Subfamily HMMS in Functional Genomics

D. Brown, N. Krishnamurthy, J.M. Dale, W. Christopher, and K. Sjölander

Pacific Symposium on Biocomputing 10:322-333(2005)

SUBFAMILY HMMS IN FUNCTIONAL GENOMICS

DUNCAN BROWN, NANDINI KRISHNAMURTHY, JOSEPH M. DALE, WAYNE CHRISTOPHER AND KIMMEN SJÖLANDER

Department of Bioengineering, University of California
Berkeley, CA 94720

The limitations of homology-based methods for prediction of protein molecular function are well known; differences in domain structure, gene duplication events and errors in existing database annotations complicate this process. In this paper we present a method to detect and model protein subfamilies, which can be used in high-throughput, genome-scale phylogenomic inference of protein function. We demonstrate the method on a set of nine PFAM families, and show that subfamily HMMs provide greater separation of homologs and non-homologs than is possible with a single HMM for each family. We also show that subfamily HMMs can be used for functional classification with a very low expected error rate. The BETE method for identifying functional subfamilies is illustrated on a set of serotonin receptors.

1 Introduction

The vast majority of proteins have no experimentally determined function, and prediction of molecular function by homology with functionally characterized proteins has become *status quo*. Such predictions are used to obtain a preliminary functional annotation, and thereby to guide wetbench experiments. However, all homology-based methods of function prediction are known to be prone to systematic errors of various types[1-4]. For a variety of reasons, including domain fusion, gene duplication, and the undeniable presence of existing database errors, inferring molecular function based on the annotated function of the top hit in database search is fraught with potential hazards. Profile and hidden Markov model (HMM) methods perform admirably in detecting homologous proteins[5], but these generally afford only a very high level of functional classification.

Phylogenomic analysis of a protein in the context of its entire family has been demonstrated to improve both the accuracy and specificity of functional annotation[2, 3, 6], but is time-consuming and not easily automated, and therefore is generally applied to single families rather than at the genomic level.

We present here a method for obtaining a classification of sequences to functional subfamilies that was used at Celera Genomics in the functional classification of the human genome[7]. Subfamily HMMs model the functional and structural variants of a protein family, so that regions of structural diversity across the family are described by subfamily-specific amino acid preferences.

The ability to assign sequences to subfamilies automatically enables the high-throughput application of phylogenomic inference of protein molecular function which might otherwise be infeasible. If subfamily HMMs are

