

Methods For Detecting Co-Evolving Sites Across Proteins

Rob Ackermann

Computer Science, Virginia Tech

The covarion hypothesis of molecular evolution states that sites within a protein do not evolve at a fixed rate. Over time, both the amino acid at a site changes as well as the rate of substitution for that site. More specifically, sites are classified as being either conserved, variant, or temporarily invariant and over time can change which of the three pools they are in. The biological explanation for this is that when a protein folds certain non-adjacent sites interact, and as such a mutation in one site can affect the substitution rate of the site it is interacting with. Sites whose interactions with each other affect their rate of evolution are said to be co-evolving, or covarions for short. Figure 1 below shows an alignment of seven protein sequences. The two bold sites are examples of covarions: a mutation in the second site has allowed the first site to vary. A method of identifying covarying sites will allow for more accurate models of evolution, which in turn could assist in finding cures for genetic diseases.

```
RDNNPVVVLENELMYGVPFEFPPEAQSKDFLIPIGKAKIER
RDNNPVVVLENELMYGVPFEFPSEAQSKDFLIPIGKAKIER
RDNNPVMLENELMYGVAFELPAEAQSKDFLIPIGKAKIER
RDDNPVVMLECELMYGVAFELPTEAFSKDFLIPIGKAKIER
RDDNPVVVFLEQELMYGVPFEMSDESFSKDFVIPIGKAKIER
RDDNPVVVFLEEELMYGVPFEMSEEVFSKDFVIPIGKAKIER
RDDNPVVMLEGELLYGVPFEMSEQAFSKDFVVPIGKAKIER
```

Figure 1: A protein alignment with two covarying sites

The goal of this research is to develop software capable of predicting which of the three rate pools any particular site is a part of, as well as making suggestions as to which sites are interacting. Unlike previous studies, the software will be tested across several proteins instead of within only one, allowing the identification of covarying sites across multiple proteins interacting in a complex or pathway. The software will require only a sequence alignment and a phylogenetic tree as input. If provided, tree branch lengths will be taken into account.

We thus far have developed a method of detecting whether a site is conserved, variant, or temporarily invariant by using tree-corrected entropy values. In a separate algorithm, mutual information has been used somewhat successfully to detect which sites are co-evolving across proteins. A combination of these two methods is currently being developed to better predict co-evolving sites. Results will be compared with previously conducted research on interactions between the proteins in question, which have largely been based on analyzing three dimensional structures. All methods that have been implemented in software thus far run at more than acceptable speeds. A flexible and expandable software package has been put together which can be used to run current methods of analysis across multiple proteins.