Predicting Gene Function Using Omics Data: From Data Preparation to Data Integration

Weidong Tian, Xinran Dong, Yuanpeng Zhou, and Ren Ren

Abstract In the post-genomic era, the continuing development of high-throughput technologies has led to the explosion of enormous amount of omics data, ranging from genomics, transcriptomics, proteomics, metabolomics, to phenomics. Integration of diverse omics data can help us to understand the complete functions of genes in the cell. However, the complexity, heterogeneity, and large-scale of the omics data have created significant challenges to the gene function prediction field. Currently, the focus of this field is to develop efficient and accurate algorithms to integrate omics data for predicting gene function. In this chapter, we first introduce the various types of omics data, and how they relate to gene functions. Then, we review current algorithms available for integrating omics data for gene function predictions. Next, we use a combined algorithm named Funckenstein as an example to further illustrate the integration process. In the final two sections, we discuss current limitations and potential improvements of this field, and offer perspectives for future directions.

Introduction

Understanding the function of genes, including the molecular function, the biological role it plays in the cell, and the impact of its malfunction on phenotypes and diseases, is a central task in biology. Traditionally, experimentalists study the function of genes by focusing on one or a few at a time. The advent of genomic era has completely revolutionized our approach to study biology. Since the initiation of the Human Genome Project (HGP) in 1990 [1], the breakthrough of modern high-throughput sequencing technologies has allowed for the decoding of the complete genomic DNA sequences of more than a thousand cellular organisms including human genome. Along with the accomplishment of complete genome sequences have emerged a diverse range of high-throughput technologies such as

W. Tian (✉)
School of Life Sciences, Fudan University, Shanghai, China; Institute of Biostatistics,
Fudan University, Shanghai, China
e-mail: weidong.tian@fudan.edu.cn

D. Kihara (ed.), *Protein Function Prediction for Omics Era*,
DOI 10.1007/978-94-007-0881-5_12, © Springer Science+Business Media B.V. 2011

oligonucleotide array, cDNA array, high-throughput two hybrid system, mass spectrometry, and so on. Thanks to the continuously reduced cost of the high-throughput technologies, it is now a routine task for many laboratories to study the properties and relationships of thousands of genes in parallel, presenting biologists an unprecedented opportunity to study the function of genes at a system level.

Given the sheer volume of the omics data, how to take advantage of the data to generate biologically meaningful insights about gene functions presents a critical challenge to the field of biology. Computational biology or bioinformatics is thus emerging as a new discipline, aiming to develop computational and statistical algorithms to effectively sort, analyze and interpret the omics data. Gene function prediction is one of the most important goals of computational biology. It can not only provide hypothesis about the function of a particular set of genes of interest that can be verified experimentally, but also uncover important mechanisms of gene function through learning the rules of predicting gene function accurately.

As the genomics era starts with the flood of genomics data, i.e., gene and protein sequences, the computational approaches initially focus on inferring gene functions by sequence comparison [2–6]. The underlying hypothesis of the sequence-based methods is that homologous proteins evolving from the same ancestor are likely to share the same function. The sequence-based methods play important roles in annotation of the newly sequenced genomes, with the majority of genes functionally inferred on the basis of the sequence similarity to previously characterized proteins. However, this approach can provide functional insights to only 50% of the genes in the genome by detecting evolutionary relationship with known proteins [7].

The sequence-based methods on gene function prediction are effective in assigning the molecular function of genes, for instance, the catalytic activity of enzymes. However, it often fails to answer what role a gene plays in a biological process, how it interacts with other genes, and where it functions in the cell, which are fundamental questions in biology. This failure is mostly because those functional aspects of the gene are determined not only by the gene sequence, but also by its relationships with other genes that may not be evolutionarily related with the target. To answer those questions, information beyond sequence alone is required. The rich trove of omics data, ranging from genomic sequence, gene expression, protein–protein interaction, genetic interaction, phenotypic change, to epigenetic information, provide information about the behaviors of a gene from various aspects. Therefore, a current challenge in gene function prediction field is to design computational algorithms to piece together information from various types of omics data, in order to obtain the whole picture of the biological role of genes in the cell.

This chapter is organized as the following sections. In the first section, we focus on omics data preparation by describing the latest high-throughput technologies to generate the data and how each type of omics data is related to gene function. In the second section, we review current algorithms available for integrating omics data to predict gene functions. In the third section, we describe in detail a combined algorithm named Funckenstein to illustrate the process of omics-based gene function prediction [8]. In the final two sections, we discuss current limitations and potential improvements of the field, and offer perspective for future directions.

Omics Data Preparation

Before describing omics data and how they relate to gene functions, let's first clarify the meaning of function. The functions of a gene essentially are observations of its behavior in the cell. For a protein kinase, from a biochemist's point of view, its function can be the phosphorylation of a hydroxyl group of a specific substrate; while in a geneticist's opinion, its function can be the signaling transduction pathway in which the gene is involved, or the disease phenotype appearing when the gene is mutated or knocked out. In order to have a complete picture of the gene function, we need to have an ontology system covering various aspects of gene functions. Gene Ontology (GO) is such an ontology system [9]. It contains three ontologies: molecular function, biological process, and cellular component. Molecular function describes the biochemical activity of a gene product, "protein tyrosine kinase" for example. Biological process refers to the biological role to which a set of genes and gene products contribute, e.g. "DNA damage pathway". Cellular component tells where in the cell a gene operates its function, for instance, "nucleus". The GO terms are organized in a directed acyclic graph, and arranged in a manner from general to specific, making it easy to be parsed by computers. GO has become the most widely used functional annotation scheme, and the current goal of gene function prediction is to predict the GO terms associated with each gene in the genome. GO term annotation of genes in different genomes can be found in the GO database.

Following the central pathway of biological information flow from the genome to cellular phenotype, we classify the omics data into five main categories: genomics, transcriptomics, proteomics, metabolomics, and phenomics (Fig. 1). Genomics represents the whole genome sequence information including gene, regulatory element, and non-coding RNA, etc. Transcriptomics covers the whole RNA transcripts in the cell, while proteomics characterizes all proteins in the cell. Metabolomics consists of proteins, mostly enzymes, and metablotes that are catalyzed or produced by enzymes in the cell. Phenomics is the combined result of genomics, transcriptomics, proteomics, and metabolomics, representing all observable cellular or organism characteristics.

Genomics

The first complete genome of a living organism was sequenced in 1995 [10]. In 2003 the complete sequence of the human genome was finished [11]. Today, there are more than 1,000 completely sequenced genomes in the public domain, and some estimates this number could reach to more than 10,000 by 2012. This owes to the introduction of the next-generation sequencing technologies which employ massively parallel sequencing strategy, capable of sequencing millions of sequence

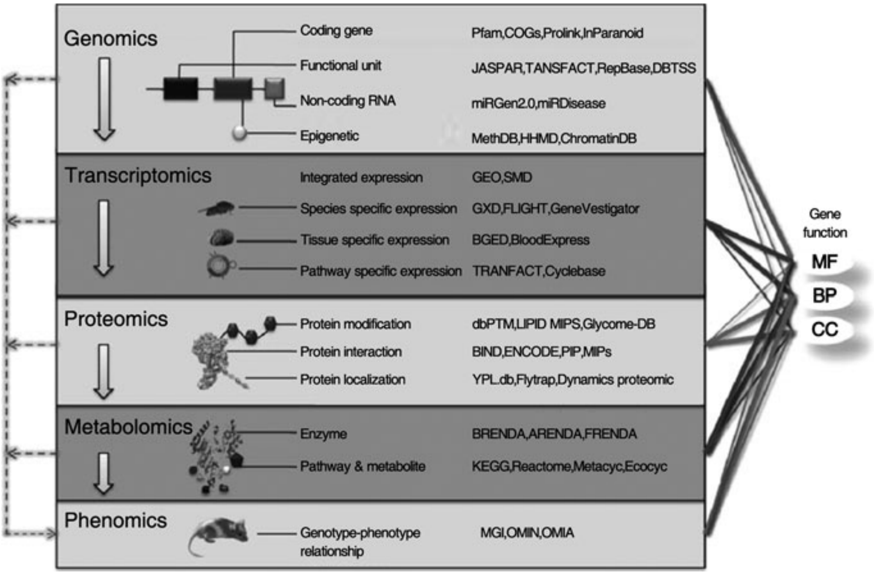

Fig. 1 Omics, database and gene function prediction. Omics data are classified into five main categories: genomics, transcriptomics, proteomics, metabolomics and phenomics. Sub-types of omics data and the representative associated databases under each category are shown. The reference for each database can be found in the main text. The link of each category of omics data to GO function is shown. Here, MF, BP, and CC are abbreviations of molecule function, biological process and cellular component, respectively. The *thickness of the line linking* omics data to gene function represents their empirical relative strength in predicting the corresponding GO functions. For example, among the three GO terms, genomics data are most effective in predicting MF, while phenomics data predict BP better than the other two GO terms. In contrast, proteomics and transcriptomics data predict BP and CC better than MF, while metabolomics data are effective in predicting MF and BP

reads in a single run, such as the 454 [12], Illumina [13] and SOLid system [14]. Using the new technology, the full genome of James Watson, the well-known DNA pioneer, was sequenced and assembled with 7.4 fold coverage in less than 2 months [15]. With such a development pace, the personalized-genomics era will be coming soon.

Model organism databases curate, manage, and store detailed up-to-date information about the gene mapping, annotation, protein domains and structures, expression data, mutant phenotypes, physical and genetic interactions, etc, of the model genomes, such as the Saccharomyces Genome Database (SGD) [16], the Mouse Genome Informatics (MGI) [17], the Arabidopsis Information Resource (TAIR) [18], the Fly Base [19], etc. Such databases are now the researcher′s starting point for informed hypothesis generation. There are also databases that store specific genomics data. According to the genome organization, we can classify those databases into coding gene, functional unit such as regulatory sequence, and non-coding RNA databases.

