Towards a better Understanding of Protein-Protein Interaction Networks

by

Tatiana A. Gutiérrez-Bunster
B.Sc., University of Bío Bío, 2001
M.Sc., University of Concepción, 2008

A Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Computer Science

© Tatiana A. Gutiérrez-Bunster, 2014
University of Victoria

All rights reserved. This dissertation may not be reproduced in whole or in part, by photocopying or other means, without the permission of the author.

Towards a better Understanding of Protein-Protein Interaction Networks

by

Tatiana A. Gutiérrez-Bunster
B.Sc., University of Bío Bío, 2001
M.Sc., University of Concepción, 2008

Supervisory Committee

Dr. Ulrike Stege, Co-Supervisor
(Department of Computer Science)

Dr. Alex Thomo, Co-Supervisor
(Department of Computer Science)

Dr. John Taylor, Co-Supervisor
(Department of Biology)

Dr. Chris Upton, Outside Member
(Department of Biochemistry)

Supervisory Committee

Dr. Ulrike Stege, Co-Supervisor
(Department of Computer Science)

Dr. Alex Thomo, Co-Supervisor
(Department of Computer Science)

Dr. John Taylor, Co-Supervisor
(Department of Biology)

Dr. Chris Upton, Outside Member
(Department of Biochemistry)

ABSTRACT

Proteins participate in the majority of cellular processes. To determine the function of a protein it is not sufficient to solely know its sequence, its structure in isolation, or how it works individually. Additionally, we need to know how the protein interacts with other proteins in biological networks. This is because most of the proteins perform their main function through interactions. This thesis sets out to improve the understanding of protein-protein interaction networks (PPINs). For this, we propose three approaches:
(1) *Studying measures and methods used in social and complex networks.*
The methods, measures, and properties of social networks allow us to gain an understanding of PPINs via the comparison of different types of network families. We studied and evaluated models that describe social networks to see which models are useful in describing biological networks. We investigate the similarities and differences in terms of the network community profile and centrality measures.
(2) *Studying PPINs and their role in evolution.*
We are interested in the relationship of PPINs and the evolutionary changes between species. We investigate whether the centrality measures are correlated with the variability and similar-

ity in orthologous proteins.
(3) *Studying protein features that are important to evaluate, classify, and predict interactions.* Interactions can be classified according to different characteristics. One of these characteristics is the energy (that is the attraction or repulsion of the molecules) that occurs in interacting proteins. We identify which type of energy values contributes better to predicting protein-protein interactions. We argue that the number of energetic features and their contribution to the interactions can be a key factor in predicting transient and permanent interactions.

Contributions of this thesis include: (1) We identified the best community sizes in PPINs. This finding will help to identify important groups of interacting proteins in order to better understand their particular interactions. We furthermore find that the generative model describing biological networks is very different from the model describing social networks A generative model is a model for randomly generating observable data. We showed that the best community size for PPINs is around ten, very different from the best community size for social and complex network (around 100). We revealed differences in terms of the network community profile and correlations of centrality measures; (2) We outline a method to test correlation of centrality measures with the percentage of sequence similarity and evolutionary rate for orthologous proteins. We conjecture that a strong correlation exists. While not obtaining positive results for our data, we believe that the reason for this is the integration problem of today's data sets. Therefore, (3) we investigate a method to discriminate energetic features of protein interactions that in turn will improve the PPIN data. The use of multiple data sets makes possible to identify the energy values that are useful to classify interactions. For each data set, we performed Random Forest and Support Vector Machine with linear, polynomial, radial, and sigmoid kernels. The accuracy obtained in this analysis reinforces the idea that energetic features in the protein interface help to discriminate between transient and permanent interactions.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I would like to thank:

Germán, porque sin tí, estos esfuerzos no valen la pena. TAM.

Mi gran familia, por hacerse cargo the Almendra y Aserrín, mientras estudio en el otro polo. Y por estar siempre para nosotros, incluso a la distancia.

Ulrike, Alex and John, for guiding me in this process of ups and downs, while giving me support and encouragement. Thank you so much.

Conicyt, Chile, por financiarme con una beca para realizar este doctorado.

Universidad del Bío-Bío, Chile, por confiar en mí y permitirme el tiempo para perfeccionarme.

Tatiana
Fall, 2014

To my everything,
Germán.

Chapter 1

Introduction

Why is it important to study proteins and protein-protein interaction networks (PPINs)?

Proteins participate in the majority of cellular processes. Moreover, the functions of proteins are specific to each and the proteins allow cells maintain their integrity, to defend against external agents, to repair damage, as well as control and regulate functions. It is not possible to determine the function of a protein knowing only its sequence or structure in isolation, or how it works individually. This is because most of the proteins perform their main function through interactions. Therefore, we need to know which proteins interact and under which conditions. For this reason characterizing the partners is crucial to understand the functional role of individual proteins and the organization of the entire biological processes. Recent progress in technology has made possible the gathering of an improved and increased amount of data such as sequence data and gene or protein interactions. This leads to the complex process of analyzing data to discover structures and functions of genes and proteins.

A long-term goal of this research is to contribute towards a better understanding of protein networks through the improvement of the analysis of existing data using network analysis. The analysis of networks using different methods permits identifying important characteristics of proteins and their interactions. Network measures contribute to the possibility of inferring or predicting functions and interactions of similar proteins in different species. PPINs provide a valuable framework to understand the functional organization of the proteome. This permits the comparison of networks coming from different species and the prediction of interaction behaviors that could be useful for better understanding of evolution, diseases, and functions.

The main goal of this thesis is to improve the understanding of protein-protein interaction networks. For this, we propose three approaches:

(1) *Studying measures and methods used in social and complex networks.*

The methods, measures, and properties of social networks allow us to gain an understanding of PPINs via the comparison of different types of network families. We studied and evaluated

models that describe social networks to see which models are useful in describing biological networks. We investigate the similarities and differences in terms of the network community profile and centrality measures.
(2) *Studying PPINs and their role in evolution.*
We are interested in the relationship of PPINs and the evolutionary changes between species. We investigate whether the centrality measures are correlated with the variability and similarity in orthologous proteins.
(3) *Studying protein features and their importance in the classification of interactions.*
We identify which types of energies contribute better to predict protein-protein interactions. We argue that the number of energetic features and their contribution to the interactions can be a key factor in predicting transient and permanent interactions.

We worked with unweighted networks, which means that every connection or interaction is assumed to have the same value or relevance in the network. The decision to use unweighted networks was due to not having enough information about the interactions for all the networks of the species that we are using. We think that point (3) will help improve the quality of the PPINs (see Figure 1.1). This improvement will consist of adding more information to the networks, specifically to the edges (interactions). As a consequence, the outcomes would be different and more meaningful from those obtained with unweighted networks, in turn improving the quality of the outcomes of point (1) and point (2).

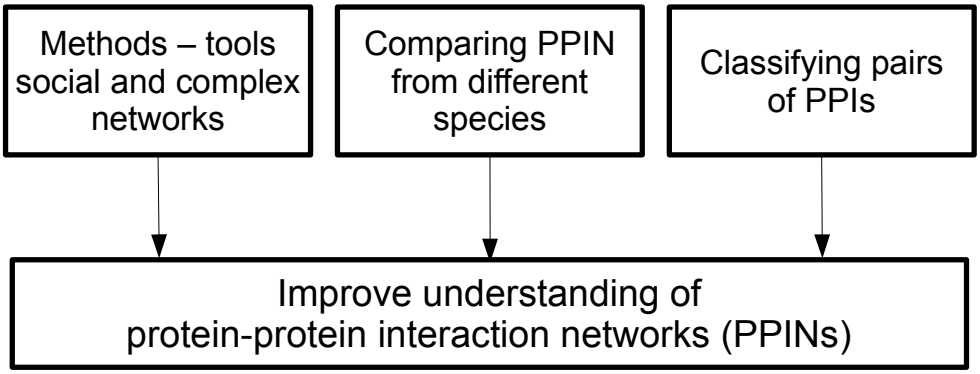


Figure 1.1: Main goal.

1.1 Research questions

Despite the large amount of information available, the search for better understanding of proteins and their interactions continues. The information available is growing with each conducted research [13, 110]; these projects contribute to the knowledge about proteins and their

interactions. As the number of research projects and the amount of data grows, an overlap of information is produced. This overlap helped create reliable PPINs or validate already existing ones. It also allows us to do a prediction of interactions without *in-vivo* experiments, thus reducing the period of time for the study and classification of interactions.

However, the use of data from different projects presents us with a challenge: the studies are done in different scenarios (see Appendix A). Examples of these scenarios are the methods used for interaction identification, data used for protein classification, and data formats used to publish results. Nevertheless, the data available is still useful for studying and learning about proteins. In this thesis, we validate our results by considering data sets for the same species, but coming from different source.

Our research is focused mainly on the study of protein-protein interaction networks. More exactly, our goal is to understand the differences and similarities of proteins from different networks for different species. For this we will use methods from Social Network Analysis, Evolutionary Analysis, and Machine Learning.

Our aim is that by comparing proteins and their PPINs using different measures, we can identify patterns in proteins and networks with relevant features and parameters, which will allow us to provide new insights on the way the proteins interact. We study PPINs of different species in order to identify those patterns that allow us to understand how proteins interact in a specific way independently of which species the proteins belong to.

We use different parameters, such as the network topology, orthologous protein relationships, centrality measures, sequence similarities, and protein features. The idea is to use multiple parameters for further analysis and more robust results.

Comparing PPI and PPINs allows us to make predictions about proteins and their interactions. One way is to do this is to determine the areas (subnetworks) of the network or specific proteins in the network with high similarities. Thanks to such patterns, we expect to be able to predict interactions. For example, consider two networks (see Figure 1.2) where one of them is well known (left, species 1) and the other one is not (right, species 2). We use the knowledge of one network to find new information about the other network, following similar patterns in both networks. In the case of having available only one network, it is possible to analyze the protein interactions to obtain good predictions of new interactions (see Figure 1.3).

To reach our research goal of better understanding the PPINs, we set out to ask the following research questions:

1. Which social network analysis methods are useful to analyze PPINs?

   There are many measures used in social networks, such as, centrality measures, topological, and conductance (measures how strong and how connected the graph is). For

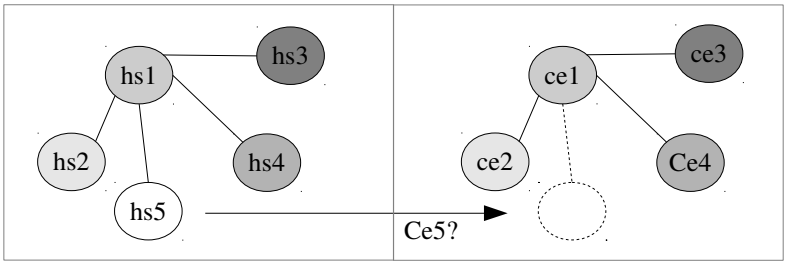


Figure 1.2: Prediction using two different Networks.

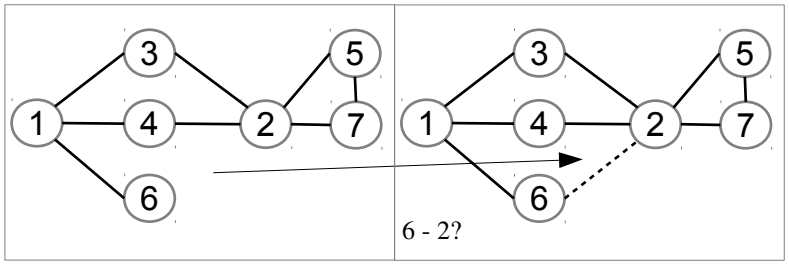


Figure 1.3: Prediction using one Network.

some of them, such as conductance for example, there are no specific studies where biological networks, or specifically protein-protein interaction networks, are the main point of study. The goal of using conductance combined with spectral algorithms is to identify which are the best subnetworks (or communities) in PPINs. We further ask if it is possible to identify differences between PPINs and social networks using a "Best-Community Analysis" type of investigation.

2. Are centrality measures (closeness, betweenness and degree), percentage of similarity, and amino acid divergence correlated for any given protein over (evolutionary) time?

   We focused on the study of the evolution of protein-protein interactions (PPIs).

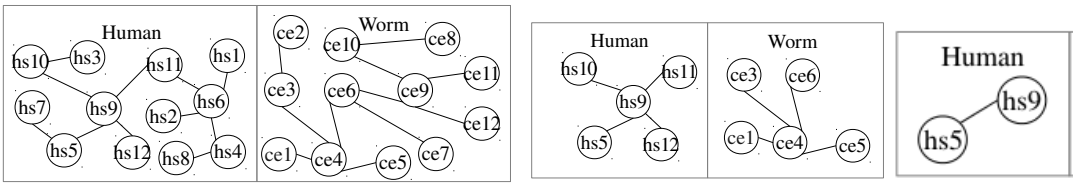
   We think that is not efficient to use centrality measures as a method to identify orthologs. It meight be possible that certain proteins could have similar centrality values but in fact they are not related at all. Instead, first we use their sequence data to do the ortholog identification. After that we can study the relations of proteins according to their centrality values in the networks.

   Some more specific questions that we address in this thesis are: Which are the protein differences between species at the amino acid sequence level?. What is the percentage of protein similarity between very different species and close species? Are the centrality values of orthologous proteins similar to each other?

3. Is it possible to improve the selection of protein characteristics in order to discriminate

between types of protein-protein interactions regarding the duration of the interaction?

Central to this are the energetic features in the surfaces of the interacting proteins that allow the discrimination between permanent and transient protein-protein interactions (different time duration). A more specific question we address is: Which specific energetic features are better predictors (with higher classification accuracy) for these types of interactions?



(a) PPINs, two species represented by graphs. (b) Subgraphs of PPIs. (c) Two proteins.

Figure 1.4: Representation of PPINs, PPIs and proteins comparison.

We study protein-protein interactions at three different levels.

- First, at the level of PPINs (species – see Figure 1.4a), we compare networks of different species (Research question 2).
- Second, at the level of subnetworks (sets of PPIs – see Figure 1.4b), we consider parameters such as the number of shared subnetworks between networks, and conductance measures to evaluate the sizes of the subnetworks in order to identify differences between PPIN and social networks (Research question 1).
- Third, at the level of proteins (see Figure1.4c), we compare individual proteins in different species (Research question 2) in order to evaluate the correlation between their centrality measures and sequence similarity (see Figure1.4c: $hs9$ and $ce3$). We also classify interactions between two proteins (Research question 3) in order to predict new interactions (see Figure1.4c: $hs9 - hs5$).

1.2 Methodology

In this section, we introduce the methodology developed to study the three research questions proposed in the previous section. This methodology is depicted in Figure 1.5. The three phases correspond to the "Data Collection" section, followed by the "Research Question" section, and "Interpretation of the data" section. Each phase is explained in detail as follows.

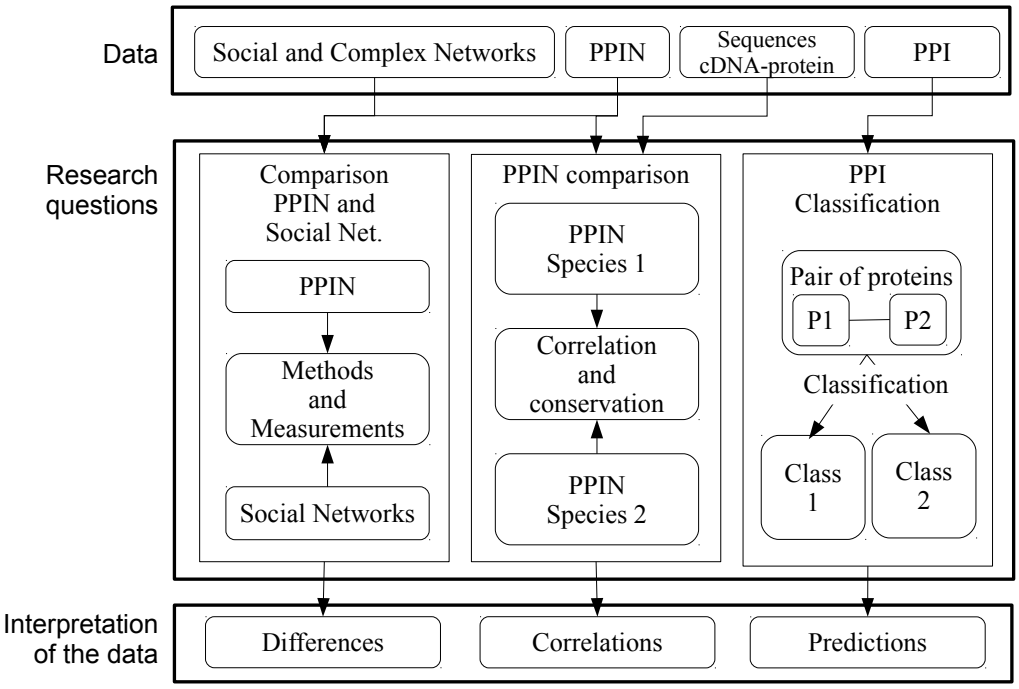


Figure 1.5: Phases for the selection of energetic features and validation of efficiency in the classification.

Methodology: Data Collection

In the data phase (Figure 1.5 – top box) we gather all the information and data necessary for the study of our research questions.

First, we gather a list of social networks to be used as a tool for comparison and validation for the best-community research.

Second, we collect a set of PPI networks from different species. We collect different species, in some cases we have species with more than one network because they come from different sources. This data will be used for research question 2 and 3 (see Figure 1.5 comparison PPIN and social networks, and PPIN comparison).

Third, we collect information about sequences of all the proteins and cDNA from the different species that we are using.

Fourth, we gather a list of protein-protein interactions to be used to predict protein interaction types. For this part, we use interactions that are already classified. In this way, we can train our methods and validate them.

In the following we outline our methodology on how we address our main research questions.

Methodology: Research Questions

- *Research question 1. Differences between PPIN and social and complex networks.*

  These classes of networks exhibit differences related to the size of the community (subnetworks). Here we use measures from Graph Theory and Social Networks to determine the differences. This is done using an algorithm to obtain subnetworks from different sizes and evaluate each of them using the eigenvalues of the subnetwork with respect to the whole network. We perform this analysis for both families of networks, PPINs and social and complex networks.

- *Research question 2. PPIN comparisons.*

  We conduct a study of proteins that present a behaviour that could be related to proteins from other species. We investigate how to identify the relevance of a protein in the network and its neighbors. For this, we combine different data sets to be able to do a crossing of the data obtained and obtain more information about proteins in different species.

- *Research question 3. PPI type classification.*

  Here, we focus on the study and analysis of energetic features of protein interactions to predict two types of interactions related to the time length of the interaction. We start with the creation of the database to be used and extract relevant features. Next, we proceed to a detailed feature selection and construction of robust Machine Learning classifiers. Lastly, we perform a thorough validation using different sizes of training and test data sets.

Methodology: Interpretation of the data

Our results from the investigation of the three research questions are as follows.

Differences between PPIN and social and complex networks. After a detailed study of community structure in different families of PPINs and social and complex networks using advanced, state-of-the-art network tools, we conclude that the best community sizes for different families are vastly different. Surprisingly, the best community size for PPINs is about ten, which is an order of magnitude smaller than the values for the other network families. Furthermore, we observe that the generating community models for the different families we study are also quite different.

PPIN comparison. We identify orthologus proteins from different pair of species and we compare their percentage of similarity, centrality measures, and evolutionary rate to

identify some patterns that could help understand the evolution of these proteins. What we found is that there is no pronounced correlation between network measures and evolutionary rate of species. In other words, the well preserved proteins over evolutionary time showed to have a variety of centrality values (low and high). While this is a negative result, we believe it is nevertheless interesting because it is contrary to the intuitive belief that network measures and evolutionary rate of proteins are correlated.

PPI classification. By considering numerous energetic features capturing the way the proteins interact in their interface with each other, we were able to build robust Machine Learning classifiers that achieved a high success rate in predicting the type (transient or permanent) of interaction between proteins. Namely, the accuracy we achieved was in the order of 87%, which is significantly better than the level achieved by previous works.

The work done in Chapter 3 and 5 have been published the conference proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014). And presented in the International Symposium on Network Enabled Health Informatics, Biomedicine and Bioinformatics 2014 (as a part of ASONAM 2014).

1.3 Thesis overview

The remainder of the thesis is as follows:

- Chapter 2 introduces different terminologies and definitions needed for a better understanding of the topics in this research. The chapter is divided into three sections: introduction to biological definitions (proteins and interactions), graph theory definition used for the analysis, and measures used in social networks.

- Chapter 3 exposes the community size differences between PPINs and social networks found using a spectral algorithm. Also are presented the similar centrality measure values for the different classes of networks.

- Chapter 4 presents the study of proteins relations between species (ortholog - conservation). This is done through comparisons on percentage of similarity, centrality measures and changes in the proteins at the amino acid level.

- Chapter 5 describe our analysis on protein-protein interactions to improve the feature selection to classify proteins according to the duration of the interactions.

- In Chapter 6 we summarize the contributions of the thesis and propose future work and open questions.

Chapter 2

Background

Our research focuses on protein-protein interaction networks. We introduce some terminology and concepts to the reader to facilitate the understanding of the present work. The chapter is divided in three sections: Proteins, interactions and networks; graph theory; and social networks.

The first section introduces terminology related to protein-protein networks. We start with background on proteins, the main focus of this research. We describe their main functions and how the proteins are structured to perform their function. Also, we describe their role in the networks. Lastly, we explaine the meaning of amino acid sequence divergence in molecular evolution.

The second section introduces the necessary graph theory definitions used for representation and subsequent analysis of the data. Here, we describe basic properties of a graph and the centrality measures that will be used to analyze the PPINs in two of the three approaches outlined. The centrality measures are betweenness, closeness and degree.

The last section introduces basics in social network. In particular, we explain the conductance which we use to analyze the research data in the first approach.

2.1 Proteins, interactions and networks

Today there is a close collaboration between computer science and biology. The study of different areas in biology has led to a large volume of data and information available especially in the area of genomics and genetics. Despite the large amount of sequence data and advances in experimental techniques to provide approximate models of the structure and dynamics of proteins (X-ray crystallography or nuclear magnetic resonance), each day the difference between the number of sequences and the number of known structures increases. Structure prediction methods aim to provide a model to conduct biological studies and provide a structural basis

for the interpretation of biological phenomena when there is no experimentally determined structure.

Most of the cell processes that support life involve interactions between genes. Proteins are encoded by genes [64]. Each gene has a unique position (or location) in the genome and a unique name, which is typically also given to the protein. This naming system, and the fact that individual genes can be cloned and expressed to generate pure protein solution, result in the fact that proteins are often studied in isolation. However, it is probably safe to say that no protein can function on its own. Even proteins known best for binding DNA, RNA, and non-protein ligands have protein binding partners. To understand species and their systems, it is not enough to identify their proteins, but also the interactions between these proteins in order to better comprehend the species. For this reason, researchers started to study known interactions and new ones [37, 94].

Notably, there is a huge amount of data obtained from sequencing projects. The data provides to the researchers with different levels of information about the species (e.g. DIP [105], BioGRID [30], HPRD [96], Genbank [11, 12], UniProt [111]).

Some of the main interests when studying proteins are how to computationally manipulate and explain the large amount of data generated from different sources. Also, the interactions of different fields, such as, biology, mathematics, computer science, and bioinformatics play an important role in creating models, algorithms, and methods to help describe, classify, analyze, and visualize data. In our research, we use different social network and graph theory tools to explain or interpret in a better way the data and results we obtain from our analysis of PPINs.

2.1.1 Proteins

Proteins determine the shape and structure of cells and control the majority of life processes. The functions of proteins are specific to each and allow cells to maintain their integrity, defend against external agents, repair damage, control and regulation functions, among others.

Proteins are molecules composed essentially of carbon, hydrogen and oxygen. They may also contain nitrogen, and certain types of proteins contain phosphorus, iron, magnesium and copper and other elements. Amino acids are characterized by having a carboxyl group (-COOH) and an amino group (-NH2). The other two parts of the carbon are saturated with an H atom and a radical group called variable R. A peptide bond is a covalent bond is established between the carboxyl group of an amino acid and the amino group of the next, resulting in the detachment of a molecule of water [62].

The bond between two amino acids results in a peptide; if the number of amino acids that form the molecule is not greater than 10, is called oligo-peptide, if it exceeds 10 is called

poly-peptide and if the number is more than 100 amino acids (approx.) it is referred to as a protein. Proteins have one or more long chains of amino acid residues. A protein chain could have a range of 50 to 2000 amino acid residues.

The organization of a protein is defined by four structural levels called: primary, secondary, tertiary and quaternary structure. Each of these levels gives the arrangement of the previous level in the space [77].

The structures are (see Figure 2.1 for representation):

- The primary structure: the polypeptide chain and the order in which these amino acids are found. The function of a protein depends on its sequence and the forms it takes.
- The secondary structure: this is the arrangement of the amino acid sequence in space. Amino acids, as they are being linked, acquire a stable spatial conformation, secondary structure. There are two types of structure: α–helix and β–sheet.
- The tertiary structure or three-dimensional structure: reports on the disposition of the secondary structure of a polypeptide to fold back on itself. This conformation remains stable thanks to the existence of links (intramolecular interactions) between the radical R of amino acids. Some examples of types of links are: the disulfide bridge between amino acid radicals having sulfur; the hydrogen bonds; the electrical bridges; and hydrophobic interactions [76].
- Quaternary structure: arrangement of multiple folded proteins unioned by weak bonds of several polypeptide chains with tertiary structure to form a protein complex.

All correct folding depends on whether a protein is able to form properly its structure. If the protein does not fold, it will not be able to fulfill its biological function. The study of the biological function of proteins and their interactions is closely related to the three-dimensional structure of a native protein, which is determined by the multiple interactions that occur between the amino acids forming the polypeptide chain. The three-dimensional structure of a protein under physiological conditions is considered the most stable of the possible structures.

A protein *domain* is a part of a given protein sequence and tertiary structure that can change, function, and exist independently of the rest of the protein chain [102]. The size of individual structural domains varies from 36 residues to 692 residues, with an average of approximately 100 residues. Many proteins only contain a single domain [119] (see Figure 2.2 for a three domain representation of a protein).

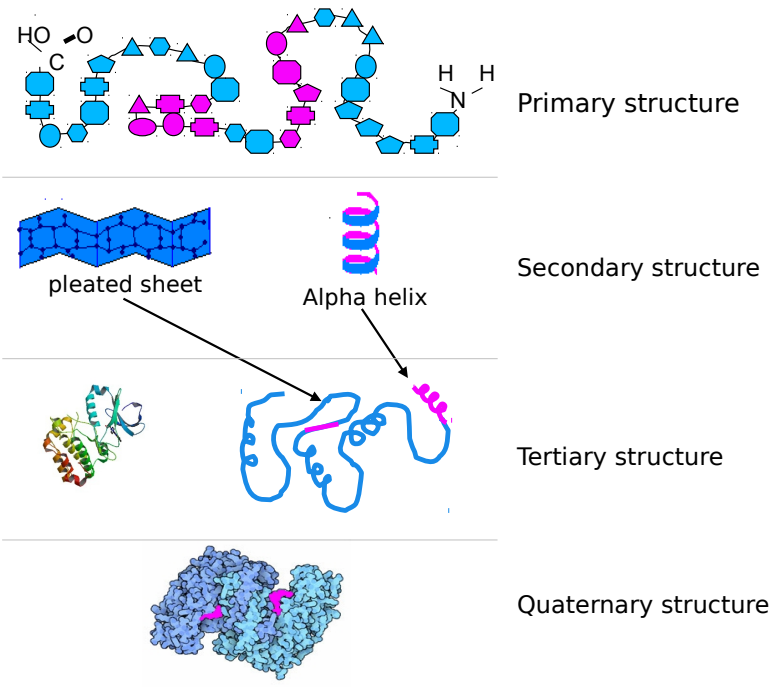


Figure 2.1: Representation of protein structures.

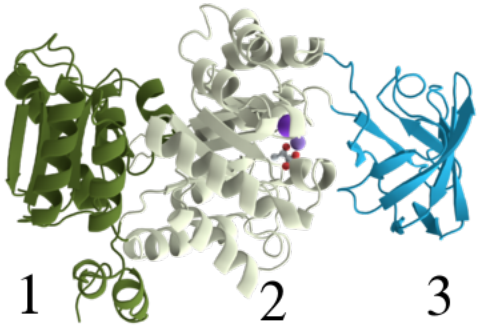


Figure 2.2: Three domains in protein 1pkn [13].

Properties of proteins

Some of the properties of proteins are: (1) Specificity, which means that each protein performs a particular function that is directed by its primary structure and spatial conformation. Any change in the protein structure may mean a loss of function; (2) Denaturation, which is the loss of tertiary structure, by breaking the bridges that form the structure. When a water-soluble protein is denatured, it becomes insoluble in water and precipitates. Denaturation may occur due to changes in temperature or pH variations. In some cases, the denatured proteins can return to their original state via a process called renaturation; (3) The influence of the type of residue and structure in the accessibility to the solvent. First it describes the analysis used

to determine the parameters for the prediction interface algorithm, which includes accessibility or structural information such as the interaction between beta structures folded or helical structures. These data can be obtained via the identification of surface regions involved in protein-protein interactions; (4) Solubility. This property is maintained as long as the strong and weak links are present. Increasing the temperature and pH, the solubility is lost; and (5) Electrolytic capacity is determined by electrolysis.

2.1.2 Protein Interactions

The interactions between atoms of the amino acids are subject to restrictions imposed by topological connectivity of the chain. Stabilizing interactions maintain the native structure formation. Destabilizing interactions interrupt the formation of the native structure and prevent the proteins from acquiring a structure that is incompatible with their biological function. The perfect balance between stabilizing and destabilizing interactions results in a native protein folding. In the presence of additional molecules or other proteins it is possible to form attractive and repulsive forces (energies) between molecules or proteins (interactions with other amino acid chains). These interactions may lead to the formation of intermolecular associations (interactions between molecules) or aggregates, such as: protein-protein interactions, protein-DNA interactions, protein-small molecules interactions [1], or protein-ligand interactions, among others. These interactions depend on the different circumstances, such as, temperature, pH, ionic strength, the entities involved, and the environment. The focus of our research is protein-protein interactions.

Despite knowing the structures of many proteins, there are no methods to predict protein-protein interactions (PPI) with high accuracy. These methods can not indicate how proteins interact with each other and in which way. For this reason, it is not possible to predict the stability of the interaction, since one cannot determine the function of a protein, knowing only the sequence or structure in isolation. It should be noted that there is still no full understanding of the folding of a protein. The structural information of proteins has not advanced as quickly as information about sequences and functions.

The technology has made it possible to study interactions between proteins using large scale experiments. However, the available information on protein interactions comes from studying three-dimensional structures of protein complexes with *in-vivo* interaction experiments (techniques performed in a living organism and stored in PDB [13]). Recall that protein-protein interactions (PPI) happen when two or more proteins bind together to carry out their biological function in a protein-protein interaction network.

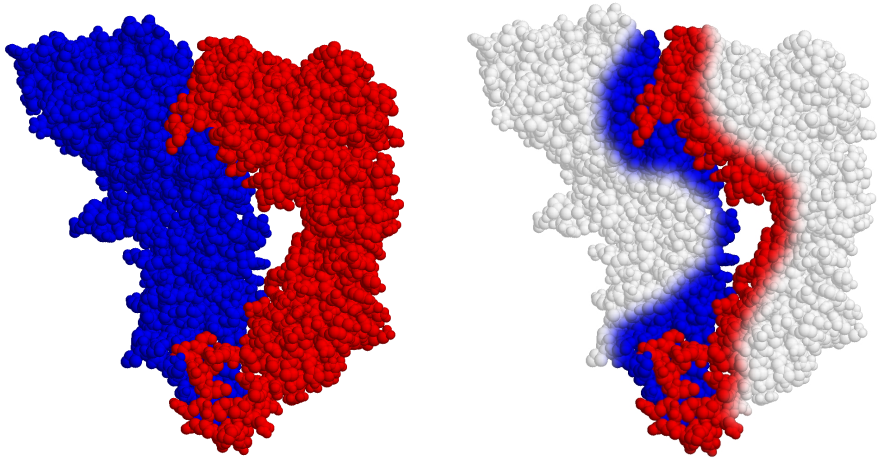
There are different methods used to identify PPIs. The first method used was co-immu-

noprecipitation [28]. After that, the yeast-2-hybrid assay (Y2H) [28] made it possible to investigate interactions between pairs of proteins or protein domains. Since 2000, mass spectrophotometry (MS) [36] has been the most common way to study PPIs. This tool allows even higher throughput than Y2H and is not limited to pairwise interactions [88]. As a result, large protein-protein interaction networks (PPINs) have been generated. According to Von Mering [114], combinations of methods to identify protein interactions (MS, Y2H, correlated mRNA expression) are typically better than the independent use of them. Also, the overlapping of interactions identified by different researchers creates a better scenario to understand the functions of proteins involved in each species. Such overlapping contributes to the creation of more robust PPINs [101].

When a protein participates in an interaction, it uses one or many parts of its surface. If we have two proteins that interact, they will interact on a portion of their surface. For example, if we have two interacting cubes each cube is using 1/6 of it surface to interact. The 1/6 surface is named *interaction zone (union site)*. The interaction zone has different features from the remaining 5/6 of the cube.

Interaction Zone

Proteins are composed of amino acids. The characteristics of amino acids who are in the area of the interaction, such as, their position and their surface geometry (shape) finally define some of the properties that characterize its mode of action or interaction capabilities of other proteins or molecules. When a protein is involved in a protein-protein interaction (PPI), the PPI involves one or more of its surfaces.



(a) Protein complex, protein A and B. (b) Interaction zone (colors).

Figure 2.3: Representation of PPI zone (Complex 1A6D, from data set used in Chapter 5).

When we have two proteins interacting, one portion of protein A is touching protein B (see in Figure 5.3, where protein A and protein B). The zone where both proteins have contact is called the interaction zone (interface). This is a general description for the binding sites (see Figure 2.3b). This interaction zone has properties different from the rest of the surface allowing the proteins to interact specifically with one or more proteins.

2.1.3 Relationship among sequences

Sequences diverge during evolution, most commonly due to the replacement of nucleotides in genes in a sequence. The amount of divergence between two sequences can tell us how closely two sequences are related. Duplications or repetitions in either sequence alter the sequence alignments. The degree of similarity between genes reflects the evolutionary relationship between them [91]. The comparison of whole genome sequences from two or more organisms can reveal the location of a previously unknown gene.

It is possible to create alignments with gene sequences (DNA code) and protein sequences (amino acid code). The alignments show the similarity of sequences.

Next, we give some definitions. Two gene sequences (in short, genes) are *homologs*, if they are related by descent from a common ancestral DNA sequence. The term homolog applies to the relationship of genes separated by the event of *speciation* or to the relationship of genes separated by the event of *gene* duplication [2, 25].

Two genes are *orthologs*, if they belong to different species that evolved from a common ancestral gene by speciation of a parental sequence. Ortholog identification is essential for reliable prediction of the functions of genes in the sequenced genomes [2, 86]. Also, there are sequences from different organisms that have a high degree of similarity (their sequences are similar) but the functional relationship between these genes has not been demonstrated.

Genes are often duplicated to generate multiple copies contained in the genome. After these duplications the genes could diverge in their function. The relationship between genes of the same species is called *paralogous* (related by duplication) [2] (see Figure 2.4). The function and sequence information of an individual gene (protein) can help to understand the relationship between and within species. If two sequences A and A' are paralogs, and both are related (common ancestral gene) to a specific sequence B from another species, then the relationship between A and A' with B is called *co-orthologous* (see Figure 2.5b).

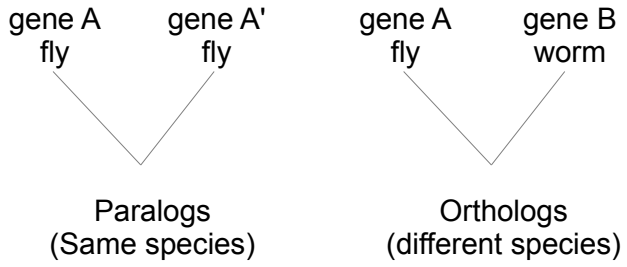
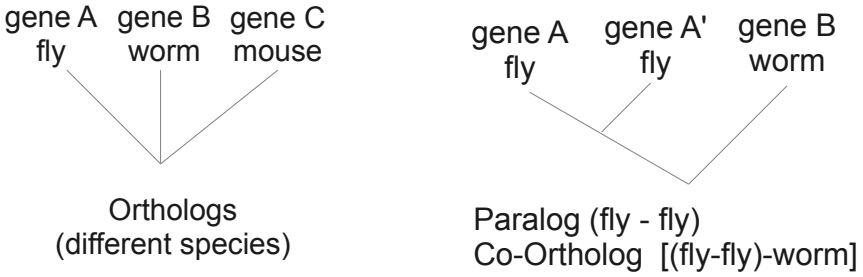


Figure 2.4: Paralogs and orthologs. Protein to the right: interactions of different species; protein to the left: interaction between genes of the same species.



(a) Interactions of proteins in three different species.

(b) Two proteins from the same species together related with a protein from a different species.

Figure 2.5: Orthologs, paralogs, and co-orthologs.

- Figure 2.4, on the left: The paralogous relationship between two sequences from the same species; and on the right, orthologous relationship between two sequences from different species.
- Figure 2.5a: An orthologous relationship is possible between many sequences.
- Figure 2.5b: There are two proteins from the fly species. They are both co-orthologs of the worm protein. The worm protein is a co-ortholog of the two fly proteins.

Orthologs can be able to maintain their function during evolution. Unlike paralogs, they evolve new functions, even if these are related to the original one. Orthologs and paralogs are also homologs [86].

2.1.4 Protein-Protein Interaction Networks

A protein can have interactions with many different proteins. Therefore, one can view the interaction of proteins as biological networks. Every protein has an important role in its net-

work, this role covers the function and interactions of the protein. PPINs provide a valuable framework to understand the functional organization of the proteome, to permit the comparison of networks from different species, and to predict some behaviors that could be useful for better understanding of evolution and functions. Knowing the protein and its environment makes it possible to predict all or some of its functions. This knowledge would contribute to identify proteins and their interactions.

One concern is how large networks are managed. As the size of a network increases also the complexity that is associated with the variability of the data. With all the data gathered from these methods, the set of proteins and interactions are evaluated to see the robustness of the data.

In Brohee et al. [18] a comparison of four algorithms is made: Markov clustering (MCL), restricted neighborhood search clustering, super paramagnetic clustering, and molecular complex detection. The algorithms are used to evaluate various methods such as MS, Y2H, genetic studies and their rates of false positives and miss fraction of the existing interactions. They analyzed the sensitivity and robustness of the algorithms and the alterations in the graphs, concluding that MCL is remarkably robust to be used with altered graphs (there are edges removed and added). Clustering methods are used in the study of PPINs because they can be effective approaches for the identification of protein complexes or functional modules [115].

A PPIN consists of a set of PPIs [51, 112]. The choice of a representation of a PPIN depends on which features are to be modeled. When using graph theory, all the proteins in an organism and all possible interactions between them are represented by a graph [93]. Each vertex represents a protein and each edge can represent a variety of interactions – physical, metabolic, genetic, or biochemical [35]. Use Graph theory approaches to analyze biological networks are important since they can detect properties that would possibly remain undetected otherwise. To compare the values of different graphs it is a challenging task, due to the fact that data stem from different sources (different projects – *in-vivo* or *in-silico*) and methods (for example, Y2H and MS).

2.1.5 The meaning of dN/dS ratio in molecular evolution

In genetics, the dN/dS ratio (also called Ka/Ks ratio) is a way of measuring the rate of sequence change in a gene that tells us something about the selective evolutionary pressures that are acting on a protein-coding gene. It tells us whether the sequence of the gene is under pressure to stay the same, change, or drift randomly. *Synonymous mutations* (SYN) are mutations that do not cause any changes in a protein (silent mutation). And *non-synonymous mutations* (NONS) are mutations that do result in changes in a protein.

Definitions

dN is the total number of non-synonymous changes (#*NONS*) divided by the number of non-synonymous sites (#*NONSsites*), making it a measure of how often these potential changes happen. $dN = \frac{\#NONS}{\#NONSsites}$.

dS is the total number of synonymous changes (#*SYN*) divided by the number of synonymous sites (#*SYN* sites), making it a measure of how often these potential changes happen (can be viewed as a proxy of background mutation). $dS = \frac{\#SYN}{\#SYNsites}$.

dN/dS ratio measures how often the average mutation in a gene is resulting in a change in the protein it produces. The ratio indicates the extent of changes at the amino acid level after normalized by silent mutational changes at the DNA level. Hence, it is a proxy for positive selection pressure in coding genes. This definition assumes selection only at the protein level, not at the DNA or RNA level. *dN/dS* ratio is used to infer the direction and magnitude of natural selection acting on protein coding genes. *dN/dS* ratio is designed to study divergence because its definition assumes fixed changes.

Next we present the interpretations of the different values for *dN/dS*

Ratio equal to one ($\frac{dN}{dS} = 1$).

If mutations in a gene are random, or equally likely to cause changes or not. A ratio of one indicates neutral evolution.

Ratio around 1 ($\frac{dN}{dS} \approx 1$).

This indicates either neutral evolution at the protein level or the average of the sites under positive and negative selective pressures. The gene or protein at different times along its evolution may cancel each other out, giving an average value that may be lower, equal or higher than one.

