A Survey of Sequence Alignment Algorithms

[1]Arunima Mishra,[2] Sudhir Singh Soam,[3]Surya Prakash Tripathi

[1] Dr A.P.J. A.K Technical University, Uttar Pradesh, [2, 3] Department of CS & Engg , I. E.T. , Lucknow

***Abstract* - Biological Sequence alignment is widely used in the field of Bioinformatics and computational biology to determine the similarity between the biological sequences. Many computational methods have been suggested for sequence alignment. For example, dynamic Programming provides a solution for aligning the biological sequences with time complexity of the order of O (MN). Heuristic approach always works to find related sequences in a database search but does not have the guarantee of an optimal solution like the dynamic programming algorithm but these methods are 50-100 times faster than dynamic programming therefore better suited to search databases. In this paper a survey of various computational approaches used for aligning the biological sequences has been given for different auxiliary data structure for read sequences or reference sequences or both.**

Keywords – Sequence alignment, Dynamic programming, Heuristic approach, auxiliary data structure.

I. INTRODUCTION

Sequence alignment [1] is one of the major tasks in the bioinformatics. It consists of aligning a query sequence to a sequence database with the aim of determining those sequences that have statistically significant matches to that of the query sequence. It is different from the classical problem of string matching [2] in computer science, where we are interested to find out exact matches. Sequence Alignment is a problem of approximate string matching or string matching allowing errors [3]. The problem in its most general form is to find the position of a text where a given pattern occurs allowing a limited number of errors in the matches. The errors are those operations that biologist knows are common to occur in genetic sequences. The distance between the two sequences is defined as the minimum sequence of operations to transform one in to the other. With regard to likelihood, the operations are assigned a cost, such that the most likely operations are cheaper and the goal is to minimize the total cost.

A. Sequence Alignment

An alignment between two sequences is simply a pairwise match between the characters of each sequence. A true alignment of nucleotide or amino acid sequences is one that reflects the evolutionary relationship between two or more homologous sequences [4] that share a common ancestor. If the same letter occurs in both sequences then this position has been conserved in evolution. If the letters differ it is assumed that the two derive from an ancestral letter (which could be one of the two or neither). Homologous sequences may have different length, though, which is nearly explained through insertions or deletions in sequences. Thus, a letter or a stretch of letters may be paired up with dashes in the other sequence to signify such an insertion or deletion. Since an insertion in one sequence can always be seen as a deletion in the other one frequently uses the term "indel".

AATCTATA	AATCTATA	AATCTATA
AAG–AT--A	AA- G–ATA	AA----GATA

Three possible gapped alignments between two short sequences.

Sequence alignment can be divided mainly in two categories-

1) Pair wise sequence alignment

Pair wise sequence analysis examines the similarities of two sequences by searching for the alignment with the highest score. There are two types of alignments for pairwise sequence analysis based on dynamic programming method:

a) Global alignment: Attempt to align every residue in every sequence, are most useful when the sequences in the query set are similar and of roughly equal size. Needleman and Wunsch [5] were the first to present a dynamic programming algorithm that could find the global alignment between two amino acid sequences.

b) Local alignment: Are more useful for dissimilar sequences that are suspected to contain regions of similarity within their larger sequence context. Smith and Waterman [6] introduced a new algorithm with a different method of scoring similarity aimed at finding optimum local alignment sub-sequences, at the expense of the global score.

A. Multiple Sequence Alignment

Multiple sequence alignment aims to find similarities between many sequences. One of the multiple sequence alignment solutions are heuristic algorithms with approximate approaches, such as the CLUSTAL family of programs created by Higgins [7][8][9]which use a progressive algorithm by Feng and Doolittle[10]. Profile Hidden Markov Models (HMMs) provide another successful solution to the problem of MSA. They were introduced by Krogh and colleagues in 1994[11]. . A set of methods to produce MSAs while reducing the errors inherent in progressive methods are classified as "iterative" because they work similarly to progressive methods but repeatedly realign the initial sequences as well as adding new sequences to the growing MSA. A popular iteration-based method called MUSCLE (multiple sequence alignment by log-expectation) improves on progressive methods with a more accurate distance measure to assess the relatedness of two sequences [12] .Both pair wise and multiple sequence alignment algorithms use substitution matrices to score the sequence alignment. Substitution matrices evaluate potential replacements for protein and nucleic acid sequences. Both pairwise and multiple sequence alignment algorithms use substitution matrices to score the sequence alignment. In substitution matrices each possible residue substitutions given a score reflecting the probability of such a change. There are two popular protein substitution matrix models: Percent Accepted Mutation (PAM)[13]and Blocks Substitution Matrix (BLOSUM)[14].

