# The CAFA Challenge

COMP150: Protein Bioinformatics Topics
Fall 2010, Final Project

Shilpa Nadimpalli, Lee Tien, Mengfei Cao, Duncan Renfrow-Symon

## Background

Perhaps the most pressing issue hindering progress in the field of protein biology is functionally annotating the large volume of data generated by high throughput sequencing methods. Multiple factors contribute to the difficulty in making these function predictions. First, while in theory the primary structure of a protein contains all the information necessary to fold into the functionally significant secondary and tertiary forms, it has so far proven computationally intractable to model this. The only reliable way to deduce a proteins structure is to crystallize it and then examine the structure with x-ray crystallography or NMR. This procedure is laborious and expensive, and consequently only a fraction of known proteins have had their structures solved. The lack of reliable structural information for unknown proteins makes is difficult to then predict the proteins function. Second, protein structures do not have a one-to-one mapping to function. Each individual protein has a unique combination of secondary structures and rearranging the structural units in a given protein would likely result in a complete change in function. Therefore, even when a protein is structurally defined, it may not be trivial to predict function.

Nevertheless, researchers have developed programs that attempt to functionally annotate proteins based on their sequence. These programs generally attempt to match new proteins against large databases of protein sequences with known function in order to find homologous proteins, i.e. proteins with a common ancestor. The theory is that evolutionarily related proteins will share aspects of functionality. Other programs attempt to match on specific protein domains, while still others attempt to harness the power of distributed computing to find energetically favorable protein folds. Unfortunately the widespread use of automated methods has raised new concerns about the quality of annotations. Researchers fear that automated methods will use these incorrect annotations and thus introduce pervasive errors throughout the databases.

The Critical Assessment of Function Annotations (CAFA) Challenge provides is an annual competition designed to verify the accuracy of automated prediction methods. The CAFA group negotiates with select researchers to keep their newly solved protein structures and functions confidential. Then the CAFA group releases a substantial list of these protein sequences, almost 50,000 this year, as targets for functional prediction. At the end of the challenge, participants are evaluated based on how well their methods predicted the solved protein functions kept confidential by CAFA.

CAFA is able to objectively assess prediction accuracy because the language for assigning protein function has been standardized by the Gene Ontology (GO) Consortium. Ontology terms fall into one of three categories – cellular component, biological process, and molecular function. Each of these categories contains a hierarchical tree of increasingly specific ontology terms. For the purpose of the CAFA Challenge, only biological process and molecular function GO terms are evaluated.

## Our Approach

We approached the CAFA Challenge in a two-pronged fashion. First, we used two different sequence alignment techniques (BLAST [Basic Local Alignment Search Tool] and Pfam) in order to transfer GO terms from known protein structures and functional domains that match the sequences of the unknown targets. We believe that this sequence alignment approach allows us to reliably annotate the “low hanging fruit” in the unknown protein targets, and is likely a common first method employed by most groups participating in the challenge.

Our second method of function prediction was based on identifying known structural domains in the target sequences. To do this, we employed the SMURF (Structural Motifs Using Random Fields) application, developed at Tufts, as a first step. SMURF allowed us to identify beta-propeller motifs in the unknown proteins. Beta-propellers are well known to facilitate protein-protein interaction, which result in their involvement in a wide variety of functions from signal transduction, to transcription, to apoptosis. Despite the lack of functional specificity that can be achieved by identifying beta-propellers in the unknown targets, the approach was still useful in verifying protein-

protein interaction GO terms identified by other methods. In addition, the use of SMURF establishes a framework for the incorporation of other structural motif prediction algorithms into our methodology.

## Technical Details

The main component of our function prediction system was a MySQL database, as shown in Figure 1. The database allowed us to run each of our prediction methods in parallel, storing the results independently and then aggregating them for the final output. The flow of data for each of our prediction methods is detailed below.

