IDENTIFICATION OF INTERACTION SITES OF G PROTEIN-COUPLED RECEPTORS USING MACHINE LEARNING TECHNIQUES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

MEHMET EMRE ŞAHIN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
COMPUTER ENGINEERING

AUGUST 2014

Approval of the thesis:

IDENTIFICATION OF INTERACTION SITES OF G PROTEIN-COUPLED RECEPTORS USING MACHINE LEARNING TECHNIQUES

submitted by **MEHMET EMRE ŞAHIN** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Adnan Yazıcı
Head of Department, **Computer Engineering**

Assoc. Prof. Dr. Tolga Can
Supervisor, **Computer Engineering Department, METU**

Examining Committee Members:

Prof. Dr. Nihan Kesim Çiçekli
Computer Engineering Department, METU

Assoc. Prof. Dr. Tolga Can
Computer Engineering Department, METU

Assoc. Prof. Dr. Çağdaş D. Son
Department of Biological Sciences, METU

Assoc. Prof. Dr. Hasan Oğul
Computer Engineering Department, Başkent University

Assist. Prof. Dr. Aybar C. Acar
Informatics Institute, METU

Date:

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: Mehmet Emre Şahin

Signature :

ABSTRACT

IDENTIFICATION OF INTERACTION SITES OF G PROTEIN-COUPLED RECEPTORS USING MACHINE LEARNING TECHNIQUES

Şahin, Mehmet Emre
Ph.D., Department of Computer Engineering
Supervisor : Assoc. Prof. Dr. Tolga Can

August 2014, 84 pages

G protein-coupled receptors (GPCRs), which play a crucial role in a host of pathophysiological pathways, form the largest and most divergent receptor family. Typically, they transmit outer signals to the inner cell by interacting with G-proteins. The emerging concept of GPCR dimerization has unsettled the classical idea that GPCRs function as monomeric units. Prediction of the interface residues of GPCR dimers is a challenging topic. The method proposed in this thesis trains itself with known interfaces from the literature and makes predictions using both the sequence and three-dimensional structural information about GPCRs. The predictions are assessed by comparison to known interfaces in the literature. Our results show that the predictions are consistent with real interactions; however, further biological validation is still needed. During the development of the method, a new database was published for the use of the community: IntGPCR, the database of interacting GPCRs. IntGPCR contains information about interacting GPCRs, where the contents are curated from the literature. Up-to-dateness and the wealth of its contents, containing 309 interacting GPCRs curated from 348 articles, make IntGPCR a valuable resource for GPCR researchers. The other proposed method is about the classification of the GPCRs, serving to the requirement of an efficient and rapid classification to group the receptors according to their functions. GPCRsort, a new classification tool for GPCRs using the structural features derived from their primary sequences is proposed. Comparison experiments with the current known GPCR classification techniques show that GPCRsort is able to rapidly (in the order of minutes) classify uncharacterized GPCRs with 97.3% accuracy whereas the best available technique's accuracy is 90.7%.

Keywords: GPCR dimerization, interface prediction, classification of GPCRs, database of interacting GPCRs, 3D modelling of proteins

ÖZ

MAKİNE ÖĞRENME TEKNİKLERİ KULLANILARAK G PROTEİN-KENETLİ RESEPTÖRLERİN ETKİLEŞİM BÖLGELERİNİN TESPİT EDİLMESİ

Şahin, Mehmet Emre
Doktora, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi : Doç. Dr. Tolga Can

Ağustos 2014 , 84 sayfa

Patofizyolojik yolaklarda önemli rol oynayan G protein-kenetli reseptörler (GPKR), en büyük ve en çok çeşitli reseptör ailesini oluşturmaktadır. Tipik olarak, G-proteinler ile etkileşerek hücre dışındaki sinyalleri hücre içerisine iletirler. Gelişmekte olan GPKR dimerleşmesi görüşü, GPKR'lerin tek parçalı bireyler halinde görevlerini yerine getirdiği klasik görüşünü geride bırakmıştır. GPKR dimerlerinin arayüzlerinin tahmini ilgi çekici bir konudur. Burada sunulan yöntem, kendisini literatürde bulunan bilinen arayüzler ile eğiterek, GPKRlerin hem sekans, hem de üç-boyutlu yapısal bilgilerini kullanarak tahminlerini yapar. Her ne kadar sonuçların biyolojik olarak tasdiklenmesi gerekiyorsa da, önerilen yöntemin bilinen arayüzler tabanlı değerlendirme sonuçları iç açıcı ve gerçek veriyle uyumludur. Bu yöntemin geliştirilmesi esnasında, araştırmacıların kullanması için IntGPCR adında yeni bir veritabanı yayınlanmıştır. İçeriği literatürden derlenen IntGPCR, etkileşen GPKR'ler hakkında bilgiler içermektedir. Güncelliği ve içeriğinin zenginliği, 348 makaleden çıkarılan 309 etkileşen GPKR bilgisi, IntGPCR veritabanını benzerleri arasında ön plana çıkarır. Tez çalışmaları kapsamında bir diğer geliştirilen metot da GPKR'lerin sınıflandırılması ile ilgilidir. Bu metot, reseptörlerin fonksiyonlarına göre hızlı ve verimli bir şekilde sınıflandırılması ihtiyacına yönelik geliştirilmiştir. GPCRsort, GPKR'lerin birincil sekanslarından elde edilen yapısal özellikleri kullanan yeni bir sınıflandırma aracıdır. Güncel GPKR sınıflandırma teknikleri ile karşılaştırma deneyleri göstermektedir ki, en iyi kullanılabilir tekniğin %90.7 doğruluğa sahip olduğu yerde, GPCRsort %97.3 doğruluk oranı ile GPKR'leri sınıflandırabilmekte ve bunu hızlıca (dakikalar içerisinde) gerçekleştirebilmektedir.

Anahtar Kelimeler: GPKR dimerleşmesi, arayüz tahmini, GPKR'lerin sınıflandırılması, etkileşen GPKR'ler veritabanı, proteinlerin üç boyutlu modellenmesi

To my family

ACKNOWLEDGMENTS

I would like to thank my supervisor Associate Professor Tolga Can for his continuous support, strong guidance, and most importantly friendship with positive attitudes. I would also express my appreciation to Associate Professor Çağdaş Son for his encouragement, help and friendship. It was a great honor to have the chance of working with them throughout my PhD years. Their insightful approaches made me feel so safe in all my studies. Thank you for always making time in your very busy agenda to talk about the evaluation of the status of the studies. I also would like to thank Professor Nihan Kesim Çiçekli for her support and guidance in every our meetings.

The true owners of this work is my family. Firstly, I would like to thank my wife for her great patience and support. Here I would like to both thank to and apologize from my little daughter Zehra declining of her play requests during my PhD studies. I believe she will understand when she will read these lines one day. And my biggest appreciation goes to my mother, father and sister. They were always behind me in all my decisions and encouraged me every time. I could not achieve any success without any of my family members.

I would like to thank my colleagues, especially Cemal Kırlılar and Tunç Tekelioğlu, and my company for supporting my thesis studies.

I would also like to thank Yusuf Talha Tamer, İnci Barut and Melike Dönertaş from Department of Biological Sciences for their help in the scope of my thesis. A special thank goes to Şemsi Cihan Yücel for reading my thesis to comment on my English.

Finally, I would like to thank TÜBİTAK-BİDEB for providing financial support during my thesis studies with the National Graduate Scholarship Program for PhD (2211).

TABLE OF CONTENTS

ABSTRACT v

ÖZ vii

ACKNOWLEDGMENTS x

TABLE OF CONTENTS xi

LIST OF TABLES xv

LIST OF FIGURES xvi

LIST OF ABBREVIATIONS xvii

CHAPTERS

1 INTRODUCTION 1

2 GPCRSORT: A CLASSIFICATION METHOD FOR GPCRS 5

  2.1 Background 5

  2.2 Related Work 6

  2.3 Materials and Methods 8

    2.3.1 GPCR representation 8

    2.3.2 Dataset preparation 9

    2.3.3 Method 10

2.3.4 Environment of experiments 11

2.3.5 Feature vector contents 12

2.3.6 Cross validation experiments 12

2.3.7 Using an independent dataset as validation data . . 13

2.3.8 Comparison with the BTP method 13

2.3.9 Comparison with other methods 13

2.4 Results 14

2.4.1 Effect of feature vector contents 14

2.4.2 Cross validation experiments 15

2.4.3 Using an independent dataset as validation data . . 16

2.4.4 Comparison with the BTP method 17

2.4.5 Comparison with other methods 19

2.4.6 Running Time Analysis 20

2.5 Discussion 21

3 INTGPCR: DATABASE OF GPCR-GPCR INTERACTIONS 23

3.1 Background 23

3.2 Related Work 24

3.3 Materials and Methods 26

3.3.1 Preparation of the Article List 26

3.3.2 Analysis of Articles 28

3.3.3 Development of the Web Interface 28

3.3.4 Modelling of GPCRs 29
3.3.4.1 Modelling Method 31
3.3.4.2 Alignment Method Decision 33
3.3.4.3 Selection of Template Proteins 35
3.3.5 Visualization of Dimer Formations 37
3.4 Presentation of the IntGPCR 38
3.5 Discussion 42
4 INTERACTION SITE PREDICTION OF GPCR DIMERS 45
4.1 Background 45
4.2 Related Work 47
4.3 Materials and Methods 50
4.3.1 Dataset 50
4.3.2 Method 51
4.3.3 Experimental Setup 55
4.4 Results 55
4.5 Discussion 61
5 CONCLUSION AND FUTURE WORK 63
REFERENCES 65

APPENDICES

A GPCR STRUCTURES 75
B INTERACTION SITE PREDICTION RESULTS 79

CURRICULUM VITAE 83

LIST OF TABLES

TABLES

Table 2.1 **Results of using different feature vectors** 15

Table 2.2 **Evaluation results of 10-fold cross validation** 15

Table 2.3 **Evaluation results of 10-fold cross validation using non-redundant dataset** 16

Table 2.4 **Evaluation results of using a separate testing data** 16

Table 2.5 **Classification performance of the method using GPCRpred dataset as the validation data** 17

Table 2.6 **Confusion matrix of the method using GPCRpred dataset as the validation data** 18

Table 2.7 **Classification performance of GPCRsort and BTP method** . . . 19

Table 2.8 **Classification performance of GPCRsort compared with the current known methods** 20

Table 3.1 **Comparison of using different alignment methods** 34

Table 3.2 **Comparison of using different template known models** 36

Table 4.1 **Predicted interface regions of some GPCRs whose interfaces are proposed in the literature** 56

Table A.1 **Experimentally solved GPCR structures** 76

Table B.1 **Prediction Results of Experiment 1** 79

Table B.2 **Prediction Results of Experiment 2** 80

LIST OF FIGURES

FIGURES

Figure 1.1 **The seven-transmembrane α-helical structure of a GPCR.** Gray colored region is the membrane. Differently colored numbered regions are transmembrane domains. (Figure is taken from [103]) 2

Figure 2.1 **Schematic diagram of a GPCR.** 9

Figure 2.2 **Comparison of running times of methods.** GPCRBind running time is taken from [12]. 20

Figure 3.1 **Schema of the IntGPCR database system.** Picture is generated using SchemaSpy [16]. 30

Figure 3.2 **A sample Browse page from the IntGPCR database system.** . . 39

Figure 3.3 **A sample Interaction Details page from the IntGPCR database system.** 40

Figure 3.4 **A sample GPCR Details page from the IntGPCR database system.** 41

Figure 4.1 **Visualization of selected residues. a.** The 3 residues are selected from R'_T. **b.** The 3 residues are selected from R'_P. 52

LIST OF ABBREVIATIONS

GPCR	G protein-coupled receptor
GPCRDB	G Protein-Coupled Receptor Data Base
TM	Transmembrane region
BTP	Binary Topology Pattern method
PDB	RCSB Protein Data Bank

CHAPTER 1

INTRODUCTION

The G-protein-coupled receptors (GPCRs) form a superfamily of integral membrane proteins and this superfamily is one of the largest, divergent and most studied families of proteins [49, 18]. The structure of a GPCR (Figure 1.1) comprises seven highly conserved α-helical transmembrane (TM) domains, three intracellular and three extracellular loops, an extracellular N-terminus and an intracellular C-terminus [27]. A protein is classified as a GPCR if two main requirements are satisfied. The first one is having seven sequence stretches, of about 25 to 35 residues, that are α-helices spanning the membrane. The second requirement is the ability to interact with a G-protein [29].

The main function of GPCRs is to transduce extracellular signals into intracellular reactions. They have a primary role in establishing the sensory and regulatory connection of the cell with the outside world [12]. For outside ligands, they act as receptors and for internal processes, they are actuators. Most GPCRs generate signals when they detect a ligand. This ligand can be from a diverse set including hormones, ions, amines, peptides, lipids, nucleotides, odors, tastes, and photons of light [51]. When the ligand interacts with the GPCR, it initiates some conformational changes and stabilizes the active configuration of the receptor that will activate a G-protein at the cytosolic side. A complex system involving a variety of mechanisms is observed by interaction of more than one type of GPCR with more than one type of G-protein.

Due to the stated functions above, GPCRs play critical roles in physiological processes such as cellular metabolism, neurotransmission, secretion, and cellular differentiation. Due to their significant role, GPCRs are involved in many major diseases

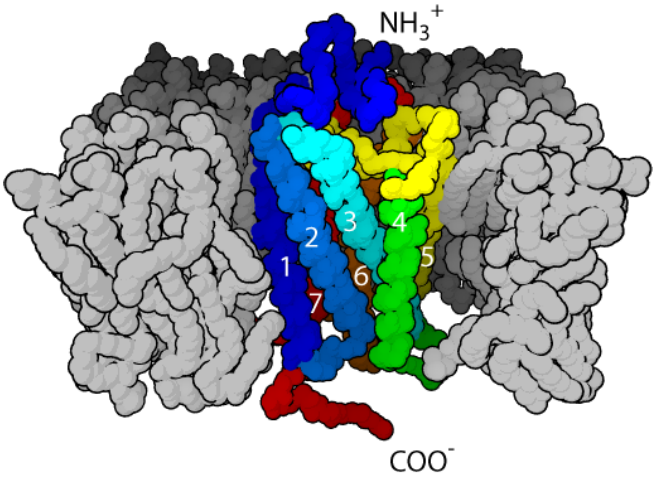


Figure 1.1: **The seven-transmembrane α-helical structure of a GPCR.** Gray colored region is the membrane. Differently colored numbered regions are transmembrane domains. (Figure is taken from [103])

including cancer, psychiatric, metabolic and infectious diseases [51]. This means that, there is a large potential in developing therapeutic drugs that could act on GPCRs [100]. For the pharmaceutical industry, GPCRs are a major target. It is estimated that 60-70% of all medicines in development today target GPCRs [51, 66]. It is also pointed out that drugs have still only been developed to affect a very small number of GPCRs [29].

Today, a large number of protein sequences are identified as GPCRs; yet, their structures and functions are not fully characterized [95]. An organization of these GPCRs into classes is necessary for efficient study and analysis of their functions. It is often desirable to classify a novel protein sequence identified as a GPCR into one of the known classes for inferring its function. There are many known GPCR sequences whose ligands remain unidentified, i.e., orphan GPCRs [91]. Natural functions of those GPCRs are in question. Classification of orphan receptors could decrease the efforts on the initial studies with these types of GPCRs.

So far, GPCRs had been considered to function as monomeric units. This assumption directed the studies of the model of ligand binding and signal transduction [51]. This idea also delimited the efficient analysis of drugs' effects and side effects. However, over the past few years, this idea has been debated. Recent findings clearly show that many GPCRs form homodimers or heterodimers [49]. In studies it is stated that, dimerization occurs early after biosynthesis. Also this is necessary for correct transport of receptor to the membrane, putting forward the idea of primary role of dimerization in the maturation of protein [93]. But the methods searching for GPCR-ligand pairs cannot find the interactions between interacting GPCRs.

Dimerization of GPCRs can be studied with three-dimensional protein structures of GPCRs. Unfortunately, tertiary structures of GPCRs are largely unavailable [66]. Experiments of detecting structures of proteins are costly and time-consuming. Computational methods help here to predict the structures and characters of GPCRs. A large number of GPCR primary sequences are known. Also there exists information about known GPCR dimers. Developing a computational method for predicting interacting GPCRs from known data is valuable.

In this study, the primary focus is the prediction of the interface residues in the GPCR dimers. A novel methodology is proposed for the purpose of this prediction. This proposed method uses both the sequence and structural data about GPCRs to make its predictions. Known interface regions from the literature are used in the operation of the method. For the structural data, created models of the GPCRs as well as the available three-dimensional structures of receptors are used. The performance of the method is evaluated with the analysis of the predictions on the interfaces that are already proposed as known in the literature.

The necessity of a dataset consisting of the known interacting regions of GPCR dimers directed the creation of a database within the studies. This database, the IntGPCR, is a database of GPCR dimers curated from the literature. The system provides researchers a portal for easy access and analysis of GPCR dimers.

Within the studies, a new methodology is proposed for the classification of GPCRs, named GPCRsort. GPCRsort is an effective method in accurately classifying GPCRs into correct classes. Its performance is evaluated with the current available GPCR classification methodologies. GPCRsort gives the highest accuracy, %97.3, among these state of the art techniques. In addition to this great performance, the running time of GPCRsort is also faster than the compared methodology proposed by Cobanoglu *et al.* [12].

To summarize, the contributions of this study can be listed as follows:

- A new GPCR classification method is proposed, named GPCRsort. This proposed method outperforms the state of the art classification techniques in the accuracy and running time.
- The IntGPCR database is created and published for the use of the researchers who studies about GPCR dimers. The IntGPCR, which contains GPCR dimer data curated from the literature, provides browsing, searching and visualization of the interacting GPCRs.
- A novel method is proposed for the prediction of interface residues of GPCR dimers. The performance of the proposed method is evaluated according to the available known interface data.

The organization of the rest of the dissertation is as follows. Chapter 2 explains the studies about GPCRsort. IntGPCR creation and presentation are explained in Chapter 3. The details of the proposed method for the interface prediction of GPCR dimers are presented in Chapter 4. Each of these chapters is organized with the background, related works and experiments about the related topic. The last chapter, Chapter 5, concludes the dissertation with overall discussions and future directions.

CHAPTER 2

GPCRSORT: A CLASSIFICATION METHOD FOR GPCRS

GPCRs require an efficient and rapid classification method to group the members according to their functions. An emerging number of orphan GPCRs demand novel, rapid and accurate classification of the receptors since the current classification tools are inadequate and slow. This chapter presents the development of a new classification tool for GPCRs using the structural features derived from their primary sequences: GPCRsort. Comparison experiments with the current known GPCR classification techniques show that GPCRsort is able to rapidly (in the order of minutes) classify uncharacterized GPCRs with 97.3% accuracy whereas the best available technique's accuracy is 90.7%.

