Reads in NGS Are Distributed over a Sequence Very Inhomogeneously

Michael Sadovsky[1,2](✉), Victory Kobets[2], Georgy Khodos[2], Dmitry Kuzmin[3], and Vadim Sharov[3]

[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
victory.kobets@gmail.com, kalcifer@list.ru

[3] Institute of Space Research and Computer Sciences, Siberian Federal University,
Kirenskogo str., 26, 660074 Krasnoyarsk, Russia
{dkuzmin,vsharov}@sfu-kras.ru
http://icm.krasn.ru

Abstract. Distribution of read starts over a sequences genetic entity is studied. Key question was whether the starts are distributed uniformly and homogeneously along a sequence, or there exist some spots of the increased local density of the starts. To answer the question, 15 bacterial genomes have been studied. It was found that some genomes exhibit extremely far distribution pattern, from an homogeneity, while others show lower level of the inhomogeneity. The inhomogeneity level was determined through the Kullback-Leibler distance between the real string distribution, and that one bearing the most probable continuations of the shorter strings.

Keywords: Order · Digitalization · Entropy · Mutual entropy · Equilibrium

1 Introduction

Currently, the sequencing technologies are growing up rapidly. These technologies are both smart and complex, thus challenging researchers to figure out the issues resulted from biology, and those resulted from the technology details. These latter may be quite complicated and not obvious, at the first glance. A variety and abundance of the problems ranging from biological issues (so called "wet protocol") to computational and ever hard mathematical (i.e. assembling and the uniqueness of that latter) points hardly could be just outlined, not speaking about a comprehensive analysis. Here we focus on the specific problem that becomes acute due to the progress in sequencing and processing of genetic data.

I. Rojas et al. (Eds.): IWBBIO 2019, LNBI 11465, pp. 271–282, 2019.
https://doi.org/10.1007/978-3-030-17938-0_25

There are many tools and pipelines to assemble the read ensemble into a set of contigs, scaffolds, and further on. All of them are based on de Bruijn graph methodology [1–4]. Regardless the specific details of the assembling algorithm, all of them have one key idea standing behind the approach: the starts of reads obtained by a sequencer from a genetic entity are supposed to be distributed (almost) uniformly and homogeneously along the sequenced genetic sequence. Our paper aims to prove (or disprove) the validity of this supposition, through a simulation of read generation.

Coverage (local coverage, to be exact) $H_L(n)$ is the most common index of a quality of sequencing. It is defined as a number of unique reads covering a given nucleotide; here L stands for the length of reads (they are supposed to be of equal length, for simplicity). Evidently, this index is not expected to the same, for different fragments of a genetic sequence under consideration, that is why the local index should be introduced [3,4]. Obviously

$$\overline{H} = N^{-1} \sum_{n=1}^{N} H_L(n) \tag{1}$$

is the average (over a genome) cover index. A quality of sequencing output could be characterized with two figures: the former is the average cover (1), and the latter is its variance (or standard deviation) determined over a genome.

Indeed, that is a common place that the figure of the standard deviation of (1) is small, and a sequence is covered rather homogeneously by reads. Such homogeneity is not observed, in reality: as a rule, local cover is extremely inhomogeneous. Of course, the up-to-date algorithms and software platforms are able to process such inhomogeneous data flows, while it takes significantly greater time and resources. Reciprocally, the assembling quality becomes doubtful, not speaking about an uniqueness.

Here we aim to simulate a sequencer operation, in order to model the distribution of read starts over a sequence. Also, we study the patterns of real distribution and compare them to simulation ensembles, in order to find out the rules standing behind the distribution. Such rules are of great value for evaluation of an assembling quality, for any genome entity, and any sequencing machine and pipeline.

2 Study of the Real Distribution of Start Points of Reads Along a Genome

To begin with, we have studied the distribution of the real read starts along a genome sequence. To do that, we downloaded the assembled genomes and the reads ensemble. Then, we mapped the reads back over the genome, and fixed the positions of the starts of the reads. Mapping has been carried out with Bowtie 2 software. Two output files were developed, due to the mapping: the former was $\{0, 1\}$-sequence of the length N (here N is the length of a genome under consideration), and the latter was the sequence of integers m_j of the length N, $1 \leq j \leq N$, where m_j was the number of reads (of various lengths) starting at the j-th position.

