Scoring Alignments

Genome 373
Genomic Informatics
Elhanan Borenstein
A quick review

- The computational bottleneck
  - Scale of biological data

- Complexity of tasks
A quick review: Informatic challenges

- Sequence comparison:
  - Find the best alignment of two sequences
  - Find the best match (alignment) of a given sequence in a large dataset of sequences
  - Find the best alignment of multiple sequences
- Motif and gene finding
- Relationship between sequences
  - Phylogeny
- Clustering and classification
- Many many many more ...
A quick review: Informatic challenges

- Sequence comparison:
  - Find the best alignment of two sequences
  - Find the best match (alignment) of a given sequence in a large dataset of sequences
  - Find the best alignment of multiple sequences

- Motif and gene finding

- Relationship between sequences
  - Phylogeny

- Clustering and classification

- Many many many more...
Motivation

• Why compare two DNA or protein sequences?
Motivation

• Why compare two DNA or protein sequences?
  – Determine whether they are descended from a common ancestor (homologous)
  – Infer a common function
  – Locate functional elements (motifs or domains)
  – Infer protein or RNA structure, if the structure of one of the sequences is known
  – Analyze sequence evolution
  – Infer the species from which a sequence originated
  – Quantify abundance/coverage
One of many commonly used tools that depend on sequence alignment.
Sequence Comparison Challenges

- Find the best *global* alignment of two sequences

- Find the best *global* alignment of multiple sequences

- Find the best *local (partial)* alignment of two sequences

- Find the best match (alignment) of a given sequence in a longer dataset of sequences
**Sequence Comparison Challenges**

- **✓** Find the best *global* alignment of two sequences
- **✗** Find the best *global* alignment of multiple sequences
- **✓** Find the best *local (partial)* alignment of two sequences
- **✓** Find the best match (alignment) of a given sequence in a longer dataset of sequences
Global Alignment Mission:
Find the best global alignment between two sequences.
Global Alignment Mission: Find the best global alignment between two sequences.

Find the best alignment of **GAATC** and **CATAC**:

<table>
<thead>
<tr>
<th>GAATC</th>
<th>GAAT–C</th>
<th>−GAAT–C</th>
<th>GAAT–C</th>
</tr>
</thead>
<tbody>
<tr>
<td>CATAC</td>
<td>C–ATAC</td>
<td>C–A–TAC</td>
<td>C–ATAC</td>
</tr>
<tr>
<td>GAATC–C</td>
<td>GAAT–C</td>
<td>GA–ATC</td>
<td>GAAT–C</td>
</tr>
<tr>
<td>CA–TAC</td>
<td>CA–TAC</td>
<td>CATA–C</td>
<td>CA–TAC</td>
</tr>
</tbody>
</table>

(some of a very large number of possibilities)
Global Alignment Mission: Find the best global alignment between two sequences.

Find the best alignment of **GAATC** and **CATAC**:

- **GAAT-C**
- **C-A-TAC**

Conceptually:

- What does a “correct” alignment mean?
- Correct vs. Best
Global Alignment Mission:
Find the best global alignment between two sequences.

Find the best alignment of **GAATC** and **CATAC**:

```
  GAAT--C
---A--TAC
```

**Technically:**

- This is a search (optimization) problem!!
- What do we need to solve this problem?
Global Alignment Mission:

Find the best global alignment between two sequences.

An algorithm for finding the alignment with the best score

A method for scoring alignments
Scoring Principles

- Score each locus independently.
- The alignment score will be the sum of the scores in all loci.
- Perfect Matches will get a positive (good) score.
- What about mismatches?
Scoring Principles

• Score each locus independently.
• The alignment score will be the sum of the scores in all loci.
• Perfect Matches will get a positive (good) score.
• What about mismatches?

<table>
<thead>
<tr>
<th>Purine</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pyrimidine</td>
<td>C</td>
<td>T</td>
</tr>
</tbody>
</table>

(transitions are typically about 2x as frequent as transversions in real sequences)
### Scoring Aligned Bases

- A reasonable **substitution matrix**:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10</td>
<td>-5</td>
<td>0</td>
<td>-5</td>
</tr>
<tr>
<td>C</td>
<td>-5</td>
<td>10</td>
<td>-5</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>-5</td>
<td>10</td>
<td>-5</td>
</tr>
<tr>
<td>T</td>
<td>-5</td>
<td>0</td>
<td>-5</td>
<td>10</td>
</tr>
</tbody>
</table>

What about gaps?

- **GAATC**
- **CATAC**

\[-5 + 10 + -5 + -5 + 10 = 5\]
What About Gaps?

- A reasonable substitution matrix:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10</td>
<td>-5</td>
<td>0</td>
<td>-5</td>
</tr>
<tr>
<td>C</td>
<td>-5</td>
<td>10</td>
<td>-5</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>-5</td>
<td>10</td>
<td>-5</td>
</tr>
<tr>
<td>T</td>
<td>-5</td>
<td>0</td>
<td>-5</td>
<td>10</td>
</tr>
</tbody>
</table>

What do gaps mean?

