Computational and Structural Biotechnology Journal 11 (2014) 1–10

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Mini Review

Using biological networks to improve our understanding of infectious diseases

Nicola J. Mulder *, Richard O. Akinola, Gaston K. Mazandu, Holifidy Rapanoel

Computational Biology Group, Department of Clinical Laboratory Sciences, IDM, University of Cape Town Faculty of Health Sciences, Anzio Road, Observatory, Cape Town, South Africa

ARTICLE INFO

Available online 27 August 2014

Keywords:
Biological networks
Tuberculosis
Pathogen
Evolution
Protein–protein interaction

ABSTRACT

Infectious diseases are the leading cause of death, particularly in developing countries. Although many drugs are available for treating the most common infectious diseases, in many cases the mechanism of action of these drugs or even their targets in the pathogen remain unknown. In addition, the key factors or processes in pathogens that facilitate infection and disease progression are often not well understood. Since proteins do not work in isolation, understanding biological systems requires a better understanding of the interconnectivity between proteins in different pathways and processes, which includes both physical and other functional interactions. Such biological networks can be generated within organisms or between organisms sharing a common environment using experimental data and computational predictions. Though different data sources provide different levels of accuracy, confidence in interactions can be measured using interaction scores. Connections between interacting proteins in biological networks can be represented as graphs and edges, and thus studied using existing algorithms and tools from graph theory. There are many different applications of biological networks, and here we discuss three such applications, specifically applied to the infectious disease tuberculosis, with its causative agent *Mycobacterium tuberculosis* and host, *Homo sapiens*. The applications include the use of the networks for function prediction, comparison of networks for evolutionary studies, and the generation and use of host–pathogen interaction networks.

Contents

1. Introduction

The biology of organisms is complex and involves the interplay between numerous factors, including proteins, nucleic acids and small

* Corresponding author. Tel.: +27 21 4066058.
E-mail addresses: Nicola.mulder@uct.ac.za (N.J. Mulder), roakinola@gmail.com (R.O. Akinola), gmazandu@cbio.uct.ac.za (G.K. Mazandu), holy@aims.ac.za (H. Rapanoel).

http://dx.doi.org/10.1016/j.csbj.2014.08.006
2001-0370/

molecules. These, in turn, are influenced by the environment and evolve to enable adaptation to environmental niches. Bacterial pathogens have evolved to infect their hosts through multiple mechanisms, including horizontal gene transfer [1], mutations [2], gene duplications [3] and gene loss [4]. In order to study infectious diseases caused by bacterial pathogens, we need to improve our understanding of the underlying molecular biology of these organisms so that we can determine how they infect, persist and cause disease, as well as better understand the pharmacokinetic and pharmacogenomic actions of anti-bacterial drugs.

The functioning of a biological system is largely driven by proteins, which interact and work together in pathways and processes. Therefore to understand the system, proteins must be studied within the context of their interactions with other proteins, rather than in isolation. Proteins can interact through direct physical binding, or through indirect associations, such as contributing to the same biological process. Protein–protein interaction networks are probably the most used example of biological networks, and can include interactions from both physical protein–protein binding as well as other functional interactions [5]. The vast amount of data generated over the years by different high-throughput biological technologies has raised the need for an integrative approach where datasets from heterogeneous sources are merged into a single network of interacting proteins. In these biological networks, the nodes are proteins and the edges represent functional interactions between proteins which can be derived from a variety of different data sources [6]. These sources include direct physical binding, for which there are a number of protein–protein interaction databases (e.g. IntAct, DIP, BIND), co-expression, functional similarity, text-mining, co-localization and other functional genomics data sources [6].

Biological networks provide the starting point for a number of analyses that aim to improve our understanding of biological systems [7]. Since biological networks are depicted as network graphs, many of these analysis tools draw on concepts and algorithms from graph theory. These allow us to, for example, determine the properties of nodes, such as their degree (number of neighbours), betweenness and centrality, which provide a feeling of how important that node is in facilitating communication between other nodes in the network and in holding connected components of the network together. We can also perform in silico knock-out studies to determine the potential impact of targeting a particular protein. Identifying the essentiality of proteins and the effect of knocking out the protein in the biological network of a pathogen has the potential to enable in silico prediction of potential drug targets when studying infectious diseases. There are many other applications of biological networks, and in this article we review some of these applications in studying human pathogens, using examples from our work on *Mycobacterium tuberculosis* and related mycobacteria. *M. tuberculosis* is the causative agent of tuberculosis (TB), an infectious disease of epidemic proportions in developing countries. First we review the use of protein–protein functional interaction (PPI) networks for protein function prediction (note, functional interactions include all functional connections between proteins, not only physical binding), and then we demonstrate how networks can facilitate evolutionary studies between pathogenic and non-pathogenic strains with differing genome sizes by comparing three different networks. Finally, we review some methods for generating host–pathogen interaction networks to improve our understanding of the interplay between host and pathogen during infection, not only using the *M. tuberculosis*–human interaction network as an example but also providing use cases from other host–pathogen studies.

2. Use of biological networks for function prediction

The completion of several sequencing projects and other high-throughput biological technologies has generated complete genome sequences and functional genomics data for several organisms. The abundance of these diverse biological data from various sources constitutes a rich source of knowledge, providing valuable insights into the dynamics driving collective and specific features of these organisms, and shedding light on the targeted organism's biology. Despite the uncontested successes recorded from comparative and functional genomics in gaining a better understanding of these organisms' biology and evolution, a number of challenges still remain. One of the main challenges is the lack of functional annotations for a relatively high proportion of genes and thus proteins within genomes. From 20 to 50% of genes within a genome are still annotated as ‘unknown’, ‘uncharacterized’ or ‘hypothetical’, and this limits our ability to exploit these data [8], leading to the paradigm of “a world which is data rich yet information poor”. *M. tuberculosis* contains a large number of “uncharacterised” or “hypothetical” proteins, which limits our ability both to understand their role in the pathogenesis of TB and to determine their potential as drug targets.

Proteins perform an astonishing range of biological functions in an organism, including roles as structural proteins, as enzymes and for the transportation of materials within and between cells. Each protein is a gene product that interacts with the cellular environment in some way to promote the cell's growth and function, implying that knowledge of protein functions and their biological pathways is crucial for understanding an organism's behaviour. Thus, one of the major tasks in the post-genomic era is genome annotation, or assigning functions to gene products in order to capitalize on the knowledge gained through different biological data produced. This requires a systematic description of the attributes of genes and proteins without any ambiguity using a standardized syntax and semantics in a format that is human readable and understandable, as well as interpretable computationally [9]. One of the biggest accomplishments in this area is the creation of the Gene Ontology (GO), which currently serves as the dominant and most popular functional classification scheme for annotation and functional representation of genes and their products [10].

The initial computational approach for assigning functions to an uncharacterized protein uses sequence similarity search tools, such as the Basic Local Alignment Search Tool (BLAST) [11]. This approach is referred to as homology-based annotation transfer, providing a straightforward scheme for suggesting possible functions for uncharacterized proteins. The key assumption driving this approach is that two proteins with significantly similar sequences are evolutionarily linked and might thus share common functions. However, some factors limit its applicability; for example, no known sequence may be similar to the novel protein sequence in the database, and above all, the most significant database hit may perform a different function due to gene duplication events [12,13], domain shuffling events (deletions), or single point mutations [14]. Several approaches that do not rely directly on sequence similarity have also been implemented, which include using information about gene fusions, phylogenetic profiles of protein families, gene adjacency in genomes and expression patterns [15]. Below we describe the concept of and algorithms for function prediction and the use of GO and biological networks to achieve this.

