Chapter 5

Conclusion

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

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

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

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

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

The methods outlined in this thesis calculate log-odds scores of a model of interest with respect to a background, or null model. While this is convenient for ranking matches relative to one another, it does not indicate the significance of a match. For protein domain identification, significance scores also enable the comparison of log-odds scores between different domain models. Calculating robust significance scores is one area in which further research is required for the techniques presented in this thesis. In the first two chapters, significance was calculating using the EVD parameterised for the standard profile HMM, without taking into account the extra information utilised in these chapters. This appears to be a satisfactory approximation for the language models, as the error vs significance curves are not substantially skewed by the language model scores. However the error vs significance curves for the phylogenetic HMM demonstrate that the EVD is not parameterised correctly for this score (see figure 3.6). Significance has not been calculated for the pseudogene or positive selection scores. An EVD could be fitted to the pseudogene scores of a large set of known functional genes in order to calculate significance for pseudogene predictions. Positive selection, on the other hand, is somewhat harder to unambiguously prove or disprove, and so no such equivalent benchmark set for positive selection exists. Hence calculating robust significance scores for the positive selection test will require simulation.

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

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

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