Computational Protein Function Prediction: Framework and Challenges

Meghana Chitale and Daisuke Kihara

Abstract Large scale genome sequencing technologies are increasing the abundance of experimental data which requires functional characterization. There is a continually widening gap between the mounting numbers of available genomes and completeness of their annotations, which makes it impractical to manually curate the genomes for function information. To handle this growing challenge we need computational techniques that can accurately predict functions for these newly sequenced genomes. In this chapter we focus on the framework required for computational function annotation and the challenges involved. Controlled vocabularies of functional terms, e.g. Gene Ontology, MIPS functional catalogues, Enzyme commission numbers, form the basis of prediction methods by capturing the available biological knowledge in the form, suitable for computational processing. We review functional vocabularies in detail along with the methods developed for quantitatively gauging the functional similarity between the vocabulary terms. We also discuss challenges in this area, first pertaining to the erroneous annotations floating in the sequence database and second regarding the limitations of the functional term vocabulary used for protein annotations. Lastly, we introduce community efforts to objectively assess the accuracy of function prediction.

Introduction

With the advances in technology, whole genome sequencing for new organisms is no longer an enormous project. Numbers of genomes are being sequenced every year adding the tremendous amount of data available for computational investigators. As shown in Fig. 1, the number of entries of genomes in KEGG database [1] have almost doubled form year 2007 (~ 600 genomes) to year 2010 (~1,200 genomes).

D. Kihara (✉)
Department of Biological Sciences; Department of Computer Science; Markey Center for Structural Biology, College of Science, Purdue University, West Lafayette, IN 47907, USA
e-mail: dkihara@purdue.edu

D. Kihara (ed.), *Protein Function Prediction for Omics Era*,
DOI 10.1007/978-94-007-0881-5_1, © Springer Science+Business Media B.V. 2011

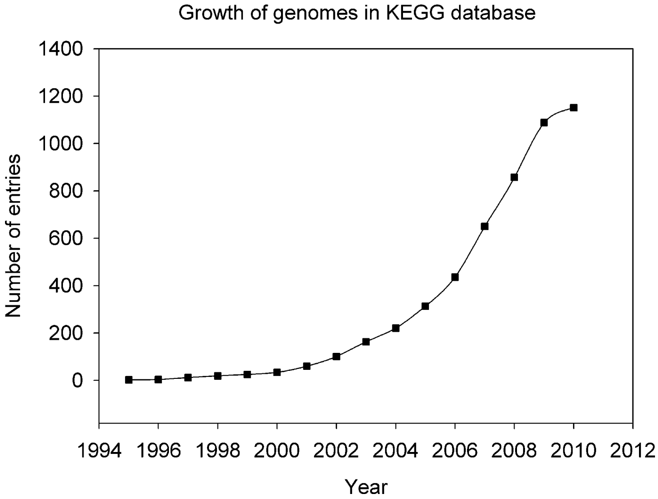


Fig. 1 Growth of genomes in KEGG database from year 1995 till 2010. Yearly release information of KEGG data was obtained from GenomeNet (http://www.genome.jp/en/db_growth.html)

The pace of accumulating sequence data will only increase, in fact, the new generation technology can sequence microbial genome within a couple of days [2, 3].

However, it is still a daunting task to correctly assign functional annotations to these newly sequenced genomes based on their sequence information. It is not feasible to conduct conventional experimental procedures on this entire stockpile of sequences for recovering the functional information, and this has triggered the need for methods that can consistently assign functions to unknown proteins [4–8]. Conventionally in this scenario researchers have focused on using homology or sequence similarity to transfer annotations to newly sequenced proteins using popular homology search algorithms such as BLAST [9] and FASTA [10, 11]. Although considering homology is a genuine way of inferring function in the light of evolution, practically, it is not always trivial to extract correct function information from a sequence database search result. Another weakness of the conventional homology searches is that a considerable portion of genes in a genome are left as unannotated. In Fig. 2, we have analyzed the number of annotated genes in the genome sequences taken from the KEGG database [1] (version March 2010). We have examined the genomes to separate the number of genes that have unknown annotations characterized by keywords mentioned in the caption for the figure. This gives us a crude idea about the percentage of unknown genes in each genome. It can be seen from Fig. 2 that for around 50% of genomes in the database we know functional characteristics of less than 60% of genes in there. Even for well studied model organisms such as *Saccharomyces cerevisiae* (82.4% annotated), *Escherichia coli K-12 MG1655* (64.9% annotated), *Arabidopsis thaliana* (66.3% annotated), a significant number of genes have no annotation. Therefore, new methods in this area are required to improve the function prediction accuracy as well as the genome annotation coverage.

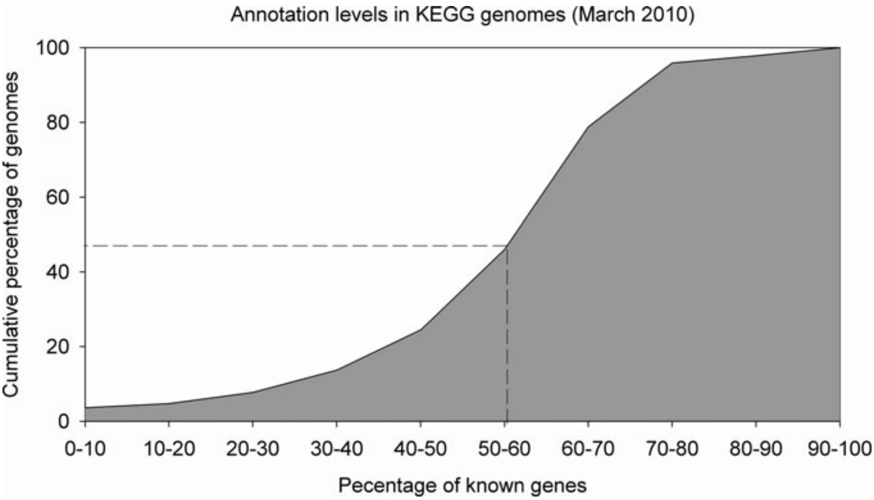


Fig. 2 Annotation levels of genomes in KEGG database. 1,172 genomes in KEGG database were analyzed to separate the number of annotated genes from unknown genes (entries in the database annotated with terms "hypothetical", "putative", "unknown", "uncharacterized", "predicted", "no hits", "codon recognized", "expressed protein", and "conserved protein"). The figure shows cumulative percentage of genomes having specified percentage of annotated genes

