BMC Research Notes

Technical Note

GPCRTree: online hierarchical classification of GPCR function

Matthew N Davies*[1], Andrew Secker[2], Mark Halling-Brown[3], David S Moss[3], Alex A Freitas[2], Jon Timmis[4], Edward Clark[4] and Darren R Flower[1]

Address: [1]The Jenner Institute, University of Oxford, Compton, Newbury, Berkshire, RG20 7NN, UK, [2]Department of Computing and Centre for BioMedical Informatics, University of Kent, Canterbury, Kent, CT2 7NF, UK, [3]Department of Crystallography, Birkbeck College, University of London, Malet Street, London, WC1E 7HX, UK and [4]Departments of Computer Science and Electronics, University of York, Heslington, York, YO10 5DD, UK

Email: Matthew N Davies* - m.davies@mail.cryst.bbk.ac.uk; Andrew Secker - andysecker@gmail.com; Mark Halling-Brown - m.halling-brown@mail.cryst.bbk.ac.uk; David S Moss - d.moss@mail.cryst.bbk.ac.uk; Alex A Freitas - A.A.Freitas@kent.ac.uk; Jon Timmis - jt517@ohm.york.ac.uk; Edward Clark - edclark@cs.york.ac.uk; Darren R Flower - darren.flower@jenner.ac.uk

* Corresponding author

Published: 21 August 2008

BMC Research Notes 2008, **1**:67 doi:10.1186/1756-0500-1-67

Received: 8 August 2008
Accepted: 21 August 2008

This article is available from: http://www.biomedcentral.com/1756-0500/1/67

Abstract

Background: G protein-coupled receptors (GPCRs) play important physiological roles transducing extracellular signals into intracellular responses. Approximately 50% of all marketed drugs target a GPCR. There remains considerable interest in effectively predicting the function of a GPCR from its primary sequence.

Findings: Using techniques drawn from data mining and proteochemometrics, an alignment-free approach to GPCR classification has been devised. It uses a simple representation of a protein's physical properties. GPCRTree, a publicly-available internet server, implements an algorithm that classifies GPCRs at the class, sub-family and sub-subfamily level.

Conclusion: A selective top-down classifier was developed which assigns sequences within a GPCR hierarchy. Compared to other publicly available GPCR prediction servers, GPCRTree is considerably more accurate at every level of classification. The server has been available online since March 2008 at URL: http://igrid-ext.cryst.bbk.ac.uk/gpcrtree/.

Background

The G protein-coupled receptors (GPCR) comprise a diverse range of integral membrane proteins regulating many important physiological functions [1-3]. Ligand binding to a GPCR on the cell surface initiates cell signaling. An extremely heterogeneous set of molecules act as GPCR ligands. The GPCRs are a common target for therapeutic drugs and approximately 50% of all marketed drugs target GPCRs [4,5]. In spite of their functional and sequence diversity, GPCRs share certain common structural features, but show a far greater conservation of three-dimensional structure than primary sequence [6]. This makes it difficult to develop for GPCR subtypes a comprehensive classification system based on sequence [7]. The most commonly-used system of classification is that implemented in the GPCRDB database [8], which divides the GPCRs into six classes (Class A: Rhodopsin-like, with over 80% of all GPCRs in humans; Class B: Secretin-like; Class C: Metabotropic glutamate receptors; Class D: Pheromone receptors; Class E: cAMP receptors; and the much smaller Class F: Frizzled/smoothened family). Classes A, B, C and F are found in mammalian species while Class D proteins are found only in fungi and Class E proteins are exclusive to *Dictyostelium*. The six classes are further

divided into sub-divisions and sub-sub-divisions based on the function of a GPCR and its specific ligand.

Previous attempts at classifying the GPCRs from its primary sequence have included motif-based classification tools [9,10] and machine learning methods such as Hidden Markov Models [11,12] and Support Vector Machines (SVMs) [13]. Several publicly-available SVM-based GPCR classifiers exist: PRED-GPCR [14,15], GPCR-PRED [16] and GPCRsclass [17]. Some predictive techniques have used a combination of SVMs and HMMs [18]. Other approaches towards GPCR Classification have included Self-Organising Maps [19], Quasi-predictor Feature Classifiers [20] and Decision Trees [21]. GPCRTree is a new publicly-available server based on the idea of selecting the best classifier (from a set of candidate classifiers) at each node of the GPCR class tree.

