Available online at www.sciencedirect.com
ScienceDirect
Computational Biology and Chemistry 31 (2007) 246–256
Computational Biology and Chemistry
www.elsevier.com/locate/compbiolchem

A combined approach for the classification of G protein-coupled receptors and its application to detect GPCR splice variants

Hae-Seok Eo [a], Jae Pil Choi [b], Seung-Jae Noh [b], Cheol-Goo Hur [b,*,1], Won Kim [a,**,1]

[a] *School of Biological Sciences, Seoul National University, Seoul 151-742, Republic of Korea*
[b] *Korea Research Institute of Bioscience and Biotechnology (KRIBB), 52, Oun-dong, Yuseong, Daejeon 305-333, Republic of Korea*

Received 8 January 2007; accepted 7 May 2007

Abstract

G protein-coupled receptors (GPCRs) constitute the largest family of cell surface receptors and play a central role in cellular signaling pathways. The importance of GPCRs has led to their becoming the targets of more than 50% of prescription drugs. However, drug compounds that do not differentiate between receptor subtypes can have considerable side effects and efficacy problems. An accurate classification of GPCRs can solve the side effect problems and raise the efficacy of drugs. Here, we introduce an approach that combines a fingerprint method, statistical profiles and physicochemical properties of transmembrane (TM) domains for a highly accurate classification of the receptors. The approach allows both the recognition and classification for GPCRs at the subfamily and subtype level, and allows the identification of splice variants. We found that the approach demonstrates an overall accuracy of 97.88% for subfamily classification, and 94.57% for subtype classification.

Keywords: G protein-coupled receptor; Transmembrane domain; Splice variant; Profile hidden Markov model; Physicochemical property

1. Introduction

G protein-coupled receptors (GPCRs) constitute the largest family of cell surface receptors, and play a central role in a cellular signaling network that regulates many basic physiological processes, such as secretion, neurotransmission, growth, cellular differentiation and the immune response (Baldwin, 1994; Lefkowitz, 2000). On account of a large diversity within the family and the diverse roles in the cellular signaling pathways, the receptors have been heralded as therapeutic drug targets (Drews, 1996; Marinissen and Gutkind, 2001).

GPCRs share a common functional unit in the form of seven transmembrane (TM) domains that are connected by alternating intracellular and extracellular loops (Baldwin, 1993; Gether, 2000; Bjarnadottir et al., 2006). These receptors can be grouped into subfamilies according to their protein sequence homology, the ligand structure, and receptor function (Horn et al., 2003) (Fig. 1). Somehow some relationship between the sequences of GPCRs and ligands of particular subfamilies seems to exist, but there is no clear correlation between sequence similarity and ligand-specificity (Papasaikas et al., 2003; Inoue et al., 2004). These evolutionary complexes of GPCRs in their families have been the main obstacle to developing effective methods for GPCR classification (Kim et al., 2000). In this report, we describe a new combined approach for the identification and classification of GPCRs from a large-scale genomic database.

Several research groups have developed methods for the recognition and classification of the receptors (Elrod and Chou, 2002; Karchin et al., 2002; Papasaikas et al., 2003; Huang et al., 2004; Bhasin and Raghava, 2005). Although their methods are useful, there are three problems in applying these methods to a large protein family. First, most research groups did not succeed in the classification of different subtypes of receptors belonging to one subfamily and only concentrated on the subfamily level. Second and more importantly, the investigators have failed to include methods for detecting splice variants of the receptors in their classification strategies. A highly accurate subdivision of subfamily and detection of splice variants are important for developing drugs and solving the problem of side effects (Bhasin and Raghava, 2005). Lastly, several methods are too computationally expensive to apply directly to whole GPCR families.

* Corresponding author. Tel.: +82 42 879 8560; fax: +82 42 879 8569.
** Corresponding author. Tel.: +82 2 887 0752; fax: +82 2 872 1993.
E-mail addresses: hurlee@kribb.re.kr (C.-G. Hur), wonkim@plaza.snu.ac.kr (W. Kim).
[1] Authors contributed equally to this work.

1476-9271/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2007.05.002

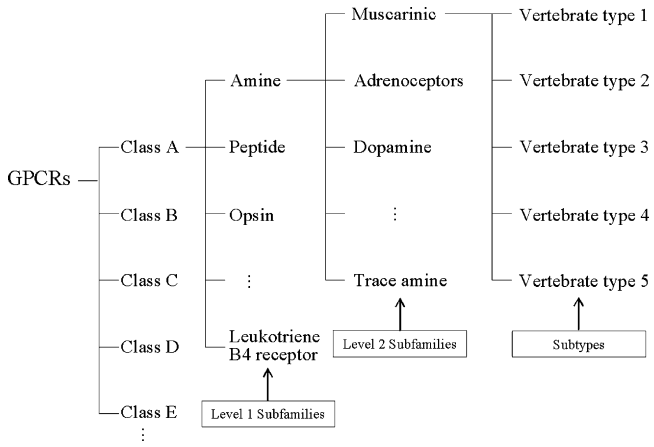

Fig. 1. Schematic view of the GPCR family tree. These classifications were taken from the GPCRDB information system (Horn et al., 2003).

Here, we describe a novel approach that overcomes the above problems and its application for identifying novel splice variants in the tissue-specific genomic database, TISA (Noh et al., 2006). For coping with the evolutionary complexes of the receptors and to enhance the diagnostic power of the approach, the fingerprint method, which uses TM domains for building diagnostic signatures, was adopted. Recently, Gaulton and Attwood (2003) showed a diagnostic sensitivity of the fingerprint method in protein classification, and Sadka and Linial (2005) introduced a forte of TM domains in characterizing membranous proteins. With the fingerprints, profile hidden Markov models (HMMs), which excel at recognizing the weak similarity between members, and also the physicochemical properties of amino acids are combined to enhance the descriptive power of the approach.

We constructed a transmembrane-hidden Markov model-Library (TM-HMM-Library) for identifying and classifying the receptors into subfamilies and adopted Grantham's physicochemical distances of amino acids for subdividing the GPCRs into subtypes (Grantham, 1974; Graur and Li, 2000). With the combined approach, we show its performance with a test set, and then illustrate its utility in the identification of novel splice variants.

