BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 7(8)

# Extracting glycan motifs using a biochemically-weighted kernel

**Hao Jiang[1], Kiyoko F Aoki-Kinoshita [2]* & Wai-Ki Ching[1]**

[1]Advanced Modeling and Applied Computing Laboratory, Department of Mathematics, University of Hong Kong, Pokfulam Road, Hong Kong; [2]Department of Bioinformatics, Faculty of Engineering, Soka University, Tokyo, Japan; Kiyoko F Aoki-Kinoshita– Email: kkiyoko@soka.ac.jp; phone/Fax: +81-42-691-4116; *Corresponding author

Selected publications from Asia Pacific Bioinformatics Network (APBioNet) 10th International Conference on Bioinformatics (InCoB 2011), Malaysia, November 30 to December 02, 2011

**Abstract:**
Carbohydrates, or glycans, are one of the most abundant and structurally diverse biopolymers constitute the third major class of biomolecules, following DNA and proteins. However, the study of carbohydrate sugar chains has lagged behind compared to that of DNA and proteins, mainly due to their inherent structural complexity. However, their analysis is important because they serve various important roles in biological processes, including signaling transduction and cellular recognition. In order to glean some light into glycan function based on carbohydrate structure, kernel methods have been developed in the past, in particular to extract potential glycan biomarkers by classifying glycan structures found in different tissue samples. The recently developed weighted q-gram method (LK-method) exhibits good performance on glycan structure classification while having limitations in feature selection. That is, it was unable to extract biologically meaningful features from the data. Therefore, we propose a biochemically-weighted tree kernel (BioLK-method) which is based on a glycan similarity matrix and also incorporates biochemical information of individual q-grams in constructing the kernel matrix. We further applied our new method for the classification and recognition of motifs on publicly available glycan data. Our novel tree kernel (BioLK-method) using a Support Vector Machine (SVM) is capable of detecting biologically important motifs accurately while LK-method failed to do so. It was tested on three glycan data sets from the Consortium for Functional Glycomics (CFG) and Kyoto Encyclopedia of Genes and Genomes (KEGG) GLYCAN and showed that the results are consistent with the literature. The newly developed BioLK-method also maintains comparable classification performance with the LK-method. Our results obtained here indicate that the incorporation of biochemical information of q-grams further shows the flexibility and capability of the novel kernel in feature extraction, which may aid in the prediction of glycan biomarkers.

**Keywords:** metacestode; rodent; internal transcribed spacer; ribosomal DNA; polymerase chain reaction

**Background:**
Supporting evidence has verified that glycans play crucial roles in cellular functions. However, the complexity in developing high-throughput techniques to characterize glycan structures poses one of the main obstacles to assess the structural elements responsible for specific functions. Thanks to the availability of glycan structure databases such as KEGG **[1]** and the Consortium for Functional Glycomics (CFG) **[2],** informatics techniques can be applied directly to glycan data to help researchers better understand the functions and structures of these complicated molecules.

Compared to the linear structures of DNA and proteins, glycans are generally nonlinear polymers that can be represented by rooted ordered trees. Several approaches have been developed to mine structural features embedded in glycans **[3, 4].** Support vector machines (SVMs) with tree kernels for analyzing glycan structures have been extensively investigated **[5, 6].** In **[5],** 3-


© 2011 Biomedical Informatics

mers were used to represent the features for each glycan structure, where more weight was applied to the matching structures of the variable region (specifically, the non-reducing terminal structures of glycans) in constructing the kernel matrix. As for **[6],** the kernel function was expressed as a sum of local kernels over all possible subtrees. One of the groundbreaking representatives is the q-gram method **[7, 8]** which considers the vector of the frequencies of all possible subtrees isomorphic to paths with q nodes as the q-gram distribution. Like the previously proposed kernels, the traditional q-gram method ignores the similarity between two different q-grams. Taking into consideration the similarity of geometric structures, monosaccharides and glycosidic bonds in q-grams, a new tree kernel was created **[9],** resulting in a weighted q-gram method: LK-method. With this method, the classification performance was improved for some important glycan classes. However, one of the limitations of this method lies in the poor performance in extracting biologically relevant glycan substructures, the most important goal of our research. Our aim is to remedy the defects of the LK-method.

