Metabolic Engineering 13 (2011) 1–10

Contents lists available at ScienceDirect

Metabolic Engineering

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

Minireview

Network-based function prediction and interactomics: The case for metabolic enzymes

S.C. Janga [a,*], J. Javier Díaz-Mejía [b,c], G. Moreno-Hagelsieb [b]

[a] MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, United Kingdom
[b] Department of Biology, Wilfrid Laurier University, 75 University Avenue West, Walterloo, Ontario, Canada N2L 3C5
[c] Terrence Donnelly Center for Cellular and Biomolecular Research, University of Toronto, 160 College Street, Toronto, Ontario, Canada M5S 3E1

ARTICLE INFO

Article history:
Received 2 June 2010
Received in revised form
15 July 2010
Accepted 16 July 2010
Available online 21 July 2010

Keywords:
Genome annotation
Function prediction
Genomic context
Protein–protein interactions
Data integration
Interactomics

ABSTRACT

As sequencing technologies increase in power, determining the functions of unknown proteins encoded by the DNA sequences so produced becomes a major challenge. Functional annotation is commonly done on the basis of amino-acid sequence similarity alone. Long after sequence similarity becomes undetectable by pair-wise comparison, profile-based identification of homologs can often succeed due to the conservation of position-specific patterns, important for a protein's three dimensional folding and function. Nevertheless, prediction of protein function from homology-driven approaches is not without problems. Homologous proteins might evolve different functions and the power of homology detection has already started to reach its maximum. Computational methods for inferring protein function, which exploit the context of a protein in cellular networks, have come to be built on top of homology-based approaches. These network-based functional inference techniques provide both a first hand hint into a proteins' functional role and offer complementary insights to traditional methods for understanding the function of uncharacterized proteins. Most recent network-based approaches aim to integrate diverse kinds of functional interactions to boost both coverage and confidence level. These techniques not only promise to solve the moonlighting aspect of proteins by annotating proteins with multiple functions, but also increase our understanding on the interplay between different functional classes in a cell. In this article we review the state of the art in network-based function prediction and describe some of the underlying difficulties and successes. Given the volume of high-throughput data that is being reported the time is ripe to employ these network-based approaches, which can be used to unravel the functions of the uncharacterized proteins accumulating in the genomic databases.

1. Introduction

Determining the functions of proteins encoded in genome sequences represents a major challenge in current biology. As of March 2010, the TrEMBL database (http://www.ebi.ac.uk/uniprot/TrEMBLstats/) contained 10,618,387 sequences. The Genomes Online Database (http://www.genomesonline.org/) reported more than 1000 published genomes with over 3700 genome projects underway. The database also reports more than 100 metagenome projects with the Venter's marine microbial communities project (Rusch et al., 2007) alone contributing more than 6,000,000 proteins to the already exploding protein repertoire. While the pace at which sequencing technologies are able to generate the DNA sequence data is increasing, our ability to unravel the functional roles of the encoded proteins therein has been rather limited.

Originally, proteins identified from genome sequencing projects were mostly annotated through homology, inferred with the aid of pair-wise alignment tools such as BLAST (Altschul et al., 1997), followed by manual intervention (Gotoh, 1999; Pearson, 1995; Procter et al., 2010). Researchers would have an idea of the function of a protein by finding a significant sequence similarity to another protein whose function had been experimentally characterized. This homology-based annotation transfer is essentially the most widely used form of computational function prediction. The rationale behind homology-based annotation is that, if two sequences have a high degree of similarity, then they have evolved from a common ancestor, and thus they should have similar, if not identical, functions. However, with increasing number of sequences as well as the effects of gene duplications, which might be followed by divergence of function, the power of homology-based annotation is being challenged. Adding to this is the problem of errors in annotation even in human curated databases, which spread misannotations when homology-based approaches are used. In addition, most of the newly identified proteins do not show significant sequence similarity with experimentally characterized proteins worsening the

Abbreviations: BLAST, Basic local alignment search tool; GC, Genomic context; PPI, Protein–protein interaction; EC, Enzyme commission; GBA, Guilt-by-association

* Corresponding author.
E-mail address: sarath@mrc-lmb.cam.ac.uk (S.C. Janga).

1096-7176/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ymben.2010.07.001

Table 1
Resources currently available for protein function prediction grouped according to the predominant method or approach implemented in them. Note that the list may be incomplete as some resources which are not directly relevant to the methods discussed here might have escaped their mention in this table.

Approach	Resource	Webpage
Sequence similarity based	GOtcha (Martin et al., 2004)	http://www.compbio.dundee.ac.uk/gotcha/gotcha.php
	PFP (Hawkins et al., 2009)	http://dragon.bio.purdue.edu/pfp/
	GOsling (Jones et al., 2008)	https://www.sapac.edu.au/gosling/
	OntoBlast (Zehetner, 2003)	http://functionalgenomics.de/ontogate/
	GOblet (Groth et al., 2004)	http://goblet.molgen.mpg.de
	Blast2GO (Conesa et al., 2005)	http://www.blast2go.de
Phylogenomics based	SIFTER (Engelhardt et al., 2005)	http://sifter.berkeley.edu
	AFAWE (Jocker et al., 2008)	http://bioinfo.mpiz-koeln.mpg.de/afawe/
	RIO (Zmasek and Eddy, 2002)	http://www.rio.wustl.edu/
	OrthoStrapper (Hollich et al., 2002)	http://www.cgb.ki.se/OrthoGUI
Domain/pattern/profile based	InterProScan (Mulder et al., 2008)	http://www.ebi.ac.uk/tools/interproscan/
	Pfam (Finn et al., 2008)	http://pfam.sanger.ac.uk
	SUPERFAMILY (Wilson et al., 2009)	http://supfam.cs.bris.ac.uk/superfamily/
	PROSITE (Sigrist et al.)	http://www.expasy.ch/prosite/
	PRINTS (Attwood et al., 2003)	http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/
	SMART (Letunic et al., 2009)	http://smart.embl-heidelberg.de/
	Gene3D (Lees et al.)	http://gene3d.biochem.ucl.ac.uk/gene3d/
	PANTHER (Mi et al.)	http://www.pantherdb.org/
	TIGRFAMs (Selengut et al., 2007)	http://www.tigr.org/TIGRFAMs/
	SCOP (Andreeva et al., 2008)	http://scop.mrc-lmb.cam.ac.uk/scop/
	CATH (Cuff et al., 2009)	http://www.cathdb.info/
	CatFam (Yu et al., 2009)	http://www.bhsai.org/downloads/catfam.tar.gz
	PIRSF (Nikolskaya et al., 2006)	http://pir.georgetown.edu/pirwww/dbinfo/pirsf.shtml
	PRODOM (Bru et al., 2005)	http://prodom.prabi.fr/prodom/current/html/home.php
	EFICAz (Arakaki et al., 2009)	http://cssb.biology.gatech.edu/skolnick/webservice/EIFICAz2/index.html
	PRIAM (Claudel-Renard et al., 2003)	http://bioinfo.genotoul.fr/priam/REL_JUL06/index_jul06.html
Sequence clustering based	ProtoNet (Kaplan et al., 2005)	http://www.protonet.cs.huji.ac.il/
	CluSTr (Petryszak et al., 2005)	http://www.ebi.ac.uk/clustr/
	eggNOG (Muller et al., 2010)	http://eggnog.embl.de
	COGs (Tatusov et al., 2003)	http://www.ncbi.nlm.nih.gov/COG/
	InParanoid (Berglund et al., 2008)	http://inparanoid.sbc.su.se/cgi-bin/index.cgi
	MultiParanoid (Alexeyenko et al., 2006)	http://multiparanoid.sbe.su.se/index.html
	OrthoMCL (Chen et al., 2006)	http://www.orthomcl.org/cgi-bin/OrthoMclWeb.cgi
Machine learning based	ProtoFun (Jensen et al., 2003)	http://www.cbs.dtu.dk/services/ProtFun/
	GOPET (Vinayagam et al., 2006)	http://genius.embnet.dkfz-heidelberg.de/menu/biounit/open-husar
	SVM-Prot (Cai et al., 2003)	http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi
	ffPred (Lobley et al., 2008)	http://bioinf.cs.ucl.ac.uk/ffpred/
	EzyPred (Shen and Chou, 2007)	http://www.csbio.sjtu.edu.cn/bioinf/EzyPred/
Network based	MCODE (Bader and Hogue, 2003)	http://baderlab.org/Software/MCODE
	MCL (Enright et al., 2002)	http://www.micans.org/mcl/
	SAMBA (Tanay et al., 2004)	http://acgt.cs.tau.ac.il/samba/
	RNSC (King et al., 2004)	King et al. 2004
	PRODISTIN (Brun et al., 2003)	http://crfb.univ-mrs.fr/webdistin/
	Cytoscape (Yeung et al., 2008)	http://www.cytoscape.org/
	STRING (Jensen et al., 2009)	http://string.embl.de/
	VisANT (Hu et al., 2009b)	http://visant.bu.edu/
	VIRGO (Massjouni et al., 2006)	http://whipple.cs.vt.edu/virgo/welcome.cgi

problem for manual curators to keep up with the influx of data (Yooseph et al., 2007). All of these factors have contributed to an increase in a varied number of automated approaches for functional inference (see Table 1) (Godzik et al., 2007; Han et al., 2006; Rentzsch and Orengo, 2009; Zhao et al., 2008a). These automated methods are based on a number of features (Table 1), starting from nucleotide or amino acid sequence, sequence patterns/profiles and protein structure patterns to chromosomal location, phylogenetic information, expression profiles, molecular interaction data, functional associations and gene co-evolution.

