Development of a Hierarchical Protein Classification Tool

by

Zhifang Wang

Presented to the Faculty of

The Graduate College at University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Master of Science

Major: Computer Science

Under the Supervision of Professor Jitender Deogun and Professor Etsuko Moriyama

Lincoln, Nebraska

Dec 2002

Development of a Hierarchical Protein Classification Tool

Zhifang Wang, M.S.

Department of Computer Science and Engineering

University of Nebraska-Lincoln, 2002

Advisor: Dr. Jitender Deogun and Dr. Etsuko Moriyama

ABSTRACT:

Computer aided protein classification is useful in mining new classes of proteins and new members of protein families from genome databases. Traditional methods *e.g.*, PFAM, PROSITE, and PRINTS, primarily rely on sequence similarity. These methods are used in protein sequence annotation and proved to be very useful. However, they are not the best options when dealing with extremely divergent protein families where sufficient sequence similarity is often not present. Recently Kim et al. (2000) developed a classification method based on discriminant analysis (DA) which that utilizes amino acid composition and physico-chemical properties of proteins. Instead of simply relying on sequence similarity, this method incorporates structural features that might be more functionally relevant. In this project we try to design a protein classification tool that incorporates the aforementioned four methods. We designed a hierarchical classification scheme that combines the strengths of the four methods. G-protein Coupled Receptor protein family is chosen as an example to examine the usefulness of this scheme of protein classification. The tool can be either used to improve the discriminant analysis method or identify new protein family members that are missed by traditional classification methods.

ACKNOWLEDGEMENTS

I want to take this opportunity to thank everyone who encouraged and helped me to finish this project. Firstly I would like to thank Dr. Jitender S. Deogun for his guidance and support throughout the course of my studying in UNL. I also like to express my sincere appreciation to Dr. Etsuko Moriyama for providing me with this nice project topic, an excellent computing environment, and tireless support and guidance during project. Her encouragement and help is very critical to the success of this project. And I also want to thank Dr. Hong Jiang and Dr. Stephen Scott for being on my committee.

Thanks Dr. Guoqin Lu for comments on my project and reports and Dr. Ruben Donis for leading me into the Bioinformatics field.

A special thank you to Dr. Allen Tsang who has provided me with a lot of valuable discussions throughout my project development. His encouragement has constantly helped me channel my potentials in the right direction. I am really grateful for that.

I also thank Yavuz F. Yavuz, the system administrator of Moriyama Bioinformatics Lab, for his wonderful support and encouragement. And thanks all the members of the Moriyama Bioinformatics Lab who helped me and made this project an enjoyable experience.

Many thanks go to all of the faculty and staff for their help during my study at the Department of Computer Science and Engineering.

Finally I wish to give deepest appreciation to my family and friends for always being here for me.

1. Introduction

Protein classification is one of the fundamental tasks in the post-genomic biological science. Its goal is to classify the proteins from large genome database into families. Within a family, member proteins share some common signatures that might represent common functional or structural properties for the family. Therefore, by means of classification, we can predict functions of new proteins obtained from genomic databases. Automated protein classification is becoming imperative because of the increasingly large amount of sequence data accumulated by multiple genome projects. This is why a great deal of efforts has been made in development of effective protein classification methods and software packages. Currently majority of the methods used in the protein classification (or annotations) relies on primary sequence similarities. Due to inefficacy in identifying extremely divergent protein family members by these methods, new attempts have been made to utilize structural features and other global features from protein sequences [1].

In this project, we designed a protein classification system that combines the strengths of both sequence-similarity based methods and secondary-structure based methods. For sequence-similarity based methods, we chose PFAM, PRINTS, and PROSITE. The secondary-structure based method used in this project was the method based on discriminant analyses (DA method, described in the next section). We first designed a MySQL based database which integrates the classification results of the four methods. Then a hierarchical classification scheme was designed. We chose the G-protein coupled receptor protein family as an example to examine the potentials of this database and the hierarchical classification scheme.

G-protein coupled receptor (GPCR) is a membrane protein. Membrane proteins play an important role in the physiological function of the organism. They facilitate transporting molecules and

signals between the cell and its environment. Many transmembrane proteins transport specific metabolites, drugs and ions. GPCRs act as receptors for a wide variety of signaling molecules including hormones, nucleotides, opiates, neurotransmitters, and odorants. GPCRs are important targets of pharmacological agents. Identifying and classifying them has a significant meaning in both basic and applied (pharmaco-medical) sciences [1].

This report first provides some background in Section 2. Section 3 describes the databases and programs used in this project. The overall design and implementation is described in Section 4. Section 5 shows the design of the MyGPCR database. And Section 6 describes the web interface. Section 7 presents the conclusion and the future work needed to be done.

2. Background and Related Work

This section introduces the background knowledge necessary to understand and carry on the project. Different protein classification methods are reviewed first. The G-protein couple receptor is described next. Databases and software packages used in our project are introduced in the next section (Section 3).

2.1 Protein Classification Methods

The rapid accumulation of sequence data requires us to develop computational methods for protein sequence annotation. One of such attempts is to organize proteins into families based on the presence of common motifs in the sequences. Motifs are originally derived from a set of known members of a protein family and are some conserved regions or patterns in the sequences. Motifs are considered to represent regions of structural or functional significance. Therefore they can be used as classifiers to find more members of the protein families [9].

Many methods are currently employed to identify members of protein families. The most basic method is to use some form of sequence comparison algorithm to assign family membership. Other more advanced methods generally start with a multiple sequence alignment of a set of known members of a family and use three approaches to characterize conserved regions in the alignment. Regular expression patterns and pattern groups, position specific matrix, profile Hidden Markov Models (HMMs). These classes of methods are well regarded in the fields and are widely used in current protein annotations. But they are limited when little sequence similarity are present. New classes of protein classification methods do not reply on multiple alignments; rather they try to use statistics or mathematics theory, combined with some structural information in protein classification. Examples of this class of methods include neural network [14], discriminant analysis [1], and support vector machines [10]. This class of methods is still in development/experiment stage so their performances are still not well evaluated. Below are short introductions to some methods that are of my interest.

- Sequence Similarity

  The most basic method for protein classification is to use some form of sequence comparison algorithm to assign family membership. BLAST (Basic Local Alignment Search Tool) is a set of similarity search programs that allow a query sequence to be compared with all the sequences in the selected database (SWISS-PROT, NR etc). The significance of a match is indicated by a score. Most closely matched sequences are listed first. BLAST search gives a hint that which sequences are similar to the sequence of the interest. So it is usually the first step in finding the function of a new or unknown sequence. The BLAST programs have been designed for speed, with a minimal sacrifice of sensitivity to distant sequence relationships. It uses a heuristic algorithm which seeks local as opposed to global alignments and is therefore

able to detect relationships among sequences which share only isolated regions of similarity [13].

- **Regular Expression Patterns**

  A pattern of highly conserved residues (presumably identifying an active site or other important structural feature) can be represented as a regular expression that characterizes the family. For example, the following is from a PROSITE pattern entry: F-x(5)-G-x(2,4)-G-H (see the next section for an introduction on the PROSITE pattern database). It defines a pattern which starts with an amino acid F, followed by any 5 amino acids, then G, followed by any 2 to 4 free character (amino acids), and ends with G and H. Another PROSITE pattern example is: {DERK}(6)-[LIVMFWSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C. This pattern consists of 6 amino acids of any type except D, E, R, and K, followed by 2 amino acids chosen from L, I, V, M, F, W, S, T, A, or G, then one amino acid chosen from L, I, V, M, F, Y, S, T, A, G, C, or Q, followed by any one of A, G, or S, and ending with a C. PROSITE has historically been used as the starting point or seed for families created by other methods. Another resource that uses regular expressions is the e-MOTIF collection (web-address for the e-MOTIF). In e-MOTIF, families may be split into subfamilies, each with a pattern highly specific to it.

- **Position Specific Scoring Matrix (PSSM)**

  The rigidity of a regular expression pattern can be limiting in representing divergent protein motifs families. First of all short patterns are insufficient for today's databases including many partial sequences (*e.g.*, ESTs). Secondly requirements of perfect matches produce many false negatives. In order to solve these problems, some methods assign a score for each amino acid at every position in the multiple alignment (along with specific gap penalties assigned for insertions at each position). PRINTS (described in the next section) is the simplest of these

approaches. Families in PRINTS are characterized by the presence of several ordered, ungapped "motifs" that together represent a fingerprint of the family. PRINTS motifs are simply a set of amino-acid frequency tables that are derived from short conserved regions. Members of a PRINTS family may have 1 or more missed motifs within the fingerprint. Hence, PRINTS has the added benefit that subfamily relationships can be discerned (as evidenced by the patterns of unmatched motifs). Blocks+ (http://www.blocks.fhcrc.org/) is another well-curated resource that features a Position-Specific Scoring Matrix, PSSM. Blocks+ collects seed sequences from numerous sources (e.g., PROSITE, PRINTS, Pfam, ProDom, DOMO), and builds a set of ungapped string of blocks (PSSMs) for each family. PSI-BLAST – Position Specific Iterated BLAST is another interesting example that uses PSSMs. PSI-BLAST uses an iterated search in which sequences found in one found of searching are used to build the score matrix for the next round of search. A subset of PROSITE families uses gapped PSSMs as well. PSSMs are sometimes referred as profiles in some literatures as well. Compared to profile hidden Markov model (described next), PSSMs are much simpler.

- **Profile Hidden Markov Models (HMMs)**

  Hidden Markov Model (HMM) is a class of probabilistic models that are generally applicable to time series or linear sequences. HMMs have been most widely applied to recognizing words in digitized sequences of the acoustics of human speech. HMMs were introduced into computational biology in late 1980's, and for use as profile models in 1994. Now HMMs have become a very popular profile model. The Pfam protein family database (described in the next section) and some PROSITE profiles were created using this method. Profile HMMs assign statistical weights, rather than scores, to each position in the profile, reflecting the probability of finding any of the 20 amino acids, an insertion, or deletion.

- **Discriminant Analysis (DA) Method**

  Discriminant analysis is a very useful statistical tool. It takes into account the different variables of an object and works out which group the object most likely belongs to. Discriminant analysis was used by Kim *et al* [1] to develop a new protein classification method to discriminates GPCR sequences from non-GPCR sequences. It uses concise statistical variables based on physico-chemical properties of protein sequences. The selected statistical variable set includes amino-acid composition and periodicity statistics based on hydrophobicity, and polarity.

- **Support Vector Machine**

  Support Vector Machines (SVM) is a statistical learning algorithm that is popular in machine learning community and pattern recognitions. Similar to DA method, a learning machine is first trained to distinguish between two categories from a series of labeled examples and is then used to predict the class membership of previously unseen examples. A SVM makes use of a mathematical tool called kernel function to measure the similarity between two examples. SVM is used to construct a GPCR subfamily classifier by a group of researchers in University of California in Santa Cruz [8]. This classifier makes use of only sequence information and is reported not very effective in GPCR super family classification.

Various Protein and Protein Family Databases

The following table lists some common protein databases, protein pattern databases and integrated tools that are relevant to the subject.

