Chapter 3

Enhanced Domain Recognition Using Phylogeny

There have been several suggestions in the literature for combining sequence based hidden Markov models (HMMs) with models of evolution [Yan95, MD95, FC96, TGJ96, SH04]. Evolutionary models model changes between homologous sequences at a site, typically with a uniform substitution process at all sites whereas sequence based hidden Markov models have site-specific models but only consider a single sequence at a time. The theme common to all of these suggestions is that integrated models will be both more realistic and more powerful for common bioinformatics tasks, such as building alignments, detecting homologues and reconstructing trees. Several of these models have been discussed in section 1.3. The goal of these methods is to improve the fit of phylogenetic models to real data, and thus to improve the reliability of phylogenetic inference made from these models.

Qian and Goldstein [QG03] have applied the tree HMM developed in [MD95] to incorporate the phylogenetic information contained in the seed alignment which is used to build a profile HMM. Recall from section 1.3 that the tree HMM only has match and delete states. Qian and Goldstein effectively re-label match states in which greater than half of the sequences in the seed alignment have a gap as insert states. The tree HMM simultaneously addresses the issue of weighting sequences to correct for redundancy and smoothing observed emission and transition counts to obtain probabilities. This approach determines a different profile HMM for each internal node in the seed tree. The process can be thought of as re-rooting the tree at a particular internal node, and using the Felsenstein algorithm (see

eq. 1.3.3) to calculate the distribution of transition probabilities and emission probabilities at the new root. This can be smoothed further by evolving these probabilities further back in time. This method does not consider the phylogenetic context of the target sequences to be scored by the profile HMM. To reduce confusion I will call methods which incorporate phylogenetic information in the seed alignment such as this one *tree profile HMMs*, consistent with the terminology introduced in [MD95], and call methods which incorporate phylogenetic information with respect to the target sequence *phylogenetic profile HMMs*, consistent with [SH04].

In this chapter I will consider whether the integration of models of evolution with profile HMMs can improve the detection of protein domains. I investigate whether it is possible to use sequences closely related to the query sequence to increase the sensitivity of the search. The motivation for this chapter was the observation that the Pfam annotation of closely related sequences is often inconsistent. It was reasoned that one could improve Pfam coverage by annotating the domain architecture of clusters of closely related homologues, rather than annotating proteins individually. As an example, figure 3.1 displays the N-terminal domain alignment and Pfam annotation for a cluster of homologues to GUDH_ECOLI. From the structure of this protein, it is known that this protein is a member of the MR_MLE Pfam family. The MR_MLE_N domain is detected as a significant hit in only two of the eight homologues, while the phylogenetic profile HMM method developed in this chapter scores the entire alignment above the Pfam threshold. The alignment also includes the consensus sequence from the profile HMM. This chapter investigates the extent to which the principle illustrated by this example can be applied on a large scale.

I will first describe in detail the phylogenetic profile HMM and in particular how it is built from a seed alignment and how it is used to score an alignment of target sequences. I will describe how site-specific frequency and rate variation is incorporated in the phylogenetic profile HMM. I discuss the time complexity of the algorithm and how the speed of the calculation can be increased by performing the calculations in an appropriate order. I also discuss the calculation of significance values. Subsequently I present the results of a SCOP test of the phylogenetic HMM on 44 Pfam families. The results of the test are given for several variations on the phylogenetic profile HMM. One of the parameterisations yields 67% more homologues above the first non-homologous sequence, thus demonstrating the potential gain

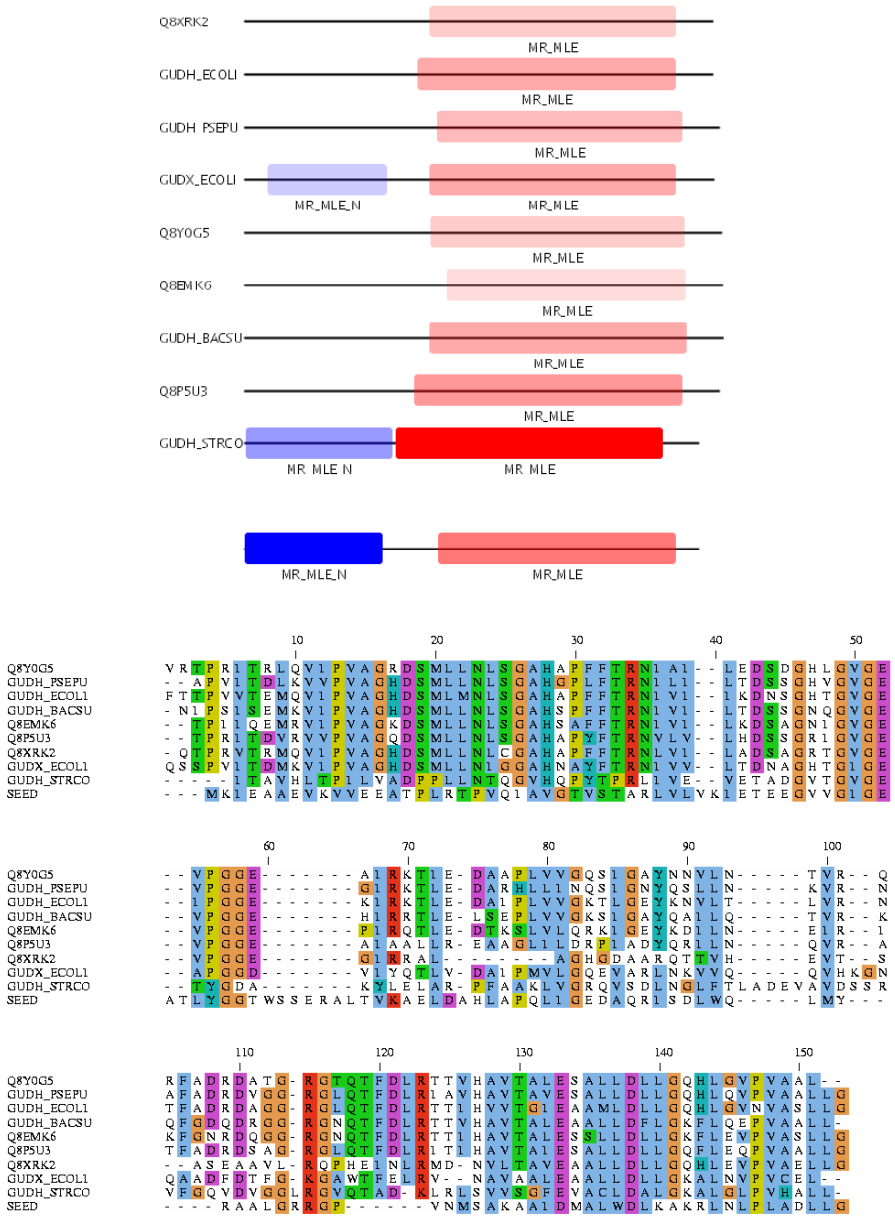

Figure 3.1: Top: Pfam full-model hits to homologues of GUDH_ECOLI. The opacity of the hit is proportional to the strength of the hit (log-odds score minus threshold) relative to the best scoring hit in this cluster. GUDH_STRCO has the strongest hits for both MR_MLE_N and MR_MLE domains. MR_MLE_N is detected in only 2 of the 9 proteins. The bottom track displays the result using the Phylogenetic HMM. It detects a MR_MLE_N signal which is stronger than any of the single protein signal, as well as a relatively strong MR_MLE signal. Bottom: Alignment of N-terminal domain using PROBCONS [DMBB]. The line marked 'seed' is the consensus sequence from the profile HMM.

in homology detection from this technique.

