Sequence Analysis (part II)

BBSI 2006: Lecture $(\chi+2)$

Takis Benos (2006)

Outline

• Sequence variation
• Distance measures
• Scoring matrices
• Pairwise alignments (global, local)
• Database searches (BLAST, FastA)
• Multiple sequence alignments

Sequence Variations
Sequence variation

- Base mutations: the source of sequence variation

Purines

A \[\rightarrow\] G

Pyrimidines

C \[\rightarrow\] T

Transitions

Transversions

Sequence variation (cntd)

tggagc\(\rightarrow\)t
att\(\rightarrow\)aatgctaagt
acatt\(\rightarrow\)cc
tgagtt\(\rightarrow\)a
atg\(\rightarrow\)aat
cc
tgagtt\(\rightarrow\)a
atg\(\rightarrow\)aat
cc
tgagtt\(\rightarrow\)a
atg\(\rightarrow\)aat
cc

silent
missense
nonsense

tggagc\(\rightarrow\)t
att\(\rightarrow\)aatgctaagt
acatt\(\rightarrow\)cc
tgagtt\(\rightarrow\)a
atg\(\rightarrow\)aat
cc
tgagtt\(\rightarrow\)a
atg\(\rightarrow\)aat
cc
tgagtt\(\rightarrow\)a
atg\(\rightarrow\)aat
cc

deletion
**Sequence variation (cntd)**

![Image of sequence variation graph]

Figure 3. Average rates of substitution in different parts of genes and in non-coding regions.


---

**Distance measures**

---

**Nucleic acid distances**

- No selection - no correction:

\[ D = \frac{k}{N} \]

![Diagram of nucleic acid distances]

Nucleic acid distances (cntd)
• Jukes-Cantor correction:
\[ D_{JC} = -0.75 \ln (1 - D/0.75) \]


Nucleic acid distances (cntd)
• Kimura’s 2-parameter model:
\[ D_{K2P} = -0.5 \ln (1 - 2P -2Q) - 0.25 \ln (1 - 2Q) \]


Scoring matrices
Nucleic acid distances (cntd)

- Nucleotide substitution matrices.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-4</td>
<td>5</td>
<td>-4</td>
<td>-4</td>
</tr>
<tr>
<td>T</td>
<td>-4</td>
<td>5</td>
<td>-4</td>
<td>-4</td>
</tr>
<tr>
<td>C</td>
<td>-4</td>
<td>-4</td>
<td>5</td>
<td>-4</td>
</tr>
<tr>
<td>G</td>
<td>-4</td>
<td>-4</td>
<td>-4</td>
<td>5</td>
</tr>
</tbody>
</table>

Identity

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

BLAST

Transition/Transversion

Amino acid distances: PAM

- Percent Accepted Mutations (PAM) matrices:
  - Frequency substitution matrix from aligned sequences (Dayhoff, 1978).
  - M(i,j): no. of a.a. i to j mutations
  - 71 groups of closely related proteins (why?); 1,572 changes.
  - PAMn: the aligned sequences have n a.a. substitutions per 100 residues.

Amino acid distances: PAM (cntd)

- Assumptions of the PAM model:
  - Replacement at any site depends only on the a.a. on that site, given the mutability table.
  - Sequences in the training set (and those compared) have average a.a. composition.
Amino acid distances: PAM (cntd)

**Score(i,j) = \log_{10} \frac{M(i,j)}{f(i)}**

Sources of error in the PAM model:

- Many proteins depart from the average a.a. composition.
- The a.a. composition can vary even within a protein (e.g., transmembrane proteins).
- A.a. positions are not “mutated” equally probably; especially in long evolutionary distances.

Sources of error in the PAM model (cntd):

- Rare replacements are observed too infrequently and...
- ...errors in PAM1 are magnified in PAM250.
A.a. distances: BLOSUM

- Blocks Substitution Matrices (BLOSUM):
- Log-likelihood matrix (Henikoff & Henikoff, 1992)
- BLOCKS database of aligned sequences used as primary source set.

A.a. distances: BLOSUM (cntd)

- Weighted contribution of similar(*) sequences in order to reduce redundancy.
- BLOSUM62 is more closely related to PAM120.

