Protein Domains: New methods for detection and evolutionary analysis

Lachlan James Murray Coin
Magdalene College
Cambridge

January 2005

A dissertation submitted for
the degree of Doctor of Philosophy
at the University of Cambridge

Preface

The work presented in this dissertation was carried out at the Sanger Institute in Cambridge between October 2001 and December 2004. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. No part of this dissertation nor anything substantially the same has been or is being submitted for any qualification at any other university.

Summary

Protein domains are the structural, functional and evolutionary units of proteins. A useful way to predict the function and structure of a new protein coding gene, or one which is poorly genetically or biochemically characterised, is to identify the domain architecture on the basis of the amino acid sequence, and then infer function and structure from other proteins with similar domain architectures.

The first part of my thesis concerns improving techniques for identification of protein domains from amino acid sequence. I investigate the application of language modelling techniques from speech recognition to integrate contextual information into domain prediction. This takes advantage of the observation that certain combinations of domains are more likely to occur than others. I also investigate using knowledge of the species in which the protein occurs to improve domain prediction, and develop an integrated model of species and domain context. Lastly, I investigate the degree to which protein domains can be identified on alignments of homologous proteins, rather than on the sequences taken individually. This method relies on the development of models of evolution which reflect the structural and functional constraints of conserved sites in the protein domain and using these models to calculate the likelihood that the given protein cluster has been evolving within these structural and functional constraints. I have tested each of these approaches on proteins of known structure, and demonstrated improvements in domain identification in each case.

The second part of my thesis concerns using annotated protein domains to understand the evolution of gene families. I look for cases in which the gene family unambiguously contains a particular protein domain, but also contains proteins which are diverging away from the domain. Using evolutionary models developed in the first part of my thesis which reflect functional/structural constraints at conserved sites, I develop a technique for scoring the degree to which evolution along a branch in the gene tree is constrained by the need to

maintain the structure and function of the protein, and conversely, the likelihood that it is not evolving under these constraints. I have used this approach as the basis of a test for pseudogenes, which has been tested against standard methods for identifying pseudogenes on the manual annotation of human chromosome six. I have also used this approach to develop a test for positive selection, and characterised positive selection in several gene families.

Acknowledgements

Firstly I thank my supervisor, Richard Durbin, for his advice, support and encouragement. I also thank Alex Bateman for useful advice and support. I thank Ewan Birney for many useful suggestions and research directions, particularly with regard to pseudogene detection. Many thanks also to Nick Goldman and Simon Whelan for stimulating discussions and critique of parts of this work. Thanks to Ian Korf, Rob Finn, Sam Griffiths-Jones, Corin Yeats, Mhairi Marshall, Ashwin Hajarnavis, David Carter, Mark Minichiello, Irmtraud Meyer, Marc Sohrmann, Thomas Down, Kevin Howe, Andy Futreal and Mike Stratton as well as all of the other members of the Wellcome Trust Genome campus.

I am grateful to Magdalene College, Cambridge, for a Leslie Wilson studentship and to the Cambridge Australia Trust for a scholarship.

Contents

Chapter 1

Introduction

Furthering our understanding of protein domains is fundamental to our knowledge of both the function of proteins and the evolutionary pressures which have shaped them. Protein domains are the basic structural and functional repertoire from which evolution produces novel function through new combinations as well as modifications within a protein domain. Thus protein domains form an important 'evolutionary crane'[1] [Den95] which have been used during the evolution of complexity.

There has been a recent explosion in the amount of molecular sequence data. As of July 2004, the Integr8 project (`http://www.ebi.ac.uk/integr8/`) contains information from 13 sequenced eukaryotic, 19 archaeal and 150 bacterial genomes. With more genomes in the sequencing pipeline, as well as environmental genome shotgun sequencing [Ven04], the growth of molecular sequence data is set to increase. The number of protein structures is also steadily increasing. Thus, there is increasing amounts of data upon which to build a firmer understanding of protein domain evolution.

Molecular sequence data, used in conjunction with structural data, has already proved to be central in furthering the goal of understanding protein function and evolution. Sequence data has been used for many diverse purposes, including but not limited to:

- detection of evolutionary relationships between proteins;
- clustering proteins into families on the basis of homology;

[1] According to Dennett, an evolutionary crane is a piece of machinery which has itself been generated via evolution, but once created has sped up evolution considerably.

- inferring detailed evolutionary relationships between homologous proteins by reconstructing phylogenetic trees, and describing orthology and paralogy relationships between proteins;
- detecting conserved regions, and inferring potential structural domains;
- inferring branchings in the tree of life;
- detecting highly conserved, functionally important residues;
- detecting correlated sites and inferring interactions;
- detecting pseudogenes;
- detecting positive selection;
- detecting recombination.

Moreover, protein sequence data has led to a better understanding of the evolutionary process both at the molecular level in terms of the mutational and insertion/deletion (indel) processes and at the domain and whole-protein level. Increasingly, with population sequence data becoming available, an understanding of population processes within species is emerging. This includes an understanding of processes responsible for disease and cancer, via studies of somatic mutations and translocations [FCM$^+$04], as well as an understanding of mutational processes which occur in germline cells but are not fixed due to non-viability.

Many of these applications rely on likelihood-based inference and probabilistic models. A probabilistic model parameterises a probability distribution over the space of all possible datasets, and thus assigns a probability (termed likelihood when the data is fixed and we are interested in the probability assigned to it by a particular model) for the particular dataset under investigation. The likelihood-based approach to inference is to construct several probabilistic models, and by comparing the likelihood of the data under one model vs another decide which best explains the data. One example of this approach is phylogenetic tree reconstruction, where each possible tree topology (together with a fixed model of molecular evolution) parameterises a different probabilistic model over alignments. An advantage of likelihood based inference is its flexibility with regard to unknown parameters of the model,

such as branch lengths in the case of tree reconstruction. Unknown parameters can be estimated as those which maximise the likelihood of the data (the maximum likelihood approach). Different models (in this case different tree topologies) can be compared on the basis of the maximum likelihood values. One pitfall of this approach is that adding parameters to a model will never decrease the maximum likelihood value and in most cases increase it, so that models with many parameters have a higher maximum likelihood than those with fewer parameters. A related problem, sometimes called the many parameter trap, occurs when parameters are introduced in relation to the size of the dataset, for instance site-specific parameters for an alignment. The Bayesian approach [Jay03], deals with these problems by comparing models via the integral of the likelihood with respect to a prior probability distribution over parameter space.

In this introductory chapter I will review some of the literature concerning protein domains, introduce mathematical models which are used to identify protein domains, as well as review the literature on parameterising evolutionary models. Accurate protein alignments and trees are fundamental to the approaches taken in this thesis; however, I will not review the vast literature surrounding these two topics, as no new techniques for either alignment or tree reconstruction will be presented in this thesis.

1.1 Protein Domains

The strict definition of a protein domain is a distinct structural and evolutionary unit of a protein. This can include the entire protein, in the case of a single domain protein, or it can include a fragment of a multi-domain protein which is observed to also occur in different contexts on different proteins. This definition assumes that the protein domain has a unique stable three dimensional structure, so the definition excludes natively disordered regions, although these regions may still be of functional importance – it has been shown that eukaryotes have 3-5 times more proteins with long regions of no secondary structure than other kingdoms [Ros02]. Not all evolutionary units of proteins have had their structures elucidated, and so in this thesis the definition will be weakened to include evolutionarily conserved protein fragments which are not homologous to an entry in the protein data bank (PDB). A domain architecture is defined as in [VBB$^+$04] as the linear arrangement of protein domains – two proteins have the same architecture if and only if they have the same domain composition,

and the domains are arranged from N-terminal to C-terminal in the same way, with linker regions of up to 30 residues allowed between domains. Inserts can be accommodated in this definition by requiring the same nesting pattern. The term *superfamily* will be used as defined in SCOP [HMBC97], which is as a collection of all domains which are related by descent. *Fold* will also be used as in SCOP to denote the collection of domains which have structural similarity, but may not be evolutionarily related.

A hypothesis forwarded in [LPR01, SL03] is that domains are the result of fusions of *antecedent domain sequences* (ADSs) which are ancient peptide sequences. The authors argue that this hypothesis explains the similarity of motifs present in different protein folds better than the alternative hypothesis of convergent evolution. Under the ADS hypothesis, motifs rather than domains are the atomic units of protein evolution. This hypothesis remains unproven and controversial. It is not considered further in this thesis.

A compelling view to emerge from several structural biology studies is that multi-domain proteins can be described at a high level by their protein domain architectures, and that most proteins are multi-domain (two-thirds in prokaryotes [TPC98], more in eukaryotes [Ger98]). Transferring functional annotation between single-domain proteins from the same superfamily can be done with 68% accuracy, whereas for multi-domain proteins sharing a domain combination the accuracy is 81% and for multi-domain proteins with almost complete residue coverage and identical domain architectures the accuracy rises to 91% [HG01]. These authors also found that ignoring the multi-domain structure of a multi-domain protein is particularly perilous for functional transfer: only 35% of multi-domain protein pairs sharing a single domain have the same function. Aloy and co-workers have shown that the geometry of interaction is generally conserved for homologues with sequence identity above 30% but is not conserved between members of the same fold without evidence of shared ancestry, although they also provide examples of very close homologues not preserving the interaction, and distant homologues which strongly preserve the geometry of interaction [ACSR03]. Thus, it appears that a substantial amount of protein function can be understood via an understanding of the structure and function of representative multi-domain proteins. On the basis of this principle, protein targets have been prioritised for structure determination in structural genomics projects [Bre00, AHT03, VBB$^+$04]. As well as predicting function, domain architectures can be used to assist cellular localization prediction [MSBP02].

Novel proteins are formed during evolution by duplication and recombination. Duplication gives rise to proteins which are freer to diverge and evolve new functions. While this process often leads to the formation of a pseudogene, it is also the main source for the creation of new genes [PP02]. It has been observed that the degree to which different domain superfamilies have been duplicated and subsequently maintained in the genome varies substantially, and this results in a power law distribution of domain superfamily occurrence[QLG01]. Recombination can lead to the formation of novel protein domain architectures, by either fusing genes or by shuffling exons via intronic recombination, leading to domain shuffling [KZNL02]. Insertions of one domain into another account for 9% of non-redundant domain architectures in the PDB – a small but likely significant subset of protein structures [ASHS04].

Apic et al. have demonstrated in [AGT01b, AGT01a, AHT03] that the observed pattern of domain combinations is highly non-random. In fact, a few domain superfamilies are highly versatile in forming multi-domain proteins with a variety of other protein domains, while most have only a single partner. A random model of recombination would predict a much flatter distribution. This suggests that the protein domain combinations which are observed are strongly selected. The authors also showed that multicellular organisms have more sequences and more domain families participating in tandem repeats.

A study of the geometry of domain combinations of Rossman domains [BC02], which are highly versatile in forming multi-domain proteins, has demonstrated that proteins which have the same domain architecture have evolved from the same ancestor. The authors confirmed the observation in [AGT01b] that superfamily combinations almost always occur in the same sequential order, and identified only 2% of cases in which both sequential orders of a domain pair occur. Moreover the authors discovered no structural reason for a particular order, and conclude the observed order is due to the single recombination event which occurred to create the combination. The authors also found extensive conservation of the relative geometry of the domain pair provided the order was conserved, and not otherwise.

Vogel et al. showed in [VBB$^+$04] that some domain combinations occur in many different domain contexts, while preserving the spatial relationships and the linear order of the combination. Such a combination is called a *supra-domain*. Two particular types of supra-domains, were identified based on the geometry of the interaction: *interface* supra-domains have an interface which is critical to the biochemical activity of the protein, whereas the

domains in *separate* supra-domains have biochemically separate but complementary activities. An example provided by the authors of a separate supra-domain is the P-loop nucleotide triphosphate hydrolase domain, which binds and hydrolyses GTP in order to drive a conformational change that is transmitted to its supra-domain partner. In the example provided for the interface supra-domain, both partners of the supra-domain are directly involved in the same cofactor binding interactions. As with domains, some supra-domains have been duplicated substantially, while others have only a few copies. A few supra-domains are very versatile with respect to other domain contexts, while most occur in only a few domain contexts. Vogel et al. note that the majority (64%) of single SCOP domains occur in all three kingdoms of life, whereas most two-domain combinations (96%) and most duplet supra-domains (85%) do not occur in at least one kingdom. Moreover, it was observed in [AGT01b] that of those superfamilies which do participate in kingdom specific domain combinations, significantly more are from all three kingdoms than not. Thus, while domains are in general ancient and common to the last common ancestor of three kingdoms of life, domain combinations have occurred largely within the evolution of specific kingdoms.

Superfamilies often display a wide diversity of function. 25% of CATH superfamilies contain members of different enzyme types [TOT01]. A recent evolutionary study into how evolution generates functional diversity from similar structures demonstrates the economy of nature: structurally conserved residues are kept intact, including residues important for cofactor binding [BBT03]. As noted in [VBK$^+$04], these studies have to a large extent focussed on residue changes within the protein domain structure, and not investigated the effect of domain context in modulating the behaviour of component protein domains. As shown in [HG01] context is of vital importance in correctly annotating domain function. Many examples of context-modulated function have been observed. One example given in [VBK$^+$04] is the winged helix domain, which is typically a DNA binding domain, and in many cases is combined with a regulatory domain, but can also be combined with a catalytic domain so that the protein function, while still acting on DNA, is changed. This is termed a *syntactic* change in [VBK$^+$04]. A more radical modulation of behaviour is observed in cases where the winged helix domain no longer has any DNA binding activity, but instead acts as a substrate specificity pocket, which the authors term a *semantic* change. Elucidating the range and extent of context-dependent domain functional change is an important area for future research

in structural biology.

1.2 Sequence based protein domain detection

An important problem is to identify the protein domain architecture of novel protein sequences, for example from genome sequencing projects. Once the domain architecture is determined, it may be possible to transfer functional annotation from biochemically and genetically characterised homologues, as well as to infer structurally important residues as well as regions of interactions with other proteins.

One potential approach is to use pairwise comparison techniques, such as BLAST [AMS+97], and to consider pairwise similarity scores with all members of a domain family. However, methods which use a profile are more sensitive than methods which look for pairwise homology [PKB+98] . A profile summarises the site-specific residue frequencies of a multiple alignment of known members of a domain family, termed the *seed* alignment. The simplest profile method is the position specific scoring matrix (PSSM) which constructs a probability distribution at each of the m sites in the alignment and does not allow gaps. A novel sequence is scanned by the PSSM by calculating at each site the probability that the next m residues in the sequence have been emitted by the corresponding distributions in the PSSM, and the highest score is taken as the overall score.

Profile hidden Markov models (HMMs) formalise PSSMs as probabilistic models and improve its sensitivity by allowing insertions and deletions relative to the profile. A profile HMM (labelled $\mathbf{D}$) is a probabilistic model which parameterises a probability distribution over all possible sequences (labelled $x = x_1 x_2 \dots x_n$). The basic idea is that the profile HMM constructed for a domain family assigns high probability to sequences which are homologous to the domain family (or more strictly contain a homologous fragment), and a low probability to non-homologous sequences. One problem with using the probability of the sequence as a score, regardless of the precise details of the profile HMM, is that long sequences (above a certain length threshold) will inevitably have lower probability than shorter sequences. From the point of view of Bayesian inference another problem is that the correct probability upon which to base the inference is the posterior probability

$$P(\mathbf{D}|x) = \frac{P(x|\mathbf{D})P(\mathbf{D})}{\sum_{\mathbf{D}'} P(x|\mathbf{D}')P(\mathbf{D}')}, \tag{1.1}$$

where the denominator is a sum of the likelihood under all possible domain models $\mathbf{D}'$ multiplied by the prior probability of that model, which is expensive to calculate. Both problems can be overcome by introducing as an alternative hypothesis a background probability distribution $\mathbf{R}$ over sequence space, and calculating the ratio of posterior probabilities, in which case the term involving the sum in the previous equation cancels out. It will be convenient to work with log probabilities

$$\log \frac{P(\mathbf{D}|x)}{P(\mathbf{R}|x)} = \log \frac{P(x|\mathbf{D})}{P(x|\mathbf{R})} + \log \frac{P(\mathbf{D})}{P(\mathbf{R})} \tag{1.2}$$

where the first term is called the log-odds score, and the second term is the log ratio of the prior probabilities of the models, and can be thought of as a threshold on the log-odds score. As long as the background model has a similar distribution over protein lengths, the scores should be normalised with respect to protein length.

The log-odds score is used to rank sequences and apply a threshold cut-off such that all sequences scoring above a threshold are taken to be members of the family. It is a useful measure for inferring relative similarities of sequences to the protein domain, but does not provide similarity scores in absolute terms, or at a particular level of significance. Empirical significance values can be obtained by calculating the log-odds scores for sequences randomly sampled from the background model. In this way a distribution of scores for random 'proteins' is obtained, and the significance level of a sequence log-odds score can be obtained by counting the fraction of random sequences which score higher. However, to get an accurate significance value for high scoring sequences in this way many hundreds of thousands of random sequences need to be scored[2]. To get around this problem it has been observed that the distribution of random scores from an profile HMM follows an extreme value distribution, which can be successfully parameterised with much less data (HMMER uses 5000 sequences)[DEKM98].

It is useful at this point to parameterise the profile HMM. I start with a description of Markov models, which will be useful at other points in this thesis.

Discrete Markov Models

Let Σ denote a state space, $\{Y_i : [0,1] \to \Sigma\}_{i=1,2\ldots}$ denote a series of random variables each of which takes values in the state space Σ. Let μ denote the uniform probability distribution

[2]particularly if the significance value is to be ascertained to the level required for annotation in Pfam, which is a significance of less than 1/N where N is the number of protein sequences scored, currently around 1.5m.

on the interval $[0,1]$ so that

$$P(Y_i = y) = \mu(\{r \in [0,1] : Y_i(r) = y\})$$

defines a probability distribution on Σ.

A k^{th} order Markov model is a probabilistic model with the property that the state at position i is only dependent on the preceding k states:

$$P(Y_i = y|Y_{i-1} = y_1, \ldots, Y_1 = y_{i-1}) = P(Y_i = y|Y_{i-1} = y_1, \ldots, Y_{i-k} = y_k). \quad (1.3)$$

In the simplest cases of a homogeneous Markov model these probabilities are independent of position in the chain

$$P(Y_i = y|Y_{i-1} = y_1, \ldots Y_{i-k} = y_k) = P(Y_{i'} = y|Y_{i'-1} = y_1, \ldots Y_{i'-k} = y_k),$$
$$\forall 1 \leq i, i' \leq n, \text{ and } y_1 \ldots y_k \in \Sigma. \quad (1.4)$$

So all that is needed to specify a Markov model is to specify the states, and the transition probabilities $P(y|y_1, \ldots y_k)$ between states. In the case of a first order Markov model, I will also write $P(y_1 \rightarrow y)$ for the transition probability. It will be useful to include a special state S in which the model starts and one for which it terminates, T. The probability distribution parameterised by a Markov model is over chains of states, which will be of finite but unbounded length provided there is a path with non-zero transition probability from every state in the model to the end state.

A hidden Markov model is a Markov model in which some states $y \in \Sigma$ themselves are allowed to be random variables, $y : [0,1] \rightarrow \Upsilon$, taking values in the space Υ. These are termed emission states. It is also useful to allow states which are not random variables, which includes the start and terminate states. The emission probability distribution for emission state y over Υ is then defined as

$$P(u|y) = \mu(\{r \in [0,1] : y(r) = u\}).$$

In the case of a profile HMM, the state space Υ will be amino-acids, codons or nucleotides. These states are hidden in the sense that they are not observed in the data, but are internal states of the overall probabilistic model. They are introduced in order to provide flexibility in parameterising an appropriate probability distribution over sequence space. So, in order to

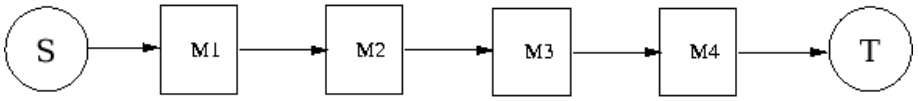

Figure 1.1: Basic architecture for profile hidden Markov model for an alignment with four amino acids and without gaps. Each M_i corresponds to a column in the multiple alignment, and emits over a distribution of amino acids. B, E correspond to begin and end states.

specify a Hidden Markov model, the hidden states Σ, transition probabilities $P(y|y_1, \ldots y_k)$ and emission probabilities $P(u|y)$ must be specified.

The PSSM reframed in this framework is shown in figure 1.1. The state space Σ consists of match states M_j for each column in the seed alignment as well as the begin and end states. Each match state emits in the space Υ of amino-acids.

A profile HMM from HMMER is shown in figure 1.2, taken from [Edd03]. In fact this comprises two HMMs – the domain model HMM and the null model HMM. Both HMMs are first order HMMs. For an alignment consisting of m conserved columns (which can be defined as columns with less than 50% gaps) the domain model state space Σ includes m match and insert emission states M_j and I_j as well as m non-emission states D_j. Σ also includes an N-terminal, C-terminal and inter-domain emission state N, C, J respectively as well as domain begin and end non-emission states B, E. The M_j emit residues according to a probability distribution estimated from the counts observed in a particular conserved column of the alignment. The insert states I_j emission probability is calculated from all insert states in the Pfam database. The match to insert transitions specify the 'cost' of opening a gap relative to the protein domain, and the insert to insert transitions specify the cost to maintain the gap, which is an affine gap scoring scheme. The delete states allow for domain states to be skipped, with a penalty controlled by match to delete and delete to match transitions. The N, C, J states allow the model to score full length proteins by allowing for N- C-terminal and inter domain regions respectively. These states emit according to a background model of residue usage in proteins. The null model HMM consists of a single emission state, which emits according to the background distribution of protein residues.

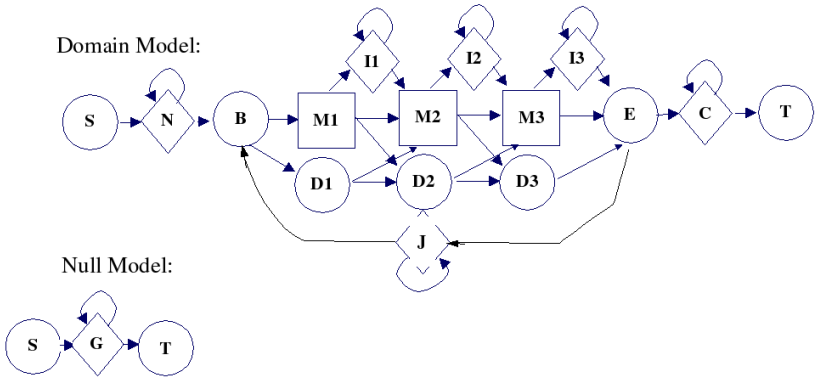

Figure 1.2: Diagram of profile Hidden Markov Model. States which emit symbols are shown as squares or diamonds; circles do not emit symbols. The core model consists of match states – which model conserved residues of a protein family; insert states – which allow for segments of the query sequence not present in the protein family; and delete states – which allow for deletions of conserved residues in the protein family from the query sequence. The model consists of several flanking states, which allow for local matches and multiple hits. The transition to the J state allows for multiple hits of the model to a single query sequence. The N, J and C states are analogous to insert states, but occur before, between and after the model hit respectively. The B and T states are states used to begin and terminate a hit to the query, while S and E states are formally required as overall start and end states. To obtain the log-odds score we also require a null model. The null model consists of an null emission state G which emits according to a background distribution, and can loop back to itself, or transition to the end state. Effectively the transitions of the null model act to negate the otherwise intrinsic penalty for scoring longer query sequences.

Calculating the likelihood of a profile HMM

The forward algorithm can be used to calculate the likelihood of the sequence given the profile HMM. Note that the profile HMM can generate a particular sequence with many different paths through the HMM architecture, although only a few will have high posterior probability[DEKM98]. The forward algorithm naturally sums over all possible paths, in contrast to the Viterbi algorithm (which will not be used, and so not described in more detail, but see [DEKM98]) which calculates the probability of the mostly likely path to have generated the sequence. The Viterbi algorithm is employed by HMMER, and has the advantage that all calculations can be done in log probability space and that only summation is required. The forward algorithm, on the other hand requires working in probability space with multiplication, which can lead to underflow errors if an adaptive scaling algorithm is not employed.

The forward algorithm proceeds by iteratively filling in eight matrices $P(x_1 \dots x_i|S \dots M_j)$, $P(x_1 \dots x_i|S \dots I_j)$, $P(x_1 \dots x_i|S \dots D_j)$, $P(x_1 \dots x_i|S \dots C)$, $P(x_1 \dots x_i|S \dots J)$, $P(x_1 \dots x_i|S \dots N)$, $P(x_1 \dots x_i|S \dots B)$, $P(x_1 \dots x_i|S \dots E)$ which are the partial probabilities of the HMM emitting subsequence up to and including the i^{th} residue and the ending in the j^{th} match, delete, insert or the C, J, N, B, E states respectively. Let ψ_i denote the state which emitted residue x_i. If the domain begin state B is interpreted as also being M_0, these scores can be calculated

recursively using

$$
\begin{aligned}
P(x_1 \dots x_i | S \dots M_j) &= P(x_i|\psi_i = M_j) \cdot \begin{pmatrix} P(x_1 \dots x_{i-1}|S \dots M_{j-1}) \cdot P(M_{j-1} \to M_j) \\ + P(x_1 \dots x_{i-1}|S \dots I_{j-1}) \cdot P(I_{j-1} \to M_j) \\ + P(x_1 \dots x_i|S \dots D_{j-1}) \cdot P(D_{j-1} \to M_j) \end{pmatrix} \\
P(x_1 \dots x_i | S \dots D_j) &= \begin{pmatrix} P(x_1 \dots x_i|S \dots M_{j-1}) \cdot P(M_{j-1} \to D_j) \\ + P(x_1 \dots x_i|S \dots D_{j-1}) \cdot P(D_{j-1} \to D_j) \end{pmatrix} \\
P(x_1 \dots x_i | S \dots I_j) &= P(x_i|\psi_i = I_j) \cdot \begin{pmatrix} P(x_1 \dots x_{i-1}|S \dots M_j) \cdot P(M_j \to I_j) \\ + P(x_1 \dots x_{i-1}|S \dots I_j) \cdot P(I_j \to I_j) \end{pmatrix} \\
P(x_1 \dots x_i | S \dots B) &= \begin{pmatrix} P(x_1 \dots x_{i-1}|S \dots C) \cdot P(C \to B) \\ + P(x_1 \dots x_{i-1}|S \dots J) \cdot P(J \to B) \end{pmatrix} \\
P(x_1 \dots x_i | S \dots E) &= \begin{pmatrix} P(x_1 \dots x_{i-1}|S \dots M_m) \cdot P(M_m \to E) \\ + P(x_1 \dots x_{i-1}|S \dots I_m) \cdot P(I_m \to E) \\ + P(x_1 \dots x_i|S \dots D_m) \cdot P(D_m \to E) \end{pmatrix}
\end{aligned}
\tag{1.5}
$$

$$
\begin{aligned}
P(x_1 \dots x_i | S \dots C) &= P(x_i|\psi_i = C) \cdot \begin{pmatrix} P(S \to C) \text{ if } i = 1 \\ P(x_1 \dots x_{i-1}|S \dots C) \cdot P(C \to C) \end{pmatrix} \\
P(x_1 \dots x_i | S \dots J) &= P(x_i|\psi_i = J) \cdot \begin{pmatrix} P(x_1 \dots x_{i-1}|S \dots E) \cdot P(E \to J) \\ + P(x_1 \dots x_{i-1}|S \dots J) \cdot P(J \to J) \end{pmatrix} \\
P(x_1 \dots x_i | S \dots N) &= P(x_i|\psi_i = N) \cdot \begin{pmatrix} P(x_1 \dots x_{i-1}|S \dots N) \cdot P(N \to N) \\ + P(x_1 \dots x_{i-1}|S \dots E) \cdot P(E \to N) \end{pmatrix}
\end{aligned}
\tag{1.6}
$$

The overall likelihood of the domain matching the sequence is equal to the score for the terminal state: $P(x|\mathbf{D}) = P(x_1 \dots x_n|S \dots C)P(C \to T)$. The probability of the sequence being emitted by the null model can be calculated as

$$P(x|\mathbf{R}) = P(G \to G)^n \cdot P(G \to T) \cdot \prod_i P(x_i|G).$$

The overall log-odds score is then calculated as $\log P(x|\mathbf{D}) - \log P(x|\mathbf{R})$.

It is also possible to calculate the backward partial scores $P(x_{i+1} \dots x_n|M_j \dots T)$, $P(x_{i+1} \dots x_n|I_j \dots T)$, $P(x_{i+1} \dots x_n|D_j \dots T)$, $P(x_{i+1} \dots x_n|C \dots T)$, $P(x_{i+1} \dots x_n|J \dots T)$,

$P(x_{i+1}\dots x_n|N\dots T)$, $P(x_{i+1}\dots x_n|B\dots T)$, $P(x_{i+1}\dots x_n|E\dots T)$ which are the partial probability of the HMM emitting the subsequence from residue $i+1$ to n and coming from j^{th} match, delete, insert or the C, J, N, B, E states respectively. The backward algorithm proceeds iteratively from the C-terminal to N-terminal end of the sequence (equations only shown for $P(x_{i+1}\dots x_n|M_j\dots T)$). Again to simplify the equations, the E state is interpreted to be the same state as M_{m+1} (where m is the number of match states).

$$P(x_{i+1}\dots x_n|M_j\dots T) = \begin{pmatrix} P(x_{i+1}|\psi_{i+1}=M_{j+1})\cdot P(x_{i+2}\dots x_n|M_{j+1}\dots T)\cdot P(M_j\to M_{j+1}) \\ +P(x_{i+1}|\psi_{i+1}=I_j)\cdot P(x_{i+2}\dots x_n|I_j\dots T)\cdot P(M_j\to I_j) \\ +P(x_{i+1}\dots x_n|D_{j+1}\dots T)\cdot P(M_j\to D_{j+1}) \end{pmatrix} \tag{1.7}$$

Using the definition of the partial forward and backward scores

$$P(x|\psi_i=M_j) = P(x_1\dots x_i|S\dots M_j)P(x_{i+1}\dots x_n|M_j\dots T)$$

and hence, using Bayes' theorem

$$P(\psi_i=M_j|x) = \frac{P(x_1\dots x_i|S\dots M_j)\cdot P(x_{i+1}\dots x_n|M_j\dots T)}{P(x)} \tag{1.8}$$

This provides a way of calculating the posterior probabilities.

Building the profile HMM

A profile HMM is constructed from a seed alignment of homologous sequences. The conserved columns in the alignment correspond to match states, and the other columns correspond to insert states. Which columns to label as match states and which to label as insert states can be either resolved heuristically (by labelling all columns with greater than 50% gaps as insert states), or using the maximum *a posteriori* architecture algorithm to find the profile HMM which optimises the likelihood of the seed alignment (see [DEKM98]).

The observed residues in a conserved column are used to estimate the emission probability distribution for that column, and the observed transitions are used to estimate the transition probabilities. The most straightforward approach is to use the observed frequencies as the emission and transition probabilities, however this assumes that the rows are independently sampled from the target probability distribution, which is not true. Some sequences

are closely related to each other whereas some are more distantly related. To get around this problem, various weighting schemes have been proposed [SA90, HH92, GSC94, KM95, HH96], which all in effect try to adjust the weights of the rows included in the counts so that sequences from parts of sequence space which are not well sampled in the seed contribute more, and sequences from parts of sequence space which are well sampled each contribute less to the overall estimation of emission and transition probabilities. Even after re-weighting the sequences to remove sampling biases, another problem is that the space of possible domain family members has been inadequately sampled, particularly for small seed alignments. This problem has been addressed using a mixture of Dirichlet priors, in which the emission probability is taken to be the mixture of k posterior probabilities, each of which are calculated as the posterior probability of the residue frequencies given the column and one of k Dirichlet priors. The mixture co-efficients are calculated as the posterior probability of each mixture component given the observed counts in the column. See [SKB$^+$96] for more details. This problem can also be addressed using the tree HMM introduced in the next section.

Other methods and extensions

One improvement made to profile HMM methods in recent years has been the introduction of iterations, whereby an initial sequence is used to build a profile HMM which is searched against a database and significant hits are used to rebuild the profile HMM. This process is then repeated until no further hits are found. This is the strategy used in SAM [HK96] and PSI-blast [AMS$^+$97].

Several profile-profile comparison techniques have been proposed in recent years [Pie96, YL02, SBG03, Sd04]. The motivation for these methods is that profile-sequence comparisons are more sensitive than sequence-sequence comparisons and so profile-profile comparisons might be expected to be even more sensitive in detecting weak homology. Indeed, these methods appear to be more sensitive than profile HMMs. The method proposed in [Pie96] was developed for the comparison or conserved ungapped alignments from the BLOCKS [HHP99] database and so does not allow gaps. PROF_SIM [YL02] and COMPASS [SBG03] both use allow gaps via a Smith-Waterman local alignment algorithm with column similarity scores based on Jensen-Shannon entropy and a symmetric log-odds ratios respectively. Söding [Sd04] generalizes the profile HMM framework to compare two profiles.

Several discriminative support vector machine (SVM) approaches have been applied to homology detection. Jaakkola et al. [JDH00] proposed a method for using profile HMMs to derive a kernel function in a SVM classifier. The motivation for this approach is that profile HMMs are trained using positive training examples only. A discriminative model, which takes both positive and negative training examples, should perform better. A support vector machine is a discriminative model which can be thought of as a classifier which can be trained to discriminate points in a high dimensional space. A support vector machine relies on a kernel function $K(x, x^k)$ which can be thought of as a measure of similarity between a sequence x and training example x^k, which can be either positive or negative. Considering the profile HMM as a likelihood function over sequence space, Jaakkola et al. define a vector U_x called a Fisher score, which is the partial derivative of the log-likelihood score at the sequence X with respect each of the parameters of the profile HMM. The vectors U_x and U_{x^k} are then used to derive the kernel function via a formula presented in [JDH00]. Leslie [LEC+04] et al. have proposed a string kernel for protein classification which maps a protein to a vector U_x called its 'k-spectrum' which is the set of all k-mers contained in the protein. The kernel function is then a vector function of U_x and U_{x^k} as before.

1.3 Models of sequence evolution

The genomic sequence of cellular organisms is in constant flux. During cell division replication introduces copying errors of which some fraction remain uncorrected. Recombination leads to exchange of genomic material between alleles in the case of eukaryotes, and between different species in bacteria. Processes such as non-allelic homologous recombination lead to genomic rearrangements including deletion, inversion, translocation and duplication. Retroviral elements are integrated into genomic DNA. Certain proteins promote genomic mutation via processes such as class switch recombination and somatic hypermutation, particularly in certain cell types such as germinal centre B cells where mutation is required in order to generate a diverse set of antibodies. External factors such as radiation also lead to genomic mutation. These mutations can occur either in somatic cells, in which case they are not passed to the next generation, or in germline cells. Most germline cell mutations are hypothesised to be neutral [Kim83], however some will be deleterious and therefore not survive. Rarely, mutations will be advantageous and selected for, resulting in a selective sweep through the

population.

The DNA or RNA sequence of viruses is typically under an even higher rate of flux than for cellular organisms due to copying errors during replication in the host cell as well as processes such as host-mediated hypermutation. In many cases these errors are not repaired by the host cell DNA repair machinery and so the rate at which mutation occurs is significantly higher. Retrotransposition is a particularly error prone step leading to high rates of mutation in retroviruses. Viruses often have particular features which enhance the rate of mutation.

Thus genomic sequences change over time and these changes can be modelled at different levels: within a single cell during the cell's lifetime; progressively during transmission from parent to daughter cells; within a population of cells (for example during the progression of a tumour, or a bacterial culture); transmission from a multicellular parent to offspring organism; within a population of multi-cellular organisms; or between different species of organisms. Each of these levels requires a different level of resolution. For instance when modelling difference between species, differences within a population will typically be ignored and the most frequent allele will be taken as representative for that species. Due to duplication, different segments of genomic DNA will be related to each other via descent, and so sequence evolution can be modelled within a single genome.

1.3.1 Probabilistic models of sequence evolution

Let Υ describe a state-space, which initially is taken to be all of sequence space, and let $u, v \in \Upsilon$ be elements of this state space. Let $|u|$ denote the length of a sequence. A probabilistic model of sequence evolution, denoted by $\mathcal{E}$, is a model which describes a probability distribution $P_{\mathcal{E}}(x^t = u)$ over sequences at each time $t \geq 0$. This can be used to describe the transition probability $P_{\mathcal{E}}(x^{t+\triangle t} = v | x^t = u)$ of observing a sequence v at time $t + \triangle t$ given that u was observed at time t. A general probabilistic model of evolution would need take into account all of the mutational processes described above, including point mutation, insertion, deletion, recombination, gene conversion and translocation. Moreover, such a model would also need to describe how the rates of each of these processes change with respect to position in the genome and time. This is clearly a very challenging task.

Given the complexity of the task, why bother constructing probabilistic models of evolution? The answer principally lies in the usefulness of the likelihood $P(\{x^k\}|\mathcal{E}, T)$ of a cluster

of homologous sequences $\{x^k\}$. If $\mathcal{E}$ is fixed, the likelihood can be used as a criterion for evaluating how well the tree fits the data, for finding optimal branch lengths, and for parameterising a posterior distribution over all possible trees. This approach can also be used to find evidence for recombination [HW01]. If T is fixed, the likelihood can be used to compare different evolutionary models, and so gain quantifiable insight into the evolutionary process itself. This approach is the basis for tests for pseudogenes and positive selection, which will be further described in section 4.1 and 4.2. Probabilistic models of evolution can also be used to align sequences [MD95, HB01, Hol03, MLH04].

Whole sequence evolutionary models

The first standard simplifying assumption is that $\mathcal{E}$ is a continuous-time Markov process over the state space Υ. This corresponds to assuming that the transition probability, $P_{\mathcal{E}}(x^{t+\triangle t} = u|x^t = v)$ is independent of t, or that evolution is homogeneous with respect to time. This simplifying assumption is clearly violated in many circumstances. One example is if a functional gene becomes non-functional, in which case the evolutionary constraints on the sequence change. One way to improve the realism of models with respect to this assumption is to have different models on different parts of the tree, as in Chapter 4. If the process $\mathcal{E}$ is assumed to be Markov, then the time evolution of the probability distributions $P_{\mathcal{E}}(t)$ is described by the differential equation

$$\frac{dP_{\mathcal{E}}(t)}{dt} = P_{\mathcal{E}}(t)\mathbf{Q}r \tag{1.9}$$

where $\mathbf{Q}$ is a fixed rate matrix describing the instantaneous transition rate between states in the state space so that $\mathbf{Q}_{u,v}$ the instantaneous rate of transition between states u and v, and satisfies

$$Q_{u,u} = -\sum_{v\in\Upsilon, v\neq u} Q_{u,v},\ \forall u \in \Upsilon. \tag{1.10}$$

An arbitrary scaling constant r representing the rate of evolution has been included for future reference and can be assumed at this stage to be equal 1. This rate matrix is scaled so that the average rate of substitution at equilibrium is 1:

$$-\sum_{u\in\Upsilon} \pi_u \mathbf{Q}_{u,u} = 1 \tag{1.11}$$

which reduces by 1 the number of parameters required to specify a rate matrix and implies that rt is measured in units of expected substitutions per site. The solution to the differential

equation is given by

$$P_{\mathcal{E}}(t) = P_{\mathcal{E}}(0)e^{\mathbf{Q}rt}, \tag{1.12}$$

where $e^{\mathbf{Q}rt}$ is the matrix exponential. The transition probabilities are given by

$$P_{\mathcal{E}}(x^{t+\triangle t} = v | x^t = u) = \left[e^{\mathbf{Q}r\triangle t}\right]_{u,v}. \tag{1.13}$$

The rate matrix $\mathbf{Q}$ is assumed to be irreducible, which requires that the there is a non-zero probability of transitioning over some time $\triangle t > 0$ between any two states u and v, and recurrent, which requires that the probability of visiting each state at least N times in an infinite amount of time is equal to 1 for all positive integers N. A stationary distribution π of $\mathbf{Q}$ is a distribution for which $\pi\mathbf{Q} = 0$. For a recurrent, irreducible rate matrix $\mathbf{Q}$ a stationary distribution π exists and is unique up to scalar multiplication (see [Nor97] for further details). Thus it makes sense to talk about the stationary probability distribution of $\mathcal{E}$, and so I will write $\mathcal{E} = (\pi, \mathbf{Q})$. Another common simplifying assumption is reversibility, which implies that the instantaneous flux between residues is the same in both directions

$$\pi_u \mathbf{Q}_{u,v} = \pi_v \mathbf{Q}_{v,u}. \tag{1.14}$$

This halves the number of parameters required to estimate the rate matrix $\mathbf{Q}$. There is no a priori reason to expect that evolution is reversible, although there is some evidence [AB97] that DNA evolution in many cases is close to reversible. As observed in [HD98], insertion events may be short and frequent, while deletion events long and rare, which would lead to a violation of the reversibility assumption. Observe that if eq. 1.14 holds then

$$\mathbf{S}_{u,v}(f) = \pi_u^f \pi_v^{f-1} \mathbf{Q}_{u,v} \tag{1.15}$$

is symmetric, i.e. $\mathbf{S}_{u,v}(f) = \mathbf{S}_{v,u}(f)$. This defines a single parameter family of symmetric matrices for $\mathbf{Q}$. The parameter f, described in [GW02], is called the +gwF parameter. $\mathbf{S}(0)$ is referred to as an exchangeability matrix. A symmetric matrix can be expressed in the form

$$\mathbf{S}(f) = \mathbf{N}(f)\mathbf{D}(f)\mathbf{N}(f)^T \tag{1.16}$$

where $\mathbf{D}(f)$ is a diagonal matrix, $\mathbf{N}(f)$ is an orthonormal matrix and $\mathbf{N}(f)^T$ is the matrix transpose (see [Lay94] for further details). Let Π be a diagonal matrix with entries $\Pi_{u,u} = \pi_u$.

If I restrict to $f = 1/2$

$$\mathbf{Q} = \Pi^{-1/2}\mathbf{S}(1/2)\Pi^{1/2} \tag{1.17}$$

$$= \mathbf{N}'(1/2)\mathbf{D}(1/2)\mathbf{N}'(1/2)^{-1} \tag{1.18}$$

where

$$\mathbf{N}'(1/2) = \Pi^{-1/2}\mathbf{N}(1/2). \tag{1.19}$$

Thus, $\mathbf{Q}$ is diagonalizable and the matrix exponential can be calculated as

$$e^{\mathbf{Q}r\triangle t} = \Pi^{-1/2}\mathbf{N}(1/2)e^{\mathbf{D}(1/2)r\triangle t}\mathbf{N}(1/2)^T\Pi^{1/2} \tag{1.20}$$

which provides a fast way to calculate the matrix exponential – first calculate the orthonormal decomposition of $\mathbf{S}(1/2)$ and then for all $t > 0$ the matrix exponential step just consists of exponentiating the diagonal entries of $\mathbf{D}(1/2)$ and two matrix multiplication steps. The columns of $\mathbf{N}'(1/2)$ are the eigenvectors of $\mathbf{Q}$ and can be interpreted as directions in state space in which information about the ancestral sequence is lost through evolution. The corresponding diagonal entries are the rate at which the information is lost.

Most methods also assume stationarity, which says that $P_{\mathcal{E}}(0) = \pi$, or that the system is at equilibrium at time 0. There is also no particular reason to expect stationarity to hold in general. In particular, a universal trend of amino acid loss and gain has been observed in all kingdoms of life [JKA$^+$], with Cys, Met, His, Ser, and Phe gaining and Pro, Ala, Glu, and Gly losing frequency. Moreover G+C content varies widely between genomes, again indicating the stationarity does not hold in general.

