HS-IKI-EA-05-201

Using semantic similarity measures across Gene Ontology to predict protein-protein interactions

Hanna Sigrún Helgadóttir

c02hanhe@student.his.se

Examensarbete i bioinformatik, C-nivå, 20 poäng

2005

HÖGSKOLAN SKÖVDE

Institutionen för kommunikation och information

Examensrapport inlämnad av Hanna Sigrún Helgadóttir till Högskolan i Skövde, för Kandidatexamen (B.Sc.) vid Institutionen för kommunikation och information.

Using semantic similarity measures across Gene Ontology to predict protein-protein interactions

Submitted by Hanna Sigrún Helgadóttir to Högskolan Skövde as a dissertation for the degree of B.Sc., in the School of Humanities and Informatics.

2005-06-07

I certify that all material in this dissertation which is not my own work has been identified and that no material is included for which a degree has previously been conferred on me.

Signed: ____

Using semantic similarity measures across Gene Ontology to predict protein-protein interactions

Hanna Sigrún Helgadóttir (c02hanhe@student.his.se)

Abstract

Living cells are controlled by proteins and genes that interact through complex molecular pathways to achieve a specific function. Therefore, determination of protein-protein interaction is fundamental for the understanding of the cell's lifecycle and functions. The function of a protein is also largely determined by its interactions with other proteins. The amount of protein-protein interaction data available has multiplied by the emergence of large-scale technologies for detecting them, but the drawback of such measures is the relatively high amount of noise present in the data. It is time consuming to experimentally determine protein-protein interactions and therefore the aim of this project is to create a computational method that predicts interactions with high sensitivity and specificity. Semantic similarity measures were applied across the Gene Ontology terms assigned to proteins in *S. cerevisiae* to predict protein-protein interactions. Three semantic similarity measures were tested to see which one performs best in predicting such interactions. Based on the results, a method that predicts function of proteins in connection with connectivity was devised. The results show that semantic similarity is a useful measure for predicting protein-protein interactions.

Keywords: Semantic similarity, Gene Ontology, Protein-protein interactions, Protein function.

Acknowledgements

First of all I would like to thank my supervisor, Zelmina Lubovac, for her interest, her comments, ideas and valuable guidance along the way. I would also like to thank my examiner, Björn Olsson, for his comments and interesting ideas for this work.

Table of Contents

1 Introduction

In traditional molecular biology, various problems have been approached by studying the function of individual genes and gene products. These studies have proven to be very successful and resulted in the discoveries of many biological principles. Despite this, many questions remain unanswered. This is mainly because several gene products usually work together to achieve a specific function in biological processes (Ge *et al.*, 2003). Marcotte *et al.* (1999) state that the lives of biological cells are controlled by proteins or genes that interact through complex molecular pathways to achieve a specific function. These functions include, among other, metabolism, gene expression regulation and signal transduction. Due to this fact, cell biology has evolved from the science that mainly focused on assigning functions to individual molecules, to a science that tries to cope with a set of molecules that act together to form functional modules. Determination of protein-protein interactions is therefore fundamental to our understanding of the lifecycle, metabolism and regulatory aspects of a cell. Also, the function of a protein is defined to a great extent by its interactions with other proteins (Winters and Day, 2003) and can be viewed as the protein's position within the cellular interaction network (Salwinski and Eisenberg, 2003).

A number of different biological networks have been derived, including protein interaction networks (PINs) that represent experimentally determined interactions between proteins in various organisms. One of the driving forces of post-genomic biology is the study of the networks of protein-protein interactions that control the lives of cells and organisms. These networks have been constructed by detecting pairwise interactions between proteins (Deane *et al.*, 2002). The majority of the knowledge of protein-protein interactions has been collected form genetic and biochemical experiments (Marcotte *et al.*, 1999). In this work the dataset that is used is downloaded from the Database of Interacting Proteins (DIP), and contains experimentally determined protein-protein interactions in yeast (*Saccharomyces cerevisiae*). Most of these interactions were identified with high-throughput yeast two-hybrid screens. The main problem with such high-throughput methods is the occurrence of spurious interactions in the output data. Both false positive and false negative interactions occur. Possible causes of false interactions from two-hybrid screens are self-activators and weak, non-specific interactions (Bader *et al.*, 2004). Even though automatic methods for determining protein-protein interactions have been developed, it can be time consuming to experimentally determine protein-protein interactions. Therefore scientists have been putting effort into investigating whether interactions can be predicted by computational means (see Marcotte *et al.*, 1999 for an example).

The aim of this project is to investigate whether semantic similarity can be used to predict interactions between proteins. If it can be shown that semantic similarity has good predictive power, the functionality of new proteins may be predicted and false annotations can be revealed. Proteins interact with each other to achieve some common purpose and therefore it is possible to draw some conclusions regarding the function of a certain protein based on the functions of its interacting partners (Deng *et al.*, 2002). To complete the project, annotation-based measures will be used, by relating protein-protein interactions to functional annotation in Gene Ontology (GO) (The Gene Ontology Consortium, 2001). Lord *et al.* (2003a) have shown that high similarity at sequence level involves high similarity in GO annotations, i.e. high semantic similarity and therefore it would be interesting to examine if semantic similarity can be used to predict protein-protein interactions and protein functions.

This dissertation is organized as follows. In chapter 2, the theoretical background of the study is described. Related work is presented in chapter 3 and in chapter 4 the problem of the study is introduced and the aims and objectives of the project are presented. Chapter 5 contains a detailed description of the approach used in the project. In chapter 6 the results of the study are presented and analysed. Chapter 7 includes a discussion of the results and in chapter 8 conclusions that were drawn from the results are presented. Finally, chapter 9 proposes directions for future work based on the findings of this project.

2 Background

This chapter describes the theoretical background of the project. Chapter 2.1 discusses the Database of Interacting Proteins and the dataset that is used in this project. Section 2.2 describes the yeast two-hybrid screening technique that has been used to experimentally determine most of the interactions in the dataset. In chapter 2.3, Gene Ontology is discussed and finally chapter 2.4 discusses semantic similarity and the semantic similarity measures that are used in the project.

2.1 The Database of Interacting Proteins (DIP)

The Database of Interacting Proteins (DIP) is a database that contains experimentally determined protein-protein interactions. The DIP aims to provide the scientific community with a single, easily accessible database containing information about protein interactions and interaction networks in biological processes (Xenarios *et al.*, 2000).

The database consists of three key tables, namely the PROTEIN, SOURCE and EVIDENCE tables (Salwinski *et al.* 2004). The PROTEIN table contains protein information, i.e. protein identification codes from SWISS-PROT, PIR and GenBank, the gene name, description, enzyme code and cellular localization, when these aspects are known (Xenarios *et al.*, 2000). The SOURCE table contains sources of the experimental information and the EVIDENCE table contains information on individual experiments. Information on protein-protein interactions is stored in two tables, INTERACTION and INT_PRT. The INTERACTION table contains an identifier for the interaction and the INT_PRT table contains both an identifier for the interaction and for the proteins that are involved in the interaction. This structure allows the description of pairwise interactions as well as more complex interactions involving several proteins. For pairwise interactions, two entries are made in the INT_PRT table for each INTERACTION entry and when more than two proteins are involved, more than two entries are made in the INT_PRT table. A table called METHOD is linked to the EVIDENCE table and provides a list of controlled vocabulary terms and a reference to the corresponding PSI ontology entries. This is used to annotate the experiments that have been used to identify the protein-protein interactions. The details of the topology of a molecular complex that was inferred from experiment is also stored when available. This information is kept in two tables, the LOCATION table states the ranges of amino acids that take part in the interaction and the TOPOLOGY table pairs these regions into records describing the observed binary interactions (Salwinski *et al*., 2004).

The DIP is not only useful for storing interactions, but also for studying the properties of protein interaction networks, benchmarking predictions of protein-protein interactions, studying the evolution of such interactions and for understanding the function of proteins and protein-protein relationships (Xenarios *et al*., 2000).

Since the reliability of experimentally determined interactions varies, methods for assessing the quality of the data have been developed and used to identify CORE sets, i.e. the most reliable subset of interactions. These CORE sets can for instance be utilized to evaluate the reliability of high-throughput protein-protein interaction datasets, to study protein interaction networks and for development of prediction methods (Salwinski *et al.,* 2004).

2.1.1 The DIP-YEAST dataset and the CORE subset

The DIP-YEAST dataset contains experimentally determined protein-protein interactions in *S. cerevisiae*. The majority of these interactions have been determined using high-throughput yeast two-hybrid screens (Ito *et al*., 2001). Since such high-throughput measures can generate a large amount of data in a relatively short time, the datasets are generally too large to verify each interaction individually by using older methods for detecting protein-protein interactions (Deane *et al*., 2002). Deane *et al*. (2002) have assessed the complete DIP-YEAST dataset, containing 8063 protein-protein interactions that are described in DIP since November 2001. Two forms of computational assessment were used in their study, expression profile reliability (EPR) and the paralogous verification method (PVM). They estimate that about 50% of the protein-protein interactions in the DIP-YEAST dataset are reliable and were able to confidently identify 3003 of these interactions with the help of PVM.

As a result of the assessment by Deane *et al*. (2002), a subset of DIP-YEAST has been created, containing these 3003 interactions. This set is referred to as CORE and all of the interactions within that set are assumed to be correct. This CORE dataset will be used in this project.

2.2 Yeast two-hybrid (Y2H) screens

Many protein-protein interactions that are present in protein interaction networks have been experimentally determined using high-throughput methods such as two-hybrid screens. For instance, the majority of the interactions present in the DIP-YEAST dataset were identified using yeast two-hybrid screens (Ito *et al*., 2001).

Two-hybrid screening is based on the character of eukaryotic RNA polymerase II transcriptional activator proteins. Such proteins consist of at least two domains, a DNA binding domain (DBD) that specifically binds to promoters of specific genes, and an activation domain (AD) that acts as a site of recruitment for several protein complexes that result in the transcription of the gene (Reece, 2004). Expression of just one of these domains, AD or DBD, in a cell does not result in transcriptional activation. The screening is performed by fusing a protein, called the bait, to a DNA binding domain and another protein, called the prey, to an activation domain. The bait is then used to attract a potential prey. If the two proteins interact, the two domains will join to form a transcriptional activator, leading to the expression of a reporter gene placed downstream of the promoter (Reece, 2004).

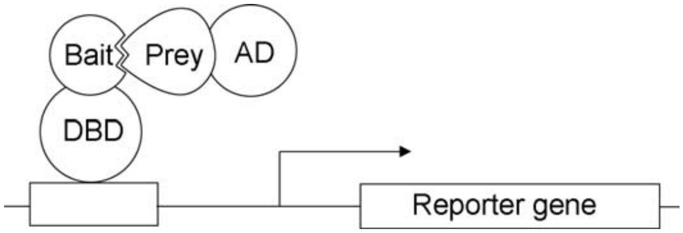


Figure 2.1: Yeast two-hybrid screen. When both of the hybrid proteins, a bait protein fused to a DNA binding domain and a prey protein fused to an activation domain, are expressed and the bait and the prey are able to interact with each other, a functional activator protein is formed. This results in the expression of the reporter gene (Redrawn from Reece (2004)).

Although this method is frequently used to detect protein-protein interactions, it suffers from several deficiencies that result in noise in the output data. False positive interactions can occur for several reasons, e.g. if the prey is a DNA binding protein itself then it does not require the recruitment of a bait protein fused to a DBD in order to initiate the expression of the reporter gene. False negative interactions can also occur, for instance because some physiologically relevant protein-protein interactions are weak and therefore unable to recruit other proteins needed to start the expression (Reece, 2004).

2.3 Gene Ontology (GO)

The amount of biological information has grown dramatically in the last years and therefore it has become ever more important to explain and classify biological objects in meaningful ways. Ontologies present a mechanism for capturing a community's view of a domain in a form that is both accessible by humans as well as computers. According to Lord *et al.* (2003b), one of the most important ontologies within the bioinformatics community is Gene Ontology (GO) (The Gene Ontology Consortium, 2001). GO provides a set of cross-species biological vocabularies to describe genes and gene products in a consistent way. The ontology is divided into three aspects that were chosen because they symbolize information sets that are universal and fundamental to the annotation of information about genes and gene products (The Gene Ontology Consortium, 2001). The first of those three aspects is molecular function. It describes what the gene product does at a biochemical level without specifying when or how the event occurs. The second is biological process, which states the biological objective that the gene product helps to fulfil. The last of these aspects is cellular component. It refers to the location in the cell where the gene product is active. The cellular component of a gene product is not necessarily a place in the general sense of the word. It can also be a term that describes complexes where several proteins are found. The GO aspects are ordered vocabularies in the form of directed acyclic graphs (DAG's) that represent a network where each term may be a "child" of one or more "parent". The relationship between a child and a parent may be of one of two types, namely the "is a" or "part of" relationships. The "is a" relationship applies when a child is an instance of the parent", and the "part of" when the child is a component of the parent (The Gene Ontology Consortium, 2001).

