Computational Biology and Chemistry 53 (2014) 1–4

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Editorial

Editorial: Complexity in genomes

Two years ago, three of us (AP, YA, WL) organized a satellite meeting in the framework of the European Conference on Complex Systems (ECCS12) (Gilbert et al., 2014) focusing on genomic complexity. Although biological life on earth is one of the most complex systems, the field of complex system studies seems to mainly deal with physical systems where mathematical description, measurement, and modelling are traditionally addressed. The idea came to us that exploration of genomics in the framework of complex systems theory is needed in print, which led to this special issue.

In the literature, the term complexity (C) in genomes is used with several meanings. Some people use the number of genes in a genome to measure its complexity (Hahn and Wray, 2002). The C in "low complexity" (e.g. Wootton and Federhen, 1993) regions and the C in "more complex" genomes (e.g. Van Oeveren et al., 2011) are both caused by repetitive sequences, the only difference being that the repeat length is shorter in the former whereas the variety of repeats is larger in the latter case. Biological complexity is also a much debated concept (McShea, 1996; McShea and Brandon, 2010). Here we use the C-word more consistently and more generically: when used on an object, a process, a system, it means that it defies simple or traditional description – full of surprises, lacking single universal law, longer (Li and Vitányi, 1997) and/or time-consuming description (Bennett, 1988) in reproducing a copy, etc.

In biology and in genomics, just when one believes a universal law should cover all organisms at all time, exceptions are always discovered. For example, the central dogma (from DNA to mRNA to protein) was violated with the discovery of reverse transcription (Temin and Mizutani, 1970). The fact that a continuous stretch of DNA is transcribed into mRNA and this last is translated to protein in prokaryotes turned out to be untrue for eukaryotes (Chow et al., 1977). When it was commonly accepted that all biological functions are carried out by proteins, and protein-coding genes are the most meaningful part of the genome, the regulatory role of RNA was discovered (Fire et al., 1998; Morris and Mattick, 2014), and non-protein-coding regions are the focus of intensive studies in recent years (The ENCODE Project Consortium, 2012). The implication that evolutionarily conserved non-coding regions (Bejerano et al., 2004) must have a regulatory function faces the reality of high turnover rate of these regulatory elements (Dermitzakis and Clark, 2002). The list goes on.

It would be impossible to cover all hard-to-describe topics in genomics. What we aim in this special issue is to bring researchers who are comfortable with the theme of complexity in physical sciences to discuss genomes. A common thread of all papers here is the quantitative nature of the analysis, not merely a qualitative description. As early as 80 years ago, it was proposed that an institute should be established in which "biologists, chemists, physicists and mathematicians will cooperate in the future opening, and beneficial use, of the vast territory of quantitative biology" (Harris, 1933). Though we are still far away from outlining "complexity in genomes" as a field, just as "quantitative biology" not being a clear field for over 80 years, at least we bring those with a complex systems background to study genomics. There are 18 papers in this special issue, which can be roughly grouped into four categories.

DNA sequences as symbolic sequences: A large group of papers are treating DNA sequence from genomes as symbolic sequences, and apply techniques from time series analysis to study them (Cocho et al., 2014; Melnik and Usatenko, 2014; Papapetrou and Kugiumtzis, 2014; Provata et al., 2014b; Suvorova et al., 2014; Wu, 2014). This topic has its own historical surprises: the simplest description of a symbolic sequence is a random sequence, and the next simplest one is short-range-correlated sequences. However, DNA sequences as symbolic sequences were shown to be much more complicated, exhibiting long-range correlations (Li and Kaneko, 1992; Peng et al., 1992; Voss, 1992).

Provata et al. (2014b) is an extension of the work on human genome in Provata et al. (2014a) to other organisms. It exemplifies a typical approach in studying symbolic sequences: 4-nucleotide to 2-symbol conversions, dimer frequency and Markov transition probability, block entropy, symbol persistence properties, etc. Quantitative markers are extracted as indicatives of evolution, since organisms with different evolutionary paths are examined and compared. This collection of the basic statistics from DNA sequences is more accessible to readers who are less familiar with biology.

Cocho et al. (2014) attempts to explain the exponential correlation function observed in bacteria genomes by showing the roles played by different codon positions, by frame-shift, and by coding region size distributions. Without mixing statistics from different codon positions, the correlation between positions will be much weaker. Without a frame-shift between neighboring coding sequences, the correlation will not decay at all. And without a broad distribution of coding sequence length, the correlation function could be linear instead of exponential.

