Pattern Recognition for Subfamily Level Classification of GPCRs Using Motif Distillation and Distinguishing Power Evaluation

Ahmet Sinan Yavuz, Bugra Ozer, and Osman Ugur Sezerman

Faculty of Engineering and Natural Sciences
Sabanci University
Istanbul, Turkey
{asinanyavuz,bozer,ugur}@sabanciuniv.edu

Abstract. G protein coupled receptors (GPCRs) are one of the most prominent and abundant family of membrane proteins in the human genome. Since they are main targets of many drugs, GPCR research has grown significantly in recent years. However the fact that only few structures of GPCRs are known still remains as an important challenge. Therefore, the classification of GPCRs is a significant problem provoked from increasing gap between orphan GPCR sequences and a small amount of annotated ones. This work employs motif distillation using defined parameters, distinguishing power evaluation method and general weighted set cover problem in order to determine the minimum set of motifs which can cover a particular GPCR subfamily. Our results indicate that in Family A Peptide subfamily, 91% of all proteins listed in GPCRdb can be covered by using only 691 different motifs, which can be employed later as an invaluable source for developing a third level GPCR classification tool.

Keywords: g-protein coupled receptors, data mining, pattern recognition.

1 Introduction

G protein coupled receptors (GPCRs) represent the largest family of membrane proteins in the human genome. As their dysfunction contributes to some of the most prevalent human diseases, they are of exceptionally high interest in various areas including the drug industry as more than 50% of modern drugs have GPCRs as their main targets [1]. An important property of the GPCRs is that certain aminoacid residues are well conserved across specific families [2]. This property has been utilized in numerous studies, such as synthesizing new GPCRs [3-6], and developing family classifiers. In addition, all GPCRs share a particular structural framework. Structure of a G-protein-coupled receptor comprises seven α-helical transmembrane domains, an extracellular N-terminus, and an intracellular C-terminus [7].

GPCRs are activated by a diverse range of ligands such as small peptides, amino acid derivatives, taste, light or smell [8]. The general classification for GPCRs in vertebrates is as follows: rhodopsin-like (Family A), secretin-like (Family B), glutamate-like (Family C), Adhesion and Frizzled/Taste2 [9, 10]. In addition to this

T. Shibuya et al. (Eds.): PRIB 2012, LNBI 7632, pp. 267–276, 2012.

classification, there are 4 levels of classification down in classification tree. Family A is the family of highest interest from a pharmaceutical research perspective as besides being more than 80% of all human GPCRs are in this family alone [11], the number of sequences in this family is significantly higher than the others. Therefore, we will also emphasize our efforts on peptides subfamily, which belong to Family A.

Due to their significant roles and their importance in drug design, it is highly crucial to be able to distinguish which ligands a specific GPCR interacts with and which regions of the sequence have a particularly crucial role in ligand binding. However, this process is complex, and it is not easy to identify corresponding regions. Despite the significant amount of pharmaceutical research done in this field, 3D structures of only few GPCR structures are known [9], whereas there are large numbers of GPCR primary sequences have been identified [12]. Therefore, in order to identify and characterize the novel receptors, it is crucial to develop *in silico* methods that only work with primary sequences to determine the ligand binding sites and motifs of these novel receptors.

Additional serious challenge is the classification of orphan GPCR sequences. An orphan GPCR is a sequence that has high similarity to known and annotated GPCR sequences but nothing is known about its structure, physiologic function or the activating ligand. As the difference between the number of annotated sequences and the number of identified sequences raises, so does the number of orphan GPCRs. Besides, considering the contribution of GPCRs to cancer initiation, growth and metastatic spread, identification of orphan GPCRs and revealing the pathways related with these GPCRs is placed in the spotlight as prime candidates for cancer prevention and treatment and the orphan GPCRs are of very high interest as they are not yet identified. Therefore, it is essential to find the rules that cover most of the GPCR sequences especially those in the Family A, which is the family most relevant to human drug design. In this work, we will focus utterly on motif coverage within the Peptides subfamily, which belongs to Family A.

As a quick summary, there are two dominant goals for *in silico* GPCR researches: first is to classify GPCR sequences according to their subfamilies, and second is to identify the key ligand-receptor binding sites and family specific motifs using only the protein sequence information. Unlike many of the previous efforts, major concern of this work is only 3rd level classification of GPCR sequences and exploration and analysis of the presence of any layered motifs that are effective in the determination of sub-subfamily classes. Hence, this work is concerned with not only aiming for an *in silico* motif mining for GPCR classification but also providing a valuable source of conserved motifs for experimentalists and other groups working in 3rd level GPCR classification.

