VOL.2 NO.2September 2013 - ISSN 2233 – 1859

UOIuBIH
ORSinBIH
Operations Research Society in
Bosnia and Herzegovina

Southeast Europe Journal of Soft Computing

Available online:www.scjournal.com.ba

DNA Sequencing

Busra Gheith, Mehmet Can
International University of Sarajevo, Faculty of Engineering and Natural Sciences, Hrasnicka Cesta 15, Ilidža
71210 Sarajevo, Bosnia and Herzegovina

Article Info
Article history:
Received 17 Sep.2013
Received in revised form 17 Oct 2013

Keywords:
DNA, sequencing, genome, algorithm, spectrum

Abstract

The sequencing of the Human Reference Genome, with Human Genome Project announced ten years ago, provided a roadmap that is the foundation for modern biomedical research. Reference Genome represents digital database founded by scientists which contains representative examples of species genomes. The need for sequencing has never been greater than it is today. Sequencing has found its applications within diverse research sectors including comparative genomics and evolution, forensics, epidemiology, and applied medicine for diagnostics and therapeutics. Arguably, the strongest rationale for ongoing sequencing is the question for identification and interpretation of human sequence variation as it relates to health and disease. The paper gives review of current DNA sequencing algorithms and techniques as well as next-generation of DNA sequencing. Since the DNA sequencing field is changing rapidly the information in this paper represent a snapshot of this particular moment.

1. INTRODUCTION

Dramatic progress in biology has been driven by knowledge about deoxyribonucleic acid (DNA). DNA is a molecule composed of deoxyribonucletides connected by phosphodiester linkages. This molecule is first observed by Frederich Miescher in the late 1800s. Importance of DNA is realized in 1953 when scientists revealed helix structure of DNA which carries biological information from one generation to the next. DNA is constructed of a double helix held together by hydrogen bonds. in this structure there are four kinds of nucleotides: A-adenine, C-cytosinem G-guanine and T-thymne [1]. Two standards of helix are complement to each other where A pairs with T and G pairs with C.

DNA sequence of living organisms is called genome. A section of the DNA coding for a protein is called a gene. The size of a gene may vary greatly, ranging from about 1,000 bases to 1 million bases in humans. Genome, for a human contains about 3 billion bases and about 20,000 genes on 23 pairs of chromosomes (Neil C.Jons,Pavel A.Pevzner, 2004).

The DNA sequencing consists of determination sequence of nucleotides of an examined DNA fragment cut out from a genome by restriction enzymes or by the shotgun approach (Waterman, M.S., 1995), (J. Blazewicz, P. Formanowicz, F.Guinand and M. Kasprzak., 2002). Modern era of DNA sequencing started in 1977 with developments of two techniques (J. Blazewicz, P. Formanowicz, F.Guinand and M. Kasprzak., 2002).

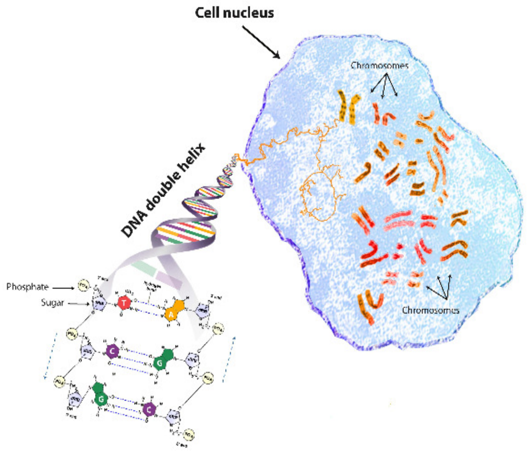

Figure 1: Structure of DNA

However, the overwhelming majority of DNA sequence production until today has relied on some version of the Sanger biochemistry (R.Idury and M.Waterman, 1995). Statistical properties of DNA sequences have been extensively studied in and mathematical tools used for evaluation of DNA sequencing is presented in the paper (J.Shendure, H.Jii, 2008).

This paper is organized as follows : section one gives introduction to DNA sequencing. In second section is giver overview of methods used for DNA sequencing. Section three introduces algorithms for DNA sequencing. Paper concludes with section four.

2. METHODS FOR DNA SEQUENCING

In this section is given overview with chronological development of the distinct approaches to DNA sequencing. These are:

1- Sangers method

2- Maxam and Gilbert method

3- Cyclic array sequencing

4- Sequencing by hybridization

5- Electrophoresis

6- Mass spectrometry method

7- Method for nanopore sequencing.

