

Math and Stats for Sequence Alignment & Pattern Searching

David Wishart

University of Alberta

david.wishart@ualberta.ca

Objectives

- **Gain an awareness of importance of mathematics and statistics in alignment**
- **Develop a “practical” understanding of some key mathematical methods**
- **Develop a “practical” vocabulary of some key mathematical and statistical methods**
- **See applications in alignment, pattern finding, and pattern identification**

Bioinformatics

- **Mathematics** **→** **Mathematician**
- **Statistics** **→** **Statistician**
- **Informatics** **→** **Informatician**
- **Bioinformatics** **→** **Bioinformatician**
~~**Bioinformaticist**~~

Definition

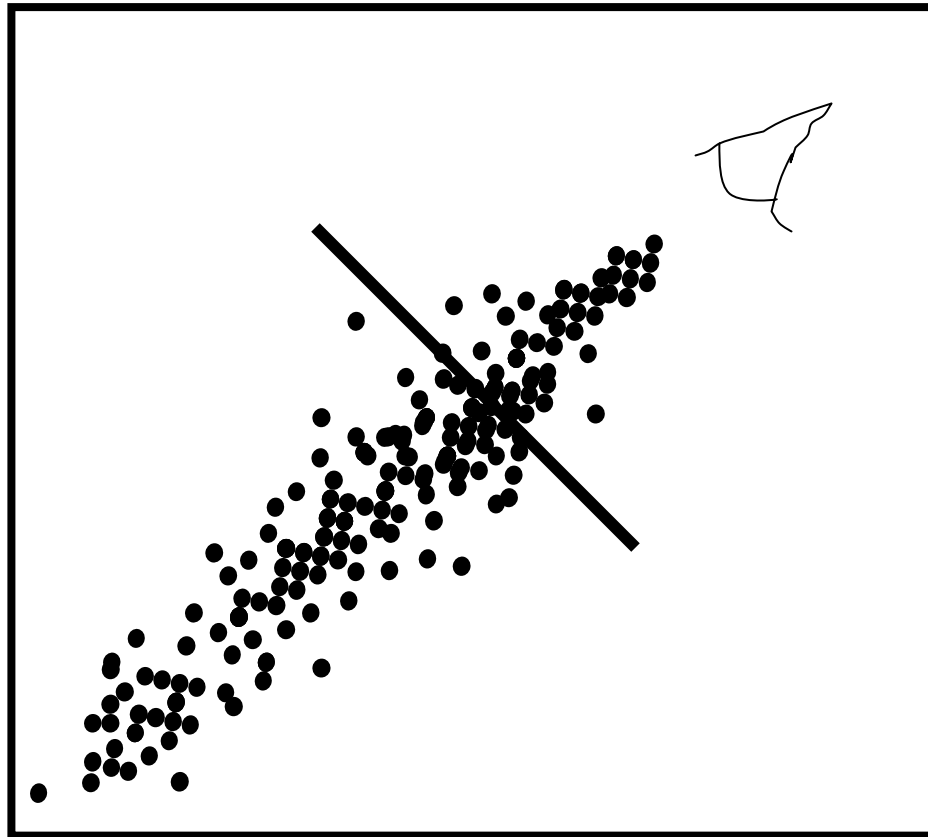
- ***Bioinformatics - The application of Mathematics, Statistics and Information Theory to the interpretation of Biological Data***

Some Perspective

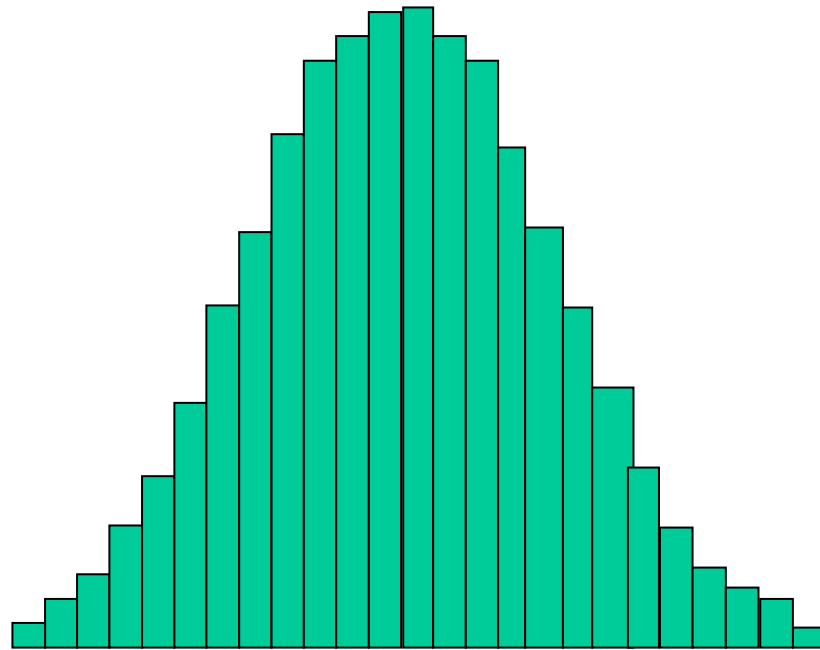
- **Mathematics and Statistics underlie nearly every aspect of Bioinformatics**
- **Most of the major breakthroughs in Bioinformatics arose through innovations in Mathematics or Statistics (i.e. FASTA, BLAST, Phred/Phrap, BLOSUM, GenScan, PSI-BLAST, Threading, GRAIL, etc.)**

Distributions & Significance

Cross Sectioning a Scatter Plot



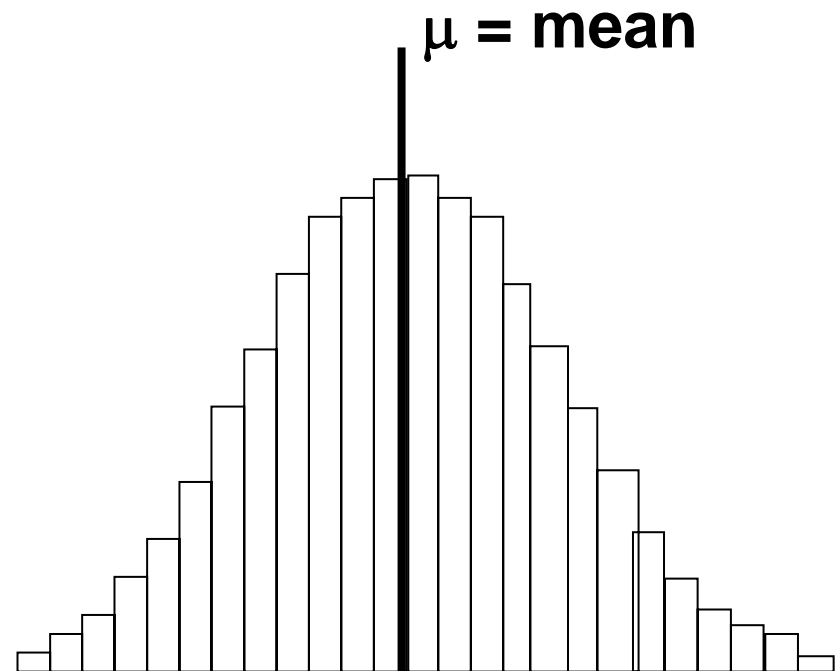
What kind of point scatter do you see?



