Detection of similar GPCRs by using protein secondary structures

Ja-Hyo Ku*, Young-Woo Yoon**
*Graduate School of Computer Engineering, YeungNam University
**Dept. of Computer Engineering, YeungNam University, Korea

Abstract

G protein-coupled receptor (GPCR) family is a cell membrane protein, and plays an important role in a signaling mechanism which transmits external signals through cell membranes into cells. Now, it is estimated that there may be about 800-1000 GPCRs in a human genome. But, GPCRs each are known to have various complex control mechanisms and very unique signaling mechanisms. GPCRs are involved in maintaining homeostasis of various human systems including an endocrine system or a neural system and thus, disorders in activity control of GPCRs are thought to be the major source of cardiovascular disorders, metabolic disorders, degenerative disorders, carcinogenesis and the like. As more than 60% of currently marketed therapeutic agents target GPCRs, the GPCR field has been actively explored in the pharmaceutical industry. Structural features, and class and subfamily of GPCRs are well known by function, and accordingly, the most fundamental work in studies identifying the previous GPCRs is to classify the GPCRs with given protein sequences. Studies for classifying previously identified GPCRs more easily with mathematical models have been mainly going on. Considering that secondary sequences of proteins, namely, secondary binding structures of amino acids constituting proteins are closely related to functions, the present paper does not place the focus on primary sequences of proteins as previously practiced, but instead, proposes a method to transform primary sequences into secondary structures and compare the secondary structures, and then detect an unknown GPCR assumed to have a same function in databases of previously identified GPCRs.

1. Introduction

As the 21st century begins, large-scale gene-related projects Including the Human Genome Project have been on an increasing trend, and a large quantity of sequence information has been rapidly generated by the help of technical development in high-throughput Sequencing. Even though primary gene sequencing is said to have been completed, it is nothing more than DNA mapping and thus, there are still pending questions which should be solved, such as 'which part is a gene; and 'if it is a gene, what are a functions and a working mechanism of the gene. Thereupon, in order to shorten a study period of various bio-industries including a new drug development, the need of efficiently analyzing a large quantity of sequence information was raised. Under such background, conversion of sequence information into databases has been progressed and methods of more speedily and exactly searching sequence information previously published have been improved and developed.

G protein-coupled receptor (GPCR) family is a cell membrane protein, and plays an important role in a signaling mechanism which transmits external signals through cell membranes into cells. Now, it is estimated that there may be about 800-1000 GPCRs in a human genome. But, GPCRs each are known to have various complex control mechanisms and very unique signaling mechanisms. GPCRs are involved in maintaining homeostasis of various human systems including an endocrine system or a neural system and thus, disorders in activity control of GPCRs are thought to be the major source of cardiovascular disorders, metabolic disorders, degenerative disorders, carcinogenesis and the like. As more than 60% of currently marketed therapeutic agents target GPCRs, the GPCR field has been actively explored in the pharmaceutical industry.

Databases in which all sort of GPCRs are classified in a subfamily unit by function, have been published, GPCRs within the subfamily being similar in function. The most fundamental work in GPCRs studies is to search the databases with a given amino acid sequence. Analytical subjects are 'is the given amino acid sequence a GPCR?; if it is a GPCR, is it the same as a previously known GPCR?, or otherwise, is it a new class of GPCR unknown to date?' and the like. If two amino acid sequences show high similarity, the given amino acid sequence is considered to be a GPCR and can be included to a corresponding subfamily.

Although among tools having been developed up to now for searching sequence information, the BLAST is most extensively used, its searching errors due to continuous sequences or low homology often have been reported. Particularly, for GPCR family, it is difficult to use existing information registered to

** Corresponding Author : Young-Woo Yoon

databases, since there are cases in which GPCRs have a similar function despite their low similarity in amino acid sequences as well as cases in which GPCRs play a totally different role despite their high similarity in amino acid sequences. Considering that functions of proteins are determined by their stereoscopic structures, it is thought that belonging to a same subfamily but being different in amino acid sequences is caused by similar change in stereoscopic structures of different classes of GPCR proteins in the course of evolution.

