Sequence Alignment and Dynamic Programming

J. R. Quine
Department of Mathematics
Part of this lecture is taken from *Calculating the secrets of life*, Chapter 3: *Seeing conserved signals: using algorithms to detect similarities between biosequences*, by Eugene W. Meyers. You can print low resolution copies of each page directly from this link. Another reference for this material is *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, by Richard Durbin.
Eugene Meyers

- Eugene Meyers was responsible for the computations by *Celera Genomics* for completing the Human Genome Project.
- Read the first 6 pages of his article in Chapter 3 of *Calculating the secrets of life*. This is a good introduction to dynamic programming.
Mutation in DNA is a natural evolutionary process.
Less often a nucleotide is deleted or inserted.
Some mutations are lethal, some are favorable, some have little impact.
Sequence similarity

We are looking for similarities between nucleotide or amino acid sequences. Some possible implications of sequence similarity are

- Proteins have a common evolutionary origin
- Proteins have a similar function
Improving the Score Using Gaps (Without Penalty)

1. ATTACG
   ATATCG

2. ATTA-CG
   A-TATCG

3. AT-TACG
   ATAT-CG

The scores are 4, 5, 5 respectively
Important concepts

- Scoring (or substitution) matrix $\delta$. The simplest is the unit scoring matrix as in dot matrix techniques.

- Total score of alignment, $\Sigma_i \delta(a_i, b_i)$.

- Local alignment, global alignment

- Edit graph. This helps to formalize the ideas of dynamic programming
Unit cost scoring scheme

<table>
<thead>
<tr>
<th>$\delta$</th>
<th>$-$</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>$-$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>0</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>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

**FIGURE 3.1** Unit-cost scoring scheme.
The edit graph

- $(M+1)(N+1)$ vertices
- A-gap edges with score $\delta(a_i, -)$
- B-gap edges with score $\delta(-, b_i)$
- Alignment edges with score $\delta(a_i, b_j)$. 
Rule for Scoring Each Vertex in the Edit Graph

\[ S(i, j) = \max \{ \\
S(i - 1, j - 1) + \delta(a_i, b_j), \quad \]
\[ S(i - 1, j) + \delta(a_i, -), \quad \]
\[ S(i, j - 1) + \delta(-, b_j) \} \]
Scores In The Edit Graph

\[ S(i-1, j-1) \]

\[ \delta(a_i, b_j) \]

\[ S(i, j) \]

\[ \delta(a_i, -) \]

\[ \delta(-, b_j) \]
The Completed Edit Graph

Completed graph
Pseudo-code for dynamic programming

0. var \( S \): array \([0..M,0..N]\) of real
1. \( S[0,0] \leftarrow 0 \)
2. for \( j \leftarrow 1 \) to \( N \) do
3. \( S[0,j] \leftarrow S[0,j-1] + \delta(-,b_j) \)
4. for \( i \leftarrow 1 \) to \( M \) do
5. \{ \( S[i,0] \leftarrow S[i-1,0] + \delta(a_i,-) \)
6. for \( j \leftarrow 1 \) to \( N \) do
7. \( S[i,j] \leftarrow \max \{ S[i-1,j-1] + \delta(a_i,b_j), S[i-1,j] + \delta(a_i,-), S[i,j-1] + \delta(-,b_j) \} \)
8. \}
9. write “Maximum score is” \( S[M,N] \).

FIGURE 3.3 The classical dynamic programming algorithm.
Dynamic Programming

As we see from the above the Meyers article, there is an algorithm called *dynamic programming*, which gives us the highest scoring alignment between two sequences. The algorithm runs in $O(MN)$ time where $M$ and $N$ are the lengths of the two sequences.
Dynamic Programming

- Dynamic programming is a general computational paradigm of wide applicability. A problem can be solved by dynamic programming if the final answer can be determined by computing a tableau of answers to progressively larger subproblems.
Alignment Algorithms

- Needleman-Wunsch
  - Global alignment
- Smith-Waterman
  - Local alignment
Who Invented Dynamic Programming?

from *Introduction to Computational Biology (Maps, sequences and genomes)* by Michael S. Waterman:
Who Invented Dynamic Programming?

- Needleman and Wunsch (1970) wrote a paper titled *A general method applicable to the search for similarities in the amino acid sequence of two proteins*. It was surely unknown to the authors that their method fit into a broad class of algorithms introduced by Richard Bellman under the name dynamic programming.
Gap Penalty

- The above unit scoring scheme scored zero for matching with a gap. Usually there is a negative score for matching with a gap, $S(a,-) = S(-,a) = -d$. The number $d$ is called the \textit{gap penalty}. This will decrease the occurrence of gaps in optimal alignments.
Homework Problem

- Calculate the dynamic programming matrix and an optimal alignment for the DNA sequences GAATTC and GATTA, scoring
  - +2 for a match
  - -1 for a mismatch
  - -2 for a gap (2 is the gap penalty)
- (Note: Do this twice for the different ways to score a gap at the beginning or end of the sequence)
Amino Acid Sequences