1.1 Related Work

There are many current GPCR classification methods involving various machine learning techniques. One of the most common methods employed in GPCR classification is support vector machines (SVM). In this sense, GPCRpred server [13] is based on 20 different SVMs for different levels of classification where the feature vectors are derived from the dipeptide arrangement of each protein. As reported in [14], SVM classification gives better results compared to BLAST and profile HMMs with around

90% valid classification level. However, there are several failures attached with this approach as it misses the physiochemical properties of the receptors which are vital in determining the matching ligand, leading to inaccurate results.

Another common approach to GPCR classification is usage of Hidden Markov Models (HMM). PRED-GPCR [15] server uses this approach with employing 265 signature profile HMMs in the classification of GPCR sequences. However, HMM based prediction methods are not optimal in predicting subfamilies. In addition, likewise SVM based methods, HMM-based methods also bears the problem of opaqueness, yet they are not straightforward to discover key ligand interacting sites of the receptors.

In addition to these techniques, a HMM/SVM hybrid method is utilized for GPCR classification. Named the GRIFFIN Project [16], this project combines the efficiency of HMM-based prediction with predictive power of SVM.

In addition to these widely used methods, there were also some other methods [17,18] proposed which use a number of metrics to make classification efforts more successful and summarize the amino acids of a sequence in a number of continuous parameters. Additionally, Davies et al. [19] proposed a method using 10 different classification algorithms, which employs the structural and physiochemical properties of amino acids, to perform a hierarchical GPCR sequence classification. In this method, best resulting classification method at each level is employed in progressing down the classification tree. Even though, they have various superiorities, all these methods lack the necessary transparency to determine the key ligand receptor interaction sites and identify specific residues.

In overall, current methods in GPCR prediction are suffering mainly from opaqueness of models and impossibility of extracting information out of models in addition to classification. Identifying key interaction sites conserved in families, sub-families or sub-sub families will be beneficial in classification of orphan GPCR sequences. Hence, this work mainly aims to extract possible ligand-receptor interaction sites for each sub-subfamily via identifying the key motifs that cover protein families.

2 Methods

In order to form our training set, we have obtained 304 peptide subfamily sequences from GPCRdb, which includes 32 different sub-subfamilies, such as angiotensin, bombesin, and bradykinin. We aimed to find covering motifs for each of these sub-subfamilies via our pattern recognition method.

Our proposed method in this work can be summarized as follows:

1. Motif distillation by Motif Specificity Measure
2. Distinguishing Power Evaluation of distilled motifs
3. Motif selection with general weighted set cover problem

Briefly, motif distillation step is used to discriminate family specific motifs from randomly generated pool of motifs. Subsequently, distinguishing power evaluation (DPE) of the distilled motifs is used to determine the efficiency of the motifs in sub-subfamily classification with assigned DP score value to enable comparison between each other. Lastly, DP score assigned top selected motifs are used in general weight set cover problem to find out the smallest set of motifs that can cover the maximum amount of proteins located in peptide subfamily.

2.1 Amino Acid Grouping

It is commonly known that there 20 amino acids present considering the proteins. It is challenging to determine fixed length conserved motifs within a protein family using 20-letter amino acid alphabet. Through the evolution, families binding to the same ligand change their sequence while preserving the physicochemical properties of the binding site the same. Therefore, it is very difficult to find identical binding signals within a family. To be able to capture similar motifs, which are different in their sequence, a common approach is to reduce this 20-letter alphabet to a smaller number by grouping the similar amino acids together. At this stage, there are several basic physicochemical properties such as hydrophobicity, charge, and mass, which can be used as an origin of grouping. For our approach, Sezerman grouping of amino acids is used, which is proposed at Cobanoglu et al. [20], and its efficiency tested over other amino acid grouping techniques [21].

Table 1. The amino acid grouping scheme in Sezerman's grouping

Groups	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*
Amino Acids	IVLM	RKH	DE	QN	ST	A	G	W	C	YF	P

2.2 Motif Definition and Motif Specificity Measure

The sequence information without any feature selection cannot be used to perform any rule extraction. The main idea behind motif specificity measure is within a sub-subfamily, certain length aminoacid sequences at specific positions of the same exocellular region would be preserved in comparison to sequences of other sub-subfamilies. The main idea behind this project is that amino acids might be fundamental to the binding process since otherwise they would not have been conserved. The motifs are essential to represent some location specific properties of the sequences, as the objective of this study is to determine key interaction sites as well as extracting set of rules for classification. For this purpose, motifs used in this work are defined similar to the motifs proposed by Cobanoglu et al. [20], which includes information of triplet of residues, the exocellular region of occurrence (n-terminus, exoloop1, exoloop2 or exoloop3) and lastly the position of first residue of triplet relative to the length of the amino acid sequence. In order to determine the transmembrane regions reported in the motif definition, we used TMHMM tool [22].