Sequence Databases	SWISS-PROT	A highly curated protein sequence database with a high level of annotations. Integrated nucleic acid sequences, protein sequences, and protein tertiary structures, as well as specialized data collections. http://www.expasy.org/sprot/
	TrEMBL	A computer-annotated supplement of SWISS-PROT that contains all the translations of EMBL nucleotide sequence entries not yet integrated in SWISS-PROT. http://www.expasy.org/sprot/
	PIR	Protein Information Resource. Another resource of protein sequences with annotations. http://www-nbrf.georgetown.edu/
Databases of Patterns and Motifs	PROSITE patterns	Regular expression based patterns. http://www.expasy.ch/PROSITE/
	PROSITE profiles	Database of protein families and domains. http://www.expasy.ch/PROSITE/
	Pfam	Database of HMMs for domains and families. http://pfam.wustl.edu/index.html

	SMART	Simple Modular Architecture Research Tool. Relies on hand-curated multiple sequence alignments of representative family members from PSI-BLAST to builds HMMs, which is used to search database for more sequences for alignment until no more homologues detected. http://smart.embl-heidelberg.de/
	PRINTS	Database of protein fingerprints made of one or more ungapped conserved regions. Built by iterative scanning of OWL database. http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/
	TIGRFAMs	Collection of protein families in HMMs built with curated multiple sequence alignments and with associated functional information. http://www.tigr.org/TIGRFAMs
	BLOCKS	Collection of multiply aligned ungapped segments corresponding to most highly conserved regions of proteins– represented in profile. PROSITE, PRINTS, Pfam, ProDom and Domo are used in building the database. http://www.blocks.fhcrc.org/
	ProDom	Groups all sequences in SWISS-PROT and TrEMBL into domains. http://prodes.toulouse.inra.fr/prodom/doc/prodom.html

	PIR-ALN	Database of annotated protein sequence alignments derived automatically from PIR PSD. Includes alignments at superfamily (whole sequence), family (45% identity) and domain (in more than one superfamily) levels. http://www-nbrf.georgetown.edu/pirwww/ search/textpiraln.html
	ProtoMap	Automatic classification of all SWISS-PROT proteins into groups of related proteins (also including TrEMBL now). http://www.protomap.cs.huji.ac.il
	Domo	Database of gapped multiple sequence alignments from SWISS-PROT and PIR. Each entry is one homologous domain. Provides annotation on related proteins, functional families, evolutionary tree etc. http://www.infobiogen.fr/srs6bin/cgi-bin/wgetz?-page+LibInfo+-lib+DOMO
	ProClass	Non-redundant protein database organized by family relationships defined by PROSITE patterns and PIR superfamilies. Facilitates protein family information retrieval, domain and family relationships, and classifies multi-domain proteins. http://pir.georgetown.edu/gfserver/proclass.html

Integrated Databazses	MetaFam	Protein family classification built with Blocks+, DOMO, Pfam, PIR-ALN, PRINTS, PROSITE, ProDom, SBASE, SYSTERS. Automatically create supersets of overlapping families using set-theory to compare databases- reference domains covering total area. Use non-redundant protein set from SWISS-PROT, TrEMBL and PIR. http://metafam.ahc.umn.edu/
	IProClass	Integrated database linking ProClass, PIR-ALN, PROSITE, Pfam and Blocks. Use non-redundant proteins from SWISS-PROT and PIR. http://pir.georgetown.edu/iproclass/
	CDD	Conserved Domain database. Database of domains derived from SMART, Pfam and contributions from NCBI. Uses reverse position-specific BLAST (matrix). Links to proteins in Entrez and 3D structure. http://www.ncbi.nlm.nih.gov:80/Structure/cdd/cdd.shtml.
	InterPro	Built from PROSITE, PRINTS, Pfam, ProDom, SMART, SWISS-PROT and TrEMBL. http://www.ebi.ac.uk/interpro

Table 1. List of Protein Databases and Protein Pattern Databases.

2.2 G-Protein Coupled Receptors (GPCRs)

G-protein coupled receptors (GPCRs), or seven-transmembrane receptors (7tm), are membrane proteins. They all have a structure that is characterized by seven regions that cross cell membrane, as illustrated by Figure 1. Members of this family include receptors for many hormones, neurotransmitters, chemokines, and calcium ions, as well as sensory receptors for various odorants, bitter and sweet taste, and even photons of light. GPCRs regulate many physiological processes and are the most common target of therapeutic drugs. Therefore, identifying GPCRs has significant meaning in both basic and applied science [1, 2].

GPCRs can be grouped into three major families, A, B and C, on the basis of sequence similarity. Sequences within each family generally share over 25% of sequence identity in the transmembrane region, and a distinctive set of highly conserved residues and motifs. Among the three families, little similarity is evident beyond the predicted 7TM architecture. Family A is by far the largest group, and includes the receptors for, for example, light (rhodopsin), adrenaline (adrenergic receptors), and, indeed, most other 7TM receptor types, including the olfactory receptor subgroup. Nearly 200 7TM receptors that recognize over 80 distinct ligands have been functionally characterized. Family B is very small and includes the receptors for the gastrointestinal peptide hormone family (secretin, glucagon, vasoactive intestinal peptide, and growth-hormone-releasing hormone), corticotropin-releasing hormone, calcitonin and parathyroid hormone. All family B receptors seem to couple mainly to activation of the effector adenylyl cyclase through the G-protein Gs. Family C is also relatively small, and contains the metabotropic glutamate receptor family, the GABAB receptor, and the calcium-sensing receptor, as well as some taste receptors. All family C members have a very large extracellular amino-terminus that seems to be crucial for ligand-binding and activation. Other minor families are Class D Fungal pheromone and Class E cAMP receptors

(Dictyostelium) Frizzled/Smoothened family. GPCRDB (http://www.gpcr.org) is a GPCR information system that collects GPCR related information. A brief introduction on GPCRDB can be found in the next section.

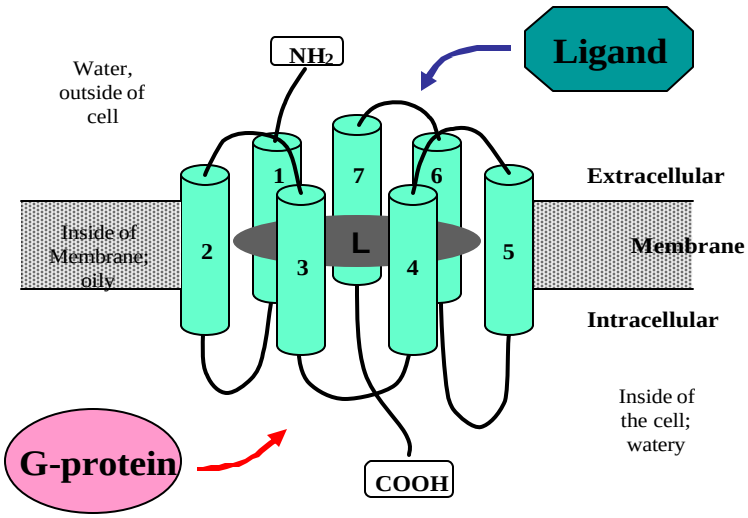

Figure 1. A model for the transmembrane regions in a G-protein coupled receptor (adapted from http://bioinfolab.unl.edu/emlab/research.html).

3. Databases and Software Packages used in the Project

3.1 SWISS-PROT

SWISS-PROT is a curated protein sequence database that provides a high level of annotations, such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc. It was established in 1986 and maintained collaboratively, since 1987, by the Department of Medical Biochemistry of the University of Geneva (and now the Swiss Institute of Bioinformatics, SIB) and the EMBL Data Library (now the European Bioinformatics

Institute, EBI). In this project we try to classify GPCRs from SWISS-PROT. Web based SWISS-PROT can be found at http://www.expasy.ch/sprot/. A flat File SWISS-PROT database is available to download from ftp://us.expasy.org/databases/swiss-prot/. SWISS-PROT Release 40 (downloaded in June 2002) used in this project has 101,602 entries. The number of entries is continually increasing. For example, SWISS-PROT Release 40.26 of 13-Aug-2002 has 112,892 entries. An example of SWISS-PROT entries is found in the Appendix 1. Each entry is composed of different line types, each with their own format. Lines begin with a two-character line code, which indicates the type of data contained in the line. The current line types and line codes and the order in which they appear in an entry, are listed in the Appendix 2. Since accession numbers are permanent, (as IDs can be changed), whenever possible, we used accession numbers as a preferred way to refer a particular SWISS-PROT entry.

3.2 GPCRDB

The GPCRDB is a G protein-coupled receptor database system aimed at the collection and dissemination of GPCR related data. It holds sequences, mutant data, and ligand binding constants as primary (experimental) data. In addition it also holds computationally derived data such as multiple sequence alignments, three-dimensional models, phylogenetic trees, and two-dimensional visualization tools. GPCRDB sequence data are imported from the SWISS-PROT database. Thus the format of GPCRDB is the same as SWISS-PROT. The GPCRDB is available via the WWW at http://www.gpcr.org/7tm [3].

In this project, the GPCRDB (March 2002 release) was used as a reference to obtain GPCR sequences for the DA method training data and GPCR specific entries from PFAM profiles, PRINTS fingerprints, and PROSITE patterns/rules/profiles. We used the list of SWISS-PROT and TrEMBL entries at http://www.gpcr.org/7tm/htmls/entries.html for sampling GPCR training data.

3.3 PFAM and HMMER

Pfam is a database of protein families. Each protein family is represented by a multiple alignment of a group of proteins that share some common protein domains or conserved protein regions. The alignments are constructed semi-automatically using hidden Markov models (see below). The alignments are considered to represent some evolutionary conserved structure that has implications for the protein function. Profile hidden Markov models (profile HMMs) built from the Pfam alignments can be very useful for automatically recognizing a new protein belonging to an existing protein family, even if the similarity is weak. Unlike standard pairwise alignment methods (e.g. BLAST and FASTA), Pfam HMMs deal sensibly with multi-domain proteins.

Pfam is constructed in two separate parts. Pfam-A includes accurate human-crafted multiple alignments, whereas the smaller Pfam-B contains protein families non-overlapped with Pfam-A and derived from the PRODOM database (an automatically compiled protein domain database, derived from SWISS-PROT and TrEMBL). Pfam-A families have permanent accession numbers and contain functional annotations and cross-references to other databases, while Pfam-B families are re-generated at each Pfam release and are un-annotated. Release 7.4 (date), the version used in our project, contains a total of 3882 families Version 7.5 of Pfam (August 2002) contains alignments and models for 4176 protein families, based on the SWISS-PROT 40 and SP-TrEMBL 18 protein sequence databases.

HMMER is an implementation of profile HMM methods for sensitive database searches using multiple sequence alignments as queries. A multiple sequence alignment is given as an input (termed as a seed alignment). A statistical model (hidden Markov model) is built based on the alignment and used as a query to find (and/or align) additional members of the sequence family from a database ("full alignment").

HMMER is most sensitive at identifying complete domains. Its preferred algorithm is a "profile alignment" algorithm that is neither fully global nor fully local alignment, but instead looks for a "glocal" alignment that is global with respect to the model, but (multiply) local with respect to the sequence -- e.g. to look for one or more complete domains in a query sequence. Fragments of domains do occur, especially in truncated sequences, translated ESTs, or because of insertions of new domains in existing ones. HMMER can also build models that do fully local alignment and tolerate fragments (indicated with the hmmbuild -f notation on the BM line of the Pfam annotation). A standard Pfam model does not tolerate fragments. Pfam therefore includes two different types of models, the "glocal" models ("ls" mode in the Pfam_ls HMM database) and Smith/Waterman models ("fs" mode in the Pfam_fs HMM database). In the "glocal" mode, only full-length complete domains are found. In Smith/Waterman mode, fragmentary domains can also be found, because fully local alignments are allowed. "ls" mode is much more sensitive than "fs" mode, but only if a complete domain is actually present; if a partially deleted fragment is present, "fs" mode will be needed [7].

Web based Pfam can be accessed from http://pfam.wustl.edu/. Pfam databases can be downloaded from ftp://ftp.genetics.wustl.edu/pub/Pfam/. HMMER package can be downloaded at http://hmmer.wustl.edu/. An example of HMM profile from Pfam_ls (the profile database for the "glocal" search model) is given in the Appendix 3.

3.4 PRINTS and FingerPRINTScan

PRINTS is a database of protein fingerprints. A "fingerprint" is a group of conserved regions called motifs by PRINTS. Fingerprints can encode protein folds and functionalities more flexibly and powerfully than can single motifs. FingerPRINTScan is a search tool for the PRINTS protein fingerprints database. It can accept either the main format of the database, *i.e.*, .dat file, or the

profile format of the PRINTS database. Profile format of the PRINTS database can be either generated using FingerPRINTScan or downloaded with FingerPRINTScan. Using the profile format with FingerPRINTScan speeds up the computation process. Web based PRINTS database services can be accessed from http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/. And FingerPRINTSCan can be downloaded for local installation at ftp://proline.sbc.man.ac.uk/pub/fingerPRINTScan/. FingerPRINTScan help can be accessed from http://www2.ebi.ac.uk/printscan/help.html and http://bioinfo.man.ac.uk/fingerPRINTScan/evalwarning.html. An example entry from PRINTS main database (.dat) that describes a fingerprint identified by "ANTISTASIN" can be found in the Appendix 5.

3.5 PROSITE and ps_scan

PROSITE is a database of protein families and domains [4]. It consists of biologically significant sites, patterns, rules, and profiles (position specific scoring matrix) that help to reliably identify to which known protein family (if any) a new sequence belongs. A profile is a table of position-specific amino acid weights and gap costs. These numbers (also referred to as scores) are used to calculate a similarity score for any alignment between a profile and a sequence, or parts of a profile and a sequence. An alignment with a similarity score higher than or equal to a given cut-off value constitutes a motif occurrence. A distinguishing feature between a pattern and a profile is that the former is usually confined to a small region with high sequence similarity whereas the latter attempts to characterize a protein family or domain over its entire length. Similar to SWISS-PROT, a PROSITE entry also comprises different types of lines indicated by two leading characters (see Appendix 6). Web based PROSITE can be accessed at http://us.expasy.org/PROSITE/. PROSITE database can be downloaded from ftp://us.expasy.org/databases/PROSITE/.

ps_scan is a perl program used to scan one or several patterns, rules, and/or profiles from PROSITE against one or several protein sequences in SWISS-PROT or FASTA format (it consists of a description line beginning with ">" and multiple lines of sequence data). It requires two compiled external programs, "pfscan" and "psa2msa", from PFTOOLS packages. ps_scan can be download at ftp://us.expasy.org/databases/PROSITE/tools/ps_scan/. Readme file for ps_scan.pl can be found at ftp://us.expasy.org/databases/PROSITE/tools/ps_scan/sources/README.

