

INFORMATION CAPACITY OF GENOMES

Anna A. Koval[†], Victor M. Lougovoy[‡]
Siberian Federal University

[†]fishka.v.banke@gmail.com [‡]vitec90@mail.ru

The analysis of statistical patterns in completely sequenced genomes is of a great interest. Despite of many researches carried out in this area, there are a lot of aspects which have not been studied yet. A complexity of patterns observed in a genetic sequence may vary significantly. The complexity itself is a matter of interests of mathematicians, biologists and biophysicists. Here we present some extreme properties of information capacity of genomes.

The set of all the words (of the given length q) observed at the sequence makes the support of that former (or q -support, if indication of the length is necessary). Providing each element of a support (i.e., each word ω) with the number of copies n_ω of that latter, one gets the dictionary of the sequence (of the thickness q). Changing the number of copies n_ω for frequency $f_\omega = n_\omega/N$, one gets a frequency dictionary W_q (of the thickness q).

A thinner frequency dictionary W_{q-1} constructed from the dictionary of the thickness q is unique, while the upward transformation yields several thicker reconstructed frequency dictionaries $\{W_q^{(1)}, W_q^{(2)}, W_q^{(3)}, \dots, W_q^{(k)}\}$. Thus, we construct the thinner frequency dictionary W_{q-1} from the real frequency dictionary and having it we reconstruct the thicker dictionary \widetilde{W}_q meeting the maximal entropy:

$$S_{\max} = - \sum \tilde{f}_\omega \cdot \ln \tilde{f}_\omega, \quad (1)$$

where $\tilde{f}_\omega \in \widetilde{W}_q$. The dictionary \widetilde{W}_q exists always, and the frequency \tilde{f}_ω is determined by the expression

$$\tilde{f}_{i_1, i_2, \dots, i_q} = \frac{f_{i_1, i_2, \dots, i_{q-1}} \times f_{i_2, i_3, \dots, i_{q-1}, i_q}}{f_{i_2, i_3, \dots, i_{q-1}}}, \quad (2)$$

where $\omega = i_1, i_2, \dots, i_{q-1}, i_q$.

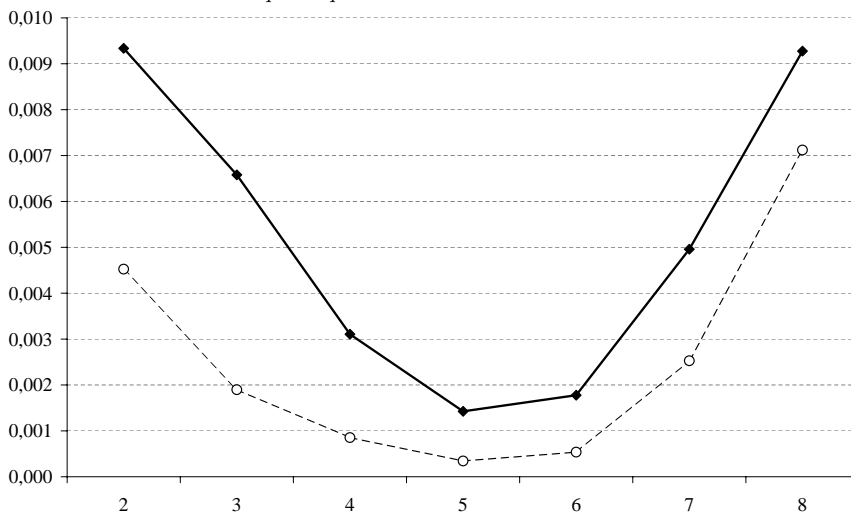


Figure 1: Information capacity (3) behaviour.

Information capacity is a measure of deviation of the reconstructed dictionary from the real one. We checked whether the information capacity has some extreme properties: it reaches the maximum for real DNA sequences. To verify this assumption, we compared the information capacity of the real DNA sequence to that latter of surrogate DNA sequences. Surrogate DNA sequences have been made as follows: codons from coding regions were replaced by synonyms ones with equal frequency and non-coding regions were intact, so that

the translation remains the same. As information capacity is the conditional entropy of the real frequency dictionary relatively to reconstructed frequency dictionary determined by the following expressions:

$$\tilde{S} = 2S_{q-1} - S_q - S_{q-2} \quad \text{and} \quad \tilde{S}_2 = 2S_1 - S_2, \quad (3)$$

where \tilde{S} is the mutual entropy of the frequency dictionary (of the thickness q), and S_q is absolute entropy of the frequency dictionary (see [1] for more details).

Figure 1 shows the behaviour of information capacity (3) for the third chromosome of *Pichia stipitis*, in comparison to the surrogate ones. There was ten surrogates generated, and the curve shows the averaged behaviour; meanwhile, the variance was very small. Here we implemented the randomization based on the equal frequency distribution of synonymous codons, while a number of other patterns of randomization might be implemented, as well.

References

- [1] Sadovsky M.G. Information capacity of nucleotide sequences and its applications // *Bulletin of Mathematical Biology* (2006). V.68, # 2. P. 156–178.