncogene homolog A (avian)
5	TGFBR1	Ransforming growth factor, beta receptor 1
6	PTK2B	PTK2B protein tyrosine kinase 2 beta
7	HIPK2	Homeodomain interacting protein kinase 2
8	RPS6KB1	Ribosomal protein S6 kinase, 70kDa, polypeptide 1
9	TGFB1	Transforming growth factor, beta 1
10	ERBB2	V-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)

6.4.6 Case Study: Long QT Syndrome

In this section, we examine Long QT Syndrome (LQTS) in more details. According to [120], LQTS is a disorder of the heart's electrical activity which can cause sudden, uncontrollable, dangerous arrhythmias in response to exercise or stress. Table 6.8

Table 6.7: 10 highest-ranked proteins predicted for cluster 4 = {zellweger-syndrome, amyotrophic-lateral-slecrosis, usher-syndrome}.

Index	Protein Symbol	Protein Full Name
1	MYO15A	Myosin XVA
2	MYO3A	Myosin IIIA
3	MYO6	Myosin VI
4	DDO	D-aspartate oxidase
5	PEX12	Peroxisomal biogenesis factor 12
6	PEX7	Peroxisomal biogenesis factor 7
7	PXMP3	Peroxisomal membrane protein 3
8	SIRT3	Sirtuin 3
9	AGXT	Alanine-glyoxylate aminotransferase
10	ABCD1	ATP-bindende cassette, sub-familie D (ALD), lid 1

Table 6.8: Proteins associated with the Long QT Syndrome. The data is taken from Berger et. al.[6].

Index	Protein sybmbol	Full Protein name
1	KCNQ1	Potassium voltage-gated channel, KQT-like subfamily, member 1
2	KCNH2	Potassium voltage-gated channel, subfamily H (eag-related), member 2
3	SCN5A	Sodium channel, voltage-gated, type V, alpha subunit
4	ANK2	Ankyrin 2, neurona l
5	KCNE1	Potassium voltage-gated channel, Isk-related family, member 1
6	KCNE2	Potassium voltage-gated channel, Isk-related family, member 2
7	KCNJ2	Potassium inwardly-rectifying channel, subfamily J, member 2
8	CACNA1C	Calcium channel, voltage-dependent, L type, alpha 1C subunit
9	CAV3	Caveolin 3
10	SCN4B	Sodium channel, voltage-gated, type IV, beta
11	AKAP9	A kinase (PRKA) anchor protein (yotiao) 9
12	SNTA1	Syntrophin, alpha 1
13	ALG10	Asparagine-linked glycosylation 10 homolog (yeast, alpha-1,2-glucosyltransferase)

shows the set of proteins involved in LQTS.

Table 6.9: 10 most discriminative functions according to $\chi^2(f_i)$ (Formula 6.8).

Index	Function	Short Description
1	GO:0008016	Regulation of heart contraction
2	GO:0060307	Regulation of ventricular cardiomyocyte membrane repolarization
3	GO:0060299	Regulation of heart contraction
4	GO:0002095	Caveolar macromolecular signaling complex
5	GO:0014819	Regulation of skeletal muscle contraction
6	GO:0031579	Membrane raft organization
7	GO:0033292	T-tubule organization
8	GO:0005251	Delayed rectifier potassium channel activity
9	GO:0005244	Voltage-gated ion channel activity
10	GO:0008015	Blood circulation

Most Relevant Features for Long QT Syndrome

The number of different functions occurring in our human dataset is 9833; this is also the dimensionality of the *Func-Indiv* method if no dimensionality reduction is used. As we discussed in section 6.3.3, we can use a χ^2-based feature selection methods to reduce this number; at the same time, this techniques rank functions according to how relevant they are for prediction of disease relatedness.

Table 6.9 shows the ten most discriminant individual functions obtained. It can be seen that the top three GO annotations are explicitly related to cardiac action potential (regulation of heart contraction, regulation of ventricular cardiomyocyte membrane repolarization and negative regulation of sarcomere organization). Positions 4 and 5 are concerning caveolar signaling (which is also very prominent in the heart) and regulation of skeletal muscle contraction, alluding to the fact that muscle contraction in the skeleton and in the heart is goverened by related processes. Membrane rafts (as well as caveolae) are important for cardiac ion channel function as has been found before, [73] which is also correctly identified in Table 6.9. T-tubule organization, while not immediately apparent, has been linked to a 'new paradigm' for human arrhythmias recently [2]. It is interesting that explicit potassium and ion channel activity are appearing only low in this list, along with the broad term of blood circulation. Hence, overall it can be said that the most discriminative functions are overall meaningful, with specific functions appearing at the top, biologically derived functions (raft organization, T-tubule organization) in the middle, and general terms at the bottom of the terms derived from the analysis.

Our dataset contains 10,282 proteins. The Anova based method uses the ANOVA measure to select the most relevant among these. More detailed information could be obtained from an ANOVA analysis of the most relevant proteins among the full set of 10,282 proteins. Table 6.10 now shows the ten proteins with the highest ANOVA measure obtained using our analysis. Interestingly, no ion channel has been most

Table 6.10: 10 most discriminative proteins according to Anova (Formula 6.7).

Index	Protein	Short Description
1	NDUFS6	NADH dehydrogenase [ubiquinone] iron-sulfur protein 6, mitochondrial
2	KCNH1	Potassium voltage-gated channel subfamily H member 1
3	KCNH5	Potassium voltage-gated channel, subfamily H (eag-related), member 5
4	KCNF1	Potassium voltage-gated channel subfamily F member 1
5	AKAP6	A-kinase anchor protein 6
6	ALG10B	Asparagine-linked glycosylation 10, alpha-1,2-glucosyltransferase homolog B
7	KCR1	A membrane Protein That Facilitates Functional Expression of Non-inactivating K+ Currents Associates with Rat EAG Voltage-dependent K+Channels
8	KCNE1	Potassium voltage-gated channel subfamily E member 1
9	KCNH2	potassium voltage-gated channel, subfamily H (eag-related), member 2
10	KCNE2	Potassium voltage-gated channel subfamily E member 2

signficiant, but the NADH dehydrogenase NDUFS6. It has been found that HDUSF6 knockouts cause mitochondrial complex I deficieny [54], causing various cardiac problems such as reduced systolic function and cardiac output. On the one hand, this might relate to a functional relationship between diseases; on the other hand it might indicate imperfect diagnosis, hence confusing different underlying disease biology. The six Potassium channels listed can be understood to be involved in direct polarization and depolarization of the cardiac action potential; however the three remaining proteins, namely AKAP6, ALG10B and KCR1 deserve particular attention here. AKAP6 (also called mAKAP) anchors Protein Kinase A to RYR2 which is able to generate Ca2+ 'sparks' due to simultaneous activation within a certain neighborhood radius [124], and hence importance to the cardiac action potential and deviations thereof. ALG10B (also known as KCR1) is interestingly thought to be able to reduce KCNH2 sensitivity to proarrhythmic drug blockade which may be due to glycosylation of this potasssium channel [60, 92], hence our method was able to not only identify protein directly involved in causing LQTS, but also modifer proteins such as AKAP6 and ALG10B.

Predicting Disease-Related Proteins using Individual method *RW-Indiv*

The following steps were performed for predicting new LQTS-related proteins:

1. A new *trainSet* was built containing all the proteins annotated as being involved in LQTS (positive set) in addition to 100 randomly selected proteins (negative set).

2. A *testSet* was built containing all the remaining proteins in the network.

3. The *RW-indiv* method was used to rank the proteins in the *testset*.

Table 6.11 lists the highest ranked newly identified LQTS-related genes. In agreement with expections, many of the genes identified are (as hERG itself) voltage-gated Potassium channels; however also Sodium channels (SNC4A), Calcium channels (CACNB3 and CACNA1A) and solute carriers (SLC8A1) appear in the list. This is in agreement with the known proteins involved in the regulation of cardiac action potential, which are known to involve all three types of ions. KCNJ8 seems to be involved in cardiovascular sudden death at least in mouse models [51], indicating that while focusing on LQTS is of high practical relevance in today's drug development environment, one can in turn also assume that other ion channels involved in drug adverse reactions are currently not receiving sufficient attention. SLC8A1, as a sodium/calcium exchanger, is known to be involved in regulating action potential as well [1], though it is not easy to find a specific link to the QT interval prolongation in this case. SCN4A mutations have bene found to be insignificant under standard conditions, but become relevant in patients treated with LQ-inducing drugs [89]. This finding is interesting since it appears also synergistic adverse relations between genes and LQTS syndrome can be identified using our network approach. One of the potassium channels newly identified to be involved in cardiac action potential regulation (and, hence, with potential LQTS liability) is KCJN12 [49], which is indeed thought to be invovled in providing the cardiac inward rectifier current (IK1). A similar observation can be made regarding KCNA1, where it is thought that a brain-driven cardiac dysfunction can be made responsible for sudden death syndrome in epilespy patients [35]. Mutations in CACNA1 are classified as 'LQTS8' and, while rare, have been shown to be linked to LQTS [75]. Hence, overall we can find associations between the genes identified here and LQTS in many cases - and, interestingly, often they are dependent on the particular genetic or drug treatment conditions of the patients (such as in case of SCN4A and KCNA1).

6.5 Compare individual and network based prediction for LQTS

Considerable differences are apparent from the proteins included in the cluster including LQTS along with Epilespy and Dystonia (Table 6.5), and the prediction of LQTS-related proteins (Table 6.11). The receptors identified in Table 6.5 are on the one hand G-Protein Coupled Receptors (GPCRs) such as the Dopamine D1, D3 and D4 receptor subtypes identified with the highest rank in the disease cluster. The only ion channel selected is KCNQ4, which has been linked to deafness [21]; however, only related potassium channels appear to have been linked to LQTS until this stage. On the other hand, KCR1 (ALG10B), which is though to modulate sensititvity to drugs causing LQTS, also appears in this list (as well as in Table 6.10, in the list of most significant proteins according to ANOVA-based selection). On the other hand, Table

Table 6.11: Newly identified LQTS-related proteins by applying *RW-Indiv* method to the original seed proteins.

index	Gene-Name	Short Description
1	KCNH1	Potassium voltage-gated channel, subfamily H (eag-related), member 1
2	KCNH5	Potassium voltage-gated channel, subfamily H (eag-related), member 5
3	KCNJ8	Potassium inwardly-rectifying channel, subfamily J, member 8
4	SLC8A1	Solute carrier family 8 (sodium/calcium exchanger), member 1
5	SCN4A	Sodium channel, voltage-gated, type IV, alpha subunit
6	KCNJ4	Potassium inwardly-rectifying channel, subfamily J, member 4
7	CACNB3	Calcium channel, voltage-dependent, beta 3 subunit
8	KCNJ12	Potassium inwardly-rectifying channel, subfamily J, member 12
9	KCNA1	potassium voltage-gated channel, shaker-related subfamily, member 1 (episodic ataxia with myokymia)
10	CACNA1A	Calcium channel, voltage-dependent, P/Q type, alpha 1A subunit

6.11 is very much dominated by the different subtypes of voltage-gated potassium channels, which occupy 6 out of the 10 positions when RW-indiv is applied to the selection of novel proteins, with the remaining genes selected being ion channels or exchangers of sodium and/or calcium ions. Hence, it can be seen that both methods arrive at a very different selection of genes involved in the disease cluster, as well as the identification of novel disease genes using the RW-Indiv method. Combined with the fact that very disease relevant genes were identified in Table 6.11 (as discussed above), we believe that this illustrates the performance of the method implemented in this work.

6.6 Conclusions

Prediction accuracy of almost all the previous work on predicting disease-related proteins depends directly on the initial disease-related proteins (seed proteins). While the initial seed proteins of each disease suffers from several 'False Negative' cases, dependency of previous methods on the incomplete seed proteins is the main drawback of these methods.

In this article, we reduced the number of the False Negative cases in the initial seed proteins by proposing informative Human Disease Network (HDN). We analyzed different *Structural* and *Functional* prediction methods and we concluded that

a hybrid method which considers both Structural and Functional information in the PPI network is the best method for building the HDN. We built a HDN based on 20 diseases and we showed that resulting HDN is biologically meaningful. Then, we clustered HDN and we augmented the seed proteins of diseases based on the cluster they belong to. Finally, we predicted disease-related proteins based on the augmented version of seed proteins. Literature mining of the newly found disease-related proteins proved the usefulness of using our proposed HDN for predicting disease-related proteins.

