



Introduction

Only approximately 51% of the human proteome can be annotated by the standard motif-based recognition systems [1]. These systems, currently aggregated into a single distributed system by InterPro [2], include PFAM, PRINTS, ProSite, ProDom, SMART, and SWIS-PROT+TrEMBL. PFAM consists of hidden Markov models based on hand-curated alignments of protein domains. PRINTS is a repository of protein fingerprints, that is, groups of conserved motifs. ProSite is a database of protein families and domains based on common structural motifs. SWIS-PROT is a highly curated and annotated protein sequence database, which is combined with the a computer annotated supplement, TrEMBL. Our goal is to provide a wider base of annotation systems in order to expand the ability to annotate the human proteome.

Within NetAffx, we are providing standard PFAM and BLOCKS motif annotations and BLASTp similarity searches. In addition, we are providing annotations based on high-level, protein-based classification schemes. The goal is to provide protein family-level annotations, which will allow the user to rapidly classify probe sets of interest based on structure and function, giving more insight into a protein's role within the context of the cell. Our approach allows us to efficiently adapt and incorporate the growing base of knowledge from various classification systems into our annotation effort. The current release includes annotations based on the following: 1. Structural Classification of Protein (SCOP) classifies proteins based on structure, function, and phylogenetic aspects; 2. Enzyme Commission numbers (EC) classify enzymes based on type of reaction catalyzed, substrate and reagent/co-factors; and 3. G protein coupled receptor families (GPCR) classifies 7-transmembrane receptors based on SWIS-PROT defined family definitions.

Families_GPCR Methods

Families_GPCR contains alignments to families of G protein coupled receptors as organized by SWIS-PROT and further sub-divided based on the following method. The GPCR classification list may be found at <http://www.expasy.ch/cgi-bin/lists?7tmrlist.txt>. Alignments within a family are generated by scoring against SAM-T99 derived HMM models, as described for the FAMILY_SCOP database. For low-quality models (containing excessive gaps) we decomposed the alignments into subfamilies using BETE [3]. Given a sequence alignment, BETE begins with each sequence in a separate subfamily, and merges similar subfamilies to minimize the overall encoding cost of the alignment. This encoding cost is defined according to the alignment column entropy summed over each subfamily and column, and relative to the number of subfamilies, as follows:

$$E = N \log_2(S) - \sum_s \sum_c \log_2(P(n_{\{c,s\}}|\theta))$$

Here, c represents each alignment column, s represents each subfamily, S represents the total number of subfamilies, and θ is the subfamily decomposition. $n_{\{c,s\}}$ is the vector of amino acid counts observed in

column c and subfamily s , and $P(n_{\{c,s\}})$ is the posterior probability of $n_{\{c,s\}}$ computed according to Dirichlet mixtures [4]. Finally, N is the total number of amino acids in the alignment. The intuition behind this method is as follows. If the subfamilies contain distantly-related sequences, then the $P(n_{\{c,s\}})$ terms will be small, resulting in a high encoding cost. If the subfamilies are partitioned too finely, then the $N \log_2(S)$ term will be large, resulting in a high encoding cost. The two halves of the equation balance each other to cluster together the sequences with the most significant relation. This yields both a subfamily decomposition, and a phylogenetic tree describing the order in which the subfamilies were merged. We used the PHYLODENDRON [5] tool to generate visualizations of these trees. We used these visualizations to identify the major sequence clusters, and partitioned the alignment accordingly.

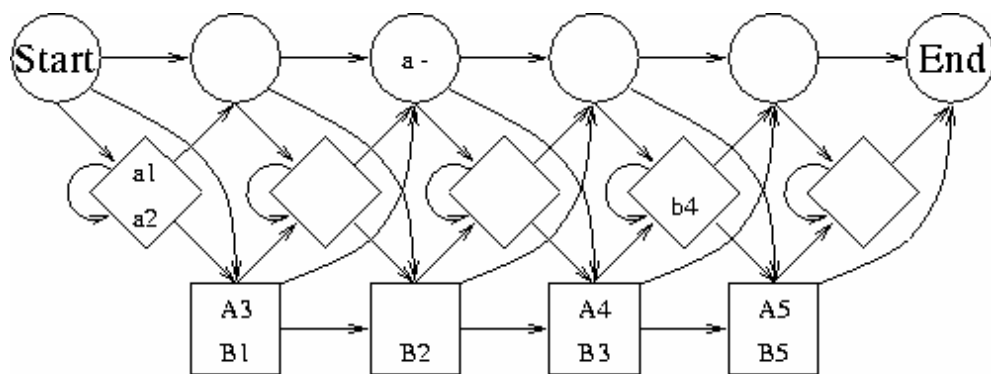
New alignments within each family are generated by scoring against SAM-T99 derived HMM models, as described for the FAMILY_SCOP database. High level description, including sequences used to build the models and review articles for each family are included in the Families_GPCR database within SRS. The models have been shown to be able to distinguish between the families in Table 1.

