BIOINFORMATICS

Vol. 21 Suppl. 1 2005, pages i359–i368
doi:10.1093/bioinformatics/bti1055

Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity

*S. Joshua Swamidass†, Jonathan Chen†, Jocelyne Bruand, Peter Phung, Liva Ralaivola and Pierre Baldi**

Institute for Genomics and Bioinformatics, School of Information and Computer Sciences, University of California, Irvine, CA, USA

Received on January 15, 2005; accepted on March 27, 2005

ABSTRACT

Motivation: Small molecules play a fundamental role in organic chemistry and biology. They can be used to probe biological systems and to discover new drugs and other useful compounds. As increasing numbers of large datasets of small molecules become available, it is necessary to develop computational methods that can deal with molecules of variable size and structure and predict their physical, chemical and biological properties.

Results: Here we develop several new classes of kernels for small molecules using their 1D, 2D and 3D representations. In 1D, we consider string kernels based on SMILES strings. In 2D, we introduce several similarity kernels based on conventional or generalized fingerprints. Generalized fingerprints are derived by counting in different ways subpaths contained in the graph of bonds, using depth-first searches. In 3D, we consider similarity measures between histograms of pairwise distances between atom classes. These kernels can be computed efficiently and are applied to problems of classification and prediction of mutagenicity, toxicity and anti-cancer activity on three publicly available datasets. The results derived using cross-validation methods are state-of-the-art. Tradeoffs between various kernels are briefly discussed.

Availability: Datasets available from http://www.igb.uci.edu/servers/servers.html

Contact: pfbaldi@ics.uci.edu

1 INTRODUCTION

Small molecules with at most a few dozen atoms play a fundamental role in organic chemistry and biology. They can be used as combinatorial building blocks for chemical synthesis, as molecular probes for perturbing and analyzing biological systems, and for the screening/design/discovery of new drugs, the majority of which are small molecules, and other useful compounds. As increasing numbers of datasets of small molecules become available, it becomes important to develop computational methods for the classification and analysis of small molecules and in particular for the prediction of their physical, chemical and biological properties. Such computational methods must be capable of dealing with data structures that go beyond the standard fixed-size vectorial representation to encompass molecules of variable structure and size. Here we develop kernels methods for small molecules and apply them to the prediction of mutagenicity, toxicity and anti-cancer activity for three publicly available datasets.

Two general classes of methods that have been proposed in the past to process variable-size structured data and applied to molecular structures are inductive logic programming (ILP) (Muggleton, 1992) and graphical models (Pearl, 1988; Lauritzen, 1996; Heckerman, 1998) together with the associated recursive neural networks (Micheli *et al.*, 2001, 2003; Baldi and Pollastri, 2003). Other, related approaches, not discussed here, include genetic algorithms (Koza, 1994) and stochastic grammars (Sakakibara *et al.*, 1994). The graphical model and even more so the ILP approach often suffer from computational complexity issues, and the recursive neural network approach requires rooting and orienting molecular graphs in ways that are not always natural or canonical. Over the last decade, kernel methods have emerged as a flexible class of machine learning methods capable of handling structured data. Kernel methods preserve the advantages of linear algorithms when modeling complex data characterized by non-linear properties (Cristianini and Shawe-Taylor, 2000; Schölkopf and Smola, 2002).

In what follows, for conciseness, we assume that the reader is already familiar with kernel methods. Intuitively a kernel defines a similarity measure between two molecules. Most of the kernels for discrete objects in the literature (Collins and Duffy, 2002; Leslie *et al.*, 2003; Lodhi *et al.*, 2002; Vert, 2002; Vishwanathan and Smola, 2003) are special cases of, or related to, convolution kernels (Haussler, 1999). Spectral kernels, in particular, are derived by (1) building feature vectors by counting occurrences of particular substructures (subsequences, subgraphs, etc.) and (2) by defining a similarity measure on these feature vectors. Convolution graph kernels published in the literature (Gärtner *et al.*, 2003; Kashima *et al.*, 2003; Mahé *et al.*, 2004) that can be applied to molecular graphs

*To whom correspondence should be addressed.

†These authors contributed equally to this work.

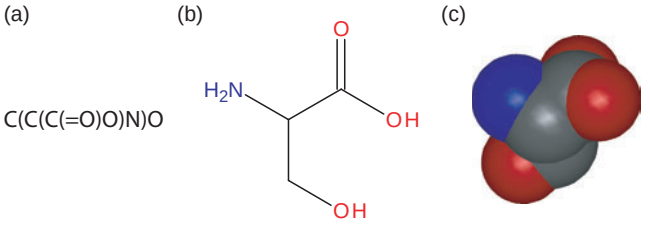

Fig. 1. Three representations for the amino acid Serine (Ser). (**a**) 1D SMILES string, (**b**) 2D representation with graph of bonds and (**c**) 3D space-filling model. A fourth representation based on fingerprint vectors is described in the text.

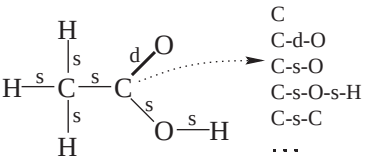

Fig. 2. Molecule represented as an undirected labeled graph. The labels on the vertices correspond to atom symbols and those on the edges describe the type of covalent bond between atoms (e.g. 's' for a single bond, 'd' for a double one). Also shown are examples of labeled paths of length 1 and 2 resulting from a depth-first search exploration of the graph, starting from one of the carbon atoms.

have several limitations, including high computational cost and/or inclusion of irrelevant or noisy substructures [referred to as 'totters' in Mahé *et al.* (2004)] at the expense of more relevant ones. Furthermore, these kernels often do not provide an easy avenue for the incorporation of chemists' background knowledge and experience and are not tuned to the specific properties of small molecule graphs.

Here we derive efficient spectral and other kernels for molecules by leveraging their multiple representations in the form of strings (1D), graphs of bonds (2D) and atom coordinates (3D) (Fig. 1). Although in some sense the three representations are equivalent, we will give particular emphasis to graph kernels associated with the 2D representation because of their novelty and because, in the current state of knowledge, this representation is the richest, least biased and most accurate, as will be explained in the Discussion. We will leverage the particular properties of small molecular graphs: these graphs are small both in terms of the number of vertices and edges—the average number of edges per node is typically only slightly above 2—and are highly constrained by the laws of chemistry.

