Scoring and heuristic methods for sequence alignment

Amino Acid Substitution Matrices

- Used to score alignments.
- Reflect evolution of sequences.

Unitary Matrix:

$$M_{ij} = \begin{cases} 1 & i=j \\ 0 & o/w \end{cases}$$

Genetic Code Matrix:

M_{ij} = min no. of base changes needed to alter codon of i to codon of j.

Scoring Matrices

- Wish evolutionary-based matrices
- More similar pairs of sequences should require different matrices than more divergent pairs.
- Several families of matrices were constructed, to be used according to the level of divergence:
  - Global approach (PAM).
  - Local approach (BLOSUM)
- Higher PAM and Lower BLOSUM for more different sequences

Log-odds

- All matrices compare the probability of the aligned sequences according to:
  - Random model: letters are independent
  - Alternative model: paired letters have some joint probability.

$$\frac{P(x,y \mid M)}{P(x,y \mid R)} = \prod_i \frac{P(x_i,y_i)}{P(x_i)P(y_i)}$$

- Taking a logarithm results in an additive scoring system.

PAM Matrices (Dayhoff et al., 78)

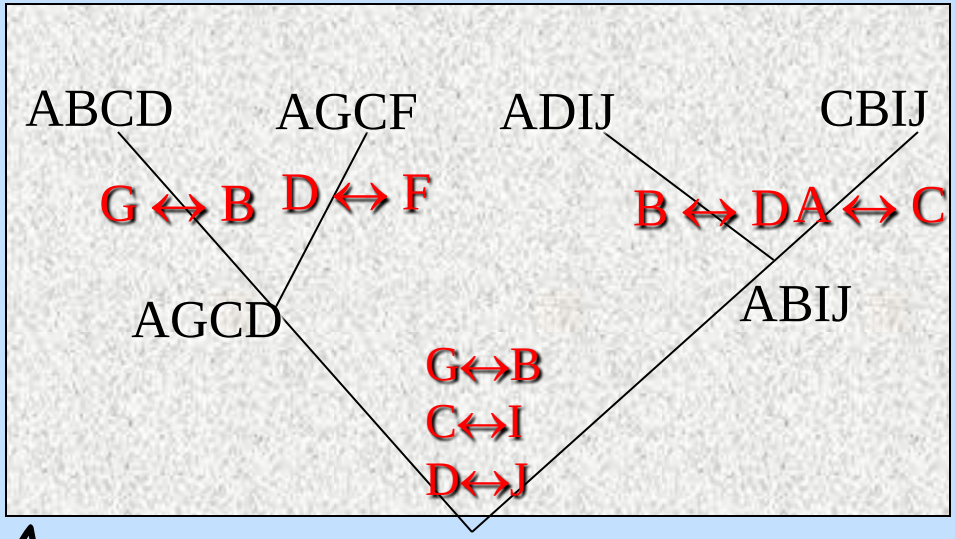
- PAM = Percent (or Point) Accepted Mutation
- Protein sequences S_1, S_2 are **at evolutionary distance of one PAM** if S_1 has converted to S_2 with an average of **one accepted point mutation per 100 AAs**.:
  - PAM1 should be used for sequences whose evolutionary distance causes 1% difference between them.
  - PAM2 should be used for sequences twice as distant…

Observed % difference	Evolutionary distance in PAMs
1	1
5	5
10	11
15	17
20	23
30	38
40	56
50	80
55	94
60	112
70	159
75	195
80	246

PAM Matrices (2)

Generating PAM:

- Start with aligned sequences, highly similar, with known evolutionary trees.



- Count exchanges $A_{ab} = A_{ba}$
- Compute matrix M_{ab} = "prob."(a changes to b in one unit) = $A_{ab} / \sum_c A_{ac}$
- Now M^k gives change probs. in k units.

$$\text{"log-odds"} = \log \frac{f(a)M^k(a,b)}{f(a)f(b)} = \log \frac{M^k(a,b)}{f(b)}$$

Dayhoff's Data

- 71 parsimony-based evolutionary trees of close sequence families.
- 1,572 substitutions overall
- Normalized matrix (multiplying all non-diagonal entries by a constant) so that:

$$\sum f(i)(1-M_{ii}) = 0.01$$

ORIGINAL AMINO ACID

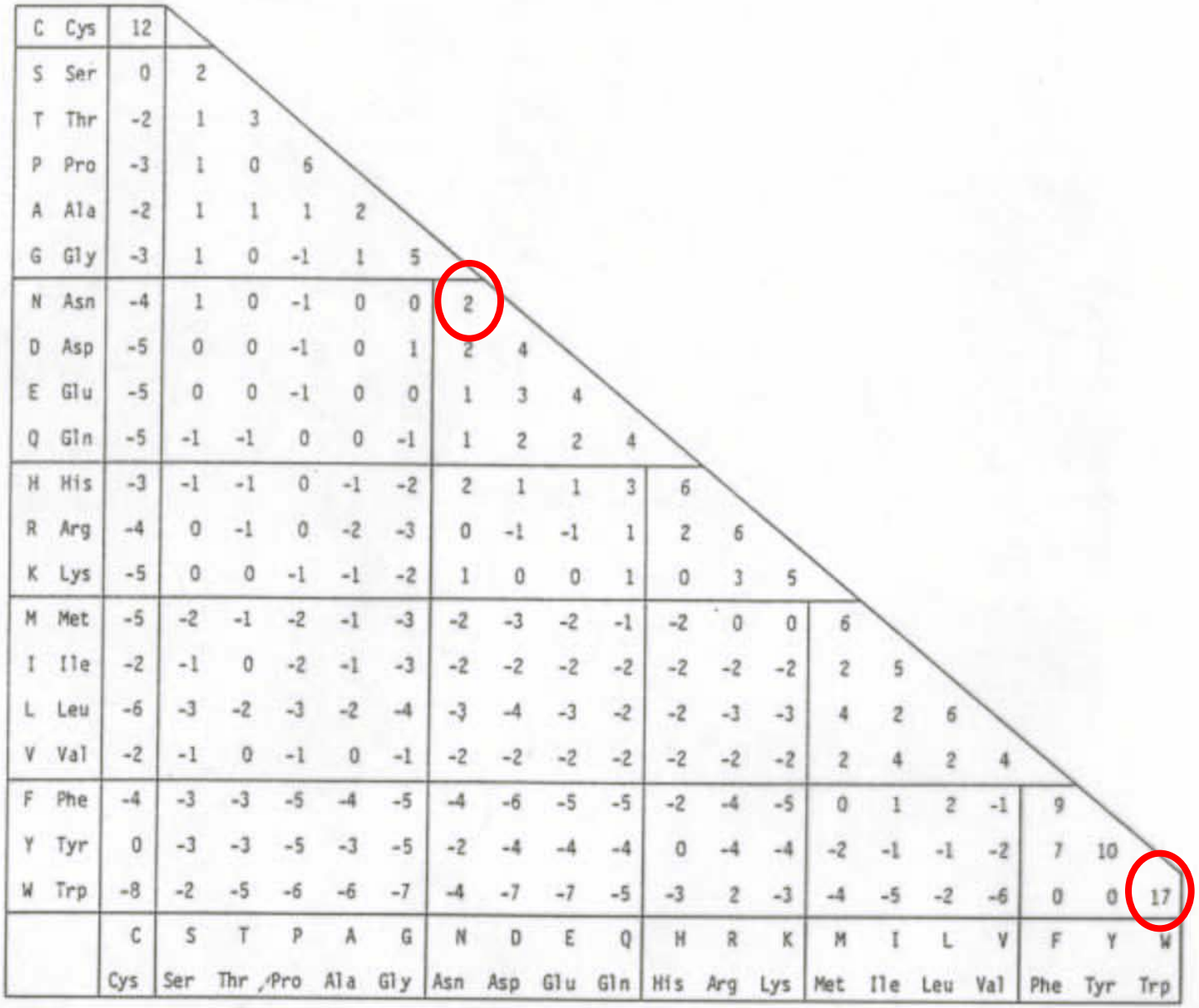
REPLACEMENT AMINO ACID	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A Ala	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R Arg	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N Asn	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D Asp	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C Cys	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q Gln	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E Glu	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G Gly	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H His	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I Ile	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L Leu	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K Lys	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M Met	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F Phe	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P Pro	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S Ser	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T Thr	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W Trp	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y Tyr	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V Val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

Figure 82. Mutation probability matrix for the evolutionary distance of 1 PAM. An element of this matrix, M_{ij}, gives the probability that the amino acid in column j will be replaced by the amino acid in row i after a given evolutionary interval, in this case 1 accepted point mutation per 100 amino acids. Thus, there is a 0.56% probability that Asp will be replaced by Glu. To simplify the appearance, the elements are shown multiplied by 10,000.

ORIGINAL AMINO ACID

REPLACEMENT AMINO ACID	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A Ala	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
R Arg	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
N Asn	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
D Asp	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
C Cys	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Q Gln	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
E Glu	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
G Gly	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
H His	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
I Ile	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
L Leu	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
K Lys	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
M Met	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
F Phe	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
P Pro	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
S Ser	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
T Thr	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
W Trp	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Y Tyr	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
V Val	7	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17

Figure 83. Mutation probability matrix for the evolutionary distance of 250 PAMs. To simplify the appearance, the elements are shown multiplied by 100. In comparing two sequences of average amino acid frequency at this evolutionary distance, there is a 13% probability that a position containing Ala in the first sequence will contain Ala in the second. There is a 3% chance that it will contain Arg, and so forth. The relationship of two sequences at a distance of 250 PAMs can be demonstrated by statistical methods.



C	Cys	12																					
S	Ser	0	2																				
T	Thr	-2	1	3																			
P	Pro	-3	1	0	6																		
A	Ala	-2	1	1	1	2																	
G	Gly	-3	1	0	-1	1	5																
N	Asn	-4	1	0	-1	0	0	2															
D	Asp	-5	0	0	-1	0	1	2	4														
E	Glu	-5	0	0	-1	0	0	1	3	4													
Q	Gln	-5	-1	-1	0	0	-1	1	2	2	4												
H	His	-3	-1	-1	0	-1	-2	2	1	1	3	6											
R	Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6										
K	Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5									
M	Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6								
I	Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5							
L	Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6						
V	Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4					
F	Phe	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9				
Y	Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10			
W	Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17		
		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W		
		Cys	Ser	Thr	Pro	Ala	Gly	Asn	Asp	Glu	Gln	His	Arg	Lys	Met	Ile	Leu	Val	Phe	Tyr	Trp		

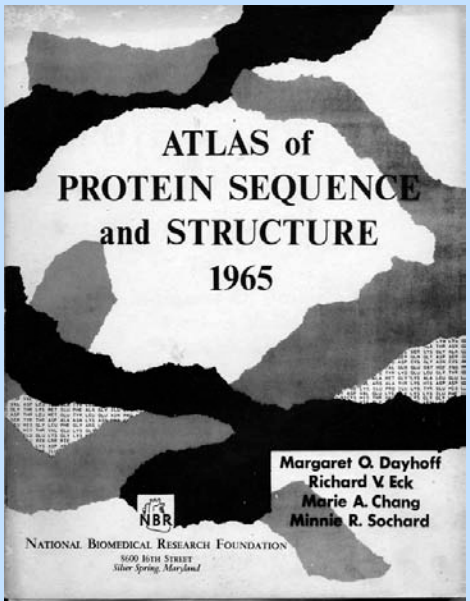
Figure 84. Log odds matrix for 250 PAMs. Elements are shown multiplied by 10. The neutral score is zero. A score of -10 means that the pair would be expected to occur only one-tenth as frequently in related sequences as random chance would predict, and a score of +2 means that the pair would be expected to occur 1.6 times as frequently. The order of the amino acids has been arranged to illustrate the patterns in the mutation data.