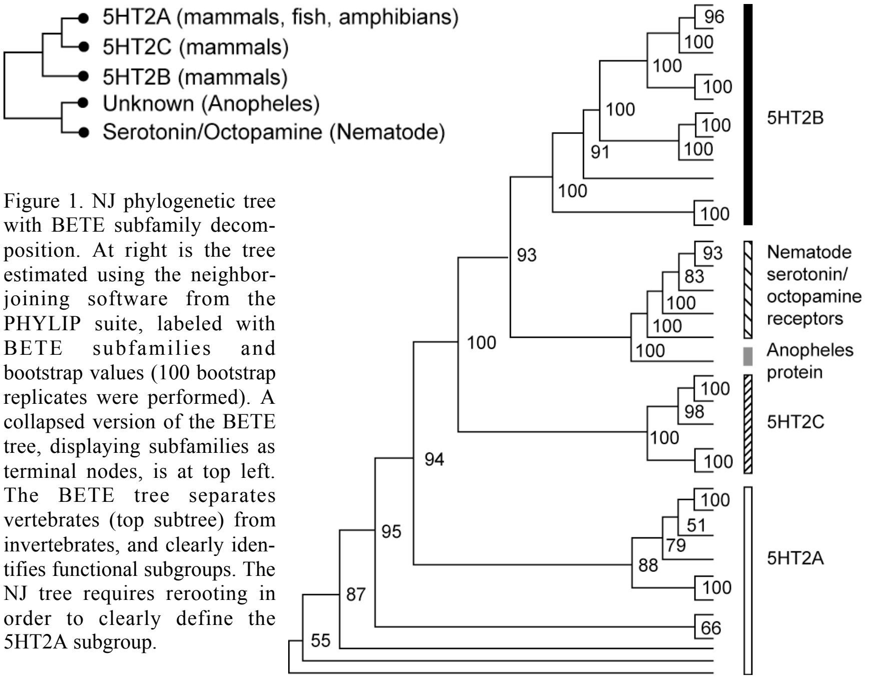

Figure 1. NJ phylogenetic tree with BETE subfamily decomposition. At right is the tree estimated using the neighbor-joining software from the PHYLIP suite, labeled with BETE subfamilies and bootstrap values (100 bootstrap replicates were performed). A collapsed version of the BETE tree, displaying subfamilies as terminal nodes, is at top left. The BETE tree separates vertebrates (top subtree) from invertebrates, and clearly identifies functional subgroups. The NJ tree requires rerooting in order to clearly define the 5HT2A subgroup.

constructed for the family, scores of sequences against subfamily HMMs can indicate a preliminary phylogenetic classification of a sequence, together with a more precise prediction of function.

The remainder of the paper is organized as follows. An illustration of the BETE subfamily decomposition is presented in section 2. Section 3 describes our method of constructing subfamily HMMs, and section 4 shows experimental results comparing the use of subfamily HMMs with single, family-level HMMs on several tasks: training sequence detection, remote homolog detection and classification accuracy. Discussion and future work are described in section 5.

2 BETE Subfamily Decomposition

In these experiments, we obtain a subfamily decomposition using Bayesian Evolutionary Tree Estimation (BETE)[8]. BETE estimates a phylogenetic tree using agglomerative clustering; subtrees are represented by profiles constructed using Dirichlet mixture densities[9] and symmetrized relative entropy is used as a distance metric between subtrees. Subfamilies are determined by a minimum-description-length cut of the tree into subtrees[8].

As presented elsewhere on *Src homology 2* (SH2) domains[8] and in the functional characterization of the proteins encoded in the human genome[7], the BETE subfamily decomposition corresponds closely to experimental data on protein function and structure. We present in this section an illustration of the subfamily classification enabled by BETE, in application to the serotonin-receptor-related family of G-protein-coupled receptors. G-protein-coupled receptors are of enormous biomedical importance and include many pharmaceutical targets. Subfamily classification is particularly valuable in the context of this group due to the number of orphan receptors with unknown ligand specificity[10].

For this example, sequence homologs to serotonin receptor type 2B from human (SwissProt accession P41595) were gathered from the NR database using the FlowerPower program (in preparation). The homologs were aligned using MUSCLE[11], and trees were constructed using BETE and neighbor-joining (NJ, from the PHYLIP suite[12]). See Figure 1.

3 Subfamily HMM Construction Method

The method requires as input a multiple sequence alignment of a set of related proteins, with a specified decomposition into subfamilies. We use the same HMM architecture for each subfamily HMM (SHMM); a general HMM (GHMM) is constructed for the family as a whole and SHMMs are created by replacing the GHMM match state amino acid distribution at each position with a subfamily-specific distribution.

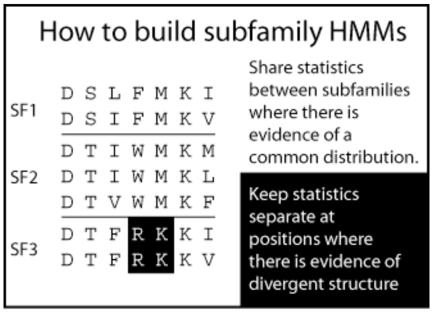
How to build subfamily HMMs
SF1
D S L F M K I
D S I F M K V
SF2
D T I W M K M
D T I W M K L
D T V W M K F
SF3
D T F R K K I
D T F R K K V
Share statistics between subfamilies where there is evidence of a common distribution.
Keep statistics separate at positions where there is evidence of divergent structure

Figure 2. How to build subfamily HMMs. Amino acid distributions for positions defining the family as a whole are estimated once, and fixed within each subfamily. For non-globally conserved positions, examine the amino acids aligned by each of the other subfamilies. If a subfamily aligns similar amino acids, share statistics. Otherwise, keep statistics separate. In this toy example, the first two subfamilies will share statistics throughout the alignment. The last subfamily will share statistics with the first two (and vice-versa) at the black-on-white positions, but not at the white-on-black positions.

