A Graph Based Contig Analysis for Big DNA Sequence

Munmun Biswas

Abstract— DNA assembly is one of the most principle and troublesome issue in bioinformatics. The read are produced by next generation sequencing faces difficulties due to large and complex repeat of human genome. Although many of the issues create for the orientation and ordination of contig generation from short reads. One of the key advances that have prompted a change in contig lengths has been mate pairs, which encourage the assembly of repeating regions. Mate pairs have been used in next generation assembler to correct the assembly graph or to interface the contig into scaffolds. Such strategies have permitted the identification of longer contigs than would be conceivable with single reads; however, they can even now neglect to determine complex repeats. In this manner, enhanced strategies for incorporating mate pairs will strongly affect contig length later on.
In our proposed method, we use pair de-brujin graph a speculation of the de Bruijn graph which incorporate mate pair information to generate contig in graph structure. This graph can able to resolve error and repeat resolution and improve contig size in assembly.
Index Terms— Bioinformatics, Contig, De-brujin graph, fragment assembly, next generation sequencing, mate pair, Pair de-brujin graph.

1 Introduction

DNA sequence of data proceed profoundly affect on biology.
DNA comprises of four alphabets sets (A, C, G and T), on account of RNA there is U (Uralic) rather than T) and Protein sequence comprises of twenty alphabets in order. Finding the sequence of base-pairs in a given DNA molecule isn't a simple task. The primary technique to identify the exact request of base-pairs in a DNA molecule was devised by Fredrick Sanger in 1977 [4] and this is as yet the most exact technique for DNA sequencing. Two critical issues around Sanger innovation are its moderate run time for extensive genomes and its cost.Next generation Sequencing (NGS) advancements [3] were proposed from 1996 with the point of reducing the cost and expanding the speed of the DNA sequencing process. Discovering overlaps between the reads, merging the right connections and expanding the reads to accomplish larger sequences are the basic task of "DNA Assembly" algorithms. The first generation of assemblers took after the overlap-layout-consensus paradigm, where overlaps were heuristically used to combine reads into contigs [1] [5]. Afterward, the presentation of de Bruijn graphs prompted significant changes in assembly [2] [12][13][14]. Instead of the overlap-layout-consensus approach, these assemblers initially fabricated a graph where the principal genome is spelled by a series of walk through the graph and non-branching walks compare to substrings (contigs) of the genome. At the point when the length of a repeat is longer than double the read length, it becomes hard to accurately match its upstream and downstream regions. In order to ease issue, sequencing technologies were stretched out to deliver mate pairs [9] the genomic distance between pairs of reads (called the insert size) is all around assessed In this article, we propose the paired de Bruijn graph, a speculation of the de Bruijn graph that consolidates the mate pair information into graph structure, instead of a post-processing step [6].
Similarly as moving from the heuristic overlap-layout consensus paradigm to the de Bruijn graph brought about better assemblies, we trust that moving from heuristic mate pair algorithms to paired de Bruijn graphs could result in a more powerful utilization of mate pair information including error correction and repeat resolution.

• *Munmun Biswas is currently Lecturer of Computer Science and Engineering in BGC Trust University, Bangladesh, PH-01797235245. E-mail: munmun-biswas29@yahoo.com*

2 Proposed System Architecture

A novel approach ContigG is proposed to generate contig using mate pair information. Our proposed structure is worked based on DNA sequence (Fig 1). DNA Nucleotides of human, animal, plant, insects or micro-organs are the data set of input sequences

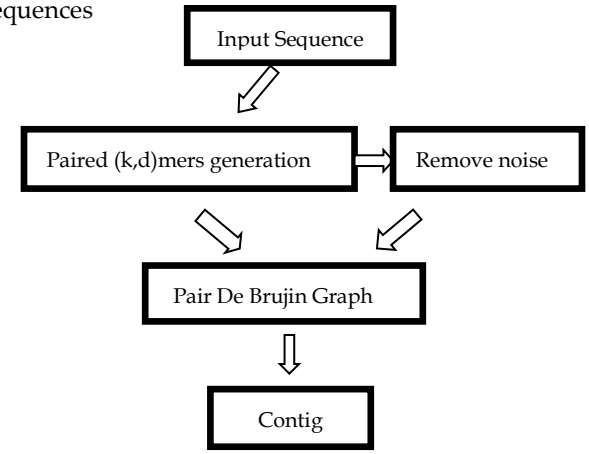

Fig1: the entire outline of the research work where the input are the DNA sequence which include nucleotide base pair of A, G, Cand T.

