Statistical Significance of Tree Similarity Scores

Kiyoko F. Aoki
kiyoko@kuicr.kyoto-u.ac.jp

Atsuko Yamaguchi
atsuko@kuicr.kyoto-u.ac.jp

Yasushi Okuno
okuno@kuicr.kyoto-u.ac.jp

Tatsuya Akutsu
takutsu@kuicr.kyoto-u.ac.jp

Nobuhisa Ueda
ueda@kuicr.kyoto-u.ac.jp

Minoru Kanehisa
kanehisa@kuicr.kyoto-u.ac.jp

Hiroshi Mamitsuka
mami@kuicr.kyoto-u.ac.jp

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

Keywords: tree-matching, carbohydrate sugar chains, glycans, glycobiology

1 Introduction

New methodologies for performing pairwise tree-matching on carbohydrate sugar chain data were introduced in [3], in which well-known sequence alignment algorithms [12] were extended and an already known polynomial-time graph algorithm for finding the maximum common subtree (MCST) of two trees [7] was used to implement what is called the KEGG Carbohydrate Matcher, or KCaM. These new methodologies are currently available on the web via KEGG Glycan [9, 15]. We make note of some appealing work related to not only KCaM but biological tree-structure matching in general.

2 Method Analysis

One interesting point to make note of from KCaM is that similarity scores (as opposed to distance measures) are calculated such that statistical analysis can be more readily performed, as done in [1, 14] for sequence similarity analysis. It was shown in [3] that the similarity score distribution of a random sampling of trees fits an extreme value distribution, similar to sequence similarity score distributions [2]. However, the methods presented are still very basic as the parameters have not been tuned and more biologically meaningful weighting parameters have not been applied. Thus, a variety of work is possible, including the interpretation of similarity scores using p-values and E-values, and more biologically meaningful similarity scores based on score matrices for glycans.

There has been a lot of work on the statistical significance of sequence alignment scores, where some claim that there is even no need to fit an extreme-value distribution [11], as long as the scoring scheme (i.e., gap penalty and substitution matrix or profile), sequence composition and length are taken into account. Therefore, for the case of tree comparison and tree alignment scores of glycans, the statistical significance of these scores will entail the analysis of the distribution of the known monosaccharides compared with the distribution of those monosaccharides in a match, and seeing how a formula similar to PAM [5] can be applied. If such a scoring matrix can be developed and verified for tree structures, then future work on the similarities of tree structures can also begin. Here we make note of the fact that there is a difference between the tree alignment methods presented in KCaM and those of, say, phylogenetic trees used for sequence homology search. In KCaM, the objective is to find similar tree structures in and of themselves, whereas evolutionary and phylogenetic trees are used for finding similar sequences, which are applied to the leaves of the input trees.

3 Discussion

The results of the matches of the small set of glycans presented in [3] is just one step towards obtaining a full understanding of the significance of similar glycans. It has taken almost twenty years to get to the point where there is a stable understanding and usage of procedures for analyzing protein sequence homologies, since pioneers such as Karlin-Altschul [10] and Smith-Waterman [13] first established alignment methods and performed similarity score analysis for sequences.

There is a plethora of work currently underway on tree structures in bioinformatics [8]. But work in glycans is slow due to the complexity of not only the structures themselves but also their biosynthesis [4, 6]. Therefore, the amount of data is limited. However, with the release of KEGG Glycan and based on the work of our pioneers, we can expect another flurry of activity in statistically analyzing the similarity scores of trees in the near future.

References

[1] Altschul, S. and Gish, G., Local alignment statistics, *Methods Enzymol.*, 266:460–480, 1996.

[2] Altschul, S.F., Bundschuh, R., Olsen, R., and Hwa, T., The estimation of statistical parameters for local alignment score distributions, *Nucleic Acids Research*, 29(2):351–361, 2001.

[3] Aoki, K.F., Yamaguchi, A., Okuno, Y., Akutsu, T., Ueda, N., Kanehisa, M., and Mamitsuka, H., Efficient tree-matching methods for accurate carbohydrate database queries, *Genome Informatics*, 14:134–143, 2003.

[4] Bertozzi, C.R. and Kiessling, L.L., Carbohydrates and glycobiology review: chemical glycobiology, *Science*, 291:2357–2364, 2001.

[5] Dayhoff, M.O., Barker, W.C., and Hunt, L.T., Establishing homologies in protein sequences, *Methods in Enzymology*, 91:524, 1983.

[6] Drickamer, K., Two distinct classes of carbohydrate-recognition domains in animal lectins, *J. Biol. Chem.*, 263:9557–9560, 1988.

[7] Edmonds, J. and Matula, D., An algorithm for subtree identification, *SIAM Rev.*, 10:273–274, 1968.

[8] Höchsmann, M., Töller, T., Giegerich, R., and Kurtz, S., Local similarity in RNA secondary structures, *Proceedings of CS Bioinformatics*, IEEE Computer Society Press, 159–168, 2003.

[9] Kanehisa, M. *et al.*, The KEGG databases at GenomeNet, *Nucleic Acids Res.*, 30:42–46, 2002.

[10] Karlin, S. and Altschul, S.F., Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes, *Proc. Natl. Acad. Sci.*, 87:2264–2268, 1990.

[11] Mott, R., Accurate formula for p-values of gapped local sequence and profile alignments, *J. Mol. Biol.*, 300:649–659, 2000.

[12] Smith, T.F. and Waterman, M.S., Identification of common molecular subsequences, *J. Mol. Biol.*, 147(1):195–197, 1981.

[13] Smith, T.F., Waterman, M.S., and Burks, C., The statistical distribution of nucleic acid similarities, *Nucleic Acids Res.*, 13:645–656, 1985.

[14] Waterman, M. and Vingron, M., Sequence comparison significance and poisson approximation, *Statistical Science*, 9:401–418, 1994.

[15] `http://glycan.genome.ad.jp/`