

Multiple Alignments for Structural, Functional, or Phylogenetic Analysis of Homologous Sequences

Higgins&Talyor Ch3 (Overview)

1. Why multiple sequence alignments?
2. The basic concepts
3. Multiple alignment methods
 - (a) Optimal methods
 - (b) Progressive global alignments
 - (c) Block-based global alignments
 - (d) Motif-based local alignments
 - (e) Aligning protein-coding DNA seqs
4. Visualizing and editing multiple alignments
5. Databases of multiple alignments

Why Multiple Alignments?

The multiple seq alignment is an effective tool for comparative analysis of homologous seqs. Studying mutation patterns thru the analysis of homologous seqs is useful not only to study evolutionary relationships among seqs, but also to identify structural or functional constraints on seqs. Table 1 on page 52 summarizes examples of usage of multiple alignments.

- Identification of functionally important sites
- Demonstration of homology between seqs
- Molecular phylogeny
- Search for weak but significant similarities in seq db
- Structural prediction
- Functional prediction
- Design of primers for PCR identification of related genes

The Basic Concepts for Multiple Alignment

1. Homology is inferred by sequence similarity. However, pairwise seq similarity may not detect distantly related seqs. In that case, multiple alignments can be used to infer homology among distantly related seqs.
2. Multiple copies of a given module may be repeated at different positions within a seq. In that case, we cannot align the whole seqs, thus the *local multiple alignment* is needed.
3. As different seqs come from different evolutionary distance, we may need to use different scoring schemes, ie, scoring matrices and gap penalties.
4. Different scoring matrices are used for modeling different evolutionary distances. Increasing indices in PAM matrices indicates larger evolutionary distances and increasing indices in BLOSUM matrices indicates shorter evolutionary distances.

Scoring Matrices

- In the CLUSTAL W programs, substitution matrices are automatically selected and varied at different alignment stages according to the divergence of the seqs to be aligned.
- Probabilities of substitutions also vary along seqs according to the local environment of AAs in the folded protein. Thus, several environment specific substitution matrices have been developed, e.g., α -helix or β -sheet. However, these matrices are rarely used.
- At DNA level, probabilities of substitution vary according to the bases. For example transitions, purines (A and G) and pyrimidines (c and T) are more frequent than transversions.
- Probabilities of nucleotide substitution also depend on neighboring bases. For example, in vertebrate, C in CG dinucleotides is hypermutable.

Insertions and Deletions

- The opening gap will get more penalty score than subsequent gaps. One model known as *affine* or linear gap penalty is $p = a + bL$. This model underestimates the prob of long indels and a better model would be $p = a + b \log(L)$. However, this model is rarely used because of computational complexity. *Why?*
- Like the substitution matrices, indels penalty scores should be varied according to the divergence of seqs.
- The indel prob also depends on the nature of seqs: proteins, structural RAN, non-coding DNA.
- Knowledge on the structure of proteins can be used to weight indels. For example, CLUSTAL W uses residue specific indel penalties and locally reduced indel penalties to encourage new gaps in potential loop regions rather than in regular secondary structure.

Searching for Homologous Seqs

Searching the seq dbs with a query seq is the first step to collect a set of homologous seqs. However, to guarantee a more comprehensive search, one may repeat the search procedure with several distantly related homologues identified in the first step.

The sensitivity of a seq similarity search may be improved by weighting sites according to their degree of conservation. Check back the family based search (Chapter 5) by Henikoff and Henikoff.

Multiple Alignment Methods

- Algorithms that gurantee to find the optimal alignment for a given scoring scheme. e.g, MSA
- Hueristic algorithms that are based on a progressive pairwise alignment approach. e.g, CLUSTAL W
- Hueristic algorithms that build a global alignment based on local alignment. e.g., DIALIGN
- Hueristic algorithms that build *local* alignments. e.g., MEME. e.g, MEME, hmalign, Gibbs, genetic aligorithms based multiple alignments.

Try all of them!

DIALIGN 2.1 is available on bio as `dialign2-1`, which improved the performance of DIALIGN2 with GABIOS-LIB. see <http://bibiserv.techfak.uni-bielefeld.d>

Optimal Methods for Global Alignments

- Scoring scheme for multiple aligned characters.
As the scoring matrices like BLOSUM and PAM provide scores for a pair of chars only, it is not trivial to score more than two aligned chars.
 - Sum-of-pairs scores
 - Sum-of-pairs scores with weighting
 - Evolutionary tree based scores
- Once we have a scoring scheme, we can compute an optimal alignment using a dynamic programming technique. However, computing an optimal alignment of n seqs of length l requires $O(2^n l^n)$ time and $O(l^n)$ space. MSA uses the branch and bound technique to reduce the computing time, but it still need too much time.

Progressive Alignment

Progressive alignment algorithms operate in three steps.

