Triplet Frequencies Implementation in Total Transcriptome Analysis

Michael Sadovsky[1,2(✉)], Tatiana Guseva[2], and Vladislav Biriukov[2]

[1] Institute of Computational Modelling of SB RAS,
Akademgorodok 660036, Krasnoyarsk, Russia
msad@icm.krasn.ru
[2] Institute of Fundamental Biology and Biotechnology,
Siberian Federal University, Svobodny prosp., 79, 660049 Krasnoyarsk, Russia
dianema2010@mail.ru, vladislav.v.biriukov@gmail.com
http://icm.krasn.ru

Abstract. We studied the structuredness in total transcriptome of Siberian larch. To do that, the contigs from total transcriptome has been labeled with the reads comprising the tissue specific transcriptomes, and the distribution of the contigs from the total transcriptome has been developed with respect to the mutual entropy of the frequencies of occurrence of reads from tissue specific transcriptomes. It was found that a number of contigs contain comparable amounts of reads from different tissues, so the chimeric transcripts to be extremely abundant. On the contrary, the transcripts with high tissue specificity do not yield a reliable clustering revealing the tissue specificity. This fact makes usage of total transcriptome for the purposes of differential expression arguable.

Keywords: Order · Probability · Triplet · Symmetry · Projection · Clustering

1 Introduction

Transcriptome is a set of all the symbol sequences from $\aleph = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$ alphabet corresponding to the entire ensemble of RNA moleculae (of mRNA moleculae) found in a cell (or in a sample). In a genome deciphering, transcriptome sequencing, assembling and annotation goes ahead. The point is that one may not be sure a transcriptome is stable, in terms of the composition of the sequences mentioned above. Indeed, the set definitely depends on a tissue, on a development stage, on a life cycle stage, and many other factors.

Stipulating a stability of a genome in an organism, one may expect that various tissues exhibit different expression of genes; this is a common place for multicellular organisms, and may take place in unicellular ones, if different stages of a life cycle are considered. Such difference is claimed *differential expression*. This latter is essential in a study of various physiological processes run in an

I. Rojas et al. (Eds.): IWBBIO 2019, LNBI 11465, pp. 370–378, 2019.
https://doi.org/10.1007/978-3-030-17938-0_33

organism, and may tell a researcher a lot concerning some peculiarities in functioning of biochemical and genetic networks.

Total transcriptome is the ensemble of all RNA (or mRNA) sequences gathered regardless their origin, through a bulky source sampled from an organism, or a tissue, etc. Since some genes in specific tissues, or cells may be suppressed or yield lowered expression due to some other reasons, one may expect that total transcriptome make a useful tool for assembling of all the genes observed in a sample, if assembled totally. Here we checked this idea on the total transcriptome of *Larix sibirica* Ledeb.

So, the goal of the study was to compare the efficiency of a "help and support" in specific transcriptome assembling, through the implementation of the total one. To do that, we have sequenced, filtered and cleaned the reads, for four specific tissues: needles, cambium, shoot, and seedling. These four specific transcriptome have been assembled; simultaneously, a total set of reads has been obtained through merging of all four specific ensembles into a single one. Then the assembling of the (total) transcriptome has been carried out. Finally, we tried to compare the total transcriptome with four specific ones to see whether some improvement in assembling "bottle neck" transcripts in specific transcriptomes takes place, or not; speaking in advance, we found greater losses than profits, in such approach.

2 Materials and Methods

Sequencing of *L. sibirica* Ledeb. total transcriptome was carried out in Labouratory of forest genomics of Siberian federal university. Four groups of tissue specific read ensembles have been obtained separately: needles, cambium, shoot and seedling. Also, later we merged all the reads ensembles into a single one, and assembled the total transcriptome.

Real transcriptomes (both tissue specific, and for the total one) comprise the contigs of various lengths. Some figures characterizing the specific (as well, as the total one) transcriptomes are shown in Table 1; the table presents the figures for the longest contig ($L_{\max}$), average length of transcripts ($\langle L \rangle$), and total abundance of contigs in a transcriptome (M). All transcripts were longer 200 b. p.

For the proposes of the clustering and analysis of transcriptomes, we selected the subsets of contigs, in each specific transcriptome (including the total one). We took into the subsets sufficiently long contigs, only. The idea standing behind such selection is following: shorter contigs would yield rather abundant subsets of points (in 64-dimensional space) that are in local quasi-equilibrium: in other words, too many short contigs would have zero frequency of some triplets. Moreover, a greater number of triplets would be presented in a single copy, in a number of such shorter contigs, thus yielding a kind of quasi-equilibrium over the subspace determined by these triplets.

