W380–W382 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh431

PRED-GPCR: GPCR recognition and family classification server

P. K. Papasaikas, P. G. Bagos, Z. I. Litou, V. J. Promponas and S. J. Hamodrakas*

Department of Cell Biology and Biophysics, Faculty of Biology, University of Athens, Panepistimiopolis, Athens 157 01, Greece

Received February 15, 2004; Revised April 2, 2004; Accepted April 13, 2004

ABSTRACT

The vast cell-surface receptor family of G-protein coupled receptors (GPCRs) is the focus of both academic and pharmaceutical research due to their key role in cell physiology along with their amenability to drug intervention. As the data flow rate from the various genome and proteome projects continues to grow, so does the need for fast, automated and reliable screening for new members of the various GPCR families. PRED-GPCR is a free Internet service for GPCR recognition and classification at the family level. A submitted sequence or set of sequences, is queried against the PRED-GPCR library, housing 265 signature profile HMMs corresponding to 67 well-characterized GPCR families. Users query the server through a web interface and results are presented in HTML output format. The server returns all single-motif matches along with the combined results for the corresponding families. The service is available online since October 2003 at http://bioinformatics.biol.uoa.gr/PRED-GPCR.

OVERVIEW

G-protein coupled receptors (GPCRs) constitute a vast cell surface receptor family, with a heterogeneous functional profile (1). However, the task of automatic classification of novel GPCRs into one of the functional families, as defined by ligand specificity, is not always straightforward. Since no clear correlation between ligand specificity and sequence similarity can be assessed, database search methods based on pairwise similarity [e.g. BLAST (2)] are not always suitable for this task. Search against pattern/motif databases highly improves the diagnostic performance in automated classification. Instead of querying against complete sequences, informative key regions are selected (either by experts or automatically) and are deployed as ''baits'' in order to distinguish between families. Several methods are used to encode the information of such regions. These methods include regular expressions [PROSITE (3)], position-specific scoring matrices [BLOCKS (4)], frequency matrices [PRINTS (5)] and profile hidden Markov models (HMM) [Pfam (6)].

In a somewhat similar approach, PRED-GPCR exploits the descriptive power of profile HMMs (7,8) along with an exhaustive discrimination method to construct a library of highly selective GPCR family signatures.

The family classification system used for the PRED-GPCR library is based mainly on the TiPs pharmacological classification for receptors (9) and the GPCRDB information system (10). However, several families are denoted by two or more synonyms in order to anthologize family descriptions alternatively used by TiPs, the GPCRDB, the Swiss-Prot (11) database and existing literature.

The PRED-GPCR library includes signatures for all well characterized GPCR families with an adequate number of member sequences to sustain the building of credible and descriptive profile HMMs.

Signatures for newly characterized GPCR families are regularly added, and complete re-training of the system is scheduled on a yearly basis (Figure 1).

The goal of the PRED-GPCR prediction system is to provide a complement to the existing pattern database analysis servers and potentially a computational tool for genome wide identification and family classification of GPCRs.

METHOD

The PRED-GPCR library is built using a naive sequential feature selection method. For all available families, multiple alignments are built, high entropy alignment stretches are excluded, and the remaining blocks are further segmented into confined overlapping fragments. The profile HMMs constructed from these fragments form a set of potential family-describing features. Consequently, for the sake of fine-tuning, all profiles are queried against a set of transmembrane receptors, including the training examples. Those profiles

*To whom correspondence should be addressed. Tel: +30 210 727 4545 or +30 210 727 4931; Fax: +30 210 727 4742; Email: shamodr@cc.uoa.gr

