

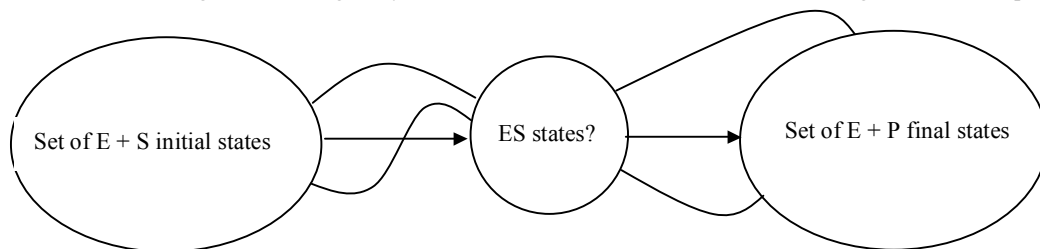
The TPS algorithm & the Evolutionary Path of Protein Structures

1.12.07

by Tom Darden, Jotun Hein, Mark Sansom, , Lee Pedersen & Willie Taylor

Comparing homologous objects are central to biology and the last decade have been dominated by this in the field of comparative genomics. However, many other objects than sequences in biology are homologous and can be subjected to evolutionary study. Examples could be networks and structures. Protein structures have been compared for decades now in a non-statistical and non-evolutionary way. Such analysis would benefit tremendously by methods that modelled the evolution over time of structures explicit. The motivation for doing this, has increased due to the large set of known structures presently, but has not materialized due to lack of models and computational power. Fortunately, such models have been explored extensively both in statistics and also in Molecular Dynamics (MD). Molecular Dynamics can simulate the behaviour of molecular systems with up to 10^6 atoms (typically 10^3 - 10^4 atoms) and for time periods of up to a microsecond (10^{-6} s) dependent on details. Applications often involve dynamic paths, where both start configuration and end configuration of the system is known – for instance the catalysis of a substrate into a product. This is a well studied problem in statistics and the natural algorithms are now being used large scale in MD under the name Transition Path Sampling (TSP) (Bolhuis et al, 2002). The modelling problem described here is very similar to the TPS problem and can be explored using the same algorithm.

The proposed project clearly is ambitious, but is worth pursuing as it represents the optimal way of studying protein evolution. It is ambitious since it involves investigate *all* possible path between two protein structures. And for each path it involves predicting *all* protein structures on the steps of the path. There are methods to do this, but great care must be taken to be efficient in computations. Both paths and structures will have great redundancy allowing for reuse of calculations. Additionally, protein structures can be represented at different levels of detail, with the coarsest being representing secondary structure elements (SSEs) as labelled sticks with relationships to other SSEs, and the (almost) finest level being a full atomic representation of the structure. Making the correct representational choice is crucial in making an interesting analysis in finite time and at the same time not having trivialized the problem.



The first models of enzymatic action were published in the first decades of the 20th century by Michaels, Menten, Henry, Haldane and others. These are models with a few states and interactions and lead to simple equations allowing explicit analysis of their dynamics. It is clear that the real picture is vastly more complicated and started to be analyzed in the last decade of the 20th century by molecular dynamics. A naïve MD simulation could have 10^3 - 10^4 atomic positions (enzyme, substrate, water and ions) and would in principle be simulated for 10^9 steps of 10^{-15} second duration. Clever techniques such as Transition Path Sampling can improve significantly on this and force dynamic trajectory toward a desired end state and still allow rates to be calculated and identification of key interacting groups.

The analysis of two protein structures will imply investigation paths from one to the other. Doing this will need four problems to be addressed:

i. Levels of representation of structures can span full atomic presentation, simplified amino acids, .. , only a graph with SSE. The first possibility is clearly preferable and should be achievable.

