Multiple sequence alignment

November 2, 2017
Why do multiple alignment?

- Gain insight into evolutionary history
  - Can assess time of divergence by looking at the number of mutations needed to change one sequence into another
  - Create phylogenetic trees to determine which sequences are most closely related
- Highly conserved regions imply structural/functional information
- Group sequences into families -> functional information
- Sometimes the only way to align distant sequences
alignment reflects structure though sequences are not well conserved
Some history

1975: Fitch and Yasunobu suggested the use of a phylogenetic tree for MSA

1984: Hogeweg and Hesper propose integrated method to generate alignment and phylogenetic tree

make tree \[\leftrightarrow\] make alignment

(iterate)
sequences and trees

MDI---------------------NKLLAS-
MDRSAKIV-------------R-EEVAELLENR

MDI---------------------NKLLAS-
MFSS-------------DIDSLLNNK

Ogata et al ISME J 2011
remove shared gaps, calculate score

\[
\begin{align*}
  &\text{MDI}---------------------\text{NKLLAS} - \\
  &\text{MDRSAKIV}-----------------------\text{R-EEVAELLENR} \\
  &\text{MDI}--------\text{NKLLAS} - \\
  &\text{MDRSAKIVREEVAELLENR} \\
  &\text{MDI}---\text{NKLLAS} - \\
  &\text{MFSSDIDSLLNNK} \\
  &\text{MDI}---\text{NKLLAS} - \\
  &\text{MFSSDIDSLLNNK}
\end{align*}
\]

\[\text{score}=x\]

\[\text{score}=y\]
sequences and trees

Ogata et al. ISME J 2011
Some history

1987: Feng and Doolittle

“once a gap, always a gap”

postulated that the best trees will account for extant sequences with the smallest number of genetic events

Used distance measures instead of similarity measures

\[ D = -\ln(S_{\text{eff}}) = -\ln \left( \frac{S_{\text{real}} - S_{\text{rand}}}{S_{\text{ident}} - S_{\text{rand}}} \right) \]
Distance measures

\[ D = -\ln(S_{\text{eff}}) = -\ln\left(\frac{S_{\text{real}} - S_{\text{rand}}}{S_{\text{ident}} - S_{\text{rand}}}\right) \]

- \( S_{\text{real}} \) is the score for the alignment itself
- \( S_{\text{rand}} \) is the score obtained from randomizing the sequences
- \( S_{\text{ident}} \) is the average score of the two sequences aligned to themselves
More history

1988: Taylor

greedy MSA algorithm

1989: Lipman, Altschul, Kececioglu

dynamic programming for MSA

idea: each MSA imposes pairwise alignments on each set of sequences
Scoring multiple alignments

- Sum-of-pairs score — assumes statistical independence of the columns
- Can simply sum all pairwise substitution scores in the column
- not substantially affected by a single misaligned sequence

MSLQRTF
MTKQ−TF
−SLQR−F
Scoring multiple alignments

**Example: three-way alignment**

Score (per position) = \( \log \left( \frac{p_{abc}}{q_a q_b q_c} \right) = \log \left( \frac{p_{ab}}{q_a q_b} \right) + \log \left( \frac{p_{bc}}{q_b q_c} \right) + \log \left( \frac{p_{ac}}{q_a q_c} \right) \)

where \( p_{ab} \) is the frequency of co-occurrence of a and b if they’re related, 
\( q_a \) is the frequency of a in the population
True multiple sequence alignment

• Dynamic programming algorithms are too slow and in fact, cannot guarantee an optimal answer

• BUT it’s interesting to see how they work

• The DP recursion is too big to write out but if you have the optimal sequence up to a point, the next step is to make the optimal move (gap must be considered for every single sequence, in all combinations)
dynamic multiple sequence alignment

current alignment

character to add

MTL  L
M-K  S
MSL  T

MTL-
M-KS
MSLT

MTLL
MSLT

MTL-
M-K-
MSLT

etc
Time considerations

• For true MSA need to find sum-of-pair scores for all possible sequence pairs.

• Time complexity: $O(2^N L^N)$ where $N$ is the number of sequences and $L$ is the sequence length (this assumes that the sequences are roughly the same length)

• BUT, in fact, MSA is NP-complete.
Carillo-Lipman algorithm (implemented by Lipman, Altschul and Kececioglu)

Heuristic method, works well for a few sequences

Idea: the number of cells in the n-dimensional alignment lattice is the product of all of the sequence lengths! this is way too big, so we can’t examine every cell

Need to find a way to reduce the number of cells examined and still get the right answer . . .
Carillo-Lipman algorithm

The MSA path projects a pairwise path onto all sequence pairs.

