Glycan Linkage Substitution Matrix Rooted on Tree Alignments

Kiyoko F. Aoki[1] **Hiroshi Mamitsuka**[1] **Minoru Kanehisa**[1]

kiyoko@kuicr.kyoto-u.ac.jp mami@kuicr.kyoto-u.ac.jp kanehisa@kuicr.kyoto-u.ac.jp

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho,
Uji, Kyoto 611-0011, Japan

Keywords: glycans, tree alignments, score matrix, substitution matrix, glycan linkages, glycobiology

1 Introduction

Glycans are chains of monosaccharides, known to be extremely vital for the development and functioning of multi-cellular organisms. Matching and alignment algorithms for glycans were introduced based on proven efficient methods in bioinformatics for protein sequences [2]. These algorithms consequently can be taken advantage of in the form of statistical analyses. These algorithms were designed to provide alignment scores for glycan linkage alignments such that the frequency of substitution of glycan linkages may be analyzed. In doing so, glycan linkages deemed to be similar to one another may be extracted from the abundant glycan data currently available [6]. This leads to not only a straightforward technique for linkage analysis of glycans, but it also leeds to a better understanding of enzymatic processes in glycan biosynthesis, an area of work which is just beginning to take off these past few years, but is yet to reach its peak in altitude.

2 Methods

Glycan Score Matrix Generation In contrast to amino acid score matrices based on proteins, for glycans, there are several considerations we needed to take in our procedure to calculate our score matrix. Firstly, we needed to determine the most meaningful, but most basic, components for glycans. Considering that differences in anomeric and hydroxyl groups between linkages is rather important in glycan recognition, it was necessary to incorporate as much linkage information as possible into our matrix. Therefore, we focused on the alignments of linkages as opposed to monosaccharides. This of course could potentially generate an exponential number of entries. However, we show that the variety of linkages actually present in nature are limited, and our procedure is able to capture the most important ones. Another challenge was in the definition of family classification. Although glycan classes are available, the data within these classes are currently rather sparse and/or redundant. Therefore, we created our own set of representative glycans and generated our own "glycan blocks" for our score matrix. We approach the concept of blocks from a slightly different perspective compared to BLOSUM in that due to the small sizes of our structures, it was not meaningful to look for "highly conserved regions" per se. Therefore, we obtained our glycan blocks based on clusterings of each class at different levels of similarity.

Matrix Calculation Within each block of glycans, a pairwise alignment was performed for every pair and the frequency of linkage alignments was counted. Let f_{ij} denote the frequency of aligning linkages i and j. The probability of occurrence q_{ij} of this i and j alignment was then calculated by dividing f_{ij} by the total number of all alignments. Next, the probability of a particular linkage i occurring in an alignment was calculated as $p_i = q_{ii} + \sum_{i \neq j} q_{ij}/2$. This takes into account the reflectivity of the alignments. The expected probability of occurrence of an alignment of i and j is thus $e_{ij} = p_i p_j$(for $i = j$) and $2p_i p_j$ (for $i \neq j$). We can then calculate the score for aligning linkages i and j by the formula $s_{ij} = \log_2(q_{ij}/e_{ij})$. Additionally, we could calculate the expected score, or E-Score, for the matrix with $E = \sum_i \sum_j^i p_i p_j s_{ij}$.

3 Results

Our final matrix consists of 1281 entries of linkage pairs and its E-Score is -0.27809, a preferable value in that it is accurately gives more weight to truly significant alignments [1].

The Area Under the ROC (Receiver Operating Characteristics) Curve, or AUC[5], compares *sensitivity* vs. *false positive rate*[1]. We plotted these values and measured the area under the given curve. For our glycans, we were able to obtain very interesting results in terms of AUC performance. We created "sub-matrices" out of glycans from within individual classes to compare with the AUC performance of our final matrix. Table 1 lists the performance of the original matching algorithm (match), the alignment scores using the selected preliminary matrices for each corresponding class (sub-matrix) and our final combined matrix (final matrix). Using this final matrix, the O-Glycans actually performed even better than when we used its own best matrix, implying that it may contain many unknown linkage information that could not be fully captured by its own matrix.

Table 1: Final score matrix performance

	N-Glycans	O-Glycans	Sphingolipids
match	98.6%	79.9%	90.8%
sub-matrix	99.0%	89.2%	93.3%
final matrix	99.0%	90.1%	91.0%

Table 2: Selected matrix entries

Linkage	Log-odds score
Fucα1-6GlcNAc	2.45254
GalNAcβ1-4Gal	1.24666
Galα1-3Gal	0.00718184

Table 2 lists some of the entries from our glycan score matrix. The highest scorer is a fucosylated linkage onto *N*-acetylglucosamine, which is often found on the chitobiose core of N-Glycans [7, 8]. Another high scorer is GalNAcβ1-4Gal, which is part of a terminal linkage and is found in the ganglioside biosynthesis pathway [3]. Finally, the Galα1-3Gal linkage structure may be found as the terminal linkage of the α-gal epitope, a common structure in mammals recognized by T-cells [4].

4 Discussion

Our score matrix has, in effect, specified pairs of similar linkages, just as amino acid score matrices produce pairs or sets of amino acids of similar chemical properties. Correspondingly, this matrix can provide confidence values such as E-values in glycan alignments, as done in BLAST. Future work may entail delving into the physico-chemical properties of glycan linkages based on our matrix.

References

[1] Altschul, S. F., Amino acid substitution matrices from an information theoretic perspective, *J. Mol. Biol.*, 219:555–565, 1991.

[2] Aoki, K.F., Yamaguchi, A., Ueda, N., Akutsu, T., Mamitsuka, H., Goto, S., and Kanehisa, M., KCaM (KEGG Carbohydrate Matcher): a software tool for analyzing the structures of carbohydrate sugar chains, *Nucl. Acids Res.*, 32:W267–W272, 2004.

[3] Dicesare, J.L. and Dain, J.A., The enzymic synthesis of ganglioside. IV. UDP-N-acetylgalactosamine: (N-acetylneuraminyl)-galactosylglucosyl ceramide N-acetylgalactosaminyltransferase in rat brain, *Biochim. Biophys. Acta*, 231(2):385–393, 1971.

[4] Galili, U., The α-gal epitope (Galα1-3Galβ1-4GlcNAc-R) in xenotransplantation, *Biochimie*, 83:557–563, 2001.

[5] Hand, D.J. and Till R.J., A simple generalisation of the area under the ROC curve for multiple classification problems, *Machine Learning*, 45:171–186, 2001.

[6] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M., The KEGG resource for deciphering the genome, *Nucl. Acids Res.*, 32:D277–D280, 2004.

[7] Lowe, J.B. and Marth, J.D., A genetic approach to mammalian glycan function, *Annu. Rev. Biochem.*, 72:643–691, 2003.

[8] Noda, K., Miyoshi, E., Gu, J., *et al.*, Relationship between elevated fx expression and increased production of gdp-l-fucose, a common donor substrate for fucosylation in human hepatocellular carcinoma and hepatoma cell lines, *Cancer Research*, 63:6282–6289, 2003.

[1] Sensitivity is the proportion of the number of correctly categorized glycans to the total number of glycans in the class, and the false positive rate is the proportion of the number of falsely categorized glycans that do not belong to the class to the total number of glycans that do not belong to the class.