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: $\Phi_q = \left\{\phi_q^1, \phi_q^2, \cdots \phi_q^{n_q}\right\}$. 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 σ 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:

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

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

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 α 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 α 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, α should be set to a smaller value because large α 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 α. Values of α that are too small, on the other hand, would not help much to differentiate different features. Thus, while greater α 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 α 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

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