Consider firstly a binary sequence obtained from mapping of reads over a genome. The key question here is whether zeros and ones are following in some (statistically revealed) manner, or they run randomly, with no order or pattern in their interlocation. Two approaches here should be explored:

1. Supposing the sequence of zeros and ones follows some probabilistic law, fit the sequence with some proper distribution function and identify the parameters of the distribution for further analysis;
2. Considering the sequence of zeros and ones as a symbol one, convert it into a series of frequency dictionaries $\{W_q\}$ of increased thickness q, $1 \leq q \leq q^*$ and figure out the most unexpected strings of the length q derived from the frequencies of the strings of the length $l < q$.

Here we follow the second approach that is completely similar to that one used to study the statistical properties of nucleotide sequences [5–11].

3 Simulation of Start Points of Reads: Theoretical Background

Let now describe the approach to study the statistical properties of the start points distribution in more detail. A digitalization described above converts a genome sequence into a symbol one, with two types of alphabet: the former is binary one $\{0, 1\}$, and the latter consists of M symbols, where M is the maximal number of reads starting at the same point, in a genome.

As soon as a genome is converted into a symbol sequence, it must be transformed into a series of frequency dictionaries $\{W_q\}$ of increasing thickness q. The thickness q is the length of words (strings) comprising a dictionary. More exactly, let q be the length of window that identifies a fragment in a sequence. Frequency dictionary W_q is the list of all the words (strings) observed within a sequence, so that each word ω in a dictionary is supplied with its frequency. The frequency

$$f_\omega = \frac{n_\omega}{N}, \tag{2}$$

where n_ω is the number of copies of a word ω, and N is the length of a sequence; to make the definition (2) feasible, one must connect a sequence into a ring, see details in [5–11]. Such closure results in appearance of $q - 1$ phantom words in a dictionary, while we neglect them.

Consider now the series

$$W_1, W_2, W_3, \ldots, W_{q-1}, W_q$$

of the frequency dictionaries in more detail. The key question here is the relation between the dictionaries observed in this series. Actually, a "downward" transfer (i.e., the transfer from W_j to W_{j-1} dictionary) is obvious: to do it, one must sum

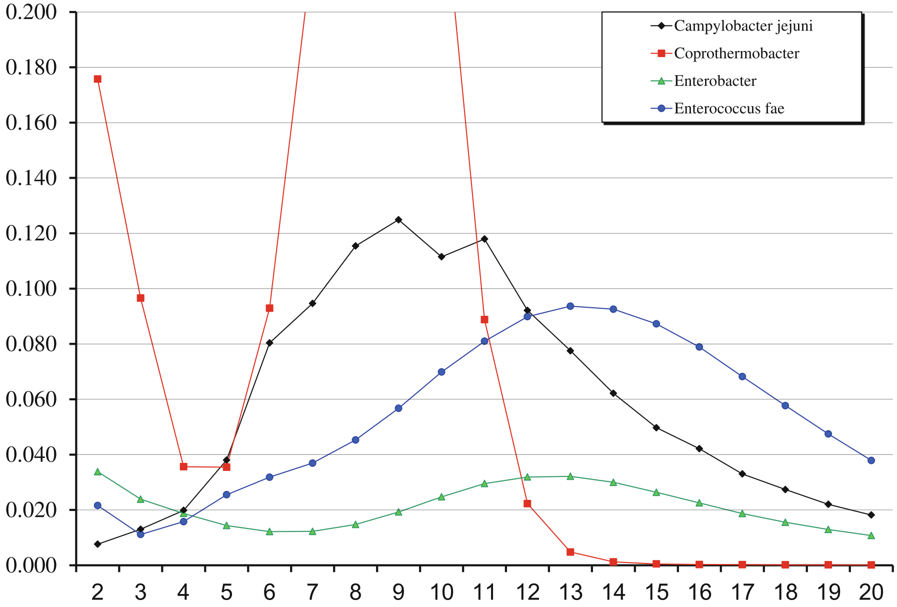

Fig. 1. Information capacity (5), vertical axis, determined for the symbol sequence representing the distribution of starts, with respect to the number of the starts observed in each nucleotide. Horizontal axis represents the thickness q.

up the frequencies of all the words differing in the first (or in the last) symbol[1]. The "upward" transfer $W_j \mapsto W_{j+1}$ is less evident.

