Present Perspectives on the Automated Classification of the G-Protein Coupled Receptors (GPCRs) at the Protein Sequence Level

Matthew N. Davies [1]

David E. Gloriam [2]

Andrew Secker [3]

Alex A. Freitas [3]

Jon Timmis [4]

Darren R. Flower [5]

[1] SGDP, Institute of Psychiatry, King's College London
De Crespigny Park, London, United Kingdom, SE5 8AF

[2] Department of Medicinal Chemistry, University of Copenhagen
Universitetsparken 2, 2100 Copenhagen, Denmark

[3] School of Computing and Centre for BioMedical Informatics
University of Kent, Canterbury, Kent CT2 7NF, U.K.

[4] Departments of Computer Science and Electronics
University of York, Heslington, York YO10 5DD, U.K.

[5] School of Life and Health Sciences,
Aston University,
Aston Triangle, Birmingham, B4 888

Corresponding Author: Dr Darren R Flower

Keywords: GPCR/ Classification/ Bioinformatics/ Alignment/ Tools

Abstract

The G-protein coupled receptors – or GPCRs - comprise simultaneously one of the largest and one of the most multi-functional protein families known to modern-day molecular bioscience. From a drug discovery and pharmaceutical industry perspective, the GPCRs constitute one of the most commercially and economically important groups of proteins known. The GPCRs undertake numerous vital metabolic functions and interact with a hugely diverse range of small and large ligands. Many different methodologies have been developed to efficiently and accurately classify the GPCRs. These range from motif-based techniques to machine learning as well as a variety of alignment-free techniques based on the physiochemical properties of sequences. We review here the available methodologies for the classification of GPCRs. Part of this work focuses on how we have tried to build the intrinsically hierarchical nature of sequence relations, implicit within the family, into an adaptive approach to classification. Importantly, we also allude to some of the key innate problems in developing an effective approach to classifying the GPCRs: the lack of sequence similarity between the six classes that comprise the GPCR family and the low sequence similarity to other family members evinced by many newly revealed members of the family.

1 Introduction

The speed at which new protein sequences are discovered seems to constantly accelerate as new technology, such as third-generation instrumental sequencing, comes on-stream. However, many of these newly-sequenced proteins show strong, or at least overt, similarity to existing sequences: in terms of novelty, the proportion of new yet redundant sequences is also constantly increasing. This may indicate that we are close to defining a final complement of biological sequences available within nature. This virtual "global proteome" is estimated at close to five million sequences [1]. In light of this, and the bewildering productivity of metagenomics [2], the efficient analysis of protein sequences to determine structure and function has become an imperative goal of genomics. Bioinformatics, the application of knowledge-driven science to biological macromolecules and the complex systems that arise from their interactions, can collate and capitalise on this wealth of information by developing efficient and incisive algorithms for its analysis.

The coming step change apparent with biomedicine presents us with a confounding *embarrassment of riches;* a term which incidentally entered English as long ago as 1738 in John Ozell's translation of a French play, L'Embarras des richesses (1726). The *embarrassment of riches* set to become ever richer in complexity and detail yet our inability to deal with it is set to become ever more embarrassing. The English language remains enthral to the common parlance of the past. Even today, in the globalised, multi-cultural twenty-first century, much of our idiomatic speech is submerged by the nautical vocabulary of the Eighteenth century or corralled by the high phrasing of William Tyndale's 1528 bible. In much the same way, our ways of thinking and of acting and of being can remain wedded to the past, hampered and constrained by unnecessary conservatism and caution, as seen in the way that so many cling to outmoded ways of tackling pressing and exigent issues. We need new and innovative approaches to functional genomics as much as we need them in business, politics, and diverse other areas of human activity. Fully rather than partially engaging with the potential of computer science is one such innovation. Experimentalists must allow themselves to be directed by data-hungry informatics analysis, abandoning their rigid instance on necessarily-restrictive experiment-first, hypothesis-driven research in favour of an approach that embraces the full potential of research that is driven instead by the informatics-led discovery and analysis of data [3].

The very large amount of data concerning protein function currently available is of incalculable valuable, since within it is the potential to understand better the nature of disease, and through that understanding to enact better treatment, instantiated by the design of more effective drugs, therapies, and vaccines. However, to properly capitalise on this rare potential, we need to use intelligent data analysis. This has been given the catchy epithet: "data mining" (DM), since it involves "mining", that is, identifying and analysing, raw data; transforming it first into information, then useful knowledge, and eventually, it may be hoped, into true understanding [4].

Biomedical discovery can often hinge upon effective computational sequence analysis revealing important and unexpected functional relationships between members of protein families. DM is a discipline within informatics that seeks to extract patterns from raw data that are useful, non-trivial, and previously unknown [5] [6]. When performing the DM task of classification, algorithms are exposed to input data, identifying therein patterns and regularities which are subsequently used to make predictions. In terms of machine learning terminology, the task of classification is a form of supervised learning where the patterns discovered by DM can be used to infer a class (or classification) of previously unseen or unclassified data; in this way the class (or function) of a protein can be predicted by these previously discovered patterns. Automated function prediction through classification algorithms is now a vital if underappreciated component of biological discovery, since new sequences are currently discovered at such a prodigious rate that classification is a task far beyond anything that manual labelling can achieve. Soon, new genomes may be delivered on a daily basis. The concomitant interpretation-gap will become huge beyond measure and, without automation, making sense of this glut of data will be impossible.

The potential confusion inherent within such a scenario has long been foreshadowed by our experience of one particular protein family: the G-protein coupled receptors or GPCRs. The GPCRs comprise a large and varied multi-gene super-family which consists of integral membrane proteins involved in seemingly innumerable physiological functions [7-8]. GPCRs are, for example, responsible for turning diverse extracellular and endogenous signals into a limited number of intracellular responses. A heterogeneous collection of molecules act as GPCR ligands: these include light, in the form of photons; ions; hormones; neurotransmitters; peptides; and proteins. Since GPCRs are implicit in physiological processes as diverse as neurotransmission, cellular metabolism, secretion, inflammatory responses, and

cellular differentiation [9], they have become a consistent target for the development of medicines: roughly 50% of all marketed drugs target a GPCR [10]. Most anti-GPCR therapies have been derived through the somewhat haphazard processes innate to medicinal chemistry; driven, as it certainly was in times past when GPCR structures were not available, by the whims and caprices of synthetic chemistry rather than the focussed, rationality of structure-based design.

Now, of course, through both sequence analysis and the determination of crystal structures, there is much that initial *in silico* approaches can tell us about newly determined GPCR sequences, including forecasts of its potential function. Despite the diversity of the superfamily, there are many commonalities among GPCRs. Every member of the GPCR superfamily contains seven highly conserved transmembrane segments (of 25-35 consecutive residues), each displaying a high degree of hydrophobicity. Rather than forming a perfect circle or regular ellipse, the seven membrane-crossing α-helical segments (TM1-7) form a flattened two-layer structure known as the transmembrane bundle, which is thought to be common to all GPCRs [11].

Compared to many apparently similar sets of proteins, the GPCRs exhibit a far greater conservation of structure than of sequence. In this review, we seek to place this assertion into context and explore some of its implications. The proteome of a cell is considerably larger, more complex, and certainly more interesting than its genome. Some estimates place the number of proteins encoded by the human genome 2–3 orders of magnitude higher than the number of genes. Mechanisms, including splice variants, inteins, post-translational modifications, cleavage of precursors, and other types of proteolytic activation, magnify the number of protein products by an order of magnitude [12].

Moreover, the proteome is more dynamic than the genome, showing greater differences between individual human patients, and exhibiting more profound differences with cell type, functional status, and disease state. Identifying, cataloguing and characterising an individual's proteome and its complement of GPCRs, will doubtless prove more challenging than did annotating the genome. Analysing proteomes can only be addressed, and addressed convincingly, by coupling the experimental to the computational. 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. By reviewing the various computational techniques used to classify and categorise the GPCRs, we will explore how the severe computational challenge presented by the complex hierarchical classification of the GPCRs has and will be met. Approaches such as those we describe will have important applications not only in discovering and characterising novel protein sequences but also in better understanding the interrelatedness apparent between known members of the GPCR superfamily.

2. Brief Overview of Nomenclature, Classification, and Repertoires

General, immanent difficulties in producing a comprehensive classification system for the protein super-families have existed for some time; as long perhaps as sufficient numbers of protein sequences to warrant such an analysis have been available [13]. Since they are so numerous, and the interrelationships within the group so complex, the GPCRs have proved especially contentious. 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 even be homologous.

Even selecting an appropriate nomenclature has provoked controversy. The term "family" has long been used to describe groupings with the GPCRs. The definition of family does not rely solely on the possession of sequence similarity, but also embraces a larger set of functional, structural, and evolutionary features. In this context, and when viewed rigorously, the term "superfamily" can seem ambivalent and confusing.

Unfortunately, and despite such difficulties, there is not - and possibly there never will be - any overarching term or collective noun able to include and subsume all GPCR sequences. Many have been suggested. Umma, for example, is used as a collective term encompassing the community, the totality of all Muslims, irrespective of all other allegiances. Thus, umma can serve as term able to sit above all other definitions of familial propinquity; it can imply relatedness in some not wholly explicit sense. Another such expression is "clan" [14]: it uses the idea of kinship leniently, recognizing both convergent and divergent evolution. Because of its pervasive usage, we retain here the terms: family and superfamily. They are useful, if

flawed, catch-all expressions and can encompass both homologous and non-homologous protein families.

One of the first GPCR superfamily classification systems was introduced by Kolakowski for the now defunct GCRDb database [15], and further developed by Vriend *et al.* for the GPCRDB database [16-18]. GPCRDB divides the superfamily into six classes. The first of which is Class A, the so-called Rhodopsin-like GPCRs, accounting for over 80% of family members across species. There are around 300 human non-olfactory Class A receptors mostly binding peptides, biogenic amines or lipids [19].

