Amino Acids (2008) 35:517–530
DOI 10.1007/s00726-008-0077-y

REVIEW ARTICLE

Protein function prediction with high-throughput data

Xing-Ming Zhao · Luonan Chen · Kazuyuki Aihara

Received: 31 January 2008 / Accepted: 13 March 2008 / Published online: 22 April 2008

Abstract Protein function prediction is one of the main challenges in post-genomic era. The availability of large amounts of high-throughput data provides an alternative approach to handling this problem from the computational viewpoint. In this review, we provide a comprehensive description of the computational methods that are currently applicable to protein function prediction, especially from the perspective of machine learning. Machine learning techniques can generally be classified as supervised learning, semi-supervised learning and unsupervised learning. By classifying the existing computational methods for protein annotation into these three groups, we are able to present a comprehensive framework on protein annotation based on machine learning techniques. In addition to describing recently developed theoretical methodologies, we also cover representative databases and software tools that are widely utilized in the prediction of protein function.

Keywords High-throughput data · Machine learning · Protein function prediction · Semi-supervised learning · Supervised learning · Unsupervised learning

This work was partly supported by the National High Technology Research and Development Program of China (2006AA02Z309), and JSPS-NSFC collaboration project.

X.-M. Zhao · L. Chen · K. Aihara
ERATO Aihara Complexity Modelling Project, JST,
Tokyo 151-0064, Japan

X.-M. Zhao · L. Chen
Institute of Systems Biology, Shanghai University,
200444 Shanghai, China

X.-M. Zhao
Intelligent Computing Lab,
Hefei Institute of Intelligent Machines,
Chinese Academy of Sciences,
230031 Hefei, Anhui, China

X.-M. Zhao · L. Chen · K. Aihara (✉)
Institute of Industrial Science,
The University of Tokyo, Tokyo 153-8505, Japan
e-mail: aihara@sat.t.u-tokyo.ac.jp

L. Chen
Department of Electrical Engineering and Electronics,
Osaka Sangyo University, Osaka 574-8530, Japan

Introduction

As the number of sequenced genomes rapidly increases, the understanding and annotation of these genomes—the post-genomic era—becomes more and more important. To date, about 25% of the genes of the well-studied yeast *Saccharomyces cerevisiae* remain uncharacterized, and only about 20% of the genes of *Homo sapiens* have been annotated. Because it would be time-consuming and expensive to determine the functions of all proteins empirically, many computational methods have been developed to tackle this problem. One example of a straightforward approach is to identify proteins homologous to the target protein; this method is based quite simply on the assumption that proteins with similar sequences carry out similar functions. Given the large numbers of homologous proteins that have been identified, it is possible to annotate many unknown proteins for function-related attributes in accordance with what is known about their homologous counterparts (Cao et al. 2006; Chen et al. 2006a, b, 2007; Chou and Zhang 1995; Diao et al. 2007, 2008; Ding et al. 2007; Du et al. 2003, 2006; Du and Li 2006; Gao and Wang 2006; Guo et al. 2006; Guo et al. 2006a, b; Jahandideh et al. 2007; Kedarisetti et al. 2006; Lin and Li 2007a, b; Mondal et al.

2006; Nanni and Lumini 2008b; Niu et al. 2006; Sun and Huang 2006; Tan et al. 2007; Wen et al. 2007; Xiao et al. 2006b; Zhang and Ding 2007; Zhang SW et al. 2006; Zhou 1998; Zhou and Assa-Munt 2001). However, alignment-based methods may not work well when the sequence similarity between known proteins and the query protein is very low, for example, lower than 30% (Emanuelsson et al. 2007; Chou and Shen 2008). Under such circumstances, the non-sequential model or discrete model (Chou and Shen 2007e) provides a feasible framework for addressing this problem through the application of various techniques (Emanuelsson et al. 2007; Huang et al. 2006; Huang and Li 2004; Jaakkola et al. 2000; Leslie et al. 2004; Liao and Noble 2002; Shi et al. 2007; Vinga and Almeida 2003; Zhao et al. 2005; Chou and Shen 2006a, b, 2007a, b; Shen and Chou 2007b, c, g; Shen et al. 2007). The discrete model has been quite successfully used to predict the subcellular localization of proteins (Cai and Chou 2003; Chen and Li 2007a, b; Chou and Shen 2006a, b, 2007a, b, e; Gao et al. 2005a, b, 2006a; Huang and Li 2004; Li and Li 2008; Mundra et al. 2007; Shen and Chou 2005a, b, 2006, 2007a, b, c, d, e, f, g, h; Shen et al. 2006, 2007; Shi et al. 2007, 2008; Tantoso and Li 2007; Xiao and Chou 2007; Xiao et al. 2005, 2006a; Zhang S et al. 2007; Zhang SW et al. 2007; Zhang T et al. 2006; Zhang ZH 2006; Zhou and Doctor 2003; Zhou et al. 2007a, b), enzyme functional class (Cai and Chou 2005; Cai et al. 2005; Chou 2005; Shen and Chou 2007a), membrane protein type (Cai and Chou 2006; Chou and Cai 2005; Chou and Shen 2007c; Liu et al. 2005a, b; Pu et al. 2007; Wang et al. 2004) and signal peptides (Wang et al. 2006; Chou and Shen 2007d; Liu et al. 2007a, b; Shen and Chou 2007e; Wang et al. 2005). The information thus obtained has provided useful insights into the functions of proteins. Most recently, a user-friendly tool called "Cell-PLoc" (Chou and Shen 2008) has been developed. Cell-Ploc is a package of web-servers that can be used for predicting the subcellular localization of proteins in various organisms. In the Cell-PLoc package, none of the proteins included have a greater than 25% sequence identity to any other proteins in the same subset. This criterium is important for avoiding homology and redundancy bias and is particularly useful for dealing with those proteins that do not have significant sequence homology to any of the character-known proteins. In the discrete model, the pseudo amino acid (PseAA) composition approach (Chou 2001, 2005) has been widely employed to predict various attributes of proteins based on their sequences (Chou and Cai 2005; Cai and Chou 2006; Chen et al. 2006a, b; Du and Li 2006; Gao et al. 2005b; Kurgan et al. 2007; Li and Li 2008; Lin and Li 2007a, b; Mondal et al. 2006; Mundra et al. 2007; Pu et al. 2007; Shen and Chou 2006, 2007f; Shen et al. 2006, 2007; Shi et al. 2007, 2008; Wang et al. 2006; Xiao et al. 2005, 2006a; Zhang SW et al. 2006; Zhang T et al. 2006; Zhou et al. 2007a, b). Structural classes of proteins have also been identified according to their sequences (Chen et al. 2006c; Liu et al. 2007a, b; Zhou 1998; Zhou and Assa-Munt 2001) under the assumption that proteins with similar structures are mostly likely to have similar functions (Bandyopadhyay et al. 2006; Hou et al. 2005).

The recent advances made in high-throughput biotechnologies, such as yeast two-hybrid systems (Chien et al. 1991), protein complex (Gavin et al. 2002; Ho et al. 2002) and microarray expression profiles (Eisen et al. 1998), have resulted in the generation of an extremely large amount of biological data. These data represent rich sources of information for deducing and understanding protein functions. Accordingly, many computational methods have been developed for protein function prediction based on these high-throughput data. For example, starting with the assumption that interacting proteins have similar function, researchers are using protein–protein interaction (PPI) data for protein annotation (Chou and Cai 2006; Fang et al. 2008; Nanni and Lumini 2008a; Pugalenthi et al. 2007; Zhang S et al. 2007; Zhang SW et al. 2007); similarly, gene expression profiles are used with the assumption that genes with similar expression profiles usually carry out similar functions, and so on. Table 1 lists the most popular databases reported in the literature in terms of their use in protein annotation; these include PPI, microarray and functional annotation databases.

