21

A Survey on Evolutionary Analysis in PPI Networks

Pavol Jancura and Elena Marchiori
Radboud University Nijmegen
The Netherlands

1. Introduction

The analysis and application of the evolutionary information, as measured by means of the conservation of protein sequences, using protein-protein interaction (PPI) networks, has become one of the central research areas in systems biology from the last decade. It provides a promising approach for better understanding the evolution of living systems, for inferring relevant biological information about proteins, and for creating powerful protein interaction and function prediction tools. The aim of this survey is to give a general overview of the relevant literature and advances in the analysis and application of evolution in PPI networks. Due to the broad scope and vast literature on this subject, the present overview will focus on a representative selection of research directions and state-of-the-art methods to be used as a solid knowledge background for guiding the development of new hypothesis and methods aiming at the extraction and exploitation of evolutionary information in PPI networks.

This survey consists of two main parts (see Fig. 1). The first part deals with research works concerning the relation between evolution and the topological structures of a PPI network, in particular trying to discover and assess the evidence of such a relation and its strength at different granularity levels. Specifically, we consider works analysing evolution at the single protein level as well as at the level of a collection of proteins present in a PPI network. The second part of this survey describes works analysing how such evolutionary evidence can be exploited for knowledge discovery, in particular for inferring relevant biological information, such as protein interaction prediction and the discovery of functional modules conserved across multiple species.

The main terms and concepts underlying protein interaction and evolution which are used throughout the survey are summarized in the sequel. In general, a protein-protein interaction can represent different types of relations, such as a true physical bond or a functional interplay between proteins. Here, if not explicitly stated, a PPI represents a physical protein interaction as detected by experimental methods, such as yeast two-hybrid (Y2H) screening, co-immunoprecipitation or tandem affinity purification.

Two proteins are called *homologous* if they share high sequence similarity. There are two main types of homologous proteins: *orthologous* and *paralogous*. Here, for simplicity, we consider a protein pair to be orthologous if the proteins of the pair are from different species. We refer to the proteins of an orthologous pair as orthologs. Analogously, a protein pair is considered to

be paralogous if its proteins belong to the same species, in this case their proteins are called paralogs. A general assumption is that the proteins of an orthologous pair originated from a common ancestor, having been separated in evolutionary time only by a speciation event, while paralogous proteins are the product of gene duplication without speciation. The concept of orthology can be directly extended to more than two species, where one can consider clusters of orthologous proteins containing at least one protein of each species.

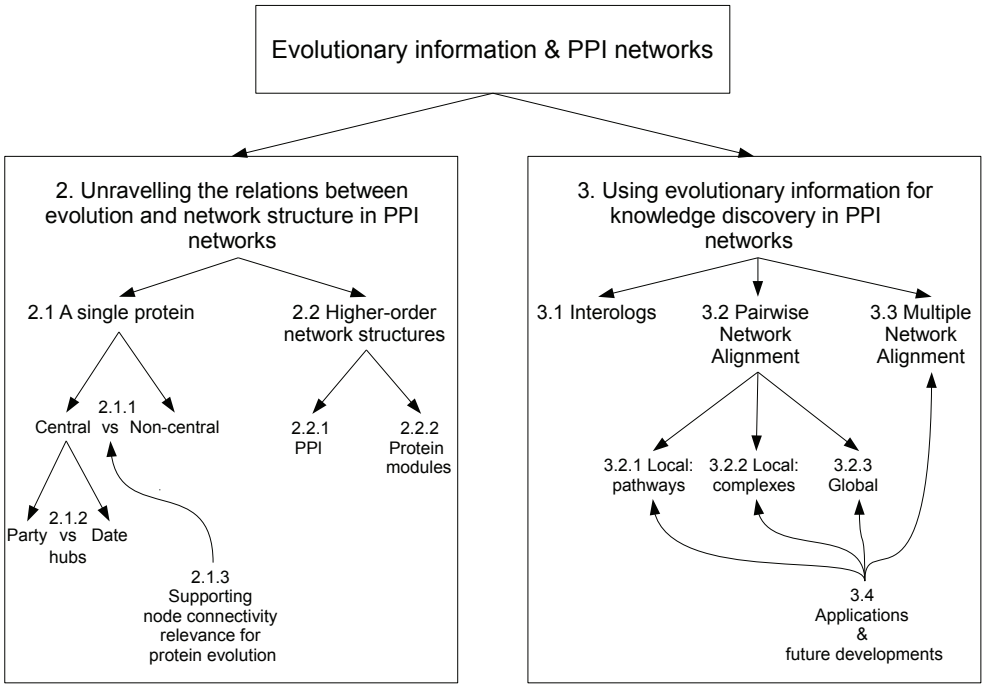


Fig. 1. The structure of the survey.

2. Unravelling the relations between evolution and network structure in PPI networks

We begin with a summary of those studies that involve the analysis of evolutionary information in a single PPI network. One can divide these works into the following two main groups. The first group studies evolutionary conservation with respect to topological properties of a PPI network. The second one primarily investigates the role of evolution with respect to the functional modules present in a PPI network.

The aim of the first group of studies is to describe how the topology of a single PPI network reflects the evolutionary signal present in the proteins it contains. This evolutionary signal is represented by the set of orthologs and it is retrieved with respect to a different species. Specifically, given a PPI network of the species to be investigated and a set of proteins of a

distinct species, those proteins of the network being a part of orthologous pairs or clusters (resulting from a sequence comparison of proteins of the two or multiple species respectively) are considered to be source of the evolutionary or orthology signal in the network. Then, having established the orthology relationship between proteins of the two or multiple species, one can estimate the evolutionary rate or distance of aligned protein sequences (see e.g. Yang & Nielsen, 2000). The higher the rate, the faster is considered the evolution of proteins. Consequently, proteins which evolve slowly are well-conserved and a little or none change to them can be observed throughout the evolution. Other protein evolutionary measures have been also considered, as propensity for gene loss, evolutionary excess retention or protein age (see Table 1).

Type of evolutionary measure	**Evolutionary measure**	**References**
Evolutionary conservation	Evolutionary Rate	e.g. Yang & Nielsen (2000), Wall et al. (2005),
	Propensity for Gene Loss	Krylov et al. (2003)
	Evolutionary Excess Retention	Wuchty (2004)
	Phyletic Retention	Gustafson et al. (2006), Chen & Xu (2005), Fang et al. (2005)
Protein age classification	Time of Origin	Kunin et al. (2004)
	Protein Age Group	Ekman et al. (2006), Kim & Marcotte (2008)

Table 1. Measures of evolutionary signal at protein level

2.1 Relation between a single protein in a PPI network and evolution

Various features of a PPI network topology can be investigated with respect to evolutionary information; the first and simplest ones are measures acting on the single nodes of the network. One can associate with a node different topological measures which estimate the relative relevance of the node within the network, here called *centrality* or *connectivity* of a node.

A basic centrality measure of a node is its degree. The degree of a node is the number of edges containing the node or, in terms of a PPI network, it is the number of proteins with which the protein represented by the node in the network interacts. It has been observed that a protein degree distribution of PPI networks follows a power law and thus PPI networks fall into a class of scale-free networks (see e.g. Jeong et al., 2001). Scale-free networks have a few highly connected nodes, called hubs, and numerous less connected nodes, which mostly interact only with one or two nodes.

2.1.1 Essentiality, centrality and conservation of a protein

As a decade ago large protein physical interaction data were not yet available, researchers mainly focussed on the study of the correlation between importance of a protein function for a living cell (essentiality, dispensability) and its evolutionary conservation rate. The generally accepted premise is that essential genes or proteins should evolve at slower rates

than non-essential ones (see e.g. Kimura, 1983). Although empirical studies have cast doubts on the validity of this hypothesis (see e.g. Hurst & Smith, 1999; Pal et al., 2003; Rocha & Danchin, 2004), in the end the vast majority and late evidences favour the existence of correlation between gene essentiality or dispensability and evolutionary conservation (see e.g. Fang et al., 2005; Fraser et al., 2002; 2003; Hahn & Kern, 2005; Hirsh & Fraser, 2001; 2003; Jordan et al., 2002; Krylov et al., 2003; Ulitsky & Shamir, 2007; Wall et al., 2005; Wang & Zhang, 2009; Waterhouse et al., 2011; Zhang & He, 2005). In particular, as recently stated by Wang & Zhang (2009), the correlation remains weak yet still conveniently sufficient for practical use.

After the growth of protein interaction data, also the correlation between essentiality and centrality, and evolutionary conservation and centrality started to be investigated. At first the *centrality-essentiality relationship* was mostly investigated by examining the degree of a node, proving the existence of the correlation (see e.g. Fraser et al., 2002; 2003; Hahn & Kern, 2005; Jeong et al., 2001; Krylov et al., 2003). However Coulomb et al. (2005) showed no correlation between essentiality and centrality, where centrality was assessed not only by the degree but also by higher order centrality measures, namely average neighbours' degree of a node and clustering coefficient of a node, suggesting that the correlation centrality-essentiality could be an artefact of the dataset. These findings were later supported by Gandhi et al. (2006) who considered a set of PPI networks and also did not observe any significant relationship between a node degree and the essentiality of the corresponding protein. Interestingly, Coulomb et al. (2005) did not test other centrality measures as betweenness and closeness, which showed a higher correlation with essentiality than just the simple degree (Hahn & Kern, 2005). Nevertheless, Batada, Hurst & Tyers (2006) reaffirmed the existence of the correlation between the node degree and essentiality taking into account Coulomb et al.'s concerns. However, Yu et al. (2008) again disputed the correlation using the compilation of Yeast high quality PPI data. Results contradicting this work appeared in two consecutive studies by Park & Kim (2009) and Pang, Sheng & Ma (2010). The first study (Park & Kim, 2009) considered also other centrality measures than just the degree of a node. As a result, the correlation could be successfully revealed, whereas the highest correlation was observed with measures based on betweenness and closeness, similarly to Hahn & Kern (2005). In the other study (Pang, Sheng & Ma, 2010) the newer, updated yeast PPI dataset was used and the correlation between degree of a node and its (protein) essentiality could be detected.

Although, the above works support that there is a connection between topological position of a node and functional importance, it seems one cannot explain this centrality-lethality rule just by the degree distribution (He & Zhang, 2006; Zotenko et al., 2008). This seems to be in accordance with the analysis conducted in (Lin et al., 2007) showing that protein domain complexity is not the single determinant of protein essentiality and that there is a correlation between the number of protein domains and the number of interactions (Schuster-Bockler & Bateman, 2007). In addition, Kafri et al. (2008) showed that highly connected essential proteins tend to have duplicates which can compensate their deletion thus decreasing the deleterious effect of their removal, a phenomenon that could possibly explain the findings that genes with no duplicates are more likely to be essential (Giaever et al., 2002). Therefore higher order topological features appear to be more appropriate for capturing gene essentiality, especially those based on node-betweenness and node-closeness (Hahn & Kern, 2005; Park & Kim, 2009; Yu et al., 2007), which are believed to estimate better the local connectivity or centrality of a

node within the network. Moreover, these features also relate with gene expression (Krylov et al., 2003; Pang, Sheng & Ma, 2010; Yu et al., 2007).