Findings

Algorithm

A previously-constructed comprehensive GPCR sequences dataset was used to train and test the classifier [22]. Proteins shorter than 280 amino acids were removed, eliminating incomplete protein sequences. All identical sequences were removed to avoid redundancy and classes with fewer than 10 examples were also removed. The dataset used to train the server contains 8222 protein sequences in 5 classes at the family level (A-E), 38 classes at the sub-family level, and 87 classes at the sub-subfamily level. Class F was not considered since it contains too few sequences to develop an accurate classification model. The system uses an alignment-independent classification system based on amino acids physical properties. Proteochemometrics uses 5 "z-values" (z1–z5) derived from 26 real physiochemical properties using principal component analysis [23,24]. These five values are calculated for each amino acid in the sequence and are used to generate the 15 attribute values described in [17], giving a purely numerical description of the protein.

The GPCRTree server classifies at the GPCR Class, Subfamily and Sub-Subfamily level. Hierarchical classification of a sequence is performed using a selective top-down approach, whereby each group of sibling nodes in the GPCR class tree becomes a flat classification problem solved using a standard classifier [25,26], obviating the need to devise a novel classifier. The full dataset trains the root classifier, while only relevant subsets of the data are used to train classifiers at the subfamily and sub-subfamily levels. When an unclassified sequence is presented to the algorithm, the root level classifier assigns it to a class, which is then passed down to an appropriate classifier at the next level until it is assigned to a subfamily and a sub-subfamily [27]. Instead of a single classification algorithm being used at each node of the class tree, many classifiers are trained using a subset of the training set called the sub-training set, and then tested using a separate part of the training set called the validation set. The classifier with the highest classification accuracy on the validation set is selected for that node. Eight standard classification algorithms were used as candidate classifiers at each node of the GPCR tree. All code was written using the open source WEKA data mining package [28,29] and the default parameters were used for each algorithm.

Testing

The GPCRTree server has been validated against three other predictive GPCR servers [22]. The GPCRTree server was trained using the full GPCRtree dataset, and then tested with each GPCR server dataset as test data. GPCRTree produced accuracies of 97% at the Class level, 84% at the Sub-family and 75% at the Sub-Subfamily level. This exceeded the PRED-GPCR server at the Class level and is comparable at the Sub-family level. It exceeds the GPCRPred server at all levels of the hierarchy. The GPCRsclass server was the most successful classifier at the most specific (sub-sub-family) level; this may be because the classifier is overly specialised, being applicable only to the Class A Amine sub-subfamily level. Of servers applicable to all GPCR classes, GPCRTree is the most accurate GPCR prediction server currently available.

Implementation

GPCRTree is available through a web interface – http://igrid-ext.cryst.bbk.ac.uk/gpcrtree/. It was implemented using PHP, dHTML and a java client. The PHP interface affords a simple and straightforward method to submit a protein sequence for evaluation. The code for the selective top-down approach, as previously published, required several changes to facilitate its effective integration into the server environment. Training was modified such that all GPCR proteins belonging to a class with 10 or more examples (protein sequences) were used. The algorithm then pauses and waits for input that will come as an auxiliary program making a TCP socket connection with the selective top down classifier. Upon connection, the auxiliary program will send the protein sequence to be classified and then pause. The classifier will make a prediction and then return the result. A TCP connection has been used for several reasons. It can allow multiple users to access the classifier. Separate users can run separate auxiliary programs, and so the classifier can queue these requests ensuring that only one will invoke the classifier at any given time. The remainder will be queued and serviced in the order of submission. Moreover, this architecture promotes portability. It may be necessary, for resource or security reasons, to run the classifier on different hardware. In this case, the server can invoke the auxiliary program which can connect via network connection to the separate machine running the classifier.

