Journal of Computational Intelligence in Bioinformatics.
ISSN: 0973-385X Volume 4 Number 1 (2011) pp. 171—176

http://www.ripublication.com/jcib.htm

Sequence Search and Alignment: Bioinformatics Tools for Bacterial Species

Shruti Sharma, Prateek Lal Shah and Pranjal Kumar

Amity Institute of Biotechnology, Amity University, Noida-125, U.P., India

Reviewers:
Pawan Kumar Maurya, E-mail: pawanbiochem@gmail.com
Hina Bansal Chaurasia, E-mail: hinabansal@gmail.com

Abstract

The interdisciplinary nature of bioinformatics will require the use of a variety of discipline-specific databases. Biomolecular sequences and structures of land, air and water species are determined rapidly and the data entries are unevenly distributed for different organisms. It frequently leads to the BLAST results of homologous search containing undesirable entries from organisms living in different environments. To reduce irrelevant search results, a separate database for comparative genomics is urgently required. A comprehensive bioinformatics tool set and an integrated database, named Bioinformatics tools for Bacterial species is proposed for comparative analyses among model species and various bacteria.

Novel matching techniques based on conserved motifs and/or secondary structure elements are designed for efficiently and effectively retrieving and aligning remote sequences through cross-species comparisons. It will be especially helpful when sequences under analysis possess low similarities and unresolved structural information. Bacterial genes are extensively used in Recombinant DNA technology and other branches of Biotechnology.

Comparisons of homologous sequences, gene finding, and prediction of gene expression are the most common techniques used on assembled datasets. In order to study how normal cellular activities are altered in different disease states, the biological data must be combined to form a comprehensive picture of these activities. Therefore, this software can be evolved such that the most pressing task now involves the analysis and interpretation of various types of data, including nucleotide and amino acid sequences, protein domains, and protein structures. In addition, the system provides core techniques of multiple sequence alignment, multiple second structure profile alignment and

iteratively refined multiple structural alignments for biodiversity analysis and verification bacterial biology.

Introduction

In today's fast advancing field of Biotechnology, new organisms are being discovered and their genomes and/or coding regions being studied at an alarming rate. The process of decoding of these genomes and sequence analysis is very important. The term "sequence analysis" in biology implies subjecting a DNA or peptide sequence to sequence alignment, sequence databases, repeated sequence searches, or other bioinformatics methods on a computer.

Since the development of methods of high-throughput production of gene and protein sequences during the 90s, the rate of addition of new sequences to the databases increases continuously. Such a collection of sequences does not, by itself, increase the scientist's understanding of the biology of organisms. However, comparing sequences with known functions with these new sequences is one way of understanding the biology of that organism from which the new sequence comes. Thus, sequence analysis can be used to assign function to genes and proteins by the study of the similarities between the compared sequences. Nowadays there are many tools and techniques that provide the sequence comparisons (sequence alignment) and analyze the alignment product to understand the biology.

Sequence analysis in molecular biology and bioinformatics is an automated, computer-based examination of characteristic fragments, e.g. of a DNA strand. It basically includes the following topics:

1. The comparison of sequences in order to find similarity and dissimilarity in compared sequences (sequence alignment)
2. Identification of gene-structures, reading frames, distributions of introns and exons and regulatory elements.
3. Finding and comparing point mutations or the single nucleotide polymorphism (SNP) in organism in order to get the genetic marker.
4. Revealing the evolution and genetic diversity of organisms.
5. Function annotation of genes.

In chemistry, sequence analysis comprises techniques used to do determine the sequence of a polymer formed of several monomers. In molecular biology and genetics, the same process is called simply "sequencing".

Objectives

Since the existing databases are all a conglomerate of various types of organisms living in different habitats and belonging to various taxa, the results for a particular search are often unrelated to the field of the search and this often results in consumption of a lot of time to search for the appropriate result.

Hence, the main aim of constructing this database is to minimize the spectrum of the search to a specific group of organisms which are linked by virtue of their

ancestry and/or habitat which will result in lesser deviations from the actual search. Bacteria have been chosen for this database since bacteria are one of the most researched organisms and maximum amount of laboratory work is being performed on them.

Methods and Materials

In bioinformatics, a sequence alignment 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. Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix. Gaps are inserted between the residues so that identical or similar characters are aligned in successive columns.

Very short or very similar sequences can be aligned by hand. However, most interesting problems require the alignment of lengthy, highly variable or extremely numerous sequences that cannot be aligned solely by human effort. Instead, human knowledge is applied in constructing algorithms to produce high-quality sequence alignments, and occasionally in adjusting the final results to reflect patterns that are difficult to represent algorithmically (especially in the case of nucleotide sequences).