We consider now works that analyse the correlation between evolution and centrality. Also in this case the two main features used to estimate this correlation are the degree of a node and the evolutionary rate. At first, it was hypothesized that proteins with a higher degree should evolve slower (Fraser et al., 2002). A main criticism to this hypothesis was based on the fact that the analysis conducted in (Fraser et al., 2002) did not take into account the presence of a possible bias and of noise in data obtained from high-throughput experiments (Bloom & Adami, 2003; Jordan et al., 2003a;b). Nevertheless Fraser et al. (2003), Fraser & Hirsh (2004) and Lemos et al. (2005) could confirm the existence of such correlation by taking into account these objections. Kim et al. (2007) also confirmed interconnection between centrality, essentiality and conservation and showed that peripheral proteins of the PPI network are under positive selection for species adaptation. Moreover, the link between the connectivity of a node and its evolutionary history was further substantiated by works studying the correlation between node degree and other evolutionary measures such as propensity for gene loss (Krylov et al., 2003), evolutionary excess retention (Wuchty, 2004) and protein age (Ekman et al., 2006; Kunin et al., 2004). However Batada, Hurst & Tyers (2006) again pointed to a lack of evidence for a significant correlation between the evolutionary rate and the connectivity of a node. Moreover, Makino & Gojobori (2006) classified proteins according to two criteria, clustering coefficient of a node and protein's multi-functionality, and showed that multi-functional proteins of sparse parts of yeast PPI network (with a low clustering coefficient) evolve at the slowest rate regardless of the degrees of the connectivity. This suggests that clustering coefficient is a better descriptor of protein evolution within the global network of protein interactions.

A possible explanation for these conflicting results was proposed by Saeed & Deane (2006) who showed that the strength and significance of the correlation between evolution and centrality varies depending upon the type of PPI data used. Also Saeed & Deane (2006) found that more accurate datasets demonstrate stronger correlations between connectivity and evolutionary rate than less accurate datasets. Another reason may be the existence of two distinct types of highly connected nodes, so-called *party* and *date hubs*, which appear to satisfy different evolutionary constraints.

2.1.2 Evolution of party and date hubs

Specifically, Han et al. (2004) observed a bimodal distribution of average Pearson correlation coefficients between the expression profiles of proteins and its interacting partners. This yielded a classification of hubs into party hubs, having similar co-expression profiles with their neighbours, and date hubs, having different co-expression profiles with their neighbours. As a consequence, party hubs tend to interact simultaneously ("permanently") with their partners and to connect proteins within functional modules while date hubs tend to interact with different partners at different time/space ("transiently") and to bridge different modules. Thus, one may also refer to party hubs as *intramodule* and to date hubs as *intermodule* (Fraser, 2005).

Fraser (2005) was the first to investigate the difference in evolution between date and party hubs and found that party hubs are highly evolutionary constrained, whereas date hubs are

more evolutionary labile. This is clearly in accordance with findings of Mintseris & Weng (2005) who argued that residues in the interfaces of permanent protein interactions tend to evolve at a relatively slower rate, allowing them to co-evolve with their interacting partners, in contrast to the plasticity inherent in transient interactions, which leads to an increased rate of substitution for the interface residues and leaves little or no evidence of correlated mutations across the interface. The work of Fraser (2005) was, in addition, later corroborated by Bertin et al. (2007). Examining three dimensional properties of proteins also supported this hypothesis, as multi-interface hubs were found to be more evolutionary conserved and essential as well as more likely to correspond to party hubs (Kim et al., 2006). Defining singlish- and multi-Motif hubs further substantiated these findings, because multi-Motif hubs were found to be more evolutionary conserved, more essential and to correlate with multi-interface hubs (Aragues et al., 2007). In addition, other features as orderness of regions in protein sequences and the solvent accessibility of the amino acid residues was shown to be different between party and date hubs and to contribute in the lowering of the evolutionary rate of party hubs (Kahali et al., 2009). Recently, Mirzarezaee et al. (2010) applied feature selection methods and machine learning techniques to predict party and date hubs based on a set of different biological characteristics including amino acid sequences, domain contents, repeated domains, functional categories, biological processes, cellular compartments, etc.

However, other researchers disputed not only the evolutionary differences between party and date hubs but the existence of hub types as such (Agarwal et al., 2010; Batada, Reguly, Breitkreutz, Boucher, Breitkreutz, Hurst & Tyers, 2006; Batada et al., 2007). Indeed, some datasets do not exhibit clear or robust bimodal distribution of hubs' gene co-expression profiles (Agarwal et al., 2010) and in some cases there is even a complete lack of bimodality (Batada, Reguly, Breitkreutz, Boucher, Breitkreutz, Hurst & Tyers, 2006; Batada et al., 2007). Therefore, Pang, Cheng, Xuan, Sheng & Ma (2010) argue that the average Pearson correlation coefficient is a weak measure of whether a protein acts transiently or permanently with its interacting partners and they propose a new measure, a co-expressed protein-protein interaction degree. This measure estimates the actual number of partners with which a protein can permanently interact. One can interpret it as a degree of 'protein party-ness' and it offers more a continuum-like estimate of the protein's interaction property. This seems to be in accordance with Nooren & Thornton (2003) who suggest that rather a continuum range exists between distinct types of protein interactions and that their stability very much depends on the physiological conditions and environment.

Pang, Cheng, Xuan, Sheng & Ma (2010) firstly corroborated the results of Saeed & Deane (2006) on the correlation variations between connectivity and evolutionary rate of a protein on different datasets and then they showed that the co-expression-dependent node degree correlates significantly with the protein's evolutionary rate irrespectively of the specific dataset used. However, their topological measure is derived by using an external source of experimental data on gene expression. The further investigation on purely topological features of a PPI network which would distinguish transient and permanent interactions, and party and date hubs could bring more insights on how the evolutionary history of a protein is wired in its position within the network of all the protein interactions in an organism. In this perspective, network path-based measures, such as betweenness and closeness, seem to be promising (Yu et al., 2007). All the more, these measures also appear to relate to

protein essentiality (Park & Kim, 2009; Yu et al., 2007) and it could clarify the link between essentiality and evolution as such. Thereafter, they could improve on the prediction of essential genes from the topology of a PPI network in combination with protein evolutionary information, such as phyletic retention (Gustafson et al., 2006), as already corroborated by several application of machine learning techniques for essential gene detection, prioritizing drug targets and determining virulence factors (see e.g. Chen & Xu, 2005; Deng et al., 2011; Doyle et al., 2010; Gustafson et al., 2006; McDermott et al., 2009).

2.1.3 Node connectivity is relevant for protein evolution

Since the factors relevant for protein evolution could be of a multiple character (Wolf et al., 2006), it is interesting to investigate whether protein connectivity plays a central or a more subtle role. In the latter case, the link between protein connectivity and evolution could be the results of spurious correlations due to other underlying biological processes (Bloom & Adami, 2003). In order to address this issue, the contribution of protein connectivity to protein evolutionary conservation has been also studied in an integrated way (Pal et al., 2006) using multidimensional methods such as principal component analysis (PCA) and principal component regression (PCR).

The first successful application of PCA was given by Wolf et al. (2006) on seven genome-related variables. The derived first component reflected a gene's 'importance' and confirmed positive correlation between lethality, expression levels and number of protein-protein interaction which at the same time constrained protein evolution measures. Interestingly, the component also showed that the number of paralogs positively contributes to gene essentiality, which contradicts the finding of Giaever et al. (2002) that non-duplicated genes tend to be essential. However, the study of Drummond et al. (2006) revealed by using PCR only single determinant of protein evolution, namely translational selection, which is almost entirely determined by the gene expression level, protein abundance, and codon bias. Later, Plotkin & Fraser (2007) re-examined the use of PCR method and showed noise in biological data can confound PCRs, leading to spurious conclusions. As a result, when they equalized for different amounts of noise across the predictor variables no single determinant of evolution could be found indicating that a variety of factors-including expression level, gene dispensability, and protein-protein interactions may independently affect evolutionary rates in yeast. This observation was further substantiated by a recent study (Theis et al., 2011) where 16 genomic variables were analysed using Bayesian PCA. The study supports the evidence for the three above-discussed correlations. It also demonstrates how different definitions of paralogs may lead to different conclusions on their effect on essentiality, and thus commenting on Wolf et al.'s conflicting result (Wolf et al., 2006).

2.2 Higher-order structures in a PPI network and evolution

Researchers have also focused on other topological structures of a PPI network than just a node and their relation to evolutionary conservation. With increasing topological complexity we may talk about a single protein-protein interaction (an edge in PPI network), topological motifs, and protein clusters or modules as detected by their interaction density or network traffic.

2.2.1 Evolution and protein-protein interaction

Unlike in the case of a single protein, where various well-established methods for measuring sequence evolution are developed, to the best of our knowledge only a recent attempt has been made in order to estimate the evolutionary rate of protein-protein interaction (Qian et al., 2011). However, this study is limited to a small set of PPIs in yeasts and can not be yet applied for large-scale studies due to the lack of data. Thus, the research has extensively focused on estimating correlated evolution of a protein pair and their functional or physical interaction (Pazos & Valencia, 2008).

It is generally assumed that proteins which co-evolve tend to participate together in a common biological function. This hypothesis is supported by many examples of functionally interacting protein families that co-evolve (see e.g. Galperin & Koonin, 2000; Moyle et al., 1994). Co-evolution of proteins may be assessed at sequence level (*sequence co-evolution*) by correlating evolutionary rates (Clark et al., 2011), or at gene family level (*gene family evolution*) by correlating occurrence vectors (Kensche et al., 2008). An occurrence vector or a phylogenetic profile (phyletic pattern) (Tatusov et al., 1997) is an encoding of protein's (homologue's) presence or absence within a given set of species of interest (Kensche et al., 2008). In general, the methods for correlating protein evolution have been successfully applied to predict a physical or functional interaction between proteins (Clark et al., 2011; Kensche et al., 2008), where sequence co-evolution is powerful in predicting the physical interaction and phylogenetic profiling is a good indicator of functional interplay between proteins in a broader sense. Large-scale co-evolutionary maps have also been constructed and analysed for better understanding the evolution of a species and its link to protein interactions (see e.g. Cordero et al., 2008; Tillier & Charlebois, 2009; Tuller et al., 2009). All these works suggest that the topology of PPIs should reflect the evolutionary processes behind the proteins which formed such network.

The first systematic study of linked genes and their evolutionary rates was done by Williams & Hurst (2000) who showed that the rates of linked genes are more similar than the rates of random pairs of genes. Pazos & Valencia (2001) performed the first successful large-scale prediction of physical PPIs based on sequence co-evolution by correlating phylogenetic trees. Another large-scale study by Kim et al. (2004) on domain structural data of interacting protein families also revealed their high co-evolution but also showed a high diversity in the correlation of rates of each family pair. Specifically, protein families with a greater number of domains were shown to be more likely to co-evolve. However, Hakes et al. (2007) argued that this correlation of evolutionary rates is not responsible for the covariation between functional residues of interacting proteins. Nevertheless, other studies have been able to predict interacting domains from co-evolving residues between domains or proteins (see e.g. Jothi et al., 2006; Yeang & Haussler, 2007) indicating that different organisms use the same 'building blocks' for PPIs and that the functionality of many domain pairs in mediating protein interactions is maintained in evolution (Itzhaki et al., 2006).

Another perspective on co-evolution of interacting partners was given by Mintseris & Weng (2005), who distinguished between transient and obligate interactions. The authors concluded that obligate complexes are likely to co-evolve with their interacting partners, while transient interactions with an increased evolutionary rate show only little evidence for a correlated evolution of the interacting interfaces. This observation was later corroborated by Brown

& Jurisica (2007) who analysed the presence of protein interactions across multiple species via orthology mapping and found that the greater the conservation of a protein interaction is, the higher the enrichment for stable complexes. Beltrao et al. (2009) also observed that stable interactions are more conserved than transient interactions, by studying evolution of interactions involved in phosphoregulation. Finally, Zinman et al. (2011) extracted protein modules from a yeast integrated protein interaction network using various source of PPI evidence, and showed that interactions within modules were much more likely to be conserved than interactions between proteins in different modules.