Indeed, in general such transfer yields a family of dictionaries $\{W_{j+1}\}$, instead of a single one. Of course, the family contains the real frequency dictionary W_{j+1}, while there is no way to identify it. Simultaneously, there exists the specific frequency dictionary $\widetilde{W}_{j+1}$ in this family that comprise the most expected continuations of the words of the length j into the words of the length $j+1$. This specific dictionary (let's call it *reconstructed* one) exhibits the maximal entropy, among others comprising the family.

This extremal principle, together with the linear constraints of the "downward" transfer in a series of frequency dictionaries, yields the expected frequency explicitly:

$$\widetilde{f}_{\nu_1\nu_2\ldots\nu_{q-1}\nu_q} = \frac{f_{\nu_1\nu_2\ldots\nu_{q-2}\nu_{q-1}} \times f_{\nu_2\nu_3\ldots\nu_{q-1}\nu_q}}{f_{\nu_2\nu_3\ldots\nu_{q-2}\nu_{q-1}}}; \tag{3}$$

here we derive $\widetilde{f}(\omega_q)$ from $f(\omega_{q-1})$, see details in [5–11]. Finally, one must compare the real frequency dictionary W_q to that one bearing the most expected continuations: $\widetilde{W}_q$. To do that, the specific entropy of real frequency dictionary W_q against the reconstructed one must be calculated:

[1] The equality of these two sums stands behind the connection of a sequence into a ring.

Table 1. Information capacity (5) for the genomes *Acinetobacter baumannii* (1), *Clostridium autoethanogenum* DSM 10061 (2), *E. coli* K12 (3), *E. coli* O157 (4), *Saccharopolyspora erythraea* (7), *Stanieria spp.* NIES-3757 (8), *Staphilococcus aureus* NCTC 8325 (9), *Yersinia pseudotuberculosis* YPIII (10) with respect to the number of starts in each nucleotide.

q	1	2	3	4	7	8	9	10
2	0.006855	0.000151	0.017722	0.000194	0.001249	0.000003	0.006578	0.001581
3	0.007378	0.000131	0.010428	0.000194	0.001073	0.000004	0.008799	0.000969
4	0.009004	0.000318	0.006923	0.000620	0.001005	0.000006	0.012815	0.000998
5	0.009519	0.000899	0.004901	0.002075	0.001076	0.000008	0.013922	0.001346
6	0.009734	0.002810	0.005316	0.006755	0.001513	0.000012	0.015945	0.002269
7	0.010228	0.007856	0.005875	0.018833	0.001909	0.000018	0.020776	0.004497
8	0.011036	0.018508	0.005971	0.044770	0.001709	0.000024	0.031730	0.009182
9	0.013226	0.038805	0.005699	0.089915	0.001422	0.000042	0.050902	0.017498
10	0.017014	0.071194	0.005139	0.151169	0.001490	0.000052	0.078325	0.029912
11	0.022700	0.111794	0.004541	0.210797	0.001954	0.000066	0.109040	0.046115
12	0.030096	0.153165	0.003988	0.242604	0.002791	0.000077	0.134495	0.064515
13	0.038545	0.181569	0.003534	0.227371	0.003967	0.000105	0.148101	0.081751
14	0.047060	0.184315	0.003184	0.174954	0.005500	0.000132	0.141853	0.095160
15	0.054321	0.160808	0.002907	0.110638	0.007324	0.000151	0.120265	0.101762
16	0.059905	0.120647	0.002617	0.058855	0.009580	0.000176	0.089713	0.101649
17	0.062156	0.077580	0.002541	0.027123	0.012024	0.000208	0.059527	0.095004
18	0.060816	0.043650	0.002395	0.011079	0.014533	0.000228	0.035781	0.083791
19	0.057186	0.021745	0.002449	0.004268	0.017319	0.000249	0.019894	0.070384
20	0.050751	0.009841	0.002410	0.001649	0.020015	0.000303	0.010248	0.056214
N	4335793	4352205	4641652	5498578	8212805	5319768	2821361	4689441
Dth	98.40	198.67	203.79	234.70	102.74	108.72	188.51	84.03
σ	58.05	25.57	20.63	35.93	41.07	18.86	74.93	22.87

$$\overline{S}\left[\widetilde{W}_q|W_q\right] = \sum_{\omega\in\Omega} f_\omega \cdot \ln\left(\frac{f_\omega}{\widetilde{f}_\omega}\right). \tag{4}$$

Keeping in mind the expression (3), one gets

$$\overline{S}_q\left[\widetilde{W}_q|W_q\right] = 2S_{q-1} - S_q - S_{q-2}; \qquad \overline{S}_q\left[\widetilde{W}_2|W_2\right] = 2S_1 - S_2. \tag{5}$$

More details on these formulae could be found in [6–8].

