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 $P(\psi_i = M_{l,j}|x_{k,\cdot})$

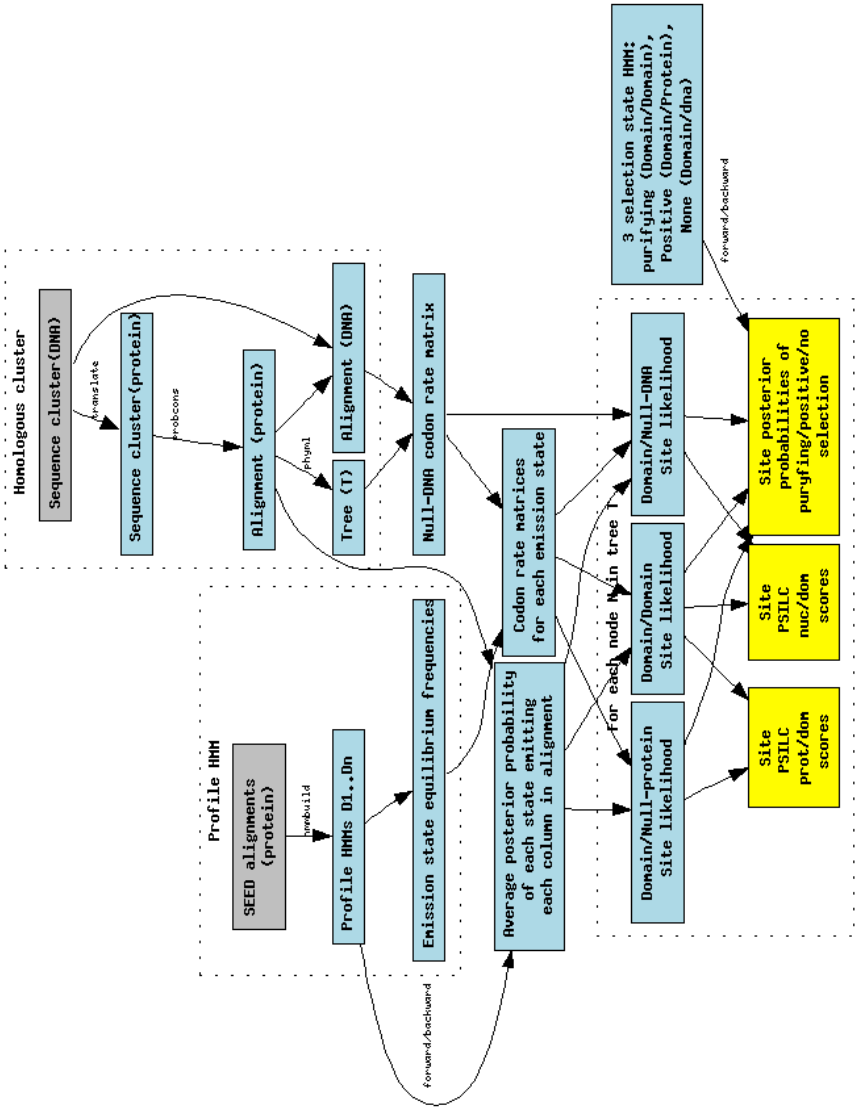

Figure 4.1: Conceptual diagram PSILC. Inputs are shown in grey, and outputs in yellow, with intermediate steps in light-blue.

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

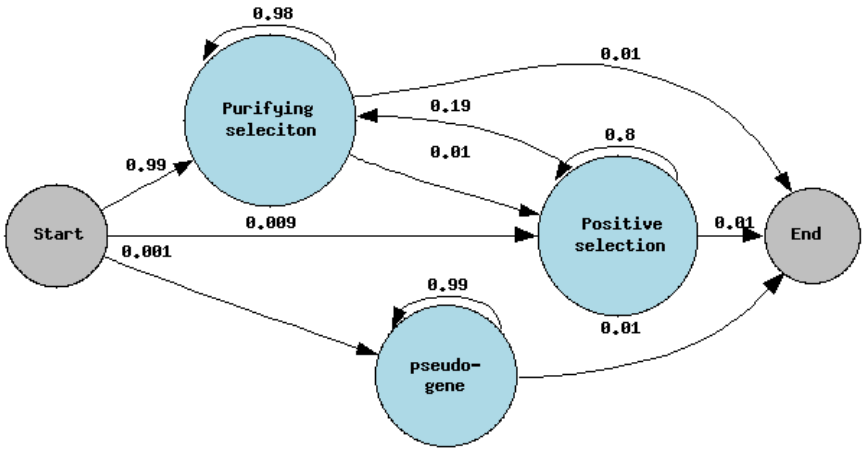

Figure 4.2: Diagram of selection HMM, comprising 3 states - purifying selection ($\mathcal{E}_{\text{dom,dom}}$), pseudogene ($\mathcal{E}_{\text{nuc,dom}}$), positive selection ($\mathcal{E}_{\text{prot,dom}}$). The transitions are given as an example only, and can be specified by the user.

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

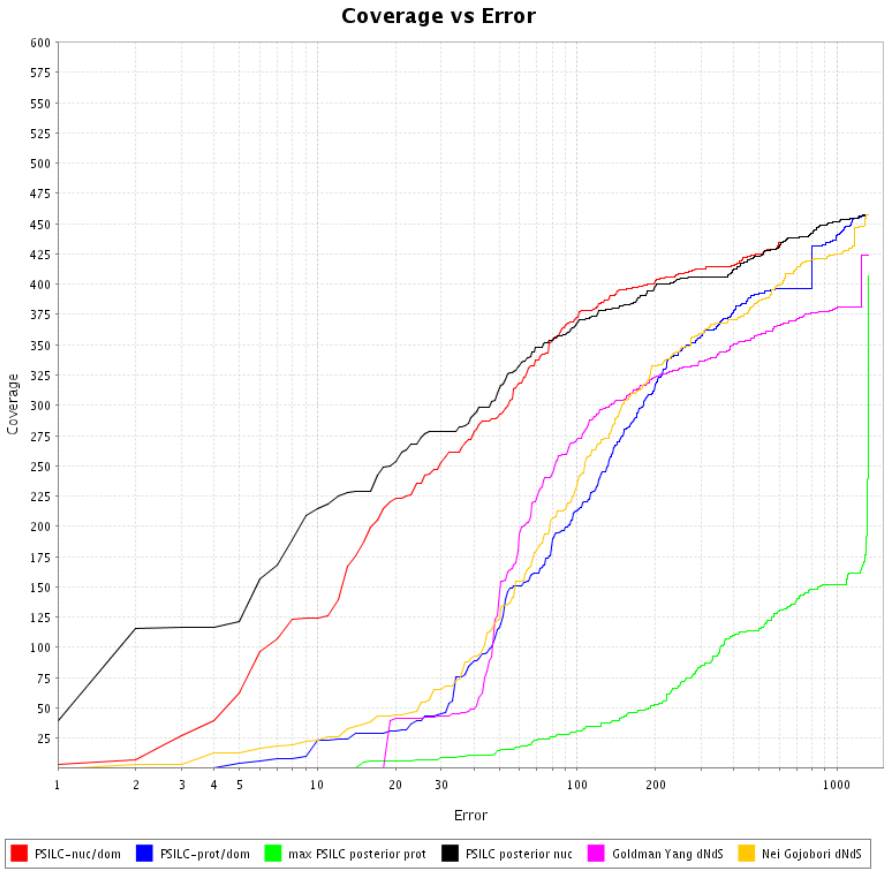

Figure 4.3: Coverage vs error curve for PSILC and dN/dS. The graph has been plotted on a log x-axis to reflect the fact that coverage level at low error rate is more important than at a high error rate. Several (pseudo)genes had a PSILC posterior nuc score of 1.0 - (pseudo)genes with the same PSILC posterior 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.

