

“Statistical Models of Protein Structure Evolution”

Background

Over recent times there has been explosive growth in the availability of protein sequence data. This data explosion, combined with the increase in the power of computers and statistical models, has both driven and enabled the use of more rigorous techniques in bioinformatics research. In the area of comparative sequence analysis, for instance, there has been increasing emphasis on statistical and stochastic modelling of the evolutionary changes connecting the related sequences. Stochastic models of sequence evolution are indispensable for a number of reasons: Firstly, you can only observe a limited number of homologous sequences and their history and relationships can only be inferred using a model. Secondly, often quantities of interest are not observable directly, such as rates, strength of selection, bias in mutation events and more. These quantities are parameters in a model. Lastly, modelling is needed for hypothesis testing, confidence intervals, and determination of how much data is needed to distinguish between different hypotheses. These models have provided a wealth of important tools in common use for biochemistry, molecular biology, and biomedical research, providing much insight into the structure, function, and physiological role of many proteins.

Evolutionary models are virtually all continuous time Markov Models in a discrete space (nucleotides, amino acids, codons, complete sequences...). The continuous time Markov model for a single nucleotide has progressed from the simplest models (Jukes and Cantor, 1969) towards more complex models with up to 12 parameters (Felsenstein, 2004). Different sites can have different models (like 3 different models for the three positions in a codon) or the same for all positions (often used for non-coding regions). All models originally assumed independence between neighbouring sites (for instance nucleotides). Recent major developments have been incorporating dependence among sites like CG avoidance (Jensen and Pedersen, 2000), allowing insertions and deletions in the stochastic model (Thorne et al., 1991, Mitchison, 1999) and including dependence on local (Koshi and Goldstein, 1995, Goldman et al., 1996) or hidden (Koshi et al., 1999) protein structure. These techniques now form the basis of phylogenetics, database search methods, and diverse forms of functional annotation.

Parallel to the growth in sequence data, but with a delay of more than a decade, protein structure data have also experienced a massive growth. Unfortunately, protein structures are more complex than their sequence counterparts, so one cannot easily adapt algorithms developed in the sequence domain to examine the structural domain. For example structure comparison, the simplest possible application, encompasses a much greater variety of methods than is the case for sequence comparison (Brown et al., 1996). Much less has been done in the fields of structural evolution and the relationship between evolution at the sequence and structure levels. It has been noted for a long time that there is a strong relationship between evolutionary distance and structural similarity (Chothia and Lesk, 1986). Modelling of this relationship has previously been confined to extremely simple models (including those of RAG) (Govindarajan and Goldstein, 1997, Bornberg-Bauer and Chan, 1999, Dokholyan and Shakhnovich, 2001), or by considering sequence evolution in the context of a fixed structure (Parisi and Echave, 2001, Bastolla et al., 2003).

In addition to being more complicated, there are other important differences between protein structure evolution and sequence evolution. Changes happen on larger time scales, selection may play a stronger role, and the underlying genetic change often involve events other than insertions, deletions and substitutions. There are, however, strong similarities. Structures also evolve by duplication, mutational change, and selection, and pose challenges in finding homologies, interpreting evolution and creating useful classifications.

Despite the obvious challenges, a model of protein structure evolution would be an invaluable tool to researchers in many areas. Because much of the selective pressure on sequences exist at the protein level, understanding how protein structures change would be invaluable for modelling sequence evolution. Structures generally change much more slowly than sequences, suggesting that a good model of structural evolution would enable the detection of increasingly remote homologous proteins. More directly, the protein structure is closer than the sequence to the protein function and mechanism, the qualities of most interest to biochemists and molecular biologists. A model of protein structure evolution could provide tools to understand how these fundamental properties evolve, and how they are related for different proteins.

Programme and Methodology

The objectives

- To develop generally algorithms and software that
 - i. Analyze a set of homologous structures, i.e. estimated parameters in the evolutionary process, test hypotheses and make probabilistic statements about the evolutionary paths of the structures.
 - ii. Simulate structure evolution generally, not tied to a specific data set, to investigate more general questions about the occurrence of structures.
- Use the methodology and software to analyze a specific series of data sets of increasingly challenging nature.