For proteins of known structure, Robinson et al. [RJK$^+$03] parameterise a whole sequence model for sequences evolving in such a way as to preserve this structure. The authors restrict their evolutionary model to DNA sequences of length N and allow only one position in the sequence change in any given mutation event, so the rate matrix $\mathbf{Q}$ is of size 4^N x 4^N and each row has no more than $3N$ non-zero off-diagonal entries. The rate of amino-acid changing substitutions is based on the propensity of the mutation to change the structure using a sequence-structure compatibility score. Transition/transversion and non-synonymous/synonymous substitution rate ratios are used to determine the underlying DNA mutation rate within these constraints.

Substitution, insertion, deletion models

Most models of sequence evolution further assume that the evolution consists of two independent processes, namely a k-mer residue substitution process and an insertion/deletion (indel) process. The k-mer residue substitution process can itself be considered as a continuous time Markov process. Let $\dot{\Upsilon}$ and $\dot{\mathcal{E}} = (\dot{\mathbf{Q}}, \dot{\pi})$ denote the state space and substitution model respectively for a single residue substitution process, with the natural extension for 2-mer and 3-mer substitution processes. The symbols u, v will be used to represent both arbitrary length sequences as well as single residues, but it will be clear from the context which is implied in each case.

Miklós et al. consider the class of evolutionary models which allow local point substitutions and multiple residue inserts and deletes (called SID models) [MLH04]. Let $\rho_I(u)$ be the context-independent rate of insertion of sequence u between two residues in an ancestral sequence and let $\rho_D(u)$ be the context-independent rate of deletion of sequence u.

The simplest SID model disallows insertions and deletions in the evolutionary model, and treats gaps as either missing data (see section 1.3.3 for a discussion on how to accommodate missing data in the likelihood calculation), or as an extra residue character. This has the effect of not allowing the sequence length to change over time. The residue substitution process can be further simplified by assuming that sites evolve independently of one another according to a single residue model $\dot{\mathcal{E}}$. Site-specific residue models are discussed in more depth below.

The TFK91 links model [TKF91] is a SID model with an arbitrary point substitution matrix $(\dot{\mathbf{Q}}, \dot{\pi})$ and an indel process governed by

$$\rho_I(u) = \begin{cases} \lambda \dot{\pi}_{u_1} \text{ if } |u| = 1 \\ 0 \text{ otherwise} \end{cases} \tag{1.21}$$

$$\rho_D(u) = \begin{cases} \mu \text{ if } |u| = 1 \\ 0 \text{ otherwise,} \end{cases} \tag{1.22}$$

where λ is the insert rate and μ is the deletion rate. The TFK92 model [TKF92] is an extension to this model but considers a sequence to consist of fixed-length indivisible fragments of variable length k. The substitution process for each of these k-mers is given by $\overset{k}{\mathcal{E}} = (\overset{k}{\mathbf{Q}}, \overset{k}{\pi})$

and the indel process is the same as for TFK91 but considered over arbitrary length k-mers. Each of the $\overset{k}{\mathcal{E}}$ can be parameterised as a product of k independent single residue models.

Miklós et al. introduce a long-indel model [MLH04] which is parameterised over the state space Υ of sequences of arbitrary length by

$$\rho_I(u) = \lambda_{|u|} \prod_{k=1}^{|u|} \dot{\pi}_{u_k} \tag{1.23}$$

$$\rho_D(u) = \mu_{|u|} \tag{1.24}$$

where $\lambda_{|u|}$ and $\mu_{|u|}$ are the rate of deletion and insertion respectively of sequences of length $|u|$. Miklós et al. derive restrictions on the insertion and deletion rates in order to preserve reversibility. Alignment algorithms using this model are also presented.

Mitchison and Durbin [MD95] propose a tree HMM to model insertions and deletions. Under this model, there is a collection of n match $\{M_j\}$ and n delete $\{D_j\}$ states of a HMM, each of which will generate a column in a multiple sequence alignment with n columns. The model does not allow insertions, so the maximum length n of sequence generated by this model is pre-specified. The HMM architecture is shown in figure 1.3. The path through the HMM is evolved as well as the residues. Thus, the sequence x^t is augmented with ψ^t describing the path through the model at time t. Mitchison and Durbin propose that each transition in the model evolves independently according to continuous, stationary, time-reversible Markov process, denoted $\bar{\mathcal{E}} = (\bar{\mathbf{Q}}, \bar{\pi})$, over the state space of transitions, denoted by $\bar{\Upsilon}$, where

$$\bar{\Upsilon} = \{M_i \to M_{i+1}, M_i \to D_{i+1}, D_i \to D_{i+1}, D_i \to M_{i+1}\}, \tag{1.25}$$

provided M_0 is interpreted as the begin state and M_{n+1} is interpreted as the end state. The rate matrix over transitions $\bar{\mathbf{Q}}$ is trained from a database of alignments. Note that the path ψ^t is hidden, and so to use the tree HMM for inference of trees and evolutionary distances, it is necessary to sum over all possible hidden states, which is computationally expensive. The tree HMM does not attempt to model novel insertions – instead it models 're-insertion' of ancestral sequence which has been temporarily lost in a lineage. For practical purposes this is not a substantial drawback but it is unsatisfactory from a theoretical point of view.

1.3.2 Models of residue substitution

The most general non-reversible DNA model is the unrestricted model (UNR), which has 11 free parameters (12 off-diagonal elements minus 1 parameter for scaling). The general time

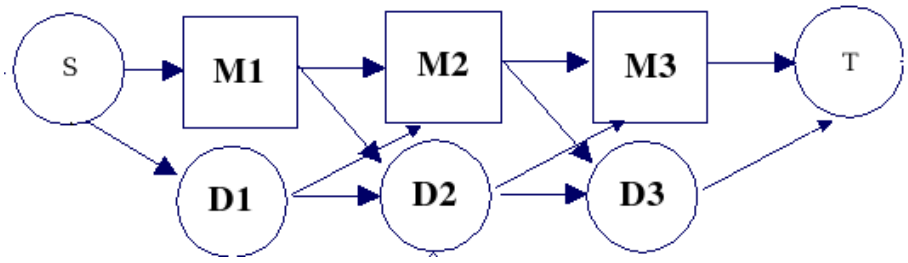

Figure 1.3: Basic architecture for tree HMM as defined in [MD95]. Each M_i corresponds to a column in the multiple alignment, and emits over a distribution of amino acids. S, T correspond to start and termination states.

reversible model (GTR) has 5 free parameters (6 upper diagonal elements minus 1 parameter for scaling). Other substitution models include the HKY model [HKY85], which has 4 free parameters and the F81 model [Fel81], which has 3 free parameters. These parameters are typically trained from an alignment with a given tree topology by jointly finding the parameters of the model and the branch lengths which maximise the likelihood of the data. Methods have been devised for simultaneously optimising over alternative tree topologies.

Goldman and Yang [GY94] and Muse and Gaut [MG94] have described models of codon evolution which take into account an underlying nucleotide model based on the HKY [HKY85] with transition/transversion ratio of κ as well as a non-synonymous/synonymous rate ω. These models were refined in [YN98]. Let $u = u_1u_2u_3$ and $v = v_1v_2v_3$ be one of the 61 non-termination codons. Then the codon rate matrix is parameterised as

$$\dddot{\mathbf{Q}}_{u,v} = \begin{cases} 0 \text{ if the codons differ at more than 1 position,} \\ \dddot{\pi}_v \text{ for a synonymous transversion,} \\ \kappa\dddot{\pi}_v \text{ for a synonymous transition,} \\ \omega\dddot{\pi}_v \text{ for a non-synonymous transversion,} \\ \omega\kappa\dddot{\pi}_v \text{ for a non-synonymous transition.} \end{cases} \tag{1.26}$$

In [GY94] codon models which take into account the chemical similarity of substituted amino acids are presented.

The codon and DNA models presented have a small number of parameters which can be trained by maximum likelihood (ML) on a given alignment. Reversible amino acid models, on the other hand, have 190 upper diagonal elements and so 189 free parameters after scaling as

in eq. 1.11. Reliable estimates of these parameters requires a significant amount of data. Thus amino-acid rate matrices are typically derived from databases of alignments, as is discussed in more detail below. However, different alignments and particularly different sites in an alignment have very different structural and functional constraints. Using a database derived rate matrix to describe all columns in an alignment leads to an inaccurate model of evolution at particular sites [Bru96, HB98]. One approach is to modify database derived rate matrix, labelled $\hat{\dot{\mathcal{E}}} = (\hat{\dot{\mathbf{Q}}}, \hat{\dot{\pi}})$ so that the stationary probabilities of the new rate matrix $\dot{\mathbf{Q}}$ are set to a given distribution $\dot{\pi}$ which is set (or trained) to reflect a particular alignment or site in an alignment. This can be achieved as described in [CAJ$^+$94, GW02] where the exchangeabilities matrix for the alignment is set to be equal to that estimated from a database, i.e. $\dot{\mathbf{S}}(f) = \hat{\dot{\mathbf{S}}}(f)$, where the equation for $\mathbf{S}(f)$ is given in eq. 1.15. This leads to the equation

$$\dot{\mathbf{Q}}_{uv} = \left(\frac{\dot{\pi}_v}{\hat{\dot{\pi}}_v}\right)^{1-f} \times \hat{\dot{\mathbf{Q}}}_{uv} \times \left(\frac{\hat{\dot{\pi}}_u}{\dot{\pi}_u}\right)^{f}. \tag{1.27}$$

The +gwF parameter f is restricted to lie between 0 and 1 [GW02]), and can be thought of as the trade-off between frequencies in the equilibrium distribution resulting from pressure to mutate from ($f = 1$) and pressure to mutate towards ($f = 0$) a particular residue/base. The most common approach [CAJ$^+$94] is to set $f = 0$ which reduces equation 1.27 to

$$\dot{\mathbf{Q}}_{uv} = \left(\frac{\dot{\pi}_v}{\hat{\dot{\pi}}_v}\right) \hat{\dot{\mathbf{Q}}}_{uv} \tag{1.28}$$

Accounting for variation between sites

In many cases a single substitution model is used for every site in an alignment, in contrast to the profile HMM methods discussed in section 1.2 which have a different frequency distribution at each site. This is readily seen to be a drastic simplifying assumption for both protein and DNA alignments. Some regions in an alignment will be evolving slowly due to functional and/or structural reasons. Regions of DNA vary greatly in composition, for example in G+C content. Codon substitution patterns at neutrally evolving or positively selected sites will be different from those under purifying selection.

Yang [Yan93] proposed the use of a mixture of substitution models of the form eq. 1.12 each with different rates r chosen from a discretised gamma distribution. Yang [Yan95] as well as Felsenstein and Churchill [FC96], further proposed correlating rates at adjacent sites via a first order spatial Markov chain. In [Yan95] it is assumed that the rate at a site is drawn

from a different discretised gamma distribution, and the variance of the gamma distribution is the hidden parameter of the spatial HMM. In [FC96] the rate (drawn from a finite set of categories) is itself the hidden parameter of the HMM. In [SH04], context dependent site-specific substitution models are used as part of an HMM framework. Their model is context dependent on the previous column:

$$P_{\mathcal{E}}(x^{t+\triangle t} = u | x^t = v) = \\ P_{\mathcal{E}}(x_1^{t+\triangle t} = u_1 | x_1^t = v) \prod_{i=2}^{n} P_{\mathcal{E}}(x_i^{t+\triangle t} = u_i | x_i^t = v_i, x_{i-1}^t = v_{i-1}, x_{i-1}^{t+\triangle t} = u_{i-1}) \quad (1.29)$$

Yang and Nielsen [NY98, YN00] describe models of codon evolution where the ratio of the rates of non-synonymous and synonymous substitution are allowed to vary between sites. These models have been successful in detection positive selection, as discussed in section 4.2.

Bruno [Bru96, HB98] learn site-specific rate matrices from a given alignment, using an amino-acid model and a codon model respectively. In both cases the site-specific rate matrices are defined in terms of the site-specific residue frequencies $\dot{\pi}$. In [Bru96] the EM algorithm is used to find the residue frequencies which optimise the likelihood of the column. In [TGJ96] the authors introduce a model for amino-acid evolution which has rate matrices specific to particular secondary structure states. A spatial HMM is used to correlate the hidden structural states along the length of the sequence. The authors demonstrated a significantly better likelihood fit to the alignment data, and used the model to derive phylogenetic trees as well as to label sites in the alignment with secondary structure states. This method was extended to accommodate more states in [LGTJ98] and to model transmembrane proteins specifically in [LG99].

Lartillot and Hervet [LP04] have recently defined a Bayesian mixture model which allows each site in an alignment to evolve according to a mixture of K distinct evolutionary models $\dot{\mathbf{Q}}^k$, where K is itself a parameter of the model and the $\dot{\mathbf{Q}}^k$ are parameterised as in eq. 1.28. Thus each class is parameterised by a different stationary probability $\dot{\pi}^k$. The authors define appropriate priors over the model parameters as well as tree space and present a MCMC sampling technique for identifying the ML model parameterisation and tree. In this way they are able to learn the optimal number of rate matrix categories in the data.

Database derived protein rate matrices

The original attempts to estimate $\hat{\mathbf{Q}}$ used maximum parsimony (MP) rather than maximum likelihood. Dayhoff et al. [DSO78] and [JTT92] used MP to estimate both the trees and ancestral sequences for multiple protein families, and counted the observed amino acid replacements along the tree to estimate the PAM matrices. Jones et al. [JTT92] extended this technique and applied it to a much larger database of protein families. To avoid observing transitions which are the product of multiple steps and to avoid assigning an ancestral sequence, the authors counted transitions based on pairwise sequence comparisons (where each sequence is used in only one comparison) between sequences which are more than 85% identical.

The maximum likelihood approach has been applied to estimating amino acid replacement rates in [AH96, YN98, AWMH00, WG01]. The first three of these estimated amino acid replacement rates in vertebrate mitochondrial, mammalian mitochondrial and chloroplast sequences respectively. Whelan and Goldman [WG01] apply an approximate form of ML training on a larger database of globular protein sequences. Holmes and Rubin [HR02] use expectation maximisation (EM) [DLR77] to train substitution models from sequence alignments and phylogenetic trees. The EM algorithm is designed to maximise the likelihood of data where some of the data is missing. In this case the missing data corresponds to the precise substitution history of the sequence. The model can also accommodate finding a pre-defined number of hidden substitution rate matrices in the data.

1.3.3 Likelihood calculation

For a given tree T with branch lengths specified and evolutionary model $\mathcal{E}$, it is desirable to calculate the likelihood of a cluster of sequences $\{x^k\}$, $P(\{x^k\}|T,\mathcal{E})$. This likelihood is useful for several purposes: to evaluate different evolutionary models on a fixed tree, with the aim of finding the model that best fits the data; or to evaluate different trees with a fixed evolutionary model, with the aim of finding the tree which best fits the data. The likelihood can be calculated efficiently using Felsenstein's algorithm [Fel81]. Felsenstein's algorithm allows the summation over unknown states at internal nodes of the tree, and is closely related to the forward algorithm for for HMMs. In fact, as several authors have noted, the algorithm also allows summation over unknown states at the leaves of the tree (which might occur, for

example if there is a gap present in the alignment). Let $p_{u,k}$ denote the partial likelihood of all sequences $\{x^{k'}\}$ below node n_k given sequence u in the ancestral sequence at n_k. This algorithm proceeds by calculating in post-order (i.e. working upwards from the leaves),

$$p_{u,k} = \begin{cases} 1 \text{ if } n_k \text{ is a leaf node and } u \text{ matches } x^k \\ 0 \text{ if } n_k \text{ is a leaf node and } u \text{ does not match } x^k \\ \prod_h \sum_v p_{v,kh} \cdot P_{\mathcal{E}}(x^{kh} = v | x_k = u) \text{ otherwise} \end{cases} \tag{1.30}$$

where $n_{k1}, n_{k2}, \ldots$ are the child nodes of n_k. The term 'matches' (following [SH04]) has been used to include cases where x^k contains a gap but is otherwise equal to u. This is effectively the same as treating the gap as missing data.

1.4 Outline of thesis

In this thesis, I focus on probabilistic modelling of protein domain evolution. Protein domain databases, such as Pfam [BCD+04] provide a valuable resource for studying protein domain evolution. To demonstrate the volume of data amenable for probabilistic analysis of the type described above, Pfam release 16.0 contains 7677 protein families covering 1.1m protein sequences and 264m residues. In the next two chapters of this thesis I investigate ways to model protein domains in order to improve protein domain detection and to extend the coverage of protein domain databases. Looking for distant homologues is important beyond simply extending residue coverage of domain databases. Arguably the most divergent members of a particular domain family are the most interesting for identifying the range of potential functions and partners for a particular domain as well as identifying fast evolving proteins. The final chapter of the thesis concerns looking for such fast evolving proteins in order to identify pseudogenes, as well as proteins and sites under positive selection.

Chapter 2

Enhanced Domain Detection Using Approaches From Speech Recognition

Most modern speech recognition techniques use probabilistic models to interpret a sequence of sounds [Cha93, Jel97]. Hidden Markov models, in particular, are used to recognize words. The same techniques have been adapted to find domains in protein sequences of amino acids [KBM$^+$94, DEKM98], as discussed in section 1.2. However in both cases, detection of individual constituent domains or words is impeded by noise. One technique which has been successfully used in speech recognition is to use language models to capture the information that certain word combinations are more likely than others, thus improving detection based on context. As discussed in section 1.1, only a limited set of all possible domain combinations are observed, and the pattern of occurrence is highly non-random ([AGT01b, AHT03]). Moreover, particular domain combinations are re-used in many domain architectures [VBB$^+$04]. Thus, language models from speech recognition may also be applicable to the problem of protein domain identification. I have successfully used this approach to improve domain prediction in Pfam [CBD03].

Furthermore, different species have different protein domain repertoires, even to the extent that certain protein domain families are kingdom specific. More strikingly, domain combinations are highly kingdom specific ([AGT01b, VBB$^+$04]). Thus, taxonomic context by

itself may also provide extra information for domain detection, and is likely to be even more useful when used in combination with language models of domain context. I have previously used taxonomic information to improve domain identification in Pfam [CBD04].

In this chapter, I will present a unified model of domain and taxonomic context, extending the approaches of [CBD03, CBD04]. I will first provide a brief overview of some of the techniques used in speech recognition, followed by a comparison of the high-level statistics of word and domain use which will help to motivate further the application of language modelling to domain detection. I then modify the speech recognition techniques in order to apply them to domain detection and to incorporate taxonomic context. The results section comprises firstly a test of the method on proteins of known structure using the SCOP classification[AHB+04], in which I will show that the combined taxonomic and domain context method performs better than the individual methods and that each perform better than a standard search which ignores context altogether. The final part of the results section consists of a scan of the combined method against all Uniprot[ABW+04] proteins to determine the number of novel Pfam domain occurrences detectable with this technique.

2.1 Statistical Speech Recognition Techniques

Speech recognition has been greatly facilitated by the application of statistical models including hidden Markov models (HMMs) and Bayesian methods. The steps in the process are illustrated in figure 2.1.

Once the acoustic signal has been parsed into discrete sound symbols, the statistical approach is to build two types of model: for each word there is a phonetic model for the emission of sounds, based on observed pronunciation patterns in terms of phonemes; above this there is a language model for the emission of a sequence of words, based on word use patterns. In order to recognize a given sentence, the method seeks the sequence of words $\mathbf{D} = \mathbf{D}_1, \ldots \mathbf{D}_n$ that maximises the probability of the sentence given the acoustic evidence x and the language model $\mathbf{M}$. This probability can be split (using Bayes' rule) into a word term based on the phonetic model (first term), and a 'context' term, based on the language model (second term):

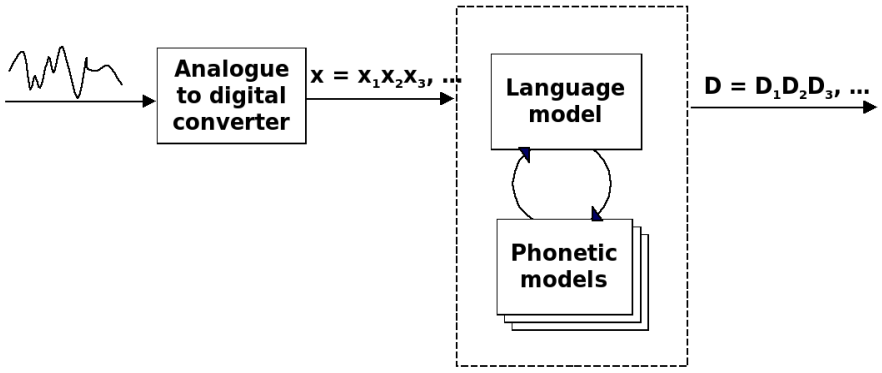

Figure 2.1: Schema for a speech recognizer. First the analogue speech waveform is converted into a sequence of phonemes, $x = x_1, x_2, \ldots$. This sequence is processed by a composite stochastic language and phoneme model.

$$P(\mathbf{D}|x, \mathbf{M}) = \frac{P(x|\mathbf{D})}{P(x|\mathbf{M})} P(\mathbf{D}|\mathbf{M}),, \tag{2.1}$$

assuming that x is conditionally independent of the language model $\mathbf{M}$ given $\mathbf{D}$. When searching for the most likely sequence of words $\mathbf{D}$, $P(x|\mathbf{M})$ is a fixed constant so it suffices to maximise

$$P(\mathbf{D}|x, \mathbf{M}) \propto P(x|\mathbf{D})P(\mathbf{D}|\mathbf{M}). \tag{2.2}$$

Referring again to figure 2.1 and equation 2.1 observe that statistical speech recognition naturally divides itself into the following sub-problems, each of which I discuss to the extent it applies to domain recognition. See [RJ93] for further details.

Conversion Convert the analogue signal into a discrete acoustic signal

Acoustic Modelling For each word, develop and parameterise an acoustic model capable of discriminating the given word from all others.

Language Modelling Develop and parameterise a single language model

Conversion to a digital signal

The basic idea is to sample the properties of the acoustic signal at some rate (e.g 100HZ), and to find the closest match to this vector of properties from a library of reference vectors. As biological sequences are already digitized, this problem is not applicable.

Acoustic Modelling

The aim is to construct an acoustic model for each word in the language model which is capable of recognizing words from acoustic signal. A phonetic encoding is determined for each word in the vocabulary as a sequence of phonemes $\phi_1, \phi_2, ...$ from a phonetic dictionary. For each phoneme in the phonetic dictionary a HMM is created which emits over the space of sound symbols obtained from the previous step. The encoded phonemes' HMMs are concatenated to form a word HMM. Training data for the word models is obtained by recording word pronunciations. The model can be trained from this data using the Baum-Welch algorithm[DEKM98]. This step corresponds to using a profile HMM in biological sequence modelling, as discussed in section 1.2.

Language Modelling

The aim of language modelling is to create a model over all possible word combinations which reflect actual word use patterns in speech. The analogy in domain recognition is a model over all possible domain combinations which reflect protein domain occurrence patterns. Mathematically this corresponds to parameterising the distribution $P(\mathbf{D}|\mathbf{M}) = P(\mathbf{D}_1, \ldots \mathbf{D}_n|\mathbf{M})$ in a tractable form. One approach is to assume that word use is a Markov process. That is, if the joint probability is expressed in terms of conditional probabilities,

$$P(\mathbf{D}_1 \ldots \mathbf{D}_n|\mathbf{M}) = P(\mathbf{D}_1)P(\mathbf{D}_2|\mathbf{D}_1) \ldots P(\mathbf{D}_n|\mathbf{D}_1, \ldots, \mathbf{D}_{n-1}),$$

to assume that

$$P(\mathbf{D}_i|\mathbf{D}_{i-1}, \ldots, \mathbf{D}_1) = P(\mathbf{D}_i|\mathbf{D}_{i-1}, \ldots, \mathbf{D}_{i-k}).$$

In speech recognition, a second order ($k = 2$) Markov model is usually found to be most effective, which is called a *trigram* model. First order methods are called *digram* methods. In principle, the higher the order k, the more long-range dependencies can be incorporated into the model. However, for a fixed data set, as k increases less and less training data becomes available for the particular context and so the probability estimates become less and less reliable. In linguistic terms, Markov models are stochastic regular grammars, and therefore do not capture the grammatical structure of a sentence. Thus they are not capable of assigning

zero probability to grammatically incorrect sentences, nor modelling long range dependencies implied by the grammatical structure. To achieve this it is necessary to use (in order of increasing complexity and ability to effectively model linguistic structures) stochastic context free grammars, tree-adjoining grammars [JY99] or context sensitive grammars [Cho59]. However, Markov models are computationally efficient and have been found to work surprisingly well in practice.

For domain recognition, it is not yet clear that there is a general higher-order grammar for domain occurrence, much less how to represent the syntax with a formal grammar. Thus, approximating the dependence of domain occurrence based on adjacent domains appears to be an appropriate way to proceed. A phenomenon which occurs in protein domain combinations but not speech is nested domains, which account for 9% of all domain combinations [ASHS04].

Training data for a language model is obtained from analysing text, typically in the subject area in which the model will be used. In one sense training a language model is straightforward, as there are no hidden variables and the transition probabilities between words can be observed directly. However, the main challenge with language modelling is data sparseness, particularly with trigram methods. The training corpus will not contain all possible trigram word combinations used in speech, and observed trigrams occur at such low frequencies that observed counts are not reliable estimators of probability. This is dealt with via smoothing, which is an integral part of language modelling and has formed the basis for much language modelling research.

Equivalence classification of words is one technique for smoothing sparse data. An example is to treat all the synonyms for a particular word as the same; another is to classify all proper names as a single word. An example from domain modelling is classifying all members of a superfamily as the same domain, or classifying regions of low complexity as a single domain. The method developed in this chapter classifies all Pfam families in the same Pfam clan as the same family.

Another smoothing technique is to interpolate lower order counts in the estimation of the trigram and digram probabilities. That is, to assign

$$P(\mathbf{D}_i|\mathbf{D}_{i-1}) = \alpha_1 P(\mathbf{D}_i) + (1-\alpha_1)\frac{\mathbf{N}(\mathbf{D}_{i-1},\mathbf{D}_i)}{\mathbf{N}(\mathbf{D}_{i-1})}, \tag{2.3}$$

$$P(\mathbf{D}_i|\mathbf{D}_{i-1},\mathbf{D}_{i-2}) = \alpha_2 P(\mathbf{D}_i|\mathbf{D}_{i-1}) + (1-\alpha_2)\frac{\mathbf{N}(\mathbf{D}_{i-2},\mathbf{D}_{i-1},\mathbf{D}_i)}{\mathbf{N}(\mathbf{D}_{i-2},\mathbf{D}_{i-1})}, \tag{2.4}$$

where $\mathbf{N}(\mathbf{D}_i)$ is the count of $\mathbf{D}_i$ in the corpus, so that even in the case the trigram is not observed it will have a non-zero probability assigned based on the digram probability. One principle often used, called back-off estimation, is that the trigram probabilities $P(\mathbf{D}_i|\mathbf{D}_{i-1}, \mathbf{D}_{i-2})$ should be more reliable if the context $\mathbf{D}_{i-2}, \mathbf{D}_{i-1}$ is observed many times in the training corpus, and similarly for the digram probabilities. Thus it makes sense that the interpolation parameters α are not constant but rather decreasing functions of the amount of context, e.g. $\alpha_2 = f(\ \mathbf{N}(\mathbf{D}_{i-2}, \mathbf{D}_{i-1}))$. A step-function is typically used to approximate f, with several categories each having a different value of α. A portion of the training data is held over to estimate the optimal values for this function.

An alternative to Markov models for approximating the joint distribution $P(\mathbf{D}_1, \ldots, \mathbf{D}_n)$, which only capture local dependencies, is the whole-sentence exponential model introduced by Rosenfeld and co-workers [RCZ01]. The whole sentence exponential model takes the form

$$P(\mathbf{D}) = \frac{1}{Z} P_0(\mathbf{D}) \exp(\sum_i \lambda_i f_i(\mathbf{D})) \tag{2.5}$$

where Z is the normalizing constant and the $f_i(\mathbf{D})$ are termed *features* of the sentence: arbitrary properties of the sentence which can be computed. $P_0(\mathbf{D})$ is an initial approximation to $P(\mathbf{D})$, which can be a uniform distribution, or the distribution obtained from the trigram model described above. It can be shown that there exists is a unique equation of the form of eq. 2.5 which satisfies the following constraints on the feature averages under $P(\mathbf{D})$,

$$E_P(f_i) = K_i, \tag{2.6}$$

provided the constraints are consistent. Moreover, among all solutions to equation 2.6 (including solutions not of exponential form), the exponential solution is closest to $P_0(\mathbf{D})$ under the Kullback-Leibler distance (see [DEKM98]). This means that in the case $P_0(\mathbf{D})$ is the uniform distribution, the exponential solution is the solution which maximises the entropy. In this sense the exponential solution is appealing because it maximises the uncertainty of the distribution while still satisfying all of the constraints presented. So, given a training corpus, the strategy of whole-sentence exponential modelling is to first choose features f_i which capture particular aspects of the data, then to calculate empirical averages of the f_i over all sentences $\mathbf{D}'$ in the training corpus

$$K_i = \frac{1}{N} \sum_{\mathbf{D}'} f_i(\mathbf{D}'),$$

and finally to find the unique equation of the form eq. 2.5 which satisfies the constraints in eq. 2.6. An iterative procedure is available to find this solution, and is given in [RCZ01]. The main challenge in implementing this procedure is that it requires calculating the average $E_p(f_i)$ over all possible sentences $\mathbf{D}$ at each step in the iteration. This is approximated by Rosenfeld and colleagues using a sampling technique.

In speech recognition the features used in an exponential model include: the number of times a particular n-gram occurs, either sequentially, or in the entire sentence; existence of particular grammatical structures; pauses etc. For domain recognition this framework could incorporate arbitrary co-occurrence patterns (not just adjacent co-occurrence), expected distribution over the number of repeats as well as protein specific information such as taxonomy, function and localisation.

In this work I focus on applying the Markov rather than exponential model approach to language modelling. The Markov model is substantially more efficient to train and to score, and has been used successfully in speech recognition. Moreover, early results from whole-sentence models do not appear to provide a significant improvement in performance [RCZ01]. However, the exponential model does appear to provide significantly more flexibility and is certainly an avenue for further investigation.

2.2 Patterns of domain occurrence and co-occurrence

To motivate the application of language models to protein domain recognition it is interesting to observe the patterns of domain occurrence in relation to the pattern of word occurrence.

Zipf [Zip35] first described the power law behaviour of word occurrence. The Zipf distribution for words is displayed in figure 2.3 and reflects the fact that some words are used very frequently while most words are used rarely. The power law distribution is of the form $\mathbf{N}(D) = aR(D)^{-b}$ where $\mathbf{N}$ is the count of a word and $R(D)$ is the rank of the word according to its count. A Zipf distribution also satisfies $b = 1$. Power law behaviour has been observed in many biological contexts, including the distribution of protein families and folds [QLG01], occurrence of DNA k-mers, occurrence of pseudogenes and levels of gene expression [LQZ+02].

The Zipf curve for words in figure 2.3 applies from rank 3 to rank 2000 but breaks down after this. Figure 2.2 shows a power law distribution for Pfam domains. The slope of this

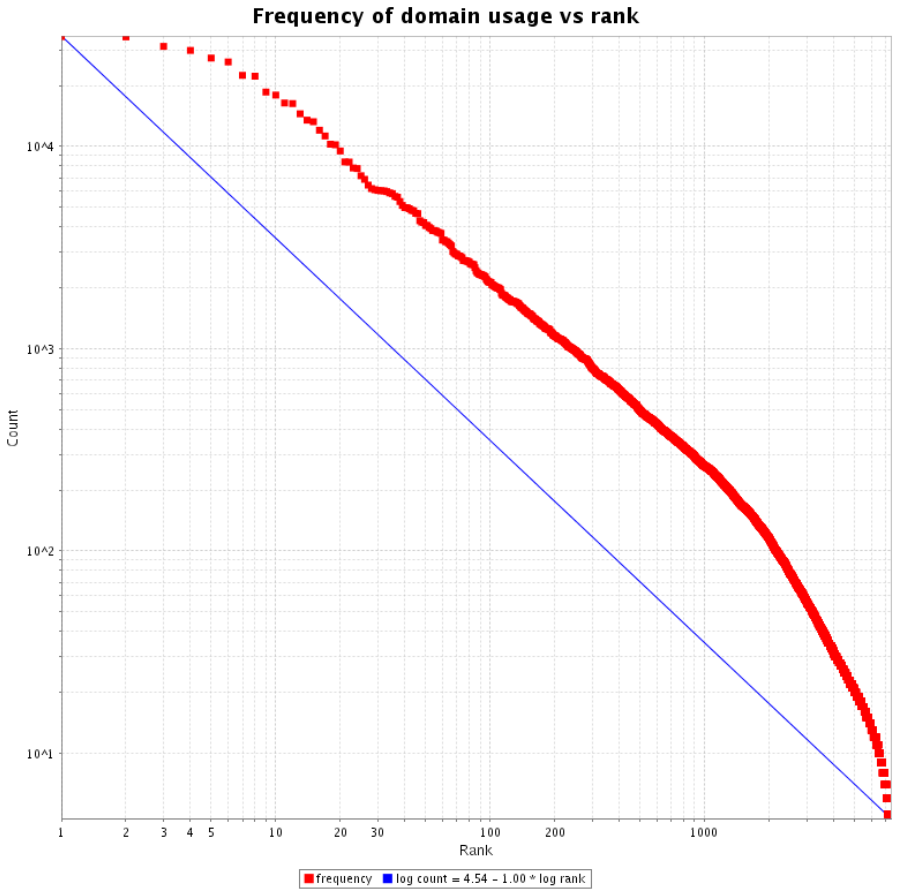

Figure 2.2: Plot of frequency of domain occurrence vs domain rank (according to frequency, decreasing from left to right). The domains models are from Pfam 15.0, and are scored over all proteins from Uniprot. Pfam clans have been used to group closely related domains into a single entry. The blue line shows the $\log y = C - 1.0 \log x$ line interpolated between the highest and lowest ranked domain.

graph is approximately 1.0 from rank 5 to rank 2000, but the gradient is higher at high-rank domains and lower at low-rank domains. As domain annotation improves, we expect to find novel small families, but for some of these small families to have more than 200 instances. These families will then be of higher rank than than those known families of rank 2000 and above. We also expect to increase the number of instances in small families as they are not as well characterised as larger families. The combined effect should be to expand the region of the graph following Zipf's law to toward the right. It appears that Zipf's law fits protein domains at least as well as words.

Next, I consider the different patterns of domain occurrence given different taxonomic

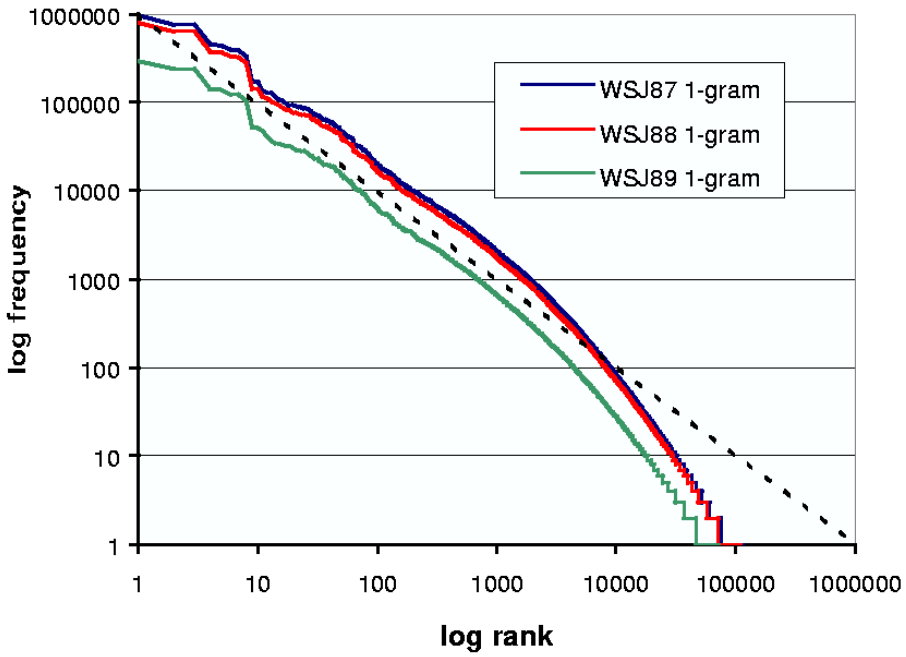

Figure 2.3: Plot of frequency of word occurrence, taken from the Wall Street Journal from 1987, 1988 and 1989 with sizes approximately 19 million, 16 million and 6 million words respectively. This graph is taken from [HSGMS02]

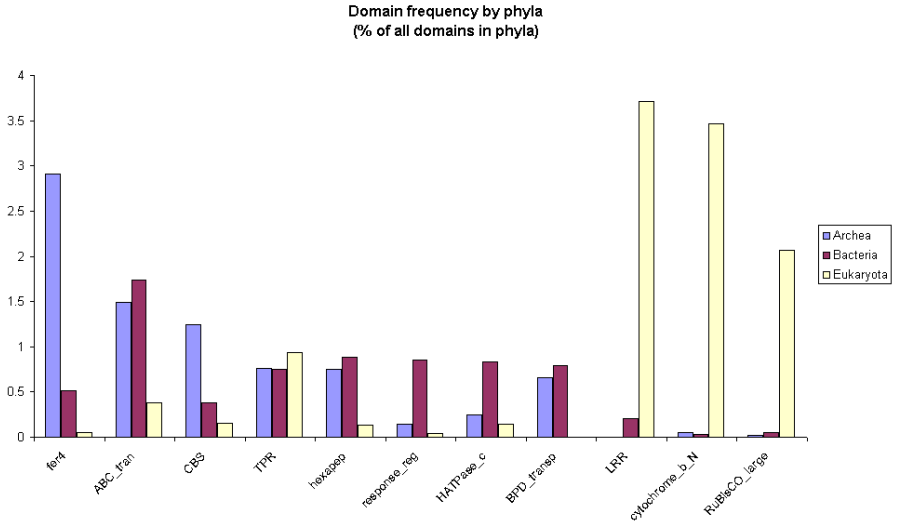

Figure 2.4: Distribution of example domains amongst archaea, eukaryota and bacteria from all proteins in Uniprot. The top 5 domains for each phyla are included. This graph was not constructed on a genome basis and redundancies in the Uniprot database have not been removed, thus the graph may display bias due to over-representation of particular sequences in Uniprot.

contexts. Fig. 2.4 shows examples of domains which have biased taxonomic distribution. For example, the 4Fe-4S binding domain comprises 2.9% of archaeal domains in Pfam, but only 0.5% of bacterial domains and 0.05% of eukaryotic domains. Therefore a weak 4Fe-4S binding domain signal in archaea is more likely to be a real signal than a weak eukaryota 4Fe-4S binding domain signal. Intuitively, less amino-acid based evidence is required to believe an 4Fe-4S binding domain in archaea than in eukaryota.

Figure 2.5 demonstrates different patterns of co-occurrence of the TPR domain across three kingdoms of life. The TPR domain mediates protein-protein interactions and is observed in eukaryota, bacteria and archaea, as can be seen in figure 2.4. In each of the three kingdoms there is a high probability of observing TPR following another TPR repeat. Uniquely in eukaryota, a TPR domain is frequently observed following an APC8 (Anaphase promoting complex sub-unit 8) domain and also following a PRP1_N (PRP splicing factor, N-terminal) domain. Uniquely in bacteria, a TPR domain has high probability following a NB-ARC (signalling motif found in bacteria and eukaryota) and following an FF domain (also involved in protein-protein interaction and found in eukaryotes and bacteria). Uniquely to Archaea, there is a high probability of observing a TPR domain following a CW_binding_2

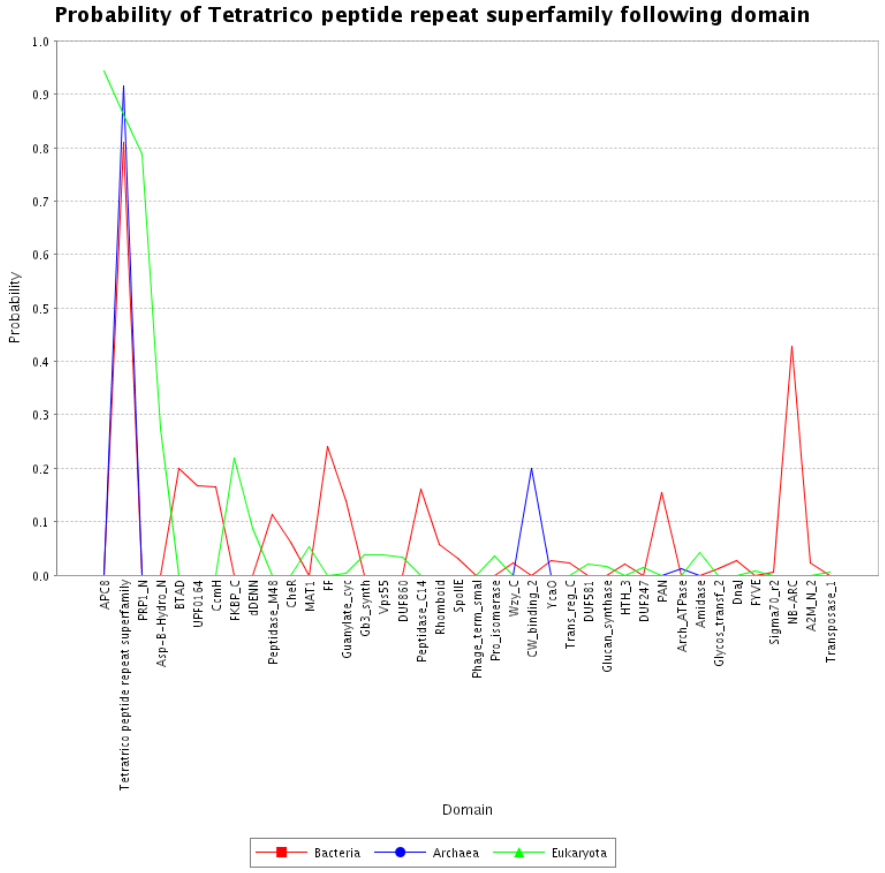

Figure 2.5: Observed probability of Tetratrico Peptide Repeats in different contexts. The probability of observing a member of the TPR clan given the combined taxonomic and domain context - Bacteria (red), Archae (blue) and Eukaryota (green) - and preceding domain.

(putatively involved in cell wall binding) domain. This reinforces the findings of [AGT01b] that domain combinations are highly kingdom specific, and also indicates the importance of building language models which take the taxonomic context into account.

2.3 Methods: Application to protein domain detection

As discussed in section 2.1, profile HMM techniques introduced in section 1.2 broadly map to the acoustic modelling problem in speech recognition [KBM+94, DEKM98]. In this section, I will modify the language modelling techniques outlined in section 2.1 to apply them to protein domain recognition.