2.1 Background

GPCRs could be classified according to their functions, ligand bindings or their structures. Currently there are several classification schemes. The most widely adopted classification scheme has the following groups: rhodopsin, secretin, glutamate, adhesion and frizzled/taste2 [83, 80]. This scheme is based upon the GPCR superfamily classification system that was introduced by Kolakowski in [45]. This defunct system divides GPCRs into seven families, specified A-F and O, using original standard similarity searches [18]. Horn *et al.* developed this system in [36] for the G Protein-Coupled Receptor Data Base (GPCRDB), which is one of the most popular databases for GPCRs. GPCRDB is organized in a hierarchical structure. GPCRs are divided into six families, stated as A-F, in the first versions of GPCRDB. Later, the database is reorganized and the latest version of the GPCRDB contains five classes at the top level; that are, Class A Rhodopsin like, Class B Secretin like, Class C Metabotropic glutamate/pheromone, Vomeronasal receptors (V1R and V3R), and Taste receptors T2R [102]. Each class is further divided into subclasses except the last one, the Taste receptors. Furthermore, in some families, division continues into further sub-subclasses. Class A is the largest and most studied family [39] which includes more than 80% of all human GPCRs [20]. Class B receptors bind to large peptides [8]. Metabotropic glutamate receptors in Class C bind to glutamate, which is an amino-acid that functions as an excitatory neurotransmitter [18]. The group of receptors named as fungal pheromone receptors include GPCRs that bind to pheromones which are used by organisms for chemical communication [17]. Finally, vomeronasal and taste receptors are putative receptors.

The 3D structure of a GPCR can be very valuable in inferring its function; however, since GPCRs are very difficult to crystallize, techniques such as X-Ray crystallography are not directly applicable. Currently, only 21 different types out of thousands of GPCR structures have been experimentally solved [105]. This makes the sequence of the protein as the primary source to work with.

2.2 Related Work

Several methodologies have been developed to classify GPCRs using their sequence data. Some of these methodologies are motif-based classification techniques, machine learning methods such as Hidden Markov Models or Support Vector Machines (SVM).

GPCRpred [3] is a SVM-based method for predicting families and subfamilies of GPCRs. Five SVMs are built to determine the top-level class of a GPCR and 14 SVMs are used to determine the subfamily of a GPCR if it belongs to the Class A GPCRs. The reported results show that GPCRs can be classified into top classes with 97.5% accuracy [3]. However, the method is insufficient to predict the exact family of the GPCR that is at the leaf of the whole GPCR class hierarchy.

Davies *et al.* propose a strategy to classify GPCRs in [20]. Their method, named GPCRTree, uses an alignment-independent classification system based on amino-acids' physical properties. It employs principal component analysis to select best components for sequence representation. At each level of the GPCR class hierarchy, 10 different classification algorithms are tested and the best performed algorithm is chosen at that level. The disadvantage of this method is its low accuracy for sub-classes and the time consuming calculation at each level for the unused classifiers.

A recent technique, proposed by Cobanoglu *et al.* in [12], uses sequence-derived motifs to classify GPCRs. The motifs they produce characterize the subfamilies by discovering receptor-ligand interaction sites. They propose Distinguishing Power Evaluation technique to select the best motifs for a subfamily. In their reported results, it is stated that their method outperformed the state-of-the-art techniques for GPCR Class A subfamily prediction. The deficiency of this algorithm is; its prediction covers only certain subfamilies of the Class A family. It cannot predict GPCRs from other classes or cannot state the exact class of the GPCR like in the case of the GPCRpred algorithm. Another point to be emphasized in this algorithm is its computational complexity. Running time of the algorithm is too long to make a GPCR prediction. Since GPCRBind is a rule extraction method, training takes time on the order of hours, i.e. 31 hours for 90.7% accuracy (Figure 8 in [12]).

Inoue *et al.* propose a method, named the Binary Topology Pattern (BTP) method, for the classification of GPCRs [37]. Their classifier is similar to the proposed method in this chapter, GPCRsort, as they also use the structural region lengths. Only loop lengths are used in the BTP method. Inoue *et al.* report the accuracy of the BTP method on training data only, which overestimate the actual accuracy of the method. The BTP method also makes use of fixed thresholds in the classifier which may lead to poor generalization performance. In the BTP method, the loop lengths are marked as short or long loops and these binary values are used in the calculations. However, binarization of loop lengths is not needed and region lengths are directly used in GPCRsort, as described in the next section.

2.3 Materials and Methods

Problem definition

Given a GPCR sequence, predict its class in a given classification scheme and a classification level.

2.3.1 GPCR representation

A GPCR representation can be seen in Equation (2.1).

$$GPCR = (FV, X) \tag{2.1}$$

where, FV is the feature vector, X is the class id

The feature vector of a protein is constructed using the structure of a GPCR. Representation of a GPCR can be seen in Figure 2.1. A feature vector is defined as a 15-dimensional vector as shown in Equation (2.2).

$$FV = [TM_1, TM_2, TM_3, TM_4, TM_5, TM_6, TM_7, N, L_1, L_2, L_3, L_4, L_5, L_6, C] \tag{2.2}$$

where TM_{1-7} : TM region lengths

N : N-terminus length

L_{1-6} : Loop lengths

C : C-terminus length

The length of a region is described as the number of amino-acids that comprise the respective region in Equation (2.2). The sum of entries in the feature vector gives the length of the GPCR represented by that vector.

GPCRDB [102] has a hierarchy of classes and defines a class id for each family in the hierarchy. As mentioned earlier, the class hierarchy starts with 5 top-most classes.

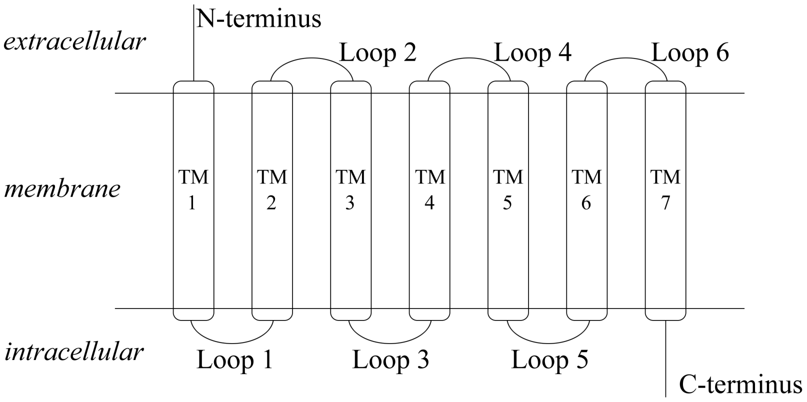


Figure 2.1: **Schematic diagram of a GPCR.**

Families are further divided into subclasses and sub-subclasses. This division is done based upon the function of the GPCR and the ligand that it binds. X, the class id in the GPCR representation in Equation (2.1), denotes the class id of the lowermost GPCRDB subfamily in which the protein is classified.

2.3.2 Dataset preparation

The protein sequences that comprise the datasets used in the methods are taken from the GPCRDB [102]. GPCRDB is a molecular-class information system that contains large amounts of heterogeneous data on GPCRs. The proteins in the GPCRDB are collected by mining for GPCRs from NR database that is compiled by the NCBI (National Center for Biotechnology Information). Hidden Markov Models are used to classify the proteins in this database. It currently contains 38525 proteins that are classified across 1272 families. The protein family members and class descriptions are easily reachable through the web site in [31].

Transmembrane regions of GPCRs are predicted using TMHMM stand-alone software package [47]. This program is for prediction of transmembrane helices in proteins. It uses a hidden Markov model to predict these regions [87]. TMHMM is selected because it has been rated best in an independent comparison of programs for prediction of transmembrane helices [57].

Transmembrane regions of all GPCRDB proteins are predicted using the TMHMM program. After the prediction process, the results showed that, 29038 proteins were labeled as having seven transmembrane regions. The feature vectors are constructed for those 29038 GPCRs and their representations are used in the following experiments. It can be seen that this number is much higher than the dataset sizes that are used in the previous studies. For instance, the GPCRpred dataset, which is used in studies GPCRpred [3] and GPCRBind [12], contains 1054 entries. As far as we know, the biggest dataset used to train and test the method is the GDS dataset, proposed by Davies *et al.* in [19], containing 8354 GPCRs. On the other hand, GPCRDB contains 38525 proteins. The dataset, used in this article, includes nearly 75% of the whole GPCRDB proteins, which shows that an up-to-date dataset is used to train and test the proposed classifier, GPCRsort. The remaining 25% of the GPCRDB proteins are either fragments or proteins which do not contain seven transmembrane regions as identified by TMHMM.

A non-redundant version of the dataset is also created. This version of the dataset is used to investigate the level of accuracy bias over certain GPCR families. Sequence redundancy is removed by intersecting the dataset with the UniRef90 database [89] which is maintained at 90% non-redundancy level. This intersection comprises the second dataset with 10216 entries. The size of this non-redundant set is still larger than the dataset sizes used in the previous works.

2.3.3 Method

Let P be the GPCR whose class is unknown. The steps to predict the class of P by the proposed method are as follows:

1. Let T be the training set consisting of GPCRs whose families (classes) are known.
2. Predict transmembrane regions of all GPCRs in T using the TMHMM tool. If a GPCR is marked as not having seven transmembrane regions by the tool, remove it from T.

3. Construct the feature vectors *FV* for each GPCR as mentioned in the Problem Definition. So *T'* will be:

$$T' = \{G : G \in T, and\, G = (FV, X)\} \tag{2.3}$$

where, X is the class of G

4. Train the Random Forest [5] classifier using *T'* and construct the classification model *M*.

5. Predict TM regions of *P* using the TMHMM tool and construct its FV_P.

6. Using the model *M* obtained in the 4^{th} step, let Random Forests determine the class of *P* using FV_P.

Random Forest classifier is used as the classification method. It is a classifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees. This method is the combination of the Breiman's 'bagging' idea [4] and the random selection of features [34]. By the power of this combination, it constructs a collection of decision trees with controlled variation.

Random Forest is chosen because of its advantages over the other classification methods. Firstly, it is one of the most accurate learning algorithms available [9]. It has methods for balancing error in class population unbalanced datasets [5]. This property is important because, the GPCRs in the constructed dataset are assigned to the classes in an unbalanced way. The method is also computationally efficient and runs very fast.

2.3.4 Environment of experiments

The experiments were performed in a PC with Intel Core i5 3.33 GHz CPU, 3 GBs of memory and 32-bit Windows 7 operating system. Weka 3 data mining software [33] is used for the construction of the classification model and the prediction of the classes using its built-in Random Forest algorithm.

Weka is a collection of machine learning algorithms for data mining tasks. It contains tools for data pre-processing and classification. In the pre-processing step, it takes the training dataset and makes analysis on the dataset. In the classification step, it presents many classifiers to choose from and constructs the classification model according to the selected classifier using the dataset given in the pre-processing step. It also presents test options for the verification of the constructed model. Random Forest classifier, with default options, is selected for the results presented in this study.

2.3.5 Feature vector contents

The idea of using the lengths of each region in the feature vector is a simple yet effective method. The structure of a GPCR contains the distinguishing properties of itself. Having seven transmembrane regions and also loops that connect these regions, brought the idea of taking benefit from this structure.

Each transmembrane region length is approximately the same, about 20-27 amino-acids. Using only these lengths could not easily distinguish proteins because of the shortness of the length of this region. On the other hand, loop and two termini lengths vary from 3 to 500 amino-acids. These lengths serve as distinguishing features for classification purposes. Besides the length, the locations of the loops, intracellular and extracellular, also play a critical role in the function of GPCRs.

An experiment is done to determine the regions which will be used in the feature vector. Seven different feature vectors are created for comparison, contents of which are: only transmembrane regions, extracellular loops, intracellular loops, transmembrane regions and extracellular loops, transmembrane regions and intracellular loops, all loops and two termini, all lengths. 10-fold cross validation experiments are performed using the created dataset for each feature vector construction. All region lengths are chosen to be in the content of a feature vector.

2.3.6 Cross validation experiments

The test dataset is constructed from the available dataset. Firstly, *k*-fold cross validation is applied. In this method, the whole dataset is partitioned into *k* equal size subsets. From these *k* subsets, *k-1* subsets are used as the training set and the remaining one set is used as the validation set for testing. This process is repeated *k* times, where each of the subsets is used once as validation data. The average of the *k* experiments is reported as the result. In this experiment, *k* is chosen as 10. The same experiment is repeated using the non-redundant version of the dataset.

2.3.7 Using an independent dataset as validation data

The performance of the method is measured using a different testing dataset as the validation data. GPCRpred dataset [3] was chosen as the test dataset. GPCRpred dataset contains subclasses of Class A GPCRs. There are total of 1054 proteins in this dataset. After the TM prediction step for the GPCRs in the set, 885 proteins are correctly classified as having a valid GPCR model by TMHMM. The proteins that exist in GPCRpred are removed from the training dataset. After the removal of GPCRpred proteins, training set contains 28204 proteins.

2.3.8 Comparison with the BTP method

To compare GPCRsort with the BTP method, the same training and testing datasets have to be used in the experiments. The dataset used in the calculations described in the BTP method article could not be obtained. Thus, the non-redundant dataset, with some modifications, is used for the results reported in this section. The classes in non-redundant dataset are reorganized according to the classes described in the BTP method dataset. BTP method employs HMMTOP [98] to extract loop lengths from GPCRs. To make the exact implementation of the method, same method, HMMTOP, is used to determine loop lengths of GPCRs in the non-redundant dataset. Total dataset size decreased to 9315 after HMMTOP TM region predictions. The BTP method is implemented exactly as described in the article.

2.3.9 Comparison with other methods

A new GPCR classification method is proposed in this chapter. There are existing GPCR classification methods in the literature. GPCRsort's classification performance should be compared with the ones in the literature. A perfect comparison could be done using the same training and testing datasets for the compared methods.

Cobanoglu *et al.* proposes a method, named GPCRBind [12], for the GPCR classification problem and compares this method with state-of-the-art GPCR classification methods reported by Davies *et al.* [19]. These methods use the GDS dataset [19] as training and the GPCRpred dataset [3] as testing datasets. To make a perfect comparison, exactly same datasets are obtained and used in the proposed method to compare its performance to the other methods' performances.

2.4 Results

Following sub-sections explain several performance scenarios of the method using different training and testing datasets. Generally, previous studies measure their classification performance on the subfamilies of the top-most families. We adopt a similar evaluation setting and report classification results for the second level of the GPCR class hierarchy. There are totally 75 classes as the subfamilies.

2.4.1 Effect of feature vector contents

Accuracy results for the seven setups that are prepared for the experiment done for the determination of feature vector contents are given in Table 2.1. About 47% of dataset entries could be correctly predicted using only the transmembrane region lengths. Loop lengths seem to be more effective than transmembrane region lengths. Additional region lengths integrated into the feature vector improve the performance of the predictor. Best accuracy is achieved with the use of all 15 lengths. The accuracy when using only the loop lengths is very close to the best accuracy. If the running time of the algorithm was important, only these 8 lengths could be chosen, but this is not a consideration for our algorithm. These results guide us to include all transmembrane, loop, N-terminus and C-terminus lengths in the feature vector.

Table 2.1: **Results of using different feature vectors**

Attribute type	**Correctly classified instances**
All [15] (TM_{1-7} + L_{1-6} + N + C)	26290 (90.54%)
Loops [8] (L_{1-6} + N + C)	25911 (89.23%)
TM regions [7] (TM_{1-7})	13688 (47.14%)
Extracellular loops [4] ($L_{2,4,6}$ + N)	22249 (76.62%)
Intracellular loops [4] ($L_{1,3,5}$ + C)	22675 (78.09%)
TMs + Extra. loops [11] (TM_{1-7} + $L_{2,4,6}$ + N)	24287 (83.64%)
TMs + Intra. loops [11] (TM_{1-7} + $L_{1,3,5}$ + C)	24413 (84.07%)

2.4.2 Cross validation experiments

Tables 2.2, 2.3 and 2.4 contain performance measures. Before going into details of the results of experiments, it would be better to define these measures. Recall (or sensitivity, corresponding to true positive rate) is the measure of the ability of GPCRsort to select instances of a certain class. Precision (or positive predictive value) is the measure of the accuracy provided that a specific class has been predicted. Fall-out (or false positive rate) value for a class is the real negatives that occur as predicted in that class. F-measure is a derived effectiveness measurement and interpreted as a weighted average of precision and recall. The area under the receiver operating characteristic (ROC) curve represents the probability that GPCRsort ranks a randomly chosen positive instance higher than a randomly chosen negative one. Recall, precision, fall-out and F-measures listed in tables are the weighted average of values for each class. The focus in the evaluations is on how confident one can be in the classifier [72].

At first step in the 10-fold cross validation studies, whole dataset is used to construct training and testing datasets. The results are presented in Table 2.2. Results show that

Table 2.2: **Evaluation results of 10-fold cross validation**

Measurements	**Values**
Correctly classified instances	90.54%
Recall*	0.905
Precision*	0.904
FP rate*	0.014
F-Measure*	0.903
ROC Area	0.983

* Weighted average of values for each class

Table 2.3: **Evaluation results of 10-fold cross validation using non-redundant dataset**

Measurements	**Values**
Correctly classified instances	80.43%
Recall*	0.804
Precision*	0.798
FP rate*	0.034
F-Measure*	0.796
ROC Area	0.953

* Weighted average of values for each class

the accuracy is very high as 90.54%. High true positive rate and low false positive rate can be easily seen in this table. In the second step of cross validation studies, the non-redundant version of the dataset is used to construct training and testing sets. Table 2.3 lists the results of this second step. High accuracy value in this table, which is 80.43%, shows that the accuracy is only affected by 10% compared to the value in the first step that uses the whole dataset. In a similar way, recall, precision and fall-out values are affected slightly. These results remove the question about biasing of the results to certain families because of the sequence redundancy in the first dataset.

2.4.3 Using an independent dataset as validation data

In the second experiment, the testing dataset is an independent dataset. The results of the classification done by GPCRsort are listed in Tables 2.4 and 2.5. The confusion matrix of the method using GPCRpred dataset as the testing data can be seen

Table 2.4: **Evaluation results of using a separate testing data**

Measurements	**Values**
Correctly classified instances	94.92%
Recall*	0.949
Precision*	0.95
FP rate*	0.013
F-Measure*	0.946
ROC Area	0.982

* Weighted average of values for each class

Table 2.5: **Classification performance of the method using GPCRpred dataset as the validation data**

Subfamily	**Total**	**Predicted**
Amine (AMN)	208	204 (98.1%)
Peptide (PEP)	305	301 (98.7%)
Cannabinoid (CAN)	11	11 (100%)
Gonadotrophin-releasing hormone (GRH)	9	9 (100%)
Hormone protein (HMP)	24	24 (100%)
Nucleotide-like (NUC)	30	29 (96.7%)
Lysosphingolipid and LPA (LYS)	8	8 (100%)
Melatonin (MEL)	13	11 (84.6%)
Olfactory (OLF)	69	68 (98.6%)
Platelet activating factor (PAF)	4	1 (25%)
Prostanoid (PRS)	8	6 (75%)
Rhodopsin (RHD)	174	163 (93.7%)
Thyrotropin-releasing hormone (TRH)	7	0 (0%)
Viral (VIR)	12	2 (16.7%)
Leukotriene B4 receptor (LEU)	3	3 (100%)
Total	**885**	**840 (94.9%)**

in Table 2.6. Table 2.4 shows that the performance of the method is very high, as the total accuracy is 94.9%. When the confusion matrix is analyzed, the problems with the prediction of Thyrotropin-releasing hormone, Platelet activating factor and Viral families stand out. The reason of the inefficient classification of these classes is the small number of those class entries in the training dataset. Another reason can be remarked as the creation time of the GPCRpred dataset. Classes in GPCRDB are reorganized several times up to date. Most of the other families are predicted with or near to 100% accuracies.

