3

Automated Prediction of Protein Function from Sequence

Meghana Chitale, Troy Hawkins and Daisuke Kihara

3.1 Introduction

Investigation of protein gene function is a central question in molecular biology, biochemistry, and genetics. Because genes evolved from the same ancestral gene retain similarity in their function in most cases, finding known genes which have sufficient sequence similarity is a powerful way for predicting function. In this chapter we review computational techniques and resources for gene function prediction from sequence. We start with an overview of widely used homology search tools, such as BLAST, and extend discussion to more recently developed methods.

3.2 Principle of Inferring Function from Sequence Similarity

The driving forces of the evolution of life include complete or partial genome duplication and rearrangement,[1] and also duplications which occur on a gene basis,[2,3] that lead speciation of organisms. While active exchange of a portion of genomes between organisms such as lateral gene transfer makes ancestral relationship of organisms far more complicated that previously thought,[4,5] on the individual gene level it is generally true that duplicated or transferred genes within or between organisms retain significant sequence similarity. Genes that have evolved from a single ancestral gene are referred as *homologous* with each other.[6] Two types of homology are distinguished. *Orthologous* genes are those are diverged from speciation events of a common gene of an ancestral organism and thus reside

Prediction of Protein Structures, Functions, and Interactions Edited by Janusz M. Bujnicki
© 2009 John Wiley & Sons, Ltd. ISBN: 978-0-470-51767-3

in different organisms. In contrast, *paralogous* genes refer to those which are duplicated in a same organism thus locate at different positions in a same genome. Thus sequence similarity is an effective way to detect homology between genes (reviewed in detail in Chapter 1 by Kaminska *et al*. in this volume).

A pair of genes which share significant sequence similarity may have diverged quite recently in the history of the evolution, or there may have been an evolutionary pressure which kept the sequence unchanged over the course of a long evolution time. Another possibility is that the two sequences converged to be similar because of structural or functional constraints. In either case, functions of such two genes usually share significant similarity considering the evolutionary scenario behind it. Thus sequence similarity between two genes strongly indicates homology, which implies functional similarity in most of the cases. However, caution is needed because there are exceptions that homologous proteins have very different functions. Recent works discuss such interesting examples.[7,8]

The relationship between the sequence similarity and function similarity is also well understood in the light of the tertiary structure of proteins (reviewed in detail in chapters by Majorek *et al*. (Chapter 2) and Kosinski *et al*. (Chapter 4) in this volume). The widely accepted Anfinsen's dogma claims that the protein sequence determines the tertiary structure of the protein.[9] Moreover, from the observation of a growing number of solved protein structures, it is well established that proteins with a similar sequence generally have a similar overall fold.[10,11] Considering that the structure of a protein has crucial roles in realizing function, e.g. to catalyze chemical reaction at an active site binding a substrate or to interact with other proteins, having the same fold can be strong evidence that the proteins share functional similarity. (But there are notable counter examples, *e.g*. *superfolds,* which are protein folds adopted by different protein families.[12])

3.3 Homology Search Methods

The strategy of a sequence-based protein function prediction for a target protein is to find known protein genes which share a significant sequence similarity from a database (reviewed in detail in Chapter 1 by Kaminska *et al*. in this volume) and make prediction with function terms associated with the protein genes found. The sequence similarity of two proteins is effectively and rigorously computed by using a dynamic programming algorithm.[13,14] The SSEARCH program[15] performs rigorous local sequence alignment by the Smith-Waterman algorithm[14] between a target sequence and each sequence in a database and lists retrieved sequences sorted by their statistical significance score, E-value. As computing rigorous local sequence alignments against a current large database by SSEARCH take a considerable amount of time on a regular desktop computer, FASTA[15] and BLAST,[16] both of which employ faster algorithms than dynamic programming algorithm for computing alignments, are more widely used. FASTA reduces computational time by restricting computation of a pairwise alignment only within highly similar regions using a lookup table, while BLAST starts with finding precomputed similar 'words' of a fixed length taken from a target sequence in the framework of the finite automaton. Benchmark studies show FASTA and BLAST deteriorate the sensitivity of database search in the tradeoff for the speed compared to SSEARCH,[11,17] but all three methods will not miss obvious homologous sequences with significant sequence similarity. A search result

will also depends on parameters used, such as the amino acid similarity matrix and gap penalties.[18]

The conventional way of using these homology search tools is to extract function annotation from top hit sequences which have a significant score either in terms of the E-value or the Smith-Waterman (SW) alignment score. The commonly used threshold for the E-value is 0.01 (or 0.001), and 200 for the SW score, which were originally established on benchmark datasets of a limited size.[19,20] This strategy is commonly used in gene function annotation in genome sequencing projects.[21,22] The advantage of using a unique threshold value is that it is easy to process automatically for a large number of genes. On the other hand, problems of this strategy include that it does not take into account that each protein family has a different degree of sequence conservation[7] and also a large portion of genes in a genome are usually left as unknown because of the rather conservative function assignment.[23]

Several interesting ideas have been proposed to identify further distantly related homologs using the homology search tools. For example, an intermediate sequence found in an initial search is used to reach further distant homologs in the second run of the search[24,25] and consensus of different methods is shown to improve search performance.[26,27]

The three homology search methods introduced above perform sequence-to-sequence comparisons. In contrast, PSI-BLAST performs profile-to-sequence comparisons, making a very sensitive database search possible.[28] PSI-BLAST iterates searches, at each time constructing a profile (multiple sequence alignment, MSA) with a target and retrieved sequences, which is used for a search in the next iteration. The iteration is halted to make the final function prediction when retrieved sequences are saturated or the predefined maximum time of iterations is reached. A profile can enhance family specific conserved sequence information in a query sequence. The flip-side of PSI-BLAST's extreme sensitivity is that it occasionally produces false positives.[29] Thus, PSI-BLAST is often used with a conservative (strict) parameter setting.[30]

Profiles can also be precomputed for sequences in a database, and a target sequence is matched against them (sequence-to-profile comparison).[31] BLOCKS[32] and ProDom[33] are databases of profiles of protein domains, where a user can search known functional domains in a sequence. A protein fingerprint is a group of conserved regions used to characterize a protein family. PRINTS[34] is a collection of such protein fingerprints. Pfam[35] and SUPERFAMILY[36] are databases which store profiles of protein domains in the form of hidden Markov models (HMMs), which are statistical representations of sequence profiles.[37] Finally, both a target sequence and database sequences are precomputed into profiles and the target profile is aligned with profiles in the database. Profile-to-profile comparison methods have been shown to be very sensitive and used not only for protein function prediction[38] but also for protein structure prediction (i.e. predicting protein fold).[39,40] Numerous methods for constructing and comparing profiles have been proposed, including ways to select sequences to be included in a profile, ways to score an alignment of two profiles, and how to handle gaps.[39–41]

3.4 Predicting Function from the Other Types of Information

Besides using sequence, various other features of genes can be used for function prediction. The global tertiary structure of proteins can indicate very distant evolutionary relationships between proteins,[42] and detecting local structure similarity is aimed to predict function

by identifying functionally importance sites, such as active sites of enzymes.[43,44] Known pathway information is used as a template for finding missing genes which fit to holes in known pathways.[45] Use of Microarray gene expression data[46] and protein–protein interaction data[47] is actively investigated in function prediction. Now that many different types of databases are established and more new experimental data are made available, combination of heterogeneous data has become an interesting and promising direction for function prediction. However, as the focus of this chapter is sequence-based approaches, refer to recent review articles[23,48 49] and also the other chapters of this book for more information.

3.5 Limitations and Problems of Function Prediction from Sequence

A practical convenience of predicting function from sequence is that most of the function information of genes resides in sequence databases, such as UniProt,[50] Pfam,[35] and also protein domain and motif databases (reviewed in Chapter 1 by Kaminska *et al*. in this volume), e.g. PROSITE,[51] BLOCKS,[32] and PRINTS.[34] A consequent intrinsic limitation is that any method can essentially only extract function information which exists in a database and it is very difficult to make a prediction which goes beyond available function description of retrieved sequences. By the same reason, if function information of a gene in a database is wrong, that wrong information will be transferred to a target gene. Thus, erroneous annotation may be propagated by being reused in subsequent function assignments.[52,53] Incorrect function prediction can happen even with having genes with correct function description because of various reasons, such as ignoring multi-domain organization of genes and non-orthologous gene displacement.[54] Indeed erroneous function annotations are frequently reported.[55] To amend wrong annotations, the research community of *Escherichia coli* has held a meeting to manually curate gene annotations.[56] A recent interesting approach is a community based annotation using wiki, allowing any researcher to participate in annotating genes.[57]

3.6 Controlled Vocabularies for Gene Function Annotation

Automation of protein function prediction requires a well-established controlled vocabulary describing the annotations, which is unified across different species and research communities. If arbitrary terms are used for describing a biological function, for example, if a gene involved in ‘bacterial protein synthesis’ is described as involved in ‘translation’ in one database and as ‘protein synthesis’ in another, an automatic procedure would easily miss the similarity between the two annotations. Even for manual annotation, non-critical use of annotations from existing database entries is a major cause of erroneous function assignment.[54] Thus we need a universal way to describe gene function in structured manner which avoids ambiguity. To allow uniform referencing for functional annotations across databases several ontologies (vocabularies) have been developed. Those ontologies include Gene Ontology (GO),[58] Enzyme Commission (EC) number[59] and MIPS functional catalogue (FunCat).[60] These ontologies provide the basis for computational prediction of protein functions as they constitute the exhaustive organized space that will be searched in order to assign the most probably function to an un-annotated protein.

