Pairwise Sequence Alignment (III)

Jianlin Cheng, PhD

School of Electrical Engineering and Computer Science
University of Central Florida

2006

Scoring Matrix

- How to accurately measure the similarity between amino acids (or nucleotides) is one key issue of sequence alignment.
- For nucleotides, a simple identical/not identical scheme is mostly ok.
- Due to various properties of amino acids, it is hard and also critical to measure the similarity between amino acids.

Evolutionary Substitution Approach

Basic idea:

During the evolutionary process, the substitution of similar (or dissimilar) amino acids is more (or less) likely to be kept / selected within protein families than random substitutions (M. Dayhoff).

So, the frequency / probability of which one residue substitutes another one is an indicator of their similarity.

Comment:

Most accurate and successful way to derive amino acid similarity. The idea is fundamental and profound. (**learn from evolution and nature**)

PAM Scoring Matrices

(M. Dayhoff)

- Select a number of protein families. Each family has a number of similar protein sequences that evolve from the same ancestor.
- Align sequences in each family and count the frequency of amino acid substitution of each column. The frequency is used to compute the empirical substitution probability of which residue i substitutes residue j (P_{ij}).
- Similarity score is ratio of observed substitution probability over the random substitution probability. $Sim(i,j) = \log(P_{ij} / (P_i * P_j))$. P_i is the observed probability of residue i and P_j is the observed probability of residue j. (logarithm of likelihood ratio / odds)
- PAM: Point Accepted Mutation

A Simplified Example

```
ACGTCGAGT
ACCACGTGT
CACACTACT
ACCGCATGA
ACCCTATCT
TCCGTAACA
ACCATAAGT
AGCATAAGT
ACTATAAGT
ACGATAAGT
```

Chars	Prob.
A	6 / 10
C	1 / 10
G	2 / 10
T	1 / 10

Substitution Frequency Table

	A	C	G	T
A	30	6	12	6
C	6	0	2	1
G	12	2	1	2
T	6	1	2	0

Total number of substitutions: 90

	A	C	G	T
A	.33	.07	.14	.07
C	.07	0	.02	.01
G	.14	.02	.01	.02
T	.07	.01	.02	0

P(A<->A) = 0.33

P(A<->C) = 0.07+0.07=0.14

A Simplified Example

```
ACGTCGAGT
ACCACGTGT
CACACTACT
ACCGCATGA
ACCCTATCT
TCCGTAACA
ACCATAAGT
AGCATAAGT
ACTATAAGT
ACGATAAGT
```

Chars	Prob.
A	6 / 10
C	1 / 10
G	2 / 10
T	1 / 10

Substitution Frequency Table

	A	C	G	T
A	30	6	12	6
C	6	0	2	1
G	12	2	1	2
T	6	1	2	0

Total number of substitutions: 90

	A	C	G	T
A	.33	.07	.14	.07
C	.07	0	.02	.01
G	.14	.02	.01	.02
T	.07	.01	.02	0

P(A<->A) = 0.33

P(A<->C) = 0.07+0.07=0.14

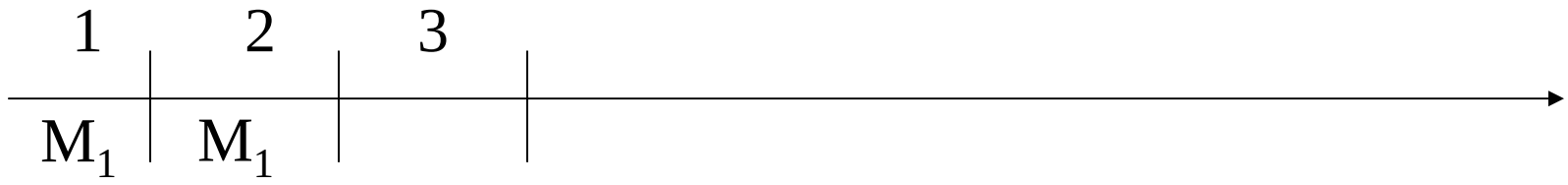
Sim(A,C) = log(0.14/(0.6*0.1))

Comments

- Generate a 20 by 20 symmetric similarity matrix
- Score > 0: positive match
- Score < 0: negative match
- Real PAM matrix is generated from very similar sequences (only 1% accepted mutations)
- This matrix is derived from families of sequences evolved during a period of time (one molecular clock (1 million years)). For more divergent sequences evolving through longer period of time, it is not appropriate. Fortunately, we can derive similarity matrix using the same idea, starting from this substitution similarity matrix.

