DOI 10.1002/pmic.200700093 *Proteomics* 2007, *7*, 2800–2814

REVIEW

Proteomic applications of automated GPCR classification

Matthew N. Davies[1], David E. Gloriam[2], Andrew Secker[3], Alex A. Freitas[3], Miguel Mendao[4], Jon Timmis[4] and Darren R. Flower[1]

[1] Edward Jenner Institute, Compton, Newbury, Berkshire, UK
[2] European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
[3] Department of Computing and Centre for BioMedical Informatics, University of Kent, Canterbury, Kent, UK
[4] Departments of Computer Science and Electronics, University of York, Heslington, York, UK

The G-protein coupled receptor (GPCR) superfamily fulfils various metabolic functions and interacts with a diverse range of ligands. There is a lack of sequence similarity between the six classes that comprise the GPCR superfamily. Moreover, most novel GPCRs found have low sequence similarity to other family members which makes it difficult to infer properties from related receptors. Many different approaches have been taken towards developing efficient and accurate methods for GPCR classification, ranging from motif-based systems to machine learning as well as a variety of alignment-free techniques based on the physiochemical properties of their amino acid sequences. This review describes the inherent difficulties in developing a GPCR classification algorithm and includes techniques previously employed in this area.

Received: January 31, 2007
Revised: April 23, 2007
Accepted: April 23, 2007

Keywords:
Alignment / Bioinformatics / Classification / GPCR / Tools

1 Introduction

The G-protein coupled receptors (GPCRs) form a large and diverse multigene superfamily of integral membrane proteins that are involved in many important physiological functions [1–3]. GPCRs are responsible for the transduction of endogenous extracellular signals into an intracellular response. 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. The GPCR protein's association with the heterotrimeric G-protein complex causes the GDP bound to the Gα subunit to be exchanged for GTP. The Gα-GTP complex then dissociates from the Gβ subunit$_\gamma$, freeing the Gα subunit to couple to an effector enzyme. An extremely heterogeneous set of molecules can act as GPCR ligands including ions, hormones, neurotransmitters, peptides, and proteins. Sensory GPCRs can also be activated by stimuli such as light, taste, or odour. More than one type of GPCR can interact with more than one kind of G-protein, creating a complex system involving a variety of mechanisms. GPCRs control and/or affect physiological processes as diverse as neurotransmission, cellular metabolism, secretion, cellular differentiation, and inflammatory responses [4]. Mutations in GPCR-coding genes have been linked to over 30 human diseases including retinitis pigmentosa, hypo- and hyperthyroidism, nephrogenic diabetes insipidus, as well as several fertility disorders [5].

The GPCR superfamily is a common target for therapeutic drugs and approximately 50% of all marketed drugs are targeted towards a GPCR [6]. Roughly speaking, ligands

Correspondence: Dr. Matthew Davies, Edward Jenner Institute, Compton, Newbury, Berkshire RG20 7NN, UK
E-mail: m.davies@mail.cryst.bbk.ac.uk
Fax: +44-(0)207-631-6803

Abbreviations: ACC, AutoCross covariance; **BLAST,** basic local alignment search tool; **cAMP,** cyclic adenosine monophosphate; **EIIP,** electron–ion interaction potential; **GABA,** gamma-amino butyric acid; **GPCR,** G protein coupled receptor; **GPCRDB,** GPCR database; **GPCRHMM,** GPCR Hidden Markov model; **GRAFS,** Glutamate Rhodopsin Adhesion Frizzled/Taste2 Secretin; **HMM,** Hidden Markov model; **HMMTOP,** HMM for topology prediction; **QFC,** quasi-predictor feature classifier; **SOM,** self-organising map; **SVM,** support vector machine; **TM,** transmembrane; **TMHMM,** transmembrane HMM; **7TMHMM,** seven transmembrane HMM

may be divided into agonists, which directly activate the receptor, and antagonists, which interfere with the process of activation, usually by inhibiting the action of agonists. Agonists and antagonists may be further divided into full and partial, the latter only producing a partial physiological response in comparison to the former. A conventional view of the interaction is that antagonists function by blocking the endogenous ligand's binding site, although recent data suggest that some forms of antagonism can be "permissive", allowing some but not all receptor-mediated signals to be blocked [7]. This would suggest that there are different forms of agonism and antagonism whereby some GPCR-mediated signals are enhanced whilst others are suppressed.

There are inherent difficulties in providing a comprehensive classification system for the GPCR superfamily [8]. Even the choice of appropriate nomenclature has proved contentious. The term "family" has long been used to describe groupings with the GPCRs. The definition of family relies not just upon the possession of sequence similarity, but also embraces a corresponding set of structural, functional, and evolutionary features. However, there is no overarching term which includes the whole set of GPCR sequences. The real issue here is a more subtle semantic one. Evolutionary relationships between different GPCR groups are not certain; some receptors may have arisen through convergent evolution to adopt a particular structural scaffold, and may not be homologous. Given such uncertainty, the term superfamily may be ambiguous and equivocal in this context. Another, less well-used term is "clan" [9]. This term makes use of a looser, more inclusive definition of "kinship" that recognizes convergent as well as divergent evolutionary processes. However, we retain use of the term superfamily to encompass both homologous and likely nonhomologous protein families and superfamilies; use of "clan" may have been appropriate, but because this term has not been widely adopted we deprecate its use to avoid further confusion over nomenclature.

One of the first GPCR superfamily classification systems was introduced by Kolakowski [10] for the now defunct GCRDb database. GPCRs were divided into seven families, designated A–F and O, derived from original standard similarity searches. This system was further developed by Vriend *et al*. [11] for the GPCR database (GPCRDB) database. The GPCRDB database divides the superfamily into six classes. These are the Class A rhodopsin-like, which account for over 80% of all GPCRs, Class B secretin-like Class C metabotropic glutamates, Class D pheromones, Class E cAMP receptors and the Class F frizzled/smoothened family (see Table 1). Class A is the largest of the human GPCR subtypes. There are at least 286 human nonolfactory Class A receptors, the majority of which bind peptides, biogenic amines, or lipid-like substances [12]. The receptors binding endogenous peptides have an import role in mediating the effects of a wide variety of neurotransmitters, hormones, and paracrine signals. The receptors that bind biogenic amines, *e.g.*, norepinephrine, dopamine, and serotonin, are very commonly

Table 1. Class A, rhodopsin-like, which account for over 80% of all GPCR; Class B, secretin-like; Class C, metabotropic glutamates; Class D, pheromones; Class E, cAMP receptors; and the Class F frizzled/smoothened family

GPCRDB family	Protein family description
Class A	Rhodopsin-like
Class B	Secretin-like
Class C	Metabotropic glutamates/pheromone
Class D	Fungal Pheromone
Class E	cAMP receptors
Class F	Frizzled/Smoothened

modulated by drugs. Pathological conditions, including Parkinson's disease, schizophrenia, drug addiction, and mood disorders are examples of where imbalances in the levels of biogenic amines cause altered brain functions. Class B receptors bind the large peptides such as secretin, parathyroid hormone, glucagon, glucagon-like peptide, calcitonin, vasoactive intestinal peptide, growth hormone releasing hormone, and pituitary adenylyl cyclase activating protein [13]. Metabotropic glutamate receptors (mGluRs), a type of glutamate receptor, are activated through an indirect metabotropic process. Like all glutamate receptors, mGluRs bind to glutamate, an amino acid that functions as an excitatory neurotransmitter. In humans, mGluRs are found in pre- and postsynaptic neurons in synapses of the hippocampus, cerebellum, and the cerebral cortex, as well as other parts of the brain and in peripheral tissues. Pheromones are used by organisms for chemical communication [14] and cAMP receptors are part of chemotactic signalling systems [15]. Frizzled receptors are necessary for Wnt binding while the smoothened receptor mediates hedgehog signalling [16, 17]. The six different classes can further be divided into subfamilies and sub–subfamilies based upon the function of the GPCR protein and the specific ligand that it binds.

