



Pairwise Sequence Alignment


PH296 Presentation

Hui Tang 11/18/02




Biological Sequence Analysis

R. Durbin et al.

- 
- Many figures in this presentation are taken from this book.
 - Authors:
 - R. Durbin, S. Eddy, A. Krogh, G. Mitchison
 - It's a good book introducing widely used algorithms in computational biology.




Outline

- 
- Motivation
 - Definitions
 - Scoring Model
 - Algorithms
 - Significance of Scores



Motivation



Sequence comparison and alignment is a central problem in computational biology. The most basic task is: given two known sequences (DNA, RNA or amino acids) and a scoring model, determine if they are related or not.

- What sorts of alignment should be considered
- The scoring model
- The algorithm used to find optimal (or good) scoring alignments
- The statistical method used to evaluate the significance of an alignment score



Definitions

- Sequences diverged from common ancestor through mutations:
 - Substitution (AAGC \longrightarrow AAGT)
 - Insertion (AAG \longrightarrow AAGT)
 - Deletion (AAGC \longrightarrow AAG) } gaps
- Substring and subsequence
 - abc is a subsequence of axbycz, but NOT a substring





(a)

```
HBA_HUMAN  GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKL
              G+ +VK+HGKKV  A++++AH+D++  +++++LS+LH  KL
HBB_HUMAN  GNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL
```

(b)

```
HBA_HUMAN  GSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL
              ++  +++++H+ KV    + +A  ++                +L+ L+++H+ K
LGB2_LUPLU NNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG
```

(c)

```
HBA_HUMAN  GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSSD----LHAHKL
              GS+ + G +    +D L  ++ H+ D+  A +AL D    ++AH+
F11G11.2   GSGYLVGDSLTFVDLL--VAQHTADLLAANAALLDEFFPQFKAHQE
```

Figure 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 anematode glutathione S-transferase homologue named F11G11.2.

Scoring Model

	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	-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
E	-1	0	0	2	-3	2	6	-3	0	-4	-3	1	-2	-3	-1	-1	-1	-3	-2	-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	-3	1	0	-2	10	-4	-3	0	-1	-1	-2	-1	-2	-3	2	-4
I	-1	-4	-3	-4	-2	-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	-1	-2	-1	2	5	-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

Figure 2.2 The BLOSUM50 substitution matrix. The log-odds values have been scaled and rounded to the nearest integer for purposes of computational efficiency. Entries on the main diagonal for identical residue pairs are highlighted in bold.

Scoring Model-cont.

- Substitution Matrices

Random model R assumes residues occurs independently with some probabilities.

$$Pr(x,y/R) = \prod_i q_{xi} \prod_j q_{yi}$$

Match model M assumes aligned pairs of residues occur with a joint probability.

$$Pr(x,y/M) = \prod_i p_{xiyi}$$

odds ratio = match model likelihood / random model likelihood

$$= \prod_i p_{xiyi} / q_{xi} q_{yi}$$

log-odds ratio

$$S = \sum_i s(x_i, y_i) \quad \leftarrow \text{Example Fig2.1a}$$

$$\text{where } s(a,b) = \log(p_{a,b}/q_a q_b)$$



Scoring Model-cont.

- Gap Penalties

linear model:

$$\chi(g) = -gd$$

affine model:

$$\chi(g) = -d - (g - 1)e$$

where g is the length of gap,

d is called gap-open penalty,

e is called gap-extension penalty.



Linear vs. affine

- Examples:


GCTACTAG-T-T--CGC-T-TAGC
GCTACTAGCTCTAGCGCGTATAGC

GCTACTAG**T**T-----CG**C**TTAGC
GCTACTAG**C**TCTAGCGCG**T**ATAGC

e is usually less than d, allowing long gaps being penalized less than they would be in linear model.



Alignment Algorithms

- 
- Global alignment (Needleman–Wunsch algorithm)
 - Local alignment (Smith-Waterman algorithm)
 - Extensions
 - Different gap models
 - Heuristic alignment algorithms-BLAST




