

Introduction to BLAST

- Basic Local Alignment Search Tool
- Used for searching large databases for sequences having good local alignments with some query sequence
- Locates very small, good alignments and then attempts to extend them.

From the website:

The initial search is done for a word of length "W" that scores at least "T" when compared to the query using a given substitution matrix. Word hits are then extended in either direction in an attempt to generate an alignment with a score exceeding the threshold of "S". The "T" parameter dictates the speed and sensitivity of the search.

Let's BLAST some Proteins

“I” before “E,”
Except after “C,”
Or when sounding like “eigh”
As in “neighbor” and “weigh,”
Or in “protein.”

Recall that proteins are strings of amino acids.
There are 20 amino acids, denoted by the letters:
ACDEFGHIKLMNPQRSTVWY

That is, all letters except BJOUXZ

But first... What is this “substitution matrix?”

Substitution Matrix

Here is a portion of the BLOSUM62 substitution matrix. It gives the scores of the columns in a protein alignment:

	A	C	D	E	F	G	H	I	...
A	4	0	-2	-1	-2	0	-2	-1	
C	0	9	-3	-4	-2	-3	-3	-1	
D	-2	-3	6	2	-3	-1	-1	-3	
E	-1	-4	2	5	-3	-2	0	-3	
F	-2	-2	-3	-3	6	-3	-1	0	
G	0	-3	-1	-2	-3	6	-2	-4	
H	-2	-3	-1	0	-1	-2	8	-3	
I	-1	-1	-3	-3	0	-4	-3	4	
...									...

For example, the score of the alignment shown to the right would be:

HIDE FACE

Smith-Waterman with a Scoring Matrix

We can find an optimal local alignment of an amino acid sequence using our scoring matrix.

	K	A	T	H	Y
C					
H					
C					
K					

What BLAST Does

Suppose HEDGE is our query sequence

- Make a list of all triples in the query sequence

HED
EDG
DGE

- Compile a list of triples which have a high score with each of those triples.

HED	HED, HDD, DED, DDD, HHD, . . .
EDG	EDG, DDG, DEG, EEG, EHG, . . .
DGE	DGE, EGE, EGD, EGD, HGE, . . .

- The value of “high” is set once the query sequence is known, so that no more than 1 in 50 of the database entries contains a high-scoring triple.

What BLAST Does

- All occurrences of a high-scoring sequence are found in the database.
- These “seeds” are then extended in each direction, usually without gaps, to find long, high scoring regions
- These regions are then returned, and are called “hits”
- BLAST returns all the hits it can find, up to some maximum number. Then it says “Found too many...”

The PAM Matrices

- PAM stands for “point accepted mutation” or “percent accepted mutation”
- This refers to the situation in which 1% of the amino acids in a protein sequence have mutated
- Let M-1 be the matrix showing the transition probabilities for each amino acid during such a period of time. For simplicity, let us imagine that there were just 4 amino acids:

M-1	A	C	D	E
A	0.99	0.002	0.003	0.005
C	0.002	0.99	0.005	0.003
D	0.003	0.005	0.99	0.002
E	0.005	0.003	0.002	0.99

- Obtained from closely related sequences in nature and computing the empirical frequency of amino acid substitution.

The PAM-1 Matrix

- PAM-1 is a *scoring* matrix useful for measuring how good an alignment is
- Each entry in PAM-1 is derived from the corresponding entry in M-1 by
 - ▶ Multiplying by 4
 - ▶ Taking the logarithm base 10
 - ▶ Multiplying by 10
 - ▶ Rounding to the nearest integer

PAM-1	A	C	D	E
A	6	-21	-19	-17
C	-21	6	-17	-19
D	-19	-17	6	-21
E	-17	-19	-21	6

What kind of alignments would be preferred under the PAM-1 scoring matrix?

The PAM-2 Matrix

The PAM-2 matrix is a scoring matrix derived from the M-2 matrix using the same four steps.