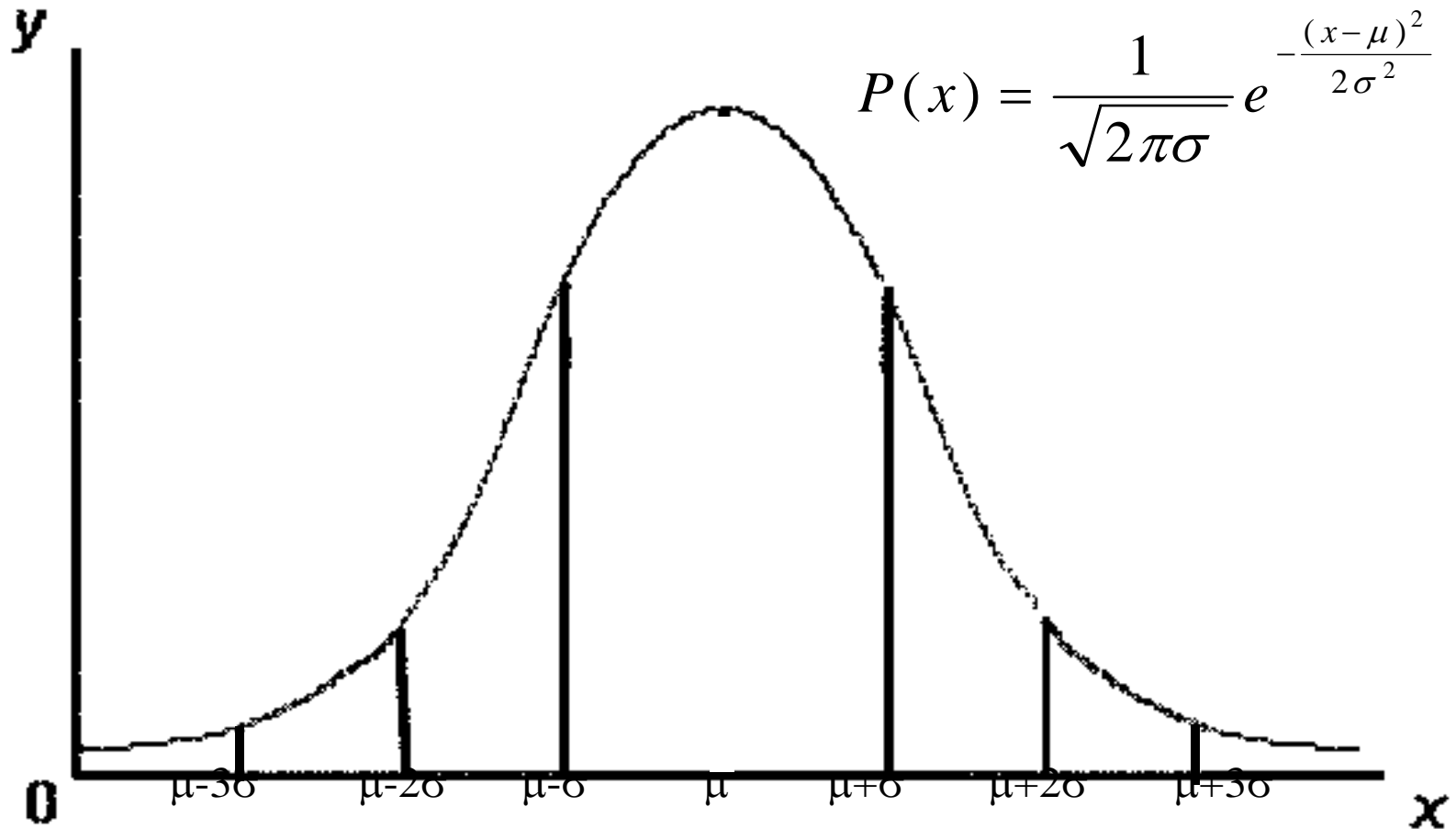
Gaussian or Normal Distribution

Features of a Normal Distribution

- **Symmetric Distribution**
- **Has an average or mean value (μ) at the centre**
- **Has a characteristic width called the standard deviation (σ)**
- **Most common type of distribution known**



Gaussian Distribution



Some Equations

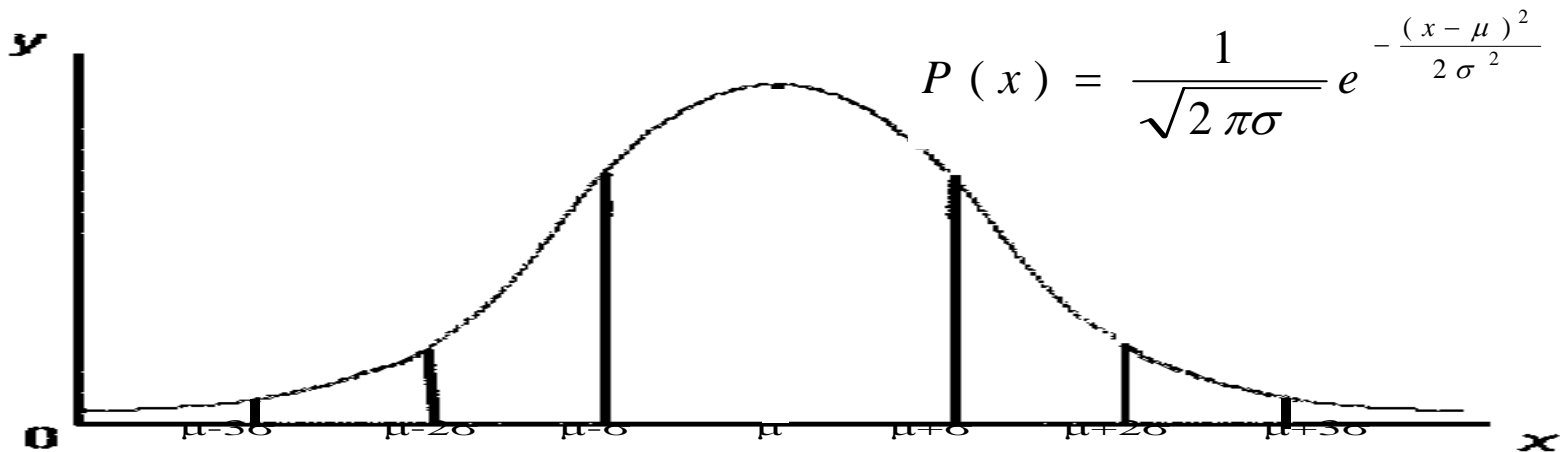
Mean $\mu = \frac{\sum x_i}{N}$

Variance $\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$

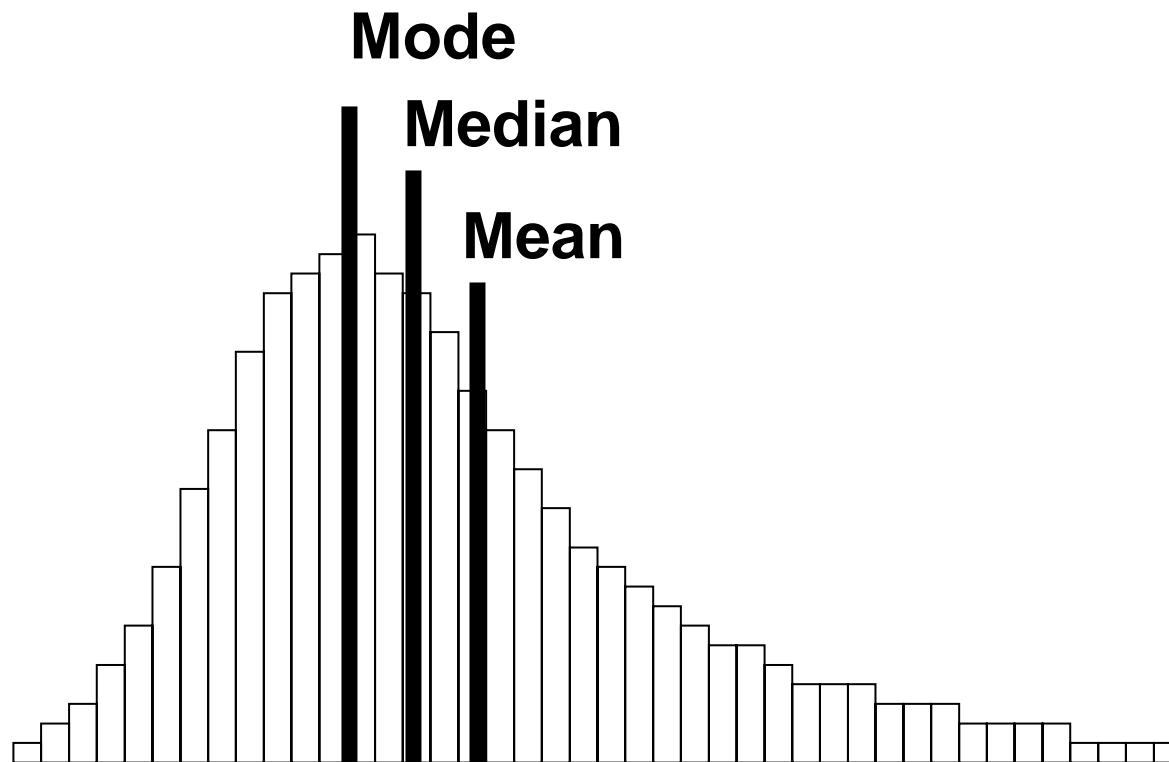
Standard Deviation $\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$

