

# Evolutionary Stepping Stones: Trajectories in Structure Space

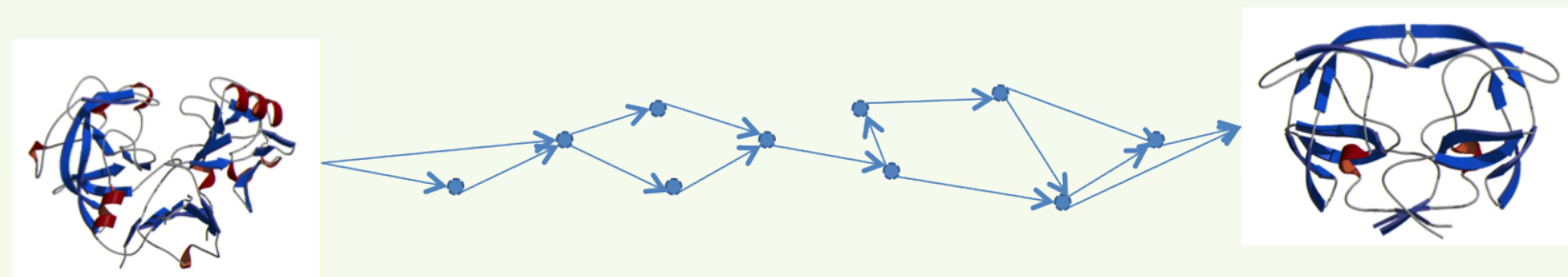
J. Domanski<sup>‡</sup>, A. FitzMaurice<sup>‡</sup>, E. Sjöland<sup>‡</sup>, J. L. Herman<sup>‡</sup>, R. Lyngsø<sup>‡</sup>, M. Sadowski<sup>\*</sup>, W. R. Taylor<sup>\*</sup> and J. Hein<sup>‡</sup>

<sup>\*</sup> Division of Mathematical Biology, MRC National Institute for Medical Research, The Ridgeway, Mill Hill, NW7 1AA, United Kingdom

<sup>‡</sup> Department of Statistics, University of Oxford, 1 South Parks Road, OX1 3TG, United Kingdom

## Motivation

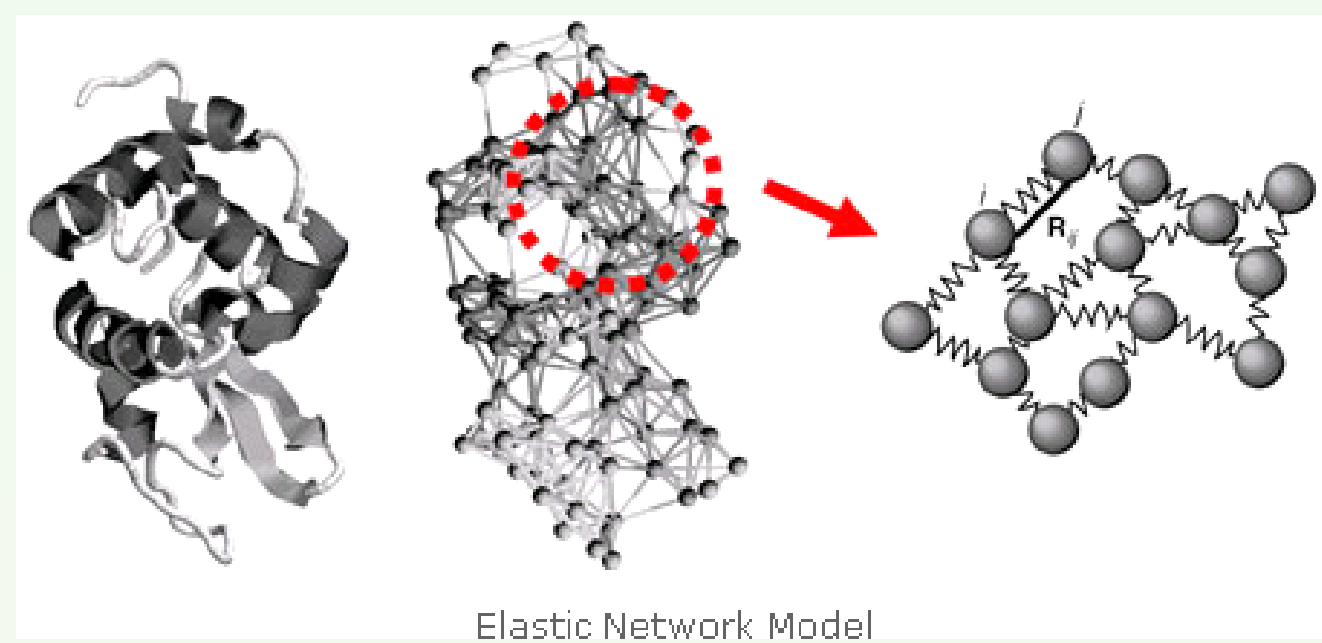
Whilst much work has been done on protein sequence evolution, fewer studies address the evolution of protein structure. Given two protein structures, how might we determine their evolutionary relationship? We have addressed this question by randomly generating evolutionary 'stepping stones' between two real, or *native*, proteins, and then considering the work required to transition between these structures to find likely evolutionary paths.