2.1. Protein function and Gene Ontology

From a mathematical point of view, transference of a functional label from a set A to a set B is a rule which associates each object (input) ‘x’ in A with at most one object (output) ‘y’ in B. In this case, ‘y’ represents the realization of ‘x’, called a function of ‘x’. For a function to be well-defined one needs to know the two sets A and B and the rule of associations of objects or realizations of all objects of A. Without loss of generality, a set is a collection of well-defined objects, and if A and B are well described, then a function is completely determined by knowing just the realizations of objects. Similarly, assuming the context and the scope of interest are known, protein function is a concept used to describe all types of realizations or activities to which the protein contributes, which take place within an organism, and which have consequences at the cellular and system levels [16]. Thus, the concept "protein function" may be subjective and ambiguous without describing the context and the scope of interest.

This observation suggests that protein function assignment requires the characterization of protein contributions using well-defined and structured vocabularies specifying the aspect and the context surrounding these contributions. GO [10] provides a way of consistently describing genes and proteins in three key biological aspects of genes in a living cell, namely, description of the tasks that are carried out by the proteins (molecular function, MF), their broad biological goals (biological process, BP), and the subcellular components, or locations where the activities are taking place (cellular component, CC) [8]. These ontologies are engineered as a directed acyclic graph (DAG), and produce a well-adapted platform to computationally process data at the functional level [17]. GO has been widely adopted and successfully deployed in several biological and biomedical applications, ranging from theoretical to experimental and computational biology [9].

2.2. Protein function prediction algorithms

Producing high-quality and accurate protein functions is challenging, as manual or experimental approaches are expensive and time-consuming, so the number of manually determined protein functions available for a particular genome is usually far fewer than that produced by computational or electronic approaches. Furthermore, finding functions of uncharacterized proteins experimentally is challenging for several reasons; for example function may be specifically related to the native environment in which a particular organism lives, the gene may have no use in the laboratory environment, or it may be impossible to imitate the natural host, with its myriad of other micro-organisms. Therefore we cannot always determine the exact function of a gene or gene product by experiments alone [16]. As a result, protein functions assigned using computational approaches are dominant and this is the most likely future trend as they currently represent more than 99% of annotations in the GO annotation (GOA-UniProtKB) dataset (http://www.ebi.ac.uk/GOA/uniprot_release).

As an illustration, for the pathogen *M. tuberculosis* (MTB) strain CDC1551, available data shows 4202 protein coding genes, but only 2694 are annotated, with a total of 1114, 2282 and 2322 proteins characterized with respect to the CC, MF and BP ontologies, respectively, as extracted from the latest version of GOA (version 130, released on 15 April, 2014: http://www.ebi.ac.uk/GOA/proteomes). From these, only 166, 44 and 98 entries contain annotations manually assigned with respect to the CC, MF and BP ontologies, respectively. It is worth mentioning that the GO evidence codes found from the Experimental category are: Inferred from Direct Assay (IDA), Inferred from Physical Interaction (IPI), Inferred from Mutant Phenotype (IMP), Inferred from Genetic Interaction (IPI) and Inferred from Expression Pattern (IEP), but the evidence code Inferred from Experiment (EXP) was not found at all for this organism. Furthermore, we identified some cases (see Table 1 for five examples for BP and eleven for MF) of functional inference in which functions predicted computationally or electronically were experimentally validated or vice versa.

Before using computational methods for protein function prediction and annotation with GO terms, we first analysed the specificity of terms inferred computationally and manually in annotated MTB proteins, using the GO-universal metric [8] to compute similarity between terms in a given ontology. Semantic similarity measures provide a numerical value for the similarity between two GO terms based on their position in the DAG and are used to differentiate between terms that are related but differ because of their different levels of specificity and those that appear in different paths and are thus unrelated. The results in Fig. 1 indicate that sometimes computational approaches produce more specific annotations, e.g. for the BP ontology, while manual inference provides more specific annotations for the CC and MF ontologies. The ideal is to predict the most specific term in the ontology when annotating an uncharacterized protein, but this is hard to achieve since it requires enough experimental evidence when using manual inference or more information for electronic inference. Fig. 1 indicates that there are instances where manual inference produces more specific annotations compared to electronic inference and others where electronic inference provides more specific annotations than manual inference. This suggests that the best route toward the elucidation of the function of uncharacterized proteins may include a combination of experimental approaches and predictions through computational analysis [5]. Here we focus on computational prediction.

Based on the fact that a protein does not achieve its function alone but cooperates with other proteins to perform that function, several approaches have been adopted for the use of the structure of protein–protein interaction networks for predicting protein functions. This approach provides the advantage of alleviating the impact of the reliability issue of data from different experiments which can be noisy. Combining information from multiple sources into one unified network should lead to higher confidence and increased coverage, and improve the prediction analyses performed on the basis of these networks [18]. There are several goals that a function prediction algorithm needs to meet, which include improvement in annotation quality and genomic coverage, i.e., to increase the proportion of genes or gene products in a genome which are annotated [16]. Despite the high degree of noise

Table 1
Proteins in MTB strain CDC1551 annotated with the same BP and MF ontology terms from electronic and manual inferences. The level indicates the level of the GO term in the GO DAG, assuming that the root of each ontology is located at level 0. The manual evidence code is provided, together with the source of electronic inferences.

Protein	GO ID	GO term	Level	Evidence code, source
Biological process				
Q7D8E1	GO:0045454	Cell redox homeostasis	5	TAS/IDA, InterPro
Q7D8E1	GO:0055114	Oxidation–reduction process	3	IDA, GOC
P95276	GO:0008152	Metabolic process	1	IDA, GOC
O07218	GO:0016998	Cell wall macromolecule catabolic process	6	IDA, InterPro
P71937	GO:0006355	Regulation of transcription, DNA-templated	8	IDA, InterPro
P71971	GO:0006979	Response to oxidative stress	3	IMP, InterPro
Molecular function				
O53294	GO:0004497	Monooxygenase activity	3	IDA, UniProt, IEA UniProt
P0CF99	GO:0043750	Phosphatidylinositol alpha-mannosyltransferase activity	5	IDA, UniProt
P96291	GO:0016747	Transferase activity, transferring acyl groups other than amino-acyl groups	4	IDA, InterPro
Q7D4L9	GO:0008745	N-acetylmuramoyl-L-alanine amidase activity	5	IDA, InterPro
P71855	GO:0016747	Transferase activity, transferring acyl groups other than amino-acyl groups	4	IDA, InterPro
P95001	GO:0004764	Shikimate 3-dehydrogenase (NADP+) activity	5	TAS/IDA, InterPro/UniProt
O33342	GO:0004356	Glutamate-ammonia ligase activity	6	IDA, InterPro
P71828	GO:0003840	Gamma-glutamyltransferase activity	6	IDA, InterPro/UniProt
Q7D8E1	GO:0015035	Protein disulfide oxidoreductase activity	5	IDA, InterPro
O53665	GO:0004316	3-Oxoacyl-[acyl-carrier-protein] reductase (NADPH) activity	6	IDA, UniProt
P96830	GO:0016791	Phosphatase activity	5	IDA, InterPro

