Dr. Ruslan Tetuev

Institute of Mathematical Problems of Biology RAS

HILBERT SPACES AND COMPUTATIONAL BIOINFORMATICS

THIS RESEARCH IS FUNDED BY RFBR, GRANT 15-29-07063

Maxim Pyatkov, Anton Pankratov, Ruslan Tetuev
Institute of Mathematical Problems of Biology RAS – the Branch of Keldysh Institute of Applied Mathematics of Russian Academy of Science
Keywords: huge approximate genetic repeats, spectral analysis

Here we are offering a new method to find huge approximate tandem and dispersed repeats in DNAs. A number of discrete methods have been developed and successfully applied to that problem with only comparatively small tandem repeats considered. Many of them are based on k-tuple match detection [1] or edit distances [2], but they all are not applicable for finding really inexact huge tandem repeats with a period over 1000 bp [3]. Apparently this size seems to be a serious limitation for all of the earlier approaches.

Since the methods of bioinformatics are often 'ported' from other fields of Computer Science (e.g. text/image data analysis) our main idea was to 'borrow' the solution for the same problem that has already been faced when image and audio data starts growing fast. To address this specific issue they created a joint technical committee that decided to switch from discrete methods to the spectral ones (i.e. switching from GIF/Tiff formats to JPEG, switching from WAV to MP3, etc.). In other words, all the matrices of discrete data should be now approximately represented as vectors of a Hilbert space.

Thus we were to develop a novel spectral-based approach for comparative analysis of genomic sequences, which was mainly devoted to finding huge and highly inexact repeats in DNAs. The novelty of our approach consisted in using specific 'spectral invariants' for genetic regions, which were based on spectral decomposition of DNA's profiles, for instance, GC-content. Those spectral invariants are actually vectors of Hilbert spaces, so they obey to all the basic rules for those vectors, which still could be easily interpreted in terms of our comparative analysis (inverted repeats, complement repeats, etc.). Altogether, all these mathematical facts are extremely useful while comparing DNAs and they could make it much easier to find similar genetic chunks across highly divergent species (and still in a little amount of time/space) [4,5,6].

[1] Benson G. Tandem cyclic alignment. Discret Appl Math. 2005; 146:124–133.
[2] Sokol D, Benson G, Tojeira J. Tandem repeats over the edit distance. Bioinformatics. 2007;23:e23-e30.
[3] [online] [access: 07.09.2017] http://tandem.sci.brooklyn.cuny.edu/Chromosomes.do
[4] Tetuev et al. Analytical methods in problems of recognition the structural and functional organisation of genetic sequences, The 2006 BGRS (Bioinformatics of the Genome Regulation and Structure) International Summer School for young scientists "Evolution, Systems Biology and High Performance Computing Bioinformatics", Novosibirsk, Russia July 12–15, 2006.
[5] Tetuev R.K. and Nazipova N.N. Consensus of repeated region of mouse chromosome 6 containing 60 tandem copies of a complex pattern. Repbase Reports. 2010. 10 (5). 776–776.
[6] [online] [access: 07.09.2017] http://mpyatkov.github.io/sbars/

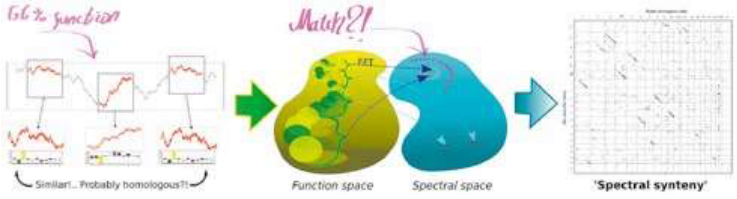

Figure 1. Understanding of 'Spectral Indexing' for lengthy genomic sequences and genome-wide visualization of local spectral similarity.