

## [P28] Automatic workflow for the classification of local DNA conformations

Petr Čech<sup>1</sup>, Daniel Svozil<sup>1</sup>, Jaromír Kukul<sup>2</sup>, Bohdan Schneider<sup>3</sup>, Jiří Černý<sup>3</sup>

<sup>1</sup> *Laboratory of Informatics and Chemistry, ICT Prague, Technická 5, 166 28, Prague 6, Czech Republic.*

<sup>2</sup> *Faculty of Nuclear Sciences and Physical Engineering, CTU Prague, Trojanova 13, 122 00, Prague 2, Czech Republic.*

<sup>2</sup> *Institute of Biotechnology AS CR, v. v. i., Vídeňská 1083, 142 00, Prague 4, Czech Republic.*

A growing number of crystal and NMR structures reveal a considerable structural polymorphism of DNA architecture going well beyond the usual image of a double helical molecule. DNA is highly variable with dinucleotide steps exhibiting a substantial flexibility in a sequence-dependent manner. An analysis of the conformational space of the DNA backbone and the enhancement of our understanding of the conformational dependencies in DNA are therefore important for full comprehension of DNA structural polymorphism. A detailed classification of local DNA conformations based on the technique of Fourier averaging was published in our previous work [1]. However, this procedure requires a considerable amount of manual work. To overcome this limitation we developed an automatic classification method [2] consisting of the combination of supervised and unsupervised approaches. A proposed workflow is composed of  $k$ -NN method followed by a nonhierarchical single-pass clustering algorithm. We applied this workflow to analyze 816 X-ray and 664 NMR DNA structures, and we identified and annotated six new conformers. Using the workflow, dinucleotides with unassigned conformations can be either classified into one of already known 24 classes or they can be flagged as unclassifiable. The results illustrate the utility of machine learning approaches in the classification of local DNA conformations.

[1] S.D; et al. *Nucleic Acids Research* 36:11 (2008) 3690-3706.

[2] Č.P; et al. *BMC Bioinformatics* 14:205 (2013).