Caveats

- Markovian model: state at time n depends only on state at time n-1
- Assumes constant molecular clock
- Same model for all AA positions
- Ignores indels

Margaret Oakley Dayhoff (1925-1983)

A pioneer in the use of computers in chemistry and biology, beginning with her PhD thesis project in 1948. Her work was multi-disciplinary, and used her knowledge of chemistry, mathematics, biology and computer science to develop an entirely new field. She is credited today as one of the founders of the field of Bioinformatics. Dr. Dayhoff was the first woman in the field of Bioinformatics.



BLOSUM (Henikoff & Henikoff, 92)

- PAM: based on highly similar global alignments
- BLOSUM (BLOcks SUbstitution Matrix): based on short, gapless local alignments
  - Identify blocks: conserved segments in alignment of proteins from the same family.
  - Eliminate sequences that are >x% identical (by clustering & representing each cluster by a single sequence)
  - Collect stats A_{ab} on pairs (a,b) in each column
  - q_{ab} = prob of AA pairs (a,b) in same column
  - p_a = prob of observing a
  - e_{ab} = freq. of pair (a,b) assuming independence $=p_a^2$ if a=b, $2p_ap_b$ if $a \neq b$
  - Log odds: $s_{ab} = \log (q_{ab}/e_{ab})$
  - BLOSUM X matrix: s_{ab} discretized

Blosum62

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

FIG. 2. BLOSUM 62 substitution matrix (*Lower*) and difference matrix (*Upper*) obtained by subtracting the PAM 160 matrix position by position. These matrices have identical relative entropies (0.70); the expected value of BLOSUM 62 is −0.52; that for PAM 160 is −0.57.

Comparing matrices

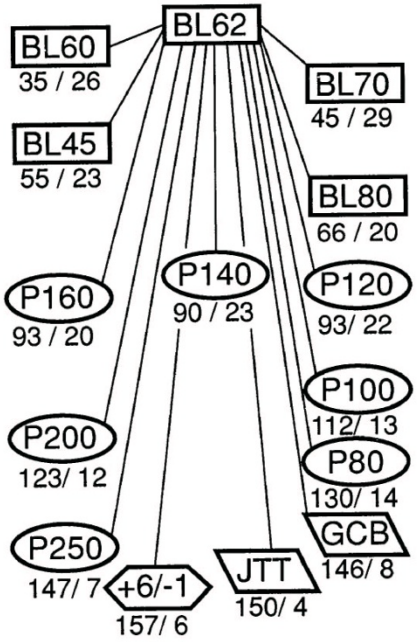


FIG. 4. Searching performance of BLAST using different matrices from the BLOSUM (BL) series, the PAM (P) series, and two recent updates of the standard Dayhoff matrix: GCB (25) and JTT (26). Results are based on searches using queries for each of 504 different groups. For each pair of numbers below a box representing a matrix, the first is the number of groups for which BLOSUM 62 missed fewer sequences than that matrix, and the second is the number of groups for which BLOSUM 62 missed more. The vertical distance between each matrix and BLOSUM 62 is proportional to the difference.

PAM vs BLOSUM in different algorithms

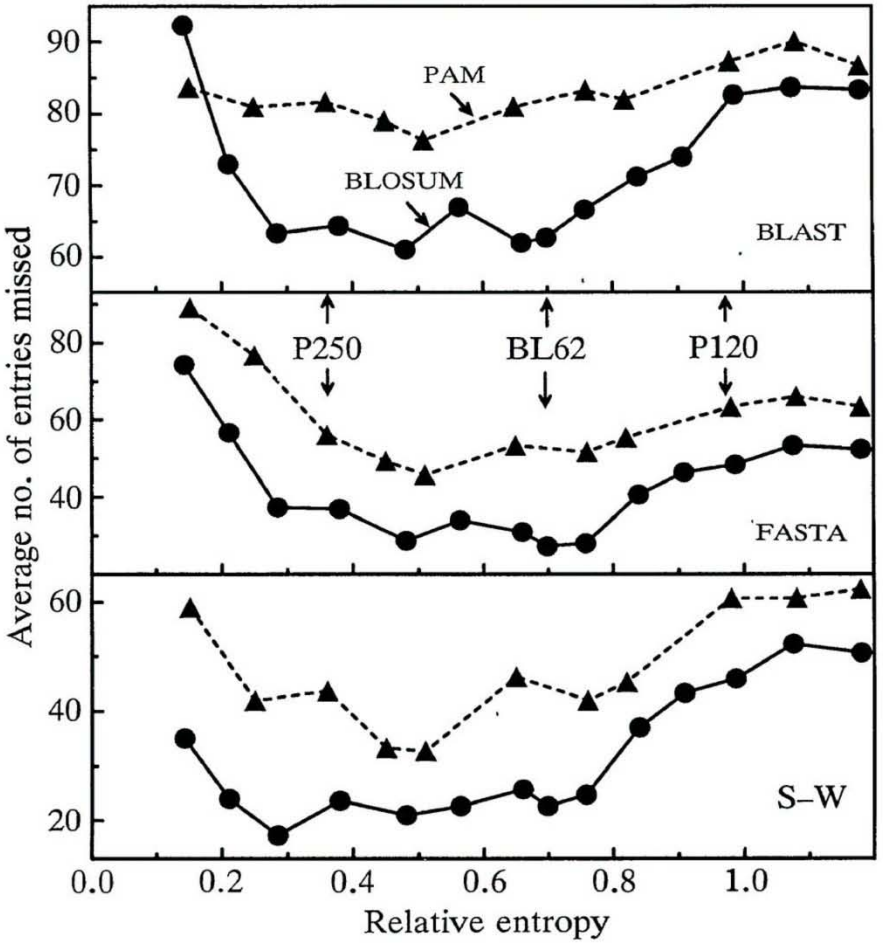


FIG. 3. Searching performance of programs using members of the guanine nucleotide-binding protein-coupled receptor family as queries and matrices from the BLOSUM and PAM series scaled in half-bits (11). Removal of this family from the BLOCKS data base led to a nearly identical matrix with similar performance. Matrices represented (left to right) are BLOSUM (BL) 30, 35, 40, 45, 50, 55, 60, 62, 65, 70, 75, 80, 85, and 90 and PAM (P) 400, 310, 250, 220, 200, 160, 150, 140, 120, 110, and 100. The average numbers of true positive Swiss-Prot entries missed are shown for LSHRRAT, RTARAT, and UL33$HCMVA versus Swiss-Prot 20. Results using BLAST and FASTA or SSEARCH (S–W) are not comparable to each other, since different detection criteria were used for the three programs.

One recipe for selecting a matrix

- Close sequences:
  PAM 100 or BLOSUM 80
- Distant sequences:
  PAM 250 or BLOSUM 45
- Database scanning:
  PAM 120 or BLOSUM 62

THERE IS NO "ONE SIZE FITS ALL" MATRIX !

Sequence Alignment Heuristics

Some slides from:

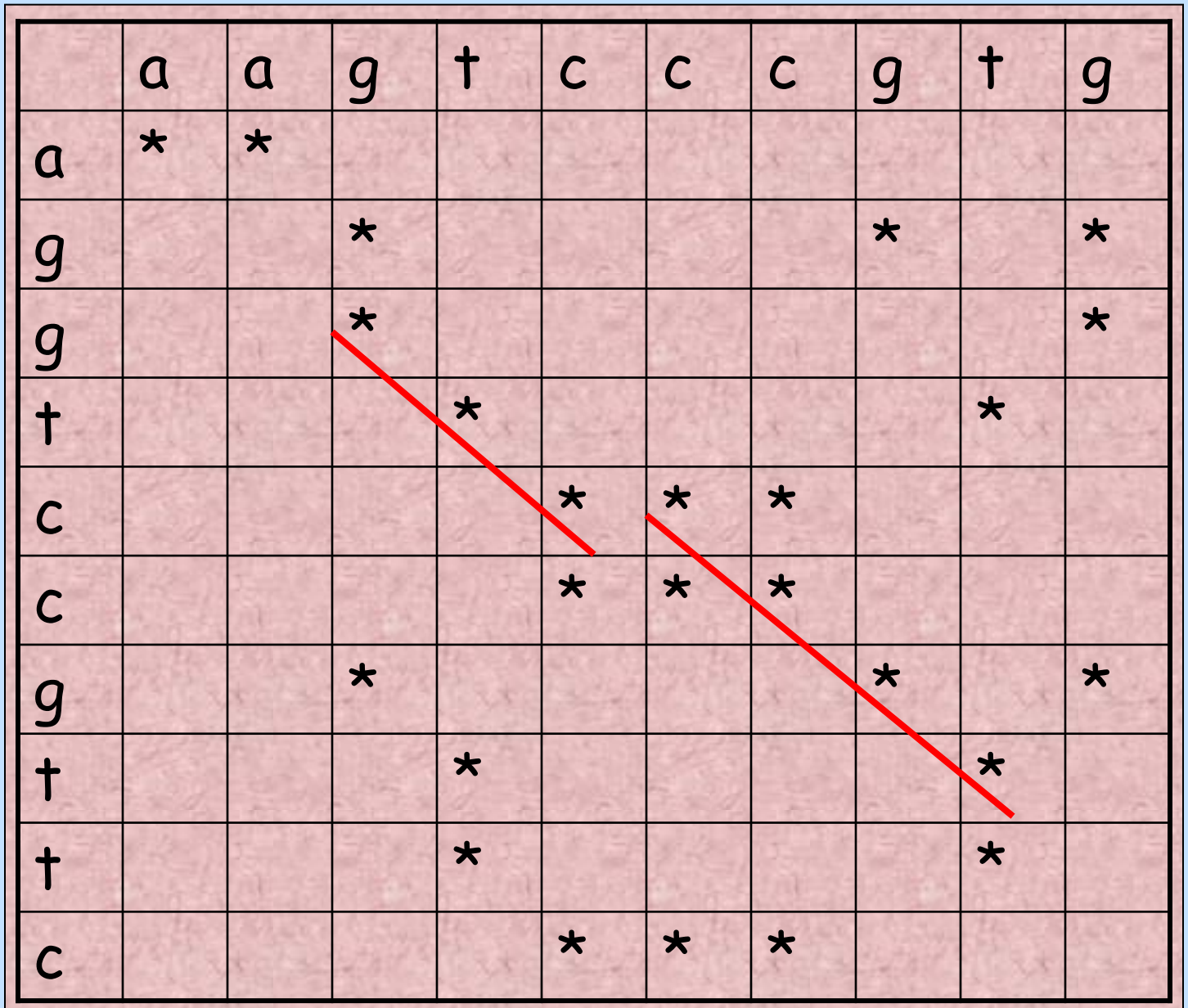
- Iosif Vaisman, GMU
  mason.gmu.edu/~mmasso/binf630alignment.ppt
- Serafim Batzoglu, Stanford
  http://ai.stanford.edu/~serafim/
- Geoffrey J. Barton, Oxford

"Protein Sequence Alignment and Database Scanning"
http://www.compbio.dundee.ac.uk/ftp/preprints/review93/review93.pdf

Why Heuristics ?