In this review, we provide a survey of the main computational methods used for protein function prediction with high-throughput data, especially from the perspective of machine learning. The machine learning methods for protein function prediction are classified into three groups: supervised methods that utilize known annotations to construct a model from training data and predict the functions of unknown proteins; unsupervised methods that group proteins with similar functions together; semi-supervised methods that have only a small number of annotations and a large amount of un-annotated data. Figure 1 is a schematic illustration of the three kinds of methods that are used in protein annotation. Note that the aim of this review is to summarize recent developments on protein annotation; as such, it is by no means comprehensive due to the rapid evolvement of the filed.

Supervised methods

Supervised learning is a machine learning technique that constructs a model from training data, where the training data consist of pairs of input objects and desired outputs. The output of the model can be a continuous value (i.e., regression), or a class label of the input object (i.e.,

Table 1 Popular biological databases for protein annotation

Database	Description
Protein–protein interaction	
HPRD	A (manually) curated database of human protein information http://www.hprd.org
IntAct	Manually curated molecular interaction data from literature http://www.ebi.ac.uk/intact/site/index.jsf
MINT	Experimentally verified protein interactions from the literature http://mint.bio.uniroma2.it/mint/Welcome.do
MPPI	A collection of manually curated high-quality PPI data collected from literature http://mips.gsf.de/proj/ppi
BioGRID	A database of protein and genetic interactions http://www.thebiogrid.org
STRING	A database of known and predicted protein–protein interactions http://string.embl.de
DIP	A collection of experimentally determined protein interactions http://dip.doe-mbi.ucla.edu/dip/Main.cgi
Microarray	
GEO	A gene expression/molecular abundance repository http://www.ncbi.nlm.nih.gov/geo
SMD	Stores raw and normalized data from microarray experiments, and provides data retrieval, analysis and visualization http://genome-www5.stanford.edu
ArrayExpress	A public repository for transcriptomics and related data http://www.ebi.ac.uk/arrayexpress
caArray	A MIAME compliant data repository http://caarraydb.nci.nih.gov/caarray
Annotation	
GO	A controlled vocabulary to describe gene and gene product attributes http://www.geneontology.org/
COG	Clusters of Orthologous Groups (COG) http://www.ncbi.nlm.nih.gov/COG/index.html
Enzyme	Enzyme Commission http://www.chem.qmul.ac.uk/iubmb/enzyme
EGAD	Expressed Gene Anatomy Database http://www.tigr.org/tdb/egad/egad.shtml
EGP	*E. coli* Genome Proteome http://genprotec.mbl.edu/
HAMAP	High-quality automated and manual annotation of microbial Proteomes http://www.expasy.org/sprot/hamap/
Uniprot	The most comprehensive catalog of information on proteins http://www.ebi.uniprot.org/index.shtml
Funcat	An annotation scheme for the functional description of proteins from prokaryotes, unicellular eukaryotes, plants and animals http://mips.gsf.de/projects/funcat
TIGRFAMs	Protein families based on Hidden Markov Models http://www.tigr.org/TIGRFAMs/

classification). In the supervised learning methods, the prediction of protein function is usually regarded as a classification problem, where a set of features are collected for each protein, and the learning algorithm is utilized to infer the association rule between the features and the biological functions. The methods described below differ in terms of the data and the learning algorithms that are used for protein annotation. Table 2 lists the most popular supervised methods that have been proposed for protein annotation.

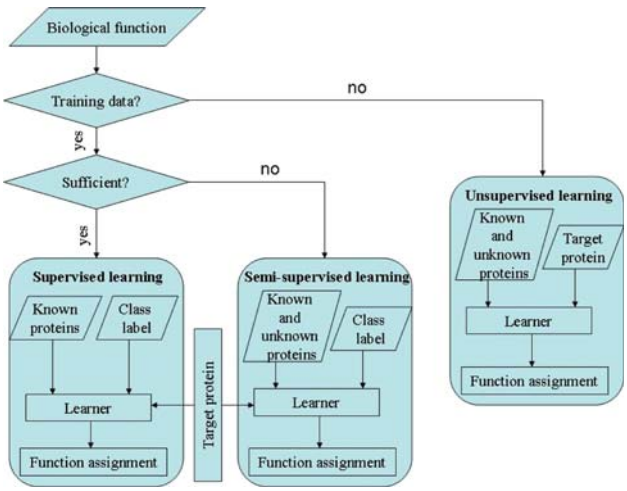

Fig. 1 Schematic illustration of the machine learning methods in protein annotation

Gene expression data

Gene expression data is generally organized as a matrix, where each gene is denoted by a vector (row) and each condition by an attribute (column). The functions of a number of the genes are usually known in advance, and these act as the class labels for the corresponding vectors representing genes. The remaining genes remain unlabeled, and the learning algorithm is utilized to assign a label to them.

Brown et al. (2000) applied support vector machines (SVMs) to predict the functions of yeast genes based on gene expression data. Subsequent numerical experiments demonstrated the promising effectiveness of SVMs in protein function prediction. Mateos et al. (2002) utilized multilayer perceptrons (MLP) (Huang 1996) to predict protein functions of the yeast genome into 96 function

Table 2 A summary of supervised learning methods for protein annotation

Supervised learning methods	References	Software (if available)
Gene expression profiles		
Support vector machines	Brown et al. (2000)	Genex http://www.soe.ucsc.edu/research/compbio/genex/
	Ng and Tan (2003)	
Multilayer perceptions	Mateos et al. (2002)	
k-nearest neighbor	Pandey and Kumar (2007)	
Protein–protein interaction		
Markov random field model	Deng et al. (2003)	
Markov random field model	Letovsky and Kasif (2003)	Netmark http://genomics10.bu.edu/netmark/
Kernel logistic regression	Lee et al. (2006)	KLR http://msms.usc.edu/hyunjul/klr/klr.html
Data integration		
Bayesian network	Troyanskaya et al. (2003)	
	Chen and Xu (2004)	Genefas http://digbio.missouri.edu/genefas/
Support vector machines	Lanckriet et al. (2004)	
	Barutcuoglu et al. (2006)	
	Zhao et al. (2007)	
	Tsuda et al. (2005)	

classes. By analyzing the performance of the MLP classifier, the authors showed that the performance of the classifier is affected not only by the learning technique but also by the nature of the data. Ng and Tan (2003) subsequently combined multiple datasets for learning with SVMs and presented a strategy for selecting the most informative datasets for learning individual classes. More recently, Pandey and Kumar (2007) presented a modified *k*-nearest neighbor learning algorithm for protein annotation based on gene expression data, where the similarity between functional classes is taken into account. The results demonstrate that the incorporation of inter-relationships between functional classes substantially improves the performance of function prediction algorithms. It is worth mentioning that the inter-relationship among functional classes has also been taken into account recently by several researchers (Barutcuoglu et al. 2006; Lee et al. 2006), who also report that the biological functions are not independent of each other.

