Mismatch string kernels for discriminative protein classification

Christina S. Leslie[1,*], Eleazar Eskin[1], Adiel Cohen[1], Jason Weston[2] and William Stafford Noble[3,†]

[1]Department of Computer Science, Columbia University, 1214 Amsterdam Avenue, Mail Code 0401, New York, NY 10027, USA, [2]Max-Planck Institute for Biological Cybernetics, Spemannstrasse 38, 72076 Tübingen, Germany and [3]Department of Genome Sciences, University of Washington, 1705 NE Pacific Street, Seattle, WA 98195, USA

ABSTRACT

Motivation: Classification of proteins sequences into functional and structural families based on sequence homology is a central problem in computational biology. Discriminative supervised machine learning approaches provide good performance, but simplicity and computational efficiency of training and prediction are also important concerns.

Results: We introduce a class of string kernels, called mismatch kernels, for use with support vector machines (SVMs) in a discriminative approach to the problem of protein classification and remote homology detection. These kernels measure sequence similarity based on shared occurrences of fixed-length patterns in the data, allowing for mutations between patterns. Thus, the kernels provide a biologically well-motivated way to compare protein sequences without relying on family-based generative models such as hidden Markov models. We compute the kernels efficiently using a mismatch tree data structure, allowing us to calculate the contributions of all patterns occurring in the data in one pass while traversing the tree. When used with an SVM, the kernels enable fast prediction on test sequences. We report experiments on two benchmark SCOP datasets, where we show that the mismatch kernel used with an SVM classifier performs competitively with state-of-the-art methods for homology detection, particularly when very few training examples are available. Examination of the highest-weighted patterns learned by the SVM classifier recovers biologically important motifs in protein families and superfamilies.

Availability: SVM software is publicly available at http://microarray.cpmc.columbia.edu/gist. Mismatch kernel software is available upon request.

Contact: cleslie@cs.columbia.edu

---

*To whom correspondence should be addressed.

†Formerly William Noble Grundy, see www.gs.washington.edu/noble/name-change.html

INTRODUCTION

One of the central problems in computational biology is the classification of protein sequences into functional and structural families based on sequence homology. Approaches based on pairwise similarity of sequences (Waterman *et al.*, 1991; Altschul *et al.*, 1990, 1997), profiles for protein families (Gribskov *et al.*, 1987), consensus patterns using motifs (Bairoch, 1995; Attwood *et al.*, 1998) and hidden Markov models (Krogh *et al.*, 1994; Eddy, 1995; Baldi *et al.*, 1994) have all been used for this problem. Recent research suggests that the best-performing methods are discriminative: protein sequences are seen as a set of labeled examples—positive if they are in the family and negative otherwise—and a learning algorithm attempts to learn a decision boundary between the different classes. In this category, several successful techniques (Jaakkola *et al.*, 2000; Liao and Noble, 2002; Leslie *et al.*, 2002a) use protein sequences to train a support vector machine (SVM) (Vapnik, 1998) classifier.

In this paper, we present a method for using SVMs for remote homology detection, based on a family of kernel functions called mismatch kernels. A kernel function measures the similarity between a pair of inputs, and defines an inner product in an implicit feature space for the SVM optimization problem. The features used by our mismatch kernel are the set of all possible subsequences of amino acids of a fixed length k. If two protein sequences contain many k-length subsequences that differ by at most m mismatches, then their inner product under the mismatch kernel will be large. More precisely, the mismatch kernel is calculated based on shared occurrences of (k, m)-patterns in the data, where the (k, m)-pattern generated by a fixed k-length subsequence consists of all k-length subsequences differing from it by at most m mismatches. Thus, the mismatch kernel extends the computationally simpler spectrum kernel presented in Leslie *et al.* (2002b), adding the biologically important notion of mismatches.

We also describe the mismatch tree data structure that we use to compute the mismatch kernel. We can efficiently compute the contributions of all instances of (k, m)-patterns occurring in the data to all entries of the kernel matrix in one pass while traversing the tree. For (k, m)-parameters that are useful in applications, the compute time to generate the kernel is fast enough for practical use on real datasets. Moreover, when mismatch kernels are used with SVMs, we can implement the classification to make linear-time prediction on test sequences.

We report results for two sets of experiments over the structural classification of proteins (SCOP) database (Murzin *et al.*, 1995). In the first set of experiments, we test our method on the benchmark dataset assembled by Jaakkola *et al.* (2000), where SCOP sequences are augmented by domain homologs of positive training sequences in order to assist HMM-based methods. We show that our mismatch kernel, in conjunction with an SVM classifier, performs competitively with state-of-the-art methods such as the the SVM-Fisher method on this dataset. In the second set of experiments, we perform a similar set of experiments on a newer version of the SCOP database (Liao and Noble, 2002); however, in this second test, we limit the training examples to proteins in the original SCOP dataset. In the absence of additional domain homologs for training the hidden Markov models, SVM-Fisher and other HMM-based approaches exhibit poorer performance than our mismatch-SVM approach. However, mismatch-SVM performs as well as SVM-pairwise, the best-performing method reported in Liao and Noble (2002) for this benchmark.

The current work is an expanded version of Leslie *et al.* (2002b), which defined the mismatch kernel and presented results on the Jaakkola *et al.* dataset. Here, in addition to reporting experiments on the second benchmark dataset and comparing with the SVM-pairwise method, we present further results that give biological motivation for use of our method. First, we show that the mismatch kernel captures a biologically meaningful notion of sequence similarity by using the kernel values in an unsupervised setting for family-level and superfamily-level homology detection and for fold recognition. Second, we describe how to extract from a trained mismatch-SVM classifier the most informative subsequences in the training set. These sequences typically correspond to highly conserved, motif regions in the positive sequence class. Thus, while we do not use a multiple alignment or HMM as an input to our method, we can recover conservation information as an output of our training.

SVMS AND KERNELS

Support Vector Machines are a class of supervised learning algorithms first introduced by Vapnik (Boser *et al.*, 1992; Vapnik, 1998). Given a set of labeled training vectors $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, m$, where the $\mathbf{x}_i$ are real vectors and y_i are ± 1, training an SVM amounts to solving an optimization problem that determines a linear classification rule $f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$.