High-order Markov chains are mathematical models that add more complexity to the simple first-order Markov chain, with the goal of better fitting complex sequences. Both Melnik and Usatenko (2014) and Papapetrou and Kugiumtzis (2014) addressed

http://dx.doi.org/10.1016/j.compbiolchem.2014.08.003

high-order Markov models. In Melnik and Usatenko (2014), the relationship between memory functions and correlation function, which reduces the number of parameters in high-order Markov (Usatenko et al., 2009), is applied to DNA sequences.

In Papapetrou and Kugiumtzis (2014), the order of Markov models in DNA sequences is estimated by a technique proposed in Papapetrou and Kugiumtzis (2013). This analysis also shows a clear difference between those DNA sequences which can be modelled by a higher-order Markov chain, and those which can not (such as those with power-law correlation). In the latter case, the estimated Markov chain order does not converge with the increasing sequence length.

Finding hidden or latent periodicity in DNA sequences has a long history in bioinformatics, starting from the periodicity-3 signal in protein-coding regions (Fickett, 1982). Suvorova et al. (2014) compares the performance of several alternative periodicity-detection methods. They found that spectra-based methods tend to shift the signal to a shorter periodicity, whereas a direct matching and test of a fuzzy motif with a fixed length, called "information decomposition" (Korotkov et al., 2003), performs better.

Wu (2014) study concerns exact repeats (unlike the latent periodicity studied in Suvorova et al., 2014) in DNA sequences. In bioinformatics community, the most common tool in detecting repeats is the dot-matrix plot (Mount, 2013). In Wu (2014), such repeats are detected by the recurrence plots borrowed from the study of dynamical systems (Wu, 2004).

Spatial position and size distribution of functional units: The second large group of papers concerns the size and/or spacing distribution of genomic units (Dios et al., 2014; Gao and Miller, 2014; Muiño et al., 2014; Tsiagkas et al., 2014). If the biologically functional units (e.g. genes) are randomly distributed in the genome, the gap length follows negative binomial and geometric distribution, with an exponential trend. On the other hand, if the functional unit is larger than a single point on the chromosome with its own size, the simplest description of sizes is still an exponential distribution. In DNA sequences, the observed distributions for both gap distances and sizes follow mostly power-laws.

Gao and Miller (2014) focuses on the size distribution of orthologs obtained from sequence alignment. Such distribution for human-chimpanzee alignment tends to be exponential, whereas that for human-mouse alignment or multi-species ultraconserved regions (Bejerano et al., 2004) tends to be power-law distribution with an exponent of −4 (Salerno et al., 2006). These can also be compared to the distribution of paralogs (by genome self-alignment) which is power-law distribution with exponent −3 (Gao and Miller, 2011; Massip and Arndt, 2013). It is argued in Gao and Miller (2014) that orthologs from closely related species contain both a component from self-aligned paralogs and one from orthologs in distant species, so its distribution is a mixture as well.

In order to detect genome clustering, Dios et al. (2014) compares gap distances of genomic elements to the geometric distribution, continuing their earlier work (Hackenberg et al., 2011, 2012). On average, close to 30% of genomic elements in the human genome are found to be within clusters. Functional and regulatory elements (genes, CpG islands, transcription factor binding sites, enhancers) show higher clustering levels, as compared to DNase sites, repeats (Alus, LINE1) or SNPs. The clusters for all these elements form in turn high-level super-clusters, thus revealing a complex genome landscape dominated by hierarchical clustering.

Muiño et al. (2014) studies a clustering of cancer somatic mutations called "kataegis" (Greek word for "storm") (Nik-Zainal et al., 2012). The gap distance between mutations is bimodal, but the tail of the peaks falls off as a power-law function. Spatial clustering of somatic mutations may imply mutational hot spots, and the targeted hypermutated genes may provide new insight on cancer biology.

Tsiagkas et al. (2014) studies the gap distance between CpG islands, both those near genes and those away from genes (orphan CpG islands). Power-law distribution is again obtained similar to those of other functional units (Sellis et al., 2007; Sellis and Almirantis, 2009; Klimopoulos et al., 2012; Polychronopoulos et al., 2014). A simple evolutionary model based on segmental duplication is used to simulate a possible scenario to explain the data.