As the first chapter in this book, we explain the fundamental information, which lays the framework of computational protein function prediction. We first summarize controlled functional vocabularies and evaluation measures for accuracy of protein function prediction. Along with this, we would like to draw readers' attention to challenges in this area, first pertaining to the erroneous annotations floating in the sequence database and second regarding the limitations of the functional term vocabulary used for protein annotations. Lastly, we introduce community efforts to objectively assess the accuracy of function prediction.

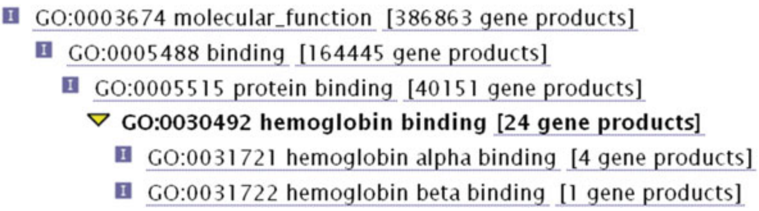
Controlled Functional Vocabularies

For managing computational function prediction we need to transform the descriptive biological knowledge into qualitative and quantitative models, which requires robust and accessible biological information system. Protein functions or annotations have long been described with vocabularies that are conventionally used within each research community or research group. Thus, there have been cases that essentially same annotations are described with different terms across different species and research communities. However, such situations hinder computational handling of functional information, including extraction of function information of genes from databases and summarizing such information to predict function. A practical solution for this is to unify the functional terms used for functional annotation of genes. In recent years controlled sets of functional vocabularies have been developed along this direction. Below we describe several ontologies, including Gene Ontology (GO) [12], Enzyme Commission (EC) number, [13], MIPS functional catalogue [14] (FunCat), Transporter Classification System [15], KEGG orthology [16], and the other efforts of constructing ontologies.

Gene Ontology

The Gene Ontology (GO) Consortium [17] of collaborating databases has developed a structured controlled vocabulary to describe gene function. GO vocabulary terms are arranged in a hierarchical fashion using a Directed Acyclic Graph (DAG) and are separated into three categories: Biological Process (BP), Molecular Function (MF), and Cellular Component (CC). One or more terms from each category can be used to describe a protein. Cellular component indicates to which anatomical part of the cell the protein belongs to, for example, *ribosome (GO:0005840)* or *nucleus (GO:0005634)*. Biological process terms indicate assemblies of molecular functions which achieve a well defined task through a series of cellular events. Examples of biological processes are *carbohydrate metabolism (GO:0003677)*, *regulation of transcription (GO:0045449)* etc. Molecular functions represent activities carried out at molecular level by proteins or complexes, for example, *catalytic activity (GO:0003824)* or *DNA binding (GO:0003677)* etc. Thus each GO term will have a category and an identifier in the format GO:xxxxxxx associated with it, along with a term definition to explain the meaning of the term. For example, term *protein binding* is referred using identifier *GO:0005515* and its definition says following *Interacting selectively and non-covalently with any protein or protein complex*. The vocabulary is arranged as a DAG where each term can have one or more parents. Figure 3 represents the tree structure obtained for the term *hemoglobin binding* showing all its parents till the root term *all*. As you go deeper in the hierarchy the terms become more specific.

All terms in GO other than the root term have either *is-a*, *is_part_of*, *positively regulates, or negatively regulates* relationship with some other more general term. For example as shown in Fig. 4 the term *glucose transport (GO:0015758)* is_a



GO:0003674 molecular_function [386863 gene products]
GO:0005488 binding [164445 gene products]
GO:0005515 protein binding [40151 gene products]
▽ **GO:0030492 hemoglobin binding [24 gene products]**
GO:0031721 hemoglobin alpha binding [4 gene products]
GO:0031722 hemoglobin beta binding [1 gene products]

Fig. 3 Structure of Gene Ontology for term *hemoglobin binding* displayed using AmiGO browser (http://amigo.geneontology.org/cgi-bin/amigo/go.cgi) for GO terms. Against each term the number of gene products that are annotated with the given term in the GO database, is displayed

