Chem-Bio Informatics Journal, Vol. 6, No. 1, pp.1-16(2006)

Physicochemical properties of amino acid sequences of G-proteins for understanding GPCR-G-protein coupling

Ganga D. Ghimire[1], Kenichiro Imai[2], Fumitsugu Akazawa[1], Toshiyuki Tsuji[2], Masashi Sonoyama[2] and Shigeki Mitaku[2,*]

1. Tokyo University of Agriculture and Technology, Department of Biotechnology
Nakacho 2-24-16, Koganei, Tokyo 184-8588, JAPAN
2. Nagoya University, School of Engineering, Department of Applied Physics
Furocho, Chikusa-ku, Nagoya, 464-8606, JAPAN
*E-mail: mitaku@nuap.nagoya-u.ac.jp

(Received December 26, 2005; accepted February 15, 2006; published online March 13)

Abstract

A G-protein binds to a G-protein-coupled-receptor (GPCR) by physical interactions between the amino acid sequences. Various physical properties of possible binding regions at the C- and N-termini of the α subunits of G-proteins were analyzed in order to classify the four families of α subunits; α_{s}, $\alpha_{i/o}$, $\alpha_{q/11}$ and $\alpha_{12/13}$, as well as the eight subfamilies. The distribution of the charge density and the hydrophobicity of 100 residues at the C-terminus enabled discrimination of the four families of α subunits, whereas the same physical properties of the 60 residues at the N-terminus enabled discrimination of the subfamilies.

Key Words: GPCR, G-protein, signal transduction, proteomics, bioinformatics

Area of Interest: Bioinformatics and Bio computing

1. Introduction

The coupling between a G-protein-coupled-receptor (GPCR) and a G-protein is one of the most important processes in signal transduction. More than 1000 genes in the human genome code for GPCRs [1]. This provides a very large superfamily of receptor proteins, which function as targets of half of all drugs [2]. GPCRs are located in the cytoplasmic membrane and mediate information from outside a cell to G-proteins located at the inner surface of the membrane. A G-protein is a trimeric protein complex consisting of α-, β- and γ-subunits. Upon binding with a GPCR, a G-protein dissociates into α- and βγ-subunits, one of which interacts with the effecter protein in the cell [3]. There are several families of Gα-proteins (α_s, $\alpha_{i/o}$, $\alpha_{q/11}$ and $\alpha_{12/13}$). These families of G-proteins induce the different kinds of signal transduction, thereby leading to different physiological responses [4]. Therefore, the mechanism of specific binding between a G-protein and a GPCR has a profound importance to the understanding of signal transduction.

The two proteins, GPCRs and G-proteins, must have complementary physical properties in order to ensure specific binding. Therefore, from the viewpoint of physical interactions, two bioinformatic approaches to the G-protein-GPCR coupling are possible: analysis of GPCRs and analysis of G-proteins. If the properties of G-proteins at the binding site are known, it will be much easier to infer the properties of the corresponding GPCR, and vice versa. Most methods to classify the specific binding between receptors and G-proteins focus on the analysis of GPCRs [5][6]. However, the sequence homology between GPCRs is typically low, and common physical features of the amino acid sequences at the binding site are difficult to characterize. On the other hand, all amino acid sequences for G-proteins show relatively high homology, and are suitable for elucidating common physicochemical properties of the binding site. Therefore, the following are addressed in this work: The types of physicochemical properties of G-proteins responsible for specific binding to GPCRs, and how to develop an accurate prediction system for the classification of G-proteins on the basis of physicochemical parameters.

The analysis of G-proteins by a combination of molecular and structural biology revealed that the C- and N-terminal regions of an α-subunit face the cytoplasmic domain of the GPCR [4]. Furthermore, site-directed mutagenesis indicated that the C-terminal segment of the α subunit contributes to the binding specificity with a GPCR. In addition, the N-terminal segment of the α-subunit is in contact with the receptor protein [4]. However, there are several difficulties involved in elucidating the mechanisms of the specific binding between a GPCR and a G-protein. (1) The three dimensional structure of a GPCR-G-protein complex is not available. Thus, the intermolecular interactions at the binding site of the two proteins cannot be analyzed with atomic resolution. (2) The binding segment of GPCR is disordered, the structure is thought to be fixed after binding with the G-protein [7], and the amino acid regions in G-proteins that bind to the GPCRs cannot be defined exactly. Therefore, this work seeks to elucidate the characteristic physical features of the C- and the N-terminal segments of the α subunits, without knowing the exact binding regions.

Exclusive discrimination of the subfamilies from amino acid sequences alone is a problem similar to predicting membrane proteins. In both cases, the exact binding regions in the amino acid sequences are not known. The difference lies in the binding counterpart: membrane proteins bind

with the lipid bilayer, whereas G-proteins bind with GPCRs. In the case of membrane proteins, a typical transmembrane helix consists of not only the central hydrophobic region, but also clusters of amphiphilic residues at the termini of the helices [8], a high performance prediction system for membrane proteins [9][10]. This approach, the "physical finger print" method, in which the distribution of various parameters are simplified to a small number of parameters by coarse graining, was successfully applied recently to the prediction of secondary structure breakers in soluble proteins [11]. When the "physical finger print" approach is applied to the problem of the molecular recognition, it has great advantage that the interactions inferred from the analysis of one protein can be used for the analysis of the other protein by assuming the complementarity of the interactions between two binding proteins.