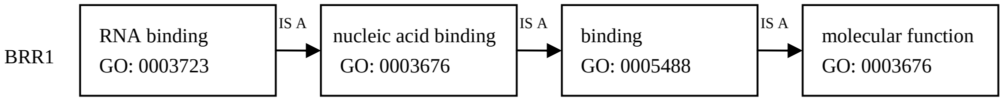


Figure 2.2: GO molecular function subgraph for the protein BRR1.

The three different aspects of GO are gathered under a top-level term, "Gene Ontology" (GO:0003673), however they are entirely orthogonal, being disconnected subgraphs (Lord *et al.*, 2003b).

2.4 Semantic similarity measures

Gene Ontology has provided a standard vocabulary to annotate entries within several biological databases such as SWISS-PROT and InterPro. This fact should allow querying across these databases, for instance by asking for proteins that are semantically similar to the input protein (Lord *et al*., 2003b).

In biology, semantic similarity usually refers to similarity between two concepts in ontology (Lin, 1998). Several different measures for calculating the semantic similarity between terms in ontology have been proposed. A natural way to calculate semantic similarity is to evaluate the distance between the nodes that correspond to the items that are being compared, i.e. the shorter the path between those nodes the more semantically similar are the terms. Measures based on that assumption rely on the concept that links in the ontology represent unvarying distance. Unvarying distance is very difficult to define and control, especially in biological categories where certain sub-ontologies are denser than other ones (Resnik, 1999).

A number of more recent measures for calculating semantic similarity are based on information content, and use the assumption that the less frequently used terms are generally more informative (Lord *et al*., 2003b). In this study three different measures of semantic similarity are applied across Gene Ontology. All of those measures are based on information content of the shared parents of the terms that are being compared. To calculate semantic similarity using information content based measures, the probability of each term assigned to the gene product must be calculated using maximum likelihood estimation. This is done by counting the number of times each term or its children occur in annotations for every protein in the dataset and dividing that number by the total number of GO term annotations in the dataset (Lord *et al*., 2003b). The equation for calculating the maximum likelihood estimate is:

$$P(t) = \frac{freq(t)}{N}$$

where N is the total number of GO terms in the dataset and freq(t) is the number of times term t or any child term of t occur in the dataset (Speer *et al*., 2004).

Because GO allows multiple parents for each term, it is possible that the terms that are being compared share parents by more than one path. Then the average term to term similarity is used, since proteins can be annotated with more than one term and the interest lies in the semantic similarity between the proteins and not the GO terms as such (Lord *et al.*, 2003a). For calculating the semantic similarity between two terms, the parent term with the lowest probability shared by those terms has to be identified. This is called the probability of the minimum subsumer, p_{ms}, and is calculated by the following equation (Lord *et al.*, 2003b):

$$p_{ms}(t_1, t_2) = \min_{t \in S(t_1, t_2)} \{p(t)\}$$

where S(t_1, t_2) is the set of terms that include both t_1 and t_2 (Resnik, 1999).

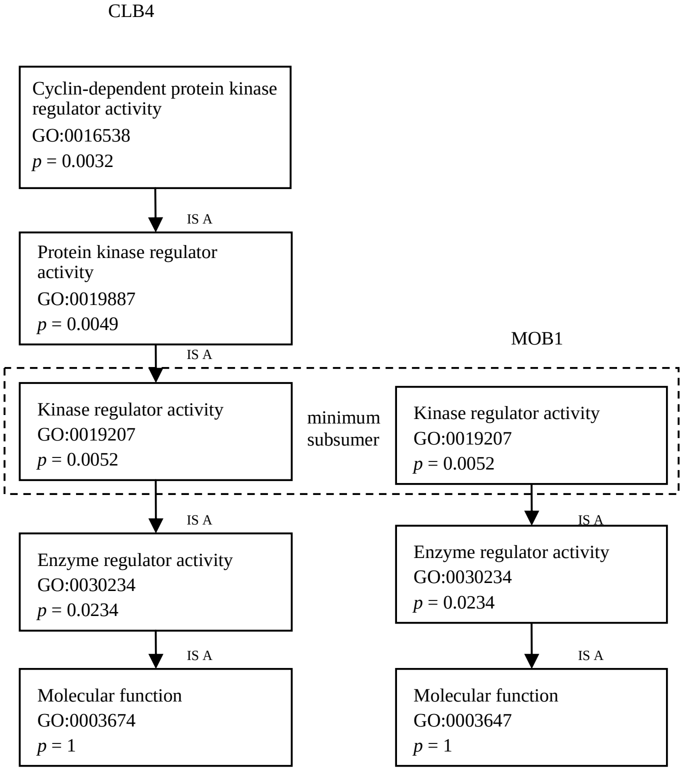


Figure 2.3: GO molecular function subgraphs including term probabilities for proteins CLB4 and MOB1. The dashed box indicates the minimum subsumer of the proteins. Subgraphs are generated with SGD Gene Ontology Term Finder[1].

In the following subchapters the three different semantic similarity measures that are used in this project are described. Those three measures were chosen since they are well documented and have been adapted to be used with Gene Ontology (Lord *et al.*, 2003b). Lord *et al.* (2003b) showed that all three measures show a strong correlation between sequence and semantic similarity. In chapter 2.3.1 the semantic similarity

[1] http://db.yeastgenome.org/cgi-bin/GO/goTermFinder

measure by Lin (1998) is discussed, the one by Resnik (1999) is described in chapter 2.3.2 and the measure by Jiang and Conrath (1998) in chapter 2.3.3.

2.4.1 Lin

The semantic similarity measure by Lin (1998) is both based on the information content of the parent terms and of the terms that are being compared. Since $p_{ms} \geq p(t_1)$ and $p_{ms} \geq p(t_2)$, the value varies between one (for identical terms) and zero (Lord *et al.*, 2003b). The semantic similarity by Lin (1998) is calculated using the following equation (Lord *et al.*, 2003b):

$$sim(t_1, t_2) = \frac{2 \times [\ln p_{ms}(t_1, t_2)]}{\ln p(t_1) + \ln p(t_2)}$$

In the equation, $p(t_i)$ is the probability of term t_i and $p_{ms}(t_1,t_2)$ is the probability of the minimum subsumer for terms t_1 and t_2.

2.4.2 Resnik

The measure by Resnik (1999) only uses the information content of the shared parents. As the value for p_{ms} varies between zero and one, this measure generates values between infinity (for similar terms) and zero. In practice, for terms that are present in the corpus, the maximum value is defined as ln(*t*), where *t* stands for the number of occurrences of any term in the corpus (Lord *et al*., 2003b).

Using the semantic similarity measure by Resnik (1999), semantic similarity can be calculated according to the equation below (Lord *et al.*, 2003b):

$$sim(t_1, t_2) = -\ln p_{ms}(t_1, t_2)$$

2.4.3 Jiang and Conrath

The measure by Jiang and Conrath (1998) is actually a semantic distance measure rather than a semantic similarity measure. It is quite similar to the semantic similarity measure by Lin, i.e. based on both the information content of the parent terms and of the terms that are being compared, but uses the terms in a different order. In theory this measure can, as the measure by Resnik, give arbitrary large values, but in practice the maximum value is 2 ln(*t*) where *t*, as before, stands for the number of occurrences of any term in the corpus (Lord *et al*., 2003b).

The equation for calculating the semantic distance between two terms according to Jiang and Conrath (1998) is as follows (Lord *et al.*, 2003b):

$$dist(t_1, t_2) = -2 \ln p_{ms}(t_1, t_2) - (\ln p(t_1) + \ln p(t_2))$$

3 Related work

This chapter contains a description of some work that is related to this project. In chapter 3.1, computational methods that have been used for predicting protein-protein interactions and protein function are discussed briefly and chapter 3.2 contains a concise summary of the comparison by Lord *et al.* (2003b) of the three different semantic similarity measures that are used in this project.

3.1 Prediction of protein-protein interactions and protein function

Most computational approaches are both cheaper and faster than experimental analysis for prediction of protein-protein interaction and protein function predictions (Franzot and Carugo, 2003). Several approaches for such predictions have been developed and in their paper, Franzot and Carugo (2003) state that they can be categorized into four different classes. 1) methods based on genomic information; 2) methods based on evolutionary relationship; 3) *ab inito* methods that only require a single protein sequence for the prediction; 4) methods that require three-dimensional information.

An example of a method that falls into the first category is domain fusion analysis by Marcotte *et al.* (1999). This computational method for predicting protein interactions from genome sequences is based on the findings that some pairs of interacting proteins have homologs in another organism fused to a single protein chain. The fusion of two protein domains into a single protein chain can greatly enhance the affinity of one domain to the other. This suggests that some pairs of interacting proteins may have evolved from proteins that included both of the domains on a single polypeptide chain (Marcotte *et al.*, 1999). The genome of *Escherichia coli* (*E. coli*) was searched for pairs of such non-homologous sequences where both members of the pair had significant similarity to a single protein sequence in the genome of another organism, termed a Rosetta Stone sequence. Three independent tests were made on the interactions predicted by the domain fusion analysis and the results show that a reasonable part of them may in fact interact (Marcotte *et al.* 1999).

Marcotte *et al.* (1999) have also devised a method for predicting protein function. It groups proteins by a combination of three independent prediction methods, namely correlated evolution, correlated mRNA expression patterns and patterns of domain fusion, to determine functional relationships among proteins in *S. cerevisiae*. These three methods allowed them to create several thousands of links between proteins of related function, thus providing a means to characterize proteins of unknown function. Using this combined method, 62% of proteins of then unknown function in yeast could be assigned a general function (Marcotte *et al.*, 1999).

Vinayagam *et al.* (2004) devised a method for Gene Ontology based protein function prediction using Support Vector Machines (SVM). They developed an automated system that on a large scale assigns molecular function GO terms to uncharacterized cDNA sequences and labels each prediction with a confidence value. This method takes care of several problems that have been associated with other approaches that have been used to functionally annotate sequences on a large scale. These problems include that some approaches can only be applied to a specific dataset, while other methods output non-formalized results or they do not calculate a confidence estimate for their predictions (Vinayagam *et al.*, 2004). For training the SVMs, information from GO-annotated protein sequences was used and to enhance the reliability of the predictions, several SVMs were used and their results combined using a voting scheme. The system was cross-validated using a large dataset from many different organisms and showed an average precision of 80% for 74% of the protein sequences and that the prediction performance was organism-independent (Vinayagam *et al.*, 2004).

3.2 Comparison of semantic similarity measures

Lord *et al.* (2003a) adapted the semantic similarity measure by Lin (1998) to be used with the Gene Ontology. They tested the measure by analyzing semantic similarity and plotting it against sequence similarity, proving their hypothesis that sequences that show high sequence similarity should be highly semantically similar (Lord *et al.*, 2003a). The correlation between sequence similarity and semantic similarity proved to be highest when the molecular function aspect of GO is considered. The other two aspects, biological process and cellular component, also showed some correlation, especially at higher levels of sequence and semantic similarities (Lord *et al.*, 2003a). This is to be expected since the sequence of a protein largely determines its molecular function, but not the cellular location or the biological process in which it is involved (Lord *et al.*, 2003b).

In another paper by Lord *et al.* (2003b), they extended their study to the semantic similarity measures by Resnik (1999) and Jiang and Conrath (1998). They related those measures to sequence similarity and compared the results from all three measures, investigating the different aspects of GO. For all three measures, semantic similarity is most strongly correlated with the molecular function aspect of GO, then the cellular component aspect and least with the biological process aspect. Of the three semantic similarity measures, the one by Resnik (1999) showed the strongest correlation between sequence and the molecular function aspect of GO, but it also showed the weakest correlation against the biological process aspect, suggesting that this measure shows more discrimination towards the different aspects of GO than the other two measures. When only looking at the molecular function aspect of GO, the measure by Resnik (1999) shows the strongest correlation with sequence similarity, followed by the measure by Lin (1998) and at last the one by Jiang and Conrath (1998) (Lord *et al.*, 2003b).

4 Problem description

This chapter presents the aims and objectives of the study. Chapter 4.1 contains a definition of the problem and chapter 4.2 declares the aim of the project. Chapter 4.3 describes the objectives that have to be met in order to reach the aim.

4.1 Problem definition

Interactions between proteins affect all processes in a living cell and proteins can rarely carry out their functions in isolation. It has been proposed that all proteins in a given cell are connected through an extensive network, where interactions are continuously forming and dissociating (Uetz and Vollert, 2005).

The focus of this project is on pairwise protein-protein interactions and at the aim is to investigate whether semantic similarity can be used to predict such interactions. By doing that, functionality of new proteins may be predicted and false annotations can be revealed. To complete the project annotation based measures will be used, by relating protein-protein interactions to functional annotation in Gene Ontology (The Gene Ontology Consortium, 2001). Lord *et al.* (2003) have shown that high similarity at sequence level involves high similarity in GO annotations, i.e. high semantic similarity. Therefore it would be interesting to examine whether semantic similarity can be used to predict protein-protein interaction and protein function.

