Cell. Mol. Life Sci. 64 (2007) 2505 – 2511
1420-682X/07/192505-7
DOI 10.1007/s00018-007-7211-y


**Cellular and Molecular Life Sciences**

# Visions & Reflections (Minireview)

# Computational protein function prediction: Are we making progress?

**A. Godzik, M. Jambon and I. Friedberg**[*, +]

Burnham Institute for Medical Research, 10901 N. Torrey Pines Rd., La Jolla, CA 92037 (USA),
Fax: +1 858 713 9949, e-mail: idoerg@ucsd.edu

Received 1 May 2007; received after revision 24 May 2007; accepted 30 May 2007
Online First 5 July 2007

**Abstract.** The computational prediction of gene and protein function is rapidly gaining ground as a central undertaking in computational biology. Making sense of the flood of genomic data requires fast and reliable annotation. Many ingenious algorithms have been devised to infer a protein's function from its amino acid sequence, 3D structure and chromosomal location of the encoding genes. However, there are significant challenges in assessing how well these programs perform. In this article we explore those challenges and review our own attempt at assessing the performance of those programs. We conclude that the task is far from complete and that a critical assessment of the performance of function prediction programs is necessary to make true progress in computational function prediction.

**Abbreviations.** DAG: Directed Acyclic Graph; CASP: Critical Assessment of Structure Prediction; GO: Gene Ontology.

**Keywords.** Protein function prediction, bioinformatics, CASP, AFP, aspartate dehydrogenase, aspartate oxidase, non-orthologous replacement, NAD synthesis.

The computational prediction of protein function is a central undertaking in computational biology. We are being deluged with genomic information from genomic and more recently metagenomic [1] projects. Although high-throughput experimental methods are making great progress, the only effective way to annotate genes and genomes *en masse* is by computational means. However, the computational prediction of function presents unique challenges. Chiefly, the increasing diversity of protein sequences requires new means of annotation beyond homology assignments [2–4]. Additional concerns are standardizing the vocabulary of functional annotation [5–7] and the assessment of prediction programs [8]. The first two topics, new means of annotation and vocabulary standardization, have been reviewed extensively. However the third problem, that of how well computational function prediction is performing, has not been addressed as widely; this is chiefly because a methodology for assessing the accuracy of function prediction programs has not yet been developed to any kind of satisfaction. Nevertheless it should be done, as function prediction is rapidly becoming a leading problem in computational molecular biology, and we should know how well individual programs and the field as a whole are performing. In this brief essay, we explain the problems and outline the major

* Corresponding author.
+ Present address: CalIT2 mail code #0436, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093 – 0436 (USA).

challenges facing the assessment of computational function prediction. We then proceed to describe one test case from our own attempt at assessing function prediction programs that took place at the first annual Automated Function Prediction meeting (AFP 2005).

Over the past two decades, many computational methods for predicting protein function from sequence or structure data have been developed. Searching annotated databases for homologous sequences is probably the method used most often, with the BLAST [9] suite of programs dominating. Other methods include searching databases that contain curated information regarding specific protein families, most popular being Pfam [10] and the Conserved Domains Database (CDD) [11]; other notable efforts include PANTHER [12], STRING [13] and TIGRFAM. In prokaryotes, chromosomal location adds a measure of predictive ability [14, 15], since chromosomally adjacent genes are often involved in the same cellular pathway. Since protein structure is more conserved than sequence, some methods use structural information to predict function (for reviews see [4, 16, 17]). However, protein structure information is usually unavailable, making prediction of function from structure an interesting problem, yet with a more limited scope than prediction from sequence. Given that so many different computational function prediction methods exist, there is a pressing need to assess the performance of those methods. A standardized assessment of functional annotation quality is important for understanding the strengths and weaknesses not only of any individual program, but also of the field in general.