6.7 Acknowledgment

This research is funded by the Dutch Science Foundation (NWO) through a VIDI grant.

Chapter 7

Conclusions

7.1 Introduction

In the previous chapters of this thesis, we presented several approaches for advancing the state-of-the-art for a number of the tasks in the Protein-Protein Interaction (PPI) networks. In this final chapter, we discuss our main contributions and possible future trends for each open problem of PPI networks.

7.2 Shortest-Path Distance and Anova-based Feature Selection

We modeled the PPI network as a graph $G(V, E)$, where V is a set of nodes (proteins in our context) and E is a set of edges (interactions in our context) connecting pairs of nodes. Shortest-path distance is a simple and still powerful feature when the input data is modeled as a graph. In the context of PPI networks, this type of feature has been used for network clustering before. In chapters 2 and 4, we proposed to use this type of feature for predicting annotation information of proteins in the PPI networks. A general predicting procedure was as follows: First, we described the proteins based on their shortest-path distance to specific, automatically selected, other proteins in the PPI network. Second, we apply machine learning for the prediction task.

Noisy nature of PPI networks and high-dimensional description vectors in large graphs are potential problems of this general predicting procedure. We proposed to reduce the noise and dimensionality in the description vectors by only retaining the shortest-path distance to a few "important" nodes. We defined node i as an "important" node, if the shortest-path distance of some node v to i is likely to be relevant for v's classification. We applied the Anova measure to select the important proteins in the PPI networks. We used shortest-path distance as a predictive feature and the Anova measure as a feature selection strategy in chapters 2 and 4 of this thesis. In both cases, the empirical results proved the usefulness of the proposed features.

7.3 Collaborative Functions

One of the main open problems of PPI networks is to predict the functional annotation of proteins in the network. Most of the previous methods predict the proteins' functions based on guilt-by-association and here, we call it Similarity Assumption: Interacting proteins tend to have the similar functions. In chapter 3 of this thesis, we considered a biological process as an aggregation of each individual protein's functions. So, we assumed that topologically close proteins tend to have collaborative functions and not necessarily similar functions (Collaboration Assumption). We defined "collaborative functions" as pairs of functions that frequently interface with each other in different interacting proteins. To our knowledge, this was the first study that considered the collaboration assumption for the task of function prediction in

PPI networks. The information about which functions collaborate, can be extracted easily from a PPI network, and using that information leads to improved predictive accuracy. We proposed two methods for this purpose: The first method calculates the collaboration value of two functions based on an iterative reinforcement strategy. The second method adopts an artificial neural network. Empirical results confirmed that the notion of collaborativeness of functions, rather than similarity, is useful for the task of predicting the functions of proteins.

As a future works, we may apply this idea to other domains, outside PPI networks. The notion of homophily is well-known in network analysis; it states that similar nodes are more likely to be linked together. The notion of collaborativeness, in this context, could also be described as "selective heterophily". It remains to be seen to what extent this notion may lead to better predictive results in other types of networks.

7.4 Network Contextual Information

PPI networks have been widely used for the task of predicting proteins involved in cancer. When the input data is a PPI network, the main challenge is to find features with good predictive power that can be computed from this network. Previous machine learning based methods have mostly focused on the functional information about the protein for which a prediction is made, or proximity of known cancer-related genes in the PPI network. In chapter 4 of this thesis, we proposed the following two types of input features and we showed that these features have good predictive power.

1. Functional Context: While previous methods have considered GO annotations of proteins as predictive features, no methods up till now have considered GO annotations of the neighbors of those proteins at the same time. One advantage of using GO annotations of the neighbors for the prediction task is that GO annotations are often incomplete, and by collecting GO information from the neighbors of a protein p, we may get more information about p itself. This argument is backed up by the fact that GO annotations of proteins can often be predicted well from the GO annotations of their neighbors. However, this is not the only effect; there is also a direct relationship between a protein's involvement in cancer and the GO annotations of the proteins it interacts with.

2. Structural Context: This context relates to the relative position of proteins in the network. Several previous methods described each protein p based on the shortest-path distance of p to some previously known cancer/disease proteins. Alternatively, we could describe a protein's position relative to other proteins than only cancer-related ones. In this thesis, we proposed a relevance measure for proteins that is inspired by statistical Anova, and showed that shortest-path distance to a relatively small number of proteins (selected according to the Anova-based measure) is informative for the task of predicting cancer-related proteins in the PPI networks.

Empirical results proved that the proposed network contextual information (functional and structural contexts) of a protein in a PPI network, offer additional information regarding the possible involvement of a protein in cancer. These features increase the accuracy of predictive models and have a biological interpretation.

7.5 Interaction-based Chi-square

The task of predicting in a PPI network which proteins are involved in cancer has received a significant amount of attention in the literature. Several approaches have been proposed based on machine learning methods. Their success depends on two main parameters: First, feature representation of the proteins and second, choosing the right machine learning method. The previous methods studied these two parameters and found that the quality of the prediction results depends only slightly on the chosen machine learning method, but strongly on the chosen features, and after considering different protein's features individually, Gene Ontology (GO) annotations turned out to be particularly important. Several authors proposed to use a χ^2-based feature selection method to select the most relevant GO terms.

Selecting individual discriminative functions based on original χ^2 does not consider the network topology and the way different functions interact with each other in the network. For the task of predicting cancer-related proteins, it is possible that a function f_i does not correlate itself with cancer-involvement, but when a protein with function f_i interacts with a protein with function f_j, this interaction may be an indication of the former protein being involved in a cancer. In chapter 5 of this thesis, we proposed a new method, "Interaction-based Chi-square", to combine the GO annotations of proteins with the information contained in the topology of a PPI network for the feature selection task. Empirical results show that our proposed interactive features are biologically meaningful and improve the prediction accuracy of these systems.

7.6 Informative Human Disease Network

Identification of novel proteins likely involved in diseases is an important issue in the area of computational biology. Previous methods assumed to have a set of proteins which are previously known to be involved in disease (i.e., seed proteins) and then, they try to extend the seed proteins by predicting new disease-related proteins. In almost all the discussed methods, prediction accuracy depends directly on the initial seed proteins. While the initial seed proteins of each disease suffers from several 'False Negative' cases (i.e., disease-related proteins which are not annotated as being involved in disease), dependency of previous methods to the incomplete seed proteins is the main drawback of these methods.

In chapter 6 of this thesis, we reduced the number of False Negative cases in the initial seed proteins by proposing an informative Human Disease Network (HDN). We analyzed different *Structural* and *Functional* prediction methods and we concluded

that a hybrid method which considers both structural and functional information in the PPI network is the best method for building the HDN. We built a HDN based on 20 diseases and we showed that it is biologically meaningful. Then, we clustered the HDN and we augmented the seed proteins of diseases based on the cluster they belong to. Finally, we predicted disease-related proteins based on the augmented version of seed proteins. Literature mining of the newly found disease-related proteins proved the usefulness of our proposed HDN for predicting disease-related proteins.

7.7 Future Works

As a future works, we could apply our contributions to other domains, outside PPI networks. For example, Rahmani et al., [101] predicted the social tags in the graph of annotated web pages based on the "selective heterophily" notion discussed in chapter 3. They observed that this idea improves the prediction accuracy. We could also use our proposed HDN for a hypothesis generation about diseases' drug targets.

Bibliography

[1] Slc8a1 solute carrier family 8 (sodium/calcium exchanger), member 1 [homo sapiens].

[2] Michael J. Ackerman and Peter J. Mohler. Defining a new paradigm for human arrhythmia syndromes. *Circulation Research*, 107(4):457–465, 2010.

[3] Ramon Aragues, Chris Sander, and Baldo Oliva. Predicting cancer involvement of genes from heterogeneous data. *BMC Bioinformatics*, 9(1):172, 2008.

[4] Kyriaki Bakirtzi, Maria Hatziapostolou, Iordanes Karagiannides, Christos Polytarchou, Savina Jaeger, Dimitrios Iliopoulos, and Charalabos Pothoulakis. Neurotensin signaling activates microrna s -21 and -155 and akt, promotes tumor growth in mice, and is increased in human colon tumors. *Gastroenterology*, Jul 2011.

[5] Bala S. Balakumaran, J. Taylor Herbert, and Phillip G. Febbo. Myc activity mitigates response to rapamycin in prostate cancer through 4ebp1-mediated inhibition of autophagy. *Autophagy*, 6(2):281–282, Feb 2010.

[6] Seth I. Berger, Avi Ma'ayan, and Ravi Iyengar. Systems Pharmacology of Arrhythmias. *Sci. Signal.*, 3(118):ra30+, April 2010.

[7] Arun Bhardwaj, Seema Singh, Sanjeev K. Srivastava, Richard E. Honkanen, Eddie Reed, and Ajay P. Singh. Modulation of protein phosphatase 2a activity alters androgen-independent growth of prostate cancer cells: therapeutic implications. *Mol Cancer Ther*, 10(5):720–731, May 2011.

[8] Giampaolo Bianchini, Yuan Qi, Ricardo H. Alvarez, Takayuki Iwamoto, Charles Coutant, Nuhad K. Ibrahim, Vicente Valero, Massimo Cristofanilli, Marjorie C. Green, Laszlo Radvanyi, Christos Hatzis, Gabriel N. Hortobagyi, Fabrice Andre, Luca Gianni, W. Fraser Symmans, and Lajos Pusztai. Molecular anatomy of breast cancer stroma and its prognostic value in estrogen receptor-positive and -negative cancers. *J Clin Oncol*, 28(28):4316–4323, Oct 2010.

[9] Hendrik Blockeel, Leander Schietgat, Jan Struyf, Saso Dzeroski, and Amanda Clare. Decision trees for hierarchical multilabel classification: A case study in functional genomics. In Johannes Fürnkranz, Tobias Scheffer, and Myra Spiliopoulou, editors, *PKDD*, volume 4213 of *Lecture Notes in Computer Science*, pages 18–29. Springer, 2006.

[10] Enrrico Bloise, Henrique L. Couto, Lauretta Massai, Pasquapina Ciarmela, Marzia Mencarelli, Lavinia E. Borges, Michela Muscettola, Giovanni Grasso, Vania F. Amaral, Geovanni D. Cassali, Felice Petraglia, and Fernando M. Reis. Differential expression of follistatin and flrg in human breast proliferative disorders. *BMC Cancer*, 9:320, Sep 2009.

[11] Annika Brendle, Haixin Lei, Andreas Brandt, Robert Johansson, Kerstin Enquist, Roger Henriksson, Kari Hemminki, Per Lenner, and Asta Försti. Polymorphisms in predicted microrna-binding sites in integrin genes and breast cancer: Itgb4 as prognostic marker. *Carcinogenesis*, 29(7):1394–1399, Jul 2008.

[12] John Brognard, Amy S. Clark, Yucheng Ni, and Phillip A. Dennis. Akt/protein kinase b is constitutively active in non-small cell lung cancer cells and promotes cellular survival and resistance to chemotherapy and radiation. *Cancer Research*, 61(10):3986–3997, 2001.

[13] Christine Brun, Carl Herrmann, and Alain Guénoche. Clustering proteins from interaction networks for the prediction of cellular functions. *BMC Bioinformatics*, 5:95, 2004.

[14] Wenjun Chang, Liye Ma, Liping Lin, Liqiang Gu, Xiaokang Liu, Hui Cai, Yongwei Yu, Xiaojie Tan, Yujia Zhai, Xingxing Xu, Minfeng Zhang, Lingling Wu, Hongwei Zhang, Jianguo Hou, Hongyang Wang, and Guangwen Cao. Identification of novel hub genes associated with liver metastasis of gastric cancer. *Int J Cancer*, 125(12):2844–2853, Dec 2009.

[15] Fei Chen. *JUNB jun B proto-oncogene*, 2011 (accessed June 8, 2011). `http://atlasgeneticsoncology.org/Genes/JUNBID178.html`.

[16] Jing Chen, Bruce Aronow, and Anil Jegga. Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinformatics*, 10(1):73, 2009.

