Automated structure modeling with Eidogen's suite of algorithms

A. Poleksic, J.F. Danzer and D.A. Debe

Eidogen-Sertanty, Inc.

apoleksic@eidogen-sertanty.com

STRUCTFAST (Structure Realization Utilizing Cogent Tips From Aligned Structural Templates) is a novel profile-profile alignment algorithm uniquely capable of incorporating important information from a structural family directly into the dynamic programming process. Query sequence profiles are generated using a modified version of NCBI's PSI-BLAST algorithm [1]. A database of profiles for representatives from the PDB are generated in a similar manner, but are augmented with information from structure based alignments for the structural family. Each query sequence is then aligned and scored against the library of structural profiles. Our profile-profile scoring function is based solely on probability theory and therefore contains no parameters to optimize (e.g. score matrix normalization parameters, zero-shift, etc.). Moreover, the basic principle of local alignment algorithms, the fact that the expected residue-pair score must be negative in order for the algorithm to stay in the local regime, is implicitly taken care of by the analytical nature of the scoring function. The gap penalties are position specific and are based on information from the template's structural family as well as on the distribution of residue-pair scores.

Statistical significance of alignment scores are assessed using a variant [2] of the island statistics method [3,4], so that the final E-value for every database hit accounts for the lengths and compositions of the sequences being compared. This is an important component of our method, since computing statistical significance of the alignment scores by doing extensive sequence shuffles for every database hit would be computationally prohibitive. The method we propose is able to recognize the lack of sequence similarity for any pair of sequences early in the shuffling process and thus save on the search time. Any given sequence will typically have a small number of significant hits in a large representative database, so the vast percentage of comparisons will be computed very efficiently. On the other hand, if a pair of sequences are related, our method approaches the complexity of a brute force method of doing extensive random shuffles, and thus is able to recognize and precisely estimate the statistics of the pair.

The core alignment algorithm in all three of our automated servers is the same. SFST outputs the alignment with the overall best E-value. BNMX and EXPM go a step further to refine alpha carbon coordinates by using multiple PDB templates and the remaining backbone atoms are reconstructed from the alpha carbon coordinates [5]. The main difference between BNMX and EXPM is in the choice of the null model for protein sequence families. BNMX assumes a fixed background model, whereas EXPM uses the

idea similar to that of Dirichlet mixtures [6] to estimate the log-odd score for a given pair of profile columns. Another difference is in the choice of several algorithms' parameters, such as the score significance cutoffs and gap penalties.

To evaluate the performance of our methods, we have used standard model quality measures (e.g. those in LiveBench [7]) as well as ones developed in-house. One way to assess the algorithm's ability to generate accurate 3-dimensional structures is to evaluate its success rate in determining intact and accurate structural models for small molecule binding sites in low-homology test sets. Binding sites typically involve multiple regions of the protein coming together in space and so their accuracy has been observed to be sensitive to even small alignment errors. Furthermore, small molecule binding sites contain the critical information required for drug design making their accuracy particularly important. In order to assess binding site accuracy, test sets composed of co-crystal structures are particularly useful, since the co-crystal data allows for direct definition of a site and also measurement of its resolution. We have compiled different sets of experimentally determined co-crystal structures with various degrees of sequence homology to other proteins deposited in the PDB and used the accuracy of the predicted coordinates in the active sites as a guiding principle in assessing the quality of the models produced by our methods.

REFERENCES:

1. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., 25, 3389-3.

2. Poleksic,A., Hambly,K., Danzer.J.F., Debe DA. Increased remote homology detection performance using a fast method for determining local alignment statistics, *Bioinformatics*. Advance Access published on April 7, 2005.

3. Olsen,R., Bundschuh,R. And Hwa,T. (1999) Rapid assessment of extremal statistics for gapped local alignment. In Lengauer,T., Schneider,R., Bork,P., Brutlag,D., Glasgow,J., Mewes,H.-W. And Zimmer,R. (eds), *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, Menlo Park, CA, pp. 211-222.

4. Altschul,S.F., Bundschuh,R., Olsen,R., Hwa,T. (2001) The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Res*. 29, 351-61.

5. Milik,M., Kolinski,A., Skolnick,J. (1997) An algorithm for rapid reconstruction of protein backbone from alpha carbon coordinates. *J. Comput. Chem*., 18, 80-85.

6. Sjolander,K., Karplus,K., Brown,M., Hughey,R., KroghA., Mian,I.S., Haussler,D. (1996) Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. Comput Appl Biosci., 12, 327-45.

7. Rychlewski,L., Fischer,D., Elofsson,A. (2003) LiveBench-6: large-scale automated evaluation of protein structure prediction servers. Proteins., 53 Suppl 6, 542-7.