2. Materials and methods

2.1. Dataset

The GPCR dataset used for training and evaluating the method was extracted from SwissProt (Release 50.0, www.expasy.org/cgi-bin/lists?7tmrlist.txt; Boeckmann et al., 2003) based on the following two conditions: description of "Mammalia" in the OC line and no description of "Fragment" in the DE line. Finally, 1021 protein sequences were obtained and classified based on GPCRDB (Release 9.0; Horn et al., 2003), as shown in Table 1 . The dataset was arranged into six classes and subdivided into 89 subfamilies. We also extracted 1021 decoy negative protein sequences, which have "TRANSMEM" in the FT lines but do not belong to GPCR family, from SwissProt and added them to the dataset for evaluating the method.

2.2. GPCR subfamily classification (construction of the TM-HMM-Library)

The subfamily classification scheme in GPCRDB is defined by the chemical difference of ligands rather than the receptor sequence homology. Many *Level 2 subfamilies* belonging to the same *Level 1 subfamily* show highly low sequence similarity that may be the result of convergent evolution (Karchin et al., 2002). For these reasons, some *Level 1 subfamilies* are subdivided into subtypes, which are specific in terms of function, via *Level 2 subfamilies* (e.g. *Level 1 subfamily* (amine) > *Level 2 subfamily* (muscarinic acetylcholine) > subtype (muscarinic acetylcholine vertebrate type 1)) and other *Level 1 subfamilies* are subdivided directly into subtypes (e.g. *Level 1 subfamily* (latrophilin) > subtype (latrophilin type 1)) (Fig. 1). Reflecting on these characteristics of the receptor subfamilies, we decided to partition the 1021 sequences into subsets of 89 subfamilies according to *Level 2 subfamilies*, and *Level 1 subfamilies*, which do not have *Level 2 subfamilies*, for constructing the classification method.

The main goal of this step was to develop the method for recognizing and classifying GPCRs into subfamilies. For constructing the TM-HMM-Library, a multiple sequence alignment was constructed for each subset of GPCR subfamilies, using ClustalW 1.83 (Thompson et al., 1994) and seven TM domains were extracted from each subset. Then 623 (=89 subfamilies × 7 TMs) profile HMM were built with the HMMER ver.2.3.2 package by using the program *hmmbuild* and option *–A*, and the TM-HMM-Library was calibrated using the program *hmmcalibrate* (Eddy, 1998) (Fig. 2). The performance of the TM-HMM-Library as a subfamily level classifier was estimated after an *hmmpfam*-search for TM domains in a query sequence against an HMM Library (Eddy, 1998). The *hmmpfam* program of the HMMER software package reads a sequence file and compares each sequence in it, one at a time, against all of the HMMs in the TM-HMM-Library to look for significantly similar sequence matches. From the output report for each query sequence, which reports the best scoring domains in order of their occurrence in the sequence, the query sequence can be classified into the correct subfamily.

2.3. Performance evaluation (n-fold cross-validation)

In order to evaluate the performance of the approach, we used an *n*-fold cross-validation. This experiment, which requires *n* members per subfamily, is problematic for a subfamily experiment, due to the large number of subfamilies, many of which are sparsely populated (Karchin et al., 2002). In this study, we decided on a 5-fold cross-validation for the subfamily classification experiment, in which the dataset of all GPCRs sequences is divided into five subsets of approximately equal size. The TM-HMM-Library has trained the net five times, each time leaving out one of the subsets from the training. The remaining subset for each time is used to estimate the performance of the trained TM-HMM-Library. We also added the same size negative sequences, which are non-GPCR transmembrane proteins, to each subset for estimating specificity. For identifying the optimal threshold,

Table 1
Details of the dataset used for training and evaluating the method

Class/subfamily	Species (mammal)				Total
	Human	Mouse	Rat	Others[a]	
Class A: Rhodopsin-like	**235**	**210**	**163**	**194**	**802**
Muscarinic acetylcholine	5	4	5	6	20
Alpha adrenoceptors	6	6	6	8	26
Beta adrenoceptors	3	3	3	6	15
Dopamine	5	5	5	5	20
Histamine	4	3	3	5	15
Serotonin	12	12	12	12	48
Trace amine	5	10	13	2	30
Angiotensin	2	3	3	2	10
Bombesin	3	3	2	3	11
Bradykinin	2	2	2	4	10
C5a anaphylatoxin	3	2	2	3	10
Fmet-leu-phe	3	2	0	2	7
APJ-like	1	1	1	1	4
Interleukin-8	2	1	2	5	10
C-C chemokine	12	12	4	8	36
C-X-C chemokine	4	3	2	6	15
C-X3-C chemokine	2	2	1	0	5
Cholecystokinin CCK	2	2	2	4	10
Endothelin	2	2	2	4	10
Melanocortin	5	5	3	10	23
Duffy antigen	1	1	0	3	5
Neuropeptide Y	5	6	2	9	22
Neurotensin	2	2	2	0	6
Opioid	4	4	4	5	17
Somatostatin	5	5	5	2	17
Tachykinin	3	3	3	5	14
Vasopressin/vasotocin	3	3	3	4	13
Oxytocin/mesotocin	1	1	1	2	5
Galanin	3	3	3	0	9
Thrombin	1	1	1	2	5
Proteinase-activated	3	3	3	0	9
Orexin	2	1	2	1	6
Neuropeptide FF	2	1	2	0	5
Urotensin II	1	1	1	1	4
GPR37/endothelin B-like	2	2	1	0	5
Chemokine receptor-like	4	3	2	0	9
Neuromedin U-like	2	2	2	0	6
Somatostatin- and angiogenin-like peptide	2	2	0	0	4
Melanin-concentrating hormone receptors	2	1	0	2	5
Prokineticin receptors	2	2	2	2	8
Follicle stimulating hormone	1	1	1	2	5
Lutropin-choriogonadotropic hormone	1	1	1	2	5
Thyrotropin	1	1	1	2	5
Rhodopsin vertebrate	4	3	3	7	17
Rhodopsin Other	4	4	0	1	9
Prostaglandin	6	6	5	7	24
Prostacyclin	1	1	1	1	4
Thromboxane	1	1	1	2	5
Adenosine	4	4	4	6	18
Purinoceptors	16	11	6	3	36
Cannabinoid	2	2	2	1	7
Platelet activating factor	1	1	1	1	4
Gonadotropin-releasing hormone	2	1	1	5	9
Thyrotropin-releasing hormone	1	1	1	2	5
Growth hormone secretagogue	1	1	1	1	4
Melatonin	3	3	0	3	9
Lysosphingolipid and LPA (EDG)	7	7	4	3	21
Leukotriene B4 receptor	2	2	2	0	6
SREB	3	3	3	0	9
Mas proto-oncogene and Mas-related (MRGs)	10	12	6	0	28
RDC1	1	1	1	1	4