3.6 Discriminant Analysis (DA) Method

In our project we used these DA methods for the first stage classification. The selected statistical variable set includes amino-acid composition and periodicity statistics based on hydrophobicity, and polarity. Instead of the non-parametric linear function used in Kim *et al.* (2000), parametric linear discriminant function (LDA), quadratic discriminant function (QDA), and logistic discriminant function (LOG) were used to discriminate GPCR sequences from the database. We retrained the DA method.

4. Design and Implementation Overview

The objective of this project is to design a hierarchical protein classification tool which makes good use of the strengths of several traditions methods (PFAM, PRINTS and PROSITE) and newly developed discriminant function analysis. Traditional methods are known to be accurate but they reply on too much on sequence similarity thus are too conservative when dealing with divergent protein family such as GPCRs. And discriminant analysis (DA) are shown to be very helpful in identifying protein families where the sequence similarity if weak. But discriminant analysis tends to have a slightly higher false positives. Therefore, we can use DA as the first screening step to separate GPCRs from Non-GPCRs. Then we use traditional methods to classify the GPCRs

predicted by DA. The resulting data sets are then reorganized in several groups of dataset, for example, GPCRs by only one method, only two methods, and only three methods. By examining each dataset, we can find weakly supported GPCRs and strongly supported GPCRs by traditional methods. The weakly supported GPCRs are more interesting since it might suggest unidentified or misidentified GPCRs by the traditional methods. The results can be presented by the Figure 2.

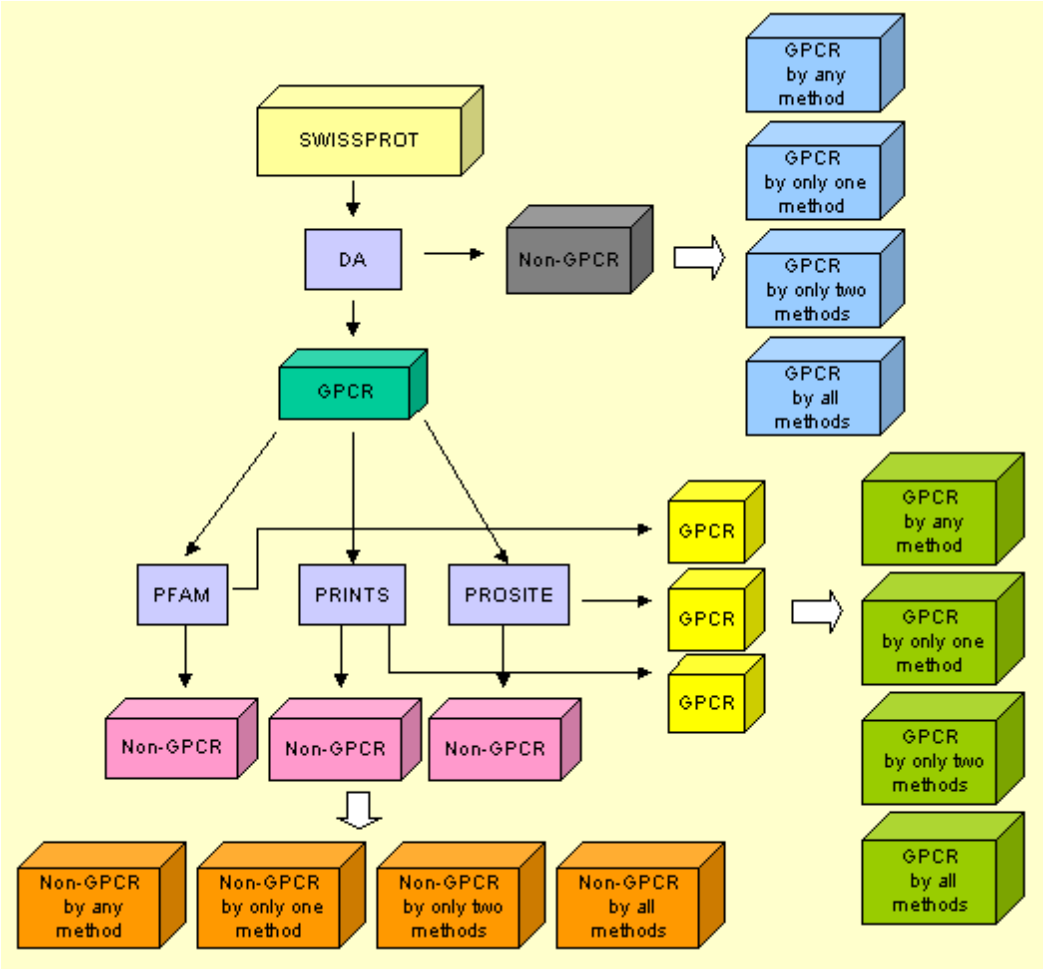

Figure 2. Hierarchical Protein Classification Scheme (the blue prints)

To realize this scheme, we design a MySQL based relational database to hold the raw classification results of SWISS-PROT by each method. Having a relational database has the following benefits:

- Easy manipulation of data using SQL language. Compared to Perl or C, SQL is much easier to learn and use.

- The database can be updated easily since it stores raw classification results by each method. Since SWISS-PROT. PFAM, PRINTS, and PROSITE databases update regularly, an easy update scheme is very important to keep our database and web site up to date.
- Allow users to use GUI interfaces, for example, ACCESS, MyCC, OpenOffice to view and query data, generate reports easily.
- The database can be easily integrated with web services to allow easier access of the data via WWW.
- The hierarchical classification scheme can be designed based on the information stored in the database. This has the added benefit that once the database is updated, the changes will propagate to the hierarchical classification scheme automatically.
- RDBMS has been around for quite long time and the technology is very mature. There are a lot of commercial or non commercial products available. Using RDBMS to manage the biological data would undoubtedly improve the biologists' productivity.

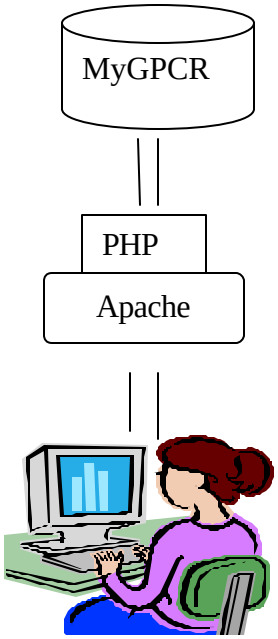

Figure 3. Configuration of Web Based Hierarchical Protein Classification Scheme.

The Figure 3. shows the configuration of the web based hierarchical classification scheme with the support of a MySQL database, MyGPCR. And Figure 4. shows the data flows in construction of MyGPCR database. According to the database flow, the project development can be broken into four important stages:

- Classifying proteins from SWISS-PROT using DA, PFAM, PRINTS, and PFAM, respectively. First the required databases (SWISS-PROT, PFAM, PROSITE, PRINTS, and GPCRDB) and software packages associated with each database need to be downloaded and installed locally. Next a training process is needed for DA methods to classify proteins. To this end, a new set of training data needs to be prepared first. For PFAM, PROSITE, and PRINTS, profiles/patters/fingerprints specific to GPCR sequences need to be extracted from GPCRDB and verified accordingly. These profiles/patters/fingerprints are then used in each method to classify the proteins.
- Parsing the classification results from the four methods, and constructing the MyGPCR database (MySQL based protein database).
- Constructing web interface and organizing the classification results in the hierarchical way.

The following sections will provide more details on the first step. The classifications by the four methods are described in detail in section 5. The construction of the MyGPCR database and the web interface are detailed in Section 6 and Section 7 respectively.

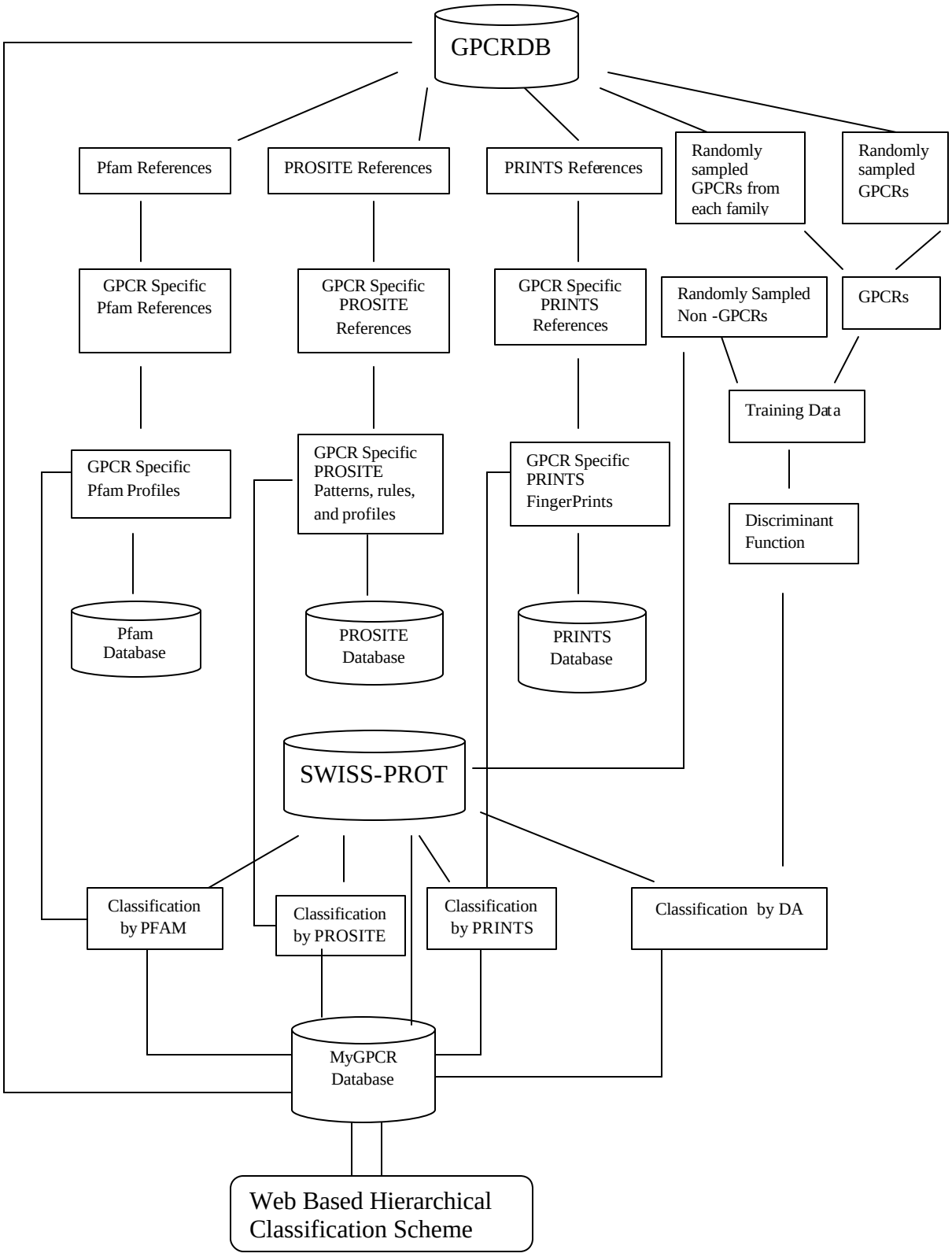

Figure 4. Data Flow.

5. Classifying Sequence Database Using Four Methods

5.1 Obtaining GPCR Specific Pfam Profiles, PRINTS Fingerprints, and PROSITE Patterns/Rules/Profiles

GPCR specific entries (profiles and patterns) from Pfam, PRINTS, and PROSITE were required to search the sequence database for GPCR candidates using these methods. We first obtained a list of PFAM, PRINTS, and PROSITE references (a list of accession numbers or IDs) from GPCRDB. These cross-references are available from the "DR" lines of the SWISS-PROT format entries. GPCR non-specific references were deleted from the lists manually. The remaining references were used to obtain profile/pattern entries from PFAM (Pfam_ls, Pfam_fs), PRINTS (pval file), and PROSITE (.dat file). In this project, 12 GPCR specific PFAM entries, 188 GPCR specific PRINTS entries, and 11 GPCR specific PROSITE entries were obtained (See Appendix 8).

The software tools developed in this process are:

- DRlist List PFAM, PRINTS, and PROSITE references from the sequence database for each sequence listed in the input file. The input file is a list of accession numbers of the protein sequences. The sequence database needs to be in SWISS-PROT format.
- DRtool Make a list of non-redundent database references by any of the sequence in the input file. The input file is a list of accession numbers of the protein sequences.
- DRMatrics List PFAM, PRINTS and PROSITE references in the matrix format for each sequence listed in the input file.
- fetchmms Fetch one or more hmm profiles from a hmm database (Pfam_ls or Pfam_fs).