Ratio greater than one ($\frac{dN}{dS} > 1$).

This indicates the positive selective pressure. Comparisons of homologous genes with a high *dN/dS* ratio are usually said to be evolving under positive or Darwinian selection.

Ratio less than one ($\frac{dN}{dS} =< 1$).

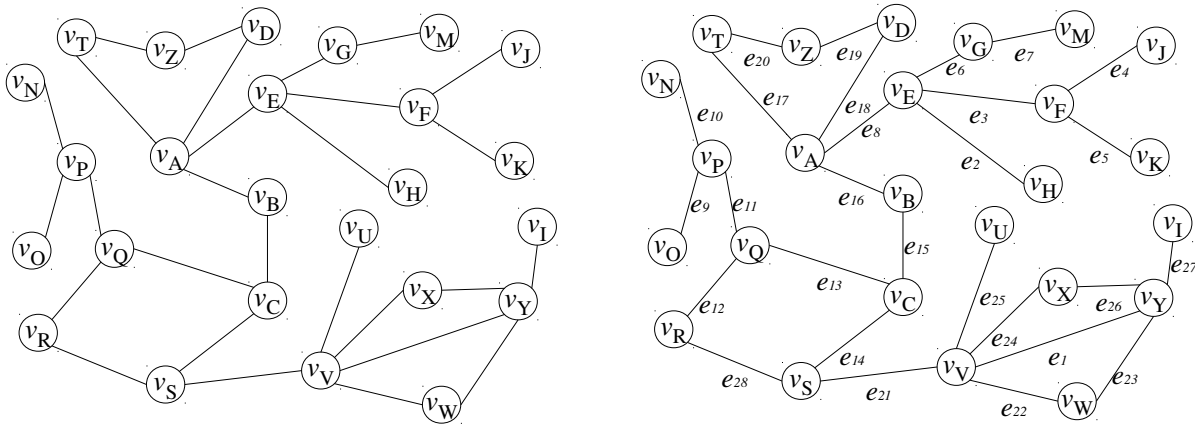
This indicates pressures to conserve protein sequence. Ratio less than one implies purifying selection (stabilizing).

2.2 Graph theory definitions

2.2.1 General terminology

This section describes some definitions from graph theory and social networks that we use in our work on protein networks. For references of the terminology we refer to [16, 26, 32, 46, 116, 128].

We define an undirected graph as $G = (V, E)$ where V is the vertex set and E is the edge set (see Figure 2.6a). The elements of $V = \{v_1, v_2, \ldots, v_n\}$ are called vertices. The size of V is the number of elements in V that is $n = |V|$.

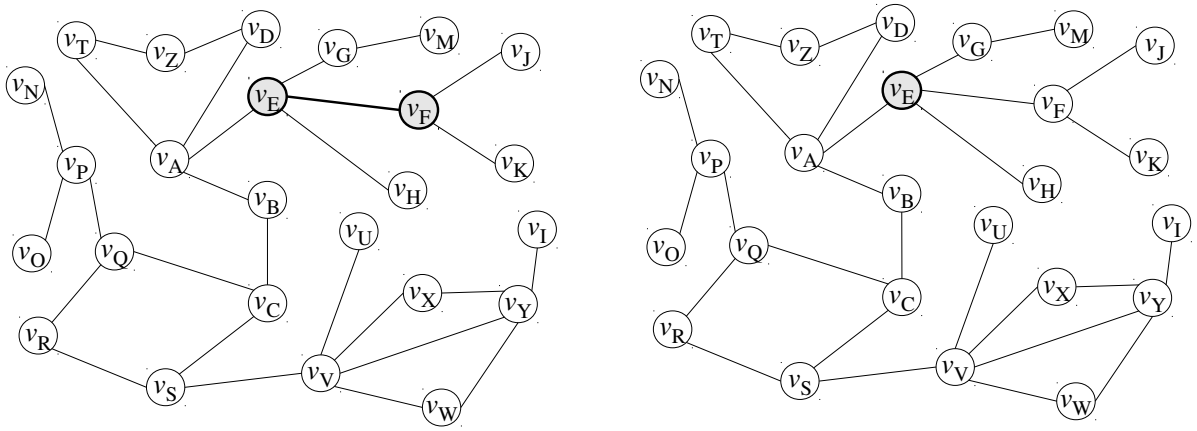


(a) Vertices in Graph G. (b) Edges in Graph G.

Figure 2.6: Graph G.

The elements of $E = \{e_1, e_2, \cdots, e_m\}$ are called edges. The size of E is the number of elements in E, that is $m = |E|$ (see Figure 2.6b).

In a graph G, two vertices v_i and v_j are *adjacent* if they are joined by an edge $(v_i, v_j) \in E$. For example, v_E and v_F in Figure 2.7a are adjacent vertices.



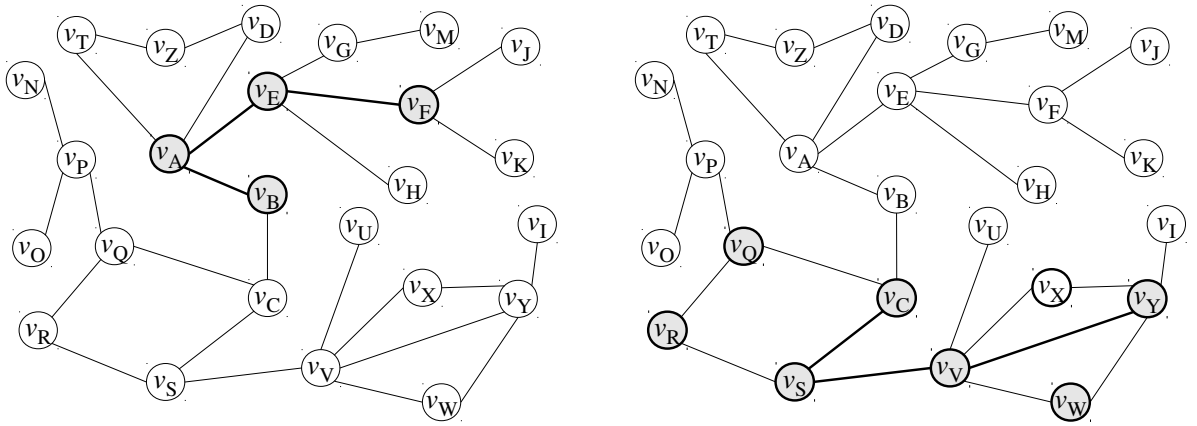
(a) Adjacent vertices in Graph G.

(b) Degree for vertex v_E in Graph G.

Figure 2.7: Adjacency and degree in Graph G.

The *degree* $dg(v)$ of a vertex v in a graph G is the number of edges incident to v [26, 46]. A vertex of degree zero denotes an *isolated vertex* or *singleton*. That is a *singleton* is a vertex v with no incident edges in G. In Figure 2.7b the degree of v_E is $dg(v_E) = 4$.

A *path* $u_1, u_2, v_3, \cdots, u_r$ is a sequence of vertices in G with $(u_i, u_{i+1}) \in E$ for $1 \leq i \leq n$. We call u_1 the *start vertex* of the path and u_r the *end vertex*. A *simple path* is a path that does not contain repeated vertices. The *length* of a simple path is the number of edges that it uses. Figure 2.8a depicts path v_B, v_A, v_E, v_F of length 3.



(a) Path in Graph G.

(b) Simple paths and shortest path from vertex v_C to vertex v_Y in Graph G.

Figure 2.8: Simple paths and shortest path in Graph G.

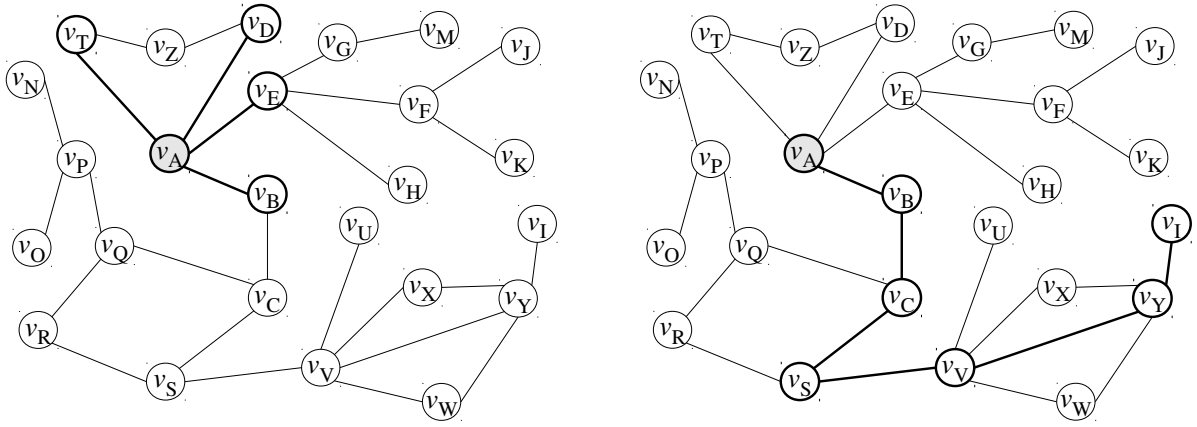
A *shortest path* between two vertices v_i and v_j in G is a path from v_i to v_j of shortest length, also called *geodesic*. The length of a shortest path from v_i to v_j is called the *distance* $dist(v_i, v_j)$ from v_i to v_j. In the case of vertices v_C and v_Y in Figure 2.8b, there are 6 simple paths with start vertex v_C and vertex v_Y (see Table 2.1). The shortest path has distance 3.

Number	Simple path	Path length
1	$\lvert\{v_C, v_S, v_V, v_Y\}\rvert$	3
2	$\lvert\{v_C, v_S, v_V, v_W, v_Y\}\rvert$	4
3	$\lvert\{v_C, v_S, v_V, v_X, v_Y\}\rvert$	4
4	$\lvert\{v_C, v_Q, v_R, v_S, v_V, v_Y\}\rvert$	5
5	$\lvert\{v_C, v_Q, v_R, v_S, v_V, v_W, v_Y\}\rvert$	6
6	$\lvert\{v_C, v_Q, v_R, v_S, v_V, v_X, v_Y\}\rvert$	6

Table 2.1: Six simple paths and one shortest path from vertex v_C to vertex v_Y in graph G.

The set of *neighbors or the neighborhood* $N_G(v)$ of v consist of all the vertices adjacent to v, not including v itself. The *closed neighborhood* $N_G[v]$ of v includes v also, that is $N_G[v]$=$N_G(v) \cup \{v\}$. The *eccentricity* $\xi_G(v)$ of a vertex v in a graph G is the maximum distance from v to any other vertex v_i in the graph, $v \neq v_i$.

Consider vertex v_A in Figure 2.9a. Its *neighborhood* is $N_G(v_A) = \{v_B, v_D, v_T, v_E\}$. The closed *neighborhood* is $N_G[v_A] = \{v_A, v_B, v_D, v_T, v_E\}$ or $N_G[v_A]$= $N_G(v_A) \cup \{v_A\}$ (see also Figure 2.10b, adjacency matrix). The *eccentricity* of v_A is $\xi_G(v_A) = 6$ (see path in Figure 2.9b). The value is obtained from the maximum value in Figure 2.10a distance matrix, column A (row I).



(a) Neighbors of vertex v_A in Graph G.

(b) Eccentricity of vertex v_A in Graph G.

Figure 2.9: Neighbors and eccentricity of vertex v_A in Graph G.

	A	B	C	D	E	F	G	H	I	J	K	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A																									
B	1																								
C	2	1																							
D	1	2	3																						
E	1	2	3	2																					
F	2	3	4	3	1																				
G	2	3	4	3	1	2																			
H	2	3	4	3	1	2	2																		
I	6	5	4	7	7	8	8	8																	
J	3	4	5	4	2	1	3	3	**9**																
K	3	4	5	4	2	1	3	3	**9**	2															
M	3	4	5	4	2	3	1	3	**9**	4	4														
N	5	4	3	6	6	7	7	7	8	8	8	8													
O	5	4	3	6	6	7	7	7	8	8	8	8	2												
P	4	3	2	5	5	6	6	6	7	7	7	7	1	1											
Q	3	2	1	4	4	5	5	5	5	6	6	6	2	2	1										
R	4	3	2	5	5	6	6	6	4	7	7	7	3	3	4	1									
S	3	2	1	4	4	5	5	5	4	6	6	6	4	4	3	2	1								
T	1	2	3	2	2	4	4	4	7	5	5	5	6	6	5	4	5	4							
U	0	4	3	6	6	7	7	7	3	8	8	8	6	6	5	4	3	2	6						
V	0	3	2	5	5	6	6	6	2	7	7	7	5	5	4	3	2	1	5	1					
W	0	4	3	6	6	7	7	7	2	8	8	8	6	6	5	4	3	2	6	2	1				
X	0	4	3	6	6	7	7	7	2	8	8	8	6	6	5	4	3	2	6	1	1	2			
Y	0	4	3	6	6	7	7	7	1	8	8	8	7	7	6	5	4	2	6	2	1	1	1		
Z	2	3	4	1	3	4	4	4	8	5	5	5	7	7	6	5	6	5	1	7	6	7	7	7	

(a) Distance matrix of Graph G in Figure 2.6a.

	A	B	C	D	E	F	G	H	I	J	K	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A		1		1	1														1						
B	1		1																						
C		1														1		1							
D	1																								1
E	1					1	1	1																	
F					1				1	1															
G					1						1														
H					1																				
I																								1	
J						1																			
K						1																			
M							1																		
N														1											
O															1										
P													1	1		1									
Q			1												1		1								
R																1		1							
S			1														1				1				
T	1																								1
U																					1		1		
V																		1		1		1	1	1	
W																					1			1	
X																				1	1			1	
Y									1												1	1	1		
Z				1															1						

(b) Adjacency matrix of Graph G in Figure 2.6a.

Figure 2.10: Neighbors, eccentricity of vertex v_A and diameter and radius of Graph G.

Using the distance matrix from Figure 2.10a we obtain the maximum and minimum eccentricity values for the rest of the vertices (see Table 2.2).

	B	C	D	E	F	G	H	I	J	K	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
$\xi_G(v)$	5	5	7	7	8	8	8	9	9	9	9	8	8	7	6	7	6	7	8	7	8	8	8	8

Table 2.2: Eccentricity for vertices of Graph G.

The *diameter* $D(G)$ of a graph G is the maximum eccentricity over all vertices in the graph. The *radius* $R(G)$ of a graph G is the minimum eccentricity over all vertices in the graph. The *periphery of a graph* is the set of vertices that has maximum eccentricity. The vertices in this set are called *peripheral vertices*. The *center of a graph* is the set of vertices that has minimum eccentricity. The vertices in this set are called *central vertices*. The *density* of a graph G is the ratio of the number of edges and the number of possible edges in G.

For our the example the measures are obtained from Table 2.2. The *diameter* of G is $D(G) = 9$ and G's *radius* is $R(G) = 5$. The *periphery* is $\{v_I, v_J, v_K, v_M\}$ and *the center* vertices are $\{v_B, v_C\}$. The average degree of G is $\bar{dg} = 2.32$. The *density* of G is $\frac{\bar{d}}{n-1} = \frac{2.32}{25-1} = 0.097$.

Next we introduce terminology that measures centralities in graphs.

2.2.2 Centrality measures

Betweenness

The *betweenness* (also sometimes called betweenness centrality) of a vertex v is based on the number of shortest paths from all vertices to all others in G that pass through v. Before defining *betweenness* formally we recall that the distance of two vertices a and b is the length of a shortest path connecting a and b. Therefore the following holds.

- $dist_G(a,a) = 0$, for every $a \in V$.
- $dist_G(a,b) = dist_G(b,a)$ for $a,b \in V$.
- A vertex $v \in V$ lies on a shortest path between vertices $a,b \in V$ if and only if $dist_G(a,b) = dist_G(a,v) + dist_G(v,b)$.

We define

NSP_{ab} is the total number of shortest paths between vertex $a \in V$ and vertex $b \in V$.

$NSP_{ab}(v)$ is the number of all those shortest paths between vertex $a \in V$ and $b \in V$ that pass through vertex $v \in V$. $a \neq b \neq v \in V$.

The *betweenness* $bc(v)$ of a vertex v in a graph G is defined as follows:

$$bc(v) = \sum_{a \neq b \neq v \in V} \frac{NSP_{ab}(v)}{NSP_{ab}}$$

The *normalized betweenness* $nbc(v)$ of a vertex v in an undirected graph G is

$$nbc(v) = \frac{bc(v)}{(N-1)(N-2)/2},$$

where N is the number of vertices in the graph G that is $N = |V|$. Note that $nbc(v) \in [0,1]$. We further note

- If a is adjacent to b, then $NSP_{ab} = 1$.
- If there is no path connecting a and b in G, then $NSP_{ab} = 0$.
- A vertex v with high *betweenness* has a strong influence over paths in the graph [32, 33, 120]. This means, if v is removed from the graph then its connectivity is affected considerably.

Example 1. Betweenness

Consider a graph T in Figure 2.11, with 6 vertices and 6 edges. We calculate the *betweenness* for every vertex in the graph. We have 15 vertex pairs (see Table 2.3). The are also 24 simple paths between the vertex pairs.

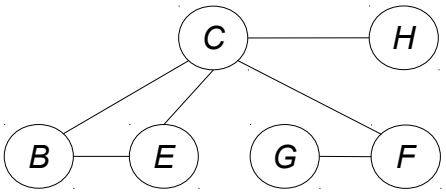


Figure 2.11: Graph T.

	C	B	E	F	G	H
C	-	$\underline{CB}, CEB$	$\underline{CE}, CBE$	$\underline{CF}$	$\underline{CFG}$	$\underline{CH}$
B	-	-	$\underline{BE}, BCE$	$\underline{BCF}, BECF$	$\underline{BCFG}, BECFG$	$\underline{BCH}, BECH$
E	-	-	-	$\underline{ECF}, EBCF$	$\underline{ECFG}, EBCFG$	$\underline{ECH}, EBCH$
F	-	-	-	-	$\underline{FG}$	$\underline{FCH}$
G	-	-	-	-	-	$\underline{GFCH}$
H	-	-	-	-	-	-

Table 2.3: Simple paths for all the pairs of Graph T in Figure 2.11. 15 of these paths (underlined) are shortest.

In Table 2.3 we underlined all the shortest paths between pairs. There are 8 shortest paths that pass through vertex C, and 4 shortest paths pass through vertex F.

All vertices that participate in shortest paths passing through C or F are shown in Table 2.4.

Vertex v	Shortest paths through vertex v
C	$\{B-H, B-F, B-G, E-H, E-F, E-G, H-F, H-G\}$
F	$\{C-G, B-G, E-G, H-G\}$

Table 2.4: Shortest paths that pass through a vertex v in Graph T from Figure 2.11.

Next, we calculate the betweenness for vertices C and F. Table 2.5a and Table 2.5b show the number of shortest paths for each vertex pair in T, as well as the number of those paths that pass through C and F, respectively.

Finally, we can calculate the *betweenness* of C and F. For graph T we have $N = 6$. In Table 2.6 we present the *betweenness* values and normalized *betweenness* values. Note that here $\frac{(N-1)(N-2)}{2} = 10$.

Vertex v	Pair	$NSP_{ab}/NSP_{ab}(v)$
C	$\{B-F\}$	$1/1=1$
	$\{B-G\}$	$1/1=1$
	$\{B-H\}$	$1/1=1$
	$\{E-F\}$	$1/1=1$
	$\{E-G\}$	$1/1=1$
	$\{E-H\}$	$1/1=1$
	$\{F-H\}$	$1/1=1$
	$\{G-H\}$	$1/1=1$

(a) Shortest paths pass through C

Vertex v	Pair	$NSP_{ab}/NSP_{ab}(v)$
F	$\{C-G\}$	$1/1=1$
	$\{B-G\}$	$1/1=1$
	$\{H-G\}$	$1/1=1$
	$\{E-G\}$	$1/1=1$

(b) Shortest paths pass through F

Table 2.5: Shortest paths pass through C and F from Table 2.4.

Vertex v	$bc(v)$	$nbc(v)$
C	$1+1+1+1+1+1+1+1=8$	$8/10=0.533$
F	$1+1+1+1=4$	$4/10=0.267$

Table 2.6: Betweennes for every vertex of Graph T in Figure 2.11.

We can see in Table 2.6 that vertex C has higher betweenness than F. This means that vertex C is more central than vertex F in the graph.

Example 2. More than one shortest path

In contrast to above example, more than one shortest path between a given pair of vertices may exist

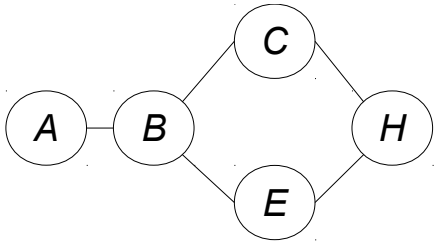


Figure 2.12: Graph Q.

For example in graph Q in Figure 2.12 there are three pairs of vertices that have more than one shortest path: $A-H$ with $ABCH$ and $ABEH$; $B-H$ with BCH and BEH; and $C-E$ with CBE and CHE.

For every case are two shortest paths, that is $NSP_{AH}=NSP_{BH}=NSP_{CE}=2$

Table 2.7 shows the different shortest paths vertex B,C,E or H, respectively. For example vertex B participates in 5 shortest paths (column *Shortest paths*), but three of them have a different shortest path through another vertex.

Vertex v	Shortest paths passing vertex v	$bc(v)$
B	$ABC, ABE, ABCH/2, ABEH/2, CBE/2$	1+1+0.5+0.5+0.5=3.5
C	$ABCH/2, BCH/2$	0.5+0.5=1
E	$ABEH/2, BEH/2$	0.5+0.5=1
H	CHE	0.5

Table 2.7: Betweenness values for each vertex in Graph Q.

We can see in Table 2.7 that vertex B has the highest betweenness (3.5) even though there is more than one short path.

Closeness or closeness centrality

The *closeness* evaluates how close a vertex v is to all other vertices in the graph [116].

More formally, the *closeness* of a vertex v is the inverse of the sum of distances from v to all other vertices:

$$cl(v) = \frac{1}{\sum_{t \in V} dist(v,t)},$$

where $\sum_{t \in V} dist(v,t)$ is the sum of shortest path distances from v to all others vertex in V.

The *normalized closeness* $ncl(v)$ of a vertex v in an undirected graph G is

$$ncl(v) = cl(v)(N-1)$$

where N is the number of vertices in the graph G [16].

Example: Closeness

Consider graph T in Figure 2.13 with 6 vertices and 6 edges. We calculate the *closeness* for every vertex in the graph.

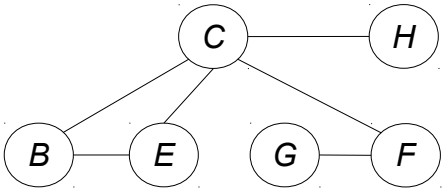


Figure 2.13: Graph T.

We first obtain the distances for all pairs of vertices. The distance matrix for graph T is ahown in Table 2.8.

	C	B	E	F	G	H
C	-	1	1	1	2	1
B	1	-	1	2	3	2
E	1	1	-	2	3	2
F	1	2	2	-	1	2
G	2	3	3	1	-	3
H	1	2	2	2	3	-
Summed values	6	9	9	8	12	10

Table 2.8: Distance matrix of Graph T in Figure 2.13.

Next, $cl(v)$ is calculated for each vertex in graph T. *Closeness* values and *normalized closeness* values are shown in Table 2.9.

	Closeness	Normalized closeness
C	$\frac{1}{6}$ =0.167	$\frac{5}{6}$ = 0.83
B	$\frac{1}{9}$ =0.111	$\frac{5}{9}$ = 0.56
E	$\frac{1}{9}$ =0.111	$\frac{5}{9}$ = 0.56
F	$\frac{1}{8}$ =0.125	$\frac{5}{8}$ = 0.63
G	$\frac{1}{12}$ =0.083	$\frac{5}{12}$= 0.42
H	$\frac{1}{10}$ =0.1	$\frac{5}{10}$= 0.5

Table 2.9: Closeness values of Graph T in Figure 2.13.

We can see in Table 2.9 that vertex C has the highest closeness (0.83) in the graph T. This means vertex C is the nearest to all other vertices in the graph T (higher values assume a positive meaning in term of node proximity, vertex C is more central).

2.3 Social networks

2.3.1 Measures from social networks

Communities in graphs

In our work, we evaluate similarities and differences between social networks and biology networks. Here, we introduce the concepts from social networks and social network analysis necessary for this work. Social networks are structures where actors (nodes) represent people or other entities, like organizations, embedded in a social context, and where the connection

between the entities (edges) represent interaction, collaboration, or influence between these entities [70].

An interesting characteristics of a graph is how nodes are grouped together. Edges may denote relationships, such as friendship, acquaintances, classmates, colleagues, or business. A community or cluster is a connected subgraph that can be viewed as a set of nodes that share some common characteristics or functions (interest, goal, or project) [21, 128].

There are studies of methods for finding communities in networks. The algorithms used to find communities are based on dividing the network in subnetworks where the number of edges inside the subnetwork is relatively large (this means communities are strongly connected internally), and the edges between communities are relatively sparse (weakly connected to the outside) [21].

In this research, we use the term *community* as a set of nodes that share a function or feature and having a strong relation between them.

There are methods from the area of data mining that identify various aspects of network organization [128]. Clustering algorithms are used to identify sets of nodes that are more likely to interact with each other than with nodes outside the set.

Next, we consider conductance, which measures the quality of a community.

Conductance

The *conductance* of a graph measures how strongly connected the graph is. A low *conductance* means that there is some bottleneck in the graph, that is, a subset of nodes is not well connected with the rest of the graph. A high *conductance* means that the graph is well connected. Before defining conductance of a graph we define the *conductance of a subset* of the vertices in the graph. Here, the measure is applied to a set of nodes with respect to the rest of the nodes in the graph: subsets of nodes are weighted to reflect their importance [38].

Let $G = (V,E)$ be an undirected graph and $S \subset V$ be a set of nodes with $|S| \leq \frac{1}{2}|V|$.

$V = \{v_1, v_2, ..., v_i, ..., v_j, ..., v_n\}$

$E = \{e_1, e_2, e_3, ..., e_m\}$

Then the *conductance* $\phi(S)$ in G is defined as

$$\phi(S) = \frac{\left|\{e_{ij} \in E : v_i \in S \text{ and } v_j \notin S\}\right|}{\min\{\sum_{v_i \in S} dg(v_i), \sum_{v_j \notin S} dg(v_j)\}},$$

where $\left|\{e_{ij} \in E : v_i \in S \text{ and } v_j \notin S\}\right|$ shows the total number of edges with endpoint in the community and one endpoint in the complement of the community and $\sum_{v_i \in S} dg(v_i)$ the sum of the degree of nodes in the community.

The *conductance of a graph* is defined as the minimum conductance over all possible subsets.

$$\phi(G) = \min_{S \subset V} \varphi(S)$$

Example: Conductance

Consider graph H in Figure 2.14 with 13 vertices and 18 edges. We calculate the *conductance* for the three communities in the graph.

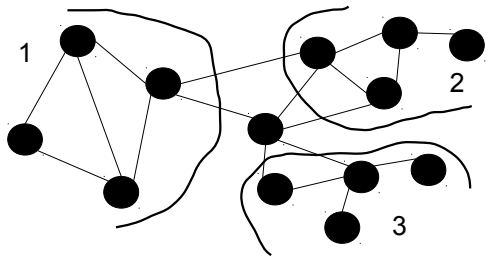


Figure 2.14: Graph H and three communities.

Communities 1, 2, and 3 from Figure 2.14 are densely connected internally and sparsely connected with the rest of the graph. In table 2.10 are the conductances for each community.

Community	$\left\lvert\{e_{ij} \in E : v_i \in S \text{ and } v_j \notin S\}\right\rvert$	$\sum_{v_i \in S} dg(v_i)$	Cd
1	2	12	$\frac{2}{12} = 0.167$
2	3	10	$\frac{2}{10} = 0.3$
3	2	8	$\frac{2}{8} = 0.25$

Table 2.10: Conductance for communities in graph H.

The best community is number one with the lowest *conductance*, $Cd(1) = 0.167$. The second best community is community number 3, $Cd(3) = 0.25$. Conductance of graph H is the minimum of the values obtained in column Cd, $Cd(G) = 0.167$.

Chapter 3

How do biological networks differ from social networks?

In this chapter we outline important differences between (1) protein interaction networks (PPIN) and (2) social and other complex networks in terms of fine-grained network community profiles. While these families of networks present some general similarities, they also have some stark differences in the way the communities are formed. Namely, we find that the sizes of the best communities in such biological networks are an order of magnitude smaller than in social and other complex networks. We furthermore find that the generative model describing biological networks is very different from the model describing social networks. While for latter the Forest-Fire model best approximates their network community profile, for biological networks it is a random rewiring model that generates networks with the observed profiles. Our study suggests that these families of networks should be treated differently when deriving results from network analysis, and a fine-grained analysis is needed to better understand their structure.

3.1 Literature review

Protein interaction networks have been the focus of numerous works in the research community. Pavlopoulos et al. [93] and Mason et al. [81] studied how to find the most important nodes in large protein networks. They utilize such information for better determining protein functions [93] and identifying drug targets [81]. Pavlopoulos et al. [93] use a different approach and different models and methods to reveal hidden properties and features of a network. They indicate that the structure of these protein networks are linked to function. The protein network topology analysis is limited because it provides a static perspective of the system and

not a dynamic perspective. Mason et al. [81] focus on network analysis to determine the role of proteins or genes of unknown function to identify potential applications in medicine (drugs for different diseases). They design effective containment strategies for infectious diseases providing early diagnosis of neurological disorders through detecting abnormal patterns of neural synchronization in specific brain regions.

Barabassi and Oltvai in [7] study the general properties of the proteins in networks coming from complex interactions. They identify the interlink between structure, topology, network usage, robustness and function. Using network tools allowed them to see a different perspective of proteins and genes. Barabassi and Oltvai in [7] make the case that with respect to the common measures of network structure, the proteins in these networks and people in social networks behave similarly. In this work, however, we show that this is not always the case.

To understand the biological significance of systems, many researchers applied different models, approaches, and methods to identify motifs or patterns that indicate common properties. They analyze the networks in detail using measures like network centralities [60, 89], network topologies [97, 124], cluster analysis [72, 100], or network models [93].

Clustering algorithms (also called community-detection algorithms) are used to understand the organization of networks and their functions through the identification of protein complexes or functional modules [81, 115]. The clustering algorithms typically join proteins in groups (communities) according to attributes that are shared by the proteins in the group. These algorithms show that identifying and predicting communities also helps identifying important nodes (proteins) in the network. There also comparative analyses of different clustering algorithms have been performed to identify those that are better at predicting relevant communities. Wang et al. in [115] present a detailed clustering algorithm comparison for extracting clusters from protein interaction networks. Most of the algorithms focus on protein complexes and functional modules. Some more recent works [63, 103] propose improving the prediction of protein function by utilizing protein community information. Also, Lee et al. [63] apply a method that improves modularity solutions to predict protein functions. However, the aforementioned works do not make use of conductance scores as we do.

Centrality measures help to analyze the different communities and evaluate how a gene or protein is relevant for its community, other communities or the complete network [126]. They did an analysis of topological properties in different networks. Also, they propose the integration of different approaches for future predictive models. Centrality measures (such as, betweenness, closeness or degree) give evidence that there is a close relationship between the centrality of a node and its essentiality in the network [81]. One example is presented by Goh et al. [41]. They examined biological networks and found that betweenness and degree of nodes are significantly correlated. Girvan and Newman [38] use edge betweenness to detect

community peripheries. They were able to detect known structure with high sensitivity and reliability.

3.2 Methodology

We focus on the structural differences of protein interaction networks versus social and other complex networks. While there are some similarities between them, there are nevertheless significant differences, mainly in the community structure of these networks. This is in contrast to the widely held belief that biological networks are very similar to social networks and thus tools and insights from the latter can be easily applied to or extended for the former [7]. We show that best communities are smaller in size by an order of magnitude in biological networks compared to those in social networks.

Community detection is very important not only for social networks, but also for biological networks. This is because communities can provide for a better understanding and insight on the fine grained structure of biological networks and the way their different parts work together.

We compute for our study the *network community profiles* (NCPs) of 11 large protein-interaction networks. NCPs are based on the notion of conductance that captures the ratio of edges connecting nodes within the community with nodes outside the community to edges inside the community. The smaller the conductance of a set of nodes, the more community-like the set is. Conductance is extensively used to measure the cohesiveness of a community and has been shown to have parallels with the theory of random walks on networks.

Along the lines of [68], we investigate the conductance of communities over all the possible size scales. The main question we explore is: What are the best community sizes and community qualities for each network family? The network community profile is one of the best tools to answer this question. Intuitively, NCP extracts the conductance of the best community as a function of the size values considered. While NCP is NP-hard to compute, there are several approximation algorithms that give satisfactory solutions [68].

We present the following empirical findings. First, the conductance of the best communities for each size scale (k) decreases initially, and the global minimum is typically achieved for $k = 10$. This is in contrast to social networks where the global minimum is reached for $k = 100$ or greater, an order of magnitude bigger than the global minimum for biological networks. Second, at the size of about $k = 10$, the NCP for biological networks exhibits an uptrend, which means that the community structure deteriorates as more nodes are considered in communities. In other words, the communities start blending with each other and gradually disappear. And third, differently from social networks, the generative model explaining this

type of behavior is not Forest Fire [66, 67], but a random rewiring model [83] conditioned on the same degree distribution as the original graph.

Knowing that the best communities in biological networks are an order of magnitude smaller than communities in social networks is very important. This is because community structure can help us to decide which are the possible missing links to further investigate. Clearly, there is a higher chance that there is a missing link between nodes within a community than between nodes not in the same community. Exploring missing links in social networks is not particularly expensive. However, it is quite expensive to do so for biological networks. Therefore, the smaller the meaningful communities, the fewer missing links we need to explore in a laboratory setting. A community profile plot helps in better understanding the costs of further investigating missing links in biological networks.

3.2.1 Conductance and community profiling

We consider the networks to be undirected graphs. The conductance of a set gives a score for the quality of the set as a community [38]. For a formal definition see subsection 2.3.1.

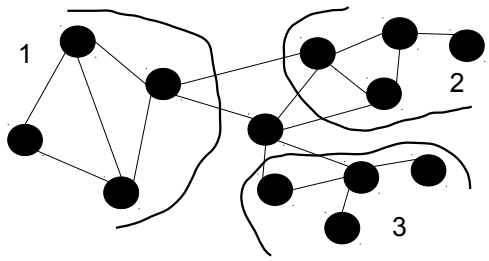


Figure 3.1: A networks and three communities. Communities 1, 2, and 3 are densely connected internally and sparsely connected with the rest of the graph.

The higher number of edges that cross the boundaries of a set S, the higher the conductance $\gamma(S)$, and the lower the community structure of S. Hence, for detecting good communities, we look for sets of low conductance. These are sets that are densely connected internally and sparsely connected with the rest of the graph. In Figure 3.1, we observe three good communities: 1, 2, and 3. The values are the following:

1. There are 2 edges that separate community 1 from the rest of the network. The sum of the degree of nodes in the community is 12. The conductance is $\frac{2}{12} = 0.167$.

2. There are 3 edges that separate community 2 from the rest of the network. The sum of the degree of nodes in the community is 10. The conductance is $\frac{3}{10} = 0.3$.

3. There are 2 edges that separate community 1 from the rest of the network. The sum of the degree of nodes in the community is 8. The conductance is $\frac{2}{8} = 0.25$.

We note that there are many community measures. However, as noted by prominent works [58, 109], the conductance captures the gestalt of communities [127], and therefore is used frequently to perform community detection [17,74,106]. In community profiling we select the best community for each size and plot their conductance scores. In order to find communities with good conductance (minimum values), we use the local spectral clustering algorithm of [4] and the bag-of-whiskers clustering algorithm of [68]. Whiskers are sets of nodes connected to the rest of the graph by one edge; bag-of-whiskers are sets of such whiskers. As shown in [68], bags-of-whiskers give communities with very good conductance scores.

3.2.2 Biological networks analyzed

We considered many of the reasonable-sized protein interaction networks[1] available. From a total of 55 networks, coming from 16 species, we focus here on 11 of them (Table 3.1). The results for other networks are comparable. The networks we consider have sizes varying from 708 to 15,337 nodes, and from 1,357 to 133,645 edges. The data sets were obtained from various sources (see column 'Reference' in Table 3.1). To evaluate the different networks sources, we work with two networks for two species, *Caenorhabditis elegans* and *Homo sapiens*. We validate our results by considering data sets for the same species, but coming from different source.

Biological Networks	Nodes	Edges	Reference
Arabidopsis thaliana	7,050	16,263	BioGrid [110]
Caenorhabditis elegans 1	3,895	7,758	BioGrid [110]
Caenorhabditis elegans 2	2,528	3,706	Harvard [24]
Drosophila melanogaster	8,127	38,839	BioGrid [110]
Echericha coli	2,874	11,538	DIP [121]
H pylo	708	1,357	DIP [121]
Homo sapiens 1	15,337	133,645	BioGrid [110]
Homo sapiens 2	6,711	17,348	Mint [71]
Mus musculus	4,602	9,841	BioGrid [110]
Saccharomyces cerevisie	5,376	24,734	Mint [71]
Schizosaccharomyces pombe	4,008	55,362	BioGrid [110]

Table 3.1: Biological networks.

[1] We will refer to these networks as *biological networks*. We note that there are also other types of biological networks that we plan to study as part of our future work.

3.2.3 Network community plot analysis

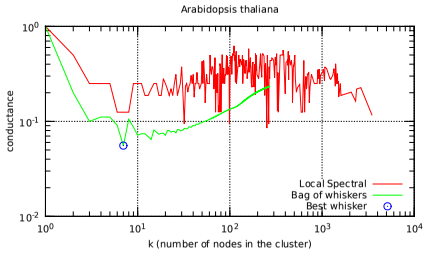
As shown in [68] most social networks exhibit the following community profile structure. Up to a certain size the slope of NCP is downward: as the size increases the conductance values decrease. This in turn means that the best sets become increasing community-like. At size of 100 or more, the NCP reaches a global minimum. This implies that the best communities in social networks are typically of size 100 or more. If the number of nodes in a community is larger than 100 or more, the NCP of most social networks is upward sloping over several orders of magnitude. This means that after a certain size, typically at least 100, the communities become less meaningful and they blend more and more with the whole network. For other networks, such a power-grid networks, the NCP is almost always slopping downwards. This means the more nodes are added to communities the better they become in terms of conductance.

Initially a spectral algorithm was implemented but the running time was not efficient. We decided to use the algorithm used in Leskovec et al. [68]. We used the SNAP Package [65], the specific function used was Plots the network Community Profile (NCP – ncpplot package). The algorithm use Spectral algorithm to define the clusters and the conductance to evaluate the communities.

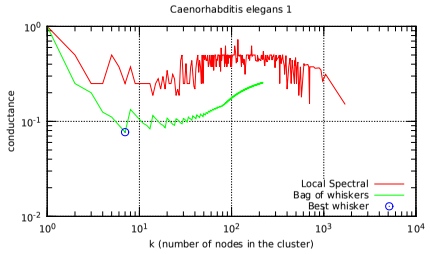
In Figure 3.2 and Figure 3.3 we show the community profiles for biological and social networks and a power-grid network. We show the conductance scores of the best communities computed, using the Local Clustering algorithm of [4] and the Bag-of-Whiskers algorithm of [68]. A local clustering of a vertex quantifies how close its neighbors are to being a clique. The local clustering information is propagated into larger communities in a subtle and location-specific manner. In the case of bag-of-Whiskers algorithm, it was created by [68] where they artificially create communities from disconnected whiskers and measure conductance of such clusters.

In Figure 3.2 we depict, network community profiles for biological networks. They are computed using the local spectral clustering and bag-of-whiskers algorithms. The conductance values are shown a long axis Y, and the number of nodes in the corresponding cluster a long axis X. Both algorithms give a network community profile that is initially downward slopping, then trending upwards. The global minimum for both methods, across most of the biological networks, is at about a community size value of ten. This is in stark contrast to network community profiles for social and other complex networks. Also, observe that whiskers give significantly better communities than local spectral clustering.

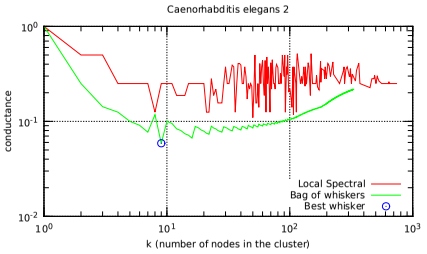
See Table 3.2 for statistical data of the networks.



