AUTOMATED FUNCTION PREDICTION

New avenues in protein function prediction

IDDO FRIEDBERG, MARTIN JAMBON, AND ADAM GODZIK

Burnham Institute for Medical Research, La Jolla, California 92037, USA

(RECEIVED October 3, 2005; FINAL REVISION February 15, 2006; ACCEPTED February 15, 2006)

The huge influx of protein sequence and structure information is becoming more a quagmire of data rather than the font of knowledge that was anticipated. The latest tally of sequences in GenBank stands at >100 gigabases, and in Protein Data Bank (PDB) there are 34,917 structures (as of January 31, 2006). Of those, ~40% and 1%, respectively, are characterized as ''unknown function.'' The comparatively low fraction of unknowns in PDB reflects the large effort spent solving each protein structure, part of which is directed to functional characterization. Nevertheless, the unknown function segment is rapidly growing in PDB and in Structural Genomics centers (Chandonia and Brenner 2006). With the advent of cheaper and faster techniques, both for sequencing and for solving protein structures, we can only expect this trend to accelerate. The best example of this trend is the recent flood of environmental genomic data (metagenomics) that is already dwarfing the output from all previous genome sequencing efforts and consists almost solely of predicted proteins with unknown functions (Tringe and Rubin 2005).

In the life sciences, this sheer volume of raw data in need of annotation is unprecedented. Computational biology is being called upon, now more than ever, to process these data and provide us with biochemical, physiological, and evolutionary context. Even though experimental high-throughput functional annotations have seen such breakthroughs as RNAi and large-scale binding studies, the time and cost of determining the function of every single gene and gene product are prohibitive. Therefore, most of the functional annotation will be done with computational tools. However, an increase in genomic data means an increase not only in the number of sequences and structures but also in their diversity. Simple homology transfer—annotation by inferring functionality from homologous sequences or structures—is telling us less and less about what proteins are actually doing.

Reprint requests to: Iddo Friedberg, Burnham Institute for Medical Research, 10901 N. Torrey Pines Rd. La Jolla, CA 92037 USA; e-mail: idoerg@burnham.org; fax: (858) 713-9949.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.062158406.

In the first Automated Function Prediction (AFP) meeting, some 100 researchers and students got together to explore new methods of computational protein function prediction. The AFP meeting was held in June 2005 in Detroit, Michigan, alongside the 2005 Intelligent Systems in Molecular Biology conference. The meeting brought people coming from very diverse backgrounds yet sharing common interests together for a day of talks, panel discussions, and poster presentations about protein function prediction. Fifteen talks were selected for oral presentation from a total of 40 submissions. These talks covered a wide range of protein function prediction based on amino acid sequence, three-dimensional structure, genomic context, and more. The proceedings' extended abstracts are available at http://BioFunctionPrediction.org/AFP/previousmeets/afp05/. For this special section, we have chosen five studies to be published as full-length articles, illustrating the breadth and depth of computational protein function prediction.

The first study in the AFP 2005 Section, performed by David Kristensen and colleagues (2006), tackles the problem of function prediction by locating functional sites in protein structures and associating them with specific enzymatic functions. This question is especially important in this age of structural genomics when protein structures are being solved but the function of the proteins is unknown. In their study, they locate important residues by using the Evolutionary Trace method. This method locates evolutionarily conserved residues, then locates spatial patterns of such conserved residues and uses a classification algorithm to successfully distinguish between different enzymes. By using the enzyme commission classification (EC), Kristensen's group shows that their two-tiered method—evolutionary conservation and then pattern matching—succeeds in discriminating between different enzymatic functions.

Bandyopadhyay and colleagues (2006) deal with a similar problem. Their study focuses on locating function-specific fingerprints. By using a graph representation of the protein, they search for subgraphs that are uniquely associated with a function. Normally, this approach can be

very expensive computationally. However, Bandyopadhyay's group has developed a fast subgraph isomorphism algorithm that manages to classify structures into different functional classes.

Note that the above two studies use function-prediction methods based on structure information. It is quite interesting that not less than five years ago, knowing a protein's structure meant that most aspects of its function were already known. This is because the expense and time associated with solving a protein structure required that individual laboratories pick their targets very carefully: Only structures that were deemed biologically interesting and/or easy to solve were targeted. The question this posed was: "We know what this protein does, so how will we solve its structure to find out how it does that?" With the advent of structural genomics, a new question arises: "We have the molecular structure of this protein, so can we infer from that what it actually does?" We are faced with a rapidly increasing number of proteins whose structures are known but whose functions are not. Consequently, predicting function from structure is a relatively young field and quite an exciting one.