A user enters a protein sequence in plain or fasta format and submits the job (Figure 1). The interface then sends an AJAX call to the java client. The GPCRTree java client submits sequences to the GPCRTree server, where they are classified and the classification returned to the java client which in turn passes this result to the interface. The sequence and classification is then displayed below the submission button (Figure 2).

Where non-standard residues are included within the sequences, substitutions are made: a sequence containing a 'B' (asparagine or aspartic acid) is assigned as an asparagine 'N'; a 'Z' (glutamine or glutamic acid) is assigned as a glutamine 'Q'; and a 'U' (selenocysteine) is assigned as a cysteine 'C'. All unknown residues 'X' were assigned as alanines 'A'.

Conclusion

GPCR classification is among the most challenging problems in bioinformatics due to the sequence diversity of the GPCR superfamily and the uneven distribution of its various family subgroups. GPCRTree is the first server to implement an alignment-independent representation of protein sequences and is also the first to classify sequences using a classifier specifically selected for each group of sibling nodes in the GPCR functional classification tree. By selecting the best classifier (from a set of candidate classifiers) at each GPCR class tree node, the selective top-down method effectively exploits the fact that different classifiers have different biases that are more suitable for different classification problems. GPCRTree is currently the most accurate publicly-available server for the prediction of GPCR sequence classification and it utilises a simple yet robust interface that can undertake multiple classifications simultaneously.