There are approximately 60 "orphan" GPCR proteins that show the sequence properties of Class A rhodopsin-like receptors but for which there are no defined ligands or functions (Gloriam *et al.*, unpublished). There are also many orphan receptors within the Class B family. Most orphan GPCRs have relatively low sequence similarity to well characterised GPCRs with known functions and/or known ligands; it is therefore often difficult to infer information about their function. It is possible that many of these orphan receptors have ligand-independent properties, specifically the regulation of ligand-binding GPCRs on the cell surface [18]. This was first suggested when a study of the Class C metabotropic γ-aminobutyric acid B ($GABA_B$) receptor showed that it was a heterodimer composed of two subunits, B1 and B2 [19]. $GABA_{B1}$ was responsible for the binding of the ligand while the $GABA_{B2}$ subunit promotes the efficient transport of $GABA_{B1}$. It is also possible that many of the orphan receptors are also responsible for the regulation of nonorphan GPCR cell surface expression, in either a positive

[20] or negative way [21]. If this is true then the relative expression of orphan and nonorphan GPCR proteins could be an important factor for the regulation of cell signaling. There has also been considerable interest in the tendency of GPCRs to form higher order oligomers in living cells [22]. Dimeric ligands linked by spacer arms have been used to identify the importance of coexpression of certain GPCR subtypes, indicating that the formation of these oligomers is a crucial part of GPCR signaling, although the extent to which oligomerisation occurs across the whole GPCR superfamily remains uncertain.

It is possible to identify a receptor's natural ligand by various experimental techniques such as the use of antagonistic antibodies, application of antisense DNA technologies, and transgenic animal studies. However, in order to focus such work, there are various initial *in silico* approaches able to characterize a GPCR sequence and forecast its potential function. Despite the diversity of the superfamily, certain commonalties remain within all GPCRs. All proteins within the GPCR superfamily contain seven highly conserved transmembrane segments (of 25–35 consecutive residues) which display a high degree of hydrophobicity. Segments can be located by a method of analysis similar to the sequence similarity search. Identifying the transmembrane regions [TM1–7] also identifies the remaining structure of the GPCR. This sequence will contain the three extracellular loops [EL1-3], three intracellular loops [IL1-3] as well as the protein termini. It can therefore be divided into the following regions:

N terminus-TM1-IL1-TM2-EL1-TM3-IL2-TM4-EL2-TM5-IL3-TM6-EL3-TM7-C terminus.

The transmembrane segments form seven α-helices in a flattened two-layer structure known as the transmembrane bundle, a structure common to all GPCRs [23]. In line with the characteristics of most related groups of proteins, the GPCR superfamily shows a far greater degree of structural conservation than it does conservation of sequence. We review here the various computational techniques used to classify GPCRs. These approaches have application not only in discovering and characterizing novel protein sequences but also in better understanding the interrelatedness apparent between known members of the GPCR superfamily.

2 Sequence based approach

One of the central dogmas of bioinformatics is that there is always a strong relation between a protein's sequence and its structure and function. This would imply a novel protein's function could be best determined by its sequence similarity to a known protein. It is remarkable that in spite of the high degree of structural similarity within the GPCR superfamily (the seven conserved transmembrane helices), there is a low degree of sequence identity, suggesting that the various classes may have originated independently during evolution. What is known is that GPCRs with the same ligands can bind to different G proteins and equally GPCRs that bind the same G protein can bind completely different ligands [24]. It is also established that, for some GPCRs, two receptors may bind the same ligand and the same G protein while having less than 25% sequence similarity. Conversely, proteins such as melanocortin, lysophosphatidic acid, and sphingosine 1-phosphate receptors all have a degree of sequence similarity but unrelated functions. Clearly, this is a challenging – that is to say difficult and complex – area in which to work. Fortunately, the importance of the GPCR as physiological agents and drug targets more than justifies our efforts in addressing this challenge. Here we outline various approaches that have been used to develop GPCR classification algorithms and attempt to highlight the strengths and weaknesses of the various approaches.

2.1 BLAST

The most obvious and straightforward approach to characterizing a protein is to run a standard basic local alignment search tool (BLAST) search [25]. The degree of relatedness is calculated by generating gapped alignments of the protein sequences with estimates of concomitant statistical significance. The program assigns a probability score for each position in an alignment that is based on the frequency with which that substitution is known to occur among consensus blocks within related proteins. The output will list proteins closely resembling the submitted sequence in descending order of expectation or "*E*" values (the *E*-value is a measure of the reliability of the *S* score, a calculation of the similarity of the query to the sequence shown). The lower the *E* value, the more significant the score and thus the higher the predicted relatedness is between the submitted sequence and the database entry. For a GPCR query sequence, it is likely that many of the proteins that show a high degree of sequence identity will also be GPCR gene/protein sequences. BLAST searches have been used to identify novel GPCR proteins in cases where there has been moderate yet detectable sequence similarity to known GPCR sequences. This, however, makes the technique of limited use for the GPCR superfamily where there is a low degree of sequence similarity between the six families.

2.2 Motif approach

While BLAST searches tend to identify generic, global similarities between protein sequences, a motif-based approach focuses on specific, length-restricted traits unique to families or subfamilies. Joost and Methner were able to suggest potential functions for a number of orphan receptors by producing multiple alignments of Class A GPCRs [26]. Chou and coworkers have shown that within the amine receptor subfamily, there is a strong correlation between the different subgroups and their amino acid composition [27–29]. Protein family databases are developed using multiple sequence alignment of the family of interest and identifying the most

conserved regions to be the basis of family motifs. PROSITE [30] characterises families with a single conserved region. More accurate are the diagnostic GPCR "fingerprints", which have been developed based on common patterns of conservation within the seven transmembrane regions [31, 32]. Rather than a single motif, the method identifies several short conserved regions within the sequence group analysed that then comprise the fingerprint. Subfamily and sub–subfamily level fingerprints are derived from segments within the TM regions, parts of the loops and parts of the N- and C-termini. This allows false positives to be more readily determined as the sequence will tend to lack several of the motifs. The fingerprint approach can also be used to design protein "signatures" at different levels of the GPCR superfamily. The database PRINTS [33] contains over 270 GPCR fingerprints and has been demonstrated to identify similarities between receptors with low sequence similarity. Interestingly, at the class level, the majority of the motifs are found in the extracellular loops while at the subfamily level, the majority are located within the intracellular loops. Huang *et al.* [34] also found the highest degree of conservation within the GPCRs occurs within the TM3 region, in particular the "DRY" motif that is common amongst amine receptors (although there are over 100 Class A proteins in which the motif is absent). However, as the number of known members for a family expands, it has become harder to precisely define the fingerprints. Understandably, it is also often the case that very atypical GPCR sequences cannot easily be identified using the fingerprint method.

Holden and Freitas 2006 [35] classified GPCRs using three different kinds of motifs: PROSITE patterns, PRINTS fingerprints and InterPro [36] entries. InterPro is a motif-based approach that combines the sequence profiles of the PRINTS, PROSITE and Pfam databases. Three different GPCR datasets were created; each dataset used one of the three types of motifs as attributes. The number of GPCR proteins and the number of attributes in each dataset was as follows: 338 proteins and 281 attributes when using Fingerprints; 194 proteins and 127 attributes when using PROSITE patterns; and 584 proteins and 448 attributes when using entries from the InterPro database. In order to classify GPCRs the authors used a swarm intelligence algorithm - a relatively new type of adaptive learning algorithm. The algorithm discovers IF-THEN classification rules of the form: *IF <a certain set of motifs is present in the protein> THEN <predict a certain class>*, where the type of motifs used in the rule antecedent (IF part) can be either PROSITE patterns, or Fingerprint signatures or InterPro entries, depending on the dataset being analysed. Hence, the goal of the swarm intelligence algorithm is to find the best possible combination of motifs to put in the antecedent of a rule, in order to create rules with the highest possible predictive accuracy. The predictive accuracy of a rule is the percentage of proteins that have the class predicted by the rule among all the proteins that have the set of motifs specified in the rule antecedent. The best results were obtained by using Fingerprint and InterPro attributes while the predictive accuracy of PROSITE patterns was relatively poor. The best classification accuracy at the level of GPCR classes (families) was 89.6% and 86.3% using Fingerprints and InterPro attributes respectively. Substantially lower accuracy rates were obtained for deeper levels of the class hierarchy.

2.3 GPCR repertoire

Various methods have been used to identify the total number of GPCRs (the so-called "repertoire") in the human genome. The increasing refinement of the human genome assembly has allowed for more sophisticated *in silico* analysis to be undertaken. The Human Genome Project used a combination of protein families and protein domains to estimate that there are 616 GPCR sequences belonging to Classes A, B and C. A motif-based approach was used whereby InterPro estimated the total number of rhodopsin-like GPCRs to be 569 [37]. Takeda and colleagues extracted approximately 950 ORFs from the human genome that had 200–1500 amino acid residues similar to those of GPCRs [38].