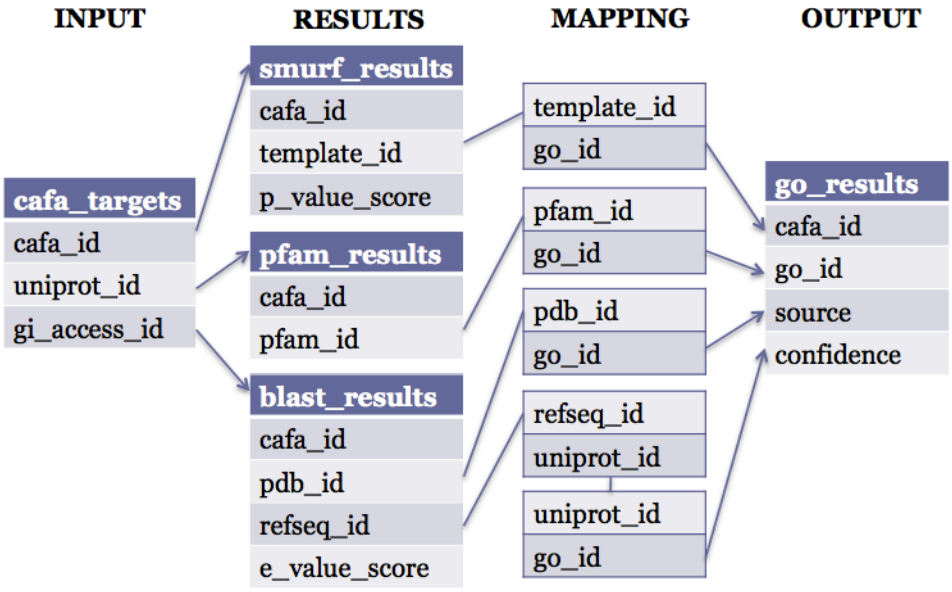


*Figure 1: Database Schema*

### BLAST

We first installed a local version of the BLAST sequence alignment tool. Due to both data and run-time constraints, we were limited to using BLAST to compare our unknown sequences to only the proteins in the Protein Databank (PDB) database, rather than the non-redundant (nr) database, which contains many more sequences. The upside to this is that GO terms transferred from matches in the PDB are likely to be much more high-quality and reliable since they are coming from proteins with solved structures.

Unknown FASTA sequences were fed to the BLAST program in batches of 100, and the results were then parsed and uploaded to the blast_results table in the database. Sequence hits were defined as having an e-value <= $10^{-5}$ and a maximum of 6 of these hits were added to the table. To produce the final GO term predictions, the BLAST results were joined with a mapping table that was pre-populated with data matching PDB IDs to GO terms. We also included a mapping table for RefSeqIDs to GO terms, which would allow for the use of the nr database with BLAST in the future.

### Pfam

We also installed a local version of Pfam. Pfam relies on the HMMER3 to generate hidden Markov models of protein domain families that are then used to find these family domains in unknown sequences. Similar to BLAST, the Pfam results were parsed and uploaded to the pfam_results table, and then joined with a pre-populated mapping table to assign final GO predictions.

### SMURF

Finally, we obtained and ran a local copy of the SMURF application to identify beta-propeller

structures. We compared our unknown sequences to 12 different propeller templates (6-bladed, 7-bladed, 8-bladed, and all permutation of double-bladed propellers (i.e. 6-6 bladed, 6-7 bladed, etc)). We defined a positive hit on a template when p <= $10^{-4}$. As with the other methods, the results were parsed and uploaded to smurf_results.

Unfortunately, due to the wide range of functions for proteins containing beta-propellers, we were not able to reliably populate the mapping table linking beta-propellers to specific GO functional terms. At a high level, we know that these proteins are involved in protein-protein interaction, but a detailed literature search did not provide a consensus for more specific propeller function. If time had allowed, we may have run a training set of propellers proteins with known functions through SMURF in order to build a probabilistic model for likely functions of a given beta-propeller fold.