In this work, we applied this method to the problem of classifying the subfamilies of α-subunits of G-proteins. We analyzed segments consisting of 100 residues and 60 residues at the C- and the N-terminal ends, respectively, in order to classify the eight subfamilies of α-subunits of G-proteins: α_s, α_i, α_o, α_t, α_q, α_{11}, α_{12} and α_{13}. The segments were equally divided into five regions, and the average values of hydrophobicity and the charge densities were calculated for all regions. Twenty numbers of parameters, five average values for hydrophobicity and five for charge density for the C- and the N-segments, were used in the discrimination analysis. The results showed that the physical finger print method was also applicable to the problem of classifying G-proteins.

2. Methods

2.1 Datasets of amino acid sequences of α subunits of G-proteins

The data for the amino acid sequences of the G-proteins were obtained from the gpDB of subfamilies, G_s, G_i, G_o, G_t, G_q, G_{11}, G_{12}, G_{13} [12]. Table 1 lists the IDs of the 99 G-proteins used in this work. The numbers of α subunits in the eight subfamilies were 13 (α_s), 27 (α_i), 24 (α_o), 8 (α_t), 23 (α_q), 10 (α_{11}), 3 (α_{12}) and 3 (α_{13}).

The pair sequence homology among the α subunits ranged from about 20 % to almost 100 %. Table 2 shows the maximum and minimum sequence homologies by the clustering system CLUSTALW for all combinations of subfamilies. The sequence homology of two α-subunits within a subfamily is statistically higher than the homology of sequences from different subfamilies. However, there are sequences with low homology in the same subfamilies, and sequences with high homology in different subfamilies. The range for the pairs in the α_{11} subfamily was between 32.3 and 98.6 %. This indicates that two sequences in the α_{11} subunit may have homology as low as about 30 %. Interestingly, the physiological effect of those sequences is the same. However, the homology for at least one pair from the α_{11} and the α_q subfamilies is as high as 90.9 %. This fact indicates that the activation of very similar G-proteins gives rise to different physiological effects in signal transduction pathways. In this analysis, we did not omit the redundant data because almost all

Table 1. Dataset for analysis.

family	subfamily	ID
α_s	α_s	GPR0114, GPR0115, GPR0116, GPR0117, GPR0118, GPR0119, GPR0121, GPR0122, GPR0125, GPR0126, GPR0127, GPR0351, GPR0405
$\alpha_{i/o}$	α_i	GPR0005, GPR0107, GPR0109, GPR0111, GPR0212, GPR0213, GPR0214, GPR0215, GPR0216, GPR0217, GPR0218, GPR0220, GPR0221, GPR0222, GPR0223, GPR0224, GPR0225, GPR0226, GPR0227, GPR0228, GPR0229, GPR0230, GPR0298, GPR0328, GPR0365, GPR0385, GPR0403
	α_o	GPR0009, GPR0010, GPR0011, GPR0012, GPR0013, GPR0014, GPR0015, GPR0016, GPR0017, GPR0018, GPR0019, GPR0021, GPR0023, GPR0024, GPR0025, GPR0026, GPR0271, GPR0334
	α_t	GPR0244, GPR0245, GPR0246, GPR0247, GPR0248, GPR0249, GPR0250, GPR0251
$\alpha_{q/11}$	α_q	GPR0001, GPR0233, GPR0234, GPR0235, GPR0236, GPR0238, GPR0239, GPR0241, GPR0242, GPR0243, GPR0407, GPR0410, GPR0411, GPR0413, GPR0415, GPR0417, GPR0418,
	α_{11}	GPR0003, GPR0027, GPR0029, GPR0031, GPR0032, GPR0033, GPR0097, GPR0314, GPR0360, GPR0408
$\alpha_{12/13}$	α_{12}	GPR0034, GPR0035, GPR0036
	α_{13}	GPR0037, GPR0038, GPR0380

Table 2. Sequence homology among subfamily. The maximum and the minimum of the scores by ClustalW are shown for all combinations of subfamilies.

	α_s	α_i	α_o	α_t	α_q	α_{11}	α_{12}	α_{13}
α_s	99.7 / 50.3							
α_i	41.5 / 32.0	99.7 / 70.3						
α_o	43.3 / 33.3	71.3 / 44.1	99.7 / 44.1					
α_t	43.6 / 34.1	70.3 / 46.0	60.6 / 44.8	99.7 / 47.3				
α_q	42.8 / 32.3	50.1 / 41.2	49.3 / 41.7	49.3 / 41.5	99.4 / 57.8			
α_{11}	41.8 / 22.7	49.3 / 34.6	48.7 / 33.3	48.4 / 34.3	90.9 / 32.6	98.6 / 32.3		
α_{12}	33.1 / 30.3	41.1 / 38.0	41.2 / 37.3	42.8 / 39.0	43.0 / 39.7	41.8 / 23.7	99.7 / 96.8	
α_{13}	35.0 / 16.7	39.7 / 20.2	41.1 / 12.1	44.8 / 27.7	45.0 / 31.2	44.0 / 25.4	61.8 / 50.9	98.8 / 94.8

data in the same subunit had the redundancy above 30 %. The problem in this work is how to discriminate very similar amino acid sequences in different subfamilies.

2.2 "Physical finger print" method for classification of proteins

In the "physical finger print" method, the average values of indices of amino acids with well-defined physicochemical meanings are calculated for several sequence regions, and a set of average values are used for discrimination analysis. The number and the length of the sequence ranges, as well as the physical properties for analysis vary according to the related phenomena. The specific binding sites in the coupling between G-protein and GPCR are located at the C- and N-terminal regions and the related sequence ranges seem as long as 100 residues according to the three dimensional structure [4][13]. In this work, we used five ranges within the 100 residues of the C-terminus and five ranges within the 60 residues of the N-terminus. The physicochemical properties used for analysis consisted of the hydrophobicity and the electric charges. The number and the length of the sequence ranges, as well as the physical properties were selected after extensive trials of discrimination analyses. Other properties did not show significant difference among subfamilies.