PAM Matrices

Time: (unit: million years)



Each unit of time clock obey substitution matrix: M_1.

What is the substitution matrix of 2 time clocks: $M_1 * M_1$

$P_2(i->j) = Sum_k(P_1(i->k)P_1(k->j)$ (row * column)

…..

What is substitution matrix of n time clocks: M_1^n

PAM250

- Compute M^n (for instance n=1,2,…,250…)
- Convert M^n to similarity matrix using $\log(P_{ij} / P_i * P_j)$
- Most widely used matrix is PAM250.
- For more similar sequences, select lower n. For more divergent sequences, select higher n.

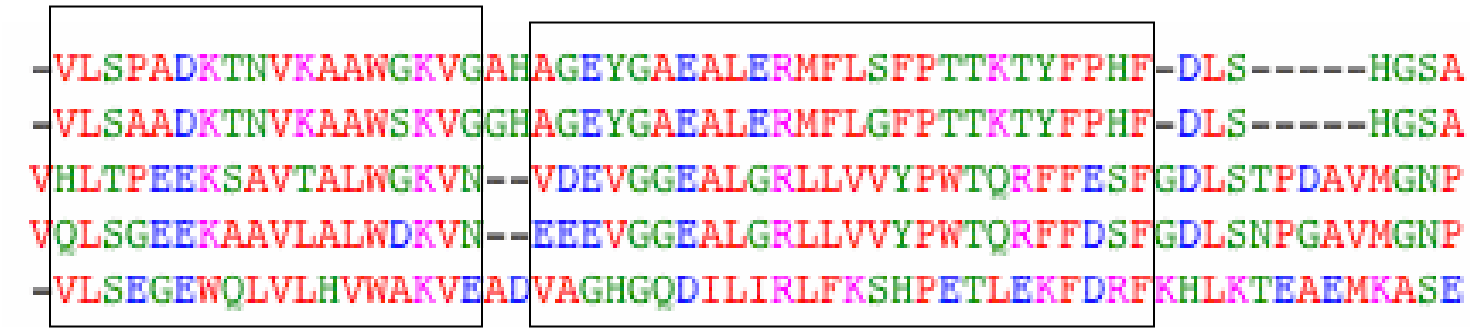
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	12																			
S	0	2																		
T	-2	1	3																	
P	-3	1	0	6																
A	-2	1	1	1	2															
G	-3	1	0	-1	1	5														
N	-4	1	0	-1	0	0	2													
D	-5	0	0	-1	0	1	2	4												
E	-5	0	0	-1	0	0	1	3	4											
Q	-5	-1	-1	0	0	-1	1	2	2	4										
H	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F	-4	-3	-3	-5	-4	-5	-3	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17

PAM250 Matrix (log odds multiplied by 10)