Protein–protein interaction

The rational behind all of the computation methods based on PPIs is that proteins close to each other in the PPI network have a high probability of possessing similar functions. Deng et al. (2003) developed a Markov random field (MRF) model for protein function prediction based on PPIs. In the MRF model, each protein *Pi* is assigned a variable X_I, where $X_I = 1$ if the protein has that function and $X_I = 0$ otherwise. The conditional probability on the functional labeling is proportional to $\exp(-U(x))$, where x_I is the value of X_I, and

$$U(x) = -\alpha N_1 - \beta_1 N_{10} - \gamma_1 N_{11} - \kappa N_{00} \quad (1)$$

where $\alpha = \log\left(\frac{\pi}{1-\pi}\right)$ and π is the prior probability of a protein having the function, and $N_{ll'}$ is the number of (l,l') interacting pairs in the PPI network. The parameters α and β_1 of the model are estimated using a pseudo-likelihood method based on the proteins whose function is known. The probability distribution of the function of unknown proteins is then estimated by Gibbs sampling. Letovsky and Kasif (2003) also proposed a MRF model for protein annotation where it is assumed that the number of neighbors of a protein that are annotated with a given function is binomially distributed and the distribution's parameter depends on whether the protein has that function. The loopy belief propagation (Murphy et al. 1999) is employed to perform inference in their model.

In the methods described above, a protein's functions are predicted based on PPI data and the functional annotations of its interaction partners, where only direct interactions are considered and the functions are considered to be independent. Lee et al. (2006) recently developed a new kernel logistic regression (KLR) method for protein function prediction based on diffusion kernels. In the KLR method, the authors incorporated the correlation among biological functions into their model by identifying a set of functions that are highly correlated with the function of interest using the χ^2 test. The prediction accuracy is comparable to another protein function classifier based on the SVMs with a diffusion kernel.

Integration of different kinds of data sources

Although various kinds of high-throughput data can provide important information on protein functions, many high-throughput data are notorious for the noise in the data and the specificity for scale. The integration of different types of biological data for utilization in protein function prediction is becoming a popular trend and is expected to improve prediction accuracy. There are many ways to combine different kinds of data sources for protein annotation, and these approaches can be classified into two groups: kernel methods and Bayesian network.

Troyanskaya et al. (2003) developed a MAGIC (multisource association of genes by integration of clusters) system in which the Bayesian network is employed to integrate different types of high-throughput biological data. The inputs to the system consist of gene–gene relationship matrices established on different high-throughput data. The numerical results demonstrate that MAGIC improves prediction accuracy compared with microarray analysis alone.

Chen and Xu (2004) also developed a Bayesian model to integrate different kinds of data sources, including PPI, microarray data, and protein complex data. In their methods, two prediction models are presented: local prediction and global prediction. In the local prediction model, the probability of protein x having function F is defined as:

$$G(F,x) = 1 - \prod_{i}^{m}(1 - P(S|D_I)) \tag{2}$$

where m is the total number of high-throughput data sources, and $P(S|D_I)$ is the probability that two genes have the same function given data D_I, which is defined as:

$$P(S|D_I) = 1 - \prod_{j=1}^{n}(1 - P_j(S|D_I)) \tag{3}$$

where n is the total number of interaction partners given D_I, and $P_j(S|D_I)$ is the probability that interacting pair j have the same function. In the local prediction method, only immediate interaction partners are considered, which may lead to local optimal solutions. Therefore, a global prediction model is presented by utilizing the Boltzmann machine to characterize the global stochastic behavior of the network. The authors show that the global model outperforms other existing methods.

Lanchriet et al. (2004) developed a kernel method for data fusion. They first constructed a kernel matrix K_I for each data source I and then combined all the kernel matrices in a linear form:

$$K = \sum_{I=1}^{m}\mu_I K_I \tag{4}$$

where m is the total number of kernel matrices. The coefficients μ_I are estimated using a semidefinite program (SDP). The authors show that the kernel fusion method outperforms the MRF method (Deng et al. 2003). Zhao et al. (2007) recently constructed a functional linkage graph from different types of data with the shortest path distance as the similarity measure and then employed SVMs to annotate proteins. Tsuda et al. (2005) proposed a kernel method by combining multiple protein networks, where the combination weights are obtained by convex optimization.

Barutcuoglu et al. (2006) recently provided another approach for integrating different types of data for protein function prediction. They combined different types of data into one vector by concatenating all feature vectors for one gene. Using all available data, they trained a SVMs classifier for each functional class. They also constructed a Bayesian net to combine the outputs of the classifiers taking the functional taxonomy into consideration.

Semi-supervised methods

In general, only a small number of proteins have actually been annotated for a certain function. Therefore, it is difficult to obtain sufficient training data for the supervised learning algorithms. Under such circumstances, semi-supervised learning methods provide an alternative approach to protein annotation. Table 3 lists the most popular semi-supervised methods for protein annotation.

Neighborhood

In semi-supervised methods, the most straightforward method for annotating an unknown protein is to determine its function by investigating the functions of its immediate partners in the PPI network. Schwikowski et al. (2000) annotated the unknown protein with the most commonly occurring functions of its interaction partners; in the literature, this approach is called the Majority rule method. The main problem of this approach is that only immediate partners are considered, and the whole topology of the network is not taken into account. To handle this problem, Hishigaki et al. (2001) defined the neighborhood of a protein with a radius of n. For an unknown protein, the functional enrichment in its n-neighborhood is investigated with χ^2 test, and the top ranking functions are assigned to the unknown proteins. This approach alleviates the constraints of the Majority rule method to some extent.

In another approach, not only the neighborhood of the protein of interest is considered but also the shared neighborhood of a pair of proteins. Chua et al. (2006) defined the functional similarity between a pair of proteins by taking both the direct and indirect neighbors of the protein pair into account. They also proposed a new method for integrating different kinds of data for protein annotation and showed

Table 3 A summary of semi-supervised learning methods for protein function prediction

Semi-supervised learning methods	References	Software (if available)
Neighborhood		
Majority rule	Schwikowski et al. (2000)	
	Chua et al. (2006)	
χ^2 test	Hishigaki et al. (2001)	
Probabilistic suffix tree	Kirac et al. (2006)	
Association analysis	Pandey et al. (2007)	
Global optimization		
Simulated annealing	Vazquez et al. (2003)	
Hopfield network	Karaoz et al. (2004)	
Integer linear program	Nabieva et al. (2005)	
Discriminative method		
Support vector machines	Zhao et al. (2008a)	
Random forest	Huang et al. (2007)	ContextAnnotation http://zhoulab.usc.edu/ContextAnnotation

promising results (Chua et al. 2007). Kirac et al. (2006) recently presented a model that considers the annotations in the paths leading to the target protein in the PPI network. This model is implemented using the probabilistic suffix tree data structure, and the results have been better than other neighborhood-based methods. To overcome the incompleteness and false positives of existing PPI networks, Pandey et al. (2007) presented a new method to transform the original interaction network into a new network with the spurious edges removed and biologically valid ones added. This reconstructed protein interaction network signicantly improved the performance of standard function prediction algorithms.

Global optimization

Despite the simplicity and effectiveness of the neighborhood methods, they do not take the topology of the whole interaction network into account. Furthermore, the neighborhood methods will not work if there are no annotations for all of the neighbors of the target protein.