3.1 Algorithm

3.1.1 Phylogenetic profile HMM

Whereas the standard profile HMM described in section 1.2 parameterises a probability distribution over all possible sequences, a phylogenetic profile HMM $\mathbf{D}$ parameterises a conditional probability distribution over all possible alignments A of k sequences given a phylogenetic tree T with k leaves, which is denoted $P(A|\mathbf{D}, T)$. Let $\mathbf{R}$ denote a background model, which also parameterises a conditional probability distribution over alignments A given a tree T, $P(A|\mathbf{R}, T)$. As for the standard profile HMM, the log-odds score

$$\log \frac{P(A|\mathbf{D}, T)}{P(A|\mathbf{R}, T)} \tag{3.1}$$

is used to classify matches to the model.

A phylogenetic profile HMM uses the same HMM model architecture as the profile HMM, as shown in figure 1.2, except that the emission states of the model emit columns of an alignment (given a tree) rather than residues. If the tree T is a single node, the phylogenetic profile HMM reduces to a standard profile HMM. The main underlying idea is that each of the match states of the profile HMM corresponds to a different evolutionary model which reflects the structural and functional constraints of this position in the protein domain. A standard profile HMM relies on detecting the biased distribution of residues at a site in a protein domain for its predictive power. A phylogenetic profile HMM also relies on detecting a specific residue distribution, but can also take into account whether the pattern of substitutions in a column is consistent with the particular match state. This is illustrated in figure 3.2, which shows an alignment of part of the MR_MLE_N model to the GUDH_ECOLI alignment discussed above. In the first column most positions in the first row match the consensus valine, and in cases where the position does not match the consensus it has mutated within the class of 'allowed' residues at this position (alanine, isoleucine and leucine). 'Allowed' is taken to mean residues which are observed in the seed alignment but at lower frequencies. In column 8, none of the sequences matches the consensus, glutamate, but the observed conserved serine and and alanine residues still appear to be consistent with this match state. Columns 10, 13, 15 correspond to a highly conserved glycine in both the seed and the target alignment.

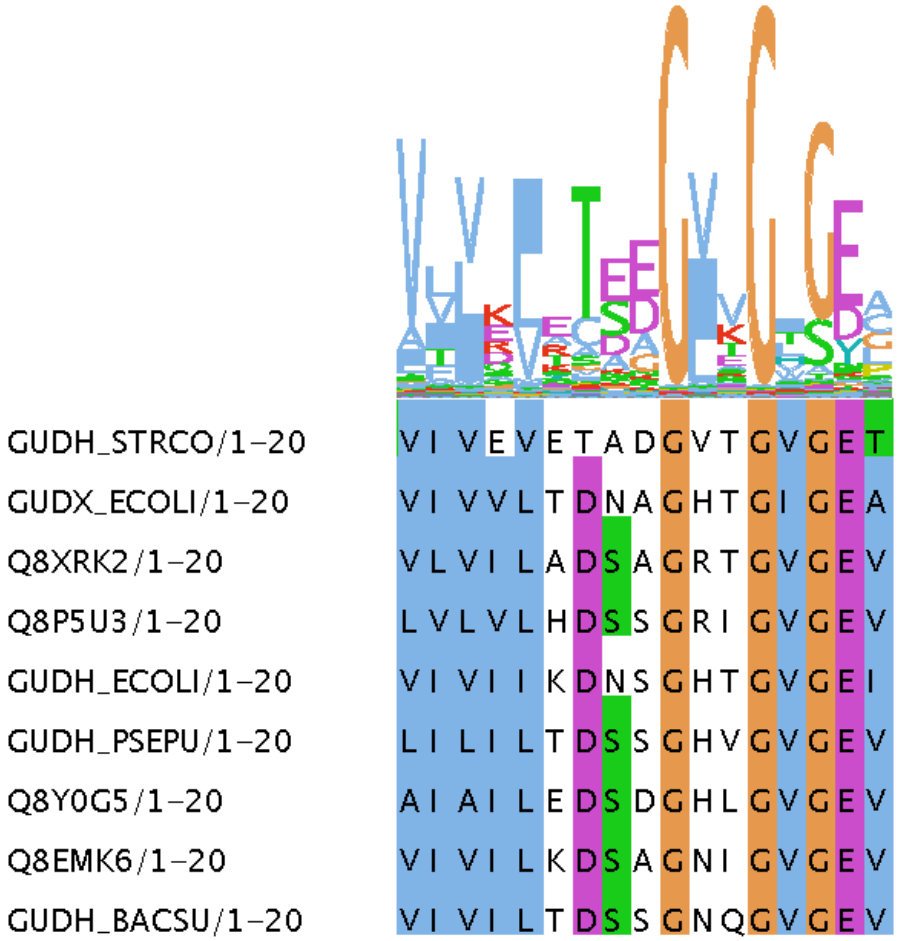

Figure 3.2: A fragment of the alignment of the GUDH_ECOLI alignment (bottom) shown together with aligned emission states from the profile HMM(top). Match states 34 to 50 are shown together with the corresponding columns from the alignment (which may not be contiguous). The total height of a column in the HMM logo is proportional to 1 - entropy of match state/maximum entropy, where maximum entropy is the entropy of the uniform distribution over 20 states. Thus a perfectly conserved column will have a height of 1 and the uniform distribution will have a height of 0. The relative heights of the residues within a column of the HMM logo are just the relative frequency. Note that the alignment in this figure has been calculated using PROBCONS [DMBB] rather than *hmmalign*, which aligns the sequences individually to the profile HMM. In other words, the alignment has been calculated without assuming a match to the HMM states.

Let $A = \{x_{k,i}\}_{1 \leq k \leq K, 1 \leq i \leq n}$ denote the target alignment to be scored by the phylogenetic profile HMM, where k indexes the sequences and i indexes the column of the target alignment. Let $S = \{s_{k,i}\}_{1 \leq k \leq K, 1 \leq i \leq n}$ refer to the *seed* alignment used to build the HMM. Let $\psi_1, \ldots, \psi_n$ denote the path the HMM takes through the alignment, so that ψ_i is the HMM state which emits column $x_{.,i}$.

All of the standard HMM algorithms (Viterbi, forward, forward-backward) will apply to the phylogenetic HMM provided the emission probabilities are interpreted as the probability of emitting an entire column of an alignment, i.e. $P(x_{.,i}|\psi_i = \mathbf{M}_j, T)$.

Estimating substitution models for match states

In order to calculate the emission probabilities $P(x_{.,i}|\psi_i = M_j|T)$, a different substitution model is constructed for each match state, with the aim of building a model of evolution at each conserved site in the seed alignment which reflects the evolutionary pressures acting at this site. The approach I take is empirical, rather than theoretical, in that the observed residues from a column in the seed alignment are used as the basis for building the site-specific models of evolution, rather than (for instance) restricting site-specific evolution within a particular class of residues (e.g. a hydrophobic site). I will assume that the substitution models are homogeneous with respect to position on the tree. This assumption will be relaxed in Chapter 4 in order to test for differential evolution along particular branches of the tree.

Once the substitution models have been parameterised, the emission probability calculation proceeds using Felsenstein's tree pruning algorithm, as described in section 1.3.3.