Table 2. Information capacity (5) for the genomes *Acinetobacter baumannii* (1), *Clostridium autoethanogenum* DSM 10061 (2), *E. coli* K12 (3), *E. coli* O157 (4), *Saccharopolyspora erythraea* (7), *Stanieria spp.* NIES-3757 (8), *Staphilococcus aureus* NCTC 8325 (9), *Yersinia pseudotuberculosis* YPIII (10) for binary genome representation.

q	1	2	3	4	7	8	9	10
2	0.000016	0.171241	0.016968	0.000051	0.032699	0.000071	0.000002	0.000690
3	0.000005	0.092698	0.009659	0.000044	0.022948	0.000067	0.000002	0.000358
4	0.000039	0.027254	0.005648	0.000078	0.017313	0.000188	0.000002	0.000296
5	0.000017	0.006960	0.002577	0.000107	0.012300	0.000296	0.000002	0.000282
6	0.000027	0.002292	0.001462	0.000147	0.008861	0.000321	0.000003	0.000252
7	0.000045	0.001216	0.000894	0.000209	0.006493	0.000287	0.000004	0.000241
8	0.000039	0.000672	0.000522	0.000241	0.004674	0.000261	0.000004	0.000224
9	0.000033	0.000516	0.000390	0.000240	0.003582	0.000209	0.000007	0.000245
10	0.000096	0.000508	0.000331	0.000246	0.002880	0.000201	0.000012	0.000256
11	0.000082	0.000485	0.000296	0.000220	0.002359	0.000165	0.000017	0.000268
12	0.000139	0.000743	0.000341	0.000238	0.001891	0.000180	0.000024	0.000323
13	0.000280	0.001145	0.000431	0.000333	0.001753	0.000216	0.000034	0.000438
14	0.000509	0.001630	0.000674	0.000481	0.001693	0.000337	0.000046	0.000638
15	0.000989	0.002188	0.001256	0.000825	0.001897	0.000604	0.000057	0.001061
16	0.001938	0.002616	0.002020	0.001567	0.002697	0.001074	0.000067	0.001930
17	0.003840	0.003202	0.002553	0.003088	0.004363	0.001666	0.000082	0.003718
18	0.007741	0.003807	0.002690	0.006113	0.008357	0.002521	0.000097	0.007337
19	0.016023	0.004396	0.002607	0.012768	0.015387	0.003796	0.000115	0.015557
20	0.035304	0.004948	0.002472	0.028037	0.021780	0.005396	0.000145	0.034443

4 Results

We examined 16 bacterial genomes that meet the criteria; namely, we need the genome that

(1) sequenced by Illumina technology;
(2) are duly assembled and annotated, and
(3) there is a set of original reads available for the further analysis.

There are few non-bacterial genomes meeting these criteria; besides, a genome consisting of several chromosomes poses some other technical and essential problems, so we kept ourselves within the prokaryotic genomes, twelve entities in total.

Table 1 shows the data on information capacity (5) obtained for eight bacterial genomes. The genomes gathered in the table have rather smooth and similar pattern of the information capacity behaviour; the upper part of the table shows the data obtained for the digitalization with number of starts taken into account.

