Introduction to Bioinformatics

Prof. Dr. Nizamettin AYDIN

Multiple Sequence Alignment

naydin@yildiz.edu.tr
http://www3.yildiz.edu.tr/~naydin

Outline

- Multiple sequence alignment
- Introduction to MSA
- Methods of MSA
  - Progressive global alignment
  - Iterative methods
  - Alignments based on locally conserved patterns

Motivation…

- Similar genes can be conserved across species that perform similar or identical functions.
- Many genes are represented in highly conserved forms across organisms.
- By performing a simultaneous alignment of multiple sequences having similar or identical functions
  - we can gain information about
    - which regions have been subject to mutations over evolutionary time
    - which regions are evolutionarily conserved.
  - Such knowledge tells
    - which regions or domains of a gene are critical to its functionality.
- Sometimes genes that are similar in sequence can be mutated or rearranged to perform an altered function.
  - By looking at multiple alignments of such sequences,
    - we can tell which changes in the sequence have caused a change in the functionality.

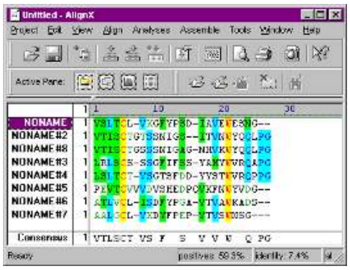
…Motivation…

- Multiple sequence alignment yields information
  - concerning the structure and function of proteins,
  - and can help lead to the discovery of important sequence domains or motifs with biological significance
    - while at the same time uncovering evolutionary relationships among genes.
- In multiple sequence alignment, the idea is
  - to take three or more sequences, and align them
  - so that the greatest number of similar characters are aligned in the same column of the alignment.

…Motivation

- The difficulty with multiple sequence alignment is that
  - now there are a number of different combinations of
    - matches,
    - insertions,
    - and deletions
  - that must be considered when looking at several different sequences.
- Methods to guarantee the highest scoring alignment are not feasible.
- Therefore, approximation methods are put to use in multiple sequence alignment.

Example MSA…



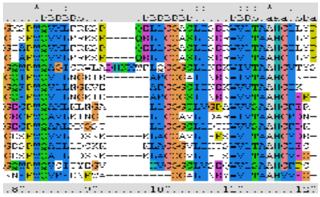
```
            1 1        10        20        30
NONAME      1 VSLTCL-VXGFYPSD-IAVEWESNG--
NONAME#2    1 VTISCTGTSSNIGS--ITVNWYQQLPG
NONAME#8    1 VTISCTGSSSNIGAG-NHVKWYQQLPG
NONAME#3    1 LRLSCS-SSGFIFSS-YAMYWVRQAPG
NONAME#4    1 LSLTCT-VSGTSFDD-YYSTWVRQPPG
NONAME#5    1 PEVTCVVVDVSHEDPQVKFNWYVDG--
NONAME#6    1 ATLVCL-ISDFYPGA-VTVAWKADS--
NONAME#7    1 AALGCL-VXDYFPEP-VTVSWNSG---

Consensus   1 VTLSCT VS F  S  V V W  Q PG
```

positives: 59.3% identity: 7.4%

- Example multiple alignment of 8 immunoglobulin sequences.
  - Immunoglobulin
    - any of a class of proteins present in the serum and cells of the immune system, which function as antibodies.

…Example MSA

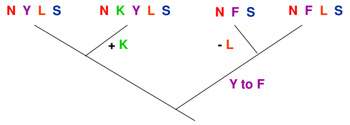


- Each row is a different protein sequence
- Each column is a different aligned position

Relationship of MSA to Phylogenetic analysis

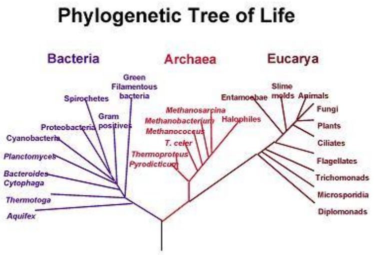
- Once the MSA has been found, the number or types of changes in the aligned sequences may be used for a phylogenetic analysis

```
seqA      N  -  F  L  S
seqB      N  -  F  -  S
seqC      N  K  Y  L  S
seqD      N  -  Y  L  S
```