2. Overview of network-based function prediction

The very definition of biological function is ambiguous with its exact meaning depending on the context in which it is used and the classification it is based on (Rison et al., 2000; Whisstock and Lesk, 2003). It is obvious that biological function has many aspects associated with it. For instance, the function of a kinase can be described from very broadly as in "enzyme," to very precisely as in "phosphorylation of the hydroxyl group of a specific substrate." A different way to understand the role of a protein within the cell is to ask where exactly it occurs in the cell. This aspect is important especially for entities that can potentially occur in a number of sub-cellular localizations. In this particular case, kinases can be identified either in the cytoplasm or in the nucleus and this information is crucial in gathering their roles and interactions with other proteins within the cellular environment. Likewise, a mutation in the kinase can be associated with a disease phenotype. Therefore, it is clear that when speaking of a protein's function, we must always specify the aspect or aspects of the functional description. In particular, in the process of developing a function prediction tool one must keep in mind which functional aspect or aspects are of interest in the prediction pipeline and use the appropriate vocabulary.

Once functional aspects of a protein are defined, the question is how function can be interpreted in computational terms. For instance, protein sequences for a long time have been represented as character strings that enable their use for many computational tasks, including pair-wise comparisons and multiple sequence alignments, motif searching, database searching and several other tasks, aimed at extracting biological information from the sequence. In contrast to sequence information, until recently the annotation of a protein has been written in human language conveying the complex descriptions and intricacies of its function as well as the experimental evidence supporting it, in terms of custom non-standard formats varying across different groups. As a result, vocabulary went on to be invented and re-invented, with many terms being synonymous. This synonymy not only raises confusion among human curators revising the annotations but also increases the chances of additional errors due to non-standard formats for annotating function. Therefore, over the years a need to convey this information in a more controlled and well-defined fashion has emerged. One of the first groups of people to appreciate this problem were the biochemists who came up with the Enzyme Commission (EC) classification (Tipton, 1994). The EC classifies metabolic reactions in a four-level hierarchy, which are noted by four-position identifiers, going from the most general in the first position to the most specific function of the enzyme in the last position. This classification not only addresses the need for a controlled vocabulary, but also a well-defined hierarchical relationship between terms allowing the comparison between annotations. While enzymes form one of the most commonly occurring protein classes in the cell, they are definitely not the only kind. Thus, EC numbers are not sufficient for annotating all protein functions in a cell. As a solution, Monica Riley and colleagues developed the MultiFunc classification system for *Escherichia coli* in 1993 (Riley, 1993; Serres and Riley, 2000). This attempt was followed by multispecies annotation systems which came later, including the Clusters of Orthologous Groups (COG) (Tatusov et al., 1997), based on manual annotation of a group of orthologous proteins by hierarchically organizing their functional descriptions; Swissprot annotations based on human curation efforts on well-annotated proteins (Apweiler, 2001; Kretschmann et al., 2001); FunCat, a hierarchically structured, scalable, controlled classification system enabling the functional description of proteins in an organism-independent fashion (Ruepp et al., 2004); and, more recently, the Gene Ontology (GO) (Ashburner et al., 2000). The common theme among these schemes is the establishment of a controlled vocabulary and in many cases a categorization that proceeds from the general to the specific. The GO is currently the dominant cross-species approach for machine-legible functional annotation and covers three major aspects of gene product function, namely "molecular function," "biological process" and "cellular component." Each GO is implemented as a directed acyclic graph (DAG) where terms are represented as nodes in the graph and are arranged from the general to the specific. The DAG arrangement means that each node can have more than a single parent enabling the description of functions associated with more than one biological activity or process. Standardizing annotations and defining the relationships between terms using a graph, makes computational analyses easier. For instance, given a GO-annotated genome a researcher can computationally identify the set of all genes with a given annotation and likewise predict functional labels of proteins using such a controlled vocabulary. However, such standardized annotations also limit the flexibility in the amount of detail an annotation can contain.

Having defined function and the means of describing function, one can start discussing function prediction. Predicting functions using network-based approaches, which is the topic of this review, essentially requires two seed components: (a) a network of functional associations amenable for graph theoretical analysis and (b) a network-based function prediction algorithm for predicting functional labels of uncharacterized genes in the graph/network under study. In what follows, we will first discuss different approaches for constructing and integrating functional association networks and then outline currently available computational methods for inferring function based on them. Fig. 1 provides an overview of the different major steps involved in the prediction of function from network context, starting from draft or complete genome sequences.

3. Methods and databases for building functional association networks

Traditionally, the function of a protein has been experimentally identified using a number of low-throughput approaches,

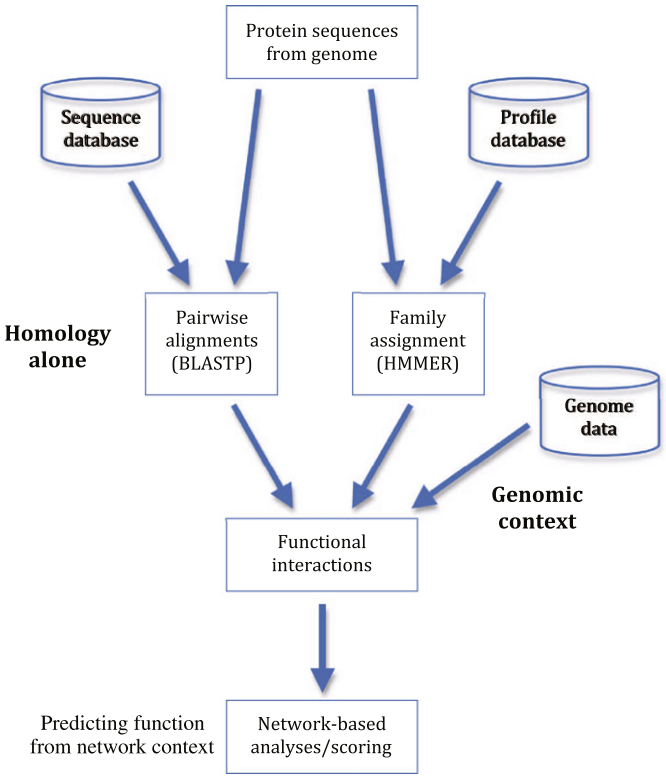

Fig. 1. Overview of the different steps involved in the prediction of function of a gene from genomic and network context. Functional assignments start with protein sequences taken from a genome. Annotation by homology can come from two sources: pairwise comparison and comparison against protein family profiles. If the pairwise alignments match a query protein to a previously functionally characterized protein, the functional annotation might be as simple as assigning the same function to the query protein. Matching to a functionally annotated protein family profile, which can detect homologies beyond pairwise alignments, can provide the same kind of information. However, more particular information might be possible, for example the query protein might contain an ATP-binding domain. Information about homologies, whether to annotated or unannotated proteins, can help determine protein function by complementing the information with presence and organization of homologs in other genomes (genomic context). Genomic context might associate a query protein with proteins working in a common cellular process, such as "translation" or common molecular function such as a particular metabolic pathway. The predicted interaction can be visualized as a network, and its involvement in a cellular function be determined with further scoring using the structure of the network, thus ensuring that the annotation takes into account the number and consistency of functional role of the proteins connected to the query. Flow chart shows the keys steps which comprise of annotating a gene product with function starting from a draft or complete genome sequences, using genomic approaches in the early stages and integrating network-based approaches as the number of data sources increase.

such as mutagenesis of residues or of whole genes, which allowed for the identification of phenotypes to follow up analysis. However, it is clear that this approach is limited in its ability to infer the function of proteins, failing for those exhibiting mild phenotypic effects, or for those not expressed under standard experimental conditions. In addition, since most proteins associate dynamically with a number of other cellular entities during their lifetime, the traditional approach of identifying function of a protein by isolating it from the rest of the cellular machinery can be misleading. This problem and the availability of high-throughput experimentally determined protein–protein interaction maps for diverse model organisms have given rise to the use of large datasets for delineating the biological processes, pathways and complexes where proteins participate (Aranda et al., 2010; Bader et al., 2003; Breitkreutz et al., 2008). Indeed, there is now significant overlap and informative variation between different types of low- and high-throughput experiments (Shoemaker and Panchenko, 2007), which provides convincing reasons for exploiting them as complementary approaches for unraveling the functions of proteins. Accordingly, there has been an explosion in the number of methods and databases providing functional associations (both direct physical and indirect contextual interactions) between proteins using both experimental and computational means (Table 2).