3.1 Estimating Subfamily Amino Acid Distributions

For each subfamily *s*, and at each column *c* in the alignment, we compute a distribution over amino acids. We first discuss two special cases.

The first special case involves family-defining positions. Amino acid distributions at positions conserving the same amino acid across all sequences (allowing gaps) are fixed for all subfamilies. This enables subfamilies containing very few sequences to share in the knowledge of the critical residues (which might otherwise become generalized to allow substitutions). We first estimate the number of independent observations in the family as a whole; this number is used to weight the observed amino acids in deriving the posterior estimate of the amino acid distribution using Dirichlet mixture densities. The second case involves handling gapped positions. In these cases, the amino acid distribution is copied from the general HMM for that position.

General case: To allow us to recognize related family members, but still maintain specificity for each individual subfamily, we combine the amino acids in subfamily s at column c with amino acids from subfamilies aligning similar amino acids to those in subfamily s at position c. This is illustrated in Figure 2.

Sequence weighting: In common usage, sequence weighting is often restricted to deriving relative weights for a set of sequences, to down-weight sequences in highly populated subgroups and up-weight subgroups with few sequences. However, in using Dirichlet mixture densities to estimate amino acid distributions, the magnitude of the counts is also critical. In deriving subfamily HMM match state distributions, our approach involves estimating the number of independent observations in the alignment. We compute for every position in the alignment the frequency of the most frequent amino acid (ignoring gap characters) to derive the positional conservation propensity, and then compute the average conservation propensity over all columns (P_{cons}). The number of independent counts (NIC) can then be defined as $\text{NIC} = N^{1-P_{\text{cons}}}$, where N is the number of sequences in the alignment. This has the effect of producing an NIC of 1 when the sequences in the alignment are 100% identical, and having NIC approach N as the diversity in the alignment increases. The relative weights can then be derived independently. In the following equations, the notation $\vec{n} = (n_1, n_2, ..., n_{20})$ refers to the *weighted* counts of the amino acids seen at column c in subfamily s, and n_i represents the weighted count of amino acid i at column c in subfamily s. The amino acid distribution at that position for subfamily s is estimated as follows:

Step 1: Obtain a Dirichlet mixture density posterior.

We obtain a full Dirichlet mixture posterior density Θ^{Post} by combining the Dirichlet mixture prior Θ^{Prior} with the observed (weighted) amino acids seen in the column[9]. The mixture coefficients q_j, denoting the prior probability of each component j and the component density parameters $\vec{\alpha}_j$ of the Dirichlet mixture Θ^{Post} are set as follows:

$$\begin{gathered} q_j = \Pr(\vec{\alpha}_j \mid \vec{n}, \Theta^{\text{Prior}}) \\ \alpha_{ji} = \alpha_{ji} + n_i \end{gathered} \qquad (1)$$

Step 2: Compute the family contribution from subfamilies $s' \neq s$.

When we compute the contribution from other subfamilies to the profile for subfamily s at a fixed position, we add in amino acids from each subfamily proportional to the probability of the amino acids aligned by each subfamily at that position. Letting $\vec{n}_{s'}$ be the amino acids aligned at that column by subfamily s', the "family contribution" is summed over all the subfamilies $s' \neq s$, creating a vector of amino acids $\vec{f}$, as follows:

$$\vec{f} = \sum_{s' \neq s} \Pr(\vec{n}_{s'} \mid \Theta^{\text{Post}}) \vec{n}_{s'} \tag{2}$$

In this equation, $\Pr(\vec{n}_{s'} \mid \Theta^{\text{Post}})$ represents the posterior probability of $\vec{n}_{s'}$ given the posterior Dirichlet mixture density for subfamily s at that position. In practice, we need to prevent the other subfamilies' contributions from swamping the amino acids observed in subfamily *s*; this is accomplished by capping the total $|\vec{f}|$ to a user-specified maximum

Step 3: Combine the family contribution with the counts in subfamily s, to obtain the total counts $\vec{t} = (t_1, t_2, \ldots t_{20})$, *where* $t_i = n_{si} + f_i$.

