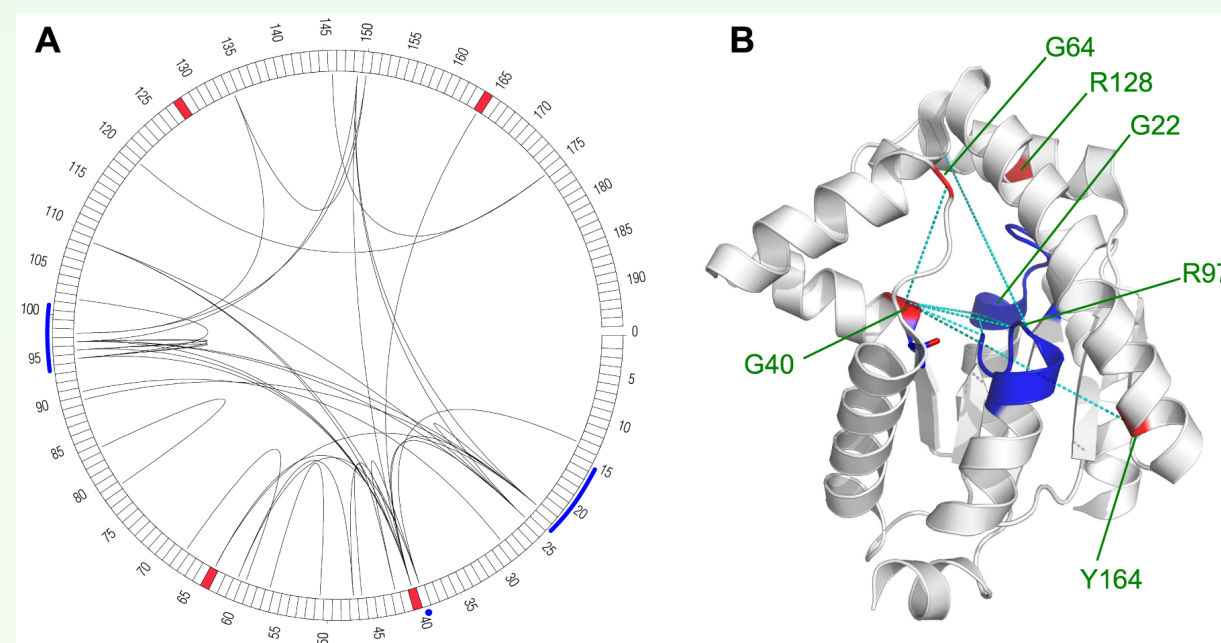


Motivation

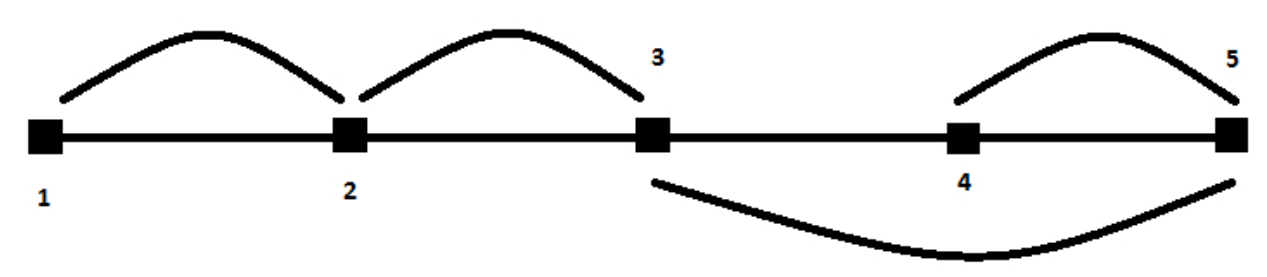
It is well known that the residues of some positions in a protein evolve in a dependent manner as a result of local proximity of these positions. After a destabilizing mutation occurred at one site it is more likely that also the correlated site will mutate to minimize these destabilizing effects.