The coding gene and protein sequence databases contain information from a wide range of gene and protein sequence features. They have been the largest sources of training data for gene function prediction. Detection of evolutionary relationships is the first step of functional inference for sequence-based methods. Pfam is a database of evolutionarily related protein sequences [4]. It currently contains more than 10,000 protein families generated from the multiple sequence alignments (MSAs) of evolutionarily related sequences using Hidden Markov Models (HMMs). Those protein families cover more than 70% protein sequences in the protein universe. Evolutionary relationships between sequences can be further distinguished as orthology, paralogy, and inparalogy [20]. Because orthologous sequences are resulted from a speciation event, and likely retain the ancestral function, detection of orthologous relationships can be effective in making gene function prediction [21]. Such databases include Cluster of Othologous Groups (COGs) [22] and InParanoid [23], etc. Phylogenetic profile shows the pattern of the presence and absence of the homolog of a given gene in different genomes. Two genes with similar phylogenetic profile tend to be functionally related, e.g., involved in the same pathway. It provides a non-homology based way to infer functions. ProLinks is a databases of phylogenetic profiles [24]. In addition, there are databases focusing on protein sequence features and patterns related to protein functions, such as Prosite [25] and PRINTS [26]. Direct functional inference can be made when a new sequence matches a known protein feature or pattern.

Functional unit databases include those containing regulatory sequences (e.g., transcription factor (TF) binding sites), repeat elements, and other functional units, such as enhancer, silencer, etc. Those functional units are not genes, but they are located in the vicinity of gene, e.g., in the promoter region, and often evolutionarily conserved. They play important roles in regulating, altering and determining gene functions. Patterns of the functional units can provide important hints about gene function [27]. Identification of genome-wide patterns of TF binding sites can be done by high-throughput technologies including CHIP-chip [28] and CHIP-seq [29]. TF binding sites can also be predicted by in silico methods, mostly based on evolutionary conservation. TRANSFAC [30] and JASPAR [31] are two large TF binding site databases, consisting of both experimentally validated and putative evolutionarily conserved TF binding sites in eukaryotic genomes. In addition, though the functional role of repeat elements remains in speculation, a recent study found that in human genome, functionally similar genes are overrepresented among genes with similar repeat element profiles in the promoter region [32], suggesting that repeat elements information is worth continuing exploration for gene function prediction. RepBase [33] is the database storing repeat element information. The promoter region contains rich information responsible for regulatory role of gene function, and DBTSS (DataBase of Transcription Start Sites (TSS)) [34], which includes precise positional information for TSS and promoter region of the eukaryotic mRNA, can be useful for predicting gene function as well.

More and more evidence have shown that the majority of transcriptome consist of non-coding RNA transcripts [35]. Thus, RNomics, the study of the structure, function, and process of non-coding RNAs, is starting to attract more and more attention. Though the function of most non-coding RNAs remains mysteries, the discovery and extensive studies of microRNA or microRNomics have led to a new paradigm of gene regulation which takes place post-transcriptionally and pre-translationally [36]. MicroRNAs regulate the process of cell development, differentiation, proliferation, mobility, and apoptosis through the regulation of its target genes. Target genes regulated by the same microRNA may be involved in the same biological process. miRGen is a database that provides information about the miRNA target genes and their corresponding TF in human and mouse [37]. miR2Disease provides comprehensive information about human diseases associated with miRNA deregulation from literatures [38].

Besides genomics data, epigenomics that study the epigenetic changes, including DNA methylation and histone modifications, across the entire genome can provide important insights about the function of genes as well [39]. Epigenetics changes can lead to activation or inactivation of genes, and play important roles in cell development, differentiation and tumorigenesis. The DNA Methylation Database (MethDB) [40] and Human Histone Modification Database (HHMD) [41] contain information about DNA methylation and histone modification in human genome, while the ChromatinDB [42] database contains genome-wide ChIP data for histone modifications in yeast genome. With more and more experimental data becoming available, mining epigenomics data will provide a novel approach to predict gene functions.

Transcriptomics

The transcriptome represents the complete set of RNA transcripts in the cell [43]. Both the expression and abundance of RNA transcripts can change in response to cellular development, physiological and environmental condition changes. The microarrays and serial analysis of gene expression (SAGE) represent the most well-used technologies to study transcriptome [44]. Recently, deep sequencing RNA transcripts using the next-generation sequencing technologies has detected RNA transcripts at single base resolution, allowing for the discovery of novel transcripts that cannot be detected with traditional technologies [45]. Transcriptomics data are invaluable to understand gene functions. By focusing on differentially expressed genes under different development stages, one may identify genes responsible for the biological process governing cellular development. In addition, genes with correlated expression patterns under different conditions are likely functionally related [46]. Gene Expression Omnibus (GEO) is the largest public repository of transcriptomics data [47]. It currently contains more than 400 thousands samples submitted from a wide range of platforms on many organisms, and this number is increasing every day. In addition, there are species-specific expression databases, such as GXD for mouse [48], FLIGHT for fly [49], and GeneVestigator [50] for Arabidopsis; tissue specific expression databases, such as BGED for brain [51], and BloodExpress for blood [52]; pathway specific expression databases, such as GermOnline for germ line development [53] and Cyclebase for cell cycle process [54]. The vast amount of transcriptomics data under a wide range of conditions makes mining of transcriptomics data an active field for gene function prediction.

Proteomics

Proteins are the main components of the metabolic pathway, and many proteins interact with each other either in a complex or transiently to function in a biological process. Proteome is the complete set of proteins encoded in the genome. Proteomics is the large-scale study of proteome, focusing on the post-translational modifications of proteins, protein abundance, protein variants, and protein-protein interactions [55]. Depending on the environmental and cellular physiological conditions, proteome may vary significantly from one cell or condition to another. Protein abundance may not be inferable from RNA expression, due to post-transcriptional regulation. Proteins are also subject to post-translational modifications, such as phosphorylation, glycosylation, and acetylation, which are critical for some proteins to be functional.

The most widely used proteomics techniques are two-dimensional gel electrophoresis [56] and mass spectrometry [57]. Both can identify and quantify cellular proteins. New technologies, such as shotgun proteomics, promise to significantly improve the accuracy and coverage of proteome detection [58]. Latest technologies to determine post-translational modifications of proteins include PROTOMAP which combines SDS-PAGE with shotgun proteomics [59]. Databases of post-translational modifications include dbPTM [60], an integrated database containing information about protein phosphorylation, glycosylation and sulfation, etc. Protein subcellular localization is one of the three ontologies of gene functions in GO. There are several species-specific databases of subcellular location, e.g., YPL.db for yeast [61] and Flytrap for Drosophila.

Interactomics is the study of all protein physical interactions in the cell. In a broad sense, the interaction can be extended to refer to the interaction between protein and DNA or RNA, or the genetic interactions between proteins as well. High-throughput interaction technologies include yeast two-hybrid system [62] and tandem Affinity purification followed by mass spectrometry (TAP) [63], etc. Genome-wide protein-protein interactomes have been reconstructed in several model organisms, including yeast [64], worm [65], and human [66]. A number of interaction databases have been established, including BIOGRID [67], MIPS [68], IntAct [69], MINT [70], DIP [71] from published literatures, PIP [72] and OPHID [73] from computational predictions, and the integrated databases, such as BIND [74], HPRD and STRING [75]. Technologies detecting protein-DNA and RNA interactions include Protein-chip [76]. BIND [74] and ENCODE [77] databases contain information about protein-DNA interactions. Genetic interactions can be captured by synthetic genetic array (SGA) [78], diploid-based synthetic lethality analysis with microarrays (dSLAM) [79], synthetic dosage-suppression and lethality and haploinsufficiency [80]. BIOGRID [67] database contains genetic interactions from literatures.

Metabolomics

Metabolomics is the study of small chemical metabolite in the cell. Enzymes are the major components of metabolism that catalyze to convert or give rise to metabolites. In response to change of environmental and cellular condition, the gene expression, translation, and catalytic activity of enzymes can change, which can lead to the change of metabolite profiles. Small metabolites in turn can play important regulatory roles in gene expression, translation, and the biological processes. Therefore, it is necessary to integrate transcriptomics, proteomics and metabolomics data in the same context, in order to obtain a complete picture of gene functions. High-throughput metabolomics technologies include gas chromatographic mass spectrometry (GC/MS) [81], liquid chromatographic mass spectrometry (LC/MS) [82], as well as nuclear magnetic resonance (NMR) [83]. Examples of Enzyme databases include BRENDA (BRaunschweig ENzyme DAtabase) [84] that contains information about classification, nomenclature, reaction, specificity and many features of enzymes. Metabolic pathway databases include KEGG (Kyoto Encyclopedia of Genes and Genomes) [85], MetaCyc [86], and EcoCyc [87].

Phenomics

A phenotype is an observable characteristic of a cell or an organism. It is the consequence of genome, transcriptome, proteome, and metabolome combined. It can be the morphology, development state, biochemical property, physiological condition, or reaction to the external environment. Phenomics, which associates the phenotype with the genotype, investigates genome-wide phenotypic manifestations at cellular and organism level. High-throughput phenotyping (HTP) is critical to phenomics. Current technologies include genome-scale RNAi screens for knock down analysis and phenotype microarray for simple assessment of microbe growth capability. Further advances in experimental technologies and computational algorithms are needed to speed up the phenomics studies. The Online Mendelian Inheritance in Man (OMIM) database has the largest collection of human genotype-disease information [88]. The online Mendelian Inheritance in Animals (OMIA) provides genotype-disease information in animals [89]. PhenomicDB [90] and GeneCards [91] databases provide heterogenous phenotypic information from a number of different model organisms. Phenotype Ontology systems are being developed to store, organize, and manage phenotype in a structured way, similar to that in GO. Mouse Phenotype Ontology (MPO) [92] and PhenoGO [93] provide such framework. Phenotype has been used for gene function prediction. Philip et al. cluster genotype-phenotype data, and assign the overrepresented functions in the cluster to the known gene [94].

