An Optimal Genome Reassembling technique by Artificial Bees system for small genome sequences

Susobhan Baidya
Department of Information Technology
Heritage Institute of Technology
Kolkata, India
Email: susobhan.baidya@heritageit.edu

Rajat Kumar De
Machine Intelligence Unit
Indian Statistical Institute
Kolkata, India
Email: rajat@isical.ac.in

Abstract—[1] **Fragment assembling problem (FAP) is an NP-complete problem. The present article presents an Artificial Bees Colony (ABC) learning system to solve Genome sequence reassembling techniques. Reference Genome sequence which is taken 99% analogous to a Genome from same organism, because of the fact the sequences from the similar organism usually have approximately 99.9% resemblance. We have used the sequences from NCBI database[2]. Then we have cloned each sequence and shear the clone to a numeral short reads. Here, we have used a different perception in Genome reassembling by Synthetic Bees System where nectar amount is relative to the accuracy of assembled reads with some reference genome sequences inside the similar creature. For local heuristics information, we have introduced local alignment of short reads instead local overlapping among the reads. The outcome depict that our methodology is more accurate than an existing Bee Colony Algorithm. Genome reassembling methodology require a huge concurrency and vast storage because of size of Genome sequences of mammalian group is $\sim 10^9$bp, and ABC is inherently concurrent in nature. We have run LSBCO in 64 bit O.S in HP proliant server with 16GB RAM, 2-quad core processor. We have computed our methodology for the Genome length up to 127429bp. We have simulated hierarchical sequencing, and finally stitched the each segments to get back the actual Genome sequence.**

Key words : BCO, NP-complete, Genome Sequence , short reads, Hierarchical BAC by BAC sequencing.

I. INTRODUCTION

Genome Sequencing is the method of forming the exact arranges of nucleotides A, C, T, G within a DNA molecule. The most powerful areas of Genome Sequencing are Biomedicine [6] and Forensics, where it finds the nucleotide sequence of an unknown DNA. Through the Genome Sequencing, we can identify diseases, similarity among the sequences, and the stream of information from ancestor to descendant.

Genome Sequencing is a pretty tough and computationally exhaustive work. A small creature bacteria, contains ~2Mbp in sequence string and the whole Genome string is wired [6]. We have simulated Hierarchical sequencing Fig. 1. In reality which is under Hierarchical sequencing, is mapped to different locations of reference Genome and then each mapped segments are slashed further by bio-chemical reaction [7] and generates a set of short fragments. That fragments are known as reads (100bp-500bp). Then the reads are stitched by some intelligent comutation to shape the actual Genome . Genome reassembling require sophisticated parallel processing and large storage [3]. Complexity of Genome reassembling process include [3], [8] the next:

Next Generation sequencing (NGS) method reassembles the reads and generates contigs. Finally, contigs are required to be stitched yet again physically by sequence finishing mechanism [3]. Overlapping among the reads might be wrong. For instance reads AACCT has an overlap with the read CTCCTGG that might not be the real overlap in the acual sequence. Throughout the cloning of Genome sequence, there may be a little removal and alteration of bases. At the time of creation of reads, there might be trouncing of bases at the boundaries of reads, and a few reads might be missing due to that cause. Repeat regions may guide to an incorrect sequence [11]. In the human genome has roughly 50% repeats Fig 2 [4]. Now consider an easy example, lacking the intricacy of cloning and overlaps Here in the Fig. 2, we have shown a string graph for a particular sequence, where CCGT and AG are repetition. Suppose ACAAC is the starting read. Then the next apparent option will be CCGT. After attainment CCGT, the accurate option should be TA. Now as per the graph, there is another option to hop to TACCT. Therefore the managing repeats are hard to meet up the correct sequence.

From the year 1997 EST sequencing, ABI370, Pyrosequencing and Phrap were the victorious reassembling tools for comparatively small Genome [12]. Human genome was first time sequenced in 2004 by IHGSC [12]. ABI-SOLid SOAPdenovo, AbySS and Velvet [1], [3], [7], [11] are a few of the renowned next generation sequencing (NGS) technologies. NGS creates short reads (100bp-500bp) for sequence $\sim 10^9$bp. The entire technologies falls into primary set. [3].