Nevertheless only very few evolutionary models deal with correlated evolution. Therefore we tried to show that residues which evolved in a correlated manner can be estimated by using a probabilistic evolutionary model on a phylogenetic tree.

General approach

Our work considers evolution to be governed by a pair-wise correlated process between amino acids. The transitions between these states are modeled with a Continuous-time Markov Chain. The dependency structure, which can be represented as a tree, is then restricted to a spanning tree to ensure that a Markov chain Monte Carlo approach is feasible.

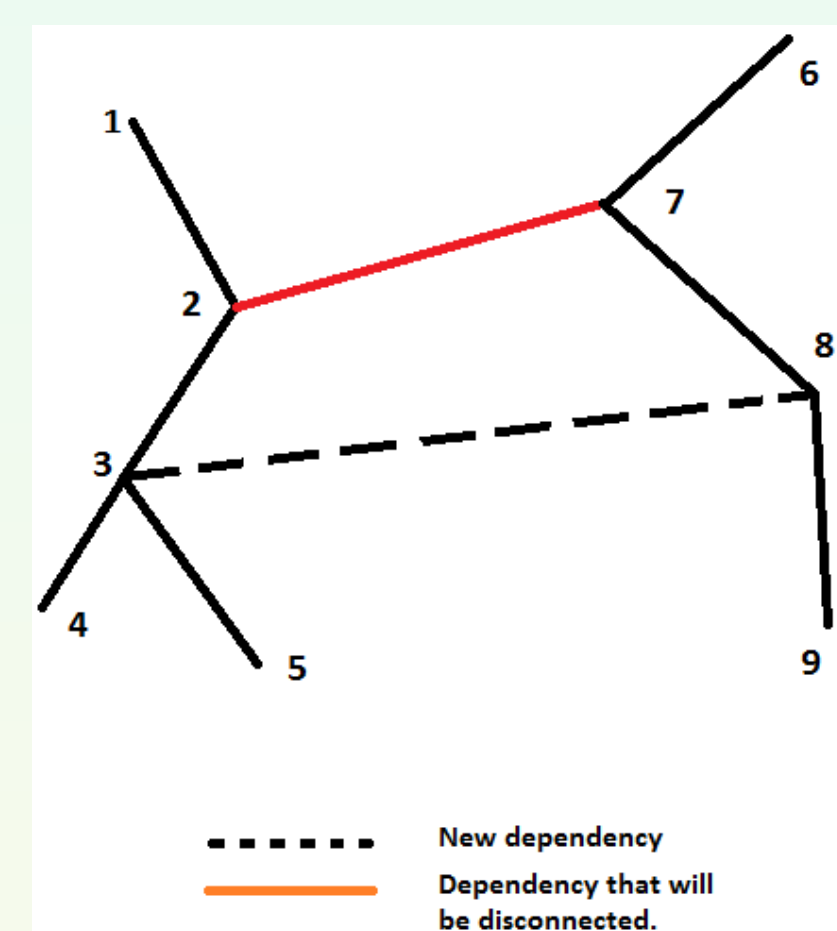


The approach is verified by comparing results when the algorithm is applied to a smaller alphabet and data that were simulated with the same model.

Simplifications to make the model computable

As we expressed $Q^{(L)}$ as a sum of matrices that do not commute, we can use the Trotter expansion of $\exp tQ^{(L)}$ in order to compute $P^{(L)}(t)$. Under the spanning tree assumption for the dependency structure, the distribution of probabilities for each sequence can be written as a product of pair marginals divided by site marginals, and these marginals can be updated efficiently using a variation of the sum-product algorithm in order to calculate the likelihood of the observed sequences given a particular dependency structure.

Since it is not possible to analytically find the maximum likelihood spanning tree, we use a Markov chain Monte Carlo approach to sample different dependency structures. The mechanism of how a new dependency structure is proposed is shown in the right picture.



