Use of Limited Suboptimal Alignment in Homology Modeling

Christopher L. Tang[1], Donald S. Petrey[1], Marc Fasnacht[1], Mickey Kosloff[1], Emil Alexov[1,2], Barry Honig[1,2]

1   Introduction.

Improving the accuracy of alignment between a query sequence and a template protein of known structure remains a significant challenge in homology modeling [1]. Although maximal alignments perform sufficiently well to detect similarity relationships between proteins in database search applications, often the correct alignment from a structural standpoint deviates from the maximal alignment. A possible reason for this is that linear alignment schemes may not contain sufficient information to properly align a query sequence onto a template – tertiary information must be evaluated, for instance, by using physical energy terms to discriminate between native and non-native alignments [2]. Such a strategy requires a method for sampling and generating suboptimal alignments – alignments which are not maximal. In recent years, "consensus methods", which look at sequence alignments from programs written by many different laboratories, have been proposed as one solution to sampling alignments (e.g. [3]). However, such methods may not sample sufficiently and may miss the correct alignment if the alignment is "counterintuitive" to the alignment programs [4]. An alternative is near optimal alignment, which may be used to generate alignments whose scores are within a preset distance of the maximal [5]. However, the space of near optimal alignments is very large and often inconsequential differences in alignment will be reported [1]. Other variations of alignment sampling methods include iterative masking [6], [7] and parametric sampling [8]. However, these do not necessarily guarantee k-best solutions for any given set of sequences, and there is no unique set of parameters.

True k-best algorithms have historically been avoided for their computational expense [4] but in the advent of increasing computer power it is possible to revisit these. In addition, it is possible to supply detailed limits over what regions of a protein should be sampled thoroughly by suboptimal alignment; for instance, it may be desired to concentrate on how secondary structure elements should be aligned between query and template. Here we present a method for generating limited k-best suboptimal alignments, which allows us to sample suboptimal alignments deeply in regions of interest and hence more meaningfully for homology modeling. We build models for these alignments and seek to determine whether we can find closer-to-native alignments in our pool of suboptimal alignments.

References

[3] Ginalski, K., Rychlewski, L. 2003. Protein structure prediction of CASP5 comparative modeling and fold recognition targets using consensus alignment approach and 3D assessment. *Protein Structure, Function &Genetics*. 53:410-417.

[1] Jaroszewski, L., Li, W. and Godzik, A. 2002. In search for more accurate alignments in the twilight zone. *Protein Science* 11:1702-1713.

[2] Petrey, D.S. and Honig, B. 2000. Free energy determinants of tertiary structure and the evaluation of protein models. *Protein Science*. 9:2181-2191.

[4] Petrey D.S., Xiang, Z., Tang, C.L., Xie, L., Gimpelev, M., Mitros, T., Soto, C.S., Goldsmith-Fischman, S., Kernytsky, A., Schlessinger, A., Koh, I., Alexov, A. and Honig, B. 2003. Using multiple structure alignments, fast homology building, and energetic analysis in fold recognition and homology modeling. *Protein Structure, Function &Genetics*. 53:430-435.

[6] Saqi, M.A. and Sternberg, M.J. 1991. A simple method to generate non-trivial alternate alignments of protein sequences. *Journal of Molecular Biology* 219:727-732.

[5] Waterman, M.S. 1983. Sequence alignments in the neighborhood of the optimum with general application to dynamic programming. *Proceedings of the National Academy of Sciences USA*. 80:3123-3124.

[7] Waterman, M.S. and Eggert, M. 1987. A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *Journal of Molecular Biology*. 197:723-728.

[8] Waterman, M.S., Eggert, M. and Lander, E. 1992. Parametric sequence comparisons. *Proceedings of the National Academy of Sciences USA*. 89: 6090-6093.