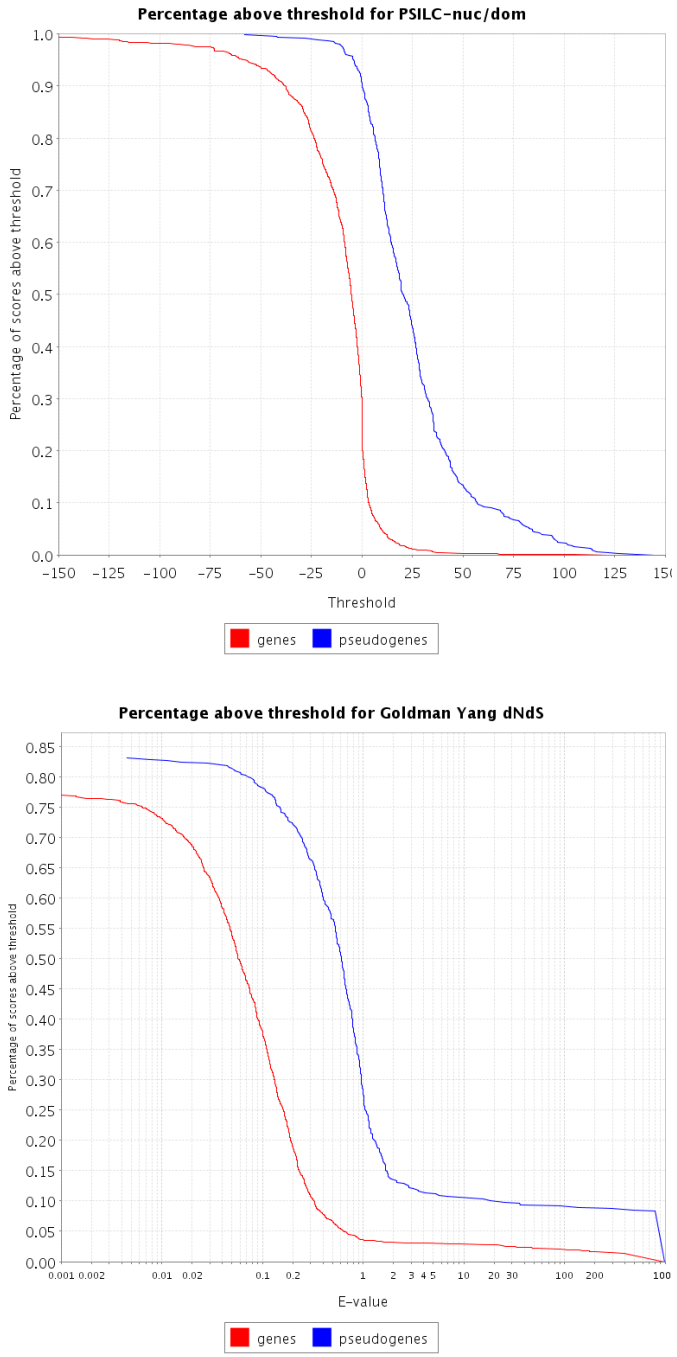

Figure 4.4: Comparison of discrimination between pseudogenes and functional genes between the PSILC-nuc/dom method (top graph) and Goldman Yang dN/dS (lower graph). In both graphs I plot the fraction of (pseudo)genes scoring above a particular threshold, with the pseudogenes represented by the blue line,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

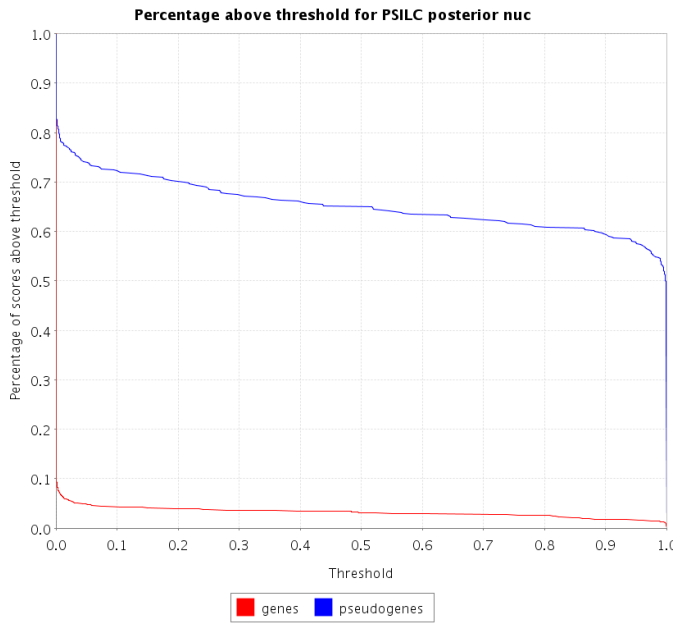

Figure 4.5: Comparison of discrimination between pseudogenes and functional genes using the PSILC posterior-nuc score.