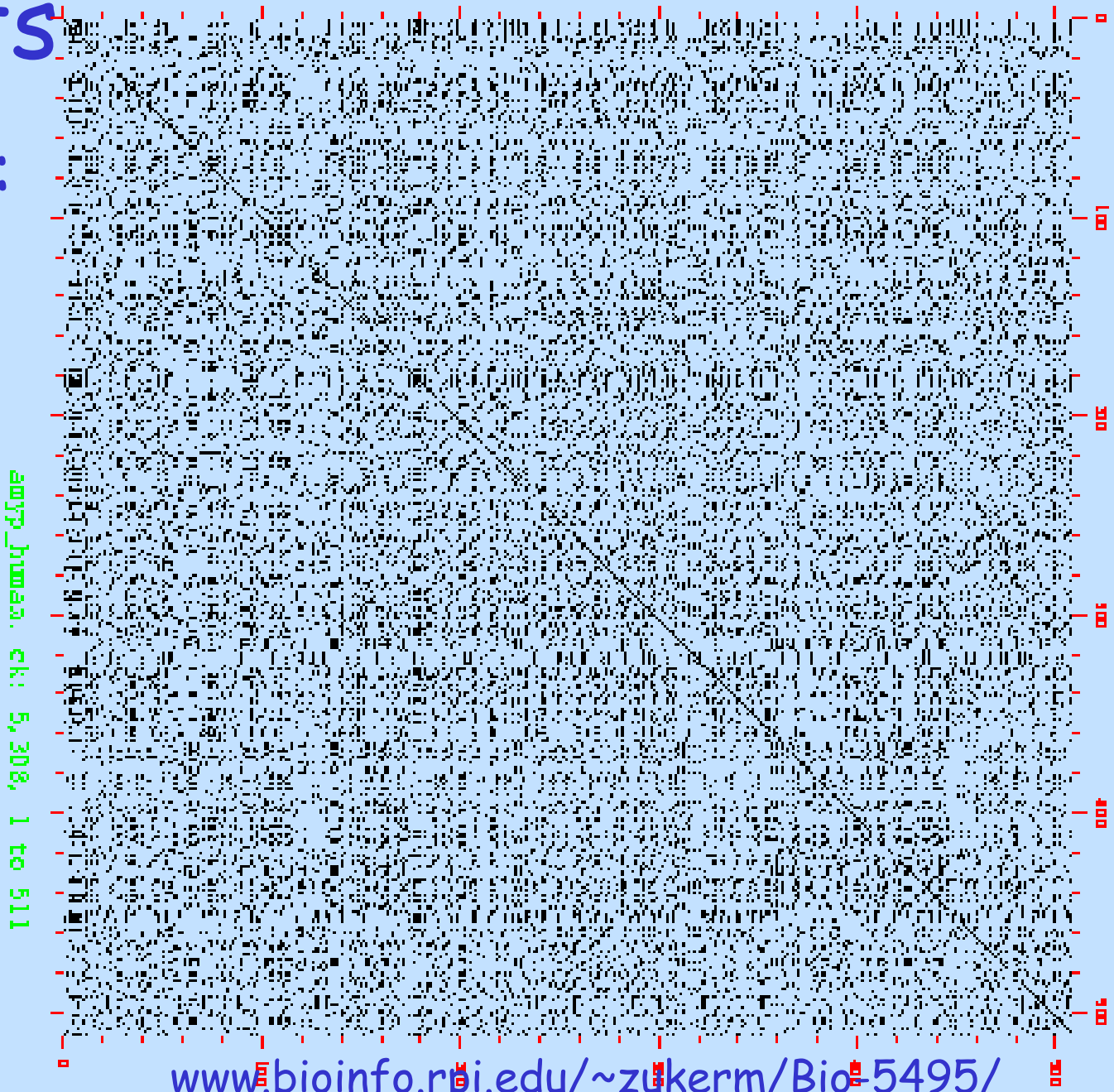
- **Motivation**:
  - Dynamic programming guarantees an optimal solution & is efficient, but
  - *Not fast enough* when searching a database of size ~10^{12}, with a query of length 200-500bp
- **Solutions**:
  - Implement on hardware. (e.g. COMPUGEN)
  - Use faster heuristic algorithms.
  - Database preprocessing
- **Common Heuristics**: FASTA, BLAST

Alignment Dot-Plot Matrix

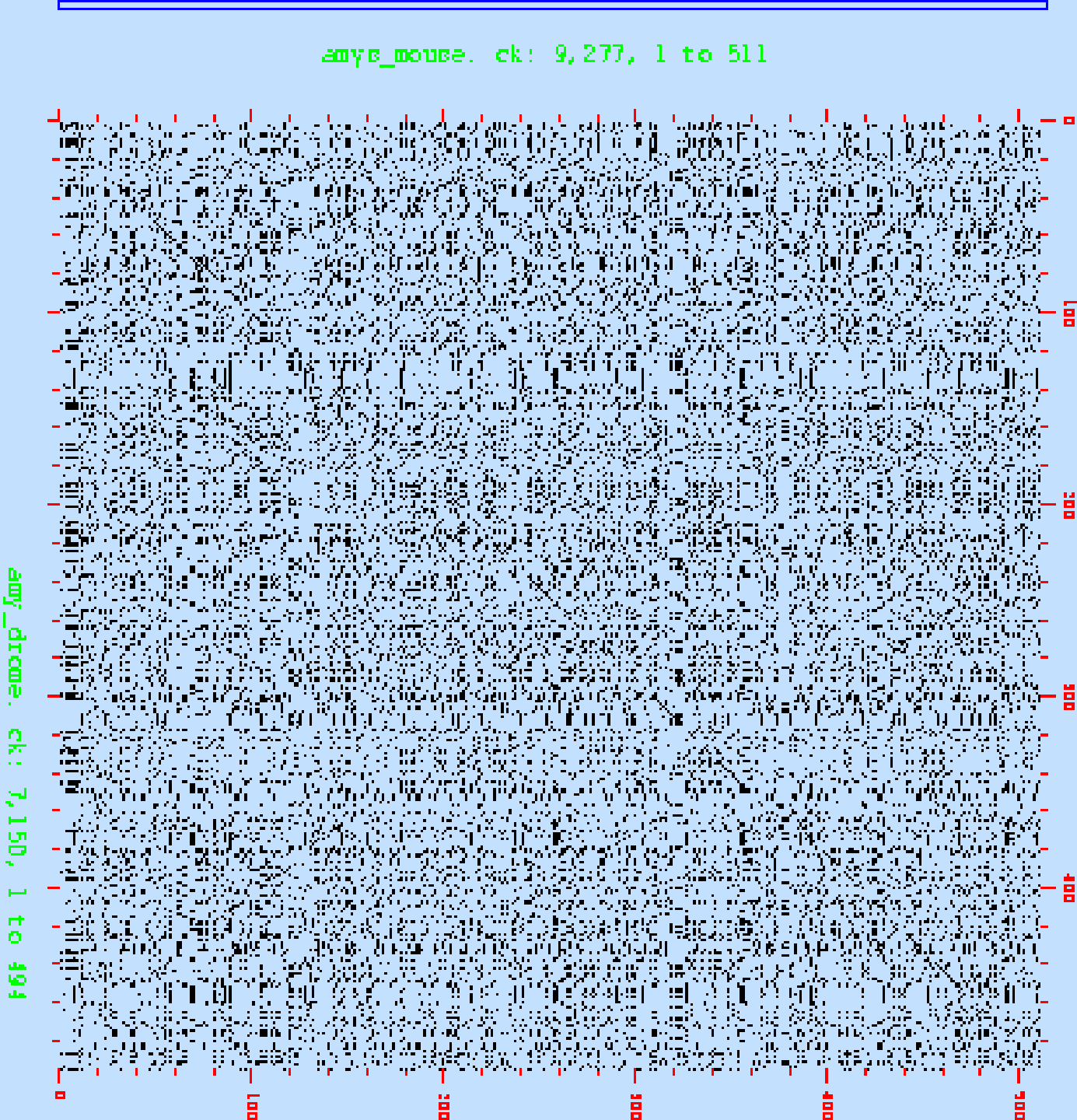


Dot plots

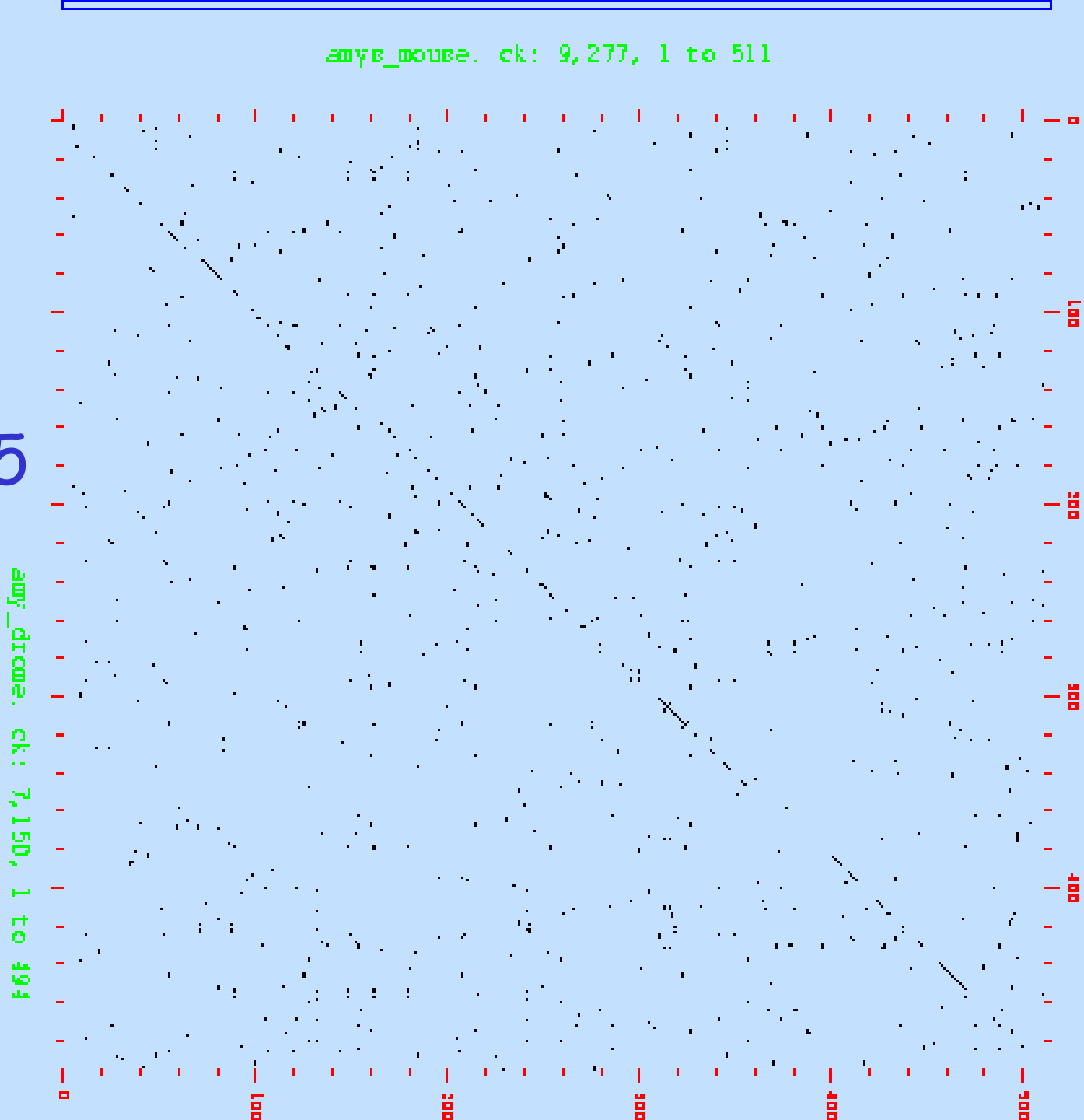
Example 1: close protein homologs (man and mouse)



Example 2: remote protein homologs (man and bacilus)