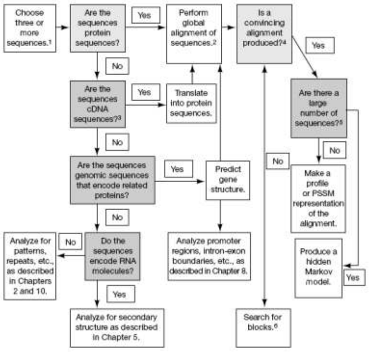


Hypothetical evolutionary tree that could have generated three sequence changes

Phylogenetic analysis



MSA method



Choose three or more sequences.[1]

Are the sequences protein sequences?

Yes

No

Perform global alignment of sequences.[2]

Is a convincing alignment produced?[4]

Yes

No

Are the sequences cDNA sequences?[3]

Yes

No

Translate into protein sequences.

Are there a large number of sequences?[5]

No

Yes

Are the sequences genomic sequences that encode related proteins?

Yes

No

Predict gene structure.

Make a profile or PSSM representation of the alignment.

Analyze for patterns, repeats, etc., as described in Chapters 2 and 10.

No

Do the sequences encode RNA molecules?

Yes

Analyze promoter regions, intron-exon boundaries, etc., as described in Chapter 8.

Produce a hidden Markov model.

Analyze for secondary structure as described in Chapter 5.

Search for blocks.[6]

Approaches to Multiple Sequence Alignment

- There are many approaches to multiple sequence alignment;
  - in the past decade many dozens of programs have been introduced
    - https://doi.org/10.1093/bib%2F6.1.6
- We will consider four approaches to multiple sequence alignment:
  - Exact methods (Dynamic Programming)
  - Progressive Alignment
  - Iterative Alignment
  - Statistical Modeling

Dynamic Programming Approach

- Exact methods of multiple alignment use dynamic programming
  - guaranteed to find optimal solutions.
- Dynamic programming with two sequences
  - Relatively easy to code
  - Guaranteed to obtain optimal alignment

- Can this be extended to multiple sequences?

DP with 3 Sequences…

- Consider the following amino acid sequences to align

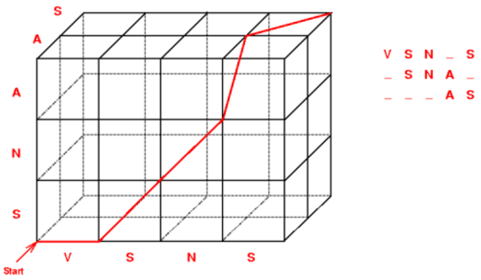
  VSNS, SNA, AS
- Instead of filling a two dimensional matrix as we did with two sequences,
  - we now fill a three dimensional space.
- Put one sequence per axis (x, y, z)
- Three dimensional structure results

…DP with 3 Sequences…

Possibilities:

- All three match;
- A & B match with gap in C
- A & C match with gap in B
- B & C match with gap in A
- A with gap in B & C
- B with gap in A & C
- C with gap in A & B

…DP with 3 Sequences



- Figure source:
  - http://www.techfak.uni-bielefeld.de/bcd/Curric/MulAli/node2.html#SECTION00020000000000000000

DP with *k* Sequences…

- Suppose the length of each sequence is n residues.
- If there are two such sequences,
  - then the number of comparisons needed to fill in the scoring matrix
    - is n^2,
      - since it is a two-dimensional matrix.
- The number of comparisons needed to fill in the scoring cube
  - when three sequences are aligned
    - is n^3
  - when four sequences are aligned,
    - is n^4

…DP with *k* Sequences

- Thus, as the number of sequences increases,
  - the number of comparisons needed increases exponentially, i.e. n^k
    - where n is the length of the sequences,
    - and k is the number of sequences.
- Thus, without any changes to the dynamic programming approach, this becomes impractical for even a small number of short sequences rather quickly.