Another sequence-based approach has been used to develop an alternative classification system to the six GPCR families [8]. The GRAFS classification system was developed using phylogenetic analysis [39]. GRAFS divides the GPCR superfamily into the *Glutamates, Rhodopsins, Adhesions, Frizzled/Taste 2* and *Secretin* families (GRAFS) [40] (see Table 2). The authors of the GRAFS classification system constructed a roadmap of all known human GPCR and separated functional genes from pseudogenes. The process, which has been iterated and improved several times, has also led to the discovery of several new GPCRs [26, 41–43]. The GPCR repertoires of several other species have also been published, including mouse [44], rat [45], chicken [46], pufferfish [47] and mosquito [48]. The GRAFS GPCR families arose before the chordate lineage diverged from the lineage leading to nematodes as the nematode *Caenorhabditis elegans* has more than 100 receptors belonging to the GRAFS GPCR families

Table 2. The human GPCR families according to the GRAFS nomenclature, designations in other classification systems and the number of functional members in human

GRAFS family	Designation in GPCRDB	Number human members
Glutamate	C	22 (including taste type 1 receptors)
Rhodopsin	A	284 (+388 olfactory receptors)
Adhesion	B	33
Frizzled	F	11
Secretin	B	15
Taste 2	–	25

The GRAFS classification system is the only system in which the *Adhesion* and *Secretin* families are recognized as separate families.

[49]. In parallel to the GRAFS families, other GPCR families have arisen and/or evolved in specific lineages/species, such as the plant MLO receptors [50], the yeast STE2 and STE3 receptors which recognize mating factors [51], the nematode chemosensory receptors [52], the insect gustatory receptors [53] and mammalian vomeronasal receptors involved in pheromone recognition [54]. GPCRs in fungi, plants and animals have no sequence similarity, except for one *Adhesion*-like GPCR found in thale cress (*Arabidopsis thaliana*) [55]. The lack of sequence similarity between GPCR families makes the question of a common origin speculative.

3 Hidden Markov models

Compared to motif-based prediction, a more computationally sophisticated, if not necessarily a more effective, approach to the GPCR classification problem is the use of Hidden Markov models (HMMs), which derive scores based on statistical analysis. An HMM is a statistical model where the system being modeled is assumed to be a Markov process with unknown parameters. In a regular Markov model, the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a Hidden Markov model, the state is not directly visible, but variables influenced by the state are visible. The aim of the HMM is to determine the hidden parameters from the parameters that are observable. The extracted model parameters can then be used to perform further analysis of data that were not part of the training process. HMMs are widely used in bioinformatics, particularly in sequence alignment and in generating profiles for protein families. An HMM profile can also be used for searching sequence databases for new members of a given protein family. The PROSITE database contains profiles, position-specific scoring matrices, for several GPCR families while the Pfam database [56] contains an extensive collection of HMM profiles for both GPCR families and domains. Another application of HMMs is protein topology prediction using secondary structure sequences. HMMs have been used to identify the transmembrane regions of the GPCR in order to identify and characterize a putative GPCR sequence. Apart from discriminating GPCRs from nonGPCRs, TM prediction is often used as a precursor to classification. Some of the more commonly used and better known programs are adumbrated in Sections 3.1–3.3.

3.1 Hidden Markov model for topology prediction (HMMTOP)

HMMTOP method (http://www.enzim.hu/hmmtop/) is based on the principle that the topology of the TM regions is determined by the maximum divergence of amino acid composition [57]. The method is based on the hypothesis that the differences between the amino acid distributions in the various structural parts are the main driving force in the folding of the membrane proteins. This means the topology of TM proteins may be determined by the amino acid compositions of the various structural parts by showing maximum differences, rather than by enforcing specific compositions in these parts. The difference between two distributions can be characterized by a divergence function. The sum of the divergence values between the distribution of amino acids in the different structural parts and the distribution of residues in the whole protein is used to measure differences in the amino acid distributions of the structural parts. This sum differs only in one constant from the log-likelihood. Thus, the topology of membrane proteins can be determined if their amino acid sequences can be segmented into specific regions in such a way that the product of the relative frequencies of the amino acids of these segments along the amino acid sequence should be maximal.

3.2 Transmembrane Hidden Markov model (TMHMM)

The TMHMM Server (version 2.0; http://www.cbs.dtu.dk/services/TMHMM/) predicts the location of transmembrane helices by dividing a protein sequence into the most probable distribution compared to known GPCRs [58]. The program uses a novel method to model and predict the location and orientation of α helices in membrane-spanning proteins. It is based on a HMM with an architecture that corresponds to a biological system. The close mapping between the biological and computational states allows the program to infer which parts of the model architecture are important to capture the information that encodes the membrane topology, and also to obtain a better understanding of the mechanisms and constraints involved. Models are estimated both by maximum likelihood and a discriminative method. Any evidence of there being seven clearly defined TM regions within the sequence might be a good indication of it being a GPCR protein. The output details each of the TM regions within the sequence identified by the HMM. This method consistently displays a high false positive rate. However, it has been observed that when the HMMTOP and TMHMM programs are used in combination they have a higher overall success rate (0.819) than when they are used separately (0.808 and 0.762, respectively) (see Table 3). Many proteins with seven transmembrane regions are incorrectly predicted as having six or eight TM regions. This is a potential weakness of the program because the alignment of the transmembrane regions must be correct for there to be an accurate comparison made between sequences. There is a variant on the TMHMM program called 7TMHMM, which can applied to GPCR prediction and always requires the identification of seven transmembrane regions within the sequence [59].

3.3 GPCRHMM

The GPCRHMM program implements an HMM that specifically recognises GPCRs based on TM topology-related features. Wistrand *et al.* [60] found distinct loop length patterns

Table 3. Topology prediction using various membrane prediction programs

TM topology prediction method	Predicted as 7tm 7-tms	Other-tms	7-tms predicted as 6 or 8 tms	Other-tms	Sensitivity	Specificity	Success Rate
HMMTOP	1941	368	455	106	0.776	0.841	0.808
TMHMM	1725	323	547	230	0.689	0.842	0.762
MEMSAT	1453	425	621	428	0.581	0.774	0.760
SOSUI	1217	309	941	344	0.486	0.798	0.623
TMAP	944	389	1074	484	0.377	0.708	0.517
Combinatorial HMMTOP + TMHMM	2114	546	245	143	0.845	0.795	0.819

HMMTOP and TMHMM are observed to be more accurate when combined than when used separately (adapted from Inoue *et al.* 2005 [58]

and differences in amino acid composition between cytosolic loops, extracellular loops, and membrane regions. In their analysis, 13 Pfam families from GPCRDB were selected and analysed alongside the nonGPCR families of bacteriorhodospin and protein kinase. A combination of UniProt annotation, TM prediction tools and profile HMMs were used to identify the membrane and loop regions. The helices were calculated as having a median length between 22–24 amino acids while the loops were assigned much more variable length. However, the length of the first intracellular loop appears to be much more conserved than the length of the second and third. HMMs were generated for each defined region of the structure. Similarities were observed within the amino acid compositions of the three extracellular loops and also within the compositions of the three intracellular loops. The *N*- and *C*-termini were judged to be too variable to be of use to the analysis but it was observed that the profile of the *C*-terminal regions adjacent to the intracellular loops was quite different from the global *C*-terminal profile. The Sensitivity (True Positive Proportion) and Specificity (True Negative Proportion) values are shown in Table 4. Other transmembrane prediction programs include TMpred, which uses a combination of several statistical preference matrices, derived from an expert-compiled dataset of membrane proteins [61], TopPred II [62], PRED-TMR2 [63], TMHMM 2.0 [64], and TM Finder [65].

Table 4. The sensitivity and specificity for various transmembrane prediction programs (adapted by Wistrand *et al.* [60])

	Sensitivity (%)	Specificity (%)
GPCRHMM		
Global Score > 15	94.4	99.07
Global Score > 5	93.7	99.72
Global Score > 0	92.8	100
HMMTOP		
7TM	79.3	98.89
6–8 TM	95.4	91.13
Phobius		
7TM	79.6	98.79
6–8 TM	94.8	90.2
QFC	95.5	88.6
7TMHMM	93.5	90

4 Support vector machines (SVMs)

SVMs are machine-learning algorithms based on statistical learning theory. In two-class problems, an 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 form of classification is known as linear classification. However, it is a powerful form of classification because, although it is linear in the higher dimensional feature, it is nonlinear 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. Finding the maximal margin is important because, if another data point is added to the data (corresponding to a data point in the test set), it is easier to classify it correctly when there is a greater separation between the two classes. The input vectors closest to the optimal hyperplane are called the support vectors. Although SVMs are more commonly used to solve two-class problems, this technique has been applied to the classification of GPCR data with more than two classes by running the algorithm multiple times (once for each class).