Example 2: dot for 4+ matches in window of 5



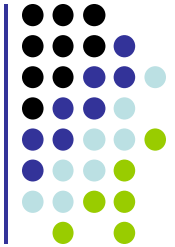
Key observations

- Substitutions are much more likely than indels
- Homologous sequences contain many matches

- Even $O(m+n)$ time would be problematic when db size is huge
- Numerous queries are run on the same db

➔ Preprocessing of the db is desirable

Banded Alignment



Assume we know that x and y are very similar

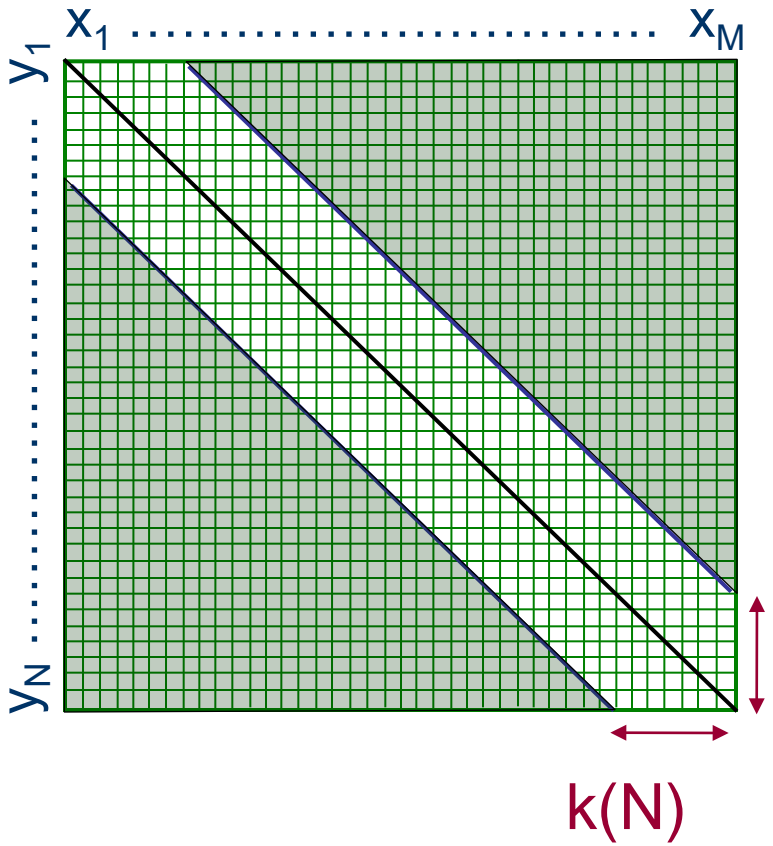
Assumption: # gaps(x, y) < k(N) (say N>M)

Then, x_i | y_j implies | i – j | < k(N)

We can align x and y more efficiently:

Time, Space: $O(N \times k(N)) << O(N^2)$

Banded Alignment



Initialization:

F(i,0), F(0,j) undefined for i, j > k

Iteration:

For i = 1…M

For j = max(1, i – k)…min(N, i+k)

$$F(i,j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i, j-1) - d, \text{ if } j > i - k(N) \\ F(i-1, j) - d, \text{ if } j < i + k(N) \end{cases}$$

Termination: same

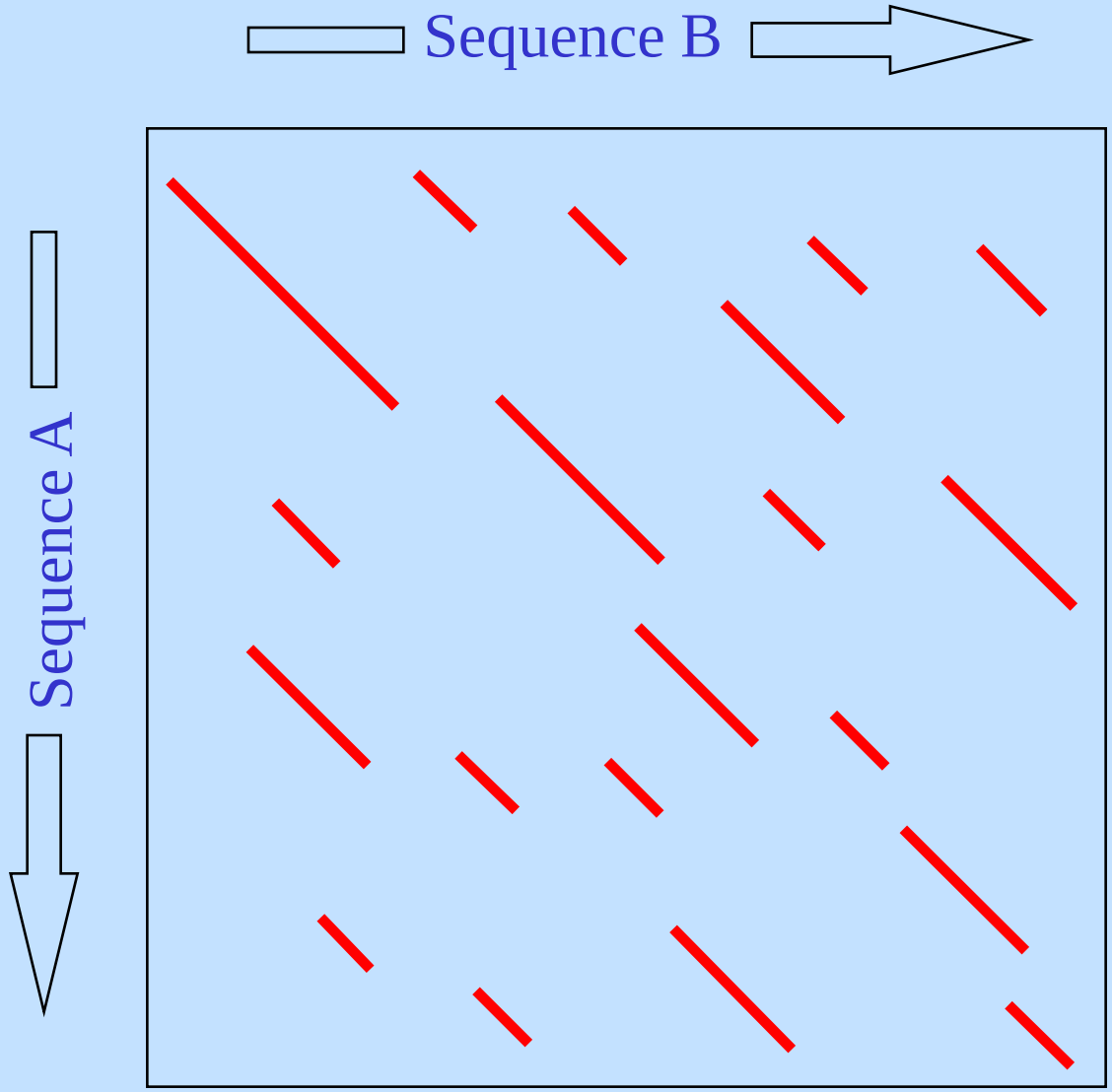
FASTA (Lipman & Pearson '88)

Key idea: Good local alignment must have exact matching subsequences.

ktup = required min length of perfect match

1. Find 10 highest-scoring diagonal runs = almost consecutive matches of length ktup on the same diagonal
2. Rescore using a subs. matrix. Best soln = init1
3. Combine close sub-alignments. best soln = initn
4. Compute best DP solution in a band around init1. result = opt