[17] Yuan Chen, Tiantian Cui, Thomas Knösel, Linlin Yang, Kristin Zöller, and Iver Petersen. Igfbp7 is a p53 target gene inactivated in human lung cancer by dna hypermethylation. *Lung Cancer*, 73(1):38–44, Jul 2011.

[18] Hon Nian Chua, Wing-Kin Sung, and Limsoon Wong. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, 22(13):1623–1630, 2006.

[19] Pinchas Cohen, Donna M. Peehl, George Lamson, and Ron G. Rosenfeld. Insulin-like growth factors (igfs), igf receptors, and igf-binding proteins in primary cultures of prostate epithelial cells. *Journal of Clinical Endocrinology and Metabolism*, 73(2):401–407, 1991.

[20] Diane J. Cook and Lawrence B. Holder. *Mining Graph Data*. John Wiley & Sons, 2006.

[21] Paul J. Coucke, Peter Van Hauwe, Philip M. Kelley, Henricus Kunst, Isabelle Schatteman, DÃľsirÃľe Van Velzen, Johan Meyers, Robbert J. Ensink, Margriet Verstreken, Frank Declau, Henri Marres, Kumar Kastury, Shalender Bhasin, Wyman T. McGuirt, Richard J. H. Smith, Cor W.R.J. Cremers, Paul Van de Heyning, Patrick J. Willems, Shelley D. Smith, and Guy Van Camp. Mutations in the kcnq4 gene are responsible for autosomal dominant deafness in four dfna2 families. *Human Molecular Genetics*, 8(7):1321–1328, 1999.

[22] Breast Cancer Database. *Breast Cancer Database*, 2011 (accessed June 8, 2011). `http://www.itb.cnr.it/breastcancer/php/geneReport.php?id=55250`.

[23] Cancer Gene Database. *Cancer Gene Database*, 2011 (accessed Sep 1, 2011). `http://ncicb.nci.nih.gov/projects/cgdcp`.

[24] GeneCards Human Gene Database. *C20orf185 Gene - GeneCards | LPLC3 Protein | LPLC3 Antibody*, 2011 (accessed June 8, 2011). `http://www.genecards.org/cgi-bin/carddisp.pl?gene=C20orf185`.

[25] Charlotte M. Deane, Łukasz Salwiński, Ioannis Xenarios, and David Eisenberg. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Molecular & cellular proteomics : MCP*, 1(5):349–356, May 2002.

[26] Zoltan Dezso, Yuri Nikolsky, Tatiana Nikolskaya, Jeremy Miller, David Cherba, Craig Webb, and Andrej Bugrim. Identifying disease-specific genes based on their topological significance in protein networks. *BMC Systems Biology*, 3(1):36+, 2009.

[27] Sinan Erten, Gurkan Bebek, Rob Ewing, and Mehmet Koyuturk. Dada: Degree-aware algorithms for network-based disease gene prioritization. *BioData Min*, 4, 2011.

[28] William E. Fisher, Laszlo G. Boros, and William J. Schirmer. Insulin promotes pancreatic cancer: Evidence for endocrine influence on exocrine pancreatic tumors. *Journal of Surgical Research*, 63(1):310 – 313, 1996.

[29] Simon Furney, Desmond Higgins, Christos Ouzounis, and Nuria Lopez-Bigas. Structural and functional properties of genes involved in human cancer. *BMC Genomics*, 7(1):3, 2006.

[30] Simon J. Furney, Borja Calvo, Pedro LarraÃśaga, Jose A. Lozano, and Nuria Lopez-Bigas. Prioritization of candidate cancer genesâĂŤan aid to oncogenomic studies. *Nucleic Acids Research*, 36(18):e115, 2008.

[31] T. K. B. Gandhi, Jun Zhong, Suresh Mathivanan, L. Karthick, K. N. Chandrika, Sujatha S. Mohan, Salil Sharma, Stefan Pinkert, Shilpa Nagaraju, Balamurugan Periaswamy, Goparani Mishra, Kannabiran Nandakumar, Beiyi Shen, Nandan Deshpande, Rashmi Nayak, Malabika Sarker, Jef D. Boeke, Giovanni Parmigiani, Jörg Schultz, Joel S. Bader, and Akhilesh Pandey. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature Genetics*, 38(3):285–293, February 2006.

[32] GeneCards. *GeneCards*, 2011 (accessed Sep 1, 2011). `http://www.genecards.org/`.

[33] Rania B. Georges, Hassan Adwan, Hadjar Hamdi, Thomas Hielscher, Ulrich Linnemann, and Martin R. Berger. The insulin-like growth factor binding proteins 3 and 7 are associated with colorectal cancer and liver metastasis. *Cancer Biol Ther*, 12(1):69–79, Jul 2011.

[34] A. Ghellal, C. Li, M. Hayes, G. Byrne, N. Bundred, and S. Kumar. Prognostic significance of tgf beta 1 and tgf beta 3 in human breast carcinoma. *Anticancer Res*, 20(6B):4413–4418, Nov/Dec 2000.

[35] Edward Glasscock, Jong W. Yoo, Tim T. Chen, Tara L. Klassen, and Jeffrey L. Noebels. Kv1.1 potassium channel deficiency reveals brain-driven cardiac dysfunction as a candidate mechanism for sudden unexplained death in epilepsy. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 30(15):5167–5175, April 2010.

[36] GO. *The Gene Ontology*, 2012 (accessed March 29, 2012). `http://www.geneontology.org/`.

[37] Heike Goehler, Maciej Lalowski, Ulrich Stelzl, Stephanie Waelter, Martin Stroedicke, Uwe Worm, Anja Droege, Katrin S. Lindenberg, Maria Knoblich, Christian Haenig, Martin Herbst, Jaana Suopanki, Eberhard Scherzinger, Claudia Abraham, Bianca Bauer, Renate Hasenbank, Anja Fritzsche, Andreas H. Ludewig, Konrad Büssow, Konrad Buessow, Sarah H. Coleman, Claire-Anne A. Gutekunst, Bernhard G. Landwehrmeyer, Hans Lehrach, and Erich E. Wanker. A protein interaction network links git1, an enhancer of huntingtin aggregation, to huntington's disease. *Mol Cell*, 15(6):853–865, 2004.

[38] Kwang-Il Goh, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, May 2007.

[39] U. Guldener, M. Munsterkotter, G. Kastenmuller, N. Strack, J. van Helden, C. Lemer, J. Richelles, S. J. Wodak, J. Garcia-Martinez, J. E. Perez-Ortin, H. Michael, A. Kaps, E. Talla, B. Dujon, B. Andre, J. L. Souciet, J. De Montigny, E. Bon, C. Gaillardin, and H. W. Mewes. Cygd: the comprehensive yeast genome database. *Nucleic Acids Research*, 33(Supplement 1):D364+, January 2005.

[40] Lucia Gullotti, Jacqueline Czerwitzki, Jutta Kirfel, Peter Propping, Nils Rahner, Verena Steinke, Philip Kahl, Christoph Engel, Roland Schüle, Reinhard Buettner, and Nicolaus Friedrichs. Fhl2 expression in peritumoural fibroblasts correlates with lymphatic metastasis in sporadic but not in hnpcc-associated colon cancer. *Lab Invest*, Aug 2011.

[41] Ji-Youn Han, Geon Kook Lee, Sun Young Yoo, Sung Jin Yoon, Eun Young Cho, Heung Tae Kim, and Jin Soo Lee. Association of sumo1 and ubc9 genotypes with tumor response in non-small-cell lung cancer treated with irinotecan-based chemotherapy. *Pharmacogenomics J*, 10(2):86–93, Apr 2010.

[42] Christine L. Hattrup and Sandra J. Gendler. Muc1 alters oncogenic events and transcription in human breast cancer cells. *Breast Cancer Res*, 8(4):R37, 2006.

[43] Xiao-Ping He, Chang-Qing Su, Xing-Hua Wang, Xue Pan, Zhen-Xing Tu, Yang-Fang Gong, Jun Gao, Zhuan Liao, Jing Jin, Hong-Yu Wu, Xiao-Hua Man, and Zhao-Shen Li. E1b-55kd-deleted oncolytic adenovirus armed with canstatin gene yields an enhanced anti-tumor efficacy on pancreatic cancer. *Cancer Lett*, 285(1):89–98, Nov 2009.

[44] Sandra Heesch, Isabelle Bartram, Martin Neumann, Jana Reins, Maximilian Mossner, Cornelia Schlee, Andrea Stroux, Torsten Haferlach, Nicola Goekbuget, Dieter Hoelzer, Wolf-Karsten Hofmann, Eckhard Thiel, and Claudia D. Baldus. Expression of igfbp7 in acute leukemia is regulated by dna methylation. *Cancer Sci*, 102(1):253–259, Jan 2011.

[45] Dominique B Hoelzinger, Ana Lucia Dominguez, Shannon E Smith, Soraya Zorro Manrique, and Joseph Lustgarten. Ccl1: a novel therapeutic target for the modulation of treg function: Implications for immmunotherapy of cancer. *The Journal of Immunology*, pages 182, 40.10, 2009.

[46] Fung Yu Huang, Pui Man Chiu, Kar Fai Tam, Yvonne K. Y. Kwok, Elizabeth T. Lau, Mary H. Y. Tang, Tong Yow Ng, Vincent W. S. Liu, Annie N. Y. Cheung, and Hextan Y. S. Ngan. Semi-quantitative fluorescent pcr analysis identifies prkaa1 on chromosome 5 as a potential candidate cancer gene of cervical cancer. *Gynecol Oncol*, 103(1):219–225, Oct 2006.

[47] Online Mendelian Inheritance in Man. *Online Mendelian Inheritance in Man*, 2011 (accessed Sep 1, 2011). `http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim`.

[48] Ronald Jansen, Haiyuan Yu, Dov Greenbaum, Yuval Kluger, Nevan J. Krogan, Sambath Chung, Andrew Emili, Michael Snyder, Jack F. Greenblatt, and Mark Gerstein. A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644):449–453, 2003.

[49] Muneshige Kaibara, Keiko Ishihara, Yoshiyuki Doi, Hideki Hayashi, Tsuguhisa Ehara, and Kohtaro Taniyama. Identification of human kir2.2 (kcnj12) gene encoding functional inward rectifier potassium channel in both mammalian cells and xenopus oocytes. *FEBS Lett*, 531(2):250–254, Nov 2002.

[50] Raghu Kalluri and Michael Zeisberg. Fibroblasts in cancer. *Nat Rev Cancer*, 6(5):392–401, 2006.

[51] Garvan C. Kane, Chen-Fuh Lam, Fearghas O'Cochlain, Denice M. Hodgson, Santiago Reyes, Xiao-Ke Liu, Takashi Miki, Susumu Seino, Zvonimir S. Katusic, and Andre Terzic. Gene knockout of the kcnj8-encoded kir6.1 katp channel imparts fatal susceptibility to endotoxemia. *The FASEB Journal*, 20(13):2271–2280, 2006.

[52] Gozde Kar, Attila Gursoy, and Ozlem Keskin. Human cancer protein-protein interaction network: A structural perspective. *PLoS Comput Biol*, 5(12):e1000601+, December 2009.

[53] Masuko Katoh and Masaru Katoh. Identification and characterization of human snail3 (snai3) gene in silico. *Int J Mol Med*, 11(3):383–388, Mar 2003.

[54] Bi-Xia Ke, Salvatore Pepe, David R. Grubb, Jasper C. Komen, Adrienne Laskowski, Felicity A. Rodda, Belinda M. Hardman, James J. Pitt, Michael T. Ryan, Michael Lazarou, Jane Koleff, Michael M. H. Cheung, Joseph J. Smolich, and David R. Thorburn. Tissue-specific splicing of an ndufs6 gene-trap insertion generates a mitochondrial complex i deficiency-specific cardiomyopathy. *Proceedings of the National Academy of Sciences*, 2012.

[55] Jung Hwa Kim, Ji Min Lee, Hye Jin Nam, Hee June Choi, Jung Woo Yang, Jason S. Lee, Mi Hyang Kim, Su-Il Kim, Chin Ha Chung, Keun Il Kim, and Sung Hee Baek. Sumoylation of pontin chromatin-remodeling complex reveals a signal integration code in prostate cancer cells. *Proc Natl Acad Sci U S A*, 104(52):20793–20798, Dec 2007.