To avoid the above mentioned effect, we have eliminated shorter contigs. We comprise sufficiently long contigs, to carry out clustering and visualization

of the data. Table 1 shows the figures used to select the contigs involved into analysis: L_d is the cut-off length of the contigs, in each specific transcriptome. That former means that we selected the contigs longer than L_d; M_d figures show the abundances of the sets of selected longer contigs.

To gain the total transcriptome, the reads ensembles obtained for each specific tissue have been merged into a single ensemble, and assembling has been carried out [1,2]. Common idea in total transcriptome implementation is to enforce the coverage level of the genes expressed in various tissues, thus improving assembling of *de novo* sequence. Not discussing here an efficiency (quite arguable, frankly speaking), we just stress that a total transcriptome still is a good first step, in any genome deciphering being a kind of *mean filed* approximation.

2.1 Frequency Dictionaries

To analyze statistical properties of transcriptomes, we used a conversion of them into frequency dictionaries; in particular, we focused on triplet frequency dictionaries, only. Formally, a triplet frequency dictionary is the list of all triplets $\omega = \nu_1\nu_2\nu_3$ observed in a sequence $\mathfrak{T}$. This is the triplet frequency dictionary $W_{(3,1)}$. More generally, let t be the step of a move of the reading frame (of the length 3) identifying a triplet ω. Then the frequency dictionary $W_{(3,t)}$ is the list of triplets identified in $\mathfrak{T}$, if the reading frame moves along $\mathfrak{T}$ with the step t. Definitely, one gets t different triplet frequency dictionaries here: there are t different starting positions of the first location of the reading frame.

Further, we shall focus on the dictionaries $W_{(3,1)}$ and $W_{(3,3)}$. In such capacity, there could be 3 triplet frequency dictionaries of $W_{(3,3)}$ type. The analysis of statistical properties of transcriptome provided here is based on the fact that three different frequency dictionaries $W_{(3,3)}$ determined over coding part of a genome differ seriously from similar dictionaries determined over non-coding ones [3–6]. This difference stands behind the analysis.

We did not derive all three versions of triplet frequency dictionaries of $W_{(3,3)}$ type for the transcripts; instead, we developed the clustering of triplet frequency dictionaries expecting them to gather into the clusters corresponding to the phase (i.e. reading frame shift figure $t = \{0; 1; 2\}$) and strand embedment (leading vs. ladder).

2.2 Clustering and Visualization

We used freely distributed software *ViDaExpert* by Andrew Zinovyev (`bioinfo.curie.fr`) for visualization data. Also, K-means clustering technique [7] has been applied, to prove a structuredness in transcriptome data. To retrieve a structure pattern in transcriptome (any of them, enlisted above), each contig was converted into frequency dictionary $W_{(3,1)}$. Everywhere further we shall denote it as W_3; to distinguish different dictionaries, we shall use an upper index in square brackets: $W_3^{[j]}$, so that $f_\omega^{[j]} \in W_3^{[j]}$. Here $f_\omega^{[j]}$ is the frequency of a triplet ω. Well known Euclidean metrics

$$\rho\left(W_3^{[1]}, W_3^{[2]}\right) = \sqrt{\sum_{\omega=\mathsf{AAA}}^{\mathsf{TTT}} \left(f_\omega^{[1]} - f_\omega^{[2]}\right)^2} \tag{1}$$

has been used to determine a distance between two triplet frequency dictionaries $W_3^{[1]}$ and $W_3^{[2]}$, for clustering and visualization purposes.

Using *ViDaExpert* software, we considered the distribution of points corresponding to frequency dictionaries in three-dimensional projection; the choice of axes for the projection was carried out automatically, since we observed the distribution in three principal components (the first one, the second one, and the third one), mainly, not in triplets.

To prove (or disprove) visually observed clustering, we used K-means, provided by the same software. The choice of K was determined by the stability of clustering: we always started from $K = 2$ and stopped at $K^\star$ where clustering became unstable. Besides, we also used elastic map technique, for the purposes of visualization, mainly. Detailed description of that methodology could be found in [8–12].

2.3 Chargaff's Parity Discrepancy

Chargaff's parity rules stipulate several fundamental properties of nucleotide sequences describing a kind of symmetry in them. We used these rules to analyze the observed cluster patterns, in transcriptomes. Tot begin with, Chargaff's substitution rule stipulates that in double stranded DNA molecule nucleotide A always opposes to nucleotide T, and vice versa. Same is true for the couple of nucleotides $\mathsf{C} \Leftrightarrow \mathsf{G}$.