The preference of conserved protein interactions to be placed in modular parts of a network was also observed by Wuchty et al. (2006) by extending the paradigm of protein's connectivity and its evolutionary conservation to the connectivity of a protein-protein interaction. Specifically, they used the hypergeometric clustering coefficient to estimate the interaction cohesiveness of the PPI's neighbourhood and orthologous excess retention in order to asses the evolutionary conservation of PPIs. They used the same clustering coefficient as that given by the presence of orthologs of interacting proteins in another organism and showed that PPIs with highly clustered environment were accompanied by an elevated propensity for the corresponding proteins to be evolutionary conserved as well as preferably co-expressed (Wuchty et al., 2006). These findings are significant all the more they were shown to be stable under perturbations. This propensity of interacting proteins to be more conserved and prevalent among taxa was later confirmed by Tillier & Charlebois (2009) who used evolutionary distances to estimate the protein's conservation. Yet another perspective on conservation of PPIs was given by Kim & Marcotte (2008) who classified proteins into four groups (from oldest to youngest) according their age and found a unique interaction density pattern between different protein age groups, where the interaction density tends to be dense within the same group and sparse between different age groups.

2.2.2 Evolution and modularity of PPI networks

All the evidences above that PPIs whose proteins are evolutionary correlated tend to form stable complexes and to be embedded in cohesive areas of a network topology support the premise that modularity of PPI networks is maintained by evolutionary pressure (Vespignani, 2003). Indeed, when examining networks solely built from sequence co-evolution, gene context analysis or gene family evolution of completely sequenced genomes, one may observe that these networks exhibit high modularity with clusters corresponding to known functional modules, thus revealing the structure of cellular organization (Cordero et al., 2008; Tuller et al., 2009; von Mering et al., 2003).

Regarding the networks of physically interacting proteins, to the best of our knowledge the first direct evidence that evolution drives the modularity of PPI networks was provided by Wuchty et al. (2003). They looked beyond a single protein pair and studied the more complex patterns of interacting proteins, called topological motifs. In general, they found that, as the number of nodes in a motif and number of links among its constituents increase, a greater and stronger conservation of the proteins could be observed. This was corroborated by Vergassola et al. (2005) who focused on specific instances of motifs known as cliques. Cliques are topological patterns where all protein constituents interact with each other. Vergassola et al. (2005) provided evidence for co-operative co-evolution within cliques of interacting

proteins. Later, Lee et al. (2006) investigated motifs at a higher resolution level, by defining for each motif different motif modes based on functional attributes of interacting proteins: again their findings indicated that motifs modes may very well represent the evolutionary conserved topological units of PPI networks. More recently, Liu et al. (2011) studied network motifs according to the age of their proteins and discovered that the proteins within motifs whose constituents are of the same age class tend to be densely interconnected, to co-evolve and to share the same biological functions. Moreover, these motifs tend to be within protein complexes.

The finding that modularity of PPI networks is constrained by evolution and that conserved interactions are enriched in dense motifs and regions of a PPI network also suggest that protein complexes present in such cohesive areas should be evolutionary driven (Jancura et al., 2012). As putative protein complexes can be extracted from a PPI network by means of clustering techniques, Jancura et al. (2012) detected such protein complexes in the PPI network consisting of only yeast proteins having an ortholog in another organism and compared them with those protein complexes derived either by using the global topology of a yeast PPI network or by using a network induced by randomly selected proteins. The in-depth examination of enriched functions in these three types of protein complexes revealed that evolutionary-driven complexes are functionally well differentiated from other two types of protein complexes found in the same interaction data. As a consequence, new complexes and protein function predictions could be unravelled from PPI data by using a standard clustering approach with the inclusion of evolutionary information. In addition, evolutionary-driven complexes were found to be differentially conserved, in particular some complexes were detected for all distinct set of orthologs as determined by comparison with different species, some exhibited only a subset of proteins identifiable in a complex across all species, and some complexes being found only for one specific set of orthologs. This suggests that presence of evolution in modularity of PPI networks is more versatile and flexible with different degrees of conservation.

The findings of Jancura et al. (2012) seem to conform with related studies that focused on evolutionary cohesiveness of protein functional modules in order to investigate whether a group of proteins which functionally interact, co-evolve more cohesively than a random group of proteins. Either known protein complexes and pathways were analysed (Fokkens & Snel, 2009; Seidl & Schultz, 2009; Snel & Huynen, 2004) or putative protein modules usually derived from integrated networks of functional link evidences (Campillos et al., 2006; Zhao et al., 2007; Zinman et al., 2011). A different strategy was employed by Yamada et al. (2006) who at first detected evolutionary modules which were afterwards compared with enzyme connectivity in a metabolic network.

Although the co-evolution of modules is assessed by the presence or absence of modules' constituents across a set of species, there is no standard method to measure the degree to which a module evolves cohesively (Fokkens & Snel, 2009). For instance, Snel & Huynen (2004) used the deviation of the number of modules' orthologs per species from the average number of modules' orthologs per species, whereas Campillos et al. (2006) measured the fraction of joined evolutionary events given the reconstructed, most parsimonious evolutionary scenario of the genes in a module over their phylogenetic profiles.

Despite this measures' diversity, the common conclusion is that the majority of modules evolve flexibly (Campillos et al., 2006; Fokkens & Snel, 2009; Seidl & Schultz, 2009; Snel & Huynen, 2004; Yamada et al., 2006). Also, it appears that curated modules evolve more cohesively than modules derived from high throughput interaction data (Fokkens & Snel, 2009; Seidl & Schultz, 2009; Snel & Huynen, 2004). Moreover, there is a different enrichment in functions which co-evolve. For example, biochemical pathways, certain metabolic and signalling processes, as well as core functions like transcription and translation, tend to have higher rate of evolutionary cohesiveness (Campillos et al., 2006; Fokkens & Snel, 2009; Zhao et al., 2007). This is also supported by methods which cluster phylogenetic profiles in order to detect biochemical pathways or to predict functional links and thus exploiting the predictive power of phylogenetic methods (Glazko & Mushegian, 2004; Li et al., 2009; Watanabe et al., 2008). These methods show a relatively good performance in characterizing biochemical pathways but seem to have a limited coverage for physically interacting proteins (Watanabe et al., 2008). A dubious result was reported on inter-connectivity of cohesive and flexible modules. Specifically, Fokkens & Snel (2009) demonstrated that components of cohesive modules are less likely to interact with each other than in the case of flexible modules, while two other studies (Campillos et al., 2006; Zinman et al., 2011) suggest cohesive modules to be more highly connected.

It is possible that the above studies underestimated the actual degree of evolutionary cohesiveness present in the modularity of protein interaction networks due to their conservative approach, the limitations in ortholog detection as well as the cohesiveness measures which are restricted to phylogenetic profiles. Nevertheless, they show that, as evolution is a complex process, its presence in modularity of protein interaction networks also exhibits a very complex nature, whose understanding is far from being complete. Evolution itself, indeed, can be expected to be asynchronous and heterotactous along the tree of life.

In general, the interim evidence shows different evolutionary pressure for different types of protein interactions and data. In particular, the slowly evolving interacting partners are enriched in stable, permanent complexes, and functional modules such as biochemical pathways and curated complexes exhibit higher evolutionary cohesiveness than high throughput complexes. It seems that the co-evolutionary degree of modules within PPI networks increases with greater integration of various sources of evidence for proteins to functionally interact (Zinman et al., 2011). Also, not all protein complexes and functional modules need to be co-evolutionary modules (Fokkens & Snel, 2009). There is a continuum from extremely conserved to rapidly changing modules, where those modules found to be co-evolving appear to be enriched in certain, specific functional categories (Campillos et al., 2006). In addition, the degree of conservation and co-evolution of functional modules within interaction networks seem to reflect cellular organization and their spatio-temporal characteristics. For instance, cohesive modules can be classified according to their evolutionary age as ancestral, intermediate and young, where one may observe ancient, ancestral modules to be highly conserved and perform essential, core processes such as information storage and metabolism of amino acids, while young modules are less conserved and responsible for the communication with the environment (Campillos et al., 2006). Therefore one might expect ancestral modules to contain static, obligate interactions as the proteins of essential functions tend to involve multiple domains with slow evolutionary

rates, whereas young modules can be enriched with dynamic, transient interactions with less but fast evolving protein domains to allow adaptation to the environment.

3. Using evolutionary information for knowledge discovery in PPI networks

The tendency of functionally linked or physically interacting proteins and densely interacting motifs to exhibit correlated evolution and/or to be conserved across species is at the core of methods for inferring relevant biological information using PPI networks. Although such biological information can be limited and biased towards specific type of known interactions and protein functions, it allows one to infer new, unknown functions of proteins, to improve the understanding of biological systems, and to guide the discovery of drug-target interaction. In its basic form, the knowledge discovery process is based on the transfer of information involving a single interaction between two organisms, while in its most complex form it involves the identification and transfer of protein complexes across multiple species. In the sequel we summarize concepts and techniques used to achieve these goals, in particular the notions of "interologs" and of multiple PPI networks alignment.

3.1 Predicting protein interaction: Interologs

If two proteins physically interact in one species and they have orthologous counterparts in another species, it is likely that their orthologs interact in that species too. If such conserved interactions exist, they are called *interologs*. This simple method of protein interaction inference was firstly introduced and tested by Walhout et al. (2000) on proteins involved in vulval development of nematode worm, where potential interactions between these proteins were identified based on interactions of their orthologs in other species. Later, Matthews et al. (2001) performed a large-scale analysis of this inference technique using the yeast PPI network as a model and proteins of worm as a target. Although the success rate of detection of inferred interactions by Y2H analysis was between 16%-31%, it represented a 600-1100-fold increase compared to a conventional approach at that time (Matthews et al., 2001).

The interologs-based protein interaction prediction has become one of the standard methods for *in silico* PPI prediction. The method can be easily extended to more PPI data from multiple species. In particular, having two groups of orthologs, where each ortholog group contains proteins from the same N species, and observing an interaction between proteins of these orthologous groups in $(N-1)$ species, the interaction between proteins of the N-th species present in the ortholog groups can be predicted. This multidimensional character of interolog inference has been extensively used to predict and build databases of the whole interactome for various species, either as a stand alone approach or in combination with other *in silico* methods, which often integrate multiple data types including the gene co-expression, co-localization, functional category, the occurrence of orthologs and other genomic context methods. In this way researchers could provide, for instance, the first sketch of human interactome (Lehner & Fraser, 2004), build the interactome of plants (Geisler-Lee et al., 2007; Gu et al., 2011), and improve the understanding of processes in a malarial parasite (Pavithra et al., 2007) or in cancer (Jonsson & Bates, 2006). Also, three, up-to-date, tools have been recently implemented and made available to perform this inference task (Gallone et al., 2011; Michaut et al., 2008; Pedamallu & Posfai, 2010).

Several algorithmic enhancements of the interologs-based approach have been introduced since the first proposal of a systematic use of interolog inference (Matthews et al., 2001). For instance, Yu et al. (2004) have strengthen the definition of ortholog by using a reciprocal best-hit approach and compared it to the original one-way best-hit approach implemented by Matthews et al. (2001). In addition, they required a minimum level for a joint similarity of orthologous sequences in order to perform interolog mapping. Their method yielded a 54% accuracy in contrast to a 30% of the previous method by Matthews et al. (2001).