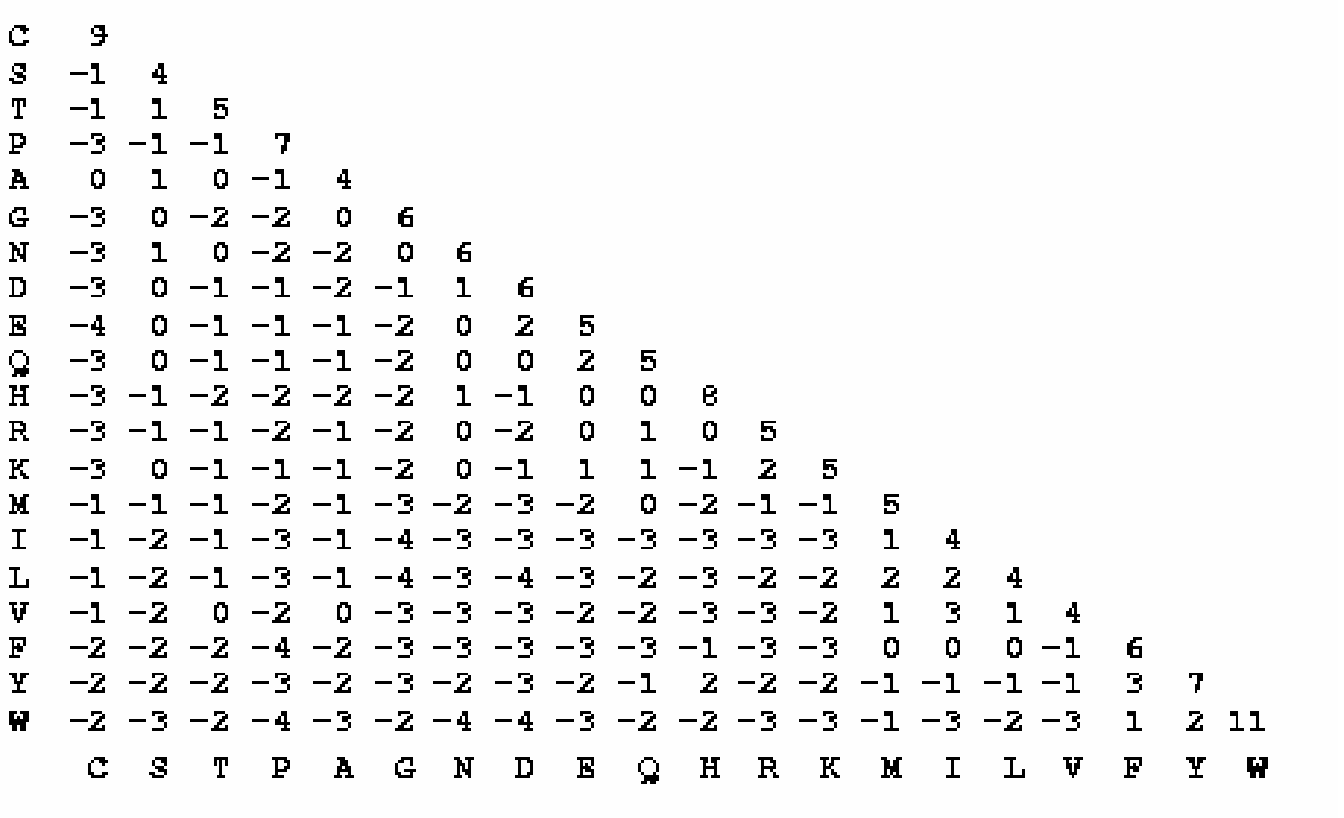
BLOSUM Matrices

(Henikoff and Henikoff)

- PAM matrices don’t work well for aligning evolutionarily divergent sequences.
- Main differences: PAM based on observed mutations throughout global alignment. BLOSUM based on highly conserved local regions /blocks without gaps. PAM uses mutation matrix multiplication; BLOSUM uses percentage of identity.
- BLOSUMn is a matrix calculated from proteins share at most n% identity. BLOSUM62 is the most widely used matrix (BLAST, PSI-BLAST, CLUSTALW)
- For more divergent sequences, choose smaller n.
- BLOSUM: BLOcks SUbstitution Matrix



