De Bruijn Graph based De novo Genome Assembly

Mohammad Ibrahim Khan
Computer Science and Engineering, Chittagong University of Engineering and Technology
Chittagong, Bangladesh
muhammad_ikhancuet@yahoo.com

Md. Sarwar Kamal
Computer Science and Engineering, Chittagong University of Engineering and Technology
Chittagong, Bangladesh
sarwar.saubdcoxbazar@gmail.com

***Abstract*—The Next Generation Sequencing (NGS) is an important process which assures inexpensive organization of vast size of raw sequence data set over any traditional sequencing systems or methods. Various aspects of NGS like template preparation, sequencing imaging and genome alignment and assembly outlines the genome sequencing and alignment .Consequently, deBruijn Graph (dBG) is an important mathematical tool that graphically analyzes how the orientations are constructed in groups of nucleotides. Basically, de Bruijn graph describe the formation of the genome segments in a circular iterative fashions. Some pivotal de Bruijn graph based de novo algorithms and software package like T-IDBA, Oases, IDBA-tran, Euler, Velvet, ABySS, AllPaths, SOAPdenovo and SOAPdenovo2 have illustrated here.**

***Index Terms*—Next Generation Sequencing (NGS), deBruijn Graph (dBG), SOAPdenove2**

I. Introduction

In the age of information superhighway all the invention, discovery, analysis, synthesis, anti-synthesis and measurements are managed by various computing devices and algorithmic steps. To make proper achievement in computing analysis change should be made in algorithms not in system. In this regards, the researcher and scientist are working all over the world towards the progress of efficient planning on algorithm and automated method. So from the last few years, the researcher and scientist have made a constitutional deviation from the use of mechanical Sanger DNA nucleotide sequencing for genome segments reasoning. It is obviously correct that the mechanical Sanger process had been monopolizes from the last twenty five years and drives a number of historic achievements, comprising the outstanding outcome of full sequencing of human genome sequence. In contempt of huge development of theory and scientific inventions all along this era, the constraint of mechanical Sanger DNA nucleotide sequencing indicates the necessity of new and updated methods and scholarly mechanisms for analyzing vast volume of genomes irrespective of human, plants, virus and microorganism. Now –a-days, numbers of unique idea and vocational attempt have been established on the road to the evolution of new techniques to overcome the problems of Sanger. The continuing innovation of DNA, RNA and Proteins sequencing techniques have accompanied the development of sequencing system under the environment of sudden reduced expenditures and better outcomes in performance comparing with the systems of tow three years back. Various sequencers from some established methods like 454 Life Science or Roche or Illumina and Applied Biosystems are in developments and Helicos have appeared in application. All these sequencers will produce more and more data and consequently increase the volume of data set. To diminish the data set National Institute of Health (NIH) are going to generate proper sequence within couple of years. Consequently, the updated systems with higher throughput assure the researcher to perform the sequencing with less time and space.

A. Influence of Next Generation Technique

Sequencing of biological data set has improved swiftly due to the appearance of monumental lateral sequencing methodologies, which together called as Next Generation Sequencing (NGS). We know that there is huge difference between serial and lateral computing systems. Lateral or parallel system used to solve multiple instructions with in a second where a single instruction is solved by serial system during the same time. Since NGS methods support parallel computations, it is very easy to get high-throughput sequencing of small length at reduced cost [1], [2]. Consequently Next Generation Sequencing technologies are accelerating the Computational Biology (CB) research in different fields like genomics, metagenomics, transcriptomics, proteogenomics, gene expression analysis, DNA sequencing, non coding RNA discovery and Protein-Proteins Interactions alignments [3]. The proper genome assembly is very cumbersome as well as costly in time and space complexity. Beside this is an NP hard problem which indicates that there are not sufficient solutions for the exact assembly. The assembly problem on sequencing

doi:10.4304/jsw.9.8.2160-2168

is occurred due to the difficulties to read complete genome sequence in one read under the environment of ongoing methods for sequencing.

B. Genome Assembler Necessity

Genome Assembly is a process which combines various small segments into complete one after making subdivisions of different segments to perform certain activity. We know that Shortgun sequencing process sub-divide the complete genome into arbitrary sections and scan each section autonomously. By implementing Next Generation Sequencing technology, any sequencer can design ample amounts of DNA sequencing data set in the meanwhile of a specific run with very minimum costs. But most of the designed data set becomes exaggeration by the over density of sequencing errors as well as genomic repetitions. In this respects, to design a genome assembler for Next Generation Sequencing is one of the difficult phenomena due to the shortages of proper exact computational methods as well as systems that incorporate the computing techniques. To handle the difficulties of NGS assembly, there should have a strong and fundamental structure that helps to efficient managements. . Genome Assembler must have a proper input data set which contains two files as sequence container and score container. NGS methods have efficient and high performance small reads and to measure the files is a space occupied process. Most of the assembler use graph oriented data structures to make easier of the assembly process and salvage time and memory cost. But some assembler can vary in respect of graph formation, graph design, graph traversing and graph resolution [4].

Scientist and researcher used two distinct types of genome assembly in DNA data sequencing research. These are:

1) The Comparative Process of Genome Assembly.
2) De novo process of Genome Assembly.