Peptide-binding receptors play important roles in mediating neurotransmitters, hormones and paracrine signals. Receptors which bind to biogenic amines, such as norepinephrine, dopamine, and serotonin, constitute a set of major drug targets, since pathological conditions - schizophrenia, Parkinson's disease, and depression amongst others – are excellent examples of where unbalanced endogenous amine levels lead to altered brain function. Likewise, diligent, sedulous, and assiduous efforts over two decades have led to a remarkable transformation in our knowledge of the structural basis of ligand-mediated functioning of the rhodopsin-like GPCRs: the structure of bovine rhodopsin published in 2000 has been followed more recently by those of ligand-bound avian and human β_1- and β_2-adrenoceptors, and the human A_{2A} adenosine receptor, which were determined in an inactive, ligand-bound conformation [20-24].

The second class is Class B or Secretin-like GPCRs; as a group they have only weak similarity at the sequence level to Class A receptors, despite a presumed similarity of more significant proportions at the structural and functional level [25]. The group is also rather smaller, with only 15 members, which bind large endogenous peptides such as: glucagon; the incretins - glucagon-like peptide 1 (GLP-1) and glucose-dependent insulinotropic polypeptide (GIP); vasoactive intestinal peptide (VIP), secretin and calcitonin; parathyroid hormone (PTH); corticotropin-releasing factor (CRF); growth-hormone releasing factor (GRF); and pituitary adenylate cyclase activating polypeptide (PACAP) [26]. The secretin-like receptors have a large N-terminal extracellular domain of 100 to 160 residues, which plays a vital part in binding ligands. No complete class B structure has been determined, however; nonetheless, a structure corresponding to the N-terminal ligand binding domain has

been solved for 6 different receptor subtypes, which, when combined with biophysical data gives real insight into the structural basis of ligand binding [27-29].

The third class is Class C and comprises the Metabotropic glutamate-like receptors (mGluRs). These excitatory neurotransmitter receptors are activated via an indirect metabotropic process [30]. In humans, mGluRs are found principally within pre- and postsynaptic neurons in the hippocampus, cerebellum and the cerebral cortex, as well as other regions of the brain and in the periphery [31].

The fourth class is Class D, which contains about 20 distinct receptors, is comprised of highly-diverged receptors for peptide pheromones [32] [33]. Class D GPCRs are split between two major subfamilies: Ste2 and Ste3. There is no obvious sequence similarity between these two subfamilies. The two subfamilies are expressed on cells with distinct phenotypes although both receptors types do activate the same G protein signalling pathway. Class D receptors lack many features characteristic of Class A GPCRs. They have no ERY or DRY motif on TM3, no NPxxY motif on TM7, and no disulfide between the end of TM3 and loop 2. The ligand of Ste2 is a small peptide that binds to a surface comprising the ends of the TM helices and the extracellular loop scaffold.

The fifth class of GPCRs is Class E, comprising cAMP receptors from the protozoan amoeba *Dictyostelium discoideum*, which form part of several chemotactic signalling systems [30]. Compared to other lower eukaryotes with sequenced genomes, *Dictyostelium* has over 55 GPCRs: including four receptors for extracellular cAMP [34-35]. cAR1, the best characterized cAMP receptor, is essential for the starvation-induced aggregation of up to 10^5 Dictyostelium amoebae and subsequent development. cAR1 mediates cellular chemotaxis along cAMP gradients and also induces critical aggregation-stage genes, including cAR1 and its G-protein partner Gα2. In addition to class D and E, other groups of GPCRs are only found exclusively outside the subphylum vertebrata, such as the large family of nematode chemosensory receptors [36].

Finally, the sixth class is Class F, which contains Frizzled/smoothened receptors from *Drosophila*, which are necessary for Wnt binding and the mediation of hedgehog signalling respectively [37]. This recently identified a group of 7 TM receptors are considered the most highly diverged, especially with respect to rhodopsin [38].

An alternative, and potentially superior, sequence-based classification system has been proposed for the GPCR family [39-40]. The GRAFS classification system was developed using phylogenetic analysis [41]. GRAFS divides the GPCR superfamily into the *Glutamates, Rhodopsins*, *Adhesions*, *Frizzled/Taste 2* and *Secretin* families, from which the acronym GRAFS is derived. The authors of GRAFS were able successfully to differentiate pseudogenes from functional genes, and were also able to classify all human GPCR and leading to the identification of several new GPCRs [42-49].

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 [50]. In parallel to the GRAFS families, other GPCR families have arisen and/or evolved in specific lineages or species, such as the plant MLO receptors [51] and the insect gustatory receptors [52]. Likewsie, GPCRs in fungi, plants and animals have no sequence similarity, except for one *Adhesion*-like GPCR found in thale cress (*Arabidopsis thaliana*) [53].

Thus, when looked at in isolation, or more typically from a purely anthropomorphic perspective, defining families that span and classify the GPCRs is more taxing than perhaps it should be. For example, the group typified by the vomeronasal type 1 receptors [54], which are involved in pheromone recognition [55], is large and its members numerous in all mammals studied, yet in humans these genes number only five. Other, less easily categorised outlier GPCRs also exist. They can and do prove problematic to both manual and automated classification systems. Examples include the Ocular albinism receptor [56], ITR [57], and GPR108 [58], which are all only present as a single copy in humans.

2.1 GPCR Repertoire

In spite of the high degree of structural similarity within the GPCR superfamily, the totality of the evidence presented above is consistent with the assertion that the discernible similarity across the whole tranche of proteins called or classified as GPCRs is so low as to make a proper, unambiguous phylogenetic analysis of these proteins next to impossible. Thus the lack of overt sequence similarity between GPCR families makes a putative common origin very much an open question. An agnostic view is perhaps warranted, suggesting that most of the various classes described above originated independently during evolution. Of the classes, adhesion and secretin families are most likely to have originated together [59]. This ambiguity is by no means a unique phenomenon in biology. The lipocalin protein is another example [60-68] with many similar features. They too have a highly conserved structure yet their similarity is often so low as to have frustrated its detection.

However, it is still perfectly possible to audit the complement of GPCRs or other proteins within a genome or proteome, irrespective of any spurious or specious evolutionary rationale that one is obliged to impose by biologist's dogma. Various methods have been used to identify this so-called "repertoire" of GPCRs within the various genomes sequenced to date. Repertoires are, in a conceptual sense, genomes within genomes, or sub-genomes: the total number of a particular variety of protein or members of a protein family.

Previous best guesses put the number of GPCRs within the human genome at approximately one percent of total genes, with other estimations putting the number of GPCRs involved in olfaction at an inaccurate and unlikely additional 1000–2000. These figures, and the more accurate numbers given below, should be set against the size of the human genome. Currently, the principal human genome sequence is a composite from five different individuals: its putative size has been whittled or winnowed down from figures in excess of 100,000 first to about 40000 genes and then to around 20,000. A recent and more reliable assessment from 2006 puts the number of coding genes at or about 25,043 [69]; while, a 2007 estimate puts the value at about 20,488, with perhaps another 100 genes yet to be found [44].

An early analysis by Fredriksson *et al*. [40] put the total number of human GPCR genes at 802, while Niimura and Nei have put the current number of olfactory receptor (OR) genes at 388 and pseudogenes at 414 [70]. Subsequently, Fredriksson

and co-workers, and indeed several other groups as well, have been able to identify many new rhodopsin-like and adhesion-like GPCRs in the burgeoning suite of genomes available for study. In light of this the size of any genome and the number of GPCRs within it must remain educated guesses. While both will alter, particularly as the genomes of individual humans are sequenced, we can be reasonably confident that the majority genes and most GPCRs have been discovered.

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 [71]. Takeda and colleagues extracted approximately 950 open reading frames from the human genome that had 200–1500 amino acid residues similar to those of GPCRs [72]. The GPCR repertoires of several other species have also been published, including mouse [41], rat [46], chicken [45], pufferfish [73], amongst several others. The recently-determined G protein-coupled receptor repertoire within the dog genome was shown to be more similar to that found in humans than that found in rodents [42].

2.2 GPCR Training and Test Sets for Classification Algorithms.

In the machine learning scenario, a classification algorithm is trained with examples (i.e. GPCRs) with known classes and the classification model discovered from this set is used to predict the classes of further examples drawn from a separate test set, which were unseen during training. Issues of clarity, precision and bias are faced when we try to define the training and test sets to be used in GCPR classification. It is clearly worth ventilating some of the more apposite issues here. Usually, one would expect verification through the use of independent test data to be ideal; however, things can be deceptive. In general, the choice of both training and testing examples is important. Predicting examples very similar to training data is typically much easier prospect than predicting instances which are wildly unalike. Consider the classification task of predicting whether or not a protein is a GPCR. In this task the training set would contain, as positive examples, proteins known to be GPCR, whilst the negative examples would be proteins known not to be GPCRs. It is possible, if tedious, to create data sets containing positive and negative examples

which will favour good validation statistics. For example, if we have a valid positive set of GPCRs, we could choose very different sequences – say small globular proteins or sequences with extreme amino acid compositions or whatever – as negative examples. However, if one chooses as negative examples proteins which are similar to GPCRs – membrane proteins of a similar size composition and a similar number of transmembrane helices – then the task would seem to become very much demanding. What can be done to circumvent such problems? We can propose that a cascade of different negative sets of increasing difficulty is likely to be a more reliable and accurate test of a method's effectiveness. Independent tests should if possible be conducted in a double blind fashion, since almost invariably when an author is party to an evaluation (and thus influences the choice of the test and the way that the test is conducted) it is never truly independent. It is well known that published comparative tests, as conducted by a particular experimenter, will typically favour their own work over that of others.

The above discussion considered the classification task of discriminating GPCRs from non-GPCRs. However, other types of classification problems can be defined for GPCRs. In particular, one can have a training set where all examples are known to be GPCRs, and then try to predict to which class a given GPCR protein belongs. For example, when developing GPCR classification algorithms, Davies *et al*. [74] built as large and comprehensive a dataset of GPCR sequences as possible with which to train and test the classifier. All protein sequences for the dataset were obtained from Entrez [75] using text-based searching and these were used to construct each GPCR sub-family and Class level dataset. Only human proteins sequences were incorporated, with the exception of Class D proteins, which are found only in fungi, and Class E, which is only found in *Dictyostelium*. Atypically short, and probably incomplete, GPCR sequences less than 280 amino acids were removed, as were all duplicate sequences. Thus the construction of this data set, which relied on accumulated annotations extant within database, also relied on the insight, and the bias, of the many, many investigators who worked on the problem over the decades.