2 METHODS

2.1 1D kernels based on SMILES strings

Small molecules can be represented in a unique way in the form of strings over a small alphabet, called SMILES strings (Weininger *et al.*, 1989; James *et al.*, 2004, http://www.daylight.com/dayhtml/doc/theory/theory.toc.html) (Fig. 1). Although SMILES strings require ordering the atoms of a molecule, they are widely utilized and are particularly useful in database organization and searches since each molecule can be associated with a unique canonical SMILES string. Because SMILES strings are sequences of letters, all the kernels that have been developed for sequences can be applied to SMILES strings. We thus propose applying sequence kernels to SMILES strings and testing their properties on the same set of problems. These kernels are usually spectral kernels counting the occurrence of all possible substrings of a certain length contained in a sequence. Variations allow for word mismatches and insertions (Leslie *et al.*, 2004).

2.2 2D kernels-based fingerprints

2.2.1 Traditional fingerprints Traditional fingerprints are bit vectors of a given size l, typically taken in the range 100–1000 (usually $l = 512$ or 1024). Given a molecule M with n atoms and m bonds, a corresponding fingerprint is derived using depth-first searches from each vertex. Thus the substructures being considered are labeled paths, which may include labeled cycles. A path may contain the same vertex twice, but not the same edge twice. Variations are obtained depending on whether two paths emanating from the same vertex are allowed to share edges or not once they have diverged. A hash value v is computed for each path described by the sequence of atoms and bonds visited (e.g. C–s–C–d–O, Fig. 2). For each such path, v is used to initialize a random number generator and b integers are produced (typically $b = 1$ or $b = 4$). The b integers are reduced modulo l and the corresponding bits are set to one in the fingerprint. If a bit in the nascent fingerprint is set to one by a path, it is left unchanged by all the other paths (i.e. '$1 + 1 = 1$' in a clash). An attractive feature of traditional fingerprint vectors is that, if the maximal path length d is set to $+\infty$, i.e. if we want to extract all the depth-first paths starting from all the atoms of a molecule, the complexity of the procedure is $O(nm)$ when paths do not share edges after divergence. In practice, d is often set to a lower value, typically in the range 8–10. Moreover, since fingerprinting is commonly used by chemists, typical useful values for l, b and d are readily available together with additional information, such as uninformative paths that can be discarded.

2.2.2 Generalized fingerprints based on paths In what follows we use traditional compact fingerprints but also expanded fingerprint bit vectors obtained in the same way but without collapsing them to a relatively short length l. Additional variations of the fingerprinting approach that we consider include examining all paths of length up to d as well as considering integer or real-valued vectors instead of bit vectors. For instance, one can use actual path counts instead of binary indicator variables. In this case, each component of the resulting vector corresponds to the number of times a particular path is encountered in the depth-first search explorations of

a molecule. It is also be possible to re-weight the vectors according to the TF–IDF weighting scheme (Salton, 1991) commonly used in text retrieval. In this case, a molecule can be viewed as a piece of text consisting of all the labeled paths of length up to d that can be retrieved by depth-first search explorations. Our preliminary experiments using the TF–IDF approach, however, did not yield significant improvements, and therefore it is not used here. An alternative to the TF–IDF scheme to preserve/enhance paths carrying the most relevant information for a given classification task is to consider a reduced set of paths selected according to the mutual information criterion (Yang and Pedersen, 1997; Dumais *et al.*, 1998). An interesting aspect of this approach, besides that of reducing path vocabulary size, is that the automatically extracted paths may be validated (or invalidated) by chemists.

Thus, to summarize the situation, let $\mathcal{P}(d)$ be the set of all possible atom–bond paths with a maximum of d bonds. Resorting to the depth-first search strategy, the feature map ϕ for a molecule $\mathbf{x}$ and a given depth d can be written as

$$\phi_d(\mathbf{x}) = (\phi_{\text{path}}(\mathbf{x}))_{\text{path}\in\mathcal{P}(d)},$$

where $\phi_{\text{path}}(\mathbf{x})$ is equal to 1 if at least one depth-first search of depth d starting from all the atoms of $\mathbf{x}$ produces the path 'path'. The feature map $\boldsymbol{\varphi}_d$, with the corresponding φ_{path}, counting the number of paths found, can be similarly defined. The feature map giving fixed-size vectors of size l corresponds to the particular feature map $\bar{\phi}_{d,l}$ given by

$$\bar{\phi}_{d,l}(\mathbf{x}) = (\phi_{\gamma_l(\text{path})}(\mathbf{x}))_{\text{path}\in\mathcal{P}(d)},$$

where $\gamma_l : \mathcal{P}(d) \to \{1,\ldots,l\}^b$ is a function mapping paths to a set of indices. Standard chemical fingerprints are a special case where the hash function, random generation and congruence operations are captured by the function $\phi_{\gamma_l(\text{path})}$.

2.2.3 Fingerprint similarity Using these feature maps or fingerprints, different kernels can be proposed by using different measures of similarity between fingerprints, including the straightforward inner product $k_d(\cdot,\cdot) = \langle\cdot,\cdot\rangle_d$ such that, for two molecules $\mathbf{u}, \mathbf{v}$,

$$k_d(\mathbf{u},\mathbf{v}) = \langle\mathbf{u},\mathbf{v}\rangle_d = \sum_{\text{path}\in\mathcal{P}(d)} \phi_{\text{path}}(\mathbf{u})\phi_{\text{path}}(\mathbf{v}),$$

and the corresponding kernel $\bar{k}_{d,l}$ is defined in the same way, using $\bar{\phi}_{d,l}$. Here we propose three normalized kernels that are closely related to what is commonly known as the Tanimoto similarity (Fligner *et al.*, 2002; Flower, 1998; Gower, 1971; Gower and Legendre, 1986).

Tanimoto kernel. Let $\mathbf{u}, \mathbf{v}$ be two molecules and d a positive integer. Consider the feature map ϕ_d and the corresponding kernel k_d. The Tanimoto kernel k_d^t is defined by

$$k_d^t(\mathbf{u},\mathbf{v}) = \frac{k_d(\mathbf{u},\mathbf{v})}{k_d(\mathbf{u},\mathbf{u}) + k_d(\mathbf{v},\mathbf{v}) - k_d(\mathbf{u},\mathbf{v})}. \quad (1)$$

If $\phi(\mathbf{u})$ is regarded as the set of features that can be extracted from $\mathbf{u}$ using depth-first search exploration, then k_d^t simply computes the ratio between $|\phi(\mathbf{u}) \cap \phi(\mathbf{v})|$, i.e. the number of elements in the intersection of the two sets $\phi(\mathbf{u})$ and $\phi(\mathbf{v})$, and $|\phi(\mathbf{u}) \cup \phi(\mathbf{v})|$, i.e. the number of elements in the set corresponding to the union of $\phi(\mathbf{u})$ and $\phi(\mathbf{v})$.

