Multiple Sequence Alignment Based Upon Statistical Approach of Curve Fitting

Vineet Jha, Mohit Mazumder*, Hrishikesh Bhuyan, Ashwani Jha, and Abhinav Nagar

InSilico Biosolution, 103B, North Guwahati College Road, Abhoypur, Near IIT-Guwahati, P.O – College Nagar, North Guwahati – 781031, Assam, India
mazumder.mohit@gmail.com

Abstract. The main objective of our work is to align multiple sequences together on the basis of statistical approach in lieu of heuristics approach. Here we are proposing a novel idea for aligning multiple sequences in which we will be considering the DNA sequences as lines not as strings where each character represents a point in the line. DNA sequences are aligned in such a way that maximum overlap can occur between them, so that we get maximum matching of characters which will be treated as our seeds of the alignment. The proposed algorithm will first find the seeds in the aligning sequences and then it will grow the alignment on the basis of statistical approach of curve fitting using standard deviation.

Keywords: Multiple Sequence Alignment, Sequence Alignment, Word Method, Statistically Optimized Algorithm, Comparative Genome Analysis, Cross Referencing, Evolutionary Relationship.

1 Introduction

Multiple sequence alignment is a crucial prerequisite for biological sequence data analysis.

It is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. A large number of multi-alignment programs have been developed during last twenty years. There are three main considerations in choosing a program: biological accuracy, execution time and memory usage. Biological accuracy is generally the most important concern amongst all. Some of the prominent and accurate programs according to most benchmarks are *CLUSTAL W* [1], *DIALIGN* [2], *T-COFFEE* [3], MAFFT, MUSCLE, PROBCONS . An overview about these tools and other established methods are given [4].

T-COFFEE is a prototypical consistency- based method which is still considered as one of the most accurate program available. MAFFT and MUSCLE have a similar design, building on work done by Gotoh in the 1990s that culminated in the PRRN

* Corresponding author.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 184–192, 2009.

and PRRP programs [10,11], which achieved the best accuracy of their time but were relatively slow and were not widely adopted.

The recently published methods ALIGN-M [12], DIALIGN [2, 13], POA [14, 15] and SATCHMO [16] have relaxed the requirement for global alignment by allowing both alignable and non-alignable regions. Although these methods are sometimes described as 'local'; alignable regions must still be co-linear (i.e. appear in the same order in each sequence). This model is appropriate for protein families with well-conserved core blocks surrounded by variable regions, but not when input sequences have different domain organizations.

2 Objective

The main purpose of undertaking this study is to develop a new algorithm which can align multiple sequences together in lesser time.

The alignment of sequences will be done on the basis of statistical approach in lieu of heuristics approach. Statistical method of curve fitting by means of standard deviation and stochastic approaches will be used.

3 Algorithm for Multiple Sequence Alignments

1. Take 2 sequences.
2. Find the seed for alignment.
3. Build the distance matrix between the sequences.
4. Plot a graph.
5. Plot the scores of sequences in distance matrix on the graph about the line x=y.
6. Find the minimum deviation of that curve or line of minimum deviation.
7. Draw the minimum deviation onto the graph and name it as master alignment.
8. Arrange the alignment according to the master sequence alignment on distance matrix.

3.1 Detail Explanation of above Algorithm

Here we consider a DNA sequence as line not as string where each character represents a point in the line. The sequences are lined up in such a way that maximum overlap occurs between them, thus giving maximum matching of characters.

3.1.1 Hypothesis

For example consider these two sequences

DNA1 ATCGGGGCTATC
DNA2 ATCCCCCTATTG

A matrix was built according to the following conditions. If the character matches in the two sequences the distance between the two point is 0 and if mismatch occurs then the distance is given by the PAM [17] or BLOSSUM [18] matrix. The distance

contributed from the deviation to the line where x=y. Then the matrix is constructed and plotted on a graph(fig 1). The graph and matrix are as following-

Table 1. This is matrix formed between the two DNA sequences in this match score=0 and mismatch score is given by PAM and BLOSSUM matrix 1st and 2nd row consist of DNA1 & DNA2 unaligned row is taken as zero at this stage ,3rd &5th rows are the score of matrix

A	T	C	G	G	G	G	C	T	A	T	C	DNA1
A	T	C	C	C	C	C	T	A	T	T	G	DNA2
0	0	0	1	1	1	1	2	1	1	0	1	DNA1
0	0	0	0	0	0	0	0	0	0	0	0	UNALIGNED SEQ
0	0	0	-1	-1	-1	-1	-2	-1	-1	0	-1	DNA2

The graph will look like fig 1 for the above sequences. The matching string will be called as seed (where our final alignment will grow) and the mismatch string will expand like a bulb. Now the main problem whice arises at this stage is how to deflate the bulb. This is where statistics comes into action. The bulb is like a scattered plot. We have to find a line or curve of minimum deviation in the bulb which will be our master alignment line.