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-

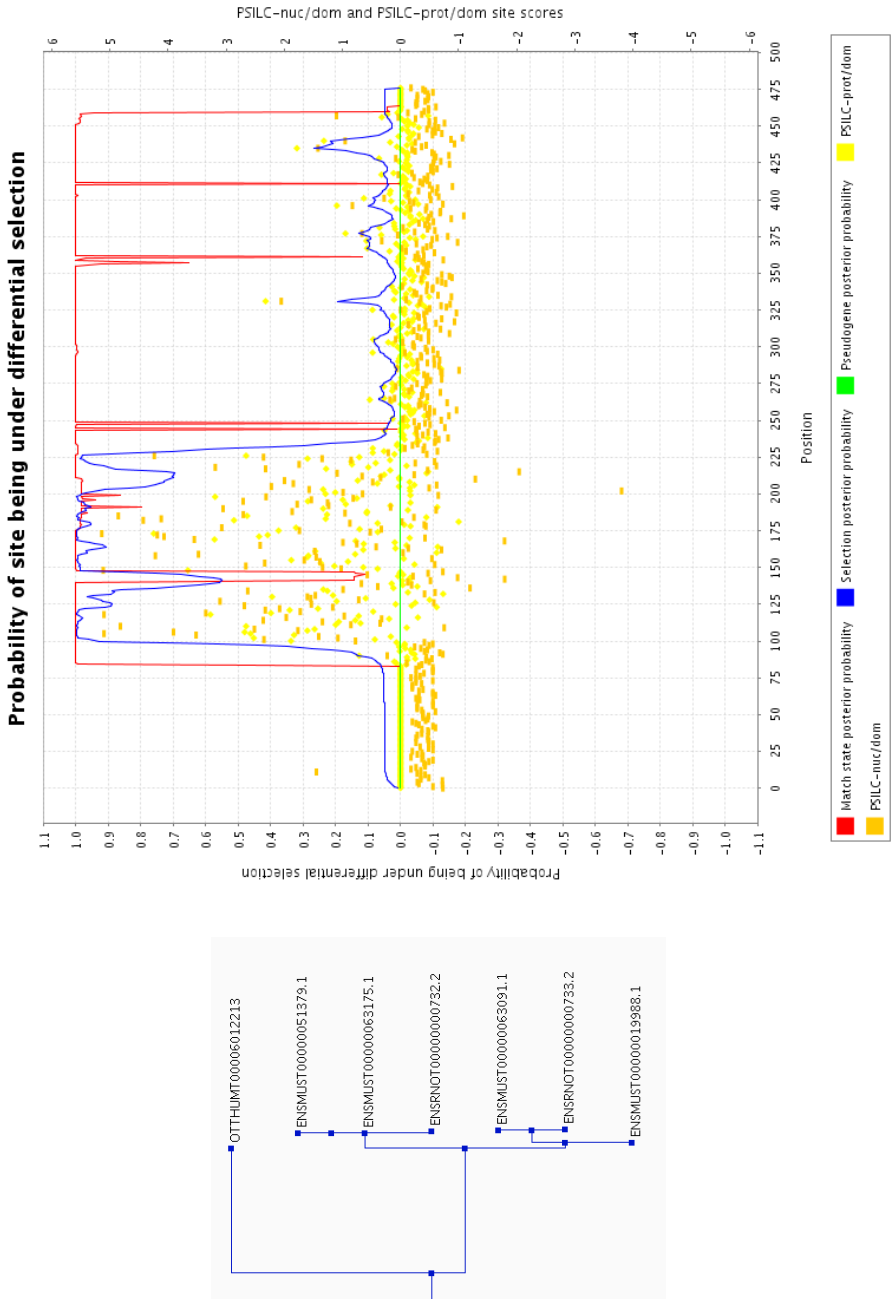

Figure 4.6: Pseudogene and selection status of Vega human (functional) gene OTTHUMT00006012213. Left: the Phyml tree of OTTHUMT00006012213 and homologues in mouse, human and rat genomes. Right: plot of PSILC nuc-dom(orange) and PSILC prot-dom(yellow) scores; Pfam domain match probability (to Pkinase,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

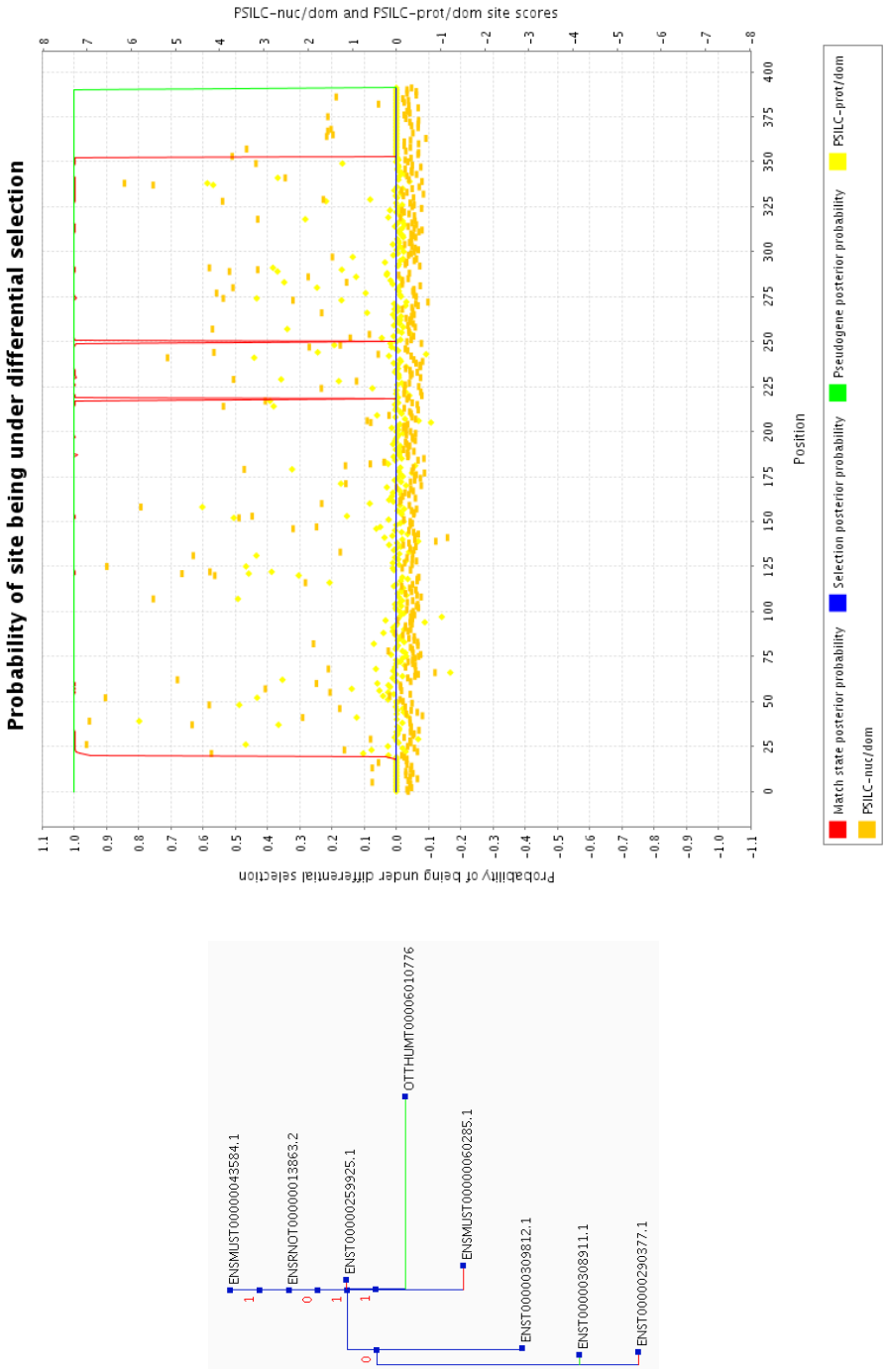

Figure 4.7: Pseudogene and selection status of Vega human pseudo-gene OTTHUMT00006009362. Left: the PhymI tree of OTTHUMT00006009362 and homologues in mouse, human and rat genomes. Right: plot of PSILC nuc-dom(orange) and PSILC prot-dom(yellow) scores; Pfam domain match probability (to Pkinase, SH3, 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