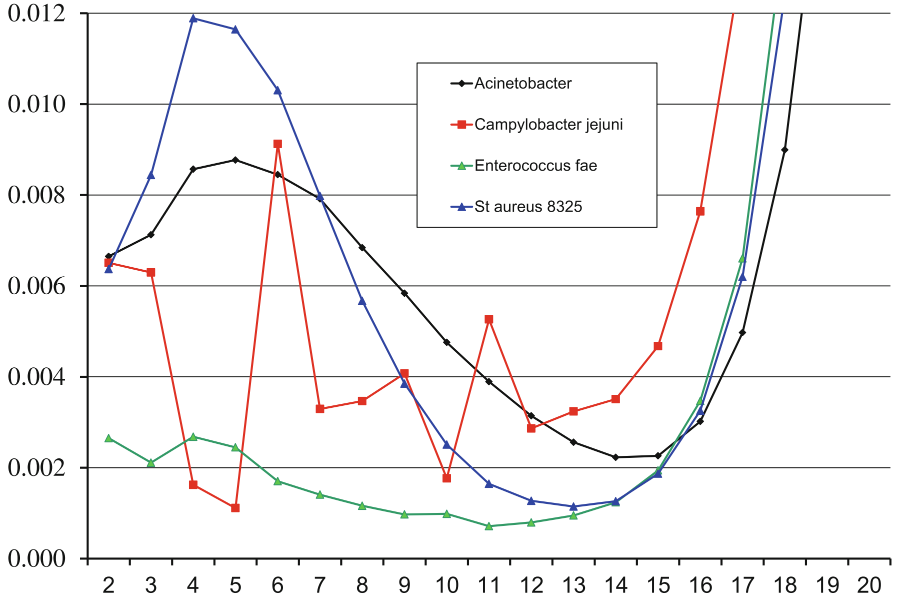

Fig. 2. Information capacity (5) determined for the binary symbol sequence representing the distribution of starts.

The lower part of that former shows similar data for binary digitalization (i.e. regardless the number of starts occurred in a nucleotide). Also, this table lengths of the bacterial genomes (denoted N), total genome cover depth (Dth), and the standard deviation for the set of local cover indices (1). All these figures are shown in the bottom of the table.

Figure 1 shows the patterns observed for four bacterial genomes; these genomes exhibit quite variable behaviour and rough pattern of the information capacity variation with the frequency dictionary thickness q growth. Similarly, Table 2 and Fig. 2 show the figures and pattern, respectively, for the same set of bacterial genomes, while transformed into a binary sequence each. *Coprothermobacter proteolyticus* genome yields a tremendous growth of information capacity (5) (see Fig. 1) with maximum figure of 0.404645, for non-binary digitalization. Reciprocally, the pattern of information capacity (5) observed for binary digitalization of the genome has four local minima; probably, these two observations make an evidence of the increased complexity of the reads starts distribution along a sequence.

Let now concentrate on Figs. 1 and 2. They show the behavioural patterns of information capacity (5), for four bacterial genomes each. First of all, all the curves are bell-shaped and it results from the finite sampling effect: an abundance of a frequency dictionary W_q grows exponentially, as q grows linearly. Hence, the greatest majority of the strings occur in a single copy, as q becomes great enough. Moreover, there exists specific figure q^* that yields no word occurred in two or more copies, at all; this figure makes a redundancy measure of the frequency dictionary of this thickness [11].

Figure 1 shows the pattern for the distribution of reads starts along a sequence, with respect to the number of the starts taken place in each nucleotide. The information capacity (5) of the frequency dictionaries of various thickness q reflects a predictability of a continuation of a word of the given length $q - 1$ into a word of the length q; if $\overline{S}_q \approx 0$, then all the words (the frequency of each word, to be exact) of the length q could be quite exactly predicted from the frequencies of the words of the length $q - 1$. The predictability goes worse, in general, as $\overline{S}_q$ grows up (see details in Sect. 5). Hence, the genomes shown in these figures exhibit quite low level of predictability of the distribution of the number of starts observed in a window of the given length q, as derived from the frequency ensemble of the starts numbers distribution observed in a shorter window.

Comparison of these two figures reveals significant smoothness in predictability of the starts numbers distribution, when counting it with respect to the specific numbers of starts observed in a nucleotide; probably, such behaviour comes from combinatorial reasons rather than from biology. Indeed, the specific numbers of starts taken into account for a dictionary implementation enlarge the alphabet capacity, thus cutting-off the tail of the distribution. Such cut-off manifests in a smoother pattern of the curve (5). Reciprocally, a multimodality of the distributions shown in these figures is of great interest. An occurrence of two (or more) local minima (and maxima, reciprocally) means an existence of some meso-scale structuredness in the starts distribution. The patterns shown in Figs. 1 and 2 differ in digitalization implemented for a study of the distribution of the reads starts numbers: the former shows the distribution with respect to the number of starts observed in a nucleotide, while the letter represents just the fact of a start, regardless to the specific number of reads starting in a nucleotide. There are only two common genomes in these Figs: *Campylobacter jejuni* and *Enterococcus faecalis* OG1RF; other genomes are different. It means that predictability of the strings representing the distribution of starts number is sensitive to digitalization version. In such capacity, those two genomes mentioned above exhibit the highest level of unpredictability in the starts numbers distribution along a genome.

