

Protein Family-based Methods for Homology Detection & Analysis Higgins&Talyor Ch5 (Overview)

1. Terms used to describe relationships among proteins
2. Displaying protein relationships
3. Block-based Methods for multiple seq. alignment
4. Position-specific scoring matrices
5. Searching family databases with seq. queries
6. Searching with family-based queries

From Pairwise to Related Multiple Seqs

So far, we discussed pairwise seq alignment methods (SW algorithm, FASTA, BLAST) to detect homologous sequences in a database.

As seq data growing at an accelerated rate, related seqs together can be used to improve homology detection.

Even when functional info. is lacking for known members of a protein family, these members can be aligned and the alignments used in searches; for example, PSI-BLAST.

Terms Describing Relationships Among Proteins

1. **Motifs** refer to any regions of shared similarity among seqs. Shared similarity does not necessarily reflect shared ancestry. For example, helix-turn-helix DNA binding motifs.
2. **Domains** are regions from a structural perspective. It is difficult to identify the limit of a domain without seeing a structure. Furthermore, domains need not be contiguous along a seq.
3. **Modules** may correspond to separately folded domains, such as C_2H_2 zinc finger motif, and may be repeated within a seq. Unlike domains, modules are necessarily contiguous along a seq.

From Pairwise to Multiple Seqs

(Continued)

1. **Family** is a generic term to describe proteins (or genes) with sufficiently high seq. similarity that common ancestry may be inferred. A multidomain protein might have modules that belong to several different families.
2. **Super-family** and **sub-family** are defined in terms of what is meant by family, which has no precise definition.
3. Somtimes, proteins fold similarly, even though no seq. similarity among them is detected. Conclusions??
4. Seq. similarity may be evident, even though common ancestry is doubtful, as in the case of coiled-coil regions of proteins.

Alternative Approaches to Inferring Functions from Seq Alignment

There are two classes of comparison tools for seq. analysis of protein families, **blockers** and **gappers**.

Blockers, motif-based tools, consider aligned protein seqs to consist of nuggets of alignment info. (blocks) separated by regions that have no certain alignments.

Gappers try to find nuggets best found by determining where to place gaps in each seq. such that the blocks correctly align.

Blockers favor the BLAST family tools and BLOSUM matrices while gappers favor gap-based tools such as the dynamic programming alignment and HMMs.

Displaying Protein Relationships

- Depiction of pairwise seq alignments are not easily extended to multiple seq. alignments. For example, displaying an alignment of three seqs, $\{s_1, s_2, s_3\}$, requires displaying three different pairwise alignments.
- Traditional multiple sequence alignments does not solve the problem either. Especially, displaying a large number of seqs in this way is very complex.
- Three representations for displaying multiple protien relationships are patterns, logos, and trees.

Displaying Multiple Seqs

- **Patterns** are the simplest position-based representations of multiple alignments. An example Prosite pattern looks like
C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H.
- **Sequence logos** are graphical displays of multiple seq alignment consisting of ordered stacks of letters representing AAs at successive positions. (see Fig.1 on page 97)
Try it at http://blocks.fhcrc.org/about_logos.html.
- **Trees** are designed to show *differences* that distinguish between seqs or groups of seqs in contrast to position-based displays, such as patterns and logos, that show alignment info. on common.
Henokoff and Henikoff emphasize on the quality of multiple seq. alignment rather than on the choice of tree construction algorithms.
Try it at http://blocks.fhcrc.org/about_trees.html.

Block-based Methods for Multiple Seq. Alignment

This section describes block-based strategies for finding multiple-sequence alignment that are then used for db searching.

- **Pairwise alignment initiated methods** perform pairwise comparisons between seqs and then asking which high-scoring local regions are in common for most or all of the seqs in the group. MACAW combines overlapping segment pairs that exceed a threshold score. The extent of the block is limited by the requirement that each column have some minimum degree of homogeneity.
- **Pattern-initiated methods** exhaustively search for patterns that are common in all related seqs and part of true motifs. For example, the Motif program examines all seqs for the presence of spaced triplets of the form $aa_1 d_1 aa_2 d_2 aa_3$ for all combinations of d_1 and d_2 out to a reasonable maximum distance (about 20).
- **Iterative methods** iteratively search for motifs with seeds, blocks of specified width, using the Expectation-Maximization or Gibbs sampling.

Gibbs Sampling

Gibbs sampling applies when the conditional prob. $P(x_i|x_j, i \neq j)$ can be computed easily or when the variables X_i take on value from a small set.

Starting from $(x_1^t, x_2^t, \dots, x_n^t)$ at time t .

1. Select x_1^{t+1} according to $P(X_1|x_2^t, \dots, x_n^t)$.
2. Select x_2^{t+1} according to $P(X_2|x_1^{t+1}, x_2^t, \dots, x_n^t)$.
3. ...
4. Select x_n^{t+1} according to $P(X_n|x_1^{t+1}, x_2^{t+1}, \dots, x_{n-1}^{t+1})$.