When approaching the task of assessing function prediction programs, we should bear in mind that the definition of biological function is highly contextual. Different aspects of biological function of the same protein may be viewed as taking place on different scales of time and space. (Although there is, of course, a lot more to functional aspects than just time and space scales). Typically, enzymatic reactions take place at a distance of a few nanometers and within microseconds. Going up in scale, the same enzyme is invariably part of one or more cellular networks: signal transduction, metabolic pathway, transcription control, etc; those take place over a distance of micrometers and seconds. Going even more up in scale, the enzyme of interest might facilitate embryonic development, a process measured in months and in centimeters. There is no single computational or experimental method that will let us know all the functional aspects of a protein. Indeed, physiological experiments conducted on a cellular or organismal level rarely reveal the biochemical function of a newly characterized protein but do tell us about its role in the life of the cell. Conversely, purification of a novel protein and subjecting it to a biochemical assay would not, by itself, tell us of its cellular role. Therefore, when communicating information regarding protein function, context is of primary importance: we should always know which functional aspect is being discussed and formalize this definition in a form amenable to computational processing. As we shall see below, the framework of a comprehensive, controlled vocabulary does exactly that.

There are two major inherent problems with comparative assessment of the ability to predict functions: First, there is the problem of the benchmark: how do we select the proteins whose functions are to be predicted? The function of the benchmark proteins should not be well known, as those would be easily discovered by a simple sequence similarity search. For that reason, well-annotated proteins or their close homologs cannot be used. On the other hand, the functions of proteins that bear no sequence or structural similarity to annotated proteins are seldom known! This inability to properly blind a test set for functional prediction we call the test-set blinding problem. One solution to the test-set blinding problem would be the selection of recently discovered proteins that have not yet been published or whose published annotation has failed to percolate into the canonical annotation databases upon which most function prediction programs rely. Another solution is backdating the annotation database that is storing predictions and then checking them at some future date.

The second challenge in assessing predictions is that of establishing a metric for scoring. This problem is best illustrated by an example: Suppose that we are investigating the ability of function prediction programs to predict the function of a protein known to be a collagenase. Two prediction programs are being compared, but neither provides a perfect prediction, *i.e.* they both fail to report the protein to be a collagenase. One program predicts that the protein is a serine peptidase, and the other predicts that it is a fibrolase. Both are wrong, but from a biochemist's point of view, the second program is less wrong, since both collagenases and fibrolases are metallo-endopeptidases, whereas serine peptidases have a different catalytic mechanism. How can we score this difference between a near miss and a wide miss? To do so, we need a system that includes (1) a controlled vocabulary to avoid errors caused by semantics and (2) a metric between terms in that vocabulary. Fortunately, the Gene Ontology (GO) [6] fulfills these conditions. GO is a hierarchical, controlled vocabulary used to describe genes and gene product attributes in any organism. Having the functional terms organized on a

hierarchical directed acyclic graph (DAG) makes it possible to set a metric based on the edge distance between terms. Specifically, in the experiment detailed here, we used the *semantic similarity* measure introduced by Lord *et al.* [18]; this metric calculates the distance between any two terms in an ontology based on the probability of both terms and the distance between them (for details see Fig. 1). Other semantic similarity measures have subsequently been published [19, 20]. As our function similarity measure is GO-based, we assessed only those programs that give results using the Gene Ontology. The programs assessed were Spearmint [21], Rulebase [22], Annolite, PFP [23], PhydBac [24], ProKnow [25], Proteome Analyst [26] and GOpet [27, 28]. Gene Ontology also formalizes the definition of functional context that has been addressed above; namely, GO has three "ontology trees": for molecular function, biological process and cellular component. Molecular function describes activities on a molecular level, such as binding or catalysis. Biological process describes functions that are series of events accomplished by assemblies of molecular functions, examples including "apoptosis" or "thiamine synthesis". Finally, cellular component is the compartment or compartments of the cell in which the protein shows its activity. Each protein can be annotated by one or more terms from each of these ontologies. To simplify matters, we only tested programs for predictions in the molecular function ontology.