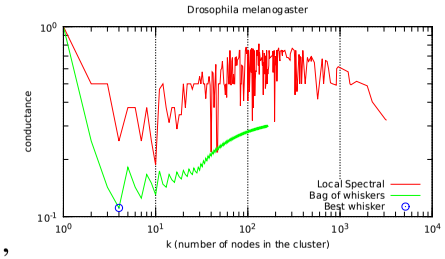
(a) *Arabidopsis thaliana.*



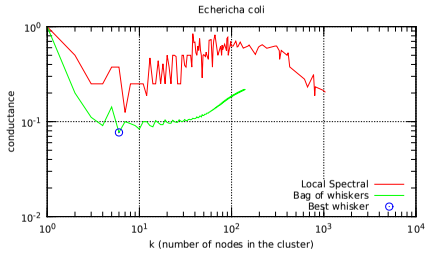
(b) *Caenorhabditis elegans 1.*



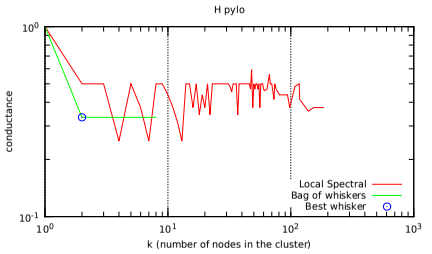
(c) *Caenorhabditis elegans 2.*



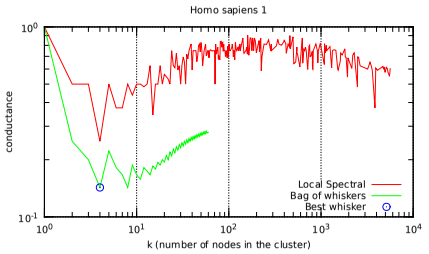
(d) *Drosophila melanogaster.*



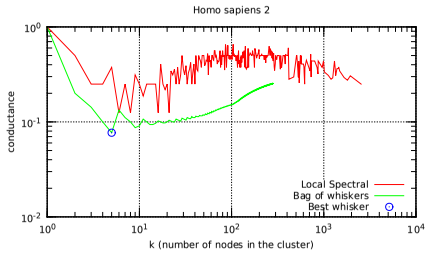
(e) *Echericha coli.*



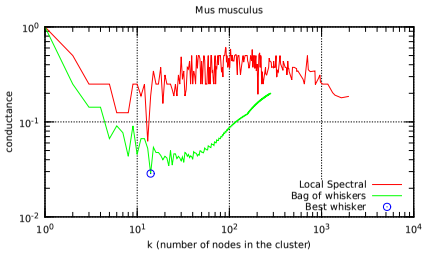
(f) *H pylo.*



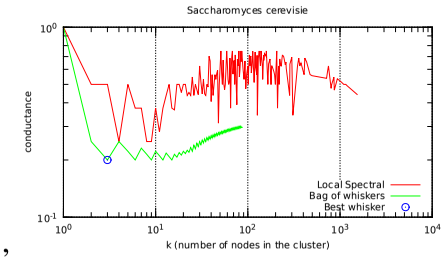
(g) *Homo sapiens 1.*



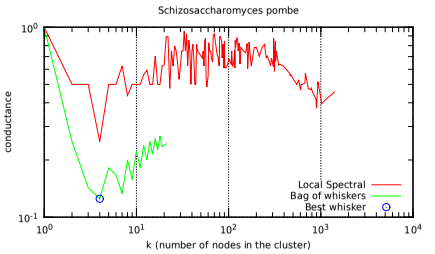
(h) *Homo sapiens 2.*



(i) *Mus musculus.*



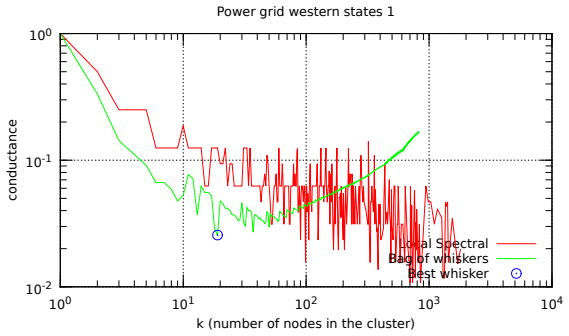
(j) *Saccharomyces cerevisie*



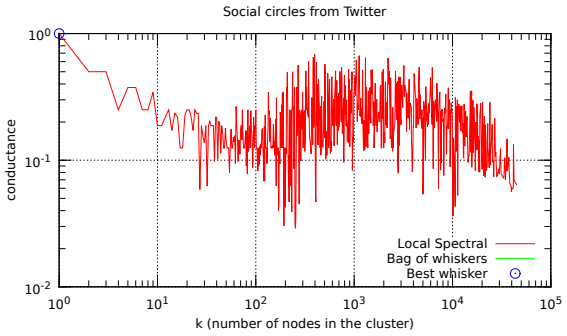
(k) *Schizosaccharomyces pombe.*

Figure 3.2: Network community profiles for biological networks computed using the local spectral clustering (red/dark) and bag-of-whiskers (green/light) algorithms.

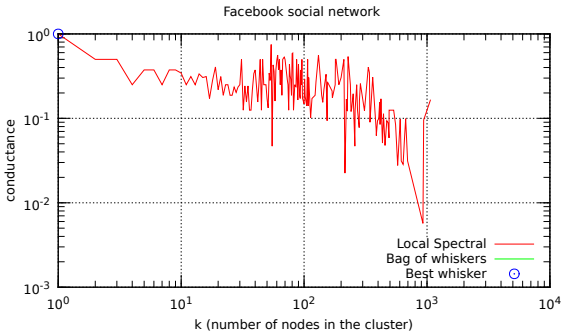
In Figure 3.3, we show the community profiles of two social networks, Twitter and Facebook, as well as the community profile of a power-grid network. We observe that the community profiles of the two social networks have a downward slope up to a certain community size, and then they trend upward. In the case of NCP of the power-grid is always going downward. Similar extensive results for social networks are presented in [68], this corroborate our results. Their global minimum orders of magnitude greater than the global minima we observe for biological networks. Also, we observe that there are no whiskers for the Twitter and Facebook networks, i.e. there are no communities that are barely connected to the rest of the network.



(a) CDG - Spectral and Whiskers algorithm.



(b) Twitter - Spectral and Whiskers algorithm.



(c) Facebook - Spectral and Whiskers algorithm.

Figure 3.3: Network community profiles (red/dark) and bag-of-whiskers (green/light) algorithms of two social networks and a power-grid network (a) $4,941$ nodes [117], (b) $81,306$ nodes [65], and (c) $4,039$ nodes [65].

We show the NCP of biological networks in Figure 3.2. We observe a similar shape of the NCPs as for the studied of social networks. Initially the slope is downward, then upward. However, the global minimum is not reached at size 100 or greater as for social networks, but surprisingly at size about 10, an order of magnitude smaller! This indicates that biological networks present a very different community structure when compared to social networks as their local structure appears much more local than social networks. We also observe that

Networks	Average Degree	Network Diameter	Connected Components	Avg.Clustering Coefficient	Average Path Length	Average #Triangles
Arabidopsis thaliana	4.61	14	154	0.16	4.46	2.79
Caenorhabditis elegans 1	3.98	13	86	0.11	4.29	1.86
Caenorhabditis elegans 2	2.93	14	147	0.04	5.32	0.37
Drosophila melanogaster	9.56	10	50	0.12	4.09	22.37
Echericha coli	8.03	12	295	0.15	3.97	19.07
H. pylo	3.83	9	17	0.03	4.13	0.33
Homo sapiens 1	17.43	8	57	0.32	2.67	85.47
Homo sapiens 2	5.17	11	135	0.11	4.41	1.98
Mus musculus	4.28	16	124	0.20	4.34	2.35
Saccharomyces cerevisie	9.20	10	28	0.13	3.87	11.69
Schizosaccharomyces pombe	27.63	8	6	0.24	2.80	159.21
Twitter	33.00	5	1	0.57	4.59	160.90
Facebook	43.69	8	1	0.62	3.69	1197.33
Power grid western states	2.67	46	1	0.11	18.99	0.40

Table 3.2: Statistical Data of the networks.

whiskers give significantly better communities than Local Clustering. This means that the best communities are only barely connected to the rest of the graph for biological networks.

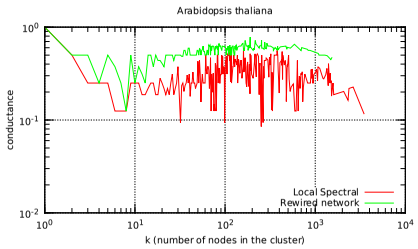
3.3 Modelling results

3.3.1 Modelling results and discussion

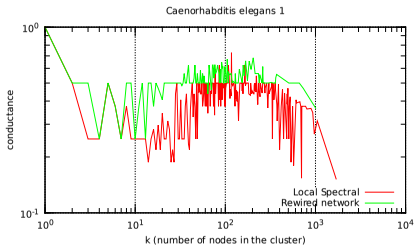
A natural question we would like to answer is: What generative model best fits biological networks? For social networks, [68] shows that a Forest-Fire model, where new edges are added via a recursive burning mechanism in an epidemic-like fashion, generates networks with network profiles that closely resemble profiles of social networks.

In contrast, a Forest-Fire model is not the right choice for biological networks [68]. Surprisingly, we observed that a *rewiring* model proposed by [83] can generate networks with a network community profile that closely resembles profiles of biological networks. The rewiring model works as follows. Starting with the original network we randomly select pairs of edges and switch their nodes. By doing this many times, we obtain a random graph with the same degree sequence as the original one.

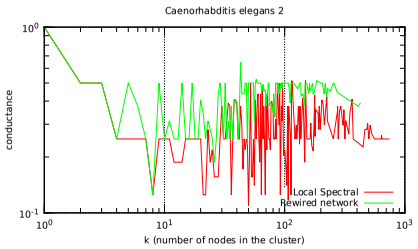
We show the NCPs with rewiring in Figure 3.4 for biological networks, and in Figure 3.5 for the two social networks and the power-grid network. We observe that the NCPs for the rewired networks behave similar to those for the original biological networks. On the other hand, the behavior of the NCPs for the rewired social networks and the power-grid network is quite different from their original counterparts. This reinforces once more the fact that the internal structure of communities in biological networks is very different from that observed in social and other complex networks.



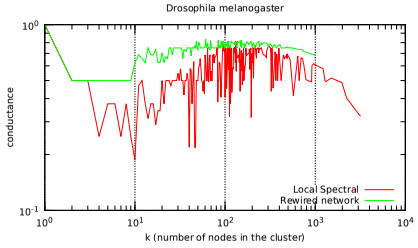
(a) *Arabidopsis thaliana.*



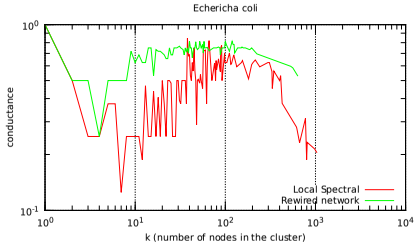
(b) *Caenorhabditis elegans 1.*



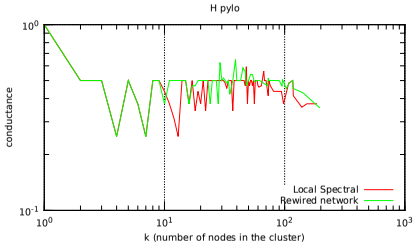
(c) *Caenorhabditis elegans 2.*



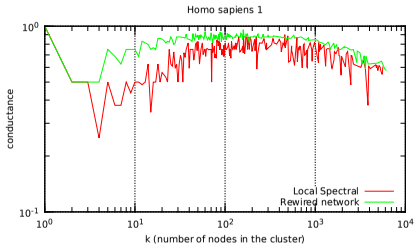
(d) *Drosophila melanogaster.*



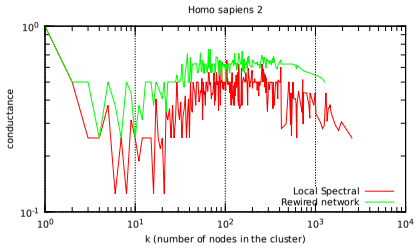
(e) *Echericha coli.*



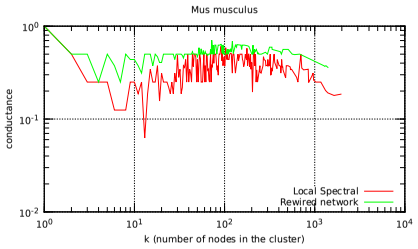
(f) *H pylo.*



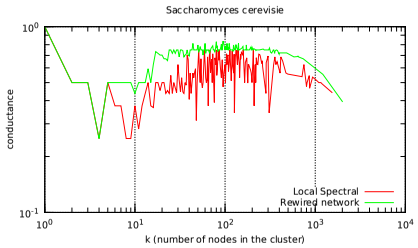
(g) *Homo sapiens 1.*



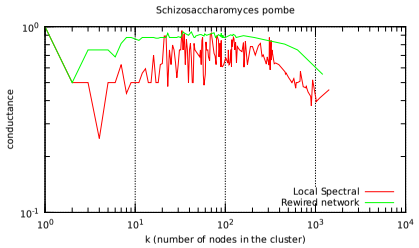
(h) *Homo sapiens 2.*



(i) *Mus musculus.*

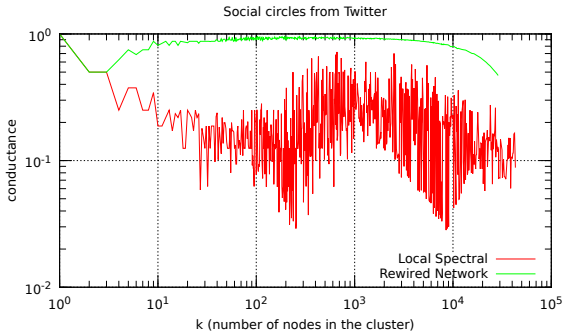


(j) *Saccharomyces cerevisie.*

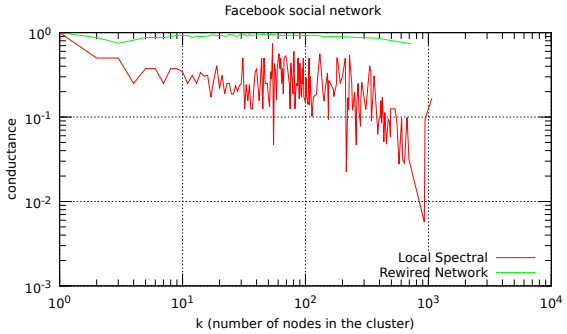


(k) *Schizosaccharomyces pombe.*

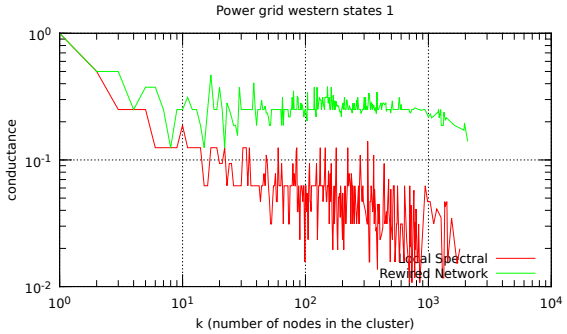
Figure 3.4: Network community profiles of biological networks (red/dark) and their rewired (green/light) networks. The profiles of the original networks and their rewired counterparts exhibit a similar nature. This is not the case for social and other complex networks.



(a) Twitter - Spectral and rewired netwwork.



(b) Facebook - Spectral and rewired netwwork.



(c) CGD - Spectral and rewired netwwork.

Figure 3.5: Network community profiles (red/dark) compared to profiles of rewired networks (green/light). The profiles of the rewired networks are different from those of the original networks. Recall, that for biological networks, we observe the opposite, the profiles of the rewired networks are the same as the originals.

3.3.2 Centrality differences between biological and other networks

We applied Spearman's rank correlation to determine the relation between the centrality measures used in the biological and social networks. We obtained Spearman's rank correlation for three centrality measures, betweenness (*nbc*), closeness (*ncl*), and degree (dg) for all biological networks (see Table 3.3), and several social and complex networks (see Table 3.4).

Networks	ncl/nbc	dg/nbc	dg/ncl
Arabidopsis thaliana	0.429	**0.863**	0.318
Caenorhabditis elegans 1	0.484	**0.926**	0.500
Caenorhabditis elegans 2	0.535	**0.963**	0.568
Drosophila melanogaster	0.832	**0.909**	0.865
Echericha coli 1	0.736	**0.916**	0.848
H. pylo	0.771	**0.969**	0.794
Homo sapiens 1	0.594	**0.852**	0.684
Homo sapiens 2	0.631	**0.864**	0.603
Mus musculus	0.520	**0.861**	0.475
Saccharomyces cerevisie	0.804	**0.859**	0.799
Schizosaccharomyces pombe	0.749	**0.867**	0.790

Table 3.3: Spearman correlation between centrality measures for biological networks.

Networks	ncl/nbc	dg/nbc	dg/ncl
Coauthor ships in science	0.382	0.486	**0.627**
AstroPhysics collaboration 1	0.650	0.714	**0.834**
AstroPhysics collaboration 2	0.562	0.645	**0.748**
Energy Physics, Phenomenology	0.523	0.624	**0.720**
Energy physics, Citation	0.472	0.596	**0.828**
Energy Physics, Theory	0.615	**0.805**	0.650
Condense Matter collaboration	0.558	**0.721**	0.698
R&Quantum Cosmology collab.	0.553	**0.676**	0.589
Enron email	0.516	**0.758**	0.506
Social circles: Facebook	0.479	**0.788**	0.430
Power grid western states 1	0.296	**0.804**	0.233

Table 3.4: Spearman correlation between centrality measures for social and complex networks.

The set of social and complex networks where selected because are the smallest of their type and can be compared with the biological networks that have similar sizes. The correlations for each comparison — ncl/nbc, dg/nbc, and dg/ncl — are presented in Table 3.3 and Table 3.4.

All networks exhibit a strong correlation between degree and betweenness centrality. However, we also observe that the biological networks show a significantly stronger correlation between degree and betweenness, and between closeness and betweenness (Figures 3.6, 3.7, and 3.8). This interesting observation suggests once more that the structure of these two families of networks is quite different, in contrast to the often held belief that they are pretty much the same in terms of structure.

In Figures 3.6, 3.7, and 3.8 we present the comparison of Spearman's rank correlations between biological networks and social networks. We observe that the degree and betweenness correlation is significantly higher than the other correlations. Also, the correlation is more pronounced for biological networks than for social networks (see in Figure 3.6). This indicates that despite the high correlation, in both cases there is a clear difference between biological and social networks.

Regarding the correlation between degree and closeness in Figure 3.7, both families of networks exhibit a similar behaviour, with the median for biological networks being slightly higher than the median for social networks.

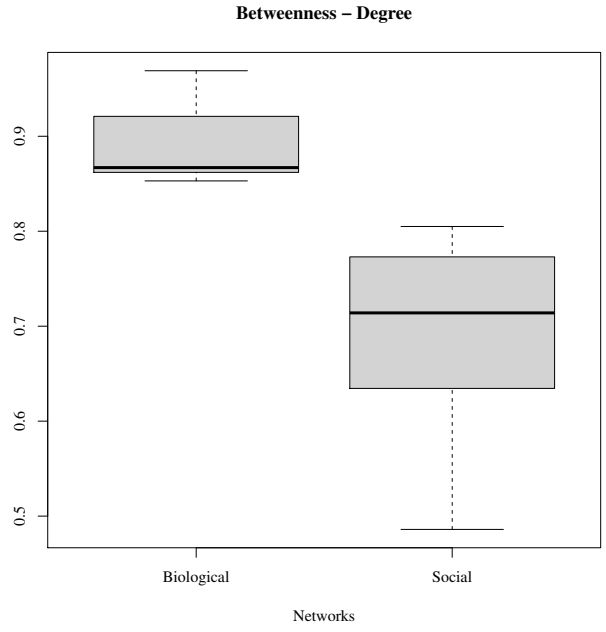


Figure 3.6: Comparison of Spearman's rank correlations between biological networks and social networks. Betweenness - Degree.

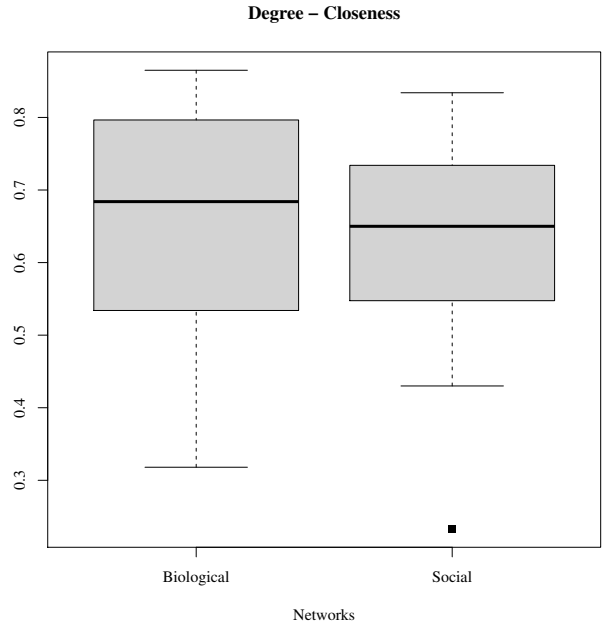


Figure 3.7: Comparison of Spearman's rank correlations between biological networks and social networks. Degree - Closeness.

Finally, betweenness and closeness correlation in Figure 3.8 shows a higher median for biological networks than for social and complex networks.

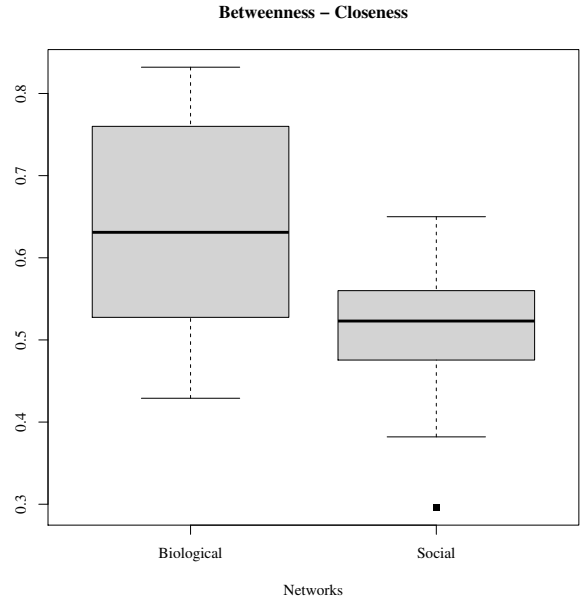


Figure 3.8: Comparison of Spearman's rank correlations between biological networks and social networks. Betweenness - closeness.

3.4 Conclusions and future work

We presented an empirical study on the fine-grained structural differences between biological networks (namely protein interaction networks) and social and other types of networks (around 100 or more). We revealed surprising differences in terms of the network community profile and correlations of centrality measures. More specifically, we showed that the best community size in terms of community conductance is at about a size value of ten, and this holds across almost all the available protein networks of a reasonable size. Such a community size is an order of magnitude lower than that for social and other networks.

The shape of NCPs for both biological and social networks is quite similar; they initially slope downward, then upward. This behaviour is different from that of other networks (neither biological nor social).

We can see from our centrality values that our protein networks are mature. The species used in this research have been studied for years.

Knowing the best size of a community in a PPIN, we can suggest to researchers to continue to focus on networks (work in the laboratory - *in-vivo*) where the size of the community is still not the best size ($\tilde{1}0$). Community sites of $k < 10$ indicate that the networks still do not have enough information (proteins or interactions).

As future work, we would like to extend our experiments to biological networks of other types, such us, metabolic, gene regulatory and neural networks, and examine a wider range of network measures at fine levels of granularity.

Chapter 4

Finding correlations between orthologs using centrality measures, percentage of similarity and rate of evolution

Sequences can be able to diverge during evolution. The amount of divergence between two sequences can tell us how closely two sequences are related (degree of similarity), and it reflects the evolutionary relationship between them [91].

Sequence divergence measures the total number of differences between two sequences. *Percentage of similarity* measures how similar two sequences are. With these two measures alone we cannot know the type of change (details of the changes) that occurred to the sequences. The higher the percentage of similarity between two sequences, the lower the number of changes in the amino acid sequence (dN/dS ratio) should be. And, the lower the percentage of similarity between two sequences, the higher the number of changes in the amino acid sequence should be. Finally, the *amino acid divergence* measures the type of changes (considering the gene position in the codon) that occurred in the sequence (dN/dS ratio or test for positive selection, see 2.1.5 for definition).

Our research is focused on the study of the evolution of protein-protein interactions (PPIs). To do this, we quantify amino acid divergence, percentage of similarity, and three traits associated (centrality measures) with protein-protein interactions for orthologous genes in different species. The centrality measures used are betweenness (bc), closeness (cl), and degree (dg). See section 2.2 for definitions. The species we considered are:

- fly (*Drosophila melanogaster* - dm),
- worm (*Caenorhabitis elegans* - ce),

- mouse (*Mus musculus* - mm), and
- human (*Homo sapiens* - hs).

We predict that the percentage similarity, dN/dS ratio divergence, and the three interaction traits will be correlated. Our prediction is based upon work published by Hahn and Kern [44]. They found that slow-evolving proteins (conserve selection) in *S. cerevisiae*, *D. melanogaster*, and *C. elegans* have high betweenness scores. Although their work provides a foundation for our research, there are several substantial differences: The first is the evolutionary time span being considered. In the work by Hahn and Kern, a slow-evolving protein was one that changed little between close relatives. For example, a slow-evolving protein in *D. melanogaster* was one that had a very similar ortholog in *D. pseudoobscura* (Dp) - both are flies, and a slow-evolving protein in *C. elegans* was a protein that had a very similar ortholog in *C. briggsae* (Cb) - both are worms (shown in Figure 4.1).

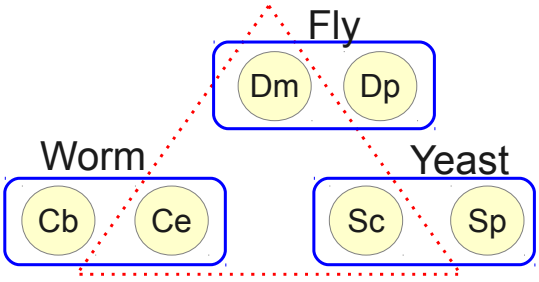


Figure 4.1: Relation between species. Red line (segmented), compare centralities between species. Blue line (continuous), alignments between species from the same family Ce with Cb, Dm with Db, and Sc with Sp [44].

Thus, despite including yeast, fly, and worm in their study, no comparisons (at the sequence level) across these species were made. A second difference is that Hahn and Kern did not report betweenness values for the protein from in the con-generic species used in the amino acid sequence comparisons. That is, while slow-evolving proteins in *D. melanogaster* tended to have high betweenness scores, we have no interaction data for these proteins in *D. pseudoobscura*.

In this research we identify the *same genes* (i.e., orthologs) in fly, worm, mouse and human, and compare sequence data to protein-protein interaction data (traits) over a much longer time scale. Another key strength of our approach is that we consider PPI data that are available for all species in this research. The three PPI measures were calculated from data downloaded from public databases. The main challenge in our approach is the identification of orthologs across such a broad evolutionary time scale.

How does gene duplication and divergence at the amino acid sequence level correlate with changes in the PPI parameters we have measured? If the interactions of a protein change over time it could mean that its sequence has changed too (the amino acids have changed) or its partners are different. In the case of changes at the amino acid level, the protein changes if mutations are non-synonym (big changes in the sequence), otherwise the protein does not change if mutations are Synonym (small changes in the sequence). This will be explained in more detail in the following sections.

4.1 Literature review

We will discuss how to compare PPINs among species using measures and tools developed for other areas, such as, math, sociology, social networks, and biology.

To understand the species and their systems it is not enough to identify the proteins in them, all the interactions between these proteins are needed to comprehend the species. For this reason, researchers started to study the interactions and trying to identify new ones [37]. The interactions can be classified according to the definition proposed by [88].

There are different methods to identify these interactions, such as, Y2H (Yeast two hybrid assay), MS (mass spectrometry), Correlated mRNA expression, and genetic interactions [28, 36, 114]. In [114] the authors compared some of these methods to study protein interactions. Since each method has its own benefits and drawbacks, their conclusion was that the combination of these methods is better than using them independently.

In [18], the authors made the comparison of four algorithms (Markov clustering, restricted neighborhood search clustering, super paramagnetic clustering, and molecular complex detection) to evaluate the resulting cluster from annotated complexes. In a protein interaction network, clustering is used as an effective approach for identifying protein complexes or functional modules [115].

The networks normally are represented by graphs where the nodes are proteins and the edges are their interactions. Using graph theory to analyze biological networks allows us to show properties and functions that were hidden in the network.

The use of graph theory in the study of the topology of biological networks is successfully applied by Pavlopoulos et al. [93]. They studied in detail the interactions, motifs and clusters using graph theory tools to explain better the behaviour of the networks. They concluded that there is a close relation between the structure of the network and the function of the nodes.

According to [61, 82, 93], the topology studies (structure of the networks) are limited because the data is from a specific time and does not reflect the dynamic system of a network.

In [22, 47, 98], the authors proposed that using a global network algorithm, it is possible to

work with only the network topology and no additional information.

According to Przulj [98], sequence and topology provide complementary ideas of biological knowledge.

There are also new algorithms for the identification of orthologs [69, 92]. The authors of [92] presented a new global alignment algorithm that evaluates functions and structure of each protein and its neighbors.

Most of the network analysis measures stems from graph theory [93]. Measures of centrality have an important role in the study of the different relationships between and within species. Centrality is measured as the connectivity of a protein, counting the total number of its interactions [54]. Greater connectivity means more direct contacts for the protein [31].

The analysis of networks using centrality measures permits to identify important elements between the proteins (nodes) and their interactions (edges). The three measures of centralities are based on: Connectivity (degree), closeness or betweenness. These are three measures that we will be using in a different part of the research (see section 2.2 for the definitions). In the case of biological networks, proteins in a good position can influence those proteins that do not have the same importance in the network [43, 44, 49, 116].

Betweenness (*bc*) is a centrality measure that indicates the frequency that a protein appears in the shortest way that connects two other proteins. Here we consider how many shortest paths are between these relevant proteins, because they control the flow of the network. According to Hahn et al. (2004), the structure and function of the protein have effects on the evolutionary rate in the protein networks [48]. Genes that are more central are more likely to be lethal when knocked out [53], and to evolve more slowly [43].

The idea is to identify which proteins are the most central in each species. So, identifying those proteins according to the centrality measures could permit to see patterns that could explain the evolution between these species.

Proteins with high betweenness are more likely to be essential and the evolutionary age of proteins is positively correlated with betweenness [57] (it in the center of the network). In the case that a protein has low centrality it is likely located in the periphery of the network. The authors exposed that rewiring of interactions via mutation is an important factor in the production of such proteins. In [59, 125], they use modularity based on *bc* values allowing to have a relationship between the functions of a network and its components.

These measures are used to identify essential proteins of PPINs. To analyze the PPIN from different species, centrality measures on the networks of all species are determined. Each network is used to determine the relative importance of a protein in the network. Moreover, we use the orthologous classification of these networks, this means we have a set of clusters composed by proteins (from different networks) grouped by percentage sequence similarity

and centrality measures. Therefore, the definition of orthologs is an additional and important factor in the analysis.

4.2 Methodology

Our goal is to determine if orthologs from two different species are similar with respect to their different traits. The traits that we are using are the centrality measures (see 2.2.2 for definitions), the sequence similarity, and the dN/dS ratio (see 2.1.5 for definition).

For a given species α we call the PPIN H_α. Each species is composed of proteins, the *protein set* is called W_α. Every protein in W_α is represented as $k_{\alpha i}$. When we compare the PPINs of two different species α and β, we refer to protein pairs as t_{ij} with $t_{ij} = \{k_{\alpha i}, k_{\beta j}\}$.

Using the general idea of Hahn and Kern's research we define: the species to be compared, the methods and constrains for the orthologous classification, the centrality measures and algorithms (to organize the data and integrate the data) to be used. All these will permit to find those proteins of H_α that has a high sequence similarity and a possible correlation with proteins of H_β.

Comparisons are done for two different species (α and β). First, Worm and Fly, and next Human and Mouse. These pairs of species were specifically chosen because they have good quality PPIs and their protein and cDNA sequences are available. We will carry out from the selection of the protein sequences, their alignments and subsequent orthologous classification using sequence analysis with BLASTp.

After finishing the comparison of the first two species, we reviewed our process and we realized that: the constraints used were not so stringent and this ultimately led to not having a fine orthologous selection; The orthologs obtained were not unique (a protein from specie A has more than one match in species B). We refine the constraints used to improve the orthologs classification. These constraints are necessary to keep only those orthologs with their best percentage of similarity. We create a procedure (see Figure 4.2) to identify orthologs, calculate the amino acid divergence and the centrality measures.

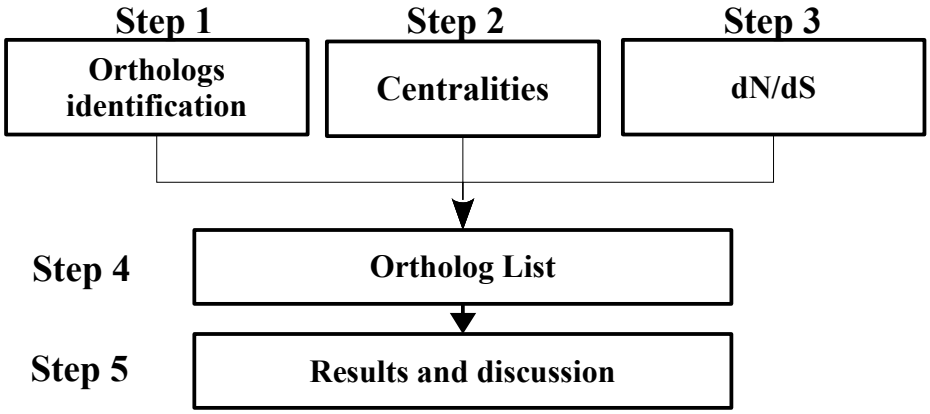


Figure 4.2: Methodology overview.

Figure 4.2 presents the five main steps of our work: Alignments of protein sequences, centrality measures, amino acid divergence (dN/dS ratio), merging of values, and the analysis of the results. In sections below, we present in detail the five steps used for the selection of the pairs of orthologs according to the proposed constrains. When we discuss about best orthologs we are referring to those pairs of proteins (A and B), where protein A is the best match to protein B, and the other way around. The best match means the pair with the higher percentage of sequence similarity. In section 4.3.5 we present the differences between the different experiments.

4.2.1 Step 1: Alignments

The goal of step 1 is to obtain a list of candidate orthologs from two different species that contain the best ortholog pairs for each protein (see Figure 4.3). From Ensembl [29] we took the protein sequences from both species. Later, we use BLASTp to do the reciprocal alignment (RBH).

BLAST is a basic local alignment search tool [86]. BLAST compares primary biological sequence information. It compares a query sequence (A) with an other sequence (B) to identify segments of the sequence (B) that resemble the query sequence A [3]. BLASTp compares specifically protein sequences (all with all).

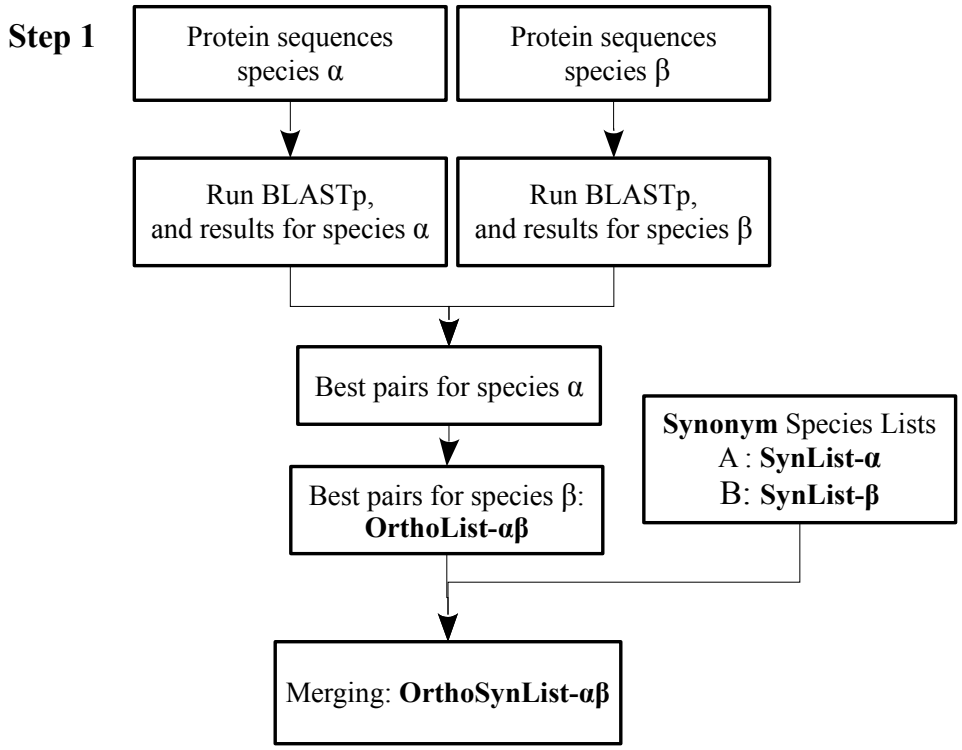


Figure 4.3: Step 1. Orthologous selection from pair of species.

For example, we have two species α and β that represent the two input files from BLASTp. The protein sequences from each species are $W_\alpha = \{k_{\alpha 1}, k_{\alpha 2}, k_{\alpha 3}, k_{\alpha 4}, k_{\alpha 9}\}$ and $W_\beta = \{k_{\beta 1}, k_{\beta 2}, k_{\beta 3}, k_{\beta 4}\}$. To identify the orthologs we need to execute the BLASTp in both directions (RBH). First $\alpha \rightarrow \beta$, were α is the query and β the subject. Second, we do the same, but we swap α and β to β is the query and α the subject. The idea is to obtain the best match for each protein taking in consideration the alignments from both directions, scores and percentage of similarity. The results from using BLASTp is a list of pairs with similarity scores, in this case we have two lists, one from each direction. Following the example, from the alignment results we have eighteen pairs with different similarity scores as follows:

$$
\begin{array}{lll}
\text{Direction } \alpha \rightarrow \beta: & \{k_{\alpha 1}, k_{\beta 2}\}: 40; & \text{Direction } \beta \rightarrow \alpha: \quad \{k_{\beta 1}, k_{\alpha 1}\}: 30; \\
& \{k_{\alpha 1}, k_{\beta 3}\}: 50; & \{k_{\beta 1}, k_{\alpha 2}\}: 40; \\
& \{k_{\alpha 1}, k_{\beta 4}\}: 60; & \{k_{\beta 1}, k_{\alpha 3}\}: 50; \\
& \{k_{\alpha 2}, k_{\beta 2}\}: 15; & \{k_{\beta 2}, k_{\alpha 4}\}: 45; \\
& \{k_{\alpha 3}, k_{\beta 3}\}: 80; & \{k_{\beta 3}, k_{\alpha 3}\}: 80; \\
& \{k_{\alpha 4}, k_{\beta 2}\}: 45; & \{k_{\beta 4}, k_{\alpha 1}\}: 60; \\
& \{k_{\alpha 2}, k_{\beta 1}\}: 40; & \{k_{\beta 2}, k_{\alpha 2}\}: 15; \\
& \{k_{\alpha 9}, k_{\beta 3}\}: 95 & \{k_{\beta 4}, k_{\alpha 8}\}: 30; \\
& & \{k_{\beta 5}, k_{\alpha 2}\}: 70; \\
& & \{k_{\beta 3}, k_{\alpha 9}\}: 95
\end{array}
$$

We create an algorithm that has five functions:

1. Concatenate the BLASTp results (get results from both sets);
2. Select best scores using the query species (the highest scores for each protein).
3. Select the maximum score without duplications;
4. Select the best score for subject species; and
5. Integration of formats.

The details of the five functions are as follows:

Concatenate BLASTp results

Initially, we took from Ensembl [29] two sets of protein sequences from different species (more exactly, we used all the sequence available). We run BLASTp with these sequences. The two sets of protein pairs (P) from BLASTp results are loaded together in memory using python.

BLASTp compares specifically protein sequences (all with all). Now, we need to take only those protein pairs t_{ij} with proteins from different species (the goal is to keep only candidate orthologs). Each t_{ij} has the following attributes: id species_1, id species_2, percent identity, alignment length, mismatches, gaps, query start and end position, subject start and end position, e-value, and bit score.

At this point, the proteins are not uniquely matched with proteins in the other species. This means, a protein might appear in more than one t_{ij}. We need to find the best match (that is, the one with highest similarity). Therefore, we organize the data by creating a group for each

of the proteins in H_α (the protein ID $k_{\alpha i}$ was used to label the groups). In each group there is one or more proteins from H_β that share similarities with this specific protein $k_{\alpha i}$.

The goal of concatenating and cleaning is to reduce the number of comparisons and duplications (this is possible to see in the results section). For each protein $k_{\beta j}$ included in a group it is added an attribute that indicate the direction of the query in BLASTp, $A = \alpha \rightarrow \beta$ or $B = \beta \rightarrow \alpha$. Following the example, here we have six groups and each consists of one to five proteins with its score values and direction.