Dynamic Programming

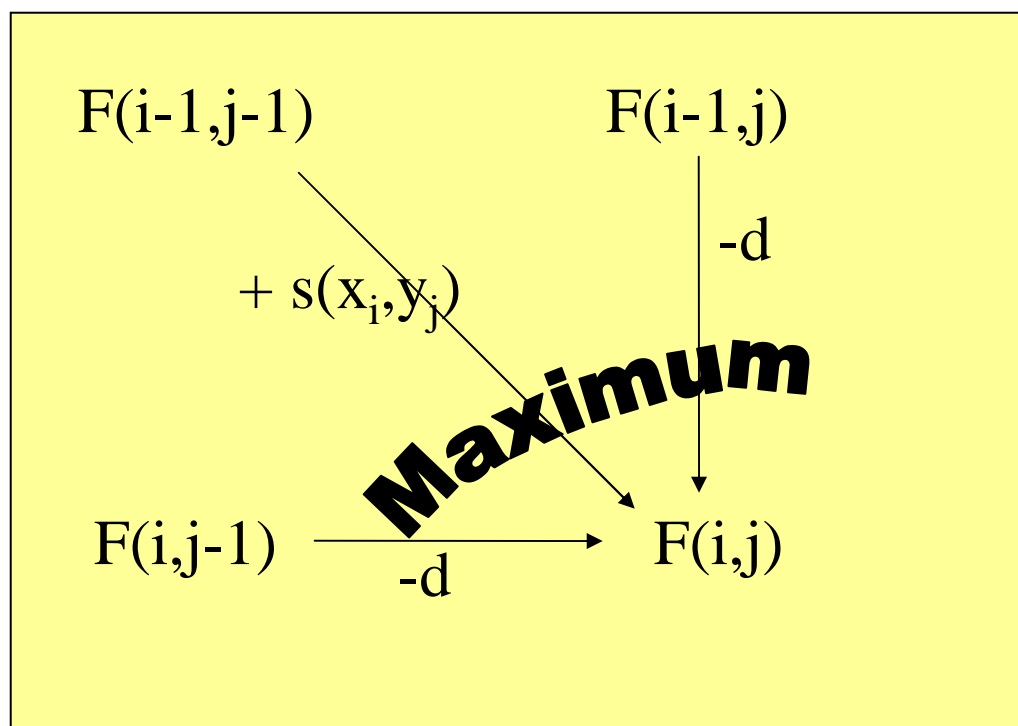
- a programming technique which can store the result of each subsubproblem. Therefore save the time to recalculate it when you met with it next time.
- Steps
 - recursive relation
 - tabular computation
 - trace-back





Global Alignment

- 
- $F(i,j)$: score of the best alignment between the initial segment $x_{1\dots i}$ of x up to x_i and the initial segment $y_{1\dots j}$ of y up to y_j .
 - Boundary conditions:
 $F(0,0)=0$; $F(0,j) = -jd$; $F(i,0) = -id$
 - $F(i,j) = \text{Max} \{ F(i-1,j-1) + s(x_i, y_j),$
 $F(i-1, j) - d,$
 $F(i, j-1) - d \}$
 - Example: align two short amino acid sequences
HEAGAWGHEE and PAWHEAE.





	H	E	A	G	A	W	G	H	E	E	
P	0 ←	-8 ←	-16 ←	-24 ←	-32 ←	-40 ←	-48 ←	-56 ←	-64 ←	-72 ←	-80
A	-8 ↑	-2 ↑	-9 ↑	-17 ←	-25 ←	-33 ←	-42 ←	-49 ←	-57	-65	-73
W	-16 ↑	-10 ↑	-3 ↑	-4 ←	-12	-20 ←	-28 ←	-36 ←	-44 ←	-52 ←	-60
H	-24 ↑	-18 ↑	-11 ↑	-6	-7	-15	-5 ←	-13 ←	-21 ←	-29 ←	-37
E	-32 ↑	-14 ↑	-18	-13	-8	-9	-13	-7	-3 ←	-11 ←	-19
A	-40 ↑	-22 ↑	-8 ←	-16	-16	-9	-12	-15	-7	3	-5
E	-48 ↑	-30 ↑	-16	-3 ←	-11	-11	-12	-12	-15	-5	2
E	-56 ↑	-38 ↑	-24	-11	-6	-12	-14	-15	-12	-9	1

HEAGAWGHE-E

--P-AW-HEAE

Figure 2.5 Above, the global dynamic programming matrix for our example sequences, with arrows indicating traceback pointers; values on the optimal alignment path are shown in bold. Below, a corresponding optimal alignment, which has total score 1.




Time Complexity

- Initialize matrix values: $O(n)$, $O(m)$
- Filling in rest of matrix: $O(nm)$
- Trace-back: $O(n+m)$

$$O(n^2)$$



Local Alignment

- 
- *Usually biological sequences under consideration are very long and will surely not be similar to each other globally. To find the best alignment for small subsequences is of interest. These are referred to as local alignments.*
 - *$F(i,j)$: score of the best alignment of a subsequences x and y .*
 - *Boundary conditions:*

$$F(0,0) = F(0,j) = F(i,0) = 0$$

- *Recursive relation:*

$$F(i,j) = \text{Max} \{ 0, \quad \leftarrow$$
$$F(i-1,j-1) + s(x_i, y_j),$$
$$F(i-1, j) - d,$$
$$F(i, j-1) - d \}$$

