International Journal of Computers & Technology
Volume 4 No. 1, Jan-Feb, 2013, ISSN 2277-3061

DNA sequence comparison based on Tabular Representation

Mrs. Archana Verma[1], Mr. R.K.Bharti[2], Prof. R.K.Singh[3]

[1]*Research Scholar, UTU, Dehradun*

Astt.Prof.CSE, BTKIT,Dwarahat

Almora, Uttarakhand, India

vermarchana05@gmail.com

[2] *Astt. Prof. CSE BTKIT, Dwarahat*

raj05_kumar@yahoo.co.in

[3]*OSD,UTU,Dehradun*

Uttarakhand, India

rksinghkec12@rediffmail.com

Abstract-

DNA sequence comparison remains as one of the critical steps in the analysis of phylogenetic relationships between species. In order to get quantitative comparison, we want to devise an algorithm that would use the tabular representation of DNA sequences. The tabular approach of representation captures the essence of the base composition and distribution of the sequence. In this contribution, we take the tabular notation for DNA sequences and then these tables are compared to find the similarity/dissimilarity measure of the sequences. We have developed algorithms for comparing DNA sequences. These programs help us to search similar segments of sequences, calculate similarity scores and identify repetitions based on local sequence similarity. There are two approaches: one is to find the exact similarity and another is to find the measurement for similarity. The first approach is more sensitive, which can be used to search DNA sequence similarities only if complete matches occurred and can compare exactly similar sequences only. This approach violates if a single mismatch for any base character appears so it is not a general solution. To find the miss matches along with the matches we have suggested another approach which compiles the information matrix based on matches and miss matches. This approach is quiet general in terms of sequences which have a large fragment common with less no of dissimilar base characters. This alternate approach includes an additional step in the calculation of the similarity score that denotes multiple regions of similarity between sequences. For both these approaches computer programs are prepared and tested on data sets. These programs can be used to evaluate the significance of similarity scores using a shuffling method that preserves local sequence composition. In addition, these programs have been generalized to allow comparison of DNA sequences based on a variety of alternative scoring matrices. We have been developing tools for the analysis of protein The method is very simple and fast, and it can be used to analyze both short and long DNA sequences.

The utility of this method is tested on the several sequences of species and the results are consistent with that reported.

***Key words*: DNA, sensitivity and selectivity, gene, NBRF**

I INTRODUCTION

Finding sequence similarities with genes of known function is a common approach to infer a newly sequenced gene's function. Various tools for the analysis of DNA sequence similarity have been developed, that achieve a balance of sensitivity and selectivity on the one hand and speed and space requirements on the other. Comparative genomics is founded on the assumption that much of life's language is contained in its linear DNA sequence. Comparisons of genomic DNA sequences present one way to understand the syntax and vocabulary of this language. One great advantage of using whole rather than partial genome sequence is that comparisons may be made between the most closely related genes or regions in the genomes compared.

One such program the FASTP program searches amino acid sequence data bases (1), which uses a rapid technique for finding identities shared between two sequences and exploits the biological constraints on molecular evolution. FASTP has decreased the time required to search the National Biomedical Research Foundation (NBRF) protein sequence data base by more than two orders of magnitude and has been used by many investigators to find biologically significant similarities to newly sequenced proteins. There is a trade-off between sensitivity and selectivity in biological sequence comparison: methods that can detect more distantly related sequences (increased sensitivity) frequently increase the similarity scores of unrelated sequences (decreased selectivity). Another program in this category is FASTA, which uses an improved algorithm that increases sensitivity with a small loss of selectivity and a negligible decrease in speed. A related program, LFASTA was also developed by the biologists, for local similarity analyses of DNA or amino acid sequences. These programs run on commonly available microcomputers as well as on larger machines. Several other works have also been done in this concern. One algorithm was also presented by authors which translates the sequences most reliably and compares the translated sequences. This method enables us to find protein sequence similarity in DNA sequences even if we do not know the protein sequences which are coded in the DNA sequence. The algorithm produces temporal DNA sequence for each original DNA sequence. They are divided into codons from the beginning and translated into protein sequences. Each temporal and original DNA sequence is