Other approaches exploited the knowledge on a higher conservation rate of PPIs in dense network motifs. For instance Huang et al. (2007) scored interologs according to the density of the topological pattern containing the respective PPI of the interolog in a model species as determined by the extraction of maximal quasi-cliques from the PPI network of the model species. This score was integrated with scores of other various features used for PPI prediction, such as tissue specificity, sub-cellular localization, interacting domains and cell-cycle stage. The use of multiple types of features was shown to yield more accurate predictions of PPIs in comparison with other interolog-based methods used to build interactome databases. More recently, Jaeger et al. (2010) proposed another interesting method based on two steps. First a set of all candidate interologs is built across the considered species. Next, interologs are assembled into maximal conserved and connected patterns by detecting frequent sub-graphs appearing in the interolog network of the candidate set. Only functionally coherent patterns were used for interolog inference.

The interolog concept was also modified and used in other ways and application domains. In particular, Tirosh & Barkai (2005) proposed a method to assess and increase the confidence of a predicted PPI by examining the co-expression of proteins of its potential interolog in other species. Chen et al. (2007) extended interolog mapping for homologous inference of interacting 3D-domains and they built a database of so-called 3D-interologs (Lo, Chen & Yang, 2010). Chen et al. (2009) used interologs to transfer conserved domain-domain interactions. Recently, Lo, Lin & Yang (2010) combined this interolog domain transfer with the former 3D-interolog detection technique and implemented an integrated tool for searching homologous protein complexes. Finally, Lee et al. (2008) exploited interologs to predict inter-species interactions.

Despite the successful use of interolog inference, a gap was observed between the actual, observed number of conserved interactions and the expected theoretical coverage (Gandhi et al., 2006; Lee et al., 2008). In order to test the reliability of interolog transfer, Mika & Rost (2006) performed a comprehensive validation of the method on several datasets. Their findings suggested that interolog transfers are only accurate at very high levels of sequence identity. In addition, they also compared the interolog transfer within species and across species. In the case of within-species interolog inference a PPI is transferred onto proteins which are sequence similar to the proteins of the considered PPI in the same species. Surprisingly, such paralogous interolog transfers of protein-protein interactions were shown to be significantly more reliable than the orthologous ones. This result was later substantiated by Saeed & Deane (2008), indicating that homology-based interaction prediction methods may yield better results when within-species interolog inference is also considered. In addition, Brown & Jurisica (2007) argued that one also needs to take into account whether all interactions have equal probability of being transferred between organisms. For example, the dynamic components of the interactomes are less likely to be accurately mapped from

distantly related organisms. Moreover, there is apparent bias of interologs to be enriched in stable, permanent complexes (Brown & Jurisica, 2007), which is completely in accordance with findings on the different evolution of transient and permanent interactions. On the other hand, it is likely that the performance of interolog inference could be underestimated since its accuracy is assessed using experimentally tests based on Y2H techniques or high-throughput datasets with a high abundance in Y2H interactions, which were found to be highly enriched in transient and inter-complex connections (Yu et al., 2008).

3.2 Pairwise protein network alignment

Detection and transfer of an interolog between species have motivated the study and exploration of interspecies conservation of protein interactions on a global scale. In particular, instead of focusing on a conserved interaction alone one can compare and align whole interactome maps of distinct species, which mimics the idea behind sequence alignment methods. This approach gave a rise to so-called *network alignment* approach (Sharan & Ideker, 2006).

Using protein network alignment, one can either search for conserved functional network structures such as protein complexes and pathways, or identify functional orthologs across species. As a result this approach should provide a greater evidence and support for protein function and protein interaction prediction for yet uncharacterized or unknown biological processes. Protein network alignment methods can be classified into two main groups:*local network alignments* and *global network alignments*.

As most of the research attention has focused on comparing PPI networks of two different species, here we discuss the successive development of methods for, so-called, *pairwise network alignment*. In sequel we survey local pairwise alignments for detecting evolutionary conserved pathways, local pairwise alignments for detecting conserved protein complexes, and global pairwise network alignment techniques.

3.2.1 Local pairwise network alignment for pathway detection and query tasks

The main goal of local protein network alignment is to detect conserved pathways and protein complexes across species, by searching for local regions of input networks having both high topological similarity between the regions and high sequence similarity between proteins of these regions. The standard approach to this task consists of two main phases: *an alignment phase* and *a searching phase*. In the first phase a merged network representation of compared PPI networks is constructed, called *alignment or orthology graph*. The second phase performs a search for the structures of interest in the orthology graph. Each output result corresponds to a pair or multiplet of complexes or pathways which are evolutionary conserved across the two or more (PPI networks of the) species, respectively.

The first alignment method of whole PPI networks of two species using protein sequence similarity was introduced by Kelley et al. (2003). In this method, called *PathBLAST*, first a many-to-many mapping between proteins of the two species is determined by considering each pair of proteins with a sequence similarity higher than a given threshold as putative orthologs. Next, every orthologous pair is encoded in one alignment node of the new alignment graph and three types of edges (direct, gap and mismatch edge) are identified

between these alignment nodes as follows. The direct edge corresponds to the case when a PPI between proteins of two orthologous pairs exists in the PPI networks of both species. The gap edge represents the case when in one species the respective proteins of alignments nodes are connected indirectly through a common neighbour. Finally, the mismatch edge between alignments nodes is formed if such indirect connection is found between the corresponding proteins in the PPI networks of both species. Gap and mismatch edges are used to describe possible evolutionary variations or account for experimental errors in data (Kelley et al., 2003). In the search phase, the alignment graph is turned into acyclic sub-graphs by random removal of alignment edges, which allows to extract high-scoring paths in linear time by a dynamic programming approach. The score of a path is computed as the sum of log probabilities of true orthology encoded in alignment nodes of the path and of true conserved interactions encoded by alignment edges contained in the path. Interestingly, the method was also applied to align a PPI network with its own copy. In this way they could identify conserved (paralogous) pathways within one species.

The work of Kelley et al. (2003) was followed by other alignment techniques for discovering conserved pathways based on evolutionary conservation. The main drawbacks of *PathBLAST* are that it detects conserved linear pathways in protein interaction data, which is represented as an undirected graph, and it has an exponentially worsening efficiency with the expected increasing length of a pathway to be detected. To circumvent these limitations Pinter et al. (2005) proposed an alignment technique designed explicitly for metabolic networks with directed links between enzymes. The method also handles more complex structures than a simple path, because the scoring of the alignment is based on sub-tree homeomorphism, which can be solved by an efficient deterministic approximation. Another enhancement for the pathway alignment problem was proposed by Wernicke & Rasche (2007) who designed a method that does not impose topological restrictions upon pathways and exploits the biological and local properties of pathways within the network. Another effective approach to metabolic network alignment was developed by Li et al. (2008) which uses an integrative score on compound and enzyme similarities. Pathway alignment has been further extensively investigated and various other techniques have been proposed (see e.g. Cheng et al., 2008; Koyutürk, Kim, Subramaniam, Szpankowski & Grama, 2006; Li et al., 2007).

The evolutionary mapping of *PathBLAST* can also be used to query a known pathway of one species into the PPI network of another species. However, due to limitations and algorithmic constraints of *PathBLAST*, many other methods have been developed with a focussed application of orthologous querying of biological functional complexes, and tools and web-services are available for querying general pathways and other types of protein functional modules across species (see e.g. Bruckner et al., 2009; Dost et al., 2008; Qian et al., 2009; Yang & Sze, 2007).

3.2.2 Local pairwise network alignment for protein complex detection

Another group of methods which followed *PathBLAST* focus on detection of conserved protein complexes across (PPI networks of two or more) species. As these methods compare networks of physical interactions, the identified complexes can be used for interolog prediction as well as for protein function prediction of yet uncharacterized proteins. The detected conserved complexes are either (putative) entire physical complexes or conserved parts of them.

To the best of our knowledge, the first method for detecting conserved complexes using pairwise comparison of PPI networks was introduced by Sharan, Ideker, Kelley, Shamir & Karp (2005) and called *NetworkBLAST*. It can be viewed as a direct extension of *PathBLAST* for the task of complex detection across species. The method employs a comprehensive probabilistic model for conservation of protein complexes and searches for heavy induced sub-graphs in the weighted orthology graph. As the maximal induced sub-graph problem is computationally intractable, *NetworkBLAST* employs a bottom-up greedy heuristic for this task.

Many alignment network techniques which followed *NetworkBLAST* are motivated by the computational intractability issue derived from the problem of a finding maximal common or induced sub-graph in an ortholog graph, and are based on different heuristics. For instance, Koyutürk, Kim, Topkara, Subramaniam, Grama & Szpankowski (2006) partitions the alignment graph into smaller clusters by performing an approximated balanced ratio-cut. In another method by Koyutürk, Kim, Subramaniam, Szpankowski & Grama (2006) the most frequent interaction motifs are extracted from an orthology-contracted graph. Liang et al. (2006) transforms the problem of maximal common sub-graph into the problem of finding all maximal cliques in the graph. Recently, Tian & Samatova (2009) introduced an algorithm based on detection of connected-components of the orthology graph solvable in a very efficient way.

Other researchers propose to restrict the search space to cope with intractability issue of searching phase instead of performing heavy heuristics. For example Li et al. (2007) pre-clusters one PPI network in order to detect candidate complexes which are afterwards aligned to the target species network with an exact integer programming algorithm. Jancura & Marchiori (2010) proposed a pre-processing algorithm based on detection of network hubs for dividing PPI networks, prior to their alignment, into smaller sub-networks containing potential conserved modules. Each possible pair of sub-networks can be later aligned with a state-of-the-art alignment method where the search phase can be performed by means of an exact algorithm, allowing one to perform network comparison in a fully modular fashion and possibly to parallelize the computation. An interesting modular approach was introduced by Narayanan & Karp (2007), where an orthology graph is not constructed but rather networks are compared and split consecutively in several recursive steps until all possible solutions, conserved sub-graphs, are found. Similarly, Gerke et al. (2007) only compares, but does not merge, local hub-centred regions of PPI networks as identified by clustering coefficients and node degrees. The method by Ali & Deane (2009) is again another example of approach where an alignment graph is not explicitly constructed; there interspecies protein similarities are considered as new edges in such a way that species PPI networks and similarity edges between them are encoded into a single global meta-graph which can be searched by standard clustering techniques.

There are also alignment methods which try to incorporate or use other types of information than just the one based on sequence similarity and interaction conservation. For instance, Guo & Hartemink (2009) exploited the findings on co-evolving interacting domains which mediate PPIs and, instead of using putatively homologous proteins for alignment, compares PPI networks across species according to conserved domains of protein-protein interactions. Ali & Deane (2009) propose a functionally guided alignment of PPI networks, where a scoring function incorporates not only sequence and topological similarity of aligned proteins but also

their gene co-expression characteristics and coherence of functional annotations. Thus, the method can be seen as detecting functional modules shared across species rather than strictly evolutionary modules. Finally, Berg & Lässig (2006) developed a generalized alignment Bayesian method applicable to different biological networks.

Despite various pairwise alignment techniques have been introduced, only a few of them embody an evolutionary model of PPI networks in the scoring scheme of an alignment. Notably, Koyutürk, Kim, Topkara, Subramaniam, Grama & Szpankowski (2006) were the first to introduce a method that builds the orthology graph following the duplication/divergence model based on gene duplications. Another interesting method was proposed by Hirsh & Sharan (2007) who extended the probabilistic score of *NetworkBLAST* to asses the likelihood that two complexes originated from an ancestral complex in the common ancestor of the two species being compared under the evolutionary pressure of duplication and link dynamics events.

