Bioinformatics Advance Access published October 22, 2007

Proteomics

On the hierarchical classification of G Protein-Coupled Receptors

Matthew N. Davies[1,*], Andrew Secker[2], Alex A. Freitas[2], Miguel Mendao[3], Jon Timmis[3] and Darren R. Flower[1]

[1]Edward Jenner Institute, Compton, Newbury, Berkshire, RG20 7NN, U.K.
[2]Department of Computing and Centre for BioMedical Informatics, University of Kent, Canterbury, Kent CT2 7NF, U.K.
[3]Departments of Computer Science and Electronics, University of York, Heslington, York YO10 5DD, U.K.

Associate Editor: Prof. John Quackenbush

ABSTRACT

Motivation

G protein-coupled receptors (GPCRs) play an important role in many physiological systems by transducing an extracellular signal into an intracellular response. Over 50% of all marketed drugs are targeted towards a GPCR. There is considerable interest in developing an algorithm that could effectively predict the function of a GPCR from its primary sequence. Such an algorithm is useful not only in identifying novel GPCR sequences but in characterising the interrelationships between known GPCRs.

Results

An alignment-free approach to GPCR classification has been developed using techniques drawn from data mining and proteochemometrics. A dataset of over 8,000 sequences was constructed to train the algorithm. This represents one of the largest GPCR datasets currently available. A predictive algorithm was developed based upon the simplest reasonable numerical representation of the protein's physicochemical properties. A selective top-down approach was developed which used a classifier to assign sequences to subdivisions within the GPCR hierarchy. The predictive performance of the algorithm was assessed against several standard data mining classifiers and further validated against Support Vector Machine-based GPCR prediction servers. The selective top-down approach achieves significantly higher accuracy than standard data mining methods in almost all cases.

Contact

m.davies@mail.cryst.bbk.ac.uk

1 INTRODUCTION

1.1 The G Protein-Coupled Receptors

The G protein-coupled receptors (GPCR) are composed of a diverse range of integral membrane proteins that regulate many important physiological functions (Christopoulos *et al*., 2002; Gether *et al*., 2002; Bissantz, 2002). GPCRs control and/or affect processes as diverse as neurotransmission, cellular metabolism, secretion, cellular differentiation and inflammatory responses (Hebert *et al*., 1998). The binding of a ligand on the cell surface causes the GPCR to become active, and subsequently bind and activate ubiquitous guanine nucleotide-binding regulatory (G) proteins within the cytosol. An extremely heterogeneous set of molecules can act as GPCR ligands including ions, hormones, neurotransmitters, peptides and proteins. The GPCRs are a common target for therapeutic drugs and approximately 50% of all marketed drugs target GPCRs (Flower, 1999; Klabunde *et al*., 2006). In spite of their functional and sequence diversity, there are certain structural features common to all GPCRs. All GPCRs contain seven highly conserved transmembrane segments. The sequences also contain three extracellular loops (EL1-3), three intracellular loops (IL1-3) as well as the protein N and C termini. The transmembrane segments form seven α-helices in a flattened two-layer structure known as the transmembrane bundle, a structure seen in all GPCRs (Milligan, 2006). The GPCRs shows a far greater conservation with regard to the three-dimensional structure than to the primary sequence.

The diversity of the GPCRs means it is difficult to develop a comprehensive classification system for all of the GPCR subtypes (Davies *et al*., 2007). One of the first GPCR classification systems was introduced by Kolakowski for the now defunct GCRDb database (Kolakowski, 1994). GPCRs were divided into seven groups, designated A-F and O, derived from original standard similarity searches. This system was further developed for the GPCRDB database (Horn *et al*., 2003), which divides the GPCRs into six classes. These are the Class A Rhodopsin-like, which account for over 80% of all GPCRs in humans, Class B Secretin-like, Class C Metabotropic glutamate receptors, Class D Pheromone receptors, Class E cAMP receptors and the Class F Frizzled/smoothened family. There are at least 286 human non-olfactory Class A receptors, the majority of which bind peptides, biogenic amines or lipid-like substances (Fridmanis *et al*., 2006). Class B receptors bind large peptides such as secretin, parathyroid hormone, glucagon, calcitonin, vasoactive intestinal peptide and pituitary adenylyl cyclase activating protein (Cardoso *et al*., 2006). Class C Metabotropic glutamate receptors (mGluRs) are a type of glutamate receptor that are activated through an indirect metabotropic process (Das *et al*., 2006). There are two other GPCR families that are considerably smaller. Class D is composed of pheromone receptors, which are used for chemical communication (Nakagawa *et al*., 2005) while Class E, the cAMP receptors, form part of the chemotactic signalling system of slime molds (Prabhu *et al*., 2006). There is also an additional minor class, the Frizzled/Smoothened receptors, which are necessary for Wnt binding and the mediation of hedgehog signalling, a key regulator of animal development (Foord *et al*., 2002). The six different classes can be further divided into sub-families and sub-subfamilies based upon the

