STUDIA INFORMATICA 2015
Volume 36 Number 4 (122)

Mateusz GARBULOWSKI, ANDRZEJ POLAŃSKI
Silesian University of Technology, Institute of Informatics

A MODEL OF GENOME LENGTH ESTIMATION BASED ON K-MERS DETECTION[1]

Summary. The genome length estimation at raw sequencing data level gives a practical knowledge about size of the DNA sequence at early stage of analysis. In our research, we created a model based on random sampling of k-mer (very short DNA fragments), that we used to predict genome size. Furthermore, we made the comparison of model results with empirical whole-genome sequencing data.

Keywords: genome length estimation, genome size, sequencing model

OSZACOWANIE DŁUGOŚCI GENOMU NA PODSTAWIE DETEKCJI FRAGMENTÓW K-MER

Streszczenie. Oszacowanie rozmiaru genomu na podstawie surowych danych pochodzących z sekwencjonowania dostarcza wiedzy na temat długości DNA na wczesnym etapie analizy. W naszej pracy stworzony został model szacujący długość genomu oparty na losowym doborze krótkich fragmentów DNA zwanych k-mer. Wyniki powstałe przy użyciu modelu zostały odniesione do danych pochodzących z eksperymentów sekwencjonowania całych genomów.

Słowa kluczowe: szacowanie długości genomu, wielkość genomu, model sekwencjonowania

[1]Calculation were carried out using GeCONiI infrastucutre (POIG.02.03.01-24-099/13)

1. Introduction

Whole-genome sequencing methods [5, 6, 7, 8] are commonly used to read the genomic DNA structure of many living organisms. The DNA sequences consists of four basic units, called nucleotides: A – adenine, T – thymine, C – cytosine and G – guanine. The shotgun sequencing technique is based on creation of short sequences (called reads) by cutting genomic DNA in random places. The main measure that describe the process of reads creation is a depth of coverage [1, 3], which means how many nucleotides cover the genome in particular place. The depth of coverage is described as NL/G [1] where N is the number of reads, L is a mean length of reads (in bases) and G is the whole genome length (in bases). The problem, which we take on is how to estimate genome length knowing only N and L values. Algorithm, that we chosen is based on sampling k-mers (shorter than reads, very small fragments of exploring genome) which are in fact character subsets of read sequences (Fig. 1).

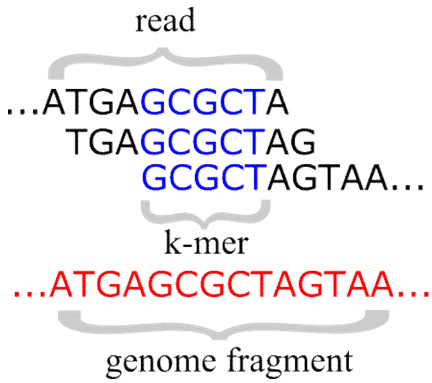

Fig. 1. Basic definitions in analysis
Rys. 1. Podstawowe pojęcia w analizie

New methods of sequencing [7, 8] (called by researchers as next-generation sequencing) are highly specialized in genome analysis. A lot of experiments during the last years, bring a lot of datasets, which may be useful to study length of the genome sequences. To produce those data, there are several sequencing platforms (fully-automatic devices) used to sequencing: Roche 454, HiSeq 2000 – Illumina, SOLiD System, Ion Personal, HeliScope and PacBio system. In spite of the different chemistry and base detection mechanisms there are two common steps for all methods: library construction (chemical preparation of sequences) and base detection. The most widely used platform is Illumina [7, 8] sequencer. Illumina technology uses sequencing by synthesis approach with bridge amplification (DNA set generation). A sequence length of reads in Illumina is about 50 to 200 bases (depending on device version). The main advantages of Illumina sequencing are: single per time, nucleotide detection and relatively high set of reads.

Such a large collection of data that whole genome sequencing produces may lead to errors [9] in order to nucleotide detection. The types of errors are different for each platform [9] e.g. for the Illumina and SOLiD most of mistakes types are substitutions (wrong nucleotide detection). Roche 454 reads contain a lot of deletion (missing of some nucleotide) and insertion (additional wrong nucleotide) and for the Helicos the main mistakes are deletions. This possible errors may lead to some inaccuracy in the analysis, so in exploring such a data with errors there may be a need to use some correction mechanisms.