3.0 Sequence based approaches to the Annotation of GPCRs

A central dogma within bioinformatics, which provides a key justification and raison d'être for the discipline, is that there exists a strong and readily-discernible

relationship between the sequence of a protein and its structure. This relationship provides a *de facto* link between sequence and function, as function arises largely from the dynamic 3-dimensional structure of a folded protein not from its sequence. This implies that the function of a newly discovered protein can be determined, at least in part, by determining the similarity of its sequence to the sequence of a known and studied protein.

As we mentioned above, the virtual global proteome nears completion [1]. That is not to say that all sequences will have been determined but that within the next few years all truly distinct sequences are expected to be discovered. From then on, with rare exceptions or the discovery of novel forms of life, such perhaps as any organic component of nanobacteria, sequencing will move from a bold and exciting era of discovery to a more mundane era of cataloguing and correction. To make a fanciful if not wholly inappropriate allusion, we will move from the era of Magellan, Cook, and Amundsen, to the era of the global positioning system. The glamour and excitement then will move inexorably away from the low hanging fruit of genome sequencing and focus instead on the use of the information in productive ways. Analysis of the GPCRs is just one such endeavour.

We do know that GPCRs with the same ligands can and do bind different G proteins. We also know that GPCRs binding the same G protein can bind completely different ligands [76-77]. Likewise, two GPCRs may bind the same ligand and bind the same G protein yet have less than 25% sequence identity. GPCRs for melanocortin, lysophosphatidic acid and sphingosine 1-phosphate have an atypically high degree of sequence similarity yet their functions are unrelated. Clearly, this is and may remain an on-going challenge; which is another way of saying that it is a difficult and complex problem demanding a solution. Fortunately, the importance of the GPCR as physiological agents and drug targets more than justifies our efforts to address and resolve this challenge.

3.1 Full-length Sequence Searching Approaches to the Discovery and Annotation of GPCRs.

Arguably, the most obvious and straightforward approach to characterising a protein sequence usually involves searching a sequence database - which contains within it sets of previously annotated sequences – using a pair-wise similarity tool,

such as FastA [78] or BLAST (Basic Local Alignment Search Tool) [79]. BLAST searches typically reveal obvious similarities between the query and one or more sequences in the database, as determined from pair-wise alignments along with concomitant statistical significance. Proteins are listed, as ranked by expectation or 'E' values. Such values are a measure of the reliability of the similarity calculated by the method. Low E values are more significant, implying the greater reliability of the identified relatedness between the two sequences.

For a GPCR query, most proteins sitting at the top of the list, and thus evincing high sequence identity, will be true GPCRs. BLAST searches have often identified new GPCR proteins, mainly where there is detectable sequence similarity to other GPCR sequences, a situation which is becoming increasingly uncommon. Currently, this kind of obvious similarity is harder and harder to find, as bioinformaticians find themselves increasingly working at the margins.

In the present era, BLAST, while well-used, is often of limited value for hunting out new members of the GPCR superfamily, since there is often a low degree of sequence similarity between the six families and between outliers within groups. An ideal result will show unambiguous similarity to a well-characterised protein over the full length of the query. However, often outputs contain no significant hits. Obviously, a more typical state of affairs would fall between such extremes, affording a list of incomplete matches to a wide variety of proteins. Many of these hits will be uncharacterised or have dubious or contradictory annotations [80]. The difficulty then lies in the reliable inference of homology (the verification of a divergent evolutionary relationship) and, from this, the extrapolation to biological function.

Why is this? As the size of sequence databases rises inexorably, and are increasingly contaminated by populations of poor-quality or partial sequences, the probability of making high-scoring yet actually random matches will also rise. Moreover, if not appropriately masked, hits matching atypical regions may swamp and thus obscure search outputs. Many sections of protein sequences are atypical: some have repetitious sequences where the same pattern is repeated many times over. Others have what is called low sequence complexity, where one or two residue types are used to the exclusion of all others [81]. This contrasts with normal protein sequences where the usage and repetition of each of the twenty amino acids varies little from the perfect average of 5%.

The modular or multi-domain structure of many proteins is also a problem. It may not be obvious, when matching to many concurrent domains, which corresponds correctly to the query. Even when the correct domain has been identified, direct transfer of extant functional annotation may not be appropriate, since the function of the domain may be quite different. Even if a wholly correct and validated match can be discovered, pairwise similarity struggles to distinguish orthologues from paralogues. Thus, in sum and to lapse into the vernacular, BLAST is very much a blunt instrument, particularly for the fine-detail analysis of large and/or complex protein families. Similar phenomena bedevil the analysis of other protein families, such as the Lipocalins [60-61, 82-85]. Similarity determined over the full sequence length simple cannot, when used alone, offer all answers to all the questions. Only the integrated and integrative use of a diversity of complementary techniques, each with their own strengths and weakness can achieve such an objective.

3.2 Motif-based approaches to the Discovery and Annotation of GPCRs.

To a first approximation, BLAST generates generic, full-length similarities between sequences, while so-called motif-based approaches focus on specific, length-restricted traits unique to families or sub-families. Many protein family databases - most famously typified by PROSITE [86] or PFAM [87], and latterly subsumed by InterPro [88], a system combining sequence profiles from several databases – are built on such an approach. They use multiple alignments to identify highly conserved regions that can form the basis of characteristic, and even diagnostic, motifs for family or subfamily membership.

Of available approaches – single motifs through to HMM models of entire sequences – perhaps the more informative are so-called “fingerprints”. GPCR fingerprints have been developed using patterns of common conservation within the seven transmembrane regions [89] [90] [91]. Rather than identifying a single, lone motif, fingerprinting looks at many short yet conserved regions within the sequence group. Sub-family and sub-sub-family level fingerprints are derived from segments within the TM regions, parts of the loops and parts of the N- and C-termini. False positives are readily determined since typically sequences will lack one or more of the motifs.

The PRINTS database system [91] contains within it hundreds of GPCR fingerprints. Individual motifs within such fingerprint can reflect structurally or functionally important sections of sequence, say a TM domain or a ligand-binding site. PRINTS has been demonstrated to identify similarities between receptors with low sequence similarity: it allows a user to find the GPCR superfamily to which a particular query sequence belongs (i.e. at the level of rhodopsin-like versus secretin-like, *etc*.); the family to which it belongs (e.g., muscarinic versus adrenergic, *etc*.); and also its subtype (i.e. muscarinic M1, M2, etc.). However, as known members become more numerous, it becomes ever harder to define fingerprints with synoptic precision. Nor can very atypical GPCR sequences be easily identified using the fingerprint method or indeed other methods.

Holden & Freitas [92] classified GPCRs using three different kinds of motifs: PROSITE patterns, PRINTS fingerprints and InterPro [88] entries. Three different GPCR datasets were created. Each dataset used a different set of attributes: 338 proteins and 281 attributes were derived from PRINTS; 194 proteins and 127 attributes from PROSITE; and 584 proteins and 448 attributes from InterPro. Holden & Freitas used a swarm intelligence algorithm [93] for GPCR classification.

Their algorithm induced sets of IF-THEN classification rules. These took the form: *IF <set of motifs is present> THEN <predict a certain class>*. The motifs forming these sets could come from either PROSITE, PRINTS, or InterPro. The goal of this work was to find the most discriminating set of motifs which formed the most accurate rule. PRINTS motifs performed best (89.6% classification accuracy at the family level), InterPro marginally worse (86.3% classification accuracy at the family level), while PROSITE patterns performed poorly. Substantially lower accuracy rates were obtained for sub-families and below.

3.3 Machine Learning and Statistical Pattern Recognition approaches to the Discovery and Annotation of GPCRs: Artificial Neural Networks, Hidden Markov Models, and Support Vector Machines

In many cases, conventional bioinformatics techniques, such as global sequence searching and/or motif matching, can determine useful information from a sequence through pair-wise alignment or by comparing the sequence to previously determined motifs. Although such an alignment or motif based approach is without question

valid, it may not always be optimal when trying to identify GPCRs. 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 and subjective manual intervention. One should also remember that conventional biochemically-based GPCR Classification schemes were created using the identity of the ligand to which the receptor binds not sequence similarity. A more computationally sophisticated, if not necessarily a more effective approach to the GPCR classification problem is through use of techniques based on Machine Learning, a branch of Artificial Intelligence or statistical pattern recognition.

As we have stressed, an intrinsic limitation of any supervised learning algorithm in the classification scenario 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. Predictive models should always be tested for accuracy before being used. A particularly suitable method of assessing a model's accuracy is by using cross-validation. This method splits the full data set into several - typically 10 - independent, non-overlapping sets. The classification algorithm is executed many times. During each cycle, one of the partitions is kept back as the test set and the algorithm is trained on the remainder. Accuracy is assessed using the single partition that had been kept back previously. In this way, it can be hoped that test and training sets become as representative as possible.

The goal of all GPCR classification studies is to take a protein or genome sequence and classify it. To achieve this, predictive methods – speaking in generalities – are in want of three things: appropriate structural representations of the sequences being used to train them; an equivalent list of biological status (i.e. membership of superfamily, family, type, and subtype) corresponding to these sequences; and some form of computational engine capable of predicting relationships between the particular description of the sequences and their associated status. The outcome of this process is a predictive model relating structure to status (class). The induction engine can take any one of many different forms, such as machine learning or statistical pattern recognition techniques, as described here.

An example of Artificial Neural Networks (ANNs) in the analysis of GPCR data is the use of Self-Organising Maps (SOMs) [94]. SOMs perform unsupervised learning (in this case, clustering) to discriminate protein families from each other. 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. The overall performance of the map can be assessed using the sensitivity and specificity values as well as calculating the total accuracy of the clustering. Otaki *et al.* [95] reported a 97.4% precision at clustering 12 Class A sub-families using SOMs.

A Hidden Markov model or HMM is a statistical model where the system being modelled is assumed to be a Markov process with unknown parameters. In a Markov process, the probability distribution describing future states depends solely on the present state not on states prior to that: the future depends upon the present not the past. In a regular Markov model, the state is seen by the observer and thus only state transition probabilities are parameters. In a HMM, the state is not visible, although variables influenced by that state can be seen, and so the aim is to determine the hidden parameters from the observable parameters. HMMs have gained significant currency, particularly when used for sequence alignment [96].