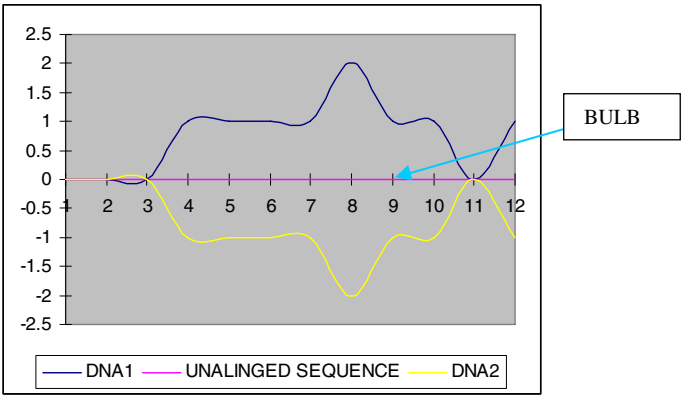

Fig. 1. This graph shows the score of matches & mismatches of both DNA1 & DNA2, In this matches are called as "seed" & mismatches will appear as "bulb"& the middle line represents the Master Alignment

A matrix was built which contains the master alignment. A score was assigned to it. The master alignment line will guide the sequences to align properly as following:

Table 2. This matrix consist Aligned sequence score along with previous score of DNA1 &DNA2 which is obtained from above graph

A	T	C	G	G	G	G	C	T	A	T	C	DNA1
A	T	C	C	C	C	C	T	A	T	T	G	DNA2
0	0	0	1	1	1	1	2	1	1	0	1	DNA1
0	0	0	0	0	0	-1	2	-2	1	-1	0	ALING SEQ
0	0	0	-1	-1	-1	-1	-2	-1	-1	0	-1	DNA2

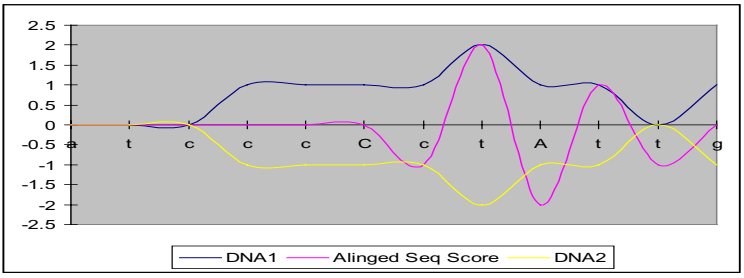

Fig. 2. This graph shows the curve of scores of DNA1 &DNA2 along with the curve obtained by the minimum deviation of these named as "MASTER ALIGNMENT"

After the matrix is prepared the final alignment is done on the basis of scoring matrix, adjusting gaps and mismatch, checking the maximum aligning score. Introduction of gaps are done to maximize the alignment so that better alignment and homology is achieved.

Table 3. This matrix shows the alignment done by the help of "MASTER ALIGNMEENT" obtained on the previous graph by the help of minimum deviation method

a	t	c	g	g	g	g	c	t	a	t	#	c	DNA1
a	t	c	c	c	c	#	c	t	a	t	t	g	DNA2
a	t	c	g/c	g/c	g/c	g	c	t	a	t	t	c/g	MASTER ALINGMENT (MA)
0	0	0	1	1	1	0	0	0	0	0		1	DNA1 SCORE
0	0	0	0	0	0		0	0	0	0	0	0	MA SCORE
0	0	0	-1	-1	-1	0	0	0	0	0	0	-1	DNA2 SCORE

Final graph based on the matrix score was obtained. Size of the bulb is reduced. The bulb that is visible in the graph is showing the mismatch between the bases present in the alignment.