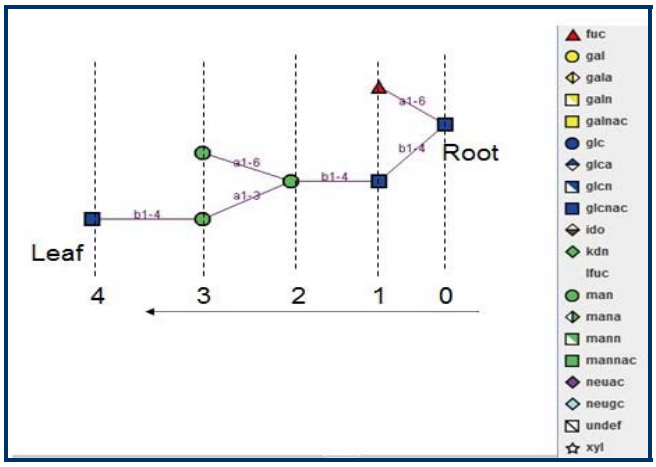


**Figure 1**: Glycan structure with layer information: Root layer is defined as 0.

Kernel methods work by embedding data instances into a feature space F. Due to their good performance in processing complicated data, kernel methods have gained increasing popularity in computational biology **[10].** The Positive Semi-Definite (PSD) property **[11]** of a kernel matrix is required to ensure the existence of a Reproducing Kernel Hilbert Space (RKHS) where a convex optimization formulation can be deduced to yield an optimal solution. However, in practice, similarity matrices can violate the PSD property. For example, in bioinformatics some popular functions evaluating pair-wise similarity between DNA and protein sequences produce non-PSD (or indefinite) kernel matrices. Unfortunately, the best way to use them in the SVM framework is not clear. The weighted q-gram method avoids the problem of the non-PSD property by constructing the similarity matrix as to ensure the PSD property of the kernel matrix. The method performs well in terms of classification accuracy for the often-used leukemia data set, but it did not perform as well on other data sets. Furthermore, the feature selection results of this method were poor in that the biologically known motifs for specific data sets were not retrieved in the results.

In order to obtain biologically meaningful results, we focused on the similarity matrix, which is symmetric and can be decomposed into $S = X \cdot P \cdot X^T$ such that is the diagonal matrix of the eigenvalues sorted in ascending order. Here is an orthogonal matrix of the corresponding eigenvectors. The weighted q-gram method deals with the similarity matrix as $S^T S$ which is in fact $X \cdot P^2 \cdot X^T$. To some extent, we may consider different eigenvalues as representing the roles that each q-gram plays in classification. Furthermore, the kernel matrix used in training the SVM should, in principle, involve the similarity matrix $S$ itself rather than $S^T S$. In this context, a negative eigenvalue $-\lambda$ $(\lambda>0)$ will then be squared, becoming $\lambda^2$, the square of its original magnitude. This suggests a possible reason why this method cannot perform well in all of the data sets as the importance of those negative eigenvalues were magnified.

Previous studies have presented methods that attempt to alter the spectrum of an indefinite kernel matrix in order to create a PSD one. Representatives include the denoising method which deems all negative eigenvalues as noise and replaces them with zero **[12],** the flipping method which flips the sign of negative eigenvalues so as to form a PSD kernel matrix **[13],** the diffusion method which takes the data distribution into account by replacing the eigenvalues with an exponential form **[14],** and the shifting method which shifts eigenvalues to ensure the nonnegativity of all the eigenvalues **[15].** The LK-method shares some similarity with the flipping method in that negative eigenvalues become the absolute values of themselves. Considering the fact that the denoising method, which neglects the negative eigenvalues, also yields good classification results, we propose a novel method treating eigenvalues in ascending order.

Another problem with the previous model was that even though the weighted q-gram method considered the similarity between two different q-grams, the importance of the q-grams in the context of the whole glycan structures themselves was not taken into account. From a biological perspective, the variability of the sugars near the leaves is larger than those near the root **[5].** Thus, employing the similarity matrix developed by the LK-method, we developed a biochemically-weighted kernel (BioLK-method) utilizing biological knowledge by adding weight based on the layer information $l^i$ of q-grams with while $e^{\alpha l^i}$ ensuring the PSD property of the similarity matrix.

The effectiveness of our BioLK-method was then compared with the LK-method, the representative of the weighted q-gram method in terms of predictive performance of glycan classification and motif extraction. Our newly developed method exhibited comparable classification performance, if not better, with the LK-method. Moreover, our new method could capture biologically meaningful glycan substructures through feature selection while the LK-method failed to do so.