The parameters and the equations for the analysis were previously described in detail. [11] The average hydrophobicity index $\langle H \rangle$ and the average charge density $\langle C \rangle$ were calculated in order to smooth the uneven distribution of parameters using the following equations.

$$\langle H(i) \rangle = \left[\sum_{k=i-3}^{i+3} H(k) \right] / 7 \qquad (1)$$

$$\langle C(i) \rangle = \left[\sum_{k=i-3}^{i+3} C(k) \right] / 7 \qquad (2)$$

Herein, the parameters $H(k)$ and $C(k)$ indicate the Kyte and Doolittle hydropathy index of amino acids at the *k*-th position in the sequence [14] and the electric charge scaled by the elementary charge, respectively.

The average values $\langle X \rangle$, in which *X* is the hydrophobicity *H* or the electric charge *C*, were further averaged for the five ranges of 20 residues in the C-terminus and 12 residues in the N-terminus.

$$\langle \langle X \rangle \rangle_j = \sum_{i \in range(j)} \langle X(i) \rangle / l$$

Double averages are denoted as $\langle \langle X \rangle \rangle_j$, in which *j* indicates the number of ranges from 1 to 5 and *l* is 20 for C-terminus and 12 for N-terminus. The five ranges in the C-terminal segments are denoted as RC1 to RC5, and the five ranges in the N-terminal segments as RN1 to RN5. When two families (subfamilies) are discriminated, the sequences in one family (subfamily) are assumed to be the positive data, and the sequences in the other are assumed to be negative data. The weight deviation DX is then calculated, in which the difference is weighted by the average difference between the positive and negative data as follows [8],

$$DX = \sum_{k=1}^{5} \left(\left\langle \left\langle X \right\rangle \right\rangle_j - \overline{\left\langle \left\langle N \right\rangle \right\rangle_j} \right) \times \left(\overline{\left\langle \left\langle P \right\rangle \right\rangle_j} - \overline{\left\langle \left\langle N \right\rangle \right\rangle_j} \right) \quad (3)$$

Herein, $\overline{\left\langle \left\langle P \right\rangle \right\rangle_j}$ and $\overline{\left\langle \left\langle N \right\rangle \right\rangle_j}$ are the averages of $\left\langle \left\langle X \right\rangle \right\rangle_j$ over all positive and all negative data, respectively. We used only the hydrophobicity and the electric charge, therefore, we only obtained four parameters for analysis: the weighted deviation of the double average of the hydrophobicity DH and that of the electric charges DC for the C- and the N-terminal segments.

The results of the discrimination analysis are described by the following discrimination function,

$$Score = a_0 + a_1 DH + a_2 DC \quad (4)$$

The families were discriminated by analysis of the C-terminal segments, and the subfamilies were discriminated by analysis of the N-terminal segments because a single segment of the C-terminal segment or the N-terminal segment did not provide sufficient information for the complete discrimination.

3. Results

For determining the sequence regions for the analysis, we observed the structural relationship between the domains within an α-subunit, the subunits in a G-protein and a GPCR. Figure 1A shows that an α-subunit is comprised of two domains: a binding domain, which is in contact with a GPCR, and a helical domain, which is connected with the binding domain by narrow loops, and located on the opposite side of the GPCR (PDB ID: 1gp2) [13]. This diagram is shown from the direction parallel to the membrane plane, and the lines in Figure 1A represent the surfaces in contact with the GPCR (B) and the helical domain (C). The binding domain is formed by two segments at the C- and N-terminal ends. The 100 residues at the C-terminal end and the 60 residues at the N-terminal end are shown in blue and green, respectively. In the current work, the entire structure of the binding domain was assumed to be closely related to the binding specificity of G-proteins, and the two regions of amino acid sequences were used for classification of the α subunits into eight subfamilies (Figure 1B).

We first calculated the moving average of the hydrophobicity $\langle H \rangle$ and the electric charges $\langle C \rangle$ with the window of seven residues using Eqs. (1) and (2). The double averages $\left\langle \left\langle H \right\rangle \right\rangle$ and $\left\langle \left\langle C \right\rangle \right\rangle$ were then calculated for five regions consisting of 20 residues at the C-terminal segments of all α-subunits in the families, α_s, $\alpha_{i/o}$, $\alpha_{q/11}$ or $\alpha_{12/13}$. Figure 2 shows the average profiles of the electric charges and the hydrophobicity for the five regions in the C-terminal segment. The shape of the average profiles differs significantly among the families. If the profiles for data within a family