Here, the data set of different species can be conceivable to take as input as character set of DNA nucleotide data stream. The input data moves from file to memory and store reads and pair information. Extract k-mers while handling each read from the input files by having pre-defined k value. To create

International Journal of Scientific & Engineering Research, Volume 9, Issue 8, August-2018
ISSN 2229-5518

contig using mate pair de brujin graph we require paired k-mers.A paired k-mers is a pair of k mer at a fixed distance d Paired k-mers can alternatively represent as (k,d)-mers.(k,d)-mers is generated by bilabel representation.

A bilabel (a|b) is a couple of strings, a and b, of equivalent length. Characterize left (a|b) = a and right (a|b) = b. A k-mer bilabel demonstrates both a and b have length k. Characterize prefix (a|b) = (a1 . . . ak-1|b1 . . . bk-1) and suffix (a|b) = (a2 . . . ak|b2 . . . bk).

The idea behind this noise detection technique is that the input reads are arbitrarily dispersed through the genome with generally even scope, along these lines all (k, d)-mers ought to be seen at least some minimum number of times, and entries that are seen less frequently than the threshold value can be thought to be noise. Discover (k-1) length overlaps between all (k, d)-mers and connection (k, d)-mers that can be the prefix or postfix of each other. Overlaps are heuristically used to combine read into contig.The pair de brujin graph is used to construct a graph where the number of repeated (k, d) - mers rapidly drops as d expands, and thus the contigs of the paired de Bruijn graph based on these (k, d)-mers could be longer. We watched that contig lengths enhanced drastically as the insert size expanded. We in this manner trust that properly utilizing mate pairs has a strong potential to increment contig lengths.

2.1 Paired (k, d) mers generation

Here, our proposed process able to manage input file that are coming with pair information. The pair information indicates which two reads are associated as pairs. We assume that all read have a similar l. In addition, the mate pair is made utilizing a pair of string of length l drawn from genome at position I and j .Commonly, the relative distance between reads is alluded in term of insert size, the quantity of nucleotide from the start nucleotide of a to the end nucleotide of b: j-i+1(fig2a). In any case, for the reasons for our development, it is more advantageous to express it as far as d = j - I, the distinction in their leftmost coordinates (fig2b). The main algorithm is used to utilize paired (K, d)-mer sets. Paired (k, d)-mers Hence reads can be expelled from the memory after the (K, d)-mer creation process.

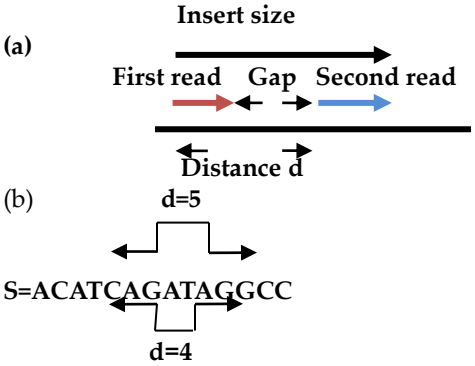

Fig2: (a) A mate pair is a pair of read with a separation of d between their begin positions.

(b) Two mate pair with d=4 and d=5.

2.2 Remove noisy (k, d) mers

Numerous assemblers utilize error detection/correction procedures to discover and resolve noise in input information, and after that run the assembly algorithm on the corrected data. It is demonstrated that utilizing error detection/correction procedures enhances the assembly outcomes, anyway there are a few issues utilizing these techniques:

Error recognition/correction is time requesting and their run time increments radically when working with substantial sources of info e.g. human genome, despite the fact that they simply should be run once.

