Progressive multiple alignments of sequence triplets using structural information

Matthias Kruspe

Department of Computer Science, Bioinformatics Group
University of Leipzig

EMBIO Meeting
Vienna
21-24 May, 2006

Typical framework of progressive alignment algorithms (e.g. *CLUSTAL*)

1. determine distances by pairwise alignment of all sequences
2. calculate phylogenetic tree from the pairwise alignment scores
3. align sequences sequentially guided by tree

Problems

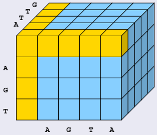
- not guaranteed to find optimal alignment
- ultimate alignment depends on initial pairwise alignments
- introduced gaps remain fixed during whole progressive alignment process
- loss of information when alignment is calculated

		agca
a−ga	ag−a	ag−a
agga	agga	agga

Idea

- try to increase information transfer from sequences to alignment
- try to increase quality of introduced gaps
- instead of comparing only two sequences in each step compare three sequences

Alignment of sequence triplets (3D alignment)

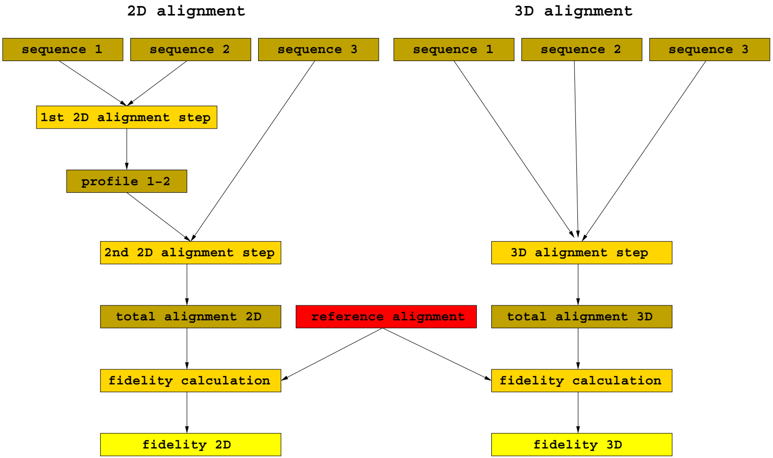


- apply standard *Needleman-Wunsch* dynamic programming algorithm with extensions to align three sequences
- use extended scoring scheme to handle all possible combinations of gap-open and -extension
- use sum of pairs cost model
- simple scoring function with fixed and position independent scoring terms (exchange costs and gap penalties)

Gotoh, O. 1986
Alignment of Three Biological Sequences with an Efficient Traceback Procedure
J. theor. Biol, 121

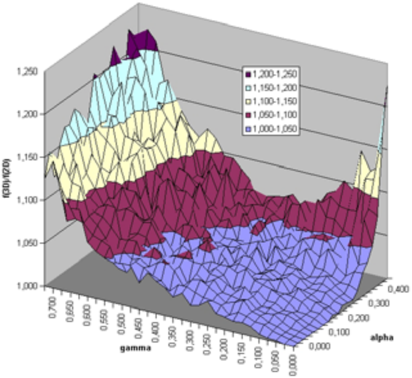
3D vs. 2D. alignment

- want to assess benefit of 3D alignment algorithm by aligning artificial sequence triplets with various distances



Results

- 500 sets of sequence triplets with average length of 200 nucleotides



- ratio of f_{3D} to f_{2D} (>1 means benefit of 3D alg.)
  - fidelity benefit increases significantly with increasing indel probability

Alignment order

- usually $n > 3$ sequences are given → must perform progressive sequence alignment
- problem: how to determine correct alignment order

Neighbor-Net

- distance based clustering method similar to *Neighbor Joining* to construct phylogenetic networks
- sequences are represented as nodes
- two steps: agglomeration and expansion
  - agglomeration: three nodes are fused to two new nodes
  - expansion: process is reversed, result is planar graph that represents re-construct phylogenetic network

Bryant, D., Moulton, V. (2004)
Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks *Mol. Biol. Evol.*, 21(2)

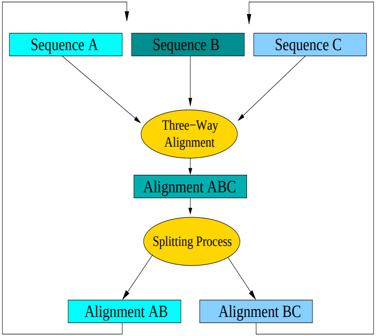
Alignment order

Getting alignment order out of phylogenetic network

- every node fusion in *Neighbor-Net* algorithm corresponds to a three-way alignment
- order of node fusion determines alignment order
- to keep framework consistent alignment must be divided into two alignments (possibility to remove mis-placed gaps)