In summary, omics data ranging from genomics, transcriptomics, proteomics, metabolomics, to phenomics, are being generated at an unprecedented pace, providing us with tremendous opportunities to tackle the biologically important questions at a whole new level. However, the complexity, heterogeneity, and scale of omics data present significant challenges to the biology community as well. Developing a standard procedure to store, manage, and share omics data is being strongly advocated [95, 96]. The establishment of a common standard will greatly facilitate the process to design better strategies to mine and integrate the omics data.

Computational Algorithms to Integrate Omics Data for Gene Function Prediction

Many computational algorithms have been developed to predict gene functions from omics data. As the omics era starts with completely sequenced genomes, early efforts on algorithm development focused on exploring genomics data for gene function prediction. With diverse sets of omics data introduced by high-throughput technologies continuously emerging, the current focus of the gene function prediction field has switched to omics data integration. Because of the high complexity, heterogeneity, and large-scale of the omics data, it is often difficult to design the integration rules beforehand. Machine learning or statistical algorithms are frequently used to learn from and integrate the complex data to make predictions. Recently, interaction networks or broadly speaking, functional linkage networks, have been used to integrate omics data. In this section, we first briefly summarize sequence-based gene function prediction methods. Then, we introduce several machine learning and statistical algorithms for omics data integration. Finally, we describe in detail the network-based integration, by introducing the construction of functional linkage network and the exploration of network topology for gene function prediction.

Sequence-Based Algorithms for Gene Function Prediction

Most sequence-based gene function prediction methods are based on a simple assumption, i.e., function tends to be conserved among evolutionary related sequences. Thus, detecting evolutionary relationships is a critical step, which is often done by a database search for homologous sequences with powerful tools, such as PSI-BLAST [2]. Function of an unknown gene can be predicted if it is found to share a significant sequence similarity with a known gene. However, this approach is often unreliable, especially for inference of specific functions [3, 6]. For example, systematic analysis of enzyme function inference using homology-based methods reveals that on average, above 60% sequence identity is required for accurate enzyme function inference [6]. With such a restrictive cut-off, however, a significant amount of false negatives would be produced. Modifications of sequence-based methods have been made and achieved significant improvement, including those by distinguishing orthology from paralogy [22, 97], those by inspecting phylogenetic profile information [98], and those by focusing on the functionally important residues in the sequences [5], etc. The sequence-based methods mostly focus on predicting the molecular function aspect of genes. Recently, Hawkins and Kihara investigated the association relationships between different GO terms [99]. They built a Function Association Matrix (FAM) between GO terms from different GO categories. By considering the FAM and PSI-BLAST hit, their PFP algorithm can make predictions of GO terms beyond the molecular function terms. In addition to sequence information, three-dimensional structural information of proteins has also been extensively explored for predicting gene functions [100–102].

Non-network Based Omics Data Integration for Gene Function Prediction

The omics data type can be very different from each other. For example, gene expression is represented by a real value, while a sequence pattern is a binary value, either “present” or “absent” in a gene, and a phenotype can be a categorical value, e.g., “normal”, “sick”, “very sick”. Some machine learning algorithms, such as neural network and Support Vector Machine (SVM), are flexible to the format of the input data. For simplicity, however, the real value and the categorical value can be transformed into binary values. For example, gene expression value can be divided into several bins, with each bin considered as a new feature. After the appropriate coding systems of the omics data are decided, gene function prediction can then be considered as a binary classification problem, for which many machine-learning algorithms are available. Popular machine-learning algorithms include SVM, Bayesian Network (BN), Decision Tree (DT), Neural Network (NN), and so on. Here, we briefly introduce these algorithms, and then focus on examples of using them for omics data integration.

SVMs represent a family of statistical machine-learning methods that aim to optimally separate data into two categories by drawing a hyperplane in an N-dimensional vector space [103]. BN is a representation of a joint conditional probabilistic distribution that encodes the probabilistic relationships among features of interest [104]. DT is essentially a series of questions from which the classification or probability of a gene having a given function can be inferred [105]. NN mimics the human neuron perception system by consisting of a large number of highly interconnected elements to solve a problem [106]. Some of the algorithms can provide the rules of how a prediction is made, making it easier for human to understand, such as BN and DT, while others act like a “black box”, such as NN. Yet, all these algorithms have been successfully applied in predicting gene functions.

Pavlidis and coworkers used a kernel-based SVM to combine gene expression profile and phylogenetic profile to infer yeast gene MIPS function categories [107]. Rather than simply concatenating both expression and phylogenetic profiles into a vector space, they used two kernel functions to transform the data into a higher dimension space separately. The new kernels were trained by SVM, with the results simply combined to make a final prediction. Lanckriet and coworkers further improved the kernel-based SVM to combine protein complex, protein domain, protein-protein interaction, genetic interaction, and gene expression information [108]. Instead of simple addition, a weighted linear combination was implemented to combine the results from each kernel. Troyanskaya and coworkers developed a BN-based algorithm named MAGIC to predict functional linked gene pairs from genetic and physical interactions, microarray, and transcription factor binding sites data [109]. Because learning the conditional probability in the BN structure is not an easy task, the authors consulted experimental experts and designed an expert-BN reflecting relationship between different evidence. The results from the BN integration were superior to unsupervised clustering algorithms significantly. Zhang and coworkers used a probabilistic DT to predict co-complex protein pairs from mRNA expression, transcription regulator, subcellular localization, phenotype and some sequence features [110]. Unlike BN, the DT does not rely on any previous assumption about conditional dependence; it automatically weights each data type when building tree. King and coworkers used the DT to make prediction of gene functions from patterns of annotation, and compare the result with that done by BN [111]. The result showed that DT is comparable to BN and in some cases better. NN has been widely used in biological data analysis. Jensen and coworkers developed a NN to predict protein function from various types of predicted potein features, including post-translational modification, sub-location and sorting [112]. Mateos and coworkers used a NN to predict gene function from gene expression data [113]. In addition, they pointed out that the poor performance of machine learning can be attributed to incomplete protein function annotations.

The algorithms introduced above employ a single model to integrate omics data. Multiple models can also be applied. Then, a new model is used to combine the prediction results. Hibbs and coworkers employed three different algorithms, bioPIXIE, MEFIT and SPELL, to predict genes involved in the process of mitochondrion organization and biogenesis [114]. bioPIXIE is a BN model aiming to integrate diverse sets of omics data. MEFIT focuses on integration of only microarray data. SPELL focuses on identifying coexpressed genes associated with the target biological process. The results of the three algorithms were combined with different weights determined based on their association with functional relationships. The combined algorithm achieved better performance than any single classifier did. Tian and coworkers developed a combined algorithm named Funckenstein which has two component classifiers [8]. The two classifies use different sets of omics data to predict gene function independently. A regression model is used to combine the results from these two classifiers. We will describe this algorithm in detail in the third section.

Network-Based Omics Data Integration for Gene Function Prediction

The wide use of high-throughput interaction technologies has allowed for the reconstruction of genome-scale protein physical interaction network in several organisms [64–66]. Extensive studies have been conducted on the interaction network, including using it to integrate omics data and for gene function prediction. In protein interaction network, the nodes are genes, while the edges are protein physical interactions (PPI). The edge can be any sorts of functional relationships as well, including genetic interaction, correlated gene expression, homologous relationship, etc. Thus, the network can be conveniently used as a framework to integrate various sources of omics data. The integrated network is often called functional linkage network (FLN) to indicate the functional links between genes. In addition, the network structure can be explored to obtain more information for gene function prediction. Here, we first introduce the reconstruction of FLN for omics data integration and gene function prediction. Then, we review current algorithms available to explore network structure, in particular the network module, for gene function prediction.

The concept of FLN was first introduced by Marcotte et al. in 1999 [115]. In their work, the functional links between proteins were constructed by combining protein-protein links from various sources: experimentally derived PPI, correlated gene expressions, related domain fusion, correlated phylogenetic profiles, and related metabolic function. Different evidences were simply combined without weight. High confidence protein links were defined as those with more than two evidences. Marcotte group further extended the idea of functional linkage by introducing a probabilistic FLN in yeast genome [116]. They computed a likelihood score of whether a pair of genes has a functional linkage defined by a common KEGG pathway given the evidence. The final FLN was a result of the integration of eight types of omics data, including physical interactions, genetic interactions, mRNA coexpression, functional linkages from literature mining, and computational linkages from gene-fusion and phylogenetic profiles. The resulted functional linkages showed a comparable accuracy in predicting KEGG pathway relationships to that by protein-protein interactions determined by small-scale experiments. Linghu and coworkers employed machine-learning algorithms to automatically integrate five types of omics data: PPI, genetic interaction, expression data, sequence similarity, phylogenetic profile and domain fusion to generate a FLN in yeast genomes [117]. The functional linkage was defined as the presence in the same KEGG pathway. Then, they designed a decision rule to infer protein pathway function from the FLN. Karaoz and coworkers constructed a FLN by using protein-protein interactions as the edges, with the weight determined by the correleated expression value of the interacting genes. A GAIN (Gene Annotation using Integrated Network) algorithm was then used to predict protein functions, by systematically propagating the labels of genes with known GO terms to unlabelled genes across the FLN [118]. Tian and coworkers applied a probabilistic decision tree (PDT) to construct FLN from various sources of experimentally determined protein physical and genetics interactions, and use this FLN to predict candidate gene with specific function annotations [8]. Reconstruction of FLN can also be found in other recent works [119, 120].

Besides integrating multiple sources of omics data into a single FLN, multiple FLNs can also be constructed. The final results can be either from the integration of the result from individual FLN, or from a new FLN integrated from multiple FLNs. For example, GeneMANIA [121], an algorithm developed by Mostafavi and coworkers, first builds multiple FLNs from various sources of omics data. Then, it employs a fast heuristic algorithm derived from linear regression to integrate multiple FLNs into a composite FLN. Finally, it applies a Gaussian field label propagation algorithms to predict gene function from the composite FLN. This algorithm was ranked one of the best methods in predicting gene function in the first critical assessment of mouse gene function based on the evaluation measurement of area under the ROC [122].