Block 1 Block2



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

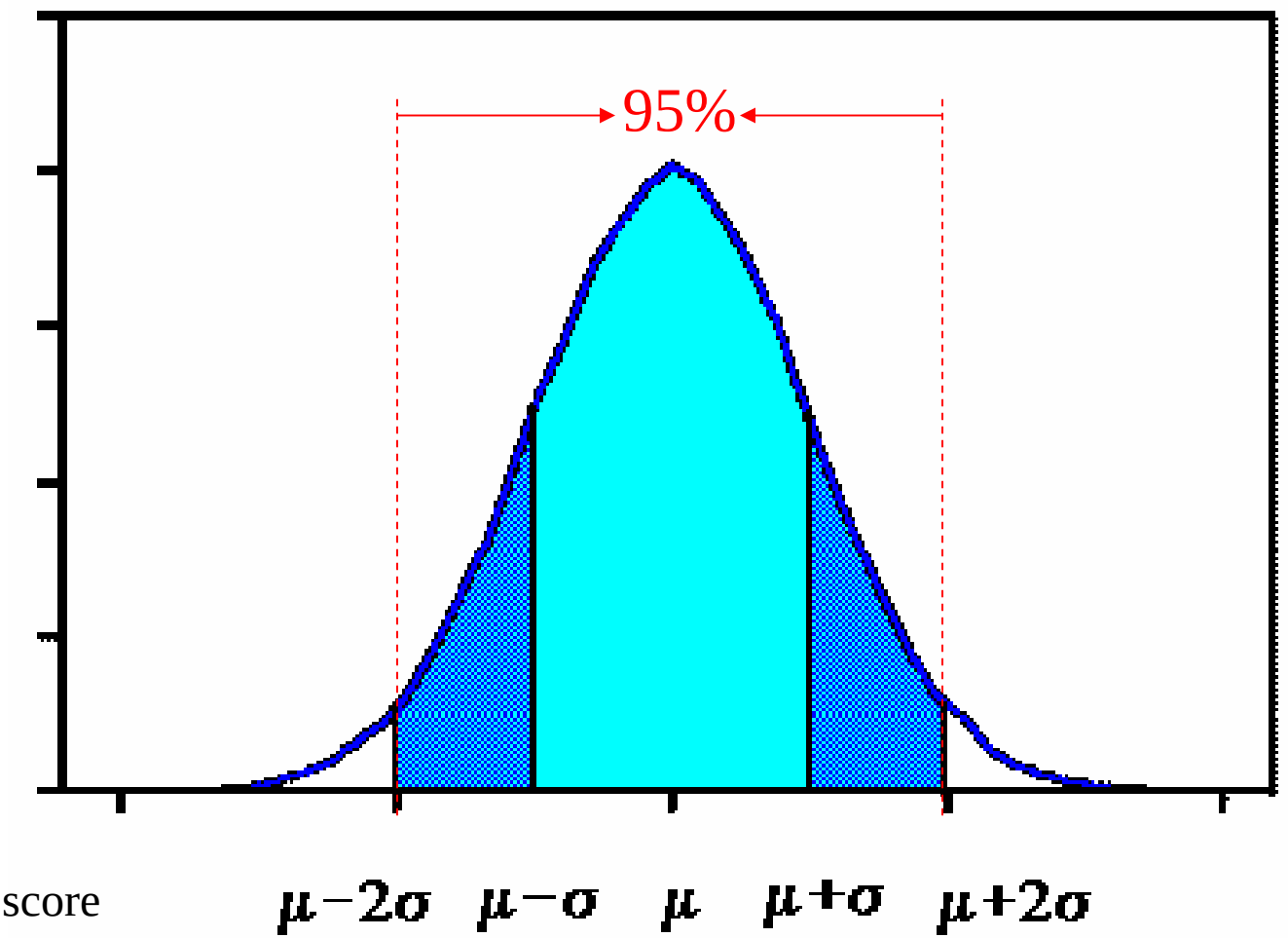
BLOSUM62 Matrix

Significance of Sequence Alignment

- Why do we need significant test?
- Mathematical view: unusual versus “by chance”
- Biological view: evolutionary related or not?
- How can we do it? (statistical methods)

