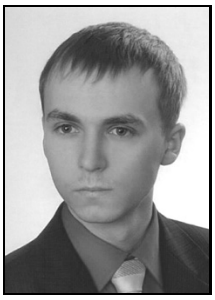
Mateusz GARBULOWSKI, Andrzej POLAŃSKI
SILESIAN UNIVERSITY OF TECHNOLOGY
Akademicka Street 16, 44-100 Gliwice, Poland

A system for simulation of DNA coverage in shotgun sequencing processes

Eng. Mateusz GARBULOWSKI

Student of Biotechnology specializing in Bioinformatics, at the Silesian University of Technology. In 2013 got the engineer title.

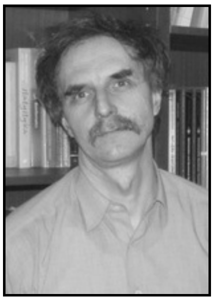
e-mail: mateuszgarbulowski@gmail.com



Prof. dr hab. inż. Andrzej POLAŃSKI

Andrzej Polanski (year of graduation in 1982, a doctorate in 1990, habilitation in 2000, titular professorship 2009). Director of Bioinformatics at the Institute of Informatics, Silesian University of Technology. Director of doctoral studies at the Silesian University of Technology computer science studies. Author and co-author of over 150 scientific publications.

e-mail: andrzej.polanski@polsl.pl



Abstract

A design of a computational environment for simulation and statistical analysis of shotgun DNA sequencing process is presented. The approach involves developing simulation procedures on the basis of the Lander-Waterman theory. The explored aspects concern numbers of gaps and contigs. Simulations allow drawing certain conclusions: the created model is very similar to the Lander-Waterman theory, simulations of k-mers maps by the Poisson process allows estimating statistics of contigs number.

Keywords: sequencing, statistic, DNA coverage.

System dla symulacji pokrycia DNA w procesach sekwencjonowania typu shotgun

Streszczenie

W artykule zawarte są informacje dotyczące statystycznej analizy metody sekwencjonowania typu „*Shotgun*". Projekt zakładał stworzenie środowiska obliczeniowego oraz modelu matematycznego, który jak najdokładniej odzwierciedla proces sekwencjonowania metodą „*Shotgun*", wykorzystując przy tym losowe powstawanie krótkich sekwencji nukleotydowych, tak zwanych *read'ów*, a co za tym idzie również losowe formowanie się *contig'ów* – w pełni odtworzonych odcinków sekwencji. Stworzony model dzielił sekwencję zasad na zadaną ilość *read'ów* o stałej długości którą następnie odtwarzał poprzez porównanie końca poprzedniego i początku kolejnego *read'a*, sprawdzając tym samym ile fragmentów zostaje w pełni złożonych w contig'i. Jako własności statystyczne metody można rozumieć wzory *Landera-Watermana* przewidujące ilość powstawania *contig'ów*, które biorą pod uwagę całkowitą ilość *read'ów*, długość *read'ów* oraz całą długość sekwencji wejściowej. Wartości uzyskane metodą *Landera-Watermana* oraz uzyskane za pomocą modelu przedstawiono w postaci wykresu zależności ilości powstających *contig'ów* do parametru ścieżki pokrycia. Dodatkowo pod względem statystycznym wykreślono histogramy przedstawiające częstość występowania zasad w danym miejscu stworzone w oparciu o model oraz wykreślono na wykresie zakres wyników dla ilości powstających *contig'ów* wyliczony jako maksima i minima dla wielu losowych prób i przedstawione jako zależność od ścieżki pokrycia.

Słowa kluczowe: sekwencjonowanie, statystyka, pokrycie DNA.

1. Introduction

The statistical analysis of shotgun technology for DNA sequencing is very important to take a better look on this and get to know how to improve the process.

DNA sequencing by the shotgun method is commonly used to sequence very large amounts of DNA such as a whole genomes. This approach shredded DNA into smaller fragments called reads. The reads are then reassembled, by taking into account their overlaps. The main result of the assembly of reads is creation of large piece of estimated DNA sequences, called contigs.