3.2.3 Global pairwise network alignment

In contrast to local network alignment, which uses many-to-many homologous mapping between proteins of distinct species to detect local conserved regions of a high topological similarity in the respective PPI networks, global protein network alignment uses this mapping to define an unique, globally optimal mapping across whole topologies of PPI networks (Singh et al., 2007), even if it were locally suboptimal in some regions of the networks. In the most strict form of this unique mapping each node in one input network is either matched to one node in the other input network or has no match in the other network. Thus the goal of global protein network alignment is to define functional orthologs across species, as the solution offers a way to resolve the ambiguity of orthology detection with the use of species interactome map. Naturally, as a by-product the global alignment can also identify conserved complexes or pathways.

To the best of our knowledge, the first method performing explicitly global alignment on pair of networks, called *IsoRank*, was introduced by Singh et al. (2007). Similarly to the local network alignment problem, the global network alignment problem is in general computationally intractable. As a consequence, *IsoRank* employs an approximation using an eigenvalue framework in a manner analogous to Google's PageRank algorithm.

Several advancements have naturally followed the introduction of *IsoRank*. For instance, Evans et al. (2008) proposed an asymmetric network matching algorithm based on a network simulation method called quantitative simulation, where a similarity score of a protein pair is iteratively updated by the similarity scores of their neighbours and vice versa until a unique global optimum is found. Other researchers focused more on formulating global alignment as combinatorial optimization problems. For instance Zaslavskiy et al. (2009) redefined the problem of global alignment as a standard graph matching problem and investigated methods using ideas and approaches from state-of-the-art graph matching techniques. Klau (2009) formalized global network alignment as an integer linear programming problem, where a near-optimal solution with a quality guarantee is found by solving a Lagrangian relaxation of the original optimization formulation. Recently, Chindelevitch et al. (2010) proposed a method where the global alignment is encoded as bipartite matching and applied a very efficient local optimization heuristic used for the well-known Travelling Salesman Problem.

3.3 Multiple protein network alignment

The methods on network alignment discussed so far perform alignment of two PPI networks of distinct species. The next natural extension is aligning more than two PPI networks, that is multiple network alignment. A first attempt to perform multiple local network alignment using three species was done by Sharan, Suthram, Kelley, Kuhn, McCuine, Uetz, Sittler, Karp & Ideker (2005), which exploited the scoring model of *NetworkBLAST*. However, the method scales exponentially with the number of input species and consequently it is ineffective for large scale comparisons.

Apart from the scalability problem, there are also other issues related to the problem of aligning more than two species. For instance, the putative orthologous mapping of certain proteins does not need to span across all species, meaning that proteins may be conserved only for a particular subset of species. This "orthology decay" is more evident when a large number of increasingly distant species are considered in the alignment. As a result, functional modules, such as pathways and complexes, can have a different degree of conservation, with some modules being strictly conserved across all species and some other modules being conserved only for a particular clade. Thus, a good alignment method should allow one to search for conserved modules at different degree of conservation. However, such requirement also increases the complexity of searching and consequently one may need to prune the number of all possible species combinations in alignment.

To the best of our knowledge, the first method capable of an efficient comparison of multiple PPI networks, called *Graemlin*, was introduced by Flannick et al. (2006). The alignment model of the method allows one to perform local as well as global alignment and is also applicable for querying tasks of particular biological modules of interest across PPI networks. It employs a rather involved scoring scheme which allows one to search for conserved pathways as well as for conserved complexes. It also outputs modules with a different conservation degree. *Graemlin* progressively aligns the closest pair of PPI networks according the species distance measured using a phylogenetic tree, until the last pair on the root of the tree is compared, corresponding to the most conserved parts of the aligned networks. The main disadvantage of this approach is that it involves to estimate many parameters. Recently, a supervised, automated parameter learner was proposed to lessen the burden of parameter tuning (Flannick et al., 2009).

Another phylogeny-guided local network alignment was proposed by Kalaev et al. (2008). Although the method uses the same probabilistic scoring for conserved complex as *NetworkBLAST*, it avoids its exponential scalability by redefining the alignment model such that it does not construct the merged representation of aligned networks but represents them as separate layers interconnected via orthologous mapping. Then a seed, that is, a group of putatively orthologous proteins spanning across all species, is selected using the species phylogeny and greedily expanded by adding other proteins being orthologous to each other in all respective species in order to maximize the alignment conservation score. The proposed method, however, identifies only protein complexes conserved across all species and does not detect complexes conserved only for a certain subset of species.

Notably, the functionally guided network alignment method of Ali & Deane (2009), previously mentioned as one of the methods for pairwise alignment, was also shown to perform efficiently local alignment of multiple networks.

All these multiple local network alignments do not reconstruct a plausible evolutionary history of PPI networks based on a model of evolution, although they might be phylogeny-aware. Motivated by this observation, Dutkowski & Tiuryn (2007) introduced a new multiple local network alignment method, called *CAPPI*, which from the given PPI networks of distinct species aims to reconstruct an ancient PPI network of the common ancestor. The method uses a Bayesian inference framework based on a duplication and divergence model of network evolution which mimics the processes by which most protein interactions are formed. After the reconstruction step, the ancestral network is decomposed into connected components which correspond to the ancestral modules of protein interactions and are projected back to the original networks to obtain the actual conserved network residues. Although the demonstrated application of the method was restricted to orthologous groups spanning across all species (Dutkowski & Tiuryn, 2007), to the best of our knowledge *CAPPI* is the only model-based approach for large-scale ancestral network reconstruction.

Among the multiple alignment methods above mentioned, only *Graemlin* was shown to perform a global multiple network alignment, yet it relies on a involved parameter estimation step and phylogeny-guided approximation. Recently Liao et al. (2009) developed another global alignment technique which is fully unsupervised and phylogeny-free. The method, called *IsoRankN*, is built on the *IsoRank* algorithm mentioned above (Singh et al., 2007) and its extension to the multiple global network alignment (Singh et al., 2008a). At first *IsoRankN* scores topological and sequence similarity matching between putatively orthologous proteins of each pair of input networks using *IsoRank*. Then, a maximum k-partite graph matching problem is formulated on the induced graph of pairwise alignment scores (Singh et al., 2008a) and the exact solution is approximated by a spectral graph partitioning algorithm. *IsoRankN* also effectively identifies one-to-one orthologous mappings for all subset of species and appears to out-perform *Graemlin* in terms of coverage and quality of functional enrichments.

3.4 Applications and future developments

Local and global alignment methods have been successfully applied to study evolution of species and to discover relevant biological knowledge. For example, Suthram et al. (2005) applied the network alignment of Sharan, Suthram, Kelley, Kuhn, McCuine, Uetz, Sittler, Karp & Ideker (2005) to examine the degree of conservation between the Plasmodium protein network and other model organisms, such as yeast, nematode worm, fruit fly and the bacterial pathogen Helicobacter pylori. They investigated whether the divergence of Plasmodium at the sequence level is reflected in the configuration of its protein network. Indeed, the alignments showed very little conservation suggesting that the patterns of protein interaction in Plasmodium, like its genome sequence, set it apart from other species (Suthram et al., 2005).

Another application of local network alignment was performed by Tan et al. (2007) who combined transcriptional regulatory interactions with protein-protein interactions and identified co-regulated complexes between yeast and fly revealing different conservation of their regulators. This finding advocates that PPI networks may evolve more slowly than transcriptional interaction networks. In addition, Schwartz et al. (2009) and Dutkowski & Tiuryn (2009) used conserved complexes detected by network alignments for protein interaction prediction in a manner similar to the interologs transfer approach and demonstrated their usefulness. In particular, Schwartz et al. (2009) provided a

comprehensive experimental design which includes PPI prediction using network alignment, and demonstrated how effectively it reduces the cost of interactome mapping.

Furthermore, Bandyopadhyay et al. (2006) presented the first systematic identification of functional orthologs based on protein network comparison. They used the pairwise local alignment model of Kelley et al. (2003) to construct the orthology graph and then they resolved ambiguity of orthology mapping by fitting a logistic function previously trained on a known set of functional orthologs. In contrast, Singh et al. (2008b) predicted functional orthologs in unsupervised manner by using explicitly a global multiple network alignment method.

Finally, Kolar et al. (2008) performed a cross-species analysis of two herpes-viruses using the generalized Bayesian network alignment of Berg & Lässig (2006). Interestingly, the performed alignment employs in its probabilistic scoring system evolutionary rates of sequences and thus it goes beyond the narrow use of orthologous mapping as done in all other alignment techniques. The method predicted meaningful functional associations that could not be obtained from sequence or interaction data alone.

Despite the recent progress and increasing number of network alignment tools, their further development remains an ongoing research issue, in particular for multiple network comparison. Only a few methods perform the scoring of alignment according to evolutionary models and there is only one of them which fully reconstructs network evolutionary history. This clearly is in contrast with the numerous techniques for the reconstruction of evolutionary history of gene families. Also, actual alignment methods do not distinguish among diverse types of interactions, specifically between transient and permanent interactions. For example, the prior knowledge on different evolutionary behaviour of these types of physical interactions could be incorporated into a scoring scheme of alignment construction.

In addition, all but one network comparison methods just rely on the straightforward use of putative orthologous mapping as identified by sequence comparison or available in orthologous databases, but they do not employ evolutionary measures, such as evolutionary distances or retentions, which can be derived from the corresponding sequence alignments. These measures assess the level of evolutionary conservation and they could potentially improve the performance of network alignments.

Mostly all current applications of network alignments have worked with networks of physical interactome. However, the power of network alignment for functional annotation and other system biology applications could be explored when one performs comparison of more general, functional interaction networks. One may expect that such alignment could reveal a higher number of conserved modules as the interspecies conservation of modularity across protein networks increases with combined, integrated evidence for a pair of proteins to be functionally linked. Finally, all available methods here considered focused on conservation of modules but not on the more general concept of module evolutionary cohesiveness or co-evolution. The evolutionary cohesiveness can be assessed especially for the case of multiple alignments. Indeed, all conserved modules are inherently very cohesive, however not all evolutionary modules need to exhibit the correlated conservation at a level as expected by actual multiple network alignments. Protein functional modules differ in the degree of conservation and also in the degree of cohesiveness.

4. References

Agarwal, S., Deane, C. M., Porter, M. A. & Jones, N. S. (2010). Revisiting date and party hubs: Novel approaches to role assignment in protein interaction networks, *PLoS Comput Biol* 6(6): e1000817.

Ali, W. & Deane, C. M. (2009). Functionally guided alignment of protein interaction networks for module detection, *Bioinformatics* 25(23): 3166–3173.

Aragues, R., Sali, A., Bonet, J., Marti-Renom, M. A. & Oliva, B. (2007). Characterization of protein hubs by inferring interacting motifs from protein interactions, *PLoS Comput Biol* 3(9): e178.

Bandyopadhyay, S., Sharan, R. & Ideker, T. (2006). Systematic identification of functional orthologs based on protein network comparison, *Genome Research* 16(3): 428–435.

Batada, N. N., Hurst, L. D. & Tyers, M. (2006). Evolutionary and physiological importance of hub proteins, *PLoS Comput Biol* 2(7): e88.

Batada, N. N., Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B.-J., Hurst, L. D. & Tyers, M. (2006). Stratus not altocumulus: A new view of the yeast protein interaction network, *PLoS Biol* 4(10): e317.