Example

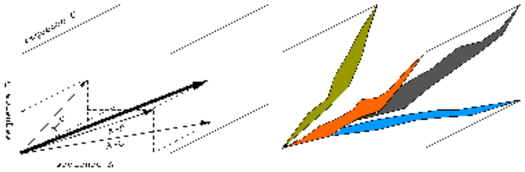
- 2 protein sequences length = 300, excluding gaps
  - number of comparisons by dynamic programming
    - $300^2 = 9 \times 10^4$
- 3 protein sequences length = 300, excluding gaps
  - number of comparisons by dynamic programming
    - $300^3 = 2.7 \times 10^7$
- Wang L, Jiang T (1994). "On the complexity of multiple sequence alignment". J Comput Biol. 1 (4): 337–348. CiteSeerX 10.1.1.408.894. doi:10.1089/cmb.1994.1.337

Reduction of space and time…

- Carrillo and Lipman:
  - – multiple sequence alignment space bounded by pairwise alignments
    - Carrillo H, Lipman DJ (1988). "The Multiple Sequence Alignment Problem in Biology". SIAM Journal on Applied Mathematics. 48 (5): 1073–1082. doi:10.1137/0148063
- Projections of these alignments lead to a bounded alignments
  - Konagurthu AS, Stuckey PJ (2006). "Optimal sum-of-pairs multiple sequence alignment using incremental Carrillo and Lipman bounds". J Comput Biol. Apr;13(3):668-85.

…Reduction of space and time…

- – Following figures show how the two dimensional search spaces can be projected into a three dimensional volume that can be searched



…Reduction of space and time…

- Idea for reduction of memory and computations:
  - – Multiple sequence alignment imposes an alignment on each of the pairs of sequences.
- Alignments found for each of the pairs of sequences can impose bounds on the location of the MSA within the cube (three sequences) or k-dimensional space (k sequences).
  - – Step 1
    - Find pairwise alignment for sequences.
  - – Step 2
    - Trial MSA produced by predicting a phylogenetic tree for the sequences
  - – Step 3
    - Sequences multiply aligned in the order of their relationship on the tree

…Reduction of space and time

- This is a heuristic alignment
- Therefore the alignment is not guaranteed to be optimal
- Alignment provides a limit to the volume within which optimal alignments are likely to be found

MSA

- MSA: Developed by Lipman, 1989
- Incorporates extended dynamic programming
- MSA calculates the multiple alignment score within the lattice by adding the scores of the corresponding pairwise alignments in the multiple sequence alignment.
- This measure is known as the sum of pairs (SP) measure.
- The optimal alignment is based on the best SP score.

Scoring of MSA's

- MSA uses Sum of Pairs (SP)
  - – Scores of pair-wise alignments in each column added together
  - – Columns can be weighted to reduce influence of closely related sequences
  - – Weight is determined by distance in phylogenetic tree

Sum of Pairs Method…

- The sum of pairs method scores all possible combinations of pairs of residues in a column of a multiple sequence alignment.
- For instance, consider the alignment

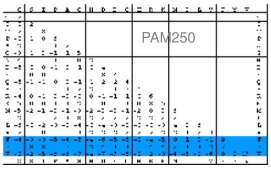
E	C	S	Q	(1)
S	N	S	G	(2)
S	W	K	N	(3)
S	C	S	N	(4)

- Since there are four sequences,
  - there will be six different alignments to consider for each column.
- The alignments, listed by the sequence number are listed as follows:

  1-2; 1-3; 1-4; 2-3; 2-4; 3-4

…Sum of Pairs Method…

E	C	S	Q	(1)
S	N	S	G	(2)
S	W	K	N	(3)
S	C	S	N	(4)



	Residues	Score	Residues	Score	Residues	Score	Residues	Score
1-2	E-S	0	C-N	-4	S-S	2	Q-G	-1
1-3	E-S	0	C-W	-8	S-K	0	Q-N	1
1-4	E-S	0	C-C	12	S-S	2	Q-N	1
2-3	S-S	2	N-W	-4	S-K	0	G-N	0
2-4	S-S	2	N-C	-4	S-S	2	G-N	0
3-4	S-S	2	W-C	-8	K-S	0	N-N	2
		6		-16		6		3

…Sum of Pairs Method

- Problem with this approach:
  - more closely related sequences will have a higher weight
- The MSA program gets around this by calculating weights to associate to each sequence alignment pair.
- The weights are assigned based on the predicted tree of the aligned sequences.