2.4.4 Comparison with the BTP method

The performance of GPCRsort with several datasets is analyzed in the first two experiments. The next experiments aim the comparison of GPCRsort with the existing GPCR classification methods. First comparison is with a similar method to GPCRsort: BTP method. Table 2.7 contains the results of this experiment. This table clearly shows that GPCRsort outperforms the BTP method. In total, GPCRsort predicts 85.2% of instances correctly, whereas BTP method predicts only 47.6%. Fur-

Table 2.6: **Confusion matrix of the method using GPCRpred dataset as the validation data**

		Predicted class														
		AMN	PEP	CAN	GRH	HMP	NUC	LYS	MEL	OLF	PAF	PRS	RHD	TRH	VIR	LEU
Actual class	AMN	204	3	0	0	0	0	0	0	0	0	0	0	1	0	0
	PEP	1	301	0	0	0	0	1	0	1	0	0	1	0	0	0
	CAN	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0
	GRH	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0
	HMP	0	0	0	0	24	0	0	0	0	0	0	0	0	0	0
	NUC	0	1	0	0	0	29	0	0	0	0	0	0	0	0	0
	LYS	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0
	MEL	0	1	0	0	0	0	0	11	0	0	0	0	0	0	0
	OLF	0	0	0	0	0	0	0	0	68	0	0	0	0	0	0
	PAF	0	2	0	0	0	0	0	0	0	1	0	0	0	1	0
	PRS	0	0	0	0	0	1	0	0	1	0	6	0	0	0	0
	RHD	0	3	0	0	0	0	0	0	1	0	0	163	7	0	0
	TRH	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0
	VIR	0	7	0	0	0	2	0	0	1	0	0	0	0	2	0
	LEU	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3

Table 2.7: **Classification performance of GPCRsort and BTP method**

Subfamily	**Total**	**GPCRsort**	**BTP**
A1	1197	972 (81.2%)	246 (20.6%)
IL-8R	20	10 (50%)	0 (0%)
Chemokine/Chemokine-like	255	185 (72.5%)	168 (65.9%)
A2	2772	2447 (88.3%)	1160 (41.8%)
Hormone	385	337 (87.5%)	98 (25.5%)
Olfactory	2510	2470 (98.4%)	2172 (86.5%)
Nucleotide-like	360	170 (47.2%)	38 (10.6%)
PAF	24	7 (29.2%)	0 (0%)
GRH	115	77 (67%)	0 (0%)
LLPA	100	67 (67%)	63 (63%)
Class A unclassified	636	405 (63.7%)	318 (50%)
B1	141	114 (80.9%)	57 (40.4%)
B2	23	15 (65.2%)	11 (47.8%)
Metabotropic glutamate	61	34 (55.7%)	45 (73.8%)
Ext. calcium-sensing	9	5 (55.6%)	0 (0%)
GABA-B	46	39 (84.8%)	0 (0%)
Class A unclassified	280	255 (91.1%)	56 (20%)
Frizzled/Smoothened	381	330 (86.6%)	0 (0%)
Total	**9315**	**7939 (85.2%)**	**4432 (47.6%)**

Class definitions are taken from [37]

thermore, the BTP method cannot predict even a single member correctly in some families, i.e. GRH or GABA-B. One reason for the bad performance of the BTP method is its overfitting of the training dataset and its poor generalization performance.

2.4.5 Comparison with other methods

We compare GPCRsort with a state of the art classifier, GPCRBind, which has been shown to outperform several other classifiers [12]. The accuracy results of this comparison are listed in Table 2.8. GPCRsort gives the highest accuracy among these classifiers. 97.3% accuracy shows how GPCRsort improves the classification performance on GPCRs. It would be also good to see the accuracy results of these classifiers with the created dataset here. However, the oldness, limited access and low flexibility of the other classifiers prevent us to create a comparison environment like that.

Table 2.8: **Classification performance of GPCRsort compared with the current known methods**

Classifier	Accuracy
GPCRsort	97.3%
GPCRBind	90.7%
GPCRTree	76.2%
PRED-GPCR	73.8%
GPCRpred	67.1%

2.4.6 Running Time Analysis

Running time of the method is the sum of running times of the steps of the method. TMHMM runs in seconds to find the transmembrane regions. In the experiments, classifier model construction took only a few seconds. Determination of the class of an unknown GPCR takes milliseconds. The whole method does not take more than a minute. This makes the method a practical method. Another important property of the method is; it can be run on any PC, not requiring a server to run. Compared to GPCRBind, this is a significant improvement in running time, since GPCRBind needs hours to construct the sequence motifs for GPCR class prediction [12]. Figure 2.2 shows the comparison of running times of GPCRBind and GPCRsort on the calculations of the experiment described in the ‘Comparison with other methods’ section.

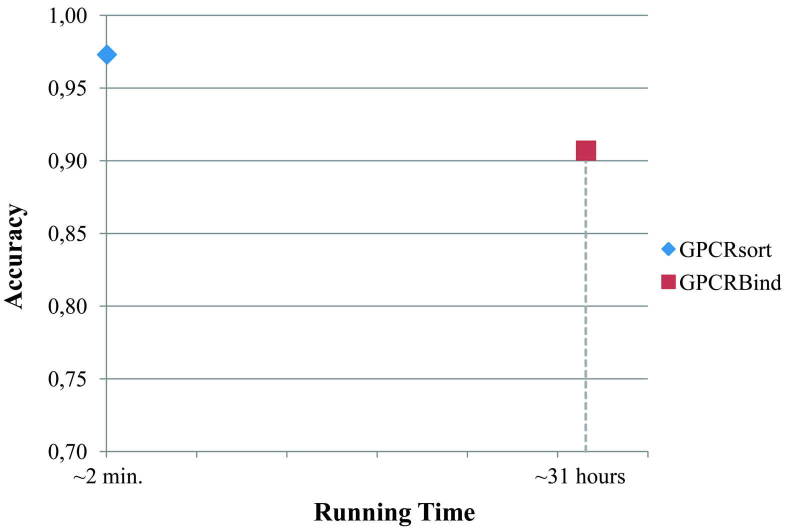


Figure 2.2: **Comparison of running times of methods.** GPCRBind running time is taken from [12].

2.5 Discussion

A new GPCR classification method is proposed, GPCRsort, which is simple but effective in accurately classifying GPCRs into correct GPCRDB classes. In fact, GPCRDB itself contains predictions: GPCR sequences in GPCRDB are selected by classifying them against a database of HMMs. These HMMs are created from the previous release of GPCRDB [102]. This observation raises an issue here that a predictor is tested against another predictor. However, GPCRDB is used as gold standard in most of the classification methods in the literature. The work done in this study is to compare GPCRsort with the other methods. This is the reason for using GPCRDB as the benchmark in this chapter.

GPCRsort can be used to classify uncharacterized GPCRs and direct further biological studies accordingly. Using the structural lengths of the GPCR substructures is a very simple idea. Similar proteins preserve the lengths of the same structural regions because of the evolutionary development of the genes [69]. Receptors that make contact with similar ligands are evolved from the same common ancestors. Therefore, the substructure region lengths remained similar. It is possible to find out how close these receptors to each other just looking at these lengths. Our experiments show that, despite its simplicity, the lengths of a GPCR's substructures is very powerful as a discriminator of GPCR classes and a Random Forest classifier based on this feature is able to significantly outperform more elaborate sequence pattern based approaches.

With GPCRsort, it is possible to characterize orphan GPCRs and conduct directed biological experiments to validate the ligands of these novel GPCRs; hence, reducing the time for related drug studies significantly. The accuracy of GPCRsort is very close to perfect except for the viral, thyrotropin-releasing hormone, and platelet activating factor receptors. The small number of those class entries in the training sample is the basic reason for their incorrect classification. Challenges related to the classification of these classes of GPCRs can be investigated as future work and the overall accuracy can be further improved.

GPCRsort shines among other approaches in the comparison experiments. GPCRsort outperforms the BTP method, predicting 85.2% of instances correctly, where only

47.6% of instances could be predicted correctly by BTP method. Having 97.3% accuracy shows the power of this method when compared to other methods under the same conditions. In addition to this power, being able to classify each single class demonstrates the generality of this classifier. Moreover, rapid running time of the method makes it easily testable when a GPCR classification is necessary.

CHAPTER 3

INTGPCR: DATABASE OF GPCR-GPCR INTERACTIONS

The increase in the number of studies on the GPCR dimerization brings a necessity of a collective database to be in the service for the researchers. The purpose of the creation of the IntGPCR is to fill that necessity for the community. IntGPCR is a database of interacting GPCRs. The contents of this database are curated from the literature. The biological and computational studies on dimerization of GPCRs are collected and carefully analyzed for any dimerization data. Giving particular importance to the up-to-dateness of the database, every valuable information, to the smallest one, from the articles are gathered. This chapter presents the steps of creation and the presentation of the IntGPCR.

3.1 Background

Chemistry defines an oligomer as a molecular complex that consists of a few monomer units. In biology, a protein oligomer is a macromolecular complex that consists of a number of protein monomers. Dimers are oligomers composed of two monomers. The two proteins are joined by either strong or weak bonds. When the two proteins are identical, the complex is named as homodimer. Heterodimers are the protein complexes in which the two proteins are not identical. Protein dimerization is the process of conversion of two proteins to a dimer formation [38].

Throughout the 1970s and 1980s, there are studies that propose G protein-coupled receptors (GPCRs) could exist as dimers or higher order oligomers [81]. However, until the end of 1990s, GPCRs are thought to be monomeric units whose main functions are interacting with G-proteins after ligand activation [1]. The number of studies about oligomerization, especially dimerization, of GPCRs is increased after this period. Now, it is accepted that GPCRs could exist as monomeric, dimeric or oligomeric complexes in the cell.

The question of why GPCRs dimerize is the subject of several studies. Researchers define the role of this dimerization in different ways. One reason for the dimerization is: trafficking of the receptor from endoplasmic reticulum to the cell surface. For instance, $\mathrm{GABA_{B1}}$ receptor constructs a heterodimer with $\mathrm{GABA_{B2}}$ receptor for the targeting to the cell membrane [65]. Another example for the trafficking is the homodimerization of $\beta 2$ adrenergic receptor [82]. Receptor activation is another role of the dimerization [14]. Again $\mathrm{GABA_{B1}}$ and $\mathrm{GABA_{B2}}$ receptors can be shown as example. $\mathrm{GABA_{B1}}$ unit is responsible for ligand activation, where $\mathrm{GABA_{B2}}$ activates G-proteins [43, 22].

The mechanism of the GPCR dimerization is also in interests of the researchers. There are several experimental studies in the literature for the determination of the mechanism of the dimers. These experiments use generally biological methods. In addition to biological experiments, there exist computational studies too. The details of these studies are mentioned in the next chapter. This chapter contains the information of how those studies are analyzed and how the data is obtained from the literature.

3.2 Related Work

The need for a database of interacting GPCRs attracted attention of some researchers. There exist three published databases that contains data about GPCR oligomerization. These are gpDB [96], GRIPDB [62] and GPCR-OKB [44].

GpDB is a database of GPCRs, G-proteins, effectors and their interactions [96]. The system contains data about coupling specificity of GPCRs to their respective G-proteins and also dimerization information of GPCRs. GpDB was last updated on March 2008. The database is primarily focused on information about interactions between GPCRs and their partner G-proteins. GPCR dimerization data remains in the background in gpDB. This data is not easily accessible by the users, because the data is listed only on the corresponding GPCR family member entries [97]. Any search cannot be made to list the interacting GPCRs. Also the database does not contain any interface details on the interacting GPCRs. Another deficiency of the database about GPCR dimers is that the specific interacting GPCR cannot be reached from the dimerization data. Only the type of the interacting GPCR is listed. After a difficult analysis of the database, navigating entry by entry, the interacting GPCRs' data statistics is obtained. GpDB contains 82 interacting GPCR entries that are gathered from 68 published articles from the literature.

G protein coupled Receptor Interaction Partners DataBase (GRIPDB) provides information about GPCR oligomerization [62]. The system is hosted from Japan. The system contains experimentally identified GPCR oligomers data from the literature. In addition to this data, the database also contains suggested interfaces of oligomerization based on the GRIP server [61] predictions. GRIP server predicts interfaces for oligomerization sites based on sequence alignment of the query GPCR and its homologs and a template structure [61]. This template structure is either rhodopsin or $\beta 2$ adrenergic receptor. The GRIPDB system is not easy to navigate and list the interacting GPCRs [60]. The system is last updated on January 2011. Besides that, containing some wrong information about interacting GPCRs makes this system not reliable. For instance, the system lists the homodimer of μ-opioid receptors from a study of George *et al.* [30]. However, this study only shows the oligomerization of μ-opioid and δ-opioid receptors. There are some other examples of errors in this database. After a long difficult period for the analysis of the system, the data statistics are: 112 interacting GPCRs that are curated from 107 published articles. Among these dimer data, there exist 17 interface details.

The third database is the G Protein Coupled Receptor Oligomer Knowledge Base (GPCR-OKB) [44]. This system contains GPCR oligomerization data derived from the literature. GPCR-OKB was last updated on November 2012. This system is the best between the three databases mentioned here, based on usability and content. However, this system also contains erroneous entries about interacting GPCRs. An example is: the entry of the heterodimer formation of human α-1b adrenergic and α-2a adrenergic receptors that are curated from the article of Xu *et al.* [104]. However, this article mentions heterodimerization of α-2a adrenergic and β-1 adrenergic receptors. Like this entry, there exist several erroneous entries. The system contains 192 interacting GPCRs information curated from 220 articles from the literature. Among these interacting GPCRs, there exist 35 interface details.

General analysis of these three systems shows that the biggest deficiency of these systems is their up-to-dateness. In addition to that, the erroneous data make the system unreliable. Besides these, a system with easy navigation and useful visualization of interacting GPCRs is needed. Also it is thought that there exist more articles in the literature that contain interacting GPCRs and their interfaces. These ideas lead to the development of IntGPCR, the Database of Interacting GPCRs. IntGPCR contains 309 interacting GPCRs, where 138 of them contain interface details, curated from 348 articles from the literature, and continuing its development to be up-to-date. The following sections describe the development of this system.

3.3 Materials and Methods

3.3.1 Preparation of the Article List

First and the very important step in the development of IntGPCR is the collection of sources that contain data about GPCR interaction. These sources are the published articles from the literature. Because the analysis of articles is a time-consuming study, the articles must be chosen carefully so that each of them should contain valuable information about GPCR dimerization. Valuable information can be anything about dimerization of GPCRs, including the GPCRs that take part in the dimerization, interface data that is a region or specific residues from GPCRs or any text stating a positive or negative information about the dimerization.

The work is started with a small list of articles, specifically 20 articles, which are known to include dimerization data. This list is used to obtain specific keywords to be used in the search parameters for getting the related article list from the literature. A simple text mining process, generating the word frequency table, is employed here to get the keywords. Abstract and introduction parts of these 20 articles are collected to be input. RapidMiner Studio program [35] is used for its text mining capabilities. RapidMiner Studio is a program useful for machine learning, data mining, text mining and predictive analytics topics [75]. The steps of the construction of word frequency table is listed below:

- Words from the 'Abstract' and 'Introduction' sections from each article are collected and uppercase letters are converted to the lowercase.
- English stop words (is, are, and, we, etc.) are removed from the word list.
- n-Grams are determined. n is set as maximum 4. In this way, it is possible to determine phrases that are constructed from up to four words (like g_protein_coupled_receptor).
- Words are grouped according to their stems.
- As the last step, frequencies of words of at least two in length are calculated.

The word frequency table suggested some useful words that can be used as characteristic keywords for the search parameters in the literature to get the articles which contain GPCR dimerization data. The constructed search string is shown in 3.1.

Seach string used in Pubmed

$$\boxed{'(gpcr \text{ AND } (oligomer \text{ OR } dimer \text{ OR } heterodimer \text{ OR } homodimer))'} \tag{3.1}$$

The search string in 3.1 is used in Pubmed [58] to list the articles that suit those keywords. Pubmed comprises a lot of citations for biomedical literature from life science journals, and online books. PubMed is a free resource that is developed and maintained by the National Center for Biotechnology Information (NCBI), at the U.S. National Library of Medicine (NLM), located at the National Institutes of Health (NIH). Most of the citations include links to full-text contents. Full texts of all articles found in the search results were downloaded if available. Finally there were 157 full-text articles waiting to be analyzed for the GPCR dimerization data.

3.3.2 Analysis of Articles

The aim of the analysis of the gathered articles is to extract information about GPCR dimerization if exists. The first thing that comes to mind is to automatize this process with some text mining methods for easy analysis. However, the structures of the dimerization data in the articles are not standardized. The data could be anything like interacting amino acids, or transmembrane or loop regions that make contact, or only a dimer proof between different GPCRs or same GPCRs. It is too difficult to distinguish these type of data in an unstructured full-text article with text mining methods. Because the aim of this study is to correctly get the interaction data, the article analysis is done manually. All downloaded articles are studied and the corresponding GPCR interaction data is obtained if available.

After the analysis, among the 157 articles mentioned in the previous section, 117 articles contain dimerization data. Furthermore, each article contains important references to other literature studies which contain GPCR dimerization information. Those references are also marked and downloaded if they are available. With those references, the total count of the studied articles became 541. Because of the time-consuming process of the analysis of these articles, no further references are followed. Among these 541 articles, 348 of them contain GPCR interaction information.

Each article is carefully studied to extract the correct data. The articles contain experiment results from biological studies as well as computational studies. Articles generally state the existence of a dimer and the GPCRs that are involved. Minority of the articles give specific interface information between the interacting GPCRs. The interface data are generally some transmembrane or loop regions that are involved in the interactions. The number of articles that contain information about specific residues that take part in the interaction is only 47.

3.3.3 Development of the Web Interface

To present the dimerization dataset created from the literature to the researchers in an informative and easy to use way, it was decided that a web interface would be the best option. For the development of the web interface, PHP [94] and MySQL [64] tools are employed. PHP is a popular general-purpose scripting language that is especially suited to web development. It is fast, flexible and practical. MySQL is a widely used [86] open-source relational database management system. It is a popular choice of database for use in web applications.

The organization of the data in the database management system is shown in Figure 3.1. Easy management and fast usage are considered when the structure of the database is designed. Each row in the *'interactions'* table contains an interaction's details. This interaction can be a homodimer or a heterodimer or an oligomer consisting of more than two monomers. If the interaction contains interface details, there are links to the *'interface'* table for the interface details. An interface entry can be any type according to the information it holds. This information could be any one of; region data, residue data or text describing the interface, or a combination of these listed data. Each interaction is the result of a biological or a computational experiment. The experiment details are linked to the *'experiment'* table. Each interaction is curated from an article from Pubmed [58]. The Pubmed Id and the publication year of the article is held in the *'pubmed_articles'* table. Some interactions can be visualized from the web interface, described in the next sections. The model file paths are held in the interactions row.

