Editorial

Machine Learning Models in Protein Bioinformatics

Lukasz Kurgan[1] and Yaoqi Zhou[2]

[1]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada; [2]Indiana University School of Informatics, Indiana University–Purdue University and Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN, USA

Bioinformatics is a relatively new field concerned with the computational analysis and prediction of properties of biomolecules, DNA, RNA, and proteins, in particular, on a genomic/proteomic scale. Machine learning models play increasingly important roles in development of novel methodologies, summarization, and high-throughput analysis in the bioinformatics field. Advances in the related area, including protein structure and function prediction [1, 2], structural bioinformatics [3], and peptide analysis [4] were recently summarized, and several works that overview specific sub-areas of protein bioinformatics, such as prediction of secondary structure [5, 6], helical transmembrane proteins [7], localization and targeting [8], binding sites [9, 10], and RNA-binding [11], were published in the last couple of years. This issue provides a comprehensive overview of current efforts related to the analysis of protein data, from sequences to structures to functions. It consists of two parts, the first with five reviews and the second that includes seven original methodology papers.

The first review by Xin and Radivojac summarizes approaches for the computational identification of functional residues in protein structures and discusses their applications in functional proteomics, including prediction of catalytic residues, post-translational modifications, and nucleic acid-binding sites. The second manuscript by Kurgan and Disfani provides a comprehensive review of ten one-dimensional structural descriptors of proteins and comparatively summarizes over eighty computational models that are used to predict these descriptors from the protein sequences, primarily focusing on the prediction of secondary structure, relative solvent accessibility, and disorder. The review by Gromiha and Huang discusses machine learning-based and statistical methods for the computational prediction of protein folding rates and stability. The fourth paper by Zhou and coworkers overviews and compares current techniques for the prediction of small open reading frames and emphasizes the need for further research in this area. The last review introduces cellular automata and concentrates on its applications in the protein bioinformatics.

The first original research paper by Kihara and coworkers describes the three-dimensional Zernike descriptor, which is used to describe molecular surfaces, and overviews several applications of this descriptor. In the next paper, Qin and Zhou introduce their DISPLAR method that aims at the accurate protein structure-based prediction of DNA binding sites. The manuscript by Xu and coworkers describes and evaluates a new sampling-based machine learning method to rank protein structural models by integrating multiple scores and features. The next two original contributions describe new methodologies for the protein model quality assessment. The work by Martin, Mirabello, and Pollastri concerns an efficient knowledge-based approach that utilizes neural network pairwise interaction fields. The paper by Meller and coworkers introduces a method based on the prediction of relative solvent accessibility using support vector regression, which is applied to soluble and alpha-helical membrane proteins. The next contribution by Hwang *et al.* investigates a relation between contact numbers and catalytic residues to build a simple and effective predictor of the catalytic residues. We close the issue with the paper by Yin, Fan, and Shen which proposes and evaluates an accurate nearest neighbor-based method for the prediction of the conotoxin superfamily.

We are excited to deliver this comprehensive issue that tackles a diverse set of developments in the area of protein bioinformatics. We hope that it will constitute an indispensable resource for bioinformaticians, computer scientists, computational biologists, biophysicists, and biochemists.

Last but not least, we would like to thank all the authors who make this issue possible. We are in great debt to the 26 anonymous reviewers from around the world who delivered timely and useful comments to the authors. The guest editors also express their gratitude to the Editor-in-Chief Prof. Ben M. Dunn for his invitation and support that resulted in the successful completion of this special issue.

REFERENCES

[1] Pavlopoulou, A.; Michalopoulos, I. State-of-the-art bioinformatics protein structure prediction tools. *Int J Mol Med.*, **2011**, doi: 10.3892/ijmm.2011.705. [Epub ahead of print]

[2] Shenoy, S.R.; Jayaram, B. Proteins: sequence to structure and function - current status. *Curr. Protein Pept. Sci..*, **2010**, *11* (7), 498-514.

[3] Hamelryck, T. Probabilistic models and machine learning in structural bioinformatics. *Stat. Methods Med. Res.*, **2009**, *18* (5), 505-526.

[4] Yang, Z.R. Peptide bioinformatics: peptide classification using peptide machines. *Methods Mol. Biol.*, **2008**, *458,* 159-183.

[5] Zhang, H.; Zhang, T.; Chen, K.; Kedarisetti, KD.; Mizianty, M.J.; Bao, Q.; Stach, W.; Kurgan L. Critical assessment of high-throughput standalone methods for secondary structure prediction. *Briefings in Bioinform.*, **2011**, 2011doi: 10.1093/bib/bbq088. [Epub ahead of print]

[6] Pirovano, W.; Heringa, J. Protein secondary structure prediction. *Methods Mol. Biol.*, **2010**, *609*, 327-348.

[7] Tusnády, G.E.; Simon, I. Topology prediction of helical transmembrane proteins: how far have we reached? *Curr. Protein Pept. Sci*., **2010**, *11* (7), 550-561.

[8] Rastogi, S.; Rost, B. Bioinformatics predictions of localization and targeting. *Methods Mol. Biol.*, **2010**, *619*, 285-305.

[9] Leis, S.; Schneider, S.; Zacharias, M. In silico prediction of binding sites on proteins. *Curr. Med. Chem*., **2010**, *17* (15), 1550-1562.

[10] Chen, K.; Mizianty, M.; Gao, J.; Kurgan, L. A critical comparative assessment of predictions of protein binding sites for biologically relevant organic compounds. *Structure*, **2011**, *19* (5), 613-621.

[11] Zhang, T.; Zhang, H.; Chen, K.; Ruan, J.; Shen, S.; Kurgan, L. Analysis and prediction of RNA-binding residues using sequence, evolutionary conservation, and predicted secondary structure and solvent accessibility. *Curr. Protein Pept. Sci.*, **2010**, *11* (7), 609-628.