

DNA sequences coding for proteins typically evolve under a number of constraints, including the need to preserve structure, function and stability. As a result of these constraints, certain mutations increase the likelihood of other mutations within the same gene, and these correlations can be detected by examining multiple homologous sequences. There is great interest in detecting such *coevolution* in order to make inference about the structural or functional features of the protein that may be responsible for the correlations. The ultimate aim of many such studies is to predict structure from sequence data alone, and this approach has been successful in the prediction of RNA secondary structure (Knudsen and Hein, 1999) and thus the position of RNA genes in DNA sequences. However, these attempts for protein structures have been much less successful (Horner *et al.*, 2008). A common issue with all these approaches is that they take as input a single multiple sequence alignment, using this as the basis of the correlation calculations, and this can have a very strong effect on the resulting inference (Dickson *et al.*, 2010). On the other hand, most alignment algorithms assume that each residue evolves independently, such that there is an inherent contradiction between the two steps.

In this project we will seek to combine alignment and detection of correlated evolution, with the aim of doing a better job at both. Building on existing probabilistic models of molecular evolution, we will allow pairwise dependencies between columns in the alignment, leading to more realistic inference of homology, and an *alignment-free* measure of correlation.