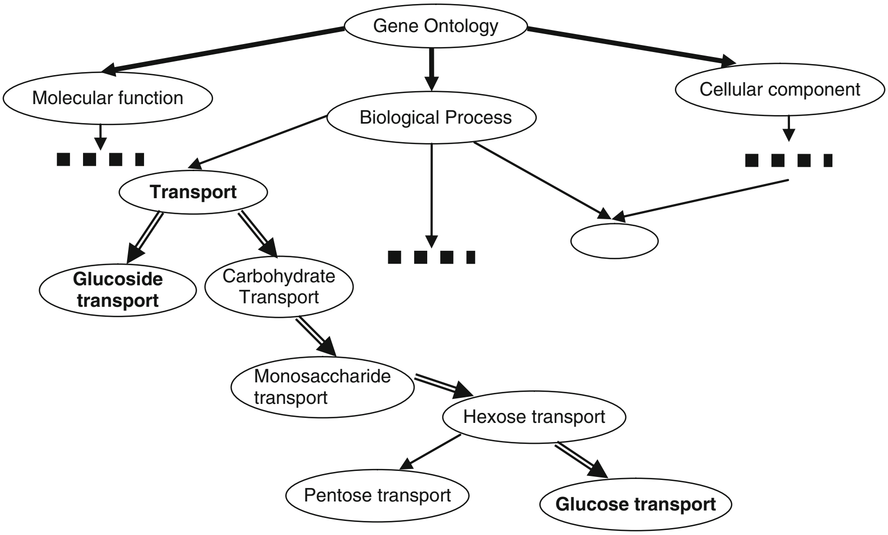
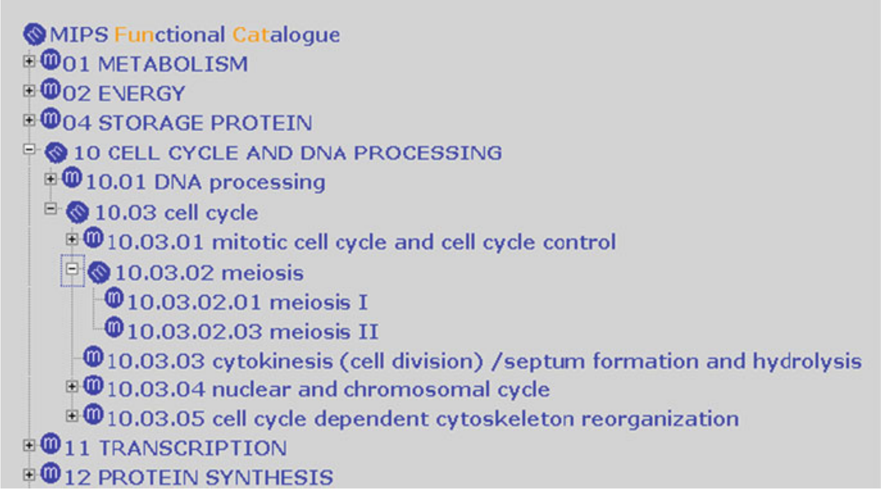


Fig. 4 Partial Gene Ontology hierarchy describing the ancestors of terms *Glucoside transport* and *Glucose transport*. Double lined arrows show the path to the Lowest Common Ancestor (LCA) of the two terms

Hexose transport (GO:0008645), which ultimately is_a *transport (GO:0006810)*. Due to this relationship when a protein is annotated by term *X* then it is automatically annotated by all ancestor terms of *X* which are basically less specific descriptions of *X*. Similarly, some more relationships have been defined in GO, e.g. *B* is part_of *A*, which implies that when *B* exists it is part of *A*. For example, *mitochondrial membrane (GO:0031966)* is part of *mitochondrial envelope (GO:0005740)*. *Regulates* relationship is used in GO to capture the fact that one process can directly affect the manifestation of another process; this relationship has two sub-relations *positively regulates* and *negatively regulates* to capture the specific forms of regulation.

Association between a gene product and its GO annotation is generally based on one or more supporting evidences. GO has defined the evidence codes that help capture information about the source from which this association is obtained (http://www.geneontology.org/GO.evidence.shtml). Inferred from Electronic Annotation (IEA) is the only evidence code that is not reviewed by a curator indicating that assignment of annotation to the gene product is automatic. All curator-assigned evidence codes fall into one of the four categories; (1) experimental (e.g. Inferred from Direct Assay (IDA), Inferred from Genetic Interaction (IGI) etc), (2) computational analysis (e.g. Inferred from Sequence or structural Similarity (ISS), Inferred from Genomic Context (IGC)), (3) author statement (Traceable Author Statement (TAS), Non-traceable Author Statement (NAS)), and (4) curatorial statement (Inferred by Curator (IC) and No biological Data available (ND)). It should be noted that evidence codes do not indicate quality of annotation but only provide information about the source of annotation.



MIPS Functional Catalogue
01 METABOLISM
02 ENERGY
04 STORAGE PROTEIN
10 CELL CYCLE AND DNA PROCESSING
10.01 DNA processing
10.03 cell cycle
10.03.01 mitotic cell cycle and cell cycle control
10.03.02 meiosis
10.03.02.01 meiosis I
10.03.02.03 meiosis II
10.03.03 cytokinesis (cell division) /septum formation and hydrolysis
10.03.04 nuclear and chromosomal cycle
10.03.05 cell cycle dependent cytoskeleton reorganization
11 TRANSCRIPTION
12 PROTEIN SYNTHESIS

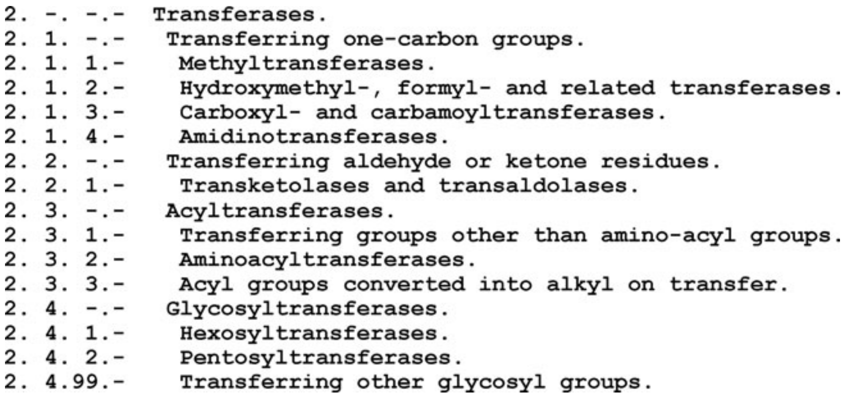
Fig. 5 Hierarchical structure of MIPS functional catalogue displayed partially using FunCat Database tool (http://mips.helmholtz-muenchen.de/proj/funcatDB/search_main_frame.html)

MIPS Functional Catalogue

Similar to Gene Ontology, MIPS Functional Catalogue (FunCat) [14] is a hierarchically organized species independent vocabulary (Fig. 5). FunCat is organized as a tree rather than a DAG. In FunCat there are 28 main catalogues, each of which is organized in a hierarchical tree structure. These main branches or catalogues cover features like *localization*, *transport*, *metabolism,* etc. FunCat currently contains 1,307 categories each of which is assigned a two digit number. FunCat identifier is represented as a series of category numbers separated by a dot based on the level in the hierarchy, for example *metabolism* is *01* and locates at first level, while *01.01.03.02.01* is *biosynthesis of glutamate* which belongs to most specific level.

