Generation of long open reading frames in prokaryotic genomes

Krystian Bączkowski

It is generally accepted that the development of new genetic information in organisms is carried out by modifications of existing information, e.g. by exon shuffling, the differentiation of gene after its duplication, lateral gene transfer, fusion and fission of genes. However, the genetic code and protein-coding sequences have some properties that enable to obtain a completely new genetic information. This is facilitated by reducing the probability of stop codon formation in alternative reading phases of protein coding sequences. It results from the low frequency of codons beginning from TA in protein coding sequences, which makes the complementary strand (read in the fifth phase of a gene) relatively poor in codons starting with these dinucleotides, including stop codons TAA and TAG. A similar effect is caused by decrease in the usage of some codons encoding serine and leucine that are complementary to stop codons in the fourth phase. As a result of these mechanisms relatively long Open Reading Frames (ORFs) can be generated in alternate phases of existing protein coding genes. Despite proposition of these mechanisms, no extensive studies were performed to answer the following questions: how significant is the role of these mechanisms in generating of new ORFs in genomes and which genomic properties are especially favorable for formation of new genetic information.

In order to find answers to these questions, many studies on over a million annotated nonoverlapping ORFs from 457 prokaryotic genomes have been conducted. Analysis of stop codon occurrence in all alternative phases of the studied ORFs showed that it is possible to create relatively long frames in these phases. In almost one third of the studied genomes, the average distance between stop codons for some alternative phases exceeded 300 nucleotides. However, the large length of a new ORF is not sufficient to become coding. The newly created frame must carry an information about encoded products. Analyzes conducted with the GeneMark program indicated that more than 8,000 alternative readings of annotated ORFs have a relatively high coding probability. Potential products of alternative reading frames showed even stronger similarity to real coding sequence (approximately several dozen percent in a given phase) in amino acid composition, average hydrophobicity, polarity, isoelectric point and the tendency to form secondary structures and transmembrane regions. If the studied sequences were generated in alternative phases, their reading in the correct phase should show similarity to

ORFs by that they were generated. In fact we find significant similarities to more than 200,000 analyzed sequences. Among them there were nearly 60,000 protein-coding sequences, which demonstrates the possibility of obtaining the function by the newly generated sequences. Further studies showed that conserved functional and structural elements of proteins such as domains or motifs are also often involved in the generation. Carried out analyzes allowed partly to explain the mystery of ORFans, i.e. ORFs that are specific to a particular genome or at most to several closely related genomes. A large part of these hypothetical frames was probably generated by protein coding genes in their alternate phases. The dissertation also presents several scenarios of ORF's generating in real examples of annotated sequences from different genomes.

All performed analyses show that the power of generation depends on alternative phase of generating sequence. New frames are usually generated in the fourth and fifth phase of annotated ORFs. Frames read in these phases characterized by the smallest number of stop codons. They also showed the greatest coding potential calculated in the GeneMark and the biggest number of found significant homologs. In addition, the potential products that were read in the fourth phase of annotated ORFs were the most similar to the real protein coding sequences especially in in the distribution of the isoelectric point and the tendency to form alpha helices and transmembrane regions.

It has been shown that the intensity of ORFs' generation process is strongly positively correlated with the global content of guanine and cytosine in studied genomes. The highest potential for generation of long ORFs have members of the Acidobacteria, delta-Proteobacteria, Deinococcus-Thermus and beta-Proteobacteria. The obtained results indicate that the identified mechanisms can operate on the large scale and the sequences (called alterlogs) arising from alternative readings of other sequences can easily obtain a new genetic information providing rapid evolution of possessing them organisms.

All received data were integrated in the relational database, which can be further expanded and updated. There is also a possibility to use it to perform additional analysis: to determine the way of overlapping of two ORFs in a genome, identify incorrectly annotated frames, search for processes of gene fusion and fission and genes that are involved in translational coupling process.