Proteins are constructed in stages. Amino acid sequences of proteins are termed a protein primary structure, and amino acids in the primary structure are connected together to form two kinds of configuration called α-helixes and β-sheets. α-helixes, β-sheets and amino acid sequences connecting them are termed a protein secondary structure. α-helixes and β-sheets are used as basic components constituting proteins, and produce a stereoscopic form termed a protein tertiary structure. At least two protein tertiary structures are complexed together to form a complex called a protein quaternary structure. At present, algorithms have been developed for predicting protein secondary structures from amino acid sequences, but their accuracy is known to be about 70%. Even though tertiary structures and quaternary structures of a few proteins have been identified by precise analysis and many studies have been focused on methods of predicting tertiary structures, such studies have not yielded satisfactory results so far.

Considering that functions of proteins are determined by their stereoscopic structures, the present paper proposes a method to compare secondary structures of two GPCRs having different amino acid sequences, and then detect an unknown GPCR assumed to have a same function in databases of previously identified GPCRs.

2. Study details and methods

2.1 Molecular biological structures of GPCRs

Most GPCRs consist of a single polypeptide comprising 400-500 residues and contain seven transmembrane-spanning α-helices. Of these α-helices, the third intracellular loop is larger than the other loops and Interacts with G-proteins (Figure 1).

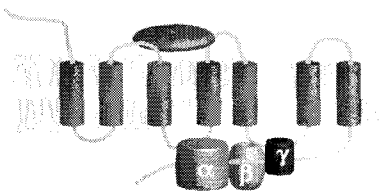


Figure 1. A schematic view of a GPCR structure

A region to which ligands bind is located in one or more α-helix segment region buried in a membrane. The receptor is different from receptors associated with ion channels having ligand binding parts at their extracellular N-terminal which small hydrophobic molecules can easily approach.

2.2 Study methods

Because updated GPCR databases about a number of various organisms have been registered to the GPCRDB operated initiatively by the CMBI, the Netherlands, information about protein sequences of certain GPCRs can be easily acquired from the databases. Furthermore, the databases have provided information about similar GPCRs having a same function in a family format. For GPCRs belonging to a same family, even though there are some cases in which primary sequences of proteins are similar, more numerous GPCRs are confirmed to have a same function despite their substantial difference in primary sequences of proteins. Considering that functions of proteins are more influenced by stereoscopic structures formed by combination of amino acids, namely, secondary structures than by simple sequences of the amino acids, there is a room for study about said finding. In other words, it is necessary to look into similarity in secondary structures for GPCRs showing low similarity in primary sequences among GPCRs belonging to a same family, and examine whether, in spite of difference in simple combination of amino acids, similar secondary stereoscopic structures formed by the amino acids have a same function. In the case of trying to observe secondary structures known to have a same function, observing the secondary structures actually requires substantially much expense in terms of time and cost. Therefore, the present paper intends to compare enormous data easily by using software packages predicting primary sequences of proteins in a form of secondary structures. Comparison of similarity in protein primary structures and secondary structures is performed with the BLAST.

2.2.1 GPCRDB

Now, the GPCRDB which constructs databases on all information of GPCRs is operated on the website. Said DB is an integrated DB comprising the DB at the CMBI, the Netherlands, the ORDB at Yale University, the United States, and the Swiss-Prot DB at Frank Kolackowski. For the Swiss-Prot DB, if GPCRs are aligned with "family" as shown in Figure 2 and a certain ID is selected, and then data are searched in a HSSP format within the GPCRDB-Family, GPCRs belonging to a same family are displayed in a screen and information about their amino acid sequence can be obtained by clicking each GPCR.

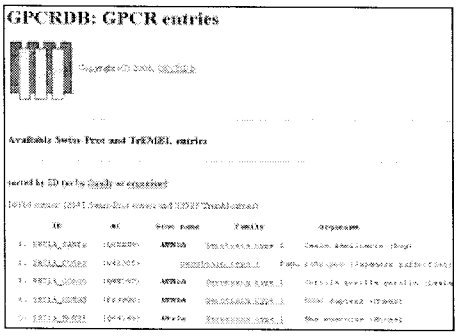


Figure 2. The GPCRDB Web picture

2.2.2 Tools and methods for comparing primary sequences

The BLAST is used for comparing protein primary structures of GPCRs belonging to a same family in the GPCRDB. As groundwork, however, databases on proteins should be constructed with the FORMATDB. As shown in Figure 3, after all amino acid sequences of GPCRs belonging to a same family are saved into one text file, databases are constructed with the FORMATDB. Next, Each amino acid sequence of GPCRs belonging to a same family is saved into a separate text file and compared with the BLASTALL to obtain a resulting file displaying similarity scores, as shown in Figure 4.

