Comparison of SVM-Based Methods for Remote Homology Detection

Hiroto Saigo[1] hiroto@kuicr.kyoto-u.ac.jp
Jean-Philippe Vert[2] Jean-Philippe.Vert@mines.org
Tatsuya Akutsu[1] takutsu@kuicr.kyoto-u.ac.jp
Nobuhisa Ueda[1] ueda@kuicr.kyoto-u.ac.jp

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji-city, Kyoto 611-0011, Japan
[2] Geostatistics Center, Ecole des Mines de Paris, Fontainebleau, France

Keywords: support vector machine, protein sequences, sequence alignment, hidden Markov model

1 Introduction

Remote homology detection for protein sequences is one of the important and well-studied problems in Bioinformatics. Many algorithms have been developed for this purpose. The Smith-Waterman (SW) dynamic programming algorithm was developed in early 1980's [8], and is still used widely today. In 1990's, many methods were developed based on *profiles* [1] and *hidden Markov models* [2, 4]. In 2000's, methods using SVMs (support vector machines) were developed such as the SVM-Fisher method [3]. Recently, Liao and Noble proposed the SVM-pairwise method [5], which uses a vector of pairwise similarities with all proteins in the training set. Quite recently, we proposed a new SVM based method (SVM-SW), which uses the SW algorithm as a kernel function [7]. Though we do not yet succeed to prove that the SW score is always a valid kernel, SVM-SW worked successfully in all cases we tested. In this poster abstract, we briefly show the results of comparison of algorithms for remote homology detection using the SCOP database [6].

2 Method and Results

We compared SVM-SW with SVM-pHMM, SVM-Fisher, SVM-pairwise, PSI-BLAST, HMMER, and SAM, where SVM-pHMM is an SVM-based method that uses the score output by a pair HMM model [7].In order to evaluate the accuracy of each method, we follow the benchmark procedure used in [5]. The algorithms are tested on their ability to classify protein domains into superfamilies in the Structural Classification of Proteins (SCOP) [6] version 1.53. We used the data set provided at `www.cs.columbia.edu/compbio/svm-pairwise/`. As a performance measure, we used ROC_{50} scores to compare different homology detection methods. The ROC_{50} score is the area under the receiver operating characterestic curve - the plot of true positives as a function of false positives - up to the first 50 false positives.

3 Conclusion

The results of comparison show that the SVM-SW method significantly outperforms all existing, state-of-the-art algorithms we tested. Moreover the CPU time of SVM-SW is an order of magnitude shorter than the CPU time of SVM-pairwise. Therefore, we can conclude that SVM-SW is currently the best method for detection of remote homology.

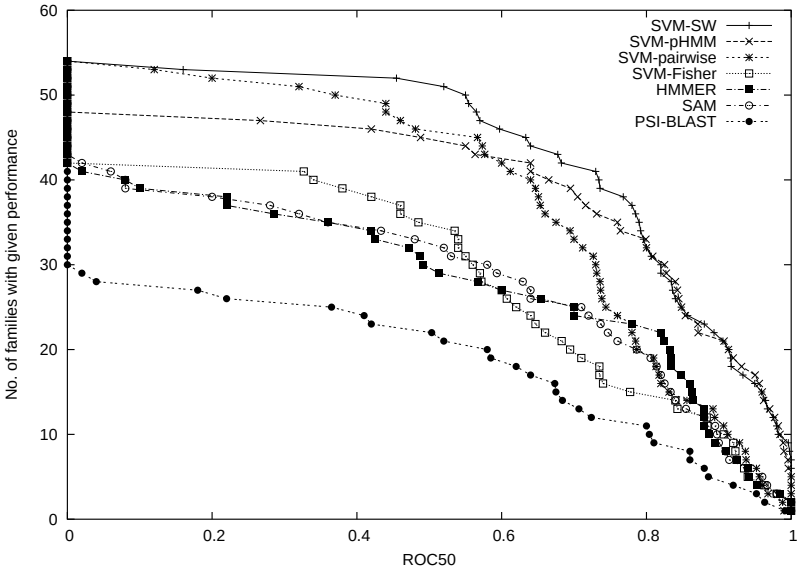

Figure 1: Comparison of seven homology detection methods.

Acknowledgments

We thank Li Liao and William Stafford Noble for making their data set and software available. We also thank Richard Hughey for making the SAM toolkit available, and Mark Diekhans for providing information of the Fisher kernel.

References

[1] Altschul, S. F. *et al.*, Gapped blast and psi-blast: A new generation of protein database search programs, *Nucleic Acids Research*, 25:3389–3402, 1997.

[2] Eddy, S.R., Multiple alignment using hidden Markov models, *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, 114–120, 1995.

[3] Jaakkola, T., Diekhans, M., and Haussler, D., A discriminative framework for detecting remote protein homologies, *Journal of Computational Biology*, 7:95–114, 2000.

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

[5] Liao, L. and Noble, W. S., Combining pairwise sequence similarity and support vector machines for remote protein homology detection, *Proc. 6th Int. Conf. Computational Molecular Biology*, 225–232, 2002.

[6] Murzin, A. G. *et al.*, SCOP: A structural classification of proteins database for the investigation of sequences and structures, *Journal of Molecular Biology*, 247:536–540, 1995.

[7] Saigo, H., Vert, J-P., Akutsu, T., and Ueda, N., Protein homology detection using string alignment kernels, *Manuscript*, 2002.

[8] Smith, T. and Waterman, M. A., Identification of common molecular subsequences, *Journal of Molecular Biology*, 147:195-197, 1981.