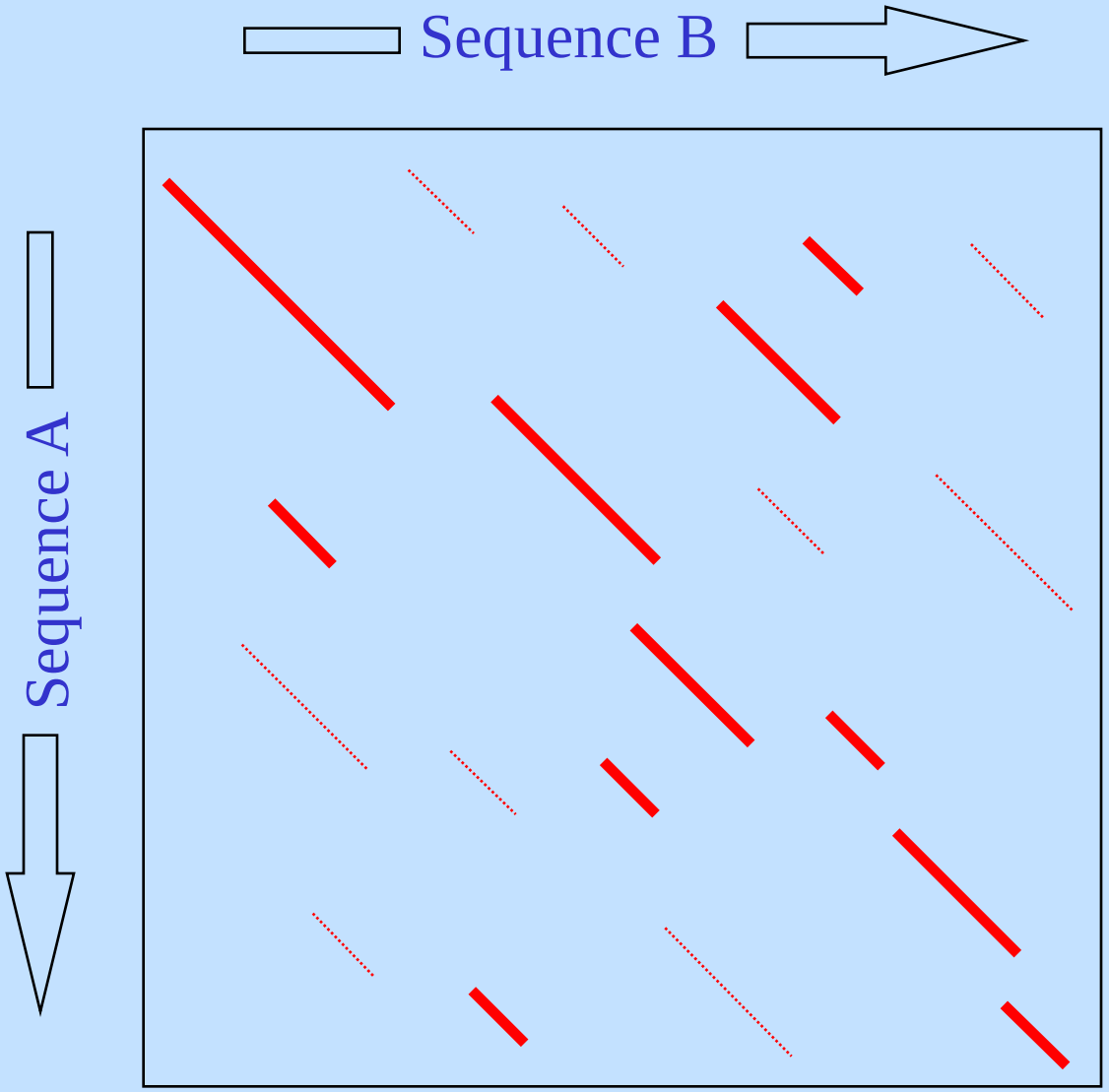
FASTA - Step 1



Find diagonal runs of matches of length **ktup**

4-6 for DNA,1-2 for AA

FASTA - Step 2



2

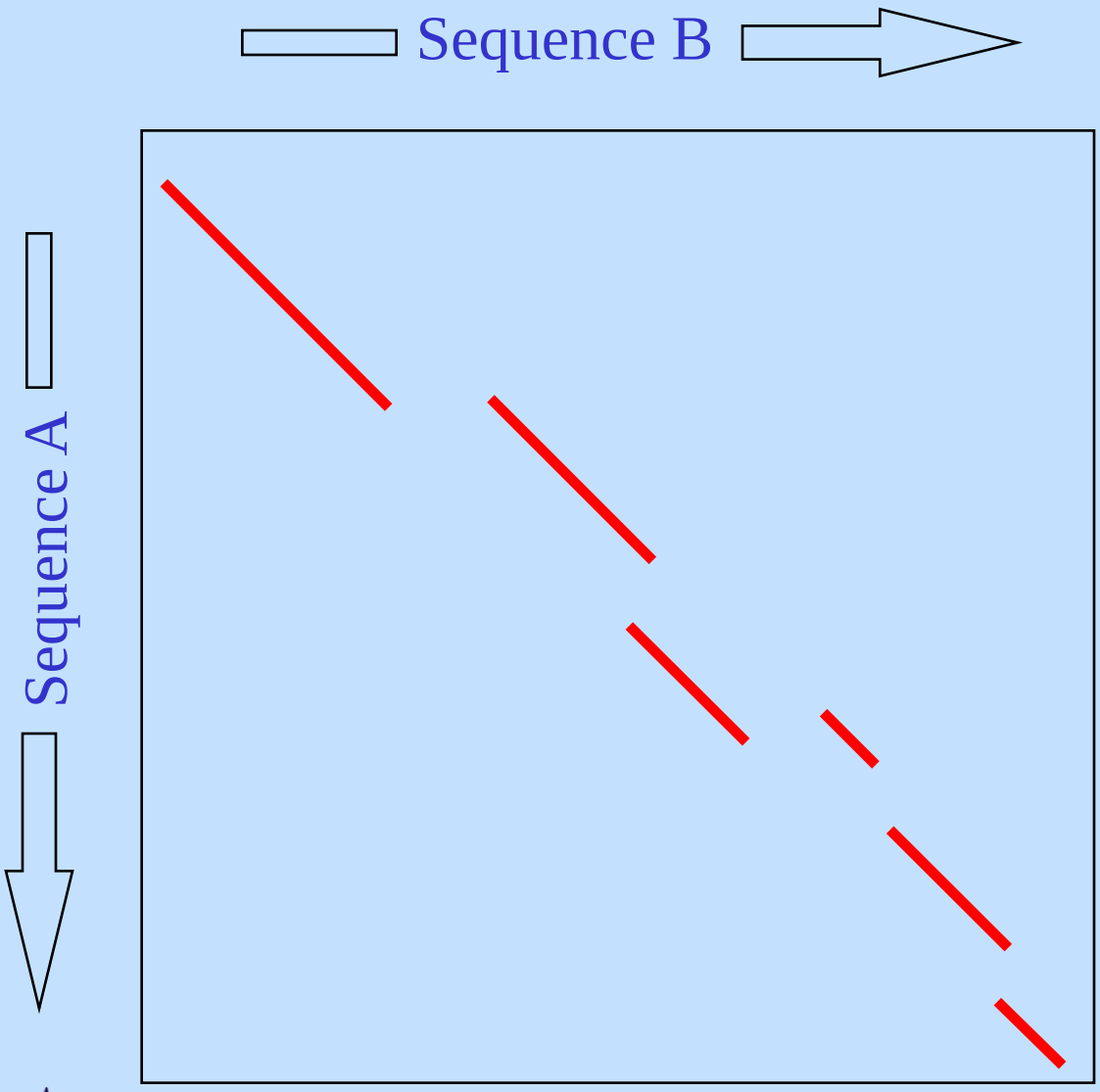
Rescoring using a subs. matrix

high score

low score

The score of the highest scoring initial region is saved as the **init1 score**.

FASTA - Step 3



3

Join sub-alignments (allow indels)

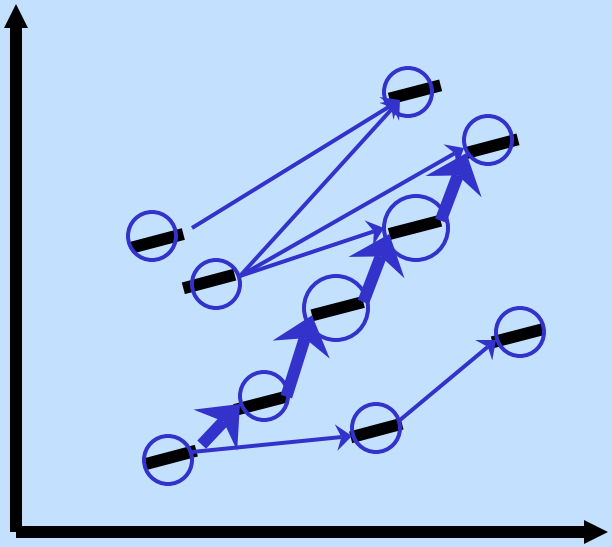
Non-overlapping regions are joined. The score equals sum of the scores of the regions minus a gap penalty. The score of the highest scoring region is the **initn score**.

Combining diagonal runs

- Construct an alignment graph:

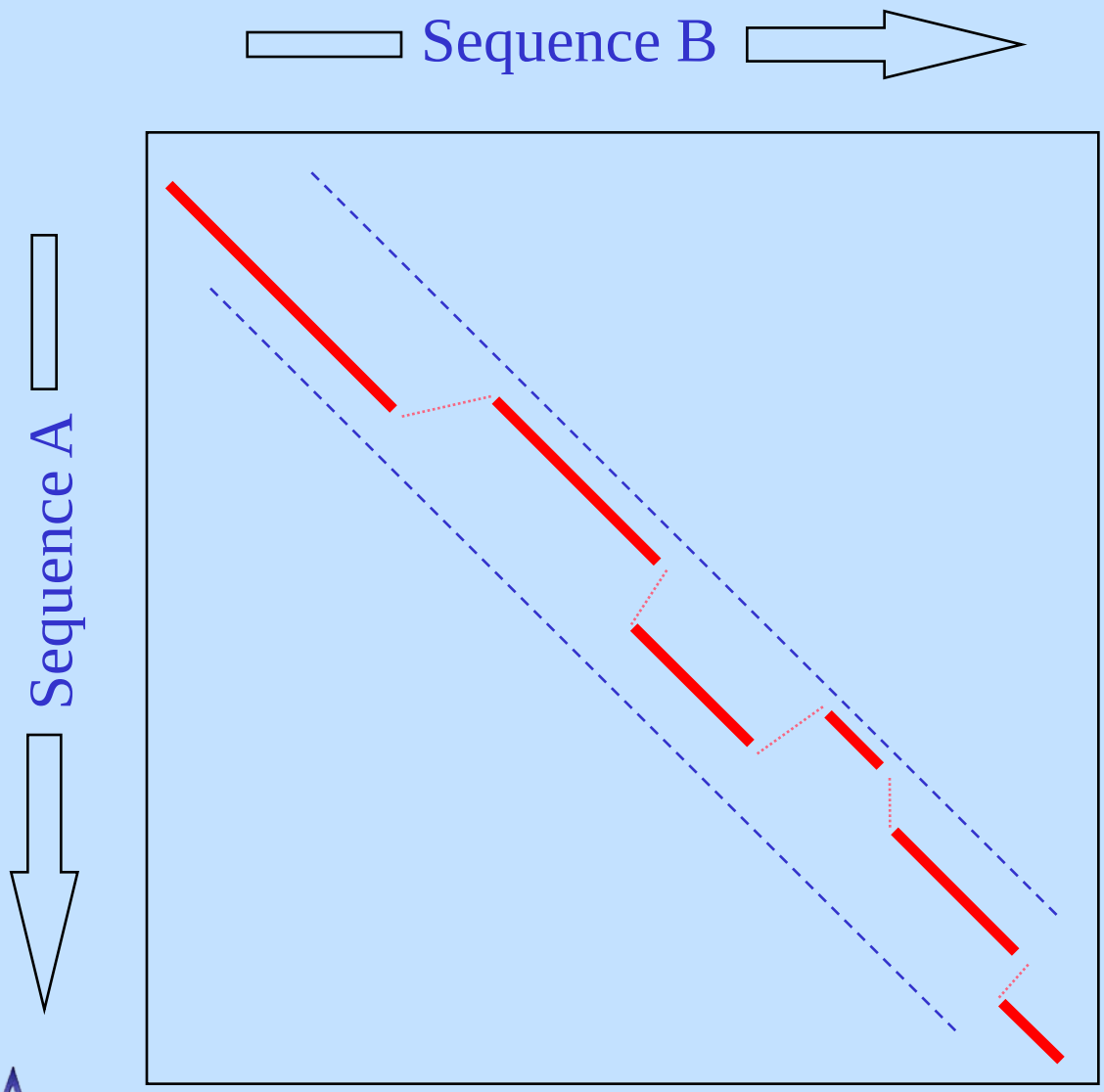
• nodes =sub-alignments (SAs)
• weight – alignment score (from 1)
• Edges btw SAs that can fit together,
• Weight - negative, depends on the size of the corresponding gap

- Find a maximum weight path in it, *initn*
- Use initn for an initial ranking of sequences.



Alignment graph

FASTA - Step 4



4

Banded alignment
Around init1
(width=16/32)

The score for this alignment is the **opt score**.

FASTA Output

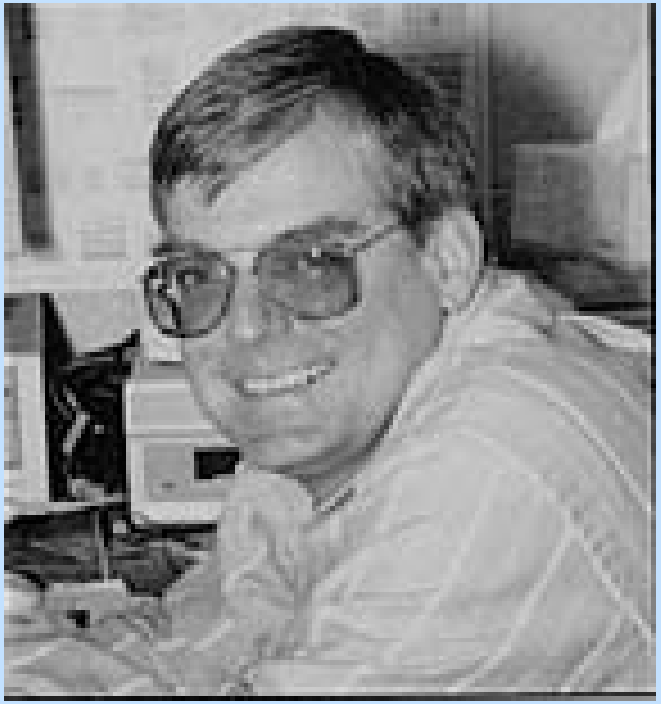
```
>>SWNEW:HBE_HYLSY Q95190 HEMOGLOBIN EPSILON CHAIN.          (146 aa)
 initn: 638 init1: 638 opt: 638 Z-score: 1255.8 expect() 5.2e-64
Smith-Waterman score: 638;  80.690% identity in 145 aa overlap (3-147:2-146)

               10        20        30        40        50        60
GGAMMA MGHFTEEDKATITSLWGKVNVEDAGGETLGRLLVVYPWTQRFFDSFGNLSSASAIMGNPK
         ::: :.::..::::.:.::.::::.:::::::::::::::::::::::::: :::.:::
SWNEW:  VHFTAEEKAAVTSLWNKMNVEEAGGEALGRLLVVYPWTQRFFDSFGNLSSPSAILGNPK
                10        20        30        40        50

               70        80        90       100       110       120
GGAMMA VKAHGKKVLTSLGDAIKHLDDLKGTFAQLSELHCDKLHVDPENFKLLGNVLVTVLAIHFG
       :::::::::::.::::..:.:: :::.:::::::::::::::::::.: .:: :::
SWNEW: VKAHGKKVLTSFGDAIKNMDNLKTTFAKLSELHCDKLHVDPENFKLLGNVMVIILATHFG
      60        70        80        90       100       110

              130       140
GGAMMA KEFTPEVQASWQKMVTGVASALSSRYH
       :::::::::.:::.:..:: ::. .::
SWNEW: KEFTPEVQAAWQKLVSAVAIALAHKYH
     120       130       140
```

The information on each hit includes:

- General information and statistics
- SW score, %identity and length of overlap

August 1997: NCBI Director David Lipman (far left) coaches Vice President Gore (seated) as he searches PubMed. NIH Director Harold Varmus (center) and NLM Director Donald Lindberg look on.