4.1 GPCR subfamily classifier

The GPCR Subfamily Classifier algorithm (http://www.soe.ucsc.edu/research/compbio/gpcr-subclass/) transforms protein sequences into fixed-length feature vectors in order to apply SVM to the data [66]. The algorithm learns to

distinguish between subfamily members and nonmembers by making several passes through a training set. The trained SVM can then be used to make a prediction for a novel protein sequence. Unfortunately, the program has not been updated since March 2002 and does not incorporate sequences or classes discovered since that time into the algorithm.

4.2 Pred-GPCR

Pred-GPCR (http://athina.biol.uoa.gr/bioinformatics/PRED-GPCR/) [67, 68] was developed as a fast Fourier transform with SVMs on the basis of the hydrophobicity of the amino acid sequence. The program was trained using 403 sequences from 17 subfamilies from GPCR Classes B, C, D, and F. Quantitative descriptions of the proteins relating to hydrophobicity, bulk and electronic properties were used based upon 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. When using the FHΦ hydrophobicity scale, the technique achieved an overall 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%. Also, the relative composition of the dataset is unusual with 105 of the 403 sequences coming from the frizzled/smoothened family and the majority of the subfamilies containing 10–20 sequences. The sequences were transformed firstly into numerical representations of the sequence based upon the EIIP values and then secondly, into the frequency domain using the discrete Fourier transform, a method by which sequences of different length can be normalized. The output of these transforms is used as the input for the SVM. In the case of a n-class classification problem where $n > 2$, as is the case for the GPCR families, each ith SVM, $i = 1,\ldots,n$, was trained. All samples in the ith subfamily were given the label "1" and all other samples were given the label "−1" (these are referred to as one-*vs.*-rest SVMs). The results in Table 5 indicate that the technique is extremely effective. However, as mentioned earlier, the dataset had an unusual class distribution. One hundred and five of the 403 sequences came from the frizzled/smoothened family and in general there was a small number (less than 20) of sequences per subgroup. An intrinsic limitation of any supervised learning (classification) algorithm is that a classification model constructed from a training set can only have a chance of good predictive accuracy on a test set that is derived from the same (or at least similar) probability distribution as the training set. Given the unusual class distribution in the training set used in this work, it seems unlikely that the classification model would have a very high predictive accuracy if applied to a large set of GPCR sequences with a more usual class distribution.

Table 5. Pred-GPCR physicochemical properties for the GPCR B, C, D, and F Classes with accuracy (ACC) and Matthew's correlation coefficient (MCC) values (adapted from Guo *et al.* [70])

Class	GPCR subfamily	Hydrophobicity model		c–p–v model		EIIP model	
		ACC (%)	MCC	ACC	MCC	ACC	MCC
Class B	Calcintonin	95	0.97	85	0.91	95	0.97
	Corticotrophin releasing factor	100	1	95.7	0.97	95.7	0.97
	Glucagon	91.7	0.95	91.7	0.95	58.3	0.75
	GHRH	84.6	0.91	76.9	0.87	69.2	0.82
	Parathyroid hormone	76.5	0.86	58.8	0.75	52.9	0.71
	PACAP	90.9	0.95	81.8	0.9	90.9	0.95
	Vasocactive intestinal	85.7	0.92	71.4	0.83	57.1	0.74
	Latrophilin	100	1	95	0.97	95	0.97
	Methuselah-like protein	61.9	0.76	57.1	0.73	47.6	0.66
	Total	87.4	0.92	80.1	0.88	75.5	0.84
Class C	Metabotropic glutamate	91.3	0.92	82.6	0.85	91.3	0.92
	Calcium-sensing like	66.7	0.79	61.1	0.75	61.1	0.75
	GABA-A	95.7	0.97	65.2	0.77	65.2	0.77
	Taste receptors	91.7	0.95	91.7	0.95	66.7	0.8
	Total	87.8	0.91	75.8	0.83	76.8	0.84
Class D	Fungal pheromone A-factor	87.5	0.91	93.8	0.95	50	0.63
	Fungal pheromone B-factor	100	1	100	1	100	1
	Total	95.8	0.97	97.9	0.98	83.3	0.88
Class F	Frizzled	100	1	100	1	100	1
	Smoothened	90.9	0.95	90.9	0.95	90.9	0.95
	Total	99	0.99	99	0.99	99	0.99
Total		91.6	0.94	86	0.91	82.7	0.88

4.3 GPCRsClass

GPCRsClass [69] is another SVM-based program that concentrates on the Class A aminergic receptor subfamily. In the first round of analysis, an SVM was generated to distinguish amines from all other GPCRs. Then multiclass SVMs were set up to classify amines into the acetylcholine, adrenoreceptor, dopamine, and serotonin sub–subfamilies. Again, the one-*vs*.-the-rest SVM was used to predict the *i*th class. The SVM requires patterns of fixed length for training and testing. The sequences were transformed to fixed length format by measuring the amino acid and dipeptide compositions, giving vectors of 20 and 400 dimensions, respectively. The dipeptide composition proved to be far more reliable than the amino acid, scoring 99.7% accuracy at discriminating amine from nonGPCRs and 92% are discriminating between the four sub–subfamilies. A similar method involving amino acid, dipeptide, and tripeptide compositions [70] claimed to get 98% accuracy at the Class level. GPCRClass gave 94% accuracy at the class level when tested with the same dataset.

4.4 GPCRPred

Similar to GPCRClass, the program determines firstly whether a sequence is or is not a GPCR, to which classes it belong and then, if it is a Class A, to which subfamily it belongs [71]. The dataset used contained 692 Class A sequences, 56 Class B, 16 Class C, 11 Class D, and 3 Class E. The vectors were based upon the dipeptide composition. Again, the one-*vs*.-rest SVM was used to characterize each Class and subfamily. GPCR *vs*. nonGPCR sequences showed 99.5% accuracy, the Class prediction showed 97.3% accuracy and the subfamily step showed on average 85% accuracy.

5 GPCR to G protein coupling specificity

It is known that each GPCR subtype couples to a subset of G proteins within a given cell. Much research has been dedicated to understanding the GPCR-G protein coupling specificity. This is important in understanding the physiological mechanisms underlying the response mediated through the activation of a specific GPCR. The G_s and $G_{i/o}$ classes are responsible for the stimulation and inhibition of adenylate cyclase, respectively, while the $G_{q/11}$ family activates phospholipase C enzymes. These three families constitute the major functional class of G proteins. It is known that the specificity of the interaction with the GPCR is determined by the interaction with the α chain of the G protein. As is the case for the extracellular ligands, no specific sequence motifs have been discovered that determined coupling specificity. Understanding GPCR to G protein coupling is thus an alternative route to classifying GPCRs.

5.1 Predicting GPCR to G protein coupling specificity with HMM

Sgourakis *et al*. [72] developed a method to predict coupling specificity to GPCR proteins using HMMs. The program focused on the three intracellular loops and *C*-terminus of the GPCR protein in order to identify commonly occurring patterns. One hundred and three receptors for which non-promiscuous coupling had been determined were grouped into three functional classes and divided up into transmembrane and loop regions using the 7TMHMM program (http://ep.ebi.ac.uk/GPCR/) [58]. Having identified the intracellular loop and *C*-terminal regions, pattern discovery was carried out using the program Sequence Pattern Exhaustive Search (SPEXS), a sequence pattern discovery tool that looks for sequence patterns occurring most often in sequences or for patterns over- or under-represented in subdatasets as compared to other sets of sequences [72]. Over 4000 patterns were discovered in the specified regions and the probability of any given pattern appearing in another group was also calculated. The pattern score was calculated as the inverse of the probability of occurring. Combinations of patterns are used to determine specificity. For a submitted sequence the number of combinations was calculated and if more than 30% belong to a given G protein group then it was declared as a putative prediction. Unfortunately, 20% of the dataset showed a very low number of matches, suggesting that the developed patterns were not sufficient to describe the three G protein subsets. Multiple alignment of the inner domains of the training set also showed the three groups to be indistinct. All three groups were shown to have a low sensitivity (<0.40) and a high specificity (>0.9) but no specific statistical analysis was carried out.