(*) n% similar; the n in BLOSUMn
**A.a. distances: BLOSUM (cntd)**

```
<table>
<thead>
<tr>
<th></th>
<th>BLOSUM 60</th>
<th>BLOSUM 62</th>
<th>BLOSUM 45</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less divergent</td>
<td>PAM 1</td>
<td>PAM 120</td>
<td>PAM 250</td>
</tr>
<tr>
<td>More divergent</td>
<td>BLOSUM 60</td>
<td>BLOSUM 62</td>
<td>BLOSUM 45</td>
</tr>
</tbody>
</table>
```


**Substitution matrices: comparison**

- **PAM vs BLOSUM**

- Matrices of choice:
  - BLOSUM62: the all-weather matrix
  - PAM250: for distant relatives

**Substit. matrices: comparison (cntd)**

- **PAM vs BLOSUM (cntd)**

  - Lower PAM/higher BLOSUM matrices identify shorter local alignments of highly similar sequences

  - Higher PAM/lower BLOSUM matrices identify longer local alignments of more distant sequences

### Substit. matrices: comparison (cntd)

#### PAM10

| A  | R  | N  | D  | C  | Q  | E  | G  | H  | I  | L  | K  | M  | F  | P  | S  | T  | W  | Y  | V |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 7  | -10| 9  | -7 | -9 | 9  | -6 | -17| -1 | 8  | -5 | -15| -5 | 0  | -20| -1 | 8  | -4 | -13| -6 | -7 | 7  |

#### PAM250

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

#### BLOSUM62

| A  | R  | N  | D  | C  | Q  | E  | G  | H  | I  | L  | K  | M  | F  | P  | S  | T  | W  | Y  | V |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 4  | -2 | 6  | -2 | -3 | 3  | -3 | -3 | 3  | -3 | -3 | -3 | -3 | -3 | -3 | -3 | -3 | -3 | -3 | -3 | -3 |
Pairwise alignments

Alignment: the problem
Given two sequences, $S$ and $T$, and a scoring matrix find their relative arrangement with the highest “score”.

Seq. #1: G A A T T C A G T T A
Seq. #2: G G A T C G A

Alignment: the problem (cntd)
Alignment: the problem (cntd)

- Scoring schemes: three possible situations...
  - Match
  - Mismatch
  - Gap
    - Gap initiation
    - Gap extension

How much??

Alignment: a naïve approach

```
Seq #1
 G A A T T C A G T T A
1   1   1   1   1
1   1   1   1
11  1   1   1
1   1
1   1   1
1   1
1   1
1   1
1   1
1   1
1   1
```

```
Seq #2
 G A T C
1   1   1
1   1
1   1
1   1
```

Insertion seq. #1

```
G A A T T C - A G T T A
|   |   |
G G A - T - C G A
```

Alignment: a naïve approach

```
Seq #1
 G A A T T C A G T T A
1   1   1   1   1
1   1   1   1
11  1   1   1
1   1
1   1   1
1   1
1   1
1   1
1   1
1   1
1   1
```

```
Seq #2
 G A T C
1   1   1
1   1
1   1
```

```
G A A T T C - A G T T A
|   |   |
G G A - T - C G A
```

Alignment: a naïve approach (cntd)

The formula:

\[ M_{ij} = \text{MAXIMUM}\{ \]

\[ M_{i-1,j-1} + S_{ij} \text{ (match/mismatch in the diagonal)}, \]

\[ M_{i,j-1} + w \text{ (gap in sequence \#1)}, \]

\[ M_{i-1,j} + w \text{ (gap in sequence \#2)} \]

\} \]

In the following example, the score for match is 1 and for mismatch and gap is 0.

Alignment: adding scores (cntd)

• In each step we need to keep track only the scores of the \((i,j)\) position and its immediate neighbours: \((i-1,j-1)\), \((i-1,j)\) and \((i,j-1)\).

• We backtrack from the right-down corner to find the actual alignment.
Alignment: adding scores (cntd)

\[ S(1,1) = \max\{ \\
S(0,0) + 1 = 1, \\
S(0,1) + w = 0, \\
S(1,0) + w = 0, \\
S(1,0) + w = 0\} = 1 \]

Source:
http://www.sbc.su.se/~per/molbioinfo2001/dynprog/dynamic.html

Alignment: adding scores (cntd)

Alignment:

(Seq #1) A

Alignment:

(Seq #2) A

Source:
http://www.sbc.su.se/~per/molbioinfo2001/dynprog/dynamic.html
Alignment: adding scores (cntd)

Alignment:

Seq #1: T A
Seq #2: - A

Source: http://www.sbc.su.se/~per/molbioinfo2001/dynprog/dynamic.html

Alignment: another example

The formula:

\[ M_{ij} = \text{MAXIMUM} \{
\begin{align*}
M_{i-1,j-1} + S_{ij} & \quad \text{(match/mismatch in the diagonal)}, \\
M_{i,j-1} + w & \quad \text{(gap in sequence #1)}, \\
M_{i-1,j} + w & \quad \text{(gap in sequence #2)}
\end{align*}\}

New scores: 2 for match, -1 for mismatch and -2 for gap.
Alignment: another example (cntd)

Source:
http://www.sbc.su.se/~per/molbioinfo2001/dynprog/adv_dynamic.html
Alignment: another example (cntd)

(Seq #1) G A A T C A G T T A
|   |   | |   |     |
(Seq #2) G G A - T C - G - - A

Alignment:

Source:
http://www.sbc.su.se/~per/molbioinfo2001/dynprog/adv_dynamic.html

Global alignment

\[
M_{ij} = \max \left\{ 
M_{i-1,j-1} + \text{Score}(S_i,T_j), 
0, 
M_{i,j-1} + \text{W}, 
M_{i-1,j} + \text{W}
\right\}
\]

Needleman & Wunsch, 1970

Local alignment

\[
M_{ij} = \max \left\{ 
M_{i-1,j-1} + \text{Score}(S_i,T_j), 
0, 
M_{i,j-1} + \text{W}, 
M_{i-1,j} + \text{W}
\right\}
\]

Smith & Waterman, 1981
Local alignment

Given two sequences, \( S \) and \( T \), find two subsequences, \( s \) and \( t \), whose alignment has the highest “score” amongst all subsequence pairs.