Standard Deviations (Z-values)

$\mu \pm 1.0$ S.D.	0.683	$> \mu + 1.0$ S.D.	0.158
$\mu \pm 2.0$ S.D.	0.954	$> \mu + 2.0$ S.D.	0.023
$\mu \pm 3.0$ S.D.	0.9972	$> \mu + 3.0$ S.D.	0.0014
$\mu \pm 4.0$ S.D.	0.99994	$> \mu + 4.0$ S.D.	0.00003
$\mu \pm 5.0$ S.D.	0.999998	$> \mu + 5.0$ S.D.	0.000001



Mean, Median & Mode

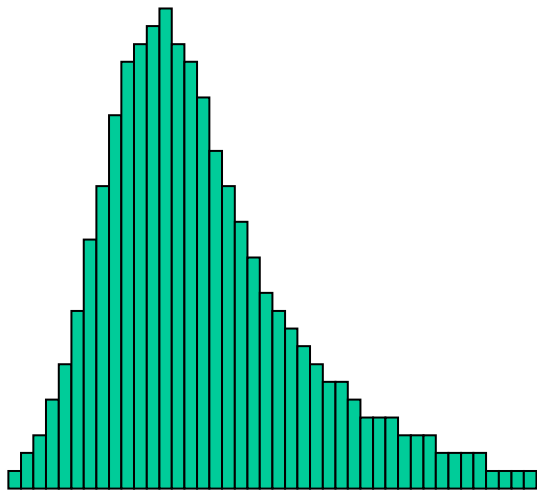


Mean, Median, Mode

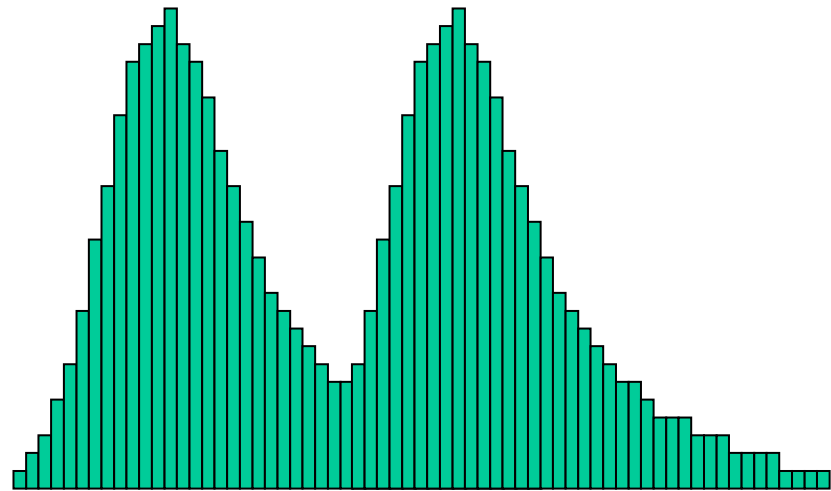
- In a Normal Distribution the mean, mode and median are all equal
- In skewed distributions they are unequal
- **Mean** - average value, affected by extreme values in the distribution
- **Median** - the “middlemost” value, usually half way between the mode and the mean
- **Mode** - most common value

Different Distributions

Unimodal



Bimodal

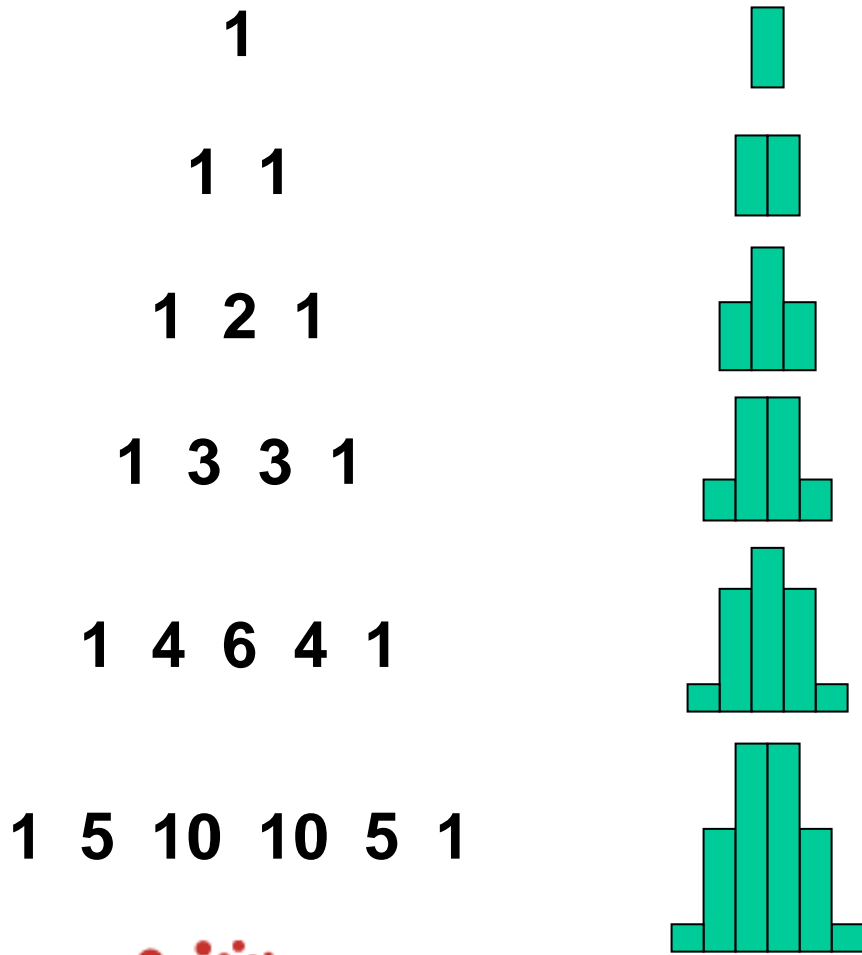


Other Distributions

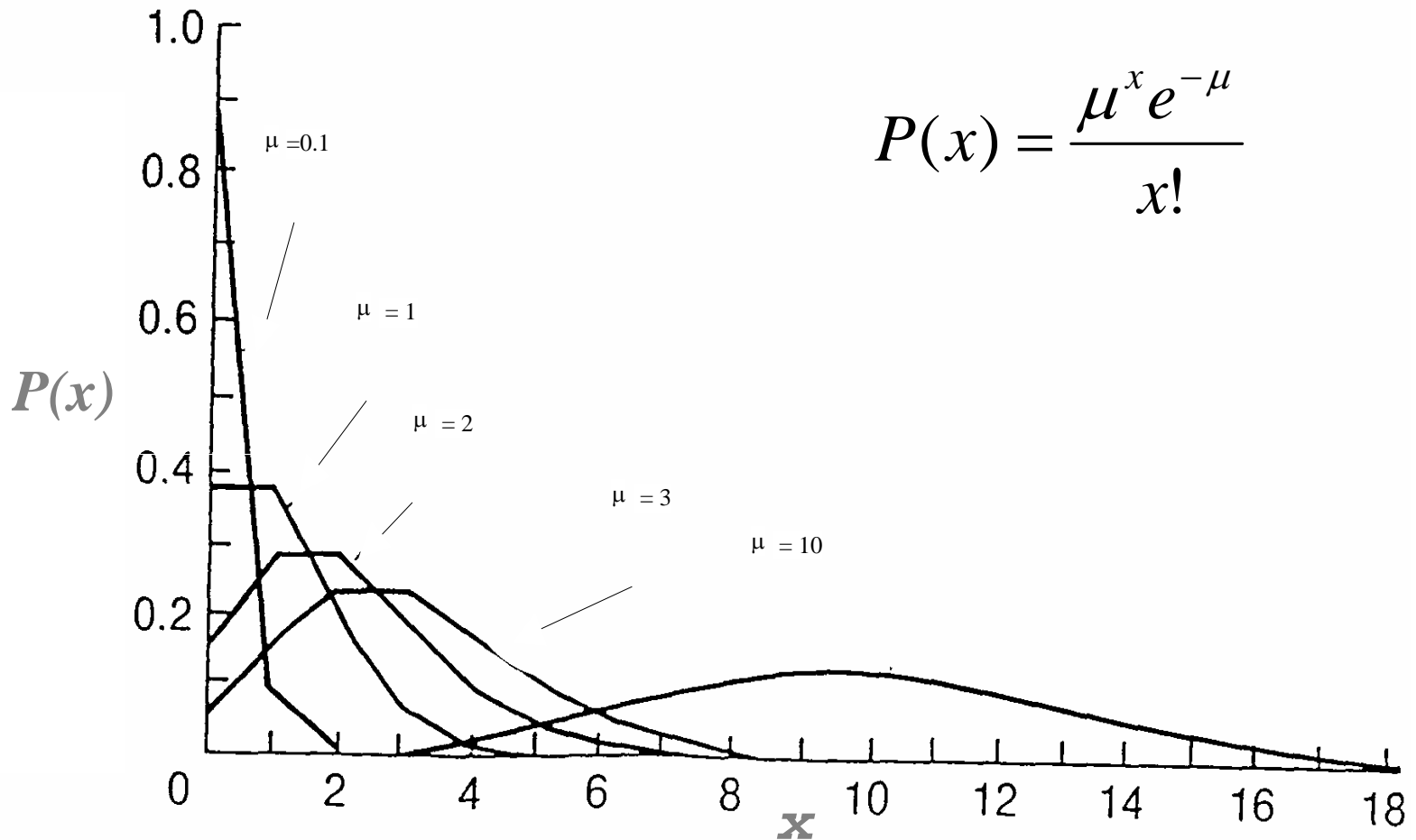
- **Binomial Distribution**
- **Poisson Distribution**
- **Extreme Value Distribution**
- **Boltzman Distribution**