GPCRs are listed in the *'gpcr'* table. Minimum information about a GPCR are held in the table, the other details can be taken from the linked databases, which are GPCRDB [31] and Uniprot [95]. *'gpcrdb_id'* is the GPCRDB entry id and *'acc_code'* is the Uniprot entry id for these links in a *'gpcr'* entry. There is a *'family_id'* link from *'gpcr'* table to *'family'* table, connecting each GPCR to its belonging family. Again, there is minimum data about a GPCR family. The other data can be reached from the supplied link to the GPCRDB that is constructed from the *'family_id'* column of the *'family'* table. The species of a GPCR is kept in the *'species'* table. The names of a GPCR and species are kept in the *'short_name'* columns of the corresponding tables.

3.3.4 Modelling of GPCRs

GPCR modelling is an interesting topic in this study. As stated previously, it is not easy to determine three-dimensional structure of GPCRs in living cells with biologi-

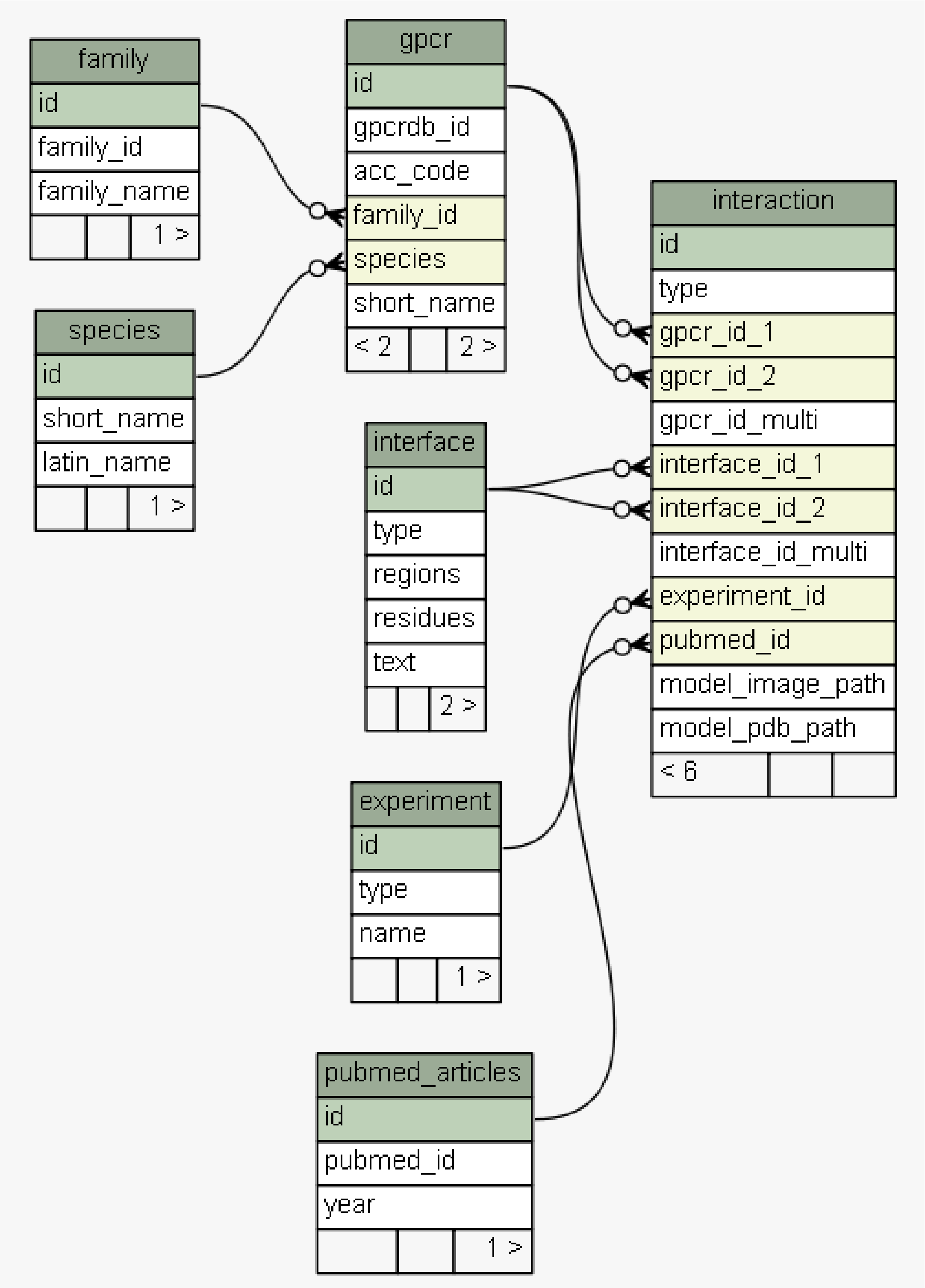


Figure 3.1: **Schema of the IntGPCR database system.** Picture is generated using SchemaSpy [16].

cal experimental methods like X-ray crystallography, NMR Spectroscopy or electron microscopy. Up to now, only 21 different types out of thousands of GPCR structures have been experimentally solved [105]. Here, computational methods help to predict the structures of GPCRs. GPCRs are modelled with the method described in the next section to be used in the IntGPCR and also interface prediction studies described in Chapter 4.

Homology modelling is employed for the prediction of the structure of GPCRs. Homology modelling is the process of construction of the atomic-resolution model of a protein from its amino acid sequence and an experimentally known three-dimensional structure of a related template homologous protein [46]. Here, the effect of the template protein on the reliability of the resulting model is important. It has been shown that protein structures are more conserved than protein sequences amongst homologues [11]. So if the template protein is a homologue of the unknown structured protein, than the resulting model can be accepted as reliable.

The proteins whose three-dimensional structures are determined experimentally are listed in Table A.1 in Appendix A. The table contains the PDB [76] ids of these proteins and also the families that they belong to. These proteins are used as template proteins in the homology modelling studies.

Two tools are used in the modelling process. First one is the MAFFT tool [42] which is a multiple sequence alignment program for protein sequences. The second one is the MODELLER [23] which is used for homology modelling of protein three-dimensional structures. Both tools run on unix-like operating systems. Because the process is very time-consuming, a workstation computer is chosen for the tools to run on. The experiments are performed on a workstation with Intel Xeon Processor E5, 64 GBs of memory and 64-bit Ubuntu Linux operating system.

3.3.4.1 Modelling Method

Let P be a GPCR whose structure is unknown. Proposed method for the modelling of P is comprised of the following steps:

1. Let A be the set of GPCRs whose three-dimensional structures are experimentally determined. The entries in this set are listed in Table A.1.
2. Determine the appropriate entries from set A to be used as template models in the process of homology modelling of P. This determination is based on the families of the GPCRs. So A' will be:

$$A' = \{S : S \in A, \textit{and S is a template for P}\} \tag{3.2}$$

3. Mark transmembrane regions of P and the elements of A': Change all residues of transmembrane regions; domain 1 to '1', domain 2 to '2', and so on.
4. Use MAFFT tool [42] to multiply align P and the GPCRs in A' using the marked sequences resulted from the previous step.
5. Change the residues in the transmembrane regions in the multiple alignment results to the corresponding original residues and prepare input file for MODELLER tool [23] from these aligned results.
6. Run MODELLER to predict the atomic-resolution model of P.

Determination of the template models in step 2 is based on the families of the proteins. Because of the increase of accuracy in homology modelling when using homologues template proteins, evolutionary analogous proteins are tried to be selected. The proteins are grouped according to their top-most families in Table A.1. For instance, 8 of the known structured GPCRs belongs to Class A \ Amine family. These 8 GPCRs are selected as template for the prediction of proteins that belong to Class A \ Amine family. For a GPCR from a Class A family, but not in any sub-family of Class A families listed in the table, all 23 known Class A family proteins are used as template. For example, to predict the model of a GPCR from Class A \ Melatonin family, all those 23 Class A proteins will be used as template. In Class B and Class C families, there exist only one known structured GPCR to be used as template. For the other GPCRs, all the proteins in the table are used. An analysis experiment is done for comparison of the use of specific proteins as describe here, and the use of all proteins as template. The details of the experiment is described in the following sections.

In steps 3 and 4, the multiple sequence alignment is done using a special alignment method: transmembrane alignment. This alignment is a specialized global sequence alignment, where the transmembrane regions are aligned in the same positions in the aligned sequences. The purpose of the alignment is to increase the reliability of the predicted models. GPCRs are special proteins whose properties are described in Chapter 1. Each GPCR contains similar seven transmembrane domains, where the terminals and loops that connect these transmembrane regions differs in length. The transmembrane domains take part inside the membrane. It can be seen that this specialty is reflected to the three-dimensional structure of the GPCRs. So here, transmembrane alignment is better than a global sequence alignment. There exist a comparison experiment about alignments in the next section.

In this method, the start and the end points of transmembrane regions and the families of GPCRs are taken from GPCRDB [31]. The models for all GPCRDB entries whose three-dimensional structures are unknown are predicted by this method. As a result, the models for 36390 GPCRs are created.

3.3.4.2 Alignment Method Decision

First experiment about the proposed homology modelling method is the determination of the alignment method in the multiple sequence alignment step. As stated in the previous section, proposed transmembrane alignment is employed in the method. To remove the questions about if using the global sequence alignment could give better results than using the transmembrane alignment, an experiment is designed for the comparison.

Only the proteins whose three-dimensional structures are known are used in this experiment. There are 25 known GPCRs that are listed in Table A.1. 22 of them, who have at least one template structure, are used in this experiment. One GPCR across the known ones is accepted as unknown structured GPCR, and the method is applied on that protein to predict its structure. Two predictions are made for this protein, where in one of them global sequence alignment is used, and in the other transmembrane alignment is used. At the end, there are two predicted structures and one structure that is determined experimentally. Those structures are compared to known structure and the closest to the real structure wins the comparison. The measurement in the comparison is the root-mean-square deviation, RMSD. RMSD is the measure of the average distance between the atoms of superimposed proteins [15]. Chimera tool [68], which has ability to easily calculate RMSD of the models, is used to calculate RMSD scores. This experiment is repeated for all 22 GPCRs and the alignment method is determined according to the results. The results of this experiment are listed in Table 3.1.

Table 3.1: **Comparison of using different alignment methods**

	Global Alignment		**TM Alignment**	
	cutoff: 2 Å	cutoff: 5 Å	cutoff: 2 Å	cutoff: 5 Å
3uon	0.995 (157)	1.416 (187)	0.724 (122)	1.147 (132)
4daj	0.745 (168)	1.364 (191)	0.912 (134)	1.792 (177)
4amj	0.996 (141)	1.700 (180)	0.856 (159)	1.471 (193)
2rh1	0.908 (124)	2.442 (211)	0.833 (146)	1.510 (178)
3pbl	1.083 (78)	2.432 (131)	0.982 (78)	2.697 (171)
3rze	1.123 (75)	2.380 (132)	0.944 (109)	2.314 (157)
4iar	0.936 (148)	1.913 (198)	1.119 (101)	2.177 (140)
4ib4	1.126 (100)	2.001 (150)	0.827 (31)	2.422 (70)
2lnl	1.348 (28)	3.401 (108)	1.403 (17)	3.565 (83)
4mbs	1.192 (24)	3.955 (74)	1.240 (15)	3.329 (66)
3odu	1.097 (178)	1.783 (233)	0.933 (113)	1.928 (168)
4grv	1.118 (78)	2.834 (194)	1.194 (70)	2.459 (117)
4ej4	0.902 (209)	1.381 (237)	0.703 (244)	1.037 (265)
4n6h	1.272 (52)	2.924 (139)	0.865 (181)	1.684 (227)
4djh	1.271 (30)	2.980 (70)	0.875 (169)	1.730 (210)
4dkl	1.123 (33)	2.514 (62)	0.834 (180)	1.637 (230)
4ea3	1.339 (18)	3.105 (53)	0.850 (175)	1.517 (212)
3vw7	1.290 (25)	3.044 (92)	1.047 (57)	2.503 (115)
1u19	1.310 (182)	1.814 (239)	1.212 (73)	2.663 (171)
3ayn	1.227 (199)	1.691 (245)	1.464 (128)	2.303 (197)
4eiy	1.019 (48)	2.341 (71)	1.305 (6)	2.348 (20)
4ntj	0.797 (45)	2.371 (65)	1.199 (15)	1.625 (19)

Before the analysis of the values in Table 3.1, this table needs a description. The results of the predictions that global sequence alignment is used are listed in the 2^{nd} and 3^{rd} columns. 4^{th} and 5^{th} columns contain the results in which transmembrane alignments are used. The RMSD calculations are made by the MatchMaker tool in the Chimera [68]. That tool has an option about whether to iteratively remove far-apart residue pairs from the *'match list'* used to superimpose the structures. It iterates by pruning long atom pairs until no atom pair exceeds x angstroms. The x is a parameter that can be given as input to the tool. In this experiment, two options are tried as the x value: 2 Å and 5 Å. 2^{nd} and 4^{th} columns list the RMSD values for the 2 Å case for each type of alignments and 3^{rd} and 5^{th} columns list the 5 Å case. Rows in the table contain the RMSD results for each GPCR that is taken for the prediction. RMSD values present in the cells are in the measurement of angstroms. There exist a value in parenthesis next to each RMSD value. That value denotes the number of atom pairs that exist in the *'match list'*. Higher the number of atom pairs and lower the RMSD value denote a good matching for the proteins that are in consideration. For each row, the winner alignment type is colored as gray. The table shows that transmembrane alignment could give better results.

3.3.4.3 Selection of Template Proteins

The next experiment about the proposed homology modelling method is the determination of the template proteins which will be used in the multiple alignment and modelling processes with the unknown GPCR. In the method, a nearest family based approach is employed. The results of using different selection of template proteins should be analyzed. For this purpose, an experiment environment is designed as in the same case described in the previous section.

Again the used proteins in this experiment are the proteins whose three-dimensional structures are known. One of them is taken as unknown protein and the modelling method is applied on that protein. In this case, three predictions are made with changing the selection method of the template proteins in each case. In the first one, the template proteins are selected as told in the method; the nearest subfamily members of the top-most family are selected. In the second case, the nearest top-most family members are selected. In the third one, all the available proteins are selected. For instance, for the human muscarinic acetylcholine receptor (pdb Id: 3uon), selected template proteins are as follows: For type I case: members from the Class A \ Amine family, for type II case: members of the Class A family and for type III case: all the known proteins. There are three predicted models for this experiment. As in the previous section, these predicted models are compared with the known structure of the protein and RMSD values are calculated. The results are shown in Table 3.2.

Table 3.2: **Comparison of using different template known models**

	Type I		**Type II**	**Type III**	
	cutoff: 2 Å	cutoff: 5 Å	cutoff: 5 Å	cutoff: 2 Å	cutoff: 5 Å
3uon	0.724 (122)	1.147 (132)	3.318 (68)	1.074 (54)	2.260 (90)
4daj	0.912 (134)	1.792 (177)	3.401 (63)	1.192 (36)	2.859 (85)
4amj	0.856 (159)	1.471 (193)	3.072 (64)	1.414 (15)	3.541 (59)
2rh1	0.833 (146)	1.510 (178)	2.849 (105)	1.049 (102)	2.113 (152)
3pbl	0.982 (78)	2.697 (171)	3.192 (70)	1.160 (46)	2.381 (78)
3rze	0.944 (109)	2.314 (157)	3.186 (72)	1.556 (19)	3.165 (67)
4iar	1.119 (101)	2.177 (140)	2.686 (76)	1.349 (39)	2.963 (115)
4ib4	0.827 (31)	2.422 (70)	3.461 (29)	1.399 (11)	3.620 (41)
2lnl	1.403 (17)	3.565 (83)	3.123 (57)	1.423 (13)	3.299 (42)
4mbs	1.240 (15)	3.329 (66)	3.025 (93)	1.210 (6)	3.286 (38)
3odu	0.933 (113)	1.928 (168)	2.934 (94)	1.279 (36)	3.002 (112)
4grv	1.194 (70)	2.459 (117)	3.388 (85)	1.212 (18)	3.088 (68)
4ej4	0.703 (244)	1.037 (265)	2.948 (171)	0.719 (242)	1.137 (263)
4n6h	0.865 (181)	1.684 (227)	2.670 (113)	1.001 (176)	1.745 (225)
4djh	0.875 (169)	1.730 (210)	3.472 (57)	0.953 (169)	1.733 (209)
4dkl	0.834 (180)	1.637 (230)	3.119 (64)	0.847 (212)	1.405 (242)
4ea3	0.850 (175)	1.517 (212)	3.440 (51)	0.985 (155)	1.773 (203)
3vw7	1.047 (57)	2.503 (115)	3.554 (64)	1.195 (25)	2.974 (61)
1u19	1.212 (73)	2.663 (171)	3.333 (41)	1.489 (13)	3.533 (51)
3ayn	1.464 (128)	2.303 (197)	3.305 (58)	1.159 (42)	2.596 (87)
4eiy	1.305 (6)	2.348 (20)	3.235 (63)	1.677 (7)	3.404 (19)
4ntj	1.199 (15)	1.625 (19)	3.081 (23)	1.400 (8)	3.471 (31)

The structure of the Table 3.2 is same as the Table 3.1 that is described in the previous section. For the location requirements, *cutoff: 2 Å* column for the Type II experiment is removed. For each protein, best predicted case is colored as gray. In most cases, the results of the experiment Type I give better results than the other cases. In Type I case, the template proteins are selected as the nearest subfamily members of the top-most classes. These results direct the use of Type I type selection of template GPCRs.

3.3.5 Visualization of Dimer Formations

A specific property of the IntGPCR database is its ability of visualization of the interacting GPCRs. This is achieved with the help of the three-dimensional structures of the GPCRs that are involved in the interactions as well as the interface information between them. The environment of the IntGPCR is suitable for this visualization. All three-dimensional structures are in-hand from the modelling process and the crystallized proteins. Furthermore, the existence interface information in the literature is included in the database. There remains one more step to visualize them in the user interface.

For the demonstration of the interacting GPCRs, a docking program is employed. Protein-protein docking is the computational modelling of the quaternary structure of protein complexes formed by two or more interacting proteins. In the scope of IntGPCR, these complexes are GPCR homodimers or heterodimers. The aim is the prediction of the three-dimensional structure of the dimer as it will occur in a living organism. There are several methods for protein-protein docking published in the literature, Monte Carlo simulations, reciprocal space methods, etc. There are also tools and servers for the docking procedure. ClusPro 2.0 [13] server is one of these programs. It is a protein-protein docking tool in which two models are given as input and the program outputs the predicted three-dimensional structures of the resulting dimer. The program also take attraction and repulsion residues to direct the resulting model. This server is used for creation of the models of dimer formations. The interface information in the database is used as input in the attraction residues field of the server.

The resulting atomic-resolution models from ClusPro server should be presented to the users. This will help the researchers who look for the details of an interaction for an initial opinion. JSmol [40] is used for this purpose. JSmol is an open-source Java viewer for chemical structures in 3D. It is used as an applet in the web interface for each interaction that the model is available. It is adjusted to initially show the *'Ribbons'* style of structures for the presented model, in which two models are colored differently for easy analysis. There are more options in the applet menu for the analysis of the models. The best created model from the ClusPro server is shown in the web applet. The other created models for the dimers can be downloaded by the researchers via the web interface for locally analysis.

3.4 Presentation of the IntGPCR

IntGPCR is presented in the web page (http://bioserver.ceng.metu.edu.tr/IntGPCR). IntGPCR system provides browsing, searching and visualizing the dimer formations. The entrance page welcomes the users and gives a brief information about the system. The menu on the left of each page provides easy navigation through pages. *Help*, *About* and *Contact* pages contain related information about IntGPCR usage, the team members' list and contact information respectively.

