# Comparing Kernels For Predicting Protein Binding Sites From Amino Acid Sequence

Feihong Wu[1,2,5,7] , Byron Olson[6,8], Drena Dobbs[3,4,5,6,9] and Vasant Honavar[1,2,4,5,6,10]

***Abstract*— The ability to identify protein binding sites and to detect specific amino acid residues that contribute to the specificity and affinity of protein interactions has important implications for problems ranging from rational drug design to analysis of metabolic and signal transduction networks. Support vector machines (SVM) and related kernel methods offer an attractive approach to predicting protein binding sites. An appropriate choice of the kernel function is critical to the performance of SVM. Kernel functions offer a way to incorporate domain-specific knowledge into the classifier.**

**We compare the performance of 3 types of kernels functions: identity kernel, sequence-alignment kernel, and amino acid substitution matrix kernel for predicting protein-protein, protein-DNA and protein-RNA binding sites. The results show that the identity kernel is quite effective in on all three tasks, with the substitution kernel based on amino acid substitution matrices that take into account structural or evolutionary conservation or physicochemical properties of amino acids yields modest improvement in the performance of the resulting SVM classifiers for predicting protein-protein, protein-DNA and protein-RNA binding sites.**

## I. Introduction

Proteins are the principal catalytic agents, structural elements, signal transmitters, transporters and molecular machines in cells. Hence, assigning them putative functions from sequences alone remains one of the most challenging problems in functional genomics. Protein-protein, protein-DNA, and protein-RNA interactions play a pivotal role in protein functions. Experimental detection of residues in protein-protein interaction surfaces must come from determination of the structure of protein-protein, protein-DNA and protein-RNA complexes. However, experimental determination of structures of such complexes is a time-consuming and expensive enterprise. Hence, there is a need for reliable computational methods for identifying protein-protein, protein-DNA and protein-RNA binding sites from the amino acid sequence of the protein. Machine learning methods in general, and support vector machines and related kernel methods in particular, offer an attractive approach to construction of sequence-based classifiers for identifying such binding sites [Schőlkopf et al., 2003], [Yan et al., 2004a], [Yan et al., 2004b], [Vert, 2005], [Terribilini et al., 2006].

The SVM [Boser et al., 1992] classifies inputs into two classes using a hyperplane in a high-dimensional space. If the patterns are not separable in the original $n$-dimensional pattern space, a suitable non-linear kernel function is used to implicitly map the patterns in the $n$-dimensional input space into a typically higher (finite or even infinite)dimensional feature space in which the patterns become separable. SVM selects the hyperplane that maximizes the margin of separation between the two classes from among all separating hyperplanes. The kernel function measures the similarity between pairs of patterns in the feature space. An appropriate choice of the kernel function is critical to the performance of SVM. An ideal kernel function assigns a higher similarity score to any pair of patterns that belong to the same class label than it does to any pair of patterns that belong to different classes. Kernel functions provide a means of incorporating domain-specific knowledge into an SVM. Hence, there is a large body of work aimed at designing suitable kernels for protein sequence classification [Leslie et al., 2002], [Leslie et al., 2004]. Against this background, we investigate the effect of incorporating various types of biological information into SVM kernels for protein-protein, protein-DNA, and protein-RNA binding site prediction.

The rest of this paper is organized as follows: Section 2 describes the 3 data sets used in the study. Section 3 introduces the kernel methods and elaborates on the design of the 3 types of kernel functions. Section 4 presents the experimental results and discusses the factors influencing classification performance. Section 5 summarizes the findings and suggests future work. The final section lists kernel method applications in computational biology field.

## II. Materials

The data sets used in this study are available for download at http://www.cild.iastate.edu/GM066387_homepage.htm.

### A. 42 Peptidase Protein-Protein Interface Data Set

A peptidase is an enzyme that digests proteins through the breaking of peptide bonds. The peptidase interface data set consists of 42 peptidase chains (with sequence identity $< 40\%$) from the MEROPS database[Rawlings et al., 2004]. Interface residues (binding sites; those amino acids in the sequence that bind to another protein ) are defined by a loss in solvent accessible surface area (ASA) from the free monomer to the bound complex. The ASA is computed using the Naccess program[Hubbard, 1993](http://wolf.bms.umist.ac.uk/naccess/). A residue is defined as a interface residue when its ASA lost

