Analysis of selection on mRNA secondary structure strength in protein-coding sequences within conserved protein families

Michael Peeri[1] and Tamir Tuller[1,2]

[1]Department of Biomedical Engineering, The Iby and Aladar Fleischman Faculty of Engineering and The Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, 6997801, Israel
[\2]Sagol School of Neuroscience, Tel-Aviv University, Tel-Aviv, 6997801, Israel

Email

Michael Peeri: mich1@post.tau.ac.il
Tamir Tuller: tamirtul@post.tau.ac.il

Abstract

The mRNA molecule can form secondary structure through short-range base-pairing interactions determined by the nucleotide sequence. These structures compete with other interactions of the mRNA strand and are suspected to influence many gene expression processes. In this study we attempt to study selection acting on mRNA secondary structure strength in finer resolution than done before, by analyzing families of orthologous proteins. Within a conserved protein family, the nucleotide sequences observed are the result of multiple selective pressures maintaining the folded protein's function, but also efficient and accurate translation, protein folding and degradation and other steps in the gene expression process. Focusing on an homologous position within the family, (i.e. at the nucleotides encoding a homologous amino-acid), we can assume many of these selective pressures act similarly on all members of the family, justifying their analysis as samples taken from a single distribution. We can therefore perform statistical tests (given sufficient data) to infer which selective pressures are needed to explain the observations at any homologous position. The strength of purifying selection on the amino-acid level can be measured against codon bias, mRNA secondary structure bias and other characteristics of the coding sequence to reveal how these processes are regulated by the coding sequence and answer the following questions:

- Which factors explain the huge variation between genes and regions within a genome?
- What are the relationships between different traits selected for in different regions of the coding sequences? do some traits tend to be selected together in the same regions?
- Does the coding sequence provide enough flexibility to allow arbitrary traits to be selected for in the same region, or are there trade-offs between traits?
- Does selection for secondary structure strength accompany specific protein domains or protein secondary structures?