Bill Pearson

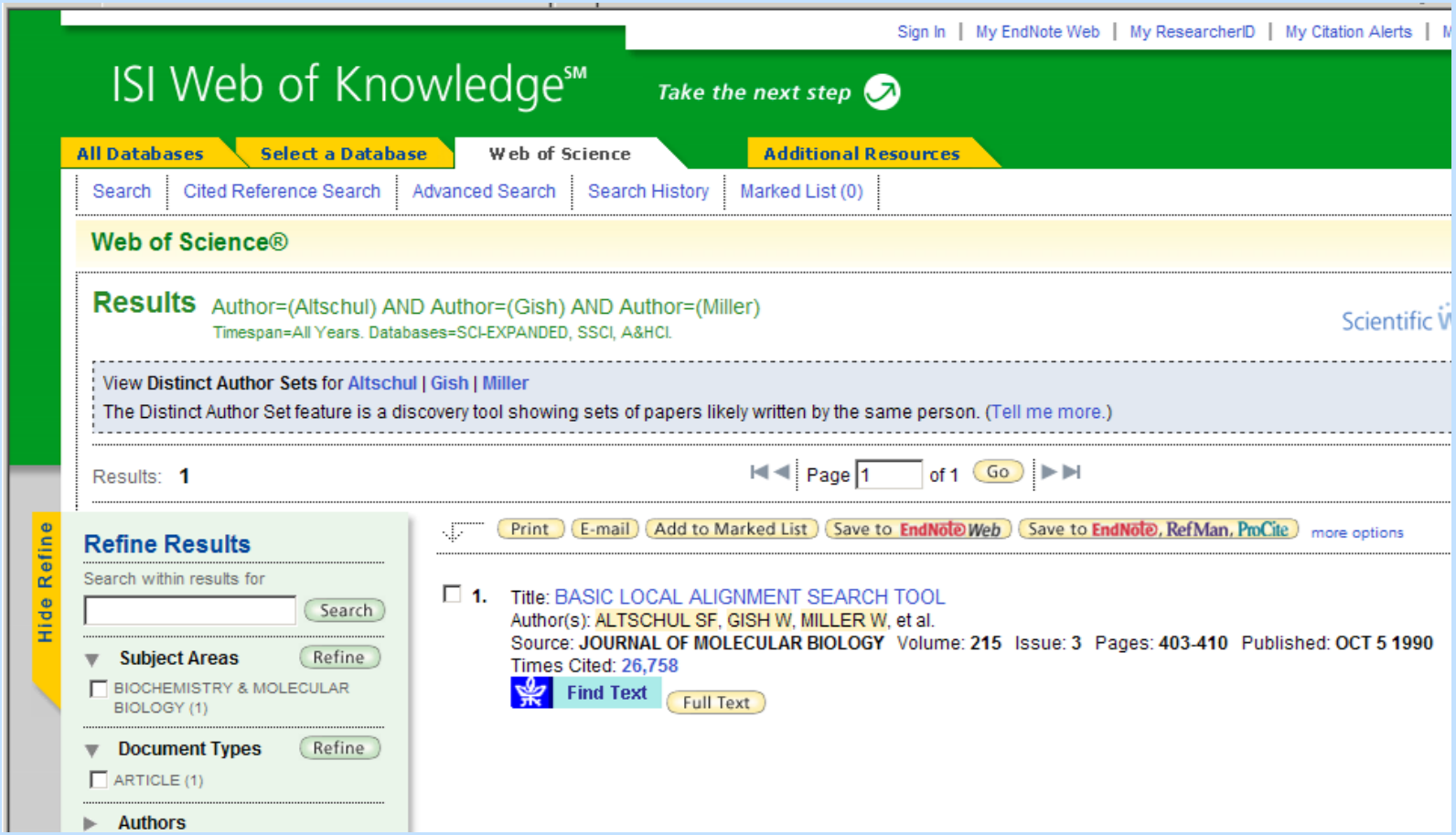


Bill Pearson received his Ph.D. in Biochemistry in 1977 from the California Institute of Technology. He then did a post-doctoral fellowships at the Caltech Marine Station in Corona del Mar, CA and at the Department of Molecular Biology and Genetics at Johns Hopkins. In 1983 he joined the Department of Biochemistry at the University of Virginia.

BLAST

Basic Local Alignment Search Tool

Altschul, Gish, Miller, Myers and Lipman 1990

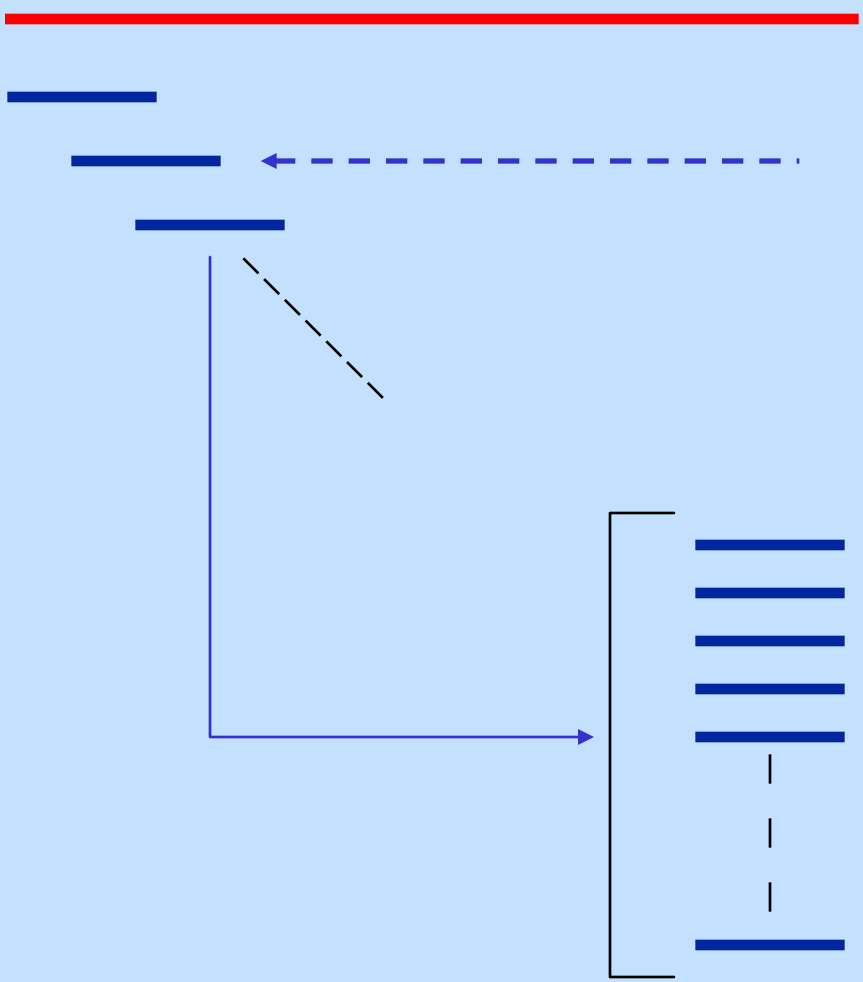


BLAST – outline

- Compile a list of high scoring words with the query
- Scan the database for hits
- Extend hits

BLAST Algorithm

1



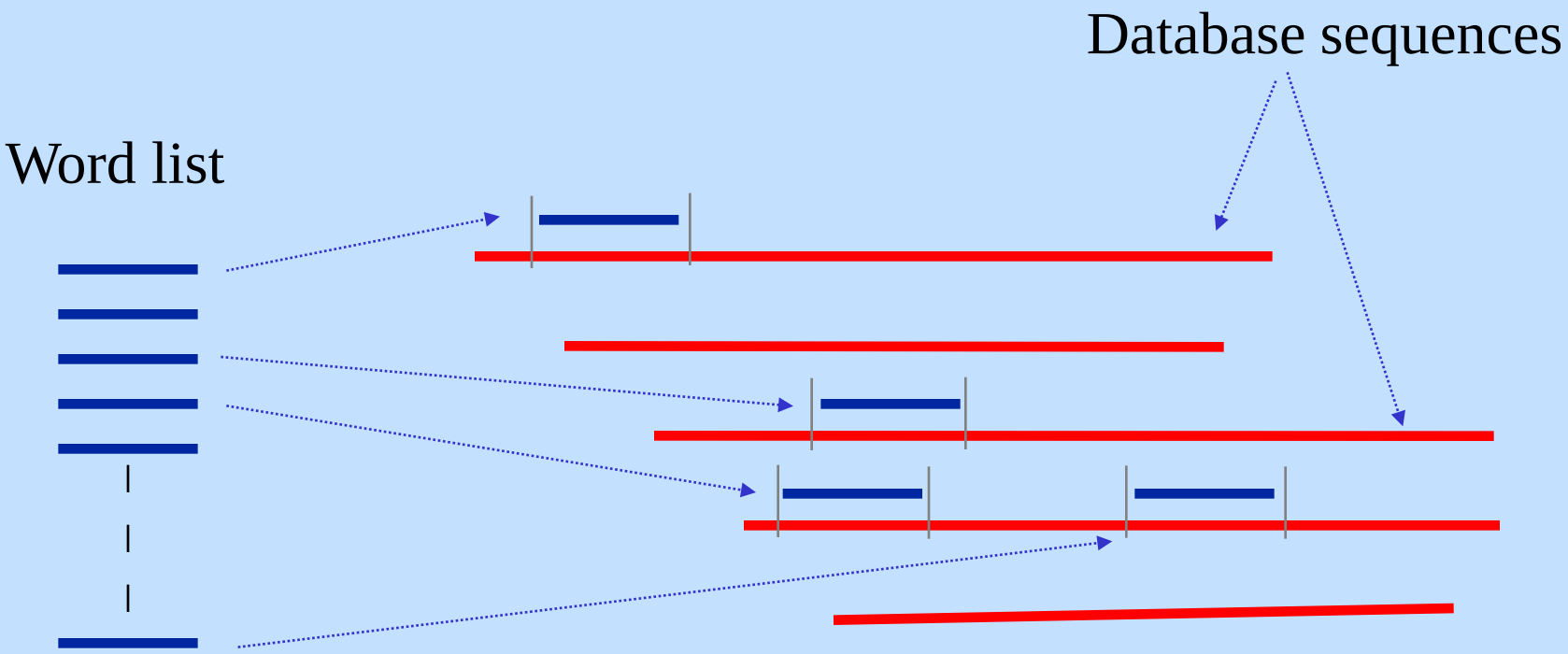
Query sequence of length L

Maximium of L-w+1 words
(typically w = 3 for proteins)