The approach: Making a useful stochastic model of protein structure evolution involves solving several hard problems. First, a tractable model must be defined, then probability of observed structures must be evaluated, and finally, parameter estimation and model testing need to be carried out. The continuous-time Markov models of evolution used for other biological phenomena share common features that could guide model building for protein structure evolution. Such a model is defined by the various processes that can occur with their respective probabilities of occurrence. Given the difficulty of this modelling and associated calculations, it is important to realize that simplifications can be made both in the evolutionary process and how to represent a protein leading to a hierarchy of modelling problems, dependent on how realistic the model is. A first most realistic version stems from a full description of how sequences evolve, are translated, the protein folded and then the structure selected against. A second more simplified version will work on proteins represented as topologies and lengths assigned to all secondary elements. Starting with the most realistic model, evaluating the likelihood of a single step in protein structure evolution can be viewed as a three stage process:

1. A traditional genetic event must occur. This would be a substitution or insertion-deletion which could be modelled according to known sequence processes. On much longer time scales other events such as fusion, duplication, and inversion might play a role.
2. This change would give rise to a new protein sequence that might fold into a new structure. Although it is likely to be highly similar to the previous structure, this new structure must still be predicted. This prediction step would have to be done very efficiently as it will be repeated many times in the simulation of an evolutionary path.
3. The viability of the resulting sequence would have to be evaluated using a fitness function evaluating how likely the protein is to fold and be functional. Such fitness functions, while highly simplified, do exist, based on how “protein like” the predicted structure is.

The model draws heavily on already established models of sequence evolution. The major problems arise due to the size of the structure space, the number of steps a structure would have to traverse to be transformed into its homologous counterpart, and the number of possible paths leading from one protein to another. This leads to a computationally expensive algorithm necessitating multiple protein structure prediction steps.

This process defines the probability of a specified evolutionary path, but to evaluate the full likelihood of the data, the probability of a structure according to the equilibrium distribution would be needed. This problem is again hard, but the probability of this could be approximated by using the topological approach below where this could be answered. It is also a possibility only to use the evolutionary paths, but this would be incorrect and the “probability of structure” is of interest in its own right and necessary for homology testing, that can be rephrased as testing the probability of picking the two structures independently versus picking one and letting it evolve into the other.

Rapid model generation.

The simplest approach that we will consider starts with a single protein of known sequence and structure. The sequence of this structure would be randomly mutated (including insertions and deletions) and remodelled back onto its own structure. The modelling method we propose to use is very fast and completely automatic. It is based only on alpha-carbon positions but despite this can give good results. It starts at many different points on the structure and grows a new protein chain using the known structure as guide-points. The guide-points are assigned by a local

threading alignment which can accommodate insertions and deletions and match like secondary structures as well as hydrophobic to buried positions. Because of its stochastic component, each model that is generated is slightly different and typically 100 will be generated for each mutation in the sequence for 1000 mutations. The generation of this number of models has proved practical for sequences up to 100 residues in an overnight run in a cluster of 128 pentium processors. For each mutation in the sequence, the secondary structure will be re-predicted using the PsiPred method, this combined with associated changes in the hydrophobicity of the sequence will result in even greater variation in the pool of models.

Each model that is generated will be assessed for its fitness. This will entail a rapid check on compactness and degree of fold complexity to eliminate poor models. The remaining will be assessed using rapid evaluation methods, such as SPREK (Taylor Jonassen, 2004) and TUNE (Lin et al., 2003) that can operate with alpha carbon data only. The remaining models will be ranked by fitness and the best 10 taken to start the process again. With increases in computer power, and combining the resources of NIMR and Oxford, we anticipate that this 10 fold increase over the starting population is practical and can be maintained for perhaps up to 1000 cycles. The balance between the number of mutations, the population size and the number of cycles will be adjusted to provide a maximum return of information. The generation and selection of models is similar to the strategy employed in the Genetic Algorithm (GA) which has been used previously for similar model generation (Petersen and Taylor, 2003). Initially, the current protocol is different as it does not involve genetic crossover events between models (this is a difficult operation on 3D structures) but we anticipate that the method may eventually include this operation along with other features commonly employed in the GA approach. Our initial aim, however, is to start as simply as possible.