2.3.1 Formulation

Let $\mathbf{M}$ denote the combined language and taxonomy model. For each amino acid sequence x with taxonomy T my approach is to annotate the sequence with the domain sentence $\mathbf{D} = \mathbf{D}_1, \mathbf{D}_2, \ldots \mathbf{D}_n$ matching amino acid segments $\mathbf{D}_i \leftrightarrow x_{[s_i,e_i]}$ if the probability $P(\mathbf{D}|x, \mathbf{M})$ is sufficiently high. Let $\mathbf{R}$ denote the background model for generating the sequence independently residue by residue according to an average compositional model. Note that in this formulation $\mathbf{D}$ stands for the N- to C-terminal linear sequence of domains as well as the particular set of (start, end) protein co-ordinates for each of the $\mathbf{D}_i$. I require that the $\mathbf{D}_i$ do not overlap, but place no restriction on the size of the gaps between the $\mathbf{D}_i$. It is not required that the $x_{[s_i,e_i]}$ completely cover the protein. It should be noted that in the case where a protein domain has not yet been modelled (for instance, it does not appear in the pdb, and it has not yet been discovered in sequence space), a relatively large gap may result in the correct domain annotation of a sequence. Also, transmembrane and low complexity regions are not modelled. Residues which are not within a $x_{[s_i,e_i]}$ will be assumed to be emitted under the model $\mathbf{D}$ according to the background distribution $\mathbf{R}$. Then

$$P(\mathbf{D}|x, T, \mathbf{M}) = \frac{P(x|\mathbf{D},T,\mathbf{M})}{P(x|T,\mathbf{M})} P(\mathbf{D}|T, \mathbf{M}) \tag{2.7}$$

$$\propto \frac{P(x|\mathbf{D},T,\mathbf{M})}{P(x|\mathbf{R})} P(\mathbf{D}|T, \mathbf{M}) \tag{2.8}$$

$$= \left(\prod_i \frac{P(x_{[s_i,e_i]}|\mathbf{D}_i,T,\mathbf{M})}{P(x_{[s_i,e_i]}|\mathbf{R})} P(\mathbf{D}_i)\right) \times \left(\prod_i \frac{P(\mathbf{D}_i|T,\mathbf{M},\mathbf{D}_1,\ldots\mathbf{D}_{i-1})}{P(\mathbf{D}_i)}\right), \tag{2.9}$$

assuming independence of the amino acid fragments $x_{[s_i,e_i]}$ from the other fragments $x_{[s_j,e_j]}$, $j \neq i$ conditional on $\mathbf{D}_i, T, \mathbf{M}$. Because I am only interested in maximising $P(\mathbf{D}|x, T, \mathbf{M})$ over all possible domain sentences and fixed x, the term $P(x|T, \mathbf{M})$, which is independent of the domain sentence, has been replaced with $P(x|\mathbf{R})$. Then residues not belonging to any sequence fragment $x_{[s_i,e_i]}$ cancel out between the numerator and denominator.

Taking logs and defining the overall sentence score $\mathbf{SS}_{x,T,M}$

$$\log P(\mathbf{D}|x, T, \mathbf{M}) \propto$$

$$\mathbf{SS}_{x,T,M}(\mathbf{D}) := \left(\sum_i \log \frac{P(x_{[s_i,e_i]}|\mathbf{D}_i)}{P(x_{[s_i,e_i]}|\mathbf{R})} - \tau_{\mathbf{D}_i}\right) + \left(\sum_i \frac{P(\mathbf{D}_i|T, \mathbf{M}, \mathbf{D}_1, \ldots \mathbf{D}_{i-1})}{P(\mathbf{D}_i)}\right), \tag{2.10}$$

with domain score threshold $\tau_{\mathbf{D}} = \log \frac{1}{P(\mathbf{D})}$. Note that $P(x_{[s_i,e_i]}|\mathbf{D}_i)$ represents the probability that the model for domain $\mathbf{D}_i$ generated the sequence $x_{[s_i,e_i]}$; and that $P(x_{[s_i,e_i]}|\mathbf{R})$ represents the probability that the sequence was generated independently residue by residue according to a background composition model. Also, $P(\mathbf{D})$ represents the probability of obtaining $\mathbf{D}$ according to a background distribution over domains. The left-hand bracket scores the fit of the domain sentence to the amino-acid sequence, while the right-hand bracket is the context dependent score.

A simplified view of the Pfam annotation process [BCD+04] is that a domain $\mathbf{D}$ annotating the sequence fragment $x_{[s_i,e_i]}$ is recognized as real if the domain log-odds ratio is greater than a manually curated threshold,

$$\log \frac{P(x_{[s_i,e_i]}|\mathbf{D}_i)}{P(x_{[s_i,e_i]}|\mathbf{R})} > \tau_{\mathbf{D}_i}. \tag{2.11}$$

This log-odds ratio is calculated using the HMMER package [Edd98]. The actual process is somewhat more complicated. As outlined in section 1.2, HMMER calculates the log-odds ratio that the model generated the full sequence x allowing for multiple matches of the domain model $\mathbf{D}_i$ to the sequence. This is called the *sequence score*. HMMER also calculates the contribution from each of the repeated domains $\mathbf{D}_i$, which is called the *domain score*. Pfam enforces a threshold on both the domain and sequence scores, whereas eq. 2.11 just shows the domain score threshold.

Comparison of eqs. 2.10 and 2.11 reveals that the standard approach is essentially equivalent to ignoring the context term . My approach is to maximise the sentence score $\mathbf{SS}_{x,T,M}$ given in eq. 2.10 over all domain sentences $\mathbf{D}$, using the Pfam domain threshold for $\tau_{\mathbf{D}_i}$, and the HMMER domain score for $\frac{P(x_{[s_i,e_i]}|\mathbf{D}_i)}{P(x_{[s_i,e_i]}|\mathbf{R})}$.

2.3.2 Context model and smoothing strategy

The combined taxonomic and language context model is parameterised by considering a different Markov language model $\mathbf{M}_T$ for each taxonomy T. Begin and end states are included in the modelling in order to capture associations of domains with the beginning and end of proteins. A Markov model of order k asserts that the conditional probability of the i^{th} domain given all preceding domains is only dependent on the k preceding domains:

$$P(\mathbf{D}_i|\mathbf{M}_T, \mathbf{D}_1 \ldots \mathbf{D}_{i-1}) = P(\mathbf{D}_i|\mathbf{D}_{i-k} \ldots \mathbf{D}_{i-1}). \tag{2.12}$$

The terms in eq. 2.12 are calculated using the observed counts in the Pfam database (denoted by $\mathbf{N}$) and are smoothed recursively using lower order domain contexts and higher taxa as described for speech recognition. In the following, T_0 denotes the species of the protein in question, T_j the j^{th} parent taxon and T_m is the root of the taxonomy. For a fixed taxon T_j the probabilities are smoothed over domain contexts:

$$\begin{aligned}\hat{P}(\mathbf{D}_i|\mathbf{M}_{T_j}, \mathbf{D}_{i-k}\dots\mathbf{D}_{i-1}) =&(1-\alpha)\cdot\left(\frac{\mathbf{N}(T_j, \mathbf{D}_{i-k}, \dots \mathbf{D}_{i-1}, \mathbf{D}_i)}{\mathbf{N}(T_j, \mathbf{D}_{i-k}, \dots \mathbf{D}_{i-1})}\right)\\ &+\alpha\cdot\hat{P}(\mathbf{D}_i|M_{T_j}, \mathbf{D}_{i-k+1}\dots\mathbf{D}_{i-1})\end{aligned} \tag{2.13}$$

$$\hat{P}(\mathbf{D}_i|\mathbf{M}_{T_j}, \mathbf{D}_{i-1}) = (1-\alpha)\cdot\left(\frac{\mathbf{N}(T_j, \mathbf{D}_{i-1}, \mathbf{D}_i)}{\mathbf{N}(T_j, \mathbf{D}_{i-1})}\right) + \alpha\cdot\hat{P}(\mathbf{D}_i|\mathbf{M}_{T_j}) \tag{2.14}$$

$$\hat{P}(\mathbf{D}_i|\mathbf{M}_{T_j})) = \frac{\mathbf{N}(T_j, \mathbf{D}_i)}{\sum_{\mathbf{D}}\ \mathbf{N}(T_j, \mathbf{D})}. \tag{2.15}$$

The sum in eq. 2.15 is over all domain occurrences in the Pfam database. The interpolation parameter α is a fixed constant between 0 and 1. Back-off estimation, as described for speech recognition, allows α to be a decreasing function of the amount of context $\mathbf{N}(T_j, \mathbf{D}_{i-k}, \dots, \mathbf{D}_{i-1})$. This was investigated and not found to significantly improve the classification.

Next, contributions from higher order taxa are recursively interpolated

$$\begin{aligned}P(\mathbf{D}_i|\mathbf{M}_{T_j}, \mathbf{D}_{i-1}\dots\mathbf{D}_{i-k}) =&(1-\beta)\cdot\hat{P}(\mathbf{D}_i|\mathbf{M}_{T_j}, \mathbf{D}_{i-k}\dots\mathbf{D}_{i-1})\\ &+\beta\cdot P(\mathbf{D}_i|\mathbf{M}_{T_{j+1}}, \mathbf{D}_{i-k}\dots\mathbf{D}_{i-1})\end{aligned} \tag{2.16}$$

$$P(\mathbf{D}_i|\mathbf{M}_{T_m}, \mathbf{D}_{i-k}\dots\mathbf{D}_{i-1}) = \hat{P}(\mathbf{D}_i|\mathbf{M}_{T_m}, \mathbf{D}_{i-k}\dots\mathbf{D}_{i-1}). \tag{2.17}$$

The parameter β represents the degree to which the estimation is based on nodes higher up in the taxonomy rather than the leaves. Note that this strategy is a smoothing strategy which recursively interpolates counts of species which are similar according to the NCBI taxonomy. In order to avoid over-fitting a taxonomy which has low coverage in Uniprot, only those nodes in the taxonomy below which there is a sufficient sample size, $10,000$ proteins in this case, are retained. For proteins which have a species T_0 which does not meet this criteria, T_0 is set equal to the first ancestor taxonomy in the modified taxonomy tree (which may be the root of the tree, if none of the kingdom-specific ancestor taxa meet the sample size criteria).

The interpolation parameters can be trained from training data which is held over from generating the counts for the context models. All that is required is some form of

objective function, and then an optimization technique can be used to find the parameters which optimise the objective function. In the results section, held data from the SCOP test are used to estimate these interpolation parameters.

2.3.3 Context score of a domain in a protein with fixed context

I need to consider how to score an arbitrary Pfam domain instance on a protein with fixed context (i.e. the other domains on the protein are already known). This is required for the SCOP test in section 2.4.1. My approach is to consider the difference between the sentence score $\mathbf{SS}_{x,T,M}$ for the domain sequence including and excluding the domain in question. Denote by $\mathbf{d}_l$ the Pfam family which I am scoring, and by $\mathbf{D}$ the fixed (pre-annotated) context of the protein such that no $\mathbf{D}_i$ in $\mathbf{D}$ overlaps with $\mathbf{d}_l$. Then, define the sentence score for a single domain as

$$\mathbf{SS}_{x,T,M}(\mathbf{d}_l) = \mathbf{SS}_{x,T,M}(\mathbf{D} \bigcup \mathbf{d}_l) - \mathbf{SS}_{x,T,M}(\mathbf{D} \setminus \mathbf{d}_l) \tag{2.18}$$

2.3.4 Dynamic programming algorithm

The space of all potential domain assignments for a particular protein is large, and hence an algorithm which concentrates on searching probable domain assignments is required. My approach is to first run HMMER against the protein for each Pfam family, keeping only those hits which have HMMER e-value less than 1000. In this way, a list $\mathbf{d} = \mathbf{d}_1 \dots \mathbf{d}_m$ of potential domains is obtained, ordered by end position, with corresponding amino acid fragments $x_{[s_i,e_i]}$. The search space is now all possible subsequences of domains in this list. The search through this reduced space is optimized using a dynamic programming technique.

Firstly, assume that the language model is a first order Markov model. In that case, the goal is to find the domain sentence $\mathbf{D} = \mathbf{D}_1 \dots \mathbf{D}_n$, a sublist of $\mathbf{d}$ which maximises the protein log-odds score $\mathbf{SS}_{x,T,M}(D)$, where

$$\mathbf{SS}_{x,T,M}(\mathbf{D}) = \sum_{i=1}^{i=n+1} H(\mathbf{D}_i) + C(\mathbf{D}_i|\mathbf{D}_{i-1}) \tag{2.19}$$

$$H(\mathbf{D}_i) = \text{HMMER}(\mathbf{D}_i) - \tau_{\mathbf{D}_i} \tag{2.20}$$

$$C(\mathbf{D}_i|\mathbf{D}_{i-1}) = \log\left(\frac{P(\mathbf{D}_i|\mathbf{D}_{i-1})}{P(\mathbf{D}_i)}\right). \tag{2.21}$$

Note that $H(\mathbf{D}_i)$ is just the HMMER score for the domain minus the threshold, and that $C(\mathbf{D}_i|\mathbf{D}_{i-1})$ is termed the transition score. Denote the begin and end states as $\mathbf{D}_0, \mathbf{D}_{n+1}$ respectively, so that $C(\mathbf{D}_1|\mathbf{D}_0)$ is the transition score coming from the begin state and $C(\mathbf{D}_{n+1}|\mathbf{D}_n)$ is the transition score going to the end state. As the end state contributes no sequence-based score, $H(\mathbf{D}_{n+1})$ is set to zero.

Define $\mathbf{D}^i$ to be the highest scoring domain sentence which ends in domain $\mathbf{d}_i$ without overlaps. The following recursion relation then applies:

$$\mathbf{SS}_{x,T,M}(\mathbf{D}^i) = H(\mathbf{d}_i) + \max_{e_j<s_i}\{\mathbf{SS}_{x,T,M}(\mathbf{D}^j) + C(\mathbf{d}_i|\mathbf{d}_j)\}, \tag{2.22}$$

where the condition $e_j < s_i$ ensures that the maximising sentence does not contain domain overlaps. Then set

$$\mathbf{D}^i = \{\mathbf{D}^j, \mathbf{d}_i\} \tag{2.23}$$

where $\mathbf{D}^j$ maximises eq. 2.22. Repeated application of eq. 2.22 and eq. 2.23 for $i = 1 \ldots m+1$ gives the maximising sentence $\mathbf{D} = \mathbf{D}^{m+1}$ required by eq. 2.19 (again, I use the convention that $\mathbf{d}_{m+1}$ is the end state, so that $\mathbf{D}^{m+1}$ is interpreted as the maximising sentence ending with the end state).

The assumption that the Markov model $\mathbf{M}$ is first order is now relaxed, and $C(\mathbf{D}_i|\mathbf{D}_{i-1})$ is replaced with $C(\mathbf{D}_i|\mathbf{D}_{i-1}\ldots\mathbf{D}_{i-k})$. Equation eq. 2.22 now becomes

$$\mathbf{SS}_{x,T,M}(\mathbf{D}^i) = H(\mathbf{d}_i) + \max_{e_{j_1}<s_{j_2}<e_{j_2}<\ldots<s_{j_k}}\{\mathbf{SS}_{x,T,M}(\mathbf{D}^{j_1,\ldots j_k}) + C(\mathbf{d}_i|\mathbf{d}_{j_k},\ldots\mathbf{d}_{j_1})\}, \tag{2.24}$$

and so the strategy outlined above is no longer guaranteed to return the highest scoring sequence under the language model. However, this strategy is still used in this case, and has been found to still work well in practice.

2.3.5 Incorporating the sequence score threshold

As mentioned above, Pfam uses a *sequence score* threshold in addition to the domain score threshold given in eq. 2.11. This thresholding is equivalent to a threshold on the sum of log-odds scores contributed by all instances of a particular domain type on a protein (for instance the sum of all of the zf-C2H2 domain scores). As the method applies Pfam thresholds, it must also apply a sequence score filter as a post-processing step to retain consistency with Pfam.

In order to do this, the maximising domain sentence is obtained as before. The total score for the maximising sentence comprises the sum of HMMER scores (left-hand bracket of eq. 2.10) and the context score (the right-hand bracket in eq. 2.10). As before, the total HMMER score for each type of domain on the maximising sentence is summed to give a sequence score for that domain type. Now, the context component of the score is distributed amongst each of the sequence scores such that as many domain types score above the sequence threshold as possible. To do this, assuming a positive context score, simply order the domain types according to sequence score and allocate to the first sub-threshold domain type as much context score is required to meet the sequence score threshold. Repeat this step until the context score has been completely distributed.

2.3.6 Variable length Markov model

The fixed-order Markov model has a significant drawback: the lengths of commonly occurring domain architectures are not fixed; some patterns are first order (CBS domains often occur in pairs), while many patterns have a higher order (the group of RNA polymerase RBP1 domains commonly occur in groups of seven). Restricting to a fixed order Markov model will degrade the ability of the model to recognize patterns of arbitrary length. Instead, for each proposed context $\mathbf{D}^j$ from eq. 2.22 in the dynamic programming algorithm, a different order k for M is chosen which is the maximum order which is observed in the training database. More precisely, labelling $\mathbf{D}^j = \mathbf{D}^j_1 \dots \mathbf{D}^j_{n_j}$ the order k is chosen to be the largest order with non-zero training set count $\mathbf{N}(\mathbf{D}^j_{n_j-k} \dots \mathbf{D}^j_{n_j})$. As this does not depend on the current domain $\mathbf{d}_i$, eq. 2.12 still defines a consistent probability distribution over domains. In practice, however, to cut down on memory requirements for storing counts of arbitrary length, I restrict $k \leq 4$.

This approach is an example of decision tree modelling which is commonly used in language modelling. Decision trees partition domain histories $\mathbf{D}^j$ into equivalence classes $\Phi_1 \dots \Phi_M$ with a corresponding probability distribution $P(\mathbf{D}_i|\Phi_l)$. My approach partitions on the basis of the longest domain context which has been observed in the training set. It is straightforward to develop more complicated decision rules, and this remains a basis for further investigation. My approach is also similar to the interpolated Markov chain approach used by Salzberg [SPD$^+$99] in gene prediction.

2.3.7 Incorporating Pfam clans

The Pfam project groups together closely related Pfam families into Pfam clans. Pfam enforces an overlap rule: the Pfam threshold must be set to ensure that no distinct significant Pfam family matches overlap. Clans were created to relax this rule – that is, two families from the same clan are allowed to have significant matches which are overlapping, and the family which scores highest above its own threshold is annotated as the matching Pfam family. From the point of view of language modelling of domains clans can be seen as variants of a single domain (in much the same way that different phonetic representations of a word are the same word). I have taken the approach that from a language modelling point of view, Pfam families from different clans are considered to be from the same family, and hence their counts are aggregated. This only applies for training and scoring the the transition scores $C(\mathbf{D}_i|\mathbf{D}_{i-1}, \mathbf{D}_{i-k})$ but the HMMER component $H(\mathbf{D}_i)$ remains specific to the domain which is being scored. Importantly, the threshold remains domain (not clan) dependent, as thresholds may still vary substantially within a clan (particularly if one clan member is a fragment of another).

Clans and context modelling have had a mutually beneficial existence in Pfam. Pfam annotators use context domain hits to guide their decisions about new clans to build, and grouping Pfam families into clans means that context modelling has more information (as more patterns are observed) with which to score domain architectures.

2.3.8 Significance scores

The Pfam database maintains for each domain hit an e-value score as well as a log-odds score. The e-value score for a domain is the number of hits which would be expected to have a score greater than or equal to the score of the domain in a random database of the same size. It is calculated for each Pfam family by fitting an extreme value distribution (EVD) to the bit scores of hits of that family against a set of randomly generated proteins, as implemented in the *hmmcalibrate* program of the HMMER package. The e-value score does not directly affect the assignment of domains in Pfam as manually created thresholds are used instead. However, the significance of domain matches is important to consider as it is used by end users when evaluating marginal hits. Moreover, significance scores can be used to compare the reliability of hits from different Pfam families, whereas log-odds scores cannot. Significance values are

required in the SCOP test to generate aggregate ranked lists of domain matches. Thus it is important to consider the effect of language modelling on significance scores.

One possibility is to use the unmodified EVD parameters calculated by *hmmcalibrate* to calculate the significance of HMMER+context scores. This is the approach pursued in the SCOP test in section 2.4.1. An alternative strategy is to score the HMMER+context model on randomly generated proteins in order to generate a modified EVD. As the significance score relates to a particular domain rather than the entire domain sentence, the method described in section 2.3.3 is used to calculate the HMMER+context score for the domain as the difference of the HMMER+context score of the maximising sentence with and without the domain in question. As in *hmmcalibrate* the HMM is required to pass through the given domain at least once. Note that in almost all cases, the language model uses a start $\rightarrow$ domain $\rightarrow$ end architecture as it finds no other domains with scores above threshold to include in the calculation. In this case, all of the start to domain and domain to end transition scores will be attributed to the domain.

This process is demonstrated on two Pfam families, WD40 and pkinase as shown in fig. 2.6. Two different types of behaviour are observed. In one case, pkinase commonly occurs by itself on a protein, and hence hits to random proteins typically have their scores enhanced slightly by the language model, so that the EVD shifts to the right. However, real hits also have their scores enhanced. Furthermore, in the case of a single domain protein, the increase will be the same as the shift in the EVD, so that the significance of the hit remains unchanged. In contrast, hits to the pkinase domain in atypical contexts will not have their scores enhanced, and so their significance will decrease. The other example, WD40 commonly occurs in repeats of 5-8 units; so that individual random hits are penalised under the language model (by about 4 bits) and so the EVD shifts to the left. The language model enhances the score of real hits (as they do occur in the appropriate repeating pattern), thus providing the compound effect of increasing the score of real hits and increasing the significance of hits at a given score. To summarise, the effect of language modelling on significance scores appears to be either neutral, in the case in which the scores of random and real hits are shifted by the same amount, or more discriminatory, in the case of decreasing random scores and increasing real scores.

A weakness of this approach to calculating significance scores is that it considers random

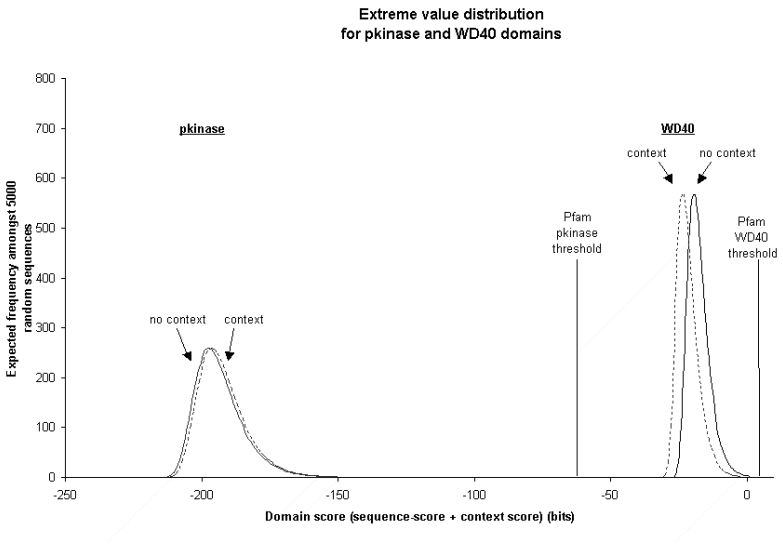

Figure 2.6: Extreme Value Distribution (EVD) curves calculated for pkinase and WD40 Pfam domains. The solid lines are the standard EVD curves calculated using HMMER. The dashed lines use the language modelling method, and hence take contextual information into account. For almost all sequences, this results in a domain sentence consisting of a BEGIN state followed by the given domain and ending in an END state. WD40 is commonly found in groups of 5 to 8 tandem repeats, so that single random WD40 hits are penalised by the language model. The WD40 EVD shifts 4.0 bits to the left. On the other hand pkinase often occurs by itself on a protein, and hence random single pkinase repeats gain slightly under the language model. The pkinase EVD shifts 1.1 bits to the right.

proteins for calculating the language model component of the score, whereas false positive hits in real proteins do not have random protein as context. This has not been further investigated.

2.3.9 Implementation

A major implementation challenge was to store efficiently in memory the counts of occurrence patterns of domains and species used in eqs. 2.13, 2.14, 2.15. These counts are central to the dynamic programming algorithm described above, and speed of accessing these counts is critical. Note that the counts are stored, rather than the smoothed probabilities, as the space of possible domain and taxonomy combinations is much vaster than the space of observed combinations. A context map is stored, which contains as keys every observed 1-mer, 2-mer .. k-mer observed in Pfam (with k normally set at 4). These keys map to a secondary map, in which each observed taxonomy from the reduced taxonomy tree maps to the number of observations of the given domain sequence k-mer in proteins of this taxonomy or with the taxonomy as ancestor. In order to facilitate rapid access to the counts to compute the smoothed probabilities, the context map is stored as a red-black tree [CLRS01]. The smoothing equations eq. 2.13 recursively interpolate from higher-order to lower-order contexts. However the counts are stored and accessed in the reverse order, progressively narrowing down from general to more and more specific contexts. This is achieved by first constructing an ordering of Pfam domains. This ordering is used to infer an ordering on domain sequences working from right to left – that is for two given domain sequences the final position is first compared, then, if this is equal the penultimate position is compared, etc. If the two domain sequences are equal in all positions, but one is shorter than the other, then the shorter sequence is ordered ahead of the longer sequence. This ordering is used to create the red-black map. Now consider eq. 2.13. The successive numerators require the counts $\mathbf{N}(T_j, \mathbf{D}_{i-k}, \ldots \mathbf{D}_i)$, $\mathbf{N}(T_j, \mathbf{D}_{i-k+1}, \ldots \mathbf{D}_i)$, ... $\mathbf{N}(T_j, \mathbf{D}_i)$, which are obtained in reverse order. Firstly the node in the red-black map is found below which all domain sequences end in $\mathbf{D}_i$, as all subsequent counts will be from this sub-tree. The first position in this sub-map is the last of the counts required. This process is continued, progressively narrowing down the sub-tree of counts. For successive denominators, the same strategy can be pursued, but starting with all domain sequences ending in $\mathbf{D}_{i-1}$. The counts over all ancestral T_j are all collected at the same stage. This search is not optimized in the same way as the maps are much smaller, so a standard

hashing strategy is sufficient.

As an example, consider scoring an example transition $C(\mathbf{D}_i|\mathbf{D}_{i-k},\dots\mathbf{D}_i)$. Consider the domain sequence $\mathbf{D}_0$ =BEGIN, $\mathbf{D}_1$ = C2-set, $\mathbf{D}_2$ = ig and I will show how to calculate eq. 2.21. Let the taxonomy of the protein be (Eutheria, Coelomata, Eukaryota, root). I assume $\alpha = 0.7$ and $\beta = 0.35$. The array of counts required for smoothing is given as follows:

	T_0	T_1	T_2	T_3
$\mathbf{N}(T_j,\mathbf{D}_0,\mathbf{D}_1)$	3224	43394	5029	5210
$\mathbf{N}(T_j,\mathbf{D}_1)$	17894	255714	29972	30256
$\mathbf{N}(T_j)$	379460	618859	1376701	3005810

(2.25)

	T_0	T_1	T_2	T_3
$\mathbf{N}(T_j,\mathbf{D}_0,\mathbf{D}_1,\mathbf{D}_2)$	2428	3355	3946	4011
$\mathbf{N}(T_j,\mathbf{D}_1,\mathbf{D}_2)$	12132	17657	21018	21119
$\mathbf{N}(T_j,\mathbf{D}_2)$	17894	25571	29972	30256

(2.26)

From these counts, I calculate the probabilities, and finally log-odds scores.

	T_0	T_1	T_2	T_3
$\frac{\mathbf{N}(T_j,\mathbf{D}_0,\mathbf{D}_1,\mathbf{D}_2)}{\mathbf{N}(T_j,\mathbf{D}_0,\mathbf{D}_1)}$	0.75	0.77	0.78	0.77
$\frac{\mathbf{N}(T_j,\mathbf{D}_1,\mathbf{D}_2)}{\mathbf{N}(T_j,\mathbf{D}_1)}$	0.68	0.69	0.70	0.70
$\frac{\mathbf{N}(T_j,\mathbf{D}_2)}{\mathbf{N}(T_j)}$	0.05	0.04	0.02	0.01
$P(\mathbf{D}_2\|T_j,\mathbf{D}_0,\mathbf{D}_1)$	0.39	0.40	0.39	0.38
$P(\mathbf{D}_2)$				0.01
$C(\mathbf{D}_2\|\mathbf{D}_0,\mathbf{D}_1)$				5.24

(2.27)

So the transition score is 5.24 bits. In other words, in eutherian mammals it is $2^{5.24} = 38$ times more likely to see a ig as the second domain in a protein following a C2-set domain than it is in a random protein.

2.4 Results

Figure 2.7 shows the processes carried out this chapter. The results are split into two sections, the SCOP test and the Pfam scan. The training set for the language model consisted of Pfam release 15 and proteins from the Uniprot [ABW+04] database consisting of Swissprot release

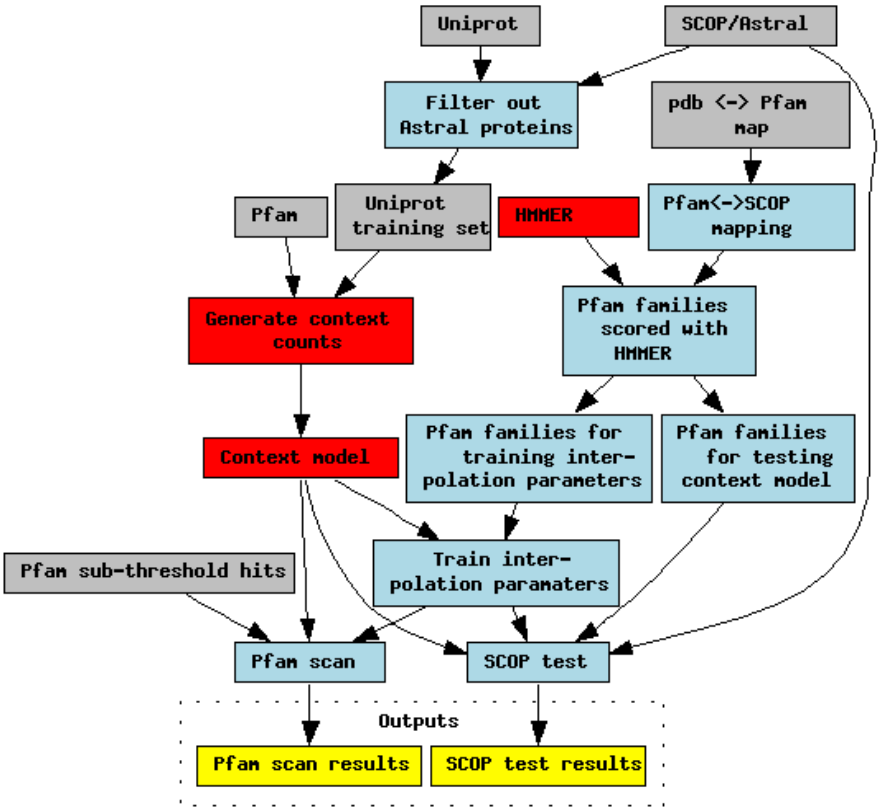

Figure 2.7: Conceptual diagram of processes and data in this chapter. Inputs are shown in grey, and outputs in light blue, with intermediate steps in yellow, and software steps in red.

44.0 and SP-TrEMBL27.0, with all proteins which match proteins from the ASTRAL protein set (filtered to a maximum of 40% identity between any two proteins in the set) removed. This set consisted of 982, 523 proteins, which only includes those proteins which have at least one annotated Pfam domain.

2.4.1 SCOP test

In order to test objectively the ability of the language model to detect protein domains, I use the SCOP test, initially developed by Brenner et al. [BCH98] and subsequently used by many authors to evaluate homology prediction algorithms (e.g. [MG02]). The SCOP database classifies all proteins of known structure [HMBC97] in terms of protein domains. Multi-domain proteins are split into component protein domains, which are classified hierarchically in four

levels: family, superfamily, fold and class. Sequences belonging to the same family share sequence similarity, suggesting a common function and implying a clear common evolutionary origin; families are clustered into superfamilies on the basis of structural similarity suggesting a probable common evolutionary origin; superfamilies are grouped into folds on the basis of similar secondary structure topology. ASTRAL is a database of protein sequence fragments of known structure, annotated with SCOP family classifications [CWC$^+$02]. ASTRAL provides protein sequences filtered to various levels of sequence similarity. The SCOP test works by running a given algorithm and domain model over all proteins classified by SCOP, and comparing domain family predictions with the known structural class. In this way it is possible to independently identify proteins homologous to the given domain family (all proteins belonging to the same SCOP superfamily) and proteins which are non-homologous (all proteins belonging to a different SCOP fold). Proteins belonging to the same fold but different superfamily are not classified as homologous or non-homologous.

The SCOP test was modified in order to apply it to the domain models in Pfam. Using the file *pdbmap* (available at ftp://ftp.sanger.ac.uk/pub/databases/Pfam/pdbmap) I obtain a list of all proteins in which a Pfam domain annotation overlaps a PDB structure . The coordinates (with respect to the Uniprot protein sequence) of both the PDB structure and the Pfam domain are provided in this file. The PDB structure is classified by SCOP. Providing a given Pfam domain overlaps one and only one SCOP superfamily, I classify all SCOP proteins in this superfamily as homologous to the Pfam domain, and all proteins outside the fold to which this superfamily belongs as non-homologous. I identify 1970 Pfam families in Pfam release 15 which satisfy this criteria. Of these, I use 500 to train the interpolation parameters of the context models. The remaining 1470 form the test set of Pfam families for the SCOP test.

For a given algorithm and a given Pfam family, the SCOP test proceeds by scoring every protein in the ASTRAL filtered sequence set (to a maximum of 40% shared identity in this case), and generating a list of proteins ranked according to model log-odds score. The ultimate goal of homology detection is to score all homologous proteins above all non-homologous proteins. One simple measure of relative success is the number of true homologies scored above the highest scoring non-homologous sequence, which I shall refer to as the 'over the top' score (OTT). An alternative is a *coverage vs error* curve which plots at each point in

the ranked list the total number of homologous proteins (true-positive) above this point on the y-axis vs the number of non-homologous (false positive) above this point on the x-axis. A randomly ranked list would give on average an equal proportion of homologous and non-homologous sequences identified. For a given error rate, a higher curve is a more effective classifier of homologous proteins, and the area under the curve is another measure of overall success. The minimum error rate (MER), which is the minimum of the sum of number of homologous sequences classified as non-homologous and non-homologous sequences classified as homologous, can also be used.

If instead of ranking according to model score the list is ranked according to e-value significance, then it is possible to generate an aggregated ranked list of significance across multiple domain models. From this list a score representing the effectiveness of the algorithm across all domain families can be obtained, using either the OTT, MER or area under the coverage versus error curve. The ranking by significance is necessary as the log-odds scores between models are not comparable.

The SCOP test was carried out on the following variants of the context models described in the previous section: HMMER alone; HMMER with a digram language model, denoted HMMER+2gram (which implies that a single domain is considered as context); HMMER+3gram; HMMER+4gram; HMMER with taxonomy context (denoted HMMER+taxonomy) and HMMER+4gram+taxonomy. It can be seen from the following results that the 4gram model is a small improvement on the 3gram model. A HMMER+7gram+species model was tested to observe the effect of longer context, but it was not found to improve results beyond the HMMER+4gram+context model.

In order to apply the language models, it was necessary to identify the protein sequence in Uniprot which matched each of the protein fragments in the ASTRAL set, so that I could use Pfam to assign the domain context and also obtain the taxonomic position of the protein from the NCBI taxonomic code assigned by Uniprot to each protein. As noted above, ASTRAL contains protein fragments, so it is also necessary to assign the correct position of the protein fragment on the Uniprot protein. This is achieved with the *pdbmap* file discussed above. The HMMER+context model score for a particular Pfam domain was obtained as the difference between the context score of the full domain sequence including the Pfam domain and the context score excluding the context domain, as given by equation eq. 2.18.

The interpolation parameters were trained on 500 of the 1970 Pfam families with the remainder forming the test set of Pfam families. The sum of individual family OTT scores was used as the objective function to train the taxonomy and domain context interpolation parameters. This score was chosen as it replicates most closely the objective of improving Pfam annotation, for which a threshold is manually curated for each family with the aim that there are no false positives. The optimal parameters from this set were $\alpha = 0.7$ for domain context, and $\beta = 0.35$ for taxonomic context.

Figure 2.8 displays the coverage versus error curve over all Pfam domains tested (with the results ranked by significance). HMMER+4gram+taxonomy identifies 3% more homologous proteins at an error rate of 1000 proteins. Table 2.4.1 shows summary measures of the performance of each of the context models. From the point of view of using the method to improve Pfam domain annotation, the important measure is the sum of family OTT scores (column 4). HMMER+4gram+taxonomy improves this measure by 2.2%, implying that if the Pfam thresholds could be optimally selected, context models could increase the number of domains annotated by 2.2%. HMMER+4gram+taxonomy is substantially better under this metric than HMMER+4gram, indicating that taxonomy is useful in improving the context models. Taxonomy on its own generates a smaller improvement than the 4-gram but better than the 3-gram language model.

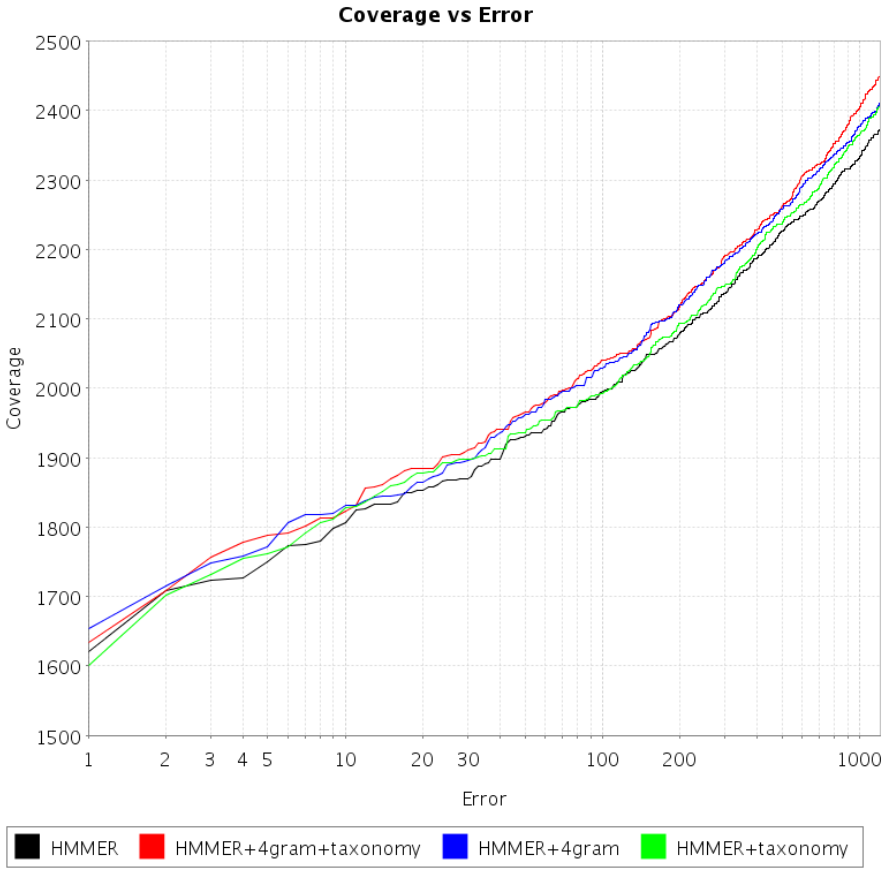

Figure 2.8: Coverage vs error curve for detection of remote homologies for aggregated results from 1470 Pfam families not used for training the interpolation parameters. The lines are black - HMMER score, green- HMMER+taxonomy , blue - HMMER+4gram, the red - HMMER+4gram+taxonomy. A higher line indicates a better classification of remote homologies. I display only up to 1000 false positives.

Method	# families with OTT		sum of family score		Aggregate score	
	Better	Worse	OTT	MER	OTT	MER
HMMER	-	-	3604	5092	1620	3692
+2gram	37	15	3638	5042	1650	3668
+3gram	46	20	3646	5041	1655	3662
+4gram	50	22	3657	5031	1654	3665
+taxonomy	53	34	3644	5064	1601	3687
+taxonomy +4gram	69	39	3682	5017	1634	3650

Table 2.1: Comparison of context models with HMMER, scored over the 1470 families not used for training the interpolation parameters.

For each method the number of false positive and false negative matches at a given e-value significance is plotted in figure 2.9. Context models improve error rates over a range of e-values less than 1.0 by reducing false negative matches with negligible impact on false positive matches. This demonstrates that at a given e-value threshold, HMMER+4gram+taxonomy has a lower error rate than HMMER alone. From the point of view of large scale classification of protein homology with profile HMMs this is an important result, as classification is often done on the basis of a global e-value threshold. This figure justifies to a certain extent the use of the same EVD on context adjusted scores, in that the false positive error curve is correctly calibrated with the HMMER false positive score. Note that no false positives are obtained with evalue of 10^{-3} or lower.

Figure 2.10 displays the domains which have the greatest increase and decrease in OTT score. In particular, C2-set gains 12 domains while Semialdehyde_dh loses 3 domains. In some cases the increase obtained by using a joint model is greater than the sum of the individual OTT score increases of the 4gram and context models (for example Laminin_EGF).

One family with significant improvement is the C2-set domain. C2-set is a member of the immunoglobulin superfamily clan in Pfam, and commonly co-occurs with other immunoglobulin superfamilies on a protein. HMMER alone scores 6 positive sequence from the

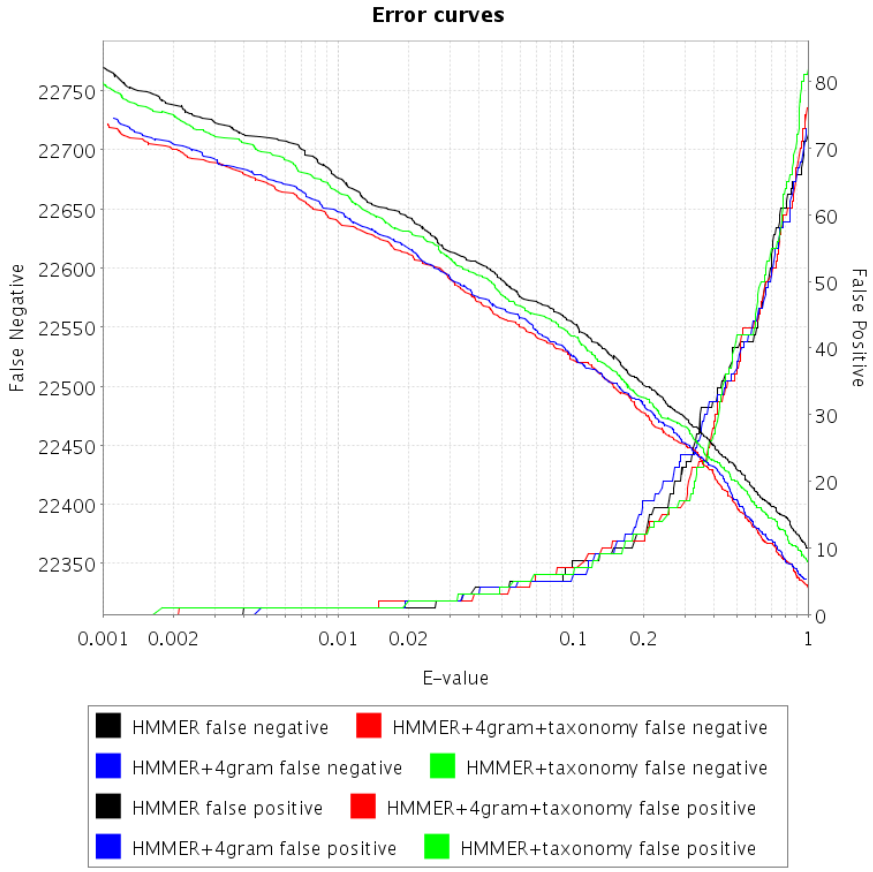

Figure 2.9: Number of false negative (upper six lines) and false positive (lower six lines) matches versus e-value threshold for HMMER (red lines) and context models. At a given e-value threshold, each of the models decreases false negative rates with negligible impact on false positive rates.

