Knowledgebase Multiple Sequence Alignment

Ken Nguyen*, Yi Pan
Department of Computer Science, GSU

Abstract
Correctly identify and align motifs in multiple DNA, RNA, or Protein sequences help from unraveling the mystery of species evolution to drug design. We present our multiple sequence alignment (MSA) technique that relies on existing knowledge base to provide a more realistic and reliable sequence alignment. Our experiments show this approach yields more practical and reliable results than many other methods, even when the knowledgebase is partial or incomplete.

Biological Sequences
Sequence - A single, continuous molecule of nucleic acid or amino acid (residues).
The types of the of the acid and the formation of the acids may hold the functionality or reproduction signature of the sequence.
Motifs – a conserved sub-structure or group of residues that repeatedly appear in related the sequences of related species.

Sequence Alignment
A technique to arrange the symbols in a given sequences so that similar motifs and structures aligned.
Motifs and structures could be unknown
Must be operationally useful (fast)
Biological meaningful
Mathematically rigorous

Multiple Sequence Alignment Problems and Techniques
Given a set of sequences of at most length n,
Find an arrangement of the symbols in the sequences that maximize the number of matches with minimal number of indels, i.e. (·) [or()] insertions.

Existing Methods
Progressive:
Pair-wise aligned each sequence (1-to-all) by dynamic programming.
Build a phylogeny tree using pair-wise alignment scores
Assemble the sequences following the tree.

Probabilistic:
Profile all the symbols in the sequences
Apply different mathematical techniques such as Hidden Markov Model, Fast Fourier transform, or Simulated Annealing.

Heuristic:
Iteratively align the sequence to a meta-alignment.
Profile the alignment and realign until no better result found.

Graph:
A sequence is a connected graph of symbol nodes.
Merge nodes with same symbols across the sequences.

Optimality doesn’t imply correctness
Input sequences

DP – Global Alignment

 DP – Local Alignment

Output alignment should be:

Ignoring the existing large sequence knowledgeable
Not all sequences are new.
Search a new sequence against annotated sequence database for similarity is common.
Correctness and Usefulness:
Mathematically correct does not equate to biologically correct.
Optimal scored alignments may not be biologically meaningful.

Motivation
Utilize known biological information of the sequences (T-COFFEE and MACSIMS aligned)
Utilize the existing annotated sequence database to predict the biological feature of the sequences.

KB-MSA Method
Preprocessing the knowledgebase
The annotated sequence databases can be processed with the “approximate nearest neighbors” [1] technique to provide a fast querying time of O(k) log l (log*l)^2 where l is the max length of a sequence in database and |Q| is the length of the querying sequence.

When knowledgebase is unavailable:
When the sequence knowledge is unavailable, an artificial knowledgebase is built from the local alignment of the input sequence.

Experimental Results
The BALiBASE benchmark is used.
KB-MSA with complete sequence knowledge yields about 10% increase in average alignment score comparing to ClustalW and T-Coffee.
KB-MSA with 10% sequence knowledge yields about 5% increase in average alignment score comparing to ClustalW and T-Coffee.
KB-MSA without knowledgebase (KB-MSA 0%) yields similar alignment score as ClustalW and T-Coffee.

Existing Knowledge Bases:
As of 09-Feb-10: UniProtKB/TrEMBL contains 10376872 sequence entries comprising 334735583 amino acids.
As of 09-Feb-10: UniProtKB/Swiss-Prot contains 514799 sequence entries, comprising 181163771 amino acids abstracted from 186824 references.

Facts

References:

* - supported by Molecular Basis of Disease Program, GSU