To summarize, experimental approaches employed for constructing functional association networks mostly comprise of data from high-throughput protein–protein interaction screens (Gavin et al., 2006; Krogan et al., 2006; Shoemaker and Panchenko, 2007; Tarassov et al., 2008; Yu et al., 2008), followed by networks built from gene pairs showing significant correlation of expression across conditions, derived from microarray datasets (Luo et al., 2007; Ruan et al., 2010; Wang et al., 2009). More recently, genetic interactions—measuring the fitness defects of double mutants compared to those of individual mutants—are also being employed for constructing these functional linkage networks (Babu et al., 2009; Butland et al., 2008; Costanzo et al., 2010). These high-throughput experimental approaches not only increase the confidence of an association but also give a cellular context to the protein providing a complementary view to the traditional functional prediction paradigm.

In addition to the experimental methods, several computational methods have been proposed for constructing protein–protein associations from sequence data alone. These mainly consist of the so-called Genomic Context (GC) methods; namely gene fusion, gene cluster or gene order conservation, phylogenetic profiles and operon rearrangements. The gene fusion approach tries to detect the fusion of two genes into a single protein coding gene in one of the sequenced genomes and thereby links them as a strong functional association (Enright et al., 1999; Marcotte et al., 1999). The method of gene order conservation aims to identify pairs of genes that consistently show a tendency to cluster in immediate vicinity in a number of genomes, suggesting a strong functional link in prokaryotic genomes, which are abundant in operons (Dandekar et al., 1998; Janga and Moreno-Hagelsieb, 2004; Overbeek et al., 1999). While this method has been abundantly exploited in prokaryotic function prediction, recent and mounting evidence supports the utility of this approach for functional inference in eukaryotes (Davila Lopez et al., 2010; Hurst et al., 2004; Liu and Han, 2009; Pignatelli et al., 2009). The method of operon rearrangement tries to identify a link between any pair of genes on a genome as long as their orthologs are predicted to be organized in an operon with a high confidence in at least one sequenced genome (Janga et al., 2005; Rogozin et al., 2002; Snel et al., 2002). The power of this approach depends on the predictive quality of operon prediction methods which have been shown to reach ~90% accuracy in most sequenced genomes (Brouwer et al., 2008; Moreno-Hagelsieb and Collado-Vides, 2002). Yet another approach not based on genomic proximity is phylogenetic profiles. In this method a vector of presence/absence of a gene across all the analyzed genomes is constructed and compared to identify genes showing correlated profiles, as a measure of functional linkage.

Table 2
Different approaches for generating functional linkage maps or networks. Typically, these networks either independently or integrated versions of them form the input for network-based functional inference algorithms.

Approach	Description	Data sources
Protein–protein interactions	Physical interactions between proteins identified either by mass spectrometry or one of the hybrid approaches are used to generate protein interaction maps on a large-scale which are used as input for function prediction algorithms (Shoemaker and Panchenko, 2007).	**HPRD** (http://www.hprd.org) **IntAct** (http://www.ebi.ac.uk/intact/site/index.jsf) **MINT** (http://cbm.bio.uniroma2.it/mint/index.html) **BioGRID** (http://www.thebiogrid.org) **DIP** (http://dip.doe-mbi.ucla.edu/dip/Main.cgi) **MPPI** (http://mips.gsf.de/proj/ppi) **eNet** (http://ecoli.med.utoronto.ca)
Co-expression networks	In these approaches gene co-expression above a significant correlation threshold is considered as a presence of a functional linkage between genes. Genome-wide inspection of these gene co-expression networks provides an intuitive way to represent complex co-expression patterns between many genes providing functional insights into uncharacterized processes (Aoki et al., 2007; Huber et al., 2007; Lasko, 2000).	**GEO** (http://www.ncbi.nlm.nih.gov/geo) **SMD** (http://genome-www5.stanford.edu) **ArrayExpress** (http://www.ebi.ac.uk/arrayexpress) **caArray** (http://caarraydb.nci.nih.gov/caarray) **M3D** (http://m3d.bu.edu)
Genetic interaction networks	In these approaches interactions between genes are constructed by linking gene pairs which show significantly reduced fitness when both the genes are knocked out compared to when each gene is knocked out independently. These lethality assays are carried out on a high-throughput scale to construct genome-scale interactions (Butland et al., 2008; Costanzo et al., 2010)	**BioGRID** (http://www.thebiogrid.org) **DRYGIN** (http://drygin.ccbr.utoronto.ca) **IM Browser** (http://proteome.wayne.edu/PIMdb.html)
GC networks	These approaches include the gene fusion, gene cluster or gene order conservation, phylogenetic profile and operon rearrangement methods (Dandekar et al., 1998; Enright et al., 1999; Janga et al., 2005; Pellegrini et al., 1999). See text for further discussion.	**STRING** (http://string.embl.de) **ProLinks** (http://prolinks.mbi.ucla.edu) **VisANT** (http://visant.bu.edu) **eNet** (http://ecoli.med.utoronto.ca)
Integration of data sources	These approaches integrate different kinds of functional association data using bayesian or kernel techniques and then construct high-confidence functional linkage networks which are then used for function prediction (Bowers et al., 2004; Chen and Xu, 2004; Hu et al., 2009a; Jensen et al., 2009; Lanckriet et al., 2004; Linghu et al., 2008; Marcotte et al., 1999; Massjouni et al., 2006; Myers et al., 2009; Troyanskaya et al., 2003; Tsuda et al., 2005; Zhao et al., 2008b)	**STRING** (http://string.embl.de) **ProLinks** (http://prolinks.mbi.ucla.edu) **VisANT** (http://visant.bu.edu) **Virgo** (http://whipple.cs.vt.edu:8080/virgo) **eNet** (http://ecoli.med.utoronto.ca)

The rationale is that two proteins showing similar profiles, i.e. coordinated in their evolutionary gain and loss, are expected to be functionally related (Gaasterland and Ragan, 1998; Pellegrini et al., 1999). Modified versions of this approach take into account the phylogenetic signal of the genomes employed and/or the redundancy in the genome sequence information (Barker and Pagel, 2005; Date and Marcotte, 2003; Moreno-Hagelsieb and Janga, 2008).

The integration of different types of interaction data into genome-wide functional linkage maps has gained popularity for functional inference as these maps not only boost coverage, but also confidence, when assessing protein function. One of the first studies demonstrating the power of integrating different types of interaction data was published by Marcotte et al., 1999 who put together diverse kinds of GC methods. This was followed by a number of other methods such as those implemented in the STRING and PROLINKS databases, among other focused studies and implementations (Bowers et al., 2004; Chen and Xu, 2004; Hu et al., 2009a; Jensen et al., 2009; Massjouni et al., 2006; Myers et al., 2009; Troyanskaya et al., 2003). Typically, in these networks edge weights correspond to the integrated interaction probability obtained by first scoring each of the methods independently against a set of gold standard interactions, and then combined with a Bayesian method that assumes the scores of each method to be independent of each other. More complex methods take into account the dependence and correlation between methods to develop a regression model for scoring the integrated interactome (Linghu et al., 2008; Zhao et al., 2008a). Kernel methods form the second group of approaches frequently used for integration of data from different sources (Lanckriet et al., 2004; Tsuda et al., 2005). Nevertheless, all of them boil down to constructing a network with either weighted or unweighted edges, which are then used for propagating annotations to uncharacterized members using approaches discussed in the section below.

4. Computational methods for predicting function from network context

Any set of functional associations, whether experimentally derived or predicted by computational methods, can be depicted as a network of nodes connected by edges, with nodes representing proteins and edges denoting the interactions between such nodes. Most network-based functional inference algorithms work under the premise that the closer the two nodes are in the network the higher is the functional similarity between them (Sharan et al., 2007). Accordingly, most computational approaches for predicting function from networks simply exploit the context of a protein within their local network-neighborhood, analogous to traditional sequence or GC methods. These approaches also generally tend to infer a broader kind of function, such as a biological process, as opposed to the molecular/biochemical function, which is typically inferred by homology-based approaches, making network-based approaches complementary methods for annotating genomes. All of these methods essentially employ machine-learning techniques and can be grouped into two major classes: those using direct network-context and those assisted by module prediction. The former infer function based on connections (direct or indirect) in the network, while the later first identify clusters, or modules, of related proteins and then annotate each protein based on the known functions of the module's members (see Table 3 for a summary of the methods belonging to either class). Since machine learning methods themselves can be classified into supervised and unsupervised techniques, direct methods fall into the supervised class, while module-based methods, which involve clustering of genes to obtain coherent groups, naturally belong to the unsupervised class. To recall, supervised techniques utilize known annotations as training data to first construct a model, and then predict the functions of unknown proteins using the model, while unsupervised methods group proteins together without the need to input any training (gold positive) data. Supervised methods perform best if there is sufficient training data available (Sharan et al., 2007). Otherwise, unsupervised (module-based) methods are regarded as an ultimate choice for function prediction for scarcely annotated datasets.