Summary of MSA

1. Calculate all pairwise alignment scores
2. Use the scores to predict tree
3. Calculate pair weights based on the tree
4. Produce a heuristic MSA based on the tree
5. Calculate the maximum weight for each sequence pair
6. Determine the spatial positions that must be calculated to obtain the optimal alignment
7. Perform the optimal alignment
8. Report the weight found compared to the maximum weight previously found

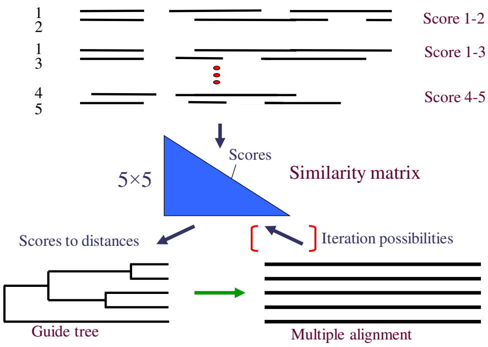
Progressive MSA…

- MSA program is limited in size
- The approach of progressive alignment is
  - to begin with an alignment of the most alike sequences,
  - and then build upon the alignment using other sequences.
    - Feng, D., Doolittle, R.F. Progressive sequence alignment as a prerequisitetto correct phylogenetic trees. J Mol Evol 25, 351–360 (1987). https://doi.org/10.1007/BF02603120
    - Grasso C, Lee C (2004). "Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems". Bioinformatics. 20 (10): 1546–56. doi:10.1093/bioinformatics/bth126
- Progressive alignments work by
  - first aligning the most alike sequences using dynamic programming,

…Progressive MSA…

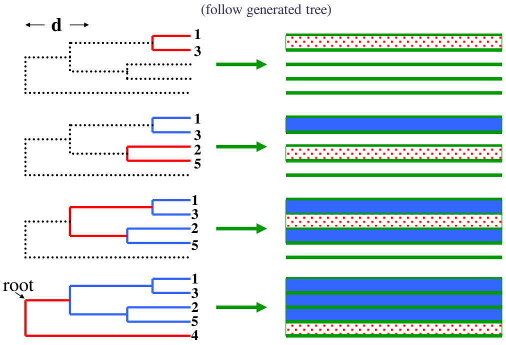
- and then progressively adding less related sequences to the initial alignment
- alignment on each of the pairs of sequences
  - next, trail MSA is produced by first predicting a phylogenetic tree for the sequences
  - sequences are then multiply aligned in order of their relationship on the tree
    - starting with the most related sequences
    - then progressively adding less related sequences to the initial alignment
- used by PILEUP and CLUSTALW
- not guaranteed to be optimal

…Progressive MSA…



…Progressive MSA…

(follow generated tree)



CLUSTALW- CLUSTALX

- The guide tree in the initial programs was constructed via a UPGMA CLUSTer analysis of the pairwise ALignments, hence the name CLUSTAL
- CLUSTALW and CLUSTALX are progressive alignment programs
  - W for Weighted ; X for X Window
  - UPGMA (unweighted pair group method with arithmetic mean)
- They follow the following steps:
  1. Perform pairwise alignments of all sequences
  2. Use alignment scores to produce phylogenetic tree
  3. Align sequences sequentially, guided by the tree

CLUSTALW

- The initial pairwise alignments are calculated using an enhanced dynamic programming algorithm

- the genetic distances used to create the phylogenetic tree are calculated by dividing the total number of mismatched positions by the total number of matched positions.

