Volume 3, No. 7, Nov-Dec 2012
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
RESEARCH PAPER
Available Online at www.ijarcs.info

DNA Sequencing by Hybridization and Shotgun Sequencing Technique

Narayan Kumar Sahu*
M.Tech Scholar
Department of Computer Science and Engg.
Disha Institute of Mgmt. and Technology
Raipur, Chhattisgarh, India
narayansah@gmail.com

Prof. Somesh Kumar Dewangan
Reader
Department of Computer Science and Engg
Disha Institute of Mgmt. and Technology
Raipur, Chhattisgarh, India
somesh_4@yahoo.com

Abstract: DNA sequencing includes several methods and technologies that are used for determining the order of the nucleotide bases adenine, guanine, cytosine, and thymine in a molecule of DNA.
Knowledge of DNA sequences has become indispensable for basic biological research, other research branches utilizing DNA sequencing, and in numerous applied fields such as diagnostic, biotechnology, forensic biology and biological systematic. The advent of DNA sequencing has significantly accelerated biological research and discovery. The rapid speed of sequencing attained with modern DNA sequencing technology has been instrumental in the sequencing of the human genome, in the Human Genome Project. Related projects, often by scientific collaboration across continents, have generated the complete DNA sequences of many animal, plant, and microbial genomes.
The first DNA sequences were obtained in the early 1970s by academic researchers using laborious methods based on two-dimensional chromatography. Following the development of dye-based sequencing methods with automated analysis DNA sequencing has become easier and orders of magnitude faster. Common challenges of DNA sequencing include poor quality in the first 15–40 bases of the sequence and deteriorating quality of sequencing traces after 700–900 bases. So we have a challenge for a good quality after sequencing [1] [7] [10].

Keywords: DNA, SHOTGUN, SBH, GA.

I. INTRODUCTION

DNA sequence assembly is one of the core areas in bioinformatics. It involves correctly aligning and joining DNA fragments to derive genome or plasmid sequences.

In Molecular, biology has been driven and guided by knowledge of DNA. The modern history began in 1953 with the discovery of the three-dimensional structure of DNA by Crick and Watson. Isolation of restriction enzymes and DNA polymerases made possible DNA sequencing, the determination of the base sequence along a strand of DNA. The modem era of "rapid" DNA sequencing began in 1977 with the development of two techniques: di-deoxy chain termination (Sanger et *01.:* 1977) and chemical degradation (Maxam and Gilbert, 1977). Beginning with Smith *et al.* **(**1986**),** DNA Sequencing machines were developed that automated gel electrophoresis, data acquisition, and base determination.

The DNA molecule to be sequenced is ***L*** base pairs (bp) in length. The sequence of one strand is a = *a1….a2...*al Generated at random are *N* shorter fragments *f1, f2* **. . .*fN*** each of approximate length ***l*** << ***L*** whose base sequences can be determined experimentally.

Currently, there are various types of sequencing technologies and some emerging ones. The Sanger dideoxy sequencing technology has been widely used since 1970; this sequencing technology produces DNA fragments of length between 400-900 bp ("long reads"). There are sequence assembly algorithms that have been designed to work with this type of data. However, since the year 2000, new sequencing technologies have emerged (and some are emerging); these are known as the next generation sequencing technologies. Some of these technologies include 454 sequencing, Illumina sequencing, Helicos sequencing, and Solid sequencing; and they produce DNA fragments of varying lengths from 35-400 bp ("short reads"). There is also sequence assemblers designed to work with these newer sequencing technologies.

When dealing with a sequence assembly project that involves a combination of the sequencing technologies, it is important to have a proven strategy to assemble the data sets from the mixed data sources. This research work involves experiments on DNA sequencing by SHOTGUN and SBH sequencing technology.

The aim of this research is to assemble the data sets from the mixed data sources. This research work involves experiments on DNA sequencing by SHOTGUN and SBH sequencing technology for sequence assembly. This helps contribute to establishing good methodologies for sequence assembly from DNA read data of different lengths. It also helps determine the best (or most appropriate) sequence assembler to use in various situations.

II. METHODOLOGY

In this paper we have proposed a solution for DNA sequence assembly problem using GA [2]. For solving any optimization problem we have to first formulate the problem according to optimization problem. In this case first we formulate the DNA sequence assembly problem according to GA. Next subsection describes how we formulate the DNA sequence assembly problem.

To solve the problem, representation of the individual and fitness value is required. GA is based on population (candidate

solution) and each population have its own fitness value according to which it is compared from others, so we have to first represent the DNA sequence assembly problem in terms of GA.