[56] Kyu Kwang Kim and Han Bok Kim. Protein interaction network related to helicobacter pyloriinfection response. *World J Gastroenterol*, 15(36):4518–28, 2009.

[57] A. D. King, Natasa Przulj, and Igor Jurisica. Protein complex prediction via cost-based clustering. *Bioinformatics*, 20(17):3013–3020, 2004.

[58] Sebastian Köhler, Sebastian Bauer, Denise Horn, and Peter N. Robinson. Walking the interactome for prioritization of candidate disease genes. *American journal of human genetics*, 82(4):949–958, April 2008.

[59] Nevan J. Krogan, Gerard Cagney, Haiyuan Yu, Gouqing Zhong, Xinghua Guo, Alexandr Ignatchenko, Joyce Li, Shuye Pu, Nira Datta, Aaron P. Tikuisis, Thanuja Punna, José M. Peregrín-Alvarez, Michael Shales, Xin Zhang, Michael Davey, Mark D. Robinson, Alberto Paccanaro, James E. Bray, Anthony Sheung, Bryan Beattie, Dawn P. Richards, Veronica Canadien, Atanas Lalev, Frank Mena, Peter Wong, Andrei Starostine, Myra M. Canete, James Vlasblom, Samuel Wu, Chris Orsi, Sean R. Collins, Shamanta Chandran, Robin Haw, Jennifer J. Rilstone, Kiran Gandi, Natalie J. Thompson, Gabe Musso, Peter St Onge, Shaun Ghanny, Mandy H. Lam, Gareth Butland, Amin M. Altaf-Ul, Shigehiko Kanaya, Ali Shilatifard, Erin O'Shea, Jonathan S. Weissman, C. James Ingles, Timothy R. Hughes, John Parkinson, Mark Gerstein, Shoshana J. Wodak, Andrew Emili, and Jack F. Greenblatt. Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, 440(7084):637–643, March 2006.

[60] Sabina Kupershmidt, Iris C-H Yang, Kenshi Hayashi, Jian Wei, Siprachanh Chanthaphaychith, Christina I Petersen, Daivid C Johns, Alfred L George, Dan M Roden, and Jeffrey R Balser. The ikr drug response is modulated by kcr1 in transfected cardiac and noncardiac cell lines. *FASEB J*, 17(15):2263–5, 2003.

[61] Ming-Tsung Lai, Chun-Hung Hua, Ming-Hsui Tsai, Lei Wan, Ying-Ju Lin, Chih-Mei Chen, I-Wen Chiu, Carmen Chan, Fuu-Jen Tsai, and Jim Jinn-Chyuan Sheu. Talin-1 overexpression defines high risk for aggressive oral squamous cell carcinoma and promotes cancer metastasis. *J Pathol*, 224(3):367–376, Jul 2011.

[62] Minalini Lakshman, Xiaoke Huang, Vijayalakshmi Ananthanarayanan, Borko Jovanovic, Yueqin Liu, Clarissa S. Craft, Diana Romero, Calvin P. H. Vary, and Raymond C. Bergan. Endoglin suppresses human prostate cancer metastasis. *Clin Exp Metastasis*, 28(1):39–53, Jan 2011.

[63] Ellen Lammerts, Pernilla Roswall, Christian Sundberg, Philip J. Gotwals, Victor E. Koteliansky, Rolf K. Reed, Nils-Erik Heldin, and Kristofer Rubin. Interference with tgf-beta1 and -beta3 in tumor stroma lowers tumor interstitial fluid pressure independently of growth in experimental carcinoma. *Int J Cancer*, 102(5):453–462, Dec 2002.

[64] Piotr Laudanski, Oksana Kowalczuk, Dagmara Klasa-Mazurkiewicz, Tomasz Milczek, Dominik Rysak-Luberowicz, Magdalena Garbowicz, Wlodzimierz Baranowski, Radoslaw Charkiewicz, Jacek Szamatowicz, and Lech Chyczewski. Selective gene expression profiling of mtor-associated tumor suppressor and oncogenes in ovarian cancer. *Folia Histochem Cytobiol*, 49(2):317–324, 2011.

[65] H.P Levy. *EhlersÐDanlos syndrome, hypermobility type. In: GeneReviews at GeneTests: Medical Genetics Information Resource (database online). Copyright, University of Washington, Seattle.*, 2011 (accessed March 5, 2012). `http://www.ncbi.nlm.nih.gov/books/NBK1279/`.

[66] Li Li, Kangyu Zhang, James Lee, Shaun Cordes, David Davis, and Zhijun Tang. Discovering cancer genes by integrating network and functional properties. *BMC Medical Genomics*, 2(1):61, 2009.

[67] Zhengyang Li, Yoko Sasaki, Masaru Mezawa, Shuang Wang, Xinyue Li, Li Yang, Zhitao Wang, Liming Zhou, Shouta Araki, Hiroyoshi Matsumura, Hideki Takai, and Yorimasa Ogata. camp and fibroblast growth factor 2 regulate bone sialoprotein gene expression in human prostate cancer cells. *Gene*, 471(1-2):1–12, Jan 2011.

[68] CiteXplore literature searching. *CiteXplore literature searching*, 2011 (accessed Sep 1, 2011). `http://www.ebi.ac.uk/citexplore/`.

[69] Huan Liu and Rudy Setiono. Chi2: Feature selection and discretization of numeric attributes. In *In Proceedings of the Seventh International Conference on Tools with Artificial Intelligence*, pages 388–391, 1995.

[70] Jian-Ping Lu, Jiao Zhang, Kwonseop Kim, Thomas C. Case, Robert J. Matusik, Yan-Hua Chen, Michael Wolfe, Jongdee Nopparat, and Qun Lu. Human homolog of drosophila hairy and enhancer of split 1, hes1, negatively regulates δ-catenin (ctnnd2) expression in cooperation with e2f1 in prostate cancer. *Mol Cancer*, 9:304, Nov 2010.

[71] Zelmina Lubovac, Jonas Gamalielsson, and Björn Olsson. Combining functional and topological properties to identify core modules in protein interaction networks. *Proteins*, 64(4):948–959, 2006.

[72] Avi Ma'ayan, Sherry L. Jenkins, Joseph Goldfarb, and Ravi Iyengar. Network analysis of fda approved drugs and their targets. *The Mount Sinai journal of medicine, New York*, 74(1):27–32, April 2007.

[73] Ange Maguy, Terence E. Hebert, and Stanley Nattel. Involvement of lipid rafts and caveolae in cardiac ion channel function. *Cardiovascular Research*, 69(4):798–807, March 2006.

[74] Guido Marcucci, Kati Maharry, Michael D. Radmacher, Krzysztof Mrózek, Tamara Vukosavljevic, Peter Paschka, Susan P. Whitman, Christian Langer, Claudia D. Baldus, Chang-Gong Liu, Amy S. Ruppert, Bayard L. Powell, Andrew J. Carroll, Michael A. Caligiuri, Jonathan E. Kolitz, Richard A. Larson, and Clara D. Bloomfield. Prognostic significance of, and gene and microrna expression signatures associated with, cebpa mutations in cytogenetically normal acute myeloid leukemia with high-risk molecular features: a cancer and leukemia group b study. *J Clin Oncol*, 26(31):5078–5087, Nov 2008.

[75] Argelia Medeiros-Domingo, Pedro Iturralde-Torres, and Michael J Ackerman. Clinical and genetic characteristics of long qt syndrome (article). *Revista EspaÚola de CardiologŠa*, 60(07):739–752, 2007.

[76] Hans-Werner Mewes, Dmitrij Frishman, Christian Gruber, Birgitta Geier, Dirk Haase, Andreas Kaps, Kai Lemcke, Gertrud Mannhaupt, Friedhelm Pfeiffer, Christine M. Schüller, S. Stocker, and B. Weil. Mips: a database for genomes and protein sequences. *Nucleic Acids Research*, 28(1):37–40, 2000.

[77] Tijana Milenkovic, Vesna Memisevic, Anand K. Ganesan, and Natasa Przulj. Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related functional genomics data. *Journal of the Royal Society, Interface / the Royal Society*, 7(44):423–437, March 2010.

[78] Tijana Milenkovic and Natasa Przulj. Uncovering biological network function via graphlet degree signatures. *Cancer informatics*, 6:257–273, 2008.

[79] Tom Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[80] Joanna R. Morris, Chris Boutell, Melanie Keppler, Ruth Densham, Daniel Weekes, Amin Alamshah, Laura Butler, Yaron Galanty, Laurent Pangon, Tai Kiuchi, Tony Ng, and Ellen Solomon. The sumo modification pathway is involved in the brca1 response to genotoxic stress. *Nature*, 462(7275):886–890, Dec 2009.

[81] Victor Neduva, Rune Linding, Isabelle Su-Angrand, Alexander Stark, Federico d. Masi, Toby J. Gibson, Joe Lewis, Luis Serrano, and Robert B. Russell. Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol*, 3(12):e405, 2005.

[82] Chor-Fung Ng, Patrick Kwok-Shing Ng, Vivian Wai-Yan Lui, Jialiang Li, Judy Yuet-Wa Chan, Kwok-Pui Fung, Yuen-Keng Ng, Paul Bo-San Lai, and Stephen Kwok Wing Tsui. Fhl2 exhibits anti-proliferative and anti-apoptotic activities in liver cancer cells. *Cancer Lett*, 304(2):97–106, May 2011.

[83] Kyoto Encyclopedia of Genes and Genomes. *Kyoto Encyclopedia of Genes and Genomes*, 2011 (accessed Sep 1, 2011). `http://www.genome.jp/kegg/disease/`.

[84] R. Ogawa, H. Ishiguro, Y. Kuwabara, M. Kimura, A. Mitsui, Y. Mori, R. Mori, K. Tomoda, T. Katada, K. Harada, and Y. Fujii. Identification of candidate genes involved in the radiosensitivity of esophageal cancer cells by microarray analysis. *Dis Esophagus*, 21(4):288–297, 2008.

[85] Hirokazu Ogino, Seiji Yano, Soji Kakiuchi, Hiroaki Muguruma, Kenji Ikuta, Masaki Hanibuchi, Hisanori Uehara, Kunihiro Tsuchida, Hiromu Sugino, and Saburo Sone. Follistatin suppresses the production of experimental multiple-organ metastasis by small cell lung cancer cells in natural killer cell-depleted scid mice. *Clin Cancer Res*, 14(3):660–667, Feb 2008.

[86] Rachelle R. Olsen and Bruce R. Zetter. Evidence of a role for antizyme and antizyme inhibitor as regulators of human cancer. *Mol Cancer Res*, Aug 2011.

[87] M Oti, B Snel, MA Huynen, and HG Brunner. Predicting disease genes using protein-protein interactions. *J Med Genet*, 43(8):691–698, 2006.

[88] V. Papa, B. Gliozzo, G.M. Clark, W.L. McGuire, D. Moore, Y. Fujita-Yamaguchi, R. Vigneri, I.D. Goldfine, and V. Pezzino. Insulin-like growth factor-i receptors are overexpressed and predict a low risk in human breast cancer. *Cancer Res*, 53(16):3736–40, 1993.

[89] Y. Pereon, G. Lande, S. Demolombe, S. Nguyen The Tich, D. Sternberg, H. Le Marec, and A. David. Paramyotonia congenita with an scn4a mutation affecting cardiac repolarization. *Neurology*, 60(2):340–2, 2003.

[90] Gizeh Pérez-Tenorio, Elin Karlsson, Marie Ahnström Waltersson, Birgit Olsson, Birgitta Holmlund, Bo Nordenskjöld, Tommy Fornander, Lambert Skoog, and Olle Staal. Clinical potential of the mtor targets s6k1 and s6k2 in breast cancer. *Breast Cancer Res Treat*, 128(3):713–723, Aug 2011.

[91] Suraj Peri, J. Daniel Navarro, Troels Z. Kristiansen, Ramars Amanchy, Vineeth Surendranath, Babylakshmi Muthusamy, T. K. Gandhi, K. N. Chandrika, Nandan Deshpande, Shubha Suresh, B. P. Rashmi, K. Shanker, N. Padma, Vidya Niranjan, H. C. Harsha, Naveen Talreja, B. M. Vrushabendra, M. A. Ramya, A. J. Yatish, Mary Joy, H. N. Shivashankar, M. P. Kavitha, Minal Menezes, Dipanwita Roy R. Choudhury, Neelanjana Ghosh, R. Saravana, Sreenath Chandran, Sujatha Mohan, Chandra Kiran K. Jonnalagadda, C. K. Prasad, Chandan Kumar-Sinha, Krishna S. Deshpande, and Akhilesh Pandey. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res*, 32(Database issue), January 2004.