Support Vector Machines (SVMs) are machine-learning algorithms based on statistical learning theory [97]. In two-class problems, an SVM maps two sets of distinct data representing sequence descriptions onto a multi-dimensional feature space and then sets about constructing a division between the classes. The optimal division is one with a maximum distance to the closest data point from each of the two classes. Finding this optimal division is important since should another data point be added, it is easier to classify it correctly when there is a significant separation between classes. The data points nearest to the optimal division are termed support vectors. Although SVMs are more commonly used to solve 2-class problems, this technique can been applied to GPCR classification with more than two classes by running the algorithm multiple times (i.e. once for each class) [98].

3.4 Alignment Free Methods approaches to the Discovery and Annotation of GPCRs: Proteochemometrics, Properties, and Statistics

Rather than aligning sequences, and from such alignments deducing pseudo-evolutionary relationships, alignment-independent classification systems use the physiochemical properties of amino acids to give insight into functionally or structurally important differences between sequences. To enable this process, we need to turn the symbolic structure of the protein into a set of numbers.

Proteochemometrics is an example of such an approach; it has been applied to the classification of the GPCR superfamily. Proteochemometrics uses Wold's five Z values which encode key properties of the twenty biogenic amino acids [99-104]. Recondite electronic effects are described by the Z4 and Z5 values. Z3 values describe amino acid polarity. Polar or hydrophilic amino acids have large positive values, while non-polar amino acids have large negative values. Size or, more properly, volume properties are accounted for by Z2 values. Large negative values correspond to low volume amino acids while large positive numbers indicate amino acids with large volume and surface area. Z1 values account for amino acid lipophilicity: a large negative value corresponds to a lipophilic amino acid, and vice-versa.

Replacing each amino acid in the sequence with these five Z values, and then transforming it in some manner, reduces a protein sequence to the required numerical description. The resulting, normalized matrix is analysed using Principal Component Analysis (PCA) and Partial Least Squares (PLS), generating a classification model. Using the proteochemometrics method, Lapnish *et al.* developed a model with an accuracy of 0.76 for a diverse set of amine GPCRs [103].

Kim *et al.* also developed a physico-chemically based classification method [105], which separates sequences into specific categories using a linear discriminant function, called a Quasi-predictor Feature Classifier (QFC) algorithm, within a statistically-defined 'feature space'. The resulting model was used to screen databases for novel GPCRs. The QFC approach was trained on 750 GPCRs from the GPCRDB and 1000 randomly chosen non-GPCR proteins of 200-1000 amino acids in length. Several amino acid property scales were examined and the values normalized using a sliding window. Windows comprising 13-16 amino acids were more effective than those of 32 or 64 amino acids. Test sets of 100 GPCRs and 100 non-GPCRs were classified with a 99% accuracy; versus 530 ion channels (non-GPCR transmembrane proteins), QFC was 96.4% accurate. QFC had a higher false positive rate than many motif-based techniques, which is consistent with the approach needing more filtering.

Huang [106] used Quinlan's C4.5 algorithm [107] to induce a decision tree partitioning 4395 GPCR sequences into 5 Classes, 39 sub-families, 93 sub-sub-families, and types. Each protein was represented as a vector comprising its normalised composition. C4.5 chooses to split data by selecting the composition feature that best discriminates the classes to be predicted. Division continues until a

defined stopping criterion is attained. The technique was 86.9% accurate at the sub-family level and 81.5% accurate for sub-sub-families.

4.0 CLASSIFYING GPCR HIERARCHICALLY

In the last decade or so, networks have emerged from relative obscurity to become a powerful, pervasive, perhaps even dominant, description of complex biological, chemical, and physical systems. Such networks are often organized as a hierarchy. Networks arise in nature in many areas, and biology in particular is replete with examples: epidemiology, neural networks for locomotion, vision, speciation and the synchronous flashing of fireflies.

The standard depiction of a hierarchy represents it in the form of a tree or dendrogram. The word dendrogram comes from the Greek *dendron* meaning "tree" and *gramma* meaning "drawing". Within such trees, collections of nodes are divided into groups; groups that split into a succession of sub-divisions, usually over several levels. Arguably, the most familiar and immediate dendrogram is to most of us a family tree. In a family tree, individuals are connected to their siblings through direct parental lineage, and are connected to their cousins through a shared lineage in prior generations: grandparents and great-grandparents, and so on.

Going beyond mere straightforward clustering, the hierarchical nature inherent within networks implies significant organization at many levels. Sub-groups within networks often correspond to well-understood functional units: modules in protein-protein interaction (PPI) networks, metabolic networks, or genetic regulatory networks; communities in social networks; or ecological niches in food webs. Hierarchies explain many general network properties, including their highly connected nature. The implication is that the hierarchy is a key organizing principle common to all complex networks, however they arise.

Traditionally, readily discernible groups within a network are deemed assortative or disassortative. In assortative networks, groups comprise members which are highly, and close to equally, interconnected, yet there are relatively fewer connections between groups. Disassortative networks are characterised by a few highly-connected hubs and many other nodes with very few connections. The majority of networks within molecular biology, including metabolic and PPI networks, are of this type.

A network, hierarchical or otherwise, can be viewed as a graph, consisting of nodes or vertices linked together by edges or connections. Nodes are instances; edges their interactions. Edges can be bidirectional (two-way interactions) or unidirectional (one-way interactions). A graph where all edges are unidirectional is known as a directed graph. In a graph, edges may be weighted to indicate how strong the corresponding interactions are. Weighted edges may be different or nominally identical.

Compared to many other problems explored so thoroughly by computer science and computer scientists, hierarchical classification is a relatively under-investigated special case of the broader and rather better-known area labelled classification. Rather than trying to predict a set of classes without discernible and overt complex mutual dependency, classes are arranged in a furcating hierarchy of many levels [108]. In an informatic context, higher-level classes correspond to general functions, whilst lower-level classes correspond to more specific functions. Specific classes will inherit the features or functions of their parents. An important distinction, when doing hierarchical classification, is whether the class hierarchy has the structure of a tree or direct-acyclic-graph (DAG). In a "tree-structured class hierarchy" each class has just one "parent class", although a parent class can have many child classes. In a "DAG-structured class hierarchy" one class can potentially have several parents. The GPCR hierarchy is tree-structured, whilst other hierarchies, such as the Gene Ontology (GO), are DAG-structured.

Such hierarchies are obvious within extant pharmacological and sequence analysis of the GPCR family. It is explicit and implicit within many databases. At the level of superfamilies, motifs within a database tend to encode universally common features such as TM helices. At the family level, motifs encode regions uniquely characteristic of a particular family, distinguishing it from others. These regions are usually small parts of TM or loop regions, which are often involved in ligand binding. Interestingly, at the sub-family level, the majority of the motifs are found in the extracellular loops yet at the sub-subfamily level, most originate within intracellular loops.

There exist several strategies for predicting hierarchical classes [108]. The simplest is to flatten the dataset to the deepest (most specific) level of the hierarchy, then use one of many standard “non-heirarchical” classification algorithms to predict classes. This strategy is wasteful: it fails to take full advantage of the information implicit

within the class hierarchy and tends to obtain smaller predictive accuracy than truly hierarchical classification approaches, especially when the number of classes is large.

Very different is the global approach. This utilises a single, but complex, hierarchical classification algorithm that considers all hierarchical relationships among all classes in a single execution of the algorithm. Perhaps due to its complexity, implementations of such an approach are scarce, although one such model has been used for gene function prediction in *Saccharomyces cerevisiae* [109].

A different hierarchical classification strategy is the so-called local method, nicknamed the top-down approach. In some sense, this represents an intermediate approach between the extremes of algorithmic simplicity and complexity associated with the flat and big-bang approaches. In the top-down approach, during training, a tree of local classifiers is constructed. This tree of local classifiers mirrors the distribution of classes in the class tree. In the testing phase, examples are classified by first exposing the unknown example to the classifier at the root node. A prediction is made and used to decide to which child (1st level) classifier the example will be transferred. This process continues until all examples have been categorised by the classifier at a leaf node. Usually, every node uses the same classification algorithm to formulate the local classifier at that node.

In light of the inherently hierarchical nature of networks - both networks constructed and observed by the human mind - we have attempted to make use of this powerful idea and constructed a means of using the concept of a hierarchy directly to help us automate the prediction of the many-levelled GPCR classification. Davies *et al*. [74] developed an alternative to this top-down hierarchical classification method by using a set of potential classification algorithms at each node, and selecting the best classifier at that stage. This 'selective top-down approach' has been found to outperform a standard top-down classifier in the GPCR classification task. However, the selective approach can be slow in training, particularly if the number of classes is large.

In the examples cited above, the data set contained 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. Moreover, rather than use the primary sequence to perform the classification, which would rule out the use of the vast majority of traditional classifiers, the system uses an alternative form of

protein data representation. Their approach used a simplified version of the representation used in proteochemometrics, making use of Wold's five z-values.

Since data mining algorithms do not tend to work with variable numbers of descriptors, it proved necessary to normalize the z-values so that all proteins had the same number of variables. Davies *et al.* [74] used a normalization method which calculated the arithmetic mean for each *z* value over the whole protein. This very simple technique was found to retain predictive accuracy while significantly reducing requirements for storage and processing time.

Later, they made use of an augmented version of this attribute creation method. In this case, 15 attributes described each protein. Five were created as described above but in addition to this, five more were created from the N-terminus of the protein and five from the C-terminus. Therefore, in the case of the N-terminus, the means for each of the five *z*-values were computed for only the first 150 amino acids, while in the case of the C-terminus, the means over the last 150 amino acids were also determined.

The selective top-down approach was implemented in WEKA [6]: and made use of 10 classification techniques, namely: Naïve Bayes method, a Bayesian network, a SVM [110], nearest neighbour, the PART rule induction algorithm, J48 (WEKA's implementation of the well-known C4.5 decision tree induction algorithm), a Naïve Bayes tree, a multi-layer neural network with back propagation, an artificial immune system, and a conjunctive rule learner. Our approach exploits the idea that some characteristics may be important to distinguish protein subsets at one classification level while being relatively unimportant at another. Classifiers were chosen in a training data-driven manner. Since different classifiers may be more suited to certain nodes of the class hierarchy, it was assumed that the overall classification accuracy should increase if a diverse set of algorithms is used within the hierarchy.