It is important to investigate the chosen problem since the interactions in protein interaction networks are derived on such a large scale, using relatively error-prone methods, so that much noise is often present in the form of false positive and false negative interactions. The amino acid sequence of proteins partially determines the structure of the proteins, and the protein structure is closely related to the protein function (Sternberg, 1996). Therefore it is worthwhile investigating whether semantic similarity can be used to predict interactions between gene products. As said before, the interacting partners of a protein can give clues regarding the function of the protein (Deng *et al.*, 2002). Consequently semantic similarity measures may be used to assign function to unknown proteins and to reveal incorrect annotations in GO. By doing this, the accuracy of interactions in current protein interaction networks can be improved.

4.2 Aim

The aim of this project is to investigate whether semantic similarity measures based on Gene Ontology (GO) can be used to predict protein-protein interactions.

Three different measures for calculating semantic similarity will be evaluated in order to determine which one is able to reconstruct the protein interaction network with respect to sensitivity and specificity.

4.3 Objectives

In order to achieve the aim of the project, each of the following objectives need to be fulfilled.

4.3.1 Identify measures for calculating semantic similarity

The first step in the project is to identify measures for calculating semantic similarity based on Gene Ontology. Several measures exist, but measures based on information content will be selected. This is because some of the older measures suffer from several drawbacks and they are not as accurate as the newer measures (Resnik, 1999).

This step also involves determining which aspects of GO will be used.

4.3.2 Identify information sources needed

Step two is to identify appropriate information sources that are needed in order to complete the project. A dataset containing validated protein-protein interactions will be required so that the achieved results can be compared to it. In this project, a dataset containing a protein interaction network from yeast will be used. It is chosen since *S. cerevisiae* is one of the most studied organisms in the world (Franzot and Carugo, 2003) and the CORE dataset contains interactions that are assumed to be correct, i.e. it is assumed that the dataset contains no false positive interactions and can thus be used for the comparison (Deane *et al*., 2002). The network will be downloaded from the Database of Interacting Proteins, or DIP (Xenarios *et al*., 2003).

4.3.3 Apply the evaluation method to the data

This objective involves applying the semantic similarity measures, chosen in step one (see chapter 4.3.1) on the proteins in the dataset downloaded in step two (see chapter 4.3.2). This will result in the creation of a pairwise matrix consisting of zeros (0) and ones (1), based on the semantic similarities. In the matrix, 1 stands for interaction between the proteins and 0 means that no interaction takes place. The threshold value of minimum semantic similarity required between a pair of proteins to get the value of 1 in the matrix will be varied with respect to sensitivity and specificity in an effort to retrieve the optimal interaction matrix. The semantic similarity measures by Lin (1998), Resnik (1999) and by Jiang and Conrath (1998) will be used in this project and their predictive power compared to see if the specificity and sensitivity of the derived interactions can be improved by using a particular semantic similarity measure.

4.3.4 Evaluate the results

The final step is to evaluate the results, by comparing the results, represented in the matrix, to the original protein interactions retrieved from DIP. The CORE DIP-YEAST dataset is assumed to be correct, i.e. it contains no false positive interactions. In other words, the matrix based on semantic similarity between GO terms will be compared to the original dataset form DIP.

The results that will be obtained by using the different semantic similarity measures will also be evaluated and compared to determine which of the measures is best capable of reconstructing the interactions in the dataset.

Results that deviate from the interactions in the original protein interaction network can undergo further analysis to determine the cause of these false interactions. Possible causes can for instance be false annotations, one of the proteins is annotated with “unknown function” or the interaction may be a false positive or negative in the original PIN. Some proteins may interact within a certain biological process without having a similar function. Since only the molecular function aspect of GO is considered in the project, some false negative interactions may be obtained because of this reason. To identify such cases, the method can be tested on the biological process aspect of GO as well to see if any of this false positive interactions become true positive, i.e. are included in the prediction.

It is also possible to combine different aspects of GO to see if the specificity and sensitivity of the predictions can be improved. All three aspects can be combined or only two at a time.

The method can also be tested on predicted functional modules that have high average hub-to-neighbour and neighbour-to-neighbour semantic similarity. Hub-to-neighbour similarity refers to the semantic similarity between the hub protein, i.e. the highly connected node that holds the module together, and its neighbours. Neighbour-to-neighbour similarity refers to the semantic similarity between the neighbours of the hub (Lubovac *et al.*, 2005). Those results can then be compared to results obtained from proteins that are known to interact but are not part of such modules. It is expected that semantic similarity based on the molecular function aspect of GO is probably more likely to accurately predict the interactions between proteins in the modules than the interactions between proteins that are not included in a module.

If semantic similarity between GO terms proves to be suitable to predict protein-protein interactions, functions of proteins that are annotated with “unknown function” can be predicted to be similar to the functions of their interacting partners, if any.

5 Method

In this chapter, the approach taken in this project is described. Chapter 5.1 contains an overview of the approach. Chapter 5.2 discusses the identification of semantic similarity measures and GO aspects to use in the project. Chapter 5.3 presents the determination of subsets of the CORE dataset to test the method on. Chapter 5.4 contains a description of the application and evaluation of the method on the data and finally, chapter 5.5 describes the approach taken to predict the function for proteins with unknown function.

5.1 An overview of the project

An overview of the steps taken to complete the project is shown in figure 5.1 below.

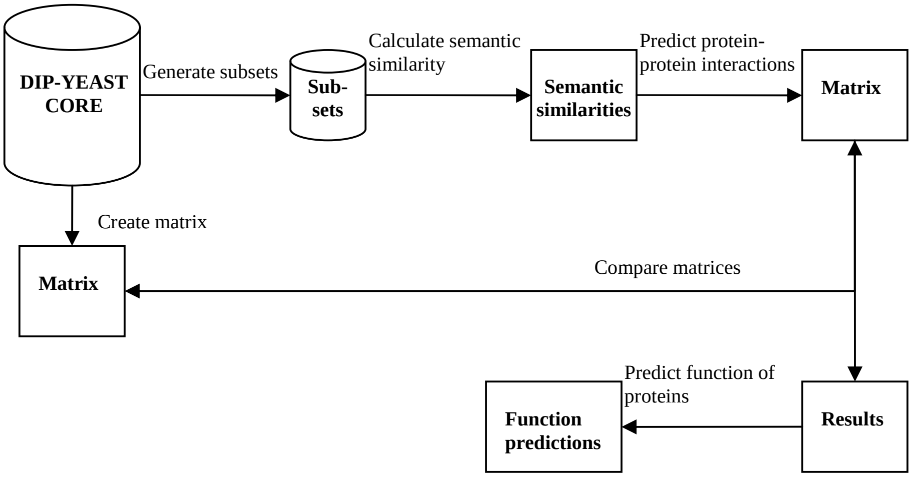


Figure 5.1: A brief overview of the project. First, smaller datasets of the DIP-YEAST CORE dataset are created and then semantic similarity between all of the proteins in each datasets is calculated. From those results a pairwise matrix consisting of 0s and 1s, where 1 stands for interaction and 0 for no interaction, is created. A similar matrix is created from the original protein interaction network and used for comparison. Function predictions for proteins of unknown function are then made, partly based on the results from the protein-protein interaction predictions.

5.2 Identification of semantic similarity measures and GO aspects

For calculating the semantic similarity between GO terms, three measures were chosen. These are the measures by Lin (1998), Resnik (1999) and Jiang and Conrath (1998). The main reason for choosing these measures is that they are well documented and, as stated in chapter 3.2, have all been shown to show correlation between sequence similarity and molecular function semantic similarity (Lord *et al.* 1999b).

Three measures are chosen since it is interesting to see whether one measure is able to predict protein-protein interactions with higher specificity and sensitivity than the other.

In the present study, only the molecular function aspect of GO is considered. Lord *et al.* (2003b) showed that sequence similarity is most strongly correlated with semantic similarity based on the molecular function aspect, even though correlation with the other two aspects exists.

5.3 Identification of datasets and other information sources

In this study, the *Saccharomyces* Genome Database (SGD[2]) is used. It contains GO annotations for all three aspects of GO, but in this study only annotations belonging to the molecular function aspect are considered.

Also, the DIP-YEAST CORE dataset is used. All of the interactions in the dataset are assumed to be correct and will therefore be used as a comparison to the results obtained when using semantic similarity as a prediction method.

Four different subsets of proteins were created from the CORE set. The results from protein-protein interaction predictions for each of the sets were then compared to see if there is any difference in the predictive power of the method when different sets of proteins with different characteristics are used. The reason for not applying the method on the whole CORE dataset is that computational power to calculate the semantic similarity between all proteins in the dataset was insufficient.

The first of the four subsets is called CORE-CDC28 and it contains the protein CDC28 and all of its neighbours, 96 proteins in total. CDC28 is one of five different cyclin-dependent protein kinases (CDKs) in yeast (Mendenhall and Hodge, 1998) and has the highest connectivity of the proteins in the CORE dataset. CDC28 is fundamental in the control of the main events of the cell division cycle in yeast (Mendenhall and Hodge, 1998) and therefore interacts with many proteins that do not necessarily have high sequence or semantic similarity to CDC28. Therefore, the results from protein-protein interaction predictions for the proteins in the CORE-CDC28 dataset are expected to contain a considerable number of false negative interactions.

[2] http://genome-www.stanford.edu/Saccharomyces

The second subset is the CORE-MOD dataset. This dataset contains proteins that have been predicted to form functional modules (Lubovac *et al.*, 2005). The module prediction was made using calculations of both semantic similarity and clustering coefficient. All of the 109 proteins in this subset belong to modules with a semantic similarity above 0.7, using the measure by Lin (1998) and a clustering coefficient above 0.6 and are shown in table 5.1. Because of the high semantic similarity of the modules, it is assumed that the protein-protein interaction prediction method will be able to predict the protein-protein interactions in the CORE-MOD dataset with high accuracy.

Table 5.1: Each row in the table represents a functional module predicted by Lubovac *et al.* (2005). The first column shows the hub of the module, i.e. the highly connected protein that holds the module together, followed by columns that represent the hub's neighbours. The final two columns show the clustering coefficient and the semantic similarity of the modules, respectively.

Hub	**N1**	**N2**	**N3**	**N4**	**N5**	**N6**	**N7**	**N8**	**CC**	**SSM**
WBP1	OST1	OST2	OST3	OST4	OST5	STT3	SWP1	PKC1	0.82	0.80
GCD6	GCD1	GCD2	GCD7	GCD11	SUI2	SUI3	GCN3		0.76	1.00
PRP24	LSM2	LSM3	LSM5	LSM6	LSM7	LSM8			1.00	0.74
RPB4	RPB7	IKI3	RPB2	RPB3	RPB5	RPO21			0.67	0.75
SET3	SIF2	ZDS1	SNT1	HOS4	HOS2	HST1			0.87	0.71
RPC53	TFC4	RPC37	RPC34	RPC40	RPO31				0.60	0.91
RET1	RPC11	RPC19	RPC40	RPC34	RPO31				0.70	1.00
RPN3	RPN12	RPC40	RPT3	RPT1	PRE1				0.60	0.78
HCR1	RPG1	PRT1	TIF35	TIF5	SUI1				0.70	0.76
ELP6	IKI1	YLR327C	ELP2	IKI3	ELP4				0.60	0.73
UTP4	UTP7	KRR1	PWP2	UTP22	UTP18				0.70	0.75
DIP2	UTP7	PWP2	UTP22	UTP18					0.83	1.00
VAM7	VAM3	YPT7	YKT6	VTI1	NYV1				0.60	0.86
SME1	SMX3	HSH155	PRP4	SMD2	SMD3				0.60	0.97
MAK11	ERB1	NOP2	TIF6	NOP7					0.67	0.75
VMA10	VMA2	VMA7	VPH1	STV1					0.67	1.00
VMA5	VMA2	VPH1	STV1	VMA8					0.67	1.00
HIR2	SNF5	SWI3	SNF2	HIR1					1.00	0.73
RTF1	CTR9	LEO1	PAF1	SPT5					1.00	0.80
SPC98	TUB4	SPC97	SPC110	SPC72					0.83	1.00

The third subset is CORE-CC06 which contains all proteins in the CORE set that have a clustering coefficient of at least 0.6, in total 318 proteins.

The fourth and last subset, CORE-CC05 contains randomly selected proteins from the CORE dataset that have a clustering coefficient below 0.6. This set also contains 318 proteins, as the CORE-CC06 set, and was created for comparison to CORE-CC06.

Clustering coefficient was a determining factor when predicting the functional modules in the CORE-MOD dataset. The modules all have a clustering coefficient of at least 0.6 and thus it would be interesting to see if the protein-protein interaction prediction method will predict the protein-protein interactions in the CORE-CC06 dataset with higher accuracy than the interactions in the CORE-CC05 dataset.