Table 3
Different methods currently available for network-based function prediction.

Method	Description	References
Direct	In simpler versions of these methods function of a protein is assigned based on the number of annotated protein neighbors in the immediate network neighborhood which are associated with a particular function. Advanced approaches take into account overall network topology and are able to give confidence scores for predictions. Techniques such as flow simulation and graph theoretic based have shown to yield high accuracies on some model systems. Other methods in this category involve the use of probabilistic markov random models.	(Chen and Xu, 2004; Chua et al., 2006; Deng et al., 2003; Hishigaki et al., 2001; Karaoz et al., 2004; Letovsky and Kasif, 2003; Nabieva et al., 2005; Schwikowski et al., 2000; Vazquez et al., 2003)
Module based	In these approaches, two major steps are involved: (1) Identification of modules which are functionally coherent using any clustering technique and (2) predicting function of uncharacterized members in a cluster using any of the direct methods or by computing enrichment for characterized functions in a given module and then transferring the annotations to other members. The first step follows the notion that genes which work in the same biological process should be homogenous in their functional roles and hence plays a crucial role in these methods. So majority of the methods in this category differ in the approach taken to identify modules.	(Altaf-Ul-Amin et al., 2006; Bader and Hogue, 2003; Brun et al., 2003; King et al., 2004; Pereira-Leal et al., 2004; Rives and Galitski, 2003; Samanta and Liang, 2003; Spirin and Mirny, 2003; Troyanskaya et al., 2003)

Among direct methods, the majority rule, or Guilt-by-Association (GBA), method is the simplest, and perhaps the most intuitive (Schwikowski et al., 2000). This method determines the function of a protein based on the known functions of proteins lying in the immediate network neighborhood. Although simple and often effective for dense networks, the method does not take into account the overall topology of the network, nor does it provide a confidence score for the predicted functional label. Therefore, over the years, researchers have produced more sophisticated methods to address these limitations (Chua et al., 2006; Hishigaki et al., 2001). To address the problem of considering the topology beyond the immediate neighbors, Hishigaki et al. defined the neighborhood of a protein with a radius of n. For an unknown protein, the functional enrichment in its n-neighborhood was investigated with a χ^2 test and the top ranking functions were assigned to the

unknown proteins. Chua et al. took a different approach by considering not only the neighborhood of a protein of interest but also the shared neighborhood of a pair of proteins. This allowed them to define a functional similarity between a pair of proteins by taking both the direct and indirect neighbors of the protein pair into account. Other direct methods involve the use of graph theoretical principles such as cuts and flow-simulation in order to take advantage of the global and/or local topology of the network under consideration (Karaoz et al., 2004; Nabieva et al., 2005; Vazquez et al., 2003). In doing so, these methods also aim at maximizing the number of edges (for a protein of interest) which connect to other proteins assigned with the same function. Some authors also employed probabilistic approaches to address the caveats of the original methods and follow the premise that the function of a protein is independent of all other proteins given the functions of its immediate neighbors—thereby leading to the use of Markov random field models for solving the problem of function prediction (Deng et al., 2003; Letovsky and Kasif, 2003; Sharan et al., 2007). Nevertheless, there is convincing evidence from recent studies that functional classes in a cell are not independent of each other and that their inter-relationships should be taken into account for improving function prediction algorithms (Barutcuoglu et al., 2006; Lee et al., 2006; Pandey et al., 2009).

Biological functional systems are thought to be inherently modular, with groups of genes being associated with a particular biological process/pathway (Hartwell et al., 1999). This has resulted in the development of module-based functional inference approaches. In these unsupervised approaches clustering methods identify coherent groups of genes, predicted to work together to achieve a common biological task. Once modules are identified, simple methods like GBA and hypergeometric enrichment, computed for every function associated with the module, are used for transferring annotations to the uncharacterized members. Therefore, in these approaches the initial clustering method employed is crucial in determining the quality of the functional predictions. As a result, different module-assisted methods mainly differ in the module detection technique employed. Module finding algorithms typically depend on the network topology information used as a distance metric. Clustering techniques can identify either a predefined number of clusters, as is the case in *k*-means clustering, or an undefined number as resulting from hierarchical clustering. Some of the module detection techniques can also detect overlapping clusters as a means of revealing the inherent plasticity of function in biological systems. As such, these techniques allow for the annotation of multiple functions to a given protein, which is becoming common for both prokaryotic and eukaryotic proteins and is referred to as the moonlighting nature of proteins (Gancedo and Flores, 2008; Jeffery, 2009; Tompa et al., 2005). Table 3 summarizes some of the commonly employed module-assisted techniques for functional inference [see references (Frades and Matthiesen, 2010) and (Zhao et al., 2008a) for more elaborate discussion on different clustering techniques currently available].

5. Network-based prediction of function from genomic context (GC): enzymes as a case study

Network-based analyses add to the power of computational prediction of functional categories for non-annotated genes. High-throughput and computational methods for inference of functional associations might clearly indicate a link between an unannotated gene and an annotated one. But that alone might not be sufficient evidence that they belong to the same category. It is also evident that a gene may be linked to more than just another gene, which can complicate the decision to label an unannotated gene. One way around this problem is to use network-based measures to assign one or more functional labels to unannotated genes with some confidence score.

The semantic similarity (SS) (Wang et al., 2007) between two genes represents the closeness of their functional annotations in databases such as GO. For instance, using the topology and child-to-parent relationships of the GO graph, one can determine whether pairs or sets of genes are closely related to each other in a functional context. For example, the genes *fumC* and *sdhA* are functionally similar to each other because they participate in the tricarboxilic-acid-cycle, which is reflected in their high SS (0.942). In contrast, the gene *aspA*, which participates in the biosynthesis of asparagine, is functionally less related with either *fumC* (SS=0.762) or *sdhA* (SS=0.704). In principle, one can expect the same behavior to hold true for non-annotated genes allowing us to make functional inferences.

To determine whether network-based approaches improve functional inferences over raw interaction/inference scores we contrasted two types of scores *versus* the SS of gene pairs in *E. coli*. The first score, called 'Raw', represents the integration of four GC-based methods: gene fusions, conservation of gene order, phylogenetic profiles and operon rearrangements, restricted to high confidence inferences (≥ 0.9) (Hu et al., 2009a). The second score, called weighted Topological Overlap Matrix (TOM) score (Ravasz et al., 2002; Zhang and Horvath, 2005), represents the proportion of 'raw' links in common between two nodes (between two genes in this case) normalized by the total number of links involving those nodes and their confidence weight.

As can be observed in Fig. 2, the TOM score allows for more detailed predictions than the Raw score. Compare for example the cutoffs: Raw ≥ 0.99 ($\sim$17,300 pairs) *versus* TOM ≥ 0.4 ($\sim$15,200 pairs). Both distributions show two peaks (bimodal) but for the raw score the highest peak is at an SS score of ~ 0.78, whereas with the TOM score the dominant peak is at an SS score of ~ 0.95. TOM plots show a very clear inflection point at TOM ≥ 0.4, where the dominating population are pairs with high SS (~ 0.95), whereas for TOM cutoffs < 0.4 the dominating SS is ~ 0.75. This inflection also occurs in raw scores but is not as striking as in TOM. Also it is noteworthy to mention that TOM scores allow to continue distinguishing the similarity between pairs when raw scores already reach their maximum value, where $\sim$9000 pairs all have raw scores=1 (thus no more stringent cutoff can be chosen), whereas TOM still allows to continue distinguishing pairs with closer SS. These observations show that network-based approaches provide more detailed information on functional annotations than their raw seeds and suggest that they are more amenable for gene function prediction. Additionally, enzymes show slightly higher peaks at the same cutoff values as the overall network (top *versus* bottom panel of Fig. 2), suggesting that metabolic functions might be predicted with more confidence than non-enzymatic functions. Further studies are necessary to determine whether these differences are influenced by the topological properties of the input network and/or other biological attributes of proteins from a particular functional class under study.

