A NEW ALGORITHM FOR GLOBAL ALIGNMENT IN DNA SEQUENCING

V. Anitha[1], B.Poorna[2]
[1]Sr.Lecturer, Panimalar Engineering College,Chennai,India
[2]Corresponding author: Professor, Easwari Engineering College, Chennai,India
Email: [1]anithamoses_06@hotmail.com

Abstract

Alignment is the most basic component of biological sequence manipulation and has diverse applications in sequence assembly, sequence annotation, structural and functional predictions for genes and proteins, phylogeny and evolutionary analysis. Classical methods like Needleman Wunsch (for global alignment) in 1970 and Smith-Waterman (for local alignment) in 1981 suffer from the drawback that it involves a large number of computational steps and has a statically allocate a large section of memory for computer implementation. This paper suggests an alternate method to obtain global alignment between two sequences using logic gates and compare the performance of our algorithm with that of the classical method.

Key words: DNA sequencing –Sequence alignment –Binary representation – Logic gates – Graph Coincidence.

I. INTRODUCTION

DNA carries the genetic information for life as we know it. Before its identification by Watson and Crick in 1953, the quantum physicist Schr¼dinger had already accurately predicted the carrier of genetic information to be an "a

periodic crystal": a structured medium (crystal) capable of storing information because of

variation allowed within the structure (a periodicity)[5]. With more and more complete genomes of prokaryotes and eukaryotes

becoming available and the completion of human genome project in the horizon, fundamental questions regarding the characteristics of these sequences arise.

Life represents order. It is not chaotic or random [7]. Thus, we expect the DNA sequences that encode life to be non random. In other words, they should be very compressible. There are also strong biological evidences that support this claim: it is well-known that DNA sequences, especially in higher eukaryotes, contain many (approximate) tandem repeats; it is believed that there are only about a thousand basic protein folding patterns; it also has been conjectured that genes duplicate themselves sometimes for evolutionary or simply for "selfish: purposes. All these give more concrete support that the DNA sequences should be reasonably compressible. However, such regularities are often blurred by random mutation, translocation, cross-over and reversal events, as well as sequencing errors. It is well recognized that the compression of DNA sequences is a very difficult task [10]. The DNA sequences only consists of 4 nucleotide bases {A, C, G, T}, 2 bits are enough to store each base. So a mapping between every pair of letters that are derived from the same ancestral letter through the replication of cells called evolutionary relationships is done on some biological sequences. This has the disadvantage that it is very restrictive in not allowing point mutation or convergent evolution and too permissive by not seeking to find the truly orthologous segments between the sequences of two species. This finds all sequence similarities that are more significant than a threshold above random similarity, implying some common function[6].

The goals of the genomics community are to sequence and align a large number of species in order to study their biology and evolution through comparative sequence analysis. The unsolved problems in alignment are improved pairwise alignment with a statistical basis improved alignment based gene prediction, effective multiple genomic sequence alignment, better alignment browsers and rigorous methods for evaluating the accuracy of alignment. Sequence evolution is complicated and difficult to model probabilistically and therefore rigorous yet practical statistical methods for alignment and objective criteria for evaluation of alignments are really hard to obtain.

II. DNA SEQUENCE

DNA molecules are composed of single or double DNA fragments or often called oligonucleotides (oligos for short) or strands. Nucleotides form the basis of DNA. A single stranded fragment has a phosphor-sugar backbone and four kinds of bases denoted by the symbols A, T, G and C for the bases adenine, thymine, guanine and cytosine respectively. These four nucleic acids, which can occur in any order combined in Watson-Crick complementary pairs to form a double strand helix of DNA. Due to the hybridization reaction, A is complementary with T and C is complementary with G. Base pairs are the most common unit for measuring the length of a DNA. A DNA can be specified uniquely by listing its sequence of nucleotides on base pairs. As an example, any sequence oligonucleotides, such as 5' – ACCTG – 3' has a complementary sequence, 3' - TGGAC – 5'. Digits 5' and 3' denote orientation of DNA oligonucleotides.

Fig. 1. Double helix

Needleman Wunsch proposed an algorithm based on dynamic programming to solve the alignment problem. The main disadvantage of this method is that the scoring matrix construction and trace back causes a significant degradation in the runtime of the above algorithm. The Altschul's method to search for similarities between a query and all the sequences in a databases matches by first looking for small segments of the input sequence to match with other sequences[13]. It then builds from those matched regions to the largest ungapped region it can find. The method proposed by Pearson and Lipman searches similarities between one sequence and any group of sequences of the same types [2].