The model

Every Continuous-time Markov chain can be characterized by a rate matrix Q . As we want to model pairwise correlations we are not longer using Q as the rate matrix but $Q^{(2)}$. Both matrices are shown below to illustrate the differences for a binary alphabet size:

$$Q = \begin{matrix} 0 & -\alpha & \alpha \\ 1 & 1/\alpha & -1/\alpha \end{matrix} \quad Q^{(2)} = \begin{matrix} 00 & -2 \cdot \alpha & \alpha & 0 \\ 01 & 1/\alpha & -(1/\alpha + \beta) & \beta \\ 10 & 1/\alpha & 0 & -(1/\alpha + \beta) \\ 11 & 0 & 1/\beta & -2/\beta \end{matrix}$$

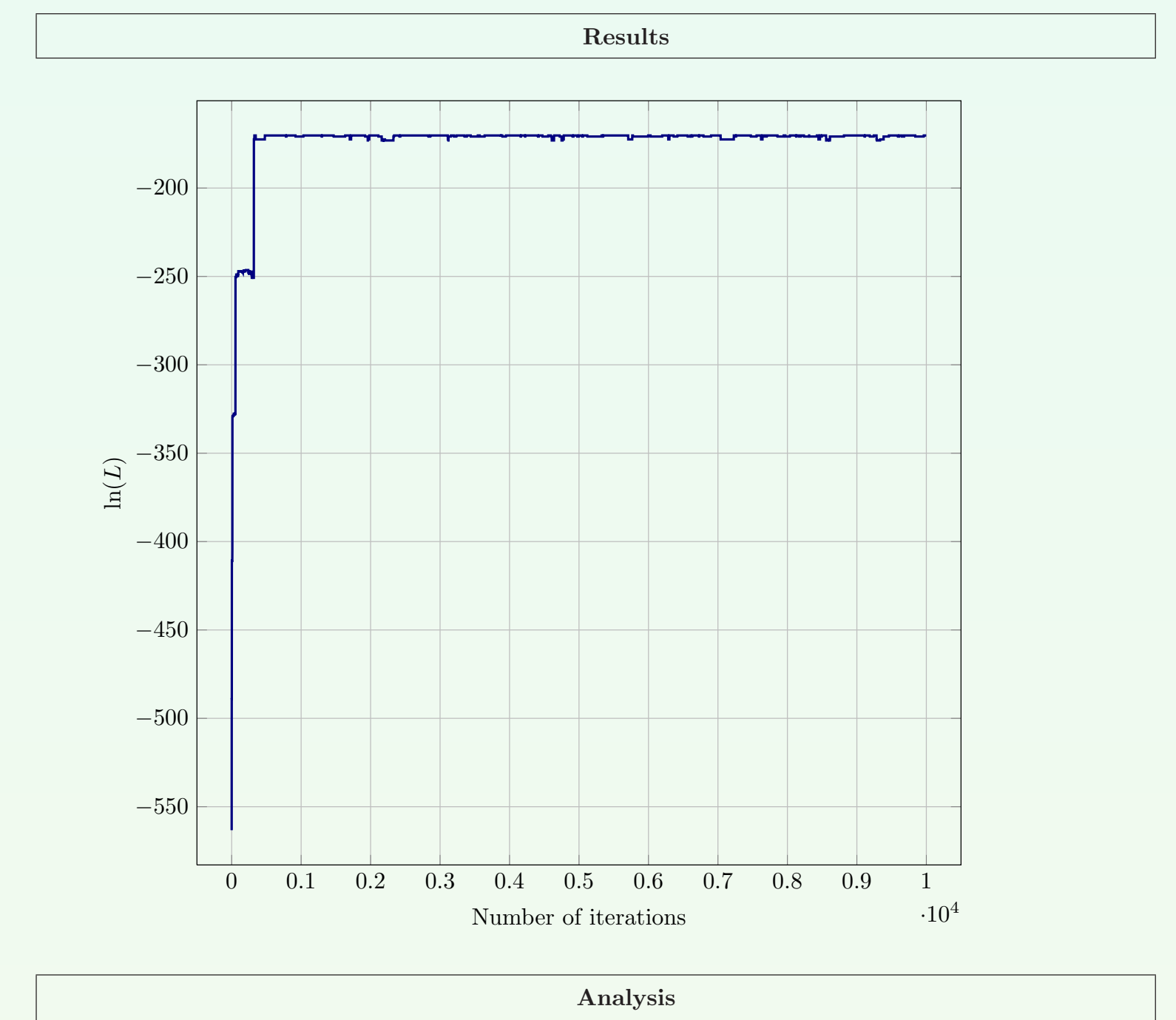
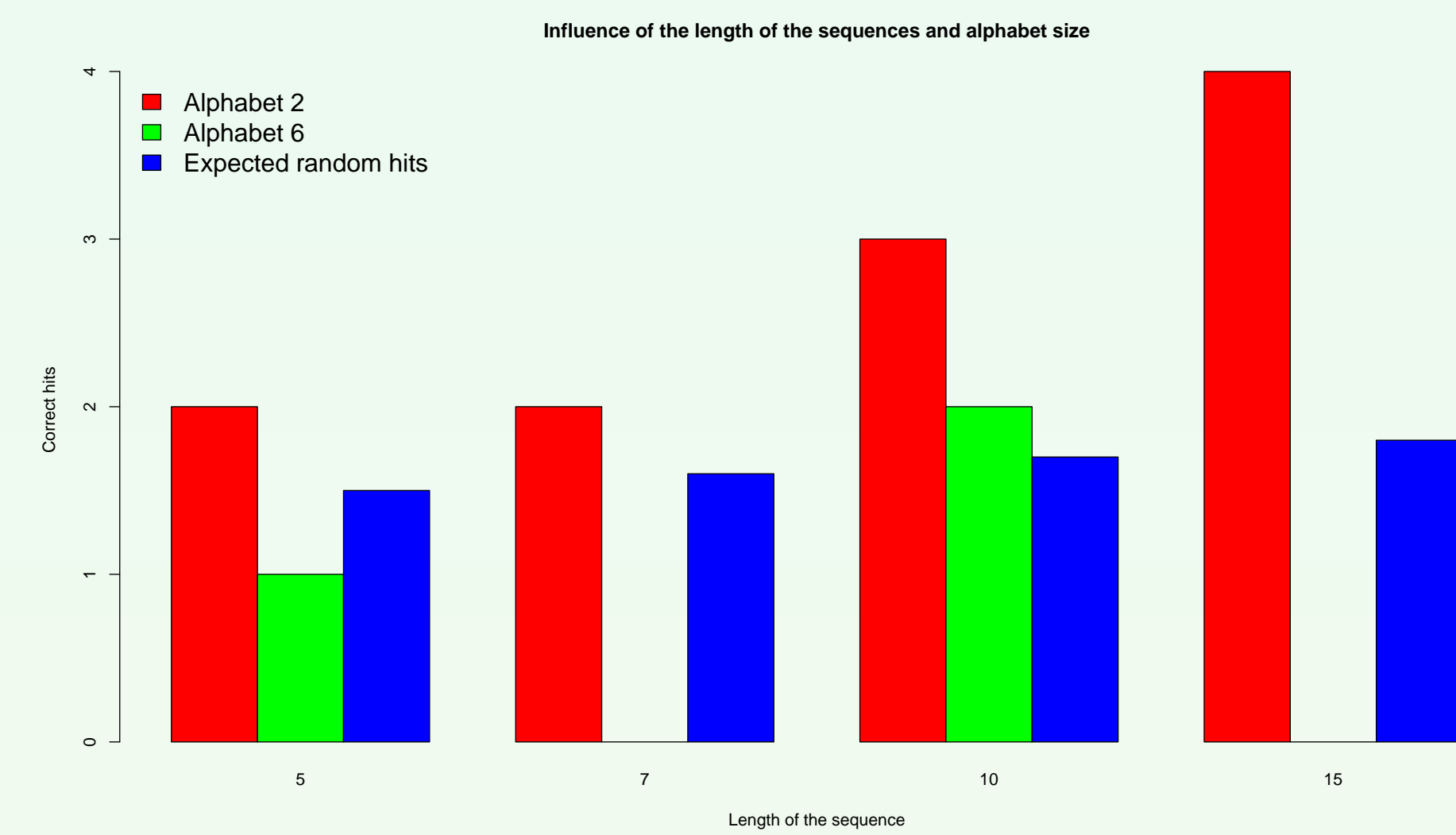
One fact that complicates calculations is that the pairs of our dependency structures always overlap at at least one site. Therefore calculating the probability of one sequence evolving into another cannot be simply calculated by the product of the transition probabilities for each pair, but by considering each pair and their neighbors as non-interacting subsystems [1]. We can then calculate the overall rate matrix $Q^{(L)}$. In the case of a linear dependency structure this is:

$$Q^{(L)} = \sum_{k=1}^{L-1} I_{|\Omega|^{k-1}} \otimes Q^{(2)} \otimes I_{|\Omega|^{L-k-1}}$$

In order to calculate the likelihood of a particular set of characters evolving into another set according to this pairwise model in a given time t , we need to compute the matrix exponential $P^{(L)}(t) = \exp(t \cdot Q^{(L)})$. By substituting in the different branch lengths, this can be extended to a set of sequences related by a phylogeny, using Felsenstein's pruning algorithm [2].

However, since the size of $Q^{(L)}$ grows exponentially with the sequence length, this process rapidly becomes infeasible, necessitating an alternative approach.

Results



The real dependency structure :
 (1,2) (2,3) (2,5) (3,4) (3,7) (5,6) (5,12) (5,13) (7,8) (7,9) (10,12) (11,12) (13,14) (13,15)

The predicted dependency structure by the frequency table :
 (11,12) (8,15) (1,2) (1,6) (2,3) (3,4) (5,9) (6,7) (7,13) (9,10) (10,14) (13,14) (5,12) (4,8)

After running the MCMC, we obtained a table with the counts of all visited edges of the dependency structure. With the help of this we were able to calculate a frequency table, from which we could construct a predicted dependency structure. To compare the true with the predicted structure we use the easiest criterion and just count the number of edges that they share.

We predicted the original dependency structure with different lengths of the sequences, the number of iterations was set to 10,000 and the sample size to 10. The longer the sequence gets, the more edges are predicted correct. The more important result is that the number of correctly predicted edges increases much faster than the number of correct edges in the random model. This means that the longer the sequence gets, the more significant the difference between our model and a random model is. In order to evaluate the change of the ratio of true edges to the total number of edges in time we would need more data for longer sequences.

Further work to be done

Many improvements to the existing model have to be made, especially dealing with insertions and deletions, to make it possible to test the model on real data. One idea would be to adapt the standard TKF91 model [3], which allows calculating the likelihood of sequences including indels. The quality of these predictions of dependencies could be assessed on data sets where tertiary protein structures are available. We often observed, that the MCMC got trapped in a local optima. In order to overcome this problem it might be worth considering advanced MCMC methods, like Simulated Annealing or Parallel Tempering. If it could be shown that this model also gives satisfying results with real data, one could think about incorporating this into existing statistical alignment programs as a measure of likelihood for the entire tree to further improve the predictions of these programs.

References

- [1] Lunter, G. and Hein, J. (2004). A nucleotide substitution model with nearest-neighbour interactions. *Bioinformatics* 4(20) Suppl 1:i216-23.
- [2] Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol.* 17(6):368-76.
- [3] Thorne, J. L., Kishino, H., Felsenstein, J. (1991) An evolutionary model for maximum likelihood alignment of DNA sequences. *J Mol Evol.* 33(2):114-24.

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 EU COGANGS (Comparative Genomics and next Generation Sequencing).