A minimum threshold can likewise be set by the user that characterizes the minimum occurrence number of (k,d)-mers in the info set. All entries that have less occurrences will be removed.

2.3 From De Brujin Graph to Paired de Brujin Graph

In De Bruijn graph based approach, our algorithms have a parameter k that directs the measure of the substrings into which the reads are spilled up. Hence, however our input is an arrangement of mate pairs of reads of any length, we quickly spell them up into smaller pieces. Formally, each mate pair of reads is replaced by its constituent l - k (sub-) mate pairs, where the reads of each (sub-) mate pair currently have length k + 1.

Characterize a k-mer as a string of length k. We accept that the parameter k is fixed. Given a string S=s1 . . . sn, let Sk(i) be the k-mer si . . . si+k-1 (where the index is taken modulo n). The set of all k-mers Sk(i) (for 1 ≤i ≤n) is known as the k-spectrum of S. For a k-mer a=a1 . . . ak, we characterize two (k - 1) - mers, prefix (a) =a1 . . . ak-1 (remove last character) and suffix (a) =a2 . . . ak (remove first character). We say that k-mer an adjusts at position i if a = Sk(i).

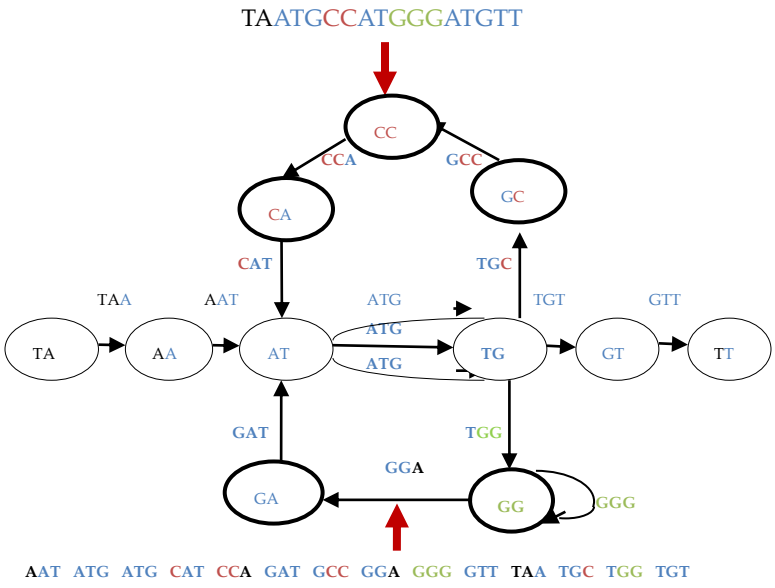

Fig 3: Representation from read to de Brujin Graph

Paired de Brujin graph assembly method creates overlap graph with each node containing a (k,d)-mer and each edge characterizing overlap between two (k,d)-mer nodes, accordingly characterizing arrangements of length (k+1). Utilizing

International Journal of Scientific & Engineering Research, Volume 9, Issue 8, August-2018
ISSN 2229-5518

the entire (k,d)-mer set, this makes a substantial and memory concentrated paired de Bruijn graph which has numerous nodes and edges that keep the algorithms from effectively rearranging the graph on the off chance that it occurs by utilizing an inappropriate k value.
A graph demonstrating mate pairs in the exceptional case that all sets are precisely the same distance d apart.

- Characterize an initial graph G_0 on m = 2|C| vertices. For each bilabel (a|b) ϵ C (representing a (k + 1, d)-mer), present two new vertices u, v and an shape an edge u→v. Label the edge by (a|b); label u by prefix (a|b) (sub-string from the first element to the one preceding the last element); and label v by suffix(a|b) (sub-string from the second element to the last element).
- Glue vertices of G0 together when they have a similar label. The graph G so acquired is known as the paired de Bruijn graph of C.