[1] Artificial Intelligence Research Laboratory.
[2] Department of Computer Science.
[3] Department of Genetics, Development and Cell Biology.
[4] Laurence H Baker Center for Bioinformatics and Biological Statistics.
[5] Bioinformatics and Computational Biology Graduate Program.
[6] Center for Computational Intelligence, Learning, and Discovery.
Iowa State University, Ames, IA 50011-1040,USA.
[7] wuflyh@cs.iastate.edu,[8] olson@iastate.edu,
[9] ddobbs@iastate.edu,[10] honavar@cs.iastate.edu.

on complex formation is $> 1\dot{A}^2$[Jones and Thornton, 1996]. Relative solvent accessibility is defined as the ratio of ASA to the nominal maximal ASA of the residue by Rost and Sander[Rost and Sander, 1994]. A residue is defined as a surface residue when the relative accessibility is greater than $25\%$. This data set consists of $1694$ interface residues out of $5513$ total surface residues.

### B. 56 Protein-DNA Interface Data Set

Specific proteins bind DNA to direct DNA replication and transcription. The $56$ protein-DNA binding data set, first published by Jones[Jones et al., 2003], includes $56$ non-homologous protein chains. The definition of interface residues is the same as in the $42$ peptidase interface data set. This results in $1752$ interface residues out of $12665$ total residues.

### C. 109 Protein-RNA Interface Data Set

The $109$ protein-RNA binding data set extracted from PDB [Berman et al., 2000] consists of $109$ non-homologous protein chains. Interface residues are determined using software ENTANGLE [Allers and Shamoo, 2001]. The data set consists of $3518$ interface residues out of $25,118$ total residues.

## III. Method

### A. Support Vector Machines and Kernel Functions

The SVM classifies inputs into two classes using a hyperplane in a high-dimensional space. If the patterns are not separable in the original *n*-dimensional pattern space, a suitable non-linear kernel function is used to implicitly map the patterns in the *n*-dimensional input space into a typically higher (finite or even infinite)dimensional feature space in which the patterns become separable. SVM selects the hyperplane that maximizes the margin of separation between the two classes $C_+$ and $C_-$ from among all separating hyperplanes. The kernel function measures the similarity between pairs of patterns in the feature space. Given the training data set with *m* labelled examples

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), (\mathbf{x}_3, y_3), ..., (\mathbf{x}_m, y_m)$$
$$where \begin{cases} y_k = 1 & \text{if } \mathbf{x}_k \in C_+; \\ y_k = -1 & \text{if } \mathbf{x}_k \in C_-, \end{cases}$$

the SVM produces a decision function:

$$\begin{aligned} D(\mathbf{x}) &= \sum_{k=1}^{m} \alpha_k \mathbf{K}(\mathbf{x}_k, \mathbf{x}) + b \\ & \text{such that} \\ & \text{ if D}(\mathbf{x}) > 0, \mathbf{x} \in A \\ & \text{ otherwise } \mathbf{x} \in B \end{aligned}$$

where the kernel function **K** is a predefined kernel function. The *weights* $\alpha_k$ $(1 \leq k \leq m)$ and the *bias b* are determined by the SVM algorithm. The training samples with non-zero weights are called the support vectors.

The optimization procedure used in training a support vector machine coefficients essentially solves a quadratic programming problem. This utilizes a *Kernel* whose elements represent the pairwise kernel evaluations between training instances (i.e. $K_{ij} = K(x_i, x_j)$). A valid kernel function needs to satisfy the Mercer conditions which requires the kernel matrix to be positive semi-definite [Lanckriet et al., 2002].

### B. Input Representation and Kernel Function Definition

In this study, the SVM was trained to predict whether or not a residue is in the interaction site. The input to the SVM consists of the identity of amino acids within a window of 11 contiguous residues, corresponding to the target residue flanked by five sequence neighbors residues on each side. The desired output of the classifier is a 1 if the target residue is an interface residue (class $C_+$) and -1 (class $C_-$) otherwise. The training set consists of 11-residue subsequences extracted from the protein sequences, with each window labelled with the corresponding class label.

A kernel function defines similarity between two fixed length sequences $S_a = a_1 a_2 ... a_n$ and $S_b = b_1 b_2 ... b_n$ in which $a_i, b_i (1 \leq i \leq n)$ are amino acids and $n$ is the width of the window. We define three kernel functions: the *identity kernel*, the *alignment kernel*, and the *substitution kernel*.