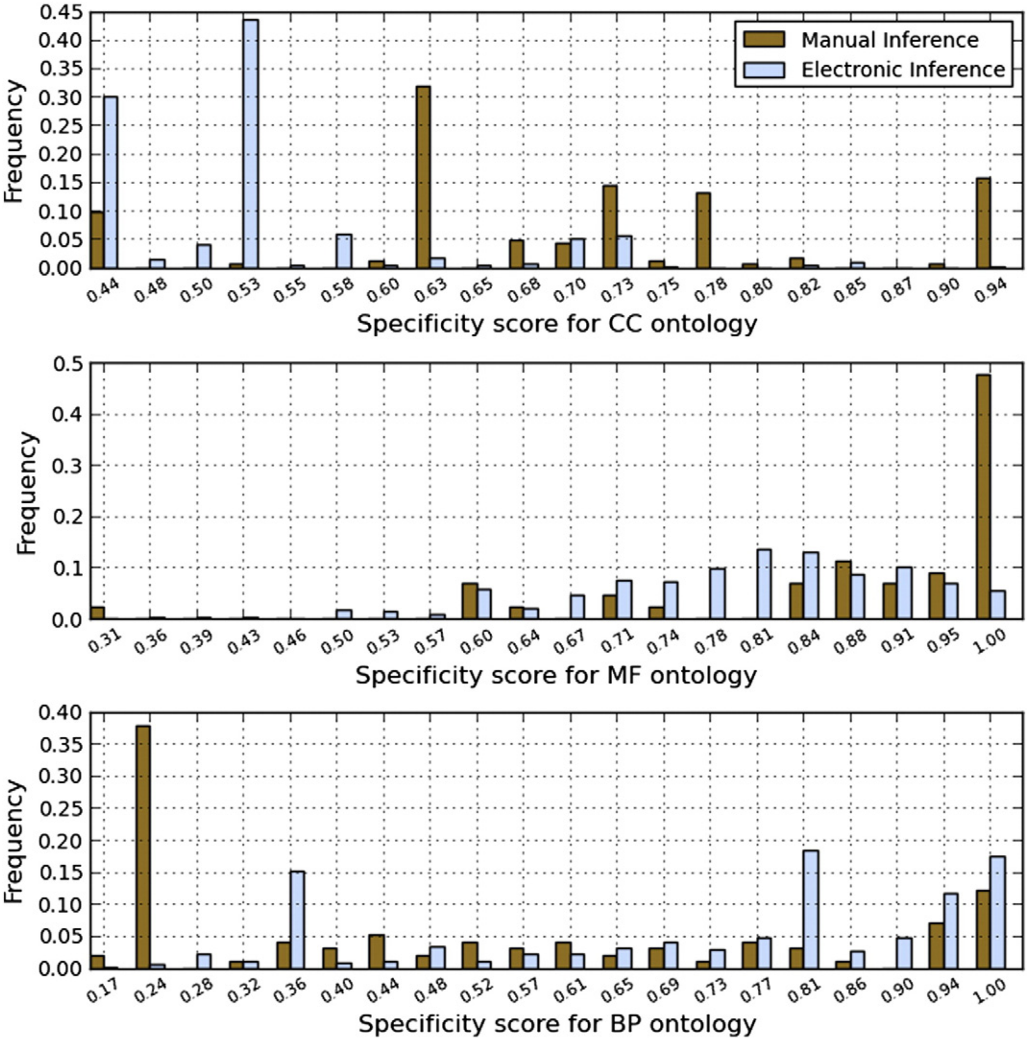

Fig. 1. Comparison of annotations inferred manually and electronically in the MTB genome strain CDC1551 in terms of term specificity score computed using the GO-universal metric.

that interaction data from high throughput experiments contain, making them potentially unreliable, uncontested successes have been recorded from the use of computational approaches to predict functions of uncharacterized proteins using these data. In fact, the progress made in the use of computational approaches to predict protein function from diverse types of biological data has positively impacted the functional genomics research field. As shown in Table 1, in the context of MTB strain CDC1551, there are cases of functional inference in which functions predicted computationally were experimentally validated and vice versa.

Several approaches for predicting protein functions from protein–protein interaction networks have been proposed and are mainly classified into two categories, namely global network topology and local neighbourhood based approaches. Global network topology based approaches use global optimization [19–21], probabilistic methods [22–25] or machine learning [26–29] to improve the prediction accuracy using the global structure of the network under consideration. A detailed review of these approaches, particularly on machine learning based approaches can be found in [30]. In the case of local neighbourhood based approaches, also known as 'Guilt-by-Association', 'Majority Voting' or 'Neighbour Counting' [31], direct interacting neighbours of proteins are used to predict protein functions. The dualism of "Guilt-by-Association" and "Global" prediction approaches for characterizing a protein has raised a debate separating the Bioinformatics community into divergent groups with differing views. On one hand, there are proponents of the "Guilt-by-Association" strategy, stating that a gene or gene product shares the function of the most closely related genes of known functions, thus predicting protein functions by observing the patterns of each protein's neighbourhood. This fraction highlights the inability of global prediction approaches to provide significant improvement over the simple and elegant local prediction approach [32]. On the other hand, the advocates of the "Global" prediction approach argue for a global view of the protein–protein interaction network to achieve efficient annotation prediction.

From the computational side, the "Guilt-by-Association" prediction approach may be an excellent and more straightforward approach since the "Global" prediction approach raises a scalability issue for large networks which may not be proportional to the prediction improvement. However, this straightforward approach may lead to systematic error especially in the case where the protein under consideration does not share functions with any of its direct neighbours [16]. This case was depicted in the Yeast Proteome Network and therefore, Jin and Cho [33] proposed a new approach for dealing with this type of local protein association behaviour by building a "Protein Interaction Network Dictionary" (PIND) in which the protein target's function is obtained from characterized proteins whose direct interacting neighbours share a certain level of similarity with the protein target's interacting neighbours. This phenomenon has also been demonstrated by Chua et al. [34,35], who showed that in many cases proteins share functional similarity with level-2 neighbours, and level-2 neighbours have an above average likelihood of sharing functional similarity. They introduced a functional similarity weight (FS-Weight) method for predicting protein functions from protein interaction data using level-1 and level-2 neighbours.

2.3. Deploying GO in genome annotation analysis

A general shortcoming of different function prediction approaches and genome annotation analyses is that they do not always effectively consider the structure of the ontology being used to predict these functions. In some cases, such as in the use of GO slim, the level at which a term can be considered to be specific or informative in the ontology hierarchy is fixed in order to do GO term comparison. However, it is evident that while using a subset of GO terms or a reduced version of GO, such as GO slim, to compare genes which makes GO terms and annotations easier to work with, valuable information is lost in the simplification [9]. This partial coverage of the annotation structure may, therefore, compromise the prediction outcomes and annotation analyses [16].

In order to incorporate the whole structure of the GO DAG, it is important to make use of a semantic similarity measure in the annotation analysis or prediction algorithm under consideration. Semantic similarity measures allow integration of the biological knowledge contained in the GO DAG, and have contributed to the improvement of biological analyses [8,17]. This was explored in [16] to predict functions of uncharacterized proteins in *M. tuberculosis* strain CDC1551 using the strategy summarized in Fig. 2. This scheme presents an annotation prediction model which uses direct interacting neighbours combined with second level interacting neighbours to achieve efficient trade-off between the scalability issue, prediction improvement and genomic coverage. It uses GO-based semantic similarity measures to propagate annotations from characterized proteins to uncharacterized proteins, incorporating relationships between terms in the GO DAG structure in the prediction process. The cut-off similarity score that enables an annotation occurring among protein neighbours to be more accurately assigned to the protein under consideration is estimated by applying the algorithm to a dataset with some known annotations removed, i.e. one can measure the performance, such as precision or accuracy, of the prediction algorithm on proteins with known annotations in the network. These semantic similarity and cut-off scores are then used to predict annotations of uncharacterized proteins in the network using protein neighbourhood annotations.

The level 1 and 2 neighbours' functional label occurrence patterns are used to identify the key principles driving the functions imposed on a protein by its neighbours, referred to as “traces” of the underlying biological organization of the system. Depending on the features of the protein under consideration obtained from its direct and level-2 interacting partners, the optimal strategy, which consists of finding the best use of ‘traces’ of underlying biological principles, is applied to predict the functions of the protein more accurately. Throughout this annotation prediction process, instead of using exact matches only, relationships between GO terms in the GO DAG structure are considered through the GO term's semantic similarity. This data-driven prediction model should undoubtedly improve the prediction quality and the genome coverage for any organism.