International Journal of Computers & Technology
Volume 4 No. 1, Jan-Feb, 2013, ISSN 2277-3061

compared using DNA scoring system, translated sequences are also compared using protein scoring system. The sum of the scores are calculated for all possible temporal DNA sequences. Gene similarities between two genes with known and unknown function alert biologists to some possibilities.Computing a similarity score between two genes tells how likely it is that they have similar functions Dynamic programming is a technique for revealing similarities between genes. The Change Problem is a good problem to introduce the idea of dynamic programming. In 1984 Russell Doolittle and colleagues found similarities between cancer-causing gene and normal growth factor (PDGF) gene.

Here in this paper we are proposing some new ideas for sequence comparison which are based on the tabular representation of biological sequences.

II MEHTODS AND MATERIAL

2-D RA Method for representing a biological sequence: It is a new approach to overcome the shortcoming in the previous graphical algorithms, used for the representation of biological sequences. As we know a biological sequence is a stream of characters. There are four base characters{a, t, g, c}.To represent a biological sequence, RA method uses a table. The table has two column namely base and position. The table composed of the base character and the position of the occurrences of that base character in the biological sequence. The table has four rows, one for each base character. The Base column will take one base character as a value at a time and the corresponding position column represent the location number of that base character in the Biological sequence. As shown in fig. 5.

Steps of RA method:

1. Create a table of four rows and two columns.

2. Initialize all rows of the base field as one of the base character i.e.{a, t, g, c}.

3. Read unprocessed character (k) in the sequence.

4. Find the match for the character (k) with the Base field value

5. Then write the location of k in the position field.

6. Go to step 3 until EOF of the sequence.

7. End

We apply this method on the following Biological sequence: Goat alanine β-globin 86 bases

ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGC
TTCTGGGGCAAGGTGAAAGTGGATGAAGTTGGTGCT
GAGGCCCTGGGCAG

Base	Positions
A	0,6,13,16,18,19,30,45,46,51,52,53,58,61,62,73,84
C	3,7,10,22,25,26,29,31,32,35,38,44,70,76,77,78,83
G	2,5,9,12,14,15,17,20,21,24,27,33,34,40,41,4243, 47,48,50,54,56,57,60,63,66,67,69,72,74,75,80,81, 82,85
T	1,4,8,11,23,28,36,37,39,49,55,59,64,65,68,69,71,7 9

Table 1 : RA representation of β-globin Biological Sequence

The comparison algorithms we have developed proceeds with the tabular representation of sequences. These algorithms searches the similarity among various biological sequences with varying approaches.

Exact Similar Algorithm : To find whether the given sequences are similar or not.

Step 1: construct the table for the given sequences using RA method(mentioned above).

Step 2: for each value of base column the position vector is checked.

Step 3: if both the position and base column values are matched then the sequences are exact similar.

The advantage of this algorithm is that it checks a wide positional area for a single base character, hence this method works very fast to identify the dissimilarity without going to the distant position of the sequence and only checks of a single base character are capable to denote the dissimilarity. So this method is very quick to find the dissimilarity. Disadvantage is that if sequences are dissimilar with only one character then this method simply denotes the dissimilarity without concerning the measure of similarity. This method is strictly bounded to test whether the sequences are similar or not it does not consider the level of similarity in a dissimilar sequence pair. Therefore another method is developed to check the similarity measure. This new algorithm counts the no of similar and dissimilar characters and then report how much similar are the sequences. Obviously this new approach is quiet efficient because it will not only check the similarity but also checks for the dissimilarity and the result is the information about the similarity/dissimilarity measure.

Similarity Measure Algorithm: To find the similarity measure of biological sequences.