Family name	number of subfamilies
Acetylcholine (muscarinic) receptors	2
Adrenergic (alpha) receptors	1
Adrenergic (beta) receptors	1
Adrenomedullin receptors	1
Angiotensin receptors	1
Bombesin receptors	1
Bradykinin receptors	1
Cannabinoids receptors	1
Chemokines and chemotactic factors receptors	7 (CC, complement, CX3C, DARC, FMLP, XC)
Cholecystokinin / gastrin receptors	1
Dopamine receptors	2
EDG receptors	1
Endothelin receptors	1
Family 2B receptors	10 (calcitonin, corticotropin, diuretic, GHRH, GIP, glucagon, parathyroid, pituitary, secretin, VIP)
Family 3C receptors	3 (calcium-sensing, GABA, metabotropic-glutamate)
Frizzled/smoothened receptors	1
Galanin receptors	1
Glycoprotein hormones receptors	1
Histamine receptors	1
MAS proto-oncogene receptors	1
Melanocortins receptors	1

Melatonin receptors	2 (other, ML1X)
Neuropeptide Y receptors	2 (FF, Y)
Neurotensin receptors	1
nucleotide-like adenosine	1
nucleotide-like purinoceptors	1
octopamine	1
Odorant/olfactory and gustatory receptors	1
Opioid peptides receptors	1
Opsins	1
Orexins receptors	1
Platelet activating factor receptors	1
Prostanoids receptors	4 (prostacyclin, prostaglandin DE, prostaglandin F, thromboxane)
Proteinase-activated receptors	1
Releasing hormones receptors	3 (GHS, gonadotropin, and thyrotropin)
Serotonin receptors	1
Somatostatin and urotensin receptors	1
Tachykinin receptors	1
Vasopressin / oxytocin receptors	1
Vomer nasal (mouse and rat only to date)	1

Table 1. G protein receptor families as classified by SWIS-PROT seven transmembrane page.

Previous work by Karchin [6] has shown that SAM-T99 HMMs can accurately distinguish between protein families of GPCRs. While other methods including Fisher kernel support vector machines built on HMMs can more accurately distinguish among families, the increased computation to train SVMs makes their application to large numbers of models prohibitive.



```

a1  a2  A3      -      A4  .  A5
.   .  B1      B2      B3  b4 B5

```

Figure 1 represents a linear hidden Markov model consisting of nodes which include match (square), insert (diamond), and delete (circle) states

<http://www.cse.ucsc.edu/research/complib/sam.html>].

For each GPCR family, a hidden Markov model was built using the Sequence Alignment and Modeling system [7] (SAM 3.2.1) target99 script. Using the SAM-T99 perl script, HMMs were built for 75 GPCR families as defined by SWIS-PROT 7tmrlist. By default, this script iteratively uses BLAST to find sequences related to the seed GPCR family alignment in the non-redundant (nr) Genbank protein sequence database. However, for the GPCR classes, the BLAST runs were suppressed and no database searching was performed. These sequences were aligned and a hidden Markov model extracted. The resulting HMMs consist of a series of nodes, one for each column in the multiple alignment. In SAM's HMMs, each node consists of a match state, an insert state, and a delete state. Each node and transition has a distribution of values; thus, these linear HMMs have position-dependent character distributions and position-dependent insertion and deletion gap penalties. The HMMs are only scored based on the length of seed sequence; therefore, biasing the models towards domain recognition of the appropriate length. These models are sensitive to the characterization of a protein superfamily, and the alignment of a sequence against a trained model will automatically yield a multiple alignment to those sequences.

Basic parameter settings to control the generality or specificity of each model include the threshold values for admitting new homologs at each iteration, the maximum number of homologs to admit during the modeling process, and selection of the sequence weighting scheme. SAM-T99 default threshold values are 0.001, 0.02, 0.2, and 1.0 at each of the four BLAST iterations against the nr database. These values have been tuned for superfamily recognition. Because we are suppressing BLAST searching of nr, we have chosen to use the "-no_search -family -aweight_bits 0.5" parameters.

In an effort to validate the models, from each of the acetylcholine, adenosine, adrenergic, family2B and odorant receptor sets one representative sequence was omitted. Then all GPCR sequences found in the SWIS-PROT list were scored against the models, including the known sequences omitted from the models.

In all cases, the omitted GPCRs scored higher in their respective families than in any other model for the 74 other GPCR categories. Karchin, in a thorough set of experiments, has shown the average percent correct at the minimum error point to be 97.90% for SAM-T99.

Whole genome gene set

A set of protein sequences covering the Golden Path of the human genome (October 7, 2000 freeze <http://genome.ucsc.edu/>) was generated by the Genie [8] programs suite [Kulp, D. & Wheeler, R., published in <http://genome.ucsc.edu/>], with the repeat regions masked out. The data set consists of three sets of amino acid sequences: (1) proteins from Genbank whose associated mRNA sequences could be mapped to genomic sequence (2) proteins predicted by *AltGenie*, an enhanced Genie program which predicts alternatively splice transcripts using merged mRNA/EST-to-genomic alignments and (3) proteins predicted by *StatGenie*, a purely *ab initio* gene-finder. These sequences are non-redundant; none of the included genes overlap the same genomic region. In cases where there were many genes overlapping the same region, the one with the longest CDS (translation) was kept [Williams, A., unpublished data]. This set, known as annot10, contains 59,378 protein sequences.