Step 4: Estimate the posterior amino acid distribution using Dirichlet mixture priors. We modify the normal method for estimating the probability of amino acid *i* at a position by substituting t_i for n_i and $|\vec{t}|$ for $|\vec{n}|$:

$$\hat{p}_i \propto \sum_j \Pr(\vec{\alpha}_j \mid \vec{n}, \Theta^{\text{Prior}}) \frac{t_i + \alpha_{ji}}{|\vec{t}| + |\vec{\alpha}_j|} \tag{3}$$

Thus, we estimate the posterior probability of each component of the Dirichlet mixture prior using both the observed subfamily counts, and counts from all other subfamilies.

4 Experimental Validation

4.1 Data Chosen for Experiments

Subfamily HMMs are expected to contribute the most towards improved homolog detection when constructed for large and diverse protein families, and we selected a limited set of protein families with these characteristics at the outset. We chose entries from the list of PFAM[13] families beginning with letters A-C, based on the following criteria: (1) Each family had to have at least one member whose structure had been solved, and the alignment had to be at least 80 residues in length. This ensured that the selected family corresponded to a structural domain and was not simply a short repeat. (2) In order to provide informative comparisons between remote-homolog detection methods, the family-level HMM had to detect at least 10 homologs in the Astral PDB90 dataset of protein structural domains, and each family had to belong to a different SCOP superfamily[14]. (3) Finally, to ensure enough diversity in the family, the PFAM full alignment had to contain at least 600 sequences and have < 30% average pairwise identity (alignments with more than 3000 sequences after being made non-redundant at 95% identity were excluded). Details of PFAM families used in these experiments are provided in Table 1.

Table 1. PFAM family data. *SCOP* is the SCOP superfamily ID for the PFAM family; *# sequences, Gaps* and *Average %ID* refer to the number of sequences, fraction of gaps and average percent identity, respectively, within the UG95 MSA. *# subfamilies* is the number of subfamilies found by BETE for that family. *Subfam BL62* is the average per-position BLOSUM62 score in each subfamily MSA. *Full BL62* is the average per-position BLOSUM62 score in the UG95 MSA. *Full BL62 ≥ 1* is the fraction of columns in the UG95 MSA with average BLOSUM62 scores ≥ 1. *TS GHMM* and *TS SHMM* give the fraction of training sequences detected by the GHMM and SHMM within an E-value cutoff of 1e-10. *PDB GHMM* and *PDB SHMM* give the fraction of PDB90 homologs detected by the GHMM and SHMM within an E-value cutoff of 1e-10. *Class. Acc.* is the fraction of sequences correctly classified in the leave-one-out experiments.

PFAM Family	SCOP	# sequences	Gaps	Average %ID	# subfamilies	Subfam BL62	Full BL62	Full BL62 ≥ 1	TS GHMM	TS SHMM	PDB GHMM	PDB SHMM	Class. Acc.
AAA	c.37.1	1573	0.07	22	238	2.65	0.79	0.32	0.55	0.90	0.04	0.06	0.99
Cadherin	b.1.6	2002	0.05	23	704	2.77	0.85	0.36	0.69	0.82	0.60	0.90	0.99
Cytochrome C	a.3.1	725	0.14	17	245	2.57	0.58	0.18	0.12	0.68	0.10	0.54	0.93
Alpha_amylase	c.1.8	874	0.13	17	184	2.48	0.52	0.21	0.98	0.99	0.34	0.34	0.98
Aminotran_1_2	c.67.1	1250	0.06	16	316	1.96	0.23	0.13	0.91	0.95	0.39	0.39	0.96
C2	b.7.1	865	0.06	21	263	2.69	0.77	0.33	0.38	0.88	0.69	0.77	1.00
Aldo_ket_red	c.1.7	755	0.12	22	236	2.64	0.94	0.35	0.94	0.96	1.00	1.00	0.97
Abhydrolase_1	c.69.1	1209	0.06	14	687	2.57	0.02	0.13	0.41	0.84	0.18	0.27	0.99
Amidohydro_1	c.1.9	626	0.21	11	221	2.60	0.05	0.10	0.72	0.82	0.50	0.65	0.96

Figure 3 shows average per-position BLOSUM62 scores plotted for selected families, displaying both the high family diversity and within-subfamily conservation.

