Using Simulated Annealing to solve the DNA Fragment Assembly Problem

By

Dr. Abdul-Wahab S. Ibrahim **Baidaa A. Atya'**
Amaal Khdum **Enas Mohammed Hessian**
Al-Mustansiriayh University, the College of Education, Computer Science Department

Abstract

Different ways have used to solve the problem of simple division of DNA and the results have been obtained by the researchers varied between good results and the other did not reach the desired results. In this research, has been proposed to use the formulation of a new method to change any form by use of heating, which relies on the principle of physics idea was to bring the heat for metals and thus gradually change shape to form the required start as high a temperature and decreasing gradually to get the required form of the metal. Applying the same principle we enable for getting simple solution to the problem of the division of DNA and the results are excellent, where it is given the value of high temperature in the beginning process of the division of several divisions to stop and start a gradual decline in the value of the temperature to get to stopping of the division.

حل مشكلة الانقسام البسيط لـ DNA باستخدام التحاسبات التأقلمية

د.عبد الوهاب سامي ابراهيم بيداء عبد الخالق عطية
امال كاظم اينـاس محمد حسين
الجامعة المستنصرية / كلية التربية / قسم علوم الحاسبات

الخلاصة:-

لقد استخدمت طرق مختلفة لحل مشكلة الانقسام البسيط للـ DNA وكانت النتائج التي حصل عليها الباحثون متفاوتة ما بين نتائج جيدة واخرى لم تصل الى النتائج المطلوبة. وفي بحثنا هذا تم استخدام طريقة جديدة تدعى صياغة او تغيير أي شكل باستخدام الحرارة، حيث تعتمد فكرتها على مبدأ فيزياوي هو تسليط حرارة على المعادن وبالتالي تدريجيا يتغير شكلها الى الشكل المطلوب حيث تبدأ بدرجة حرارة عالية وتتناقص تدريجيا الى ان نحصل على الشكل المطلوب للمعدن. وبتطبيق نفس المبدأ اسستطعنا حل مشكلة الانقسام البسيط للـ DNA وبنتائج ممتازة، حيث يتم اعطاء قيمة حرارة عالية في البداية تتم عملية انقسام عديدة وتبدأ الانقسامات بالتوقف تدريجيا بهبوط قيمة درجة الحرارة الى ان يحصل التوقف بعدعدد من دورات الانقسام.

1-Introduction:-

Simulated Annealing is a well-known optimization method for finding the global optimum , developed by Metropolis and Kirkpatrick et al [1]. The basic idea of the method is to sample the space using a Gaussian distribution.

DNA fragment assembly is a technique that attempts to reconstruct the original DNA sequence from a large number of fragments, each one having several hundred base-pairs (bps) long. The DNA fragment assembly is needed because current technology, such as gel electrophoresis, cannot directly and accurately sequence DNA molecules longer than 1000 bases. However, most genomes are much longer. For example, a human DNA is about 3.2 billion nucleotides in length band cannot be read at once [2][3]. The (CSBH) technique was developed to deal with this limitation.

2-The DNA Fragment Assembly Problem

With the advance of computational science, bioinformatics has become more and more attractive to researchers in the field of computational biology. Genomic data analysis using computational approaches is very popular as well. As we know, the primary goal of any genomic project is to determine the complete sequence of the genome and its genetic content. Thus, a genome project is accomplished in two steps, the first is the genome sequencing and the second is the genome annotation (i.e., the process of identifying the boundaries between genes and other features in raw DNA sequence)[4].

In this paper, we focus on the genome sequencing, which is also known as the DNA fragment assembly problem. The fragment assembly occurs in the very beginning of the process and that other steps depend on its accuracy. The input of the DNA fragment assembly problem is a set of fragments that are randomly cut from a DNA sequence. The DNA (Deoxyribonucleic Acid) is a double helix of two anti-parallel and complementary nucleotide sequences (see Figure 1) . One strand is read from 5’ to 3’ and the other from 3’ to 5’. A DNA sequence is always read in the 5’ to 3’ direction. There are four kinds of nucleotides in any DNA sequence: Adenine (A), Thymine (T), Guanine (G), and Cytosine (C) [5].