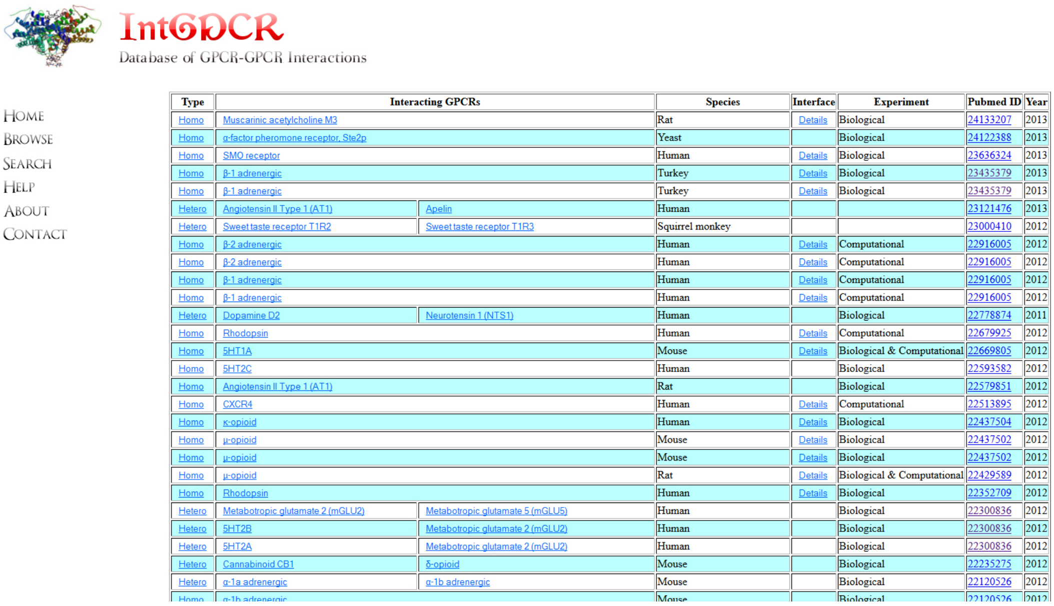
The database of interacting GPCRs can be accessed by browsing the whole list or by searching for specific GPCRS using the two links in the navigation menu: *Browse* and *Search*. *Browse* page provides a list of the interactions. A sample browse page is provided in Figure 3.2. The list consists of entries that contain a brief information about the interactions. The column definitions are listed below:

- **Type:** Type of the interaction. Can be either Homodimer or Heterodimer. Clicking on this field opens the *Interaction Details* page of that specific interaction.
- **Interacting GPCRs:** The GPCR short names that are involved in the interaction. If the interaction type is Homodimer, this field contains only one GPCR name since there is only one type of GPCR in the interaction. If the interaction type is Heterodimer, this field contains two or more GPCR names. The GPCR names are clickable and provide a link for the *GPCR Details* page for that specific GPCR.
- **Species:** Name of the species that the interacting GPCRs belong to.
- **Interface:** If the interaction contains any interface information, this field is written as 'Details' and can be clicked to list the interaction details as in the first column 'Type'. If this field is empty, that means the interaction do not contain interface information.

- **Experiment:** The type of the experiment studied for this interaction. In can be 'Biological', 'Computational' or both. The experiment details are listed in the *Interaction Details* page.

- **Pubmed ID:** The related Pubmed Id of the published article that provides the interaction. Clicking on this field will direct to related article in Pubmed.

- **Year:** Publication year of the article.



Type	Interacting GPCRs		Species	Interface	Experiment	Pubmed ID	Year
Homo	Muscarinic acetylcholine M3		Rat	Details	Biological	24133207	2013
Homo	α-factor pheromone receptor, Ste2p		Yeast		Biological	24122388	2013
Homo	SMO receptor		Human	Details	Biological	23636324	2013
Homo	β-1 adrenergic		Turkey	Details	Biological	23435379	2013
Homo	β-1 adrenergic		Turkey	Details	Biological	23435379	2013
Hetero	Angiotensin II Type 1 (AT1)	Apelin	Human			23121476	2013
Hetero	Sweet taste receptor T1R2	Sweet taste receptor T1R3	Squirrel monkey			23000410	2012
Homo	β-2 adrenergic		Human	Details	Computational	22916005	2012
Homo	β-2 adrenergic		Human	Details	Computational	22916005	2012
Homo	β-1 adrenergic		Human	Details	Computational	22916005	2012
Homo	β-1 adrenergic		Human	Details	Computational	22916005	2012
Hetero	Dopamine D2	Neurotensin 1 (NTS1)	Human		Biological	22778874	2011
Homo	Rhodopsin		Human	Details	Computational	22679925	2012
Homo	5HT1A		Mouse	Details	Biological & Computational	22669805	2012
Homo	5HT2C		Human		Biological	22593582	2012
Homo	Angiotensin II Type 1 (AT1)		Rat		Biological	22579851	2012
Homo	CXCR4		Human	Details	Computational	22513895	2012
Homo	κ-opioid		Human	Details	Biological	22437504	2012
Homo	μ-opioid		Mouse	Details	Biological	22437502	2012
Homo	μ-opioid		Mouse	Details	Biological	22437502	2012
Homo	μ-opioid		Rat	Details	Biological & Computational	22429589	2012
Homo	Rhodopsin		Human	Details	Biological	22352709	2012
Hetero	Metabotropic glutamate 2 (mGLU2)	Metabotropic glutamate 5 (mGLU5)	Human		Biological	22300836	2012
Hetero	5HT2B	Metabotropic glutamate 2 (mGLU2)	Human		Biological	22300836	2012
Hetero	5HT2A	Metabotropic glutamate 2 (mGLU2)	Human		Biological	22300836	2012
Hetero	Cannabinoid CB1	δ-opioid	Mouse		Biological	22235275	2012
Hetero	α-1a adrenergic	α-1b adrenergic	Mouse		Biological	22120526	2012
Homo	α-1b adrenergic		Mouse		Biological	22120526	2012

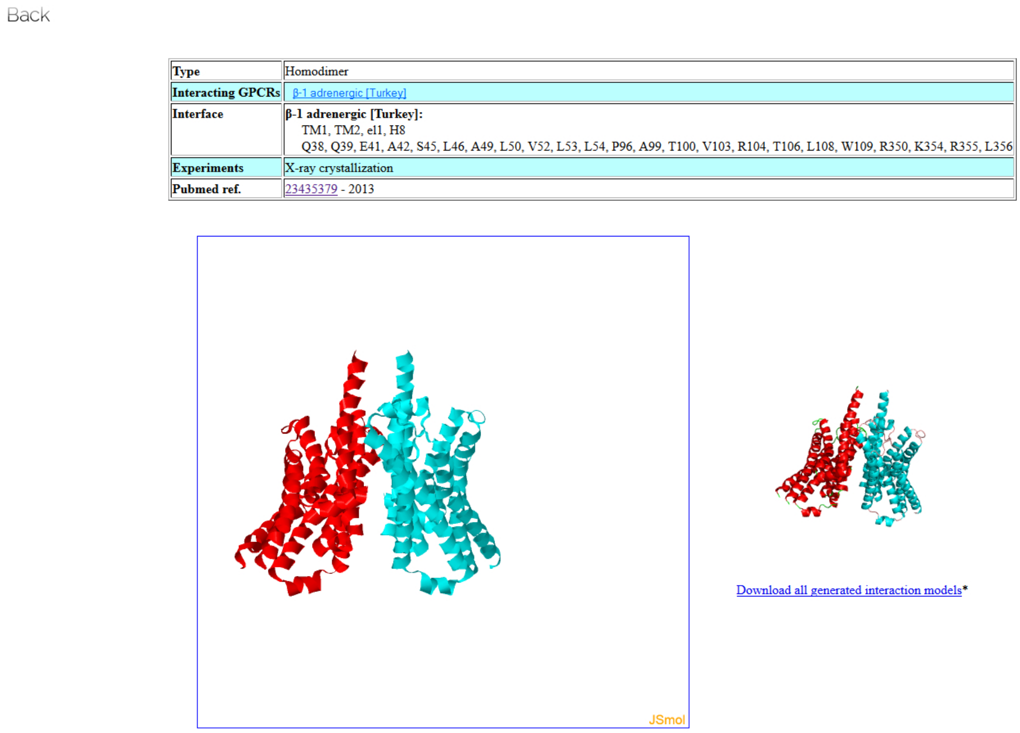
Figure 3.2: **A sample Browse page from the IntGPCR database system.**

Interaction Details and *GPCR Details* pages are the other informative pages that can be reachable using the links in the listed table. A sample *Interaction Details* page of turkey β-1 adrenergic receptor homodimer is shown in Figure 3.3. *Interaction Details* page lists the following extra information in addition to the listed above information about the clicked interactions:

- **Interface:** The interface details are listed in this field if they are included in the article. The interface can be interacting regions (Transmembrane, loops or C/N terminals) or specific interacting residues.

- **Interaction model:** The dimer model in the JSmol [40] window. The methods for creation of this model is described in the previous sections in this chapter.

- **Downloadable models:** The other interaction models created by the ClusPro 2.0 server [88]. These models are easily downloadable in .zip archive format file for further analysis.



Back

Type	Homodimer
Interacting GPCRs	β-1 adrenergic [Turkey]
Interface	**β-1 adrenergic [Turkey]:** TM1, TM2, el1, H8 Q38, Q39, E41, A42, S45, L46, A49, L50, V52, L53, L54, P96, A99, T100, V103, R104, T106, L108, W109, R350, K354, R355, L356
Experiments	X-ray crystallization
Pubmed ref.	23435379 - 2013

Download all generated interaction models*

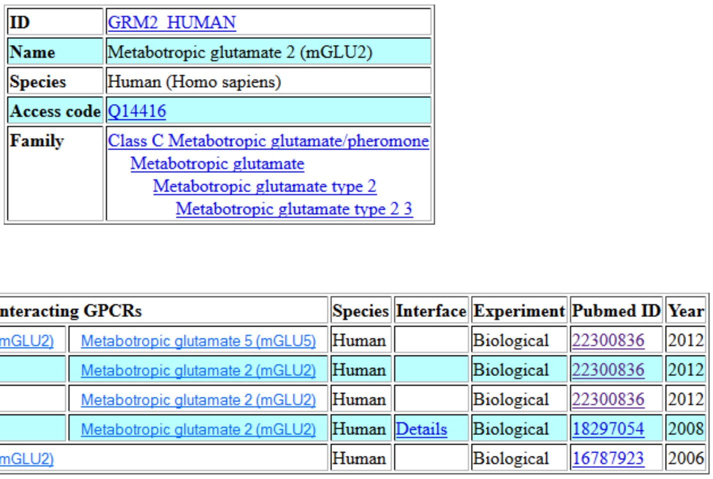
Figure 3.3: **A sample Interaction Details page from the IntGPCR database system.**

Figure 3.4 shows a sample page of details page of the human metabotropic glutamate 2 receptor. *GPCR Details* page lists the following information about the GPCR of interest:

- **ID:** GPCRDB [31] database ID of the GPCR. There is a link on this field to access the GPCRDB page.
- **Name:** Short name of the GPCR.
- **Species:** The species that this GPCR belongs to.
- **Access code:** Uniprot [95] access code of the GPCR. There is a link on this field to access the Uniprot page.

- **Family:** Family hierarchy of the GPCR. Every ancestor family member has a link to easily access to their GPCRDB page.

- **Involved interactions:** The table at the bottom of the page lists the interactions that this GPCR is involved in. The format of this table is the same as the table in the Browse page.



ID	GRM2_HUMAN
Name	Metabotropic glutamate 2 (mGLU2)
Species	Human (Homo sapiens)
Access code	Q14416
Family	Class C Metabotropic glutamate/pheromone Metabotropic glutamate Metabotropic glutamate type 2 Metabotropic glutamate type 2 3

Type	Interacting GPCRs		Species	Interface	Experiment	Pubmed ID	Year
Hetero	Metabotropic glutamate 2 (mGLU2)	Metabotropic glutamate 5 (mGLU5)	Human		Biological	22300836	2012
Hetero	5HT2B	Metabotropic glutamate 2 (mGLU2)	Human		Biological	22300836	2012
Hetero	5HT2A	Metabotropic glutamate 2 (mGLU2)	Human		Biological	22300836	2012
Hetero	5HT2A	Metabotropic glutamate 2 (mGLU2)	Human	Details	Biological	18297054	2008
Homo	Metabotropic glutamate 2 (mGLU2)		Human		Biological	16787923	2006

Figure 3.4: **A sample GPCR Details page from the IntGPCR database system.**

Search page is used to search the IntGPCR database for specific data. Results of the search are listed in the Browse page. The search can be performed by specifying the following parameters:

- **Interaction type:** Specifies which type of interactions can be listed. The options are: All, Homodimers or Heterodimers.

- **GPCR:** The names of the GPCRs are listed. A specific GPCR can be searched in the interacting GPCRs.

- **Species:** The species of the GPCRs are listed.

- **Pubmed ID:** A Pubmed ID can be entered in the specified textbox. If the related article is in the database, the interaction data contained in that article will be listed.

- **Experiment type:** The type of the experiment can be specified as All, Biological or Computational.

- **Interface:** The interface search parameter states if the interactions have interface data or not.

3.5 Discussion

In this chapter, the creation of a new database for the use of researchers is described. IntGPCR is very useful for the biological and computational studies about GPCRs. It provides researchers a portal for professional analysis of dimerizations of specific GPCRs. The interface is designed for easy access and the system is open to development further, meaning that updating the database can be accomplished easily.

The IntGPCR system differs from the other similar databases in different aspects. First and important of all is the reliability of the database. As stated earlier, erroneous entries make existing databases unreliable. The creation of the IntGPCR from the literature articles is processed meticulously. It was not easy to analyze those biological articles at first, because of the distance between the two disciplines, biology and computer science. However after getting familiarized about the topics, it became easy. This resulted in careful analysis of the articles.

Up-to-dateness of IntGPCR is another advantage when compared with the other databases. The first version of the IntGPCR database has been available in April 2014 and the database is continuously updated. However, the last update time of the other databases is in November 2012.

IntGPCR shines through the other databases according to the data size it contains. IntGPCR contains 309 interaction data curated from 348 PubMed articles. The biggest database between the others, GPCR-OKB, contains 192 interactions curated from 220 articles. Except the two interactions, all interactions in the other databases are contained in IntGPCR, whereas IntGPCR provides 119 more interaction data, which increases the value of IntGPCR.

Three-dimensional visualization of the dimers is a new concept proposed in IntGPCR. This eases the studies of researchers very much. One can easily visualize the formation of a dimer interactively. Interactive here is a key word, providing the users to see the big picture of their studies in different ways.

In the scope of studies about IntGPCR, the three dimensional structures of 36390 GPCRs are modelled. In view of the difficulty of crystallization of GPCRs, this number is very big in size. Although the models are predictions, for the initial studies of GPCRs, biologically and also computationally, these models are thought to be very helpful. These models and the interaction data from IntGPCR are used in the next study described in Chapter 4, Interaction Site Prediction of GPCR dimers.

CHAPTER 4

INTERACTION SITE PREDICTION OF GPCR DIMERS

This chapter includes the details of the proposed method for the prediction of interface residues in GPCR dimers and the evaluation of the performance of this proposed method operating on two different experimental environments. Known interface information from the literature and the three-dimensional structures are used for the prediction. The method employs searching of the interacting residues on three-dimensional contact interface after sequence alignment of regions where the training is done using a published known interface. The prediction results for several GPCRs are discussed and the performance of the method is analyzed. The results show that the proposed method performs its job well. The exact evaluation could be done after the biological verification of the results.

4.1 Background

There exist several mechanisms for the formation of GPCR dimers: coiled coil domains' interactions within the C-terminal tails of two receptors, intramolecular disulphide bonds between the cysteine residues in the long N-terminal regions, or interaction between the transmembrane regions of the proteins [55]. 'Domain swapping' and 'lateral packing' models are two types proposed for the transmembrane region interactions of the receptors [90]. In domain swapping, two independent regions separate from the receptor and recombine between the two receptors of the dimer [32]. The integrity of the interacting proteins are maintained in lateral packing model, where GPCRs contact each other via interaction sites on the transmembrane domains. Some studies suggest domain swapping in GPCR dimerization [2] while in some studies no evidence found about domain swapping despite dimerization takes place [48, 84]. Vast majority of studies approve lateral packing model, which is contact dimerization, for the formation of the GPCR dimers.

For the identification of the GPCR dimers various biological methods are applied. Those can be categorized as pharmacological, biochemical, biophysical and structural methods [78, 90]. The first indirect evidences for the existence of GPCR dimers are provided by pharmacological methods in 1980s. Type of these early studies are complex radioligand-binding experiments. Biochemical methods are employed generally in 1990s to observe the dimerization between GPCRs. Co-immunoprecipitation is the most used biochemical method to detect GPCR dimerization [90], where co-expression of differentially epitope-tagged receptors are in study. Several drawbacks of co-immunoprecipitation methods directed researchers to use additional methods to verify the detected interaction. Resonance energy transfer methods, which are able to detect protein-protein interactions in living cells, are applied for the demonstration of GPCR dimers. These type of experiments are biophysical methods, which are bioluminescence (BRET) and fluorescence (FRET) resonance energy transfer methods. Homo- or hetero-dimerization of a great number of GPCRs have been reported using these biophysical techniques [70]. Cell imaging and photo-bleaching protocols are combined with FRET to examine the cellular location of the interactions [56]. There are also alternative approaches, like bimolecular fluorescence complementation, to detect GPCR dimers in living cells. A combination of these techniques are employed most commonly to determine GPCR dimers because of the existence strengths and limitations of each technique. Besides that, the most acceptable GPCR dimerization evidences came from the result of structural studies such as atomic force microscopy and their crystal structures [50].

Bioinformatics techniques have been applied to predict likely interfaces of GPCR dimers. These studies begin with multiple sequence alignments and assume that the evolutionary related proteins display common structural and functional features [77]. Generally there exist two categories for the computational methods that predict oligomerization interfaces [28]. Docking methods are used if three-dimensional structural information is available. If this information is not available, bioinformatics methods based on sequence and genomic information are employed. Both of these categories have limitations and accuracy considerations specific to each other reviewed in [99, 21].

Receptor sequences are studied in the computational methods to predict dimerization of GPCRs. Correlated mutation analysis, subtractive correlated mutation, entropy, variants of the evolutionary trace method and hidden-site class model of evolution are sequence based methods used for predictions [85, 27, 101]. It was thought that the increase in the number of GPCR sequences will bring a significant increase in the power of these sequence based methods [28]. However in this area, there has been a comparatively little additional sequence-based study because of the limitations of these studies [85]. Docking, molecular dynamics simulations and coarse-grained simulations are other types of computational methods studied for the determination of GPCR dimerization interfaces. These methods, especially docking, are still in development. The results of these types of approaches need experimental validation [85]. Besides that, these methods have an important role in the determination of the interfaces.

4.2 Related Work

Computational methods are employed by some researchers for the prediction of the interaction site regions of GPCR dimers. The available information about GPCRs directed the studies in this field. Three-dimensional structure of GPCRs are mostly used in the computational prediction studies. Especially, recent increase of the number of experimentally determined three-dimensional structures showed their effects on the studies.