Comparative Genome Assembly works based on the reference genome from the same parts of the organisms or similar species is used as a scaling to direct the assembly approach by making pair wise alignment of the genome segments for assembled data set. This process mainly used in genome re-sequencing [5]. On the other hand, de novo assembly does not maintain any scaling process and this is used to reconstruct genomes which are not identical to previously sequenced genomes [6]. Comparisons between comparative and de novo genome assembly (Table 1) is illustrate here.

TABLE I.
COMPARISONS OF DE NOVO GENOME ASSEMBLY AND COMPARATIVE GENOME ASSEMBLY

Comparative Genome Assembly	De novo Genome Assembly
Performs scaling based on reference genomes.	Works based on fragmented segments.
It combines overlap reads.	It is feasible for long data segments.
Generate Multiple Sequences alignment	It demands more complete coverage
Generate Consensus Sequences.	Generate fragmented assemblies
Small read can scale several locations	It perform good when there is no reference data set.
Repeated segments are problematic.	Two important paradigms as Overlap Graph and K-mer Graph.
Accuracy rate is high	Accuracy rate is relatively less than comparative genome assembly

C. Next-Generation Sequencing Technologies

The latest methods that assure reduced cost, accurate sequencing, alignment and high-throughput computing are called Next Generation. These update approaches and concepts arrange some blueprint that depends on the integration of template design, sequencing, pairwise genome alignment and genome assembly methods. The improvement of NGS approaches on Computational Biology, Computational Chemistry, Genetic Engineering and Genetic Birding has switched the policy we realize regarding automated and scientific process in all aspects of research and analysis. The pivotal development endeavor through Next Generation Sequencing is the possibilities to yield immense size of data reasonably like few billions of short reads within per execution rate. This environment permits to make dimension to sequence the orientation of the genome bases. As for example, microarrays techniques have been reinstated by sequence based techniques which can determine and verify the unusual translation without previous idea of any specific gene and can present clue in relation to different interlace and sequence change in determined genes [7],[8]. The strength to sequence complete genome of different similar creature has permitted substantial relative and innovative consideration are being computed which were inconceivable just a two year back. The multi scale application of NGS makes easy our analysis over different animals towards the exact answer of health and diseases influence.

The diversity of NGS characteristics makes it vital in new inventions of genome irrespective of any part of the research lab. Some of the remarkable outcomes and implementations of various labs are as 454 , GA, MiSe, HiSeq, SOLiD, Ion Torrent, RS system and Heliscope have general features to process the data set in parallel mode which improve the vast size of data generation in a single execution[9],[10].Some implementation produce short reads as near about 75 bp by SOLiD [11], 100 to 150 bp by Illunina [12], 200 bp by Ion Torrent [12] and 400 to 500 bp by 454 [12] and long read by Pacific Bioscience but it suffer multiple errors [13]. Besides, each of the implementation has some limitations. As 454, Ion Torrent and Pacific Bioscience suffers for Indel (Insert and Delete), on the other hand SOLiD and Illumina suffer for substitutions problem. The general process of each implementation is that they generate two distinct type of data set as short read sequences and quality score value for each and every base in the given or taken read. The computed score is imposed to measure the performance of sequence quality, trim reads and finally discard low quality bases. Some NGS environment can design paired-end reads and each this

reads contain a distinction distance. The distinction distance or separation distance is called clone length. Basically the pair-end-reads are applied on combining process of contigs in the last stage of genome assembly. It is verified that Next Generation Sequence reads are generally accessible at internet through the Sequence Read Archive (SRA) [14] and assembled reads are accessible through Assembly Archive (AA) [15],[16],[17].

II. Definition of De Bruijn Graph

The directed graph which maintain edge from all nodes of the graph as a fashion of A=(a_1,a_2,a_3,……..a_n) to B = (b_1,b_2,b_3,…….b_n) with B is being a left-sided part of A, or b1=a2, b2=b3 and so on is called de Bruijn graph. To formal define de Bruijn graph depends on two set as nodes and their relation as dimensions. In the light of Bioinformatics the nodes set can be set of nucleotides of DNA and the dimensions is similar with k-mer length.

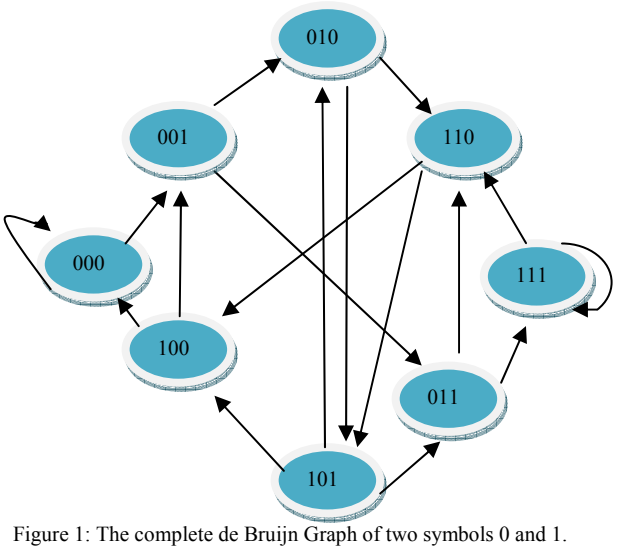

Figure 1: The complete de Bruijn Graph of two symbols 0 and 1.