In this article we review general alignment techniques, their improvements, applications and their shortcomings. Finally we will discuss the future of alignment algorithms.

II. OVERVIEW OF SEQUENCE ALIGNMENT ALGORITHMS

Needleman-wunsch is a well-known algorithm for global alignment based on the concept of dynamic programming but it is suitable for only short sequences and becomes very slow in case of long sequences(Time complexity for Needleman Wunsch algorithm is O(MN)). Smith-Waterman algorithm does local alignment and compares all bases against to all bases which is clearly too slow its time complexity is same as to Needleman-wunsch algorithm. One solution of this approach is given by BLAST[15] that find exact short seed matches which are then extended to longer alignment

A. FASTA and the BLAST Family

FASTA[16] and BLAST both are Heuristic Methods they prune the search space by using fast approximate methods to select the sequences of the database that are likely to be similar to the query and to locate the similarity region inside them. FASTA algorithm performs local optimization so that dissimilar portions of the sequence outside the optimized alignment region do not affect the score of the alignment. To make search faster FASTA algorithm uses lookup table to locate all identities or groups of identities between two DNA or amino acid sequences for the first step of comparison.

In conjunction with the lookup table it uses the diagonal method to find all regions of similarity between the two sequences. The speed and sensitivity is controlled by the parameter called ktup (k touple) which specifies the size of the word. Lesser the ktup value more sensitive the search by default ktup=2 is for protein search and 4 or 6 is for nucleotide.

This method identifies region of a diagonal that have the highest density of ktup matches. FASTA uses a formula for scoring ktup matches that incorporate the actual pam 250 values for aligned residues. FASTA saves some best local regions regardless of whether there are on same or on different diagonals. These few high scoring regions are partial alignments without gaps. Then FASTA algorithm checks that if there are several initial regions with score greater than the cutoff it checks whether they can be joined to get a gapped alignment.

BLAST also works on local similarity of sequences. It Seeks maximum segment pair which is the highest scoring pair of identical length segment chosen from two sequences.(The MSP should be locally maximum means its score cannot be improved through extending or by shortening both segments and is above some cutoff). Then it does a rapid approximation of MSP scores. Sequence database is enormous but only handful of sequences are homologous to the query sequence. To make the search fast BLAST searches for a word of fixed length w. The main strategy of Blast is to seek only segment pair that contains a word pair with a score of at least T. Any such hit is extended to determine that if it is contained within a segment pair whose score is greater than or equal to S. Smaller value of T time of algorithm. FASTA searches generally require significantly more execution time than the BLAST searches. However FASTA algorithms are considered by some to be more sensitive than BLAST, particularly when the query sequence is repetitive.

B. Gapped Alignment

Gapped blast or PSI-BLAST [17] is a .refined version of blast it runs three times faster than original BLAST. It uses a method that convert statistically significant alignments produced by BLAST in to a position specific score matrix. The Extension step in BLAST algorithm takes the maximum time, to reduce this time PSI-BLAST proposes a two hit method that seeks the existence of two non-overlapping words in the same diagonal and within a distance A to one another. For the sake of sensitivity the parameter T must be lowered. The BLAST 2 SEQUENCES [18] program finds multiple local alignments between two sequences, allowing the user to detect homologous protein domains or internal sequence duplications. `BLAST 2 SEQUENCES' is an interactive tool that utilizes the BLAST engine for pairwise DNA-DNA or protein-protein sequence comparison and is based on the same algorithm and statistics of local alignments. The BLAST 2.0 algorithm generates a gapped alignment by using dynamic programming to extend the central pair of aligned residues. This tool is better suited mostly to compare two sequences that are already known to be homologous.