5.2 GRIFFIN

Both an SVM and an HMM are used to predict coupling specificity in G-protein and receptor interaction feature finding instrument (GRIFFIN) [24]. The program features both ligand and G protein prediction capabilities in the same algorithm, suggesting that there is a relationship between the identity of the extracellular ligand and the type of G protein it stimulates which appears to transcend the identity of the GPCR transducing the signal. The training dataset was composed of opsins and olfactory receptors from Class A as well as receptors from GPCR Classes B, C, and F. Using an HMM, the subfamilies can be differentiated according to G protein type. All of the GPCRs used in the dataset bind to one of the three G proteins ($G_{i/o}$, $G_{q/11}$, or G_s) mentioned previously. Instead of using transmembrane prediction software, the TM regions of the GPCRs were determined through conventional multiple alignment. From this, an HMM profile of multiple GPCR binding to a given ligand was generated. The SVM classifies the vector representations of GPCRs using the maximal margin hyperplane. Each G protein is compared to the rest of the data set using the one-

vs.-the-rest procedure. At the first stage, a query sequence is searched against the HMM profiles of the defined GPCR classes. If the computed HMM profile score is larger than the threshold of a certain subfamily then the processes is stopped and the sequence is assigned to that subfamily. If not, the program compares the scores of the three trained SVMs (with the G_s, the $G_{i/o}$, and the $G_{q/11}$ profiles) and assigns the GPCR to the class of the SVM with the largest score. The average discrimination sensitivities and specificities were 87 and 88% for $G_{i/o}$, 85 and 84% for $G_{q/11}$, and 85 and 89% for G_s, respectively (see Table 6). This showed a higher overall sensitivity and specificity than Moller *et al.* [59]. However, there is no indication as to how the dataset would accommodate a poorly aligned sequence or protein that showed low sequence similarity to the GPCRs used to generate the HMM profiles.

6 Alignment free methods

In most cases, conventional bioinformatics techniques determine information about a protein sequence through alignment or by comparing the sequence to previously determined motifs. While this approach is certainly valid, it may not necessarily be the most effective form of analysis when dealing with the problem of GPCR identification. Firstly, the sequence of the GPCR superfamily varies between 290–834 amino acids in length, meaning that many of the subfamilies cannot be effectively aligned without significant manual correction. It is also important to bear in mind that the conventional GPCR classification system is based not on sequence similarity but by the ligand to which the receptor is bound. Alignment-independent classification systems use the physiochemical properties of amino acids to determine differences between protein sequences.

6.1 Proteochemometrics

Proteochemometrics is a technique whereby twenty-six separate physicochemical properties of the protein are used to calculate five empirical "z" values for all twenty amino acids [74]. The z1 value accounts for the amino acid's lipophilicity and is determined by TLC variables, log *P* values and the nonpolar surface area. A large negative value corresponds to a lipophilic amino acid, whilst a large positive value corresponds to a polar/hydrophilic amino acid. Steric properties are accounted for by the z2 values which summarizes the residue's steric bulk/polarisability. In this case, a large negative value corresponds to a lower molecular weight and small surface area while a large positive value corresponds to a higher molecular weight and large surface area. The z3 value describes the polarity of the amino, this is determined by the log *P* values and nonpolar surface area. A lipophilic amino acid corresponds to a large negative value while a polar/hydrophilic amino acid has a large positive value. The electronic effects, determined by electronegativity, heat of formation, electrophilicity, are described by the z4-5 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. AutoCross Covariance (ACC) is used to normalize the uneven size of the z matrices and then principal component analysis (PCA) and partial least squares (PLS) are carried out in order to provide

Table 6. Average discrimination sensitivities and specificities for the three G protein families (adapted from Yabuki *et al.* [24]

G-protein type	*n*	Sensitivity	Specificity	Number of crossvalidations	Best kernel function
Gi/o	61	77	78.3	4	RBF
Gq/11	47	68.1	72.7	4	RBF
Gs	24	83.3	95.2	4	RBF
G-protein type	*n*	Sensitivity	Specificity	Number of crossvalidations	Best kernel function
Gi/o	61	91.8	94.9	4	Polynomial
Gq/11	47	93.6	89.8	4	Polynomial
Family	G-protein type	Sensitivity	Specificity	Threshold of bit score	
Opsin	Gt	99.7	100	153.9	
Olfactory	Golf	100	100	151.2	
Class B	Gs	100	100	68	
Class C	Gi/o	93.5	100	1054.6	
	Gq/11	100	100	1325.3	
Frizzled	Unclear	100	100	168.7	
Smoothened	Unclear	100	100	627.6	

a classification system for the various classes of protein. Key to the ACC approach is balancing the two factors of maximum lag, *L*, and the degree of normalization, *p*. The *L* value is a summation, over the sequence, of the product of z values for all pairs of two amino acids separated by the defined lag value. The value is then normalized by the number of terms in the summation. Optimal parameter values for the dataset are determined by trying various combinations and assessing their total classification accuracy. Using the proteochemometrics method, Lapnish *et al*. developed a model with an accuracy of 0.76 for a diverse set of amine GPCRs.

6.2 Self-organising maps (SOMs)

SOMs [75] can be used for GPCR classification using z values. SOMs are artificial neural networks (ANNs) that perform unsupervised learning (in this case, clustering) to discriminate one protein family from another. Unlike PCA, which relies upon establishing linear relations, SOMs can accommodate nonlinear relations into their algorithms. The SOMs can locate samples from an input space to particular "neurons" in a 2-D lattice through an adaptive process. Each neuron is fully connected to the input space and has a synaptic weight vector for the connection. The output space is 50 by 50 neurons on a square lattice. The learning process causes closely located neurons to become responsive to similar input data. Sequences from the same family are expected to form a cluster although it cannot be assumed that the clusters will be visually recognized on the SOM output map. In order to make a family map, it is necessary to determine a family area that contains the most frequent "activator" family samples. The feature map used for GPCR classification was amended to include an objective border between clusters, developing a map to clearly distinguish each family. The overall performance of the map can be assessed using the sensitivity and specificity values as well as calculating the total accuracy of prediction. Otaki *et al*. reported a 97.4% precision at classifying 12 Class A subfamilies using SOMs.

6.3 Quasi-predictor feature classifier

Kim *et al*. also used physicochemical information as the basis of a predictive technique [76, 77]. The classification method places sequences in a "feature space" and creates discrimination functions that classify the sequences into specific categories. Here, the feature space uses statistical measures of physicochemical properties and then a linear discriminant function to extract a potential GPCR sequence from a genome. The Quasi-predictor Feature Classifier (QFC) algorithm was designed to statistically characterize the differentiating features of the physicochemical properties of protein sequences using heuristic data reduction principles. The parameters derived from the characterization were then used to screen databases for novel GPCRs. The training data set was generated from 750 GPCRs from the GPCRDB (no information was given about Class distribution but it may be assumed the majority were Class A) and 1000 randomly chosen nonGPCR proteins of 200–1000 amino acids in length (it is also not stated how many of these were transmembrane proteins). The amino acid properties examined were the Goldman Engelman Steitz (GES) hydropathy, the Kyte-Doolittle index, polarity, p*I*, molecular weight, solubility, and the alpha helix index. The values were also normalized using the Sliding Windows Recogniser, a technique similar to lags used in the AutoCross Covariance approach, except that the separated values (defined here by a "window") are summed rather than multiplied. Window lengths of 13–16 amino acids were shown to be more effective than lengths of 32 and 64 amino acids. An independent test set of 100 GPCRs and 100 nonGPCRs was classified with an accuracy of 99%. When tested against 530 ion channels proteins (transmembrane but not GPCRs), the algorithm achieved an accuracy of 96.4%. It was suggested that the lack of reliance on motifs and sequence similarity had allowed the techniques to avoid the pitfalls of biased sampling. However, the QFC algorithm was shown to have a higher false positive rate than most motif-based techniques, suggesting that the technique would benefit from another stage of filtering.

7 Miscellaneous classification

7.1 GPCR classification tree

Huang's GPCR classification tree [78] is a decision tree that divided up 4395 sequences into Classes, Subfamilies, sub–subfamilies and types. A total of 39 subfamilies and 93 sub–subfamilies were discovered in this way. Each protein is represented by a vector of 20 values, where each *i*th value, $i = 1,\ldots,20$, represents the percentage of the *i*th amino acid in the composition of the protein's sequence. The *k*th protein X_k is explicitly formulated as:

$$X_k = \begin{bmatrix} X_k, 1 \\ X_k, 2 \\ \cdot \\ \cdot \\ \cdot \\ X_k, 20 \end{bmatrix}, \quad k = 1, 2 \ldots N$$

where *N* is the total number of proteins in the dataset. The program used the C4.5 algorithm, which is a divide and conquer approach that splits the training sets into subsets based upon the amino acid compositions of the proteins. Each splitting point becomes a node in the decision tree. This choice of when to split is made by selecting the amino acid composition feature that best discriminates among the classes to be predicted, and then splitting the training data according to the values of the chosen feature. The division continues until the stopping criteria are satisfied. The default

feature-selection criterion is based on the "information gain ratio", which is a measure of how successfully a given feature separates the data. The technique recorded 86.9% accuracy for the subfamily level and 81.5% accuracy for the sub–subfamily level.