Availability and requirements

Project name: GPCRTree

Project home page: http://igrid-ext.cryst.bbk.ac.uk/gpcrtree/

Operating system(s): Platform independent

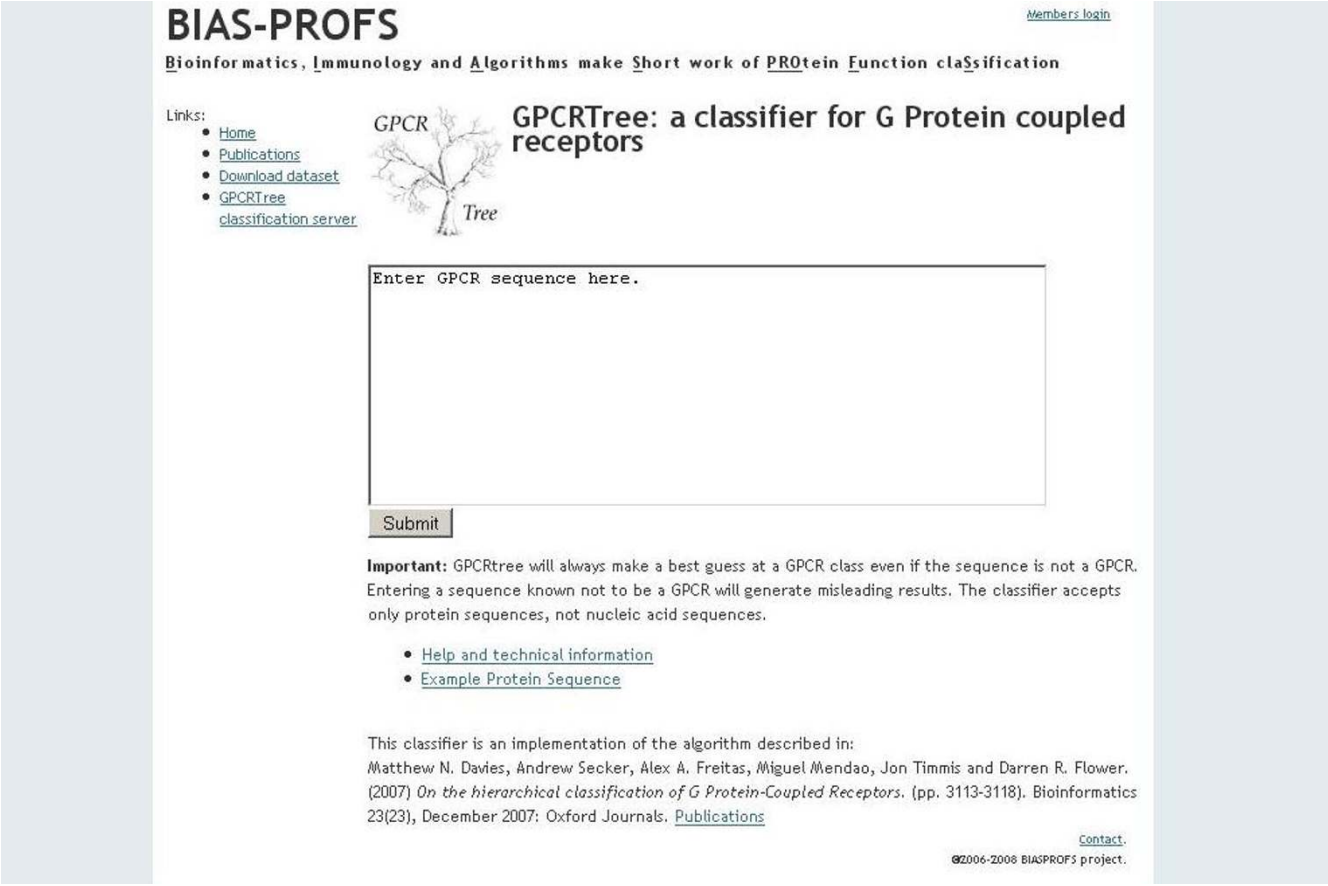

Figure 1
Input page for GPCRTree server.