We note that, if the feature map used is $\bar{\phi}_{d,l}$, for a given $l \in \mathbb{N}$, instead of ϕ_d—and thus, $\bar{k}_{d,l}$ instead of k_d—in the above formula, the corresponding kernel $\bar{k}_{d,l}^t$ is exactly the Tanimoto similarity measure that chemists use for fast molecular comparison and retrieval with fixed-size bit vectors of size l.

MinMax kernel. Let $\mathbf{u}, \mathbf{v}$ be two molecules and d a positive integer. Consider the feature map $\boldsymbol{\varphi}_d(\cdot)$ and the corresponding $\varphi_{\text{path}}(\cdot)$. The MinMax kernel k_d^m is defined by

$$k_d^m(\mathbf{u},\mathbf{v}) = \frac{\sum_{\text{path}\in\mathcal{P}(d)} \min(\varphi_{\text{path}}(\mathbf{u}), \varphi_{\text{path}}(\mathbf{v}))}{\sum_{\text{path}\in\mathcal{P}(d)} \max(\varphi_{\text{path}}(\mathbf{u}), \varphi_{\text{path}}(\mathbf{v}))}. \quad (2)$$

This kernel function is closely related to the Tanimoto kernel in two different ways. First, it is identical to the Tanimoto kernel when applied to binary vectors. Second, in a more subtle way, the MinMax kernel can be viewed as a Tanimoto kernel on a different set of binary vectors obtained by transforming the vector of counts. More precisely, for a given d, consider the two integer-valued feature vectors $\boldsymbol{\varphi}_d(\mathbf{u})$ and $\boldsymbol{\varphi}_d(\mathbf{v})$ and an integer q larger than any count in $\boldsymbol{\varphi}_d(\mathbf{u})$ and $\boldsymbol{\varphi}_d(\mathbf{v})$. As we deal with sets of relatively small-sized molecules, a convenient value of q can be found easily. Assume that $p = |\mathcal{P}(d)|$. If $\boldsymbol{\varphi}_d(\mathbf{u})$ and $\boldsymbol{\varphi}_d(\mathbf{v})$ are expanded to the two binary feature vectors $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{v}}$ of size pq such that $\tilde{u}_i$ (resp. $\tilde{v}_{i'}$) is set to one if and only if $i \bmod q < \tilde{u}_i$ (resp. $i' \bmod q < \tilde{v}_{i'}$), then (Fig. 3)

$$k_d^m(\mathbf{u},\mathbf{v}) = \frac{\langle\tilde{\mathbf{u}},\tilde{\mathbf{v}}\rangle}{\langle\tilde{\mathbf{u}},\tilde{\mathbf{u}}\rangle + \langle\tilde{\mathbf{v}},\tilde{\mathbf{v}}\rangle - \langle\tilde{\mathbf{u}},\tilde{\mathbf{v}}\rangle}.$$

This similarity measure has the nice property that it can take into account the actual frequencies of the different paths in a molecule while still being strongly related to the Tanimoto kernel and still taking values between 0 and 1. Using path counts, this kernel produces a more reliable way to assess the similarity between molecules of different sizes (see Results). In addition, we also tested a hybrid kernel which include information about common absent paths (Fligner *et al.*, 2002).

Definition Hybrid kernel. Let $\mathbf{u}$ and $\mathbf{v}$ be two molecules. Let d and l be two positive integers, and θ a real number in the interval $[-1,+2]$. The Hybrid kernel k_d^h between $\mathbf{u}$ and $\mathbf{v}$ is defined by

$$k_{d,l}^h(\mathbf{u},\mathbf{v}) = \tfrac{1}{3}\left[(2-\theta)\cdot\bar{k}_{d,l}^t(\mathbf{u},\mathbf{v}) + (1+\theta)\cdot\neg\bar{k}_{d,l}^t(\mathbf{u},\mathbf{v})\right], \quad (3)$$

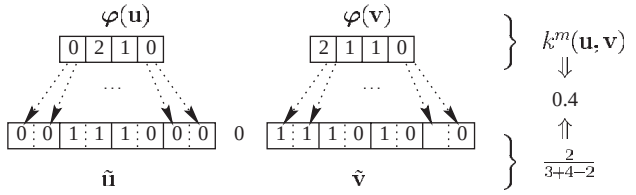

Fig. 3. Connection between the Tanimoto and MinMax kernels. If the feature vectors $\boldsymbol{\varphi}(\mathbf{u})$ and $\boldsymbol{\varphi}(\mathbf{v})$ are transformed into the bit vectors $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{v}}$, then $k_d^m(\mathbf{u},\mathbf{v})$ is exactly the ratio between the number of bits set to one both in $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{v}}$ divided by the total number of (unique) bits set to one in $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{v}}$. Given a large integer q, each path in $\mathcal{P}(d)$ is associated with a distinct set of q consecutive indices. If $\varphi_{\text{path}}(\mathbf{u}) > 0$, then $\varphi_{\text{path}}(\mathbf{u})$ consecutive bits are set to one in $\tilde{\mathbf{u}}$ starting at the indices corresponding to 'path'. The same holds for $\mathbf{v}$ and $\tilde{\mathbf{v}}$.

where $\neg\bar{k}^t_{d,l}$ is the kernel based on the feature map $(\neg\bar{\phi}_{d,l}(\mathbf{x})) = (\neg\bar{\phi}_{\gamma_l(\text{path})}(\mathbf{x}))_{\text{path}\in\mathcal{P}(d)}$, where $\neg$ is a logical 'not' and $\gamma : \mathcal{P}(d) \to 1, \ldots, l$. Thus the Hybrid kernel is a convex combination of two kernels, respectively measuring the number of common paths and common absent-paths between two molecules. When $\theta = -1$, the Hybrid kernel reduces to the Tanimoto kernel. In practice, θ is typically set to the average density of the bit vectors, which is in the interval [0,1]. It should be clear that a hybrid version of the MinMax kernel is possible along the same lines. However, in simulations the hybrid kernel did not yield to significant improvements and therefore is mentioned only briefly in what follows.