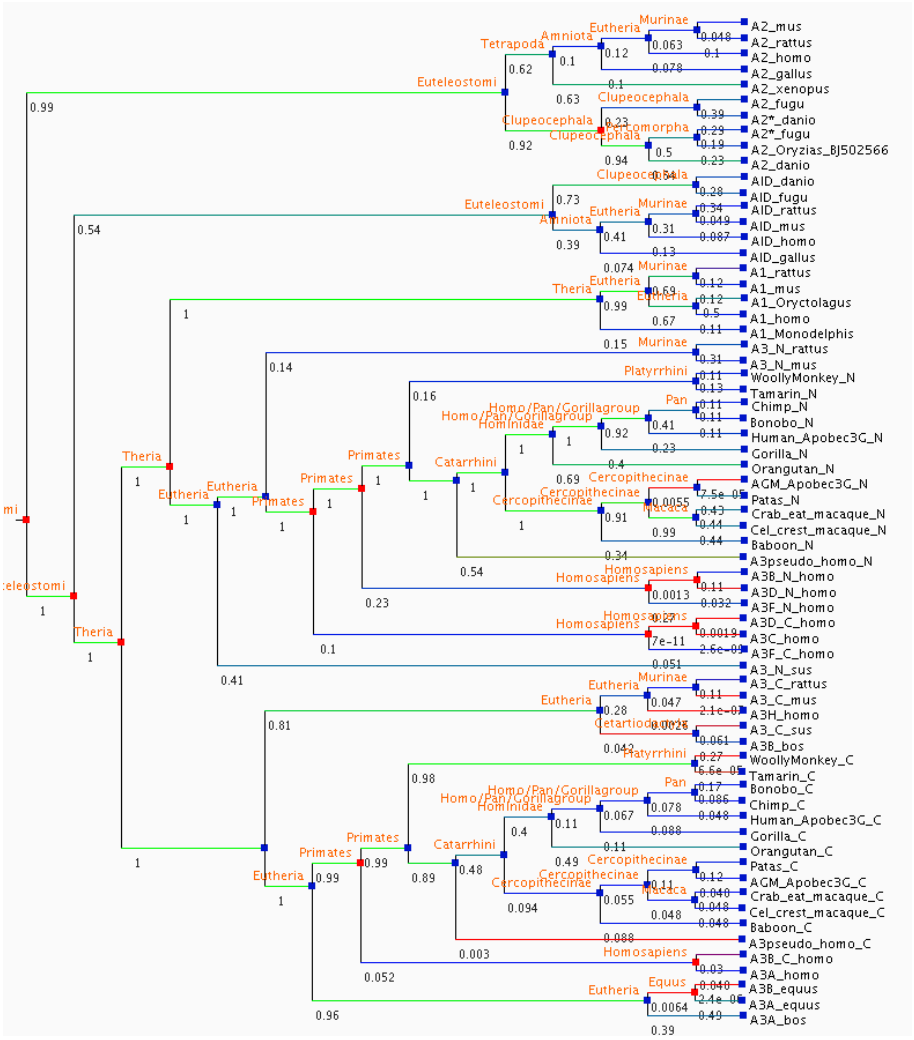

Figure 4.8: Tree of APOBEC/AID family, showing extensive positive selection. Green branches to a node indicate strong evidence for positive selection below this node, whereas red branches indicate strong evidence for pseudogene evolution below a given node. Blue branches indicate lack of evidence for 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.

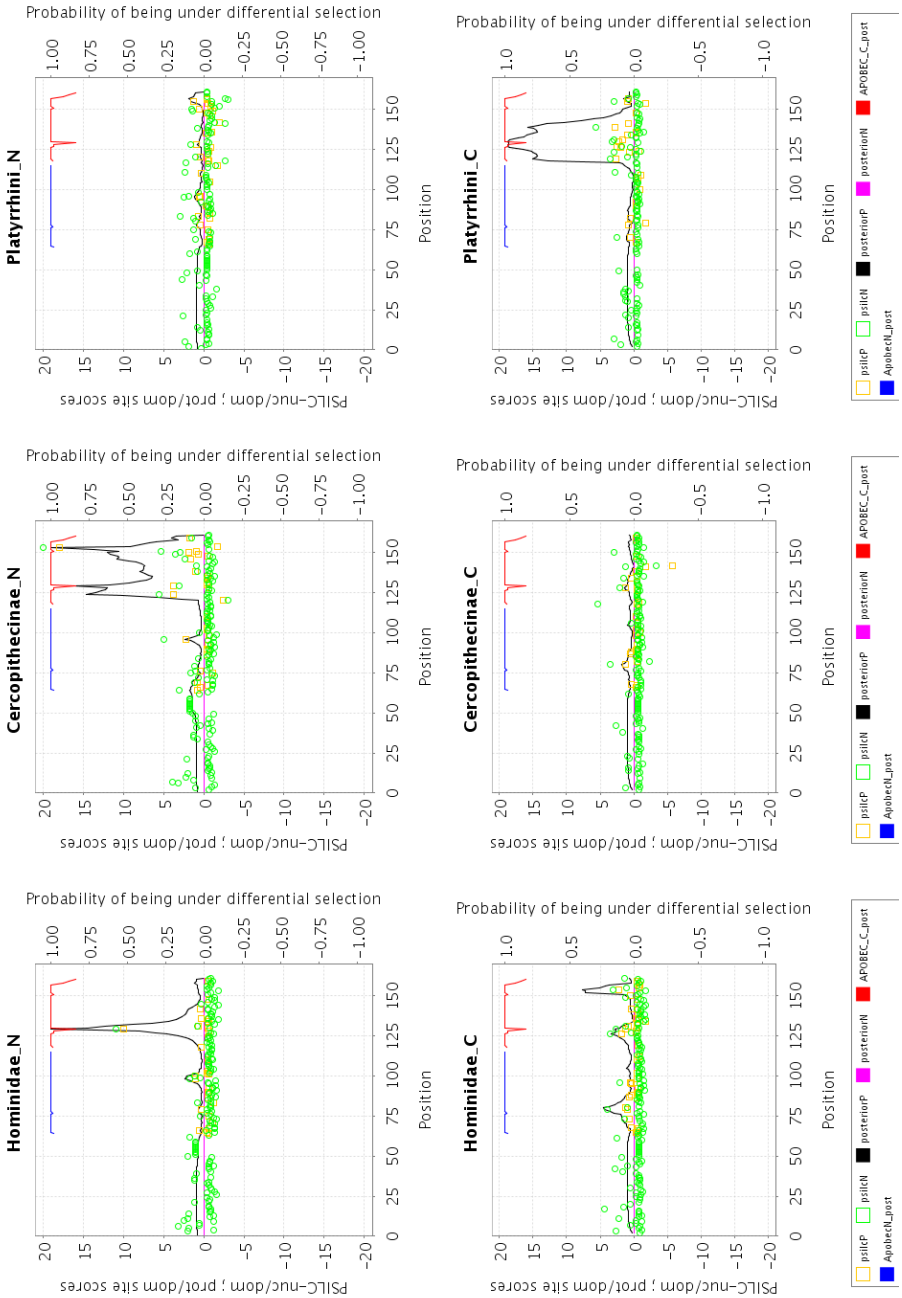

Figure 4.9: Site-specific graphs of selection acting on APOBEC3G genes. Green circles/orange squares indicate the site specific PSILC-nuc/dom and PSILC-prot/dom scores respectively, which, for clarity, are only plotted if less than -0.3 or greater than 0.3. The black/purple line indicates the posterior 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.