Binomial Distribution

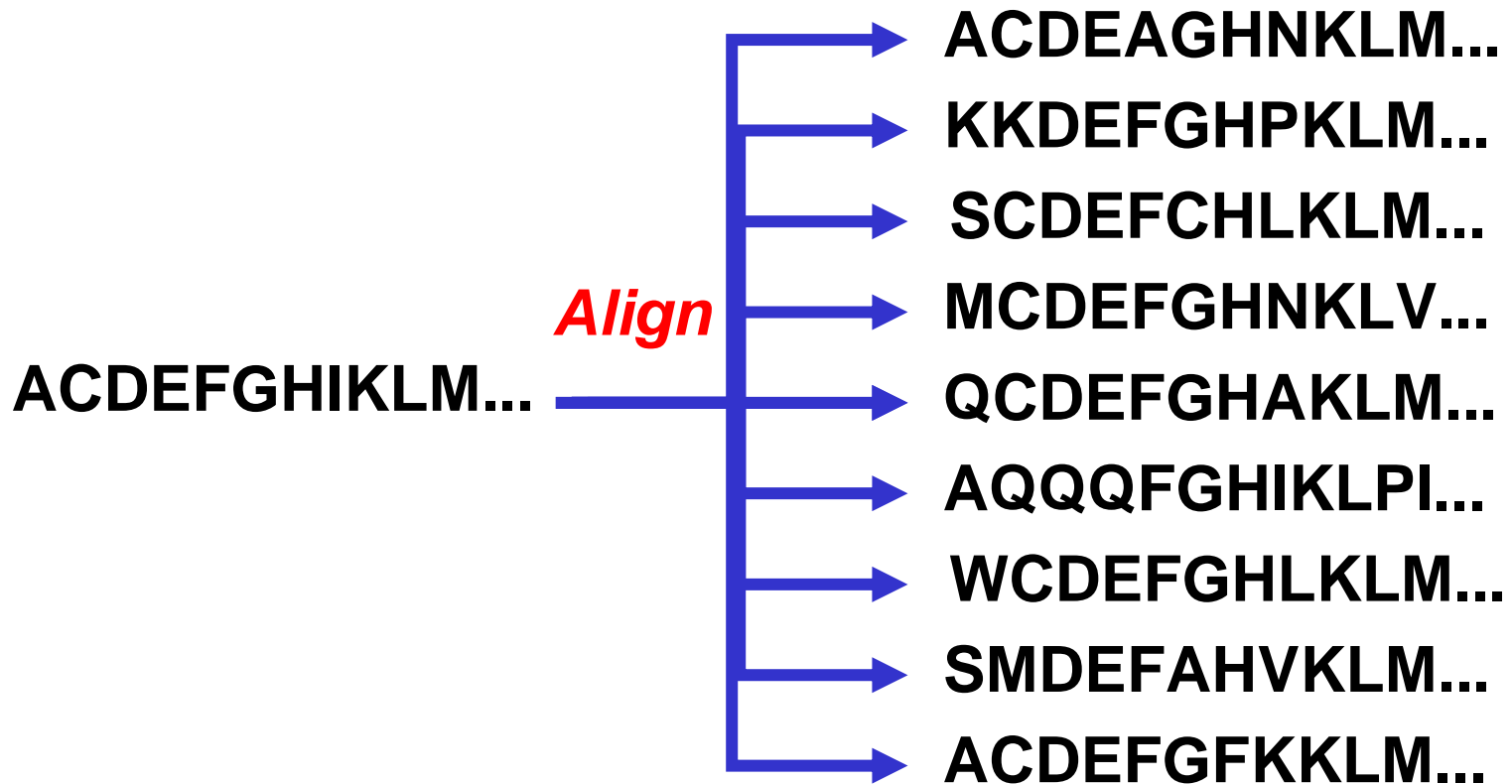
$$P(x) = (p + q)^n$$



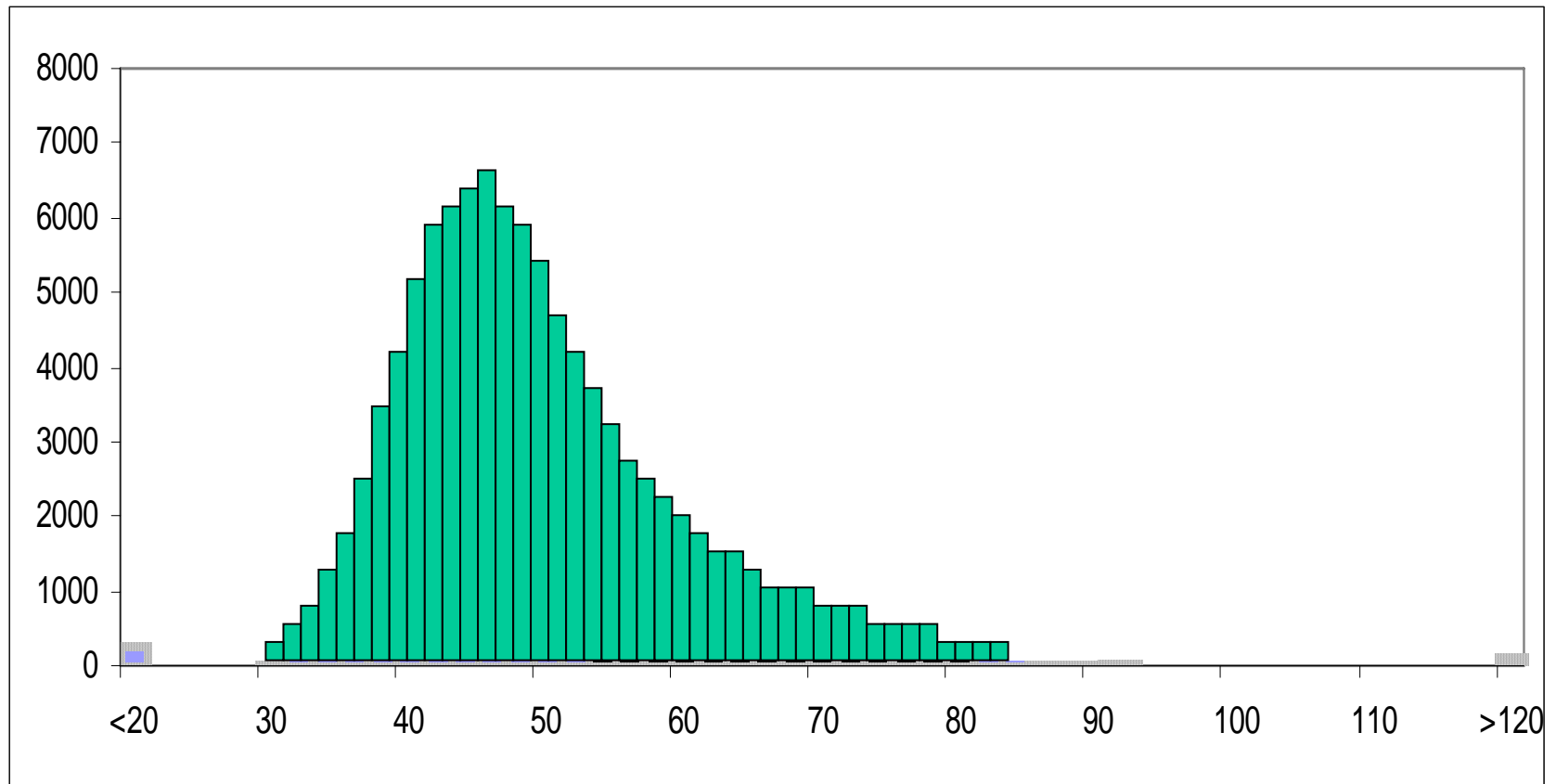
Poisson Distribution



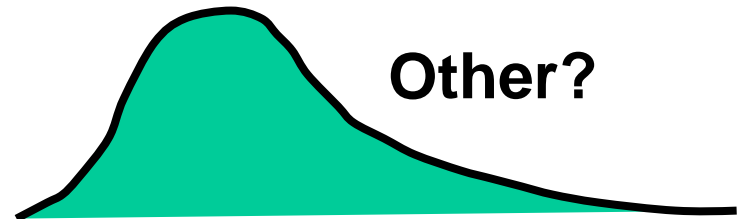
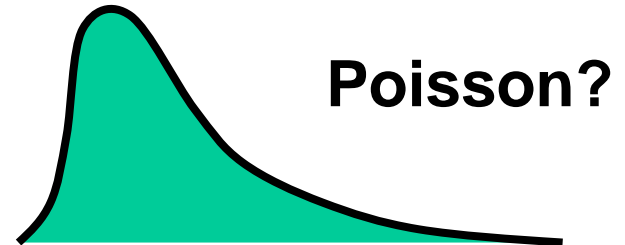