Additionally, this section introduces key parameters that should be considered when choosing the DNA sequencing strategy most appropriate for a given application. It should be emphasized that the DNA sequencing field is changing rapidly, so the information in this unit represents a snapshot of this particular moment.

2.1 Sangers method

Sangers sequencing is also known as dideoxy sequencing. This method is introduced for the first time in 1975 and in the next 30 years has completely dominated modern methods for sequencing (S. Grumbach, F.Tahi, 1994). Sanger sequencing is a DNA sequencing method in which target DNA is denatured and annealed to an oligonucleotide primer, which is then extended by DNA polymerase using a mixture of dideoxynucleotide triphosphates (normal dNTPs) and chain-terminating dideoxynucleotide triphosphates (ddNTPs). Novelty in this approach was use of polyacrylamide gels to separate the products or primed synthesis by DNA polymerase in order to increase chain length (D. Huson, 2010) (J. Shendure, G.Porreca, G. Church , 2008) (F. Sanger, S. Nicklen, and R. Coulson , 1977). The enzyme method for DNA sequencing has been used for genomic research as the main tool to generate the fragments necessary for sequencing, regardless of the sequencing strategy. Two different approaches, shotgun and primer walking sequencing, are the most used for DNA sequencing. Shotgun sequencing is a random process where DNA is randomly fragmented into smaller pieces, which produces a high level of redundancy which increases total cost.

Second approach represents direct sequencing of unknown DNA where sequence is known. Unknown sequence of DNA is inserted into a vector and amplified (F.Sanger, S.Nicklen, and R.Coulson , 1977). This walking approach has major advantage in the reduced redundancy. These methods are described more in detailed in (F.Sanger, S.Nicklen, and R.Coulson , 1977)

Since this is very complex method there are few variation of this method like chain-terminator sequencing, dye-terminator sequencing (S.Grumbach, F.Tahi, 1994).

2.2 Maxam and Gilbert method

Maxam and Gilbert (1977) developed a DNA sequencing method that was similar to the Sanger and Coulson method in using polyacrylamide gels to resolve bands that terminated at each base throughout the target sequence, but very different in the way that products ending in a specific base were generated. The are three main advantages of the Maxam Gilbert and other chemical methods compared with Sanger's chain termination reaction method. First advantage is that a fragment can be sequenced from the original DNA fragment, instead of from enzyme copies. Second advantage is no subcloning and no PCR reactions are required. Consequently, for the location of rare bases, the chemical cleavage analysis cannot be replaced by the dideoxynucleotide terminator method, as the latter analyses the DNA of interest via its complementary sequence, it can, thus, only give sequence information in terms of the four canonical bases. Third advantage relies on the fact that this method is less susceptible to mistakes with regard to sequencing of secondary structures or enzyme mistakes. Some of the chemical protocols are recognized by different authors as being simple, easy to control, and the chemical distinctions between the different bases are clear (F .Sanger, S. Nicklen, and R. Coulson, 1977).

2.3Cyclic array sequencing

Platforms performing cyclic array are performing simultaneously decoding of a two-dimensional array. Main property here is that features are not necessarily separated into individual wells (D.Huson, 2010). Complete process is cycle because enzymatic process is applied to interrogate the identity of a single base position for all features in parallel within each cycle.

Discussion about three distinct approached that are currently available as commercial or open-source platforms, is given in (D. Huson, 2010). New developed parallel methods increased number of sequence reads from a single experiment than using capillary electrophoresis sequencers. This effort is achieved with sacrifices in length and accuracy of the individual reads.

The most popular and widely used platforms that found commercial use are Roche 454, Illumina, ABI's SOLiD, Pacific Biosciencs sequencing.

Roche 454 sequencing allows shotgun sequencing of genomes without cloning in E.coli or any host cell. The 454 system was the first next-generation sequencing platform available as a commercial product 454 Genome Sequencer FLX is able to produce 100Mb sequence with 99.5 % accuracy in over 250 bases length.

Illumina sequencing is similar to Roche 454 only difference is that this method uses chain-terminating nucleotides. This machine can analyse more than one billion bases of 75 base reads in a single run.

ABI's SOLiD sequencing is realized in 2008. Technology is based on hybridization-ligation chemistry. This method can generate 2-3 Gb of sequence per run.

Pacific Biosciences SMRT DNA sequencing offers long reads, ultra-fast cycle tune and flexibility. SMRT unifies separate applications of real time single molecule DNA sequencing and methylation sequencing. These first instruments are launched in 2010 (D.Huson, 2010).