*Definition 1 (identity kernel):* The identity kernel counts the number of matching residues between the two strings $S_a, S_b$.

$$\begin{aligned} \mathbf{K}_i(S_a, S_b) &= \sum_{k=1}^{n} \mathbf{e}(a_k, b_k) \\ where & \begin{cases} \mathbf{e}(a_k, b_k) = 1, & \text{if } a_k = b_k; \\ \mathbf{e}(a_k, b_k) = 0 & \text{otherwise.} \end{cases} \end{aligned}$$

It is easy to show that the resulting kernel matrix $\mathbf{K}_i$ is a positive semidefinite matrix.

*Definition 2 (alignment kernel):* Let **A** be a matrix of alignment scores obtained by locally aligning each pair of strings $S_a, S_b$, in the training set.

$$\mathbf{A}(S_a, S_b) = \mathbf{align}(S_a, S_b)$$

where $\mathbf{align}(S_a, S_b)$ is the alignment score based on local alignment of $S_a$ and $S_b$. The **align** function, and hence the matrix **A** is not guaranteed to be positive definite. To circumvent this problem, we define the alignment kernel $\mathbf{K}_a$ as follows:

$$\mathbf{K}_a(S_a, S_b) = \begin{cases} \mathbf{A}(S_a, S_b) - \lambda_g & \text{if } S_a = S_b; \\ \mathbf{A}(S_a, S_b) & \text{otherwise} \end{cases}$$

where $\lambda_g$ is the smallest eigenvalue of the matrix of pairwise alignment scores **A**. The resulting matrix $\mathbf{K}_a$ is a positive semidefinite matrix.

*Definition 3 (substitution kernel):* Let $\mathbf{M}_s$ be an amino acid substitution matrix [Henikoff and Henikoff, 1992]. Substitution matrices are not typically positive definite. We can create a positive semidefinite matrix $\mathbf{M}$ from a substitution matrix $\mathbf{M}_s$ as follows:

1) Find the minimal entry *min* of $\mathbf{M}_s$
2) Find the maximal entry *max* of $\mathbf{M}_s$
3) $\mathbf{M}(i,j) = \frac{\mathbf{M}_s(i,j)-min}{max-min}$
4) Find the least eigenvalue $\lambda$ of $\mathbf{M}$
5) $\mathbf{M}(i,i) = \mathbf{M}(i,i) - \lambda$

The substitution kernel is defined as follows:

$$\mathbf{K}_s(S_a, S_b) = \sum_{k=1}^{n} \mathbf{M}(a_k, b_k)$$

Substitution matrixes of amino acid are symmetric matrices expressing the rate of substitution of one amino acid by another. A variety of substitution matrices that are based on physical, chemical and biological properties of amino acids as well as evolutionary and structural considerations are available in the AAindex database[Kawashima and Kanehisa, 2000]. For example, HENS920102, a well known BLOSUM62 matrix, is based on evolutionary considerations; The substitution matrix JOHM930101 is based on structural considerations, and MCLA720101 is based on chemical properties of amino acids.

### C. Performance Measures

Let *TP* be the number of true positives(residues predicted to be interaction sites that are actually interaction sites); *FP* the number of false positives(residues predicted to be interaction sites that are actually non-interaction sites); *TN* the number of true negatives; *FN* the number of false negatives. the performance measures *ac*(accuracy), *re*(recall), *pr*(precision) and *cc*(correlation coefficient) defined as follows:

$$\begin{aligned} ac &= \frac{TP + TN}{TP + FP + TN + FN} \\ re &= \frac{TP}{TP + FN} \\ pr &= \frac{TN}{TP + FP} \\ cc &= \frac{TP * TN - FN * FP}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}} \end{aligned}$$

## IV. Experimental Results

We trained SVM classifiers for predicting whether or not a target residue is a (protein-protein, protein-DNA, or protein-RNA) interface residue based on the amino acid identities of its sequence neighbors using the identity kernel $\mathbf{K}_i$, alignment kernel $\mathbf{K}_a$ and substitution kernel $\mathbf{K}_s$. The classifiers were trained and evaluated (using leave-one-out cross-validation) on the 3 data sets: P(42 peptidase protein-protein interface data set), D(56 protein-DNA interface data set) and R(109 protein-RNA interface data set). The alignment kernel was derived using the BLOSUM62 (HENS920102) substitution matrix. The substitution kernel was derived using 3 different substitution matrixes and get 3 substitution kernels: $\mathbf{K}_{sh}$ with evolution based substitution matrix HENS920102, $\mathbf{K}_{sj}$ with structure based matrix JOHM930101 and $\mathbf{K}_{sm}$ with chemical similarity based matrix MCLA720101. Our SVM classifiers with different kernels were implemented based on WEKA machine learning package[Witten and Frank, 2005].