Compared to the standard top-down approach with the same classifier at every node, this selective approach involves additional steps: at each node, training data is randomly assigned between sub-training and validation sets; several classifiers are then trained using this sub-training data set and tested using the validation data set. The classifier with the greatest resulting classification accuracy is then selected for that particular node. Finally, the selected classifier is re-trained using the full original training data.

The headline accuracy statistics for the selective top-down approach were 95.87% at the family level (5 classes), 80.77% at the sub-family level (38 classes), and

69.98% at the sub-subfamily level (87 classes). Overall, our approach out-performed all single-classifier methods, and also compared favourably with several extant servers using both our data sets and theirs.

Recently, this approach has also been extended in 2 ways. First, we used our adaptive approach to reduce the innate complexity involved in representing the 20 standard amino acids. There exists no universally-applicable reduced alphabet despite the numerous criteria upon which to group the amino acids. An optimization algorithm was developed to identify the most efficient grouping when classifying GPCRs [111-112].

Secondly, through the addition of an attribute selection step, where selection occurs independently at each node in a data-driven manner, leaving only those attributes that best discriminate classes at that node [113]. The number of attributes selected was highly variable between classifier nodes, but varied little for the same node during repeated cross-validation. Thus, it was postulated that the attribute selection method is reacting to the varying levels of difficulty of predicting particular classes at different positions in the class tree. It was found that the addition of this attribute selection stage made no significant difference to the predictive accuracy of the hierarchical classification system, yet did afford the advantage that the processing time was reduced significantly.

5.0 Prediction Servers

A question commonly articulated by those tasked with discovering and understanding biology, rather than those tasked with developing tools to analyse data, is this: what is the value of creating methods that cannot be used? Few have the time to implement all published methods that might prove useful, so it is that the prevalence of internet servers comes to resolve this dilemma. Many such servers are now available for the identification and classification of GPCRs. We adumbrate a few below.

HMMTOP (Hidden Markov Model for TOpology Prediction), available at URL: http://www.enzim.hu/hmmtop/, uses the idea that TM regions have the most atypical amino acid composition [114], and that such region may be determined by looking for maximum divergence rather than identifying regions with a specific composition. 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.

TMHMM (Transmembrane Hidden Markov Model), URL: http://www.cbs.dtu.dk/services/TMHMM/, predicts transmembrane helices by using a HMM to partition a protein sequence into the most probable distribution compared to known GPCRs [115]. Models are estimated using a maximum likelihood and a discriminative method. This method consistently displays a high false positive rate, and many proteins with seven transmembrane helices are incorrectly predicted as possessing six or eight TM regions. Interestingly, when HMMTOP and TMHMM are combined they have a higher overall success rate (0.819) than when used separately (0.808 and 0.762).

GPCRHMM, URL: http://noble.gs.washington.edu/~lukall/gpcrhmm/, implements an HMM that specifically recognises GPCRs. Wistrand *et al.* [116] found distinct loop length patterns and differences in amino acid composition between cytosolic loops, extracellular loops and membrane regions.

Pred-GPCR (http://athina.biol.uoa.gr/bioinformatics/PRED-GPCR/) [117] combined fast fourier transforms with SVMs to leverage sequence hydrophobicity in the identification of GPCRs. 403 sequences from 17 sub-families from GPCR Classes B, C, D and F were used to train the program. Optimal performance reached an accuracy of 93.3%, and the accuracies for different subfamilies varied between 66.7 and 100%. However, 105 of the 403 sequences originated from the frizzled/smoothened family and there were typically 10-20 sequences per subgroup. Given this unusual distribution within the training set, it seems unlikely that the classification model would prove highly predictive when applied to larger GPCR datasets with a less atypical class distribution.

GPCRsClass [118], URL: http://www.imtech.res.in/raghava/gpcrsclass/, is a SVM-based server that focuses on Class A GPCRs. Using dipeptide composition, GPCRsClass is 99.7% accurate at dividing amine from non-GPCRs, and 92% accurate when splitting sequences into sub-subfamilies. A similar program from the same stable, GPCRPred, URL: http://www.imtech.res.in/raghava/gpcrpred/, first determines if a sequence is a GPCR, then which class it belongs to, and finally, assuming it is a Class A GPCR, to which subfamily it belongs [119]. GPCR vs non-GPCR sequences had 99.5% accuracy, the Class prediction was 97.3% accurate, and the sub-family was on average 85% accurate.

The hierarchical approach to GPCR classification developed by Secker, Davies, and co-workers, and described at length above, has also been made available freely over the world wide web, implemented within the webserver GPCRTree [120]; URL: http://igrid-ext.cryst.bbk.ac.uk/gpcrtree/.

Certain other servers, such as The GPCR Subfamily Classifier (former URL: http://www.soe.ucsc.edu/research/compbio/gpcr-subclass/), have now been retired from active service, while others, including a variant of the TMHMM program, called 7TMHMM, URL: http://tp12.pzr.uni-rostock.de/~moeller/7tmhmm/, are only available for download. Many, many servers exist for predicting transmembrane proteins, notably: PRED-TMR2 [121] and TMHMM 2.0 [122]. For a putative GPCR, such servers can be used as a semi-independent check on the location of TM helices.

6.0 Discussion and Conclusion

Currently, molecular biology is generating prodigious quantities of genomic data, as a result of a number of automated experiments. It is tempting to conceive of the genome (the entire complement of genes comprising an organism) as an "encoded text" that is progressively decoded to produce proteins. But to decode this properly, that is in terms of function and interaction we need to leverage our knowledge in a dynamic and adaptive fashion. Knowledge is information that has been properly organised, integrating raw and visceral fact with complex, subtle, multi-levelled meaning. Understanding is the deepest and most profound cognitive level available to us; where we would be able not just to catalogue bald observations, but gain true and pervasive insight into not just how something happens but why and in what way it happens. In other words, beyond knowledge is understanding, an almost intuitive grasp of the whys that enables us to manipulate and to manage the world of our existence, not just experience it. With the emergence of knowledge and understanding comes the ability to make intuitive predictions, to organize rationally, and to design and create; in short, to escape the limitations imposed by precedent and provenance, and remake the familiar world in the image of our aspirations.

However, it is, is it not always worth challenging our assumptions? Thus, at this stage, it might be appropriate to ask an important question: why is it important to address so well studied a superfamily as the GPCRs? There are several answers to this question. One answer is at once societal, economic, and commercial in nature. Unmet

medical need, real or perceived, is a strong driver of the world economy. GPCRs remain important drug targets still dominating a major, if diminishing, proportion of global pharmaceutical sales [10] [123] [124].

During 2009, 26 new drugs were formally approved by the FDA. Of these, around a fifth were molecules which directly targeted one or more GPCRs. Saphris or asenapine, a product from Merck-Organon, is a mixed antagonist which is active at a variety of aminergic rhodopsin-like GPCRs including serotonin receptors, dopamine receptors, adrenergic receptors, and histamine receptors. It acts as a treatment for acute episodes of bipolar disorder and schizophrenia. Bepreve or bepotastine, another of the five new GPCR-targetted drugs, is an ophthalmic H1-receptor antagonist able to inhibit histamine release from mast cells. Marketed by Ista pharmaceuticals it treats itchy eyes triggered by pollen, plants and other irritants. Fanapt or iloperidone, another treatment for schizophrenia, is a product of Vanda Pharmaceuticals licensed in the USA and Canada by Novartis. It is a dopamine, serotonin, and norepinephrine receptor antagonist. Prasugrel is a P2Y12 receptor antagonist that helps manage acute coronary syndrome; it is marketed by Eli Lilly and Daiichi Sankyo's and is intended to rival the Sanofi-Aventis and Bristol-Myers Squibb blockbuster plavix. Samsca or tolvaptan is a vasopressin V2-receptor antagonist which acts as a treatment of hyponatremia.

Additionally, three compounds – Bromocriptine, Imitrex, and Tenex – have been approved as switched therapies, and again all three target GPCRs. Together, this indicates that despite the emergence of a veritable plethora of alternative target families, from the Kinases onwards, GPCRs retain their place amongst the elite of accessible, profitable drug targets.

Having said all that, it must be remembered that the pharmaceutical industry has ceased to be the unstoppable engine of unalloyed, ever-increasing profitability that once it was. 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%. In 2006, for example, worldwide sales of prescription medicines rose by a modest 7% to around $602 billion; and recall that 2006 was economically a very good year long before stock market turbulence and the cupidity of the wealthy plunged the world in financial turmoil. Established pharmaceutical companies have been inconvenienced by the coincidence of incipient product droughts, caused 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. 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 dollars a year.

Available biological databanks are now remarkably numerous; indeed, they require a database of their own to catalogue them [125], thus necessitating a significant degree of automated rather than manual data analysis. Manual analysis is, to lapse into the vernacular, slow but sure; automated analysis can be rapid yet uncertain. Annotation or labels within sequence databases are often inferred from observed similarities to homologous, annotated proteins. This may create error, particularly when sequence similarity is marginal. As a result, it is widely believed that there are now substantial numbers of incorrect annotations throughout commonly-used databases [126].

This problem is compounded by the Markov process of 'error percolation' [127], whereby the functional annotations of similar proteins may themselves have been acquired through chains of similarity to sets of other proteins. Such chains of inference are seldom recorded, so it is generally impossible to determine how a particular database annotation has been acquired. Such transitive approaches necessarily lead to a systematic deterioration of database quality, and pose a continuing threat to information reliability through the propagation of incorrect annotations. Although curators are aware of these problems, and constantly strive to reduce errors, nonetheless, databases are historical products, and users should always bear in mind their possible imperfections.

Other answers are more biological in nature. Another confounding issue for GPCRs within the human genome is the presence of polymorphisms, where numerous variant forms, or alleles, of a single gene exist within a population. Such "mutations", both inactivating and activating, are implicated in many genetic disorders necessitating genome-wide association studies of SNPs to identify critical target receptors for disease susceptibility. Because pharmacogenomic and pharmacogenetic studies on GPCRs are currently very rare, the therapeutic relevance of receptor alleles often remains unclear. Anti-schizophrenic drugs, for example, bind to a number of receptors, several with potential therapeutic value. Without a proper understanding of

the role of receptor alleles in clinical outcome, or of how sequence variation affects signal transduction, it remains difficult to predict their individual relevance.