Why do we need local alignment, if we have the global one?

Local alignment: an example

EGR4_HUMAN   KA [FACPVESCVRSFARSDELNRHLRIH] TGHKP [FQCRICLRNFSRSDHLSHVRTH] TGEKP [FACDV--CGRRFA

EGR4_RAT     KA [FACPVESCVRTFARSDELNRHLRIH] TGHKP [FQCRICLRNFSRSDHLSHVRTH] TGEKP [FACDV--CGRRFA

EGR3_HUMAN   RP [HACPAEGCDRRFSRSDELTRHLRIH] TGHKP [FQCRICMRNSFSRSDHLSHVRTH] TGEKP [FACEF--CGRKFA

EGR3_RAT     RP [HACPAEGCDRRFSRSDELTRHLRIH] TGHKP [FQCRICMRNSFSRSDHLSHVRTH] TGEKP [FACEF--CGRKFA

EGR1_HUMAN   RP [YACPVESCDRRFSRSDELTRHIRIH] TGQKP [FQCRICMRNSFSRSDHLSHVRTH] TGEKP [FACDI--CGRKFA

EGR1_MOUSE   RP [YACPVESCDRRFSRSDELTRHIRIH] TGQKP [FQCRICMRNSFSRSDHLSHVRTH] TGEKP [FACDI--CGRKFA

EGR1_RAT     RP [YACPVESCDRRFSRSDELTRHIRIH] TGQKP [FQCRICMRNSFSRSDHLSHVRTH] TGEKP [FACDI--CGRKFA

EGR1_BRARE   RP [YACPVETCDRRFSRSDELTRHIRIH] TGQKP [FQCRICMRNSFSRSDHLSHVRTH] TGEKP [FACEI--CGRKFA

EGR2_RAT     RP [YPCPAEGCDRRFSRSDELTRHIRIH] TGHKP [FQCRICMRNSFSRSDHLSHVRTH] TGEKP [FACDY--CGRKFA

EGR2_XENLA   RP [YPCPAEGCDRRFSRSDELTRHIRIH] TGHKP [FQCRICMRNSFSRSDHLSHVRTH] TGEKP [FACDY--CGRKFA

EGR2_MOUSE   RP [YPCPAEGCDRRFSRSDELTRHIRIH] TGHKP [FQCRICMRNSFSRSDHLSHVRTH] TGEKP [FACDY--CGRKFA

EGR2_HUMAN   RP [YPCPAEGCDRRFSRSDELTRHIRIH] TGHKP [FQCRICMRNSFSRSDHLSHVRTH] TGEKP [FACDY--CGRKFA

MIG1_KLULA   -- [-------------------------] ---RP [YVCPICQRGFHELHRIRIHM] TGERP [HACDFPGCSKRFSRSDHLSHVRTH] TGEKP [HACDFPGCVKRFSRSDHLSHVRTH]

MIG1_KLUMA   -- [-------------------------] ---RP [YMCPICHRGFHELHRIRIHM] TGERP [HACDFPGCAKRFSRSDHLSHVRTH] TGEKP [HACDFPGCAKRFSRSDHLSHVRTH]

MIG1_YEAST   -- [-------------------------] ---RP [HACPICHRAFH] TGERP [HACDFPGCVKRFSRSDHLSHVRTH] TGEKP [HACDFPGCVKRFSRSDHLSHVRTH]

MIG2_YEAST   -- [-------------------------] ---RP [FRCDTCHRGFHRLRTH] TGERP [HACDFPGCVKRFSRSDHLSHVRTH] TGEKP [HACDFPGCVKRFSRSDHLSHVRTH]

Local vs. global alignment


Local vs. global alignment (cntd)

- Characteristics of local alignments:
  - The alignment can start/end at any point in the matrix.
  - No negative scores.
  - The mean value of the scoring matrix (e.g. PAM, BLOSUM) should be negative.
  - There should be positive scores in the scoring matrix.

Local alignment (cntd)

- Characteristics of local alignments:
  - The alignment can start/end at any point in the matrix.
  - No negative scores.
  - The mean value of the scoring matrix (e.g. PAM, BLOSUM) should be negative.
  - There should be positive scores in the scoring matrix.