

LANL_PFIG

Nearest Neighbor Categorization for Function Prediction

K. Verspoor, J. Cohn, S. Mniszewski and C. Joslyn

Los Alamos National Laboratory

verspoor@lanl.gov

We present the methods utilized in a system aimed at predicting the function of CASP targets, as represented by a node in the Gene Ontology². The strategy we follow is to (1) identify close neighbors of a target sequence in sequence space, (2) collect the Gene Ontology nodes associated with these neighbors in a curated data set (Swiss-Prot), and (3) categorize the collection of Gene Ontology nodes based on their distribution in the Gene Ontology structure, utilizing a technology called the Gene Ontology Categorizer⁴. The resulting set of Gene Ontology nodes is interpreted as the most representative nodes for the function of the original target sequence.

To identify close neighbors of a target sequence, we performed a PSI-BLAST (Position-Specific Iterated BLAST)¹ search on the target against the NCBI NR database, with 5 iterations. We used the default e-value threshold of 10.

Once the nearest neighbors in sequence space of the target sequence have been identified, we must collect the Gene Ontology (GO) nodes associated with these sequences. To achieve this, we first obtain the Swiss-Prot identifiers annotated to each PSI-BLAST match using a parsed listing of the NR database headers. Then, using the SIB/EBI Swiss-Prot to GO mappings, we find all of the Gene Ontology nodes related to the corresponding proteins. Finally, we build a weighted collection of Gene Ontology nodes, where each node in the collection is given a weight according to the PSI-BLAST e-value. Since several near neighbors of the original target sequence may map to the same Gene Ontology nodes, the collection we build can have redundancy. In this case, each occurrence of a Gene Ontology node will be weighted individually according to its source.

This collection of weighted Gene Ontology nodes becomes the input query to a categorization technology called the Gene Ontology Categorizer (GOC)⁴. This technology aims to identify a set of nodes in the Gene Ontology which best summarize or categorize a given list of input nodes. The technology is based on

a view of bio-ontologies as combinatorially structured databases rather than facilities for logical inference, and draws on the discrete mathematics of finite partially ordered sets (posets) to develop data representations and algorithms appropriate for the Gene Ontology. Briefly (for more detail, see references 4,6), after identifying the set of input nodes in Gene Ontology space, GOC traverses the structure of the Gene Ontology, percolating hits upwards, and calculating scores for each Gene Ontology node. GOC then returns a rank-ordered list of Gene Ontology nodes representing cluster heads. In the end, this provides an assessment of which nodes best cover the input set.

We consider the set of cluster heads returned by GOC to be indicative of the function of the collection of nearest neighbors of the target sequence, and hence indicative of the function of the target sequence itself. These are returned as the predictions for the functions of the target sequence (subject to thresholding of the GOC results) and submitted to the CASP assessors.

The GOC system has many parameters that need to be specified in order to run effectively. To establish appropriate parameter settings for the CASP predictions, we created a “gold standard” test set of protein sequences for which mappings to Gene Ontology nodes were known. The test set consisted of the distinct set of Swiss-Prot sequences associated with entries in the 1.65 version of the SCOP dataset⁵ through Protein Data Bank² annotations. This set was filtered to include only those sequences that had mappings in Swiss-Prot to the Gene Ontology, resulting in 774 test sequences. We measured precision and recall results for the GO function predictions over this test set for different parameter values, making sure to eliminate a PSI-BLAST match to the original sequence itself to avoid biasing the GOC analysis. For the system used to generate the submitted results for the CASP targets, we selected the parameter values which corresponded to the best empirical balance of precision and recall over the test set.

Acknowledgements: This work was sponsored by the Department of Energy under contract W-7405-ENG-36 to the University of California. We would like to thank the Los Alamos National Laboratory Protein Function Inference Group for their contributions to this work.

1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. & Lipman,D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402.

2. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G. Bhat,T.N., Weissig,H., Shindyalov,I.N., Bourne,P.E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28 pp. 235-242.
3. The Gene Ontology Consortium (2000). Gene Ontology: Tool For the Unification of Biology, *Nature Genetics*, 25:1:25-29.
4. Joslyn,C., Mniszewski,S., Fulmer,A., Heaton,G. (2004). The Gene Ontology Categorizer. *Bioinformatics*, vol. 20, supplement 1, i169-i177.
5. Murzin,A.G., Brenner,S.E., Hubbard,T., Chothia,C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540.
6. Verspoor,K., Cohn,J., Joslyn,C., Mniszewski,S., Rechtsteiner,A., Rocha,L.M., Simas,T. (2004). Protein Annotation as Term Categorization in the Gene Ontology using Word Proximity Networks. To appear in *BMC Bioinformatics*.