Table 1 (*Continued*)

Class/subfamily	Species (mammal)				Total
	Human	Mouse	Rat	Others[a]	
LGR-like (hormone receptors)	5	2	1	0	8
GPR	17	4	3	8	32
GPR45-like	2	2	0	0	4
Cysteinyl leukotriene	2	2	2	2	8
Free fatty acid receptor (GP40, GP41, GP43)	4	3	0	0	7
Class B: Secretin-like	**29**	**23**	**18**	**13**	**83**
Calcitonin	2	2	2	2	8
Corticotropin releasing factor	2	2	2	1	7
Glucagon	3	2	3	0	8
Growth hormone-releasing hormone	1	1	1	1	4
Parathyroid hormone	2	2	2	2	8
PACAP	1	1	1	1	4
Secretin	1	1	1	1	4
Vasoactive intestinal polypeptide	2	2	2	1	7
EMR1	5	3	0	1	9
Latrophilin	4	1	3	3	11
Brain-specific angiogenesis inhibitor (BAI)	3	3	0	0	6
Cadherin EGF LAG (CELSR)	3	3	1	0	7
Class C: Metabotropic glutamate/pheromone	**18**	**12**	**14**	**3**	**47**
Metabotropic glutamate group I	2	1	2	0	5
Metabotropic glutamate group II	2	1	2	0	5
Metabotropic glutamate group III	4	1	4	0	9
Extracellular calcium-sensing	1	1	1	1	4
GABA-B	2	1	2	0	5
Orphan GPRC5	4	4	0	0	8
Taste receptors (T1R)	3	3	3	2	11
Frizzled/smoothened family	**10**	**10**	**3**	**0**	**23**
Frizzled group A (Fz 1, 2, 4, 5, 7–9)	8	8	3	0	19
Frizzled group B (Fz 3 and 6)	2	2	0	0	4
Vomeronasal receptors (V1R and V3R)	5	0	0	4	9
Taste receptors T2R	25	10	6	16	57
Total	322	265	204	230	1021

The classification is based on GPCRDB (Release 9.0).

[a] Others include *Sus scrofa* (pig, 35 sequences), *Bos Taurus* (bovine, 51), *Macaca mulatta* (rhesus macaque, 22), *Cavia porcellus* (guinea pig, 20), etc.

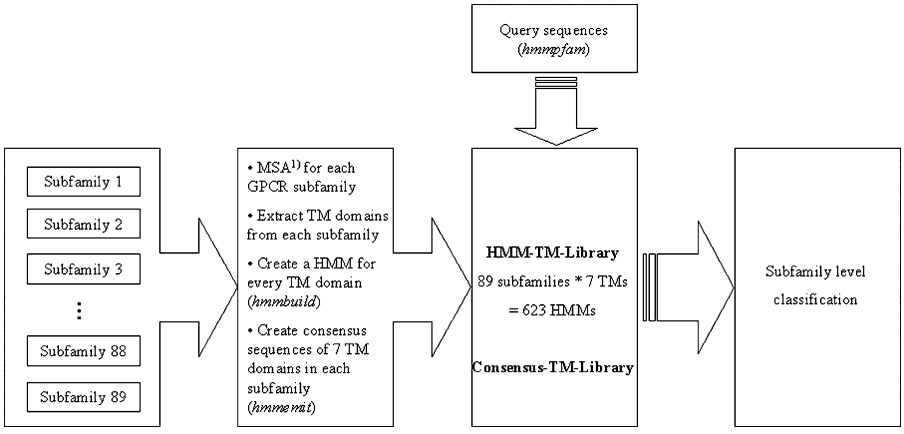

Fig. 2. Schematic diagram of the GPCR subfamily classification algorithm based on the information of the TM domains. The *hmmpfam* in the HMMER package reads a sequence file and compares each sequence in it against all of the HMMs in the TM-HMM-Library. For generating a single majority-rule consensus in each subfamily, which used for the classification of GPCRs into its subtypes, the program *hmmemit*, and option *–C* was applied and Consensus-TM-Library was constructed. (1) Multiple sequence alignment.

Table 2
Physicochemical distances for each amino acid pair

Arg **R**	Leu **L**	Pro **P**	Thr **T**	Ala **A**	Val **V**	Gly **G**	Ile **I**	Phe **F**	Tyr **Y**	Cys **C**	His **H**	Gln **Q**	Asn **N**	Lys **K**	Asp **D**	Glu **E**	Met **M**	Trp **W**		
110	145	74	58	99	124	56	142	155	144	112	89	68	46	121	65	80	135	177	Ser	**S**
	102	103	71	112	96	125	97	97	77	180	29	43	86	26	96	54	91	101	Arg	**R**
		98	92	96	32	138	5	22	36	198	99	113	153	107	172	138	15	61	Leu	**L**
			38	27	68	42	95	114	110	169	77	76	91	103	108	93	87	147	Pro	**P**
				58	69	59	89	103	92	149	47	42	65	78	85	63	81	128	Thr	**T**
					64	60	94	113	112	195	86	91	111	106	126	107	84	148	Ala	**A**
						109	29	50	55	192	84	96	133	97	152	121	21	88	Val	**V**
							135	153	147	159	98	87	80	127	94	98	127	184	Gly	**G**
								21	33	198	94	109	149	102	168	134	10	61	Ile	**I**
									22	205	100	116	158	102	177	140	28	40	Phe	**F**
										194	83	99	143	85	160	122	36	37	Tyr	**Y**
											174	154	139	202	154	170	196	215	Cys	**C**
												24	68	32	81	40	87	115	His	**H**
													46	53	61	29	101	130	Gln	**Q**
														94	23	42	142	174	Asn	**N**
															101	56	95	110	Lys	**K**
																45	160	181	Asp	**D**
																	126	152	Glu	**E**
																		67	Met	**M**

Mean chemical distance is 100.

the performance of the method was estimated using different values of a threshold parameter (0.0001, 0.001, 0.01, and 0.1) and the performance quality was evaluated by the accuracy and Matthew's correlation coefficient (MCC) (Matthews, 1975) (Table 4). The performance qualities of each subfamily in the optimal threshold were evaluated by the determination of the sensitivity, specificity, and accuracy.