The M-2 matrix gives the transition probabilities after two units of evolution represented by M-1

Let's compute M-2

Computing M-2

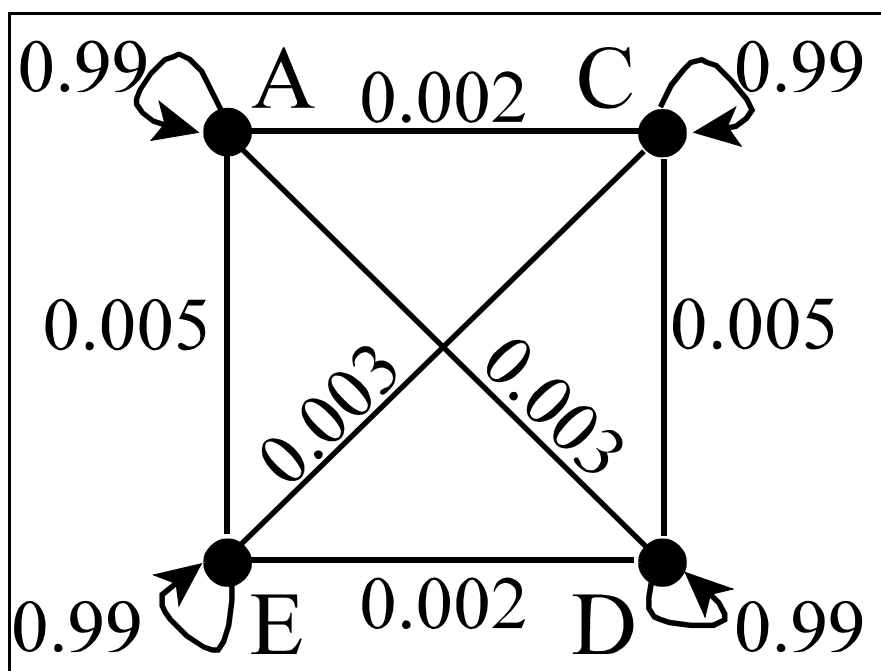
M-1	A	C	D	E
A	0.99	0.002	0.003	0.005
C	0.002	0.99	0.005	0.003
D	0.003	0.005	0.99	0.002
E	0.005	0.003	0.002	0.99

What is the probability that after two M-1 units of time, amino acid A is still amino acid A?

Notice: A did not have to *stay* A in between...

M-2	A	C	D	E
A				
C				
D				
E				

Computing M-2



$$\begin{aligned}
 P(A \rightarrow A) &= P(A \rightarrow A \rightarrow A) + P(A \rightarrow C \rightarrow A) + \\
 &\quad P(A \rightarrow D \rightarrow A) + P(A \rightarrow E \rightarrow A) \\
 &= (0.990)(0.990) + (0.002)(0.002) + \\
 &\quad (0.003)(0.003) + (0.005)(0.005) \\
 &= 0.9801
 \end{aligned}$$

$$\begin{aligned}
 P(A \rightarrow B) &= P(A \rightarrow A \rightarrow C) + P(A \rightarrow C \rightarrow C) + \\
 &\quad P(A \rightarrow D \rightarrow C) + P(A \rightarrow E \rightarrow C) \\
 &= 0.0040
 \end{aligned}$$

The Easy Way to Compute M-2

M-2 is obtained from M-1 by standard matrix multiplication.