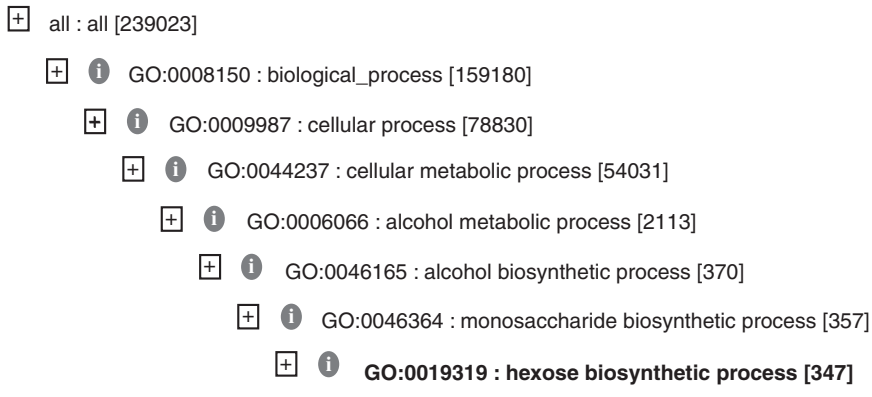

Figure 3.1 *Hierarchical organization for term GO:0019319 in Gene Ontology as displayed by Amigo(http://www.geneontology.org/) tool for searching and browsing Gene Ontology.*

3.7 Gene Ontology

GO consists of hierarchically structured vocabulary divided into three basic subcategories: molecular function, biological process and cellular component. Each term in GO is referred by an identifier of the form GO:xxxxxxx, a subcategory, and an associated textual description for that term. For example, the identifier GO:0019319 is of subcategory biological process and has short description as 'hexose biosynthetic process' (Figure 3.1). GO organizes the terms in a directed acyclic graph (DAG) structure where terms are associated by is_a or part_of relationships. The is_a classifier represents a subclass relationship where 'A is_a B' means A is description of B but at higher depth or more narrower description. 'A part_of B' indicates that whenever A is present it is part of B.

A gene can be described as performing one or more molecular functions, being part of one or more biological process and located in one or more cellular components. Another important feature of GO is that it supports association of an evidence code with each annotation indicating the nature of evidence sources that are used to support that annotation. Examples of the evidence codes are IDA (Inferred from Direct Assay), which indicates that a direct assay was carried out to determine the function, and ISS (Inferred from Sequence or Structural Similarity), which clarifies that any analysis based on sequence alignment, structure comparison, or evaluation of sequence features such as composition is performed.

3.8 Other Functional Ontologies

EC numbers are used for classifying enzymes based on the reactions they catalyze. The nomenclature of enzyme number has the form of EC x.x.x.x, consisting of four level hierarchies describing the activity of the enzyme. Partial EC numbers with only initial parts out of the four subparts will be used to refer to a class of enzymes describing a biochemical activity at a broader level. The FunCat scheme for functional description of proteins divides the annotations into 28 main categories that cover general fields. The FunCat version 2.1

includes 1362 functional categories where main categories are further subdivided up to six levels with increase in the specificity. A difference between FunCat and GO is that FunCat is organized in a hierarchical tree, while GO is structured into a DAG. A difference of enzymatic function description between FunCat and EC number is that EC number classifies catalytic activities based on the chemical reaction, while FunCat classification is based on the pathway where an enzyme acts. TCDB (Transport classification database)[61] is a database of Transporter Classification (TC) system that gives detailed comprehensive IUBMB (International Union of Biochemistry and Molecular Biology) approved classification system for membrane transport proteins. The TC system is analogous to the Enzyme Commission system for classification of enzymes, but additionally incorporates phylogenetic information. It consists of a set of representative protein sequences, most of which have been functionally characterized. These transporters are classified with a five-character designation, as follows: $D_1.L.D_2.D_3.D_4$. The letters in sequence correspond to transporter class, subclass, family, subfamily and transporter itself. The TCDB website also offers several tools specifically designed for analyzing the unique characteristics of transport proteins. The KEGG orthology (KO)[62] is both an ontology arranged around binary relations and an ontology giving annotations of class of gene products. KO decomposes the universe of all genes in all organisms into groups of functionally identical genes (orthologs). They define relationships between KEGG database objects such as reactions, substrates and products; relationships between enzyme and its location in the pathway; relationship between enzyme and protein super family to which it belongs.

3.9 Quantifying Functional Similarity

To compute the prediction accuracy of a function prediction we need to compare the similarity of predicted and actual ontology terms. The hierarchical nature of GO provides natural mechanism for comparing the terms. The basic idea is to consider the closest common parental node between predicted and correct GO terms. The scoring scheme used in the function prediction category in Critical Assessment of Techniques for Protein Structure Prediction 7 (CASP7) computes fraction of the path depth of the common parent compared with the path depth of the correct annotated GO term.[63] Resnik uses the maximum information content computed as maximum negative logarithm of any common ancestor term probability for pair of GO terms being compared.[64] Probability of occurrence of each term is defined as frequency of its occurrence in the annotation database as compared to the frequency of root term in the GO. Lord *et al*.[65] were first ones to compute the semantic similarity between a pair of proteins using Resnik's measure. Semantic similarity between two proteins was computed as the average similarity of the GO terms that annotate both the proteins. Schlicker *et al*.[66] further extend the Resnik's measure to include probabilities of both terms being compared for normalizing the semantic similarity score and also use the relevance (that decreases with probability) of the common ancestor term. Poze *et al*.[67] take a completely different approach to compute a functional distance between a pair of GO terms based on co-occurrence of terms in a same set of Interpro entries. A profile is constructed for GO terms representing its association with a set of Interpro domains taking into account the is_a relationships for GO terms and its ancestors. The profiles are used to generate a matrix of co-occurrences between GO terms.

3.10 Automated GO Term Prediction Methods

Recent years have observed development of new generation of function prediction algorithms. It is triggered by the growing need of function annotation of genes in an increasing number of newly sequenced genomes and newly solved protein tertiary structures. Moreover, large scale experimental data, such as protein–protein interaction and gene expression data, further add the urgency of developing different techniques to predict reliable annotations even at broad levels of detail for new genes. Many of the new generation of function prediction algorithms have some common features. First, they take advantage of controlled vocabulary of Gene Ontology, which facilitates computational handling of function terms. Second, most of them use BLAST or PSI-BLAST search results as the primary source of function information, realizing (or expecting) that homology search results contain more information than conventionally extracted by applying a unique E-value threshold to select significant hits. Third, some of the methods employ machine learning techniques, such as Support Vector Machines (SVM), that have recently become popular in bioinformatics area. Below we will discuss some of such methods.

Goblet[68,69] provides a web platform which assists users to analyze a BLAST search result of an input protein sequence in terms of GO terms. GO terms of retrieved sequences are displayed on the GO tree, which facilitates comparison of the GO terms. GOFigure[70] uses an idea of a minimum covering graph (MCG), which is a graph on the GO tree rooted at the GO terms that subsumes all extracted GO annotations from BLAST hits for a query sequence. The score assigned to each GO term is a weighted score of all the hits that map to it as well as the scores of all its children term. As a consequence of using MCG, not only the GO terms which directly associate to the retrieved BLAST hits but also their children terms have possibility of being final GO prediction to the query sequence. Verspoor *et al.*[71] use an ontology categorizer named POset Ontology Categorizer, which summarizes weighted collection of GO terms taken from PSI-BLAST hits. The weight of a GO term reflects the E-value of the sequence hit. For an evaluation metric of prediction, they introduce hierarchical precision and recall, which considers accuracy at each ancestral node of predicted and actual GO term.

GOtcha[72] runs BLAST for a query sequence, and GO terms are extracted from each BLAST hit. The set of GO terms and all ancestral terms are assigned a score of negative logarithm of the E-value of the BLAST hit (R-score). The sum of the R-score for all matches is normalized to the total R-score of the root node of each category in the GO tree.

