



## A profile hidden Markov model for signal peptides generated by HMMER

Zemin Zhang\* and William I. Wood

Department of Bioinformatics, Genentech Inc., South San Francisco, CA 94080, USA

Received on June 19, 2002; revised on August 15, 2002; accepted on August 16, 2002

### ABSTRACT

**Summary:** Although the HMMER package is widely used to produce profile hidden Markov models (profile HMMs) for protein domains, it has been difficult to create a profile HMM for signal peptides. Here we describe an approach for building a complex model of eukaryotic signal peptides by the standard HMMER package. Signal peptide prediction with this model gives a 95.6% sensitivity and 95.7% specificity.

**Availability:** The profile HMM for signal peptides, data sets, and the scripts for analyzing data are available for non-commercial use at <http://share.gene.com/>.

**Contact:** [zemin@gene.com](mailto:zemin@gene.com)

### INTRODUCTION

HMMER is an implementation of profile HMM methods for sensitive database searches using multiple sequence alignment profiles as queries (Eddy, 1998). A wide collection of protein domain models has been generated using the HMMER package and these models have largely comprised the Pfam protein family database (Bateman *et al.*, 2002) widely used for protein domain detection and function prediction. However, the application of HMMER has not been successfully extended to creating more complicated models as is required for secretion signal peptides. Despite the importance of predicting protein signal peptides (Nakai, 2000), a signal peptide model is noticeably absent from the Pfam database (Pfam 7.4, July 2002). The major difficulties in constructing such a model include great variation in length and a lack of obviously conserved residues in signal peptides, except  $-1$  and  $-3$  positions relative to the cleavage site (von Heijne, 1983, 1985). The few commonly available prediction servers for signal peptides, including SignalP Server V1.1 and V2.0, are based on neural networks and/or hidden Markov models (Nielsen *et al.*, 1997; Nielsen and Krogh, 1998); however, the specialized tools for reproducing those models are not generally available so it can be difficult to update or modify the prediction methods for most molecular biologists. Here, we describe up-to-date data

sets of signal peptides and the application of the HMMER package in producing a reliable prediction method for signal peptides.

### METHODS

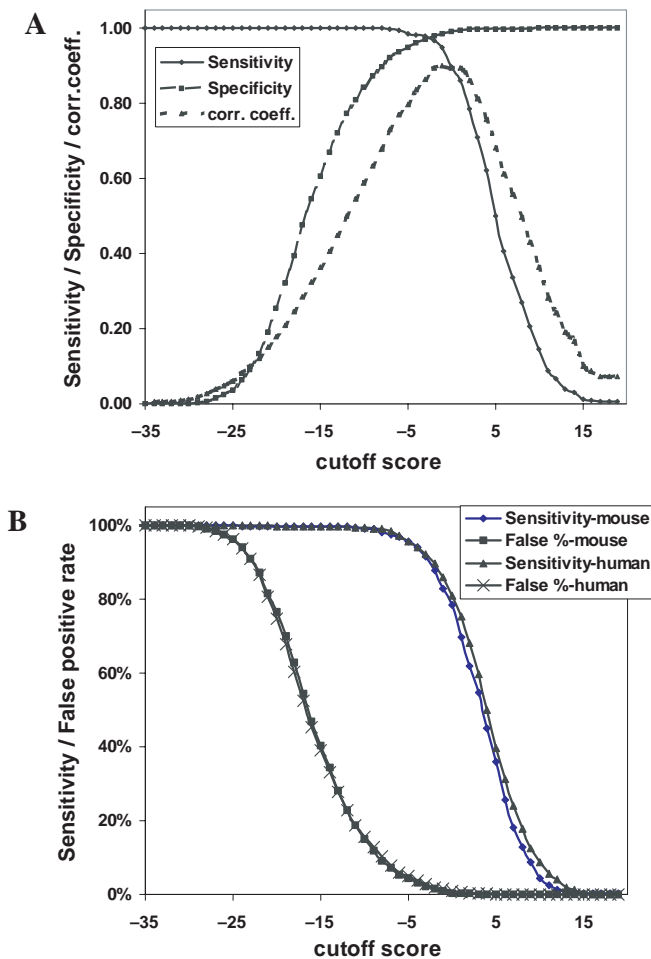
Training data set of human protein sequences and test data set of mouse sequences were extracted from Release 40 of SWISS-PROT (Bairoch and Apweiler, 2000) (details described at <http://share.gene.com>). A Perl script was used to define n-, h-, and c-regions (von Heijne, 1985) of signal peptides for the 363 human signal peptides by following the logic described previously (Nielsen and Krogh, 1998). The maximum lengths for n-, h-, and c-regions were set to 17, 20 and 12, respectively, and 12 sequences that did not show typical sub-regions were excluded. To create multiple sequence alignment, n-domains were aligned to the left, h- and c-regions were aligned to the right, and sufficient gaps were inserted between these regions to allow shorter sequences to align properly with longer sequences. The aligned sequences were then used to build a model for global alignment using the 'hmmbuild' program in the HMMER 2.2 package (Eddy, 1998). It was critical to tune the architecture prior parameter since the default setting at 0.85 failed to give a model with correct domain structures. The optimal value was found empirically to be 0.95.