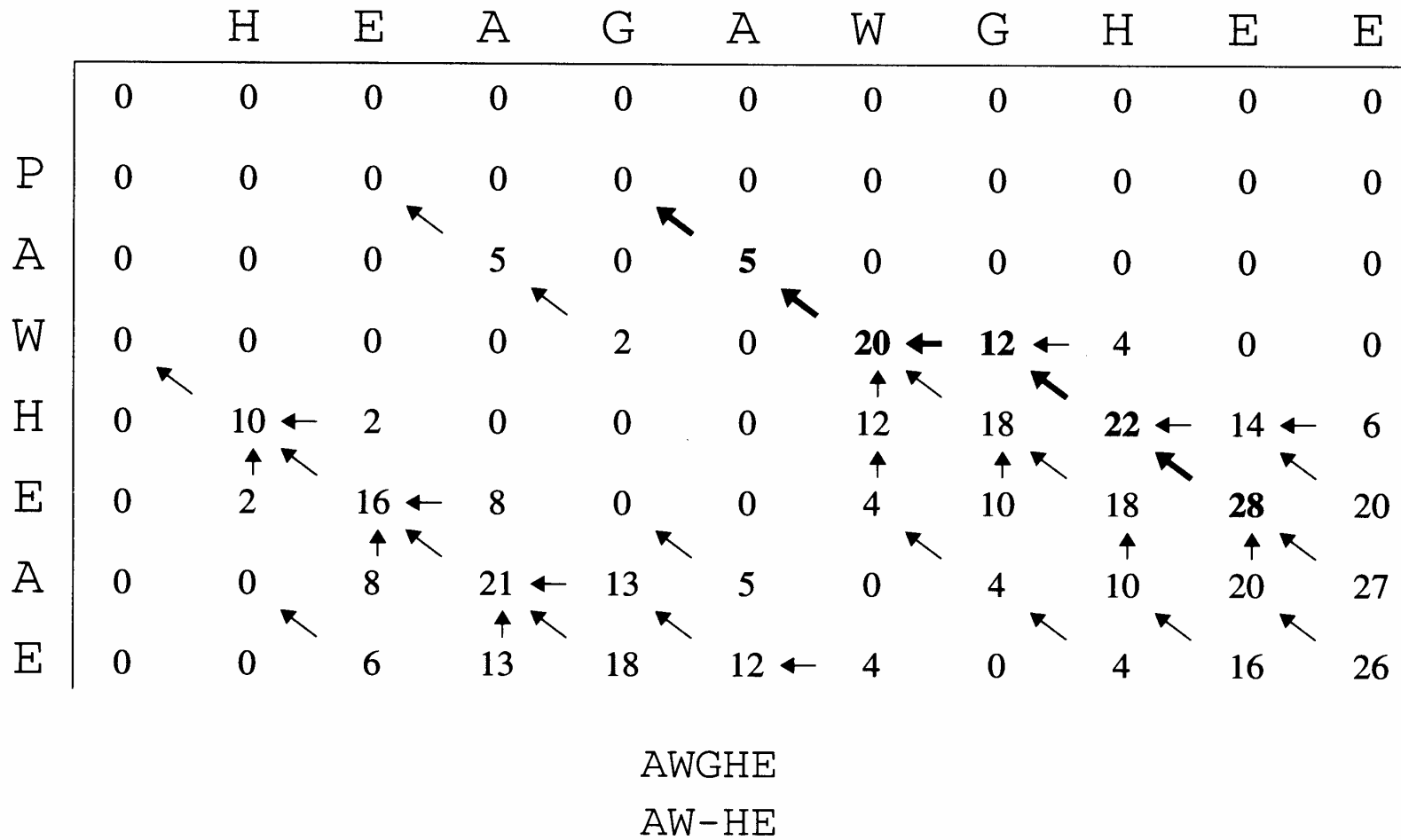



Figure 2.6 Above, the local dynamic programming matrix for the example sequences. Below, the optimal local alignment, with score 28.



Global vs. local

- 
- Same basic method
 - Difference:
 - boundary conditions;
 - trace-back;
 - Interesting URLs:

<http://www.cse.ucsc.edu/research/kestrel/runkestrel.html>

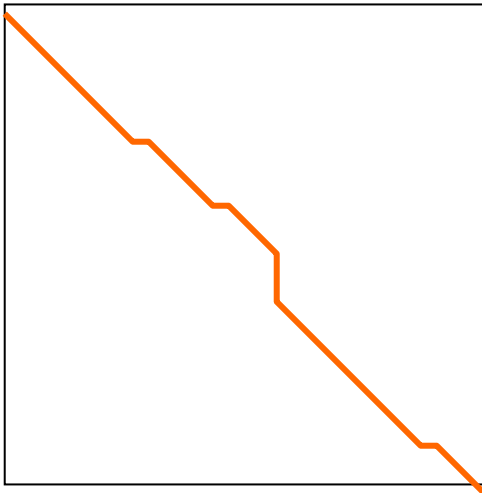
Smith-Waterman by using different substitution matrices.