Given an interaction network or a FLN, network information can be explored to assist in the prediction of gene functions. The approaches exploring network information can be generally classified into two categories: the direct approach and the module-assisted approach. The direct approach utilizes the local or global network information to predict function. The module-assisted approach is inspired by the observation that interacting or functional linked genes tend to be localized in a dense region in the network, i.e., module [123]. It involves two steps: the first step is to identify the module, and the second step is to predict the function of unknown genes based on the distribution of known genes present in the same module. Here, we introduce the algorithms for both approaches.

The simplest method of the direct approach is the neighbor counting method. For example, Schwikowski et al. counted the neighbor proteins of an unknown protein, and simply assigned the three most frequent functions of the known neighbor proteins to the unknown protein [124]. Hishigaki et al. implemented a χ^2 test for the enrichment of known functions among the neighbor interacting proteins, and assigned the statistically significant functions to the known [125]. Further optimization was done by considering not only the direct interacting proteins, but also the near-neighbor proteins and their distances in the network graph [126]. These methods consider the local information and employ simple statistical test to make predictions. More sophisticated models that consider the global network information have also been developed, including the graph theory based methods. Graph theorey-based methods take the global and full topology of the network into account and employ either a cut- or flow-based algorithm to assign function, which can be generalized as a minimum multi-way cut problem. Vazquez et al. applied this theory to the yeast protein physical interaction network to predict functional class of unknown proteins, by minimizing the number of protein interactions among different functional categories with simulated annealing [127]. In contrast to Vazquez's approach that considers multiple functions at once, Karaoz et al. handled one function at a time, and employ a propagation algorithm to allow the flow of functional information in the network, and assign a score to candidate genes of having the function. Other attractive methods include the Markov Random Field (MRF) theory-based method, which assumes the function of a protein is dependent only on its neighbors and independent of all other proteins. Deng at el was the first group to formalize the idea of MRF in predicting protein function from protein interaction network [128]. Their approach was further generalized by allowing for the use of multiple networks, such as protein physical interaction, genetic interaction and coexpression network [129]. The MRF model is based on a sophisticated statistical theory, and mathematically sound. However, a number of machine learning algorithms have been reported to outperform the MRF model with the same benchmark data used by Deng et al. [8, 130].

The module-assisted approaches involve the identification of modules or dense local structure in the network, which was originally proposed by HartWell [123]. A number of algorithms have been developed for module identification, which can generally be classified as clustering based and non-clustering based. Clustering based methods include algorithms based on the pairwise distance of protein pairs defined as the shortest path length in the network [131], or more sophisticated ways, e.g., using the graph theory. Spirin and Mirny developed two algorithms, SPC (superparamagnetic) and a Monte Carlo-base method, to maximize the density of the obtained clusters [132]. Bader and Hogue developed a molecular complex detection algorithm (MCODE) to isolate the dense regions into modules [133]. The MCODE consists of three steps: vertex weighting based on the core clustering coefficient, prediction of complex memberships, and an optional post-processing filtering or addition of proteins based on connectivity data. Sharan et al. developed a NetworkBlast algorithm to assign a likelihood ration score for each candidate set of proteins in the network [134]. This method uses a greedy network search algorithm and can identify conserved region over several networks. The non-clustering based methods involve the use of prior information about protein-protein interaction or complex information. This information is used to seed a module, which is then expanded based on network connectivity. The Complexpander software developed by Asthana et al., first produces a rank of core proteins from complex data; then, it assigns a probability to the involvement of each protein in the core, and then computes a weighted score for each pair of proteins in the end [135]. Information other than protein physical interactions can also be utilized to identify network modules. For example, Segal et al. proposed a probabilistic model to identify modules not only enriched for interactions, but also enriched for high sequence similarity [136]. Hanisch et al. used the expression information as a filtering process [137], while Tanay et al. integrated the PPI data with gene expression, phenotypic sensitivity and TF binding site, to identify modules [138]. Once the modules are identified, usually statistical tests of the enrichment of known functions are conducted to infer function of the unknown proteins.

Funckenstein, a Combined Algorithm for Omics-Based Gene Function Prediction

Having described various types of omics data and a number of algorithms available for predicting gene function by integrating omics data, here we use a combined algorithm named Funckenstein [8] as an example to further illustrate the process of integrating omics data for gene function prediction. Most algorithms described in the previous section can generally fall into two categories: the "guilt-by-profiling" approach and the "guilt-by-association" approach. The "guilt-by-profiling" approach focuses on mining the gene characteristics, e.g., a conserved sequence motif. The "guilt-by-association" approach explores the relationships between genes for functional association, e.g., orthologous relationship, correlated expression profile, etc. Either approach has its own merit. Funckenstein is an algorithm that combines both approaches to achieve a synergistic performance better than either approach alone does. It has been applied for predicting gene functions (GO) in both yeast and mouse genomes [8, 139].

There are three steps in Funckenstein (see Fig. 2 for the flow chart). The first step is to classify omics data. Following the definition of guilt-by-profiling and guilt-by-association approaches, the collected diverse sources of omics data are classified into two categories: one describing gene characteristics, and another

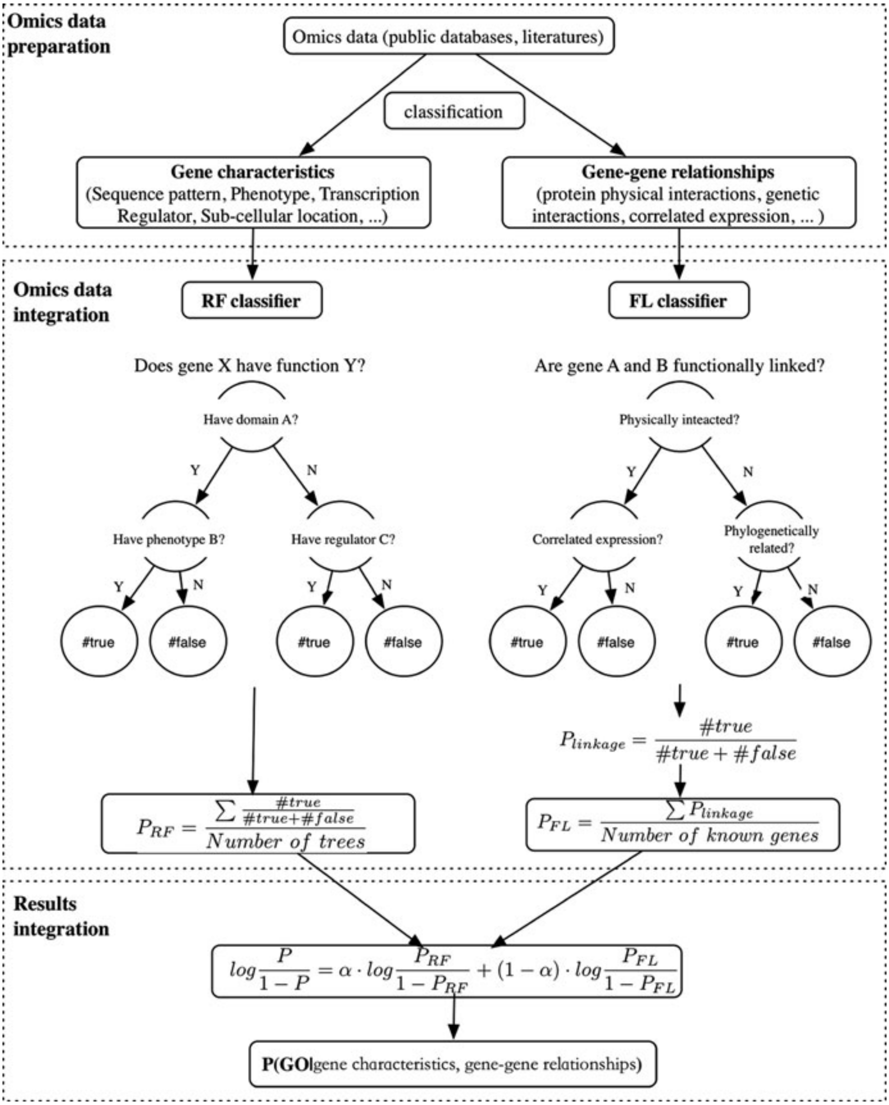
Omics data preparation
Omics data (public databases, literatures)
classification
Gene characteristics (Sequence pattern, Phenotype, Transcription Regulator, Sub-cellular location, ...)
Gene-gene relationships (protein physical interactions, genetic interactions, correlated expression, ...)
Omics data integration
RF classifier
FL classifier
Does gene X have function Y?
Have domain A?
Y
N
Have phenotype B?
Have regulator C?
#true
#false
Are gene A and B functionally linked?
Physically inteacted?
Correlated expression?
Phylogenetically related?
$P_{linkage} = \frac{\#true}{\#true + \#false}$
$P_{RF} = \frac{\sum \frac{\#true}{\#true+\#false}}{Number\ of\ trees}$
$P_{FL} = \frac{\sum P_{linkage}}{Number\ of\ known\ genes}$
Results integration
$log\frac{P}{1-P} = \alpha \cdot log\frac{P_{RF}}{1-P_{RF}} + (1-\alpha) \cdot log\frac{P_{FL}}{1-P_{FL}}$
P(GO|gene characteristics, gene-gene relationships)

Fig. 2 Flowchart of the Funckenstein algorithm. There are three steps in Funckenstein: omics data preparation, omics data integration, and results integration. RF and FL refer to the random forest and functional linkage classifiers, respectively. The decision tree under the RF is an example of many decision trees in the forest, while that under the FL is an example of 12 decision trees specific for different GO categories

describing gene-gene relationships. Take yeast gene function prediction for an example, the gene characteristics include protein sequence patterns, gene phenotypes, the common transcriptional regulators, protein sub-cellular localization, and protein complex memberships. Some of those characteristics were collected from databases, such as UniProt database, while others were obtained from the supplementary materials of the published literatures. The gene-gene relationships include various types of protein-protein interactions (both physical and genetics) determined by different experimental technologies, which were downloaded from the BIOGRID database directly. In the second step, two component classifiers of Funckenstein (the random forest (RF) and the functional linkage (FL) classifier) are trained to make predictions from the gene characteristics and gene-gene relationships, respectively. The RF classifier employs a random forest algorithm [140] to build hundreds of decision trees from the gene characteristics. Each decision tree outputs a probability of a gene having a given function, which is then averaged across all decision trees. The FL classifier first builds a FLN from gene-gene relationships using a decision tree. Then, it computes the functional linkage score of a query gene with the genes known to have the function, which are then averaged to output a probability of the query gene having the function. In the final step of Funckenstein, a regression model is implemented to combine the probability scores from both the RF and FL classifiers and output the final probability.

