1. Uses of MSA

2. Technical difficulties
   1. Select sequences
   2. Select objective function
   3. Optimize the objective function
      1. Exact algorithms
      2. Progressive algorithms
      3. Iterative algorithms
      4. Consistency-based algorithms

3. Tools to view alignments
   1. MEGA
   2. BOXSHADE & Seq. LOGOS
USES of Multiple Sequence Alignment

- Sequence relationships
- Conservation patterns
  - Mutable regions
  - Conserved residues
  - Conserved properties
  - Conserved sequence patterns
  - Domain boundaries
  - [...]

MSA can be used to:
- Infer function
- Predict secondary structure
- Phylogenetic reconstruction
- Identify residues important for function
- Prioritize mutations for functional analyses
- Sensitive database searching algorithms (PSI BLAST)

If the MSA is incorrect, all the above inferences will be incorrect!

Mutants found in the flowering gene VRN1
- Which mutant is likely to result in a non-functional protein?

<table>
<thead>
<tr>
<th>BLOSUM62 Mutations</th>
<th>2</th>
<th>-3</th>
<th>-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>TaVRN-B1</td>
<td>R</td>
<td>P</td>
<td>V</td>
</tr>
<tr>
<td>TaVRN-A1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oryza</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zea</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sorghum</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arabidopsis</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Original

- GNCHEYRKLKAVETIQKCQHKLMGEDLESLNLKELQLEQLES
- GNCHEYRKLKAVETIQKCQHKLMGEDLESLNLKELQLEQLES
- GNCHEYRKLKAVETIQKCQHKLMGEDLESLNLKELQLEQLES
- GNCHEYRKLKAVETIQKCQHKLMGEDLESLNLKELQLEQLES
- GNCHEYRKLKAVETIQKCQHKLMGEDLESLNLKELQLEQLES
- QSSRNEYLRKAVVLQRTQHRLDGLSGEIKELEQLQELDLDS
- TNNSTEVNRLKAKIELERNORHVLGEDLQASLPKEQONLEQLQDLT

Original 5 MYA
50 MYA
60 MYA
60 MYA
100 MYA
**PSI-BLAST** (Position Specific Iterative BLAST)

- Designed to detect weak relationships
- The added sensitivity comes from the use of a **profile** that is constructed (automatically) from a multiple alignment.
- The profile is generated by calculating a Position-Specific Scoring Matrix (**PSSM**) for every position in the alignment. Also called profiles of **Hidden Markov Models**
  - **PSSM** are numerical representations of a multiple alignment
  - A highly conserved position receives a high score.
  - The profile is used to perform additional searches (iteration) and the results of each iteration used to refine the profile.
  - Each iteration uses a **PSSM** built from the previous iteration.
  - Continue search iteratively until no new matches are identified: "convergence".

**PSI-BLAST steps**
- BLASTP
- Multiple Alignment
- Construct PSSM
- Use PSSM to search

**Construction of a PSSM**

Each columns in the alignment is a row in the PSSM
Frequency of occurrence of a residue at each position
Calculate Pb of each aa at each position
T at position 8 conserved= highest score 150
P at position 9 less conserve= score 89
Note low scores of aromatic FYW relative to A at P row
Cluster analysis of the CCT domains