Simplified representation of the paths through which one protein can evolve into another.

## Basic Model

Our dataset contains both the true protein structures and the generated intermediate stepping stones. Each stepping stone is represented by just its  $C_\alpha$  atoms, which are thought of as being connected by springs, creating a so-called elastic network. The work (and hence the probability) to transition from a protein  $i$  to a protein  $j$  can then be calculated from elastic network deformations;  $work = ENM(j|i)$ . This work is referred to as the *ENM potential*.



Representing protein structure using the Elastic Network Model.

We can think of the set of protein structures together with the energy required to transition between them as a weighted graph. The transition energies are stored in the energy matrix  $V$ , where  $V_{i,j} = ENM(j|i)$ .

## Shortest path: Dijkstra's algorithm

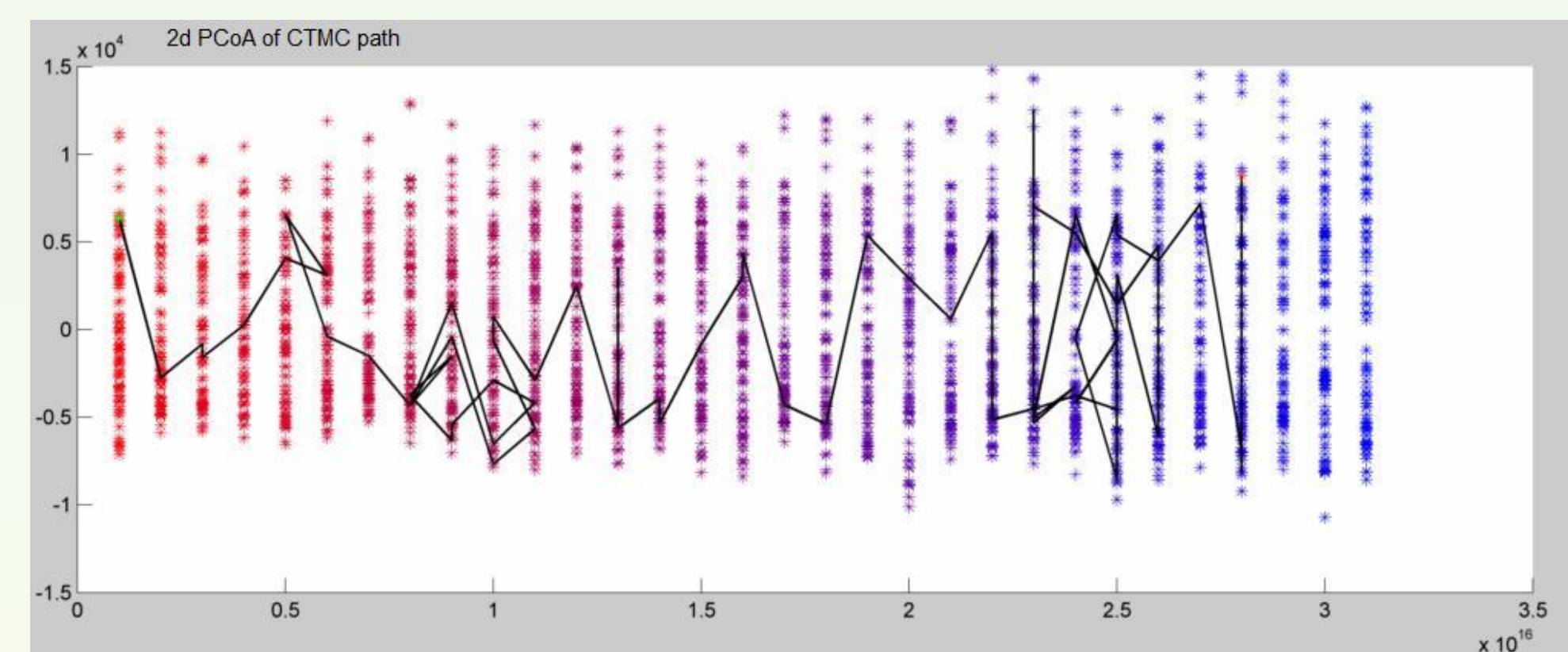
Having assigned weights (i.e. ENM potentials) to the edges of our stepping stones graph, we compute the least energy path from protein  $i$  to protein  $j$  using Dijkstra's algorithm. This is the global optimum route.