In general, total number of possible motifs is over hundreds of thousands; nevertheless, most of them occur very infrequently. The ideal motif would be the one that occurs in all the sequences that belongs to a particular sub-subfamily but never in a sequence from another sub-subfamily. In other words, motifs that are unique to a sub-subfamily would be rewarded, whereas motifs that occur either in few sequences or numerous sub-subfamilies, would be penalized. The Term Frequency Inverse Document Frequency (TF-IDF) [23] weight is a metric that measures the occurrences of a word in a family in relation to the overall number of the family members, thus enabling determination of highly family specific motifs. TF-IDF is designed for a parallel purpose and considered as a valid tool at text mining applications, and in this work, the pre-defined TF-IDF weights are used in defining the Motif Specificity Measure, originally with detailed definitions given in Cobanoglu et al. [20]. In short, as its

name suggests, motif specificity measure quantifies the specificity of a motif to a family; hence, indicates that the motif is a random motif or a possibly useful one.

2.3 Distinguishing Power Evaluation

Distinguishing power evaluation (DPE) method aims to determine the best motifs for classification in the training set. Main notion of DPE is repeatedly building decision trees from randomly partitioned test and training data, and looking for those motifs that occur very frequently in each of these decision trees [20]. During this process, DPE picks the motifs initially determined by TFIDF with the highest sub-subfamily specificity using the motif specificity measure.

Apart from its specificity to a certain sub-subfamily, there is a need for an independent comparison criterion between motifs in their distinguishing power. To create such a comparable criterion for assessing a motif's importance in classification, DPE method calculates a distinguishing power (DP) score, which is simply the sum of the accuracies of the decision trees in which that motif occurs [20]. By this score, it is possible to identify the motifs with high information gain and which may be vital in classification. More detailed instructions on DPE can be found in Cobanoglu et al. [20].

In the use of DPE method, three different total number of motifs are tested, 250, 500 and 1000 motifs, from the top of the list of distilled motifs with descending DP scores and assessed their power to cover the whole dataset, individually.

2.4 General Weighted Set Covering Model

DPE selected motif set contains various weak motifs that have a limited contribution in covering the subfamily dataset. These weak motifs may cause overlearning of training data. Therefore, frequently occurring and a complete sub-subfamily covering minimum set of motifs would be more reliable in correct classification of unseen data. Otherwise, motifs that have smaller coverage will optimize the performance on training data and decrease the accuracy of classification algorithm in unseen data.

In order to achieve a minimum number of motifs that can explain maximum portion of each given sub-subfamily datasets, it has been implemented a general weighted set covering model on DPE selected motifs. For each motif, a set of proteins which have that motif is defined separately. Additionally, considering these sets, we applied general weight set covering model to determine the minimum number of motifs which cover all of the proteins in sub-subfamilies, but not all the sub-subfamilies. In detail, this model initially calculates occurrences of each motif in all dataset proteins and each sub-subfamily proteins separately. Afterwards, calculated presence counts were used for calculating ratio1 and assessing the weight, or importance, of that motif in sub-subfamily coverage. Motifs were then sorted according to their weights, and for each sub-subfamily, highest ranked motifs were selected until no further improvement in coverage occurs.

Motif weights have been calculated via different weighting schemes including but not limited to equal weighting of motifs and maximum cardinality of motifs [26]; however, both of these criterions lack the information on specificity of a motif to a subfamily. In other words, these criterions do not provide sufficient information to distinguish subfamilies from each other, but they merely provide information on their presence in whole dataset. In order to overcome this problem, we used a maximum ratio1 criterion, which represents motif coverage in a particular family of proteins in comparison with its existence in all other families [24].

A weight of a motif for all sub-subfamilies based on maximum ratio1 criterion is calculated as:

$$W_i = \frac{|Proteins\ covered\ in\ sub{-}subfamily\ i|}{|Protein\ covered\ in\ whole\ subfamily|}, \forall i\ \epsilon\ Selected\ Subfamily \quad (1)$$

where, i is a sub-subfamily in given subfamily dataset. According to this criterion, if a motif only appears in one subfamily with high coverage, which is a desirable result for classification purposes, its ratio1 value will be become 1, and it will be regarded as an important motif in the model.