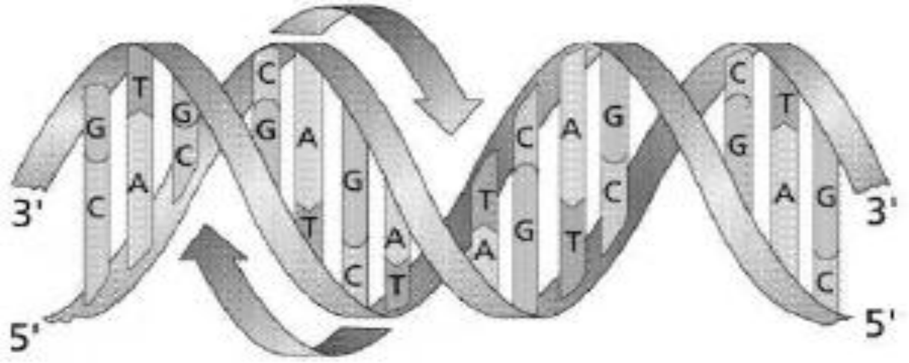

Figure. 1: Double Stranded DNA

We can think of the DNA target sequence as being the original text and the DNA fragments are the pieces cut out from the book. To further understand the problem, we need to know the following basic terminology:

- Fragment: A short sequence of DNA with length up to 1000 bps.
- Shotgun data: A set of fragments.
- Prefix: A substring comprising the first n characters of fragment f.
- Suffix: A substring comprising the last n characters of fragment f.
- Overlap: Common sequence between the suffix of one fragment and the prefix of another fragment.
- Layout: An alignment of collection of fragments based on the overlap order, i.e., the fragment order in which the fragments must to be joined.
- Contig: A layout consisting of contiguous overlapping fragments, i.e., a sequence in which the overlap between adjacent fragments is greater than a predefined threshold.
- Consensus: A sequence or string derived from the layout by taking the majority vote for each column of the layout.

To measure the quality of a consensus, we can look at the distribution of the coverage. Coverage at a base position is defined as the number of fragments at that position. It is a measure of the redundancy of the fragment data. It denotes the number of fragments, on average, in which a given nucleotide in the target DNA is expected to appear. It is computed as the number of bases read from fragments over the length of the target DNA [6].

$$Converage = \frac{\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathit{length} \text{ of fragment i}}{\text{target sequence length}} \qquad (1)$$

where n is the number of fragments. TIGR uses the coverage metric to ensure the correctness of the assembly result. The coverage usually ranges from 6 to 10 [7]. The higher the coverage, the fewer the gaps are expected, and the better the result.

3- Change the Shape By Heating (CSBH)

This method work is similar with Simulated Annealing(SA), so we can define the (SA) as a technique to find a good solution to an optimization problem by trying random variations of the current solution. A worse variation is accepted as the new solution with a probability which decreases as the computation proceeds. The slower the cooling schedule, or rate of decrease, the more likely the algorithm is to find an optimal or near-optimal solution[8].

Following is the pseudo code of SA method. What is important to note is the management of the best, current and trial states combined with the generation of random trial states and exponential random acceptance of a poorer state[9].

```
Generate an initial trail solution: trail(x)
Best(x) = curr(x) = trail(x)
For I = 1 to maxiterations
  Begin
    'generate a new random trail(x) solution from curr(x)'
     if trailCost < bestCost
       best(x) = curr(x) = trail(x)
     else if trailCost < currCost
       curr(x) = trail(x)
     else
       begin
         anneal = exp((currCost-trailCost)/t(i))
         'generate random number r between 0 and 1
         if  r < anneal
            curr(x) = trail(x)
       end
   end
```

4- DNA Fragment Assembly Using SA

The cooling schedule controls the values of the temperature parameter. It specifies the initial value and how the temperature is updated at each stage of the algorithm. Several schemata are been proposed in the literature [10]:

$$proprationSA: T_k = \alpha * T_{k-1}$$

$$CSA: T_k = \frac{T_o}{\ln(1+k)}$$

$$FSA: T_k = \frac{T_o}{1+k} \qquad (2)$$

$$VFSA: T_k = \frac{T_o}{e^{k}}$$

The proportional scheme is the most used one. In this case, the decreasing function is controlled by the α factor where α∈ (0; 1).

