Bioinformatics

Sequence Analysis: Part I. Pairwise alignment and database searching

Fran Lewitter, Ph.D.
Head, Biocomputing
Whitehead Institute

Bioinformatics Definitions

"The use of computational methods to make biological discoveries."

Fran Lewitter

Bioinformatics Definitions

"The use of computational methods to make biological discoveries."

Fran Lewitter

"An interdisciplinary field involving biology, computer science, mathematics, and statistics to analyze biological sequence data, genome content, and arrangement, and to predict the function and structure of macromolecules."

David Mount

Course Syllabus

Jan 7	Sequence Analysis I. Pairwise alignments, database searching including BLAST (FL) [1, 2, 3]
Jan 14	Sequence Analysis II. Database searching (continued), Pattern searching(FL)[7]
Jan 21	No Class - Martin Luther King Holiday
Jan 28	Sequence Analysis III. Hidden Markov models, gene finding algorithms (FL)[8]
Feb 4	Computational Methods I. Genomic Resources and Unix (GB)
Feb 11	Computational Methods II. Sequence analysis with Perl. (GB)
Feb 18	No Class - President's Birthday
Feb 25	Computational Methods III. Sequence analysis with Perl and BioPerl (GB)
Mar 4	Proteins I. Multiple sequence alignments, phylogenetic trees (RL) [4, 6]
Mar 11	Proteins II. Profile searches of databases, revealing protein motifs (RL) [9]
Mar 18	Proteins III.Structural Genomics:structural comparisons and predictions (RL)
Mar 25	Microarrays: designing chips, clustering methods (FL)

Course Information

- Lectures
- Text book
- Supplemental reading
- Homework
- Course project
- Office hours - M 2-4, W 2-4
- http://fladda.wi.mit.edu/bioinfo
- Mailing list, lewitter@wi.mit.edu, Subject: course

Topics to Cover

- Introduction
- Scoring alignments
- Alignment methods
- Significance of alignments
- Database searching methods
- Demo

Topics to Cover

- Introduction
  - Why do alignments?
  - A bit of history
  - Definitions
- Scoring alignments
- Alignment methods
- Significance of alignments
- Database searching methods
- Demo

Simian sarcoma virus onc gene, v-sis, is derived from the gene (or genes) encoding a platelet-derived growth factor.

Doolittle RF, Hunkapiller MW, Hood LE, Devare SG, Robbins KC, Aaronson SA, Antoniades HN. *Science* 221:275-277, 1983.

Cancer Gene Found

Cell, Vol. 75, 1027–1038, December 3, 1993, Copyright © 1993 by Cell Press

The Human Mutator Gene Homolog *MSH2* and Its Association with Hereditary Nonpolyposis Colon Cancer

Richard Fishel,* Mary Kay Lescoe,* M. R. S. Rao,§ Neal G. Copeland,† Nancy A. Jenkins,† Judy Garber,‡ Michael Kane,§ and Richard Kolodner§
*Department of Microbiology and Molecular Genetics
Markey Center for Molecular Genetics

can give rise to mismatched bases (Friedberg, 1985). For example, the deamination of 5-methylcytosine creates a thymine and, therefore, a G·T mispair (Duncan and Miller, 1980). Second, misincorporation of nucleotides during DNA replication can yield mismatched base pairs and nucleotide insertions and deletions (Modrich, 1991). Finally,

Homology to bacterial and yeast genes shed new light on human disease process

Evolutionary Basis of Sequence Alignment

- Similarity - observable quantity, such as per cent identity
- Homology - conclusion drawn from data that two genes share a common evolutionary history; no metric is associated with this

Some Definitions

An *alignment* is a mutual arrangement of two sequences, which exhibits where the two sequences are similar, and where they differ.

An *optimal alignment* is one that exhibits the most correspondences and the least differences. It is the alignment with the highest score. May or may not be biologically meaningful.

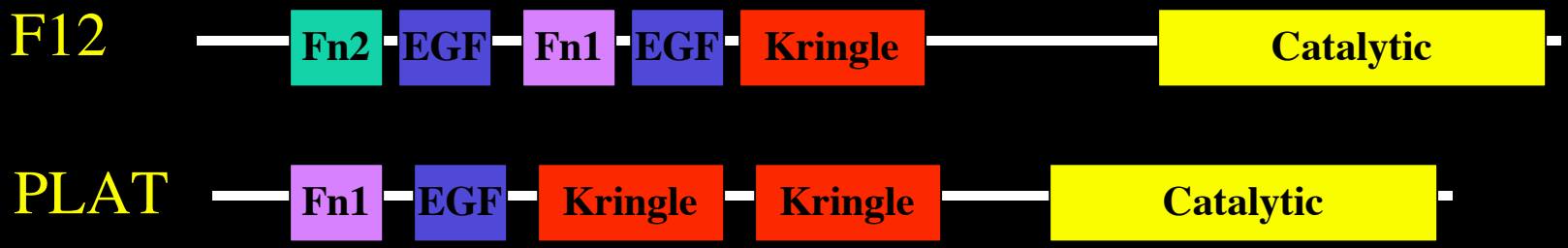
Alignment Methods

- Global alignment - Needleman-Wunsch (1970) maximizes the number of matches between the sequences along the entire length of the sequences.
- Local alignment - Smith-Waterman (1981) is a modification of the dynamic programming algorithm gives the highest scoring local match between two sequences.

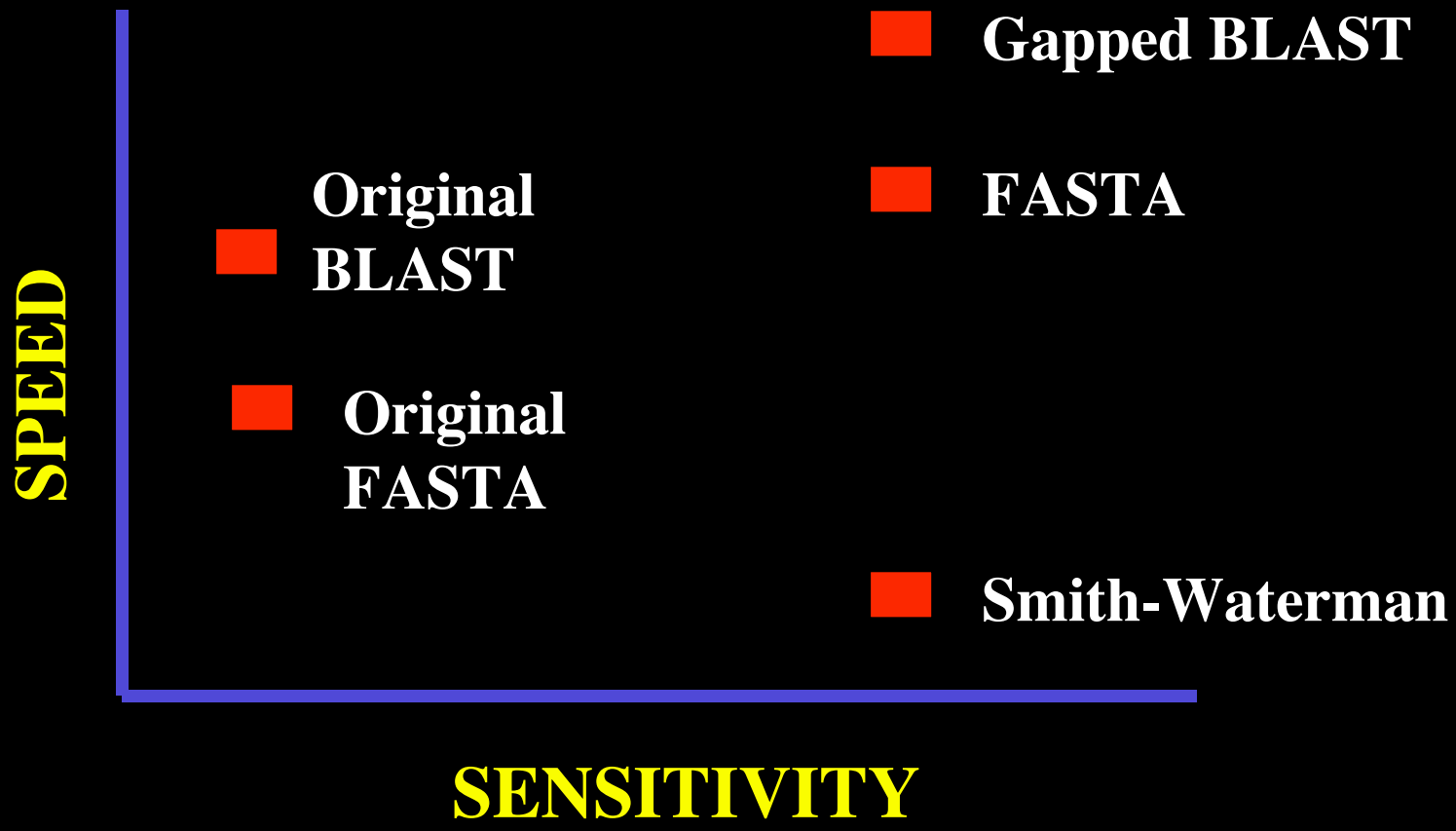
Alignment Methods

Global vs Local

Modular proteins



Database Searching Methods: Local Alignments



Possible Alignments

```
A:    T C A G A C G A G T G
B:    T C G G A G C T G
```

Possible Alignments

```
A:    T C A G A C G A G T G
B:    T C G G A G C T G

I.    T C A G A C G A G T G
      T C G G A - - G C T G
```

Possible Alignments

```
A:    T C A G A C G A G T G
B:    T C G G A G C T G

I.    T C A G A C G A G T G
      T C G G A - - G C T G

II.   T C A G A C G A G T G
      T C G G A - G C - T G
```