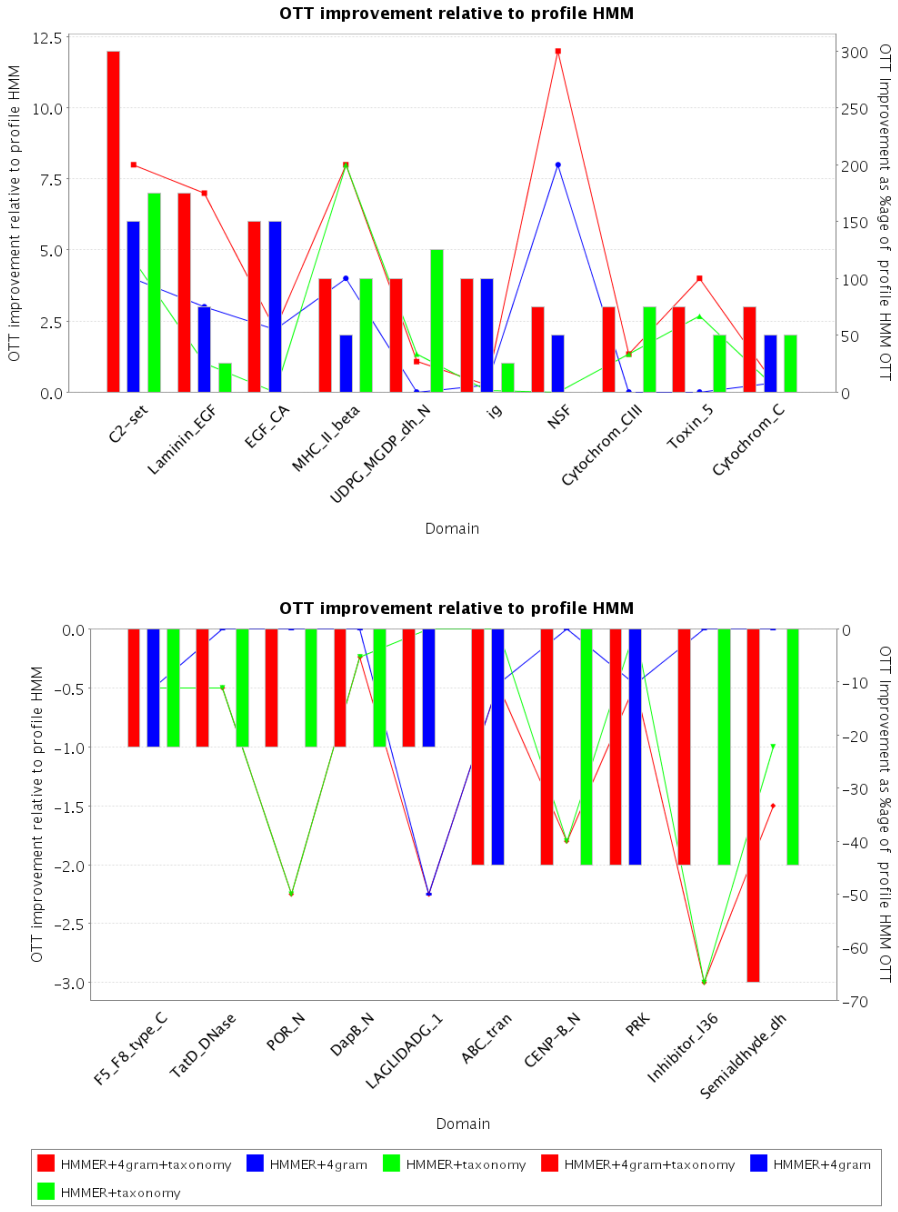

Figure 2.10: Pfam domains which have their OTT scores improved the most (upper graph) and decreased the most (lower graph), with OTT improvements relative to HMMER alone plotted for HMMER+taxonomy (blue), HMMER+4gram(red) and HMMER+4gram+taxonomy(black). The bars indicate absolute increase or decrease in OTT score, indicated on the left-hand y-axis. The lines indicate the percentage increase (or decrease) in OTT score, indicated on the right-hand y-axis.

ASTRAL test set above the first negative sequence, whereas HMMER+context+species scores 18 sequences above the first negative sequence. This improvement is obtained by increasing the significance of 12 homologous low significance scores and decreasing the significance of 3 non-homologous high significance scores. In the Pfam annotation, this domain is restricted to eutheria (placental mammals), however other members of immunoglobulin superfamily clan occur frequently in other verterbrata, and less commonly in other metazoa. The improvement in classification includes 11 vertebrate proteins and 1 insect protein. Figure 2.11 displays the significance scores for both HMMER and HMMER+4gram+taxonomy on this family.

2.4.2 Pfam scan

I scanned the Uniprot [ABW+04] database with all Pfam models to search for novel hits to these models. The same interpolation parameters were used as in the SCOP test. A HMMER+4gram+taxonomy language model was used, as the SCOP test demonstrated that this is the most sensitive of those context models tested.

The Pfam scan identifies 44792 new domain instances, which corresponds to 2.8% of the total number of domains previously scored as significant in Pfam under full-length models (Pfam also scores partial matches to Pfam domains). The new domain instances occur on 26458 proteins (which corresponds to 1.8% of the total number of proteins in Uniprot) and 3479 proteins which previously had no Pfam annotation (which corresponds to a 0.2% increase in sequence coverage). The new domain instances cover an additional 1.8m residues (Pfam full-length models previously covered 246m of 470m residues in Uniprot) which corresponds to a 0.38% increase in residue coverage. The new predictions are limited to 1245 domains, of which 344 domains contribute 95% of the new domain instances.

Figure 2.12 displays the families that the method detects. Figure 2.13 displays the length distribution of both new domains detected using context and the current Pfam annotation. Context domains have average length of 44 residues; the average length of Pfam domains is 183 residues. This is due to the over-representation of repeats in short Pfam families (and hence better contextual information) and a lower sequence-based signal-to-noise ratio for short families so that extra information is more likely to make a difference in detecting them.

Figure 2.14 shows how the impact of context varies across the taxonomic tree. In

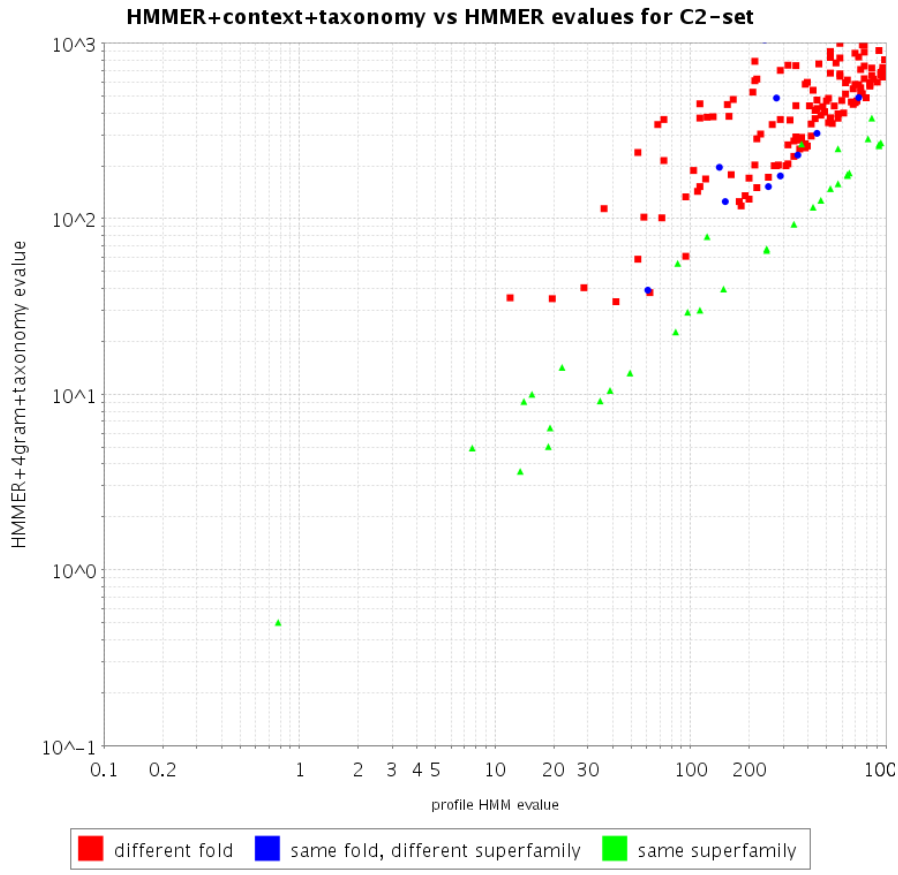

Figure 2.11: E-value significance scores for HMMER+4gram+taxonomy vs HMMER for C2-set domain, plotted on a log-log scale. The green dots represent sequences in the same SCOP superfamily (which are treated as homologous). The red dots represent sequences in different SCOP folds (which are treated as non-homologous). The blue dots represent sequences in the same SCOP fold but different superfamily (which are treated as neither homologous or non-homologous). Note that the four most significant matches (with e-value less than $1e-8$ under both HMMER and HMMER+4gram+taxonomy) are not shown. All 31 homologous sequences shown on this graph (green dots) fall below the $y = x$ line, and hence are more significant under HMMER+4gram+taxonomy than under HMMER.

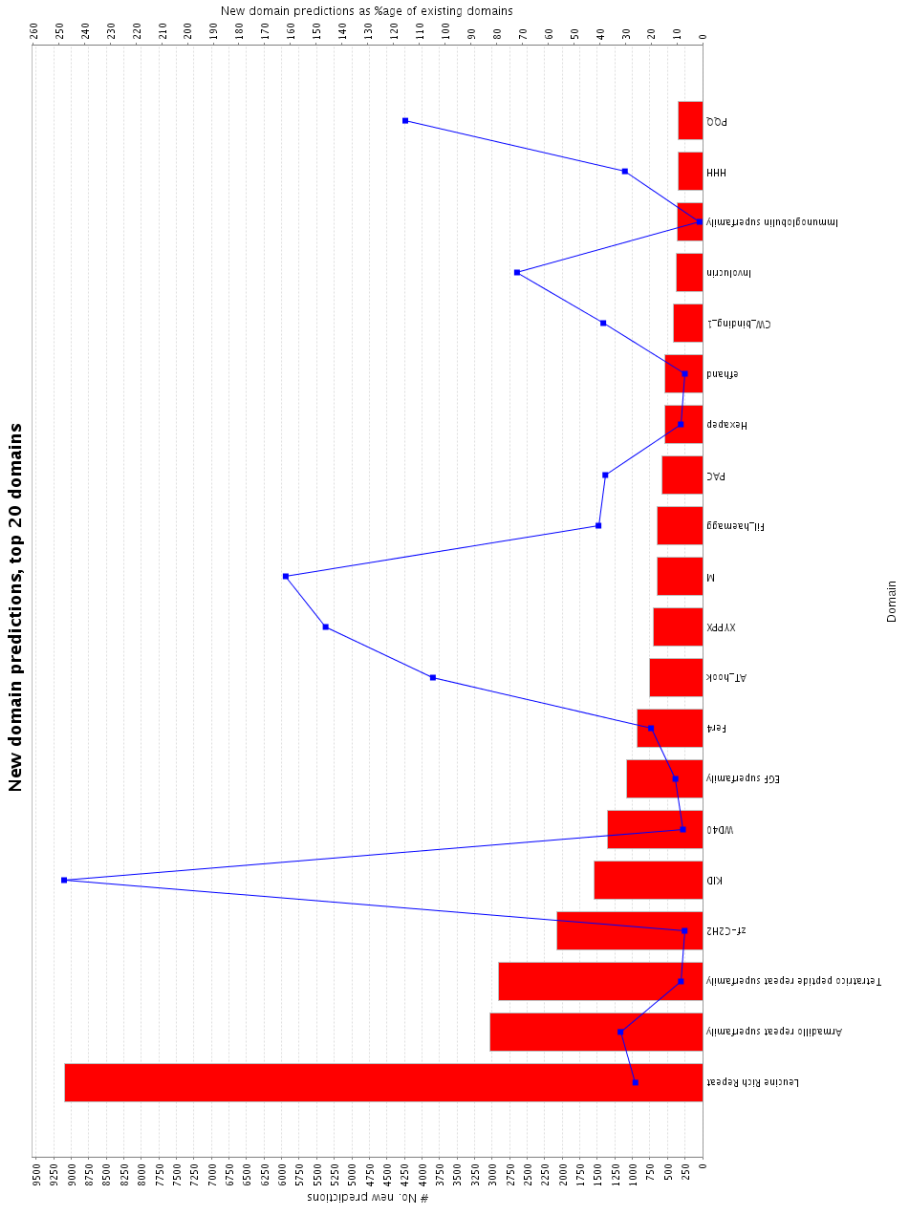

Figure 2.12: Domain occurrences amongst top 20 'context' families. The bars shows the absolute number of new predictions; the line line shows the percentage increase in that family.

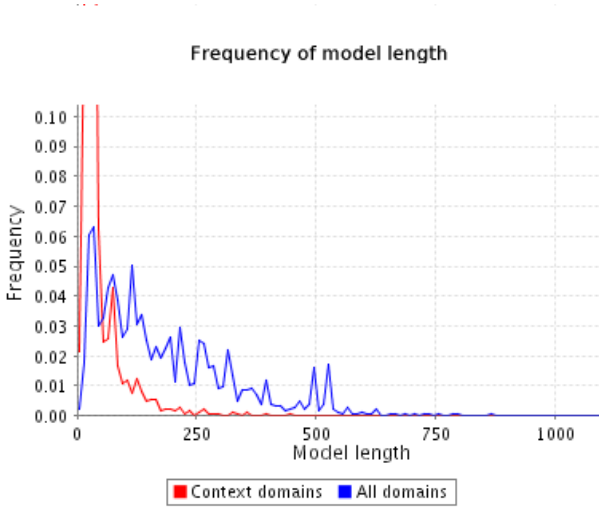

Figure 2.13: Length distribution for context domains(red) and Pfam domains (blue).

particular, context is not particularly effective in annotating Virus proteins. One possible explanation is that almost half the virus proteins in Uniprot are HIV proteins, and most of these are homologous proteins from different HIV strains, hence represents a much smaller pool of proteins with different domain architectures, each of which is already well understood. Context increases the number of Pfam annotations in bacteria and archaea by approximately 2% which is slightly below the average result. Context performs particularly well on eukaryotic proteins, increasing coverage by up to 6%. Table 2.2 suggests a weak relationship between the average number of domains per protein annotated with at least one Pfam domain and the increase in context domains.

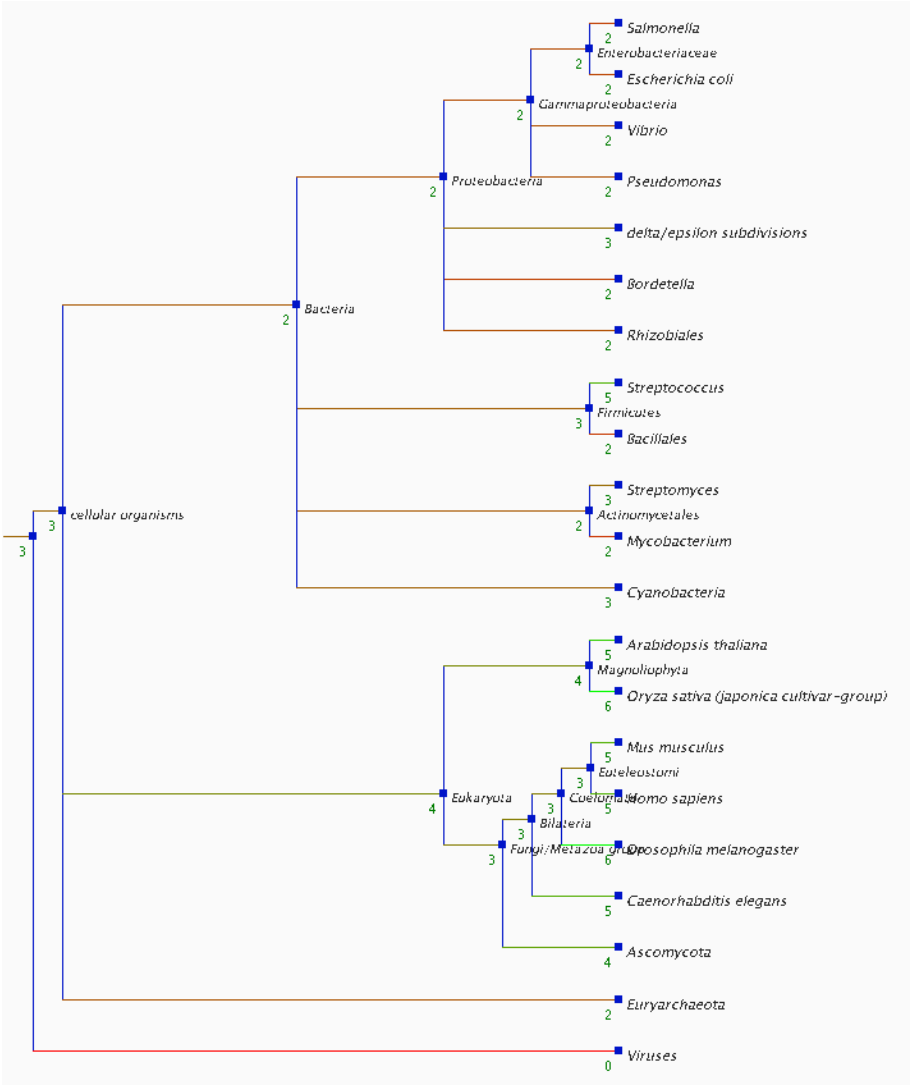

Figure 2.14: Percentage increase in domain occurrences by position in taxonomic tree. Each of the taxa displayed have more than 10,000 proteins in Uniprot (counting nodes which have the given node as ancestor). Nodes which have a single parent have been removed (for example HIV). Each node is annotated with the percentage increase in domain instances given by context at that level in the taxonomic tree. The branches above a given node are coloured according to the percentage increase, from green (high increase) to red (low increase).

Taxonomy	Percentage increase due to context domains	Average no. of domains per Pfam annotated protein
Drosophila melanogaster	6.4	2.5
Oryza sativa	6.4	2.0
Homo sapiens	4.6	2.6
Eukaryota	3.6	1.9
Bacteria	2.5	1.5
Archaea	2.4	1.5
Viruses	0.4	1.4

Table 2.2: Percentage increase in domain annotations due to context and average number of domains per protein annotated with at least one domain.

Figure 2.15 shows several examples of domains found by the context models without taxonomic context. Two TPR domains are found on the SR68_HUMAN protein, which has no TPR domains annotated in any of the protein databases. This protein is known to interact with SR72_HUMAN in the signal recognition particle [LPA+93], which itself has a pair of annotated TPR domains. As TPRs are protein-protein interaction motifs, this suggests that the interaction between SR68 and SR72 may be mediated by this region. On the previously un-annotated E2BG_CAEEL protein I find an NTP_transferase domain, followed by three hexapep repeats, all raised above the noise by their mutual compatibility.

The method also predicts a previously un-annotated Tf_Otx domain in the cone rod homeobox protein (CRX), in *H. sapiens*, *R. norvegicus* and *M. musculus* (figure 2.15). CRX is a 299 amino acid homeodomain transcription factor which is primarily expressed in the rod and cone receptors of the retina [CWN+97, FMC97]. CRX is highly conserved amongst mammalian species. CRX is known to share homology with Otx1 and Otx2, and contains a homeodomain near the N-terminus followed by a glutamine rich region, a basic region, a WSP motif, and an Otx-tail motif. The new Tf_Otx prediction extends over the un-annotated region: amino acids 164 to 250. This region encloses a valine to methionine mutation at position 242 associated with autosomal dominant cone rod dystrophy, which leads to early

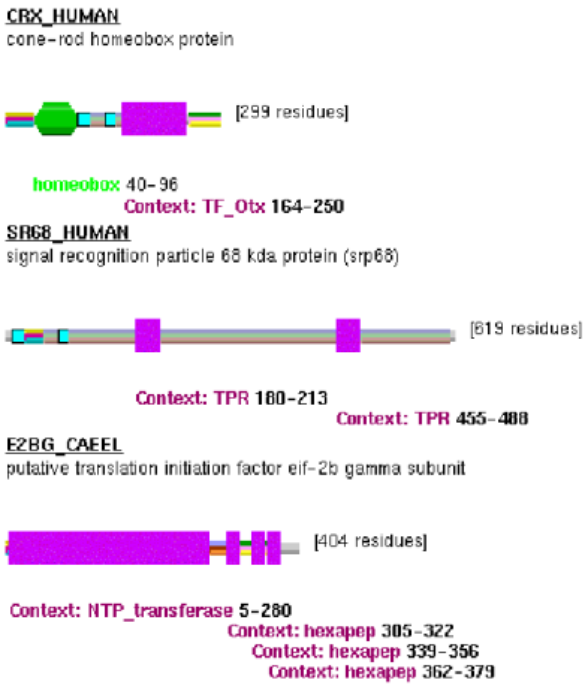

Figure 2.15: Examples of new context domains, indicated by rectangles. Standard Pfam domains are indicated by angled boxes. These domains can be identified using only a domain context model, without considering taxonomic context.

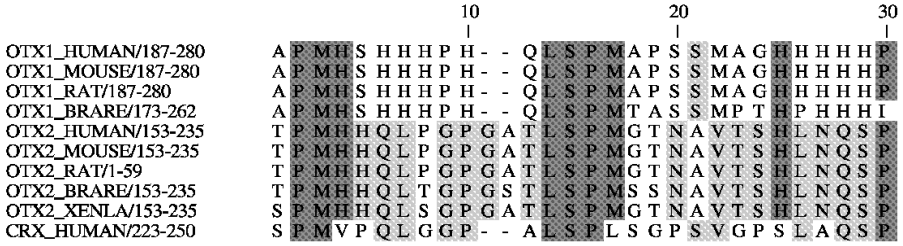

Figure 2.16: Part of multiple alignment of Tf_Otx domain in members of the Otx1 and Otx2 sub-families. Position 22 in this alignment - corresponds to position 242 on CRX_HUMAN. This position is methionine for all members of the Otx1 subfamily while it is valine for all members of the Otx2 sub-family.

blindness [SCW+97, RBD01]. Recent research demonstrates that a region coinciding with the new prediction (amino acids 200 to 284) is essential for transcriptional activation of the photo-receptor genes, and supports the hypothesis that the V242M mutation acts by impairing this transactivation process [CWX+02]. An analysis of the multiple alignment of the Tf_Otx domains (figure 2.16), demonstrates the existence of two sub-families of the domain, the first of which has a methionine at position 105 and contains all Otx1 proteins, the second of which has a valine at position 105 and contains all Otx2 proteins. Furthermore, the CRX V242M mutation aligns with this position and hence transfers the CRX Tf_Otx domain from the Otx2 to Otx1 sub-family. Finally, note that it has been demonstrated that both Otx2 and CRX transactivate the inter-photo receptor binding protein (IRBP) [BBI+99], while this has not been demonstrated for Otx1. This suggests that the V242M mutation loss of function is due to loss of IRBP transactivation ability, and conversely that position 105 in the Tf_Otx motif is critical for IRBP transactivation.

Figure 2.17 shows further examples of new domain occurrences found by considering taxonomic context only. A pair of TPR repeats are found in Aspartyl (asparaginyl) beta-hydroxylase (Q9Y4J0). This protein has been shown to be over-expressed in an enzymatically active form in hepatocellular carcinoma and cholangiocarcinoma[LJN+96]. The enzyme acts by catalyzing post-translational hydroxylation of β carbons of aspartyl and asparaginyl residues in EGF-like domains with the appropriate consensus sequence. In particular, the Notch homologues – which are known to be involved in cell differentiation and have been shown to be oncogenic – have the appropriate consensus sequence. TPR domains are thought to be involved in protein-protein interactions[DCB98], and may therefore help to mediate this

interaction.

The method also identifies novel antistasin domain on the theromin protein (THBI_THETS) in *Theromyzon tessulatum*, a leech. This protein has important medical applications as a potent thrombin inhibitor, and is found in the head of the leech [SCB$^+$00]. The antistasin family is an inhibitor of trypsin-family proteases and is often found in anti-coagulants. Thus the function of the protein concurs with the novel domain occurrence. Taxonomic modelling also find a novel occurrence of the toxin_2, or scorpion short toxin domain on the ErgToxin protein (Q9GQ92) in *Centuroides noxius* (Mexican scorpion). The ErgToxin protein blocks the ERG-K+-channels of nerve, heart and endocrine cells [SBF$^+$00]. Other members of the toxin_2 family also inhibit potassium channels.

Finally, in the fertilization 18kda protein (Q25063) in *Haliotis fulgens* (Green Abalone), a novel Egg_lysin domain is identified. Egg_lysin is found in other *Haliotidae*, as well as other *Archaeogastropoda*. The 18kda fertilization protein acts in conjunction with a paralogous 16kda lysin protein on the egg vitelline envelope. The 16kda protein creates a hole in the vitelline envelope. The 18kda protein is a potent fusagen of liposomes, and is thought to mediate membrane fusion between the gametes, a step in gamete recognition which is important in restricting heterospecific fertilization with other species [SV95]. These authors also found very high divergence amongst the group of orthologous 18kda proteins in California abalone; together with a high frequency of non-synonymous to synonymous substitution, indicating a high selective pressure toward differentiation between species and thus furthering the gamete recognition hypothesis. Furthermore, the 18kda protein exhibits a rate of evolution 2-3x that of the 16kda protein. The 18kda protein in *Haliotis fulgens* is the most distantly related of this group (with 27%–34% identity to the others), and hence standard profile methods fail to detect the similarity.

I had validated the predictions of an earlier version of this method using a Psi-blast [AMS$^+$97] test (table 2.4.2). This test was performed on a set of new domain predictions using Pfam 7.7, and an earlier version of the language modelling software which did not take into account the taxonomic context of sequences. For each novel predicted domain occurrence, Psi-blast was used to generate a set of similar sequence fragments. These sequences were then searched for matches to Pfam families. For 30.7% of novel domain occurrences Psi-blast found matches that are annotated in Pfam. In 90.0% of these the majority of annotations matched the

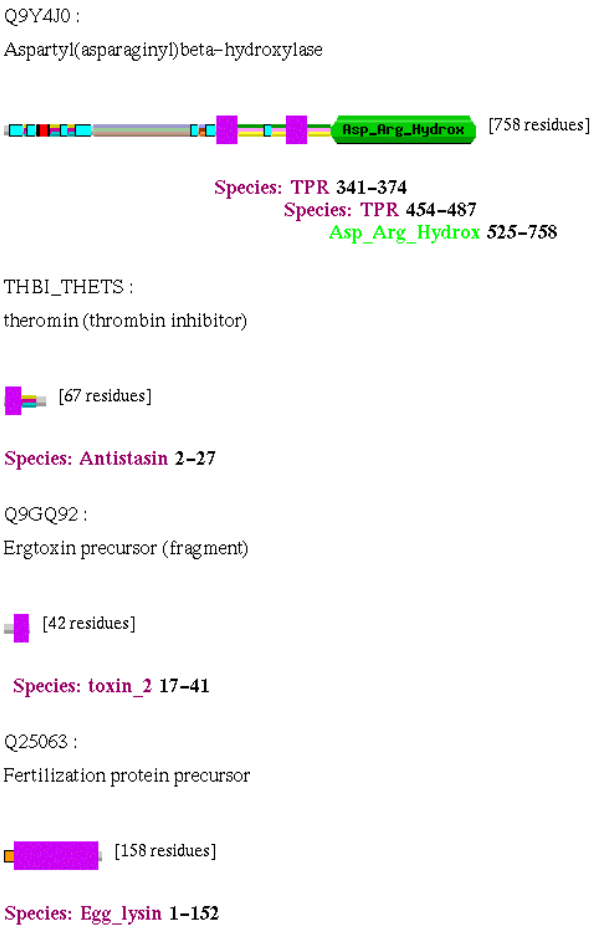

Figure 2.17: Emergence of new domains occurrences, identified using HMMER+taxonomy, indicated by magenta boxes and 'Species:' labels. Standard Pfam domains are indicated by angled boxes. These domains can be found by modelling taxonomic context without also considering domain context.

identified family; a further 7.6% had at least one match to the correct family; 0.8% matched a related family and for the remaining 1.5% all matches were to incorrect families. By inspection, the assignment due to the language modelling method of this paper appears to be correct for the overwhelming majority of the 7.6% and 0.8%, and many of the 1.5%. This suggests that the false positive rate is no more than a few percent. Since many of the 69.3% novel predictions for which Psi-blast does not find a match have higher scores than those for which it does, this also indicates the approach can detect matches which Psi-blast does not.

Psi-blast does not find match in Pfam Family		10,575	69.3%
Majority of matches to correct Pfam family		4,220	27.6%
Majority of matches to incorrect family	Has 1 match to correct family	358	2.3%
	Has matches to related family	38	0.3%
	All matches to unrelated families	72	0.5%

Table 2.3: Blast Results For New Positives Predicted By Model.

2.5 Discussion

I have demonstrated that significant improvement in protein domain detection is possible through modelling domain context using techniques inspired by speech recognition methodology. I have shown several examples in which the increased predictive power has discovered domains which further understanding of human disease and biology, and expect there will be many others. From a theoretical point of view, this method provides an integrated prediction of domain annotation for a given protein, evaluating in a strictly probabilistic fashion the appropriate trade-off between amino-acid signal strength and contextual information. Lastly, from a pragmatic perspective, the method significantly increases sequence coverage. The predictions of the method are available via the Pfam web-pages.

Further improvements to the language models are possible, motivated by similar techniques in speech recognition. Modifications to the decision trees used to classify domain contexts are possible, for example I could classify domain contexts on the basis of the longest potentially non-contiguous preceding subsequence which is also observed in the training database. Alternatively, standard classification techniques to learn optimal decision trees

can be employed. Other annotated regions on the protein could be used in our search: for example regions of low complexity and transmembrane regions. Explicitly modelling the length distribution of spacers between domains may also increase sensitivity. Lastly, alternative classes of generative grammars may be used – although it remains unclear which level is appropriate for domain modelling. The language modelling could also be adapted to take into account nested domains.

An alternative approach to language modelling, such as the exponential model introduced in section 2.1 might provide more flexibility for modelling long-range domain interactions as well as providing an alternative method for integrating taxonomic information. This method is more computationally expensive but also more flexible with regard to modelling arbitrary features.

Extra information other than taxonomy regarding the protein may also prove a useful guide in domain annotation. For example the techniques used to incorporate taxonomic information can also be used to incorporate protein localisation or even functional information such as phenotype in a systematic RNAi screen.

This type of approach may also be applicable to the discovery of cis-regulatory modules (CRMs) and transcription factor (TF) binding sites. Identification of TF binding sites using weight matrices is difficult, as they can lie kilobases away from the transcription start-site, and the motifs occur at random throughout the genome. Several authors have built organizational models which take motif positioning and orientation into account [DSW01, PFL$^+$01], while others have attempted to identify functional motifs on the basis of high local density of potential binding sites [BNP$^+$02]. Language modelling is related to some of these methods, and may provide an alternative strategy.

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

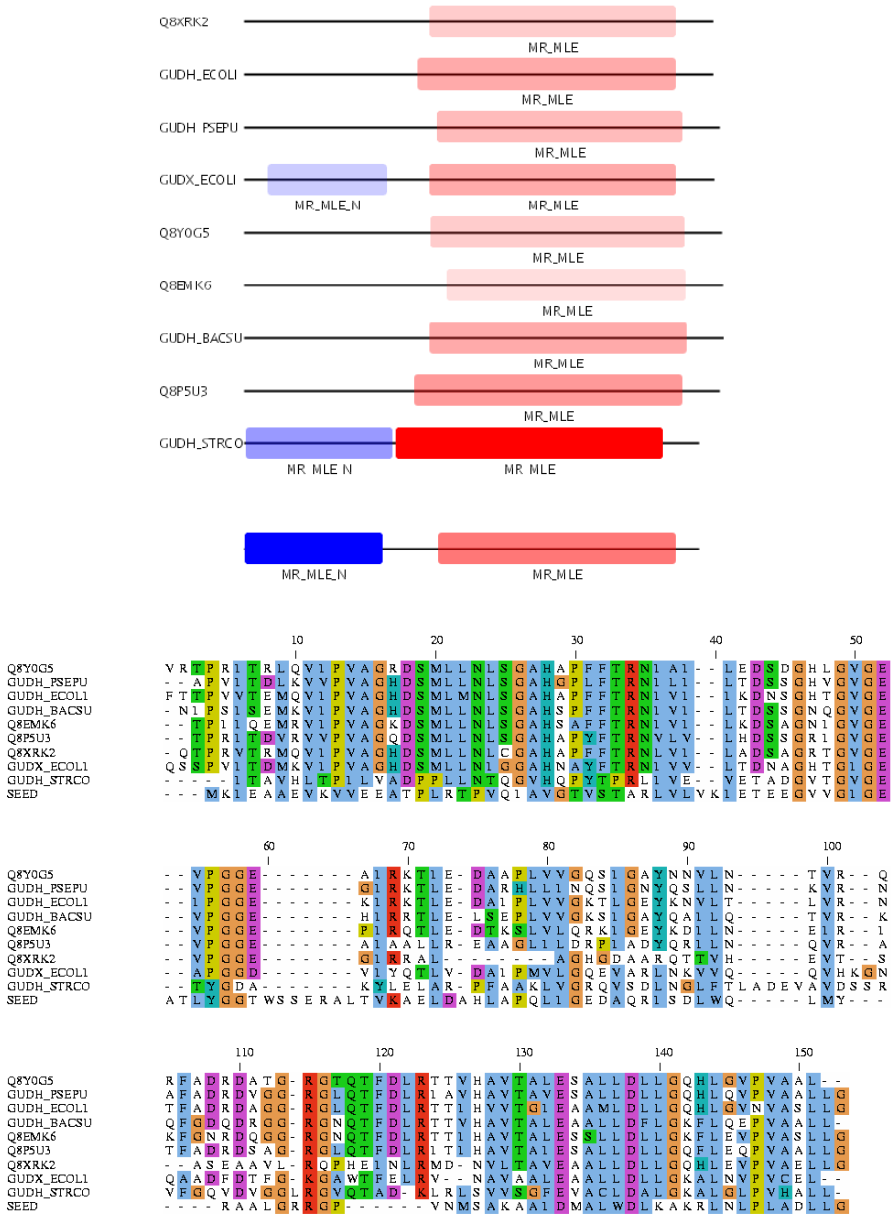

Figure 3.1: Top: Pfam full-model hits to homologues of GUDH_ECOLI. The opacity of the hit is proportional to the strength of the hit (log-odds score minus threshold) relative to the best scoring hit in this cluster. GUDH_STRCO has the strongest hits for both MR_MLE_N and MR_MLE domains. MR_MLE_N is detected 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.

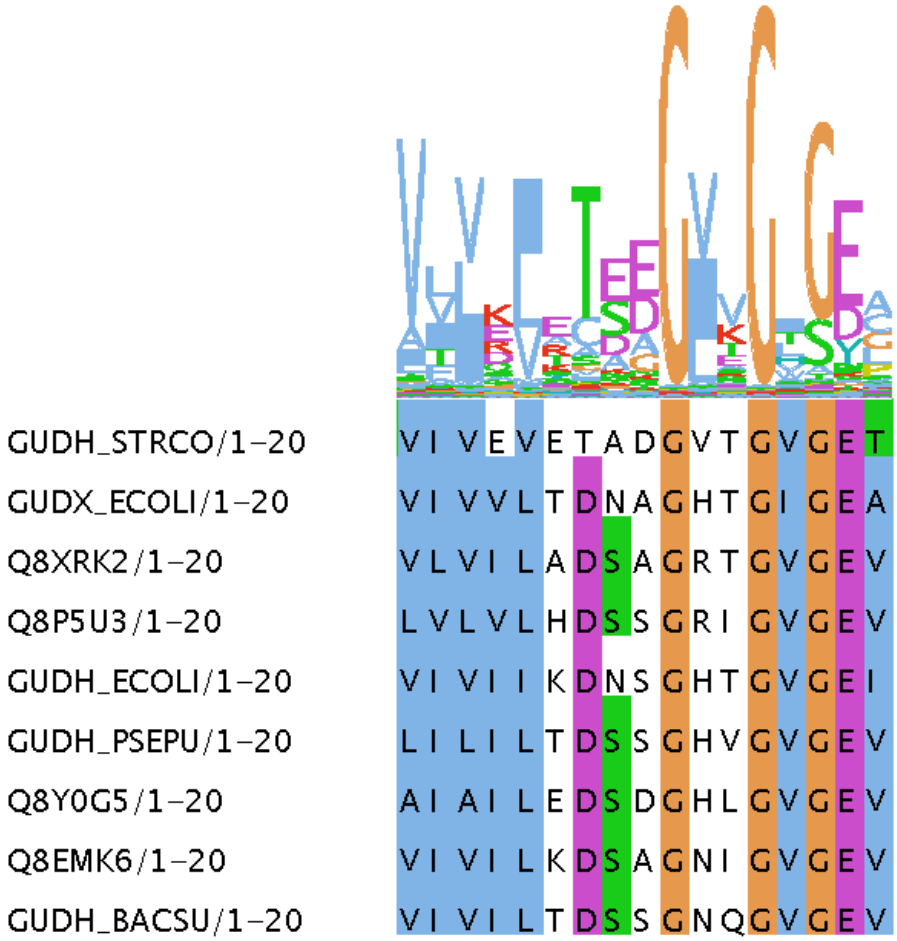

Figure 3.2: A fragment of the alignment of the GUDH_ECOLI alignment (bottom) shown together with aligned emission states from the profile HMM(top). Match states 34 to 50 are shown together with the corresponding columns from the alignment (which may not be contiguous). The total height of a column in 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,

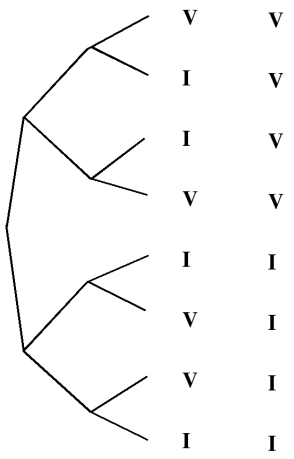

Figure 3.3: Illustration of the effect of different rates of evolution and/or different gwF modes of evolution at sites with similar functional constraints. Two site are shown for the same tree, the first site appears to be evolving much slower than the second site, however the stationary probability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

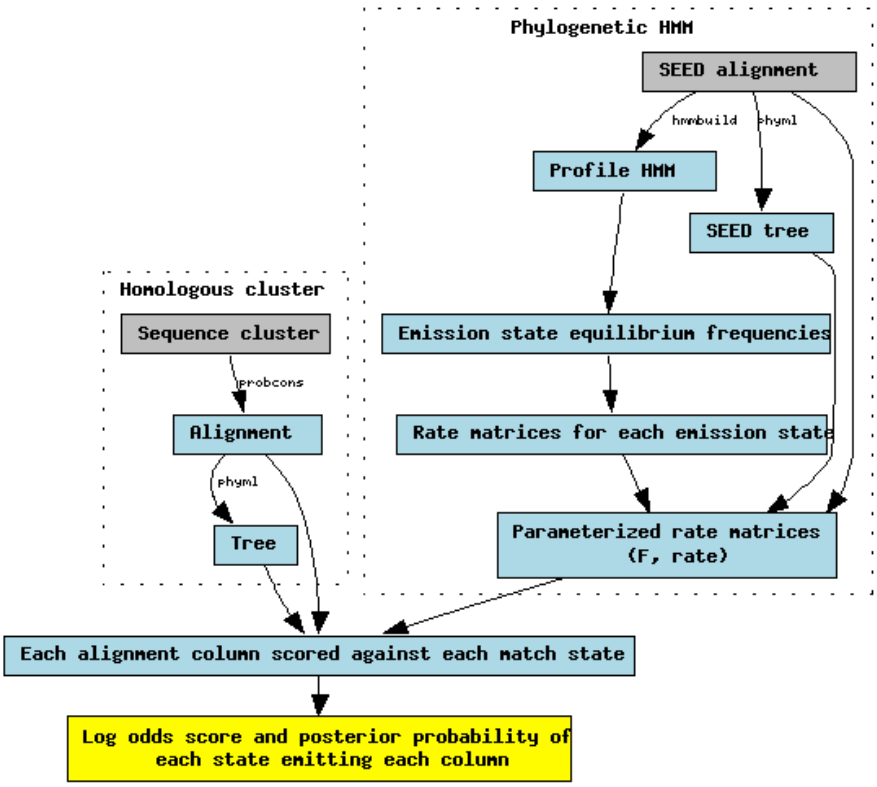

Figure 3.4: Diagram of processes involved in the phylogenetic profile HMM. Inputs are shown in grey, and outputs in yellow, with intermediate steps in light-blue.

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.

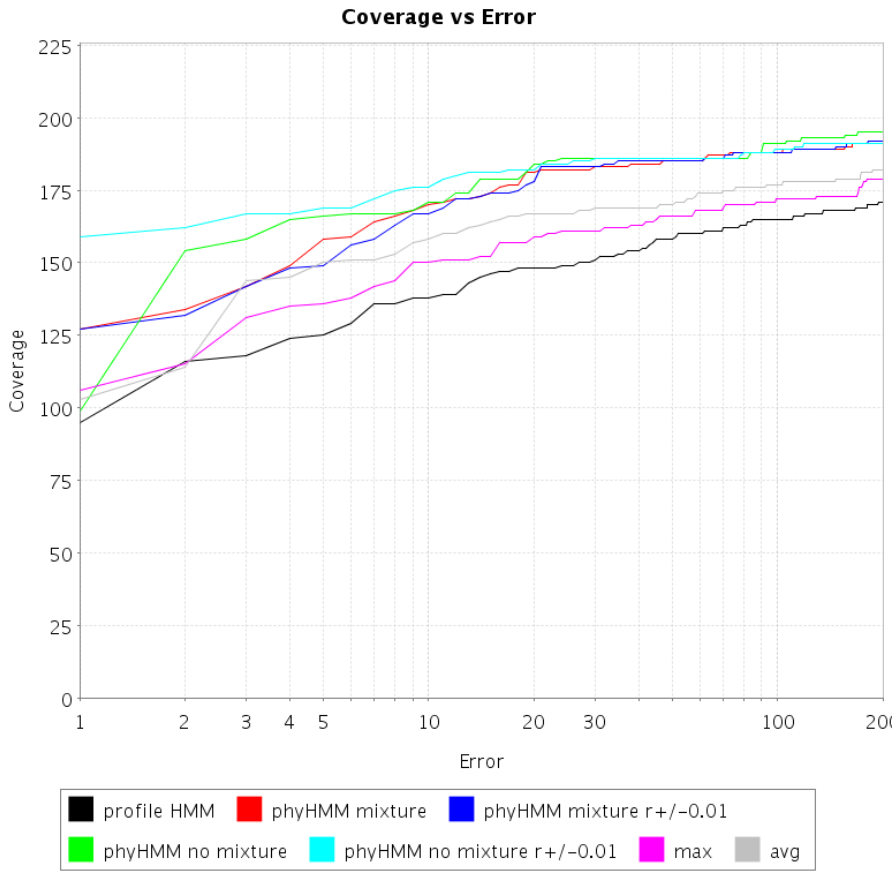

Figure 3.5: Coverage vs error curve for phylogenetic HMM vs standard profile HMM on ASTRAL test set and 44 Pfam families. The coverage at error rate of n is defined as the number of homologous sequences before the n^{th} false positive (or non-homologous sequence). The black line is the standard profile 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.