Enzyme Commission Numbers

The Enzyme Commission (EC) numbers [13] are another functional classifiers that are used to classify enzymes based on reactions they catalyze. Thus as compared to the GO vocabulary, the EC numbers are reaction oriented and describe only the biochemical activity of proteins. In the enzyme nomenclature, each EC number consists of four numbers, i.e. EC x.x.x.x, each describing the enzyme at different levels of detail. There are six top levels of EC numbers from 1 to 6 which represent *oxidoreductases, transferases, hydrolases, lyases, isomerases*, and *ligases*, respectively. The next level of depth contains more details about the reaction, for example, EC number 2.1 indicates *transferase* (2 at the top level) involved in transferring



```
2. -. -.-  Transferases.
2. 1. -.-   Transferring one-carbon groups.
2. 1. 1.-    Methyltransferases.
2. 1. 2.-    Hydroxymethyl-, formyl- and related transferases.
2. 1. 3.-    Carboxyl- and carbamoyltransferases.
2. 1. 4.-    Amidinotransferases.
2. 2. -.-   Transferring aldehyde or ketone residues.
2. 2. 1.-    Transketolases and transaldolases.
2. 3. -.-   Acyltransferases.
2. 3. 1.-    Transferring groups other than amino-acyl groups.
2. 3. 2.-    Aminoacyltransferases.
2. 3. 3.-    Acyl groups converted into alkyl on transfer.
2. 4. -.-   Glycosyltransferases.
2. 4. 1.-    Hexosyltransferases.
2. 4. 2.-    Pentosyltransferases.
2. 4.99.-    Transferring other glycosyl groups.
```

Fig. 6 EC number hierarchy displayed partially as shown by ExPASy Proteomics Server (http://ca.expasy.org/enzyme/enzyme-byclass.html)

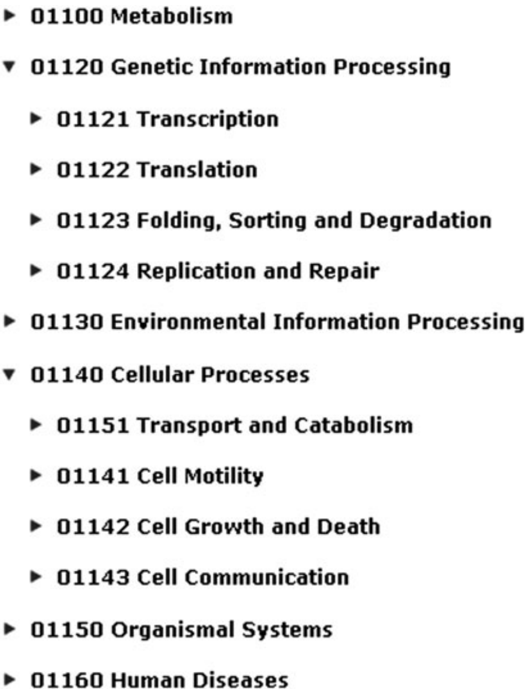
one carbon groups (1 at the second level) as shown in Fig. 6. The KEGG pathway database [1, 18] uses the EC numbers to indicate enzymes involved in metabolic pathways.

Transport Classification (TC) System

Almost all transmembrane transport processes are mediated by integral membrane proteins which are classified using Transporter Classification System [15] (http://tcdb.ucsd.edu/tcdb/). As compared to EC numbers which are focused only on function, TC classification is based on both function and phylogeny. According to this system, the transporters are classified based on five criteria and each of these provides one component of TC number for a protein. A TC number has usually five components, A, B, C, D, and E, where A corresponds to the transporter class, B corresponds to the transporter subclass, C corresponds to the family (or superfamily), D corresponds to subfamily, and E specifies the substrate transported as well as polarity of transport (in or out).

KEGG Orthology (KO)

The KEGG database includes the KEGG Orthology (KO) [16] database as one of its components [1, 18]. The primary purpose of KO is to provide the list of orthologous genes in genomes. KO is structured as a DAG hierarchy that can be effectively used for the definition of the function of ortholog groups. It has four levels with the first one consisting of five classes; *metabolism, genetic information processing, environmental information processing, cellular processes, and human diseases*, as shown in Fig. 7. The second level consists of finer functional sub-categories, third



- ► 01100 Metabolism
- ▼ 01120 Genetic Information Processing
  - ► 01121 Transcription
  - ► 01122 Translation
  - ► 01123 Folding, Sorting and Degradation
  - ► 01124 Replication and Repair
- ► 01130 Environmental Information Processing
- ▼ 01140 Cellular Processes
  - ► 01151 Transport and Catabolism
  - ► 01141 Cell Motility
  - ► 01142 Cell Growth and Death
  - ► 01143 Cell Communication
- ► 01150 Organismal Systems
- ► 01160 Human Diseases

Fig. 7 KEGG orthology displayed partially (http://www.genome.jp/kegg/ko.html)

level consists of KEGG pathways and fourth one corresponds to functional terms. The unique feature of the KO is that each entry has links to pathways and reactions as well as orthologous genes and hence it is convenient to annotate a set of genes with KO function terms and identify pathways where the genes belong to [16].