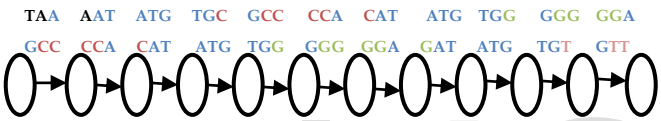

Fig4 (a) Glue nodes with identical levels

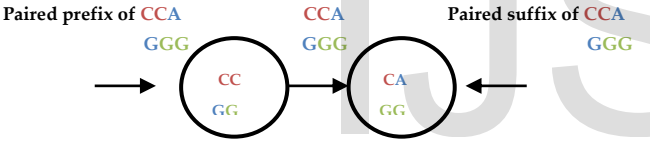

Fig4 (b) Glue nodes by paried prefix and suffix

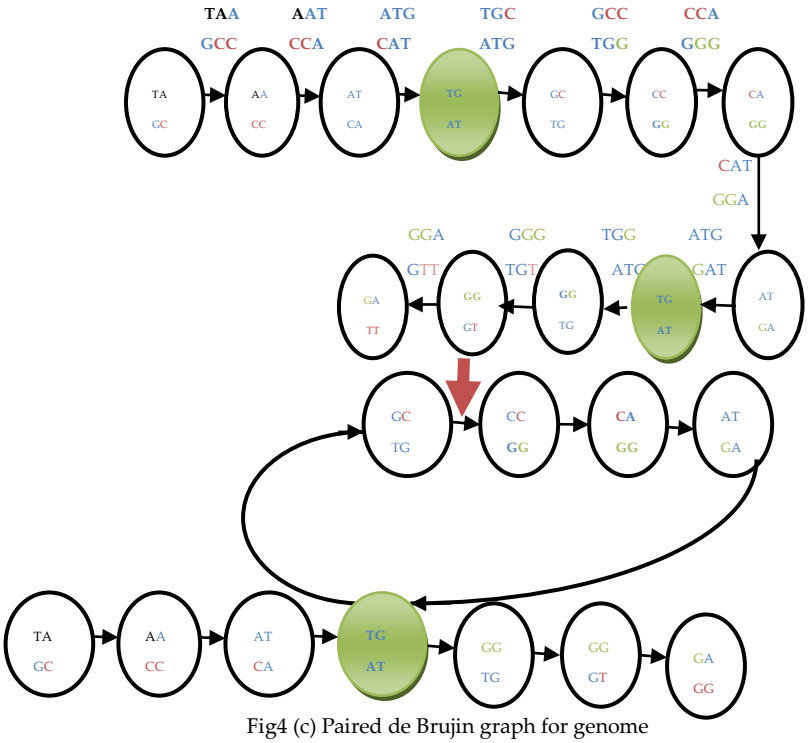

Fig4 (c) Paired de Brujin graph for genome

Likewise with the standard de Bruijn graph, in this development, each vertex of G acquires the label regular to all the vertices of G_0 that were glued together to shape it and this label is unique to that vertex. Any walk through the graph on edge sequence e1, . . . , er explains a (r + k)- mer bilabel (L|R) where L is shaped from the left labels, L=walkword(left(e1), . . . , left(er)), and R is framed from the right labels, R=walkword(right(e1), . . . , right(er)).The (k, d) - spectrum of a string S is {(Sk(i)|Sk(i+d)) : i=1, . . . , n}. At the point when C is the (k + 1, d) - mer range of S, there is a covering cycle whose left labels spell S in G. The cycle comprises of successive edges (Sk(i)|Sk(i+d)) →(Sk(i+1)|Sk(i+d+1)) for i=1, . . . , n:
Similarly likewise with the de Bruijn graph, this is a key property that makes the paired graph valuable for spelling contigs

3 Graph Complexities

The value of a graph representation of a genome can change broadly. By and large, the quantity of vertices can fill in as an unpleasant pointer of how valuable the graph is—as the quantity of vertices develops (and the number of edges remains the same), the graph is probably going to end up less entangled, and the contigs are probably going to turn into longer. Figure 5a demonstrates that in the de Bruijn graph, the quantity of repeated k-mers in E. coli drops as k increments, demonstrating that the de Bruijn diagram has more vertices and likely turns out to be less entangled.