A test example $\mathbf{x}$ is then classified as positive (negative) if $f(\mathbf{x}) > 0$ [$f(\mathbf{x}) < 0$]. Such a classification rule corresponds to a hyperplane decision boundary between positive and negative points with normal vector $\mathbf{w}$ and bias term b.

A key feature of the SVM optimization problem is that it depends only on the inner products $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ of the training vectors, allowing us to use kernel techniques. To introduce a kernel, we now suppose that our training data are simply labeled examples (x_i, y_i), where the x_i could belong to a vector space or to a space of discrete structures like sequences of characters from an alphabet. Given any feature map Φ from the input space into a (possibly high-dimensional) vector space called the feature space, we obtain a kernel defined by $K(x, y) = \langle \Phi(x), \Phi(y) \rangle$. By replacing $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ by $K(x_i, x_j)$ in the dual SVM optimization problem, we implicitly train in feature space. Moreover, if we can directly and efficiently compute the kernel values $K(x, y)$ without explicitly representing the feature vectors, we gain tremendous computational advantage for high-dimensional feature spaces.

THE MISMATCH KERNEL

Below we introduce a class of string kernels, called mismatch kernels, that can be used on biological sequence data with various kernel-based machine learning methods. Our eventual goal is to use these kernels to train SVMs for protein classification: we outline the use of kernels in SVM training, describe how to compute the kernel efficiently, and apply our mismatch kernels in an SVM approach to remote homology detection.

In this section, we define the mismatch kernels and provide some intuition for why they capture meaningful sequence similarity information. For motivation, we interpret the mismatch value for a pair of protein sequences as a sequence similarity score, and as a simple application, we show that this score can be used directly for protein classification.

Mapping protein sequences to k-mer feature space

Recall that, given a number $k \geq 1$, the k-spectrum of a biological sequence is the set of all k-length (contiguous) subsequences that it contains; we refer to such a k-length subsequence as a k-mer. In order to capture significant information from sequence data without first building a generative model, we represent the spectrum information in a sequence as a vector in 'k-mer feature space'.

If we are dealing with sequences of characters from an alphabet $\mathcal{A}$ of size $|\mathcal{A}| = l$ ($l = 20$ for the alphabet of amino acids), we represent the sequences as vectors in an l^k-dimensional vector space, or feature space, where the coordinates are indexed by the set of all possible k-mers. For a very simple feature map, we can assign to a sequence x a vector given as follows: for each k-mer α, the coordinate indexed by α will be the number of times α occurs in x. This gives the

k-spectrum feature map defined in Leslie *et al.* (2002a):

$$\Phi_k(x) = (\phi_\alpha(x))_{\alpha \in \mathcal{A}^k},$$

where $\phi_\alpha(x) = $ no. of occurrences of α in x. Now the k-spectrum kernel $K(x, y)$ for two sequences x and y is obtained by taking the inner product in feature space:

$$K_k(x, y) = \langle \Phi_k(x), \Phi_k(y) \rangle.$$

Note that this kernel gives a simple notion of sequence similarity: two sequences will have a large k-spectrum kernel value if they share many of the same k-mers. One can extend this idea by taking weighted sums of k-spectrum kernels for different values of k, as described in (Vishwanathan and Smola, 2002).

For a more sensitive and biologically realistic kernel, we want to allow some degree of mismatching in our feature map. That is, we want the kernel value between two sequences x and y to be large if they share many similar k-mers. For a fixed k-mer $\alpha = a_1, a_2, \ldots, a_k$, with each a_i a character in $\mathcal{A}$, the (k, m)-pattern generated by α is the set of all k-length sequences β from $\mathcal{A}$ that differ from α by at most m mismatches. We denote this set by $N_{(k,m)}(\alpha)$, the 'mismatch neighborhood' around α. Intuitively, when we see an instance of a k-mer α in our input sequence x, we would like to contribute not only to the α-coordinate in feature space but also to all coordinates corresponding to k-mers in the mismatch neighborhood of α.

We can now define our feature map into the l^k-dimensional feature space, indexed as before by the set of all possible k-mers. If α is a fixed k-mer, we first define the $\Phi_{(k,m)}$ on α by:

$$\Phi_{(k,m)}(\alpha) = (\phi_\beta(\alpha))_{\beta \in \mathcal{A}^k},$$

where $\phi_\beta(\alpha) = 1$ if β belongs to $N_{(k,m)}(\alpha)$ [or equivalently, if α belongs to $N_{(k,m)}(\beta)$], and otherwise $\phi_\beta(\alpha) = 0$.

We define the feature map on an input sequence x in $\mathcal{X}$ is as the sum of the feature vectors for the k-mers in x:

$$\Phi_{(k,m)}(x) = \sum_{k\text{-mers } \alpha \text{ in } x} \Phi_{(k,m)}(\alpha). \tag{1}$$

Note that $\Phi_{(k,0)}$ coincides with the spectrum kernel feature map Φ_k defined above. The (k, m)-mismatch kernel is once again just the inner product in feature space[1]:

$$K_{(k,m)}(x, y) = \langle \Phi_{(k,m)}(x), \Phi_{(k,m)}(y) \rangle.$$

These kernels measure sequence similarity based on shared occurrences of (k, m)-patterns in the data: the kernel value $K_{(k,m)}(x, y)$ will be large if the sequences x and y share many k-length subsequences differing by at most m mismatches.

---

[1] Note that since we have defined $K_{(k,m)}$ from an explicit feature map, it is automatically a valid kernel function, and there is no need, e.g. to show that the conditions of Mercer's Theorem apply.

A simple application: ROC analysis of protein classification

As a simple application to demonstrate that our kernels capture useful biological information about sequence similarity, we use the spectrum kernel $K_k(x, y)$ and mismatch kernel $K_{(k,m)}(x, y)$ to derive distance measures between protein sequences x and y, and we use these kernel-derived distances for homology detection. A receiver operating characteristic (ROC) analysis shows favorable comparison with BLAST (Altschul *et al.*, 1997) for certain homology detection problems in terms of overall ROC score. Here, our purpose is simply to motivate our use of mismatch kernels in the SVM experiments described later in the paper; we do not claim that mismatch kernels are more appropriate or effective than BLAST for measuring sequence similarity.