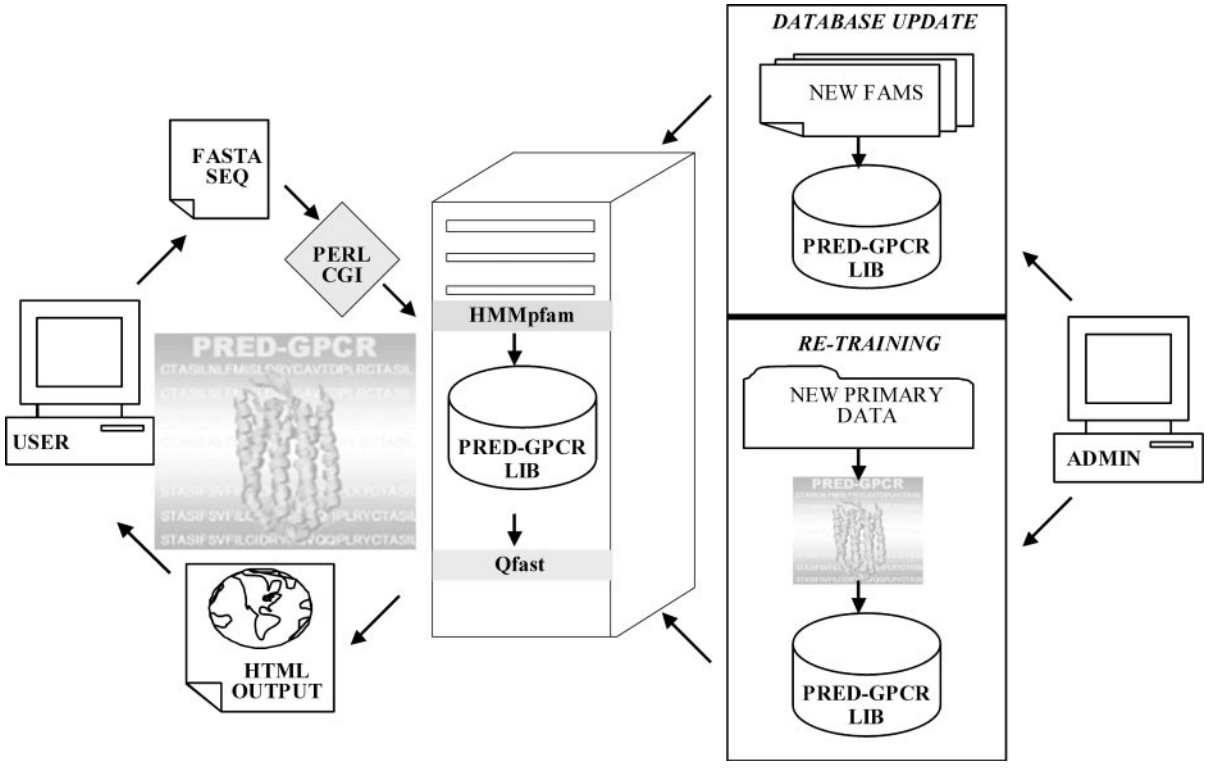


Figure 1. Data Flow in the PRED-GPCR server. Users submit their query sequence through a web interface. Results are presented in HTML. The PRED-GPCR library is regularly updated while complete retraining of the system is scheduled on a yearly basis.

that minimize an empirical error function are finally selected as family signatures. A detailed description of this method has been published elsewhere (12).

PERFORMANCE OF THE METHOD

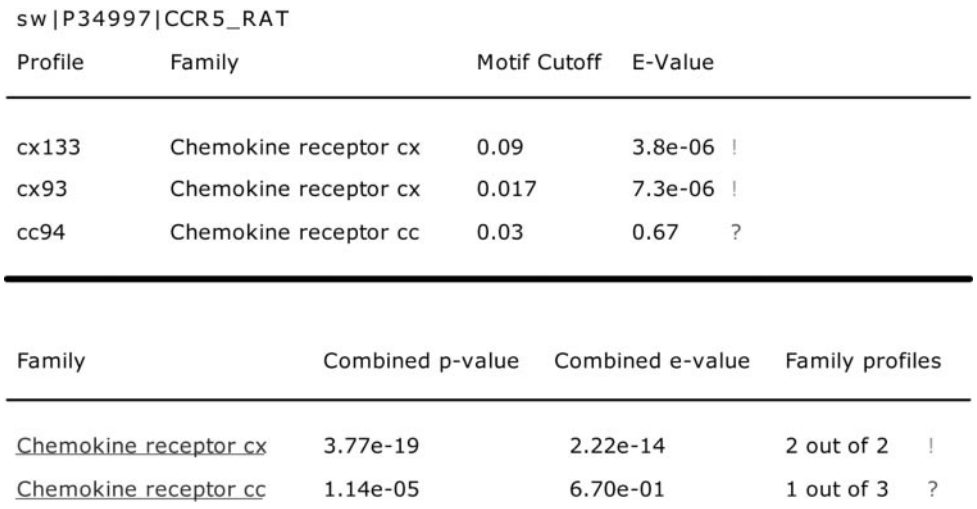
The specificity of the PRED-GPCR signature library for GPCRs was tested on a set of 1239 globular and 1361 non-GPCR transmembrane proteins with less than 25% pairwise similarity, unseen during all steps of training and optimization. Our method misclassified only 0.4% of these negative examples. The family combined *E*-value threshold was set to 0.03, which was the Minimum Error Point (MEP) for the transmembrane receptors data set. MEP is the *E*-value threshold where a classifier makes the fewest errors (false positives plus false negatives).

To demonstrate the efficacy of the method we have applied it to an independent set of 310 well-annotated sequences of GPCRs recently deposited in the Swiss-Prot database, excluding fragments. These sequences were not included in the primary data set. Application of our method with the MEP family combined *E*-value threshold correctly assigned 96% of these sequences to a GPCR family.