Another interesting question concerns the variation of the number of starts to be observed in the same nucleotide, in different bacteria. Table 3 shows these data, for nine bacterial genomes. The table contains a union of the records for those genomes; blank cells in this Table mean that there was not a nucleotide with such number of starts, in the genome. Definitely, the greatest majority of nucleotides yields no start of a read; we shall not consider them. At a glance, the number nucleotides with multiple starts decreases, as that latter grows up (see Table 3). Here *E. coli* K12 genome completely falls out of the common pattern: it shows permanent and consistent non-monotony in the number of starts distribution. Moreover, it looks like a kind of a cycle of the length 2; some reasons of such behaviour are discussed below (see Sect. 5).

Table 3. *Acinetobacter baumannii* (1), *Clostridium autoethanogenum* DSM 10061 (2), *Saccharopolyspora erythraea* (3), *Staphilococcus aureus* NCTC 8325 (4), *Stanieria spp.* NIES-3757 (5), *Yersinia pseudotuberculosis* YPIII (6), *E. coli* K12 (7), *E. coli* O157 (8), *Enterobacter cloacae* (9).

n_s	1	2	3	4	5	6	7	8	9
1	835013	1421481	1449801	736737	213175	1387796	34206	1687399	298955
2	228570	595816	190999	319300	13853	429958	66101	907832	221129
3	47061	195018	21748	105585	488	103255	8093	391783	49898
4	9859	55149	3205	33897	27	21449	19111	152246	27531
5	1788	14236	837	9888	1	4103	2899	54214	6285
6	463	3457	274	3041		815	5695	18568	3314
7	129	812	99	868		182	1094	6256	866
8	52	201	29	311		70	1748	1940	519
9	22	34	6	96		39	378	728	197
10	9	11	2	48		22	546	213	123
11	6	1	1	12		13	152	72	66
12	3	1		8		12	169	22	40
13	4			4		11	44	11	17
14				2			62	5	23
15	1	1				2	28	1	12
16						1	25	1	11
17				1		1	5	1	8
18						2	2		1
19							6		2
20							2		
21							1		
22							1		2
23		1							2
24									1
25									1
26									2
27							1		
28							1		1
29							1		1
32							1		
35								1	
95				1					
116				1					

The table shows significant variation in the maximal number of starts found in a nucleotide; probably, this fact results from the peculiarities of sequencing procedure and may represent a quality of the sequencing rather than the biological issues. Extremely variable highest number of the starts (95 and 116 observed for *Staphylococcus aureus*) supports indirectly this idea. In general, the number of nucleotides giving the increasing starts number in a genome follows an exponential law: indeed, one may calculate the ratio of the numbers in two subsequent cells in Table 3 and find them to be quite proximal.

5 Discussion

The distribution of read starts along a nucleotide sequence is studied. This question is rather acute, since numerous inhomogeneities in this distribution may bring problems in assembling, annotation and further analysis of genetic entities. We explore the generalized approach to reveal some inhomogeneities in the starts distribution similar to [5–11]. Here a genome is considered as a symbol sequence, and we refrain from implementation of any biological knowledge "till the end"; in other words, we seek for the highly unexpected sites in the symbol sequences and the procedure is free from any biological knowledge. As soon as the sites are found, their biological role is studied. Basically, the hypothesis is that the sites tend to be located non-randomly, with a sounding preference to some biologically charged loci. It was found that the sites are distributed along a genome very non-randomly; whether the sites are located in the biologically important parts of a genome, still awaits for the answer.

The results provided above definitely show that the distribution of start points over a genetic entity is rather far from any equilibrium, or homogenous one. Any experimentalist knows that sequencing may skip some (rather extended) areas in a genetic sequence; the reasons of such distortion may follow both from biological issues of a matter, and from peculiarities of the sequencing technology. Here we tried to answer the question towards the character of this inhomogeneity in starts distribution.