2.4. GPCR subtype classification

Only highly accurate identification of receptor subtypes can solve the problem of efficacy and side effects of various drugs (Bhasin and Raghava, 2005). The replacements of amino acids by divergent evolution or the distinct usages of amino acids by convergent evolution, which were separately adopted in each subtype, produced subtle or lots of sequence differences in many regions among them. For coping with the complex evolutionary background of each subtype, the second step aimed at subdividing GPCR subfamily adopts Grantham's physicochemical distances between amino acids, which based on the properties of amino acids such as polarity, molecular volume, and chemical composition (Grantham, 1974; Graur and Li, 2000) (Table 2). A replacement of an amino acid by a similar one, which is called a conservative replacement, is indicated by small distance, and a replacement of an amino acid by a dissimilar one is called a radical replacement and indicated by large distance (Grantham, 1974; Graur and Li, 2000). If there is a pair of amino acid sequences, a numerical vector can be produced by Grantham's physicochemical distances between them.

T	Y	I	T	L	E	L	V	I	A	...
58	0	0	59	5	0	32	32	0	0	...
A	Y	I	G	I	E	V	L	I	A	...

From the consensus TM sequences in each subfamily, which were constructed using the HMMER ver.2.3.2 package, the program *hmmemit*, and option *–C* (Eddy, 1998) (Fig. 2), the numerical vectors of each of the subtypes and query sequence can be obtained and the strength of the linear relationships between each subtype and query sequence can be calculated by the Pearson correlation coefficient, R (Table 3). R^2 close to 1 indicates a strong linear relationship; values close to 0 a weak one. The query sequence in Table 3 shows a strong relationship with *Adenosine subtype 1* (*AA1R*), indicating the reasonableness of locating the query sequence in the *Adenosine subtype 1* (*AA1R*) (Figs. 3 and 4).

$$R = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

where X_i is the physicochemical distance between ith residue of query sequence and ith residue of consensus sequence and Y_i is the physicochemical distance between ith residue of subtype sequence and ith residue of consensus sequence.

3. Results and discussion

3.1. TM-HMM-Library thresholds selection and GPCR subfamily classification

The procedure for constructing the TM-HMM-Library and its application for classifying GPCRs into the subfamily level is illustrated in Fig. 2. The performance of the method was evaluated through 5-fold cross-validation and was estimated using four different E-value thresholds. The results are summarized in Table 4 which give the sensitivity, specificity, accuracy and MCC of classification for each different E-value threshold. The accuracy and MCC of the method reached 97.88% and 0.96,

Table 3
The numerical vectors of each subtype and query TM sequence

	Consensus sequence (adenosine subfamily TM1): TYITLELVIALLAVVGNVLV	R^2 (R)
Query TM1 (adenosine subfamily): AYIGIEVLIALVSVPGNVLV	58,0,0,59,5,0,32,32,0,0,0,32,99,0,68,0,0,0,0,0	–
Adenosine subtype 1 (AA1R) TM1: AYIGIEVLIALVSVPGNVLV	58,0,0,59,5,0,32,32,0,0,0,32,98,0,68,0,0,0,0,0	1[a] (1)
Adenosine subtype 2 (AA2AR) TM1: VYITVELAIAVLAILGNVLV	69,0,0,0,32,0,0,64,0,0,32,0,0,29,32,0,0,0,0,0	0.07303876[b] (0.2702568)
Adenosine subtype 2 (AA2BR) TM1: LYVALELVIAALAVAGNVLV	92,0,29,58,0,0,0,0,0,0,96,0,0,0,64,0,0,0,0,0	0.1263553[c] (0.3554649)
Adenosine subtype 3 (AA3R) TM1: TYITMEAAIGLCAVVGNMLV	0,0,0,0,15,0,96,64,0,60,0,198,0,0,0,0,0,21,0,0	0.00988843[d] (0.09944062)

[a] Linear relationship between query and adenosine subtype 1 (AA1R).
[b] Linear relationship between query and adenosine subtype 2 (AA2AR).
[c] Linear relationship between query and adenosine subtype 2 (AA2BR).
[d] Linear relationship between query and adenosine subtype 3 (AA3R).