III. INDIVIDUAL REPRESENTATION

In DNA sequence assembly problem inputs are the set of fragments which need to be assembled and build a common sequence which does not have any repeated pattern. The common sequence is considered as output for DNA sequence assembly problem. To find out the function or property of specific genes, the reading of nucleotide or chemical base (A, T, C, G) sequence is done. Large nucleotide sequences are called DNA sequence. The large DNA sequence consists of repeated patterns of nucleotide that's why the DNA sequence becomes large. In DNA sequence assembly repeated patterns are removed and one consensus sequence builds.

In DNA sequence assembly large DNA sequence of a particular gene is taken for assembly process. Large DNA file is split randomly in different fragments of DNA sequence which are used in assembly process. After getting the set of fragment, fragments are aligned and the longest match between the suffix of one sequence and the prefix of another is determine. All possible pair combination of fragments is compared and matching score is determined. On the basis of matching score fragment order is determined. At last the consensus sequence is found out from the fragment order. We have performed experiments with the nucleotide sequences of homosapiens (human) and mouse viz. MACF1, TNFRSF19 and ZFA. The DNA data is taken from the NCBI.

We have solved DNA sequence assembly problem using the continuous version of GA. In DNA sequence assembly problem fragment order in which fragments are aligned is very important but very hard to find the best order from the large possible combinations of fragment order. Using GA we determine the fragment order. GA is based on the concept of population and each individual represents a solution for a problem. In case of DNA sequence assembly problem each individual of GA represents the fragment order on which fragments are aligned to find out a consensus sequence. Each individual has certain dimension value for DNA sequence assembly problem each individual has a dimension value equal to the number of fragments taken for assembly.

DNA sequence assembly is a discrete optimization problem. In the proposed solution continuous version of GA is used instead of discrete version. To change the continuous version to real version for DNA sequence assembly problem SPV rule is used. Using the SPV (shortest position value) rule continuous position generated by GA is converted to discrete value.

Each individual or particle of GA is represented by a Position vector *Xid*= {x1, x2, x3 ...*xd}* where *i* is the particular individual and *d* represents the dimension index. Each individual of GA contain the real value for a particular dimension and on the basis of this real values new sequence vector is generated using shortest position value rule (SPV). New generated sequence vector using SPV is represented as *Sid= [fi1, fi2 ….fid]*. *Sid* is a fragment order of *i* particle in the processing order containing *d* dimension and *fi1, fi2* represent the fragment number in a fragment order.

For example the individual generated by PSO is *Xid*= {4.83, -.55, 1.90, 4.46, 1.05, 2.47, -1.28, 0.192, 3.56, 2.28} which has dimension value equal to 10 that means the number of fragments taken is 10. It is clear that *Xid* contains the real values and for DNA sequence assembly we need fragment sequence order from the set of possible combination of fragment order. SPV rule is used to generate the new sequence vector *Sid.* Dimension values of *Xid* is used to generate sequence vector, the dimension index which has the shortest value in *Xid* represents the first fragment that is *f0*, second shortest value represents the second fragment and so on. The sequence vector *Sid* generated for *Xid* using SPV is {9, 1, 4, 8, 3, 6, 0, 2, 7, 5} here 9 represents the fragment 10 and 1 represents the fragment 2 and so on. *Sid* represents the fragment order in which fragments are aligned for determining the consensus sequence. For each individual of GA, sequence vector is calculated using the SPV rule.

IV. FITNESS FUNCTION

After representation of each individual we have to calculate fitness value of each individual. On the basis of fitness value we determine the optimal solution. In case of DNA sequence assembly problem optimal solution is the maximum matching score of fragment order. First we have to align the fragments according to the fragment order *Sid* then longest match between the suffix of one fragment and the prefix of another is determine. Matching score is calculated by counting the matching nucleotide of fragments. The matching score for a pair of fragment is calculated using

$$score_{i,i+1} = \begin{cases} 0, & \text{if nucleotide does not matched} \\ score_{i,i+1} + 1, & \text{otherwise} \end{cases}$$

eq.4

In eq. (4) *scorei, i+1* is a matching score of two consecutive fragments of sequence vector *Sid*, *i* and *i+1* is the index of sequence vector *Sid*. After calculating the score of fragment pair total score is calculated for a particular individual of GA. Total score is calculated by eq. (5).

$$max\ f_i(x) = \sum_{j=0}^{D-1} score_{j,j+1}$$

In eq. (5) denotes the fitness value for individual *i* of GA. In eq. (5) *max* denotes that our objective is to maximize the value of. Individual who has the maximum value of i considered as optimum solution. Fitness function is the summation of all scores calculated by eq. (4) for an individual.