**Methodology:**

Our work incorporates two innovations. The first one is to perform a delicate transformation on the non-PSD similarity matrix constructed in one of the representatives of the weighted q-gram method: the LK-method. The second is to incorporate


© 2011 Biomedical Informatics

existing biological information when computing the kernel matrix. Major contribution of this paper is to propose a biologically significant kernel that is robust in classification as well as in motif selection. We first describe the similarity matrix construction method used for the LK-method that considers the similarity of layers, monosaccharides, glycosidic bonds and geometric tree structures among the q-grams. Based on the existing similarity matrix, a PSD similarity matrix using techniques of spectrum transformation is created. We further develop the novel kernel by combining the biological importance of different q-grams. Different experiments of binary classification and feature selection are performed on the new kernel with SVMs **(See supplementary material for detailed description).**

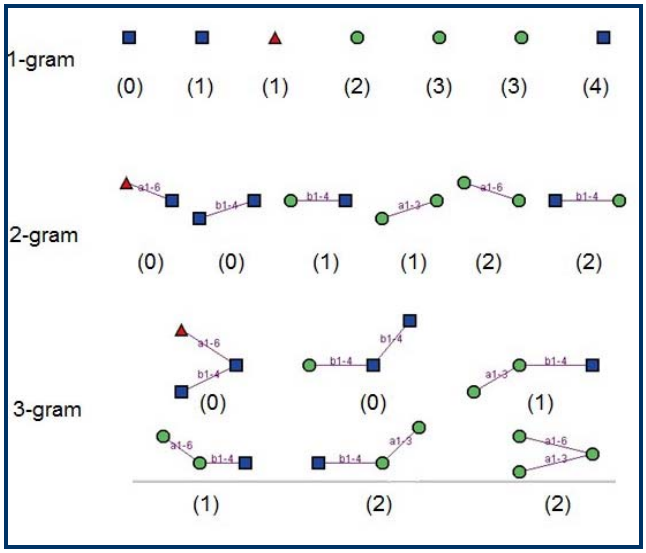


**Figure 2:** q-gram decomposition of glycan in Fig 1: q=1, 2, 3

**Discussion:**

*Materials*

Three sets of glycan data are used to evaluate the classification and feature selection performance of our developed method. They are illustrated in **(Table 1, see supplementary material)**. Glycan structures in two of the data sets are retrieved from the KEGG/GLYCAN database **[1]** with annotations from the CarbBank/CCSD database **[16].** One pertains to leukemia consisting of 355 structures originating from four human blood components: leukemic cells, erythrocytes, serum and plasma, containing 162, 111, 85 and 73 examples respectively. Another data set pertains to cystic fibrosis, containing 89 glycans related to cystic fibrosis, 107 related to respiratory mucin and 101 related to bronchial mucin. For these leukemia and cystic data sets, the total number of glycans is not the sum of each subclass because some glycans belong to several classes. In order to assess the generality of our kernel method in extracting meaningful substructures, we further utilized another data set obtained from the CFG **[2].** We obtained O-linked and N-linked glycan profile data extracted from the brain of mouse strain C57BL/6 (Mouse Strain, http://www.functionalglycomics.org/glycomics/common/jsp/samples/searchSample.jsp?templateKey=1&12=Tissue&operation=refine), which consisted of 47 structures in Wildtype and 50 structures in FucTIV+VII knockout mice.

*Classification and Feature Selection*

The effectiveness of our BioLK-method was evaluated through comparison with the LK-method in terms of performance of both classification and feature selection. Because the BioLK-method involves the determination of $\alpha$ beforehand, a program was run to find an optimal $\alpha$ in a statistical sense. The optimal $\alpha$ for the leukemia data set was 0.1, with 0.35 for the mouse data set and 0.85 for the cystic fibrosis data set. These results were consistent with our previous analysis. Since for the leukemia data set, there are in total 6527 features involved, it is very sensitive to large $\alpha$, while for the cystic fibrosis data set, which contains only 1260 features, it is reasonable that the optimal $\alpha$ is relatively large. In the mouse data set, the number of features altogether was 4214, and the corresponding optimal also lies in $\alpha = 0.35$ between.

*Classification Performance*

Table's 2-4 lists the performance of the SVM classifier for the LK-method and the BioLK-method as tested on our three data sets. We employed the Area Under the ROC Curve (AUC) measured by five-fold cross-validation run 10 times to evaluate the performance. For each $q$ ($q = 1, 2, \ldots, 9$), the tables illustrate the average AUC value over the 10 runs with standard deviations. It is clear to see that both LK-method and BioLK-method show comparable classification performance.

For the leukemia data, the classification performance always achieves accuracy greater than 89%. In the cystic fibrosis data set, the classification accuracy decreases slightly, but still achieves around 80% on average. For $q = 9$ in this data set, the performance goes down to 53% which is reasonable since this data set is much less complex when compared to the other two data sets, reflecting the fact that the number of features involved in 9-gram classification are few. For the mouse data set, the classification performance is also high, achieving accuracies in the 80% range.