The widely used .FASTQ [10] format stores all the reads sequences coming from whole-genome sequencing experiments. This data format contains two section. First part is used to storing reads sequences (line with A,T,C,G characters) and the second part contain quality PHRED scores of base calls represented in ASCII code. The PHRED describes probability error for each base.

The main objective of this work is to predict the genome size using only the raw whole-genome sequencing data, according to the assumptions of k-mer detection and knowing, that in empirical data may occur some errors. To check our method properties and correctness we made the model and then we check the real data with proper parameters. The main idea of our work and methods can be found in [1, 2].

2. Methods

2.1. A model algorithm

The model assumptions [1, 2, 3, 4] imitates the shotgun sequencing process to create raw sequencing dataset. To do the implementation of algorithm we use R environment in 3.1.0 version. As an input to create the dataset, there are couple of parameters that we can set, N as the number of reads, L as the length of reads, G to know what is the genome length at the beginning, N_k to declare the number of k-mers and at least k to set the k-mers length.

First step of algorithm is to generate collection of reads by setting above parameters. By making a set of reads we mean creating a N-element vector of L-length of sequences containing four basic nucleotides (A,T,C,G). The reads sequences are extracted from the human genome sequence stored in .FASTA file. Knowing this basic parameters we are able to count in easy way the depth of coverage, which is very important parameter in further analysis. According to [2] the article we assumed that the depth of coverage should be greater or equal to 2. This assumption is related to the sampling k-mers, when we have a relatively high depth of coverage then is bigger chance to find more significant k-mers e.g. when the

depth of coverage is 1, then we find k-mers one or zero times, so the prediction won't give the proper result, but on the other hand when the depth of coverage is 5 then the maximum number of found k-mers in reads will be also 5, what is statistically more important in analysis.

The next step of algorithm is to randomize the reads, which we use to gain the k-mers sequences and some random values to establish the beginning positions of k-mers. After receiving the k-mer sequences we checked the k-mer content in reads (defined for one k-mer sequence, as how many reads contain whole k-mer sequence). There may happen a situation, when we choose such a k-mer sequence that may occur more in one place of the genome. For example on the picture (Fig. 2) the number of k-mer CGCGC contained in reads should be 3 and chosen in first position, but there is another group of reads, where those k-mers occur. So in fact we get greater number of reads containing the k-mer, as in picture – 6.

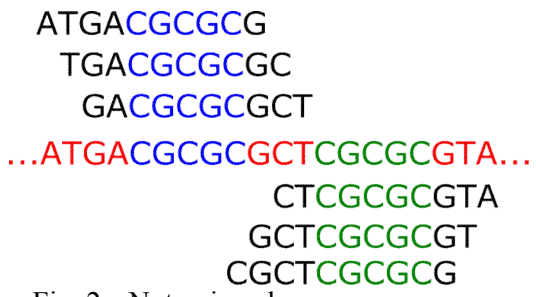

Fig. 2. Not unique k-mer sequences
Rys. 2. Powtarzalne sekwencje k-mer

Counting of non unique sequences, that in fact are describe as Poisson mixtures [1, 2] we treated as outliers. The outlier detection [4] which we applied is based on IQR rate caluclation and the outlier values are removing by simple equation Q+1.5*IQR [2] (where Q is a quartile value). The IQR and quantiles values are callculated by function contained in *stats* package. The example showed below presents those outliers detection in model.

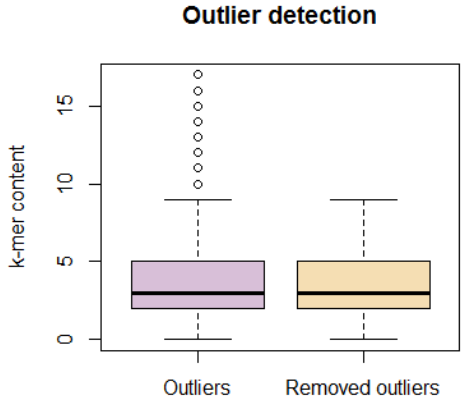

Fig. 3. Outlier detection
Rys. 3. Wykrywanie wartości odstających

As we can see on the above picture (Fig. 3) there are some outliers, that are detected by chosen method, which we eliminated by giving calculated thresholds.