PROPOSITION 1. *The Tanimoto kernel, MinMax kernel and Hybrid kernel are Mercer kernels.*

PROOF (SKETCH). The proof that k^t_d, k^m_d, $k^h_{d,l}$ are Mercer kernels follows from a result given by Gower (1971) showing that, for any integer p and any set of ℓ binary vectors $\mathbf{x}_1, \ldots, \mathbf{x}_\ell \in \mathbb{R}^p$, the similarity matrix $S = (k^t(\mathbf{x}_i, \mathbf{x}_j)_{1\le i,j\le\ell}$ is positive semi-definite; thus k^t and k^t_d are positive definite kernels or Mercer kernels. Given that, for any Mercer kernel $k \in \mathbb{R}^{\chi\times\chi}$ and any mapping $g \in \chi^{\chi'}$, $k(g(\cdot), g(\cdot)) \in \mathbb{R}^{\chi'\times\chi'}$ is a Mercer kernel [see, for instance Schölkopf and Smola (2002)], the MinMax kernel is also a Mercer kernel. Finally, using the same argument, and the fact that a convex combination of Mercer kernels is a Mercer kernel, it follows that the Hybrid kernel is also a Mercer kernel.

2.2.4 Fast computation of 2D kernels At first glance, the computations of these kernels may appear prohibitive as the feature vectors produced by the feature map are of large dimension. However, using a suffix tree data structure (Ukkonen, 1995; Weiner, 1973; as also proposed in Leslie *et al.*, 2002; Vishwanathan and Smola, 2003) allows us to compute each of the proposed kernels in time $O(d(n_1m_1+n_2m_2))$, where d is the depth of the search and n_i (resp. m_i) is the number of vertices/atoms (resp. edges/bonds) of the two molecules considered. The complexity of computing kernels based on all the paths of depth d, allowing paths emanating from the same vertex to share edges once they have diverged, is not much higher, since computing all the paths of length d has complexity $O(n\alpha^d)$, where the branching factor α is typically only slightly above 1. In short, these kernels and their variants can be computed very rapidly and hence can be used to tackle large-scale chemical classification problems.

Table 1. Leave-one-out accuracy (%) results for the Mutag and PTC datasets using different kernels

Kernel/Method	Mutag	MM	FM	MR	FR
PD Kashima *et al.* (2003)	**89.1**	61.0	61.0	62.8	66.7
MK Kashima *et al.* (2003)	85.1	64.3	63.4	58.4	66.1
1D SMILES spectral	84.0	66.1	61.3	57.3	66.1
1D SMILES + variants spectral	85.6	**66.4**	63.0	57.6	**67.0**
2D Tanimoto	*87.8*	**66.4**	*64.2*	63.7	66.7
2D MinMax	86.2	64.0	**64.5**	*64.5*	66.4
2D Tanimoto, $l = 1024, b = 1$	87.2	*66.1*	62.4	**65.7**	*66.9*
2D Hybrid $l = 1024, b = 1$	87.2	65.2	61.9	64.2	65.8
2D Tanimoto, $l = 512, b = 1$	84.6	**66.4**	59.9	59.9	66.1
2D Hybrid $l = 512, b = 1$	86.7	65.2	61.0	60.7	64.7
2D Tanimoto, $l = 1024$ + MI	84.6	63.1	63.0	61.9	66.7
2D Hybrid $l = 1024$ + MI	84.6	62.8	63.7	61.9	65.5
2D Tanimoto, $l = 512$ + MI	85.6	60.1	61.0	61.3	62.4
2D Hybrid $l = 512$ + MI	86.2	63.7	62.7	62.2	64.4
3D Histogram + Gaussian	81.9	59.8	61.0	60.8	64.4

b denotes the number of bits set to one for a given path and *MI* indicates that paths have been selected using the mutual information criterion. Depth of search is set to $d = 10$. The value of θ used for the Hybrid kernel is the average density of the fingerprints contained in the training set. Best results are in bold face and second best are italicized.

2.3 3D kernels based on atomic distances

Finally, small molecules are 3D objects described by the xyz coordinates of each atom. Programs exist, such as CORINA (Sadowski *et al.*, 1994; Gasteiger *et al.*, 1996), that can derive fairly accurate 3D coordinates for small molecules starting from their 1D or 2D representations. We propose representing a molecule as a small set of histograms, one histogram per pair of atom labels. For instance, the CC histogram is the histogram of distances between all pairs of carbon atoms contained in a given molecule. Likewise, the CO histogram is the histogram of the distances between all pairs of carbon and oxygen atoms in the molecule, and so forth. Similarity between molecules is then measured by measuring similarity between histograms, which can be done in different ways. One simple approach is to use the sum of squared differences between histogram bins (squared Euclidean distance). This squared Euclidean distance v^2 can then be used to form a normalized kernel of the form $\exp[-v^2/\lambda^2]$. Variations on this idea are possible, for instance by introducing different weights for different histogram types (CC, CO, etc).

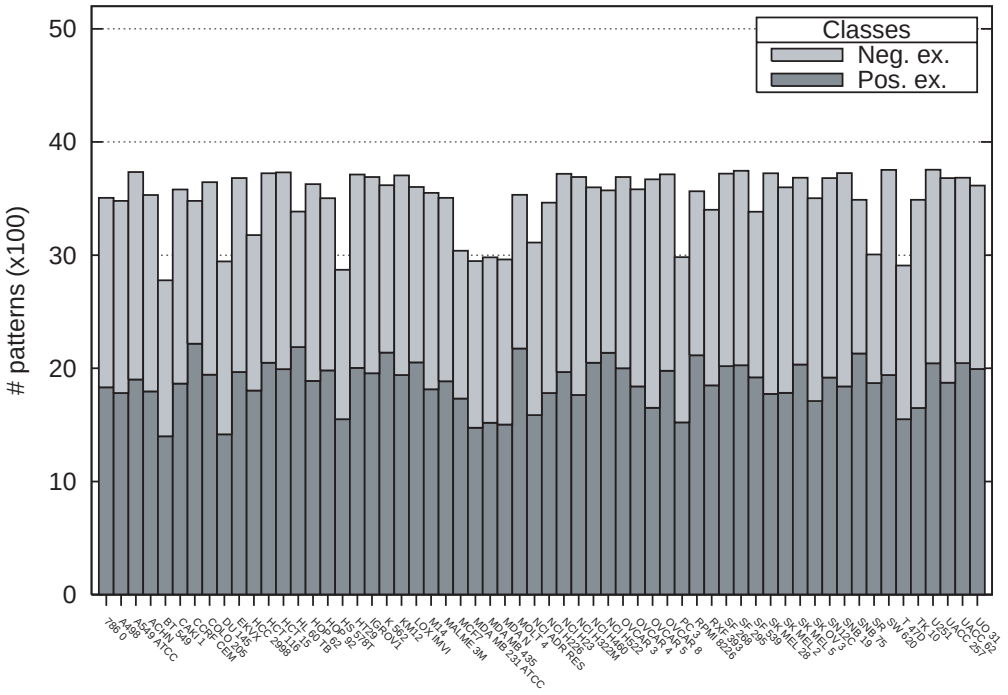

Fig. 4. Distribution of the 60 NCI screens. Positive examples correspond to cancer inhibition.