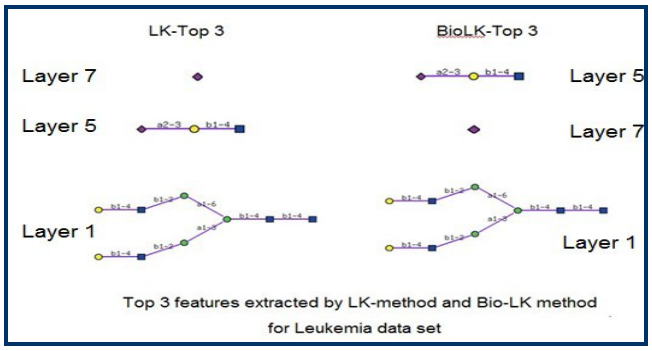


**Figure 3:** Top 3 features on the leukemia dataset. The top-scoring features extracted by the BioLK-method are the trimer structure 'NeuAc$\alpha$2-3Gal$\beta$1-4GlcNAc'found at layer 5. The substructure with the second highest score is the monomer structure 'Neu5Ac' found at layer 7. In contrast, the LK-method captures the features in reverse order.

*Feature Selection*

Both the direct usage of the similarity matrix and the incorporation of the BioWeight matrix in kernel construction enhance our confidence in extracting accurate features. The effectiveness of our BioLK-method in feature selection is assessed in comparison with the LK-method on the three glycan


© 2011 Biomedical Informatics

data sets. Figures 3-5 illustrate the top three features extracted by the LK-method and the BioLK-method. For better illustration, the corresponding figures of the features can be accessed (available with authors).

As shown in **Figure 3**, the top-scoring features extracted by the BioLK-method is the trimer structure 'NeuAc$\alpha$2-3Gal$\beta$1-4GlcNAc' found at layer 5. This precisely corresponds to the substructure in previous works **[5, 6].** The substructure with the second highest score is the monomer structure ' Neu5Ac ' found at layer 7, which is also consistent with the literature **[6].** In contrast, the LK-method captures the features in reverse order. In fact, our results are more reasonable due to the fact that *A.cylindracea* galectin (ACG) is known to specifically bind to the trimer structure, whereas sialic acid is known to appear in many tumor cells. Thus ' Neu5Ac ' is considered to be a more generalized result, whereas the trimer is more specific.

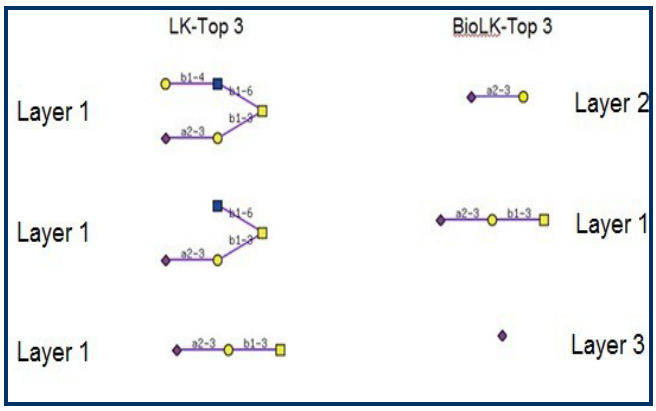


**Figure 4:** Top 3 features on the cystic dataset. The highest score using the BioLK-method is achieved by a dimmer ' NeuAc$\alpha$2-3Gal 'at layer 2, which is often found at the non-reducing end of glycan structures. The top three structures captured by the BioLK-method are all $\alpha$2-3 sialylated structures which are consistent with the literature as well. However, the features captured by the LK-method are structures which include the root, which may indicate that it is overfitting to the data.

**Figure 4** lists the top three motif candidates extracted by the LK-method and the BioLK-method in the cystic fibrosis data set. The highest score using the BioLK-method is achieved by a dimer 'NeuAc2-3Gal' at layer 2, which is often found at the non-reducing end of glycan structures. Although this is slightly different from the result predicted by **[8]** which captures this structure as the second highest score, it is acceptable since in their method, information indicating root and leaf nodes is incorporated directly into the q-gram data. Our method is still consistent with the result that the top scoring CF-related structure is α2-3 sialylated structures, which corresponds with the literature **[17, 18].** It is also consistent with the result that the top scoring features extracted included monomers and dimers. We note that the top three structures captured by the BioLK-method are all α2-3 sialylated structures which are consistent with the literature as well. However, the features captured by the LK-method are structures which include the root, which may indicate that it is overfitting to the data. Biologically speaking, one would also assume that the structures at the terminal end, and in particular the non-reducing end, are those that would be considered to be drug targets, as opposed to the larger structures containing common core structures. Indeed, the results of the LK-method all contained a common O-glycan core structure, whereas the BioLK-method extracted the common terminal structure from the non-reducing end of these results.