3. Use of networks for evolutionary studies

Bacterial pathogens often evolve from non-pathogenic species through the gain or loss of genes. More often than not this occurs through the gain of pathogenicity islands from horizontal gene transfer [1] or gene duplication. When genes are lost, their protein–protein interactions are also lost and thus certain adaptations to the biological networks need to be made. By the same token, when new genes are gained by either gene duplication or horizontal gene transfer, they need to adapt to fit into the existing set of protein–protein interactions. Thus rewiring of networks occurs during evolution to adapt to changes in the gene repertoire, and we can use this to study the evolution of pathogens.

Comparing PPI networks between closely related organisms can help us to identify how network rewiring occurs, although previous studies have shown that PPIs tend to change at a slower rate than protein sequence evolution [36]. Network comparison is not straightforward, however, as comparing graphs and sub-graphs computationally requires analysis of topologies and content of sub-graphs. In addition, two networks with similar degree distributions can differ substantially in their topological structure [37]. Nevertheless, several methods have been developed to compare networks, an example being PathBLAST [38], which is a tool for alignment of interaction networks across species. PathBLAST includes networks for six target organisms that can be compared. Shou et al. [39] proposed a different method for studying rewiring of biological networks to see how they evolved. Since we are interested in mycobacteria, which are not available in PathBLAST, we illustrate the use of biological network comparison for evolutionary analysis through an example of networks from three mycobacterial species with varying genome sizes using an adaptation of a method proposed by [39].

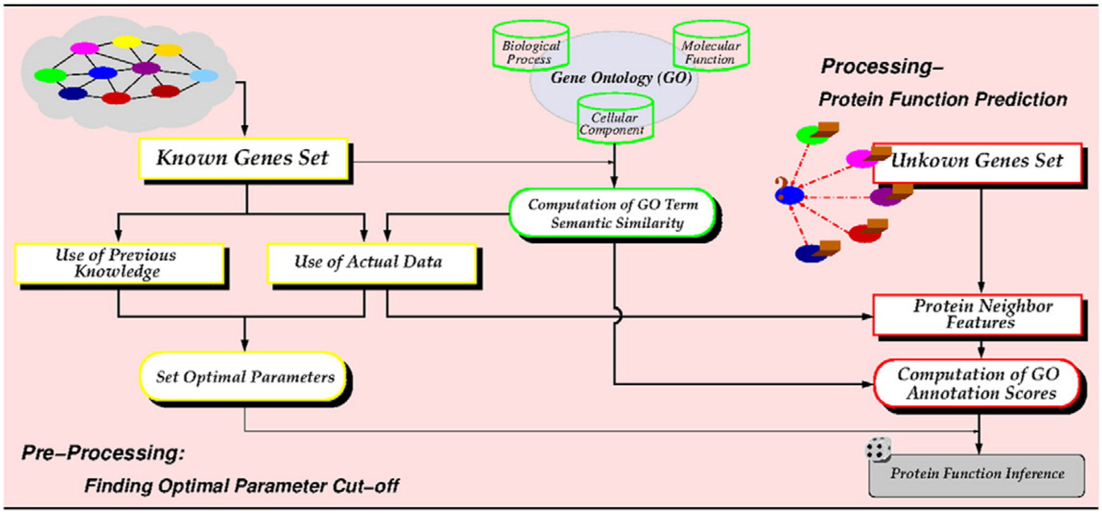
Biological Process
Molecular Function
Gene Ontology (GO)
Cellular Component
Processing–
Protein Function Prediction
Known Genes Set
Unkown Genes Set
?
Computation of GO Term Semantic Similarity
Use of Previous Knowledge
Use of Actual Data
Protein Neighbor Features
Set Optimal Parameters
Computation of GO Annotation Scores
Pre–Processing:
Finding Optimal Parameter Cut–off
Protein Function Inference

Fig. 2. Protein function prediction system. Protein–protein interaction network and semantic similarity scores between terms annotating known proteins are used to determine optimal cut-off scores.

3.1. Functional interaction networks for the three mycobacteria

Functional interaction networks were generated for *M. tuberculosis*, strain CDC1551 (MTB), *Mycobacterium leprae* (MLP) and *Mycobacterium smegmatis* (MSM) as described in [40]. Data used included protein–protein functional interaction data from STRING [41,42] (Search Tool for the Retrieval of Interacting Genes/Proteins), a database of predicted and known PPIs derived from conserved genomic neighbourhood, gene fusion, imports from database (knowledge), phylogenetic co-occurrence, high-throughput experiments and text mining. Additional interactions were predicted from sequence similarity and signatures (shared domains), microarray data (co-expression), and interologs predicted using data from public PPI databases [43–45]. Interologs are based on the premise that orthologs of interacting proteins should themselves interact. Ortholog data was extracted from Ensembl genomes, which uses Compara for ortholog prediction. For all three mycobacterial functional interaction networks, a total score was calculated for each interaction, or edge, using the combined scores of the different evidence types [40], and only interactions with medium and high confidence scores were used for the analysis. Table 2 summarizes important structural properties of the three networks for comparison. The numbers of proteins in the MLP and MSM networks are 1412 and 4953, while the numbers of protein–protein functional interactions are 20,742 and 66,543 respectively. Therefore, the numbers of proteins and edges in MSM are roughly three times that of MLP, though they share a common ancestor. MLP is a smaller genome that has lost genes through reductive evolution [46]. There are 201 structural hubs (proteins that connect sub-networks and are thus able to disconnect the network if removed), in the MTB network, while MLP and MSM have 103 and 755 hubs respectively. The high number of hubs in the MSM network may simply be a reflection of the larger genome and network. The average path length is computed by finding the mean over all shortest paths between all pairs of proteins in the network. While the MLP network has an average shortest path length of approximately 3, the MTB and MSM networks have an average shortest path length of approximately 4. These values give an indication of information spread in their respective networks, all of which exhibit a 'small world property', as their shortest path lengths are a function of the log of the number of nodes [47,48].

Since the three networks are of different sizes, we need the following definition of the clustering and average clustering coefficients which does not depend on the size of the network to compare how ortholog nodes cluster in the respective networks [49]. Let p be a node with n_p neighbours. The total number of possible edges between p's neighbours is $\frac{n_\mathrm{p}(n_\mathrm{p}-1)}{2}$ (i.e., when every neighbour of p is linked with every of its other neighbours). Thus, the clustering coefficient of p is the ratio of the actual number of edges a_p between p's neighbours to the total number of possible edges. Hence, for undirected networks, the clustering coefficient of a node p is defined as [49]

$$C_\mathrm{p} = \frac{2a_\mathrm{p}}{n_\mathrm{p}\left(n_\mathrm{p}-1\right)}.$$

The clustering coefficient of a node is between 0 and 1. A value of zero means that there is no clustering and one signifies maximal clustering. A high clustering coefficient indicates that neighbours of a node are likely to interact with each other. The average clustering coefficient describes the overall ability of nodes in a network to form clusters. It also depends on the number of nodes and edges in the network and it is defined as [50]

$$C = \frac{1}{n}\sum C_\mathrm{p}.$$

3.2. Pairwise network comparison of important proteins

Given any two networks, the approach used by Shou et al. [39] in measuring the evolutionary rewiring rate of biological networks was to name one as the reference network and the other as the compared network. Firstly, all orthologous nodes from both networks were determined, and then we identified three sets of nodes: Common nodes (CN), lost nodes (LN) and gained nodes (GN). Common nodes are those that have orthologous counterparts in both networks, lost nodes are those present in the reference network but absent from the compared network, and gained nodes are those present in the compared network that do not have orthologous counterparts in the reference network. Three types of edges were distinguished as: gained edges from gained nodes, lost edges from lost nodes, and common edges from common nodes. Network identity was defined as the ratio of the number of common edges between orthologous nodes present in both networks to the total number of edges in both networks times 100% [39].