4.2 Technical Details

These experiments used the UCSC SAM software[15] for scoring and aligning proteins and to construct the general HMM. Several preprocessing steps were performed on the input sequences. The PFAM full alignment for each family was made non-redundant at 95% identity to create the NR95 multiple sequence alignment (MSA). We removed all columns having > 70% gaps and then all sequences matching fewer than 70% of the PFAM HMM match states, to give final un-gapped alignments (UG95). The UG95 MSA was used to derive a general HMM using the SAM w0.5 tool. The BETE algorithm[8] was used to identify subfamilies and SHMMs were constructed as described in Section 3.

SAM reverse scores were obtained using local-local scoring throughout. Sequences were assigned a GHMM score and a SHMM score (the best of the scores against all the SHMMs for the family). Subfamily HMM E-values were computed with respect to an extreme-value distribution fitted to SHMM scores against random sequences, as in the HMMER package. To ensure that E-values were comparable across families, an assumed database size of 100,000 (the

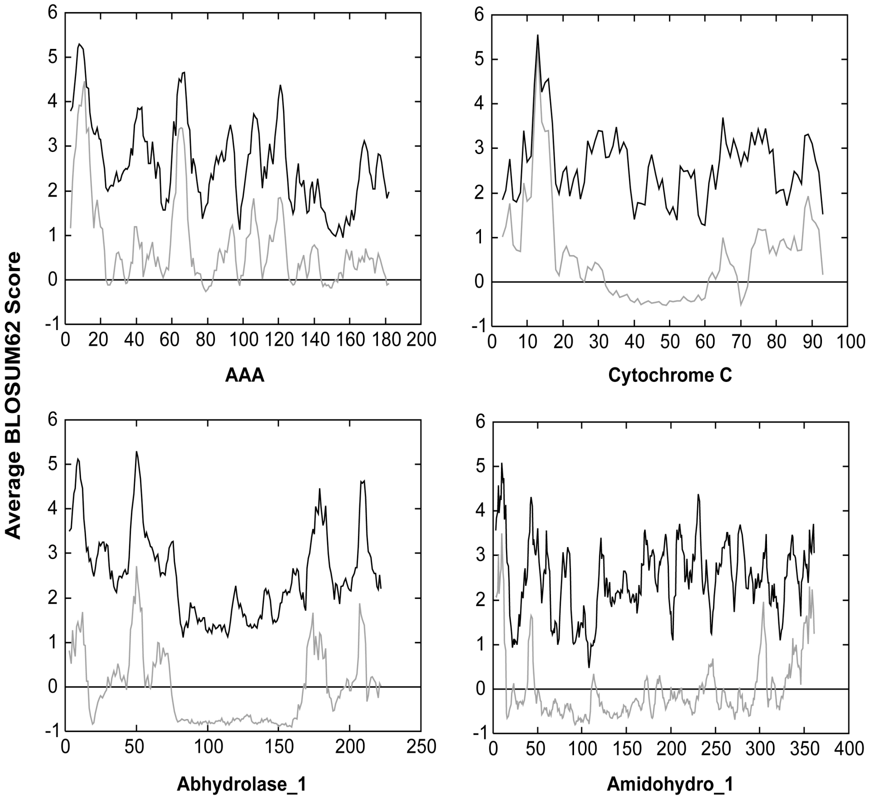

Figure 3. Average per-column pairwise BLOSUM62 scores for selected families. The X-axis shows the alignment column, and the Y-axis shows the average pairwise BLOSUM62 score for amino acids in that column. Within-subfamily scores are in black; scores across the whole family are in grey. Alignment columns with low average BLOSUM62 produce noisy HMM match-state distributions. Alignment quality of novel sequences to HMMs constructed from alignments with these characteristics can be correspondingly poor in regions of high structural variability across the family as a whole.

approximate size of SWISSPROT release 40) was used for training sequence detection experiments. For remote homolog detection experiments, we scored against Astral PDB90 release 1.65, and we used the true database size (8888).

4.3 Training Sequence Detection

Unaligned training sequences were extracted from the NR95 MSA and scored against the GHMM and SHMMs for that family. Figure 4 shows sequence coverage versus E-value summed over all nine PFAM families for both methods (results using an E-value cutoff of 1e-10 are summarized in Table 1). P-values were computed using the Wilcoxon signed-rank test to determine the significance of the differences between methods at different E-value cutoffs.