3 DATA

3.1 Mutag dataset

The Mutag dataset (Debnath *et al.*, 1991) consists of 230 chemical compounds along with information indicating whether they have mutagenicity in *Salmonella typhimurium*. Among the 230 compounds, however, only 188 (125 positive examples and 63 negative) are considered to be learnable (Debnath *et al.*, 1991) and thus are used in our simulations. The results from other groups, including those of Kashima *et al.* (2003), reported in Table 1, were obtained on the same subset of 188 molecules. The accuracy reported in the simulations—which assesses the ability of a classifier to assign the correct label to a molecule—is estimated through a leave-one-out procedure.

3.2 Predictive Toxicology Challenge dataset

The Predictive Toxicology Challenge (PTC) dataset (Helma *et al.*, 2001) reports the carcinogenicity of several hundred chemical compounds for male mice (MM), female mice (FM), male rats (MR) and female rats (FR). As with the Mutag dataset, the accuracies reported in Table 1 are estimated through a leave-one-out procedure.

3.3 NCI dataset

The Mutag and PTC datasets are useful but somewhat small. The NCI dataset, made publicly available by the National Cancer Institute (NCI), provides screening results for the ability of ∼70 000 compounds to suppress or inhibit the growth of a panel of 60 human tumor cell lines. We use the dataset corresponding to the concentration parameter GI50, essentially the concentration that causes 50% growth inhibition.[1] For each cell line, ∼3500 compounds, described by their 2D structures, are provided together with information on their cancer-inhibiting action. The distributions of positive examples (associated with cancer inhibition) and negative examples for the 60 cell lines are reported in Figure 4. Not only is the NCI dataset considerably larger than the Mutag and PTC datasets, but overall it is also more balanced. Thus the trivial background statistical predictor always predicting the class encountered most frequently has poorer performance on the NCI dataset, e.g. close to 50%. Performance on the NCI dataset is analyzed by cross-validation methods using 20 random 80/20 training/test splits of each subset. Values reported are averaged across these 20 splits.

4 RESULTS

We conducted several experiments to compare the various classes of kernels using different parameter settings. Here we report representative subsets of results together with the main findings. In addition to the usual accuracy measure, we also report the ROC score, which is the normalized area under the ROC curve plotting the number of true positive (TP) predictions as a function of the number of false positive

[1] A complete description of the cell lines is available at http://dtp.nci.nih.gov/docs/misc/common_files/cell_list.html

(FP) predictions for varying classification thresholds (Hanley and McNeil, 1982). Precision [TP/(TP + FP)] and recall [TP/(TP + FN)] (FN = false negative) measures are also computed, together with their harmonic mean (F-measure).

4.1 Mutag and PTC datasets

Results on the Mutag and PTC datasets obtained using 1D, 2D and 3D kernels with various parameter settings are reported in Table 1. In this table, we also report the results obtained by Kashima *et al.* (2003) using their marginalized kernels and a frequent pattern mining approach (Kramer and De Raedt, 2001). Overall, the three classes of kernels—1D, 2D and 3D—introduced here produce good results, well above chance level and comparable to, or better than, the state-of-the-art in the field. In general, the 2D kernels seem to perform best on these tasks. The 1D SMILES kernels are not far behind the 2D kernels and both seem to perform better than the 3D kernels. The best 1D kernels results are obtained using spectral sequence kernels but adding five SMILES string variants to each molecule by randomly selecting different starting points (SMILES strings impose an ordering on the atoms of a molecule and a starting point but the user can force different starting points and obtain slightly different SMILES strings for the same molecule). In Table 1, we see also that using the 2D kernels with the traditional molecular fingerprinting procedure, i.e. using fixed-size bit vectors of length l, still allows us to obtain acceptable results. In general, however, these results are not as good as those obtained with unbounded l and the deterioration in performance is accentuated when l is reduced from 1024 to 512 (except for MM), most likely as a result of the increase in the number of clashes.

Among the 2D kernels, the Tanimoto and MinMax kernels for depth $d = 10$ always rank among the top two methods and produce results that are consistently above those reported previously in the literature. The only exception is on the Mutag dataset, where the PD (Pattern Discovery) algorithm achieves 89.1% accuracy versus 87.9% for the Tanimoto kernel. Performances of up to 89.4% have actually been reported in King *et al.* (1996). Such a difference, however, may not be significant, because the Mutag dataset is too small. This is confirmed by further refining the 2D kernels and implementing an exhaustive search of all possible paths up to depth d, which can still be efficiently implemented owing to the small size and small degree of these graphs (the average degree for the NCI dataset, for instance, is 2.11). The corresponding results in Table 2 show, for example, that the MinMax kernel achieves a cross-validated performance accuracy of 91.5%, higher than all previously reported results (King *et al.*, 1996; Mahé *et al.*, 2004).

We also conducted tests using a mutual information criterion (Dumais *et al.*, 1998) to select informative paths. We computed the mutual information between the binary (0–1) variable associated with a given path and the binary (± 1) variable associated with the class. We ranked the paths accordingly and removed those paths with low mutual information. Surprisingly, for vectors of equal size, the results with mutual information are not better than those obtained using the simple limited-size fingerprint approach. We conjecture that this is a size effect (the set of informative paths that is retained is too small) and that a difference would become detectable with a larger value of l.

Table 2. Classification accuracies (%) obtained on Mutag using exhaustive paths extraction and the Tanimoto and the MinMax kernel with a depth set to $d = 10$

Kernel	Type	Atom #	Val.
Tanimoto	87.8	90.4	90.4
Tanimoto + cycle	86.2	87.8	89.9
MinMax	89.4	91.0	**91.5**
MinMax + cycle	89.9	**91.5**	89.4

'Type' corresponds to the case where atom descriptions are fully retained (e.g. in the paths constructed, a carbon connected to two hydrogens is different from one connected to three), 'Atom #' corresponds to the equivalence class where only the atomic numbers are used to label atoms and 'Val.' to the situation where all atoms having the same valence are considered equivalent. The 'cycle' option refers to different ways of including and counting cycles in molecular graphs.

