

Prediction of Glycosyltransferases Synthesizing Glycoconjugates Using Variable-length N-gram Model

Yeon-Dae Kwon¹
yd-kwon@aist.go.jp

Shougo Shimizu²
shimizu@sel.cs.hiroshima-cu.ac.jp

Hisashi Narimatsu¹
h.narimatsu@aist.go.jp

¹ Glycogene Function Team, Research Center for Glycoscience, National Institute of Advanced Industrial Science and Technology, Open Space Laboratory Central-2, 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan

² Department of Computer and Media Technologies, Hiroshima City University, 3-4-1 Ozuka-Higashi, Asa-minami-Ku, Hiroshima 731-3194, Japan

Keywords: glycoconjugate, glycosyltransferase, acceptor, variable-length N-gram model

1 Introduction

We consider the problem of predicting the most likelihood sequence of glycosyltransferases that synthesize a given glycoconjugate. Such information is useful for synthesis of a targeted glycoconjugate. It is empirically known that what kind of glyco is added in some synthesis step is mostly determined depending on only a small portion of the previous structure. Therefore, we adopt a variable-length N-gram model as a prediction model, and evaluate its precision using known human acceptor structure data.

2 Variable-length N-gram Model for Prediction of Glycosyltransferases

Variable-length N-gram model is an extension of N-gram model where the value of N is variable depending on a word history, which enhances precision of N-gram model without losing reliability [1]. We adopted this model as a prediction model for glycoconjugate structures.

We have constructed a database system for glycogene, GGDB [2]. GGDB contains various information about 240 human glycosyltransferases, including their known acceptors. For applying these glycoconjugate data to this model, we encode glycoconjugate data as follows: (i) one *word* consists of one combination of glyco, linkage, and glycosyltransferase, (ii) a *history* is a path from a targeted glyco to its root. That is, we only consider path structures of acceptors, although they have tree structures in general, (iii) the occurrence probability of each path structure is assumed to be equal, (iv) we ignore non-related structures for prediction, which are described in various sources such as papers and books [3]. It is expected that this model works effectively because what kind of glyco is added in a next synthesis step is mostly determined depending on only a small fraction of previous substructures.

Figure 1 shows an example of prediction suffix tree for glycoconjugates. The labels of nodes mean the history (the leftmost glyco is the immediate previous one). The probability distribution attached to each node represents conditional probability of what kind of glyco is added after the history. For example, Gal-labeled node means the previous glyco is Gal, and the probability that GlcNAc is added by b3GnT5 in a next synthesis is 0.2 and so on. When the relative entropy between two probability distributions of parent-child

relationship nodes is within a given threshold, the construction of a prediction suffix tree terminates.

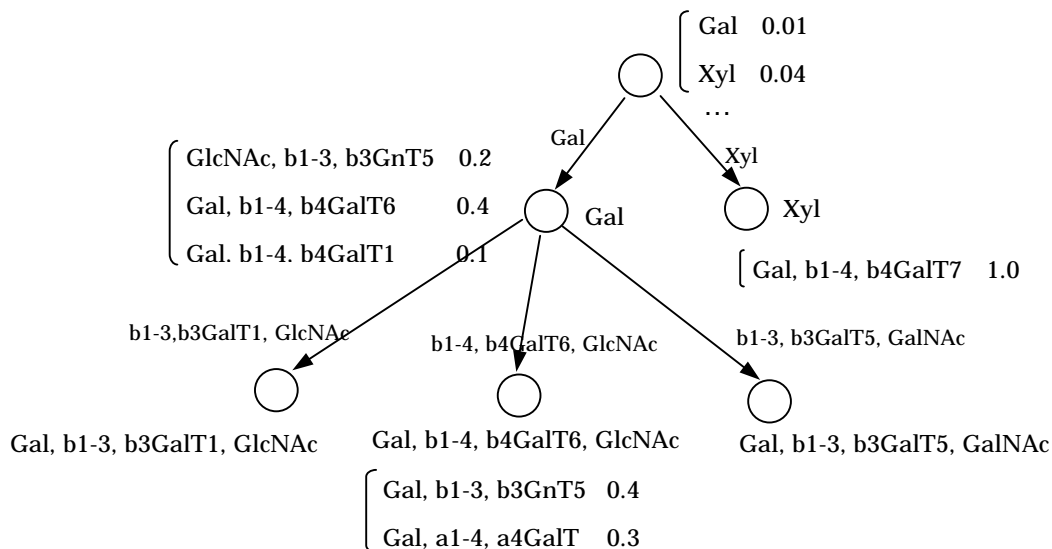


Figure 1: An example of prediction suffix tree for glycoconjugates

3 Experiment

We use 437 paths and 120 paths in acceptors as training sets and test sets, respectively. Queries consist of a sequence of glyco names. A prediction suffix tree is constructed from training sets and outputs likelihood structures using this model. The resulting tree have 36 nodes, where 9 nodes of them consist of 1 word, 23 nodes consist of 2, and 4 nodes consist of 3. The maximum length is set to 4 as a parameter.

This model predicts structures precisely when glycos in a query strongly depends on the previous glycos in turn. For example, for a query GalNAc,Gal,Gal,Glc, the model outputs GalNAc , β 1-3,b3GalNAcT1,Gal , α 1-4,a4GalT,Gal, β 1-4,b4GalT6,Glc as the first rank. However, for queries that contain glycosyltransferases which work only for specific glycoproteins (such as C1GalT, ppGalNAcT, and b3GnT6) or glycolipids (such as b3GalT4, b3GnT5, and b4GalNAcT1), it tends to produce incorrect results. For example, for a query Gal,GalNAc,Gal,Glc, which occurs frequently in glycolipids, the model outputs Gal, β 1-3,C1GalT1,GalNAc, β 1-3,b3GalNAcT2,Gal, β 1-4,b4GalT6,Glc as the third, and this is not the case.

4 Conclusion

We applied variable-length N-gram model for prediction of glycosyltransferases that synthesizes a given glycoconjugate. For further improvements, we plan to incorporate glycoproteins/glycolipids dependencies and tree structures of glycoconjugates into prediction model.

References

- [1] Kita, K., *Computation and Language*, vol. 4, Probabilistic Language Model, University of Tokyo Press, 1999.
- [2] Kwon, Y-D., GGDB: A database system for glycogene, *The Second Symposium of Japanese Consortium for Glycobiology and Glycotechnology*, pp.42-43, 2004.
- [3] Taniguchi, N., Honke, K., and Fukuda, M., *Handbook of Glycosyltransferases and Related Genes*, Springer, 2002.