1) Graph based Greedy approach
2) Overlap-layout-consensus (OLC)
3) De Bruijn graph based approach or Eulerian path based approach
4) Align-layout-consensus
5) BAC-by-BAC (hierarchical) sequencing

In 2005, scientist Denis Karaboga has developed the ABC algorithm, which was based on the smart foraging activities

[1]978-1-5090-1047-9/16/$31.00 2016 IEEE
[2]www.ncbi.nlm.nih.gov

[3]https://www.cbcb.umd.edu/research/assembly_primer

of honey bee swarm. The flow of information amongst bees leads to the arrangement of a tuned cooperative knowledge, which is done by **Waggle** dance of bees in the hive. The entire process is based on path exploration and path exploitation for nectar or pollen food sources. The exploration is done by scout bees and employed bees. The exploitation procedure is done by onlookers bees. The scout bees explore the new food source randomly, The employed bees gather the exact information of food sources in the hive and finally according to that information the onlookers bees collect the nectar or pollen from the food sources. In this paper, we have projected a novel Genome reassembling technique founded on Artificial Bees Colony Algorithm. We have given our algorithm name Local Score guided Bees Colony Optimization (LSBCO). Here, we have not processed any k-mer graph like De Bruijn graph, So it is become simpler for treatment repeats. We have assumed each read as a place, and a read can be repeated within the sequence. We represent local guidance for displacement of bees by local score of a read. We have compared our method with present Bee Colony Optimization algorithm [13], which is shortened by OVBCO (Overlapped base pair guided Bees Colony Optimization). In the existing methodology, a bee initiates from a starting place and flies to a place, which is not yet visited. The movement is biased with the information of nectar attentiveness and overlapping base pairs among the reads. Here in LSBCO algorithm, we have used local score of a read with nectar attentiveness for displacement of a bee from one place to a different place. At last bee reaches to destination and modify memory (nectar concentration) derived from score of the solution. The effectiveness of the methodology along with its superior performance over existing algorithms is demonstrated. Our methodology reaches 99.9%-100% perfection with the actual sequence. We restrict our data set into $\sim 10^5$bp (the sequences like *Homo sapiens* breast cancer) because of lack of suitable computation.

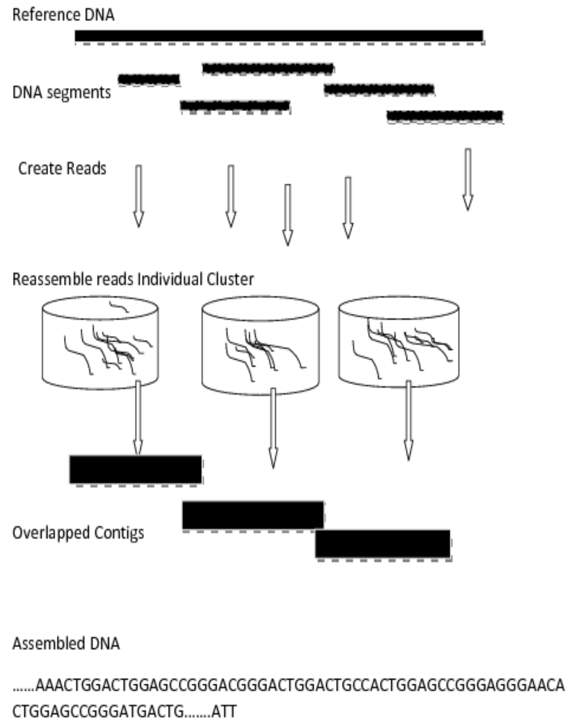

Fig. 1. Hierarchical BAC by BAC sequencing and Whole Genome Sequencing

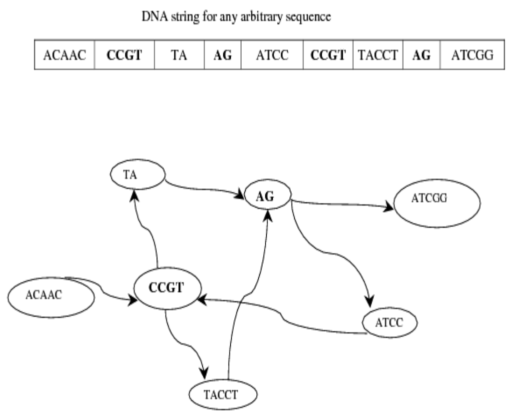

Fig. 2. Equivalent string graph for the above sequence

II. Methodology

Before describing the methodology, we will make some assumptions as mentioned below.

