## Can proteins read letters?

Anatoly Sorokin, Evgenia Temlyakova, Timur Dzhelyadin, Svetlana Kamzolova
*Institute of Cell Biophysics RAS, 142290 Pushchino, Russia,* lptolik@gmail.com

Analysis of nucleotide sequences is ubiquitous in bioinformatics and structural biology. It is used to solve wide variety of problems: analyze evolutionary relationshipse between species and between proteins, predict protein functions and even bacterial lifestyle. DNA linguistics is widespread and provide us with a lot of reliable information, therefore we rely on four A, T, G and C letter so much that it really looks like proteins that process DNA "texts" can read.

Analysis of sequence specific physical properties of DNA double helix has a long history [1-4]. All properties of DNA double helix could be devided into three groups. First one is a long range properties, like electrostatic potential or duplex thermodynamical stability, that are influenced by base pair as far as several tens from point of consideration. The second group is medium range properties, like local geometry of double helix, that are controlled by sequence about two-six pairs away. The last group of physical properties are local properties, like coordinates of hydrogen bond donors, that depends upon “letter” at the point of consideration. So, in theory, all results obtained with text base methods and even more could be described with analysis of DNA physical properties. The benefit of physical properties over DNA sequence is that they give us better understanding and provide quantitative description of protein-DNA recognition mechanisms.

There is a view that in practice full analysis would require knowledge of physical properties for the protein that involved in interaction with DNA and that together with lack of dedicated resources preclude wide use of physical properties in analysis of gene functioning.

At the moment there are methods of prediction of bacterial promoters using profile of DNA electrostatic potential [5] with sensitivity and specificity comparable to methods that use DNA sequence [6].

We will show that there are cases when despite evolution matter is the DNA sequences, evolution driver is clearly the DNA structure and physical properties.

1. Yeramian, E. (2000). Genes and the physics of the DNA double-helix. *Gene*, *255*(2), 139–150.
2. Jensen, L. J., Friis, C., & Ussery, D. W. (1999). Three views of microbial genomes. *Research in Microbiology*, *150*(9-10), 773–777.
3. Choo, Y., & Klug, A. (1997). Physical basis of a protein-DNA recognition code. *Current Opinion in Structural Biology*, *7*(1), 117–125.
4. Oshchepkov et al. (2004). SITECON: a tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for site recognition. *Nucleic Acids Research*, *32*(Web Server issue), W208–12.
5. Polozov, R. V., Dzhelyadin, T. R., Sorokin, A. A., Ivanova, N. N., Sivozhelezov, V. S., & Kamzolova, S. G. (1999). Electrostatic potentials of DNA. Comparative analysis of promoter and nonpromoter nucleotide sequences. *Journal of Biomolecular Structure & Dynamics*, 16(6), 1135–1143.
6. E. A. Temlyakova, S. G. Kamzolova, A. A. Sorokin. (2012) Clustering of E.coli promoter electrostatic profiles. Proceedings of the 8th International Conference on Bioinformatics of genome Regulation and Structure\Systems Biology, p.318