Example

Two symbol 0 and 1.
Dimensions =3, so the possible words= 2^3=8 as 000,001,010,011,100,101,110,111. Two nodes are only related if last two symbols of A are same with first two symbols with B (A→B). On the same it can be define that the de Bruijn graph maintain a relation only if last k-1 symbols of node A are similar as the first k-1 symbols of node B.

A. Important for Bioinformatics

In the current research of Computational Biology data set of genomes and DNA nucleotides are increasing day by day with exponential speed. To utilize and mange these vast volume of data, de Bruijn graph is one of the important mathematical tool that helps to remove assembly errors and to get optimal solution. The crucial hurdle for new Bioinformatics challenge is to drain allocated Random Access Memory (RAM). It has been experimentally proved that the analysis which requires large memory is not good for proper scaling of genome data set. In this regards, de Bruijn graph has become important for transcriptome genome data assembly.

B. Necessity of De Bruijn Graphs in Genome Assembly

The results of various experiments have showed that genome data set of or nucleotides base have seldom duplicate regions in genome sequences and this duplicate genomes data are very problematic for genome assembly. Besides, Overlaps Layout Consensus (OLC) processes are also not enough to address the problem accurately and efficiently. To overcome the problems of OLC processes, De Bruijn graph helps to manage the repetitive genome data set more efficiently with feasible result. The figure 3 below shows a basic scenario about the repetitive genome segments which initiate genome assembly problems for proper pair wise alignments as well genome sequencing. From genome Read1 and Genome Read 2, we see that two read portions of these reads are approximately uniform which indicates the overlaps between two genome segments. The Blue and Purple sections of the figure are dissimilar part of the sequences. To assemble these types of segments, OLC based assembler might suffer with wrong connections between contigs and may overlook some parts of the complete genome segments.

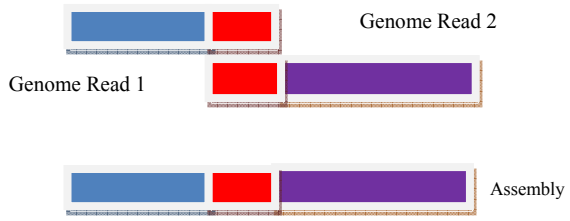

Figure 2: The overlaps between genome segments

Overlap Layout Consensus based package manage such problems with costly computing process to relate contigs and several genomes frame. Besides, an OLC assembler demands continuous predictions either any small change is exist between any two consecutive overlapped genome segments for repetitions and error.

On the other hand de Bruijn graph solve the repeats data set without making any predictions and measure the genome regions in uniform consideration. It incorporates the concepts of Next Generation Sequencing (NGS) and capable to reduce errors for shorter sequence errors input reads. It is widely know that NGS process can decode sequences promptly and cheaply of thousands small genome segments. It depends on K-mer graph which solve the overlap data set in short reads perfectly. Beside, k-mer graph does not maintain any all beyond all overlap identifications, does not keep data set in memory for overlapped data and integrate duplicate data set. It can manage large volume of data set by making distributed memory facility. De Bruijn graph oriented solution become feasible due to its feature that divides large volume data set into small length where all segments will be equal length. Basically, de Bruijn graph perform the activity in two fundamental steps. At first, small input

reads are divided into more sub-divisions and de Bruijn graph is formed from those sub-divided portions. At final stage, genome sequences are built from formatted de Bruijn graph.

C. Transcription Assembler Iterative De Bruijn graph Approach (T-IDBA)

Transcription Assembler Iterative De Bruijn graph Approach (T-IDBA) **[**18] is an updated algorithm of IDBA which re-establish known isoforms beyond association of any genome segments. It resolves the difficulties of IDBA by applying the clue of pair wise alignments of end side. It also fixed up the drawbacks of long repeats in separate genes and the difficulties of subdivision in the identical gene. The complete input genome graph is clustered into unique regions where each region refers to gene groups or a specific gene. The T-IDBA finds out majority of the isoforms by using the heuristic concepts under dept first search algorithm. Besides by applying accumulated de Bruijn graph [19] it overcome the Indel or gap problem of Next Generation Sequencing technology by considering unique end input reads up to the length of fifty base pairs.

T-IDBA algorithm

1) Implement Iterative De Bruijn graph Approach on given input data set as $I=I_{min}$ *to* I_{mod} for a graph G. Here, I= Number of repeated genomes segments. I_{mod}=Feasible repeats set.
2) Pair wise genome input reads (pair end reads) alignments and determine the relations R(x,y) between any two nodes of the graph G.
3) For any relation R (x,y), if there maintain a specific edge e which relate x and y, then it consider the edge is long read for remaining part of the relations of given data set.
4) Repeat step 1 for $I=I_{mod}$ *to* I_{max} to get a reconstructed de Bruijn graph.