Exon 1

Cluster analysis of the CCT domains

Photoperiod response

I HvCO4

II HvCO7

III OsC

IV HvCO3

I OsB

II OsF

III OsHD1

IV OsH

ZCCT2 Ha

ZCCT2 Hb

ZCCT2 Tm

ZCCT2 Td

ZCCT1 Tm

G3116

ZCCT1 Tm DV92

ATCOL3

OsN

II ATCOL9

III ATCOL9

IV ATCOL9

OsG

II OsG

III OsG

IV OsG

0.1

0.2

0.3

0.4

0.5

0.6

0.7

0.8

0.9

1.0

1.1

1.2

1.3

1.4

1.5

1.6

1.7

1.8

1.9

2.0

2.1

2.2

2.3

2.4

2.5

2.6

2.7

2.8

2.9

3.0

3.1

3.2

3.3

3.4

3.5

3.6

3.7

3.8

3.9

4.0

4.1

4.2

4.3

4.4

4.5

4.6

4.7

4.8

4.9

5.0

5.1

5.2

5.3

5.4

5.5

5.6

5.7

5.8

5.9

6.0

6.1

6.2

6.3

6.4

6.5

6.6

6.7

6.8

6.9

7.0

7.1

7.2

7.3

7.4

7.5

7.6

7.7

7.8

7.9

8.0

8.1

8.2

8.3

8.4

8.5

8.6

8.7

8.8

8.9

9.0

9.1

9.2

9.3

9.4

9.5

9.6

9.7

9.8

9.9
# Multiple Sequence Alignment (MSA)

## 1. Uses of MSA

## 2. Technical difficulties

1. **Select sequences**
2. **Select objective function**
3. **Optimize the objective function**
   - 1. Exact algorithms
   - 2. Progressive algorithms
   - 3. Iterative algorithms
   - 4. Consistency-based algorithms

## 3. Tools to view alignments

1. MEGA
2. BOXSHADE & Seq. LOGOS
**Sequence selection**

- Use database searching to identify related proteins.
- Select **homologous** sequences.
- If sequences are similar along all the sequence use **Global MSA programs**.
- If not, PSI-Blast can help to define conserved regions.
- Or use **local MSA programs** (Gibbs sampler, Match-Box, or MACAW)
- Select a “representative” group of sequences. Large number of sequences are slow to compute and hard to analyze.
- The inclusion of a large number of a particular group of closely related proteins will dominate the alignment or profile (in spite of weighting schemes).
Multiple Sequence Alignment (MSA)

1. Uses of MSA
2. Technical difficulties
   1. Select sequences
   2. Select objective function
   3. Optimize the objective function
      1. Exact algorithms
      2. Progressive algorithms
      3. Iterative algorithms
      4. Consistency-based algorithms
3. Tools to view alignments
   1. MEGA
   2. BOXSHADE & Seq. LOGOS
Problem: Optimal pairwise alignment ≠ optimal multiple alignment

We need an Objective Function (OF) to measure which alignment is better.
Objective functions (OF)
Define the mathematical objective of the search

Most widely used MSA packages use the OF “sum-of-pairs”

- Define a mathematical optimum
- Use sum of scores from pairwise alignments and affine gaps
- Use a Mutation Data Matrix (e.g. Blosum 62)
- Some add weighting proportional to the information in the seq.

A biologically ideal OF should

- Maximize similarity
- Minimize the number of gaps (over their length)
- Retain conserved motifs and patterns
- Retain functionally important alignments
- Recapitulate phylogeny
- Concentrate on alignable regions, not in gapped regions
- Consider the limitations imposed by the 3D structure

It is a non-trivial task to test the biological correctness of an objective function.
Simplest OF: Sum-of-pairs (SP)

After the best MSA is obtained, non-alignable sequences and spaces facing spaces are removed and a score is calculated for the induced MAS using any chosen scoring scheme (distance or similarity).

<table>
<thead>
<tr>
<th>Seq. 1</th>
<th>AT-AATG</th>
<th>Induced Seq. 1-3 alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq. 2</td>
<td>CTGAG-G</td>
<td>Seq. 1 AT-AATG</td>
</tr>
<tr>
<td>Seq. 3</td>
<td>ATGAA-G</td>
<td>Seq. 3 ATGAA-G</td>
</tr>
</tbody>
</table>

**Distance scheme:**

```
Seq.1 AT-AATG
Seq.2 CTGAG-G
Seq.3 ATGAA-G
```

**Sum-of-pairs score:** The SP of a MSA is the sum of the scores of all the scores of the *induced* pairwise global alignments

```
<table>
<thead>
<tr>
<th>Seq. 1</th>
<th>AT-AATG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq. 2</td>
<td>CTGAG-G</td>
</tr>
<tr>
<td>Seq. 3</td>
<td>ATGAA-G</td>
</tr>
</tbody>
</table>
```