First we compared important proteins in the MTB and MLP networks using the approaches described in [39] and [40], since these two organisms are closely related, despite the large difference in genome size. Important proteins were those possessing certain topological properties such as having high betweenness, closeness and eigenvector centralities, which make them important in the functionality of the network [50]. A protein in a functional network belongs to the gravity centre, if its closeness centrality measure is strictly greater than the reciprocal of the average shortest path length. This value corresponds to 1/3.62739 or 0.27568 in the MTB and 1/3.16955 or 0.31550 in the MLP network. In using the betweenness centrality to determine important proteins, we considered those proteins in which their betweenness is greater than the total number of shortest paths, obtained by multiplying the average shortest path length by the total number of proteins in the functional network. We then combined these criteria with the requirement that the eigenvector centrality should be greater than 10^{-5}.

We obtained a set of 355 and 116 proteins which have a high centre of gravity and thus may be potentially interesting as drug targets in the MTB and MLP networks respectively, and compared their functional classes obtained from Tuberculist (http://genolist.pasteur.fr/Tuberculist, accessed: 28 October, 2011) & Leproma (http://genolist.pasteur.fr/Leproma, accessed: 28 September, 2012). Proteins belonging to the intermediary metabolism and respiration functional class are the

Table 2
Comparing network properties in the MTB, MLP and MSM networks.

Parameters	*Mycobacterium tuberculosis*	*Mycobacterium leprae*	*Mycobacterium smegmatis*
Number of proteins (nodes)	4136	1412	4953
Number of functional interactions (edges)	59,919	20,742	66,543
Number of hubs	201	103	755
Density	0.007	0.0208	0.0054
Average degree	28	29	26
Average shortest path length	3.62739	3.16955	4.2224
Number of connected components	23	19	166
% of nodes in largest component	98.7%	97.5%	91.7%

most represented in these lists of potential drug targets, followed by the unknown classes for MTB and information pathways for MLP. Among these potential drug targets, we extracted those proteins with high closeness which are classified as central proteins, and influential proteins, which are those with high eigenvector centralities. 241 and 69 are the central and influential targets respectively in the MTB network, while 95 and 37 are the central and influential targets respectively in MLP. Due to the unavailability of curated functional classes for MSM at the time of writing, we were unable to make the same kind of comparison as with MLP. However, we identified 294 potential drug targets in the MSM network and of these, 184 were central target proteins and 16 were influential target proteins.

We used the technique described in [39] to identify a total of 2859 proteins in the MTB network without a corresponding ortholog in the MLP network and 135 proteins in the MLP network without corresponding orthologs in the MTB network. In total, 1277 proteins have orthologous counterparts in both networks as shown in Table 3. Similarly, 2148 distinct proteins belong to the MTB network without corresponding orthologs in the MSM network, 2965 proteins in the MSM network have no corresponding orthologs in the MTB network, and 1988 proteins have orthologous counterparts in both networks. Out of the 1412 proteins in the MLP proteome, only 342 have no orthologous counterpart in the MSM network, and 3883 proteins in MSM have no corresponding ortholog in MLP. Table 3 includes the total number of common edges to the two organisms being compared. A common edge is an edge in which both protein pairs are corresponding orthologs in both networks and are interologs. There are 3693 functional interactions or edges common to the MTB and MLP networks, 2284 edges are common to the MSM and MTB networks, while 1901 edges are common to MLP and MSM. We found a total of 1001 proteins which have orthologous counterparts in all three organisms, and these networks all share 297 common edges. Based on the classification of proteins as drug, central and influential targets and using the 1001 orthologs in their intersection, we found eight drug targets and one influential target overlapping among the three organisms.

3.3. Evolutionary differences between the three mycobacterial species

We compared the three networks using orthologs [39] and following a three-way approach: slow grower MTB versus slow grower MLP, fast grower MSM versus MLP, and MTB versus MSM [51]. From the original networks, we removed proteins and functional interactions involving proteins that were not among the 1001 orthologs shared by all three organisms to produce three sub-networks each consisting of 1001 proteins. We then determined the number of shared edges for the orthologs and used this to calculate network identities for the three sub-networks (Table 4). The results in Tables 3 and 4 show that the MLP and MTB sub-networks are more similar than the MTB versus MSM and MLP versus MSM sub-networks, which makes sense as they are more closely related and both are slow growers [51].

We computed the clustering coefficients of each of the networks for the 1001 common orthologous proteins [39]. The boxplot in Fig. 3 shows that the average clustering coefficients are 0.4257, 0.4758 and 0.3292 for MTB, MLP and MSM sub-networks, respectively. By the definition of the average clustering coefficient, this means that the MLP sub-network nodes are most likely to cluster together followed by MTB and then the MSM sub-network. This confirms that the MTB and MLP sub-networks are more similar to each other than the MSM network [40].

This example illustrates how we can compare biological networks between closely related organisms and how protein–protein networks can be used for evolutionary studies. The shared edges, i.e. interactions that are conserved between all 3 species may form a core of proteins and interactions required by all organisms, and central proteins which overlap between the two pathogenic strains would be potentially interesting drug targets. Since *M. leprae* is a highly reduced genome, one can assume that many of the remaining proteins must be reasonably essential for infection and survival. A deeper look into the differences between the full networks in these three mycobacterial species may shed light on how the networks rewire as genes are gained or lost over evolution.

Table 3
Number of ortholog proteins shared, common edges and network identity of the compared networks.

A	B	# of proteins in A only	# of proteins in B only	# of common proteins	# of common edges	Network identity
MLP	MTB	135	2859	1277	3693	4.5%
MLP	MSM	342	3883	1070	1901	2.1%
MSM	MTB	2965	2148	1988	2284	1.5%

Table 4
Number of common nodes, edges and network identity of the compared sub-networks containing only orthologous proteins.

A	B	# of edges in A only	# of edges in B only	# of common proteins	# of common edges	Network identity
MLP	MTB	13,670	9941	1001	2820	11.9%
MLP	MSM	13,670	5086	1001	1849	9.8%
MSM	MTB	5086	9941	1001	656	4.3%

4. Prediction and use of host–pathogen interaction networks

Above we have demonstrated several uses of the study of biological networks for individual organisms and comparisons between them. PPI networks can also be used for improving our understanding of the interplay between the pathogen and its host during infection. Infection and disease progression occur as a result of the interaction, usually via specific protein–protein interactions, between host and pathogen proteins. Experimental detection of host–pathogen interactions has thus far been limited, evidenced by the limited data available in the literature and public interaction databases on experimentally verified interactions. Some host–pathogen databases are available for certain interaction data, e.g. PHI-base (Pathogen Host Interactions: http://www.phi-base.org/) and PATRIC (Pathosystems Resource Integration Center: http://patricbrc.org/portal/portal/patric/HPITool), but again they do not cover all organisms and are not comprehensive. Therefore,

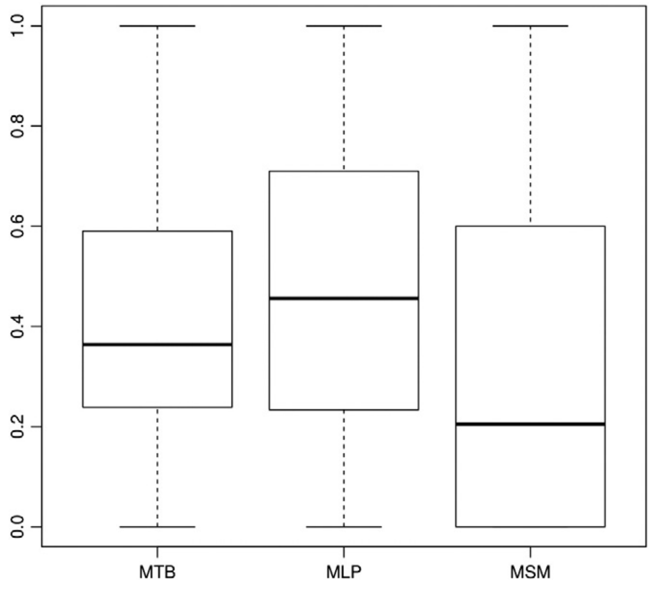

Fig. 3. Boxplot of the clustering coefficients of the 1001 proteins of the three sub-networks.

many computational biologists have turned to the computational prediction of these interactions. Prediction of host–pathogen protein–protein interaction networks can be broadly classified into three groups depending on the methods used to predict the interactions. The first group consists of methods using interologs, the second uses protein domains and structure, and the last group uses machine learning.