Vazquez et al. (2003) proposed a new global method to annotate a protein which takes the topology of the whole network into consideration. A function is assigned to an unknown protein so that the number of the same annotations associating with its neighbors is maximized by minimizing the score function:

$$E = -\sum_{i,j} J_{ij}\delta(\sigma_I, \sigma_j) - \sum_i h_I(\sigma_I) \quad (5)$$

where J_{ij} is the element of the adjacency matrix of the interaction network, $\delta(i,j)$ is the discrete δ function, and $h_I(\sigma_I)$ is the number of partners of protein I annotated with function σ_I. The simulated annealing is employed to minimize the score function above. Karaoz et al. (2004) subsequently developed a similar method by assigning a state $S_u \in \{0,1\}$ to an unknown protein u so as to maximize the score function $\sum_{(u,v)\in E} S_u S_v$, where u and v are nodes in the PPI network, and $(u,v) \in E$ signifies that there is an interaction between protein u and protein v. The optimization problem is handled by employing a discrete Hopfield network, where only one function is considered each time.

Another method recently proposed by Nabieva et al. (2005) also formulates the annotation problem as global optimization problem, where a unique function is assigned to an unknown protein so as to minimize the cost of edges connecting proteins with different assignments. More specifically, these authors formulated the optimization problem as an integer linear program (ILP) model. In the ILP model, each protein annotated with the target function in the PPI network is regarded as the source of functional flow. By simulating the spread of this functional flow through the network, each unknown protein obtains a score for having the function based on the amount of flow it received.

Discriminative method

The semi-supervised methods described above are actually generative models that construct a model with only positive samples, i.e., annotated proteins. In general, those proteins outside of the target functional class are seen as negative samples, whereas the proteins in the functional class are positive samples. However, this setting may be not valid because each protein is usually annotated with multiple functions. Although some proteins are not yet annotated with the target function, they still may actually have that function. Furthermore, the imbalanced problem will arise if all the proteins outside of the functional class are seen as negative samples, which will degrade the performance of the classifier (Zhao et al. 2008b).

Zhao et al. (2008a) proposed a new algorithm to define the negative samples in protein function prediction. In detail, the one-class SVMs and two-class SVMs are used as the core learning algorithm in order to identify the representative negative samples so that the positive samples hidden in the unlabeled data can be recovered. The experiments demonstrate that with the negative samples generated, the performance of prediction methods is improved compared with other methods defining negative samples (Carter et al. 2001).

Huang et al. (2007) recently described a data mining technique with the aim of identifying the frequent itemsets in the gene network by integrating 65 human microarray datasets. The authors then utilized the random forest classifier to predict the functions of proteins by considering the network topology score, recurrence, density, size, average node degree, percentage of unknown genes, and functional enrichment of network modules.

Unsupervised methods

The supervised and semi-supervised learning methods annotate proteins utilizing prior information on the annotated proteins. However, in some cases, no such information is available. The unsupervised learning methods provide a means to approach the problem in this case. Unsupervised learning refers to learning algorithms that do not utilize any prior information about the class label. Clustering is a popular unsupervised learning algorithm that has been widely used in the prediction of protein function. In clustering, the first step is to detect the module (or complex) in which the unknown protein is located; the second step is to transfer the functions—to a greater of lesser extent—of known proteins in the same module to the unknown protein. The unsupervised methods described herein differ in the data used and the definition of the cluster. Table 4 lists the most popular unsupervised methods that have been used for protein annotation.

Clustering based on network topology

Bader and Hogue (2003) proposed the molecular complex detection algorithm (MCODE) to predict complexes in the PPI network. In MCODE, the vertices of the PPI network are weighted first, and the tense interconnected modules are then detected. Spirin and Mirny (2003) presented two algorithms for extracting the complexes and functional modules in the PPI network: the first is based on the superparamagnetic clustering (SPC) (Blatt et al. 1996), and

Table 4 A summary of unsupervised learning methods for protein function prediction

Unsupervised learning methods	References	Software (if available)
Clustering based on network topology		
MCODE	Bader and Hogue (2003)	MCODE http://cbio.mskcc.org/bader/software/mcode/
Superparamagnetic clustering	Spirin and Mirny (2003)	
Markov clustering	Pereira-Leal et al. (2004)	MCL http://micans.org/mcl/
	Krogan et al. (2006)	
HCS	Przulj et al. (2004)	
RNSC	King et al. (2004)	
Edge-betweenness clustering	Dunn et al. (2005)	
Similarity-based clustering		
k-means like clustering	Samanta and Liang (2003)	
Hierarchical clustering	Rives and Galitski (2003)	
	Arnau et al. (2005)	
The shortest path	Zhou et al. (2002)	
Hierarchical clustering	Brun et al. (2003)	PRODISTIN http://crfb.univ-mrs.fr/webdistin/
Data integration		
Hierarchical clustering	Hanisch et al. (2002)	
Probabilistic graph model	Segal et al. (2003)	
Superparamagnetic clustering	Tornow and Mewes (2003)	
Biclustering	Tanay et al. (2004)	SAMBA http://acgt.cs.tau.ac.il/samba/
Gene annotation using integrated networks	Massjouni et al. (2006)	Virgo http://whipple.cs.vt.edu:8080/virgo
Hidden modular random field model	Shiga et al. (2007)	

MCODE, Molecular complex detection algorithm; HCS, highly connected subgraphs algorithm; RNSC, restricted neighborhood search clustering algorithm

the second is a Monte Carlo algorithm that aims to maximize the density of the predicted clusters. The SPC algorithm has been shown to be able to detect the sparsely connected functional modules, such as the MAPK signaling cascade. Pereira-Leal et al. (2004) applied the Markov clustering (MCL) (Enright et al. 2002) algorithm to predict complexes in the PPI network. These authors showed that the pathways can be inferred from the hierarchical network of modular interactions. Krogan et al. (2006) subsequently also applied the MCL algorithm to find complexes in the PPI network, where the strength of PPI is estimated in advance.

Przulj et al. (2004) presented the highly connected subgraphs (HCS) algorithm (Hartuv and Shamir 2000) for detecting complexes in the PPI network. King et al. (2004) proposed the restricted neighborhood search clustering (RNSC) algorithm for partitioning the PPI network into clusters based on a cost function that is used to evaluate the partitioning. These authors showed that the RNSC algorithm outperforms the MCODE algorithm. Dunn et al. (2005) subsequently utilized the edge-betweenness clustering to predict complexes in PPI network.

The algorithms described above predict complexes in the PPI network based on the topology structure of the network. A recent study by Brohee and van Helden (2006) compared four different clustering methods—RNSC, SPC, MCODE, and MCL—for their usefulness in predicting complexes in the PPI network. Using a set of known complexes in the PPI network, these authors compared the performance of the four methods in detecting complexes in perturbed PPI networks. Their results indicated that the MCL algorithm is more robust than the other three algorithms.

Similarity-based clustering

Instead of using the topology of the molecular network to predict the complexes, several authors have defined the similarity between a pair of proteins and applied the standard clustering algorithms to predict complexes. In this approach, the similarity between a pair of proteins can be defined utilizing their direct neighbors or neighborhood.

Samanta and Liang (2003) defined the similarity between a pair of proteins (*A*,*B*) with a *P* value, which is defined as:

$$P(N, N_A, N_B, m) = \frac{\binom{N}{m}\binom{N-m}{N_A-m}\binom{N-N_B}{N_B-m}}{\binom{N}{N_A}\binom{N}{N_B}} \qquad (6)$$

where N is the total number of proteins in the PPI network, N_A and N_B are respectively the number of interaction partners of A and B, and m is the number of proteins in common between N_A and N_B. After obtaining the similarity metric, the PPI network is partitioned into different groups, where each group is seen as a complex.