We first normalize the kernels via

$$K(x, y) \leftarrow \frac{K(x, y)}{\sqrt{K(x, x)}\sqrt{K(y, y)}}. \tag{2}$$

We then consider the induced distance

$$d(x, y) = \sqrt{K(x, x) - 2K(x, y) + K(y, y)}$$

between pairs of protein sequences x and y. We wish to see whether the kernel-induced pairwise distance $d(x, y)$ can be used directly to predict whether x and y belong to the same protein family, superfamily or fold.

We evaluate the performance of this distance for homology detection and compare with the performance of Smith–Waterman and BLAST over the same dataset using default parameters. Sequences for these experiments were extracted from the SCOP (Murzin *et al.*, 1995) version 1.53 using the Astral database (http://astral.stanford.edu/, Brenner *et al.*, 2000), removing similar sequences using an E-value threshold of 10^{-25}. This procedure resulted in 4352 distinct sequences, grouped into families, superfamilies and folds. All pairwise E-values are computed by querying the database of 4352 iteratively with each sequence.

We perform separate experiments to detect family-level and superfamily-level homology and fold-membership using Smith–Waterman, BLAST, the spectrum kernel with $k = 4$ and the mismatch kernels with the following (k, m) parameters $(6, 1)$ and $(8, 2)$. In experiments using kernel-derived distances, we compute ROC curves by varying the distance threshold: a pair of homologous proteins that is below the threshold is considered a true positive, while a pair of non-homologous proteins below the threshold is considered a false positive. We also compute an ROC curve for Smith–Waterman and BLAST using E-value thresholds.

Figure 1 shows ROC curves for all methods in the superfamily task. ROC curves for the family and fold tasks are available on the supplemental data website for this paper (www.cs.columbia.edu/compbio/mismatch). Table 1 summarizes the ROC scores for all methods at each task. The
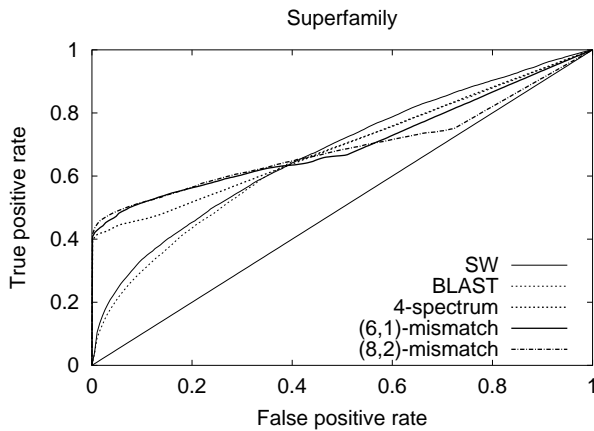
Superfamily

Fig. 1. ROC curves comparing mismatch kernels with Smith–Waterman (SW) and BLAST for unsupervised homology detection. 'True' relationships are defined by SCOP superfamilies.

Table 1. ROC scores for homology experiments comparing kernels with BLAST

	SW	BLAST	Spectrum $k=4$	Mismatch $k=6$, $m=1$	Mismatch $k=8$, $m=2$
Fold	0.713	0.619	0.623	0.623	0.621
Superfamily	0.679	0.660	0.702	0.704	0.695
Family	0.820	0.737	0.741	0.784	0.708

The ROC score is the area under the ROC curve.

ROC score is the area under the ROC curve: perfect discrimination between positives and negatives gives an ROC score of 1, while a random classifier has an expected score close to 0.5. We first note that in this unsupervised setting, neither the kernel-induced distances nor the E-values from BLAST and Smith–Waterman gave very strong performance in any of the homology/fold detection tasks. However, for family-level homology detection, the best mismatch kernel performs better than BLAST. For the superfamily homology and fold recognition tasks, the mismatch kernels outperform both BLAST and Smith–Waterman at low false positive rates, though Smith–Waterman is more sensitive at fold detection at higher false positive rates. In all tasks, the mismatch-induced distance gives stronger performance than the spectrum-induced distance at low false positive rates.

EFFICIENT COMPUTATION OF KERNEL MATRIX

Because our feature vectors are sparse vectors in a very high-dimensional feature space, we do not compute and store the feature vectors for our input sequences. Instead, we directly and efficiently compute the kernel matrix using a data structure called a mismatch tree.

Mismatch tree data structure

The mismatch tree data structure is similar to a trie or suffix tree (Gusfield, 1997). We use the mismatch tree to represent the feature space (the set of all k-mers) and to organize a lexical traversal of all instances of k-mers that occur (with mismatches) in the data. The entire kernel matrix is computed in one traversal of the tree. Our algorithm is similar to the approach presented in (Sagot, 1998; Pavesi *et al.*, 2001) for finding subsequence patterns that occur with mismatches. A related data structure was also used for sparse prediction trees (Eskin *et al.*, 2000; Pereira and Singer, 1999).

A (k, m)-mismatch tree is a rooted tree of depth k, where each internal node has 20 (more generally l) branches, each labeled with an amino acid (symbol from $\mathcal{A}$). A leaf node represents a fixed k-mer in our feature space, obtained by concatenating the branch symbols along the path from root to leaf. An internal node represents the prefix for those k-mer features that are its descendants in the tree. We perform a depth-first traversal of the data structure and store, at a node of depth d, n pointers to all substrings ('k-mer instances') from the sample dataset whose d-length prefixes are within m mismatches from the d-length prefix represented by the path down from the root; this set of substrings represents the valid instances of the d-length prefix in the data. We also keep track, for each valid instance, of how many mismatches it has when compared with the prefix. Note that the set of valid instances for a node is a subset of the set of valid instances for the parent of the node; when we descend from a parent to a child, each instance is either passed down (with 0 or 1 additional mismatch), or it is eliminated because it has exceeded m mismatches. When we encounter a node with an empty list of pointers (no valid occurrences of the current prefix), we do not need to search below it in the tree. When we reach a leaf node—corresponding to a particular feature k-mer α—we have pointers to all instance k-mers occurring in the source sequences that are up to m mismatches from α. Because for a source sequence x, the instances with mismatches of α in x—the k-mers in x belonging to $N_{(k,m)}(\alpha)$—are exactly the ones that contribute to the α-coordinate of the feature vector $\Phi(x)$, we can now sum the contributions of all instances occurring in each source sequence and update the kernel matrix entries $K(x,y)$ for those source sequences x and y having non-zero α-features. That is, if $n_\alpha(x)$ and $n_\alpha(y)$ are the number of instances (with mismatches) of k-mer α in x and y, respectively, at the leaf node, we perform the update $K(x,y) \leftarrow K(x,y) + n_\alpha(x) \cdot n_\alpha(y)$. Figure 2 gives an example of the mismatch tree traversal.