A. Existing Method

The genome is the complete set of DNA molecules inside any cell of a living organism that is passed from one generation to its offspring. The DNA is considered the blue print of life because it encodes the information necessary to produce the proteins required for all cellular processes. This makes living beings biologically similar or distinct. Early sequence alignment programs used the unitary scoring matrix. A unitary matrix scores all matches the same and penalizes all mismatches the same. Although this scoring is sometimes appropriate for DNA comparisons, using a unitary matrix amounts to proclaiming ignorance about DNA evolution and structure. Thirty years of research in aligning protein sequences have shown that different matches and mismatches among the pairs that are found in alignments require different scores. Many alternatives to the unitary scoring matrix have been suggested [12]. Swagatam Das and Debangshu Dey [15] is designed a new algorithm for local alignment in DNA and Subhra Sundar Bandyopadhyay et al. [14] have designed a algorithm for direct comparison and compare the performance of our algorithm with that of the classical method and also propose an alternate scoring scheme based on fuzzy concept , discuss its advantage over conventional PAM matrix. Guralnik et al. [3] have designed a K-means based algorithm for clustering protein sequences. Eva Bolten et al. [16] have used transitive homology, a graph theoretic based approach for clustering. Protein sequences for structure prediction. Single linkage clustering method is computationally very expensive for large set of protein sequences and it also suffers from chaining effect, Even in graph based approaches the distance matrix values are to be calculated and is also expensive for large data sets. In both the cases, distance matrix may not be accomodated in main memory and it increases the disk I/O operations. Here, we propose an algorithm as simple and the experimental results are discussed in the following sections.

B Proposed Method

The main kinds of information stored in biological databases are DNA and protein sequences. The International Nucleotide Sequence Database Collaboration, which is comprised of the DNA DataBank of Japan(DDBJ), the European Molecular Biology Laboratory(EMBL), and the GenBank at the National Center for Biotechnology Information, hosts more than 25 million sequence records comprising more than 32 billion nucleotides. The dynamic programming algorithm for alignment between two DNA sequences proposed by Smith and Waterman, Needleman-Wunsch is a very well known and versatile algorithm, and has been widely been referred in the domain of Bioinformatics. The scoring matrix construction and trace-back causes a significant degradation in the runtime of the above algorithm. Sequence similarity is actually a well-known problem in computer science. For the computer scientist, biomolecular sequences are just another source of data. As biological databases grow in size, faster algorithms and tools are needed. The information is saved in binary strings that are made up of 0 and 1 integers at computers, similarly it is saved in DNA strings that are build of A, T, C and G molecules in living individuals. In the other words, the DNA is capable to perform computing and processing tasks and in the method of saving the information on genes. Clustering is an active topic in pattern recognition, data mining, statistics and machine learning with diverse emphasis[1]. The earlier approaches do not adequately consider the fact that the data set can be too large and may not fit in the main memory of some computers. In bioinformatics, the number of DNA sequences is now more than half a million. It is necessary to devise efficient algorithms to minimize the disk I/O operations. The problem we have considered here is : Given a set of DNA sequences, implement efficient logic gates to find similarity so as to improve the accuracy and reduce the disk I/O operations, computation time and space requirements. Also, find an alternative and efficient scheme to generate a graph of DNA sequences.

C. Encoding of Nitrogen Base

Let the four kinds of nitrogen bases Adenosine, Cytosine, Guanine and Thymine be represented by a pair of binary numbers 00, 01, 10 and 11 respectively. Thus the given sequence AGCTAT will be represented by the binary string 00 10 01 11 00 11. Thus any given DNA sequence may be represented by a binary string.

III. SIMILARITY BETWEEN SEQUENCE

Let S_1 and S_2 be two DNA sequences the similarity between which needs to be determined. For example

S_1= AGTCATGGCCAA and its binary equivalent is 001011010011101001010000. Also S_2=AGTCCTGCCCAC and its binary equivalent is 001011010111100101010001. A NOR function operation between two binary variables in Boolean algebra satisfies the condition that if both the binary variables are the same the output is 1, else it is equal to zero [8]. Let X and Y are the two binary variables. Applying NOR function between the two sequences gives

Table 1. NOR gate

X	Y	X NOR Y
0	0	1
0	1	0
1	0	0
1	1	1

Now when this NOR function is applied between the binary strings corresponding to the DNA sequences results in another binary string, with a 1 value when the two corresponding bits are one.