There are several things about Funckenstein that need attention. First of all, Funckenstein predicts each GO term independently, i.e., the parent-child GO term relationships are not considered. Secondly, Funckenstein does not allow GO term annotation to be used as a feature in the training to avoid the issue of circularity. Third, rather than building one FLN, Funckenstein builds 12 FLNs by considering the type of ontology, i.e., Molecular Function, Biological Process and Cellular Component, and the specificity of GO terms which is defined by the number of genes annotated with the GO term and ranges from 3 to 10, 11 to 30, 31 to 100, and 101 to 300, respectively. Fourth, when measuring the prediction performance, the area under the precision-recall curve instead of the ROC curve is used. ROC curve has been widely used as a measure of performance, which plots the true positive rate against the false positive rate [141]. In comparison, the precision-recall curve is the plot of precision against the true positive rate. Suppose the number of true positives, false positives, true negatives, and false negatives are TP, FP, TN and FN, respectively, then the true positive rate = TP/(TP+FN), the false positive rate = FP/(FP+TN), and the precision = TP/(TP+FP). When the number of real negatives, (FP+TN), is far more than the number of real positives, (TP+FN), the false positive rate can be very small, even though FP is much larger than TP. In that case, the predictions may not be useful to biologists. In fact, to most biologists, they may be concerned more with the positive predictions the computational biologists made than the negatives. In contrast, the precision-recall curve is independent of the number of real negatives, and is more intuitive to biologists. Accordingly, Funckenstein is optimized based on the area under the precision-recall curve.

Funckenstein has been benchmarked with the same dataset used by a previous integrated algorithm for yeast gene function prediction. That algorithm, developed by Deng et al., uses a Markov Random Field (MRF) to integrate protein-protein interaction, coexpression, and genetic interaction networks, and estimates the prior probability of a gene having a given function by a Naïve Bayes method from protein complex memberships [128]. Funckenstein outperformed this algorithm by a significant margin in predicting yeast gene MIPS functions [8]. In the first critical assessment of the mouse gene function prediction which was participated by nine leading groups in the omics-based gene function prediction field, on average, for most GO categories evaluated, Funckenstein outperformed all other groups in terms of the precision at 20% recall [122]. In sum, Funckenstein achieves state-of-the-art performance in integrating omics data for gene function prediction.

Here we'd like to describe several interesting points during the development of Funckenstein. First of all, to achieve best synergistic effects in performance, it is better to use as different omics data as possible to train the guilt-by-profiling and the guilt-by-association methods separately. For example, a sequence pattern can be considered as a gene characteristic, but it can also be used to link two genes that have the same pattern. In yeast gene function prediction, we tested to code gene characteristics as additional gene relationships to train the FL classifier. Although we could improve the performance of the FL classifier greatly with the new additions, the combined results were worth than before. This suggests that the same omics data should not be utilized more than once. Second, more interactions data can substantially improve the performance of the FL classifier and consequently that of Funckenstein. In the benchmark with Deng et al.'s dataset, there were only a few thousands interactions available; while in the BIOGRID database, there are nearly a hundred thousands interactions curated from various high-throughput studies. The relative contribution of the FL classifier to Funckenstein's performance is significantly increased in the latter benchmark. This suggests by adding more gene-relationships from new omics data, we could further improve Funckenstein's performance. Third, building a FLN helps the FL classifier play a bigger role in predicting specific gene functions. When a GO term is associated with only a few known genes, it is difficult to train from the "positive" samples. In contrast, the 'transfer rules' are learned from the many GO terms within the specific GO category in the FLN. This stresses the importance of reconstructing a FLN in predicting gene functions.

Current Limitations and Potential Improvements

Omics Data Are Not Thoroughly Used

Figure 3 shows the frequency of different types of omics data used in the published "gene function prediction" algorithms since 2001. It is apparent that protein sequence, gene expression, and protein-protein interaction are the dominant omics data for gene function prediction, with the rest of omics data seldom or not used. For the three most used types of omic data, protein-protein interaction and gene expression data are becoming the focus in current algorithm development, which is

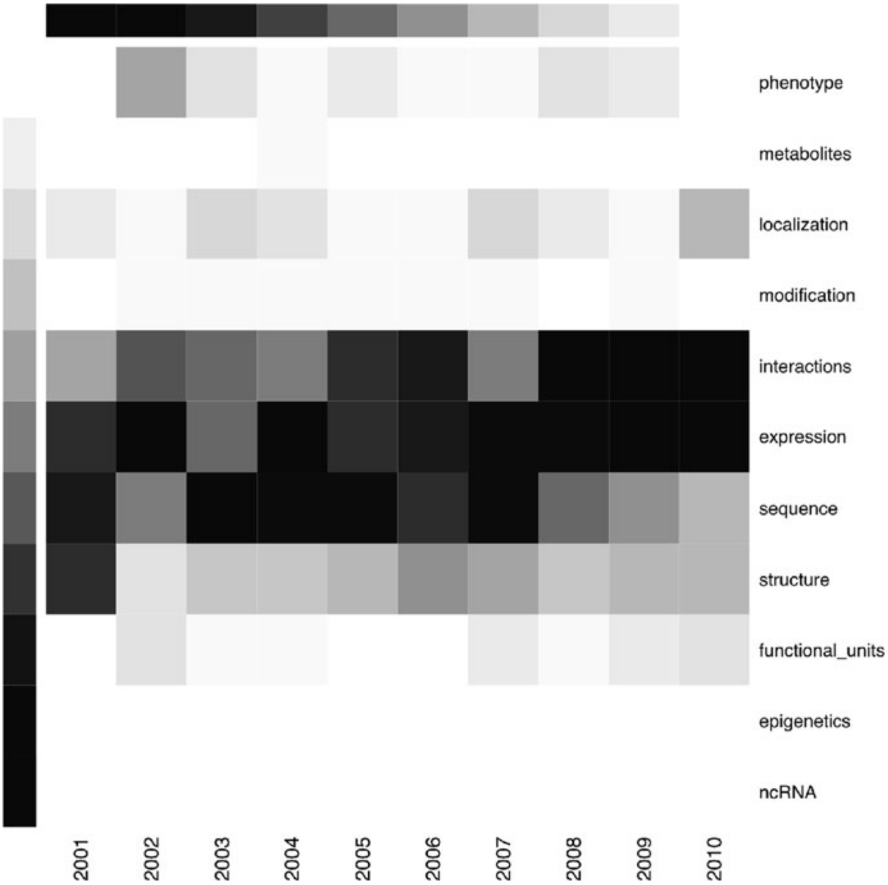

Fig. 3 Heat-map of the number of published gene function prediction algorithms using different types of omics data from 2001 to 2010. A Pubmed search with the "eutils.pl" script obtained from NCBI using different synonyms of "gene function prediction" from 2001 to 2010 results in over 800 literatures. Synonyms corresponding to different types of omics data are then used to count the number of publications using the corresponding omics data each year. The number is plotted in the heatmap. The *blackness* of each square in the heatmap represents the relative frequency of the corresponding publications each year

consistent with the trend that microarray and two-hybrid high-throughput technologies are becoming widely used. The lack of use of other omics data by current algorithms can be attributed to the fact that some omics data are not abundant enough. For example, the phenomics data are still lacking because developing an efficient high-throughput screen for phenotypic change is not an easy task. However, even the genomics data are not fully used. For example, although some algorithms integrate the TF binding site information, only the presence or absence information of the TF binding sites is used. In fact, the combination of TF binding sites, its relative position and the number of occurrence of TF sites in the promoter region all contribute to the target gene functions. In addition, the 5′ UTR and 3′ UTR of the target gene may also contain important functional unit information necessary for the function of the genes. Therefore, a more thorough use of omics data should be done in order to make further improvements. On the other hand, the metabolite, non-coding RNA, and epigenetics information are completely ignored by the current algorithms, which also points out where a potential improvement of the current algorithms can be made.

Omics Data Sharing Is Urgent and Needs to Be Standardized

Another reason why the omics data are not thoroughly used by current algorithms is because of the problem of omics data sharing. Although we have listed a large number of databases storing specific omics data in the first section, these databases may not be updated frequently enough to include the most recent high-throughput studies. In those cases, computational biologists often have to collect a large fraction of omics data from the supplementary of the published literatures by themselves, which is very time-consuming and laborious. In some cases, it may deter computational biologist from using the data. For example, the gene-naming system is often inconsistent from one high-throughput study to another, making automatic cross comparison almost impossible. With more and more omics data accumulated, this issue has become so serious that a number of algorithms for gene name translation have been published lately [142, 143]. In addition, the omics data are often lack of appropriate annotation, making it difficult for computational biologists to use or to interpret the results. With large amount of omics data being generated every day, standardization of omics data for sharing has never been so urgent. The advocate for a guideline like Minimum Information Requested In the Annotation of biochemical Models (MIRIAM) for omics data sharing is becoming louder than ever [95, 144]. The establishment and enactment of such a common standard for omics data sharing will greatly facilitate the improvement of current algorithms.

Omics data sharing is also an issue among computational biologists. A common benchmark omics dataset is important for computational biologists to test their algorithms and compare with others, so that they can make proper improvement. However, most times the benchmark omics dataset used by one algorithm is not accessible to others. CASP (Critical Assessment of Techniques for Protein Structure Prediction) has been successfully conducted for evaluating protein structure prediction methods [145]. A similar project can be extremely useful to the gene function prediction community. The first critical assessment of mouse gene function prediction project (MouseFunc) has been conducted [122], and more such project should follow. However, unlike protein structure prediction which can be compared with an experimentally determined structure, function is difficult to measure in an objective and timely manner, making effective benchmark for function prediction comparison not an easy task.