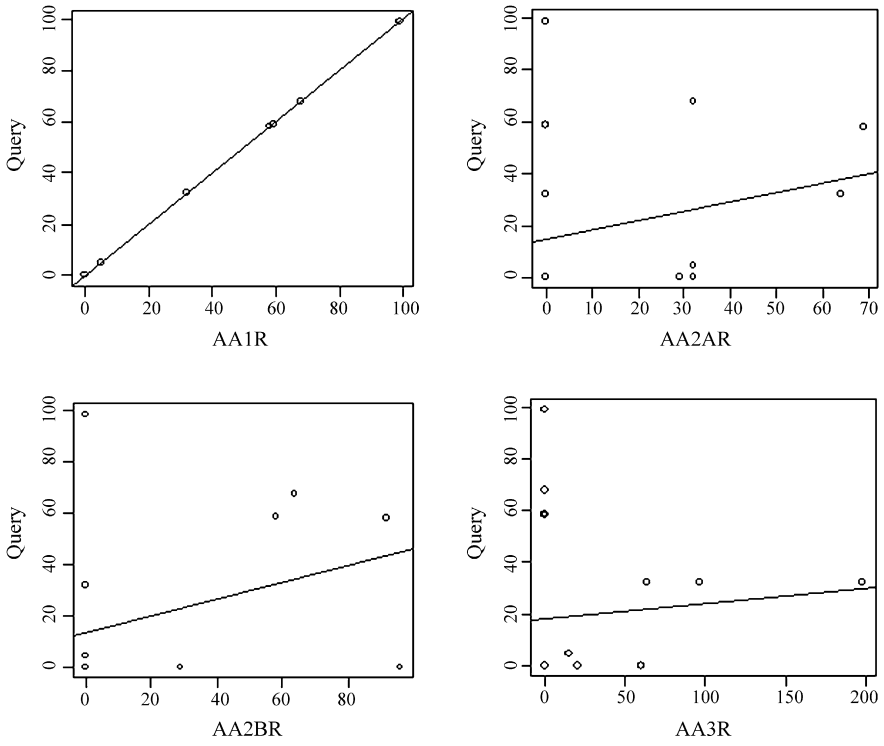

Fig. 3. Linear relationships between each subtype and query TM 1 sequence. In the upper left scatter plot, the points fall on a straight line; a strong relationship between query and subtype, AA1R, is seen.

respectively, at a threshold value of 0.01. The results show the high discriminative capacity of the method in distinguishing GPCRs from other non-GPCR transmembrane proteins and in the classification of GPCRs into subfamilies. We selected an *E*-value threshold equal to 0.01 for subfamily classification and subsequence analysis.

The detailed results of the 5-fold cross-validation experiments at a threshold value of 0.01 are summarized in Table 5 which give the sensitivity, specificity and Acc for each subfamily. A fair number of subfamilies were classified perfectly

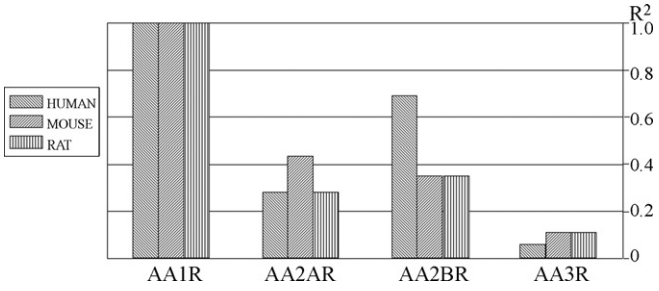

Fig. 4. Histogram of Pearson correlation coefficient, R^2, between each subtype and query. The histogram shows a high correlation relationship between query and AA1R subtype.

Table 4
The performance of the TM-HMM-Library in recognizing and classifying the GPCRs at different thresholds

Threshold[a]	Sensitivity	Specificity	Acc	MCC
0.0001	93.00	100	96.50	0.93
0.001	95.25	100	97.63	0.95
0.01	96.75	99.00	97.88	0.96
0.1	96.75	95.50	96.13	0.92

Acc, accuracy; MCC, Matthew's correlation coefficient.
[a] Hits with *E*-values better than the threshold are detected.

Table 5
The performance of the combined approach in classifying the GPCRs at the subfamily and subtype levels

Subfamily	Subfamily level			Subtype level
	Sen[a]	Spe[b]	Acc[c]	Sen
Muscarinic acetylcholine	1.00	1.00	1.00	1.00
Alpha adrenoceptors	1.00	1.00	1.00	1.00
Beta adrenoceptors	1.00	1.00	1.00	1.00
Dopamine	1.00	1.00	1.00	0.83
Histamine	1.00	1.00	1.00	1.00
Serotonin	1.00	0.86	0.93	1.00
Trace amine	1.00	1.00	1.00	1.00
Angiotensin	1.00	1.00	1.00	0.50
Bombesin	1.00	1.00	1.00	1.00
Bradykinin	1.00	1.00	1.00	1.00
C5a anaphylatoxin	1.00	1.00	1.00	1.00
Fmet-leu-phe	1.00	1.00	1.00	1.00
APJ-like	1.00	1.00	1.00	1.00
Interleukin-8	1.00	1.00	1.00	0.83
C-C chemokine	0.86	1.00	0.93	0.83
C-X-C chemokine	1.00	1.00	1.00	1.00
C-X3-C chemokine	1.00	1.00	1.00	0.33
Cholecystokinin CCK	1.00	1.00	1.00	1.00
Endothelin	1.00	1.00	1.00	1.00
Melanocortin	1.00	1.00	1.00	1.00
Duffy antigen	1.00	1.00	1.00	1.00
Neuropeptide Y	1.00	1.00	1.00	1.00
Neurotensin	1.00	0.67	0.82	1.00
Opioid	1.00	1.00	1.00	1.00
Somatostatin	1.00	1.00	1.00	1.00
Tachykinin	1.00	1.00	1.00	1.00
Vasopressin/vasotocin	1.00	1.00	1.00	1.00
Oxytocin/mesotocin	1.00	1.00	1.00	1.00
Galanin	1.00	1.00	1.00	1.00
Thrombin	1.00	1.00	1.00	1.00
Proteinase-activated	1.00	1.00	1.00	1.00
Orexin	1.00	1.00	1.00	1.00
Neuropeptide FF	1.00	1.00	1.00	1.00
Urotensin II	1.00	1.00	1.00	1.00
GPR37/endothelin B-like	1.00	1.00	1.00	1.00
Chemokine receptor-like	1.00	1.00	1.00	1.00
Neuromedin U-like	1.00	1.00	1.00	1.00
Somatostatin- and angiogenin-like peptide	1.00	1.00	1.00	1.00
Melanin-concentrating hormone receptors	1.00	1.00	1.00	1.00
Prokineticin receptors	1.00	1.00	1.00	1.00
Follicle stimulating hormone	1.00	1.00	1.00	1.00
Lutropin-choriogonadotropic hormone	0.50	1.00	0.71	1.00
Thyrotropin	1.00	1.00	1.00	1.00
Rhodopsin vertebrate	1.00	1.00	1.00	1.00
Rhodopsin Other	0.40	1.00	0.63	1.00
Prostaglandin	1.00	1.00	1.00	0.89
Prostacyclin	1.00	1.00	1.00	1.00
Thromboxane	1.00	1.00	1.00	1.00
Adenosine	1.00	1.00	1.00	1.00
Purinoceptors	1.00	1.00	1.00	1.00
Cannabinoid	1.00	1.00	1.00	1.00
Platelet activating factor	1.00	1.00	1.00	1.00
Gonadotropin-releasing hormone	1.00	1.00	1.00	1.00
Thyrotropin-releasing hormone	1.00	1.00	1.00	1.00
Growth hormone secretagogue	1.00	1.00	1.00	1.00
Melatonin	1.00	1.00	1.00	1.00
Lysosphingolipid and LPA (EDG)	1.00	1.00	1.00	1.00
Leukotriene B4 receptor	1.00	1.00	1.00	1.00
SREB	1.00	1.00	1.00	1.00
Mas proto-oncogene and Mas-related (MRGs)	1.00	1.00	1.00	1.00
RDC1	1.00	1.00	1.00	1.00
LGR-like (hormone receptors)	1.00	1.00	1.00	1.00