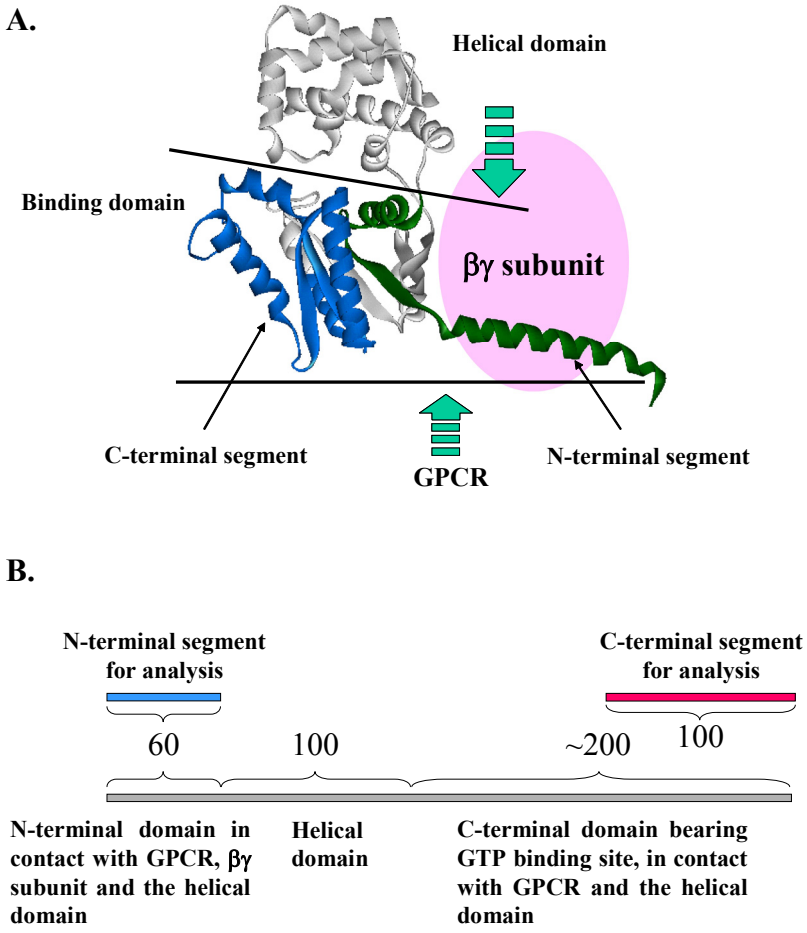

Figure 1. The structure of the α-subunit of a G-protein (PDB ID: 1gp2) [13] viewed along the direction parallel to the membrane plane (A).
The position of the βγ subunits is shown with an ellipse. The binding domain consists of the N-terminal 60 residues and the C-terminal 200 residues, whereas the helical domain consists of the remaining 100 residues. The C-terminal 100 residues and the N-terminal 60 residues were used for the analyses (B).

are very similar to the average profile of the family in Figure 2, it is easy to classify the data based on data from other families by a combination of ten values of the charge density and the average hydrophobicity.

Figures 3A, 3B and 3C show the results of the classification of four families based on the scores for discrimination between α_s versus $\alpha_{q/11}$, $\alpha_{q/11}$ versus $\alpha_{12/13}$ and $\alpha_{i/o}$ versus $\alpha_{q/11}$, respectively. As shown in Figure 3A, the score (α_s versus $\alpha_{q/11}$) can distinguish not only the data within the families α_s and $\alpha_{q/11}$, but also the data within the families $\alpha_{i/o}$ and $\alpha_{12/13}$. All but one of the data set for the

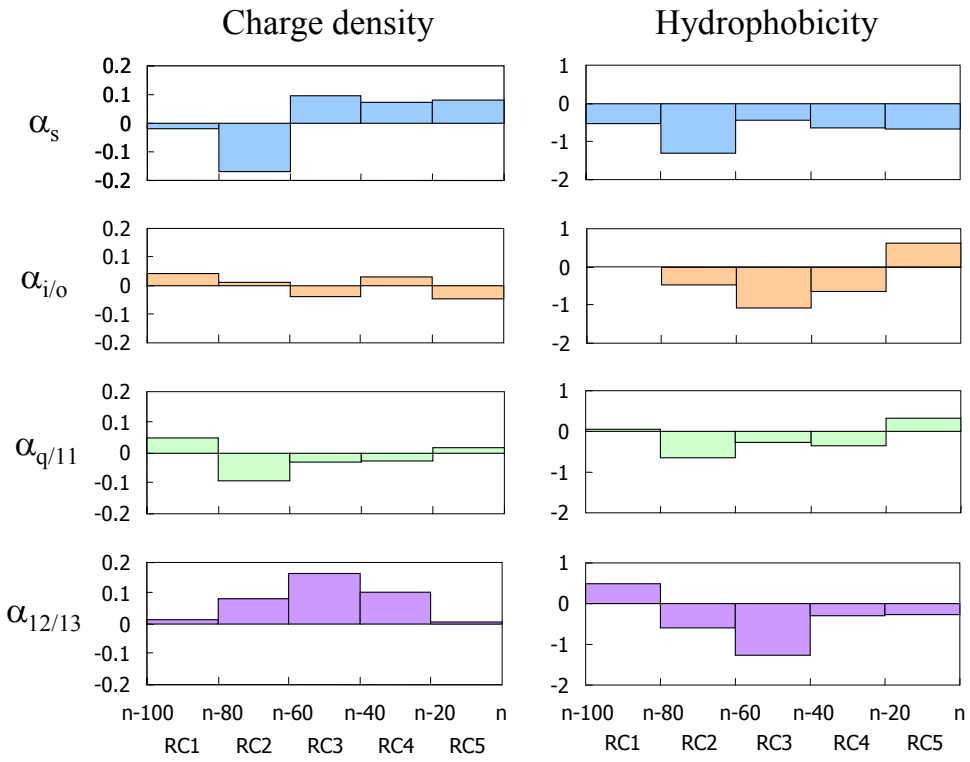

Figure 2. Profiles of the average charge density and the average hydrophobicity for five regions (RC1 – RC5) consisting of 20 residues in the C-terminal segment.

