**Course: Advanced Bioinformatics**

**Module title: MSA other Methods**

**Module no. : 40**

### In this module, discussion involves the other methods for multiple sequence alignment

#### Group approach

In the group approach, sequences are clustered into related groups. A consensus sequence is produced to make alignments between the groups. Examples of programs implementing the group approach are PIMA and MULTAL.

PIMA

MULTAL

#### Tree approach

The tree method uses the distance method of phylogenetic analysis to arrange the sequences. The two closest sequences are then aligned, and the consensus of these two is aligned with the next best sequence (or group of sequences) until an alignment is produced that includes all of the sequences. This approach is a popular approach used by PILEUP, CLUSTALW and ALIGN.

TREEALIGN is a program that uses the tree approach, but rearranges the tree as sequences are aligned to produce the tree by maximum parsimony of the tree.

### Localized Alignments in sequences

Just like with pairwise alignments, we may not be interested in the global alignment of multiple sequences, but rather only specific regions that are conserved. For instance, given regions of genomic DNA occurring upstream or before a certain gene, there might be sequences where transcription factors bind to the DNA so that the gene can be transcribed. Thus, if we are interested in determining if there is any signal in the regions upstream of a certain family of genes across several different organisms, it would be important to only find the conserved region, and not try to align all of the genomic DNA.

Localized alignments of protein sequences can yield information about conserved domains found in otherwise unrelated proteins.

Programs to detect localized alignments typically use one of the following three approaches: Profile Analysis; block analysis; pattern-searching or statistical methods

### Profile analysis

Profiles are found by first multiply aligning the sequences, determining which regions are the most highly conserved, and then creating a scoring matrix for the alignment of the highly conserved region. The profile is composed of columns, and may include matches, mismatches, insertions, and deletions found in a particular column.

Once a profile is created, it can be used to search a target sequence or database for possible matches to the profile using the profiles scores to evaluate the likelihood at each position.

The drawback of profiles is that the profile is only as representative as the variation in the sequences used to construct it. Thus, there is a bias in the profile towards the training data.

For each position in a profile, there is a column for each amino acid, plus a column for an unknown amino acid (z), and a column for gap opening and gap extension. There is a row for each position in the multiple alignment.

Block Analysis

Expectation-Maximization

Gibbs Sampling

Hidden Markov Models

Position Specific Scoring Matrix

Sequence Logos

Profile: Scores for substitutions and gaps in each column

Blocks: ungapped aligned regions

Alignments based on locally conserved patterns found in the same order in the sequences (synteny)

Use of statistical methods and probabilistic models of the sequences

Multiple sequence alignments yield information into the evolutionary history of the sequences – sequences that are most similar are likely to be recently derived from a common ancestor sequence

If the sequences in a multiple alignment have quite a bit of variation then it is difficult to create a multiple sequence alignment due to the different combinations of substitutions, insertions, and deletions that can be used

Local Alignment of proteins

ECSQ

SNSG

SWKN

SCSN

Profiles and Position-Specific scoring matrices

Motif-Based Approaches

Gibbs Sampling Algorithm

Describe this – hand out papers

MEME

Meta-MEME

Hidden Markov Models

Scoring Multiple Alignments

Programs for multiple sequence alignment

Progressive Alignment programs:

CLUSTALW, CLUSTALX

MSA

PRALINE

Iterative Alignment Programs:

DIALIGN

MULTALIGN

## PRRP

SAGA

Local Alignment of proteins

Asset

BLOCkS

EMOTIF

Gibbs Sampler

HMMER

MACAW

MEME

SAM

MSA

ClustalX

ClustalW

Viewing Multiple Alignments

sequence logos

Once an alignment is made, they can be compared using Hidden Markov Models

IBM’s MUSCA

<http://cbcsrv.watson.ibm.com/Tmsa.html>

ClustalW

<http://www.ebi.ac.uk/clustalw/>

<http://clustalw.genome.ad.jp/>

DIALIGN

<http://bibiserv.techfak.uni-bielefeld.de/cgi-bin/dialign_submit>

Web Logo

<http://www.bio.cam.ac.uk/cgi-bin/seqlogo/logo.cgi>

DNA Sequence Data sets

Fasta File Format

GenBank File Format

ASN.1 File Format

XML File Format