Other Biological Ontologies

Along with the aforementioned vocabularies for protein function, there are some other interesting ontologies that provide annotations to proteins in different domains specifically for particular species or research communities. Smith et al. [19] have developed Open Biological and Biomedical Ontologies (OBO) Foundry which consists of a collaborative effort to merge ontologies, where we can find a wide variety of open biological ontologies listed on their project website (http://www.obofoundry.org/). The ontologies include Protein Ontology developed by Protein Information Resources (PIR, http://pir.georgetown.edu/pro/), which encompasses evolution and multiple protein forms of a gene, Chemical Entities of Biological Interest (CHEBI) developed by the European Bioinformatics Institute, which classifies structures of biologically relevant chemical compounds, and ontologies for phenotype and anatomy of individual organisms. Such efforts are helping standardize the representation of domain knowledge across research communities and increase its application. By combining different ontologies, function prediction methods which output GO terms could be expanded to predict other types of ontology terms, such as phenotype.

Definition of Functional Similarity

Definition of functional similarity for protein pairs is important when comparing predictions with actual annotations of proteins to compute the prediction accuracy. A quantitative functional similarity score is also used as the target function to be optimized in the course of developing a function prediction method. In this section we overview several metrics proposed for quantifying functional similarity using the function ontology. We use the GO here since the proposed metrics are developed for the GO. However, application of the metrics to the other ontologies should be straightforward. For a review on this topic, refer to Sheehan et al. [20].

The simplest technique that can be used to compare annotations is head to head comparisons [21, 22] where we check for exact matches. Its key disadvantage is that the information embedded in the vocabulary structure is not used. Vocabulary structure relates terms to each other and with head to head comparisons we will be penalizing inexact predictions that are close to the actual ones on the GO DAG. Set based similarity measures have been developed based on head to head comparisons to match the two objects described using a set of features. Tversky et al. [23] use Eq. (1) to describe similarity between two objects *a* and *b* which have feature sets *A* and *B* respectively, as some function *F* of features that are common, that only belong to *A* and that only belong to *B*.

$$sim(a,b) = F(A \cap B, A - B, B - A) \tag{1}$$

Another technique [21, 24, 25] that is commonly used for GO annotations is to base the similarity on the minimum path length between a pair of terms on DAG or on the fact that ancestors are less specific representation of the same term in DAG hierarchy. This technique can suffer from drawback that not all parts of GO are developed equally and not all terms at the same depth in the structure represent same biological details.

Some techniques describe a protein as a binary vector with 1's and 0's specifying presence and absence of terms in the annotation set of a protein. The similarity between two such vectors can be defined as a cosine distance (Eq. (2)), where p_i and p_j are vectors describing annotations of two proteins. Instead of binary values, the terms can also be represented as weights based on their frequency of occurrence in the database reflecting how specific they are [26, 27].

$$sim(p_i,p_j) = \frac{p_i \cdot p_j}{|p_i|\,|p_j|} = \frac{p_i \cdot p_j}{\sqrt{p_i \cdot p_i} \cdot \sqrt{p_j \cdot p_j}} \tag{2}$$

In the function prediction category in CASP7 [21], the assessors designed a score based on the depth of common ancestor between predicted and actual GO terms as shown in Eq. (3). Each annotation is compared to its closest target prediction which forms a "computable pair", and the total score is given by the sum of depths of common ancestor of all computable pairs normalized by the maximum possible value of score. Along with this they have also used the head to head comparison of GO term predictions for comparing different methods.

$$\text{GOscore} = \frac{\text{sum of common ancestor depths of computable pairs}}{\text{sum of annotated terms depth}} \tag{3}$$

Resnik [28] has defined the Information Content (IC) of a term c based on the frequency of the occurrence of that term in the database as explained in the Eqs. (4), (5), and (6), where each term's frequency depends on its children node in the vocabulary structure because of the is_a relationships in the GO.

$$\text{freq}(c) = \text{annot}(c) + \sum_{h \in \text{children}(c)} \text{freq}(h) \tag{4}$$

$$p(c) = \text{freq}(c)/\text{freq}(\text{root}) \tag{5}$$

$$IC(c) = –\log(p(c)) \tag{6}$$

He has developed a graphical method to compute similarity between two terms (say *c1* and *c2*) in the taxonomy, by using the IC of their Lowest Common Ancestor (LCA) term (Eq. (7)). Figure 4 illustrates the concept of LCA by showing that the LCA of terms *Glucoside transport* and *Glucose transport* in the GO hierarchy is the term *transport* which is common ancestor for both terms and is located at the maximum depth in the DAG. Lin [29] further extended this semantic similarity measure to include information content of both terms being compared along with the information content of the ancestor term (Eq. (8)). Lord et al. [30] have first applied this IC based semantic similarity technique from Eq. (7) to Gene Ontology vocabulary to compute functional similarity based on protein annotations.

$$Sim_{Lin}(c1, c2) = \max_{c \in \{\text{common ancestors of } c1 \text{ and } c2\}} (-\log(p(c))) \tag{7}$$

$$Sim_{Lin}(c1, c2) = \max_{c \in \{\text{common ancestors of } c1 \text{ and } c2\}} \left(\frac{2 \log p(c)}{\log p(c1) + \log p(c2)} \right) \tag{8}$$

These term based similarity scores were extended to develop a pair-wise protein similarity score by Schlicker et al. [31]. They combined Resnik's and Lin's scores to compute a semantic similarity score for a pair of GO terms as shown in Eq. (9). To compute the semantic similarity between pair of proteins A and B they used the pair wise similarity values between the GO annotations GO_A and GO_B of both proteins respectively. Then scores from two different GO categories were combined to finally compute the overall similarity between the given two proteins (Eq. (12)). As shown in Eq. (10), semantic similarity matrix S_{ij} holds the pair wise similarity scores for all pairs of annotations from GO_A and GO_B where set GO_A has N annotations and GO_B has M annotations. For these two sets the overall similarity score referred as *GOscore* is computed by finding best matched hits for annotations in one of the directions using either row wise or column wise average of maximums (Eq. (11)). Further as shown in Eq. (12) *BPscore* and *MFscore* values computed using annotation sets from each of these categories are combined to yield the final *funsim* score that represents semantic similarity between pair of proteins under consideration.

