International Journal of Systems Biology, ISSN: 0975–2900, Volume 1, Issue 2, 2009, pp-43-45

Multiple sequence alignment: a methodology for protein identification

Srinivasa Rao V.*[1], Das S. K.[2], Nageswara Rao K.[3] and Kusuma Kumari E.[4]
*[1,3]Computer Science and Engineering, PVP Siddhartha Institute oOf Technology, Vijayawada, India, akrgvsr@gmail.com
[2] Department of Computer Science, Berhampur University, Berhampur, Orissa, India
[4] Electronics and Communication Engineering, Nova College of Engineering, Jangareddygudem, India

Abstract- Multiple sequence alignment is the foundation of many important applications in bioinformatics that aim at detecting functionally important regions, predicting protein structures, building phylogenetic trees etc. Although the automatic construction of a multiple sequence alignment for a set of remotely related sequences cause a very challenging and error-prone task, many downstream analyses still rely heavily on the accuracy of the alignments.
Keywords- Molecular Biology, Global Alignment, ClustalW

Introduction

In molecular biology it is more common to consider more than two sequences for alignment. This is called multiple sequence alignment. The complexity of the problem grows fast when the number of sequences to be aligned grows. Sequence aligning is used for sequence comparisons in molecular biology. By aligning sequences we can see how similar the sequences are, and from that information draw some conclusions on how related the sequences are. High sequence similarity in a bimolecular sequence such as DNA, RNA and protein usually indicates similar function and or structure [1]. An alignment of two sequences(strings) *S*1 and *S*2 is obtained by inserting gaps(spaces), at the beginning, into or at the ends of *S*1 and *S*2, and then placing the two resulting sequences one above the other so that every character or gap in either sequence is opposite a unique character or a unique gap in the other sequence

*S*1: aa-t-gactagatta-ca

*S*2: -agtagac-ag-ttatc-

Types of alignment

There are several types of alignment. Here is a short description of some (but not all) of them.

Global alignment This is the general problem of aligning sequences. The whole sequence is considered when several sequences are aligned against each other. In biology, the problem was _rst solved by [1].

Local alignment With local alignment we only consider a smaller part of one or several sequences for alignment, i.e. we want to _nd subsequences that are aligned. An algorithm was proposed by Smith and Watermann [2] and local alignment is sometimes referred to as Smith-Watermann alignment.

k-**di_erence global alignment** This is a restricted version of global alignment. Given strings *S*1 and *S*2 and a _xed number k you would like to _nd the best global alignment of *S*1 and *S*2 containing at most k mismatches and spaces.

***k*-di_erence inexact matching** Given strings *P* and *T*, _nd all (if any) occurrences of *P* in *T* using at most k mismatches and spaces.

Sequence alignments include alignments between DNA sequences, protein sequences, or between EST sequence and un-spliced DNA sequence. These are called pairwise sequence alignments. There is also multiple sequence alignment, which aligns multiple sequences to find their common similarity. As mentioned earlier, sequence alignment algorithms can be generalized into global and local alignments. They are based on Needleman-Wunsch global alignment [1970] and Smith-Waterman local alignment [1981]. However, faster algorithms can be used to speed up the process.

Research approach and design

The present research aims at finding the proteins responsible for Insulin Resistance Syndrome in two phases. The first phase of the research attempts to identify the candidate proteins that cause Insulin Resistance Syndrome. The data pertaining to these proteins is extracted from the databases that are available online for free access. The functional protein sequences of these proteins in FASTA are to be extracted from (National Center for Biotechnology Information (NCBI), (http\\www.ncbi.nih.nlm.gov).

The second phase of the research analyzes the data by employing Multiple Sequence Alignment using ClustalW online tool. These alignments produce a Phylogenetic Tree along with the alignment scores. From the tree the results of the research are to be inferred in the last phase of the research.

Software and algorithms

The present research uses ClustalW, a web based progressive alignment tool for Multiple

Sequence Alignment (MSA). ClustalW adds sequences one by one to the existing alignment to build a new alignment because of its progressive nature. Progressive in this context means, it starts with using pair wise method to determine the most related sequences and then progressively adding less related sequences initial alignment. The order of the sequences to be added to the new alignment is indicated by a precomputed phylogenetic tree called a guide tree. The guide tree is constructed using the similarity of all possible pairs of sequences.

The ClustalW algorithm has three important Phases. They are
Phase I: All pairs of sequences are aligned separately to calculate a Distance Matrix based on the percentage of mismatches of each pair of sequences.
Phase II: The guide tree is constructed from the distance matrix using the Neighbour Joining algorithm.
Phase III: The sequences are progressively aligned following the guide tree.

The complexity involved in the process of ClustalW is
Complexity of phase–I: (Distance Matrix) is O(N2L2).
Complexity of phase–II: (Neighbour joining) is O(N4).
Complexity of phase–III: (Progressive Alignment) is O(N3+NL2)

The total complexity of the total process is O(N4+L2)
Where L is the length of the sequence and N is the number of sequences.
The alignment of two sequences at each step in the final progressive alignment is an important aspect of Multiple Sequence Alignment (MSA) in terms of Computer Processing and memory usage. In ClustalW, to align long sequences in a reasonable amount of memory space, the memory efficient Dynamic Programming concept is used (Myers and Miller). Dynamic Programming sacrifices some processing time but it makes very large alignments practical in very little memory. In spite of all these merits, there are some demerits in this strategy. One important disadvantage is that it does not allow different gap opening and extension penalties at each position. Another disadvantage is the propagation of errors from the initial alignments because of its progressive nature.