[92] C. I. Petersen, T. R. Mcfarland, S. Z. Stepanovic, P. Yang, D. J. Reiner, K. Hayashi, A. L. George, D. M. Roden, J. H. Thomas, and J. R. Balser. In vivo identification of genes that modify ether-a-go-go-related gene activity in Caenorhabditis elegans may also affect human cardiac arrhythmia. *Proceedings of the National Academy of Sciences USA*, 101:11773–11778, 2004.

[93] Shirley M. Potter, Roisin M. Dwyer, Marion C. Hartmann, Sonja Khan, Marie P. Boyle, Catherine E. Curran, and Michael J. Kerin. Influence of stromal-epithelial interactions on breast cancer in vitro and in vivo. *Breast Cancer Res Treat*, Feb 2011.

[94] Dale Powner, Petra M. Kopp, Susan J. Monkley, David R. Critchley, and Fedor Berditchevski. Tetraspanin cd9 in cell migration. *Biochem Soc Trans*, 39(2):563–567, Apr 2011.

[95] Cancer Genome Project. *Cancer Genome Project*, 2011 (accessed Sep 1, 2011). `http://www.sanger.ac.uk/genetics/CGP/Census/`.

[96] Foster J. Provost and Tom Fawcett. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *KDD*, pages 43–48, 1997.

[97] Predrag Radivojac, Kang Peng, Wyatt T. Clark, Brandon J. Peters, Amrita Mohan, Sean M. Boyle, and Sean D. Mooney. An integrated approach to inferring gene-disease associations in humans. *Proteins*, 72(3):1030–1037, August 2008.

[98] Hossein Rahmani, Hendrik Blockeel, and Andreas Bender. Predicting the functions of proteins in protein-protein interaction networks from global information. *Journal of Machine Learning Research - Proceedings Track*, 8:82–97, 2010.

[99] Hossein Rahmani, Hendrik Blockeel, and Andreas Bender. Collaboration-based function prediction in protein-protein interaction networks. In João Gama, Elizabeth Bradley, and Jaakko Hollmén, editors, *IDA*, volume 7014 of *Lecture Notes in Computer Science*, pages 318–327. Springer, 2011.

[100] Hossein Rahmani, Hendrik Blockeel, and Andreas Bender. Interaction-based feature selection for predicting cancer-related proteins in protein-protein interaction networks. In Stefan Kramer and Neil Lawrence, editors, *Machine Learning in Systems Biology, Proceedings of the Fifth International Workshop, Vienna, Austria, July 20-21, 2011,*, pages 68–73, 2011.

[101] Hossein Rahmani, Behrooz Nobakht, and Hendrik Blockeel. Collaboration-based social tag prediction in the graph of annotated web pages. In Ruggero Pensa, Francesca Cordero, Celine Rouveirol, Rushed Kanawati, Jose Troyano, and Paolo Rosso, editors, *DyNaK 2010: Dynamic Networks and Knowledge Discovery,*, pages 1–12, 2010.

[102] Joachim Rassow, Wolfgang Voos, and Nikolaus Pfanner. Partner proteins determine multiple functions of hsp70. *Trends Cell Biol*, 5(5):207–212, May 1995.

[103] Alexander W. Rives and Timothy Galitski. Modular organization of cellular networks. *Proceedings of the National Academy of Sciences*, 100(3):1128–1133, February 2003.

[104] Heinz Ruffner, Andreas Bauer, and Tewis Bouwmeeste. Human protein-protein interaction networks and the value for drug discovery. *Drug Discov Today*, 12(17-18):709–716, 2007.

[105] Shinichi Sakamoto, Richard O. McCann, Rajiv Dhir, and Natasha Kyprianou. Talin1 promotes tumor invasion and metastasis via focal adhesion signaling and anoikis resistance. *Cancer Res*, 70(5):1885–1895, Mar 2010.

[106] Lee Sam, Yang Liu, Jianrong Li, Carol Friedman, and Yves A. Lussier. Discovery of protein interaction networks shared by diseases. *Pac Symp Biocomput*, pages 76–87, 2007.

[107] Sampoorna Satheesha, Victoria J. Cookson, Louise J. Coleman, Nicola Ingram, Brijesh Madhok, Andrew M. Hanby, Charlotte A. B. Suleman, Vicky S. Sabine, E. Jane Macaskill, John M. S. Bartlett, J. Michael Dixon, Jim N. McElwaine, and Thomas A. Hughes. Response to mtor inhibition: activity of eif4e predicts sensitivity in cell lines and acquired changes in eif4e regulation in breast cancer. *Mol Cancer*, 10:19, Feb 2011.

[108] Mikkel H. Schierup, Thomas Mailund, Heng Li, Jun Wang, Anne Tjønneland, Ulla Vogel, Lars Bolund, and Bjørn A. Nexø. Haplotype frequencies in a subregion of chromosome 19q13.3, related to risk and prognosis of cancer, differ dramatically between ethnic groups. *BMC Med Genet*, 10:20, Mar 2009.

[109] Andreas Schlicker, Thomas Lengauer, and Mario Albrecht. Improving disease gene prioritization using the semantic similarity of Gene Ontology terms. *Bioinformatics*, 26(18):i561–i567, September 2010.

[110] Wolfgang A. Schulz, Marc Ingenwerth, Carolle E. Djuidje, Christiane Hader, Jörg Rahnenführer, and Rainer Engers. Changes in cortical cytoskeletal and extracellular matrix gene expression in prostate cancer are related to oncogenic erg deregulation. *BMC Cancer*, 10:505, Sep 2010.

[111] Benno Schwikowski, Peter Uetz, , and Stanley Fields and. A network of protein-protein interactions in yeast. *Nat Biotechnol*, 18(12):1257–1261, December 2000.

[112] Serena Scollen, Craig Luccarini, Caroline Baynes, Kristy Driver, Manjeet K. Humphreys, Montserrat Garcia-Closas, Jonine Figueroa, Jolanta Lissowska, Paul D. Pharoah, Douglas F. Easton, Robin Hesketh, James C. Metcalfe, and Alison M. Dunning. Tgf-β signaling pathway and breast cancer susceptibility. *Cancer Epidemiol Biomarkers Prev*, 20(6):1112–1119, Jun 2011.

[113] Christof Seidl, Matthias Port, Christos Apostolidis, Frank Bruchertseifer, Markus Schwaiger, Reingard Senekowitsch-Schmidtke, and Michael Abend. Differential gene expression triggered by highly cytotoxic alpha-emitter-immunoconjugates in gastric cancer cells. *Invest New Drugs*, 28(1):49–60, Feb 2010.

[114] Martha L. Slattery, Abbie Lundgreen, Jennifer S. Herrick, and Roger K. Wolff. Genetic variation in rps6ka1, rps6ka2, rps6kb1, rps6kb2, and pdk1 and risk of colon or rectal cancer. *Mutat Res*, 706(1-2):13–20, Jan 2011.

[115] Sara Stahl, Rui Mm Branca, Ghazal Efazat, Maria Ruzzene, Boris Zhivotovsky, Rolf Lewensohn, Kristina Viktorsson, and Janne Lehtiö. Phosphoproteomic profiling of nsclc cells reveals that ephrin b3 regulates pro-survival signaling through akt1-mediated phosphorylation of the epha2 receptor. *J Proteome Res*, 10(5):2566–2578, May 2011.

[116] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. Biogrid: a general repository for interaction datasets. *Nucleic Acids Research*, 34(Database-Issue):535–539, 2006.

[117] M.S. Starr. The role of dopamine in epilepsy. *Synapse*, 22:159–94, 1996.

[118] Ulrich Stelzl, Uwe Worm, Maciej Lalowski, Christian Haenig, Felix H. Brembeck, Heike Goehler, Martin Stroedicke, Martina Zenkner, Anke Schoenherr, Susanne Koeppen, Jan Timm, Sascha Mintzlaff, Claudia Abraham, Nicole Bock, Silvia Kietzmann, Astrid Goedde, Engin Toksöz, Anja Droege, Sylvia Krobitsch, Bernhard Korn, Walter Birchmeier, Hans Lehrach, and Erich E. Wanker. A human protein-protein interaction network: A resource for annotating the proteome. 122(6):957–968, September 2005.

[119] Shiwei Sun, Yi Zhao, Yishan Jiao, Yifei Yin, Lun Cai, Yong Zhang, Hongchao Lu, Runsheng Chen, and Dongbo Bu. Faster and more accurate global protein function assignment from protein interaction networks using the mfgo algorithm. *FEBS Lett*, 580(7):1891–1896, Mar 2006.

[120] What Is Long QT Syndrome? *What Is Long QT Syndrome?*, 2011 (accessed October 27, 2011). `http://www.nhlbi.nih.gov/health/health-topics/topics/qt/`.

[121] Alexei Vazquez, Alessandro Flammini, Amos Maritan, and Alessandro Vespignani. Global protein function prediction from protein-protein interaction networks. *Nat Biotechnol*, 21(6):697–700, June 2003.

[122] Maria Claudia Vigliani, Patrizia Polo, Adriano ChiŸ, Bruno Giometto, Letizia Mazzini, and Davide Schiffer. Patients with amyotrophic lateral sclerosis and cancer do not differ clinically from patients with sporadic amyotrophic lateral sclerosis. *Journal of Neurology*, 247:778–782, 2000. 10.1007/s004150070092.

[123] Christian von Mering, Roland Krause, Berend Snel, Michael Cornell, Stephen G. Oliver, Stanley Fields, and Peer Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403, May 2002.

[124] Shi-Qiang Wang, Long-Sheng Song, Edward G. Lakatta, and Heping Cheng. Ca2+ signalling between single l-type ca2+ channels and ryanodine receptors in heart cells (article). *Nature*, 410(6828):592–596, 2001.

[125] T. Wang, Y-H Chen, H. Hong, Y. Zeng, J. Zhang, J-P Lu, B. Jeansonne, and Q. Lu. Increased nucleotide polymorphic changes in the 5'-untranslated region of delta-catenin (ctnnd2) gene in prostate cancer. *Oncogene*, 28(4):555–564, Jan 2009.

[126] Thomas C Wehler, Kirsten Frerichs, Claudine Graf, Daniel Drescher, Katrin Schimanski, Stefan Biesterfeld, Martin R Berger, Stephan Kanzler, Theodor

Junginger, Peter R Galle, Markus Moehler, Ines Gockel, and Carl C Schimanski. Pdgfralpha/beta expression correlates with the metastatic behavior of human colorectal cancer: a possible rationale for a molecular targeting strategy. *Oncology Reports*, 19(3):697–704, 2008.

[127] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999.

[128] Lee Lee Wong, Daohai Zhang, Chan Fong Chang, and Evelyn S. C. Koay. Silencing of the pp2a catalytic subunit causes her-2/neu positive breast cancer cells to undergo apoptosis. *Exp Cell Res*, 316(20):3387–3396, Dec 2010.

[129] SL Wong, LV Zhang, AH Tong, Z Li, DS Goldberg, OD King, G Lesage, M Vidal, B Andrews, and H Bussey. Combining biological networks to predict genetic interactions. *Proc Natl Acad Sci USA*, 101(44):15682–15687, 2004.

[130] Xuebing Wu and Shao Li. *Cancer Gene Prediction Using a Network Approach. Chapter 11 Mathematical and Computational Biology*. Cancer Systems Biology (Ed. Edwin Wang). Series: Chapman and Hall/CRC, 2010.

[131] Zhihong Xiong, Lihua Ding, Junzhong Sun, Jia Cao, Jing Lin, Zhaohui Lu, Yufei Liu, Cuifen Huang, and Qinong Ye. Synergistic repression of estrogen receptor transcriptional activity by fhl2 and smad4 in breast cancer cells. *IUBMB Life*, 62(9):669–676, Sep 2010.

[132] Jianzhen Xu and Yongjin Li. Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics*, 22(22):2800–2805, October 2006.