Intricacies in next-generation sequencing: The next group of papers concern the high-throughput (next-generation) sequencing (Gallo et al., 2014; Li and Freudenberg, 2014; Zhu and Zheng, 2014). The current sequencing biotechnology involves a mechanical breakage of genome into fragments, sequencing either the whole or the two ends of the fragment (the sequenced piece is called a read), and either aligning the reads back to the reference genome, if such a reference is available, or "de novo" constructing the genome sequence from overlapping reads. When one region of the genome is identical to another, that redundancy creates tremendous difficulties in either reads alignment or in de novo assembly.

Gallo et al. (2014) addresses a seldom discussed topic of hidden parameters in a de novo assembly. Using the SOAPdenovo program (Luo et al., 2012) as an example, Gallo et al. (2014) shows that assembly results can be altered if the parameter values are not chosen optimally, which can be a problem as many users of a de novo assembly program simply use the default setting. A particular ignored parameter is the k of k-mer length in the de Bruijn graph. The optimal choice of k is a function of fragment size, read length, and the level of redundancy in the genome.

Li and Freudenberg (2014) locates all exact repeats of length 1000 bases (kb) to the human genome, previously identified in Li et al. (2014). More than 1% of the human genome are covered by these unmappable 1000-mer reads. The unmappable regions are compared to those of twenty or so genomic annotations. About 4% of human genes overlap with these unmappable regions. And more than 90% of the unmappable regions were in the segmental duplicated regions (Bailey et al., 2002). On the other end, there is zero overlap between unmappable regions and the ultraconserved elements (Bejerano et al., 2004).

Zhu and Zheng (2014) does not attempt to align or assemble reads, but to identify a specific bacterium in a mixture of many bacteria (i.e., meta-genomes). Their approach is based on the idea that species-specific codon usage leads to characteristic k-mer frequencies in six reading frames of the coding region and in non-coding regions. Collecting k-mer frequencies from the reads, feeding them as inputs to a learning algorithm (Zheng and Wu, 2003), will indicate the presence or absence of specific types of bacterial genomes.

Specific biological and genomic topics: These papers are grouped together as they address specific biological applications: (Junier, 2014; Nikolaou, 2014; Pratanwanich and Lio, 2014; Zaghloul et al., 2014; Zuo et al., 2014).

Junier (2014) is an overview of different forces and mechanisms that shape the organization and structure of bacterial genomes at different levels. At protein level, interactions between amino acids determine the co-evolution of protein sequences. At genome level, genes cluster into operons, with complicated co-regulation and co-expression for various biological processes (transcription, translation, replication, cell division). Junier (2014) aims at discussing all relevant mechanisms in a single work.

Nikolaou (2014) explores biological explanation of a linguistic-motivated regularity in the genome, the Menzerarth's law at the gene-exon-base level (Li, 2012). The Menzerath law in this context states that if a gene contains more exons, the average exon size tends to be smaller. This Menzerath law was shown to be true for human genes (Li, 2012). Using mouse genes, Nikolaou (2014) shows that only genes with low conservation tend to follow the Menzerath law. These genes also tend to have less alternative splicing, fewer exons, and larger exon sizes.

Profiling genome-wide gene expressions at different conditions becomes easier by the microarray technology. Besides focusing on individual genes, more and more analyses focus on collection of genes such as genes involved in a given biochemical pathway. Pratanwanich and Lio (2014) investigates yet another method, called latent Dirichlet allocation (Blai et al., 2003), in the context of drug treatment, following a similar work using the Bayesian sparse factor model (Ma and Zhao, 2012). Although this work is within the scope of machine learning, the topic of multi-scales and multi-levels remain a favorite in the complex system study.

The finding of strand asymmetry at the replication origin in bacterial genomes (Lobry, 1996) led to search of GC or AT skew in the human genome (Brodie Of Brodie et al., 2005). The so-called skew-N domain is certain pattern in the skew series which is proposed as an indication of the replication origin (Touchon et al., 2005). Zaghloul et al. (2014) follows this long line of research to propose a new type of patterns in the skew series, called skew-split-N domains which is reminiscent of a letter N but split in half. Skew-N domains cover 1/3, whereas skew-split-N domains cover 12%, of the human genome. It is proposed that skew-split-N domains contain random replication initiations.