Generally, molecular dynamics simulations and rigid-body docking simulations are used in the literature. The first thing that is performed in these type of approaches is the homology modelling of the GPCR that is studied on using other available crystal structures of GPCRs. Rhodopsin crystal structure is widely used as template in the modelling of GPCRs like human muscarinic acetylcholine M1 [52], human neurotensin 1 [10], human lutropin [24] and human dopamine D2 and adenosine A2A

[7] receptors. Some examples of other studies that used homology modelling are; Johnston *et al.* builds the models of human β1-adrenergic and β2-adrenergic receptors using the crystal structures of turkey β1-adrenergic and β2-adrenergic receptors [41], mouse δ-opioid receptor models are generated using the human β2-adrenergic receptor x-ray crystal structure as template in [74]. In some experiments, modelling is used to model the missing regions of the known structures as in the case of human CXCR4 receptor [79].

Molecular dynamics simulations are performed in most of the studies. Most of the studies that use molecular dynamics, focus on dimeric interfaces that have received experimental validation according to publications. For instance, transmembrane (TM) domain 4 and TM domain 1 of β1-adrenergic and β2-adrenergic homodimers in [41], symmetrical TM1/helix 8 interface of rhodopsin homodimer in [67] and TM4 of the mouse δ-opioid receptor homodimers [74] are the focused regions on the prediction of the interfaces taking into account the previous experimental studies. Coarse-grained molecular dynamics are also applied in some predictions [67, 74]. The tools that are used for the molecular dynamics methodologies also differ. Groningen Machine for Chemical Simulations (GROMACS) [73] or Chemistry at Harvard Macromolecular Mechanics (CHARMM) [6] packages are used for the simulations whereas MARTINI force field [53] is seen to be used in coarse-grained versions.

In some studies, a computational approach based on rigid-body docking, ad hoc filtering and cluster analysis has been combined a protocol for dimerization free energy estimations to predict the likely interfaces in the dimers [10, 24]. Docking simulations are also widely used as in the case of molecular dynamics in the predictions. Docking is carried out on two identical copies of the monomers, i.e., one monomer is used as a fixed protein, which is target, and the other as a mobile protein, which is probe [26, 25, 24]. However, other versions of docking can also be seen in some studies. Data driven docking of a ligand on two muscarinic acetylcholine M1 receptors is performed in [52], where a dimeric interface is presented between the two receptors as a result. In another study, Canals *et al.* applies the rigid-body docking in [7] as this; one dopamine D2 is subjected to docking simulations with 11 different arrangements of adenosine A2A receptor. ZDOCK software [71] is generally employed in the studies for the rigid-body docking experiments.

Some researchers apply biological experiments in addition to their computational methods to validate the prediction results. Fanelli *et al.* employ an integrated approach of *in silico* and *in vitro* experiments [26]. *In vitro* experiments, which are site-directed mutagenesis, FRET and ligand binding studies, are employed to validate the predictions of their computational experiments. The same case is seen in the study of Canals *et al.* [7], that FRET and BRET techniques are used to demonstrate the adenosine A2A and dopamine D2 heterodimers in living cells in addition to rigid-body docking simulations.

The studies about the prediction of dimerization interfaces using computational experiments are generally seen to be focused on specific GPCRs on each study. Namely, the researchers make their experiments to resolve the interface of a specific homodimer or heterodimer. The examples are as follows: human β1-adrenergic and β2-adrenergic homodimers [41], human CXCR4 homodimer [79], human adenosine A2A homodimer [25], mouse δ-opioid homodimer [74], human dopamine D2 and adenosine A2A heterodimer [7] etc.

Studies that propose a methodology for the prediction of the interfaces for all GPCRs are not many. The studies of Taylor *et al.* [92] and Nemoto and Toh [63] are examples of this type of studies. Taylor *et al.* propose a novel method for *de novo* protein design in [92]. The method involves two main stages. In the first stage, a rank-ordered list of amino acid sequences is selected with flexible templates. These sequences are the inputs to the second stage. In this second stage, sequences are selected from a list of sequence positions as predicted to be in the interface region. The dimer of the glycophorin A is used as a model system to test the efficacy of this method and the model's results are found to be consistent with the experimental findings. A flexible template is developed for the rhodopsin homodimer and used to predict sets of three and five mutations. Their results are found to be consisted across the case studies.

Nemoto and Toh develops a new method to predict the interface for the GPCR oligomers [63]. Their method combines the information about the residue conservation with the structure data. Because of the lack of structural coordinate data at the time of their study, only the coordinate data from the bovine rhodopsin crystal structure is used in their proposed method. The accuracy of the method is checked with the available interface information from some Class A GPCRs that are experimentally suggested. In their method, there exist two assumptions; if a residue is involved in the oligomerization, the residue is expected to be conserved within the same subtype and also that conserved residues are more abundant at the interface region that at the non-interface surface. There exist two parts in their prediction procedure. In the first part, the sequence and structure data are preprocessed. The three-dimensional coordinates are reduced to two-dimensional plane. The information obtained from the first part is integrated in the second part where the prediction is carried out. The benefits and the pitfalls of the method for predicted interfaces in each subfamily are analyzed. Again the poorness of the dimer information at the time of the study limits the accuracy of the proposed method.

A novel method is proposed in this chapter for the prediction of interfaces of GPCR dimers. The proposed method combines the structure data with the sequence alignment data. The available interface information from the literature are used to evaluate the proposed method. The following sections contain the details of the proposed methodology and the evaluation of these results.

4.3 Materials and Methods

4.3.1 Dataset

The interface details of interacting GPCRs are taken from IntGPCR, which is described in detail in Chapter 3. The residues that are proposed to be in the dimer interface are important for the studies of this chapter.

The coordinates of the residues are taken from the PDB [76] file that contains three-dimensional structure of the known structured GPCRs, which are listed in Appendix A. The construction of models of the GPCRs, whose three-dimensional structures are unknown, is described in *Modelling of GPCRs* section in Chapter 3. Those models are used for the coordinates of the residues from the unknown GPCR structures.

4.3.2 Method

The steps of the proposed method for the prediction of the interaction sites of GPCR dimers are as follows:

1. Select the residues from a known interface:

T:	Selected GPCR
$IR_1, IR_2, ..., IR_n$:	Interface regions that contains the residues (For instance TM1, TM2 ...)
$R = \{R_1, R_2, ..., R_m\}$:	The set that contains the selected residues

2. Say P is the GPCR whose interacting sites are wanted to be learned. Get the same regions of the interface regions (IRs) of T from P. Make global sequence alignments of the corresponding regions separately. Local version of 'EMBOSS Needle' [54] that runs on PC is chosen to make the sequence alignments. At the end of this step, each region is aligned.

3. Create the matching residues sets R'_T and R'_P. For each R_x from R in aligned IR_T, get the corresponding residue from the aligned IR_P. If it is in the same group of amino acids as R_x, than put R_x to R'_T and put the other residue to R'_P. So final sets will be:

$$R'_T \subseteq R = \{\textit{Matching Rs from T}\}$$
$$R'_P = \{\textit{Matching Rs from P}\}$$

4. Calculating the distances and angles step. Take each R'_T and R'_P residues' coordinates from the pdb files (CA atoms' coordinates). Begin with the first 3 residues in each set. Name these residues as:

   First 3 from R'_T: P_{11}, P_{12}, P_{13}
   First 3 from R'_P: P_{21}, P_{22}, P_{23}

   These residues can be visualized as shown in the Figure 4.1a and 4.1b.

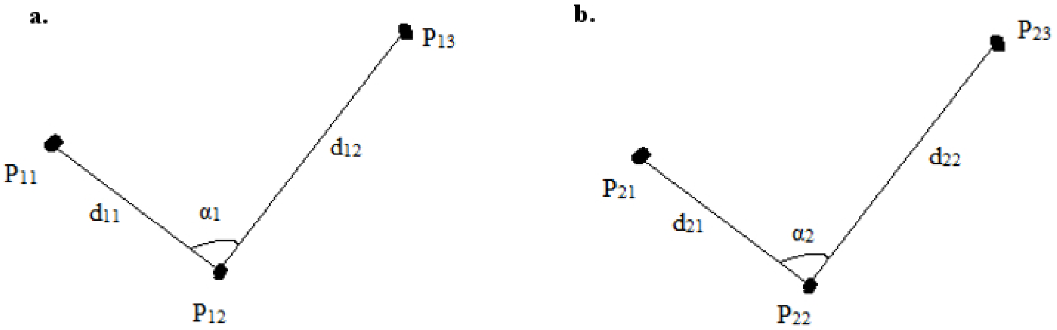


Figure 4.1: **Visualization of selected residues. a.** The 3 residues are selected from R'_T. **b.** The 3 residues are selected from R'_P.

Next step is the calculation of distances d_{11}, d_{12}, d_{21}, d_{22} and angles α_1 and α_2 in Figure 4.1. Calculation of d_{11}, d_{12} and α_1 is defined in Equation Box 4.1, calculation of the others is same.

Equation Box 4.1: **Calculation of distances and angles**

$$v_{11} = \vec{P_{12}P_{11}} = (P_{11}.x - P_{12}.x, P_{11}.y - P_{12}.y, P_{11}.z - P_{12}.z)$$
$$v_{12} = \vec{P_{12}P_{13}} = (P_{13}.x - P_{12}.x, P_{13}.y - P_{12}.y, P_{13}.z - P_{12}.z)$$
$$\boldsymbol{d_{11}} = \sqrt{v_{11}.x^2 + v_{11}.y^2 + v_{11}.z^2}$$
$$\boldsymbol{d_{12}} = \sqrt{v_{12}.x^2 + v_{12}.y^2 + v_{12}.z^2}$$
$$v_{11norm} = \left({}^{v_{11}.x}/_{d_{11}}, {}^{v_{11}.y}/_{d_{11}}, {}^{v_{11}.z}/_{d_{11}}\right)$$
$$v_{12norm} = \left({}^{v_{12}.x}/_{d_{12}}, {}^{v_{12}.y}/_{d_{12}}, {}^{v_{12}.z}/_{d_{12}}\right)$$
$$res = v_{11norm}.x \times v_{12norm}.x + v_{11norm}.y \times v_{12norm}.y + v_{11norm}.z \times v_{12norm}.z$$
$$\boldsymbol{\alpha_1} = \arccos(res)$$

After the calculations of the distances and angles that are shown in Figure 4.1, there will be 3 differences that will be used in the evaluations:

$$diff_1 = |d_{11} - d_{21}| \tag{4.1}$$
$$diff_2 = |d_{12} - d_{22}| \tag{4.2}$$
$$diff_3 = |\alpha_1 - \alpha_2| \tag{4.3}$$

Differences calculated from the equations 4.1, 4.2 and 4.3 will determine the interface residues. These values are compared with the thresholds defined below:

$$distance\_threshold_1 = d_{11} \times 0.15$$
$$distance\_threshold_2 = d_{12} \times 0.15$$
$$angle\_threshold = 10^\circ$$

The algorithm written in Algorithm 1 determines which residues are in the interface. After the process of Algorithm 1, the set R'_P contains the predicted interface residues for P in a dimer.

In Step 2 of the proposed algorithm, sequence alignments are applied between the determined regions of GPCRs. EMBOSS Needle tool [54] is employed to make these sequence alignments. This tool uses Needleman-Wunsch alignment algorithm [59] to find the optimum alignment of two sequences along their entire length. The regions

Algorithm 1 Interface residues prediction for GPCR dimers

```
while $P_{13}$ != NULL do
    if $P_{11}$ is the first entry in $R'_T$ then
        if $diff_1 \leq distance\_threshold_1$ && $diff_2 \leq distance\_threshold_2$ && $diff_3 \leq angle\_threshold$ then
            $P_{11} \leftarrow P_{12}$; $P_{12} \leftarrow P_{13}$;
            $P_{13} \leftarrow$ next residue from $R'_T$ if exists, else NULL;
            $P_{21} \leftarrow P_{22}$; $P_{22} \leftarrow P_{23}$;
            $P_{23} \leftarrow$ next residue from $R'_P$ if exists, else NULL;
        else if $diff_1 \leq distance\_threshold_1$ && ($diff_2 > distance\_threshold_2$ || $diff_3 > angle\_threshold$) then
            remove $P_{13}$ from $R'_T$
            $P_{13} \leftarrow$ next residue from $R'_T$ if exists, else NULL;
            remove $P_{23}$ from $R'_P$
            $P_{23} \leftarrow$ next residue from $R'_P$ if exists, else NULL;
```

Algorithm 1 Interface residues prediction for GPCR dimers (continued)

```
    else if $diff_1 > distance\_threshold_1$ then
      $d_{13} \leftarrow$ distance of vector $\overrightarrow{P_{11}P_{13}}$
      $d_{23} \leftarrow$ distance of vector $\overrightarrow{P_{21}P_{23}}$
      $diff_4 \leftarrow |d_{13} - d_{23}|$
      $distance\_threshold_3 \leftarrow d_{13} \times 0.15$
      if $diff_4 \leq distance\_threshold_3$ then
        remove $P_{12}$ from $R'_T$; $P_{12} \leftarrow P_{13}$;
        $P_{13} \leftarrow$ next residue from $R'_T$ if exists, else NULL;
        remove $P_{22}$ from $R'_P$; $P_{22} \leftarrow P_{23}$;
        $P_{23} \leftarrow$ next residue from $R'_P$ if exists, else NULL;
      else
        remove $P_{11}$ from $R'_T$; $P_{11} \leftarrow P_{12}$; $P_{12} \leftarrow P_{13}$;
        $P_{13} \leftarrow$ next residue from $R'_T$ if exists, else NULL;
        remove $P_{21}$ from $R'_P$; $P_{21} \leftarrow P_{22}$; $P_{22} \leftarrow P_{23}$;
        $P_{23} \leftarrow$ next residue from $R'_P$ if exists, else NULL;
      end if
    end if
  else
    if $diff_2 \leq distance\_threshold_2$ && $diff_3 \leq angle\_threshold$ then
      $P_{11} \leftarrow P_{12}$; $P_{12} \leftarrow P_{13}$;
      $P_{13} \leftarrow$ next residue from $R'_T$ if exists, else NULL;
      $P_{21} \leftarrow P_{22}$; $P_{22} \leftarrow P_{23}$;
      $P_{23} \leftarrow$ next residue from $R'_P$ if exists, else NULL;
    else
      remove $P_{13}$ from $R'_T$
      $P_{13} \leftarrow$ next residue from $R'_T$ if exists, else NULL;
      remove $P_{23}$ from $R'_P$
      $P_{23} \leftarrow$ next residue from $R'_P$ if exists, else NULL;
    end if
  end if
end while
```

are aligned separately. Gaps may be included in the resulting alignment. The default options of the tool are used in the operations.

There is a statement in the step 3 of the above algorithm mentioning about the amino acid groupings. A scheme of amino acid groupings is used to match similar amino acids in the matching phase of the algorithm. There are several properties of amino acids which could help to group them in similar clusters. These can be hydrophobicity, charge, mass, volume, etc. The amino acid grouping scheme defined by Cobanoglu *et. al.* [12] is used in the algorithm proposed here. This scheme groups amino acids into 11 groups which are: IVLM, RKH, DE, QN, ST, A, G, W, C, YF, and P.

4.3.3 Experimental Setup

Two experiments are studied. In each of them, a known interface is selected. Possible interface residues are found according to the selected known interface with the help of the proposed algorithm. The predicted interface residues are compared with the known ones that are available in the IntGPCR. Selected known interfaces for the two experiments are as follows:

- **Exp. 1:** *GPCR*: ADRB1_MELGA (P07700)
  Interface: TM1, TM2, el1, H8
  Residues: Q38, Q39, E41, A42, S45, L46, A49, L50, V52, L53, L54, P96, A99, T100, V103, R104, T106, L108, W109, R350, K354, R355, L356
- **Exp. 2:** *GPCR*: ADRB1_MELGA (P07700)
  Interface: il2, TM4, el2, TM5
  Residues: Y140, L141, T144, S145, F147, R148, S151, L152, L171, W181, R183, R205, A206, A210, I218, R229

4.4 Results

The Table 4.1 contains results of the predictions from the two experiments whose setups are explained in the previous section.

Table 4.1: **Predicted interface regions of some GPCRs whose interfaces are proposed in the literature**

GPCR	**Exp.**	**Known Interfaces**	**Predicted Interfaces (Exp. 1)**	**Predicted Interfaces (Exp. 2)**
P08483	1	ICL2, H8		TM3, ICL2, ICL3 Y166, T170, R252
	1	ICL3 [K259, K262, E263, L264, A265]		
	1	C140, C220		
	1	ICL3		
P07550	2	TM3, TM4	TM1, TM2, H8 I38, L42, V44, L45, P88, A91, L95, R333, L339	TM3, ICL2, TM4, ECL2, TM5 Y132, T136, S137, L163, W173, R175, A198, A202, V210, R221
	2	TM1, H8		
	1	TM1, H8		
	1	TM1, H8		
	2	K273, G280, L284, L287, P288, Y308		
P08588	2	TM3, TM4	TM1, TM2, H8 Q55, Q56, A59, L63, A66, L67, V69, L70, L71, P113, A116, T117, V120, R384	TM3, ICL2 Y157, L158, T161, S162
	2	TM1, H8		
P08100	2	TM1, H8	TM1	TM3, TM5

Table 4.1 – continued from previous page

GPCR	Exp.	Known Interfaces	Predicted Interfaces (Exp. 1)	Predicted Interfaces (Exp. 2)
	1	TM1, H8 [C316]	S38, M39, A42, L46, L47	Y136, V137, I214
	2	TM4, TM5 [H152, F159, A166, S202, Y206]		
	2	TM4, TM5		
	1	TM4, TM5		
	1	W175, Y206		
Q64264	3	TM4, el2, TM5 [W175, Y198, R151, R152]	TM1, TM2 S40, L41, I47, P91, A94	TM3, TM4, TM5 Y135, T139, L166, H193, I206
P61073	2	TM5, TM6	TM1, TM2 N35, I47, L50, P92, A95, V99	TM3, TM4, TM5 Y135, L136, L165, L210
	1	TM3, ICL2, TM4 [Y135, L136, H140, P147]		
	1	TM5, TM6 [L194, V197, V198, F201, M205, L210, W195-L267, N192-E268, L266-W195]		
P41145	1	TM1, TM2, H8		TM3, TM5 Y157, I158, I237
P42866	1	TM5, TM6	TM1	TM3, ICL2, TM5

Table 4.1 – continued from previous page

GPCR	Exp.	Known Interfaces	Predicted Interfaces (Exp. 1)	Predicted Interfaces (Exp. 2)
	1	TM1, TM2, H8	M65, A68, I69, I71, M72	Y166, I167, K174, M243
P33535	3	TM4	TM1 M65, A68, I69, I71, M72	TM3, TM5 Y166, I167, M243
P11229	2	TM6, TM7 [I383, L402, W405, L372, I413, C417, L420]	TM1, TM2 A26, I30, L34, L37, T84	TM3, ICL2, TM5 Y124, T128, L199, R210
P20309	1	TM1, TM2, TM4, ICL3, TM5, TM7	TM1, TM2	TM3, ICL2, ECL2, TM5
	2	F164, Y167, K183, R184, G186, V187, V194, F206, W207, G358, R362, K370, L371	I77, L80, V81, I130	Y167, T171, W207, R253
P32300	1	TM4, TM5 [V181, T213]	TM1, TM2	TM3, ECL2
	2	TM4 [P162, A163, K166, I170, W173, S177, V181, V185]	L48, A51, I52, L55, P103	Y147, I148, W207
	1	C-terminus		
O95665	1	TM2, ECL1, TM3, ICL2, TM4	TM1 T36, A40, L41	TM3, ICL2, TM4 L135, R142, L164

Table 4.1 – continued from previous page

GPCR	Exp.	Known Interfaces	Predicted Interfaces (Exp. 1)	Predicted Interfaces (Exp. 2)
P14416	2	K149, R150, R151, T153, V154, V161, F172, G173		
	1	ICL3 [R217, R218, R219, R220, K221, R222]		
	2	TM5, TM6, N-terminus		
	1	TM5, ICL3, TM6, ECL3		
	1	TM6, TM7		
Q13639	1	TM3, TM4 [C112, C145]	TM1, TM2 I30, L31, M32, A78, V82	TM3, TM5 Y120, K190, I203, R214
	2	TM2, TM4, TM6		
P25025	1	ECL1, TM3, ICL2		TM3, ICL2, TM4 Y145, L146, R153, L175
Q17094	1	H8	TM1, H8 I32, A35, I36, I39, I40, R316	
P29403	1	TM1, H8	TM1 S38, A42, L46, L47	TM3, TM5 Y136, I137, I214, R225

Table 4.1 contains the interface regions of GPCRs that are curated from the literature. The experiments which produce these interfaces can be biological, computational or both. The table also contains this information in the *Exp.* column. The values in this column can be either 1, 2 or 3 indicating biological, computational and both respectively. There can be more than one proposed interface for a GPCR. These interfaces are listed with their experiment types in several rows under the spanning GPCR row. For instance, there exist four proposed interfaces for the first GPCR, P08483 that is rat Muscarinic acetylcholine receptor M3, each are found by the biological experiments. The *Known Interfaces* column contains those proposed interfaces. The information on that column can be either only regions or amino acids with their corresponding regions. The last two columns in the table show the prediction results of the proposed method in this study operated on the two experiments, if they exist. The method predicts the amino acid residues that are in the interface region. The corresponding regions of those predicted residues are also included in those columns.