7.2 OET-KNN

Optimised evidence-theoretic K-nearest neighbour (OET-KNN) is a classification method similar to SVMs [79]. The method was used to classify proteins represented by features referring to amino acid composition and the order of amino acids in the protein sequence. This is incorporated into a 20 dimensional amino acid composition vector. The technique has not been applied specifically to the GPCR Classification problem but to various transmembrane protein families with an accuracy >90%.

8 GPCR virtual screening

In the last 15 years, drug discovery has used high-throughput screening (HTS) to search for molecules with specific biological activities. The experimental HTS is, however, both very expensive and time consuming. More recently, it has been possible to supplement experimental work with *in silico* methods that significantly reduce the time and effort required to identify lead compounds for further optimization. So-called virtual screening (VS) is a technique that can explore large compound databases for drug-like properties by computational analysis of receptor–ligand interaction. In the case of GPCRs, there is considerable interest in developing drug-like small molecules that are agonists or antagonists of a specific receptor. The synthetic agonists and antagonists can, in a loose sense, be modeled on the structure and properties of endogenous GPCR ligands, as they are likely to share similar characteristics. It is possible to identify GPCR protein subtypes that have a general affinity for the endogenous antagonists by using synthetic molecules. It is also possible to use VS to classify the various GPCR subtypes by identifying subtype specificity through receptor–ligand interactions.

Exploration of the affinity of a specific ligand for different GPCRs or of different ligands for the same GPCR can reveal a lot of information about GPCR identity and function. An example of this is the Urotension II ligand, an 11 amino acid peptide that functions as a mammalian vasoconstrictor, which has been implicated in the regulation of cardiovascular homeostasis [80]. Quantitative structure–activity relationship (QSAR) studies have suggested that the functional attribute of the ligands lies in the Trp9-Lys10-Trp11 region. In particular, it is believed that the position and distance between the positive ionisable residues relative to the adjacent aromatic residues may be key to determining the ligands affinity [81, 82]. It is desirable to be able to replicate this structure using a smaller nonconformationally dependent ligand. To this end, various hexapeptides as well as molecules showing the benzidine or izide moieties have been developed to try to replicate the activity of the endogenous ligand. The selection of a small molecule dataset to be screened against a GPCR model occurs in several stages. Firstly, filters which are not specific for the target are applied, removing structures which are not drug-like, such as reactive compounds [83]. When general properties of a structure are known (such as the fact that GPCR agonists and antagonists tend to be large heterocyclic compounds), it is possible to construct a pharmacophore to select a manageable number of structures from a larger dataset. GPCR models have been successfully used for a number of different GPCR proteins. Potential antagonists were also discovered for the Neurokinin-1 receptor [84], the alpha1a adrenergic receptors [85], and the dopamine 3 (D3) subtype receptor [86] using the VS technique.

8.1 Homology modelling

One of the most effective ways of carrying out VS, other than *via* QSAR analysis, is to generate a three dimensional (3-D) structure of the receptor and then dock a succession of ligands into its binding groove. Homology modeling is a technique that takes the amino acid sequence of an unknown structure and the solved structure of a similar protein and computationally mutates each amino acid in the solved structure into the corresponding amino acid from the unknown structure. The newly generated structure can be optimized by the use of molecular dynamics simulations. This technique cannot realistically be used to identify a GPCR sequence but if the sequence has been classified then generating a 3-D structure of the protein can be useful in trying to characterize ligands that might bind the receptor. In conventional homology modeling, one or more templates can be used. Therefore, although target and templates are likely to be correctly aligned if sharing more than 40% identity, they need to be realigned if they are in the "twilight zone" sharing less than 30% identity. Unfortunately, due to the difficulties of overexpression, purification, and concentration of membrane proteins, there is only one experimentally determined structure of a GPCR currently available. The structure is that of bovine rhodopsin which was elucidated to 3.5 Å using X-ray crystallography in 2000 [87]. Therefore, there are several techniques that try to generate 3-D models of GPCRs from the primary sequence using a combination of what is known of their structure – *e.g.*, the seven transmembrane regions – and information obtained from the bovine rhodopsin model. These 3-D structures can then be used for VS. Very few members of the GPCRs are sufficiently similar to bovine rhodopsin to have more than 30% sequence identity; the vast majority are within the "twilight" zone of less than 30%. Despite this, all GPCRs show a common pattern of hydrophobicity and similarity of structure even when there is sequence divergence (see Section 1). For this reason, the bovine rhodopsin structure has frequently been used as

the basis for GPCR homology modeling and has generated structures that have successfully been used in VS experiments.

Conventional homology modeling is of little or no use if, as is most typically the case, there is low sequence similarity between the unknown protein sequence and the sequence of the bovine rhodopsin protein. However, the transmembrane sequences can be docked together using the bovine rhodopsin structure as a scaffold so that hydrophobic faces are orientated into the membrane phase and hydrophilic faces point into the lumen of the protein. Hydrophobic profiles of multisequence alignment of GPCRs can be used to assign helical transmembrane regions. The programs WHATIF (http://swift.cmbi.kun.nl/whatif/) [88] and MembStruk [89] can both be used for this purpose (MembStruk has been validated using the bovine rhodopsin structure.). This presupposes that the location of the transmembrane regions has already been determined. The next stage is to incorporate the extracellular and intracellular loops as well as the termini of the molecule into the transmembrane scaffold. The loops are harder to model because not only will there be low sequence similarity between bovine rhodopsin and the protein sequence but the respective loops may be of a very different length. It is for this reason the termini and loops are usually added in an extended conformation. The program DRAWBRIDGE [90] builds loop regions onto protein structures using the conformational propensities of amino acids to generate novel candidates for protein loop regions. An alternative to standard homology modeling is threading assembly requirement (TASSER). The program combines threading, whereby the target sequence is threaded through the backbone structures of a collection of template proteins and a "goodness of fit" score is calculated for each sequence-structure alignment. *Ab initio* algorithms are used to span the similar and dissimilar regions [91]. A benchmark test of 2234 PDB protein structures revealed that approximately two-thirds of single domain proteins could be modeled with a Cα RMSD that is within 6.5 Å of the native structure. These improved threading templates may offer a significant advantage over traditional homology modeling techniques. Two forms of TASSER were applied, the first designed to construct a protein on the basis of it being a membrane protein, the second without any prior knowledge of the protein's identity. Out of 907 GPCR sequences tested, 819 appeared to show the correct global fold. Also, a structural consistency was observed between GPCRs binding the same or similar ligands. The collected structures therefore represent a new possibility in the field of GPCR Classification based upon predicted structural folds.

A different approach is taken by the PREDICT method, which combines the properties of protein sequences with that of their membrane environments [92]. PREDICT searches through receptor conformations for the most stable 3-D structure of the transmembrane domain. The algorithm takes into account the membrane environment, the membrane lipophilic core and the polar head groups. Once generated, the structures are ranked by the PREDICT energy score. The process begins with coarse modeling, followed by fine modeling of the most stable coarse structures generated. The structure then undergoes further refinement with molecular dynamics. The VS performed on the structures is evaluated by enrichment factors, the capacity of an *in silico* screening procedure to identify known binders from a background of random components.

8.2 Molecular docking

3-D VS requires selecting a manageable number of small molecule candidates from a database such as CHEMBRIDGE [93] and docking them using a program like GOLD [94]. This technique was successfully applied to characterize the antagonist binding of the CCR1 Chemokine receptor. A combination of MembStruk [89] and the HierDock program [95] was used to build a model of CCR1 and dock into it the BX 471 antagonist [96]. The program generates the transmembrane regions and hydrophobic maxima potential for the target structure before optimizing the translational and rotational orientation of the side chains. A void space of BX741, include 14 regions of $7 \times 7 \times 7$ Å^3 volume, were used to determine the binding site of BX471. Analysis of the protein–ligand interaction revealed a strong hydrophobic character to the groove by which the urea group of the molecule is able to form hydrogen bonds with water molecules. Validation of the technique used 35 known CCR1 antagonists added to 51 000 compounds selectively filtered from the MayBridge database (http://cds.dl.ac.uk/cds/datasets/orgchem/isis/maybridge.html). The program placed 43% of the known CCR1 antagonists in the top 2% of predicted binders and 63% within the top 5%.