2.4Sequencing by hybridization

The main idea here is to build a two-dimensional grid with k-tuples which is a word of length k. This matrix represents sequencing chip. DNA is labeled with radioactive material, and each k-tuple present in the simple is hybridized with its reverse complement in the matrix. Unhybridized DNA is removed from the matrix (J. Blazewicz, P. Formanowicz, F.Guinand and M.Kasprzak., 2002). Mathematical aspects of this method are nontrivial. Eulerian method gives one solution to this method (J.Shendure, G. Porreca, G. Church , 2008).

2.5Electrophoresis

Because of the problems related to preparation of gel, new method for sequencing was invented. Capillary electrophoresis is a fast technique for analysis. This technique can resolve complex mixtures of biopolymers in a high electric field (F.Sanger, S.Nicklen, and R.Coulson , 1977).

2.6Mass-spectrometry method

Mass spectrometry has found its place in DNA sequencing. This method relies on the precise measurement of the masses of DNA fragments.

This method shows better results for RNA analysis and it is predicted that this method will not displace conventional methods for most DNA sequence applications (F.Sanger, S.Nicklen, and R.Coulson , 1977).

2.7Nanopore sequencing

The development of nanopore devices to sequence DNA is an area of vigorous research that promises exciting results. The essential idea is to make single stranded DNA (ssDNA) pass through (or 'translocate') a nanopore whose diameter is of the order of a few atoms. The passage of individual bases in the ssDNA is sensed using electrical or optical means. The resulting signal is processed to reveal the identity (i.e. A, C, G or T) of the individual bases as they pass through the nanopore (L.Franca, E.carrilho, and T.Kist , 2002).

3. DNA SEQUENCING ALGORITHMS

In this section will be given review of current algorithms used for DNA sequencing.

3.1 A compression algorithm for DNA sequencing

Grumbach and Tahi (S.Grumbach, F.Tahi 1994 , D.Huson 2010) proposed two lossless compression algorithms for DNA sequences, namely *Biocompress* and *Biocompress-2. Biocompress-2* detects exact repeats and complementary palindromes located earlier in the target sequence, and then encode them by repeat length and the position of a previous repeat occurrence. Author Rivals gives another compression algorithm *Cfact*, which searches the longest exact matching repeat using suffix tree data structure in an entire sequence. The idea of *Cfact* is basically the same as *Biocompress-2* except that *Cfact* is a two-pass algorithm. The lossless compression algorithm *GenCompress* in this paper achieves significantly higher compression ratios than both *Biocompress-2* and *Cfact*.

GenCompress is a one-pass algorithm. Algorithm works in next manner:

1- For input w, some part v has already been compressed.

2- GenCompress finds optimal prefix that approximately matches substring v.

3- After outputting the code, remove the prefix and append it to the suffix of v.

After defining compression gain function G algorithm is performed in 4 steps in order to find optimal prefix.

In order to reduce search space few observations are made based on two lemmas (S.Bokhari, J.Sauer, 2004).

The compression results of *GenCompress* for DNA sequences indicate that our method based on approximate matching is more effective than others. *GenCompress* is able to detect more regularity and achieve best compression results by using this observation (S.Bokhari, J.Sauer, 2004).

3.2 A greedy algorithm for aligning DNA sequencing

Greedy algorithm is used for aligning DNA sequences which are different only from sequencing errors or by equivalent errors from other sources. Algorithm works in a way that uses two sequences with sequence errors and does iterations in three loops. Iterations are done until 30 identical nucleotides are found. Problem can be when loops are finished quickly because next sequencing error appears after few basepairs, and or loop can be iterated appreciable number of times because search has reached low complexity sequence region. X-drop approach gives solution for these problems. X-drop algorithm gives different termination condition for the loops and region of the search are expanded at regions of low-sequence complexity or concentrated sequencing errors. More about this can be found in (X.Chen, S.Kwong, M.Li, 1999).

A greedy algorithm uses measurement of difference between two sequences. An alignment is done counting differences or columns that do not align identical nucleotides. Main problem is to find difference D (i ,j), between two strings which is than defined as the minimum number of differences. It can happen that two sequences have the same number of differences and same score. In this case there are two lemmas described in (X.Chen, S.Kwong, M.Li, 1999) for determining. For some algorithms only score alignment is adequate. Implementations of this algorithm have been incorporated in the National Center for Biotechnology Information. Most popular one is called Blast-family tool.