Data collection and analysis
Computational models have crucial roles in the analysis of complex biological systems. A single type of large-scale experimental analysis of molecular interactions or cellular states is insufficient for elucidating biological functions in specific biological contexts. Multiple, often heterogeneous sets of experimental data must be integrated and comprehensively analyzed to untangle the complexity of signaling, regulatory and metabolic pathways, together with their crosstalk. Toward this goal, systems biology advocates the iterative process of making models, obtaining experimental data, and reconciling discrepancies between model predictions and experimental outcomes. For such comprehensive studies, target species are often bacteria with known genome sequences and a small number of genes. In systems biology, computational analysis is expected to contribute in three ways. First, is the manifestation of each pathway's role within the context of others. The declarative representation of biological knowledge, in a visual or graphical format, enables us to view the model and has a major role in systems data analysis (Y. Tao et al, 2004). Second, computational analysis is used to estimate model components or to assign parameters that are experimentally indeterminable. Abstract, high-level models are refined to more physicochemical, low-level models through this step [9]. Third, computer simulations can predict a system's behavior under perturbation. Simulation results are used to generate new testable hypotheses and to improve models, and thus close the iterative cycle of systems biology. The data requirements of the present research are met with an access to the public databases available online. These databases are the repositories of a variety of data like protein sequence data. This research has used the functional protein sequences of the 27 candidate proteins causing Insulin Resistance Syndrome. The functional protein sequences in FASTA format for these proteins were collected from NCBI (National Center for Biotechnology Information, (http\\www.ncbi.nih.nlm.gov). The protein sequences of these molecules were compared against each other using multiple sequence alignment techniques, looking for similarity in the sequences and functionality. For this purpose ClustalW ver1.83 was used and their respective alignment scores were elucidated, which calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Based on these results, the scores table and phylogenetic tree that show the distance between the protein sequences was constructed.

Dynamic programming
By definition, dynamic programming is an algorithmic technique in which an optimization problem is solved by avoiding the sub-problem solutions and re-computing them. Dynamic programming was the brainchild of an American Mathematician, Richard Bellman, who described the way of solving problems where people need to find the best decisions one after another. In the forty-odd years since this development, the number of uses and applications of dynamic

programming has increased enormously. In fact, the word "Programming" in the name has nothing to do with writing computer programs. Mathematicians use the word to describe a set of rules which anyone can follow to solve a problem. They do not have to be written in a computer language.

Conclusion

The process of aligning multiple biological sequences is inherently difficult. This difficulty is compounded by the many competing choices for the alignment parameters, in choosing the alignment algorithms. In multiple sequence alignment, similar sequence motifs are identified and protein families are analyzed. The general method of multiple alignment has been to extend the pair wise alignment method into a simultaneous n-wise alignment by using a DP algorithm

References

[1] Needleman. S.B. and Wunch. C.D. (1970) *Journal of Molecular Biology*, 48, 443-453, 1970.

[2] Smith. T.F. and Waterman. M.S. (1981) *Journal of Molecular Biology*, 147, 195-197,

[3] Ishii M., Robert Y., Nakayama A., Kanai and Tomita M. (2004) *J. Biotechnol* .113, 281–294.

[4] Sharom J.R., Bellows D.S. nd Tyers M. (2004) *Curr Opin Chem Biol,* 8, 81–90.

[5] Eungdamrong N.J. and Iyengar R. (2004) *Biol Cell* 96,355–362.

[6] Vidal M., (2001) *Cell*, 104, 333–339.

[7] Selinger D.W., Wright M.A. and Church G.M. (2003) *Trends Biotechnol*, 21, 251–254.

[8] Tao Y., Liu Y., Friedman C. and Lussier Y.A. (2004) *Drug Discovery Today Biosilico*, 2, 237–245.

[9] Ideker T. and Lauffenburger D. (2003) *Trends Biotechnol,* 21, 255–262.

Biography

Dr. V. Srinivasa Rao received the degree M. Tech in Computer Science and Technology from Andhra University. He received the Ph.D. degree in Computer Science and Engineering from the Berhampur University. Currently, he is a Professor at PVP Siddhartha Institute of Technology, Vijayawada, A.P, India

Dr. S. K. Das presently working in the department of Computer Science, Berhampur University, Orissa, India.

Dr. K. Nageswara Rao received the M.Tech degree in Computer Science and Engineering from the Andhra University. He received the Ph.D. degree in Computer Science and Engineering from Andhra University. Currently, he is a Professor and Head at PVP Siddhartha Institute of Technology, Vijayawada, India.

Smt. E. Kusuma Kumari received the degree M. Tech in Electronics and Instrumentation Engineering from JNT University Hyderabad. Currently she is a Associate Professor in ECE Dept at Nova College of Engineering, Jangareddygudem, India. Her research interests are Antenna communications related and in Bioinformatics Instrumentation. Now she is doing Ph.D in Electronics and Communication Engineering.