In order to overcome the problems with benchmarks described above, we reviewed recent literature for findings such as new protein function and non-orthologous replacements. Since our group is involved in the Protein Structure Initiative effort, we also looked at proteins with new folds whose structures have been recently determined in structural genomics projects. We compiled a small set of five proteins that are not obviously homologous to any other annotated proteins. Here we report on one interesting case study that illustrates some of the challenges associated with automated function prediction and the assessment thereof. The full list of proteins and the assessment results are available at http://biofunctionprediction.org/AFP/previousmeets/afp05/results.

**TM1643: a non-orthologous replacement for L-aspartate oxidase.** The open reading frame (ORF) TM1643 was identified in the genome of the hyperthermophilic bacterium *Thermotoga maritima.* This ORF encodes a soluble, 241-residue protein that is conserved in several (15) organisms, including in humans and in *Caenorhabditis elegans.* A function of this protein cannot be inferred from its sequence, as it has no similarity to other proteins with known function. The structure of TM1643 was solved in 2002 [29]. Its function was deduced from its location next to homologs of NadA and NadC from *Escherichia coli*, two enzymes that catalyze the *de novo* synthesis of nicotinamide adenine dinucleotide (NAD) from L-aspartate. In prokaryotes, the first reaction in this pathway is catalyzed by the flavin adenine dinucleotide (FAD)-containing L-aspartate oxidase (NadB), which oxidizes L-aspartate to iminoaspartate using fumarate or oxygen as electron acceptors. TM1643 occupies the position of a third enzyme in this operon, NadB, yet shares no sequence- or structure-based similarity with NadB. Furthermore, purified TM1643 showed no evidence for the presence of an FAD cofactor, and the enzyme has no L-aspartate oxidase activity. Therefore, it is highly unlikely that TM1643 is an aspartate oxidase. Kinetic studies have shown, however, that TM1643 does have specific L-aspartate dehydrogenase activity, which produces iminoaspartate, the first product required for this pathway [29]. Based on this strong *in vitro* evidence, it was suggested that TM1643 is an L-aspartate dehydrogenase and a non-orthologous replacement for NadB. The difficulty in evaluating TM1643 function predictions was that, being a novel function, "L-aspartate dehydrogenase" was not a term that appeared in GO. However, the obvious parent term is "Oxidoreductase CH-NH2 bonds NAD/NADP acceptor" (see Fig. 2). Aspartate dehydrogenase activity has since been incorporated into GO in this fashion.

We submitted the amino acid sequence and 3D structure (for those programs accepting 3D structures) of TM1643 to several function prediction programs. The best minimal subsuming terms given were generalized ones such as "catalytic activity" or, more specifically, "oxidoreductase activity" (Fig. 2b). However, there was an interesting miss that provided a good clue to the function: Phydbac predicted TM1643 to have a "nicotinate nucleotide dephosphorylase activity". *En face*, this prediction was wrong, and the minimal common subsumer it shared with the true activity was "catalytic activity", a wider miss than the other servers whose minimal common subsumer was "oxidoreductase activity" (Fig. 2b). However, it was an informative error: PhydBac predicted nicotinate-binding activity, which does help provide a partial functional picture: TM1643 uses NAD as a cofactor. The reason that PhydBac provided this prediction is that it looks at co-evolution, co-localization and gene fusion events, and it detected a fusion event with quinolinate phosphoribosyl transferase, another enzyme involved in the microbial NAD/NADP biosynthetic pathway. For that reason, PhydBac correctly predicted that TM1643 is involved in NAD/NADP synthesis.