This means that we can calculate an upper bound for the cost of this projection; this upper bound limits the points through which the projection can pass.
Carillo-Lipman algorithm

Lipman et al. wrote MSA, a multiple sequence alignment program using dynamic programming with heuristic modifications.

Can align 6-8 sequences of 200-300 residues in length; practical limit is closer to 4-5 sequences.
Progressive alignment methods

Feng and Doolittle’s idea: create pairwise alignments, use them to make an ad hoc tree, then use that tree to align the alignments into one big MSA.

Underlying assumption: closely related sequences will give the most robust alignments, so we should start with those (and the tree highlights which sequences those are).
Generic progressive alignment
original version released in 1994

Nice combination of algorithms and heuristics to create a very usable program. Still widely used today and is a benchmark for new programs (though it shouldn’t be)
1. Align all pairs of sequences separately to calculate distance matrix from divergence of each pair of sequences

2. Calculate guide tree from distance matrix

3. Align sequences progressively according to branching order in guide tree
CLUSTAL W
Progressive alignment

- “once a gap, always a gap”

example:
CLUSTAL Omega

Improvements:

- align sequences to randomly chosen subset first
- guide tree made by k-means clustering
- align along the guide tree until two alignments remain; use HMM to consolidate
CLUSTAL Omega

Improvements, continued:

• Dynamically vary gap penalties by position and in a residue-specific manner
  • ≥5 hydrophilic residues -> lower gap penalty by 1/3
  • Increase gap penalty less than 8 residues away from an existing gap
  • Lower gap penalty in an existing gap
• Substitution matrices varied at difference alignment stages according to the divergence of the sequences to be aligned
Iterative MSA

• CLUSTAL W is a progressive algorithm

• Many methods use format of CLUSTAL W but then rearrange the alignment iteratively to improve its score
MAFFT

- Iterative multiple sequence alignment based on fast Fourier transform
- first published in 2002
- cleverly restricts scope of dynamic programming used
- can incorporate other data (structure etc) while making the alignments
MAFFT

- first steps are progressive -- use kmer counting as first measure of similarity (k=6)
MAFFT

- then iterates to improve the trees & alignments
MAFFT

allows addition of unaligned sequences into existing multiple sequence alignment

has RNA-specific modes

can incorporate structural information into scoring
bigger issues . . .

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Basidiol2_f</td>
<td>YAAALG-DEVAAYAS-SDWRDNLCASALALALATNN-----SAYY</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Arabidop_f</td>
<td>YSDSLSSSVCPPYCSYSGBKDELWAGSWLLRATNN-----PYY</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cyanobac_f</td>
<td>YSDSIP-EVRNYNSWGSYGEDELAYGAALRSAVNSAGGDSAYL</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Firmicu1_f</td>
<td>YTAANG-----YYSSTS-FYDDLWACWLYMATND-----KSYL</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Actinoba_f</td>
<td>YSDCVP--AGAFYNSWSGYQDELVGWGAYWLYKATGD-----DSYL</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Firmicu2_f</td>
<td>YSDCIT-DAQQYNSWSGYKDELTWGAVWLYLATEE-----QYQL</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Proteoba_f</td>
<td>YSDCIT-DASSYKWSGYQDELVWSALWLYRATGEA-----ASYL</td>
</tr>
</tbody>
</table>

| Gaps in f   | XXXXX | X|X   | XXXXX |
| Differences | 000000000000 000 | 0 0 |
| Gaps in r   | XXXXX | X|X   |
| Basidiol1_r | YQTSVPSVADAYASS--GFQDELAIAALFISLAGNSS-----DAY |
| Basidiol2_r | YAAALGDEVAAYASS--DWRDNLCASALALALATNN-----AYY  |
| Arabidop_r  | YSDSLSSSVCPPYCSYSGBKDELWAGSWLLRATNN-----PYY  |
| Cyanobac_r  | YSDSIP-EVRNYNSWGSYGEDELAYGAALRSAVNSAGGDSAYL  |
| Firmicu1_r  | YTA-----ANGYYS-TSFYDDLWACWLYMATNDK-----SYL  |
| Actinoba_r  | YSDCVP--AGAFYNSWSGYQDELVGWGAYWLYKATGD-----SYL |
| Firmicu2_r  | YSDCITD-AQQYNSWSGYKDELTWGAVWLYLATEEQ-----QYQL |
| Proteoba_r  | YSDCITD-ASSYKWSGYQDELVWSALWLYRATGEA-----SYL  |