What if gaps have no penalty?

GAAT-C

CA-TAC

-5 + 10 + ? + 10 + ? + 10 = ?
Scoring Gaps?

- **Linear** gap penalty: every gap receives a score of $d$:

  \[
  \begin{align*}
  \text{GAAT} - \text{C} & \quad d = -4 \\
  \text{CA} - \text{TAC} & \\
  -5 + 10 + -4 + 10 + -4 + 10 = 17
  \end{align*}
  \]
Scoring Gaps?

- **Linear** gap penalty: every gap receives a score of $d$:
  
  $$\text{GAAT} - \text{C} \quad d = -4$$
  
  $$\text{CA} - \text{TAC}$$

  $$-5 + 10 + -4 + 10 + -4 + 10 = 17$$

- **Affine** gap penalty: opening a gap receives a score of $d$; extending a gap receives a score of $e$:

  $$\text{G} -- \text{AATC} \quad d = -4$$

  $$\text{CATA} -- \text{C} \quad e = -1$$

  $$-5 + -4 + -1 + 10 + -4 + -1 + 10 = 5$$
**Same Method Applies to AA**

**BLOSUM62 Score Matrix**

|   | A | R | N | D | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V | B | Z | X |
| A | 4 | -1 | -2 | -2 | 0 | -1 | -1 | 0 | -2 | -1 | -1 | -1 | -1 | -1 | -1 | 1 | 0 | -3 | -2 | 0 | -2 | -1 | 0 | -4 |
| R | -1 | 5 | 0 | -2 | -3 | 1 | 0 | -2 | 0 | -3 | -2 | 2 | -1 | -3 | -2 | -1 | -1 | -3 | -2 | -3 | -1 | 0 | -1 | -4 |
| N | -2 | 0 | 6 | 1 | -3 | 0 | 0 | 0 | 1 | -3 | -3 | 0 | -2 | -3 | -2 | 1 | 0 | -4 | -2 | -3 | 3 | 0 | -1 | -4 |
| D | -2 | -2 | 1 | 6 | -3 | 3 | 0 | 2 | -1 | -1 | -3 | -4 | -1 | -3 | -3 | -1 | 0 | 1 | -4 | -3 | -3 | 4 | 1 | -1 | -4 |
| C | 0 | -3 | -3 | -3 | 9 | -3 | -4 | -3 | -3 | -1 | -1 | -3 | -1 | -2 | -3 | -1 | -1 | -2 | -2 | -1 | -3 | -3 | -2 | -4 |
| Q | -1 | 1 | 0 | 0 | -3 | 5 | 2 | -2 | 0 | -3 | -2 | 1 | 0 | -3 | -1 | 0 | 1 | -2 | -1 | -2 | 0 | 3 | -1 | -4 |
| E | -1 | 0 | 0 | 2 | -4 | 2 | 5 | -2 | 0 | -3 | -3 | 1 | -2 | -3 | -1 | 0 | 1 | -3 | -2 | -2 | 1 | 4 | -1 | -4 |
| G | 0 | -2 | 0 | -1 | -3 | -2 | -2 | 6 | 2 | -4 | -4 | -2 | -3 | -3 | -2 | 0 | 2 | -2 | -3 | -3 | -1 | -2 | -1 | -4 |
| H | -2 | 0 | 1 | -1 | -3 | 0 | 0 | 2 | 8 | -3 | -3 | -1 | -2 | -1 | -2 | -1 | 2 | -2 | 2 | -2 | 3 | 0 | 0 | -1 | -4 |
| I | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -3 | 4 | 2 | -3 | 1 | 0 | -3 | -2 | -1 | -3 | -1 | 3 | -3 | -3 | -3 | -1 | -4 |
| L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | -2 | 4 | -2 | 2 | 0 | -3 | -2 | -1 | -2 | -1 | 1 | 4 | -3 | -1 | -4 |
| K | -1 | 2 | 0 | -1 | -3 | 1 | 1 | 2 | -1 | -3 | -2 | 5 | -1 | -3 | -1 | 0 | 1 | -3 | -2 | -2 | 0 | 1 | -1 | -4 |
| M | -1 | -1 | -2 | -3 | -1 | 0 | 2 | -3 | -2 | 1 | 2 | -1 | 5 | 0 | 2 | -1 | -1 | -1 | -1 | 1 | 3 | -1 | -3 | -1 |
| F | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0 | 0 | 0 | 3 | 0 | 6 | -4 | -2 | -2 | 1 | 3 | -1 | -3 | -3 | -1 |
| P | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | -1 | -1 | -4 | -3 | -2 | -2 | -1 | -2 | -4 |
| S | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | -1 | 2 | -2 | 0 | -1 | -2 | 1 | 4 | 1 | -3 | -2 | -2 | 0 | 0 | 0 | -4 |
| T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | 1 | 5 | -2 | -2 | 0 | -1 | 1 | 0 | -4 |
| W | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -3 | -2 | -3 | -1 | 1 | 4 | -3 | -2 | 1 | 1 | 2 | -3 | -4 | -3 | -2 | -4 |
| Y | -2 | -2 | -3 | -2 | -1 | -2 | -3 | -2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | -1 | -3 | -2 | -1 | -4 |
| V | 0 | -3 | -3 | -3 | -1 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | 1 | -2 | -2 | 0 | 3 | -1 | 4 | -3 | -2 | -1 | -4 |
| B | -2 | -1 | 3 | 4 | -3 | 0 | 1 | -1 | 0 | -3 | -4 | 0 | -3 | -3 | -2 | 0 | 1 | -4 | -3 | -3 | 4 | 1 | -1 | -4 |
| Z | -1 | 0 | 0 | 1 | -3 | 3 | 4 | -2 | 0 | -3 | -3 | 1 | -1 | -3 | -1 | 0 | 1 | -3 | -2 | -2 | 1 | 4 | -1 | -4 |
| X | 0 | -1 | -1 | -1 | -2 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | 0 | 2 | -1 | -1 | -1 | -1 | -1 | -4 |