M-1	A	C	D	E	M-1	A	C	D	E
A	0.99	0.002	0.003	0.005	A	0.99	0.002	0.003	0.005
C	0.002	0.99	0.005	0.003	C	0.002	0.005	0.003	0.003
D	0.003	0.005	0.99	0.002	D	0.003	0.99	0.002	0.002
E	0.005	0.003	0.002	0.99	E	0.005	0.002	0.99	0.99

M-2	A	C	D	E
A	0.9801	0.0040		
C				
D				
E				

M-2

M-2	A	C	D	E
A	0.9801	0.0040	0.0060	0.0099
C	0.0040	0.9801	0.0099	0.0060
D	0.0060	0.0099	0.9801	0.0040
E	0.0099	0.0060	0.0040	0.9801

PAM-2

PAM-2	A	C	D	E
A	6	-18	-16	-14
C	-18	6	-14	-16
D	-16	-14	6	-18
E	-14	-16	-18	6

Comparison Between PAM-1 and PAM-2

PAM-1	A	C	D	E
A	6	-21	-19	-17
C	-21	6	-17	-19
D	-19	-17	6	-21
E	-17	-19	-21	6

PAM-2	A	C	D	E
A	6	-18	-16	-14
C	-18	6	-14	-16
D	-16	-14	6	-18
E	-14	-16	-18	6

How do you think we would obtain PAM-250?

Raise M-1 to the 250th power and do that log stuff.

Comparison Between PAM-2 and PAM-250

PAM-2	A	C	D	E
A	6	-18	-16	-14
C	-18	6	-14	-16
D	-16	-14	6	-18
E	-14	-16	-18	6

PAM-250	A	C	D	E
A	5	-6	-5	-3
C	-6	5	-3	-5
D	-5	-3	5	-6
E	-3	-5	-6	5

Notice how PAM-250 penalizes much less heavily for mismatches than PAM-2. Thus it is much better suited for aligning distantly related sequences.

Handout #1 — Smith-Waterman with an Amino Acid Substitution Matrix

Here is the BLOSUM62 scoring matrix:

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	4	0	-2	-1	-2	0	-2	-1	-1	-1	-1	-2	-1	-1	-1	1	0	0	-3	-2
C	0	9	-3	-4	-2	-3	-3	-1	-3	-1	-1	-3	-3	-3	-3	-1	-1	-1	-2	-2
D	-2	-3	6	2	-3	-1	-1	-3	-1	-4	-3	1	-1	0	-2	0	-1	-3	-4	-3
E	-1	-4	2	5	-3	-2	0	-3	1	-3	-2	0	-1	2	0	0	-1	-2	-3	-2
F	-2	-2	-3	-3	6	-3	-1	0	-3	0	0	-3	-4	-3	-3	-2	-2	-1	1	3
G	0	-3	-1	-2	-3	6	-2	-4	-2	-4	-3	0	-2	-2	-2	0	-2	-3	-2	-3
H	-2	-3	-1	0	-1	-2	8	-3	-1	-3	-2	1	-2	0	0	-1	-2	-3	-2	2
I	-1	-1	-3	-3	0	-4	-3	4	-3	2	1	-3	-3	-3	-3	-2	-1	3	-3	-1
K	-1	-3	-1	1	-3	-2	-1	-3	5	-2	-1	0	-1	1	2	0	-1	-2	-3	-2
L	-1	-1	-4	-3	0	-4	-3	2	-2	4	2	-3	-3	-2	-2	-2	-1	1	-2	-1
M	-1	-1	-3	-2	0	-3	-2	1	-1	2	5	-2	-2	0	-1	-1	-1	1	-1	-1
N	-2	-3	1	0	-3	0	1	-3	0	-3	-2	6	-2	0	0	1	0	-3	-4	-2
P	-1	-3	-1	-1	-4	-2	-2	-3	-1	-3	-2	-2	7	-1	-2	-1	-1	-2	-4	-3
Q	-1	-3	0	2	-3	-2	0	-3	1	-2	0	0	-1	5	1	0	-1	-2	-2	-1
R	-1	-3	-2	0	-3	-2	0	-3	2	-2	-1	0	-2	1	5	-1	-1	-3	-3	-2
S	1	-1	0	0	-2	0	-1	-2	0	-2	-1	1	-1	0	-1	4	1	-2	-3	-2
T	0	-1	-1	-1	-2	-2	-2	-1	-1	-1	-1	0	-1	-1	-1	1	5	0	-2	-2
V	0	-1	-3	-2	-1	-3	-3	3	-2	1	1	-3	-2	-2	-3	-2	0	4	-3	-1
W	-3	-2	-4	-3	1	-2	-2	-3	-3	-2	-1	-4	-4	-2	-3	-3	-2	-3	11	2
Y	-2	-2	-3	-2	3	-3	2	-1	-2	-1	-1	-2	-3	-1	-2	-2	-2	-1	2	7