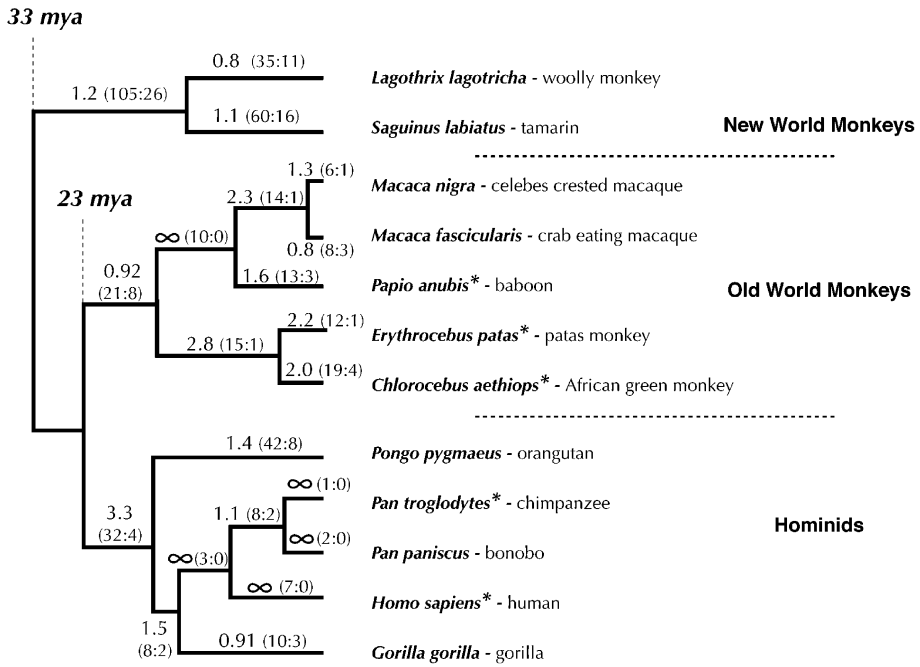

Figure 4.10: Diagram of the tree of full-length APOBEC3G sequences taken from [SEM04]. The starred species are those which are infected by HIV/SIV. The numbers on the branch indicate the maximum likelihood value of dN/dS estimated by PAML using the free-branches model. The numbers in brackets are the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

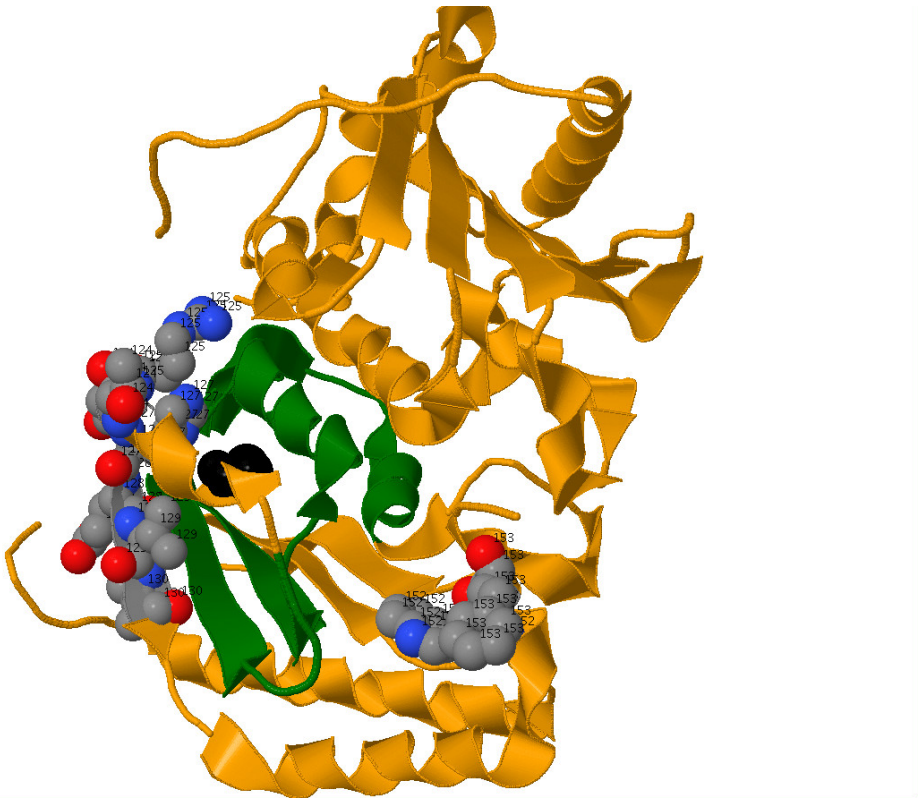

Figure 4.11: The structure of yeast cytosine deaminase, which is homologous to the APOBEC/AID family. The structure is of the homo-dimer. The region homologous to the Vif binding region in human APOBEC3G is drawn in green. The residue which aligns with residue 128 in the human APOBEC3G family is mapped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

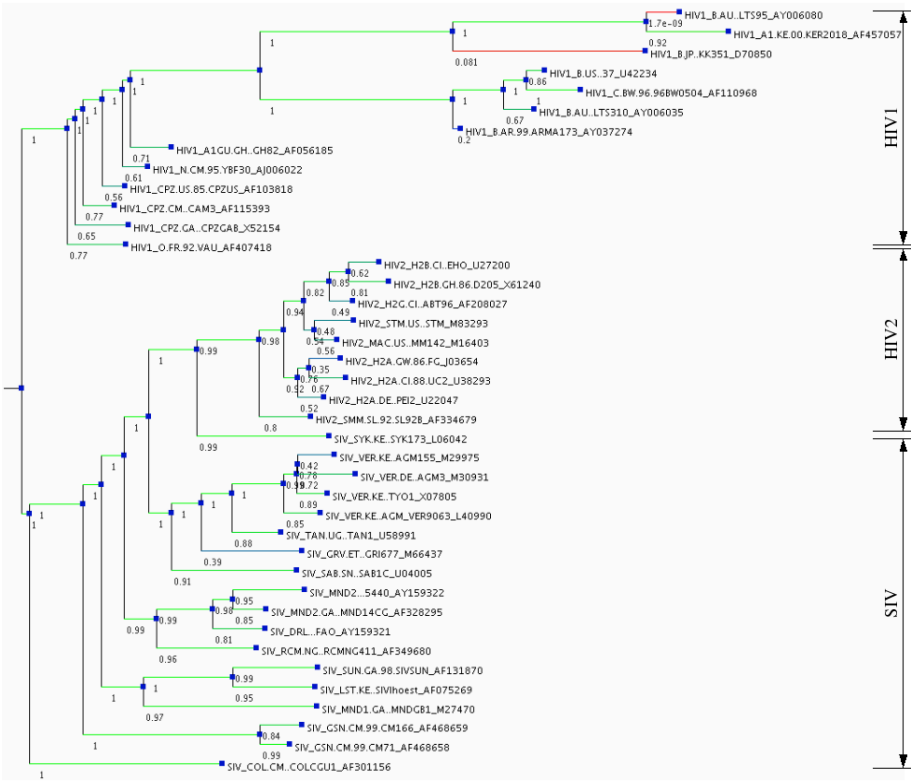

Figure 4.12: Tree of Vif proteins showing extensive selection. A green branch indicates strong evidence for selection on the branch to and below that node, whereas a red branch indicates the gene is evolving under a neutral DNA model. The numbers given below the branches indicate the maximum posterior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

