Protein Superfamily Building Methods Comparison

Kunbin Qu[1]
kqu@rigel.com
Jun Zhu[2]
junz@amgen.com
Scott Pegg[3]
spegg@mako.cgl.ucsf.edu
Patricia Babbitt[3]
babbitt@cgl.ucsf.edu

[1] Department of Genomics and Target Discovery, Rigel, Inc., South San Francisco, CA 94080, USA
[2] Department of Computational Biology, Amgen, Inc., Thousand Oaks, CA, 91320, USA
[3] Departments of Biopharmaceutical Science and Pharmaceutical Chemistry, University of California at San Francisco, CA 94143, USA

Keywords: protein family classification, database search, homology search

1 Introduction

With the enormous amount of sequence data generated from the genome projects, function prediction and sequence annotation become critical for understanding and effectively utilizing the information potentially available. Classifying unknown sequences into appropriate, well curated families plays an important role in fulfilling such a task. Here, we compare several protein superfamily building algorithms and propose a new strategy to improve the current methods.

2 Methods and Results

The following methods were compared: 1) straight-forward Hidden Markov Model (HMM) package (HMMER2.1 from WushU), 2) iterative HMM package (SAM3.0 from UCSC), 3) Gibbs sampling method with database search (PROBE, Neuwald *et al.*, 1997), 4) PSI-blast (Altshul *et al.*, 1997), 5) BayesianProfiler (Zhu *et al.*, 1999), 6) congruence analysis by Shotgun (Pegg & Babbitt, 1999). The methods were tested on sequences from several superfamilies representing different fold classes. Each superfamily includes very divergent sequences, including, in some cases, members whose homologies to each other are difficult to find without comparison of three-dimensional structures. The hit list for each of the methods was generated by choosing a reasonable cutoff based on E-values (E = 0.001 for most of the cases), or Bayesian evidence for BayesianProfiler. True positives were verified by examination of the hit list by an expert in the biology of each superfamily and/or as reported in the published literature. Because searching a large database such as the non-redundant protein database at NCBI (over 400,000 sequences) is very time consuming for HMM and BayesianProfiler (ranges from 8 hours to one full day), the search space was limited to a much smaller set as suggested by the current SAM distribution. This set was generated from Blast analysis of a seed sequence using a high cutoff E-value (= 300). However, we found that by limiting the search space in this fashion, HMM and BayesianProfiler performed much worse than PSI-blast and PROBE. The performance of these two approaches can be greatly improved, however, by correctly expanding their respective search spaces. From previous experience, Shotgun appears to offer a valuable method to expand and recruit such limited search sets. We propose to combine HMM and Shotgun in the future to achieve better performance in both speed and sensitivity.

The following table lists the number of true positive (TP) hits for the tested methods for one of the superfamilies investigated, the enolase superfamily (Babbitt *et al.*, 1996), of the alpha/beta barrel fold class. Preliminary analysis of the number of putative false positives found by each method shows considerable variability with each approach. More rigorous analysis will be required, however, to determine the true false positive rate for each method.

	Shotgun	PROBE	PSI-blast	HMMER	SAM	Bayes
TP	184	153	172	50^a, 152^b	55^a, 179^b	55^a, 172^c

The seed sequence is the enolase superfamily domain of P31458 (gi: 2851654) as described by Babbitt *et al* (1995). The search space is denoted by superscripts. ^{a} : search space is the set created by Blasting (BLASTP 2.0a WashU, 1998) the seed sequence against a static version of the NR protein database as of March 11, 1999 with cutoff of E = 300 (this is the default setting in the SAM3.0 distribution). ^{b} search space is the whole static NR. ^{c} search space is 184 true positives plus 1943 random sequences from the static NR. The 184 true positives were originally generated as reported in Babbitt *et al.* (1996), and confirmed using Shotgun as reported in Pegg & Babbitt. Examination of the hits found by the other approaches evaluated in this study found a few additional true positives, and we expect that additional sequence analysis or structural characterization of proteins in the NR database may eventually reveal other superfamily members missed by these methods. Most of the additional true positives found by the methods we tested are fragments of the enolase protein itself, however.

Similar tests have been performed for other superfamilies representing other fold classes and additional analysis of these superfamilies suggests that a critical parameter in building divergent superfamilies is the generation of a sufficiently divergent starting set, which is dependent, in turn, on the seed sequence with which the analysis is started. This result implies that giving more attention to correctly weighting starting sequences for use in each model through appropriate sub-classification methods tuned to each superfamily is likely to result in better performance.

References

[1] Altschul, S.F., Madden, T.L., Schffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, 25(17):3389–3402, 1997.

[2] Babbitt, P.C., Mrachko, G. T. *et al.*, A functionally diverse enzyme superfamily that abstracts the a-protons of carboxylic acids, *Science*, 267: 1159–1161, 1995

[3] Babbitt, P. C., M. Hasson *et al.*, The enolase superfamily: A general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids, *Biochem.*, 35:16489–16501, 1996.

[4] HMMER: hmmer.wustl.edu

[5] Neuwald, A.F., Liu, J.S., Lipman, D.J., and Lawrence, C.E., Extracting protein alignment models from the sequence database, *Nucleic Acids Res.*, 25(9):1665–1677, 1997

[6] Pegg, S.C. and Babbitt, P.C., Shotgun: getting more from sequence similarity searches, *Bioinformatics*, 15(9):729–740, 1999.

[7] SAM3.0: www.cse.ucsc.edu/research/compbio/sam.html

[8] Zhu, J., Luthy, R., and Lawrence, C.E., Database search based on Bayesian alignment, *ISMB*, 7:297–305, 1999.