*To whom correspondence should be addressed.

function of a GPCR and the specific ligand to which it binds. In this paper, the following terminology is used to describe the classification of GPCR sequences. The six major GPCR families are termed 'Classes', the secondary level of classification is termed 'Sub-families' and the third level of classification is termed 'Sub-subfamilies'. Not all human GPCRs can be effectively classified using this system, there are approximately 60 "orphan" GPCRs that show the sequence properties of Class A Rhodopsin-like receptors yet have no defined ligands or functions (Gloriam *et al*., 2005). It is possible that many of these orphan receptors have ligand-independent properties, for example, the regulation of ligand-binding GPCRs on the cell surface.

1.2 GPCR Prediction Servers

Previous attempts at predicting the function of a GPCR from its primary sequence, and therefore its position within a given hierarchical system, have included motif-based classification tools (Attwood, 2001; Flower *et al*., 2002) and machine learning methods such as Hidden Markov Models (Wistrand *et al*., 2006). These approaches have applications not only in discovering and characterising novel protein sequences but also in better understanding relationships between known GPCRs. The majority of predictive techniques, however, have used Support Vector Machines (SVMs) (Karchin *et al*., 2002), machine-learning algorithms based on statistical learning theory. In two-class problems, a SVM maps the input vectors (data points representing protein descriptions) into a higher dimensional feature space and then constructs the optimal hyperplane to separate the classes, while avoiding overfitting. This is a powerful form of classification because, although it is linear in the higher dimensional feature, it is non-linear in the original attribute space of the input vectors. The optimal hyperplane is the one with a maximum distance to the closest data point from each of the two classes. The distance is called the margin, and the optimal hyperplane is called the maximal margin hyperplane. The input vectors closest to the optimal hyperplane are called the support vectors. Although SVM are more commonly used to solve 2-class problems, this technique can be applied to the classification of GPCR data by successively trying to classify one class against all others (Karchin *et al*., 2002).

Several publicly available SVM-based GPCR classifiers exist. PRED-GPCR (http://athina.biol.uoa.gr/bioinformatics/PRED-GPCR/) (Papasaikas *et al*., 2004; Guo *et al*., 2005) was developed as a fast fourier transform with SVMs on the basis of the hydrophobicity of the amino acid sequence. Quantitative descriptions of the proteins relating to hydrophobicity, bulk and electronic properties were derived from the hydrophobicity model, composition-polarity-volume (c-p-v) model and the electron-ion interaction potential (EIIP) model. Three different hydrophobicity scales - the Kyte-Doolittle Hydrophobicity (KDHΦ), Mandell Hydrophobicity (MHΦ) and Fauchére Hydrophobicity (FHΦ) - were used. The sequences are transformed, first, into numerical representations of the sequence based upon the EIIP values and, second, into the frequency domain using the discrete Fourier transform, a method by which sequences of different length can be normalised. The output of these transformations is used as the input for the SVM. In the case of an n-class classification problem, where $n > 2$, as is the case for the GPCR families, each ith SVM, i=1,…,n, is trained. When using the FHΦ hydrophobicity scale, the technique achieved a reported accuracy of 93.3% and a Matthew's correlation coefficient of 0.95. However, the range of accuracies between the subfamilies varied between 66.7 and 100% (Papasaikas *et al*., 2004).