5.4 Application and evaluation of the method

For each of the subsets, the following procedure was applied. First, semantic similarity between all of the proteins in the dataset was calculated using a script written in PHP[3]. This script uses the information obtained from the SGD and calculates the semantic similarity across the GO terms that are assigned to the proteins, using one of the three semantic similarity measures (Resnik, 1999; Lin, 1998; Jiang and Conrath, 1998) at a time. The results are written to a file so that they can be analysed manually and accessed by other applications.

Another application, written in Java [4] then reads the file and creates a pairwise matrix consisting of 1s and 0s where 1 stands for interaction between the proteins and 0 stands for no interaction. A pair of proteins receives a 1 if the semantic similarity between them is above a specific threshold value, otherwise it receives a 0.

This application also creates a similar matrix from the original CORE dataset. This matrix is considered to be correct, i.e. is assumed that it does not contain any false positive or false negative interactions, and can therefore be used to calculate the sensitivity and specificity of the interactions in the matrix based on semantic similarity.

When a term is compared to itself using the semantic similarity measures by Resnik (1999) and Jiang and Conrath (1998), the score depends on where in the ontology the term is, i.e. terms that occur less frequently have higher scores than those that are more common (Lord *et al.*, 1999b). The measure by Lin (1998) hides this information and a term compared to itself always scores 1 (Lord *et al.*, 1999b), which as said in chapter 2.4.1 is the highest possible score using this measure.

[3] http://www.php.net

[4] http://java.sun.com

When creating the pairwise matrix from the CORE dataset the diagonal of the matrix will be filled with 0s, since a term is not considered to be its own neighbour. In the matrix based on semantic similarity, whether a term compared to itself receives a 1 or a 0 depends on the semantic similarity measure and the threshold value used. If the measure by Lin (1998) is used, all terms compared to themselves will receive a 1, but with the other two measures the score will only be 1 if the semantic similarity is above the threshold value used.

This creates problems when comparing the two matrices, for instance when using the measure by Lin (1998). Since the diagonal in that matrix is filled with 1s and the matrix created from the CORE dataset has only 0s in its diagonal, all of the predictions in the diagonal of the matrix based on semantic similarity would be considered to be false positive predictions.

To avoid this kind of problems, the application fills the diagonals of both matrices with 1s resulting in that all predictions in the diagonal will be considered to be true positives.

The threshold value of the minimum semantic similarity required between a pair of proteins to receive a 1 in the matrix was varied to see which threshold value is generates results with the highest specificity and sensitivity. The interval within which the threshold value was varied was determined by manually inspecting the results from the PHP script that calculates the semantic similarity between the proteins. Different intervals were used for the three different semantic similarity measures.

The Java application compares the two matrices; the one created from the CORE dataset to the matrix based on semantic similarity, and calculates several measures to determine the predictive power of the method. First it calculates the specificity and sensitivity of the prediction method. Sensitivity is defined as the probability that the method correctly identifies positives and is calculated using the equation:

$$\frac{TP}{TP + FN}$$

where TP is the number of true positives and FN is the number of false negatives. The specificity is the method’s power to correctly identify negatives, calculated as:

$$\frac{TN}{TN + FP}$$

where TN is the number of true negatives and FP is the number of false positives (Piatt, 2004).

Predictive values, both positive and negative are also calculated. The positive predictive value (PPV) is the probability that two proteins that are predicted to interact based on semantic similarity, do in fact interact. It is calculated as:

$$\frac{TP}{TP + FP}.$$

The negative predictive value (NPV) is the probability that two proteins that are not predicted to interact do not do so in the CORE dataset and is calculated as (Piatt, 2004):

$$\frac{TN}{TN + FN}$$

The sensitivity and specificity of a predictive method varies with the chosen threshold value. Receiver Operating Characteristic (ROC) curves are used to display the range of sensitivities and specificities of a prediction method. An application written in MATLAB[5] was implemented to create ROC curves for each subset of data and using all three measures of semantic similarity. A ROC curve is a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) and shows the trade-off between sensitivity and specificity of a prediction method, i.e. an increase in sensitivity will result in decrease in specificity. The area under the ROC curve is a measure of the accuracy of the prediction method. An area of 1 signifies a perfect prediction method whereas an area of 0.5 represents a method of very poor predictive power (Tape, 2003).

5.5 Predicting the function of proteins

All four subsets that were used in this study contain several proteins that are annotated with the term GO:0005554, which stands for unknown molecular function. An effort towards determining the function of these proteins based on the functions of their interacting partners was made, with the main focus on the proteins belonging to the CORE-CC06 dataset. The predictions were made in connection with the connectivity of the node’s neighbours. Connectivity refers to the number of neighbours that a

[5] http://www.mathworks.com/products/matlab/

specific protein has. It was hypothesized that neighbours with higher connectivity carry more semantic function than neighbours with lower connectivity and thus give better clues about the function of the protein of unknown function. Figure 5.2 below illustrates this hypothesis.

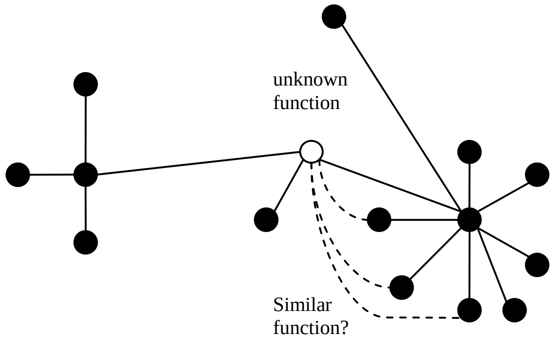


Figure 5.2: A protein interaction network. The filled circles represent proteins of known function and the unfilled one represents a protein of unknown function. The neighbours of the protein of unknown function have connectivities of 9, 1 and 4, respectively. According to the hypothesis, it is assumed probable that the protein has a function similar to the function of the neighbours of the protein with a connectivity of 9.

Based on the findings of the protein-protein interaction prediction experiments and this hypothesis, it was decided to select the neighbour with the highest connectivity and base the function predictions on the minimum subsumer of the neighbours of this neighbour.

Thus, for each protein of unknown function in the dataset, the neighbour with the highest connectivity was chosen. SGD Gene Ontology term finder[6] was used to find the minimum subsumer for as many neighbours as possible. All neighbours with a connectivity of ten or higher were considered. The reason for that was that it was considered that a node with a connectivity of ten carries that much semantic function that it has to be taken into account. If a protein of unknown function had many neighbours with connectivity above ten, the top three were considered.

[6] http://db.yeastgenome.org/cgi-bin/GO/goTermFinder

When selecting a function for the protein of unknown function from the variety of GO terms that are assigned to the proteins, terms with low *p*-values and high frequency were favoured. The frequency is calculated by dividing the number of neighbours that are annotated with the specific term by the total number of neighbours. When more than one neighbour was considered, the protein of unknown function was assigned a functional term with relatively high frequency, i.e. the term should be present in a large fraction of the neighbours' neighbours, if possible. If no one representative term was present among all neighbours, more than one function was assigned to the protein. The function of the neighbours of the neighbour with the highest connectivity was, however, assumed to be most probable.

This prediction method was primarily applied to the proteins of unknown function in the CORE-CC06 dataset since it was considered to be most appropriate for this analysis. The CORE-MOD dataset contained only a few proteins of unknown function. All of the proteins of unknown function in the CORE-CC05 dataset have one single neighbour and the proteins in the CORE-CDC28 dataset all have CDC28 as their neighbour with highest connectivity. As previously stated, CDC28 has the highest connectivity of the proteins in the CORE dataset and interacts with many proteins that do not necessarily share high sequence similarity or the similar semantic function with CDC28.

This function prediction method was also applied to a small dataset called CORE-TEST. This datasets contains 30 randomly selected proteins with known functions. The results were then compared to the correct functions of the proteins to gain some clues about the accuracy of the method.

6 Results and analysis

In this chapter the results from protein-protein interaction predictions based on semantic similarity are presented and analysed along with the results from predicting function of proteins. Chapter 6.1 presents the results from the protein-protein interaction predictions and chapter 6.2 contains the results from the function predictions.

6.1 Prediction of protein-protein interactions

All of the three semantic similarity measures that were used to predict pairwise interactions between proteins performed with very high specificity and sensitivity in three out of four datasets. There is a difference in predictive power between the measures but as expected, all three produced the best results when predicting the protein-protein interactions in the CORE-MOD dataset and the worst when predicting the interactions in the CORE-CDC28 dataset (see figures 6.1, 6.2 and 6.3). It was expected that the prediction method would predict the interactions in the CORE-CC06 datasets with more accuracy than the interactions in the CORE-CC05 dataset, but that was only the case with the measures by Lin (1998) and Jiang and Conrath (1998). The measure by Resnik (1999) predicted the interactions in the CORE-CC05 with higher accuracy. The difference in the accuracy between the two datasets when the Resnik measure is used is not as large as when the other two measures are used (see figures 6.1, 6.2 and 6.3).

This information about the difference in the accuracy of the predictions between the datasets is based on the area under the ROC curves for each dataset and semantic similarity measure. This area can be used as a measure of the accuracy of a prediction method and Tape (2003) proposes that a traditional academic point system can be used as a rough guide for classifying the accuracy of a prediction method. This system is presented in table 6.1.

Table 6.1: A rough guide for classifying the accuracy of a prediction method based on the area under a ROC curve (Tape, 2003).

Area	Predictive power
0.9 – 1.0	Excellent (A)
0.8 – 0.9	Good (B)
0.7 – 0.8	Fair (C)
0.6 – 0.7	Poor (D)
0.5 – 0.6	Fail (E)

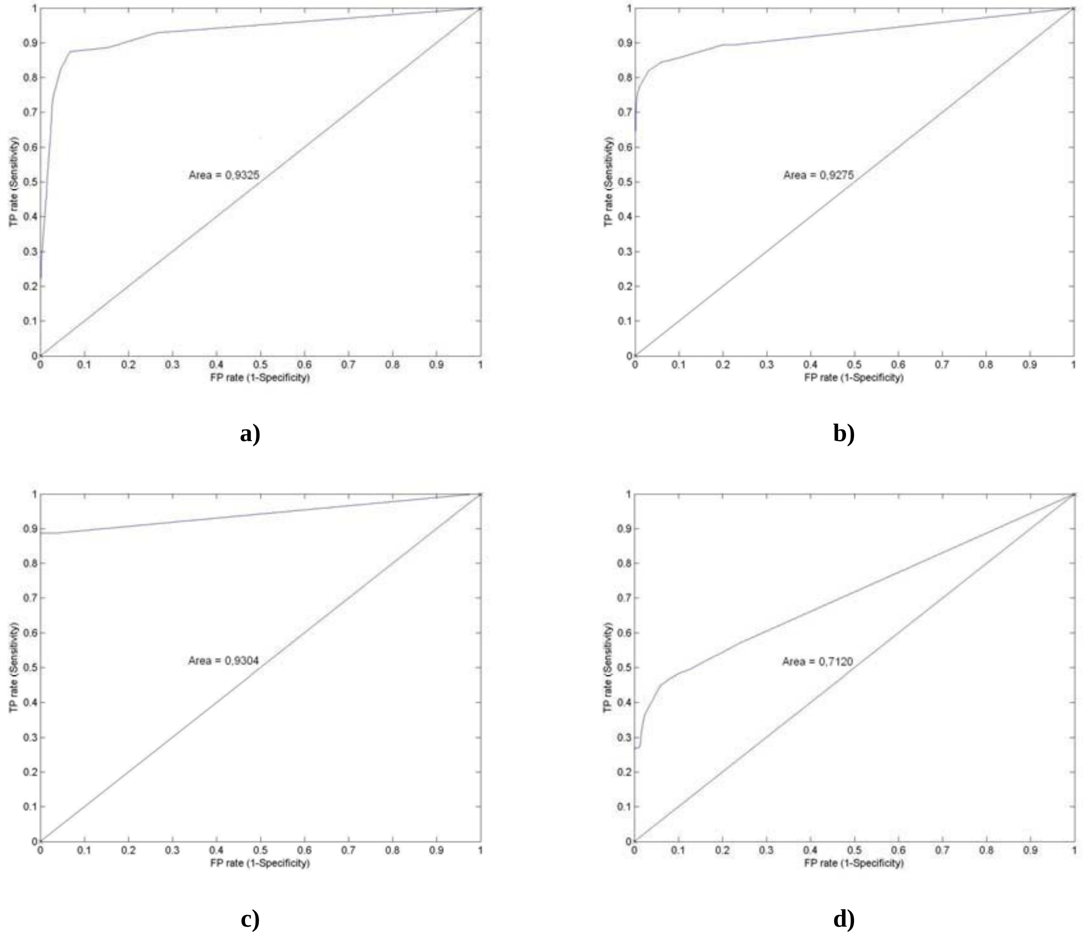
The measure by Resnik (1999) performed best in predicting the protein-protein interactions in three of four datasets. Lord *et al.*, (2003b) illustrated that this measure showed the strongest correlation between sequence similarity and the molecular function aspect of GO (see chapter 3.2). The only dataset where the Resnik measure did not perform best of all the measures was the CORE-CDC28 dataset.

