Abstract

In this work we present articles which are connected by the topic of linguistic-like approaches in genomics, which allow to treat genetic sequences as a "text" containing potential "words" (oligonukleotides, oligopeptides) of length *n.* Such an approach stands on the border of quantitative and qualitative analysis and, contrary to standard comparative bioinformatics methods, it is possible to compare phylogenetically distant individuals. Central article of my work (Zemková et al., 201..) is an analysis of peptide vocabularies of parasites and free-living organisms which showed significant differences in diversity of 4-6 amino acids long peptides of these compared groups. Parasites generally display reduction of pentapeptides, which is partly compensated by increased diversity of hexapeptides. This result is in accordance with our *a priori* hypothesis that parasites use immune evasion strategy to escape from MHC-based immunity system of its vertebrate host. Results also suggest that the length of key region for peptide recognition is about 4-5 amino acids and hence only short part of longer peptide bound in MHC participate on reaction with T-receptor.

In other two articles which arose as a product of cooperation with Prof Trifonov from the University of Haifa, we again used an analysis of genomic vocabularies. In the first article we detected potentially amphipathic structure. Distribution of these structures in proteome show some general regularities and they seem to be species – specific, so it could be possible to use them as so called "proteomic signatures" (Pietrokovski et al. 1990). Further we build an analogy between the preference of amphipathic peptides and the principle of its formation with the structure of words in human languages where consonant and vowels alternate due to natural constraints of pronounceability of their combinations.

Second article is dealing with possible reconstruction of ancestral sequences of DNA from recent human genome. An algorithm for potential reconstruction of the so-called generator sequence is suggested. The last provided article (Trifonov and Zemková, 2015) builds an analogy between the origin of DNA from simple tandem repeats of invasive character (Frenkel and Trifonov 2012; Trifonov and Bettecken 1997) and the origin of human speech from the utterances of so called *canonical babbling*. There is an extension of this metaphor of origin of life and language/ speech: In the history of science there is a parallel between natural sciences and linguistics. In both fields the living entities and languages were explained only from one invariant level of genome (Monod, 1971) or the language is considered to be an abstract system given by inner inherent scheme (Chomsky, 1986)