BIOINFORMATICS

Vol. 00 no. 00 2006
Pages 1–7

Glycan classification with tree kernels

Yoshihiro Yamanishi [a]*, Francis Bach [b], Jean-Philippe Vert [c]

[a]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan, [b]Center of Mathematical Morphology, Ecole des Mines de Paris, Fontainebleau, France, [c]Center for Computational Biology, Ecole des Mines de Paris, Fontainebleau, France

ABSTRACT

Motivation: Glycans are covalent assemblies of sugar that play crucial roles in many cellular processes. Recently, comprehensive data about the structure and function of glycans have been accumulated, therefore the need for methods and algorithms to analyze these data is growing fast.

Results: This paper presents novel methods for classifying glycans and detecting discriminative glycan motifs with support vector machines (SVM). We propose a new class of tree kernels to measure the similarity between glycans. These kernels are based on the comparison of tree substructures, and take into account several glycan features such as the sugar type, the sugar bound type, or layer depth. The proposed methods are tested on their ability to classify human glycans into four blood components: leukemia cells, erythrocytes, plasma, and serum. They are shown to outperform a previously published method. We also applied a feature selection approach to extract glycan motifs which are characteristic of each blood component. We confirmed that some leukemia-specific glycan motifs detected by our method corresponded to several results in the literature.

Availability: Softwares are available upon request.

Supplementary information: Datasets are available at the following webisite: http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/glycankernel/.

Contact: yoshi@kuicr.kyoto-u.ac.jp

1 INTRODUCTION

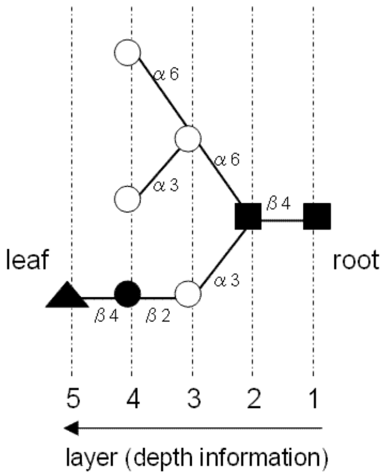
Glycans, or carbohydrate sugar chains, are covalent assemblies of sugars (oligosaccharides and polysaccharides) that exist in either free form or in covalent complexes with proteins or lipids. Although a growing body of evidence supports crucial roles for glycans in many cellular processes, including cell-cell communication, immune system, protein interaction or tumor progression [Varki et al., 1999, Fuster and Esko, 2005], understanding the biological functions of glycans and relating them to their structure remains challenging experimentally. As databases such as KEGG/Glycan [Kanehisa et al., 2004, Hashimoto et al., 2006] have started accumulating information about the structure and function of glycans, the need for methods and algorithms to analyze these data is growing fast.

Unlike genes or protein sequences, for which a number of well-established algorithms are now available for various data mining tasks such as similarity detection, clustering, supervised classification, structure prediction, or functional motif extraction, glycans are generally not linear polymers. They have more complex structures, that can be represented by rooted ordered trees, with monosaccharides as labeled vertices, and sugar bounds as labeled edges (see Figure 1). As a result, specific approaches have recently been developed for comparison of glycans [Aoki et al., 2004, 2005], probabilistic modeling of glycan families [Ueda et al., 2005], and analysis of MS/MS spectra of oligosaccharides [Tang et al., 2005]. There is still an incentive to develop efficient methods for the automatic classification of glycans, and the extraction of biologically relevant substructures.

Glycans exhibit a large diversity of structures in different organisms, and in different tissues and organs of a given organism. Recently, a computational approach to the supervised classification of glycans into blood components and to the detection of leukemia-specific glycan substructures has been proposed [Hizukuri et al., 2005]. The approach is based on the extraction of short linear substructures from the glycan structures, resulting in a quantitative measure of similarity between glycans based on the count of shared substructures. This measure of similarity is then used as an input to a support vector machine (SVM) classifier which is trained to discriminate between different blood components and blood types. The goal of this paper is to extend this framework to a broader class of structure representations, with the motivation of both increasing the accuracy of glycan classification, and providing a framework for the extraction of biologically relevant glycan substructures.