$k_{\alpha 1}$: [[$k_{\beta 4}$, 60, A], [$k_{\beta 1}$, 30, B], [$k_{\beta 4}$, 60, B], [$k_{\beta 6}$, 60, B]]
$k_{\alpha 2}$: [[$k_{\beta 2}$, 70, A], [$k_{\beta 1}$, 70, A], [$k_{\beta 1}$, 70, B], [$k_{\beta 2}$, 70, B], [$k_{\beta 5}$, 70, B]]
$k_{\alpha 3}$: [[$k_{\beta 3}$, 80, A], [$k_{\beta 1}$, 50, B], [$k_{\beta 3}$, 80, B]]
$k_{\alpha 4}$: [[$k_{\beta 2}$, 45, A], [$k_{\beta 2}$, 45, B]]
$k_{\alpha 8}$: [[$k_{\beta 4}$, 30, B]]
$k_{\alpha 9}$: [[$k_{\beta 3}$, 95, A], [$k_{\beta 3}$, 95, B]]

Select best score for proteins in species H_α

Now, that the groups are sorted we use filters to keep only those t pairs with the highest score in each group. This is necessary due to the possibility of many $k_{\beta j}$ for one $k_{\alpha i}$. In this case, we take out the pairs $[k_{\alpha 1}, k_{\beta 1}, 30, B]$ and $[k_{\alpha 3}, k_{\beta 1}, 50, B]$ due to lower scores. This results in six groups, and each group contains elements of highest score only.

$k_{\alpha 1}$:[[$k_{\beta 4}$, 60, A], [$k_{\beta 4}$, 60, B], [$k_{\beta 6}$, 60, B]]
$k_{\alpha 2}$:[[$k_{\beta 2}$, 70, A], [$k_{\beta 1}$, 70, A], [$k_{\beta 1}$, 70, B], [$k_{\beta 2}$, 70, B], [$k_{\beta 5}$, 70, B]]
$k_{\alpha 3}$:[[$k_{\beta 3}$, 80, A], [$k_{\beta 3}$, 80, B]]
$k_{\alpha 4}$:[[$k_{\beta 2}$, 45, A], [$k_{\beta 2}$, 45, B]]
$k_{\alpha 8}$:[[$k_{\beta 4}$, 30, B]]
$k_{\alpha 9}$:[[$k_{\beta 3}$, 95, A], [$k_{\beta 3}$, 95, B]]

Select maximum score

Because we want to use the RBH to select the orthologous candidates we only need to keep the alignments or the pairs that have the same proteins, same score and different directions (A and B). For example, in $k_{\alpha 4}$ there are two t, $[k_{\alpha 4}, k_{\beta 2}, 45, A]$ and $[k_{\alpha 4}, k_{\beta 2}, 45, B]$, where in both cases the protein is $k_{\beta 2}$, the scores are the same, and the directions are different. We keep only one $[k_{\alpha 4}, k_{\beta 2}, 45]$ and this pair is one of the best match for both proteins ($k_{\alpha 4}$ and $k_{\beta 2}$) at this point.

$$
\begin{array}{lcl}
k_{\alpha 1}:[[k_{\beta 4},60,A],\,[k_{\beta 4},60,B] & & \\
k_{\alpha 2}:[[k_{\beta 2},70,A],\,[k_{\beta 1},70,A], & & k_{\alpha 1}:[k_{\beta 4},60] \\
\quad [k_{\beta 1},70,B],\,[k_{\beta 2},70,B]] & & k_{\alpha 2}:[k_{\beta 2},70] \\
k_{\alpha 3}:[[k_{\beta 3},80,A],\,[k_{\beta 3},80,B]] & \Rightarrow & k_{\alpha 3}:[k_{\beta 3},80] \\
k_{\alpha 4}:[[k_{\beta 2},45,A],\,[k_{\beta 2},45,B]] & & k_{\alpha 4}:[k_{\beta 2},45] \\
k_{\alpha 8}:[] & & k_{\alpha 9}:[k_{\beta 3},95] \\
k_{\alpha 9}:[[k_{\beta 3},95,A],\,[k_{\beta 3},95,B]] & &
\end{array}
$$

In the case of our groups $k_{\alpha 1}$, $k_{\alpha 2}$, $k_{\alpha 3}$, $k_{\alpha 4}$, and $k_{\alpha 9}$ group, $k_{\alpha 2}$ has two proteins with the same values $[k_{\alpha 2},k_{\beta 2},70]$ and $[k_{\alpha 2},k_{\beta 1},70]$. We choose one of them $[k_{\alpha 2},k_{\beta 2},70]$. The final groups are $k_{\alpha 1}$, $k_{\alpha 2}$, $k_{\alpha 3}$, $k_{\alpha 4}$, $k_{\alpha 9}$.

There is a possibility that a group contains duplicates (more than two pairs with the same score and same direction), the duplicates are removed.

Select best score for proteins in species H_β

Now that we have the maximum scores for each protein in H_α, we need to select the maximum score for each protein in H_β. We call a pair a final pair $(k_{\alpha i},k_{\beta j})$ when each protein is the best match of the other protein (for both proteins the other one is the highest score). We take the groups one by one and we organize them by proteins of H_β. So, three groups are created $k_{\beta 4}, k_{\beta 3}$ and $k_{\beta 2}$.

$$
\begin{array}{lcll}
k_{\alpha 1}:[\,k_{\beta 4},60\,] & & & \\
k_{\alpha 2}:[\,k_{\beta 2},70\,] & & k_{\beta 4}\text{: }[[k_{\alpha 1},60]] & k_{\beta 4}\text{: }[k_{\alpha 1},60] \\
k_{\alpha 3}:[\,k_{\beta 3},80\,] & \Rightarrow & k_{\beta 3}\text{: }[[k_{\alpha 3},80],[k_{\alpha 9},95]] & \Rightarrow \quad k_{\beta 3}\text{: }[k_{\alpha 9},95] \\
k_{\alpha 4}:[\,k_{\beta 2},45\,] & & k_{\beta 2}\text{: }[[k_{\alpha 2},70],[k_{\alpha 4},45]] & k_{\beta 2}\text{: }[k_{\alpha 2},70] \\
k_{\alpha 9}:[\,k_{\beta 3},95\,] & & &
\end{array}
$$

Here it is possible to obtain more than one pair for a protein $([k_{\beta 3},k_{\alpha 3},80],[k_{\beta 3},k_{\alpha 9},95])$. To deal with such situations we applied the same functions as aboce (point 2 and 3). We eliminate directions because all of them have already been verified. Finally, we have a set of protein pairs ($\{k_{\beta 4},k_{\alpha 1}\}$, $\{k_{\beta 3}k_{\alpha 9}\}$, and $\{k_{\beta 2}k_{\alpha 2}\}$) were each protein in the pair is the best for each other, this set is named "*OrthoList-* $\alpha\beta$".

Integration of formats

Our final goal is to correlate alignment score, centrality and divergence values from proteins of different species. The data sets need to be merged into one large data set to proceed to search for correlations. The data used for alignment scores and centrality measures are from different

sources. An inconvenience is how to integrate or merge the data when the proteins have different identification names. A match is not possible with these notations (original names). This is not an uncommon problem, Gabaldon mentioned in [34] that the lack of standardized formats makes comparisons or integration of data sets challenging and time-consuming.

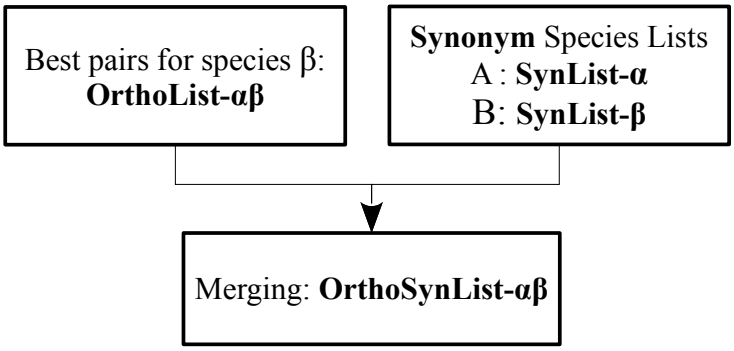


Figure 4.4: Integration of formats.

Each data set has the following format:

- alignments data format {ENS-id-α, ENS-id-β, ..., values} and
- centrality format {entrez-id, ..., values}.

It is necessary to obtain a list of synonyms that unifies these two formats. We got the lists of proteins for each species (α, β) from Ensembl (ENS) with the following format {ENS-id-α, ENS-id-β, entrez-id}. Next, we merged these three lists: *OrthoList-*$\alpha\beta$, *SynList-*α, *SynList-*β.

There are three types of results: First, there is no entrez-id for one of the two ensembl id ENS-id-α or ENS-id-β); second, there is no entrez-id for both proteins (ENS-id-α and ENS-id-β) and third, there is entrez-id for both proteins.

Here, we only use the third one because we need both to try to discover any correlation between the centrality measures from the species networks and its alignment relation. For this reason, there are orthologs that are not used in the next step. The three lists that are intersected, we created a new one *OrthoSynList-*$\alpha\beta$ that includes all the values from *OrthoList-*$\alpha\beta$ plus an extra field with the entrez-id for each protein from α or β.

4.2.2 Step 2: Centralities

Definitions and examples of centrality measures are discussed in section 2.2. A vertex with high *bc* (betweenness) has a strong influence over paths in the network. The *bc* value is not only a consequence of vertex positions, we can also see how centralized the graph is

[32, 33, 120]. Recall that *bc* of a vertex is the proportion of all geodesics between pairs of other vertices that include this vertex (that is, it measures how frequently a vertex appears on all shortest path between two other vertices in a network). The relationship between the *bc* of all vertices can expose much about the overall network structure.

The individual *bc* values show the type of control that a vertex can have in the network. If the network is low in density there is not much control, in the opposite case when the density is high there is the potential to control. In the case of biological networks, having a protein in a good position permits the protein to influence (that is produce changes with its interactions) the proteins that do not have the same importance [43, 44, 49, 116].

Using *bc* allows us to reveal proteins that, over evolutionary time, have conserved roles in the network. The centrality values (in the case of betweenness-*bc* and closeness-*cl*) in each species need to be normalized first (normalized betweenness-*nbc* and normalized closeness-*ncl*), to be able to compare them. Due to, the different network sizes (proteins and interactions). Because of the normalization, in the big networks the *nbc* or *ncl* values could be in average smaller.

For example, if an ortholog from two different species have *bc* values that are similarly high, this could suggest that the proteins conserve their network role. So, identifying these orthologs will allow us to see patterns (or any relation) that explain the evolution of different species. Further, this allows us to identify orthologs that have not conserved their network roles.

Now, the goal of this step is to calculate the centrality measures for all proteins of each species α and β (see Figure 4.5). For this research we are using the following centrality measures: degree, betweenness, and closeness. To be able to calculate the values it is necessary to have a network of the species. A network is composed of proteins ($k_{\alpha i}$) and the interactions between them (e.g. a protein-protein interaction $k_{\alpha i}, k_{\alpha x}$).

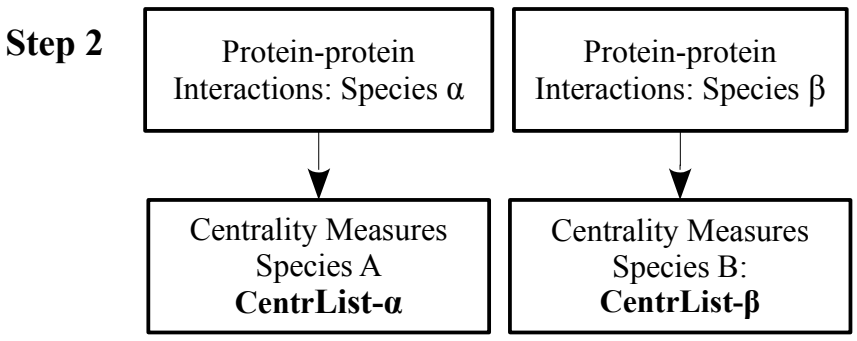


Figure 4.5: Step 2. Obtaining centrality measures.

For the construction of the network, a list of protein-protein interactions (PPI) were ex-

tracted for each species from BioGRID [30]. Every PPI pair has the next attributes: BioGRID Interaction ID and Pub Code. In the case of each protein has the next attributes: Entrez Gene Interactor, BioGRID ID Interactor, Systematic Name Interactor, Official Symbol Interactor. For our research, we are using all the attributes except the Systematic Name Interactor (an attribute that its not useful for our research). Each protein can be in more than one interaction (as the start or end) or be interacting with itself. Using the PPI, we get a list of unique proteins.

Now, that the data are complete, it is necessary to change it to Gephi's format[1] (new correlative identification names), a program that is used to represent the data and calculate the centralities [9]. For each protein we have the next attributes: Id, Label (Entrez Gene), Eccentricity (ξ), normalized closeness (*ncl*), and normalized betweenness (*nbc*). After the centrality values are calculated, we integrated them with the interaction attributes.

The result of this step are two lists of PPIs with the information from BioGRID and the centrality values. The final lists are *CentrList-*α and *CentrList-*β.

4.2.3 Step 3: Amino acid divergence.

In this step we calculate the *amino acid divergence* $\left(\frac{dN}{dS} ratio\right)$ for all the pairs resulting from step one ($OrthoSynList - \alpha\beta$, see Figure 4.3). Here we measure how often the average mutation in a gene is resulting in a change of the protein it produces.

Here we are using the same protein sequences from step one, a list of the pair interactions, and we the set of cDNA[2] (complementary DNA) sequence also from the Enembl site [29] (see Figure 4.6).

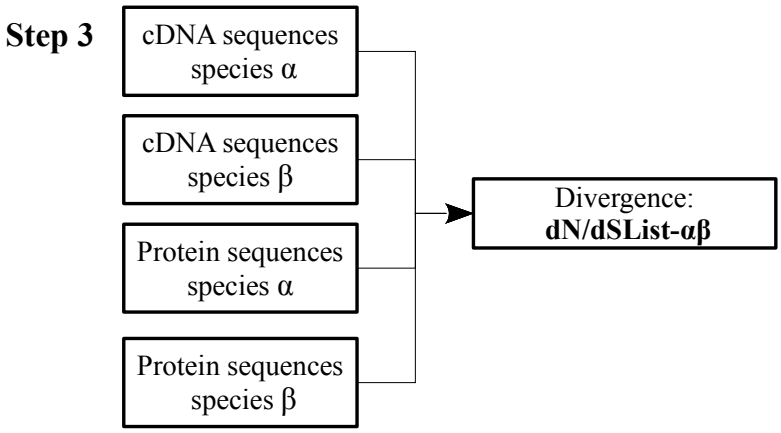


Figure 4.6: Step 3. Obtaining dN/dS ratio.

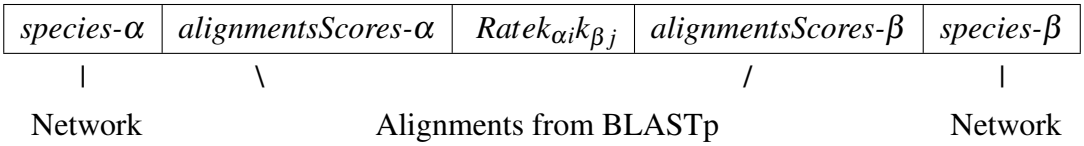
[1]Gephi, Graph Visualization and Manipulation software. http://www.gephi.org

[2]This *intron-free* DNA is constructed using *intron-free* mRNA as a template. It is, therefore, a *complementary* copy of the mRNA, and is called complementary DNA (cDNA)

To do the calculations we use the package PAML [123]. This program contains the algorithms to calculate synonymous (dS) and non-synonymous (dN) substitution rates using: Needleman—Wunsch dynamic programming algorithm for sequence alignment; scoring matrix BLOSUM50 with every possible substitution has assigned a score base on its observed frequencies in the alignment of related proteins. The BLOSUM50 matrix is used for alignments with gaps; PAL2NAL, a program that converts sequence alignment of proteins and the corresponding DNA (or mRNA) sequences into a codon alignment; and Codeml is a wrapper for estimating evolutionary rate (dN/dS) using two algorithms, first to implement the codon substitution model, and second to implement the amino acid substitution model. Both use maximum likelihood.

In our results we observed some errors, mainly because of the sequence files, where the proteins have some extra amino acids, or the initial or final amino acid is missing. So there are differences between the protein and cDNA sequences. In some cases the errors were not expected because the files came from the same server, but the data are from different versions and sources. This problem did not occur in step one because BLASTp recognized the problem and avoided using the extra amino acids. We discarded those pairs with errors.

Here the rate ratio results are integrated in one only set that contain a merge between *species*-α, *alignmentsScores*-α, *rate*-$\alpha\beta$, *alignmentsScores*-β, and *species*-β. The list is called *dN/dSList*-$\alpha\beta$.



Due to the presence of errors the resulting list *dN/dSList*-$\alpha\beta$ has fewer number of pairs that the initial list *OrthoSynList*-$\alpha\beta$. The list *dN/dSList*-$\alpha\beta$ will be integrated to the list resulting from step two and three on step four.

4.2.4 Step 4: Merging values

Finally with all the data obtained from alignments, centrality measures and sequence divergence we proceed to merge all this data from the two species, using as a connector the alignment obtained. This was done in three steps: first, alignment results with synonyms; second, centrality with the previous one; and finally this result with the divergence dN/dS ratio. We use this order to reduce the sizes of the data sets and the processing time (see Figure 4.7).

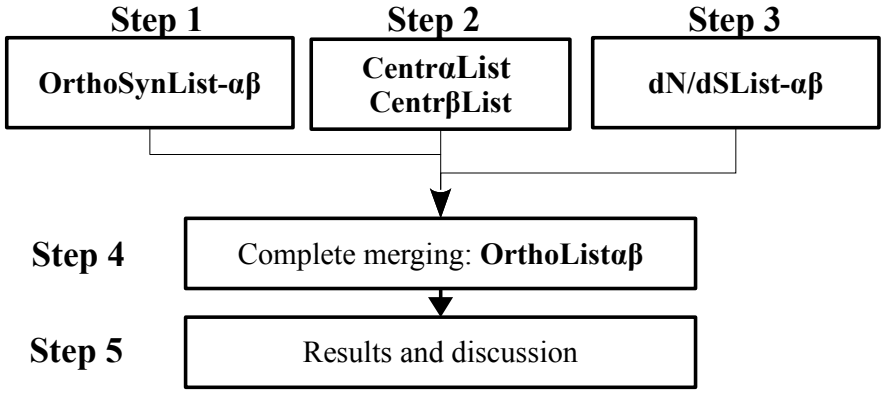


Figure 4.7: Matching the values from steps 1, 2, and 3.

- Step 1: *OrthoSynList-*$\alpha\beta$ has the following fourteen attributes:
  - $k_{\alpha i}$, $k_{\beta j}$,
  - $alignmentLength\, k_{\alpha i} k_{\beta j}$,
  - $PercentIdentity\, k_{\alpha i} k_{\beta j}$,
  - $mismatches\, k_{\alpha i} k_{\beta j}$,
  - $iniQuery\, k_{\alpha i}$,
  - $gap\, k_{\alpha i} k_{\beta j}$,
  - $endQuery\, k_{\alpha i}$,
  - $iniQuery\, k_{\beta j}$,
  - $endQuery\, k_{\beta j}$,
  - $BitScore\, k_{\alpha i} k_{\beta j}$,
  - $e-value\, k_{\alpha i} k_{\beta j}$,
  - $geneEntrez\, k_{\alpha i}$, and
  - $geneEntrez\, k_{\beta j}$.
- Step 2. *CentrList-*α and *CentrList-*βList has the following fifteen attributes:
  - $k_{\alpha i}, k_{\beta j}$,
  - $PercentIdentity\, k_{\alpha i} k_{\beta j}$,
  - $BitScore\, k_{\alpha i} k_{\beta j}$,
  - $e-value\, k_{\alpha i} k_{\beta j}$,
  - $geneEntrez\, k_{\alpha i}$,
  - $geneEntrez\, k_{\beta j}$,
  - $Label\, k_{\alpha i}$,
  - $Label\, k_{\beta j}$,
  - $\xi\, k_{\alpha i}$,
  - $\xi\, k_{\beta j}$,
  - $Cl\, k_{\alpha i}$,
  - $ncl\, k_{\beta j}$,
  - $nbc\, k_{\alpha i}$, and
  - $nbc\, k_{\beta j}$.
- Step 3. *dN/dSList-*$\alpha\beta$ has the following three attributes:
  - $k_{\alpha i}$,
  - $k_{\beta j}$,
  - $Rate\, k_{\alpha i} k_{\beta j}$

- Step 4: MergeList. The final list is with the following fourteen attributes:

  - $k_{\alpha i}$, $k_{\beta j}$,
  - $PercentIdentity k_{\alpha i} k_{\beta j}$,
  - $BitScore k_{\alpha i} k_{\beta j}$,
  - $e-value k_{\alpha i} k_{\beta j}$,
  - $geneEntrez k_{\alpha i}$,
  - $geneEntrez k_{\beta j}$,
  - $nbck_{\beta j}$,
  - $\xi k_{\alpha i}$,
  - $\xi k_{\beta j}$,
  - $Clk_{\alpha i}$,
  - $nclk_{\beta j}$,
  - $nbck_{\alpha i}$, and
  - $Rate k_{\alpha i} k_{\beta j}$.

Here we can say in detail that protein $k_{\alpha i}$ has a betweenness of $bck_{\alpha i}$ and it has a percentage of similarity of $PercentIdentity k_{\alpha i} k_{\beta j}$ and a rate ratio of $Rate k_{\alpha i} k_{\beta j}$, with protein $k_{\beta j}$ which has a betweenness of $Bck_{\beta j}$.

4.3 Results

We apply our methodology to two pairs of species: Human-mouse (hs-mm) and Fly-worm (dm-ce). For each pair, we performed the five steps explained in section 4.2.

4.3.1 Data management

Human and Mouse orthologs

We use four types of data sets for each species: number of cDna sequences, number of peptide sequences for orthologs and dN/dS ratio, protein-protein interactions (PPI), and synonyms files.

In step 1, we use only the protein sequences of size (see Figure 4.8):

Species	Human (hs)	Mouse (mm)
Number of peptide sequences*	95,639	50,877

$*$ Fasta format.

From the BLASTp results we obtain two files: one with 8,397,674 alignments pairs $\alpha\beta$ (the comparisons are all with all) and the second file with 3,668,268 alignments pairs $\beta\alpha$. After eliminating all the paralog pairs (focusing in orthologs) and selection of the best match for each protein (highest scores), the result is the *OrthoList*-$\alpha\beta$ of 7,173 alignments pairs (human-mouse).

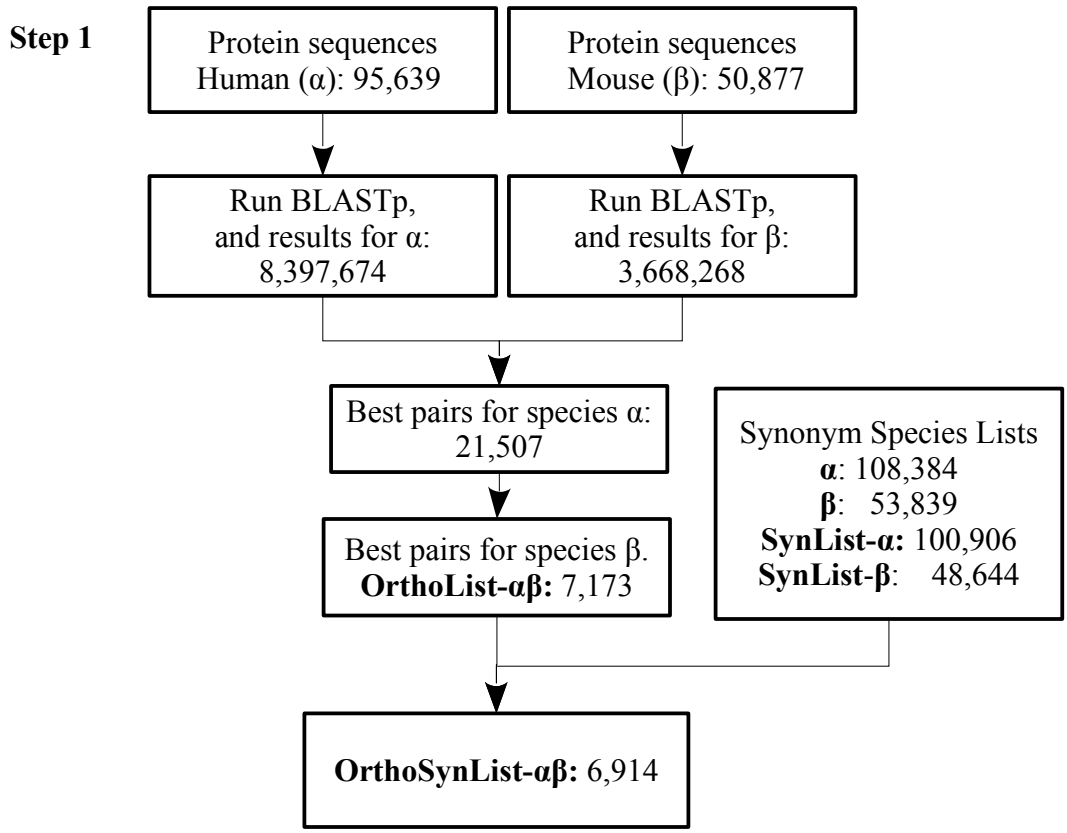


Figure 4.8: Step 1: obtaining orthologs of human and mouse.

Next, we merge these results with the synonym files (100,906 human and 48,644 mouse). The synonym files have a high number of proteins. Here, an inconvenience is that not all of the proteins are in *OrthoList-*$\alpha\beta$. We create a final file with 6,914 alignments pairs (*OrthoSynList-*$\alpha\beta$). Every pair has the data from BLASTp and we add the synonym name to be able to merge/concatenate this data with the centrality values that we obtain from (step 2).

In step 2, we use the PPI networks from Biogrid [30].

Species	Human (hs)	Mouse (mm)
Protein-protein interactions (PPI)	106,159	9,410

We calculate the three centrality measures (*betweenness*, *closeness*, and $degree - dg$) for both networks. In the case of *betweenness* and *closeness* the values were normalized (nbc, ncl). We obtain a set of 14,473 proteins in human and 4,555 proteins in mouse (see Figure 4.9).

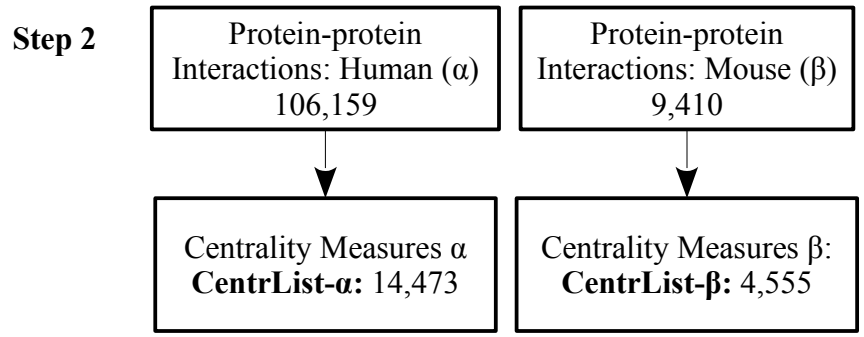


Figure 4.9: Step 2: obtaining centrality values of human and mouse.

In step 3 we use again the protein sequences from step 1, and the cDNA data sets of (see Figure 4.10):

Species	Human (hs)	Mouse (mm)
Number of cDna sequences*	211,716	92,484

∗ Fasta format.

For the dN/dS calculations we have a set of all combinations between the 211,716 cDNA with 95,639 of human and 92,484 cDNA with 50,877 of mouse.

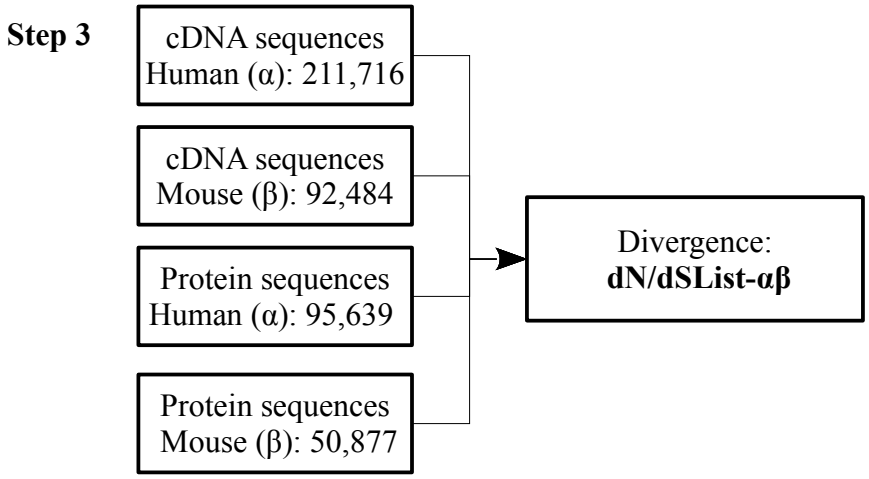


Figure 4.10: Step 3: obtaining dN/dS ratio values of human and mouse.

In step 4, we merge the results with the sequence divergence values (see Figure 4.11). In this case the final value is reduced from 1,168 to 947 pairs. Here we have problems with the sequences as we mentioned in section "Amino acid divergence" 4.2.3 (proteins have some extra amino acids, or the initial or final amino acid is missing). The total number is reduced because an inconsistency between a peptide and a nucleotide sequences.

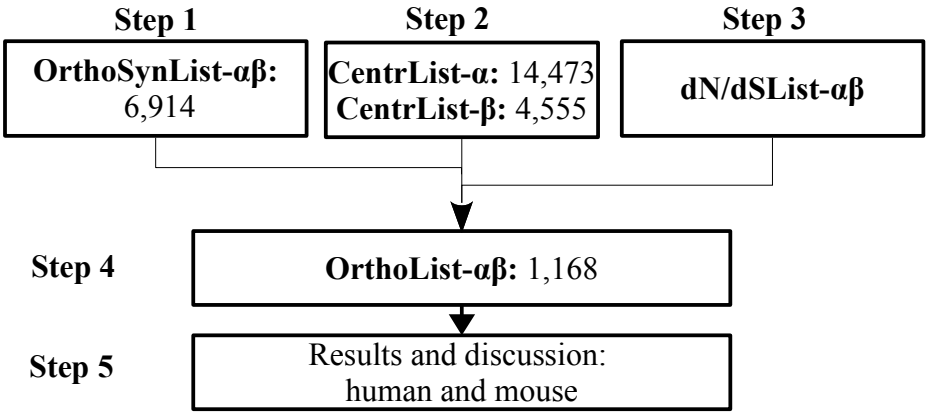


Figure 4.11: Step 4: Merge the three sets from step 1, 2, and 3 of human and mouse.

The final merge is a file with 947 alignments pairs. Every pair has an identity score, centrality values, and dN/dS ratio.

Worm and Fly orthologs

We use the same four types of data sets for each species: number of cDna sequences, number of peptide sequences, protein-protein interactions (PPI), and synonyms files.

In step 1, we use only the protein sequences of size (see Figure 4.12):

Species	Worm (ce)	Fly (dm)
Number of peptide sequences*	31,234	24,719

$*$ Fasta format.

From the BLASTp results we obtain two files: one with 2,019,871 alignments pairs (the comparisons are all with all) and the second file with 2,563,523 alignments pairs. After eliminate all the paralog pairs (focus in orthologs) and selection of the best match for each protein (highest scores), the result is the *OrthoList-*$\alpha\beta$ of 4,316 alignments pairs (worm-fly).

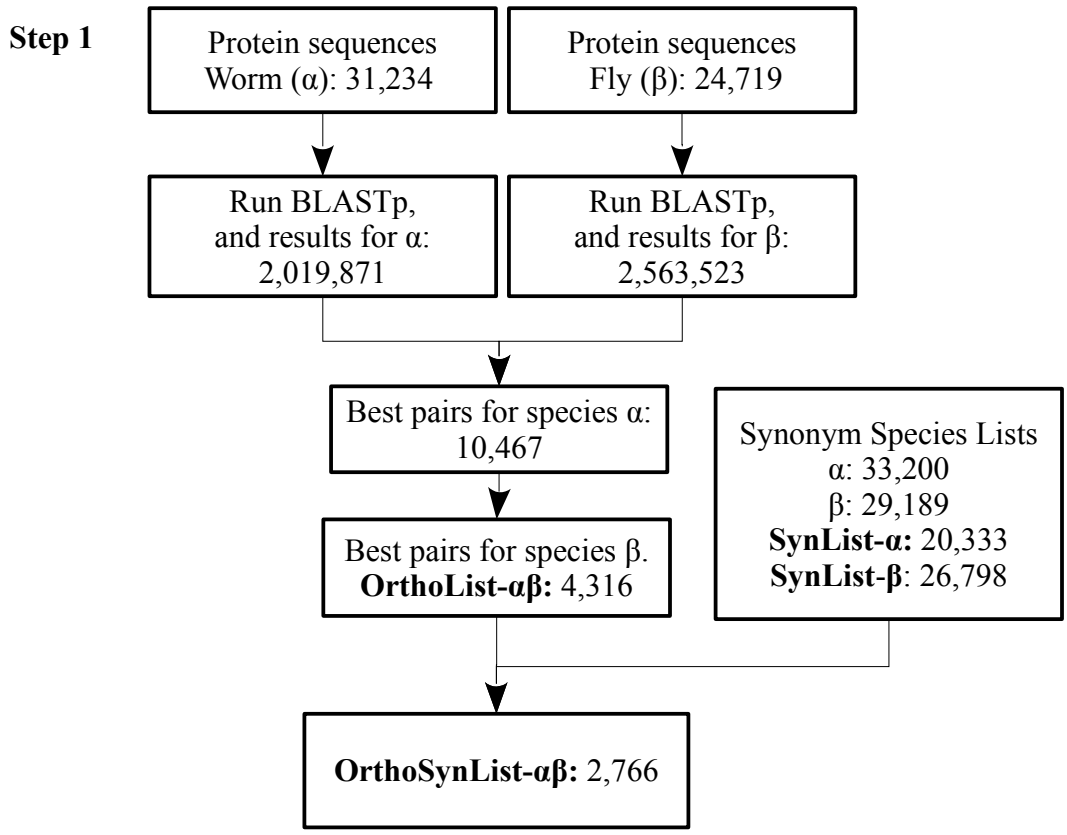


Figure 4.12: Step 1: obtaining orthologs of worm and fly.

Next, we merge these results with the synonym files (20,333 worm and 26,798 fly). We create a final file with 2,766 alignment pairs (*OrthoSynList-*$\alpha\beta$). Every pair has the data from BLASTp and we add the synonym name to be able to merge/concatenate this data with the centrality values that we obtain from step 2.

In step2, we use the PPI networks from Biogrid [30].

Species	Worm (ce)	Fly (dm)
Protein-protein interactions (PPI)	7,215	34,798

We calculate the three centrality measures (*betweenness*, *closeness*, and $degree - dg$) for both networks. In the case of *betweenness* and *closeness* the values were normalized (nbc, ncl). We obtain a set of 3,611 proteins in worm and 7,568 proteins in fly (see Figure 4.9).

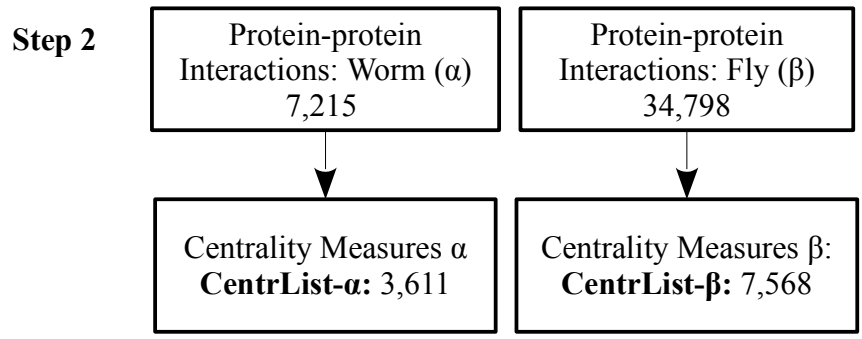


Figure 4.13: Step 2: obtaining centrality values of worm and fly.

In step 3 we use again the protein sequences from step 1, and the cDNA data sets of (see Figure 4.14):

Species	Worm (ce)	Fly (dm)
Number of cDna sequences*	57,844	29,173

∗ Fasta format.

For the dN/dS calculations we have a set of all combinations between the 57,844 cDNA with 31,234 of worm and 29,173 cDNA with 24,719 of fly.

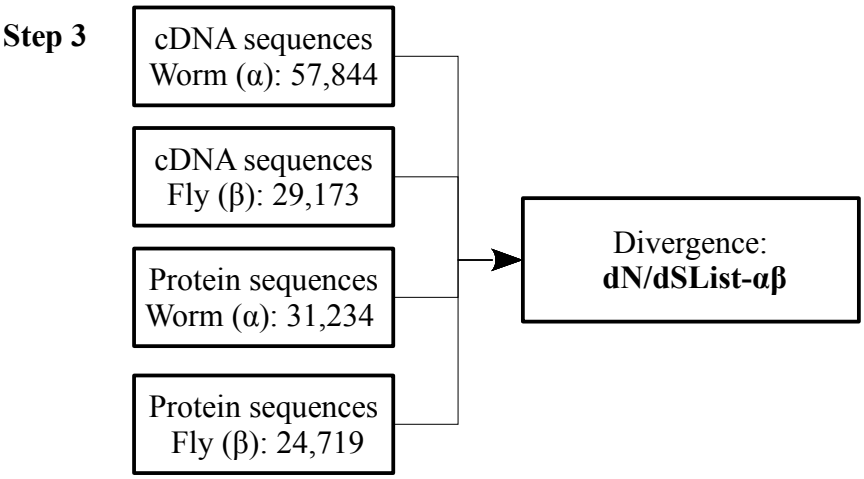


Figure 4.14: Step 3: obtaining dN/dS ratio values of worm and fly.

In step 4, we merge the results with the sequence divergence values (see Figure 4.15), in this case the final value is reduced for only 1 pair from 582 to 581 pairs. The total number is reduced because an inconsistency between a peptide and a nucleotide sequences.

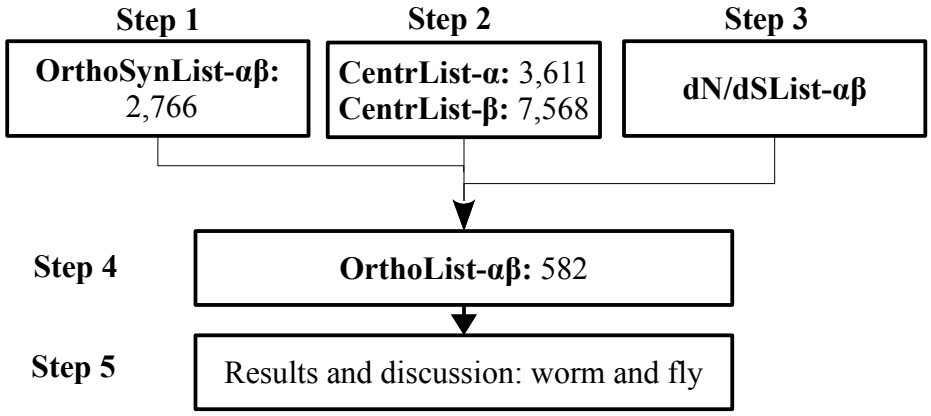


Figure 4.15: Step 4: Merge the three sets from steps 1, 2, and 3 of worm and fly.

The final merge is a file with 581 alignments pairs. Every pair has an identity score, centrality values, and dN/dS ratio.

After finishing the 5 steps methodology, we proceed to analyze the two sets of orthologs, human-mouse and worm-fly. In the next sections we describe and interpret the data using the different measures, centrality, percentage of similarity, and dN/dS ratio.

We start evaluating each measure separately to be able to identify the individual characteristics of the data. For a summary of the networks' data see Table 4.1.

	Worm	Fly	Human	Mouse
Protein	3,611	7,568	14,476	4,555
Interactions	7,215	34,708	106,159	9,410

Table 4.1: Network information.

The networks present different sizes in proteins and interaction number. We calculate the modularity class for the four species to see how sparse or concentrated the proteins are in the network. There are 29 and 48 classes (subnetworks) in fly and worm respectively. This could indicate that the proteins in worm are more disperse than the proteins in the fly. Human and mouse present a different observation, both have a similar distribution between number of classes and proteins (88 and 19 classes respectively).

4.3.2 Centrality measures

In this section we describe the results from the three centrality measures betweenness, closeness, and the degree (see section 2.2 for definitions). The number of pairs evaluated is reduced because we identify proteins with the same genes in one species. For example, there are two

orthologs A-A' and B-B'. (A, B) from species one and (A', B') from species 2. In this case the names are different (A,B) but the genes associated with A and B are the same. Because of this we expect similar *bc* and degree for both.

For each centrality we evaluate the data in the species and in the pair of species.

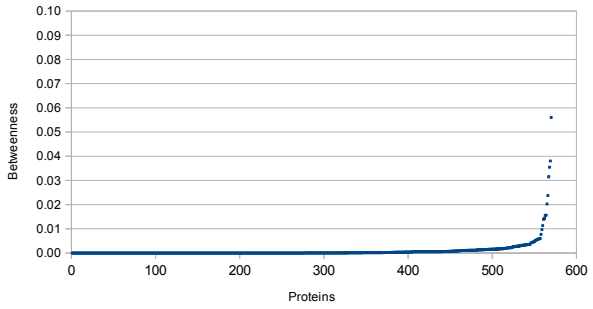
4.3.2.1 Betweenness

In Table 4.2 we have a summary of the betweenness values for the four species. In all the cases there are proteins with betweenness zero. This means, the proteins are in the periphery area or not-well connected in the network and therefore do not occur in many paths. The human species has almost 49% of proteins with betweenness zero, mainly because human is the species with higher number of proteins with respect to the number of interactions. The average betweenness is the same for all the networks, although they have different sizes (proteins and interactions) mainly because of the outliers. The median shows that most of their values are concentrated in a low range. We need to consider that. when large networks are normalized the values could be very low.