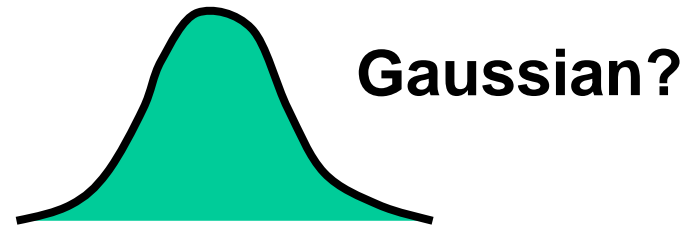
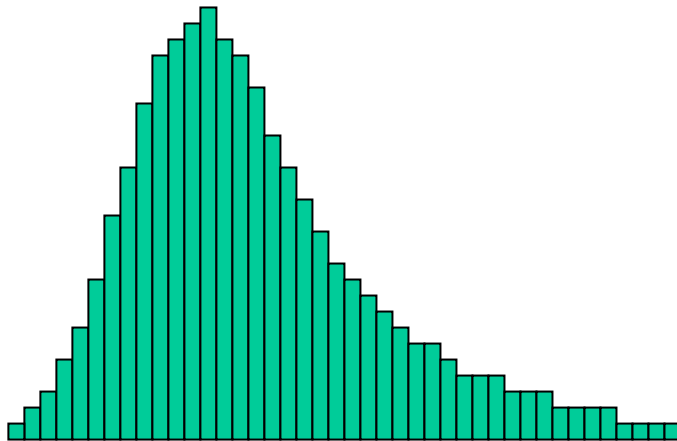
Let's try an experiment...



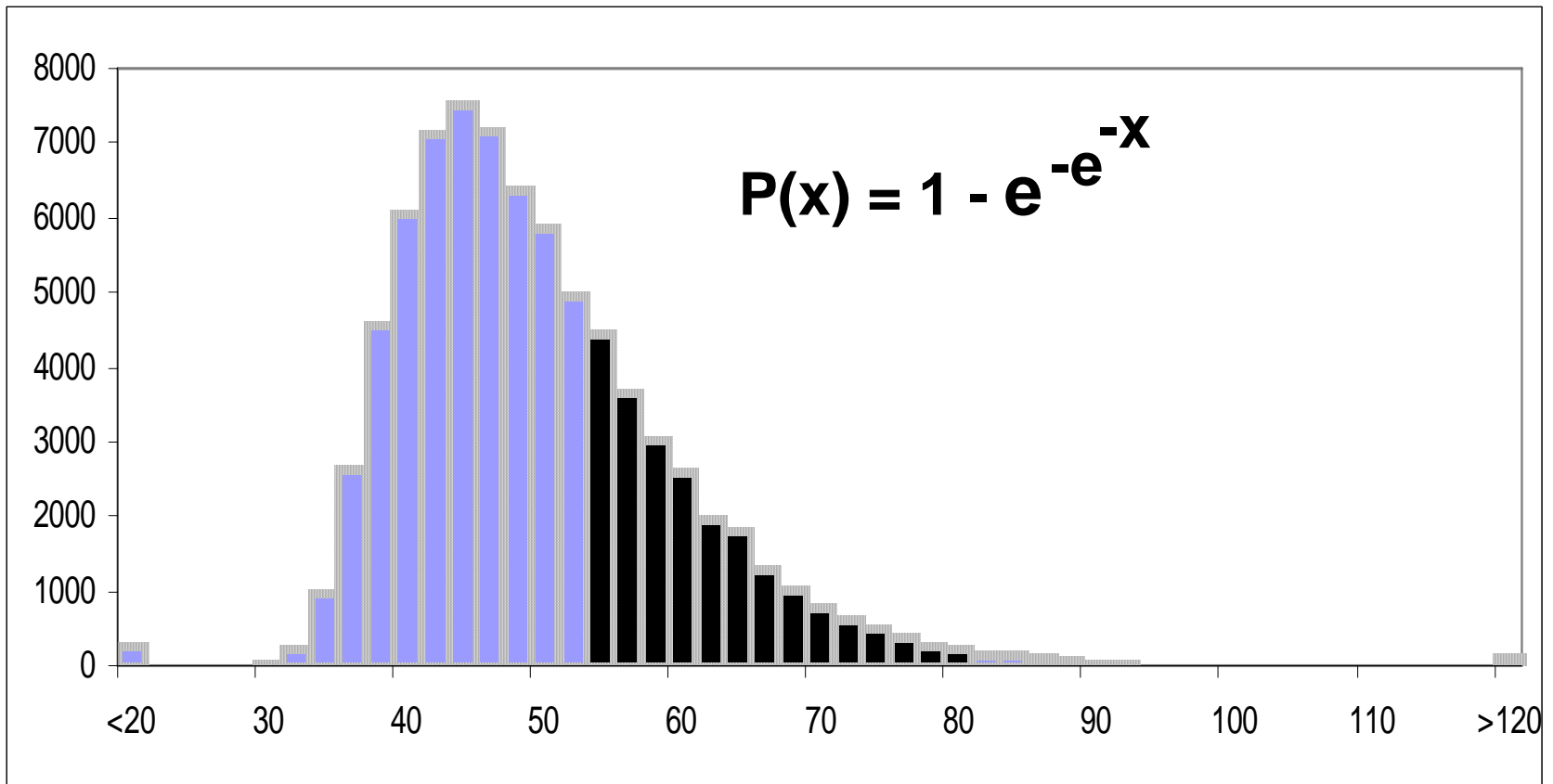
What kind of score distribution do you get?



