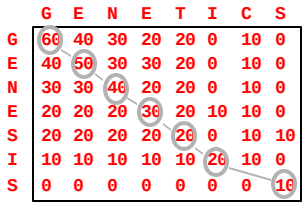
Multiple Sequence Alignment: An Introduction



	G	E	N	E	T	I	C	S
G	60	40	30	20	20	0	10	0
E	40	50	30	30	20	0	10	0
N	30	30	40	20	20	0	10	0
E	20	20	20	30	20	10	10	0
S	20	20	20	20	20	0	10	10
I	10	10	10	10	10	20	10	0
S	0	0	0	0	0	0	0	10

Resources for this lecture

- Web page
  http://www.carleton.ca/~jcheetha/topics
  - Lecture Slides
  - Lab
  - Notes
  - Papers
  - Programs

Alignments tell us about...

- Function or activity of a new gene/protein
- Structure or shape of a new protein
- Location or preferred location of a protein
- Stability of a gene or protein
- Origin of a gene or protein
- Origin or phylogeny of an organelle
- Origin or phylogeny of an organism

Factoid:

Sequence comparisons lie at the heart of all bioinformatics

Similarity versus Homology

- Similarity refers to the likeness or % identity between 2 sequences
- Similarity means sharing a statistically significant number of bases or amino acids
- Similarity does not imply homology
- Homology refers to shared ancestry
- Two sequences are homologous is they are derived from a common ancestral sequence
- Homology usually implies similarity

Similarity versus Homology

- Similarity can be quantified
- It is correct to say that two sequences are X% identical
- It is correct to say that two sequences have a similarity score of Z
- It is generally incorrect to say that two sequences are X% *similar*

Similarity versus Homology

- Homology cannot be quantified
- If two sequences have a high % identity it is OK to say they are homologous
- It is incorrect to say two sequences have a homology score of Z
- It is incorrect to say two sequences are X% homologous

Sequence Complexity

MCDEFGHIKLAN…. High Complexity

ACTGTCACTGAT…. Mid Complexity

NNNNTTTTTNNN…. Low Complexity

Translate those DNA sequences!!!

Assessing Sequence Similarity

```
THESTORYOFGENESIS
THISBOOKONGENETICS
```
Two Character Strings

```
THESTORYOFGENESI-S
* * * * * * * * * *
THISBOOKONGENETICS
```
Character Comparison

```
THE STORY OF GENESIS
THIS BOOK ON GENETICS
```
Context Comparison

Is This Alignment Significant?

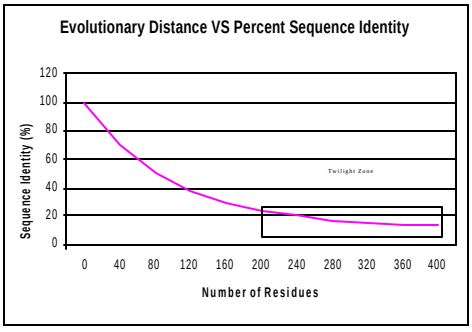
```
Gelsolin  89   L G N E L S Q D E S G A A A I F T V Q L  108
Annexin   82   L P S A L K S A L S G H L E T V I L G L  101
         154   L E K D I I S D T S G D F R K L M V A L  173
         240   L E - S I K K E V K G D L E N A F L N L  258
         314   L Y Y Y I Q Q D T K G D Y Q K A L L Y L  333

Consensus      L x P x x x P D x S G x h x x h x V L L
```


Some Simple Rules

- If two sequence are > 100 residues and > 25% identical, they are likely related
- If two sequences are 15-25% identical they may be related, but more tests are needed
- If two sequences are < 15% identical they are probably not related
- If you need more than 1 gap for every 20 residues the alignment is suspicious

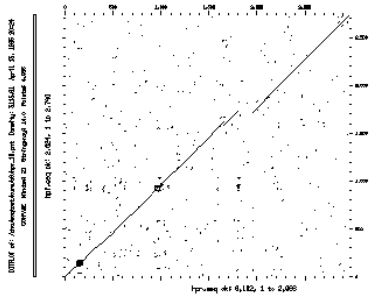
Doolittle's Rules of Thumb



Sequence Alignment - Methods

- Dot Plots
- Dynamic Programming
- Heuristic (Fast) Local Alignment
- Multiple Sequence Alignment