Another interesting point to note is that different proxies for function are being used in these two articles: the enzyme commission classification in the Kristensen article, and SCOP superfamilies in the Bandyopadhyay article. Choosing different proxies for function description is not as idiosyncratic as it may seem at first: Different aspects of function are described by using different devices. The EC is well suited for functional description when the functional aspect of interest is enzymatic function. SCOP superfamilies provide a more general proxy when dealing with protein structures. Different proxies for function, while understandable, result in a different choice of vocabulary to describe function. A standardized vocabulary is essential in the field of functional annotation. To solve this problem, controlled vocabularies for different aspects of protein function were created. The most widely accepted of such controlled vocabularies is the Gene Ontology (GO) database and informatics resource (Ashburner et al. 2000). GO is a collection of well-defined terms describing three different aspects of protein function: molecular function (such as catalytic activity or transporter activity), cellular location (such as mitochondria or Golgi), and biological process (such as apoptosis or transcription). Each aspect is also called an ontology, and each ontology is represented as a semi-hierarchical graph, with more specific terms being derived from more general terms. Thus, the ontological framework provides a standard and comprehensive way for describing protein function. Having a standard also enables us to assess computational predictions using a distance measure based on the similarity of GO terms. In the third article in this section, Karin Verspoor and colleagues (2006) investigate a way of doing so. They suggest a method that assesses predictions based on the edge distance of terms in the GO graph but takes into account that specific terms should be treated differently than more general terms. By using a gold standard of GO annotated proteins, they evaluate their own function-prediction method that functionally annotates proteins by using the results derived from PSI-BLAST. Several alternatives for a GO-based distance measure were suggested recently (Lord et al. 2003; Shakhnovich 2005), and it would be interesting to compare and contrast them.

For the simple reason that we have 100-fold more protein sequences than structures, most functional annotation takes place in sequence space rather than in structure space. Here we include two such methods. The first, by Troy Hawkins and Daisuke Kihara (Hawkins et al. 2006), relies on understanding the frequency of which certain functions are associated with other functions. Hawkins and Kihara have scanned GO annotated protein function databases and have determined the frequencies with which each GO term is associated with other GO terms. By doing so, they increase the sensitivity of traditional sequence similarity search to function determination, and they manage to attribute GO terms from weakly found sequence similarities.

In the second study, Ori Sasson and colleagues (2006) show an example of how classification algorithms can be easily adapted to infer annotation. This group has adapted a hierarchical protein classification system to infer the function from protein sequences. Their system, Protonet, provides an unsupervised hierarchical clustering of protein sequence space. By localizing the unknown protein to a given cluster and using local similarities, they manage to assign function with high accuracy.

To conclude, the field of automated prediction of gene and protein function is emerging as a new and exciting discipline in computational biology. This section provides a sample of the diversity of the methods used in predicting protein function. It also serves to highlight various perils and pitfalls, such as the problems associated with defining function and with assessing the accuracy of prediction schemes. Having a glut of sequences and structures to annotate, computational function prediction requires faster and stronger computational tools without sacrificing annotation accuracy. It is a pleasure to see that so many researchers are interested in this field, and we hope that you enjoy and benefit from reading the following studies as much as we have.

Acknowledgments

There were many people involved in setting up the AFP 2005 meeting and the resulting section in *Protein Science*. First and foremost, we are grateful to the International Society of Computational Biology for providing us with an excellent setting for the AFP meeting. Thanks to Hershel Safer, B.J. Morrison McKay, Steven Leard, and Stephanie Hagstrom for constant support and for streamlining the process of setting up

the meeting. We are indebted to Michael Sternberg, Russ Altman, Patricia Babbitt, and Olivier Lichtarge for their encouragement and superb plenary talks; Sri Krishna Subramanian for his thorough functional analysis of proteins used for function prediction; Cindy Cook for editorial work on the AFP 2005 booklet and on this special section; and Ilan Samish and Aviv Regev for invaluable advice. We would like to thank 20 referees for their hard work in reviewing the manuscripts submitted to this special *Protein Science* section. Finally, we are grateful to the 100 attendees and especially the 25 poster presenters and speakers at AFP 2005. Thank you all for sharing your work and for making an excellent and enjoyable meeting. The AFP 2005 meeting was sponsored by The International Society for Computational Biology, the Burnham Institute for Medical Research, SAIC, and NIH grants P01GM63208 and U54GM074898.

References

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25:** 25–29.

Bandopadhyay, D., Huan, J., Liu, J., Prins, J., Ssnoeyink, J., Wang, W., and Tropsha, A. 2006. Structure-based function inference using protein family-specific fingerprints. *Protein Sci.* (this issue).

Chandonia, J.M. and Brenner, S.E. 2006. The impact of structural genomics: expectations and outcomes. *Science* **311:** 347–351.

Hawkins, T., Luban, S., and Kihara, D. 2006. Enhanced automated function prediction using distantly related sequences and contextual association by PFP. *Protein Sci.* (this issue).

Kristensen, D.M., Chen, B.Y., Fofanov, V.Y., Ward, R.M., Lisewski, A.M., Kimmel, M., Kavraki, L.E., and Lichtarge, O. 2006. Recurrent use of evolutionary importance for functional annotation of proteins based on local structural similarity. *Protein Sci.* (this issue).

Lord, P.W., Stevens, R.D., Brass, A., and Goble, C.A. 2003. Investigating semantic similarity measures across the Gene Ontology: The relationship between sequence and annotation. *Bioinformatics* **19:** 1275–1283.

Sasson, O., Kaplan, N., and Linial, M. 2006. Functional annotation prediction: All for one and one for all. *Protein Sci.* (this issue).

Shakhnovich, B.E. 2005. Improving the precision of the structure-function relationship by considering phylogenetic context. *PLoS Comput. Biol.* **1:** e9.

Tringe, S.G. and Rubin, E.M. 2005. Metagenomics: DNA sequencing of environmental samples. *Nat. Rev. Genet.* **6:** 805–814.

Verspoor, K., Cohn, J., Mniszewski, S., and Joslyn, C. 2006. A categorization approach to automated ontological function annotation. *Protein Sci.* (this issue).