Rives and Galitski (2003) investigated the organization of complexes in the PPI network with the shortest path distance as the similarity between a pair of proteins. Similarly, Zhou et al. (2002) defined the similarity measure as the shortest path distance between a pair of genes based on gene expression data. Arnau et al. (2005) subsequently used the shortest path length among proteins as a similarity measure for hierarchical clustering. Brun et al. (2003) defined the similarity measure between a protein pair with Czekanovski–Dice distance and utilized the hierarchical clustering to predict complexes.

Integration of various kinds of data sources

The clustering methods described above predict protein functions with single data source, i.e., PPI or gene expression data. Despite the usefulness of PPI and gene expression profiles, the integration of different data sources may improve the prediction accuracy and reliability.

Hanisch et al. (2002) constructed a distance function by combining information from gene expression data and biological networks. Based on the distance function, it is then possible to perform a joint clustering of genes and vertices of the network. Segal et al. (2003) described a probabilistic model that is derived from the data using the expectation maximization (EM) algorithm to detect pathways based on gene expression and protein interaction data. Tornow and Mewes (2003) calculated the correlation strength of a group of genes by considering the probability of these genes belonging to the same module in a different network. The rational behind the method is that in any given group of genes, there is a high probability that those with a significant correlation strength in different networks carry out the same function.

Tanay et al. (2004) presented an integrative framework SAMBA (statistical-algorithmic method for bicluster analysis) for various kinds of data, including protein interaction, gene expression, phenotypic sensitivity, and transcription factor (TF) binding. These authors proved the effectiveness of SAMBA by predicting the functions of more than 800 uncharacterized genes. Massjouni et al. (2006) subsequently constructed a functional linkage network (FLN) from gene expression and molecular interaction data and propagated the functional labels across the FLN in order to precisely predict the functions of unlabeled genes. More recently, Shiga et al. (2007) proposed a hidden modular random field model for protein annotation by combining gene expression data and gene network.

Conclusions and perspectives

The prediction of protein function by computational means has become an active field in computational and systems biology in recent years. Despite the large number of computational methods that have been developed, however, it is still a difficult task to annotate proteins with precision using high-throughput data. In this section, we briefly address the obstacles and possible solutions for protein annotation with respect to computational methods and data, respectively.

A large number of machine learning algorithms are currently available; however, only a small number of these have been applied to protein function prediction. Among the supervised learning algorithms, SVMs are particularly popular due to their good performance and strong statistical background. However, other popular supervised learning algorithms, such as the Radial basis function neural network (Huang 1999) and decision tree (Quinlan 1993), have not yet been applied to this field. Among the unsupervised learning algorithms, hierarchical clustering is preferred by computational biologists due to its simplicity and good performance. One of the possible reasons for the absence of other learning algorithms in protein annotation is the scarcity of corresponding software, whereas SVMs have many free versions of software available online. Therefore, it is necessary for scientists in the machine learning field to make their algorithms freely accessible to biologists.

Although there is a large number of computation methods for protein annotation, it is difficult for newcomers to the field to determine which method should be used to predict the functions of unknown proteins. Therefore, a systematic comparison of different methods is needed. A comparison of prediction methods with the gold standard dataset, such as gene ontology (GO) (Ashburner et al. 2000) and MIPS Funcat (Ruepp et al. 2004), may provide insights or advice to the newcomer that will facilitate the choice of methods to be used. For example, it has been shown (Sharan et al. 2007) that a semi-supervised method, i.e., Majority rule (Schwikowski et al. 2000), generally performs better than an unsupervised method, i.e., MCODE (Bader and Hogue 2003). However, choosing an appropriate method involves many factors, and the choice should be made with care. In general, the supervised methods perform best if there are sufficient training data available. Otherwise, it is better to try the semi-supervised methods first if there is at least a small number of proteins that have been annotated for a certain function. The unsupervised methods are generally regarded as the final choice. On the other hand, different prediction methods have been proposed for different data and functional schemas. For example, some methods perform well in one case but badly in another. One possible solution is to construct meta-servers by combining a set of best-performing prediction methods.

Given the abundance of different methods proposed in literatures, a comprehensive comparison of these different methods is necessary. In statistical prediction, the independent dataset test, sub-sampling test, and jackknife (or leave-one-out) test are the three cross-validation tests widely employed to examine the accuracy of a predictor (Chou and Zhang 1995). For the independent dataset test, as pointed out in Chou and Shen (2008), although none of the proteins to be tested occurs in the training dataset, the selection of proteins for the testing dataset could be quite arbitrary unless the latter is sufficiently large. This kind of arbitrariness may directly affect the final conclusion. In the sub-sampling test, the training data are classified into several subsets, where one subset is used as testing set while the remainder make up the training set. The problem with cross-validation (such as fivefold and tenfold cross-validation) is that the number of possible selections in dividing a dataset is an astronomical figure even for a very simple dataset (see Eq. 50 of Chou and Shen 2007e). Therefore, any practical result by the cross-validation test alone represents only one of many possible results and, consequently, cannot mitigate the arbitrariness. In contrast, the jackknife test can always yield a unique result for a given benchmark dataset. Therefore, the jackknife cross-validation is the most objective and is being widely applied by more and more investigators to test the quality of various predictors.

At the present time, most of the existing computation methods consider only one function each time. In other words, the biological functions are treated independently. As recently pointed out by several researchers (Barutcuoglu et al. 2006; Lee et al. 2006; Pandey and Kumar 2007), biological functions are not independent of each other, and the incorporation of correlations among functions can significantly improve the performance of the prediction methods. Therefore, the correlation among functions should be taken into account by methods developed in the future.

As a general rule, most of the prediction methods described here focus on single data sources, for example, PPI or gene expression data. However, PPI data are notorious for false positives and incompleteness. One possible solution is to reconstruct the PPI network by removing spurious edges and adding biologically valid ones, as described in Pandey et al. (2007). On the other hand, the correlation coefficients among genes are generally used as the similarity measures for proteins pairs, where genes having high correlations are assumed to have similar functions. The problem with this assumption is that the correlation coefficients cannot capture the functional linkages among proteins in some cases. For example, the

correlations among all protein are very high or very low. One possible solution is to utilize the higher order statistics to capture the functional linkages, as suggested in Zhou et al. (2005).

Although the integration of different kinds of data has been used for protein annotation and shown promising results, the diverse kinds of data sources should be combined carefully. Which kinds of data should be combined? Does integration of all kinds of data really improve prediction accuracy? As mentioned in Myers and Troyanskaya (2007), each dataset has its own specific structure. In some cases, combining different kinds of data adds noise to existing data and degrades the performance of the prediction method. Therefore, the nature of the data should be taken into account while developing new prediction methods. That said, new effective data fusion methods are needed in the future for protein annotation.

Despite the limitations of the existing computational methods and the noise lying in the high-throughput data, protein annotation based on high-throughput data is a promising and active research field. With the rapid advances in biotechnology, more biological data of high quality and reliability will be expected. Accordingly, the prediction accuracy of function prediction methods will be improved. Although many computational methods have been developed, there is still much room for improvement. The scientists in machine learning field are expected to develop more accurate prediction methods and make them easily accessible to biologists.

Acknowledgments This work was partly supported by the National High Technology Research and Development Program of China (2006AA02Z309), and JSPS-NSFC collaboration project.