Dot Plots



Dot Plots

- "Invented" in 1970 by Gibbs & McIntyre
- Good for quick graphical overview
- Simplest method for sequence comparison
- Inter-sequence comparison
- Intra-sequence comparison
  - Identifies internal repeats
  - Identifies domains or "modules"

Sequence Alignments

- Align two or more sequences to show maximum similarity: DNA, RNA, or protein.
- Implied homology, descent from common ancestor. Implied similar functions.
  - Implied for each residue independently
- Align total length sequences, or shorter regions first, or just domains.
- Align left and right ends, with internal gaps for sequences of unequal length.
- aa-alignments: both identity and similarity (based on chemical properties)

Alignment can be easy or difficult

```
GCGGCCCA TCAGGTAGTT GGTGG
GCGGCCCA TCAGGTAGTT GGTGG
GCGTTCCA TCAGCTGGTT GGTGG
GCGTCCCA TCAGCTAGTT GGTGG
GCGGCGCA TTAGCTAGTT GGTGA
******** ********** *****

TTGACATG CCGGGG---A AACCG
TTGACATG CCGGTG--GT AAGCC
TTGACATG -CTAGG---A ACGCG
TTGACATG -CTAGGGAAC ACGCG
TTGACATC -CTCTG---A ACGCG
******** ?????????? *****
```

Easy

Difficult due to insertions or deletions (indels)

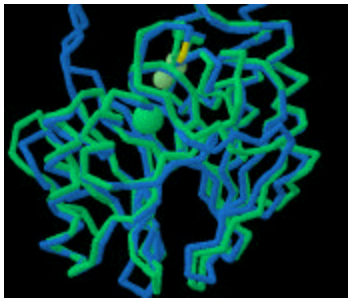
Multiple Sequence Alignment

- More than two sequences
- Example: protein families
- Implied common ancestor
  - Makes sense to align sequences only if they share a common evolutionary ancestor.
- Orthologs as organisms diverged
- Paralogs by gene duplication within one organism
- Across broad taxonomic groupings

Multiply Aligned Proteins

- Ideally, the amino acids in each column of the alignment occupy a similar 3D structural position. All aa's in that position descend from a common ancestral amino acid, i.e. are homologous.
- But, protein sequences and structures diverge, so it is not always possible to know if an alignment is "correct," i.e. represents homology.
- See the structural example next.

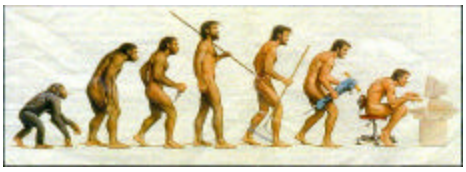
Superposition of Polypeptide Backbones of Aligned Proteins



Why Do Multiple Alignments?

- To represent sequence similarities and differences within a group of related sequences.
- To show evolutionary pattern of nature's successful experiments in generating genetic diversity.
- To indicate which residues can be changed without destroying adaptiveness.
- For DNA: To find regions for PCR primers.
- For proteins: To predict regions of conserved 2D and 3D structures.
- First step in molecular phylogenetics analysis.

Molecular Phylogenetics



Purposes of Multiple Alignments of Proteins

- Discover conserved regions.
- Uncover patterns of α-helix, β-sheet.
- Uncover patterns of hydrophobicity and hydrophilicity.
- Find "gappy" regions of surface loops, hypervariability.
- To search for family members and remote homologs in databases.
- Do phylogenetic analysis, tree drawing.