D. Oases

Oases [20], the software package which assembles the input reads heuristically without the presence of reference genomes throughout the genome sequences along with substitute isoforms. This program store RNA sequences reads into the lookup table or hash table, reduce the gaps dynamically, strong settlement of different mesh of the nodes and accurate merging of numerous alignment assemblies. The fundamental parameters for various assemblers are the length of k-mer. The activity of the assembly algorithm depends mainly on this parameter. The complete processes of Oases are briefly described below. The steps are:

1) Contig Assembly
2) Contig Correction
3) Boundary Construction
4) Boundary Filtering
5) Locus Construction
6) Transitive reduction of the loci
7) Extracting transcript assemblies
8) Merging assemblies with Oases-M

(1)Contig Assembly

Oases perform this step primarily by taking the input which is the output generated by Velvet assembler [21] or program. This data set then used to generate genome boundary or scaffolds from genome sequences.

(2) Contig correction

To have accurate contig is very good for assumptions of the input reads of the genome sequences. This software package performs accurate contig selection by applying automated and fixed refining on repeated data set. The automated filtering is an updated error correction process of Velvet algorithm done by TourBus. Besides, Oases maintain a regional edge removal for every node of the genome sequence graph. An open edge is deleted if its broadcasting area is less than ten percent of the total broadcast edges from the similar node. It is analogous to the algorithm Yassour **[**22**].** In last stage the contigs which has less broadcast or coverage than fixed broadcast will deleted from the genome segments or assembly.

(3)Scaffold construction

Scaffold is the integration of the set of distance information between any two consecutive contigs. It is very essential for de Bruijn graph because this graph is subdivided into various components and the distance between two components or contigs can help to built scaffold.

(4)Scaffold filtering

Two types of filtering are implemented to the built scaffold in previous stage. One is fixed filtering and other is mechanical filtering. Fixed or static filtering is used to the very little coverage of the genome sequences and mechanical filtering is used to the contigs which has regional genome coverage.

(5)Locus construction

Locus is the clustered set of the contigs which is called loci in short. This idea is considered as conceptually when there is no overlaps or gaps in given input reads of the genome sequences. But practically it is not possible.

(6)Transitive reduction of the loci

Oases delete extra duplicate distances that generate incorrect result from the given input genome sequences.

E. Iterative De Bruijn graph Approach transcriptomes (IDBA-tran)[23]

IDBA-tran is a statistical and dynamic software package that iteratively deletes all misleading nodes and edges with regional or local threshold value. The deletion

process permits to divide the complete graph into various small discontinuous parts where each part contains a group of genes or a single gene with some correct nodes or edges of lower valued known isoforms. It can integrate lower valued known transcripts with higher valued known transcripts. This software package performs significantly better with other tools in respect of sensitivity and specificity. Analogous to the IDBA, this program also maintain an accumulated de Bruijn graph to maintain complete information regarding high and low expressed transcriptome.

It is very difficult processes to combine various exons of an individual gene to encode to collective isoforms [24]. But the existing tools under go from losing or generate incorrect information on isoforms. Existing tools solve this problem either on alignment base assembly or based on reference genomes or annotation information. Besides, existing tool suffer following two problems as Exons shared by several isoforms. Numerous expression steps of isoforms of the same genomes. In this respects, IDBA-tran computes the probability that any k-mer holds the error either in cluster of k-mer or in complete probability distribution of the genomes data segments. According to the various probability distribution and size of the contig, this software package measures regional threshold value to select if the k-mer or contig contain any errors. When it finds an incorrect nodes or edges, then the misleading edge is deleted from the graph constantly.

F. The De Bruijn Graph in Euler [25]

At first Eulerian path scan the input reads to establish a graph. During the scanning it has been find out incorrect nodes or edges by counting low density k-mers. The scanning process depends on mainly input reads duplications and irregularities of the sequencing errors. The Euler scanning or filter is maintained with multiple k-mers and their measured densities of input read data set. This scanning process or filtering approach increases the accuracy by correcting low density k-mers. It deletes the wrong k-mers and numbered the nodes in the graph. But sometime the scanning or filtering process may accidently remove the corrected k-mers.

This program performs spectral alignment [26] to identify the proper reads of input data set. It constantly verifies the K-mers information between any two reads and entire reads in given input data set. The continuous verification enable the deletion of K-mers which has lowest score value of threshold. It is a regular process to select a threshold value from reads after measuring the distribution of K-mers density exists in the given input reads. There are two frequency distributions in Euler process or it is bi-modal. One mode indicates the situation where some K-mers appear once or twice in response of genome sequence error or lowest frequency limit. Other mode indicates the duplication of the K-mers exist in the given input reads. Euler de Bruijn graph based package find a threshold limit in between the maximum threshold value and feasible numbers of all K-mers as correct or incorrect. For every incorrect K-mers reads Ruler de Bruijn graph approach implements a greedy inspection for base read replacement which decrease the incorrect K-mers numbers. Euler performs replacement to make appropriate of incorrect reads but it never perform insertions and deletions.