6. Conclusion and discussion

While the post-genomic era has introduced the genomic complement of hundreds of microbes, it has also provided us with several unanswered questions regarding the functional relevance of the genes discovered therein. It is noteworthy to mention that even in a model organism like *E. coli*, which has been the workhorse for molecular genetics for more than 100 years, nearly one-third of the genes have no experimental evidence

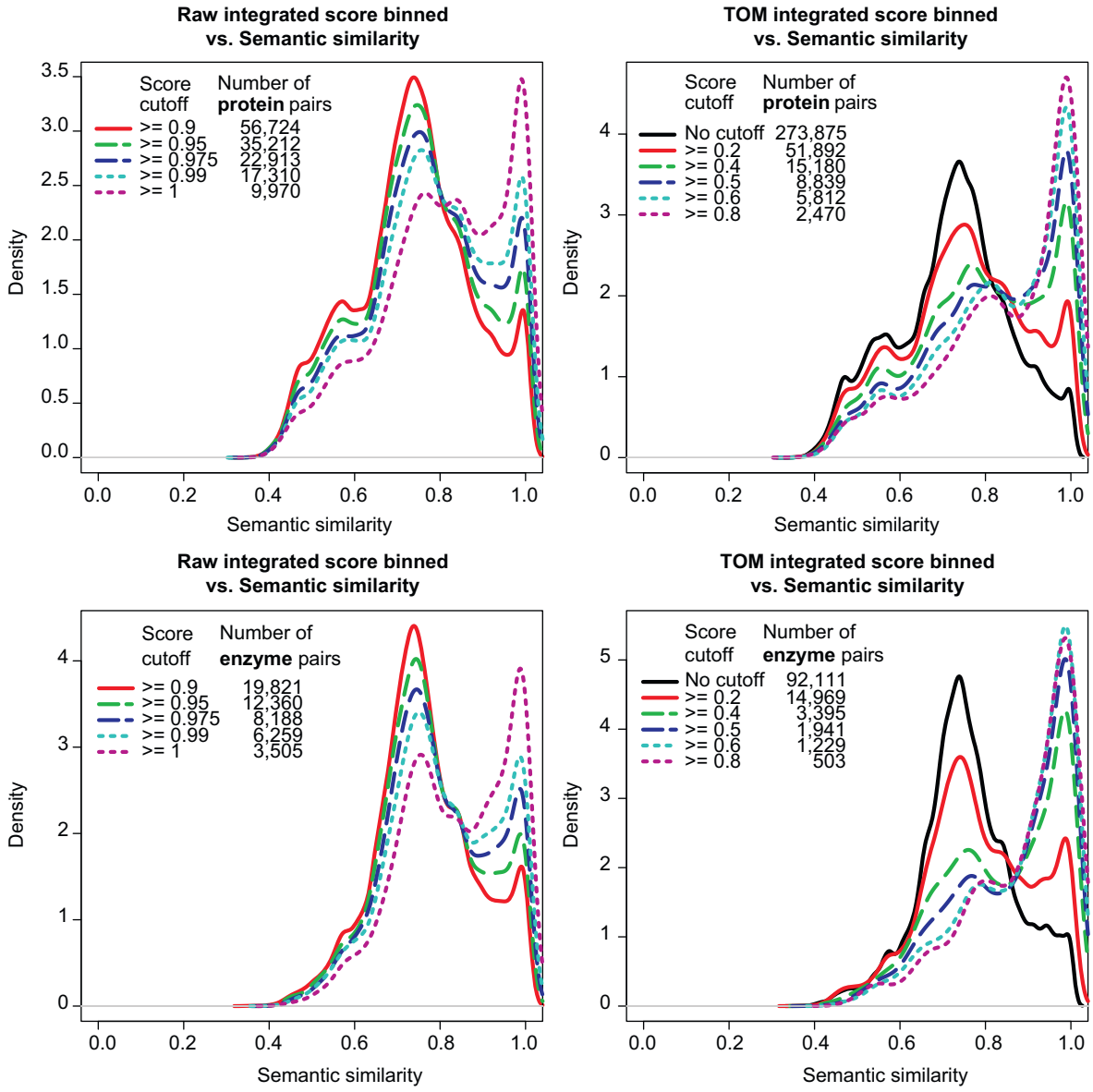

Fig. 2. Semantic similarity (SS) of GC-based functional inference networks for gene pairs in *E. coli*. Two types of scores were used, the raw integrated GC score (Hu et al., 2009a) [left panels]; and a weighted Topological Overlap Matrix (TOM) score (Ravasz et al., 2002; Zhang and Horvath, 2005) [right panels]. Note that TOM scores allow us to go deeper into functional associations and to further relate the genes predicted to functionally interact beyond the point where the raw score has reached a maximum, thus highlighting the importance of using network-based measures for analyzing functional associations, predicting a functional category, and advancing our understanding of functional modules. While the top panel shows the comparison of raw and TOM scores for varying SS scores between all pairs of genes, the bottom panel represents the same for enzymatic gene pairs.

supporting their biological role (Hu et al., 2009a; Riley et al., 2006) with other model systems harboring higher fractions of unannotated gene complements in their genomes (Sharan et al., 2007). Novel approaches are therefore needed not only to complement existing functional inference techniques but to also enable the prediction of functional associations of uncharacterized genes either to already characterized processes, or as new groups with yet to be known processes. Network-based approaches discussed here can be a big leap forward in uncovering the functional complements of genomes in years to come with the number of uncharacterized genes across the prokaryotic lineage growing at a considerable rate. Availability of high-throughput data at various levels from genomic to metagenomic and transcriptomic to system-wide interaction maps should enable the integration of data for better functional inferences at least in model systems in the next decade.

A major issue currently faced in the automated functional inference field is the lack of a systematic comparison of the number of different methods, thereby hindering bench scientists to choose the most appropriate tools for the prediction of function of a protein of their interest. Since different methods have been proposed to perform well for different functional schemas, a comparison of prediction methods with gold standard datasets may provide insights, or at least advice newcomers, about the appropriate choice of methods for the specific problem of interest. One possibility could be to construct user-friendly servers to automatically direct the user to the best performing method depending on the input, and/or provide a user with a set of best-performing methods for a functional schema of interest to the user.

In addition to the lack of appropriate tools for predicting function, another evident problem with network-based or more

generally automated functional inference algorithms is the depth at which they can predict function. Since these algorithms rely on existing annotations for transferring labels with certain probability to unannotated genes, one can not predict the exact function – for instance, at the level of molecular role – but rather prediction is limited to the currently available depth in GO hierarchy. This is especially an issue for processes which are poorly characterized and are also lower in representation in the genome, making it hard to identify their role. Nevertheless, in such cases the grouping of functional related genes into modules by these methods, can aid in designing more focused experiments to unravel the phenotypes.

Yet another key issue in predicting function from networks is the incompleteness of data and the inherent noise in the interactomes. To address this, several groups are increasingly using integration of data as a means of reducing noise (false positives) and to increase coverage. However, integration of data is in itself a challenging problem in data mining as several questions come in whenever data needs to be integrated. For instance, data integration often demands to have individual gold standards for each type of data being employed, it also requires an estimate of the dependency between the data types and the inherent noise in each data type. Add to this the fact that integration of data does not always increase accuracy. Therefore, factors like the nature of the data and their compatibility with existing data sources, availability of benchmarking datasets, and dependency between the sources, are some of the issues that need to be considered in developing techniques for integration of data. Nevertheless, given the volume of high-throughput data that is being reported for understanding diverse model systems the time is ripe to employ these network-based approaches to unravel the functions of the ever-increasing number of uncharacterized proteins accumulating in sequence databases.

Acknowledgements

SCJ acknowledges financial support from MRC Laboratory of Molecular Biology and Cambridge Commonwealth Trust. GM-H acknowledges research support from the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canadian Institutes of Health Research (CIHR) and computational facilities from the Shared Hierarchical Academic Research Computing Network (SHARCNET). JJD-M has been supported by funds from a Canadian Institutes of Health Research (CIHR) Operational Grant and the Mexican Science and Technology Research Council (CONACYT). We would also like to thank AJ Venkatakrishnan and Guilham Chalancon for critically reading the manuscript and providing helpful comments.

References

Alexeyenko, A., Tamas, I., Liu, G., Sonnhammer, E.L., 2006. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. Bioinformatics 22, e9–15.

Altaf-Ul-Amin, M., Shinbo, Y., Mihara, K., Kurokawa, K., Kanaya, S., 2006. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. BMC Bioinform. 7, 207.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Andreeva, A., Howorth, D., Chandonia, J.M., Brenner, S.E., Hubbard, T.J., Chothia, C., Murzin, A.G., 2008. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res. 36, D419–D425.

Aoki, K., Ogata, Y., Shibata, D., 2007. Approaches for extracting practical information from gene co-expression networks in plant biology. Plant Cell Physiol. 48, 381–390.

Apweiler, R., 2001. Functional information in SWISS-PROT: the basis for large-scale characterisation of protein sequences. Brief Bioinform. 2, 9–18.

Arakaki, A.K., Huang, Y., Skolnick, J., 2009. EFICAz2: enzyme function inference by a combined approach enhanced by machine learning. BMC Bioinform. 10, 107.

Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A.T., Kerrien, S., Khadake, J., Kerssemakers, J., Leroy, C., Menden, M., Michaut, M., Montecchi-Palazzi, L., Neuhauser, S.N., Orchard, S., Perreau, V., Roechert, B., van Eijk, K., Hermjakob, H., 2010. The IntAct molecular interaction database in 2010. Nucleic Acids Res. 38, D525–D531.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25–29.

Attwood, T.K., Bradley, P., Flower, D.R., Gaulton, A., Maudling, N., Mitchell, A.L., Moulton, G., Nordle, A., Paine, K., Taylor, P., Uddin, A., Zygouri, C., 2003. PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res. 31, 400–402.

Babu, M., Musso, G., Diaz-Mejia, J.J., Butland, G., Greenblatt, J.F., Emili, A., 2009. Systems-level approaches for identifying and analyzing genetic interaction networks in *Escherichia coli* and extensions to other prokaryotes. Mol. Biosyst. 5, 1439–1455.

Bader, G.D., Betel, D., Hogue, C.W., 2003. BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res. 31, 248–250.

Bader, G.D., Hogue, C.W., 2003. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinform. 4, 2.

Barker, D., Pagel, M., 2005. Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. PLoS Comput. Biol. 1, e3.

Barutcuoglu, Z., Schapire, R.E., Troyanskaya, O.G., 2006. Hierarchical multi-label prediction of gene function. Bioinformatics 22, 830–836.

Berglund, A.C., Sjolund, E., Ostlund, G., Sonnhammer, E.L., 2008. InParanoid 6: eukaryotic ortholog clusters with inparalogs. Nucleic Acids Res. 36, D263–D266.

Bowers, P.M., Pellegrini, M., Thompson, M.J., Fierro, J., Yeates, T.O., Eisenberg, D., 2004. Prolinks: a database of protein functional linkages derived from coevolution. Genome Biol. 5, R35.

Breitkreutz, B.J., Stark, C., Reguly, T., Boucher, L., Breitkreutz, A., Livstone, M., Oughtred, R., Lackner, D.H., Bahler, J., Wood, V., Dolinski, K., Tyers, M., 2008. The BioGRID interaction database: 2008 update. Nucleic Acids Res. 36, D637–D640.

Brouwer, R.W., Kuipers, O.P., van Hijum, S.A., 2008. The relative value of operon predictions. Brief Bioinform. 9, 367–375.

Bru, C., Courcelle, E., Carrere, S., Beausse, Y., Dalmar, S., Kahn, D., 2005. The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acids Res. 33, D212–D215.

Brun, C., Chevenet, F., Martin, D., Wojcik, J., Guenoche, A., Jacq, B., 2003. Functional classification of proteins for the prediction of cellular function from a protein–protein interaction network. Genome Biol. 5, R6.

Butland, G., Babu, M., Diaz-Mejia, J.J., Bohdana, F., Phanse, S., Gold, B., Yang, W., Li, J., Gagarinova, A.G., Pogoutse, O., Mori, H., Wanner, B.L., Lo, H., Wasniewski, J., Christopolous, C., Ali, M., Venn, P., Safavi-Naini, A., Sourour, N., Caron, S., Choi, J.Y., Laigle, L., Nazarians-Armavil, A., Deshpande, A., Joe, S., Datsenko, K.A., Yamamoto, N., Andrews, B.J., Boone, C., Ding, H., Sheikh, B., Moreno-Hagelsieb, G., Greenblatt, J.F., Emili, A., 2008. eSGA: *E. coli* synthetic genetic array analysis. Nat. Methods 5, 789–795.

Cai, C.Z., Han, L.Y., Ji, Z.L., Chen, X., Chen, Y.Z., 2003. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res. 31, 3692–3697.

Chen, F., Mackey, A.J., Stoeckert Jr., C.J., Roos, D.S., 2006. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res. 34, D363–D368.

Chen, Y., Xu, D., 2004. Global protein function annotation through mining genome-scale data in yeast *Saccharomyces cerevisiae*. Nucleic Acids Res. 32, 6414–6424.

Chua, H.N., Sung, W.K., Wong, L., 2006. Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. Bioinformatics 22, 1623–1630.

Claudel-Renard, C., Chevalet, C., Faraut, T., Kahn, D., 2003. Enzyme-specific profiles for genome annotation: PRIAM. Nucleic Acids Res. 31, 6633–6639.

Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M., Robles, M., 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21, 3674–3676.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L., Toufighi, K., Mostafavi, S., Prinz St., J., Onge, R.P., VanderSluis, B., Makhnevych, T., Vizeacoumar, F.J., Alizadeh, S., Bahr, S., Brost, R.L., Chen, Y., Cokol, M., Deshpande, R., Li, Z., Lin, Z.Y., Liang, W., Marback, M., Paw, J., San Luis, B.J., Shuteriqi, E., Tong, A.H., van Dyk, N., Wallace, I.M., Whitney, J.A., Weirauch, M.T., Zhong, G., Zhu, H., Houry, W.A., Brudno, M., Ragibizadeh, S., Papp, B., Pal, C., Roth, F.P., Giaever, G., Nislow, C., Troyanskaya, O.G., Bussey, H., Bader, G.D., Gingras, A.C., Morris, Q.D., Kim, P.M., Kaiser, C.A., Myers, C.L., Andrews, B.J., Boone, C., 2010. The genetic landscape of a cell. Science 327, 425–431.

Cuff, A.L., Sillitoe, I., Lewis, T., Redfern, O.C., Garratt, R., Thornton, J., Orengo, C.A., 2009. The CATH classification revisited—architectures reviewed and new ways to characterize structural divergence in superfamilies. Nucleic Acids Res. 37, D310–D314.

Dandekar, T., Snel, B., Huynen, M., Bork, P., 1998. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem. Sci. 23, 324–328.