- A food source is an arc between two places.
- Now consider number of employed bees be m_e and the number of scout bees be m_{sc}. Then total number of bees in the hive will be $m_h = m_e + m_{sc} + 1$, assuming only one onlooker bee. It is also assumed that 75% of bees are employed and the remaining 25% of bees are scout.
- Initially, the nectar amount of a particular arc is not known to employed bees.
- Each employed bees are allowed to move to only a certain number of places for a particular tours.
- Employed bees jump to different food sources according to amount of nectar and number of unvisited places.
- After collection of nectar information, employed bees come back to the hive and update the effectiveness of the tour in terms of nectar amount.
- The onlooker bee collects the nectar from the arcs on currently available optimal path.
- The onlooker bee collects an equal percentage of nectar from different food sources on the best possible optimal path, and equal percentage of nectar will be decayed from the remaining food sources by natural loss. Thus an uniform decay factor ρ is considered for exploration process.
- After the tour of the onlooker bee, scout bees start moving randomly and explore new paths.
- Number of places is equal to the number of reads.
- Entire bees system will find out an optimal path, where the sequences corresponding to the food sources will

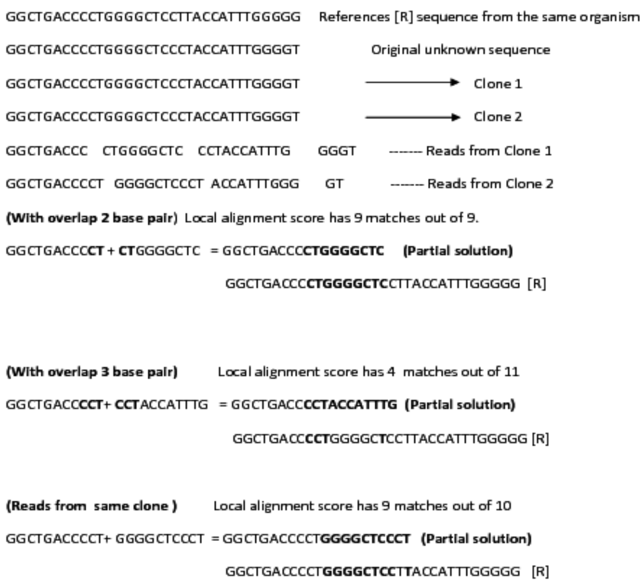

Fig. 3. Clones and Reads

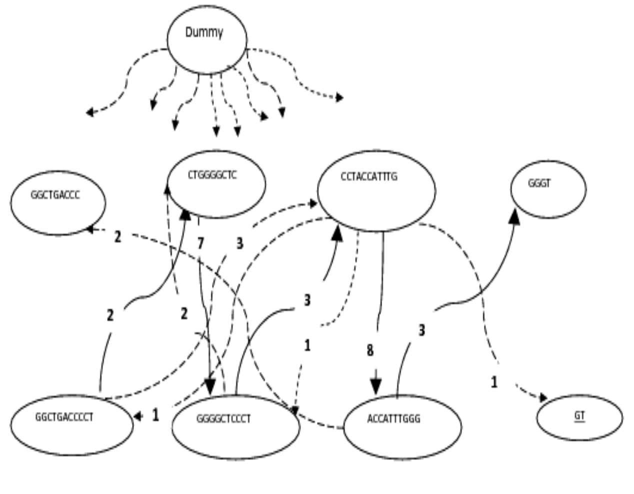

Fig. 4. Corresponding graph with probable movement by the matrix W

determine the assembled Genome sequence.

Let us assume a set of m reads such that $m = m_h - 1$, produced from a given sequence. All the scout bees or employed bees will move, in parallel, through the places. The percentage of different types of bees can be varied in different algorithm implementation, but the total number of bees in the system should be nearly equal to the number of places [9], [10]. Here we have considered one more bee compared to the number of places. Suppose i^{th} and j^{th} are the pair of reads and the common bases towards tail of i^{th} read and head of j^{th} read is represented by w_{ij}, where i^{th} and j^{th} reads are from separate clones. If reads are belong to same clone then set to -1 . Our aim is to assemble these reads and reproduce the Unknown Sequence. Every reads are considered as a place. Let a nectar matrix of the order $m \times m$, which indicates the nectar density on the link from i^{th} place to j^{th} place, represented by $= [\psi_{ij}]_{m \times m}$. Here ψ_{ij}, which will provide. Primarily, these ψ_{ij}'s presume arbitrary integer in [0, 10]. Now, we do not know the starting read, we assume a duplicate starting place.