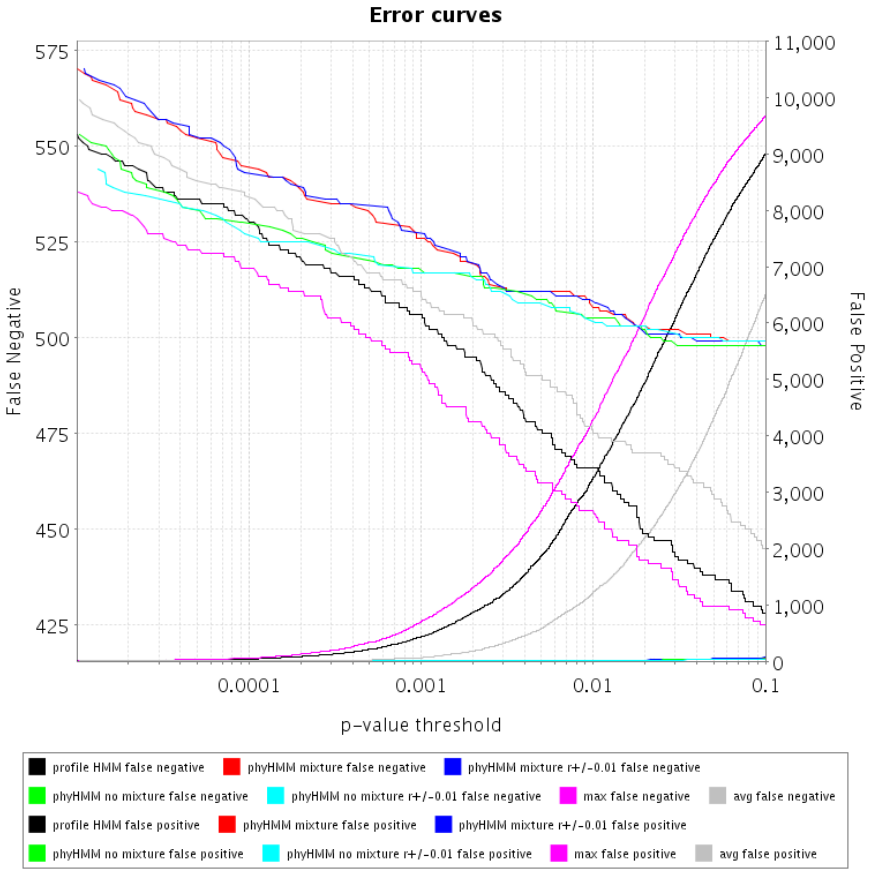

Figure 3.6: Error rate vs p-value score threshold according to HMMER extreme value distribution. False negatives are shown on the left x-axis (which fall from left to right as the p-value threshold increases). False positives are shown on the right x-axis (which rise as the p-value threshold increases). 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.

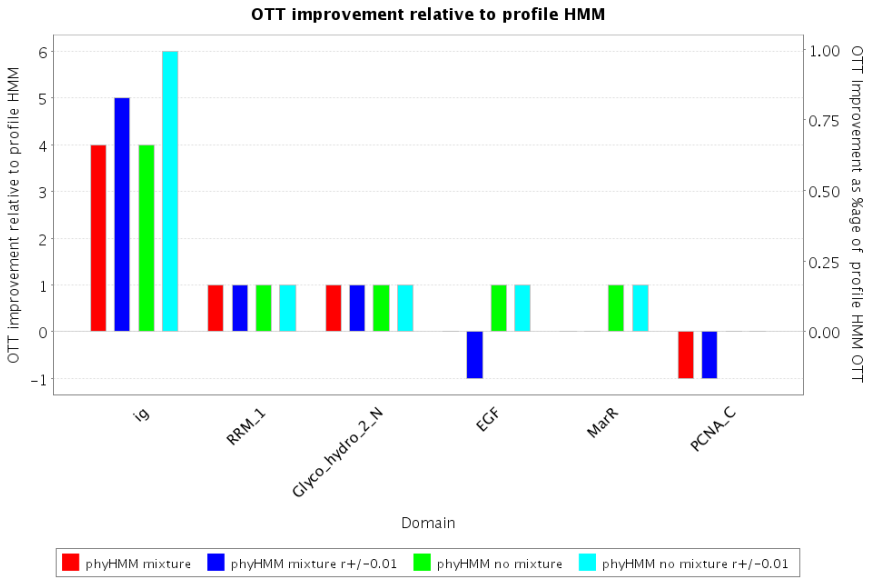

Figure 3.7: Improvement in OTT score relative to the standard profile HMM on a family basis. The red/blue bars are the scores for the phylogenetic HMM with mixture null model and without/with rate variation respectively. The green and cyan bars are for models without a mixture model background, and without/with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

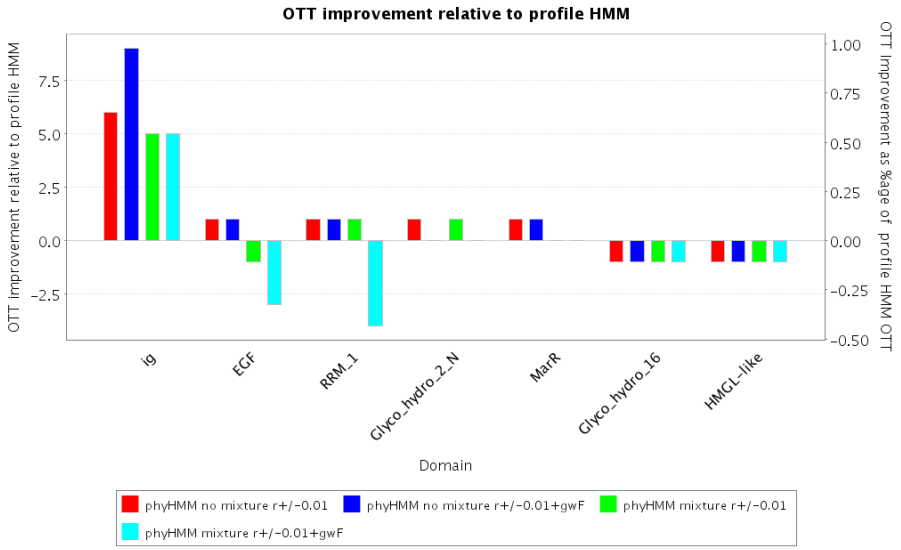

Figure 3.8: Impact of gwF variation on detecting homology for six families. +gwF variation degrades performance for the mixture model, and in one case improves performance for the non-mixture model.

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

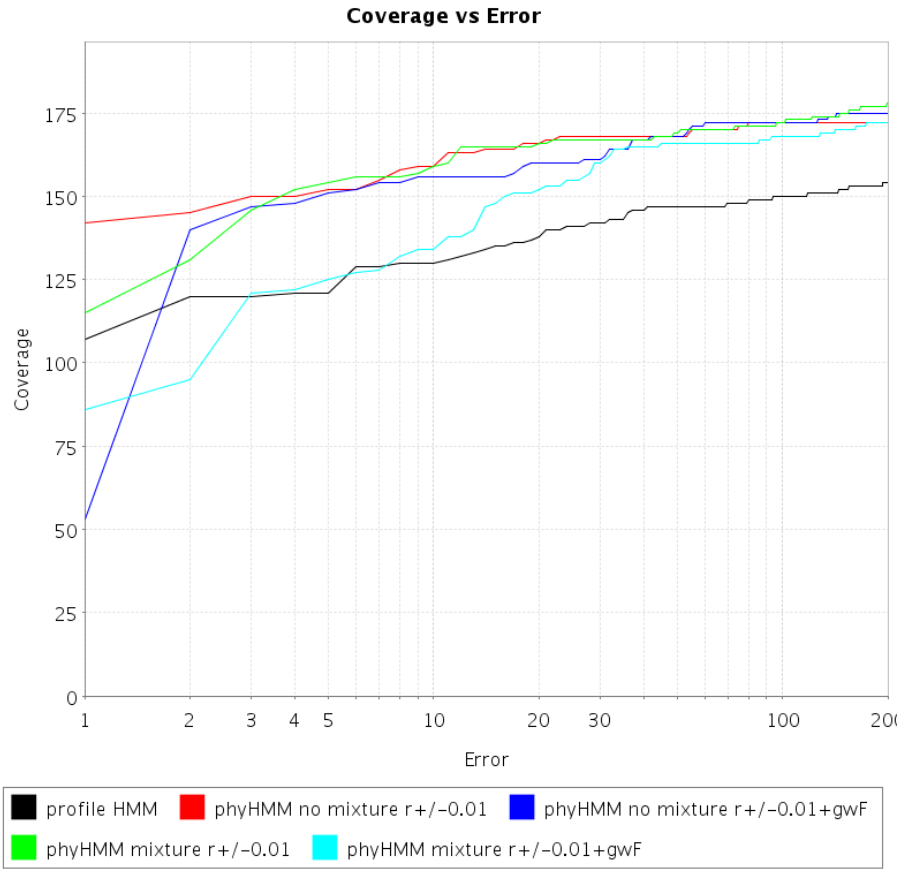

Figure 3.9: Coverage vs error curve for models which include +gwF variation (light and dark blue) vs models which do not include +gwF variation (red and green), scored on 30 Pfam families. The curve for a standard profile HMM is included in black for reference. Both models which have a mixture null model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

Chapter 4

Using protein domains to identify pseudogenes and positive selection

The detection of pseudogenes and genes under positive selection are both important challenges for bioinformatics.

From the point of view of functionally annotating eukaryotic genomes it is crucial to separate protein coding genes from genes which are not translated to give functional proteins. Moreover, while not translated, transcribed pseudogenes are increasingly thought to play an important regulatory role [HYC+03]. While experimental techniques for detecting gene transcripts are well developed and amenable to high throughput analysis (including EST libraries, RT-PCR, Northern blots, microarray analysis), this is not yet the case for detecting protein products. Standard techniques (such as a Western blot) require an antibody for the protein to be available, which in turn requires an expressed protein, or a synthetic peptide. Furthermore, these techniques would have to distinguish the protein from any close homologues that may not be pseudogenes. Thus, bioinformatics has an important role to play in identifying likely pseudogenes.

The identification of genes under positive selection is an important tool for understanding the evolutionary pressures acting on various organisms. Moreover, identifying sites under selection can help pinpoint the molecular basis for adaption in processes such as drug resistance, immune defense, speciation, brain size, etc. This also leads to biologically testable hypotheses regarding the functional importance of particular mutations.

Compositional methods for the identification of pseudogenes are often related to meth-

ods for the detection of positive selection. This is certainly true for methods which estimate the ratio of the rates of non-synonymous (dN) to synonymous substitution (dS). In this case the factor distinguishing pseudogene evolution from positive selection is a dN/dS ratio of around 1 across the length of the gene, rather than several sites of the gene with $dN/dS > 1$.

In this chapter I introduce a new compositional method for the detection of pseudogenes and positive selection, using the techniques developed in chapter 3. The motivation for this method is that not all non-synonymous substitutions are equally detrimental – or transformational – to the function of the protein. With knowledge of the functional importance of a site as well as the degree to which a site is conserved in related functional proteins, it should be possible to weight amino-acid changing mutations based on how likely they are to change the structure and function of the protein. Thus, mutations in sites which are highly conserved and structurally/ functionally important contribute greater evidence to either positive selection or pseudogene evolution than do amino-acid changing mutations in a poorly conserved site.

I will first demonstrate that this method is a better predictor of pseudogene status than current techniques, to the extent that strong assertions about the pseudogene status of particular genes can now be made, rather than weaker assertions about sets of genes which are enriched for pseudogenes. I then investigate the application of the technique to the identification of positive selection, and discover positive selection in proteins implicated in the immune response to HIV infection as well as in the HIV protein which counteracts this response. I re-analyse the abalone sperm lysin set in which positive selection has been previously identified, and show that despite significant non-synonymous mutation, the mutations are mostly consistent with maintaining the protein domain, and thus unlikely to result in major conformational changes. Finally, I carry out a large scale scan for positive selection in 11 genomes, and identify Pfam domains which are over-represented in positively selected genes. The results are compared between species.

The algorithm and program developed in this chapter is called PSILC, which is a double acronym: {Pseudogene / Positive Selection} Inference from Loss of Constraint. The method presented here extends the algorithm first introduced in [CD04], which was only concerned with pseudogene annotation. The extensions presented in this chapter allow the method to differentially detect positively selected genes from pseudogenes, which has the effect of improving pseudogene classification as well as providing site and lineage specific predictions

of positive selection.

4.1 Pseudogenes

Pseudogenes have been defined as sequences of genomic DNA which are originally derived from functional genes but are no longer translated into functional protein products. Pseudogenes are thought to have arisen by two distinct processes. Unprocessed pseudogenes are believed to have arisen from genome duplication, with a subsequent loss of function of one copy due to the accumulation of disabling mutations in the coding or regulatory sequence. Processed pseudogenes lack introns, and are thought to have arisen by reverse transcription of processed mRNA, followed by integration back into the genome. There is an increasing number of examples where pseudogenes play an important biological role, particularly in eukaryotic genomes [BA03]. It had been assumed that pseudogenes will rapidly degenerate and become indistinguishable from surrounding genomic sequence, due to non-functionality. Although this process has been observed in prokaryotic genomes [AA01], eukaryotic genomes contain many pseudogenes which have avoided full degeneration, and there appears to be less pressure to delete pseudogenes in eukaryotes than prokaryotes [Mig00, HG02]. A regulatory role for a human pseudogene has been observed experimentally [HYC$^+$03]. Moreover it has been calculated that $2-3\%$ of all human processed pseudogenes are expressed, and that $0.5-1\%$ of mouse processed pseudogenes are expressed [Yano04].

Pseudogenes are often mis-annotated as functional genes in sequence databases [Mou02]. Two recent surveys [TSZB03, HHB$^+$02] both estimate ≈ 20000 human pseudogenes. Sequence based methods for identifying pseudogenes include methods which rely on the presence of truncations by mutation to stop codon or frame-shift, and compositional methods which are based on estimating the ratio of the rates of substitution at synonymous sites to the rate of substitution at non-synonymous sites (dN/dS). Torrents *et al.* [TSZB03] concluded that half of human pseudogenes have no detectable frame-shifts or internal stop codons, and hence compositional methods are required to identify pseudogenes. The dN/dS methods are based on the assumption that amino acid changes in a protein coding gene are in general detrimental to its function, and hence less common, whereas a pseudogene has no functional constraints, and hence the ratio of the rates of synonymous and non-synonymous mutation should be equal. There are many ways to estimate the rates of synonymous and non-synonymous substitution

(see [BEW03] for a review). In this chapter, I test the method in [GY94] as well as the method of [NG86] as calculated by PAML. The method in [GY94] was used in the survey from [TSZB03].

The method of Goldman and Yang [GY94] uses the model of codon evolution described in equation 1.26. I use the free dN/dS ratios for branches model, in which each branch in the tree is allowed to have a different dN/dS ratio, and the branch dN/dS ratios which maximise the likelihood under equation 1.26 are reported.

4.2 Positive selection

Natural selection can be defined as the process by which the relative frequencies of alleles in a population change to reflect their relative fitness. The action of natural selection can be verified, for example, by mutation fluctuation experiments as developed by Luria and Delbrück [LD43], in which a bacteriophage introduced into bacterial culture induces phage-resistant colonies. Luria and Delbruck demonstrated that this was due to random mutations conferring resistant genotypes. Natural selection is thought to act on new alleles generated by mutations in one of three ways. If the mutation decreases fitness it will be removed from the population, which is called purifying selection. Positive selection occurs when the mutation enhances fitness and so the frequency of the allele increases in subsequent generations. This results in a selective sweep as regions linked to the advantageous mutation also increase in frequency which also reduces variation in linked regions. If the mutation is selectively neutral it will persist in subsequent generations at some low alleleic frequency, possibly disappearing from the population at some stage due to random drift or a selective sweep at a linked site. Kimura [Kim83] proposes that most polymorphisms are selectively neutral. However, there are many examples of positive selection acting at the amino acid level.

Tests for positive selection can be loosely divided into those which are based on alleleic variance within a population, and those which are based on comparisons of homologous sequence between different species. These techniques have been used to detect selection in a wide variety of gene families, for example [HN88, LOV95, SV95, YSV00, YNGP00, SEM04].

One of the most popular and direct ways for detecting positive selection in protein coding genes is to identify an excess of non-synonymous to synonymous substitutions. There have been many methods proposed for using dN/dS to detect selection which can be split into

methods which use parsimony to reconstruct ancestral sequences (e.g. [SG99]) and methods which estimate dN/dS as a parameter in a probabilistic model using maximum likelihood (e.g. [NY98]). In the method of [NY98] different probabilistic models are created, each based on the formulation in equation 1.26. One such model is a mixture model of three different site categories, with invariable sites ($\omega = 0$), neutral sites ($\omega = 1$) and positively selected sites ($\omega > 1$). The mixture co-efficients and the value of ω for positively selected sites are those which maximise the likelihood. The maximum likelihood of this model is compared to the maximum likelihood of the constrained model in which the frequency of positively selected sites is set to zero under a likelihood ratio test. If the test result is significant and $\omega > 1$ for positively selected sites then selection is inferred. This method has been extended in [YSV00] to accommodate more realistic models of variation of ω amongst sites. These methods have been shown to be accurate and powerful methods for detecting positive selection [WYGN04]. In [YN02] the branch-site model was developed for detecting positive selection at individual sites along a specific lineage. It has been suggested that the branch-site model detects false-positives in some evolutionary scenarios [Zha04].

Guindon et al. recently extended the maximum likelihood framework for detecting selection by allowing the model to switch between different ω categories at some rate, and calculating the expected fraction of time the selection process spends in a particular category to infer positive selection. Tests for using evolutionary rate shifts in order to detect positive selection have also been proposed [KM01, Gu01, GMB01]. These tests are based on the observation that subsequent to duplication, a rate change often occurs in residues of the protein responsible for its new function.

4.3 Algorithm

The PSILC algorithm uses the protein domain match state specific rate matrices defined in section 3.1.1, which will be referred to collectively as a *domain model* of evolution. Recall that this collection of rate-matrices defines a different model of evolution at each site in the alignment which matches a match state of the profile HMM. In chapter 3 these evolutionary models were used to test whether the domain model of evolution was more likely to have generated the alignment than a null protein model of evolution. In this chapter, however, the starting assumption is that the domain model has generated the alignment, and I test whether

evolution below a particular node in the tree is better explained by either a null protein or null DNA model of evolution. Thus, for a given node, the domain model of evolution now takes the role of background model, and a composite evolutionary model consisting of a null protein or null DNA model below the node under consideration and a domain model on all other branches, is tested against this new background model. This can be thought of as inverting the log-odds ratio in equation 3.1 used in chapter 3 below the given node.

If a pseudogene is present in the tree T, then evolution along the final branch to this gene is expected to be explained better by the composite domain/null-DNA model of evolution than the background domain model, and so the composite model should provide a higher likelihood. This is the basis for the pseudogene score. If, on the other hand, a single site in a gene is positively selected, then the site-specific likelihood under the composite domain/null-protein model should be higher than under the background domain model. This forms the basis for the positive selection score.

Figure 4.1 provides an overview of the PSILC algorithm. The two inputs to PSILC consist of a homologous cluster of in-frame protein coding nucleotide sequences without internal stop codons (top right hand side), and a collection of profile HMMs D_l matching sequences in the homologous cluster (top left-hand side). An alignment and tree are built for the homologous cluster. Each of the profile HMMs D_l is aligned to the alignment via the forward-backward algorithm. A rate matrix is built for each match state, and a null DNA and null protein rate matrix are constructed. Via the alignment of the HMMs to the protein alignment, site-specific likelihoods under the background domain evolutionary model (the domain/domain likelihood) as well as under the composite domain/null-DNA and domain/null-protein models are calculated. These are summed to give an overall log-likelihood for each of the three evolutionary models from which the PSILC-prot/dom and PSILC-nuc/dom log-odds ratio are calculated by subtracting the domain/domain log-likelihood from the domain/null-protein and the domain/null-DNA log-likelihoods respectively. Thus a high PSILC-nuc/dom score reflects a better fit to the alignment of the composite domain/null-DNA model than the domain model, and so this is taken to be the principal pseudogene score. The site-specific likelihoods are also integrated via a three state *selection HMM* to obtain site-specific posterior probabilities of positive selection. Each of these steps is described in more detail below.

There are two important differences with respect to chapter 3. The first is that all

of the evolutionary models score codon alignments rather than protein alignments. This is necessary so that the likelihood can be calculated in a consistent manner over both DNA (as required by the domain /null-DNA models) as well as protein sequences (as required by the domain/null-protein and background domain models). The second is that each site in the alignment will be assumed to be evolving under a mixture of each of the profile HMM emission state evolutionary models according to the posterior probability of each state emitting this site. Thus, the model marginalises over the alignment of the profile HMM to the alignment according to this posterior probability.

Building the alignment and tree

PSILC translates the DNA sequences into protein sequences, which are then aligned using either PROBCONS [DMBB] or MUSCLE [Edg04], and back-translated (referencing the original DNA sequences) into a codon alignment, $A = \{x_{k,i}\}$. PSILC also produces a tree T from the protein alignment using either Phyml [GG03], or neighbour joining with maximum likelihood distances. In both cases an amino acid rate matrix (such as WAG[WG01]) is used. An amino-acid rate matrix, rather than nucleotide rate matrix is used to estimate distances as the background assumption is that the cluster is evolving as protein. More accurately, the background assumption is that the cluster is evolving according to the site specific rate matrices specified in the protein domain model of evolution, and so a more consistent approach is to calculate distances based on the protein domain model. This may make some difference to the branch length estimates [HB98, LP04], but this is not investigated here. PSILC also accepts user defined trees.

Aligning the Profile HMM to the protein alignment

For each sequence $x_{k,\cdot}$ in the protein cluster, and each profile HMM D_l, PSILC calculates the log-odds score (relative to a null model given in the HMMER HMM) of the model matching the sequence, using the forward algorithm described in chapter 1. From the log-odds score, and using the parameters for the extreme value distribution given in the HMMER model, PSILC calculates an empirical p-value. If this p-value is greater than a user-specified threshold (or the default value of 1e-5), for all sequences in the cluster, the model is not further considered in the PSILC calculation. PSILC also calculates the posterior probability $P(\psi_i = M_{l,j}|x_{k,\cdot})$

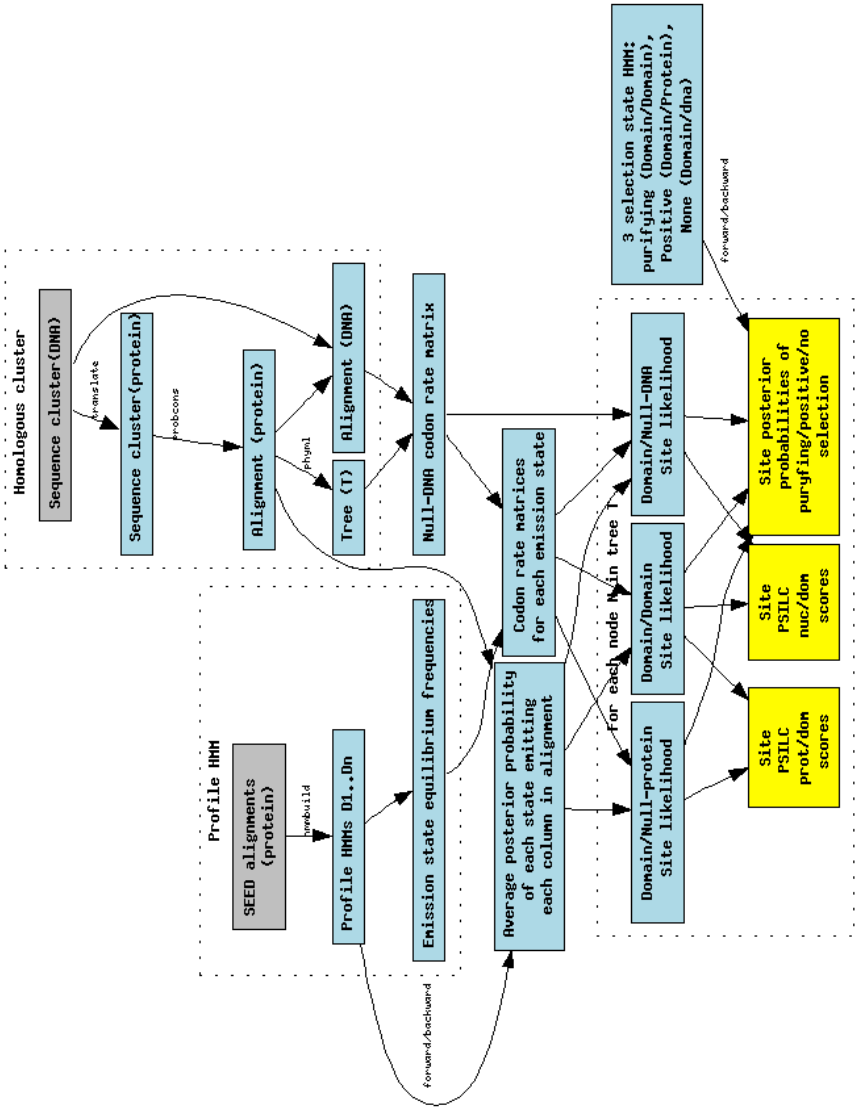

Figure 4.1: Conceptual diagram PSILC. Inputs are shown in grey, and outputs in yellow, with intermediate steps in light-blue.

that the match state $M_{l,j}$ emitted residue $x_{k,i}$ in the sequence x_k. By averaging across all sequences which matched the model D_l below the p-value threshold, PSILC calculates the posterior probability $P(\psi_i = M_{l,j}|A)$ of each match state emitting each column in the alignment. Although this procedure will guarantee that $\sum_j P(\psi_i = M_{l,j}|A) <= 1$ for each HMM D_l, it cannot guarantee that

$$S_i = \sum_{l,j} P(\psi_i = M_{l,j}|A) <= 1, \tag{4.1}$$

which is required below. This may happen if two profile HMMs are included which are closely related, for instance two HMMs from the same SCOP superfamily or Pfam clan. Hence each posterior probability is divided by S_i if $S_i > 1$.

PSILC is robust to the inclusion of profile HMMs which do not match sequences in the cluster for example models which have a low e-value score but are false matches, as these will be removed in the previous step. PSILC is also robust to the inclusion of models which partially match the protein cluster (i.e. they match in HMMER's 'fs' mode), as PSILC considers a match state at a position in proportion to its posterior probability of emitting this state. PSILC will also run if no profile HMMs are provided, in which case it will effectively compare a null protein model to a null DNA model. The profile HMMs can be downloaded directly from a profile HMM database, such as Pfam, or can be built directly from a seed alignments using *hmmbuild* from the HMMER package. The Profile HMMs should be first calibrated using the *hmmcalibrate* program from HMMER, so that PSILC can calculate empirical e-value significance scores.

Models of substitution

All PSILC likelihoods are calculated on the basis of codon rate matrices and codon alignments. Hence, it is necessary to devise models of codon substitution which reflect

(i) the null DNA model of evolution, labelled $\mathcal{E}_{nuc}$;

(ii) the null protein model of evolution, labelled $\mathcal{E}_{prot}$;

(iii) the match state specific models of evolution, labelled $\mathcal{E}_{M_j}$.

The null-DNA codon rate matrix uses one of the HKY[HKY85], TN[TN93], F81/F84 [Fel81], GTR [LPSS84] nucleotide models (as specified by the user), and the observed nu-

cleotide frequencies in the alignment A as the steady-state probabilities. The parameters in each of the nucleotide models are trained using the tree T and the alignment A. The null-DNA codon rate matrix is not calculated directly. Instead, PSILC calculates the codon transition probability as

$$P_{\mathcal{E}_{\text{nuc}}}(x^{t+\triangle t} = u_1u_2u_3|x^t = v_1v_2v_3) = \prod_{i=1,2,3} P_{\mathcal{E}_{\text{nuc}}}(x_i^{t+\triangle t} = u_i|x_i^t = v_i) \tag{4.2}$$

assuming independence between codon sites.

The null protein codon rate matrix is calculated using one of WAG[WG01], WAG+gwF[GW02], JTT[JTT92] models with the observed amino acid frequencies in the alignment A as the steady-state probabilities (using eq. 1.27). The f parameter in the WAG+gwF model is trained using the tree T and the alignment A. Codon transition probabilities are calculated as:

$$P_{\mathcal{E}_{\text{prot}}}(x^{t+\triangle t} = u|x^t = v) =$$
$$\begin{cases} \text{n.a if } v \text{ is a stop codon} \\ 0 \text{ if } u \text{ is a stop codon} \\ P_{\mathcal{E}_{\text{prot}}}(a(x^{t+\triangle t}) = a(u)|a(x^t) = a(v)) * \frac{P_{\mathcal{E}_{\text{nuc}}}(x^{t+\triangle t}=u|x^t=v)}{\sum_{w:a(w)=a(u)} P_{\mathcal{E}_{\text{nuc}}}(x^{t+\triangle t}=w|x^t=v)} \text{ otherwise} \end{cases} \tag{4.3}$$

where $a(x)$ is the amino acid translation of x. This equation splits the transition probability from amino acid $a(v)$ to $a(u)$ amongst all possible codons corresponding to $a(u)$ according to the relative probability of transitioning (at a DNA level) to each of these possible codons.

The match state protein rate matrices are calculated as described in section 3.1.1. Rate variation between match states was not modelled, and the f parameter of the WAG+gwF model, if used, is set to the same value as for the null protein rate matrix. These are converted into codon models using the technique described in the previous paragraph.

Site specific likelihood scores

For a given leaf node n in the tree T, let T_n denote the branch to node n, and $T \setminus T_n$ denote all other branches on the tree. The following evolutionary hypothesis are considered:

(i) $\mathcal{E}_{\text{nuc,dom}}$: neutral DNA evolution along T_n, domain constrained evolution on $T \setminus T_n$ (pseudogene evolution);

(ii) $\mathcal{E}_{\text{prot,dom}}$: protein constrained evolution along T_n, domain constrained evolution on $T \backslash T_n$ (evolution under positive selection);

(iii) $\mathcal{E}_{\text{dom,dom}}$: domain constrained evolution on all T, including T_n (purifying selection).

The likelihood of each site $x_{.,i}$ is calculated under each of the evolutionary hypotheses, weighting the contribution of each HMM match state according to the posterior probability of being in the match state at the alignment position, and also including the contribution of the insert states of the profile HMM with weight $1 - S_i$ where S_i is given by equation 4.1.

$$P(x_{.,i}|T, \mathcal{E}_{\text{nuc,dom}}) = \sum_{j,l} P(x_{.,i}|\mathcal{E}_{\text{nuc},M_{l,j}}, T) * P(\psi_i = M_{l,j}|x) + (1 - S_i) * P(x_{.,i}|\mathcal{E}_{\text{nuc,prot}}) \tag{4.4}$$

$$P(x_{.,i}|T, \mathcal{E}_{\text{prot,dom}}) = \sum_{j,l} P(x_{.,i}|\mathcal{E}_{\text{prot},M_{l,j}}, T) * P(\psi_i = M_{l,j}|x) + (1 - S_i) * P(x_{.,i}|\mathcal{E}_{\text{prot,prot}}) \tag{4.5}$$

$$P(x_{.,i}|T, \mathcal{E}_{\text{dom,dom}}) = \sum_{j,l} P(x_{.,i}|\mathcal{E}_{M_{l,j},M_{l,j}}, T) * P(\psi_i = M_{l,j}|x) + (1 - S_i) * P(x_{.,i}|\mathcal{E}_{\text{prot,prot}}) \tag{4.6}$$

The calculation of the likelihoods $P(x_{.,i}|T, \mathcal{E}_*)$ can be carried out according to the Felsenstein algorithm [Fel81], as described in section 1.3.3. Note that the term $P(x_{.,i}|\mathcal{E}_{M_{l,j},M_{l,j}}, T)$ is just the emission state probability under the match state $M_{l,j}$ used in section 3.1.1, which is written there as $P(x_{.,i}|\psi_i = M_{l,j}, T)$. The notation has been modified here to emphasise the evolutionary models used on each branch in the tree.

Integrating site specific scores

At this point, PSILC proceeds in two distinct ways in order to integrate site specific likelihoods into an overall PSILC score. One is to assume that a single evolutionary hypothesis applies

to all sites in the alignment, and calculate the log-odds ratios

$$\begin{aligned}\text{PSILC-nuc/dom} &= \log \frac{P(A|\mathcal{E}_{\text{nuc,dom}}, T)}{P(A|\mathcal{E}_{\text{dom,dom}}, T)} \\ &= \sum_i \log \frac{P(x_{.,i}|T, \mathcal{E}_{\text{nuc,dom}})}{P(x_{.,i}|T, \mathcal{E}_{\text{dom,dom}})},\end{aligned} \tag{4.7}$$

$$\begin{aligned}\text{PSILC-prot/dom} &= \log \frac{P(\mathcal{E}_{\text{prot,dom}}|A, T)}{P(\mathcal{E}_{\text{dom,dom}}|A, T)} \\ &= \sum_i \log \frac{P(x_{.,i}|T, \mathcal{E}_{\text{prot,dom}})}{P(x_{.,i}|T, \mathcal{E}_{\text{dom,dom}})},\end{aligned} \tag{4.8}$$

assuming that the sites of the alignment are conditionally independent given the tree T and each of the evolutionary hypotheses. These scores are both pseudogene scores as pseudogenes have lost both the domain-encoding and protein-encoding constraint. These scores may be misleading for positively selected genes, particularly if a strongly conserved site is mutated (which would give rise to a strong PSILC score for a single site that might not be outweighed by the domain constrained evolution along the remainder), or if many conserved sites are mutated.

An alternative approach is to regard the evolutionary hypotheses as hidden states of a hidden Markov model (which I shall call a *selection HMM*), and to use posterior decoding (outlined in the introduction) to calculate the posterior probability of being in each state at each site in the alignment. The hidden Markov model used is shown in figure 4.2. The emission probabilities for each evolutionary state and each site are given by eqs. 4.4-4.6. PSILC uses the forward-backward algorithm to calculate the posterior probabilities of being in each of the evolutionary states at each site. Sites with gaps (or unknown characters) at all positions below the target node are non-informative (the emission probabilities are all equal) and so are removed from this calculation. In this way, for example, the selection HMM does not have to 'pay' the higher transition cost for staying in a positive selection state without accumulating log-odds score.

The transition probabilities of the selection HMM can be configured in different ways based on prior knowledge of a particular gene family, and on the particular test. The configuration used to test for pseudogenes is shown in 4.2. In this configuration, a path through the HMM must be either exclusively in the pseudogene state, or not in the pseudogene state at all. Hence the posterior probability of being in a pseudogene state is uniform across the length of the gene, and this probability can be used as a metric of pseudogene status (I

will call this the 'PSILC posterior nuc' score). In this configuration, the maximum posterior probability of being in a selection state can be used as a metric for selection (I will call this 'max PSILC posterior prot' score). Another possibility is to allow a small transition probability (e.g. 1e-5) from purifying to pseudogene models and a small transition back from pseudogene to purifying, and use the average posterior probability as a pseudogene metric. A third alternative would be applicable once pseudogene status had been ruled out, and the user wished to account for any nucleotide favoured evolution as positive selection. In this case the model could be reconfigured such that selection and pseudogene states are treated equally in the Markov model: the purifying state can transition to the pseudogene state with the same probability as to the selection state, and the pseudogene state can transition back to purifying with the same probability as the selection state. The maximum of the posterior probabilities of selection and pseudogene can then be used as a metric for selection.

Note that for sites $x_{.,i}$ in the alignment which do not match any of the profile HMMs, the contribution to the likelihood (eqs. 4.4-4.6) made by the match states will be small (provided the posterior probability of these match states matching the site is small). In this case, the score under $\mathcal{E}_{\text{dom,dom}}$ and under $\mathcal{E}_{\text{prot,dom}}$ both reduce to that under $\mathcal{E}_{\text{prot,prot}}$, and the score under $\mathcal{E}_{\text{nuc,dom}}$ reduces to that under $\mathcal{E}_{\text{nuc,prot}}$. Hence, outside the region matched by the profile HMMs the contribution to PSILC-prot/dom is 0, and the contribution to PSILC-nuc/dom is determined by comparing a nucleotide model along the final branch to a protein encoding model, which is in general non-zero. Thus, PSILC-nuc/dom captures extra information relative to PSILC-prot/dom outside the protein domain region.

Complexity and optimizing the algorithm

The computational complexity of the algorithm is driven by calculating the likelihoods

$$P(x_{.,i}|\mathcal{E}_{M_{l,j},M_{l,j}}) \tag{4.9}$$

$$P(x_{.,i}|\mathcal{E}_{\text{prot},M_{l,j}}) \tag{4.10}$$

$$P(x_{.,i}|\mathcal{E}_{\text{nuc},M_{l,j}}). \tag{4.11}$$

Equation 4.9 must be calculated for each site and each match state. Eqs 4.10, 4.11 must be calculated for each site, match state and each node on the tree. The likelihood calculation is linear in the number of sequences for a fixed size alphabet. Hence the order of the computation

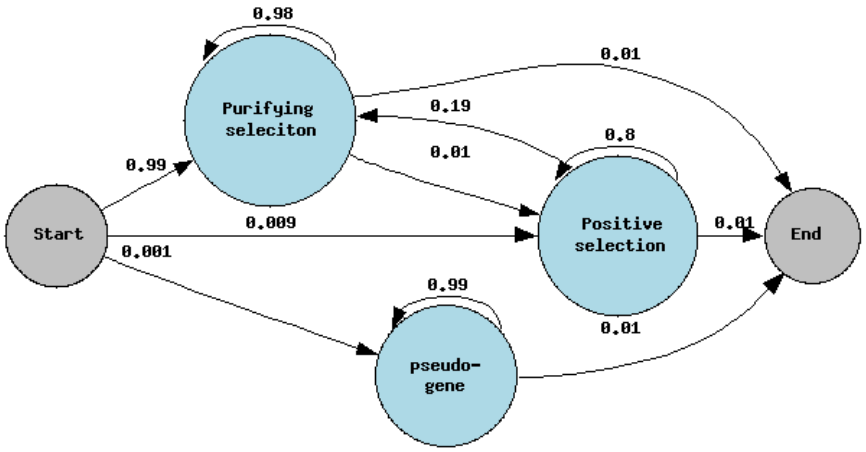

Figure 4.2: Diagram of selection HMM, comprising 3 states - purifying selection ($\mathcal{E}_{\text{dom,dom}}$), pseudogene ($\mathcal{E}_{\text{nuc,dom}}$), positive selection ($\mathcal{E}_{\text{prot,dom}}$). The transitions are given as an example only, and can be specified by the user.

is $O(JIK^2)$, where J is the total number of match states in all HMMs matching the sequence, I is the length of the alignment, and K is the number of sequences in the alignment. This can be improved if these likelihoods are only calculated for sites i with $P(M_{l,j}, x_{.,i}) > 0.01$, and equations 4.4 - 4.6 modified accordingly. Models with low site posterior probability make only a minor contribution to the PSILC site specific scores. As only a few HMM match sites will match a given site with posterior probability greater than 0.01, this reduces the complexity to $O(JK^2)$.

PSILC speeds up the calculation by the order in which the calculations are done. For a fixed HMM l, match state j and node n, PSILC calculates eqs. 4.9, 4.10, 4.11 over all i with posterior match probability $P(M_{l,j}, x_{.,i}) > 0.01$ simultaneously. In this way the matrix exponential for each edge is only calculate once, instead of multiple times (depending on the number of sites with posterior match probability greater than 1). Another observation which provides a speed-up is that the equations 4.9, 4.10, 4.11 only differ in the rate matrix on the branch to the target node, and hence the partial likelihoods from equation 1.30 only change between these calculations for nodes which are ancestral to the target node n. Hence the calculation can be sped up by first calculating eq. 4.9 and in the Felsenstein tree pruning algorithm only recalculating those probabilities which have changed relative to eq. 4.9 in eqs. 4.10. Moreover, when making the calculation for different nodes n and n' (with the same

l and j), the partial likelihoods only change at nodes which are ancestral to either n or n', and hence the same principle of only recalculating partial likelihoods which have changed can be applied. While these optimizations do not reduce the order of the calculation, they do provide a significant practical speed-up.

4.3.1 Allowing for a single frame-shifted nucleotide sequence

The requirement of in-frame nucleotide sequences without internal stop codons can be relaxed for a single sequence in the input cluster. In this case PSILC will pairwise align as DNA sequence (using MUSCLE) this sequence and its closest homologue from the cluster (which is assumed to be in-frame). PSILC removes any columns in this alignment which are gaps in the second sequence, and replaces the original frame-shifted sequence with its aligned version (including inferred gaps). The frame-shifted sequence is now in-frame with respect to its closest homologue. If stop codons still exist in this sequence, each position in the stop codon is replaced with a gap character. Each position which is part of an incomplete codon in this sequence (due to inferred gaps) in this sequence is replaced with a gap character. The alignment of the nucleotide sequences with the modified sequence proceeds as before.

4.3.2 Restricting the size of the input cluster

In the case where only the PSILC score of a single target node is of interest, most nodes in a large tree are of small incremental importance to testing alternative hypotheses of evolution along the final branch to this node. A large tree will slow down the likelihood calculations, and moreover a large number of nodes will slow down the inference of the ML tree using Phyml. PSILC provides a level of control over the number of nodes used in building the tree, and also in calculating the PSILC scores.

The first level of control is in the tree building stage. The user can specify the maximum numbers M_1, M_2 of nodes to include in Phyml tree building and in the PSILC score calculations. If the number of sequences in the input cluster exceeds M_1, PSILC builds a guide neighbour joining tree using maximum likelihood distances (calculated using a WAG rate matrix), which is significantly faster than Phyml tree inference. PSILC passes sequences corresponding to the M_1 nodes closest (according to tree distance) to the target node in the tree to Phyml for maximum likelihood tree inference. If the number of nodes in the Phyml

inferred tree exceeds M_2, PSILC restricts to the subtree of M_2 nodes closest (according to tree distance) to the target node.

4.3.3 Calculating PSILC scores for internal nodes

If the user provides PSILC with a rooted tree, it is possible to calculate PSILC scores for internal nodes of the tree. The restriction to a rooted tree is necessary to ensure that the directionality of evolution is known (otherwise it is not possible to know a priori in which direction is the root, and in which direction are the leaves of the tree). All the above equations can then be applied to the (rooted) tree T, with T_n now interpreted as the subtree below node n together with the branch to node n[1]. The PSILC scores now reflect the log-likelihood ratio that evolution from the parent of the target node through the target node and along the subtree of the target node is evolving as a pseudogene rather than as a domain encoding gene.

4.4 Results: Vega pseudogene test set

4.4.1 Test data

The manual annotation of human chromosome 6 [Mun03] (NCBI34 human genome build), which can be obtained from `http://www.vega.sanger.ac.uk`, was used as the principal test set for the method and is called the Vega set. Vega annotates both functional genes and pseudogenes, and as such is an ideal test set. In general, Vega pseudogenes are categorised on the basis of homology to known genes/proteins with a disrupted ORF due to frame-shifts and/or in-frame stop codons. Vega contains 1887 coding transcripts on chromosome 6 and 633 pseudogenes. Of these, I extracted 1325 coding transcripts and 457 pseudogenes which could be aligned to at least one different ENSEMBL transcript using the protocol described below. Of these, 1105 coding transcripts and 422 pseudogenes matched a Pfam domain, via one or more members of the cluster. Note, however, that PSILC can be applied to clusters

[1]The user can specify one of two PSILC modes - recursive, or non recursive. The discussion here applies to the recursive model, in which the divergent evolutionary hypothesis is applied to the branch to the given node and all branches below the node. The non-recursive mode just applies the divergent hypothesis to the branch leading to the given node. These two approaches are equivalent at leaf nodes. In order to apply the recursive model at inner nodes of the tree, a rooted tree is required.

which do not match Pfam domains, but that the test reverts to distinguishing a protein coding evolutionary constraint from a null DNA model. Pfam release 15.0 was used.

