DNA SEQUENCING

Belarusian State University of Informatics and Radioelectronics
Department of Information and Computer-Aided Systems Design
Minsk, Belarus

Romantsov A.

Liahushevich S. Candidate of philological sciences, associate professor

Abstract
DNA sequencing is the process of determining the precise order of nucleotides within a DNA molecule. The canonical structure of DNA has four nucleotides: thymine, adenine, cytosine, and guanine. In this paper we observe current approaches and one of the most perspective method under development, and present the results that we obtained during our research.

The first DNA sequences were investigated in the early 1970s by academic researchers using laborious methods based on two-dimensional chromatography. Following the development of fluorescence-based sequencing methods with automated analysis,[1] DNA sequencing has become easier and orders of magnitude faster.[2]
Allan Maxam and Walter Gilbert published a DNA sequencing method in 1977 based on chemical modification of DNA and subsequent cleavage at specific bases.[3]
The chain-termination method developed by Frederick Sanger and coworkers in 1977 soon became the method of choice, owing to its relative ease and reliability.[4][5] When invented, the chain-terminator method used fewer toxic chemicals and lower amounts of radioactivity than Maxam and Gilbert's method. Because of its comparative ease, Sanger's method was soon automated and was the method used in the first generation of DNA sequencers.

Sanger's sequencing is the method which prevailed from the 1980s until the mid-2000s. Over that period, great advances were made in the technique such as fluorescent labelling, capillary electrophoresis and general automation. These developments allowed much more efficient sequencing, leading to lower costs. Sanger's method, in mass production form, is the technology which produced the first human genome in 2001, ushering in the age of genomics. However, later in the decade, radically different approaches reached the market, bringing the cost per genome down from $100 million in 2001 to $10,000 in 2011. We call such approaches Next-generation methods.
The high demand for low-cost sequencing has driven the development of high-throughput sequencing (or next-generation sequencing) technologies that parallelize the sequencing process, producing thousands or millions of sequences concurrently. High-throughput sequencing technologies are intended to lower the cost of DNA sequencing beyond what is possible with standard dye-terminator methods.
One of this new approach is nanopore sequencing. It has been under development since 1995. A nanopore is simply a small hole of the order of 1 nanometer in internal diameter. Certain porous transmembrane cellular proteins act as nanopores, and nanopores have also been made by etching a somewhat larger hole (several tens of nanometers) in a piece of silicon, and then gradually filling it in using ion-beam sculpting methods which results in a much smaller diameter hole: the nanopore. Graphene is also being explored as a synthetic substrate for solid-state nanopores.
This method is based on the readout of electrical signals occurring at nucleotides passing by biological or solid-state pores. The DNA passing through the nanopore changes its ion current. This change is dependent on the shape, size and length of the DNA sequence. Each type of the nucleotide blocks the ion flow through the pore for a different period of time. The method has a potential of development as it does not require modified nucleotides.
Such new technologies aim to increase throughput and decrease the time of obtaining the result and cost by eliminating the need for excessive reagents and harnessing the processivity of DNA polymerase.
Two main areas of nanopore sequencing in development are solid state nanopore sequencing and protein based nanopore sequencing. Protein nanopore sequencing utilizes membrane protein complexes alfha-Hemolysin and MspA (Mycobacterium SmegmatisPorin A), which show great promise given their ability to distinguish between individual and groups of nucleotides. Whereas, solid-state nanopore sequencing utilizes synthetic materials such as silicon nitride and aluminum oxide and it is preferred for its superior mechanical ability and thermal and chemical stability. The fabrication method is essential for this type of sequencing given that the nanopore array can contain hundreds of pores with diameters smaller than eight nanometers.
The sequencing technologies described above produce raw data that need to be assembled into longer sequences such as complete genomes (sequence assembly). There are many computational challenges to achieve this such as the evaluation of the raw sequence data which is done by programs and algorithms such as Phred and Phrap. Other challenges have to deal with repetitive sequences that often prevent complete genome assemblies because they occur in many places of the genome. As a consequence, many sequences may not be assigned to particular chromosomes. The production of raw sequence data is only the beginning of its detailed bioinformatical analysis. Yet new methods for sequencing and correcting sequencing errors were developed.
We made a simulation of the whole sequencing process. In our simulation we took a one molecule of Saccharomyces Cerevisiae with the length of 250 thousand bases. It was cloned 50 times. Then it was randomly cuted into pieces with the length of 300-400 nucleotides. We assume that some percentage of information can be lost during the sequencing that is why we removed the portion of nucleotides so that the remaining number averaged corresponds to predetermined percentage. Then we assembled the sequence from the pieces. If 100% of the nucleotides are known, the resulting consensus sequence is 100% identical to the original. If 75% of the nucleotides are known, the resulting

consensus sequence is 100% identical to the original. In this case a significant number (>70%) relatively short (<20 nucleotides) reads occurred, which must be rejected during assembly.
When the only 50% of the nucleotides are known, the consensus sequence obtained is 86% identical to the original. In this case we have a bigger number (> 90%) relatively short (<20 nucleotides) reads, which must be rejected during the assembly. Divergent the portions correspond to the fragments with low coverage.

REFERENCES:

Olsvik, O. Use of automated sequencing of polymerase chain reaction-generated amplicons to identify three types of cholera toxin subunit B in Vibrio cholerae O1 strains / O. Olsvik, J. Wahlberg, B. Petterson, M. Uhlén, T. Popovic, I. Wachsmuth // Journal of Clinical Microbiology, 1993.
Pettersson, E. Generations of sequencing technologies / E. Petterson, J. Lundeberg, A. Ahmadian // Genomics, 2009.
Maxam, A. A new method for sequencing DNA / A. Maxam, W. Gilbert // Proceedings of the National Academy of Sciences of the U.S.A., 1977.
Sanger, F. DNA sequencing with chain-terminating inhibitors / F. Sanger, S. Nicklen, A. Coulson // Proceedings of the National Academy of Sciences of the U.S.A., 1977.
Sanger, F. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase / F. Sanger, A. Coulson // Journal of Molecular Biology, 1975.