Pairwise sequence alignment (global and local)

Multiple sequence alignment (global and local)

Sequence evolution models

Database searching

BLAST

Substitution matrices

Sequence statistics

Evolutionary tree reconstruction

Global Multiple Sequence Alignment

**HUMAN**
MKWVTISLL FLFSSAYSRG V...FRRDA.R KSEIAHRFND LGEEFPKALV

**RABIT**
MKWVTISLL FLFSSAYSRG V...FRREA.H KSEIAHRFND VGEHFPGLV

**PIG**
--WVTISLL FLFSSAYSRG V...FRRED.Y KSEIAHRFND LGEQYFPGLV

**CHICK**
MKWVTISFI FLFSSATSRG V..FRRDA.. KSEIAHRFND LEKETHFALV

Align k sequences, so that residues in each column share a property of interest:

- a common ancestor
- a structural or functional role

Examples...

---

**Protein Function**

**Rho-GAPs:** GTPase activator proteins

**Plant Rho-GAP**
Act as molecular switches in signal transduction.

**Mammalian Rho-GAP**

<table>
<thead>
<tr>
<th>Protein</th>
<th>Function</th>
<th>Plant</th>
<th>Mammalian</th>
</tr>
</thead>
<tbody>
<tr>
<td>OctaFH1</td>
<td>EELRHTGVEL GTKLDNAMV NKSHIGARALE EGRIFIACDG</td>
<td>EELRHTGVEL GTKLDNAMV NKSHIGARALE EGRIFIACDG</td>
<td></td>
</tr>
<tr>
<td>OmluFH1</td>
<td>GELKHRGMEL GAELEQASE ADALDSKAVS GEGLYAV-G</td>
<td>GELKHRGMEL GAELEQASE ADALDSKAVS GEGLYAV-G</td>
<td></td>
</tr>
<tr>
<td>Phypa7</td>
<td>MFLEIYAIOL EGLISIPPI LDGLDVRKCE FDIDCPLLTK</td>
<td>MFLEIYAIOL EGLISIPPI LDGLDVRKCE FDIDCPLLTK</td>
<td></td>
</tr>
<tr>
<td>Selmo2a/b</td>
<td>ATLOHEAELL DKLKLPFPV VHELYLRSL LQIQHQQEL</td>
<td>ATLOHEAELL DKLKLPFPV VHELYLRSL LQIQHQQEL</td>
<td></td>
</tr>
<tr>
<td>GgrGAP</td>
<td>HAVEINGSK -QGLPVYVGRVQKSSLI LMDPETRICA</td>
<td>HAVEINGSK -QGLPVYVGRVQKSSLI LMDPETRICA</td>
<td></td>
</tr>
<tr>
<td>HsRacGAP</td>
<td>REILELDGLK -EQLPGSFT THLEVEKVF F-DRDGEKAD</td>
<td>REILELDGLK -EQLPGSFT THLEVEKVF F-DRDGEKAD</td>
<td></td>
</tr>
<tr>
<td>Hs p85</td>
<td>EAIEKKGDK -ETLQPOIQS SN-LEVRQL L-DCDPFVD</td>
<td>EAIEKKGDK -ETLQPOIQS SN-LEVRQL L-DCDPFVD</td>
<td></td>
</tr>
<tr>
<td>Hs p50</td>
<td>ATLOHALLT -EGLRSAN TVQVREVQQ H-NMLPVDF</td>
<td>ATLOHALLT -EGLRSAN TVQVREVQQ H-NMLPVDF</td>
<td></td>
</tr>
</tbody>
</table>

---

Grunt et al. BMC Evolutionary Biology 2008 8:115

---

‘...replacement of the central arginine of the “arginine finger”... confirmed a local alteration in shape and charge of the conserved GTPase interaction interface, while the overall conformation of the molecule appears preserved.’
This plant Rho-GAP ‘might not function as a GTPase-activating protein, while its ability to bind a Rho-class GTPase may be retained.’

Reconstructing Evolutionary History

Early mammal

human

rat

mouse

Global Multiple Sequence Alignment

- Multiple alignments versus pairwise alignments
  - Global multiple sequence alignment
    - Formal definition
    - Sum-of-pairs scoring
  - An exact dynamic program algorithm
  - Heuristic “progressive alignment” strategy

RNA Secondary Structure

...CCGUGUACGUGUAACCGGAUCUUUAUCC... ...CCGUGUACGUGUAACCGGAUCUUUAUCC... ...CCGUGUACGUGUAACCGGAUCUUUAUCC... ...CCGUGUACGUGUAACCGGAUCUUUAUCC... ...CCGUGUACGUGUAACCGGAUCUUUAUCC... ...CCGUGUACGUGUAACCGGAUCUUUAUCC...
Observations

1. A multiple alignment induces pairwise alignments

2. A column in the induced pairwise alignment may contain all gaps, even though no column in the MSA contains all gaps.

3. The pairwise alignments induced by the optimal multiple alignment are not the same as the optimal pairwise alignments.