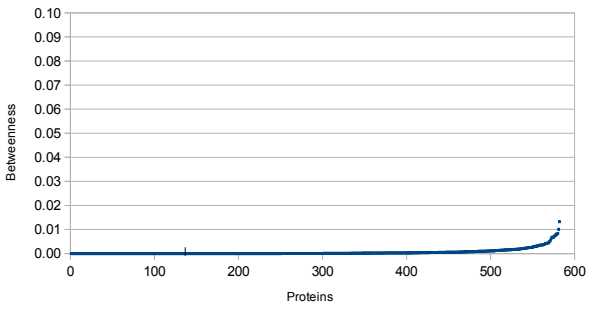
Betweenness	Worm	Fly	Human	Mouse
Minimum	0	0	0	0
Maximum	0.28	0.01	0.85	0.13
Average	0.001	0.001	0.001	0.001
Median	0.0003	0.0001	0.0001	0.0000001
Standard deviation	0.01	0.001	0.03	0.01

Table 4.2: Betweenness statistics.

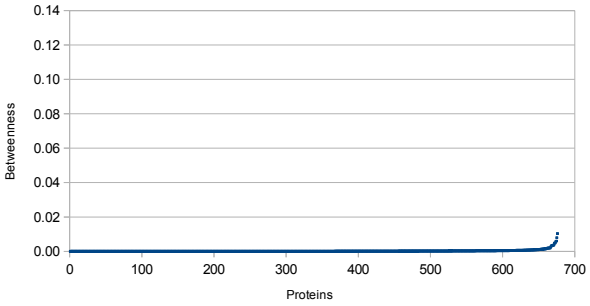
We can see from Table 4.2 that the proteins in these networks do not act as bridges between clusters in the network, as the proteins have low betweenness centrality. In this case, we can say that the majority of the proteins does not participate in shortest paths that necessarily have to go through them (that is they are not so central on average). In Figure 4.16, we have a representation of the betweenness values for the four species. We can see that the pattern of low values is repeated in all four, except for the high values in the right side of the worm and fly charts (between 0.01-0.08 in worm and between 0.01-0.14 in mouse).



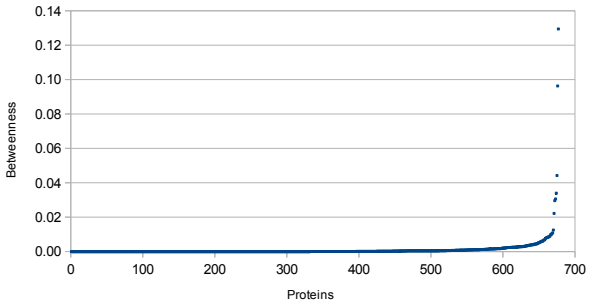
(a) Worm.



(b) Fly.



(c) Human.



(d) Mouse.

Figure 4.16: Betweenness centrality in the four species.

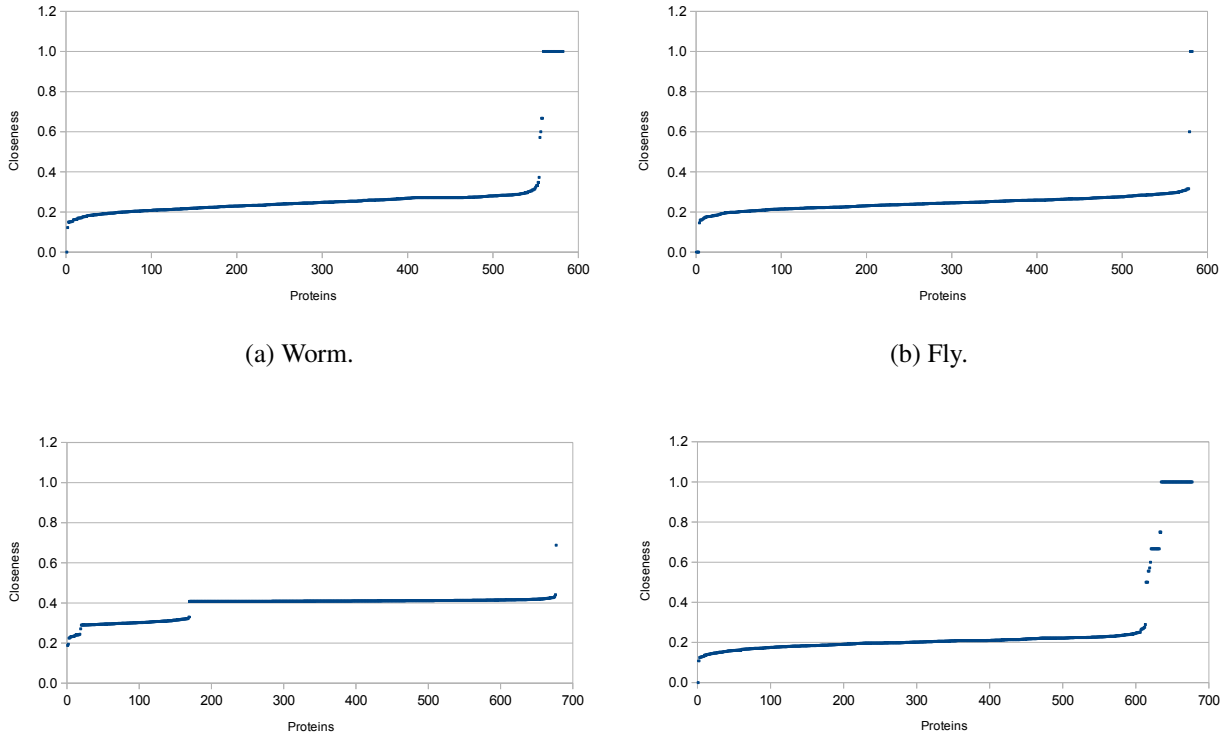
4.3.2.2 Closeness

As we did in the section above, in Table 4.3 are the closeness values for the four species. The average of the closeness is low because most of the values are concentrating on a low (median) range. Except for human that presents higher values, the average is similar to the median. According to the standard deviation with 0.05 the rest of the values are not dispersing with respect to the mean.

Closeness	Worm	Fly	Human	Mouse
Minimum	0	0	0.19	0
Maximum	1	1	0.69	1
Average	0.27	0.25	0.38	0.26
Median	0.25	0.24	0.41	0.21
Standard deviation	0.16	0.07	0.05	0.21

Table 4.3: Closeness statistics.

In Figure 4.17 we can see that the four charts present outliers, that could change the values obtained (average and standard deviation).



(a) Worm.

(b) Fly.

(c) Human.

(d) Mouse.

Figure 4.17: Closeness centrality in the four species.

4.3.2.3 Degree

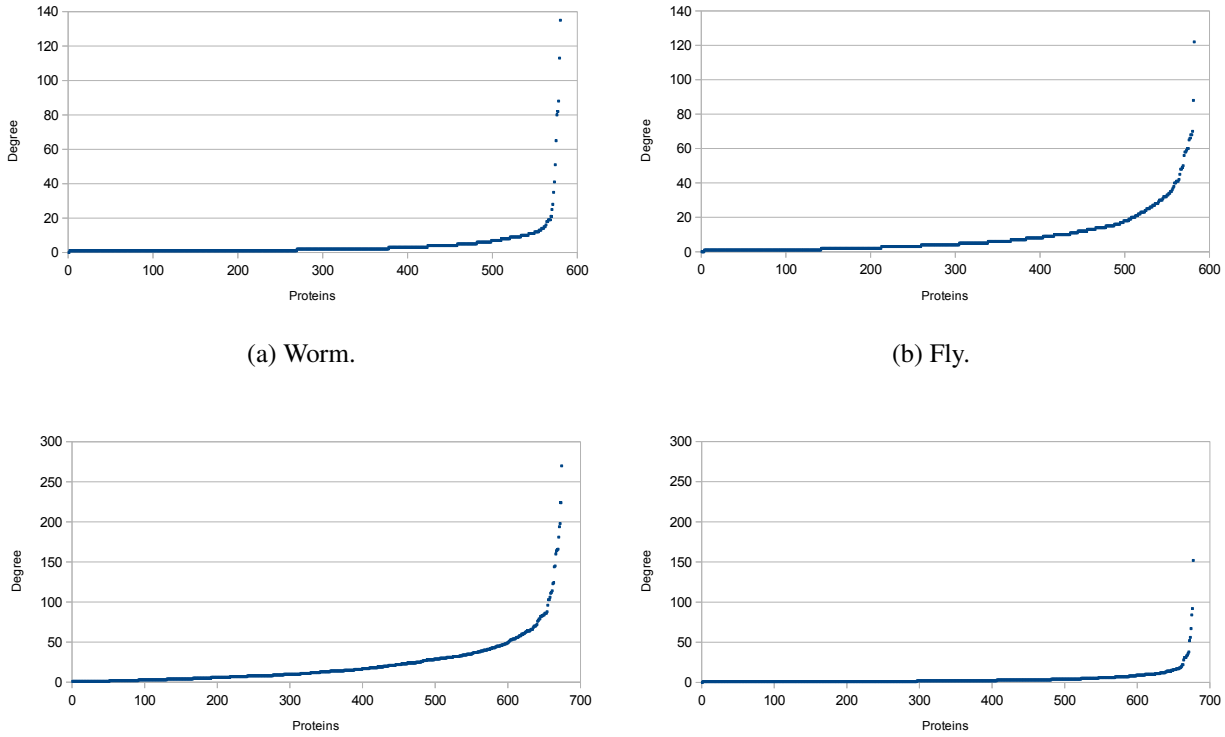
In Table 4.4 we have the summary of the degree values for the four species. The average degree for human is higher than the rest of the species mainly because of the size of the network, but also because of the high number of interactions known. However, we need to consider that the highest degree for human is only one protein (isolated case), the next highest degree is 417, a number that is closer to the degree of the other species.

Degree	Worm	Fly	Human	Mouse
Minimum	0	0	0	0
Maximum	524	122	8605	152
Average	5.5	9.15	39.8	4.36
Median	2	4	12	2
Standard deviation	25.34	13.21	331.75	9.48

Table 4.4: Degree statistics.

Worm and fly networks

The worm has the highest degree (with a few proteins) but fly has a greater number of proteins with high degree. The fly's network has three proteins with degree zero where their orthologs in the worm network have degree one. In the case of worm's network there is only one protein with degree zero and its orthologs in fly network has degree six. Over all the pairs of proteins (582) there are 80 orthologs where both have degree 1. Also, there are 100 orthologs with the same degree value (the range is between degree 1 and 18). There are 104 orthologs with the difference of one degree. This means the proteins did not change their degree although these are in different species. However, there are 292 orthologs that made changes in their degree with a difference between 2 and 15. Of two orthologs the variation of the degree is over 200. It is important to mention that the highest degree in fly is 122. In the case of worm the three highest degrees are 524, 202, 135. In both species these values have the highest *nbc* of the respective network.



(a) Worm.

(b) Fly.

(c) Human.

(d) Mouse.

Figure 4.18: Degree centrality in the four species.

Human and mouse networks

Mouse network has one protein with degree zero where their orthologs in the human network has degree 56 with 100% percentage of similarity. In the case of human network there are two proteins with degree zero and its orthologs in the mouse network both have degree 1. The degree values in mouse are gradually increasing until the 10^{th} highest degree from 52 to 152. In the case of human network the 10^{th} highest degree range from 300 to 8000.

The highest degree in human network is 8605, degree of two proteins (ENSP344818 and ENSP442800), also they have the same *nbc* (0.85). These proteins have different pairs of proteins (ENSMUSP19649 with ENSP344818; and ENSMUSP115578 with ENSP442800), both cases have 100% sequence similarity. Overall the pairs of proteins, there are 53 pairs where both have degree 1. Also, there are 79 pairs with the same degree between 1 and 10 (26 of those are between 2 and 10). In this case, proteins with high degrees have high *nbc*, those with low degree have low *nbc*. The separation was made according to the number of repetitions of the differences.

It is important to mention that the highest degree in mouse is 152. In the case of human

the three highest degrees are 8605, 417, and 338. In both species these values have the highest *nbc* of the respective network. Except for a protein in human that has degree 144 and the *nbc* is the second highest.

4.3.2.4 Difference between the measures

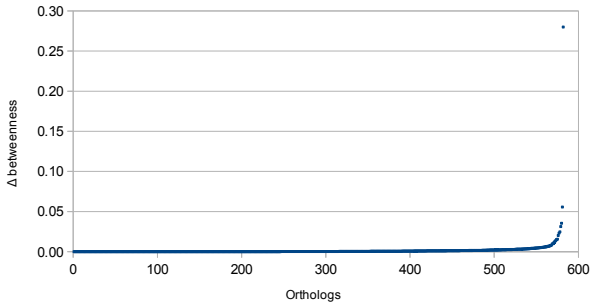
We create three new attributes to observe how close or far the centrality values are between these orthologs.

- $\Delta nbc = |nbc\ (species_\alpha) - nbc\ (species_\beta)|$,
- $\Delta ncl = |nbc\ (species_\alpha) - nbc\ (species_\beta)|$, and
- $\Delta dg = |nbc\ (species_\alpha) - nbc\ (species_\beta)|$.

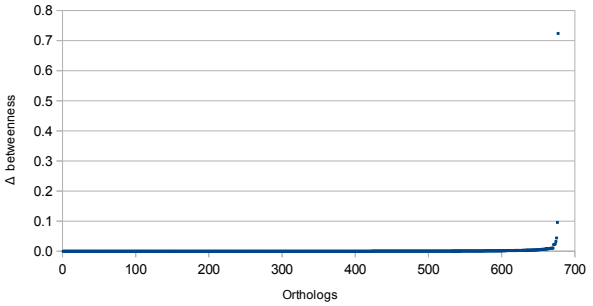
These attributes are the differences between each measure from the pairs of species.

In Figure 4.19 are the representation of the three deltas centrality measures for the orthologs. The smaller the Δvalue is, the more similar the values are between the orthologous proteins. Here in Figure 4.19b and 4.19b, we can observe that almost all the betweenness values are close to zero. This means their values are almost the same.

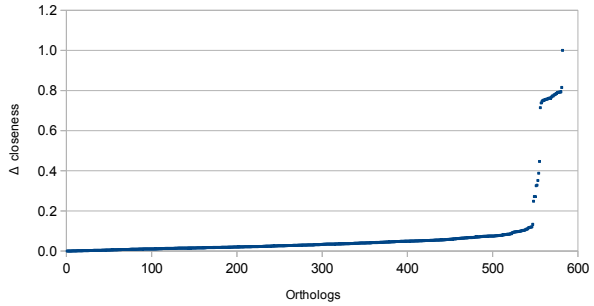
The orthologs with larger differences are in human-mouse (see Figure 4.19d) where the closeness shows a different behaviour when compared with the rest of the charts. It is important to mention that although the values are very low (close to zero), there are variations in the values (not all the values are the same).



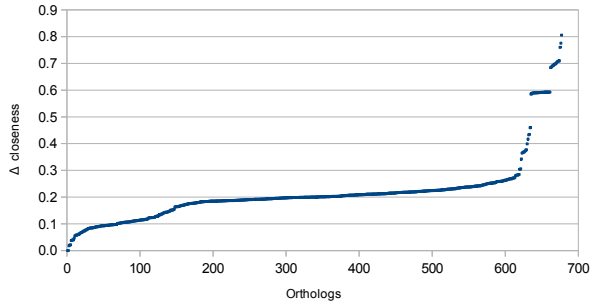
(a) Δ*nbc* worm-fly.



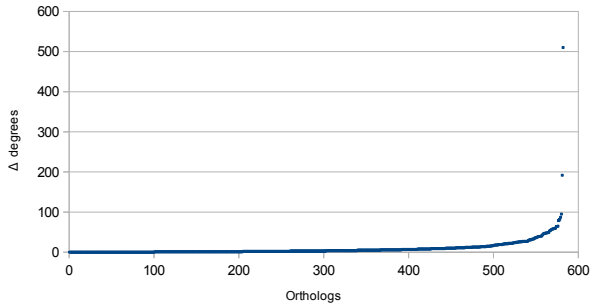
(b) Δ*nbc* human-mouse.



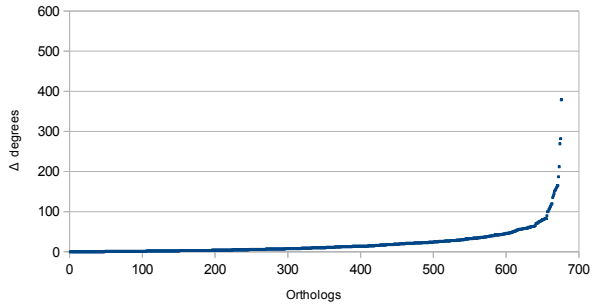
(c) Δ*ncl* worm-fly.



(d) Δ*ncl* human-mouse.



(e) Δ*dg* worm-fly.



(f) Δ*dg* human-mouse.

Figure 4.19: Difference between centrality measures.

Using the data of the deltas, we create categories for one of the centrality measures, the Δ*nbc* values. Every category has different ranges, the idea is to be able to obtain more information than the charts (see Table 4.5).

Range	Worm-Fly		Human-Mouse	
	Frequency	Percentage	Frequency	Percentage
0	85	14.6%	81	6.93%
1.00E-09-1.00E-08	0	0%	2	0.17%
1.00E-08-1.00E-07	0	0%	19	1.63%
1.00E-07-1.00E-06	4	0.69%	40	3.42%
1.00E-06-1.00E-05	34	5.84%	101	8.65%
1.00E-05-1.00E-04	94	16.15%	303	25.94%
0.0001 -0.001	203	34.88%	403	36.82%
0.001 -0.01	149	25.60%	175	14.98%
0.01 -0.09	12	2.06%	14	1.20%
0.09	1	0.17%	3	0.20%
Total	582	100%	1168	100%

Table 4.5: Number of pairs of proteins categorized by difference of betweenness.

The Δnbc values are very small (Table 4.5 column Range), this is a consequence of the normalization when calculating the centrality measures. From the data we can see that worm and fly have an important number of proteins (85 of 582) that do not differ, 14.6%, that means proteins have same *nbc*. As we have mentioned before, this could mean that the orthologs have the same betweenness or both have value zero. This is very different in human-worm, where the big group is in the range 0.0001 - 0.001 with a 36.82% difference over the total. Moreover, there is 6.93% with the same *nbc*. Comparing both pairs of species the differences are similarly distributed. The pairs with both *nbc* values equal to zero were not considered for. The remaining pairs, all have for one of the two proteins nonzero *nbc*.

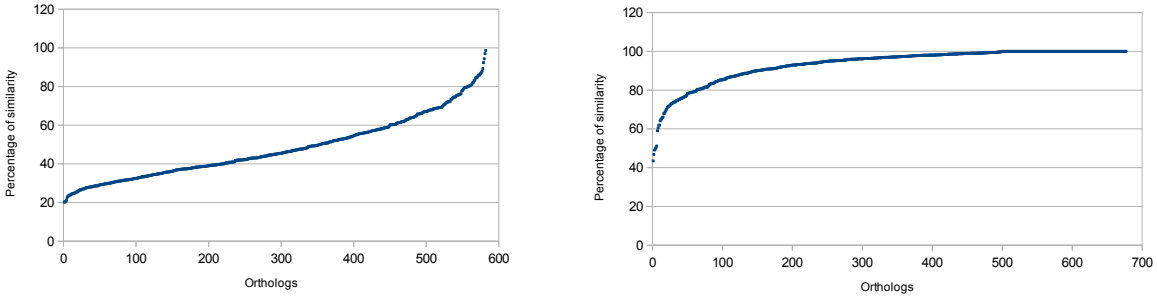
4.3.3 Percentage of similarity

These values were calculated using BLASTp. In Table 4.6 is a summary of the data obtained from the alignments.

Similarity	HS/MM	CE/DM
Minimum	43.48	20.19
Maximum	100	98.68
Average	93.46	48.03
Median	96.95	44.98
Standard deviation	9.06	15.83

Table 4.6: Statistics of percentage of similarity between pairs of species.

We can conclude that most of the orthologs in human-mouse have high similarity, contrary to what happens in worm-fly. Also, the similarity values in human-mouse are closer (low standard deviation), this means, are not disperse. See Figure 4.20b.



(a) Worm-fly.

(b) Human-mouse.

Figure 4.20: Percentage of similarity.

Here, we also create categories where the data were classified according to the percentage of similarity as shown in Table 4.7. First, we count the number of pairs for each category and then calculate the percentage of this number over the total number of pairs. The majority of the similarity values are close to 100% and in the 100 percentage category (see Figures 4.20a and 4.20b). If we do a protein analysis (see every protein sequence), it would be better to choose a limit or threshold over the 75% percentage of similarity so the analysis could be made deeply in less number of pairs.

Range	Worm-Fly		Human-Mouse	
	Frequency	Percentage	Frequency	Percentage
0 - 19	0	0%	0	0%
20- 29	63	10.82%	0	0%
30- 39	154	26.46%	0	0%
40- 49	135	23.20%	5	0.43%
50- 59	97	16.67%	6	0.51%
60- 69	73	12.54%	25	2.14%
70- 79	35	6.01%	77	6.59%
80- 89	21	3.61%	154	13.18%
90- 99	4	0.69%	649	55.57%
100	0	0%	252	21.58%
Total	582	100%	1168	100%

Table 4.7: Number of pairs of proteins categorized by percentage of identity.

The high ranges of human-mouse (between 80-100) contain more than 90% of the alignments pairs. Moreover, the range between 90-99% has more than 50% of the orthologs (649 of 1168). In worm and fly there is not a high percentage of identity between these two species (see Figure 4.7). More than the 50% of the pair alignments are between 30-59 percentage of identity. The high ranges (between 80 - 100) only have a 4.3% of the pairs. We could conclude according to these results that human and mouse are similarly closer in evolution (high percentage of similarity) than worm and fly (low percentage of similarity).

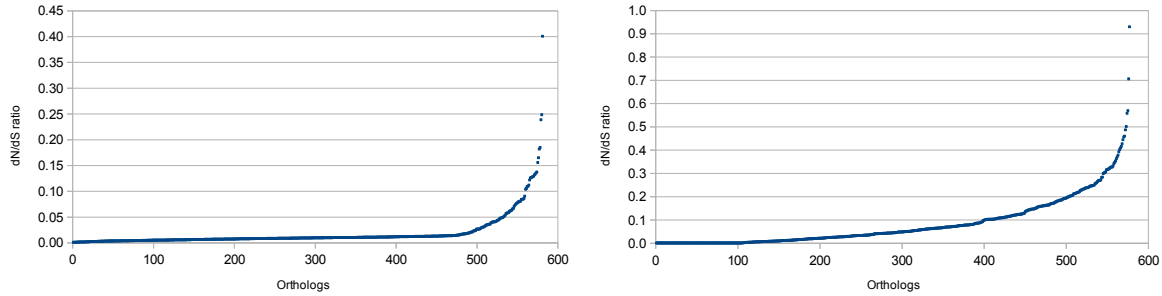
4.3.4 Divergence: dN/dS ratio

In Table 4.8 we present the summary of the values obtained in the dN/dS ratio. Here we observe that most of the orthologs do not present big changes. When the dN/dS ratio is small (dN/dS<1) means that there are not so many changes at the protein sequence level (purifying selection).

dN/dS ratio	Worm-fly	Human-mouse
Minimum	0.001	0.001
Maximum	0.40	0.93
Average	0.02	0.09
Median	0.01	0.04
Standard deviation	0.03	0.11

Table 4.8: dN/dS ratio statistics.

Now, if we compare the results from both pairs of species, we would expect low values in human-mouse because of their similarities. However, the values for worm-fly are very low, close to zero. Our interpretation of these results is that the proteins are the same in both species, they have not undergone any change in their amino acid sequence since the event of speciation. This is very unusual. These lower values will need a further evaluation, which we proposed as future work. There is a very small set of orthologs (see Figure 4.21 a and b) that contains outliers with higher dN/dS ratio values.



(a) Worm-fly.

(b) Human-mouse.

Figure 4.21: dN/dS ratio.

Next, we present more details of the data obtained from the dN/dS ratio. The highest degree (ortholog ENSMUSP104347 ENSP379180, 100% - 152) in mouse has a low $dN/dS = 0.001$. The highest degree (ortholog ENSMUSP19649 ENSP344818 100% - 8605) in human has a $dN/dS = 0.003$, below the median.

The highest degree in fly (FBpp0081600 ZK792.6 76.19% -122) has a low dN/dS of 0.002 The highest degree in worm (FBpp0070141 C32F10.2 23.81% - 524) has a $dN/dS = 0.0111$, above the median.

The general observation with respect to the dN/dS results is that both pairs of species (fly-worm and humna-mouse) have a low divergence with a rate below 1 (purifying selection).

That the sequences tend to have low betweenness values is interesting. As we mentioned, few connections appear in the proteins on the periphery of a network, which gives them the freedom to vary. When the proteins are highly connected, the ability to change is constrained. Orthologs with a low dN/dS means that the amino acids are not to free to made changes. Now, we evaluate the centrality measures with respect to the other two measure percentage of similarity and dN/dS ratio.

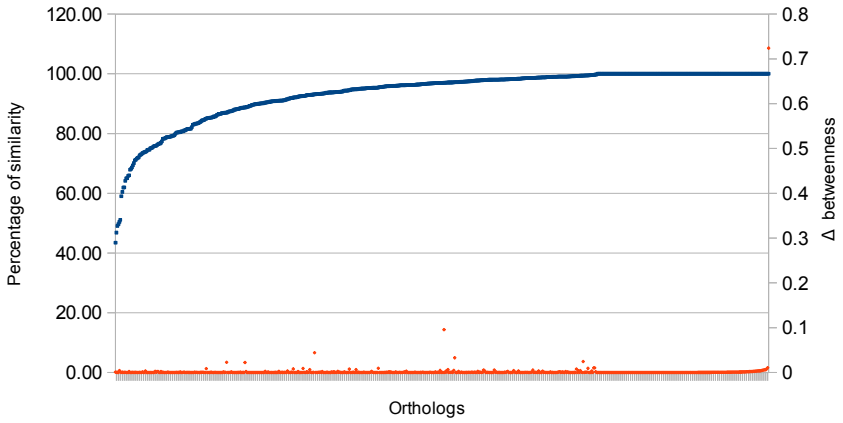
4.3.5 Centrality measures, percentage of similarity and dN/dS ratio

In this section we evaluate the Δvalue of betweenness, closeness and degree. Every chart has two representations, the percentage of similarity or dN/dS ratio (curve) and the centrality Δ (dots). Along the Y-axis (left) we have the scale for the percentage of similarity or dN/dS ratio and on the Y-axis (right) the scale for the centralities. On the X-axis we have the set of orthologs. All the charts are sorted by the percentage of similarity or dN/dS ratio.

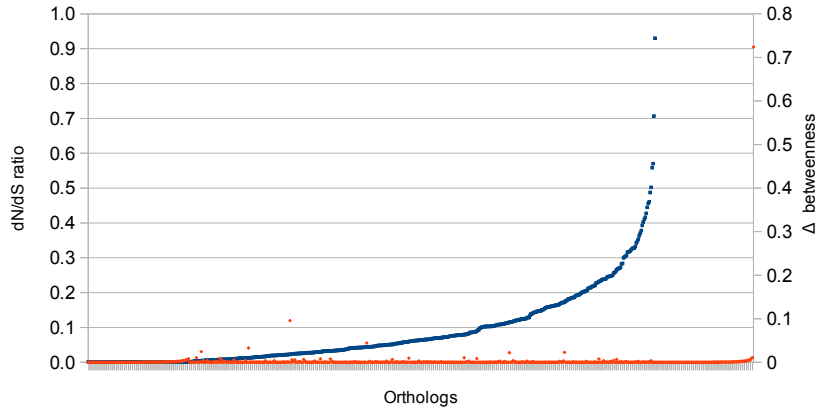
4.3.5.1 Betweenness

Human and mouse orthologs

First, we have the betweenness values for human-mouse related to percentage of similarity and dN/dS ratio. Figure 4.22a shows that high percentage of similarity values do not change the difference between the betweenness values of the proteins in the orthologs (Δ). Figure 4.22b shows that do not matter if the dN/dS ratio is high or low (the protein have changed or not) the difference between the *nbc* is still low.



(a) Δ*nbc* and percentage of similarity.

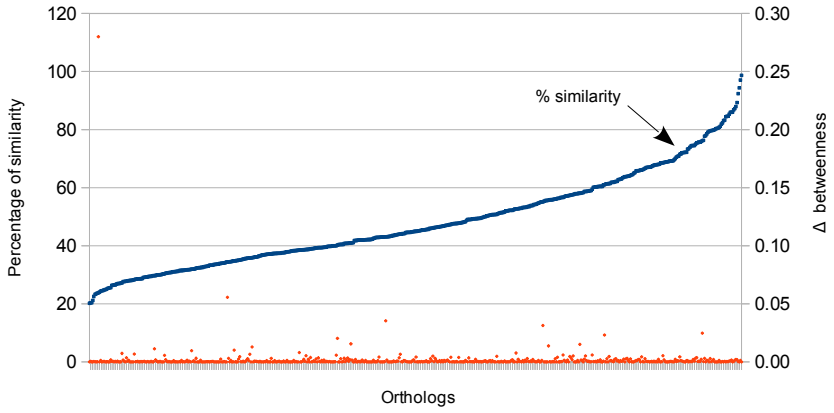


(b) Δ*nbc* and dN/dS ratio.

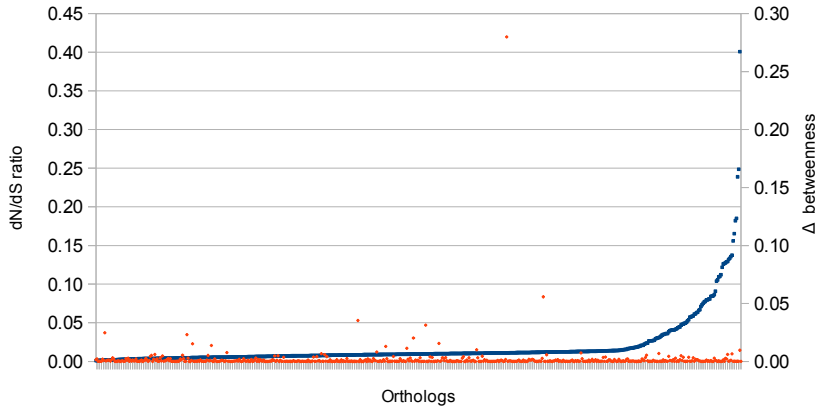
Figure 4.22: Betweenness in human and mouse. (a) Percentage similarity and (b) dN/dS ratio.

Worm and fly orthologs

Next we consider the betweenness values for worm-fly. Figure 4.23a shows that is does not matter whether the percentage of similarity is high or low the Δ*nbc* is still low (proteins with very similar *nbc* values). Figure 4.23b shows that low dN/dS ratio values do not change the Δ*nbc* of the proteins in the orthologs (Δ).



(a) Δnbc and percentage of similarity.



(b) Δnbc and dN/dS ratio.

Figure 4.23: Betweenness in worm and fly. (a) Percentage similarity and (b) dN/dS ratio.

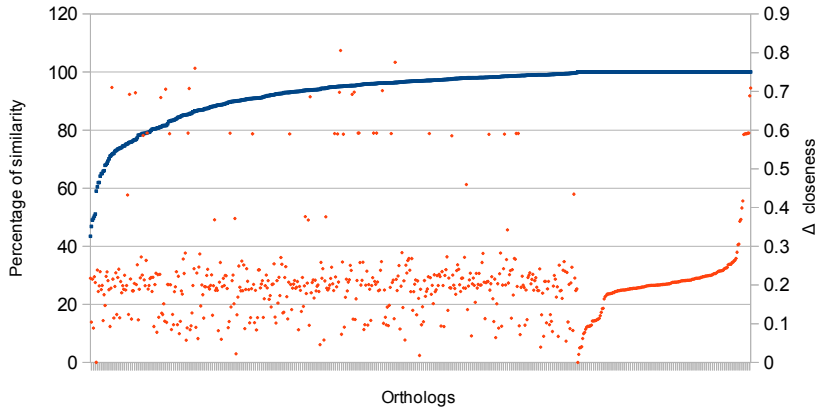
Observing the four species we can conclude that betweenness does not vary depending on the percentage of similarity or the dN/dS ratio. The proteins in the orthologs keep similar betweenness values.

4.3.5.2 Closeness

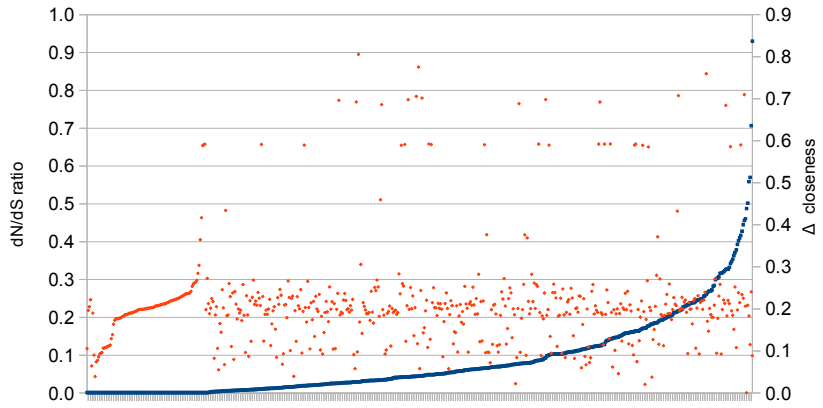
Human and mouse orthologs

Second, we consider the closeness values. Figure 4.24a shows that those orthologs with 100% of similarity have low Δncl values. Closeness values show variations between low and high (outliers), when the percentage of similarity values are below 100%. In this case the closeness values are more disperse than the betweenness values and the Δncl values are higher than the

Δncl values. Figure 4.24b shows that is does not matter whether the dN/dS ratio is high or low: the difference between the Δncl values does not show any variation. In the right side of the plot, we can see that when the similarity is 100% the closeness maintains an increase in the values.



(a) Δncl and percentage of similarity.



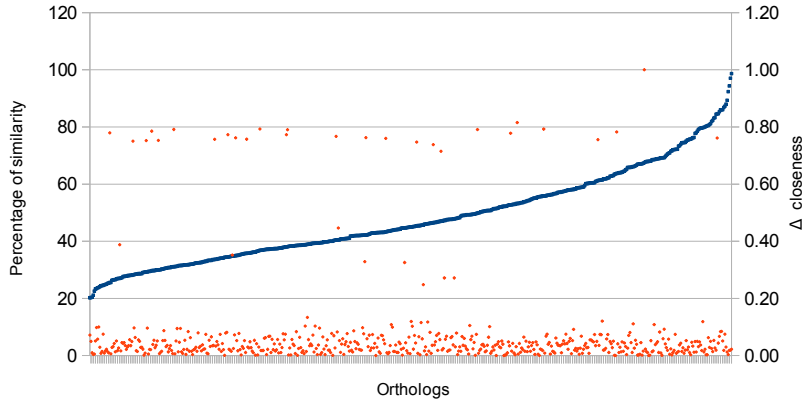
(b) Δncl and dN/dS ratio.

Figure 4.24: Closeness in human and mouse. (a) Percentage similarity and (b) dN/dS ratio.

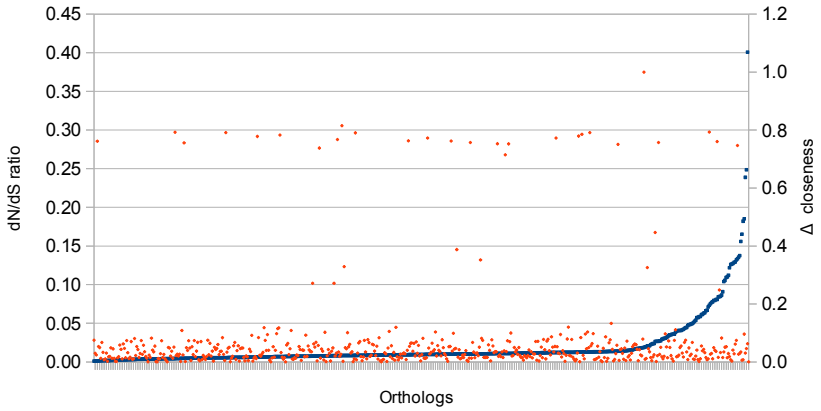
Worm and fly orthologs

Figure 4.25a shows that most of the Δncl values are small and a set of approximate 30 orthologs have with high Δncl values. These do not have a correlation with the percentage of similarity values. Figure 4.25b shows the same behaviour as the previous chart. There is no visible relation between the Δncl values and dN/dS ratio values. In the left side of the plot, we

can see that when the dN/dS ratio values are close to zero the closeness maintains an increase in the values.



(a) Δncl and percentage of similarity.



(b) Δncl and dN/dS ratio.

Figure 4.25: Closeness in worm and fly. (a) Percentage similarity and (b) dN/dS ratio.

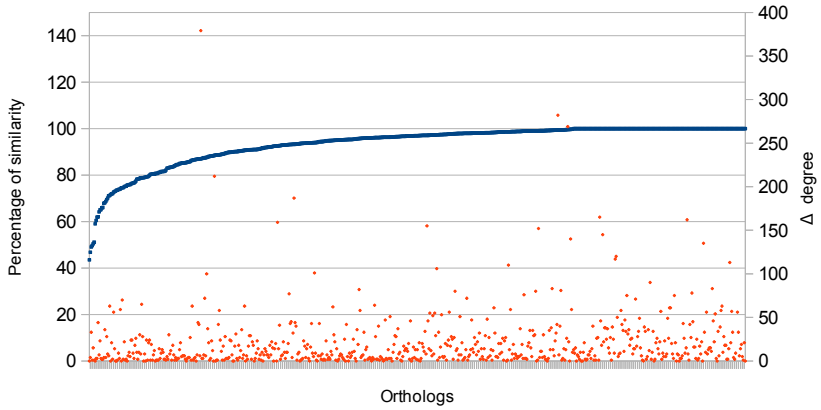
After studying the four species we can conclude that closeness does not vary depending on the percentage of similarity or the dN/dS ratio. The closeness of the proteins in the orthologs have more variations than the betweenness, although small variations.

4.3.5.3 Degree

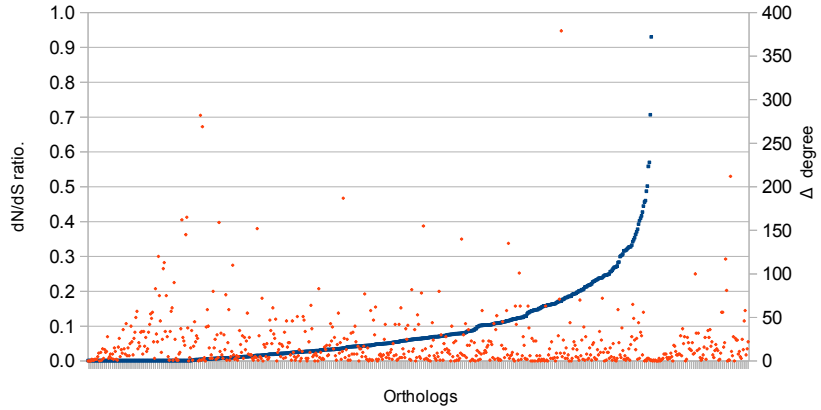
Human and mouse orthologs

Third, we have the degree values. Figure 4.26a shows that the Δdg values are dispersed. This means we have low Δdg values with similarity values between 60% and 100%. Further, we

have high Δdg values (outliers) with percentage of similarity close to 80% and 100%. Figure 4.26b shows the same behaviour as the previous chart. There is no visible relation between the Δdg values and dN/dS ratio values.



(a) Δdg and percentage of similarity.

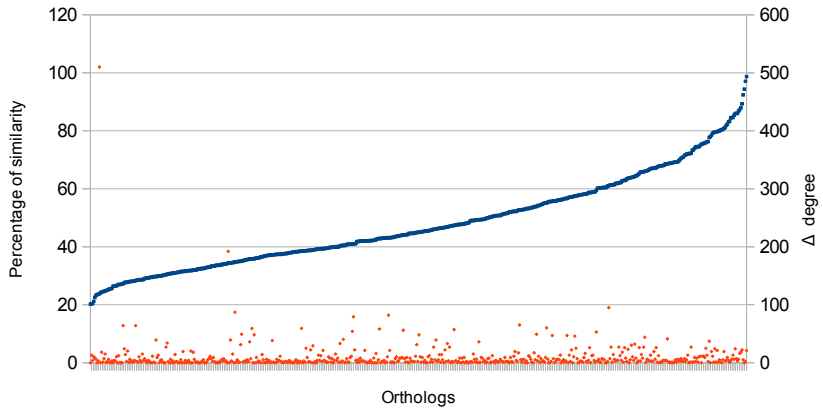


(b) Δdg and dN/dS ratio.