GPCRPred is another SVM-based classifier that determines whether a sequence is or is not a GPCR; if it is a GPCR, to which class it belongs; and then, if it is a Class A protein, to which subfamily it belongs (Bhasin *et al*., 2004). The vectors are based upon the dipeptide composition, whereby each of the 400 possible pairs of amino acids is associated with a vector component representing the percentage of the primary sequence consisting of that pair. Again, the one-vs-rest SVM is used to characterise each Class and subfamily. The program was reported as having a 99.5% predictive accuracy at the GPCR vs non-GPCR level, 97.3% accuracy at the Class level and 85% accuracy at the subfamily level. A third server, GPCRsclass (Bhasin *et al*., 2005), concentrates on the Class A aminergic receptor subfamily. In the first round of analysis, an SVM is generated to distinguish amines from all other GPCRs. Then multiclass SVMs are set up to classify amines into the acetylcholine, adrenoreceptor, dopamine and serotonin subgroups. The SVM requires patterns of fixed length for training and testing. The sequences are transformed to fixed length format by measuring the amino acid and dipeptide compositions, giving vectors of 20 and 400 dimensions, respectively. The dipeptide composition has proved to be far more reliable than the amino acid, scoring 99.7% accuracy at discriminating amine from non-GPCRs and 92% accuracy at discriminating between the four sub-subfamilies. A similar method involving amino acid, dipeptide and tripeptide compositions (Guo *et al*., 2006) claimed 98% accuracy at the Class level. GPCRsclass gave 94% accuracy at the Class level when tested with the same dataset.

Here, a new selective top-down approach using a hierarchical classifier is applied to GPCR classification. The technique was validated, first, against standard data mining classifiers and, second, against several SVM-based GPCR predictive servers.

2 METHODS

2.1 GPCR DATASET (GDS)

In order to develop an effective algorithm for the classification of GPCR sequences, it was necessary to build as large and comprehensive a dataset of GPCR sequences as possible with which to train and test the classifier. Protein sequences for the dataset were identified using the Entrez search and retrieval system (Wheeler *et al*., 2007). The system searches protein databases such as SwissProt, PIR, PRF, PDB, as well as translations from annotated coding regions in DNA databases, such as GenBank and RefSeq. Text-based searching was used to identify all sequences within each sub-subfamily of the hierarchy. These composite groups were then used to build each GPCR sub-family and Class level dataset. The datasets contain only human protein sequences, with the exception of Class D proteins, which are found only in fungi and Class E, which are found in *Dictyostelium*. All proteins shorter than 280 amino acids in length were removed in order to eliminate incomplete protein sequences, and all identical sequences within the dataset were removed to avoid redundancy. This left 8354 protein sequences in 5 classes at the family level (A-E), 40 classes at the sub-family level, and 108 classes at the sub-subfamily level. Class F was not considered as it contains too few sequences from which to develop an accurate

classification algorithm (this class has also been excluded from the PRED-GPCR and GPCRPred predictive programs). For the sake of convenience, this dataset will be referred to as the GDS (GPCR Data Set).

2.2 Sequence Representation

Rather than using the primary sequence to perform the classification, the system uses an alternative form of protein data representation. Alignment-independent classification systems use the physiochemical properties of amino acids to determine differences between protein sequences. Proteochemometrics is a technique whereby 5 "z-values" (z1-z5) are derived from 26 real physiochemical properties through the application of principal component analysis (Sandberg *et al*., 1998, Lapinsh *et al*., 2002). The z1 value accounts for the amino acid's lipophilicity, the z2 values account for steric properties, such as bulk and polarisability, and the z3 value describes the polarity of the amino acid. The electronic components of the amino acids are described by the z4 and z5 values. These five values are calculated for each amino acid in the sequence, generating a matrix that provides a purely numerical description of the protein's character. Several sequences in the GDS contain non-standard amino acid codes that are not present in the table of z-values. In such cases, the following substitutions were made. Where the sequence contained a 'B' (either an asparagine or aspartic acid) the residue was assigned as an asparagine 'N'. Where the sequence contained a 'Z' (i.e. either a glutamine or a glutamic acid), the residue was assigned as a glutamine 'Q'. Where the sequence contained a 'U', indicating selenocysteine, the sequence was changed to cysteine 'C'. All unknown residues 'X' were given as alanines 'A'.

