GPCRsort—Responding to the Next Generation Sequencing Data Challenge: Prediction of G Protein-Coupled Receptor Classes Using Only Structural Region Lengths

Mehmet Emre Sahin,¹ Tolga Can,¹ and Cagdas Devrim Son²

Abstract

Next generation sequencing (NGS) and the attendant data deluge are increasingly impacting molecular life sciences research. Chief among the challenges and opportunities is to enhance our ability to classify molecular target data into meaningful and cohesive systematic nomenclature. In this vein, the G protein-coupled receptors (GPCRs) are the largest and most divergent receptor family that plays a crucial role in a host of pathophysiological pathways. For the pharmaceutical industry, GPCRs are a major drug target and it is estimated that 60%–70% of all medicines in development today target GPCRs. Hence, they require an efficient and rapid classification to group the members according to their functions. In addition to NGS and the Big Data challenge we currently face, an emerging number of orphan GPCRs further demand for novel, rapid, and accurate classification of the receptors since the current classification tools are inadequate and slow. This study 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 showed 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%. *GPCRsort* is available in the public domain for postgenomics life scientists engaged in GPCR research with NGS: <http://bioserver.ceng.metu.edu.tr/GPCRSort>.

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 (Davies et al., 2007a; Levoye et al., 2006). The structure of a GPCR (Fig. 1) comprises seven highly conserved α -helical transmembrane domains, three intracellular and three extracellular loops, an extracellular N-terminus, and an intracellular C-terminus (Filizola, 2010). 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, which are α -helices spanning the membrane. The second requirement is the ability to interact with a G-protein (Fredriksson et al., 2003).

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 (Cobanoglu et al., 2011). For outside ligands, they act as receptors, and for internal pro-

cesses, 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 (Liu et al., 2009). 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 including cancer, psychiatric, metabolic, and infectious diseases (Liu et al., 2009). This means that there is a large potential in developing therapeutic drugs that could act on GPCRs (van der Horst et al., 2010). For the pharmaceutical industry, GPCRs are a major target. It is estimated that 60%–70% of all medicines in development today target

Departments of ¹Computer Engineering, and ²Biological Sciences, Middle East Technical University, Ankara, Turkey.

FIG. 1. Schematic diagram of a GPCR.

GPCRs (Liu et al., 2009; Peng et al., 2010). It is also pointed out that drugs have still only been developed to affect a very small number of GPCRs (Fredriksson et al., 2003).

Today, a large number of protein sequences are identified as GPCRs; yet their structures and functions are not fully characterized. 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) (Tang et al., 2012). 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.

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 (Rosenbaum et al., 2009; Sanders et al., 2011). This scheme is based upon the GPCR superfamily classification system that was introduced by Kolakowski (1994). This defunct system divides GPCRs into seven families, specified A–F and O, using original standard similarity searches (Davies et al., 2007a). Horn et al. (2003) developed this system 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 was reorganized and the latest version of the GPCRDB contains five classes at the top level: Class A Rhodopsin like, Class B Secretin like, Class C Metabotropic glutamate/pheromone, Vomeronasal receptors (V1R and V3R), and Taste receptors T2R (Vroling et al., 2011). 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 (Jacoby et al., 2006), which includes more than 80% of all human GPCRs (Davies et al., 2008). Class B receptors bind to large peptides (Cardoso et al., 2006). Metabotropic glutamate receptors in Class C bind to glutamate,

which is an amino acid that functions as an excitatory neurotransmitter (Davies et al., 2007a). The group of receptors named as fungal pheromone receptors include GPCRs that bind to pheromones that are used by organisms for chemical communication (Das and Banker, 2006). 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 (Yang and Zhang, 2014). This makes the sequence of the protein as the primary source with which to work. In the past, several methodologies were 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 (Bhasin and Raghava, 2004) 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 (Bhasin and Raghava, 2004). 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. (2008) proposed a strategy to classify GPCRs. Their method, named GPCRTree, uses an alignment-independent classification system based on physical properties of amino acids. It employs principal component analysis to select the 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 subclasses and the time taking calculation at each level for the unused classifiers.

