

Classification of Eukaryotic 7-tms Transmembrane Proteins by Binary Topology Pattern

Yasuhito Inoue¹

s498061@si.hirosaki-u.ac.jp

Yoshiaki Sugiyama¹

gs01610@si.hirosaki-u.ac.jp

Masami Ikeda²

srikeda@cc.hirosaki-u.ac.jp

Toshio Shimizu¹

slsimi@si.hirosaki-u.ac.jp

¹ Department of Electronic Information System Engineering, Faculty of Science and Technology, Hirosaki University, 3, Bunkyo-cho, Hirosaki 036-8561, Japan

² Department of Science of Bioresources, The United Graduate School of Agricultural Sciences, Iwate University, 18-1, Ueda 3-chome, Morioka 020-8550, Japan

Keywords: G protein-coupled receptor, transmembrane segment, length of loop domain, transmembrane binary topology pattern, functional identification

1 Introduction

G protein-coupled receptors (GPCRs) play extremely important functions in our body which are mainly to transduce chemical signals across cell membranes, therefore GPCRs are one of the target proteins in the biomedical field. GPCRs are found in large numbers in most eukaryotic genomes, although the only completely known three-dimensional structure in these families is that of rhodopsin from bovine at 2.8 angstroms resolution by X-ray diffraction [3]. Most GPCRs share a conserved secondary structure that consists of seven transmembrane (TM) helices with extracellular N-terminal loop region.

In this study, we applied our method for functional identification of TM proteins, by using a TM topology pattern of the number of TM segments (tms) and the length of each loop domains [4], to a dataset of eukaryotic 7-tms TM proteins with non-GPCR classes and sub-families classified according to GPCRDB [2] protein family classification.

2 Dataset and Methods

We used only eukaryotic 7-tms TM protein entries with known TM topology extracted from SWISS-PROT (Release 39) [1], of which entries with FRAGMENT are excluded, then these entries were classified according to with GPCRDB (Release 5.2) [2]. The final dataset consists of GPCR class A (772 entries), B (46), C (20), D (0), E (3), Frizzled/Smoothed families (3) as defined by GPCRDB classification, and others (non-GPCR eukaryotic 7-tms TM proteins, e.g. bovine cytochrome *c* oxidase subunit III, 34).

We defined threshold lengths *l* for loop domains in the sequence. If the loop length is longer than the defined threshold length, a topology pattern of “1” is assigned to the loop. Otherwise, the topology pattern is “0”. A wild character “*” is assigned to the loop allowed to accept either “1” or “0” pattern assignment.

3 Results and Discussion

We determined the threshold lengths for all classes defined in GPCRDB and other eukaryotic 7-tms TM proteins in our dataset. Then, we formulated the topology pattern for each of the classes based on the identified threshold length. Using different threshold lengths for the various stages of classification resulted in successful classification of function by our topology pattern (Fig. 1). First, we determined threshold lengths and topology pattern of class A, B, C, E, Frizzled/Smoothed families,

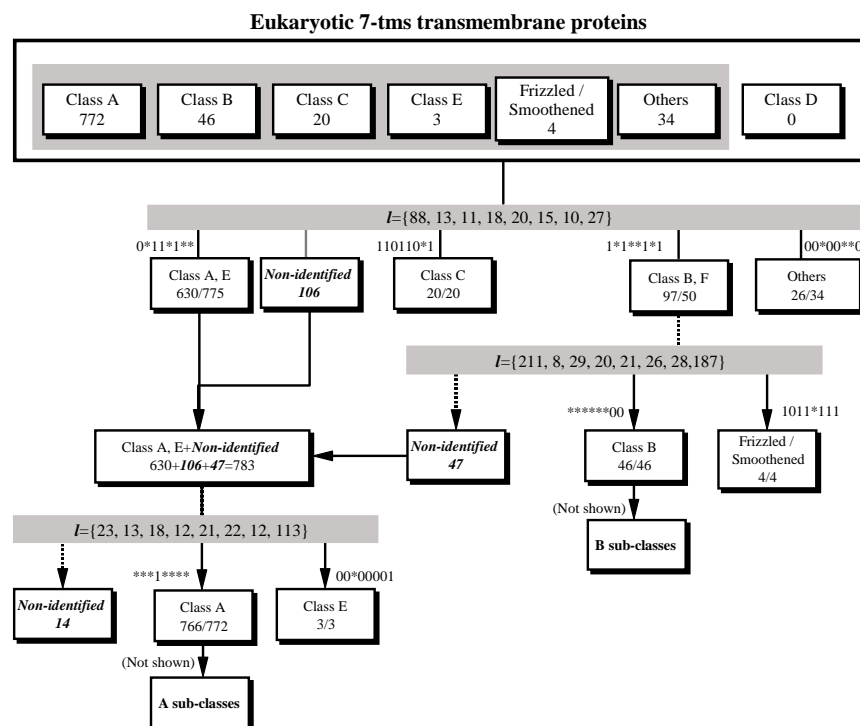


Figure 1: Figure 1: The flowchart of the classification procedure for eukaryotic 7-tms transmembrane proteins by binary transmembrane topology patterns. The number in the numerator corresponds to the detected function under a particular class, while the denominator refers to the reference entries (by GPCRDB classification).

and others (non-GPCR 7-tms TM proteins). Since topology patterns for class A and E are similar, they were grouped together. Although 106 entries did not correspond to any of the topology patterns at this classification stage. Likewise, class B and Frizzled/Smoothened families were similarly grouped together. By using another set of threshold lengths, we were able to classify completely the grouped class of B and Frizzled/Smoothened families into a distinct corresponding class. Moreover, under class B, it is possible to carry out functional classification up to all the sub-classes. However, there were entries that did not correspond to any of the topology patterns (47 entries). For grouped class of A and E, our topology pattern classified almost completely class E with 14 entries which could not be identified completely into distinct classes.

In addition, classes B, C, E and Frizzle/Smoothened families were perfectly classified by our topology pattern, while class A and others were classified also correctly but not 100 percent. Thus, using this method, it is possible to predict the GPCR functional groups comprehensively in the human genome.

References

- [1] Bairoch, A. and Apweiler, R., The SWISS-PROT protein sequence database and its supplement TrEMBL in 1999, *Nucleic Acid Res.*, 28 (1):45–48, 2000.
- [2] Horn, F., Vrend, G., and Cohen, E. F., Collecting and harvesting biological data: the GPCRDB and NucleaRDB information systems, *Nucleic Acid Res.*, 29 (1):346–349, 2001.
- [3] Palczewski, K., Kumasaka, T., Hori, T., Behnke, C.A., Motoshima, H., Fox, B.A., Le Trong, I., Teller, D.C., Okada, T., Stenkamp, R. E., Yamamoto, M., and Miyano, M., Crystal structure of rhodopsin: A G protein-coupled receptor, *Science*, 289 (5480):739–745, 2000.
- [4] Sugiyama, Y. and Shimizu, T., Detection of transmembrane protein function by a binary transmembrane topology pattern, *4th International Conference on Biological Physics*: 60, 2001.