3 Experimental Results

Associated DPE runs with motif count 250, 500 and 1000 resulted in 43%, 51%, and 91% coverage of sub-subfamily proteins, respectively. In DPE count = 250 case, the set covering model only selected 174 motifs for maximum coverage, while for DPE count = 500, there were 329 selected motifs selected. For DPE count = 1000 experiment, our model picked 691 motifs for maximum coverage, which as a result came out to be the most efficient DPE count. Full list of the motifs is available as supplementary material (Supplementary Table 1). Although increasing DPE motif count shows parallel behavior to the sub-subfamily coverage trend, each step increases in DPE motif count results in a significant increase in computational time. Besides, in order to avoid overlearning, we decided to keep motif count in a limit. Therefore, we chose DPE count = 1000 as our best result for further studies. The detailed results of the selected DPE experiment and applied general weighted set covering model is included in Table 2.

Several sub-subfamilies, namely Duffy-antigen and GPR37 endothelin B-like, shows a consistent low coverage between different DPE counts, indicating motifs that are effective on classification of these sub-subfamilies have low distinguishing power (DP score) and therefore are not selected within given DP motif counts. Hence, classification to these sub-subfamilies can be more difficult, since covering motifs are not specific enough. On the other hand, adrenomedullin family indicates a high coverage for each count set (100%, 76%, 73% respectively for DP count = 1000, 500, and 250). Consistent high coverage for adrenomedullin sub-subfamily indicates that family-specific motifs have a high DP scores showing family's distinct nature, and these motifs ranked mostly in top 250 motifs. Similar kind of behavior is also seen in anaphylatoxin sub-sub family with 96%, 74%, 43% protein coverage for given DPE counts respectively. The significant reduction in DPE count = 250 indicates that these sub-subfamily specific motifs are mostly ranked in 250-500 range. Another significant sub-subfamily result has been obtained at the case of allostatin C. Within the same trend of adrenomedullin and anaphylatoxin, 95%, 66%, 46% protein coverages were obtained for tested DPE counts. As a summary, these results validate our findings on the DPE motif selection thresholds and have an important effect on the scope of possible protein coverage. The most important 5 motifs and their locations for adrenomedullin, anaphylotoxin, and allostatin C sub-subfamily are given in the Table 3. Also, location distributions of selected motifs for these three sub-subfamilies are summarized in a histogram, Figure 1. Via analyzing the difference between high and low DP scored motifs for different sub-subfamilies, the complex sub-subfamilies can be identified and used as an additional insight in developing 3rd level GPCR classification methods.

Table 2. General weight set covering model on DPE evaluated motifs resulted in listed coverage for each sub-subfamily in GPCRdb

Family	Total Protein	Covered Protein	Percentage
Adrenomedullin	33	33	1.00
Allatostatin C	41	39	0.95
Anaphylatoxin	98	94	0.96
Angiotensin	180	147	0.82
APJ like	90	59	0.66
Bombesin	163	149	0.91
Bradykinin	223	222	1.00
Chemokine	1286	1080	0.84
Chemokine receptor-like	77	73	0.95
Cholecystokinin	170	163	0.96
Duffy antigen	51	25	0.49
Endothelin	144	130	0.90
Fmet-leu-phe	305	279	0.91
Galanin-like	405	371	0.92
GPR37 endothelin B-like	78	44	0.56
Interleukin-8	118	112	0.95
Melanin-CHormone Recep family	228	219	0.96
Melanocortin	789	749	0.95
Neuromedin U-like	215	177	0.82
Neuropeptide Y	1329	1262	0.95
Neurotensin	59	57	0.97
Opioid	288	256	0.89
QRFP family	86	81	0.94
Prokineticin receptors	98	93	0.95
Prolactin-releasing peptide	113	89	0.79
Proteinase-activated like	250	243	0.97
Somatostatin- and angiogenin-like	96	95	0.99
Somatostatin	376	361	0.96
Sulfakinin CCKLR	9	9	1.00
Tachykinin	270	265	0.98
Urotensin II	47	31	0.66
Vasopressin-like	436	412	0.94
TOTAL:	8151	7419	0.91

Table 3. Highest ranked 5 reduced alphabet motifs and their location for adrenomedullin, allastostatin C and anaphylotoxin. Position within a loop is defined as being the sequential position of the first letter of triplet, normalized by length of the corresponding loop. 0,1,2,3 correspond to the first, second, third and fourth quarter of the exoloops and n-terminus respectively.