A recent technique, proposed by Cobanoglu et al. (2011), uses sequence-derived motifs to classify GPCRs. The motifs they produce characterize the subfamilies by discovering

receptor–ligand interaction sites. They propose a 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 that 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 very 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) (Fig. 8 in Cobanoglu et al., 2011).

Inoue et al. propose a method, named the Binary Topology Pattern (BTP) method, for the classification of GPCRs (Inoue et al., 2004). Their classifier is similar to the proposed method in this article, 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, we do not need binarization of loop lengths and use region lengths directly in GPCRsort, as described in the next section.

In this study, we propose a method named GPCRsort, which determines the class of a GPCR using its structural properties. Specifically, we use the lengths of the transmembrane and loop regions of the GPCR structure. The method first determines the transmembrane regions of GPCRs, and constructs feature vectors using the lengths of the regions. Random Forest (Breiman, 2001) classifier is employed in the learning and decision parts of the method. The method can predict GPCRs at every level of the GPCR class hierarchy tree. This shows the generality of GPCRsort, as the other techniques are inadequate to make the exact classification. The experimental results show that GPCRsort is very effective and outperforms the current known techniques for GPCR classification. Highest accuracy, 97.3%, is achieved with GPCRsort method when compared to the other techniques under the same experimental conditions. Besides these advantages, running in minutes makes GPCRsort a faster prediction algorithm among others.

Materials and Methods

Problem definition

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

GPCR representation

A GPCR representation can be seen in (Eq. 1).

$$GPCR = (FV, X) \quad (\text{Eq. 1})$$

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 1. A feature vector is defined as a 15-dimensional vector as shown in (Eq. 2).

$$FV = [TM_1, TM_2, TM_3, TM_4, TM_5, TM_6, TM_7, N, L_1, L_2, L_3, L_4, L_5, L_6, C] \quad (\text{Eq. 2})$$

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 (Eq. 2). The sum of entries in the feature vector gives the length of the GPCR represented by that vector.

GPCRDB (Vroling et al., 2011) has a hierarchy of classes and defines a class id for each family in the hierarchy. As mentioned earlier, the class hierarchy starts with the five top-most classes. 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 (Eq. 1), denotes the class id of the lowermost GPCRDB subfamily in which the protein is classified.

Dataset preparation

The protein sequences that comprise the datasets used in the methods are taken from the GPCRDB (Vroling et al., 2011). 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 38,525 proteins that are classified across 1272 families. The protein family members and class descriptions are easily reachable through the web site (<http://www.gpcr.org/7tm>).

Transmembrane regions of GPCRs are predicted using TMHMM stand-alone software package (Krogh et al., 2001). This program is for prediction of transmembrane helices in proteins. It uses a hidden Markov model to predict these regions (Sonnhammer et al., 1998). TMHMM is selected because it has been rated best in an independent comparison of programs for prediction of transmembrane helices (Möller et al., 2001).

Transmembrane regions of all GPCRDB proteins are predicted using the TMHMM program. After the prediction process, the results showed that 29,038 proteins were labeled as having seven transmembrane regions. The feature vectors are constructed for those 29,038 GPCRs and their representations are used in the following experiments. If we look at the datasets used in previous studies, it can be seen that this number is much higher than the dataset sizes they use. For instance, the GPCRpred dataset, which is used in studies GPCRpred (Bhasin and Raghava, 2004) and GPCRBind (Cobanoglu et al., 2011), 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. (2007b), containing 8354 GPCRs. On the other hand, GPCRDB contains 38,525 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 our proposed classifier, GPCRsort. The remaining 25% of the GPCRDB proteins are either fragments or proteins that 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 (Suzek et al., 2007) which is maintained at 90% non-redundancy level. This intersection comprises the second dataset with 10,216 entries. The size of this non-redundant set is still larger than the dataset sizes used in the previous works.

Method

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

$$T' = \{G : G \in T, \text{ and } G = (FV, X)\} \quad (\text{Eq. 3})$$

where X is the class of G

4. Train the Random Forest (Breiman, 2001) classifier using T' and construct the classification model M .
5. Predict transmembrane regions of P using the TMHMM tool and construct its FV_P .
6. Using the model M obtained in the 4th 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 (Breiman, 1996) and the random selection of features (Ho, 1995). By the power of this combination, it constructs a collection of decision trees with controlled variation.