References

Arnau V, Mars S, Marn I (2005) Iterative cluster analysis of protein interaction data. Bioinformatics 21:364–378

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 25:25–29

Bader G, Hogue C (2003) An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4:2

Bandyopadhyay D, Huan J, Liu J, Prins J, Snoeyink J, Wang W, Tropsha A (2006) Structure-based function inference using protein family-specific fingerprints. Protein Sci 15:1537–1543

Barutcuoglu Z, Schapire RE, Troyanskaya OG (2006) Hierarchical multi-label prediction of gene function. Bioinformatics 22:830–836

Blatt M, Wiseman S, Domany E (1996) Superparamagnetic clustering of data. Phys Rev Lett 76:3251–3254

Brohee S, van Helden J (2006) Evaluation of clustering algorithms for protein–protein interaction networks. BMC Bioinformatics 7:488

Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Manuel AJ, Haussler D (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci USA 97:262–267

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

Cai YD, Chou KC (2003) Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition. Biochem Biophys Res Comm 305:407–411

Cai YD, Chou KC (2005) Predicting enzyme subclass by functional domain composition and pseudo amino acid composition. J Proteome Res 4:967–971

Cai YD, Chou KC (2006) Predicting membrane protein type by functional domain composition and pseudo amino acid composition. J Theor Biol 238:395–400

Cai YD, Zhou GP, Chou KC (2005) Predicting enzyme family classes by hybridizing gene product composition and pseudo-amino acid composition. J Theor Biol 234:145–149

Cao Y, Liu S, Zhang L, Qin J, Wang J, Tang K (2006) Prediction of protein structural class with Rough Sets. BMC Bioinformatics 7:20

Carter RJ, Dubchak I, Holbrook SR (2001) A computational approach to identify genes for functional RNAs in genomic sequences. Nucleic Acids Res 29:3928–3938

Chen YL, Li QZ (2007a) Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition. J Theor Biol 248:377–381

Chen YL, Li QZ (2007b) Prediction of the subcellular location of apoptosis proteins. J Theor Biol 245:775–783

Chen C, Tian YX, Zou XY, Cai PX, Mo JY (2006a) Using pseudo-amino acid composition and support vector machine to predict protein structural class. J Theor Biol 243:444–448

Chen C, Zhou X, Tian Y, Zou X, Cai P (2006b) Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. Anal Biochem 357:116–121

Chen J, Liu H, Yang J, Chou KC (2007) Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids 33:423–428

Chen L, Wu LY, Wang Y, Zhang S, Zhang XS (2006c) Revealing divergent evolution, identifying circular permutations and detecting active-sites by protein structure comparison. BMC Struct Biol 6:18

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

Chien C, Bartel P, Sternglanz R, Fields S (1991) The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. Proc Natl Acad Sci USA 88:9578–9582

Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. Proteins 43:246–255 (Erratum: ibid., 2001, vol 44, 60)

Chou KC (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21:10–19

Chou KC, Cai YD (2005) Prediction of membrane protein types by incorporating amphipathic effects. J Chem Inform Model 45:407–413

Chou KC, Cai YD (2006) Predicting protein–protein interactions from sequences in a hybridization space. J Proteome Res 5:316–322

Chou KC, Shen HB (2006a) Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochem Biophys Res Commun 347:150–157

Chou KC, Shen HB (2006b) Large-scale predictions of Gram-negative bacterial protein subcellular locations. J Proteome Res 5:3420–3428

Chou KC, Shen HB (2007a) Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. J Proteome Res 6:1728–1734

Chou KC, Shen HB (2007b) Large-scale plant protein subcellular location prediction. J Cell Biochem 100:665–678

Chou KC, Shen HB (2007c) MemType-2L: a Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem Biophys Res Comm 360:339–345

Chou KC, Shen HB (2007d) Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochem Biophys Res Comm 357:633–640

Chou KC, Shen HB (2007e) Review: recent progresses in protein subcellular location prediction. Anal Biochem 370:1–16

Chou KC, Shen HB (2008) Cell-PLoc: a package of web-servers for predicting subcellular localization of proteins in various organisms. Nat Protoc 3:153–162

Chou KC, Zhang CT (1995) Review: prediction of protein structural classes. Crit Rev Biochem Mol Biol 30:275–349

Chua HN, Sung WK, Wong L (2006) Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. Bioinformatics 22:1623–1630

Chua HN, Sung WK, Wong L (2007) An efficient strategy for extensive integration of diverse biological data for protein function prediction. Bioinformatics 23:3364–3373

Deng M, Zhang K, Mehta S, Chen T, Sun F (2003) Prediction of protein function using protein–protein interaction data. J Comput Biol 10:947–960

Diao Y, Li M, Feng Z, Yin J, Pan Y (2007) The community structure of human cellular signaling network. J Theor Biol 247:608–615

Diao Y, Ma D, Wen Z, Yin J, Xiang J, Li M (2008) Using pseudo amino acid composition to predict transmembrane regions in protein: cellular automata and Lempel-Ziv complexity. Amino Acids 34:111–117

Ding YS, Zhang TL, Chou KC (2007) Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. Protein Pept Lett 14:811–815

Du P, Li Y (2006) Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. BMC Bioinformatics 7:518

Du QS, Wei DQ, Chou KC (2003) Correlation of amino acids in proteins. Peptides 24:1863–1869

Du QS, Jiang ZQ, He WZ, Li DP, Chou KC (2006) Amino acid principal component analysis (AAPCA) and its applications in protein structural class prediction. J Biomol Struct Dyn 23:635–640

Dunn R, Dudbridge F, Sanderson C (2005) The use of edge-betweenness clustering to investigate biological function in protein interaction networks. BMC Bioinformatics 6:39

Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95:14863–14868

Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using targetp, signalp and related tools. Nat Protoc 2:953–971

Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30:1575–1584

Fang Y, Guo Y, Feng Y, Li M (2008) Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features. Amino Acids 34:103–109

Gao QB, Wang ZZ (2006) Classification of G-protein coupled receptors at four levels. Protein Eng Des Sel 19:511–516

Gao QB, Wang ZZ, Yan C, Du YH (2005a) Prediction of protein subcellular location using a combined feature of sequence. FEBS Lett 579:3444–3448

Gao Y, Shao SH, Xiao X, Ding YS, Huang YS, Huang ZD, Chou KC (2005b) Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, Bessel function, and Chebyshev filter. Amino Acids 28:373–376

Gavin AC, BÄosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, HÄofert C, Schelder M, Brajenovic M, Rufiner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415:141–147

Guo J, Lin Y, Liu X (2006) GNBSL: a new integrative system to predict the subcellular location for Gram-negative bacteria proteins. Proteomics 6:5099–5105

Guo YZ, Li M, Lu M, Wen Z, Wang K, Li G, Wu J (2006a) Classifying G protein-coupled receptors and nuclear receptors based on protein power spectrum from fast Fourier transform. Amino Acids 30:397–402

Guo YZ, Li M, Lu M, Wen Z, Wang K, Li G, Wu J (2006b) Classifying G protein-coupled receptors and nuclear receptors based on protein power spectrum from fast Fourier transform. Amino Acids 30:397–402

Hanisch D, Zien A, Zimmer R, Lengauer T (2002) Co-clustering of biological networks and gene expression data. Bioinformatics 18:S145–S154

Hartuv E, Shamir R (2000) A clustering algorithm based on graph connectivity. Inform Process Lett 76:175–181

Hishigaki H, Nakai K, Ono T, Tanigami A, Takagi T (2001) Assessment of prediction accuracy of protein function from protein–protein interaction data. Yeast 18:523–531

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorfi S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, S¿rensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M (2002) Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. Nature 415:180–183