- fetchPROSITE Fetch one or more entries from PROSITE main database, ie, PROSITE.dat.
- fetchPrints Fetch one or more print profiles from profile format of PRINTS databases (prints34_0.pval_blos62, prints34_0.pval_blos45, prints34_0.pval_blos80 etc., can be downloaded with FingerPRINTScan or created with FingerPRINTScan with –C option).

5.2 Classifying Sequence Database Using PFAM, PRINTS, PROSITE, and DA Methods, Respectively

The SWISS-PROT database was classified by using each classification method.

5.2.1 Classifying Sequence Database Using PFAM

We used "hmmpfam", (a program from HMMER) to search GPCR specific PFAM profiles against SWISS-PROT. The commands and options used are:

```
hmmpfam -E 1.0 -A0 --cut_ga -Z 101602 pfam.hmms.ls sprot40.fasta >
pfam.hmms.ls-sp&
```

```
hmmpfam -E 1.0 -A0 --cut_ga -Z 101602 pfam.hmms.fs sprot40.fasta >
pfam.hmms.fs-sp&
```

- -E 1.0 sets the E-value cutoff for the per-sequence ranked hit list to be 1.0. (Default E-value cutoff is 10.0, and Pfam web default is 1.0)
- -A0 shuts off the alignment output.
- --cut_ga instructs the "hmmpfam" to use Pfam GA (gathering threshold) score cutoffs set in HMM files. (Pfam web also uses this cutoff strategy)

- -Z 101602 informs "hmmpfam" to calculate the E-value scores as if we has seen a sequence database of size 101602 (the size of SWISS-PROT release 40 used in this project).

"pfam.hmms.fs" is the GPCR specific HMM database extracted from Pfam_fs using the list of GPCR specific PFAM entries mentioned earlier. Similarly "pfam.hmm.ls" is the GPCR specific HMM database extracted from Pfam_ls. "sprot40.fasta" is the SWISS-PROT database file in the FASTA format, and "pfam.hmms.ls-sp" and "pfam.hmms.fs-sp" are the output files for search results. An explanation of the HMM scores can be found at http://pfam.wustl.edu/scores.shtml.

5.2.2 Classifying Sequence Database Using PRINTS

“FingerPRINTScan” is the program to classify the SWISS-PROT database using PRINTS method. The commands and options used are:

```
FingerPRINTScan prints.profiles.62 sprot40.fasta -e 0.001 -E 257043 84355444 -o
4 -R > prints-sp-results
```

- -e sets the E-value threshold to 0.001.
- -E #1 #2 specifies the E-value calculation parameters.

  #1 is the number of sequences in the primary database (default: 80000). #2 is the number of residues in the primary database (default: 2.96103e+07)). In our case, we specifically set these two numbers to those used in "P-val FPScan" (http://bioinf.man.ac.uk/fingerPRINTScan/evalwarning.html). These two values are based on SWISS-PROT release 37 and TrEMBL release 9.
- -o 4 selects the "Table 2" output format, which includes a medium description of the results, and the ten top scoring hits detailed by fingerprint.
- -R restricts all results in all tables to those which score below the E-value threshold.

"prints.profiles.62" is the GPCR-specific fingerprint database in the profile format extracted from "prints34_0.pval_blos62", (the profile format file of prints34_0). The search result is saved in the file "prints-sp-results".

5.2.3 Classifying Sequence Database Using PROSITE

"ps_scan.pl" is a Perl program used to scan one or several patterns, rules and/or profiles from PROSITE against one or several protein sequences in SWISS-PROT or FASTA format. It requires two compiled external programs from the PFTOOLS package: "pfscan" used to scan a sequence against a profile library and "psa2msa" that is necessary for the "-o msa" output format only. The command and options used are:

```
ps_scan.pl -d PROSITE.entries sprot40.fasta -o pff > PROSITE-sp-
results &
```

- The E-value cutoff level is set to 0 (also the default cutoff level) to obtain trusted positive matches. (Cutoff level of -1 is used for potential weak matches, and is not selected in this project.)
- -o pff selects the tabular format output that lists bounding positions on the sequence and the profile, the raw and normalized profile score, and the cut-off level.
- -d option sets the PROSITE profile file used to scan the database file (sprot40.fastaThe search result is saved in the file "PROSITE-sp-results".

5.2.4 Classifying Sequence Database Using DA methods

Three DA methods are used in this project:

- LDA (Linear Discriminant Analysis)

- QDA (Quadratic Discriminant Analysis)
- LOG (Logistic Discriminant Analysis)

The three methods can be used independently. But in this project, we combined the results as follows. A protein is classified as a GPCR if any of the three DA methods classified it as a GPCR. On the other hand, a protein is classified as a non-GPCR if none of the three methods classified it as a GPCR. Therefore, the DA methods were used in the most conservative way in this project. The results were passed to other classification methods for further analysis.

The DA method classification includes the following process:

1. Preparing the training data set in FASTA format. Each entry was marked either with G (for GPCR sequences) or R (for random non-GPCR sequences).
2. Training each of the three DA methods. Variable sets were extracted from the training data. Following Kim *et al*. (2000) [1], "amino-acid index" (a linear discriminant score based on the amino-acid composition from each sequence) and moving-window variables as periodicity statistics based on hydrophobicity and polarity were used in this project.
3. Preparing the test sequences in FASTA format. The same set of variables described above was extracted from each of the test sequences. Trained functions generated in the step 2 were used to classify the test sequences.

The following subsections give more details on each step.

Preparation of the Training Data for DA methods

We used 1000 sequences in the training data set: 500 GPCR sequences and another 500 randomly selected non-GPCR sequences. All the sequences were obtained from SWISS-PROT release 40 and

their lengths were restricted to between 200 and 1000 amino acids. No partial (fragment) sequence was included in the training data set.

Among 500 GPCR sequences, 217 were chosen by sampling one from every 10 entries in each subgroup. The GPCR family list (http://www.gpcr.org/7tm/multali/multali.html) and the family member list (http://www.gpcr.org/7tm/htmls/entries.html) were used for this sampling. We picked from the lowest level subfamilies (*e.g.*, Acetylcholine Vertebrate type 5). Other 283 sequences were randomly sampled from GPCRDB. Classes Y and Z were excluded from sampling GPCRs. The putative /unclassified (orphan) members were used only in the random sampling of GPCRs.

Data Training for DA methods

The linear, quadratic, and logistic discriminant functions were trained based on the training data set. The trained functions were used to classify the SWISS-PROT data This process included the following steps (with used program names):

1. Compute amino acid frequencies from each protein. [aafstat]

2. Compute moving windows variables from each sequence. Three variables are computed: the logarithm of the average GES hydrophobicity periodicity, the logarithm of the average polarity periodicity, and the variance of the first derivative of polarity. The moving window size was set to 16 amino acids (see Kim *et al.* 2000 for more details on variables used). [mw_new and mwstat]

3. Import input files (*e.g.*, xxx.aafout and xxx.mwout) for "S-plus" statistical package.

4. Check the amino-acid frequency file (*e.g.*, xxx.aafout) and the moving window variable file (*e.g.*, xxx.mwout) contain the same set of entries in the same order.

5. Prepare a label file where each sequence is marked with "G" or "R". The label file must have the same set of entries as the other input files described in the step 4.

6. Do the linear discriminant function analysis on the amino acid frequencies. This step generates the LDA1 variable.

7. Do the linear, quadratic, and logistic discriminant function analyses on the four variables: LDA1 (computed in the step 6), the logarithm of the average GES periodicity, the logarithm of the average polarity periodicity, and the variance of the first derivative of the polarity (computed in the step 2). This step creates discriminant functions to be used later for protein classification.

The steps 4 to 7 were done with S-plus (version 6.2, and MASS library functions, lda, qda, and multinom).

A Perl script "train" was used to do these steps on a Linux command line in one step.

```
>./train train.fst
```

Running this command creates results files in S-plus objects: XXX.mat and XXX.aafmat.lda. "train.fst" is the prepared training data set.

Classification

This step classifies the SWISS-PROT data set into two groups: GPCR or non-GPCR. The discriminant functions produced in the training section are used. Similar to the training process, we also have a Perl script that does the classification in one step on a Linux command line.

```
>./predict train.fst sprot40.fst
```

Upon successful completion, a file called "emdb.txt" is generated is used to populate the "emdb" table in MyGPCR database using "load data local infile" command in the MySQL client.

1. Construction of MyGPCR Database

MyGPCR is a MySQL database. It is designed to integrate the classification results of the four classification methods described above. Currently MyGPCR has 18 tables. Creation and population of the majority of tables involve programming. The purpose of these programs is to create MySQL statements, which can be used with MySQL client to update the MyGPCR database. Table 2. lists data tables in MyGPCR database and their purposes, data source and related programs. Figure 5. shows the relationship between the data tables. Primary key is indicated by a "*". Database schemas can be found in Appendix 7.

TABLE	DESCRIPTION	NOTES
emdb	Classification result by DA Method	Parsed from emdb.txt, the output from Splus for DA; Can be loaded into MyGPCR using “load local in file” command from the MySQL Client
gpcr_family_db	GPCR Families	Parsed from http://www.gpcr.org/7tm/multali/multali.html
gpcr_family_info	Families each GPCR belongs to	Parsed from http://www.gpcr.org/7tm/htmls/entries.html
gpcr_subfamilies	GPCR subfamilies	Extracted from gpcr_family_db
gpcrdb	GPCR sequences collected in SWISS-PROT and TrEMBL	http://www.gpcr.org/7tm/htmls/entries.html
pfam	GPCR Specific PFAM entries	Refer to relevant section for obtaining these entries
pfam_fs_domains	Classification Result by PFAM using the local mode	Parsed from HMMPFAM result. The parser is: hmmpfam_result_parser
pfam_fs_scores	Classification Result by PFAM using local mode	Parsed from HMMPFAM result. The parser is: hmmpfam_result_parser
pfam_ls_domains	Classification Result by PFAM using the glocal mode	Parsed from HMMPFAM result. The parser is: hmmpfam_result_parser
pfam_ls_scores	Classification Result by PFAM using the glocal mode	Parsed from HMMPFAM result. The parser is: hmmpfam_result_parser
prints	GPCR Specific PRINTS entries	Refer to relevant section for obtaining these entries
prints_sp_predictions	Classification result by PRINTS	Parsed from FingerPRINTScan result. The parser is: FingerPrint_result_parser
PROSITE	GPCR Specific PROSITE entries	Refer to relevant section for obtaining these entries
PROSITE_sp_predictions	Classification result by PROSITE	Parsed from ps_scan.pl result. The parser is: ps_scan_result_parser
SWISS-PROT	SWISS-PROT Database	Parsed from SWISS-PROT Database. The parser is: DBtool (option 7)
pfam_matrics	Pfam reference matrix by GPCRDB	Parsed from DRMatrix output for Pfam
prints_matrics	PRINTS reference matrix by GPCRDB	Parsed from DRMatrix output for PRINTS
PROSITE_matrics	PROSITE reference matrix by GPCRDB	Parsed from DRMatrix output for PROSITE

Table 2. MyGPCR tables.