The histograms of the charge density and the hydrophobicity index differ significantly among the families of α-subunits: α_s, $\alpha_{i/o}$, $\alpha_{q/11}$ and $\alpha_{12/13}$. The total number of amino acid of an α-subunit is denoted by *n*.

families $\alpha_{i/o}$ and $\alpha_{12/13}$ were classified to the $\alpha_{q/11}$ family. The discrimination score for α_s versus $\alpha_{q/11}$ is applicable to the data of the different families of $\alpha_{i/o}$ and $\alpha_{12/13}$, corresponding to a kind of cross validation. After subtracting the data for the α_s family, analysis by score for $\alpha_{12/13}$ versus $\alpha_{q/11}$ successfully distinguished the data for these families (Figure 3B). Again, all data within the $\alpha_{i/o}$ family were classified into the $\alpha_{q/11}$ family. This is also a kind of cross validation. Finally, Figure 3C shows the results of the discrimination between the $\alpha_{i/o}$ and $\alpha_{q/11}$ families. The results of the analyses of the C-terminal segments are summarized in Table 3A. The numbers at the diagonal positions correspond to the data correctly discriminated by the three stages of analyses. Among 99 α-subunits, 97 proteins were correctly discriminated with only two errors.

However, the subfamilies could not be discriminated by analysis of the C-terminal segments (data not shown). An N-terminal segment is known to contribute to the binding between a GPCR and a G-protein [4]. Thus, we analyzed the N-terminal segment of α subunits by the same method

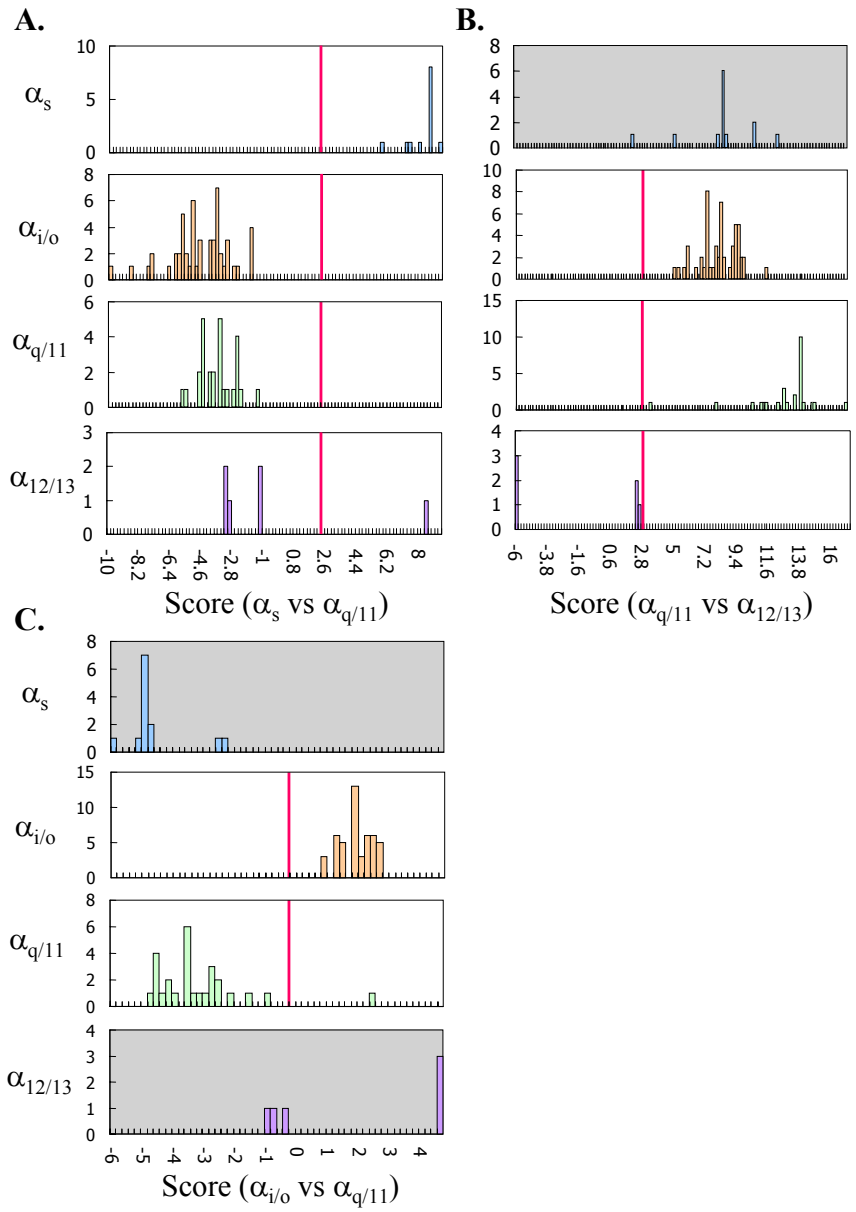

Figure 3. Three sets of histograms in which four groups of subfamilies of α subunits are discriminated by primary component analysis of the C-terminal 100 residues.

The first step (A) is represented by the score for discriminating α_s versus $\alpha_{q/11}$. Almost all data of $\alpha_{i/o}$ and $\alpha_{12/13}$ are also well discriminated against α_s. The discrimination score in the second step (B) is calculated by primary component analysis of $\alpha_{12/13}$ against $\alpha_{q/11}$. All data within $\alpha_{i/o}$ are outside that of $\alpha_{q/11}$. Analysis of the data for $\alpha_{i/o}$ and $\alpha_{q/11}$ is the final step (C), in which all but one subunit are correctly classified.

Table 3. Results of classification of Gα proteins. Classification of four families by analysis of C-terminal segments by the "physical finger print analysis (A).
Analysis of N-terminal segment (B).

A.

	Results of discrimination			
	α_s	$\alpha_{i/o}$	$\alpha_{q/11}$	$\alpha_{12/13}$
α_s	13	0	0	0
$\alpha_{i/o}$	0	53	0	0
$\alpha_{q/11}$	0	1	26	0
$\alpha_{12/13}$	1	0	0	5

B.