Random Forest was chosen because of its advantages over the other classification methods. First, it is one of the most accurate learning algorithms available (Caruana et al., 2008). It has methods for balancing error in class population unbalanced datasets (Breiman, 2001). 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.

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 (Hall et al.,

2009) 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 paper.

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.

Cross validation experiments

The test dataset is constructed from the available dataset. First, 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.

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 (Bhasin and Raghava, 2004) 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

TABLE 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%)

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, the training set contains 28,204 proteins.

Comparison of GPCRsort 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 re-organized according to the classes described in the BTP method dataset. The BTP method employs HMMTOP (Tusnady and Simon, 2001) to extract loop lengths from GPCRs. To make the exact implementation of the method, the 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.

Comparison with other methods

A new GPCR classification method is proposed in this study. 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. (2011) proposed a method, named GPCRBind, for the GPCR classification problem and compared this method with state-of-the-art GPCR classification

TABLE 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 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.

methods reported by Davies et al. (2007b). These methods use the GDS dataset (Davies et al., 2007b) as training, and the GPCRpred dataset (Bhasin and Raghava, 2004) as testing datasets. To make a perfect comparison, exactly identical datasets are obtained and used in the proposed method to compare its performance to the other methods' performances.

Results

The 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 a total of 75 classes as the subfamilies.

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 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 were 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.

Cross validation experiments

Tables 2, 3, and 4 contain performance measures. Before going into details of the results of experiments, it would be

TABLE 4. EVALUATION RESULTS OF USING 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 5. CLASSIFICATION PERFORMANCE OF METHOD USING GPCRpred DATASET AS VALIDATION DATA

<i>Subfamily</i>	<i>Total</i>	<i>Predicted</i>
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%)

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 the 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 (Powers, 2008).

TABLE 7. CLASSIFICATION PERFORMANCE OF GPCRsort AND BTP METHOD

<i>Subfamily</i>	<i>Total</i>	<i>GPCRsort</i>	<i>BTP</i>
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/Smoothed	381	330 (86.6%)	0 (0%)
Total	9315	7939 (85.2%)	4432 (47.6%)

*Class definitions are taken from Inoue et al., (2004).

As the first step in the 10-fold cross validation studies, the whole dataset is used to construct training and testing datasets. The results are presented in Table 2. Results show that the accuracy is very high at 90.54%. High true positive rates and low false positive rates 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 3 lists the results of this second step. The 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.

TABLE 6. CONFUSION MATRIX OF METHOD USING GPCRpred DATASET AS VALIDATION DATA

[illegible]

TABLE 8. CLASSIFICATION PERFORMANCE OF GPCRsort COMPARED WITH CURRENT KNOWN METHODS

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

The classification results for each GPCR are presented in Supplementary Table S1. Table S1 also contains the probability distribution across classes while classifying each GPCR.

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 4 and 5. The confusion matrix of the method using GPCRpred dataset as the testing data can be seen in Table 6. Table 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.

Comparison with the BTP method

The performance of GPCRsort with several datasets is analyzed in the first two experiments. The next experiments

compare the GPCRsort with the existing GPCR classification methods. First comparison is with a similar method to GPCRsort: BTP method. Table 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%. Furthermore, 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.

Comparison with other methods

We compare GPCRsort with a state of the art classifier, GPCRBind, which has been shown to outperform several other classifiers (Cobanoglu et al., 2011). The accuracy results of this comparison are listed in Table 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 such a comparison environment.

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 that it can be run on any PC, not requiring a server to run. Compared to GPCRBind, this is a significant

FIG. 2. Comparison of running times of methods. GPCRBind running time is taken from Cobanoglu et al., (2011).

improvement in running time, since GPCRBind needs hours to construct the sequence motifs for GPCR class prediction (Cobanoglu et al., 2011). Figure 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.

Discussion

We propose a new GPCR classification method, 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 (Vroling et al., 2011). 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 article.