Table 5 (*Continued*)

Subfamily	Subfamily level			Subtype level
	Sen[a]	Spe[b]	Acc[c]	Sen
GPR	0.15	1.00	0.39	1.00
GPR45-like	1.00	1.00	1.00	1.00
Cysteinyl leukotriene	1.00	1.00	1.00	1.00
Free fatty acid receptor (GP40, GP41, GP43)	1.00	1.00	1.00	1.00
Calcitonin	1.00	1.00	1.00	1.00
Corticotropin releasing factor	1.00	1.00	1.00	1.00
Glucagon	1.00	1.00	1.00	1.00
Growth hormone-releasing hormone	1.00	1.00	1.00	1.00
Parathyroid hormone	1.00	1.00	1.00	1.00
PACAP	1.00	1.00	1.00	1.00
Secretin	1.00	1.00	1.00	0.67
Vasoactive intestinal polypeptide	1.00	1.00	1.00	1.00
EMR1	1.00	0.67	0.82	1.00
Latrophilin	1.00	1.00	1.00	1.00
Brain-specific angiogenesis inhibitor (BAI)	1.00	1.00	1.00	0.67
Cadherin EGF LAG (CELSR)	1.00	1.00	1.00	1.00
Metabotropic glutamate group I	1.00	1.00	1.00	1.00
Metabotropic glutamate group II	1.00	1.00	1.00	1.00
Metabotropic glutamate group III	1.00	1.00	1.00	0.67
Extracellular calcium-sensing	1.00	1.00	1.00	1.00
GABA-B	1.00	1.00	1.00	1.00
Orphan GPRC5	1.00	1.00	1.00	1.00
Taste receptors (T1R)	1.00	1.00	1.00	1.00
Frizzled Group A (Fz 1, 2, 4, 5, 7–9)	1.00	1.00	1.00	0.33
Frizzled Group B (Fz 3 and 6)	1.00	1.00	1.00	1.00
Vomeronasal receptors (V1R and V3R)	1.00	0.80	0.89	1.00
Taste receptors T2R	0.86	1.00	0.93	1.00

[a] Sensitivity.
[b] Specificity.
[c] Accuracy.

with a 100% success rate and six subfamilies, i.e., serotonin, C-C chemokine, neurotensin, EMR1, vomeronasal receptors and taste receptors T2R, were classified with success rates higher than 80%. In the case of three subfamilies, i.e., Lutropin-choriogonadotropic hormone, Rhodopsin Other and GPR, a perfect specificity of 100% was seen, but with recorded low sensitivities of 0.5, 0.4 and 0.15, respectively. About half of the Lutropin-choriogonadotropic hormone receptors were classified into other subfamilies, including follicle stimulating hormone or Thyrotropin belonging to the same hormone receptor group. The GPR subfamily belonging to the Class A Orphan/other group include so many kinds of subtypes; the diffuseness in the GPR subfamily lowered the sensitivity.

3.2. Performance comparison of GPCR classifiers

In order to ascertain the quality of our approach, we compared our scheme with three published methods, namely the covariant discriminant algorithm by Elrod and Chou (2002), the bagging classification tree by Huang et al. (2004) and the support vector machine (SVM) by Bhasin and Raghava (2005). To guarantee an objective comparison, we also applied our scheme to the dataset described by Elrod and Chou (2002) in the same manner as the above methods. The amine family dataset include 167 GPCRs, of which 31 are acetylcholine, 44 are adrenoceptors, 38 are dopamine and 54 are serotonin types. As shown in Table 6, the overall accuracy of the TM-HMM-Library is similar to the

Table 6
The performance of the covariant discriminant algorithm, bagging classification tree, GPCRsclass and TM-HMM-Library in classifying the amine receptors

Amine receptors	CovDis[a]	Bagging[b]		GPCRsclass[c]		TM-HMM-Library	
	Acc	Acc	MCC	Acc	MCC	Acc	MCC
Acetylcholine	67.7	96.8	0.94	93.6	0.96	100	1
Adrenoceptor	88.6	90.9	0.82	100	0.93	94.2	0.88
Dopamine	81.6	84.2	0.73	92.1	0.95	97.4	0.95
Serotonin	88.9	81.5	0.77	98.2	0.97	96.3	0.93
Overall	83.2	87.4		96.4	0.95	96.6	0.93

[a] Covariant discriminant algorithm (Elrod and Chou, 2002).
[b] Bagging classification tree (Huang et al., 2004).
[c] GPCRsclass (Bhasin and Raghava, 2005).