Efficiency of kernel computation

During the kernel computation, we need only search down paths corresponding to k-mers that occur (with mismatches) in the data. The number of k-mers within m mismatches of any given fixed k-mer is $p(k, m, l) = \sum_{i=0}^{m} \binom{k}{i}(l-1)^i = O(k^m l^m)$. Thus, the effective number of k-mer instances that
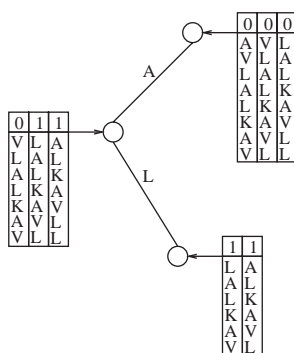
Fig. 2. An (8, 1)-mismatch tree for a sequence `AVLALKAVLL` used for computing the kernel matrix with k-mers of length 8 allowing 1 mismatch. The path from the root to the node is the 'prefix' of a particular k-mer feature. The leaf node stores the number of mismatches between the prefix of a k-mer instance and the prefix of a feature and a pointer to the tail of the k-mer. The figure shows the tree after expanding the path AL.

we need to traverse grows as $O(Nk^m l^m)$, where N is the total length of the sample data. If we are computing the kernel value $K(x, y)$ for a single pair of sequences, we make just one kernel update per leaf node that we reach in the traversal, and each k-mer instance that is counted at a leaf node is processed k times as it is passed down the path from the root. Thus to calculate a single kernel value, the complexity is $O[k^{m+1} l^m (|x| + |y|)]$.

To compute the kernel matrix for M sequences each of length n (total length $N = nM$), the traversal time is $O(nMk^{m+1} l^m)$, but the running time of the algorithm is dominated by the kernel updates made at the leaf nodes. If exactly c input sequences contain valid instances of the current k-mer at a particular leaf node, one performs c^2 updates to the kernel matrix. One can show that the worst case for the kernel computation occurs when the M feature vectors are all equal and have the maximal number of non-zero entries, so that the worst case running time for the leaf updates (and hence the full kernel computation) is $O[M^2np(k, m, l)] = O(M^2 nk^m l^m)$. Thus, the kernel computation worst case running time grows with m, the number of mismatches, at the same rate as the mismatch neighborhood and scales linearly with the length n of the input sequences. For remote homology detection, small values of m are most useful, and the kernel calculations are quite inexpensive.

Another advantage of the mismatch algorithm is its efficient use of memory, which also leads to faster running time in practice. Because we perform a depth-first traversal, the only expanded nodes are along the current search path, and thus there is a maximum of k stored nodes (counting the root node) at any time. In fact, the kernel computation can be achieved by a recursive function, without explicitly building and storing the tree.

FAST PREDICTION WITH SVMS

Mismatch kernels provide an additional advantage when used with SVMs: the particular form of the SVM solution combined with the definition of the mismatch feature map allow us to implement fast prediction on test sequences. The learned SVM classifier is given by

$$f(x) = \sum_{i=1}^{r} y_i a_i \langle \Phi_{(k,m)}(x_i), \Phi_{(k,m)}(x) \rangle + b,$$

where x_i are the training sequences that map to support vectors, y_i are labels (± 1), and a_i are weights obtained from the dual SVM optimization problem. Note that the classification function evaluated on the test sequence x is the sum of classification 'scores' $f(\alpha)$ for the k-mers α it contains. We can therefore precompute and store all the non-zero k-mer scores. Then the prediction $f(x)$ can be calculated in linear time [i.e. $O(|x|)$] by scanning through the k-mers in x and looking up the precomputed k-mer scores.

One way to compute the k-mer scores is to use two passes of the mismatch tree data structure. In the first pass, we compute the non-zero coordinates of the normal vector $w = \sum_{i=1}^{r} y_i a_i \Phi_{(k,m)}(x_i)$. We traverse the support sequences x_i, and at leaf node corresponding to k-mer β, we compute the weighted sum of valid instances to obtain the coordinate w_β of the normal vector. In the second pass, we use the normal coordinates w_β to obtain the k-mer scores. We traverse the set of k-mers β having non-zero w_β, and at leaf node corresponding to k-mer α, we compute the sum $\sum_{\beta \text{ in } N_{(k,m)}(\alpha)} w_\beta$ to obtain the score for α.

EXPERIMENTS: PROTEIN CLASSIFICATION
SCOP experiments with domain homologs

We first test the mismatch SVM method using the SCOP (Murzin *et al.*, 1995, version 1.37) datasets designed by Jaakkola *et al.* (2000) for the remote homology detection problem. In this test, remote homology is simulated by holding out all members of a target SCOP family from a given superfamily. Positive training examples are chosen from the remaining families in the same superfamily, and negative test and training examples are chosen from disjoint sets of folds outside the target family's fold. The held-out family members serve as positive test examples. Details of the datasets are available at www.soe.ucsc.edu/research/compbio/discriminative

Because Jaakkola *et al.* needed to train hidden Markov models for the protein families represented in the positive training sets for these experiments, they used the SAM-T98 algorithm to pull in domain homologs from the non-redundant protein database. These additional domain homologs were added to the dataset as positive examples in the experiments.

