RAPID AND SENSITIVE DOT MATRIX METHODS FOR GENOME ANALYSIS

Dr.M.Manimekalai,
Director of MCA,
Shrimati Indira Gandhi College , Trichy.

S.Regha, Research Scholar,
Assistant Professor in computer Science
Shrimati Indira Gandhi College, Trichy.
Email Id:reghaaravinth@gmail.com
Ph:9443286543

ABSTRACT

Dot – matrix plots are widely used for similarity analysis of biological sequences. Many algorithms and computer software tools have been developed for this purpose. Two dot-matrix comparison methods have been developed for analysis of large sequences. The methods initially locate similarity regions between two sequences using a fast word search algorithm, followed with an explicit comparison on these regions. The methods produce high quality dot-matrix plots with low background noise. Space requirements are linear, so the algorithms can be used for comparison of genome size sequences.

Simple sequence repeats(SSR) or micro-satellites are becoming standard DNA markers for plant genome analysis and are being used as markers in marker assisted breeding.De novo generation of micro-satellite markers through laboratory-based screening of SSR-enriched genomic libraries is highly time consuming and expensive. A tandem repeat in DNA is two or more adjacent, approximate copies of a pattern of nucleotides. Tandem Repeats Finder is a program **"to locate and display tandem reports in DNA sequences".** In order to use the program, the user submits a sequence in **FASTA** format.

Keywords: Dot – Plots and Scoring Matrices, Importance of scoring matrices, Similarity versus Distance, Similarity Versus Homology, Global Versus Local Similarity, Simple Sequence Repeats(SSR's)

INTRODUCTION

Understanding the structure, function and evolution of genes is one of the main goals of genome sequencing projects.Classically,gene function has been investigated experimentally through the analysis of homologous sequences has proved to be a very efficient approach to study gene function (this approach has been coined 'comparative genomic' or 'phylogenomics'.For more than three billion years, genomes have continuously undergone mutations.Globally,advantageous mutations are very rare, and hence residues that are poorly conserved during evolution generally correspond to regions that are weakly constrained by selection.

Thus, studying mutation patterns through the analysis of homologous sequences is useful not only to study evolutionary relationships between sequences, but also to identify structural or functional constraints on sequences consists of trying to place residues(nucleotides or amino acids)in columns that derive from a common ancestral residue. The best alignment will be the one that represents the most likely be used to provide reliable alignments.

SCOPE OF THE PAPER

The scoring matrices solve the analysis of sequence comparision.The choice of the matrix(model used to build the matrix) can strongly influence the outcome of the analysis. The biological equivalence of a scoring matrix is an implicit particular theory or evolution. This method utilizes the dot plot for sequence comparision.It finds out the simple sequence repeat in nucleotide sequence. It also finds out the repeats present in the genome of soyabean.

Dot – Plots and Scoring Matrices

Dot plots are used to visually compare two sequences and detect of regions of close similarity between them.At every point where the two sequences are identical, a dot is placed. To detect more detect distant similarities, it may be better to use much larger window(i.e.20,30,or even 50 bases) and some suitable percentage of identities(perhaps 50%)

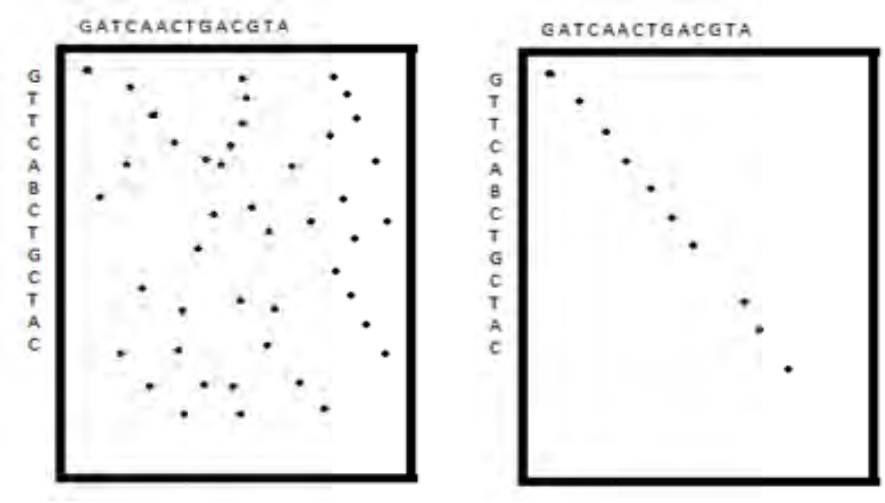
Scoring matrices: construction and analysis

We shall use the following notation for this discussion:

- A is the alphabet,
- A * is the set of all finite sequences of characters from,A

- a,b,c are variables denoting individual characters,
- s,t,u,v are variables denoting sequences from,A
- |s| denotes the length of the sequence s(in characters).

To address subsequence's and individual characters in a sequence s,we mark the boundaries between the characters of s by numbers, as shown here for s: **$ACCTGA_0A_1C_2C_3T_4G_5A_6$**



The simplest notion of distance is the Hamming distance: for two sequences of equal length, we just count the character positions in which they differ.

For example	AATA	TGCATG	GTA
Sequences	TAAT	AGCATA	ATA
Hamming distance(s,t)	3	2	1

This distance mensure is very useful in some cases, but it not flexible enough.First,the sequences may have different length and second, there is generally no fixed correspondence between their character positions.