G. The De Bruijn Graph in Velvet [27]

Velvet [26], an integrated environment that promptly operate de Bruijn graph to reduce incorrect reads and to decrease the repetitive input reads in genome segments. Velvet is a stable and simple de Bruijn graph based assembler. Velvet comprises the features of error reduction algorithms and repeat reduction algorithms. At the beginning of the analysis, error reduction process combined the genome sequences that reside together and in next step the repeat reduction process isolates the path which share regional overlaps. It manages the complete graph without any loss of existing data set. This system deletes a single node from k-mer graph length to ease the analysis process. Analogous to the Euler path assembly, Velvet also shorten the K-mer graph by eliminating impetus recursively. The impetus elimination or spurs elimination process shorten the graph length on given input reads. In contrast of Euler path, Velvet maintains an extra parameter to handle the low coverage of the input reads in genome sequences.

Velvet apply breadth first search for bubbles in the entire graph and tries to cover the total graph starting from the points with different out going edges. There are inner bubbles within the graph and to find out all the bubble is a troublesome process. So the searching is restricted in a way that the desired roads are visited in a process that moving forward from one node to all roads per iteration up to the road distance crossed the limit of the threshold. After being detection of bubble, Velvet eliminates the entire road of the graph by representing a limited reads and worked beyond the graph to repeat the alignment reads from the eliminated road to the existing road. It also eliminates the graph roads that shows very low coverage reads than a certain threshold

In conclusion, Velvet provides a complete analysis of de Bruijn graph assembly. Velvet involves error deflected reads scanning or filtering process. To overcome the graph complexity problems it implements list of heuristics approaches. Heuristics process manipulates set of activity as regional graph structure, input reads area selection, sequence verification and pair ends conditions. The Velvet programming package select de novo assembly from small input reads under paired ends from the Solexa package.

H. ABySS (Assembly By Short Sequences)[28]

Assembly By Short Sequences (ABySS) [28] is a distributed environment that is used to manipulate memory area for large volume of data set using the de Bruijn Graph simplification. This algorithm circulates the K-mer graph and complete graph estimation towards a fixed node whose occupy huge memory space. This program capable to assemble about four billions of Solexa reads from human data set. It computes graph nodes in parallel computations define in Euler and Velvet software package. Besides, ABySS eliminates spurs

repetitively whose lengths are shorter than the threshold value.

Basically it is very difficult to solve numerous data sets by a single processor or any specific devices. Instead using a single processor, distributed or parallel processor or algorithms will helps to solve the problem efficiently. ABySS divides the graph at the grate of the particular graph node. The allocation of a graph node to a processor is manipulated conversion of K-mer to an integer value. This concept is works as unbiased where K-mer and its opposite counterpart scale to the similar integer value. The concept is usable if scaled adjacent K-mers to the similar processor.

Analogous to the Velvet, it also perform restricted bubble search throughout the graph. Consequently, this program distributes easy and non overlaps paths into set of contigs and implements mate threading. It applies a solid illustration of the K-mer graph. Every graph node illustrates a K-mer and its opposite counterpart. Here every graph node maintains eight bits additional information to indicate the presence or non presence of four possible DNA nucleotides at the end of the genome sequences. Though the extra information, the graph edges are constant for the graph. This process computes the graph in distributed way starting from any random nodes per processor. It scale from any starting node and find out its neighbors gracefully as the last K-1 bases and one extra base extension denotes by a connection or edge. Besides, ABySS does not maintain scaffolds or genome boundary. It is capable of making assembly on Solexa small reads data set as well as paired end input reads.

I. AllPaths

AllPaths [29] is de Bruijn Graph DNA sequencing or genome sequencing assembly program that can handle large volume of genome data set. It works for both simulated and real data set. It occupies a preprocessor or algorithms that use a read corrected input data set. AllPaths consider those K-mers which have high frequency reads. The scanner or filter implements the k-mers for low, high and average value of K. consequently, AllPaths holds another pre-processor which creates unipaths. This process start with the analysis of accurate read overlaps seeds by K-mers. It allocates an integer value to the K-mers in a way that several K-mers are visible in a series in input reads and overlaps gains serial indicators. It introduces a database of indicator intervals and an input reads links. The database initialization decreases the Random Memory Access demands for the graph formations. Then it builds a de Bruijn graph form the database reads. The primary operation of AllPaths is spurs erosion. It subdivides the complete graph and reduces the data set repeats in genome sequences that are regionally non-repetitive. It applies heuristics algorithm to select subdivisions on the entire data set. It seed selection in various nodes of the graph defines long, reasonably covered and widely divided contigs. AllPaths tries to reduce the gaps between any paired ends reads by examining the K-mer graph in response to a condition where only one graph paths satisfy that conditions. This package or program divides the total data set into various sections and executes in parallel. After making partitions, AllPaths connects the regional graph or local graphs where the nodes are overlapped. For global graph, AllPaths deletes spurs, discontinuous nodes and paths not spanned by paired ends.

J. Short Oligonucleotide Analysis Package de novo (SOAP de novo)

The SOAPdenovo assembler [30] produce better character oriented human genome sequences from hundreds millions of Solexa reads of length seventy five base pairs and less [31],[32]. It works by scanning and correcting input data reads by implementing predefined threshold value of K-mer densities. It can attire the input reads and subdivide the paths that shows identical frazzle line components. Similar to the Velvet algorithm, SOAP de novo delete bubbles with peak reads coverage and find out a traversal road or path. Unlike Velvet and Euler Path algorithm, SOAP de novo provides more space optimization facilities.