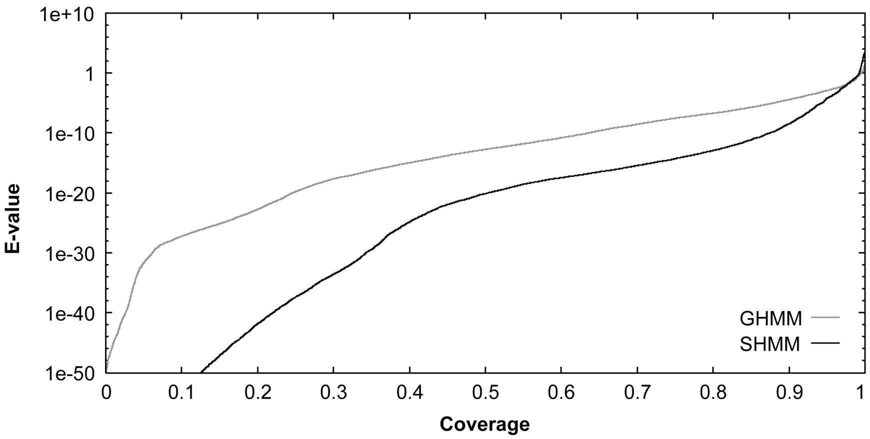

Figure 4. Training sequence detection using General and Subfamily HMMs. Shown here are results scoring training sequences against the general and subfamily HMMs using local-local scoring. The Y-axis gives the E-values for sequences against the different HMMs. The X-axis shows the fraction of sequences across all the families that scored at that level or better.

The SHMM method detects significantly more sequences than the GHMM at E-values below 1e-10, with a P-value of 0.002.

4.4 Remote Homolog Detection Tests

In these experiments, we compared subfamily and general HMMs on the ability to discriminate between homologs and non-homologs, using the Astral PDB90 database of structural domains[16], as classified by the Structural Classification of Proteins (SCOP) database[17]. The Astral PDB90 dataset is a subset of protein domains chosen so that no two are more than 90% identical when aligned. The Astral datasets have been widely used by the computational biology community to assess homology detection methods[5, 18, 19].

For computational efficiency, we first scored PDB90 with the general HMM for the family, and retrieved all sequences with E-values less than 100. These sequences were then scored against the subfamily HMMs, and the results were combined with the remaining GHMM scores. Preliminary results comparing this method with all-*vs*-all scoring of sequences against SHMMs indicated that there was little difference between the two. A more complete comparison is in preparation.

Each of the matches was marked as either True (classified to the same SCOP superfamily), False (classified to different SCOP folds) or Indeterminate (in the same SCOP fold but different SCOP superfamilies). We calculated normalized coverage and errors per query (EPQ) as described[20]. Results for each method were combined, sorted by e-value and assessed beginning with the most significant score. True positives were weighted such that each superfamily

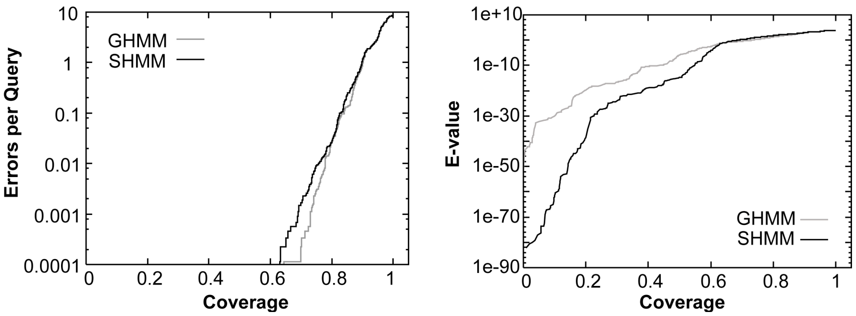

Figure 5. PDB90 discrimination experiments, comparing subfamily and general HMMs. For both methods, false positives appear at E-values of 0.1 and above, and both methods obtain similar coverage at this cutoff, indicating that SHMMs do not identify more homologs, but do afford a better separation between identified true positives and rejected sequences. SHMMs give stronger scores to clear homologs and reject non-homologs with large E-values.

contributed equally to the total coverage. The EPQ was calculated as the cumulative number of false positives divided by the total number of sequences in the database (8888 sequences for Astral PDB90 release 1.65).