To begin with, it should be said that the results shown above are biased. The problem may arise from the structure of reads ensemble. Indeed, we used the assembled genome, and the reads used to do it. The point is that the reads are obtained from both strands of DNA, while we used the leading one to align them. We used BowTie 2 to map the reads, and some of them might be mapped at the leading strand, while the have been sequenced from the ladder one. Thus, it might increase, to some extent, the number of observed starts (both unique, and multiple ones). The pattern of the number of starts distribution observed for *E. coli* K12 genome (see Table 3) proves indirectly this assumption. Hence, we plan to reconsider the starting points pattern with respect to the detailed analysis of the reads from the point of view of their strand origin.

To reveal the structuredness in the strings containing the nucleotides with various numbers of starts of reads, we used the idea of information capacity (3–5); this is an averaged measure telling on the distribution character in general, but nothing could be understood on individual level. To enhance the analysis, an idea of information valuable words [5–10] could be implemented. The idea is based on the detail analysis of (4) definition: if $\widetilde{f}_\omega \approx f_\omega$, then the corresponding term in the sum (4) is close to zero. On the contrary, the greatest contribution into the sum (4) is provided by the terms with the greatest deviation of $\widetilde{f}_{\omega}$ from f_{ω}. Such words are claimed **information valuable** ones.

So, the idea of further analysis is as following:

(1) Count the expected frequency $\widetilde{f}_{\omega}$ for each $\omega \in W_q$;
(2) Identify those with the deviation of $\widetilde{f}_{\omega}$ from f_{ω} exceeding some given level α;
(3) Match all such information valuable words over the genome, and check it against the annotation.

The hypothesis is that such words would match some peculiar sites, within a genome.

Another very important issue that falls beyond the scope of this paper, while is expected to be done soon is the approximation of the distribution of starts points located along a genome sequence with a number of various patterns, among them are Poisson distribution, LaPlace distribution, geometric distribution, negative binomial one, and many others. The idea is to fit the observed data best of all, with some specific distribution, so that some biologically sounding results might be retrieved from this fitting. In particular, the patterns shown in Figs. 1 and 2 support the hypothesis towards the feasible simulation of those distributions by Markov chains of the order 5 to 7, and around.

All these data and observations would be used for further simulation studies of the sequencing procedures implemented in various machines. Such simulation is of great value for better understanding of the details of assembling, annotation and comparison of sequenced genetic entities.

Acknowledgement. This study was supported in part by RFBR grant 18-29-13044/18.

References

1. Van Dijk, E.L., Auger, H., Jaszczyszyn, Y., Thermes, C.: Ten years of next-generation sequencing technology. Trends Genet. **30**(9), 418–426 (2014)
2. Li, H., Homer, N.: A survey of sequence alignment algorithms for next-generation sequencing. Brief. Bioinform. **11**(5), 473–483 (2010)
3. Buermans, H., den Dunnen, J.: Next generation sequencing technology: advances and applications. Biochimica et Biophysica Acta (BBA)—Mol. Basis Dis. **1842**(10), 1932–1941 (2014)
4. Conesa, A., et al.: A survey of best practices for RNA-seq data analysis. Genome Biol. **17**(1), 13 (2016)
5. Sadovsky, M.G.: Information capacity of nucleotide sequences and its applications. Bull. Math. Biol. **68**(4), 785–806 (2006)
6. Sadovsky, M.G.: Comparison of real frequencies of strings vs. the expected ones reveals the information capacity of macromoleculae. J. Biol. Phys. **29**(1), 23–38 (2003)
7. Sadovsky, M.G., Putintseva, J.A., Shchepanovsky, A.S.: Genes, information and sense: complexity and knowledge retrieval. Theory Biosci. **127**(2), 69–78 (2008)

8. Sadovsky, M.G.: Information capacity of symbol sequences. Open Syst. Inf. Dyn. **9**(01), 37–49 (2002)
9. Borovikov, I., Sadovsky, M.G.: Sliding window analysis of binary n-grams relative information for financial time series. In: Center for Advanced Signal and Image Sciences (CASIS) at LLNL 18th Annual Workshop, p. 1 (2014)
10. Sadovsky, M., Nikitina, X.: Strong inhomogeneity in triplet distribution alongside a genome. In: Ortuño, F., Rojas, I. (eds.) IWBBIO 2015. LNCS, vol. 9044, pp. 248–255. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16480-9_25
11. Bugaenko, N.N., Gorban, A.N., Sadovsky, M.G.: Maximum entropy method in analysis of genetic text and measurement of its information content. Open Syst. Inf. Dyn. **5**(3), 265–278 (1998)