International Journal of Applied Science and Advance Technology
January-June 2012, Vol. 1, No. 1, pp. 25-28

REVIEW OF TECHNIQUES FOR GENE SEQUENCING, ANNOTATION AND COMPARATIVE GENOMICS

Mohammed Onimisi Yahaya[1,2]

Abstract: The availability and complete sequencing of many organisms has made comparative analysis of gene a new field of research. The explosion in sequenced genome data on daily basis made this task an enormous one. Several techniques and methods have been devised and applied to carry out genome comparison. In this work, we surveyed and presented an overview of common methods, techniques, tools and challenges of Gene Sequencing, Annotation and Comparative genomics

Keywords: Gene, Sequencing, Genomics

I. INTRODUCTION

Comparison has been a major method in drawing inferences, functions and evolution or changes in many organisms. Comparative genomics is the study of the relationship of genome structure and functions across different biological specie or strain [1].The fundamental of comparative genome analysis is to establish a correspondence between gene and other genome feature in different organism and identify any set of conserved gene and regulatory motif [1].Our objective is to find sequences that have less divergence than the others, they may have a functional role to play. The hope is similarity in identity may mean similarity in function. The availability of many whole genome sequences, comparative analysis has become a powerful tool in the study of evolution, to identify relationship between organisms, gene finding and so on. Comparative genomics traditionally concentrated on inferring function of protein, but recently, it has also played a great role in determining regulatory motifs. This can be attributed mainly to the success of the human genome project. The primary reasons for the entire human genome project were determination of intron/exon structure of all genes, complete sequencing of all genes, genes mapping and other sequences, reveal non coding regulatory sequences, Identification of polymorphisms, development of methods for other genomes and to be able to uncover the unexpected [2]. In order to discuss the idea of comparative genomics in general, we will begin our work with the presentation of sequencing techniques which is the starting point and most important part of any aspect of Bioinformatics.

[1] Mathematical Sciences Program, Abubakar Tafawa Balewa University, Bauchi, Nigeria, *E-mail: mdonimisi@gmail.com*

[2] Information and Computer Science Dept. King Fahd University of Petroleum and Mineral Dhahran, Saudi Arabia, *E-mail: mdonimisi@kfupm.edu.sa*

2. SEQUENCING METHODS

In this section, we present an overview of methods and techniques for sequencing which serves as prelude to most other aspects such as gene finding, gene identification, gene annotation and comparative genomics. Sequencing is the process of determining the exact order of the chemical building block (called base pairs *A*, *T*, *C*, *G*) of organisms. For example, the human has nearly 3 billion base pairs. How enormous is the task of determining the exact order? Maxam-Gilbert method [3].This involves chemical cleavage at different specific nucleotides of four samples of an end-labelled DNA restriction fragment creating subfragments that can easily be separated into subfragment by gel electrophoresis It is the first methods to sequence DNA molecule of length upto 500 bp. But the major drawback of this techniques, is it reliance on the use of toxic chemicals.Sanger's method [4] -also referred to as dideoxy sequencing or chain termination, is based on the use of dideoxynucleotides (ddNTP's). Starting from a synthetic 5'-end-labelled oligoideoxynucleotides as a primer; the single DNA sequence to be sequenced is extended by sequence of polymerization reaction.In addition to the normal nucleotides (NTP's) found in DNA. Dideoxynucleotides are essentially the same as nucleotides except if it contains a hydrogen group on the 3' carbon instead of a hydroxyl group (OH). These modified nucleotides, when integrated into a sequence, prevent the addition of further nucleotides [5].These approaches are classified as the older methods of sequencing, but a version of the Sanger method is now the most commonly used because it practicability, efficiency and ease of automation.

Whole-genome shot gun sequencing: One of the strategies suggested for human genome project. This involves sequencing of random small clone fragment called reads [4] from both sides of the strand. A variation of this method preferred by the Human Genome Project is the hierarchical shotgun sequencing method [6]. In this

approach, genomic DNA is cut into pieces of about 150 Mb and inserted into BAC(bacterial artificial chromosomes) vectors, transformed into E. coli where they are replicated and stored. The BAC inserts are isolated and mapped to determine the order of each cloned 150 Mb fragment. The advantage of this is that sequencers are less likely to make mistakes when assembling the shotgun fragments into contigs. Others are detection of large numbers of DNA polymorphisms, more complete and less artifactual coverage of the genome, and improved speed and cost.