Consequently, SOAP de novo establish contigs from reads under the environment of de Bruijn graph formation technique. Then it deletes the de Bruijn graph to generate scaffolds or genome boundary. All paired reads are scaled to the contig consensus sequences and sequentially construct a contig graph where the edges of the constructed graph denotes the inter contig of same label restriction. To overcome the complexity SOAP de novo deletes edges that create transitivity relation. One important feature of SOAP is that it can handle numerous data set that blend Overlapped Layout Consensus (OLC) and de Bruijn graph methods from Next Generation Sequence assembly.

K. SOAP de novo2 [33]

SOAP de novo2 is an integrated environment which has updated Short Oligonucleotide Analysis Package de novo (SOAP de novo) software package with more accurately and efficiently. This is a memory efficient software package with the support of dynamic algorithms. The dataset of de novo genome sequencing are increasing day by day with the collaboration of Next Generation Sequencing (NGS). SOAP de novo perform the assembly widely on genome data set and results some unique findings. But it is possible to perform the same research with more accuracy, continuity and data coverage.

In this regards, SOAP de novo2 have progressed some factors that can improve the genome assembly process. These factors are: decrease over memory use, overcome the problems regarding input reads repeats, improve genome data inclusion, boost up scaffolds formation, advance in gap reductions process and finally perform optimal result in genome assembly.

TABLE II.
COMPARISONS OF SOAP DE NOVO2 AND SOAP DE NOVO

SOAP de novo	SOAP de novo2
The genome segments inclusions is about 81.16% [34].	Genome segments coverage is 93.91% [35],[36]
It absorbs too much computational time and memory for long input reads.	It is relatively faster than the SOAP de novo and takes less memory for solving same data set.
SOAP de novo implements de Bruijn Graph for genome assembly. Each genome read is stored in genome database individually.	SOAP de novo2 use Sparse de Bruijn Graph [34] where input reads are grouped to store in database instead of storing individually.
It only utilizes the large K-mers.	It utilizes the benefits of both large K-mers and small K-mers by using Multiple K-mers approach [18]. This approach first deletes sequencing errors by using small K-mers to construct graph and then re-construct the graph by applying large K-mers in a repetitive fashions.
To establish Scaffolds SOAP de novo manage heterozygous contigs irregularly.	SOAP de novo identify heterozygous contig pair by implementing contig depth and regional contig connections. So it can maintain only those contigs which has maximum depth in heterozygous pairs.
It suffers for chimeric scaffolds incorrectly constructed.	SOAP de novo solve this problem by making strong fixed up the shorter insert size. Instead of shorter insert size it use larger insert size library.
Pseudo connections between contigs without enough Paired Ends (PE) information.	SOAP de novo designed a topology based approach that to overcome the pseudo connection between contigs.
Use GapCloser to handle the problem of repetitive segments longer than read length.	SOAP de novo2 enhanced the process by implementing a concept as each iterative loop the prior input of GapCloser is accepted only the reads which can by align in ongoing loop.

III. CONCLUSION

We have gone through some important de novo genome assembly software package and algorithms. There are more than forty genome assemblies performing the genome data alignments from last few years. Despite the development of the genome assembly, still short reads generated by Next Generation Sequence technology is problematic. The particular behavior of genome assemblers is cumbersome to determine under the circumstances of base structures, genome size, data redundancy and occurrence of polymorphism. REAPR (Recognition of Errors in Assemblies using Paired Reads) is novel de novo genome assembly package that is reference independent algorithm is very useful to handle vast size genomes and NGS data sets. It can generate score for every genome base and mechanically reduce the incorrect assemblies. Due to the strict time schedule we were unable to complete this algorithm but it has very important appeal for new NGS technology based genome sequence data set. Beside CGAL (Computing Genome Assembly Likelihoods) is an important algorithm that can analyze data set without the support of any universal truth data set. CGAL compute the likelihood score of assembly that permits a system to measure data without specific value. In future research it will be very effective and economical. ALE (Assembly Likelihood Evaluation) is also enables the assembly process using comprehensive statistical procedures. This package is also not mentioned here due to the same reasons, lack of time. REAPR, CGAL and ALE must helps to design novel techniques for handling unpredictable genome assembly data set.

De novo Short Read Assembly is also applicable on optimization of Seabuckthorn Transcriptome. So genome assembly is also a very essential part of new discovery in plants history and life cycle. The process is very vital for maintaining a track for wonder plants and it will be possible to save the plants from their destructions. Similarly, several new platforms such as medical science, agricultural science, food science and pharmacy research are also depends on genome sequencing or genome data set assembly. Diseases identifications and Drugs designs also solely depends on de novo genome assembly. Future platforms will fixed up the errors rates reduction procedures comparing to current standards. Designers of assembly software or algorithms will march on by setting challenges to handle vast volume data set in medicine, agricultures, food chain, fisheries and diseases verifications. We do believe that genome assembly especially de novo genome assembly will be remarkable tool to mitigate the problem of current era. The ultimate target should pinpoints meaningful information from each and every sequence with minimum input and maximizes the outputs.