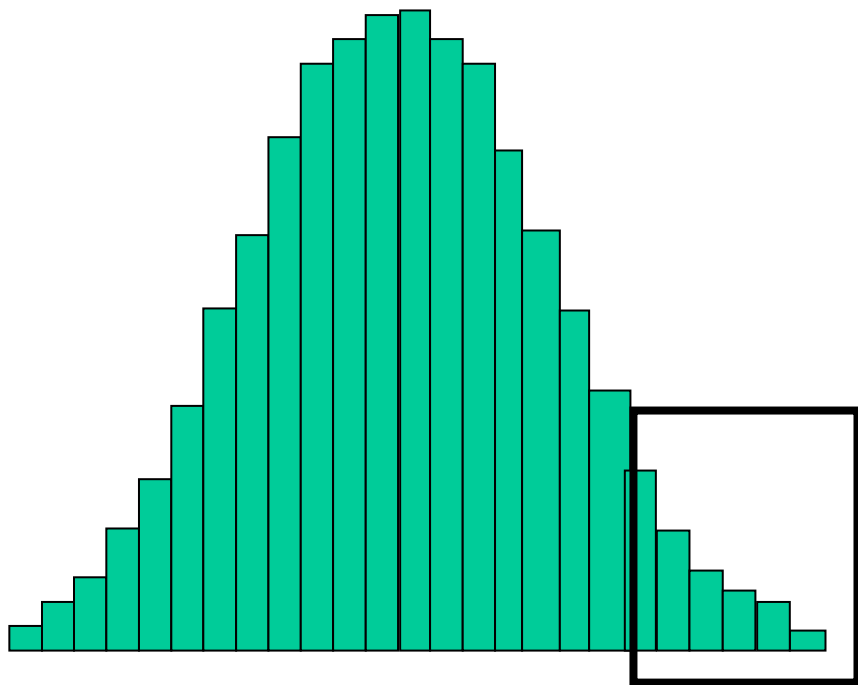
What kind of distribution?



Extreme Value Distribution



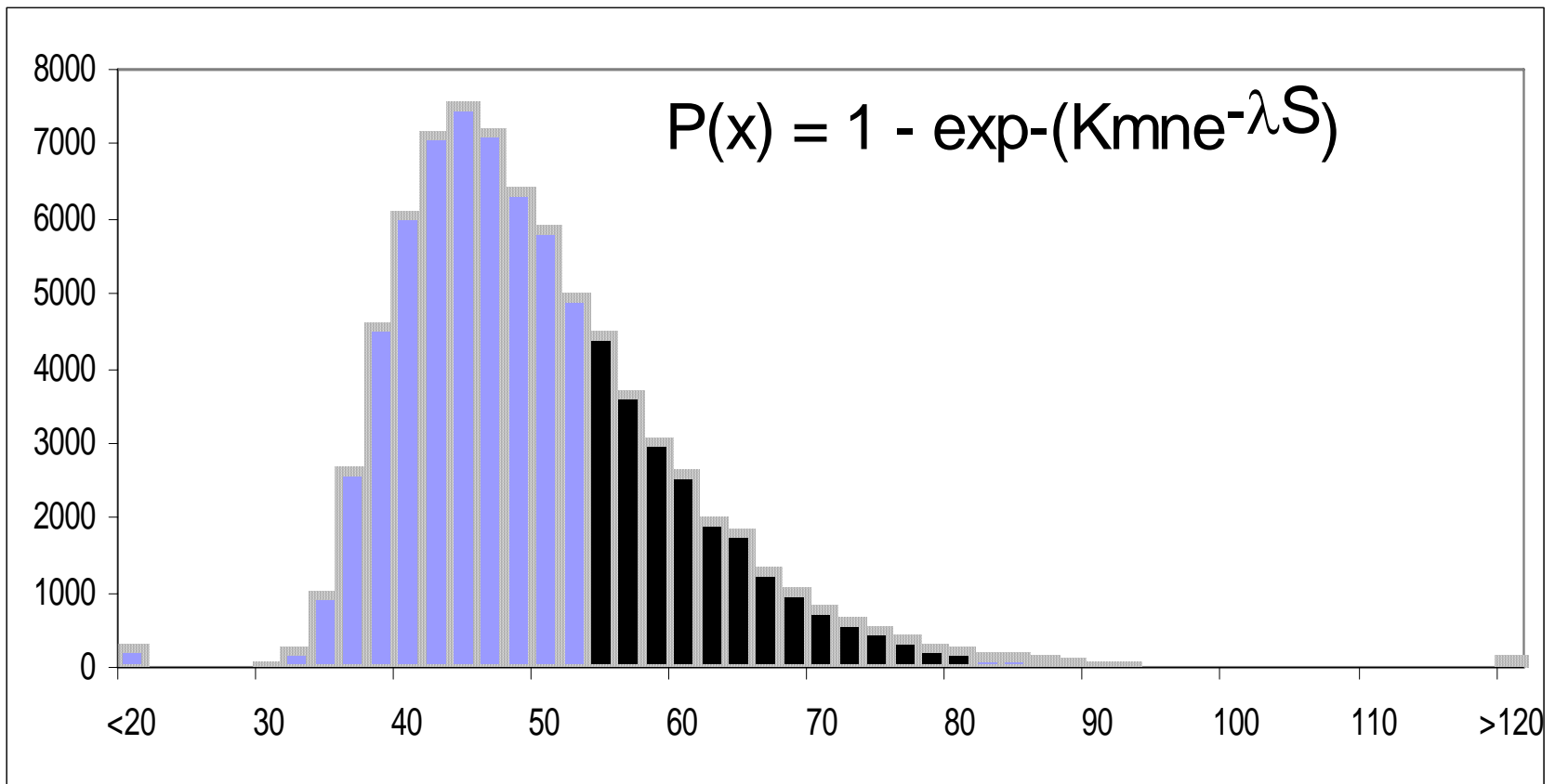
Extreme Value Distribution



Gaussian Distribution

- Arises from sampling the extreme end of a normal distribution
- A distribution which is “skewed” due to its selective sampling
- Skew can be either right or left

Extreme Value Distribution for BLAST



Extreme Value Distribution

- $Kmne^{-\lambda S}$ is called Expect or E-value
- In BLAST, Expect = 10 so $P = 0.99995$
- $P = 1 - \exp(-Kmne^{-\lambda S}) = 1 - e^{-10} = 0.99995$
- $E = 10$ means you'll find 10 random hits in your BLAST output with HSP values above the threshold score
- The lower the E value the lower the likelihood of getting a “random” hit

BLAST - Rules of Thumb

- Expect (E-value) is equal to the number of BLAST alignments with a given Score that are expected to be seen simply due to chance
- Don't trust a BLAST alignment with an Expect score > 0.01 (**Grey zone is between 0.01 - 1**)
- Expect and Score are related, but Expect contains more information. Note that %Identities is more useful than the bit Score
- Recall Doolittle's Curve (%ID vs. Length, next slide)
%ID > 30 - numres/50
- If uncertain about a hit, perform a PSI-BLAST search

BLAST Statistics

Database: All non-redundant GenBank CDS
translations+PDB+SwissProt+PIR+PRF
Posted date: Feb 3, 2004 9:48 AM
Number of letters in database: 534,067,077
Number of sequences in database: 1,624,011

Lambda	K	H
0.316	0.135	0.360

Gapped

Lambda	K	H
0.267	0.0410	0.140

Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Hits to DB: 16,762,026
Number of Sequences: 1624011
Number of extensions: 515013
Number of successful extensions: 1776
Number of sequences better than 10.0: 14
Number of HSP's better than 10.0 without gapping: 10
Number of HSP's successfully gapped in prelim test: 4
Number of HSP's that attempted gapping in prelim test: 1766
Number of HSP's gapped (non-prelim): 14
length of query: 77
length of database: 534,067,077
effective HSP length: 53
effective length of query: 24
effective length of database: 447,994,494
effective search space: 10751867856
effective search space used: 10751867856
T: 11
A: 40
X1: 16 (7.3 bits)
X2: 38 (14.6 bits)
X3: 64 (24.7 bits)
S1: 41 (21.6 bits)
S2: 66 (30.0 bits)

Extreme Value Distribution

- $Kmne^{-\lambda S}$ is called Expect or E-value
- In BLAST default $E = 10$ so $P = 0.99995$
- If E is small (<0.01) then P is small
- If Matches = 1 and Mismatches = -1 then:
 - ♦ $\lambda = \ln q/p$ and $K = (q-p)^2/q$
 - $p =$ probability of match = 0.05
 - $q =$ probability of not match = 0.95
 - Then $\lambda = 2.94$ and $K = 0.85$
 - $m =$ length of sequence & $n =$ length of database
 - $S =$ score for given HSP