As mentioned in chapter 5.4, a Java application creates a matrix of predicted protein-protein interactions and evaluates it by comparing it to the original protein network. It calculates the sensitivity and specificity of the predictions along with negative and positive predictive values (NPV and PPV). The sensitivity of the predictions is highest at lower threshold values while the specificity is highest at higher threshold values. Therefore, for each threshold value, the average of the two was calculated and the predictions are assumed to be best at the threshold value where the average specificity and sensitivity was highest. The best results from using the semantic similarity measure by Resnik on all four datasets are shown in table 6.2.

Table 6.2: The results from the Java application using the measure by Resnik (1999). Threshold values where the average of the specificity and the sensitivity was highest were chosen.

Dataset	Results
CORE-MOD	Measure: resnik Threshold value: 3.0 Number of proteins: 109 Total number of predictions: 11881 Number of correct predictions: 11255 (94.73%) Number of false positive predictions: 510 (4.29%) Number of false negative predictions: 116 (0.98%) Specificity: 0.9545 Sensitivity: 0.8245 Positive Predictive Value (PPV): 0.5166 Negative Predictive Value (NPV): 0.9893 Average spec & sens: 0.8895 Average PPV & NPV: 0.7529
CORE-CC06	Measure: resnik Threshold value: 2.0 Number of proteins: 318 Total number of predictions: 101124 Number of correct predictions: 95010 (93.95%) Number of false positive predictions: 6026 (5.96%) Number of false negative predictions: 88 (0.09%) Specificity: 0.9401 Sensitivity: 0.8445 Positive Predictive Value (PPV): 0.0735 Negative Predictive Value (NPV): 0.9991 Average spec & sens: 0.8923 Average PPV & NPV: 0.5363
CORE-CC05	Measure: resnik Threshold value: 8.0 Number of proteins: 318 Total number of predictions: 101124 Number of correct predictions: 101085 (99.96%) Number of false positive predictions: 0 (0.0%) Number of false negative predictions: 39 (0.03%) Specificity: 1.0 Sensitivity: 0.8908 Positive Predictive Value (PPV): 1.0 Negative Predictive Value (NPV): 0.9996 Average spec & sens: 0.9454 Average PPV & NPV: 0.9998
CORE-CDC28	Measure: resnik Threshold value: 4.0 Number of proteins: 96 Total number of predictions: 9216 Number of correct predictions: 8768 (95.14%) Number of false positive predictions: 218 (2.37%) Number of false negative predictions: 230 (2.50%) Specificity: 0.9754 Sensitivity: 0.3681 Positive Predictive Value (PPV): 0.3807 Negative Predictive Value (NPV): 0.9741 Average spec & sens: 0.6718 Average PPV & NPV: 0.6774

The ROC curves for the results obtained when using the semantic similarity measure by Resnik to predict protein-protein interactions in all four datasets are shown in figure 6.1.



a) b) c) d)

Figure 6.1: The ROC curves of the protein-protein interaction predictions using the semantic similarity measure by Resnik with the threshold value varied between 0 and 10. The area below the diagonal line is 0.5, as well as the area above it. The ROC curve for dataset a) CORE-MOD, b) CORE-CC06, c) CORE-CC05, d) CORE-CDC28.

According to the guide for classification of predictive power, presented in table 6.1, the predictions for three of the datasets, CORE-MOD, CORE-CC06 and CORE-CC05, fall into category A, which means that the predictions have excellent accuracy. The predictions for dataset CORE-CDC28 falls into category C which indicates fair accuracy.

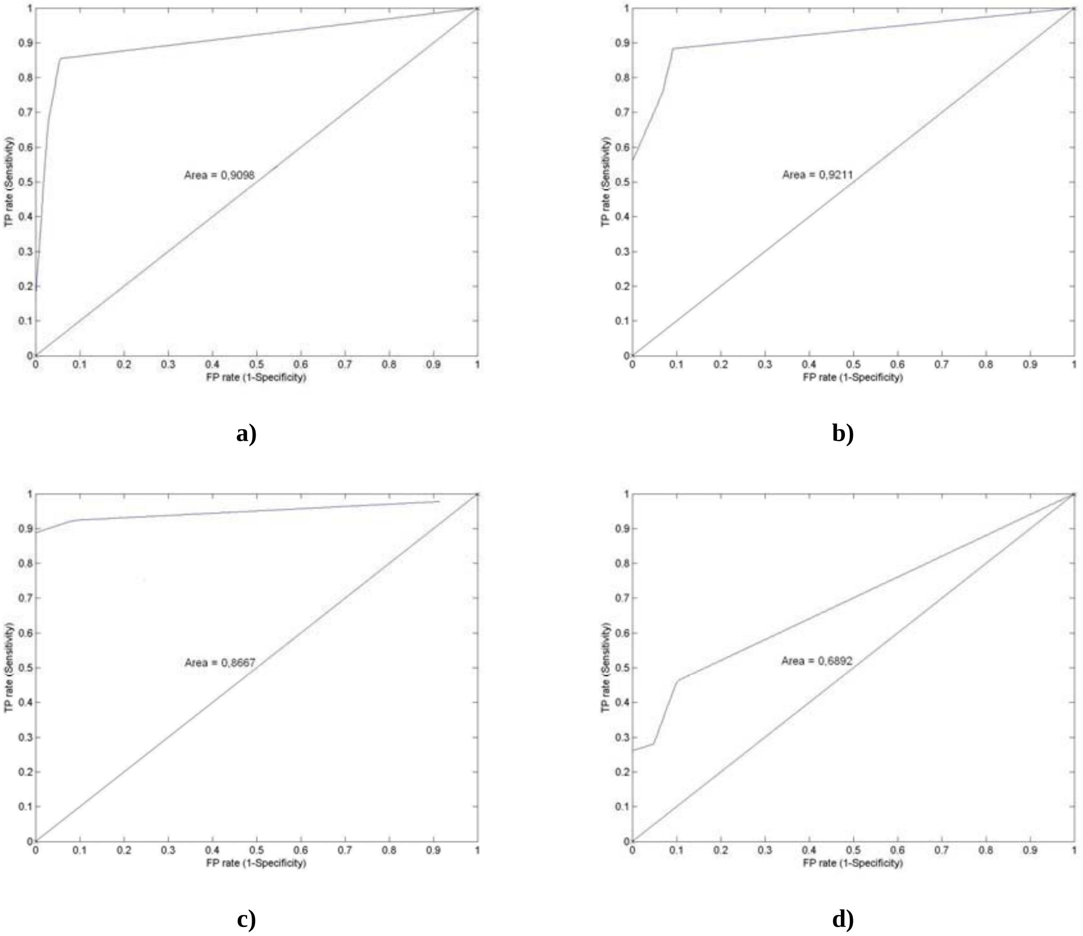
The semantic similarity measure by Lin (1998) performed second best in three of four datasets. Again CORE-CDC28 is the dataset that deviates from the others and in it the Lin measure predicted the protein-protein interactions with the lowest accuracy. This measure has though its advantages over the Resnik measure. For instance it takes the whole function of the proteins into account, i.e. the average of all terms that are assigned to the protein and their parent’s functions. The Resnik measure only makes use of the information content of the shared parents (see chapter 2.4 for further details).

The best results produced by applying the semantic similarity measure by Lin on all four datasets are shown in table 6.3.

Table 6.3: The results from the Java application using the measure by Lin (1998). Threshold values where the average of the specificity and the sensitivity was highest were chosen.

Dataset	**Results**
CORE-MOD	Measure: lin Threshold value: 0.5 Number of proteins: 109 Total number of predictions: 11881 Number of correct predictions: 11169 (94.01%) Number of false positive predictions: 616 (5.18%) Number of false negative predictions: 96 (0.81%) Specificity: 0.9451 Sensitivity: 0.8548 Positive Predictive Value (PPV): 0.4784 Negative Predictive Value (NPV): 0.9910 Average spec & sens: 0.8999 Average PPV & NPV: 0.7347
CORE-CC06	Measure: lin Threshold value: 0.7 Number of proteins: 318 Total number of predictions: 101124 Number of correct predictions: 93728 (92.69%) Number of false positive predictions: 7290 (7.21%) Number of false negative predictions: 106 (0.10%) Specificity: 0.9275 Sensitivity: 0.8127 Positive Predictive Value (PPV): 0.0594 Negative Predictive Value (NPV): 0.9989 Average spec & sens: 0.8701 Average PPV & NPV: 0.5291
CORE-CC05	Measure: lin Threshold value: 1.0 Number of proteins: 318 Total number of predictions: 101124 Number of correct predictions: 93707 (92.67%) Number of false positive predictions: 7388 (7.31%) Number of false negative predictions: 29 (0.03%) Specificity: 0.9267 Sensitivity: 0.9188 Positive Predictive Value (PPV): 0.0425 Negative Predictive Value (NPV): 0.9997 Average spec & sens: 0.9227 Average PPV & NPV: 0.5211
CORE-CDC28	Measure: lin Threshold value: 0.7 Number of proteins: 96 Total number of predictions: 9216 Number of correct predictions: 8332 (90.41%) Number of false positive predictions: 680 (7.38%) Number of false negative predictions: 204 (2.21%) Specificity: 0.9232 Sensitivity: 0.4396 Positive Predictive Value (PPV): 0.1905 Negative Predictive Value (NPV): 0.9756 Average spec & sens: 0.6814 Average PPV & NPV: 0.5831

The ROC curves for the results obtained when using the semantic similarity measure by Lin to predict protein-protein interactions in all four datasets are shown in figure 6.2.



a)

b)

c)

d)

Figure 6.2: The ROC curves of the protein-protein interaction predictions using the semantic similarity measure by Lin with the threshold value varied between 0 and 11. The area below the diagonal line is 0.5, as well as the area above it. The ROC curve for dataset a) CORE-MOD, b) CORE-CC06, c) CORE-CC05, d) CORE-CDC28.

For this semantic similarity measure, predictions for two of the datasets fall into category A in table 6.1, i.e. excellent accuracy in the predictions. These datasets are CORE-MOD and CORE-CC06. The CORE-CC05 dataset belongs to category B, good accuracy, and for the CORE-CDC28 the results fall into category D, poor accuracy.

The third and final measure is the one by Jiang and Conrath (1998). It showed the worst predictive accuracy for three of the four datasets, but the best for the CORE-CDC28 dataset. This may suggest that the properties of the datasets that are used do not matter as much to the Jiang and Conrath measure as to the other two measures. Furthermore, this measure has the same advantage as the measure by Lin over the one by Resnik, that is, it includes the whole function of the proteins and not only the one of the shared parents.

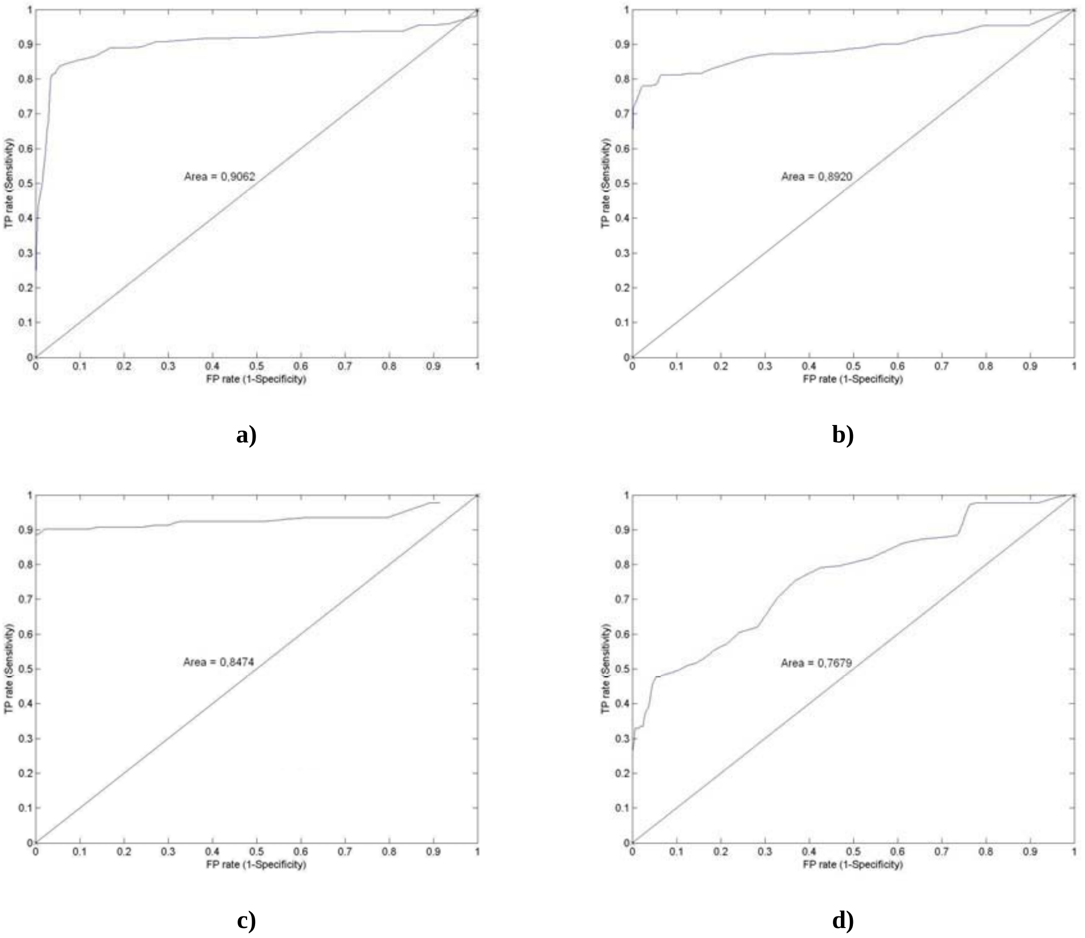
The best results from the Java application using the semantic similarity measure by Jiang and Conrath on all four datasets are shown in table 6.4.