Using our model, the 'hmmpfam' program was run to generate a score for each of the protein sequences. These scores were used to determine the signal potential, and the alignment coordinates to estimate the cleavage sites. Typically we only used the first 50 amino acid residues of a sequence for hmmpfam analysis.

### RESULTS AND DISCUSSION

Figure 1A shows the self-consistency test results based on the analysis of the signal-containing human sequences used in model building and human non-signal sequences. The separation of these two groups was striking. When the hmmpfam cutoff score is set between  $-5$  and  $-1$ , both the sensitivity and specificity are at least 95%, and Matthews' correlation coefficient (MCC; Matthews, 1975) is about

\*To whom correspondence should be addressed.



**Fig. 1.** Self-consistency test and validation of the signal peptide model. (A) Sensitivity (solid line), specificity (dashed), and Matthews' correlation coefficient (dotted) as a function of the cutoff score from hmmpfam. These were based on the analyses of 3234 non-signal human sequences and 363 signal-containing human proteins used for model building. (B) Discrimination of mouse and human signal-containing proteins from non-signal proteins as a function of the cutoff score. False positive rate is measured as the percentage of predicted positive sequences out of all non-signal sequences. These were based on the analyses of 3234 human and 1958 mouse non-signal proteins, and a wider collection of signal-containing proteins from human (892) and mouse (644).

0.90. We therefore set the cutoff score at  $-5$ , which gives 98.6% sensitivity and 95.1% specificity.

This model was then validated with multiple data sets. Figure 1B shows the performance of this model in separating human or mouse signal-containing sequences from non-signal sequences. Both the sensitivity and false positive curves are almost super-imposable between human and mouse, supporting the predictive ability of this model. Based on the mouse data, sensitivity and

specificity for signal prediction are 95.6 and 95.7% respectively, and MCC is 0.89. Similar results (92.4, 96.5%, and 0.88, respectively) were observed when this prediction method was applied on all eukaryotic test data sets used in SignalP studies (Nielsen and Krogh, 1998). Furthermore, 67% of the known cleavage sites in our sequence collection were predicted precisely, and 78% were predicted to within  $\pm 2$  residues from the sites given in SWISS-PROT. As a comparison, SignalP correctly predicts 69.5% cleavage sites, and, when applied to our test data sets, gives a sig/nonsig MCC value of 0.91. Overall, the signal/non-signal discrimination ability and cleavage-site recognition ability are comparable to previous methods (Nielsen and Krogh, 1998), even though our model is based on human sequences alone. Models built from mouse sequences or a subset of human sequences gave similar performance.

In conclusion, we have used a simple approach to build a reliable profile hidden Markov model compatible with the Pfam database and the HMMER package. The success of building such a model depends on unbiased data selection, appropriate alignment of the three domains within the signal peptides, and maintaining the model structure by optimizing the architecture prior parameter.

## ACKNOWLEDGEMENTS

We thank Colin Watanabe, Yan Zhang and Mike Ward for helpful discussion and technical assistance.

## REFERENCES

- Bairoch, A. and Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
- Bateman, A., Birney, E., Cerruti, L., Durbin, R., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M. and Sonnhammer, E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
- Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
- von Heijne, G. (1983) Patterns of amino acids near signal-sequence cleavage sites. *Eur. J. Biochem.*, **133**, 17–21.
- von Heijne, G. (1985) Signal sequences. The limits of variation. *J. Mol. Biol.*, **184**, 99–105.
- Matthews, B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.
- Nakai, K. (2000) Protein sorting signals and prediction of subcellular localization. *Adv. Protein Chem.*, **54**, 277–344.
- Nielsen, H., Engelbrecht, J., Brunak, S. and von Heijne, G. (1997) A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int. J. Neural Syst.*, **8**, 581–599.
- Nielsen, H. and Krogh, A. (1998) Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **6**, 122–130.