C. *Algorithms that Preprocesses Database*

There are sequence comparison methods that preprocesses the database such that BLAT[19] and SSAHA[20] that are designed to find matches when the query and the database subsequences are highly similar. Index MegaBLAST[21] preprocesses the database into a data structure for rapid seed searching. Index Mega BLAST is the part of BLASTn program in the NCBI C++ tool kit it preprocesses the database rather than the query to build a data structure for the seed search phase. It makes a database index which contains compressed sequence data and locations of kmers. Database index is composed in three sections header, sequence data and offset data. The header section contains the range of sequence in database, format version etc., sequence data stores sequence data in compressed form and offset data contains lookup table and offset list.

The miBLAST[22] index stores only the sequence identifiers containing that kmer but not the offset. Result of Index. MegaBLAST show that its performance is better than miBLAST by at least 2.5 times in most of the cases. miBLAST is better in small queries.

D. *Spaced Seed Method: Improvement in Seeding*

A seed allowing internal mismatches is called spaced seed. The number of matches in the seed is its weight. SOAP [23] allows either a certain number of mismatches or one continuous gap for aligning a read onto the reference sequence. The best hit of each read which has minimal number of mismatches or smaller gap is reported. RMAP [24] is another program for ungapped mapping, which takes read qualities into account. SeqMap[25] and MAQ[26]extends the method to allow k-mismatches. SeqMap offers more flexibility in the mapping. It allows up to five mixed substitutions and inserted/deleted nucleotides in the mapping.

SOAP, RMAP and SeqMap are based on the pigeon hole principle the idea is to split each read in to several parts by requiring some of the parts instead of all of them to be perfectly matched in the mapping so the non-candidates can be filtered out very quickly and the sequence combined from the matched part can be used as a key to index all the candidates. A hash table is an effective and efficient data structure to implement this task.

Pattern hunter[27]introduces a novel seeding scheme and hit processing method it suggest to use nonconsecutive k letters as seed , for example for weight six model 1110111, zero shows don't care position then ACTGACT versus ACTAACT is a seed match. Result shows that this little change increases sensitivity and speed over BLAST.

III. CONCLUSION

Despite its long history, research in sequence alignment continues to flourish. Sequence alignment in modern computational biology form the basis of many bioinformatics studies, and advances in alignment methodology can confer sweeping benefits in a wide variety of application domains. Although many of these approaches rely on the same basic principles, the details of the implementations can have tremendous effects on the performance, both in terms of accuracy and speed. Although dynamic programming produces optimal alignment but it's not desirable for more than two sequences due to high processing time. Blast gave a solution of this problem using heuristic approach that approximate Smith-waterman algorithm .PSI-BLAST, MegaBLAST, miBLAST are different variants of BLAST algorithms gives better result in given conditions. Experiments in seeding method have been done in pattern hunter that shows better result. Preprocessing the database is an idea incorporated in megaBLAST, miBLAST ,SSAHA and BLAT, mostly works in small queries.

REFERENCES

[1] Burr Settles "Sequence Alignment" www.cs.wisc.edu/bsettles/ibs08.

[2] Thomas h cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford stein "Introduction to algorithms" Second Edition.

[3] Navarro G, " A guided tour to approximate string matching." *ACM Comput Surv* 2001;33:31–88.

[4] Eugene V Koonin and Michel Y. Calperin "Sequence- Evolution-Function*" Boston Kluwer Academic*-2003, ISBN-10 :40207-274-0.

[5] Needleman SB, Wunsch CD (March 1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins",*J.Mol. Biol*., vol. 48, pp.443-453.

[6] Smith, T. F. & Waterman, M. S. (1981), "Identification of common molecular subsequences", *J. Mol. Biol.*, vol. 147, pp 195-197.

[7] Higgins, D.G. & Sharp P.M. (1988), "CLUSTAL: a package for performing multiple sequence alignment on a microcomputer", *Gene*, vol. 73,pp237-44.