GOPET[73] employs SVMs to analyze a BLAST search for a query sequence. GO terms are extracted from each retrieved sequence with attached features, including the E-value, the bit-score, the sequence identity, the coverage score, the alignment length, GO term frequency, and the evidence code of GO annotation, all of which are used as input parameters to SVMs. 99 SVM classifiers, each of which predicts a particular GO term, are constructed. An advantage of using SVM is that many different properties of retrieved sequences can be considered. On the other hand, a drawback is that a limited number of GO terms can be predicted by this implementation because a SVM needs to be constructed for individual GO term, and a sufficient number of instances (sequences) are needed for training a SVM.

ProtFun[74] is an interesting method of protein function prediction that is not based on sequence similarity but on sequence based protein features such as predicted post translational modifications, protein sorting signals, and physical/chemical properties calculated

from amino acid composition. They use the InterPro database which maps protein families to GO terms. For each GO class a standard feed-forward neural network with a single layer of hidden neurons was trained with different combinations of sequence derived features.

JAFA[75] is protein function meta-server that provides joint assembly of function predictions from five different prediction servers, namely, GOFigure,[70] Gotcha,[72] Goblet,[68] InterProScan,[76] and PhydBac2.[77] The score provided with each GO terms is the product of the GO level multiplied by the fraction of agreeing servers. Hence the scoring function rewards the predictions that are more specific and predicted by multiple servers.

SIFTER[78] models a phylogenomics procedure of annotating molecular function of genes in a probabilistic method. For a given query protein, a rooted phylogenetic tree is constructed using homologs taken from the Pfam database. Annotated GO terms to the proteins in the tree are represented as a vector, and the probabilities with which known GO terms are propagated to descendants are computed.

Another approach by Cai *et al*.[79] for predicting enzyme subclasses is based on the amino acid composition of a protein sequence. This is particularly useful when it is not possible to identify a subfamily class for protein using the sequence similarity approach. They have developed FunD-PseAA Intimate Sorting (ISort) predictor using domain information obtained from InterPro database and amino acid frequencies in the sequence.

Pattern analysis of the distributions of disordered regions has shown that functions of intrinsically disordered proteins are both length and position dependent. Lobley *et al*.[80] used location descriptors to encode the position of disordered regions in proteins and showed their correlations with GO categories by calculating the average frequency of disordered residues within different location windows for proteins sequences annotated by GO term. Their results suggest that disorder regions are more indicative of biological process than the molecular function and the information content of disorder feature set is comparably lower than that for secondary structure or amino acid composition.

3.11 Protein Function Prediction (PFP) Algorithm

Our group has developed PFP algorithm for function prediction which extends a conventional PSI-BLAST search[81] (Figure 3.2). Along with strong PSI-BLAST hits which have significant E-value, PFP also uses weak hits that are not generally considered for transferring annotations. Weakly similar hits that are not recognized as homologous to the query sequence are also used in PFP because they often share common functional domains or some functional similarity at a broader level. GO terms extracted from retrieved sequences are ranked according to the following equation considering the E-value assigned to the retrieved sequences. Currently sequences of an E-value of up to 100 are used:

$$s(f_a) = \sum_{i=1}^{N} \sum_{j=1}^{Nfunc(i)} \left((-\log(E\_value(i)) + b) P(f_a | f_j) \right), \tag{3.1}$$

where $s(f_a)$ is the final score assigned to the GO term, f_a, N is the number of the similar sequences retrieved by PSI-BLAST, $Nfunc(i)$ is the number of GO terms assigned to sequence j, $E\_value(i)$ is the E-value given to the sequence i, f_j is a GO term assigned to

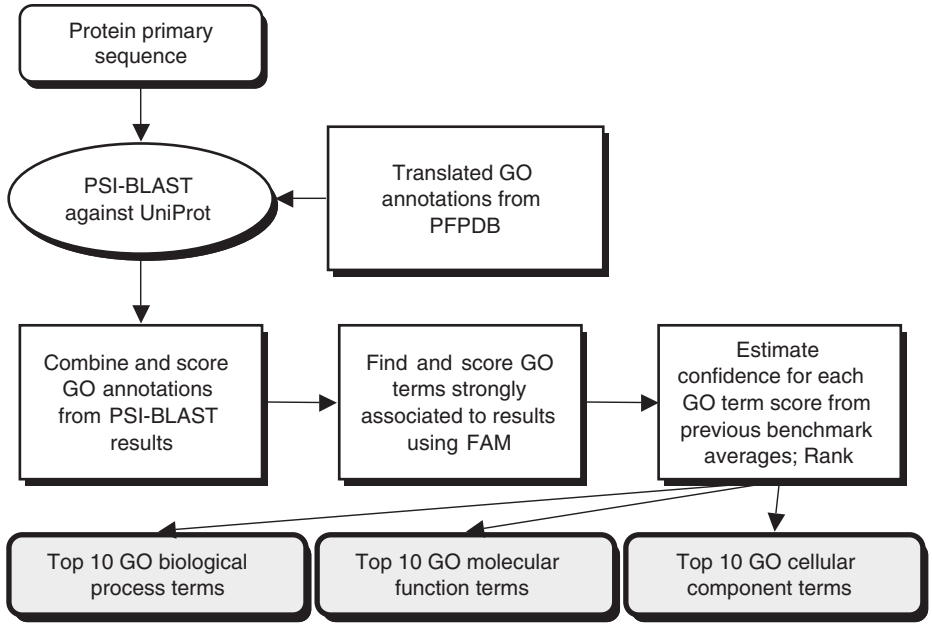

Figure 3.2 *Flowchart describing prediction method of PFP.*

the sequence i, and b is the constant value, 2 (=$\log_{10}100$), which keeps the score positive. $P(f_a|f_j)$ is the conditional probability that f_a is associated with f_j. This conditional probability is computed from co-occurrence of GO terms in single sequences in the UniProt database and stored in a two dimensional matrix named Function Association Matrix (FAM):

$$P(f_a|f_j) = \frac{c(f_a, f_j) + \varepsilon}{c(f_j) + \mu \cdot \varepsilon}, \tag{3.2}$$

$c(f_a, f_j)$ is number of times f_a and f_j are assigned simultaneously to each sequence in UniProt, and $c(f_j)$ is the total number of times f_j appeared in UniProt, μ is the size of one dimension of the FAM (i.e. the total number of unique GO terms), and ε is the pseudo-count.

The pre-computed FAM allows PFP to extract information about strongly associated terms in the database across the categories of GO which may be intuitive for biologists but not directly retrieved from the sequence database searched. For example, the (GO: 0008234) 'cysteine-type peptidase activity' in the molecular function category shows high association score with biological process term (GO:0006508) 'proteolysis' in the biological process. And molecular function (GO:0015662) 'ATPase activity, coupled to trans-membrane movement of ions, phosphorylative mechanism' is highly associated with the cellular component term (GO:0016020) membrane.

Moreover, scores given to each GO term are propagated to parent terms in the GO tree according to the number of genes associated to the predicted term relative to the parent

term:

$$s(f_p)\sum_{i=1}^{N_c}\left(s(f_{ci})\left(\frac{c(f_{ci})}{c(f_p)}\right)\right). \tag{3.3}$$

where $s(f_p)$ is the score of the parent term f_p, N_c is the number of child GO term which belong to the parent term f_p, $s(f_{ci})$ is the score of a child term ci, and $c(f_{ci})$ and $c(f_p)$ is the number of known genes which are annotated with function term f_{ci} and f_p in the Gene Ontology Annotation (GOA) database released at the European Bioinformatics Institute (EBI).

Since prediction crucially depends on available GO term annotations assigned to sequences in the database to be searched, we enriched annotated GO terms in the GOA database by adding GO terms from other databases including HAMAP, InterPro,[82] Pfam,[35] PRINTS,[34] ProDom,[33] PROSITE.[51] SMART,[83] and TIGRFam[84] as well as SwissProt Key Words.

Once a raw score of a GO term is obtained according to the equations above, its statistical significance is computed in terms of the P-value by considering the score distribution of that GO term taken from a benchmark dataset. And finally, predicted GO terms are ranked by their P-value in each of the three categories. It is important to consider the P-value rather than a raw score because some GO terms occur more frequently in a database, and thus tend to have a high raw score. For example, GO terms at a higher level in the GO tree (thus have more general function) have a high score also because scores given to its child terms are propagated to it.[85]

3.12 PFP Benchmark Results

In the paper published in 2006, we have benchmarked PFP on a set of randomly selected 2000 proteins from UniProt[81] (Figure 3.3). Three methods are compared: PFP using FAM to incorporate the GO term associations, PFP without using FAM, and transferring GO annotations from the top PSI-BLAST hit (top PSI-BLAST method). For the PFP predictions, five GO terms with the highest raw scores are predicted, and the top PSI-BLAST method predicts all the GO terms assigned to the top hit sequence. The performance was compared in terms of the sequence coverage, which reports the percentage of sequences for which correct biological process (sharing a common parent with a target annotation at GO depth ≥ 4) were predicted. To mimic a realistic situation that no significant homologs are found for a query protein sequence, the most significant sequence hits up to several E-value cutoffs in a PSI-BLAST search are ignored and only sequences with an E-value of the cutoff or larger (E-value $>$ 0, 0.01, 0.1, 1, 2, 3, 5,10, 15, 20, 25, 50, 100) were used.