Batada, N. N., Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B.-J., Hurst, L. D. & Tyers, M. (2007). Still stratus not altocumulus: Further evidence against the date/party hub distinction, *PLoS Biol* 5(6): e154.

Beltrao, P., Trinidad, J. C., Fiedler, D., Roguev, A., Lim, W. A., Shokat, K. M., Burlingame, A. L. & Krogan, N. J. (2009). Evolution of phosphoregulation: Comparison of phosphorylation patterns across yeast species, *PLoS Biol* 7(6): e1000134.

Berg, J. & Lässig, M. (2006). Cross-species analysis of biological networks by Bayesian alignment, *Proceedings of the National Academy of Sciences* 103(29): 10967–10972.

Bertin, N., Simonis, N., Dupuy, D., Cusick, M. E., Han, J.-D. J., Fraser, H. B., Roth, F. P. & Vidal, M. (2007). Confirmation of organized modularity in the yeast interactome, *PLoS Biol* 5(6): e153.

Bloom, J. & Adami, C. (2003). Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets, *BMC Evolutionary Biology* 3(1): 21.

Brown, K. & Jurisica, I. (2007). Unequal evolutionary conservation of human protein interactions in interologous networks, *Genome Biology* 8(5): R95.

Bruckner, S., Hüffner, F., Karp, R. M., Shamir, R. & Sharan, R. (2009). Torque: topology-free querying of protein interaction networks., *Nucleic Acids Research* 37(Web Server issue): W106–108.

Campillos, M., von Mering, C., Jensen, L. J. & Bork, P. (2006). Identification and analysis of evolutionarily cohesive functional modules in protein networks, *Genome Research* 16(3): 374–382.

Chen, C.-C., Lin, C.-Y., Lo, Y.-S. & Yang, J.-M. (2009). Ppisearch: a web server for searching homologous protein-protein interactions across multiple species, *Nucleic Acids Research* 37(suppl 2): W369–W375.

Chen, Y.-C., Lo, Y.-S., Hsu, W.-C. & Yang, J.-M. (2007). 3d-partner: a web server to infer interacting partners and binding models, *Nucleic Acids Research* 35(suppl 2): W561–W567.

Chen, Y. & Xu, D. (2005). Understanding protein dispensability through machine-learning analysis of high-throughput data, *Bioinformatics* 21(5): 575–581.

Cheng, Q., Berman, P., Harrison, R. & Zelikovsky, A. (2008). Fast alignments of metabolic networks, *BIBM '08: Proceedings of the 2008 IEEE International Conference on Bioinformatics and Biomedicine*, IEEE Computer Society, Washington, DC, USA, pp. 147–152.

Chindelevitch, L., Liao, C.-S. & Berger, B. (2010). Local optimization for global alignment of protein interaction networks, *Pacific Symposium on Biocomputing* 15: 123–132.

Clark, G. W., Dar, V.-u.-N., Bezginov, A., Yang, J. M., Charlebois, R. L. & Tillier, E. R. M. (2011). Using coevolution to predict protein-protein interactions, *in* G. Cagney, A. Emili & J. M. Walker (eds), *Network Biology*, Vol. 781 of *Methods in Molecular Biology*, Humana Press, pp. 237–256.

Cordero, O. X., Snel, B. & Hogeweg, P. (2008). Coevolution of gene families in prokaryotes, *Genome Research* 18(3): 462–468.

Coulomb, S., Bauer, M., Bernard, D. & Marsolier-Kergoat, M.-C. (2005). Gene essentiality and the topology of protein interaction networks, *Proceedings of the Royal Society B: Biological Sciences* 272(1573): 1721–1725.

Deng, J., Deng, L., Su, S., Zhang, M., Lin, X., Wei, L., Minai, A. A., Hassett, D. J. & Lu, L. J. (2011). Investigating the predictability of essential genes across distantly related organisms using an integrative approach, *Nucleic Acids Research* 39(3): 795–807.

Dost, B., Shlomi, T., Gupta, N., Ruppin, E., Bafna, V. & Sharan, R. (2008). Qnet: A tool for querying protein interaction networks, *Journal of Computational Biology* 15(7): 913–925.

Doyle, M., Gasser, R., Woodcroft, B., Hall, R. & Ralph, S. (2010). Drug target prediction and prioritization: using orthology to predict essentiality in parasite genomes, *BMC Genomics* 11(1): 222.

Drummond, D. A., Raval, A. & Wilke, C. O. (2006). A single determinant dominates the rate of yeast protein evolution, *Molecular Biology and Evolution* 23(2): 327–337.

Dutkowski, J. & Tiuryn, J. (2007). Identification of functional modules from conserved ancestral protein-protein interactions, *Bioinformatics* 23(13): i149–158.

Dutkowski, J. & Tiuryn, J. (2009). Phylogeny-guided interaction mapping in seven eukaryotes, *BMC Bioinformatics* 10(1): 393.

Ekman, D., Light, S., Björklund, A. K. & Elofsson, A. (2006). What properties characterize the hub proteins of the protein-protein interaction network of saccharomyces cerevisiae?, *Genome Biology* 7(6): R45.

Evans, P., Sandler, T. & Ungar, L. (2008). Protein-protein interaction network alignment by quantitative simulation, *BIBM '08: Proceedings of the 2008 IEEE International Conference on Bioinformatics and Biomedicine*, IEEE Computer Society, Washington, DC, USA, pp. 325–328.

Fang, G., Rocha, E. & Danchin, A. (2005). How essential are nonessential genes?, *Molecular Biology and Evolution* 22(11): 2147–2156.

Flannick, J., Novak, A., Do, C. B., Srinivasan, B. S. & Batzoglou, S. (2009). Automatic parameter learning for multiple local network alignment, *Journal of Computational Biology* 16(8): 1001–1022.

Flannick, J., Novak, A., Srinivasan, B. S., McAdams, H. H. & Batzoglou, S. (2006). Graemlin: General and robust alignment of multiple large interaction networks, *Genome Res.* 16(9): 1169–1181.

Fokkens, L. & Snel, B. (2009). Cohesive versus flexible evolution of functional modules in eukaryotes, *PLoS Comput Biol* 5(1): e1000276.

Fraser, H. B. (2005). Modularity and evolutionary constraint on proteins, *Nat Genet* 37(4): 351 – 352.

Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C. & Feldman, M. W. (2002). Evolutionary rate in the protein interaction network, *Science* 296(5568): 750–752.

Fraser, H. & Hirsh, A. (2004). Evolutionary rate depends on number of protein-protein interactions independently of gene expression level, *BMC Evolutionary Biology* 4(1): 13.

Fraser, H., Wall, D. & Hirsh, A. (2003). A simple dependence between protein evolution rate and the number of protein-protein interactions, *BMC Evolutionary Biology* 3(1): 11.

Gallone, G., Simpson, T. I., Armstrong, J. D. & Jarman, A. (2011). Bio::homology::interologwalk - a perl module to build putative protein-protein interaction networks through interolog mapping, *BMC Bioinformatics* 12(1): 289.

Galperin, M. Y. & Koonin, E. V. (2000). Who's your neighbor? new computational approaches for functional genomics, *Nat Biotech* 18(6): 609–613.

Gandhi, T. K. B., Zhong, J., Mathivanan, S., Karthick, L., Chandrika, K. N., Mohan, S. S., Sharma, S., Pinkert, S., Nagaraju, S., Periaswamy, B., Mishra, G., Nandakumar, K., Shen, B., Deshpande, N., Nayak, R., Sarker, M., Boeke, J. D., Parmigiani, G., Schultz, J., Bader, J. S. & Pandey, A. (2006). Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets, *Nat Genet* 38(3): 285 – 293.

Geisler-Lee, J., O'Toole, N., Ammar, R., Provart, N. J., Millar, A. H. & Geisler, M. (2007). A predicted interactome for arabidopsis, *Plant Physiology* 145(2): 317–329.

Gerke, M., Bornberg-Bauer, E., Jiang, X. & Fuellen, G. (2007). Finding common protein interaction patterns across organisms, *Evolutionary bioinformatics online* 2: 45–52.

Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B., Arkin, A. P., Astromoff, A., El Bakkoury, M., Bangham, R., Benito, R., Brachat, S., Campanaro, S., Curtiss, M., Davis, K., Deutschbauer, A., Entian, K.-D., Flaherty, P., Foury, F., Garfinkel, D. J., Gerstein, M., Gotte, D., Guldener, U., Hegemann, J. H., Hempel, S., Herman, Z., Jaramillo, D. F., Kelly, D. E., Kelly, S. L., Kotter, P., LaBonte, D., Lamb, D. C., Lan, N., Liang, H., Liao, H., Liu, L., Luo, C., Lussier, M., Mao, R., Menard, P., Ooi, S. L., Revuelta, J. L., Roberts, C. J., Rose, M., Ross-Macdonald, P., Scherens, B., Schimmack, G., Shafer, B., Shoemaker, D. D., Sookhai-Mahadeo, S., Storms, R. K., Strathern, J. N., Valle, G., Voet, M., Volckaert, G., Wang, C.-y., Ward, T. R., Wilhelmy, J., Winzeler, E. A., Yang, Y., Yen, G., Youngman, E., Yu, K., Bussey, H., Boeke, J. D., Snyder, M., Philippsen, P., Davis, R. W. & Johnston, M. (2002). Functional profiling of the saccharomyces cerevisiae genome, *Nature* 418: 387–391.

Glazko, G. & Mushegian, A. (2004). Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns, *Genome Biology* 5(5): R32.

Gu, H., Zhu, P., Jiao, Y., Meng, Y. & Chen, M. (2011). Prin: a predicted rice interactome network, *BMC Bioinformatics* 12(1): 161.

Guo, X. & Hartemink, A. J. (2009). Domain-oriented edge-based alignment of protein interaction networks, *Bioinformatics* 25(12): i240–1246.

Gustafson, A., Snitkin, E., Parker, S., DeLisi, C. & Kasif, S. (2006). Towards the identification of essential genes using targeted genome sequencing and comparative analysis, *BMC Genomics* 7(1): 265.

Hahn, M. W. & Kern, A. D. (2005). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks, *Molecular Biology and Evolution* 22(4): 803–806.

Hakes, L., Lovell, S. C., Oliver, S. G. & Robertson, D. L. (2007). Specificity in protein interactions and its relationship with sequence diversity and coevolution, *Proceedings of the National Academy of Sciences* 104(19): 7999–8004.

Han, J.-D. J., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J. M., Cusick, M. E., Roth, F. P. & Vidal, M. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network, *Nature* 430: 88–93.

He, X. & Zhang, J. (2006). Why do hubs tend to be essential in protein networks?, *PLoS Genet* 2(6): e88.

Hirsh, A. E. & Fraser, H. B. (2001). Protein dispensability and rate of evolution, *Nature* 411: 1046–1049.

Hirsh, A. E. & Fraser, H. B. (2003). Genomic function (communication arising): Rate of evolution and gene dispensability, *Nature* 421(6922): 497–498.

Hirsh, E. & Sharan, R. (2007). Identification of conserved protein complexes based on a model of protein network evolution, *Bioinformatics* 23(2): e170–176.

Huang, T.-W., Lin, C.-Y. & Kao, C.-Y. (2007). Reconstruction of human protein interolog network using evolutionary conserved network, *BMC Bioinformatics* 8(1): 152.

Hurst, L. D. & Smith, N. G. (1999). Do essential genes evolve slowly?, *Current biology* 9: 747–750.