1. Compute alignment score for all pairs of seqs.
2. Build a *guide tree* that reflects the similarities between seqs using the pairwise alignment distances.
3. Align progressively the seqs following the guide tree.

CLSUTAL W used a UPGMA (Unweighted Pair Group method using Arithmetic average) to build the guide tree, but switched to the neighbor joining method.

Progressive Alignment

Continued

- While progressively aligning a pair of seqs or alignments, we need to align alignments. There are two choices for that. One is to consider the consensus seqs and the other is to use a profile for each alignment.
- Try the Figure 2 on page 62!
- The progressive alignment algorithm is a *greedy* algorithm, meaning that any mistake that appear during early alignments cannot be corrected as a seq info is added. See p63.
- The greedy nature can be improved using an iterative optimization strategy.
- A serious problem arises when a set of *unrelated* seqs are aligned, i.e., the alignment of unrelated seqs which we do not know in advance.

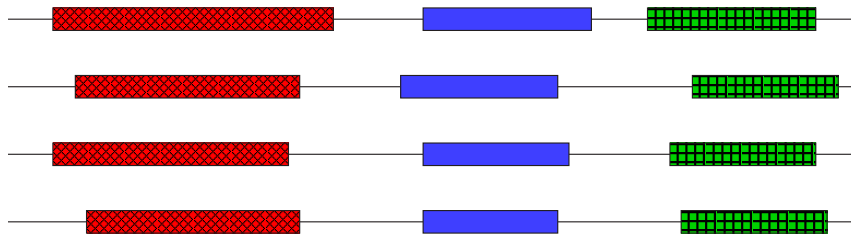
Block-based Global Alignments

- The seqs to be compared may share conserved blocks, separated by non-conserved regions. In that case, the result of optimal or progressive alignment methods will depend greatly on the choice of gap penalties. *Recal the gapper vs. blocker approaches by Henikoff.*
- Block-based alignment methods operate in two steps.
 - Gap-free blocks are identified. Blocks may be *uniform* (found in every seq) or not. We say blocks are *consistent* when blocks are in the same order, i.e., not crossed.
 - From a set of uniform blocks, an optimal consistent set of blocks is computed. DIALIGN is a heuristic algorithm to compute a good consistent set of blocks. For a set of seqs of the total length L , DIALIGN works in (L^2) .

Motif-based Local Alignments

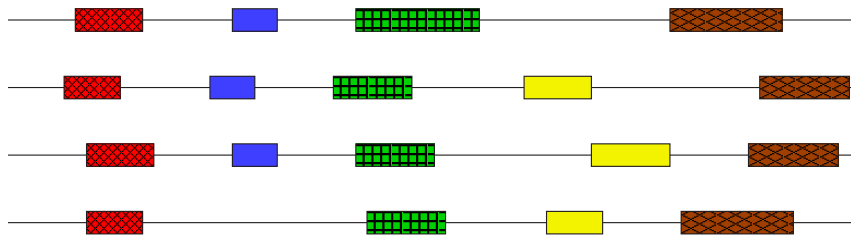
- Calculating local alignments consists of finding approximate repeated patterns in a set of seqs.
- PRALIGN compute consensus words for a given word length. For each possible word w of length k one may define the *neighborhood* of w as the set of k length. Then PRALIGN tries to compute the best score words.
- MACAW method combines pairwise comparisons in order to compute multiple local alignments. For each pair of seqs, all the diagonals with significant scores are marked and then the diagonals are merged into local alignments. $O(L^2)$ for a set of seqs of the total length L .
- Statistical methods like GIBBS or MEME.

Summary of Multiple Alignment Methods



Progressive alignment method

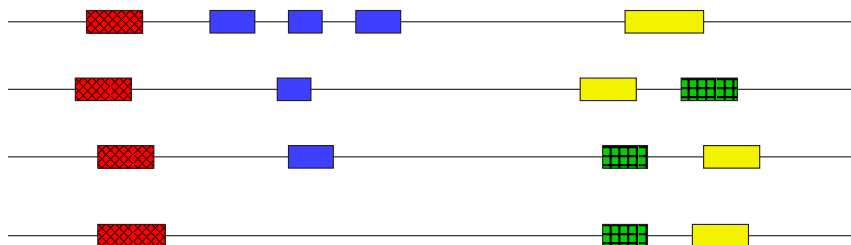
CLUSTALW



Block-based global alignment method

DIALIGN

ITERALIGN



Motif-based local alignment method

MEME

Visualizing and Editing Multiple Alignments

As computing multiple alignments is very difficult in terms of computational complexity and biological context, no existing method is reliable. Thus we often need to look at the alignments and modify them.

Experiment with the COG0551 family.

SEAVIEW, CINEMA, JALVIEW, CLUSTALX, PHYLIP, MPSA, DCSE, MEGA are among the tools you can visualize and edit multiple alignments.

Look at Table 9 for the databases of the multiple alignments.