For each (pseudo)gene transcript in the test set a blast search against the ENSEMBL [BAB+04] NCBI34 transcripts for human, rat and mouse was carried out. The query transcript and ENSEMBL transcripts with blast match e-value less than 10^{-7} and a cumulative match length greater than 80% of the query transcript were included in the input cluster of homologous sequences. Transcripts with greater than 99% match on more than 80% of the original sequence were removed from the alignment, to avoid the inclusion of sequences from ENSEMBL which are effectively the same regions in Vega. The procedure in section 4.3.1 is carried out with respect to the Vega (pseudo)gene to ensure that Vega pseudogenes are adjusted to remove frame-shifts and stop codons. Each Pfam family which matched at least one sequence in the cluster was identified (using the ENSEMBL *ensj* API, available from `http://www.ensembl.org/java`), and included in the analysis. As discussed above, the algorithm is robust to the inclusion of Pfam families which are not homologous to sequences in the input cluster. The list of Pfam families and the homologous cluster of nucleotide sequences form the inputs for the PSILC algorithm. A maximum of 10 sequences closest to the sequence of interest were used to build the tree using Phyml [GG03]. These sequences where determined on the basis of an initial neighbour joining tree. A maximum of 6 sequences closest to the sequence of interest where used to calculate the PSILC score (see section 4.3.2), with those closest chosen on the basis of the Phyml derived tree.

The dN/dS score was calculated on the full extent of the alignment. The PAML program 'codeml' was used to calculate dN/dS, using both the method of Nei and Gojobori [NG86], as well as the method of Goldman and Yang [GY94] as implemented in PAML. The method of Nei and Gojobori calculates pairwise dN/dS scores. The Goldman/Yang method incorporates $\omega = \mathrm{dN/dS}$ as a parameter in the rate matrix, and finds the value of ω which maximises the likelihood of the data. For each cluster, a maximum of 3 sequences closest to the sequence of interest (according to the Phyml derived tree) together with the target sequence were extracted from the nucleotide alignment constructed as part of the PSILC algorithm (i.e with any frame-shifts corrected) and provided as input to PAML. The PAML configuration file was set to allow branch specific ω, and the ω calculated for the final branch to the target sequence was taken as the Goldman-Yang dN/dS score. The average of all of the pairwise

Nei-Gojobori dN/dS with the target sequence was taken as the Nei-Gojobori dN/dS score.

Figure 4.3 shows the receiver operating curve for PSILC and dN/dS on the Vega chromosome 6 test set. Table 4.1 shows summary statistics for each method. The PSILC posterior-nuc score has been modified for this graph by adding to the score 1/1000 * PSILC-nuc/dom. This was done because PSILC posterior-nuc scores a small fraction of functional genes as pseudogenes with probability 1, and so some means of distinguishing genes with identical scores was required. With this modification, PSILC posterior-nuc performs the best up to an error rate of 80, beyond which PSILC-nuc/dom performs best. Most significantly from the point of view of pseudogene annotation, PSILC posterior-nuc manages to correctly identify 40 pseudogenes before it incorrectly identifies a real gene as a pseudogene, whereas all of the other methods (aside from PSILC-nuc/dom, which identifies 3) scored a functional gene ahead of all pseudogenes. Thus, as previously mentioned, PSILC can be used to make assertions about the pseudogene status of genes, whereas other methods can only identify sets which are enriched for pseudogenes. The results from this curve can be compared to the similar results from the paper [CD04], which were obtained from an earlier version of PSILC. In this paper the approach was to calculate PSILC-prot/dom likelihoods purely on the basis of the amino acid sequence and amino acid rate matrices, and it was reported that this approach does better than dN/dS. The approach outlined in this chapter is different in that all likelihoods are calculated on codon sequences, which appears to have a negative impact on the PSILC-prot/dom results. However, PSILC-nuc/dom is more effective than both PSILC prot-dom from the earlier work and much more effective than PSILC nuc-dom from the earlier work.

Figure 4.4 and 4.5 shows the fraction of (pseudo)genes scoring above threshold vs threshold for the PSILC-nuc/dom score, Goldman Yang dN/dS and PSILC posterior-nuc. The dN/dS graph is plotted on a log x-axis for clarity – the PSILC scores are effectively already log based scores. The dN/dS pseudogene distribution is centered on $dN/dS \approx 1$ as expected, and at $dN/dS \approx 0.1$ for functional genes. However, both distributions are spread over a large range of dN/dS values, which makes a clean separation on this score difficult. On the other hand, the functional genes have a much sharper distribution under the PSILC-nuc/dom score, with most of the weight located at PSILC-nuc/dom ≈ 0, and the pseudogene distribution has most of its weight greater than 0, making a clean separation more effective. The separation is less pronounced in PSILC posterior-nuc (figure 4.5). In this case less than 4% of functional

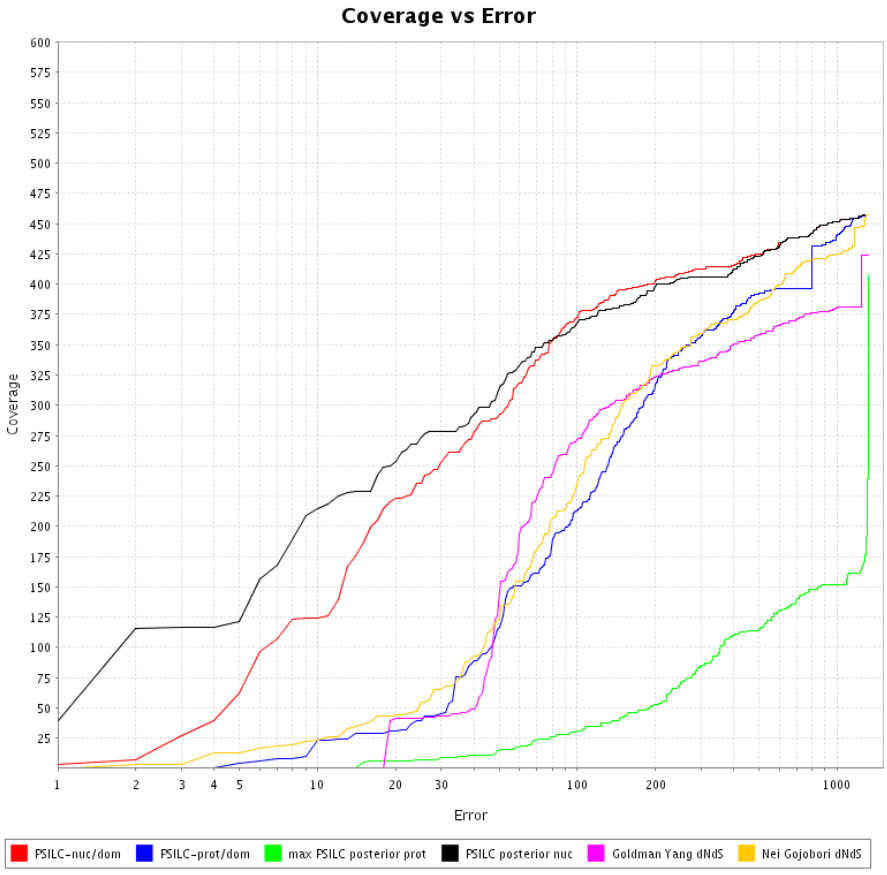

Figure 4.3: Coverage vs error curve for PSILC and dN/dS. The graph has been plotted on a log x-axis to reflect the fact that coverage level at low error rate is more important than at a high error rate. Several (pseudo)genes had a PSILC posterior nuc score of 1.0 - (pseudo)genes with the same PSILC posterior nuc score were ranked amongst themselves according to PSILC-nuc/dom score. A larger area under the curve represents a better discrimination between true and false pseudogenes.

genes have a PSILC posterior-nuc score greater than 0.5, whereas 67% of all pseudogenes score above 0.5. A small fraction of functional genes have PSILC posterior-nuc score of 1.0.

	Area under curve	OTT	MER
PSILC-nuc/dom	92.3%	3	180
PSILC posterior nuc	92.2%	40	177
PSILC-prot/dom	82.5%	0	328
Nei Gojobori dN/dS	82.4%	0	304
Goldman Yang dN/dS	81.7%	0	279
max PSILC posterior prot	29.3%	0	457

Table 4.1: Area under the coverage vs error curve, OTT (number of pseudogenes scored above the first functional gene) and MER (minimum error rate) for the different methods for classifying pseudogenes. For the PSILC-posterior nuc ranking, (pseudo)genes with the same PSILC posterior nuc score were ranked amongst themselves according to PSILC-nuc/dom score.

Figure 4.6 and figure 4.7 display the difference between a gene under selective pressure, and one which is evolving as a pseudogene. Figure 4.6 is a protein coding gene, while figure 4.7 is a pseudogene. Both have high PSILC-nuc/dom and PSILC-prot/dom scores (19,94 and 41,30 respectively). However the high-scoring region of figure 4.6 is limited to the N-terminal region, while it extends across the length of the protein for figure 4.6. The raw PSILC score would lead to the incorrect conclusion that both are pseudogenes, while the selection HMM correctly identifies the pseudogene and the gene under positive selection.

4.5 Results: detection of positive selection

In this section, I analyse the evolutionary pressures acting on three gene families: the APOBEC/AID family, occurring in vertebrates; the HIV Vif family; the Abalone sperm lysin family.

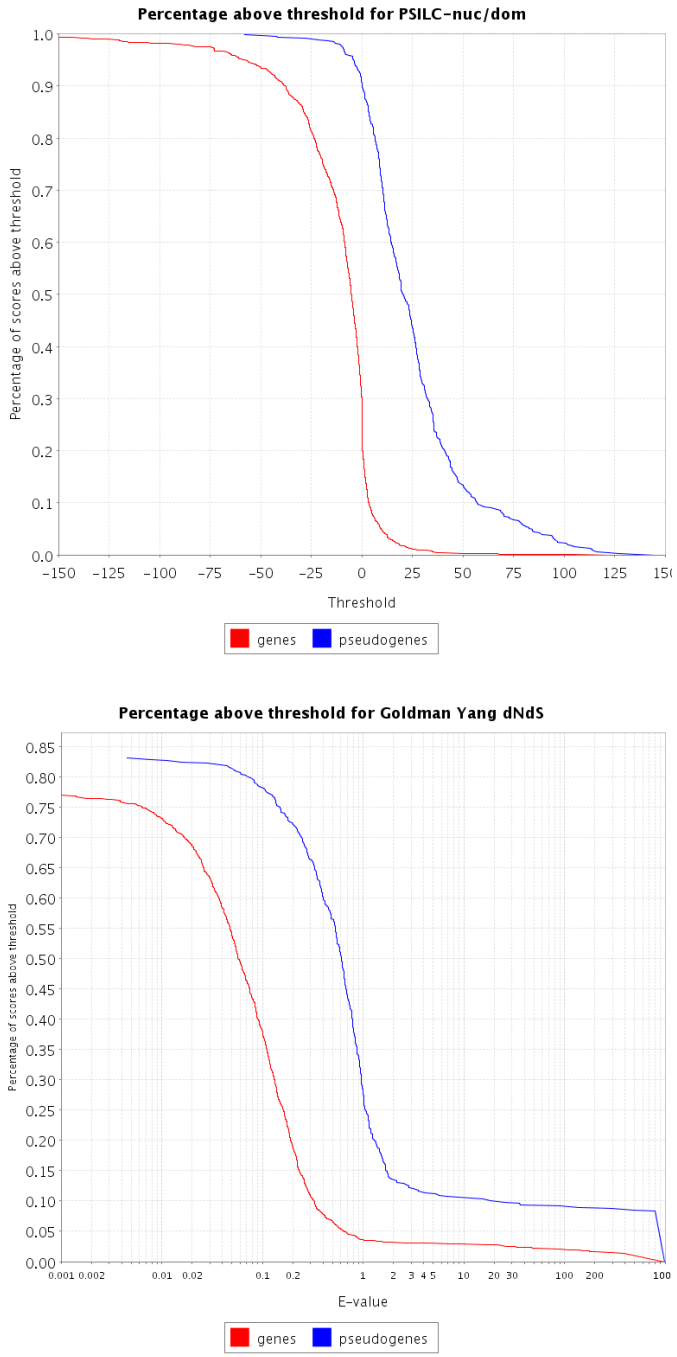

Figure 4.4: Comparison of discrimination between pseudogenes and functional genes between the PSILC-nuc/dom method (top graph) and Goldman Yang dN/dS (lower graph). In both graphs I plot the fraction of (pseudo)genes scoring above a particular threshold, with the pseudogenes represented by the blue line, and functional genes represented by the red line.

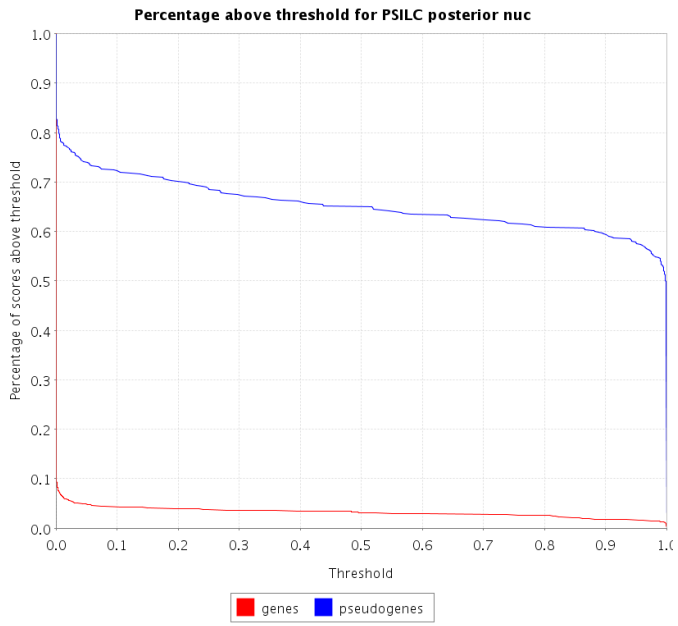

Figure 4.5: Comparison of discrimination between pseudogenes and functional genes using the PSILC posterior-nuc score.

4.5.1 Analysis of selective pressures on APOBEC/AID enzymes

Extensive evidence for positive selection within the APOBEC family has previously been found by Sawyer and co-workers[SEM04] using analysis of the ratio of the rate of synonymous and non-synonymous substitutions. I have reanalysed their data using PSILC, in order to compare results with those obtained by the authors, and to shed further light on the selective pressures driving APOBEC evolution. I have also analysed the selective pressures acting on HIV-1/HIV-2 and SIV Vif, which have been found to interact with APOBEC3G.

Background

The APOBEC/AID enzymes are part of a group of enzymes which deaminate cytosine to uracil on a polynucleotide molecule (such as single or double-stranded RNA or DNA). They are related to the cytosine and cytidine deaminases which deaminate a single nucleotide (or nucleoside or free base). In humans, the APOBEC family comprises eleven genes - APOBEC 1,2,3A,B,C, D/E, F, G, H and activation induced deaminase (AID). The APOBEC family is found throughout the vertebrates, including bony fish [CTPMN04].

APOBEC1 (apolipoprotein B mRNA editing complex catalytic subunit 1) is the cat-

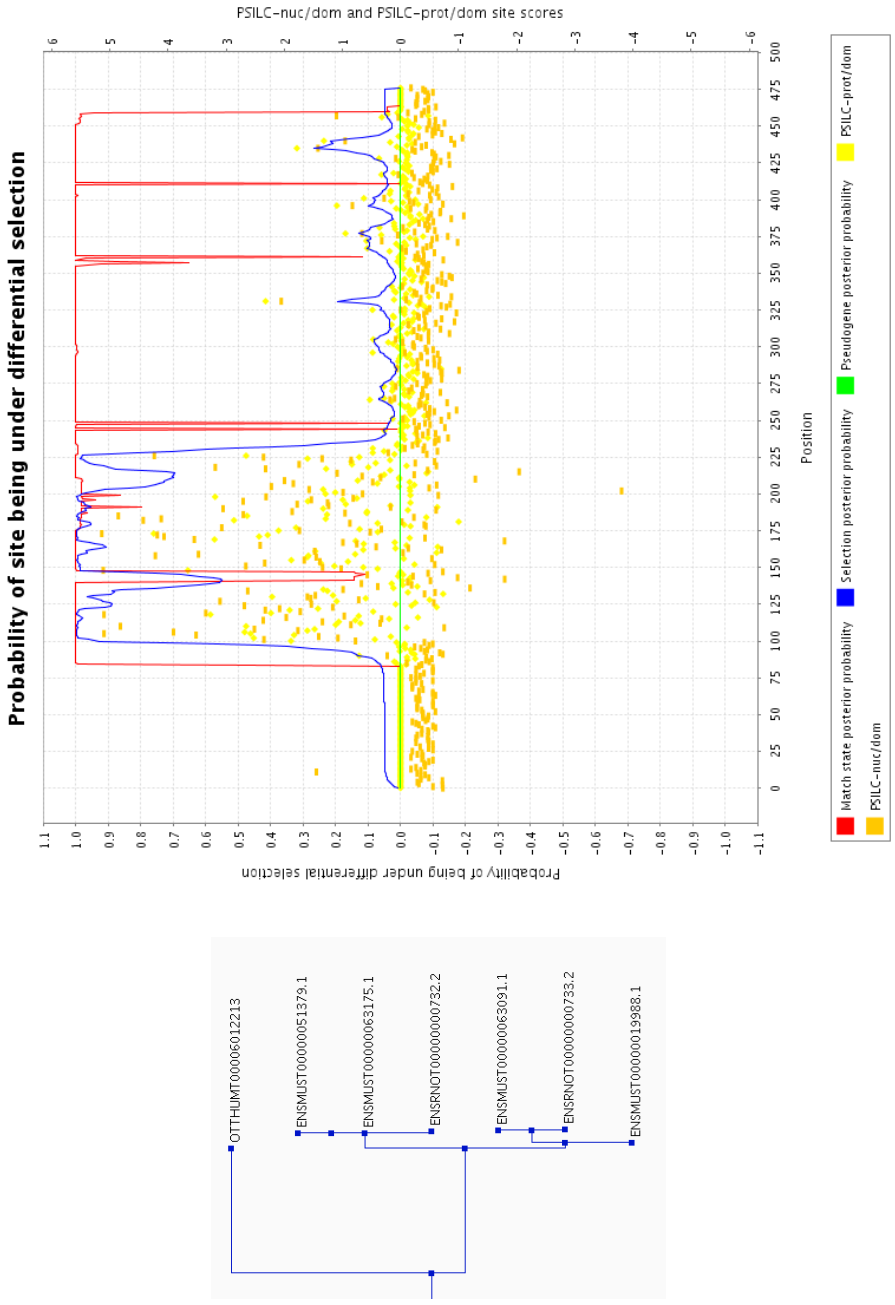

Figure 4.6: Pseudogene and selection status of Vega human (functional) gene OTTHUMT00006012213. Left: the Phyml tree of OTTHUMT00006012213 and homologues in mouse, human and rat genomes. Right: plot of PSILC nuc-dom(orange) and PSILC prot-dom(yellow) scores; Pfam domain match probability (to Pkinase, SH3, SH2 domains) (red); posterior probability of being under selection (blue); posterior probability of being pseudogene (green). Coordinates are relative to OTTHUMT00006012213 sequence.

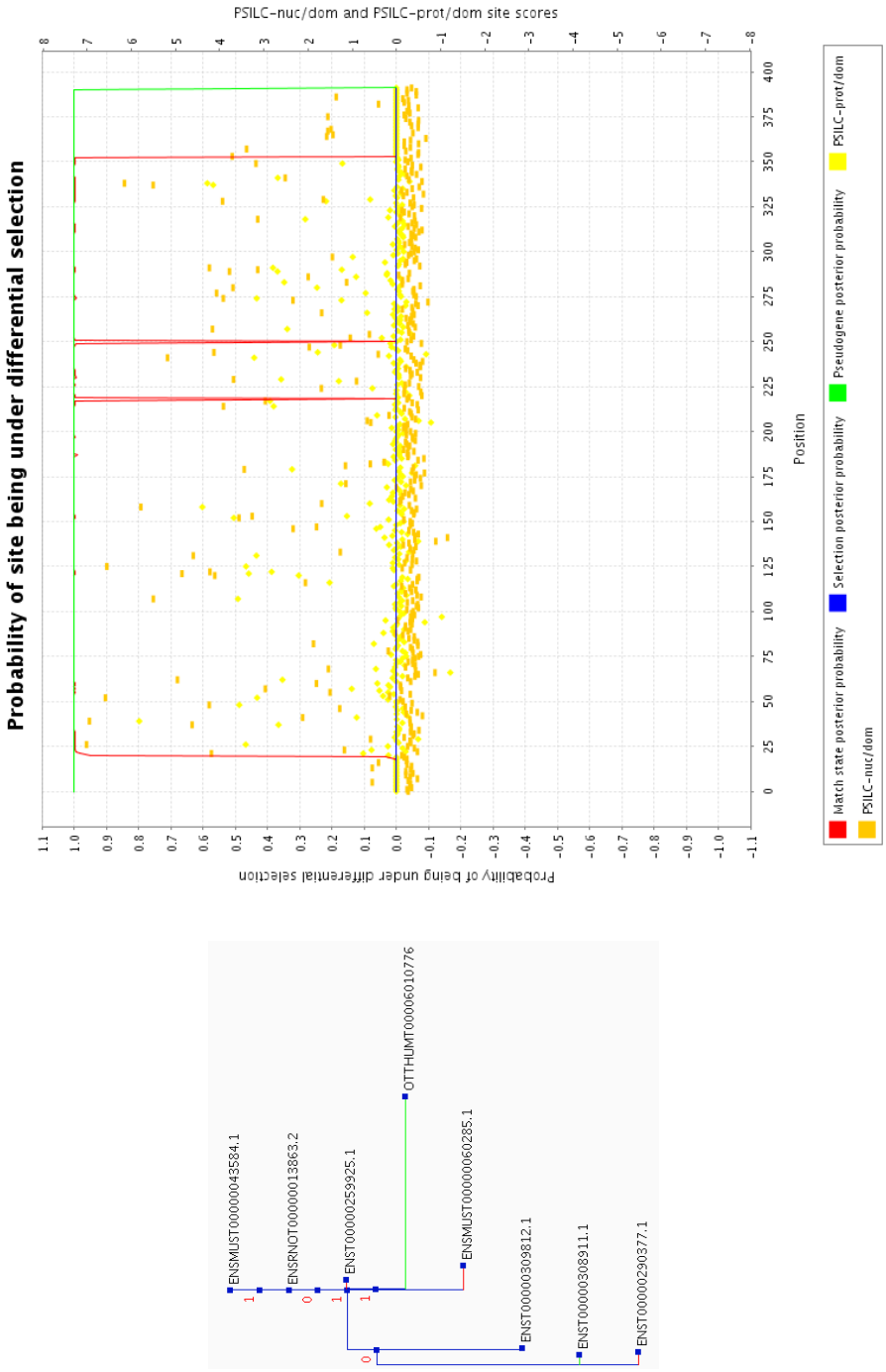

Figure 4.7: Pseudogene and selection status of Vega human pseudo-gene OTTHUMT00006009362. Left: the PhymI tree of OTTHUMT00006009362 and homologues in mouse, human and rat genomes. Right: plot of PSILC nuc-dom(orange) and PSILC prot-dom(yellow) scores; Pfam domain match probability (to Pkinase, SH3, SH2 domains) (red); posterior probability of being under selection (blue); posterior probability of being pseudogene (green). Coordinates are relative to OTTHUMT00006009362 sequence.

alytic subunit of a complex which deaminates cytidine[6666] of the mRNA of apolipoprotein B (ApoB) in the liver, thus creating a premature stop codon and a truncated form (48%) of the protein [TBD93]. Both the truncated and full length ApoB protein are involved in the transport of lipids and cholesterol. AID is expressed in germinal center B cells where it is required for immunoglobulin class switch recombination, somatic hyper-mutation and gene conversion. AID was initially proposed to also act as an RNA editing enzyme, however subsequent experiments have demonstrated the ability and preference for AID to deaminate cytosine in single stranded DNA [PMHN02].

The APOBEC3 family is only found in mammals. Non-primate mammals have a single APOBEC3 gene; however 8 are present in primates. APOBEC3A-APOBEC3G are encoded on a 130kb stretch of chromosome 22 in the same orientation[JCB$^+$02]. The APOBEC3 locus is rich in repetitive retroviral elements, which suggests that the rapid expansion in primates was facilitated by retroviral elements. According to EST evidence, APOBEC3D and APOBEC3E are likely part of the same protein. A probable processed APOBEC3 pseudogene has been detected on chromosome 12, due to the fact that it has no introns.

APOBEC3G has been identified as the gene which inhibits infection with HIV-1 strains lacking the virion infectivity factor (Vif) [SGCM02]. In the absence of Vif, APOBEC3G is packaged into retroviral particles in the producer cell. After infection of target cells by viruses produced in APOBEC3G expressing cells, APOBEC3G deaminates cytosine to uracil in the nascent viral minus strand during reverse transcription [ZYP$^+$03] . These mutations cannot be repaired correctly as the viral RNA template is simultaneously degraded during reverse transcription. Hence APOBEC3G does not affect the viral output from a producer cells, but rather protects the target cell from infection. In wild-type HIV encoding the Vif protein, APOBEC3G mediated mutation of viral cDNA is prevented by Vif inducing polyubiquitination of ABOBEC3G and so making it a target for degradation by the 26S proteasome[CHN03]. Human APOBEC3G is resistant to African green monkey SIV induced degradation but susceptible to HIV-1 Vif, and conversely African green monkey SIV is resistant to HIV-1 Vif but susceptible to African green monkey SIV Vif. The difference in sensitivity has been mapped to residue 128 in Human APOBEC3G, which is aspartic acid(D) in human APOBEC3G and lysine(K). Mutating D $\rightarrow$ K in human APOBEC3G renders it resistant to HIV1-Vif but sensitive to African green monkey SIV Vif [BDWC04]. The region of interaction between Vif

and APOBEC3G has been mapped to the residues 54-124 [CTPMN04].

APOBEC3F is adjacent to APOBEC3G, shares over 90% similarity in the upstream promoter region, and is widely co-expressed in human cells, suggesting that APOBEC3F is co-regulated with APOBEC3G. APOBEC3F is also packaged into retroviral particles; also has an effect than on viral infectivity (although smaller than APOBEC3G) and also interacts with Vif. APOBEC3B and APOBEC3C are also packaged into retroviral particles and have a weak effect on viral infectivity. ABOBEC3B and APOBEC3C are completely and partially resistant respectively to HIV Vif induced degradation.

Thus, the APOBEC3 family is in genetic conflict with the HIV/SIV Vif protein. This type of genetic interaction could be expected to lead to fixation of mutations which change the conformation of the APOBEC3G protein, as well as mutations in the Vif protein. Sawyer et al. find that the signal for APOBEC3G positive selection predates the appearance of modern lentiviruses, and conclude that APOBEC3G evolution is only partially caused by modern lentiviruses. APOBEC3G is also abundantly expressed in the germline[JCB+02]. It has been suggested that APOBEC3G is required in the germline to restrict the activity of the long-terminal bearing(LTR) human endogenous retroviruses (HERVs). The life-cycle of HERVs is similar to retroviruses, including expression in the cytoplasm (where APOBEC3G is active) and a reverse transcription stage which would be susceptible to APOBEC mediated cDNA editing. Sawyer et al. suggest that the HERVs may be a more important driving force for the evolution of APOBEC3 than the primate lentiviruses.

Method

Primate APOBEC3G DNA sequence was obtained from Sarah Sawyer, which was published in [SEM04]. DNA sequence for vertebrate APOBEC, AID sequences was obtained from Silvo Conticello, which was published in [CTPMN04]. APOBEC3 is internally duplicated with respect to APOBEC2 and so has two homologous copies of APOBEC2, whereas AID and APOBEC1 only have one homologous copy. The N-terminal and C-terminal copies within APOBEC3 proteins were split into separate sequences. A protein alignment was created using MUSCLE [Edg04], and the DNA alignment was inferred from the protein alignment. A tree was generated using Phyml [GG03] from the DNA sequence, using a HKY evolutionary model and 4 rate categories, and training the transition to transversion ratio. A nucleotide

rather than protein rate matrix was used because the primate APOBEC3G sequences are highly similar at a protein level. Two minor edits were applied to the Phyml tree so that the phylogeny within APOBEC3 was consistent between the N- and C-terminal sequences (which was obtained from a Phyml derived tree of full length APOBEC3 sequences). The tree obtained agreed with the widely accepted taxonomy, but differed slightly from the tree published in [SEM04] in the relative position of baboon and macacques (this branching is undefined in the NCBI taxonomy). Branch lengths were derived as those which maximised the likelihood under the compound WAG+gwF : HKY model (as discussed in section 4.3) and the assumption of a molecular clock. The maximum likelihood transition/transversion ratio is 2.1 and the maximum likelihood f value is 0.83.

HIV1, HIV2 and SIV Vif DNA sequences were obtained from the Los Alamos national laboratory at `http://www.hiv.lanl.gov/content/hiv-db/`. This data set consists of 558 HIV-1 Vif sequences, 47 HIV-2 Vif sequences and 21 SIV sequences. These sequences were aligned as protein using MUSCLE, and the DNA alignment was inferred from the protein alignment. These sequences were filtered so that only the 40 most diverse Vif proteins were kept in the set, resulting in 13 HIV-1 genes, 9 HIV-2 genes and 18 SIV genes. The tree for this protein set was built using Phyml, with a WAG rate matrix and 4 rate categories. The tree was re-rooted so that the HIV1 and HIV2 genes each formed a cluster, which was possible given the original Phyml tree. The maximum likelihood transition/transversion ratio was 2.4 and f value is 0.63.

Each of the sequences in the APOBEC/AID alignment had a significant match to the Pfam APOBEC-C family (e-values in range 1e-12 to 1e-20). Some of the sequences had a significant match to the Pfam dCMP_cyt_deam family, however several members did not, and none of the matches were particularly strong. This family is much longer (144 match states) than the highly conserved zinc co-ordinating motif discovered in structural studies of bacterial cytidine deaminases and of yeast cytosine and cytidine deaminases. Hence, a new HMMER HMM – which I will call APOBEC-N – was built from an alignment the N-terminal regions of the APOBEC family, dCMP-cytidine deaminases and adenosine deaminases which act on RNA (ADAR1-3) or tRNA (ADAT1-3). This family had 58 match states. The sequences all had very significant matches to this new family (1e-18 to 1e-21). Each of the sequences in the Vif alignment had a significant match to the Pfam Vif domain with e-value in the range

1e-7 to 1e-40.

The tree and HMMER hidden Markov models for both APOBEC/AID and Vif were given as input to PSILC, which was run in recursive mode with selection transition probabilities given by the diagram 4.2.

Results

The tree obtained by PSILC is shown in figure 4.8, and can be compared with the tree obtained by Sawyer et al. [SEM04] in figure 4.10. The C-terminal of an APOBEC3 pseudogene included in the dataset is correctly detected, while the N-terminal has 34% PSILC posterior-nuc score. The analysis also suggests that APOBEC3H is a pseudogene. There is a strong selection signal in the N-termini of APOBEC3G in both Cercopithecinae (old world monkeys) and Hominidae, but not the C-termini, whereas the pattern is reversed for Platyrhinni (new world monkeys). It is interesting to note that there is no lentivirus which targets new world monkeys, but we might speculate the N-terminal evolution is driven by interaction with either HERVs or other reverse-transcribed viruses. The site-specific likelihood ratios and posterior probabilities at these nodes have been plotted in figure 4.9. The position of the peaks in posterior probability (above 0.75) for both Cercopithecinae and Hominidae have been mapped to the structure of Yeast cytosine deaminase in 4.11. The position of human APOBEC3G residue 128 critical for the species specificity of Vif effectiveness maps to position 118 in this structure. All co-ordinates are given in terms of the yeast structure. It can be seen that the predicted selected sites, as well as the Vif specificity site could potentially be involved in conformational changes, or steric hindrance of the Vif APOBEC3G interaction. The Hominidae peak at 129 corresponds to a glycine {GGA, GGG, GGT} $\rightarrow$ arginine (CGT) mutation at this node. Glycine is strongly conserved at this position according to the profile HMM, and chemically quite different from Arginine, so this change would appear to change the conformation of the protein. The Cercopithecinae peak at 153 corresponds to tryptophan(TGG) $\rightarrow$ arginine (CGG) mutation, again tryptophan is strongly conserved at this position in the profile.

Other members of the APOBEC3 family as well as APOBEC1 also appear to be under strong selection. However, AID and APOBEC2 positive selection in mammals appears to be not as strong, which is consistent with the findings of Sawyer et al.

The analysis of the Vif proteins is displayed in figure 4.12. Again, extensive positive

selection has been detected in the tree. The HIV-1 Vif proteins display a stronger and more consistent signal for positive selection than the HIV-2 Vif proteins. This can also be seen in figure 4.13, in which the site specific likelihoods and posterior probabilities are plotted for two external nodes in each of HIV-1, HIV-2 and SIV-1. The HIV-1 Vif protein in the top line (HIV-1_C.BW.) displays a positive selection signal across the length of the protein. HIV-1_B.AU displays a strong pseudogene signal (green circles and purple line) as well as a strong positive selection signal (orange squares). This is an example where PSILC incorrectly (although the functionality of this protein has not been tested) identifies a gene as a pseudogene due to a high rate of positive selection across the length of the protein. The HIV-2 Vif proteins in this diagram, on the other hand, only display a selection signal at the C-terminus, and the N-terminus appears to be relatively well conserved. Some SIV proteins appear to be very highly selected (e.g. SIV_GSN in the top line) while others (SIV_GRV) display less positive selection and a higher level of conservation.

4.5.2 Analysis of selective pressures on Abalone lysin protein

I investigate the selective pressures acting on the Abalone lysin protein, which is a 16kda protein found in Abalone, and acts in conjunction with a paralogous 18kda lysin protein on the egg vitelline envelope (VE). The 18kda protein was discussed in section 2.4.2 where the Pfam Egg_lysin domain was identified in the divergent *Haliotis fulgens* protein. As discussed in this section, the 16kda protein creates a hole in the vitelline envelope and the 18kda protein is thought to mediate membrane fusion between the gametes[SV95].

The cDNA sequences for lysin from 20 abalone species has been sequenced and analysed for positive selection (using the method of Nei and Gojobori [NG86]) by [LOV95]. The authors identified a $\omega = d_n/d_s$ ratio greater than 1 when closely related species are compared, but less than 1 when distantly related species are compared, providing evidence for positive selection. The authors also hypothesised that the small ω values for distantly related species may be due to saturation effects. Subsequently, Yang and co-workers [YSV00] showed that saturation was unlikely to account for low ω values in divergent species and that ω varies greatly between sites on the lysin protein. These authors also identified regions of the protein under positive selection.

I re-analysed the data set analysed by Yang et al in [YSV00], to determine whether

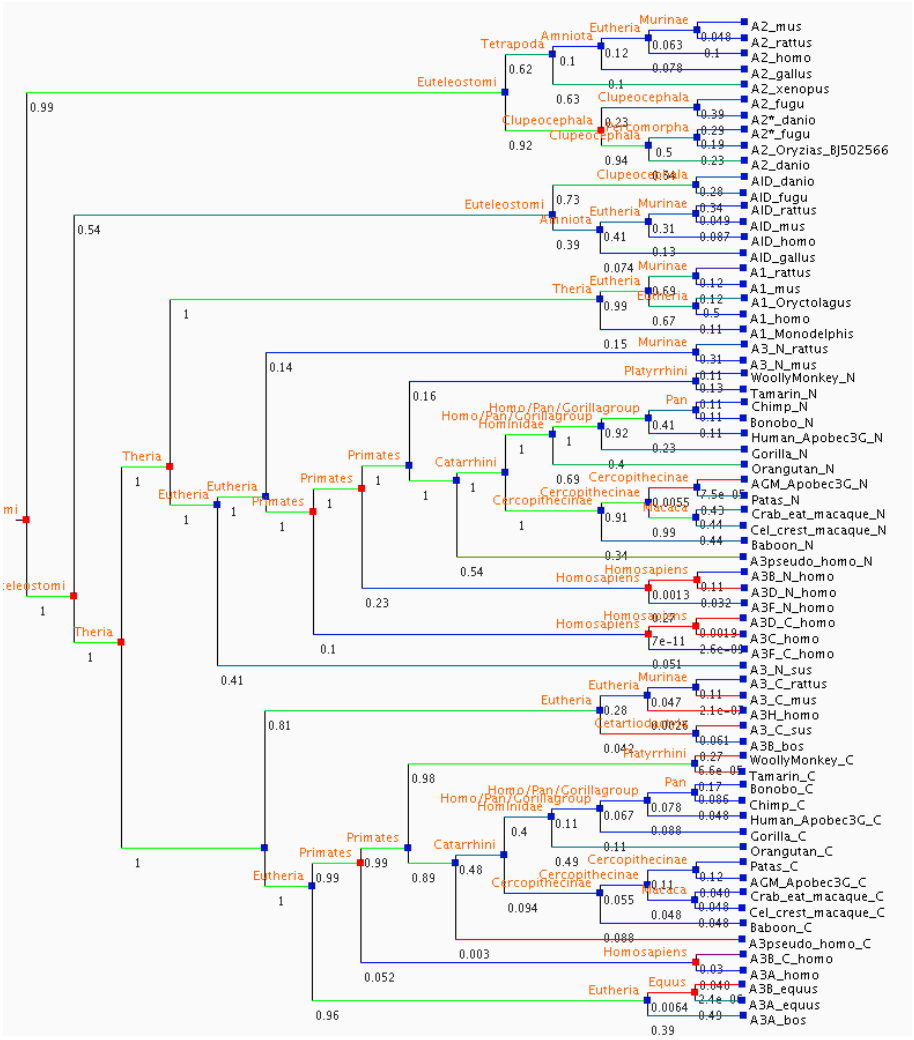

Figure 4.8: Tree of APOBEC/AID family, showing extensive positive selection. Green branches to a node indicate strong evidence for positive selection below this node, whereas red branches indicate strong evidence for pseudogene evolution below a given node. Blue branches indicate lack of evidence for selection and pseudogene evolution, and hence purifying selection. The numbers below a branch are the max PSILC posterior-prot score below and including that branch, which is used here as a score of positive selection.

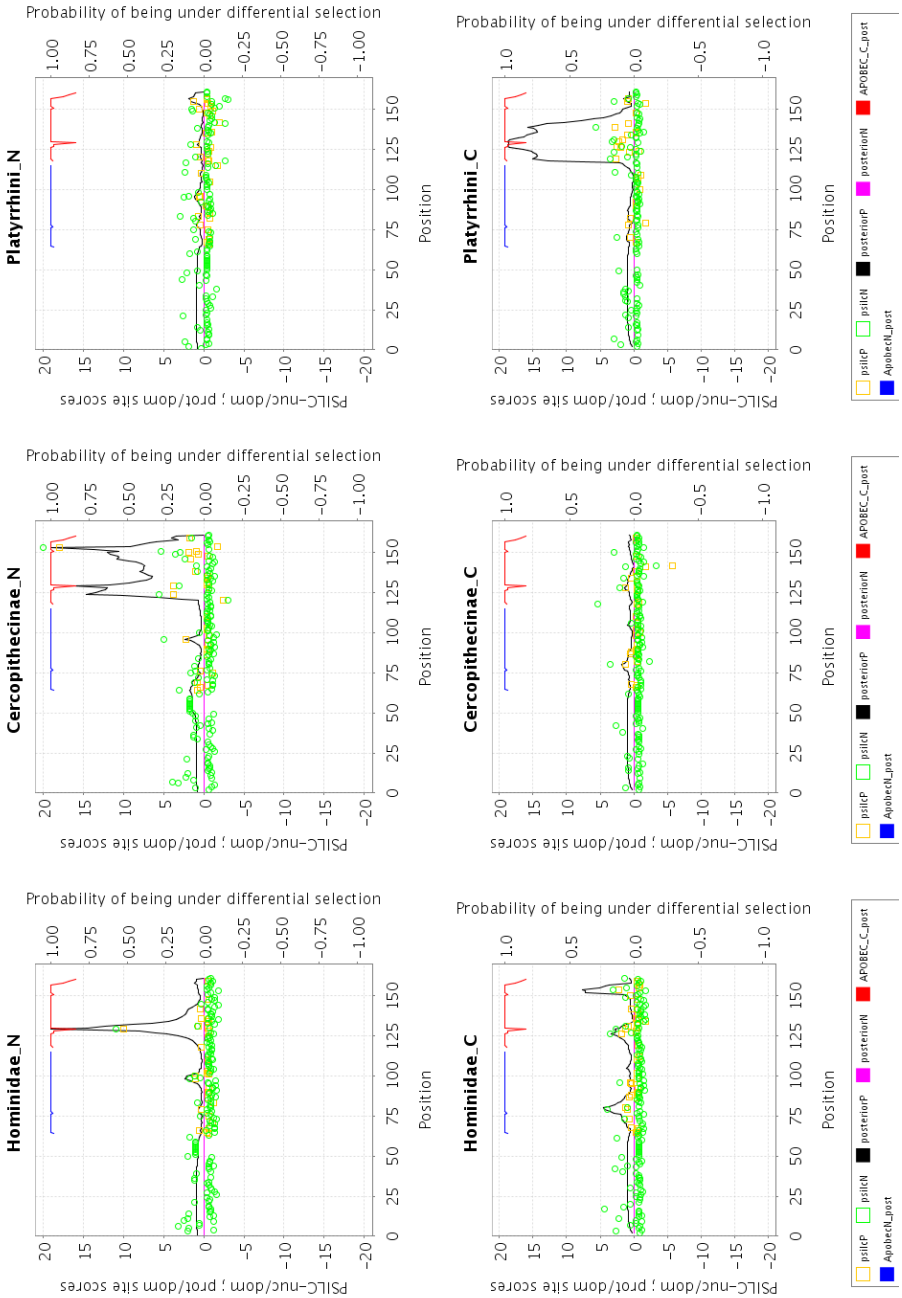

Figure 4.9: Site-specific graphs of selection acting on APOBEC3G genes. Green circles/orange squares indicate the site specific PSILC-nuc/dom and PSILC-prot/dom scores respectively, which, for clarity, are only plotted if less than -0.3 or greater than 0.3. The black/purple line indicates the posterior probability of being in a positive selection or pseudogene state respectively. Note that the purple line runs along the x axis in all of the diagrams, and hence is not clearly visible. The blue/red line is the posterior probability of being in a match state of the APOBEC_N/APOBEC_C families respectively. For clarity, these lines are only plotted for probability greater than 0.5.