The first Chargaff's parity rule stipulates that the number of A's matches the number of T's with a good accuracy, when counted over a single strand; obviously, similar proximal equity is observed for C's and G's. Finally, the second Chargaff's parity rule stipulates a proximal equity of frequencies of the strings comprising complementary palindrome: $f_\omega \approx f_{\overline{\omega}}$. Here ω and $\overline{\omega}$ are two strings counted over the same strand, so that they are read equally in opposite directions, with respect to the substitution rule, e.g., $\mathsf{CTGA} \Leftrightarrow \mathsf{TCAG}$; see [13–18] for details.

Genomes differ in the figures of discrepancy of the second Chargaff's parity rule [19]; same is true for various parts of a genome. Thus, one can compare the transcriptomes in terms of this discrepancy. To do it, let's introduce that former:

$$\mu\left(W_3^{[1]}, W_3^{[2]}\right) = \frac{1}{64}\sqrt{\sum_{\omega=\mathsf{AAA}}^{\mathsf{TTT}} \left(f_\omega^{[1]} - f_{\overline{\omega}}^{[2]}\right)^2}, \tag{2}$$

where ω and $\overline{\omega}$ are two triplets comprising complementary palindrome. Here we must take into account both couples: $f_\omega^{[1]} - f_{\overline{\omega}}^{[2]}$ and $f_\omega^{[2]} - f_{\overline{\omega}}^{[1]}$, since they exhibit different figures, in general.

Formula (2) measures a deviation between two frequency dictionaries; thus, one may expect that two dictionaries $W_3^{[1]}$ and $W_3^{[2]}$ may comprise the triplets from the opposite strands, if $\mu \to 0$. An inner discrepancy measure determined within a dictionary is another important characteristics of a dictionary. To measure it, one should change the formula (2) for

$$\xi\left(W_3\right) = \frac{1}{32}\sqrt{\sum_{\omega\in\Omega^*}\left(f_\omega - f_{\overline{\omega}}\right)^2}, \tag{3}$$

where Ω^* is the set of 32 couples of triplets comprising complementary palindromes. Obviously, here $|f_\omega - f_{\overline{\omega}}| \equiv |f_{\overline{\omega}} - f_\omega|$. We shall use the figures determined by (2) and (3) for transcriptome analysis.

2.4 Mutual Entropy to Measure the Quality of Total Transcriptome

The key aim of this paper is to compare tissue specific transcriptomes vs. the total one. To do it, we implemented a measure based on the mutual entropy calculation of the reads distribution over contigs of the total transcriptome. Describe this point in more detail. We used four tissues to get the tissue specific transcriptomes: needles, cambium, shoot and seedling. Surely, the abundance of the reads sets is different, for various tissues. So great difference in the abundances of the reads ensembles gathered for different tissues must be taken into account, and we have done it in the following way.

Table 1. Some figures characterizing transcriptomes. $L_{\min}$ is the minimal contig length, $L_{\max}$ the maximal contig length, $\langle L\rangle$ is average contig length, L_d is the selection length, and M_d is the abundance of contig set taken into consideration and N_R is the reads set abundance.

Transcriptome	$L_{\max}$	$\langle L\rangle$	L_d	M	M_d	N_R
Needles	9880	354	1000	59317	1851	2 504 853
Shoot	17893	532	5000	590240	1754	23 986 314
Seedlings	11008	455	2500	174805	1943	8 698 074
Cambium	20596	497	5000	628197	1455	9 563 901

Let N_{needles}, N_{shoot}, N_{seeding} and N_{cambium} be the numbers of the reads, in each read ensemble, respectively. Let then change the numbers for frequencies of the tissue specificity, as it occurs in the joint set of the reads:

$$f_{\text{needles}} = \frac{N_{\text{needles}}}{N_{\text{total}}}, f_{\text{shoot}} = \frac{N_{\text{shoot}}}{N_{\text{total}}}, f_{\text{seedings}} = \frac{N_{\text{seedings}}}{N_{\text{total}}}, f_{\text{cambium}} = \frac{N_{\text{cambium}}}{N_{\text{total}}},$$

where N_{total} is the sum of all N's shown above. The figures of f_{needles}, f_{shoot}, f_{seedings} and f_{cambium} provide the background to study the difference between total transcriptome and the specific ones.