V. HYBRIDIZATION SEQUENCING

Recently, a new approach to sequencing DNA was presented, sequencing by hybridization or SBH. *SBH* is a non-enzymatic method that uses a DNA microarray. A single pool of DNA whose sequence is to be determined is fluorescently labeled and hybridized to an array containing known sequences. Strong hybridization signals from a given spot on the array identify its sequence in the DNA being sequenced. Mass spectrometry may be used to determine mass

differences between DNA fragments produced in chain-termination reactions.

DNA sequencing methods currently under development include labeling the DNA polymerase, reading the sequence as a DNA strand transits through nanopores, and microscopy-based techniques, such as AFM or transmission electron microscopy that are used to identify the positions of individual nucleotides within long DNA fragments (>5,000 bp) by nucleotide labeling with heavier elements (e.g., halogens) for visual detection and recording.

Third generation technologies aim to increase throughput and decrease the time to result and cost by eliminating the need for excessive reagents and harnessing the processivity of DNA polymerase. The best automated-gene sequencers claim output of up to 7200 bases per hour. A new technique, SBH (pairing) promises to sequence more than four times faster [3] [6].

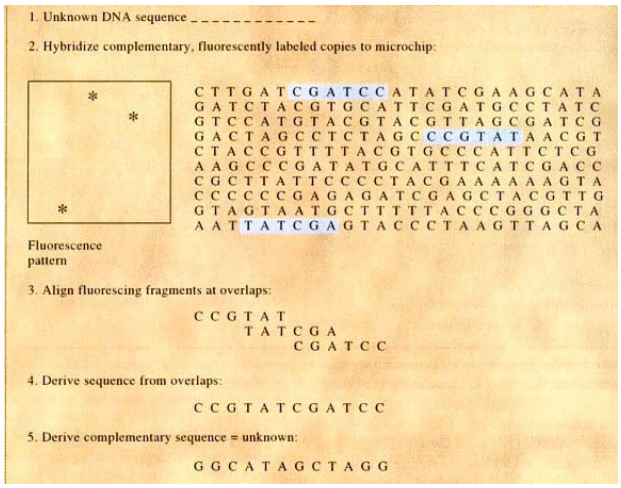


Figure.1.Sequence by Hybridization micro Array.

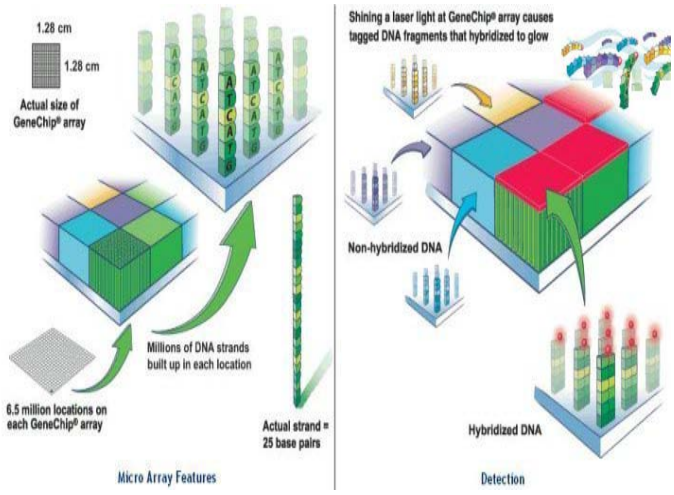


Figure.2. Affymetrix Micro Array

VI. SHOTGUN SEQUENCING

One of the great accomplishments of the late twentieth and early twenty-first centuries was the sequencing of the human genome. Biologists and geneticists were able to piece together the specific sequence of DNA bases (sub-pieces of DNA) that serve as the blueprint for human life. Remarkably, the Human Genome Project was completed ahead of schedule, in part due to advances in technology that enabled the application of new techniques that greatly accelerated the process of reading DNA sequences. In this interactive, we will get a sense of how one of these advanced techniques, called SHOTGUN sequencing, works.