$$Sim_{\mathrm{R}el}(c1,c2) = \max_{c\,\in\,\{\text{common ancestors of } c1 \text{ and } c2\}} \left(\frac{2\log p(c)\cdot(1-p(c))}{\log p(c1)+\log p(c2)}\right) \tag{9}$$

$$Sij = sim(GO_A^i, GO_B^j), \forall i \in \{1...N\} \; and \; \forall j \in \{1...M\} \tag{10}$$

$$GOscore = \max\left\{\left(\frac{1}{N}\sum_{i=1}^{N}\max_{1\le j\le M} Sij\right), \left(\frac{1}{M}\sum_{j=1}^{M}\max_{1\le i\le N} Sij\right)\right\} \tag{11}$$

$$funsim = \frac{1}{2}\cdot\left[\left(\frac{BPscore}{\max(BPscore)}\right)^2 + \left(\frac{MFscore}{\max(MFscore)}\right)^2\right] \tag{12}$$

Methods developed in the last few years have mainly focused on pair-wise protein similarity, but with the development of high throughput techniques we are frequently required to functionally interpret a computationally or experimentally determined set of proteins and check if they are functionally homogeneous [27, 32–37]. Earlier coherence of set of proteins was based mostly on the enrichment of annotations in the set [38, 39], but it has been shown that average number of enriched GO annotations in random groups is more than the number in coherent groups of proteins [37]. This has put forth the need to further develop better protein group coherence detection methods that can segregate groups of biologically relevant proteins from random ones.

Chagoyen et al. [27] use Eq. (2) for computing pair wise similarity between proteins in the set under consideration. Later they aggregate the scores across all pairs of proteins in the set S to obtain coherence score for the set as shown in Eq. (13). Statistical significance of this coherence score is computed in the context of reference set using hypergeometric distribution.

$$\mathrm{score}(S) = \frac{\sum_{i=1}^{|S|}\sum_{j=i+1}^{|S|} sim(p_i,p_j)}{|S|(|S|-1)/2} \tag{13}$$

Pandey et al. [36] performed similar aggregation basing their pair wise protein similarity score on the information content of minimum common ancestor set to the sets of terms annotating two proteins. For annotations of proteins p_i and p_j they compute minimum common ancestor term set and find the number of proteins annotated by all of those terms, which is given by $\mid G_{\Lambda}(p_i, p_j) \mid$. Further the pair wise protein functional similarity score is given by Eq. (14) where G_r is set of all proteins. The pair wise scores for all pairs of proteins in a set S are averaged in Eq. (15) to obtain the coherence score for S.

$$\rho_I(\mathrm{p_i}, \mathrm{p_j}) = -\log_2\left(\frac{|G_{\Lambda(pi,pj)|}}{|G_r|}\right) \tag{14}$$

$$\sigma_A(S) = \frac{\sum_{i=1}^{|S|}\sum_{j=i+1}^{|S|}\rho_I(p_i,\, p_j)}{|S|(|S|-1)/2} \tag{15}$$

Zheng et al. [37] use probabilistic model to extract biologically relevant topics from GO annotation corpus and classify each word from MEDLINE document abstracts into these topics. A document is semantically represented as count of the number of words belonging to each of the topics. A bipartite graph called ProtSemNet is constructed by joining topics obtained from each document with the proteins associated with that document, where edge weights in the graph are based on the count of words for the topic. For evaluating functional coherence of group of proteins, they construct Steiner tree from ProtSemNet for the given group of proteins where the number of edges and total distance of the tree are used as two metrics for computing protein group coherence.

Aforementioned techniques offer an interesting new avenue in the domain of functional similarity by complimenting high throughput techniques which require formal analysis of groups of proteins.

Limitations of Homology Based Function Transfer and Erroneous Database Annotations

As an increasing number of genomes are being sequenced, more and more genes are annotated computationally mainly by using homology search tools, i.e. BLAST [9] or PSI-BLAST [40], and assigned annotations will be eventually stored in the public sequence databases [41, 42]. Once these annotations are included in the databases, they will be used as a source of function information in the annotation of new genomes. Computational annotations based on homology, however, are not always trivial [43–45]. There are numerous cases where proteins with high sequence identity have different functions [46]. Galperin and Koonin discussed major causes of questionable function assignments. These include taking into account only the annotation of the best scoring database hit, insufficient masking of low complexity regions, ignoring multi-domain organization of the query proteins or the database hits, and non-orthologous gene displacement [47]. It should be also reminded that proteins which have multiple seemingly unrelated functions in a single region (moonlighting proteins) further add complications to description of protein function [48].

Indeed several studies report potential wrong annotations to genes in genomes. Brenner compared annotations by three groups to the *Mycoplasma genitalium* genome and found that 8% of the genes have serious disagreement [49]. Devos and Valencia analyzed the different functional descriptions in genes of *M. genitalium*, *Haemophilus influenzae*, and *Methanococcus jannashii* relative to the sequence identity and estimated the error rate of annotations [50]. A recent study by Schnoes et al. [51] analyzed public databases for misannotations. Their results indicate that there are significantly less potential misannotations in Swiss-Prot [41], which is manually curated, as compared with GeneBank [42], TrEMBL [41], and KEGG [1] for the six superfamilies they studied.