If Table 4.1 is analyzed carefully, it can be seen that the performance of the proposed method is good. The method can find some exact amino acids that are also marked as the known interface. For example, for the P61073 GPCR, three of the four predicted residues are also found to be in the interface by two different biological experiments. This situation can also be seen in the P20309 GPCR, but in that case, the detection method of the known interface is a computational study. Besides these, the method can predict the interface regions for most of the GPCRs, such as P08483, P07550, O95665, etc.

The table also contains some interesting results. One of them is in the case of the GPCRs P08483 and P25025. For these GPCRs, the method could not find any residue in the interface in the first experiment. The method gives the prediction results for the second experiment. The biological experiments approve this, saying that the only interacting domains are those. On the other hand, for the P41145 GPCR, the method cannot predict any residue from the proposed known interface, but gives some other predictions from another regions. For some GPCRs, there exist only computational predictions in the literature, like P08588. The proposed method in this study confirms those computational predictions too as can be seen in the table. There exist different proposed known interface regions as in the case of P42866 GPCR. The method could predict both those regions in the two experiments.

The method cannot predict any interface residue for GPCR P14416 in both of the experiments. This is because of the difference of the actual interacting domains from the domains that the predictions are based on. For this protein, ICL3, TM6 and ECL3 domains are proposed as the known interfaces by the biological experiments. These are different regions from the used regions, which are explained in the previous section, in both of the experiments. Indeed the proposed method supports the biological experiments here, saying no other interface region exists for this GPCR.

Table 4.1 lists prediction results of the GPCRs whose interfaces are known from the literature. The method proposed here can make its predictions on each type of GPCR. The Table B.1 and B.2 in Appendix B, which list the prediction results of some selected GPCRs from the experiment 1 and 2 respectively, could be helpful for the researchers in their studies.

4.5 Discussion

A new interface prediction algorithm is proposed in this study. The proposed method is able to predict interacting site residues in GPCR dimers using the known proposed interface data from the literature. The method uses both the sequence and three-dimensional structure information about the GPCRs. The structural data used in the method can be either the known crystal structure of the receptor or the created model of the receptor, where the details of the creation are explained in the previous chapter.

The proposed algorithm has some advantages as well as its disadvantages. The most valuable part of the method is its ability of searching of the interface residues in three-dimensions, which makes the results more realistic. Here comes the lack of the three-dimensional structures problem, where only a small number of GPCRs are crystallized. This problem is overcome with the use of the models for GPCRs. It is obvious that the performance of the method will increase with the increasing number of the crystallized GPCRs.

The method uses the evolutionary data about the GPCRs with the alignment process of the sequence of the receptors. Evolutionary development of the genes resulted with the same structural regions between the similar receptors [69]. This also suggests that the existence of similar interface residues in these GPCRs. In the light of this information, the power of the method is increased with the use of sequence alignment of regions.

The method cannot predict any interface regions for some GPCRs, which are known to include interface data. This is because of the minority of the known interface information that is curated from the literature. The method uses known interface data for the training. The known interface is specific to some local regions, for instance in one interface, the regions are TM1 and TM2, where in another interface the regions are TM4 and TM5. The method predicts the interface residues from these regions, if exists. If the known interacting region is different, no prediction can be made by the method. This deficiency of the method will be resolved with more known interface data from the literature.

The two experiments are applied for the evaluation of the proposed method in this study. The biggest problem in the evaluation is the count of the known interfaces. The low count of available known interfaces, especially biologically determined ones, makes it difficult to evaluate the results. In general, predicted regions are matched up with the corresponding known regions. The predictions should be validated by biological techniques. According to the results of those biological evaluations, some improvements could be applied to the proposed method to increase its performance.

CHAPTER 5

CONCLUSION AND FUTURE WORK

Three contributions, whose details are explained in this dissertation, are provided in this study. First of them is the presentation of a new method, GPCRsort, for classification of G-protein couple receptor sequences into GPCRDB classes. GPCRsort is solely based on the lengths of secondary structure elements of a GPCR sequence as identified by a secondary structure prediction tool specialized for transmembrane proteins. The lengths of the secondary structure elements of different GPCR classes are used to train a Random Forest classifier. GPCRSort is evaluated on several experimental setups and outperforms many state-of-the-art GPCR classifiers in terms of both prediction accuracy and running time performance. Specifically, GPCRSort is able to attain 97.3% prediction accuracy on the average and is able to predict the class of a novel GPCR sequence in seconds.

Second work published to the community is the IntGPCR, the database of interacting GPCRs. The IntGPCR contains information about interacting GPCRs and their interfaces if exist. The data is curated from the literature search from the PubMed. The published articles from the literature are analyzed carefully for each piece of information about any GPCR dimer. The contents of the database are proposed interaction data resulting from biological or computational experiments. The data is presented to the use of researchers in a web site. The IntGPCR is an easily browsable and searchable portal, in addition to be able to visualize the interacting GPCRs interactively. The database differs from similar databases with its up-to-dateness and the wealth of its contents. IntGPCR contains 309 interacting GPCRs curated from 348 articles. 138 of those interacting GPCRs contain interface information.

The last contribution is the proposed novel method for the prediction of the interface residues of GPCR dimers. This method uses the sequence and the three-dimensional data of GPCRs to predict its interacting sites in a dimer. This approach is based on the known interface information that is published in the literature. Searching the amino acids in three-dimensional structures adds realistic prediction to the method. Because of the nature of the interacting domains, coordinates in the three-dimensional space are the best distinguishable properties of the residues. The predictions are made based on the locations of the residues according to each other, using the distances and the angles between them. The performance of the proposed method is evaluated with designed experiments on the known interface data. The results are very promising and consistent with the real data.

Besides these explained information about the proposed studies, there remains some future works to improve the performance of these studies. For instance, the results of the proposed method for the prediction of the interacting sites of GPCR dimers should be approved with the biological experiments. This will show the perfect accuracy of the method. Moreover, it will also direct the improvement studies on the proposed method to make adjustments on it.

The increase on the number of the crystallized GPCRs will bring performance increase on the proposed method also. Also, this will provide more realistic three-dimensional models for GPCRs. Besides this, if the number of studies on published known interface residues increases, the accuracy will be higher. This will also affect the value of the IntGPCR in a positive manner.

IntGPCR portal is designed to be easily updatable. Every administrator could be able to update the database through an interface. To actualize this, a new section should be added to the portal for updating the IntGPCR. Users, who are not experienced in technical details of the database, could easily add a new entry, edit or delete an entry by this way. This will bring the continuous up-to-dateness to the IntGPCR.

REFERENCES

[1] S. Angers, A. Salahpour, and M. Bouvier. Dimerization: an emerging concept for G protein-coupled receptor ontogeny and function. *Annual Review of Pharmacology and Toxicology*, 42:409–435, 2002.

[2] R. A. Bakker, G. Dees, J. J. Carrillo, R. G. Booth, J. F. López-Gimenez, G. Milligan, P. G. Strange, and R. Leurs. Domain Swapping in the Human Histamine H1 Receptor. *Journal of Pharmacology and Experimental Therapeutics*, 311:131–138, 2004.

[3] M. Bhasin and G. P. S. Raghava. GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. *Nucleic Acids Research*, 32:383–389, 2004.

[4] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[5] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[6] B. R. Brooks, C. L. B. III, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. CHARMM: The Biomolecular Simulation Program. *Journal of Computational Chemistry*, 30:1545–1614, 2009.

[7] M. Canals, D. Marcellino, F. Fanelli, F. Ciruela, P. de Benedetti, S. R. Goldberg, K. Neve, K. Fuxe, L. F. Agnati, A. S. Woods, S. Ferré, C. Lluis, M. Bouvier, and R. Franco. Adenosine A2A-dopamine D2 receptor-receptor heteromerization: qualitative and quantitative assessment by fluorescence and bioluminescence energy transfer. *Journal of Biological Chemistry*, 278:46741–46749, 2003.

[8] J. C. Cardoso, V. C. Pinto, F. A. Vieira, M. S. Clark, and D. M. Power. Evolution of secretin family GPCR members in the metazoa. *BMC Evolutionary Biology*, 6:108, 2006.

[9] R. Caruana, N. Karampatziakis, and A. Yessenalina. An empirical evaluation of supervised learning in high dimensions. *Proceedings of the 25 International Conference on Machine Learning*, 25:96–103, 2008.

[10] D. Casciari, D. Dell'Orco, and F. Fanelli. Homodimerization of neurotensin 1 receptor involves helices 1, 2, and 4: insights from quaternary structure predictions and dimerization free energy estimations. *Journal of Chemical Information and Modeling*, 48:1669–1678, 2008.

[11] C. Chothia and A. M. Lesk. The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, 5:823–826, 1986.

[12] M. C. Cobanoglu, Y. Saygin, and U. Sezerman. Classification of GPCRs using family specific motifs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8:1495–1508, 2011.

[13] S. R. Comeau, D. W. Gatchell, S. Vajda, and C. J. Camacho. ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics*, 20:45–50, 2004.

[14] M. Cottet, O. Faklaris, D. Maurel, P. Scholler, E. Doumazane, E. Trinquet, J.-P. Pin, and T. Durroux. BRET and time-resolved FRET strategy to study GPCR oligomerization: from cell lines toward native tissues. *Frontiers in Endocrinology*, 3(92), 2012.

[15] E. A. Coutsias, C. Seok, and K. A. Dill. Using quaternions to calculate RMSD. *Journal of Computational Chemistry*, 25:1849–1857, 2004.

[16] J. Currier. SchemaSpy: Graphical Database Schema Metadata Browser. http://schemaspy.sourceforge.net, last visited on July 2014.

[17] S. S. Das and G. A. Banker. The role of protein interaction motifs in regulating the polarity and clustering of the metabotropic glutamate receptor mGluR1a. *The Journal of Neuroscience*, 26:8115–8125, 2006.

[18] M. N. Davies, D. E. Gloriam, A. Secker, A. A. Freitas, M. Mendao, J. Timmis, and D. R. Flower. Proteomic applications of automated GPCR classification. *Proteomics*, 7:2800–2814, 2007.

[19] M. N. Davies, A. Secker, A. A. Freitas, M. Mendao, J. Timmis, and D. R. Flower. On the hierarchical classification of G protein coupled receptors. *Bioinformatics*, 23:3113–3118, 2007.

[20] M. N. Davies, A. Secker, M. Halling-Brown, D. S. Moss, A. A. Freitas, J. Timmis, E. Clark, and D. R. Flower. Gpcrtree: online hierarchical classification of GPCR function. *BMC Research Notes*, 1:67, 2008.

[21] J. Dixon. Evaluation of the casp2 docking section. *Proteins*, Supplement 1:198–204, 1997.

[22] B. Duthey, S. Caudron, J. Perroy, B. Bettler, L. Fagni, J.-P. Pin, and L. Prézeau. A single subunit (GB2) is required for G-protein activation by the heterodimeric GABA(B) receptor. *Journal of Biological Chemistry*, 277:3236–3241, 2002.

[23] N. Eswar, B. Webb, M. A. Marti-Renom, M. Madhusudhan, D. Eramian, M. yi Shen, U. Pieper, and A. Sali. *Comparative Protein Structure Modeling Using Modeller, in Current Protocols in Bioinformatics*, chapter 5.6, page 5.6.1–5.6.30. John Wiley & Sons, Inc., 2002.

[24] F. Fanelli. Dimerization of the lutropin receptor: Insights from computational modeling. *Molecular and Cellular Endocrinology*, 260–262:59–64, 2007.

[25] F. Fanelli and A. Felline. Dimerization and ligand binding affect the structure network of A(2A) adenosine receptor. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1808:1256–1266, 2011.

[26] F. Fanelli, M. Mauri, V. Capra, F. Raimondi, F. Guzzi, M. Ambrosio, G. E. Rovati, and M. Parenti. Light on the structure of thromboxane A2 receptor heterodimers. *Cellular and Molecular Life Sciences*, 68:3109–3120, 2011.

[27] M. Filizola. Increasingly accurate dynamic molecular models of G-protein coupled receptor oligomers: Panacea or Pandora's box for novel drug discovery? *Life Sciences*, 86:590–597, 2010.

[28] M. Filizola and H. Weinstein. The study of G-protein coupled receptor oligomerization with computational modeling and bioinformatics. *FEBS Journal*, 272:2926–2938, 2005.

[29] R. Fredriksson, M. C. Lagerström, L.-G. Lundin, and H. B. Schiöth. The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Molecular Pharmacology*, 63:1256–1272, 2003.

[30] S. R. George, T. Fan, Z. Xie, R. Tse, V. Tam, G. Varghese, and B. F. O'Dowd. Oligomerization of mu- and delta-opioid receptors: Generation of novel functional properties. *Journal of Biological Chemistry*, 275:26128–26135, 2000.

[31] D. Gloriam. GPCRDB: information system for G protein-coupled receptors. http://www.gpcr.org/7tm, last visited on July 2014.

[32] P. R. Gouldson, C. Higgs, R. E. Smith, M. K. Dean, G. V. Gkoutos, and C. A. Reynolds. Dimerization and Domain Swapping in G-Protein-Coupled Receptors: A Computational Study. *Neuropsychopharmacology*, 23:S60–S77, 2000.

[33] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11:10–18, 2009.

[34] T. K. Ho. Random decision forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, 3:278–282, 1995.

[35] M. Hofmann and R. Klinkenberg. *RapidMiner: Data Mining Use Cases and Business Analytics Applications*. CRC Press, 2013.

[36] F. Horn, E. Bettler, L. Oliveira, F. Campagne, F. E. Cohen, and G. Vriend. GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Research*, 31:294–297, 2003.

[37] Y. Inoue, M. Ikeda, and T. Shimizu. Proteome-wide classification and identification of mammalian-type GPCRs by binary topology pattern. *Computational Biology and Chemistry*, 28:39–49, 2004.

[38] IUPAC. *Compendium of Chemical Terminology, 2nd ed. (the "Gold Book")*. Blackwell Scientific Publications, Oxford, 1997.

[39] E. Jacoby, R. Bouhelal, M. Gerspacher, and K. Seuwen. The 7 TM G-protein-coupled receptor target family. *ChemMedChem*, 1:761–782, 2006.

[40] Jmol Community. JSmol: an open-source HTML5 viewer for chemical structures in 3D. http://wiki.jmol.org/index.php/JSmol#JSmol, last visited on July 2014.

[41] J. M. Johnston, H. Wang, D. Provasi, and M. Filizola. Assessing the relative stability of dimer interfaces in g protein-coupled receptors. *PLoS Computational Biology*, 8:e1002649, 2012.

[42] K. Katoh and D. M. Standley. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30:772–780, 2013.

[43] K. Kaupmann, B. Malitschek, V. Schuler, J. Heid, W. Froestl, P. Beck, J. Mosbacher, S. Bischoff, A. Kulik, R. Shigemoto, A. Karschin, and B. Bettler. GABA(B)-receptor subtypes assemble into functional heteromeric complexes. *Nature*, 396:683–687, 1998.

[44] G. Khelashvili, K. Dorff, J. Shan, M. Camacho-Artacho, L. Skrabanek, B. Vroling, M. Bouvier, L. A. Devi, S. R. George, J. A. Javitch, M. J. Lohse, G. Milligan, R. R. Neubig, K. Palczewski, M. Parmentier, J.-P. Pin, G. Vriend, F. Campagne, and M. Filizola. GPCR-OKB: the G Protein Coupled Receptor Oligomer Knowledge Base. *Bioinformatics*, 26:1804–1805, 2010.

[45] L. Kolakowski. GCRDb: a G-protein-coupled receptor database. *Receptors Channels*, 2:1–7, 1994.

[46] E. Krieger, S. B. Nabuurs, and G. Vriend. *Homology Modeling, in: PE Bourne and H Weissig (Eds.), Structural Bioinformatics, p. 507-521*. Wiley-Liss, 2003.

[47] A. Krogh, B. Larsson, G. von Heijne, and E. L. Sonnhammer. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology*, 305:567–580, 2001.

[48] S. P. Lee, B. F. O'Dowd, G. Y. Ng, G. Varghese, H. Akil, A. Mansour, T. Nguyen, and S. R. George. Inhibition of Cell Surface Expression by Mutant Receptors Demonstrates that D2 Dopamine Receptors Exist as Oligomers in the Cell. *Molecular Pharmacology*, 58:120–128, 2000.

[49] A. Levoye, J. Dam, M. A. Ayoub, J.-L. Guillaume, and R. Jockers. Do orphan G-protein-coupled receptors have ligand-independent functions? New insights from receptor heterodimers. *EMBO Reports*, 7:1094–1098, 2006.

[50] Y. Liang, D. Fotiadis, S. Filipek, D. A. Saperstein, K. Palczewski, and A. Engel. Organization of the G protein-coupled receptors rhodopsin and opsin in native membranes. *Journal of Biological Chemistry*, 278:21655–21662, 2003.

[51] X. Liu, M. Kai, L. Jin, and R. Wang. Computational study of the heterodimerization between μ and δ receptors. *Journal of Computer-Aided Molecular Design*, 23:321–332, 2009.