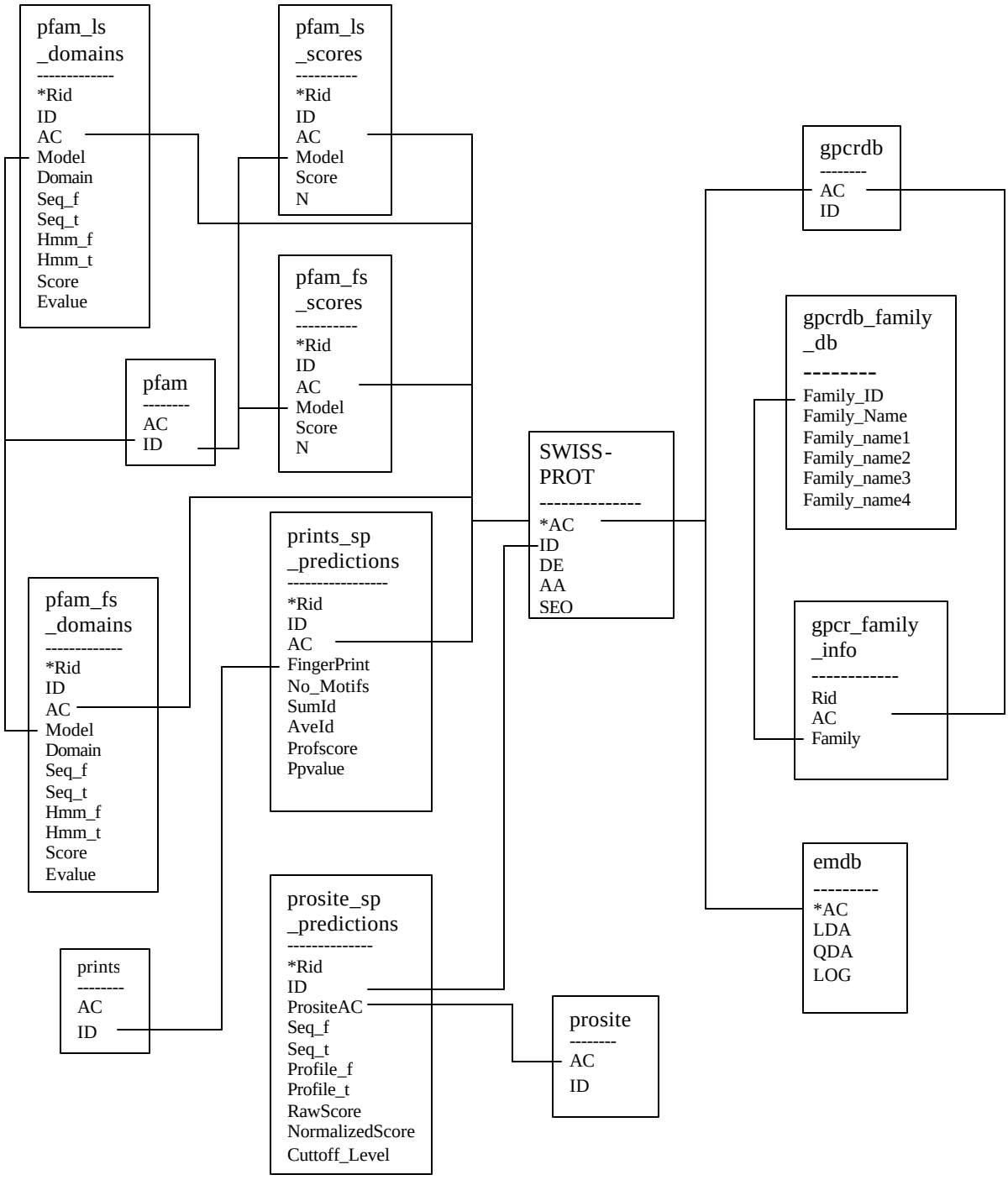

Figure 5. MyGPCR Tables and their Relationship.

7. Designing the Web Site

MyGPCR database can be easily accessed via any MySQL client using SQL language. It is a very flexible tool to do various bioinformatics research. We can also deploy the OpenOffice so that MyGPCR can be accessed thought a graphical user interface. MyGPCR can be accessed also in PHP, Perl, or C/C++ for more flexible manipulation or web site design.

In this project a web based hierarchical classification scheme illustrated in Figure 6. is designed. It can be accessed through the Internet. In this figure, every box is linked to a page. 3D boxes are datasets, while the light blue boxes are linked to descriptions relevant to the classification methods used. We first classified the SWISS-PROT database using the DA methods, which resulted in two datasets, GPCRs (G) and non-GPCRs (R). Next we used PFAM, PROSITE, and PRINTS to classify GPCR (G) candidates into several categories: for example proteins that were classified by all methods as G, proteins that were classified by some of the methods as G, and proteins that were classified as G only by DA methods. By examining each subset, we can find proteins, for example, that are more likely to be GPCRs, that are only weakly supported to be GPCRs by more conservative methods (PFAM, PROSITE, and PRINTS), and that are not classified as GPCRs by these methods but only by the DA method. The last protein group would be of most interest. They could include new members of GPCRs or those misidentified by other methods.

The numbers attached with datasets in the Figure 6. are not present in the actual web page. They are listed in this figure just to show the number of sequences in each dataset. As can be seen, the three more conservative methods (PFAM, PRINTS, and PROSITE) produce similar results while DA is very liberal in this discrimination. In each dataset, we present the number of pattern/profiles/fingerprints matched by each method in a table form for every sequence. Users can follow the links to see more detailed information for each match. The table also indicates whether

the sequence is included in GPCRDB or not. If it is included in GPCRDB, users can follow the link to see which GPCR family it belongs to.

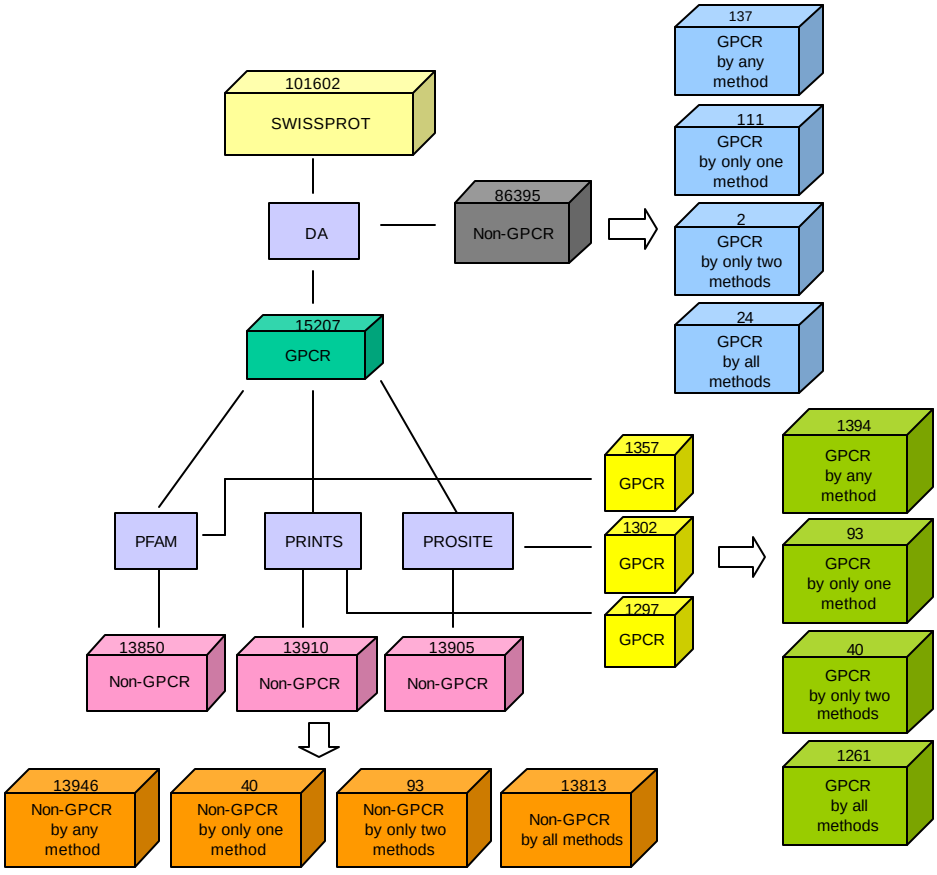

Figure 6. Hierarchical Protein Classification.

The Figure 7. is a hard copy from the web page that displays sequences classified as GPCR by all the methods (DA, PFAM, PRINTS, and PROSITE). The first two fields are for SWISS-PROT accession number and identification. LDA, QDA, and LOG are three methods in DA method. For PFAM method, there are two modes of search: ls mode for glocal search mode, ie, glocal to the profile but local to the sequence; and fs mode for local search mode, ie, local to both profile and sequence. The numbers in the columns of PFAM_LS and PFAM_FS indicate number of Pfam profiles matched in the “ls” mode and “fs” mode, respectively. When these numbers are not “0”,

you can follow the link to see what profile(s) are matched in the sequence, and how they are matched, as in Figure 8. The score section shows the models, ie, PFAM profiles, matched with the sequence. The score and E-value indicates how good the matches are. Only matches that are better than the specified E-value are shown in this table. See "Command and Options" for PFAM method to see the E-value cutoff. "Num. of Domain" indicates how many times the particular model appears in the given sequence. The domain section gives more details as to how the profiles and the sequence are matched. For "ls" mode, the whole HMM profile should be found somewhere in the sequence. For "fs" mode, you will expect partial HMM profile to be found in the sequence. The column titled "PRINTS" shows the number of fingerprints found in the sequence. If the number is non-zero, you may click the link and find the fingerprints matched in the sequence and details of the matches, as in Figure 9. In the column title "PROSITE", you are expected to see the number of PROSITE patterns, rules, or profiles matched with the sequence. Follow the link to see what patterns, rules or profiles are matched and how they are matched, as in Figure 10. (Some of the columns in Figure 10. apply only to PROSITE profiles only). The column title "GPCRDB" shows whether or not the sequence is included in GPCRDB. When a sequence is indeed included in the GPCRDB, follow the link and you will see the GPCR family or families the sequence belongs to. Sometimes a sequence can be in several families, as in Figure 11.

AC	ID	LDA	QDA	LOG	PFAM_LS	PFAM_FS	PRINTS	PROSITE	GPCRDB	DESC
O42385	5H1A_FUGRU	G	G	G	1	1	5	2	Y	5-hydroxytryptamine …
P08908	5H1A_HUMAN	G	G	G	1	1	5	2	Y	5-hydroxytryptamine …
Q64264	5H1A_MOUSE	G	G	G	1	1	5	2	Y	5-hydroxytryptamine…
P19327	5H1A_RAT	G	G	G	1	1	5	2	Y	5-hydroxytryptamine …

Figure 7. Information Shown for each Dataset.

Scores:

AC	ID	Model	Score	E-value	Num. of Domains
O42385	5H1A_FUGRU	7tm_1	353.4	4.3e-102	1

Domains:

AC	ID	Model	Domain	Seq_f	Seq_t	Hmm_f	Hmm_t	Score	E-value
O42385	5H1A_FUGRU	7tm_1	1/1	62	401	1	275	353.4	4.3e-102

Figure 8. PFAM Match Details.

AC	ID	FingerPrint	No_Motifs	SumId	AveId	Profscore	Ppvalue	E-value	GraphScan
O42385	5H1A_FUGRU	GPCRRHODOPSN	7/7	236.57	33.8	2722	1.2e-59	3.1e-54	iIIiIII
O42385	5H1A_FUGRU	5HT1ARECEPTR	4/7	206.16	51.54	1772	8.3e-23	2.1e-17	.III.I.
O42385	5H1A_FUGRU	5HTRECEPTOR	4/4	218.29	54.57	1016	2.9e-16	7.5e-11	IIII
O42385	5H1A_FUGRU	NRPEPTIDEYR	4/5	126.1	31.52	925	2.5e-11	6.4e-06	IiI.I
O42385	5H1A_FUGRU	ADRENERGICR	4/4	177.29	44.32	922	4.7e-11	1.2e-05	iIII

Figure 9. PRINTS Match Details.

PROSITE Scores for 5H1A_FUGRU (produced by "ps_scan.pl")

PROSITEID	Seq_f	Seq_t	Profile_f	Profile_t	RawScore	NormalizedScore	Cuttoff_Level
G_PROTEIN_RECEP_F1_1	131	147					
G_PROTEIN_RECEP_F1_2	62	401	1	-1	2147	45.006	0

Figure 10. PROSITE Match Details.

AC: O42385

Family: Class A Rhodopsin like

Amine

Serotonin

Serotonin Vertebrate type 1

Figure 11. GPCR Family a Sequence Belongs to.

The web site also gives some other tools to direct access the classification results for a particular sequence by different method(s). For more details please visit: http://em-13.unl.edu.

8. Conclusion and Future Work

We now have a complete set of databases and basic tools for updating the database. The majority of hierarchical classification scheme was also implemented. The next steps are:

1. Analyze and examine the results obtained from SWISS-PROT classification, and try to find any new GPCR family members or possible miss-classifications by other tradition methods.
2. Improve DA method based on the results we examine.
3. Extend the current system so that it can be used to classify different transmembrane or other protein families.
4. Provide a web-based interface for easy access to the MyGPCR database. It might be something that allows MyGPCR to be queried using SQL language. See SMART (http://smart.embl-heidelberg.de/sql_selective.shtml) for example.
5. Provide a web based interface for training for DA method.

References

[1] Kim, J., Moriyama, E. N., Warr, C. G., Clyne, P. J., Carlson, J. R. (2000). Identification of novel multi-transmembrane proteins from genomic databases using quasi-periodic structural properties. *Bioinformatics* 16: 767-775.

[2] Pierce,K., L., Premont, R., T., Lefkowitz, R., J., (2002). Seven-Transmembrane Receptors. *Nature Reviews: Molecular Cell Biology* 3: 639-650.

[3] Horn, F., Weare, J., Beukers, M.W., Hörsch, S., Bairoch, A., Chen, W., Edvardsen, Ø., Campagne, F., Vriend, G. (1998). GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Research*. 26(1): 277-281.

[4] Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C. J. A., Hofmann, K., Bairoch., A. (2002). The PROSITE database, its status in 2002. *Nucleic Acids Research*, 30: 235-238**.**

[5] Attwood, T.K. (2002**).** The PRINTS database: a resource for identification of protein families. *Briefings in Bioinformatics*, **3**(3), 252-263.