I follow the approach to modelling evolution outlined in section 1.3, in which mutations are viewed as part of a continuous time Markov process where the instantaneous rate of mutation between amino acids is given by a 20x20 rate matrix Q. Within this framework, there are many possibilities for parameterising a rate matrix on the basis of observed residues at a particular site. The challenge in formulating the right rate matrix is one of accurately describing the evolutionary process without over-fitting the model. Database derived rate-matrices (such as WAG [WG01], JTT [JTT92]) contain a lot of information about amino acid exchangeabilities, which presumably still apply to constrained sites. My approach is to use the observed residue frequencies in a column to estimate the stationary probabilities of the site-specific rate matrix, which are then used in equation 1.27 to calculate the terms $Q_{u,v}$ in

the rate-matrix. This equation has an extra parameter f called the +gwF parameter which can either be set to 0, resulting in equation 1.28, or can be modelled specifically for each state. Similarly, the rate r in equation 1.13 can be set to 1 or can be modelled specifically for each state. Both of these possibilities will be discussed below.

Estimating the site-specific stationary probabilities from an alignment column

The stationary distribution π of a continuous-time Markov process can be shown (see [Nor97]) to be equal to the frequency distribution of residues which would be observed if the evolutionary process was allowed to run for an infinite amount of time. The simplest approach to estimating this distribution from a column of residues is to set the probability of a residue to the frequency at which each residue occurs in the column. However, this approach suffers from two problems: firstly, it over-fits the model to the data given, and automatically disallows unobserved residues to occur at this site, even if they may occur but with low probability; secondly it assumes that each sequence is sampled independently from the target distribution and hence weights them equally, when in fact the observations are highly correlated.

The problem of over-fitting to the data has already been solved for standard profile HMMs using Dirichlet priors as discussed in section 1.2 and the same type of approach can be applied here. The problem of differentially weighting sequences has also been addressed in the profile HMM literature. Dirichlet priors and several sequence weighting schemes including maximum entropy are incorporated into the *hmmbuild* program in HMMER. The approach used in this chapter is to obtain the stationary probabilities from HMMER using *hmmbuild*, using a mixture of Dirichlet priors and a maximum entropy weighting scheme. An alternative approach which has not been investigated is to use the tree HMM. This approach explicitly incorporates the phylogeny of the tree of the seed alignment.

Estimating substitution models for the non-match emission states

The discussion so far has focussed on modelling match states of the HMM. Equally important are the non-match states: insert states I, linking states N, J, C and the null emission state G. One option, which I shall call the *mixture* model, is to regard each of the non-match emission

states as a mixture of the match emission states and to score

$$P(x_{.,i}|\psi_i = I_j, T) = P(x_{.,i}|\psi_i = J, T) = P(x_{.,i}|\psi_i = C, T) = P(x_{.,i}|\psi_i = N, T) = P(x_{.,i}|\psi_i = G, T) = \frac{1}{M}\sum_{j'} P(x_{.,i}|\psi_i = M_{j'}, T). \quad (3.2)$$

This strategy requires no extra likelihood calculations as the algorithm is already scoring each of the $P(x_{.,i}|\psi_i = M_{j'}, T)$ for the match state emission probabilities. However, the method for taking an unweighted average over the match state emissions is somewhat ad-hoc, but was found via experimentation to work reasonably well. The second, *non-mixture* model uses the same approach used for the match states, and calculates substitution models using equation 1.27. The stationary probabilities are again taken to be the Dirichlet smoothed frequency distributions calculated by HMMER.

Incorporating rate and gwF variation in match states

As discussed above, the equations used to calculate the match state substitution models eqs. 1.12, 1.27 allow the possibility of site-specific rates and +gwF mode. Figure 3.3 displays two sites which have the same stationary distribution but different rates and/or +gwF mode. Capturing this variation in the phylogenetic HMM may improve sensitivity.

As described in section 1.3 the gwF parameter f takes values between 0 and 1 and describes the degree to which the stationary probabilities are explained by the probability of mutating from or mutating to a residue. In the 'from' model, once a favoured residue is discovered, it is unlikely to be changed; while in the 'to' model, a favoured residue is likely to be re-discovered and mutated away from several times. The optimal +gwF parameter for a column will depend to some degree on the rate – figure 3.3 can be viewed either as demonstrating the difference between a 'from' and a 'to' (top vs bottom respectively) or as a fast vs slow column.

Here I describe how to model the rate and +gwF variation jointly, but the equations presented apply equally well to fixing the +gwF parameter at 0 and only allowing the rate to vary, or fixing the rate at 1.0 and allowing f to vary. Using a standard gradient descent algorithm [PTTF92], it is possible to find the values of r and f which maximise the likelihood of the column of a seed alignment under the site-specific rate-matrix obtained above. However,

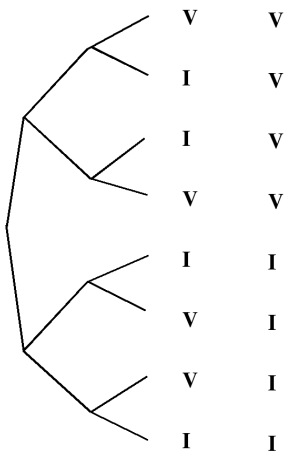

Figure 3.3: Illustration of the effect of different rates of evolution and/or different gwF modes of evolution at sites with similar functional constraints. Two site are shown for the same tree, the first site appears to be evolving much slower than the second site, however the stationary probability distribution is identical. Alternatively, the first site is evolving according to a 'from' model and the second according to a 'to' model.

this approach will over-fit the data to the extent that a uniformly conserved site will have a rate of 0, thus precluding transition to any other residue in this column. To avoid this, a prior $P(r, f)$ was introduced over the rates and gwF parameter. Experimentally calculating maximum likelihood values of f over large Pfam seed alignments revealed a preference for f values close to 0 or 1. Thus f was constrained via an indicator prior $\mathcal{I}_{\{0,1\}}(f)$, in which f takes values 0 or 1 each with probability 0.5. The gamma distribution $\gamma_{\sigma_r^2,1/\sigma_r^2}(r)$ with mean 1 and variance σ_r^2 was chosen as the prior distribution over rates, as it has been used successfully in modelling rate variation [Yan93]. The site specific rate and gwF parameters were then chosen to be those that maximised the posterior probability

$$P(r, f|Q, s_{.,i}) = \frac{P(s_{.,i}|Q, r, f) \cdot P(r, f)}{P(s_{.,i})}. \tag{3.3}$$

The prior was parameterised as

$$P(r, f) = \gamma_{\sigma_r^2,1/\sigma_r^2}(r) \cdot \mathcal{I}_{\{0,1\}}(f) \tag{3.4}$$

where the gamma distribution is given by

$$\gamma_{b,c}(r) = \frac{r}{b}^{c-1} \cdot \frac{\exp(-r/b)}{b\Gamma(c)}, \tag{3.5}$$

and where $\Gamma(c), c > 0$ is the gamma function (see [EHP00]).

The variance σ_r^2 controls the trade-off between fitting the observed pattern of evolution and over-fitting this pattern. Experimentally it was discovered that setting $\sigma_r^2 = 0.01$ provided a good trade-off. Values of r much higher than this (e.g. 0.05) degraded the performance of the algorithm by over-fitting. In essence, a small value of r encourages most site-specific rates to be close to 1 but allows some deviation if there is evidence of an elevated or decreased rate. In passing, I note an alternative to the gamma distribution is the log Gaussian distribution $N_{\mu,\sigma}(\log(r))$. The conceptual advantage of this distribution is that the probability is symmetric with respect to its inverse: a rate of y has the same probability as a rate of $1/y$, or in other words a site is as likely to be evolving y times slower as y times faster than average. This prior was not investigated further.