Since the problem is symmetrical, an insertion in u can be seen as a deletion in v and vice versa. An alignment of two sequences u and v is an arrangement of u and v by position, where u and v can be padded with gap symbols to achieve the same length:

u: AGCACAC-A or AG-CACACA

v: A-CACACTA or ACACACT-A

Importance of scoring matrices

- All analysis involving sequence comparison are in general solved by using scoring matrices.
- The choice of matrix (model used to build the matrix) can strongly influence the outcome of the analysis.
- The biological equivalence of a scoring matrix is an implicit particular theory or evolution.

Similarity versus Distance

1. Elements of the matrices specify the weight to assign a given comparison by:
   - The cost of replacing one residue with another(*distance*); or
   - A measure of the *similarity* for the replacement.
2. Similarity is used for database searching.
3. Distance is more applicable for phylogenetic tree reconstruction.
4. Maximizing a similarity is fundamentally the same as minimizing a distance.

Similarity	Distance
Local Alignments	Evolution and Phylogeny
Suited for comparing proteins	Triangle inequality

Similarity Versus Homology

A database search is frequently, but incorrectly, referred to as homology searching. The term homology implies a common evolutionary relationship between two traits-whether they are DNA sequences or bristly patterns on a fly's nose .A very high level of similarity is a strong indication of homology.

Global Versus Local Similarity

Global algorithms are not sensitive for highly diverged sequences, a better (and faster)method focuses on short regions of "local" similarity. The three most widely used local similarity algorithms are:Smith-Waterman,BLAST and FASTA.The Smith-Waterman algorithm is a rigorous "dynamic programming" approach that does not make use of heuristic shortcuts.

FASTA(developed by Lipman and Pearson in 1985),considers exact matches between short sub strings, for a given parameter.FASTA uses the dynamic-programming algorithm to compute optimal alignments.

BLAST(developed by Altschul et al.in 1990),is another heuristic based on a similar idea. BLAST focuses on no-gap alignments of a certain, fixed length. Rather than requiring exact matches, BLAST uses a scoring function to measure similarity(rather than distance).

Simple Sequence Repeats(SSR's)

Simple Sequence Repeats DNA(SSR),also known as micro satellite DNA,is composed of non-coding, repetitive nucleotide sequences abundantly distributed throughout eukaryotic genomes. Motifs of SSR's are 2,3 or 4 nucleotides and these motifs are tandemly repeated many times in nucleotide sequences. The number repetitive motifs is very variable not only between species, but even between closely related individuals. Because SSR polymorphism often exhibits a codominant Mendelian inheritance.

"TANDEM REPEATS FINDER"(trf)

A tandem repeat in DNA is two or more adjacent, approximate copies of a pattern of nucleotides. Tandem Repeats Finder is a program(G.Benson,NucleicAcidResearch(1999)vol.27.)"to locate and display tandem repeats in DNA sequences".

Types of Tandem repeats finder:

The types are been programmed under the conditions, such as the type of user, the parameter like errors,gaps,copy numbers, Maximum and minimum period size, Flanking and Masked sequences, Alignment Parameters(match, mismatch and indels)and Data file etc.,

CONCLUSION

Searching for the best alignment consists of finding the one that represents the most likely evolutionary scenario(substitutions, insertion and deletion).Different alignment algorithms have been developed, but none of them is ideal. Because of time and memory requirements algorithms that guarantee to find the best alignment for a given evolutionary model can be used in practice only with a very limited number of short sequences. Therefore non-optimal algorithms based on heuristics have been proposed to gain speed and limit memory requirements.

These methods produce high quality dot-matrix plots with low background noise. Space requirements are linear, so the algorithms can be used for comparison of genome size sequences. Computing speed may be affected by highly repetitive sequence structures of eukaryote genomes. A dot-matrix plot of yeast genome(12 Mb)with both strands was generated in 80 s with a 1 GHz personal computer.

REFERENCES

[1] R.Durbon,S.Eddy,A.Krogh & G.Mitchison(1999)Biological Sequence Analysis.Cambridge University Press.
[2] D.W.Mount(2001)Bioinformatics:Sequence and Genome Analysis Cold Spring Harbor Laboratory Press.
[3] S.C.Rastogi,Namita Mandiratta,Parag Rastogi.Bioinformatics concepts,Skills&Applications.
[4] D.Higgins , W.Taylor Bioinformatics:Sequence,Structures and Data banks A Practical Approach.
[5] Kooning,Eugene V.and Galperin,Micheal Y.,Sequence Evolution Function.Computational Approaches in Comparative Genomics.
[6] Dan.E.Krane,Michael L.Raymer,Fundamental Concepts of Bioinformatics.
[7] T.K.Attwood & D.J.Parry-Smith,Introduction to Bioinformatics.
[8] Bishop M.J.(Ed.) Guide to human genome computing Second Edition Academic Press,London(1998).
[9] Peruski L.F.Jr.,Harwood Peruski.,The Internet and new molocular biology:tools for genomic and molecular research.American society for microbiology,Washington DC(1998).
[10] Suhai S.(ED.)Computational methods in genome research,Plenum Press.New York(1999).
[11] Smith D.W.(Ed.)Biocomputing.Informatics and genome projects.Academic Press.(1994).
[12] Waterman M.S.Introduction to computational biology:maps,sequences,and genomes Chapman and Hall,London(1995).
[13] Yap T.K.Frieder O.,Martino R.I.High performance computational methods for biological sequence analysis.Kluwer Acedemic Publisher,Dordrecht(1996).