Optimal Pairwise Alignments

- (1) \text{ACT}
- (2) \text{AGT}

Optimal Multiple Alignment

- (1) \text{AC_T}
- (2) \text{A_GT}
- (3) \text{ACGT}

Although this costs more, it may be a biologically more realistic alignment.

Multiple alignments can identify patterns that are conserved in a family but not apparent in a pairwise alignment of two family members.

The pairwise alignments induced by the optimal MSA
- are not the same as the optimal pairwise alignments,
- are more biologically realistic.
Global Multiple Sequence Alignment

- Multiple alignments versus pairwise alignments
- Global multiple sequence alignment
  - Formal definition
  - Sum-of-pairs scoring
- An exact dynamic program algorithm
- Heuristic “progressive alignment” strategy

Given sequences $s_1...s_k$ of lengths $n_1...n_k$ seek $s'_1...s'_k$ of length $l \geq \max(n_i)$ such that

- Obtain $s_i$ from $s'_i$ by removing gaps
- No column contains all gaps
- The score of the alignment is optimal

Scoring function: Sum-of-Pairs

$$Score = \sum_{i=1}^{\ell} \sum_{a=1}^{k} \sum_{b>a} S(s'_a[i], s'_b[i])$$

(1) $\text{A} \_	ext{T} \text{T}$
(2) $\text{A} \_ \text{T}$
(3) $\text{AC} \text{A} \text{T}$

$S[\_\_\_] = 0$

$$Score = S[s_1s_2] + S[s_1s_3] + S[s_2s_3]$$
$$= 0 + g + g = 2g$$

Note: this example uses a similarity function. We can also use Sum-of-Pairs with distance scoring for global alignment.

Score = Sum-of-Pairs

$$Score = \sum_{i=1}^{\ell} \sum_{a=1}^{k} \sum_{b>a} S(s'_a[i], s'_b[i])$$

(1) $\text{A} \_ \text{T} \text{T}$
(2) $\text{A} \_ \text{T}$
(3) $\text{A} \text{C} \text{A} \text{T}$

$S[\_\_\_] = 0$

$$Score = S[s_1s_2] + S[s_1s_3] + S[s_2s_3]$$
$$= M + m + m = 2m + M$$

Note: this example uses a similarity function. We can also use Sum-of-Pairs with distance scoring.
Scoring function: Sum-of-Pairs

\[ \text{Score} = \sum_{i=1}^{k} \sum_{a=1}^{h} \sum_{b > a} S(s'_a[i], s'_b[i]) \]

(1) A_T T
(2) A_T
(3) ACAT

\[ S_{[\_\_]} = 0 \]

\[ \text{Score} = S[s_1 s_2] + S[s_1 s_3] + S[s_2 s_3] \]
\[ = g + M + g = 2g + M \]

Note: this example uses a similarity function. We can also use Sum-of-Pairs with distance scoring.

Sum-of-pairs scoring

Advantages: Fast, easy to use
Disadvantages:
• Assumes positional independence
• Over estimates the number of mutations

Global Multiple Sequence Alignment

• Multiple alignments versus pairwise alignments
• Global multiple sequence alignment
  ➢ An exact dynamic program algorithm
    • Calculating the alignment matrix for \( k \) sequences
      (see class notes)
    • Complexity
    • Heuristic “progressive alignment” strategy

Dynamic Programming for Multiple Alignment

Number of cells in matrix: \( O(n^k) \)
Each cell has \( O(2^k) \) neighboring cells
Calculating the sum-of-pairs score for each neighbor is \( O(k^2) \)

Total computational complexity:
\( O(n^k \cdot 2^k \cdot k^2) \)

MSA is NP-complete for Sum-of-Pairs scoring

Limits:
\( n = 500 \) residues
\( k = 8 - 10 \) sequences
Global Multiple Sequence Alignment

• Multiple alignments versus pairwise alignments
• Global multiple sequence alignment
  – Formal definition
  – Sum-of-pairs scoring
• An exact dynamic program algorithm
  ➢ Heuristic “progressive alignment” strategy

The progressive alignment heuristic for global multiple sequence alignment

Basic progressive alignment strategy

• Construct alignments for all pairs of sequences
• Compute $D$, a matrix of distances between all pairs
• Use $D$ to construct a “guide tree” $T$
• Construct MSA by
  – pairwise alignment of partial alignments (“profiles”)
  – guided by $T$
  – Using the “once a gap, always a gap” rule.
• Improve alignment by post-processing steps
Optimal Pairwise Alignments

(1) ACTCAT
(2) AGTCAT
(3) ACGTCCT

(1) ACTCAT
(2) AGTCAT
(3) ACGTCCT

\[ d(x, y) = 3 \]
\[ d(x, "\_") = 2 \]

Progressive Alignment

- Use profile alignment to merge sequences according to a guide tree.
- Typically, the most similar sequence pairs are merged first.