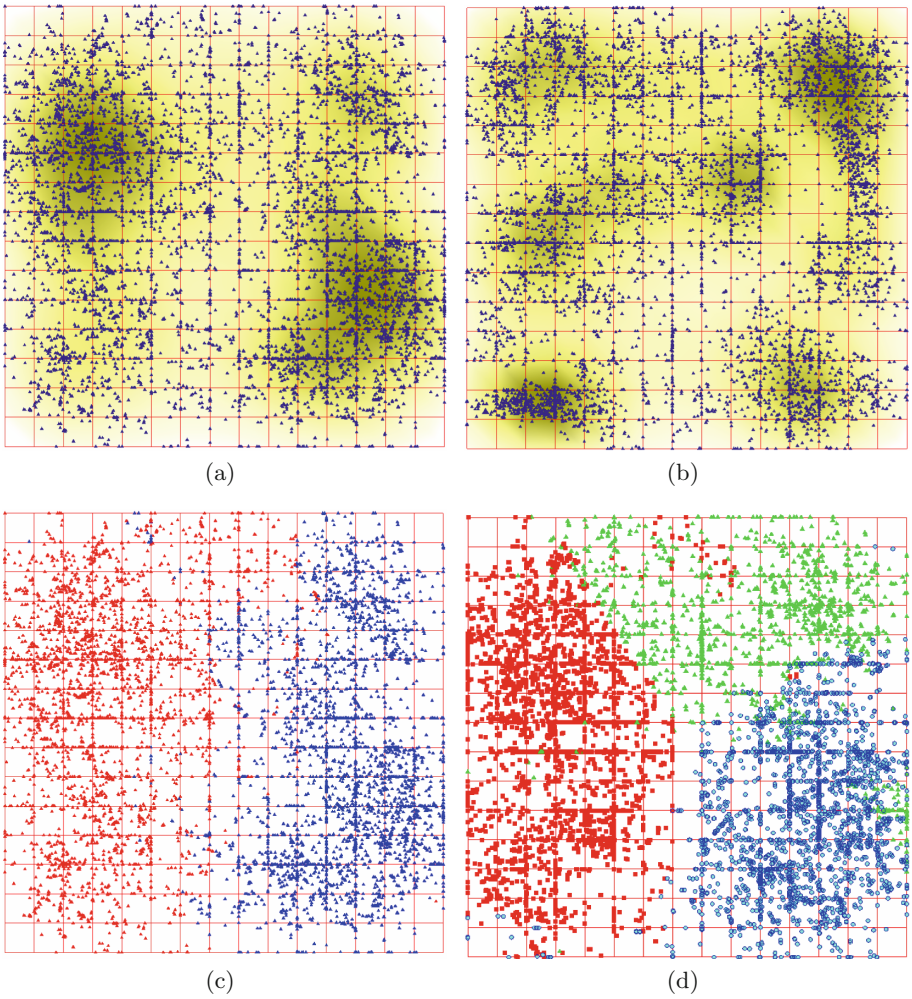


Fig. 1. Distribution of transcripts with greater mutual entropy (4) from total transcriptome; (a) is the case of $W_{(3,1)}$, (b) case of $W_{(3,3)}$. K-means is shown in (c) ($K = 2$) and in (d) $K = 3$; both cases are of $W_{(3,1)}$ type.

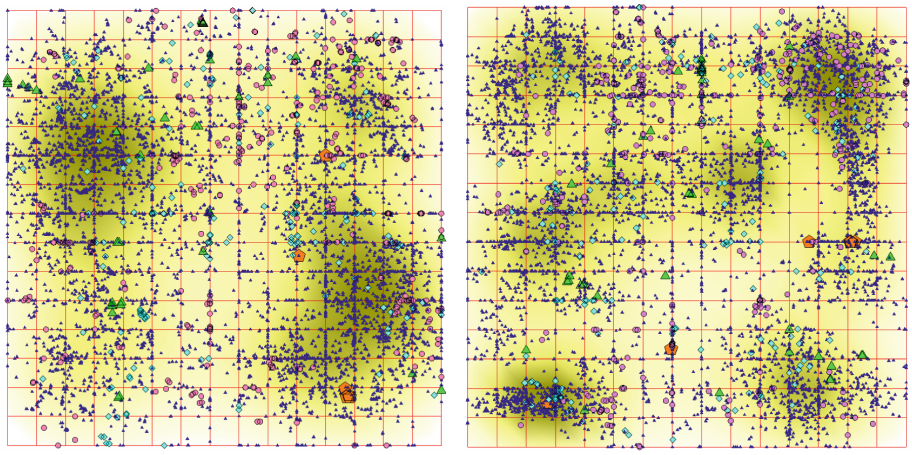
At the next stage, the numbers M_{tissue} (frequencies φ_{tissue}, respectively) of each tissue specific reads set observed over each transcript from the total transcriptome were obtained; to do it, we used back reads mapping over the total transcriptome transcripts. Thus, each transcript from total transcriptome was converted into a point in four-dimensional Euclidean space with the frequencies of tissue specific reads being the coordinates.