For each word from the query sequence find the list of words with score >=T using a substitution matrix

BLAST Algorithm

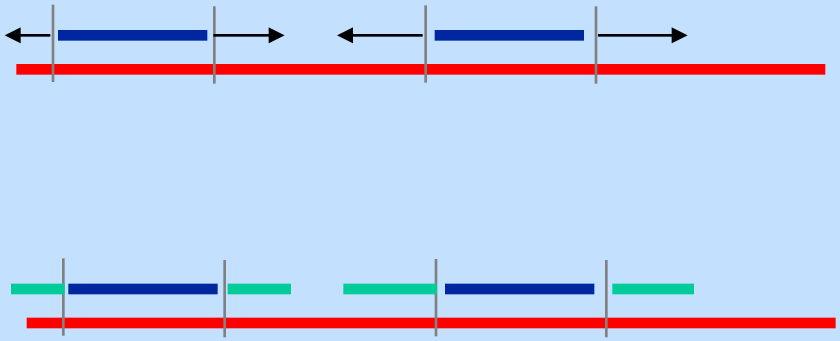
2



Exact matches of words from the word list
to the database sequences (linear time)

BLAST Algorithm

3



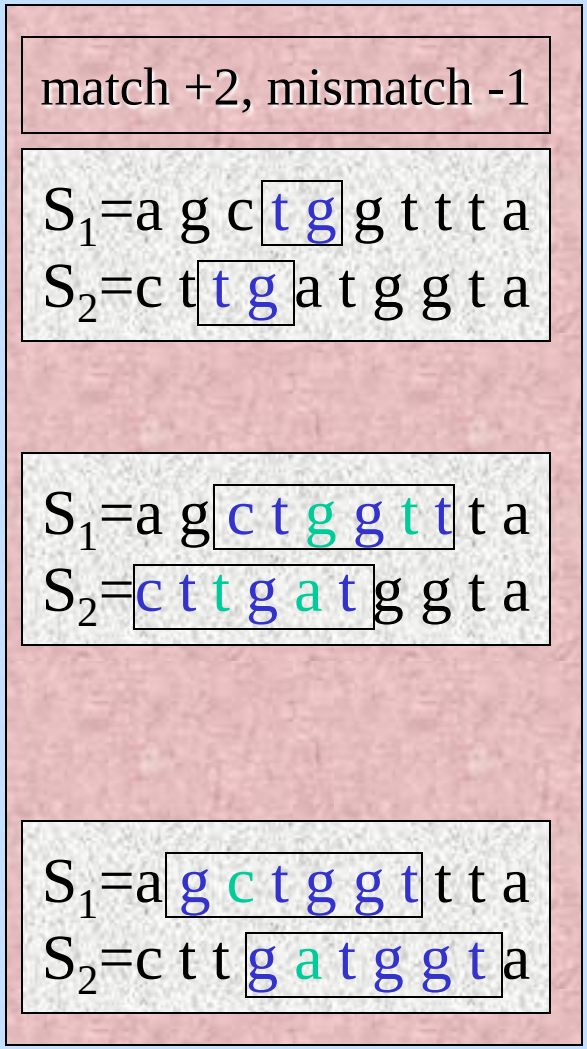
Locally Maximal Segment Pairs (MSPs)

For each exact word match, alignment is extended in both directions to find high scoring segments

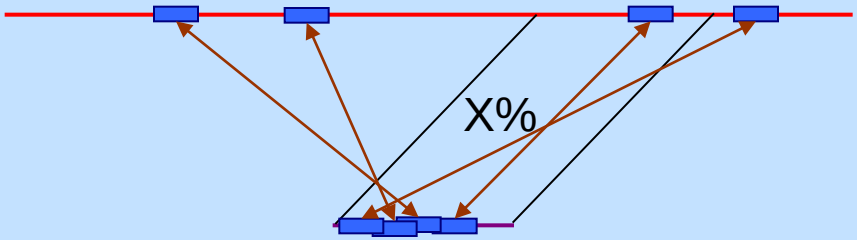
In more detail

BLAST - Basic Definitions

- Given two sequences S_1 and S_2, a ***segment pair*** is a pair of equal length subsequences of S_1 and S_2, aligned without spaces.
- A ***locally maximal segment pair*** is a pair aligned without spaces whose alignment score cannot be improved by extending it or shortening it.
- A ***maximal segment pair*** (***MSP***) in S_1, S_2 is a segment pair with the maximum score over **all** segment pairs in S_1, S_2.



Sensitivity-Speed Tradeoff



	long words (k = 15)	short words (k = 7)
Sensitivity		✓
Speed	✓	

Table 3. Sensitivity and Specificity of Single Perfect Nucleotide K-mer Matches as a Search Criterion

	7	8	9	10	11	12	13	14
A. 81%	0.974	0.915	0.833	0.726	0.607	0.486	0.373	0.314
83%	0.988	0.953	0.897	0.815	0.711	0.595	0.478	0.415
85%	0.996	0.978	0.945	0.888	0.808	0.707	0.594	0.532
87%	0.999	0.992	0.975	0.942	0.888	0.811	0.714	0.659
89%	1.000	0.998	0.991	0.976	0.946	0.897	0.824	0.782
91%	1.000	1.000	0.998	0.993	0.981	0.956	0.912	0.886
93%	1.000	1.000	1.000	0.999	0.995	0.987	0.968	0.957
95%	1.000	1.000	1.000	1.000	0.999	0.998	0.994	0.991
97%	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999
B. K	7	8	9	10	11	12	13	14
F	1.3e+07	2.9e+06	635783	143051	32512	7451	1719	399

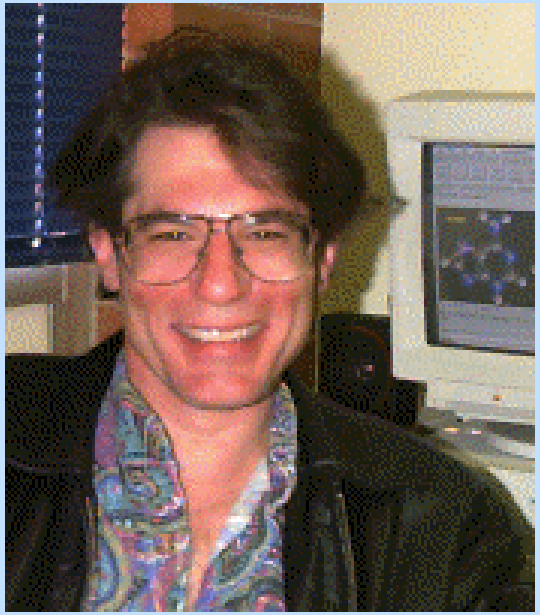
(*A*) Columns are for K sizes of 7–14. Rows represent various percentage identities between the homologous sequences. The table entries show the fraction of homologies detected as calculated from equation 3 assuming a homologous region of 100 bases. The larger the value of K, the fewer homologies are detected.
(*B*) K represents the size of the perfect match. F shows how many perfect matches of this size expected to occur by chance according to equation 4 in a genome of 3 billion bases using a query of 500 bases.

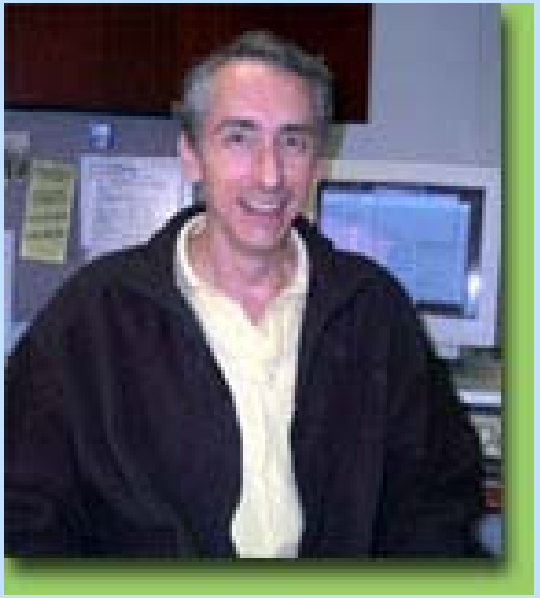
Sens.

Speed

Kent WJ, Genome Research 2002

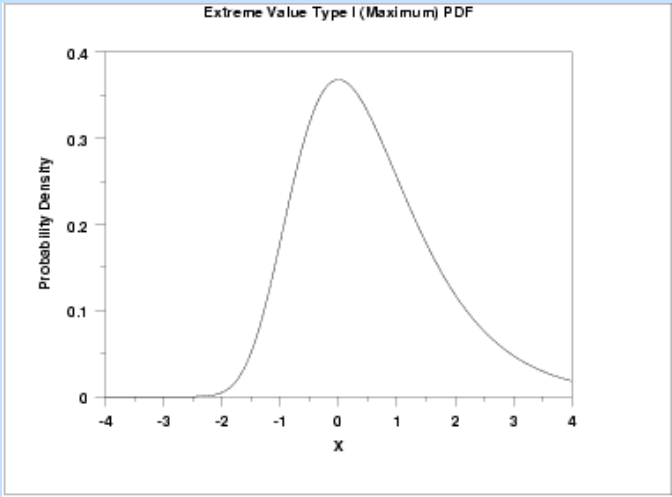
Gene Myers, Webb Miller, Warren Gish



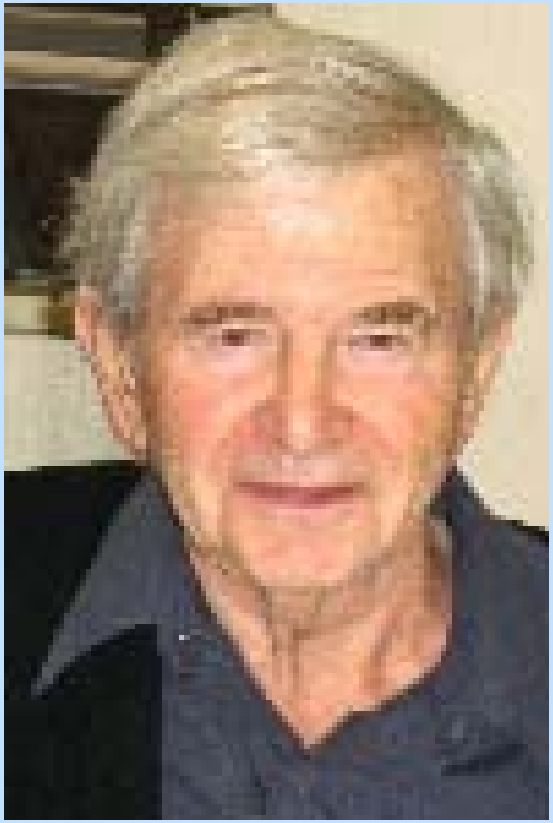


BLAST statistics

- Theory of Karlin, Altschul, and Dembo on the distribution of the MSP score at random: the maximum of mn local match scores has an **Extreme value distribution**
- Define parameters K, λ (depending on AA distribution and scoring matrix).
- Pr (finding a pair of score >S in comparing two random seqs of length m, n) = $1 - e^{-Y}$ where $Y = Kmn\, e^{-\lambda s}$
- Generalizes to db search: n->N