9 Conclusion

Examination of the current literature shows that no real consensus exists for tackling the problem of *in silico* GPCR classification. GPCR prediction is a complicated problem that may be beyond conventional bioinformatics techniques. Classification models based upon motifs are both simple and comprehensible to the user but have been observed to have false positive and false negative prediction rates that are erratic. Models constructed by SVMs or ANNs are typically opaque to the user but are often more effective. The alignment-independent methods, while showing some of the highest overall accuracy, do not allow the user to infer any information about the protein sequence other than to which family it likely belongs. Therefore, there is arguably a trade-off between the accuracy of the predictive technique and the comprehensibility of its results. A possible way to combine both accuracy and comprehensibility would be 3-D VS, which has been shown to be an effective technique for

characterizing GPCR Classes, subfamilies and sub–subfamilies and the ligands with which they can specifically interact.

It should be noted that while many of the algorithms described show a high degree of accuracy, in most cases the technique has not been assessed independently. Further benchmarking of the techniques with several different GPCR datasets seems necessary. It may also be the case that a technique that is effective at determining GPCRs from nonGPCRs would be less effective at the class, subfamily or sub–subfamily level. Different approaches could therefore be employed at each level of the classification. Furthermore, all the predictive techniques have hitherto been assessed using the GPCRDB classification system. Future work in this field may need to be directed towards training algorithms based upon alternative classification systems, such as GRAFS, in order to determine the most comprehensive approach to classifying the GPCR superfamily.

GPCRs remain important drug targets; they still account for a significant proportion of global pharmaceutical sales [6, 97]. The pharmaceutical industry is, however, no longer the engine of unalloyed, ever-increasing profitability that it once was. During 2006, worldwide sales of prescription medicines rose by a modest 7% to around $602 billion. Established pharmaceutical companies all suffer from the inconvenient coincidence of incipient product droughts, caused, in the main, by weak or dwindling internal pipelines, coupled to severe earnings pressures resulting from the expiry of major remunerative patents on flagship products. In particular, growth in the traditional markets of Japan, North America, and Europe has been slowing for several years. In 2004, North American sales grew at a rate of 8.3% to $235.4 billion, compared with 11.5% growth from 2002 to 2003. By 2006, annual sales in North America were $252 billion, increasing by only 5.7%. Yet, over half of marketed drugs concentrate on a single class of biological targets: GPCRs. This includes 25% the top 100 drugs, many of which are so-called blockbusters, each earning over $1 billion a year. Many commentators have questioned the long-term viability of the blockbuster, suggesting that the already fragmented pharmaceutical market is moving towards an even more focused market dominated by a legion of niche products. Because of their hydrophobic binding site and their vital biological roles GPCRs remain the ultimate druggable target. While we still need drugs, we will continue to explore the unique properties of the GPCR. This review has shown how we will take the next step on that road allowing us to more fully exploit as drug targets the as yet untapped potential of the entire GPCR family.

The authors should like to gratefully acknowledge funding under the ESPRC grant EP/D501377/1. We should also like to thank the sadly anonymous referees who have made such a fundamental difference to the writing of this paper. In particular, we should like to graciously acknowledge the first referee for his or her pedantic yet fulsome and painstaking criticism of the manuscript; his or her contribution has utterly transformed the paper.

10 References

[1] Christopoulos, A., Kenakin, T., *Pharmacol. Rev.* 2002, *54*, 323–374.

[2] Gether, U., Asmar, F., Meinild, A. K., Rasmussen, S. G., *Pharmacol. Toxicol.* 2002, *91*, 304–312.

[3] Bissantz, C., *J. Recept. Signal Transduct. Res.* 2003, *23*, 123–153.

[4] Hebert, T. E., Bouvier, M., *Biochem. Cell Biol.* 1998, *76*, 1–11.

[5] Schoneberg, T., Schulz, A., Biebermann, H., Hermsdorf, T. *et al.*, *Pharmacol. Ther.* 2004, *104*, 173–206.

[6] Klabunde, T., Hessler, G., *Chem. Bio. Chem.* 2002, *3*, 928–944.

[7] Kenakin, T., *Nat. Rev. Drug Discov.* 2005, *4*, 919–927.

[8] Cheng, B. Y., Carbonell, J. G., Klein-Seetharaman, J., *Proteins* 2005, *58*, 955–970.

[9] Attwood, T. K., Blythe, M. J., Flower, D. R., Gaulton, A. *et al.*, *Nucleic Acids Res.* 2002, *30*, 239–241.

[10] Kolakowski, L. F., Jr., *Receptors Channels* 1994, *2*, 1–7.

[11] Horn, F., Bettler, E., Oliveira, L., Campagne, F. *et al.*, *Nucleic Acids Res.* 2003, *31*, 294–7.

[12] Fridmanis, D., Fredriksson, R., Kapa, I., Schioth, H. B., Klovins, J., *Mol. Phylogenet. Evol.* 2007, *43*, 864–880.

[13] Cardoso, J. C., Pinto, V. C., Vieira, F. A., Clark, M. S., Power, D. M., *BMC Evol. Biol.* 2006, *6*, 108.

[14] Das, S. S., Banker, G. A., *J. Neurosci.* 2006, *26*, 8115–8125.

[15] Nakagawa, T., Sakurai, T., Nishioka, T., Touhara, K., *Science* 2005, *307*, 1638–1642.

[16] Prabhu, Y., Eichinger, L., *Eur. J. Cell Biol.* 2006, *85*, 937–946.

[17] Gloriam, D. E., Schioth, H. B., Fredriksson, R., *Biochim. Biophys. Acta* 2005, *1722*, 235–246

[18] Levoye, A., Dam, J., Ayoub, M. A., Guillaume, J. L. *et al.*, *EMBO Rep.* 2006, *7*, 1094–1098.

[19] Pin, J. P., Kniazeff, J., Liu, J., Binet, V. *et al.*, *FEBS J.* 2005, *272*, 2947–2955.

[20] Milasta, S., Pediani, J., Appelbe, S., Trim, S. *et al.*, *Mol. Pharmacol.* 2006, *69*, 479–491.

[21] Levoye, A., Dam, J., Ayoub, M. A., Guillaume, J. L. *et al.*, *EMBO J.* 2006, *25*, 3012–3023.

[22] Milligan, G., *Drug Discov. Today* 2006, *11*, 541–549.

[23] Yeagle, P. L., Albert, A. D., *Biochim. Biophys. Acta.* 2007, *1768*, 808–824.

[24] Yabuki, Y., Muramatsu, T., Hirokawa, T., Mukai, H., Suwa, M., *Nucleic Acids Res.* 2005, *33*, W148–W153.

[25] Lopez, R., Silventoinen, V., Robinson, S., Kibria, A., Gish, W., *Nucleic Acids Res.* 2003, *31*, 3795–3798.

[26] Joost, P., Methner, A., *Genome Biol.* 2002, *3*, RESEARCH0063.

[27] Chou, K. C., Elrod, D. W., *J. Proteome Res.* 2002, *1*, 429–433.

[28] Chou, K. C., *J. Proteome Res.* 2005, *4*, 1413–1418.

[29] Gao, Q.-B., Wang, Z.-Z., *Protein Eng., Des. Select.* 2006, *19*, 511–516.

[30] Hulo, N., Sigrist, C. J., Le Saux, V., Langendijk-Genevaux, P. S. *et al.*, *Nucleic Acids Res.* 2004, *32*, 134–137.

[31] Attwood, T. K., *Pharmacol. Sci.* 2001, *22*, 162–165.

[32] Flower, D. R., Attwood, T. K., *Semin. Cell Dev. Biol.* 2004, *15*, 693–701.

[33] Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A. *et al.*, *Nucleic Acids Res.* 2007 *35*, D224–D228.

[34] Huang, E. S., *Protein Sci* 2003, *12*, 1360–1367.

[35] Holden, N., Freitas, A. A., *Proc. IEEE Swarm Intell Sympos* 2006. *SIS-06*, 77–84.

[36] Attwood, T. K., Bradley, P., Flower, D. R., Gaulton, A. *et al.*, *Nucleic Acids Res.* 2003, *31*, 400–402.

[37] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C. *et al.*, *Nature* 2001, *409*, 860–921.

[38] Takeda, S., Kadowaki, S., Haga, T., Takaesu, H., Mitaku, S., *FEBS Lett.* 2002, *520*, 97–101.

[39] Schiöth, H. B., Nordström, K. J., Fredriksson, R., *Acta Physiol. (Oxf.)*, 2007, *190*, 21–31.

[40] Fredriksson, R., Hoglund, P. J., Gloriam, D. E., Lagerstrom, M. C., Schioth, H. B., *FEBS Lett.* 2003, *554*, 381–388.