- For proteins we work with strings from a 20+ letter alphabet
- A Ala  Alanine
- R Arg  Arginine
- N Asn  Asparagine
- D Asp  Aspartic acid
- C Cys  Cysteine
Amino Acid Sequences

- Q Gln  Glutamine
- E Glu  Glutamic acid
- G Gly  Glycine
- H His  Histidine
- I Ile  Isoleucine
- L Leu  Leucine
Amino Acid Sequences

- K Lys  Lysine
- M Met  Methionine
- F Phe  Phenylalanine
- P Pro  Proline
- S Ser  Serine
- T Thr  Threonine
Amino Acid Sequences

- W Trp  Tryptophan
- Y Tyr   Tyrosine
- V Val   Valine
- B Asx   Aspartic acid or Asparagine
- Z Glx   Glutamic acid or Glutamine
- X Xaa   Any amino acid
Scoring Matrices for Amino Acids

- We want a *scoring matrix* or *substitution matrix* $S$ for amino acids. A scoring matrix should reflect amino acid properties. A better score should result if amino acids with similar properties are aligned. So $S(a,b)$ should be positive if the residues are very similar and negative if very unsimilar.
# Amino Acid Properties

<table>
<thead>
<tr>
<th>Type of Amino Acid</th>
<th>Properties</th>
<th>Amino Acids</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amino acids with aliphatic hydrophobic side chains</td>
<td>The hydrophobic side chains of these amino acids will not form hydrogen bonds or ionic bonds with other groups. These hydrophobic amino acids tend to be buried in the centre of proteins away from the surrounding aqueous environment.</td>
<td>Ala, Val, Leu, Ile, Met, Pro, Phe, Trp.</td>
</tr>
<tr>
<td>Amino acids with uncharged but polar side chains</td>
<td>The side chains of these amino acids are uncharged at physiological pH.</td>
<td>Ser, Tyr, Asp, Gln, Cys.</td>
</tr>
<tr>
<td>Amino acids with acidic side chains</td>
<td>These have a carboxylic acid group in their side chain and are very hydrophilic.</td>
<td>Asp, Glu.</td>
</tr>
<tr>
<td>Amino acids with basic side chains</td>
<td>The positive charge on these side chains makes them hydrophilic and they are likely to be found at the protein surface</td>
<td>Lys, Arg, His.</td>
</tr>
<tr>
<td>Neutral side chain</td>
<td>The single hydrogen atom side chain has no strong hydrophobic or hydrophilic properties.</td>
<td>Gly</td>
</tr>
</tbody>
</table>
Scoring Matrix

- Should the scoring matrix be decided by scientists with a knowledge of biochemistry, or should it be computed from an analysis of the current database of sequences?

- Bioinformatics or Biochemistry?
A frequently used scoring matrix for amino acid sequences is a BLOSUM matrix such as the one shown below. The matrix is based on a set of multiply aligned, ungapped segments corresponding to the most highly conserved regions of proteins. These are called blocks.
## Blosum 50 Matrix

|    | A  | R  | N  | D  | C  | Q  | E  | G  | H  | I  | L  | K  | M  | F  | P  | S  | T  | W  | Y  | V  |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| A  | 5  | -2 | -1 | -2 | -1 | -1 | 0  | -2 | -1 | -2 | -1 | -1 | -3 | -1 | 1  | 0  | -3 | -2 | 0  |
| R  | -2 | 7  | -1 | -2 | -4 | 1  | 0  | -3 | 0  | -4 | -3 | 3  | -2 | -3 | -3 | -1 | -1 | -3 | -1 | -3 |
| N  | -1 | -1 | 7  | 2  | -2 | 0  | 0  | 0  | 1  | -3 | -4 | 0  | -2 | -4 | -2 | 1  | 0  | -4 | -2 | -3 |
| D  | -2 | -2 | 2  | 8  | -4 | 0  | 2  | -1 | -1 | -4 | -4 | -1 | -4 | -5 | -1 | 0  | -1 | -5 | -3 | -4 |
| C  | -1 | -4 | -2 | -4 | 13 | -3 | -3 | -3 | -3 | -2 | -2 | -3 | -2 | -2 | -4 | -1 | -1 | -5 | -3 | -1 |
| Q  | -1 | 1  | 0  | 0  | -3 | 7  | 2  | -2 | 1  | -3 | -2 | 2  | 0  | -4 | -1 | 0  | 1  | -1 | -1 | -3 |
| G  | 0  | -3 | 0  | -1 | -3 | -2 | -3 | 8  | -2 | -4 | -4 | -2 | -3 | -4 | -2 | 0  | -2 | -3 | -3 | -4 |
| H  | -2 | 0  | 1  | -1 | -5 | 1  | 0  | 2  | 10 | -4 | -3 | 0  | -1 | -1 | -2 | -1 | -2 | -3 | 2  | 4  |
| I  | -1 | -4 | -3 | -4 | -3 | -4 | -4 | -4 | 5  | 2  | -3 | 2  | 0  | -3 | -3 | -1 | -3 | -1 | 4  |
| L  | -2 | -3 | -4 | -4 | -2 | -2 | -3 | -4 | -3 | 2  | 5  | 3  | 3  | 1  | -4 | -3 | -1 | -2 | -1 | 1  |
| K  | -1 | 3  | 0  | -1 | -3 | 2  | 1  | -2 | 0  | -3 | -3 | 6  | 2  | -4 | -1 | 0  | -1 | -3 | -2 | 3  |
| M  | -1 | -2 | -2 | -4 | -2 | 0  | -2 | -3 | -1 | 2  | 3  | -2 | 7  | 0  | -3 | -2 | -1 | -1 | 0  | 1  |
| F  | -3 | -3 | -4 | -5 | -2 | -4 | -3 | -4 | -1 | 0  | 1  | -4 | 0  | 8  | -4 | -3 | -2 | 1  | 4  | -1 |
| P  | -1 | -3 | -2 | -1 | -4 | -1 | -1 | -2 | -2 | 3  | -4 | -1 | -3 | -4 | 10 | -1 | -1 | -4 | -3 | -3 |
| S  | 1  | -1 | 1  | 0  | -1 | 0  | -1 | 0  | 1  | -3 | -3 | 0  | -2 | -3 | -1 | 5  | 2  | -4 | -2 | -2 |
| T  | 0  | -1 | 0  | -1 | 1  | -1 | -1 | -2 | -2 | -1 | 1  | -1 | -2 | -1 | 2  | 5  | 3  | -3 | -2 | 0  |
| W  | -3 | -3 | -4 | -5 | -5 | -1 | -3 | -3 | -3 | -3 | -2 | -3 | -1 | 1  | -4 | -4 | -3 | 15 | 2  | -3 |
| Y  | -2 | -1 | -2 | -3 | -3 | -1 | -2 | -3 | 2  | 1  | -1 | -2 | 0  | 4  | -3 | -2 | -2 | 2  | 8  | -1 |
| V  | 0  | -3 | -3 | -4 | -1 | -3 | -3 | -4 | -4 | 4  | 1  | -3 | 1  | -1 | -3 | -2 | 0  | -3 | -1 | 5  |
Amino acids D, E and K are all charged; V, I, and L are all hydrophobic. What is the average BLOSUM50 score within the charged group of three? Within the hydrophobic group? Between the two groups?
Three Alignments, Good, Less Good, and Ugly

