A New Perspective on an Old Tool: Extending the Coverage of Sequence Similarity-Based Function Prediction with PFP

Troy Bartholomew Hawkins\textsuperscript{1} \\
thawkins@purdue.edu

Stanislav Luban\textsuperscript{2,1} \\
sluban@purdue.edu

Daisuke Kihara\textsuperscript{1,2} \\
dkihara@purdue.edu

\textsuperscript{1} Department of Biological Sciences, Purdue University, 915 West State St., West Lafayette, IN 47907, USA
\textsuperscript{2} Department of Computer Science, Purdue University, 250 N. University St., West Lafayette, IN 47907, USA

Keywords: function annotation, protein function prediction, PSI-BLAST, Gene Ontology

1 Introduction

The ultimate aim of molecular biology is to define functional roles for all proteins. In the last decade, several methods have been developed which produce large amounts of data, including correlated expression analysis by microarrays, fast genome sequencing and large-scale proteomics screens. Computational biologists have been called upon by the experimental community to aid in the organization, analysis and interpretation of this data so that its utility may be maximized. Thus far, proposed computational gene function prediction methods can be grouped into four distinct categories: evolutionary methods, which use conserved global sequence or structure to imply homology and motifs to assign biochemical function and binding sites; genomic methods, which link proteins through domain fusion events, phylogenetic profiling, conserved gene order and common regulatory elements; cellular methods, which use large proteomics datasets to define protein-protein interaction patterns and complexes; and metabolic methods, which utilize the structured networks of biochemical pathways to match proteins to uncharacterized reactions.

The most successful techniques combine clues from multiple contexts to make reliable predictions, but a major limit to function prediction is limited coverage. Even BLAST [1] searches can only cover half of the genes in a genome. In order to provide functional clues that can spark analysis of large proteomics datasets, we need a method that expands coverage by lowering prediction resolution, i.e. a method that can provide accurate (but more generalized) predictions for proteins falling outside of the coverage range for current techniques.

2 Methods

We have designed and implemented a public web server for automated protein function prediction, PFP [2], which uses a sequence-based approach. Our method extends the functionality of a typical PSI-BLAST search in three distinct ways: First, we extract and score Gene Ontology (GO) annotations based on the frequency of their occurrence in highly similar sequences. The GO is a curated hierarchical vocabulary describing the function of proteins in three categories: molecular function, biological process, and cellular component [3]. Second, we utilize relatively weak hits produced by a PSI-BLAST query, not conventionally used for transfer of function annotation. Weakly similar, lower scoring sequences output by PSI-BLAST are not recognized as orthologs to the query sequence, but often represent proteins sharing a common functional domain. These annotations are considered (i.e., not ignored) in the scoring scheme of PFP. Third, we additionally consider those functions which are strongly associated with the highest scoring annotations as described previously. To score these annotations, we designed a novel data mining tool, the Function Association Matrix (FAM), which quantifies the co-occurrence of GO annotations in proteins whose sequences are included in UniProt. Approximately two thirds of associated function pairs mined from...
UniProt bridge annotations from different GO categories. Thus, we can assign function using the FAM that cannot be retrieved directly from PSI-BLAST hits.

Each sequence retrieved by PSI-BLAST is considered as a unique set of GO function annotations, with each annotation retaining the expect (e)-value assigned to the originating sequence. The total score for an individual annotation is essentially the sum of the scores from each occurrence of that annotation in the retrieved sequences. Each of these annotations is also queried against the UniProt FAM to find commonly associated functions. The ten highest ranking annotations in each category are produced as output and represent the most likely annotations assigned to the query sequence.

3 Results and Discussions

In benchmark testing with a set of 2000 randomly chosen protein sequences, PFP has an annotation recovery rate of ~88%. We count an annotation as correct if it shares a biologically significant (GO depth >= 2, category root has GO depth = 0) common parent with the real annotation in the GO graph. PFP’s recovery rate is comparable to automated function prediction methods that use artificial intelligence techniques such as Bayesian-weighted decision tables and support vector machines to combine several sequence-derived functional clues [4,5]. More impressive, however, is that there is no significant drop in performance even when annotations from PSI-BLAST hits with e-values of 20 and below are not included. This result implies that our method of scoring the frequency of occurrence of annotations over even distantly related sequences is more reliable than direct transfer of function annotation between two highly similar sequences. In addition to self-imposed benchmarking, PFP was selected to participate in a third-party assessment of automated protein function prediction servers hosted by the Burnham Institute and presented in a special interest group meeting at ISMB 2005 [6,7]. In the assessment, PFP outscored all other participants, predicting functions for five manually selected target protein sequences.

With PFP, we extend the limits of sequence-based function prediction and provide a backbone algorithm for a comprehensive, automated, context-based method of predicting GO annotations from a single query.

References