Now consider all the employed bees or scout bees are to be unite at the duplicate starting place. Bees are permissible to hope to the other unvisited places , which depends on some probability. The chance that q^{th} employed bee will hope from i^{th} place to j^{th} place is specified by

$$p_{ij}^{q} = \frac{\psi_{ij}^{\alpha} \times s_{ij}^{\beta}}{\sum\limits_{j' \epsilon N_i} \psi_{ij'}^{\alpha} \times s_{ij'}^{\beta}} \tag{1}$$

The above equation is also used in case of scout bees, where the values of α and β are both considered as zero. Now, N_i is the total number of unvisited neighboring places from i^{th} place. The local alignment score is presented by the expression s_{ij}. This alignment is done by Needleman-Wunsch algorithm. Here we align the overlapped portion at the tail end of the i^{th} place with the head end of the j^{th} place to the fraction of the reference sequence among $l - \triangle$ and $l + n_{ij} + \triangle$. $l - \triangle$ and $l + n_{ij} + \triangle$.

Now, l is intermediate length of a solution for a particular bee reaches up to i^{th} place, and $n_{ij} = |r_j| - w_{ij}$, being the total bases as the bee hopes from i^{th} place to j^{th} place, and r_j is the total bases in j^{th} read. The $\triangle$ is an integer within the range in [1, 5]. Now consider cases of α and β with some boundary cases. Where if $\alpha = 0$, then choice of j^{th} place is relative to s_{ij}^{β} It indicates that the probability of j^{th} place, is being selected if the reads matching to i^{th} and j^{th} places have the maximum local alignment score over the considered region. In this case, Artificial Bees Colony (ABC) converges to a traditional greedy algorithm. Now consider the case, where $\beta = 0$. Now simply nectar concentration is in focus. So the probability to hope at j^{th} place is maximum if the nectar affinity on the link i^{th} and j^{th} places is the maximum, where j belongs to unvisited neighboring places of i^{th} place. The most steady values of α, β have been found are 1 and 2 respectively for the problem under consideration. Next we have to find an optimal solution. Bee will starts from a duplicate starting place and finally reaches to a place, where the solution length l is less or equals to threshold θ. Here θ is the of the reference sequence. Finally, q^{th} bee completes the tour by reassembling the reads. After this the reassembled sequence is aligned with the reference sequence by Needleman-Wunsch algorithm. This alignment score is then measured by q^{th} bee in terms of SC_q.

Then the bees (employed and scout) will come back to the duplicate starting place through the similar trail by which they have reached their destinations, and finally, they update the nectar information ψ_{ij}. It is done by modifying ψ_{ij} as q^{th} bee moves from i^{th} place to j^{th} place, on its way back, modifying ψ_{ij} is given by

$$\psi_{ij} = \psi_{ij} + \triangle\psi_{ij}^{q} \quad \forall i, j \ \ \epsilon P^{q}, \tag{2}$$

the quantity $\triangle\psi_{ij}^{q}$ is the nectar information collected by q^{th} bee across the link i^{th} and j^{th} , which is a arcs of the tour along the path P^q . Now $\triangle\psi_{ij}^{q}$ is defined as $\triangle\psi_{ij}^{q} = SC_q$, if q^{th} bee has made a tour along P^q. Otherwise it is zero.