3.3 An algorithm for heuristic managing error for DNA sequencing

In hybridization experiment are some errors (positive and negative) produced in the spectrum. Spectrum represents oligonucleotides written as words of equal length {A,C,G,T}. This algorithm proposes new superstring which should optimize a cost function based on its thickness. Algorithm, itself is based on overlapping windows of oligonucleotides and fusions of sub-segments. From this set algorithm produces sequence or several fragments included in the original sequence. Cost function for the choice of oligonucleotides is used thickness, which leads that the larger thickness is the better its current position fits into reconstructed sequence. Algorithm can be explained to work in three steps:

1- computation of overlapping windows,

2- building sub-segment and

3- reduction

In general algorithm takes specificity of SBH and returns one solution. With this positive and negative errors can be detected. This algorithm compared with previously used one has more advantages, like generating solutions much faster, it returns optimal solution, and this algorithm shows good results working with very large oligonucleotides (Waterman,M.S., 1995).

3.4 A parallel graph decomposition algorithm for DNA sequencing with nanopores

This algorithm can be considered as an extreme variation of Eulerian path approach. Eulerian path algorithm is well known by its use in hybridization method for DNA sequencing. This algorithm works so that search over a space of de Bruijin graphs until it finds one where impact of errors is eliminated or possible orientations of the two ssDNA sequences can be identified separately and unambiguously. Reconstruction of a DNA sequence is done using technique which break up each read into a set of overlapping k-mers, creating a de Bruijn graph and finding Eulerian path in the graph. Problem showing here is in orientation and complementarity of the reads. With increased complexity Bruijn graph transforms from a convulated to one that contains four sequences (3’-5’ string, 5’-3’ string and two Watson-Clark complements) where graph is formed of four disjoin paths but where paths are all equal length. The overall algorithm is performed in seven steps described in (Z. Zhang, S.Schwartz, L.Wagner and W. Milleri, 2000), (D. Lavenier, 2008) ,(L.Li and S. Khuri, 1995). Main target of this is to develop CrayMTA-2 supercomputer which could support multithreading and large flat shared memory without locality. Made algorithm is able to reconstruct sequence in about 40 min. Total time is affected by mer size and graph generation.

3.5 Ordered index seed algorithm for DNA sequence comparison

This algorithm introduces new approach in manipulating seeds focusing on faster execution time. Two DNA sequences are compared on the base of banks or full genomes. This algorithm follows four steps:

1- Indexing two banks

2- Hit extensions

3- Computing gap alignment

4- Display alignment.

These steps are performed globally and sequentially. Seed of two banks are first indexed, then following seed criteria ordering an ungained extension is started. Process stops if a seed is smaller than the starting one. This means that this seed has been already detected and there is no need to be processes again. This algorithm provides shorter execution time because it extensively uses cache memory of processors and index ordered seed technique delivers unique HSP avoiding complex data structured (Z.Zhang, S.Schwartz, L.Wagner and W.Milleri, 2000).

3.6 Algorithm for shotgun sequencing

This algorithm combines features of shotgun sequencing and sequencing by hybridization. From known fragments k-tuples or segments can be determined. In a case of ideal data which are data without errors, sequence is covered with fragments and overlap between fragments is at least k, then union of all k-tuples from the fragments will be the same ad spectrum. Main idea is to build a graph with k-tuples of the union and to perform Eulerian tours. The algorithm can be summarized into four steps:

1- Formation of union of spectrum of fragments and reverse complements

2- Construction of the spectrum graph

3- Perform Eulerian tour and infer the sequence

4- Align fragments to sequence

This algorithm is not embedded in any software yet even some prototype exists. Experiments with the prototype showed that error rate is ranging between 0.5% and 3.0%. Main advantage of this algorithm is high coverage and low sequencing errors (J. Blazewicz, P.Formanowicz, F.Guinand and M.Kasprzak., 2002), (V.Bansal, 2010), (Blazewicz,J., Formanowicz,P., Kasprak,M., Markiewicz,W.T. and Weglarz,J., 1999), (Dramanac,R., Labat,I. and Crkvenjakov,R, 1991), (Fleischner,H, 1990).

4. CONCLUSIONS

Demand for DNA sequence information has never been greater, yet current Sanger technology is too costly, time consuming, and labor intensive to meet this ongoing demand. There are good prospects for the emergence of new and non-conventional methods of DNA sequencing, which may one day revolutionize the field of DNA sequencing. Applications span numerous research interests, including sequence variation studies, comparative genomics and evolution, forensics, and diagnostic and applied therapeutics. Integration of multidisciplinary technologies will translate into practical and affordable sequencing devices capable of whole-genome analyses. The highest capacity instruments currently available require 8–14 days to produce data. Unfortunately, these run times and the ensuing analysis may not permit the return of information in a suitable timeframe, relative to the patient's need for a diagnosis. Several emerging technologies show promise of delivering next-generation solutions for fast and affordable genome sequencing. However, more applications of next-generation sequencing, beyond those covered here, are yet to come. Perhaps the most interesting possibilities for the future work and development of algorithms and methods for DNA sequencing is potential to combine the results of different experiments or correlative analyses of genome-wide methylation, histone binding patterns, and gene expression, since in today's techniques time for reading is significantly decreased and trend is in using parralel DNA sequencing with multithread processes.