4.2 NCI dataset

The trends observed on the Mutag and PTC dataset were confirmed on the larger NCI dataset. For conciseness, here we report primarily the results obtained with the 2D Tanimoto and MinMax kernels since these gave the best preliminary results on the Mutag and PTC datasets. The results for the 60 screens of the NCI datasets are plotted in Figures 5 and 6 using Tanimoto and MinMax kernels of depth $d = 10$. Figure 5 also contains results for 1D SMILES and 3D histogram kernels. Figure 5 plots the results in terms of accuracy and ROC score, and Figure 6 plots the results in terms of precision and recall. The mean accuracy for the Tanimoto kernel is 71.55% versus 72.29% for the MinMax kernel. Likewise, the mean ROC score is 77.86% versus 78.74%, the mean precision is 72.55% versus 73.02% and the mean recall is 74.90% versus 76.05%.

On the NCI dataset, we observe that these two 2D kernels have performance accuracies >70% in general, well above chance level for this balanced set. The MinMax kernel gives better results than the Tanimoto kernel almost consistently, albeit by a fairly small margin. We believe this results from the fact that this kernel uses actual counts rather than binary indicator variables. Using counts in the future ought to provide a richer representation and lead to better precision in retrieval, comparison and classification of molecules.

5 DISCUSSION

We have developed several classes of efficient kernels for small molecules by extracting corresponding feature vectors

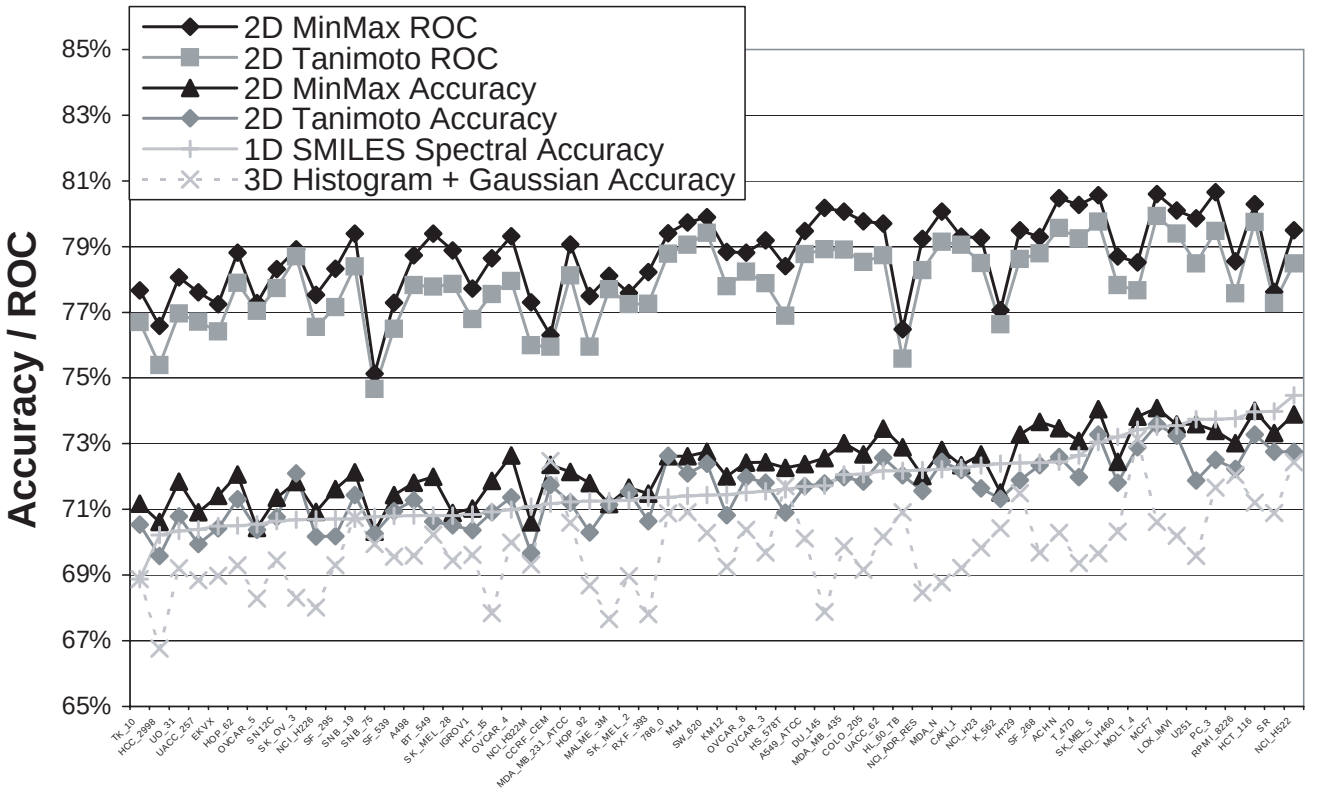

Fig. 5. Classification accuracy and ROC score obtained on the NCI problem using k^t_{10}, k^m_{10}, SMILES spectral and 3D histogram kernels. For readability purposes, all results are displayed in increasing order of accuracy for the 1D SMILES kernel, and ROC curves for the 1D SMILES and 3D histogram kernels are omitted.

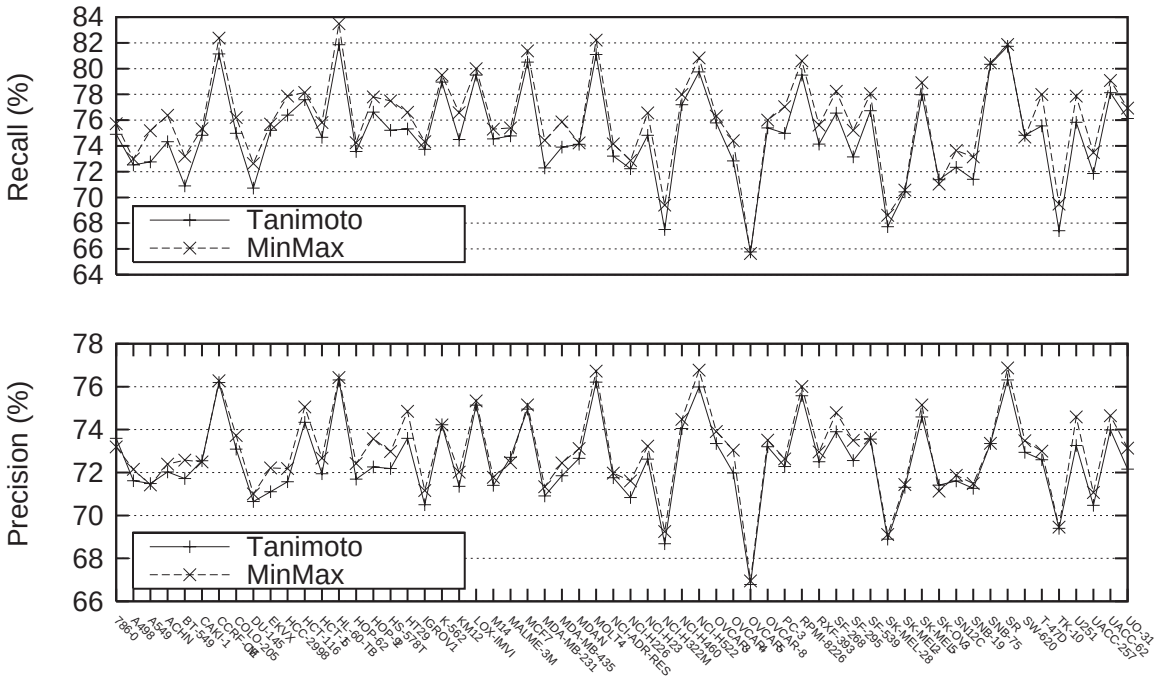

Fig. 6. Precision and recall obtained on the 60 screens using k^t_{10} and k^m_{10}.

