MULTIPLE SEQUENCE ALIGNMENT

Course code: MIM 5531
Course Title: Bioinformatics and Biostatistics
Course Teacher: Jennifer Michellin Kiruba N

- A multiple alignment is the simultaneous alignment of three or more nucleic acid or amino acid sequences.
- The procedure involves the insertion of gaps in the sequences so as to maximize the overall similarity (Higgins and Sharp 1988).
- Multiple alignments are rarely used for their own sake but are usually created for another purpose – for instance, primer design or analysis of phylogeny.
- Users select a favorite program or program package and try to optimize program settings for that.
- A multiple sequence alignment is a collection of three or more protein (or nucleic acid) sequences that are partially or completely aligned.
- Homologous residues are aligned in columns across the length of the sequences.
- These aligned residues are homologous in an evolutionary sense: they are presumably derived from a common ancestor.
- The residues in each column are also presumed to be homologous in a structural sense: aligned residues tend to occupy corresponding positions in the three-dimensional structure of each aligned protein.

Table 9-1 Main Criteria for Building a Multiple Sequence Alignment

Criterion	*Meaning*
Structural similarity	Amino acids that play the same role in each structure are in the same column. Structure-superposition programs are the only ones that use this criterion.
Evolutionary similarity	Amino acids or nucleo tides related to the same amino acid (or nucleotide) in the common ancestor of all the sequences are put in the same column. No automatic program explicitly uses this criterion, but they all try to deliver an alignment that respects it.
Functional similarity	Amino acids or nucleotides with the same function are in the same column. No automatic program explicitly uses this criterion, but if the information is available, you can force some programs to respect it — or you can edit your alignment manually.
Sequence similarity	Amino acids in the same column are those that yield an alignment with maximum similarity. Most programs use sequence similarity because it is the easiest criterion. When the sequences are closely related, their structural, evolutionary, and functional similarities are equivalent to sequence similarity.

Table 9-2 Main Applications of Multiple Sequence Alignments

Application	*Procedure*
Extrapolation	A good multiple alignment can help convince you that an uncharacterized sequence is really a member of a protein family. Alignments that include Swiss-Prot sequences are the most informative. Use the ExPASyBLAST server (at `www.expasy.ch/tools/blast/`) to gather and align them.
Phylogenetic analysis	If you carefully choose the sequences you include in your multiple alignment, you can reconstruct the history of these proteins. Use the Pasteur Phylip server at `bioweb.pasteur.fr/seqanal/phylogeny/phylip-uk.html`.
Pattern identification	By discovering very conserved positions, you can identify a region that is characteristic of a function (in proteins or in nucleic-acid sequences). Use the logo server for that purpose: `www-lmmb.ncifcrf.gov/~toms/sequencelogo.html`.
Domain identification	It is possible to turn a multiple sequence alignment into a profile that describes a protein family or a protein domain (PSSM). You can use this profile to scan databases for new members of the family. Use NCBI-BLAST to produce and analyze PSSMs: `www.ncbi.nlm.nih.gov/blast/blastcgihelp.shtml#pssm`.

DNA regulatory elements	You can turn a DNA multiple alignment of a binding site into a weight matrix and scan other DNA sequences for potentially similar binding sites. Use the Gibbs sampler to identify these sites: `bayesweb.wadsworth.org/gibbs/gibbs.html`
Structure prediction	A good multiple alignment can give you an almost perfect prediction of your protein secondary structure for both proteins and RNA. Sometimes it can also help in the building of a 3-D model.
nsSNP analysis	Various gene alleles often have different amino-acid sequences. Multiple alignments can help you predict whether a Non-Synonymous Single-Nucleotide Polymorphism is likely to be harmful. See the SIFT site for more details: `blocks.fhcrc.org/sift/SIFT.html`.
PCR analysis	A good multiple alignment can help you identify the less-degenerated portions of a protein family, in order to fish out new members by PCR (polymerase chain reaction). If this is what you want to do, you can use the following site: `blocks.fhcrc.org/codehop.html`.

Aligned columns of amino acid residues characterize a multiple sequence alignment.

This alignment may be determined due to features of the amino acids, such as:

•• There are highly conserved residues such as cysteines that are involved in forming disulfide bridges.

•• There are conserved motifs such as a transmembrane span or an immunoglobulin domain. We encounter examples of protein domains and motifs (such as the PROSITE dictionary)

•• There are conserved features of the secondary structure of the proteins, such as residues that contribute to α helices, β sheets, or transitional domains.

•• There are regions that show consistent patterns of insertions or deletions.

We consider five algorithmic approaches:
(1)exact methods;
(2)progressive alignment (e.g., ClustalW);
(3)iterative approaches (e.g., PRALINE, IterAlign, MUSCLE);
(4) consistency-based methods (e.g., MAFFT, ProbCons); and
(5) structure-based methods that include information about one or more known three-dimensional protein structures to facilitate creation of a multiple sequence alignment (e.g., Expresso).
The programs we describe in categories (3) to (5) are often overlapping; for example, all rely on progressive alignment and some combine iterative and structure-based approaches.
All the programs offer tradeoffs in speed and accuracy.
MUSCLE and MAFFT are fastest, and are therefore most useful for aligning large numbers of sequences.
ProbCons and T-COFFEE, although slower, are more accurate in many applications.