GPCRsort can be used to classify uncharacterized GPCRs and direct further biological studies accordingly (Supplementary Table S1). 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 (Pevzner and Shamir, 2011). 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 study, and the overall accuracy can be further improved.

GPCRsort shines among other approaches in the comparison experiments. 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.

Conclusion

In this article, we present a novel 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 special-

ized 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.

Acknowledgments

This work is in part supported by the Middle East Technical University Internal Research Fund [BAP-07-02-2010-02]; the Scientific and Technological Research Council of Turkey [110T414]; and International Reintegration Grants Marie Curie Actions [PIRG07-GA-2010-268336]. The authors also wish to acknowledge M.C. Cobanoglu and U. Sezerman for providing datasets for comparison purposes.

Author Disclosure Statement

The authors declare that there are no conflicting financial interests.

References

- Bhasin M, and Raghava GP. (2004). GPCRpred: An SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. *Nucleic Acids Res* 32, 383–389.
- Breiman L. (1996). Bagging Predictors. *Mach Learn* 24, 123–140.
- Breiman L. (2001). Random Forests. *Mach Learn* 45, 5–32.
- Cardoso JC, Pinto VC, Vieira FA, Clark MS, and Power DM. (2006). Evolution of secretin family GPCR members in the metazoa. *BMC Evol Biol* 6, 108.
- Caruana R, Karampatziakis N, and Yessenalina A. (2008). An empirical evaluation of supervised learning in high dimensions. *Proc Int Conf Mach Learn* 25, 96–103.
- Cobanoglu MC, Saygin Y, and Sezerman U. (2011). Classification of GPCRs using family specific motifs. *IEEE/ACM Trans Comput Biol Bioinform* 8, 1495–1508.
- Das SS, and Banker GA. (2006). The role of protein interaction motifs in regulating the polarity and clustering of the metabotropic glutamate receptor mGluR1a. *J Neurosci* 26, 8115–8125.
- Davies MN, Gloriam DE, Secker A, et al. (2007a). Proteomic applications of automated GPCR classification. *Proteomics* 7, 2800–2814.
- Davies MN, Secker A, Freitas AA, Mendao M, Timmis J, and Flower DR. (2007b). On the hierarchical classification of G protein-coupled receptors. *Bioinformatics* 23, 3113–3118.
- Davies MN, Secker A, Halling-Brown M, et al. (2008). GPCRTree: Online hierarchical classification of GPCR function. *BMC Res Notes* 1, 67.
- Filizola M. (2010). Increasingly accurate dynamic molecular models of G-protein coupled receptor oligomers: Panacea or Pandora's box for novel drug discovery? *Life Sci* 86, 590–597.
- Fredriksson R, Lagerström MC, Lundin LG, and Schiöth HB. (2003). The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol Pharmacol* 63, 1256–1272.
- Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, and Witten IH. (2009). The WEKA data mining software: An update. *SIGKDD Explor* 11, 10–18.