Itzhaki, Z., Akiva, E., Altuvia, Y. & Margalit, H. (2006). Evolutionary conservation of domain-domain interactions, *Genome Biology* 7(12): R125.

Jaeger, S., Sers, C. & Leser, U. (2010). Combining modularity, conservation, and interactions of proteins significantly increases precision and coverage of protein function prediction, *BMC Genomics* 11(1): 717.

Jancura, P. & Marchiori, E. (2010). Dividing protein interaction networks for modular network comparative analysis, *Pattern Recognition Letters* 31(14): 2083 – 2096.

Jancura, P., Mavridou, E., Carrillo-De Santa Pau, E. & Marchiori, E. (2012). A methodology for detecting the orthology signal in a ppi network at a functional complex level, *BMC Bioinformatics* 13(Suppl 1). In press.

Jeong, H., Mason, S. P., Barabasi, A.-L. & Oltvai, Z. N. (2001). Lethality and centrality in protein networks, *Nature* 411: 41–42.

Jonsson, P. F. & Bates, P. A. (2006). Global topological features of cancer proteins in the human interactome, *Bioinformatics* 22(18): 2291–2297.

Jordan, I. K., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. (2002). Essential genes are more evolutionarily conserved than are nonessential genes in bacteria, *Genome Research* 12(6): 962–968.

Jordan, I. K., Wolf, Y. & Koonin, E. (2003a). Correction: No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly, *BMC Evolutionary Biology* 3(1): 5.

Jordan, I. K., Wolf, Y. & Koonin, E. (2003b). No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly, *BMC Evolutionary Biology* 3(1): 1.

Jothi, R., Cherukuri, P. F., Tasneem, A. & Przytycka, T. M. (2006). Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions, *Journal of Molecular Biology* 362(4): 861 – 875.

Kafri, R., Dahan, O., Levy, J. & Pilpel, Y. (2008). Preferential protection of protein interaction network hubs in yeast: Evolved functionality of genetic redundancy, *Proceedings of the National Academy of Sciences* 105(4): 1243–1248.

Kahali, B., Ahmad, S. & Ghosh, T. C. (2009). Exploring the evolutionary rate differences of party hub and date hub proteins in saccharomyces cerevisiae protein-protein interaction network, *Gene* 429(1-2): 18 – 22.

Kalaev, M., Bafna, V. & Sharan, R. (2008). Fast and accurate alignment of multiple protein networks, *Research in Computational Molecular Biology*, pp. 246–256.

Kelley, B. P., Sharan, R., Karp, R. M., Sittler, T., Root, D. E., Stockwell, B. R. & Ideker, T. (2003). Conserved pathways within bacteria and yeast as revealed by global protein network alignment, *Proceedings of the National Academy of Science* 100: 11394–11399.

Kensche, P. R., van Noort, V., Dutilh, B. E. & Huynen, M. A. (2008). Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution, *Journal of The Royal Society Interface* 5(19): 151–170.

Kim, P. M., Korbel, J. O. & Gerstein, M. B. (2007). Positive selection at the protein network periphery: Evaluation in terms of structural constraints and cellular context, *Proceedings of the National Academy of Sciences* 104(51): 20274–20279.

Kim, P. M., Lu, L. J., Xia, Y. & Gerstein, M. B. (2006). Relating three-dimensional structures to protein networks provides evolutionary insights, *Science* 314(5807): 1938–1941.

Kim, W. K., Bolser, D. M. & Park, J. H. (2004). Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (psimap), *Bioinformatics* 20(7): 1138–1150.

Kim, W. K. & Marcotte, E. M. (2008). Age-dependent evolution of the yeast protein interaction network suggests a limited role of gene duplication and divergence, *PLoS Comput Biol* 4(11): e1000232.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*, Cambridge University Press.

Klau, G. (2009). A new graph-based method for pairwise global network alignment, *BMC Bioinformatics* 10(Suppl 1): S59.

Kolar, M., Lassig, M. & Berg, J. (2008). From protein interactions to functional annotation: graph alignment in herpes, *BMC Systems Biology* 2(1): 90.

Koyutürk, M., Kim, Y., Subramaniam, S., Szpankowski, W. & Grama, A. (2006). Detecting conserved interaction patterns in biological networks, *Journal of Computational Biology* 13(7): 1299–1322.

Koyutürk, M., Kim, Y., Topkara, U., Subramaniam, S., Grama, A. & Szpankowski, W. (2006). Pairwise alignment of protein interaction networks, *Journal of Computional Biology* 13(2): 182–199.

Krylov, D. M., Wolf, Y. I., Rogozin, I. B. & Koonin, E. V. (2003). Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution, *Genome Research* 13(10): 2229–2235.

Kunin, V., Pereira-Leal, J. B. & Ouzounis, C. A. (2004). Functional evolution of the yeast protein interaction network, *Molecular Biology and Evolution* 21(7): 1171–1176.

Lee, S.-A., Chan, C.-h., Tsai, C.-H., Lai, J.-M., Wang, F.-S., Kao, C.-Y. & Huang, C.-Y. (2008). Ortholog-based protein-protein interaction prediction and its application to inter-species interactions, *BMC Bioinformatics* 9(Suppl 12): S11.

Lee, W.-P., Jeng, B.-C., Pai, T.-W., Tsai, C.-P., Yu, C.-Y. & Tzou, W.-S. (2006). Differential evolutionary conservation of motif modes in the yeast protein interaction network, *BMC Genomics* 7(1): 89.

Lehner, B. & Fraser, A. (2004). A first-draft human protein-interaction map, *Genome Biology* 5(9): R63.

Lemos, B., Bettencourt, B. R., Meiklejohn, C. D. & Hartl, D. L. (2005). Evolution of proteins and gene expression levels are coupled in drosophila and are independently associated with mrna abundance, protein length, and number of protein-protein interactions, *Molecular Biology and Evolution* 22(5): 1345–1354.

Li, H., Kristensen, D. M., Coleman, M. K. & Mushegian, A. (2009). Detection of biochemical pathways by probabilistic matching of phyletic vectors, *PLoS ONE* 4(4): e5326.

Li, Y., de Ridder, D., de Groot, M. & Reinders, M. (2008). Metabolic pathway alignment between species using a comprehensive and flexible similarity measure, *BMC Systems Biology* 2(1): 111.

Li, Z., Zhang, S., Wang, Y., Zhang, X.-S. & Chen, L. (2007). Alignment of molecular networks by integer quadratic programming, *Bioinformatics* 23(13): 1631–1639.

Liang, Z., Xu, M., Teng, M. & Niu, L. (2006). Comparison of protein interaction networks reveals species conservation and divergence, *BMC Bioinformatics* 7(1): 457.

Liao, C.-S., Lu, K., Baym, M., Singh, R. & Berger, B. (2009). IsoRankN: spectral methods for global alignment of multiple protein networks, *Bioinformatics* 25(12): i253–258.

Lin, Y.-S., Hwang, J.-K. & Li, W.-H. (2007). Protein complexity, gene duplicability and gene dispensability in the yeast genome, *Gene* 387(1-2): 109 – 117.

Liu, Z., Liu, Q., Sun, H., Hou, L., Guo, H., Zhu, Y., Li, D. & He, F. (2011). Evidence for the additions of clustered interacting nodes during the evolution of protein interaction networks from network motifs, *BMC Evolutionary Biology* 11(1): 133.

Lo, Y.-S., Chen, Y.-C. & Yang, J.-M. (2010). 3d-interologs: an evolution database of physical protein- protein interactions across multiple genomes, *BMC Genomics* 11(Suppl 3): S7.

Lo, Y.-S., Lin, C.-Y. & Yang, J.-M. (2010). Pcfamily: a web server for searching homologous protein complexes, *Nucleic Acids Research* 38(suppl 2): W516–W522.

Makino, T. & Gojobori, T. (2006). The evolutionary rate of a protein is influenced by features of the interacting partners, *Molecular Biology and Evolution* 23(4): 784–789.

Matthews, L. R., Vaglio, P., Reboul, J., Ge, H., Davis, B. P., Garrels, J., Vincent, S. & Vidal, M. (2001). Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs", *Genome Research* 11(12): 2120–2126.

McDermott, J. E., Taylor, R. C., Yoon, H. & Heffron, F. (2009). Bottlenecks and hubs in inferred networks are important for virulence in salmonella typhimurium, *Journal of Computational Biology* 16: 169–180.

Michaut, M., Kerrien, S., Montecchi-Palazzi, L., Chauvat, F., Cassier-Chauvat, C., Aude, J.-C., Legrain, P. & Hermjakob, H. (2008). Interoporc: automated inference of highly conserved protein interaction networks, *Bioinformatics* 24(14): 1625–1631.

Mika, S. & Rost, B. (2006). Protein-protein interactions more conserved within species than across species, *PLoS Comput Biol* 2(7): e79.

Mintseris, J. & Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein-protein interactions, *Proceedings of the National Academy of Sciences* 102(31): 10930–10935.

Mirzarezaee, M., Araabi, B. & Sadeghi, M. (2010). Features analysis for identification of date and party hubs in protein interaction network of saccharomyces cerevisiae, *BMC Systems Biology* 4(1): 172.

Moyle, W. R., Campbell, R. K., Myers, R. V., Bernard, M. P., Han, Y. & Wang, X. (1994). Co-evolution of ligand-receptor pairs, *Nature* 368(6468): 251–255.

Narayanan, M. & Karp, R. M. (2007). Comparing protein interaction networks via a graph match-and-split algorithm, *Journal of Computational Biology* 14(7): 892–907.

Nooren, I. M. & Thornton, J. M. (2003). Diversity of protein-protein interactions, *EMBO J* 22(14): 3486–3492.

Pal, C., Papp, B. & Hurst, L. D. (2003). Genomic function (communication arising): Rate of evolution and gene dispensability, *Nature* 421(6922): 496–497.

Pal, C., Papp, B. & Lercher, M. J. (2006). An integrated view of protein evolution, *Nat Rev Genet* 7: 337–348.

Pang, K., Cheng, C., Xuan, Z., Sheng, H. & Ma, X. (2010). Understanding protein evolutionary rate by integrating gene co-expression with protein interactions, *BMC Systems Biology* 4(1): 179.

Pang, K., Sheng, H. & Ma, X. (2010). Understanding gene essentiality by finely characterizing hubs in the yeast protein interaction network, *Biochemical and Biophysical Research Communications* 401(1): 112 – 116.

Park, K. & Kim, D. (2009). Localized network centrality and essentiality in the yeast-protein interaction network, *PROTEOMICS* 9(22): 5143–5154.

Pavithra, S. R., Kumar, R. & Tatu, U. (2007). Systems analysis of chaperone networks in the malarial parasite plasmodium falciparum, *PLoS Comput Biol* 3(9): e168.

Pazos, F. & Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction, *Protein Engineering* 14(9): 609–614.

Pazos, F. & Valencia, A. (2008). Protein co-evolution, co-adaptation and interactions, *EMBO J* 27(20): 2648–2655.

Pedamallu, C. S. & Posfai, J. (2010). Open source tool for prediction of genome wide protein-protein interaction network based on ortholog information, *Source Code for Biology and Medicine* 5(1): 8.

Pinter, R. Y., Rokhlenko, O., Yeger-Lotem, E. & Ziv-Ukelson, M. (2005). Alignment of metabolic pathways, *Bioinformatics* 21(16): 3401–3408.

Plotkin, J. B. & Fraser, H. B. (2007). Assessing the determinants of evolutionary rates in the presence of noise, *Molecular Biology and Evolution* 24(5): 1113–1121.