[133] Hideo Yasukawa, Atsuo Sasaki, and Akihiko Yoshimura. Negative regulation of cytokine signaling pathways. *Annu Rev Immunol*, 18, 2000.

[134] Peng Yue, William F. Forrest, Joshua S. Kaminker, Scott Lohr, Zemin Zhang, and Guy Cavet. Inferring the functional effects of mutation through clusters of mutations in homologous proteins. *Hum Mutat*, 31(3):264–271, Mar 2010.

[135] Pierre Zindy, Yann Bergé, Ben Allal, Thomas Filleron, Sandra Pierredon, Anne Cammas, Samantha Beck, Loubna Mhamdi, Li Fan, Gilles Favre, Jean-Pierre Delord, Henri Roché, Florence Dalenc, Magali Lacroix-Triki, and Stéphan Vagner. Formation of the eif4f translation-initiation complex determines sensitivity to anticancer drugs targeting the egfr and her2 receptors. *Cancer Res*, 71(12):4068–4073, Jun 2011.

Titles in the IPA Dissertation Series since 2006

E. Dolstra. *The Purely Functional Software Deployment Model.* Faculty of Science, UU. 2006-01

R.J. Corin. *Analysis Models for Security Protocols.* Faculty of Electrical Engineering, Mathematics & Computer Science, UT. 2006-02

P.R.A. Verbaan. *The Computational Complexity of Evolving Systems.* Faculty of Science, UU. 2006-03

K.L. Man and R.R.H. Schiffelers. *Formal Specification and Analysis of Hybrid Systems.* Faculty of Mathematics and Computer Science and Faculty of Mechanical Engineering, TU/e. 2006-04

M. Kyas. *Verifying OCL Specifications of UML Models: Tool Support and Compositionality.* Faculty of Mathematics and Natural Sciences, UL. 2006-05

M. Hendriks. *Model Checking Timed Automata - Techniques and Applications.* Faculty of Science, Mathematics and Computer Science, RU. 2006-06

J. Ketema. *Böhm-Like Trees for Rewriting.* Faculty of Sciences, VUA. 2006-07

C.-B. Breunesse. *On JML: topics in tool-assisted verification of JML programs.* Faculty of Science, Mathematics and Computer Science, RU. 2006-08

B. Markvoort. *Towards Hybrid Molecular Simulations.* Faculty of Biomedical Engineering, TU/e. 2006-09

S.G.R. Nijssen. *Mining Structured Data.* Faculty of Mathematics and Natural Sciences, UL. 2006-10

G. Russello. *Separation and Adaptation of Concerns in a Shared Data Space.* Faculty of Mathematics and Computer Science, TU/e. 2006-11

L. Cheung. *Reconciling Nondeterministic and Probabilistic Choices.* Faculty of Science, Mathematics and Computer Science, RU. 2006-12

B. Badban. *Verification techniques for Extensions of Equality Logic.* Faculty of Sciences, Division of Mathematics and Computer Science, VUA. 2006-13

A.J. Mooij. *Constructive formal methods and protocol standardization.* Faculty of Mathematics and Computer Science, TU/e. 2006-14

T. Krilavicius. *Hybrid Techniques for Hybrid Systems.* Faculty of Electrical Engineering, Mathematics & Computer Science, UT. 2006-15

M.E. Warnier. *Language Based Security for Java and JML.* Faculty of Science, Mathematics and Computer Science, RU. 2006-16

V. Sundramoorthy. *At Home In Service Discovery.* Faculty of Electrical Engineering, Mathematics & Computer Science, UT. 2006-17

B. Gebremichael. *Expressivity of Timed Automata Models.* Faculty of Science, Mathematics and Computer Science, RU. 2006-18

L.C.M. van Gool. *Formalising Interface Specifications.* Faculty of Mathematics and Computer Science, TU/e. 2006-19

C.J.F. Cremers. *Scyther - Semantics and Verification of Security Protocols.* Faculty of Mathematics and Computer Science, TU/e. 2006-20

J.V. Guillen Scholten. *Mobile Channels for Exogenous Coordination of Distributed Systems: Semantics, Implementation and Composition.* Faculty of Mathematics and Natural Sciences, UL. 2006-21

H.A. de Jong. *Flexible Heterogeneous Software Systems.* Faculty of Natural Sciences, Mathematics, and Computer Science, UvA. 2007-01

N.K. Kavaldjiev. *A run-time reconfigurable Network-on-Chip for streaming DSP applications.* Faculty of Electrical Engineering, Mathematics & Computer Science, UT. 2007-02

M. van Veelen. *Considerations on Modeling for Early Detection of Abnormalities in Locally Autonomous Distributed Systems.* Faculty of Mathematics and Computing Sciences, RUG. 2007-03

T.D. Vu. *Semantics and Applications of Process and Program Algebra.* Faculty of Natural Sciences, Mathematics, and Computer Science, UvA. 2007-04

L. Brandán Briones. *Theories for Model-based Testing: Real-time and Coverage.* Faculty of Electrical Engineering, Mathematics & Computer Science, UT. 2007-05

I. Loeb. *Natural Deduction: Sharing by Presentation.* Faculty of Science, Mathematics and Computer Science, RU. 2007-06

M.W.A. Streppel. *Multifunctional Geometric Data Structures.* Faculty of Mathematics and Computer Science, TU/e. 2007-07

N. Trčka. *Silent Steps in Transition Systems and Markov Chains.* Faculty of Mathematics and Computer Science, TU/e. 2007-08

R. Brinkman. *Searching in encrypted data.* Faculty of Electrical Engineering, Mathematics & Computer Science, UT. 2007-09

A. van Weelden. *Putting types to good use.* Faculty of Science, Mathematics and Computer Science, RU. 2007-10

J.A.R. Noppen. *Imperfect Information in Software Development Processes.* Faculty of Electrical Engineering, Mathematics & Computer Science, UT. 2007-11

R. Boumen. *Integration and Test plans for Complex Manufacturing Systems.* Faculty of Mechanical Engineering, TU/e. 2007-12

A.J. Wijs. *What to do Next?: Analysing and Optimising System Behaviour in Time.* Faculty of Sciences, Division of Mathematics and Computer Science, VUA. 2007-13

C.F.J. Lange. *Assessing and Improving the Quality of Modeling: A Series of Empirical Studies about the UML.* Faculty of Mathematics and Computer Science, TU/e. 2007-14

T. van der Storm. *Component-based Configuration, Integration and Delivery.* Faculty of Natural Sciences, Mathematics, and Computer Science,UvA. 2007-15

B.S. Graaf. *Model-Driven Evolution of Software Architectures.* Faculty of Electrical Engineering, Mathematics, and Computer Science, TUD. 2007-16

A.H.J. Mathijssen. *Logical Calculi for Reasoning with Binding.* Faculty of Mathematics and Computer Science, TU/e. 2007-17

D. Jarnikov. *QoS framework for Video Streaming in Home Networks.* Faculty of Mathematics and Computer Science, TU/e. 2007-18

M. A. Abam. *New Data Structures and Algorithms for Mobile Data.* Faculty of Mathematics and Computer Science, TU/e. 2007-19

W. Pieters. *La Volonté Machinale: Understanding the Electronic Voting Controversy.* Faculty of Science, Mathematics and Computer Science, RU. 2008-01

A.L. de Groot. *Practical Automaton Proofs in PVS.* Faculty of Science, Mathematics and Computer Science, RU. 2008-02

M. Bruntink. *Renovation of Idiomatic Crosscutting Concerns in Embedded Systems.* Faculty of Electrical Engineering, Mathematics, and Computer Science, TUD. 2008-03

A.M. Marin. *An Integrated System to Manage Crosscutting Concerns in Source Code.* Faculty of Electrical Engineering, Mathematics, and Computer Science, TUD. 2008-04

N.C.W.M. Braspenning. *Model-based Integration and Testing of High-tech Multi-disciplinary Systems.* Faculty of Mechanical Engineering, TU/e. 2008-05

M. Bravenboer. *Exercises in Free Syntax: Syntax Definition, Parsing, and Assimilation of Language Conglomerates.* Faculty of Science, UU. 2008-06

M. Torabi Dashti. *Keeping Fairness Alive: Design and Formal Verification of Optimistic Fair Exchange Protocols.* Faculty of Sciences, Division of Mathematics and Computer Science, VUA. 2008-07

I.S.M. de Jong. *Integration and Test Strategies for Complex Manufacturing Machines.* Faculty of Mechanical Engineering, TU/e. 2008-08

I. Hasuo. *Tracing Anonymity with Coalgebras.* Faculty of Science, Mathematics and Computer Science, RU. 2008-09

L.G.W.A. Cleophas. *Tree Algorithms: Two Taxonomies and a Toolkit.* Faculty of Mathematics and Computer Science, TU/e. 2008-10

I.S. Zapreev. *Model Checking Markov Chains: Techniques and Tools.* Faculty of Electrical Engineering, Mathematics & Computer Science, UT. 2008-11

M. Farshi. *A Theoretical and Experimental Study of Geometric Networks.* Faculty of Mathematics and Computer Science, TU/e. 2008-12

G. Gulesir. *Evolvable Behavior Specifications Using Context-Sensitive Wildcards.* Faculty of Electrical Engineering, Mathematics & Computer Science, UT. 2008-13

F.D. Garcia. *Formal and Computational Cryptography: Protocols, Hashes and Commitments.* Faculty of Science, Mathematics and Computer Science, RU. 2008-14

P. E. A. Dürr. *Resource-based Verification for Robust Composition of Aspects.* Faculty of Electrical Engineering, Mathematics & Computer Science, UT. 2008-15

E.M. Bortnik. *Formal Methods in Support of SMC Design.* Faculty of Mechanical Engineering, TU/e. 2008-16

R.H. Mak. *Design and Performance Analysis of Data-Independent Stream Processing Systems.* Faculty of Mathematics and Computer Science, TU/e. 2008-17

M. van der Horst. *Scalable Block Processing Algorithms.* Faculty of Mathematics and Computer Science, TU/e. 2008-18

C.M. Gray. *Algorithms for Fat Objects: Decompositions and Applications.* Faculty of Mathematics and Computer Science, TU/e. 2008-19

J.R. Calamé. *Testing Reactive Systems with Data - Enumerative Methods and Constraint Solving.* Faculty of Electrical Engineering, Mathematics & Computer Science, UT. 2008-20

E. Mumford. *Drawing Graphs for Cartographic Applications.* Faculty of Mathematics and Computer Science, TU/e. 2008-21

E.H. de Graaf. *Mining Semi-structured Data, Theoretical and Experimental Aspects of Pattern Evaluation.* Faculty of Mathematics and Natural Sciences, UL. 2008-22

R. Brijder. *Models of Natural Computation: Gene Assembly and Membrane Systems.* Faculty of Mathematics and Natural Sciences, UL. 2008-23

A. Koprowski. *Termination of Rewriting and Its Certification.* Faculty of Mathematics and Computer Science, TU/e. 2008-24

U. Khadim. *Process Algebras for Hybrid Systems: Comparison and Development.* Faculty of Mathematics and Computer Science, TU/e. 2008-25

J. Markovski. *Real and Stochastic Time in Process Algebras for Performance Evaluation.* Faculty of Mathematics and Computer Science, TU/e. 2008-26

H. Kastenberg. *Graph-Based Software Specification and Verification.* Faculty of Electrical Engineering, Mathematics & Computer Science, UT. 2008-27

I.R. Buhan. *Cryptographic Keys from Noisy Data Theory and Applications.* Faculty of Electrical Engineering, Mathematics & Computer Science, UT. 2008-28

R.S. Marin-Perianu. *Wireless Sensor Networks in Motion: Clustering Algorithms for Service Discovery and Provisioning.* Faculty of Electrical Engineering, Mathematics & Computer Science, UT. 2008-29

M.H.G. Verhoef. *Modeling and Validating Distributed Embedded Real-Time Control Systems.* Faculty of Science, Mathematics and Computer Science, RU. 2009-01

M. de Mol. *Reasoning about Functional Programs: Sparkle, a proof assistant for Clean.* Faculty of Science, Mathematics and Computer Science, RU. 2009-02

M. Lormans. *Managing Requirements Evolution.* Faculty of Electrical Engineering, Mathematics, and Computer Science, TUD. 2009-03