In Figure 5, we show two plots to describe the combined results over all nine PFAM families: a standard Coverage *vs*. Error plot, and Coverage *vs*. E-value. Consistent with the results on training sequence detection, subfamily HMMs provide stronger scores to clearly homologous sequences, but general HMMs provide a small improvement over subfamily HMMs at detecting distant homologs. This improvement occurs at weak E-values of 0.1 and higher.

4.5 Classification Experiments

Homology-based functional annotation is at its heart a classification problem. Given a protein, even one for which the general family is known, determination of the functional subgroup to which it belongs is not a trivial task. As described in Section 2, the BETE subfamily decomposition correlates highly with subtypes already identified by biologists, and it is natural to use this breakdown to give a more precise prediction of protein molecular function. With this application in mind, we examined subfamily HMM performance in classifying previously unseen sequences. Note that although we used the BETE subfamily decomposition, the SHMM construction algorithm is independent of any particular sequence grouping; subfamilies identified by other methods (for example, manual annotation based on experimentally determined function) may also be used as input.

For each family, we took the BETE subfamily decomposition based on the UG95 MSA and removed one sequence at random. We required only that the subfamily to which the sequence belonged contained at least two sequences. Since the algorithm for constructing subfamily HMMs shares information across all subfamilies, we then rebuilt *all* of the subfamily HMMs using the modified alignment as input, but keeping the subfamily decomposition unchanged. We assumed that the *family* identification of a test sequence was given; withheld sequences were scored against the SHMMs from their family, and the top-scoring SHMM was identified. A sequence whose top-scoring subfamily was the one from which it had been extracted was counted as a success; any other result was a failure. We tested 10% of the sequences from each family in this way, for a total of 1035 sequences tested.

Results are shown in Table 1. Clearly, subfamily HMMs are proficient in recognizing sequences from their subgroup. The average success rate across the nine families was 97.4%, and the average for all sequences was 97.9% (22 sequences were incorrectly classified). In previous experiments (data not shown) classification errors typically come from one of several sources: alignment errors, subfamilies with few sequences losing sequences to larger subfamilies, fragments being misclassified and input multiple sequence alignments containing many gaps.

5. Discussion

Functional classification using homology-based methods is known to be prone to systematic errors of various types. Phylogenomic inference of protein molecular function has been shown by numerous investigators to improve the accuracy of functional classification, but is difficult to automate for high-throughput application.

This paper has described two tools to help automate phylogenomic inference of protein function, and demonstrated the use of these tools in predicting protein molecular function.

Bayesian Evolutionary Tree Estimation (BETE) identifies functional subfamilies given a multiple sequence alignment. BETE uses Dirichlet mixture densities and information theory to construct a phylogenetic tree and cut the tree into subtrees to obtain a subfamily decomposition. BETE has been demonstrated on a set of serotonin (and related) receptors; the subfamilies produced by BETE have been shown to correspond to known ligand receptor subtypes.

A novel method for constructing hidden Markov models for functional subfamilies has been described, given a subfamily decomposition and a multiple sequence alignment. Results have been provided on nine large and divergent PFAM families to demonstrate the use of subfamily HMMs for homolog detection, database discrimination and classification of novel sequences.

Subfamily HMMs have been demonstrated to provide better separation between homologs and non-homologs in database search than is possible using a single HMM for the family alone, recognizing more homologs with stronger scores and definitively rejecting non-homologs with large E-values. However, detection of true remote homologs is somewhat superior using HMMs constructed for the family as a whole, albeit with weak E-values and with a higher number of false positives. The advantages of using subfamily HMMs in detecting homologs are greater when the multiple sequence alignment used as the input for HMM construction contains a large number of variable sequences than when the sequences in the input alignment are more closely related.

Classification accuracy is high for subfamily HMMs. Given the biological validity of BETE-identified subfamilies, high classification accuracy makes subfamily HMMs a powerful tool for high-throughput functional genomics.

