

COMBINATORIAL PROPERTIES OF GENOMES

Eugenia A. Temlyakova[†], Victor M. Lougovoy[‡]

Siberian Federal University

[†]shotlife@gmail.com [‡]vitec90@mail.ru

The study of combinatorial patterns in completely sequenced genomes is one of the great importance in biophysics and bioinformatics. Many researches have been carried out in this area, still a lot is conspired in this field. Here we present some preliminary results of the study of combinatorial properties of genomes.

Consider a coherent DNA sequence of the length N . A word (of the length q) is any subsequence of that length found in the sequence under consideration; the set of all the words found within the sequence makes the q -support. Providing each word ω from the support with the number n_ω of its copies, one gets the finite dictionary W_q of the sequence. Changing the number n_ω for frequency $f_\omega = n_\omega/N$, one gets a frequency dictionary \widetilde{W}_q (of the thickness q). Any frequency dictionary W_q may yield the reconstructed frequency dictionary \widetilde{W}_q containing the most expected frequencies \tilde{f}_ω . Hence, the ratio

$$p_\omega = \frac{f_{\nu_1\nu_2\nu_3\dots\nu_q}}{\tilde{f}_{\nu_1\nu_2\nu_3\dots\nu_q}} \quad (1)$$

is the information value of the word ω .

A complementary palindrome is a couple of words (of the length q each) read identically in opposite directions, if the nucleotides in one of them are changed for dual ones, according to the Chargaff's complementary rule $A \Leftrightarrow T$ and $C \Leftrightarrow G$. The equivalence of the information values p_ω and $p_{\bar{\omega}}$ of the words ω and $\bar{\omega}$ combining a complementary palindrome may not be a matter of surprise, if a frequency dictionary was developed over two strands. Apparently, the symmetry is observed for the words enlisted at the frequency dictionary W_q developed over the main strand, solely; a single strand "knows nothing" about the existence of the second one. The symmetry described above is observed for any length of words. It is a well-known fact, that any symmetry is strongly related to a conservation law. Yet, there is no answer towards the meaning and functional role of the difference of the information value observed for the same words in various genomes.

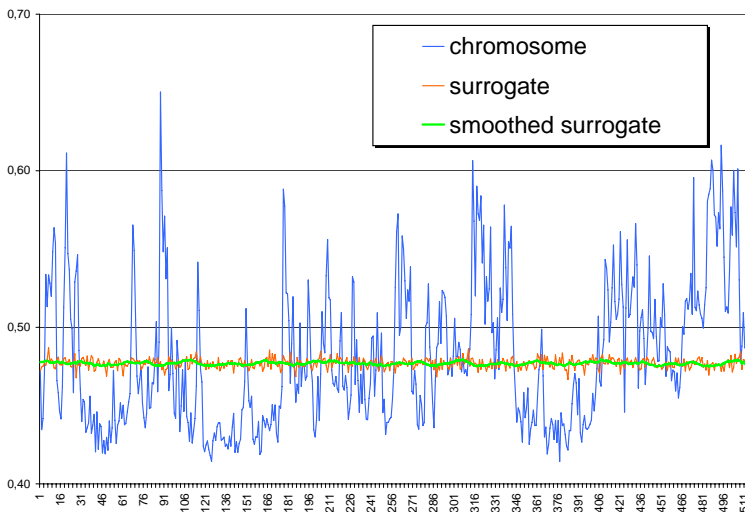


Figure 1: The discrepancy (2) of symmetry observed for human XIV chromosome, at $q = 8$.

been made due to the replacement of the synonymous codons within coding regions, and non-coding regions were intact, so that the translation remains the same.

Then, for these surrogate sequences the frequency dictionaries of the thickness q , $3 \leq q \leq 8$ have been generated, and the information value of the words were determined. Then, the measure μ of symmetry bias

$$\mu = \frac{1}{\|\Omega\|} \sum_{\omega} |p_\omega - p_{\bar{\omega}}| \quad (2)$$

has been calculated. Here Ω is the set of complementary palindromes observed within a sequence, and $\|\Omega\|$ is the cardinality of this set.

Surely, the growth of the length brings more violations of the symmetry. Finite sampling effect is the basic reason for a violation of the symmetry. Another reason for violation of the symmetry is the DNA structure features, themselves. This latter is the most important from the point of view of understanding the biological issues standing behind the effect of symmetry violation.

To figure out the finite sampling effect from the impact of nucleotide sequence features, we have carried out the computational experiments. We generated two types of surrogate nucleotide sequences; the former is developed through Bernoullian random process with the same frequencies of symbols, as observed at the real sequence, the latter has