Is a Complex Model Better than a Simple Model?

The network-based data integration for gene function prediction has attracted the attention of many computational biologists. Various sophisticated algorithms have been developed to explore global network information, including those based on graph theory, and those based on identification of network modules. However, Murali et al. found that a simple local guilt-by-association method outperforms a graph-theory based global method to predict gene function from protein interactions [146]. In addition, Song and Singh recently tested the efficacy of various clustering algorithms in clustering protein interaction networks and predicting protein function [147]. They also compared the clustering algorithms with a simple guilt-by-association algorithm based on neighbor counting. Surprisingly, the simple guilt-by-association algorithm outperformed the sophisticated clustering algorithms in predicting gene functions. This thus raises an interesting question: Is a complex model better than a simple model?

The sophisticated algorithms are often backed by strong mathematics and statistics theories, while a simple model is usually based on empirical observations. However, the sophisticated algorithms often have to make an assumption that the current knowledge about the protein interaction network is complete, which is usually not the case. Take protein interaction network for an example, the interaction network is reconstructed by collecting interactions from various experiments and literatures; i.e., it is an ensemble of protein interactions all kinds of cellular conditions. However, in reality, it is unlikely that all protein interactions in the network are present in the cell at the same time. For example, protein A interacts with both B and C according to current knowledge. But it is possible that B and C may be expressed at different developmental stages. In such case, the presumed information flow from B→A→C or from C→A→B based on network structure is not be true. Accordingly, the label propagations based on network structure would lead to the wrong answer. Therefore, it is not that a simple model is better than a complex model; instead, it is whether a complex model is applicable to the omics data.

Model Driven or Biology Driven?

We have described many machine-learning and statistical algorithms for omics based gene function prediction. A beginner may be confused of which algorithm to choose. Should he choose SVM, BN, DT, . . ., or RF? In fact, before any model is applied, the raw omics data has to be pre-processed or selected. Different groups may use different tricks to treat the raw omics data, which would lead to different outcomes. Take Funckenstein for an example, it classify the omics data into gene characteristics and gene-gene relationships categories before the application of the

RF and FL classifiers. This classification is critical to Funckenstein's success, as can be shown in yeast gene function in which the performance is worse without such classification. But how to process the raw omics data? The rational behind Funckenstein's classification is that the biological function of a gene is not only determined by its sequence, but also by what other genes it "interacts" with. As we can see from Funckenstein, perhaps a thorough understanding of the biology behind omics data and make appropriate treatment of omics data may be more effective than trying out a different model.

Prospective of Future Directions

Non-coding RNA Function Prediction

With more omics data emerging, the future of gene function prediction field will be continually focused on integrating newly added the data. However, coding gene sequence only accounts for a tiny fraction in the genome, while current results have shown that more than 70% of the genome are transcribed, with most of the transcripts being non-coding RNA [35]. The important biological role of non-coding RNA in the cell needs to be investigated. Many algorithms have been dedicated to coding gene function prediction. With the development of non-coding RNA experimental technologies, the next wave of gene function prediction will be the omics driven non-coding RNA function prediction.

Gene Function in a Dynamic Context

Gene Ontology provides an excellent system to describe the functions of a gene at three aspects, i.e., molecular function, biological process, and cellular component. However, these definitions do not take the dynamic cellular environment into account. Take catching a terrorist as an example, it is important to know what and where he is going to take actions. But it will be even more useful if we know when he is going to take actions. Similarly, besides knowing that two proteins interact with each other, it would be more interesting for biologists to know at what developmental stage, or by what environmental stimuli, they will interact with each other? Therefore, put gene functions in a dynamic context should be one of the most important and challenging directions in the future.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant No. 30971643), and Shanghai Pujiang Program (Grant No. 09PJ1401000) to WT.

References

1. Adams, M., Kelley, J., Gocayne, J., Dubnick, M., Polymeropoulos, M., Xiao, H., Merril, C., Wu, A., Olde, B., Moreno, R. Complementary DNA sequencing: expressed sequence tags and human genome project. Science **252**(5013): 1651 (1991).

2. Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**(17): 3389 (1997).
3. Rost, B. Enzyme function less conserved than anticipated. J. Mol. Biol. **318**(2): 595–608 (2002).
4. Sonnhammer, E., Eddy, S., Durbin, R. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins Struct. Funct. Genet. **28**(3): 405–420 (1997).
5. Tian, W., Arakaki, A.K., Skolnick, J. EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference. Nucleic Acids Res. **32**(21): 6226–6239 (2004).
6. Tian, W., Skolnick, J. How well is enzyme function conserved as a function of pairwise sequence identity? J. Mol. Biol. **333**(4): 863–882 (2003).
7. Hawkins, T., Kihara, D. Function prediction of uncharacterized proteins. J. Bioinform. Comput. Biol. **5**(1): 1–30 (2007).
8. Tian, W., Zhang, L., Ta an M, Gibbons, F., King, O., Park, J., Wunderlich, Z., Cherry, J., Roth, F. Combining guilt-by-association and guilt-by-profiling to predict Saccharomyces cerevisiae gene function. Genome Biol. **9**(Suppl 1): S7 (2008).
9. Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J. Gene ontology: tool for the unification of biology. Nat. Genet. **25**(1): 25–29 (2000).
10. Fleischmann, R., Adams, M., White, O., Clayton, R., Kirkness, E., Kerlavage, A., Bult, C., Tomb, J., Dougherty, B., Merrick, J. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science **269**(5223): 496 (1995).
11. ConsortiumInternational, H. G. S. Finishing the euchromatic sequence of the human genome. Nature **431**(7011): 931–945 (2004).
12. Rothberg, J., Leamon, J. The development and impact of 454 sequencing. Nat. Biotechnol. **26**(10): 1117–1124 (2008).
13. Oliphant, A., Barker, D., Stuelpnagel, J., Chee, M. BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. Biotechniques (JUN): 56–61 (2002).
14. Hultman, T., Stahl, S., Homes, E., Uhlen, M. Direct solid phase sequencing of genomic and plasmid DNA using magnetic beads as solid support. Nucleic Acids Res. **17**(13): 4937 (1989).
15. Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., et al. The diploid genome sequence of an individual human. PLoS Biol. **5**(10): e254 (2007).
16. Cherry, J., Adler, C., Ball, C., Chervitz, S., Dwight, S., Hester, E., Jia, Y., Juvik, G., Roe, T., Schroeder, M. SGD: saccharomyces genome database. Nucleic Acids Res. **26**(1): 73 (1998).
17. Blake, J., Richardson, J., Bult, C., Kadin, J., Eppig, J. MGD: the mouse genome database. Nucleic Acids Res. **31**(1): 193 (2003).
18. Rhee, S., Beavis, W., Berardini, T., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M. The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Res. **31**(1): 224 (2003).
19. Drysdale, R., Crosby, M. FlyBase: genes and gene models. Nucleic Acids Res. **33**(Database Issue): D390 (2005).
20. Sonnhammer, E.L., Koonin, E.V. Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet. **18**(12): 619–620 (2002).
21. Clark, A.G., Glanowski, S., Nielsen, R., Thomas, P.D., Kejariwal, A., Todd, M.A., Tanenbaum, D.M., Civello, D., Lu, F., Murphy, B., et al. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science **302**(5652): 1960–1963 (2003).
22. Tatusov, R., Galperin, M., Natale, D., Koonin, E. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. **28**(1): 33 (2000).

23. O'Brien K, Remm, M., Sonnhammer, E. Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res. **33**(Database Issue): D476 (2005).
24. Bowers, P., Pellegrini, M., Thompson, M., Fierro, J., Yeates, T., Eisenberg, D. Prolinks: a database of protein functional linkages derived from coevolution. Genome Biol. **5**(5): R35 (2004).
25. Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., De Castro, E., Langendijk-Genevaux, P., Pagni, M., Sigrist, C. The PROSITE database. Nucleic Acids Res. **34**(Database Issue): D227 (2006).
26. Attwood, T., Beck, M. PRINTS-a protein motif fingerprint database. Protein Eng. Des. Sel. **7**(7): 841 (1994).
27. Berman, B., Nibu, Y., Pfeiffer, B., Tomancak, P., Celniker, S., Levine, M., Rubin, G., Eisen, M. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc. Natl. Acad. Sci. USA **99**(2): 757 (2002).
28. Buck, M., Lieb, J. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics **83**(3): 349–360 (2004).
29. Schmid, C., Bucher, P. ChIP-Seq data reveal nucleosome architecture of human promoters. Cell **131**(5): 831–832 (2007).
30. Wingender, E., Dietze, P., Karas, H., Knüppel, R. TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res. **24**(1): 238 (1996).
31. Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W., Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. **32**(Database Issue): D91 (2004).
32. Huda, A., MariÒo-RamÌrez, L., Landsman, D., Jordan, I. Repetitive DNA elements, nucleosome binding and human gene expression. Gene **436**(1–2): 12–22 (2009).
33. Jurka, J. RepBase update: a database and an electronic journal of repetitive elements. Trends Genet. **16**(9): 418–420 (2000).
34. Suzuki, Y., Yamashita, R., Nakai, K., Sugano, S. DBTSS: database of human Transcriptional Start Sites and full-length cDNAs. Nucleic Acids Res. **30**(1): 328 (2002).
35. Guttman, M., Amit, I., Garber, M., French, C., Lin, M., Feldser, D., Huarte, M., Zuk, O., Carey, B., Cassady, J. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature **458**(7235): 223 (2009).
36. Bartel, D. MicroRNAs genomics, biogenesis, mechanism, and function. Cell **116**(2): 281–297 (2004).
37. Megraw, M., Sethupathy, P., Corda, B., Hatzigeorgiou, A.G. miRGen: a database for the study of animal microRNA genomic organization and function. Nucleic Acids Res. **35**(Suppl 1): D149–D155 (2006).
38. Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G., Liu, Y. miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res. **37**(Database issue): D98 (2009).
39. Bernstein, B., Meissner, A., Lander, E. The mammalian epigenome. Cell **128**(4): 669–681 (2007).
40. Grunau, C., Renault, E., Rosenthal, A., Roizes, G. MethDB – a public database for DNA methylation data. Nucleic Acids Res. **29**(1): 270 (2001).
41. Zhang, Y., Lv, J., Liu, H., Zhu, J., Su, J., Wu, Q., Qi, Y., Wang, F., Li, X. HHMD: the human histone modification database. Nucleic Acids Res. **38**(Suppl 1): D149–D154 (2009).
42. O'Connor T, Wyrick, J. ChromatinDB: a database of genome-wide histone modification patterns for Saccharomyces cerevisiae. Bioinformatics **23**(14): 1828 (2007).
43. Caron, H., Schaik, B., Mee, M., Baas, F., Riggins, G., Sluis, P., Hermus, M., Asperen, R., Boon, K., Voute, P. The human transcriptome map: clustering of highly expressed genes in chromosomal domains. Science **291**(5507): 1289 (2001).