Hou J, Jun SR, Zhang C, Kim SH (2005) From The Cover: global mapping of the protein structure space and application in structure-based inference of protein function. Proc Natl Acad Sci USA 102:3651–3656

Huang D (1996) Systematic theory of neural networks for pattern recognition. Publishing House of Electronic Industry of China, Beijing

Huang D (1999) Radial basis probabilistic neural networks: model and application. Int J Pattern Recognit Artif Intell 13:1083–1101

Huang Y, Li Y (2004) Prediction of protein subcellular locations using fuzzy k-nn method. Bioinformatics 20:21–28

Huang DS, Zhao XM, Huang GB, Cheung YM (2006) Classifying protein sequences using hydropathy blocks. Pattern Recogn 39:2293–2300

Huang Y, Li H, Hu H, Yan X, Waterman MS, Huang H, Zhou XJ (2007) Systematic discovery of functional modules and context-

specific functional annotation of human genome. Bioinformatics 23:i222–i229
Jaakkola T, Diekhans M, Haussler D (2000) A discriminative framework for detecting remote protein homologies. J Comput Biol 7:95–114
Jahandideh S, Abdolmaleki P, Jahandideh M, Asadabadi EB (2007) Novel two-stage hybrid neural discriminant model for predicting proteins structural classes. Biophys Chem 128:87–93
Karaoz U, Murali TM, Letovsky S, Zheng Y, Ding C, Cantor CR, Kasif S (2004) Whole-genome annotation by using evidence integration in functional-linkage networks. Proc Natl Acad Sci USA 101:2888–2893
Kedarisetti KD, Kurgan LA, Dick S (2006) Classifier ensembles for protein structural class prediction with varying homology. Biochem Biophys Res Commun 348:981–988
King AD, Przulj N, Jurisica I (2004) Protein complex prediction via cost-based clustering. Bioinformatics 20:3013–3020
Kirac M, Ozsoyoglu G, Yang J (2006) Annotating proteins by mining protein interaction networks. Bioinformatics 22:e260–e270
Kurgan LA, Stach W, Ruan J (2007) Novel scales based on hydrophobicity indices for secondary protein structure. J Theor Biol 248:354–366
Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrín-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MH, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. Nature 440:637–643
Lanckriet GR, Deng M, Cristianini N, Jordan MI, Noble WS (2004) Kernel-based data fusion and its application to protein function prediction in yeast. In: Pac Symp Biocomput. Division of Electrical Engineering. University of California, Berkeley, pp 300–311
Lee H, Tu Z, Deng M, Sun F, Chen T (2006) Diffusion kernel-based logistic regression models for protein function prediction. OMICS: J Integr Biol 10:40–55
Leslie CS, Eskin E, Cohen A,Weston J, Noble WS (2004) Mismatch string kernels for discriminative protein classification. Bioinformatics 20:467–476
Letovsky S, Kasif S (2003) Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics 19:i197–i204
Li FM, Li QZ (2008) Using pseudo amino acid composition to predict protein subnuclear location with improved hybrid approach. Amino Acids 34:119–125
Liao L, Noble WS (2002) Combining pairwise sequence similarity and support vector machines for remote protein homology detection. In: RECOMB '02: Proc 6th Annu Int Conf Comput Biol. ACM, New York, pp 225–232
Lin H, Li QZ (2007a) Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant. Biochem Biophys Res Commun 354:548–551
Lin H, Li QZ (2007b) Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 Dipeptide components. J Comput Chem 28:1463–1466
Liu DQ, Liu H, Shen HB, Yang J, Chou KC (2007a) Predicting secretory protein signal sequence cleavage sites by fusing the marks of global alignments. Amino Acids 32:493–496
Liu H, Wang M, Chou KC (2005a) Low-frequency Fourier spectrum for predicting membrane protein types. Biochem Biophys Res Commun 336:737–739
Liu H, Yang J, Wang M, Xue L, Chou KC (2005b) Using Fourier spectrum analysis and pseudo amino acid composition for prediction of membrane protein types. Protein J 24:385–389
Liu Z,Wu LY,Wang Y, Zhang XS, Chen L (2007b) Predicting gene ontology functions from protein's regional surface structures. BMC Bioinformatics 8:475
Massjouni N, Rivera CG, Murali TM (2006) VIRGO: computational prediction of gene functions. Nucleic Acids Res 34:W340–W344
Mateos A, Dopazo J, Jansen R, Tu Y, Gerstein M, Stolovitzky G (2002) Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons. Genome Res 12:1703–1715
Mondal S, Bhavna R, Mohan Babu R, Ramakumar S (2006) Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. J Theor Biol 243:252–260
Mundra P, Kumar M, Kumar KK, Jayaraman VK, Kulkarni BD (2007) Using pseudo amino acid composition to predict protein subnuclear localization: approached with PSSM. Pattern Recognit Lett 28:1610–1615
Murphy KP, Weiss Y, Jordan M (1999) Loopy belief propagation for approximate inference: an empirical study. In: Laskey KB, Prade (eds) Proc Uncertainty Artificial Intelligence. Morgan Kaufmann, San Mateo, pp 467–475
Myers CL, Troyanskaya OG (2007) Context-sensitive data integration and prediction of biological networks. Bioinformatics 23:2322–2330
Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M (2005) Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics 21:302–310
Nanni L, Lumini A (2008a) Combing ontologies and dipeptide composition for predicting DNA-binding proteins. Amino Acids. doi:10.1007/s00726-007-0018-1
Nanni L, Lumini A (2008b) Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. Amino Acids. doi:10.1007/s00726-007-0016-3
Niu B, Cai YD, Lu WC, Zheng GY, Chou KC (2006) Predicting protein structural class with AdaBoost learner. Protein Pept Lett 13:489–492
Ng SK, Tan SH (2003) On combining multiple microarray studies for improved functional classification by whole-dataset feature selection. Genome Inform 14:44–53
Pandey G, Kumar V (2007) Incorporating functional inter-relationships into algorithms for protein function prediction. In: ISMB Satellite Meet Automated Function Prediction.
Pandey G, Steinbach M, Gupta R, Garg T, Kumar V (2007) Association analysis-based transformations for protein interaction networks: a function prediction case study. In: KDD '07: Proc 13th ACM SIGKDD Int Conf Knowledge Discovery and data mining. ACM, New York, pp 540–549
Pereira-Leal JB, Enright AJ, Ouzounis CA (2004) Detection of functional modules from protein interaction networks. Proteins 54:49–57
Przulj N, Wigle D, Jurisica I (2004) Functional topology in a network of protein interactions. Bioinformatics 20:340–348
Pu X, Guo J, Leung H, Lin Y (2007) Prediction of membrane protein types from sequences and position-specific scoring matrices. J Theor Biol 247:259–265
Pugalenthi G, Tang K, Suganthan PN, Archunan G, Sowdhamini R (2007) A machine learning approach for the identification of odorant binding proteins from sequence-derived properties. BMC Bioinformatics 8:351
Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaufmann, San Francisco
Rives AW, Galitski T (2003) Modular organization of cellular networks. Proc Natl Acad Sci USA 100:1128–1133

Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, Tetko I, Guldener U, Mannhaupt G, Munsterkotter M, Mewes HW (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res 32:5539–5545

Samanta MP, Liang S (2003) Predicting protein functions from redundancies in large-scale protein interaction networks. Proc Natl Acad Sci USA 100:12579–12583

Schwikowski B, Uetz P, Fields S (2000) A network of protein–protein interactions in yeast. Nat Biotechnol 18:1257–1261