Because the test sets are designed for remote homology detection, we use small values of k. Without mismatches, the only reasonable values are $k = 3$ and $k = 4$, since $k \geq 5$
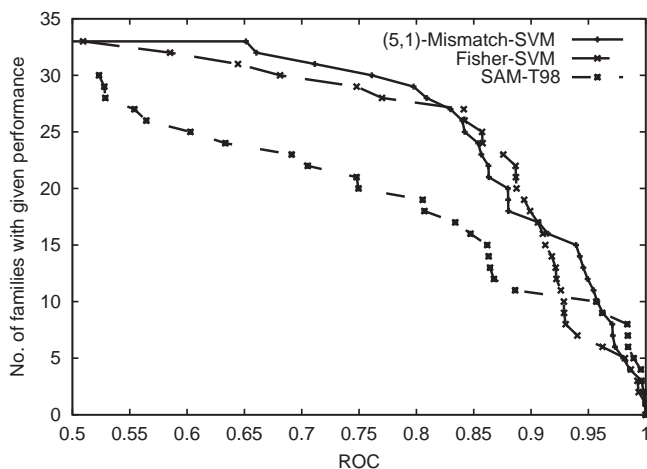
Fig. 3. Comparison of three homology detection methods for the SCOP 1.37 benchmark dataset. The graph plots the total number of families for which a given method exceeds an ROC score threshold. Each series corresponds to one of the homology detection methods described in the text.

Fig. 4. Family-by-family comparison of mismatch and Fisher kernels for SCOP 1.37 benchmark dataset. The coordinates of each point in the plots are the ROC scores for one SCOP family, obtained using the mismatch-SVM with $k = 5$, $m = 1$ (x-axis) and Fisher-SVM (y-axis). The dotted line is $y = x$.

results in a spectrum kernel matrix that is almost everywhere 0 off the diagonal (and $k < 3$ is not informative). For the mismatch kernel, we were therefore interested in slightly longer k and a very small number of mismatches for efficiency in training. We tested $(k, m) = (5, 1)$ and $(6, 1)$, where we normalized the kernel by Equation (2). Our results show that $(k, m) = (5, 1)$ yields slightly better performance, though results for both choices were similar. [Data for $(k, m) = (6, 1)$ not shown.] We use a publicly available SVM software implementation (http://microarray.cpmc.columbia.edu/gist/), which implements the soft margin optimization algorithm described in Jaakkola *et al.* (2000). Note that for this variant of the SVM optimization problem, the bias term b is fixed to 0.

For comparison, we include the original experimental results from Jaakkola *et al.* for two methods: the SAM-T98 iterative HMM, and the SVM-Fisher method. We note that, more recently, a newer version of the SAM HMM software has become available (Karplus *et al.*, 2001), modifications to the Fisher kernel have been explored (Tsuda *et al.*, 2002), and other novel approaches to homology detection have been introduced (Spang *et al.*, 2002). Figure 3 illustrates the mismatch kernel's performance relative to the profile HMM and SVM-Fisher homology detection methods. The figure includes results for all 33 SCOP families, and each series corresponds to one homology detection method. Qualitatively, the curves for SVM-Fisher and mismatch-SVM are quite similar. When we compare the overall performance of two methods using a two-tailed signed rank test (Henikoff and Henikoff, 1997; Salzberg, 1997), we find that almost none of the differences between methods are statistically significant. Using a p-value threshold of 0.05 and including a Bonferroni

adjustment to account for multiple comparisons, we find only the following significant differences: Fisher-SVM and mismatch-SVM perform better than SAM-T98 (with p-values $6.7e - 03$ and $1.3e - 02$, respectively). There is no statistically significant difference between the performance of Fisher-SVM and mismatch-SVM.

Figure 4 shows a family-by-family comparison of the ROC scores of the $(5, 1)$-mismatch-SVM and SVM-Fisher. The points fall approximately evenly above and below the diagonal, indicating little difference in performance between the two methods, though there are a handful of families on which SVM-Fisher has somewhat higher ROC scores.

Figure 5 shows the improvement provided by including mismatches in the SVM kernel. The figure plots ROC scores for two kernel SVM methods: using $k = 5$, $m = 1$ mismatch kernel, and using $k = 3$ spectrum kernel (no mismatches), the best-performing spectrum kernel reported in Leslie *et al.* (2002a). Most of the families perform better with mismatching than without, and in the cases where the 3-spectrum kernel wins over the $(5, 1)$-mismatch kernel, the difference is usually small.

SCOP experiments without domain homologs

In order to investigate the performance of our method against competing methods in the more difficult setting of very limited positive training data, we completed a similar set of SCOP experiments described in Liao and Noble (2002). Liao and Noble developed this benchmark dataset (available at www.cs.columbia.edu/compbio/svm-pairwise) to test their SVM-pairwise method, which uses explicit feature vectors
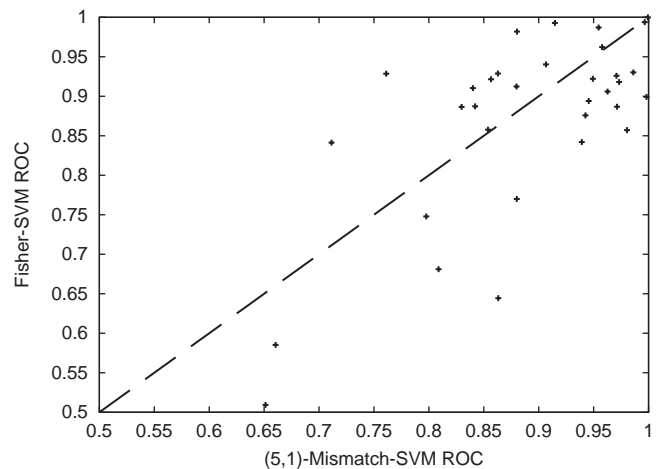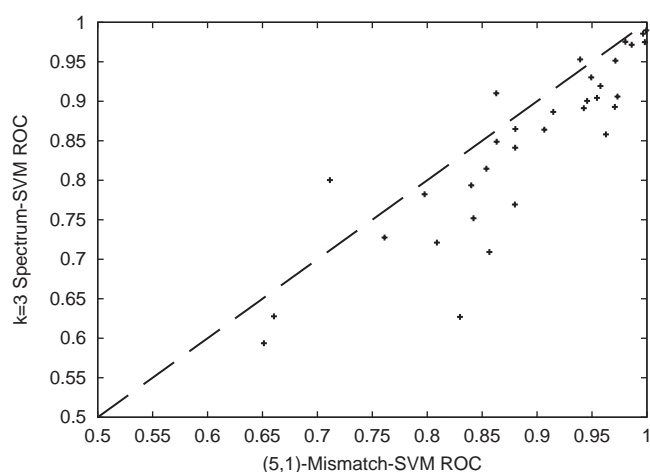
Fig. 5. Family-by-family comparisons mismatch and spectrum kernels for SCOP 1.37 benchmark dataset. The coordinates of each point in the plots are the ROC scores for one SCOP family, obtained using mismatch-SVM with $k = 5$, $m = 1$ (x-axis) and spectrum-SVM with $k = 3$ (y-axis). The dotted line is $y = x$.