Qian, W., He, X., Chan, E., Xu, H. & Zhang, J. (2011). Measuring the evolutionary rate of protein-protein interaction, *Proceedings of the National Academy of Sciences* 108(21): 8725–8730.

Qian, X., Sze, S.-H. & Yoon, B.-J. (2009). Querying Pathways in Protein Interaction Networks Based on Hidden Markov Models, *Journal of Computational Biology* 16(2): 145–157.

Rocha, E. P. C. & Danchin, A. (2004). An analysis of determinants of amino acids substitution rates in bacterial proteins, *Molecular Biology and Evolution* 21(1): 108–116.

Saeed, R. & Deane, C. (2006). Protein protein interactions, evolutionary rate, abundance and age, *BMC Bioinformatics* 7(1): 128.

Saeed, R. & Deane, C. (2008). An assessment of the uses of homologous interactions, *Bioinformatics* 24(5): 689–695.

Schuster-Bockler, B. & Bateman, A. (2007). Reuse of structural domain-domain interactions in protein networks, *BMC Bioinformatics* 8(1): 259.

Schwartz, A. S., Yu, J., Gardenour, K. R., Finley Jr, R. L. & Ideker, T. (2009). Cost-effective strategies for completing the interactome, *Nat Meth* 6(1): 55–61.

Seidl, M. & Schultz, J. (2009). Evolutionary flexibility of protein complexes, *BMC Evolutionary Biology* 9(1): 155.

Sharan, R. & Ideker, T. (2006). Modeling cellular machinery through biological network comparison, *Nature Biotechnology* 24(4): 427–433.

Sharan, R., Ideker, T., Kelley, B. P., Shamir, R. & Karp, R. M. (2005). Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data, *Journal of Computional Biology* 12(6): 835–846.

Sharan, R., Suthram, S., Kelley, R. M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R. M. & Ideker, T. (2005). From the Cover: Conserved patterns of protein interaction in multiple species, *Proceedings of the National Academy of Sciences* 102(6): 1974–1979.

Singh, R., Xu, J. & Berger, B. (2007). Pairwise global alignment of protein interaction networks by matching neighborhood topology, *Research in Computational Molecular Biology* pp. 16–31.

Singh, R., Xu, J. & Berger, B. (2008a). Global alignment of multiple protein interaction networks, *Pacific Symposium on Biocomputing* 13: 303–314.

Singh, R., Xu, J. & Berger, B. (2008b). Global alignment of multiple protein interaction networks with application to functional orthology detection, *Proceedings of the National Academy of Sciences* 105(35): 12763–12768.

Snel, B. & Huynen, M. A. (2004). Quantifying modularity in the evolution of biomolecular systems, *Genome Research* 14(3): 391–397.

Suthram, S., Sittler, T. & Ideker, T. (2005). The plasmodium protein network diverges from those of other eukaryotes, *Nature* 438(7064): 108–112.

Tan, K., Shlomi, T., Feizi, H., Ideker, T. & Sharan, R. (2007). Transcriptional regulation of protein complexes within and across species, *Proceedings of the National Academy of Sciences* 104(4): 1283–1288.

Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997). A genomic perspective on protein families, *Science* 278(5338): 631–637.

Theis, F. J., Latif, N., Wong, P. & Frishman, D. (2011). Complex principal component and correlation structure of 16 yeast genomic variables, *Molecular Biology and Evolution* 28(9): 2501–2512.

Tian, W. & Samatova, N. F. (2009). Pairwise alignment of interaction networks by fast identification of maximal conserved patterns, *Pacific Symposium on Biocomputing* 14: 99–110.

Tillier, E. R. & Charlebois, R. L. (2009). The human protein coevolution network, *Genome Research* 19(10): 1861–1871.

Tirosh, I. & Barkai, N. (2005). Computational verification of protein-protein interactions by orthologous co-expression, *BMC Bioinformatics* 6(1): 40.

Tuller, T., Kupiec, M. & Ruppin, E. (2009). Co-evolutionary networks of genes and cellular processes across fungal species, *Genome Biology* 10(5): R48.

Ulitsky, I. & Shamir, R. (2007). Pathway redundancy and protein essentiality revealed in the saccharomyces cerevisiae interaction networks, *Mol Syst Biol* 3: 1–7.

Vergassola, M., Vespignani, A. & Dujon, B. (2005). Cooperative evolution in protein complexes of yeast from comparative analyses of its interaction network, *PROTEOMICS* 5(12): 3116–3119.

Vespignani, A. (2003). Evolution thinks modular, *Nature Genetics* 35(2): 118–119.

von Mering, C., Zdobnov, E. M., Tsoka, S., Ciccarelli, F. D., Pereira-Leal, J. B., Ouzounis, C. A. & Bork, P. (2003). Genome evolution reveals biochemical networks and functional modules, *Proceedings of the National Academy of Sciences* 100(26): 15428–15433.

Walhout, A. J. M., Sordella, R., Lu, X., Hartley, J. L., Temple, G. F., Brasch, M. A., Thierry-Mieg, N. & Vidal, M. (2000). Protein interaction mapping in c. elegans using proteins involved in vulval development, *Science* 287(5450): 116–122.

Wall, D. P., Hirsh, A. E., Fraser, H. B., Kumm, J., Giaever, G., Eisen, M. B. & Feldman, M. W. (2005). Functional genomic analysis of the rates of protein evolution, *Proceedings of the National Academy of Sciences* 102(15): 5483–5488.

Wang, Z. & Zhang, J. (2009). Why is the correlation between gene importance and gene evolutionary rate so weak?, *PLoS Genet* 5(1): e1000329.

Watanabe, R., Morett, E. & Vallejo, E. (2008). Inferring modules of functionally interacting proteins using the bond energy algorithm, *BMC Bioinformatics* 9(1): 285.

Waterhouse, R. M., Zdobnov, E. M. & Kriventseva, E. V. (2011). Correlating traits of gene retention, sequence divergence, duplicability and essentiality in vertebrates, arthropods, and fungi, *Genome Biology and Evolution* 3: 75–86.

Wernicke, S. & Rasche, F. (2007). Simple and fast alignment of metabolic pathways by exploiting local diversity, *Bioinformatics* 23(15): 1978–1985.

Williams, E. J. B. & Hurst, L. D. (2000). The proteins of linked genes evolve at similar rates, *Nature* 407(6806): 900–903.

Wolf, Y. I., Carmel, L. & Koonin, E. V. (2006). Unifying measures of gene function and evolution, *Proceedings of the Royal Society B: Biological Sciences* 273(1593): 1507–1515.

Wuchty, S. (2004). Evolution and topology in the yeast protein interaction network, *Genome Research* 14(7): 1310–1314.

Wuchty, S., Barabasi, A.-L. & Ferdig, M. (2006). Stable evolutionary signal in a yeast protein interaction network, *BMC Evolutionary Biology* 6(1): 8.

Wuchty, S., Oltvai, Z. N. & Barabási, A.-L. (2003). Evolutionary conservation of motif constituents in the yeast protein interaction network, *Nature Genetics* 35(2): 176–179.

Yamada, T., Kanehisa, M. & Goto, S. (2006). Extraction of phylogenetic network modules from the metabolic network, *BMC Bioinformatics* 7(1): 130.

Yang, Q. & Sze, S.-H. (2007). Path matching and graph matching in biological networks, *Journal of Computational Biology* 14(1): 56–67.

Yang, Z. & Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models, *Molecular Biology and Evolution* 17(1): 32–43.

Yeang, C.-H. & Haussler, D. (2007). Detecting coevolution in and among protein domains, *PLoS Comput Biol* 3(11): e211.

Yu, H., Braun, P., Yildirim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., Hao, T., Rual, J.-F., Dricot, A., Vazquez, A., Murray, R. R., Simon, C., Tardivo, L., Tam, S., Svrzikapa, N., Fan, C., de Smet, A.-S., Motyl, A., Hudson, M. E., Park, J., Xin, X., Cusick, M. E., Moore, T.,

Boone, C., Snyder, M., Roth, F. P., Barabási, A.-L., Tavernier, J., Hill, D. E. & Vidal, M. (2008). High-quality binary protein interaction map of the yeast interactome network, *Science* 322(5898): 104–110.

Yu, H., Kim, P. M., Sprecher, E., Trifonov, V. & Gerstein, M. (2007). The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics, *PLoS Comput Biol* 3(4): e59.

Yu, H., Luscombe, N. M., Lu, H. X., Zhu, X., Xia, Y., Han, J.-D. J., Bertin, N., Chung, S., Vidal, M. & Gerstein, M. (2004). Annotation transfer between genomes: Protein-protein interologs and protein-dna regulogs, *Genome Research* 14(6): 1107–1118.

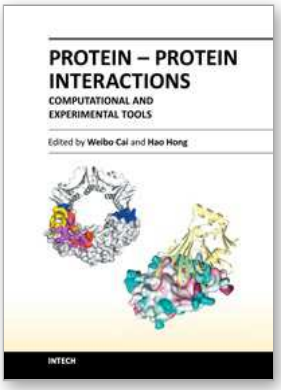
Zaslavskiy, M., Bach, F. & Vert, J.-P. (2009). Global alignment of protein-protein interaction networks by graph matching methods, *Bioinformatics* 25(12): i259–1267.

Zhang, J. & He, X. (2005). Significant impact of protein dispensability on the instantaneous rate of protein evolution, *Molecular Biology and Evolution* 22(4): 1147–1155.

Zhao, J., Ding, G.-H., Tao, L., Yu, H., Yu, Z.-H., Luo, J.-H., Cao, Z.-W. & Li, Y.-X. (2007). Modular co-evolution of metabolic networks, *BMC Bioinformatics* 8(1): 311.

Zinman, G., Zhong, S. & Bar-Joseph, Z. (2011). Biological interaction networks are conserved at the module level, *BMC Systems Biology* 5(1): 134.

Zotenko, E., Mestre, J., O'Leary, D. P. & Przytycka, T. M. (2008). Why do hubs in the yeast protein interaction network tend to be essential: Reexamining the connection between the network topology and essentiality, *PLoS Comput Biol* 4(8): e1000140.



Protein-Protein Interactions - Computational and Experimental Tools

Edited by Dr. Weibo Cai

ISBN 978-953-51-0397-4
Hard cover, 472 pages
Publisher InTech
Published online 30, March, 2012
Published in print edition March, 2012

Proteins are indispensable players in virtually all biological events. The functions of proteins are coordinated through intricate regulatory networks of transient protein-protein interactions (PPIs). To predict and/or study PPIs, a wide variety of techniques have been developed over the last several decades. Many in vitro and in vivo assays have been implemented to explore the mechanism of these ubiquitous interactions. However, despite significant advances in these experimental approaches, many limitations exist such as false-positives/false-negatives, difficulty in obtaining crystal structures of proteins, challenges in the detection of transient PPI, among others. To overcome these limitations, many computational approaches have been developed which are becoming increasingly widely used to facilitate the investigation of PPIs. This book has gathered an ensemble of experts in the field, in 22 chapters, which have been broadly categorized into Computational Approaches, Experimental Approaches, and Others.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Pavol Jancura and Elena Marchiori (2012). A Survey on Evolutionary Analysis in PPI Networks, Protein-Protein Interactions - Computational and Experimental Tools, Dr. Weibo Cai (Ed.), ISBN: 978-953-51-0397-4, InTech, Available from: http://www.intechopen.com/books/protein-protein-interactions-computational-and-experimental-tools/a-survey-on-evolutionary-analysis-in-ppi-networks

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821