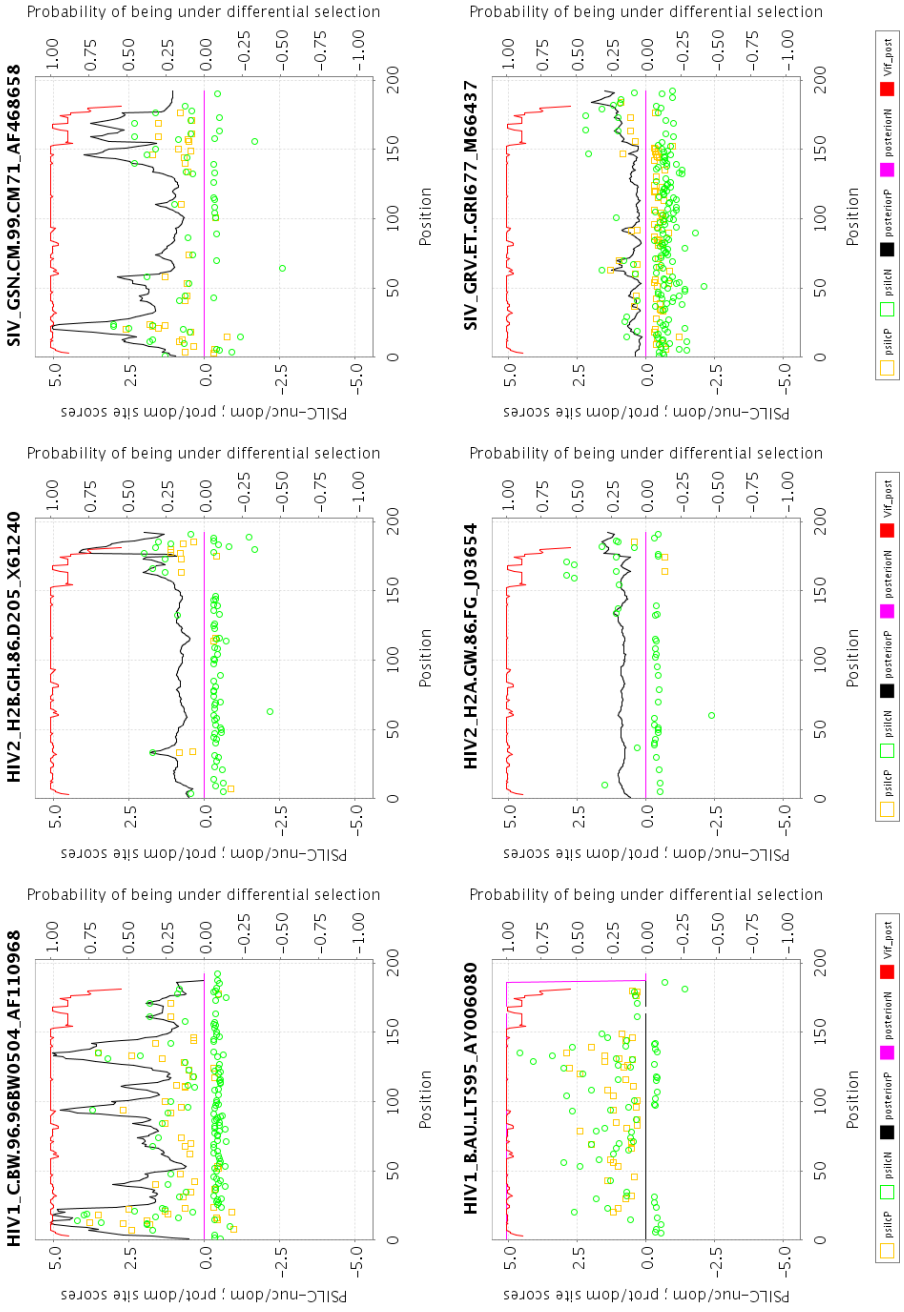

Figure 4.13: Site-specific graphs of selection acting on Vif genes. Green circles/orange squares indicate the site specific PSILC-nuc/dom and PSILC-prot/dom scores respectively, which, for clarity, are only plotted if less than -0.3 or greater than 0.3. The black/purple line indicates the posterior probability 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

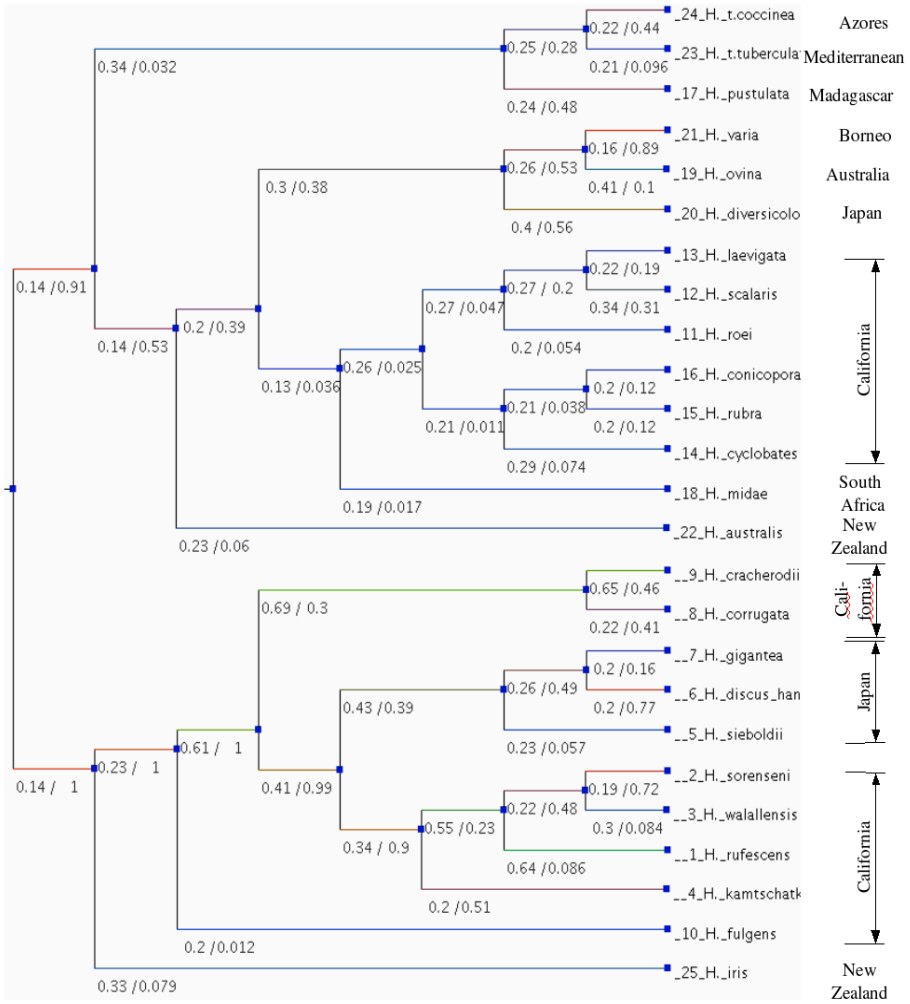

Figure 4.14: Tree of sperm lysin family, showing extensive 'non-structural' positive selection. Green branches to a node indicate support for evolution according to a neutral protein model rather than a domain constrained protein model, whereas red branches indicate support for a neutral DNA model rather 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).