More precisely, we investigate different high-dimensional representations for glycan structures and use them for supervised classification with SVM. In SVM jargon, we define new *kernels* for trees, adapted to the classification of glycans. Our tree kernels are based on the indexation of a glycan tree structure by a set of subtrees it contains. We investigate different variants in the definition of subtrees, in the importance placed on the depth of a subtree in the glycan tree, and in the size of the subtrees. In spite of their large dimensions, these vector representations are adapted to the supervised classification of glycans by kernel methods and in particular the SVM, in the spirit of earlier work on convolution kernels for tree structures [Collins and Duffy, 2001, Haussler, 1999]. We perform a thorough analysis of the classification performance obtained by different representations on the problem of predicting the blood origin of glycans among leukemia cell, erythrocyte, plasma, and serum. We then apply feature selection methods to find discriminative subtree motifs in glycans, and relate the selected substructures to biologically known facts. Finally, in order to combine the information provided by different representations we apply recently developed methods for multiple kernel learning [Bach et al., 2005], resulting in an optimal weighting of each representation for each classification task. As shown in Section 4, the methods

*to whom correspondence should be addressed



Symbol	Monosaccharide
▲	Glucose
●	Galactose
○	Mannose
◆	N-acetyl neuraminic/sialic acid
■	N-acelylglucosamine
□	N-acetylgalactosamine
△	Fucose

Fig. 1. An example of a glycan structure.

developed in this paper not only lead to better classification performance than previously reported results, they also provide additional biological insights in the role and structure of glycans, through the substructures extracted by feature selection and the weights learned by multiple kernel learning.

2 MATERIALS

All glycan structures used in this study are obtained from the KEGG/Glycan database [Kanehisa et al., 2004, Hashimoto et al., 2006]. We use the same dataset as Hizukuri et al. [2005], who took care to remove non-carbohydrate moieties such as phosphate and sulfate, and gathered glycan structures consisting of the following seven monosaccharides (sugars): glucose (Glc), galactose (Gal), mannose (Man), fucose (Fuc), N-acetylglucosamine (GlcNAc), N-acetylgalactosamine (GalNAc), and N-acetyl neuraminic/sialic acid (Neu). The linkages between these monosaccharides also have variables, such as the anomer (α or β) and the hydroxyl group numbers to which they are attached on the monosaccharides. The dataset consists of 365 glycan structures originating from four human blood components: leukemic cells, erythrocytes, serum, and plasm, with respectively 162, 112, 85, and 73 examples. Note that some glycans belong to several blood components.

3 METHODS

The main contribution of this paper is to propose various embeddings of glycan structures into Euclidean spaces of possibly large dimensions, trying to both capture biologically relevant features from the glycan structures and make them amenable to further processing. More precisely, we investigate the development of *kernel functions* for glycans, that is, of similarity functions which correspond to inner products of such Euclidean embeddings [Schölkopf and Smola, 2002]. Indeed, once such an embedding is chosen, various algorithms from machine learning can be used that can take advantage of the representation in terms of kernel—those algorithms are usually referred to as *kernel methods*, and are enjoying an increasingly popularity in computational biology due to their generally good performance and ability to process complex and structured data [Schölkopf et al., 2004]. In this paper, we present different experiments of binary classification of glycans, but the framework developed could easily take into account more complex supervised or unsupervised learning tasks, such as multi-label classification or clustering, by simply choosing the appropriate kernel algorithm [Shawe-Taylor and Cristianini, 2004].

3.1 Kernel methods for classification

Kernel methods work by first embedding each data point $x \in \mathcal{X}$ (x represent a glycan structure in our case) to a vector space $\mathcal{F}$ through a feature map $\Phi : \mathcal{X} \to \mathcal{F}$; the vector space $\mathcal{F}$ is referred to as the *feature space*. Then, linear pattern recognition algorithms are applied in this feature space, on the mapped data points. The first key characteristic of these methods is that the mapping $x \mapsto \Phi(x)$ into feature space may not be explicit; rather, the mapping is defined implicitly through the inner product $k(x, y) = \langle \Phi(x), \Phi(y) \rangle$, defined for any x and y in $\mathcal{X}$, and referred to as the *kernel function*. Given data points, $x_i \in \mathcal{X}$, $i = 1, \dots, n$, kernel learning algorithms will mainly use the kernel function evaluated at pairs of data points (x_i, x_j). The $n \times n$ matrix composed of those kernel function evaluations, defined as $K_{ij} = k(x_i, x_j)$, is referred to as the *kernel matrix*. Note that in many cases, the dimension of the feature space is large and potentially infinite; by manipulating only kernel matrices, kernel methods make possible to deal with very large numbers of features. In our experiments, the total number of features is 3330 (when considering co-rooted subtrees) and 18681 (when considering all subtrees).

The second key characteristic of kernel methods is that they can be applied to non-vectorial data, in our case labeled tree structures. We only need to be able to define a valid positive semi-definite kernel function, i.e., a function which corresponds to an inner product in some potentially large feature space. In Section 3.2 we present how such kernel functions can be defined for glycans.