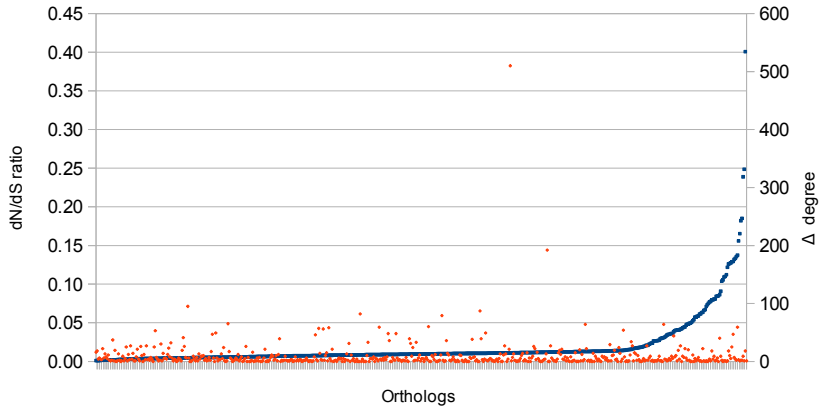
Figure 4.26: Degree in human and mouse. (a) Percentage similarity and (b) dN/dS ratio.

4.3.5.2 Worm and fly orthologs

Figure 4.27a shows that it does not matter whether the percentage of similarity is high or low: the Δnbc is still low, with some outliers. Figure 4.27b shows that low dN/dS ratio values do not change the difference between the Δdg of the proteins in the orthologs.



(a) Δdg and percentage of similarity.



(b) Δdg and dN/dS ratio.

Figure 4.27: Degree in worm and fly. (a) Percentage similarity and (b) dN/dS ratio.

We can conclude that protein sequence variation is not correlated with network parameters. This is confirmed in the correlation section. We observe that very similar proteins (high percentage of similarity) may have similar or different network values. Conversely, different proteins (orthologs) may have different or very similar network parameters

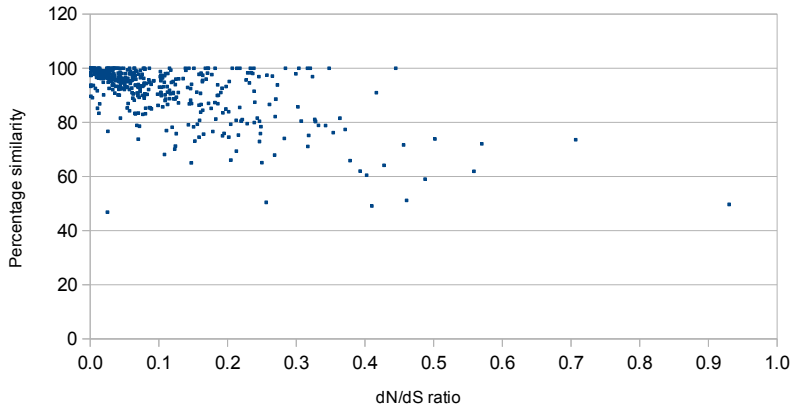
This is interesting because we predicted that protein sequence conservation with function conservation protein sequences (orthologs) are much more variable than orthologous network values.

Among network parameters (measures used): few orthologs have different degrees (even though they do have different sequences); and few orthologs have different closeness values. But an interesting set of around 25 orthologs has very different closeness values (even though the sequence similarity is not correlated with closeness).

4.3.6 Percentage of similarity and dN/dS ratio

The next two charts present the relation or not between dN/dS ratio and percentage of similarity.

In the case of human and mouse in Figure 4.28 the dN/dS ratio values are more concentrated because around the 80% of the pairs have a percentage of identity over 75%. Moreover, there are dN/dS ratio values close to one and many other close to zero. So, we can say that between human and mouse we have a high percentage of conserved protein sequences and a few pairs that are close to a neutral evolution.

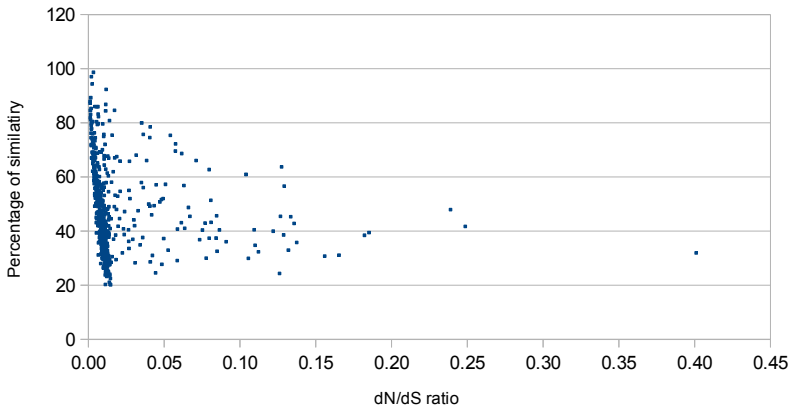


(a) Human-mouse.

Figure 4.28: Percentage of similarity and dN/dS ratio.

There are 815 of 947 orthologs with less than 0.2 dN/dS ratio, this is a 86.1% over the total of orthologs. Also, 916 of 947 orthologs with less than 0.4 dN/dS ratio, this is a 96.7% over the total of orthologs. Last, 4 of 947 orthologs around one dN/dS ratio, this is a 0.4% over the total of orthologs.

For worm and fly in Figure 4.29 the dN/dS ratio values are spread because the differences in the percentage of similarity of each orthologs. However, all of them present values that are below 0.4 dN/dS ratio, this shows conserve protein sequences. Also, the majority of the values (531 orthologs) are close to zero, that is the 92% of the total number of orthologs. The values are spread between 20% and 100% percentage of similarity. The unexpected low dN/dS ratio values showed in Figure 4.21 a) made this plot has an unusual behaviour. Where it is not relevant if the percentage similarity is high or low, the dN/dS ratio values are always low.



(a) Worm-fly.

Figure 4.29: Percentage of similarity and dN/dS ratio.

4.3.7 Correlations

In this section we review the correlation between the measures studied above. Also, we did a comparison with the results of article [44] that motivates this research.

Table 4.9 shows the results obtained in Hahn's paper and our results. We compare only worm and fly because these were the species that Hahn used. The values obtained from our networks follow the same behaviour as Hahn's networks. There are some differences in the values and this is likely due to difference in size of proteins and interactions.

	Hahn		TGB	
	Fly	Worm	Fly	Worm
$dg - nbc$	0.94	0.96	0.96	0.94
$dg - ncl$	0.84	0.55	0.87	0.36
$ncl - nbc$	0.78	0.54	0.83	0.31
$dN - nbc$	-0.07	-0.12	-0.104	-0.119
$dN - dg$	-0.06	-0.11	-0.098	-0.095
$dN - ncl$	-0.05	-0.03	-0.068	-0.034

Table 4.9: Comparison between Spearman's results of Hahn's paper and our research.

In Table 4.10 and 4.11 we present our results after using Spearman's correlation on all the measures used. From Table 4.10 we conclude that the measures used are not correlated in their specific networks, and that is transferred to the relation between the networks, too. These

results confirm the results presented in the previous sections. We have the exceptions showed in Table 4.9.

	Fly	Worm	Mouse	Human
$dg - nbc$	0.943	0.878	0.868	0.823
$ncl - nbc$	0.826	0.298	0.174	0.683
$dg - ncl$	0.872	0.360	0.181	0.817
$dN - nbc$	-0.100	-0.133	-0.064	-0.178
$dN - dg$	-0.095	-0.090	-0.036	-0.173
$dN - ncl$	-0.068	-0.036	0.004	-0.149
$dS - nbc$	0.056	0.088	-0.032	-0.066
$dS - dg$	0.072	0.091	-0.025	-0.054
$dS - ncl$	0.061	0.095	0.019	-0.016
$dN/dS - nbc$	-0.064	-0.091	-0.050	-0.157
$dN/dS - dg$	-0.065	-0.081	-0.021	-0.160
$dN/dS - ncl$	-0.037	-0.068	0.008	-0.142

Table 4.10: Spearman's correlation by species.

From Table 4.11 we conclude that none of the centrality values exhibit a direct correlation with the sequence similarity. This means that does not matter the level of similarity the centrality values could be high or low. This happens in both pairs of species.

	Fly-Worm	Mouse-Human
Similarity - dN/dS	-0.509	-0.644
Similarity - Δnbc	0.171	0.094
Similarity - Δnl	-0.056	0.039
Similarity - Δdg	0.148	0.211
dN/dS - Δnbc	-0.056	-0.096
dN/dS - Δnl	0.045	0.011
dN/dS - Δdg	-0.061	-0.150

Table 4.11: Spearman's correlation by pair of species.

According to the Spearman's values we cannot predict centrality values of one species using the percentage of similarity with another one. For example, to predict centrality values of mouse using human network. We have the same situation with the dN/dS ratio, there is no correlation. Between percentage of similarity and the dN/dS ratio exists correlation and this confirms their meaning. The higher the similarity, the lower the changes in the orthologs (low dN/dS ratio) and vice versa.

4.4 Conclusions and future work

In this chapter we initially predicted that protein sequence conservation would be correlated with function conservation.

Our goal was to determine if orthologs are similar with respect to the different traits. This means the values obtained from the networks parameters (centrality measures) would be correlated with percentage of sequence similarity and sequence divergence (dN/dS ratio).

We defined a methodology to gather the different data for the network parameters and the centrality measures. One of the main challenges in this research is the identification of orthologs across such a broad evolutionary time scale.

We studied four PPINs and their further role in evolution, specifically the study of orthologous proteins between human and mouse, and worm and fly. We selected the orthologs for these two pairs of species using percentage of sequence similarity.

Next, we calculated the three centrality measures (betweenness, closeness, and degree) to identify which proteins are the most central in each species and if they have functions or similarities in common in different species. At last, we calculated the dN/dS ratios to see the differences between the proteins that are orthologs.

Our results show that on average the three centrality measures are very similar between the orthologous proteins. The difference of the centrality values is small, this indicates the centrality values are very similar between the orthologs. This happens in both pairs of species worm-fly and human-mouse.

Both cases, fly-worm and human-mouse, present very low values in the amino acid divergence. This means that both were present purifying selection. The number of synonymous changes is higher than non-synonymous changes.

The results indicate that there is no correlation between network parameters (centrality measures) and protein sequence variation (percentage of sequence similarity and amino acid sequence divergence). This means, the role that proteins have in the network is not directly correlated with their sequence variation.

Next, we present some interpretations of our results and ideas for future work:

- Because the amino acid variation values are too low, we think that by calculating the dN/dS ratio of only the segments of the sequence that interact with other proteins we could obtain a more accurate value of the amino acid variation of the protein.

- The proteins (orthologs) that have sequence changes and maintain their role (centrality values) in the network could be because the proteins in the neighborhood had changed too. Due to the changes in the neighboring proteins, this made possible to keep the same structure of the neighbourhood for the orthologs.

- The dismissed data in this chapter because of the integration problem could be an important factor in the results obtained. As we can see in Figure 4.30 (in the case of worm and fly) there is a big amount of data that are dismissed because we do not have the specific names to identify the proteins from different sources. Improving the amount of data could improve the results obtained with respect to the correlation between centrality measures and percentage similarity or dN/dS ratio.

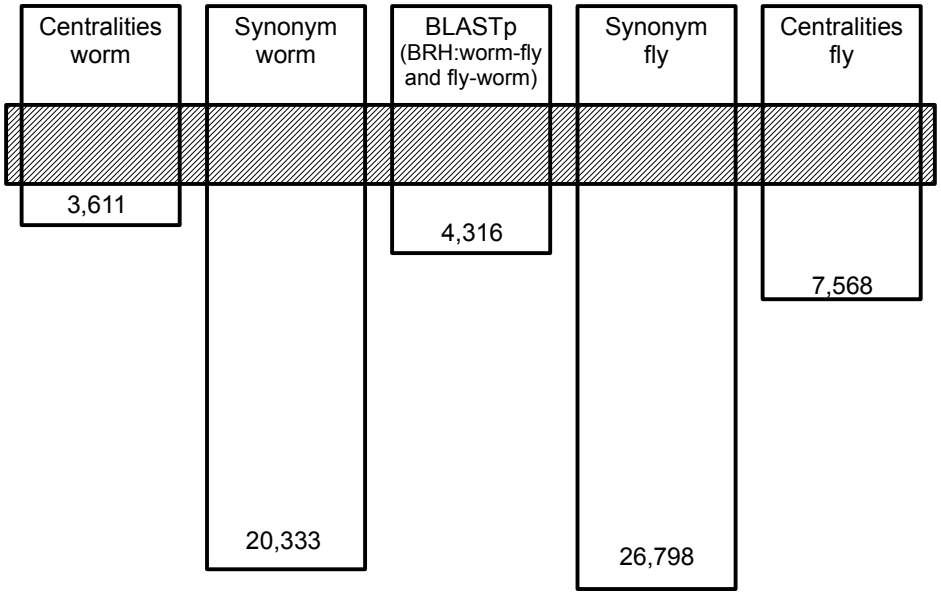


Figure 4.30: Data integration problem.

- There is still missing information in these networks. There are proteins and interactions that have not been studied yet. This could affect the centrality measures obtained.

- The dN/dS ratio values obtained for worm-fly are very low. These lower values will need a further evaluation, mainly because the model that we are using for the calculations underestimate ds and overestimate dn values. So, it is used a maximum-likelihood method to reduce the over or under estimating of substitution rates. We propose to do a simple dN/dS ratio so we compare both results and see what is producing these low dN/dS ratio values.

Chapter 5

Improving feature selection to predict protein-protein interactions

Protein–protein Interactions (PPIs) are known for their important role in diverse biological processes. Characterizing the partners of PPIs is crucial to understand the functional role of individual proteins and the organization of entire biological processes. One of the issues to understand and classify PPIs is to characterize their interfaces (interaction zone, defined in the previous section 2.1) in order to discriminate between types of interactions. For this research we are working with one type of interaction classification defined by Nooren and Thornton [88]. This classification sorts interactions by the duration time of the interaction, in this case transient and permanent interactions (these terms will be explained in detail in the next sections).

The stability of PPIs depends on many features. One of the features is the energetic features of interacting surfaces. This chapter explores the interaction zone between pairs of proteins in order to identify features that can differentiate between types of interactions. In this research we will focus only on energetic features related to the interaction zone. This zone is located where the interaction occurs between the two proteins (see section 2.1.2 for the definition). In the following sections we assume that the interacting proteins form a protein complex.

We identify energetic features that discriminate protein complexes in two classes: transient and permanent. Focusing only on the interaction zone. The classification is made using energetic features from the three-dimensional structure of the complex using the FastContact [19] application. To obtain the energetic features of an interaction we use an application named FastContact [19]. The set of protein complexes used for this research were obtained from Mintseris' study [84]. The energetic features for this study were extracted from this set.

We argue in this chapter that the number of energetic features and their contribution to

the interactions can be a key factor to predict between transient and permanent interactions. Moreover, the number of energetic features used can be adjusted according to the size of the complex. This is because not all the proteins and interaction zones studied have the same size. We evaluate different Machine Learning classifiers to predict these interactions, using a set of 298 complexes obtained from other study [84], in terms of their known three-dimensional structure.

5.1 Literature review

Protein-protein interactions (PPIs) are involved in multiple cellular processes such as the regulation of gene expression, and different processes where the oligomerization is a requirement in order to function. These interactions can be attractive or repulsive, which may result in the formation of intermolecular clusters or aggregates. PPIs depend on the protein surfaces and on the environmental conditions, such as, temperature and pH [2].

Over the past years, there have been studies and research on the functions and interactions of proteins. In these studies the methods (characteristics, geometry, probability) and technologies (computational, biological) used have changed [56,104]. Although there is no established method that allows to know in one procedure a protein in whole and allow the management of protein-protein interactions in a real environment (*in vivo*) [2].

The methods currently used are artificial and are identified as: *in silico* work performed on computer (via computer simulation) and *in vitro* technique to develop in a test tube, or controlled environment outside a living organism [2]. For those reasons, efforts are being made to understand the responsible factors for interactions between proteins at the atomic level (in detail) primarily responsible for these interactions [52, 56, 104, 113].

Efforts have been directed to characterize the geometry (shape) [62], and the physicochemical properties (energy of interaction interface) [20], and the preference of residues to appear on the surface [39], the role of hydrogen bonds, saline bridges and hydrophobic interactions [122]. Others include the loss of surface accessible to the solvent [108] as a result of the interaction and the analysis of the conservation of residues on the interaction surface [75].

The most studied feature is the amino acids composition of protein-protein interfaces (interaction zone). A comprehensive study was conducted by Ofran et al. [90], who introduced a theory-based analysis method to study six types of interfaces (functional). There are two types of internal interactions (intra and inter-domains) and four types of external interactions (homo and hertero obligomers; and homo and hetero complexes). Intra-domain are the interactions between residues in the same domain in the same structural domain in the case of inter-domain is in different structural domains in the same chain. Homo complexes and homo obligomers

are interactions between residues on two different chains that have identical sequence, the difference is that the first are transient and the second are permanent. Hetero complexes and hetero obligomers are interactions between two non-identical chains, complexes are from two different proteins and obligomers are from the same protein respectively. The study [90] concluded that the amino acid composition of these surfaces are different, as there is only 1.5% of similarity between the internal and external surfaces, and 0.2% similarity between hetero surfaces belonging to homo complexes.

Nooren and Thornton [88], performed a classification of different types of interactions, proposing the followings interactions:

- Homo and hetero oligomeric complexes. These two groups are differentiated based on their composition. Homo-oligomers are when the interactions occurs between identical chains; and the complexes are hetero-oligomer when the interactions occurs between non-identical chains.

- Obligate and non-obligate complexes. These two groups are discriminated based on their affinity. Obligate interactions are when the components can not exist independently (unstable on their own *in vivo*); and non-obligate interactions are when the components of the interaction can exist independently (interact alone or paired).

- Transient and permanent complexes. These two groups are differentiated based on the lifetime (or stability) of the complex. Transient interactions associate and dissociate temporarily *in-vivo* (short-term interaction and subject to major mutations) – less stable. Therefore, they have a temporary nature; and permanent interactions are usually more stable and irreversible (longer duration) [40]. The transient are more difficult to discriminate and understand, due to their short life [55].

Permanent interactions (lasting over time) are composed of multiple subunits (identical or different), and usually more stable because there are not many changes in this conformation. Also, transient interactions (temporary) can produce strong and weak links, but of very short period of time, which makes their analysis difficult [55, 88]. Therefore, the proteins involved in the interaction undergo changes that are difficult to reproduce in order to be studied [55, 88, 107]. The stability of the interactions depends on the energetic features of protein surfaces, which can change under different environmental and physiological conditions.

Qi et al. [99], conducted research on protein-protein interactions to determine the accuracy of predictions of interactions using six methods of classification: Random Forest (RF), RF-based k-Nearest-Neighbor, Naive Bayes, decision trees, logistic regression, and Support Vector Machine. These six classifiers were used with different features and then the results

of the performances were compared. Finding these features may have different significance depending on the type of prediction. In this case, the classifier Random Forest was shown to be the most robust and favourable for the three above-mentioned tasks among the six methods for predicting protein-protein interactions.

The classification of a new element using Random Forest is performed by each tree (each tree performs its own classification) on the forest. Then, the forest classification based on the most lending made by subtrees. Random Forest classify a new object analysis it for subjected to analysis by each of the trees in the forest. Each tree provides a classification for the new object. Next, the forest chooses the classification based on majority vote. Random Forest can combine different types of data, do not assume characteristics in the data, and it can manage very well the noise and the missing values.

The protein-protein interaction interface has been widely studied to predict protein interaction sites. Nevertheless, a success rate of 70% correct prediction has been independently achieved by several different groups [15,27,87,129]. A success rate of 77.78% correct prediction has been achieved by Maleki et al. [78–80] with SVM when using desolvation energies of atom type features (one type of energy). Gutiérrez et al. [42] achieved 81% using the forward selection strategy and the Chernoff distance to measure the class separability (ranking of the features), without discriminating by type of energy (all the energies provided by the interaction were used). Also, the best accuracy was obtained using the Loog-Duin Linear method [73].

Several studies point out that to find the characteristics that determine the best way interaction, it is necessary to evaluate the physico-chemical criteria present in most proteins [6]. This means that we must study the characteristics involved in the interaction and in the rest of the protein.

5.2 Method and data

In this section we present our developed methodology. To select the relevant features and their validation we devised a three-phase method, which is depicted in Figure 5.1. The three phases correspond to data retrieval and formatting, followed by the selection of characteristics, and the evaluation of the selected characteristics. Each of these phases are composed of different steps that will be described in the next subsections.

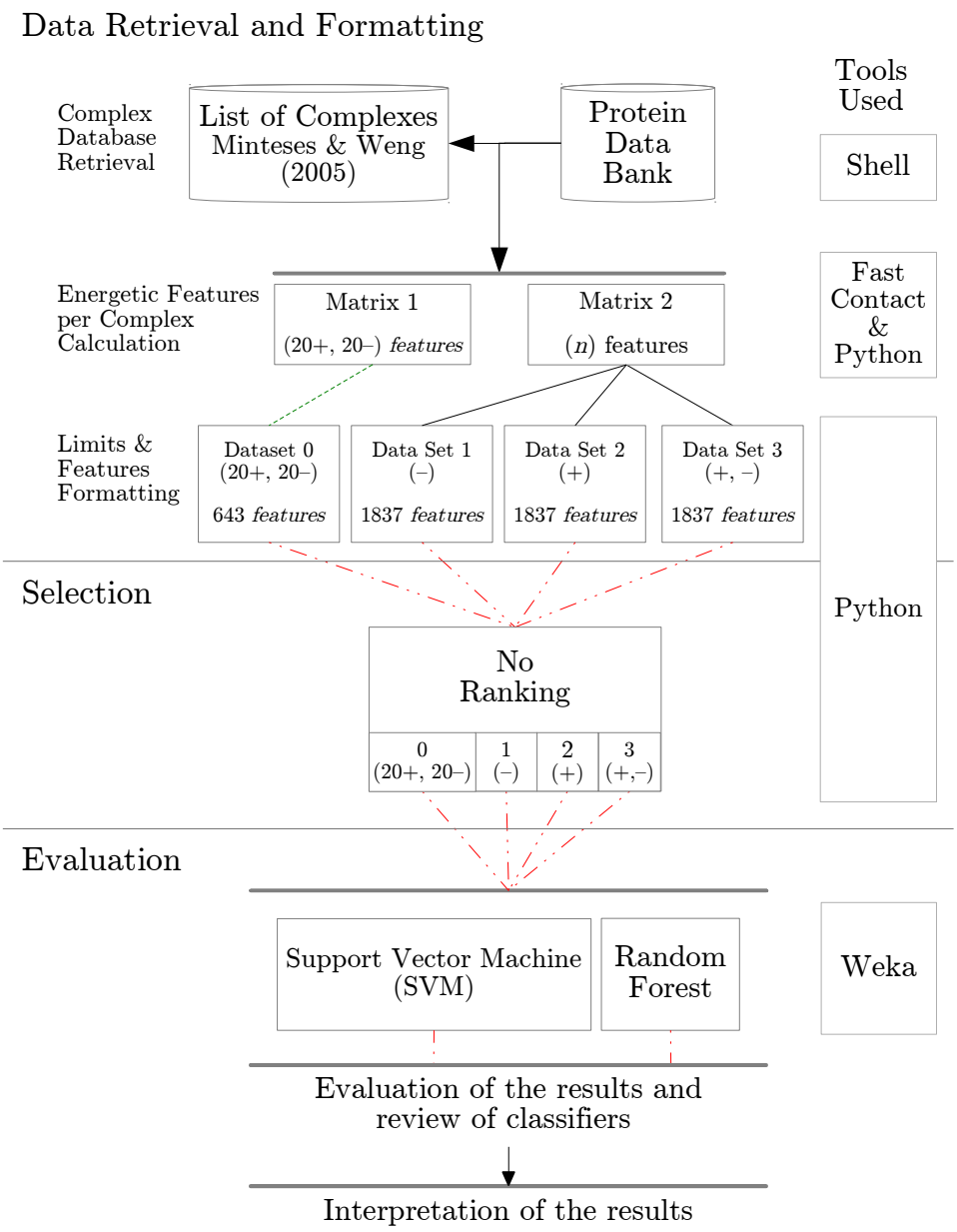


Figure 5.1: Phases for the selection of energetic features and validation of efficiency in the classification.

5.2.1 Data Retrieval and Formatting

5.2.1.1 Complex Database Retrieval

To begin identifying the relevant features and subsequent classification, we created a combined database of the information provided by the work of Mintseris and Weng (list of complexes) [85] and the Protein Data Bank [13] (Figure 5.2).

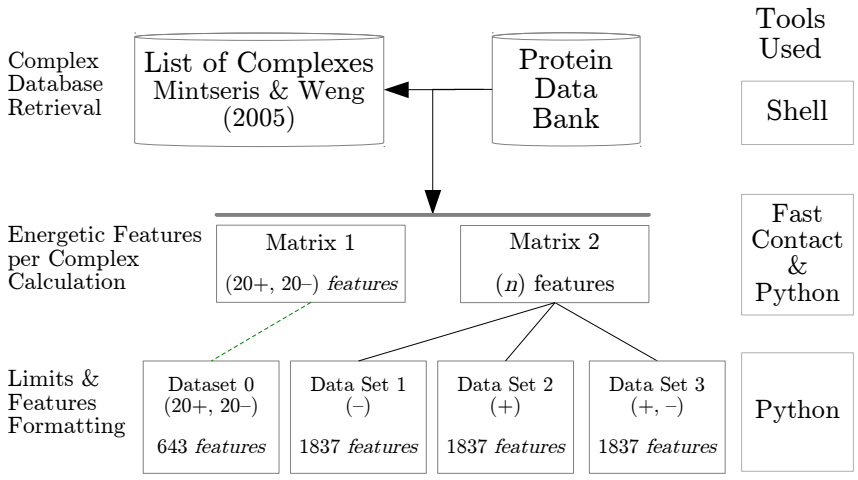
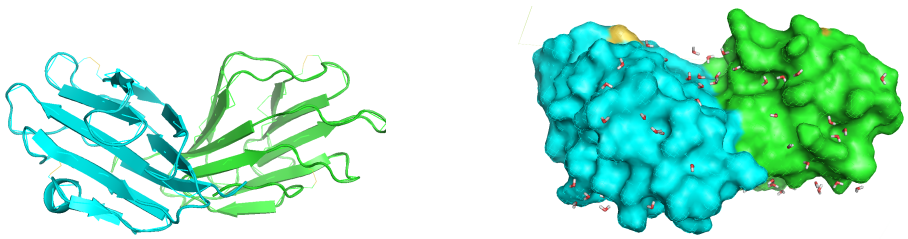


Figure 5.2: Data Retrieval and Formatting Phase.

Complexes

Mintseris and Weng studied the surface of PPIs complexes using 326 complexes. They manually (laboratory work) separated the 326 complexes in 211 transient and 115 permanent complexes [84]. The transient and permanent classification of interactions used in Mintseris' work was defined by Nooren and Thornton [88] on complex interaction of known three dimensional structure.

The list of complexes classified in transient and permanent interactions by [84] are pairs of proteins. This means every complex has two interacting proteins. Figure 5.3 has three different representations of a complex [13]. Figure 5.3a shows the chains and proteins (in different colors) in the complex; and Figure 5.3b shows the surface of the complex, the colors indicate where the proteins are.



(a) Complex chains (b) Surface

Figure 5.3: Complex 1spp.

When proteins participate in a complex, one is called *ligand* and the other one *receptor*. These names are assigned solely to differentiate the proteins from each other. Figure 5.4 has a representation of a complex and its proteins.

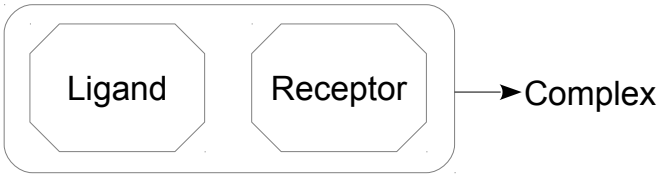


Figure 5.4: Complex, ligand and receptor.

The complexes from Mintseris work [84] are from different species. This means, every complex (or pair of proteins) was classified independently from each other.

An interaction has its own characteristics or features and these depend on: the individual features of each protein and the features of the chains participating directly in the interaction (details in section 2.1). Proteins have more than one chain, but for this research we focus only on those chains that participate actively in the interaction. The names of the chains involved in the interaction were given by [84] in the complexes list. These chains are used to evaluate the interaction of the complex. Figure 5.5 shows an example of chains in complex 1spp [13]. The chains are in different scale of colors.

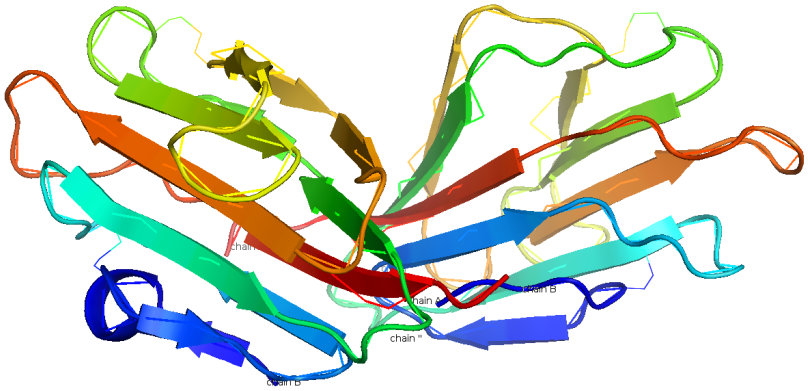


Figure 5.5: Complex chains. Chain are visualized in different colors.

As was mentioned above, to obtain the features it is necessary to know the three–dimensional structure of all proteins participating in a complex.

Three–dimensional structures

The structural information of these complexes is stored in the repository of three-dimensional structural data of large biological molecules named Protein Data Bank (PDB) [13].

All the proteins in PDB were determined experimentally in a laboratory using different experiment methods, such as, ray crystallography (x-RAY) or NMR spectroscopy. These

proteins are contribution from biologist and biochemists from around the world. Currently the repository has around 96,322 proteins structures [14].

Every molecule or protein submitted to the repository require a specific format, defined in [118]. Some of the information required are related to identification, authorship, sequences, atoms and coordinates (see Figure 5.6). The Cartesian coordinates information are the position (x,y,z) of each atom in the protein which forms the atomic structure of a protein (see Figure 5.6b).

```
1  HEADER    COMPLEX (SEMINAL PLASMA PROTEIN/SPP)    19-JUN-97   1SPP
2  TITLE     THE CRYSTAL STRUCTURES OF TWO MEMBERS OF THE SPERMADHESIN
3  TITLE    2 FAMILY REVEAL THE FOLDING OF THE CUB DOMAIN
4  COMPND    MOL_ID: 1;
5  COMPND   2 MOLECULE: MAJOR SEMINAL PLASMA GLYCOPROTEIN PSP-I;
6  COMPND   3 CHAIN: A;
7  COMPND   4 MOL_ID: 2;
8  COMPND   5 MOLECULE: MAJOR SEMINAL PLASMA GLYCOPROTEIN PSP-II;
9  COMPND   6 CHAIN: B
10 SOURCE    MOL_ID: 1;
11 SOURCE   2 ORGANISM_SCIENTIFIC: SUS SCROFA;
12 SOURCE   3 ORGANISM_COMMON: PIG;
13 SOURCE   4 TISSUE: SPERM;
14 SOURCE   5 MOL_ID: 2;
15 SOURCE   6 ORGANISM_SCIENTIFIC: SUS SCROFA;
16 SOURCE   7 ORGANISM_COMMON: PIG;
17 SOURCE   8 TISSUE: SPERM
18 KEYWDS    SEMINAL PLASMA PROTEINS, SPERMADHESINS, CUB DOMAIN
19 KEYWDS   2 ARCHITECTURE, COMPLEX (SEMINAL PLASMA PROTEIN/SPP)
20 EXPDTA    X-RAY DIFFRACTION
21 AUTHOR    A.ROMERO,M.J.ROMAO,P.F.VARELA,I.KOLLN,J.M.DIAS,
22 AUTHOR   2 A.L.CARVALHO,L.SANZ,E.TOPFER-PETERSEN,J.J.CALVETE
23 REVDAT   1   24-JUN-98 1SPP    0
24 JRNL        AUTH   A.ROMERO,M.J.ROMAO,P.F.VARELA,I.KOLLN,J.M.DIAS,
```

(a) Initial description.

```
883 ATOM    532  N   GLU A  59      66.447  37.068  25.358  1.00 15.38           N
884 ATOM    533  CA  GLU A  59      66.966  37.037  24.014  1.00 17.89           C
885 ATOM    534  C   GLU A  59      65.841  37.362  23.043  1.00 19.34           C
886 ATOM    535  O   GLU A  59      64.715  36.911  23.232  1.00 20.80           O
887 ATOM    536  CB  GLU A  59      67.527  35.646  23.745  1.00 18.32           C
888 ATOM    537  CG  GLU A  59      67.858  35.368  22.307  1.00 21.04           C
889 ATOM    538  CD  GLU A  59      68.595  34.060  22.117  1.00 27.09           C
890 ATOM    539  OE1 GLU A  59      68.471  33.158  22.971  1.00 35.35           O
891 ATOM    540  OE2 GLU A  59      69.314  33.929  21.109  1.00 32.08           O
892 ATOM    541  H   GLU A  59      65.653  36.545  25.552  1.00  0.00           H
893 ATOM    542  N   VAL A  60      66.134  38.202  22.054  1.00 20.82           N
894 ATOM    543  CA  VAL A  60      65.161  38.608  21.033  1.00 20.97           C
895 ATOM    544  C   VAL A  60      65.518  37.928  19.703  1.00 21.28           C
896 ATOM    545  O   VAL A  60      66.656  38.032  19.233  1.00 19.50           O
897 ATOM    546  CB  VAL A  60      65.141  40.175  20.825  1.00 22.62           C
898 ATOM    547  CG1 VAL A  60      64.158  40.545  19.712  1.00 15.27           C
899 ATOM    548  CG2 VAL A  60      64.756  40.914  22.134  1.00 14.60           C
900 ATOM    549  H   VAL A  60      67.049  38.548  22.009  1.00  0.00           H
901 ATOM    550  N   LEU A  61      64.562  37.206  19.121  1.00 22.53           N
902 ATOM    551  CA  LEU A  61      64.777  36.513  17.846  1.00 24.02           C
903 ATOM    552  C   LEU A  61      63.967  37.154  16.719  1.00 22.59           C
904 ATOM    553  O   LEU A  61      62.849  37.627  16.914  1.00 21.49           O
905 ATOM    554  CB  LEU A  61      64.410  35.029  17.953  1.00 26.91           C
906 ATOM    555  CG  LEU A  61      65.123  34.167  18.994  1.00 26.13           C
907 ATOM    556  CD1 LEU A  61      64.432  32.828  19.048  1.00 31.14           C
908 ATOM    557  CD2 LEU A  61      66.577  33.989  18.657  1.00 23.01           C
```

(b) Atoms section.

Figure 5.6: Protein data bank format.

The list of complexes consists of the name of the complex and the two chains that participate directly in the interaction. The name of the complex is in a format used by PDB. The chains – representing one from each protein – named chain 1 and chain 2. For example the complex $1EYX\ A:B$ is representing to *R-phycoerythrin from Gracilaria chilensis* [23]. Where chain *A* from protein one (ligand) interact with chain *B* from protein two (receptor).

We retrieve structural information for every complex in our list from PDB. For the preparation of the data, every complex was checked manually to eliminate duplications of residues; we then separated the proteins in two different files (following the requirements for the next step).

The result of this step is a complex database with consolidated data from [85] and [13]. The next step consisted of calculating the energetic features per complex.

5.2.1.2 Energetic Features per Complex Calculation

FastContact, Estimation of Contact and Binding Free Energies

The extraction of the energetic properties of the surface and interface of the interaction zone was performed using FastContact [19], a PPI analysis program. FastContact is a free energy scoring function and estimated the interaction between two proteins [19]. This program was chosen because it produces data such as the contribution of amino acids to the electrostatic energies, desolvation energies and free energies in complexes.

The electrostatic energy (E_{elec}). The electrostatic energies are associated[1] with the particular configuration of a set of point charges in a defined system. It is the force with which the positive and negative charges attract. Electrostatic energies of the same sign (for example two positive) are repulsive while between opposite sign (one positive and one negative) are attractive.

Desolvation energy (G_{des}). Desolvation energy are the amount of energy needed to move the water molecules interacting with the protein or what is the same, the force with which the protein retains water molecules.

Binding free energy (G_{bind}). Free energy is the internal energy of a system minus the amount of energy that cannot be used to perform work. When a protein interacts with another: there is a balance between the complex and the dissociated state and the complex. The difference in binding free energy is the energy required to dissociate the complex. If the complex formation is favored then the ΔG_{des} is negative; If dissociation of the complex

[1] Rupture of a molecule into simpler molecules or atoms.

is favored ΔG_{des} is positive. Ie., is the amount of energy required to dissociate the complex.

Estimation of the binding free energy, electrostatics energies, and residue contact free energies shows the active parts of the protein in the interaction [19].

The interaction between two protein is estimated as

$$\Delta G_{bind} = \Delta E_{elec} + \Delta G_{des},$$

where ΔE_{elec} is the total electrostatic energy [95]. ΔG_{des} is the total desolvation energy. And last, ΔG_{des} is calculated by an empirical contact potential of the form $\Delta G_{des} = g(r)\sum\sum e_{ij}$, where e_{ij} denotes the atomic contact potential between atom i of the receptor and j of the ligand. $g(r)$ is a smooth function varying between two limits defined by [19]

Set of features

For each complex, FastContact provides five different types of energies distributed in proteins, complex and interaction. Table 5.1 shows the summary of the energies availables. Also, we have the type of values in the column *Values*. The *E*-value is the energy, the *R* value is the residue and *AA* value is the amino acid where the energy is coming from. Column *Values used* indicates that we use only the numeric values for our research.

Figure 5.7 shows where the energies from Table 5.1 are located.

Type of contribution	Values	Values used
(1) Residues contributing to the binding free energy	E, R, AA	E, R
(2) Ligand residues contributing to the desolvation free energy	E, R, AA	E, R
(3) Ligand residues contributing to the electrostatics energy	E, R, AA	E, R
(4) Receptor residues contributing to the desolvation free energy	E, R, AA	E, R
(5) Receptor residues contributing to the electrostatics energy	E, R, AA	E, R
(6) Receptor–ligand residue electrostatic contacts	E, R, AA, R, AA	E, R, R
(7) Receptor–ligand residue free energy contacts	E, R, AA, R, AA	E, R, R

Table 5.1: Types and contributions calculated by FastContact for each complex. Energy (*E*) and Residue (*R*).

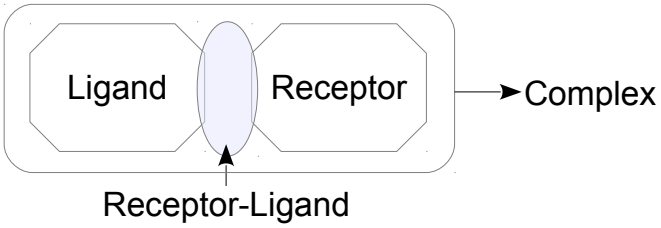


Figure 5.7: Location of the energies for the complex, ligand, receptor, and ligand-receptor.

The FastContact output data is shown in Figure 5.8. Every column represents one of the seven energies presented in Table 5.1. From columns 1 to 5 are two values (*R, AA*) associated with each energy (*E*) and in columns 6 and 7 are four values associated with each energy (see table Table 5.1 rows 6 and 7, values=R, AA, R, AA). The last two columns have more values because the energies are generated by the interaction of the ligand and receptor. So, the first (*R, AA*) represents to ligand values and the second (*R, AA*) represents to receptor values for the energy *E*.

FastContact provides the top 20 positive and top 20 negative residues contributing to the different energies. Figure 5.9 shows a visual example of the different residues associated to an amino acid in a protein (1spp). The data obtained need to be organized to create the data sets to do the tests.

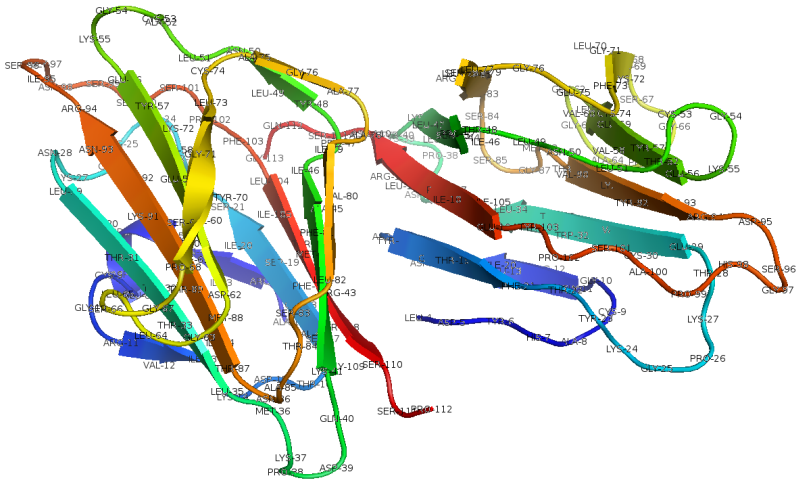


Figure 5.9: 1spp complex. The residues and amino acid are labeled in the chains.

There are two values that were not mentioned, desolvation free energy and electrostatic (4r) energy. Both explained in the FasContact section. This values are obtained from the complex, no from individual parts. Table 5.2 shows an example of the energies obtained from this two values.