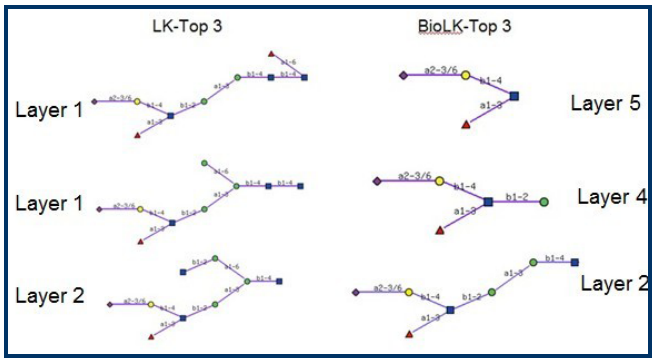


**Figure 5:** Top 3 features on the mouse_fuc dataset. The feature with the top score extracted by the BioLK-method was ' NeuAc$\alpha$2-3/6Gal$\beta$1-4(Fuc$\alpha$1-3)GlcNAc 'at layer 5, which is sialyl-Lewis X, a previously discovered motif for this sample **[2].** On the other hand, the LK-method always captured larger structures from the core.

In order to show the robustness of our method in feature extraction, we tested it on glycan profile data of mouse brain collected from FucTIV+VII knockout mice, as provided by the CFG. We then compared the feature selection results as performed by the LK-method and the BioLK-method. The top three features extracted by both methods are listed in **Figure 5**. The feature with the top score extracted by the BioLK-method was' NeuAc$\alpha$2-3/6Gal$\beta$1-4(Fuc$\alpha$1-3)GlcNAc 'at layer 5, which is sialyl-Lewis, a previously discovered motif for this sample **[2].** On the other hand, the LK-method always captured larger structures from the core. Similarly to the cystic fibrosis sample, the top results of the BioLK-method contained the common non-reducing end structure of the top results of the LK-method, thus indicating that the LK-method is probably overfitting to the data, whereas our method produced precisely the unique substructures (features) of the target data set.

**Conclusion:**

In this work, we developed a new tree kernel based on the linkage kernel constructed using the weighted q-gram method, but we included two major novelties that enabled us to obtain highly accurate results, which previous methods were unable to obtain. First, the techniques of direct usage of the non-PSD similarity matrix to form a positive one largely aided in maintaining the biological properties of the data. Many kernels developed in bioinformatics ignore this important property in kernels, and we show that this is indeed important. Secondly, the incorporation of weighted layer information of q-grams together enables high accuracy in discriminating between classification groups as well as in the correct detection of glycan motifs with flexible size. This confirms the necessity of including weighted layer information of q-grams in order to construct more biologically meaningful tree kernels.


© 2011 Biomedical Informatics

Indeed, our results were shown to correspond well with known glycan motifs obtained through experimental results, whereas the previous methods were unable to obtain the same results. Thus, we claim that our new kernel contributes greatly to the field of glycoinformatics to obtain a greater understanding of glycan functions in various areas of biological research.

**Authors contributions**:
JH came up with the idea. JH and KFA designed the research. KFA gave invaluable suggestions and created q-grams of the data sets. JH performed the research and analyzed the results. WKC supported the provided guidance on how to conduct the research. JH, KFA and WKC wrote the paper. All authors read and approved the final manuscript.

**Acknowledgement**:
Research supported in part by GRF Grant, HKU Strategy Research Theme fund on Computational Sciences, National Natural Science Foundation of China Grant No. 10971075 and Guangdong Provincial Natural Science Grant No. 9151063101000021.