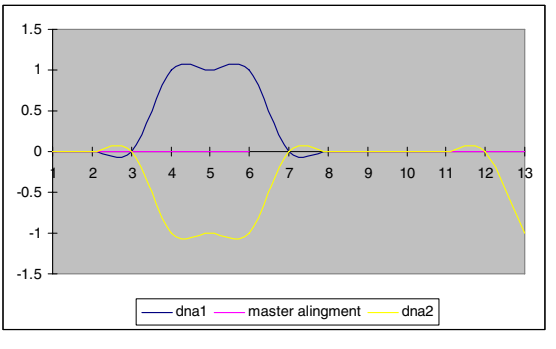

Fig. 3. This figure shows the alignment of DNA1 & DNA2 with the final score along with the Master Alignment curve

Final alignment of the sequences will be **Colour Notaions**

ATCGGGGCTAT- C **DNA1** ATCG Shows the mismatch in the sequences

ATCCCC - CTATTG **DNA2** ATCG shows the gaps in the sequences

Consider the alignment of tree or more sequences

DNA1	**ATCGGGGCTATC**
DNA2	**ATCCCCCTATTG**
DNA3	**CGCCCGGCTATG**

A matrix was built according to the following conditions. If the character matches in the two sequences the distance between the two point is 0 and if mismatch occurs then the distance is given by the PAM [17] or BLOSSUM [18] matrix. The distance contributed from the deviation to the line where x=y. Then the matrix is constructed and plotted on a graph(fig 4). The graph and matrix are as following-

Table 4. This matrix is same as that in previous example but this is for aligning 3 DNA sequences

a	t	c	g	g	g	g	c	t	a	t	c	DNA1
a	t	c	c	c	c	c	t	a	t	t	g	DNA2
c	g	c	c	c	g	g	c	t	a	t	g	DNA3
2	2	0	1	1	1	1	2	1	1	0	1	DNA1SCORE
2	2	0	-1	-1	-1	-1	-2	-1	-1	0	-1	DNA2 SCORE
-2	-2	0	-1	-1	1	1	2	1	1	0	-1	DNA3 SCORE
0	0	0	0	0	0	0	0	0	0	0	0	UNALING SCORE

The graph will look like fig 4 for the above sequences. The matching string will be called as seed (where our final alignment will grow) and the mismatch string will expand like a bulb. Now the main problem whice arises at this stage is how to deflate the bulb. This is where statistics comes into action. The bulb is like a scattered plot. We have to find a line or curve of minimum deviation in the bulb which will be our master alignment line.

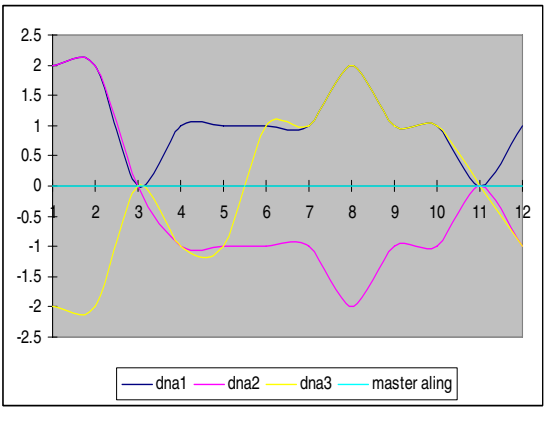

Fig. 4. This graph shows the score of matches & mismatches of both DNA1, DNA2, DNA3similarly as in the early case & the middle line represents the Master Alignment

A matrix was built which contains the master alignment. A score was assigned to it. The master alignment line will guide the sequences to align properly as following:

Table 5. This matrix consist Unaligned sequence score along with score of DNA1, DNA2& DNA3 obtained from above graph

a	t	c	g	g	g	g	c	t	a	t	c	DNA1
a	t	c	c	c	c	c	t	a	t	t	g	DNA2
c	g	c	c	c	g	g	c	t	a	t	g	DNA3
2	2	0	1	1	1	1	2	1	1	0	1	DNA1SCORE
2	2	0	-1	-1	-1	-1	-2	-1	-1	0	-1	DNA2 SCORE
-2	-2	0	-1	-1	1	1	2	1	1	0	-1	DNA3 SCORE
2	2	0	-1	-1	1	1	-2	1	1	0	-1	UNALING SCORE

After the matrix is prepared the final alignment is done on the basis of scoring matrix, adjusting gaps and mismatch, checking the maximum aligning score. Introduction of gaps are done to maximize the alignment so that better alignment and homology is achieved.

