A Study On Pair-Wise Local Alignment Of Protein Sequence For Identifying The Structural Similarity

G. Pratyusha,
Department of Computer Science & Engineering, V.R.Siddhartha Engineering College(Autonomous)
Affiliated to JNTUK, Vijayawada , Andhra Pradesh, India.

S. Jayaprada
Department of Computer Science & Engineering, V.R.Siddhartha Engineering College(Autonomous)
Affiliated to JNTUK, Vijayawada , Andhra Pradesh, India.

Dr. S. Vasavi
Department of Computer Science & Engineering, V.R.Siddhartha Engineering College(Autonomous)
Affiliated to JNTUK, Vijayawada , Andhra Pradesh, India.

Abstract

Pair wise sequence alignment methods are used to find the best-matching pair wise local or global alignments of two query sequences. Protein sequence alignment is one of the crucial tasks of computational biology which forms the basis of many other tasks like protein structure prediction, protein function prediction and phylogenetic analysis. In this paper we made a study on Pair Wise Local alignment and consider:(1) what sorts of alignment should be considered (2) the scoring system used to rank alignments (3) the algorithm used to find optimal (or good) scoring alignments and scoring measurements such as Bayesian approach, Classical approach (4) the statistical methods used to evaluate the significance of an alignment score.

Keywords: *Pair wise local alignment, Score significant, Sequence Alignment, Statistical measures, optimal alignment.*

1. Introduction:

Sequence alignment is the procedure of comparing two (pair-wise alignment) or more (Multiple sequence alignment) sequences by searching for a series of individual characters or character patterns that are in the same order in the sequences. Two sequences are aligned by writing them across a page in two rows. Identical or similar characters are placed in the same column, and non identical characters can either be placed in the same column as a mismatch or opposite a gap in the other sequence. In an optimal alignment, non-identical characters and gaps are placed to bring as many identical or similar characters as possible into vertical register. Sequences that can be readily aligned in this manner are said to be similar. There are two types of sequence alignment: global alignment and local alignment. In the global alignment an attempt is made to align the entire sequence using as many characters as possible up to the both ends of each sequence. Sequences which are quite similar and approximately the same length are suitable candidates for global alignment. In the local alignment, the alignment stops at the ends of regions of identity or strong similarity, and a much higher priority is given to finding these local regions than to extending the alignment to include more neighboring amino acid pairs. Local alignments are more suitable for aligning sequences that are similar along some of their lengths but dissimilar in others, sequences that differ in length or sequences that share a conserved region or domain. As Protein sequence alignment is the crucial task in bioinformatics, so the protein alignment problem has been studied for many years, the recent studies have demonstrated that considerable progress in improving the accuracy or scalability of multiple and pair-wise alignment tools. The paper is organized as follows: section 2 discusses about related work, section 3 about Proposed Work, section 4 about Conclusions and Future Work.

2. Related Work:

1) ALAE guarantees correctness and accelerate BLAST for most of parameters [2]. The ALAE do the process as follows (1) In dynamic programming matrixes the entries calculating takes more time, especially for the long text & query. So, calculation of most of entries without affecting the accuracy of the alignment results should be avoided. This method analyzes the entries property in matrixes and also proposes a filtering techniques family to avoid the unnecessary & meaningless calculations. And also find that in each matrix there is much duplication & for reusing those duplicates a new algorithm is proposed. (2) For large bio-sequences the space requirement satisfaction is necessary especially for both query & text. This method consider in-memory algorithm for using the recent results on a compressed suffix array to make ALAE approach possible in memory. Even though

the idea is similar to BWT-SW, but it is fit for our filtering techniques and reusing approaches. (3) This method provides an upper bound value of calculated entries. This gives a mathematical analysis and also proves that this approach could provide a better time efficiency guarantee across the representative ranges of user specified schemes. The upper bound values using ALAE could vary from $4.50mn^{0.520}$ to $9.05mn^{0.896}$ for random DNA sequences and also vary from $8.28mn^{0.364}$ to $7.49mn^{0.723}$ for random proteins sequences. This shows experimental results on the real biosequence databases which includes DNAs and proteins to demonstrate that the space and time efficiency of ALAE approach. And also shown that ALAE makes an significant performance improvement on BWTSW for all the scoring schemes and thresholds. This also accelerates BLAST for most of scoring schemes and guarantees the correctness.