Additionally, the Swiss-Prot (11) and Trembl (11) databases are regularly scanned with the PRED-GPCR system for sequences with significant matches to the existing motifs. All these results are available on the PRED-GPCR web server from the 'Taxonomy' page. Currently (April 2004), 2620 sequences from Swiss-Prot and Trembl have been assigned to PRED-GPCR families.

INPUT OPTIONS—OUTPUT FORMAT

Sequences may be submitted either in FASTA format (default) or as plain text. Available input options include filtering by two different thresholds; the first one applies to family combined *E*-values and is user-defined. The second one filters individual motifs either by a global user-defined threshold or by preset, motif-specific empirical cutoffs. Additionally, the user can optionally pre-process the input sequences by implementing the CAST algorithm (13), which allows low-complexity region detection and selective masking. The submitted sequences are scanned against the PRED-GPCR library using the hmmpfam program from the HMMER software package (http://hmmer.wustl.edu). Statistically valid combined *P*-values for all matches derived from the same family are obtained by implementing the Qfast algorithm (14) (Figure 1). Results are returned in an HTML output format. The output report consists of a separate record for each input sequence. Each record comprises two sections: (i) a ranked list of all profile HMMs matching to the query sequence below the selected *E*-value threshold along with their corresponding family and their motif specific empirical cutoffs, and (ii) a ranked list of the combined *P*-values, *E*-values and the number of profiles matched for each family (Figure 2). Trusted results for the first section are assumed those single motif matches with an *E*-value below the individual motif-specific cutoff. For the second section, we consider as significant those combined *E*-values below a corrected MEP (see Performance of the Method), weighted according to the current population of the motif database. These results are distinctly indicated in the results page (Figure 2). Users can



sw|P34997|CCR5_RAT

Profile	Family	Motif Cutoff	E-Value	
cx133	Chemokine receptor cx	0.09	3.8e-06	!
cx93	Chemokine receptor cx	0.017	7.3e-06	!
cc94	Chemokine receptor cc	0.03	0.67	?

Family	Combined p-value	Combined e-value	Family profiles	
Chemokine receptor cx	3.77e-19	2.22e-14	2 out of 2	!
Chemokine receptor cc	1.14e-05	6.70e-01	1 out of 3	?

Figure 2. Search report for a query sequence of the Chemokin receptor cx family. For each query sequence the report consists of two sections: (i) a ranked list of all profile HMM matches below the selected *E*-value threshold, along with their corresponding family and their motif specific empirical cutoffs, and (ii) a ranked list of the combined *P*-values, *E*-values and the number of profiles matched for each family.

cross-evaluate their results by browsing relevant entries from Swiss-Prot, Trembl, Pfam and Prosite databases.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their constructive citicism and valuable comments.

REFERENCES

1. Pierce,K.L., Premont,R.T. and Lefkowitz,R.J. (2002) Seven-transmembrane receptors. *Nature Rev. Mol. Cell Biol.*, **9**, 639–650.
2. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
3. Falquet,L., Pagni,M., Bucher,P., Hulo,N., Sigrist,C.J., Hofmann,K. and Bairoch,A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.*, **30**, 235–238.
4. Henikoff,J.G., Greene,E.A., Pietrokovski,S. and Henikoff,S. (2000) Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res.*, **28**, 228–230.
5. Attwood,T.K., Croning,M.D. and Gaulton,A. (2002) Deriving structural and functional insights from a ligand-based hierarchical classification of G protein-coupled receptors. *Protein Eng.*, **15**, 7–12.
6. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, 138–141.
7. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
8. Hughey,R. and Krogh,A. (1996) Hidden Markov models for sequence analysis: extension and analysis of the basic method. *Comput. Appl. Biosci.*, **12**, 95–107.
9. Alexander,S.P.H. and Peters,J.A. (2000) *TiPs Receptor and Ion Channel Nomenclature Supplement*, Elsevier, Vol. XI.
10. Horn,F., Bettler,E., Oliveira,L., Campagne,F., Cohen,F.E. and Vriend,G. (2003) GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res.*, **31**, 294–297.
11. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S. and Schneider,M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
12. Papasaikas,P.K., Bagos,P.G., Litou,Z.I. and Hamodrakas,S.J. (2003) A novel method for GPCR recognition and family classification from sequence alone using signatures derived from profile hidden Markov models. *SAR QSAR Environ. Res.*, **14**, 413–420.
13. Promponas,V.J., Enright,A.J., Tsoka,S., Kreil,D.P., Leroy,C., Hamodrakas,S., Sander,C. and Ouzounis,C.A. (2000) CAST: an iterative algorithm for the complexity analysis of sequence tracts. Complexity analysis of sequence tracts. *Bioinformatics.*, **16**, 915–922.
14. Bailey,T.L. and Gribskov,M. (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics.*, **14**, 48–54.