BLAST Options

- **Composition-based statistics (Yes)**
- **Sequence Complexity Filter (Yes)**
- **Expect (E) value (10)**
- **Word Size (3)**
- **Substitution or Scoring Matrix (Blosum62)**
- **Gap Insertion Penalty (11)**
- **Gap Extension Penalty (1)**

Composition Statistics

- **Recent addition to BLAST algorithm**
- **Permits calculated E (Expect) values to account for amino acid composition of queries and database hits**
- **Improves accuracy and reduces false positives**
- **Effectively conducts a different scoring procedure for each sequence in database**

Scoring Matrices

- **BLOSUM Matrices**
 - Developed by Henikoff & Henikoff (1992)
 - **B**LOcks **S**Ubstitution **M**atrix
 - Derived from the BLOCKS database
- **PAM Matrices**
 - Developed by Schwarz and Dayhoff (1978)
 - **P**oint **A**ccepted **M**utation
 - Derived from manual alignments of closely related proteins

How to Make Your Own Matrix

ACDEFGH..
 ACDEFGK..
 AADEFGH..
 GCDEFGH..
 ACAEY GK..
 ACAEF AH..

$$f(A,A) = \frac{\#A_{\text{obs}}}{\#A_{\text{exp}}}$$

$$f(C,A) = \frac{\#C/A_{\text{obs}}}{\#A_{\text{exp}} + \#C_{\text{exp}}}$$

	A	C	D ...
A	0.8	--	--
C	0.2	0.8	--
D	0.0	0.3	1.0
E	--	--	--

Perform
Alignment

Calculate
Frequencies

Fill Sub
Matrix

The Odds Ratio

- Odds ratio is the ratio of the number of times residue "A" is observed to replace residue "B" divided by the number of times residue "A" would be expected to replace residue "B" if replacements occurred at random
- Positive scores in a PAM matrix designate a pair of residues that replace each other more often than expected by chance

How to Make a PAM Matrix

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

X

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

X
=

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

Multiply Matrices N times to make PAM "X"

log

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

Take the log

PAM versus BLOSUM

- **First useful scoring matrix for protein**
- **Assumed a Markov Model of evolution (i.e. all sites equally mutable and independent)**
- **Derived from small, closely related proteins with ~15% divergence**
- **Much later entry to matrix “sweepstakes”**
- **No evolutionary model is assumed**
- **Built from PROSITE derived sequence blocks**
- **Uses much larger, more diverse set of protein sequences (30% - 90% ID)**

PAM versus BLOSUM

- Higher PAM numbers to detect more remote sequence similarities
- Lower PAM numbers to detect high similarities
- 1 PAM ~ 1 million years of divergence
- Errors in PAM 1 are scaled 250X in PAM 250
- Lower BLOSUM numbers to detect more remote sequence similarities
- Higher BLOSUM numbers to detect high similarities
- Sensitive to structural and functional substitution
- Errors in BLOSUM arise from errors in alignment

PAM Matrices

- **PAM 40** - prepared by multiplying PAM 1 by itself a total of 40 times
best for short alignments with high similarity
- **PAM 120** - prepared by multiplying PAM 1 by itself a total of 120 times
best for general alignment
- **PAM 250** - prepared by multiplying PAM 1 by itself a total of 250 times
best for detecting distant sequence similarity

PAM250

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

BLOSUM Matrices

- **BLOSUM 90** - prepared from BLOCKS sequences with >90% sequence ID
best for short alignments with high similarity
- **BLOSUM 62** - prepared from BLOCKS sequences with >62% sequence ID
best for general alignment (default)
- **BLOSUM 30** - prepared from BLOCKS sequences with >30% sequence ID
best for detecting weak local alignments

BLOSUM62

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

Sequence Patterns

AHGQSDFILDEADGMMKSTVPN..
HGFDSA AVLDEADHILQWERTY..
GGGNDEYIVDEADSVIASDFGH..

*** [LIVM] [LIVM] DEAD * [LIVM] [LIVM] ***

(EIF 4A ATP DEPENDENT HELICASE)

Expectation Values

- **Expectation value (ϵ) is the expected number of hits for a given sequence pattern or motif**
- **$\epsilon = N \times f_1 \times f_2 \times f_3 \times \dots \times f_k$**
- **N is the number of residues in DB (10^8)**
- **f_i is the frequency of a given amino acid(s)**

Residue	Frequency	Residue	Frequency
A	8.80%	M	1.97%
C	2.05%	N	4.58%
D	5.91%	P	4.48%
E	5.89%	Q	3.84%
F	3.76%	R	4.22%
G	8.30%	S	6.50%
H	2.15%	T	5.91%
I	5.40%	V	7.05%
K	6.20%	W	1.39%
L	8.09%	Y	3.52%

Pattern Example #1

ACIDS

$$\varepsilon = 10^8 * 0.088 * 0.021 * 0.054 * 0.059 * 0.065$$

$$\varepsilon = 38.3$$

#Found in OWL database = 14

Pattern Example #2

A*ACI[DEN]S

$$\varepsilon = 10^8 * 0.088 * 1.000 * 0.088 * 0.021 * 0.054 * \{0.059 + 0.059 + 0.046\} * 0.065$$

$$\varepsilon = 9.4$$

#Found in OWL database = 9

Minimum Pattern Lengths

$$f = 0.08 \quad \varepsilon = 10^8 * 0.08^8 = 0.17 \quad \text{min} = 8$$

$$f = 0.05 \quad \varepsilon = 10^8 * 0.05^7 = 0.08 \quad \text{min} = 7$$

$$f = 0.03 \quad \varepsilon = 10^8 * 0.03^6 = 0.07 \quad \text{min} = 6$$

How Long Should a Sequence Motif or Sequence Block Be?

- How many matching segments of length “l” could be found in comparing a query of length M to a DB of N ?
- Answer:
 $n(l) = M \times N \times f^l$
- Assume $f = 0.05$, $M = 300$, $N = 100,000,000$

n	l
3,750,000	3
187,500	4
9375	5
469	6
23	7
1.2	8
0.058	9

What's the Longest Random Block of Pairwise Matches in OWL?

- How many matching segments of length “l” could be found in comparing each to each in a DB of N ?
- Answer:
 $n(l) = (N-1) \times N/2 \times f^l$
- Assume $f = 0.05$, $N = 100,000,000$

n	l
3,906,000	7
195,312	8
9765	9
488	10
24.4	11
1.22	12
0.061	13

Rule of Thumb

**Make your
protein sequence
motifs at least
8 residues long**

PROSITE Pattern Expressions

C - [ACG] - T - Matches CAT, CCT and CGT only

C - X - T - Matches CAT, CCT, CDT, CET, etc.

C - {A} - T - Matches every CXT except CAT

C - (1,3) - T - Matches CT, CCT, CCCT

C - A(2) - [TP] - Matches CAAT, CAAP

[LIV] - [VIC] - X(2) - G - [DENQ] - X - [LIVFM] (2) - G

Protein Sequence Motifs

- **Pattern-Based (PQL) Sequence Motifs**
- **>*TCP&NLGT***
- **DOOLITTLE, R.F., OF URFS AND ORFS (1986)**
- **GUANIDINE KINASE ACTIVE SITE**
- **Profiles or Position Scoring Matrix (PSSM)**