Advantages:

- ALAE utilize a family of filtering techniques to prune meaningless calculations and an algorithm for reusing score calculations.
- ALAE improve the time efficiency of the state-of-the-art exact BWT-SW approach significantly and accelerate BLAST for most of the scoring schemes.

Disadvantage:

- Further improvement is necessary for the performance of ALAE for all scoring schemes and exploits algorithms using external memory.

Possible Extension:

- The ALAE should concentrate only on the Protein sequences so as to improve in terms of space and time complexity.

2) To discover the structural and functional similarities between the biological sequences a local sequence alignment is widely used. The BLAST and FASTA are the faster heuristic methods and these are used to compare the single sequence to hundreds or a even thousands of sequences in genetic databases. This yields the pair wise alignments with a high sensitivity. And these heuristic methods are paradigm for narrowing down the good sequences. The Rigorous alignments are used for an in-depth comparison between the query sequence and the newly found sequence subset. This presents data-parallel algorithm for a an local sequence alignment algorithm which was based on Smith waterman algorithm using an associative model of computation known as ASC[4]. This work extends the local sequence alignment to the ASC model and also serves as the foundation for finding the top *k* local alignments which were similar to SIM and LALIGN in the FASTA package but with a much faster running time In this method there are $m + 1$ PEs(Processing elements) are used to calculate the scoring "matrix." as a substitute of the 2-D contiguous matrix which resides in the memory of a sequential computer, each PE holds the information of a row in the sequential Smith-Waterman matrix. The data dependencies beside the anti-diagonals impose a strict processing order within the PEs. No Di,j value may be computed prior to D entries with row and column indexes of $(i\text{-}1,j)$, $(i,j\text{-}1)$ and $(i\text{-}1,j\text{-}1)$, or its N, W, and NW neighbors. Active PEs compute in parallel the matrix values along the anti-diagonal in a wave front method.

Advantages:

- By applying this algorithm, it finds the best local alignment in O (m+ n) time using m + 1 processing elements. So, time complexity reduces.
- It has much faster running time.

Disadvantage:

- This approach is not that much efficient because, it can't allow the return of Multiple highly-conserved regions between two sequences in a single run.

Possible Extension:

- The Smith- Waterman implementations do not store the entire computed matrix.It must be extended to store the entire computed matrix in a single processing element.

3) In biological sequence data the Sequence similarity searching is one of the most demanding techniques for using the explosively growing quantity. The Sequence similarities which are obtained from a database search with a newly sequenced protein as a query that provides useful information for the detection of homologies and for discovering functional, structural and evolutionary information of the sequence. For searching the similarities there are two types of tools: global and local, which were differ in their terms of the strategy of the algorithms. Global similarity searches seek to maximize the similarity score along the whole sequences. Local similarity searches seek to maximize the score of some isolated regions. The protein sequences of Pair-wise alignment local similarity searches produce many false positives because of compositionally biased regions, also called low-complexity regions (LCRs), of amino acid residues. By masking and filtering such regions significantly improves the reliability of homology searches and, consequently, functional predictions.[5] The CARD algorithm was proposed to investigate the structural properties of LCRs in biological sequences and filter them and improve the quality of database searches. Now we present our algorithm, named CARD. The CARD procedure as fallows. Given a sequence *s*, the algorithm detects all LCRs in *s* and masks them in two phases. The algorithm first constructs a suffix tree *T* for *s* and for every internal node *v*, computes the position

list *Pv*. In the second phase, the algorithm, using the position lists, iteratively detects LCRs and masks them.

Advantage:

- The CARD algorithm has good running time.

Disadvantage:

- If we want to filter a large number of sequences and are concerned with the processing time, CARD is not the optimal choice.

Possible extension:

- This should be extended to filter a large number of sequences also.