When all retrieved sequences are used, PFP with FAM correctly predicted biological process over 80 % of the tested query sequences, while PFP without FAM and top PSI-BLAST method made correct prediction to approximately 72 % of the query sequences. The strength of PFP is more evident when top hit sequences up to a certain E-value are not used. When only retrieved sequences with an E-value of 10 or higher are used, PFP with FAM made correct predictions to around 50 % of the query sequences, which is about five times larger than the top PSI-BLAST method. Interestingly, the sequence coverage by PFP

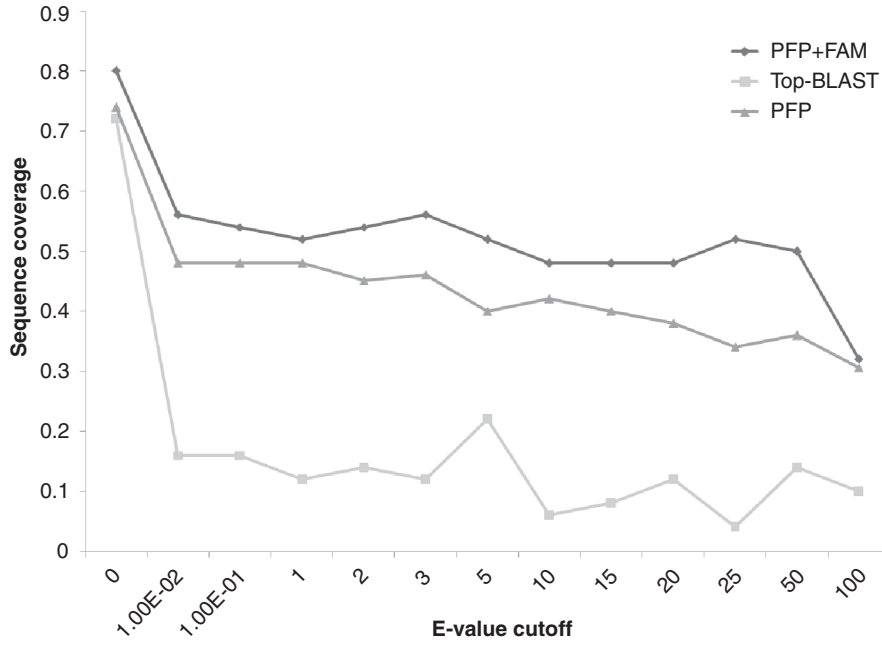

Figure 3.3 *Benchmark of PFP on a data set of 2000 sequences. Three methods are compared, PFP with FAM, PFP without FAM, and the top PSI-BLAST method. The data used in Figure 1 of our paper in 2006[81] is replotted.*

with FAM stays almost the same when the sequence hits of even larger E-value > 10 are used.

A characteristic advantage of PFP is that it can often predict a broader function or a 'low-resolution' function by identifying consensus GO terms which occur in retrieved sequences with a wide range of E-value by PSI-BLAST. Note that it is not trivial for conventional methods to make this kind of low-resolution function prediction, because there are no apparent sequence patterns for low-resolution functions. Conventional ways of using (PSI-)BLAST or motif searches are rather yes/no type prediction methods, meaning that a prediction is made when a clear functional sequence pattern is found, but no prediction is made otherwise. In contrast, PFP is able to make low-resolution function prediction when detailed function prediction cannot be made by taking consensus between function annotations of weakly similar sequences. In other words, PFP tries to give some functional clue to a query sequence by lowering resolution of function when necessary without sacrificing accuracy. An important point revealed by the benchmark study (Figure 3.3) is that the top hit by PSI-BLAST is not necessarily accurate and PFP outperforms the top PSI-BLAST method even when all retrieved sequences (with an E-value ≥ 0) are used. The pitfall of relying on only the top hit sequence has been pointed out by Galperin and Koonin.[54] PFP can often avoid transferring irrelevant annotations of the top hit sequence in a search by summarizing consensus GO terms which occur in a large number of hits in a PSI-BLAST search.

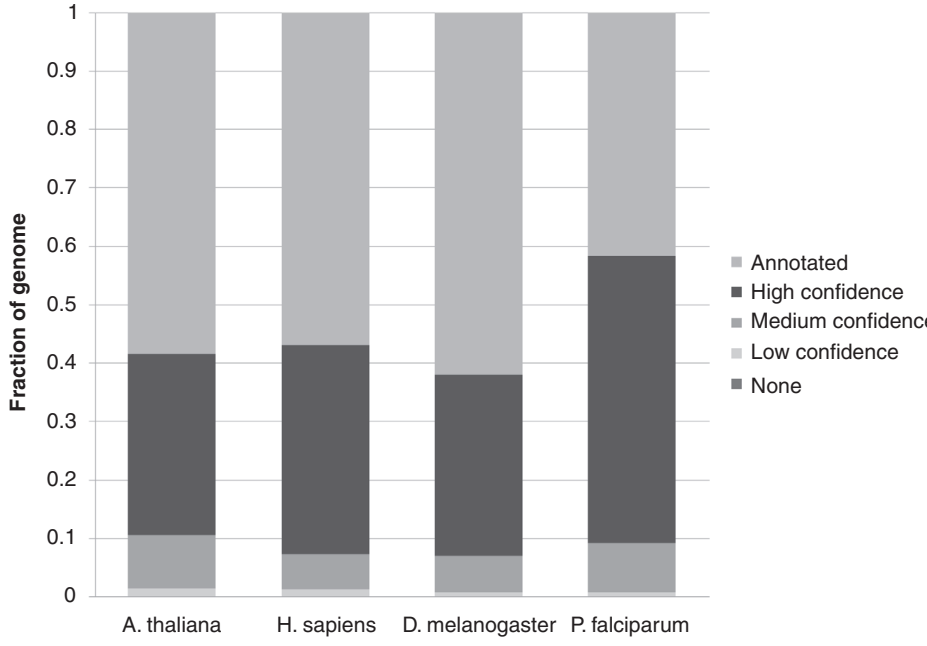

Figure 3.4 *Distribution of predictions done by PFP for four genomes classified based on the confidence score for the predicted annotations.* A. thaliana*,* H. sapiens*,* D. melanogaster*, and* P. falciparum. *Annotations of theses genomes are taken from the GOA database.*

A practical strength of PFP is that it can give function annotation to a larger number of genes in a genome by predicting low resolution functions, while typically BLAST searches can cover up to half of genes in a genome.[23] Very general function, e.g. transporter or enzyme, is not very helpful for designing biochemical experiments, but may be helpful for interpreting a large-scale data, such as gene expression data or protein–protein interaction data.[86] In Figure 3.4, fractions of genes with PFP annotations along with annotated genes in the GOA database for four organisms are shown. Predictions made by PFP are classified into three groups according to confidence level of the predictions, which are estimated by the correlation with the P-value and the accuracy in a benchmark dataset used.[85] For these genomes, PFP can provide function predictions to an additional 30–50 % of the total genes in a genome with a high confidence.

3.13 Comparative Genomics Based Methods

Completely different approaches for sequence-based function prediction use the genomic context of genes, taking advantage of the increasing number of available complete genomes. There are three major methods for this category. The first approach is to examine conservation of gene clusters in multiple genomes. Because gene locations tend to be dynamically shuffled during evolution,[87] if proximity between genes is evolutionarily

conserved across species (conserved gene clusters), there is a high likelihood of functional association between the genes.[88,89] Bacterial genomes have operon structures, which is a transcription unit with multiple genes,[90] but more conserved gene clusters are found which are not known operons. Another evidence of functional association of genes is domain fusion events.[91] If two separate genes in one organism are seen as fused domains occurring in a single protein in another organism, apparently the fusion does not interfere with function of the two genes, and most likely the two genes are involved in the same functional context. Similarity in the pattern of existence and absence of orthologous genes in genomes, which is called phylogenetic profiles,[92] also indicates functional association of genes. Bork's group has implemented these three comparative genomics based approaches in the STRING server.[93]

These comparative genomics-based methods will become more useful as the number of sequenced genomes will further increase. However, what can be predicted by these methods is functional association of genes but not functional terms of each gene. Thus, homology-based function prediction is still needed for the starting point of a genome scale annotation.

