Integration of Structural and Sequence Information for Homology-Based Modeling of Proteins

By

Michael Gribskov

Prepared under
Contract number DE-FG03-96ER62264
For the U. S. Department of Energy - Oakland

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Integration of Structural and Sequence Information for Homology-Based Modeling of Proteins

By

Michael Gribskov

Prepared under
Contract number DE-FG03-96ER62264
For the U. S. Department of Energy - Oakland

Project Number 03752.241.140

Final Report Integration of Structural and Sequence Information for Homology Modeling of Proteins (DE-FG03 96 ER 62264)

**Michael Gribskov
San Diego Supercomputer Center**

In the six month supplement to our project, "Combining sequence and structural information for homology modeling of proteins," we continued developing the approaches defined in the original proposal (evolutionary profile), and the family pairwise search (FPS) method for defining sequence patterns. This work resulted in several fundamental publications regarding methods for making statistical evaluations of sequence match scores (Bailey and Gribskov, 1998, 1999, 2002), and a submitted manuscript validating the evolutionary profile approach (Veretnik and Gribskov, submitted).

Homology modeling allows one to predict the three-dimensional structure of a novel *query* sequence based on the known three-dimensional structure of a homologous *template* sequence. This approach requires that one find a three-dimensional structure that is homologous to the query sequence. The query sequence must then be mapped onto the template sequence such that each amino acid residue appears in the correct position in the modeled structure. Both the identification of homologous sequences and the mapping of the query onto the template structure are achieved through sequence comparisons.

Standard sequence comparison methods use a single amino acid residue comparison matrix, such as BLOSUM 62 or PAM250, to identify proteins that are more similar than average. When the similarity is high enough, one can infer that these proteins are homologous and thus likely to have similar structure and function.

However, it is well known that amino acid residues in proteins show strong position specific differences in conservation depending on their position in the three-dimensional structure. These position specific differences can be observed in multiple sequence alignments and form the basis of profile (Gribskov and Veretnik, 1996) and hidden Markov model (e.g. Durbin et al., 1998) methods. In the case of the evolutionary profile, an evolutionary model is estimated at every aligned column in a multiple alignment. This model provides additional information that is used to calculate a profile with higher sensitivity and specificity. More importantly, the evolutionary profile is able to generalize, from a set of closely related sequences, a profile that reflects a more evolutionarily diverse set of related proteins (submitted). This procedure is quite useful and we term evolutionary projection.

The family pairwise search (FPS) method (Bailey and Grundy, 1999) is a new development that is useful for finding sequences that belong to a homologous family. Rather than relying on position specific scoring matrices, FPS uses traditional pairwise comparisons to achieve a similar effect. In FPS, each sequence of interest is compared to a *panel* of sequences that define a homologous family. The P-values for matching to

each of the panel sequences are combined, taking into account the similarity between the panel sequences themselves, to calculate an overall P-value for matching to the entire family. This combination of several quasi-independent measures of a proteins membership in the family greatly improves the sensitivity and accuracy of family assignment. In the six month supplementary period, we will continue to improve the publicly available FPS server (<http://fps.sdsc.edu>) to allow one to identify a broader spectrum of sequences, and to assign them to structural families. The FPS approach was the basis for the *ab initio* classification of over 1000 protein kinases identified in the *Arabidopsis thaliana* genomic sequence (see the PlantsP website).

- Durbin, R., Eddy S., Krogh, A., Mitchison G., "Biological sequence analysis: probabilistic models of proteins and nucleic acids", Cambridge University Press, 1998.
- Bailey, T.L. and Grundy, W.N., "Classifying proteins by family using the product of correlated p-values", in *Third International Conference On Computational Molecular Biology (RECOMB99)*, Association for Computing Machinery, New York, pp. 10-14, 1999.
- Bailey, T.L. and Gribskov, M., "Combining evidence using p-values: Application to sequence homology searches", *Bioinformatics*, 14, 48-54, 1998..
- Bailey, T.L. and Gribskov, M., "Methods and statistics for combining motif match scores", *Journal of Computational Biology*, 5, 211-221, 1998.
- Bailey, T.L.; and Gribskov, M. "Estimating and evaluating the statistics of gapped local-alignment scores." *Journal of Computational Biology*, 9, 575-593, 2002.
- Family Pairwise Search Website, <http://fps.sdsc.edu>.
- Gribskov, M. and Veretnik, S., "Identification of sequence patterns with profile analysis", *Methods in Enzymology* 266, 198-212, 1996.
- PlantsP Website, <http://plantsp.sdsc.edu>.
- Veretnik, S, and Gribskov, M., "Evolutionary Profiles as a Temporal Model of Protein Family Evolution", submitted.