Randomization Approach

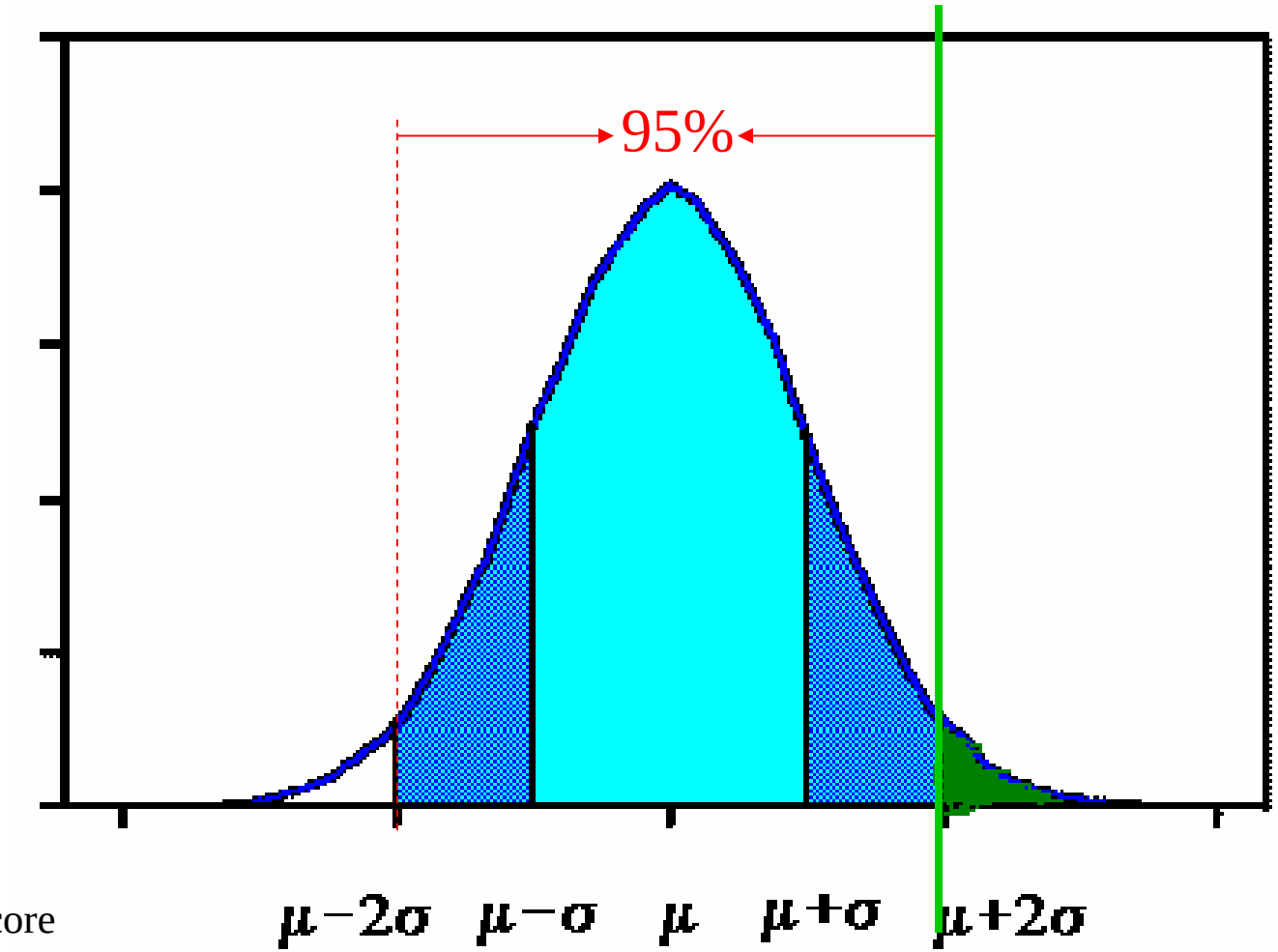
- Randomization is a fundamental idea due to Fisher.
- Randomly exchange chars (permute) with sequence P and Q to generate new sequences (P’ and Q’). Align new sequences and record alignment scores.
- Assuming these scores obey normal distribution, compute mean (u) and standard derivation (σ) of alignment scores



Normal distribution of alignment scores of two sequences

$P(u-\sigma <= S <= u+\sigma) = 0.66$

$P(u-2\ \sigma <= S <= u+2\ \sigma) = 0.95$



Normal distribution of alignment scores of two sequences

•If S = u+2 σ, the probability of observing the alignment score equal to or more extreme than this by chance is 2.5%, e.g., P(S>=u+2 σ) = 2.5%.
Thus we are 97.5% confident that the alignment score is significant (not by chance).
•For any score x, we can compute P(S >= x), which is called p-value.