When data sets have unbalanced class representation (as in the case with the data sets used in this study), the traditional performance measure of accuracy can present a misleading picture of the effectiveness of the classifier. Hence we report multiple performance measures including accuracy, recall, precision, and correlation coefficient. The results are summarized in Table I.

TABLE I

Comparison of the amino acid identity kernel $\mathbf{K}_i$, the alignment kernel $\mathbf{K}_a$, and several substitution kernels $\mathbf{K}_{sh}$ $\mathbf{K}_{sj}$ and $\mathbf{K}_{sm}$ (derived from HENS920102, JOHM930101, and MCLA720101 substitution matrices respectively). Accuracy (*ac*), recall(*re*), precision ( *pr*), and correlation coefficient (*cc*) shown are estimated using leave-one-out cross-validation.

| data set | kernel function | *ac* | *re* | *pr* | *cc* |
|---|---|---|---|---|---|
| P | $\mathbf{K}_i$ | 60.3% | 54.9% | 42.0% | 16.6% |
| | $\mathbf{K}_a$ | 63.7% | 47.6% | 43.9% | 16.6% |
| | $\mathbf{K}_{sh}$ | 63.4% | 48.1% | 44.0% | 17.7% |
| | $\mathbf{K}_{sj}$ | 63.6% | 49.7% | 44.5% | **18.9%** |
| | $\mathbf{K}_{sm}$ | 62.0% | 51.4% | 42.7% | 17.0% |
| D | $\mathbf{K}_i$ | 64.0% | 69.6% | 30.0% | 25.0% |
| | $\mathbf{K}_a$ | 63.9% | 66.0% | 29.4% | 22.7% |
| | $\mathbf{K}_{sh}$ | 64.1% | 69.3% | 29.7% | 24.4% |
| | $\mathbf{K}_{sj}$ | 64.4% | 68.1% | 29.8% | 24.3% |
| | $\mathbf{K}_{sm}$ | 65.1% | 69.6% | 30.3% | **25.7%** |
| R | $\mathbf{K}_i$ | 71.2% | 60.3% | 34.8% | 25.1% |
| | $\mathbf{K}_a$ | 69.2% | 53.1% | 31.9% | 18.0% |
| | $\mathbf{K}_{sh}$ | 72.1% | 58.4% | 35.3% | 24.9% |
| | $\mathbf{K}_{sj}$ | 72.2% | 58.9% | 35.5% | **25.3%** |
| | $\mathbf{K}_{sm}$ | 71.6% | 58.6% | 34.8% | 24.3% |

The performance of the identity kernel is competitive with that of other kernels on all three prediction tasks.

The substitution kernel, depending on the data set used, and the specific substitution kernel chosen, sometimes outperforms the identity kernel. In the case of the peptidase protein-protein interface data set, the substitution kernel yields a 13.9% **relative** improvement in correlation coefficient over the identity kernel when the JOHM930101 substitution matrix is used; In the case of the other two data sets, the relative improvement in correlation coefficient offered by the substitution kernel is quite small: 2.8% (using MCLA720101 substitution matrix on the protein-DNA interface data set) and 0.8% (using JOHM930101 substitution matrix on the protein-RNA interface data set) respectively.

The alignment kernel does not perform as well as the other kernels on these data sets. This might be due to the fact that the substitution matrix used for aligning sequences (BLOSUM62) may be suboptimal for the data sets used. (Note that the results of the substitution kernel varies with on the specific substitution matrix used).

## V. Related Work

Kernel methods have been widely applied in computational biology, and many kernel functions have been specifically designed for biological data [Schölkopf et al., 2003], [Vert, 2005]. Several authors have explored the use of support vector machines for secondary structure prediction [Hua and Sun, 2001] [Guo et al., 2004]. Bram et al. [Vanschoenwinkel and Manderick, 2004] have examined the effects of amino acid substitution matrix on the effectiveness of SVM kernels for secondary structure prediction. Jaakkola et al. [Jaakkola et al., 2000] have derived a Hidden Markov Model (HMM) profile based SVM-Fisher kernel for remote homology detection. Leslie et al [Leslie et al., 2002] have explored the $p$-spectrum kernel and a mismatch kernel[Leslie et al., 2004] for protein function classification. Saigo et al. [Saigo et al., 2004] have proposed a string alignment kernel for protein remote homology detection. Lanckriet et al.[Lanckriet et al., 2004] have developed a method based on semi-definite programming for optimal linear combination of multiple kernels for protein function prediction.