**Sum-of-pairs distance** = 4 + 3 + 2 = 9

**Weighted Sum-of-pairs score:** each score can be multiplied by a weight to reflect evolutionary distances.
New Objective Functions

**Multiple MSA:** Depending on the Mutation Data matrix selected and on the selected gap penalties (opening and extension) very different MSA will be obtained. *Which one is the correct one?*

<table>
<thead>
<tr>
<th>Seq.1</th>
<th>AT–AATG</th>
<th>Seq.1</th>
<th>ATAATG</th>
<th>Clustal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq.2</td>
<td>CTGAG–G</td>
<td>Seq.2</td>
<td>CTGAGG</td>
<td>Gap open= 11</td>
</tr>
<tr>
<td>Seq.3</td>
<td>ATGAA–G</td>
<td>Seq.3</td>
<td>ATGAAG</td>
<td>Gap ext.= 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Seq.1</th>
<th>AT–AATG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq.2</td>
<td>CTGAG–G</td>
</tr>
<tr>
<td>Seq.3</td>
<td>ATGAA–G</td>
</tr>
</tbody>
</table>

**New generation of Objective functions:** less sensitive to gap penalty estimations thanks to the incorporation of local information

- **Segment-to-segment comparisons** of the sequences (instead of character-to-character) without gap penalties is the strategy used by **DiAlign**. This approach is efficient for *local similarities*, (genomic DNA, many protein families)
  
  http://bibiserv.techfak.uni-bielefeld.de/dialign/

- **Consistency objective function:** (e.g. **T-Coffee & DiAlign**). The optimal MSA is defined as the one that agrees the most with all the optimal pair-wise alignments. “*Given a set of independent observations the most consistent are often closer to “the truth”.*
Multiple Sequence Alignment (MSA)

1. Uses of MSA

2. Technical difficulties
   1. Select sequences
   2. Select objective function
   3. Optimize the objective function
      1. Exact algorithms
      2. Progressive algorithms
      3. Iterative algorithms
      4. Consistency-based algorithms

3. Tools to view alignments
   1. MEGA
   2. BOXSHADE & Seq. LOGOS
Problem: Optimal multiple alignment is intractable for \( n \gg 2 \)

- Pairwise alignment: \( O(L^2) \)  →  2D-Pathmatrix
- \( n \)-wise alignment: \( O(L^n) \)  →  \( n \)-D-Path-hypercube

E.g. 10 GlnRS sequences – \( L = 200 \), \( n = 10 \): \( 10^{23} \) operations

Despite more than thirty years of history, multiple sequence alignment is still a topic of ongoing research in bioinformatics.

From Boris Steipe Univ. of Toronto
MSA: Exact algorithm

**MSA program**
- Multidimensional dynamic programming
- Optimizes sum-of-pairs
- More accurate than progressive methods
- BUT… Time proportional to $L^n$
- Practical to ~10 seq. of $L<200$-bp

**DCA (Divide & Conquer Algorithm)**
- Sits on top of MSA program
- Produces simultaneous MSA
- Cuts seq. in subsets, that are fed into MSA
- Practical to ~20-30 seq. of $L<200$-bp
- Easy WEB submission

**OMA (Optimal Multiple Alignment)**
- Iterative implementation of DCA
- Speeds up DCA
- Decreases memory requirements

Pairwise alignments bound search volume!

http://bibiserv.techfak.uni-bielefeld.de/dca/

http://bibiserv.techfak.uni-bielefeld.de/oma/
Multiple Sequence Alignment (MSA)

1. Uses of MSA

2. Technical difficulties
   1. Select sequences
   2. Select objective function
   3. Optimize the objective function
      1. Exact algorithms
      2. Progressive algorithms
      3. Iterative algorithms
      4. Consistency-based algorithms

3. Tools to view alignments
   1. MEGA
   2. BOXSHADE & Seq. LOGOS
Progressive algorithms (ClustalW, MultAlign, AMPS)

**CLUSTALW**
- Non-iterative & deterministic.
- OF: weighted sum-of-pairs.
- Affine gap penalties.
- Automatic substitution matrix choice.
- Most popular.
- Performs well in dense trees without obvious outliers (needs stepping stones).
- Can use SwissProt secondary structure information for gap penalty estimation.

