The SAM-T99 Protein-search Method Works Well as a Multiple Aligner

Kevin Karplus and Birong Hu
University of California, Santa Cruz
`http://www.cse.ucsc.edu/research/compbio/HMM-apps.html/T99-query.html`

31 May 2000

The SAM-T99 method was developed as a way to find similar proteins given a single sequence or a small seed alignment. It is an evolutionary improvement over SAM-T98, which has done very well in superfamily classification tests [3, 2]. Results showing the improvement in superfamily recognition will be included on the poster.

Because the SAM-T99 method generates a multiple alignment of the sequence it finds, we decided to evaluate the method as a multiple aligner, using the BAliBASE multiple-alignment test suite [4, 5], especially since other researches have questioned the quality of alignments done by hidden Markov models [1].

The initial tests used the previously untested `-tuneup` option of the script, which turns off the search of the protein database, does an initial HMM training to build the seed alignment from unaligned sequences, and uses the unaligned sequences as the set to search for homologs. This technique aligned 124 of the 141 alignments, but dropped one or more sequences in the other 17 cases, since the sequences too dissimilar from the others for SAM-T99 to recognize them as being in the same family.

The 141 alignments that were created were then used to build hidden Markov models that the sequences were forcibly aligned to, to get multiple alignments for the entire test set. These alignments were compared to the BAliBASE reference alignments, and the resulting scores compared with published results [5]. In these tests, SAM-T99(tuneup) seems comparable to the other multiple aligners such as Clustal and PRPP (much better on reference 2, slightly worse on reference 1v1, comparable on the others).

The quality of the SAM-T99 multiple alignments seems to be high enough that little or no benefit would be obtained from realigning them using a different multiple alignment tool.

References

[1] O. Gotoh. Multiple sequence alignment: algorithms and applications. *Advances in Biophysics*, 36(1):159–206, 1999.

[2] Kevin Karplus, Christian Barrett, and Richard Hughey. Hidden markov models for detecting remote protein homologies. *Bioinformatics*, 14(10):846–856, 1998.

[3] J. Park, K. Karplus, C. Barrett, R. Hughey, D. Haussler, T. Hubbard, and C. Chothia. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *JMB*, 284(4):1201–1210, 1998. Paper available at `http://www.mrc-lmb.cam.ac.uk/genomes/jong/assess_paper/assess_paperNov.html`.

[4] Julie D. Thompson, Frederick Plewniak, and Oliver Poch. BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, 15(1):87–8, 1999.

[5] Julie D. Thompson, Frederick Plewniak, and Oliver Poch. A comprehensive comparison of multiple sequence alignment programs. *NAR*, 27(13):2682–90, 1999. Additional detailed results at `http://www-igbmc.u-strasbg.fr/BioInfo/BAliBASE/prog_scores.html`.