In this paper we focus on the task of binary classification. Namely, we assume that we have labeled data points $(x_i, y_i) \in \mathcal{X} \times \{-1, 1\}$, $i = 1, \dots, n$. We use two different learning algorithms, kernel logistic regression and the support vector machine (SVM). Those two methods look for classifiers of the form $f(x) =$

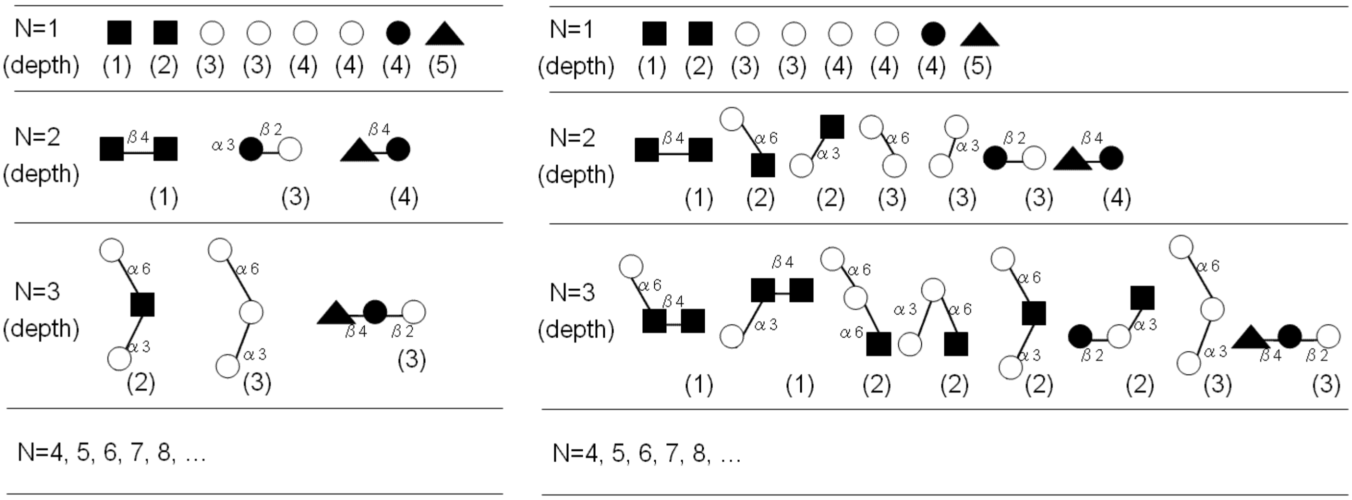


Fig. 2. An illustration for the decomposition of the glycan tree in Fig.1 into all possible co-rooted subtrees (left) and all possible subtrees (right).

$\text{sign}(\langle w, \Phi(x)\rangle + b)$, where $w \in \mathcal{F}$ and $b \in \mathbb{R}$, but with different optimization formulations (see Shawe-Taylor and Cristianini [2004] for details). The two methods work solely on the kernel matrix K and usually lead to similar classification accuracies [Hastie et al., 2001].

All experimental results reported below are obtained from a common methodological framework. For each classification problem a 5-fold cross-validation is performed, and the measures are averaged over 3 repeats of the whole process. Since our binary classification tasks involve unbalanced classes, the SVM and the logistic regression are trained with asymmetric cost, C^- for false positives and C^+ for false negatives. If n_+ and n_- are the number of positive and negative examples, then we use C^+ and C^- such that that $C^+ + C^- = 1$ and $C^+n^+ = C^-n^-$. Classification performance on test data is reported using areas under the ROC curve (AUC).

We performed the computation for single kernel learning with the *libsvm*[1] implementation of SVM in the PyML[2] machine learning environment. For multiple kernel learning, we used the Matlab toolbox of Bach et al. [2005][3].

3.2 Tree kernels for glycans

The molecular structure of a glycan x is characterized by a *labeled ordered rooted tree* $T(x) = (V(x), E(x))$, where $V(x)$ and $E(x)$ are respectively the set of vertices and the set of edges of the tree $T(x)$. Both vertices and edges are labeled. For glycans, the vertex labels represents the monosaccharide type and can take seven categorical values, while the edge labels characterize the sugar bounds and may take 12 values in the blood data used in our study.

We use rooted trees, because for glycans, the same sugar is always bound to proteins or cells during interactions between the glycan and proteins or cells; it can thus be isolated as a root. Moreover, the tree $T(x)$ is considered ordered because for each vertex the order of the children is significant. All trees which are considered in this paper are labeled ordered rooted trees, and from now on, we refer to those as simply trees. Note that glycan trees are not always represented by binary trees. There are some glycans in which the nodes have multiple children (more than two children).