Possible Alignments

```
A:    T C A G A C G A G T G
B:    T C G G A G C T G

I.    T C A G A C G A G T G
      T C G G A - - G C T G

II.   T C A G A C G A G T G
      T C G G A - G C - T G

III.  T C A G A C G A G T G
      T C G G A - G - C T G
```

Topics to Cover

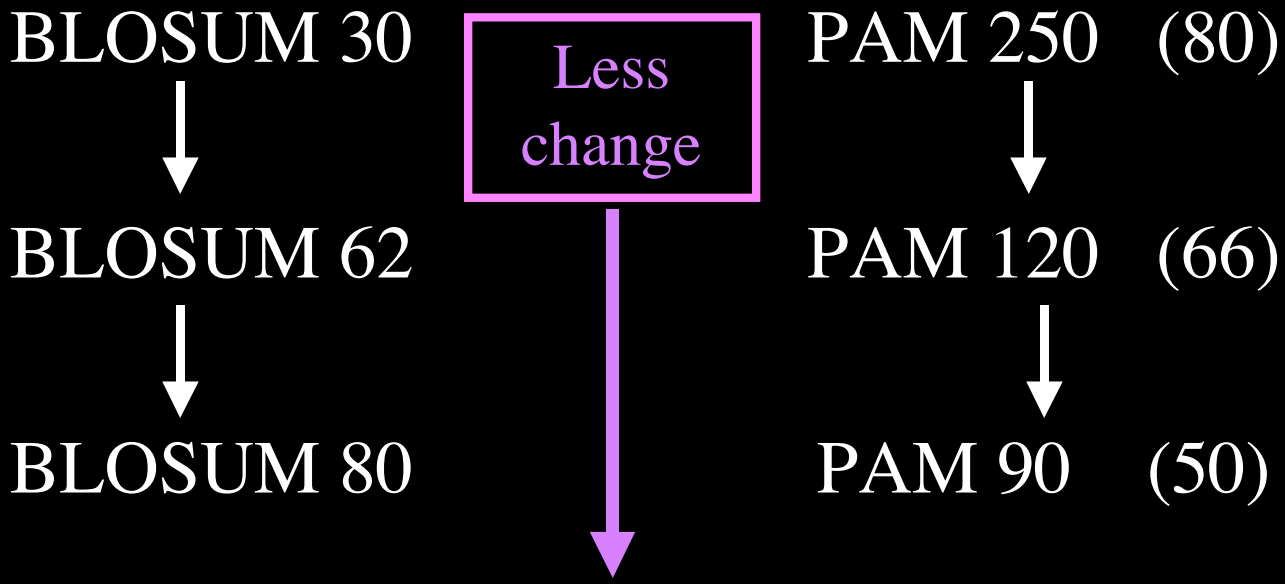
- Introduction
- Scoring alignments
  - Nucleotide vs Proteins
- Alignment methods
- Significance of alignments
- Database searching methods
- Demo

Amino Acid Substitution Matrices

PAM - point accepted mutation based on *global* alignment [evolutionary model]

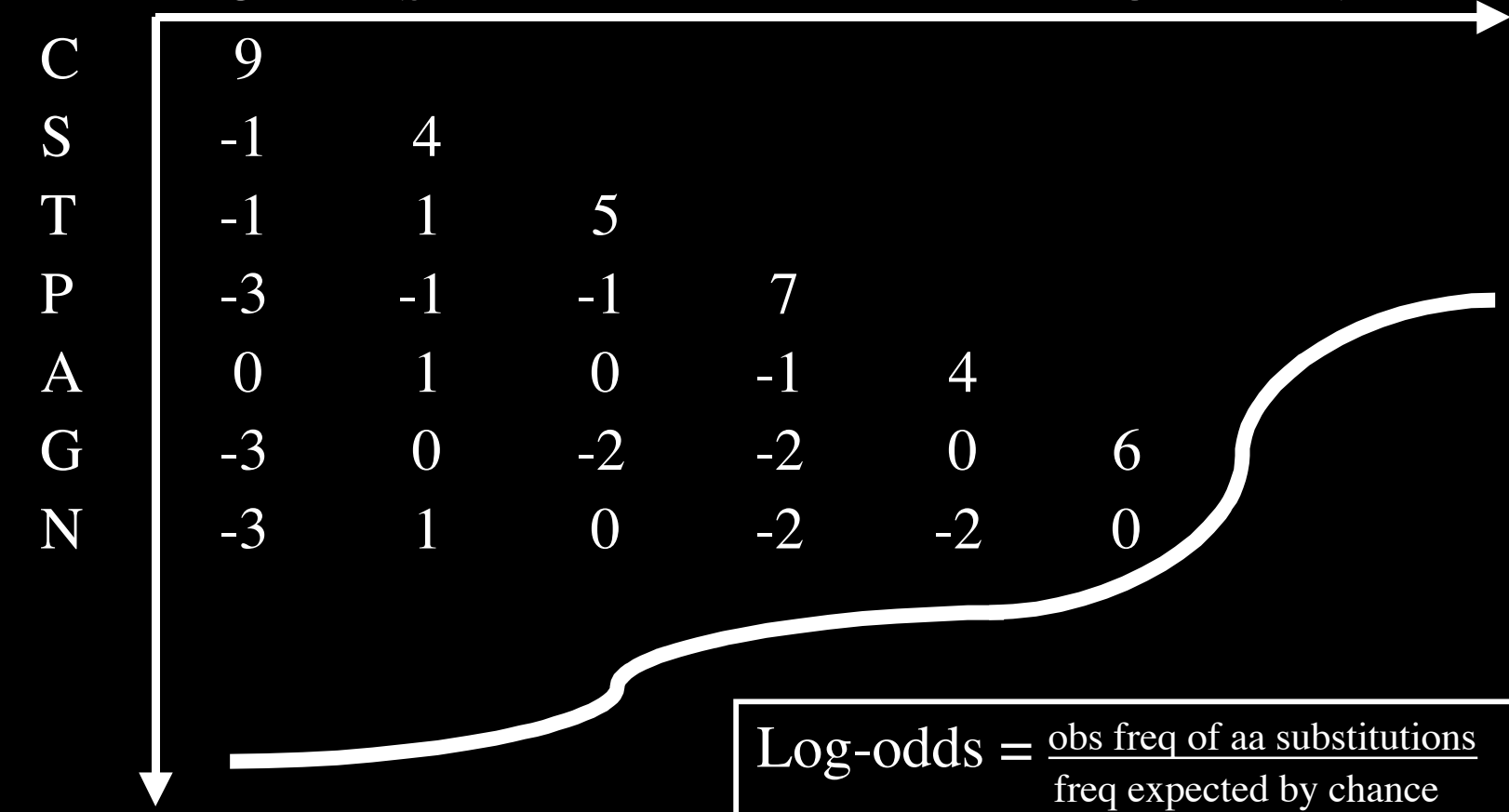
BLOSUM - block substitutions based on *local* alignments [similarity among conserved sequences]

Substitution Matrices



Part of BLOSUM 62 Matrix

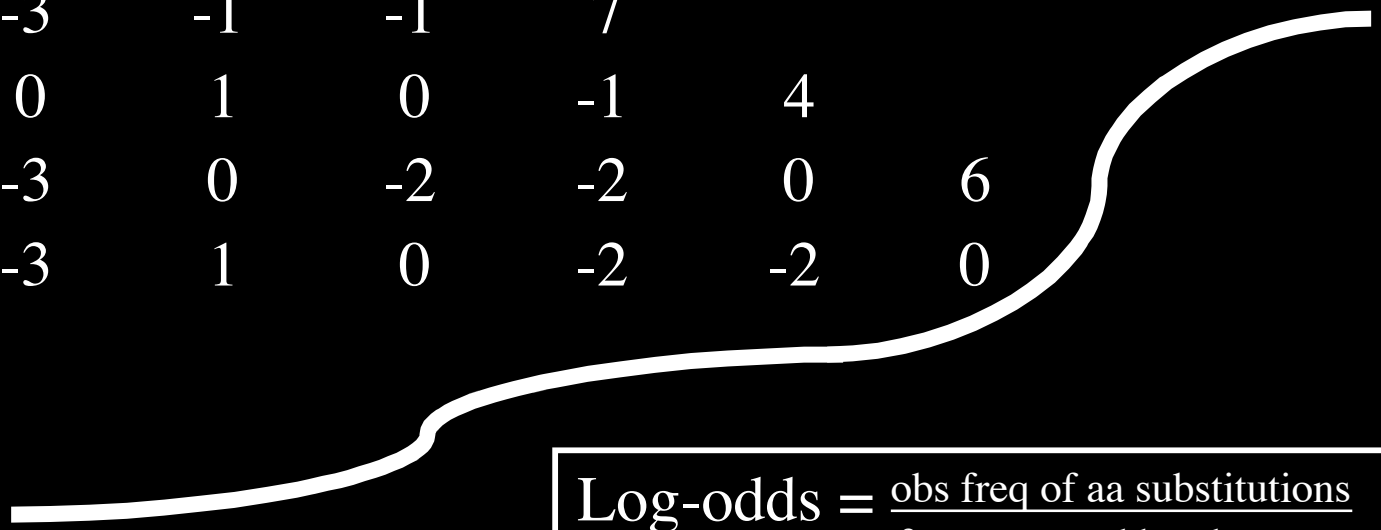
	C	S	T	P	A	G	N
C	9						
S	-1	4					
T	-1	1	5				
P	-3	-1	-1	7			
A	0	1	0	-1	4		
G	-3	0	-2	-2	0	6	
N	-3	1	0	-2	-2	0	



$$\text{Log-odds} = \frac{\text{obs freq of aa substitutions}}{\text{freq expected by chance}}$$

