THE TPI (TRNA PAIRING INDEX), A MATHEMATICAL MEASURE OF REPETITION IN A (BIOLOGICAL) SEQUENCE

Gaston H. Gonnet
Dept. of Computer Science, ETH Zurich
Swiss Federal Institute of Technology
ETH Zentrum RZ F2, CH-8092 Zurich

DNA sequences contain, among other information, the encoding of amino acids for proteins. Coding for the amino acids is redundant, that is most amino acids are coded by more than one codon (base-triplet). Usage of different codons coding for the same amino acids is called codon bias. Codon bias can be easily observed in most genomes, i.e. the probabilities of the codons is not uniform, quite skewed very often. The function of codon bias is still unknown, although error correction, DNA stability, speed of translation are usually quoted as possible reasons for it.

Different codons are translated to amino acids by different tRNA molecules. Depending on the species, codons are mapped to tRNA molecules, one-to-one or many-to-one, and tRNA molecules map to amino acids, again one-to-one or many-to-one.

One of the aspects of codon usage which is suspected of affecting the translation efficiency is whether tRNA molecules (or codons) are reused more or less frequently. To study this effect we have to design an index that will measure how much reuse of a particular tRNA (or codon) there is compared to random distribution. This is called the tPI. The tPI must be independent of particular (skewed) frequency distributions. In the end, the tPI is a probabilistic measure over a sequence of symbols from a finite alphabet.

We describe the formulation of the tPI which has the desirable properties. It is relatively straightforward to find a recursion formula to compute its values. Less straightforward is to compute the moments of its distribution, and even more complicated is to compute it efficiently. It should be noted that this is not purely a theoretical question, biologists want to compute the tPI of most sequences, so an efficient algorithm is required.

To make the computation more effective, we transformed the recursion formulas to have a desirable property that makes its computation require less

intermediate storage and hence tractable. tPI indices of entire genomes have been computed.