The main problem of erroneous annotations is that they will be reused in annotating newer genes and thus will be propagated in the databases [8]. A model of error propagation throughout the database shows that it can significantly degrade overall quality of annotations [52]. Then, how can we avoid the catastrophic deterioration of annotation of databases? First, it is important to examine the validity of annotations by experts of each protein and organism. Researchers of *E. coli* K-12 have held a meeting to examine annotations of this important model organism [53]. A recent attempts to use wiki [54] as a tool for community annotation are along the same direction [55, 56]. Another important direction is to make information and procedure transparent, which are used to make individual annotation. The aforementioned evidence codes available in the GO database provide such important information. Also, as a future direction, the architecture of biological database may need to be improved so that the lineage of annotation, i.e. the software or evidences used to make a particular annotation, homologous sequences from which the annotation are transferred, etc. can be dynamically tracked [57].

Critical Assessment of Function Prediction Methods

For the last section of this chapter, we would like to introduce community efforts for objective assessment of protein function prediction methods. As observed in the structural bioinformatics field, namely, the protein structure prediction and the protein docking prediction, evaluating methods by a quantitative score using blind prediction targets can help assessing the status of the field and also stimulates researchers' motivation for method development. In the protein structure prediction field, the Critical Assessment of Techniques for Protein Structure Prediction (CASP, http://predictioncenter.org/) while the Critical Assessment of Predictions of Interactions (CAPRI, http://www.ebi.ac.uk/msd-srv/capri/capri.html) for the protein docking prediction have served well for these purposes.

For the protein function prediction, there are two such critical assessments. The first one is as a Special Interest Group (SIG) held alongside the Intelligent Systems in Molecular Biology (ISMB) meetings. In 2005, the first meeting for the Automatic Function Prediction Special Interest Group (AFP-SIG) (http://biofunctionprediction.org/) was held at the ISMB conference at Detroit, Michigan; later meetings were followed in 2006, 2007 and 2008. The meetings are focused on exchanging ideas for automatic function predictions, which use protein sequence similarity, motifs, structures, protein-protein interactions, phylogeny, and combined data sources [58]. In 2005, they had set up a blind prediction contest where each participating research group had to provide a web interface where query sequences can be submitted and prediction results were evaluated by the organizers (thus fully automatic function prediction). The predictions were made in terms of GO terms, which were evaluated by using Eq. (7). The subsequent past AFP-SIG meetings consisted of only presentations but it was recently announced that the critical assessment of the methods will be held in the meeting of 2011.

The CASP has also started the function prediction category from CASP6 in 2004 [59]. In CASP6, predictors were allowed to provide GO terms from all three categories, binding site, binding, residue role and posttranslational modifications for each of the targets. As an exploratory category, the prediction groups were not scored and ranked at that time. In the subsequent CASP7 (2006), predictions were accepted for GO molecular function terms, EC numbers, and binding sites [21]. The aforementioned Eq. (3) in the previous section was used to assess the GO term predictions. In the CASP8 (2008) and CASP9 (2010), the function prediction is only restricted to ligand binding residue prediction, mainly because binding residues can be obtained from protein structures solved by experiments and thus can be easily assessed. In future there are many challenges in front of such blind prediction competitions: First of all, there should be availability of new functional knowledge from experimental data to evaluate the results. Also better automatic evaluation techniques may need to be developed to compare predictions with actual annotations. Finally, there should be good consensus on what types of functions will be predicted.

Summary

This chapter started with stating the motivation for development function prediction methods. Then, we overviewed fundamental technical issues for function prediction methods, including the functional ontologies and metrics for assessing accuracy for function prediction. Although steady continuous works are needed, these frameworks, especially functional ontologies, have made it possible to handle protein function computationally and also have opened up ways to for bioinformatics researchers to enter this field.

Acknowledgements MC is supported by grants from Purdue Research Foundation and Showalter Trust. DK also acknowledges a grant from National Institutes of Health (GM075004) and National Science Foundation (DMS800568, EF0850009, IIS0915801).

References

1. Kanehisa, M., Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. **28**(1): 27–30 (2000).
2. Flicek, P., Birney, E. Sense from sequence reads: methods for alignment and assembly. Nat. Methods **6**(11 Suppl): S6–S12 (2009).
3. Reeves, G.A., Talavera, D., Thornton, J.M. Genome and proteome annotation: organization, interpretation and integration. J. R. Soc. Interface **6**(31): 129–147 (2009).
4. Bujnicki, J.M. *Prediction of protein structures, functions, and interactions.* Chichester, West Sussex: Wiley. xiv, 287p., [2] p. of plates (2009).
5. Eisenberg, D., et al. Protein function in the post-genomic era. Nature **405**(6788): 823–826 (2000).
6. Friedberg, I. Automated protein function prediction – the genomic challenge. Brief Bioinform. **7**(3): 225–242 (2006).
7. Hawkins, T., Chitale, M., Kihara, D. New paradigm in protein function prediction for large scale omics analysis. Mol. Biosyst. **4**(3): 223–231 (2008).
8. Karp, P.D. What we do not know about sequence analysis and sequence databases. Bioinformatics **14**(9): 753–754 (1998).
9. Altschul, S.F., et al. Basic local alignment search tool. J. Mol. Biol. **215**(3): 403–410 (1990).
10. Pearson, W.R. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol. **183**: 63–98 (1990).
11. Pearson, W.R., Lipman, D.J. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA **85**(8): 2444–2448 (1988).
12. Harris, M.A., et al. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. **32**(Database issue): D258–261 (2004).
13. Nomenclature committee of the international union of biochemistry and molecular biology (NC-IUBMB), Enzyme Supplement 5 (1999). Eur. J. Biochem. **264**(2): 610–650 (1999). http://www.ncbi.nlm.nih.gov/pubmed/10491110
14. Ruepp, A., et al. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res. **32**(18): 5539–5545 (2004).
15. Saier, M.H., Jr. A functional-phylogenetic classification system for transmembrane solute transporters. Microbiol. Mol. Biol. Rev. **64**(2): 354–411 (2000).
16. Mao, X., et al. Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. Bioinformatics **21**(19): 3787–3793 (2005).
17. Ashburner, M., et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. **25**(1): 25–29 (2000).
18. Kanehisa, M., et al. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. **38**(Database issue): D355–360 (2010).
19. Smith, B., et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol. **25**(11): 1251–1255 (2007).
20. Sheehan, B., et al. A relation based measure of semantic similarity for Gene Ontology annotations. BMC Bioinformatics **9**: 468 (2008).
21. Lopez, G., et al. Assessment of predictions submitted for the CASP7 function prediction category. Proteins **69**(Suppl 8): 165–174 (2007).
22. Vinayagam, A., et al. GOPET: a tool for automated predictions of Gene Ontology terms. BMC Bioinformatics **7**: 161 (2006).
23. Tversky, A. Features of similarity. Psychol. Rev. **84**(4): 327–352 (1977).
24. Hawkins, T., Luban, S., Kihara, D. Enhanced automated function prediction using distantly related sequences and contextual association by PFP. Protein Sci. **15**(6): 1550–1556 (2006).
25. Wass, M.N., Sternberg, M.J. ConFunc – functional annotation in the twilight zone. Bioinformatics **24**(6): 798–806 (2008).