Rate and gwF variation can also be incorporated into the non-match emission states. The approach I take is to incorporate rate and/or gwF variation in the non-match emission states if and only if it is also used in the match emission states. Using the mixture model, rate and gwF variation will automatically be incorporated into the calculation if it is incorporated into the match states. If, instead, I use the non-mixture approach, rate and gwF variation can be incorporated by marginalising over a rate and gwF distribution. For consistency with the treatment of match states, the gamma distribution is used to marginalise over rates, and $\mathcal{I}_{\{0,1\}}(f)$ is used to marginalise of f, so that the equation used is

$$P(x_{.,i}|\psi_i = I_j, T, \gamma_{\sigma_r^2}(r), \mathcal{I}_{\{0,1\}}(f)) = \frac{1}{2}\sum_{f=0,1}\sum_{r_l} P(r_l)P(x_{.,i}|\psi_i = I_j, T, r_l, f) \tag{3.6}$$

where r_l are the rate categories used in the discrete approximation to the gamma function. The value of σ_r^2 for the gamma distribution was 1. Note that this value is larger than that used in the prior over the match state rates. This is because the choice of small r in that case was to avoid over-fitting, whereas the concern for modelling non-match emission states is to correctly represent the range of rate variation present in real data.

Building the profile HMM

The *hmmbuild* program in HMMER builds profile HMM architecture and transition probabilities using the maximum a posteriori (MAP) architecture algorithm [DEKM98], as explained in section 1.2. This technique builds the profile HMM architecture which maximises the sum of the probabilities of each sequence in the training alignment. This strategy solely uses pos-

itive training data. It has been shown in [WS04] that a more sensitive approach is to re-train transition probabilities (on a fixed architecture) using both positive and negative training data. The negative training data is generated by the null model and the highest scoring random sequences are used to re-train the transition probabilities.

The MAP architecture algorithm could be adapted to build the profile HMM which gives maximum probability to the alignment, using site specific rate matrices, provided it uses the non-mixture model for the non-match emission states. This might seem more internally consistent than using HMMER on the seed alignment. As before, residue emission probabilities would be replaced with column emission probabilities. I have not investigated this option further.

Restricting the path of the phylogenetic profile HMM

Occasionally the non-mixture model gave a non-homologous sequence cluster a high score because it contained a few columns which fit particularly match states well, such as a conserved cysteine column. The model would give these columns very high scores, and would use insert states to traverse the remaining sequence. The mixture model partially addresses this problem by including a fraction of this high scoring contribution in the null model score. A simple heuristic approach was used to solve this problem. The matrix of column emission probabilities $P(x_{.,i}|\psi_i = M_j, T)$ for the phylogenetic HMM is calculated as before, and then adjusted via

$$P(x_{.,i}|\psi_i = M_j|T) := \begin{cases} P(x_{.,i}|\psi_i = M_j|T) \text{ if } \max\limits_{1\leq k\leq K} P(\psi_i = M_j|x_k) > 0.01. \\ 0 \text{ otherwise} \end{cases} \tag{3.7}$$

The posterior probabilities in the previous equation are calculated using the forward-backward algorithm. This has the effect of restricting the path through the dynamic programming matrix that the phylogenetic HMM can take. This solves the problem of random sequence clusters matching a few columns strongly.

An alternative approach is to use a strategy based on HMMER's null2 model. In addition to the original null model, a second alignment-specific null model is calculated based on the Viterbi path taken by the model through the sequence, which is the mixture model of all of the emission states traversed by the path. If, as in the example above, the model

matches a conserved cysteine column with match state 1 and then exclusively uses insert states and delete states, the mixture would consist of 1 copy of M_1 and $n-1$ copies of an insert state, where n is the number of columns in the alignment. Denote the likelihood of the alignment under the second null model as $S2$. HMMER incorporates this score with the original log-odds score by subtracting $\log(1 + S2/256)$ from the original log-odds score to arrive at a corrected score. The factor 256 represents the prior belief that the main null model is 256 times more likely than the second null model. See [Edd03] for more details. Note that no additional tree-likelihood calculation has to be performed as every emission state has already been scored against each model.

Complexity

First I consider the complexity of searching the model against a sequence. The complexity of the likelihood calculation for a fixed alphabet is $O(K)$ where K is the number of sequences. The forward algorithm has complexity $O(NM)$ where N is the number of residues and M is the number of states. Thus the complexity of the phylogenetic HMM is $O(KNM)$. Note, however, that the algorithm simultaneously scores K sequences, and hence the average complexity per sequence is the same as for the profile HMM.

The order in which the $P(x_{.,i}|\psi_i = M_j)$ are calculated impacts on the speed of the implementation. In this implementation, M_j is first fixed, and then Felsenstein's algorithm proceeds for each site in the alignment simultaneously. That is, as the Felsenstein algorithm proceeds upwards from the leaves, the transition probability matrix $e^{\mathbf{Q}rt(n_k)}$ over branch length $t(n_k)$ is calculated via equation 1.20, and this is applied to each column i in the alignment to calculate the terms $P_{\mathcal{E}_{M_j}}(x_{kh} = v|x_k = u)$. This order of calculation avoids unnecessarily exponentiating the same rate matrix multiple times for the same branch length. This does not improve the complexity of the overall algorithm.

The most time consuming step in model construction for a standard profile HMM is the maximum entropy sequence weighting step, which is unchanged for the phylogenetic HMM. If the phylogenetic HMM incorporates differential rates and gwF values, there is an extra step of optimising these two parameters for each match state. The search consists of two one-dimensional searches optimising r, one for $f = 0$ and one for $f = 1$. Constructing a ML tree from the seed alignment is the rate limiting step in this process.

Significance calculation

In this work significance is calculated using the extreme value distribution (EVD) parameterised by *hmmcalibrate* acting on the standard profile HMM. As described in section 1.2, this works by generating 5000 random sequences, each of length 350, and parameterises the extreme value distribution to fit the distribution of these scores. A more robust approach is to calculate the EVD directly from alignments scored with the phylogenetic HMM, and this remains an area for further research. This could be achieved by first simulating (say) 5000 trees of varying numbers of sequences, according to a distribution over the number of sequences in an alignment. Clock trees can be rapidly sampled by a coalescent approach, where k sequences are generated, and recursively two nodes are chosen randomly to 'coalesce' at height t above the highest node of the pair, where t is sampled from an exponential distribution. An alignment can be simulated on this tree according to a background model by sampling the sequence at the root node from the equilibrium distribution, and progressively evolving the residues of this sequence to the leaves, determining the sequence at inner nodes along the way. These alignments can be scored against the model and the resulting scores used to parameterise an EVD.