- Ho TK. (1995). Random Decision Forest. *Proc Int Conf Doc Anal Recognit* 3, 278–282.
- Horn F, Bettler E, Oliveira L, Campagne F, Cohen FE, and Vriend G. (2003). GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res* 31, 294–297.
- Inoue Y, Ikeda M, and Shimizu T. (2004). Proteome-wide classification and identification of mammalian-type GPCRs by binary topology pattern. *Comput Biol Chem* 28, 39–49.
- Jacoby E, Bouhelal R, Gerspacher M, and Seuwen K. (2006). The 7 TM G-protein-coupled receptor target family. *Chem-MedChem* 1, 761–782.
- Krogh A, Larsson B, von Heijne G, and Sonnhammer ELL. (2001). Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J Mol Biol* 305, 567–580.
- Kolakowski LF Jr. (1994). GCRDB: A G-protein-coupled receptor database. *Receptors Channels* 2, 1–7.
- Levoye A, Dam J, Ayoub MA, Guillaume JL, and Jockers R. (2006). Do orphan G-protein-coupled receptors have ligand-independent functions? New insights from receptor heterodimers. *EMBO Rep* 7, 1094–1098.
- Liu X, Kai M, Jin L, and Wang R. (2009). Computational study of the heterodimerization between mu and delta receptors. *J Comput Aided Mol Des* 23, 321–332.
- Möller S, Croning MDR, and Apweiler R. (2001). Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* 17, 646–653.
- Papasaïkas PK, Bagos PG, Litou ZI, and Hamodrakas SJ. (2003). A novel method for GPCR recognition and family classification from sequence alone using signatures derived from profile hidden Markov models. *SAR QSAR Environ Res* 14, 413–420.
- Pevzner P, and Shamir R. (2011). *Bioinformatics for Biologists*. Cambridge University Press, New York.
- Powers DMW. (2008). Evaluation Evaluation: A Monte Carlo study. *ECAI 2008, Front Artif Intell Appl*, IOS Press 178, 843–844.
- Peng ZL, Yang JY, and Chen X. (2010). An improved classification of G-protein-coupled receptors using sequence-derived features. *BMC Bioinform* 11, 420–432.
- Rosenbaum DM, Rasmussen SGF, and Kobilka BK. (2009). The structure and function of G-protein-coupled receptors. *Nature* 459, 356–363.
- Sanders MPA, Fleuren WWM, Verhoeven S, et al. (2011). ss-TEA: Entropy based identification of receptor specific ligand binding residues from a multiple sequence alignment of class A GPCRs. *BMC Bioinform* 12, 332–344.
- Sonnhammer ELL, von Heijne G, and Krogh A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol* 6, 175–182.
- Suzek BE, Huang H, McGarvey P, Mazumder R, and Wu CH. (2007). UniRef: Comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 23, 1282–1288.
- Tang XL, Wang Y, Li DL, Luo J, and Liu MY. (2012). Orphan G protein-coupled receptors (GPCRs): Biological functions and potential drug targets. *Acta Pharmacol Sin* 33, 363–371.
- Tusnady GE, and Simon I. (2001). The HMMTOP transmembrane topology prediction server. *Bioinformatics* 17, 849–850.
- van der Horst E, Peironcelly JE, Ijzerman AP, et al. (2010). A novel chemogenomics analysis of G protein-coupled receptors (GPCRs) and their ligands: A potential strategy for receptor de-orphanization. *BMC Bioinform* 11, 316–327.
- Vroling B, Sanders M, Baakman C, et al. (2011). GPCRDB: Information system for G protein-coupled receptors. *Nucleic Acids Res* 39, 309–319.
- Yang J, and Zhang Y. (2014). GPCRSD: A database for experimentally solved GPCR structures. <http://zhanglab.cmb.med.umich.edu/GPCRSD>. Last access: July 2, 2014.

Address correspondence to:

Dr. Cagdas Devrim Son

Department of Biological Sciences, Room 227

Middle East Technical University

Universiteler Mah. Dumlupinar Boulevard

No: 1 Cankaya

Ankara, 06800

Turkey

E-mail: cson@metu.edu.tr

Abbreviations Used in Text

BTP	= Binary Topology Pattern method
GDS dataset	= BIAS-PROFS GPCR dataset
GPCR	= G protein-coupled receptor
GPCRDB	= G protein-coupled receptor database
GPCRpred	= method for predicting families of GPCRs
NCBI	= National Center for Biotechnology Information
NR database	= non-redundant database
TMHMM	= membrane protein topology prediction method
UniRef	= UniProt Reference Clusters
AMN	= amine family
CAN	= cannabinoid family
FP	= false positive rate
GPCR	= G protein-coupled receptor
GRH	= gonadotropin-releasing hormone family
HMP	= hormone protein family
LEU	= leukotriene B4 receptor family
LYS	= lysosphingolipid and LPA (EDG) family
MEL	= melatonin family
NUC	= nucleotide-like family
OLF	= olfactory family
PAF	= platelet activating factor family
PEP	= peptide family
PRS	= prostanoid family
RHD	= rhodopsin family
ROC area	= area under Receiver Operating Characteristics diagram
TM	= transmembrane
TRH	= thyrotropin-releasing hormone and secretagogue family
VIR	= viral family
HMM	= Hidden Markov Model