The non-redundant complete proteome set of SWISS-PROT plus TrEMBL entries for *Drosophila melanogaster* (13844 entries), *Caenorhabditis elegans* (18870 entries), and *Saccharomyces cerevisiae* (6186 entries) were obtained on June 15, 2001 from the EBI proteome analysis site (<http://www.ebi.ac.uk/proteome/>).

Screening genes against GPCR HMMs

The entire set of protein sequences is screened against all the GPCR HMMs, such that each sequence is given a distance score. The distance score consists of the negative log-likelihood minus NULL model scores for each sequence against a given model [9]. An E-value is calculated based on the size of the database and the reverse-sequence score.

The SCOP annotation pipeline has also been adapted to a panel of HMMs which identify family members as defined by the GPCR database. For the final scoring run, all known GPCR proteins (controls) were included in the screening set to give the optimal distance E-value scores. The relationship between a GPCR family and a gene was determined by the E-value scored by SAM3.2.1 hmmscore, with the following parameters "-select_score 4 -Emax 0.01 -sw 2 -select_mdalign." The effect is to record all hits with an E-value threshold of 0.01. A significant difference in this release is that hmmscore allows multiple hits of the same HMM against the query protein sequence.

The GRAPA method [10, 11] was applied to the all scored sequences to refine the EC family assignments. The program distsieve examines all the sequences scored against a particular SCOP model. Because the interpretation of distance scores, expressed as logarithmic E-values, is dependent upon the HMM generated for each sequence, the distance score files from each of the HMMs is edited by curve analysis to determine an E-value cutoff. The gene set includes controls representing the proteins used to build all the models; thus,

the best hit is often the seed sequence for the HMM. One can look for trends in the E-values based on this high-scoring control. Four criteria are used to distinguish the hits to be kept: (1) the hits within 20% of the maximum form an upper limit cutoff; (2) beyond the 20%, the region where the E-values flatten out is used as a cutoff; (3) when the E-values rise above a value of e^{-05} , a cutoff is imposed; (4) no more than a predetermined maximum of hits (e.g., 500) are kept.

The set of HMMs are then grouped into Superfamily sets. Since SCOP is a hierarchical classification scheme with a tree structure, a set of hits from the HMMs representing families may be grouped by Superfamilies. From these Superfamily groupings, hits are assigned to a single Family HMM within each Superfamily. Next, several lists are collated to create an alignment between each successful hit and the corresponding SCOP sequence, using the SCOP sequence's HMM for the alignment. Finally, alignments are screened for identity scores and annotations created in XML database format.

References

1. Lander, E.S., et al., *Initial sequencing and analysis of the human genome*. Nature, 2001. **409**(6822): p. 860-921.
2. Apweiler, R., et al., *The InterPro database, an integrated documentation resource for protein families, domains and functional sites*. Nucleic Acids Res, 2001. **29**(1): p. 37-40.
3. Sjolander, K., *Phylogenetic inference in protein superfamilies: analysis of SH2 domains*. Proc Int Conf Intell Syst Mol Biol, 1998. **6**: p. 165-74.
4. Sjolander, K., K. Karplus, M. Brown, R. Hughey, A. Krogh, I.S. Mian, and D. Haussler, *Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology*. Comput Appl Biosci, 1996. **12**(4): p. 327-45.
5. Gilbert, D.G., *Phylo dendron - Phylogenic tree printer*. 1999.
6. Karchin, R., K. Karplus, and D. Haussler, *Classifying G-protein Coupled Receptors with Support Vector Machines*. 2001.
7. Karplus, K., C. Barrett, and R. Hughey, *Hidden Markov models for detecting remote protein homologies*. Bioinformatics, 1998. **14**(10): p. 846-56.
8. Reese, M.G., D. Kulp, H. Tammana, and D. Haussler, *Genie--gene finding in Drosophila melanogaster*. Genome Res, 2000. **10**(4): p. 529-38.
9. Hughey, R. and A. Krogh, *Hidden Markov models for sequence analysis: extension and analysis of the basic method*. Comput Appl Biosci, 1996. **12**(2): p. 95-107.
10. Shigeta, R., G. Liu, M. Cline, A. Loraine, D. Kulp, and M.A. Siani-Rose, *Generalized Rapid Automated Protein Analysis (GRAPA): annotating the human genome based on SCOP domain-derived hidden Markov models*. submitted, 2001.
11. Shigeta, R., M.A. Siani-Rose, and D. Kulp, *RAKE: Accurate Automated Annotation of the Human Genome Based on SCOP Domain-derived Hidden Markov Models*, in *Currents in Computational Molecular Biology 2001*. 2001. p. 247-248.