(*a*) *Clone-By-Clone Sequencing*: In this approach the genome is divided into sections of 150 kb. Each section is fragmented, sequenced, and assembled. Then the sequenced sections are appended to reconstruct the entire genomic sequence, the fragments are first aligned into contigs; it is also called directed sequencing of RAC contigs. A contig consists of a series of clones that contain overlapping pieces of DNA covering a specific region of a chromosome or even the entire chromosome. Contigs are usually constructed using BAC (bacterial artificial chromosome) and cosmid clones. The general approach in creation of contigs is to identify clones that have adjacent DNA segments from the chromosome, e.g., chromosome jumping, chromosome walking, etc.

(*b*) *Sequence tagged connector sequencing*: This approach is similar to the clone-by-clone approach but involves sequencing the ends of more than half a million BACs (bacterial artificial chromosome). Since the BAC endpoints are cleaved by restriction enzymes, they are not distributed identically over the genome.

(*c*) *Pyrosequencing*: This technique is based on sequencing-by-synthesis [7]. Sequencing of DNA is carried out through detection of enzymatic activity to identify the bases. This unique method of short-read DNA sequencing, its ease of use, sequence validation and flexibility make it ideally suited for applied genomics research.

(*d*) *Next generation sequencing*: This is the current technology for sequencing that uses cost effective high throughput sequencing. This arises from the unprecedented increase in the requirement for throughput sequencing that led to the development of this automated capillary electrophoresis. The techniques sequenced nucleotide by arraying several thousands of sequencing template in either picotiter template or aparose thin layers. So that the sequence can be analyze in parallel using massive parallel computers. These strategy reduces, necessary reaction volume and extensively increases the number of sequencing reaction [8].The challenges introduces by the technology is sequencing fidelity, read length, infrastructure cost and know how to handle this enormous data. Commercial providers of Next generation sequencing system are Roche's (454), GS FLX Genome Analyzer by Roche applied Science, Illumina's Solexa 1G sequencer [9] and applied Biosystem's Solid.

(*e*) *Third generation sequencing* (*Next-Next generation Sequencing*)

The third generation of sequencing technology sees single molecules of DNA being sequenced without the need for cloning or PCR amplification and the inherent biases these procedures introduce. There are generally two Maintaining the Integrity of the Specifications forms of detection methods for single molecule sequencing: those that rely on fluorescence and CCD capture, and those that don't. Instruments that use the first of these detection methods include the Helicos Heliscope, Pacific Biosciences single molecule real time sequencing (SMRT) machines and Life Technologies-VisiGen system, which relies on fluorescence resonance energy transfer (FRET). The Pac Bio platform average reads is over 1000 bases long reaching and exceeding the read length of Sanger sequencing.

3. GENOME ASSEMBLY

In this section, we present some genome assembly techniques and tools, which also is an integral part of comparative genomics. This involves base calling and reconstruction of the fragment into meaningful order and determination of 'junk' non coding DNA. In genome assembly, the assembler must account for insertions, deletions, inversion and sequence divergence. So far, there is no single assembler that can handle the above challenges.

(*a*) *Overlap- Layout consensus approach*: In this approach, reads are compared with one another to identify overlapping region. Then pair wise alignment and overlapping graphs are created, from the graph layout, a single path (Hamiltonian path- a path that visit each edge in a graph exactly one [10]. a multiple sequence alignment and consensus is generated. Examples of tools for these are Celera, Newbler, Arachne [11] and so on.

(*b*) *Hierarchical Approach*: In this approach mapping of overlapping clones (usually BAC) is carried out using fingerprinting analysis for identifying the clones. Note that a substantial number of overlap will occur. Each clone typically (4-200kb) fragment and sequenced using shortgun approach[4]. Each reads quality can be evaluated using phred [12], Trace tuner. The level accuracy determine from the quality score is used for alignment evaluation and assembly generation. The sequence can then be assembled to recreate the insertion sequence of the clone. Part of Phred called Phrap [13] is mostly used to do this. Most times gaps exist in the insert clones, finishing or gap closure is required to close the gaps that is-human manual intervention typically with the use of PCR-based approach. The whole genome can then be combined or assembled by aligning adjacent sequences and determine a path through avoiding redundancy. Usually evaluation map is used to guide the alignment called tiling path (TPF) [14]. Examples of programs to use for these are GigaAssembler [15] and TPF analyser [14].

4. WHOLE GENOME ASSEMBLY

Here the entire genome is fragmented to form libraries of varying inserts ranging from (2, 4, or 6kb) for smaller size, (10-40kb) for intermediate and >100 for large size. The ends are sequenced generating sequence reads. The Whole genome assembler used for bacterial, viral and BAC clone were Phrap. TIGr assembler [16], and CAP3. These are the older tools; recent programs that were successfully used for large genome are Celera Assembler [2] Arachne [11] and PCAP [17].