26. Chabalier, J., Mosser, J., Burgun, A. A transversal approach to predict gene product networks from ontology-based similarity. BMC Bioinformatics **8**: 235 (2007).
27. Chagoyen, M., Carazo, J.M., Pascual-Montano, A. Assessment of protein set coherence using functional annotations. BMC Bioinformatics **9**: 444 (2008).
28. Resnik, P. Using information content to evaluate semantic similarity in a taxonomy. Proceedings of International Joint Conference on Artificial Intelligence **1**: 448–453 (1995).
29. Lin, D. An information-theoretic definition of similarity. Proceedings of the 15th International Conference on Machine Learning **1**: 296–304 (1998).
30. Lord, P.W., et al. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics **19**(10): 1275–1283 (2003).
31. Schlicker, A., et al. A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinformatics **7**: 302 (2006).
32. Martin, D., et al. GOToolBox: functional analysis of gene datasets based on Gene Ontology. Genome Biol. **5**(12): R101 (2004).
33. Pehkonen, P., Wong, G., Toronen, P. Theme discovery from gene lists for identification and viewing of multiple functional groups. BMC Bioinformatics **6**: 162 (2005).
34. Huang da, W., et al. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol. **8**(9): R183 (2007).
35. Carmona-Saez, P., et al. GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. Genome Biol. **8**(1): R3 (2007).
36. Pandey, J., Koyuturk, M., Grama, A. Functional characterization and topological modularity of molecular interaction networks. BMC Bioinformatics **11**(Suppl 1): S35 (2010).
37. Zheng, B., Lu, X. Novel metrics for evaluating the functional coherence of protein groups via protein semantic network. Genome Biol. **8**(7): R153 (2007).
38. Curtis, R.K., Oresic, M., Vidal Puig A. Pathways to the analysis of microarray data. Trends Biotechnol. **23**(8): 429–435 (2005).
39. Draghici, S., et al. Global functional profiling of gene expression. Genomics **81**(2): 98–104 (2003).
40. Altschul, S.F., et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**(17): 3389–3402 (1997).
41. Boeckmann, B., et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. **31**(1): 365–370 (2003).
42. Benson, D.A., et al. GenBank. Nucleic Acids Res. **37**(Database issue): D26–31 (2009).
43. Devos, D., Valencia, A. Practical limits of function prediction. Proteins **41**(1): 98–107 (2000).
44. Valencia, A. Automatic annotation of protein function. Curr. Opin. Struct. Biol. **15**(3): 267–274 (2005).
45. Bork, P., Koonin, E.V. Predicting functions from protein sequences – where are the bottlenecks? Nat. Genet. **18**(4): 313–318 (1998).
46. 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).
47. Galperin, M.Y., Koonin, E.V. Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. In Silico Biol. **1**(1): 55–67 (1998).
48. Jeffery, C.J. Moonlighting proteins – an update. Mol. Biosyst. **5**(4): 345–350 (2009).
49. Brenner, S.E. Errors in genome annotation. Trends Genet. **15**(4): 132–133 (1999).
50. Devos, D., Valencia, A. Intrinsic errors in genome annotation. Trends Genet. **17**(8): 429–431 (2001).
51. Schnoes, A.M., et al. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput. Biol. **5**(12): e1000605 (2009).
52. Gilks, W.R., et al. Modeling the percolation of annotation errors in a database of protein sequences. Bioinformatics **18**(12): 1641–1649 (2002).

53. Riley, M., et al. Escherichia coli K-12: a cooperatively developed annotation snapshot – 2005. Nucleic Acids Res. **34**(1): 1–9 (2006).
54. Hu, J.C., et al. The emerging world of wikis. Science **320**(5881): 1289–1290 (2008).
55. Florez, L.A., et al. A community-curated consensual annotation that is continuously updated: the Bacillus subtilis centred wiki SubtiWiki. Database (Oxford) **2009**: bap012 (2009).
56. Huss, J.W., 3rd, et al. The Gene Wiki: community intelligence applied to human gene annotation. Nucleic Acids Res. **38**(Database issue): D633–639 (2009).
57. Zhang, M., Kihara, D., Prabhakar, S. Tracing lineage in multi-version scientific databases. Proceedings of IEEE 7th International Symposium on Bioinformatics & Bioengineering (BIBE) **1**: 440–447 (2007).
58. Friedberg, I., Jambon, M., Godzik, A. New avenues in protein function prediction. Protein Sci. **15**(6): 1527–1529 (2006).
59. Soro, S., Tramontano, A. The prediction of protein function at CASP6. Proteins **61**(Suppl 7): 201–213 (2005).