Constraints on the simulations

Limited evolution: As outlined above, the method would generate an evolving population of protein structures. Providing this did not wander too far from the starting structure, the method will generate sequence and structural variation typical of that observed in known protein families. This will be used to tune the method by comparing the results of the simulation to that observed in known families. Similar work has been carried out before but either on a more detailed level of representation (Baker et al. 2005) or on greatly simplified lattice models. It is our expectation that the current method will have sufficient detail to emulate the detailed studies and the power to go well beyond the lattice simulations.

Directed evolution: A direct extension of the limited evolution described above is to incorporate a directional bias in the selection to evolve the population towards a specific goal. This could be done most simply by specifying a target structure and incorporating an element into the fitness score that measures similarity to the target. This would allow one protein to be evolved into another. The pathway that connects the two structures would not be unique and the analysis of the variety of changes observed along different paths would be a measure of the similarity of the two structures.

This approach of connecting known endpoints in the evolutionary process can be extended to sets of more than two proteins related by a known phylogenetic tree. Although this would increase the complexity of the trajectories, it could greatly reduce the amount of Monte Carlo sampling required, as the resulting constraints on ancestral states would greatly reduce the number of reasonably-likely evolutionary trajectories.

Experimental checks: The constraints described above derive from fully solved protein structures, however, this degree of structural detail is not necessary as we need only know if a sequence can fold. Experimental methods using the Green Fluorescent Protein (GFP) as a probe have been developed and can be used to rapidly determine the folding state of a protein. Data has already been gathered by one of us (Stuart) on a number of proteins and we will investigate the degree to which this data is recapitulated by the evolution simulations. After an analysis stage we hope to be able to employ the evolutionary model in a predictive mode to anticipate sequences that will fold.

Topological approach

Extending the model to more distant relationships introduces the possibility that elements of secondary structure may be inserted, deleted, or changed. To deal with these larger structural changes, we plan to incorporate higher level representation of structure called the topological level. In this representation, the overall fold of the protein will be encoded in a symbolic description. Changes in the fold will be modelled by transitions in a network of labelled nodes, representing the secondary structure elements, and edges, indicating contacts interactions. For a given set of secondary structure elements, an exhaustive listing can be given of the possible structure topologies. Such topologies are well defined combinatorial objects and can be enumerated using standard combinatorial techniques. Similar structures will have

similar structure topologies. As one structure evolves into another, the corresponding structure topologies will similarly evolve.

In detail, this topological level of change will be implemented in an identical way to the more detailed (residue level) modelling described above. The only difference will be that the topological description of the protein will provide the framework (or template) over which the population of models are generated. Mutational operation at the topological level will therefore map into quite large shifts in the model structures. This will allow very large distances in protein fold-space to be traversed in a practical time. We will concentrate on families of proteins where there is thought to be very ancient evolutionary connections, and in particular on the helicases and polymerases where one of us (Stuart) has expertise.

The comparison of combinatorial objects has been much explored answering question such as what is the shortest distance in terms of some basic events (removing/adding/relabeling). (Waterman et al.(1976), Semple and Steel (2003)) The computational method would be very similar to the method developed by the principal investigator (Song and Hein, 2003), that allows us to rapidly answer a number of fundamental questions. What are the most distant objects of a given size? How many neighbours does an object have in terms of these basic events? The answers to these questions depend upon the relative probabilities of these various events. These probabilities are *a priori* unknown, but can be estimated based on the results of the previously-defined MCMC calculations. In addition, given any set of such probabilities we can calculate the likelihood of the observed structures resulting from the structural evolution process.

Protein fold-space

After exploring the trajectory between known proteins, we then plan to move towards evaluating “free evolution” from a known protein. If the underlying process is irreducible, then all known structures would be visited if observing a sufficiently long evolutionary trajectory. This would involve simulating a number of evolutionary trajectories with the methods described before, and analysing the resulting distribution of fold types. The resulting folds can be quantified by their similarity to idealised folds held in a “Periodic Table” of possible folds, as has been used to classify the variety of folds generated for a small protein using a Genetic Algorithm based search developed by WRT (Petersen and Taylor, 2003). The same approach can be used to model structural change by allowing a more constrained structural evolution to develop from each of two different structures. The resulting models can then all be compared to each other and their degree of similarity will generate a network that can be analysed for the best path from one structure to the other using standard graph analysis algorithms (Johannissen and Taylor, 2004). This would not lead to data analysis as it only involves the evolution of a single protein, but it would be of great interest to see how large a set of known structures would be reached and observing the long term behaviour of this process would also define probability to different classes of proteins.