	Results of discrimination							
	α_s	α_i	α_o	α_t	α_q	α_{11}	α_{12}	α_{13}
α_s	13							
α_i		27	0	0				
α_o		1	17	0				
α_t		0	0	8				
α_q					17	0		
α_{11}			1		0	9		
α_{12}							3	0
α_{13}	1						0	2

used for C-terminal segment analysis. Figure 4 shows the profiles of the average values of the electric charges and hydrophobicity for five regions consisting of 12 residues in the N-terminal segments. The difference in the profiles among the subfamilies is subtle, but certainly observable. Analysis of the difference in the profiles of the physicochemical parameters, enabled calculation of the scores for the discrimination between the subfamilies: α_o versus α_t, α_i versus α_o, α_q versus α_{11} and α_{12} versus α_{13}. Figures 5A, 5B and 5C show the results of the discrimination between the subfamilies within the groups, $\alpha_{i/o}$, $\alpha_{q/11}$ and $\alpha_{12/13}$, respectively. Figure 5A indicates that the data for

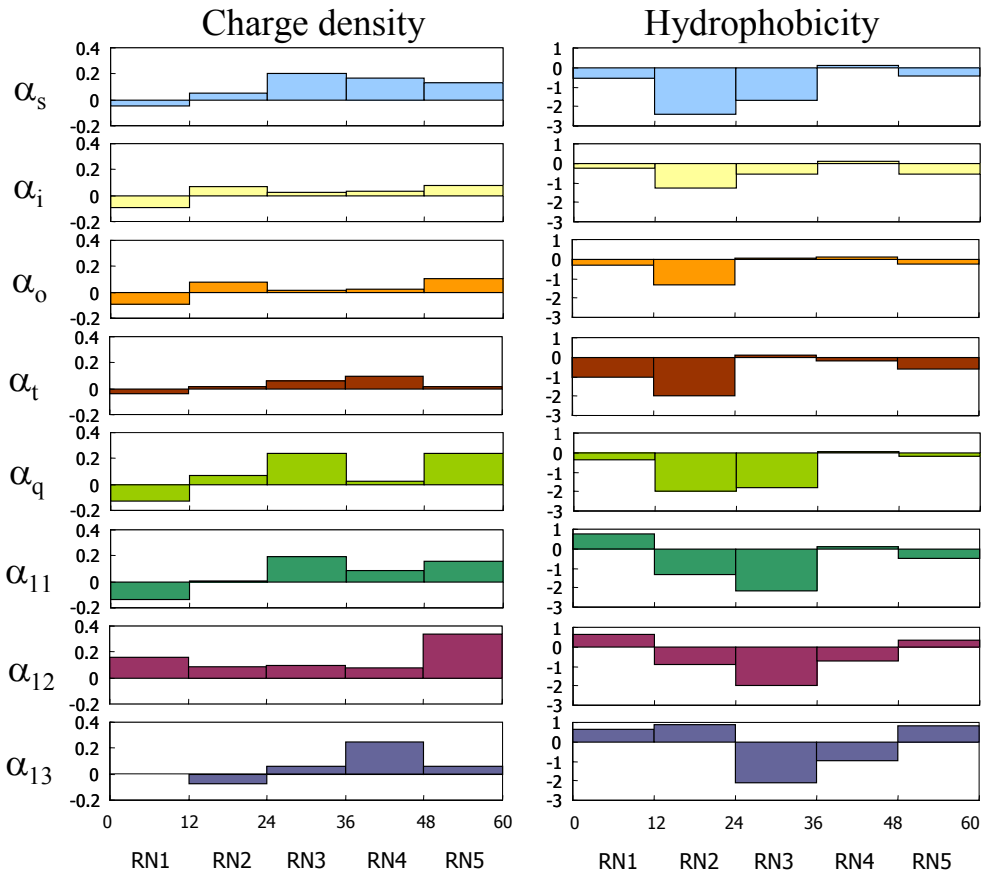

Figure 4. Profiles of the average charge density and the average hydrophobicity for five regions consisting of 12 residues in the N-terminal segment.
The histograms of the charge density and the hydrophobicity index show the difference among the subfamilies of the α subunits: α_s, α_i, α_o, α_t, α_i, α_q, α_{11}, α_{12} and α_{13}.

the subfamily α_o can be distinguished from that of subfamily α_t, and subfamily α_i outside that of α_o. Discrimination between subfamilies α_i and α_o was then carried out. It is important to note that only one datum of the α_o subfamily was incorrectly classified into the subfamily of α_i. The discrimination between α_q and α_{11} was perfect, and the discrimination between α_{12} and α_{13} was also very good. Therefore, combination of the analyses of the C- and N-terminal segments lead to a method for very accurate classification of the α subunits of G-proteins into eight subfamilies, as shown in Table 3B.