Fig. 6. Comparison of homology detection methods for the SCOP 1.53 benchmark dataset. The graph plots the total number of families for which a given method exceeds an ROC score threshold for the second set of remote homology detection experiments. Each series corresponds to one of the homology detection methods described in the text.

based on Smith–Waterman alignment scores against the set of training sequences for SVM training.

These experiments are based on SCOP version 1.53 and also test for remote homology detection, but the positive training data is limited to the original SCOP sequences. In particular, no method has the benefit of training on additional homologs of the positive training sequences pulled in from a large unlabeled protein database. For this more difficult discrimination problem, Liao and Noble found that their SVM-pairwise method outperformed SVM-Fisher, SAM-T98 and PSI-BLAST; these methods could not successfully perform in the presence of so little positive training data (Fig. 6).

In our experiments, we find that the our SVM-mismatch method performs as well as SVM-pairwise on this dataset: Figure 7 shows the comparison of the (5,1)-mismatch kernel against SVM-pairwise. A signed-rank test confirms that there is no significant difference between performance. Note, however, that the mismatch method has the advantage that its kernel is fast and efficient to compute, while the SVM-pairwise method requires expensive computation of Smith–Waterman alignment scores for every pair of training sequences and between each training and test sequence (computation of each feature is quadratic in sequence length). Also, the computation of the kernel matrix from the Smith–Waterman matrix is $O(M^3)$ in the number of training examples. We did not test SVM-pairwise on the Jaakkola dataset because the large number of experiments and training sequences make computation of the SVM-pairwise kernel prohibitively expensive.

One can argue that this set of experiments is an unrealistic comparison of methods, because one would generally not attempt to use hidden Markov models, either directly
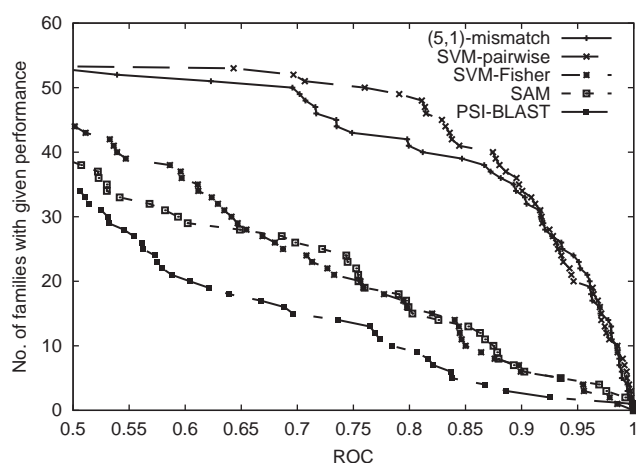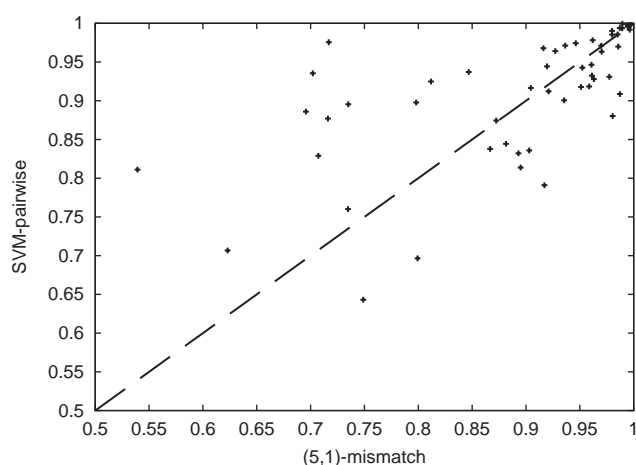
Fig. 7. Family-by-family comparison of (5, 1)-mismatch-SVM with SVM-pairwise for the SCOP 1.53 benchmark dataset. The coordinates of each point in the plot are the ROC scores for one SCOP family, obtained using the SVM-mismatch with $k = 5$, $m = 1$ (x-axis) and SVM-pairwise (y-axis). The dotted line is $y = x$.

or in the SVM-Fisher approach, that had been trained on a limited and divergent set of positive examples. However, we view these experiments as a demonstration of the power of our model-independent discriminative approach in a difficult classification problem with limited training data.

Extracting explanations for SVM predictions

In general, SVM classifiers do not easily yield explanations for the predictions that they make. In particular, when a kernel function is used to map the data into a high-dimensional

```
1cb1_4:78               saq..---ksp..aelksifek..yaakeg....dpnqlsk...eel..........kqliqaef.................p...........
SWN:CA22_HUMAN_116:179  snklh------.----fafrl..yd-ldk....d-ekisr...del..........lqvlrmmv.................gvnisdeqlgsi
SWN:S106_CHICK_4:82     maa..pldqai..gllvatfhk..ysgkeg....dknslsk...gel..........keliqkelti...............g...........
SW:BTV3_BETVE_140:197   sfggf------.----kvfd-..---edg....d-gyisa...rel..........qmvlgk-l.................g...........
SW:CA22_RAT_116:179     snklh------.----fafrl..yd-ldk....d-dkisr...del..........lqvlrmmv.................gvnisdeqlgsi


                                                                                                  *********
1cb1_4:78               sll...................k......g...................prtlddl....fqeldkn...gdgevSFEEFQvlvkkisq......
SWN:CA22_HUMAN_116:179  ---.................-......-...................---adrt....iqeadqd...gdsaiSFTEFVkvl-----ekvdv.
SWN:S106_CHICK_4:82     pkl...................k......d...................ae-iagl....medldrn...kdqeVNFQEyvtflgalamiynea.
SW:BTV3_BETVE_140:197   ---..................-fsegs.e...................idrvekm....ivsvdsn...rdGRVDFFEFkdmm-----rsvlv.
SW:CA22_RAT_116:179     ---.................-......-...................---adrt....iqeadqd...gdsaiSFTEFVkvl-----ekvdv.
```