Merging strategy:

Align the profile (1,2) with sequence (3)

\[ \_ A C T C A T \]
\[ \_ A G T C A T \]
\[ \_ A C G T C C T \]

\[ d(x, y) = 3 \]
\[ d(x, "\_") = 2 \]
Note: no penalty for substitutions in the profile. We paid for those in a previous step.
\[ d(x, y) = 3 \]
\[ d(x, "\_\") = 2 \]
Two solutions:

ACGTCCT
A_CTCAT
A_GTCAT
ACGTCCT
AC_TCAT
AG_TCAT
Result:
Progressive alignment of the profile (1,2) with sequence (3)

(1) **ACTCAT**
(2) **AGTCAT**
(3) **ACGTCCT**

\[d(x,y) = 3\]
\[d(x, "\_\") = 2\]

**Progressive alignment**

- “Once a gap, always a gap”
  - You can’t go back and correct a bad decision at an earlier step.
- Progressive alignment is not guaranteed to give the optimal alignment.
- But it does have better complexity…

**Optimal Pairwise Alignments**

<table>
<thead>
<tr>
<th>Alignment</th>
<th>Progressive alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) <strong>ACTCAT</strong></td>
<td>(1,2) + (3)</td>
</tr>
<tr>
<td>(2) <strong>AGTCAT</strong></td>
<td>(1) <strong>AC_TCAT</strong> [4m+2g = 16]</td>
</tr>
<tr>
<td>(3) <strong>ACGTCCT</strong></td>
<td>(2) <strong>AG_TCAT</strong></td>
</tr>
</tbody>
</table>

An alternate alignment

<table>
<thead>
<tr>
<th>Alignment</th>
<th>Progressive alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) <strong>AC_TCAT</strong></td>
<td>(1) <strong>AC_TCAT</strong></td>
</tr>
<tr>
<td>(2) <strong>A_GTCAT</strong></td>
<td>(2) <strong>A_GTCAT</strong></td>
</tr>
<tr>
<td>(3) <strong>ACGTCCT</strong></td>
<td>(3) <strong>ACGTCCT</strong></td>
</tr>
</tbody>
</table>

\[4m+2g = 16\]

How would you score this case?
How would you score this case?
How would you score this case?
Complexity of progressive alignment

- Distance matrix
  - Each pairwise alignment $O(n^2)$
  - Number of pairwise alignments $O(k^2)$
- Iterative construction of MSA
  - Merging by profile alignment $O(k^2n^2)$
    (see class notes for details)
Entire method $O(k^2n^2)$
**Summary:**

**Progressive alignment heuristics**

- Not guaranteed to give the optimal MSA
- Bad choice of gaps propagates
- Complexity
  - Progressive: \( O(k^2n^2) \)
  - versus DP: \( O(n^k 2^k k^2) \)
- Typically, merge the most closely related sequences first.

---

**Mathematical correctness is not a guarantee of biological accuracy.**

The performance of MSA programs is typically evaluated using benchmarks based on biological data:

- Curated structural alignment
- Automated structural alignment
- Real or simulated sequence

---

**Various benchmarks are designed to mimic properties of different types of data sets encountered in practice, especially those that are challenging to align:**

- Highly divergent sequences, e.g., <50% or <30% identity
- A family of related sequences plus several outliers, or “orphan” sequences
- Related sequences that differ due to large N or C terminal extensions or large internal insertions or deletions
BAliBase: Reference MSAs based on structural alignment.

Note that implementation choices result in substantial differences in running time:

<table>
<thead>
<tr>
<th>Aligner</th>
<th>Performance</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIALIGN</td>
<td>57.2</td>
<td>12h, 25 min</td>
</tr>
<tr>
<td>CLUSTALW</td>
<td>58.9</td>
<td>2h, 57 min</td>
</tr>
<tr>
<td>T-Coffee</td>
<td>63.6</td>
<td>144h, 51 min</td>
</tr>
<tr>
<td>MUSCLE</td>
<td>64.8</td>
<td>3h, 11 min</td>
</tr>
<tr>
<td>MAFFT</td>
<td>64.8</td>
<td>2h, 36 min</td>
</tr>
<tr>
<td>ProbCons</td>
<td>66.9</td>
<td>19h, 41 min</td>
</tr>
<tr>
<td>ProbCons-ext</td>
<td><strong>68.0</strong></td>
<td>37h, 46 min</td>
</tr>
</tbody>
</table>

* Fraction of correctly aligned residue pairs

Do et al, Genome Research, 2005

Which program to choose?

Do and Katoh, 2008