In the next step lets denote that in our set of reads are N_k number of k-mers with k length. Any of k-mers is choosing and comparing with all reads sequences, then the amount of k-mer is counting. After searching substrings in reads and counting them, we calculate the mean (1) as x(w) (sum of all the reads contained k-mers after outliers removing) divided by number of used k-mers:

$$\bar{x}(w) = \frac{\sum_{i=1}^{N_k} x(w_i)}{N_k}, \tag{1}$$

According to the theory we may count the genome length as the maximum likelihood function give as the following:

$$\hat{G} = k - 1 + \frac{N(L-k+1)}{\bar{x}(w)}, \tag{2}$$

Whole algorithm was repeated several times to make the results more accurate. So the general steps of algorithm may be presented as the flowchart, where the last stage of analysis (genome length calculation) is based on the above (1, 2) equations:

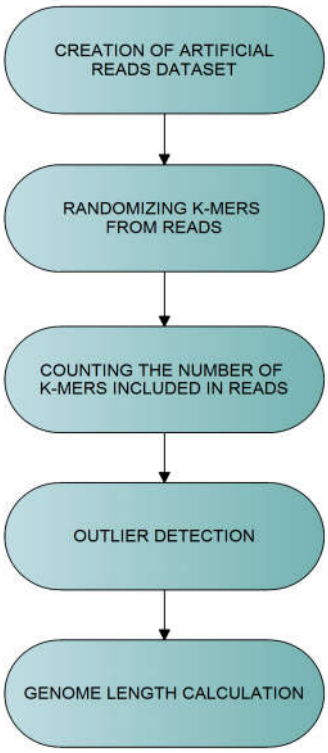

Fig. 4. Algorithm of model
Rys. 4. Algorytm działania modelu

2.2. Algorithm on the empirical data

Algorithm, which we created to estimate the genome size in empirical data is almost the same as in model. As we described in reads may occur some errors (e.g. wrong base calling and A replace to G) that may cause the situation, when we choose the k-mer sequence with error and non of the reads may contain such a pattern. Moreover, when we have a lot of mistakes in data, then we surely get the lower mean, than we should get. This case was eliminated by applying simple correction. Every of 0 values (0 means that non of reads does not include k-mer) was replaced by the mean value of the rest elements (non 0 elements).

The number of k-mer was set as 100 and the length of the k-mers as the 29 bp (which is 80% of the read). Parameters we set according to the results given by model. The estimation of the genome length according to real sequencing data run with 5 times repetition, and the number of k-mers was 100. The main steps of algorithm we presented as the flowchart:

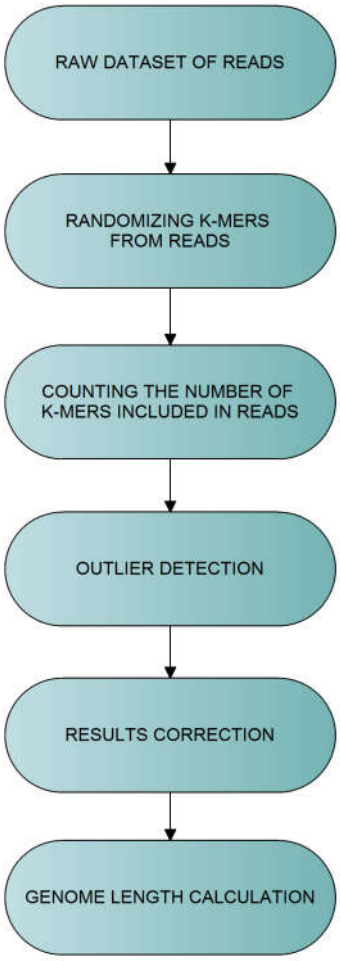

Fig. 5. Algorithm of genome length estimation on real data
Rys. 5. Algorytm oszacowania długości genomu dla danych rzeczywistych