(*a*) *Greedy Assemblers*: This is a simple but effective techniques were the assembler greedily adds reads that are similar to form a long contigs. Here missassembly may occur incase complex repeats since local information is used to add reads.

(*b*) *Align- Layout consensus*: The availability of already assembled genes in databases is explored to align a new read by inference through a process called comparative assembly.

(*c*) Eulerian path- This is based on earlier attempts of sequencing by hybridization. Instead of generating reads, strings of *k*-mers are generated. The *k*-mers are represented with graphs. An Eulerian path in the graph provides a viable assembly of the genome represented [18].

(*d*) *De Bruign Graph*: This approach does not require all reads to be compared to all other reads. Programs used for this are Euler, Velvet, Allpath, Abyss, SOAPdenovo. For details discussion on this approach see [19].

Hybrid approach: This approach combines the whole genome sequencing (WGS) and hierarchical approach. It involve complementing the clone based WGS due to low coverage of the method and the whole genome is added to provide a wide range of coverage called e-BAC (enriched BAC). An example program that uses this technique is Atlas [20].

5. GENOME ANNOTATION

This is one of the main steps in understanding gene. Annotation is the process of interpreting raw sequence data into useful biological information or is the process of attaching biological information to sequences [21]. Annotations describe the genome and transform raw genome sequences into biological information by integrating computational analyses, other biological data and biological expertise. It consists of two main steps: identifying elements on the genome, a process called gene prediction, and attaching biological information to these elements. Automatic annotation tools exist that try to perform all this by computer analysis, as opposed to curation i.e manual annotation which involves human expertise. Ideally, these approaches co-exist and complement each other in the same annotation.

Basically genome annotation has three levels: nucleotide level annotation, protein level and process level annotation. Nucleotide level annotation has sub level as gene mapping: these identifying known genes, markers and landmark within the genome using BLAST. At Protein level, characterization is done, genes are named function assigned by means of comparison to already sequenced genome .this may results into unknown protein and hypothetical protein. Structural annotation consists of the identification of genomic elements such as intron and exons, ORFs and their localization, gene structure, coding regions, location of regulatory motifs. Functional annotation consists of attaching biological information to genomic elements such as biochemical function, biological function and expression. Process level, biological processes affected by the genes are identified Process category could be cell cycle, cell death, immune response, cell metabolism. These are inferred through information that is already available. GO [21] and panther have a naming convention.

Closing the genome: this is the last part of any genome sequencing project. It is carried out to ensure accuracy and reliability. Programs such as Euler and Arachne are assembler with error checking capability. Other approaches uses error correction algorithm such as Autofinish, CONSED [22], MisEd, ReDit and so on.

6. COMPARATIVE GENOMICS

Comparative genomics is the study of the relationship of genome structure and functions across different biological specie or strain. It is the natural step that follows genome sequencing. It can also be viewed as the analysis, comparison of genetic material from different species to study evolution, gene function, and inherited disease. Comparative genomics provides a platform for understanding the uniqueness between different species. Comparative genomics exploits both similarities and differences in the proteins, RNA, and regulatory regions of different organisms to infer how selection has acted upon these elements. For example, similarity in structure might infer similarity in functions. The main purpose is to gain a better understanding of how species have evolved and to determine the function of genes and non coding regions of the genome. Comparative genomics faces theoretical problems. The fundamental part of comparative genomics is alignment; hence the problems in alignment are also the problems of comparative genomics. How do we place regulatory elements?, how can we determine the evolutionary dynamics, how to determine homologs, paralogs. The commonly used software is PSI-BLAST, a new version of BLAST that uses score and e-value to determine similarity.

7. CHALLENGES

Though significant progress has been made in the field of Bioinformatics research and application, it still faces a lot of challenges such mining genomic data. It takes different techniques and rigorous activity to sequence a genome,

managing the data generated is one major challenges due to lack of technical know, difficulty in training Bioinformatics expert, the storage of data in different format also complicated the field. Other challenges are those that are inherent to the field due to nature of gene, gene duplication, assembly, transposon (jumping genes), location of coding and non coding part of a genome just a mention a few.

Table 1
Definition peculiar to comparative genomics

Homology: Trait is any characteristic of organisms that is derived from a common ancestor.