REFERENCES

[1] T.P.Niedringhaus,D.Milanova,M.B. Kerby, M.P.Snyder MP,A,E,Barron , Landscape of next-generation sequencing technologies. Anal Chem 83: 4327– 4341, 2011.

[2] K.V.Voelkerding, S.A.Dames, J.D.Durtschi, Next-generation sequencing: from basic research to diagnostics. Clin Chem 55: 641–658, 2009.

[3] M.Helmy,N.Sugiyama, M.Tomita, Y.Ishihama, Mass spectrum sequential subtraction speeds up searching large peptide MS/MS spectra datasets against large nucleotide databases for proteogenomics. Genes Cells 17: 633–644, 2012.

[4] W.Zhang, J.Chen, Y.Yang,Y. Tang,J. Shang, A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. PLoS ONE 6: e17915. doi:10.1371/journal.- pone.0017915, 2011.

[5] M. Pop,A. Phillippy, A.L.Delcher,S.L. Salzberg SL, Comparative genome assembly. Brief Bioinform 5: 237–248, 2004.

[6] J.AMartin, Z.Wang,Next-generation transcriptome assembly. Nat Rev Genet 12: 671–682, 2011.

[7] B.Wold, & R.M.Myers, Sequence census methods for functional genomics. *NatureMethods* 5, 19–21 2008.

[8] Z.Wang, M.Gerstein,& M.Snyder, RNA-Seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet.* 10, 57–63 2009.

[9] X.Zhou, L.Ren, Q.Meng, Y.Li, Y.Yu, The next-generation sequencing technology and application. Protein Cell 1: 520–536, 2010.
[10] M.A.Quail, M.Smith, P.Coupland, T.D.Otto, S.R.Harris,A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13: 341, 2012.
[11] J.M.Mille, R.M.Malenfant, S.S.Moore, D.W.Coltman, Short reads, circular genome: skimming solid sequence to construct the bighorn sheep mitochondrial genome. J Hered 103: 140–146, 2012.
[12] N.J.Loman, R.V.Misr, T.J.Dallman, C.Constantinidou, S.E.Gharbia,Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol 30: 434–439, 2012.
[13] N.Nagarajan, M. Pop, Sequence assembly demystified. Nat Rev Genet14: 157–167, 2013.
[14] Sequence Read Archive. Available: http://www.ncbi.nlm.nih.gov/sra. Accessed 4 February 2013.
[15] Assembly Archive. Available: http://www.ncbi.nlm.nih.gov/Traces/ assembly/.Accessed 4 February 2013.
[16] Z.Derong, A Heuristic Optimization Algorithm in System Structure Optimization ,Journal of Computer, vol 9, pp743-747,2014.
[17] T.Jian Tao, Optimal Parameters and Performance Assessment for Detecting a Fluctuating Distributed-target, Journal of Computer ,pp 653-360, Vol, 8, 2013.
[18] Y.Peng, C.M.Henry, S.M.Yiu, Y.L.Francis, T-IDBA: a de novo Iterative de Bruijn Graph Assembler for Transcriptome. In: RECOMB. Vancouver, BC, Canada, 2011.
[19] Y.Peng, C.M.Henry, S.M.Yiu, Y.L.Francis: IDBA- A Practical Iterative de Bruijn Graph De Novo Assembler. In: *RECOMB: Lisbon*; 2010.
[20] M.H.Schulz, R.Daniel, V.Martin, B.Ewan, Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics, 28, 1086–1092, 2012.
[21] D.R.Zerbino, E. Birney, Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.*, 18, 821–829, 2008.
[22] G.G.Manfred ,J.H.Brian ,Y.Moran , L.Joshua, T.A.Dawn A. Ido A.,Xian , F.Lin , R. Raktima, Z.Qiandong , C. Zehua , M.Evan ,H. Nir , G.Andreas, R.Nicholas , P.Federica, B. Bruce , N.Chad , T. Kerstin, F.Nir,R.Aviv,Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.,* 29, 644–652, 2011.
[23] P Yu, C. M. Henry, Y.S. Ming , J. L. Ming, Z. X. Guang, Y. L. Francis, IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels, Bioinformatics, Vol. 29 ISMB/ECCB, pages i326–i334, 2013.
[24] C.Trapnell, B.A.Williams,G. Pertea, A.Mortazavi, G.Kwan, M.J.. Baren, S.L.Salzberg, B.J.Wold, L.Pachter, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol., 28, 511–515, 2010.
[25] P. A. Pevzner, H.Tang, M.S.Waterman, An Eulerian path approach to DNA fragments assembly. Proc Natl Acad Sci USA; 98:9748–53, 2001.
[26] D.R. Zerbino, E. Birney. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res; 18:821–9, 2008.
[27] D.R. Zerbino, G.K. McEwen, E.H. Margulies, E. Birney. Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler. PLoS One; 4:e8407, 2009.
[28] J.T. Simpson, K. Wong, K. Jackman, J.E. Schein, S. J. Jones, I. Birol. A parallel assembler for short read sequence data. Res. ABySS: 2009.
[29] J. Butler ,I. MacCallum ,M. Kleber ,I. A. Shlyakhter , M.K. Belmonte, E.S. Lander, C.Nusbaum, D.B. Jaffe . ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res; 18:810– 20, 2008.
[30] R. Li , H. Zhu,J. Ruan,W. Qian, X. Fang ,Z. Shi ,Y. Li ,S. Li ,G. Shan, K. Kristiansen, H.Yang, J.Wang. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 2009.
[31] R. Li, W. Fan,G. Tian ,H. Zhu ,L. He ,J. Cai ,Q. Huang ,Q. Cai ,B. Li ,Y. Bai ,Z. Zhang , Y. Zhang ,W. Wang,J. Li , F.Wei ,H. Li ,M. Jian ,R. Nielsen,D. Li ,W. Gu ,Z. Yang ,Z. Xuan , O.A.Ryder ,F.C. Leung ,Y. Zhou ,J. Cao, X. Sun ,Y. Fu ,X. Fang ,X. Guo ,B. Wang ,R. Hou , F.Shen, B. Mu ,P. Ni ,R. Lin ,W. Qian ,G. Wang ,C. Yu ,W. Nie ,J. Wang ,Z. Wu ,H. Liang , J. Min ,Q. Wu ,S. Cheng ,J. Ruan, M. Wang ,Z. Shi ,M. Wen ,B. Liu ,X. Ren ,H. Zheng ,D. Dong , K. Cook , G. Shan ,H. Zhang ,C. Kosiol ,X. Xie ,Z. Lu ,Y. Li ,C. C. Steiner, T. T. Lam, S. Lin Q. Zhang, G. Li ,T. Tian ,T. Gong ,H. Liu ,D. Zhang ,L. Fang ,C. Ye ,J. Zhang ,W. Hu ,A. Xu , Y.Ren ,G. Zhang ,M.W. Bruford,Q. Li , L. Ma, Y. Guo, N. An , Y. Hu ,Y. Zheng ,Y. Shi ,Z. Li ,Q. Liu ,Y. Chen , J. Zhao, N. Qu ,S. Zhao ,F. Tian , X. Wang ,H. Wang ,L. Xu ,X. Liu ,T. Vinar , Y. Wang , T. W. Lam ,S. M. Yiu. The sequence and de novo assembly of the giant panda genome. Nature; 463:311–7, 2009.
[32] R. Li, Y. Li, H. Zheng, R. Luo, H. Zhu, Q. Li, W. Qian, Y. Ren, G. Tian, J. Li, G. Zhou, X. Zhu, H. Wu, J. Qin, X. Jin, D. Li, H. Cao, X. Hu, H. Blanche, H. Cann, X. Zhang, S. Li, L. Bolund, K. Kristiansen, H. Yang, J. Wang. Building the sequence map of the human pan-genome. Nat Biotechnol; 28:57–63, 2009.
[33] L. Ruibang, L. Binghang, X. Yinlong, L. Zhenyu, H. Weihua, Y. Jianying, H. Guangzhu, C. Yanxiang, P. Qi, L. Yunjie, T. Jingbo, W. Gengxiong, Z. Hao, S. Yujian, L. Yong, Y. Chang, W. Bo, L. Yao, H. Changlei, W. C. David, Y. Siu-Ming, P. Shaoliang, X. Zhu, L. Guangming, L. Xiangke, L. Yingrui, Y. Huanming, W. Jian, L. Tak-Wah, W. Jun. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler, GigaScience, 1:18, 2012.
[34] C. Ye, Z. S. Ma, C. H. Cannon, M. Pop, and D.W. Yu: Exploiting sparseness in de novo genome assembly. BMC Bioinformatics, 13 Suppl 6:S1, 2012.
[35] Y. Peng, HC. Leung, SM. Yiu, FY. Chin: IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics, 28:1420–1428, 2012.
[36] Q.Wangren, X.Xuan, W.Lidong, and G. Dianxuan, A Novel Pseudo Amino Acid Composition for Predicting Subcellular Location of Proteins, Journal of Computer, pp 764-371, Vol: 8, 2013.