(1)Binding free energy			(2)Ligand desolvation free energy			(3)ligand electrostatics energy			(4)Receptor desolvation free energy			(5)Receptor electrostatics energy			(6)Receptor-ligand electrostatic contacts						(7)Receptor-ligand free energy contacts					
-5.396	43	ARG	-1.465	47	PRO	-7.263	43	ARG	-3.034	14	TYR	-3.678	100	GLU	-3.630	100	GLU	43	ARG		-3.338	100	GLU	43	ARG	
-2.456	106	TYR	-1.300	106	TYR	-1.156	106	TYR	-2.807	104	LEU	-2.048	109	GLY	-2.836	2	ASP	43	ARG		-2.248	2	ASP	43	ARG	
-1.796	47	PRO	-1.252	80	VAL	-1.125	102	PRO	-1.026	40	LEU	-1.572	2	ASP	-1.571	109	GLY	2	ARG		-1.369	109	GLY	2	ARG	
-1.441	80	VAL	-1.060	78	ALA	-1.118	2	ARG	-0.813	102	ILE	-1.103	108	GLN	-1.293	105	ARG	106	TYR		-1.151	105	ARG	106	TYR	
-1.254	102	PRO	-0.846	45	ALA	-0.703	41	LYS	-0.767	107	SER	-0.944	16	THR	-0.996	108	GLN	102	PRO		-1.005	108	GLN	102	PRO	
-1.239	108	TYR	-0.648	108	TYR	-0.590	108	TYR	-0.638	77	LEU	-0.667	13	ASP	-0.660	16	THR	108	TYR		-0.907	14	TYR	106	TYR	
-1.236	78	ALA	-0.606	77	ALA	-0.413	17	SER	-0.302	1	LEU	-0.424	12	ASP	-0.572	1	LEU	108	TYR		-0.796	104	LEU	47	PRO	
-0.885	2	ARG	-0.576	104	LEU	-0.335	21	SER	-0.291	108	GLN	-0.409	1	LEU	-0.404	109	GLY	21	SER		-0.743	16	THR	108	TYR	
-0.739	103	PHE	-0.509	103	PHE	-0.330	47	PRO	-0.085	15	GLY	-0.277	106	ASP	-0.399	13	ASP	2	ARG		-0.739	14	TYR	108	TYR	
-0.736	45	ALA	-0.318	46	ILE	-0.255	22	ASN	-0.013	41	VAL	-0.227	34	ASP	-0.387	14	TYR	43	ARG		-0.737	1	LEU	108	TYR	
-0.703	41	LYS	-0.266	44	MET	-0.243	1	ALA	-0.010	42	SER	-0.179	42	SER	-0.385	100	GLU	41	LYS		-0.658	102	ILE	80	VAL	
-0.656	77	ALA	-0.246	107	PHE	-0.230	103	PHE	-0.001	17	ILE	-0.161	21	LYS	-0.336	2	ASP	41	LYS		-0.592	104	LEU	106	TYR	
-0.547	104	LEU	-0.129	102	PRO	-0.218	110	SER	-0.001	103	PHE	-0.151	107	SER	-0.328	21	LYS	112	PRO		-0.518	14	TYR	80	VAL	
-0.470	46	ILE	-0.096	79	ILE	-0.208	37	LYS	0.000	27	CYS	-0.103	14	TYR	-0.301	16	THR	43	ARG		-0.445	108	GLN	104	LEU	
-0.459	17	SER	-0.046	17	SER	-0.190	80	VAL	0.000	26	GLU	-0.093	17	ILE	-0.266	13	ASP	1	ALA		-0.425	108	GLN	103	PHE	
-0.287	22	ASN	-0.039	16	THR	-0.182	25	ARG	0.000	25	THR	-0.073	101	ILE	-0.229	12	ASP	17	SER		-0.399	13	ASP	2	ARG	
-0.243	1	ALA	-0.031	22	ASN	-0.176	78	ALA	0.000	24	LYS	-0.068	3	TYR	-0.208	12	ASP	43	ARG		-0.396	104	LEU	45	ALA	
-0.208	37	LYS	-0.030	23	THR	-0.152	46	ILE	0.000	23	PRO	-0.055	45	THR	-0.197	13	ASP	19	SER		-0.385	100	GLU	41	LYS	
-0.183	107	PHE	-0.008	81	PHE	-0.053	5	GLY	0.000	22	GLY	-0.053	103	PHE	-0.197	34	ASP	2	ARG		-0.368	107	SER	104	LEU	
-0.182	25	ARG	-0.001	105	ILE	-0.051	42	VAL	0.000	21	LYS	-0.042	44	PRO	-0.195	79	ARG	47	PRO		-0.359	14	TYR	45	ALA	
0.066	85	ALA	0.000	13	ILE	0.063	107	PHE	0.000	11	THR	0.004	46	LEU	0.090	109	GLY	7	ASP		0.142	79	ARG	48	TYR	
0.069	4	ASN	0.000	14	LYS	0.066	85	ALA	0.000	19	THR	0.004	97	SER	0.102	79	ARG	48	TYR		0.148	109	GLY	106	TYR	
0.078	16	THR	0.000	15	ASP	0.066	20	ILE	0.000	20	TYR	0.005	31	LEU	0.112	13	ASP	15	ASP		0.152	16	THR	43	ARG	
0.085	7	ASP	0.000	24	ASP	0.069	4	ASN	0.000	21	LYS	0.006	47	ASN	0.113	109	GLY	3	ILE		0.200	36	LYS	2	ARG	
0.099	3	ILE	0.000	25	ARG	0.073	48	TYR	0.000	22	GLY	0.006	10	LEU	0.118	4	HIS	112	PRO		0.227	105	ARG	1	ALA	
0.130	49	LEU	0.001	42	VAL	0.077	49	LEU	0.000	23	PRO	0.008	11	THR	0.120	21	LYS	43	ARG		0.230	105	ARG	104	LEU	
0.136	62	ASP	0.003	51	LEU	0.085	7	ASP	0.000	24	LYS	0.012	15	GLY	0.122	79	ARG	77	ALA		0.232	106	ASP	47	PRO	
0.161	111	SER	0.031	112	PRO	0.099	3	ILE	0.000	3	TYR	0.017	41	VAL	0.134	109	GLY	106	TYR		0.245	2	ASP	39	ASP	
0.166	18	GLY	0.039	18	GLY	0.110	45	ALA	0.001	37	TYR	0.017	43	ILE	0.155	14	TYR	44	MET		0.316	108	GLN	50	ASN	
0.168	83	SER	0.043	101	SER	0.117	16	THR	0.013	4	HIS	0.017	18	PHE	0.177	13	ASP	106	TYR		0.339	13	ASP	106	TYR	
0.178	81	PHE	0.053	49	LEU	0.127	18	GLY	0.070	79	ARG	0.019	33	VAL	0.200	36	LYS	2	ARG		0.349	109	GLY	24	ASP	
0.200	19	SER	0.146	109	GLY	0.136	62	ASP	0.142	12	ASP	0.020	37	TYR	0.215	2	ASP	109	GLY		0.363	2	ASP	110	SER	
0.208	15	ASP	0.193	111	SER	0.168	83	SER	0.218	105	ARG	0.026	39	LEU	0.227	105	ARG	1	ALA		0.365	1	LEU	43	ARG	
0.258	39	ASP	0.200	19	SER	0.187	81	PHE	0.331	13	ASP	0.032	98	PRO	0.245	2	ASP	39	ASP		0.386	106	ASP	48	TYR	
0.298	50	ASN	0.233	2	ARG	0.208	15	ASP	0.355	100	GLU	0.036	105	ARG	0.349	109	GLY	24	ASP		0.399	2	ASP	109	GLY	
0.300	110	SER	0.273	21	SER	0.208	44	MET	0.358	109	GLY	0.058	102	ILE	0.355	1	LEU	43	ARG		0.445	105	ARG	19	SER	
0.364	109	GLY	0.292	50	ASN	0.218	109	GLY	0.364	16	THR	0.096	79	ARG	0.364	105	ARG	19	SER		0.515	18	PHE	43	ARG	
0.379	24	ASP	0.400	48	TYR	0.258	39	ASP	0.441	106	ASP	0.112	4	HIS	0.637	2	ASP	108	TYR		0.724	2	ASP	112	PRO	
0.473	48	TYR	0.519	110	SER	0.379	24	ASP	0.518	18	PHE	0.134	104	LEU	0.696	2	ASP	112	PRO		0.831	105	ARG	2	ARG	
0.583	112	PRO	1.866	43	ARG	0.551	112	PRO	1.798	2	ASP	0.198	36	LYS	0.800	105	ARG	2	ARG		0.952	2	ASP	108	TYR	

Figure 5.8: Example of FastContact output data.

Desolvation Free Energy:	1.31434864
Electrostatic (4r) Energy:	-53.0412104

Table 5.2: Position, name and energy of the complex.

As was mentioned above each energy has associated one residue and one amino acid, in the case of the contact energies it has associated two residues and two amino acids. An energy may or may not contribute to a protein-protein interaction. The energy value can be negative, zero or positive. When the energy value is negative, the residue of the amino acid contributes to the interaction (attraction between two proteins). The more negative the value, the higher the contribution. When the energy value is zero, the amino acid does not produce either attraction or repulsion between two proteins. And when the energy value is positive, the amino acid contributes less to the interaction (repulsion between two proteins)

Complex	(1)
Energy	(2)
Energy	(3)
Energy	(4)
Residue	(5)
aa	(6)
Energy	(7)
Residue	(8)
aa	(9)
⋮	⋮
Energy	(638)
Residue	(639)
aa	(640)
Residue	(641)
aa	(642)

(a) Organization of values for one complex (X)

Complex	feature 1	feature 2	⋯	feature 642

(b) Vector (X^t)

Table 5.3: Representation of the features to be used: (a) vector X and (b) Vector transposed X^t.

The data obtained from this application served to create a database of energetic features, which was cleaned to work with the most relevant data. The data is represented with a vector of static dimension. X is a vector with n number of variables required, $X = [x_1, x_2, ..., x_n]^t$ Where t is the transposed of X. Each component of the vector represents one feature, this means, a characteristic that it is expected to be significant for the classification (see Table 5.3

for representation of the data). $x_1, x_2, ..., x_n$ are the features $(i = 1, 2, ..., n)$ and C the set of classes used $(c_1, c_2, ..., c_n)$.

These specific features describe a specific complex. In this case we know the classes so this feature can be added to the vector of each complex. The final data set is represented with a matrix of m complex (rows) and n features (columns) and the additional columns for the class. As an example, if we take the output from Figure 5.8, all these data would be in one row of the matrix. The final matrix is represented in Equation (5.1):

$$M = \begin{bmatrix} c_1 & x_{1,1} & x_{1,2} & \dots & x_{1,n} \\ c_2 & x_{2,1} & x_{2,2} & \dots & x_{2,n} \\ \dots & \dots & \dots & \dots & \dots \\ c_c & x_{m,1} & x_{m,2} & \dots & x_{m,n} \end{bmatrix} \tag{5.1}$$

Using this matrix as a references and the most relevant data, two matrices were created containing the necessary features for their selection. From the two matrices, we obtained 4 different data sets to evaluate.

Matrix 1

The values for Matrix 1 were obtained from the output of the standard version of FastContact making a total of 642 features, plus the class of the interaction —permanent or transient for each of the 296 complexes. The matrix has dimensions 296×642 in Equation (5.2).

$$M_1 = \begin{bmatrix} c_1 & x_{1,1} & x_{1,2} & \dots & x_{1,642} \\ c_2 & x_{2,1} & x_{2,2} & \dots & x_{2,642} \\ \dots & \dots & \dots & \dots & \dots \\ c_c & x_{296,1} & x_{296,2} & \dots & x_{296,642} \end{bmatrix} \tag{5.2}$$

The details are shown in Table 5.1. Every type of energies has the 20 Max and 20 Min, so in total we have 40 energies by each type (column "$20+, 20-$ Features: Matrix 1"). And, if we add the two values (E, R) for each energy, we have a total of 80 features by type (column "$20+, 20-$ Features: Total").

Chain	Type of energy	Energies, residues	(20+,20−) Features	
			Matrix 1	Total
Complex	Desolvation free energy			1
	Electrostatic energy			1
	Free energy union	2 (E & R)	40 (20+, 20−)	80
Ligand	Desolvation free energy	2 (E & R)	40 (20+, 20−)	80
	Electrostatics energy	2 (E & R)	40 (20+, 20−)	80
Receptor	Desolvation free energy	2 (E & R)	40 (20+, 20−)	80
	Electrostatics energy	2 (E & R)	40 (20+, 20−)	80
Receptor—Ligand	Electrostatic energy contacts	3 (E & R & R)	40 (20+, 20−)	120
	Free energy contacts	3 (E & R & R)	40 (20+, 20−)	120
Class				1
Grand total of features calculated				643

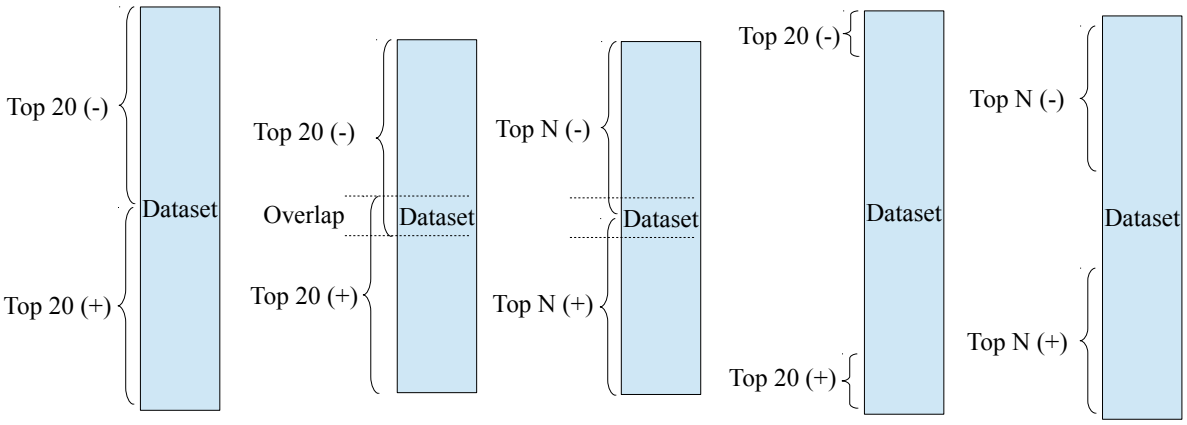
Table 5.4: Matrix 1, types of energies and features calculated by FastContact per each complex.

5.2.1.3 Limits and features formatting

During the creation of Matrix 2, we determined that a complex can have less than 40 energetic values, and the standard version of FastContact provides the top 20 positive and negative values (see Figure 5.10a). As it is shown in Table 5.5, there is at least one complex with 21 energetic values; that is, there would be an overlap of 19 features for that complex (see Figure 5.10b). Instead of using 40 energetic values, it should be used either 21 or add missing values, otherwise, there would be 19 features being double counted (see Figure 5.10c).

The previous example helps to highlight the relevance of our contribution: by having the whole population of features, it is possible to expose the overlap of energetic values. Therefore, one contribution of this work is based on the number of features studied. Additionally, it was also possible to consider the differences between negative, zero and positive energies; and a bigger number of energetic features in every complex to determine their contribution in the protein–protein interaction (see Figures 5.10d and 5.10e). With this information, we were able to compare the results of different data sets, having confidence that there was no overlap in the data (see *Matrix 2* box in Figure 5.2).

There are big complexes in the list, so the number of features is higher than the average, as we can see in Table 5.5 row Receptor–Ligand. For this reason, we evaluated the sizes of the complex and we left out a set of 28 complex from the set of complex to avoid noise in our data.



(a) Case 1. (b) Case 2. (c) Case 3. (d) Case 4. (e) Case 5.

Figure 5.10: Cases.

Chain	Type of energy	Min	Max	Used
Complex	Free energy union	29	1,520	29
Ligand	Desolvation free energy	29	1,520	21
	Electrostatics energy	29	1,520	21
Receptor	Desolvation free energy	21	1,479	21
	Electrostatics energy	21	1,479	21
Receptor—Ligand	Electrostatic energy contacts	2,064	129,999	268
	Free energy contacts	2,064	129,999	268

Table 5.5: Minimum and maximum numbers of energetic and residues values obtained among 298 complexes.

Matrix 2

From this group of 326 complexes, we conducted our study in a sample of 298 complexes, 94 complexes representing permanent protein-protein interactions, and 204 complexes representing the transient protein-protein interactions. As we explain above we left 28 complexes outside the sets becasue are too big to be compare with the rest of the complex.

The second matrix (Column *Matrix 2* in Table 5.1) has dimensions 298×1836, yielding a total of 1836 features for each of the 298 complexes –Equation (5.3).

$$M_2 = \begin{bmatrix} c_1 & y_{1,1} & y_{1,2} & \cdots & y_{1,1,836} \\ c_2 & y_{2,1} & y_{2,2} & \cdots & y_{2,1,836} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ c_c & y_{298,1} & y_{298,2} & \cdots & y_{298,1,836} \end{bmatrix} \tag{5.3}$$

The values for Matrix 2 were obtained from a custom version of FastContact. Instead of the top 20 positive and negative energetic values, FastContact was customized to provide an arbitrary number of such energies. We were motivated to explore the number of features that FastContact could determine per complex, and to utilize such information to obtain better results. As a preliminary result, we obtained an overview of the range of energetic features available, which is summarized in Table 5.6. The criteria of features selection in this study relies on the minimum number energetic values available. We chose the minimum number in order to have the same number of features —of a same type— per complex.

Table 5.6 provides a reference for the types of energy and the features available for Matrix 2. In the case of Complex (column *Chain*), we chose to study the top 29 features (column *Used*), which corresponds to the minimum number of features of free energy union available in every complex. Next, both ligand and receptor (column *Chain*) should be equally represented as these are the proteins interacting with each other. To maintain an even number of features in both ligand and receptor, we chose the minimum value available among all of them, in this case the minimum is between 21 and 29.

Chain	Type of energy	Features	*n* Features Matrix 2	Total
Complex	Desolvation free energy			1
	Electrostatic energy			1
	Free energy union	2 (E & R)	29 (14+, 15−)	58
Ligand	Desolvation free energy	2 (E & R)	21 (10+, 11−)	42
	Electrostatics energy	2 (E & R)	21 (10+, 11−)	42
Receptor	Desolvation free energy	2 (E & R)	21 (10+, 11−)	42
	Electrostatics energy	2 (E & R)	21 (10+, 11−)	42
Receptor—Ligand	Electrostatic energy contacts	3 (E & R & R)	268 (134+, 134−)	804
	Free energy contacts	3 (E & R & R)	268 (134+, 134−)	804
Class				1
Grand total of features calculated				1837

Table 5.6: Matrix 2, types of energies and features calculated by FastContact per each complex.

In all these cases, there are zero–values in those complexes with small number of non-zero values. From our point of view, an attribute with zero values is better than a missing value and there is no overlap because we have already considered different data sets for positive and negative values.

However, when the number of features available was high (as in the case of Receptor–Ligand), we considered the minimum number of non–zero values available that were either positive or negative. In this case, to keep a balance between positive and negative energetic values, we also considered the minimum common number of features available. For example, a complex with 268 negative energy values, 687 positive values and the remaining zeroes, we chose 268 because this would enable us to work with a data set of only negatives values. We applied these set of criteria to extract the data sets explained below.

From Matrix 2, as explained above, we extracted 3 data sets:

1. A list with only negatives energies. Only the energies that contribute (separately) the most to the interaction with a total of 1837 features including the feature class (see in Table 5.6 row Class). This mean, .

2. A list with only positive energies. Only the energies that contribute (separately) the least to the interaction with a total of 1837 features including the class.

3. A list with both top negative and positive energies. The number of energies proportional

to the size of the complex, considering the energies contribute (separately) the most and the least to the interactions, with a total of 1837 features including the class. For data set 3, we split the number of energetic values by half; that is, if the limit were 29 energetic values, then we would select the top 15 of negative values and the top 14 of positive values (with a reduced number of energetic values available, we granted a preference to negative values because they contribute more to an interaction).

5.2.2 Selection

The data sets obtained in the previous step are unranked (not sorted by any type of relevance) (Figure 5.11). Thus, the top energy values —either positive or negative— of every selected feature provides the same predictor value. We are not excluding any feature, except for the amino acids because are text variables and not numbers. It might be possible to rank the features in a data set, however, this study is focused on unranked data as a first step towards a major goal.

An unranked data set enables to identify the best classification method to the data provided by FastContact, only considering the negative, positive or both type of values. In a future work, we plan to explore ranking the features used in this research, and determine which might be the most relevant to discriminate between transient and permanent protein-protein interactions.

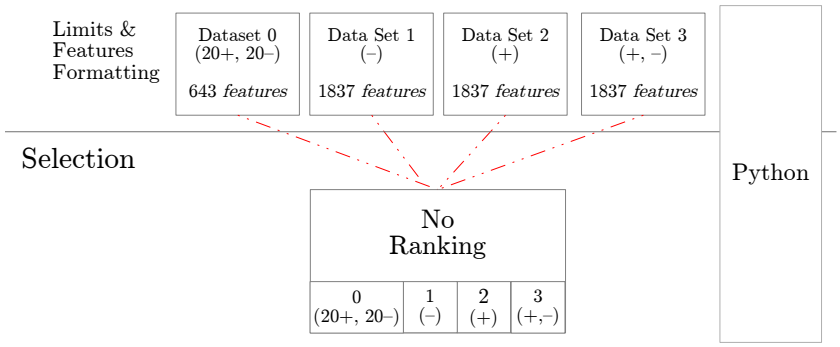


Figure 5.11: Selection Phase.

5.2.3 Evaluation

In this step we applied two classifiers to each of the four data sets: Support Vector Machine (SVM) and Random Forest (Figure 5.12).

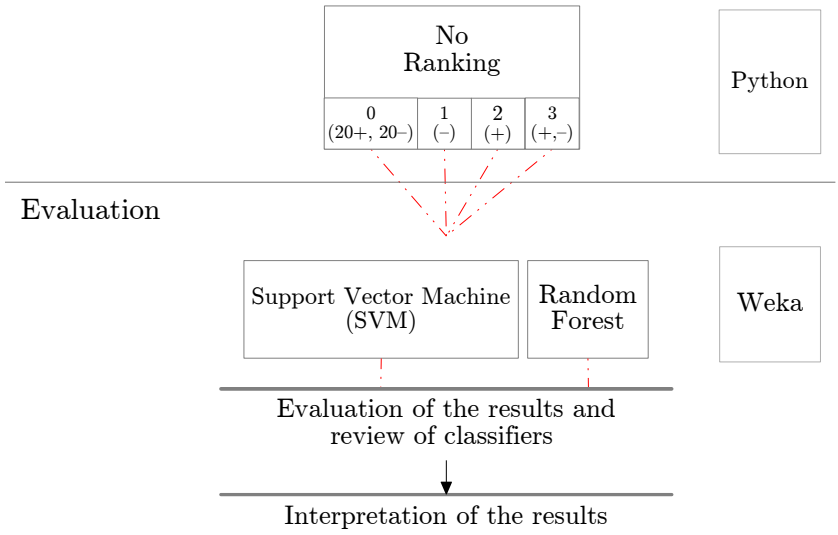


Figure 5.12: Evaluation Phase.

Support Vector Machine is a system to train linear learning machines efficiently for classification. The nonlinear SVM learning is obtained by using the so-called kernel-based functions for transforming the input attribute space in a working space of much higher dimensionality (thereby increasing the computational capacity of the linear learning machines). A hyperplane is needed to separate the two classes, so that the distance between the optimal hyperplane and nearest training pattern is maximum. Thus, we obtained six result outputs per each data set. These classifiers were performed using software Weka [45].

Random Forest is a learning method for classification that works by creating a multitude of decision trees at training time and giving the class by individual trees. For this case the training data is no random divided by us, instead the algorithm for the classification does the division using a random algorithm. In the algorithm it is possible to define different parameters, such as, maximum depth of the trees produced, selection from top p features, and number of trees as output. The Random Forest error rate depends on the correlation between any two trees in the forest. The higher the correlation the higher the Random Forest error rate, this means for all is a weak classifier. A way to reduce the error rate is to increase the strength of the individual trees. In Random Forest, the test options cross-validation or percentage split are estimated internally, during the run.

To obtain the best accuracy, we explored two test options to train and test each data set and to evaluate the classifiers: *q-Folds* Cross Validation and Percentage Split. *q-Folds* Cross validation divides the data set in q disjoint groups of equal size. It starts with the first iteration, where the first group q_1 is used for test and the remaining $q_2, ...q_n$ for training. The second iteration for test group q_2 two and the remaining $q_1, ..., q_n$ for training is taken, and so on, until

each group has been used as a test set (see Figure 5.13). After the q iterations an average is calculated with the q accuracy results. The classifier is trained q times ($\frac{n_i}{q}$, if $n_i \geq q$). In the case of our research the q was changed according to the classifier and kernel used.

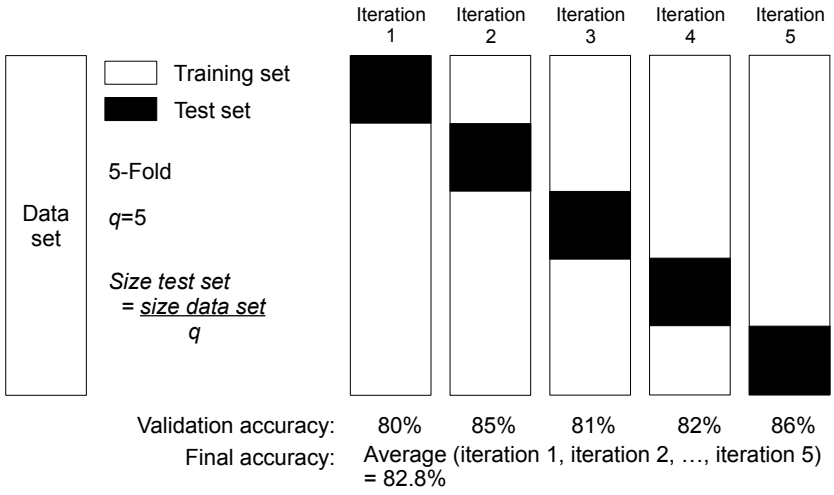


Figure 5.13: Cross validation.

A split percentage denotes the total percentage (100%) divided into two percentages to select the size of the data set that will be used as a training and test set. For example, we have a set of 20 objects divided in two classes (c1 and c2). Next, we select a percentage split of 80% for training and 20% for testing (see Figure 5.14). The 4 objects in the test set are predicted individually obtaining a probability distribution for each instance. Later, those instances with high probability are selected as correctly classified. In the case of the example, we have 3 objects correctly classified over 4 object with a 75% of precision. In addition, we applied multiple folds and percentage split sizes.

Data set size=20
80% Training set
20% Test set

Size test set = 4
Size training set = 16

Instances #:	1	2	3	4
Actual class:	c1	c1	c2	c2
Predicted:	c2	c1	c2	c2
Error:	0.7	0.2	0.1	0.3
Probability distribution:	0.3	0.8	0.9	0.7

Correctly classified instances: 75%

Figure 5.14: Percentage split.

Finally we selected those that provided higher accuracy, which are presented in section 5.3. The purpose is to compare the data set 0 $(20+, 20-)$ (with 643 features) against the other

three data sets (with 1,837 features each). That is, if the use of a larger number of features could improve the prediction between transient and permanent interactions.

5.3 Findings and discussion

In this section we present the results obtained from the application of two classifiers SVM and Random Forest and the methods used to split the data for training and testing data sets. The accuracies obtained from all the classifications are from the test data sets.

As stated in section 5.2.3, we applied multiple sizes of training and test sizes to train and test the data, selecting the ones that provided higher accuracy. These results are summarized in Tables 5.8 and 5.9, which show the maximum accuracies obtained for each data set and each classifier, respectively. Both tables are explained in detail later in this section.

Table 5.7 presents the highest accuracies obtained for each classifier, distributed across the four data sets described in section 5.2.1. The highest values in each row and columns are presented in bold.

Datasets	Support Vector Machine (SVM)					Random
	Linear	Polyn 2	Polym 3	Radial	Sigmoid	Forest
1 $(-)$	77.52%	**86.67%**	81.62%	**83.33%**	71.91%	**83.33%**
2 $(+)$	**86.67%**	80.00%	80.00%	80.00%	74.03%	77.52%
3 $(+,-)$	**86.67%**	77.92%	**81.82%**	**83.33%**	**80.00%**	80.53%
0 $(20+,20-)$	72.97%	67.79%	69.49%	77.52%	74.15%	**83.33%**

Table 5.7: Maximum accuracies of the data sets (including old list/Data set $0(20+,20-)$), according to classifiers.

From the point of view of classifiers,

- the data sets $1(-)$ to $3(+,-)$ present the best accuracy when using SVM;
- the data set $1(-)$ when used with kernel Polynomial 2, and
- the data sets $2(+)$ and $3(+,-)$ when using kernel Linear.

We can conclude that to classify complexes whose only energetic values available are negatives, then it is recommended to perform SVM Polynomial 2. Similarly, to classify complexes with only positive values $(2(+))$, or both negative and positive combined $(3(+,-))$, then it is recommended to perform SVM Linear.

From the point of view of data sets, Table 5.7 shows that data set $3(+,-)$ presents the best accuracies with respect to other data sets. All of them were obtained using the SVM classifier with different kernels: Linear, Polynomial 3, Radial, and Sigmoid (86.60%, 81.80%, 83.30%, and 80.00% respectively). We can conclude that having both positive and negative energetic values in the same data set might lead to better classification; that is, better than classifying them separately. Besides, the data sets with more features (1,837 in $1(-)$, $2(+)$, and $3(+,-)$) perform better than the data set with less features (643 in $0(20+,20-)$). The later only performs well with Random Forest classifier.

Table 5.8 presents the maximum accuracies of the classifiers, according to different sizes of training and testing data set. At the top, a list of multiple sizes used in to split the data into training and testing data sets. This part is subdivided in two: cross validation foldings (the numbers ended on F–Fold, e.g. 10–Fold) and percentage splits (the remaining ones, e.g. 80–20). At the bottom —last two rows— is the summary of the highest values found in the top part, separated by each split method.

In Table 5.8, Percentage Split presents better accuracy than Cross Validation for every classifier. The best ones with splits 80–20 (83.30% performing Random Forest) and 90–10 (86.6% performing SVM with kernels Linear and Polynomial 2). We can conclude that the training set using Percentage Split provides better results than Cross Validation for the data sets studied.

Table 5.9 contains the maximum accuracies obtained for the data sets $1(-)$, $2(+)$, and $3(+,-)$; according to different sizes of training and testing data. As in Table 5.8, the table is split in two. At the top, a list of multiple sizes used to split the data into training and testing data sets, and at the bottom, a summary of the highest values of each data set. The purpose of this table is to compare the three data sets of the same size and different types of energetic values, therefore, the data set $0(20+,20-)$ does not fit in this analysis.

Training	Support Vector Machine (SVM)					Random
Test	Linear	Polyn 2	Polyn 3	Radial	Sigmoid	Forest
02–Fold			69.46%			
05–Fold	74.83%					
08–Fold			77.85%			79.87%
10–Fold	75.83%	75.16%	68.45%	71.48%	68.46%	80.53%
12–Fold		77.47%	74.83%			
14–Fold						79.87%
15–Fold				70.47%		
20–80					69.75%	
38–62						80.54%
50–50						80.53%
66–34				72.28%		
70–30	77.52%		71.91%		71.91%	77.52%
74–26		77.92%			74.03%	
78–22			**81.82%**	75.76%		
80–20		80.00%	81.67%			**83.33%**
81–19	77.19%					
90–10	**86.67%**	**86.67%**	80.00%	**83.33%**	**80.00%**	
Cross-validation	75.83%	77.47%	77.85%	71.48%	68.46%	80.53%
Split	**86.67%**	**86.67%**	**81.82%**	**83.33%**	**80.00%**	**83.33%**

Table 5.8: Maximum accuracies of the classifiers, according to different sizes of training and testing data sets

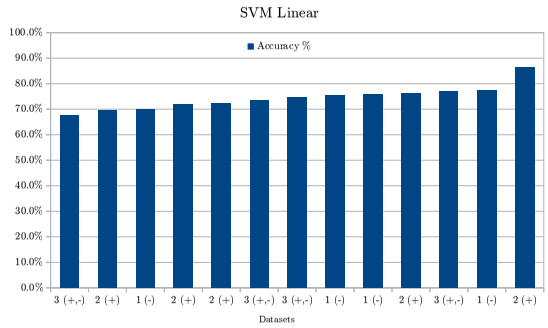
Overall, sets $1(-)$ and $3(+,-)$ have better results than set $2(+)$. Both of them present accuracy equal or higher than 80% in four splits, whereas $2(+)$ only in one. As in Table 5.8, Percentage Split presents better accuracy than Cross Validation for every data set. We can conclude that regardless of the data sets used, the best choice to split the data studied is Percentage Split.

Figure 5.15 presents a different representation of the results obtained in the classifications.

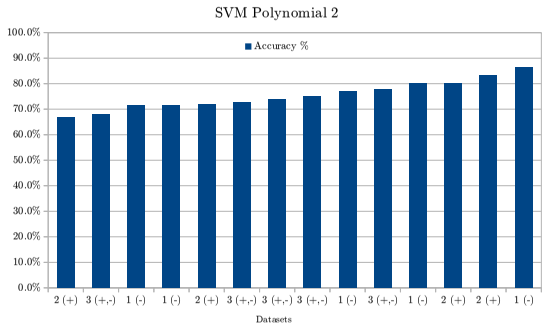
Training		Datasets	
Test	1 $(-)$	2 $(+)$	3 $(+,-)$
02–Fold		69.46%	
05–Fold		73.49%	74.83%
08–Fold	79.87%	76.51%	
10–Fold	80.54%	76.51%	80.54%
12–Fold	77.18%		68.45%
14–Fold			79.86%
15–Fold			70.47%
20–80	69.75%		
38–62			80.54%
50-50	80.53		
63-34	72.28		
70–30	77.53%	77.52%	77.53%
74–26		74.03%	77.92%
78–22	75.76%		81.82%
80–20	83.33%		
81–19			77.19%
90–10	**86.67%**	**86.67%**	**86.67%**
Cross Validation	80.55%	76.50%	80.5%
Split	**86.67%**	**86.67%**	**86.67%**

Table 5.9: Maximum accuracies of data sets, according to different sizes of training and testing data sets

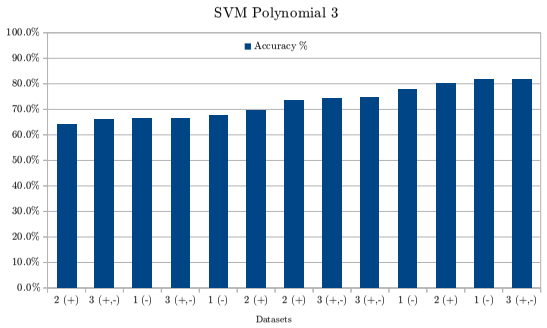
Each chart corresponds to the performance of each classifier: SVM Linear, SVM Polynomial 2, SVM Polynomial 3, SVM Radial, SVM Sigmoid, and Random Forest. We observe that data set $2(+)$ is a good data set when is used with SVM linear kernel (Figure 5.15a), providing an accuracy of 86.6%. The data set $1(-)$ is a good data set when is used with SVM polynomial 2 kernel, SVM Radial, and Random forest obtaining the highest accuracy (Figure 5.15b, 5.15d, and 5.15f, respectively) . Finally, data set $3(+,-)$ is a good data set when is used with SVM Polynomial 3 kernel, SVM Radial, and SVM Sigmoid (Figure 5.15c, 5.15d, and 5.15e, respectively). We can conclude that knowing the composition of a data set it is possible to recommend a classifier that delivers a better classification.



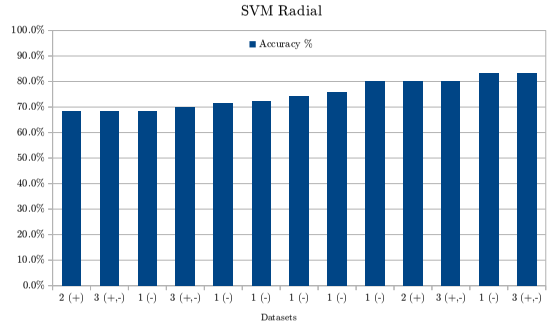
(a) Support Vector Machine - Linear.



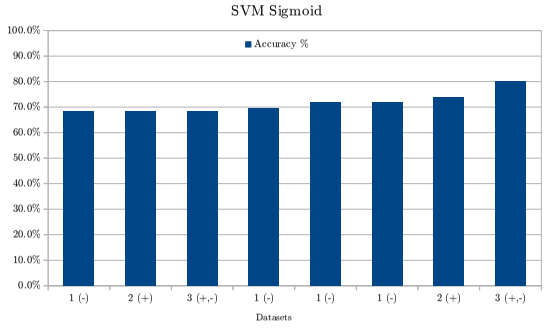
(b) Support Vector Machine - Polynomial 2.



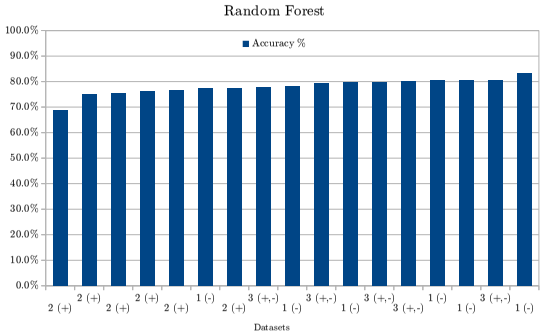
(c) Support Vector Machine - Polynomial 3.



(d) Support Vector Machine - Radial.



(e) Support Vector Machine - Sigmoid.



(f) Random Forest.

Figure 5.15: Classifiers accuracy per data set. (a) SVM Linear, (b) SVM Polynomial 2, (c) SVM Polynomial 3, (d) SVM Radial, (e) SVM Sigmoid, (f) Random Forest. The Y axis corresponds to the percentage of accuracy. The X axis corresponds to the data sets $1(-)$, $2(+)$, and $3(+,-)$; which are repeated because of the use of multiple split sizes of the training and testing data.

5.4 Conclusions and future work

We have proposed an approach to improve the prediction of the protein-protein interaction type, transient or permanent. We have investigated three additional criteria to define a data set, based on a proportional selection of energies according to the total of features delivered by FastContact for each complex. We considered data sets of positive energies, negatives energies, and both altogether. For each data set, we performed Random Forest and SVM with linear, polynomial, radial, and sigmoid kernels. The traditional data sets found in the literature consist of a fixed number of features. In complexes with a small area of interaction, there might be overlap of results if they are not considered carefully. Our approach can help to prevent such situation reducing the minimum number of features for each complex.

Energetic features of protein complex interfaces are able to discriminate between transient and permanent classes of interaction. The use of more features to classify these complexes permits to improve in 5.6% the accuracy of the classification. We obtained a 86.6% of accuracy performing SVM with Linear kernel. This result was obtained using percentage split, 90% of the data set for training (268 complexes) and 10% of the data set for testing (30 complexes). In this case, 86.67% of the 30 complexes were classified correctly. The use of multiple data sets $1(-)$, $2(+)$ and $3(+,-)$ makes possible to identify the energy values that are useful to classify complexes. Not only to determine the type of interaction, but also to identify which type of energy value (positive or negative) contributes better to describe the protein-protein interaction.

The accuracy obtained in this analysis reinforce the idea that energetic features in the interface helps to discriminate interactions between transient and permanent. Although, more work is needed in order to calculate error rates and perform more validations. Nevertheless, the current results, of a work in progress, are encouraging.

From here, we can now further combine algorithms to rank features according to their relevance, such as the forward search proposed in [42] but in this case with more features. This might enable better classification with smaller data sets.

It might be possible to rank the features in a data set. As a future work, we plan to explore ranking the features (see Figure 5.16 in a continuous blue line – *Selection* step) and evaluate the most relevant features to discriminate between transient and permanent protein-protein interactions.