3.14 Subcellular Localization Prediction

Subcellular localization can be considered as a type of gene function. Indeed the Gene Ontology organizes terms for describing localization in a DAG named cellular component. Some proteins have a signal peptide typically at its N-terminal region, which are recognized by a transporting protein and later often cleaved off. Therefore a direct way to predict subcellular localization is to recognize these signals.[94] Since molecular protein sorting mechanism differs in prokaryotes and eukaryotes, prediction methods is usually specifically designed for either one of them or for a sub-category, such as plants. PSORT is one of the earliest prediction methods, which uses multiple sequence features including signal peptides, amino acid composition, sequence motifs, and predicted trans-membrane domains in the form of a decision rule or a classifier.[95,96] They have an extensive collection of links to prediction methods and related resources at their web site, http://www.psort.org,[97] Nair *et al*.[98] demonstrate that cellular localization is an evolutionarily conserved property and homologs tend to occur at the same cellular sites. Proteome Analyst[99] obtains annotations corresponding to homologous sequences detected using BLAST and then uses them with an organism specific Bayesian classifier to classify the query protein to localization sites. Some methods[100–102] use SVM to classify proteins across different cellular components based on the frequency of twenty amino acids. The phylogenetic profile can be also used to predict localization.[103]

3.15 Identification of Functionally Important Residues

Usually molecular function of proteins, such as catalytic activity of enzymes, is carried out by a small number of residues in a protein sequence. These functionally indispensable residues are identified experimentally by constructing point mutation/deletion or domain deletion mutants, or from the tertiary structure in a ligand bound form solved by X-ray

crystallography or NMR. Databases such as PROSITE[51] and ELM[104] (for eukaryotes) store such short sequence motifs. Since a local alignment of these short motifs does not result in an alignment score which yields a significant E-value in a BLAST search, searching against a motif database is a complementary method to homology search for function prediction. If the tertiary structure of the target protein is known, conservation of residues which are not close on the sequence but locate in spatial proximity can be further detected and compared against a database of three-dimensional motifs.[44,105–107] See Chapter 7 by Kinoshita in this volume for more details on structure-based function prediction.

Functionally important residues are generally well conserved among orthologous proteins, thus, selecting conserved residues from a carefully constructed MSA of a protein family is a fundamental procedure of identifying functionally important residues.[108–112] Besides sequence conservation, combining local structure information helps accurately identifying functionally important residues.[113] Some methods are developed that identify residue positions in a MSA which discriminate predefined subfamilies thus considered to be functional residues specific to subfamilies.[114–116] In contrast, Pei's method starts with constructing a phylogenetic tree for a given set of sequences, and identifies residue positions in the MSA which have a high likelihood that follows evolution along the tree.[117] Casari *et al.*[118] apply principal component analysis to a matrix representing sequences of a family to identify groups of residues that are conserved in the whole family and also those which are specific to subfamilies. MINER is based on the finding by La and Livesay that sequence regions which show a mutation pattern that conserves the overall familial phylogeny correspond to functional sites.[119,120]

3.16 Function Prediction Competitions

Responding to the increasing need of automatic function prediction, the bioinformatics community has held function prediction contests in the last few years. Friedberg, Godzik, and their co-workers have held the Automated Function Prediction Special Interest Group meeting at ISMB 2005,[121] where they summarized the results of a blind prediction contest of protein gene function. The participants were required to set up an automatic web server which accepts protein sequences, to which the organizers submitted target sequences and evaluated returned predictions. The Critical Assessment of Techniques for Protein Structure Prediction (CASP) competitions included a function prediction category in CASP6 (2004)[122] and CASP7 (2006).[63] Target protein sequences were given to predict EC numbers, GO terms or active site/ligand binding site residues. In both AFP-SIG and CASP7, PFP had the highest overall score[63] (no ranking was given in CASP6). Objective evaluation of existing methods is essential for enhancing continuous improvements of the methods and for keeping the field active. A larger number of participants are expected to participate in these competitions in the future.

3.17 Summary

We have reviewed recent advances of sequence-based function prediction methods. Figure 3.5 summarizes different techniques for predicting function from sequence. The

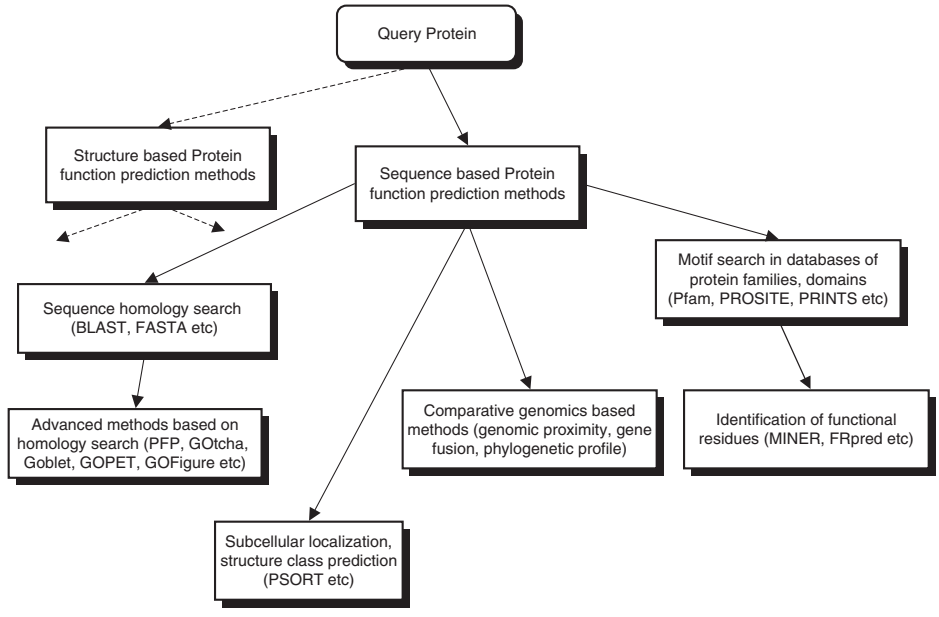

Figure 3.5 *Summary of sequence-based function prediction methods.*

first step is to perform homology search using BLAST, PSI-BLAST or FASTA. Also it is recommended to perform motif and domain searches, such as Pfam and PROSITE. If significant hits are not found, some of recent methods which expand homology search, such as PFP, could be performed. If reasonable results are still not obtained, we recommend the STRING server, which performs comparative genomics based approaches. However, note that comparative genomics methods don't predict specific functional terms of a query protein, rather shows a set of proteins which are predicted to be functionally related to the query protein. If knowing a broad class of protein is useful, subcellular localization prediction and some local structure class predictions, such as prediction of transmembrane proteins[123] will be worthwhile to try. Finally, functional residue prediction methods, e.g. MINER, will be informative for some purposes, but note that these methods are aimed to select residues for function, not to predict functional terms. Refer to Table 3.1 for available online tools.

The need of function prediction is increasing, especially for interpreting large-scale omics data. This situation is very different from more than ten years ago when BLAST, FASTA, and PSI-BLAST were developed. Automatic function prediction methods will evolve in harmony with new developments of experimental methods by incorporating those experimental data in prediction algorithms and by helping biological reasoning of experimental data. More advances in this field are expected in the near future keeping pace with the other bioinformatics areas described in the other chapters in this book.

Table 3.1 *Protein function prediction methods*

Name	WWW Address	Description
BLAST[16], PSI-BLAST[28]	http://www.ncbi.nlm.nih.gov/blast/	Sequence homology search
FASTA[15]	http://www.ebi.ac.uk/fasta33/	Sequence homology search
PFP[81]	http://dragon.bio.purdue.edu/pfp/	BLAST-based GO term prediction + association mining
GOtcha[72]	http://www.compbio.dundee.ac.uk/gotcha/gotcha.php	BLAST-based GO term prediction
GOblet[68,69]	http://goblet.molgen.mpg.de/	BLAST-based GO term prediction
GOPET[73]	http://genius.embnet.dkfz-heidelberg.de/menu/biounit/open-husar	BLAST-based GO term prediction by SVM
ProtFun[74]	http://www.cbs.dtu.dk/services/ProtFun/	Sequence feature based function classification
OntoBlast[124]	http://functionalgenomics.de/ontogate/	BLAST-based GO term prediction
FIGENIX[125]	http://sites.univ-provence.fr/evol/figenix/	Genomic annotation using phylogenomic approaches
JAFA[75]	http://jafa.burnham.org/	GO term prediction metaserver
Pfam[35]	http://pfam.sanger.ac.uk/	Protein family HMM database
SMART[126]	http://smart.embl-heidelberg.de/	Sequence fingerprint scanning
ProDom[33]	http://prodom.prabi.fr	Protein domain sequence database
BLOCKS[32]	http://blocks.fhcrc.org/	Protein domain sequence database
PRINTS[34]	http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/	Protein fingerprint database
ELM[104]	http://elm.eu.org/	Functional motif
PROSITE[51]	http://ca.expasy.org/prosite/	Database of protein domains, families and functional sites
InterProScan[82]	http://www.ebi.ac.uk/InterProScan/	Functional motif search
ScanProsite[127]	http://www.expasy.ch/prosite/	Functional motif scanning
STRING[93]	http://string.embl.de/	Comparative genomics approaches
FRpred[128]	http://toolkit.tuebingen.mpg.de/frpred	Prediction of protein functional residues
MINER[120]	http://coit-apple01.uncc.edu/MINER/	Functional residue prediction
PSORT[97]	http://www.psort.org/	PSORT family of programs for subcellular localization prediction
SignalP[94]	http://www.cbs.dtu.dk/services/SignalP/	Prediction of the presence and location of signal peptide cleavage sites