General Application Areas.

There are many possible questions that could be addressed:

- **The nature of protein evolution:** Using the model, we can investigate whether changes come in clumps or are they evenly distributed both in time and sequence. We can estimate the importance of “neutral” steps (changes not involved in structural change yet required for such change to occur) and how quickly such transitions occur. Of interest is the relationship between the constraints placed on the protein and the resulting evolutionary trajectories. In particular, what constraints are necessary to mimic the observed relationships between protein structures?
- **The relationship between structural changes sequence:** We can monitor how independent the changes at various locations are, giving a measure of the strength of correlations between different sites in the protein. If we consider change at the DNA level, this will provide insight into the various methods used to evaluate “positive” selection (Nielsen and Yang, 1998) and also how the underlying triplet code can influence the evolution of the protein sequence and its resulting structure. In addition, we can monitor how often amino acids that have the same structural role are actually homologous.
- **Improved bioinformatics tools:** Comparative sequence analysis increasingly relies on good models of sequence evolution. We can use these structure-evolution models to build better models for change at the underlying sequence level and so develop better methods for using structure comparisons for the alignment of protein sequences. We should also be able to identify the types of sequence modifications that cause or accompany structural change, allowing us to better detect these patterns in homologous sets of proteins.

- **Defining the appropriate “metric” between structures.** Our approach should give a better understanding of the relationship between structural differences and evolutionary distances. This provides a basis to decide when structural similarity can be used to provide evidence for distant homology.
- **A limited number of folds for proteins?** We can assess whether a structure is avoided for physicochemical reasons, or has it just not had the evolutionary opportunity to arise. We should see if some structures are less likely than others, and if so, whether this is because of the properties of these structures, or their relationship to others in structure or sequence space.

Work Plan:

For the 1 year duration of this project we will have 4 goals:

i. Implementation of the topological model. Since this is very well defined and will use techniques we are very familiar with we anticipate this to be 20 % of the activity.

ii. Free Evolution of Structure Evolution. This will be an ongoing activity through the whole period demanding about 10% of the activity. It does not require the integration over possible evolutionary paths as in iii.below, but only evaluation of the probabilities of the set of next possible steps – a much easier task. There will during the period be improvement to the basic evolutionary model so the process will generate more protein-like structures.

iii. Implementation and exploration of the MCMC method for comparing 2 (possibly more) structures. The Hein – group have much experience and present activity (A.Rocco (100%), A.Caldeira (100%), Hein, Lunter, Drummond, Miklos – all 10-20%) devoted to a simpler version of this problem: namely stochastic evolution of strings in terms of insertions, deletions and substitutions. This is the hard problem and will be attacked very step-wise, starting with very small and similar structures and then moving to larger more dissimilar structures. There are standard tests for convergence of the MCMC chains that will allow monitoring of the computational limits and demands of this approach. This should be 70% of the activity.

iv. Given the expected success and clarification of i-iii, Oxford and Mill Hill will prepare for a larger proposal describing in detail how stochastic models of evolution could be applied to determined protein structures in the period 2006-2009 to answer fundamental questions concerning the appearance, homology and evolution of (almost) all proteins.

Division of Labour: The PostDoc and the Hein group will do the majority of the programming and contribute MCMC expertise, Mill Hill will contribute expertise in computational evaluation of structures and sharing relevant software they have developed. David Stuart will provide test examples and biological interpretation.

Dissemination and Exploitation

We will disseminate our work through a number of national and international conferences targeting meetings that attract both academic and industrial participation. The results of these studies will also be transferred to the scientific community through publications in international journals.

Justification of resources

Manpower There is an acute shortage of trained bioinformatics researchers as recently documented by separate reports from the Research Councils and the Wellcome Trust. An RA employed on this project would gain invaluable bioinformatics skills. The candidate would need a strong mathematical/computational background and at least a proven interest in biological problems, preferably structural biology. Bioinformatics is a difficult area to recruit in due to competition from industry and the Public sector research institutes. The group as a whole has succeeded in making outstanding appointments to recently advertised postdoctoral positions. It has made a successful case to various research councils and charitable trusts that in order to continue to attract and retain such candidates the post-docs be paid above the wage for age, with this in mind we propose the start salary as point 8 on the RSA1 scale.