Find an optimal local alignment using the Smith-Waterman algorithm with the scores given above.

		K	A	T	H	Y
C						
H						
C						
K						

Handout #2 --- Computing M-2 and PAM-2 from M-1

M-1	A	C	D	E
A	0.99	0.002	0.003	0.005
C	0.002	0.99	0.005	0.003
D	0.003	0.005	0.99	0.002
E	0.005	0.003	0.002	0.99

M-2 is the matrix of transition probabilities between nucleotides during the time period represented by two M-1 transitions. It can thus be computed from M-1, once we've accepted M-1 as a model of sequence evolution for that time period. Take a moment now to compute M-2.

M-2	A	C	D	E
A				
C				
D				
E				

Finally, compute PAM-2 by following the PAM-generation steps:

- ▶ Multiplying by 4
- ▶ Taking the logarithm base 10
- ▶ Multiplying by 10
- ▶ Rounding to the nearest integer

PAM-2	A	C	D	E
A				
C				
D				
E				

Exercises — BLAST

Quick Concept:

1. Just for the fun of it, show that if a 4×4 matrix is symmetric, has “0.99” on the diagonal and has row sums equal to 1, then it must be of the form:

0.99	a	b	x
a	0.99	x	b
b	x	0.99	a
x	b	a	0.99

where $a + b + x = 0.01$

Presentation Problems:

2. Here is the BLOSUM62 scoring matrix. What is the score of the following amino acid alignment if the start gap penalty is -11 and the gap continuation penalty is -1?

AMINE---ACID
GLIDERBRACES

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	4	0	-2	-1	-2	0	-2	-1	-1	-1	-1	-2	-1	-1	-1	1	0	0	-3	-2
C	0	9	-3	-4	-2	-3	-3	-1	-3	-1	-1	-3	-3	-3	-3	-1	-1	-1	-2	-2
D	-2	-3	6	2	-3	-1	-1	-3	-1	-4	-3	1	-1	0	-2	0	-1	-3	-4	-3
E	-1	-4	2	5	-3	-2	0	-3	1	-3	-2	0	-1	2	0	0	-1	-2	-3	-2
F	-2	-2	-3	-3	6	-3	-1	0	-3	0	0	-3	-4	-3	-3	-2	-2	-1	1	3
G	0	-3	-1	-2	-3	6	-2	-4	-2	-4	-3	0	-2	-2	-2	0	-2	-3	-2	-3
H	-2	-3	-1	0	-1	-2	8	-3	-1	-3	-2	1	-2	0	0	-1	-2	-3	-2	2
I	-1	-1	-3	-3	0	-4	-3	4	-3	2	1	-3	-3	-3	-3	-2	-1	3	-3	-1
K	-1	-3	-1	1	-3	-2	-1	-3	5	-2	-1	0	-1	1	2	0	-1	-2	-3	-2
L	-1	-1	-4	-3	0	-4	-3	2	-2	4	2	-3	-3	-2	-2	-2	-1	1	-2	-1
M	-1	-1	-3	-2	0	-3	-2	1	-1	2	5	-2	-2	0	-1	-1	-1	1	-1	-1
N	-2	-3	1	0	-3	0	1	-3	0	-3	-2	6	-2	0	0	1	0	-3	-4	-2
P	-1	-3	-1	-1	-4	-2	-2	-3	-1	-3	-2	-2	7	-1	-2	-1	-1	-2	-4	-3
Q	-1	-3	0	2	-3	-2	0	-3	1	-2	0	0	-1	5	1	0	-1	-2	-2	-1
R	-1	-3	-2	0	-3	-2	0	-3	2	-2	-1	0	-2	1	5	-1	-1	-3	-3	-2
S	1	-1	0	0	-2	0	-1	-2	0	-2	-1	1	-1	0	-1	4	1	-2	-3	-2
T	0	-1	-1	-1	-2	-2	-2	-1	-1	-1	-1	0	-1	-1	-1	1	5	0	-2	-2
V	0	-1	-3	-2	-1	-3	-3	3	-2	1	1	-3	-2	-2	-3	-2	0	4	-3	-1
W	-3	-2	-4	-3	1	-2	-2	-3	-3	-2	-1	-4	-4	-2	-3	-3	-2	-3	11	2
Y	-2	-2	-3	-2	3	-3	2	-1	-2	-1	-1	-2	-3	-1	-2	-2	-2	-1	2	7