Now an onlooker bee moves through the arcs on a currently available optimal path. On its travel, it consumes a part (ρ) of nectar available on the arcs. There will be natural loss of same part (ρ) of the currently available nectar amount on the arcs of remaining paths. Thus, there will be an consistent decay of nectar on all the links among a pair of i^{th} and j^{th} places, and the modified nectar concentration is given by

$$\psi_{ij} = (1-\rho)\psi_{ij}, \forall i, j. \tag{3}$$

Here ρ is known as nectar decay constant and $0 < \rho < 1$. Decay constant enables to overlook the non optimal path. In the present methodology we have considered a little diverse expression for p_{ij}^{q} (equation (1)) in comparison to previous literature [13]. p_{ij}^{q} is defined in previous literature as

$$p_{ij}^{q} = \frac{\psi_{ij}^{\alpha} \times w_{ij}^{\beta}}{{}_{j'\epsilon N_i}\psi_{ij'}^{\alpha} \times w_{ij'}^{\beta}} \tag{4}$$

In previous methodology, the authors used the term w_{ij} as heuristics for reassembling of a split sequence [13]. w_{ij} is the overlapping among i^{th} and j^{th} reads. In the next segment, we have shown the process of reconstruct the fragmented sequence. Here we have used the concept of concatenation of reads, if consecutive reads fit in to the similar clone. In PET or Pair Ended Read [14] model we can distinguish the reads from same clone and different model. In the result segment, we have shown that correctness of the assembled sequence has been improved by our methodology over the existing Bees Colony approach . The correctness of final assembled sequence is evaluated by alignment of the original sequence and assembled Genome sequence using Needleman-Wunsch algorithm.

a) Explanation with an example :

Lest us take an easy instance described by Figs. 3 and 4 for our methodology . We try to describe here by examples, that maximum overlap is not forever a better choice for a move from one place to another. The source of the reads (source clone information) is also fundamental information for rearrangement the reads. Here we have assumed a unknown reference sequence and an known sequence. For simplicity we have shown that unknown sequence has been cloned two times. After that clones have been split into eight reads. We have used the fact during the simulation of clone process that almost 99% of the sequence is the similar inside the same organism. Now we have eight places equivalent to the eight reads lacking of knowledge of the initial place and ending place. As per the Fig. 3, the initial place perhaps GGCTGACCC or GGCTGACCCCT, and final place possibly GGGT or GT, which is unknown to us. For that reason , a duplicate initial place has been used. The links has been established between duplicate place and all the places with some logical overlaps (here in the example those connections are not shown for simplicity). A bee will initiate from the duplicate starting place, then the next movement is based on the probability (in equation (1)). We consider $W = [w_{ij}]_{m\times m}$, such a matrix that the path connecting i^{th} and j^{th} places is represented by w_{ij} Fig. 4. The trail indicated by solid arrow is indicating the accurate solution. The weight w_{ij} among the places i and j is indicating the degree of overlapping. Conventionally, $w_{ij} \neq w_{ji}$. Since we have considered every clone independently and created reads alone. We have to verify the reads, if they belongs to the similar clone. For the matrix $W = [w_{ij}]_{m\times m}$ will follow the law.

- If i^{th} and j^{th} reads belong to dissimilar clones and has an overlap of k ($k \geq 0$) bases, then $w_{ij} = k$, where $i \neq j$.
- If the i^{th} and j^{th} reads belongs to similar clone then $w_{ij} = -1$, and $i \neq j$.
- If i=j then $w_{ij} = 0$