3.1.2 Using the phylogenetic profile HMM

Figure 3.4 shows an overview of how the phylogenetic profile HMM is used in practice. which broadly consists of four steps

- Identifying, aligning and constructing a tree for a homologous cluster of sequences.
- Building a phylogenetic profile HMM.
- Calculating the emission probabilities for each column and match state.
- Dynamic programming to find the overall log-odds score

The homologous cluster of sequences can be obtained from a global clustering of proteins (using, for example, PHIGS [Deh] or Tribe-MCL [EKO03]). Alternatively, for a single target query sequence, the homologous cluster can be obtained via a blast [AMS$^+$97] search of Uniprot [ABW$^+$04]. In this case, only proteins which have blast hits of significance less that

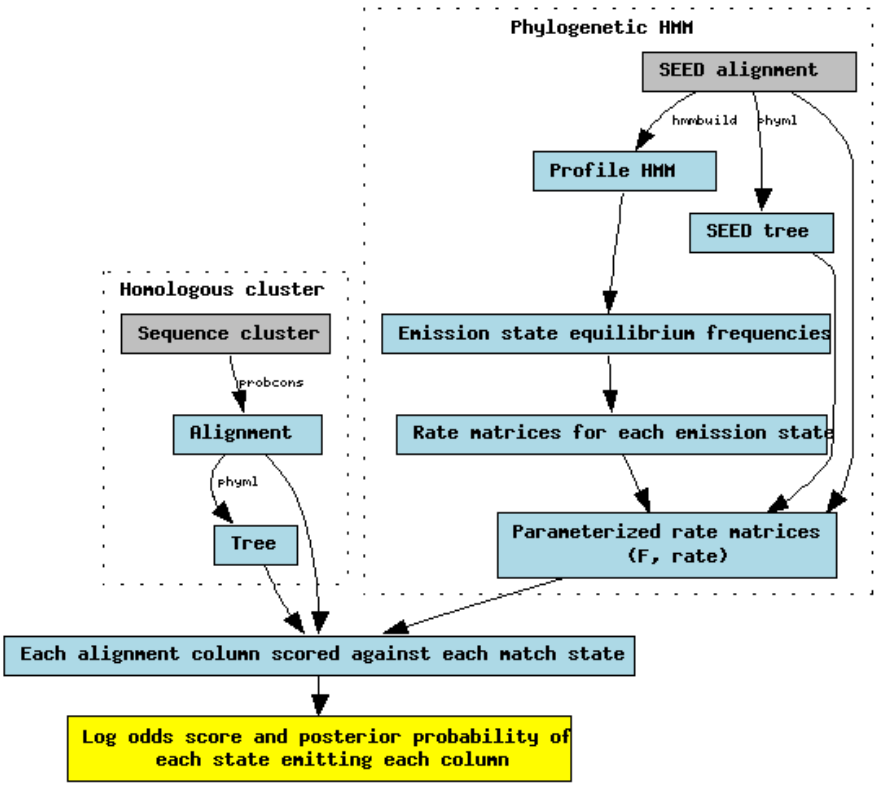

Figure 3.4: Diagram of processes involved in the phylogenetic profile HMM. Inputs are shown in grey, and outputs in yellow, with intermediate steps in light-blue.

10^{-7} covering at least 80% of the query sequence (this could comprise multiple local hits) are accepted into the cluster.

For the SCOP test described in the results section, I use PROBCONS [DMBB] to align the homologous cluster of sequences. Phyml [GG03] is used to build a maximum likelihood tree, using a WAG matrix, and 4 gamma rate categories. If the tree has more than 5 leaf nodes, it is trimmed back to 5 nodes in such a way as to include the original query sequence and to include the most diverse collection of remaining sequences. This set of 5 sequences is calculated recursively – a set of $k+1$ sequences is generated from a set of k sequences by adding the sequence which has the largest average pairwise distance to the k sequences. This trimming step was performed in order to minimize the computational time taken.

The phylogenetic profile HMM is built as described in the previous section, using the HMMER smoothed emission probabilities as the stationary probabilities of state-specific rate matrices, and determining site-specific rates and +gwF modes as described above. The emission probabilities are calculated using Felsenstein's tree pruning algorithm [Fel81] and the forward algorithm is used to find the overall log-odds score.

3.2 Results

I compare the detection of homologues by a phylogenetic HMM to a standard profile HMM using the same Pfam derived SCOP test presented in section 2.4.1. For each Pfam family tested, HMMER is used to find all sequences from the ASTRAL set filtered to 40% identity which have an evalue match of less than 100. These proteins form the test set for the method, with correct homologues assigned on the basis of belonging to the same SCOP superfamily as the Pfam domain, and incorrect homologues assigned on the basis of belonging to a different fold. As discussed in the previous section, for each query protein from ASTRAL the target cluster of homologous proteins is constructed via a blast search.

The phylogenetic HMM is compared to a custom implementation of HMMER's *hmmsearch* rather than *hmmsearch* itself. The custom implementation was used so that the phylogenetic HMM is compared to a profile HMM which is identical in all respects except for the fact that it scores columns rather than residues. In other words, all of the dynamic programming routines used are the same, the only difference is in the way in which the emission probabilities are calculated. The scoring of the profile HMM differs from the HMMER scoring

scheme in two ways. Firstly, the forward algorithm rather than the Viterbi algorithm is used to score the model. One reason for using the Viterbi algorithm in HMMER is speed and memory usage, it allows all calculations to be done in log space using integers, rather than in probability space which requires extra memory and time to progressively scale rows in the dynamic programming matrix to avoid underflow errors. The disadvantage, as discussed in section 1.2, is that the the Viterbi algorithm calculates the probability of the most likely path through the model, rather than the full probability of the model emitting the sequence. Another advantage of the forward algorithm is that it allows, in conjunction with the backward algorithm, the calculation of posterior probabilities of a state of the model emitting a particular site. These probabilities are used in the model scoring step to restrict the path of the phylogenetic HMM (see section 3.1.1), and will be useful for detecting positively selected sites in Chapter 4. Moreover, the Felsenstein algorithm requires working in probability space, as it involves a summation over probabilities (although by replacing the $\sum$ in the Felsenstein algorithm with a max a Viterbi algorithm could in theory be applied to approximating the tree likelihood with the probability of the most likely ancestral reconstruction). Using Viterbi rather than forward does not impact the speed of the algorithm, as the calculation of the tree likelihood is the slowest step. Thus I have decided to use forward rather than Viterbi algorithm. The second difference with respect to HMMER is that the model does not incorporate a null2 model. The null2 model has been shown to increase performance and will be incorporated into this implementation at a later date. In this study the phylogenetic HMM and the profile HMM are consistent in that they both do not use a null2 model. As discussed in the previous section, I use a heuristic technique to limit the potential path of the phylogenetic HMM, using the matches to the individual sequences in the alignment. Use of a null2 model may render this technique unnecessary.

I score three variations of the standard profile HMM on the target protein cluster:

(i) standard profile HMM log-odds score on the ASTRAL query sequence,

(ii) average of the log-odds scores on each of the proteins in the cluster,

(iii) maximum of the log-odds scores on each of the proteins in the cluster.

The second and third scores can be seen as simpler alternatives to the phylogenetic profile HMM for integrating information from closely related proteins.

I score several variations of the phylogenetic profile HMM:

(i) Non-mixture model: non-match state emissions (including null model) not a mixture of the match state models, but rather parameterised using the relevant HMMER emission probabilities as stationary probabilities; no rate or gwF variation.

(ii) Non-mixture model+rate variation: using a gamma distribution over as prior with variance 0.01 for match state emissions, and marginalising over a 3-category discrete gamma distribution with variance 1 for the non-match state emissions.

(iii) Non-mixture model+ rate and gwF variation: using a gwF prior of $I_{0,1}$ for determining match state gwF values, and marginalising over the same distribution for the non-match emissions.

(iv) Mixture model: non-match state emission probabilities are calculated as the average of the match state emission probabilities; no rate or gwF variation.

(v) Mixture model+rate variation: rate variation as in (ii), although the non-match emission states no longer need to be calculated

(vi) Mixture model+rate and gwF variation: gwF and rate variation in (iii); again the non-match emission states are not calculated.