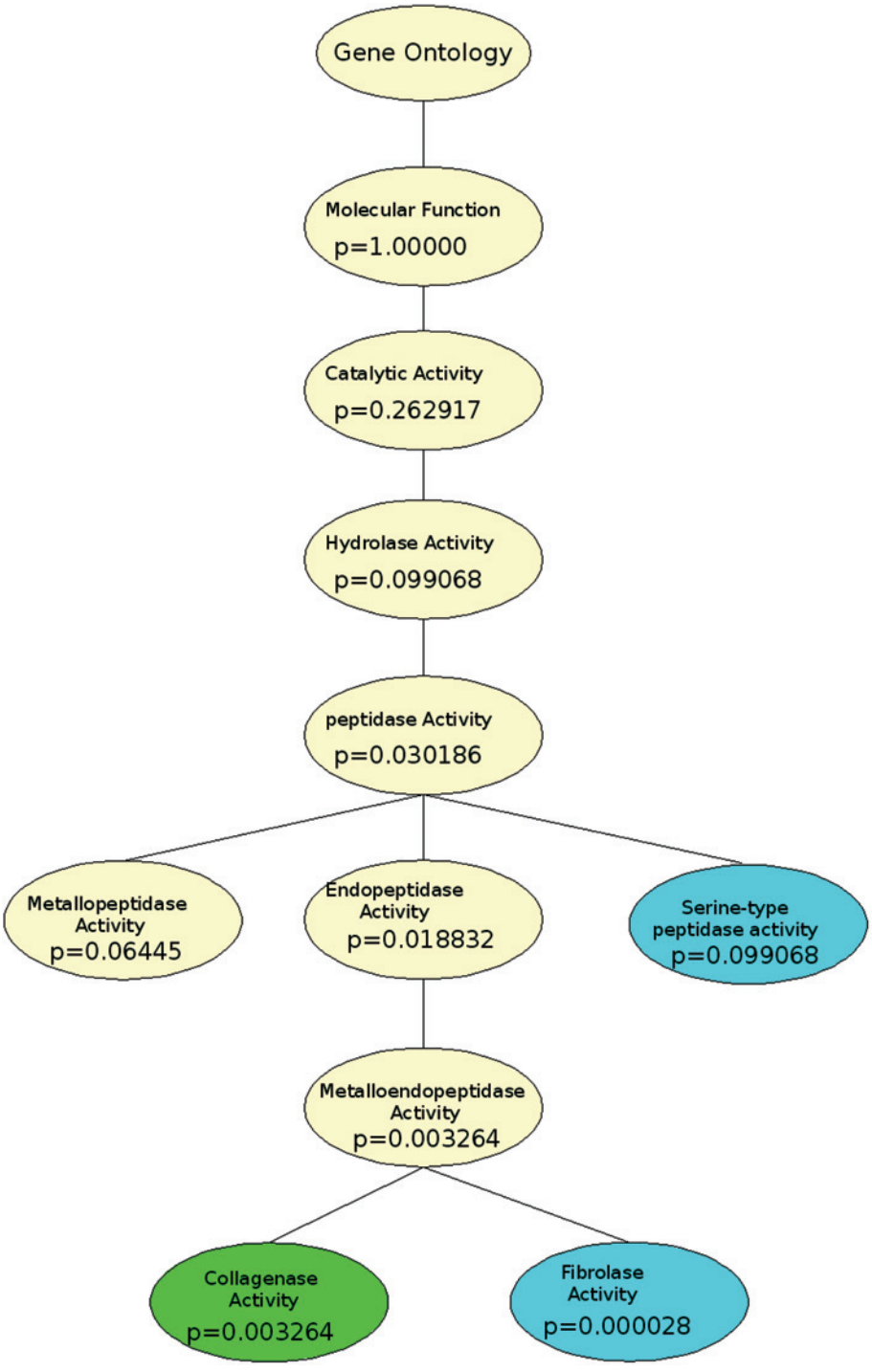


**Figure 1.** Semantic similarity between Gene Ontology (GO) terms using the method of Lord *et al.* [18]. Each node on the GO graph contains a term. Each node is annotated with its frequency in the UniProt database. The frequency is cumulative and includes all the nodes it subsumes. Thus, the root "molecular function" is always $p$=1. The semantic similarity between any two nodes is the minus logarithm of the frequency of the minimal subsuming node of any two nodes in question. Thus, the semantic similarity of "collagenase activity" and "fibrolase activity" is –log($p$("metalloendopeptidase activity")) =5.724. Likewise, the semantic similarity between "serine type peptidase activity" and "collagenase activity" is –log($p$("peptidase activity")) =3.500. The frequency of the term "peptidase activity" is, by definition, higher than that of "metallopeptidase activity", as it subsumes and includes it. Thus, the semantic similarity between "fibrolase activity" and "collagenase activity" is higher than between "collagenase activity" and "serine-type peptidase activity".