CLUSTALW

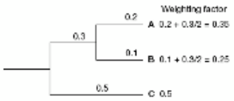
- Alignments are associated a weight based on their distance from the root node (next slide)

- Gaps are added to an existing profile in progressive methods

- CLUSTALW incorporates a statistical model in order to place gaps where they are most likely to occur

CLUSTALW - example

A. Calculation of sequence weights



B. Use of sequence weights

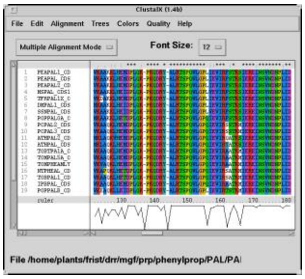
```
                          Column in alignment 1
Sequence A (weight a)     ..........K.........
Sequence B (weight b)     ..........L.........

                          Column in alignment 2
Sequence C (weight c)     ..........L.........
Sequence D (weight d)     ..........V.........
```

Score for matching these two column in an msa =

[a x c x score (K,L) +
a x d x score (K,V) +
b x c x score (L,L) +
b x d x score (L,V)] / 4

CLUSTALW / CLUSTALX



File /home/plants/frist/drr/mgf/prp/phenylprop/PAL/PAl

- ‘W’ stands for “weighting”
  - ability to provide weights to sequence and program parameters
- CLUSTALX – with graphical interface
- provides global MSA
- Not constructed to perform local alignments.
- Similarity in small regions is a problem.
- Problems with large insertions.
- Problems with repetitive elements, such as domains.
- ClustalW does not guarantee an optimal solution

http://www.ebi.ac.uk/clustalw/

CLUSTALW / CLUSTALX

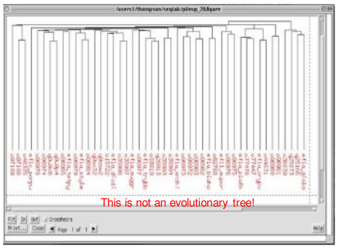
- Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTALW: improving the sensitivity of progressivemultiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research, 22, 4673–4680.
- Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., & Higgins, D. G. (1997). The CLUSTAL X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research, 25, 4876–4882.

PILEUP

- PILEUP is the multiple sequence alignment program that is part of the Genetics Computer Group (GCG) package developed at the University of Wisconsin.
- Sequences initially aligned in a pair-wise fashion using Needleman-Wunsch algorithm.
- Scores used to produce tree using unweighted pair group method using arithmetic averages (UPGMA)
- The resulting tree is then used to guide the alignment of the most closely related sequences and groups of sequences

PILEUP

- very similar to CLUSTALW
- part of the genetic computer group (GCG)
- does not guarantee optimal alignment
- plots a cluster dendogram of similarities between sequences


/users1/thompson/seqlab/pileup_28.figure
Fit In Out Crosshairs
Print... Close Page 1 of 1
Help

This is not an evolutionary tree!

Shortcoming of Progressive Approach

- Dependence upon initial pair-wise sequence alignments
  - Ok if sequences are similar
  - Errors in alignment propagated if not similar
- Suitable scoring matrices and gap penalties must be chosen to apply to the sequences as a set

Iterative Methods

- Iterative alignment methods begin by making an initial alignment of the sequences.
- These alignments are then revised to give a more reasonable result.
- The objective of this approach is to improve the overall alignment score
- Alignment is repeatedly refined
- Selection of groups is based on the phylogenetic tree
  - Hirosawa, M., Totoki, Y., Hoshida, M., & Ishikawa, M. (1995). Comprehensive study on iterative algorithms of multiple sequence alignment. Bioinformatics, 11(1), 13–18. doi:10.1093/bioinformatics/11.1.13
- Programs using iterative methods:
  - MultAling
  - PRRP (Profile-based Randomazed iterative Refinement method for alignment of Protein sequences)
  - DIALIGN (DIagonal ALIGNment)

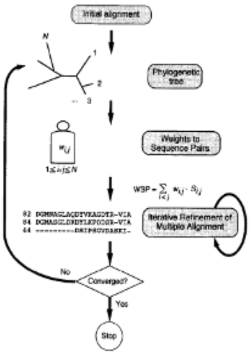
MultAlign

- Pairwise scores recalculated during progressive alignment
- Tree is recalculated
- Alignment is refined