According to the IUPHAR database, URL: http://www.iuphar-db.org/index.jsp, there are currently over ninety so-called orphan GPCRs, showing palpable sequence similarity to Class A Rhodopsin-like receptors, for which there are no known ligands or functions [43]. There are also over 30 orphan receptors in the Class B GPCR family, and around 10 in Class C. Such numbers are likely to decrease, at least within the human genome, as tenacious experimentation slowly chips away at such lacuna in our knowledge.

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 [128-129]. 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 [130]. $GABA_{B1}$ 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 non-orphan GPCR cell surface expression, in either a positive [131] or negative way [128]. If this is true then the relative expression of orphan and non-orphan GPCR proteins could be an important factor for the regulation of cell signalling. There has also been considerable interest in the tendency of GPCRs to form higher order oligomers in living cells [132]. Dimeric ligands linked by spacer arms have been used to identify the importance of co-expression of certain GPCR subtypes, indicating that the formation of these oligomers is a crucial part of GPCR signalling, although the extent to which oligomerisation occurs across the whole GPCR superfamily remains uncertain.

Speaking more generally, the search for novel GPCRs in a genome of interest, whatever it may be, is confounded by issues that arise from the complex nature of multi-gene families: database search techniques cannot easily differentiate between proteins that have arisen by a process of speciation (so-called orthologues, where the functional counterpart of a sequence is found in another species) and those that have

arisen via intra-species duplication and divergence (so-called paralogues, which may perform related but distinct functions within the same organism).

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, allowing the user to see why a GPCR has been assigned a particular classs, but have been observed to have false positive and false negative prediction rates that are erratic. Models constructed by SVMs (Support Vector Machines) or ANNs (Artificial Neural Networks) 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 [133].

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 non-GPCRs would be less effective at the class, sub-family or sub-sub-family 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.

Many commentators have questioned the long term viability of the so-called "blockbuster" drug, suggesting that the already fragmented pharmaceutical market is moving towards an era characterised by even more extensive fragmentation, where a plethora of individual, highly-focused markets are dominated not by a handful of big sellers but by a legion of niche products. Because of their well-characterised, easily druggable binding site, coupled to their vital biological roles, GPCRs remain the ultimate drug 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.

References

1. Perez-Iratxeta, C., G. Palidwor, and M.A. Andrade-Navarro, *Towards completion of the Earth's proteome.* EMBO Rep, 2007. **8**(12): p. 1135-41.
2. Simon, C. and R. Daniel, *Achievements and new knowledge unraveled by metagenomic approaches.* Appl Microbiol Biotechnol, 2009. **85**(2): p. 265-76.
3. Kell, D.B. and S.G. Oliver, *Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era.* Bioessays, 2004. **26**(1): p. 99-105.
4. Weber, G.W., S. Ozogur-Akyuz, and E. Kropat, *A review on data mining and continuous optimization applications in computational biology and medicine.* Birth Defects Res C Embryo Today, 2009. **87**(2): p. 165-81.
5. Fayyad, U., G. PiatetskyShapiro, and P. Smyth, *From data mining to knowledge discovery in databases.* AI Magazine, 1996. **17**(3): p. 37-54.
6. Witten, I.H. and E. Frank, *Data mining : practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann series in data management systems. 2000, San Francisco, Calif.: Morgan Kaufmann. xxv, 371 p.
7. Bissantz, C., *Conformational changes of G protein-coupled receptors during their activation by agonist binding.* J Recept Signal Transduct Res, 2003. **23**(2-3): p. 123-53.
8. Tuteja, N., *Signaling through G protein coupled receptors.* Plant Signal Behav, 2009. **4**(10): p. 942-7.
9. Hebert, T.E. and M. Bouvier, *Structural and functional aspects of G protein-coupled receptor oligomerization.* Biochem Cell Biol, 1998. **76**(1): p. 1-11.
10. Flower, D.R., *Modelling G-protein-coupled receptors for drug design.* Biochim Biophys Acta, 1999. **1422**(3): p. 207-34.
11. Yeagle, P.L. and A.D. Albert, *G-protein coupled receptor structure.* Biochim Biophys Acta, 2007. **1768**(4): p. 808-24.
12. Korner, M. and L.J. Miller, *Alternative splicing of pre-mRNA in cancer: focus on G protein-coupled peptide hormone receptors.* Am J Pathol, 2009. **175**(2): p. 461-72.
13. Cheng, B.Y., J.G. Carbonell, and J. Klein-Seetharaman, *Protein classification based on text document classification techniques.* Proteins, 2005. **58**(4): p. 955-70.
14. Attwood, T.K., et al., *PRINTS and PRINTS-S shed light on protein ancestry.* Nucleic Acids Res, 2002. **30**(1): p. 239-41.
15. Kolakowski, L.F., Jr., *GCRDb: a G-protein-coupled receptor database.* Receptors Channels, 1994. **2**(1): p. 1-7.
16. Horn, F., et al., *GPCRDB information system for G protein-coupled receptors.* Nucleic Acids Res, 2003. **31**(1): p. 294-7.
17. Horn, F., G. Vriend, and F.E. Cohen, *Collecting and harvesting biological data: the GPCRDB and NucleaRDB information systems.* Nucleic Acids Res, 2001. **29**(1): p. 346-9.
18. Horn, F., et al., *GPCRDB: an information system for G protein-coupled receptors.* Nucleic Acids Res, 1998. **26**(1): p. 275-9.
19. Fridmanis, D., et al., *Formation of new genes explains lower intron density in mammalian Rhodopsin G protein-coupled receptors.* Mol Phylogenet Evol, 2007. **43**(3): p. 864-80.

20. Bokoch, M.P., et al., *Ligand-specific regulation of the extracellular surface of a G-protein-coupled receptor.* Nature, 2010. **463**(7277): p. 108-12.
21. Cherezov, V., et al., *High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor.* Science, 2007. **318**(5854): p. 1258-65.
22. Rosenbaum, D.M., et al., *GPCR engineering yields high-resolution structural insights into beta2-adrenergic receptor function.* Science, 2007. **318**(5854): p. 1266-73.
23. Day, P.W., et al., *A monoclonal antibody for G protein-coupled receptor crystallography.* Nat Methods, 2007. **4**(11): p. 927-9.
24. Rasmussen, S.G., et al., *Crystal structure of the human beta2 adrenergic G-protein-coupled receptor.* Nature, 2007. **450**(7168): p. 383-7.
25. Parthier, C., et al., *Passing the baton in class B GPCRs: peptide hormone activation via helix induction?* Trends Biochem Sci, 2009. **34**(6): p. 303-10.
26. Chapter, M.C., et al., *Chemical modification of class II G protein-coupled receptor ligands: frontiers in the development of peptide analogs as neuroendocrine pharmacological therapies.* Pharmacol Ther, 2010. **125**(1): p. 39-54.
27. Dong, M., R.F. Cox, and L.J. Miller, *Juxtamembranous region of the amino terminus of the family B G protein-coupled calcitonin receptor plays a critical role in small-molecule agonist action.* J Biol Chem, 2009. **284**(33): p. 21839-47.
28. Miller, L.J. and F. Gao, *Structural basis of cholecystokinin receptor binding and regulation.* Pharmacol Ther, 2008. **119**(1): p. 83-95.
29. Dong, M., et al., *Insights into the structural basis of endogenous agonist activation of family B G protein-coupled receptors.* Mol Endocrinol, 2008. **22**(6): p. 1489-99.
30. Pin, J.P., T. Galvez, and L. Prezeau, *Evolution, structure, and activation mechanism of family 3/C G-protein-coupled receptors.* Pharmacol Ther, 2003. **98**(3): p. 325-54.
31. Brauner-Osborne, H., P. Wellendorph, and A.A. Jensen, *Structure, pharmacology and therapeutic prospects of family C G-protein coupled receptors.* Curr Drug Targets, 2007. **8**(1): p. 169-84.
32. Burkholder, A.C. and L.H. Hartwell, *The yeast alpha-factor receptor: structural properties deduced from the sequence of the STE2 gene.* Nucleic Acids Res, 1985. **13**(23): p. 8463-75.
33. Eilers, M., et al., *Comparison of class A and D G protein-coupled receptors: common features in structure and activation.* Biochemistry, 2005. **44**(25): p. 8959-75.
34. Eichinger, L. and A.A. Noegel, *Comparative genomics of Dictyostelium discoideum and Entamoeba histolytica.* Curr Opin Microbiol, 2005. **8**(5): p. 606-11.
35. Eichinger, L., et al., *The genome of the social amoeba Dictyostelium discoideum.* Nature, 2005. **435**(7038): p. 43-57.
36. Troemel, E.R., et al., *Divergent seven transmembrane receptors are candidate chemosensory receptors in C. elegans.* Cell, 1995. **83**(2): p. 207-18.
37. Prabhu, Y. and L. Eichinger, *The Dictyostelium repertoire of seven transmembrane domain receptors.* Eur J Cell Biol, 2006. **85**(9-10): p. 937-46.
38. Chou, K.C. and D.W. Elrod, *Bioinformatical analysis of G-protein-coupled receptors.* J Proteome Res, 2002. **1**(5): p. 429-33.