Another issue with the assessment of PhydBac's prediction was the inability of the GO-based metric to accurately gauge the value of its prediction. Obviously this prediction would have rated much better had the AFP 2005 challenge used the GO pathway ontology as well, where PhydBac correctly predicted TM1643 to be in the NAD synthetic pathway. This information is very useful, to the point that an observant human annotator can infer the correct molecular function from the cellular pathway prediction simply by treating TM1643 as a missing puzzle piece, as there is no homolog of L-aspartate oxidase in *T. maritima.* As a matter of fact, that was the process that led to the discovery of aspartate oxidase, as detailed in Yang *et al.* (2003).

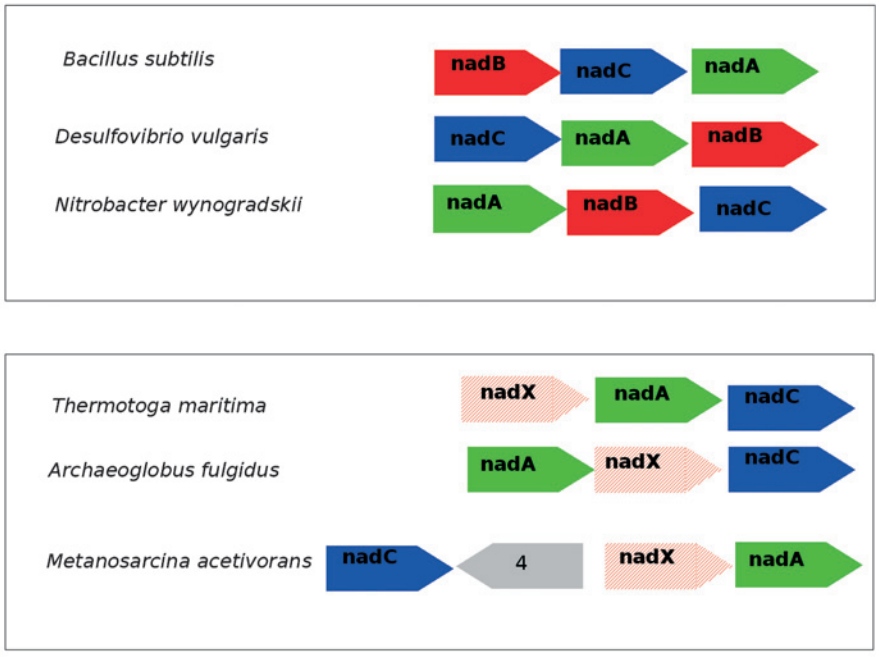


**Figure 2.** (*a*) Three-gene cluster coding for the initial stage of the NAD synthesis pathway. Top: the *Bacillus subtilis* operon and rearrangements in *D. vulgaris* and *N. wynogradskii.* Bottom: nadX, the non-orthologous replacement for nadB, is shown in *T. maritima* and two archea (*A. fulgidus* and *M. acetivorans).* (*b*) Predictions of the different programs in AFP 2005 shown on a Gene Ontology (GO) directed acyclic graph (DAG). The true function "aspartate dehydrogenase activity" is shown in orange in the place it should be given in the GO DAG. Predictions are shown in light green and common subsumers in red. All other nodes in the GO DAG are colered blue. For legibility, some nodes have been omitted and are represented as small blank circles. The names of the programs that made the predictions are parenthesized.