**Example of Progressive algorithm**
- Calculate distances/similarities between sequences
- Construct a tree
- Add sequentially, following tree

http://www.ebi.ac.uk/tools/clustalw2
Problems with progressive algorithms

1. Bad decisions taken in the initial alignments will persist throughout the process and cannot be corrected

2. For numerous sequences may take long time to calculate tree. MUSCLE is a good rapid alternative
Multiple Sequence Alignment (MSA)

1. Uses of MSA
2. Technical difficulties
   1. Select sequences
   2. Select objective function
   3. **Optimize the objective function**
      1. Exact algorithms
      2. Progressive algorithms
      3. **Iterative algorithms**
      4. Consistency-based algorithms
3. Tools to view alignments
   1. MEGA
   2. BOXSHADE & Seq. LOGOS
Iterative algorithms

Recurrent modifications of suboptimal solutions

SAGA
- Uses a ‘Genetic Algorithm’
- Can use different objective functions (e.g. Coffee)
- Mutations randomly insertion or shift gaps
- Sequences can recombine
- Sequences evolve, higher OF scores survive

Hammer (also SAM)
- Build Hidden Markov Models based on seq.
- Align sequences to HMM
- http://hmmer.janelia.org/

Gibbs sampler (Local MSA)
- Finds un-gapped motifs
- Segments are removed or added to increase a P value

GAs and HMMs have been rather disappointing in ab initio alignments.
Better: Pre-compute MSA with other program and then use this ones for optimization
Multiple Sequence Alignment (MSA)

1. Uses of MSA

2. Technical difficulties
   1. Select sequences
   2. Select objective function
   3. Optimize the objective function
      1. Exact algorithms
      2. Progressive algorithms
      3. Iterative algorithms
      4. **Consistency-based algorithms**

3. Tools to view alignments
   1. MEGA
   2. BOXSHADE & Seq. LOGOS
1.- Consistency based algorithm.
2.- Constructs pairwise and multiple alignments by comparing whole segments of the sequences rather than individual residues.
3.- No gap penalty is used.
4.- Efficient where sequences are not globally related but share only local similarities.

**METHOD**

1.- Starts with identification of highly similar segment pairs
2.- Each pair is weighted by its $P$ values (similar to BLAST) and by a second score based on its compatibility with the complete set of segment pairs.
3.- The MSA is progressively assembled by adding the pairs of segments according to their weight.
**T-COFFEE: consistency-based algorithm**

**T-Coffee** *(Consistency Objective Function For alignment Evaluation)*

Version 2.00 and higher can mix sequences and structures

Local (Lalign) and global (ClustalW) pair-wise alignments can come from different programs and can be redundant

The EL is a position-specific substitution matrix (PSSM) where the score associated with each pair of residues depends on its compatibility with the rest of the library. This library replaces the Mutation data Matrix used in ClustalW (e.g. BLOSUM62).

Pair-wise distances are computed
A Neighbor joining tree is estimated
Sequences are aligned progressively following the topology of the tree

Progressive alignment

This figure indicates the layout of T-Coffee. Local and global pairwise alignments are first computed and then combined into a primary library that is extended in order to be used for computing the multiple sequence alignment in a progressive manner.
Expresso (3D-COFFEE): incorporates structures


Expresso is a multiple sequence alignment server that aligns sequences using structural information. The user provides sequences and the server runs BLAST to identify close homologues of the sequences within the PDB structure database. Good for sequences with low similarity (<40%) because STRUCTURE EVOLVES SLOWER THAN SEQUENCE!