Date, S.V., Marcotte, E.M., 2003. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. Nat. Biotechnol. 21, 1055–1062.
Davila Lopez, M., Martinez Guerra, J.J., Samuelsson, T., 2010. Analysis of gene order conservation in eukaryotes identifies transcriptionally and functionally linked genes. PLoS One 5, e10654.
Deng, M., Zhang, K., Mehta, S., Chen, T., Sun, F., 2003. Prediction of protein function using protein–protein interaction data. J. Comput. Biol. 10, 947–960.
Engelhardt, B.E., Jordan, M.I., Muratore, K.E., Brenner, S.E., 2005. Protein molecular function prediction by Bayesian phylogenomics. PLoS Comput. Biol. 1, e45.
Enright, A.J., Iliopoulos, I., Kyrpides, N.C., Ouzounis, C.A., 1999. Protein interaction maps for complete genomes based on gene fusion events. Nature 402, 86–90.
Enright, A.J., Van Dongen, S., Ouzounis, C.A., 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30, 1575–1584.
Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L., Bateman, A., 2008. The Pfam protein families database. Nucleic Acids Res. 36, D281–D288.
Frades, I., Matthiesen, R., 2010. Overview on techniques in cluster analysis. Methods Mol. Biol. 593, 81–107.
Gaasterland, T., Ragan, M.A., 1998. Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. Microb. Comp. Genom. 3, 199–217.
Gancedo, C., Flores, C.L., 2008. Moonlighting proteins in yeasts. Microbiol Mol. Biol. Rev. 72, 197–210 table of contents.
Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dumpelfeld, B., Edelmann, A., Heurtier, M.A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A.M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J.M., Kuster, B., Bork, P., Russell, R.B., Superti-Furga, G., 2006. Proteome survey reveals modularity of the yeast cell machinery. Nature 440, 631–636.
Godzik, A., Jambon, M., Friedberg, I., 2007. Computational protein function prediction: are we making progress? Cell Mol. Life Sci. 64 2505–2511.
Gotoh, O., 1999. Multiple sequence alignment: algorithms and applications. Adv. Biophys. 36, 159–206.
Groth, D., Lehrach, H., Hennig, S., 2004. GOblet: a platform for Gene Ontology annotation of anonymous sequence data. Nucleic Acids Res. 32, W313–W317.
Han, L., Cui, J., Lin, H., Ji, Z., Cao, Z., Li, Y., Chen, Y., 2006. Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity. Proteomics 6, 4023–4037.
Hartwell, L.H., Hopfield, J.J., Leibler, S., Murray, A.W., 1999. From molecular to modular cell biology. Nature. 402, C47–52.
Hawkins, T., Chitale, M., Luban, S., Kihara, D., 2009. PFP: automated prediction of gene ontology functional annotations with confidence scores using protein sequence data. Proteins 74, 566–582.
Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., Takagi, T., 2001. Assessment of prediction accuracy of protein function from protein–protein interaction data. Yeast 18, 523–531.
Hollich, V., Storm, C.E., Sonnhammer, E.L., 2002. OrthoGUI: graphical presentation of Orthostrapper results. Bioinformatics 18, 1272–1273.
Hu, P., Janga, S.C., Babu, M., Diaz-Mejia, J.J., Butland, G., Yang, W., Pogoutse, O., Guo, X., Phanse, S., Wong, P., Chandran, S., Christopoulos, C., Nazarians-Armavil, A., Nasseri, N.K., Musso, G., Ali, M., Nazemof, N., Eroukova, V., Golshani, A., Paccanaro, A., Greenblatt, J.F., Moreno-Hagelsieb, G., Emili, A., 2009a. Global functional atlas of *Escherichia coli* encompassing previously uncharacterized proteins. PLoS Biol. 7, e96.
Hu, Z., Hung, J.H., Wang, Y., Chang, Y.C., Huang, C.L., Huyck, M., DeLisi, C., 2009b. VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. Nucleic Acids Res. 37, W115–W221.
Huber, W., Carey, V.J., Long, L., Falcon, S., Gentleman, R., 2007. Graphs in molecular biology. BMC Bioinform. 8 (Suppl. 6), S8.
Hurst, L.D., Pal, C., Lercher, M.J., 2004. The evolutionary dynamics of eukaryotic gene order. Nat. Rev. Genet. 5, 299–310.
Janga, S.C., Collado-Vides, J., Moreno-Hagelsieb, G., 2005. Nebulon: a system for the inference of functional relationships of gene products from the rearrangement of predicted operons. Nucleic Acids Res. 33, 2521–2530.
Janga, S.C., Moreno-Hagelsieb, G., 2004. Conservation of adjacency as evidence of paralogous operons. Nucleic Acids Res. 32, 5392–5397.
Jeffery, C.J., 2009. Moonlighting proteins—an update. Mol. Biosyst. 5, 345–350.
Jensen, L.J., Gupta, R., Staerfeldt, H.H., Brunak, S., 2003. Prediction of human protein function according to Gene Ontology categories. Bioinformatics 19, 635–642.
Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P., von Mering, C., 2009. STRING 8—a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res. 37, D412–D416.
Jocker, A., Hoffmann, F., Groscurth, A., Schoof, H., 2008. Protein function prediction and annotation in an integrated environment powered by web services (AFAWE). Bioinformatics 24, 2393–2394.
Jones, C.E., Schwerdt, J., Bretag, T.A., Baumann, U., Brown, A.L., 2008. GOSLING: a rule-based protein annotator using BLAST and GO. Bioinformatics 24, 2628–2629.
Kaplan, N., Sasson, O., Inbar, U., Friedlich, M., Fromer, M., Fleischer, H., Portugaly, E., Linial, N., Linial, M., 2005. ProtoNet 4.0: a hierarchical classification of one million protein sequences. Nucleic Acids Res. 33, D216–D218.
Karaoz, U., Murali, T.M., Letovsky, S., Zheng, Y., Ding, C., Cantor, C.R., Kasif, S., 2004. Whole-genome annotation by using evidence integration in functional-linkage networks. Proc. Natl. Acad. Sci. USA 101, 2888–2893.
King, A.D., Przulj, N., Jurisica, I., 2004. Protein complex prediction via cost-based clustering. Bioinformatics 20, 3013–3020.
Kretschmann, E., Fleischmann, W., Apweiler, R., 2001. Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. Bioinformatics 17, 920–926.
Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., Punna, T., Peregrin-Alvarez, J.M., Shales, M., Zhang, X., Davey, M., Robinson, M.D., Paccanaro, A., Bray, J.E., Sheung, A., Beattie, B., Richards, D.P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M.M., Vlasblom, J., Wu, S., Orsi, C., Collins, S.R., Chandran, S., Haw, R., Rilstone, J.J., Gandi, K., Thompson, N.J., Musso St, G., Onge, P., Ghanny, S., Lam, M.H., Butland, G., Altaf-Ul, A.M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J.S., Ingles, C.J., Hughes, T.R., Parkinson, J., Gerstein, M., Wodak, S.J., Emili, A., Greenblatt, J.F., 2006. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. Nature 440, 637–643.
Lanckriet, G.R., Deng, M., Cristianini, N., Jordan, M.I., Noble, W.S., 2004. Kernel-based data fusion and its application to protein function prediction in yeast. Pac. Symp. Biocomput., 300–311.
Lasko, P., 2000. The drosophila melanogaster genome: translation factors and RNA binding proteins. J Cell Biol. 150, F51–F56.
Lee, H., Tu, Z., Deng, M., Sun, F., Chen, T., 2006. Diffusion kernel-based logistic regression models for protein function prediction. OMICS 10, 40–55.
Lees, J., Yeats, C., Redfern, O., Clegg, A., Orengo, C., Gene3D: merging structure and function for a thousand genomes. Nucleic Acids Res. 38, D296–D300.
Letovsky, S., Kasif, S., 2003. Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics 19 (Suppl. 1), i197–204.
Letunic, I., Doerks, T., Bork, P., 2009. SMART 6: recent updates and new developments. Nucleic Acids Res. 37, D229–32.
Linghu, B., Snitkin, E.S., Holloway, D.T., Gustafson, A.M., Xia, Y., DeLisi, C., 2008. High-precision high-coverage functional inference from integrated data sources. BMC Bioinform. 9, 119.
Liu, X., Han, B., 2009. Evolutionary conservation of neighbouring gene pairs in plants. Gene 437, 71–79.
Lobley, A.E., Nugent, T., Orengo, C.A., Jones, D.T., 2008. FFPred: an integrated feature-based function prediction server for vertebrate proteomes. Nucleic Acids Res. 36, W297–302.
Luo, F., Yang, Y., Zhong, J., Gao, H., Khan, L., Thompson, D.K., Zhou, J., 2007. Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. BMC Bioinform. 8, 299.
Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O., Eisenberg, D., 1999. A combined algorithm for genome-wide prediction of protein function. Nature 402, 83–86.
Martin, D.M., Berriman, M., Barton, G.J., 2004. GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. BMC Bioinform. 5, 178.
Massjouni, N., Rivera, C.G., Murali, T.M., 2006. VIRGO: computational prediction of gene functions. Nucleic Acids Res. 34, W340–344.
Mi, H., Dong, Q., Muruganujan, A., Gaudet, P., Lewis, S., Thomas, P. D., PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. Nucleic Acids Res. 38, D204–D210.
Moreno-Hagelsieb, G., Collado-Vides, J., 2002. A powerful non-homology method for the prediction of operons in prokaryotes. Bioinformatics 18 (Suppl 1), S329–336.
Moreno-Hagelsieb, G., Janga, S.C., 2008. Operons and the effect of genome redundancy in deciphering functional relationships using phylogenetic profiles. Proteins 70, 344–352.
Mulder, N.J., Kersey, P., Pruess, M., Apweiler, R., 2008. In silico characterization of proteins: UniProt, InterPro and Integr8. Mol. Biotechnol. 38, 165–177.
Muller, J., Szklarczyk, D., Julien, P., Letunic, I., Roth, A., Kuhn, M., Powell, S., von Mering, C., Doerks, T., Jensen, L.J., Bork, P., 2010. eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. Nucleic Acids Res. 38, D190–195.
Myers, C.L., Chiriac, C., Troyanskaya, O.G., 2009. Discovering biological networks from diverse functional genomic data. Methods Mol. Biol. 563, 157–175.
Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., Singh, M., 2005. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics 21 (Suppl. 1), i302–310.
Nikolskaya, A.N., Arighi, C.N., Huang, H., Barker, W.C., Wu, C.H., 2006. PIRSF family classification system for protein functional and evolutionary analysis. Evol. Bioinform. Online 2, 197–209.
Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., Maltsev, N., 1999. The use of gene clusters to infer functional coupling. Proc. Natl. Acad. Sci. USA 96, 2896–2901.
Pandey, G., Myers, C.L., Kumar, V., 2009. Incorporating functional inter-relationships into protein function prediction algorithms. BMC Bioinform. 10, 142.
Pearson, W.R., 1995. Comparison of methods for searching protein sequence databases. Protein Sci. 4, 1145–1160.
Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., Yeates, T.O., 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. USA 96, 4285–4288.
Pereira-Leal, J.B., Enright, A.J., Ouzounis, C.A., 2004. Detection of functional modules from protein interaction networks. Proteins 54, 49–57.
Petryszak, R., Kretschmann, E., Wieser, D., Apweiler, R., 2005. The predictive power of the CluSTr database. Bioinformatics 21, 3604–3609.