*TRENDS in Genetics*
bigger issues . . .
Mind the Gaps: Evidence of Bias in Estimates of Multiple Sequence Alignments

Tanya Golubchik,* Michael J. Wise,† Simon Easteal,‡ and Lars S. Jermiin*§¶

*School of Biological Sciences, University of Sydney, Sydney, Australia; †School of Biomolecular, Biomedical and Chemical Sciences, University of Western Australia, Perth, Australia; ‡John Curtin School of Medical Research, Australian National University, Canberra, Australia; §Sydney Bioinformatics, University of Sydney, Sydney, Australia; and ¶Centre for Mathematical Biology, University of Sydney, Sydney, Australia
PRANK: phylogeny-aware alignment

- length of alignment should grow with phylogenetic distance
- gaps are “free” after one gap is placed
- distinguishes insertions from deletions
- needs relatively large number of sequences for accurate phylogeny
- relatively slow
PRANK: phylogeny-aware alignment

- Length of alignment should grow with phylogenetic distance.
- Gaps are "free" after one gap is placed.
- Distinguishes insertions from deletions.
- Needs relatively large number of sequences for accurate phylogeny.
- Relatively slow.
Sequence analysis

PSAR-Align: improving multiple sequence alignment using probabilistic sampling

Jaebum Kim\textsuperscript{1,2,*} and Jian Ma\textsuperscript{3,4,*}

\textsuperscript{1}Department of Animal Biotechnology, \textsuperscript{2}UBITA Center for Biotechnology Research (CBRU), Konkuk University, Seoul 143-701, Korea, \textsuperscript{3}Department of Bioengineering and \textsuperscript{4}Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

Associate Editor: Ziv Bar-Joseph
Fig. 1. Overview of the PSAR-Align algorithm. Given input MSA, PSAR-Align first samples SAs (A and B). These SAs are analyzed by a pair of sequences at a time (C), and posterior probabilities of aligning two residues from two different sequences are computed (D). By using these probabilities, PSAR-Align finds the revised alignment based on the maximization technique of expected accuracy (E).
Assessment and refinement of eukaryotic gene structure prediction with gene-structure-aware multiple protein sequence alignment

Osamu Gotoh¹,²*, Mariko Morita¹ and David R Nelson³
Figure 8 The outline of the preparation and analyses of data. (A) The workflow of data processing. (B) Conceptual demonstration of the Refgs.pl strategy. The phase 1 and phase 2 introns are indicated by magenta and blue triangles, respectively.
still a VERY active field

• ~1-2 new algorithms published per month

• with more genomes, MSA is increasingly necessary and increasingly useful

• how to compare MSA algorithms?
curated databases of alignments

• BALiBASE
• SABmark
• PREFAB
• HOMSTRAD

but these aren’t as relevant now . . . new benchmarking methods published almost as often!
benchmarks

simulated sequences
consistency-based (compare to a bunch of MSA results)
structure-based
phylogeny-based

Who Watches the Watchmen? An Appraisal of Benchmarks for Multiple Sequence Alignment

Stefano Iantorno, Kevin Gori, Nick Goldman, Manuel Gil, and Christophe Dessimoz
benchmarking
Billions of “human-brain peta-flops” of computation are wasted daily playing games that do not contribute to advancing knowledge.
Whole-genome multiple alignment (UCSC 44-way Multiz MSA)

Extract dubious alignment region

Reinsertion into original alignment + Evaluation

Video game:
- Computers
- Tablets
- Cell phones

Database of interesting puzzles

http://phylo.cs.mcgill.ca

Figure 1. Phylo crowd-sourcing system for local improvement of multiple genome alignments.
doi:10.1371/journal.pone.0031362.g001
PHYL
-SOLVE A PUZZLE AND HELP GENETIC DISEASE RESEARCH-

Heart and muscles diseases
Cancers
Metabolic diseases
Digestive and respiratory system diseases
Blood and immune system diseases
Brain, nervous and sensory system diseases
Infectious diseases
Other diseases