Sub-subfamily	Motif	Location	Position within Loop
Adrenomedullin	CAA	exoloop 2	0+1
Adrenomedullin	JEA	exoloop 1	0
Adrenomedullin	KCA	exoloop 2	0
Adrenomedullin	JBE	exoloop 3	0
Adrenomedullin	GFA	n-terminus	2
Allatostatin C	JAA	exoloop 1	0+1
Allatostatin C	AAA	exoloop 3	0
Allatostatin C	EEJ	n-terminus	1+2
Allatostatin C	ACD	n-terminus	1+2
Anaphylatoxin	AEA	exoloop 2	0+1
Anaphylatoxin	EJA	exoloop 2	1
Anaphylatoxin	AAC	exoloop 3	2
Anaphylatoxin	BBA	exoloop 2	2
Anaphylatoxin	CJC	n-terminus	2

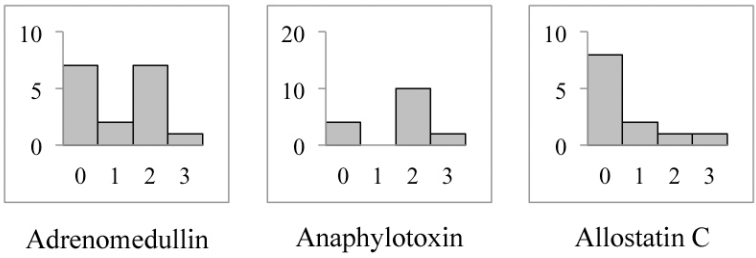

Fig. 1. Histograms of motif locations present in selected motifs for adrenomedullin, anaphlotoxin and allostatin C sub-subfamilies. In x-axis locations are denoted as 0 for n-terminus, 1 for exoloop 1, 2 for exoloop 2 and 3 for exoloop 3.

As DPE counts and selected motif counts differ notably, it can be concluded that our motif selection step helps to eliminate the motifs with high DP score and limited in the information they bring to coverage of sub-subfamily. Besides, large numbers in selected motif sets with large coverage rates indicate that these motifs can be used rule out complex patterns in transmembrane regions of GPCR receptors determining the sub-subfamily of the protein.

4 Conclusions and Future Work

In the light of recent findings, DPE method and combined applied general weight set model can be used for determining the motif set that can be used for developing classifiers for 3rd level GPCR classification problem. As 3rd level GPCR motif identification has not explored extensively in literature before, we hope that our method of obtaining minimal set of important motifs with high specificity will be a stepping stone for further developments in sub-subfamily GPCR classification. Our example case of Peptide sub-subfamily showed that our method can find important motifs for obtaining significantly large family coverage.

As can be seen from Figure 1, adrenomedullin family mostly binds from the motifs in the n-terminus and exoloop 2. These motifs mostly include negatively charged residues followed by aliphatic hydrophobic residues or ring structures and positively charged residues and/or Serine or Threonine (Table 3). Whereas anaphylotoxin mostly binds from the motifs occurring at exoloop 2 and allostatin C mostly binds from the n-terminus. Our method provides location and binding motif information each of the peptide sub-subfamilies, which are very valuable for drug development.

The future work lies in quantifying the actual predictive performance of selected motifs and developing a classification server via generalizing motif sets for each and every sub-subfamily present in the GPCRdb.

Acknowledgments. Authors would like to express their gratitude to Cem Meydan (Sabanci University) and Ceyda Sol (Sabanci University) for their valuable discussions.

Supplementary Material

Supplementary Table 1 – A complete list of motifs for all sub-subfamilies can be accessed online via http://bit.ly/O2Kk6N.

References

1. Filmore, D.: It's a GPCR World. Modern Drug Discovery 7(11), 24–28 (2004)
2. Joost, P., Methner, A.: Phylogenetic analysis of 277 human G-protein- coupled receptors as a tool for the prediction of orphan receptor ligands. Genome Biology 3(11), research0063.1–research0063.16 (October 2002)
3. Davey, J., Ladds, G.: Heterologous Expression of GPCRs in Fission Yeast. Methods in Molecular Biology 746, 113–131 (2011)
4. Gerber, S., Krasky, A., Rohwer, A., Lindauer, S., Closs, E., Rognan, D., Gunkel, N., Selzer, P.M., Wolf, C.: Identification and characterisation of the dopamine receptor II from the cat flea Ctenocephalides felis (CfDo- pRII). Insect Biochemistry and Molecular Biology 36(10), 749–758 (2006)
5. Libert, F., Parmentier, M., Lefort, A., Dinsart, C., Van Sande, J., Maenhaut, C., Simons, M.J., Dumont, J.E., Vassart, G.: Selective amplification and cloning of four new members of the G protein-coupled receptor family. Science 244(4904), 569–572 (1989)