Table 6.4: The results from the Java application using the measure by Jiang and Conrath (1998). Threshold values where the average of the specificity and the sensitivity was highest were chosen.

Dataset	**Results**
CORE-MOD	Measure: jiang Threshold value: 17.0 Number of proteins: 109 Total number of predictions: 11881 Number of correct predictions: 11157 (93.91%) Number of false positive predictions: 618 (5.20%) Number of false negative predictions: 106 (0.89%) Specificity: 0.9449 Sensitivity: 0.8396 Positive Predictive Value (PPV): 0.4731 Negative Predictive Value (NPV): 0.9901 Average spec & sens: 0.8923 Average PPV & NPV: 0.7316
CORE-CC06	Measure: jiang Threshold value: 16.0 Number of proteins: 318 Total number of predictions: 101124 Number of correct predictions: 94518 (93.47%) Number of false positive predictions: 6500 (6.43%) Number of false negative predictions: 106 (0.10%) Specificity: 0.9354 Sensitivity: 0.8127 Positive Predictive Value (PPV): 0.0661 Negative Predictive Value (NPV): 0.9989 Average spec & sens: 0.8740 Average PPV & NPV: 0.5325
CORE-CC05	Measure: jiang Threshold value: 34.0 Number of proteins: 318 Total number of predictions: 101124 Number of correct predictions: 101085 (99.96%) Number of false positive predictions: 0 (0.0%) Number of false negative predictions: 39 (0.04%) Specificity: 1.0 Sensitivity: 0.8908 Positive Predictive Value (PPV): 1.0 Negative Predictive Value (NPV): 0.9996 Average spec & sens: 0.9454 Average PPV & NPV: 0.9998
CORE-CDC28	Measure: jiang Threshold value: 17.0 Number of proteins: 96 Total number of predictions: 9216 Number of correct predictions: 8464 (91.84%) Number of false positive predictions: 562 (6.10%) Number of false negative predictions: 190 (2.06%) Specificity: 0.9365

Sensitivity: 0.4780
Positive Predictive Value (PPV): 0.2364
Negative Predictive Value (NPV): 0.9776
Average spec & sens: 0.7073
Average PPV & NPV: 0.6070

The ROC curves for the results obtained when using the semantic similarity measure by Jiang and Conrath to predict protein-protein interactions in all four datasets are shown in figure 6.3.



a)

b)

c)

d)

Figure 6.3: The ROC curves of the protein-protein interaction predictions using the semantic similarity measure by Jiang and Conrath with the threshold value varied between 0 and 35. The area below the diagonal line is 0.5, as well as the area above it. The ROC curve for dataset a) CORE-MOD, b) CORE-CC06, c) CORE-CC05 and d) CORE-CDC28.

For this semantic similarity measure, predictions for one dataset, CORE-MOD, falls into category A in table 6.1, i.e. have excellent accuracy. The accuracy of the predictions in the CORE-CC06 and CORE-CC05 datasets belong to category B, good accuracy, and for the CORE-CDC28 the results fall into category C, fair accuracy.

Table 6.5 summarizes the accuracy of the predictions after applying all three semantic similarity measures on each dataset.

Table 6.5: The accuracy of the predictions after each semantic similarity measure has been applied on all four datasets. The class that the accuracy belongs to is shown within brackets (Tape, 2003).

	Lin	**Resnik**	**Jiang and Conrath**
CORE-MOD	0.9098 (A)	0.9325 (A)	0.9062 (A)
CORE-CC06	0.9211 (A)	0.9275 (A)	0.8920 (B)
CORE-CC05	0.8667 (B)	0.9304 (A)	0.8474 (B)
CORE-CDC28	0.6892 (D)	0.7120 (C)	0.7679 (C)

The measure by Resnik gives the best overall results, but no statistical test has been applied on the data to test the significance of the difference in predictive power between the semantic similarity measures.

6.2 Prediction of function of proteins

According to the results from the protein-protein interaction predictions, the semantic similarity measure by Resnik (1999) is able to predict protein-protein interactions with the highest accuracy among the three measures. As said in chapter 2.4.2, semantic similarity calculated with this measure is only based on the probability of the minimum subsumer of the two terms being compared. Based on that, it was decided to use the probability of the minimum subsumer of neighbour proteins in combination with connectivity to predict the function of proteins of unknown function. The CORE-CC06 dataset was used and it contains a total of 70 proteins for which the function is not known. Function predictions for 10 randomly selected proteins from the dataset are shown in table 6.6, while the complete results for all of the proteins of unknown function in the CORE-CC06 dataset can be found in appendix A.

Table 6.5: The results from function predictions for 10 randomly selected proteins of unknown function in the CORE-CC06 dataset.

Node **(# neighbours)**	**Neighbour with highest connectivity** **(# neighbours)**	**Predicted function**	**Frequency**	***p*-value**
MPE1 (7)	PFS2 (14) CFT2 (13) PAP1 (12)	GO:0003723 RNA binding	50.0% 61.5% 66.7%	0.0488
PEP5 (5)	PEP3 (7) -- VPS16 (7)	GO:0005484 SNAP receptor activity -- GO:0005085 guanyl-nucleotide exchange factor activity	28.6% -- 42.9%	0.0038 -- 0.0029
SGF29 (4)	TAF6 (20) NGG1 (17) SPT7 (17)	GO:0030528 transcription regulator activity	65.0% 82.4% 70.6%	0.0518
ELG1 (4)	RFC4 (10) RFC2 (10) RFC5 (10)	GO:0003689 DNA clamp loader activity	30.0% 40.0% 40.0%	0.0007
YBL046W (3)	SPT5 (9)	RNA polymerase II transcription elongation factor activity	44.4%	0.0031
COG1 (3)	SED5 (18) COG3 (10)	GO:0005485 v-SNARE activity	38.9% 30.0%	0.0020
PEP8 (3)	VPS35 (3)	all unknown		
YIP3 (2)	YIF1 (14)	GO:0003924 GTPase activity	71.4%	0.0075
YNL056W (2)	SIW14 (2) OCA1 (2)	no terms found		
RSC6 (2)	ISW1 (17)	GO:0016887 ATPase activity	17.6%	0.0230

Of the 70 proteins of unknown function in the dataset it was impossible to assign a function to 9. For some of these proteins this was due to all of the neighbours of their immediate neighbours have unknown molecular function. For other proteins the reason was that the SGD GO term finder did not find any relevant terms in common for the neighbours, i.e. they had no significant minimum subsumer.

Six of the proteins could not be assigned only one function. These proteins all have several neighbours with high connectivity and thus the minimum subsumer of more than one of the neighbour’s neighbours is considered. These different neighbours then didn’t have any applicable minimum subsumers in common and thus all minimum subsumers were assigned to the protein.

The method was also applied on 30 known randomly selected proteins to test its performance. The method was able to predict the functions of all 30 proteins in this CORE-TEST dataset. For 22 of these proteins, i.e. 73.3%, the predicted functions were correct. Further analysis of the eight proteins whose function is different from the predicted revealed that three of these false predictions were good estimates of the correct function. Those three proteins were assigned a functional term that is the parent of the correct functional term. This means that the predictions give good clues about the functions of the proteins. The complete results from the function predictions for the proteins in the CORE-TEST dataset are found in appendix B.

7 Discussion

In this chapter, the results from the protein-protein interaction and protein function predictions are discussed. Chapter 7.1 summarizes and briefly discusses the results obtained from these experiments and chapter 7.2 discusses the various potential causes of errors in the study.

7.1 A summary of the results

All three measures of semantic similarity are able to predict protein-protein interactions with good specificity and sensitivity and thus, in combination with Gene Ontology, seem to be a good tool for such predictions. The measure by Resnik (1999) generated results with the highest accuracy of the three measures when applied on three of the four datasets. The measure by Lin (1998) performed second best on the same three datasets, followed by the measure by Jiang and Conrath (1998). For the last dataset, CORE-CDC28, the measure by Jiang and Conrath generated the best results followed by the measure by Resnik and then the one by Lin. This may suggest that the measure by Jiang and Conrath is less affected by the different properties of the datasets it is applied on than the other two measures. The difference in accuracy is though quite large between the datasets that the measure by Jiang and Conrath performed the best and the worst on.

All of the measures performed best on the CORE-MOD dataset and the worst on the CORE-CDC28. CDC28 is a large hub that interacts with many proteins that can have very different functions and thus low semantic similarity. These results suggest that the predictive power of the method can be increased if it takes predicted functional modules into consideration. The measures by Lin and Jiang and Conrath performed better on the CORE-CC06 dataset than the CORE-CC05 set, but the measure by Resnik did otherwise. Beforehand it was considered more probable that the results generated when the CORE-CC06 dataset was used would be better than the results obtained using the CORE-CC05 dataset. That would have supported the module hypothesis since the predicted functional modules all have a clustering coefficient above 0.6. Since the CORE-CC05 dataset was not completely randomly created resulting in that all of the proteins in it have only a single neighbour, a recreation of the dataset might give other results.

In this study, semantic similarity is a quite general property. As long as there is any kind of similarity between two proteins in terms of function, and the similarity is above the threshold value used, an interaction is predicted. It would be interesting to inspect the predicted interactions further to see if protein-protein interactions are more likely if the function shared by the proteins is of a particular type.

It is difficult to discuss and summarise the results obtained from the protein function predictions since a complete validation has not been performed on the data. Some clues about the accuracy of the method were obtained from the results produced from the CORE-TEST dataset, but the dataset is considered too small to draw any reliable conclusions.

However, under the hypothesis that an unknown protein has similar function as the neighbours of its neighbour with the highest connectivity, 61 of the 70 proteins of unknown function in the CORE-CC06 dataset were assigned a function. Some clues towards determining the accuracy of these predictions can be obtained by looking at the probability of the minimum subsumer (*p*-value) and the frequency of that term. Good predictions have a high frequency and a low *p*-value.

The method was able to predict the function of all 30 proteins in the CORE-TEST dataset and 22 of those predictions were accurate. Of the eight proteins that were not assigned the correct function, three proteins were assigned a function that is very close to the correct one. This suggests that the method predicts the function of proteins with high accuracy.

7.2 Potential sources of errors in the study

There are several factors that may have contributed to decreased accuracy in the obtained results.

The versions of the SGD dataset and the GO annotations that are used in this project are from year 2004. Thus, new information may have been added since the release of these datasets, i.e. some proteins may have been assigned more specific GO terms or the hypothetical proteins may have been assigned functional annotation. It would be interesting to see if this information would affect the results of this study.

Another fact that could possibly affect the outcome of the predictions is the lack of randomness when creating the CORE-CC05 dataset. This dataset was created for comparison to the CORE-CC06 dataset and contains 318 proteins with a clustering coefficient below 0.6. All of the proteins in the CORE-CC05 dataset only have a single neighbour since most of them were present at the end of the file containing all proteins in the DIP-YEAST CORE dataset with a clustering coefficient below 0.6. This file was sorted so that the proteins with the highest number of neighbours were at the top of the file and the ones with the lowest number were at the bottom. This file should have been rearranged before selecting 318 proteins from the top or the bottom of it. That would have made the dataset more reliable for comparison to the CORE-CC06 dataset since CORE-CC06 contains proteins with various numbers of neighbours.

It would also have been interesting to test the method on the whole CORE dataset and not just on the various subsets created. The areas under the ROC curves for the whole CORE dataset could then be compared to the areas for the smaller datasets to see which dataset generated better results than the whole dataset and which generated worse results. This was not possible since the computational power to calculate the semantic similarity between all proteins in the dataset was not sufficient.

A random subset of proteins from the CORE dataset could have been used instead of the whole dataset. If correctly created, the accuracy of the predictions in that dataset would probably be very similar to the accuracy of the predictions when using the whole dataset.