[6] Attwood, T.K. & Flower, D.R. (2002**).** Trawling the genome for G protein-coupled receptors: the importance of integrating bioinformatic approaches. *Drug Design - Cutting Edge Approaches*, 60-71.

[7] Bateman, A., Birney, E., Durbin R., Eddy., S., R., Howe, K., L., Sonnhammer, E., L., L. (2000). The Pfam Protein Family Database. *Nucleic Acids Research*. 28(1): 263-266.

[8] Karchin, R., Karplus K., Haussler, D., (2002), Classifying G-protein coupled receptors with support vector machines, *Bioinformatics* 18(1):147-159.

[9] Brejova, B., DiMarco, C., Vinar, T., Hidalgo, S. R., Holguin, G., Patten, C. (2000). Project report for CS798g: Finding Patterns in Biological Sequences. University of Waterloo.

[10] Brief Tutorial on Family Classification. (from a MetaFam help page) http://metafam.ahc.umn.edu/protFamilyAlgo.html

[11] Predictive Tools Available on the Web for Protein Sequence and Structure Analysis

http://www.niehs.nih.gov/Connections/2002/mar/protein.htm

[12] Motifs, Profiles, Hidden Markov Model.

http://www.ludwig.edu.au/course/course2002/talks/crc02motif_edit/sld001.htm

[13] Description of NCBI BLAST.

http://www.ncbi.nlm.nih.gov/blast/html/BLASThomehelp.html#AABLAST

[14] Huang, G.M, Farkas, J., Hood, L. (1996) High-throught cDNA screening utilizing a low order neural network filter. *Biotechniques* 21: 1110-1114.

Related Documentations and Links

Bioinformatics Related

1. ftp://ftp.genetics.wustl.edu/pub/eddy/papers/hmmreview-bioinformatics-98.pdf
2. http://www.paracel.com/publications/hmm_white_paper.html
3. PSTOOLS, http://www.isrec.isb-sib.ch/ftp-server/pftools/pft2.2/
4. ps_scan, ftp://us.expasy.org/databases/PROSITE/tools/ps_scan/
5. SWISS-PROT Protein Knowledgebase User Manual, Release 40, October 2001.

   http://us.expasy.org/sprot/userman.html.
6. THE PRINTS USER GUIDE.

   http://bioinf.man.ac.uk/dbbrowser/PRINTS/printsman.html.
7. Bioinformatics Frequently Asked Questions

   http://www.hgmp.mrc.ac.uk/CCP11/bioinformaticsfaq.jsp

Computing Related

1. How to use Access 2000 as a database front end.

http://builder.com.com/article.jhtml?id=u00320021007SKJ01.htm&page=1&vf=tt

2. HOWTO provide Microsoft Access-like functionality on Linux with open-source tools - OpenOffice.org 1.0, unixODBC, and MySQL.

   http://www.unixodbc.org/doc/OOoMySQL.pdf

3. MyODBC, the MySQL ODBC driver.

   http://www.mysql.com/products/myodbc/index.html

4. MySQL

   http://mysql.com

Appendix 1: A Sample SWISSPROT Entry

```
ID   5H1A_FUGRU     STANDARD;      PRT;   423 AA.
AC   O42385;
DT   30-MAY-2000 (Rel. 39, Created)
DT   30-MAY-2000 (Rel. 39, Last sequence update)
DT   30-MAY-2000 (Rel. 39, Last annotation update)
DE   5-hydroxytryptamine 1A-alpha receptor (5-HT-1A-alpha) (Serotonin
DE   receptor) (5-HT1A-alpha) (F1A).
OS   Fugu rubripes (Japanese pufferfish) (Takifugu rubripes).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Actinopterygii; Neopterygii; Teleostei; Euteleostei; Neoteleostei;
OC   Acanthomorpha; Acanthopterygii; Percomorpha; Tetraodontiformes;
OC   Tetraodontidae; Takifugu.
OX   NCBI_TaxID=31033;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Testis;
RX   MEDLINE=97361762; PubMed=9218723;
RA   Yamaguchi F., Brenner S.;
RT   "Molecular cloning of 5-hydroxytryptamine (5-HT) type 1 receptor
RT   genes from the Japanese puffer fish, Fugu rubripes.";
RL   Gene 191:219-223(1997).
CC   -!- FUNCTION: THIS IS ONE OF THE SEVERAL DIFFERENT RECEPTORS FOR 5-
CC       HYDROXYTRYPTAMINE (SEROTONIN), A BIOGENIC HORMONE THAT FUNCTIONS
CC       AS A NEUROTRANSMITTER, A HORMONE, AND A MITOGEN. THE ACTIVITY OF
CC       THIS RECEPTOR IS MEDIATED BY G PROTEINS THAT INHIBITS ADENYLATE
CC       CYCLASE ACTIVITY (BY SIMILARITY).
CC   -!- SUBCELLULAR LOCATION: INTEGRAL MEMBRANE PROTEIN.
CC   -!- SIMILARITY: BELONGS TO FAMILY 1 OF G-PROTEIN COUPLED RECEPTORS.
CC       STRONGEST TO THE OTHER 5HT-1 SUBTYPE RECEPTORS.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between  the Swiss Institute of Bioinformatics  and the  EMBL outstation -
CC   the European Bioinformatics Institute.  There are no  restrictions on  its
CC   use  by  non-profit  institutions as long  as its content  is  in  no  way
CC   modified and this statement is not removed.  Usage  by  and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; X95936; CAA65175.1; -.
DR   GCRDb; GCR_2429; -.
DR   InterPro; IPR000276; GPCR_Rhodpsn.
DR   Pfam; PF00001; 7tm_1; 1.
DR   PRINTS; PR00237; GPCRRHODOPSN.
DR   PROSITE; PS00237; G_PROTEIN_RECEP_F1_1; 1.
DR   PROSITE; PS50262; G_PROTEIN_RECEP_F1_2; 1.
KW   G-protein coupled receptor; Transmembrane; Glycoprotein;
KW   Multigene family.
FT   DOMAIN        1     45       EXTRACELLULAR (POTENTIAL).
FT   TRANSMEM     46     71       1 (POTENTIAL).
FT   DOMAIN       72     82       CYTOPLASMIC (POTENTIAL).
FT   TRANSMEM     83    107       2 (POTENTIAL).
```

```
FT   DOMAIN      108    118       EXTRACELLULAR (POTENTIAL).
FT   TRANSMEM    119    141       3 (POTENTIAL).
FT   DOMAIN      142    161       CYTOPLASMIC (POTENTIAL).
FT   TRANSMEM    162    186       4 (POTENTIAL).
FT   DOMAIN      188    200       EXTRACELLULAR (POTENTIAL).
FT   TRANSMEM    201    226       5 (POTENTIAL).
FT   DOMAIN      227    346       CYTOPLASMIC (POTENTIAL).
FT   TRANSMEM    347    368       6 (POTENTIAL).
FT   DOMAIN      369    379       EXTRACELLULAR (POTENTIAL).
FT   TRANSMEM    380    404       7 (POTENTIAL).
FT   DOMAIN      405    423       CYTOPLASMIC (POTENTIAL).
FT   DISULFID    118    196       BY SIMILARITY.
FT   CARBOHYD      9      9       N-LINKED (GLCNAC...) (POTENTIAL).
FT   CARBOHYD     12     12       N-LINKED (GLCNAC...) (POTENTIAL).
FT   CARBOHYD     30     30       N-LINKED (GLCNAC...) (POTENTIAL).
SQ   SEQUENCE   423 AA;  47000 MW;  7B1308626B40190F CRC64;
     MDLRATSSND SNATSGYSDT AAVDWDEGEN ATGSGSLPDP ELSYQIITSL FLGALILCSI
     FGNSCVVAAI ALERSLQNVA NYLIGSLAVT DLMVSVLVLP MAALYQVLNK WTLGQDICDL
     FIALDVLCCT SSILHLCAIA LDRYWAITDP IDYVNKRTPR RAAVLISVTW LIGFSISIPP
     MLGWRSAEDR ANPDACIISQ DPGYTIYSTF GAFYIPLILM LVLYGRIFKA ARFRIRKTVK
     KTEKAKASDM CLTLSPAVFH KRANGDAVSA EWKRGYKFKP SSPCANGAVR HGEEMESLEI
     IEVNSNSKTH LPLPNTPQSS SHENINEKTT GTRRKIALAR ERKTVKTLGI IMGTFIFCWL
     PFFIVALVLP FCAENCYMPE WLGAVINWLG YSNSLLNPII YAYFNKDFQS AFKKILRCKF
     HRH
```

Appendix 2: SWISS-PROT Line Codes and their Meaning

Line code	Content	Occurrence in an entry
ID	Identification	Once; starts the entry
AC	Accession number(s)	Once or more
DT	Date	Three times
DE	Description	Once or more
GN	Gene name(s)	Optional
OS	Organism species	Once or more
OG	Organelle	Optional
OC	Organism classification	Once or more
OX	Taxonomy cross-reference(s)	Once or more
RN	Reference number	Once or more
RP	Reference position	Once or more
RC	Reference comment(s)	Optional
RX	Reference cross-reference(s)	Optional
RA	Reference authors	Once or more
RT	Reference title	Optional
RL	Reference location	Once or more
CC	Comments or notes	Optional
DR	Database cross-references	Optional
KW	Keywords	Optional
FT	Feature table data	Optional
SQ	Sequence header	Once
	(blanks) sequence data	Once or more
//	Termination line	Once; ends the entry

Appendix 3: A Sample PFAM HMM Profile

```
HMMER2.0  [2.2g]
NAME  Octapeptide
ACC   PF03373
DESC  Octapeptide repeat
LENG  8
ALPH  Amino
RF    no
CS    no
MAP   yes
COM   hmmbuild -F HMM_ls.ann SEED.ann
COM   hmmcalibrate --seed 0 HMM_ls.ann
NSEQ  3
DATE  Sun May 26 17:32:42 2002
CKSUM 7926
GA    25.0 0.0
TC    90.9 7.9
NC    16.1 9.4
XT      -8455    -4  -1000  -1000  -8455    -4  -8455    -4
NULT    -4  -8455
NULE     595  -1558     85    338   -294    453  -1158    197    249    902  -1085   -142    -21   -313     45    531    201    384  -1998   -644
EVD   -4.960945   0.671333
HMM       A     C     D     E     F     G     H     I     K     L     M     N     P     Q     R     S     T     V     W     Y
        m->m  m->i  m->d  i->m  i->i  d->m  d->d  b->m  m->e
        -585     *  -1585
    1  -934 -1220 -1352 -1451 -2187 -1314 -1456 -2160 -1488 -2268 -1827 -1359  3760 -1471 -1571 -1129 -1216 -1780 -2050 -2023     1
    -  -149  -500   233    43  -381   399   106  -626   210  -466  -720   275   394    45    96   359   117  -369  -294  -249
    -   -33 -6045 -7087  -894 -1115  -701 -1378  -585     *
    2  -361 -1240   -91    84 -1933  1816  -171 -1618  1608 -1698  -942   -98 -1399   173   114  -363  -425 -1231 -1850 -1384     2
    -  -149  -500   233    43  -381   399   106  -626   210  -466  -720   275   394    45    96   359   117  -369  -294  -249
    -   -33 -6045 -7087  -894 -1115  -701 -1378     *     *
    3  -565 -1785   177  1387 -2153 -1202   100 -1803  2167 -1757  -974   125 -1405   502   515  -433  -493 -1455 -1882 -1346     3
    -  -149  -500   233    43  -381   399   106  -626   210  -466  -720   275   394    45    96   359   117  -369  -294  -249
    -   -33 -6045 -7087  -894 -1115  -701 -1378     *     *
    4  -988 -1921   379  3052 -2309 -1196  -495 -2108  -350 -2194 -1612  -103 -1621  -210  -770  -862 -1042 -1799 -2207 -1742     4
    -  -149  -500   233    43  -381   399   106  -626   210  -466  -720   275   394    45    96   359   117  -369  -294  -249
    -   -33 -6045 -7087  -894 -1115  -701 -1378     *     *
    5 -1130 -2063  3441   243 -2502 -1157  -662 -2514  -843 -2578 -2029   -96 -1658  -421 -1434  -970 -1239 -2138 -2388 -1915     5
    -  -149  -500   233    43  -381   399   106  -626   210  -466  -720   275   394    45    96   359   117  -369  -294  -249
    -   -33 -6045 -7087  -894 -1115  -701 -1378     *     *
    6  -176 -1067   254   189 -2018  1150  -376 -1848  -236 -1962 -1191  2581 -1301   -75  -656  -200  -360 -1320 -2101 -1523     6
    -  -149  -500   233    43  -381   399   106  -626   210  -466  -720   275   394    45    96   359   117  -369  -294  -249
    -   -33 -6045 -7087  -894 -1115  -701 -1378     *     *
    7  -496 -1628   203   370 -2003 -1139    44 -1727  1587 -1727  -953  2277 -1389   426   371  -395  -465 -1371 -1851 -1287     7
    -  -149  -500   233    43  -381   399   106  -626   210  -466  -720   275   394    45    96   359   117  -369  -294  -249
    -   -33 -6045 -7087  -894 -1115  -701 -1378     *     *
    8  -420 -1360  -157   111 -1979   907   -67 -1648  2256 -1685  -919   -80 -1413   302   354  -394  -436 -1278 -1813 -1352     8
    -     *     *     *     *     *     *     *     *     *     *     *     *     *     *     *     *     *     *     *     *
    -     *     *     *     *     *     *     *     *     0
//
```

Appendix 4: PFAM Database Line Code and their Meaning

ACC	Accession number. PFxxxxx for Pfam-A accession numbers, and PBxxxxxx for Pfam-B accession numbers
Name	Name of the profile.
DESC	Description of the entry.
LENG	Length of the profile.
AL	Alignment method of seed, ie, the method used to align the seed members.
COM	Command used to generate the profile.
CKSUM	Checksum.
GA	Gathering Threshold, ie, the search threshold to build the full alignment. The order of the thresholds is ls mode sequence, ls mode domain for glocal profile database (Pfam_ls) and fs mode sequence, fs domain mode for local profile database (Pfam_fs).
TC	Trusted Threshold. The lowest scoring match in the full alignment The order of the thresholds is ls mode sequence, ls mode domain for glocal profile database (Pfam_ls) and fs mode sequence, fs domain mode for local profile database (Pfam_fs).
NC	Noise Cutoff. The highest score match not in the full alignment. The order of the thresholds is ls mode sequence, ls mode domain for glocal profile database (Pfam_ls) and fs mode sequence, fs domain mode for local profile database (Pfam_fs).
HMM	Start of HMM profile.