Travel Central to this proposal is a close collaboration between the Oxford group (Hein, Stuart) and London Mill Hill group (Taylor, Goldstein) that has complementary expertises and the proposed research would lose much of its quality if determined effort was not mustered to keep this frequent. We thus anticipate on average for the 12 months a London-Oxford visit per week – i.e. 30 such visits. This could be the postdoc or the whole group visiting each other.

Equipment. The equipment asked for is 1 computer, 1 laptop and large hard disk. It is clear that these investigations are going to be very computer intensive and will need good equipment and data storage facilities.

References

- Brown, N., C. Orengo and Taylor (1996) "A Protein Structure Comparison Methodology" *Computers Chem.* 20:359-380.
- Chothia, C. and Lesk, A. M. (1986) The relationship between the divergence of sequence and structure in proteins. *EMBO J.*, 5, 823-826.
- Felsenstein, J. (2004) "Inferring Phylogenies" Sinauer
- Gillespie, J. (1991) "The Causes of Natural Selection" Oxford University Press.
- Goldman N., Thorne J and Jones (1996) "Using Evolutionary trees in protein secondary structure prediction and other comparative sequence analyses" *J. Mol. Biol.* 263:196-208.
- J. Hein, J.L. Jensen and C. Storm (2003) "Recursions for Multiple Statistical Alignment" (*PNAS* 100(25):14960-14965.)
- Jensen, J.L. and Pedersen, A.-M.K. (2000) "Probabilistic Models of DNA Sequence Evolution with context dependent rates of substitution" *Adv. Appl. Prob.* 32:499-517.
- Johannissen, L. O. and Taylor, W. R., "Protein fold comparison by the alignment of topological strings", *Prot. Engng.* 2004 16:949-955"
- Jukes, T.H. and Cantor, C.R. (1969) "Evolution of Protein Molecules" pp21-132 in *Mammalian Protein Metabolism* vol. III ed MN Munro. Academic Press. New York.
- Lin, K. and Taylor, W. R. and Klienjung, J. and Heringa, J. "Testing homology with (Cao et al.): A contact-based Markov model of protein evolution", *J. Compu. Biol. Chem.*, 27, 93-102, "2003"
- Jun Liu: "Monte Carlo Strategies in Scientific Computing" Springer 2001
- Lunter, Miklos, Drummond, and Hein (2004) "Alignment, Statistics and Evolution" (in "Statistical Methods in Molecular Evolution" ed. Rasmus Nielsen, in Press)
- Petersen, K. and Taylor, W. R., "Modelling zinc-binding proteins with GADGET: Genetic algorithm and distance geometry for exploring topology", *J. Molec. Biol.* 2003, 325:1039-1059"
- Douglas M. Robinson, David T. Jones, Hirohisa Kishino, Nick Goldman, and Jeffrey L. Thorne Protein Evolution with Dependence Among Codons Due to Tertiary Structure *Molecular Biology and Evolution* 2003 20: 1692-1704
- Seemple, C. and M. Steel (2003) "Phylogenetics" Oxford University Press
- Song, Y and J.J. Hein (2003) "Parsimonious Reconstruction of Evolution and Haplotype Blocks" (WABI03, Hungary. *Lecture Notes in Bioinformatics* vol. 2812. p287-302)
- Taylor, W. R. and Munro, R. E. J. and Petersen, K. and Bywater, R. P. Ab initio modelling of the N-terminal domain of the secretin receptors", *Comp. Biol. Chem.* 27:103-114, "2003"
- Taylor, W. R. and Jonassen, I., "A Structural Pattern-based Method for Protein Fold Recognition", *Proteins: struc. funct. gene.*, " *Proteins* 56:222-234" Taylor, W. R. and Jonassen, I
- Thorne, J.L., Kishino, H. and J. Felsenstein An Evolutionary Model for Maximum Likelihood Alignment of DNA Sequences. In *Journal of Molecular Evolution*, 33:114-124, 1991.
- Waterman, M.S., Smith, T.F. and Beyer, W.A. (1976) "Some Biological Sequence Metrics" *Adv. In Math.* 20:367-87.