Zuo et al. (2014) overviews the authors' work on alignment-free phylogeny using composition vector of k-mers (Hao and Qi, 2004), implemented in the computer program CVTree (Xu and Hao, 2009). The large number of bacterial genomes being sequenced provides an opportunity to compare alignment-free phylogeny with the standard Bergey's Manual of Systematic Bacteriology (Garrity et al., 2001). The importance of subtracting the expected k-mer frequencies from $(k-1)$-mer data is emphasized. The effect of k on phylogenetic tree is discussed.

Admittedly, our collection of articles in this special issue is limited in scopes. We hope to attract more authors from a more diverse background if we produce a similar special issue in the future. However, there is no denying that biology is complicated and genomes are complex. François Jacob commented in his article "Evolution and tinkering" (Jacob, 1977): "natural selection does not work as an engineer works. It works like a tinkerer – a tinkerer who does not know exactly what he is going to produce but uses whatever he finds around him..." This ad hoc nature of the evolution, prolonged tinkering process, and the resulting imperfection, might be the root cause of complexity in genomes.

References

Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W., Eichler, E.E., 2002. Recent segmental duplications in the human genome. Science 297, 1003–1007.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., Haussler, D., 2004. Ultraconserved elements in the human genome. Science 304, 1321–1325.

Bennett, C.H., 1988. Logical depth and physical complexity. In: Herken, R. (Ed.), The Universal Turning Machine - A Half Century Survey. Oxford University Press, pp. 227–257.

Blai, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Brodie Of Brodie, E.B., Nicolay, S., Touchon, M., Audit, B., d'Aubenton-Carafa, Y., Thermes, C., Arneodo, A., 2005. From DNA sequence analysis to modeling replication in the human genome. Phys. Rev. Lett. 94, 248103.

Chow, L.T., Gelinas, R.E., Broker, T.R., Roberts, R.J., 1977. An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. Cell 12, 1–8.

Cocho, G., Miramontes, P., Mansilla, R., Li, W., 2014. Bacterial genomes lacking long-range correlations may not be modeled by low-order Markov chains: the role of mixing statistics and frame shift of neighboring genes. Comput. Biol. Chem. 53, 15–25.

Dermitzakis, E.T., Clark, A.G., 2002. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. Mol. Biol. Evol. 19, 1114–1121.

Dios, F., Barturen, G., Lebrón, R., Rueda, A., Hackenberg, M., Oliver, J.L., 2014. DNA clustering and genome complexity. Comput. Biol. Chem. 53, 71–78.

Fickett, J.W., 1982. Recognition of protein coding regions in DNA sequence. Nucleic Acids Res. 10, 5303–5318.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., Mello, C.C., 1998. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegan*. Nature 391, 806–811.

Gallo, J.E., Muñoz, J.F., Misas, E., McEwen, J.G., Clay, O.K., 2014. The complex task of choosing a de novo assembly: lessons from fungal genomes. Comput. Biol. Chem. 53, 97–107.

Gao, K., Miller, J., 2011. Algebraic distribution of segmental duplication lengths in whole-genome sequence self-alignments. PLoS ONE 6, e18464.

Gao, K., Miller, J., 2014. Human-chimpanzee alignment: ortholog exponentials and paralog power-laws. Comput. Biol. Chem. 53, 59–70.

Garrity, G., Boone, D.R., Castenholz, R.W., 2001. Bergey's Manual of Systematic Bacteriolog. Springer.

Gilbert, T., Kirkilionis, M., Nicolis, G., 2014. Proceedings of the European Conference on Complex Systems 2012. Springer.

Hackenberg, M., Carpena, P., Bernaola-Galván, P., Barturen, G., Alganza, A.M., Oliver, J.L., 2011. WordCluster: detecting clusters of DNA words and genomic elements. Algorithms Mol. Biol. 6, 2.

Hackenberg, M., Rueda, A., Carpena, P., Bernaola-Galván, P., Barturen, G., Oliver, J.L., 2012. Clustering of DNA words and biological function: a proof of principle. J. Theor. Biol. 297, 127–136.

Hahn, M.W., Wray, G.A., 2002. The *g*-value paradox. Evol. Dev. 4, 73–75.

Hao, B., Qi, J., 2004. Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance. J. Bioinform. Comput. Biol. 2, 1–19.

Harris, R.G., 1933. Introduction. In: Surface Phenomena, vol. I, Cold Spring Harbor Symposia on Quantitative Biology. Cold Spring Harbor Laboratory.