Appendix 5: A Sample PRINTS Entry

The fingerprint is composed of 2 motifs, designated ANTISTASIN1 and ANTISTASIN2 respectively. There are 8 sequences that match all of the 2 motifs in this fingerprint. A short description is provided for each of these 8 sequences in lines that start with "tt". The initial alignment contains 4 sequences. And final alignment contains 8 sequences, ie, 8 truth positives.

```
gc; ANTISTASIN
gx; PR01706
gn; COMPOUND(2)
ga; 01-APR-2002
gt; Antistasin signature
gp; INTERPRO; IPR004094
gp; PDB; 1SKZ
gp; SCOP; 1SKZ
gp; CATH; 1SKZ
bb;
gr; 1. NUTT, E., GASIC, T., RODKEY, J., GASIC, G.J., JACOBS, J.W., FRIEDMAN, P.A.
gr; AND SIMPSON, E.

…

si; SUMMARY INFORMATION
si; --------------------
sd;    8 codes involving  2 elements
bb;
bb;
ci; COMPOSITE FINGERPRINT INDEX
ci; ---------------------------
cr;
cd;  2   8    8
cd; --+-----------
cd;      1    2
bb;
bb;
tp; ANTA_HAEGH     Q9TWX3        Q9TWQ7        Q25065
tp; ANTA_HIRME     GUAM_HIRNI     ANTA_HYDMA     ANTA_HAEOF
bb;
tt; ANTA_HAEGH       Ghilanten - Haementeria ghilianii (Amazon leech).
tt; Q9TWX3           GHILANTEN - Haementeria ghilianii (Amazon leech).
tt; Q9TWQ7           ANTISTASIN ISOFORM B, ATS ISOFORM B=BLOOD COAGULATION FACTOR XA INHIBITOR - Haem
tt; Q25065           GHILANTEN - Haementeria ghilianii (Amazon leech).
tt; ANTA_HIRME       Hirustasin - Hirudo medicinalis (Medicinal leech).
tt; GUAM_HIRNI       Guamerin - Hirudo nipponia (Leech).
tt; ANTA_HYDMA       Antistasin precursor (ATS) (Blood coagulation factor Xa/proclotting enzyme inhib
tt; ANTA_HAEOF       Antistasin precursor (ATS) (Blood coagulation factor Xa/proclotting enzyme inhib
bb;
bb;
sh; SCAN HISTORY
sh; ------------
dn; SPTR40_18f  2   100 NSINGLE
bb;
bb;
im; INITIAL MOTIF-SETS
```

```
im; ------------------
ic; ANTISTASIN1
il; 21
it; Antistasin motif I - 1
id; DTHGLCGEKTCSPAQVCLNNE          GUAM_HIRNI     7    7
id; TQGNTCGGETCSAAQVCLKGK          ANTA_HIRME     1    1
id; VDANGCQICRCKRSALEAPEK          ANTA_HYDMA    69   69
id; PFGPGCEEAGCPEGSACNIIT          ANTA_HAEOF    20   20
bb;
ic; ANTISTASIN2
il; 16
it; Antistasin motif II - 1
id; CMIFCPNGFKVDENGC               GUAM_HIRNI    35    7
id; CRIRCKYGLKKDENGC               ANTA_HIRME    29    7
id; CKMHCENGFVRDENGC               ANTA_HYDMA    97    7
id; CRMHCPHGFQRSRYGC               ANTA_HAEOF    50    9
bb;
fm; FINAL MOTIF-SETS
fm; ----------------
fc; ANTISTASIN1
fl; 21
ft; Antistasin motif I - 2
fd; PMKATCDISECPEGMMCSRLT          ANTA_HAEGH    57   57
fd; PMKATCDISECPEGMMCSRLT            Q9TWX3      57   57
fd; PMKATCDISECPEGMMCSRLT            Q9TWQ7      57   57
fd; PMKATCDISECPEGMMCSRLT            Q25065      58   58
fd; TQGNTCGGETCSAAQVCLKGK          ANTA_HIRME     1    1
fd; DTHGLCGEKTCSPAQVCLNNE          GUAM_HIRNI     7    7
fd; VDANGCQICRCKRSALEAPEK          ANTA_HYDMA    69   69
fd; PFGPGCEEAGCPEGSACNIIT          ANTA_HAEOF    20   20
bb;
fc; ANTISTASIN2
fl; 16
ft; Antistasin motif II - 2
fd; CRKTCPNGLKRDKLGC               ANTA_HAEGH    88   10
fd; CRKTCPNGLKRDKLGC                 Q9TWX3      88   10
fd; CRKTCPNGLKRDKLGC                 Q9TWQ7      88   10
fd; CRKTCPNGLKRDKLGC                 Q25065      89   10
fd; CRIRCKYGLKKDENGC               ANTA_HIRME    29    7
fd; CMIFCPNGFKVDENGC               GUAM_HIRNI    35    7
fd; CKMHCENGFVRDENGC               ANTA_HYDMA    97    7
fd; CRMHCPHGFQRSRYGC               ANTA_HAEOF    50    9
```

Appendix 6: PROSITE Line Code and Sample Entries

```
PROSITE Line Code:

   ID  Identification                    (Begins each entry; 1 per entry)
   AC  Accession number                  (1 per entry)
   DT  Date                              (1 per entry)
   DE  Short description                 (1 per entry)
   PA  Pattern                           (>=0 per entry)
   MA  Matrix/profile                    (>=0 per entry)
   RU  Rule                              (>=0 per entry)
   NR  Numerical results                 (>=0 per entry)
   CC  Comments                          (>=0 per entry)
   DR  Cross-references to SWISS-PROT    (>=0 per entry)
   3D  Cross-references to PDB           (>=0 per entry)
   DO  Pointer to the documentation file (1 per entry)
   //  Termination line                  (Ends each entry; 1 per entry)
```

Example of PROSITE pattern entry:

```
   ID   T4_DEIODINASE; PATTERN.
   AC   PS01205;
   DT   NOV-1997 (CREATED); JUL-1999 (DATA UPDATE); JUL-1999 (INFO UPDATE).
   DE   Iodothyronine deiodinases active site.
   PA   R-P-L-[IV]-x-[NS]-F-G-S-[CA]-T-C-P-x-F.
   NR   /RELEASE=40.7,103373;
   NR   /TOTAL=16(16); /POSITIVE=16(16); /UNKNOWN=0(0); /FALSE_POS=0(0);
   NR   /FALSE_NEG=0; /PARTIAL=0;
   CC   /TAXO-RANGE=??E??; /MAX-REPEAT=1;
   CC   /SITE=12,active_site;
   DR   P49894, IOD1_CANFA, T; O42411, IOD1_CHICK, T; P49895, IOD1_HUMAN, T;
   DR   Q61153, IOD1_MOUSE, T; O42449, IOD1_ORENI, T; P24389, IOD1_RAT  , T;
   DR   P79747, IOD2_FUNHE, T; Q92813, IOD2_HUMAN, T; Q9Z1Y9, IOD2_MOUSE, T;
   DR   P49896, IOD2_RANCA, T; P70551, IOD2_RAT  , T; O42412, IOD3_CHICK, T;
   DR   P55073, IOD3_HUMAN, T; P49898, IOD3_RANCA, T; P49897, IOD3_RAT  , T;
   DR   P49899, IOD3_XENLA, T;
   DO   PDOC00925;
   //
```

Example of a PROSITE rule entry:

```
    ID   GLYCOSAMINOGLYCAN; RULE.
    AC   PS00002;
    DT   APR-1990 (CREATED); APR-1990 (DATA UPDATE); APR-1990 (INFO UPDATE).
    DE   Glycosaminoglycan attachment site.
    PA   S-G-x-G.
    RU   Additional rules:
    RU   There must be at least two acidic amino acids (Glu or Asp) from -2 to
    RU   -4 relative to the serine.
    CC   /TAXO-RANGE=??E??;
    CC   /SITE=1,glycosaminoglycan;
    CC   /SKIP-FLAG=TRUE;
```

```
DO   PDOC00002;
//
```

Example of a PROSITE profile (matrix) entry:

```
ID   HSP20; MATRIX.
AC   PS01031;
DT   JUN-1994 (CREATED); JUN-1994 (DATA UPDATE); NOV-1995 (INFO UPDATE).
DE   Heat shock hsp20 proteins family profile.
MA   /GENERAL_SPEC: ALPHABET='ACDEFGHIKLMNPQRSTVWY'; LENGTH=97;
MA   /DISJOINT: DEFINITION=PROTECT; N1=2; N2=96;
MA   /NORMALIZATION: MODE=1; FUNCTION=GLE_ZSCORE;
MA    R1=239.0; R2=-0.0036; R3=0.8341; R4=1.016; R5=0.169;
MA   /CUT_OFF: LEVEL=0; SCORE=400; N_SCORE=10.0; MODE=1;
MA   /DEFAULT: MI=-210; MD=-210; IM=0; DM=0; I=-20; D=-20;
MA   /M: SY='R'; M=-12,-44,-11,-13,-13,-22,-2,-7,18,-12,5,-3,-11,0,21,-6,-5,-11,-16,-34;
MA   /M: SY='D'; M=1,-41,17,16,-41,-3,3,-11,-1,-22,-12,8,-7,12,-7,0,-2,-19,-53,-36;
MA   /M: SY='D';  M=2,-37,15,13,-36,2,5,-15,-3,-26,-17,10,-6,7,-10,3,2,-17,-53,-28;
MA   /M: SY='P'; M=1,-41,6,8,-38,-4,2,-20,9,-30,-14,6,13,9,8,3,0,-22,-48,-45;
MA   /M: SY='D'; M=2,-43,23,20,-42,2,9,-18,2,-30,-18,14,-5,14,-6,2,0,-21,-57,-35;
MA   /M: SY='D'; M=4,-34,9,8,-34,6,0,-17,5,-29,-14,8,-1,5,1,5,2,-17,-47,-38;
MA   /M: SY='F'; M=-28,-32,-38,-38,50,-42,-1,2,-11,6,-6,-21,-35,-27,-27,-24,-23,-14,-3,47;
MA   /M: SY='Q'; M=0,-33,-2,-7,-26,-9,-4,1,1,-10,1,-1,-5,2,0,-2,1,0,-44,-37;
MA   /M: SY='L'; M=-13,-36,-34,-37,23,-31,-21,28,-15,29,24,-24,-25,-24,-27,-20,-10,22,-33,0;
MA   /M: SY='K'; M=-8,-32,-5,-5,-19,-16,3,-11,13,-19,-2,1,-9,2,12,-3,-3,-15,-32,-28;
MA   /M: SY='L'; M=-10,-39,-30,-32,15,-26,-20,20,-16,27,20,-21,-20,-21,-27,-17,-9,16,-32,-5;
MA   /M: SY='D'; M=3,-48,33,27,-51,4,6,-19,0,-35,-22,18,-10,13,-13,2,0,-16,-65,-41;
MA   /I: MI=-55; MD=-55; I=-5;
MA   /M: SY='V'; D=-5; M=-3,-33,-23,-32,-5,-19,-21,28,-16,26,30,-17,-14,-15,-19,-12,-1,30,-48,-28;
MA   /I: MI=-55; MD=-55; I=-5;
MA   /M: SY='P'; D=-5; M=1,-2,-1,0,-3,0,0,-1,-1,-2,-2,0,4,0,0,1,0,-1,-4,-4;
MA   /I: MI=-55; MD=-55; I=-5;
..
... Some lines omitted..
..
MA   /M:  SY='K'; M=-11,-52,1,-1,-1,-17,2,-18,43,-28,3,9,-10,8,33,-2,-1,-23,-33,-43;
MA   /I: MI=*; MD=*; I=0;
NR   /RELEASE=40.7,103373;
NR   /TOTAL=181(180); /POSITIVE=176(175); /UNKNOWN=5(5); /FALSE_POS=0(0);
NR   /FALSE_NEG=0; /PARTIAL=4;
CC   /MATRIX_TYPE=protein_domain;
CC   /SCALING_DB=reversed;
CC   /AUTHOR=P_Bucher;
CC   /TAXO-RANGE=A?EP?; /MAX-REPEAT=2;
DR   P30223, 14KD_MYCTU, T; P46729, 18K1_MYCAV, T; P46730, 18K1_MYCIT, T;
DR   P46731, 18K2_MYCAV, T; P46732, 18K2_MYCIT, T; P12809, 18KD_MYCLE, T;
DR   P80485, ASP1_STRTR, T; O30851, ASP2_STRTR, T; P02497, CRA2_MESAU, T;
DR   P24622, CRA2_MOUSE, T; P24623, CRA2_RAT  , T; P15990, CRA2_SPAEH, T;
..
... Some lines omitted..
..
DR   P96193, IBPB_AZOVI, T; P29210, IBPB_ECOLI, T; P29778, OV21_ONCVO, T;
DR   P29779, OV22_ONCVO, T; Q06823, SP21_STIAU, T; P34328, YKZ1_CAEEL, T;
DR   P12812, P40_SCHMA , T;
DR   P81083, HS11_PINPS, P; P81161, HS2M_LYCES, P; P30220, HS3E_XENLA, P;
```