Given a tree T, the tree S is a *subtree* of T if and only if it is a connected subgraph of T and the labels of the edges and vertices of S match the corresponding ones in T. Note that the ordered rooted structure of a tree T is naturally inherited by subtrees. The subtree is said to be *co-rooted* if a vertex is always included in the subtree with all its siblings in T. See Figure 2 for examples of co-rooted subtrees and all subtrees of a given glycan. We let denote $\mathcal{S}(T)$ the set of subtrees of the tree T and $\mathcal{S}_C(T)$ the set of co-rooted subtrees of the tree T.

All kernels that we define exhibit the same structure, i.e., the kernel functions evaluated at x_1 and x_2 can be expressed as a sum of local kernels $q(S_1, S_2)$ over all possible subtrees (or co-rooted subtrees) S_1 and S_2 of the tree $T(x_1)$ and $T(x_2)$:

$$k(x_1, x_2) = \sum_{S_1 \in \mathcal{S}(T(x_1)),\ S_2 \in \mathcal{S}(T(x_2))} q(S_1, S_2)$$

or

$$k_C(x_1, x_2) = \sum_{S_1 \in \mathcal{S}_C(T(x_1)),\ S_2 \in \mathcal{S}_C(T(x_2))} q(S_1, S_2)$$

where $q(S_1, S_2)$ is itself a kernel between trees. This immediately shows that the kernels k and k_C are valid positive semi-definite functions. Indeed, q being a kernel we can write $q(S_1, S_2) = \langle \Phi(S_1), \Psi(S_2)\rangle$, and thus $k(x_1, x_2) = \left\langle \sum_{S_1 \in \mathcal{S}(T(x_1))} \Psi(S_1), \sum_{S_2 \in \mathcal{S}(T(x_2))} \Psi(S_2) \right\rangle$, which shows that $k(x_1, x_2)$ is indeed an inner product.

When the local kernel $q(S_1, S_2)$ is a Dirac function between S_1 and S_2, then the value of the kernel $k(x_1, x_2)$ is simply the number of common subtrees that verified some additional properties. Geometrically, this amount to representing a tree by the vector of indicator functions for each subtree, and taking the inner product between such vector representations. This fact will be used when computing those kernels efficiently in Section 3.2.2.

[1] http://www.csie.ntu.edu.tw/ cjlin/libsvm/

[2] http://pyml.sourceforge.net

[3] http://cmm.ensmp.fr/ bach/path/

3.2.1 Local kernel functions We consider the following local kernel functions $q(S_1, S_2)$ between subtrees:

- $q^0(S_1, S_2) = \delta(S_1 = S_2)$, that is, 1 if $S_1 = S_2$ and 0 otherwise, equality between trees being defined as equality of structure and all vertex and edge labels. The kernel then simply counts the number of common subtrees or the number of common co-rooted subtrees.
- $q^N(S_1, S_2) = \delta(S_1 = S_2)\delta(n(S_1) = N)\delta(n(S_2) = N)$, where $n(S_i)$ denotes the number of nodes of S_i and N is a pre-specified number of nodes from $N = 1$ to 10. The resulting kernel simply counts the number of common subtrees of a given size or the number of common co-rooted subtrees of a given size.
- $q^D(S_1, S_2) = \max(D + 1 - |d(S_1) - d(S_2)|, 0)\delta(S_1 = S_2)$ where $d(S_i)$ is the depth in T_i of the root of S_i, and D is the maximal allowed difference in depths. When $D = 0$, only subtrees with identical depths are matched; as D increases, pairs of subtrees with increasingly different depths contribute to the kernel. We note than q^0 (matching subtrees whatever their depths) can be seen to some extent as the limit of q^D when D tend to infinity. The fact that q^D is a valid positive definite function between subtrees results from the classical results that the function $(x, y) \in \mathbb{R}^2 \mapsto \max(D - |x - y|, 0)$ is a positive definite function, usually referred to as the *triangular kernel* [Berg et al., 1984]
- Any product of q^N and q^D: common subtrees of size N and with more or less close depths are counted.