4) In bioinformatics the biopolymer sequence comparison methods are the most commonly used tools. Heuristic method & Local dynamic programming method are the most significant advances in the analyses of biological sequence. These algorithms require an scoring matrix specification, gap penalties set ,only a single alignment return and an associated score that must be adjusted for the of the sequences lengths. The Smith–Waterman algorithm also yields a single alignment, which has the albeit optimal, and it is strongly affected by the scoring matrix choice and the gap penalties & these scores are dependent upon the aligned sequences lengths, which require a conversion of post-analysis. To overcome these shortcomings, a Bayesian algorithm was developed for the local sequence alignment (BALSA)[6]. This returns the both marginal optimal and joint alignments. From the posterior distribution and the posterior probabilities of gap penalties and scoring matrices the samples of these alignments are drawn. After that, all required sums can be completed by using a modified dynamic programming method & exact inferences on all the variables are available. The procedure of BLASA: The Bayesian model will capture the idea of local alignment, specifically aligning the related subsequences and then ignoring the unrelated sections of the sequence on both the ends. The summation of overall alignments needs to take into account of all alignments that will begin at any point and end at any point in the two sequences while adhere to the constraint that an alignment may not end before it has begun, the summation of overall alignments can be achieved through a recursive algorithm

Advantages:

- Bayesian statistics provides a means to relax the requirements and to achieve an automatic length adjustment.
- This also found that no adjustment for length was necessary as there was little relationship between the BLASA score and the sequence length.

Disadvantages:

- Local Bayesian alignment algorithm should be implemented to identify structural neighbors with little cost.
- Time complexity is more.

Possible Extension:

- The BLASA algorithm when applied to protein sequences of local alignment searching it should yields better results.

5) The Local alignment algorithms statistical properties with gaps are theoretically analyzed for correlated and uncorrelated random sequences. In the area of log-linear phase transition, the alignment statistics with gaps are shown to be characteristically differing from the gap less alignment. Uncorrelated sequences optimal scores will obey the certain robust scaling laws. The sequence homology is used to guide the observed selection of scoring parameters for the optimal detection of scoring parameters. The optimal detection similarity is possible to occur in a region close to the log side of the log linear transition. This in the local alignment with gaps in this it deecribes the local alignment statistical properties in the recursive relation. The Local alignment is necessary only for an subsequence of one sequence is correlated with a subsequence of another. [3]If the subsequences positions were known, the correlations will be detectable by global alignment. If positions were not known also local alignment is applied.

Advantages:

- The advantage of local alignment is by cutting of the length of aligned segment, it restrictions the background roughness to a finite value such that the correlation peak can still be detected.
- This is a very efficient way to optimize the scoring parameters empirically and it may be useful in the alignment of a vast number of a weakly correlated sequences.

Disadvantage:

- This takes more time and occupies more space.

Possible Extension:

- Instead of using the smith waterman algorithm for local alignment, the BLASA can be used to get better results.

3. Proposed Work:

Figure 1 presents out proposed method for finding protein sequence similarity Initially we choose the sequences. If the sequences are protein sequences (20 amino acids sequences), we perform the local alignment by BLAST and FASTA[9,10] and also by Smith Waterman algorithm, to overcome the inconsistencies and drawbacks in this algorithm we use Data parallel algorithm and BLASA[3,5]. Then we check whether the alignment is of high quality, if it is high quality examine the sequences for

presence of repeats or low-complexity sequences or regions by using CARD[4] method. Otherwise alter parameters by using ALAE [1] method to speed up BLAST, the parameters such as scoring matrix, gap penalties, repeat alignment etc. After improvement of alignment then sequences are examined for the presence of repeats or low complexity sequences. Perform the statistical tests such as BLOSUM MATRIX dynamic programming [9,10] for alignment score. Bayesian or classical approaches can be used to verify whether the alignment is score significant. If the alignment is not improved and the score is not significant then the sequences are detectably similar. If the alignment is score significant, we will give ranking by deriving the score parameters from probabilities by using Day hoffs Pam matrix .After that we say the sequences are significantly similar or not.

