SignalP 4.0: discriminating signal peptides from transmembrane regions

To the Editor: The secretory signal peptide is a ubiquitous protein-sorting signal that targets its passenger protein for translocation across the endoplasmic reticulum membrane in eukaryotes and the cytoplasmic membrane in prokaryotes[1]. Many methods have been published for predicting signal peptides from the amino acid sequence, including SignalP[2–4], PrediSi[5], SPEPlip[6], Signal-CF[7], Signal-3L[8] and Signal-BLAST[9]. A benchmark study done in 2009 found SignalP 3.0 to be the best method[10].

All these methods, however, have only limited ability to distinguish between signal peptides and N-terminal transmembrane helices. Both peptides are hydrophobic, but transmembrane helices typically have longer hydrophobic regions. Also, transmembrane helices do not have cleavage sites, but the cleavage-site pattern is in itself not sufficient to distinguish the two types of sequence. This is a substantial problem because a scan for signal peptides in any complete genome will yield a lot of false positive predictions from N-terminal transmembrane regions.

The hidden Markov model method included in SignalP versions 2.0 (ref. 3) and 3.0 (ref. 4) partially took this issue into account by including three submodels of eukaryotic sequences: signal peptide, signal anchor and other proteins. Other methods such as Phobius[11], Philius[12], membrane protein structure and topology 3 (MEMSAT3)[13], support vector machine–based MEMSAT (MEMSAT-SVM)[14] and Spoctopus[15] try to solve the problem by predicting transmembrane topology as well as signal peptides by joint models.

Here we present SignalP version 4.0, which we designed to discriminate between signal peptides and transmembrane regions. In training SignalP 4.0, we used two kinds of negative data: the first correspond to the negative data used in training earlier versions of SignalP, consisting of cytoplasmic and, for the eukaryotes, nuclear proteins; the second comprise sequences not containing signal peptides but containing transmembrane regions within the first 70 residues of the sequence.

SignalP 4.0 is a purely neural network–based method. We initially retrained the hidden Markov model part of SignalP 3.0 on the SignalP 4.0 data, including the transmembrane sequences, but the neural networks performed better than the hidden Markov model on all of the parameters tested (results not shown).

We used two types of networks in SignalP 4.0: we used transmembrane sequences as negative data to train SignalP-TM networks but trained SignalP-noTM networks without these data. A simple decision scheme is used to select between the networks: if SignalP-TM predicts four or more positions as being transmembrane positions, SignalP-TM is used for the final prediction, otherwise SignalP-noTM is used. We trained and tested the networks separately on three datasets: eukaryotes, Gram-positive bacteria and Gram-negative bacteria (**Supplementary Methods**).

We benchmarked SignalP 4.0 against SignalP 3.0 and ten other signal peptide prediction algorithms (**Fig. 1**). We compared prediction performance using the Matthews correlation coefficient[16], for which each sequence was counted as a true or false positive or negative. To test SignalP 4.0 performance, we did not use data that had been used in training the networks or selecting the optimal architecture, and the test data did not contain homologs to the training and optimization data (**Supplementary Methods**). The test set for SignalP 3.0 was also independent of the training set because we removed sequences used to construct SignalP 3.0 and their homologs from the benchmark data. For other algorithms more recent than SignalP 3.0, the benchmark data may include data used to train the methods, possibly leading to slight overestimations of their performance.

Our results show that SignalP 4.0 was the best signal-peptide predictor for all three organism types (**Fig. 1**). This comes at a price, however, because SignalP 4.0 was not in all cases as good as SignalP 3.0 according to cleavage-site sensitivity or signal-peptide correlation when there are no transmembrane proteins present (**Supplementary Results**). An ideal method would have the best