2.3. Empirical data sources and using methods

The real data from whole-genome sequencing stored in .fastq files were downloaded from the European Nucleotide Archive[2]. To access the data we search Gene Expression Omnibus - GEO[3] database taking into account only whole genome sequencing data. Data, which we collected are coming from microorganisms and have relatively short genome sequences. Such a short genomes are very easy to analyze at the early stage of tests. The main information and statistic about the data is described in Table 1. We collect in this table such content as the species names from where the genome come from, SRR number (Sequence Read Archive run accession number), sequencing device used to create data, length of whole genome and one single read, number of reads and two set parameters the number of k-mers (N_k) and the length of k-mer.

Table 1

Statistic of used whole-genome sequencing data

Species	*Candida glabrata*	*Candida albicans*	*Lachancea waltii*	*Mycoplasma agalactiae*
Strain	CLIB 138	SC 5314	NCYC 2644	PG2
SRR number	SRR059730	SRR059732	SRR059728	SRR006331
Genome size [bp]	12338308	15213099	10912112	877438
Sequencing device	Illumina Genome Analyzer			
Depth of coverage	14,3	7,7	6,9	69,5
Read len.[bp]	36			
Read number	4912244	3265677	2095254	1693848
N_k	100			
k [bp]	29			

To analyze empirical data we use packages called *seqinr*[4] and *ShortRead*[5]. Easy way to load .fastq files are functions (e.g. *readFastq* allows to read the .fastq format file) included in *ShortRead*. After reading the file we extracted only the nucleotide sequences, which gives us all the information we needed. As regards pattern matching, we apply *grepl* function which is in basic package in R. To show the data results in a barplot we use package *ggplot2*[6].

[2] http://www.ebi.ac.uk/ena
[3] http://www.ncbi.nlm.nih.gov/geo/
[4] http://cran.r-project.org/web/packages/seqinr/
[5] http://www.bioconductor.org/packages/release/bioc/html/ShortRead.html
[6] http://cran.r-project.org/web/packages/ggplot2/

3. Results

This section reports results, which we obtained for model and empirical data. First three charts shows influence of the basic parameters into the genome length estimation. The barchart (Fig. 8) shows how the algorithm is working with data coming from real experiments of whole-genome sequencing.

On the first figure (Fig. 6) we can see the influence of the depth of coverage into the estimated genome length. This shows how a size and number of reads is important in genome length estimation.

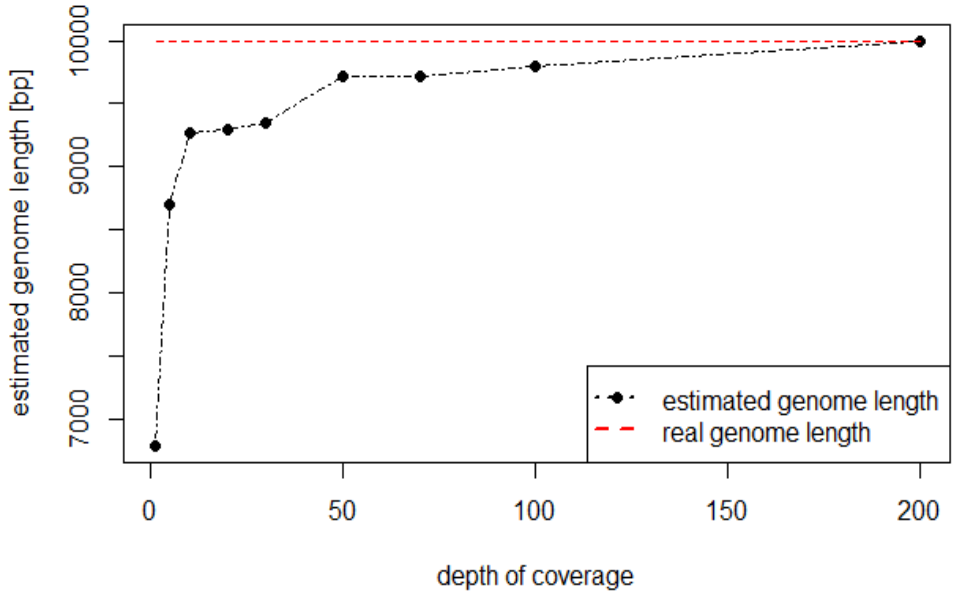

Fig. 6. Influence of the depth of coverage into genome length estimation
Rys. 6. Wpływ pokrycia na oszacowanie długości genomu