M.P.W.J. van Osch. *Automated Model-based Testing of Hybrid Systems.* Faculty of Mathematics and Computer Science, TU/e. 2009-04

H. Sozer. *Architecting Fault-Tolerant Software Systems.* Faculty of Electrical

Engineering, Mathematics & Computer Science, UT. 2009-05

M.J. van Weerdenburg. *Efficient Rewriting Techniques.* Faculty of Mathematics and Computer Science, TU/e. 2009-06

H.H. Hansen. *Coalgebraic Modelling: Applications in Automata Theory and Modal Logic.* Faculty of Sciences, Division of Mathematics and Computer Science, VUA. 2009-07

A. Mesbah. *Analysis and Testing of Ajax-based Single-page Web Applications.* Faculty of Electrical Engineering, Mathematics, and Computer Science, TUD. 2009-08

A.L. Rodriguez Yakushev. *Towards Getting Generic Programming Ready for Prime Time.* Faculty of Science, UU. 2009-9

K.R. Olmos Joffré. *Strategies for Context Sensitive Program Transformation.* Faculty of Science, UU. 2009-10

J.A.G.M. van den Berg. *Reasoning about Java programs in PVS using JML.* Faculty of Science, Mathematics and Computer Science, RU. 2009-11

M.G. Khatib. *MEMS-Based Storage Devices. Integration in Energy-Constrained Mobile Systems.* Faculty of Electrical Engineering, Mathematics & Computer Science, UT. 2009-12

S.G.M. Cornelissen. *Evaluating Dynamic Analysis Techniques for Program Comprehension.* Faculty of Electrical Engineering, Mathematics, and Computer Science, TUD. 2009-13

D. Bolzoni. *Revisiting Anomaly-based Network Intrusion Detection Systems.* Faculty of Electrical Engineering, Mathematics & Computer Science, UT. 2009-14

H.L. Jonker. *Security Matters: Privacy in Voting and Fairness in Digital Exchange.* Faculty of Mathematics and Computer Science, TU/e. 2009-15

M.R. Czenko. *TuLiP - Reshaping Trust Management.* Faculty of Electrical Engineering, Mathematics & Computer Science, UT. 2009-16

T. Chen. *Clocks, Dice and Processes.* Faculty of Sciences, Division of Mathematics and Computer Science, VUA. 2009-17

C. Kaliszyk. *Correctness and Availability: Building Computer Algebra on top of Proof Assistants and making Proof Assistants available over the Web.* Faculty of Science, Mathematics and Computer Science, RU. 2009-18

R.S.S. O'Connor. *Incompleteness & Completeness: Formalizing Logic and Analysis in Type Theory.* Faculty of Science, Mathematics and Computer Science, RU. 2009-19

B. Ploeger. *Improved Verification Methods for Concurrent Systems.* Faculty of Mathematics and Computer Science, TU/e. 2009-20

T. Han. *Diagnosis, Synthesis and Analysis of Probabilistic Models.* Faculty of Electrical Engineering, Mathematics & Computer Science, UT. 2009-21

R. Li. *Mixed-Integer Evolution Strategies for Parameter Optimization and Their Applications to Medical Image Analysis.* Faculty of Mathematics and Natural Sciences, UL. 2009-22

J.H.P. Kwisthout. *The Computational Complexity of Probabilistic Networks.* Faculty of Science, UU. 2009-23

T.K. Cocx. *Algorithmic Tools for Data-Oriented Law Enforcement*. Faculty of Mathematics and Natural Sciences, UL. 2009-24

A.I. Baars. *Embedded Compilers*. Faculty of Science, UU. 2009-25

M.A.C. Dekker. *Flexible Access Control for Dynamic Collaborative Environments*. Faculty of Electrical Engineering, Mathematics & Computer Science, UT. 2009-26

J.F.J. Laros. *Metrics and Visualisation for Crime Analysis and Genomics*. Faculty of Mathematics and Natural Sciences, UL. 2009-27

C.J. Boogerd. *Focusing Automatic Code Inspections*. Faculty of Electrical Engineering, Mathematics, and Computer Science, TUD. 2010-01

M.R. Neuhäußer. *Model Checking Nondeterministic and Randomly Timed Systems*. Faculty of Electrical Engineering, Mathematics & Computer Science, UT. 2010-02

J. Endrullis. *Termination and Productivity*. Faculty of Sciences, Division of Mathematics and Computer Science, VUA. 2010-03

T. Staijen. *Graph-Based Specification and Verification for Aspect-Oriented Languages*. Faculty of Electrical Engineering, Mathematics & Computer Science, UT. 2010-04

Y. Wang. *Epistemic Modelling and Protocol Dynamics*. Faculty of Science, UvA. 2010-05

J.K. Berendsen. *Abstraction, Prices and Probability in Model Checking Timed Automata*. Faculty of Science, Mathematics and Computer Science, RU. 2010-06

A. Nugroho. *The Effects of UML Modeling on the Quality of Software*. Faculty of Mathematics and Natural Sciences, UL. 2010-07

A. Silva. *Kleene Coalgebra*. Faculty of Science, Mathematics and Computer Science, RU. 2010-08

J.S. de Bruin. *Service-Oriented Discovery of Knowledge - Foundations, Implementations and Applications*. Faculty of Mathematics and Natural Sciences, UL. 2010-09

D. Costa. *Formal Models for Component Connectors*. Faculty of Sciences, Division of Mathematics and Computer Science, VUA. 2010-10

M.M. Jaghoori. *Time at Your Service: Schedulability Analysis of Real-Time and Distributed Services*. Faculty of Mathematics and Natural Sciences, UL. 2010-11

R. Bakhshi. *Gossiping Models: Formal Analysis of Epidemic Protocols*. Faculty of Sciences, Department of Computer Science, VUA. 2011-01

B.J. Arnoldus. *An Illumination of the Template Enigma: Software Code Generation with Templates*. Faculty of Mathematics and Computer Science, TU/e. 2011-02

E. Zambon. *Towards Optimal IT Availability Planning: Methods and Tools*. Faculty of Electrical Engineering, Mathematics & Computer Science, UT. 2011-03

L. Astefanoaei. *An Executable Theory of Multi-Agent Systems Refinement*.

Faculty of Mathematics and Natural Sciences, UL. 2011-04

J. Proença. *Synchronous coordination of distributed components.* Faculty of Mathematics and Natural Sciences, UL. 2011-05

A. Moralı. *IT Architecture-Based Confidentiality Risk Assessment in Networks of Organizations.* Faculty of Electrical Engineering, Mathematics & Computer Science, UT. 2011-06

M. van der Bijl. *On changing models in Model-Based Testing.* Faculty of Electrical Engineering, Mathematics & Computer Science, UT. 2011-07

C. Krause. *Reconfigurable Component Connectors.* Faculty of Mathematics and Natural Sciences, UL. 2011-08

M.E. Andrés. *Quantitative Analysis of Information Leakage in Probabilistic and Nondeterministic Systems.* Faculty of Science, Mathematics and Computer Science, RU. 2011-09

M. Atif. *Formal Modeling and Verification of Distributed Failure Detectors.* Faculty of Mathematics and Computer Science, TU/e. 2011-10

P.J.A. van Tilburg. *From Computability to Executability – A process-theoretic view on automata theory.* Faculty of Mathematics and Computer Science, TU/e. 2011-11

Z. Protic. *Configuration management for models: Generic methods for model comparison and model coevolution.* Faculty of Mathematics and Computer Science, TU/e. 2011-12

S. Georgievska. *Probability and Hiding in Concurrent Processes.* Faculty of Mathematics and Computer Science, TU/e. 2011-13

S. Malakuti. *Event Composition Model: Achieving Naturalness in Runtime Enforcement.* Faculty of Electrical Engineering, Mathematics & Computer Science, UT. 2011-14

M. Raffelsieper. *Cell Libraries and Verification.* Faculty of Mathematics and Computer Science, TU/e. 2011-15

C.P. Tsirogiannis. *Analysis of Flow and Visibility on Triangulated Terrains.* Faculty of Mathematics and Computer Science, TU/e. 2011-16

Y.-J. Moon. *Stochastic Models for Quality of Service of Component Connectors.* Faculty of Mathematics and Natural Sciences, UL. 2011-17

R. Middelkoop. *Capturing and Exploiting Abstract Views of States in OO Verification.* Faculty of Mathematics and Computer Science, TU/e. 2011-18

M.F. van Amstel. *Assessing and Improving the Quality of Model Transformations.* Faculty of Mathematics and Computer Science, TU/e. 2011-19

A.N. Tamalet. *Towards Correct Programs in Practice.* Faculty of Science, Mathematics and Computer Science, RU. 2011-20

H.J.S. Basten. *Ambiguity Detection for Programming Language Grammars.* Faculty of Science, UvA. 2011-21

M. Izadi. *Model Checking of Component Connectors.* Faculty of Mathematics and Natural Sciences, UL. 2011-22

L.C.L. Kats. *Building Blocks for Language Workbenches.* Faculty of Electrical Engineering, Mathematics, and Computer Science, TUD. 2011-23

S. Kemper. *Modelling and Analysis of Real-Time Coordination Patterns.* Faculty of Mathematics and Natural Sciences, UL. 2011-24

J. Wang. *Spiking Neural P Systems.* Faculty of Mathematics and Natural Sciences, UL. 2011-25

A. Khosravi. *Optimal Geometric Data Structures.* Faculty of Mathematics and Computer Science, TU/e. 2012-01

A. Middelkoop. *Inference of Program Properties with Attribute Grammars, Revisited.* Faculty of Science, UU. 2012-02

Z. Hemel. *Methods and Techniques for the Design and Implementation of Domain-Specific Languages.* Faculty of Electrical Engineering, Mathematics, and Computer Science, TUD. 2012-03

T. Dimkov. *Alignment of Organizational Security Policies: Theory and Practice.* Faculty of Electrical Engineering, Mathematics & Computer Science, UT. 2012-04

S. Sedghi. *Towards Provably Secure Efficiently Searchable Encryption.* Faculty of Electrical Engineering, Mathematics & Computer Science, UT. 2012-05

F. Heidarian Dehkordi. *Studies on Verification of Wireless Sensor Networks and Abstraction Learning for System Inference.* Faculty of Science, Mathematics and Computer Science, RU. 2012-06

K. Verbeek. *Algorithms for Cartographic Visualization.* Faculty of Mathematics and Computer Science, TU/e. 2012-07

D.E. Nadales Agut. *A Compositional Interchange Format for Hybrid Systems: Design and Implementation.* Faculty of Mechanical Engineering, TU/e. 2012-08

H. Rahmani. *Analysis of Protein-Protein Interaction Networks by Means of Annotated Graph Mining Algorithms.* Faculty of Mathematics and Natural Sciences, UL. 2012-09

Summary

The main goal of this thesis is mining annotated graphs. We chose Protein-Protein Interaction (PPI) networks as a specific graph domain to apply our methods. We modeled the PPI network as a graph where each node is a protein and each edge is a physical interaction between two proteins. There are different types of annotation information for each protein in the PPI network. "Functional annotation" states the biological functions of proteins in the PPI network and "disease/cancer relatedness annotation" indicates if one protein is involved in disease/cancer or not. We worked on these prediction tasks to improve the annotation information of proteins in the PPI network.

The task of function prediction in the PPI network is trying to predict the functions of un-annotated proteins based on the information in the network. We proposed two approaches for this task. In the first approach, we used shortest-path distances among different proteins as protein description features and Anova (Analysis of variance) as a feature selection method for reducing the noise and dimensionality in the description vectors. Then, we applied machine learning for the prediction task. In the second approach, we introduced novel functional features that indicate so-called "Collaborative Functions": Pairs of functions that frequently interface with each other in different interacting proteins. Most of the previous methods predict the proteins' functions based on guilt-by-association: Interacting proteins tend to have similar functions. We proposed two methods to extract collaborative functions from the PPI network. The first method calculates the collaboration value of two functions based on an iterative reinforcement strategy. The second method adopts an artificial neural network. Empirical results confirmed that the notion of collaborativeness of functions, rather than similarity, is useful for the task of predicting the functions of proteins.

