Optimal Hidden Markov Models for All Sequences of Known Structure

Julian Gough[1] `jgough@mrc-lmb.cam.ac.uk`
Cyrus Chothia[1] `chc1@mrc-lmb.cam.ac.uk`
Kevin Karplus[2] `karplus@cse.ucsc.edu`
Christian Barrett[2] `cbarrett@cse.ucsc.edu`
Richard Hughey[2] `rph@cse.ucsc.edu`

[1] Structural Studies Division, MRC Laboratory of Molecular Biology, Hills rd, Cambridge, CB2 2QH, England
[2] Baskin Center for Computer Engineering and Science, University of California, Santa Cruz, CA 95064, USA

Keywords: hidden Markov model, sequence analysis, protein sequence homology

1 Introduction

Hidden Markov Models (HMMs) are probably the most powerful tool for the detection of protein sequence homology [4]. Maximization of their capabilities and biological usefulness requires the correct interpretation of their scores, and sufficient coverage of the sequence variations that exist in different protein families. Using information available from the SCOP database we investigated optimal measures for these two aspects of HMMs [5].

The SCOP database is a hierarchical classification of the domains that are found in all proteins of known structure. Those evolutionary relationships of proteins which cannot be detected from sequence can usually be inferred from structure. These are listed in the SCOP database where superfamilies contain sets of proteins with evolutionary relationships derived from both sequence and structure.

Searches performed using a set of HMMs produced for each member of a superfamily of homologous sequences worked more effectively than a single HMM built from an alignment of those sequences.

Using the SAM-Target99 procedure (see below), HMM models were made for all PDB95 sequences, that is the 4591 sequences of proteins of known structure which have sequence identities of 95% or less. These sequences in the context of their SCOP classification form the basis of the results described below.

2 Method

The SAM-Target99 method which is described in detail by the authors [3], takes as input a single sequence and searches the non-redundant sequence database (nrdb90) [2] using WU-BLASTP to produce a large set of potential homologues and a small set of close homologues. The small set is scored against an HMM built on the single starting sequence, and all but low scoring sequences aligned to it to provide the initial alignment for the first iteration which is carried out on the large set. Four iterations are used of a selection, training, and alignment procedure. Each iteration begins with an alignment from which it creates: a model, and a large set of sequences found from a WU-BLASTP search of nrdb90 which becomes broader with each iteration producing progressively larger sets. The model is searched against the set, and the additional sequences found are added to the growing alignment.

The models which comprise the library are built from the final alignments and associated sequence weights produced by the SAM-Target99 procedure.

3 Results

A complete model library was built from the sequence domain database of less than 95% identity [1] comprising of 4591 models. These models were then scored against the 21828 sequences of protein domains in SCOP, i.e. all available sequence domains of known structure including those with over 95% identity. The models themselves and the scores they made in matching the full set of PDB sequences were then examined:

1. For all HMMs in each superfamily we determined the extent to which the sequences and weights from which they were built are the same.

2. We determined the extent to which different models from each superfamily detected different homologues.

3. We determined the scores each model made for correct and incorrect hits.

From the data available from 1. and 2. we were able to determine the redundancy of the models produced from PDB95, that is how different the starting sequences have to be before they create models that will detect different homologues. The information available from 3. provides the best overall score to use for the detection of homologues at different error rates and also the extent to which these scores differ for models produced for different superfamilies.

On the basis of these results a complete and searchable library of superfamily profiles can be created. Profiles consist of several redundancy-filtered hidden Markov models each with tailored homology decision parameters. Although structural information is not explicitly imposed on the multiple model profiles, it is included by their grouping and assessment according to the SCOP classification. All of the models are subject to ongoing validation, and although they are initially produced by the same automatic procedure they are open to individual or group revision where it is found that improvements can be made.

We intend to make the library publicly available and updated with new releases of SCOP.

References

[1] Brenner, S.E., Koehl, P., and Levitt, M., The ASTRAL compendium for protein structure and sequence analysis, *Nucleic Acids Res.*, 28(1):254–256, 2000.

[2] Holm, L., and Sander, C., Removing near-neighbour redundancy from large protein sequence collections, *Bioinformatics*, 14(5):423–429, 1998.

[3] Karplus, K., Barrett, C., and Hughey, R., Hidden Markov models for detecting remote protein homologies, *Bioinformatics*, 14(10):846–856, 1998.

[4] Krogh, A., Brown, M., Mian, I.S., Sjölander, K., and Haussler, D., Hidden Markov models in computational biology applications to protein modeling, *J. Mol. Biol.*, 235(5016):1501–1531, 1994.

[5] Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C., SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.*, 247(4):536–540, 1995.