Limiting ourselves to all combinations of the 7 values for D ($D = no, 0, 1, 2, \cdots, 5$) and 19 values for N ($N = 1, 2, 3, \cdots, 17, 18, all$) , and applying the local kernels either on the set of subtrees or the set of co-rooted subtrees, already provides us with a basic set of $2 \times 7 \times 19$ different kernel functions. In this case, $D = no$ means that we do not consider any depth information, and $N = all$ means that we use all the nodes. We observe that with these notations, the representation of glycans proposed in Hizukuri et al. [2005] involving the matching of linear subtrees of length 3 with identical depth corresponds to the kernel obtained by taking the product of q^N, for $N = 3$, with q^D, for $D = 0$, and summing over all subtrees.

3.2.2 Kernel computations For large trees, the kernels presented earlier can be computed by dynamic programming in time quadratic in the number of vertices [Collins and Duffy, 2001, Shawe-Taylor and Cristianini, 2004]. In our situation, the number of vertices is usually small (always less than 18) and thus a complete recursive enumeration of the subtrees is feasible, which leads to a basis representation of the feature space (except for $q^D, D > 0$), and thus allows us to perform feature selection as presented in Section 3.3. The case of $q^D(D > 0$ is just slightly different, because although one can write it as an inner product in an explicit feature space (e.g., each subtree S with depth d can be indexed by the features $\{S_d, S_{d+1}, \ldots, S_{d+D-1}\}$ to obtain the kernel q^D by inner product), which allows to compute this kernel by explicit vector representations of the trees, the dimensions of this feature space are not associated to a precise subtree which makes feature selection in that case meaningless.

3.3 Feature selection

Kernel methods and support vector machines have the drawback that, although they generally give good classification accuracy, the biological facts that make this classification efficient are often hidden. In the case of glycans, however, a particular incentive besides good classification is to extract substructures characteristic of different classes of glycans. Indeed, the determination of such substructures might shed light on the biochemical mechanisms involved in a given process, for example.

Among the kernels proposed in 3.2, those involving the local kernels q^D for $D > 0$ do not correspond to obvious embedding of the glycans based on their substructures, and therefore do not lend themselves particularly well the the extraction of discriminative substructures. However, those involving only q^0, q^N and q^D for $D = 0$ can be written as explicit dot product in a space indexed by various substructures. As a result, instead of keeping the complete embedding, one can instead focus its attention on the selection of a small number of informative features, for a given classification task, which could then correspond to discriminative substructures.

The problem of feature selection is a well-known problem in statistics and machine learning, with many existing algorithms. A thorough comparison of many methods being beyond the scope of this paper, we focused on a simple feature selection method proposed in Golub et al. [1999] in the framework of gene selection from microarray data. Given a training set of positive and negative examples (in the binary classification framework), each feature is ranked according to the value of the statistics: $\frac{|\mu_+ - \mu_-|}{\sigma_+ + \sigma_-}$, where μ_+, μ_-, σ_+ and σ_- are the mean and standard deviation of the values of the features on the positive and negative sets, respectively. In order to evaluate the relevance of this approach to feature selection for the classification of glycans, we first estimate the classification accuracy when a varying number of features are selected; we then propose an analysis of the features themselves in the context of glycobiology in Section 4.

3.4 Multiple kernel learning

In Section 3.2 we have defined a large set of 266 different kernels k_j, $j = 1, \ldots, Q = 266$, by matching subtrees with different structures. In this situation, it has recently been proved advantageous to combine the different kernels k_j into a a single kernel k by using a convex combination $k(x, y) = \sum_{j=1}^{Q} \eta_j k_j(x, y)$, with nonnegative coefficients (η_j) which sum to one [Lanckriet et al., 2004]. During training, both the parameters of the convex combination (η_j) and the resulting classifier using the combined kernels are estimated in a single convex optimization problem. In this paper, we used the kernel matrix normalization procedure and the code of Bach et al. [2005], which is based on kernel logistic regression and provides an efficient way of searching over the regularization parameter. In particular, the entire set of solutions for all values of the regularization parameter is obtained with complexity $O(Qn^3)$. Combining kernels leads to better classification performance as well as a selection of the kernels which are relevant for the classification task at hand. In our case, those kernels are characterized by a number of nodes N and a depth parameter D; the weights of the kernel combination thus give information on the most discriminative size of subtrees as well as on the importance of depth.

Table 1. Classification results for each class: erythrocyte, leukemia, plasma and serum.