Finally, the mutual entropy

$$\overline{S}_k = \sum_{j=1}^{4} \varphi_j \cdot \ln\left(\frac{\varphi_j}{f_j}\right) \tag{4}$$

was determined for each transcript taken into consideration from the total transcriptome; here the index j enlists the tissues. The transcripts list was descending ordered, and the top part of the list has been analyzed. Index k in (4) enumerates the transcripts in the total transcriptome. Obviously, φ_j figures were determined for each transcript from the total transcriptome individually, while f figures were the same. Mutual entropy (4) measures a deviation of the distribution of the tissue specificity of reads observed within a transcript: if $\overline{S}_k = 0$, then the k-th transcript does not differ from the ensemble of the reads of the total transcriptome, and, in such capacity, is stipulated to be the most chimeric one. On the contrary, if a transcript yields the maximal deviation of (4) from zero, then it means the highest level of tissue specificity. It should be born in mind that the maximum of (4) depends on the specific tissue: in particular,

$$\max\left\{\overline{S}_k\right\} = -\ln f_k\,. \tag{5}$$



(a) (b)

Fig. 2. Distribution of contigs with higher preference of the tissue specific reads occurrence; the case of $W_{(3,1)}$ is left and the case of $W_{(3,3)}$ is right. (Color figure online)

3 Results and Discussion

The visualization of the total transcriptome (via transformation of sufficiently long contigs into triplet frequency dictionaries $W_{(3,1)}$ and $W_{(3,3)}$) reveals a structuredness in that latter. Figure 1 shows the distribution of the contigs. Apparently, there are two clusters in the Fig. 1(a) and six clusters in Fig. 1(b). The

clusters shown in Fig. 1 are provided by elastic map technique. Clustering with K-means for $K = 2$ and $K = 3$ is shown in Fig. 1(c) (two classes pattern) and Fig. 1(d) (three classes pattern). It should be said that these two patterns are very stable: more than 85 % of the runs of K-means converted to the same points distribution. The distributions provided by K-means with $K \geq 4$ were quite unstable.

Figure 2 answers the key question of the paper, whether the total transcriptome supports better assembling of tissue-specific ones, or not. Here we traced the distribution of the contigs with increased content of tissue-specific reads. To do that, we firstly identified the contigs with high level of mutual entropy (4), then checked what tissue reads prevail in a contig, and labeled it according to the tissue prevalence. Figure 2 shows the obtained distribution; here rosy circles represent cambium, green triangles represent needles and brown pentagons represent seedlings. Evidently, there is no preference in the tissue-specific enriched contigs over the clusters.

Also, Chargaff's discrepancies behaviour looks quite remarkable: for K-means classification with $K = 2$ the intraclasses discrepancies are $\xi_1 = 5.45 \times 10^{-4}$ and $\xi_1 = 5.90 \times 10^{-4}$, respectively, with the interclass discrepancy $\mu_{(1,2)} = 8.20 \times 10^{-4}$. Here the discrepancy between two classes seems to exceed those figures observed within a class. The situation is different, for $K = 3$. Here the intraclass discrepancies differ rather apparently, for three classes: $\xi_1 = 6.29 \times 10^{-4}$, $\xi_2 = 2.64 \times 10^{-4}$ and $\xi_3 = 5.91 \times 10^{-4}$, respectively. Obviously, the second class falls out of the general pattern of Chargaff's discrepancies. This fact may tell that the second class comprises the contigs from the opposite strands, unlike the first one and the third one. Same idea is supported by the figures of the interclass discrepancies; these are $\mu_{(1,2)} = 3.32 \times 10^{-4}$, $\mu_{(2,3)} = 4.27 \times 10^{-4}$, but $\mu_{(1,3)} = 3.71 \times 10^{-5}$.