Lawrence et. al applied Gibbs sample technique to motif finding *Science*, 262:208-214, Oct. 1993

Try Protocol 1 on page 100

1. Extract 10 random seqs from SwissProt.
2. Construct alignments using BlockMaker.
3. Perform MAST search.
4. Compare the names of submitted seqs to the significance hits from your MAST search.

The first question would be “do you find using MAST the original seqs when searched with the alignments?”

This experiment shows that blocks from a set of *randomly* chosen seqs can be used for seq searching! Are those blocks, however, functionally conserved regions?

In addition, how do we know if blocks are significant?

Position-specific Scoring Matrices (PSSMs)

To search with the multiple seq. alignments, we need a method to score matches with the alignments. A PSSM is the most widely used method for this purpose.

A PSSM consists of columns of weights for each amino acid derived from corresponding columns of a multiple seq alignment.

Profiles are PSSMs constructed using average score method.

Profile HMMs are PSSMs that are constructed using an iterative probabilistic algorithm for determination of position-specific gap penalties.

Two Major Issues for Constructing PSSMs

- **Sequence weights** are needed to reduce redundancy resulting from non-representative sampling of seqs.
Very similar seqs get low weights and more divergent seqs get higher weights. Weights can be measured in terms of distances in a tree, the average pairwise distances, or maximum discrimination between true positives and background.
- **PSSM column scores** involves more than two (a pair of) AAs, thus scoring matrices like BLOSUM and PAM cannot be used directly. One simple way is to compute the average of all possible pairwise scores.
However, the average score method is *insensitive* to the number of seqs in the multiple alignment. An effective way of dealing with this problem is to add *pseudocounts* to the observed counts of residue occurrences.

Searching Family DBs with Sequence Queries

Searches of family dbs can be more sensitive than searches of seq dbs because multiple alignment information is utilized.

- **Curated family dbs:** Prosite, Prints, and Pfam
 - Prosite is a db of curated patterns and also provides a profile PSSM.
 - Prints provides curated ‘fingerprints’ multiple alignments (ordered sets of blocks) that can be searched using pattern or PSSM method.
 - Pfam is a db of HMM PSSMs. Unlike Prosite and Prints, Pfam is automatically expanded using manually curated *seed* alignment: PfamA and PfamB.

Searching Family DBs with Sequence Queries *(Continued)*

- **Clustering dbs:** ProDom, DOMO, Protomap, and Prof_pat
An alternative to curation is to search a db against itself, then cluster similar seqs into families automatically.
 - ProDom is constructed from Swiss-Prot and it can be searched with multiple alignments or consensus seqs.
 - DOMO is based on Swiss-Prot and PIR. DOMO clusters tend to be longer and fewer in number than ProDom clusters.
 - Protomap is based on Swiss-Prot. It does not yield multiple seq alignments but generates a graphical tree-like view of the clustering.
 - Prof_pat extracts patterns from clustering Swiss-Prot/TrEMBL.
- **Derived family dbs:** Blocks and Proclass
Derived family dbs utilize protein family grouping provided by other resources.

Other Tools for Searching Family DBs

- *Identify* searches seqs versus pattern-based representations of individual blocks and fingerprints derived from the Blocks and Prints dbs.
- *ProfileScan* allows searching using generalized HMM-like profile PSSMs from Prosite, Pfam, and other sources.

None is perfect. it is worthwhile to try several of them for analysing a seq of interest. See Figure 3 on page 104.

Searching with Family-based Queries

As more seqs fall into families, opportunities increase for using family info. for identifying modules and new family members.

- **Searching with embedded queries:** A potential drawback to block-based approaches is that regions of uncertain alignment are not scored and the loss of this alignment info. can potentially reduce searching sensitivity. This problem has been addressed by implementation of a simple *embedding* strategy: a consensus id determined for a set of related seqs, the seq that is closest to the consensus is chosen, and blocks are embedded into that seq. COBBLER (COnsensus Biasing By Locally Embedding Residues) is a tool of this kind.
- **Searching with PSSMs:** The MULTiple Alignment Searching Tool (MAST) program searches block-based multiple alignments against the standard seq db sets.
- **Iterated PSSM searching:** PSI-BLAST constructs a PSSM after the first iteration. The PSI-BLAST PSSM employs position-based seq weights and pseudo-counts that are modelled upon AA substitution prob. *You have to use carefully the automated iterations!*

Multiple Alignment-based Searching of Protein Family DBs

Multiple alignments can be used for searching multiple alignment dbs.

LAMA is a program that searches ungapped blocks versus family dbs. In LAMA, PSSM columns are scored against one another by calculating a correlation coefficient, and a high scoring alignment is one in which seq-weighted distribution of residues is highly similar overall between aligned columns.