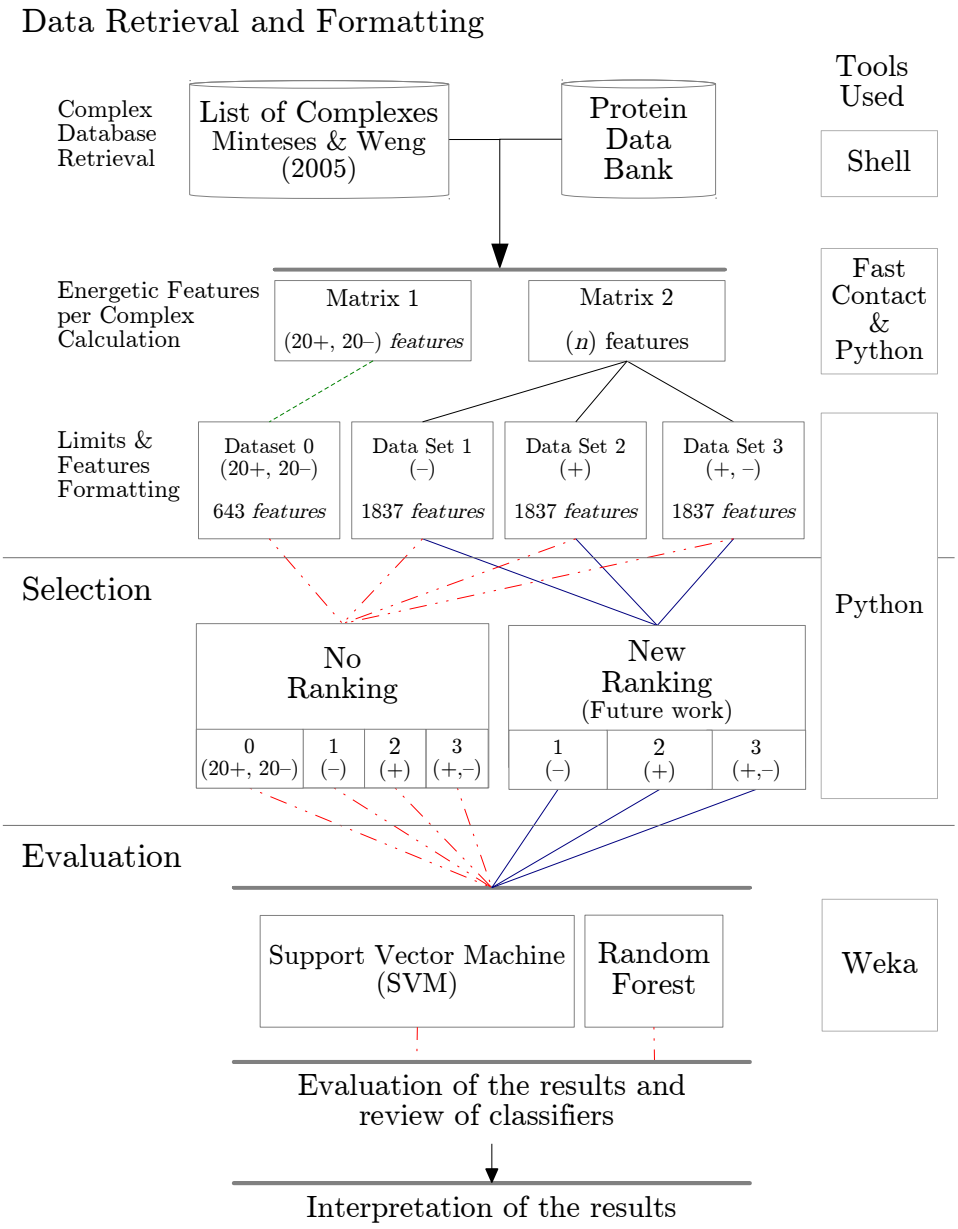


Figure 5.16: Phases for the Selection of energetic features and validation of efficiency in the classification with ranking.

Chapter 6

Conclusions and future work

6.1 Contributions

Our main goal of this thesis was to improve the understanding of protein-protein interaction networks. For this, we proposed three approaches:

The first approach focused on an empirical study on the fine-grained structural differences between protein protein interaction networks and social and complex networks. When identifying best community sizes in social and complex networks, a successful approach is to combine methods finding the conductance with spectral algorithms. We found that these methods work well for PPINS. We identified the best sizes for a group of related proteins for performing their function. We determine that the best community size in terms of community conductance is around size ten, and that this holds across most available protein networks of a reasonable size. The PPINs best community size is 10 times smaller than the best community size for social and complex networks. We these results it is possible to identify differences between these two types of networks in terms of the network community profile and correlations of centrality measures.

The second approach focuses on the study of PPINs and their further role in evolution, specifically the study of orthologous proteins. We are interested in the relationship of PPINs and the evolutionary changes in different species. This will contribute to the understanding of specific PPINs. For this we investigated whether the centrality measures would exhibit correlation with evolutionary changes or conservation. We worked with unweighted networks, which means that every connection or interaction is assumed to have the same value or relevance in the network. The decision to use unweighted networks is due to the fact that there is not enough information about the interactions for all the networks (species) that we are using. We believe that without the use of the orthologous classification it can happen that identified proteins that exhibit similar centralities are in fact not related in sequence and func-

tion. We found that the percentage of sequence similarity varies for very distant species and similar species, in our case worm-fly and human-mouse. The centrality values of orthologous proteins, however, are similar.

Despite conjecturing that there exists a correlation between the centrality values and percentage of sequence similarity and/or dN/dS ratios we could not confirm such a correlation. A possible reason for this could be the use unweighted graphs.

The third approach focuses on addressing this problem. Is it possible to improve the selection of protein characteristics to discriminate protein-protein interactions according to the duration of the interaction? There are energetic features in the surfaces of the interacting proteins that allow the discrimination between permanent and transient protein-protein interactions. The use of more features to classify these complexes permits to improve the accuracy of the classification by 5.6%. For each data set, we performed Random Forest and SVM with linear, polynomial, radial, and sigmoid kernels. We obtained an accuracy of 86.6% when performing SVM with linear kernel. The accuracy obtained in this analysis reinforces the idea that energetic features in the interface help to discriminate interactions between transient and permanent.

The good results in the last approach provide the idea of evaluating the possibility to incorporate the features selected in the networks used in approaches 1 and 2. We believe that our third approach helps to improve the quality of the PPINs by adding more information or value to the interaction of the networks. This information will help to create more reliable weighted PPINs, highlighting the proteins and interactions that are more relevant. Having a network with these values will improve our centrality measures used in approach 2 and also give us more accurate results to investigate possible correlations between centrality measures and percentage of sequence similarity and/or dN/dS ratios. As a consequence, the results of this improved approach would be more meaningful compared to those obtained with unweighted networks, that is improving the quality of the outcomes of approaches one and two. Recall that our long-term goal was to contribute to the understanding of specific functions of proteins to prevent or mitigate the effect of diseases in sick organisms. Our finding, namely best community sizes in PPINs, may contribute to improving the understanding of protein functions because we have a better knowledge of the group (community) of proteins their interact with. The ability to discriminate interactions according to their energetic features will lead to a better understanding of the PPIs in general. This may further contribute to providing essential information of PPIs to prevent or control diseases through an improved understanding of protein-protein interaction networks.

6.2 Future Work

We would like to extend further our experiments associated with social networks (chapter 3). One possibility is to investigate more species than the once we already studied. Another possibility is to investigate biological networks of other types (such as, gene regulatory or metabolic networks), and finally examine a wider range of network measures at fine levels of granularity.

We further suggest focusing on orthologs with only high percentage of sequence similarity and to analyze in more detail the centrality values and the proteins that are participating in these comparisons.

Our methodology in chapter 5 can be refined using ranked features. The ranking should be assigned by the relevance of the feature. Also, the features selected can be subclassified for a better classification with smaller data sets.

We already proposed the idea to create more reliable weighted PPINs using energetic features (see Figure 6.1). We belief that our third approach (study of PPI) will improve the quality of PPINs, as it will add more information or value to interaction in PPIs. Also, we may be able to highlight more relevant proteins and interactions. This information could allow us to create more reliable weighted PPINs. This means that more reliable weighted networks could improve the quality of outcomes of chapters 3 (Community profile network and Centrality measures) and 4 (comapring PPIN from different species),

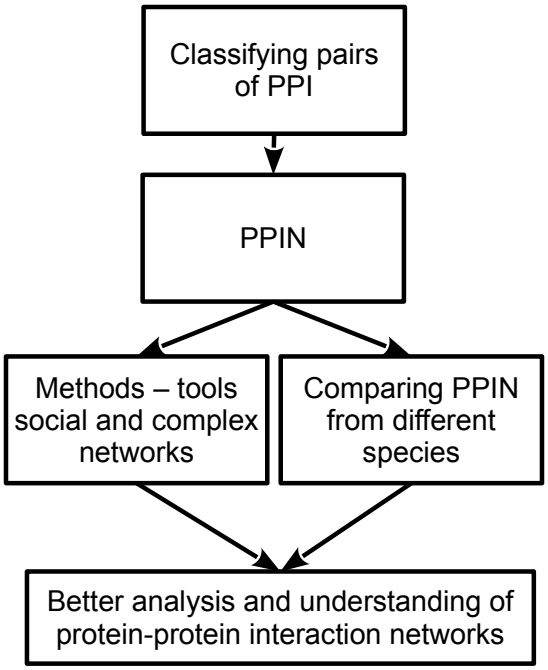


Figure 6.1: Future work.

Bibliography

[1] B. Alberts. The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, 92(3):291–4, 1998.

[2] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*. Garland Science, 4th edition, 2002.

[3] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–402, September 1997.

[4] Reid Andersen, Fan Chung, and Kevin Lang. Local graph partitioning using pagerank vectors. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '06, pages 475–486, Washington, DC, USA, 2006. IEEE Computer Society.

[5] R. Apweiler, T.K. Attwood, Amos Bairoch, Alex Bateman, Ewan Birney, M Biswas, P Bucher, Lorenzo Cerutti, F Corpet, M.D.R. Croning, and Others. InterPro - an integrated docuemntation resource for protein families , domains and functional sites. *Bioinformatics*, 16(12):1145, 2000.

[6] A.I. Archakov, V.M. Govorun, A.V. Dubanov, Y.D. Ivanov, A.V. Veselovsky, P. Lewi, and P. Janssen. Protein-protein interactions as a target for drugs in proteomics. *Proteomics*, 3(4):380–91, 2003.

[7] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

[8] M Barthelemy. Betweenness centrality in large complex networks. *Physical Journal B-Condensed Matter and Complex*, pages 1–6, 2004.

[9] Jacomy M. (2009). Gephi: an open source software for exploring Bastian M., Heymann S., manipulating networks. International AAAI Conference on Weblogs, and Social Media. From AAAI. Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*, 2009.

[10] Alex Bateman, Ewan Birney, Lorenzo Cerutti, Richard Durbin, Laurence Etwiller, Sean R Eddy, Sam Griffiths-Jones, Kevin L Howe, Mhairi Marshall, and Erik L L Sonnhammer. The pfam protein families database. *Nucleic acids research*, 30(1):276–80, January 2002.

[11] D. Benson, D. J. Lipman, and J. Ostell. GenBank. *Nucleic Acids Res*, 21(13):2963–5, 1993.

[12] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. GenBank: update. *Nucleic Acids Res*, 32(Database issue):D23–6, 2004.

[13] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–42, January 2000. (http://www.rcsb.org/pdb/Welcome.do).

[14] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. Protein data bank, statistics, 2014.

[15] A. J. Bordner and R. Abagyan. Statistical analysis and prediction of protein-protein interfaces. *Proteins*, 60(3):353–66, 2005.

[16] Ulrik Brandes. A Faster Algorithm for Betweenness Centrality. *Journal of Mathematical Sociology*, 25(2):163–177, 2001.

[17] Ulrik Brandes and Thomas Erlebach. *Network Analysis: Methodological Foundations (Lecture Notes in Computer Science)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[18] Sylvain Brohée and Jacques Van Helden. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC bioinformatics*, 7(1):488, January 2006.

[19] Carlos J Camacho and Chao Zhang. Fastcontact: rapid estimate of contact and binding free energies. *Bioinformatics (Oxford, England)*, 21(10):2534–2536, May 2005.

[20] P. Chakrabarti and J. Janin. Dissecting protein-protein recognition sites. *Proteins*, 47(3):334–43, 2002.

[21] R. Cohen and S. Havlin. *Complex Networks: Structure, Robustness and Function*. Cambridge University Press, 2010.

[22] R Colak, F Hormozdiari, F Moser, A Schönhuth, J Holman, M Ester, and S C Sahinalp. Dense graphlet statistics of protein interaction and random networks. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 178–89, January 2009.

[23] C Contreras-Martel, J Martinez-Oyanedel, M Bunster, P Legrand, C Piras, X Vernede, and J. C. Fontecilla-Camps. Crystallization and 2.2 a resolution structure of r-phycoerythrin from gracilaria chilensis: a case of perfect hemihedral twinning. *Acta Crystallographica Section D: Biological Crystallography*, 57(1):52–60, 2001.

[24] Dana-farber cancer institute and harvard medical school. Worm Interactome Database, 2011.

[25] Richard Deonier, Simon Tavaré, and Michael Waterman. *Computational Genome Analysis. An introduction*. Springer, 2005.

[26] Reinhard Diestel. *Graph Theory*, volume 173. Springer, 4th edition, 2010.

[27] Caffrey DR, Somaroo S, and Hughes JD. Are protein protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci*, 13(1):190–202, 2004.

[28] Stanley Fields. The two-Hybrid System to detect protein-protein interactions. *Department of Microbiology, state University of New york at Stony Brook.*, 5(2):116–24, 1993.

[29] Paul Flicek, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Stephen Fitzgerald, Laurent Gil, Carlos García Girón, Leo Gordon, Thibaut Hourlier, Sarah Hunt, Nathan Johnson, Thomas Juettemann, Andreas K. Kähäri, Stephen Keenan, Eugene Kulesha, Fergal J. Martin, Thomas Maurel, William M. McLaren, Daniel N. Murphy, Rishi Nag, Bert Overduin, Miguel Pignatelli, Bethan Pritchard, Emily Pritchard, Harpreet S. Riat, Magali Ruffier, Daniel Sheppard, Kieron Taylor, Anja Thormann, Stephen J. Trevanion, Alessandro Vullo, Steven P. Wilder, Mark Wilson, Amonida Zadissa, Bronwen L. Aken, Ewan Birney, Fiona Cunningham, Jennifer Harrow, Javier Herrero, Tim J.P. Hubbard, Rhoda Kinsella, Matthieu Muffato, Anne Parker, Giulietta Spudich, Andy Yates, Daniel R. Zerbino, and Stephen M.J. Searle. Ensembl 2014. *Nucleic Acids Research*, 42:749–755, 2014.

[30] Biological General Repository for Interaction Datasets. Biological general repository for interaction datasets. http://thebiogrid.org/, 2011.

[31] Hunter B. Fraser, Aaron E. Hirsh, Lars M. Steinmetz, Curt Scharfe, and Marcus W. Feldman. Evolutionary Rate in the Protein Interaction Network. *Science*, 296(5568):750–752, 26 April 2002.

[32] Linton C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[33] Stephen P. Freeman, Linton C. Borgatti and Douglas R. White. Centrality in valued graphs: A measure of betweenness based on network flow. *Social Networks*, 13(2):141–154, June 1991.

[34] Toni Gabaldon, Christophe Dessimoz, Julie Huxley-Jones, Albert J Vilella, Erik L L Sonnhammer, and Suzanna Lewis. Joining forces in the quest for orthologs. *Genome biology*, 10(9):403, January 2009.

[35] Allan Gibbons. *Algorithmic Graph Theory*. Cambridge University Press, 26 July 1985.

[36] Anne-Claude Gingras, Ruedi Aebersold, and Brian Raught. Advances in protein complex analysis using mass epectrometry. *The Physiological Society*, 563(1):11–21, 2004.

[37] Giot, Bader, and Brouwer. A protein interaction map of drosophila malanoganster. *Sciences*, 302(1727), December 2003.

[38] M Girvan and M E J Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821–6, June 2002.

[39] A. L. Gnatt, P. Cramer, J. Fu, D. A. Bushnell, and R. D. Kornberg. Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 A resolution. *Science*, 292(5523):1876–82, 2001.

[40] Cher-Sing Goh, Duncan Milburn, and Mark Gertein. Conformational changes associated with protein-protein interactions. *Current opinion in structural biology*, 14:104–109, 2004.

[41] K-I Goh, E Oh, B Kahng, and D Kim. Betweenness centrality correlation in social networks. *Physical Review E*, 67(1):017101, 2003.

[42] Tatiana Gutiérrez-Bunster, Luis Rueda, José Martínez-Oyanedel, and Marta Bunster. Study of energetic features in protein–protein interaction interfaces. Master's thesis, University of Concepción, 2008.

[43] Matthew W Hahn, Gavin C Conant, and Andreas Wagner. Molecular evolution in large genetic networks: does connectivity equal constraint? *Journal of molecular evolution*, 58(2):203–11, 2004.

[44] Matthew W Hahn and Andrew D Kern. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular biology and evolution*, 22(4):803–806, 2005.

[45] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA data mining software. *ACM SIGKDD Explorations*, 11(1):10–18, 2009.

[46] Robert A. Hanneman and Mark Riddle. *Introduction to social network methods*. University of California, 2005.

[47] Desmond J Higham, Marija Rasajski, and Natasa Przulj. Fitting a geometric graph to a protein-protein interaction network. *Bioinformatics (Oxford, England)*, 24(8):1093–9, April 2008.

[48] Hsiang Ho, Tijana Milenković, Vesna Memisević, Jayavani Aruri, and Anand K Ganesan. Protein interaction network topology uncovers melanogenesis regulatory network components within functional genomics datasets. *BMC systems biology*, 4:84, January 2010.

[49] Franz Huber. *Social networks and knowledge spillovers: networked knowledge workers and localised knowledge spillovers*. Peter Lang, 2007.

[50] Nicolas Hulo, Amos Bairoch, Virginie Bulliard, Lorenzo Cerutti, Edouard De Castro, Petra S Langendijk-Genevaux, Marco Pagni, and Christian J a Sigrist. The PROSITE database. *Nucleic acids research*, 34(Database issue):D227–30, January 2006.

[51] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensize two-hybrid analysis to explore the yeast protein interaction interactome. *Proc. Natl. Acad. Sci.*, 98(8):4569–4574, 2001.

[52] J. Janin. *Protein-Protein Recognition.*, chapter Kinetics and thermodynamics of protein-protein interactions from a structural perpective. Oxford University Press, 2000. 344pp.

[53] H. Jeong, S.P. Mason, Albert-László Barabási, and Zoltán N Oltvai. Lethality and centrality in protein networks. *Arxiv preprint cond-mat/0105306*, page 41, 2001.

[54] Jesse D. Bloom and Christoph Adam. Apparent dependence of protein evolutionary rate on number of protein-protein interactions is linked to biases in protein-protein interaction data sets. *BMC Evolutionary Biology*, 3:21, 2003.

[55] S. Jones and J. M. Thornton. Principles of protein-protein interactions. *Proc Natl Acad Sci U S A*, 93(1):13–20, January 1996.

[56] S Jones and J.M. Thornton. *Protein-Protein Recognition*, chapter Analysis and classification of protein-protein interactions from a structural perspective. Oxford University Press, 2000.

[57] Maliackal Poulo Joy, Amy Brock, Donald E Ingber, and Sui Huang. High-betweenness proteins in the yeast protein interaction network. *Journal of biomedicine & biotechnology*, 2005(2):96–103, June 2005.

[58] Ravi Kannan, Santosh Vempala, and Adrian Vetta. On clusterings: Good, bad and spectral. *J. ACM*, 51(3):497–515, 2004.

[59] Philip M Kim, Jan O Korbel, and Mark B Gerstein. Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. *Proceedings of the National Academy of Sciences of the United States of America*, 104(51):20274–9, December 2007.

[60] Dirk Koschützki and Falk Schreiber. Comparison of centralities for biological networks. In *German Conference on Bioinformatics*, pages 199–206, 2004.

[61] Oleksii Kuchaiev, Tijana Milenković, Vesna Memisević, Wayne Hayes, and Natasa Przulj. Topological network alignment uncovers biological function and phylogeny., September 2010.

[62] M. C. Lawrence and P. M. Colman. Shape complementarity at protein/protein interfaces. *J Mol Biol*, 234(4):946–50, 1993.

[63] Juyong Lee, Steven P Gross, and Jooyoung Lee. Improved network community structure improves function prediction. *Scientific reports*, 3, 2013.

[64] A.L. Lehninger, D.L. Nelson, and M.M. Cox. *Lehninger Principles of Biochemistry*. W. H. Freeman, 2005.

[65] Jure Leskovec. Stanford network analysis project. http://snap.stanford.edu/index.html, 2013.

[66] Jure Leskovec, Jon M. Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD*, pages 177–187, 2005.

[67] Jure Leskovec, Jon M. Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *TKDD*, 1(1), 2007.

[68] Jure Leskovec, Kevin J. Lang, Anirban Dasgupta, and Michael W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009.

[69] Chung-Shou Liao, Kanghao Lu, Michael Baym, Rohit Singh, and Bonnie Berger. Isorankn: Spectral methods for global alignment of multiple protein networks. *Bioinformatics*, 25(12):i253–i258, June 2009.

[70] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *ournal of the American society for information science and technology*, 58(7):1019–1031, 2007.

[71] Luana Licata, Leonardo Briganti, Daniele Peluso, Livia Perfetto, Marta Iannuccelli, Eugenia Galeota, Francesca Sacco, Anita Palma, Aurelio Nardozza, Elena Santonico, Luisa Castagnoli, and Gianni Cesareni. Mint, the molecular interaction database: 2012 update. *Nucleic Acids Research*, 40(Database-Issue):857–861, 2012.

[72] Chuan Lin, Young-rae Cho, Woo-chang Hwang, Pengjun Pei, and Aidong Zhang. Clustering methods in protein-protein interaction network. *Knowledge Discovery in Bioinformatics: Techniques, Methods and Application*, pages 1–35, 2007.

[73] Marco Loog and Robert P W Duin. Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion. *IEEE transactions on pattern analysis and machine intelligence*, 26(6):732–9, June 2004.

[74] Ulrike Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, December 2007.

[75] B. Ma, T. Elkayam, H. Wolfson, and R. Nussinov. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc Natl Acad Sci U S A*, 100(10):5772–7, 2003.

[76] J.R. Macdonald and . Johnson, Jr. Environmental features are important in determining protein secondary structure. *Protein Sci*, 10(6):1172–7, 2001.

[77] M. Madigan, J. Martinko, D. Stahl, and D.P. Clark. *Brock Biology of Microorganisms*. Pearson Education, 2011.

[78] Mina Maleki, Md. Mominul Aziz, and Luis Rueda. Analysis of obligate and non-obligate complexes using desolvation energies in domain-domain interactions. *Proceedings of the Tenth International Workshop on Data Mining in Bioinformatics - BIOKDD '11*, pages 1–6, 2011.

[79] Mina Maleki, Md. Mominul Aziz, and Luis Rueda. Analysis of relevant physicochemical properties in obligate and non-obligate protein-protein interactions. In *2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, pages 345–351. IEEE, November 2011.

[80] Mina Maleki and Luis Rueda. Domain-domain interactions in obligate and non-obligate protein-protein interactions. In *2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, pages 907–908. IEEE, November 2011.

[81] Oliver Mason and Mark Verwoerd. Graph theory and networks in biology. *Systems Biology, IET*, 1(2):89–119, 2007.

[82] Tijana Milenković and Natasa Przulj. Uncovering biological network function via graphlet degree signatures. *Cancer informatics*, 6:257–73, January 2008.

[83] R. Milo, N. Kashtan, S. Itzkovitz, M. E. J. Newman, and U. Alon. On the uniform generation of random graphs with prescribed degree sequences. *eprint arXiv:cond-mat/0312028*, December 2003.

[84] Julian Mintseris and Zhiping Weng. Atomic contact vectors in protein-protein recognition. *Proteins*, 53(3):629–639, November 2003.

[85] Julian Mintseris and Zhiping Weng. Structure, function, and evolution of transient and obligate protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 102(31):10930–5, August 2005.

[86] NCBI. Orthology. http://www.ncbi.nlm.nih.gov/education/BLASTinfo/Orthology.html, June 2011.

[87] Hani Neuvirth, Raz Ran, and Gideon Schreiber. ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *Journal of molecular biology*, 1(338), 2004.

[88] Irene M. A. Nooren and Janet M. Thornton. Diversity of protein-protein interactions. *The EMBO journal*, 22(14):3486–92, July 2003.

[89] Alberto Ochoa and Leticia Arco. Differential Betweenness in Complex Networks Clustering. In *Progress in Pattern Recognition, Image Analysis and Applications Applications*, pages 227–34. Springer, 2008.

[90] Yanay Ofran and Burkhard Rost. Analysing six types of protein-protein interfaces. *Journal of Molecular Biology*, 325(2):377–387, 2003.

[91] Roderick D.M. Page and Edward C. Holmes. *Molecular Evolution: A Phylogenetic Approach*. John Wiley & Sons, 4 August 2009.

[92] Daniel Park, Rohit Singh, Michael Baym, Chung-Shou Liao, and Bonnie Berger. Isobase: a database of functionally related proteins across ppi networks. *Nucleic acids research*, 39(Database issue):D295–300, January 2011.

[93] Georgios Pavlopoulos, Maria Secrier, Charalampos N Moschopoulos, Theodoros G Soldatos, Sophia Kossida, Jan Aerts, Reinhard Schneider, and Pantelis G Bagos. Using graph theory to analyze biological networks. *BioData mining*, 4(1):10, January 2011.

[94] Matteo Pellegrini, David Haynor, and Jason M Johnson. Protein interaction networks. *Expert review of proteomics*, 1(2):239–49, 2004.

[95] Richard W Pickersgill. A rapid method of calculating charge–charge interaction energies in proteins. *Protein engineering*, 2(3):247–248, 1988.

[96] Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, Lavanya Balakrishnan, Arivusudar Marimuthu, Sutopa Banerjee, Devi Somanathan, Aimy Sebastian, Sandhya Rani, Somak Ray, Harrys Kishore, Sashi Kanth, Mukhtar Ahmed, Manoj Kashyap, Riaz Mohmood, Y. L. Ramachandra, V. Krishna, B. Abdul Rahiman, Sujatha Mohan, Prathibha Ranganathan, Subhashri Ramabadran, Raghothama Chaerkady, and Akhilesh Pandey. Human Protein Reference Database 2009 update. *Nucleic Acids Research*, 37(1):767–772, 6 November 2008.

[97] Nataša Pržulj. Protein-protein interactions: Making sense of networks via graph-theoretic modeling. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 33(2):1–9, December 2010.

[98] Natasa Przulj, D G Corneil, and I Jurisica. Efficient estimation of graphlet frequency distributions in protein-protein interaction networks. *Bioinformatics (Oxford, England)*, 22(8):974–80, April 2006.

[99] Yanjun Qi, Ziv Bar-joseph, and Judith Klein-seetharaman. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins*, 63(3):490–500, 2006.

[100] Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2658–2663, 2004.

[101] Karthik Raman. Construction and analysis of protein-protein interaction networks. *Automated experimentation*, 2(1):2, 2010.

[102] Jane S. Richardson. The anatomy and taxonomy of protein structure. *Advances in protein chemistry*, 34:167–339, 1981.

[103] Ronald Rousseau and Lin Zhang. Betweenness centrality and q-measures in directed valued networks. *Scientometrics*, 75(3):575–590, June 2008.

[104] R. B. Russell, F. Alber, P. Aloy, F. P. Davis, D. Korkin, M. Pichaud, M. Topf, and A. Sali. A structural perspective on protein-protein interactions. *Curr Opin Struct Biol*, 14(3):313–24, 2004.

[105] Lukasz Salwinski, Christopher S. Miller, Adam J. Smith, Frank K. Pettit, James U. Bowie, and David Eisenberg. Database of Interacting Proteins. http://dip.doe-mbi.ucla.edu/dip/Main.cgi, 2004.

[106] Satu Elisa Schaeffer. Survey: Graph clustering. *Comput. Sci. Rev.*, 1(1):27–64, August 2007.

[107] H.A. Scheraga. Contribution of physical chemistry to an understanding of protein structure and function. *Protein Sci*, 1(5):691–3, 1992.

[108] H. P. Shanahan and J. M. Thornton. Amino acid architecture and the distribution of polar atoms on the surfaces of proteins. *Biopolymers*, 78(6):318–28, 2005.

[109] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.

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

[111] The UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic acids research*, 40(D1):71–75, November 2011. 1) The European Bioinformatics Institute (EMBL) 2) Swiss Institute of Bioinformatics (SIB), Centre Medical Universitaire 3) Georgetown University Medical Center, Protein Information Resource 4) University of Delaware.

[112] P Uetz, L Giot, G Cagney, T A Mansfield, R S Judson, J R Knight, D Lockshon, V Narayan, M Srinivasan, P Pochart, a Qureshi-Emili, Y Li, B Godwin, D Conover, T Kalbfleisch, G Vijayadamodar, M Yang, M Johnston, S Fields, and J M Rothberg. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403(6770):623–627, February 2000.

[113] A. Valencia and F. Pazos. Computational methods for the prediction of protein interactions. *Curr Opin Struct Biol*, 12(3):368–73, 2002.

[114] Christian Von Mering, Lars J Jensen, Berend Snel, Sean D Hooper, Markus Krupp, Mathilde Foglierini, Nelly Jouffre, Martijn a Huynen, and Peer Bork. String: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic acids research*, 33(Database issue):D433–7, January 2005.

[115] Jianxin Wang, Min Li, Youping Deng, and Yi Pan. Recent advances in clustering methods for protein interaction networks. *BMC genomics*, 11(Suppl 3):S10, January 2010.

[116] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. University of Cammbridge, 1994.

[117] Duncan J Watts and Steven H Strogatz. Collective dynamics of small-world networks. *nature*, 393(6684):440–442, 1998.

[118] J Westbrook, Z Feng, S Jain, TN Bhat, N Thanki, and et al. The protein data bank: Unifying the archive. *Nucleic Acids Res*, 30:245–248, 2002.

[119] Sarah J Wheelan, Aron Marchler-Bauer, and Stephen H Bryant. Domain size distributions can predict domain boundaries. *Bioinformatics*, 16(7):613–618, 2000.

[120] Douglas R. White and Stephen P. Borgatti. Betweenness centrality measures for directed graphs. *Social Networks*, 16(4):335–346, 1994.

[121] Ioannis Xenarios, Esteban Fernandez, Lukasz Salinski, Xiaoqun Joyce Duan, Michael J. Thompson, Edward M. Marcotte, and David Eisenberg. Dip: The database of interacting proteins: 2001 update. *Nucleic Acids Research*, 29(1):239–241, 2001.

[122] D. Xu, C.J. Tsai, and R. Nussinov. Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Eng*, 10(9):999–1012, 1997.

[123] Ziheng Yang. Paml 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8):1586–1591, 2007.

[124] Soon-Hyung Yook, Zoltán N Oltvai, and Albert-László Barabási. Functional and topological characterization of protein interaction networks. *Proteomics*, 4(4):928–942, 2004.

[125] Jeongah Yoon, Anselm Blumer, and Kyongbum Lee. An algorithm for modularity analysis of directed and weighted biological networks based on edge-betweenness centrality. *Bioinformatics (Oxford, England)*, 22(24):3106–8, December 2006.

[126] Haiyuan Yu, Kim Philip M., Emmett Sprecher, Valery Trifonov, and Mark Gerstein. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS computational biology*, 3(4):e59, 2007.

[127] C. T. Zahn. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Trans. Comput.*, 20(1):68–86, January 1971.

[128] Aidong Zhang. *Protein Interaction Networks. Computational Analysis*. Cammbridge University Press, 2009.

[129] H. X. Zhou and Y. Shan. Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins*, 44(3):336–43, 2001.

Appendix A

Limitations due to data sets available

Below are some of the inconveniences and limitations of the data available from different sources (websites and servers).

Network size. A limitation of working with several species (biological networks) is the amount of information available from each of them (different sizes, number of nodes and interactions). For some species, it is possible to have more than one network due to the large amount of studies on protein-protein interactions.

It is difficult to make comparisons between the networks because of the different sizes in their data sets (proteins and interactions), for this reason we normalized the data. The biological data (protein-protein interaction networks) used in this research is from databases BIOGRID [13] and Ensembl [29]. We prioritize the standardization of the data over its size. In the case of the social and complex networks used, we selected those networks with sizes similar to the biological networks to be able to compare them.

Balance. The number of proteins is very different from the number of interactions between species. For example in our data set, the mouse network has 24,855 proteins, but the number of interactions is only 776. In the case of yeast, there are 6,659 proteins, the number of interactions is 164,718, which is 24 times the number of its proteins [92].

PPINs have characteristics in common with social and complex networks, for example, Facebook network has 4,039 nodes and 88,234 interactions, almost 22 times the number of proteins [65].

Integration. A lack of standardized definitions among the names of genes and proteins in different networks is an important limitation to the current research and should be given consideration when drawing conclusions about the final findings. It is necessary to have some variable names in common to be able to make comparisons between different

species. As many of these data sets are non-standard and also not well documented, integrating and sharing biological data becomes a challenge.

Data format. Most of the information that we are interested in is available in public reference databases. The inconvenience is that the data are stored in a variety of formats and nomenclatures in a multitude of different systems. Finding and extracting the correct important data, combining data sources and coping with their distribution and differences is a difficult task. For example, the variety of experimental studies (yield related) over non-identical data could have inherent weaknesses that might affect the results. Examples of databases with different formats are: PROSITE [50], DIP [105], and Pfam [10]. This format problem is common, therefore attempts exist to integrate these data. An example of this is the database InterPro [5].

Appendix B

Example *dN/dS* ratio

Suppose we have the following two nucleotide sequences with their amino acid translation (protein sequences) using the universal genetic code. We need to identify the similarities and differences between this two homologous sequences and classify them as synonymous or non-synonymous.

Nucleotide Sequence 1	ACT	CCG	TTA	AGC	GTA	GGA	CAT	AGG
	- - *	- - -	* - -	- - *	- - *	- * -	* - -	* - -
Nucleotide Sequence 2	ACA	CCG	ATA	AGG	GTG	GAA	TAT	CGG
Protein Sequence 1	T	P	L	S	V	G	H	R
	-	-	*	*	-	*	*	-
Protein Sequence 2	T	P	I	R	V	E	Y	R
Codon	1	2	3	4	5	6	7	8

Table B.1: Nucleotide sequences and the respectively protein sequence.

In Table B.1, we have 24 nucleotides in each sequence, three nucleotides per codon. The asterisks in the nucleotide sequences indicate the nucleotide site where the two sequences differ. In the case of the protein sequence, we have 8 amino acids and the asterisks indicate that the amino acid are different.

From the protein sequences we can observe that four of the nucleotide substitutions cause an amino acid change. There are 4 non-synonymous (NONS) substitutions and three synonymous (SYN) substitutions, with a total of seven nucleotide substitutions.

We need to count the number of NONS and SYN nucleotide sites. We start with the first codon for both sequences, ACT (T–Thr) and ACA (T–Thr) respectively. Next, we consider the first position (a NONS site) in both sequences, the first nucleotide is *A*. Changes in this nucleotide will cause an amino acid change, for example Table B.2 (first two columns below the column 1st. position) shows the possible changes in the case of a substitution at *A*. For sequence 1 and 2, the possible changes are S–Serine, P–Proline, A–Alanine. In the last two

rows of the table are the number of NONS and SYN for the possible mutation, in this case we have zero SYN (none of the possibilities are synonymous changes) and one NONS (all the possibilities are non-synonymous changes) in every sequence.

Next, consider the second position (a NONS site), we have the same situation that first position. Any change of these nucleotides C (seq.1 and seq.2), always have a change in the amino acid in both sequences (see Table B.2 column 2nd. position). If we consider the third position (a SYN site), we can see that all the possible changes of the nucleotides T (seq.1) and A (seq.2) all form the same amino acid, T. The proportion of SYN and NONS for these sequences is SYN=1 and NONS=0.

1st pos.		2nd pos.		3rd pos.	
Seq. 1	Seq. 2	Seq. 1	Seq. 2	Seq. 1	Seq. 2
TCT=S	TCA=S	ATT=I	ATA=I	ACA=T	ACT=T
CCT=P	CCA=P	AAT=A	AAA=L	ACC=T	ACC=T
GCT=A	GCA=A	AGT=S	AGA=A	ACG=T	ACG=T
S=$\frac{0}{3}$=0	S=$\frac{0}{3}$=0	S=$\frac{0}{3}$=0	S=$\frac{0}{3}$=0	S=$\frac{3}{3}$=1	S=$\frac{3}{3}$=1
N=$\frac{3}{3}$=1	N=$\frac{3}{3}$=1	N=$\frac{3}{3}$=1	N=$\frac{3}{3}$=1	N=$\frac{0}{3}$=0	N=$\frac{0}{3}$=0

Table B.2: Analysis first site from Table B.1, first codon from both sequences ACT (amino acid T) and ACA (amino acid T).

We did the same counting for the eight codons. However, there are cases where changes in the third position made caused in the amino acid, too. When a nucleotide in the third position (in a codon) mutate and the possible resulting codons code for different amino acids than the original we need to indicate the proportion of the changes in SYN and NONS count. We code the three possible resulting codons (as we did before) and next we consider the proportion of SYN and NONS of the possibilities.

1st position		2nd position		3rd position	
Seq. 1	Seq. 2	Seq. 1	Seq. 2	Seq. 1	Seq. 2
TGC=C	TGG=W	ATC=I	ATG=M	AGT=S	AGT=S
CGC=R	CGG=R	ACC=T	ACG=T	AGA=R	AGA=R
GGC=G	GGG=G	AAC=N	AAG=K	AGG=R	AGC=S
S=0; N=1	S=$\frac{1}{3}$; N=$\frac{2}{3}$	S=0; N=1	S=0; N=1	S=$\frac{1}{3}$; N=$\frac{2}{3}$	S=$\frac{1}{3}$; N=$\frac{2}{3}$

Table B.3: Analysis third codon from Table B.1 for both sequences, TTA (amino acid L) and ATA (amino acid I).

We have this situation in codon four in Table B.3 of the example (see Table B.1). In codon four we have sequence one, AGC (Ser–S) and sequence two, AGG (Arg–R). In the $3rd position$ column we have the possible codons for sequence 1, with AGT, AGA, AGG is coding Ser–S, Arg–R, and Arg–R respectively. With 1/3 SYN (we have only one Ser–S) and 2/3 NONS (we have two other amino acid). Sequence 2 with AGT, AGA, AGC is coding Ser–S, Arg–R, and

Ser–S respectively. With 1/3 SYN (we have only one Arg–R) and 2/3 NONS (we have two other amino acid).

The overall average for the site, considering both sequences as the starting point for mutation, it is shown in the third row of Table B.4. This site made a portion SYN and a portion NONS.

Pos.	Sequence 1 and 2	S	N
1st	$\frac{1}{2}(0S+1N)+\frac{1}{2}(\frac{1}{3}S+\frac{2}{3}N)$	$\frac{1}{6}$	$\frac{5}{6}$
2nd	$\frac{1}{2}(1N)+\frac{1}{2}(1N)$	0	1
3rd	$\frac{1}{2}(\frac{1}{3}S+\frac{2}{3}N)+\frac{1}{2}(\frac{1}{3}S+\frac{2}{3}N)$	$\frac{1}{3}$	$\frac{2}{3}$

Table B.4: Proportion of SYN and NONS in codon 4.

In Table B.5 are the proportions for each site that are SYN and NONS respectively (last two rows).

Seq.1	ACT	CCG	TTA	AGC	GTA	GGA	CAT	AGG
	- - *	- - -	* - -	- - *	- - *	- * -	* - -	* - -
Seq.2	ACA	CCG	ATA	AGG	GTG	GAA	TAT	CGG
SYN	001	001	$\frac{1}{6}0\frac{1}{2}$	$\frac{1}{6}0\frac{1}{3}$	001	$00\frac{2}{3}$	$0\frac{1}{6}\frac{1}{6}$	$\frac{1}{3}0\frac{2}{3}$
NONS	110	110	$\frac{5}{6}1\frac{1}{2}$	$\frac{5}{6}1\frac{2}{3}$	110	$11\frac{1}{3}$	$1\frac{5}{6}\frac{5}{6}$	$\frac{2}{3}1\frac{1}{3}$

Table B.5: Proportions of SYN and NONS for each site.

Now we can calculate the ratio $\frac{dN}{dS}$. First, it is necessary the sum all the SYN substitutions:

$$1+1+\frac{1}{6}+\frac{1}{2}+\frac{1}{6}+\frac{1}{3}+1+\frac{2}{3}+\frac{1}{6}+\frac{1}{6}+\frac{1}{3}+\frac{2}{3}=6.1667$$

Next, the sum all the NONS substitutions:

$$1+1+1+1+\frac{5}{6}+1+\frac{1}{2}+\frac{5}{6}+1+\frac{2}{3}+1+1+1+1+\frac{1}{3}+1+\frac{5}{6}+\frac{5}{6}+\frac{2}{3}+1+\frac{1}{3}=17.8333$$

Now, we calculate the total number of NONS divided by the number of NONS sites, $dN=\frac{\#NONS}{\#NONSsites}$

$$dN=\frac{4}{17.8333}=0.224$$

And, the total number of SYN divided by the number of SYN sites, $dS=\frac{\#SYN}{\#SYNsites}$

$$dS=\frac{3}{6.1667}=0.489$$

Finally the ratio $\frac{dN}{dS}$ for these two sequences is $\frac{dN}{dS} = \frac{0.224}{0.489} = 0.458$.

According to the two initial DNA sequences, we have 16 non-synonymous sites and eight synonymous sites. From the proteins sequence view, there are four proteins substitutions between the two sequences from a total of seven proteins. We could interpret as a positive selection because more than 50 percent (four protein change from a total of seven) of the proteins in the sequences change.

However, according to our ratio results we can say that this case is a purifying selection (with a dN/dS ratio of 0.458) and not a positive selection. The possible mutations are not 16 NONS and 8 SYN but 17.8333 NONS and 6.166 SYN. The difference between the values are those cases where the nucleotides in the third position made NONS changes instead of the SYN and the cases where the amino acid tolerate changes in the first position (Arg and Leu according to the universal genetic code). The four NONS changes are over 17.8333 NONS sites and these are 22% of the total, and the three SYN changes are over 6.1667 SYN sites and these are 48.9% of the total. Proportionally there are more SYN changes than NONS, this means, there are more SYN changes. At the end, there are not making substantial changes in the sequence (conserve protein sequence).