These results make sense in the light of a simple metric which can be used to compare the information content of BETE subfamily and whole-family alignments: the average pairwise BLOSUM62 score of the alignment columns, either within subfamilies or across the full alignment. As shown in Figure 3 and summarized in Table 1, within-subfamily scores are consistently higher than whole-family scores. This within-family similarity explains the high specificity of subfamily HMMs for closely-related sequences (such as training sequences, homologs, and sequences being classified). Conversely, the whole-family diversity explains the ability of general HMMs to improve upon SHMMs for remote homolog detection.

Our homolog detection results comparing subfamily and general HMMs illustrate the synergy between the two approaches for modeling protein families. General HMMs can be used as a first pass to detect related family members; subfamily HMMs then confer a more specific classification. Such a combined approach also minimizes the additional computational time required by our method, as only a very small fraction of sequences will be scored against the full set of subfamily HMMs.

Our results suggest several directions for future work. In particular, we chose the families in the current dataset based on our beliefs about what type of families would benefit from subfamily decomposition. Investigation of SHMM performance on a more representative set of families is an immediate priority. This will allow us to better ascertain how family characteristics such as size and diversity contribute to increased SHMM performance over GHMMs.

SHMMs require as input a subfamily decomposition; we have used the BETE method to derive this cut in these experiments. Other methods of obtaining this cut, including the use of standard phylogenetic tree construction algorithms, are also under investigation.

With regard to our SHMM construction algorithm, several issues should be investigated. Here we used a single method of sequence weighting; alternate algorithms may provide increased performance. Also, the homolog detection and classification performance of our SHMMs should be assessed against the 'naïve' SHMM method that does not share information between subfamilies. Such a method simply builds an HMM directly from the sequences in the subfamily.

Acknowledgements

This work was supported by National Science Foundation Grant No. 0238311, and by the National Institutes of Health under Grant R01 HG002769-01. The authors wish to thank the reviewers for their insightful comments.

References

1. M. Y. Galperin, E. V. Koonin, *In Silico Biol* **1**, 55 (1998).
2. J. A. Eisen, *Genome Res* **8**, 163 (Mar, 1998).
3. K. Sjölander, *Bioinformatics* (2004).
4. S. E. Brenner, *Trends Genet* **15**, 132 (Apr, 1999).
5. J. Park *et al.*, *J Mol Biol* **284**, 1201 (Dec 11, 1998).
6. J. A. Eisen, C. M. Fraser, *Science* **300**, 1706 (Jun 13, 2003).
7. J. C. Venter *et al.*, *Science* **291**, 1304 (Feb 16, 2001).
8. K. Sjölander, *Proc Int Conf Intell Syst Mol Biol* **6**, 165 (1998).
9. K. Sjölander *et al.*, *Comput Appl Biosci* **12**, 327 (Aug, 1996).
10. A. Gaulton, T. K. Attwood, *Curr Opin Pharmacol* **3**, 114 (Apr, 2003).
11. R. C. Edgar, *Nucleic Acids Res* **32**, 1792 (2004).
12. J. Felsenstein. (2003).
13. A. Bateman *et al.*, *Nucleic Acids Res* **30**, 276 (Jan 1, 2002).
14. L. Lo Conte *et al.*, *Nucleic Acids Res* **28**, 257 (Jan 1, 2000).
15. R. Hughey *et al.*, "Sequence Alignment and Modeling Software System" *Tech. Report No. UCSC-CRL-99-11* (2000).
16. S. E. Brenner, P. Koehl, M. Levitt, *Nucleic Acids Res* **28**, 254 (Jan 1, 2000).
17. T. J. Hubbard, B. Ailey, S. E. Brenner, A. G. Murzin, C. Chothia, *Nucleic Acids Res* **27**, 254 (Jan 1, 1999).
18. J. Gough, C. Chothia, *Nucleic Acids Res* **30**, 268 (Jan 1, 2002).
19. L. Lo Conte, S. E. Brenner, T. J. Hubbard, C. Chothia, A. G. Murzin, *Nucleic Acids Res* **30**, 264 (Jan 1, 2002).
20. R. E. Green, S. E. Brenner, *Proc. IEEE.* **9**, 1834 (2002).