As we see the best results are obtained when the depth coverage is relatively high, so the best case is to have a datasets with relatively high depth of coverage. The next results (Fig. 7) presents how to choose the k-mer length. We can see that, when we have the k-mer length as the 60%-80% of read then we get the result very close to the real genome length. Moreover, above the length of k-mer at 80% level of read, we see decrease of estimation genome length. It may happen because of very unique sequences that may occur very rarely in reads, so it's hard to apply the outlier detection here.

We also checked the influence of the number of k-mer parameters, but above the small values of 10-50 there was no differences between estimated genome length. The very high value e.g. 1000 of k-mers given the same effect as the smallest number e.g. 100.

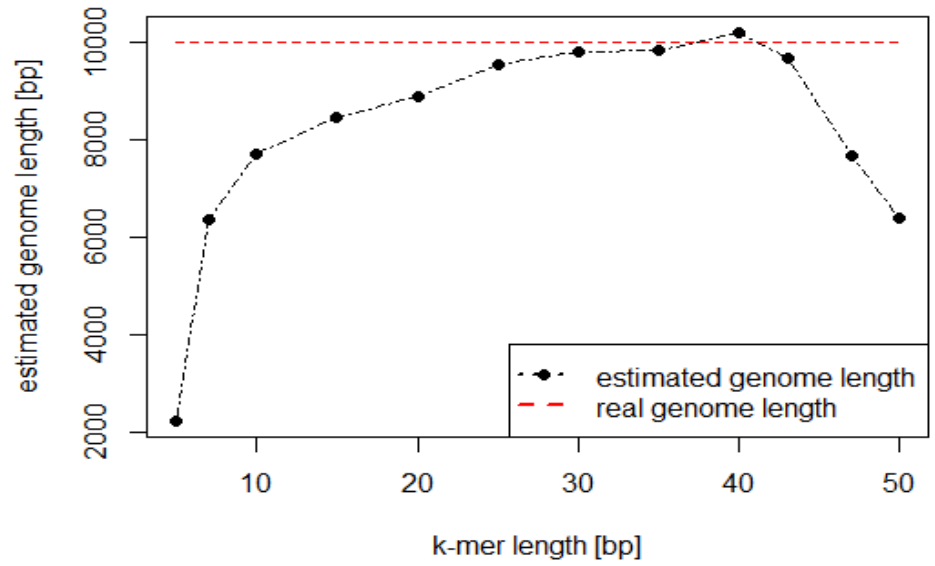

Fig. 7. Influence of k-mer length into genome length estimation
Rys. 7. Wpływ długości k-mer na oszacowanie długości genomu

On the next chart we wanted to check how the algorithm is working on the real data. We collect three groups of results (estimated genome length with and without correction, and the real genome length founded in NCBI[7] database) for each of four organisms.