- Progressive Sequence Alignment
- The most commonly used algorithms that produce multiple alignments are derived from the progressive alignment method.
- This was proposed by Fitch and Yasunobu (1975) and described by Hogeweg and Hesper (1984) who applied it to the alignment of 5S ribosomal RNA sequences.
- The method was popularized by Da-Fei Feng and Russell Doolittle (1987, 1990).
- It is called “progressive” because the strategy entails calculating pairwise sequence alignment scores between all the proteins (or nucleic acid sequences) being aligned, then beginning the alignment with the two closest sequences and progressively adding more sequences to the alignment.
- A benefit of this approach is that it permits the rapid alignment of hundreds or even thousands of sequences. A major limitation is that the final alignment depends on the order in which sequences are joined.
- It is therefore not guaranteed to provide the most accurate alignments.

Iterative Approaches

Iterative methods compute a suboptimal solution using a progressive alignment strategy, and then modify the alignment using dynamic programming or other methods until a solution converges.

An initial tree is divided and profiles from each side are re-aligned.-

These methods therefore create an intial alignment and then modify it to try to improve it, using some objective function to maximize a score

Consistency-Based approaches

In progressive alignments using the Feng–Doolittle approach, pairwise alignment scores are generated and used to build a tree.

Consistency-based methods adopt a different approach by using information about the multiple sequence alignment as it is being generated to guide the pairwise alignments.

We discuss two consistency-based multiple sequence alignment programs: ProbCons (Do et al., 2005) and T-COFFEE (Notredame et al., 2000).

MAFFT also includes an iterative refinement approach with consistency-based scores (Katoh et al., 2005), and the Ensembl program Pecan (discussed in “Analyzing Genomic DNA Alignments via Ensembl”) applies a consistency approach to aligning genomic DNA.

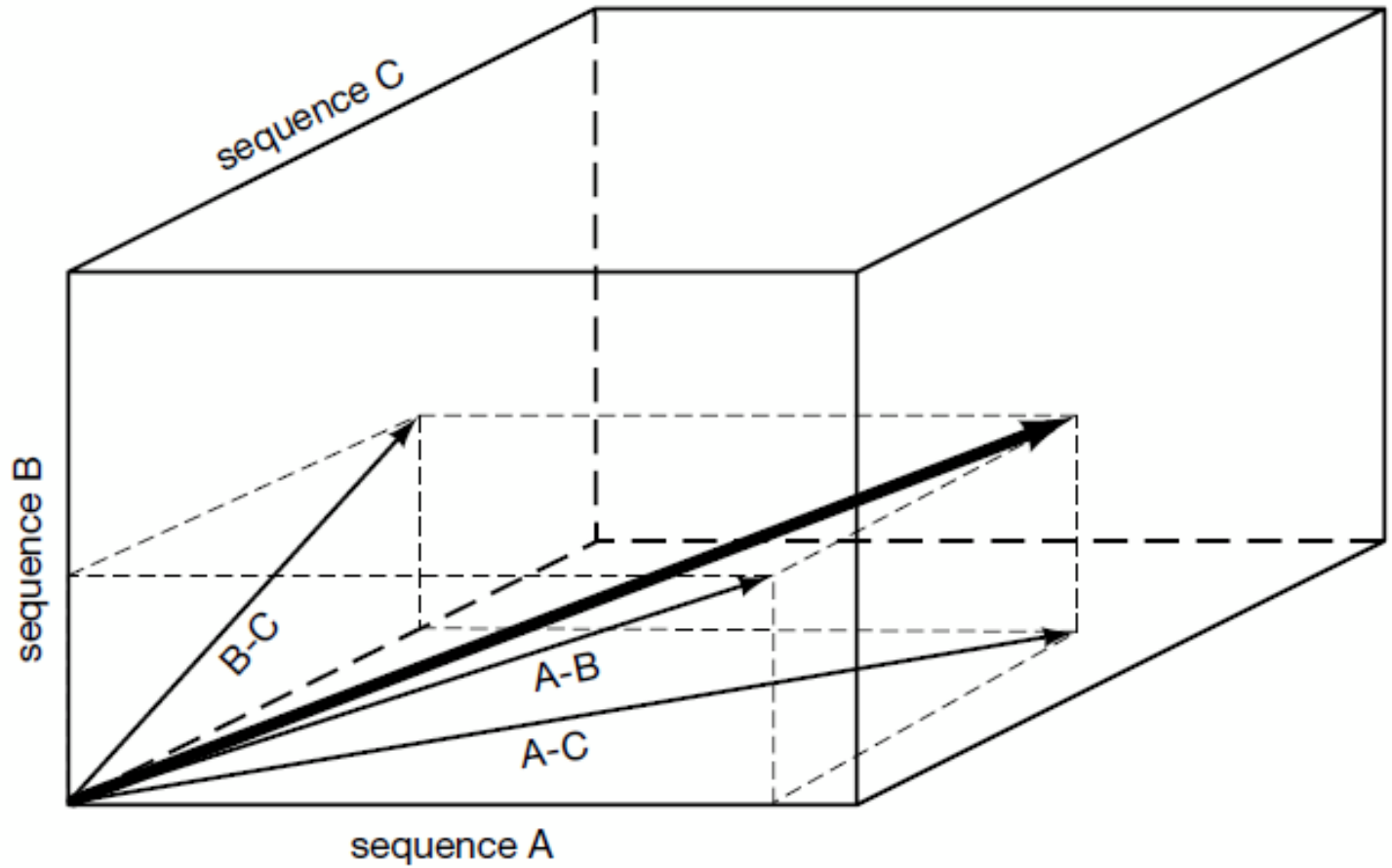

Figure 4.2. Alignment of three sequences by dynamic programming. Arrows on the surfaces of the cube indicate the direction for filling in the scoring matrix for pairs of sequences, A with B, etc., performed as previously described. The alignment of all three sequences requires filling in the lattice of the cube space with optimal alignment scores following the same algorithm. The best score at each interior position requires a consideration of all possible moves within the cube up to that point in the alignment. The trace-back matrix will align positions in all three sequences including gaps.