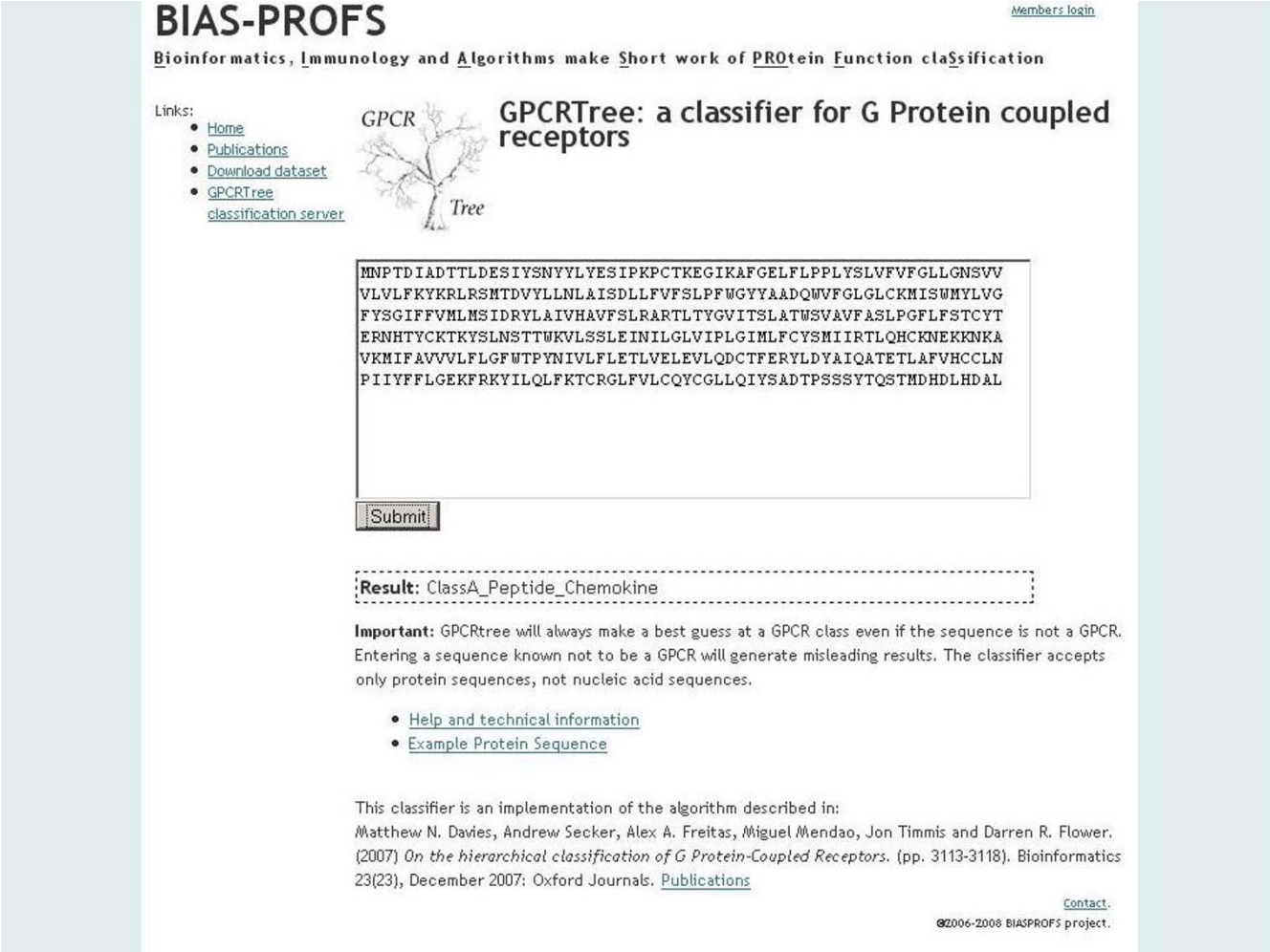

Figure 2
Results page for the GPCRTree server showing the prediction for the sequence of Chemokine CCR4 receptor.

Programming language: PHP, dHTML, Java

Other requirements: None

License: None

Any restrictions to use by non-academics: None

Abbreviations

GPCR: G protein coupled receptor; TCP: Tranmission Control Protocol; WEKA: Waikato Environment for Knowledge Analysis; SVM: Support Vector Machine

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MD Built GPCRtree datatset, created alignment-free representations of protein sequences, and wrote the paper. AS Designed and implemented the selective top-down method for hierarchical classification. Implemented method of turning raw protein sequences into numerical attributes. Assisted in writing the paper and implementation of the code on the GPCRTree web server. MHB Constructed and implemented GPCRTree server, currently maintains server at Birkbeck College, University of London DM Supervised construction of GPCRTree server at Birkbeck College AF Supervised the design of the selective top-down method for hierarchical classification. JT Supervised a mathematical analysis of data mining algorithms for hierarchical classification EC Performed a mathematical analysis of data mining algorithms for hierarchical classification. DRF Supervised design and construction of GPCRtree dataset, development of representations of pro-

tein sequences, and co-wrote the paper. All authors read and approved the paper.

Acknowledgements

The authors should like to gratefully acknowledge funding under the ESPRC grant EP/D501377/1 and the European Union ImmunoGrid project FP6-2004-IST-4 (contract no. 028069).