4.1. Prediction using interologs

Interologs, which are conserved interactions between a pair of proteins which have interacting orthologs in another organism, have been widely applied to predict inter-species interactions. Starting from an intra-species (or inter-species) protein interaction AB (also referred to as a template interaction), the orthologs A′ and B′ of A and B in the species of interest are determined and an interaction between A′ and B′ is inferred. Interactions between *Homo sapiens* and various pathogens such as *Plasmodium falciparum* (PF) and *M. tuberculosis* have been predicted using this method [52,53]. Although the interolog methods yield a large number of interactions, most of them are not likely to take place in vivo. Therefore the interactions predicted by interologs need to be filtered to obtain plausible interactions. For instance, Lee et al. [52] applied two filters to the human–PF predicted interactions. In the first filter, human proteins annotated with GO cellular component making their interactions with PF proteins impossible were removed, and in the second, only PF proteins having the specific translocation signal RELXE/Q required to translocate the proteins into erythrocyte cytoplasm, at the N-terminal were retained. Wuchty [54] applied a random forest algorithm to assess the quality of the interactions based on the sequence composition of the proteins. The interactions were further filtered based on molecular characteristics of the PF proteins that allow interactions with the human host and localization of the protein pair which are more probable for an interaction to take place. Rapanoel et al. [53] applied various filters to the human–MTB interactions. Using experimentally verified interactions, only predicted interactions which are neighbours of the known ones in the PPI network were retained, and interactions where both proteins were differentially expressed in microarray data relevant to the infection process were prioritized.

4.2. Prediction using protein domain and structure

Protein domains determine the structure and function of proteins. Interactions between two proteins are mediated by these domains; this idea has been used to predict PPIs. For instance, Dyer et al. [55] used protein domain profiles combined with Bayesian statistics to predict interactions between human and *P. falciparum*. Starting with intra-species PPI datasets, they use Bayesian statistics to first calculate the probability that two proteins containing two functional domains interact. Then, the domains in all human and PF proteins were found using InterProScan [56] and the previously computed probability was used to compute the probability of every human–PF protein pair to interact. A cut-off of 0.5 was used to decide whether two proteins interact or not.

Another hypothesis used in PPI prediction is that a pair of proteins having structures similar to known interacting protein pairs is likely to interact. Davis et al. [57] used sequence and structural similarities to predict interactions between human and 10 different pathogens. The protein structure of the host and parasite proteins was modelled using MODPIPE [58] and pairs of human–pathogen proteins with similarity to known interactions from PIBASE [59] were then identified. Sequence similarity was used in the absence of structure. In this case, two human and pathogen proteins interact if each of the two proteins has at least 50% sequence similarity to components of binary protein interactions with joint sequence identity >80% from IntAct [45], where the joint sequence identity is the geometric mean of individual percent identities [60]. Biological filters such as expression in tissues targeted by the pathogen and expression on the cell surface were further applied. Predicted interactions based on templates used in more than 1% of the total predictions were also removed, since they lack specificity.

Protein structural similarity has also been used by Doolittle and Gomez to predict interactions between HIV and human [61], and between dengue virus (DENV) and its two hosts, *H. sapiens* and the mosquito vector *Aedes aegypti*[62]. In this method, human proteins structurally similar to HIV (or DENV) proteins were identified using the Dali Database [63]. The interactions of these HIV-similar (or DENV-similar) proteins with other human proteins, referred to as target proteins were then identified from the Human Protein Reference Database (HPRD: http://www.hprd.org/). After refinement, a physical interaction was predicted between the HIV protein and the human target protein. Interaction prediction between DENV and *A. aegypti* followed a similar procedure, except that *Drosophila melanogaster* was first used to find the DENV-similar and target proteins. The orthologs of the *D. melanogaster* target proteins in the real host *A. aegypti* were then used.

Evans et al. [64] used sequence motifs to predict HIV-1 and human protein interactions. In this case, the two previous methods by Dyer et al. and Davis et al. could not be applied because HIV-1 proteins have few domains and their structures are hard to find. HIV-1 proteins were annotated with short eukaryotic linear motifs (ELMs) using the ELM resource, and counter domains (CDs) or proteins known to interact with these ELMs were retrieved. The CDs were then mapped to PROSITE domains. In addition, PROSITE domains and ELMs on human protein sequences from HPRD were also determined, and a host protein was predicted to interact with a virus protein if the former binds to or competes with the latter. More precisely, a host protein P having a CD known to interact with the ELM of a virus protein can bind to it, whereas a host protein having an ELM similar to a virus protein competes with it for interaction with P.

4.3. Prediction using machine learning

Tastan et al. [65] trained a random forest classifier to classify an HIV-1–human protein pair as interacting or non-interacting. The features for the classification are based on human protein features such as expression during or presence in tissues susceptible to HIV-1 infection. Graph properties of the human interactome, such as degree, clustering coefficient and betweenness centrality were also used as features. They also looked at some properties of HIV-1/human protein pairs such as GO and sequence similarity. Qi et al. [66], on the other hand, used multi-task learning for predicting interactions between HIV-1 and human proteins. Having two reference sets, labelled and partially labelled, they train a supervised classification using the labelled data and the partially labelled data was then used to improve the supervised classification in a semi-supervised auxiliary task.

Kshirsagar et al. [67] applied a multi-task pathway-based learning method to computationally predict interactions between human and four pathogens, namely *Yersinia pestis*, *Francisella tularensis*, *Salmonella* and *Bacillus anthracis*. In this case, a task was the set of host–pathogen proteins involved in one disease. They integrated interactions from several tasks by using the task regularization framework and modifying the regularization term to encode the biological hypothesis, which is that the bacterial species will target the same biological pathway in their human host. An advantage of the machine-learning approach is its ability to integrate various biological information resources in a statistical learning framework. However, a machine-learning approach requires a large amount of training data, which is not available for many organisms. Furthermore, defining the features is not straightforward and sometimes does not carry any biological meaning.

4.4. Assessment and use of host–pathogen PPIs

Host–pathogen interactions are important for understanding disease mechanisms and developing new drugs, and several computational methods have been used to predict them. However, the sparseness of known host–pathogen interactions makes it hard to assess the various methods. For instance, Davis et al. [57] found only 33 host–pathogen protein interactions in the literature for 10 pathogen species and 47 human–MTB interactions were retrieved from the literature by Rapanoel et al. [53]. In assessing their methods, there was little to no overlap with known interactions. The human–HIV protein interactions are probably the best documented interactions with the existence of the HIV-1 Human Interaction Database. Indirect methods have therefore been used to assess host–pathogen predicted PPIs. The most commonly used is enrichment in GO terms and KEGG pathways pertaining to the infection [52–54,64]. Predicted proteins have also been compared to genes expressed during infection to determine the overlap between these sets [57].