Assume we have two reads TGGACTCCAC and ACTAAAATTTG from dissimilar clones, now the greatest prefix cum suffix is two and subsequent to assembling thepartial solution will be TGGACTCCACTAAAATTTG. If they are from identical clone then the solution will be j^{th} read is concatinated to i^{th} read. The partial will be TGGACTCCACACTAAAATTTG. Here the Fig. 3 is mapped to the Fig. 4. For better understanding, we assume just the link like $w_{ij} = k$, where $k > 0$. Suppose after starting from a duplicate initial place, a bee jumps to GGCTGACCCCT. The partial solution is GGCTGACCCCT and the partial length (l) of the tour is 11. At the present neighboring places are CTGGGGCTC, GGGGCTCCCT, CCTACCATTTG, ACCATTTGGG and GT. Now consider the case where the next read is CTGGGGCTC then i will be GGCTGACCCCT and j will be CTGGGGCTC. Now they are from separate clones with an overlap 2. Following that the partial solution will become GGCTGACCCCTGGGGCTC and l becomes 18. The darken region is to be used for estimate s_{ij}, which is equivalent to $((L_\theta/(L_l)) \times 100$, where L_θ is score over the common portion which is mentioned by dark region. We have shown the calculation of the local alignment score in Fig. 3. L_l is the length of the considered portion which is 9. It means the similar to 9bp out of 9bp within the considered region. But if we seek for bigger overlap from existing place GGCTGACCCCT to next one CCTACCATTTG, in that case similarity reduced by 4bp out of 11bp. So it is clear that It might guide to a non optimal solution (that is factual for the present example) because thereafter rearrangement the reads GGCTGACCCCT and CCTACCATTTG, we will find partial solution GGCTGACCCCTACCATTTG. In this case the length of the partial solution is 19. Actually we have to assume a modest drift $\triangle$ in left corner and right corner over the considered region. This hoping for the unvisited places will be continued until $l \leq \theta$. From the Figs. 3 and 4, it is obvious that only the information of overlapping among would not guide to the best solution. So the case where we don't have any information weather the reads are belonging to the identical clone or from different clones. In that situation, there may be a incorrect partial solution towards the global solution. For example, suppose GGCTGACCC, CTGGGGCTC, CCTACCATTTG and GGGT are the places, and belong to the same clone, which are selected by a bee one after another. The bee will rearrange individuals reads by the information of overlaps simply, and the final stitched sequence becomes GGCTGACC**C**TGGGGCT**C**CTACCATTT**G**GGT. This sequence is not a real partial solution, because we have considered a overlaps among the reads GGCTGACCC, CTGGGGCTC, CCTACCATTTG and GGGT, but actually there is none overlap.

III. ALGORITHM LSBCO

Input: $W = [w_{ij}]_{m \times m}$ represents the degree of overlap among the reads from dissimilar clones;
Length of reference sequence θ;
the nectar matrix, $= [\psi_{ij}]_{m \times m}$
Output: A most favorable Genome sequence which has to be reassembled from the set of reads

Repeat until there is an additional development in successive two results .

- *Step 1*: Put random values for $\psi_{ij}s$, in [1, 5].
- *Step 2*: Initiate m_e employed bees at starting place (duplicate) and start their movements concurrently.
- *Step 3*: For each q^{th} employed bee do the steps
  - *Step 3.a*: Let the bee to hope from present place i to next probable unvisited place j by equation (1), and $w_{ij} > 0$ or $w_{ij} = -1$, $\forall j$.
    - *Step 3.b.i*: If q^{th} bee has found the final place for which partial solution length $l \geq \theta$ then calculate score SC_q by Needleman-Wunsch algorithm with the reference sequence seq and *Go to Step 4*.
    - *Step 3.b.ii*: Else *Go to Step 3.a*.
- *Step 4*: For all q^{th} employed bee on its return to the hive, update nectar information by equation (2) for all links on the trail of q^{th} employed bee.
- *Step 5*: Keep the most excellent sequence up to now estimated.
- *Step 6*: Decay nectar amount along all the arcs by equation (3).
- *Step 7*: Put m_{sc} scout bees at starting place (dummy place) and start their movements concurrently.
- *Step 8*: For all q^{th} scout bee do the steps
  - *Step 8.a*: Let the scout bee to move from present place i to next probable unvisited place j randomly, and $w_{ij} > 0$ or $w_{ij} = -1$, $\forall j$.
    - *Step 8.b.i*: If q^{th} bee has found the final place, where the partial solution length $l \geq \theta$ then calculate score SC_q by Needleman-Wunsch algorithm by reference sequence seq and *Go to Step 9*.
    - *Step 8.b.ii*: Else *Go to Step 8.a*.
- *Step 9*: For all q^{th} scout bee on its return to the hive, update nectar information by equation (2) for all links on the trail of q^{th} scout bee.
- *Step 10*: Decrease m_e by 1 for decreasing old food source and increase m_{sc} by 1, for searching for new food source.
- *Step 11*: Keep the most excellent sequence up to now anticipated and *Go to Step 2*.