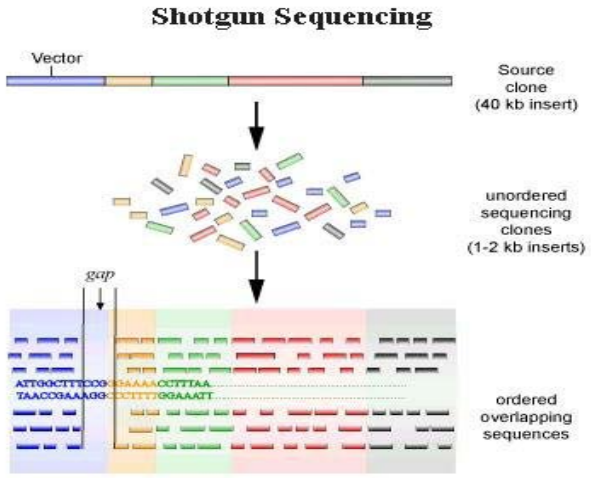
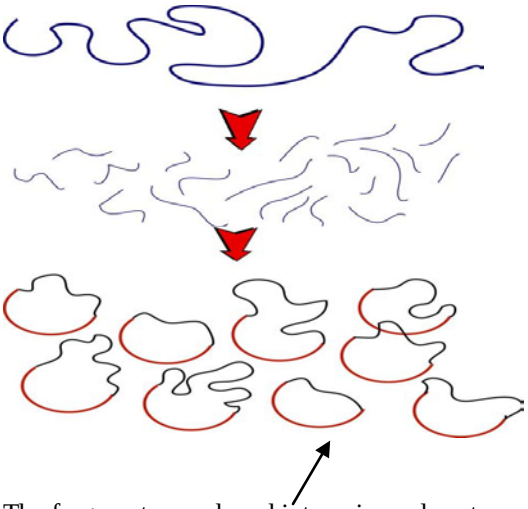


Figure.3. Shotgun Sequencing

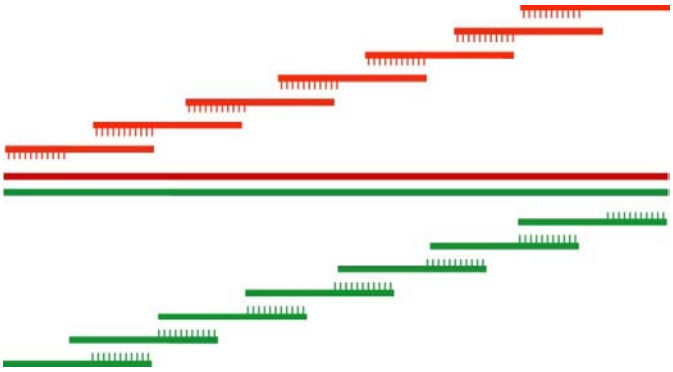
a. ***Making a shotgun library*****:** Genomic DNA is sheared or restricted to yield random fragments of the required size [4].



The fragments are cloned into universal vector

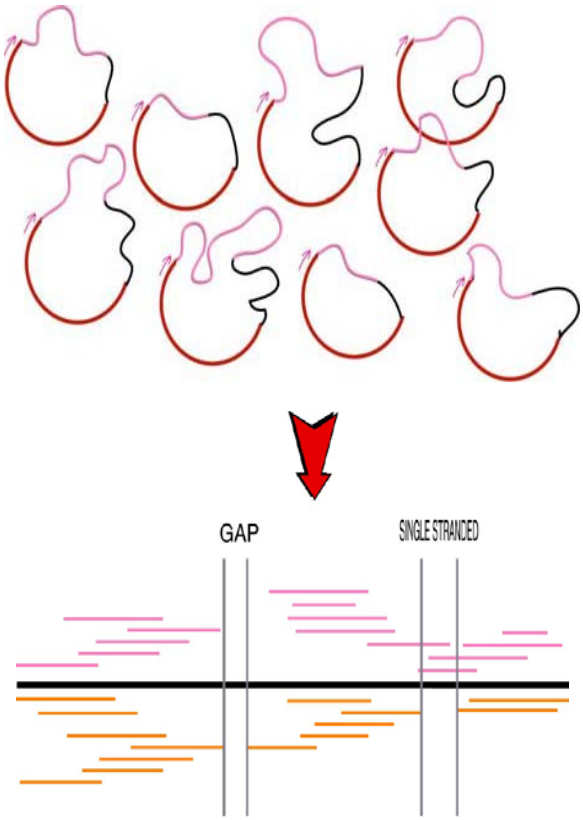
Below is a **contig** (for "contiguous sequence"). The red-green hybrid in the center is the original dsDNA to be sequenced. It was broken up, and smaller pieces cloned into plasmids. The inserts in various randomly chosen plasmids were then sequenced to give the smaller fragments shown. Note that it is important to sequence ***both strands***. While this may seem a waste of effort given the rules of Watson-Crick base pairing, the fact is that certain areas on one strand may be difficult to sequence accurately (for example, because of local secondary structure formation). The complementary strand,

however, may sequence well. Using primers from opposite ends will give you sequence for both strands.