Several authors have explored the application of machine learning approaches to classification of protein-protein, protein-DNA, and protein-RNA interface sites from amino acid sequences. Yan et al. [Yan et al., 2004a], [Yan et al., 2004b] have used SVM for identifying protein-protein interface residues among surface residues using amino acid sequence information. Sen et al. [Sen et al., 2004] have proposed an approach to combining several different sources of information (including amino acid sequence, evolutionary conservation, and structure comparison) to improve the accuracy of protein-protein interface residues. Yan et al. [Yan et al., 2006] have explored the use of several types of information derived from amino acid sequences to train a Naive Bayes classifier on the 56 protein-DNA data set used in this study. The result obtained using amino acid sequence identity alone (correlation coefficient of 24%) is comparable to that of the SVM reported here. However, addition of residue entropy of the target residue (obtained from multiple sequence alignment) with other sequences in the training data set as an additional input to the classifier improved the correlation coefficient to 28%. Terribilini et al. [Terribilini et al., 2006] have used a Naive Bayes classifier to predict protein-RNA interface residues from amino acid sequence. On the data set of 109 protein-RNA interfaces which is same as the protein-RNA interface data set used in our study, the Naive Bayes classifier yields a correlation coefficient of 35%, which is better than that of SVM trained using sequence kernels. However, the reported performance of Naive Bayes classifier for protein-RNA interface prediction was obtained with a window size of 25 (as opposed to a window size of 11 used in our study).

## VI. Summary

We have compared the performance of 3 types of kernels to predict protein-protein, protein-DNA, and protein-RNA interfaces from amino acid sequence information alone. Our results suggest that the identity kernel is competitive with apparently more sophisticated kernels on all three prediction tasks. Our results also suggest the possibility of improving the performance of the SVM classifiers using kernel functions derived using amino acid substitution matrices. Yan et al. [Yan et al., 2006] have shown that it is possible to improve the accuracy of protein-DNA interface prediction by using sequence entropy of the target residue as an additional input to the Naive Bayes classifier. Sen et al. [Sen et al., 2004] have reported improved accuracy of protein-protein interface prediction using multiple types of information. Hence, there is reason to expect that the performance of the SVM classifiers reported in this paper can be further improved by using other types of information such as sequence conservation score [Glaser et al., 2003], predicted or known secondary structure, sequence entropy, sequence disorder, sequence entropy, among others. Work in progress is aimed at exploring these possibilities.

**Acknowledgements:** This research was supported by a grant from the National Institutes of Health (GM066387) to Vasant Honavar. We thank C. Yan and M. Terribilini for providing data sets and discussions.

## References

[Allers and Shamoo, 2001] Allers, J. and Shamoo, Y. (2001). Structure-based analysis of protein-rna interactions using the program entangle. *J Mol Biol*, 75-86:311(1).

[Berman et al., 2000] Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., and Bourne, P. (2000). The protein data bank. *Nucleic Acids Res*, 28:235–242.

[Boser et al., 1992] Boser, B. E., Guyon, I., and Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *Computational Learing Theory*, pages 144–152.

[Glaser et al., 2003] Glaser, F., Pupko, T., Paz, I., Bell, R., Bechor-Shental, D., Martz, E., and Ben-Tal, N. (2003). Consurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, 19(1):163–4.

[Guo et al., 2004] Guo, J., Chen, H., Sun, Z., and Lin, Y. (2004). A novel method for protein secondary structure prediction using dual-layer svm and profiles. *Proteins*, 54(4):738–43.

[Henikoff and Henikoff, 1992] Henikoff, S. and Henikoff, J. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–9.

[Hua and Sun, 2001] Hua, S. and Sun, Z. (2001). A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *J Mol Biol*, 308(2):397–407.

[Hubbard, 1993] Hubbard, S. (1993). Naccess. department of biochemistry and molecular biology, university college, london.