REFERENCES

Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. ,"A basic local alignment search tool". *Journal of molecular biology, vol.*215,pp: 403–410.

Blazewicz,J., Formanowicz,P., Kasprak,M., Markiewicz,W.T. and Weglarz,J., " DNA sequencing with positive and negativeErrors". *Journal of computational biology. Vol.*6, pp:113–123.

D.Huson, "Overview of DNA sequencing" Journal of bioinformatics, pp. 145–167, Jan. 2010.

D.Lavenier, "Ordered index seed algorithm for intensive DNA sequence comparison", IEEE, 2008.

Dramanac,R., Labat,I. and Crkvenjakov,R, " An algorithm for the DNA sequence generation from *k*-tuple word contents of the minimal number of random fragments", journal of molecular biology, vol.8, pp 1085-1102.

E.Mardis, "A decade's perspective on DNA sequencing technology" Nature, vol.470, pp1-6, 2011

F.Sanger, S.Nicklen, and R.Coulson "DNA sequencing with chain-terminating inhibitors" Proc.Natl.Acad.Sci. USA,vol 74., pp. 1-5.

Fleischner,H, " *Eulerian Graphs and Related Topics*" Elsevier,Amsterdam.

Fologea, D., Gershow, M., Ledden, B., McNabb, D.S., Golovchenko, J.A., and Li, J, ". Detecting single stranded DNA with a solid state nanopore". *Nano Lett.* pp:1905-1909,2005.

G.pavesi, G.mauri, G.Pesole, "An algorithm for finding signals of unknown length in DNA sequences" Journal of bioinformatics, vol.17, pp.207-214, 2001.

Gotoh, O, ". An improved algorithm for matching biological sequences", Journal of molecular biology. *vol* 162, 705–708.

J. Blazewicz, P.Formanowicz, F.Guinand and M.Kasprzak, "A heuristic managing errors for DNA sequencing" *Bioinformatics.*, vol. 18, May 2002, pp. 632-650.

J.Blazewicz, J.Kaczmarek, M.Kasprzak, “Sequential and parallel algorithms for DNA sequencing”, Cabins, vol.13, pp 151-158, 1997

J.Shendure, G.Porreca, G.Church “Overview of DNA Sequencing Strategies” Current protocols in molecular biology, pp. 1–11. Jan. 2008.

J.Shendure, H.Ji, “Next-generation DNA sequencing” *Nature biotechnology*, vol. 26, pp. 1135–1145, Oct. 2008.

J.Zhang, L.Wu and X.Zhang, “Reconstruction of DNA sequencing by hybridization”, Journal of bioinformatics, vol.19, pp14-21, 2003.

L.Franca, E.carrilho, and T.Kist “A review of DNA sequencing techniques” Quaterly reviews of biophysics, vol 2., pp.169-200. 2002.

L.Li and S.Khuri, “A comparison of DNA fragment assembly algorithms”

M. Metzker, “Emerging technologies in DNA sequencing”, CSH press genome research, vol15., pp 1767-1776.

Neil C.Jons, Pavel A. Pevzner, “An introduction to bioinformatics algorithms” Massachusetts Institute of Technology, 2004.

R.Idury and M. Waterman, “A new algorithm for DNA Sequence Assembly” Journal of computiational biology, vol. 2, pp. 291–306, Jan. 1995.

S.Bokhari, J.Sauer, “A parallel graph decomposition algorithm for DNA sequencing with nanopores” Journal of bioinformatics, vol.21, pp. 889-896, 2004.

S.Grumbach, F.Tahi, “A new challenge for compression algorithms: genetic sequences”.

S.Shataba, A.Rahman, “Algorithm in Bionformatics,” Department of Computer Science and Engineering, Banglades University of Engineering and Technology.

V.Bansal, “A statistical method for the detection of variants from next-generation resequencing of DNA pools” Journal of bioinformatics, vol.26, pp.318-324, 2010.

Waterman, M.S. , “*Introduction to Computational Biology” C*hapman and Hall, New York.

X. Chen, S. Kwong, M.Li, “A compression algorithm for DNA sequences and its applications in genome comparison”.

Z. Zhang, S. Schwartz, L. Wagner and W. Miller “A greedy algorithm for aligning DNA sequences”, Journal of computational biology, vol.7, pp.203-214, 2000.