The data mining algorithms used cannot cope with variable numbers of predictor attributes. It is therefore essential to normalise these values such that each protein is described by a set number of predictor attributes. Normalisation of sequences has previously been carried out using Auto Cross Covariance (ACC) (Wold *et al*., 1993). In previous work (Secker *et al.,* 2007) we described a normalization method where the arithmetic mean for each z value is computed over the whole protein. This was found to retain predictive accuracy while significantly reducing processing time and storage requirements, compared to ACC. For each attribute (z-value) x, the mean value for that attribute $\overline{x}$ is the mean of the values of that attribute in a protein over all amino acids (a) where the total number of amino acids in the protein is represented as N.

$$\overline{x} = \frac{1}{N}\sum_{a=1}^{N} x_a$$

The equation above is therefore applied five times, once for each attribute, where each attribute corresponds to a z-value.

In this investigation, we use an augmented version of this attribute creation method. In this case, 15 attributes are used to describe each protein. Five are created as described above but in addition to this, five more are created from the N-terminus of the protein while a further five are created using the C-terminus. The termini of a GPCR protein have the ability to be powerful predictors of function since the ends of the GPCR will be involved in either intracellular or extracellular binding. Therefore, in the case of the N-terminus, the means for each of the five z-values are computed for only the first 150 amino acids while in the case of the C-terminus, the means over the last 150 amino acids are determined. In reality, the actual lengths of the N and C-termini will vary between GPCRs; the value of 150 amino acids was found, in controlled experiments, to give in the largest improvement in predictive accuracy.

2.3 Classification Algorithm

In order for the algorithm to be effective, it must be able to predict protein function based on an established classification system for the GPCRs. The GPCRDB database suggests a workable hierarchy for GPCR sequences and so it is the one used by GDS (although alternative hierarchies exist, such as the GRAFS Classification system (Schiöth *et al*., 2005)). In the data mining literature, there exists a range of strategies for predicting hierarchical classes (Freitas *et al*., 2007). The simplest is to flatten the dataset to the most specific level of the hierarchy, then use one of the plethora of standard classification algorithms to predict to what class each sequence belongs. However, this strategy does not take advantage of the information implicit in the class structure. An alternative is the so-called "big bang" approach, which uses a single, and typically complex, hierarchical classification algorithm. In the test phase, each example may be assigned to one class at each level of the hierarchy by one single application of the learned model. Perhaps due to its complexity, implementations of such an approach are scarce, although such a model has been used to predict gene function in *Saccharomyces Cerevisiae* (Claire *et al*., 2003).

A middle ground between these two strategies is the top-down approach, where the hierarchical classification process is converted into a number of flat classification problems that may be solved independently by running a standard classifier for each (Freitas *et al*., 2007; Costa *et al*., 2007). The advantage with this strategy over the others is, as is the case with flat classification, no special classifier must be written to perform the task (other than the scaffolding required to support a classifier tree). The structure of the tree aids the classifier and reduces the number of classes that must be considered at the most specific level (see Figure 1).

The standard top-down approach proceeds as follows. Given, for example, the class tree in Figure 1 (a), a tree of classifiers is built to reflect the structure of the classes, as shown in Figure 1 (b). Thus a tree of classifiers is generated such that the output of one classifier constitutes the input for another. To train the classifiers in the hierarchy, all data in the training set is used to train the root classifier while only the relevant subsets of data are used to train at the levels of the subfamily and the sub-subfamily. When an unknown sequence is presented to the classifier tree, the root level classifier will assign it a class then pass it down to the appropriate classifier at the next level until the sequence is assigned to a subfamily and a sub-subfamily.

A novel version of the top-down approach was developed and used as the chosen strategy for classifying the GDS. The top-down approach takes advantage of the hypothesis that some characteristics may be important to discern between two protein subsets at one classification level while being less important at another. The top-down approach exploits this, as any classifier in the tree is trained using only data instances of the classes they are required to classify between. In the standard top-down approach the same classification algorithm is used in each node in

the class tree. It is, however, possible that different classifiers may be more suited to different nodes in the class tree and that therefore the classification accuracy may be increased by using different algorithms in the classifier hierarchy. Importantly, these classifiers are selected in a data-driven manner using training data. This is referred to as the selective top-down approach.

