Poster C-31

Evolutionary Patterns At Different Codon Positions

Authors:
Lee Bofkin (*Cambridge University, European Bioinformatics Institute*)
Nick Goldman (*European Bioinformatics Institute*)

Short Abstract: Functional constraints imposed by the genetic code affect evolutionary patterns at different codon positions. We analysed >5000 protein-encoding sequence alignments and assessed differences in common evolutionary parameters (e.g. evolutionary rate, base composition) at the three codon positions. Our results are important to understanding evolutionary processes and phylogeny inference.

Long Abstract:

The key unit of the genetic code is the codon, a nucleotide triplet that unambiguously codes for a single amino acid or the termination of a polypeptide chain. A single amino acid may be coded for by several codons, leading to redundancy in the genetic code. In terms of encoding a specific amino acid, the first codon position is more functionally constrained than the third codon position but less functionally constrained than the second codon position, on average.

Thus, protein-encoding DNA sequences have to encode a protein that functions according to the genetic code. Mutations that occur over evolutionary time are affected differently by selection depending on the effect they have on the protein, and the genetic code affects the changes we observe. Conversely, since we know the genetic code, we can learn about the functional pressures on proteins by observing changes and can potentially make better evolutionary inferences using phylogenetic models that are aware of the genetic code and its typical effects on protein-coding sequence evolution.

This study uses large amounts of data and applies a wide range of evolutionary models to quantify differences between the three codon positions accurately. Some of the factors we investigate have not been studied before; others are investigated more thoroughly than in previous studies. We quantify the differences in evolutionary rates, nucleotide frequencies, transition: transversion biases and the levels of heterogeneity of evolutionary rates at the different codon positions.

Maximum likelihood models were applied to over 5000 protein-coding multiple sequence alignments from the PANDIT database (Whelan et al., 2006). We test how well different models explain the evolution of these datasets using likelihood ratio tests, identifying the percentage of families that show significant differences in the estimates of the parameters of our models (e.g. evolutionary rate) at the different codon positions. The distributions of parameter estimates reveal fascinating aspects of the effects of the genetic code on the basic properties of codon evolution. Furthermore, we investigate how using different evolutionary models may affect our conclusions and assess the interactions between different parameter estimates of each model.

We present results that illustrate the following findings. On average, the third codon position evolves fastest and the second codon position evolves the slowest. The differences in evolutionary rates can be related to the propensity of mutations to cause synonymous or non-synonymous amino acid changes at the different codon positions and the properties of amino acids that are likely to result from different types of non-synonymous mutation. The third codon position has the least heterogeneity in evolutionary rate, followed by the first codon position; the second codon position has the most rate heterogeneity. Frequently, strong negative selection at the second codon position means that many such sites evolve very slowly and a small fraction of second codon position sites evolve much faster, causing high rate heterogeneity. Third codon positions tend to evolve at a relatively similar rate to each other, leading to reduced heterogeneity in evolutionary rate at the third codon position.

First and second codon positions have virtually identical distributions of transition: transversion biases but the third codon position has a much higher average transition: transversion bias. Transversions at the third codon position are much more frequently non-synonymous than transitions and subsequent selection against transversions leads to a high transition: transversion bias. The effects on the amino acid encoded caused by a transition or transversion is very similar at the first and second codon positions, hence their similar transition: transversion biases. In terms of nucleotide frequencies, the second codon position has a bias against the more mutagenic GC nucleotides. This result reflects selection for low rates of mutation at the second codon position. The first codon positions have a slight purine bias and third codon positions have a slight pyrimidine bias.

Our investigation has shown how the genetic code and amino acid exchangeabilities have affected the evolutionary properties of the three codon positions. The resulting distributions of parameter estimates could be used to help develop novel gene-finding models because we are now more aware of the nucleotide-level evolutionary patterns that we expect to see in protein-encoding sequences. We do not expect non-protein-encoding sequences to have evolved in such a manner. Furthermore, the distributions of the parameter values can be used to give appropriate prior distributions for model parameters in Bayesian studies. Currently, many Bayesian studies use flat or uniform prior distributions of ‘plausible’ parameter values to estimate the posterior distributions of these parameters. More realistic prior distributions, such as we have provided, may lead to more reasonable posterior distributions of parameter values for specific datasets.

References:

Whelan S, de Bakker P, Quevillon E, Rodriguez N and Goldman N, 2006. PANDIT: an evolution-centric database of protein and associated nucleotide domains with inferred trees. Nucleic Acids Research 34:D327-D331