39. Schioth, H.B. and R. Fredriksson, *The GRAFS classification system of G-protein coupled receptors in comparative perspective.* Gen Comp Endocrinol, 2005. **142**(1-2): p. 94-101.
40. Fredriksson, R., et al., *The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints.* Mol Pharmacol, 2003. **63**(6): p. 1256-72.
41. Bjarnadottir, T.K., et al., *Comprehensive repertoire and phylogenetic analysis of the G protein-coupled receptors in human and mouse.* Genomics, 2006. **88**(3): p. 263-73.
42. Haitina, T., et al., *The G protein-coupled receptor subset of the dog genome is more similar to that in humans than rodents.* BMC Genomics, 2009. **10**: p. 24.
43. Gloriam, D.E., R. Fredriksson, and H.B. Schioth, *The G protein-coupled receptor subset of the rat genome.* BMC Genomics, 2007. **8**: p. 338.
44. Nordstrom, K.J., et al., *Comprehensive comparisons of the current human, mouse, and rat RefSeq, Ensembl, EST, and FANTOM3 datasets: identification of new human genes with specific tissue expression profile.* Biochem Biophys Res Commun, 2006. **348**(3): p. 1063-74.
45. Lagerstrom, M.C., et al., *The G protein-coupled receptor subset of the chicken genome.* PLoS Comput Biol, 2006. **2**(6): p. e54.
46. Gloriam, D.E., H.B. Schioth, and R. Fredriksson, *Nine new human Rhodopsin family G-protein coupled receptors: identification, sequence characterisation and evolutionary relationship.* Biochim Biophys Acta, 2005. **1722**(3): p. 235-46.
47. Gloriam, D.E., et al., *The repertoire of trace amine G-protein-coupled receptors: large expansion in zebrafish.* Mol Phylogenet Evol, 2005. **35**(2): p. 470-82.
48. Gloriam, D.E., et al., *High species variation within the repertoire of trace amine receptors.* Ann N Y Acad Sci, 2005. **1040**: p. 323-7.
49. Bjarnadottir, T.K., et al., *The human and mouse repertoire of the adhesion family of G-protein-coupled receptors.* Genomics, 2004. **84**(1): p. 23-33.
50. Fredriksson, R. and H.B. Schioth, *The repertoire of G-protein-coupled receptors in fully sequenced genomes.* Mol Pharmacol, 2005. **67**(5): p. 1414-25.
51. Devoto, A., et al., *Molecular phylogeny and evolution of the plant-specific seven-transmembrane MLO family.* J Mol Evol, 2003. **56**(1): p. 77-88.
52. Hill, C.A., et al., *G protein-coupled receptors in Anopheles gambiae.* Science, 2002. **298**(5591): p. 176-8.
53. Josefsson, L.G. and L. Rask, *Cloning of a putative G-protein-coupled receptor from Arabidopsis thaliana.* Eur J Biochem, 1997. **249**(2): p. 415-20.
54. Dulac, C. and R. Axel, *A novel family of genes encoding putative pheromone receptors in mammals.* Cell, 1995. **83**(2): p. 195-206.
55. Kouros-Mehr, H., et al., *Identification of non-functional human VNO receptor genes provides evidence for vestigiality of the human VNO.* Chem Senses, 2001. **26**(9): p. 1167-74.
56. Bassi, M.T., et al., *Cloning of the gene for ocular albinism type 1 from the distal short arm of the X chromosome.* Nat Genet, 1995. **10**(1): p. 13-9.
57. Tsukada, S., et al., *Inhibition of experimental intimal thickening in mice lacking a novel G-protein-coupled receptor.* Circulation, 2003. **107**(2): p. 313-9.

58. Edgar, A.J., *Human GPR107 and murine Gpr108 are members of the LUSTR family of proteins found in both plants and animals, having similar topology to G-protein coupled receptors.* DNA Seq, 2007. **18**(3): p. 235-41.
59. Nordstrom, K.J., et al., *The Secretin GPCRs descended from the family of Adhesion GPCRs.* Mol Biol Evol, 2009. **26**(1): p. 71-84.
60. Flower, D.R., A.C. North, and C.E. Sansom, *The lipocalin protein family: structural and sequence overview.* Biochim Biophys Acta, 2000. **1482**(1-2): p. 9-24.
61. Akerstrom, B., D.R. Flower, and J.P. Salier, *Lipocalins: unity in diversity.* Biochim Biophys Acta, 2000. **1482**(1-2): p. 1-8.
62. Flower, D.R., *The lipocalin protein family: structure and function.* Biochem J, 1996. **318 (Pt 1)**: p. 1-14.
63. Flower, D.R., et al., *The first prokaryotic lipocalins.* Trends Biochem Sci, 1995. **20**(12): p. 498-9.
64. Flower, D.R., *Multiple molecular recognition properties of the lipocalin protein family.* J Mol Recognit, 1995. **8**(3): p. 185-95.
65. Flower, D.R., *The lipocalin protein family: a role in cell regulation.* FEBS Lett, 1994. **354**(1): p. 7-11.
66. Flower, D.R., *Structural relationship of streptavidin to the calycin protein superfamily.* FEBS Lett, 1993. **333**(1-2): p. 99-102.
67. Flower, D.R., A.C. North, and T.K. Attwood, *Structure and sequence relationships in the lipocalins and related proteins.* Protein Sci, 1993. **2**(5): p. 753-61.
68. Flower, D.R., A.C. North, and T.K. Attwood, *Mouse oncogene protein 24p3 is a member of the lipocalin protein family.* Biochem Biophys Res Commun, 1991. **180**(1): p. 69-74.
69. Clamp, M., et al., *Distinguishing protein-coding and noncoding genes in the human genome.* Proc Natl Acad Sci U S A, 2007. **104**(49): p. 19428-33.
70. Niimura, Y. and M. Nei, *Evolution of olfactory receptor genes in the human genome.* Proc Natl Acad Sci U S A, 2003. **100**(21): p. 12235-40.
71. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
72. Takeda, S., et al., *Identification of G protein-coupled receptor genes from the human genome sequence.* FEBS Lett, 2002. **520**(1-3): p. 97-101.
73. Metpally, R.P. and R. Sowdhamini, *Genome wide survey of G protein-coupled receptors in Tetraodon nigroviridis.* BMC Evol Biol, 2005. **5**: p. 41.
74. Davies, M.N., et al., *On the hierarchical classification of G protein-coupled receptors.* Bioinformatics, 2007. **23**(23): p. 3113-8.
75. Wheeler, D.L., et al., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2008. **36**(Database issue): p. D13-21.
76. Muramatsu, T. and M. Suwa, *Statistical analysis and prediction of functional residues effective for GPCR-G-protein coupling selectivity.* Protein Eng Des Sel, 2006. **19**(6): p. 277-83.
77. Yabuki, Y., et al., *GRIFFIN: a system for predicting GPCR-G-protein coupling selectivity using a support vector machine and a hidden Markov model.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W148-53.
78. Lipman, D.J. and W.R. Pearson, *Rapid and sensitive protein similarity searches.* Science, 1985. **227**(4693): p. 1435-41.

79. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
80. Altschul, S.F., et al., *Issues in searching molecular sequence databases.* Nat Genet, 1994. **6**(2): p. 119-29.
81. Wootton, J.C., *Non-globular domains in protein sequences: automated segmentation using complexity measures.* Comput Chem, 1994. **18**(3): p. 269-85.
82. Salier, J.P., et al., *Lipocalins in bioscience: the first family gathering.* Bioessays, 2004. **26**(4): p. 456-8.
83. Paine, K. and D.R. Flower, *The lipocalin website.* Biochim Biophys Acta, 2000. **1482**(1-2): p. 351-2.
84. Flower, D.R., *Experimentally determined lipocalin structures.* Biochim Biophys Acta, 2000. **1482**(1-2): p. 46-56.
85. Flower, D.R., *Beyond the superfamily: the lipocalin receptors.* Biochim Biophys Acta, 2000. **1482**(1-2): p. 327-36.
86. Sigrist, C.J., et al., *PROSITE, a protein domain database for functional characterization and annotation.* Nucleic Acids Res, 2010. **38**(Database issue): p. D161-6.
87. Finn, R.D., et al., *The Pfam protein families database.* Nucleic Acids Res, 2010. **38**(Database issue): p. D211-22.
88. Hunter, S., et al., *InterPro: the integrative protein signature database.* Nucleic Acids Res, 2009. **37**(Database issue): p. D211-5.
89. Attwood, T.K., *A compendium of specific motifs for diagnosing GPCR subtypes.* Trends Pharmacol Sci, 2001. **22**(4): p. 162-5.
90. Flower, D.R. and T.K. Attwood, *Integrative bioinformatics for functional genome annotation: trawling for G protein-coupled receptors.* Semin Cell Dev Biol, 2004. **15**(6): p. 693-701.
91. Mulder, N.J., et al., *New developments in the InterPro database.* Nucleic Acids Res, 2007. **35**(Database issue): p. D224-8.
92. Holden, N. and A.A. Freitas, *A hybrid particle swarm/ant colony algorithm for the classification of hierarchical biological data.* 2005 IEEE Swarm Intelligence Symposium, 2005: p. 100-107.
93. Holden, N. and A.A. Freitas, *Hierarchical classification of protein function with ensembles of rules and particle swarm optimisation.* Soft Computing, 2009. **13**(3): p. 259-272.
94. Yan, A.X., *Application of self-organizing maps in compounds pattern recognition and combinatorial library design.* Combinatorial Chemistry & High Throughput Screening, 2006. **9**(6): p. 473-480.
95. Otaki, J.M., et al., *Alignment-free classification of G-protein-coupled receptors using self-organizing maps.* Journal of Chemical Information and Modeling, 2006. **46**(3): p. 1479-1490.
96. Gollery, M., *Handbook of hidden Markov models in bioinformatics*. 2008, Boca Raton, Fla. ; London: Chapman & Hall/CRC. xix, 156 p.
97. Cristianini, N. and J. Shawe-Taylor, *An introduction to Support Vector Machines : and other kernel-based learning methods*. 2000, Cambridge: Cambridge University Press. xi, 189 p.
98. Lorena, A.C. and A.C.P.L.F. de Carvalho, *Comparing techniques for multiclass classification using binary SVM predictors.* Micai 2004: Advances in Artificial Intelligence, 2004. **2972**: p. 272-281.