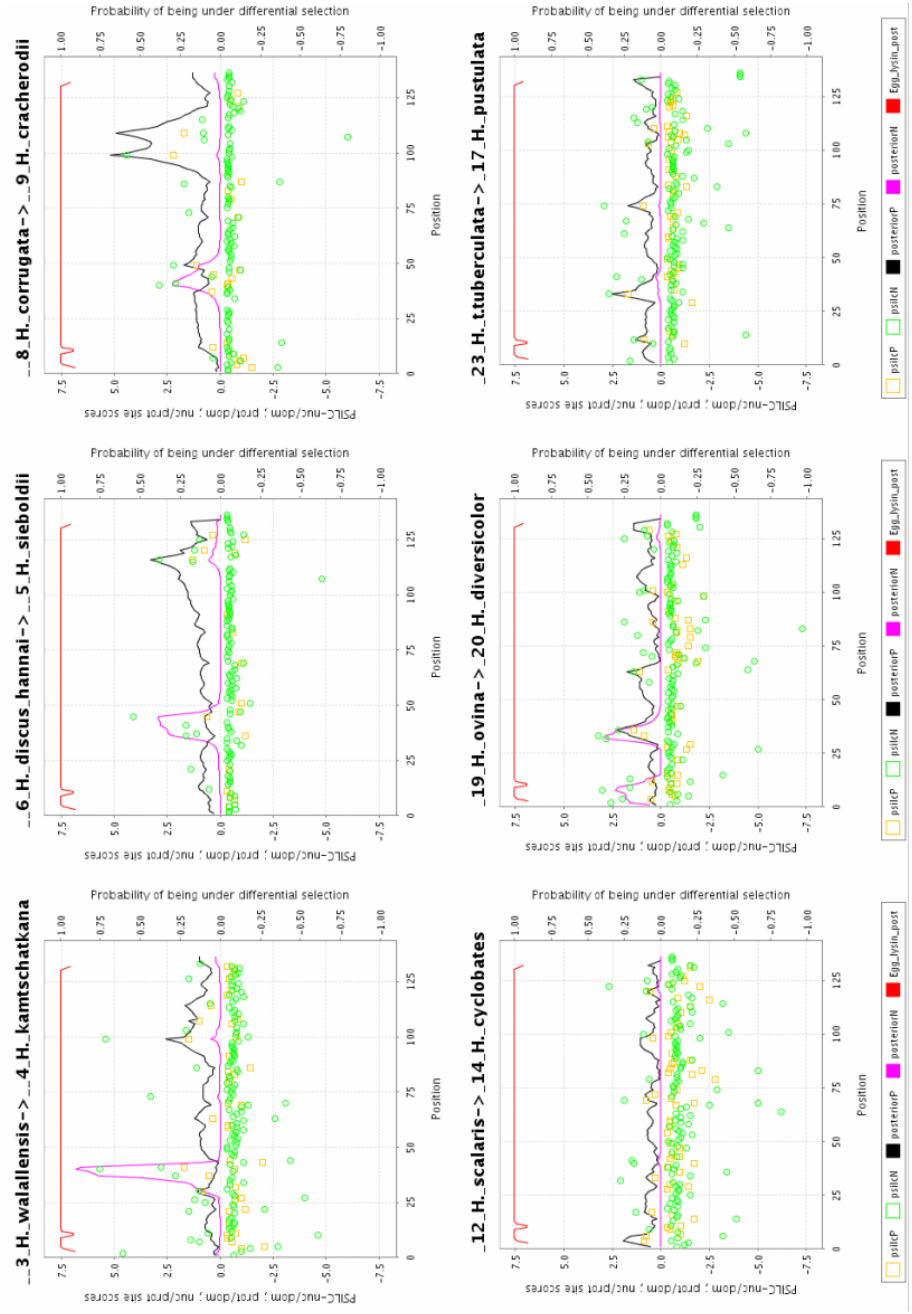

Figure 4.15: Site-specific graphs of selection acting on lysin genes. Green circles/orange squares indicate the site specific PSILC-nuc/dom and PSILC-prot/dom scores respectively, which, for clarity, are only plotted if less than -0.3 or greater than 0.3. The black/purple line indicates the posterior 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.

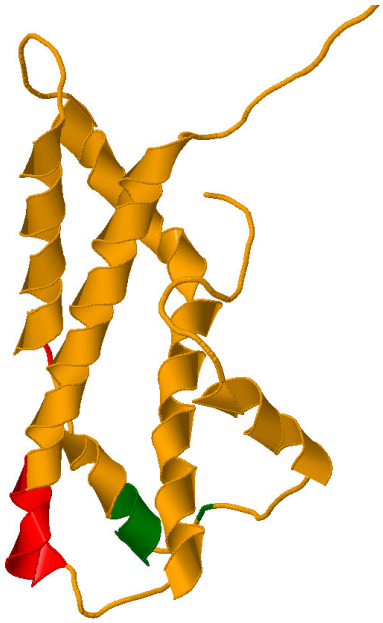

Figure 4.16: Structure of lysin, with regions of posterior probability of neutral DNA (red) or neutral protein evolution (green) greater than 50% in the clade *H. cracherodii* → *H. kamtschatkana*.

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

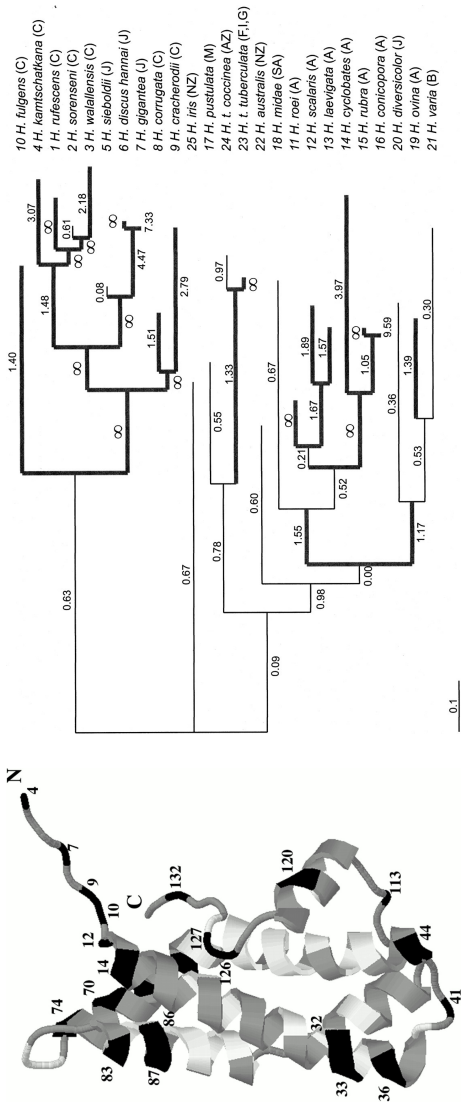

Figure 4.17: Taken from [YSV00]. Top: lysin tree, with the maximum likelihood estimates of dN/dS using PAML in the free ratios model on the branches of the tree. The thick lines indicate those branches with $dN/dS > 1$. Bottom: structure of lysin with sites inferred to be under positive selection (with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