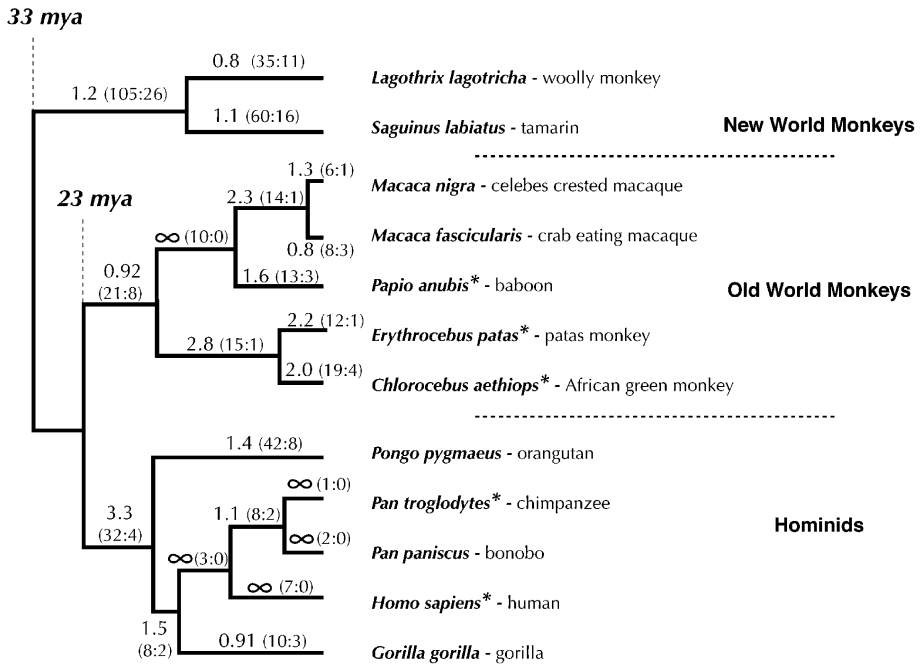

Figure 4.10: Diagram of the tree of full-length APOBEC3G sequences taken from [SEM04]. The starred species are those which are infected by HIV/SIV. The numbers on the branch indicate the maximum likelihood value of dN/dS estimated by PAML using the free-branches model. The numbers in brackets are the number of synonymous and non-synonymous substitutions, calculated by inferring the ancestral sequences, us

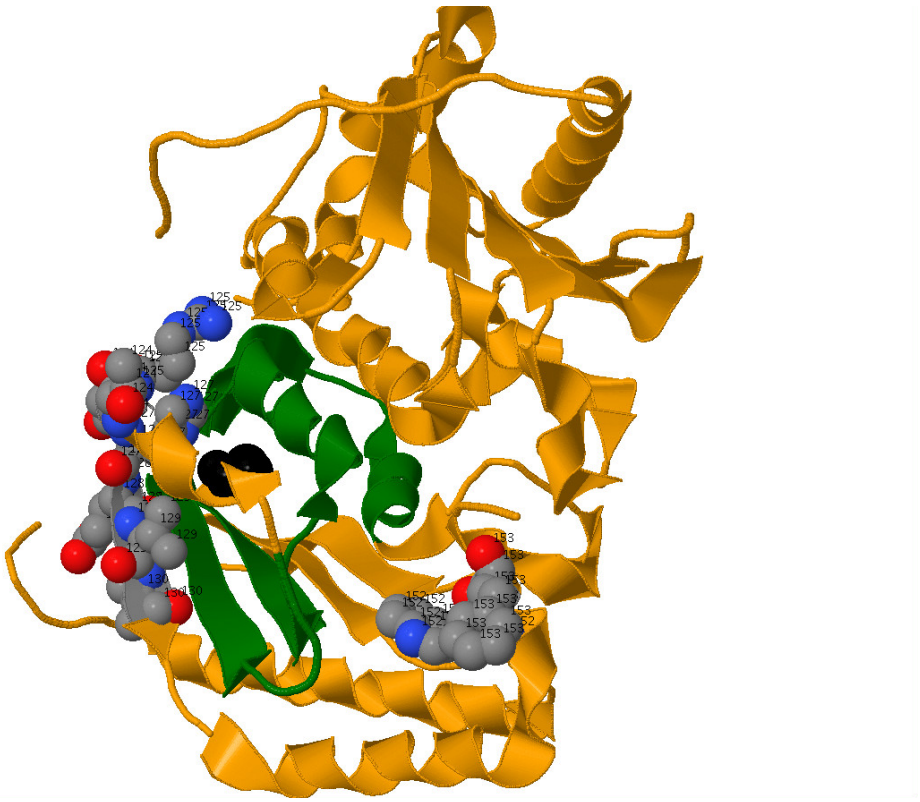

Figure 4.11: The structure of yeast cytosine deaminase, which is homologous to the APOBEC/AID family. The structure is of the homo-dimer. The region homologous to the Vif binding region in human APOBEC3G is drawn in green. The residue which aligns with residue 128 in the human APOBEC3G family is mapped to position 118 in this structure, and shown in black. PSILC predictions of positively selected regions are shown via the space-fill representation.

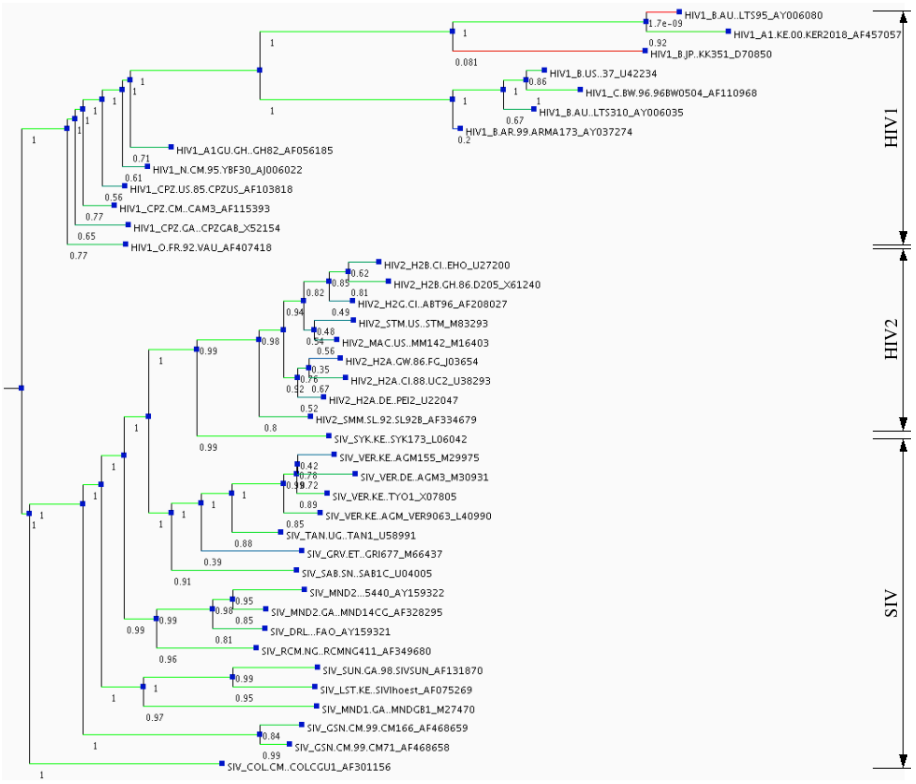

Figure 4.12: Tree of Vif proteins showing extensive selection. A green branch indicates strong evidence for selection on the branch to and below that node, whereas a red branch indicates the gene is evolving under a neutral DNA model. The numbers given below the branches indicate the maximum posterior probability of selection acting on the branch to this node and the subtree below the node.

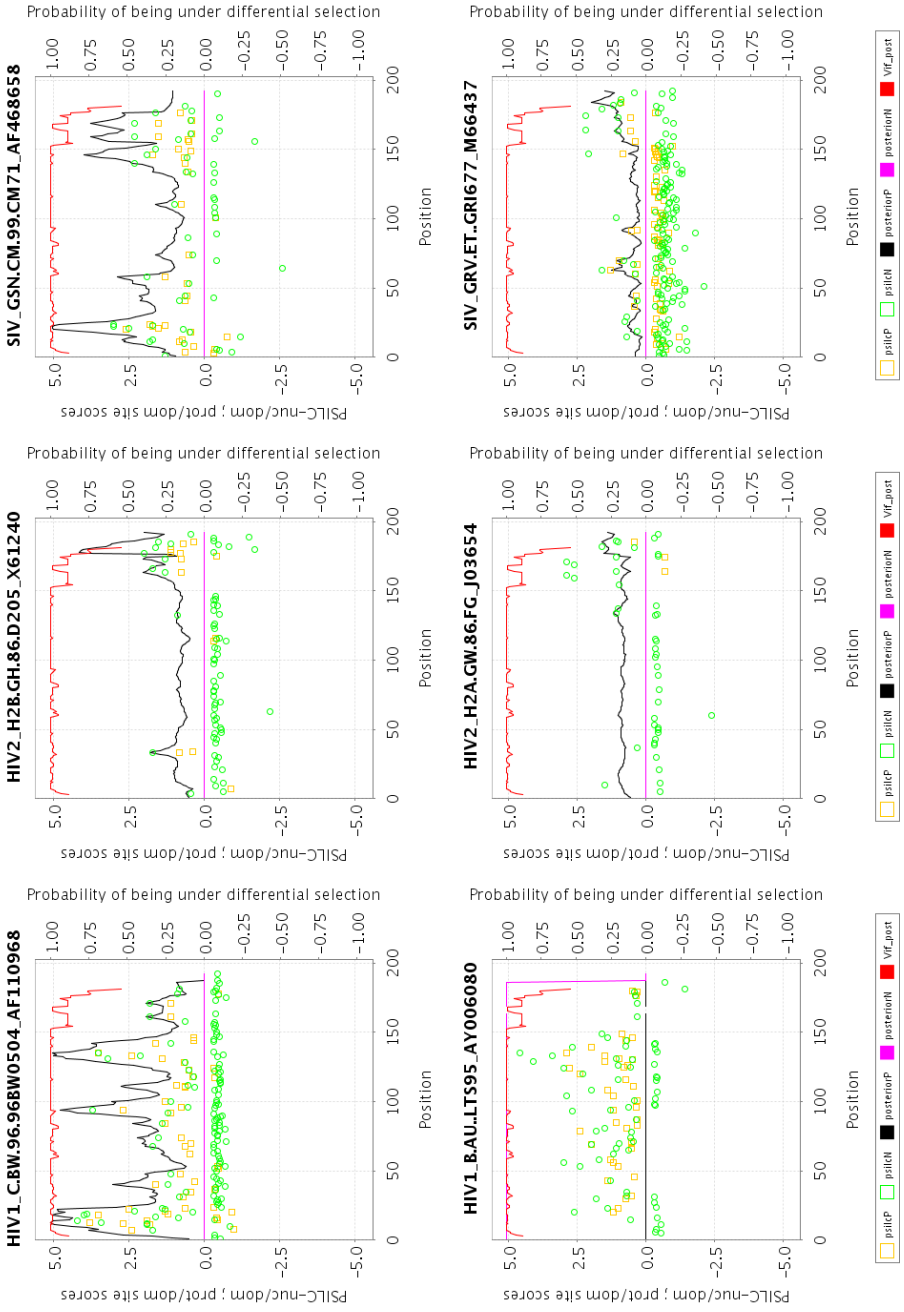

Figure 4.13: Site-specific graphs of selection acting on Vif genes. Green circles/orange squares indicate the site specific PSILC-nuc/dom and PSILC-prot/dom scores respectively, which, for clarity, are only plotted if less than -0.3 or greater than 0.3. The black/purple line indicates the posterior probability of being in a positive selection or pseudogene state respectively. Note that the purple line runs along the x axis in all of the diagrams, and hence is not clearly visible. The blue/red line is the posterior probability of being in a match state of the Vif family, which are only plotted for probability greater than 0.5.

positive selection can be identified by looking for amino acid changes which disrupt the profile HMM consensus and hence also the protein conformation. The tree topology was obtained from the paper [YSV00], and maximum likelihood branch lengths and substitution model parameters were estimated under the compound WAG+gwF : HKY model (discussed in section 4.3) and the assumption of a molecular clock. The maximum likelihood transition/transversion ratio is 1.6 and the maximum likelihood f value is 0.77.

PSILC was run in recursive mode with selection HMM transitions probabilities as shown in 4.2. With these transition probabilities, PSILC only detected a positive selection signal in the C-terminus of H. cracherodii and H. rufescens (with posterior probabilities of 30% and 25% respectively). This suggests that the lysin proteins are not under positive selection from the point of view of large structural changes. It may, however, still be the case that the lysin proteins are evolving under a weaker diversifying pressure for changes which do not disrupt the protein structure. To investigate this second hypothesis in more detail, the transition probabilities were adjusted to allow transitions in and out of the neutral DNA model from the domain model, and to relax the transition probabilities to the positively selected state. The probabilities used were start → {selection 0.05, pseudogene 0.01, purifying 0.94}; purifying →{selection 0.05, pseudogene 0.01, purifying 0.93, end 0.01}; selection →{purifying 0.2, selection 0.49, end 0.01}; pseudogene →{pseudogene 0.98, purifying 0.01, end 0.01}.

Figure 4.14 shows the overall results with the relaxed transition parameters, and can be compared with the tree in figure 4.17. Again *H. cracherodii* and *H. rufescens* display the strongest signal for positive selection as determined by a protein coding model. Several branches have high posterior probability of neutral DNA evolution, supporting the hypothesis that although the evolution of the lysin has been largely conserved with respect to structure, it has been freer to explore alternative amino-acids which do not affect the structure. Figure 4.15 displays the site specific scores at particular nodes in the lysin tree. Each of the three graphs in the top line, as well as the first graph in the second line are of clades with all species from the same geographic region (California, Japan, California and California respectively). The remaining two graphs are of clades with all descendants dispersed geographically. If, as hypothesised in previous papers, evolution is driven pressure to reduce heterospecific fertilization amongst abalone within the same geographical region, then the geographically restricted clades should exhibit more selection. Although the first three of the geographically

restricted clades appear to display more selection than the two geographically diverse clades, the geographically restricted *H. scolaris* → *cyclobates* clade breaks the rule. The top 3 clades display selection in similar regions of the protein. The selection peaks from the three graphs on the top line are plotted on the structure of lysin in figure 4.16. It is interesting that the N-terminal lysin segment evolving as neutral DNA and the C-terminal section evolving as neutral protein are spatially adjacent and external to the protein structure. This figure should be compared with the predicted positions of positive selection in [YSV00] displayed in 4.17. The PSILC predictions agree with the PAML predictions at sites 36, 41, 113, but PSILC also predicts sites 107-109 to be positively selected.

4.6 Results: Global scan for pseudogenes and positive selection

I conducted a global scan for positive selection and pseudogenes in the genomes of 4 mammals (*H. sapiens, P. troglodytes, M. musculus, R. norvegicus*), 1 bird (*G. gallus*), 2 fish (*F. rubripes, D. rerio*), 2 insects (*D. melanogaster and A. Gambiae*) and 2 nematodes (*C. Briggsae and C.Elegans*). The PHIGS database `http://phigs.jgi-psf.org` clusters proteins from complete Opisthokont (Fungi and Metazoa) genomes into protein gene families. I consider only those genomes which are also in the ENSEMBL database. All PHIGS clusters containing at least one human protein, at least 3 members in total and matching at least on Pfam domain, were extracted from the PHIGS database. Protein coding DNA sequence for any sequence from the above 11 genomes in the clusters was extracted from the ENSEMBL database, and formed the inputs for PSILC. Trees for each of the clusters were built as neighbour joining trees based on maximum likelihood distances calculated using the WAG protein rate matrices and a single rate category. PSILC was only applied to the leaf nodes due to the difficulty in rooting trees and to reduce running time. PSILC used a WAG model of protein evolution and a HKY model of DNA evolution.

Figure 4.18 shows the number of human genes in scored PHIGS clusters with high pseudogene scores. There are 282 genes with PSILC posterior nuc score of 1.0, and 110 genes with PSILC posterior nuc score of 1 and PSILC-nuc/dom score of greater than 50. No functional genes in the Vega test set scored above this combined threshold, thus each of these

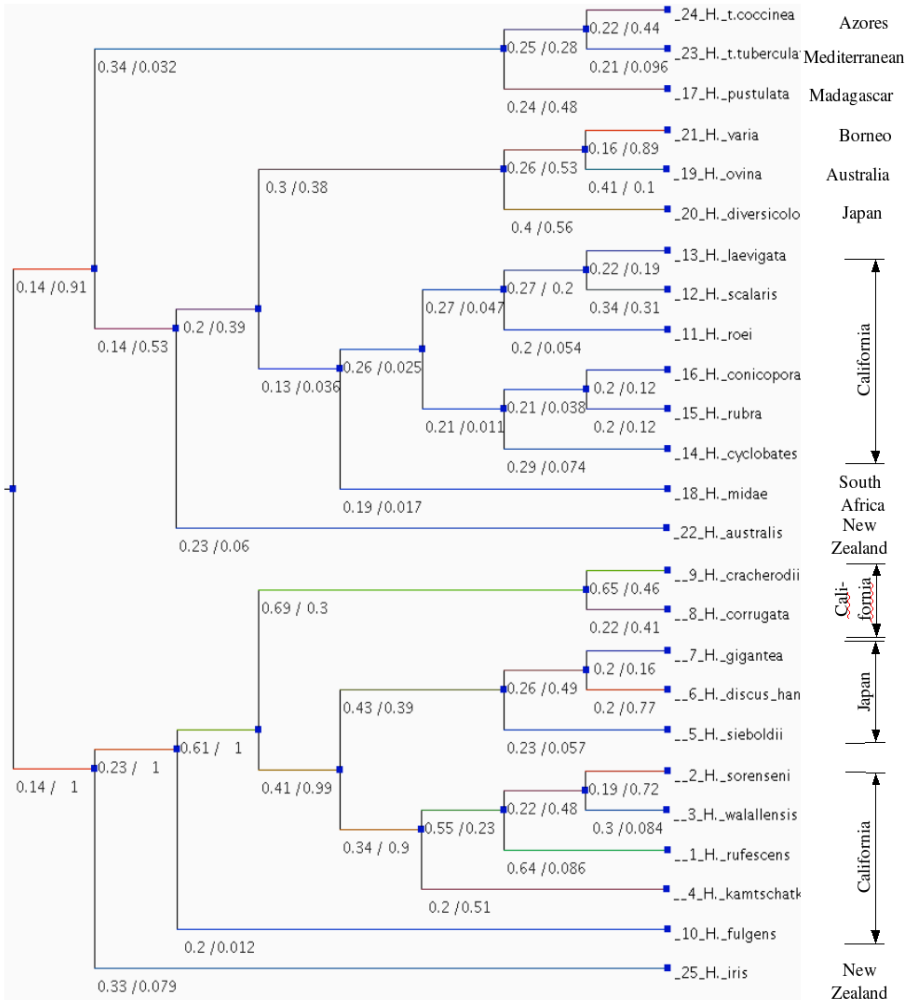

Figure 4.14: Tree of sperm lysin family, showing extensive 'non-structural' positive selection. Green branches to a node indicate support for evolution according to a neutral protein model rather than a domain constrained protein model, whereas red branches indicate support for a neutral DNA model rather than a protein domain constrained model. Blue branches indicate lack of evidence for positive selection and pseudogene evolution. The numbers on a branch are the maximum posterior probability of being in a neutral protein model (first number) and the maximum posterior probability of being in neutral DNA model (second number).

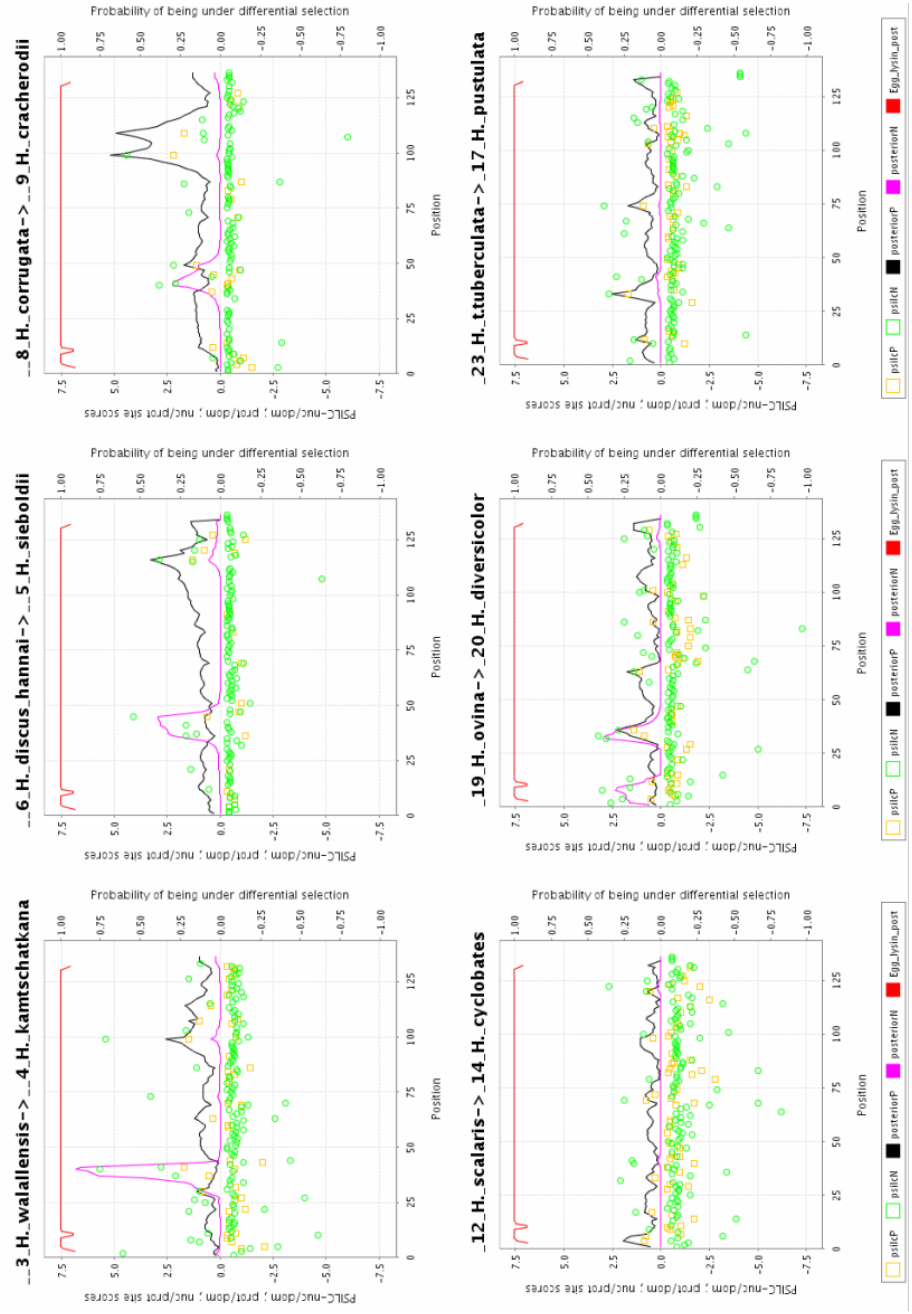

Figure 4.15: Site-specific graphs of selection acting on lysin genes. Green circles/orange squares indicate the site specific PSILC-nuc/dom and PSILC-prot/dom scores respectively, which, for clarity, are only plotted if less than -0.3 or greater than 0.3. The black/purple line indicates the posterior probability of being in a positive selection or pseudogene state respectively. The blue/red line is the posterior probability of being in a match state of the Egg_lysin domain. For clarity, these lines are only plotted for probability greater than 0.5. The relevant nodes in the tree for each graph are the most recent common ancestor of the two leaf nodes given in the graph titles.

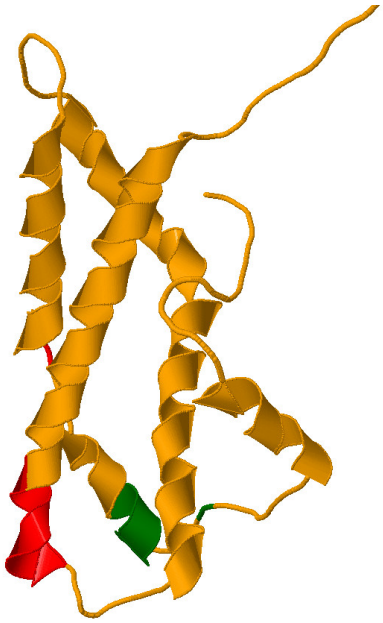

Figure 4.16: Structure of lysin, with regions of posterior probability of neutral DNA (red) or neutral protein evolution (green) greater than 50% in the clade *H. cracherodii* → *H. kamtschatkana*.

110 genes is highly likely to be a pseudogene. ENSEMBL [BAB+04] builds genes by searching for homology to known proteins using GeneWise [BD00]. When this procedure is applied to a pseudogene with a frame-shift, GeneWise will in some instances introduce a small intron to compensate for a frame-shift. Thus a short minimum intron length in an ENSEMBL gene is an indication that the gene is in fact a frame-shifted pseudogene. The frequency distribution of minimum intron lengths for multi-exon genes with PSILC posterior-nuc score of 1.0 has been plotted in figure 4.19. As would be expected for a pseudogene set, a significant number of members (28%) have minimum intron length of less than 5 base-pairs, whereas a small fraction of genes in the full set have intron lengths less than 5 base-pairs.

Figure 4.20 shows for each of 11 species the number of clusters with a protein in that species with maximum posterior probability of being under selection greater than a given threshold on max PSILC posterior-prot. As clusters are included only if they contain a human protein, the total number of clusters with a protein in each species loosely reflects the evolutionary distance from that species to human. For instance the other mammals occur in approximately 86% of clusters, while the nematode worms only occur in approximately

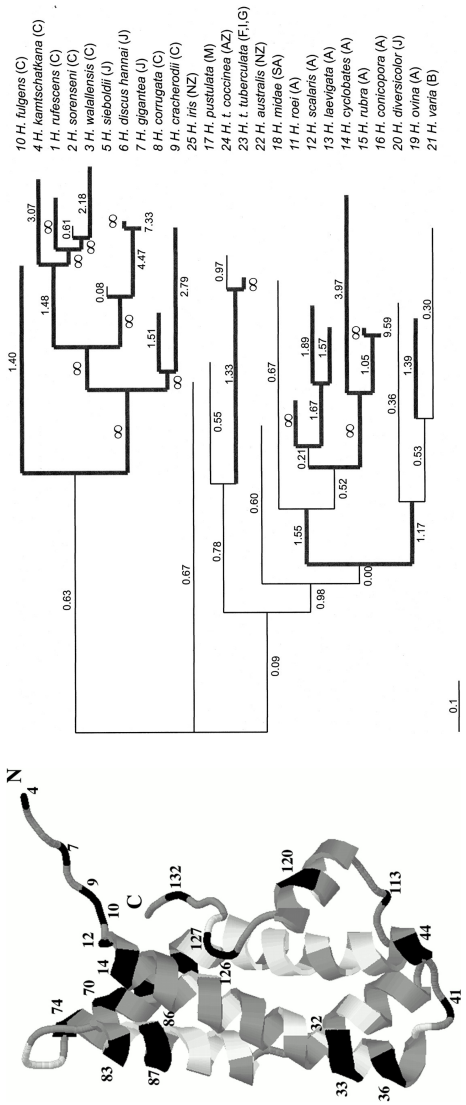

Figure 4.17: Taken from [YSV00]. Top: lysin tree, with the maximum likelihood estimates of dN/dS using PAML in the free ratios model on the branches of the tree. The thick lines indicate those branches with $dN/dS > 1$. Bottom: structure of lysin with sites inferred to be under positive selection (with greater than 99% posterior probability) coloured in black. Sites in white are under purifying selection.

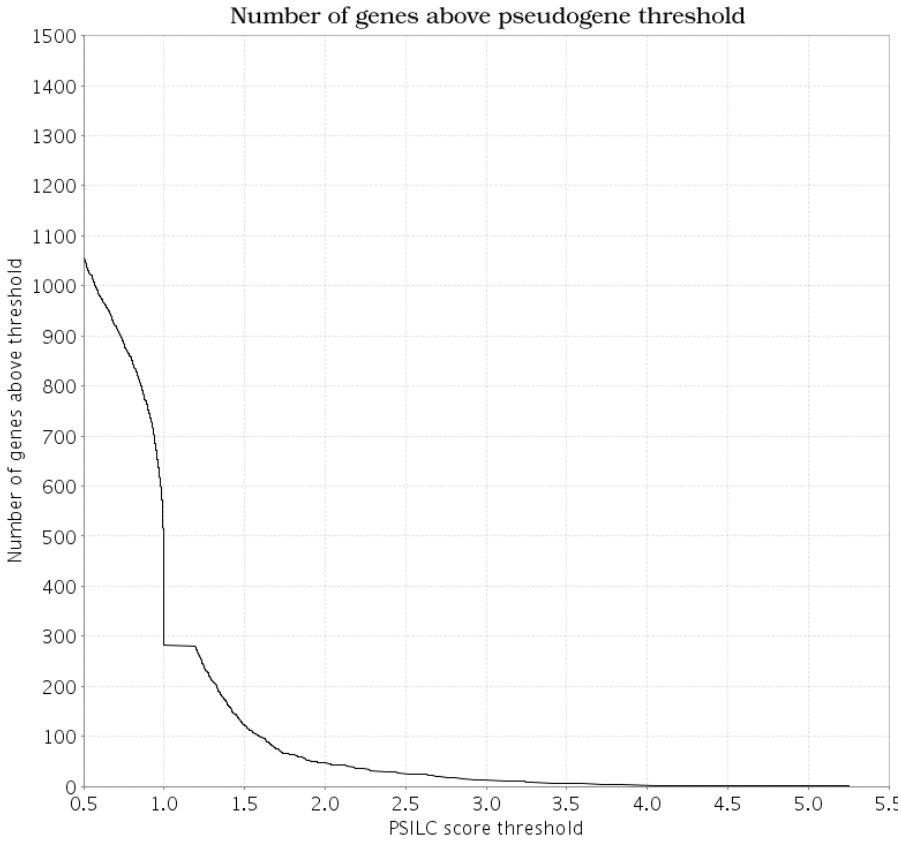

Figure 4.18: Number of human genes in clusters with combined PSILC posterior-nuc threshold/ PSILC nuc-dom score above threshold. The combined score was calculated by adding PSILC-nuc/dom / 100 to all genes with a PSILC posterior-nuc score of 1. Thus a score of 1.5 indicates a PSILC posterior-nuc score of 1 and a PSILC nuc/dom score of 50. No functional genes scored above this combined threshold in the Vega chromosome six test set. A small fraction of functional genes in the Vega chromosome 6 test set had PSILC posterior-nuc score of 1. Scores are only plotted if greater than 0.5

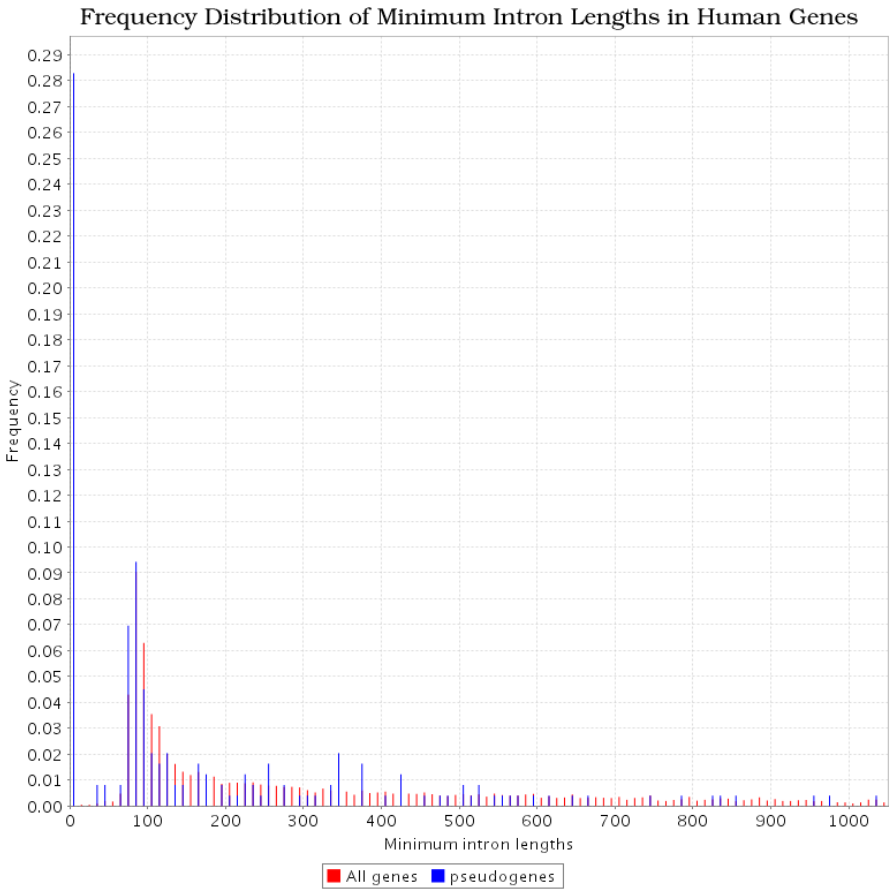

Figure 4.19: Frequency distribution of minimum intron lengths for multi-exon human pseudogene candidates as determined by a PSILC posterior nuc score of 1.0 (blue bars), versus all genes included in the study (red bars). Only intron lengths up to 1000 base-pairs are shown.

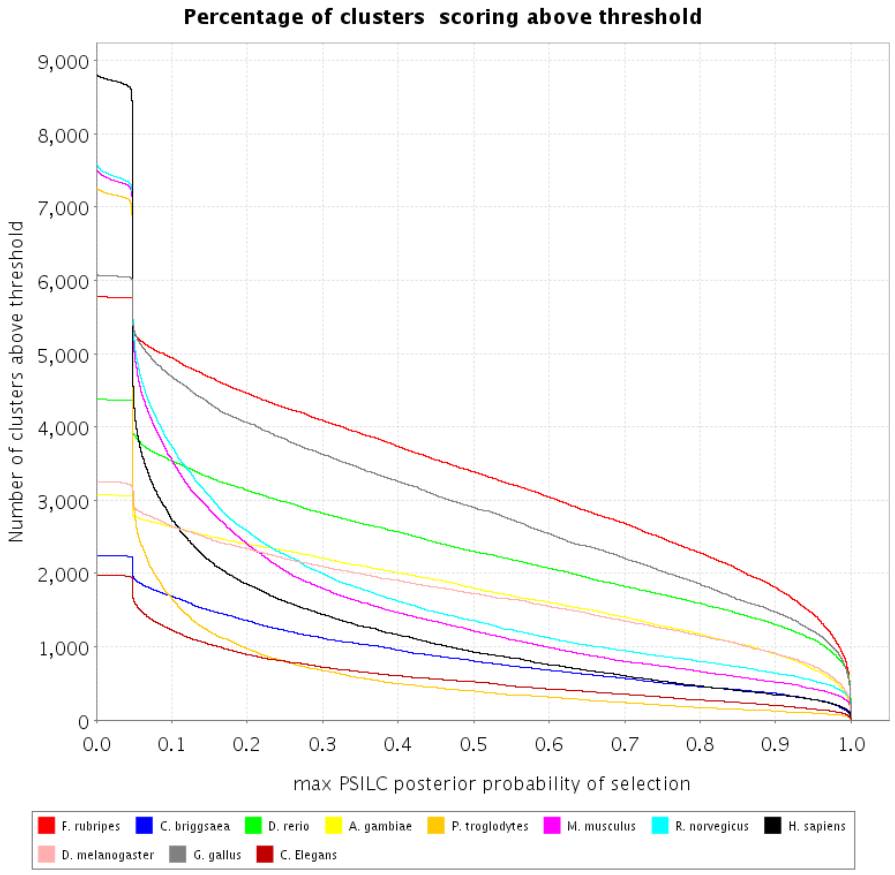

Figure 4.20: Number of clusters with a protein with maximum posterior probability of being under positive selection greater than a given max PSILC posterior-prot threshold in each of 11 species.

25% of clusters. Approximately 6% of clusters contain a human protein which is positively selected at a 75% max PSILC posterior-prot threshold, which falls to 3% at a 95% threshold.

PHIGS clusters were taken to be either weakly or strongly positively selected in a particular species if there was a protein in the cluster from that species which had a max PSILC posterior-prot score greater than either 75% or 95% respectively. For each species, the count of each Pfam domain occurring in positively selected clusters in that species was compared to the number expected by chance, and a p-value was calculated using the binomial distribution. The p-value represents the probability that the same or greater number of clusters with a particular Pfam domain would be obtained if the same number of positively selected clusters were drawn at random. To correct for the fact that multiple hypothesis are tested simultaneously, the 5% threshold for significance is divided by the number of Pfam

domains counted in at least one positively selected cluster. This cluster based approach to calculating significance avoids including protein domains purely on the basis of expansion and positive selection within a single cluster.

Table 4.6 displays the Pfam domains which are significantly over or under-represented for positively selected proteins in mammalian genomes. All domains which are statistically significant below 30% after the correction for multiple hypothesis testing are listed, and domains which are significant at 5% are displayed in bold. All of the domains detected as significantly over-represented are extracellular excluding the calponin homology CH domain but including Immunoglobulin (ig) superfamily domains, epidermal growth factor (EGF), 7 transmembrane receptor rhodopsin family (7tm_1), trypsin, and CUB. 45 of 464 immunoglobulin superfamily clusters have a selected human protein at 75% threshold, versus 19 expected clusters. Ig is also over-represented at the 95% threshold for selection, but not at a significant level. Immunoglobulin domains are found in proteins with a diverse set of functions, including antibodies and signalling proteins such as tyrosine kinases, both of which would be expected to be under positive selection. EGF repeats are commonly found in the extracellular region of membrane bound proteins. 7tm_1 proteins transduce extracellular signals, and include hormone, neurotransmitter and light receptors. CH is involved in signal transduction, and is also found in cytoskeletal proteins. Trypsin is a secreted proteolytic enzyme. CUB is an extracellular domain often occurring in developmentally regulated proteins, as well as in peptidases. CUB is the only domain which is significantly over-represented at the 95% threshold for selection. Two further domains which do not make the cut-off for significant over-representation are also extracellular domains: Laminin_G-like module and the scavenger receptor cysteine rich domain (SRCR) domain. WD40 repeats – found in proteins acting as transmembrane receptor signal transduction intermediaries – are significantly under-represented.

The 7 transmembrane receptor (secretin family) (7tm_2) and DUF887 are the only significantly overrepresented domain in positively selected chimpanzee clusters. The lack of success in finding chimpanzee proteins under positive selection may be due to the low quality of the current sequence. The list of over-represented mouse and rat domains is a similar to the human list, but excludes trypsin (although this is still over-represented), CUB and CH domains (both of which occur roughly at expected levels). 7tm_1 domains are particularly over-represented in rat, occurring in 79 versus 42 expected selected clusters. Somewhat

surprisingly, zf-C2H2 – a nucleic acid binding domain – is significantly over-represented in mouse, and over-represented (but not significantly) in rat. This repeat is under-represented (not significantly) in human positively selected protein clusters at both thresholds.

As well as ig, 7tm_1 and EGF, the pleckstrin homology (PH) domain, protein kinase superfamily and SRC homology-3 (SH3) domains are significantly over-represented in chicken, zebrafish and pufferfish. The PH domain occurs in proteins involved in intracellular signalling, as well as constituents of the cytoskeleton. SH3 domains are found in proteins involved in signal transduction related to cytoskeletal organisation. The PH and SH3 domain occurs at and less than, respectively, the level expected by chance in positively selected human clusters, whereas protein kinases are over-represented.

Protein kinase and ig domains are also over-represented in fruit-fly and mosquito clusters. No statistically significant over-representation was found in either nematode genomes, however the percentage of genes included in this study is less than a quarter of the full complement of nematode genes.

Hence it appears that extracellular,membrane bound and signalling proteins are particularly strong candidates for positive selection in several eukaryotic genomes. Positive selection is expected in families of paralogous proteins which bind peptide or protein ligands, as these proteins need to evolve specificity to different ligands after duplication, in order to mediate different responses to different inputs. The CUB and CH domains appear to be the only domains significantly over-represented in human selected proteins which is not over-represented in other selected proteins of other vertebrate genomes.

These results can be compared to other whole genome scans for positive selection. In [Cla03a], a scan of chimp and human genomes, using mouse as a reference genome, was carried out. These authors also discovered a strong positive selection signal in the human genome in G protein coupled receptor proteins, other protein receptors and extracellular matrix proteins. The strongest signal was discovered in olfactory proteins, which was also discovered using PSILC (data not shown). Other molecular functions also show a positive selection signal, including ion channel and transport proteins. Also corresponding to the results shown above, these authors found far fewer molecular functional categories in chimp under positive selection. The categories which were identified were chaperones, cell adhesion and extracellular matrix proteins. The authors identified amino acid metabolism as a biological process

showing significant positive selection in chimp, which might corroborate the positive selection signal discovered in the Gln-synt protein domain described above.

	tot.	sel. >75%	sel. >95%	exp. >75%	exp. >95%	sig. >75%	sig. >95%
H. sapiens	8882	529	276				
Immunoglobulin s.f.	**464**	**45**	**19**	**28**	**14**	**2.3e-13**	**1.4e-01**
EGF s.f.	**165**	**29**	**17**	**9.8**	**5.1**	**5.4e-07**	**2.7e-05**
7tm 1	**437**	**50**	**17**	**26**	**14**	**1.1e-05**	**2.0e-01**
Trypsin	**71**	**15**	**8**	**4.2**	**2.2**	**3.7e-05**	**2.0e-03**
CH	**28**	**9**	**3**	**1.7**	**0.87**	**6.2e-05**	**5.8e-02**
CUB	**35**	**9**	**8**	**2.1**	**1.1**	**3.2e-04**	**1.9e-05**
WD40*	188	1	1	11	5.8	1.5e-04	2.0e-02
SRCR	23	6	5	1.4	0.71	2.9e-03	8.6e-04
Laminin G-like module	40	8	6	2.4	1.2	3.2e-03	1.8e-03
P. troglodytes	7315	200	95				
7tm 2	**22**	**5**	**4**	**0.6**	**0.29**	**4.0e-04**	**2.2e-04**
Gln-synt C	2	2	0	0.055	0.026	1.4e-03	1.0e+00
Gln-synt N	2	2	0	0.055	0.026	1.4e-03	1.0e+00
DUF887	2	**2**	**2**	**0.055**	**0.026**	**1.4e-03**	**3.3e-04**
SAM PNT	8	3	0	0.22	0.1	1.5e-03	1.0e+00
Lipocalin	19	4	3	0.52	0.25	2.0e-03	2.1e-03
AMOP	3	2	2	0.082	0.039	3.2e-03	7.4e-04
M. musculus	7775	734	421				
7tm 1	**332**	**45**	**19**	**31**	**18**	**3.2e-08**	**4.4e-01**
zf-C2H2	**296**	**39**	**25**	**28**	**16**	**5.0e-06**	**2.1e-02**
Protein kinase C, C1 domain	24	10	7	2.3	1.3	1.3e-04	4.0e-04
Protein kinase s.f.	198	35	19	19	11	4.0e-04	1.3e-02
Immunoglobulin s.f.	414	48	25	39	22	6.6e-04	3.2e-01
Lectin C	48	13	10	4.5	2.6	8.6e-04	3.7e-04
PH	103	19	16	9.7	5.6	5.4e-03	2.3e-04
R. norvegicus	7836	867	536				
7tm 1	**378**	**79**	**33**	**42**	**26**	**0.0e+00**	**9.4e-02**
Immunoglobulin s.f.	**387**	**55**	**32**	**43**	**26**	**4.8e-06**	**2.2e-02**
EGF s.f.	137	31	17	15	9.4	2.0e-04	1.6e-02
Protein kinase s.f.	202	38	27	22	14	1.4e-03	9.2e-04
Laminin G-like module	30	10	9	3.3	2.1	2.3e-03	2.9e-04
DUF667	5	4	4	0.55	0.34	2.5e-03	4.3e-04

Table 4.2: Significantly over/under-represented Pfam domains in clusters with a positively selected human, chimp, mouse, rat proteins respectively. Results for 75% and 95% posterior probability thresholds are shown. Pfam domains which are significant at 5% after adjusting for testing multiple hypotheses are in bold. An asterix indicates under-representation.

	tot.	sel. >75%	sel. >95%	exp. >75%	exp. >95%	sig. >75%	sig. >95%
G. gallus	6122	2029	1207				
EGF s.f.	**141**	**80**	**52**	**47**	**28**	**0.0e+00**	**0.0e+00**
Immunoglobulin s.f.	**261**	**146**	**95**	**87**	**51**	**0.0e+00**	**0.0e+00**
WD40*	**156**	**32**	**20**	**52**	**31**	**7.1e-14**	**2.1e-06**
7tm 1	**127**	**61**	**29**	**42**	**25**	**7.9e-13**	**2.4e-01**
Protein kinase s.f.	**185**	**79**	**43**	**61**	**36**	**9.3e-10**	**1.2e-02**
PH	**91**	**43**	**31**	**30**	**18**	**2.0e-07**	**3.0e-03**
Src homology-3 domain	**142**	**54**	**41**	**47**	**28**	**1.1e-02**	**7.6e-08**
F. rubripes	5810	2478	1452				
Protein kinase s.f.	**171**	**100**	**65**	**73**	**43**	**0.0e+00**	**0.0e+00**
EGF s.f.	**114**	**84**	**55**	**49**	**28**	**0.0e+00**	**0.0e+00**
Immunoglobulin s.f.	**172**	**119**	**80**	**73**	**43**	**0.0e+00**	**0.0e+00**
PH	**84**	**58**	**40**	**36**	**21**	**0.0e+00**	**1.3e-04**
WD40*	**153**	**42**	**28**	**65**	**38**	**8.8e-17**	**1.7e-05**
Homeobox*	**91**	**22**	**19**	**39**	**23**	**6.8e-12**	**2.5e-01**
zf-C2H2*	**186**	**60**	**37**	**79**	**46**	**3.3e-11**	**1.3e-04**
Src homology-3 domain	**145**	**80**	**57**	**62**	**36**	**3.7e-10**	**2.2e-16**
fn3	**62**	**41**	**30**	**26**	**15**	**1.0e-09**	**6.5e-04**
Ank*	**105**	**32**	**14**	**45**	**26**	**3.5e-07**	**2.8e-08**
DEAD-like superfamily*	**69**	**20**	**9**	**29**	**17**	**2.4e-05**	**2.2e-02**
D. rerio	4438	1710	1098				
Protein kinase s.f.	**152**	**85**	**52**	**59**	**38**	**0.0e+00**	**2.4e-08**
Immunoglobulin s.f.	**144**	**79**	**59**	**55**	**36**	**0.0e+00**	**0.0e+00**
PH	**73**	**50**	**37**	**28**	**18**	**0.0e+00**	**5.2e-05**
7tm 1	**78**	**47**	**30**	**30**	**19**	**3.9e-12**	**1.4e-02**
WD40*	**119**	**31**	**19**	**46**	**29**	**4.6e-09**	**3.2e-06**
Ank*	**82**	**19**	**14**	**32**	**20**	**4.7e-08**	**9.2e-02**
Src homology-3 domain	**116**	**58**	**36**	**45**	**29**	**8.3e-07**	**3.1e-03**

Table 4.3: Significantly over/under-represented Pfam domains in clusters with a positively selected chicken, pufferfish and zebrafish proteins respectively. Results for 75% and 95% posterior probability thresholds are shown. Pfam domains which are significant at 5% after adjusting for testing multiple hypotheses are in bold. An asterix indicates under-representation.

	tot.	sel. >75%	sel. >95%	exp. >75%	exp. >95%	sig. >75%	sig. >95%
D. melanogaster	3303	1253	737				
WD40*	**125**	**32**	**14**	**47**	**28**	**1.5e-09**	**5.2e-10**
Protein kinase s.f.	**112**	**57**	**36**	**42**	**25**	**4.7e-08**	**2.0e-02**
Immunoglobulin s.f.	**34**	**27**	**26**	**13**	**7.6**	**3.7e-04**	**1.3e-07**
EGF s.f.	15	11	11	5.7	3.3	3.1e-02	7.1e-04
A. gambiae	3111	1285	692				
WD40*	**114**	**32**	**15**	**47**	**25**	**3.1e-09**	**1.7e-02**
Protein kinase s.f.	**100**	**56**	**33**	**41**	**22**	**2.6e-08**	**1.8e-02**
Immunoglobulin s.f.	**31**	**24**	**23**	**13**	**6.9**	**3.2e-03**	**1.1e-06**
C. Elegans	1995	311	147				
WHEP-TRS	5	4	4	0.78	0.37	8.3e-03	5.7e-04
Amidase	3	3	3	0.47	0.22	1.2e-02	1.5e-03

Table 4.4: Significantly over/under-represented Pfam domains in clusters with a positively selected fruit-fly, mosquito and nematode proteins respectively. Results for 75% and 95% posterior probability thresholds are shown. Pfam domains which are significant at 5% after adjusting for testing multiple hypotheses are in bold. An asterix indicates under-representation.