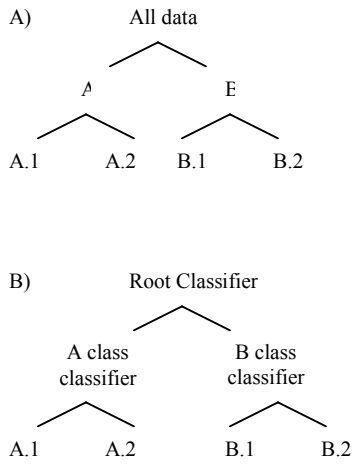

Figure 1. Example of a hierarchical dataset (a) and how that hierarchy may be reflected in a tree of classifiers (b) ready for a top-down approach to classification.

The selective top-down approach generates a tree of classifiers in a similar manner to the standard top-down approach but with some additional stages during training.. At each node, the training data for that node is split into sub-training and validation sets, with data instances being assigned randomly. A number of different classifiers are then trained using this sub-training data and tested using the validation set. The classifier that yields the highest classification accuracy in the validation set is selected as the classifier for this node in the class tree. The sub-training and validation sets are then merged to produce the original training set (for that node), and the selected classifier is then re-trained. Eight standard classification techniques were used (Witten *et al*., 2005). These were the Naïve Bayes method, a Bayesian network, an SVM (Keerthi *et al*., 2001), nearest neighbour (using Euclidean distance), a decision list (Frank *et al*., 2007), J48 (a decision tree algorithm much like C4.5), a Naïve Bayes tree (a decision tree with a naïve Bayes classifier at each node), AIRS2 (a classifier based on the Artificial Immune System paradigm (Watkins *et al*., 2004)) and a conjunctive rule learner. This list of classifiers was carefully chosen to include a wide range of paradigms. All code was written using the WEKA data mining package (Brownlee, 2006) and the default parameters were used for each algorithm.

3 RESULTS

Two separate studies were undertaken to assess the quality of the selective top-down technique. The first (Section 3.1) was to compare the effectiveness of the approach in comparison with the standard top-down technique. The second (Section 3.2) tests the accuracy of the algorithm in comparison with three publicly available GPCR classifiers and tests it against datasets that have been used to train the three servers. Where accuracies are reported for each level, the accuracy is computed as the percentage of correct classifications at that level.

3.1 Cross-validation experiments

In order to compare the quality of the selective top-down classifier, it was tested on the prepared GDS dataset. All experiments were carried out using a 10-fold cross-validation method. As data instances are added randomly to each fold, each test was repeated 30 times and the mean values are reported. Whilst data instances were randomly assigned to folds, care was taken to ensure that at least one instance of each class was present in each fold. For this reason the decision was taken that any class containing fewer than 10 examples was discarded for this test. This left 87 classes at sub-subfamily level, 38 at the sub-family level and 5 classes at the family level. In total, 8222 proteins remained in the dataset. When training the selective top-down classifier, each of the 9 classifiers was trained using 80% of the training data (sub-training set) available to that node, and evaluated using the remaining 20% (validation set).

To validate the algorithm, results for a standard top-down approach are shown for each classifier that the selective top-down algorithm has a choice between. A value denoting the significance of the difference between the accuracy of the selective approach and each particular algorithm was computed using the corrected resampled t-test (Witten *et al*., 2005). This test attempts to eliminate the issues encountered when a standard t-test is used over multiple runs of a cross-validation procedure. In Table 1, a shaded cell indicates that the corresponding accuracy value of the selective top-down classifier is significantly greater than the shaded value. The significance threshold was set at 1% and a 2-tailed test was used.

The results show that the novel selective top-down approach compares favourably with the standard top-down approach and in almost all cases surpasses these established data mining techniques. The nearest neighbour classifier was the classifier predominantly chosen by the selective approach at the top level and as such it is no surprise that there is no statistically significant difference between the nearest neighbour classifier at this level. One disadvantage of the top-down approach (both the selective and standard types) is that any example misclassified at one level has no possibility of being correctly classified at deeper levels and therefore misclassifications can be seen to accrue as the level depth increases.

Table 1: Predictive accuracy (%) of the selective top-down technique at each level compared against several standard classifiers. A shaded cell indicates that the corresponding accuracy value of the selective top-down classifier is significantly greater than the shaded value.