Part of BLOSUM 62 Matrix

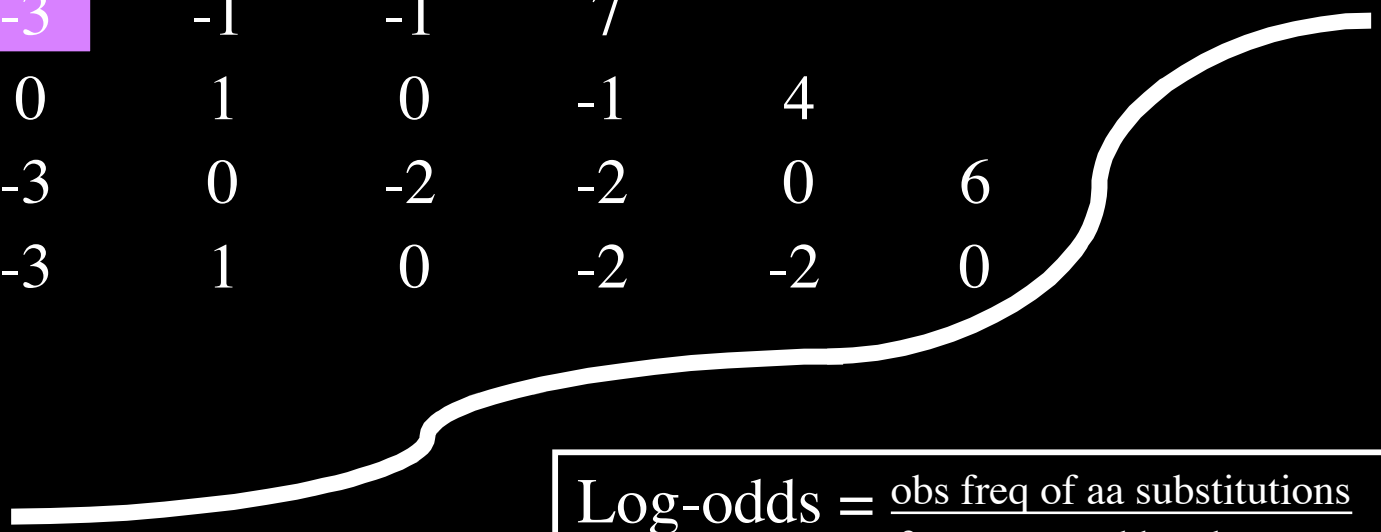
	C	S	T	P	A	G	N
C	9						
S	-1	4					
T	-1	1	5				
P	-3	-1	-1	7			
A	0	1	0	-1	4		
G	-3	0	-2	-2	0	6	
N	-3	1	0	-2	-2	0	



$$\text{Log-odds} = \frac{\text{obs freq of aa substitutions}}{\text{freq expected by chance}}$$

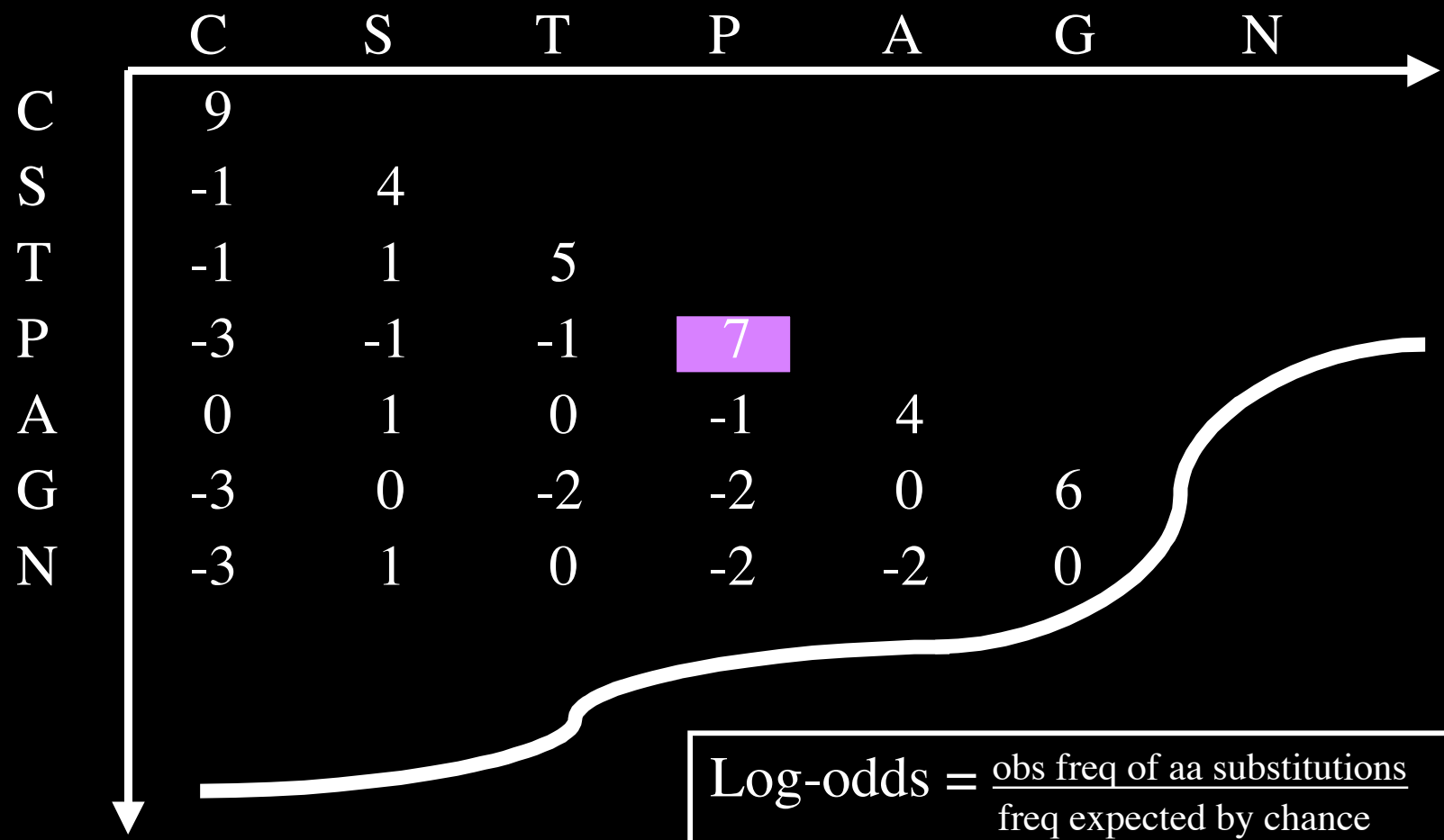
Part of BLOSUM 62 Matrix

	C	S	T	P	A	G	N
C	9						
S	-1	4					
T	-1	1	5				
P	-3	-1	-1	7			
A	0	1	0	-1	4		
G	-3	0	-2	-2	0	6	
N	-3	1	0	-2	-2	0	



$$\text{Log-odds} = \frac{\text{obs freq of aa substitutions}}{\text{freq expected by chance}}$$

Part of BLOSUM 62 Matrix

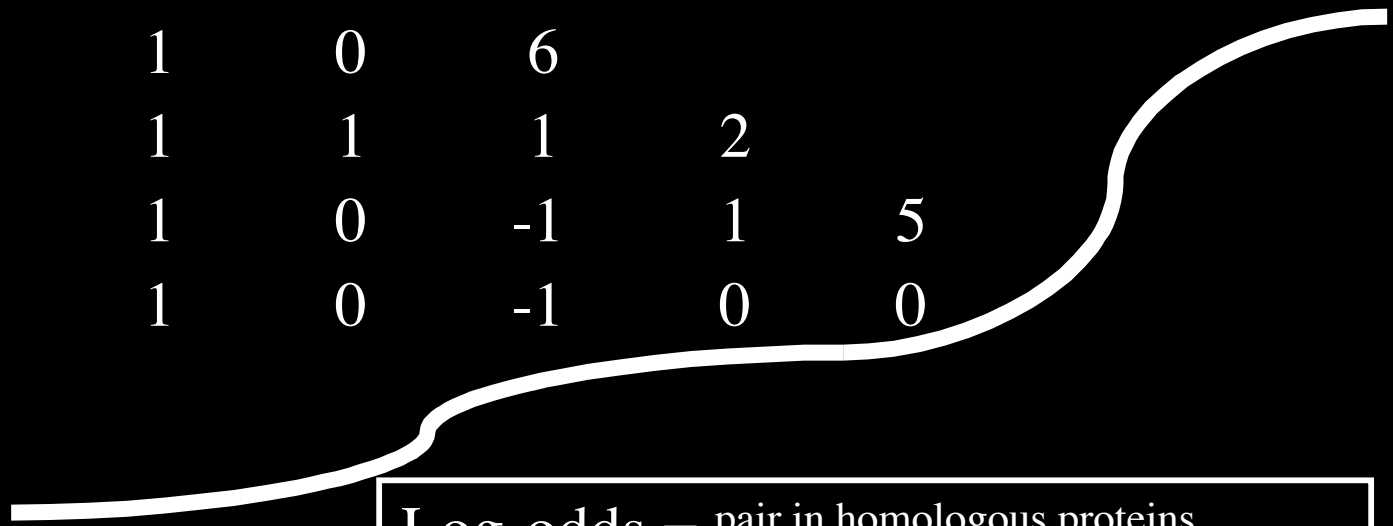


	C	S	T	P	A	G	N
C	9						
S	-1	4					
T	-1	1	5				
P	-3	-1	-1	7			
A	0	1	0	-1	4		
G	-3	0	-2	-2	0	6	
N	-3	1	0	-2	-2	0	

$$\text{Log-odds} = \frac{\text{obs freq of aa substitutions}}{\text{freq expected by chance}}$$