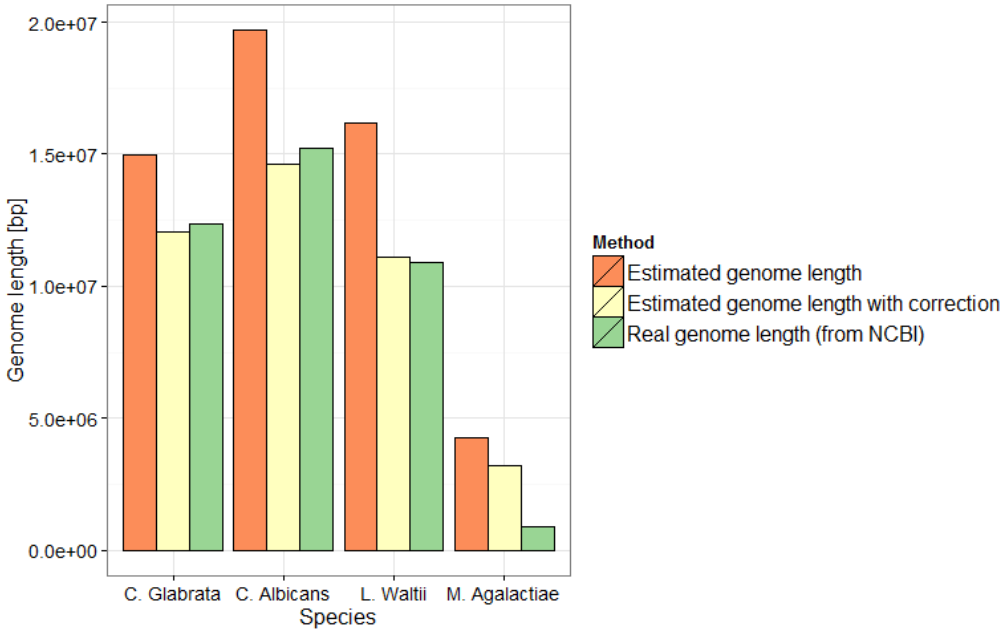

Fig. 8. Empirical data estimation results compared with the real genome size
Rys. 8. Porównanie oszacowanych wartości długości genomu z rzeczywistymi

[7] http://www.ncbi.nlm.nih.gov/assembly

Estimation of genome length let us to calculate the depth of coverage of datasets. To calculate the depth of coverage we use genome length value after correction. As we see the genome sizes without correction given in the chart and table are higher than the real genome length. This overestimation is caused by errors in real data. Sampling randomly k-mers we may choose the k-mer with error that may not occur in any read. This defective k-mer make the mean lower, that in fact influence to the higher genome size.

Table 2

Results for the empirical data

Species	*Candida glabrata*	*Candida albicans*	*Lachancea waltii*	*Mycoplasma agalactiae*
Estimated genome size [bp]	14970133	19688496	16185406	4252096
Estimated genome size with correlation [bp]	12038619	14620517	11081666	3218311
Calculated depth of coverage	11.81	5.97	4.66	14.34

4. Conclusions

The whole-genome sequencing data analysis is a wide field for creating bioinformatics tools and algorithms. The genome length estimation according to raw sequencing data is a problem, which we researched by creating our model. A created model can estimate the genome length in very good results according only to such parameters as: the number of sequences, length of sequences, k-mer number and k-mer length. To check some properties of parameters we draw two plots. First plot (Fig. 3) shows that increasing depth of coverage causes better estimation to known genome length. Furthermore, very small coverage make the estimated genome length lower than the real length. Next plot (Fig. 4) present influence of k-mer length into estimated values. Moreover, we see that in parameter k length is 80% of L and, that is the upper bound and the estimated values greater than this k-mer length are decreasing. Created results are showing that to obtain good estimated values we should choose the data with high coverage and should have unique k-mers sequences (not so long and also not so short).

Applied correction into the model let us to check the real data from WGS can let us know the estimated values of four chosen organisms. Estimated values without correction are higher than real genome length for all organisms, that is related to choosing k-mers with errors that making it not unique. The correction, which we apply can estimate very similar genome length values as the real genome lengths from NCBI database. The estimation of the genome length my depend on quality of sequencing data, as we observe the *M. Agalactiae* organism estimated values are very high in comparison with real data.

Above analysis explain the process of genome length estimation and show the influence of main parameters. The estimated genome length values may be used to calculate the depth of coverage, which is very good measure of the data quality and usefulness. The corrected values of the study organisms genome length, gives very good results that provides the model as a good predictor of genome length.