		Erythrocyte		Leukemia		Plasma		Serum	
		Co-rooted	All	Co-rooted	All	Co-rooted	All	Co-rooted	All
D	N	AUC	AUC	AUC	AUC	AUC	AUC	AUC	AUC
0	1	91.7 ± 0.2	89.1 ± 0.3	92.0 ± 0.6	89.7 ± 1.0	81.6 ± 0.8	77.3 ± 0.5	87.2 ± 0.1	84.4 ± 0.5
0	3	91.8 ± 0.4	93.3 ± 0.5	91.1 ± 0.4	93.3 ± 0.2	81.0 ± 0.8	83.6 ± 1.2	86.2 ± 0.8	86.5 ± 0.2
0	all	92.7 ± 0.3	92.5 ± 0.3	92.5 ± 0.2	91.5 ± 0.6	82.7 ± 0.7	83.5 ± 0.4	88.7 ± 0.7	89.6 ± 0.9
2	1	92.2 ± 0.3	90.1 ± 0.2	91.7 ± 0.1	89.9 ± 0.5	83.1 ± 0.8	76.6 ± 0.9	87.0 ± 1.3	83.9 ± 0.5
2	3	93.4 ± 0.6	94.6 ± 0.5	93.3 ± 0.3	94.6 ± 0.2	82.7 ± 1.6	83.1 ± 1.7	86.5 ± 0.7	86.6 ± 0.3
2	all	93.3 ± 0.3	93.6 ± 0.2	93.2 ± 0.2	92.1 ±0.5	84.5 ± 1.0	83.5 ± 0.6	88.7 ± 1.1	88.7 ± 1.0
no	1	93.2 ± 0.3	88.5 ± 0.3	91.5 ± 0.2	91.0 ± 0.6	81.6 ± 0.9	72.8 ± 1.2	85.2 ± 0.3	85.5 ± 1.6
no	3	93.8 ± 0.7	94.8 ± 0.6	92.9 ± 0.3	94.0 ± 0.2	82.2 ± 1.3	82.0 ± 1.7	85.9 ± 0.3	86.2 ± 0.7
no	all	93.6 ± 0.5	93.3 ± 0.4	92.2 ± 0.1	92.7 ± 0.0	83.3 ± 1.4	84.3 ± 1.2	87.6 ± 0.8	89.0 ± 1.1

Table 2. Classification results for multiple kernels (AUC).

Erythrocyte	Leukemia	Plasma	Serum
94.4 ± 2.3	96.0 ± 1.5	83.7 ± 4.9	91.2 ± 2.7

4 RESULTS AND DISCUSSIONS

4.1 Classification results

Table 1 presents the experimental results of a few kernels on the four problems of discriminating each blood origin from the other ones. We focus our attention on the following kernels, both for the set of subtrees and for the set of co-rooted subtrees: $N = 1$ (single sugar), $N = 3$ (sugar trimers), or no N (all subtrees); $D = 0$ (strict depth matching), $D = 2$ (similar depth matching), or no D (all matches whatever the depths). In each case we report the area below the ROC curve (AUC) of true positives as a function of false positives. The classification results for multiple kernel learning are presented in Table 2.

A few points are worth mentioning; first, the AUCs are generally high for all problems, which confirms that the glycan structure contains a lot of information on their role. Also, predicting the erythrocyte and leukemia classes is easier than predicting the plasma and serum class. In addition, results published by Hizukuri et al. [2005] corresponds to $D = 0$, $N = 3$ and the choice "all subtrees". This shows that some kernels introduced in this paper usually lead to similar or improved performance, although the best kernel depends on the problem. Finally, apart from the plasma class, multiple kernel learning leads to improved classification performance.

4.2 Feature and kernel selection

In order to confirm the trends observed in the results of supervised classification, we performed a systematic analysis of the classification performance when only a small number of features are selected. The results presented in Figure 3 suggests that most of the discriminative power is obtained from only a few glycan substructures for all classes. This also shows that our feature selection procedure provides a reasonable selection of discriminative features. The effect of depth feature seems to differ between classes (blood components in this study). This phenomenon might be due to the biological function of glycans in each blood component.

In order to analyze relevant feature, we performed a single feature extraction on the whole dataset and examined the first few discriminative features selected in each class. Because of the space limitation we focus here on leukemia cell specific motifs obtained from the tree kernel with all subtrees and $D = 0$. Figure 4 shows the five subtrees with highest scores. The high scoring glycan substructures can be considered to be characteristic motifs of each blood component; the substructure with the highest score is α-Neu$p$-(2→3)-β-Gal$p$-(1→4)-Glc$p$NAc at the fifth layer, which exactly corresponds to the substructure in previous work [Hizukuri et al., 2005]. The substructure with the second highest score is α-Neu$p$ at the seventh layer. This result suggests that a sialic acid (represented by Neu$p$) attaching to the leaf part of glycans can be involved in cancer. This hypothesis is actually consistent with an experimental report that sialic acid tends to appear in many tumor cells [Kannagi et al., 1986]. All the results for extracted substructures in each blood component can be obtained from the online supplement.