## Other paths: CTMC

Dijkstra's algorithm (used to determine the global optimum path) gives us little information about other likely paths, or the probability of transitioning from  $i$  to  $j$ . To address this limitation, we use a Continuous-Time Markov Chain (CTMC) to generate paths between  $i$  and  $j$  on the space of stepping stones. Transition state theory is used to claim that the rate of transition from  $i$  to  $j$  is proportional to  $e^{-\alpha V_{ENM}(j|i)}$ , for a spring constant  $\alpha$ . So the transition rate matrix  $Q$  is such that for  $i \neq j$ ,  $q_{ij} = e^{-\alpha V_{ENM}(j|i)}$ . Further we require the diagonal elements  $q_{ii} = -\sum_{j \neq i} q_{ij}$  to ensure that every row sums to zero.

To compute a path on the state space of stepping stones that begins at structure  $i$  and ends at structure  $j$ , we must discretise the CTMC. We achieve this using an algorithm known as uniformization, which is endpoint-conditioned to ensure the given end structure is reached in finite time.

Having obtained a selection of paths on the space of stepping stones, we consider a variety of methods of visualisation, including *Principal Coordinate Analysis (PCoA)*.



The 2d PCoA visualisation of a path on the space of stepping stones generated using a CTMC.

By running the CTMC a large number of times, we can gather statistics about the frequency with which different paths occur, and hence about the probability of the evolutionary path between two protein structures passing through a given stepping stone.

## Results

With 27,000 stepping stones, we found that CTMC paths generated on the state space didn't tend to overlap sufficiently frequently to draw many significant conclusions. For this reason we considered transitions between protein *folders*, as opposed to between individual protein structures. Having generated a large number of paths, we visualised the flow density between different folds as a weighted graph, the width of whose edges represent the density of transitions between the two connected folds. Note that to account for different folds containing different numbers of stepping stones, these flux densities are weighted by comparison with the expected transitions given a uniform probability distribution of transitions. The graph is directed; the green edges are those going from the lower to the higher indexed vertex, and the red edges are those going from the higher to the lower.