[Jaakkola et al., 2000] Jaakkola, T., Diekhans, M., and Haussler, D. (2000). A discriminative framework for detecting remote protein homologies. *J Comput Biol*, 7(1-2):95–114.

[Jones et al., 2003] Jones, S., Shanahan, H., Berman, H., and Thornton, J. (2003). Using electrostatic potentials to predict dna-binding sites on dna-binding proteins. *Nucleic Acids Res*, 31(24):7189–98.

[Jones and Thornton, 1996] Jones, S. and Thornton, J. (1996). Principles of protein-protein interactions. *Proc Natl Acad Sci USA*, 93(1):13–20.

[Kawashima and Kanehisa, 2000] Kawashima, S. and Kanehisa, M. (2000). Aaindex: amino acid index database. *Nucleic Acids Res*, 28(1):374.

[Lanckriet et al., 2004] Lanckriet, G., Deng, M., Cristianini, N., Jordan, M., and Noble, W. (2004). Kernel-based data fusion and its application to protein function prediction in yeast. *Pac Symp Biocomput*, pages 300–11.

[Lanckriet et al., 2002] Lanckriet, G. R. G., Cristianini, N., Bartlett, P. L., Ghaoui, L. E., and Jordan, M. I. (2002). Learning the kernel matrix with semi-definite programming. In *ICML*, pages 323–330.

[Leslie et al., 2004] Leslie, C., Eskin, E., Cohen, A., Weston, J., and Noble, W. (2004). Mismatch string kernels for discriminative protein classification. *Bioinformatics*, 20(4):467–76.

[Leslie et al., 2002] Leslie, C., Eskin, E., and Noble, W. S. (2002). The spectrum kernel: A string kernel for svm protein classification. In *Pacific Symposium on Biocomputing*, pages 564–575.

[Rawlings et al., 2004] Rawlings, N., Tolle, D., and Barrett, A. (2004). Merops: the peptidase database. *Nucleic Acids Res*, 32:D160–4.

[Rost and Sander, 1994] Rost, B. and Sander, C. (1994). Conservation and prediction of solvent accessibility in protein families. *Proteins*, 216-26:20(3).

[Saigo et al., 2004] Saigo, H., Vert, J., Ueda, N., and Akutsu, T. (2004). Protein homology detection using string alignment kernels. *Bioinformatics*, 20(11):1682–9.

[Schölkopf et al., 2003] Schölkopf, B., Tsuda, K., and Vert, J.-P. (2003). *Kernel methods in computational biology. MIT Press*.

[Sen et al., 2004] Sen, T. Z., Kloczkowski, A., Jernigan, R. L., Yan, C., Honavar, V., Ho, K.-M., Wang, C.-Z., Ihm, Y., Cao, H., Gu, X., and Dobbs, D. (2004). Predicting binding sites of hydrolase-inhibitor complexes by combining several methods. *BMC Bioinformatics*, 5(1):205.

[Terribilini et al., 2006] Terribilini, M., Lee, J.-H., Yan, C., Jernigan, R. L., Carpenter, S., Honavar, V., and Dobbs, D. (2006). Identifying interaction sites in “recalcitrant” proteins: predicted protein and RNA binding sites in Rev proteins of HIV-1 and EIAV agree with experimental data. *Pacific Symposium on Biocomputing*, 11:415–426.

[Vanschoenwinkel and Manderick, 2004] Vanschoenwinkel, B. and Manderick, B. (2004). Substitution matrix based kernel functions for protein secondary structure. *The Procedding of International Conference on Machine Learning and Applications*.

[Vert, 2005] Vert, J.-P. (2005). Kernel methods in genomics and computational biology. *Technical Report HAL:ccsd-00012124, October 2005*.

[Witten and Frank, 2005] Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco*.

[Yan et al., 2004a] Yan, C., Dobbs, D., and Honavar, V. (2004a). Identification of interface residues in protease-inhibitor and antigen-antibody complexes: a support vector machine approach. *Neural Comput. Applic.*, 13:123–129.

[Yan et al., 2004b] Yan, C., Dobbs, D., and Honavar, V. (2004b). A two-stage classifier for identification of protein-protein interface residues. *Bioinformatics*, Suppl 1:I371–I378.

[Yan et al., 2006] Yan, C., Terribilini, M., Wu, F., Jernigan, R. L., Dobbs, D., and Honavar, V. (2006). Predicting dna-binding sites of proteins from amino acid sequence. *BMC Bioinformatics*, in press.