Databases of Protein Alignments

- Pfam
- Prosite
- Prints
- Blocks

```
ID   p99.1.4782; BLOCK
AC   BP04782A; distance from previous block=(0,11)
DE   PROTEIN SEX-DETERMINING REGION Y TESTIS-DETERMI
BL   DPA; width=17; seqs=8; 99.5%=837; strength=1238
SRY_CALJA|P51501   (   8) MLRVFNSDEYNPAALQN 100

SRY_BISBO|Q27949   (   1) MFRVLNDDVYSPAVVQQ  55
SRY_BOVIN|Q03255   (   1) MFRVLNDDVYSPAVVQQ  55
SRY_CAPHI|Q03256   (  12) MFRVLKDDVYSPAVVQQ  57
SRY_PIG|P36393     (   9) MFRVLKADDYSPAAQQQ  73
SRY_SHEEP|Q03257   (  12) MFRVLKDDVYSPAVVQQ  57

SRY_GORGO|P48046   (   8) MLSVFNSDDYSPAVQQT  77
SRY_HUMAN|Q05066   (   8) MLSVFNSDDYSPAVQEN  85
```

Dynamic Programming

	G	E	N	E	T	I	C	S
G	10	0	0	0	0	0	0	0
E	0	10	0	10	0	0	0	0
N	0	0	10	0	0	0	0	0
E	0	0	0	10	0	10	0	0
S	0	0	0	0	0	0	0	10
I	0	0	0	0	0	10	0	0
S	0	0	0	0	0	0	0	10

	G	E	N	E	T	I	C	S
G	50	40	30	20	20	0	10	0
E	40	60	30	30	20	0	10	0
N	30	30	70	30	20	0	10	0
E	20	20	20	80	30	20	10	0
S	20	20	20	20	20	10	10	10
I	10	10	10	10	10	20	10	0
S	0	0	0	0	0	0	0	10

```
G E N E T I C S
| | | | * | |   |
G E N E S I   S
```


Dynamic Programming

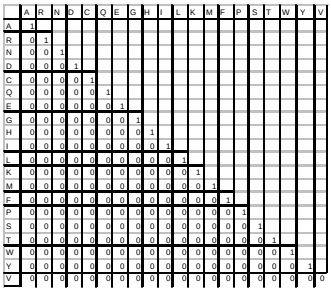
- Developed by Needleman & Wunsch (1970)
- Refined by Smith & Waterman (1981)
- Ideal for quantitative assessment
- Guaranteed to be mathematically optimal
- Slow N^2 algorithm
- Performed in 2 stages
  - Prepare a scoring matrix using recursive function
  - Scan matrix diagonally using traceback protocol

The Recursive Function

$$S_{ij} = s_{ij} + \max \begin{cases} S_{i-1,j-1} & \text{or} \\ \max_{2<x<i} S_{i-x,j-1} + w_{x-1} & \text{or} \\ \max_{2<y<i} S_{i-1,j-y} + w_{y-1} \end{cases}$$

W = gap penalty
S = alignment score

Identity Scoring Matrix (S_{ij})



A Simple Example...

```
  A A T V D
A 1
V
V
D
```

```
  A A T V D
A 1 1
V
V
D
```

```
  A A T V D
A 1 1 0 0 0
V
V
D
```

```
  A A T V D
A 1 1 0 0 0
V 0
V
D
```

```
  A A T V D
A 1 1 0 0 0
V 0 1 1
V
D
```

```
  A A T V D
A 1 1 0 0 0
V 0 1 1 2
V
D
```


A Simple Example...

```
  A A T V D
A 1 1 0 0 0
V 0 1 1 2 1
V
D
```

```
  A A T V D
A 1 1 0 0 0
V 0 1 1 2 1
V 0 1 1 2 2
D 0 1 1 1 3
```

```
  A A T V D
A 1 1 0 0 0
V 0 1 1 2 1
V 0 1 1 2 2
D 0 1 1 1 3
```

```
  A A T V D
  |   | |
A V - V D
```

```
  A A T V D
  |   | |
A - V V D
```

```
  A A T V D
  |   | |
A V V D
```

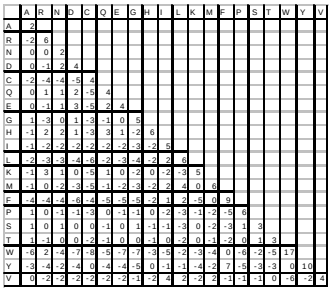
Could We Do Better?