[41] Bjarnadottir, T. K., Fredriksson, R., Hoglund, P. J., Gloriam, D. E. *et al.*, *Genomics* 2004 *84*, 23–33.

[42] Nordström, K. J., Mirza, M. A., Larsson, T. P., Gloriam, D. E. *et al.*, *Biochem. Biophys. Res. Comm.* 2006, *348*, 1063–1074.

[43] Vassilatis, D. K., Hohmann, J. G., Zeng, H., Li, F. *et al.*, *Proc. Natl. Acad. Sci. USA* 2003 *100*, 4903–4908.

[44] Bjarnadottir, T. K., Gloriam, D. E., Hellstrand, S. H., Kristiansson, H. R., Schioth, H. B., *Genomics* 2006, *88*, 263–273.

[45] Gloriam, D. E., Schioth, H. B., Fredriksson, R., *Biochim. Biophys. Acta* 2005, *1722*, 235–246.

[46] Lagerstrom, M. C., Hellstrom, A. R., Gloriam, D. E., Larsson, T. P. *et al.*, *PLoS Comput. Biol.* 2006, *2*, e54.

[47] Metpally, R. P., Sowdhamini, R., *BMC Evol. Biol.* 2005, *5*, 41.

[48] Hill, C. A., Fox, A. N., Pitts, R. J., Kent, L. B. *et al.*, *Science* 2002, *298*, 176–178.

[49] Fredriksson, R., Schioth, H. B., *Mol. Pharmacol.* 2005, *67*, 1414–1425.

[50] Devoto, A., Hartmann, H. A., Piffanelli, P., Elliott, C. *et al.*, *J. Mol. Evol.* 2003, *56*, 77–88.

[51] Hagen, D. C., McCaffrey, G., Sprague, G. F., Jr., *Proc. Natl. Acad. Sci. USA* 1986, *83*, 1418–1422.

[52] Robertson, H. M., *Genome Res.* 1998, *8*, 449–463.

[53] Hill, C. A., Fox, A. N., Pitts, R. J., Kent, L. B. *et al.*, *Science* 2002, *298*, 176–178.

[54] Kouros-Mehr, H., Pintchovski, S., Melnyk, J., Chen, Y. J. *et al.*, *Chem. Senses* 2001 *26*, 1167–1174.

[55] Josefsson, L. G., Rask, L., *Eur. J. Biochem.* 1997, *249*, 415–420.

[56] Bateman, A., Birney, E., Durbin, R., Eddy, S. R. *et al.*, *Nucleic Acids Res.* 2000, *28*, 263–266.

[57] Tusnady, G. E., Simon, I., *Bioinformatics* 2001, *17*, 849–850.

[58] Inoue, Y., Yamazaki, Y., Shimizu, T., *Biochem. Biophys. Res. Commun.* 2005, *338*, 1542–1546.

[59] Möller, S., Vilo, J., Croning, M. D., *Bioinformatics* 2001, *17*, S174–S181

[60] Wistrand, M., Kall, L., Sonnhammer, E. L., *Protein Sci.* 2006, *15*, 509–521.

[61] Hofmann, K., Stoffel, W., *Biol. Chem. Hoppe-Seyler* 1993, *347*, 166.

[62] Claros, M. G., von Heijne, G., *Comput. Appl. Biosci.* 1994, *10*, 685–686.

[63] Pasquier, C., Promponas, V. J., Palaios, G. A., Hamodrakas, J. S., Hamodrakas, S. J., *Protein Eng.* 1999, *12*, 381–385.

[64] Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E. L., *J. Mol. Biol.* 2001, *305*, 567–580.

[65] Deber, C. M., Wang, C., Liu, L. P., Prior, A. S. *et al.*, *Protein Sci.* 2001, *10*, 212–219.

[66] Karchin, R., Karplus, K., Haussler, D., *Bioinformatics* 2002, *18*, 147–159.

[67] Papasaikas, P. K., Bagos, P. G., Litou, Z. I., Promponas, V. J., Hamodrakas, S. J., *Nucleic Acids Res.* 2004, *32*, W380–W382.

[68] Guo, Y. Z., Li, M. L., Wang, K. L., Wen, Z. N. *et al.*, *Acta Biochim. Biophys. Sin. (Shanghai)* 2005, *37*, 759–766.

[69] Bhasin, M., Raghava, G. P., *Nucleic Acids Res.* 2005, *33*, W143–W147.

[70] Guo, Y. Z., Li, M., Lu, M., Wen, Z. *et al.*, *Amino Acids* 2006, *30*, 397–402.

[71] Bhasin, M., Raghava, G. P., *Nucleic Acids Res.* 2004, *32*, W383–W389.

[72] Sgourakis, N. G., Bagos, P. G., Papasaikas, P. K., Hamodrakas, S. J., *BMC Bioinformatics* 2005, *6*, 104.

[73] Vilo, J., Kapushesky, M., Kemmeren, P., Sarkans, U., in: Parmigiani, G., Garett, E. S., Irizarry, R. A., Zeger, S. L. (Eds), *The Analysis of Gene Expression Data: Methods and Software*, Springer Verlag, New York, NY 2003.

[74] Lapinsh, M., Gutcaits, A., Prusis, P., Post, C. *et al.*, *Protein Sci.* 2002 *11*, 795–805.

[75] Otaki, J. M., Mori, A., Itoh, Y., Nakayama, T., Yamamoto, H., *J. Chem. Inf. Model* 2006, *46*, 1479–90.

[76] de Trad, C. H., Fang, Q., Cosic, I., *Protein Eng.* 2002, *15*, 193–203.

[77] Kim, J., Moriyama, E. N., Warr, C. G., Clyne, P. J, Carlson, J. R., *Bioinformatics* 2000, *16*, 767–775.

[78] Huang, Y., Cai, J., Ji, L., Yanda, L., *Comput Biol Chem* 2004, *28*, 275–280.

[79] Shen, H. B., Chou, K. C., *Biochem. Biophys. Res. Commun.* 2005, *337*, 752–756.

[80] Brkovic, A., Hattenberger, A., Kostenis, E., Klabunde, T., Flohr, S. *et al.*, *J. Pharmacol. Exp. Ther.* 2003, *306*, 1200–1209.

[81] Flohr, S., Kurz, M., Kostenis, E., Brkovich, A. *et al.*, *J. Med. Chem.* 2002, *45*, 1799–1805.

[82] Grieco, P., Carotenuto, A., Campiglia, P., Marinelli, L. *et al.*, *J. Med. Chem.* 2005, *48*, 7290–7297.

[83] Sadowski, J., Kubinyi, H., *J. Med. Chem.* 1998, *41*, 3325–3329.

[84] Evers, A., Klebe, G., *J. Med. Chem.* 2004, *47*, 5381–5392.

[85] Evers, A., Klabunde, T., *J. Med. Chem.* 2005, *48*, 1088–1097.

[86] Varady, J., Wu, X., Fang, X., Min, J., Hu, Z. *et al.*, *J. Med. Chem.* 2003, *46*, 4377–4392.

[87] Palczewski, K., Kumasaka, T., Hori, T., Behnke, C. A. *et al.*, *Science* 2000, *289*, 739–745.

[88] Vriend, G., *J. Mol. Graph.* 1990, *8*, 52–56.

[89] Vaidehi, N., Floriano, W. B., Trabanino, R., Hall, S. E. *et al.*, *Proc. Natl. Acad. Sci. USA* 2002, *99*, 12622–12627.

[90] Ring, C. S., Cohen, F. E., *Israel J. Chem* 1994, *34*, 245–252.

[91] Zhang, Y., Devries, M. E., Skolnick, J., *PLoS Comput. Biol.* 2006, *2*, e13.

[92] Shacham, S., Marantz, Y., Bar-Haim, S., Kalid, O. *et al.*, *Proteins* 2004, *57*, 51–86.

[93] *ChemBridge* (Express Pick, October 2001: ChemBridge Corporation, San Diego, CA).

[94] Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., Taylor, R. D., *Proteins* 2003, *52*, 609–623.

[95] Floriano, W. B., Vaidehi, N., Goddard, W. A., III, Singer, M. S., Shepherd, G. M., *Proc. Natl. Acad. Sci. USA* 2000, *97*, 10712–10716.

[96] Vaidehi, N., Schlyer, S., Trabanino, R. J., Floriano, W. B. *et al.*, *J. Biol. Chem.* 2006, *281*, 27613–27620.

[97] Flower, D. R., *Biochim. Biophys. Acta* 1999, *1422*, 207–234.