Level	Family	Sub-family	Sub-sub-family
Selective top-down	95.87%	80.77%	69.98%
Naïve Bayes	77.29%	52.60%	36.66%
Bayesian Network	85.54%	64.27%	50.69%

SMO	80.21%	56.67%	35.96%
Nearest Neighbour	95.87%	78.68%	69.40%
PART	93.27%	78.73%	65.68%
J48	92.93%	77.49%	64.30%
Naïve Bayesian Tree	93.07%	76.92%	64.78%
AIRS2	91.98%	74.58%	62.68%
Conjunctive Rules	76.19%	49.93%	16.49%

3.2 Empirical comparison with GPCR classification servers

While there is evidence that the novel selective top-down approach may lead to better classification accuracy compared with standard top-down classifiers, it is important to validate the novel approach by testing with other datasets and against other classifiers specific for GPCR prediction. The PRED-GPCR, GPCRpred and GPCRsclass servers, all three of which are publicly available, were selected for this purpose. Additionally, the datasets that were used to train and test the three servers were kindly supplied by their authors. The GPCRPred dataset is composed of 1008 Class A sequences, 56 Class B, 16 Class C, 11 Class D and 3 Class E, making a total of 1096 sequences. The PRED-GPCR program was trained using 403 sequences from 17 sub-families from GPCR Classes B, C, D and F. GPCRsclass dataset is composed of four amine sub-subfamilies, 31 Acetylcholine sequences, 44 Adrenoreceptors, 38 Dopamine and 54 Serotonin, making a total of 167 sequences. For a full assessment of the technique, it was necessary to run all of the datasets against the developed algorithm and the three servers.

For this test, the selective top-down classifier was trained using the full GDS dataset (8354 protein sequences) then tested using each of the GPCR server datasets as test data. This simulates the situation in which the selective top-down approach, trained with the GDS dataset, could be deployed as a public server. The predictive accuracy at each level of the hierarchy is shown in Table 2. A separate sub-table is displayed for each dataset so the quality of the classification can be directly compared between each server. In the experiments, every classification method has been tested using every dataset and the resultant classification accuracies are presented. For the sake of completeness, each sub-table includes the instances where a method has been tested using its own dataset although it is acknowledged that these values are of limited use has been trained and tested using the same data. Rows in the table where this occurs have been italicised, as the figures contained in this row will represent results heavily biased in favour of that particular classifier.

The selective top-down approach generally exceeds PRED-GPCR at the Class level and is comparable at the Sub-family level. Both the selective top-down and PRED-GPCR are shown to be notably better than GPCRPred at all levels of the hierarchy. GPCRsclass was the most successful classifier at the most specific level but this is likely to be due to the fact that the classifier can only be applied at the sub-subfamily level and is therefore highly specialised. The other classifiers, however, have to classify at all three levels and in the case of the selective top-down classifier, accuracy at the sub-subfamily level will suffer from misclassification at the Class and Subfamily stage.

Table 2. Benchmark results of the GPCR datasets comparing the GPCR servers against the Selective Top-down Approach.

GDS dataset

Server	Class	Sub-family	Sub-sub-family
Selective top-down	*99.6%*	*91.8%*	*87.0%*
PRED-GPCR	73.2%	72.2%	67.6%
GPCRpred	64.7%	46.1%	-
GPCRsclass	-	-	94.0%

PRED-GPCR dataset

Server	Class	Sub-family	Sub-sub-family
Selective top-down	96.3%	85.7%	76.6%
PRED-GPCR	*95.1%*	*95.1%*	*94.5%*
GPCRpred	70.1%	55.6%	-
GPCRsclass	-	-	83.0%

GPCRpred dataset

Server	Class	Sub-family	Sub-sub-family
Selective top-down	92.1%	76.2%	57.8%
PRED-GPCR	80.7%	73.8%	59.9%
GPCRpred	*87.2%*	*67.1%*	-
GPCRsclass	-	-	100.0%

GPCRSCLASS dataset

Server	Class	Sub-family	Sub-sub-family
Selective top-down	100.0%	82.3%	78.1%
PRED-GPCR	100.0%	100.0%	92.8%
GPCRpred	65.2%	59.7%	-
GPCRsclass	-	-	*82%*