[52] C. Marquer, C. Fruchart-Gaillard, G. Letellier, E. Marcon, G. Mourier, S. Zinn-Justin, A. Ménez, D. Servent, and B. Gilquin. Structural model of ligand-G protein-coupled receptor (GPCR) complex based on experimental double mutant cycle data: MT7 snake toxin bound to dimeric hM1 muscarinic receptor. *Journal of Biological Chemistry*, 286:31661–31675, 2011.

[53] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries. The MARTINI force field: coarse grained model for biomolecular simulations. *The Journal of Physical Chemistry B*, 111:7812–7824, 2007.

[54] H. McWilliam, W. Li, M. Uludag, S. Squizzato, Y. M. Park, N. Buso, A. P. Cowley, and R. Lopez. Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Research*, 41:W597–W600, 2013.

[55] G. Milligan. Oligomerisation of G-protein-coupled receptors. *Journal of Cell Science*, 114:1265–1271, 2001.

[56] G. Milligan and M. Bouvier. Methods to monitor the quaternary structure of G protein-coupled receptors. *FEBS Journal*, 272:2914–2925, 2005.

[57] S. Moller, M. D. Croning, and R. Apweiler. Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, 17:646–653, 2001.

[58] National Center for Biotechnology Information. PubMed Database. http://www.ncbi.nlm.nih.gov/pubmed, last visited on July 2014.

[59] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453, 1970.

[60] W. Nemoto, K. Fukui, and H. Toh. GRIP Database. http://grip.cbrc.jp/GDB/index.html, last visited on July 2014.

[61] W. Nemoto, K. Fukui, and H. Toh. GRIP: A server for predicting interfaces for GPCR oligomerization. *Journal of Receptors and Signal Transduction*, 29:312–317, 2009.

[62] W. Nemoto, K. Fukui, and H. Toh. GRIPDB - G protein coupled Receptor Interaction Partners Database. *Journal of Receptors and Signal Transduction*, 31:199–205, 2011.

[63] W. Nemoto and H. Toh. Prediction of interfaces for oligomerizations of G-protein coupled receptors. *Proteins: Structure, Function, and Bioinformatics*, 58:644–660, 2005.

[64] Oracle Corp. MySQL. http://www.mysql.com, last visited on July 2014.

[65] A. Pagano, G. Rovelli, J. Mosbacher, T. Lohmann, B. Duthey, D. Stauffer, D. Ristig, V. Schuler, I. Meigel, C. Lampert, T. Stein, L. Prézeau, J. Blahos, J.-P. Pin, W. Froestl, R. Kuhn, J. Heid, K. Kaupmann, and B. Bettler. C-terminal interaction is essential for surface trafficking but not for heteromeric assembly of GABA(b) receptors. *The Journal of Neuroscience*, 21:1189–1202, 2001.

[66] Z.-L. Peng, J.-Y. Yang, and X. Chen. An improved classification of G-protein-coupled receptors using sequence-derived features. *BMC Bioinformatics*, 11:420–432, 2010.

[67] X. Periole, A. M. Knepp, T. P. Sakmar, S. J. Marrink, and T. Huber. Structural determinants of the supramolecular organization of G protein-coupled receptors in bilayers. *Journal of the American Chemical Society*, 134:10959–10965, 2012.

[68] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin. UCSF Chimera—A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25:1605–1612, 2004.

[69] P. Pevzner and R. Shamir. *Bioinformatics for Biologists*. Cambridge University Press, New York, 2011.

[70] K. D. Pfleger and K. A. Eidne. Monitoring the formation of dynamic G-protein-coupled receptor-protein complexes in living cells. *The Biochemical Journal*, 385:625–637, 2005.

[71] B. G. Pierce, K. Wiehe, H. Hwang, B. H. Kim, T. Vreven, and Z. Weng. ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics*, 30:1771–1773, 2014.

[72] D. M. Powers. Evaluation Evaluation: a Monte Carlo study. *ECAI 2008, Frontiers in Artificial Intelligence and Applications, European Conference on Artificial Intelligence*, 178, 2008.

[73] S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, , and E. Lindahl. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29:845–854, 2013.

[74] D. Provasi, J. M. Johnston, and M. Filizola. Lessons from free energy simulations of delta-opioid receptor homodimers involving the fourth transmembrane helix. *Biochemistry*, 49:6771–6776, 2010.

[75] RapidMiner. RapidMiner Studio. http://www.rapidminer.com/products, last visited on July 2014.

[76] RCSB. PDB: Protein Data Bank. http://www.rcsb.org, last visited on July 2014.

[77] P. H. Reggio. Computational methods in drug design: modeling G protein-coupled receptor monomers, dimers, and oligomers. *AAPS Journal*, 8:E322–E336, 2006.

[78] C. D. Rios, B. A. Jordan, I. Gomes, and L. A. Devi. G-protein-coupled receptor dimerization: modulation of receptor function. *Pharmacology & Therapeutics*, 92:71–87, 2001.

[79] D. Rodríguez and H. G. de Terán. Characterization of the homodimerization interface and functional hotspots of the CXCR4 chemokine receptor. *Proteins: Structure, Function, and Bioinformatics*, 80:1919–1928, 2012.

[80] D. M. Rosenbaum, S. G. Rasmussen, and B. K. Kobilka. The structure and function of g-protein-coupled receptors. *Nature*, 459:356–363, 2009.

[81] A. Salahpour, S. Angers, and M. Bouvier. Functional significance of oligomerization of G-protein-coupled receptors. *Trends in Endocrinology & Metabolism*, 11:163–168, 2000.

[82] A. Salahpour, S. Angers, J.-F. Mercier, M. Lagacé, S. Marullo, and M. Bouvier. Homodimerization of the β2-adrenergic receptor as a prerequisite for cell surface targeting. *Journal of Biological Chemistry*, 279:33390–33397, 2004.

[83] M. P. Sanders, W. W. Fleuren, S. Verhoeven, S. van den Beld, W. Alkema, J. de Vlieg, and J. P. Klomp. ss-TEA: Entropy based identification of receptor

specific ligand binding residues from a multiple sequence alignment of class A GPCRs. *BMC Bioinformatics*, 12:332–344, 2011.

[84] A. Schulz, R. Grosse, G. Schultz, T. Gudermann, and T. Schöneberg. Structural implication for receptor oligomerization from functional reconstitution studies of mutant v2 vasopressin receptors. *Journal of Biological Chemistry*, 275:2381–2389, 2000.

[85] L. M. Simpson, B. Taddese, I. D. Wall, and C. A. Reynolds. Bioinformatics and molecular modelling approaches to GPCR oligomerization. *Current Opinion in Pharmacology*, 10:30–37, 2010.

[86] solid IT. DB-Engines Ranking. http://www.db-engines.com/en/ranking, last visited on July 2014.

[87] E. L. Sonnhammer, G. von Heijne, and A. Krogh. A hidden markov model for predicting transmembrane helices in protein sequences. *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology*, 6:175–182, 1998.

[88] Structural Bioinformatics Lab, Boston University. ClusPro 2.0: protein-protein docking. http://cluspro.bu.edu, last visited on July 2014.

[89] B. E. Suzek, H. Huang, P. McGarvey, R. Mazumder, and C. H. Wu. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, 23:1282–1288, 2007.

[90] L. Szidonya, M. Cserző, and L. Hunyady. Dimerization and oligomerization of G-protein-coupled receptors: debated structures with established and emerging functions. *Journal of Endocrinology*, 196:435–453, 2008.

[91] X.-L. Tang, Y. Wang, D.-L. Li, J. Luo, and M.-Y. Liu. Orphan G protein-coupled receptors (GPCRs): biological functions and potential drug targets. *Acta Pharmacol Sin*, 33:363–371, 2012.

[92] M. S. Taylor, H. K. Fung, R. Rajgaria, M. Filizola, H. Weinstein, and C. A. Floudas. Mutations affecting the oligomerization interface of G-protein-coupled receptors revealed by a novel de novo protein design framework. *Biophysical Journal*, 94:2470–2481, 2008.

[93] S. Terrillon and M. Bouvier. Roles of G-protein-coupled receptor dimerization. *EMBO reports*, 5:30–34, 2004.

[94] The PHP Group. PHP: Hypertext Preprocessor. http://www.php.net, last visited on July 2014.

[95] The UniProt Consortium. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, 42:D191–D198, 2014.

[96] M. C. Theodoropoulou, P. G. Bagos, I. C. Spyropoulos, and S. J. Hamodrakas. gpDB: a database of GPCRs, G-proteins, Effectors and their interactions. *Bioinformatics*, 24:1471–1472, 2008.

[97] M. C. Theodoropoulou, A. L. Elefsinioti, P. G. Bagos, I. C. Spyropoulos, and S. J. Hamodrakas. gpDB: a database of GPCRs, G-proteins, Effectors and their interactions. http://bioinformatics.biol.uoa.gr/gpDB, last visited on July 2014.

[98] G. E. Tusnády and I. Simon. The hmmtop transmembrane topology prediction server. *Bioinformatics*, 17:849–850, 2001.

[99] A. Valencia and F. Pazos. Computational methods for the prediction of protein interactions. *Current Opinion in Structural Biology*, 12:368–373, 2002.

[100] E. van der Horst, J. E. Peironcely, A. P. IJzerman, M. W. Beukers, J. R. Lane, H. W. van Vlijmen, M. T. Emmerich, Y. Okuno, and A. Bender. A novel chemogenomics analysis of G protein-coupled receptors (GPCRs) and their ligands: a potential strategy for receptor de-orphanization. *BMC Bioinformatics*, 11:316–327, 2010.

[101] S. Vohra, S. V. Chintapalli, C. J. R. Illingworth, P. J. Reeves, P. M. Mullineaux, H. S. X. Clark, M. K. Dean, G. J. G. Upton, and C. A. Reynolds. Computational studies of Family A and Family B GPCRs. *Biochemical Society Transactions*, 35:749–754, 2007.

[102] B. Vroling, M. Sanders, C. Baakman, A. Borrmann, S. Verhoeven, J. Klomp, L. Oliveira, J. de Vlieg, and G. Vriend. GPCRDB: information system for G protein-coupled receptors. *Nucleic Acids Research*, 39:309–319, 2011.

[103] Wikimedia Foundation, Inc., media is originally uploaded by Bensaccount. G protein-coupled receptor. http://en.wikipedia.org/wiki/G_protein-coupled_receptor, last visited on July 2014.

[104] J. Xu, J. He, A. M. Castleberry, S. Balasubramanian, A. G. Lau, and R. A. Hall. Heterodimerization of alpha 2a- and beta 1-adrenergic receptors. *Journal of Biological Chemistry*, 278:10770–10777, 2003.

[105] J. Yang and Y. Zhang. GPCRSD: a database for experimentally solved GPCR structures. http://zhanglab.ccmb.med.umich.edu/GPCRSD, last visited on July 2014.

APPENDIX A

GPCR STRUCTURES

Table A.1: **Experimentally solved GPCR structures**

GPCR	**Species**	**UniProt Id**	**PDB Id**	**Family**
Class A \ Amine Family				
Muscarinic acetylcholine receptor M2	Homo sapiens (Human)	P08172	3uon	Musc. acetylcholine Vertebrate type 2
Muscarinic acetylcholine receptor M3	Rattus norvegicus (Rat)	P08483	4daj	Musc. acetylcholine Vertebrate type 3
β-1 adrenergic receptor	Meleagris gallopavo (Turkey)	P07700	4amj	Beta Adrenoceptors type 1
β-2 adrenergic receptor	Homo sapiens (Human)	P07550	2rh1	Beta Adrenoceptors type 2
D(3) dopamine receptor	Homo sapiens (Human)	P35462	3pbl	Dopamine Vertebrate type 3
Histamine H1 receptor	Homo sapiens (Human)	P35367	3rze	Histamine type 1
5-hydroxytryptamine receptor 1B	Homo sapiens (Human)	P28222	4iar	Serotonin type 1b
5-hydroxytryptamine receptor 2B	Homo sapiens (Human)	P41595	4ib4	Serotonin type 2b
Class A \ Peptide Family				
C-X-C chemokine receptor type 1	Homo sapiens (Human)	P25024	2lnl	Interleukin-8 type A
C-C chemokine receptor type 5	Homo sapiens (Human)	P51681	4mbs	C-C Chemokine type 5
C-X-C chemokine receptor type 4	Homo sapiens (Human)	P61073	3odu	C-X-C Chemokine type 4
Neurotensin receptor type 1	Rattus norvegicus (Rat)	P20789	4grv	Neurotensin type 1
δ-type opioid receptor	Mus musculus (Mouse)	P32300	4ej4	Opioid type D
δ-type opioid receptor	Homo sapiens (Human)	P41143	4n6h	Opioid type D
κ-type opioid receptor	Homo sapiens (Human)	P41145	4djh	Opioid type K

Table A.1 – continued from previous page

GPCR	Species	UniProt Id	PDB Id	Family
μ-type opioid receptor	Mus musculus (Mouse)	P42866	4dkl	Opioid type M
Nociceptin receptor	Homo sapiens (Human)	P41146	4ea3	Opioid type X
Proteinase-activated receptor 1	Homo sapiens (Human)	P25116	3vw7	Proteinase-activated type 1
Class A \ (Rhod)opsin Family				
Rhodopsin	Bos taurus (Bovine)	P02699	1u19	Vertebrate blue/green opsin
Rhodopsin	Todarodes pacificus (Japanese flying squid)	P31356	3ayn	Vertebrate opsin
Class A \ Nucleotide-like Family				
Adenosine receptor A2a	Homo sapiens (Human)	P29274	4eiy	Adenosine type 2
P2Y purinoceptor 12	Homo sapiens (Human)	Q9H244	4ntj	Purinoceptor P2RY12-14
Class A \ Lysosphingolipid and LPA (EDG) Family				
Sphingosine 1-phosphate receptor 1	Homo sapiens (Human)	P21453	3v2y	Sphingosine 1-phosphate Edg-1
Class B Family				
Glucagon receptor	Homo sapiens (Human)	P47871	4l6r	Glucagon 7
Class C Family				
Metabotropic glutamate receptor 1	Homo sapiens (Human)	Q13255	4or2	Metabotropic glutamate type 1 3

APPENDIX B

INTERACTION SITE PREDICTION RESULTS

Table B.1: **Prediction Results of Experiment 1**

GPCR	**Family**	**Predicted Interfaces**
P08588	Beta Adrenoceptors type 1	Q55, Q56, A59, L63, A66, L67, V69, L70, L71, P113, A116, T117, V120, R384
A7BJV8	Beta Adrenoceptors type 1	Q55, Q56, A59, A66, V69, L70, L71, P113, A116, T117, V120, R375
B4PBQ3	Serotonin Insect	A225, S228, V229, L233, I235, L236, V237, A282, I286
D2HHT0	Histamine type 2	S22, V23, V27, I29, L30, I31, P73, A76, L80
Q8HZF6	Beta Adrenoceptors type 3	A10, A18, L19, V21, L22, P65, A68, T69, L72
Q9CRR2	Beta Adrenoceptors type 1	L2, A5, L6, V8, L9, L10, A55, T56, V59
Q6GN84	Angiotensin type 2	Q27, E29, I37, P82, A85, T86, R310, R314, H315
Q8SPN2	Prokineticin receptors	A61, I65, A68, L69, M73, P117, V124, K355
C3Z7B5	Serotonin type 1a	L8, L12, I14, L15, V16, A61, T62, L65
Q8JG07	Alpha Adrenoceptors type 2d	A25, T28, I35, L36, I37, P79, L86, K378
D8PJS9	Kiss receptor (GPR54)	L37, L41, M43, L44, V45, P87, R320, K325

Table B.1 – continued from previous page

GPCR	Family	Predicted Interfaces
A9JRC2	Free fatty acid receptor 3	E8, A9, T12, I13, L19, I20, P62, K305
Q9TST5	Beta Adrenoceptors type 2	I38, L42, V44, L45, P88, A91, S92, L95
Q2QKU5	Tachykinin like 2	L46, L53, P96, R323, R327, R328, L329
A4GZ86	Melanin-concentrating hormone receptors	A28, V32, I36, L40, P84, K306, R311
Q8MJV3	Somatostatin type 5	V45, A48, V49, V53, R318, R322, L324
Q9BMA9	Octopamine type 6	E51, A52, L61, I63, I64, P107, I114
A3QNZ9	Taste 2	T566, I567, A570, L571, L575, P619

Table B.2: **Prediction Results of Experiment 2**

GPCR	Family	Predicted Interfaces
Q9TST6	Beta Adrenoceptors type 1	Y157, L158, T161, S162, F164, R165, S168, L169, L188, W198, R200, R222, A223, A227, V235, R246
P47899	Beta Adrenoceptors type 1	Y157, L158, T161, S162, F164, R165, L169, L188, W198, R200, R222, A223, A227, V235, R246
Q8HZG1	Beta Adrenoceptors type 2	Y121, T125, S126, F128, K129, L133, L152, W162, R164, A187, A191, V199, R210
B8YLW8	Dopamine Vertebrate type 1	Y121, S125, S126, F128, R129, L152, W162, K164, R191, I204, R215
Q60483	Beta Adrenoceptors type 3	Y133, L134, T137, R141, V145, W174, R176, L212, R223
Q4S3N2	Histamine type 2	Y99, L100, T103, R107, L111, R165, I178, R189

Table B.2 – continued from previous page

GPCR	Family	Predicted Interfaces
P30989	Neurotensin type 1	Y167, L168, F174, K175, T178, L179, L198
C3YRD2	Serotonin type 7	Y98, L99, S102, L129, K155, I168, R179
B4HFU8	PRXamide-like	Y162, I163, F169, R170, T173, M174, H222
Q8VGU2	Olfactory 85	Y121, V122, S125, S126, V133, L152
D2HL27	Trace amine type 17	Y131, M132, T135, T142, I208, K219
Q9GL18	Alpha Adrenoceptors type 2b	Y98, S102, S109, V129, A162, R181
C3YFQ1	Histamine type 3	Y106, T110, K114, L137, W146, K192
A7SQR4	GPR74 like	F93, M94, R101, L123, A151, L163
Q28BQ5	Proteinase-activated type 2	Y196, V197, F203, L208, L226

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Sahin, Mehmet Emre
Nationality: Turkish (TC)
Date and Place of Birth: 11.10.1985, Ankara
Marital Status: Married
Phone: 0 536 3234887
Fax: 0 312 5921403

EDUCATION

Degree	Institution	Year of Graduation
B.S.	Computer Engineering Dept., METU	2007
High School	Kocaeli Körfez Fen Lisesi	2003

PROFESSIONAL EXPERIENCE

Year	Place	Enrollment
2007 - Current	Aselsan A.Ş.	Expert Software Engineer

PUBLICATIONS

International Publications

Mehmet Emre Sahin, Tolga Can, and Cagdas D. Son, GPCRsort – Responding to the Next Generation Sequencing Data Challenge: Prediction of G protein-coupled receptor classes using only structural region lengths, OMICS: A Journal of Integrative Biology, 2014 (in press).

Mehmet Emre Sahin, and Tolga Can, Predicting protein-protein interactions from protein sequences using two different classifiers with auto covariance, 4^{th} International Symposium on Health Informatics and Bioinformatics, Ankara, 2009.