- Key to the performance of Dynamic Programming is the scoring function
- Dynamic Programming always gives the mathematically correct answer
- Dynamic Programming does not always give the biologically correct answer
- The weakest link -- The Scoring Matrix

Scoring Matrices

- An empirical model of evolution, biology and chemistry all wrapped up in a 20 X 20 table of integers
- Structurally or chemically similar residues should ideally have high diagonal or off-diagonal numbers
- Structurally or chemically dissimilar residues should ideally have low diagonal or off-diagonal numbers

A Better Matrix - PAM250



PAM Matrices

- Developed by M.O. Dayhoff (1978)
- PAM = Point Accepted Mutation
- Matrix assembled by looking at patterns of substitutions in closely related proteins
- 1 PAM corresponds to 1 amino acid change per 100 residues
- 1 PAM = 1% divergence or 1 million years in evolutionary history

Using PAM250...

A T V D
A 2
T 1 3
V 0 0 4
D 0 0 -2 4

Gap Penalty = -1

```
  A A T V D      A A T V D      A A T V D
A 2            A 2 2          A 2 1 1 0 0
V              V              V
V              V              V
D              D              D

  A A T V D      A A T V D      A A T V D
A 2 1 1 0 0    A 2 1 1 0 0    A 2 1 1 0 0
V 0 2          V 0 2 1        V 0 2 1 5
V              V              V
D              D              D
```


Using PAM250...

A T V D
A 2
T 1 3
V 0 0 4
D 0 0 -2 4

Gap Penalty = -1

```
  A A T V D      A A T V D      A A T V D
A 2 1 1 0 0    A 2 1 1 0 0    A 2 1 1 0 0
V 0 2 1 5 -1   V 0 2 1 5 -1   V 0 2 1 5 -1
V              V 0 1 2 5 3    V 0 1 2 5 3
D              D 0 1 1 0 9    D 0 1 1 0 9

      A A T V D
      |   | |
      A V - V D
```

Multiple Alignment Algorithm

- *Take all "n" sequences and perform all possible pairwise (n/2(n-1)) alignments*
- *Identify highest scoring pair, perform an alignment & create a consensus sequence*
- *Select next most similar sequence and align it to the initial consensus, regenerate a second consensus*
- *Repeat step 3 until finished*

Multiple Sequence Alignment

- Developed and refined by many (Doolittle, Barton, Corpet) through the 1980's
- Used extensively for extracting hidden phylogenetic relationships and identifying sequence families
- Powerful tool for extracting new sequence motifs and signature sequences

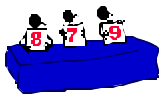
Multiple Alignment

- Most commercial vendors offer good multiple alignment programs including:
  - GCG (Oxford Molecular Group)
  - PepTool/GeneTool (BioTools Inc.)
  - LaserGene99 (DNAStar)
- Popular web servers include MULTALIN and CLUSTALW
- Popular freeware includes PHYLIP & PAUP

Mutli-Align Websites

- Match-Box
  http://www.fundp.ac.be/sciences/biologie/bms/matchbox_submit.shtml
- MUSCA http://cbcsrv.watson.ibm.com/Tmsa.html
- MSA http://stateslab.wustl.edu/ibc/msa.html
- MULTALIN
  http://www.toulouse.inra.fr/multalin.html
- CLUSTALW http://www2.ebi.ac.uk/clustalw/

Scoring Alignments



- Scoring matrix like BLOSUM or PAM
  - BLOSUM62 useful standard
  - May be altered as user wishes
- Consider both identity and similarity
- Common ancestry, conserved sequence, implies conserved function.
- Gap penalties

BLOSUM 62 Matrix

```
A Ala  4
R Arg -1  5
N Asn -2  0  6
D Asp -2 -2  1  6
C Cys  0 -3 -3 -3  9
Q Gln -1  1  0  0 -3  5
E Glu -1  0  0  2 -4  2  5
G Gly  0 -2  0 -1 -3 -2 -2  6
H His -2  0  1 -1 -3  0  0 -2  8
I Ile -1 -3 -3 -3 -1 -3 -3 -4 -3  4
L Leu -1 -2 -3 -4 -1 -2 -3 -4 -3  2  4
K Lys -1  2  0 -1 -3  1  1 -2 -1 -3 -2  5
M Met -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5
F Phe -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6
P Pro -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4  7
S Ser  1 -1  1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4
T Thr  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5
W Trp -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1  1 -4 -3 -2 11
Y Tyr -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2  2  7
V Val  0 -3 -3 -3 -1 -2 -2 -3 -3  3  1 -2  1 -1 -2 -2  0 -3 -1 4
     Ala Arg Asn Asp Cys Gln Glu Gly His Ile Leu Lys Met Phe Pro Ser Thr Trp Tyr Val
       A   R   N   D   C   Q   E   G   H   I   L   K   M   F   P   S   T   W   Y   V
```