PRRP

- Initial pairwise alignment predicts tree
- Tree produces weights
- Locally aligned regions considered to produce new alignment and tree
- Continue until alignments converge

Iterative procedure used by PRRP to compute MSA


Initial alignment
N
1
2
3
Phylogenetic tree
w_{ij}
$1 \le i < j \le N$
Weights to Sequence Pairs
$WSP = \sum_{i<j} w_{ij} \cdot S_{ij}$
82 DGMHAGLAQDTVKAGDYR-VIA
84 DGMASGLDRDTLKPDDHR-VTA
44 ---------DRIPSGVDASKI-
Iterative Refinement of Multiple Alignment
No
Converged?
Yes
Stop

Gotoh, O. (1996). Significant Improvement in Accuracy of Multiple Protein Sequence Alignments by Iterative Refinement as Assessed by Reference to Structural Alignments. Journal of Molecular Biology, 264(4), 823 – 238. doi:10.1006/jmbi.1996.0679

GOTOH, O. (1999). Multiple sequence alignment: Algorithms and applications. Advances in Biophysics, 36, 159–206. doi:10.1016/s0065-227x(99)80007-0

DIALIGN

- Pairs of sequences aligned to locate ungapped aligned regions
- Diagonals of various lengths identified
- Collection of weighted diagonals provide alignment

Genetic Algorithm Approach…

- The goal of genetic algorithms used in sequence alignment is to generate as many different multiple sequence alignments by rearrangements that simulate gaps and genetic recombination events.
- SAGA (Serial Alignment by Genetic Algorithm) is one such approach that yields very promising results, but becomes slow when more than 20 sequences are used.

…Genetic Algorithm Approach…

1) Sequences (up to 20) written in row, allowing for overlaps of random length – ends padded with gaps (100 or so alignments)

XXXXXXXXXXX-----
---------XXXXXXXX
--XXXXXXXXX-----

…Genetic Algorithm Approach…

2) The initial alignments are scored by the sum of pairs method.
   - Standard amino acid scoring matrices and gap open, gap extension penalties are used

3) Initial alignments are replaced to give another generation of multiple sequence alignments
   - One half of the multiple sequence alignments are chosen to proceed to the next generation unchanged (natural selection).
     - This half is chosen by assigning probabilities to each sequence based on an inverse proportion of their SP scores (the best alignments, since the SP scores are weighted according to their distance from the parent).
   - The other half of the alignments are sent to the next generation, but are first subject to mutation.

…Genetic Algorithm Approach…

4) MUTATION:
   - In the mutation process, gaps are inserted into the sequences subject to mutation and rearranged in an attempt to create a better scoring alignment
   - In this step
     - the sequences subject to mutation split into two sets based on estimated phylogenetic tree
     - gaps of random lengths inserted into random positions in the alignment

…Genetic Algorithm Approach…

- Mutations:

- XXXXXXXX XXX---XXX—XX
- XXXXXXXX XXX---XXX—XX
- XXXXXXXX X—XXX---XXXX
- XXXXXXXX X—XXX---XXXX
- XXXXXXXX X—XXX---XXXX

…Genetic Algorithm Approach

5) Recombination of two parents to produce next generation alignment is accomplished

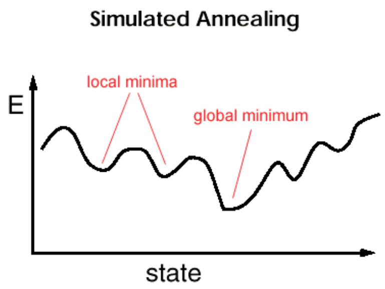
6) The next generation is evaluated going back to step 2, and steps 2-5 are repeated a number (100-1000) times.
   - The best scoring multiple sequence alignment is then obtained
     - note that it may not be the optimal scoring alignment.

7) The entire process is repeated several times, starting from a different initial alignment each time.
   - The best scoring multiple sequence alignment is then chosen and reported to the user.