As said in chapter 3.2, Lord *et al.* (1999b) showed that semantic similarity based on the molecular function aspect of GO showed strongest correlation to sequence similarity of all of the GO aspects. This was the main reason for choosing this aspect in this study. Not all proteins that interact have similar functions or sequence and therefore it would be interesting to take the other two aspects into consideration, especially the biological process aspect. The accuracy of the predictions would probably be increased if a combination of two or three aspects would be used.

8 Conclusions

The aim of this project was to investigate whether semantic similarity measures based on Gene Ontology could be used to predict protein-protein interactions. Three measures were evaluated in order to determine which one could predict the interactions in the CORE dataset of protein-protein interactions in *S. cerevisiae* with the highest sensitivity and specificity. Those three measures were the originally proposed by Lin (1998), Resnik (1999) and Jiang and Conrath (1998). In addition to calculating the sensitivity and specificity of the predicted interactions, ROC curves were created to assess the accuracy of the predictions.

Based on the results of the protein-protein interaction predictions it is possible to conclude that semantic similarity can be used to predict such interactions. Further testing and analysis is required to determine optimal parameter values, e.g. to find the optimal threshold values for each measure.

It is also possible to conclude that the measure by Resnik is the measure that is best capable of predicting protein-protein interactions, even though the other two measures have their advantages. This is based on the accuracy of the predictions in this study, i.e. the area under the ROC curves since there is not much difference in the average specificity and sensitivity of the derived interactions between the different measures.

As mentioned before, it is difficult to make any reliable conclusions about the function predictions for the proteins of unknown function since the predictions have not been completely validated using any standardized measure. The fact that 73.3% of the function predictions in the CORE-TEST dataset were accurate does though suggest that the method is able to predict protein function with high accuracy.

9 Future work

Several ideas for future research can be based on the findings of this project. The optimal goal would be to compensate for all the shortcomings of this study.

Probably the most interesting work would be to investigate the predictive power of the protein-protein interaction prediction method if the biological process and cellular component aspects of GO are used instead of the molecular function aspect. A combination of two or even all three aspects would probably be of most interest.

It would also be interesting to inspect the method further with respect to predicted functional modules, i.e. predict the protein-protein interactions in such modules correctly by allowing different threshold values within a single module.

In this project, the method was only tested on proteins from *S. cerevisiae*, but it would also be interesting to test its performance on data from other organisms.

The aspect of this project that requires the most future research is the function predictions. These predictions were based on a hypothesis that requires more analysis and testing for validation. The predictions that were made in this project have not been completely validated using any standardized measure or confidence values. The results obtained from the small CORE-TEST dataset suggest that the method is very accurate and the frequency and *p*-value for all predictions give some clues towards determining the probability of the predictions being correct. The method that was used in this project could for instance be applied to a larger dataset of proteins of known function to see how accurate the method is.

It would also be very interesting to compare the accuracy of the function prediction method that was used in this project to some other methods used for function predictions. For instance the method by Vinayagam *et al.* (2004), which is discussed in chapter 3.1 and also assigns GO terms to proteins, could be applied on the datasets that were used in this study and the accuracy of those predictions compared to the accuracy of the function prediction method used in this project.

References

Bader, J.S., Chaudhuri, A., Rothberg, J.M. and Chant, J., 2004. Gaining confidence in high-throughput protein interaction networks. *Nature biotechnology*, 22(1), pp. 78-85.

Barabási, A-L. and Oltvai, Z.N., 2004. Network biology: understanding the cell's functional organization. *Genetics*, 5, February, pp. 101-113.

Deane, C.M., Salwinski, L., Xenarios, I. and Eisenberg, D., 2002. Protein Interactions: Two methods for assessment of the reliability of high-throughput observations. *Mol. Cell. Proteomics*, 1(5), pp. 349-356.

Deng, M.m Zhang, K., Metha, S., Chen, T. and Sun, F., 2002. Prediction of protein function using protein-protein interaction data. In *Proceedings of the IEEE Computer Society Bioinformatics Conference*, Stanford, California, pp.197-206.

Franzot, G. and Carugo, O., 2003. Computational approaches to protein-protein interaction. *J Struct Funct Genomics*, 4(4), pp. 245-255.

Ge, H., Walhout, A.J.M. and Vidal, M., 2003. Integrating 'omic' information: a bridge between genomics and systems biology. *TRENDS in genomics*, 19(10), pp. 551-560.

Hartwell, L.H., Hopfield, J.J., Leibler, S. and Murray. A.W., 1999. From molecular to modular cell biology. *Nature*, 402, December, pp. 47-52.

Ito, T., Chiba, T., Ozawa, R.,Yoshida, M., Hattori, M. and Sakaki, Y., 2001. Comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA*, 97, pp. 4569-4574.

Jiang, J.J. and Conrath, D.W., 1998. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of International Conference on Research in Computational Linguistics*, Taiwan, pp. 19-33.

Lin, D., 1998. An Information-Theoretic Definition of Similarity. In *Proceedings of the 15th international conference in machine learning*, Madison, Wisconsin, pp. 296-204.

Lubovac, Z., Gamalielsson, J., Olsson, B. and Lindlöf, A., 2005. Exploring protein networks with a semantic similarity measure across Gene Ontology. In *Proceedings of the 6^th International Symposium on Computational Biology and Genome Informatics*, USA, July 2005.

Lord, P.W., Stevens, R.D., Brass, A. and Goble, C.A., 2003a. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10), pp. 1275-1283.

Lord, P.W., Stevens, R.D., Brass, A. and Goble, C.A., 2003b. Semantic similarity measures as tools for exploring the Gene Ontology. *Pac Symp Biocomput*, pp. 601-612.

Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates T.O. and Eisenberg, D, 1999. Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. *Science,* 285, July, pp. 751-753.

Marcotte, E.M., Pellegrini, M., D.W., Thompson, M.J., Yeates, T.O., Eisenberg, D., 1999. A combined algorithm for genome-wide prediction of protein function*. Nature,* 402, pp. 83-86.

Piatt, J.H., 2004. *Sensitivity, Specificity, Likelihood Ratios, and ROC Curves, with a Few Neurosurgical Applications* [online]. Available from: http://www.drexel.edu/med/neurosurgery/ped/tests_talk.pdf [Accessed 13 March 2005].

Poyatos, J.F. and Hurst, L.D., 2004. How biologically relevant are interaction-based modules in protein networks?. *Genome Biology,* 5(11), pp. R93.

Reece, R.J., 2004. *Analysis of Genes and Genomes*. John Wiley & Sons Ltd, West Sussex.

Resnik, P., 1999. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, 11, pp. 95-130.

Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U. and Eisenberg, D., 2004. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research*, 32(Database issue), pp. D449-D451.

Salwinski, L. and Eisenberg, D., 2003. Computational methods of analysis of protein-protein interactions. *Curr Opin Struct Biol.*, 13(3), pp. 377-382.

Speer, N., Spieth, C. and Zell, A., 2004. A Memetic Clustering Algorithm for the Functional Partition of Genes Based on the Gene Ontology. In *Proceedings of the*

References

2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, La Jolla, California, pp. 252-259.

Sternberg, M.J.E. 1996. *Protein Structure Prediction.* Oxford University Press, Oxford.

Tape, T.G., 2003. *Interpreting Diagnostic Tests* [online]. Available from: http://gim.unmc.edu/dxtests/Default.htm [Accessed 13 March 2005].

Teichmann, S.A., Murzin, A.G. and Cothia, C., 2001. Determination of protein function, evolution and interactions by structural genomics. *Curr Opin Struct Biol.*, 11(3), pp. 354-363.

The Gene Ontology Consortium, 2001. Creating the gene ontology resource: design and implementation. *Genome Res.*, 11, pp. 1425-1433.

Tornow, S. and Mewes, H.W., 2003. Functional modules by relating protein interaction networks and gene expression. *Nucleic Acids Research*, 31(21), pp. 6283-6289.

Uetz, P. and Vollert, C.S., 2005. *Protein-Protein Interactions.* Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine. Springer Verlag.

Vinayagam, A., Konig, R., Moormann, J., Schubert, F., Eils, R., Glatting, K.H. and Suhai, S., 2004. Applying Support Vector Machines for Gene ontology based gene function prediction. *Bioinformatics,* 5(1), pp. 116-129.

Winters, M.S. and Day, R.A, 2003. Detecting Protein-Protein Interactions in the Intact Cell of *Bacillus subtilis* (ATCC 6633). *J. Bacteriol.*, 185(14), pp. 4268-4275.

Xenarios, I., Rice, D.W., Salwinski, L., Baron, M.K., Marcotte, E.M. and Eisenberg, D., 2000. DIP: The Database of Interacting Proteins. *Nucleic Acid Research*, 28(1), pp. 289-291.

Appendix A – Function predictions for CORE-CC06

The complete results from the function predictions for all of the proteins of unknown function in the CORE-CC06 dataset.

Table A.1: The results from function predictions for all proteins of unknown function in the CORE-CC06 dataset. In the first column the name of the protein and the number of neighbours it has is displayed. In the second column the neighbour or neighbours with the highest connectivity are presented. The predicted function of the proteins in the first column is displayed in the third column and the frequency and *p*-value for the term in column three are displayed in columns four and five respectively.

Node **(# neighbours)**	**Neighbour with highest connectivity** **(# neighbours)**	**Predicted function**	**Frequency**	***p*-value**
CAF40 (10)	CDC39 (19) POP2 (16) -- CCR4 (15)	GO:0000175 3’-5’-exoribonuclease activity -- GO:0030528 transcription regulator activity	31.6% 25.0% -- 26.7%	0.0023 -- 0.0518
MPE1 (7)	PFS2 (14) CFT2 (13) PAP1 (12)	GO:0003723 RNA binding	50.0% 61.5% 66.7%	0.0488
CTF8 (6)	RFC4 (10) RFC2 (10) RFC5 (10)	GO:0003689 DNA clamp loader activity	30.0% 40.0% 40.0%	0.0007
BNI5 (5)	CDC12 (14) GIN4 (10)	GO:0005200 structural constituent of cytoskeleton	28.6% 50.0%	0.0064
ERV25 (5)	SEC23 (14) SEC13 (12) SEC24 (10)	GO:0005198 structural molecule activity	28.6% 58.3% 30.0%	0.0460
PEP5 (5)	PEP3 (7) -- VPS16 (7)	GO:0005484 SNAP receptor activity -- GO:0005085 guanyl- nucleotide exchange factor activity	28.6% -- 42.9%	0.0038 -- 0.0029
RAV1 (5)	SKP1 (24)	GO:0008324	16.7%	0.0202

	TFP1 (12)	cation transporter activity	33.3%	
	VMA8 (11)		54.5%	
NAS6 (5)	RPT3 (29)	GO:0004175	37.9%	0.0114
	RPT1 (24)	endopetidase activity	57.9%	
	RPN10 (21)		71.4%	
SEC28 (5)	SEC22 (18)	GO:0005478	16.7%	0.0046
	BOS1 (17)	intracellular transporter activity	17.6%	
	BET1 (15)		40.0%	
SGF29 (4)	TAF6 (20)	GO:0030528	65.0%	0.0518
	NGG1 (17)	transcription regulator activity	82.4%	
	SPT7 (17)		70.6%	
MAK11 (4)	ERB1 (25)	GO:0003723	28.0%	0.0488
	TIF6 (21)	RNA binding	28.6%	
	NOP2 (15)		46.7%	
	NOP7 (15)		20.0%	
ELG1 (4)	RFC4 (10)	GO:0003689	30.0%	0.0007
	RFC2 (10)	DNA clamp loader activity	40.0%	
	RFC5 (10)		40.0%	
CWC23 (4)	PRP8 (17)	GO:0003723	52.9%	0.0488
	CEF1 (15)	RNA binding	60.0%	
	PRP43 (15)		60.0%	
HUA1 (4)	RVS167 (42)	GO:0005515	26.2%	0.0529
	SLA1 (21)	protein binding	28.6%	
	YSC84 (11)		63.6%	
MPP10 (4)	PWP2 (43)	GO:0003676	51.2%	0.0899
	KRR1 (23)	nucleic acid binding	60.9%	
GFD1 (3)	NUP42 (26)	GO:0005198	26.9%	0.0460
		structural molecule activity		
YBP1 (3)	NUP116 (39)	GO:0005198	25.6%	0.0460
	NUP100 (27)	structural molecule activity	37.0%	
YBL046W (3)	SPT5 (9)	RNA polymerase II transcription elongation factor activity	44.4%	0.0031
COG6 (3)	SED5 (18)	GO:0005485	38.9%	0.0020
	COG3 (10)	v-SNARE activity	30.0%	
ATG3 (3)	ATG12 (11)	GO:0016887	36.4%	0.0230
		ATPase activity		