Firstly, models without gwF variation are considered. The coverage vs. error curves scored on 44 Pfam families are shown in figure 3.5. Statistics summarizing the performance of each of the methods are shown in table 3.1. Each of the phylogenetic HMM methods has a higher coverage at a given error from after the first false positive onwards, and each improves the classification in more families than they degrade it (as assessed by the number of homologous sequences scored above the first non-homologous sequence, the over the top or OTT score). Each of the phylogenetic methods improves the sum of family OTT and MER (minimum error rate scores) relative to a standard profile HMM. If the scores are ranked globally, according to a p-value criteria, then all of the phylogenetic methods have a higher aggregate OTT score and all have a lower aggregate minimum error rate. The best performing method is the phylogenetic HMM with no mixture and with rate variation, which scores 67% more homologous sequences above the first non-homologous sequence relative to

the standard profile HMM, and reduces the error rate by 29%. Rate variation appears to improve the performance of the non-mixture model but does not impact the mixture model, suggesting that the biggest impact may be due to marginalising over several possible rates in the null model. The performance of the maximum and average profile HMM scores is mixed – they have a lower aggregate MER but higher OTT scores, improve the classification in more families than degrade it, and improve the sum of family OTT and MER scores. However the improvements are not as pronounced as for the phylogenetic HMM. The change in a performance on a family by family level can be seen in figure 3.7. The largest family improvement in the 44 families tested is in the immunoglobulin (ig) domain.

The error versus significance curves for the phylogenetic HMMs versus the profile HMM are shown in figure 3.6. The phylogenetic HMMs each have false positive rates at a fixed p-value threshold which are much lower than the standard profile HMM, as well as higher false negative rates. The increase in false negative rate is smaller than the decrease in false positive rate such that the phylogenetic profile HMMs overall perform better. The phylogenetic HMM false negative and false positive rate increasingly diverge from those of the standard profile HMM as the p-value increases. This is due to the e-value not being calibrated very well for the phylogenetic HMM at high p-values. Within the different types of phylogenetic profile HMM, the non-mixture models have a lower false negative rate at low p-value thresholds, and modelling rate variation does not appear to influence error rates substantially, although for the non-mixture model at low p-value thresholds, the false negative rate is below even the profile HMM false negative rate. As expected, using the maximum of the standard profile HMM scores has a lower false negative rate but higher false positive rate, while using the average score gives higher false negative but lower false positive rates.

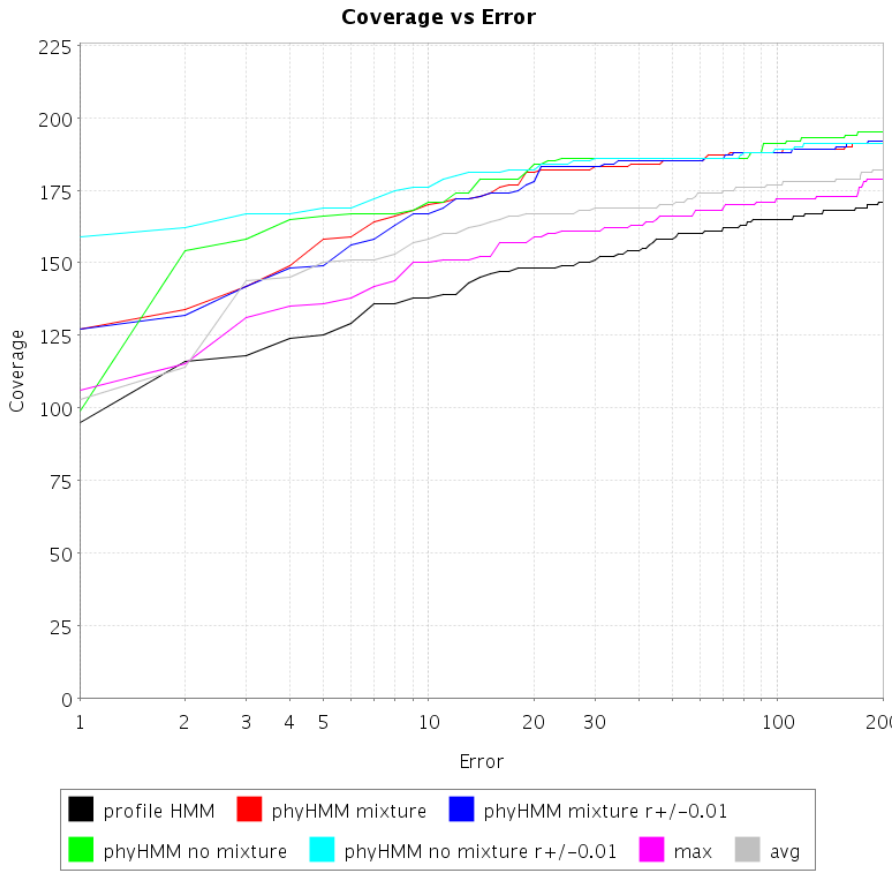

Figure 3.5: Coverage vs error curve for phylogenetic HMM vs standard profile HMM on ASTRAL test set and 44 Pfam families. The coverage at error rate of n is defined as the number of homologous sequences before the n^{th} false positive (or non-homologous sequence). The black line is the standard profile HMM score on the sequence from ASTRAL, the purple line is the maximum of all sequence scores in the same homologous cluster, and the grey line is the average sequence score in the cluster. The green and red lines are scores for a phylogenetic HMM without rate variation with a mixture null model and non-mixture null model respectively. The dark and light blue lines are scores for a phylogenetic HMM with rate variation modelled according to a gamma distribution, and with a mixture null model and non-mixture null model respectively. The best performing method is the non-mixture model with rate variation.