1.- BLAST PDB structures (>60% identity, >70% coverage)
2.- LIBRARY CONSTRUCTION
   • Align sequences using lalign
   • Align templates to structures using SAP
   • Align structure induced alignments

PDB Oct/2010 = 68,288 structures
http://www.rcsb.org/pdb/
Benchmark tests (from Notredame 2002)

<table>
<thead>
<tr>
<th>Method</th>
<th>Ref1</th>
<th>Ref2</th>
<th>Ref3</th>
<th>Ref4</th>
<th>Ref5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>DiAlign</td>
<td>71.0</td>
<td>25.2</td>
<td>35.1</td>
<td>74.7</td>
<td>80.4</td>
<td>57.3</td>
</tr>
<tr>
<td>ClustalW</td>
<td>78.5</td>
<td>32.2</td>
<td>42.5</td>
<td>65.7</td>
<td>74.3</td>
<td>58.7</td>
</tr>
<tr>
<td>T-Coffee</td>
<td>80.7</td>
<td>37.3</td>
<td>52.9</td>
<td>83.2</td>
<td>88.7</td>
<td>68.7</td>
</tr>
</tbody>
</table>

BAliBASE reference database

http://bips.u-strasbg.fr/fr/Products/Databases/BAliBASE/prog_scores.html

Ref1: homogenous set of sequences
Ref2: homogenous group of sequences and an out-layer
Ref3: contains two distantly related groups of sequences.
Ref4: contains sequences that require long internal gaps
Ref5: contains sequences that require long-terminal gaps

ClustalW: performed well on Ref. sets 1-3, but poorly on 4-5 when long internal or terminal gaps are required.

When large gaps required T-Coffee and DiAlign perform better
Simulation tests (from Kumar & Filipski. 2007. Genome Research 17:127–135)

A) Free change and indels.
B) Constrain that 20% of DNA positions are occupied by conserved blocks (10 time slower evolution).

The homology accuracy is calculated as a fraction of simulated positions that were aligned correctly, and it is plotted against the fraction of sites different.
## Summary strategies

### Strategies for multiple alignment algorithms reflect biological heuristics

- **Exact**
  - Bounded optimal solution search:
    - "Maximize similarity, Minimize gaps"

- **Progressive**
  - Align most similar first, then add together
    - "Align according to phylogeny"

- **Consistency based**
  - Conserved regions guide alignment
    - "Retain conserved regions"

- **+/- Iterated**
  - Improve alignment from draft alignments
    - "Conserved regions of significance can be extended"

- **Combinations**
  - Empirical combinations of above

*From Boris Steipe Univ. of Toronto*
Alignment Visualization

Alignment Visualization: e.g. BOXSHADE

Web server or local installation ...

MSF or CLUSTAL ALN format ...

Absolutely Conserved
Mostly Conserved
Somewhat Conserved

BOXSHADE: http://www.ch.embnet.org/software/BOX_form.html
http://bioweb.pasteur.fr/seqanal/interfaces/boxshade.html

http://www.ch.embnet.org/software/BOX_form.html
Multiple Sequence Alignment

Representation of multiple alignments: the consensus sequence

DIVMTQSPSSL...
DIVMSQSPSSL...
DIVMTQSPTFL...
DVVMQTPTLL...
NIVLTQSPASL...
DIOMTOSTSSL...

DIVMTQSPSSL...

"Consensus sequence"

BUT ... this is lossy.

consider e.g. the TATAAT box:

TAT...

49% 58% 54%

Only ~5% of bacterial promoters actually have the sequence TATAAT.

From Boris Steipe Univ. of Toronto
**Sequence logos** are a graphical representation of a MSA. Each logo consists of stacks of symbols, one stack for each position in the sequence. The height of symbols within the stack indicates the relative frequency of each amino or nucleic acid at that position.

**WebLogo:** application to generate sequence logos
http://weblogo.berkeley.edu/
Multiple Sequence Alignment

At the end of the day ...

• Make sure you have the right sequences.
• Use more than one alignment method.
• Don't align parts of sequences that can't be aligned (because they are not homologous).
• Realize problems from multi-domain proteins.
• Above all, use your common sense.