Fig. 8. Mapping high-scoring k-mers onto a multiple alignment. The figure shows the first five sequences from a profile HMM multiple alignment of the EF-hand superfamily and its homologs. The 200 k-mers that receive the highest weights from a (5,1)-mismatch SVM have been mapped onto the multiple alignment. Occurrences of these k-mers are indicated by uppercase letters. Periods ('.') in the alignment correspond to insertions and hyphens ('-') correspond to deletions with respect to the HMM. The location of a motif-like region of clustered k-mer occurrences is indicated by asterisks.

feature space, the hyperplane decision boundary located by the SVM training algorithm is only implicitly defined by dot products between images of data points in the feature space. Most kernel functions make it very difficult to compute explicitly the location of this hyperplane.

However, the mismatch kernel differs from most kernel functions in that the mismatch kernel's mapping into feature space is explicit [as in Equation (1)]. Therefore, for this kernel, the normal of the separating hyperplane **w** can be represented directly as $\mathbf{w} = \sum_i y_i a_i \Phi_{(k,m)}(x_i)$, where each x_i is a training example, y_i its training label, and a_i its corresponding weight obtained by the SVM algorithm. Note that $a_i > 0$ only for a subset of training examples, which are referred to as support vectors. The location of the hyperplane **w** is thus given by a set of weights, one for each k-mer. Using the mismatch data structure we can efficiently compute these k-mer weights by summing the support vector weights of all instances at each leaf node.

A sequence that contains many k-mers with large, positive weights is very likely to be a member of the positive class defined by the classifier. Therefore, in the case of protein classification, we expect these highly informative k-mers to correspond to conserved regions in the protein family alignment. Figures 8 and 9 illustrate this phenomenon. We selected at random an experiment in which the SVM-mismatch method performed well from the first set of SCOP experiments (Jaakkola *et al.*, 2000); the positive set for the chosen experiment was the EF-hand superfamily. For this experiment, we computed the 200 k-mers that are assigned the highest weights by the (5,1)-mismatch SVM. Many of these k-mers are overlapping or slight variations of one another. The resulting set of 200 k-mers was then mapped onto the positive training set. In Figure 8, five sequences from the multiple alignment for training sequences from this experiment are shown in detail, with the high-scoring k-mers indicated by uppercase letters. In these five sequences, the high-scoring k-mers all appear within a nine-letter region of the alignment. Note that this is the only nine-letter block within the alignment that does not contain any insertions or deletions. Figure 9

Fig. 9. Schematic of k-mers mapped onto a multiple alignment. The figure shows the first 150 sequences from the same alignment as Figure 8, in schematic form. In the figure, insertions and deletions are white, unmatched amino acids are gray, and amino acids that map onto one of the top 200 k-mers are black. The entire alignment (not shown) contains 847 sequences.

shows a larger portion of the same alignment, in schematic form. Here, the motif-like nature of the clusters of k-mer occurrences is apparent as vertical black bars in the alignment.

This analysis shows that, although the mismatch SVM does not provide sequence alignments *per se*, the classifier can be used to identify sequence regions that are highly conserved in the positive class.

DISCUSSION

We have presented a class of string kernels that measure sequence similarity without requiring alignment and without depending upon a generative model. We have also presented a method for efficiently computing these kernels. For the remote

homology detection problem, we are encouraged that our discriminative approach—combining SVMs with the mismatch kernel—performs competitively in the SCOP experiments when compared with state-of-the-art methods. When we make the discrimination problem more difficult by providing only limited training data (in particular, when no domain homologs of positive training sequences are used), our method performs as well as SVM-pairwise, an approach that previously outperforms other known methods in this setting.

Our method has several advantages over other SVM approaches to protein classification, such as SVM-Fisher and SVM-pairwise. The SVM-Fisher method is appealing because it combines the rich biological information encoded in a profile hidden Markov model with the discriminative power of the SVM algorithm. However, one generally needs a lot of data or prior knowledge to train the hidden Markov model. In addition, because calculating the Fisher scores depends on dynamic programming (quadratic in sequence length for profile HMMs), in practice it is very expensive to compute the kernel matrix. Our mismatch kernel approach gives efficient kernel computation, linear time prediction and maintains good performance even when there is little training data. Hence, our method could be advantageous in a number of settings: for training on large datasets, where kernel evaluations are the bottleneck; in applications where fast prediction is important, such as web-based tools; and in cases where positive examples are scarce and a good model is unavailable.

The SVM-pairwise method has a different computational problem, in that the computation of the kernel matrix is cubic in the number of training sequences. Furthermore, when pairwise Smith–Waterman scores are used, the computation time of each of these scores depends quadratically on sequence length. It is possible to use a faster algorithm for the sequence comparisons and to use a smaller set of sequences for comparison, at some cost in accuracy (Liao and Noble, 2003). However, in addition to computational cost, the SVM-pairwise classifier does not offer any biological insight about the sequences that it classifies. By contrast, one can extract high-scoring k-mers from a trained SVM-mismatch classifier in order to look for discriminative motif regions in the positive sequence family.

Several authors have previously defined kernel functions over strings (Watkins, 1999; Haussler, 1999), and one of these string kernels has been applied to text classification (Lodhi *et al.*, 2000). However, the cost of computing each kernel entry $K(x, y)$ scales as $O(|x||y|)$ in the length of the input sequences, making them too slow for most biological applications. For the (k, m)-mismatch kernel, each kernel evaluation is $O(k^{m+1}l^m(|x| + |y|))$, where l is the size of the alphabet. Since small values of m, the number of mismatches, are most useful in biological applications, the worst case running time is feasible.

