Mathematical Modeling and Statistical Analysis in biomedical Research
(February 1-2, 2002, Miyajima, Hiroshima, Japan)

Support Vector Machines (SVMs) in bioinformatics

Jean-Philippe Vert *
Bioinformatics Center, Institute for Chemical Research, Kyoto University

Support Vector Machines (SVMs) and related kernel methods ([8]) are a powerful class of algorithms first introduced by Vapnik and coworkers in the 90's for classification and regression problems. They are based on a sound theoretical framework, namely statistical learning theory, and have been shown to provide very powerful algorithms for several real-world applications, such as optical character recognition or various prediction problems.

These methods have been gaining popularity recently in bioinformatics for at least two reasons:

- SVMs can easily replace currently used algorithms (such as neural networks) for classification problems where the input is a vector and the output is a class, and appear to usually outperform other algorithms. Examples of such classification problems include for example the prediction of the cellular localization of a protein from a vector of features, such as the frequency of each amino-acid in the protein ([5]). SVMs are particularly resistant to overfitting even in large dimensions, when only a few training examples are available.

- SVMs can be adapted to problems where the objects to classify are not real vectors thanks to the possibility of designing *kernel functions*. A kernel function $K(x, y)$ where x and y are two objects is meant to represent the similarity between the objects, and is the only information SVM uses about the objects. As an example, if one wants to classify variable-length sequences (such as proteins) which can not easily be represented as vectors, it suffices to define a kernel function between the sequences (related for instance to the alignment score of the sequences) to be able to use SVMs

In this talk I will first review the SVM algorithm itself, and present various applications of SVMs to bioinformatics which were developed in the recent years. These applications include analysis of microarray data for gene function prediction ([1, 7]) or tissue classification ([3]), automatic protein classification into functional families ([6]), translation initiation site recognition ([10]), protein secondary structure prediction ([4]) or protein fold recognition ([2]).

I will then propose new methods to adapt SVMs to particular problems arising in bioinformatics through the design of new kernels ([9]). I will consider the case when one wants to classify structured objects such as sequences or graphs, and has been able to estimate a probability distribution on the space of objects, such that objects with a high probability are

*Uji, Kyoto 611-0011, Japan. e-mail: Jean-Philippe.Vert@mines.org more likely to belong to one particular class. This case is very usual in bioinformatics when one uses for instance hidden Markov models to characterize a family of proteins, or weight matrices to characterize particular motifs in DNA. I will show how it is possible to design a Euclidean geometry on the space of objects such that two objects are close to each other whenever they share rare common subparts. This Euclidean geometry is concretely defined through a kernel function $K(x, y)$ which represents the inner product between any two objects x and y in the Euclidean space, and can be used by SVMs to linearly discriminate the objects in this space. I will give experimental results on the prediction of signal peptide cleavage site in protein.

References

[1] Michael P. S. Brown, William Noble Grundy, David Lin, Nello Cristianini, Charles Walsh Sugnet, Terence S. Furey, Jr. Manuel Ares, and David Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA*, 97:262–267, 2000.

[2] Chris Ding and Inna Dubchak. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, 17:349–358, 2001.

[3] Terrence S. Furey, Nigel Duffy, Nello Cristianini, David Bednarski, Michel Schummer, and David Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.

[4] Sujun Hua and Zhirong Sun. A novel method of protein secondary structure prediction with high segment overlap measure: Support vector machine approach. *Journal of Molecular Biology*, 308(2):397–407, April 2001.

[5] Sujun Hua and Zhirong Sun. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, 17(8):721–728, 2001.

[6] Tommi Jaakkola, Mark Diekhans, and David Haussler. A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*, 7(1,2):95–114, 2000.

[7] Paul Pavlidis, Jason Weston, Jinsong Cai, and William Noble Grundy. Gene functional classification from heterogeneous data. In *Proceedings of the Fifth Annual International Conference on Computational Biology*, pages 249–255, 2001.

[8] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, New-York, 1998.

[9] Jean-Philippe Vert. Support vector machine prediction of signal peptide cleavage site using a new class of kernels for strings. In Russ B. Altman, A. Keith Dunker, Lawrence Hunter, Kevin Lauerdale, and Teri E. Klein, editors, *Proceedings of the Pacific Symposium on Biocomputing 2002*, pages 649–660. World Scientific, 2002.

[10] A. Zien, G. Ratsch, S. Mika, B. Schölkopf, T. Lengauer, and K.-R. Muller. Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*, 16(9):799–807, 2000.