In order to study this process and its statistics, we have created a model which contains all the assumptions of the shotgun design, described in the literature [1, 2], and we have compared it to Lander-Waterman number of contigs prediction. Furthermore by creating a histogram containing a frequency of the occurrence of a base at each place, we have confirmed that formation of a contig is indeed well modelled by a Poisson process. The last result presented in this research is the range that showing the bounds created by model results and it may be used to compare the empirical data. In the forthcoming sections of the paper we will show how the computational environment for DNA coverage simulation was designed, how it compares to the Lander – Waterman theory and how the main problems was solved.

2. Model and Lander-Waterman comparison

The first aim of the project was the use of computer simulations for the verification of the Lander-Waterman prediction equation. The Lander-Waterman model is based on three basic parameters: the length of reads called *L*, the numbers of all reads called *N* and the region of analyzing genome called G. The Lander – Waterman (LW) equation [1] for the average number of contigs is as follows:

$$E[\#contigs] = N \cdot e^{\left(-\frac{N \cdot L}{G}\right)} \qquad (1)$$

The LW prediction shown in (1) was compared to empirical statistics created with the use of a model based on the knowledge of the shotgun sequencing process [1, 2].

The short description of the created algorithm is as follows:

- the procedure reads the true input DNA sequence
- the next step is randomization of *N* positions of k-mers (each of length *L*) along the DNA input sequence (of length *G*)
- a collection of read sequences (k-mers) is created.

The set of the obtained k-mers and date on their positions, is used by the developed computational program to count the number of contigs as those fragments which contain overlaps (ranging at least T bps) between subsequent reads.

In order to compare the results of the model and Lander-Waterman prediction, there was used the parameter of the depth coverage called "*a*" [1] and described as:

$$a = \frac{N \cdot L}{G} \qquad (2)$$

The depth coverage parameter tells us how much of the bases falls on the whole genome. It is important that "*a*" should be more than one for sufficient coverage.

In order to obtain a more reliable conclusions, the main algorithm was repeated 200 times for each variable value *N*.

The result of comparison model and prediction for length *G* as 100 000 bases, read length 200 bases and coverage 20 bases for variable *N* parameter is the following graph:

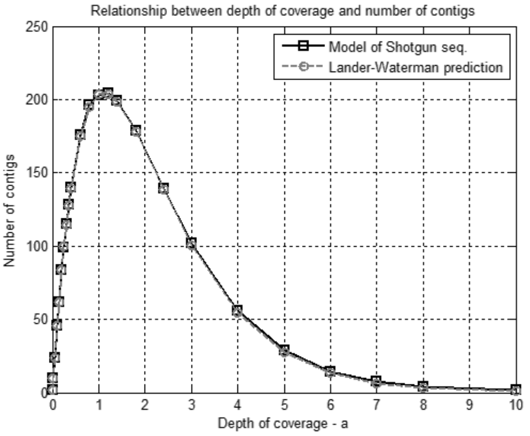


Fig. 1. Relationship between the model of Shotgun sequencing process and Lander-Waterman prediction
Rys. 1. Związek pomiędzy stworzonym modelem a wzorem predykcyjnym Landera-Watermana

The graph shows that the maximum number of contigs is for the depth of coverage about 1. For "*a*" more than 1 the number of contings the decreases to one contig.