The number of the iterations between two consecutive changes of the temperature is given by the parameter Markov Chain length, whose name alludes the fact that the sequence of accepted solutions is a Markov chain (a sequence of states in which each state only depends on the previous one).

This operator generates a new neighbor from current solution. For our experimental runs, we use the edge swap operator. This operator randomly selects two positions from a permutation and then invert the order of the fragment between these two fragment positions.

5- the results:-

To test and analyze the performance of our algorithm, we generated two problem instances with GenFrag. GenFrag takes a known DNA sequence and uses it as a parent strand from which to randomly generate fragments according to the criteria (mean fragment length and coverage of parent sequence) supplied by the user. The first problem instance, 842596 4, contains 442 fragments with average fragment length of 708 bps and coverage 4. The second problem instance, 842596 7, contains 773 fragments with average fragment length of 703 bps and coverage 7. We evaluated the results in terms of the number of contigs assembled.

Table 1: Parameters when heading and optimum solution of the problem.

Move operator	**Edge Swap**
Markov chain length	***Tota number valuations* /100**
Coolin scheme	**Proportional (∈ =0:99)**
38524243_4(Results on the first Instances.)	**b f e t** **225744 223994 504850 7.92**
38524243_7 (Results on the second Instances.)	**b f e t** **416838 411818 501731 12.52**
Final Best Contigs. (38524243_4)	**4**
Final Best Contigs (38524243_7)	**2**

From this table we notes that in all problems SA reached to optimal solution and cost of the problem is increased when the problems become complex and the generation is increasing too.

6-the Conclusion

The DNA fragment assembly is a very complex problem in computational biology. Since the problem is NP-hard, the optimal solution is impossible to find for real cases, except for very small problem instances. Hence, computational techniques of affordable complexity such as heuristics are needed for it.

For the future, we plan to study new and more adequate fitness functions for this problem by including explicitly the number of contigs. These new functions are quite time consuming and then we are also planning to extend this study by using parallel versions of these algorithms. Previous experiments showed that the parallelism allows not only to reduce the execution time, but it also improves the accuracy in computing solution.

7-Reference

[1] P. Green. "Phrap". http://www.phrap.org/2003.

[2] G.G. Sutton, O. White, M.D. Adams, and A.R. Kerlavage. "TIGR Assembler: A new tool for assembling large shotgun sequencing projects", Genome Science & Tech., pp. 9–19, 1999.

[3] T. Chen and S.S. Skiena. "Trie-based data structures for sequence assembly. In The Eighth Symposium on Combinatorial Pattern Matching", pp. 206–223, 2001.

[4] X. Huang and A. Madan. "CAP3: A DNA sequence assembly program, Genome Research", 9, pp. 868–877, 1999.

[5] E.W. Myers. "Towards simplifying and accurately formulating fragment assembly", Journal of Computational Biology, 2(2), pp. 275–290, 2004.

[6] P.A. Pevzner. "Computational molecular biology: An algorithmic approach". The MIT Press, London, 2002.

[7] E. Alba, G. Luque, and S. Khuri. "Assembling DNA Fragments with Parallel Algorithms. In B. McKay", editor, Proceedings of the IEEE Congress on Evolutionary Computation (CEC-2005), pp. 57–65, Edinburgh, UK, 2005.

[8] J. Setubal and J. Meidanis. "Fragment Assembly of DNA, in Introduction to Computational Molecular Biology", pp. 105–139. University of Campinas, Brazil, 2000.

[9] S. Kim. "A structured Pattern Matching Approach to Shotgun Sequence Assembly". PhD thesis, Computer Science Department, The University of Iowa, 1997.

[10] C.F. Allex. "Computational Methods for Fast and Accurate DNA Fragment Assembly". UW technical report CS-TR-99-1406, Department of Computer Sciences, University of Wisconsin-Madison, 2003.