



## Guest Editor's Introduction

## Computational protein function predictions



What is the role of this protein? Where does this protein localize in a cell? Are there any ligands that bind to this protein? If so, what are they? Which residues constitute a functional site of the protein? These questions, which in a broader sense seek the biological function of a protein, are fundamental and central in modern biology. Ultimately, the biological function of a protein needs to be determined by experiments. However, a hypothesis is needed to design an assay because it determines whether a target protein has a particular function or not. Biologists come up with hypotheses of protein function from circumstantial evidence, and computational function prediction can play an important part. Computational function prediction methods are also useful for analyzing protein function in a proteomic scale since methods can be applied to a large number of proteins in a realistic time. As more protein function annotations accumulate in various databases, and algorithms advance in the machine learning field, computational protein function prediction methods have become more accurate and reliable in recent years. Moreover, it is noticeable that various different types of predictions have emerged, which also indicates the maturity of the field. Features of proteins that can be used for function prediction ranges from conventional sequence information, structures, to networks of protein associations.

To capture the current landscape of the quite diverse field of computational protein function prediction, this issue collected state-of-the-art function prediction methods of different types. The first three articles [1–3] describes sequence-based methods. The first paper by the Tian group describes their sequence-based function prediction method named GOFDR [1]. GOFDR takes query protein sequence as input, and predicts Gene Ontology (GO) terms for the query from similar sequences to the query that are retrieved from a sequence database, which is similar per se to other existing sequence-based methods. A notable device in GOFDR is that it considers residues that are specific for a particular GO term in transferring the GO term to the query. Argot2.5 by the Toppo group retrieves similar sequences to a query from databases by two methods, BLAST and HMMER3, a hidden Markov Model-based tool, and takes GO terms from the retrieved sequences [2]. An interesting idea implemented in Argot2.5 is that it considers taxonomy information of sequences, namely, GO terms that seem incompatible with the taxon of a sequence are filtered out. The next article by Das and Orengo [3] reviews sequence and structure-based function prediction methods with a focus on their protein classification database, CATH-Gene3D, and associated FunFHMmer server. CATH-Gene3D is a classification of sequences to the CATH protein structural domain classification database and FunFHMmer matches a query sequence to a sequence family

in CATH-Gene3D using a hidden Markov model and thereby predict function of the query.

The next three methods [4–6] take a protein tertiary structure as input and predict binding ligands for the query protein. Binding ligand prediction is not only for function prediction of proteins but it is also useful for drug design. The method developed by Nakamura and Tomii represents a binding pocket by an ensemble of all triangles consisting of three  $C\alpha$  atoms in the pocket [4]. The triangles are classified into 171,700 types considering amino acid types of vertices and edge distances. Thus, a pocket is represented by a vector that shows the frequency of each triangle type in the pocket, which is further reduced to a vector of 11 elements by multi-dimensional scaling. A query pocket is compared with pockets of known binding ligands in terms of the 11-element vector, and a binding ligand is predicted from identified similar pockets to the query. The Kihara group used a different representation of ligand binding pockets [5]. They used 3D Zernike descriptor (3DZD), a mathematical series expansion, for representing surface of binding pockets and ligand molecules. 3DZD was applied for pocket-to-pocket comparison in the method named Patch-Surfer as well as pocket-to-ligand comparison in PL-PatchSurfer. The last paper in this category, written by Ondrechen et al., describes their method, SALSA, and its application to functional subclass prediction of glycoside hydrosidases [6]. SALSA compares predicted active site residues of a query protein to the known consensus active site residues of functional families. Active site residue prediction for a query is performed by a combination of three methods, a sequence-based, a pocket-structure-feature-based, and an electrostatics-based method named THEMATICs. THEMATICs is a unique method developed earlier by the author's group, which identifies ionizable residues with a perturbed titration curve as functional residues by solving the Poisson–Boltzmann equation.

The subsequent two methods provide different types of annotations to query protein structures, namely, protein interaction sites and peptide docking prediction. Maheshwari and Brylinski report eFindSitePPI, which predicts protein binding site residues in query protein structure by considering five properties of residues in machine learning frameworks [7]. CABS-dock, developed by the Kolinski group, docks a peptide onto a receptor protein structure using a coarse-grained protein structure model, CABS [8].