4.7 Discussion

I have demonstrated in this chapter that PSILC is a useful tool for identifying pseudogenes and positive selection. There are several potential shortcomings of the method. Firstly, PSILC relies heavily on having a good alignment. For example if a protein was conserved in a particular position but the alignment program did not align the conserved column properly, PSILC will incorrectly find evidence for either a pseudogene or positive selection. Identification of positive selection will be more prone to this sort of error than pseudogene identification, as several such errors would need to be present across the length of the gene for PSILC to infer pseudogene evolution incorrectly. This underlines the importance of accurate alignment programs, and I have endeavoured to minimize this problem by using the most accurate alignment programs available, such as MUSCLE and PROBCONS. One way to deal with this problem would be to calculate PSILC scores over many high likelihood alignments. However this is a very computationally expensive approach. PSILC also relies on having an accurate tree. For identifying genes and positive selection at external nodes, the main contribution to the

PSILC score will be from close neighbours. Thus, it is most important to have the topology close to the leaves correct, which is more easily achieved than deep internal branchings. Finally, PSILC relies on an accurate and representative protein domain HMM. Pfam HMMs are hand-curated and thus more reliable than automatically generated profile HMMs. However, as was evident in the study of APOBEC3G, there is not always an appropriate profile HMM in the database. PSILC automatically corrects if a poorly scoring HMM is included in the dataset, so that this problem usually leads to a loss of information regarding conserved sites, rather than incorrect inference of selection or pseudogenes.

One direction for further investigation is the development of significance values for PSILC scores for pseudogenes and positive selection. Significance of scores is currently gauged by reference to the small high-quality benchmark test set used – the Vega test set. It would be relatively straightforward to fit an extreme value distribution (provided this is the appropriate distribution) to scores of functional genes from this test set, and to use this to score significance of pseudogene hits. However, it is likely that proteins matching different HMMs have markedly different distributions of PSILC pseudogene scores, in much the same way that different HMMs have different log-odds score EVD parameters. If this is the case, then a more appropriate strategy may be to simulate evolution of functional proteins with a particular Pfam domain, and use the scores of these sets to parameterise a different distribution for each HMM. This second strategy may also be amenable for parameterizing an EVD for positive selection.

Another analysis for which PSILC would be useful is a large scale scan of genome segments not annotated as protein coding genes for pseudogenes, following [TSZB03] and [HMZ+03]. The approach here is to scan the genome for similarity to known coding regions in non-coding DNA, using – for example – BLASTX [GS93]. PSILC would then be used to confirm that the genome fragments found in this approach were genuinely evolving as neutral DNA.

PSILC could also prove useful in scoring non-synonymous coding SNPs for loss of function. This approach could also be applied to somatic mutations identified as part of the Cancer Genome project for impact on protein function, using data from the Catalogue of Somatic Mutations in Cancer (COSMIC) database [BDF+04]. In fact, it has already been shown that protein kinases are over-represented in somatically mutated genes which are implicated

in cancer [FCM$^+$04]. The protein kinases are also over-represented in the set of positively selected human PHIGS clusters in section 4.6, and hence it may be interesting to investigate the relationship – if any – between sites which are selected with sites which are implicated as oncogenic.

Chapter 5

Conclusion

There has been substantial progress made in recent years in describing patterns of evolution of protein domains [TPC98, AGT01b, TOT01, Ros02, VBK$^+$04]. Significant progress has also been made in developing models which describe the molecular evolution of proteins [GY94, Bru96, TGJ96, HR02, MLH04, LP04]. In this thesis, I have focussed on using this increased understanding of protein domain evolution to infer biologically important signals from sequence data. The first application in this work used information regarding species specific patterns of domain co-occurrence in order to infer the domain architecture of a protein from its sequence. The second application used information regarding patterns of substitution in conserved domain sites between closely related proteins to enhance the detection of protein domains in a cluster of homologous sequences. Whereas these two applications both look for patterns which have been conserved by evolution via purifying selection, the last application in this work looks for cases when this conservation has been lost in order to infer neutral evolution acting on a pseudogene as well as positive selection. I have demonstrated in each case that the extra sources of information can be used to improve inference, however the way in which the models are parameterised and trained is of critical importance. An over-trained or poorly parameterised model can substantially degrade inference.

Using observed patterns of occurrence in sequence data to infer biologically important signals is a common theme in bioinformatics, including amongst many other applications the detection of RNA genes from secondary structure folding potential and of protein coding genes from similarity to known gene structures. Thus, this work can be seen as part of a general approach to bioinformatics in which our understanding of a particular process is transformed

into a predictive probabilistic model, and this model is refined over time as our understanding of the process increases.

The first chapter was motivated by the observation that evolution has selected and preserved a restricted repertoire of patterns of domain co-occurrence. This is similar to the language modelling problem in speech recognition, and so language modelling techniques were used to incorporate information regarding patterns of domain occurrence into a framework for enhanced domain detection. A variable length Markov model was used to capture taxonomic-specific domain co-occurrence patterns. To avoid over-training, database counts of domain co-occurrence patterns were smoothed by recursively interpolating shorter domain contexts and higher-order taxa. The method resulted in a 2.2% improvement in the prediction of true-positive domain occurrences before the first false-positive at a family-by-family (i.e. non-aggregated) level. This improvement varied substantially by species, with the largest improvement in eukaryotes and a negligible improvement in virus protein annotation, which probably reflects the number of domains per protein as well as the flexibility of the repertoire of domain co-occurrence. This method is currently being used to extend the coverage of Pfam.

The motivation for the second chapter was an observation that some closely related proteins did not share the same annotated domain architecture in Pfam. In general, this was because the proteins were distantly related to the Pfam domain so although some of the proteins scored above threshold, most scored below the Pfam threshold. I investigated whether it was possible to take into account the pattern of substitution between closely related proteins in order to annotate a cluster of homologous proteins. This technique was found to be particularly sensitive to the way in which the site-specific evolutionary models were parameterised, and so several alternative parameterisations were investigated. The best performing of these resulted in a 67% improvement in detection of Pfam domains on an aggregated list of hits across multiple families, ranked by significance. On a family by family level the improvement was 5.3%. This method incorporated site-specific rate as well as residue frequency usage information. It has been observed that site-specific evolutionary models improve the likelihood of an alignment (after accounting for a penalty for increasing the number of free parameters), but as far as I am aware this is the first demonstration that site-specific evolutionary models can improve detection of protein domains.

In contrast to the first two chapters, the final chapter was concerned with identifying

cases of protein domain evolution where the observed conservation has been lost on a branch. The site-specific models of protein domain evolution from the previous chapter were used to describe the expected residue at a particular site if the protein was evolving to preserve the structure and function of the domain. An alternative model consisting of a composite protein domain model until the parent node and a neutral DNA model on the final branch was used to describe the residues expected if the protein was evolving as a pseudogene on this final branch. This technique for pseudogene identification was shown to be more successful at detecting pseudogenes on a human annotated test set of genes and pseudogenes than standard techniques based on the ratio of the rates of synonymous and non-synonymous substitution. The feasibility of integrating this technique into the ENSEMBL pipeline is currently under investigation. By identifying sites which appear to be evolving under a neutral protein model rather than a domain constrained model or a neutral DNA model of evolution, this method also predicts sites under positive selection. This approach was used to identify sites under positive selection in the APOBEC3 proteins, which have been implicated in the immune response against retroviruses, as well the HIV Virion Infectivity factor (Vif) and the Abalone lysin protein. The approach was also used in a global scan for positive selection which primarily identified several classes of extracellular proteins as under significant positive selection in vertebrates.

The methods outlined in this thesis calculate log-odds scores of a model of interest with respect to a background, or null model. While this is convenient for ranking matches relative to one another, it does not indicate the significance of a match. For protein domain identification, significance scores also enable the comparison of log-odds scores between different domain models. Calculating robust significance scores is one area in which further research is required for the techniques presented in this thesis. In the first two chapters, significance was calculating using the EVD parameterised for the standard profile HMM, without taking into account the extra information utilised in these chapters. This appears to be a satisfactory approximation for the language models, as the error vs significance curves are not substantially skewed by the language model scores. However the error vs significance curves for the phylogenetic HMM demonstrate that the EVD is not parameterised correctly for this score (see figure 3.6). Significance has not been calculated for the pseudogene or positive selection scores. An EVD could be fitted to the pseudogene scores of a large set of known functional

genes in order to calculate significance for pseudogene predictions. Positive selection, on the other hand, is somewhat harder to unambiguously prove or disprove, and so no such equivalent benchmark set for positive selection exists. Hence calculating robust significance scores for the positive selection test will require simulation.

This thesis has focussed on protein domain evolution, however similar techniques may be applicable for other conserved biological signals. Transcription factor binding sites and cis-regulatory modules may be amenable to some of the techniques presented here. Indeed in [BNP+02] the authors look for functional motifs on the basis of high local density as well as a sequence match score, which is related to language modelling. Moses et al. [MCP+04] use an evolutionary model similar to the phylogenetic profile HMM to identify conserved transcription factor binding sites. Gene prediction is another area in which these ideas might be applied, use of a phylogenetic HMM in this area has been explored by [MPJ03].

The techniques in Chapter 3 and 4 rely on parameterising a different evolutionary model at each match state for each profile HMM in Pfam. We might expect that there is really a much smaller *vocabulary* of evolutionary models which could account for the variation in each of these match states. It would be interesting to try discover this vocabulary of match states, and to build phylogenetic profile HMMs for each domain family in Pfam which restricted to using match states from this vocabulary. This would make calculating phylogenetic profile HMM scores for all of Pfam a feasible task, given that emission probabilities for each site would only have to be calculated for each match state in the vocabulary, rather than for each match state of each profile HMM. Moreover, this would lead to a robust definition of a null model as the mixture model of all of the states in the vocabulary.

The work presented in this thesis demonstrates the usefulness of modelling protein domain evolution in addressing core problems in bioinformatics such as homology detection, pseudogene detection and the detection of positive selection.

Bibliography

[AA01] J. O. Andersson and S. G. Andersson. Pseudogenes, junk, and the dynamics of rickettsia genomes. *Molecular Biology and Evolution*, 18(5):829–39., 2001.

[AB97] L. Arvestad and W.J. Bruno. Estimation of reversible substitution matrices from multiple pairs of sequences. *Journal of Molecular Evolution*, 45(6):696–703, 1997.

[ABW+04] R. Appweiler, A. Bairoch, C.H. Wu, Barker W.C., Boeckmann B., S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi, and L.S. Yeh. Uniprot: the universal protein knowledgebase. *Nucleic Acids Research*, 32:D115–D119, 2004.

[ACSR03] P. Aloy, H. Ceulemans, A. Stark, and R. B. Russell. The relationship between sequence and interaction divergence in proteins. *Journal of Molecular Biology*, 332(5):989–98, 2003.

[AGT01a] G. Apic, J. Gough, and S. A. Teichmann. An insight into domain combinations. *Bioinformatics*, 17(1):S83–S89, 2001.

[AGT01b] G. Apic, J. Gough, and S.A. Teichmann. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *Journal of Molecular Biology*, 310(2):311–25, 2001.

[AH96] J. Adachi and M. Hasegawa. Model of amino acid substitution in proteins encoded by mitochondrial DNA. *Journal of Molecular Evolution*, 42(4):459–68, 1996.

[AHB+04] A. Andreeva, D. Howorth, S. E. Brenner, T. J. Hubbard, C. Chothia, and A. G. Murzin. database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Research*, 32 Database issue:D226–9., 2004.

[AHT03] G. Apic, W. Huber, and S. A. Teichmann. Multi-domain protein families and domain pairs: comparison with known structures and a random model of domain recombination. *Journal of Structural and Functional Genomics*, 4(2-3):67–78, 2003.

[AMS+97] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.

[ASHS04] R. Aroul-Selvam, T. Hubbard, and R. Sasidharan. Domain insertions in protein structures. *Journal of Molecular Biology*, 338(4):633–41, 2004.

[AWMH00] J. Adachi, P.J. Waddell, W. Martin, and M. Hasegawa. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *Journal of Molecular Evolution*, 50(4):348–58, 2000.

[BA03] E. S. Balakirev and F. J. Ayala. Pseudogenes: are they "junk" or functional? *Annual Review of Genetics*, 37:123–51., 2003.

[BAB+04] E. Birney, D. Andrews, P. Bevan, M. Caccamo, G. Cameron, Y. Chen, L. Clarke, G. Coates, T. Cox, J. Cuff, V. Curwen, T. Cutts, T. Down, R. Durbin, E. Eyras, X. M. Fernandez-Suarez, P. Gane, B. Gibbins, J. Gilbert, M. Hammond, H. Hotz, V. Iyer, A. Kahari, K. Jekosch, A. Kasprzyk, D. Keefe, S. Keenan, H. Lehvaslaiho, G. McVicker, C. Melsopp, P. Meidl, E. Mongin, R. Pettett, S. Potter, G. Proctor, M. Rae, S. Searle, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, A. Ureta-Vidal, C. Woodwark, M. Clamp, and T. Hubbard. Ensembl 2004. *Nucleic Acids Research*, 32(1):D468–70., 2004.

[BBI+99] N. Bobola, P. Briata, C. Ilengo, N. Rosatto, C. Craft, G. Corte, and R. Ravazzolo. Otx2 homeodomain protein binds a element necessary for interphotore-

ceptor retinoid binding protein gene expression. *Mechanisms of Development*, 82(1-2):165–169., 1999.

[BBT03] G. J. Bartlett, N. Borkakoti, and J. M. Thornton. Catalysing new reactions during evolution: economy of residues and mechanism. *Journal of Molecular Biology*, 331(4):829–60, 2003.

[BC02] M. Bashton and C. Chothia. The geometry of domain combination in proteins. *Journal of Molecular Biology*, 315(4):927–39, 2002.

[BCD+04] A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, D. J. Studholme, C. Yeats, and S. R. Eddy. The pfam protein families database. *Nucleic Acids Research*, 32(1):D138–41., 2004.

[BCH98] S. E. Brenner, C. Chothia, and T. J. Hubbard. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proceedings of the National Academy of Sciences of the USA*, 95(11):6073–8., 1998.

[BD00] E. Birney and R. Durbin. Using Genewise in the Drosophila annotation experiment. *Genome Research*, 10:547–548, 2000.

[BDF+04] S. Bamford, E. Dawson, S. Forbes, J. Clements, R. Pettett, A. Dogan, A. Flanagan, J. Teague, P.A. Futreal, M.R. Stratton, and R. Wooster. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *British Journal of Cancer*, 91(2):355–8, 2004.

[BDWC04] H. P. Bogerd, B. P. Doehle, H. L. Wiegand, and B. R. Cullen. A single amino acid difference in the host APOBEC3G protein controls the primate species specificity of HIV type 1 virion infectivity factor. *Proceedings of the National Academy of Sciences of the USA*, 101(11):3770–4, 2004.

[BEW03] N. Bierne and A. Eyre-Walker. The problem of counting sites in the estimation of the synonymous and nonsynonymous substitution rates. implications for the

correlation between the synonymous substitution rate and codon usage bias. *Genetics*, 165(3):1587–97., 2003.

[BNP+02] B. P. Berman, Y. Nibu, B. D. Pfeiffer, P. Tomancak, S. E. Celniker, M. Levine, G. Rubin, and M. Eisen. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the drosophila genome. *Proceedings of the National Academy of Sciences of the USA*, 99(2):757–762, 2002.

[Bre00] S. E. Brenner. Target selection for structural genomics. *Nature Structural Biology*, 7 Suppl:967–9., 2000.

[Bru96] W.J. Bruno. Modeling residue usage in aligned protein sequences via maximum likelihood. *Molecular Biology and Evolution*, 13(10):1368–74, 1996.

[CAJ+94] Y. Cao, J. Adachi, A. Janke, S. Paabo, and M. Hasegawa. Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. *Journal of Molecular Evolution*, 39(5):519–27., 1994.

[CBD03] L. Coin, A. Bateman, and R. Durbin. Enhanced protein domain discovery by using language modeling techniques from speech recognition. *Proceedings of the National Academy of Sciences of the USA*, 100(8):4516–20, 2003.

[CBD04] L. Coin, A. Bateman, and R. Durbin. Enhanced protein domain discovery using taxonomy. *BMC Bioinformatics*, 5(1):56, 2004.

[CD04] L. Coin and R. Durbin. Improved techniques for the identification of pseudogenes. *Bioinformatics*, 20 Suppl 1:I94–I100, 2004.

[Cha93] E. Charniak. *Statistical Language Learning*. The MIT Press, Cambridge, Massachusetts, 1993.

[CHN03] S. G. Conticello, R. S. Harris, and M. S. Neuberger. The Vif protein of HIV triggers degradation of the human antiretroviral DNA deaminase APOBEC3G. *Current Biology*, 13(22):2009–13, 2003.

[Cho59] N. Chomsky. On certain formal properties of grammars. *Information and Control*, 2:137–167, 1959.

[CKM+03] D. Chivian, D. E. Kim, L. Malmstrm, P. Bradley, T. Robertson, P. Murphy, C. E. M. Strauss, R. Bonneau, C. A. Rohl, and D. Baker. Automated prediction of CASP-5 structures using the Robetta server. *Proteins*, 53 Suppl 6:524–33, 2003.

[Cla03a] A. G. Clark et al. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science*, 302(5652):1960–3, 2003.

[CLRS01] T.H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. The MIT Press, 2001.

[CTPMN04] S. G Conticello, C. J. F. Thomas, S. Petersen-Mahrt, and M. S. Neuberger. Evolution of the AID/APOBEC Family of Polynucleotide (Deoxy)Cytidine Deaminases. *Molecular Biology and Evolution*, 2004.

[CWC+02] J. M. Chandonia, N. S. Walker, Lo L. Conte, P. Koehl, M. Levitt, and S. E. Brenner. Astral compendium enhancements. *Nucleic Acids Research*, 30(1):260–3., 2002.

[CWN+97] S. Chen, Q. L. Wang, Z. Nie, H. Sun, G. Lennon, N. G. Copeland, D. J. Gilbert, N. A. Jenkins, and D. J. Zack. Crx, a novel otx-like paired-homeodomain protein, binds to and transactivates photoreceptor cell-specific genes. *Neuron*, 19(5):1017–1030., 1997.

[CWX+02] S. Chen, Q. L. Wang, S. Xu, I. Liu, L. Y. Li, Y. Wang, and D. J. Zack. Functional analysis of cone-rod homeobox) mutations associated with retinal dystrophy. *Human Molecular Genetics*, 11(8):873–884., 2002.

[DCB98] A. K. Das, P. W. Cohen, and D. Barford. The structure of the tetratricopeptide repeats of protein phosphatase 5: implications for-mediated protein-protein interactions. *EMBO Journal*, 17(5):1192–9., 1998.

[Deh] P. Dehal. Phylogenetically inferred groups. http://phigs.jgi-psf.org.

[DEKM98] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK, 1998.

[Den95] D.C. Dennett. *Darwin's dangerous idea*. The Penguin Press, 1995.

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B 39:1–38, 1977.

[DMBB] C. B. Do, M. S. P. Mahabhashyam, M. Brudno, and S. Batzoglou. Probcons: Probabilistic consistency-based multiple alignment of amino acid sequences. *Genome Research*. In press.

[DSO78] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt. A model of evolutionary change in proteins. In M. O. Dayhoff, editor, *Atlas of Protein Sequence and Structure*, volume 5, supplement 3, pages 345–352. National Biomedical Research Foundation, Washington, DC, 1978.

[DSW01] V. G. Durner, M. Scherf, and T. Werner. Experimental data of a single promoter can be used for in silico detection of genes with related regulation in the absence of sequence similarity. *Mammalian Genome*, 12:67–72, 2001.

[Edd98] S. R. Eddy. Profile-hidden Markov models. *Bioinformatics*, 14:755–763, 1998.

[Edd03] S. Eddy. Hmmer users guide, 2003.

[Edg04] R. C. Edgar. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5(1):113, 2004.

[EHP00] M. Evans, N. Hastings, and B. Peacock. *Statistical Distributions*. Wiley, 2000.

[EKO03] A. J. Enright, V. Kunin, and C. A. Ouzounis. Protein families and TRIBES in genome sequence space. *Nucleic Acids Research*, 31(15):4632–8, 2003.

[FC96] J. Felsenstein and G. A. Churchill. A hidden Markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution*, 13:93–104, 1996.

[FCM+04] P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton. A census of human cancer genes. *Nature Reviews Cancer*, 4(3):177–83, 2004.

[Fel81] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.

[FMC97] T. Furukawa, E. M. Morrow, and C. L. Cepko. Crx, a novel otx-like homeobox gene, shows photoreceptor-specific expression and regulates photoreceptor differentiation. *Cell*, 91(4):531–541., 1997.

[Ger98] M. Gerstein. How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Folding and Design*, 3(6):497–512, 1998.

[GG03] S. Guindon and O. Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5):696–704, 2003.

[GMB01] E.A. Gaucher, M.M. Miyamoto, and S.A. Benner. Function-structure analysis of proteins using covarion-based evolutionary approaches: Elongation factors. *Proceedings of the National Academy of Sciences of the USA*, 98(2):548–52, 2001.

[GS93] W. Gish and D.J. States. Identification of protein coding regions by database similarity search. *Nature Genetics*, 3(3):266–72, 1993.

[GSC94] M. Gerstein, E. L. L. Sonnhammer, and C. Chothia. Volume changes in protein evolution. *Journal of Molecular Biology*, 236:1067–1078, 1994.

[Gu01] X. Gu. Maximum-likelihood approach for gene family evolution under functional divergence. *Molecular Biology and Evolution*, 18(4):453–64, 2001.

[GW02] N. Goldman and S. Whelan. novel use of equilibrium frequencies in models of sequence evolution. *Molecular Biology and Evolution*, 19(11):1821–31., 2002.

[GY94] N. Goldman and Z. Yang. codon-based model of nucleotide substitution for protein-coding sequences. *Molecular Biology and Evolution*, 11(5):725–36., 1994.

[HB98] A.L. Halpern and W.J. Bruno. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Molecular Biology and Evolution*, 15(7):910–7, 1998.

[HB01] I. Holmes and W.J. Bruno. Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics*, 17(9):803–20, 2001.

[HD98] I. Holmes and R. Durbin. Dynamic programming alignment accuracy. *Journal of Computational Biology*, 5(3):493–504, 1998.

[HG01] H. Hegyi and M. Gerstein. Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Research*, 11(10):1632–40, 2001.

[HG02] P. M. Harrison and M. Gerstein. Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. *Journal of Molecular Biology*, 318(5):1155–74., 2002.

[HH92] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the USA*, 89:10915–10919, 1992.

[HH96] J. G. Henikoff and S. Henikoff. Using substitution probabilities to improve position-specific scoring matrices. *Computer Applications in the Biosciences*, 12:135–143, 1996.

[HHB+02] P. M. Harrison, H. Hegyi, S. Balasubramanian, N. M. Luscombe, P. Bertone, N. Echols, T. Johnson, and M. Gerstein. Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Research*, 12(2):272–80., 2002.

[HHP99] J.G. Henikoff, S. Henikoff, and S. Pietrokovski. New features of the Blocks Database servers. *Nucleic Acids Research*, 27(1):226–8, 1999.

[HK96] R. Hughey and A. Krogh. Hidden Markov models for sequence analysis: extension and analysis of the basic method. *Computer Applications in the Biosciences*, 12:95–107, 1996.

[HKY85] M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial. *Journal of Molecular Evolution*, 22(2):160–74., 1985.

[HMBC97] T. J. P. Hubbard, A. Murzin, S. Brenner, and C. Chothia. SCOP: a structural classification of proteins database. *Nucleic Acids Research*, 25:236–239, 1997.

[HMZ+03] P. M. Harrison, D. Milburn, Z. Zhang, P. Bertone, and M. Gerstein. Identification of pseudogenes in the drosophila melanogaster genome. *Nucleic Acids Research*, 31(3):1033–7., 2003.

[HN88] A.L. Hughes and M. Nei. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, 335(6186):167–70, 1988.

[Hol03] I. Holmes. Using guide trees to construct multiple-sequence evolutionary HMMs. *Bioinformatics*, 19 Suppl 1:i147–57, 2003.

[HR02] I. Holmes and G.M. Rubin. An expectation maximization algorithm for training hidden substitution models. *Journal of Molecular Biology*, 317(5):753–64, 2002.

[HSGMS02] L. Q. Ha, E. I. Sicilia-Garcia, J. Ming, and F. J. Smith. Extension of zipf's law to words and phrases. In R. Altman, D. Brutlag, P. Karp, R. Lathrop, and D. Searls, editors, *Coling proceedings*, pages 315–320, Taipei, Taiwan, 2002.

[HW01] D. Husmeier and F. Wright. Detection of recombination in DNA multiple alignments with hidden Markov models. *Journal of Computational Biology*, 8(4):401–27, 2001.

[HYC+03] S. Hirotsune, N. Yoshida, A. Chen, L. Garrett, F. Sugiyama, S. Takahashi, K. Yagami, A. Wynshaw-Boris, and A. Yoshiki. An expressed pseudogene regulates the messenger stability of its homologous coding gene. *Nature*, 423(6935):91–6., 2003.

[Jay03] E. Jaynes. *Probability Theory. The logic of science.* Cambridge University Press, Cambridge, U.K., 2003.

[JCB+02] A. Jarmuz, A. Chester, J. Bayliss, J. Gisbourne, I. Dunham, J. Scott, and N. Navaratnam. An anthropoid-specific locus of orphan C to U RNA-editing enzymes on chromosome 22. *Genomics*, 79(3):285–96, 2002.

[JDH00] T. Jaakkola, M. Diekhans, and D. Haussler. A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*, 7:95–114, 2000.

[Jel97] F. Jelinek. *Statistical Methods For Speech Recognition*. The MIT Press, Cambridge, Massachusetts, 1997.

[JKA+] I. K. Jordan, F. A. Kondrashov, I. A. Adzhubei, Y. I. Wolf, A. S. Kondrashov, S. Sunyaev, and E. V. Koonin. A universal trend of amino acid gain and loss in protein evolution: the modern echo of code origin. To be published.

[JTT92] D. T. Jones, W. R. Taylor, and J. M. Thornton. The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences*, 8(3):275–82., 1992.

[JY99] A. Joshi and Y.Schabes. Tree-adjoining grammars. Technical report, University of Pennsylvania, 1999. http://www.cis.upenn.edu/ joshi/.

[KBM+94] A. Krogh, M. Brown, I. S. Mian, K. Sjölander, and D. Haussler. Hidden Markov models in computational biology: applications to protein modeling. *Journal of Molecular Biology*, 235:1501–1531, 1994.

[Kim83] M. Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK, 1983.

[KM95] A. Krogh and G. Mitchison. Maximum entropy weighting of aligned sequences of proteins or. In 1995/01/01, editor, *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, pages 215–21., Laboratory of Molecular Biology, Cambridge, England., 1995. AAAI Press.

[KM01] B. Knudsen and M.M. Miyamoto. A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. *Proceedings of the National Academy of Sciences of the USA*, 98(25):14512–7, 2001.

[KZNL02] H. Kaessmann, S. Zllner, A. Nekrutenko, and W-H. Li. Signatures of domain shuffling in the human genome. *Genome Research*, 12(11):1642–50, 2002.

[Lay94] D. C. Lay. *Linear algebra and its applications*. Addison Wesley, 1994.

[LD43] S.E. Luria and M. Delbrück. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, 28(6):491–511, 1943.

[LEC+04] C. S. Leslie, E. Eskin, A. Cohen, J. Weston, and W. S. Noble. Mismatch string kernels for discriminative protein classification. *Bioinformatics*, 20(4):467–76, 2004.

[LG99] P. Lio and N. Goldman. Using protein structural information in evolutionary inference: transmembrane proteins. *Molecular Biology and Evolution*, 16(12):1696–710., 1999.

[LGTJ98] P. Lio, N. Goldman, J. L. Thorne, and D. T. Jones. : PASSML: combining evolutionary inference and protein secondary structure prediction. *Bioinformatics*, 14(8):726–33., 1998.

[LJN+96] L. Lavaissiere, S. Jia, M. Nishiyama, de la S. Monte, A. M. Stern, J. R. Wands, and P. A. Friedman. Overexpression of human aspartyl(asparaginyl)beta-hydroxylase in hepatocellular carcinoma and cholangiocarcinoma. *Journal of Clinical Investigation*, 98(6):1313–23., 1996.

[LOV95] Y.H. Lee, T. Ota, and V.D. Vacquier. Positive selection is a general phenomenon in the evolution of abalone sperm lysin. *Molecular Biology and Evolution*, 12(2):231–8, 1995.

[LP04] N. Lartillot and H. Philippe. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21(6):1095–109, 2004.

[LPA+93] H. Lutcke, S. Prehn, A. J. Ashford, M. Remus, R. Frank, and B. Dobberstein. Assembly of the 68- and 72-kd proteins of signal recognition particle with 7s. *Journal of Cell Biology*, 121(5):977–985., 1993.

[LPR01] A.N. Lupas, C.P. Ponting, and R.B. Russell. On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *Journal of Structural Biology*, 134(2-3):191–203, 2001.

[LPSS84] C. Lanave, G. Preparata, C. Saccone, and G. Serio. A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, 20(1):86–93, 1984.

[LQZ+02] N. M. Luscombe, J. Qian, Z. Zhang, T. Johnson, and M. Gerstein. The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. *Genome Biology*, 3(8):RESEARCH0040, 2002.

[MCP+04] A. M. Moses, D. Y. Chiang, D. A. Pollard, V. N. Iyer, and M. B. Eisen. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biology*, 5(12):R98, 2004.

[MD95] G. J. Mitchison and R. Durbin. Tree-based maximal likelihood substitution matrices and hidden Markov models. *Journal of Molecular Evolution*, 41:1139–1151, 1995.

[MG94] S. V. Muse and B. S. Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*, 11(5):715–24., 1994.

[MG02] M. Madera and J. Gough. Comparison of profile hidden markov model procedures for remote homology detection. *Nucleic Acids Research*, 30(19):4321–8., 2002.

[Mig00] A.J. Mighell et al. Vertebrate pseudogenes. *FEBS Letters*, 468:109–14., 2000.

[MLH04] I. Miklós, G.A. Lunter, and I. Holmes. A "Long Indel" model for evolutionary sequence alignment. *Molecular Biology and Evolution*, 21(3):529–40, 2004.

[Mou02] A. Mounsey et al. Evidence suggesting that a fifth of annotated caenorhabditis elegans genes may be pseudogenes. *Genome Research*, 12:770–75., 2002.

[MPJ03] J.D McAuliffe, L. Pachter, and M.I. Jordan. Multiple-sequence functional annotation and the generalized hidden Markov phylogeny. Technical report, University of California Berkley, 2003.

[MSBP02] R. Mott, J. Schultz, P. Bork, and C.P. Ponting. Predicting cellular localization using a domain projection method. *Genome Research*, 8:1168–1174, 2002.

[Mun03] A. J. Mungall et al. The sequence and analysis of human chromosome 6. *Nature*, 425(6960):805–11., 2003.

[NG86] M. Nei and T. Gojobori. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, 3(5):418–26., 1986.

[Nor97] J.R. Norris. *Markov Chains*. Cambridge University Press, 1997.

[NY98] R. Nielsen and Z. Yang. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, 148(3):929–36, 1998.

[PFL+01] P. Pavlidis, T. S. Furey, M. Liberto, D. Haussler, and W. N. Grundy. Promoter region based classification of genes. *Proceedings of the Pacific Symposium on Biocomputing*, pages 151–163, 2001.

[Pie96] S. Pietrokovski. Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Research*, 24(19):3836–45, 1996.

[PKB+98] J. Park, K. Karplus, C. Barrett, R. Hughey, D. Haussler, T. Hubbard, and C. Chothia. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *Journal of Molecular Biology*, 284(4):1201–10., 1998.

[PMHN02] S. K. Petersen-Mahrt, R. S. Harris, and M. S. Neuberger. AID mutates E. coli suggesting a DNA deamination mechanism for antibody diversification. *Nature*, 418(6893):99–103, 2002.

[PP02] V. E. Prince and F. B. Pickett. Splitting pairs: the diverging fates of duplicated genes. *Nature Reviews Genetics*, 3(11):827–37, 2002.

[PTTF92] W. H. Press, S. A. Teukolsky, W. Vetterling T., and B. P. Flannery. *Numerical Recipes in C*. Cambridge University Press, Cambridge, UK, 1992.

[QG03] B. Qian and R. A. Goldstein. Detecting distant homologs using phylogenetic tree-based HMMs. *Proteins*, 52(3):446–53., 2003.

[QLG01] J. Qian, N.M. Luscombe, and M. Gerstein. Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *Journal of Molecular Biology*, 313(4):673–81, 2001.

[RBD01] C. Rivolta, E. L. Berson, and T. P. Dryja. Dominant leber congenital amaurosis, cone-rod degeneration, and retinitis pigmentosa caused by mutant versions of the transcription factor. *Human Mutation*, 18(6):488–498., 2001.

[RCZ01] R. Rosenfeld, S.F. Chen, and X. Zhu. Whole-sentence exponential language models: a vehicle for linguistic statistical integration. *Computer Speech and Language*, 15(1), 2001.

[RJ93] L. R. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, Englewood Cliffs, NJ, 1993.

[RJK+03] D. M. Robinson, D. T. Jones, H. Kishino, N. Goldman, and J. L. Thorne. Protein evolution with dependence among codons due to tertiary structure. *Molecular Biology and Evolution*, 20(10):1692–704, 2003.

[Ros02] B. Rost. Did evolution leap to create the protein universe? *Current Opinion in Structural Biology*, 12(3):409–16, 2002.

[SA90] P. R. Sibbald and P. Argos. Weighting aligned protein or nucleic acid sequences to correct for unequal representation. *Journal of Molecular Biology*, 216:813–818, 1990.

[SBF+00] A. Scaloni, C. Bottiglieri, L. Ferrara, M. Corona, G. B. Gurrola, C. Batista, E. Wanke, and L. D. Possani. Disulfide bridges of ergtoxin, a member of a new sub-family of peptide blockers of the ether-a-go-go-related K+ channel. *FEBS Letters*, 479(3):156–7., 2000.

[SBG03] R. I. Sadreyev, D. Baker, and N. V. Grishin. Profile-profile comparisons by COMPASS predict intricate homologies between protein families. *Protein Science*, 12(10):2262–72, 2003.

[SCB+00] M. Salzet, V. Chopin, J. Baert, I. Matias, and J. Malecha. Theromin, a novel leech thrombin inhibitor. *Journal of Biological Chemistry*, 275(40):30774–80., 2000.

[SCW+97] P. K. Swain, S. Chen, Q. L. Wang, L. M. Affatigato, C. L. Coats, K. D. Brady, G. A. Fishman, S. G. Jacobson, A. Swaroop, E. Stone, P. A. Sieving, and D. J. Zack. Mutations in the cone-rod homeobox gene are associated with the cone-rod dystrophy photoreceptor degeneration. *Neuron*, 19(6):1329–1336., 1997.

[SEM04] S. L. Sawyer, M. Emerman, and H. S. Malik. Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. *PLoS Biology*, 2(9):E275, 2004.

[SG99] Y. Suzuki and T. Gojobori. A method for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution*, 16(10):1315–28., 1999.

[SGCM02] A. M. Sheehy, N. C. Gaddis, J. D. Choi, and M. H. Malim. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature*, 418(6898):646–50, 2002.

[SH04] A. Siepel and D. Haussler. Combining phylogenetic and hidden Markov models in biosequence analysis. *Journal of Computational Biology*, 11(2-3):413–28, 2004.

[SKB+96] K. Sjölander, 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. *Computer Applications in the Biosciences*, 12(4):327–345, 1996.

[SL03] J. Sding and A. N. Lupas. More than the sum of their parts: on the evolution of proteins from peptides. *Bioessays*, 25(9):837–46, 2003.

[SPD+99] S. L. Salzberg, M. Pertea, A. L. Delcher, M. J. Gardner, and H. Tettelin. Interpolated markov models for eukaryotic gene finding. *Genomics*, 59(1):24–31., 1999.

[SV95] W. J. Swanson and V. D. Vacquier. Extraordinary divergence and positive darwinian selection in a fusagenic protein coating the acrosomal process of abalone spermatozoa. *Proceedings of the National Academy of Sciences of the USA*, 92(11):4957–61., 1995.

[Sd04] J. Sding. Protein homology detection by HMM-HMM comparison. *Bioinformatics*, 2004.

[TBD93] B. Teng, C.F. Burant, and N.O. Davidson. Molecular cloning of an apolipoprotein B messenger RNA editing protein. *Science*, 260(5115):1816–9, 1993.

[TGJ96] J. L. Thorne, N. Goldman, and D. T. Jones. Combining protein evolution and secondary structure. *Molecular Biology and Evolution*, 13:666–673, 1996.

[TKF91] J. L. Thorne, H. Kishino, and J. Felsenstein. An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution*, 33:114–124, 1991.

[TKF92] J. L. Thorne, H. Kishino, and J. Felsenstein. Inching toward reality: an improved likelihood model of sequence evolution. *Methods in Enzymology*, 34:3–16, 1992.

[TN93] K. Tamura and M. Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, 10(3):512–26, 1993.

[TOT01] A.E. Todd, C.A. Orengo, and J.M. Thornton. Evolution of function in protein superfamilies, from a structural perspective. *Journal of Molecular Biology*, 307(4):1113–43, 2001.

[TPC98] S.A. Teichmann, J. Park, and C. Chothia. Structural assignments to the Mycoplasma genitalium proteins show extensive gene duplications and domain rearrangements. *Proceedings of the National Academy of Sciences of the USA*, 95(25):14658–63, 1998.

[TSZB03] D. Torrents, M. Suyama, E. Zdobnov, and P. Bork. A genome-wide survey of human pseudogenes. *Genome Research*, 13(12):2559–67., 2003.

[VBB+04] C. Vogel, C. Berzuini, M. Bashton, J. Gough, and S. A. Teichmann. Supra-domains: evolutionary units larger than single protein domains. *Journal of Molecular Biology*, 336(3):809–23, 2004.

[VBK+04] C. Vogel, M. Bashton, N. D. Kerrison, C. Chothia, and S. A. Teichmann. Structure, function and evolution of multidomain proteins. *Current Opinion in Structural Biology*, 14(2):208–16, 2004.

[Ven04] J. C. Venter et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304(5667):66–74, 2004.

[WG01] S. Whelan and N. Goldman. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution*, 18(5):691–9., 2001.

[WS04] M. Wistrand and E. L. L. Sonnhammer. Improving profile HMM discrimination by adapting transition probabilities. *Journal of Molecular Biology*, 338(4):847–54, 2004.

[WYGN04] W. S. W. Wong, Z. Yang, N. Goldman, and R. Nielsen. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics*, 168(2):1041–51, 2004.

[Yan93] Z. Yang. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution*, 10:1396–1401, 1993.

[Yan95] Z. Yang. A space-time process model for the evolution of dna sequences. *Genetics*, 139:993–1005, 1995.

[Yano04] Y. Yano et al. A new role for expressed pseudogenes as ncRNA: regulation of mRNA stability of its homologous coding gene. *Journal of Molecular Medicine*, 82(7):414–22, 2004.

[YL02] G. Yona and M. Levitt. Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *Journal of Molecular Biology*, 315(5):1257–75, 2002.

[YN98] Z. Yang and R. Nielsen. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of Molecular Evolution*, 46(4):409–18., 1998.

[YN00] Z. Yang and R. Nielsen. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution*, 17(1):32–43., 2000.

[YN02] Z. Yang and R. Nielsen. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution*, 19(6):908–17, 2002.

[YNGP00] Z. Yang, R. Nielsen, N. Goldman, and A.M. Pedersen. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155(1):431–49, 2000.

[YSV00] Z. Yang, W. J. Swanson, and V.D. Vacquier. Maximum-likelihood analysis of molecular adaption in abalone sperm lysin reveals variable selective pressures among lineages and sites. *Molecular Biology and Evolution*, 2000.

[Zha04] J. Zhang. Frequent false detection of positive selection by the likelihood method with branch-site models. *Molecular Biology and Evolution*, 21(7):1332–9, 2004.

[Zip35] G.K. Zipf. *Psycho-Biology of Languages*. Houghton-Mifflin, 1935.

[ZYP+03] H. Zhang, B. Yang, R. J. Pomerantz, C. Zhang, S. C. Arunachalam, and L. Gao. The cytidine deaminase CEM15 induces hypermutation in newly synthesized HIV-1 DNA. *Nature*, 424(6944):94–8, 2003.