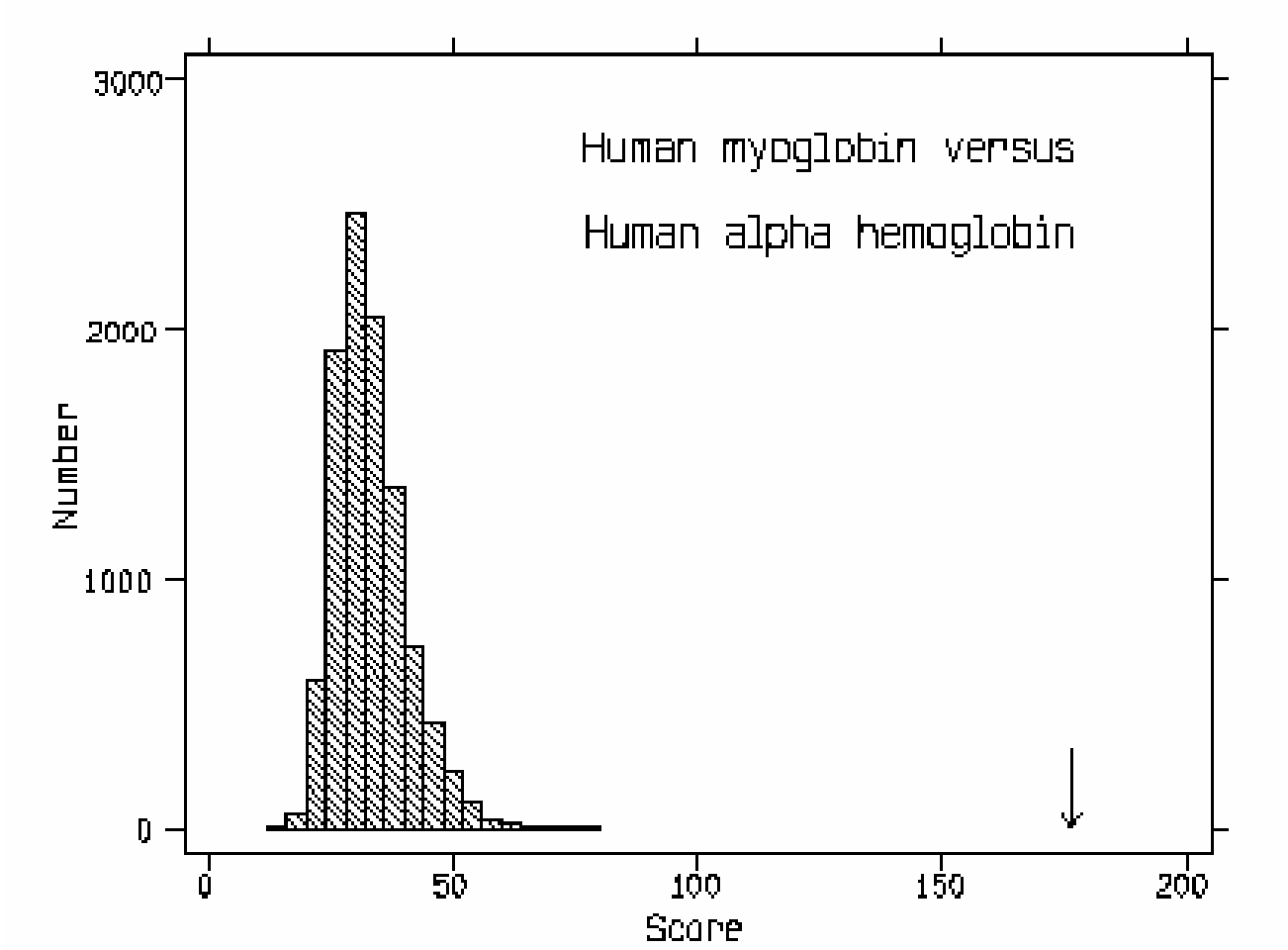


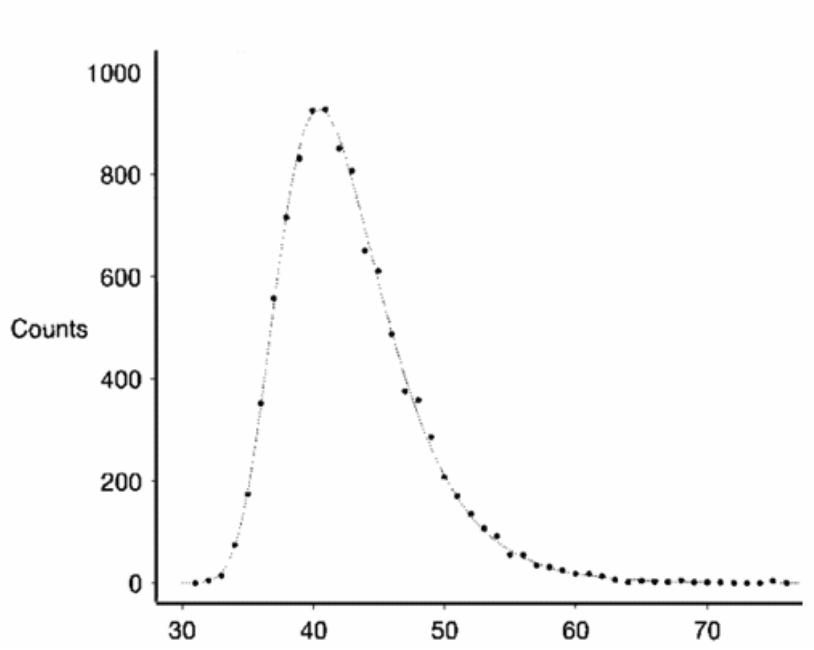
Figure: Histogram of alignment scores

Model-Based Approach
(Karlin and Altschul)