IV. RESULTS

Here we have shown the efficiency of LSBCO, in sequencing a number of Genome sequences. We have used the sequences of dissimilar organisms, like *Taenia solium*, *Influenza*, *HIV*, *Human Breast cancer gene*. The sequences are collected from www.ncbi.nlm.nih.gov. First, we have changed the known sequence with the help of random removals and changes in sequence base pairs. Finally, the known sequence become comparable to the fresh one by 99.9%. Now, this new sequence is our given unknown sequence. After this, the unknown sequence has been split into short reads. The identified sequence was considered as the reference. We have shaped reads of length 100bp to 300bp by 3X coverage. After this we have removed a small number of reads arbitrarily, due to the reason that a few information might be missing during the creation of reads by some biochemical process. Since we were unable to compute high performance computation, we had to detain by the sequences-length of 10^5bp. We have used our Local Score guided Bee Colony algorithm (LSBCO) to abovementioned data sets. We have compared our outcomes with earlier methods using Bee colony optimization algorithms (OVBCO) [13]. The overlap base pair were used as local heuristics for the earlier algorithms. Here, the local score of the reads s_{ij} is considered for local heuristics. The output sequences are then aligned with the original sequence to validate our solutions. Finally, the alignment scores are rewarded to the solutions.

A. Implementation issues

Here we have used the reads with coverage up to 3X and simple reads with 1% to 3% information missing at the time of creation reads. Normalization of nectar value is necessary following perticular number of iterations, or else an overflow will occur . We have taken $s_{ij} \geq 1$ and $SC_q \geq 0$. We have restricted the value of SC_q in [1, 1000], s_{ij} in [1, 100] and $\psi_{ij}s$ in [1, 50]. As we have used a duplicate starting point, the size of each matrices Ψ and W is $(m+1) \times (m+1)$. We have considered the notion of Roulette wheel choice for equations (1) and (4). For the sequence span above 10Kbp. Here, the idea of Hierarchical sequencing is used by us, where the actual sequence is split and mapped to a exacting position. The locations can be mapped by STS probe or by other location marker using specialized bio-chemical process. The whole sequence is cut into number of segments of length about 10Kbp, and every segment has overlapping with neighboring segments. Again we have split each segments arbitrarily. The reassembled reads mapped to analogous individual segments. In conclusion, we have stitched each reassembled segments and produce an most favorable sequence

TABLE I
DATA SETS

Sequence Name	Short form	Length	No. of reads(Avg)
Taenia solium (CLONE PAT6) ACTIN GENE, COMPLETE CDS	Taenia1837	1837	38
INFL-Sequence 41 from PATENT WO2010036948	INFL6125	6125	123
HIV-1 ISOLATE 99ET8, ETHIOPIA GAG POLYPROTEIN	HIV8746	8746	175
Homo sapiens BRCA1, (RPL21) PSEUDOGENE	BRCA117143	117143	2420

B. Analysis

We have achieved 99.9%-100% precision (Table (II)) for the above mentioned data set. Even though the outcome of LSBCO and the offered OVBCO are approximately similar, OVBCO has resulted in only 75% precision for *Homo sapiens* BARC1 data. It is practical that the running time of LSBCO is expensive with respect to OVBCO. It is observed that LSBCO is more stable and accurate. Again, it is observed that for shorter sequence the results almost identical for both the methodology. In the other hand, it is also true for the large sequences, the methodology OVBCO is slightly unbalanced with respect to LSBCO to the extent of accurateness sequence. OVBCO has less time complexity because it considered, the overlap W for hopping one place to another. The cost of w_{ij} was estimated before starting of tours by bees. Thus the complexity of computing p^q_{ij} (equation (4)) was little. Therefore there is a probability to fail to spot the real adjacent read which has no overlap with the existing read , it may happens when reads are belong to same clone. Thus, from the optimality point of view, the quality of solution has been compromised by OVBCO for the better time complexity. On the other hand, LSBCO process all the unexplored places consequent to overlapping and non-overlapping reads for calculation of s_{ij}, it has additional time complexity with respect to OVBCO , which has been considered for each movement of a bee. For n number of unvisited reads and reads of average size $m$bp, the time complexities to choose the next place is $O(n)$ and $O(m \times n)$ for OVBCO and LSBCO in that order. The result (Table II) for our methodology is within the range 99.9% to 100%. The cases where we are getting 99.9% only for the reason of different length among reference Genome and unknown Genome. The LSBCO will reconstruct the sequence from the reads with target sequence length θ. θ is the length of identified reference frame length. OVBCO also follow this rule. In a few case we give this difference by 1-5 bp but in the case of equal length of unknown sequence and reference sequence , we are getting 100% accuracy. In all the cases, the alignment score between unknown sequence and reference sequence is 99%-99.9%.

