# Second Generation Sequence Assembly Review

Sun Zhao

January 6, 2013

**Abstract**

The abstract abstract.

# 1 Before Everything

I would like to first explain the motivation of this article at the beginning. The theme of this article is about sequence assembly which is a computer aided process for constructing gene sequence. It is actually belonging to the scope of bioinformatics. As a pure computer science undergraduate student, I have started researching in this field since the summer in 2010. After reading lots of papers, discussing with biology researchers from China as well as foreign ones, trying kinds of bioinformatic tools, I approached nothing new but valuable experiences of it. In this article, I will answer the questions of "what is gene sequencing", "what is sequence assembly and its challenge", "how to assembly sequence", "how to evaluate sequence assembly softwares" and my own contribution to sequence assembly in a computer science researcher's perspective. I'am not going to tell the deep details but giving initiations about what to do and how to do about sequence assembly. Moreover, I will provide valuable references of related papers and materials to help you get a full view of sequence assembly.

# 2 What is gene sequencing

In genetics and biochemistry, sequencing means to determine the primary structure of an unbranched biopolymer. The biopolymer can be DNA, RNA, protein. In this article, I will focus on DNA and RNA sequence analysis excluding protein. DNA sequencing is the process of determining the nucleotide order of a given DNA chain. Concretely, DNA is a chain of four types nucleotide, represented by letter of 'A', 'G', 'C', 'T'. DNA sequencing is trying to produce the corresponding string of 'A', 'G', 'C', 'T' for a sample DNA chain. However, the most popular biology sequencing method called shot gun, randomly cut the original DNA chain into fragments and a set of 'A', 'G', 'C', 'T' strings. Each nucleotide string which is the sequence of a fragment is called a read. To increase the read coverage and read quality, copies of DNA made by PCR amplification with a typical bacteria template are sequenced. Fig. 1 shows the shot gun process, and you may find that reads are sequenced from the two ends of DNA fragments instead of the complete one. The end sequencing phenomena is caused by biology sequencing methods, however, if particular reaction and methods are included, the distance between the two ends can be estimated. In this case, the two end reads are called pair-end reads and the distance is called insert length.

DNA molecules are double-stranded helices, consisting of two long complement strands. According to base paring principle–'A' complements with 'T' and 'G' complements with 'C', the sequence of a strand can be inferred from its opposite

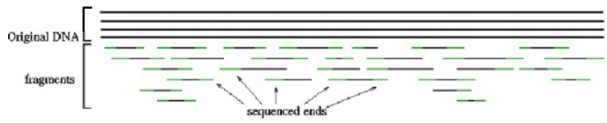


Figure 1:

one. Particular sequences denoted by 3' and 5' in the DNA strand is used to specify the orientation of the two strand, and for simplicity, if one strand is specified as 3' to 5', then the opposite one is 5' to 3'. Example 1 shows a double strain DNA fragment with pair-end reads(red string). Note that the pair-end reads are positioned on opposite strand and will be sequenced all from 3' to 5'. So the two pair-end reads string should be 'AGCTAA' and 'GCCAA'.

3'→5'
AGCTAATGCTATCTTGGC
TCGATTACGATAGAACCG
5'→3'
Example 1

Sequencing methods and platform is developing as time goes. The typical genome analyser of first generation is Sanger[4] producing read lengths of approximately 800bp (typically 500-600bp with non-enriched DNA).Recently, new sequencing methods have emerged [2]. Commercially available technologies include 454 Sequencing [3], Illumina genome analyser [1] and SOLiD sequencing(www.appliedbiosystems.com). Compared to traditional Sanger methods, these technologies function with significantly lower production costs and higher throughput. However, the reads produced by these next-generation sequencing technologies are much shorter than traditional Sanger reads, currently around 400-500 base pairs (bp) for 454, 50bp for Illumina and 35bp for SOLiD. Because of their length, they must be produced in large quantities and at greater coverage depths than earlier sequencing projects.
Reads are saved as a file by sequencing chip and the most popular format of read file is FASTA and FASTQ. A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line (def-line) is distinguished from the sequence data by a greater-than (">") symbol at the beginning. It is recommended that all lines of text be shorter than 80 characters in length. An example sequence in FASTA format is shown in Fig. 2.

>AB000263 |acc=AB000263|descr=Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.|len=368
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGG
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGGACAGAATGCC
CTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAG
TTTAATTACAGACCTGAA

Figure 2:

# 3 What is sequence assembly

# Reference


[1] D.R. Bentley. Whole-genome re-sequencing. *Current opinion in genetics & development*, 16(6):545–552, 2006.

[2] E.R. Mardis et al. The impact of next-generation sequencing technology on genetics. *Trends in genetics*, 24(3):133, 2008.

[3] M. Margulies, M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005.

[4] F. Sanger, GM Air, BG Barrell, NL Brown, AR Coulson, JC Fiddes, PM Slocombe, and M. Smith. Nucleotide sequence of bacteriophage (d x174 dna. 1977.