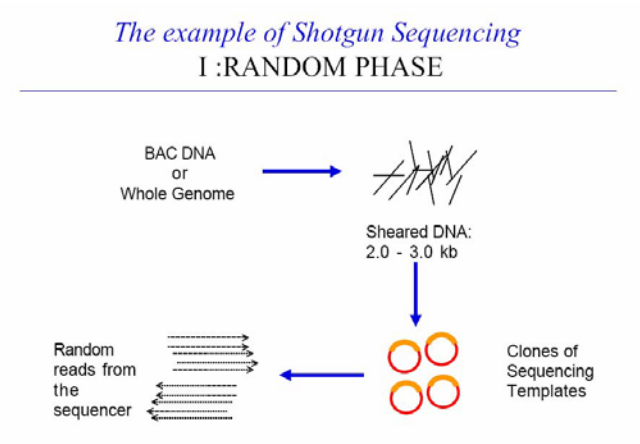


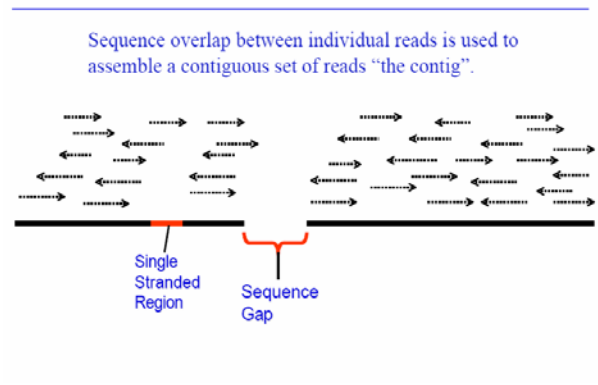
Once you have sequenced a bunch of small fragments, a computer can find regions of overlap (shown as hatch marks above) and properly align them into the complete original sequence

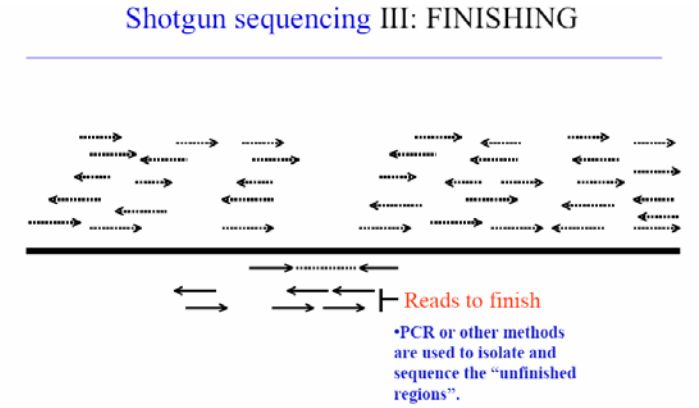
VII. SHOTGUN SEQUENCING



These sequencing reads are assembled in to contigs, identifying gaps (where there is no sequence available) and single-stranded regions (where there is sequence for only one strand).The gaps and single-stranded regions are then targeted for additional sequencing to produce the full sequenced molecule[5].







We have proposed a new algorithm for DNA sequence assembly that combines the features of shotgun Sequencing and sequencing by hybridization [8] [9]. The algorithm takes advantage of high coverage and low sequencing error rates made possible with the advent of advanced DNA sequencing machines.

VIII. CONCLUSION

We have proposed a new algorithm for DNA sequence assembly that combines the features of shotgun sequencing and sequencing by hybridization. The algorithm takes advantage of high coverage and low sequencing error rates made possible with the advent of advanced DNA sequencing machines. Table 1 shows the results of our experiment [11]

[12] [13].First, we generated a random DNA sequence of length 20,000 bp. We then generated a fragment set of mean length 400 bp for this sequence, with a mean depth of coverage of 7 using the enzyme program of the GenFrag utility (Engle and Burks, 1993), version 2.1 (Engle and Burks, 1994). Sequencing errors were simulated on this fragment set with error rates ranging from 0.5% to 3.0% using the mutate program of the same utility. Each entry in Table 1 corresponds to a different error rate. These mutated fragments are then given as input to our prototype program. All runs were performed with k = 15. All runs were completed less than ten seconds on a SUN SPARCstation 10. Our prototype builds the sequence graph on a given fragment set and performs the graph reductions mentioned above. It does not perform Inc Eulerian tours to handle repeats, and reports the inferred sequence on each edge remaining in the irreducible graph. It does not yet perform the multiple alignments on the fragments.