**References:**
**[1]** Hashimoto K *et al*. *Glycobiology*. 2006 **16**: 63R [PMID: 16014746].
**[2]** Parry S *et al*. *Glycobiology*. 2007 **17**: 646 [PMID: 17341505].
**[3]** Aoki-Kinoshita KF *et al*. *Bioinformatics*. 2006 **15**: e25 [PMID: 16873479].
**[4]** Hashimoto K *et al*. *ACM Transactions on Knowledge Discovery from Data*. 2008 **2**: 6.
**[5]** Hizukuri Y *et al*. *Carbohydrate Res*. 2005 **340**: 2270 [PMID: 16095580].
**[6]** Yamanishi Y *et al*. *Bioinformatics*. 2007 **23**: 1211 [PMID: 17344232].
**[7]** Kuboyama T *et al*. *Information and Media Technologies*. 2007 **2**: 292.
**[8]** Kuboyama T *et al*. *Genome Inform*. 2006 **17**: 25 [PMID: 17503376].
**[9]** Li LM *et al*. *BMC Bioinformatics*. 2010 **18**: 11 Suppl [PMID: 20122206].
**[10]** Ben-Hur A *et al*. *PLoS Computational Biol*. 2008 **4**: e10000173 [PMID: 18974822].
**[11]** Hofmann T *et al*. *Annals of Statistics*. 2008 **36**: 1171.
**[12]** Pekalska E *et al*. *Journal of Machine Learning Research*. 2002 2, 175.
**[13]** Graepel T *et al*. *NIPS*. 1999 **11**: 438
**[14]** Neuhaus M *et al*. *Spatial Vision*. 2009 **22**: 425 [PMID: 19814905].
**[15]** Roth V *et al*. *IEEE Trans on PAMI*. 2003 **25**: 1540.
**[16]** Doubet S & Albersheim P. *Glycobiology*. 1992 **2**: 505 [PMID: 1472756].
**[17]** Degroote S *et al*. *Glycobiology*. 1999 **9**: 1199 [PMID: 10536036].
**[18]** Mawhinney TP *et al*. *Carbohydrate Res*.1992 **235**: 179 [PMID: 1473102].
**[19]** Hattori M *et al*. *J Am Chem Soc*. 2003 **125**: 11853 [PMID: 14505407].
**[20]** Wu G et al. Technical Report. UCSB. 2005

**Edited by TW Tan**
**Citation: Jiang *et al.*** Bioinformation 7(8): 405-412 (2011)
**License statement:** This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited

BIOINFORMATION open access

# Supplementary materials:

**Methodology:**

***q-gram Representation***

The following terminology will be used throughout this paper. We use a labeled ordered rooted tree to characterize the molecular structure of a glycan. For glycans, the vertex labels stand for the monosaccharide type while the edge labels represent glycosidic bonds. Since the order of the children is significant, the tree of glycans is considered ordered. The monosaccharide at the reducing end is considered the root. We also define the concept of a *layer* for subtree rooted at a monosaccharide (i.e. a vertex) as the distance of the vertex from the root.

We formulate q-grams for *labeled ordered rooted* trees. A q-gram is defined as a tree with q nodes isomorphic to a path where every node has at most two adjacent nodes, for $q \geq 1$. A q-gram representation of a specific glycan is denoted as a vector of length N, where N is the total number of q-grams within the glycan data set being investigated. **Figure 2** shows the q-gram decomposition of the given glycan structure (data not shown, please check with authors). In total, if the glycan data set contains N glycans $\{g_1, g_2, \cdots g_N\}$, we denote the set of all q-grams existing in these N glycans to be a q-gram set: $\Phi_q = \{\phi_q^1, \phi_q^2, \cdots \phi_q^{n_q}\}$. For a specific glycan $g_i$ in the data set, the q-gram representation is a column vector $x_i^q = \left[x_{1i}^q, x_{2i}^q, \cdots x_{n_q i}^q\right]^T$ where $x_{li}^q$ is the number of lth q-gram in the glycan $g_i$.

***q-gram Similarity in LK-method***

Next, we describe the concept of similarity between two glycans (each represented as a q-gram) as defined in the LK-method. For each q-gram, there are $q$ monosaccharides and $q-1$ glycosidic bonds linking them to one another. When $q=1$, we just consider a single monosaccharide instead. Suppose a q-gram is characterized by $\phi_q = \{l, M, B, \sigma\}$, where $l$ is the layer of the q-gram, $M$ is the ordered set of monosaccharides it contains, $B$ stands for the corresponding chemical bonds and $\sigma$ represents the structure shape (i.e., linear, branched, etc.) of this q-gram.

Given two q-grams $\phi_q^i = \{l^i, M^i, B^i, \sigma^i\}$ and $\phi_q^j = \{l^j, M^j, B^j, \sigma^j\}$, the similarity between the two q-grams are defined as:

$$S_q(\phi_q^i, \phi_q^j) = S^\sigma(\sigma^i, \sigma^j) \cdot S^l(l^i, l^j) \cdot \prod_{k=1}^{q} S^M(m_k^i, m_k^j) \cdot \prod_{k=1}^{q-1} S^B(b_k^i, b_k^j)$$

Where $S^\sigma(\sigma^i, \sigma^j)$ is the similarity between the shapes of the two q-grams, $S^l(l^i, l^j)$ is the similarity between the layers of the two q-grams, $S^M(m_k^i, m_k^j)$ is the similarity of the corresponding monosaccharides, and $S^B(b_k^i, b_k^j)$ is the similarity of the chemical bonds.