That is a common place that a researcher is not guaranteed against the necessity to study total transcriptome, instead of a (tissue) specific one. Such situations may take place when a new (or rare) specimen is under analysis. Hence, one has to have a tool to evaluate the limits of knowledge that could be retrieved from the total transcriptome. Indeed, one may prefer to add sugar to a salty solution; others may want to add salt to a sweety sirup; nobody is able to distinguish the results. Meanwhile, significant number of chimeric transcripts may make a problem in analysis of a total transcriptome, say, in differential expression evaluation. If the tissue specificity of various reads is known *á priori* then one may eliminate the chimeric contigs from the ensemble due to specific entropy evaluation. The results presented above show some patterns revealed through clustering; this structuredness may be used for elimination of chimeric contigs. Nonetheless, the reliable approach to do it still awaits for further implementations.

References

1. Rahman, M.A., Muniyandi, R.C.: Review of GPU implementation to process of RNA sequence on cancer. Inf. Med. Unlocked **10**, 17–26 (2018)
2. Johnson, M.T.J., et al.: Evaluating methods for isolating total RNA and predicting the success of sequencing phylogenetically diverse plant transcriptomes. PLoS ONE **7**(11), 1–12 (2012)
3. Gorban, A., Popova, T., Zinovyev, A.: Codon usage trajectories and 7-cluster structure of 143 complete bacterial genomic sequences. Phys. A Stat. Mech. Appl. **353**, 365–387 (2005)
4. Gorban, A.N., Popova, T.G., Zinovyev, A.Y.: Seven clusters in genomic triplet distributions. Silico Biol. **3**(4), 471–482 (2003)
5. Sadovsky, M., Senashova, M., Malyshev, A.: Chloroplast genomes exhibit eight-cluster structuredness and mirror symmetry. In: Rojas, I., Ortuño, F. (eds.) Bioinformatics and Biomedical Engineering. LNCS, pp. 186–196. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-78723-7_16
6. Sadovsky, M.G., Senashova, M.Y., Putintseva, Y.A.: Chapter 2. In: Chloroplasts and Cytoplasm: Structure and Functions, pp. 25–95. Nova Science Publishers, Inc. (2018)
7. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press, London (1990)
8. Gorban, A.N., Zinovyev, A.: Principal manifolds and graphs in practice: from molecular biology to dynamical systems. Int. J. Neural Syst. **20**(03), 219–232 (2010). PMID: 20556849
9. Mirkin, B.: The iterative extraction approach to clustering. In: Gorban, A.N., Kégl, B., Wunsch, D.C., Zinovyev, A.Y. (eds.) Principal Manifolds for Data Visualization and Dimension Reduction. LNCS, vol. 58. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-73750-6_6
10. Gorban, A.N., Zinovyev, A.Y.: Fast and user-friendly non-linear principal manifold learning by method of elastic maps. In: 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015, Campus des Cordeliers, Paris, France, 19–21 October 2015, pp. 1–9 (2015)
11. Akinduko, A.A., Gorban, A.: Multiscale principal component analysis. J. Phys. Conf. Ser. **490**(1), 012081 (2014)
12. Mirkes, E.M., Zinovyev, A., Gorban, A.N.: Geometrical complexity of data approximators. In: Rojas, I., Joya, G., Gabestany, J. (eds.) Advances in Computational Intelligence. LNCS, pp. 500–509. Springer, Berlin (2013). https://doi.org/10.1007/978-3-642-38679-4_50
13. Mascher, M., Schubert, I., Scholz, U., Friedel, S.: Patterns of nucleotide asymmetries in plant and animal genomes. Biosystems **111**(3), 181–189 (2013)
14. Morton, B.R.: Strand asymmetry and codon usage bias in the chloroplast genome of Euglena gracilis. Proc. Nat. Acad. Sci. **96**(9), 5123–5128 (1999)
15. Forsdyke, D.R.: Symmetry observations in long nucleotide sequences: a commentary on the discovery note of Qi and Cuticchia. Bioinformatics **18**(1), 215–217 (2002)
16. Mitchell, D., Bridge, R.: A test of Chargaff's second rule. Biochem. Biophys. Res. Commun. **340**(1), 90–94 (2006)
17. Sobottka, M., Hart, A.G.: A model capturing novel strand symmetries in bacterial DNA. Biochem. Biophys. Res. Commun. **410**(4), 823–828 (2011)
18. Nikolaou, C., Almirantis, Y.: Deviations from Chargaff's second parity rule in organellar DNA: insights into the evolution of organellar genomes. Gene **381**, 34–41 (2006)
19. Albrecht-Buehler, G.: Fractal genome sequences. Gene **498**(1), 20–27 (2012)