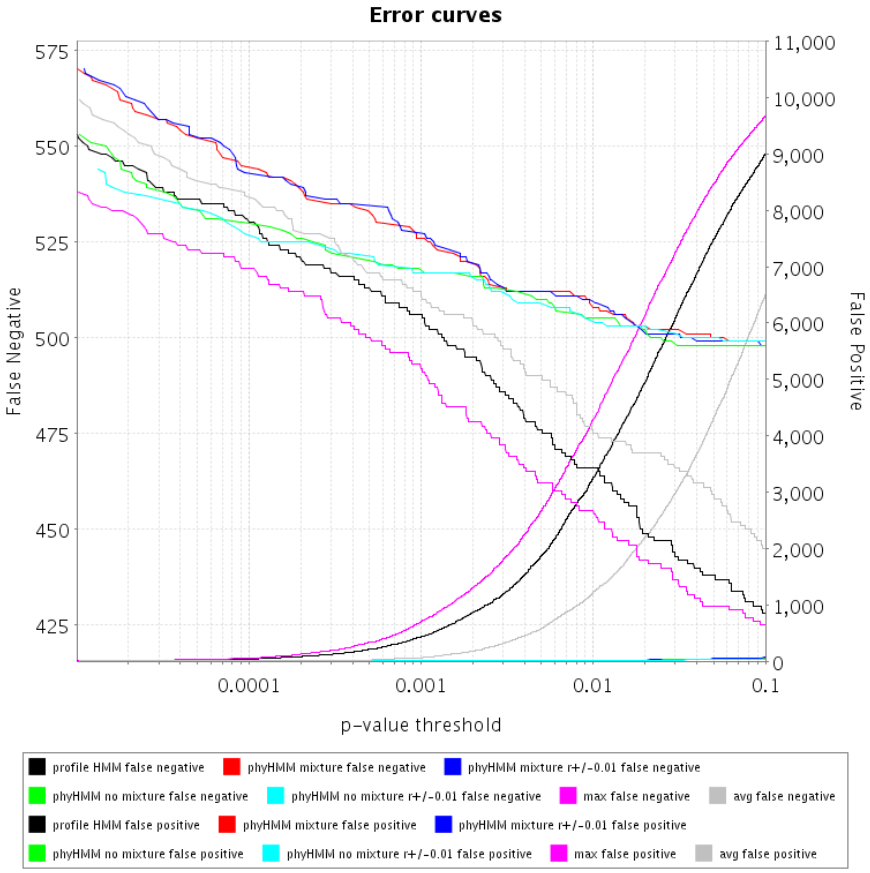

Figure 3.6: Error rate vs p-value score threshold according to HMMER extreme value distribution. False negatives are shown on the left x-axis (which fall from left to right as the p-value threshold increases). False positives are shown on the right x-axis (which rise as the p-value threshold increases). The curves are: standard profile HMM score (black), average of profile HMM score (grey), maximum of profile HMM score (purple) and the following phylogenetic HMM scores: mixture, no rate variation (red); mixture with rate variation (dark blue); non-mixture with no rate variation (green); non-mixture with rate variation (light blue). The rate variation if used was according to a gamma prior with variance 0.01. This figure is plotted on a log x-axis to emphasise the behaviour of the algorithms at low false positive rates, which is the range in which most applications – including Pfam – use homology detection.

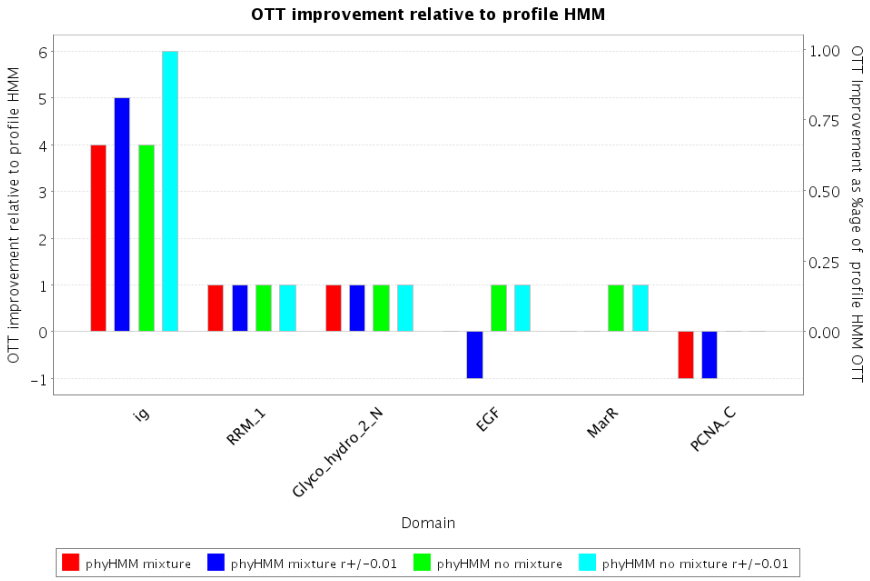

Figure 3.7: Improvement in OTT score relative to the standard profile HMM on a family basis. The red/blue bars are the scores for the phylogenetic HMM with mixture null model and without/with rate variation respectively. The green and cyan bars are for models without a mixture model background, and without/with rate variation. The biggest improvement is seen in the immunoglobulin (ig) family.

Method	# families with OTT		sum of family score		Aggregate score	
	Better	Worse	OTT	MER	OTT	MER
profile HMM	0	0	174	86	95	129
phyHMM mixture	3	1	179	79	127	98
phyHMM mixture r+/-0.01	3	2	179	78	127	98
phyHMM no mixture	5	0	182	75	99	95
phyHMM no mixture r+/-0.01	5	0	184	77	159	92
max	4	0	179	80	106	119
avg	5	0	186	74	103	111

Table 3.1: Comparison of phylogenetic models with a standard profile HMM scored over 44 families.

Now I consider the effect of including +gwF rate variation. Figure 3.8 displays the effect

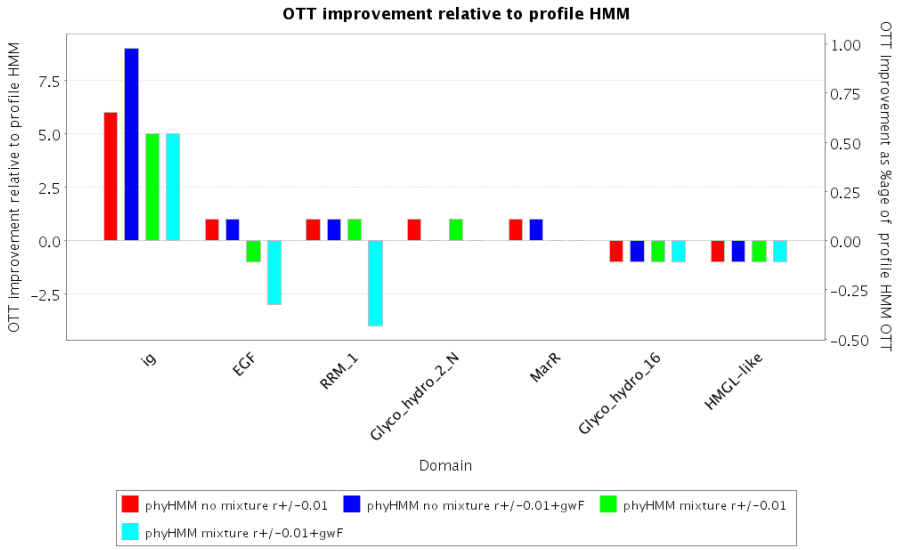

Figure 3.8: Impact of gwF variation on detecting homology for six families. +gwF variation degrades performance for the mixture model, and in one case improves performance for the non-mixture model.

of gwF variation on an individual Pfam family basis. In all cases rate variation is included in the model. For the non-mixture model, +gwF variation provides a substantial improvement in detecting immunoglobulin (ig) domains. However, for the mixture model, gwF variation systematically degrades homology detection. Figure 3.9 displays the aggregate results over 30 Pfam models. Including +gwF variation degrades detection of homology in both cases. Unfortunately, I cannot conclude from this that modelling +gwF variation is always detrimental to performance. A possibility is that +gwF will improve performance if rate variation is not also incorporated. Moreover, there are many ways to model +gwF variation and this result could be due to they way the models described in section 3.1 incorporate this information. Two alternative priors on +gwF have been experimented with, including a uniform prior and a beta distribution parameterised to best fit the seed alignment, neither of which yielded better results.