Host–pathogen PPIs may shed light on how the pathogen attacks the host and how the host responds to these attacks. Human–MTB predicted interactions, for example, have been used to filter potential drug targets in MTB [68]. An initial list of 881 potential protein drug targets was first identified in the MTB network [69] using network centrality measures. The reasoning behind the method is that the MTB network exhibits “scale-free” and “small world” properties, making the system vulnerable against targeted attack and the network navigability easy, independently of the size of the network. In such a system, a few proteins play critical roles and are essential for the survival of the system. Therefore, proteins with high centrality measures (betweenness, closeness, etc.) may be considered to be potential drug targets. These proteins were then overlaid onto a human–MTB PPI network to filter out those which have direct interactions with human proteins predicted by sequence similarity, and those predicted by interologs which have paralogs or are not essential for MTB [68]. Among the predicted drug targets with no direct interactions with human proteins, those which were direct neighbours of the predicted drug target directly interacting with human proteins and those having paralogs were also filtered out. The final list contained 67 drug targets which include previously identified targets from other sources.

5. Summary and outlook

Biological networks describe functional interactions between genes or proteins within an organism or between organisms, and can include connections that do not necessarily require physical binding. These networks provide a means for studying an organism at the system level and identifying potentially important proteins through their network properties, as well as for finding interesting modules or subnetworks. Biological networks tend to be modular in structure, have a small world property with few average path lengths, and demonstrate scale-free topologies following power-law degree distributions. This makes them robust in the face of perturbations [37]. Networks can therefore help to improve our understanding of biological systems and communication flow within them. Above we have demonstrated some of the uses for biological networks, but others include the use of networks in analysing gene lists from high-throughput biology, and in post-genome-wide association study analysis, among other applications. An understanding of the biological organization of an organism from its PPI network can play a crucial role in vaccine or drug target discovery by highlighting important proteins. For example, network centrality measures can be used to locate central proteins that play important roles in the biological processes and molecular functions of the organism, and in silico knock-out studies can predict the impact of targeting a protein. In some of the above examples we demonstrate how potential drug targets are selected in MTB in using network properties and refined using the MTB–human interaction network. Increasing the number of organisms studied, for example by including two additional mycobacteria, one pathogenic and one non-pathogenic to humans can facilitate further refining of potential drug targets.

Although there are many uses and applications of biological networks, the field is still developing and has a number of challenges. The first challenge is the quality and quantity of data used to generate the networks. Functional interaction networks integrate a wide variety of high-throughput data which is often intrinsically noisy. Does the integration of this data lead to even noisier data or does it help to increase our confidence in interactions that are supported by multiple potentially noisy data sources? It is important that networks are assessed to determine their likely accuracy. For a single organism PPI, this can be done using e.g. GO annotations, if we assume that proteins involved in the same biological process and located in the same subcellular location are more likely to interact. This, of course, relies on the assumption being correct, and the availability and quality of GO annotations. Assessment of host–pathogen PPIs is also difficult due to the scarcity of known interactions to compare and evaluate predictions. In some cases experimentally derived PPIs from public databases can be used for assessment, but even these should be used with caution if they were generated from high-throughput yeast two hybrid experiments, for example.

The applications of biological networks also have their challenges. Protein function prediction is one application for which a number of algorithms have been developed. In our example of three mycobacteria, the fact that many of the central or important proteins in the networks were from the unknown functional class supports the need for function prediction tools to try to determine the roll of these proteins as they may be important for pathogenesis. Since traditional sequence similarity methods failed to predict functions for these proteins we had to turn to using biological networks. However, recent studies have shown that while in some interactions the proteins share functions or biological processes, this is not necessarily carried throughout the network [70].

Evolutionary studies on networks require the comparison between two or more networks, which requires identification of orthologs and comparison of network topologies around these orthologs. Again, previous studies have demonstrated that unless there is strong sequence conservation, protein–protein interactions are not necessarily very well conserved [71]. Therefore, although biological networks are often used for different applications, and studies have generated interesting insights into biological systems from these, the results should be considered with caution, and new developments need to be made in this field to increase our confidence in both the predicted interactions and the applications thereof.

Acknowledgements

Some of the work presented in this review was funded by the National Research Foundation in South Africa (Grant No: 86934), a Claude Leon Foundation Bursary fellowship and a DAAD bursary.

References

[1] Wang J, Behr MA. Building a better bacillus: the emergence of *Mycobacterium tuberculosis*. Front Microbiol 2014;5:139.

[2] Alland D, Whittam TS, Murray MB, Cave MD, Hazbon MH, et al. Modeling bacterial evolution with comparative-genome-based marker systems: application to *Mycobacterium tuberculosis* evolution and pathogenesis. J Bacteriol 2003;185(11):3392–9.

[3] Magadum S, Banerjee U, Murugan P, Gangapur D, Ravikesavan R. Gene duplication as a major force in evolution. J Genet 2013;92(1):155–61.

[4] Merhej V, Georgiades K, Raoult D. Postgenomic analysis of bacterial pathogens repertoire reveals genome reduction rather than virulence factors. Brief Funct Genomics 2013;12(4):291–304.

[5] Yellaboina S, Goyal K, Mande SC. Inferring genome-wide functional linkages in *E. coli* by combining improved genome context methods: comparison with high-throughput experimental data. Genome Res 2007;17:527–35.

[6] Harrington ED, Jensen LJ, Bork P. Predicting biological networks from genomic data. FEBS Lett 2008;582:1251–8.

[7] Browne F, Wang H, Zheng H, Azuaje F. GRIP: a web-based system for constructing gold standard datasets for protein–protein interaction prediction. Source Code Biol Med 2009;4:2.

[8] Mazandu GK, Mulder NJ. A topology-based metric for measuring term similarity in the gene ontology. Adv Bioinf 2012;2012:975783.

[9] Mazandu GK, Mulder NJ. Information content-based gene ontology semantic similarity approaches: toward a unified framework theory. Biomed Res Int 2013; 2013:292063.

[10] GO-Consortium. Gene ontology: tool for the unification of biology. Nat Genet 2000; 25(1):25–9.

[11] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. A basic local alignment search tool. J Mol Biol 1990;215(3):403–10.

[12] Eisen JA. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res 1998;8:163–7.

[13] Bork P, Koonin EV. Predicting functions from protein sequences—where are the bottlenecks? Nat Genet 1998;18:313–8.

[14] Galperin MY, Koonin EV. Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. Silico Biol 1998;1(1):55–67.

[15] Galperin MY, Koonin EV. Who's your neighbor? New computational approaches for functional genomics. Nat Biotechnol 2000;18:609–13.

[16] Mazandu GK, Mulder NJ. Using the underlying biological organization of the *Mycobacterium tuberculosis* functional network for protein function prediction. Infect Genet Evol 2012;12(5):922–32.

[17] Mazandu GK, Mulder NJ. DaGO-Fun: tool for Gene Ontology-based functional analysis using term information content measures. BMC Bioinf 2013;14:284.

[18] Mazandu GK, Mulder NJ. Scoring protein relationships in functional interaction networks predicted from sequence data. PLoS One 2011;6(4):e18607.

[19] Vazquez A, Flammini A, Maritan A, Vespignani A. Global protein function prediction from protein–protein interaction networks. Nat Biotechnol 2003;21(6):697–700.

[20] Tsuda K, Shin H, Scholkopf B. Fast protein classification with multiple networks. Bioinformatics 2005;21:ii59–65.

[21] Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics 2005;21(1):i302–10.

[22] Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). Proc Natl Acad Sci U S A 2003;100(14):8348–53.

[23] Deng M, Chen T, Sun F. An integrated probabilistic model for functional prediction of proteins. J Comput Biol 2004;11(2–3):463–75.

[24] Letovsky S, Kasif S. Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics 2003;19(1):i197–204.

[25] Cho YR, Shi L, Ramanathan M, Zhang A. A probabilistic framework to predict protein function from interaction data integrated with semantic knowledge. BMC Bioinf 2008;9:382.