Other attempts were made at assessing protein function predictions. Notably, the last two meetings of the Critical Assessment of Techniques for Protein Structure Prediction (CASP) for assessing protein *structure* predictions have also had short protein *function* predictions sessions. The CASP meetings are an example of a well-structured collaboration between experimental and computational structural biologists aimed at assessing the capabilities of protein structure predictions. As such, it is interesting to note how the CASP organizers dealt with the function prediction problem. In the CASP6 (which took place in 2004) and CASP7 (2006) experiments, there were attempts to predict protein function using the same protein whose structure was being predicted in the main CASP venue, that of structure prediction. However, the main problem faced by the CASP function prediction assessors was that they knew no more of the protein's function than the predictors did. As such, the CASP benchmark did not obey the condition of functional foreknowledge by the assessors that is blinded from the predictors. Therefore, the primary goal in the CASP6 and 7 function prediction experiments was essentially exploratory, to provide useful suggestions for future endeavors [8, 30]. Another goal was to provide researchers studying those proteins with useful information to help them prioritize experiments for functional assignment. This was done using a consensus approach, which the assessors estimated would reflect upon the true nature of the protein's function. Following the CASP6 experiment, the conclusion was made that function prediction should be limited to a specific context – enzymes – and that the vocabulary used should be that of Enzyme Commission Classification numbers (EC) rather than GO terms. This was decided in order to simplify the comparative assessment among different predictions. The conclusion from the CASP7 experiment was that another category for predicting functional sites (as opposed to function) should be established.