http://www.people.virginia.edu/~wrp/cshl02/Altschul/Altschul-3.html

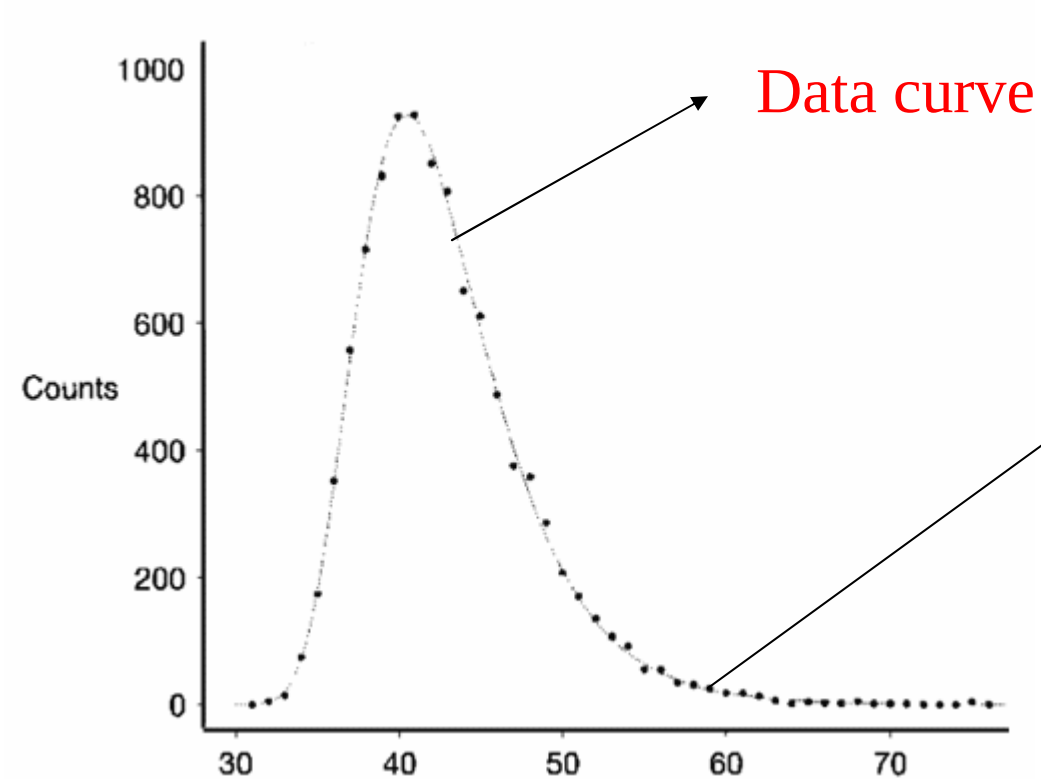
- Extreme Value Distribution

$$P(S \geq x) = 1 - \exp(-Kmn\, e^{-\lambda x})$$



K and lamda are statistical parameters depending on substitution matrix. For BLOSUM62, lamda=0.252, K=0.35

How good is the model?



Comments

- Extreme value distribution is developed from local alignment (random walk), but often is also used with global alignment.
- $P(S \geq x)$ is called **p-value**. It is the probability that random sequences has alignment score equal to or bigger than x. Smaller -> more significant.

Ten Topics

- 1. Introduction to Molecular Biology and Bioinformatics
- 2. Pairwise Sequence Alignment Using Dynamic Programming
- 3. Practical Sequence/Profile Alignment Using Fast Heuristic Methods (BLAST and PSI-BLAST)
- 4. Multiple Sequence Alignment
- 5. Gene and Motif Identification
- 6. Phylogenetic Analysis
- 7. Protein Structure Analysis and Prediction
- 8. RNA Secondary Structure Prediction
- 9. Clustering and Classification of Gene Expression Data
- 10. Search and Mining of Biological Databases, Databanks, and Literature