The last category is network-based methods [9–13]. The first paper in this category, written by Cao and Cheng, presents a GO term prediction method for a query protein that uses a combination of three information sources: sequence similarity, protein–protein and gene-expression networks, and local sequence statistics [9]. In the next article, the Pandey group systematically

analyzes ensemble approaches that use heterogeneous information sources. They provide a software package, DataSink, for generating diverse ensembles of classifiers [10]. Brun and her colleagues discuss a method to identifying network modules [11], which have different expression profiles. They discuss that such “dysregulated” modules identified through different stages of cancer progression can be considered as potential cancer proteins. Next, Xianghong Zhou and her colleagues review two methods developed in their lab, which predict function of isoforms of proteins using gene co-expression networks constructed from RNA-Seq data [12]. Liu and Hu developed a method for predicting subcellular localization of genes from gene networks (e.g. protein–protein interaction network, gene co-expression network) using a kernel-logistic regression [13]. They further applied the method to gene co-expression networks of disease (cancer) and normal states, and predict genes that differ their localization in the two states as candidates that are responsible for the disease.

The computational protein function prediction field will certainly evolve further and will be more routinely used by biology labs. This special issue exhibits a snapshot of such active developments. It is the editor's pleasure if the issue can invite experimental biologists to use the methods introduced in the articles and also provide useful hints for computational biologists who develop new prediction methods.

## References

- [1] Qingtian Gong, Wei Ning, Weidong Tian, GoFDR: a sequence alignment based method for predicting protein functions, *Methods* 93 (2016) 3–14.
- [2] Enrico Lavezzo, Marco Falda, Paolo Fontana, Luca Bianco, Stefano Toppo, Enhancing protein function prediction with taxonomic constraints - the Argot2.5 web server, *Methods* 93 (2016) 15–23.
- [3] Sayoni Das, Christine A. Orengo, Protein function annotation using protein domain family, resources, *Methods* 93 (2016) 24–34.
- [4] Tsukasa Nakamura, Kentaro Tomii, Protein ligand-binding site comparison by a reduced vector representation derived from multidimensional scaling of generalized description of binding sites, *Methods* 93 (2016) 35–40.
- [5] Woong-Hee Shin, Mark Gregory Bures, Daisuke Kihara, PatchSurfers: two methods for local molecular property-based binding ligand prediction, *Methods* 93 (2016) 41–50.
- [6] Ramya Parasuram, Caitlyn L. Mills, Zhouxi Wang, Saroja Somasundaram, Penny J. Beuning, Mary Jo Ondrechen, Local structure based method for prediction of the biochemical function of proteins: applications to glycoside hydrolases, *Methods* 93 (2016) 51–63.
- [7] Surabhi Maheshwari, Michal Brylinski, Template-based identification of protein–protein interfaces using eFindSitePPI, *Methods* 93 (2016) 64–71.
- [8] Maciej Blaszczyk, Mateusz Kurcinski, Maksim Kouza, Lukasz Wieteska, Aleksander Debinski, Michal Jamroz, Andrzej Kolinski, Sebastian Kmiecik, Modeling of protein–peptide interactions using the CABS-dock web server for binding site search and flexible docking, *Methods* 93 (2016) 72–83.
- [9] Renzhi Cao, Jianlin Cheng, Integrated protein function prediction by mining function associations, sequences, and protein–protein and gene–gene interaction networks, *Methods* 93 (2016) 84–91.
- [10] Sean Whalen, Om P. Pandey, Gaurav Pandey, Predicting protein function and other biomedical characteristics with heterogeneous ensembles, *Methods* 93 (2016) 92–102.
- [11] Andreas Zanzoni, Christine Brun, Integration of quantitative proteomics data and interaction networks: identification of dysregulated cellular functions during cancer progression, *Methods* 93 (2016) 103–109.
- [12] Wenyuan Li, Chun-Chi Liu, Shuli Kang, Jian-Rong Li, Yu-Ting Tseng, Xianghong Jasmine Zhou, Pushing the annotation of cellular activities to a higher resolution: predicting functions at the isoform level, *Methods* 93 (2016) 110–118.
- [13] Zhonghao Liu, Hu Jianjun, Mislocalization-related disease gene discovery using gene expression based computational protein localization prediction, *Methods* 93 (2016) 119–127.

*Editor*

Daisuke Kihara

*Department of Biological Sciences/Computer Science,  
Purdue University, West Lafayette, IN 47907, USA  
E-mail address: dkihara@purdue.edu*