3. Find an optimal local alignment using the above scoring matrix with gap penalty -4.

		W	R	A	P	S
		0	0	0	0	0
S		0				
W		0				
I		0				
P		0				
E		0				

4. Organisms on Pentworld have just 5 amino acids. This matrix shows the transition probabilities for one percent of accepted mutation. What is the PAM-1 matrix? PAM-250?

	A	C	D	E	F
A	0.900	0.001	0.002	0.050	0.047
C	0.015	0.920	0.020	0.030	0.015
D	0.002	0.017	0.880	0.050	0.051
E	0.010	0.020	0.045	0.920	0.005
F	0.020	0.070	0.001	0.050	0.859

5. Use the PAM matrices from the previous problem to make some conjecture about the evolutionary distance between the strings ACDAAACAE and ADACAACFE, aligned gaplessly.

6. A typical scoring scheme for the alignment of nucleotide sequences is “+1” for a match and “-1” for a mismatch. This could be written as a 4×4 scoring matrix similar to our PAM matrices, but with just two values for its entries. It could thus be thought of as the PAM-1 matrix associated with some hypothetical M-1 matrix, obtained by following the four steps we did in class with the multiplying by 4, taking the log, and so on. Assuming the M-1 matrix has the form:

	A	C	D	E
A	x	y	y	y
C	y	x	y	y
D	y	y	x	y
E	y	y	y	x

what base should you use for the logarithm step to obtain the ± 1 , PAM-1 scoring matrix, without having to do any rounding?

7. One organism on Pentaworld (see problem 4) has a protein consisting of about 10,000 amino acids which has been evolving for millions of years, amounting to about a thousand PAM-1 time frames. Approximately how many of each type of amino acid does this protein contain, assuming:
- It initially consisted of 9000 A's and 1000 F's?
 - It initially consisted of an equal number of each of its 5 amino acids?

8. The version of the Smith-Waterman algorithm we've seen may introduce gaps into an alignment in order to achieve the optimal score. How would you adapt the Smith-Waterman algorithm to find the optimal *gapless* local alignment between two strings. Test your method here: (Assign +1 for a match and -1 for a mismatch.)

		A	A	C	G	C	A	T	A
	☐	☐	☐	☐	☐	☐	☐	☐	☐
C	☐	☐	☐	☐	☐	☐	☐	☐	☐
G	☐	☐	☐	☐	☐	☐	☐	☐	☐
G	☐	☐	☐	☐	☐	☐	☐	☐	☐
A	☐	☐	☐	☐	☐	☐	☐	☐	☐
T	☐	☐	☐	☐	☐	☐	☐	☐	☐
A	☐	☐	☐	☐	☐	☐	☐	☐	☐
T	☐	☐	☐	☐	☐	☐	☐	☐	☐
C	☐	☐	☐	☐	☐	☐	☐	☐	☐