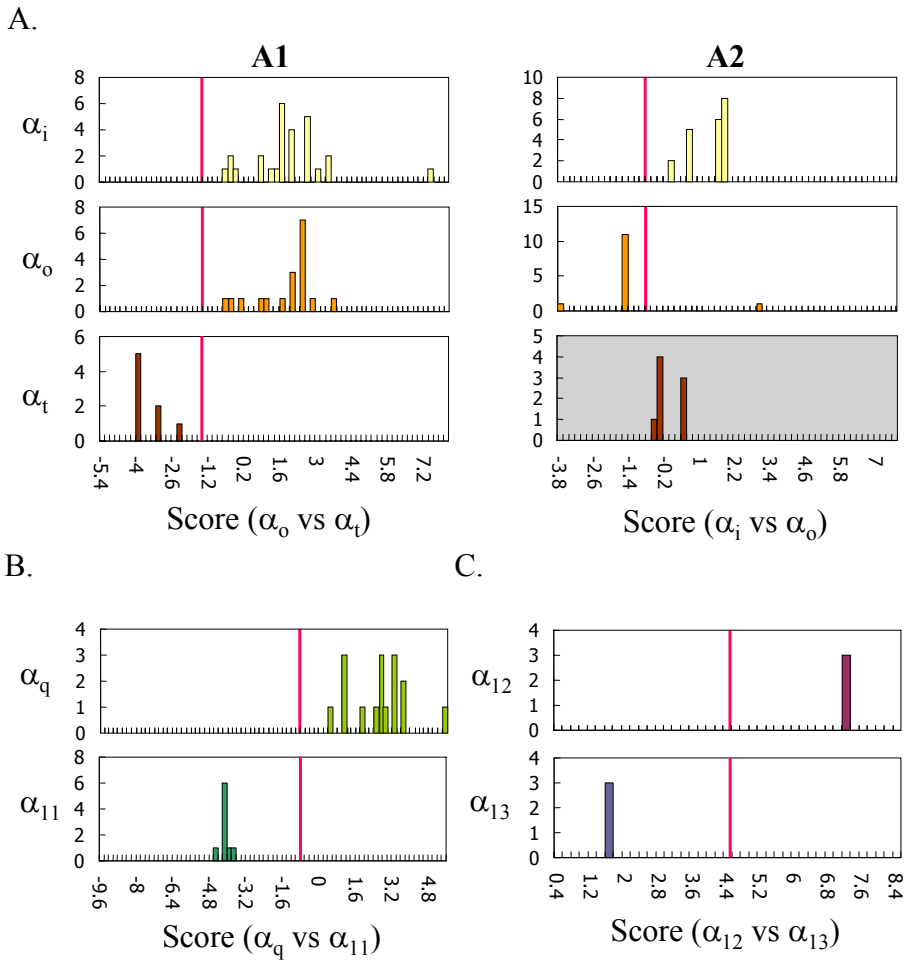

Figure 5. Results of the analyses of the N-terminal segments. The discrimination among subfamilies α_i, α_o and α_t was carried out in two steps (A).
First discrimination between α_o and α_t, and all the data for α_i outside that for α_o (A1). In the second step, the subunits of α_i and α_o were well discriminated (A2). The subfamilies in the families $\alpha_{q/11}$ (B) and $\alpha_{12/13}$ (C) could also be discriminated by analyses of the N-terminal segments.

As the control analysis, the sequence homology of 100 residues at the C-terminus and 60 residues at the N-terminus was calculated. Table 4A shows the homology analysis of the C-terminus. The incorrect classification was observed for three data. Table 4B is the result of the homology analysis of the N-terminus which corresponds to Table 3B. The total number of errors of the classification by the homology analysis was six. This number is small enough and the accuracy of the classification is as good as 95 %. However, the accuracy for the “physical finger print” method was better than

Table 4. Results of classification of Gα proteins. Classification of four families by homology analysis of C-terminal segments (A).
Homology analysis of N-terminal segment (B).

A.

	Results of discrimination			
	α_s	$\alpha_{i/o}$	$\alpha_{q/11}$	$\alpha_{12/13}$
α_s	13	0	0	0
$\alpha_{i/o}$	0	52	1	0
$\alpha_{q/11}$	0	1	26	0
$\alpha_{12/13}$	0	0	1	5

B.

	Results of discrimination							
	α_s	α_i	α_o	α_t	α_q	α_{11}	α_{12}	α_{13}
α_s	12	1						
α_i		26	0	1				
α_o		1	16	0	1			
α_t		0	0	8				
α_q					17	0		
α_{11}		1			0	9		
α_{12}							3	0
α_{13}					1		0	2

the homology analysis. The number of parameters in the homology analysis corresponds to the number of amino acids, 100 for the C-terminus and 60 for the N-terminus, while the number of parameters in the "physical finger print" method in this work is only ten for both analyses in the C- and N-terminal segments. Namely, small number of physicochemical parameters realized better performance than the homology analysis. This fact strongly suggests that the hydrophobicity and the electric charges are the essential properties for the specificity of the binding.

4. Discussion

A heterotrimeric G-protein links the activation of a GPCR and the cell's physiological responses. Therefore, the typing of the G-protein-GPCR coupling according to the information in amino acid sequences is one of the most important problems in the field of the bioinformatics. In this work, we analyzed the amino acid sequences of the α-subunits of various G-proteins, focusing on the C- and N-terminal segments, which are in direct contact with GPCRs.

The meaning of the "physical finger print" approach will be clarified by comparing the present work with the sequence homology analysis. Since the classification of subfamilies of Gα-proteins is in fact performed by the sequence homology analysis of total amino acid sequences, the phylogenetic tree of all 99 α-subunits by the system CLUSTALW gives very good clustering of the subfamilies. Only two data missed the clustering. This level of the accuracy is almost the same as the present work. However, since the specific binding between a G-protein and a GPCR occurs at the interface between the two proteins, the clustering of α-subunits by the two methods should be compared, using only the binding regions. Thus, we analyzed the homology of the partial sequences of 100 residues at the C-terminus and 60 residues at the N-terminus which correspond to the putative binding sites. The results showed that the clustering by the homology analysis of partial sequences was a little worse than the "physical finger print" method. This is physically quite reasonable in that the physical properties of the binding regions determine the specificity of the coupling between the two proteins.

Then, the same approach can be applied to the classification of GPCRs on the basis of the information about the "physical finger print" of the hydrophobicity index and electric charges of Gα-proteins. For example, the charge density at the C-terminus of α_s subunit shows a significant negative cluster as shown in Figure 2, whereas $\alpha_{12/13}$ family has a positive peak of the charge density. Due to the complementarity of the interactions at the binding site, GPCRs which bind with α_s and $\alpha_{12/13}$ subunits are expected to have significant positive and negative peaks of the charge densities, respectively. This type of analysis is impossible in the approach based on the sequence homology. We are now analyzing the amino acid sequences of the cytoplasmic side of GPCRs which will be reported elsewhere.