The similarity of shape between two q-grams is defined as:

$$S^\sigma(\sigma^i, \sigma^j) = \begin{cases} 1, & \sigma^i = \sigma^j \\ 0, & otherwise \end{cases}$$

The similarity of layers is defined using the distance of layers:

$$S^l(l^i, l^j) = 1 - \frac{|l^i - l^j|}{\max(l)}$$

The similarity among monosacccharides is obtained from the chemical structure comparison method SIMCOMP developed by **[19].** For the bond similarity, it is defined according to their chemical meanings (additional data available with authors).

***The linkage kernel in the LK-method then can be created by:***

$$K_q^{LK} = V_q^T \cdot S_q^T \cdot S_q \cdot V_q$$

Where $V_q$ is the q-gram representation matrix of the glycan data set.

***Biochemically-Weighted Kernel Construction: BioLK-method***

In order to bypass the issue of the non-PSD property in kernel construction, the LK-method uses $S^T S$ as a replacement for the similarity matrix $S$ However, from a biological standpoint, the kernel should be constructed as follows:

$K_q = V_q^T \cdot S_q \cdot V_q$ kq = $v^t$q

Here our objective is to directly use the indefinite similarity measures to construct both a new one that is PSD and that biologically shares more similarity with the original similarity matrix.

***Mathematically, the similarity matrix S can be decomposed as follows:***

$$S = X \cdot P \cdot X^T$$

Where $X$ is the unit eigenvector matrix corresponding to the eigenvalues sorted in ascending order, P is the diagonal matrix of eigenvalues sorted in ascending order. Usually the similarity matrix constructed is non-PSD which means there are negative

 © 2011 Biomedical Informatics

eigenvalues. Taking into consideration the fact that the **denoising method** and the **flipping method** (described in the Introduction part) both can yield high precision in classification for protein datasets **[20],** we may get some clues in constructing a new similarity matrix based on the original non-PSD one. Basically, we should keep the original positive eigenvalues while avoiding the magnification of negative eigenvalues. Therefore, the new similarity matrix is proposed as:

$$\hat{S} = X \cdot \hat{P} \cdot X^T$$

where

$$\hat{P} = \begin{pmatrix} \hat{\lambda}_1 & 0 & \cdots & 0 \\ 0 & \hat{\lambda}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\lambda}_n \end{pmatrix}$$

and $\hat{\lambda}_i$, $i = 1, 2, \ldots n$ are defined as:

$$\hat{\lambda}_i = \begin{cases} e^{\lambda_i - 1}, & \lambda_i \le 1 \\ \lambda_i, & otherwise \end{cases}$$

The newly developed similarity matrix in this context is PSD. It preserves the ascending property of eigenvalues without changing most of the positive eigenvalues. Moreover, the effect of negative eigenvalues is also included without magnification.

However, the similarity matrix only considers the similarity of the geometric structure, monosaccharides and glycosidic bonds among q-grams. Glycans exhibit the property that substructures near the leaf are more variable. It is therefore desirable that we include this biological information in kernel construction. This may play a pivotal role in capturing exact motifs in feature selection.

We measure the importance of q-grams by defining BioWeight for them according to the layer of q-grams.

$$BioWeight(\phi_q^i) = e^{\alpha \cdot l^i}, \ \alpha \in [0,1]$$

The kernel therefore can be constructed as follows:

$$K_q^{BioLK} = V_q^T \cdot BioWeight \cdot \hat{S} \cdot BioWeight \cdot V_q$$

For the **BioWeight** matrix $\alpha$ is a parameter to be predetermined. It endows the q-grams as a unit with significance in the whole feature set. The function we choose for **BioWeight** originates from a weight function used in constructing the similarity matrix for the leukemia data set **[5]**. The two functions $e^{\alpha l^i}$ and $1 - e^{-\alpha l^i}$ share similarity in putting more weight on the substructures in the variable region. The reason for $\alpha$ as a parameter to be predetermined in our paper is that for different data sets, the number of features embedded varies from one to another. In the case of large data sets with numerous complicated features, $\alpha$ should be set to a smaller value because large $\alpha$ will pose too much significance on the variable part, thereby bringing about side effects to extract wrong substructures. On the other hand, relatively smaller data sets contain fewer and simpler structures, under which circumstance the data would be less sensitive to large $\alpha$. Values of $\alpha$ that are too small, on the other hand, would not help much to differentiate different features. Thus, while greater $\alpha$ may contribute to better feature selection results, they must not be too large, but not so small that feature selection cannot be performed well. We have thus developed an algorithm to select the appropriate values for $\alpha$ given the size of the feature set (data not shown).