Part of PAM 250 Matrix

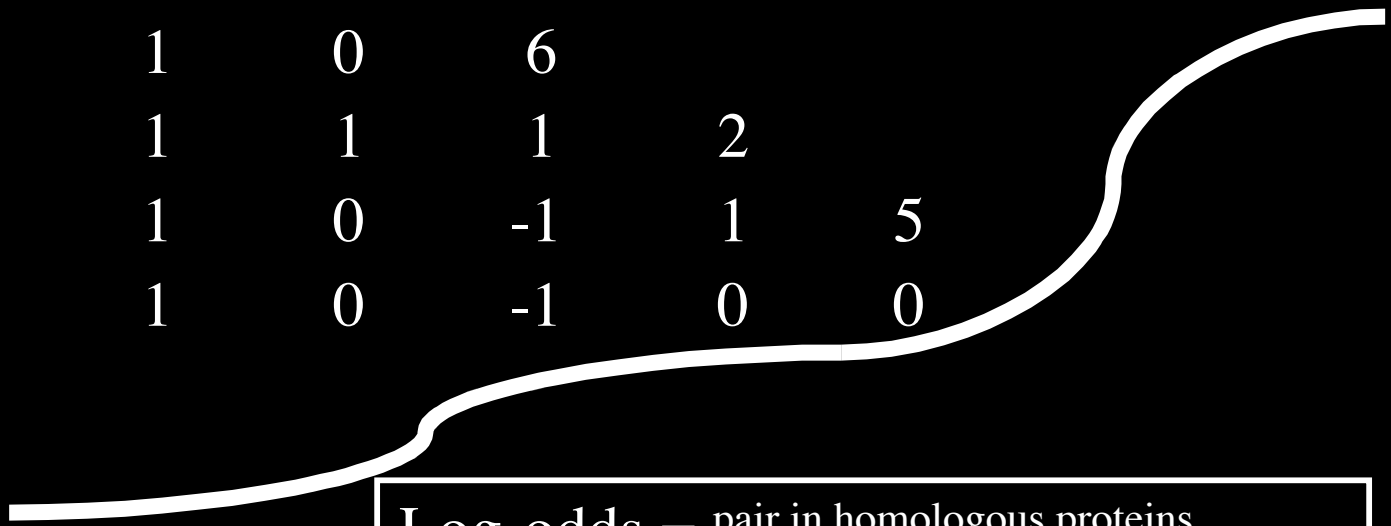
	C	S	T	P	A	G	N
C	1	2					
S	0	2					
T	-2	1	3				
P	-3	1	0	6			
A	-2	1	1	1	2		
G	-3	1	0	-1	1	5	
N	-4	1	0	-1	0	0	



$$\text{Log-odds} = \frac{\text{pair in homologous proteins}}{\text{pair in unrelated proteins by chance}}$$

Part of PAM 250 Matrix

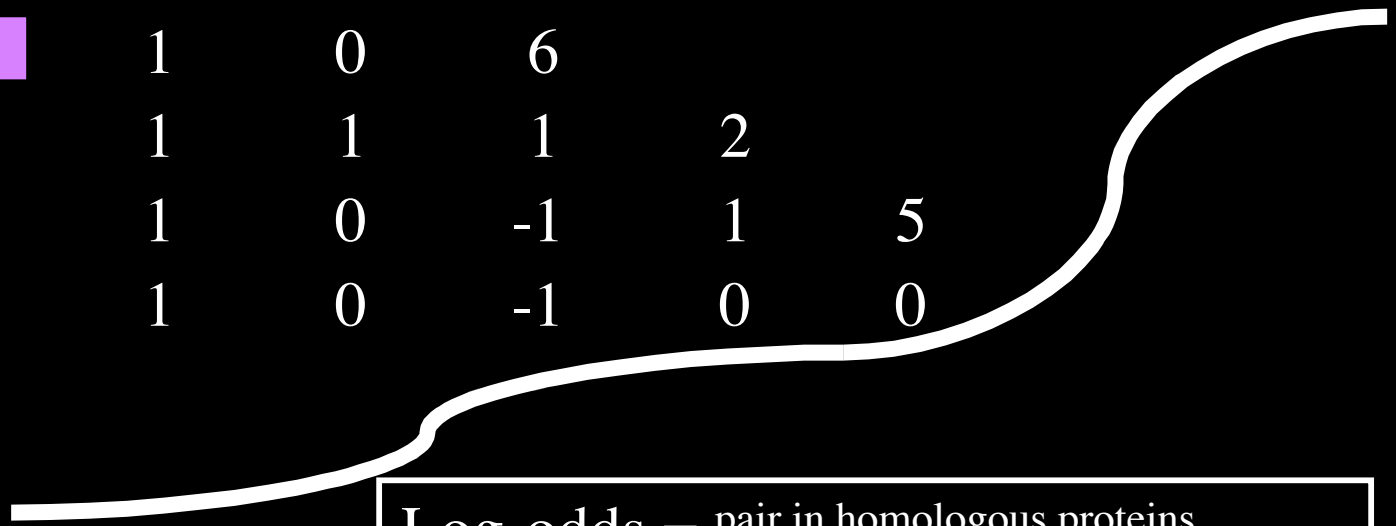
	C	S	T	P	A	G	N
C	1	2					
S	0	2					
T	-2	1	3				
P	-3	1	0	6			
A	-2	1	1	1	2		
G	-3	1	0	-1	1	5	
N	-4	1	0	-1	0	0	



$$\text{Log-odds} = \frac{\text{pair in homologous proteins}}{\text{pair in unrelated proteins by chance}}$$

Part of PAM 250 Matrix

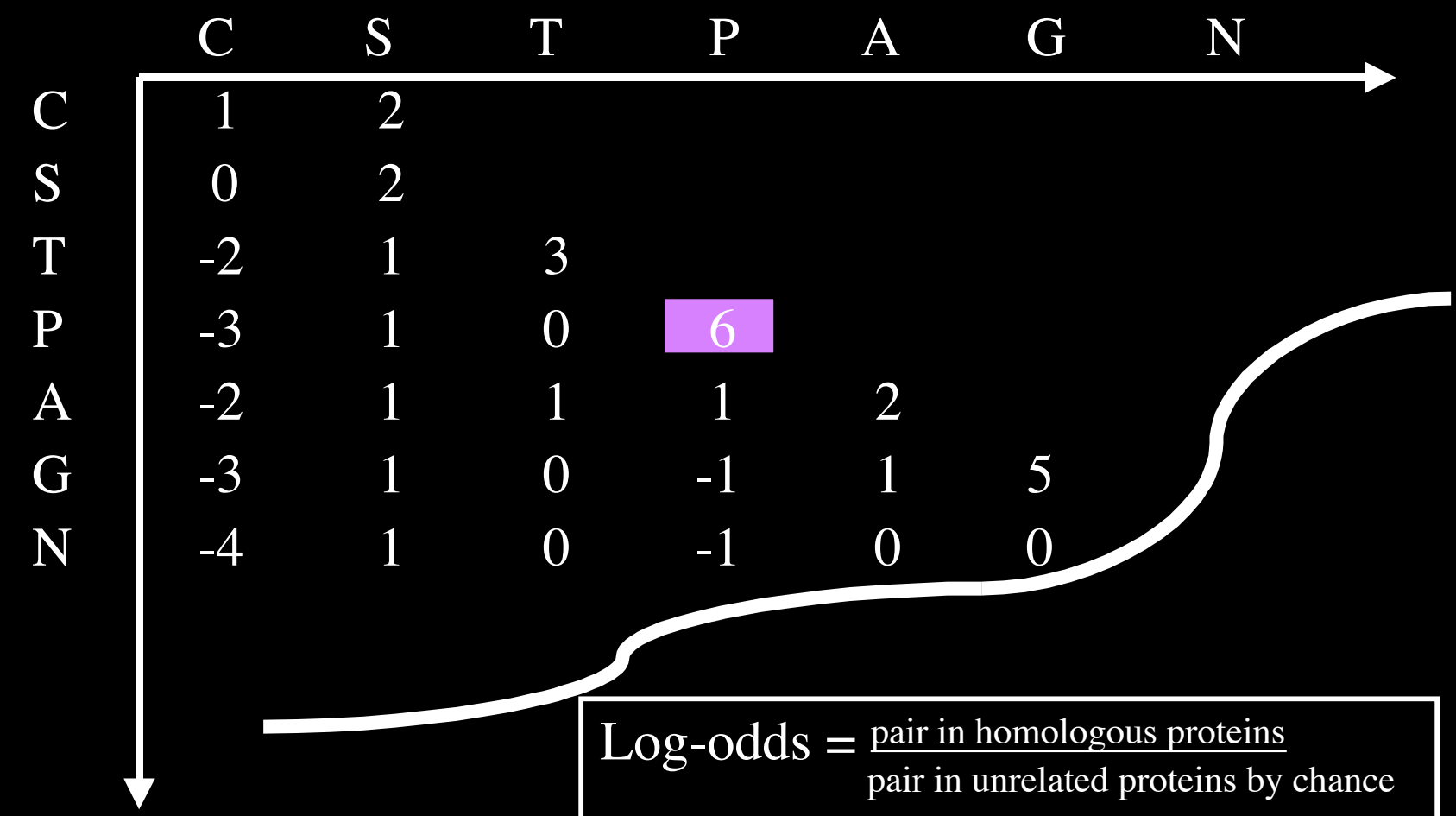
	C	S	T	P	A	G	N
C	1	2					
S	0	2					
T	-2	1	3				
P	-3	1	0	6			
A	-2	1	1	1	2		
G	-3	1	0	-1	1	5	
N	-4	1	0	-1	0	0	



$$\text{Log-odds} = \frac{\text{pair in homologous proteins}}{\text{pair in unrelated proteins by chance}}$$

Part of PAM 250 Matrix

	C	S	T	P	A	G	N
C	1	2					
S	0	2					
T	-2	1	3				
P	-3	1	0	6			
A	-2	1	1	1	2		
G	-3	1	0	-1	1	5	
N	-4	1	0	-1	0	0	



$$\text{Log-odds} = \frac{\text{pair in homologous proteins}}{\text{pair in unrelated proteins by chance}}$$

Gap Penalties

- Insertion and Deletions (indels)
- Affine gap costs - a scoring system for gaps within alignments that charges a penalty for the existence of a gap and an additional per-residue penalty proportional to the gap's length