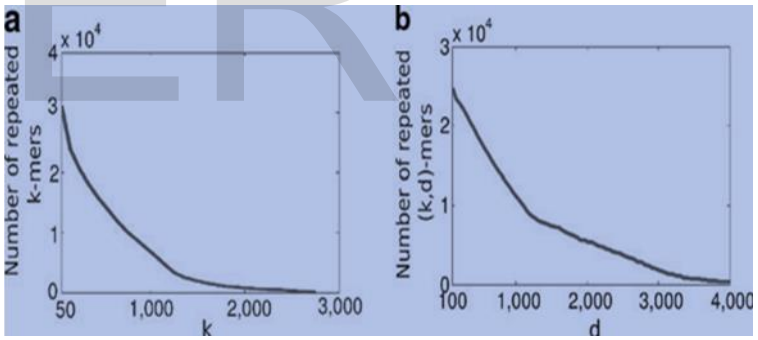

Fig5: The impact of expanding k and d. (a) The quantity of repeated k-mers in the E. coli genome, for different estimations of k. (b) The number of repeated (k, d)- mers, for different estimations of d with k = 50

On the other hand, think about pairs of k-mers, i.e., (k, d) - mers. Figure 5b demonstrates that, after fixing k = 50, the number of repeated (k, d) - mers drops as d increments. This isn't surprising because of the repeat structure of genomes—the greater the d, the less normal it is to have pairs of repeats dispersed a separation of d separated.
Figure 5a, b represents alternatives for enhancing contig lengths: expanding the read length as opposed to expanding the insert size.
In this way, we could assemble a graph whose vertices speak to (k, d)- mers rather than k-mers, at that point the length of the contigs is probably going to increment as the insert size develops. This is the essential inspiration for the paired de brujin graph, and, as we are showed in Section 2, the contig

International Journal of Scientific & Engineering Research, Volume 9, Issue 8, August-2018
ISSN 2229-5518

lengths in the paired de Bruijn graph do in reality increment with d.

4 Experimental Results

We actualized a model assembly algorithm to test the effectiveness of the paired de Bruijn graph approach under the ideal conditions of perfect coverage and error free reads. We tested with E. coli (4.6 Mbp) and Human chromosome 22 (35 Mbp after removal of ambiguous bases). The reads were produced with perfect coverage, which means for each situation in the genome we created a single (k, d, Δ) - mer adjusting to it. The insert size was picked consistently at arbitrary from the predetermined range. We inform as contigs the (left) words spelled by every maximal walk of the graph whose inside vertices have only one out-neighbor. We approved that any created contigs mapped flawlessly back to the original genome—this was the situation for all the contigs.

Developing the de Bruijn graph and discovering all its non-branching ways requires some time O (n log n), where n is the quantity of k-mers. The development of the pair de Bruijn graph has an extra cost of looking through all neighbors inside a separation 2Δ of every node. Subsequently, the running time of the algorithms is O (n log n + n min {2Δ, n}), where n is the quantity of (k, d, Δ) – mers.

Our inspiration for the paired de Bruijn graph approach was that the quantity of repeated (k, d)- mers rapidly drops as d increases (Fig. 5b), and consequently the contigs of the paired de Bruijn graph in view of these (k, d)- mers could be longer. To test this hypothesis, we created a set of mate pairs with differing insert sizes and plotted the length of the got contigs (Fig. 6a). To separate the impact of the insert size, the scope of the data was perfect (the (k, d) - range), the insert sizes were perfect (D = 0), and the read length was fixed to 50. We watched that contig lengths enhanced significantly as the insert size expanded.