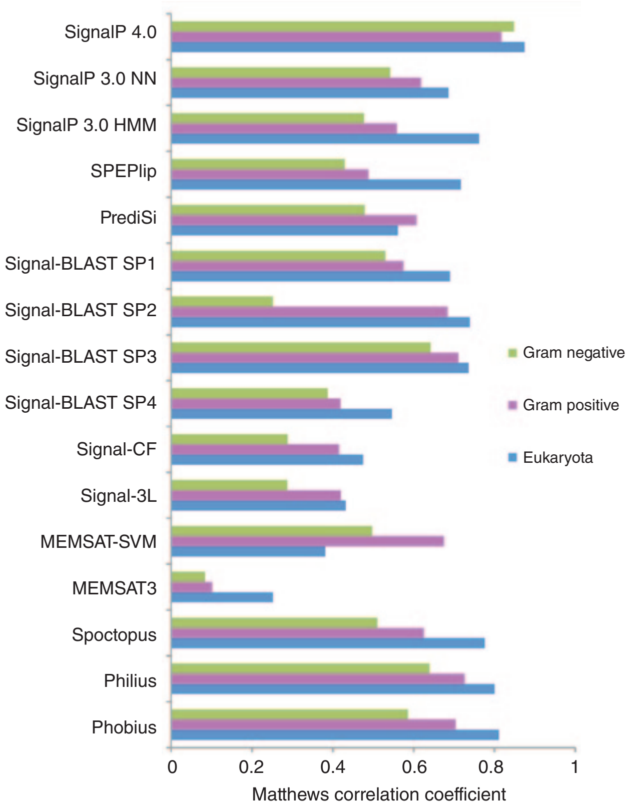

Figure 1 | Comparison of signal peptide prediction algorithm performance shown as the Matthews correlation coefficient for eukaryotes, Gram-positive bacteria and Gram-negative bacteria. HMM, hidden Markov model.

correlation for all negative sets and the best cleavage-site sensitivity, but SignalP 4.0 shows the best ability to discriminate between signal peptides and non–signal peptides in a realistic setting, such as a proteome-wide situation where the data include transmembrane sequences.

SignalP is available as a web tool and as a downloadable version at http://www.cbs.dtu.dk/services/SignalP/. Data used to train and test SignalP 4.0 as well as proteome-wide signal-peptide predictions for three proteomes (one from each organism group) are available under the heading 'Data.' The web version of SignalP 4.0 is free for all users. The downloadable version is free for academic users but is provided for a fee to commercial users.

Note: Supplementary information is available on the Nature Methods website.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

Thomas Nordahl Petersen[1], Søren Brunak[1,2], Gunnar von Heijne[3,4] & Henrik Nielsen[1]

[1]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark. [2]Novo Nordisk Foundation Center for Protein Research, Health Sciences Faculty, University of Copenhagen, Copenhagen, Denmark. [3]Center for Biomembrane Research, Department of Biochemistry and Biophysics, Stockholm University, Stockholm, Sweden. [4]Science for Life Laboratory, Stockholm University, Solna, Sweden.
e-mail: hnielsen@cbs.dtu.dk

1. von Heijne, G. *J. Membr. Biol.* **115**, 195–201 (1990).
2. Nielsen, H., Brunak, S., Engelbrecht, J. & von Heijne, G. *Protein Eng.* **10**, 1–6 (1997).
3. Nielsen, H. & Krogh, A. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6**, 122–130 (1998).
4. Bendtsen, J.D., Nielsen, H., von Heijne, G. & Brunak, S. *J. Mol. Biol.* **340**, 783–795 (2004).
5. Hiller, K., Grote, A., Scheer, M., Munch, R. & Jahn, D. *Nucleic Acids Res.* **32**, W375–3W79 (2004).
6. Fariselli, P., Finocchiaro, G. & Casadio, R. *Bioinformatics* **19**, 2498–2499 (2003).
7. Chou, K.-C. & Shen, H.-B. *Biochem. Biophys. Res. Commun.* **357**, 633–640 (2007).
8. Shen, H.-B. & Chou, K.-C. *Biochem. Biophys. Res. Commun.* **363**, 297–303 (2007).
9. Frank, K. & Sippl, M.J. *Bioinformatics* **24**, 2172 (2008).
10. Choo, K., Tan, T. & Ranganathan, S. *BMC Bioinformatics* **10**, S2 (2009).
11. Käll, L., Krogh, A. & Sonnhammer, E.L.L. *J. Mol. Biol.* **338**, 1027–1036 (2004).
12. Reynolds, S.M., Käll, L., Riffle, M.E., Bilmes, J.A. & Noble, W.S. *PLoS Comput. Biol.* **4**, e1000213 (2008).
13. Jones, D.T. *Bioinformatics* **23**, 538–544 (2007).
14. Nugent, T. & Jones, D. *BMC Bioinformatics* **10**, 159 (2009).
15. Viklund, H., Bernsel, A., Skwark, M. & Elofsson, A. *Bioinformatics* **24**, 2928–2929 (2008).
16. Matthews, B.W. *Biochim. Biophys. Acta* **405**, 442–451 (1975).