Example of simple scoring system for nucleic acids

- Match = +1 (ex. A-A, T-T, C-C, G-G)
- Mismatch = -1 (ex. A-T, A-C, etc)
- Gap opening = - 2
- Gap extension = -1

Example of simple scoring system for nucleic acids

- Match = +1 (ex. A-A, T-T, C-C, G-G)
- Mismatch = -1 (ex. A-T, A-C, etc)
- Gap opening = - 2
- Gap extension = -1

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

Example of simple scoring system for nucleic acids

- Match = +1 (ex. A-A, T-T, C-C, G-G)
- Mismatch = -1 (ex. A-T, A-C, etc)
- Gap opening = - 2
- Gap extension = -1

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

Example of simple scoring system for nucleic acids

- Match = +1 (ex. A-A, T-T, C-C, G-G)
- Mismatch = -1 (ex. A-T, A-C, etc)
- Gap opening = - 2
- Gap extension = -1

```
T  C  A  G  A  C  G  A  G  T  G
T  C  G  G  A  -  -  G  C  T  G
-------------------------------
+1 +1 -1 +1 +1 -2 -1 -1 -1 +1 +1 = 0
```

Scoring for BLAST 2 Sequences

```
Identities = 45/101 (44%), Positives = 54/101 (52%), Gaps = 7/101 (6%)

Query: 204 YTGPFCDV----DTKASCYDGRGLSYRGLARTTLSGAPCQPWASEATYRNVTAEQ---AR 256
           Y+  FC      +  + CY G G +YRG    T SGA C PW S      V   Q   A+
Sbjct: 198 YSSEFCSTPACSEGNSDCYFGNGSAYRGTHSLTESGASCLPWNSMILIGKVYTAQNPSAQ 257

Query: 257 NWGLGGHAFCRNPDNDIRPWCFVLNRDRLSWEYCDLAQCQT 297
             GLG H +CRNPD D +PWC VL   RL+WEYCD+  C T
Sbjct: 258 ALGLGKHNYCRNPDGDAKPWCHVLKNRRLTWEYCDVPSCST 298
```

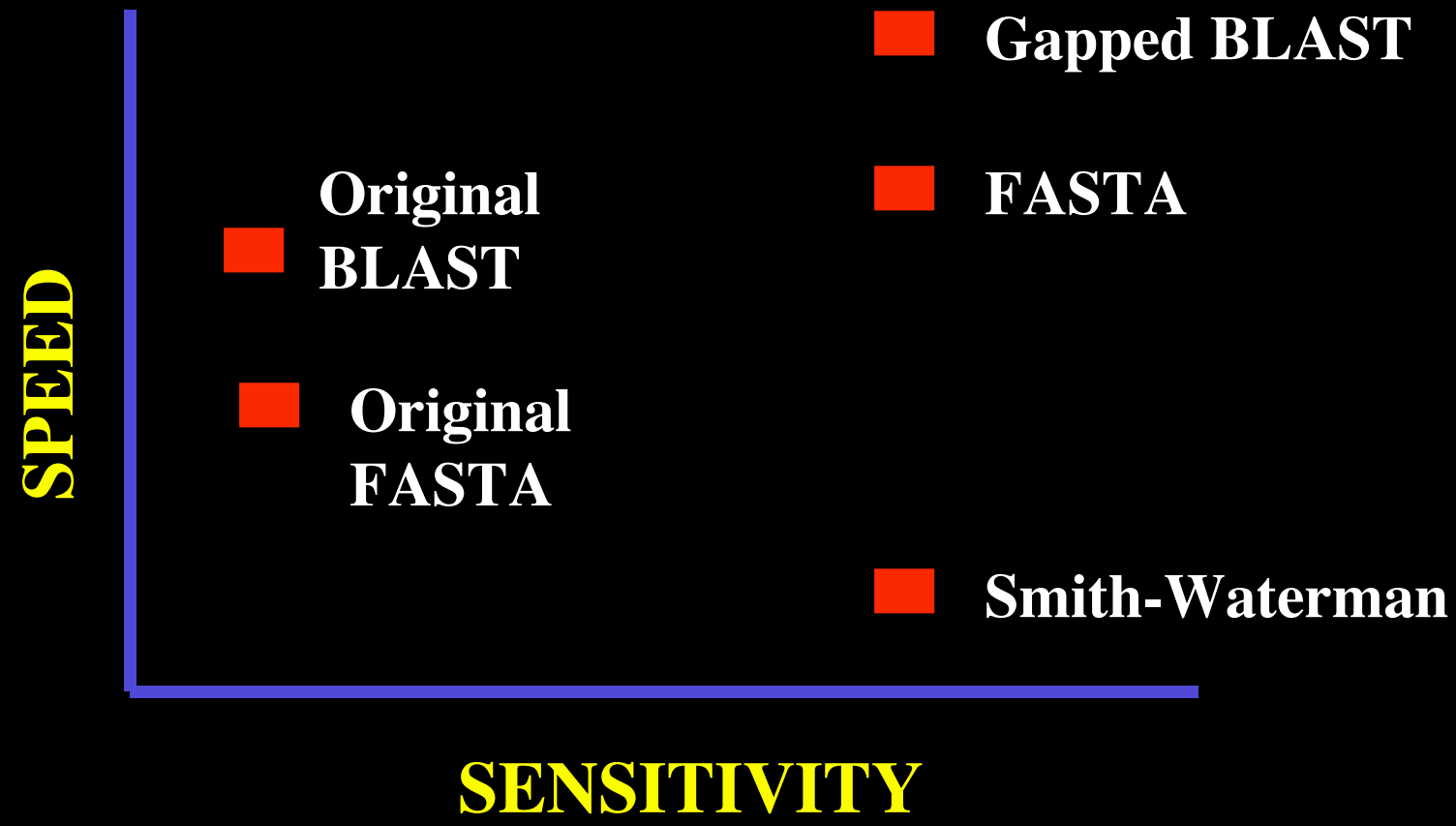
Based on BLOSUM62

```
Position  1: Y - Y =   7
Position  2: T - S =   1
Position  3: G - S =   0
Position  4: P - E =  -1
         . . .
Position  9: - - P = -11
Position 10: - - A =  -1
         . . .
              Sum     230
```

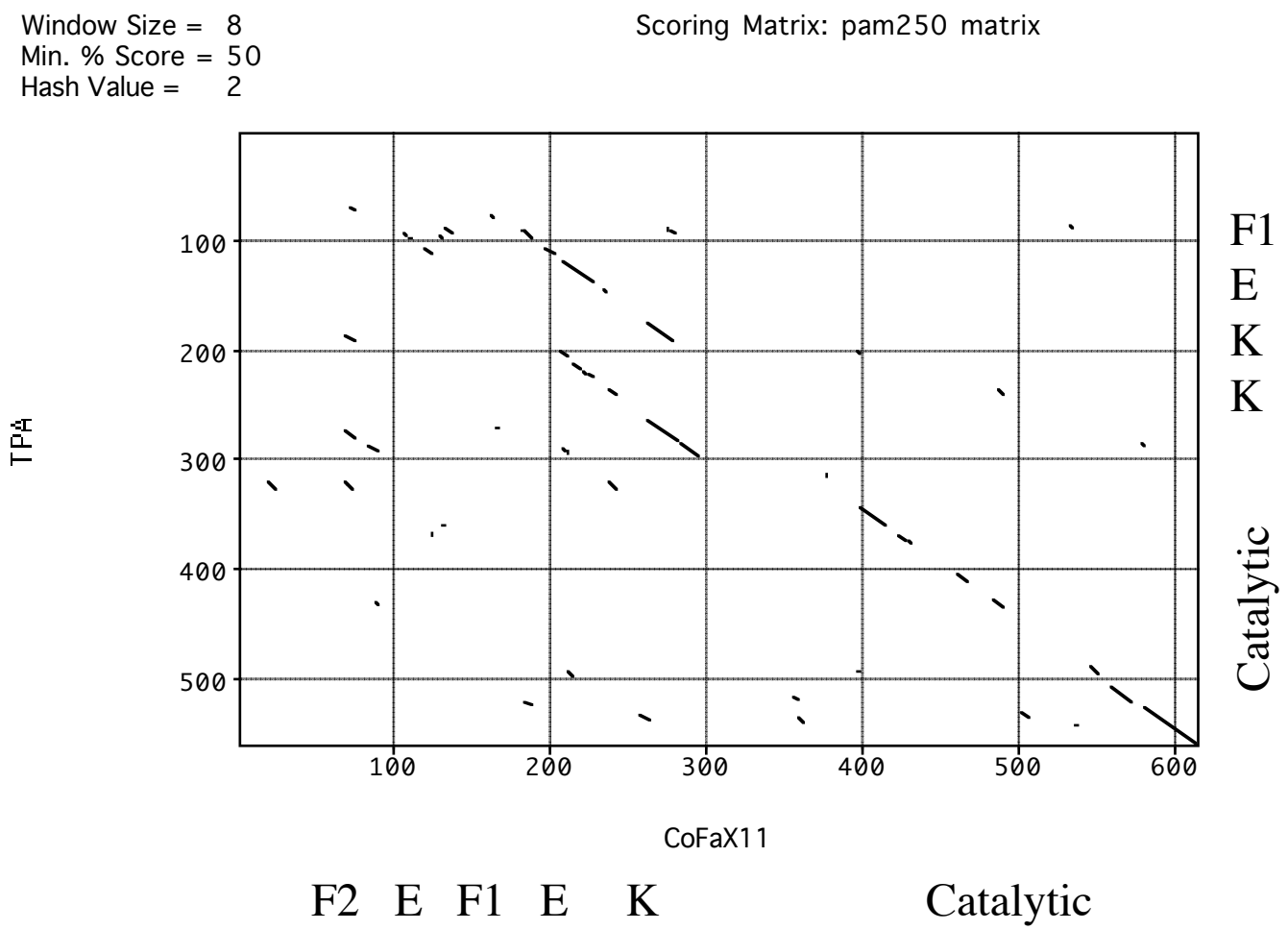
Topics to Cover

- Introduction
- Scoring alignments
- Alignment methods
  - Dot matrix analysis
  - Exhaustive methods; Dynamic programming algorithm (Smith-Waterman (Local), Needleman-Wunsch (Global)
  - Heuristic methods; Approximate methods; word or k-tuple (FASTA, BLAST)
- Significance of alignments
- Database searching methods
- Demo