Matrix for pairing rewards and mispairing penalties

Alignment Methods

- By hand: requires expert knowledge of protein evolution and structure; also tedious.
- Automatic: done by a computer program; requires a scoring method that can be used to measure how good a particular alignment is. ClustalW is this type of method.
- Refinement: after automatic alignment, improve with advanced methods (softwa or by hand (requires an editor).

Methods: Multidimensional Dynamic Programming

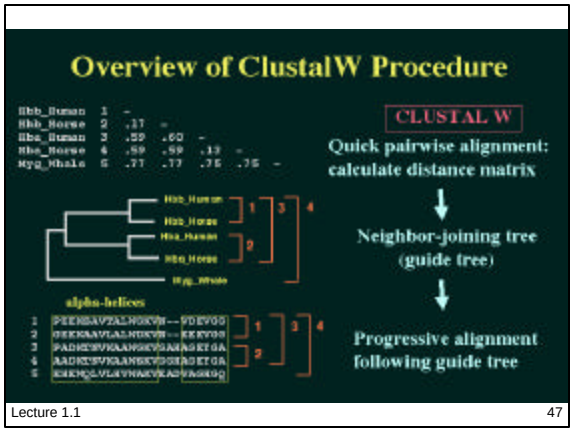
- Guarantees an optimal solution.
- Memory and computational time required increase exponentially with the product of the sequence lengths.
- **MSA** algorithm is the foremost implementation; can align 5-7 proteins of 200-300 residues. But servers have limitations:
  - Runs > 10 minutes are terminated.
  - If > 4 sequences, limit lengths to 100 aas.

Methods: Progressive Alignment

- General method is to construct a succession of pairwise alignments.
- Most common approach
- Fast and efficient due to use of heuristics
- **Feng-Doolittle** method: construct a "guide tree" of similarities; pairwise align most similar sequences first.
- **ClustalW**: adds more sophisticated methods, but still works by adding one sequence at a time

Clustal Algorithm

1. Compare *n* sequences in pairs. Do all possible pairs [*n x (n-1) / 2*].
2. Construct a similarity matrix [*n x n*] of pairwise alignments.
3. Rank sequences according to how well they align pairwise.
4. Do progressive alignment: Start with best pair of sequences and continue aligning, until all sequences are aligned.



Clustal Algorithm II

4. Do progressive alignment:
   1) start with best pair of sequences
   2) align next most similar sequence to first pair
   3) continue aligning next most similar, until all sequences are aligned.
5. No iteration is done to refine the final alignment.
6. ClustalW may not give best possible MSA. Usually need to refine MSA by hand.

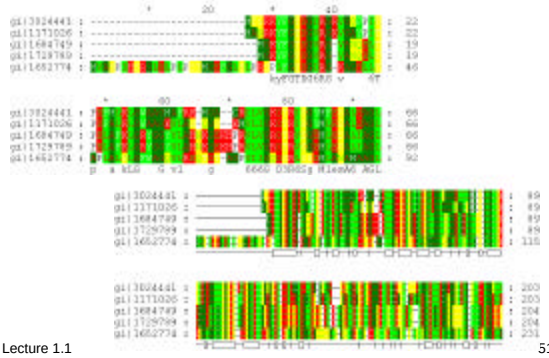
ClustalW Features

- Global alignment
- Progressive alignment
- Fast, heuristic, but not guaranteed to find the best alignment
  - Should adjust MSA by hand
- Good for similar length sequences
- Good for protein families

Alignment Editors

- Programs that allow the user to manipulate a multiple sequence alignment that has been produced by software.
- Sections of sequences can be moved to improve the alignment (possibly).
- Most importantly, sections of the MSA can be shaded and highlighted to show which aas are the most similar (conserved), and what the chemical properties are.
- The GeneDoc and JalView programs are examples.