6. Methner, A., Hermey, G., Schinke, B., Hermans-Borgmeyer, I.: A novel G protein-coupled receptor with homology to neuropeptide and chemoattractant receptors expressed during bone development. Biochemical and Biophysical Research Communications 233(2), 336–342 (1997)
7. Horn, F., Bettler, E., Oliveira, L., Campagne, F., Cohen, F.E., Vriend, G.: GPCRDB information system for G protein-coupled receptors. Nucleic Acids Research 31(1), 294–297 (2003)
8. Gether, U.: Uncovering molecular mechanisms involved in activation of G protein-coupled receptors. Endocrine Reviews 21(1), 90–113 (2000)
9. Rosenbaum, D.M., Rasmussen, S.R.G.F., Kobilka, B.K.: The structure and function of G-protein-coupled receptors. Nature 459(7245), 356–363 (2009)
10. Foord, S.M., Bonner, T.O.M.I., Neubig, R.R., Rosser, E.M., Pin, J.P., Davenport, A.P., Spedding, M., Harmar, A.J.: International Union of Pharmacology. XLVI. G Protein-Coupled Receptor List. Pharmacological Reviews 57(2), 279–288 (2005)
11. Davies, M.N., Secker, A., Halling-Brown, M., Moss, D.S., Freitas, A.A., Timmis, J., Clark, E., Flower, D.R.: GPCRTree: online hierarchical classification of GPCR function. BMC Research Notes 1, 67 (2008)
12. Gaulton, A., Attwood, T.K.: Bioinformatics approaches for the classification of G-protein-coupled receptors. Current Opinion in Pharmacology 3(2), 114–120 (2003)
13. Bhasin, M., Raghava, G.P.S.: GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. Nucleic Acids Research 32(Web Server Issue), W383–W389 (2004)
14. Karchin, R., Karplus, K., Haussler, D.: Classifying G-protein coupled receptors with support vector machines. Bioinformatics 18(1), 147–159 (2002)
15. Papasaikas, P.K., Bagos, P.G., Litou, Z.I., Promponas, V.J., Hamod- Rakas, S.J.: PRED-GPCR: GPCR recognition and family classification server. Nucleic Acids Research 32(Web Server Issue), W380–W382 (2004)
16. Yabuki Y., Muramatsu T., Hirokawa T., Mukai H., Suwa M.: GRIFFIN: a system for predicting GPCR–G-protein coupling selectivity using a support vector machine and a hidden Markov model. Nucleic Acids Research, 33(Web server issue), W148–W153 (2005)
17. Cui, J., Han, L.Y., Li, H., Ung, C.Y., Tang, Z.Q., Zheng, C.J., Cao, Z.W., Chen, Y.Z.: Computer prediction of allergen proteins from sequence-derived protein structural and physicochemical properties. Molecular Immunology 44(4), 514–520 (2007)
18. Atchley, W.R., Zhao, J., Fernandes, A.D., Druke, T.: Solving the protein sequence metric problem. PNAS 102(18), 6395–6400 (2005)
19. Davies, M.N., Secker, A., Freitas, A.A., Mendao, M., Timmis, J., Flower, D.R.: On the hierarchical classification of G protein-coupled receptors. Bioinformatics 23(23), 3113–3118 (2007)
20. Cobanoglu, M.C., Saygin, Y., Sezerman, U.: Classification of GPCRs using family specific motifs. IEEE Transactions on Computational Biology 8(6), 1495–1508 (2011)
21. Davies, M.N., Secker, A., Freitas, A.A., Clark, E., Timmis, J., Flower, D.R.: Optimizing amino acid groupings for GPCR classification. Bioinformatics 24(18), 1980–1986 (2008)
22. Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L.L.: Predicting trans- membrane protein topology with a hidden Markov model: application to complete genomes. Journal of Molecular Biology 305(3), 567–580 (2001)
23. Salton, G.: Developments in automatic text retrieval. Science 253(5023), 974–980 (1991)
24. Sol, C.: Identification of disease related significant SNPs. M.Sc. Thesis. Faculty of Engineering and Natural Sciences. Sabanci University (2010)