Segal E, Wang H, Koller D (2003) Discovering molecular pathways from protein interaction and gene expression data. Bioinformatics 19:264–272

Sharan R, Ulitsky I, Shamir R (2007) Network-based prediction of protein function. Mol Syst Biol 3:88

Shen HB, Chou KC (2005a) Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. Biochem Biophys Res Comm 337:752–756

Shen HB, Chou KC (2005b) Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types. Biochem Biophys Res Commun 334:288–292

Shen HB, Chou KC (2006) Ensemble classifier for protein fold pattern recognition. Bioinformatics 22:1717–1722

Shen HB, Chou KC (2007a) EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. Biochem Biophys Res Comm 364:53–59

Shen HB, Chou KC (2007b) Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. Protein Eng Design Select 20:39–46

Shen HB, Chou KC (2007c) Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. Biochem Biophys Res Commun 355:1006–1011

Shen HB, Chou KC (2007d) Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. Protein Eng Design Select 20:561–567

Shen HB, Chou KC (2007e) Signal-3L: a 3-layer approach for predicting signal peptide. Biochem Biophys Res Comm 363:297–303

Shen HB, Chou KC (2007f) Using ensemble classifier to identify membrane protein types. Amino Acids 32:483–488

Shen HB, Chou KC (2007g) Virus-PLoc: a fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. Biopolymers 85:233–240

Shen HB, Chou KC (2007h) Using ensemble classifier to identify membrane protein types. Amino Acids 32:483–488

Shen HB, Yang J, Chou KC (2006) Fuzzy KNN for predicting membrane protein types from pseudo amino acid composition. J Theor Biol 240:9–13

Shen HB, Yang J, Chou KC (2007) Euk-ploc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. Amino Acids 33:57–67

Shi JY, Zhang SW, Pan Q, Cheng Y-M, Xie J (2007) Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. Amino Acids 33:69–74

Shi JY, Zhang SW, Pan Q, Zhou GP (2008) Using pseudo amino acid composition to predict protein subcellular location: approached with amino acid composition distribution. Amino Acids doi: 10.1007/s00726–007–0623-z

Shiga M, Takigawa I, Mamitsuka H (2007) Annotating gene function by combining expression data with a modular gene network. Bioinformatics 23:468–478

Spirin V, Mirny LA (2003) Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci USA 100:12123–12128

Sun XD, Huang RB (2006) Prediction of protein structural classes using support vector machines. Amino Acids 30:469–475

Tanay A, Sharan R, Kupiec M, Shamir R (2004) Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. Proc Natl Acad Sci USA 101:2981–2986

Tan F, Feng X, Fang Z, Li M, Guo Y, Jiang L (2007) Prediction of mitochondrial proteins based on genetic algorithm—partial least squares and support vector machine. Amino Acids 33:669–675

Tantoso E, Li XB (2007) AAIndexLoc: predicting subcellular localization of proteins based on a new representation of sequences using amino acid indices. Amino Acids doi:10.1007/s00726–007-0616-y

Tornow S, Mewes HW (2003) Functional modules by relating protein interaction networks and gene expression. Nucleic Acids Res 31:6283–6289

Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). Proc Natl Acad Sci USA 100:8348–8353

Tsuda K, Shin H, Scholkopf B (2005) Fast protein classification with multiple networks. Bioinformatics 21:59–65

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

Vinga S, Almeida J (2003) Alignment-free sequence comparison–a review. Bioinformatics 19:513–523

Wang M, Yang J, Liu GP, Xu ZJ, Chou KC (2004) Weighted-support vector machines for predicting membrane protein types based on pseudo amino acid composition. Protein Eng Design Select 17:509–516

Wang M, Yang J, Chou KC (2005) Using string kernel to predict signal peptide cleavage site based on subsite coupling model. Amino Acids 28:395–402 (Erratum, ibid. 2005, 29:301)

Wang SQ, Yang J, Chou KC (2006) Using stacked generalization to predict membrane protein types based on pseudo amino acid composition. J Theor Biol 242:941–946

Wen Z, Li M, Li Y, Guo Y, Wang K (2007) Delaunay triangulation with partial least squares projection to latent structures: a model for G-protein coupled receptors classification and fast structure recognition. Amino Acids 32:277–283

Xiao X, Chou KC (2007) Digital coding of amino acids based on hydrophobic index. Protein Pept Lett 14:871–875

Xiao X, Shao S, Ding Y, Huang Z, Huang Y, Chou KC (2005) Using complexity measure factor to predict protein subcellular location. Amino Acids 28:57–61

Xiao X, Shao SH, Ding YS, Huang ZD, Chou KC (2006a) Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. Amino Acids 30:49–54

Xiao X, Shao SH, Huang ZD, Chou KC (2006b) Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. J Comput Chem 27:478–482

Zhang S, Jin G, Zhang XS, Chen L (2007) Discovering functions and revealing mechanisms at molecular level from biological networks. Proteomics 7:2856–2869

Zhang SW, Pan Q, Zhang HC, Shao ZC, Shi JY (2006) Prediction protein homo-oligomer types by pseudo amino acid composition: approached with an improved feature extraction and naive Bayes feature fusion. Amino Acids 30:461–468

Zhang SW, Zhang YL, Yang HF, Zhao CH, Pan Q (2007) Using the concept of Chou's pseudo amino acid composition to predict protein subcellular localization: an approach by incorporating

evolutionary information and von Neumann entropies. Amino Acids. doi:10.1007/s00726–007–0010-9

Zhang TL, Ding YS (2007) Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes. Amino Acids 33:623–629

Zhang T, Ding Y, Chou KC (2006) Prediction of protein subcellular location using hydrophobic patterns of amino acid sequence. Comput Biol Chem 30:367–371

Zhang ZH, Wang ZH, Zhang ZR, Wang YX (2006) A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine. FEBS Lett 580:6169–6174

Zhao XM, Cheung YM, Huang DS (2005) A novel approach to extracting features from motif content and protein composition for protein sequence classification. Neural Networks 18:1019–1028

Zhao XM, Chen LN, Aihara K (2007) Gene function prediction with the shortest path in functional linkage graph. Lect Notes Oper Res 7:68–74

Zhao XM, Chen LN, Aihara K (2008a) Gene function prediction using labeled and unlabeled data. BMC Bioinformatics 9:57

Zhao XM, Chen LN, Aihara K (2008b) Protein classification with imbalanced data. Proteins 70:1125–1132

Zhou GP (1998) An intriguing controversy over protein structural class prediction. J Protein Chem 17:729–738

Zhou GP, Assa-Munt N (2001) Some insights into protein structural class prediction. Proteins 44:57–59

Zhou GP, Doctor K (2003) Subcellular location prediction of apoptosis proteins. Proteins 50:44–48

Zhou XB, Chen C, Li ZC, Zou XY (2007a) Improved prediction of subcellular location for apoptosis proteins by the dual-layer support vector machine. Amino Acids. doi:10.1007/s00726–007-0608-y

Zhou XB, Chen C, Li ZC, Zou XY (2007b) Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. J Theor Biol 248:546–551

Zhou X, Kao MCJ, Wong WH (2002) From the Cover: transitive functional annotation by shortest-path analysis of gene expression data. Proc Natl Acad Sci USA 99:12783–12788

Zhou XJ, Kao MCJ, Huang H, Wong A, Nunez-Iglesias J, Primig M, Aparicio OM, Finch CE, Morgan TE, Wong WH (2005) Functional annotation and network reconstruction through cross-platform integration of microarray data. Nat Biotechnol 23:238–243