4 CONCLUSION

The classification of GPCR sequences has proven difficult for conventional bioinformatics classification approaches such as sequence similarity or the identification of specific motifs. However, the structural and functional consistency of GPCR proteins suggests that there is an overall conservation of certain key properties that are necessary to maintain the transmembrane bundle that characterises the group. The effectiveness of proteochemometrics for this type of analysis has already been demonstrated by previous research. However, this is the first time an alignment-free approach has been used on a dataset of this size. A straightforward representation was used that was a development over previously published work. While it appeared to work well, we expect that other more complex representations will be necessary as the work is extended to other problems in bioinformatics. The advantages of the selective top-down approach over standard ("flat" classification) data mining techniques and the current GPCR servers is clearly demonstrated by the accuracies achieved. It demonstrates that each stage of the classification problem is dependent on unique criteria.

Any supervised learning (classification) algorithm has intrinsic limitations. For example, a classification model constructed from a particular training set will only have good predictive accuracy on a test set if that set has the same (or at least similar) probability distribution to the training set. If an unusual class distribution in the training set was used to build a classification model, it is unlikely that the model would have a very high predictive accuracy if applied to a large set of GPCR sequences with a more usual class distribution. Both PRED-GPCR and GPCRPred struggled to accommodate the full diversity of the GDS, while the selective top-down approach proved to be adaptable to both a generalised dataset (PRED-GPCR) and a specialised one (GPCRsclass).

ACKNOWLEDGEMENTS

The authors should like to gratefully acknowledge funding under the ESPRC grant EP/D501377/1. The authors also wish to extend their thanks to G.P.S. Raghava and P.K. Papasaikas for providing their datasets for the purposes of comparison. We are also deeply indebted to Professor Teresa K Attwood for her detailed critique of the manuscript.

REFERENCES

Attwood,T.K. (2001) A compendium of specific motifs for diagnosing GPCR subtypes. *Pharmacol Sci.*, 22,162-5.

Attwood,T.K. *et al.* (2002) PRINTS and PRINTS-S shed light on protein ancestry. *Nucleic Acids Res,.* 30, 239-41.

Bhasin,M. and Raghava,G.P. (2004) GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. *Nucleic Acids Res.*, 32, W383-9.

Bhasin,M. and Raghava,G.P. (2005) GPCRsclass: a web tool for the classification of amine type of G protein-coupled receptors *Nucleic Acids Res.*, 33, W143-7.

Bissantz,C. (2003) Conformational changes of G protein-coupled receptors during their activation by agonist binding. *J Recept Signal Transduct Res.,* 23, 123-153.

Brownlee,J. (2007). WEKA Classification Algorithms, Version 1.6, http://sourceforge.net/projects/wekaclassalgos.

Cardoso,J.C. et al. (2006) Evolution of secretin family GPCR members in the metazoa.*BMC Evol Biol.*, 6, 108.

Christopoulos,A. and Kenakin,T. (2002) G protein-coupled receptor allosterism and complexing. *Pharmacol Rev.*, 54, 323-374.

Claire,A. and King,R.D. (2003) Predicting Gene Function in *Saccharomyces Cerevisiae*. *Bioinformatics*, 19, 42-49

Costa,E.P. *et al.* (2007) Comparing several approaches for hierarchical classification of proteins with decision trees. *Proc. of the 2007 Brazilian Symposium on Bioinformatics (BSB-2007).*

Das,S.S. and Banker,G.A. (2006) The role of protein interaction motifs in regulating the polarity and clustering of the metabotropic glutamate receptor mGluR1a. *J Neurosci.*, 26, 8115-25.

Davies,M.N. *et al.* (2007). Proteomic applications of automated GPCR classification. *Proteomics,* 7, 2800-14.

Flower,D.R. and Attwood,T.K. (2004) Integrative bioinformatics for functional genome annotation: trawling for G protein-coupled receptors. *Semin Cell Dev Biol.* 15, 693-701.

Flower DR. (1999) Modelling G-protein-coupled receptors for drug design. *Biochim Biophys Acta*. 1422, 207-234.

Foord,S.M. *et al.* (2002) Bioinformatics and type II G-protein-coupled receptors. *Biochem. Soc. Trans.*, 30, 473-9.

Frank,E. and Witten,I.H. (1998) Generating Accurate Rule Sets without Global Optimization. *Fifteenth International Conference on Machine Learning*.