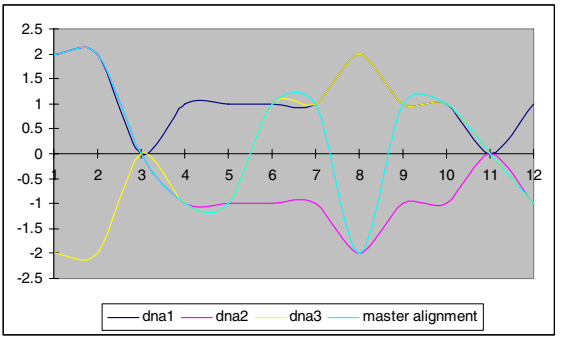

Fig. 5. This graph shows the curve of scores of DNA1, DNA2 &DNA3 along with the curve obtained by the minimum deviation of these curve for the alignment purpose named as "MASTER ALIGNMENT"

Table 6. It consist of the final alignment score of DNA1, DNA2& DNA3 which is used to draw the final alignment curve , these scores are obtained by aligning the DNA's with the help of Master Alignment by shifting of bases

a	t	c	g	g	g	g	c	t	a	T	#	c	DNA1
a	t	c	c	c	c	#	c	t	a	T	t	g	DNA2
c	g	c	c	c	g	g	c	t	a	T	#	g	DNA3
2	2	0	1	1	1	0	0	0	0	0		1	DNA1SCORE
2	2	0	-1	-1	-1		0	0	0	0	0	-1	DNA2 SCORE
-2	-2	0	-1	-1	1	0	0	0	0	0		-1	DNA3 SCORE
2	2	0	-1	-1	1	0	0	0	0	0	0	-1	ALIGNMNT SCORE

Final graph based on the matrix score was obtained. Size of the bulb is reduced. The bulb that is visible in the graph is showing the mismatch between the bases present in the alignment.

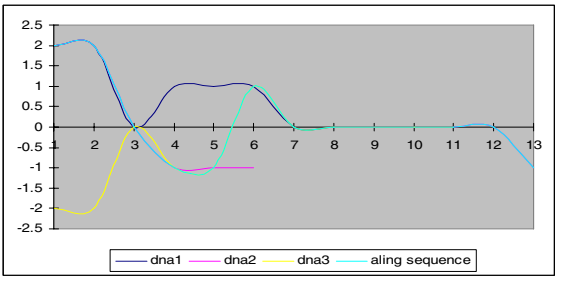

Fig. 6. This figure shows the alignment curve of DNA1, DNA2& DNA3 with the final score in the Matrix along with the Master Alignment curve

Table 7. This matrix shows the alignment of all 3 DNA sequences done by the help of "MASTER ALIGNMENT" obtained on the previous graph by the help of minimum deviation method

a	t	c	g	g	g	g	c	t	a	t	#	c	DNA1
a	t	c	c	c	c	#	c	t	a	t	t	g	DNA2
c	g	c	c	c	g	g	c	t	a	t	#	g	DNA3 SCORE
a/c	t/g	c	g/c	g/c	g/c	g	c	t	a	t	t	c/g	MASTER ALIGNMENT

Final alignment of the sequences will be:

DNA1 ATCGGGGCTAT –C

DNA2 ATCCCC – CTATAG

DNA3 CGCCCGGCTAT -G

COLOUR NOTAION

ATCG Shows the mismatch in the sequences

ATCG shows the gaps in the sequences

4 Result and Discussion

The main feature of the algorithm is sequential execution through seed finding and growing those seeds on the basis of statistical approach of curve fitting using standard deviation.

The seed of alignment can be found either by using heuristics approach of BLAST [19] or FASTA algorithm or using DOTPLOT approach. Here we have used DOTPLOT approach to find out the short or long inter sequences match (seed). Seeds are then filtered out and un-aligned sequences will be aligned using this algorithm.

This algorithm is unique as it is not based on progressive method and also doesn't divide the sequences in bits and pieces for alignment so it gives better alignment. As it uses seeds to align the sequences it gives better homology.

Results have shown that this algorithm worked well even if seed is absent or the sequences are far apart from each other.

It can align cDNA with gene which most of MSA algorithms fails to do. Advantage of cDNA not containing the introns can be used here. Exons on the cDNA are aligned with coding genes in the DNA sequence which can than be used as seeds.

This leads to a phenomenon of DNA looping. These looping DNA are aligned by creating gap in the cDNA.Here these gaps indicates introns in the genome.

Here we have aligned the EST with gene to find out point mutation. The data was downloaded from UCSC genome browser[20,21]. UCSC refgene contains the co-ordinates for exon and intron in gene. While aligning EST with gene the EST must match with exon (as a biological phenomenon). Thus we have calibrated our MSA verses (Stretcher[22], Matcher[22,23], Needle[24] and Waunch[25]) to find match and mismatch in EST verses gene alignment.