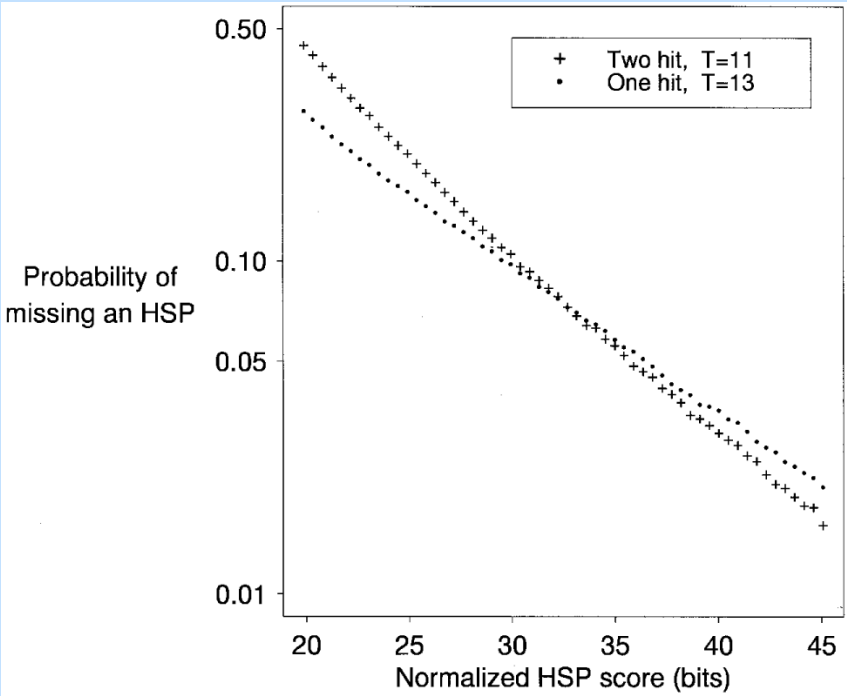


Sam Karlin, Steve Altschul, Amir Dembo



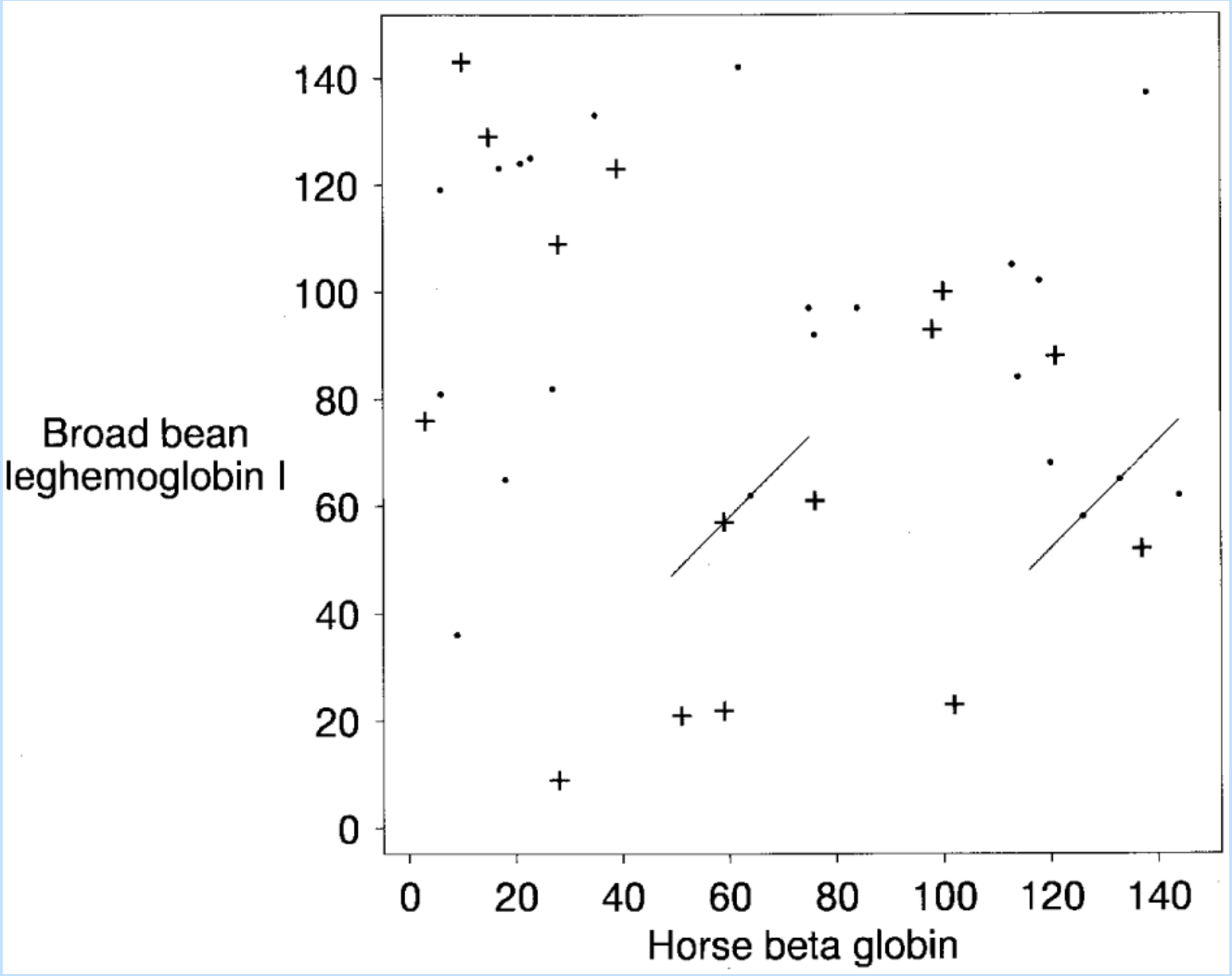
Gapped BLAST (Altschul et al. 97)

- The original BLAST extends high-scoring SPs (HSPs) without gaps.
- The new version allows gapped extensions for the best segments passing the two hit condition: two close hits on the same diagonal



Gapped BLAST outline

- Find two non-overlapping w-long words with:
  - score $\geq$ T, each
  - on same diagonal
  - within distance $\leq$ A
- Perform ungapped extension
- If score exceeds S (1:50 sequences), perform gapped extension; use center pair as seed.
- Apply DP on a changing region: stop extension when score falls X_g below best score attained so far



- **Figure** 2. The BLAST comparison of broad bean leghemoglobin I (87) (SWISS-PROT accession no. P02232) and horse [beta]-globin (88) (SWISS-PROT accession no. P02062). The 15 hits with score at least 13 are indicated by plus signs. An additional 22 non-overlapping hits with score at least 11 are indicated by dots. Of these 37 hits, only the two indicated pairs are on the same diagonal and within distance 40 of one another. Thus the two-hit heuristic with $T = 11$ triggers two extensions, in place of the 15 extensions invoked by the one-hit heuristic with $T = 13$.

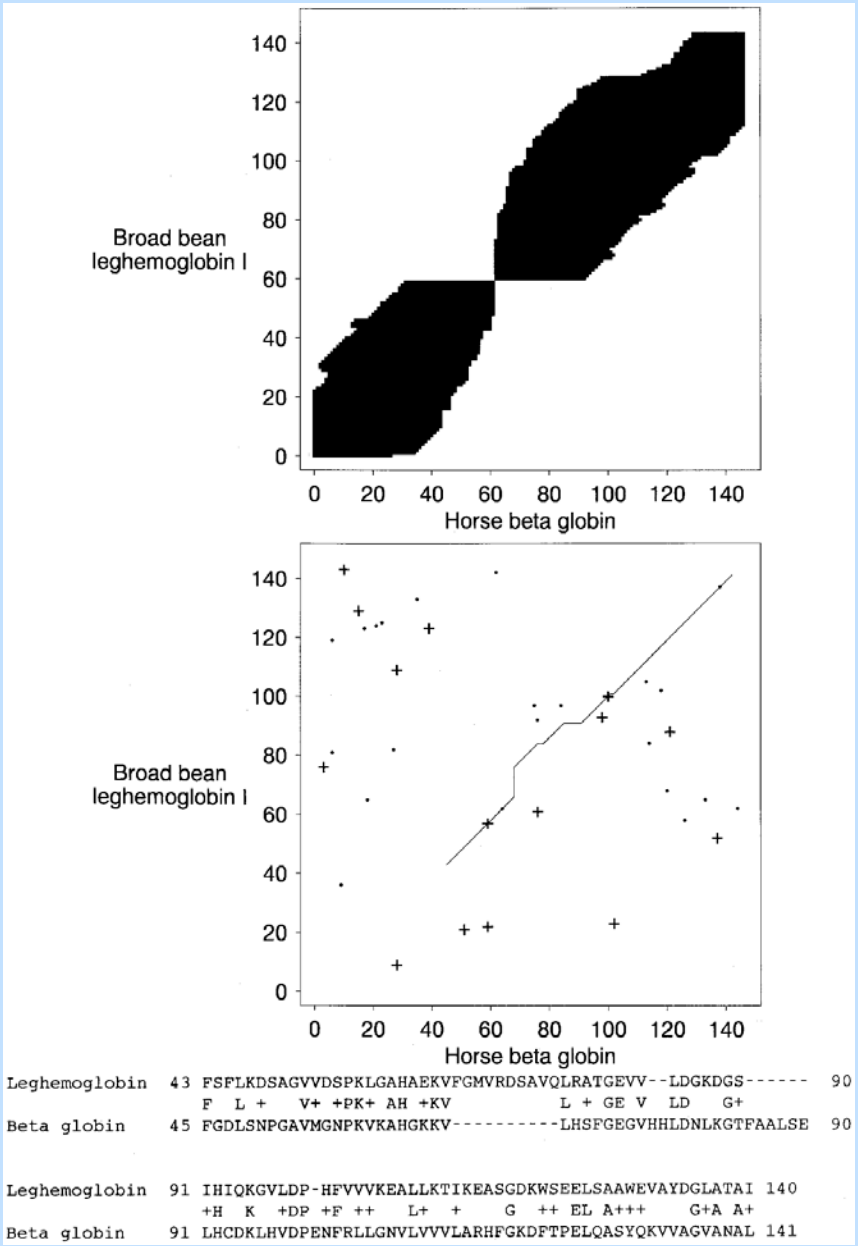


Figure 3. A gapped extension generated by BLAST for the comparison of broad bean leghemoglobin I (87) and horse [beta]-globin (88). (**a**) The region of the path graph explored when seeded by the alignment of alanine residues at respective positions 60 and 62. This seed derives from the HSP generated by the leftward of the two ungapped extensions illustrated in Figure 2. The *X*g dropoff parameter is the nominal score 40, used in conjunction with BLOSUM-62 substitution scores and a cost of 10 + *k* for gaps of length *k*. (**b**) The path corresponding to the optimal local alignment generated, superimposed on the hits described in Figure 2. The original BLAST program, using the one-hit heuristic with *T* = 11, is able to locate three of the five HSPs included in this alignment, but only the first and last achieve a score sufficient to be reported. (**c**) The optimal local alignment, with nominal score 75 and normalized score 32.4 bits. In the context of a search of SWISS-PROT (26), release 34 (21 219 450 residues), using the leghemoglobin sequence (143 residues) as query, the *E*-value is 0.54 if no edge-effect correction (22) is invoked. The original BLAST program locates the first and last ungapped segments of this alignment. Using sum-statistics with no edge-effect correction, this combined result has an *E*-value of 31 (21,22). On the central lines of the alignment, identities are echoed and substitutions to which the BLOSUM-62 matrix (18) gives a positive score are indicated by a `+'

Time analysis

	Overhead: database scanning, output, etc.	Calculating whether hits qualify for ungapped extension	Ungapped extensions	Gapped extensions
Original BLAST	**8 (8%)**		**92 (92%)**	
Gapped BLAST	**8 (24%)**	**12 (37%)**	**5 (15%)**	**8 (24%)**

Speed: ~3 times faster than the original BLAST

Thomas Madden, David Lipman, Alex Schaeffer, Steve Altschul