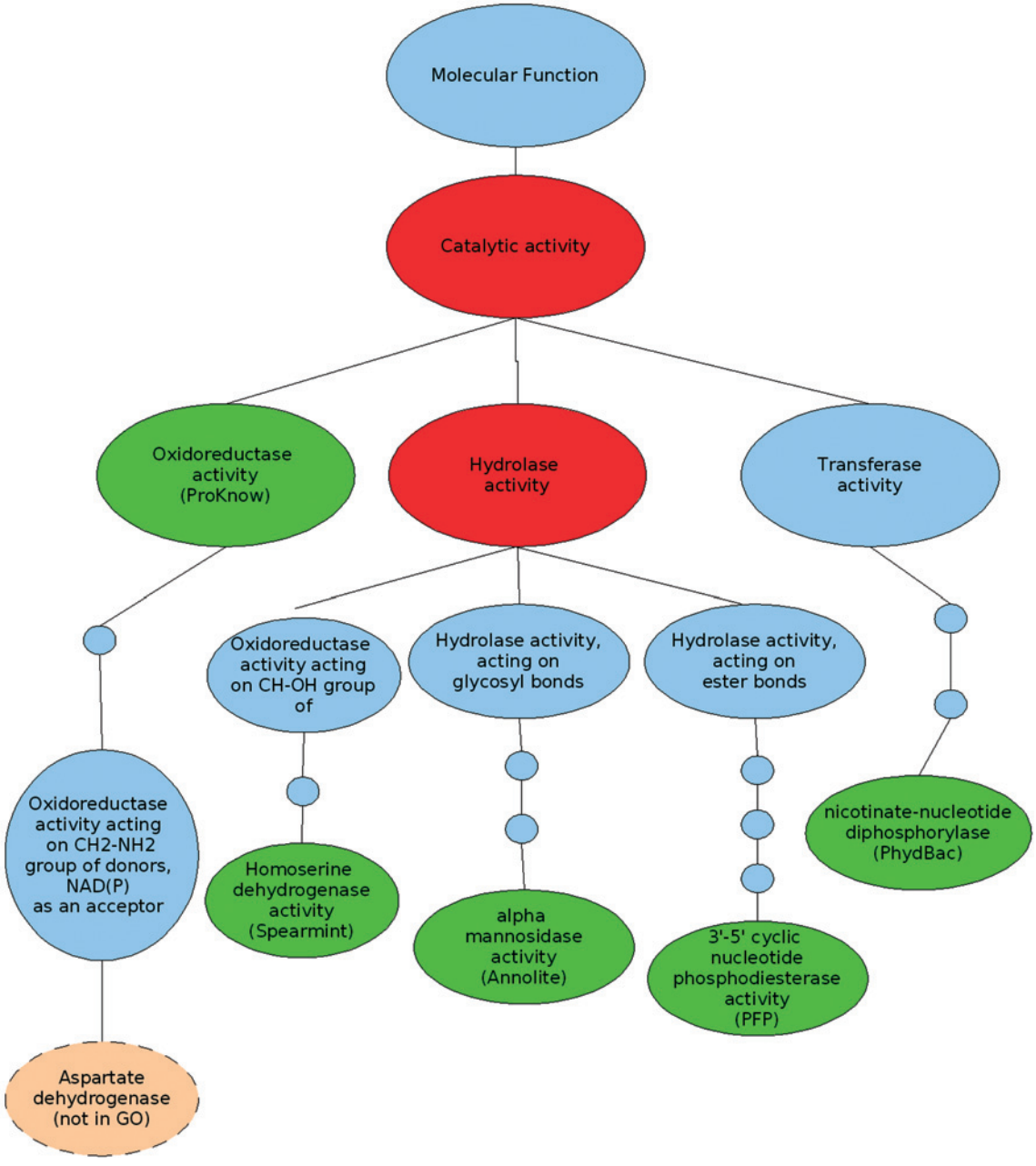


**Figure 2.** *(Continued)*

## Conclusions

The proper assessment of programs for protein function prediction is dependent upon our ability to produce a viable, blind benchmark. Sadly, we are not at that stage yet. Achieving such a benchmark would require a close and well-managed collaboration between computational and experimental scientists to provide challenging cases for function prediction programs. A similar rapport has been established for some years now by computational and experimental structural biologists for the benchmarking of protein structure. The result is the ongoing experiments of CASP for assessing protein structure predictions and Critical Assessment of Protein Interaction Prediction Techniques (CAPRI) for assessing the predictions of protein-protein interactions. Both CASP and CAPRI are very successful undertakings and are considered to be gold standards for the evaluation of structure or interaction prediction programs. It is outside the scope of this article to establish the framework for such an endeavor. However, it appears that the computational biology community is moving in the right direction: both the AFP and the CASP meetings have identified

the problems and have established milestones towards a proper assessment. We hope that this short essay will help facilitate collaborations between experimental and computational biologists so that computational protein function prediction is able to advance using the same critical tools that have helped advance computational protein structure prediction.

*Acknowledgments.* We would like to thank the participants of the AFP 2005 meeting for making a great meeting happen. We are especially grateful to the authors of the prediction programs mentioned in this article for making them available for assessment. We would also like to thank Hershel Safer and the Special Interest Group coordinators of the International Society for Computational Biology (ISCB) for giving us the opportunity to conduct the AFP meeting under ISCB sponsorship. The organizers of AFP 2005 gratefully acknowledge the support of National Institutes of Health grant number 5 U54GM074898–02, Joint Center for Structural Genomics 2.