To investigate the part that read length plays in respect to the insert size, we created sets of mate pairs with changing read lengths however with a fixed insert size (1000 nt). To confine the impact of the read length, we had perfect coverage and no variety in the insert size. For E. coli, we found that, for an insert size of 1000 nt, once the read length became over a small threshold of 10– 20 nt, the contig lengths about came to the hypothetical optimum that could be accomplished by essentially creating reads of length equivalent to the insert size (Fig. 6b). For Human, we expected to expand the read length to 300 nt keeping in mind the end goal to achieve the optimum with 1000 nt insert size (Fig. 6b).However, for a more drawn out insert size (5000 nt), a read length of 50 approached (Fig. 6a) to accomplishing the optimum (which, with 5000 nt peruses, was a single contig). Subsequently, by legitimately utilizing mate pairs with extensive enough insert size, one can fundamentally lessen the constraints caused by short read length. We measure the impact of expanding variability in the insert size (D) on the assembly. We fix the insert size to be 1000 nt and create 50-long reads with idealize coverage, while shifting D (Fig. 6c). We found that the assembly breaks down with expanding D, particularly for the Human genome. At the point when D is vast, the chance of two vertices of the de Bruijn graph being associated increments, and, henceforth, the quantity of vertices (bilabels) that don't adjust yet by the by get glued together increments. In this circumstance, the read length is as yet essential in deciding the complexity of the (non-paired) de Bruijn graph.

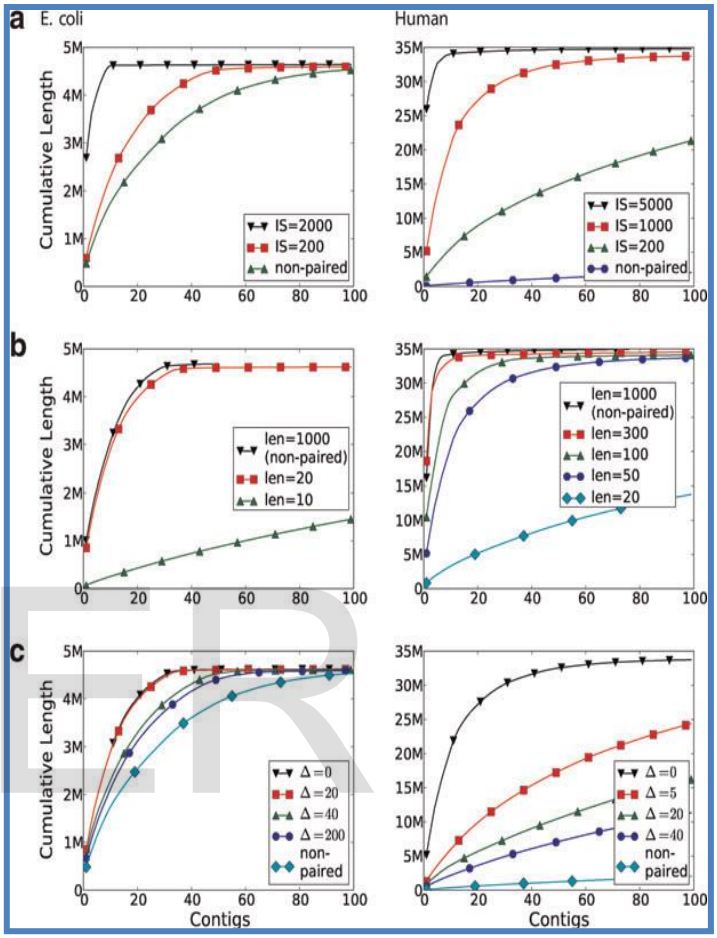

Fig5.Contig lengths. Combined contig lengths (for standard and paired de Bruijn diagrams) on simulated information with perfect scope. Contigs are arranged all together from largest to smallest. Point (x,y) implies the largest x contigs have combined length y. (a) To analyze the impact of the insert size (IS) on the assembly, we kept the read length fixed at 50, yet shifted the insert size. We additionally produced non-paired reads of length 50. For E. coli, the curve for insert size 6000 isn't demonstrated in light of the fact that there was just a single contig, speaking to the entire genome. (b) To break down the impact of read length on contig lengths, we fixed the iinsert size to 1000 yet differed the read length.We likewise produced non-paired reads of length 1000, giving an upper bound on how great the assembly can be for this situation. (c) To examine the impact of varieties in the insert size (D), we fixed the mean insert size (1000) and read length (50). We additionally demonstrate the benchmark contig lengths in a nonpaired dataset, with read length 50 and great scope.