Freitas,A.A. and de Carvalho,A.C.P.L.F. (2007) A Tutorial on Hierarchical Classification with Applications in Bioinformatics. *Research and Trends in Data Mining Technologies and Applications*, D. Taniar Ed. Idea Group, 175-208.

Fridmanis,D. *et al.* (2006) Formation of new genes explains lower intron density in mammalian Rhodopsin G protein-coupled receptors. *Mol Phylogenet Evol.*, 43, 864-80.

Gether,U. *et al.* (2002) Structural basis for activation of G-protein-coupled receptors. *Pharmacol Toxicol.*, 91, 304-312.

Gloriam,D.E. *et al.* (2005) Nine new human Rhodopsin family G-protein coupled receptors: identification, sequence characterisation and evolutionary relationship. *Biochim Biophys Acta*, 1722, 235-46

Guo,Y.Z. *et al.* (2006) Classifying G protein-coupled receptors and nuclear receptors on the basis of protein power spectrum from fast Fourier transform. *Amino Acids*, 30, 397-402.

Guo,Y.Z. *et al.* (2005) Fast fourier transform-based support vector machine for prediction of G-protein coupled receptor subfamilies *Acta Biochim Biophys Sin (Shanghai),* 37, 759-66.

Hebert,T.E. and Bouvier,M. (1998) Structural and functional aspects of G protein-coupled receptor oligomerization. *Biochem Cell Biol.* 76, 1-11.

Horn,F. *et al.* (2003) GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res*., 31, 294-7.

Karchin,R. *et al*. (2002) Classifying G-protein coupled receptors with support vector machines. *Bioinformatics*, 18, 147-159.

Keerthi,S.S. et al. (2001) Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Computation*, 13, 637-649.

Klabunde,T. and Hessler,G. (2002) Drug design strategies for targeting G-protein-coupled receptors. *ChemBioChem* 2002, *3*, 928–944.

Kolakowski,L.F. Jr (1994) GCRDb: a G-protein-coupled receptor database. *Receptors Channels*, 2, 1-7.

Lapinsh,M. *et al. (*2002) Proteochemometrics Modelling of the Interaction of Amine G-Protein Coupled Receptors with a Diverse Set of Ligands. *Molecular Pharmacology*, 61, 1465-1475.

Milligan,G. (2006) G-protein-coupled receptor heterodimers: pharmacology, function and relevance to drug discovery. *Drug Discov Today*, 11, 541-9.

Nakagawa,T. *et al. (*2005) Insect Sex-Pheromone Signals Mediated by Specific Combinations of Olfactory Receptors. Science**,** 307**,** 1638–1642.

Papasaikas,P.K. *et al. (*2004) PRED-GPCR: GPCR recognition and family classification server. *Nucleic Acids Res.,* 32, W380-2.

Prabhu,Y. and Eichinger,L. (2006) The Dictyostelium repertoire of seven transmembrane domain receptors *Eur J Cell Biol.*, 85, 937-46.

Sandberg,M. *et al.* (1998) New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *Journal of Medicinal Chemistry*, 41, 2481-2491.

Schiöth,H.B. and Fredriksson,R. (2005) The GRAFS classification system of G-protein coupled receptors in comparative perspective. *Gen. Comp. Endocrinol*., 142, 94-101.

Secker,A. *et al.* (2007) An Experimental Comparison of Classification Algorithms for the Hierarchical Prediction of Protein Function. *3rd UK Knowledge Discovery and Data mining Symposium (UKKDD 2007)*,13-18.

Watkins,A. *et al.* (2004) Artificial Immune Recognition System (AIRS): An Immune-Inspired Supervised Learning Algorithm. *Genetic Programming and Evolvable Machines*, 5, 291-317.

Wheeler,D.L. *et al.* (2007) Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 35, D5-12.

Wistrand,M. *et al.* (2006) A general model of G protein-coupled receptor sequences and its application to detect remote homologs. *Protein Sci.*, 15, 509-21.

Witten,I.H. and Frank,E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann,.San Francisco.

Wold,S. *et al.* (1993) DNA and peptide sequences and chemical processes mutlivariately modelled by principal component analysis and partial least-squares projections to latent structures. *Anal. Chim. Acta*, 277, 239–25.