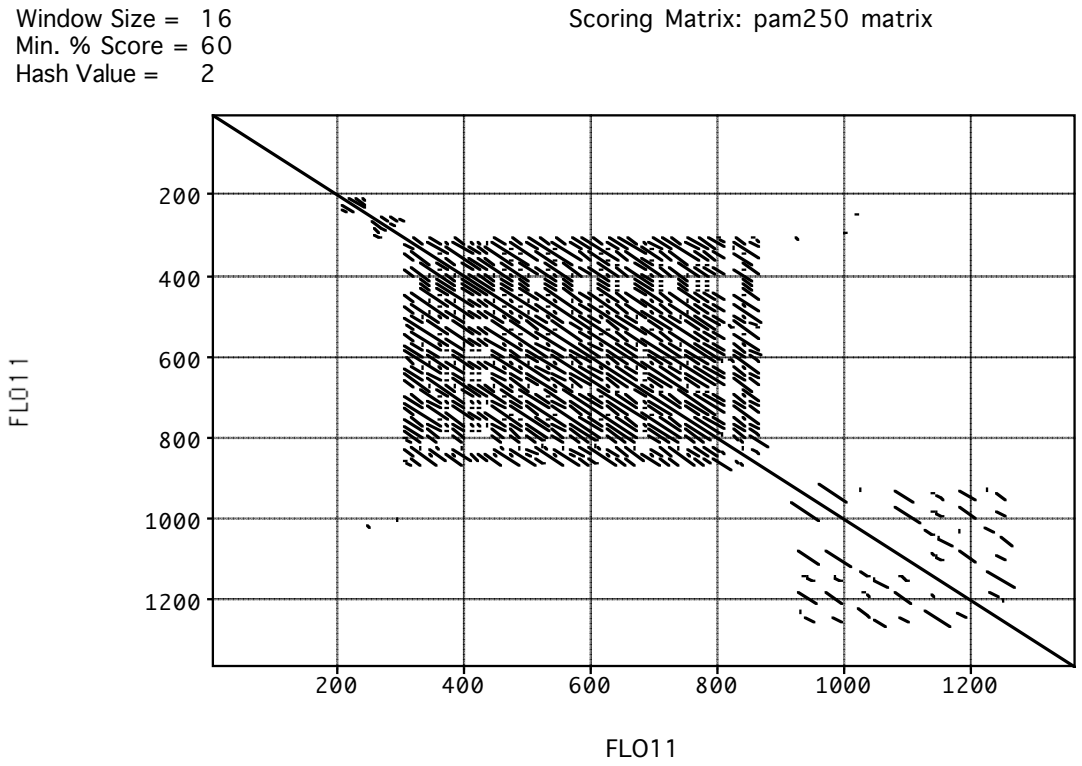
Database Searching Methods: Local Alignments



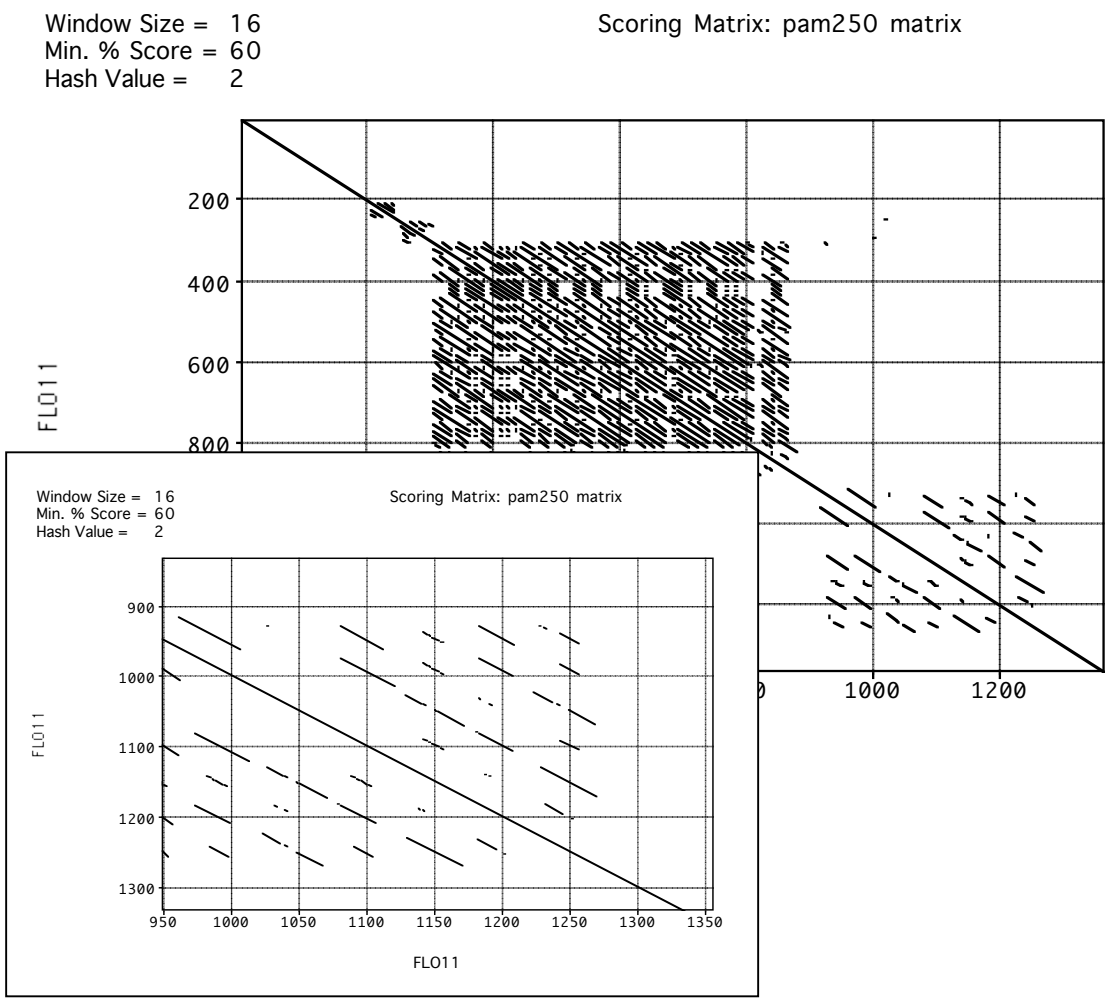
Dot Matrix Comparison



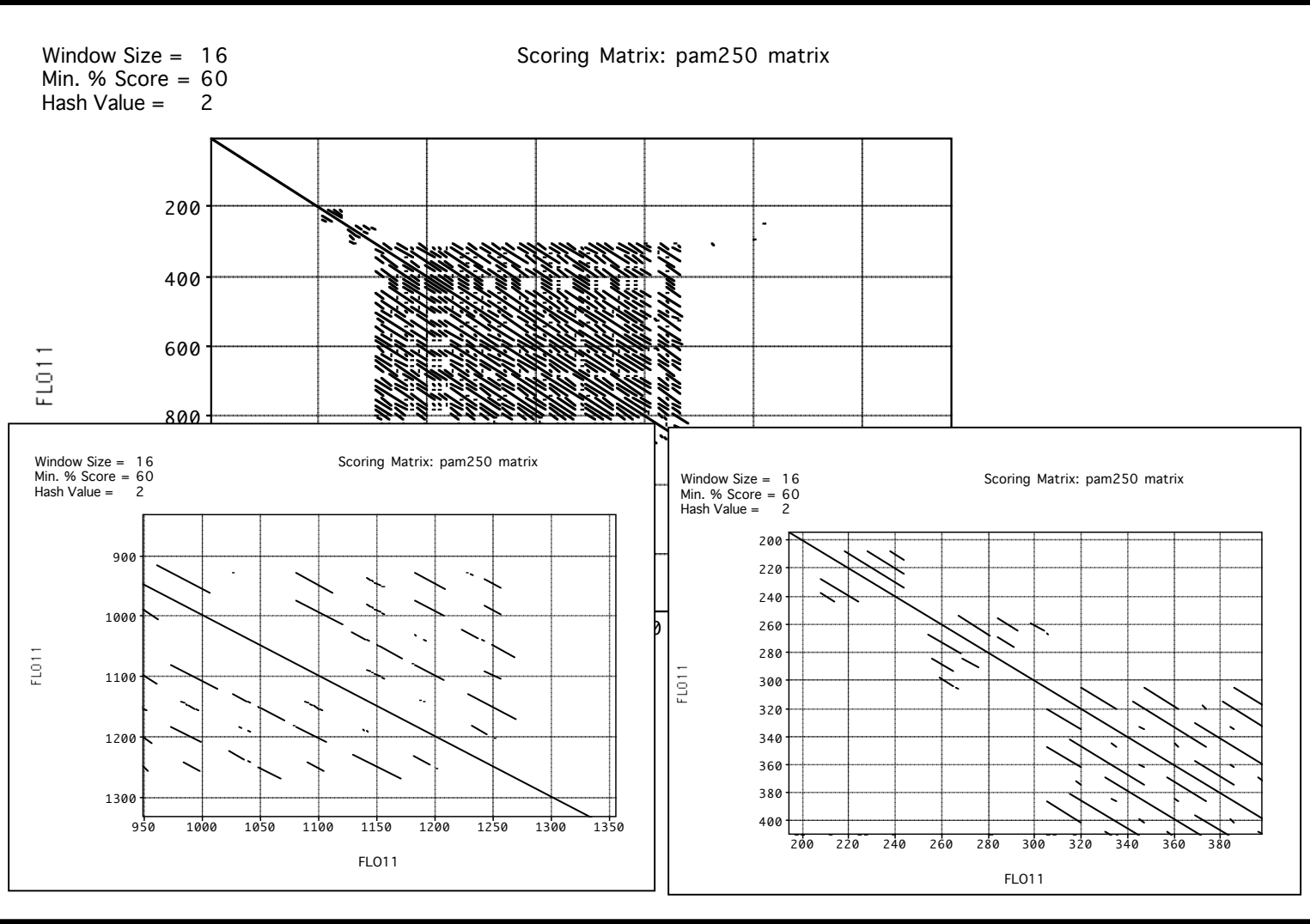
Dot Matrix Comparison



Dot Matrix Comparison



Dot Matrix Comparison



Dynamic Programming

Dynamic Programming

- Provides very best or optimal alignment

Dynamic Programming

- Provides very best or optimal alignment
- Compares every pair of characters (e.g. bases or amino acids) in the two sequences