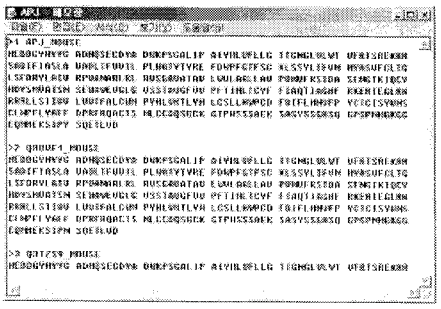


Figure 3. Saving amino acid sequences of GPCR Family into a text file for DB construction

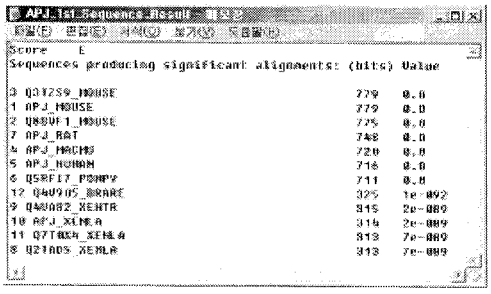


Figure 4. A similarity result of primary sequences

2.2.3 Tools and methods for comparing secondary structures

Secondary structures of GPCRs are transformed by using softwares for secondary structure prediction which transform primary sequences into secondary structures. Various softwares were developed, of which the nnPredict (the UCSF) is utilized in the paper. The nnPredict is a program to predict amino acid sequences in a form of secondary structures, and use two layers of a forward directional neural networking structure. The nnPredict receives an input as an amino acid code (A C D E F G H I K L M N P Q R S T V W Y) and produces an output as a protein structure code (H E -), as shown in Figure 5. On this occasion, 'H' represents a helix structure, 'E' represents a beta strand structure, and '-' represents a turn structure. After such transformation into secondary structures, similarity in secondary structures should be compared. This process also uses the BLAST identically as in comparison of primary sequences. But, as the BLAST can recognize only symbolic codes of nucleic acids or amino acids, secondary structure codes 'H', 'E', and '-' are transformed into 'A', 'T', and 'G', respectively. Data transformed like this are processed with the same method as in comparison of primary structures so as to obtain similarity scores of secondary structures, as shown in Figure 6.

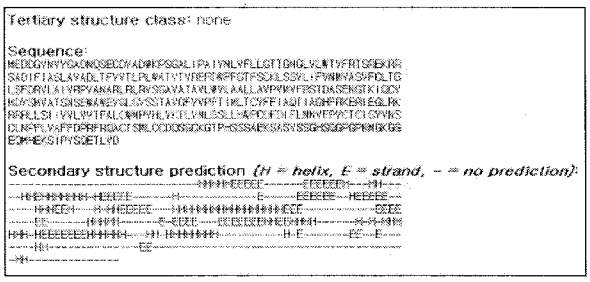


Figure 5. Transformation of primary sequences into secondary structures with the nnPredict.

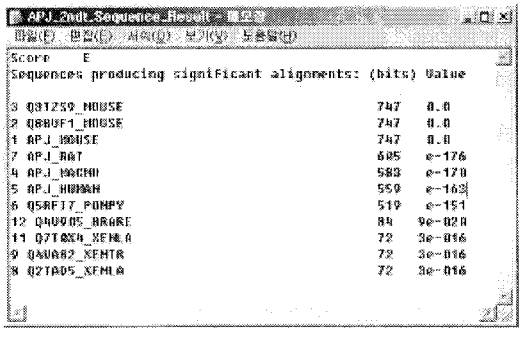


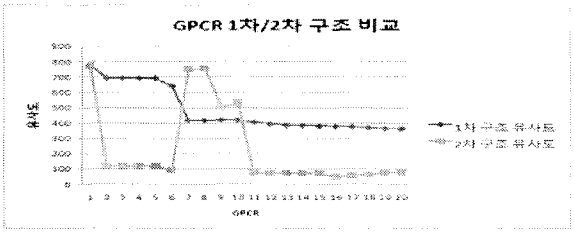
Figure 6. A similarity result of secondary sequences