The different methods of alignment used are as follows:

a. Pair-wise alignment

Pair-wise sequence alignment methods are used to find the best-matching piecewise (local) or global alignments of two query sequences. Pair-wise alignments can only be used between two sequences at a time, but they are efficient to calculate and are often used for methods that do not require extreme precision (such as searching a database for sequences with high similarity to a query). The three primary methods of producing pair-wise alignments are dot-matrix methods, dynamic programming, and word methods; however, multiple sequence alignment techniques can also align pairs of sequences. Although each method has its individual strengths and weaknesses, all three pair-wise methods have difficulty with highly repetitive sequences of low information content - especially where the number of repetitions differ in the two sequences to be aligned. One way of quantifying the utility of a given pair-wise alignment is the 'maximum unique match', or the longest subsequence that occurs in both query sequence. Longer MUM sequences typically reflect closer relatedness.

The primary method which we will be using are the word methods.

Word methods

Word methods, also known as *k*-tuple methods, are heuristic methods that are not guaranteed to find an optimal alignment solution, but are significantly more efficient than dynamic programming. These methods are especially useful in large-scale database searches where it is understood that a large proportion of the candidate sequences will have essentially no significant match with the query sequence. Word methods identify a series of short, non-overlapping subsequences ("words") in the

query sequence that are then matched to candidate database sequences. The relative positions of the word in the two sequences being compared are subtracted to obtain an offset; this will indicate a region of alignment if multiple distinct words produce the same offset. Only if this region is detected do these methods apply more sensitive alignment criteria; thus, many unnecessary comparisons with sequences of no appreciable similarity are eliminated.

BLAST and L-align are softwares which utilise the Word methods.

BLAST

In bioinformatics, Basic Local Alignment Search Tool, or BLAST, is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences. A BLAST search enables a researcher to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold. Different types of BLASTs are available according to the query sequences. For example, following the discovery of a previously unknown gene in a bacterium, a scientist will typically perform a BLAST search of the human genome to see if humans carry a similar gene; BLAST will identify sequences in the human genome that resemble the bacterial gene based on similarity of sequence.

BLAST is one of the most widely used bioinformatics programs, because it addresses a fundamental problem and the algorithm emphasizes speed over sensitivity. This emphasis on speed is vital to making the algorithm practical on the huge genome databases currently available, although subsequent algorithms can be even faster.

Input sequences are in FASTA format or Genbank format whereas BLAST output can be delivered in a variety of formats. These formats include HTML and plain text. For NCBI's web-page, the default format for output is HTML. When performing a BLAST on NCBI, the results are given in a graphical format showing the hits found, a table showing sequence identifiers for the hits with scoring related data, as well as alignments for the sequence of interest and the hits received with corresponding BLAST scores for these.

L align

L align is the ideal complement to BLAST. It is slower but more accurate, and it returns as many local alignments as specified by the user (the best scoring one first, followed by the second best scoring and so on.) In general, L align is very effective when it comes to analyzing complicated proteins which are full of repeats.