Let R be the resultant string. The number of 1's in R gives a measure of the similarity between the two sequences. Replace two successive one's by one if they are same and replace two successive bits by 0 if they are different. Thus the resultant binary string will have a 1 when the corresponding characters match else it is a zero.

A. Proposed Algorithm

```
Step 0 : Start
Step 1 : Input two binary strings $B_1$, $B_2$ corresponding to
         the two DNA sequences $S_1$ and $S_2$
Step 2 : Form the binary string of length 'n' for $S_1$
Step 3 : Form the binary string of length 'm' for $S_2$
Step 4 : Perform exclusive NOR operation between
         corresponding bits of $S_1$ and $S_2$ to form string R of
         length p, where p= min (n, m)
         j=0
Step 5:  For i = 1 to p step 2
         Begin
         If R (i) = R(i+1) then $O_j$ = 1
                 else $O_j$ = 0
         j = j+1
         End
         cnt = 0
         half = p/2
Step 6:  For j = 1 to half
         If $O_j$ = 0 then cnt = cnt+1
Step 7:  mtch = ((half – cnt) / half ) x 100
Step 7:  The two strings match in mtch %
Step 8:  End
```

Thus for the example given above

S_1=A G T C A T G G C C A A

B_1=00 10 11 0100 11 10 10 01 01 00 00

S_2= A G T C C T G C C C A C

B_2= 00 10 11 01 01 11 10 01 01 01 00 01

R(B_1**NOR**B_2)=11 11 11 11 10 11 11 00 11 11 11 10

O = 1 1 1 1 0 1 1 0 1 1 1 0

Match = 75%

A simple subtraction of the corresponding bits will not result in a binary string, hence the logical function NOR is chosen.

IV. GRAPHICAL REPRESENTATION

The binary representation of A, C, G, T which are 00,01,10,11 can be used to construct a graph as follows. Let (0, 0) be the first point on a two dimensional graph. Then for each binary value equivalent of the alphabet the next point is suitably marked and a straight line is drawn connecting them. For example when the sequence is S_1 = AGTCATGGCCAA its binary value is 001011010011101001010000. Let p_i be the point representing the alphabet. For A (0, 0) we are drawing a slanting line in negative direction to get back the original sequence. Then the graph for the above sequence would be represented by solid line. For the sequence S_2 given above the graph for other sequence may be obtained. Then the graph for the above sequence would be represented by dashed line.

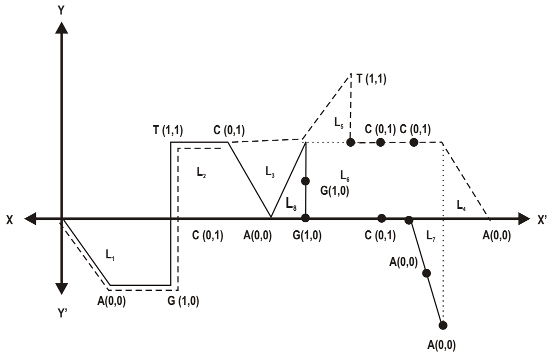


Fig. 2. Graphical representation for the sequence S_1 = AGTCATGGCCAA and the sequence S_2 = AGTCCTGCCCAC

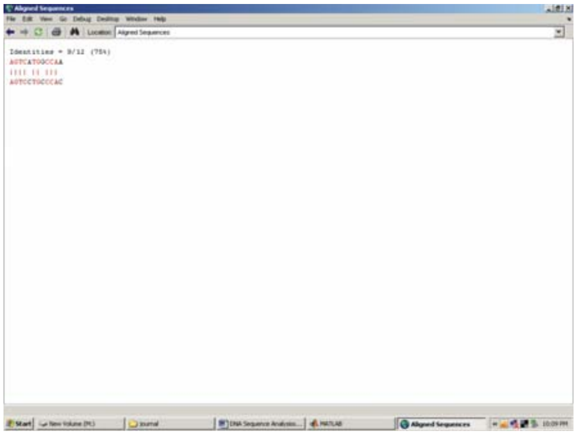
The above graph has 8 segments (L_i, i= 1 to 8). The boundary areas of the 8 segments are obtained. The following inferences are possible. By comparing both the graphs the mismatch occurs from the segment $L_{3.}$ The two graph were found to be nearly the same hence the two sequences are likely to match.