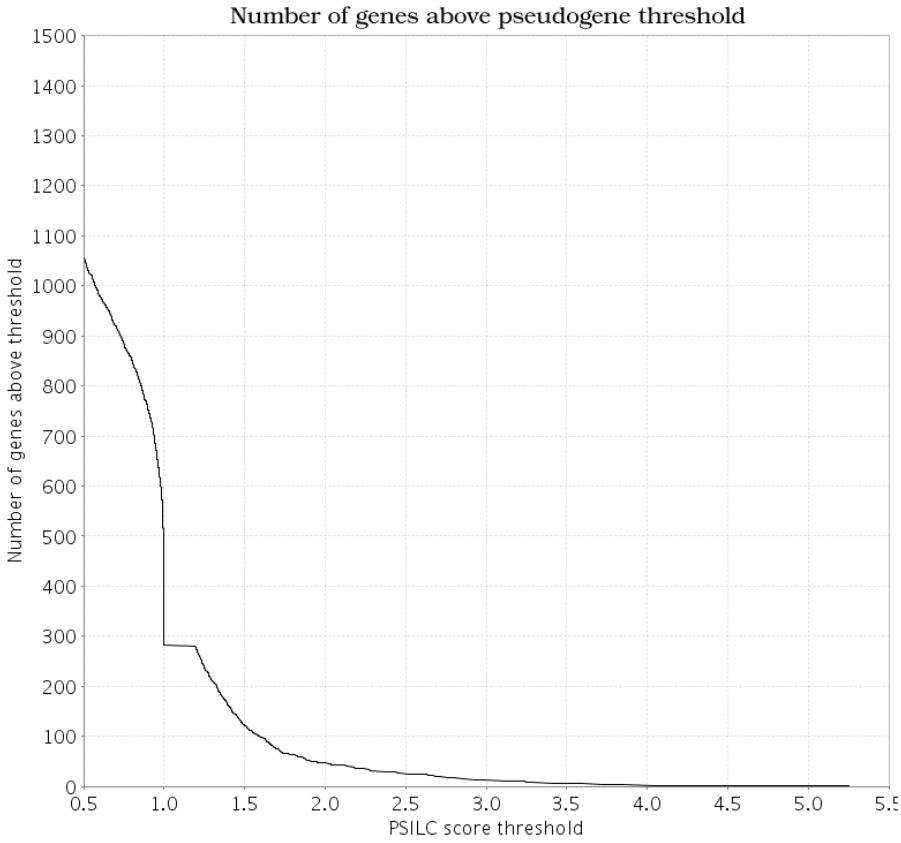

Figure 4.18: Number of human genes in clusters with combined PSILC posterior-nuc threshold/ PSILC nuc-dom score above threshold. The combined score was calculated by adding PSILC-nuc/dom / 100 to all genes with a PSILC posterior-nuc score of 1. Thus a score of 1.5 indicates a PSILC posterior-nuc score 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

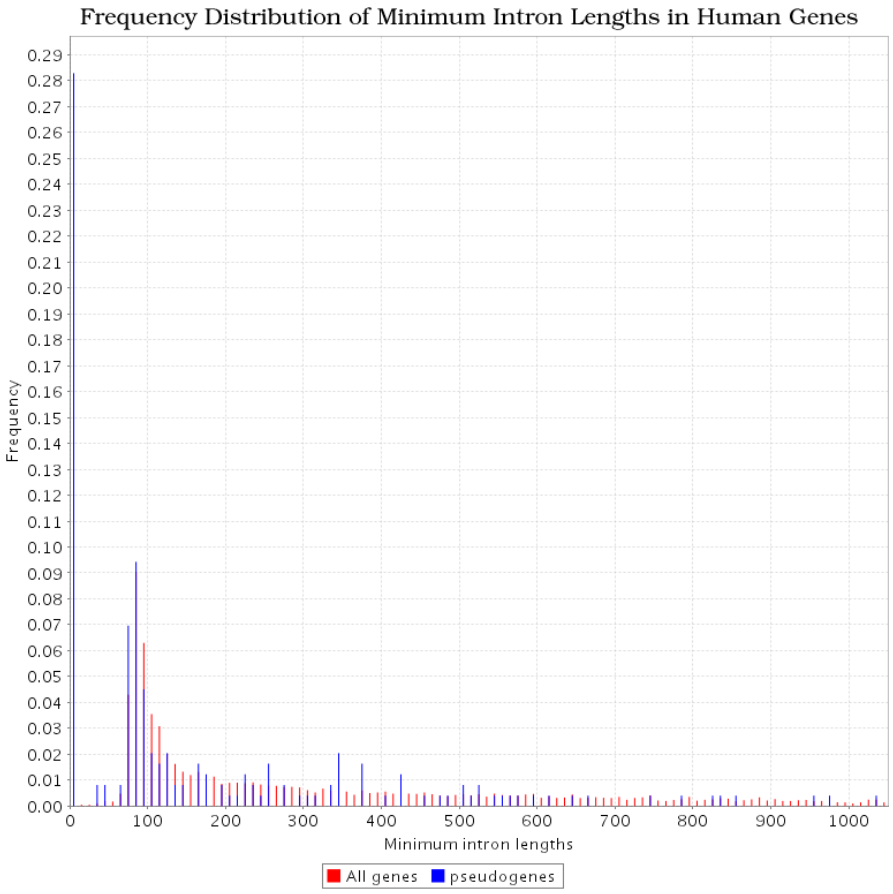

Figure 4.19: Frequency distribution of minimum intron lengths for multi-exon human pseudogene candidates as determined by a PSILC posterior nuc score of 1.0 (blue bars), versus all genes included in the study (red bars). Only intron lengths up to 1000 base-pairs are shown.

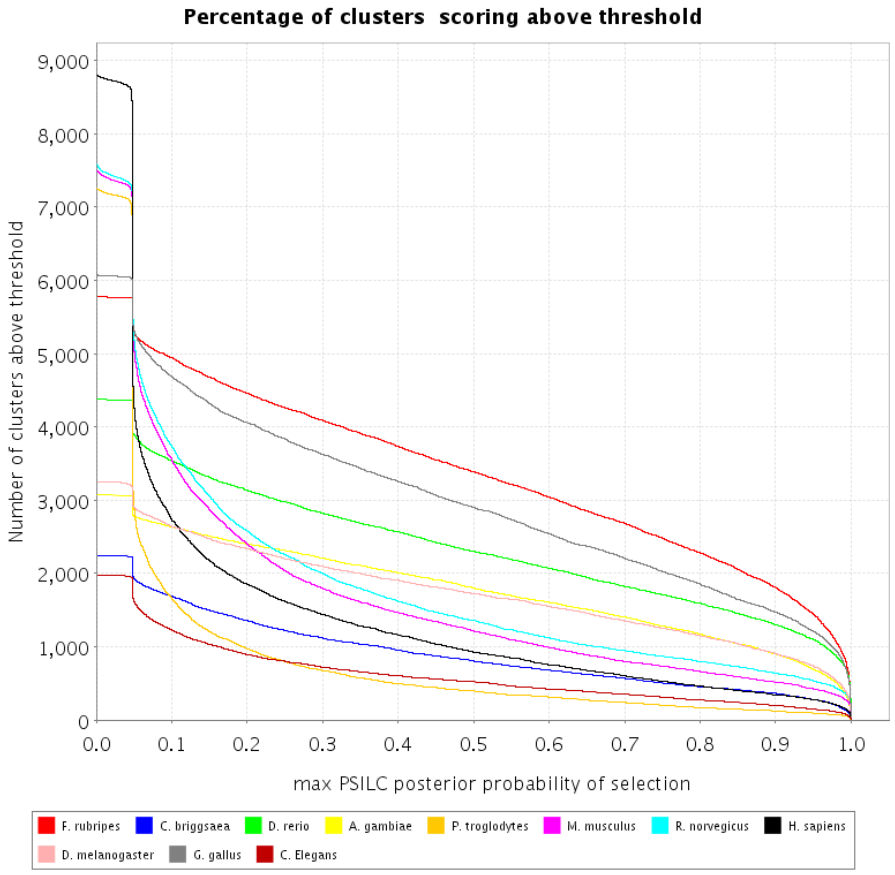

Figure 4.20: Number of clusters with a protein with maximum posterior probability of being under positive selection greater than a given max PSILC posterior-prot threshold in each of 11 species.

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.