Comparison of Profile and Hidden Markov Model Methods for Detecting Degenerate Motifs in Protein Sequences

Krzysztof M. Rajkowski rajkow@kb.inserm.fr
Etienne-Emile Baulieu baulieu@kb.inserm.fr
Michael Schumacher schuma@kb.inserm.fr
INSERM Unité 488, Hôpital de Bicêtre, 80 rue du Général Leclerc,
F-94276 Le Kremlin-Bicêtre, France

Keywords: profile, hidden Markov model, tetratricopeptide

1 Introduction

Large scale protein sequence searches for motifs which might indicate protein identities and/or functions has become important with the large number of unknown proteins revealed by genome sequencing. The primary criterion for using a given search method is that of efficacy - minimization of the number of false negative 'misses' and false positive 'hits'. Criteria of simplicity and rapidity of use are important but secondary. For highly degenerate motifs the only search methods widely employed involve the use of Profile analysis [2] or Hidden Markov Models (HMMs) [3]. The present work compares the efficacy of the two methods using the example of searches for the degenerate 34-residue tetratricopeptide repeat (TPR) motif [4]. Results obtained with a Profile method, described below, were compared with those given in the Pfam4.4 database, obtained using an HMM [1].

2 Method and Results

For the Profile method a 20×34 Bayesian probability matrix P_{ij} was constructed where each element was $\log(f_{ij}/q_i)$ [5], f_{ij} being the relative frequency of occurence of each amino acid i at each of the 34 positions j in 462 known TPR sequences from 70 proteins and q_i was the relative frequency of occurence of each amino acid i in the SWISSPROT37 database (excluding N-terminal methionines). (Relative frequencies could be equated with probabilities because of the large numbers of sequences used). Sequences were searched using a 34-residue sliding window with each window scored as $\sum_{j=1}^{34} P_{ij}$ for the corresponding amino acids. Scores were converted to expectations that they would be obtained by chance using parameters obtained by scoring 100,000 pseudo-random 34-residue sequences in which, on average overall, each amino acid appeared with frequency q_i. A protein was considered to be a TPR protein if it contained at least 2 non-overlapping sequences each with expectation $e < 10^{-5}$ that its score would not be obtained by chance *in a protein of that length* (equivalent to $e < 2.10^{-7}$ for a protein of 250 residues). The same sequence database (SWISSPROT37 plus SP-TREMBL9, obtainable as the file pfamseq.gz from ftp.sanger.ac.uk; information kindly provided by Dr. A. Bateman, Sanger Institute, UK) was searched as that for the Pfam4.4 HMM search.

The Profile search detected all but 4 of the 173 non-redundant TPR proteins found by the Pfam HMM method, though 3 of these proteins only contain 1 TPR motif and thus would not be found given the Profile method's cutoff criterion of a minimum of 2 TPRs (they are tetratricopeptide *repeat* motifs). The other protein (nucleoporin nup358, not previously reported as a TPR protein) may thus be a false negative for the Profile search. On the other hand, the Profile search detected 74 proteins that are known to contain TPRs (because they are named as such, they have been reported as such in the literature or in the annotations in their database sequence entries or because they are ortho/paralogs of known TPR proteins) not detected by the Pfam HMM, and are thus false negatives for the HMM method. The Profile search also detected a further set of 90 sequences not so far known to contain TPRs (20 of these were ortho/paralogs of others in the set). Most of these sequences were for hypothetical proteins and further investigation of their homologies are required to determine whether or not they are true positives. Nevertheless, false negatives are more troublesome than false positives because, by definition, they remain unseen following a search and thus cannot be checked later. The question is thus posed whether the 74 true positives found by the Profile but not by the HMM is due to the large difference between the number of sequences used to construct the Profile (462) versus those to seed the HMM (29), to differences in cutoff scores or to the methods themselves. One advantage, however, of seeding with a large number of sequences is that if some have wrongly been reported to be, in this case, TPRs then their effects on the scoring matrix are diluted out.

3 Conclusions

The Profile method does have the advantage that all the calculations are transparent, as opposed to the HMM's 'black box' approach, but the HMM method is clearly superior in terms of simplicity and rapidity of parametising a search - it trains itself once the seed is provided. It remains that, for degenerate motifs, the Profile method compares extremely favourably with the HMM method in terms of the primary criterion, efficacy, according to this example. It is thus suggested that, until conclusive evidence is obtained, Profile searches should at least accompany HMM searches for such degenerate motifs.

References

[1] Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Finn, R.D., and Sonnhammer, E.L.L., Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins, *Nucleic Acids Res.* 27(1): 260–262, 1999.

[2] Gribskov, M., Lüthy, R., and Eisenberg, D., Profile analysis: Detection of distantly related proteins. Identification of common molecular subsequences, *Proc. Natl. Acad. Sci. USA*, 84(13):4355–4358, 1987.

[3] Krogh, A., Brown, M., Mian, I.S., Sjolander, K., and Haussler, D., Hidden Markov models in computational biology: Applications to protein modelling, *J. Mol. Biol.*, 235(5):1501–1531, 1994.

[4] Sikorski, R.S., Michaud, W.A., Wootton, J.C., Boguski, M.S., Connelly, C., and Heiter, P., TPR proteins as essential components of the yeast cell cycle, *Cold Spring Harbour Symp. Quant. Biol.*, 56:663–673, 1991.

[5] Tatusov, R.L., Altschul, S.F., and Koonin, E.V., Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks, *Proc. Natl. Acad. Sci. USA*, 91(25):12091–12095, 1994.