V. EXPERIMENTAL RESULTS

To evaluate the performance, a DNA sequence is considered

A. DNA Sequence Data Set

DNA sequences of human mitochondrial genome have been collected from Genbank. It contains 16571 sequences. From randomly 1919 sequences were selected for training and 200 for testing. The experiments were done on Intel Pentium-IV processor based machine having a clock frequency of 1700 Mhz and 512 MB RAM using MATLAB 7.0. MATLAB is a high-performance language for technical computing. It integrates computation, visualization, and programming in an easy-to-use environment where problems and solutions are expressed in familiar mathematical notation. MATLAB is an interactive system whose basic data element is an array that does not require dimensioning. This allows you to solve many technical computing problems. Using the showalignment method in MATLAB it finds the alignment same as the proposed algorithm. It is depicted in the figure below.



```
Identities = 9/12 (75%)
AGTCATGGCCAA
|||| || |||
AGTCCTGCCCAC
```

(b)

Fig. 3. Sequence alignment for the sequence (a) S1 = AGTCATGGCCAA and the sequence (b) S2 = AGTCCTGCCCAC

VI. CONCLUSION

Experimental results on a DNA sequence data set show that it performs well by comparing at low computation cost and this may be used to find the superfamily, family and subfamily relationships in DNA sequences. A binary representation for DNA sequence is given. The comparison of sequence using logic gates is proposed. This method gives the accuracy with which the two sequences match. It also provides information about which portion of the sequence do not match. Using the same binary representation a graph method for comparing the sequence is given which also provides the percentage of matching and the portions of the sequence that do not match. This is a simple method of comparing two DNA sequences. This method not only determines the percentage of matching of two sequences, but also identified the portion of the sequence that does not match. This information may be used to align the two sequences. We further aim at clustering a large set of DNA sequences consisting of sequences from various families and evaluate the performance of the proposed algorithm with necessary modifications. Also, we would like to compare the computation time for large data sets with few more algorithms. The proposed algorithm can also be used on text and web document collection.

REFERENCES

[1] A. K. Jain, M.N. Murty, and P.J. Flynn, Data clustering: A Review, ACM Computing Surveys, Vol. 31, 3, pp.264-323, 1999.

[2] E.Rivals, J-P. Delahaye, M. Dauchet and O.Delgrange. A Guaranteed Compression Scheme

for Repetitive DNA Sequences. LIFL Lille I University, technical report IT-285, 1995.

[3] Eva Bolten, Alexander Schliep and Sebastian Schneckener, Clustering Protein Sequences-Structure prediction by transitive homology, GCB,1999.

[4] Harshawardhan P.Bal, Bioinformatics principles and applications.

[5] J.H.M. Dassen, DNA Computing promises, problems and perspective, report (http://citeseer.nj.nec.com/jhm.html)

[6] K. Lanctot, M. Li, E. H. Yang. Estimating DNA sequence entropy, to appear in SODA' 2000.

[7] M. Li, P. Vitanyi, An introduction to kolmogorov complexity and its applications, Springer, 2nd ed., 1997.

[8] M. Morris Mano, Computer System Architecture

[9] Needleman Wunsch Algorithm –Global Alignment Algorithm using Dynamic Programming.

[10] R.Curnow and T. Kirkwood, Statistical analysis of deoxyribonucleic acid sequence data-a review. J. Royal Statistical Society, 152, 1989, 199-200.

[11] S.C. Rastogi, Namita Mendiratta, Parag Rastogi, Bioinformatics concepts, skills and applications.

[12] Sequence alignment, Journal of computational biology, 147, pp.195-197, 1970.

[13] S. Grumbach and F. Tahi, “A new challenge for compression algorithms: genetic sequences”, Journal of Information Processing and Management, 30:6(1994), 875-866.

[14] Subhra Sundar Bandyopadhyay, Somnath Paul and Amit Konar, Improved algorithm for DNA sequence alignment and revision of Scoring matrix, Proc. of ICISIP 2005.

[15] Swagatam Das & Debangshu Dey, A new algorithm for local alignment in DNA sequencing, Proc. of IEEE conference, INDICON 2004.

[16] V. Guralnik and G.karypis, A scalable algorithm for clustering sequential data, Proc. of Ist IEEE conference on Data Mining, 2001.

V. Anitha, pursuing her Ph.D in the area of Data mining in Mother Teresa Women's University, received her MCA and M.Phil degree from Bharathidasan University. Her area of interests include Bioinformatics, Data Mining, Parallel Computing. She has published several papers in International and National conferences.