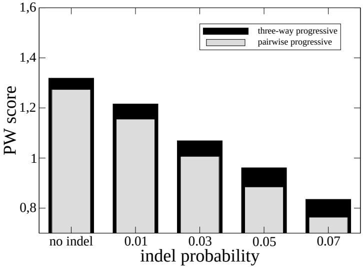
Setup

1. determine sequence distances by pair-wise alignment
2. build a phylogenetic network using *Neighbor-Net*
3. align sequences sequentially according to phylogenetic network
   - align three sequences in each alignment step
   - split alignment into two during progressive steps until final alignment is reached



PW scores

- generated a set of artificial sequences that evolve along a phylogenetic tree
- various indel probabilities to obtain different sequence distances
- aligned sequences using standard pairwise progressive alignment as well as triple alignment



- difference of PW score increases with increasing indel probability

Using structural information

- `Vienna RNA package (RNAfold)`
- given a RNA molecule compute for every base pair (i, j) probability P_{ij} that base i pairs with base j when molecule is folded (*McCaskill's* algorithm)
- define following three terms
  - $p_1(i) = \sum_{j=1}^{i-1} P_{ij}$ (base paired downstream)
  - $p_2(i) = \sum_{j=i+1}^{n} P_{ij}$ (base paired upstream)
  - $p_3(i) = 1 - p_1(i) - p_2(i)$ (base un-paired)

Score calculation

- given sequence x and y as well as base pair x_i and y_j
- final score S_{final} of a base pair is sum of weighted sequence score S_{seq} and weighted structure score S_{struct} with weighting factor $\psi \in [0, 1]$

$$S_{final}(x_i, y_j) = \psi \cdot S_{seq}(x_i, y_j) + (1 - \psi) \cdot S_{struct}(x_i, y_j)$$

Dataset

- Group II introns
- rRNA
- tRNA
- U5 spliceosomal RNA
- miRNA
- all sequences are obtained from Rfam database

Gardner, P.P., Wilm, A., Washietl, S. (2005)
A benchmark of multiple sequence alignment programs upon structural RNAs
Nucleic Acids Res, 28

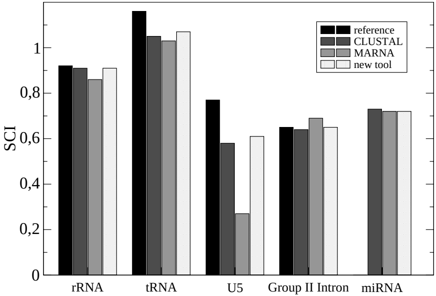
Structure conservation index (SCI)

- provides a measure of conserved secondary-structure information contained within alignment
- derivative of MFE calculated by consensus folding algorithm (`RNAalifold`)

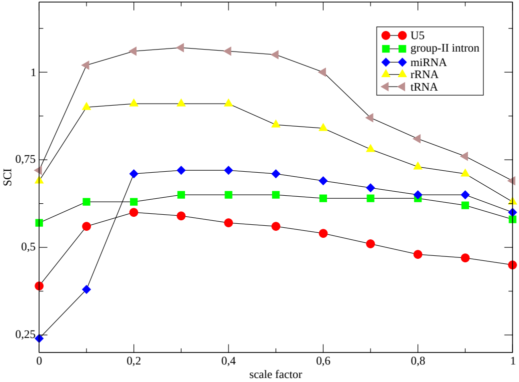
$$\mathrm{SCI}(A) = \frac{\mathrm{MFE}(A)}{\frac{1}{n}\sum_{i=1}^{n}\mathrm{MFE}(S_i)}$$

- SCI close to zero: no common RNA structure
  SCI close to one: perfectly conserved structure
  SCI larger one: conserved structure that is, in addition, supported by compensatory and/or consistent mutations preserving common structure
- measure of alignment quality independent from any reference alignment

Washietl, S., Hofacker, I., Stadler, P. (2005)
Fast and reliable prediction of noncoding RNAs *Proc. Natl Acad. Sci*, 102



	reference	ClustalW	MARNA	new tool
rRNA	0.92	0.91	0.86	0.91
tRNA	1.16	1.05	1.03	1.07
U5	0.77	0.58	0.27	0.61
g-II intron	0.65	0.64	0.69	0.65
miRNA	–	0.73	0.71	0.72



structure balanced sequence

- using both sequence and structure information increases SCI
- impact of ψ depends on
  - sequence identity (sequence with higher identity reach maximum SCI for higher values of ψ)
  - structure conservation

Wrong gap removal

- splitting alignment offers possibility to remove mis-placed gaps
- F is number of deleted gap columns, F_c is number of deleted gap columns that do not exists in final alignment

RNA family	F_c/F
Group II Intron	0.06
miRNA	0.14
rRNA	0.19
tRNA	0.12
U5	0.11

Thank you for your attention!