Table 7
The number of novel candidates with the previously known isoforms detected in the TISA database

Class	Subfamily	Subtype	Known[a]	Novel[b]	Species
A	Histamine	HRH3	5	2	Human
	Serotonin	5HT4R	5	1	Human
		5HT7R	1	1	Mouse
	Fmet-leu-phe	FPRL1	1	2	Mouse
	APJ-like	APJ	1	1	Human
	C-X-C chemokine	CXCR3	1	1	Human
	Cholecystokinin CCK	CCKAR	1	1	Mouse
	Endothelin	EDNRA	3	1	Human
	Melanocortin	MSHR	1	1	Human
	Duffy antigen	Duffy	1	1	Mouse
	Neuropeptide Y	NPY2R	1	1	Mouse
	Neurotensin	NTR2	1	1	Mouse
	Opioid	OPRM	3	4	Human
		OPRX	4 (4)	1 (7)	Human (mouse)
	Vasopressin/vasotocin	V2R	1 (1)	1 (1)	Human (mouse)
	Proteinase-activated-like	PAR2	1	1	Mouse
	Orexin	OX1R	2	2	Human
		NPFF2	2	2	Human
	Melanin-concentrating hormone receptors	MCHR1	1	1	Human
	Rhodopsin vertebrate	OPSB	1	2	Mouse
		OPSD	1	3	Mouse
		OPSG	1	5	Mouse
	Rhodopsin Other	OPSX	3	1	Mouse
		RGR	3 (1)	1 (3)	Human (mouse)
	Prostaglandin	PE2R1	1	1	Mouse
		PE2R4	2	2	Mouse
	Adenosine	AA1R	4	1	Human
		AA3R	2 (1)	1 (1)	Human (mouse)
	Purinoceptors	P2RY6	6	1	Human
		P2Y10	3	1	Human
	Gonadotropin-releasing hormone	GNRHR	1	1	Mouse
	Lysosphingolipid and LPA (EDG)	EDG7	1	1	Human
	Leukotriene B4 receptor	LT4R1	1	1	Mouse
	LGR-like	RXFP1	1	2	Human
		LGR4	2	1	Human
		LGR8	1	1	Human
	GPR	GPR19	7	1	Mouse
B	Secretin	SCTR	1	1	Mouse
	Vasoactive intestinal polypeptide	VIPR1	2	1	Human
		VIPR2	2	2	Human
	EMR1	EMR1	2	5	Mouse
	Latrophilin	LPHN2	5	4	Human
	Brain-specific angiogenesis inhibitor (BAI)	BAI1	2	5	Mouse
		BAI2	2 (2)	5 (5)	Human (mouse)
		BAI3	3 (2)	1 (2)	Human (mouse)
	Cadherin EGF LAG (CELSR)	CELR2	2	4	Human
C	Metabotropic glutamate group III	MGR4	2	3	Human
		MGR7	3	1	Human
	GABA-B	GABR1	4	1	Human
	GPRC5	GPC5C	3	7	Human
		RAI3	1	1	Human
	Taste receptors (T1R)	TS1R2	1	1	Mouse
T[c]	Taste receptors T2R	T2R14	1	2	Human

[a] Previously known GPCRs detected in TISA.
[b] Novel candidate splice variants detected in TISA.
[c] Taste receptors T2R.

success rate of GPCRsclass with 96.6%, and higher than both the covariant discriminant algorithm and bagging classification tree. As for the adrenoceptor and serotonin receptors, the accuracy is 5.8 and 1.9% lower than that of GPCRsclass. Although their web tool GPCRsclass showed a higher accuracy for the adrenoceptor and serotonin receptors in the comparison test, unlike the TM-HMM-Library that can be applied to the whole GPCR family, the application of GPCRsclass is only limited to the classification of the amine family. The comparison study indicates that the TM-HMM-Library can be applied to the recognition of the GPCR subfamily with high discriminative potency.

3.3. GPCR subtype classification

The results of the 5-fold cross-validation experiments classifying the test set into subfamilies appear in Tables 4 and 5. We designed our subtype experiment with the test set correctly classified into its subfamilies. The overall classification sensitivity of the subtype experiment reached 94.57% and the detailed results are shown in Table 5. The scatter plots and histograms in Figs. 3 and 4 indicate the results of the subtype experiment of the test sequence (AA1R_RAT) and show that a strong relationship between the test sequence and the subtype, AA1R (Table 3). The strong point of the classification algorithm is that it can also show a high discriminative potency in the recognition of the truncated forms caused by alternative splicing.

3.4. Identification of novel splice variants

Encouraged by the performance of the combined approach, we decided to apply the scheme to discover novel GPCR splice variants in a list of human and mouse transcript isoforms. In a previous study, we had performed in silico approach, i.e. a splice graph analysis based on a genomic cluster of mRNA/ESTs to generate a full spectrum of possible transcript variants of human and mouse. As a result, we obtained 97,286 and 66,022 valid transcripts from 26,143 human and 27,741 mouse genes, respectively. In addition, we tested the tissue-specificity of each gene and transcript isoform statistically based on library tissue information of the clustered ESTs. By integrating the information of alternative splicing and tissue-specificity of genes and transcripts, we have developed the tissue-specific alternative splicing (TISA) database (http://tisa.kribb.re.kr/AGC/) (for a description of the splice graph and transcript reconstruction methods in detail, please refer to Noh et al., 2006). Based on 97,286 human and 66,022 mouse transcripts, protein sequences were deduced by choosing the maximum length ORFs among all possible three frame translations from their corresponding RNA sequences.

Using the combined approach, 563 human transcripts and 435 mouse transcripts, which were generated from 305 human genes and 259 mouse genes, respectively, were identified as encoding GPCRs and classified into their subtypes. From these 998 protein sequences, we searched for novel splice variants which differed from previously known proteins in the splicing pattern of their genomic alignment structures that caused any differences in the protein coding region such as domain/motif insertion, deletion, substitution, and N-terminal and/or C-terminal truncation. Splicing variants, which cause alternations only in the 5′ or 3′ UTR were ignored. We found 60 human transcript isoforms and 56 mouse transcript isoforms, which were generated from 33 human genes and 26 mouse genes, respectively, as candidates of novel splice variants and confirmed their novelty by querying 116 protein sequence against the GenBank nr-protein database (09/2006) with the BLASTP program (version 2.2.14), which gave a no match result to the already known protein entries. Table 7 shows the number of novel candidates in each GPCR subtype with the previously known transcripts detected in the TISA database. A complete list of the novel candidates with the illustrated figures showing the alternative splicing patterns and differences in the coding regions that are compared to known isoform(s) for each gene unit is available in Supplementary data and the detailed descriptions are linked directly through the TISA database.