44. Velculescu, V., Zhang, L., Vogelstein, B., Kinzler, K. Serial analysis of gene expression. Science **270**(5235): 484 (1995).
45. Jarvie, T. Next generation sequencing technologies. Drug Discov. Today Technol. **2**(3): 255–260 (2005).
46. Le Roch, K., Zhou, Y., Blair, P., Grainger, M., Moch, J., Haynes, J., De la Vega, P., Holder, A., Batalov, S., Carucci, D. Discovery of gene function by expression profiling of the malaria parasite life cycle. Science **301**(5639): 1503 (2003).
47. Edgar, R., Domrachev, M., Lash, A. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. **30**(1): 207 (2002).
48. Ringwald, M., Mangan, M., Eppig, J., Kadin, J., Richardson, J. GXD: a gene expression database for the laboratory mouse. The Gene Expression Database Group. Nucleic Acids Res. **27**(1): 106 (1999).
49. Sims, D., Bursteinas, B., Gao, Q., Zvelebil, M., Baum, B. FLIGHT: database and tools for the integration and cross-correlation of large-scale RNAi phenotypic datasets. Nucleic Acids Res. **34**(Database Issue): D479 (2006).
50. Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L., Gruissem, W. GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. Plant Physiol. **136**(1): 2621 (2004).
51. Kato, K., Matoba, R., Saito, S., Matsubara, K. BGED-Brain Gene Expression Database. http://genome.mc.pref.osaka.jp/BGED/index.html
52. Miranda-Saavedra, D., De, S., Trotter, M., Teichmann, S., Gottgens, B. BloodExpress: a database of gene expression in mouse haematopoiesis. Nucleic Acids Res. **37**(Database issue): D873 (2009).
53. Primig, M., Wiederkehr, C., Basavaraj, R., Sarrauste de Menthiere, C., Hermida, L., Koch, R., Schlecht, U., Dickinson, H.G., Fellous, M., Grootegoed, J.A., et al. GermOnline, a new cross-species community annotation database on germ-line development and gametogenesis. Nat. Genet. **35**(4): 291–292 (2003).
54. Gauthier, N., Larsen, M., Wernersson, R., de Lichtenberg, U., Jensen, L., Brunak, S., Jensen, T. Cyclebase org a comprehensive multi-organism online database of cell-cycle experiments. Nucleic Acids Res. **36**(Database issue): D854 (2008).
55. Gorg, A., Weiss, W., Dunn, M. Current two-dimensional electrophoresis technology for proteomics. Proteomics **4**(12): 3665–3685 (2004).
56. Raymond, S., Aurell, B. Two-dimensional gel electrophoresis. Science **138**(3537): 152 (1962).
57. Perkins, D., Pappin, D., Creasy, D., Cottrell, J. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis **20**(18): 3551–3567 (1999).
58. Wu, C., MacCoss, M. Shotgun proteomics: tools for the analysis of complex biological systems. Curr. Opin. Mol. Ther. **4**(3): 242–250 (2002).
59. Yona, G., Linial, N., Linial, M. ProtoMap: automatic classification of protein sequences and hierarchy of protein families. Nucleic Acids Res. **28**(1): 49 (2000).
60. Lee, T., Huang, H., Hung, J., Huang, H., Yang, Y., Wang, T. dbPTM: an information repository of protein post-translational modification. Nucleic Acids Res. **34**(Database Issue): D622 (2006).
61. Habeler, G., Natter, K., Thallinger, G., Crawford, M., Kohlwein, S., Trajanoski, Z. YPL. db: the Yeast Protein Localization database. Nucleic Acids Res. **30**(1): 80 (2002).
62. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. **98**(8): 4569 (2001).
63. Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M., Séraphin, B. The tandem affinity purification (TAP) method: a general procedure of protein complex purification. Methods **24**(3): 218–229 (2001).

64. Yu, H., Braun, P., Yildirim, M., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N. High-quality binary protein interaction map of the yeast interactome network. Science **322**(5898): 104 (2008).
65. Li, S., Armstrong, C., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P., Han, J., Chesneau, A., Hao, T. A map of the interactome network of the metazoan C. elegans. Science **303**(5657): 540 (2004).
66. Rual, J., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G., Gibbons, F., Dreze, M., Ayivi-Guedehoussou, N. Towards a proteome-scale map of the human protein®Cprotein interaction network. Nature **437**(7062): 1173–1178 (2005).
67. Stark, C., Breitkreutz, B., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. **34**(Database Issue): D535 (2006).
68. Mewes, H., Heumann, K., Kaps, A., Mayer, K., Pfeiffer, F., Stocker, S., Frishman, D. MIPS: a database for genomes and protein sequences. Nucleic Acids Res. **27**(1): 44 (1999).
69. Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A. IntAct: an open source molecular interaction database. Nucleic Acids Res. **32**(Database Issue): D452 (2004).
70. Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., Cesareni, G. MINT: a Molecular INTeraction database. FEBS Lett. **513**(1): 135–140 (2002).
71. Xenarios, I., Salwinski, L., Duan, X., Higney, P., Kim, S., Eisenberg, D. DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res. **30**(1): 303 (2002).
72. Yang, L., Jin, G., Zhao, X., Zheng, Y., Xu, Z., Wu, W. PIP: a database of potential intron polymorphism markers. Bioinformatics **23**(16): 2174 (2007).
73. Brown, K., Jurisica, I. Online predicted human interaction database. Bioinformatics **21**(9): 2076 (2005).
74. Bader, G., Donaldson, I., Wolting, C., Ouellette, B., Pawson, T., Hogue, C. BIND – the biomolecular interaction network database. Nucleic Acids Res. **29**(1): 242 (2001).
75. Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., Snel, B. STRING: a database of predicted functional associations between proteins. Nucleic Acids Res. **31**(1): 258 (2003).
76. Zhu, H., Snyder, M. Protein chip technology. Curr. Opin. Chem. Biol. **7**(1): 55–63 (2003).
77. Thomas, D., Rosenbloom, K., Clawson, H., Hinrichs, A., Trumbower, H., Raney, B., Karolchik, D., Barber, G., Harte, R., Hillman-Jackson, J. The ENCODE Project at UC Santa Cruz. Nucleic Acids Res. **35**(Database issue): D663 (2007).
78. Tong, A., Evangelista, M., Parsons, A., Xu, H., Bader, G., Page, N., Robinson, M., Raghibizadeh, S., Hogue, C., Bussey, H. Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science's STKE **294**(5550): 2364 (2001).
79. Pan, X., Yuan, D., Ooi, S., Wang, X., Sookhai-Mahadeo, S., Meluh, P., Boeke, J. dSLAM analysis of genome-wide genetic interactions in Saccharomyces cerevisiae. Methods **41**(2): 206–221 (2007).
80. Boone, C., Bussey, H., Andrews, B. Exploring genetic interactions and networks with yeast. Nat. Rev. Genet. **8**(6): 437–449 (2007).
81. Dauner, M., Sauer, U. GC-MS analysis of amino acids rapidly provides rich information for isotopomer balancing. Biotechnol. Prog. **16**(4): 642–649 (2000).
82. Jemal, M. High-throughput quantitative bioanalysis by LC/MS/MS. Biomed. Chromatogr. **14**(6): 422–429 (2000).
83. Laskowski, R., Rullmann, J., MacArthur, M., Kaptein, R., Thornton, J. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J. Biomol. NMR **8**(4): 477–486 (1996).
84. Schomburg, I., Chang, A., Schomburg, D. BRENDA, enzyme data and metabolic information. Nucleic Acids Res. **30**(1): 47 (2002).

85. Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F, Itoh, M., Kawashima, S., Katayama, T., Araki, M., Hirakawa, M. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. **34**(Database issue): D354–357 (2006).
86. Krieger, C., Zhang, P., Mueller, L., Wang, A., Paley, S., Arnaud, M., Pick, J., Rhee, S., Karp, P. MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res. **32**(Database Issue): D438 (2004).
87. Karp, P., Riley, M., Paley, S., Pellegrini-Toole, A., Krummenacker, M. EcoCyc: enyclopedia of Escherichia coli genes and metabolism. Nucleic Acids Res. **25**(1): 43 (1997).
88. Hamosh, A., Scott, A., Amberger, J., Bocchini, C., McKusick, V. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. **33**(Database Issue): D514 (2005).
89. Nicholas, F. Online Mendelian Inheritance in Animals (OMIA): a comparative knowledgebase of genetic disorders and other familial traits in non-laboratory animals. Nucleic Acids Res. **31**(1): 275 (2003).
90. Kahraman, A., Avramov, A., Nashev, L., Popov, D., Ternes, R., Pohlenz, H., Weiss, B. PhenomicDB: a multi-species genotype/phenotype database for comparative phenomics. Bioinformatics **21**(3): 418 (2005).
91. Rebhan, M., Chalifa-Caspi, V., Prilusky, J., Lancet, D. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. Bioinformatics **14**(8): 656 (1998).
92. Gkoutos, G., Green, E., Am Mallon, J., Davidson, D. *Building mouse phenotype ontologies*. Singapore: World Scientific, p. 178 (2004).
93. Lussier, Y., Borlawsky, T., Rappaport, D., Liu, Y., Friedman, C. PhenoGO: assigning phenotypic context to Gene Ontology annotations with natural language processing. Pac. Symp. Biocomput. **2006**: 64–75 (2006).
94. Philip, G., Bertram, W., Hans-Dieter, P., Ulf, L. Mining phenotypes for gene function prediction. BMC Bioinformatics **9**: 136.
95. Field, D., Sansone, S., Collis, A., Booth, T., Dukes, P., Gregurick, S., Kennedy, K., Kolar, P., Kolker, E., Maxon, M. ′Omics data sharing. Science **326**(5950): 234 (2009).
96. Laibe, C., Le Novère, N. MIRIAM Resources: tools to generate and resolve robust cross-references in Systems Biology. BMC Syst. Biol. **1**(1): 58 (2007).
97. Goodstadt, L., Ponting, C.P. Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. PLoS Comput. Biol. **2**(9): e133 (2006).
98. Date, S.V., Marcotte, E.M. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. Nat. Biotechnol. **21**(9): 1055–1062 (2003).
99. Hawkins, T., Kihara, D. PFP: automatic annotation of protein function by relative GO association in multiple functional contexts. ISMB, June 25–29, Detroit, Michigan. pp. 117: 1471–2105 (2005).
100. Watson, J., Sanderson, S., Ezersky, A., Savchenko, A., Edwards, A., Orengo, C., Joachimiak, A., Laskowski, R., Thornton, J. Towards fully automated structure-based function prediction in structural genomics: a case study. J. Mol. Biol. **367**(5): 1511–1522 (2007).
101. Sadowski, M., Jones, D. The sequence-structure relationship and protein function prediction. Curr. Opin. Struct. Biol. **19**: 357–362 (2009).
102. Vaidehi, N., Floriano, W., Trabanino, R., Hall, S., Freddolino, P., Choi, E., Zamanakos, G., Goddard, W. Prediction of structure and function of G protein-coupled receptors. Proc. Natl. Acad. Sci. **99**(20): 12622 (2002).
103. Hearst, M., Dumais, S., Osman, E., Platt, J., Scholkopf, B. Support vector machines. IEEE Intell. Syst. **13**(4): 18–28 (1998).
104. Jensen, F. *An introduction to Bayesian networks*. London: UCL press (1996).
105. Quinlan, J. Induction of decision trees. Mach. Learn. **1**(1): 81–106 (1986).
106. Funahashi, K. On the approximate realization of continuous mappings by neural networks. Neural Netw. **2**(3): 183–192 (1989).

107. Pavlidis, P., Weston, J., Cai, J., Grundy, W. Gene functional classification from heterogeneous data. New York, NY: ACM, pp. 249–255 (2001).
108. Lanckriet, G., De Bie, T., Cristianini, N., Jordan, M., Noble, W. A statistical framework for genomic data fusion. Bioinformatics **20**(16): 2626–2635 (2004).
109. Troyanskaya, O., Dolinski, K., Owen, A., Altman, R., Botstein, D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). Proc. Natl. Acad. Sci. **100**(14): 8348 (2003).
110. Zhang, L., Wong, S., King, O., Roth, F. Predicting co-complexed protein pairs using genomic and proteomic data integration. BMC Bioinformatics **5**(1): 38 (2004).
111. King, O., Foulger, R., Dwight, S., White, J., Roth, F. Predicting gene function from patterns of annotation. Genome Res. **13**(5): 896 (2003).
112. Jensen, L., Gupta, R., Blom, N., Devos, D., Tamames, J., Kesmir, C., Nielsen, H., Staerfeldt, H., Rapacki, K., Workman, C. Prediction of human protein function from post-translational modifications and localization features. J. Mol. Biol. **319**(5): 1257–1265 (2002).
113. Mateos, A., Dopazo, J., Jansen, R., Tu, Y., Gerstein, M., Stolovitzky, G. Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons. Genome Res. **12**(11): 1703 (2002).
114. Hibbs, M.A., Myers, C.L., Huttenhower, C., Hess, D.C., Li, K., Caudy, A.A., et al. Directing experimental biology: a case study in mitochondrial biogenesis. PLoS Comput. Bio. **5**(3): e1000322 (2009).
115. Marcotte, E., Pellegrini2 M, Thompson, M., Yeates, T., Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. Proc. Natl. Acad. Sci. USA **93**: 4787–4792 (1996).
116. Lee, I., Date, S., Adai, A., Marcotte, E. A probabilistic functional network of yeast genes. Science **306**(5701): 1555 (2004).
117. Linghu, B., Snitkin, E., Holloway, D., Gustafson, A., Xia, Y., DeLisi, C. High-precision high-coverage functional inference from integrated data sources. BMC Bioinformatics **9**(1): 119 (2008).
118. Karaoz, U., Murali, T., Letovsky, S., Zheng, Y., Ding, C., Cantor, C., Kasif, S. Whole-genome annotation by using evidence integration in functional-linkage networks. Proc. Natl. Acad. Sci. **101**(9): 2888 (2004).
119. Guan, Y., Myers, C., Lu, R., Lemischka, I., Bult, C., Troyanskaya, O. A genomewide functional network for the laboratory mouse. PLoS Comput. Biol. **4**(9) (2008).
120. Kim, W., Krumpelman, C., Marcotte, E. Inferring mouse gene functions from genomic-scale data using a combined functional network/classification strategy. Genome Biol. **9**(Suppl 1): S5 (2008).
121. Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., Morris, Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biol. **9**(Suppl 1): S4 (2008).
122. Pena-Castillo, L., Tasan, M., Myers, C.L., Lee, H., Joshi, T., Zhang, C., Guan, Y., Leone, M., Pagnani, A., Kim, W.K., et al. A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. Genome Biol. **9**(Suppl 1): S2 (2008).
123. Hartwell, L., Hopfield, J., Leibler, S., Murray, A. From molecular to modular cell biology. Nature **402**(6761): 47 (1999).
124. Schwikowski, B., Uetz, P., Fields, S. A network of protein-protein interactions in yeast. Nat. Biotechnol. **18**(12): 1257–1261 (2000).
125. Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., Takagi, T. Assessment of prediction accuracy of protein function from protein-protein interaction data. Yeast **18**(6): 523–531 (2001).
126. Chua, H., Sung, W., Wong, L. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. Bioinformatics **22**(13): 1623 (2006).
127. Vazquez, A., Flammini, A., Maritan, A., Vespignani, A. Global protein function prediction from protein-protein interaction networks. Nature Biotechnology **21**: 697–700 (2003).

128. Deng, M., Zhang, K., Mehta, S., Chen, T., Sun, F. Prediction of protein function using protein-protein interaction data. J. Comput. Biol. **10**(6): 947–960 (2003).
129. Deng, M., Chen, T., Sun, F. An integrated probabilistic model for functional prediction of proteins. J. Comput. Biol. **11**(2–3): 463–475 (2004).
130. Lanckriet, G.R., Deng, M., Cristianini, N., Jordan, M.I., Noble, W.S. Kernel-based data fusion and its application to protein function prediction in yeast. Pac. Symp. Biocomput. **2004**: 300–311 (2004).
131. Arnau, V., Mars, S., Marín, I. Iterative cluster analysis of protein interaction data. Bioinformatics **21**(3): 364 (2005).
132. Spirin, V., Mirny, L. Protein complexes and functional modules in molecular networks. Proc. Natl. Acad. Sci. **100**(21): 12123 (2003).
133. Bader, G., Hogue, C. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics **4**(1): 2 (2003).
134. Sharan, R., Ideker, T., Kelley, B., Shamir, R., Karp, R. Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. J. Comput. Biol. **12**(6): 835–846 (2005).
135. Asthana, S., King, O., Gibbons, F., Roth, F. Predicting protein complex membership using probabilistic network reliability. Genome Res. **14**(6): 1170 (2004).
136. Segal, E., Wang, H., Koller, D. Discovering molecular pathways from protein interaction and gene expression data. Bioinformatics **19**(1): 264–272 (2003).
137. Hanisch, D., Zien, A., Zimmer, R., Lengauer, T. Co-clustering of biological networks and gene expression data. Bioinformatics **18**: 145–154 (2002).
138. Tanay, A., Sharan, R., Kupiec, M., Shamir, R. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. Proc. Natl. Acad. Sci. **101**(9): 2981 (2004).
139. Tasan, M., Tian, W., Hill, D.P., Gibbons, F.D., Blake, J.A., Roth, F.P. An en masse phenotype and function prediction system for Mus musculus. Genome Biol. **9**(Suppl 1): S8 (2008).
140. Breiman, L. Random forests. Mach. Learn. **45**(1): 5–32 (2001).
141. Hanley, J.A., McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology **143**(1): 29–36 (1982).
142. Berriz, G., Roth, F. The Synergizer service for translating gene, protein and other biological identifiers. Bioinformatics **24**(19): 2272 (2008).
143. van Iersel, M., Pico, A., Kelder, T., Gao, J., Ho, I., Hanspers, K., Conklin, B., Evelo, C. The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. BMC Bioinformatics **11**(1): 5 (2010).
144. Le Novore, N., Finney, A., Hucka, M., Bhalla, U., Campagne, F., Collado-Vides, J., Crampin, E., Halstead, M., Klipp, E., Mendes, P. Minimum information requested in the annotation of biochemical models (MIRIAM). Nat. Biotechnol. **23**(12): 1509–1515 (2005).
145. Moult, J., Fidelis, K., Rost, B., Hubbard, T., Tramontano, A. Critical assessment of methods of protein structure prediction (CASP) – round 6. Proteins **61(**Suppl 7): 3–7 (2005).
146. Murali, T.M., Wu, C.J., Kasif, S. The art of gene function prediction. Nat. Biotechnol. **24**(12): 1474–1475; author reply 1475–1476 (2006).
147. Song, J., Singh, M. How and when should interactome-derived clusters be used to predict functional modules and protein function? Bioinformatics **25**(23): 3143–3150 (2009).