2.2.4 Comparison of primary structures and secondary structures

Among a same family of GPCRs, there are ones having a similar primary structure, in which prediction

of secondary structures indicates naturally high similarity. However, there are still GPCRs which turned out to have an identical mechanism despite no similarity in primary structures. The purpose of the present paper is to look for GPCR pairs showing high similarity in secondary structures despite no similarity in primary structures. Data obtained in the above 2.2.2 and 2.2.3 are compared to selectively identify and collect GPCRs showing low similarity in primary structures but high similarity in secondary structures. Table 1 illustrates a graph representing similarity in primary/secondary structures between GPCR1 and GPCR2~GPCR20. Data between GPCR1 and GPCR7 and between GPCR1 and GPCR8 in table 1 demonstrate that similarity in primary structures is low, but similarity in secondary structures has a high score.

Table 1. Comparison of GPCR primary/secondary structures.



3. Result

3.1 Experimental data (32 families, and 447 subjects)

Experimental data were obtained from 32 families which were randomly selected and consisted of 447 GPCR subjects.

Table 2. The list of families used in the experiment.

Name of GPCR Family	No. of subjects	Name of GPCR Family	No. of subjects	Name of GPCR Family	No. of Subjects
OX1R	14	Q3MJB1	14	SSR1	11
OXYR	14	Q3S2J4	14	SSR4	5
PAR1	10	Q4VBP0	12	TSHR	14
PAR2	26	Q6NWR3	19	UR2R	6
PE2R1	6	Q53RV4	9	GP174	37
PE2R4	14	Q53T00	6	MASS1	4
PI2R	7	Q541E0	9	P2RY1	40
PRLHR	9	Q99463	5	5HT2A	
PTAFR	9	QRFPR	5	O00325	
Q2M339	14	SMO	13	OPRK	
OPSB	69	O11A1	32		

3.2 Families including subjects showing high similarity in secondary structures

Table 3. The identified GPCR family and the Nos. of subject pairs.

Name of GPCR Family	No. of subject pairs
GP174	27
MASS1	2
P2RY1	7

Of the thirty-two families, three families were found to include subjects showing low similarity in primary structures but high similarity in secondary structures. The abundance ratio of these subjects was 8% on a basis of total families (3/32) and also, 8% on a basis of total subjects (36/447).

4. Conclusion

In the present period in which enormous data about GPCRs have been accumulated, it has become crucial to GPCR studies to classify unknown GPCRs into GPCRs having a similar mechanism, simply by using information about amino acid sequences. Up to now, studies of classifying GPCRs have been principally carried out by using primary sequences of GPCRs. However, there are many GPCRs capable of achieving an identical mechanism even if their primary sequences are different. These GPCRs can not be classified by commonly available methods using primary sequences alone but only by expensive experimental methods.

The present paper proposed a new method which comprises transforming primary amino acid sequences of GPCRs into protein secondary structures, based on the general principle that proteins with identical secondary structures have an identical mechanism, and classifying GPCRs having an identical mechanism, based on similarity in secondary structures. The experimental data demonstrated that the percentage of GPCRs with a similar secondary structure was no more than 8%, but in consideration of no more than 70% of the accuracy of softwares predicting GPCRs in a form of secondary structures, it is thought that there are substantially more GPCRs with a similar secondary structure despite their difference in a primary sequence, which provides sufficient grounds that attention be paid to secondary structures of GPCRs despite the low percentage of GPCRs with a similar secondary structure. In conclusion, in the case of classifying GPCRs, both primary sequences and secondary structures should be considered, and once a secondary structure of an unknown GPCR is experimentally identified, it is possible to effectively classify the unknown GPCR in a short period by searching GPCRs with a similar secondary structure in databases.

5. References

[1] http://www.GPCR.org
[2] http://ncbi.nlm.nih.gov
[3] http://blast.wustl.edu