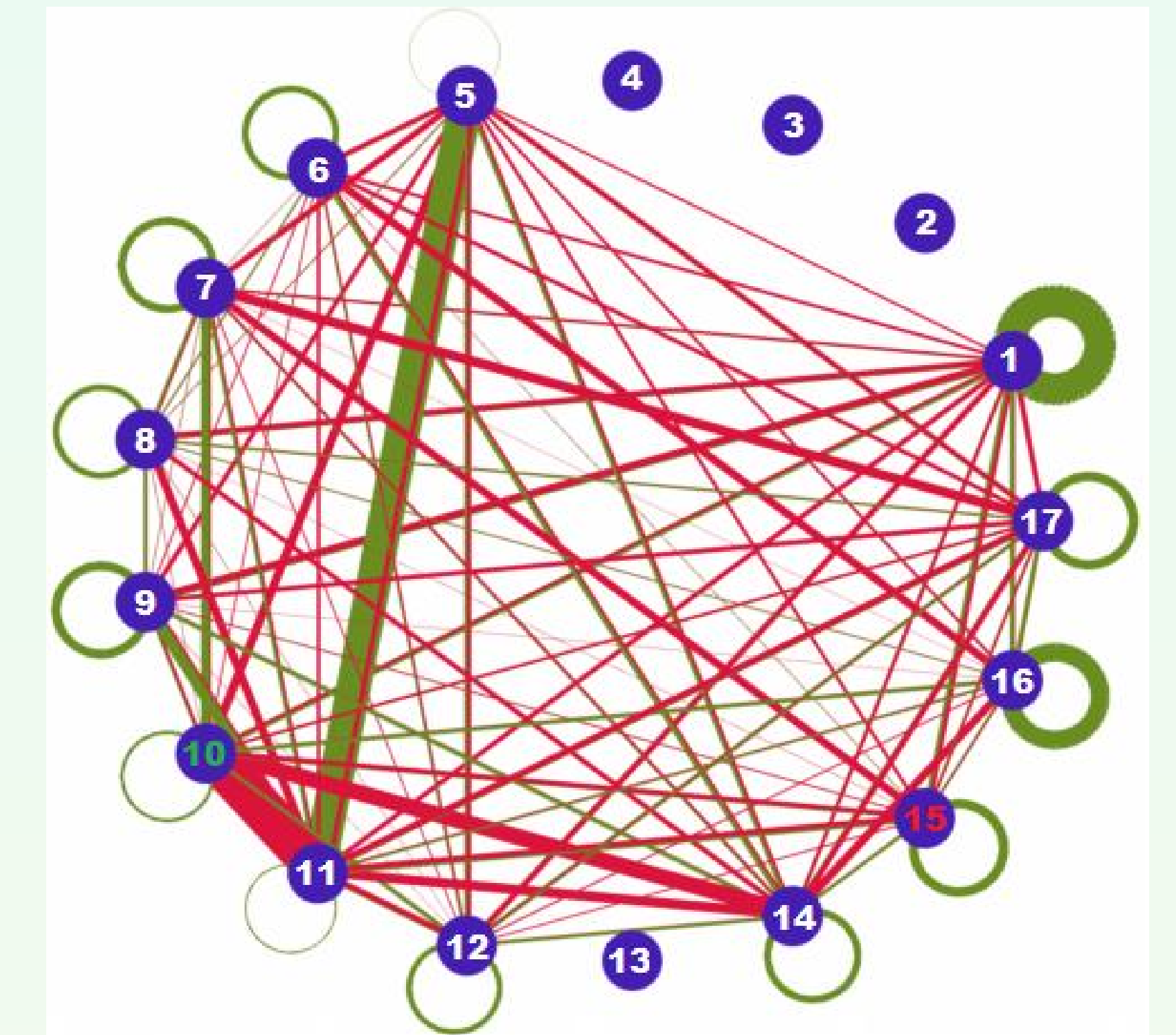
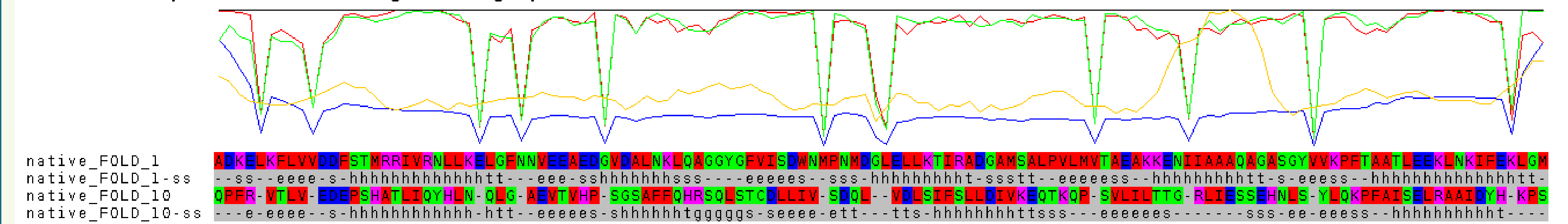


Diagram representing flow density between different protein folds.

From each path we obtain in structure space we can generate an alignment, and a set of paths between two points in the space can be used to construct a consensus alignment. Using our elastic network model, we expect the gaps (insertion and deletions) to fall in the loop regions of a protein. This is illustrated below: many of the gaps occur between two secondary structure elements (secondary structure, or "-ss" rows); vice-versa, the structurally defined elements of the proteins are aligned with much higher probability (blue line for probability, green and red for forward and reverse marginals of the probability, respectively).

native1and10/paths-native1to10-alignment.log.mpd



Alignment diagram.

## Conclusions

Whilst our results so far indicate probable intermediate stepping stones on the evolutionary path between our two native proteins, more work is needed to determine which of these is statistically significant. Having achieved this, our model could be assessed by carrying out a comparison between our results and known protein phylogenies.

## References

- [1] Bolhuis, P. G., Chandler, D., Dellago C. and Geissler, P. L. (2002) Transition path sampling: throwing ropes over rough mountain passes, in the dark *Ann. Rev. Phys. Chem.*
- [2] Marchal, S. *et al.* (2008) Asymmetric kinetics of protein structural changes *Acc. Chem. Res.*
- [3] Dellago C., Bolhuis, P. G., Csajka, F. S. and Chandler, D. (1998). Transition path sampling and the calculation of rate constants *J. Chem. Phys.*

## Acknowledgements

This work was carried out as part of the Oxford Summer School in Computational Biology, 2011, in conjunction with the Department of Plant Sciences, and with support from the Department of Zoology. Funding was provided by the Nuffield Foundation and EU grants. We thank Dr Steven Kelly for providing computational resources.