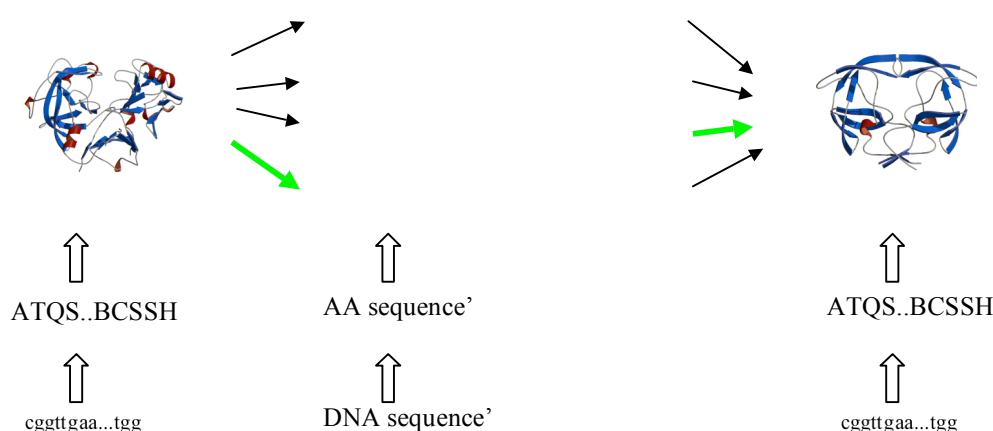
ii. The fitness functions possible on structures are huge and could consider detail functionality, but computational realism will force such functions only to consider simple indicators of “structureness”. Typical for such functions would be radius of gyration, statistics of secondary structure and more.

iii. The evolutionary steps possible from any structure/sequence are well known as are the biochemical events (substitutions and indels). If a bound is placed on insertion length, there will only be finite number of possible structure/sequence neighbors.

iv. Fast Evaluation of Structures. 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 and 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.



The basic problem of evolutionary structure analysis of two homologous proteins is simple to formulate: A DNA gene sequence is translated to a protein that folds up in a structure. In evolving one structure into another, the corresponding genes are changed by substitution and insertion-deletions from the gene behind the first structure to the gene behind the second structure. Each gene on this unobservable evolutionary path has been translated into a protein that folded into a structure. There are methods that can sum over all paths from one sequence to another if we ignore structure. However, if structure is considered each step on this evolutionary path will be influenced by the quality/fitness of the structures. In this problem cannot be address without assigning some fitness to structures.

The TSP algorithm and Markov Chain Monte Carlo (MCMC) – Let $x(t)$ be the configuration at time t . $x(0)$ will be the configuration of the first protein and $x(T)$ the configuration of the second protein. For each $x(t)$ there is a set of possible neighbours, $N(x(t))$, defined by applying to the sequence of $x(t)$ and predicting its structure. Let $S=[x(0),x(\delta),x(2\delta),\dots,x(k\delta)=x(T)]$ be a proposed evolutionary trajectory. Let $F(x)$ be the fitness of x , then $P[x(i\delta)\rightarrow x((i+1)\delta)] = F(x((i+1)\delta))/\sum F(x')$, where the summation is over all neighbours to $x(i\delta)$. This summation must be done stochastically. The $P[S]$ is the product of the probabilities of all the individual steps. Given a way to define a neighbourhood to a path, then an alternative path S' can be proposed and this path can be chosen as current with probability $\max[1,P[S']/P[S]]$. This defines a random walk in path space that will eventually visit all paths according to their probability.

An initial path can be obtained by aligning the two sequences and then choosing an arbitrary order of the proposed events in the alignment.

Objectives

Proper evolutionary models should be able to do for structure analysis, what such models did for sequence analysis:

- i. Allow analysis of 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. To simulate structure evolution generally, not tied to a specific data set, to investigate more general questions about the occurrence of structures.

References

Basner, JE and SD Schwartz (2005) "How Enzyme Dynamics Helps Catalyze a Reaction in Atomic Detail: A Transition Path Sampling Study" JACS 127.13822-31

Bolhuis, PG, -D Chandler, C Dellago and PL Geissler "TRANSITION PATH SAMPLING: Throwing Ropes Over Rough Mountain Passes, in the Dark" Annual Review of Physical Chemistry Vol. 53: 291-318

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.

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"

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"