To illustrate problems in evaluating the significance of an alignment, the figure below shows examples of three pairwise alignments, all to the same region of the human alpha globin protein sequence (SWISS-PROT database identifier HBA_HUMAN). The central line in each alignment indicates identical positions with letters, and similar positions with a plus sign. (Similar pairs of residues are those which have a positive score in the substitution matrix used to score the alignment).
Figure 2.1 Three sequence alignments to a fragment of human alpha globin. (a) Clear similarity to human beta globin. (b) A structurally plausible alignment to leghaemoglobin from yellow lupin. (c) A spurious high-scoring alignment to a nematode glutathione S-transferase homologue named F11G11.2.
First Alignment (a), Good

In the first alignment, human alpha globin to human beta globin, there are many positions at which the two corresponding residues are identical; many others are functionally conservative, such as the pair D—E towards the end, representing an alignment of an aspartic acid residue with a glutamic acid residue, both negatively charged amino acids.
Second Alignment (b), Less Good

(b) also shows a biologically meaningful alignment of human alpha globin to leghaemoglobin from yellow lupin. These two sequences are evolutionarily related, have the same three-dimensional structure, and function in oxygen binding. However, there are many fewer identities than for (a), and in a couple of places gaps have been inserted into the alpha globin sequence to maintain the alignment across regions where the leghaemoglobin has extra residues.
Third Alignment (c), Ugly

(c) shows an alignment with a similar number of identities or conservative changes as (b). However, in this case we are looking at a spurious alignment to a protein, a nematode glutathiamine S-transferase homologue, that has a completely different structure and function.
Role of Statistics

- We would like some probabilistic measure, an alignment score, indicating how closely two sequences are related. We want the score to tell us if there is a significant relationship between the sequences or if what looks good is just a random occurrence.

- A warm up, the birthday paradox. The probability of $n$ random people having different birthdays is \( \frac{365!}{((365-n)!365^n)} \). This is \( \frac{1}{2} \) for $n = 23$. 
Log Odds

- The score for the maximal scoring alignment between two amino acid sequences using the BLOSUM substitution matrix is based on a log odds system. The theory does not take account of gaps. Some possibilities for scoring gaps are given by the linear gap penalty and the gap extension penalty.
As the animals left the ark, Noah told them to go forth and multiply. After some while, Noah happened upon two snakes sunning themselves. "Why aren't you multiplying?" Noah asked. The snakes replied, "We can't, we're adders."

So Noah and his sons went into the nearby forest and felled some trees. They made a platform of logs onto which they placed the snakes. You see, even adders can multiply on a log table.
Scoring Using *Maple*

- Maple has procedures for dealing with strings, and these can be used to write a procedure for scoring alignments using BLOSUM70
Heuristic algorithms aim for speed rather than absolute accuracy. They may use dynamic programming at various stages. Two popular ones are

- **BLAST** (Basic Local Alignment Search Tool)
- **FASTA**

**Glossary of terms**