99. Strombergsson, H., et al., *Rough set-based proteochemometrics modeling of G-protein-coupled receptor-ligand interactions.* Proteins-Structure Function and Bioinformatics, 2006. **63**(1): p. 24-34.
100. Lapinsh, M., et al., *Improved approach for proteochemometrics modeling: application to organic compound - amine G protein-coupled receptor interactions.* Bioinformatics, 2005. **21**(23): p. 4289-4296.
101. Lapinsh, M., et al., *Proteochemometrics modeling of the interaction of amine G-protein coupled receptors with a diverse set of ligands.* Molecular Pharmacology, 2002. **61**(6): p. 1465-1475.
102. Lapinsh, M., et al., *Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions.* Biochimica Et Biophysica Acta-General Subjects, 2001. **1525**(1-2): p. 180-190.
103. Lapinsh, M., et al., *Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences.* Protein Science, 2002. **11**(4): p. 795-805.
104. Freyhult, E., et al., *Unbiased descriptor and parameter selection confirms the potential of proteochemometric modelling.* BMC Bioinformatics, 2005. **6**: p. -.
105. Kim, J., et al., *Identification of novel multi-transmembrane proteins from genomic databases using quasi-periodic structural properties.* Bioinformatics, 2000. **16**(9): p. 767-75.
106. Huang, Y., et al., *Classifying G-protein coupled receptors with bagging classification tree.* Comput Biol Chem, 2004. **28**(4): p. 275-80.
107. Quinlan, J.R., *C4.5 : programs for machine learning*. 1993, San Mateo, Calif.: Morgan Kaufmann. x,302p.
108. Silla, C.N. and A.A. Freitas, *A survey of hierarchical classification across different application domains.* Data Mining and Knowledge Discovery, 2010.
109. Clare, A. and R.D. King, *Predicting gene function in Saccharomyces cerevisiae.* Bioinformatics, 2003. **19 Suppl 2**: p. ii42-9.
110. Keerthi, S.S. and E.G. Gilbert, *Convergence of a generalized SMO algorithm for SVM classifier design.* Machine Learning, 2002. **46**(1-3): p. 351-360.
111. Davies, M.N., et al., *Optimizing amino acid groupings for GPCR classification.* Bioinformatics, 2008. **24**(18): p. 1980-6.
112. Secker, A., et al., *An artificial immune system for evolving amino acid clusters tailored to protein function prediction.* Artificial Immune Systems, Proceedings, 2008. **5132**: p. 242-253.
113. Secker, A.A., et al., *Hierarchical classification of G-Protein-Coupled Receptors with data-driven selection of attributes and classifiers.* . Int. J. Data Mining and Bioinformatics, 2010. **4**(2): p. 19.
114. Tusnady, G.E. and I. Simon, *The HMMTOP transmembrane topology prediction server.* Bioinformatics, 2001. **17**(9): p. 849-850.
115. Inoue, Y., Y. Yamazaki, and T. Shimizu, *How accurately can we discriminate G-protein-coupled receptors as 7-tms TM protein sequences from other sequences?* Biochem Biophys Res Commun, 2005. **338**(3): p. 1542-6.
116. Wistrand, M., L. Kall, and E.L. Sonnhammer, *A general model of G protein-coupled receptor sequences and its application to detect remote homologs.* Protein Sci, 2006. **15**(3): p. 509-21.
117. Papasaikas, P.K., et al., *PRED-GPCR: GPCR recognition and family classification server.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W380-2.

118. Bhasin, M. and G.P. Raghava, *GPCRsclass: a web tool for the classification of amine type of G-protein-coupled receptors.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W143-7.
119. Bhasin, M. and G.P. Raghava, *GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W383-9.
120. Davies, M.N., et al., *GPCRTree: online hierarchical classification of GPCR function.* BMC Res Notes, 2008. **1**: p. 67.
121. Pasquier, C., et al., *A novel method for predicting transmembrane segments in proteins based on a statistical analysis of the SwissProt database: the PRED-TMR algorithm.* Protein Eng, 1999. **12**(5): p. 381-5.
122. Krogh, A., et al., *Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.* J Mol Biol, 2001. **305**(3): p. 567-80.
123. Drews, J., *Drug discovery: a historical perspective.* Science, 2000. **287**(5460): p. 1960-4.
124. Klabunde, T. and G. Hessler, *Drug design strategies for targeting G-protein-coupled receptors.* Chembiochem, 2002. **3**(10): p. 928-44.
125. Discala, C., et al., *DBcat: a catalog of 500 biological databases.* Nucleic Acids Res, 2000. **28**(1): p. 8-9.
126. Linial, M., *How incorrect annotations evolve--the case of short ORFs.* Trends Biotechnol, 2003. **21**(7): p. 298-300.
127. Gilks, W.R., et al., *Percolation of annotation errors through hierarchically structured protein sequence databases.* Math Biosci, 2005. **193**(2): p. 223-34.
128. Levoye, A., et al., *Do orphan G-protein-coupled receptors have ligand-independent functions? New insights from receptor heterodimers.* EMBO Rep, 2006. **7**(11): p. 1094-8.
129. Levoye, A., et al., *Are G protein-coupled receptor heterodimers of physiological relevance?--Focus on melatonin receptors.* Chronobiol Int, 2006. **23**(1-2): p. 419-26.
130. Pin, J.P., et al., *Allosteric functioning of dimeric class C G-protein-coupled receptors.* FEBS J, 2005. **272**(12): p. 2947-55.
131. Milasta, S., et al., *Interactions between the Mas-related receptors MrgD and MrgE alter signalling and trafficking of MrgD.* Mol Pharmacol, 2006. **69**(2): p. 479-91.
132. Casado, V., et al., *GPCR homomers and heteromers: a better choice as targets for drug development than GPCR monomers?* Pharmacol Ther, 2009. **124**(2): p. 248-57.
133. Freitas, A.A., D.C. Weiser, and R. Appweiler, *On the importance of comprehensible classification model for protein function prediction. .* IEEE/ACM Trans. on Computational Biology and Bioinformatics, 2010. **7**(1): p. 10.

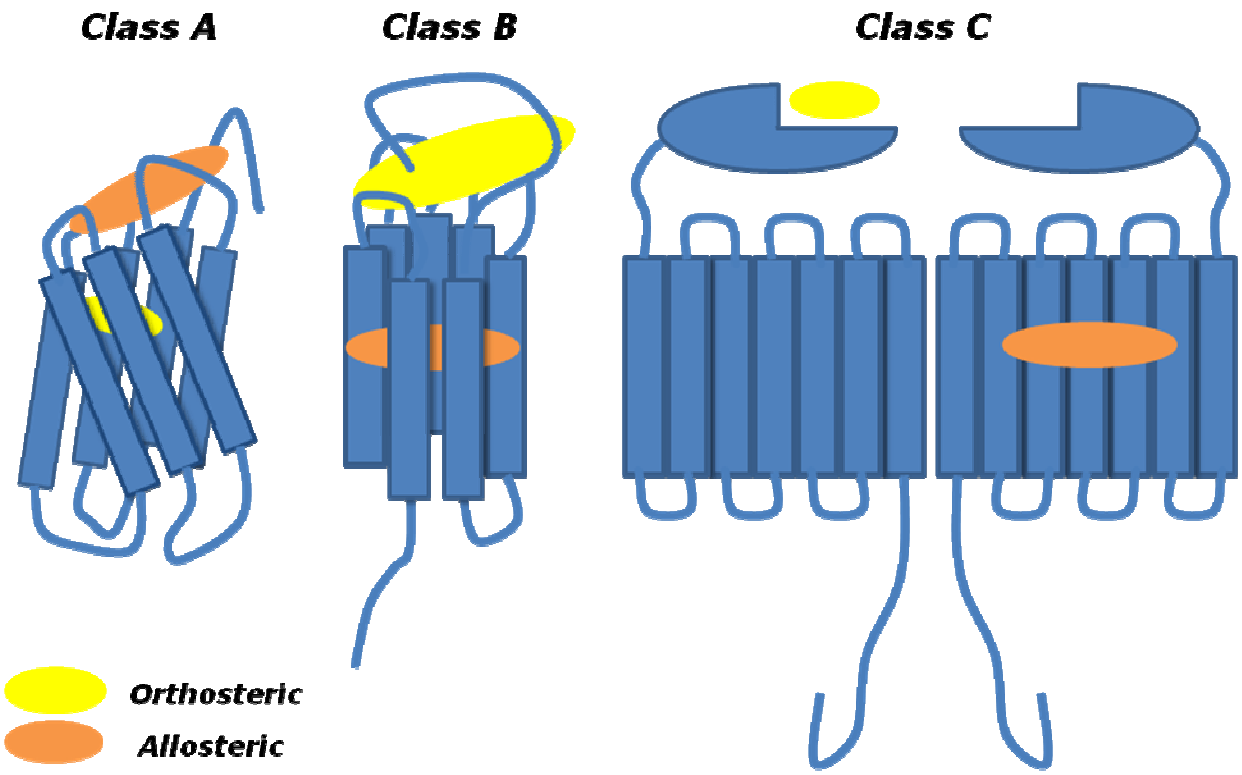

Figure 1. Schematic of the three principal Classes of the GPCRs.

A schematic comparison of the structural differences between main classes of GPCR stressing the divergent nature of ligand binding. By definition, the binding site of the natural, physiological ligand is called orthosteric, whereas binding sites located elsewhere are termed allosteric.

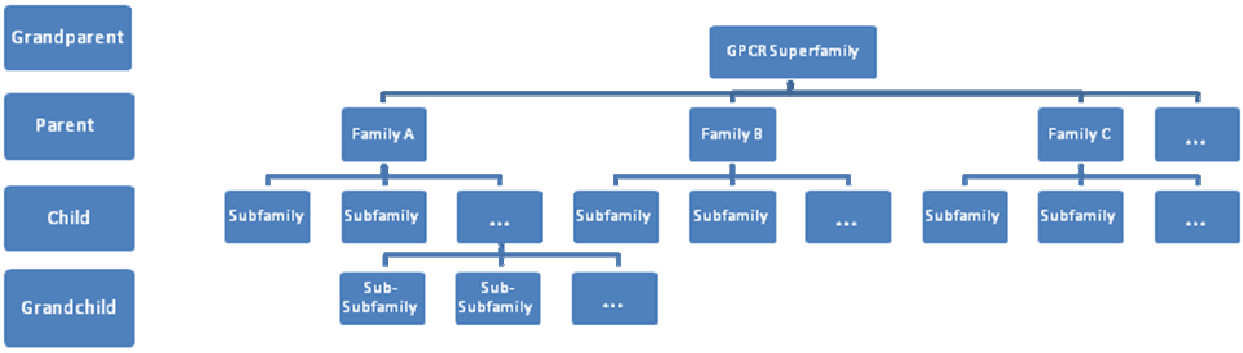

Figure 2. Four level hierarchy characteristic of the G Protein Coupled receptors (GPCRs)

A schematic dendrogram that illustrates the four level hierarchy that characterises the G Protein Coupled receptor (GPCR) umma, as used in our analysis [74].

Figure Error! No text of specified style in document.**. Prototypical schematic of a data and classifier hierarchy**

Example of a hierarchical dataset (1) and how that hierarchy may be reflected in a tree of classifiers (2) ready for a top-down approach to classification.