Furthermore, the present method for the classification of Gα-proteins can be abstracted to more general concept for the molecular recognition. (1) We hypothesized that the physical interactions closely related to the G-protein-GPCR coupling are the well-defined electrostatic and hydrophobic interactions. The complementarity of the interactions in the molecular complexes has been already reported for several different systems: the hydrogen bonding interactions in the double helix of DNA and the hydrophobic interactions together with the steric effect in the supercoil of the leucine zipper. The present work provides the more general expression of the distribution of interacting residues in proteins. (2) Another concept is that the coarse-graining, rather than the detailed distribution of the physicochemical parameters, is substantial for the binding specificity. When the same partial sequences of a Gα-protein are analyzed by the sequence homology and the "physical finger print" method, the accuracy of the classification by the latter approach was better than the former. This fact implicitly suggests that the unit of the G-protein-GPCR coupling is clusters of

amino acid rather than individual amino acids. If the coarse-grained parameters have enough information for the molecular recognition, the "physical finger print" method can be the principle of the analysis of amino acid sequences in general.

The use of the indices of various physical interactions and the coarse graining of amino acid sequences, the "physical finger print" method, have been successfully applied to other types of prediction: membrane proteins [9] and dumbbell-type proteins [15]. The work on membrane proteins primarily relied on hydrophobicity and amphiphilicity indices for amino acids analysis, with accuracy greater than 95 % [8][10]. On the other hand, the work on dumbbell-type proteins mainly relied on electrostatic repulsion for the prediction of the extended structure. Furthermore, all predicted proteins have an extended shape when amino acid sequences from pdb were tested by the physical finger print approach [15]. In the present work, the electric charges and the hydrophobicity index (when these parameters were coarse grained) were sufficient for the classification of α-subunits of G-proteins. In spite of the success in classification, the current work has a weak point; the length of the regions for analysis was determined artificially: 20 residues in the C-terminal segment and 12 residues in the N-terminal segment. The coarse graining approach to the classification of proteins is promising from the three prediction systems of membrane proteins, dumbbell type proteins and subfamilies of G-proteins. However, the determination of the length of the regions appropriate for the accurate classification is a problem to be solved, in applying the "physical finger print" method to other systems.

References

[1] S. Takeda, S. Kadowaki, T. Haga, H. Takaesu, and S. Mitaku, Identification of G protein-coupled receptor genes from the human genome sequence, *FEBS Letters* **520**, 97-101 (2002).

[2] J. Drews, Genomic sciences and the medicine of tomorrow, *Nature biotechnology*, **14**, 1516-1517 (1996).

[3] B. R. Conklin and H. R. Boume, Structural elements of G alpha subunits that interact with G beta gamma, receptors, and effectors, *Cell* **73**, 631-641 (1993).

[4] J. Wess, Molecular basis of receptor/G-protein-coupling selectivity, *Pharmacol. Theor.*, **80**, 231-264 (1998).

[5] N.G. Sgourakis, P. G.. Bagos and S. J. Harmodrakas, Prediction of the coupling specificity of GPCRs to four families of G-proteins using hidden Markov models and artificial neural networks, *Bioinformatics* **21**, 4101-4106 (2005).

[6] Y. Yabuki, T. Muramatsu, T. Hirokawa, H. Mukai and M. Suwa, GRIFFIN: a system for predicting GPCR-G-protein coupling selectivity using a support vector machine and a hidden Markov model, *Nucleic Acids Research*, **33**, W148-W153 (2005).

[7] V.P. Jaakoka, J. Prilusky, J. L. Sussman and A.Goldman, G protein-coupled receptors show

unusual patterns of intrinsic unfolding, *Protein Eng.*, **18,** 103-110 (2005).

[8] S. Mitaku, T. Hirokawa, and T. Tsuji, Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces, *Bioinformatics*, **18,** 608-616 (2002).

[9] T. Hirokawa, S. Boon-Chieng, and S. Mitaku, SOSUI: classification and secondary structure prediction system for membrane proteins, *Bioinformatics*, **14,** 378-379 (1998).

[10] S. Mitaku and T. Hirokawa, Physicochemical factors for discriminating between soluble and membrane proteins: hydrophobicity of helical segments and protein length, *Protein Eng.*, **11,** 953-957 (1999).

[11] K. Imai and S. Mitaku, Mechanisms of secondary structure breakers in soluble proteins, *BIOPHYSICS*, **1**, 55-65 (2005).

[12] A. L. Elefsinioti, P. G. Bagos, I. C. Spyropoulos and S. J. Hamodrakas, A database for G proteins and their interaction with GPCRs, *BMC Bioinformatics*, **5**, 208 (2004).

[13] M. A. Wall, D. E. Coleman, E. Lee, J. A. Iniguez-Lluhi, B. A. Posner, A. G. Gilman and S. R. Sprang, The structure of the G protein heterotrimer Gi alpha 1 beta 1 gamma 2, *Cell*, **83**, 1047-1058 (1995).

[14] J. Kyte and R. F. Doolittle, A simple method for displaying the hydropathic character of a protein, *J Mol Biol*., **157,** 105-132 (1982).

[15] N. Uchikoga, S. Takahashi, R. Ke, M. Sonoyama and S. Mitaku, Electric charge balance mechanism of extended soluble proteins, *Protein Sci.*, **14,** 74-80 (2005).