**Results**

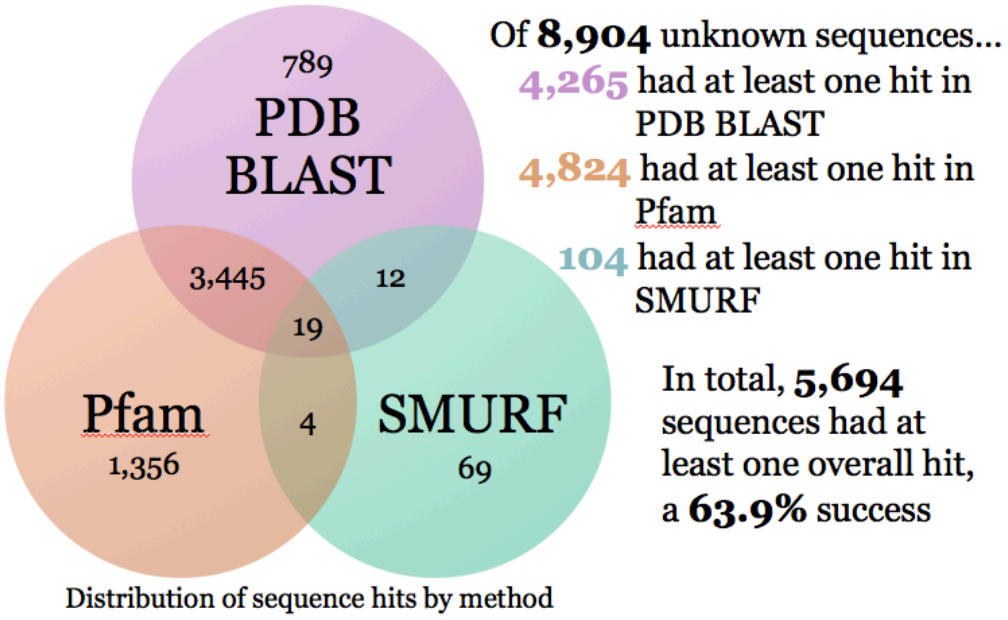


*Figure 2: Summary of results*

The results for one eukaryotic target file and one prokaryotic target file are detailed in Figure 2. In many cases, the BLAST and Pfam prediction techniques produced redundant GO terms, reinforcing the likelihood that these are correct functional predictions. In total, 5,694 unique sequences had a hit in at least one of our methods, giving a 63.9% success rate of transferring at least one ontology term.

The complementarity of our approach can be demonstrated by examining specific target proteins. For example, the eukaryotic target sequence, T38114, matched 6 PDB structures in BLAST resulting in the transfer of GO terms specific for transcription, mitosis, methylation and protein binding. The same sequence matched a Pfam family that was associated with GO terms for zinc ion binding, and nucleic acid binding. The zinc binding term was a new addition to BLAST, while the nucleic acid binding provided a more specific term for the transcription and mitosis terms identified with BLAST. Finally, the protein was also identified as containing a 7-bladed propeller in SMURF, potentially leading to further functional annotation.

**Conclusion**

It is manifestly clear that our results so far, while significant, would not be a competitive entry into the CAFA challenge. To move beyond the realm of “easy” predictions we would need to employ several more tools, including those that searched for more functionally specific structural domains or motifs. Several programs immediately present themselves. Beta wrap pro, which was developed at MIT,

searches for beta barrels which have been linked to infectious disease. Raptor, which is one of the few proprietary tools, uses protein threading and performs very well when no homologous models are available. Finally, Rosetta is an ab initio program which uses distributed computing to attempt to find the minimum energy fold state of a protein. Using these programs in addition to our prior work would help fill annotations of more “novel” or “harder” protein sequences, which is key to performing well in the challenge.

**Work Consulted**


[1] Stephen F. Altschul et al., “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs,” Nucleic Acids Research 25, no. 17 (1997): 3389 -3402.

[2] Ana Conesa et al., “Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research,” Bioinformatics (Oxford, England) 21, no. 18 (September 15, 2005): 3674-3676.

[3] R. D. Finn, “Pfam: clans, web tools and services,” Nucleic Acids Research 34, no. 90001 (1, 2006): D247-D251.

[4] Laurent Falquet et al., “The PROSITE database, its status in 2002,” Nucleic Acids Research 30, no. 1 (January 1, 2002): 235-238.

[5] Matt Menke, Bonnie Berger, and Lenore Cowen, “Markov random fields reveal an N-terminal double beta-propeller motif as part of a bacterial hybrid two-component sensor system,” Proceedings of the National Academy of Sciences 107, no. 9 (March 2, 2010): 4069 -4074.

[6] Alexey G. Murzin et al., “SCOP: A structural classification of proteins database for the investigation of sequences and structures,” Journal of Molecular Biology 247, no. 4 (April 7, 1995): 536-540.

[7] A Krogh et al., “Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes,” Journal of Molecular Biology 305, no. 3 (January 19, 2001): 567-580.

[8] M. Menke, B. Berger and L. Cowen, "Markov random fields reveal an N-terminal double beta-propeller motif as part of a bacterial hybrid two-component sensor system" PNAS March 2, 2010 vol. 107 no. 9 4069-4074.