[26] Lanckriet GRG, Deng M, Cristianini N, Jordan MI, Noble WS. Kernel-based data fusion and its application to protein function prediction in yeast. Pac Symp Biocomput 2004;9:300–11.

[27] Chen Y, Xu D. Global protein function annotation through mining genome-scale data in yeast *Saccharomyces cerevisiae*. Nucleic Acids Res 2004;32(21):6414–24.

[28] Xiong J, Rayner S, Luo K, Li Y, Chen S. Genome wide prediction of protein function via a generic knowledge discovery approach based on evidence integration. BMC Bioinf 2006;7:268.

[29] Xiong W, Liu H, Guan J, Zhou S. Protein function prediction by collective classification with explicit and implicit edges in protein–protein interaction networks. BMC Bioinf 2013;14(Suppl. 12):S4.

[30] Bernardes JS, Pedreira CE. A review of protein function prediction under machine learning perspective. Recent Pat Biotechnol 2013;7:122–41.

[31] Schwikowski B, Uetz P, Fields S. A network of protein–protein interactions in yeast. Nat Biotechnol 2000;18(12):1257–61.

[32] Murali TM, Wu CJ, Kasif S. The art of gene function prediction. Nat Biotechnol 2006; 24(12):1474–5.

[33] Jin HJ, Cho HG. Computational method for protein function prediction by constructing protein interaction network dictionary. Int J Pattern Recognit Artif Intell 2006;20(2): 285–95.

[34] Chua HN, Sung WK, Wong L. Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. Bioinformatics 2006; 22:1623–30.

[35] Chua HN, Sung WK, Wong L. Using indirect protein interactions for the prediction of gene ontology functions. BMC Bioinf 2007;8(4):S8.

[36] Qian W, He X, Chan E, Xu H, Zhang J. Measuring the evolutionary rate of protein–protein interaction. Proc Natl Acad Sci 2011;108:8725–30.

[37] Jin Y, Turaev D, Weinmaier T, Rattei T, Makse HA. The evolutionary dynamics of protein–protein interaction networks inferred from the reconstruction of ancient networks. PLoS One 2013;8(3):e58134.

[38] Kelley BP, Yuan B, Lewitter F, Sharan R, Stockwell BR, Ideker T. PathBLAST: a tool for alignment of protein interaction networks. Nucleic Acids Res 2004;32:W83–8.

[39] Shou C, Bhardwaj N, Lam HYK, Yan K, Kim PM, et al. Measuring the evolutionary rewiring of biological networks. PLOS Comp Biol 2011;7:1–14.

[40] Akinola RO, Mazandu GK, Mulder NJ. A systems level comparison of *Mycobacterium tuberculosis*, *Mycobacterium leprae* and *Mycobacterium smegmatis* based on functional interaction network analysis. J Bacteriol Parasitol 2013;4(4):173.

[41] von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, et al. STRING: known and predicted protein–protein associations, integrated and transferred across organisms. Nucleic Acids Res 2005;33:D433–7.

[42] Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, et al. STRING 8: a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 2009; 37:D412–6.

[43] Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, et al. MINT, the molecular interaction database: 2012 update. Nucleic Acids Res 2011:D857–61.

[44] Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, et al. DIP: the database of interacting proteins. Nucleic Acids Res 2000;28(1):289–91.

[45] Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, et al. The MIntAct project — IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res 2013;42:D358–63.

[46] Singh P, Cole ST. *Mycobacterium leprae*: genes, pseudogenes and genetic diversity. Future Microbiol 2011;6(1):57–71.

[47] Gursoy A, Keskin O, Nussinov R. Topological properties of protein interaction networks from a structural perspective. Biochem Soc Trans 2008;36(14):1398–403.

[48] Mason O, Verwoerd M. Graph theory and networks in biology. IET Syst Biol 2007; 1(2):89–119.

[49] Watts DJ, Strogatz SH. Collective dynamics of small world networks. Nature 1998; 393(6684):440–2.

[50] Barabasi A, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet 2004;5(2):101–13.

[51] Stahl DA, Urbance JW. The division between fast- and slow-growing species corresponds to natural relationships among the mycobacteria. J Bacteriol 1990;172(1): 116–24.

[52] Lee SA, Chan CH, Tsai CH, Lai JM, Wang FS, et al. Ortholog-based protein–protein interaction prediction and its application to inter-species interactions. BMC Bioinf 2008;9(Suppl. 12):S11.

[53] Rapanoel HA, Mazandu GK, Mulder NJ. Predicting and analyzing interactions between *Mycobacterium tuberculosis* and its human host. PLoS One 2013;8:e67472.

[54] Wuchty S. Computational prediction of host–parasite protein interactions between *P. falciparum* and *H. sapiens*. PLoS One 2011;6:e26960.

[55] Dyer MD, Murali TM, Sobral BW. Computational prediction of host–pathogen protein–protein interactions. Bioinformatics 2007;23:i159–66.

[56] Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, et al. InterProScan: protein domains identifier. Nucleic Acids Res 2005;33:W116–20.

[57] Davis F, Barkan D, Eswar N, McKerrow JH, Sali A. Host–pathogen protein interactions predicted by comparative modeling. Protein Sci 2007;16:2585–96.

[58] Eswar N, John B, Mirkovic N, Fiser A, Ilyin VA, et al. Tools for comparative protein structure modeling and analysis. Nucleic Acids Res 2003;31:3375–80.

[59] Davis FP, Sali A. PIBASE: a comprehensive database of structurally defined protein interfaces. Bioinformatics 2005;21:1901–7.

[60] Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, et al. Annotation transfer between genomes: protein–protein interologs and protein–DNA regulogs. Genome Res 2004;14(6):1107–18.

[61] Doolittle JM, Gomez SM. Structural similarity-based predictions of protein interactions between HIV-1 and *Homo sapiens*. Virol J 2010;7:82.

[62] Doolittle JM, Gomez SM. Mapping protein interactions between dengue virus and its human and insect hosts. PLoS Negl Trop Dis 2011;5:e954.

[63] Holm L, Kaariainen S, Rosenstrom P, Schenkel A. Searching protein structure databases with DaliLite v. 3. Bioinformatics 2008;24:2780–1.

[64] Evans P, Dampier W, Ungar L, Tozeren A. Prediction of HIV-1 virus–host protein interactions using virus and host sequence motifs. BMC Med Genomics 2009;2:27.

[65] Tastan O, Qi Y, Carbonell JG, Klein-Seetharaman J. Prediction of interactions between HIV-1 and human proteins by information integration. Pac Symp Biocomput 2009; 527:516–27.

[66] Qi Y, Tastan O, Carbonell JG, Klein-Seetharaman J, Weston J. Semi-supervised multi-task learning for predicting interactions between HIV-1 and human proteins. Bioinformatics 2010;26:i645–52.

[67] Kshirsagar M, Carbonell J, Klein-Seetharaman J. Multitask learning for host–pathogen protein interactions. Bioinformatics 2013;29:i217–26.

[68] Mazandu GK, Rapanoel HA, Mulder NJ. Using host–pathogen functional interactions for filtering potential drug targets in *Mycobacterium tuberculosis*. J Mycobact Dis 2013;3:126.

[69] Mazandu GK, Mulder NJ. Generation and analysis of large-scale data-driven *Mycobacterium tuberculosis* functional networks for drug target identification. Adv Bioinf 2011;2011:801478.

[70] Gillis J, Pavlidis P. "Guilt by association" is the exception rather than the rule in gene networks. PLoS Comput Biol 2012;8(3):e1002444.

[71] Lewis ACF, Jones NS, Porter MA, Deane CM. What evidence is there for the homology of protein–protein interactions? PLoS Comput Biol 2012;8(9):e1002645.