[8] Higgins, D.G et al. (1992), “ClustalV—improved software for multiple sequence alignment” *Comput. Appl. Biosci*., vol. 8, pp. 189-91.

[9] Higgins, D.G et al. (1994), “ClustalW—improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice”, *Nucl. Acid Res*., vol. 22, no. 22, pp 4673-80.

[10] Feng D. & Doolittle R. F (1987), “Progressive sequence alignment as a prerequisite to correct phylogenetic trees”), *J. Mol. Evol*., vol. 60, pp 351-360.

[11] Krogh, A. et al. (1994), “Hidden Markov models in computational biology: applications to protein modeling”, *J. Mol. Biol*., vol. 235, pp. 1501-1531.

[12] Robert C Edgar (2004), “ MUSCLE: multiple sequence alignment with high accuracy and high throughput” *Oxford Journals Science & Mathematics Nucleic Acids Research* Volume 32, Issue 5Pp. 1792-1797.

[13]Dayhoff, M. O.; Schwartz, R. M.; Orcutt, B. C. (1978), “A model of evolutionary change in proteins". *Atlas of Protein Sequence andStructure*, Vol. 5, Suppl. 3 National Biomedical Reserach Foundation, Washington D.C. U.S.A, pp 345-352.

[14] Henikoff, S; Henikoff, JG (1992), “Amino acid substitution matrices from protein blocks". *PNAS* , vol. 89, pp 10915-10919.

[15] Altschul, S.F. et al (1990), “Basic local alignment search tool”, *Journal of Molecular Biology*, vol. 215,pp 403-410.

[16] William J. Pearson and David J. Lipman, “Improved tools for biological sequence comparison”, *Proc. Natl. Acad. Sci. USA* Vol. 85, pp. 2444-2448, April 1988 Biochemistry

[17] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer1, Jinghui Zhang, heng Zhang, Webb Miller2 and David J. Lipman “Gapped BLAST and PSI-BLAST: A new generation of protein database search programs”, *Nucleic Acids Research*, 1997, Vol. 25, No. 17 3389–3402.

[18] Tatiana A. Tatusova , Thomas L. Madden, “ BLAST 2 SEQUENCES, a new tool for comparing protein and nucleotide sequences”, *FEMS Microbiology Letters* 174 (1999) 247250.

[19] W. James Kent, “BLAT-The BLAST like alignment Tool”, *Genome. Res.* 2002 Apr, 12(4):656-669.

[20] Zemin Ning, Anthony J. Cox, and James C. Mullikin “SSAHA: A Fast Search Method for Large DNA Databases”, *Genome Res*. 2001 Oct; 11(10): 1725–1729.

[21] Aleksandr Morgulis, George Coulouris, Yan Raytselis, Thomas L. Madden, Richa Agarwala and Alejandro A. Schäffer, “Database indexing for production MegaBLAST searches”, B*ioinformatics*Vol. 24 no. 16 2008, pages 1757–1764 doi:10.1093/ btn322.

[22] You Jung Kim, Andrew Boyd, Brian D. Athey, and Jignesh M. Patel, “miBLAST: scalable evaluation of a batch of nucleotide sequence queries with BLAST”, *Nucleic Acids Res*. 2005; 33(13): 4335–4344.

[23] Li R, Li Y, Kristiansen K, etal, “SOAP: short oligonucleotide alignment program.”, *Bioinformatics* 2008;24:713–4.

[24] Smith AD, Chung WY, Hodges E, et al. “Updates to the RMAP short-read mapping software”, *Bioinformatics* 2009;25: 2841–2.

[25] Jiang H, Wong WH, “ SeqMap: mapping massive amount of oligonucleotides to the genome”, *Bioinformatics* 2008;24: 2395–6.

[26] . Li H, Ruan J, Durbin R., “ Mapping short DNA sequencing reads and calling variants using mapping quality scores”, *Genome Res* 2008;18:1851–8.

[27] Bin Ma, John Tromp and Ming Li, “Pattern Hunter: Faster and more sensitive Homology search”, *Bioinformatics* Vol 18 no. 3 2002 pages 440-445.