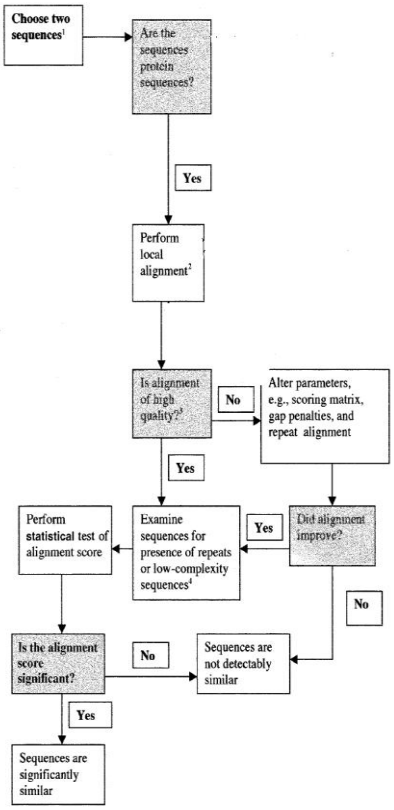

Figure 1: Proposed method for finding protein sequence similarity

Figure 2 presents sample input and output for our proposed method. It Takes the protein names PSEN1, APOE and takes the identifier names of the proteins and then align these sequences and find the sequence similarity of the protein sequences. This figure shows that between two proteins there is 0% sequence similarity.

Sample input: P49768**,** P02649

Sample output:

1 MTELPAPLSYFQNAQMSEDNHLSNTVRSQNDNR ERQEHNDRRSLGHPEPLSNGRPQGNSR 60 P49768 PSN1_HUMAN

1 - - - -- -- - -- -- --- - --- - - -- -- - --- -- ---- -- -- -- --- -- ---- ---

- -------- - - - - - - - ---- - - - - - - - - - - -- - -

0 P02649 APOE_HUMAN

Figure 2: sample input and output for our proposed method

4. Conclusions And Future Work:

Pair wise sequence alignment methods are used to find the best-matching pair wise local or global alignments of two query sequences. Protein sequence alignment is one of the crucial tasks of computational biology which forms the basis of many other tasks like protein structure prediction, protein function prediction and phylogenetic analysis. This paper studied various Pair Wise Local alignment approaches and figured the possible extensions to each of the existing approach. The paper also proposed a new method by using different techniques such as CARD, ALAE, BLASA/Smith waterman algorithm, BLAST, FASTA at various stages to find out the protein sequences similarity. This proves to be an efficient technique to overcome the drawbacks in previous methods. Even though it takes much time, in due course we identify techniques that can reduce time

References

[1] ParshantManohar, Shailendra Singh" Protein Sequence Alignment: A Review *",* March 2012. Vol (2), No (3), 141-145.

[2] Xiaochun Yang, Hong lei Liu, Bin Wang" ALAE: Accelerating Local Alignment with Affine Gap Exactly in Biosequence Databases", 2010.

[3]Terence Hwa, Michael Lasing "Optimal Detection Of Sequence Similarity By Local Algnment", 1996.

[4] Shannon I. Steinfadt, Michael Scherger, Johnnie W. Baker," A Local Sequence Alignment Algorithm Using an Associative Model Of Parallel Computation". 2006.

[5] Sung W. Shin∗ and Sam M. Kim' A new algorithm for detecting low-complexity regions in protein sequences' Vol. 21 no. 2 2005, pages 160–170.

[6] Bobbie-Jo M. Webb1,2, Jun S. Liu3 and Charles E. Lawrence1,4,* 'BALSA: Bayesian algorithm for local sequence alignment' 2002, Vol. 30, No. 5.

[7] I. Sadowski and W. R. Taylor "Evolutionary inaccuracy of pair wise structural alignments", Bioinformatics. Oxford journals.org 2012 page 1209 – 1215.

[8] Tobias Hamp, Fabian Buchwald and Stefan Kramer "Improving structure alignment-based of SCOP families using Vorlign Kernels", Bioinformatics. Oxford journals.org 2010, page 204-210.

[9] David W. Mount. "Bioinformatics Sequence and Genome Analysis "3rd chapter page 53-96,7th chapter page 283-307.The Cambridge University Press

[10] Durbin-Et-Al-Biological-Sequence-Analysis-CUP-2002-No-OCR., 2nd chapter page 12-45.The Cold Spring Hardor Labaratory Press.