Dynamic Programming

- Provides very best or optimal alignment
- Compares every pair of characters (e.g. bases or amino acids) in the two sequences
- Puts in gaps and mismatches

Dynamic Programming

- Provides very best or optimal alignment
- Compares every pair of characters (e.g. bases or amino acids) in the two sequences
- Puts in gaps and mismatches
- Maximum number of matches between identical or related characters

Dynamic Programming

- Provides very best or optimal alignment
- Compares every pair of characters (e.g. bases or amino acids) in the two sequences
- Puts in gaps and mismatches
- Maximum number of matches between identical or related characters
- Generates a score and statistical assessment

Dynamic Programming

Match = +2

Mismatch = -1

Gap = -3

	Gap	T	A	T	A	A	T
Gap	0	-3	-6	-9	-12	-15	-18
G	-3						
T	-6						
T	-9						
A	-12						
C	-15						
G	-18						
T	-21						
A	-24						
A	-27						

Dynamic Programming

Match = +2

Mismatch = -1

Gap = -3

	Gap	T	A	T	A	A	T
Gap	0	-3	-6	-9	-12	-15	-18
G	-3						
T	-6						
T	-9						
A	-12						
C	-15						
G	-18						
T	-21						
A	-24						
A	-27						

Dynamic Programming

Match = +2

Mismatch = -1

Gap = -3

	Gap	T	A	T	A	A	T
Gap	0	-3	-6	-9	-12	-15	-18
G	-3						
T	-6						
T	-9						
A	-12						
C	-15						
G	-18						
T	-21						
A	-24						
A	-27						

G T = -1	T- - G = -6
- T G - = -6	-1

Dynamic Programming

Match = +2

Mismatch = -1

Gap = -3

	Gap	T	A	T	A	A	T
Gap	0	-3	-6	-9	-12	-15	-18
G	-3	-1 -6 -6 **-1**					
T	-6						
T	-9						
A	-12						
C	-15						
G	-18						
T	-21						
A	-24						
A	-27						

Dynamic Programming

Match = +2

Mismatch = -1

Gap = -3

	Gap	T	A	T	A	A	T
Gap	0	-3	-6	-9	-12	-15	-18
G	-3	-1 -6 -6 **-1**					
T	-6						
T	-9						
A	-12						
C	-15						
G	-18						
T	-21						
A	-24						
A	-27						

Dynamic Programming

Match = +2

Mismatch = -1

Gap = -3

	Gap	T	A	T	A	A	T
Gap	0	-3	-6	-9	-12	-15	-18
G	-3	-1 -6 -6 **-1**	-4 -9 -4 **-4**				
T	-6						
T	-9						
A	-12						
C	-15						
G	-18						
T	-21						
A	-24						
A	-27						

Dynamic Programming

Match = +2

Mismatch = -1

Gap = -3

	Gap	T	A	T	A	A	T
Gap	0	-3	-6	-9	-12	-15	-18
G	-3	-1 -6 -6 **-1**	-4 -9 -4 **-4**				
T	-6						
T	-9						
A	-12						
C	-15						
G	-18						
T	-21						
A	-24						
A	-27						

Dynamic Programming

Match = +2

Mismatch = -1

Gap = -3

	Gap	T	A	T	A	A	T
Gap	0	-3	-6	-9	-12	-15	-18
G	-3	-1 -6 / -6 **-1**	-4 -9 / -4 **-4**				
T	-6						
T	-9						
A	-12						
C	-15		-8 -5 / -13 **-5**				
G	-18						
T	-21						
A	-24						
A	-27						

Dynamic Programming

Match = +2

Mismatch = -1

Gap = -3

	Gap	T	A	T	A	A	T
Gap	0	-3	-6	-9	-12	-15	-18
G	-3	-1 -6 -6 **-1**	-4 -9 -4 **-4**				
T	-6						
T	-9						
A	-12						
C	-15		-8 -5 -13 **-5**				
G	-18						
T	-21						
A	-24						
A	-27						

Dynamic Programming

Match = +2

Mismatch = -1

Gap = -3

	Gap	T	A	T	A	A	T
Gap	0	-3	-6	-9	-12	-15	-18
G	-3	-1 -6 / -6 **-1**	-4 -9 / -4 **-4**				
T	-6						
T	-9						
A	-12						
C	-15		-8 -5 / -13 **-5**				
G	-18						
T	-21						
A	-24						
A	-27						-6 -6 / -5 **-5**

Dynamic Programming

Match = +2

Mismatch = -1

Gap = -3

	Gap	T	A	T	A	A	T
Gap	0	-3	-6	-9	-12	-15	-18
G	-3	-1 -6 / -6 **-1**	-4 -9 / -4 **-4**				
T	-6						
T	-9						
A	-12						
C	-15		-8 -5 / -13 **-5**				
G	-18						
T	-21						
A	-24						
A	-27						-6 -6 / -5 **-5**

Dynamic Programming

	0	T	A	T	A	A	T
0	0	-3	-6	-9	-12	-15	-18
G	-3	-1 -6 -6 **-1**	-4 -9 -4 **-4**	-7	-10	-13	-16
T	-6	-1	-2	-2	-5	-8	-11
T	-9	-4	-2	0	-3	-6	-6
A	-12	-7	-2	-3	2	-1	-4
C	-15	-10	-8 -5 -13 **-5**	-3	-1	1	-2
G	-18	-13	-8	-6	-4	-2	0
T	-21	-16	-11	-6	-7	-5	0
A	-24	-19	-14	-9	-4	-5	-3
A	-27	-22	-17	-12	-7	-2	-6 -6 -5 **-5**

Dynamic Programming

	0	T	A	T	A	A	T
0	0	-3	-6	-9	-12	-15	-18
G	-3	-1 -6 -6 **-1**	-4 -9 -4 **-4**	-7	-10	-13	-16
T	-6	-1	-2	-2	-5	-8	-11
T	-9	-4	-2	0	-3	-6	-6
A	-12	-7	-2	-3	2	-1	-4
C	-15	-10	-8 -5 -13 **-5**	-3	-1	1	-2
G	-18	-13	-8	-6	-4	-2	0
T	-21	-16	-11	-6	-7	-5	0
A	-24	-19	-14	-9	-4	-5	-3
A	-27	-22	-17	-12	-7	-2	-6 -6 -5 **-5**

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

Global vs Local Alignment

Examples of aligning the same two proteins both globally and locally.

See Chapter 3, example 1 on the online site for *Bioinformatics* by Mount.

Original "Ungapped" BLAST Algorithm

- To improve speed, use a word based hashing scheme to index database
- Limit search for similarities to only the region near matching words
- Use Threshold parameter to rate neighbor words
- Extend match left and right to search for high scoring alignments

Original BLAST Algorithm

Query word (W=3)

Query: GSVEDTTGSQSLAALLNKCKT*PQG*QRLVNQWIKQPLM

Neighborhood words

PQG	18	PHG	13
PEG	15	PMG	13
PNG	13	PTG	12
PDG	13	Etc.	

Neighborhood Score threshold (T=13)

```
Query: 325    SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA
              +LA++L+   TP G R++ +W+  P+ D   + ER   I A
Sbjct: 290    TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA
```