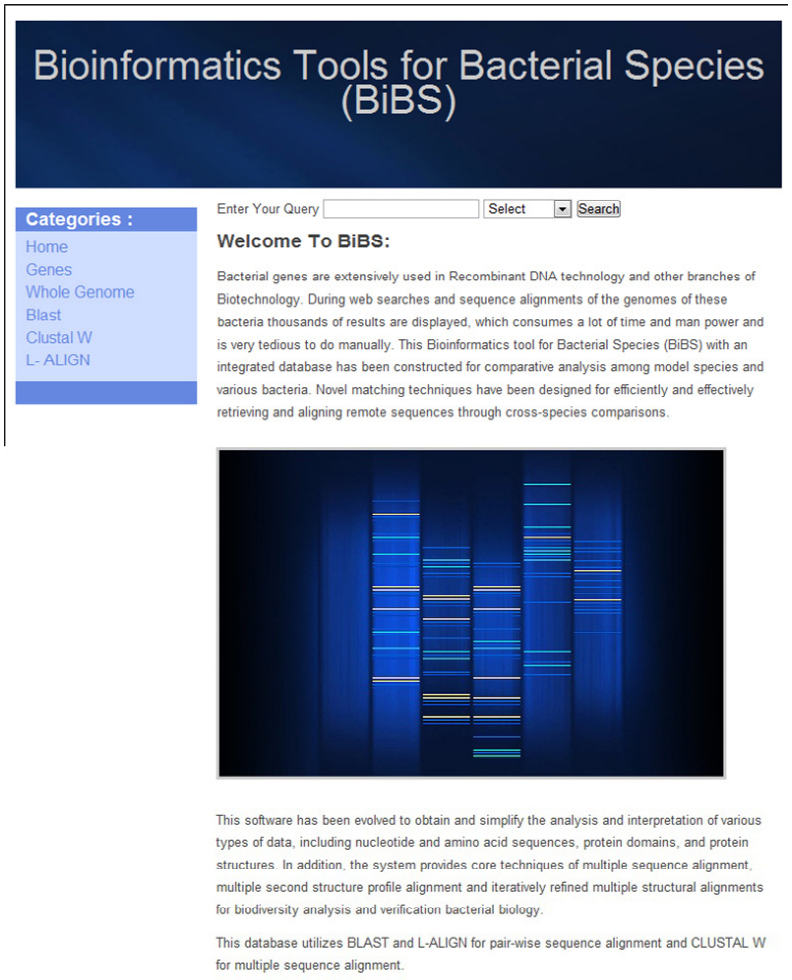
b. Multiple sequence alignment

Multiple sequence alignment is an extension of pair-wise alignment to incorporate more than two sequences at a time. Multiple alignment methods try to align all of the sequences in a given query set. Multiple alignments are often used in identifying conserved sequence regions across a group of sequences hypothesized to be evolutionarily related. Such conserved sequence motifs can be used in conjunction with structural and mechanistic information to locate the catalytic active sites of

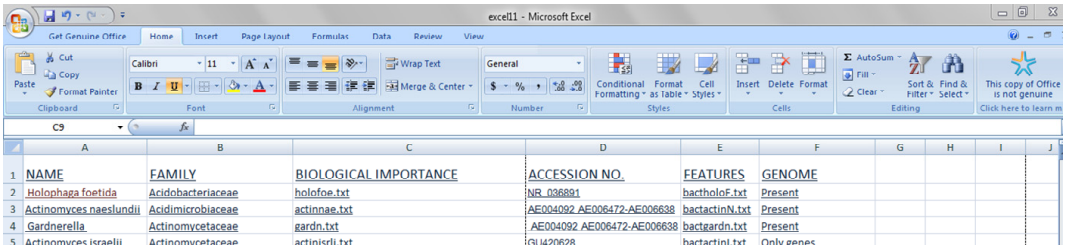
enzymes. Alignments are also used to aid in establishing evolutionary relationships by constructing phylogenetic trees. Multiple sequence alignments are computationally difficult to produce and most formulations of the problem lead to problems. Nevertheless, the utility of these alignments in bioinformatics has led to the development of a variety of methods suitable for aligning three or more sequences.

ClustalW is a software which utilizes this principle in sequence alignment.

Results and Conclusion



The above displayed homepage of the database has links to the Excel spreadsheets which contain the information about the Bacterial species. However, all the information cannot be displayed herewith due to the constraints of space. Hence, a single bacterial species has been selected to display the features and contents of the database.



The homepage also contains links to run BLAST, L align and CLUSTALW from their respective webpages. The data may be taken from the database and compared using these tools. Also, the database contains extensive information about the bacteria, like the habitat, physical characteristics, biological importance, the Family which it belongs to and the general characteristics of the Family.

References

[1] en.wikipedia.org/wiki/Sequence_alignment
[2] en.wikipedia.org/wiki/BLAST
[3] www.ncbi.nlm.nih.gov/
[4] GASSST: global alignment short sequence search tool. Rizk G, Lavenier D. Bioinformatics. 2010 Oct 15;26(20):2534-40. Epub 2010 Aug 24.PMID: 20739310 [PubMed - in process]
[5] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410. PubMed
[6] Edgar RC and S. Batzoglou. Multiple sequence alignment .,Curr Opin Struct Biol. 2006 Jun, 16(3):368-73. Epub 2006 May 5.
[7] Jung S, Jesudurai C, Staton M, Du Z, Ficklin S, Cho I, Abbott A, Tomkins J, Main D. BMC Bioinformatics. 2004 Sep 9; 5:130. Epub 2004 Sep 9
[8] L. Ruth, EMBO. Rep. 7, 17 (2006), A. R. Cossins and D. L. Crawford, Nat. Rev.Genet. 6, 324 (2005). G. J. Lieschke and P. D. Currie, Nat. Rev. Genet. 8, 353(2007).
[9] C. Pantzartzi et al., PLoS Comput. Biol. 6, e1000847 (2010).
[10] N. V. Vinithkumar, Adv. Bio.Tech. 12,(2006).
[11] R. Pushker et al, BMC Bioinformatics 6,222 (2005).
[12] P. Fariselli et al, Brief Bioinformatics 8, 78 (2007).
[13] D. M. Kristensen et al., BMC Bioinformatics 9, 17 (2008).
[14] I. Van Walle, I. Lasters and L. Wyns, Bioinformatics 21, 1267 (2005).
[15] I. Ilinkin, J. Ye and R. Janardan, BMC Bioinformatics 11, 71 (2010).
[16] Y. Ye and A. Godzik, Bioinformatics 21, 2362 (2005).