The quality of the inferred contigs is judged by running the dynamic programming algorithm. Described in the section A New Algorithm for Shotgun Sequencing that fits a given contig to the true sequence (that we know already) at the appropriate location. For each entry in Table 1. We show the number of contigs generated (the prototype actually generates twice the number of contigs because of the complementarily of edges). For each contig, we computed a similarity score from the dynamic programming algorithm by giving a positive score of 1 for each match, and a negative score of I for each mismatch or indel. We also computed the number of matches or correct base calls for each contig. When there are multiple contigs, we reported the sum of similarity scores, and total number of correct calls. The length of inferred sequence covered is the sum of lengths of all contigs. Table 1 shows that our prototype, even without performing any multiple alignment and consensus, has a very high percent of correct calls (> 99.8%), and a very high similarity score in all cases.

Table1: Result of Sequence Assembly on a Simulated Data

Error Rate (%)	Number of Contigs	Similarity score	Length of inferred Sequence	Number of Correct Calls	Length of Contigs
0.5	1	19,987	19,995	19,991	19,995
1.0	1	19,962	19,972	19,968	19,972
1.5	1	19,950	19,964	19,958	19,964
2.0	3	19,913	19,949	19,934	11,520, 6,058, 2,335
2.5	4	19,904	19,945	19,927	13,147, 6,757
3.0	5	19,899	19,981	19,945	13,063, 6,836

IX. REFERENCES

[1] Bains, ii'.a,n d Smith, G.C. 1988. A novel method for DNA sequence determination. J *Theorer. Biol.* 135, 303-307.

[2] Bean, James C. 1992. Genetics and random keys for sequencing and optimization. Technical Report 92-43,The University of Michigan.

[3] Drmanac. R., and Crkvenjakov, R. DNA sequence determination by hybridization: a strategy for efficient large-scale sequencing.*Science* 263, 596-566. 1987. Yugoslav Patent Application 570.Dumas. J.P., and Ninio, J. 1982.

[4] Kececioglu. J.D., and Myers, E.W "Exact and approximation algorithms for DNA sequence reconstruction." Ph.D.' Thesis. Depart-. 1995. Cbmbinatorial algorithms for DNA sequence assembly. *Algofithmicu* 13,7-5 1.Lander, E.S., and Waterman, M.S. 1988.

[5] Lysov. Y.P., Horentiev, V.L., Khorlin, A.A.. Khrapko, K.R., Shik, V.V., and Mirzabekov, A.D. Genomic mapping by fingerprinting random clones: a mathematical analysis.*Gerioniics* 2, 23 1-239. 1988.

[6] Macevicz, S.C. DNA sequencing by hybridization with oligonucleotides. A novel method. *Dokl. Acad. Sci. USSR* 303, 1508-1511. 1989. International Patent Application PS US89 04741,

[7] Burks, C. ―DNA sequence assembly‖, Engineering in Medicine and Biology Magazine, IEEE, Vol. 13, pp. 771—773, 1994.

[8] G.G. Sutton, O. White, M.D. Adams, and A.R. Kerlavage.―TIGR Assembler: A new tool for assembling large shotgun sequencing projects‖, Genome S cience & Tech., Vol. 1, pp. 9–19, 1995.

[9] T. Ch en and S. Skiena. ― Trie-based data structures for sequence assembly‖, Combinatorial Pattern Matching, pp. 206–223, 1998.

[10] X. Huang and A. Madan. ―CAP3: A DNA sequence assembly program‖, Genome Research, Vol. 9, p p. 868–877, 1999.

[11] E.W. Myers. ―Towards simplifying and accurately formulating fragment assembly‖, Journal of Computational .

[12] P.A. Pevzner. ―Computational molecular biology: An algorithmic approach‖, The MIT Press, Vol. 1, 2000.

[13] Allex, C.F. and Baldwin, S.F. and Shavlik, J.W. and Blattner, F.R. ―Improving the quality of automatic DNA sequence assembly using fluorescent trace-data classifications‖, Proceedings, Fourth International Conference on Intelligent Systems for Molecular Biology.