from their 1D, 2D and 3D representations. We have shown on several empirical datasets that with the proper tuning these kernels yield state-of-the-art results that are comparable or superior to those previously published in the literature. Although we do not claim to have conducted exhaustive comparisons, and ultimately different kernels may be more appropriate for different tasks, current evidence suggests that, of all the kernel classes introduced, the 2D kernels may have a slight advantage. This may seem surprising in light of the fact that, for instance, the 1D and 2D representations ought to contain the same implicit information. However, 1D kernels may suffer from the arbitrariness of the vertex ordering imposed during the construction of SMILES strings. Furthermore, simple k-mers made up of contiguous letters may not capture well the branching patterns of molecular structures. Inclusion of other parsing substructures, such as trees, may lead to better SMILES convolution kernels. In contrast, 3D kernels may suffer from the loss of information introduced by the histograms and the fact that the 3D coordinates are predicted using the CORINA program. Although the prediction of 3D coordinates is much easier for small molecules than for large molecules, such as proteins, and is believed to be fairly reliable, one cannot rule out the possibility that the noise introduced by such predictions weakens the quality of the corresponding kernels.

The accuracies obtained, for instance in the region of 72% on the NCI dataset, are very encouraging, but clearly there is room and hope for improvement in many directions. Even the better performing 2D path kernels discard information, for instance regarding the location of the paths (phases) or the handedness of the molecules. Thus there is room for developing, testing and combining new 1D, 2D and 3D kernels and other kernels, such as Fisher kernels (Jaakkola *et al.*, 1999). For instance, we are currently exploring fingerprints based on counting shallow trees, rather than paths, using depth-first or breadth-first searches combined with efficient methods for tree comparison (Vishwanathan and Smola, 2003). And in applications where molecular surfaces are the most important, '2.5D' kernels may be developed to characterize molecular surfaces. Another obvious direction of research is to apply these kernels to regression problems in chemistry to predict, for instance, physical properties (e.g. the boiling point of alkanes) or pharmacological properties (e.g. the quantitative structure–activity relationship of benzodiazepines) (Cherqaoui and Villemin, 1994; Hadjipavlou-Litina and Hansch, 1994; Micheli *et al.*, 2003).

Although chemical toxicity and activity in general can vary significantly by stereochemical isomers of the same chemical, none of the methods described so far actually considers stereochemistry. To address this problem, 1D SMILES kernels can be augmented by using isomeric SMILES, which include additional special characters to describe stereocenters. For 2D kernels, we can use additional atom-type labels indicating whether an atom is a stereocenter and, if so, which configuration it is in. Likewise, we can construct other 3D kernels that are sensitive to stereochemical properties. However, the particular datasets used in this study do not include stereochemistry specifications and, thus, such extensions could not add any value to the analyses presented here. In this regard, the annotation and public availability of larger datasets with sterochemistry information would be invaluable for the development of better predictive machine learning methods in chemistry. Finally, it is worth noting also that 'toxicity' is an extremely complex and multi-faceted concept that may vary with multiple dimensions, such as organism (e.g. yeast versus human), genome (wild-type versus mutant), environment and so forth. In time, more specialized datasets and predictors are likely to be developed.

The penetration of computational, artificial intelligence and machine learning methods in chemistry has been slower than in physics or biology for many historical and sociological reasons, including the single-investigator nature of chemical research and the dominance of high-throughput projects in physics and biology, such as the human genome project. However, large datasets of chemical information are progressively becoming publicly available and large training sets could be derived over time for a variety of problems. Overall, the results obtained suggest that automatic classification of compounds may become a viable alternative to, but not a substitute for, slower and more expensive experimental characterization in the near future. Furthermore, even if the performance of individual computational filters is not perfect, batteries of such filters could be assembled for efficient molecular screening and discovery protocols. We hope that developing efficient kernels and other machine learning methods for molecular structures will help to address some of the outstanding problems in the field and to better understand what remains a rather sparsely explored chemical space. With current estimates of a universe containing at least 10^{60} potential small molecules, of which only $\sim 10^{7}$ have been discovered or synthesized, computational methods are likely to become a major tool of chemical astronomers.

ACKNOWLEDGMENT

Work supported by NIH (LM-07443-01) and NSF (EIA-0321390) grants to P.B., by the UCI Medical Scientist Training Program, and by a Harvey Fellowship to S.J.S.

REFERENCES

Baldi,P. and Pollastri,G. (2003) The principled design of large-scale recursive neural network architectures–DAG-RNNs and the protein structure prediction problem. *J. Mach. Learn. Res.*, **4**, 575–602.

Cherqaoui,D. and Villemin,D. (1994) Use of neural network to determine the boiling point of alkanes. *J. Chem. Soc. Faraday Trans.*, **90**, 97–102.

Collins,M. and Duffy,N. (2002) Convolution Kernels for Natural Language. *Adv. in Neural Information Processing Systems*, **14**, 625–632.

Cristianini,N. and Shawe-Taylor,J. (2000) *An Introduction to Support Vector Machines and other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge, UK.

Debnath,A.K., Lopez de Compadre,R.L., Debnath,G., Shusterman,A.J. and Hansch,C. (1991) Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. Correlation with molecular orbital energies and hydrophobicity. *J. Med. Chem.*, **34**, 786–797.

Dumais,S., Platt,J., Heckerman,D. and Sahami,M. (1998) Inductive learning algorithms and representations for text categorization. In *Proceedings of the 7th International Conference on Information and Knowledge Management*, Bethesda, Maryland, pp. 148–155.

Fligner,M.A., Verducci,J.S. and Blower,P.E. (2002) A modification of the Jaccard/Tanimoto similarity index for diverse selection of chemical compounds using binary strings. *Technometrics*, **44**, 1–10.