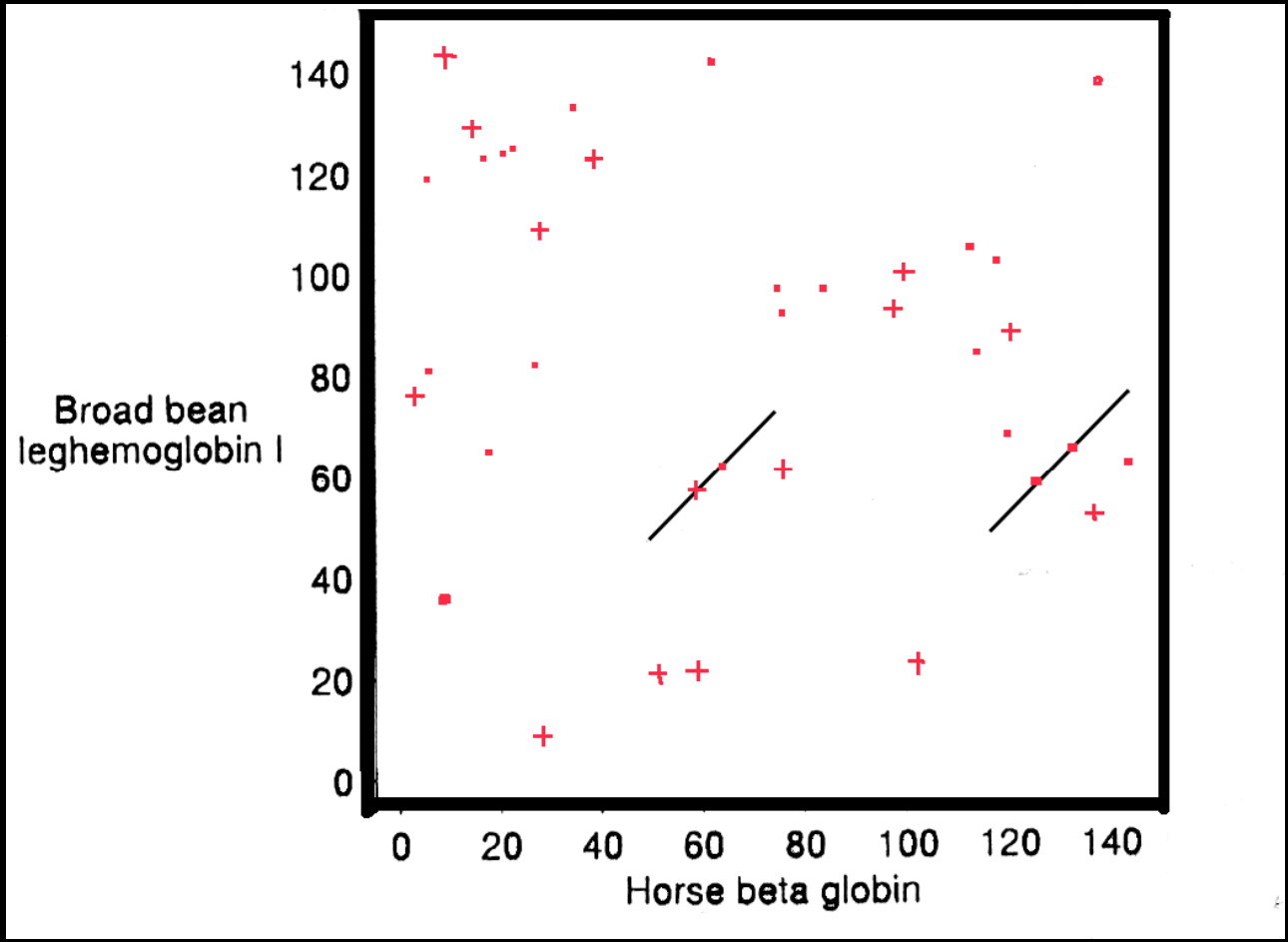
Recent BLAST Refinements

- "two-hit" method for extending word pairs
- Gapped alignments
- Iterate with position-specific matrix (PSI-BLAST)
- Pattern-hit initiated BLAST (PHI-BLAST)

Gapped BLAST

15(+) > 13
22(•) > 11

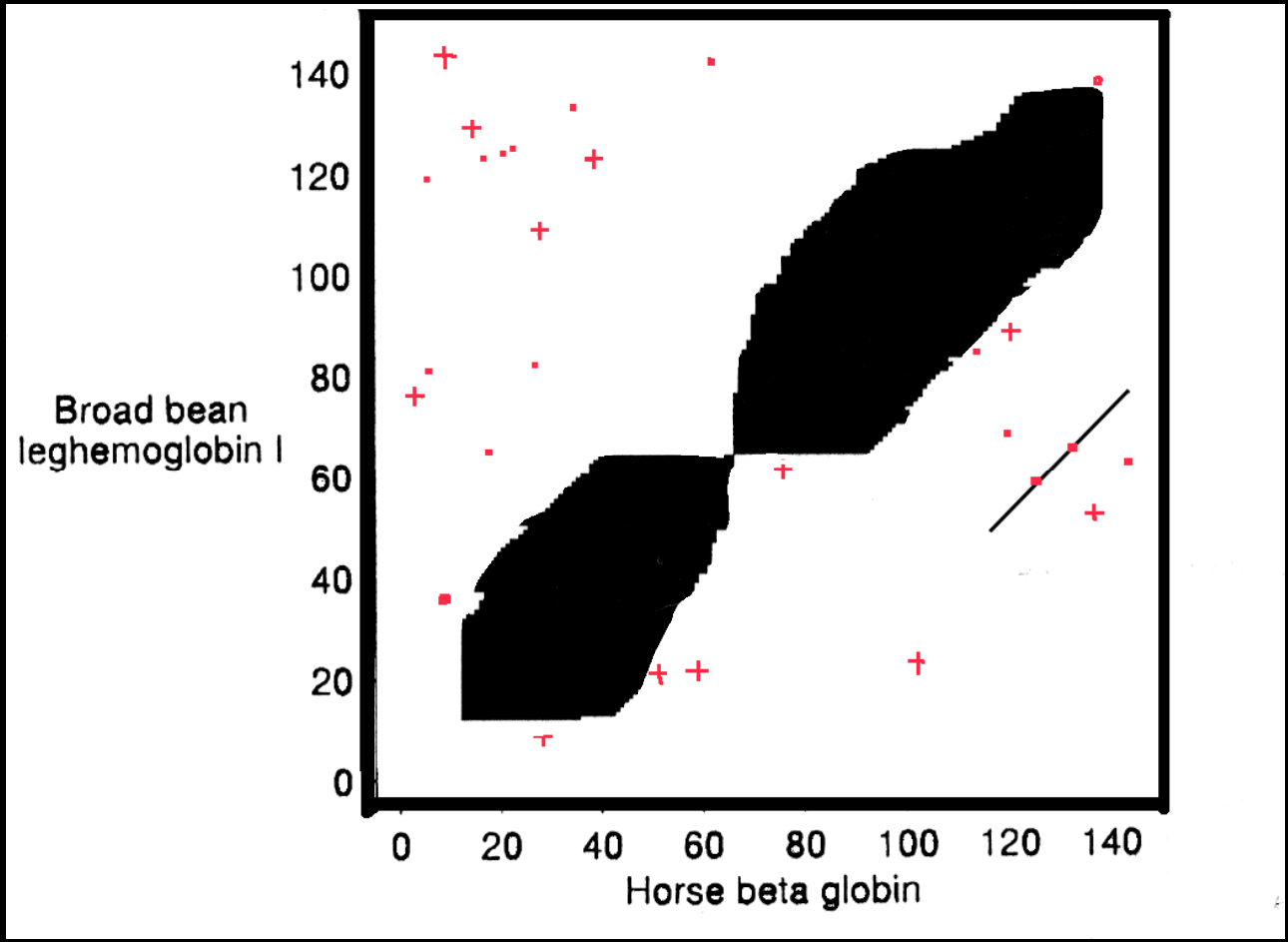
From: *Nucleic Acids Research, 1997, Vol. 25, No. 17* ***3389–3402***



Gapped BLAST

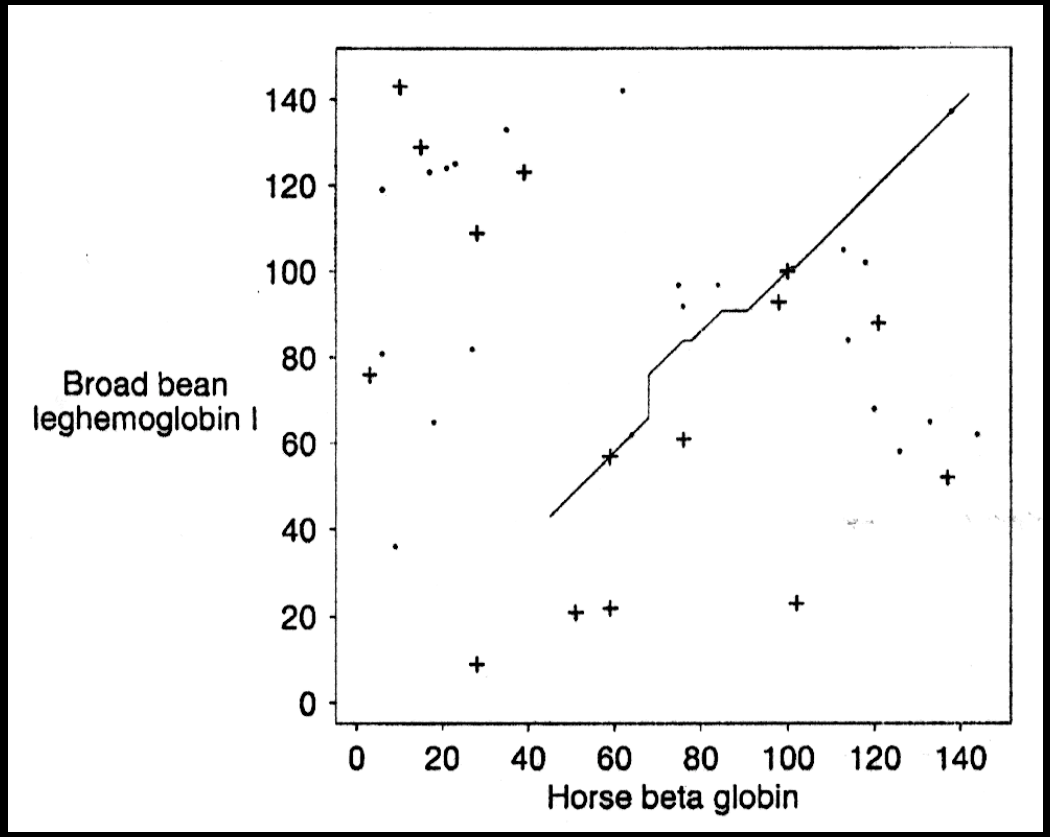
15(+) > 13
22(•) > 11

From: *Nucleic Acids Research, 1997, Vol. 25, No. 17* ***3389–3402***



Gapped BLAST

From: *Nucleic Acids Research, 1997, Vol. 25, No. 17* ***3389–3402***



Programs to Compare two sequences

Macintosh

- MacVector - Pustell Protein Matrix (DotPlot)

Web

- BLAST 2 Sequences
- RepeatFinder
- lalign

GCG/Unix

- BestFit - Smith-Waterman (randomize)
- Gap - Needleman -Wunsch (randomize)
- Dotter (dot plot)