3. Formation of contig

As derived in the literature [1, 2], the formation of contig is described as the Poisson point process. To verify the theory, we used the created computational environment to count the frequency of occurrences of a base in each location and based on this we created the histogram (frequency of repetitions) for the following genomic data: genome length 100 000 bases, read length 700 bases and number of reads 500. The result is shown below:

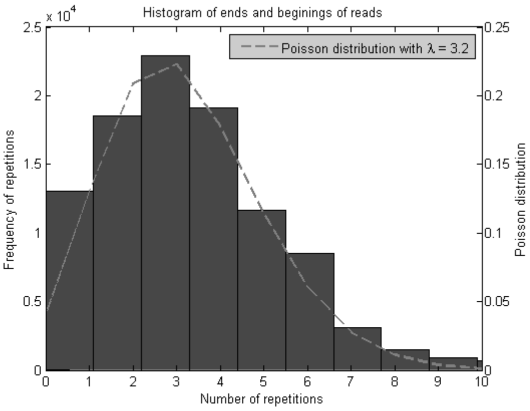


Fig. 2. Relationship between the number of repetitions and their frequency of occurence
Rys. 2. Powiązanie ilości występujących zasad w danym miejscu z częstością ich występowania

As shown in Fig. 2, the obtained histogram is enveloped by the theoretical Poisson probability distribution function. The maximum is about 3 repetition of about 2500 frequency which proves that mostly the three reads are above each other.

4. Bounds of the model range

Since the results of the model are random, a zone of tolerance such as range bounds of many tests may be created. The range is calculated as the maximum and minimum number of contigs for 500 repetitions to make the effect more likely. The result of these bounds for length *G* as 100 000 bases, read length 200 bases and coverage 20 bases for variable *N* parameter, is the following graph:

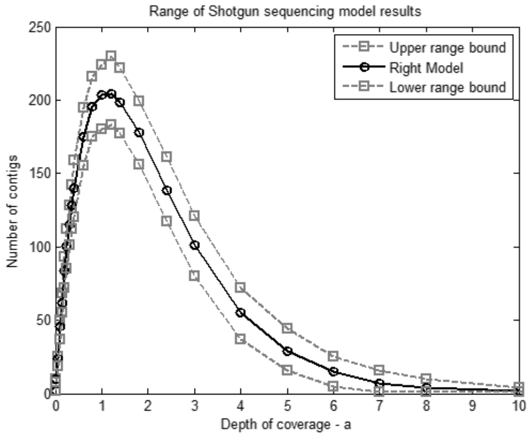


Fig. 3. Definition of the upper and lower range bound
Rys. 3. Przedstawienie wyliczonych zakresów dla modelu

The estimated intervals show that for the depth coverage about one to five there are very similar distances between the upper and lower range bound. In less than one and more than six the bounds are smaller and decrease to zero.

5. Conclusions

The created model show a huge similarity to the Lander-Waterman prediction theory. For the depth of coverage equal to about 8, the accuracy of getting contig is one, so that in this point is formed one whole contig and by increasing the depth of coverage to more than 8 the size of the coverage of the full DNA fragment is extended.

The bar configuration in Fig. 2 proves that the formation of contig is a Poisson process.

Computation of the range bounds by the created model shows that the interval is very different for each point along the horizontal axis. The difference between the bounds and the right model over the depth coverage at six decreases to zero because by the increasing the number of reads, there is formed one whole contig so that after a certain time the minimum and maximum get closer to the right model. The computation of the results range is very important. It may be helpful to compare the empirical data to theoretical predictions.

The created model can augment the Lander-Waterman prediction to better analyze the empirical data, according to its randomness and computed range bounds of results.

The research presented in this paper was supervised by professor Andrzej Polanski, Institute of Informatics, Silesian University of Technology and was partly supported by POIG.02.03.01-24-099/13 grant: "GCONiI - Upper-Silesian Center for Scientific Computation".

6. References

[1] Polański A., Kimmel M.: Bioinformatics, Statistics of the Genome Coverage, 243-252, 2007.

[2] Eric S. Lander, Michael S. Waterman: Genomic Mapping by Fingerprinting Random Clones: A Mathematical Analysis, Genomics 2, 231-239, 1988.

[3] Michael C. Wendl, Shiaw-Pyng Yang: Gap statistics for whole genome shotgun DNA sequencing projects, Bioinformatics, 1527-1534, vol. 20, no. 10, 2004.

otrzymano / received: 12.03.2014
przyjęto do druku / accepted: 02.05.2014 *artykuł recenzowany / revised paper*