Flower,D.R. (1998) On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.*, **38**, 378–386.

Gärtner,T., Flach,P.A. and Wrobel,S. (2003) On graph kernels: hardness results and efficient alternatives. In *Proceedings of the 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop*, Springer Verlag, NY, pp. 129–143.

Gasteiger,J., Sadowski,J., Schuur,J., Selzer,P., Steinhauer,L. and Steinhauer,V. (1996) Chemical information in 3D-space. *J. Chem. Inf. Comput. Sci.*, **36**, 1030–1037.

Gower,J.C. (1971) A general coefficient of similarity and some of its properties. *Biometrics*, **27**, 857–871.

Gower,J.C. and Legendre,P. (1986) Metric and euclidean properties of dissimilarity coefficients. *J. Classif.*, **3**, 5–48.

Hadjipavlou-Litina,D. and Hansch,C. (1994) Quantitative structure-activity relationship of the benzodiazepines. A review and reevaluation. *Chemical Reviews*, **94**, 1483–1505.

Hanley,J.A. and McNeil,B.J. (1982) The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, **143**, 29–36.

Haussler,D. (1999). Convolution Kernels on Discrete Structures. *Technical Report UCS-CRL-99-10*, University of California, Santa Cruz.

Heckerman,D. (1998) A tutorial on learning with Bayesian networks. In Jordan,M., (ed.), *Learning in Graphical Models*, Kluwer Dordrecht.

Helma,C., King,R.D., Kramer,S. and Srinivasan,A. (2001) The predictive toxicology challenge 2000-2001. *Bioinformatics*, **17**(1), 107–108.

Jaakkola,T., Diekhaus,M. and Haussler,D. (1999) Using the Fisher kernel method to detect remote protein homologies. *Intelligent Systems for Molecular Biology*, 1999, 149–158.

James,C.A., Weininger,D. and Delany,J. (2004) *Daylight Theory Manual*.

Kashima,H., Tsuda,K. and Inokuchi,A. (2003) Marginalized Kernels between Labeled Graphs. In *Proceedings of the 20th International Conference on Machine Learning*, Washington, DC, pp. 321–328.

King,R.D., Muggleton,S.H., Srinivasan,A. and Sternberg,M.J.E. (1996) Structure-activity relationships derived by machine learning: the use of atoms and their bond connectivities to predict mutagenicity by inductive logic programming. *Proc. Natl Acad. Sci.*, **93**, 438–442.

Koza,J. (1994) Evolution of a computer program for classifying protein segments as transmembrane domains using genetic programming. In Altman,R., Brutlag,D., Karp,P., Lathrop,R. and Searls,D. (eds), *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA, pp. 244–252.

Kramer,S. and De Raedt,L. (2001) Feature construction with version spaces for biochemical application. In *Proceedings of the 18th International Conference on Machine Learning*, Williams College, MA, pp. 258–265.

Lauritzen,S.L. (1996) *Graphical Models*. Oxford University Press, Oxford, UK.

Leslie,C., Eskin,E., Cohen,A., Weston,J. and Noble,W. (2004) Mismatch string kernels for discriminative protein classification. *Bioinformatics*, **20**, 467–476.

Leslie,C., Eskin,E., Cohen,A., Weston,J. and Noble,W.S. (2003) Mismatch string kernels for SVM protein classification. In Becker,S., Thrun,S. and Obermayer,K. (eds), *Advances in Neural Information Processing Systems*, Vol. 15, MIT Press, Cambridge, MA, pp. 1417–1424.

Leslie,C., Eskin,E. and Noble,W.S. (2002) The spectrum kernel: a string kernel for svm protein classification. *Pac. Symp. Biocomput.*, **2002**, 564–575.

Lodhi,H., Saunders,C., Shawe-Taylor,J., Cristianini,N. and Watkins,C. (2002) Text classification using string kernels. In *The Journal of Machine Learning Research*, Vol. 2, MIT Press, Cambridge, MA, pp. 419–444.

Mahé,P., Ueda,N., Akutsu,T., Perret,J.L. and Vert,J.P. (2004) Extension of marginalized graph kernels. In *Proceedings of the 21st International Conference on Machine Learning*, New York, NY.

Micheli,A., Sperduti,A., Starita,A. and Biancucci,A.M. (2003) A novel approach to QSPR/QSAR based on neural networks for structures. In Cartwright,H. and Sztandera,L.M. (eds), *Soft Computing Approaches in Chemistry*. Springer Verlag, Heidelberg, Germany, pp. 265–296.

Micheli,A., Sperduti,A., Starita,A. and Bianucci,A.M. (2001) Analysis of the internal representations developed by neural networks for structures applied to quantitative structure-activity relationship studies of benzodiazepines. *J. Chem. Inf. Comput. Sci.*, **41**, 202–218.

Muggleton,S. (ed.) (1992), Inductive logic programming, Vol. 38 of APIC Series. Academic Press, London.

Pearl,J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA.

Sadowski,J., Gasteiger,J. and Klebe,G. (1994) Comparison of automatic three-dimensional model builders using 639 X-ray structures. *J. Chem. Inf. Comput. Sci.*, **34**, 1000–1008.

Sakakibara,Y., Brown,M., Hughey,R., Mian,I.S., Sjölander,K., Underwood,R.C. and Haussler,D. (1994) Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res.*, **22**, 5112–5120.

Salton,G. (1991) Developments in automatic text retrieval. *Science*, **253**, 974–980.

Schölkopf,B. and Smola,A.J. (2002) *Learning with Kernels, Support Vector Machines, Regularization, Optimization and Beyond*. MIT University Press, Cambridge, MA.

Ukkonen,E. (1995) On–line construction of suffix trees. *Algorithmica*, **14**, 249–260.

Vert,J.P. (2002) A tree kernel to analyze phylogenetic profiles. *Bioinformatics*, **18**(Suppl.1), S276–S284.

Vishwanathan,S.V.N. and Smola,A.J. (2003) Fast Kernels for Strings and Tree Matching. In *Adv. in Neural Information Processing Systems*, Vol. 15.

Weiner,P. (1973) Linear Pattern Matching Algorithms. In *Proceedings of the 14th IEEE Annual Symposium on Switching and Automata Theory*, pp. 1–11.

Weininger,D., Weininger,A. and Weininger,J.L. (1989) SMILES. 2. algorithm for generation of uniques SMILES notation. *J. Chem. Inf. Comput. Sci.*, **29**, 97–101.

Yang,Y. and Pedersen,J.O. (1997) A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning*, Nashville, TN, pp. 412–420.