BIBLIOGRAPHY

1. Polański A., Kimmel M.: Bioinformatics. Springer, 2006, p. 243÷252.
2. Li X., Waterman M. S.: Estimating the repeat structure and length of DNA sequence using l-tuples. Genomes Res., vol. 13, 2003, p. 1916÷1922.
3. Lander E. S., Waterman M. S.: Genomic mapping by fingerprinting random clones: A mathematical analysis. Genomics, vol. 2, 1988, p. 231÷239.
4. Koronacki J., Mielniczuk J.: Statystyka dla studentów kierunków technicznych i przyrodniczych. Wydawnictwo Naukowo-Techniczne, Warszawa 2006.
5. Zhang J. et al.: The impact of next-generation sequencing on genomics. J Genet Genomics, vol. 38(3), 2011, p. 95÷109.
6. Schneeberger K., Weigel D.: Fast forward genetics enabled by new sequencing technologies. Trends in Plant Science, vol. 16, 2011.
7. Liu L. et al.: Comparison of Next-Generation sequencing systems. Journal of Biomedicine and Biotechnology, vol. 2012, 2012.
8. Schokralla et al.: Next-generation sequencing technologies for environmental DNA research. Molecular Ecology, 2012, p. 1794÷1805.
9. Samella L.: Correction of sequencing errors in a mixed set of reads. Bioinformatics, vol. 26, 2010, p. 1284÷1290.
10. Cock P. J. A.: The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variant. Nucleic Acids Research, vol. 38, 2010, p. 1767÷1771.

Omówienie

Artykuł prezentuję jedną z propozycji metody szacowania długości genomu na podstawie danych pochodzących z sekwencjonowania. Proces sekwencjonowania pozwala na odczyt sekwencji DNA (zawierającej cztery podstawowe nukleotydy oznaczane jako A, T, C, G), który współcześnie został w pełni zautomatyzowany. W wyniku sekwencjonowania otrzymuje się zestaw krótkich sekwencji DNA zwanych odczytami. Zaproponowana metoda bazuje na losowym generowaniu krótkich fragmentów (krótszych niż odczyty) zwanych k-mer, które pozyskuje się z losowo wybranych sekwencji odczytów. W analizie zaproponowano dwa kierunki badań: stworzono model generujący sztuczny zestaw danych (rys. 4) oraz przebadano wybrany zestaw danych rzeczywistych (rys. 5, tabela 1). Model bazuje na rzeczywistych właściwościach sekwencjonowania oraz zakłada, iż może zaistnieć sytuacja, podczas której zostanie wybrana powtarzalna sekwencja (rys. 2) znajdująca się na więcej niż jednym miejscu w genomie. Taki przypadek został rozwiązany przez wykrycie wartości odstających (rys. 3) oraz usunięcie ich z dalszej analizy. Samo szacowanie wielkości genomu zostało oparte na wzorze (2). Zaproponowany model pozwolił określić ilość oraz długość sekwencji k-mer (rys. 7), jaką należy ustalić w badaniach. Dzięki modelowi przebadano również relację (rys. 6) stopnia pokrycia (ilości nukleotydów zawartych w odczytach na danym miejscu w genomie) z oszacowaną wartością długości genomu. Ze względu na pojawiające się w danych rzeczywistych błędy (źle odczytany nukleotyd) zastosowano korekcję, która polegała na wypełnieniu wartości zerowych (przypadek gdy sekwencja k-mer z błędem nie została znaleziona w żadnym z odczytów) – wartością średnią pochodzącą z wektora wartości niezerowych. Zastosowanie korekcji pozwoliło na otrzymanie wyników bardzo zbliżonych do rzeczywistych długości genomu (rys. 8, tabela 2). Niniejsza praca pozwoliła na stworzenie algorytmu do szacowania długości genomu, co jest ważnym parametrem podczas analizy danych pochodzących z sekwencjonowania. Dzięki pozyskanym długościom genomu obliczono stopień pokrycia danych (tabela 2), który jest również bardzo istotnym parametrem mówiącym o przydatności i jakości danych w analizie.

Addresses

Mateusz GARBULOWSKI: Silesian University of Technology, Institute of Informatics, ul. Akademicka 16, 44-100 Gliwice, Poland, mateusz.garbulowski@polsl.pl.
Andrzej POLAŃSKI: Silesian University of Technology, Institute of Informatics, ul. Akademicka 16, 44-100 Gliwice, Poland, andrzej.polanski@polsl.pl