**Y mutates to V receives -1**
**M mutates to L receives 2**
**E gets deleted receives -10**
**G gets deleted receives -10**
**D matches D receives 6**

**Total score = -13**

**YMEGDLEIAPDAK**

**VL--DKELSPDGT**
Global Alignment Mission:
Find the best global alignment between two sequences.

- An algorithm for finding the alignment with the best score
- A method for scoring alignments

### Substitution matrix:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10</td>
<td>-5</td>
<td>0</td>
<td>-5</td>
</tr>
<tr>
<td>C</td>
<td>-5</td>
<td>10</td>
<td>-5</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>-5</td>
<td>10</td>
<td>-5</td>
</tr>
<tr>
<td>T</td>
<td>-5</td>
<td>0</td>
<td>-5</td>
<td>10</td>
</tr>
</tbody>
</table>

### Gap penalty:
- Linear gap penalty
- Affine gap penalty

Example:

- Sequence 1: GAAT
- Sequence 2: C

- Sequence 1: CA
- Sequence 2: TAC

Distance:

\[-5 + 10 + -4 + 10 + -4 + 10 = 17\]

Distance = 17
A simple algorithm ....

• **Align the two sequences: GAATC and CATA**

```
GAATC   GAAT-C  -GAAT-C  GAAT-C
CATA    C-ATAC  C-A-TAC  C-ATAC
GAATC-  GAAT-C  GA-ATC   GAAT-C
CA-TAC  C-ATAC  CATA-C   CA-TAC
```

**Simple (exhaustive search) algorithm**

1) Construct all possible alignments
2) Use the substitution matrix and gap penalty to score each alignment
3) Pick the alignment with the best score
How many possibilities?

- Align the two sequences: GAATC and CATAC

- How many different possible alignments of two sequences of length \( n \) exist?
How many possibilities?

- **Align the two sequences: GAATC and CATAC**

<table>
<thead>
<tr>
<th>GAATC</th>
<th>GAAT-C</th>
<th>-GAAT-C</th>
<th>GAAT-C</th>
</tr>
</thead>
<tbody>
<tr>
<td>CATAC</td>
<td>C-ATAC</td>
<td>C-A-TAC</td>
<td>C-ATAC</td>
</tr>
<tr>
<td>GAATC-CA-TAC</td>
<td>GAAT-C</td>
<td>GA-ATC</td>
<td>GAAT-C</td>
</tr>
<tr>
<td>CA-TAC</td>
<td>CA-TAC</td>
<td>CATA-C</td>
<td>CA-TAC</td>
</tr>
</tbody>
</table>

- How many different possible alignments of two sequences of length \( n \) exist?

<table>
<thead>
<tr>
<th>( n )</th>
<th>( \times10^\text{ } )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2.5x10^2</td>
</tr>
<tr>
<td>10</td>
<td>1.8x10^5</td>
</tr>
<tr>
<td>20</td>
<td>1.4x10^11</td>
</tr>
<tr>
<td>30</td>
<td>1.2x10^17</td>
</tr>
<tr>
<td>40</td>
<td>1.1x10^23</td>
</tr>
</tbody>
</table>
Mission:

Find the best alignment between two sequences.

- Needleman–Wunsch Algorithm
- Dynamic programming

A algorithm for finding the alignment with the best score

A method for scoring alignments

Substitution matrix:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10</td>
<td>-5</td>
<td>0</td>
<td>-5</td>
</tr>
<tr>
<td>C</td>
<td>-5</td>
<td>10</td>
<td>-5</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>-5</td>
<td>10</td>
<td>-5</td>
</tr>
<tr>
<td>T</td>
<td>-5</td>
<td>0</td>
<td>-5</td>
<td>10</td>
</tr>
</tbody>
</table>

Gap penalty:
- Linear gap penalty
- Affine gap penalty

GAAT-C

CA-TAC

\[ d = -4 \]

\[ -5 + 10 + -4 + 10 + -4 + 10 = 17 \]