Appendix A – Function predictions for CORE-CC06

APS2 (3)	APL1 (4)	GO:0030528 transcription regulator activity	50.0%	0.0518
APM4 (3)	APL1 (4)	GO:0030528 transcription regulator activity	50.0%	0.0518
LRP1 (3)	DIS3 (14) RRP4 (10)	GO:0000175 3’- 5’-exoribonuclease activity	71.4% 50.0%	0.0023
COG1 (3)	SED5 (18) COG3 (10)	GO:0005485 v-SNARE activity	38.9% 30.0%	0.0020
COG4 (3)	SED5 (18) COG3 (10)	GO:0005485 v-SNARE activity	38.9% 30.0%	0.0020
KEL2 (3)	CDC28 (95) -- KIN2 (10)	GO:0004672 protein kinase activity -- GO:0004674 protein serine/threonine kinase activity	17.9% -- 20.0%	0.0193 -- 0.0110
PEP8 (3)	VPS35 (3)	all unknown		
VPS35 (3)	PEP8 (3)	all unknown		
SNO1 (3)	SNZ3 (6)	GO:0005515 protein binding	50.0%	0.0529
YBR108W (2)	RVS167 (42) RVS161 (12)	GO:0005515 protein binding	31.0% 25.0%	0.0529
SNO3 (2)	SNZ3 (6)	GO:0005515 protein binding	50.0%	0.0529
MSI1 (2)	RLF2 (5)	GO:0030528 transcription regulator activity	40.0%	0.0518
CAC2 (2)	RLF2 (5)	GO:0030528 transcription regulator activity	40.0%	0.0518
MAD2 (2)	CDC20 (12)	GO:0005515 protein binding	41.7%	0.0529
YIP3 (2)	YIF1 (14)	GO:0003924 GTPase activity	71.4%	0.0075
UPF3 (2)	NAM7 (13)	GO:0003676 nucleic acid binding	69.2%	0.0899

SNO2 (2)	SNZ3 (6)	GO:0005515 protein binding	50.0%	0.0529
TVP23 (2)	YIP5 (14) YIP4 (14)	GO:0003924 GTPase activity	64.3% 64.3%	0.0075
YNL056W (2)	SIW14 (2) OCA1 (2)	no terms found		
NIS1 (2)	NAP1 (14) -- RIM11 (14)	GO:0004672 protein kinase activity -- GO:0016301 kinase activity	21.4% -- 21.4%	0.0193 -- 0.0305
ERV46 (2)	ERO1 (7)	GO:0016758 transferase activity, transferring hexosyl groups	28.6%	0.0119
SRL3 (2)	CDC28 (95) CKS1 (12)	GO:0004693 cyclin-dependent protein kinase activity	10.5% 58.3%	0.0013
NOP16 (2)	ERB1 (25) TIF6 (21)	GO:0003723 RNA binding	28.0% 28.6%	0.0488
ERV41 (2)	ERO1 (7)	GO:0016758 transferase activity, transferring hexosyl groups	28.6%	0.0119
RSC6 (2)	ISW1 (17)	GO:0016887 ATPase activity	17.6%	0.0230
SGF73 (2)	TAF5 (20) SPT7 (17)	GO:0030528 transcription regulator activity	50.0% 70.6%	0.0518
PCI8 (2)	RPG1 (14) PRT1 (11)	GO:0003743 translation initiation factor activity	42.9% 36.4%	0.0049
VPS8 (2)	PEP3 (7) -- VPS16 (7)	GO:0005484 SNAP receptor activity -- GO:0005085 guanyl- nucleotide exchange factor	28.6% -- 42.9%	0.0038 -- 0.0029

		activity		
BET5 (2)	TRS20 (8)	no terms found		
	BET3 (6)			
VPH2 (2)	VPH1 (11)	GO:0046961	81.8%	0.0030
		hydrogen-transporting ATPase activity, rotational mechanism		
SEC5 (2)	EXO84 (8)	GO:0005515	62.5%	0.0529
		protein binding		
RSC2 (2)	ISW1 (17)	GO:0016887	23.5%	0.0230
	NHP10 (12)	ATPase activity	25.0%	
RSC58 (2)	ISW1 (17)	GO:0016887	23.5%	0.0230
	P89501 (10)	ATPase activity	40.0%	
RCO1 (2)	SIN3 (21)	GO:0005515	19.0%	0.0529
	RPD3 (16)	protein binding	25.0%	
RXT2 (2)	SIN3 (21)	GO:0005515	19.0%	0.0529
	RPD3 (16)	protein binding	25.0%	
VPS17 (2)	VPS35 (3)	all unknown		
	PEP8 (3)			
YBR267W (2)	LSG1 (6)	GO:0003676	33.3%	0.0899
		nucleic acid binding		
LSB5 (2)	LAS17 (31)	GO:0005515	38.7%	0.0529
	SLA1 (21)	protein binding	28.6%	
TRS130 (2)	TRS20 (8)	no terms found		
	BET3 (6)			
TRS120 (2)	TRS20 (8)	no terms found		
	BET3 (6)			
RRP14 (2)	ZDS2 (22)	GO:0005515	18.2%	0.0529
	--	protein binding	--	--
	GIS1 (11)	--	18.2%	0.0075
		GO:0003924		
		GTPase activity		
APL3 (2)	APM4 (3)	all unknown function		
	APS2 (3)			
RNQ1 (2)	RFC4 (10)	GO:0003689	30.0%	0.0007
	RFC3 (8)	DNA clamp loader activity	37.5%	
SCO2 (2)	COX2 (3)	GO:0008379	66.7%	0.0018
		thioredoxin		

		peroxidase activity		
YJR115W (2)	RVS167 (42)	GO:0005515	31.0%	0.0529
	RVS161 (12)	protein binding	25.0%	
YHR115C (2)	YNL311C (5)	no terms found		
AVO2 (2)	LST8 (9)	GO:0005515	33.3%	0.0529
	TOR2 (8)	protein binding	37.5%	
YJL149W (2)	SKP1 (24)	GO:0005515	37.5%	0.0529
	CDC53 (11)	protein binding	54.5%	
KAR5 (2)	SMC3 (8)	GO:0016887	25.0%	0.0230
	SMC2 (7)	ATPase activity	28.6%	
SCO1 (2)	COX2 (3)	GO:0008379	66.7%	0.0018
		thioredoxin peroxidase activity		

Appendix B – Function predictions for CORE-TEST

The complete results from the function predictions for all of the 30 proteins in the CORE-TEST dataset.

Table B.1: The results from function predictions for all proteins of in the CORE-TEST dataset. In the first column the name of the protein and the number of neighbours it has is displayed along with its functional annotation found in SGD. In the second column the neighbour or neighbours with the highest connectivity are presented. The predicted function of the proteins in the first column is displayed in the third column and the frequency and *p*-value for the term in column three are displayed in columns four and five respectively.

Node **(# neighbours)** **Function**	**Neighbour with highest connectivity (# neighbours)**	**Predicted function**	**Frequency**	***p*-value**
LSM8 (33)	SMD3 (23)	GO:0003723	91.3%	0.0488
GO:0003723	LSM3 (18)	RNA binding	77.8%	
	LSM2 (17)		64,7%	
	PRP8 (17)		52.9%	
	PRP31 (17)		52.9%	
RPT1 (24)	RPT3 (29)	GO:0004175	37.9%	0.0114
GO:0004175	PRE1 (26)	endopetidase activity	84.6%	
	RAD23 (23)		47.8%	
UTP7 (19)	RPT3 (29)	GO:0004175	37.9%	0.0114
GO:0030515	PRE1 (26)	endopetidase activity	84.6%	
	RAD23 (23)		47.8%	
RPG1 (14)	TIF6 (21)	GO:0017111	23.8%	0.0001
GO:0003743	SUA7 (17)	nucleoside- triphosphatase activity	--	--
	TIF5 (17)		41.2%	0.0049
		--	52.9%	
		GO:0003743		
		translation initiation factor activity		
SIF2 (13)	SRP1 (56)	GO:0005515	12.5%	0.0529
GO:0017136	KAP95 (23)	protein binding	26.1%	
GO:0045129	ZDS2 (22)		18.2%	
OST1 (13)	PKC1 (13)	GO:0004579	30.8%	0.0013
GO:0004579	SEC13 (12)	dolichyl-diphosphooligosaccharide-protein glycotransferase activity	-- 58.3%	-- 0.0460
		--		
		GO:0005198		
		structural molecule activity		

Appendix B – Function predictions for CORE-TEST

SMD2 (12)	SMD3 (23)	GO:0003723	91.3%	0.0488
GO:0031202	LSM2 (17)	RNA binding	64.7%	
	LSM7 (16)	(parent of 31202)	75.0%	
SUI1 (12)	RPT6 (19)	GO:0004299	42.1%	0
GO:0003743	TIF4631 (19)	endopeptidase activity	--	--
	TIF5 (17)	--	21.5%	0.0049
		GO:0003743	52.9%	
		translation initiation factor activity		
CDC20 (12)	CDC28 (95)	GO:0004672	17.9%	0.0193
GO:0008047	CCT2 (19)	protein kinase activity	--	--
	TCP1 (17)	p: 0,0193	15.8%	0.0044
		--	11.8%	
		GO:0004722		
		protein serin/threonine phosphatase activity		
VPH1 (11)	TFP1 (12)	GO:0046961	33.3%	0.0031
GO:0046961	VMA8 (11)	hydrogen-transporting ATPase activity, rotational mechanism	54.5%	
	VMA4 (10)		50.0%	
	VMA6 (10)		30.0%	
LSM6 (11)	LSM8 (33)	GO:0003723	51.5%	0.0488
GO:0003723	SMD3 (23)	RNA binding	91.3%	
	LSM3 (18)		77.8%	
IKI3 (10)	RPO21 (27)	GO:0003899	29.6%	0.0040
GO:0016944	RPB3 (13)	DNA-directed RNA polymerase activity	61.5%	
GO:0004402				
ZDS1 (10)	PPH22 (26)	GO:0004299	50.0%	0
GO:0005515	ZDS2 (22)	endopeptidase activity	--	--
	KAP95 (23)	--	18.2%	0.0529
		GO:0005515	26.1%	
		protein binding		
RPB7 (10)	RPB3 (13)	GO:0003899	61.5%	0.0040
GO:0003899		DNA-directed RNA polymerase activity		
VAM3 (9)	VTI1 (10)	GO:0005484	70.0%	0.0038
GO:0005486		SNAP receptor activity (parent of 5486)		
STV1 (9)	TFP1 (12)	GO:0046961	33.3%	0.0031
GO:0046961	VMA8 (11)	hydrogen-transporting ATPase activity, rotational mechanism	54.5%	
	VMA4 (10)		50.0%	
	VMA6 (10)		30.0%	
GCD6 (7)	GCN3 (16)	GO:0003743	43.8%	0.0049
GO:0003743		translation initiation factor activity		

Appendix B – Function predictions for CORE-TEST

HST1 (7)	SIF2 (13)	GO:0045129	38.5%	0.0013
GO:0017136		NAD-independent histone deacetylace activity		0.0014
GO:0045129				
		GO:0017136		
		NAD-dependent histone deacetylace activity		
HSH155 (7)	CEF1 (15)	GO:0003723	60.0%	0.0488
GO:0003729	PAT1 (14)	RNA binding	71.4%	
	SMX2 (13)		84.6%	
	SMX3 (13)		76.9%	
SWP1 (7)	OST1 (13)	GO:0004579	61.5%	0.0013
GO:0004579	PKC1 (13)	dolichyl-diphosphooligosaccharide-protein glycotransferase activity	30.8%	
PRP24(6)	LSM8 (33)	GO:0003723	51.5%	0.0488
GO:0031202	LSM3 (18)	RNA binding	77.8%	
	LSM2 (17)	(parent of 31202)	82.4%	
RET1 (5)	RPC40 (27)	GO:0003899	59.3%	0.0040
GO:0003899	RPO31 (18)	DNA-directed RNA polymerase activity	77.8%	
	RPC34 (11)		91.0%	
SWI3 (5)	SNF5 (9)	GO:0016251	44.4%	0.0079
GO:0016251		general RNA polymerase II transcription regulator activity		
ELP2 (5)	IKI3 (10)	GO:0003899	60.0%	0.0040
GO:0016944		DNA-directed RNA polymerase activity		
VMA5 (4)	VMA8 (11)	GO:0046961	54.5%	0.0031
GO:0046961	VPH1 (11)	hydrogen-transporting ATPase activity, rotational mechanism	81.8%	
SPC98 (4)	TUB4 (9)	GO:0015631	44.4%	0.0018
GO:0005200		tubulin binding		
TOP1 (4)	RPD3 (16)	GO:0004407	18.8%	0.0051
GO:0003917		histone deacetylase activity		
TFC4 (3)	RPC53 (5)	GO:0003899	80.0%	0.0040
GO:0003709	BRF1 (5)	DNA-directed RNA polymerase activity	--	--
			80.0%	0.0013
		--		
		GO:0003709		
		RNA polymerase III transcription regulator activity		
RAD9 (3)	CDC28 (95)	GO:0005515	14.7%	0.0529

GO:0005515	RAD53 (19)	protein binding	21.1%	
	CHK1 (10)		30.0%	
INO2 (2) GO:0003704	INO4 (11)	GO:0003704 specific RNA polymerase II transcription factor activity	18.2%	0.0074