DR Q9QUK5, HSB7_RAT , P;
DR Q29438, ODFP_BOVIN, ?; Q14990, ODFP_HUMAN, ?; Q61999, ODFP_MOUSE, ?;
DR Q29077, ODFP_PIG , ?; P21769, ODFP_RAT , ?;
DO PDOC00791;
//

Appendix 7: MyGPCR Database Schemas

Table= swissprot

Field	Type	Null	Key	Default	Extra
ID	varchar(10)				
AC	varchar(10)		PRI		
DE	text				
AA	int(10) unsigned	YES		NULL	
SEQ	text	YES		NULL	

Table= gpcrdb

Field	Type	Null	Key	Default	Extra
ID	varchar(10)				
AC	varchar(10)		PRI		
DE	text				

Table= gpcr_family_db

Field	Type	Null	Key	Default	Extra
ID	varchar(10)				
AC	varchar(10)		PRI		
DE	text				

Table= gpcr_family_info

Field	Type	Null	Key	Default	Extra
RID	int(11)	YES		NULL	
AC	varchar(10)	YES		NULL	
Family	int(11)	YES		NULL	

Table= pfam

Field	Type	Null	Key	Default	Extra
AC	varchar(10)		PRI		
ID	varchar(50)	YES		NULL	

Table= pfam_ls_scores

Field	Type	Null	Key	Default	Extra
Rid	int(11)		PRI	0	
ID	varchar(15)	YES		NULL	
AC	varchar(10)	YES		NULL	
Model	varchar(50)	YES		NULL	
Score	double	YES		NULL	
Evalue	double	YES		NULL	
N	int(10) unsigned	YES		NULL	

Table= pfam_ls_domains

Field	Type	Null	Key	Default	Extra
Rid	int(11)		PRI	0	
ID	varchar(15)	YES		NULL	
AC	varchar(10)	YES		NULL	
Model	varchar(50)	YES		NULL	
Domain	varchar(15)	YES		NULL	
Seq_f	int(10) unsigned	YES		NULL	
Seq_t	int(10) unsigned	YES		NULL	
Hmm_f	int(10) unsigned	YES		NULL	
Hmm_t	int(10) unsigned	YES		NULL	
Score	double	YES		NULL	
Evalue	double	YES		NULL	

Table= pfam_ls_scores

Field	Type	Null	Key	Default	Extra
Rid	int(11)		PRI	0	
ID	varchar(15)	YES		NULL	
AC	varchar(10)	YES		NULL	
Model	varchar(50)	YES		NULL	
Score	double	YES		NULL	
Evalue	double	YES		NULL	
N	int(10) unsigned	YES		NULL	

Table= pfam_ls_domains

Field	Type	Null	Key	Default	Extra
Rid	int(11)		PRI	0	
ID	varchar(15)	YES		NULL	
AC	varchar(10)	YES		NULL	
Model	varchar(50)	YES		NULL	
Domain	varchar(15)	YES		NULL	
Seq_f	int(10) unsigned	YES		NULL	
Seq_t	int(10) unsigned	YES		NULL	
Hmm_f	int(10) unsigned	YES		NULL	
Hmm_t	int(10) unsigned	YES		NULL	
Score	double	YES		NULL	
Evalue	double	YES		NULL	

Table= prints

Field	Type	Null	Key	Default	Extra
AC	varchar(10)		PRI		
ID	varchar(50)	YES		NULL	

Table= prints

Field	Type	Null	Key	Default	Extra
AC	varchar(10)		PRI		
ID	varchar(50)	YES		NULL	

Table= prints_sp_predictions

Field	Type	Null	Key	Default	Extra
Rid	int(10) unsigned		PRI	0	
AC	varchar(10)				
ID	varchar(50)	YES		NULL	
FingerPrint	varchar(80)	YES		NULL	
No_Motifs	varchar(10)	YES		NULL	
SumId	float	YES		NULL	
AveId	float	YES		NULL	
Profscore	int(11)	YES		NULL	
Ppvalue	double	YES		NULL	
Evalue	double	YES		NULL	
GraphScan	varchar(50)	YES		NULL	

Appendix 8: GPCR Specific Entries

12 GPCR specific PFAM entries:

PF00001	7tm_1	PF03094	Mlo
PF00002	7tm_2	PF01461	7tm_4
PF02793	HRM	PF02949	7tm_6
PF02076	STE3	PF02354	Latrophilin
PF00003	7tm_3	PF02101	Ocular_alb
PF01534	Frizzled	PF02116	STE2

188 GPCR specific PRINTS entries:

PR00237	GPCRRHODOPSN	PR00247	GPCRCAMP	PR00245	OLFACTORYR
PR00512	5HT1ARECEPTR	PR00248	GPCRMGR	PR00530	HISTAMINEH1R
PR01012	NRPEPTIDEYR	PR00362	CANNABINOIDR	PR01471	HISTAMINEH3R
PR01101	5HTRECEPTOR	PR00522	CANABINOID1R	PR00427	INTRLEUKIN8R
PR01103	ADRENERGICR	PR00523	CANABINOID2R	PR00572	INTRLEUKN8AR
PR00513	5HT1BRECEPTR	PR00524	CCYSTOKNINAR	PR00573	INTRLEUKN8BR
PR00514	5HT1DRECEPTR	PR00527	GASTRINR	PR01144	LSHRECEPTOR
PR00515	5HT1FRECEPTR	PR01532	CXCCHMKINER3	PR00250	GPCRSTE2
PR00516	5HT2ARECEPTR	PR00564	BURKITTSLYMR	PR00533	MASONCOGENE
PR00651	5HT2BRECEPTR	PR01106	CHEMOKINER1	PR00535	MELNOCORTINR
PR00517	5HT2CRECEPTR	PR01107	CHEMOKINER2	PR01061	MELNOCORTN3R
PR01059	5HT4RECEPTR	PR00526	FMETLEUPHER	PR01062	MELNOCORTN4R
PR00518	5HT5ARECEPTR	PR00645	LCR1ORPHANR	PR01063	MELNOCORTN5R
PR00519	5HT5BRECEPTR	PR01108	CHEMOKINER3	PR00857	MELATONINR
PR01102	5HT6RECEPTR	PR01110	CHEMOKINER5	PR01149	MELATONIN1AR
PR00652	5HT7RECEPTR	PR01529	CHEMOKINER6	PR01150	MELATONIN1CR
PR00557	ADRENRGCA1AR	PR00641	CHEMOKINER7	PR01151	MELATONIN1XR
PR00556	ADRENRGCA1BR	PR01530	CHEMOKINER8	PR00536	MELNOCYTESHR
PR00240	ADRENRGCA1DR	PR01531	CHEMOKINER9	PR00244	NEUROKININR
PR00560	ADRENRGCA2CR	PR01533	CYSLTRECPTR	PR01024	NEUROKININ1R
PR00424	ADENOSINER	PR01126	DEZORPHANR	PR01025	NEUROKININ2R
PR00552	ADENOSINEA1R	PR00568	DOPAMINED3R	PR01026	NEUROKININ3R
PR00553	ADENOSINA2AR	PR00569	DOPAMINED4R	PR00639	NEUROMEDINBR
PR00554	ADENOSINA2BR	PR00565	DOPAMINED1AR	PR01479	NEUROTENSINR
PR00555	ADENOSINEA3R	PR00566	DOPAMINED1BR	PR01480	NEUROTENSN1R
PR00243	MUSCARINICR	PR01523	S1PRECEPTOR	PR01481	NEUROTENSN2R
PR00538	MUSCRINICM1R	PR00642	EDG1ORPHANR	PR01013	NRPEPTIDEY1R
PR00540	MUSCRINICM3R	PR01148	LPARECEPTOR	PR01014	NRPEPTIDEY2R
PR00541	MUSCRINICM4R	PR01527	LPARECEPTOR	PR01018	GPR10RECEPTR
PR00542	MUSCRINICM5R	PR01524	EDG3RECEPTOR	PR01015	NRPEPTIDEY4R
PR00520	ACTROPHINR	PR01128	EMR1HORMONER	PR01016	NRPEPTIDEY5R
PR00534	MCRFAMILY	PR00366	ENDOTHELINR	PR01017	NRPEPTIDEY6R
PR00241	ANGIOTENSINR	PR00570	ENDOTHELINAR	PR01528	EDG4RECEPTOR
PR00636	ANGIOTENSN2R	PR00571	ENDOTHELINBR	PR01535	VOMERONASL2R
PR00635	ANGIOTENSN1R	PR00489	FRIZZLED	PR00238	OPSIN
PR00561	ADRENRGCB1AR	PR00373	GLYCHORMONER	PR01477	LTB1RECEPTOR
PR00242	DOPAMINER	PR01143	FSHRECEPTOR	PR01476	LTBRECEPTOR
PR00563	ADRENRGCB3AR	PR01176	GABABRECEPTR	PR01525	EDG5RECEPTOR
PR00899	GPCRSTE3	PR01177	GABAB1RECPTR	PR01526	EDG6RECEPTOR
PR00901	PHEROMONEBAR	PR01178	GABAB2RECPTR	PR00965	OCULARALBNSM

PR00425	BRADYKININR	PR01146	GPR1ORPHANR	PR00664	OCTOPAMINER
PR00994	BRADYKINNB2R	PR00644	GPRORPHANR	PR00384	OPIOIDR
PR00358	BOMBESINR	PR00648	GPR3ORPHANR	PR00525	DELTAOPIOIDR
PR00637	BOMBESIN3R	PR01147	GPR4ORPHANR	PR00532	KAPPAOPIOIDR
PR01060	C3ANPHYLTXNR	PR00649	GPR6ORPHANR	PR00246	SOMATOSTATNR
PR01104	ANPHYLATOXNR	PR00650	GPR12ORPHANR	PR00537	MUOPIOIDR
PR00657	CCCHEMOKINER	PR01507	MCH1RECEPTOR	PR00547	XOPIOIDR
PR00249	GPCRSECRETIN	PR00640	GASTRINRELPR	PR00576	OPSINRH1RH2
PR00361	CALCITONINR	PR00529	GNADOTRPHINR	PR00578	OPSINLTRLEYE

11 GPCR specific PROSITE entries:

PS00237	G_PROTEIN_RECEP_F1_1
PS50262	G_PROTEIN_RECEP_F1_2
PS00649	G_PROTEIN_RECEP_F2_1
PS00650	G_PROTEIN_RECEP_F2_2
PS50227	G_PROTEIN_RECEP_F2_3
PS50261	G_PROTEIN_RECEP_F2_4
PS00979	G_PROTEIN_RECEP_F3_1
PS00980	G_PROTEIN_RECEP_F3_2
PS00981	G_PROTEIN_RECEP_F3_3
PS50259	G_PROTEIN_RECEP_F3_4
PS00238	OPSIN