Step 1: Construct the table for given sequences using RA method.

Step 2: Compare the base column value of the tables and store the dissimilarity into count variable(by incrementing the count whenever the base character and position column values are different).

Step 3: If value of count is less than or equal to the half of the length of the smaller sequence then report that the sequences are similar less than or equal to 50%.

Both these algorithms achieve much of their speed and selectivity due to the use of tabular representation. A table is helpful to locate all identities or groups of identities between two DNA sequences in its base and position vectors. The ExactSimilar algorithm picks consecutive base-position pair and prepare a report which only denotes the similarity/dissimilarity of the sequences. This method works fast by just checking a single base character for a wide positional area. For example while checking for character A if position vector contains 89 in it then the check for single character i.e. A travels the long distances in a sequence without actually exploring the sequence upto that length.

III DISCUSSION ABOUT THE EXISTING COMPARISON ALGORITHM AND USE OF EARLIER DEVELOPED METHODOLOGIES

The growth of Database used for DNA sequences is exponential in nature i.e. the number of DNA sequences are increasing very rapidly. Comparison among various segments or sequences provides us a meaningful information which will be helpful in the study of DNA sequences. The comparative study makes it easy to identify the behavior of a genome and disease identification will be faster in such a manner. Several techniques have been developed for comparing DNA sequences. There have been developed several computer programs for comparisons of protein and DNA sequences. They can be used to search sequence data bases, evaluate similarity scores and identify periodic structures based on local sequence similarity. Some of the comparison methodologies are FASTA, FASTP and LFASTA programs. The FASTP program can be used to search protein or DNA sequence data bases and can compare a protein sequence to a DNA sequence data base by translating the DNA data base as it is searched. The FASTA program is a more sensitive derivative of FASTP. FASTA includes an additional step in the calculation of the initial pairwise similarity score that allows multiple regions of similarity to be joined to increase the score of related sequences. The LFASTA program can display all the regions of local similarity between two sequences with scores greater than a threshold, using the same scoring parameters and a similar alignment algorithm; these local similarities can be helpful for individual alignments. In addition, these programs have been generalized to allow comparison of DNA or protein sequences based on a variety of alternative scoring matrices. One method for DNA sequence comparison is based on amino acid similarity, which translates the sequences most reliably and compares the translated sequences. This method enables us to find protein sequence similarity in DNA sequences even if we do not know the protein sequences which are coded in the DNA sequences. In order to get quantitative comparison another approach is used in which one has to derive some mathematical descriptors that would capture the essence of the base composition and distribution of the sequence. In this strategy, DNA sequence double helix is taken into consideration, and a mathematical descriptor is introduced for each DNA sequence based on the frequencies of codons it contains.

The above discussed comparison methodologies are few of the several existing predefined algorithms. Each and every method provides different experimental results with some more or little shortcomings. We have captured those shortcomings and try to resolve them with our proposed algorithms(i.e. ExactSimilar algo and SimilarityMeasure algo).

IV Experimental results

Testing the proposed algorithms on DNA sequences taken by NCBI provides us the information regarding their comparison. We are considering rat β-globin and goat ananine β-globin sequences as example sequences for applying the algorithms. Both the sequences and their tabular representations are as follows:

Sequence of Rat β-globin:

ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTT
AGTGGCCTGTGGGGAAAGGTGAACCCTGATAATGTT
GGCGCTGAGGCCCTGGGCAG

Base	Positions
A	0,7,11,12,16,22,24,25,30,36,50,51,52,57,58,64,66,67,79 ,90
C	6,8,9,13,19,28,31,41,42,59,60,61,74,76,82,83,84,89
G	2,3,5,15,18,21,23,26,27,33,37,39,40,44,46,47,48,49,53, 54,56,63,69,72,73,75,78,80,81,86,87,88,91
T	1,4,10,14,17,20,29,32,34,35,38,43,45,55,62,65,68,70,71 ,77,85