The task of predicting cancer-related proteins in the PPI network is trying to predict the new proteins involved in cancer. We generalized the previous methods as a two-steps algorithm. First, they select some features based on the training data to describe the proteins in the test data. Second, they apply machine learning methods to the description features in order to predict the new cancer-related proteins. Empirical results show that prediction accuracy depends more on the discriminative features rather than the machine learning methods and among different features ap-

plied individually, biological functions seems to be the most discriminative features. We proposed two approaches to select the novel features from the PPI network. In the first approach, we considered functional and structural contexts of proteins in the PPI network using the Anova measure and the chi-square method. In the second approach, we proposed a new method, "Interaction-based Chi-square", to combine the functional annotations of proteins with the information contained in the topology of a PPI network for the feature selection task. Empirical results showed that our proposed feature selection approaches are biologically meaningful and improve the prediction accuracy of these systems.

The task of predicting disease-related proteins in PPI network is an important issue in the area of computational biology. Previous methods assume to have a set of proteins which are previously known to be involved in disease (i.e., seed proteins) and then, they try to extend the seed proteins by predicting new disease-related proteins. While the initial seed proteins of each disease is incomplete and suffers from 'False Negative' cases (i.e., disease-related proteins which are not annotated as being involved in disease), dependency of previous methods on the incomplete seed proteins is the main drawback of these methods. We proposed an informative Human Disease Network (HDN) considering both functional and structural information in the PPI network to reduce the number of False Negative cases in the initial seed proteins of 20 analyzed diseases. Literature mining of newly predicted proteins proved the usefulness of the proposed HDN.

Samenvatting

Dit proefschrift beschrijft onderzoek op het gebied van datamining in geannoteerde grafen. Wij kozen Proteïne-Proteïne-Interactie (PPI) netwerken als domein om onze methodes op toe te passen. Het PPI-netwerk werd gemodelleerd als een graaf waarbij elke knoop een proteïne is en elke tak staat voor een fysieke interactie tussen twee proteïnes. Er zijn verschillende soorten informatie waarmee elk proteïne in het PPI-netwerk geannoteerd is. De "functionele annotatie" geeft de biologische functies van de proteïnes in het PPI-netwerk en de "ziekte/kanker gerelateerde annotatie" geeft aan of een proteïne bij een ziekte of kanker betrokken is. We probeerden deze annotaties te voorspellen met als doel de informatie over proteïnes in het PPI-netwerk te verbeteren.

De taak van het voorspellen van functies in het PPI-netwerk slaat op het voorspellen van functies van niet-geannoteerde proteïnes met behulp van de informatie in het netwerk. Hiervoor worden twee manieren voorgesteld. Bij de eerste manier gebruikten we kortste-pad afstanden tussen verschillende proteïnes als beschrijvende eigenschappen van de proteïnes, en variantie-analyse als selectiemethode om de ruis en de dimensionaliteit van de vectoren te verminderen. Daarna pasten we hier machinaal leren op toe om de eigenschappen voor de niet-geannoteerde proteïnes te voorspellen. Bij de tweede manier introduceerden we nieuwe functionele eigenschappen die de zogenaamde “Samenwerkende Functies” aangeven. Deze samenwerkende functies zijn paren functies die vaak samen voorkomen bij proteïnes waartussen een fysieke interactie bestaat. De meeste van de al bestaande methodes voorspellen de functies van de proteïnes door middel van *guilt-by-association*, waarbij er vanuit wordt gegaan dat proteïnes waartussen een interactie bestaat geneigd zijn om dezelfde functies te hebben. Wij onderzochten twee methodes om de samenwerkende functies uit het PPI-netwerk te halen. De eerste methode berekent de samenwerkingswaarde van twee functies door middel van een iteratieve versterkingsstrategie. De tweede methode gebruikt een kunstmatig neuraal netwerk. Empirisch onderzoek bevestigt dat het concept van samenwerkende functies beter werkt voor het voorspellen van functies van proteïnes, dan het concept dat proteïnes waartussen een interactie bestaat vaak dezelfde functies hebben.

Bij het voorspellen van aan kanker gerelateerde proteïnes in het PPI-netwerk wordt

gezocht naar nieuwe proteïnes die waarschijnlijk betrokken zijn bij het veroorzaken van kanker. We beschouwen reeds bestaande methodes als een tweestaps-algoritme. Bij de eerste stap worden aan de hand van de trainingsdata enkele eigenschappen geselecteerd die de proteïnes in de test data moeten beschrijven. Daarna wordt machinaal leren toegepast op deze eigenschappen om zo te voorspellen welke proteïnes aan kanker gerelateerd zijn. Empirisch onderzoek heeft aangetoond dat de kwaliteit van de voorspelling meer wordt bepaald door de beschrijvende eigenschappen waaruit geleerd wordt, dan door de gebruikte methode van machinaal leren. Als deze verschillende eigenschappen onafhankelijk van elkaar worden bekeken, lijken de biologische functies het beste te werken. Wij bedachten twee manieren om nieuwe eigenschappen te selecteren uit het PPI-netwerk. Bij de eerste manier wordt de functionele en structurele context van proteïnes bekeken met behulp van variantie-analyse en de χ-kwadraat methode. De tweede manier bestaat uit een geheel nieuwe methode, "Interactiegebaseerde chi-kwadraat", die de functionele annotaties van proteïnes combineert met de informatie die genesteld zit in de topologie van een PPI-netwerk, om zo de juiste eigenschappen te selecteren. Empirisch onderzoek heeft aangetoond dat onze manieren om eigenschappen te selecteren een duidelijke biologische betekenis hebben en ervoor zorgen dat het systeem betere voorspellingen doet.

Het voorspellen van aan ziekte gerelateerde proteïnes in het PPI-netwerk is een belangrijk onderzoeksgebied binnen de computationele biologie. Eerdere methodes gaan uit van een verzameling proteïnes waarvan al bekend is dat zij gerelateerd zijn aan deze ziekte (zogenaamde bronproteïnes), en vervolgens wordt geprobeerd deze verzameling uit te breiden door van andere proteïnes te voorspellen of deze ook aan die ziekte gerelateerd zijn. De initiële verzamelingen van bronproteïnes voor een ziekte zijn onvolledig: er zijn zo goed als zeker valse negatieven. Dit heeft een nadelige invloed op de resultaten bekomen met de vermelde methoden. Wij creëerden een nieuw "Human Disease Network" (HDN), een netwerk dat verbanden legt tussen ziekten, met behulp van zowel functionele als structurele informatie van het PPI-netwerk, dit om het aantal valse negatieven in de initiële verzameling bronproteïnes te verminderen voor 20 verschillende ziektes. Met behulp van literatuuronderzoek van nieuw voorspelde proteïnes, is bewezen dat het door ons gecreëerde HDN zeer bruikbaar is.

List Of Publications

1. H. Rahmani, H. Blockeel and A. Bender: Predicting Disease-Related Proteins using Informative Human Disease Network. Submitted To IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB).

2. H. Rahmani, H. Blockeel and A. Bender: Predicting Genes Involved in Human Cancer Using Network Contextual Information. Journal of Integrative Bioinformatics, 9(1):210, 2012.

3. H. Rahmani, H. Blockeel and A. Bender: Collaboration-Based Function Prediction in Protein-Protein Interaction Networks. In: Advances in Intelligent Data Analysis X - 10th Internationl Symposium, Lecture notes in Computer Science 7014: 318-327, Springer (2011)

4. H. Rahmani, H. Blockeel and A. Bender: Interaction-based feature selection for predicting cancer-related proteins in protein-protein interaction networks. In: Proceedings Fifth International Workshop on Machine Learning in System Biology (2011)

5. H. Blockeel, B. Piccart, H. Rahmani and D. Fierens: Three complementary approaches to context aware movie recommendation. In: Proceedings of the Workshop on Context-Aware Movie Recommendation: 57-60, ACM (2010)

6. H. Blockeel, H. Rahmani and M. Witsenburg: On the importance of similarity measures for planning to learn. In: P. Brazdil, A. Bernstein and J. Kietz (Eds.), 19th European Conference on Artificial Intelligence, 3rd Planning to Learn workshop (PlanLearn-2010): 69-74 (2010)

7. H. Rahmani, H. Blockeel and A. Bender: Collaboration based function prediction in protein-protein interaction networks. In: Proceedings of the 7th International Symposium on Networks in Bioinformatics (2010)

8. H. Rahmani, H. Blockeel and A. Bender: Collaboration-based function prediction in protein-protein interaction networks. In: Machine Learning in Systems Biology: Proceedings of the Fourth International Workshop: 55-58 (2010)

9. H. Rahmani, H. Blockeel and A. Bender: Collaboration-based function prediction in protein-protein interaction networks. In: European Conference on Computational Biology (2010)

10. H. Rahmani, H. Blockeel and A. Bender: Predicting the functions of proteins in protein-protein interaction networks from global information. In: JMLR: Workshop and Conference Proceedings: International Workshop on Machine Learning in Systems Biology 8: 82-97 (2010)

11. H. Rahmani, B. Nobakht and H. Blockeel: Collaboration-based social tag prediction in the graph of annotated web pages DyNaK 2010: Dynamic Networks and Knowledge Discovery. In: DyNaK 2010: Dynamic Networks and Knowledge Discovery: 1-12 (2010)

12. S. Aliakbary, H. Abolhassani, H. Rahmani and B. Nobakht: Web Page Classification Using Social Tags. In: IEEE 2009 International Conference on Computational Science and Engineering: 588-593 (2009)

13. H. Rahmani, H. Blockeel and A. Bender: Predicting the functions of proteins in PPI networks from global information. In: Proceedings of the Third International Workshop on Machine Learning in Systems Biology: 85-94, Helsinki University Printing House (2009)

Curriculum Vitae

Hossein Rahmani was born in Tehran, Iran, on March 25, 1983. He studied software engineering at the University of Tehran, Faculty of Engineering. His bachelor thesis was *Model Driven Architecture (MDA)*. In 2005, he was accepted in Artificial Intelligence department of Sharif University and in his master thesis: *Semantic web service composition using AI planning methods*, he succeeded to use AI Planning techniques to solve the problem of web service composition. The results of his master thesis encouraged him to explore the multidisciplinary research activities. Since November 2008, he started his PhD at the Leiden Institute of Advanced Computer Science (LIACS), Leiden University, in the section Algorithms and Data Mining under the supervision of Dr. Hendrik Blockeel. His PhD project was funded by Dutch Science Foundation (NWO) through a VIDI grant. The title of his PhD thesis was *mining annotated graphs* and he chose Protein-Protein Interaction (PPI) networks to apply his proposed graph algorithms. This PhD thesis is the result of four years of research in the Netherlands.

Acknowledgments

Thank you all. This, at the highest level of abstraction, shows my thanks and gratitude to all kind people around me who support, inspire and help me to accomplish my PhD. I will try to refine it a bit here, but I am sure this refinement will miss some traces. My apologies in advance.

I would like to thank my colleagues at the Leiden Institute of Advanced Computer Science (LIACS). Tijn Witsenburg, I really enjoyed sharing an office with you in the last three years. Rudy van Vliet, it was a pleasure to have you in our new office. I would like also to thank my friends from LIACS Data Mining Group.

I am very grateful to Mohammad Mahdi Jaghoori who helped and supported me in my beginning days in Netherlands. I would like to thank Behrooz Nobakht, my best friend in ten years. I owe a debt of gratitude to Mohammad Mirnezhad, who made my life in Leiden very easy, so far away from my family, Mohammad was the person who felt always responsible to provide a home-like atmosphere for me. I spent most of my days at Netherlands with different people. As the smallest sign of appreciation, I will name them: Ramin Etemaadi, Nafiseh Nazemi, Soghra Akbari, Mohammad Tayyebi, Zeinab Neshati, Faezeh Nami, Omid Karami, Masoud Mosallanejad, Shadab Gharaati, Moeen Mosleh, Akbar Shafiee and Mahdi Lamee.

I ow my deepest gratitude to my family. Shayesteh, your love and kindness is the best thing a man could ever want. I thank my parents, Aziz and Masoumeh for bringing me into this world and supporting me even if they knew I may not succeed. I would like to show my gratitude to my brothers and sisters. Thanks for the good care of parents, sending all those nice books and delicious food from far far away. I will be very grateful to you for all my life.

Lastly, I offer my regards and blessings to all of those who supported me in any respect during the completion of the thesis.

Hossein Rahmani
April 2012