We have introduced here an approach for the recognition and classification of GPCRs with novel candidate splice variants. The classifier combining fingerprint, statistical profiles and physicochemical properties of TM domains shows a higher accuracy of 97.88% for subfamily classification, and 94.57% for subtype classification. A comparative experiment, which was applied to the classification of amine receptors, also showed a strong power of our approach. A good performance in the classification of the receptors motivated us to explore novel candidate splice variants that differed in particular intracellular and extracellular domains. The functional distinction of the receptor isoforms by different tissue-specific distribution, ligand-binding profile and coupling efficiency to G protein has been previously reported (Kilpatrick et al., 1999; Minneman, 2001). To our knowledge, this is the first attempt to identify potential splice variants by a computational approach and from the application of the method to the TISA database, 116 novel candidates were identified from a list of human and mouse transcript isoforms of the TISA database. Table 7 lists the number of novel candidates for each GPCR subtype with the previously known transcripts that were also detected in this study, and Supplementary data shows their alternative patterns and more detailed information.

The GPCR classification experiment, which places the receptors into functionally related subfamilies and subtypes, is beyond mere academic curiosity. The central role of the receptors in regulating crucial cellular processes form a major part of the major pathophysiological conditions, including cardiovascular disease and cancer, and has placed them in the pharmaceutical spotlight. Consequently, GPCR classification algorithms locating the receptors into correct subfamilies and subtypes can be used to help identify and characterize receptors. Thus, the combined approach can be complementary to the characterization of GPCRs and thereby may be useful for discerning side effects and efficacy problems and should facilitate drug discovery.

Acknowledgements

This work was supported by the second stage of the Brain Korea 21 Project in 2007 and was also supported by a fund (grant no. FGM0300512) from the Korea Science and Engineering Foundation for the international joint research project

(project no. M6-0401-00-0178) of the Korea Research Institute of Bioscience and Biotechnology (Korea) and the Weizmann Institute of Science (Israel).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.compbiolchem.2007.05.002.

References

Baldwin, J.M., 1993. The probable arrangement of the helices in G protein-coupled receptors. EMBO J. 12, 1693–1703.

Baldwin, J.M., 1994. Structure and function of receptors coupled to G proteins. Curr. Opin. Cell Biol. 6, 180–190.

Bhasin, M., Raghava, G.P.S., 2005. GPCRsclass: a web tool for the classification of amine type of G protein-coupled receptors. Nucleic Acids Res. 33, W143–W147.

Bjarnadottir, T.K., Gloriam, D.E., Hellstrand, S.H., Kristiansson, H., Fredriksson, R., Schioth, H.B., 2006. Comprehensive repertoire and phylogenetic analysis of the G protein-coupled receptors in human and mouse. Genomics 88, 263–273.

Boeckmann, B., et al., 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31, 365–370.

Drews, J., 1996. Genomic sciences and the medicine of tomorrow. Nat. Biotechnol. 14, 1516–1518.

Eddy, S.R., 1998. Profile hidden Markov models. Bioinformatics 14, 755–763.

Elrod, D.W., Chou, K.C., 2002. A study on the correlation of G-protein coupled receptor types with amino acid composition. Protein Eng. 15, 713–715.

Gaulton, A., Attwood, T.K., 2003. Bioinformatics approaches for the classification of G-protein-coupled receptors. Curr. Opin. Pharmacol. 3, 114–120.

Gether, U., 2000. Uncovering molecular mechanisms involved in activation of G protein-coupled receptors. Endocr. Rev. 21, 90–113.

Grantham, R., 1974. Amino acid difference formula to help explain protein evolution. Science 185, 862–864.

Graur, D., Li, W.H., 2000. Fundamentals of Molecular Evolution, second ed. Sinauer Associates, Inc., Sunderland, Massachusetts.

Horn, F., Bettler, E., Oliveira, L., Campagne, F., Cohen, F.E., Vriend, G., 2003. GPCRDB information system for G protein-coupled receptors. Nucleic Acids Res. 31, 294–297.

Huang, Y., Cai, J., Ji, L., Li, Y., 2004. Classifying G-protein coupled receptors with bagging classification tree. Comput. Biol. Chem. 28, 275–280.

Inoue, Y., Ikeda, M., Shimizu, T., 2004. Proteome-wide classification and identification of mammalian-type GPCRs by binary topology pattern. Comput. Biol. Chem. 28, 39–49.

Karchin, R., Karplus, K., Haussler, D., 2002. Classifying G-protein coupled receptors with support vector machines. Bioinformatics 18, 147–159.

Kilpatrick, G.J., Dautzenberg, G.J., Martin, G.R., Eglen, R.M., 1999. 7TM receptors: the splicing on the cake. Trends Pharmacol. Sci. 20, 294–301.

Kim, J., Moriyama, E.N., Warr, C.G., Clyne, P.J., Carlson, R., 2000. Identification of novel multi-transmembrane proteins from genomic databases using quasi-periodic structural properties. Bioinformatics 16, 767–775.

Lefkowitz, R.J., 2000. The superfamily of heptahelical receptors. Nat. Cell Biol. 2, E133–E136.

Marinissen, M.J., Gutkind, J.S., 2001. G-protein-coupled receptors and signaling networks: emerging paradigms. Trends Pharmacol. Sci. 22, 368–376.

Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta 405, 442–451.

Minneman, K.P., 2001. Splice variants of G protein-coupled receptors. Mol. Interv. 1, 108–116.

Noh, S.J., Lee, K., Paik, H., Hur, C.G., 2006. TISA: tissue-specific alternative splicing in human and mouse genes. DNA Res. 13, 229–243.

Papasaikas, P.K., Bagos, P.G., Litou, Z.I., Hamodrakas, S.J., 2003. A novel method for GPCR recognition and family classification from sequence alone using signatures derived from profile hidden Markov models. SAR QSAR Environ. Res. 14, 413–420.

Sadka, T., Linial, M., 2005. Families of membranous proteins can be characterized by the amino acid composition of their transmembrane domains. Bioinformatics 21, i378–i386.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.