TABLE II
RESULTS OF LSBCO AND OVBCO ALGORITHMS

SeqName	% of accuracy by LSBCO	% of accuracy by OVBCO	Execution time by LSBCO(Sec)	Execution time by OVBCO(Sec)
Taeania1837	99.9	99.9	8	6
Infl6125	99.9	99.9	263	223
HIV8746	99.9	98.5	1432	561
BRCA11714	100	75	5704	1508

V. CONCLUSION

Computational complexity is the main challenges of next generation sequencing (NGS). Needleman-Wunsch sequence alignment process has the space and time complexity $O(N^2)$. N is the number of base pairs presents in reference sequence. It is tough to grip the sequences-length above 10Kbp, because for each explore by a bee of our algorithm process Needleman-Wunsch algorithm which has the time complexity $O(N^2)$. In our algorithm, we have created m reads from an unknown sequence. For simplicity we put 3X coverage of that unknown sequence. In every round, a bee in its tour has rearranged the reads with the assist of a reference sequence, which is from the similar organism. Movement of a bee is depends on a local score and global score. At last the score of rearranged sequence has been measured for further modification nectar concentration. We have applied LSBCO and OVBCO to a variety of sequences obtained from sequence of dissimilar organism. Since we were unable to process by high end computation , we limit the sequences length by $\sim 10^5$bp. LSBCO has been capable to rearrange the reads with 99.9% to 100% precision with higher execution time with respect to that of OVBCO. The results clearly describe that If we can run n tours independently by using a high performance computing machine, then the running time will be reduced sharply, even though there will be of $O(N^2)$ predominant factor for evaluation of the score. This shows that the time complexity in global score estimation will be a novel ray to next generation sequencing (NGS).

REFERENCES

[1] Yuriy Brun, Solving NP-complete problems in the tile assembly model,Theoretical Computer Science, 395, 31–46, 2008.

[2] O. Tal, Two complementary perspectives on inter-individual genetic distance, Biosystems, 111,1836, 2013.

[3] Jason R. Miller and Sergey Koren and Granger Sutton, Assembly algorithms for next generation sequencing data, Genomics, 95, 315–327, 2010.

[4] J. Butler, I. MacCallum, M. Kleber, I. A. Shlyakhter, M. K.Belmonte, E. S. Lander, C. Nusbaum, and D. B. Jaffe, Allpaths: De novo assembly of whole-genome shotgun micro reads, Genome Research, 18, 810820, 2008.

[5] G. MYERS, Whole-Genome Dna Sequencing, Computing in Science and Engineering, 1, 3343, 1999.

[6] W. J. Ansorge, Next generation dna sequencing techniques, New Biotechnology, 25, 18716784, 2009.

[7] O. Isakov and N. Shomron, Challenges and Solutions: Bioinformatics-Trends and Methodologies. Intech, November 2011, Chapter 29: Deep Sequencing Data Analysis.

[8] T. J. Treangen and S. L. Salzberg, Repetitive dna and next-generation sequencing: computational challenges and solutions, Nature Reviews Genetics, 113(1), 3646, 2011.

[9] D. KARABOGA and B. Akay, A comparative study of artificial bee colony algorithm, Applied Mathematics and Computation, 214, 108132, 2009.

[10] D. Karaboga, C. Ozturk, N. Karaboga, and B. Gorkemli, Artificial bee colony programming for symbolic regression, Information Sciences, 209, 0115, 2012.

[11] D. R. Zerbino and E. Birney, Velvet: Algorithms for de novo short read assembly using de bruijn graphs, Genome Research, 1, 821829, 2008.

[12] K. Scheibye-Alsing, S. Hoffmann, A. M. Frankel, P. Jensen, and P. F. Stadler, Sequence assembly, Computational Biology and Chemistry, 33, 2009.

[13] J. S. Firoz, M. S. Rahman, and T. K. Saha, Bee algorithms for solving dna fragment assembly problem with noisy and noiseless data, in GECCO 12 Proceedings of the 14th annual conference on Genetic and evolutionary computation. New York, NY, USA: ACM, 2012, 201208.

[14] M. J. Fullwood, C.-L. Wei, and E. T. Liu, Next-generation dna sequencing of paired-end tags (pet) for transcriptome and genome analyses. Genome Research,19, 521532, April 2009.