References

1. Christopoulos A, Kenakin T: **G protein-coupled receptor allosterism and complexing.** *Pharmacol Rev* 2002, **54:**323-374.
2. Gether U, Asmar F, Meinild AK, Rasmussen SG: **Structural basis for activation of G-protein-coupled receptors.** *Pharmacol Toxicol* 2002, **91:**304-312.
3. Bissantz C: **Conformational changes of G protein-coupled receptors during their activation by agonist binding.** *J Recept Signal Transduct Res* 2003, **23:**123-153.
4. Flower DR: **Modelling G-protein-coupled receptors for drug design.** *Biochim Biophys Acta* 1999, **1422:**207-234.
5. Klabunde T, Hessler G: **Drug design strategies for targeting G-protein-coupled receptors.** *ChemBioChem* 2002, **3:**928-944.
6. Milligan G: **G-protein-coupled receptor heterodimers: pharmacology, function and relevance to drug discovery.** *Drug Discov Today* 2006, **11:**541-549.
7. Davies MN, Secker A, Freitas AA, Mendao M, Timmis J, Flower DR: **Proteomic applications of automated GPCR classification.** *Proteomics* 2007, **7:**2800-14.
8. Horn F, Bettler E, Oliveira L, Campagne F, Cohen FE, Vriend G: **GPCRDB information system for G protein-coupled receptors.** *Nucleic Acids Res* 2003, **31:**294-7.
9. Attwood TK: **A compendium of specific motifs for diagnosing GPCR subtypes.** *Trends Pharmacol Sci* 2001, **22(4):**162-165.
10. Flower DR, Attwood TK: **Integrative bioinformatics for functional genome annotation: trawling for G protein-coupled receptors.** *Semin Cell Dev Biol* 2004, **15:**693-701.
11. Wistrand M, Kall L, Sonnhammer EL: **A general model of G protein-coupled receptor sequences and its application to detect remote homologs.** *Protein Sci* 2006, **15:**509-21.
12. Sgoruakis NG, Bagos PG, Papasaikas PK, Hamodrakas SJ: **A method for GPCRs coupling specificity to G-proteins using refined profile Hidden Markov Models.** *BMC Bioinformatics* 2006, **6:**104.
13. Karchin R, Karplus K, Haussler D: **Classifying G-protein coupled receptors with support vector machines.** *Bioinformatics* 2002, **18:**147-159.
14. Papasaikas PK, Bagos PG, Litou ZI, Promponas VJ, Hamodrakas SJ: **PRED-GPCR: GPCR recognition and family classification server.** *Nucleic Acids Res* 2004, **32:**W380-382.
15. Guo YZ, Li ML, Wang KL, Wen ZN, Lu MC, Liu LX, Lin J: **Fast fourier transform-based support vector machine for prediction of G-protein coupled receptor subfamilies.** *Acta Biochim Biophys Sin (Shanghai)* 2005, **37:**759-66.
16. Bhasin M, Raghava GP: **GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors.** *Nucleic Acids Res* 2004, **32:**W383-9.
17. Bhasin M, Raghava GP: **GPCRsclass: a web tool for the classification of amine type of G protein-coupled receptors.** *Nucleic Acids Res* 2005, **33:**W143-7.
18. Yabuki Y, Muramatsu T, Hirokawa T, Mukai H, Suwa M: **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:**W148-53.
19. Vilo J, Kapushesky M, Kemmeren P, Sarkans U, Brazma A: **Expression Profiler.** In *The Analysis of Gene Expression Data: Methodsand Software* Edited by: Parmigiani G, Garret ES, Irizarry R, Zeger SL. Springer Verlag, New York; 2003.
20. Kim J, Moriyama EN, Warr CG, Clyne PJ, Carlson JR: **Identification of novel multi-transmembrane proteins from genomic databases using quasi-periodic structural properties.** *Bioinformatics* 2000, **16:**767-75.
21. Huang Y, Cai J, Li L, Yanda L: **Classifying G-protein coupled receptors with bagging classification tree.** *Computational Biology and Chemistry* 2004, **28:**275-280.
22. Davies MN, Gloriam DE, Secker A, Freitas AA, Mendao M, Timmis J, Flower DR: **On the hierarchical classification of G Protein Coupled Receptors.** *Bioinformatics* 2007, **23:**3113-3118.
23. Lapinsh M, Prusis P, Lundstedt T, Wikberg JE: **Proteochemometrics modeling of the interaction of amine G-protein coupled receptors with a diverse set of ligands.** *Mol Pharmacol* 2002, **61:**1465-75.
24. Freyhult E, Prusis P, Lapinsh M, Wikberg JE, Moulton V, Gustafsson MG: **Unbiased descriptor and parameter selection confirms the potential of proteochemometric modelling.** *BMC Bioinformatics* 2005, **6:**50.
25. Freitas AA, de Carvalho ACPLF: **A Tutorial on Hierarchical Classification with Applications in Bioinformatics.** In *Research and Trends in Data Mining Technologies and Applications* Edited by: Taniar D. Idea Group; 2007:175-208.
26. Costa EP, Lorena AC, Carvalho ACPLF, Freitas AA, Holden N: **Comparing several approaches for hierarchical classification of proteins with decision trees.** *Proc. of the 2007 Brazilian Symposium on Bioinformatics (BSB-2007)* .
27. Davies MN, Secker A, Freitas AA, Mendao M, Timmis J, Flower DR: **On the hierarchical classification of G protein-coupled receptors.** *Bioinformatics* 2007, **23:**3113-8.
28. Witten IH, Frank E: **Data Mining: Practical Machine Learning Tools and Techniques.** Morgan Kaufmann, San Francisco; 2005.
29. Brownlee J: **WEKA Classification Algorithms, Version 1.6.** [http://sourceforge.net/projects/wekaclassalgos].