International Journal of Scientific & Engineering Research, Volume 9, Issue 8, August-2018
ISSN 2229-5518

5 Conclusions and Future Work

5.1Conclusion

In this article, we presented the paired de Bruijn graph and inspired its utilization in genome assembly. Rather than incorporating mate pairs into a post-graph development step, we have utilized them to build the graph itself. Any methods that could be performed on the regular de Bruijn graph (e.g. error correction) can be performed in a similar way on the paired de Bruijn graph. For example, notwithstanding when there are repeats that the paired de Bruijn graph does not resolve, mate pair transformation can even now be applied to the graph to help resolve the rest of the repeats.

By figuring an alternative to mate pair changes, the paired de Bruijn graph approach gives a potential technique for assembly with short read mate pairs, like the one created by Complete Genomics [3] and Helicos [11]. By not requiring extraordinary ways between paired reads in the de Bruijn graph, the paired approach could in any case resolve repeats notwithstanding the short read length (Fig 4c).

Besides, the algorithms we depict can be reached out to the strobes produced by Pacific Biosciences, which broaden the thought of the mate pair to an arrangement of numerous (more than two) reads separated by a few distances**.**

5.2 Future Work

The utilization of the right labels on edges of the paired de Bruijn graph is one thought that we need to investigate in future. As of now; we spell out every contig utilizing just thle left label. The places of the right labels are as it were known around, yet this is often adequate to shape a right hand word displaced roughly d from the left hand word. In addition, in the wake of experiencing an edge (a|b) in a walk, we should experience some edge (b|c) around d edges away (except if it is past the end of the walk). This similarity necessity may limit the decision of substantial ways while experiencing branching vertices, along these lines resolving longer repeats and enhancing contig lengths.

6 References

[1] Batzoglou, S., Jaffe, D.B., Stanley, K., et al. 2002. ARACHNE: a whole-genome shotgun assembler. Genome Res. 12,177–189.

[2] Idury, R.M., and Waterman, M.S. 1995. A new algorithm for DNA sequence assembly. J.Comput. Biol. 2, 291–306.

[3] Drmanac, R., Sparks, A.B., Callow, M.J., et al. 2010. Human genome sequencing using unchained base reads on selfassembling DNA nanoarrays. Science 327, 78.

[4] Chaisson, M.J., Brinza, D., and Pevzner, P.A. 2009. De novo fragment assembly with short mate-paired reads: does the read length matter? Genome Res. 19, 336–346.

[5] Myers, E.W. 1995. Toward simplifying and accurately formulating fragment assembly. J. Comput. Biol. 2, 275–290.

[6] Medvedev, P., Pham, S., Chaisson, M., et al. 2011. Paired de Bruijn graphs: a novel approach for incorporating matepair information into genome assemblers. Proc. RECOMB 2011.

[7] Zerbino, D.R., and Birney, E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs.
Genome Res. 18, 821–829.

[8] Medvedev, P., and Brudno, M. 2008. Ab initio whole genome shotgun assembly with mated short reads. Proc.RECOMB 2008, 50–64.

[9]Weber, J.L., and Myers, E.W. 1997. Human whole-genome shotgun sequencing. Genome Res. 7, 401–409.

[10]Pevzner, P.A., Tang, H., and Tesler, G. 2004. De novo repeat classification and fragment assembly. Genome Res. 14, 1786–1796.

[11] Harris, T.D., Buzby, P.R., Babcock, H., et al. 2008. Single-molecule DNA sequencing of a viral genome. Science 320,106.

[12] Pevzner, P.A. 1989. L-Tuple DNA sequencing: computer analysis. J. Biomol. Struct. Dyn. 7, 63–73.

[13]Pevzner, P.A., and Tang, H. 2001. Fragment assembly with double-barreled data. Bioinformatics 17, S223–S225.

[14]Pevzner, P.A., Tang, H., and Waterman, M.S. 2001. Eulerian path approaches to DNA fragment assembly. Proc. Natl.Acad. Sci. USA 98, 9748–9753.