CELLO[100]	http://cello.life.nctu.edu.tw/	subCELlular LOcalization predictor
SubLoc[102]	http://www.bioinfo.tsinghua.edu.cn/SubLoc/	Prediction of Protein Subcellular Localization by Support Vector Machine
LOCtree[129]	http://cubic.bioc.columbia.edu/cgi/var/nair/loctree/query	Prediction of Protein Subcellular Localization by Support Vector Machine
TMHMM[123]	http://www.cbs.dtu.dk/services/TMHMM-2.0/	Prediction of transmembrane helices in proteins
BOMP[130]	http://www.bioinfo.no/tools/bomp	Tool for prediction of beta-barrel integral outer membrane proteins
PROFtmb[131]	http://rostlab.org/cgi-bin/var/bigelow/proftmb/query	Per-residue and whole-proteome prediction of bacterial transmembrane beta barrels
LipoP[132]	http://www.cbs.dtu.dk/services/LipoP/	Prediction of lipoproteins and signal peptides in Gram negative bacteria

Acknowledgement

This work is partially supported by National Institute of General Medical Sciences of the National Institutes of Health (U24 GM077905 and R01GM075004), and the National Science Foundation (DMS 0604776).

References

1. J.P. Gogarten, and L. Olendzenski, Orthologs, paralogs and genome comparisons, *Curr Opin Genet Dev*, **9**, 630–636 (1999).
2. Z. Gu, L.M. Steinmetz, X. Gu, C. Scharfe, R.W. Davis, and W.H. Li, Role of duplicate genes in genetic robustness against null mutations, *Nature*, **421**, 63–66 (2003).
3. S. Ohno, *Evolution by Gene Duplication*, George Allen & Unwin, London, 1970.
4. Y. Boucher, C.J. Douady, R.T. Papke, et al. Lateral gene transfer and the origins of prokaryotic groups, *Annu Rev Genet*, **37**, 283–328 (2003).
5. W.F. Doolittle, Phylogenetic classification and the universal tree, *Science*, **284**, 2124–2129 (1999).
6. W.M. Fitch, Distinguishing homologous from analogous proteins, *Syst Zool*, **19**, 99–113 (1970).
7. W. Tian, and J. Skolnick, How well is enzyme function conserved as a function of pairwise sequence identity? *J. Mol. Biol.*, **333**, 863–882 (2003).
8. Y. Van de Peer, Evolutionary genetics: When duplicated genes don't stick to the rules, *Heredity*, **96**, 204–205 (2006).
9. C.B. Anfinsen, Principles that govern the folding of protein chains, *Science*, **181**, 223–230 (1973).
10. C. Chothia, and A.M. Lesk, The relation between the divergence of sequence and structure in proteins, *EMBO J.*, **5**, 823–826 (1986).
11. S.E. Brenner, C. Chothia, and T.J. Hubbard, Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships, *Proc. Natl. Acad. Sci. U.S.A*, **95**, 6073–6078 (1998).
12. C.A. Orengo, D.T. Jones, and J.M. Thornton, Protein superfamilies and domain superfolds, *Nature*, **372**, 631–634 (1994).
13. S.B. Needleman, and C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, *J. Mol. Biol.*, **48**, 443–453 (1970).
14. T.F. Smith, and M.S. Waterman, Identification of common molecular subsequences, *J. Mol. Biol.*, **147**, 195–197 (1981).
15. W.R. Pearson, and D.J. Lipman, Improved tools for biological sequence comparison, *Proc Natl Acad Sci U S A*, **85**, 2444–2448 (1988).
16. S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, Basic local alignment search tool, *J Mol Biol*, **215**, 403–410 (1990).
17. T. Hulsen, J. de Vlieg, and P.M. Groenen, Phylopat: Phylogenetic pattern analysis of eukaryotic genes, *BMC Bioinformatics*, **7**, 398 (2006).
18. W.R. Pearson, Comparison of methods for searching protein sequence databases, *Protein Sci*, **4**, 1145–1160 (1995).
19. W.R. Pearson, Effective protein sequence comparison, *Methods Enzymol.*, **266**, 227–258 (1996).
20. W.R. Pearson, Flexible sequence similarity searching with the Fasta3 program package, *Methods Mol Biol*, **132**, 185–219 (2000).
21. E.S. Lander, L.M. Linton, B. Birren, *et al.*, Initial sequencing and analysis of the human genome, *Nature*, **409**, 860–921 (2001).
22. S.G. Oliver, Q.J. Van Der Aart, M.L. Agostoni-Carbone, et al. The complete DNA sequence of yeast chromosome Iii, *Nature*, **357**, 38–46 (1992).
23. T. Hawkins, and D. Kihara, Function prediction of uncharacterized proteins, *J. Bioinform. Comput. Biol.*, **5**, 1–30 (2007).

24. B. John, and A. Sali, Detection of homologous proteins by an intermediate sequence search, *Protein Sci*, **13**, 54–62 (2004).
25. J. Park, S.A. Teichmann, T. Hubbard, and C. Chothia, Intermediate sequences increase the detection of homology between sequences, *J Mol Biol*, **273**, 349–354 (1997).
26. I. Alam, A. Dress, M. Rehmsmeier, and G. Fuellen, Comparative homology agreement search: An effective combination of homology-search methods, *Proc Natl Acad Sci U S A*, **101**, 13814–13819 (2004).
27. C. Webber, and G.J. Barton, Increased coverage obtained by combination of methods for protein sequence database searching, *Bioinformatics*, **19**, 1397–1403 (2003).
28. S.F. Altschul, T.L. Madden, A.A. Schaffer, *et al*., Gapped Blast and Psi-Blast: A new generation of protein database search programs, *Nucleic Acids Res*, **25**, 3389–3402 (1997).
29. W.R. Pearson, and M.L. Sierk, The limits of protein sequence comparison? *Curr Opin Struct Biol*, **15**, 254–260 (2005).
30. A.A. Schaffer, L. Aravind, T.L. Madden, *et al*., Improving the accuracy of Psi-Blast protein database searches with composition-based statistics and other refinements, *Nucleic Acids Res*, **29**, 2994–3005 (2001).
31. A.A. Schaffer, Y.I. Wolf, C.P. Ponting, E.V. Koonin, L. Aravind, and S.F. Altschul, Impala: Matching a protein sequence against a collection of Psi-Blast-constructed position-specific score matrices, *Bioinformatics*, **15**, 1000–1011 (1999).
32. J.G. Henikoff, E.A. Greene, S. Pietrokovski, and S. Henikoff, Increased coverage of protein families with the Blocks database servers, *Nucleic Acids Res*., **28**, 228–230 (2000).
33. C. Bru, E. Courcelle, S. Carrere, Y. Beausse, S. Dalmar, and D. Kahn, The Prodom database of protein domain families: More emphasis on 3D, *Nucleic Acids Res*., **33**, D212–D215 (2005).
34. T.K. Attwood, P. Bradley, D.R. Flower, *et al*., Prints and its automatic supplement, Preprints, *Nucleic Acids Res*., **31**, 400–402 (2003).
35. R.D. Finn, J. Mistry, B. Schuster-Bockler, *et al*., Pfam: Clans, web tools and services, *Nucleic Acids Res*., **34**, D247–D251 (2006).
36. D. Wilson, M. Madera, C. Vogel, C. Chothia, and J. Gough, The Superfamily database in 2007: Families and functions, *Nucleic Acids Res*, **35**, D308–313 (2007).
37. S.R. Eddy, Hidden Markov models, *Curr Opin Struct Biol*, **6**, 361–366 (1996).
38. R.I. Sadreyev, D. Baker, and N.V. Grishin, Profile-profile comparisons by Compass predict intricate homologies between protein families, *Protein Sci*, **12**, 2262–2272 (2003).
39. K. Ginalski, N.V. Grishin, A. Godzik, and L. Rychlewski, Practical lessons from protein structure prediction, *Nucleic Acids Res*., **33**, 1874–1891 (2005).
40. L. Rychlewski, L. Jaroszewski, W. Li, and A. Godzik, Comparison of sequence profiles. Strategies for structural predictions using sequence information, *Protein Sci*., **9**, 232–241 (2000).
41. R.L. Dunbrack, Jr., Sequence comparison and protein structure prediction, *Curr. Opin. Struct. Biol*., **16**, 374–384 (2006).
42. D. Kihara, and J. Skolnick, Microbial genomes have over 72 % structure assignment by the threading algorithm prospector_Q, *Proteins*, **55**, 464–473 (2004).
43. K. Kinoshita, and H. Nakamura, Protein informatics towards function identification, *Curr. Opin. Struct. Biol*., **13**, 396–400 (2003).
44. J.S. Fetrow, A. Godzik, and J. Skolnick, Functional analysis of the escherichia coli genome using the sequence- to-structure-to-function paradigm: Identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity, *J Mol Biol*, **282**, 703–711 (1998).
45. M.L. Green, and P.D. Karp, A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases, *BMC. Bioinformatics*, **5**, 76 (2004).
46. R.K. Curtis, M. Oresic, and A. Vidal-Puig, pathways to the analysis of microarray data, *Trends Biotechnol*, **23**, 429–435 (2005).
47. R. Sharan, I. Ulitsky, and R. Shamir, Network-based prediction of protein function, *Mol Syst Biol*, **3**, 88 (2007).
48. J.D. Watson, R.A. Laskowski, and J.M. Thornton, Predicting protein function from sequence and structural data, *Curr. Opin. Struct. Biol*., **15**, 275–284 (2005).
49. D. Kihara, D.Y. Yang, and T. Hawkins, Bioinformatics resources for cancer research with an emphasis on gene function and structure prediction tools, *Cancer Informatics*, **2**, 25–35 (2006).

