

POPULATION GENOMICS OF BACTERIA AND YEAST

Michael G.Sadovsky

Institute of computational modelling of SD RAS

msad@icm.krasn.ru

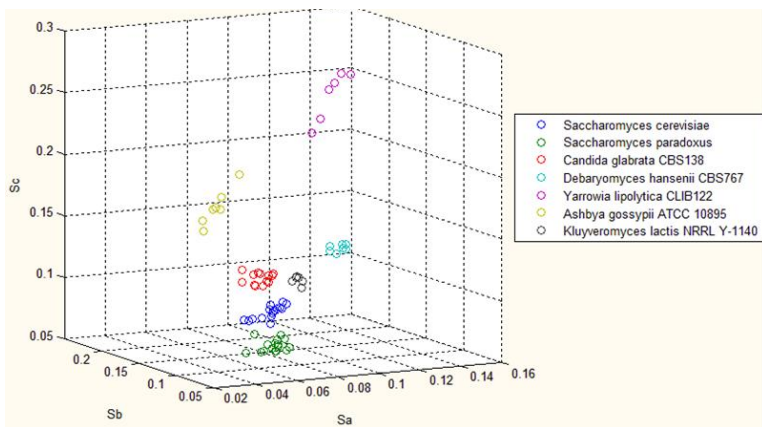
Population genomics is a rapidly growing new area of bioinformatics and system biology. It aims to reveal the population and evolutionary aspects of the genomic data observed over an ensemble of the entities. Here we propose three new indices of codon usage bias revealing some relation, both among genomes, and among the specific genes. The indices are based on the calculation of mutual entropy of codon usage frequency f_ω determined against three versions of “quasi-equilibrium” codon distribution \tilde{f}_ω :

$$I = \sum_{\omega=1}^{64} f_\omega \cdot \ln \left(\frac{f_\omega}{\tilde{f}_\omega} \right).$$

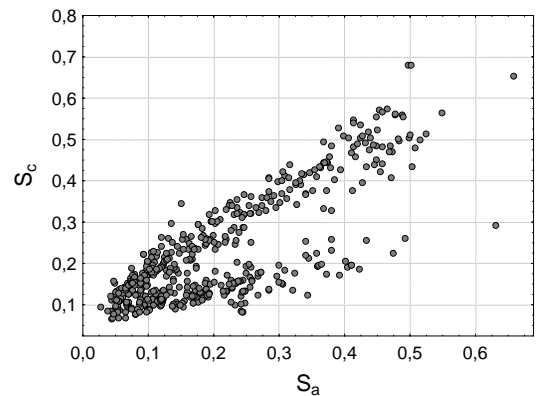
These latter are the following: locally equilibrium codon distribution, triplet distribution, and the most expected codon distribution.

The first quasi-equilibrium means that the frequency of synonymous codons are equal (forced to the arithmetic mean, within a group of synonymous codons), while the frequency of the relevant amino acid remains the same. The distribution of triplets calculated over the entire sequence (with no respect to a location of a triplet) is supposed to be the second reference distribution; and, finally, the third one is developed due to the principle of the the maximum entropy of the reconstructed frequency dictionary [1], that yields the most probable continuations of a dinucleotide originating a codon.

We studied the distribution of the yeast genomes within the space of these three indices. Fig. 1 shows the distribution of the chromosomes of seven yeast species, at the space. A distinct discretion of the genomes in this space has been observed. Moreover, a relative distribution of the genetic entities (chromosomes, of the genes families) may bring a new knowledge towards the relation between the organisms, or the functions peculiar to various gene families.



(a) Yeast Genomes



(b) Bacterial Genomes

Figure 1: Distribution of genomes in three-dimensional space determined by the indices.

Besides, a distribution of the bacterial genomes at the space of three indices has been studied. Totally, 532 bacterial genomes have been studied, from the point of view of the relative location at the space determined by the three indices. The distribution of the genomes within the space was found to look like a swallow tail. It fits quite exactly some plane at the space; two tails could be parameterized by the general C + G content of entire genome, quite precisely.

Both biological, and statistical (combinatorial) issues standing behind the patterns shown above are discussed.

References

- [1] Sadovsky M.G., Shchepanovsky A.S., Putintzeva J.A. Genes, Information and Sense: Complexity and Knowledge Retrieval // *Theory in Biosciences* (2008). V.127, P. 69–78.