***Feature Selection***

For $q = 1, 2, \ldots, 9$, we use the discriminant score $\delta(x)$ obtained from the trained SVM to represent the contribution of each q-gram pattern. The feature score representing the importance of feature f is defined as follows:

$$F(f) = \sum_{x \in X} \delta(x) \cdot I_x(f)$$

where $x$ is the glycan, and $X$ is the whole glycan data set being investigated.

$$I_x(f) = \begin{cases} 1, & If \ x \ contains \ feature \ f \\ 0, & otherwise. \end{cases}$$

The features with higher feature scores may be potential motifs. We select the most likely substructures under this mechanism.

**Table 1**: Data set composition

| | | | | |
|---|---|---|---|---|
| Leukemia 162 | Erythrocyte 111 | Plasma 73 | Serum 85 | Total 355 |
| Cystic 107 | Respiratory 89 | Bronchial 101 | | Total 177 |
| Wildtype 47 | FucTIV+VII 50 | | | Total 97 |


© 2011 Biomedical Informatics

**Table 2**: For each $q$ ($q = 1, 2, \ldots, 9$), the table illustrates the average AUC value over the 10 runs with standard deviations. Both LK-method and BioLK-method show comparable classification performance. For the leukemia data, the classification performance always achieves accuracy greater than 89%.

| q | LK-method | BioLK-method |
|---|---|---|
| 1 | 0.906±0.002 | 0.914±0.004 |
| 2 | 0.952±0.004 | 0.959±0.003 |
| 3 | 0.964±0.002 | 0.959±0.005 |
| 4 | 0.957±0.003 | 0.951±0.005 |
| 5 | 0.948±0.003 | 0.948±0.005 |
| 6 | 0.924±0.004 | 0.934±0.003 |
| 7 | 0.927±0.003 | 0.925±0.006 |
| 8 | 0.900±0.007 | 0.904±0.004 |
| 9 | 0.893±0.008 | 0.893±0.006 |

**Table 3**: For each $q$ ($q = 1, 2, \ldots, 9$), the table illustrates the average AUC value over the 10 runs with standard deviations. Both LK-method and BioLK-method show comparable classification performance. In the cystic fibrosis data set, the classification accuracy decreases slightly, but still achieves around 80% on average. For $q = 9$ in this data set, the performance goes down to 53% which is reasonable since this data set is much less complex when compared to the other two data sets, reflecting the fact that the number of features involved in 9-gram classification are few.

| q | LK-method | BioLK-method |
|---|---|---|
| 1 | 0.777±0.011 | 0.792±0.014 |
| 2 | 0.78±0.020 | 0.792±0.016 |
| 3 | 0.798±0.018 | 0.798±0.014 |
| 4 | 0.793±0.015 | 0.815±0.022 |
| 5 | 0.788±0.017 | 0.801±0.021 |
| 6 | 0.746±0.022 | 0.755±0.020 |
| 7 | 0.700±0.025 | 0.691±0.030 |
| 8 | 0.613±0.024 | 0.612±0.031 |
| 9 | 0.527±0.028 | 0.521±0.033 |

**Table 4**: For each $q$ ($q = 1, 2, \ldots, 9$), the table illustrates the average AUC value over the 10 runs with standard deviations. Both LK-method and BioLK-method show comparable classification performance. For the mouse data set, the classification performance is also high, achieving accuracies in the 80% range.

| q | LK-method | BioLK-method |
|---|---|---|
| 1 | 0.718±0.019 | 0.726±0.02 |
| 2 | 0.735±0.022 | 0.742±0.014 |
| 3 | 0.787±0.016 | 0.804±0.031 |
| 4 | 0.916±0.017 | 0.905±0.015 |
| 5 | 0.880±0.02 | 0.885±0.012 |
| 6 | 0.860±0.012 | 0.878±0.023 |
| 7 | 0.875±0.015 | 0.889±0.019 |
| 8 | 0.879±0.021 | 0.897±0.013 |
| 9 | 0.868±0.013 | 0.872±0.024 |


© 2011 Biomedical Informatics