While performing the alignment we noticed that exon can be subdivided if it contains gaps with respect to EST and vice versa. Thus we divided the gene in such a way that no gaps occur in EST and exon. Now two sequences were taken containing only mismatches. Any gap here is considered as false positive(FP) as no sequence contains gap. Any mismatch which is missed will be reported as true negative(TN) . Any wrong mismatch reported will be false negative(FN). All true matches is true positive(TP).

Sensitivity
TP (true positive)
FN (False negative)
Sensitivity= TP/TP+FN

Specificity
TN(true negative)
FP(false positive)
Specificity= TN/TN+FP

Table 8. Showing sensitivity and specificity of the algorithm in comparison to other popular algorithms

Program	Sensitivity	Specificity
Our Algorithm	99.91	98.73
Stretcher	99.51	107.37
Matcher	99.44	106.25
Needelman-Waunch	98.14	61.17
Smith-Waterman	95.9	60.96

Above comparison shows that this algorithm gives better sensitivity among all other algorithms taken into account and gives specificity comparable to all other algorithms.

References

1. Thompson, J.D., Higgins, D.G., Gibson, T.J.: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22, 4673–4680 (1994)
2. Morgenstern, B.: DIALIGN: Multiple DNA and Protein Sequence Alignment at BiBiServ. Nucleic Acids Research 32, W33–W36 (2004)
3. Notredame, C., Higgins, D., Heringa, J.: T-Coffee: a novel algorithm for multiple sequence alignment. J. Mol. Biol. 302, 205–217 (2000)

4. Notredame, C.: Recent progress in multiple sequence alignment: a survey. Pharmacogenomics 3, 131–144 (2002)
5. Lee, C., Grasso, C., Sharlow, M.F.: Multiple sequence alignment using partial order graphs. Bioinformatics 18(3), 452–464 (2002)
6. Edgar, R.: MUSCLE: Multiple sequence alignment with high score accuracy and high throughput. Nuc. Acids Res. 32, 1792–1797 (2004)
7. Do, C.B., Mahabhashyam, M.S., Brudno, M., Batzoglou, S.: ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Research 15, 330–340 (2005)
8. Katoh, K., Misawa, K., Kuma, K., Miyata, T.: MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059–3066 (2002)
9. Edgar, R.C.: MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5, 113 (2004)
10. Gotoh, O.: Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. J. Mol. Biol. 264, 823–838 (1996)
11. Gotoh, O.: A weighting system and algorithm for aligning many phylogenetically related sequences. Comput. Appl. Biosci. 11, 543–551 (1995)
12. Van Walle, I., Lasters, I., Wyns, L.: Align-m-a new algorithm for multiple alignment of highly divergent sequences. Bioinformatics 20, 1428–1435 (2004)
13. Morgenstern, B.: DIALIGN: 2 improvement of the segment-tosegment approach to multiple sequence alignment. Bioinformatics 15, 211–218 (1999)
14. Grasso, C., Lee, C.: Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. Bioinformatics 20, 1546–1556 (2004)
15. Lee, C., Grasso, C., Sharlow, M.F.: Multiple sequence alignment using partial order graphs. Bioinformatics 18, 452–464 (2002)
16. Edgar, R.C., Sjölander, K.: SATCHMO: sequence alignment and tree construction using hidden Markov models. Bioinformatics 19, 1404–1411 (2003)
17. Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C.: A model of evolutionary change in proteins. In: Dayhoff, M.O. (ed.) Atlas of Protein Sequence and Structure, vol. 5(3), pp. 345–352 (1978)
18. Henikoff, S., Henikoff, J.: Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA 89(biochemistry), 10915–10919 (1992)
19. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. J. Mol. Biol. 215(3), 403–410 (1990)
20. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., Haussler, D.: Genome Res. 12(6), 996–1006 (June 2002)
21. University of California santa Cruz, `http://genome.ucsc.edu/`
22. Rice, P., Longden, I., Bleasby, A.: EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16, 276–277 (2000)
23. MacLaughlin, D.S.: MATCHER: a program to create and analyze matched sets. Comput. Programs Biomed. 14(2), 191–195 (1982)
24. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48(3), 443–453 (1970)
25. Smith, T.F., Waterman, M.S., Fitch, W.M.: Comparative biosequence metrics. J. Mol. Evol. 18(1), 38–46 (1981)