Table 2: RA representation of Rat β-globin

Sequence of Goat alanine β – globin:

ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGC
TTCTGGGGAAGGTGAAAGTGGATGAAGTTGGTGCTG
AGGCCCTGGGCAG

Base	Positions
A	0,6,13,16,18,19,30,45,46,51,52,53,58,61,62,73,84
C	3,7,10,22,25,26,29,31,32,35,38,44,70,76,77,78,83
G	2,5,9,12,14,15,17,20,21,24,27,33,34,40,41,42,43,4 7,48,50,54,56, 57,60,63,66,67,69,72,74,75,80,81,82,85
T	1,4,8,11,23,28,36,37,39,49,55,59,64,65,68,69,71,7 9

Table 3: RA representation of Goat alanine β - globin

Viewing these tables and inspecting corresponding base and position vectors we can easily identify that both the sequences are dissimilar. The tabular representation makes the comparison easier as we have experienced while comparing the above sequences. ExactSimilar algorithm displays that these sequences are not similar then we use SimilarityMeasure algorithm which provides the information regarding the dissimilarity of both the sequences.

V Conclusion

In this paper we have proposed algorithms for comparing biological sequences. These algorithms are based on tabular representation of biological sequences. After testing we have experienced that due to the tabular representation approach comparison becomes less complex and faster. As an application, this method is tested on several sequences, and the results under these methods have an overall agreement which all are consistent with that reported previously, thus proves the utility of this new approach. Experimental results show that our algorithm uses fewer comparisons to perform searches and has a shorter elapsed searching time. Our proposed algorithm is therefore applicable for searching biological sequence databases as well as any other string searching applications.

REFERENCES

[1] Y. Zhang, B. Liao, K. Ding, On 2D graphical representation of DNA sequence of nondegeneracy, Journal of Chemical Information and Computer Science 411 (2005), 28-32.

[2] M. Randic, M. Vra'cko, N. Ler's, D. Plav'si�c, Analysis of similarity/dissimilarity of DNA sequence based on novel 2- D graphical representation, Journal of Chemical Information and Computer Science 371 (2003), 202-207.

[3] B. Liao, T. Wang, 3-D graphical representation of DNA sequences and their numerical characterization, Journal of Molecular Structure (THEOCHEM) 681 (2004), 209-212.

[4] Y. Zhang, B. Liao, K. Ding, On 3DD-curves of DNA sequences, Mol. Simul. 32 (2006), 29-34.

[5] R. Chi, K. Ding, Novel 4D numerical representation of

DNAsequences Chemical Physics Letters 407 (2005), 63-67.

[6] B. Liao, R. Li, W. Zhu, On the similarity of DNA primary sequences based on 5-D representation, Journal of Mathematical Chemistry 42 (2007), 47-57.

[7] B. Liao, T. Wang, Analysis of similarity/dissimilarity of DNA sequences based on nonoveriapping trinucleotides of nucleotide bases, Journal of Chemical Information and Computer Science 44 (2004), 166-1670.

[8] E. Hamori, J. Ruskin, H curves: a novel method of representation of nucleotide series especially suited for long DNA sequences, Journal of Biological Chemistry 258 (1983), 1318- 1327.

[9] A. Nandy, A new graphical representation and analysis of DNA sequence structure: I. Methodology and Application to Globin Genes,Curro Sci. 66 (1994), 309-314.

[10] M. Randi�c, M. Vracko, A. Nandy, S.C. Basak, On 3-D graphical representation of DNA primary sequences and their numerical characterization. Journal of Chemical Information and Computer Science 40 (2000), 1235-1244.

[11] R.K.Bharti, A.Verma. Prof. R.K.Singh" A New 2-D RA Method of Representation and Analysis of DNA Sequences", International Conference of Networking and InformationTechnology(ICNIT), PHILIPINES,10.1109/ICNIT.2010.5508475