Pignatelli, M., Serras, F., Moya, A., Guigo, R., Corominas, M., 2009. CROC: finding chromosomal clusters in eukaryotic genomes. Bioinformatics 25, 1552–1553.
Procter, J.B., Thompson, J., Letunic, I., Creevey, C., Jossinet, F., Barton, G.J., 2010. Visualization of multiple alignments, phylogenies and gene family evolution. Nat. Methods 7, S16–25.
Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., Barabasi, A.L., 2002. Hierarchical organization of modularity in metabolic networks. Science 297, 1551–1555.
Rentzsch, R., Orengo, C.A., 2009. Protein function prediction—the power of multiplicity. Trends Biotechnol. 27, 210–219.
Riley, M., 1993. Functions of the gene products of *Escherichia coli*. Microbiol. Rev. 57, 862–952.
Riley, M., Abe, T., Arnaud, M.B., Berlyn, M.K., Blattner, F.R., Chaudhuri, R.R., Glasner, J.D., Horiuchi, T., Keseler, I.M., Kosuge, T., Mori, H., Perna, N.T., Plunkett 3rd, G., Rudd, K.E., Serres, M.H., Thomas, G.H., Thomson, N.R., Wishart, D., Wanner, B.L., 2006. Escherichia coli K-12: a cooperatively developed annotation snapshot—2005. Nucleic Acids Res. 34, 1–9.
Rison, S.C., Hodgman, T.C., Thornton, J.M., 2000. Comparison of functional annotation schemes for genomes. Funct. Integr. Genom. 1, 56–69.
Rives, A.W., Galitski, T., 2003. Modular organization of cellular networks. Proc. Natl. Acad. Sci. USA 100, 1128–1133.
Rogozin, I.B., Makarova, K.S., Murvai, J., Czabarka, E., Wolf, Y.I., Tatusov, R.L., Szekely, L.A., Koonin, E.V., 2002. Connected gene neighborhoods in prokaryotic genomes. Nucleic Acids Res. 30, 2212–2223.
Ruan, J., Dean, A.K., Zhang, W., 2010. A general co-expression network-based approach to gene expression analysis: comparison and applications. BMC Syst. Biol. 4, 8.
Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Guldener, U., Mannhaupt, G., Munsterkotter, M., Mewes, H.W., 2004. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res. 32, 5539–5545.
Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., Wu, D., Eisen, J.A., Hoffman, J.M., Remington, K., Beeson, K., Tran, B., Smith, H., Baden-Tillson, H., Stewart, C., Thorpe, J., Freeman, J., Andrews-Pfannkoch, C., Venter, J.E., Li, K., Kravitz, S., Heidelberg, J.F., Utterback, T., Rogers, Y.H., Falcon, L.I., Souza, V., Bonilla-Rosso, G., Eguiarte, L.E., Karl, D.M., Sathyendranath, S., Platt, T., Bermingham, E., Gallardo, V., Tamayo-Castillo, G., Ferrari, M.R., Strausberg, R.L., Nealson, K., Friedman, R., Frazier, M., Venter, J.C., 2007. The sorcerer II global ocean sampling expedition: Northwest Atlantic through Eastern tropical Pacific. PLoS Biol. 5, e77.
Samanta, M.P., Liang, S., 2003. Predicting protein functions from redundancies in large-scale protein interaction networks. Proc. Natl. Acad. Sci. USA 100, 12579–12583.
Schwikowski, B., Uetz, P., Fields, S., 2000. A network of protein-protein interactions in yeast. Nat. Biotechnol. 18, 1257–1261.
Selengut, J.D., Haft, D.H., Davidsen, T., Ganapathy, A., Gwinn-Giglio, M., Nelson, W.C., Richter, A.R., White, O., 2007. TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. Nucleic Acids Res. 35, D260–264.
Serres, M.H., Riley, M., 2000. MultiFun, a multifunctional classification scheme for *Escherichia coli* K-12 gene products. Microb. Comp. Genomics. 5, 205–222.
Sharan, R., Ulitsky, I., Shamir, R., 2007. Network-based prediction of protein function. Mol. Syst. Biol. 3, 88.
Shen, H.B., Chou, K.C., 2007. EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. Biochem. Biophys. Res. Commun. 364, 53–59.
Shoemaker, B.A., Panchenko, A.R., 2007. Deciphering protein–protein interactions. Part I. Experimental techniques and databases. PLoS Comput. Biol. 3, e42.
Sigrist, C. J., Cerutti, L., de Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A., Hulo, N., PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res. 38, D161–D166.
Snel, B., Bork, P., Huynen, M.A., 2002. The identification of functional modules from the genomic association of genes. Proc. Natl. Acad. Sci. USA 99, 5890–5895.
Spirin, V., Mirny, L.A., 2003. Protein complexes and functional modules in molecular networks. Proc. Natl. Acad. Sci. USA 100, 12123–12128.
Tanay, A., Sharan, R., Kupiec, M., Shamir, R., 2004. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. Proc. Natl. Acad. Sci. USA 101, 2981–2986.
Tarassov, K., Messier, V., Landry, C.R., Radinovic, S., Serna Molina, M.M., Shames, I., Malitskaya, Y., Vogel, J., Bussey, H., Michnick, S.W., 2008. An in vivo map of the yeast protein interactome. Science 320, 1465–1470.
Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J., Natale, D.A., 2003. The COG database: an updated version includes eukaryotes. BMC Bioinform. 4, 41.
Tatusov, R.L., Koonin, E.V., Lipman, D.J., 1997. A genomic perspective on protein families. Science 278, 631–637.
Tipton, K.F., 1994. Nomenclature committee of the international union of biochemistry and molecular biology (NC-IUBMB). Enzyme nomenclature. Recommendations 1992. Supplement: corrections and additions. Eur. J. Biochem. 223, 1–5.
Tompa, P., Szasz, C., Buday, L., 2005. Structural disorder throws new light on moonlighting. Trends Biochem. Sci. 30, 484–489.
Troyanskaya, O.G., Dolinski, K., Owen, A.B., Altman, R.B., Botstein, D., 2003. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). Proc. Natl. Acad. Sci. USA 100, 8348–8353.
Tsuda, K., Shin, H., Scholkopf, B., 2005. Fast protein classification with multiple networks. Bioinformatics 21 (Suppl 2), ii59–65.
Vazquez, A., Flammini, A., Maritan, A., Vespignani, A., 2003. Global protein function prediction from protein–protein interaction networks. Nat. Biotechnol. 21, 697–700.
Vinayagam, A., del Val, C., Schubert, F., Eils, R., Glatting, K.H., Suhai, S., Konig, R., 2006. GOPET: a tool for automated predictions of Gene Ontology terms. BMC Bioinform. 7, 161.
Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S., Chen, C.F., 2007. A new method to measure the semantic similarity of GO terms. Bioinformatics 23, 1274–1281.
Wang, K., Narayanan, M., Zhong, H., Tompa, M., Schadt, E.E., Zhu, J., 2009. Meta-analysis of inter-species liver co-expression networks elucidates traits associated with common human diseases. PLoS Comput. Biol. 5, e1000616.
Whisstock, J.C., Lesk, A.M., 2003. Prediction of protein function from protein sequence and structure. Q. Rev. Biophys. 36, 307–340.
Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., Chothia, C., Gough, J., 2009. SUPERFAMILY-sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Res. 37, D380–386.
Yeung, N., Cline, M. S., Kuchinsky, A., Smoot, M. E., Bader, G. D., 2008. Exploring biological networks with Cytoscape software. Curr. Protoc. Bioinformatics. Chapter 8, Unit 8 13.
Yooseph, S., Sutton, G., Rusch, D.B., Halpern, A.L., Williamson, S.J., Remington, K., Eisen, J.A., Heidelberg, K.B., Manning, G., Li, W., Jaroszewski, L., Cieplak, P., Miller, C.S., Li, H., Mashiyama, S.T., Joachimiak, M.P., van Belle, C., Chandonia, J.M., Soergel, D.A., Zhai, Y., Natarajan, K., Lee, S., Raphael, B.J., Bafna, V., Friedman, R., Brenner, S.E., Godzik, A., Eisenberg, D., Dixon, J.E., Taylor, S.S., Strausberg, R.L., Frazier, M., Venter, J.C., 2007. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol. 5, e16.
Yu, C., Zavaljevski, N., Desai, V., Reifman, J., 2009. Genome-wide enzyme annotation with precision control: catalytic families (CatFam) databases. Proteins 74, 449–460.
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., Barabasi, A.L., Tavernier, J., Hill, D.E., Vidal, M., 2008. High-quality binary protein interaction map of the yeast interactome network. Science 322, 104–110.
Zehetner, G., 2003. OntoBlast function: from sequence similarities directly to potential functional annotations by ontology terms. Nucleic Acids Res. 31, 3799–3803.
Zhang, B., Horvath, S., 2005. A general framework for weighted gene co-expression network analysis. Stat. Appl. Genet. Mol. Biol. 4, Article 17.
Zhao, X.M., Chen, L., Aihara, K., 2008a. Protein function prediction with high-throughput data. Amino Acids 35, 517–530.
Zhao, X.M., Chen, L., Aihara, K., 2008b. Protein function prediction with the shortest path in functional linkage graph and boosting. Int. J. Bioinform. Res. Appl. 4, 375–384.
Zmasek, C.M., Eddy, S.R., 2002. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. BMC Bioinform. 3, 14.