1 Schloss, P. D. and Handelsman, J. (2005) Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. Genome Biol. 6, 229.
2 Rost, B., Liu, J., Nair, R., Wrzeszczynski, K. O. and Ofran, Y. (2003) Automatic prediction of protein function. Cell. Mol. Life Sci. 60, 2637 – 2650.
3 Ofran, Y., Punta, M., Schneider, R. and Rost, B. (2005) Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery. Drug Discovery Today 10, 1475 – 1482.
4 Friedberg, I. (2006) Automated protein function prediction – the genomic challenge. Brief. Bioinform. 7, 225 – 242.
5 Riley, M. (1993) Functions of the gene products of *Escherichia coli.* Microbiol. Rev. 57, 862 – 952.
6 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., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 25, 25 – 29.
7 Lomax, J. (2005) Get ready to GO! A biologist's guide to the Gene Ontology. Brief. Bioinform. 6, 298 – 304.
8 Soro, S. and Tramontano, A. (2005) The prediction of protein function at CASP6. Proteins 61 Suppl. 7, 201 – 213.
9 Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389 – 3402.
10 Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C. and Eddy, S. R. (2004) The Pfam protein families database. Nucleic Acids Res. 32, D138 – D141.
11 Marchler-Bauer, A., Panchenko, A. R., Shoemaker, B. A., Thiessen, P. A., Geer, L. Y. and Bryant, S. H. (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res. 30, 281 – 283.
12 Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A., Doremieux, O., Campbell, M. J., Kitano, H. and Thomas, P. D. (2005) The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res. 33, D284 – D288.
13 von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. A. and Bork, P. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res. 33, D433 – D437.
14 Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. and Yeates, T. O. (1999) Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proc. Nat. Acad. Sci. USA 96, 4285 – 4288.
15 Enault, F., Suhre, K., Abergel, C., Poirot, O. and Claverie, J. M. (2003) Annotation of bacterial genomes using improved phylogenomic profiles. Bioinformatics 19 Suppl 1, i105 – i107.
16 Bartlett, G. J., Todd, A. E. and Thornton, J. M. (2003) Inferring protein function from structure. Methods Biochem. Anal. 44, 387 – 407.
17 Watson, J. D., Laskowski, R. A. and Thornton, J. M. (2005) Predicting protein function from sequence and structural data. Curr. Opin. Struct. Biol. 15, 275 – 284.
18 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.
19 Schlicker, A., Domingues, F. S., Rahnenfuhrer, J. and Lengauer, T. (2006) A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinformatics 7, 302.
20 Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S. and Chen, C. F. (2007) A new method to measure the semantic similarity of GO terms. Bioinformatics 23, 1274 – 1281.
21 Kretschmann, E., Fleischmann, W. and Apweiler, R. (2001) Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. Bioinformatics 17, 920 – 926.
22 Biswas, M., O'Rourke, J. F., Camon, E., Fraser, G., Kanapin, A., Karavidopoulou, Y., Kersey, P., Kriventseva, E., Mittard, V., Mulder, N., Phan, I., Servant, F. and Apweiler, R. (2002) Applications of InterPro in protein annotation and genome analysis. Brief. Bioinform. 3, 285 – 295.
23 Hawkins, T., Luban, S. and Kihara, D. (2006) Enhanced automated function prediction using distantly related sequences and contextual association by PFP. Protein Sci. 15, 1550 – 1556.
24 Enault, F., Suhre, K. and Claverie, J.-M. (2005) Phydbac "Gene Function Predictor": a gene annotation tool based on genomic context analysis. BMC Bioinformatics 6, 247.
25 Pal, D. and Eisenberg, D. (2005) Inference of protein function from protein structure. Structure (Camb) 13, 121 – 130.
26 Lu, P., Szafron, D., Greiner, R., Wishart, D. S., Fyshe, A., Pearcy, B., Poulin, B., Eisner, R., Ngo, D. and Lamb, N. (2005) PA-GOSUB: a searchable database of model organism protein sequences with their predicted Gene Ontology molecular function and subcellular localization. Nucleic Acids Res. 33, D147 – D153.
27 Vinayagam, A., Konig, R., Moormann, J., Schubert, F., Eils, R., Glatting, K.-H. and Suhai, S. (2004) Applying support vector machines for gene ontology-based gene function prediction. BMC Bioinformatics 5, 116.
28 Vinayagam, A., del Val, C., Schubert, F., Eils, R., Glatting, K. H., Suhai, S. and Konig, R. (2006) GOPET: a tool for automated predictions of Gene Ontology terms. BMC Bioinformatics 7, 161.
29 Yang, Z., Savchenko, A., Yakunin, A., Zhang, R., Edwards, A., Arrowsmith, C. and Tong, L. (2003) Aspartate dehydrogenase, a novel enzyme identified from structural and functional studies of TM1643. J. Biol. Chem. 278, 8804 – 8808.
30 Pellegrini-Calace, M., Soro, S. and Tramontano, A. (2006) Revisiting the prediction of protein function at CASP6. FEBS J. 273, 2977 – 2983.