Jacob, F., 1977. Evolution and tinkering. Science 196, 1161–1166.

Junier, I., 2014. Conserved patterns in bacterial genomes: a conundrum physically tailored by evolutionary tinkering. Comput. Biol. Chem. 53, 125–133.

Klimopoulos, A., Sellis, D., Almirantis, Y., 2012. Widespread occurrence of power-law distributions in inter-repeat distances shaped by genome dynamics. Gene 499, 88–98.

Korotkov, E.V., Korotkova, M.A., Kudryashov, N.A., 2003. Information decomposition method to analyze symbolic sequences. Phys. Lett. A 312, 198–210.

Li, M., Vitányi, P.M.B., 1997. An Introduction to Kolmogorov Complexity and Its Applications, 2nd ed. Springer.

Li, W., 2012. Menzerath's law at the gene-exon level in the human genome. Complexity 17, 49–53.

Li, W., Freudenberg, J., 2014. Characterizing regions in the human genome unmappable by next-generation-sequencing at the read length of 1000 bases. Comput. Biol. Chem. 53, 108–117.

Li, W., Kaneko, K., 1992. Long-range correlations and partial $1/f^{\alpha}$ spectrum in a non-coding DNA sequence. Europhys. Lett. 17, 655–660.

Li, W., Freudenberg, J., Miramontes, P., 2014. Diminishing return for increased Mappability with longer sequencing reads: implications of the k-mer distributions in the human genome. BMC Bioinform. 15, 2.

Lobry, J.R., 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. 13, 660–665.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., et al., 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1, 18.

Ma, H., Zhao, H., 2012. FacPad: Bayesian sparse factor modeling for the inference of pathways responsive to drug treatment. Bioinformatics 28, 2662–2670.

Massip, F., Arndt, P.F., 2013. Neutral evolution of duplicated DNA: an evolutionary stick-breaking process causes scale-invariant behavior. Phys. Rev. Lett. 110, 148101.

McShea, D.W., 1996. Metazoan complexity and evolution: is there a trend? Evolution 50, 477–492.

McShea, D.W., Brandon, R.N., 2010. Biology's First Law: The Tendency for Diversity and Complexity to Increase in Evolutionary Systems. University of Chicago Press.

Melnik, S.S., Usatenko, O.V., 2014. Entropy and long-range correlations in DNA sequences. Comput. Biol. Chem. 53, 26–31.

Morris, K.V., Mattick, J.S., 2014. The rise of regulatory RNA. Nat. Rev. Genet. 15, 423–437.

Mount, D., 2013. Bioinformatics. Sequence and Genome Analysis, 2nd ed. Cold Springer Harbor Laboratory Press.

Muiño, J.M., Kuruǒgli, E.E., Arndt, P.F., 2014. Evidence of a cancer type-specific distribution for consecutive somatic mutation distances. Comput. Biol. Chem. 53, 79–83.

Nikolaou, C., 2014. Menzerath-Altmann law in mammalian exons reflects the dynamics of gene structure evolution. Comput. Biol. Chem. 53, 134–143.

Nik-Zainal S., S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., et al., 2012. Mutational processes molding the genomes of 21 breast cancers. Cell 149, 979–993.

Papapetrou, M., Kugiumtzis, D., 2013. Markov chain order estimation with conditional mutual information. Physica A 392, 1593–1601.

Papapetrou, M., Kugiumtzis, D., 2014. Investigating long range correlation in DNA sequences using significance tests of conditional mutual information. Comput. Biol. Chem. 53, 32–42.

Peng, C.K., Buldyrev, S., Goldberger, A., Havlin, S., Sciortino, F., Simons, M., Stanley, H.E., 1992. Long-range correlations in nucleotide sequences. Nature 356, 168–171.

Polychronopoulos, D., Sellis, D., Almirantis, Y., 2014. Conserved noncoding elements follow power-law-like distributions in several genomes as a result of genome dynamics. PLOS ONE 9, e95437.

Pratanwanich, N., Lio, P., 2014. Exploring the complexity of pathway-drug relationships using latent Dirichlet allocation. Comput. Biol. Chem. 53, 144–152.

Provata, A., Nicolis, C., Nicolis, G., 2014a. DNA viewed as an out-of-equilibrium structure. Phys. Rev. E 89, 052105.