Dr. Mohammad Ibrahim Khan received the B.S. degree in Electrical and Electronic Engineering from Bangladesh University of Engineering and Technology (BUET), Bangladesh in 1999. He received M.S. degree in Computer Science and Engineering from the same University in 2002. He received his Ph.D. degree in Computer Science and Engineering from Jahangirnagar University in 2010. Since 1999, he has been serving as a faculty member in the Department of Computer Science and Engineering at Chittagong University of Engineering & Technology (CUET), Chittagong, Bangladesh. Currently his research project on Bioinformatics Local sequence alignments algorithms analysis with his research group. His research interest includes Digital Image Processing, Graph Theory, Cryptography, Digital Watermarking, Multimedia Systems, and Digital Signal Processing.

Md.Sarwar Kamal received the B.Sc (Hons) in computer science and engineering from University of Chittagong Bangladesh in 2009. Since 2009, he has been serving as a faculty member in the Department of Computer Science and Engineering at BGC Trust University Bangladesh Chittagong. Now he is MS (Engineering) student in the Department of Computer Science and Engineering at Chittagong University of Engineering & Technology (CUET), Chittagong, Bangladesh. Currently his research project on Bioinformatics Local sequence alignments algorithms analysis under the guideline of Dr. Mohammad Ibrahim Khan .His research interest includes Bioinformatics and Data mining.