3.3 Conclusion

Scoring clusters of closely related proteins with phylogenetic profile HMMs can provide significant improvement in homology detection. However, the degree of improvement is sensitive

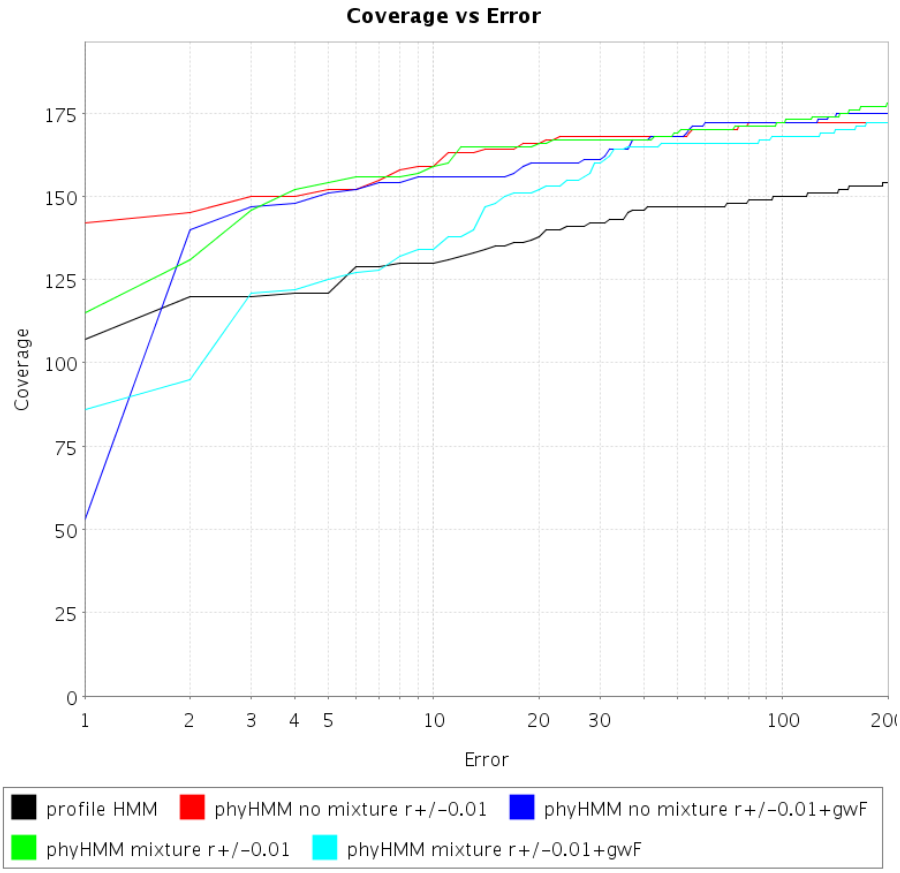

Figure 3.9: Coverage vs error curve for models which include +gwF variation (light and dark blue) vs models which do not include +gwF variation (red and green), scored on 30 Pfam families. The curve for a standard profile HMM is included in black for reference. Both models which have a mixture null model (cyan and green), as well as models do not have a mixture null model (blue and red) are shown. All models incorporate rate variation.

to the way in which the phylogenetic profile HMM is parameterised. One particular parameterisation yielded 68% more homologues scoring above the first non-homologue in a SCOP test, while other parameterisations yielded a more moderate improvement, or in some cases degraded performance.

The original motivation of this chapter is to increase coverage of domain databases such as Pfam. The results of this chapter suggest that the phylogenetic profile HMM has greatest impact when assessing scores based on a significance threshold. In particular it dramatically reduces false positive rates at a given p-value threshold. The annotation strategy in Pfam, however, is based on family specific thresholds and so the potential increase in Pfam coverage should be assessed based on the increase in the sum of family based OTT scores. On this basis, the phylogenetic HMM could produce a 5.7% increase in Pfam coverage, which can be compared with the 2.2% increase achieved with a combined domain and taxonomic context model in the previous chapter. However, a direct study into the potential improvement in Pfam is required.

A feasible strategy for introducing the phylogenetic profile HMM into Pfam would be possible, if computationally expensive. The first step would be to globally cluster proteins in Uniprot using a clustering algorithm such as Tribe-MCL [EKO03] and reciprocal best blast scores, or alternatively using a phylogenetically derived clustering such as PHIGS [Deh] (PHIGS, however, only clusters proteins from fully sequenced organisms and so the clusters would need to be extended to non-sequenced organisms, potentially using a HMMER search built from a PHIGS derived seed alignment). The alignment and tree building step currently involves relatively slow but accurate algorithms, whereas a global Pfam strategy would require faster algorithms and the impact of this on the sensitivity of the method would need to be further investigated. Pfam stores sub-threshold hits which have evalue significance less than 1000. In order to minimize the search with the phylogenetic profile HMM, which despite having the same time complexity is practically substantially slower, only hits less than e-value threshold 100 would be re-scored with the profile HMM. The analysis in this chapter involved scoring 44 families on 4418 clusters, each with up to 5 members and was carried out on a single-processor computer over the course of a few hours, so it would be possible to scale up the number of proteins by 50 (to reach 1m proteins) and the number of families by 100 provided a cluster of computers was available for the analysis.

It would be interesting to investigate the effect of the number of sequences in the tree and the divergence of members of the homologous cluster. Further investigation into how to best incorporate site-specific variation of the +gwF factor in match states is required – one option is to parameterise the +gwF factor relative to the maximum likelihood value obtained using a single model over the entire alignment, both in training and scoring, rather than in absolute terms. Another possibility is to incorporate context dependence via a first order Markov model at the training stage, following the procedure defined in [Yan95] for rate variation. Similarly, further investigation into the inclusion of rate variation is possible, again investigating different priors as well as context dependence via a HMM.

The 'null2' model from HMMER has not been implemented in this chapter, but would possibly provide a useful alternative to the heuristic rule used to limit the path of the phylogenetic HMM.

Context dependent models of substitution have not been incorporated in the model presented here. It would be straightforward and efficient to incorporate context dependent models of amino-acid substitution using the method proposed in [SH04]. These authors discovered a significant improvement in model fit with the introduction of context dependent models, thus suggesting this is a high priority for further investigation. Such a model could reflect correlation in residues between adjacent sites. In fact, context dependence of emission probabilities can be incorporated into the standard profile HMM architecture as well as the phylogenetic HMM architecture. Again, the only difference concerns whether the probability distribution is over residues, or columns of residues. In the profile HMM setting, it would be interesting to model context dependence between adjacent sites which are both emitted by match states, and not otherwise. This technique can be easily incorporated into the scoring algorithms used (such as Viterbi, forward and forward-backward) as well as the HMM building algorithm (such as the MAP architecture algorithm). For example, in the scoring step of the forward algorithm, as shown in equation 1.5, the emission score $P(x_i|\psi_i = M_j)$ in the term $P(x_i|\psi_i = M_j) \cdot P(x_1 \dots x_{i-1}|S \dots M_{j-1}) \cdot P(M_{j-1} \rightarrow M_j)$ is replaced with the context dependent emission score $P(x_i|\psi_i = M_j, x_{i-1})$. This score can be calculated as

$$P(x_i|\psi_i = M_j, x_{i-1}) = \frac{P(x_i, x_{i-1}|\pi_i = M_j, \pi_{i-1} = M_{j-1})}{Z}$$

where the normalising constant is

$$Z = P(*, x_{i-1} | \pi_i = M_j, \pi_{i-1} = M_{j-1})$$

and * is used to indicate missing data, so that the equation for Z turns into a sum over all residues in the case x_i denote residues. In the case of a phylogenetic HMM, this sum can be calculated efficiently using the Felsenstein algorithm as outlined in section 1.3.3. Thus, to incorporate context dependence in both the phylogenetic HMM and the standard profile HMM, the joint emission probabilities $P(x_i, x_{i-1}, M_j, M_{j-1})$ must be estimated. These probabilities can be obtained from the counts observed in the labelled columns of the seed alignment. These counts must be smoothed using priors to avoid over-fitting – one possibility for the pseudo-counts is the cross-product of the normal Dirichlet prior probabilities.

In summary, the phylogenetic HMM has been shown to be a valuable tool in modelling homology, and, provided it is correctly parameterised, can outperform traditional HMMs substantially. Many research directions are open for investigation, each with the potential to further improve performance. Moreover, the techniques of this chapter form the basis for pseudogene and positive selection detection in the next chapter.