Global vs. local-cont.



Optimal global
alignment

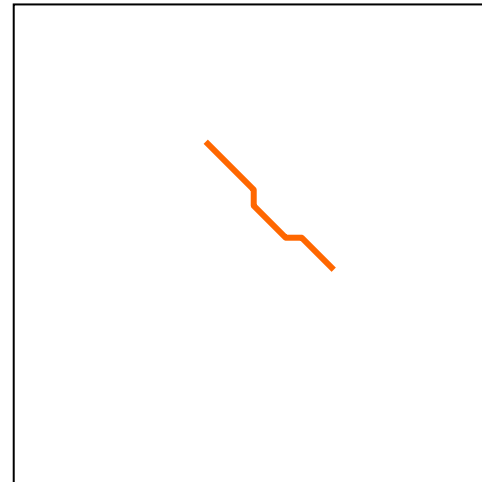
Needleman & Wunsch (1970)



Sequences align
through the whole
region

Optimal local
alignment

Smith & Waterman (1981)



Sequences align
only in small,
isolated regions




Extensions

- Affine gap model
- Overlap matches
 - one sequence contains the other, or that they overlap.





Heuristic Alignment Algorithms

- 
- **Goal:**
Search as small as possible of the cells in the dynamic programming matrix, while still looking at the high scoring alignment.
 - **Benefits:**
 - save time, $\ll O(n^2)$
 - **Drawback:**
 - might miss the best scoring alignment



Blast

Basic Local Alignment Search Tool

- Function:

Finding high scoring local alignment between a query sequence and a target database.

- Interesting URLs:

<http://www.ncbi.nlm.nih.gov/BLAST/>


<http://www.dina.dk/~sestoft/bsa/bsapplet.html>

--type in two amino acid sequences in the top-most two windows, then press the button, it'll give different optimal alignments by using different models, including the methods we mentioned above: global alignment and local alignment.





Significance of Scores



$Pr(M/x,y)$ → the probability that the sequences are related as opposed to being unrelated.

$Pr(x,y/M)$ → the one we calculated above

By using Bayes' rule, we can calculate one from another.

Assumptions:

1. Specify the prior probabilities of the two models $Pr(R)$ and $Pr(M)$;
2. $Pr(R) = 1 - Pr(M)$;

$$\begin{aligned} Pr(M/x,y) &= Pr(x,y/M)Pr(M)/Pr(x,y) \\ &= Pr(x,y/M)Pr(M)/(Pr(x,y/M)Pr(M) + Pr(x,y/R)Pr(R)) \\ &= (Pr(x,y/M)Pr(M)/ Pr(x,y/R)Pr(R)) \\ &\quad / (1 + Pr(x,y/M)Pr(M)/ Pr(x,y/R)Pr(R)) \end{aligned}$$



Significance of Scores-cont.

Set

$$S' = S + \log(\Pr(M)/\Pr(R))$$

where $S = \log(\Pr(x,y/M)/\Pr(x,y/R))$

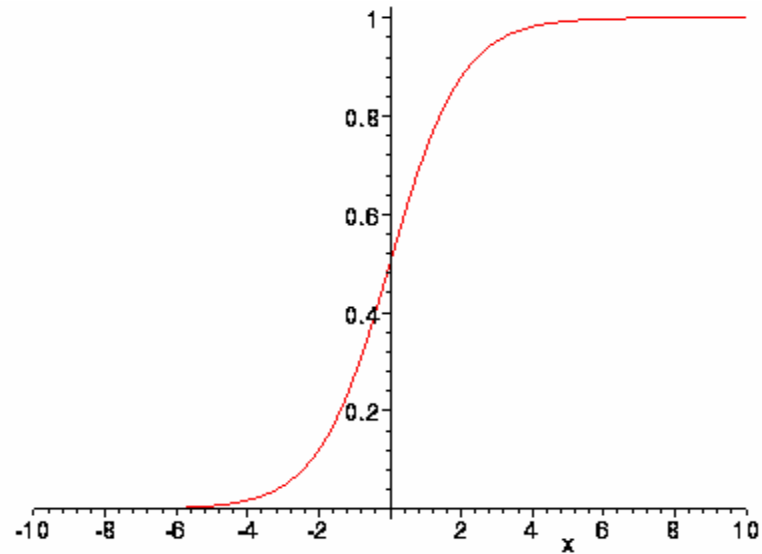
Then

$$\Pr(M/x,y) = \sigma(S')$$

Where $\sigma(x) = e^x/(1 + e^x)$, known as the logistic function.



Significance of Scores-cont.



- Compare the score to 0,1 to see if they related or not.



Further...

- Pairwise alignment with HMMs
 - One advantage is we could explore the reliability of the alignment we obtained by using DP.
- Multiple sequence alignment