Homologs: Gene or protein with similar sequences that can be attributed to common anscestor

Paralog: Homologous sequences separated by a gene duplication event. They have evolved to perform different function

Orthologs: Homologues series that have evolved from common ancestor by speciation. Mostly they evolved to perform similar function

Xenologs: Homologs resulting from horizontal gene transfer between two organisms. i.e relationship by inter specie

Analog: Non- homologous gene that have descended convergently from an unrelated anscestor

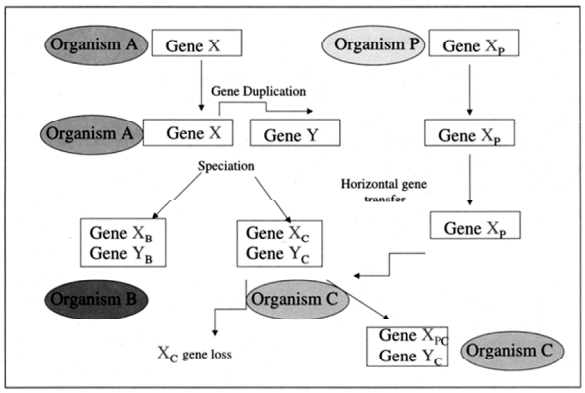

Figure 1: Schematic Representation of Difference Between Homolog, Ortholog, Paralog. See [23]

REFERENCES

[1] Manolis K., P. Nicks, et al. "Methods in Comparative Genomics: Genome Correspondence, Gene Identification and Motif Discovery". *MIT*. 2004.

[2] Weber J.L. and E.U. Meyer, "Human Whole Genome Shotgun Sequencing". *Genome Research*, **7**, pp. 401-409. 1997.

[3] Maxam A.M. and W. Gilbert, "A New Method for Sequencing DNA". *Proc. Natl. Acad. Sci. USA,* **74(2)**, 560-564. 1977.

[4] Sanger F., S. Nicklen, et al. "DNA Sequencingwith Chain-terminating Inhibitors". *Proc. Natl. Acad. Sci. U.S.A,* **74(12)**, pp. 5463-5467, 1977.

[5] Speed T. (2003). "Sanger Methods". erk:// statwww.berkeley.edu/users/terry/Classes/s260.1998/).

[6] Anderson S. "Shotgun DNA Sequencing Using Closed DNase I-generated Fragments". *Nucleic Acids Research,* **9(13)**, 3015-27. 1981.

[7] Ronaghi M. "Pyrosequencing Sheds Light on DNA Sequencing". *Genome Research*, **11(1)**, pp. 3-11. 2001.

[8] Stephen C.S. "Next Generation Sequencing Transform Today's Biology". *Nature*, **5(1)**, pp. 16-18. 2008.

[9] Potera C. "New Gene Sequencer Targets Productivity-Solexa Says Its Novel System Offers Better Cost Effectiveness Via Use of Short-reads Sequences". *Genet Eng News*, **26(17)**, 2006.

[10] Alsuwaiyel M.H., "Algorithms: Design Techniques and Analysis", *World Scientific*. 2006.

[11] Batzoglou S. e. a. "ARACHNE: A Whole-genome Shotgun Assembler". *Genome Research* **12(1)**, pp. 177-89, 2002.

[12] Ewing B. e.a. "Base-calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment". *Genome Research,* **8(3)**, pp. 175-85, 1998.

[13] de la Bastide, M. and R. McCombie, "Assembling Genomic DNA Sequences with PHRAP". *Current Protoc Bioinformatics*. 2007.

[14] Agarwala R. Assembling the Human Genome (Chapter Nine) in Handbook of Computational Molecular Biology. 2003.

[15] Kent W.J. and D. Haussler . "Assembly of the Working Draft of the Human Genome with GigaAssembler". *Genome Research*, **11(9)**, 1541-8, 2001.

[16] Sutton G.G., O. White, et al. "TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects". *Genome Science and Technology*, **1(1)**, pp. 9-19, 1995.

[17] Huang X. e.a. "PCAP: A Whole-genome Assembly Program". *Genome Research* **13(9)**, 2164-70, 2003.

[18] Pevzner P., e.a. "An Eulerian Path Approach to DNA Fragment Assembl". *Proc Natl Acad Sci.* USA **98(17)**, pp. 9748-53, 2001.

[19] Miller J.R. e. a. "Assembly Algorithms for Next-generation Sequencing Data". *Genomics*, **95(6)**, pp. 315-27. 2010.

[20] Havlak P. "The Atlas Genome Assembly System". *Genome Research*, **14(4)**, 721-32, 2004.

[21] Stein L. "Genome Annotation: From Sequence to Biology". *Nature Reviews Genetics*, **2(7)**, 493-503, 2001.

[22] GO (www.ebi.ac.uk/GO/), 2011.

[22] CONSED www.phrap.com/consed/

[23] Anand K.B. "Comparative Genomics: A Powerful New Tool in Biology". *Resonance,* **11(8)**, 22-40, 2006.