GeneDoc Screens



Progressive Alignments

- True for ClustalW and all progressive alignment methods: heuristic, not exhaustive, may not find "best" MSA.
- Mathematical process, not connected with biological reality.
- Output MSA can usually be improved by hand. Inspect MSA closely, adjust.
- Use viewer with color-coding.
- Make judgments based on biology.
- Emphasize most reliable regions of MSA, based on convincing positional homology.

Various Methods of Multiple Alignment

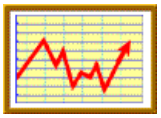
- MSA program: if few, short sequences
- ClustalW: usually best
- ClustalX: graphical
- PILEUP: in GCG package
- DIALIGN: local gap-free alignments first
- Match-Box: good for conserved domains
- SAGA: "genetic algorithm"
- MEME: iterative realignment

Test of Alignment Method

1. Align related protein sequences that have known 3D structures.
2. Match structures, identify common domains and structural motifs.
3. Check if sequence alignment has correctly aligned the common motifs.
4. Evaluate alignment methods by comparing with each other.

Still Improvements to Seek

- Better algorithms: faster, more efficient
- Better scoring: reasonable biologically
- Able to deal with repeats
- Able to deal with domain swaps
- Incorporate external information
  - eg. active sites, critical residues



Web Sites

- ClustalW servers:

http://www.clustalw.genome.ad.jp/;
http://dot.imgen.bcm.tmc.edu:9331/multi-align/multi-align.html;
http://www.genebee.msu.su/genebee.html;
http://www.bionavigator.com;
also others

- multiple sequence analysis tools:
  http://www.techfak.uni-bielefeld.de/bcd/Curric/MulAli/welcome.html

More Web Sites

- Jalview: http://www.ebi.ac.uk/~michele/jalview/
- BoxShade viewer: http://www.ch.embnet.org/software/BOX_form.html
- GeneDoc viewer: http://www.psc.edu/biomed/genedoc/
- Phylip phylogenetic analysis: http://evolution.genetics.washington.edu/phylip.html

The problem: Sequence comparison

- How to compare one sequence (target) to many sequences (database search)
- How to compare more than two sequences simultaneously

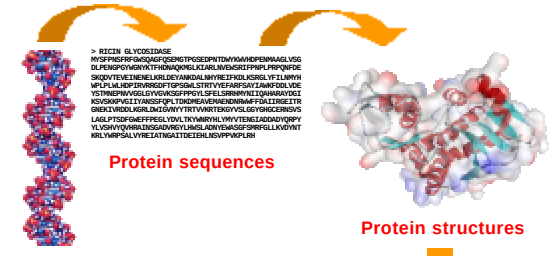
What is a Multiple Sequence Alignment (MSA) ?

- MSA is the alignment of N sequences (Protein or DNA) simultaneously, where $N > 2$.
- Let S_i denote a sequence in the Global Multiple Sequence Alignment of $N > 2$ sequences $S = \{ S_1, \ldots, S_N \}$ is obtained by inserting gaps denoted by " - "possibly at the beginning or end, positions.
- The new set of N sequences denoted by $S' = \{ S_1', \ldots, S_N' \}$ will all have length L

Reasons for aligning sets of sequences

- Infer *phylogenetic* trees from *homologous* sequences
- Highlight conserved sites/regions
  - Residues conserved during evolution play an important role
- Highlight variable sites/regions
- Uncover changes in gene/protein structure
- Prediction of protein structure and function
  - Proteins which are very similar in sequence generally have similar 3D structure and function as well (*Homology modeling*)
  - By searching a sequence of unknown structure against a database of known proteins the structure and/or function can in many cases be predicted

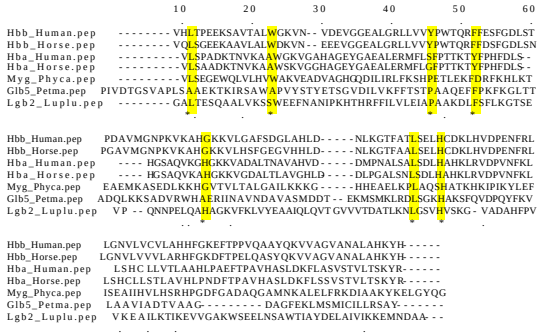
From Sequence to Function



Global vs. Local Alignments

- *Global alignment* algorithms start at the beginning of two sequences and add gaps to each until the end of one is reached.
- *Local alignment* algorithms finds the region (or regions) of highest similarity between two sequences and build the alignment outward from there.