Simulated Annealing…

- Another approach to sequence alignment that works in a manner similar to genetic algorithms is simulated annealing.
- In these approaches, you begin with a heuristically determined multiple sequence alignment that is then changed using probabilistic models that identifies changes in the alignment that increase the alignment score.
- The drawback of simulated annealing approaches is that you can get stuck finding only the locally optimal alignment rather than the alignment score that is globally optimal
- Rearranges current alignment using probabilistic approach to identify changes that increase alignment score

…Simulated Annealing



http://www.cs.berkeley.edu/~amd/CS294S97/notes/day15/day15.html

Group Approach

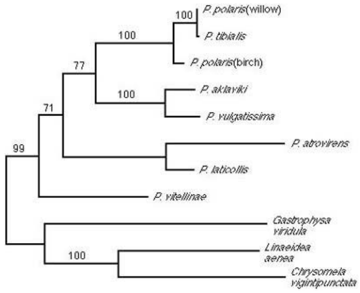
- Sequences aligned into similar groups
- Consensus of group is created
- Alignments between groups is formed

- EXAMPLES: PIMA, MULTAL

Tree Approach to MSA…

- Tree created
- Two closest sequences aligned
- Consensus aligned with next best sequence or group of sequences
- Proceed until all sequences are aligned

…Tree Approach to MSA…



- www.sonoma.edu/users/r/rank/ research/evolhost3.html

…Tree Approach to MSA

- PILEUP, CLUSTALW and ALIGN

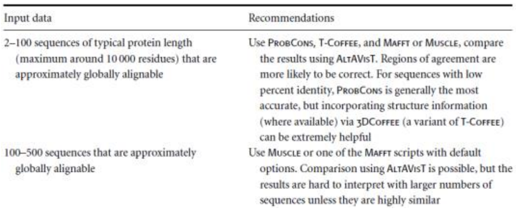
- TREEALIGN
  - rearranges the tree as sequences are added, to produce a maximum parsimony tree (fewest evolutionary changes)

Profile Analysis

- Create multiple sequence alignment
- Select conserved regions
- Create a matrix to store information about alignment
  - one row for each position in alignment
  - one column for each residue; gap open; gap extend
- Profile can be used to search target sequence or database for occurrence
- Drawback:
  - profile is skewed towards training data

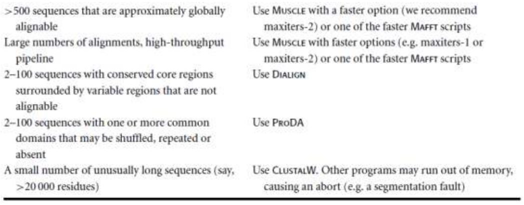
Typical alignment tasks and recommended procedures…

- The following table is taken from "Current Opinion in Structural Biology, 16, Edgar R.C. & Batzoglou S.,Multiple sequence alignment, 368–373, Copyright Elsevier (2006)«
- https://doi.org/10.1016/j.sbi.2006.04.004



Input data	Recommendations
2–100 sequences of typical protein length (maximum around 10 000 residues) that are approximately globally alignable	Use ProbCons, T-Coffee, and Mafft or Muscle, compare the results using AltAVisT. Regions of agreement are more likely to be correct. For sequences with low percent identity, ProbCons is generally the most accurate, but incorporating structure information (where available) via 3DCoffee (a variant of T-Coffee) can be extremely helpful
100–500 sequences that are approximately globally alignable	Use Muscle or one of the Mafft scripts with default options. Comparison using AltAVisT is possible, but the results are hard to interpret with larger numbers of sequences unless they are highly similar

…Typical alignment tasks and recommended procedures



>500 sequences that are approximately globally alignable	Use Muscle with a faster option (we recommend maxiters-2) or one of the faster Mafft scripts
Large numbers of alignments, high-throughput pipeline	Use Muscle with faster options (e.g. maxiters-1 or maxiters-2) or one of the faster Mafft scripts
2–100 sequences with conserved core regions surrounded by variable regions that are not alignable	Use Dialign
2–100 sequences with one or more common domains that may be shuffled, repeated or absent	Use ProDA
A small number of unusually long sequences (say, >20 000 residues)	Use ClustalW. Other programs may run out of memory, causing an abort (e.g. a segmentation fault)