For the discrimination of leukemia cells, the multiple kernel learning algorithm assigned highest weights on the kernels (D=no, N=3) and the kernel (D=0, N=10), with respectively $\eta = 0.18$ and $\eta = 0.15$. The high weight on the kernel (D=no, N=3) supports an observation that glycosyltransferases, which are involved in glycan synthesis, physically interact with about three monosaccharides at the leaves of glycans [Varki et al., 1999]. The high weight on the kernel (D=0, N=10) implies that there might exist a big leukemia-specific glycan motif which depends strictly on the localization (layer). This is also reflected by the result of our feature extraction result in Figure 4.

Moreover, if we concatenate the high scoring subtrees in Figure 4 based on their layer information, we can reconstruct one big sugar

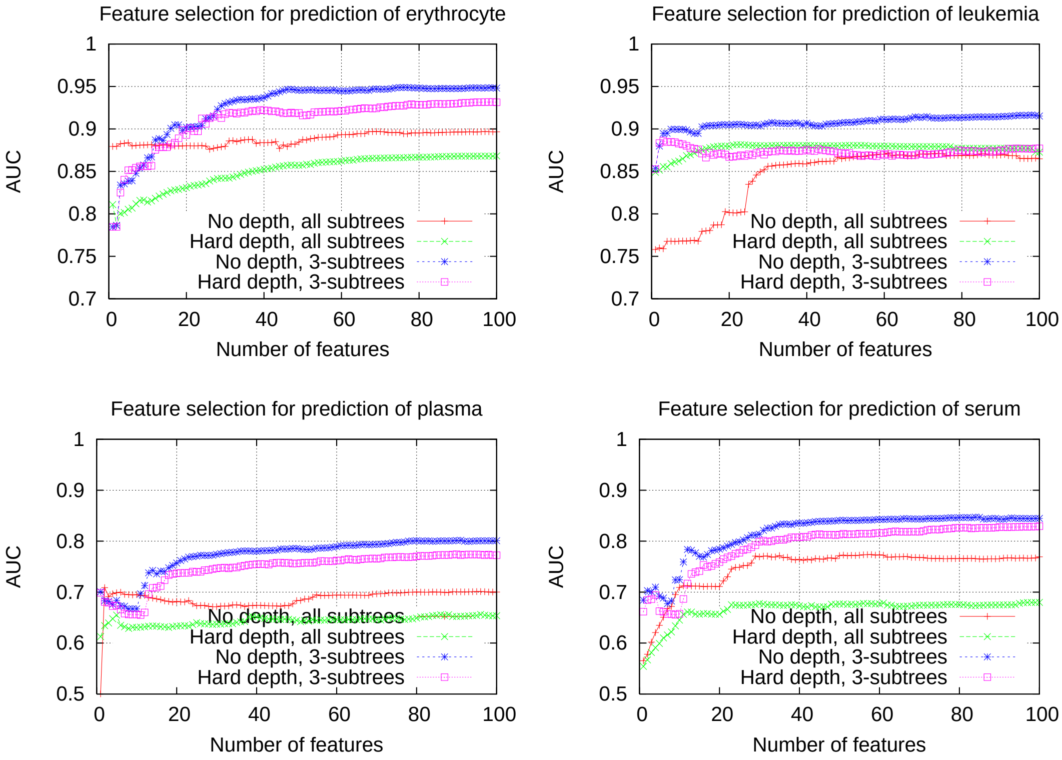


Fig. 3. Feature selection for each class: erythrocyte, leukemia, plasma and serum

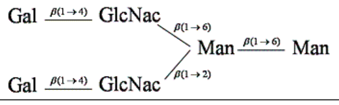
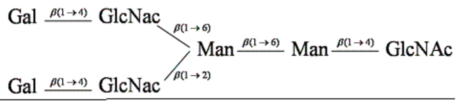
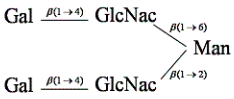
No.	score	layer	substructure
1	0.552	5	$\mathrm{Neu} \xrightarrow{\alpha(2\to3)} \mathrm{Gal} \xrightarrow{\beta(1\to4)} \mathrm{GlcNac}$
2	0.502	7	Neu
3	0.467	3	
4	0.467	2	
5	0.467	4	
6	0.463	5	$\mathrm{Neu} \xrightarrow{\alpha(2\to6)} \mathrm{Gal} \xrightarrow{\beta(1\to4)} \mathrm{GlcNac}$

Fig. 4. Characteristic substructures for leukemic cells.

chain. Therefore it suggests that glycans including such motif structure might work as a signal in the discrimination of leukemic cells from the normal cells. These results suggest that 3-mer might be an appropriate size as a glycan motif in general, but the localization information is also important in the case of big glycan motifs.