Multiple Sequence Alignment



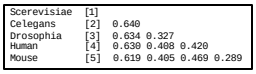
The challenge: pairwise to multiple alignment

- in principle pairwise approach applicable to multiple alignment - but not practical
- CPU time proportional to sequence length
  e.g. if aligning 2 sequences of
  300 positions = 1 second
  … 3 sequences = 300 seconds
  and 10 sequences = 300^8 seconds
  = 951 years

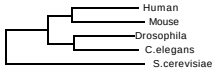
Progressive Multiple Alignment (Clustal)

- Any Exact Method would be *TOO SLOW*
- We use a *Heuristic Algorithm*.
- *Progressive Alignment Algorithms* are the most Popular
  - -Fast
  - -Clustal
  - -Greedy Heuristic (No Guarranty).

Progressive alignment



1. **Do pairwise alignment** of all sequences and calculate distance matrix.



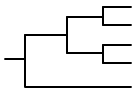
2. **Create a guide tree** based on this pairwise distance matrix.

3. **Align progressively following guide tree.**
   - start by aligning most closely related pairs of sequences
   - (keep gaps that appeared in sequence pairs fixed)
   - at each step align two sequences or one to an existing subalignment

Parallel Clustal

•**Pairwise (PW) alignment matrix**
average alignment calculation spends most of its time here
can parallelize as all elements are independent

•**Guide tree calculation**
Calculation of closest sequences (branch)can be parallelized. Together with PW matrix calculation Clustal W is ~85-92% parallel



•**Progressive alignment**
Remaining ~5-10% of the code can be parallelized at this stage by calculating profile scores in parallel. As a result the whole application can be ~93-98% parallel depending on a size of a problem

Parallel Clustal: extensions

Optimization of input parameters - scoring matrices, gap penalties - requires many repetitive Clustal W calculations with various input parameters.

Minimum Vertex Cover – use k-minimum vertex cover to remove erroneous sequences, and identify clusters of highly similar sequences.

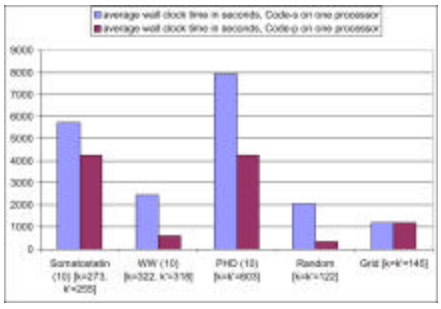
Minimum Vertex Cover

Minimum vertex cover (classic problem)

Definition: A set of vertices in an undirected graph where every edge connects at least one vertex. The minimum vertex cover problem is to find a minimum size set and is NP-complete.

Cheetham, J.J., Dehne, F., Rau-Chaplin, A., Stege, U. and Taillon, P.J. (2003) **Solving Large FPT Problems on Course Grained Parallel Machines**, *J. Computer and Systems Science (in press).*

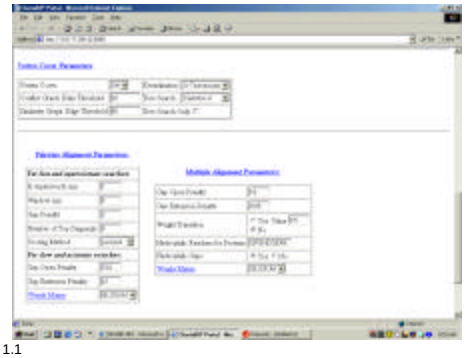
Parallel FPT Minimum Vertex Cover



Clustal XP

http://134.117.206.42:8000/

Clustal XP



Conclusions and Summary

- Parallel computing can greatly speed up multiple sequence alignments
- Minimum vertex cover can be very useful in multiple sequence alignment to detect misplaced sequences and detect groups of very similar sequences
- A fixed parameter tractable algorithm can be used to solve minimum k-vertex cover problems for large values of k.