50. C.H. Wu, R. Apweiler, A. Bairoch, *et al*., The Universal Protein Resource (Uniprot): An expanding universe of protein information, *Nucleic Acids Res*, **34**, D187–191 (2006).
51. N. Hulo, A. Bairoch, V. Bulliard, *et al*., The 20 years of prosite, *Nucleic Acids Res*, **36**, D245–249 (2008).
52. S.E. Brenner, Errors in genome annotation, *Trends Genet*, **15**, 132–133 (1999).
53. W.R. Gilks, B. Audit, D. de Angelis, S. Tsoka, and C.A. Ouzounis, Percolation of annotation errors through hierarchically structured protein sequence databases, *Math Biosci*, **193**, 223–234 (2005).
54. M.Y. Galperin, and E.V. Koonin, Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption, *In Silico Biol*, **1**, 55–67 (1998).
55. D. Devos, and A. Valencia, Intrinsic errors in genome annotation, *Trends Genet*, **17**, 429–431 (2001).
56. M. Riley, T. Abe, M.B. Arnaud, *et al*., Escherichia Coli K-12: A cooperatively developed annotation snapshot – 2005, *Nucleic Acids Res*, **34**, 1–9 (2006).
57. S.L. Salzberg, Genome re-annotation: A Wiki solution? *Genome Biol*, **8**, 102 (2007).
58. M.A. Harris, J. Clark, A. Ireland, *et al*, The Gene Ontology (Go) database and informatics resource, *Nucleic Acids Res*, **32**, D258–261 (2004).
59. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (Nc-Iubmb), Enzyme Supplement 5 (1999), *Eur J Biochem*, **264**, 610–650 (1999).
60. A. Ruepp, A. Zollner, D. Maier, *et al*., The Funcat, a functional annotation scheme for systematic classification of proteins from whole genomes, *Nucleic Acids Res*, **32**, 5539–5545 (2004).
61. M.H. Saier, Jr., C.V. Tran, and R.D. Barabote, Tcdb: The Transporter Classification Database for membrane transport protein analyses and information, *Nucleic Acids Res*., **34**, D181–186 (2006).
62. M. Kanehisa, M. Araki, S. Goto, *et al*., Kegg for linking genomes to life and the environment, *Nucleic Acids Res*, **36**, D480–484 (2008).
63. G. Lopez, A. Rojas, M. Tress, and A. Valencia, Assessment of predictions submitted for the Casp7 function prediction category, *Proteins*, **69 Suppl 8**, 165–174 (2007).
64. P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (1995).
65. P.W. Lord, R.D. Stevens, A. Brass, and C.A. Goble, Investigating semantic similarity measures across the gene ontology: The relationship between sequence and annotation, *Bioinformatics*, **19**, 1275–1283 (2003).
66. A. Schlicker, F.S. Domingues, J. Rahnenfuhrer, and T. Lengauer, A new measure for functional similarity of gene products based on gene ontology, *BMC Bioinformatics*, **7**, 302 (2006).
67. A. Del Pozo, F. Pazos, and A. Valencia, Defining functional distances over gene ontology, *BMC Bioinformatics*, **9**, 50 (2008).
68. D. Groth, H. Lehrach, and S. Hennig, Goblet: A platform for gene ontology annotation of anonymous sequence data, *Nucleic Acids Res*, **32**, W313–317 (2004).
69. S. Hennig, D. Groth, and H. Lehrach, Automated gene ontology annotation for anonymous sequence data, *Nucleic Acids Res*, **31**, 3712–3715 (2003).
70. S. Khan, G. Situ, K. Decker, and C.J. Schmidt, Gofigure: Automated gene ontology annotation, *Bioinformatics*, **19**, 2484–2485 (2003).
71. K. Verspoor, J. Cohn, S. Mniszewski, and C. Joslyn, A categorization approach to automated ontological function annotation, *Protein Sci*, **15**, 1544–1549 (2006).
72. D.M. Martin, M. Berriman, and G.J. Barton, Gotcha: A new method for prediction of protein function assessed by the annotation of seven genomes, *BMC Bioinformatics*, **5**, 178 (2004).
73. A. Vinayagam, C. del Val, F. Schubert, *et al*., Gopet: A tool for automated predictions of gene ontology terms, *BMC Bioinformatics*, **7**, 161 (2006).
74. L.J. Jensen, R. Gupta, H.H. Staerfeldt, and S. Brunak, Prediction of human protein function according to gene ontology categories, *Bioinformatics*, **19**, 635–642 (2003).
75. I. Friedberg, T. Harder, and A. Godzik, Jafa: A protein function annotation meta-server, *Nucleic Acids Res*, **34**, W379–381 (2006).