5 CONCLUSION

In this paper we developed an SVM-based approach for classifying glycans with new tree kernels, and detecting discriminative glycan motifs for each classification task. The originality of our tree kernel relies on the richness of the substructures that are considered. Our works extends the previous work of Hizukuri et al. [2005], who focused on the use of k-mer representation (3-mer in their case). Our results suggest that informative and discriminative glycan motifs do not always constitute chains of sugars of fixed length. Our framework enabled us not only to discriminate classification groups with higher accuracy but also to detect more flexible size of glycan motifs. We also confirmed that some leukemia-specific glycan motifs detected by our method corresponded to the results in the literature. It should be also pointed out that our method is applicable to classification for any types of targets such as tissues, organs, and organisms. Our future work includes more comprehensive glycan classification and motif detection for other targets.

ACKNOWLEDGMENTS

We thank Yoshiyuki Hizukuri for help with the dataset.

Conflict of Interest: None declared.

REFERENCES

K. F. Aoki, H. Mamitsuka, T. Akutsu, and M. Kanehisa. A score matrix to reveal the hidden links in glycans. *Bioinformatics*, 21(8):1457–63, Apr 2005.

K. F. Aoki, A. Yamaguchi, N. Ueda, T. Akutsu, H. Mamitsuka, S. Goto, and M. Kanehisa. KCaM (KEGG Carbohydrate Matcher): a software tool for analyzing the structures of carbohydrate sugar chains. *Nucleic Acids Res.*, 32(Web Server issue): W267–72, Jul 2004.

F. R. Bach, R. Thibaux, and M. I. Jordan. Computing regularization paths for learning multiple kernels. In *Adv. Neural. Inform. Process Syst.*, volume 17, 2005.

C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic analysis on semigroups*. Springer-Verlag, 1984.

M. Collins and N. Duffy. Convolution kernels for natural language. In *Adv. Neural. Inform. Process Syst.*, volume 14, pages 625–632, 2001.

M. N. Fuster and J. D. Esko. The sweet and sour of cancer: glycans as novel therapeutic targets. *Nat. Rev. Cancer*, 5(7):526–42, Jul 2005.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

K. Hashimoto, S. Goto, S. Kawano, K. F. Aoki-Kinoshita, N. Ueda, M. Hamajima, T. Kawasaki, and M. Kanehisa. Kegg as a glycome informatics resource. *Glycobiology*, 16:63R–70R, 2006.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag, 2001.

D. Haussler. Convolution kernels on discrete structures. Technical Report UCSC-CRL-99-10, UC Santa Cruz, 1999.

Y. Hizukuri, Y. Yamanishi, O. Nakamura, F. Yagi, S. Goto, and M. Kanehisa. Extraction of leukemia specific glycan motifs in humans by computational glycomics. *Carbohydr. Res.*, 340(14):2270–8, Oct 2005.

M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, 32(Database issue):D277–80, Jan 2004.

R. Kannagi, Fukushi Y., Tachikawa T., Shin S. Noda A., Shigeta K., Hiraiwa N., Fukuda Y., Inamoto T., and HakomoriAoki S. Quantitative and qualitative characterization of human cancer-associated serum glycoprotein antigens expressing fucosyl or sialyl-fucosyl type 2 chain polylactosamine. *Cancer Res.*, 46:2619–2626, 1986.

G. R. G. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan, and W. S. Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20:2626–2635, 2004.

B. Schölkopf and A.J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

B. Schölkopf, K. Tsuda, and J.P. Vert. *Kernel Methods in Computational Biology*. MIT Press, 2004.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Camb. Univ. Press, 2004.

H. Tang, Y. Mechref, and M.V. Novotny. Automated interpretation of ms/ms spectra of oligosaccharides. *Bioinformatics*, 21:i431–i439, 2005.

N. Ueda, K. F. Aoki-Kinoshita, A. Yamaguchi, T. Akutsu, and H. Mamitsuka. A probabilistic model for mining labeled ordered trees: Capturing patterns in carbohydrate sugar chains. *IEEE Transactions on Knowledge and Data Engineering*, 17 (8):1051–1064, 2005.

A. Varki, R. Cummings, J. Esko, H. Freeze, G. Hart, and J. Marth. *Essentials of glycobiology*. Cold Spring Harbor Laboratory Press, 1999.