Provata, A., Nicolis, C., Nicolis, G., 2014b. Complexity measures for the evolutionary categorisation of organisms. Comput. Biol. Chem. 53, 5–14.

Salerno, W., Havlak, P., Miller, J., 2006. Scale-invariant structure of strongly conserved sequence in genomic intersections and alignments. Proc. Natl. Acad. Sci. U. S. A. 103, 13121–13125.

Sellis, D., Almirantis, Y., 2009. Power-laws in the genomic distribution of coding segments in several organisms: an evolutionary trace of segmental duplications, possible paleopolyploidy and gene loss. Gene 447, 18–28.

Sellis, D., Provata, A., Almirantis, Y., 2007. Alu and LINE1 distributions in the human chromosomes. evidence of global genomic organization expressed in the form of power laws. Mol. Biol. Evol. 24, 2385–2399.

Suvorova, Y.M., Korotkova, M.A., Korotkov, E.V., 2014. Comparative analysis of periodicity search methods in DNA sequences. Comput. Biol. Chem. 53, 43–48.

Temin, H.M., Mizutani, S., 1970. Viral RNA-dependent DNA polymerase: RNA-dependent DNA polymerase in virions of rous sarcoma virus. Nature 226, 1211–1213.

The ENCODE Project Consortium, 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74.

Touchon, M., Nicolay, S., Audit, B., Brodie of Brodie, E.B., d'Aubenton-Carafa, Y., Arneodo, A., Thermes, C., 2005. Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins. Proc. Natl. Acad. Sci. U. S. A. 102, 9836–9841.

Tsiagkas, G., Nikolaou, C., Almirantis, Y., 2014. Orphan and gene related CpG Islands follow power-law-like distributions in several genomes: evidence of function-related and taxonomy-related modes of distribution. Comput. Biol. Chem. 53, 84–96.

Usatenko, O.V., Apostolov, S.S., Mayzelis, Z.A., Melnik, S.S., 2009. Random Finite-valued Dynamical Systems: Additive Markov Chain Approach. Cambridge Scientific Publisher.

Van Oeveren, J., de Ruiter, M., Jesse, T., van der Poel, H., Tang, J., Yalcin, F., Janssen, A., Volpin, H., Stormo, K.E., Bogden, R., van Eijk, M.J., Prins, M., 2011. Sequence-based physical mapping of complex genomes by whole genome profiling. Genome Res. 21, 618–625.

Voss, R.F., 1992. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. Phys. Rev. Lett. 68, 3805–3808.

Wootton, J.C., Federhen, S., 1993. Statistics of local complexity in amino acid sequences and sequence database. Comput. Chem. 17, 149–163.

Wu, Z.B., 2004. Recurrence plot analysis of DNA sequences. Phys. Lett. A 232, 250–255.

Wu, Z.B., 2014. Analysis of correlation structures in the Synechocystis PCC6803 genome. Comput. Biol. Chem. 53, 49–58.

Xu, Z., Hao, B., 2009. CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genome. Nucleic Acids Res. 37, W174–W178, web server issue.

Zaghloul, L., Drillon, G., Boulos, R.E., Argoul, F., Thermes, C., Arneodo, A., Audit, B., 2014. Large replication skew domains delimit GC-poor gene deserts in human. Comput. Biol. Chem. 53, 153–165.

Zheng, W.M., Wu, F., 2003. In-phase implies large likelihood for independent codon model: distinguishing coding from non-coding sequences. J. Theor. Biol. 223, 199–203.

Zhu, J., Zheng, W.M., 2014. Self-organizing approach for meta-genomes. Comput. Biol. Chem. 53, 118–124.

Zuo, G., Li, Q., Hao, B., 2014. On K-peptide length in composition vector phylogeny of prokaryotes. Comput. Biol. Chem. 53, 166–173.

Yannis Almirantis
Theoretical Biology and Computational Genomics Laboratory, Institute of Biosciences and Applications, National Center for Scientific Research "Demokritos", Athens, Greece

Peter Arndt
Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany

Wentian Li
Robert S Boas Center for Genomics and Human Genetics, Feinstein Institute for Medical Research, North Shore LIJ Health Systems, Manhasset, NY, USA

Astero Provata
Statistical Mechanics and Complex Dynamical Systems Laboratory, Institute of Nanoscience and Nanotechnology, National Center for Scientific Research "Demokritos", Athens, Greece

Available online 19 August 2014