76. E.M. Zdobnov, and R. Apweiler, Interproscan: An integration platform for the signature-recognition methods in Interpro, *Bioinformatics*, **17**, 847–848 (2001).
77. F. Enault, K. Suhre, and J.M. Claverie, Phydbac 'Gene Function Predictor': A gene annotation tool based on genomic context analysis, *BMC Bioinformatics*, **6**, 247 (2005).
78. B.E. Engelhardt, M.I. Jordan, K.E. Muratore, and S.E. Brenner, Protein molecular function prediction by Bayesian phylogenomics, *PLoS Comput Biol*, **1**, e45 (2005).
79. Y.D. Cai, and K.C. Chou, Predicting enzyme subclass by functional domain composition and pseudo amino acid composition, *J Proteome Res*, **4**, 967–971 (2005).
80. A. Lobley, M.B. Swindells, C.A. Orengo, and D.T. Jones, Inferring function using patterns of native disorder in proteins, *PLoS Comput Biol*, **3**, e162 (2007).
81. T. Hawkins, S. Luban, and D. Kihara, Enhanced automated function prediction using distantly related sequences and contextual association by PFP, *Protein Sci*, **15**, 1550–1556 (2006).
82. N.J. Mulder, R. Apweiler, T.K. Attwood, *et al*., New developments in the Interpro Database, *Nucleic Acids Res*, **35**, D224–228 (2007).
83. I. Letunic, R.R. Copley, S. Schmidt, *et al*., Smart 4.0: Towards genomic data integration, *Nucleic Acids Res*, **32**, D142–144 (2004).
84. D.H. Haft, J.D. Selengut, and O. White, The Tigrfams database of protein families, *Nucleic Acids Res*., **31**, 371–373 (2003).
85. T. Hawkins, M. Chitale, S. Luban, and D. Kihara, PFP: Automated prediction of gene ontology functional annotations with confidence scores, *Proteins*, Epub. (2008). DOI 10.1002/prot.22172
86. T. Hawkins, M. Chitale, and D. Kihara, New paradigm in protein function prediciton for large scale omics analysis, *Molecular BioSystems*, **4**, 223–231 (2008).
87. H. Watanabe, H. Mori, T. Itoh, and T. Gojobori, Genome plasticity as a paradigm of eubacteria evolution, *J. Mol. Evol*., **44 Suppl 1**, S57–64 (1997).
88. R. Overbeek, M. Fonstein, M. D'Souza, G.D. Pusch, and N. Maltsev, The use of gene clusters to infer functional coupling, *Proc. Natl. Acad. Sci. U.S.A*, **96**, 2896–2901 (1999).
89. T. Dandekar, B. Snel, M. Huynen, and P. Bork, Conservation of gene order: A fingerprint of proteins that physically interact, *Trends Biochem. Sci*., **23**, 324–328 (1998).
90. H. Salgado, G. Moreno-Hagelsieb, T.F. Smith, and J. Collado-Vides, Operons in escherichia coli: genomic analyses and predictions, *Proc Natl Acad Sci U S A*, **97**, 6652–6657 (2000).
91. A.J. Enright, I. Iliopoulos, N.C. Kyrpides, and C.A. Ouzounis, Protein interaction maps for complete genomes based on gene fusion events, *Nature*, **402**, 86–90 (1999).
92. M. Pellegrini, E.M. Marcotte, M.J. Thompson, D. Eisenberg, and T.O. Yeates, Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles, *Proc. Natl. Acad. Sci. U.S.A*, **96**, 4285–4288 (1999).
93. B. Snel, G. Lehmann, P. Bork, and M.A. Huynen, String: A web-server to retrieve and display the repeatedly occurring neighbourhood of a gene, *Nucleic Acids Res*., **28**, 3442–3444 (2000).
94. O. Emanuelsson, S. Brunak, G. von Heijne, and H. Nielsen, Locating proteins in the cell using Targetp, Signalp and related tools, *Nat Protoc*, **2**, 953–971 (2007).
95. K. Nakai, and P. Horton, Psort: A program for detecting sorting signals in proteins and predicting their subcellular localization, *Trends Biochem Sci*, **24**, 34–36 (1999).
96. J.L. Gardy, C. Spencer, K. Wang, *et al*., Psort-B: Improving protein subcellular localization prediction for gram-negative bacteria, *Nucleic Acids Res*., **31**, 3613–3617 (2003).
97. J.L. Gardy, and F.S. Brinkman, Methods for predicting bacterial protein subcellular localization, *Nat Rev Microbiol*, **4**, 741–751 (2006).
98. R. Nair, and B. Rost, Sequence conserved for subcellular localization, *Protein Sci*, **11**, 2836–2847 (2002).
99. Z. Lu, D. Szafron, R. Greiner, *et al*., Predicting subcellular localization of proteins using machine-learned classifiers, *Bioinformatics*, **20**, 547–556 (2004).
100. C.S. Yu, C.J. Lin, and J.K. Hwang, Predicting subcellular localization of proteins for gram-negative bacteria by support vector machines based on N-peptide compositions, *Protein Sci*, **13**, 1402–1406 (2004).
101. J. Wang, W.K. Sung, A. Krishnan, and K.B. Li, protein subcellular localization prediction for gram-negative bacteria using amino acid subalphabets and a combination of multiple support vector machines, *BMC. Bioinformatics*, **6**, 174 (2005).

102. S. Hua, and Z. Sun, Support vector machine approach for protein subcellular localization prediction, *Bioinformatics*, **17**, 721–728 (2001).
103. E.M. Marcotte, I. Xenarios, A.M. Van Der Bliek, and D. Eisenberg, Localizing proteins in the cell from their phylogenetic profiles, *Proc Natl Acad Sci U S A*, **97**, 12115–12120 (2000).
104. P. Puntervoll, R. Linding, C. Gemund, *et al*. Elm server: A new resource for investigating short functional sites in modular eukaryotic proteins, *Nucleic Acids Res*, **31**, 3625–3630 (2003).
105. J.W. Torrance, G.J. Bartlett, C.T. Porter, and J.M. Thornton, Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families, *J. Mol. Biol*., **347**, 565–581 (2005).
106. O. Lichtarge, and M.E. Sowa, Evolutionary predictions of binding surfaces and interactions, *Curr. Opin. Struct. Biol*., **12**, 21–27 (2002).
107. S. Jones, and J.M. Thornton, Searching for functional sites in protein structures, *Curr Opin Chem Biol*, **8**, 3-7 (2004).
108. W. Tian, A.K. Arakaki, and J. Skolnick, Eficaz: A comprehensive approach for accurate genome-scale enzyme function inference, *Nucleic Acids Res*., **32**, 6226–6239 (2004).
109. M.N. Wass, and M.J. Sternberg, Confunc – functional annotation in the twilight zone, *Bioinformatics*, (2008).
110. J.A. Capra, and M. Singh, Predicting functionally important residues from sequence conservation, *Bioinformatics*, **23**, 1875–1882 (2007).
111. S. Chakrabarti, and C.J. Lanczycki, Analysis and prediction of functionally important sites in proteins, *Protein Sci*, **16**, 4–13 (2007).
112. B. Sterner, R. Singh, and B. Berger, Predicting and annotating catalytic residues: An information theoretic approach, *J Comput Biol*, **14**, 1058–1073 (2007).
113. J.D. Fischer, C.E. Mayer, and J. Soding, Prediction of protein functional residues from sequence by probability density estimation, *Bioinformatics*, **24**, 613–620 (2008).
114. O.V. Kalinina, A.A. Mironov, M.S. Gelfand, and A.B. Rakhmaninova, Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families, *Protein Sci*, **13**, 443–456 (2004).
115. S.S. Hannenhalli, and R.B. Russell, Analysis and prediction of functional sub-types from protein sequence alignments, *J Mol Biol*, **303**, 61–76 (2000).
116. F. Pazos, A. Rausell, and A. Valencia, Phylogeny-independent detection of functional residues, *Bioinformatics*, **22**, 1440–1448 (2006).
117. J. Pei, W. Cai, L.N. Kinch, and N.V. Grishin, Prediction of functional specificity determinants from protein sequences using log-likelihood ratios, *Bioinformatics*, **22**, 164–171 (2006).
118. G. Casari, C. Sander, and A. Valencia, A method to predict functional residues in proteins, *Nat Struct Biol*, **2**, 171–178 (1995).
119. D. La, and D.R. Livesay, Predicting functional sites with an automated algorithm suitable for heterogeneous datasets, *BMC. Bioinformatics*., **6**, 116 (2005).
120. D. La, B. Sutch, and D.R. Livesay, Predicting protein functional sites with phylogenetic motifs, *Proteins*, **58**, 309–320 (2005).
121. I. Friedberg, M. Jambon, and A. Godzik, New avenues in protein function prediction, *Protein Sci*, **15**, 1527–1529 (2006).
122. S. Soro, and A. Tramontano, The prediction of protein function at Casp6, *Proteins*, **61 Suppl 7**, 201–213 (2005).
123. A. Krogh, B. Larsson, G. von Heijne, and E.L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes, *J Mol Biol*, **305**, 567–580 (2001).
124. G. Zehetner, Ontoblast function: From sequence similarities directly to potential functional annotations by ontology terms, *Nucleic Acids Res*., **31**, 3799–3803 (2003).
125. P. Gouret, V. Vitiello, N. Balandraud, A. Gilles, P. Pontarotti, and E.G. Danchin, Figenix: Intelligent automation of genomic annotation: Expertise integration in a new software platform, *BMC Bioinformatics*, **6**, 198 (2005).
126. I. Letunic, R.R. Copley, B. Pils, S. Pinkert, J. Schultz, and P. Bork, Smart 5: Domains in the context of genomes and networks, *Nucleic Acids Res*., **34**, D257–260 (2006).

127. N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, C.E. De, P.S. Langendijk-Genevaux, M. Pagni, and C.J. Sigrist, The Prosite Database, *Nucleic Acids Res.*, **34**, D227–230 (2006).
128. J.D. Fischer, C.E. Mayer, and J. Soding, Prediction of protein functional residues from sequence by probability density estimation, *Bioinformatics*, (2008).
129. R. Nair, and B. Rost, Mimicking cellular sorting improves prediction of subcellular localization, *J Mol Biol*, **348**, 85–100 (2005).
130. F.S. Berven, K. Flikka, H.B. Jensen, and I. Eidhammer, Bomp: A program to predict integral beta-barrel outer membrane proteins encoded within genomes of gram-negative bacteria, *Nucleic Acids Res*, **32**, W394–399 (2004).
131. H.R. Bigelow, D.S. Petrey, J. Liu, D. Przybylski, and B. Rost, Predicting transmembrane beta-barrels in proteomes, *Nucleic Acids Res*, **32**, 2566–2577 (2004).
132. W.T. Doerrler, Lipid trafficking to the outer membrane of gram-negative bacteria, *Mol. Microbiol.*, **60**, 542–552 (2006).