Many variations on the mismatch kernel can be defined to introduce other notions of inexact string matching. In Leslie

and Kuang (2003), kernels are obtained by allowing k-mers to occur with a restricted number of gaps, using features with wildcards, or using substitution probabilities to score inexact matches to a feature k-mer, rather than counting instances of k-mers with mismatches. We note that in the case of the exact-matching spectrum kernel, suffix trees have been used to eliminate the constant factor of k in the complexity, yielding $O(|x| + |y|)$ computation time (Vishwanathan and Smola, 2002). If such techniques could be extended to inexact matching kernels, we could hope for more computational savings.

Finally, in certain biological applications, the k-length subsequence features that are 'most significant' for discrimination can themselves be of biological interest. For such problems, it would be interesting to use the mismatch data structure for feature selection on the set of k-mer features, so that we identify a feature subset that both allows for accurate discrimination and gives biologically interesting information about the spectrum differences between positive and negative examples.

ACKNOWLEDGEMENTS

REFERENCES

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) A basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Altschul,S.F., Madden,T.L., Schaffer,A.A. Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Attwood,T.K., Beck,M.E., Flower,D.R., Scordis,P. and Selley,J.N. (1998) The prints protein fingerprint database in its fifth year. *Nucleic Acids Res.*, **26**, 304–308.

Bairoch,A. (1995) The prosite database, its status in 1995. *Nucleic Acids Res.*, **24**, 189–196.

Baldi,P., Chauvin,Y., Hunkapiller,T. and McClure,M.A. (1994) Hidden markov models of biological primary sequence information. *PNAS*, **91**, 1059–1063.

Boser,B.E., Guyon,I. and Vapnik,V. (1992) A training algorithm for optimal margin classifiers. In *Fifth Annual Workshop on Computational Learning Theory*, Pittsburg, PA, pp. 144–152. ACM Press.

Brenner,S.E., Koehl,P. and Levitt,M. (2000) The ASTRAL compendium for sequence and structure analysis. *Nucleic Acids Res.*, **28**, 254–256.

Eddy,S.R. (1995) Multiple alignment using hidden markov models. In *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, pp. 114–120.

Eskin,E., Grundy,W.N. and Singer,Y. (2000) Protein family classification using sparse Markov transducers. In *Proceedings*

of the Eighth International Conference on Intelligent Systems for Molecular Biology, AAAI Press, Menlo Park, CA, pp. 134–145.

Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: detection of distantly related proteins. *PNAS* **84**, 4355–4358.

Gusfield,D. (1997) *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*, Cambridge University Press, New York.

Haussler,D. (1999) Convolution kernels on discrete structure. *Technical Report*, UC Santa Cruz.

Henikoff,S. and Henikoff,J.G. (1997) Embedding strategies for effective use of information from multiple sequence alignments. *Protein Sci.*, **6**, 698–705.

Jaakkola,T., Diekhans,M. and Haussler,D. (2000) A discriminative framework for detecting remote protein homologies. *J. Comput. Biol.* **7**, 95–114.

Karplus,K., Karchin,R., Barrett,C., Tu,S., Cline,M., Diekhans,M., Grate,L., Casper,J. and Hughey,R. (2001) What is the value added by human intervention in protein structure prediction? *Proteins*, **45**, 86–91.

Krogh,A., Brown,M., Mian,I., Sjolander,K. and Haussler,D (1994) Hidden markov models in computational biology: applications to protein modeling. *J. Mol. Biol.*, **235**, 1501–1531.

Leslie,C., Eskin,E. and Noble,W.S. (2002a) The spectrum kernel: a string kernel for SVM protein classification. *Proc. Pac. Biocomput. Symp.*, **7**, 566–575.

Leslie,C., Eskin,E., Weston,J. and Noble,W.S. (2002b) Mismatch string kernels for SVM protein classification. *Neural Inform. Process. Sys.*, **15**, 1441–1448.

Leslie,C. and Kuang,R. (2003) Fast kernels for inexact string matching. *Conference on Learning Theory and Kernel Workshop*, 114–128.

Liao,L. and Noble,W.S. (2002) Combining pairwise sequence similarity and support vector machines for remote protein homology detection. In *Proceedings of the Sixth Annual International Conference on Computational Molecular Biology*, Washington, DC, pp. 225–232. ACM press.

Liao,L. and Noble,W.S. (2003) Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. *J. Comput. Biol.*, in press.

Lodhi,H., Shawe-Taylor,J., Cristianini,N. and Watkins,C. (2000) Text classification using string kernels. *Neural Inform. Process. Sys.*, **13**, 563–569.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.

Pavesi,G., Mauri,G. and Pesole,G. (2001) An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics*, **17**, S207–S214. Proceedings of the Ninth International Conference on Intelligent Systems for Molecular Biology.

Pereira,F. and Singer,Y. (1999) An efficient extension to mixture techniques for prediction and decision trees. *Machine Learning*, **36**, 183–199.

Sagot,M. (1998) Spelling approximate or repeated motifs using a suffix tree. *Lecture Notes in Computer Science*, **1380**, 111–127.

Salzberg,S.L. (1997) On comparing classifiers: pitfalls to avoid and a recommended approach. *Data Mining Knowl. Discov.*, **1**, 371–328.

Spang,R., Rehmsmeier,M. and Stoye,J. (2002) A novel approach to remote homology detection: jumping alignments. *J. Comput. Biol.*, **9**, 747–760.

Tsuda,K., Kawanabe,M., Rätsch,G., Sonnenburg,S. and Müller,K. (2002) A new discriminative kernel from probabilistic models. *Neural Comput.*, **14**, 2397–2414.

Vapnik,V.N. (1998) *Statistical Learning Theory*, Springer.

Vishwanathan,S.V.N. and Smola,A. (2002) Fast kernels for string and tree matching. *Neural Inform. Process. Sys.*, **15**, 569–576.

Waterman,M.S., Joyce,J. and Eggert,M. (1991) Phylogenetic analysis of DNA sequences. *Computer Alignment of Sequences*, pp.59–72. Oxford University Press.

Watkins,C. (1999) Dynamic alignment kernels. *Technical Report*, UL Royal Holloway.