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

Finding Sequence Motifs

	1				50
P43871-1IKKLDSN	SIHAIISDIP	YGIDYDDWDI	LHSNTNSALG
S18997-1	LMSKIYQMDA	VDWLKTLENC	SVDLFITDPP	YESL.EKYRQ	IGTTTRLKES
P23192-1	EINKIHQMNC	FDFLDQVENK	SVQLAVIDPP	YNL.....
P29538-1	MDQRLICNA	IKALKNLEEN	SIDLIITDPP	YNLG.KDY..
P14751-1	TRHVYDVCDC	LDTLAKLPDD	SVQLIICDPP	YNI.....
P34721-1	KNFNIIYQGNC	IDFMSHFQDN	SIDMIFADPP	YFLS.NDG.L	TFKNSIIQ..
P50178-1	ENAILVHADS	FKLLEKIKPE	SMDMIFADPP	YFLS.NGG.M	SNSGGQIV..
P20590-1	FLNTILKGDC	IEKLKTIPNE	SIDLIFADPP	YFMQ.TEGKL	LRTNGDEF..
S43876-1	GPETIIHGDC	IEQMNALPEK	SVDLIFADPP	YNLQ.LGGDL	LRPDNSKV..
P28638-1	EAKTIIHGDA	LAELKKIPAE	SVDLIFADPP	YNIG.KNF..
P23941-1	DLGKLYNGDC	LELFKQVPDE	NVDTIFADPP	FNLD.KEY..
P14230-1	RSCKIIVGDA	REAVQGLDSE	IFDCVVTSP	YWGL.RDY..
P14243-1	NGATLFEGDA	LSVLRRLPSG	SVRCIVTSP	YWGL.RDY..
Q04845-1	LNNMLLQGNC	AETLKKLPDE	SVNLVFTSP	YY.....
S53866-1	WVNDIHEGDA	EEVLAELPES	SVHVMVTSP	YFGL.RDY..
P29568-1MNELKDK	SINLVVTSP	YPMV.EIWDR	LFSELNPKIE

Signature Sequence:

DPP Y

Rules of Thumb

- **Pattern-based sequence motifs should be determined from no fewer than 5 multiply aligned sequences**
- **A good degree of sequence divergence is needed. If “S” is the %similarity and “N” is the no. of sequences then $1 - S^N > 0.95$**
- **A good sequence pattern should have no fewer than 8 defined amino acid positions**

Sequence Pattern Databases

- **PROSITE** - <http://ca.expasy.org/prosite/>
- **BLOCKS** - <http://www.blocks.fhcrc.org/>
- **DOMO** - <http://www.infobiogen.fr/services/domo/>
- **PFAM** - <http://pfam.wustl.edu>
- **PRINTS** - <http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/>
- **SEQSITE** - PepTool

PSSM's & Profiles



A N V R Q C
A
P I A Y A
A F G
H A A K
A L A T A
G A

Profiles & PSSMs Need Multi-sequence Alignment

```

1                                                                                               50
P43871-1 ..... ...IKKLDSN SIHAIISDIP YGIDYDDWDI LHSNTNSALG
S18997-1 LMSKIYQMDA VDWLKTLENC SVDLFITDPP YESL.EKYRQ IGTTRRLKES
P23192-1 EINKIIHQMNC FDFLDQVENK SVQLAVIDPP YNL.....
P29538-1 MDQRILCSNA IKALKNLEEN SIDLIITDPP YNLG.KDY..
P14751-1 TRHVYDV CDC LDTLAKLPDD SVQLIICDPP YNI.....
P34721-1 KNFNIIYQGNC IDFM SHFQDN SIDMIFADPP YFLS.NDG.L TFKNSIIQ..
P50178-1 ENAILVHADS FKLLEKIKPE SMDMIFADPP YFLS.NGG.M SNSGGQIV..
P20590-1 FLNTILKGDC IEKLKTIPNE SIDLIFADPP YFMQ.TEGKL LRTNGDEF..
S43876-1 GPETIIHGDC IEQMNALPEK SVDLIFADPP YNLQ.LGGDL LRPDNSKV..
P28638-1 EAKTIIHGDA LAELKKIPAE SVDLIFADPP YNIG.KNF..
P23941-1 DLGKLYNGDC LELFKQVPDE NVDTIFADPP FNLD.KEY..
P14230-1 RSCKIIVGDA REAVQGLDSE IFDCVVTSP YWGL.RDY..
P14243-1 NGATLFEGDA LSVLRRLPSG SVRCIVTSP YWGL.RDY..
Q04845-1 LNNMLLQGNC AETLKKLPDE SVNLFVTSPP YY.....
S53866-1 WVNDIHEGDA EEVLAELPES SVHVMVTSPP YFGL.RDY..
P29568-1 ..... ...MNELKDK SINLVVTSPP YPMV.EIWDR LFSELNPKIE

```

Signature Sequence:

DPP Y

Building a PSSM - Step 1

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

*Multiple
Alignment*



A	3	2	0	0	1	0	0	5	2	1
C	1	0	0	2	0	0	0	0	1	4
G	1	0	1	0	0	5	0	0	1	0
T	0	3	4	3	4	0	5	0	1	0

*Table of
Occurrences*

Building a PSSM - Step 2

A	3	2	0	0	1	0	0	5	2	1
C	1	0	0	2	0	0	0	0	1	4
G	1	0	1	0	0	5	0	0	1	0
T	0	3	4	3	4	0	5	0	1	0

Table of Occurrences



A	.6	.4	0	0	.2	0	0	1	.4	.2
C	.2	0	0	.4	0	0	0	0	.2	.8
G	.2	0	.2	0	0	1	0	0	.2	0
T	0	.6	.8	.6	.8	0	1	0	.2	0

PSSM with no pseudocounts

Pseudocounts

- **Method to account for small sample size of multi-sequence alignment**
- **Gets around problem of having “0” score in PSSM or profile**
- **Defined by a correction factor “B” which reflects overall composition of sequences under consideration**
- **$B = \sqrt{N}$ or $B = 0.1$ which falls off with N where N = # sequences**

Pseudocounts

- **Score(X_i) = $(q_x + p_x)/(N + B)$**
- **q = observed counts of residue X at pos. i**
- **p = pseudocounts of X = B*frequency(X)**
- **N = total number of sequences in MSA**
- **B = number of pseudocounts (assume \sqrt{N})**

$$\text{Score}(A_1) = (3 + \sqrt{5(0.32)}) / (5 + \sqrt{5}) = 0.51$$

Including Pseudocounts - Step 2

A	3	2	0	0	1	0	0	5	2	1
C	1	0	0	2	0	0	0	0	1	4
G	1	0	1	0	0	5	0	0	1	0
T	0	3	4	3	4	0	5	0	1	0

*Table of
Occurrences*



A	.51	.38	.09	.09	.24	.09	.09	.79	.38	.24
C	.19	.06	.06	.33	.06	.06	.06	.06	.19	.61
G	.19	.06	.19	.06	.06	.75	.06	.06	.19	.06
T	.09	.51	.65	.51	.65	.09	.79	.09	.24	.09

*PSSM with
pseudocounts*

Calculating Log-odds - Step 3

A	.51	.38	.09	.09	.24	.09	.09	.79	.38	.24
C	.19	.06	.06	.33	.06	.06	.06	.06	.19	.61
G	.19	.06	.19	.06	.06	.75	.06	.06	.19	.06
T	.09	.51	.65	.51	.65	.09	.79	.09	.24	.09

***PSSM with
pseudocounts***

 **Log₁₀**

A	0.2	0.4	2.2	2.2	0.7	2.2	2.2	0.0	0.4	0.7
C	0.7	2.5	2.5	0.4	2.5	2.5	2.5	2.5	0.7	0.1
G	0.7	2.5	0.7	2.5	2.5	0.0	2.5	2.5	0.7	2.5
T	2.2	0.2	0.1	0.2	0.1	2.2	0.0	2.2	0.7	2.2

***Log-odds
PSSM***

A Sample Protein PSSM/Profile (using Log_2)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	
1 W G V L	3	-2	3	4	0	4	-1	3	-1	4	4	1	1	1	-2	1	2	6	-6	-
2 L L S P	2	-2	-2	-1	3	0	-1	3	-1	6	5	-1	3	0	-1	3	1	4	1	-
3 V V V V	2	2	-2	-2	2	2	-3	11	-2	8	6	-2	1	-2	-2	0	2	15	-9	-
4 K E A T	6	-2	5	6	-5	4	1	0	5	-2	0	3	3	3	1	3	6	0	-6	-
5 A P L P	6	-1	0	1	-2	2	0	1	0	2	2	0	8	2	0	2	2	3	-5	-
6 G G G G	7	1	7	5	-6	15	-1	-3	0	-4	-3	4	3	6	1	6	2	-1	-6	-
7 S S Q E	4	-1	7	7	-6	7	2	-2	2	-3	-2	4	3	6	1	6	2	-1	-6	-
8 S S T P	4	4	2	2	-4	4	-1	0	2	-3	-2	2	7	0	1	10	6	0	-2	-

Scoring a Query Sequence

		A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W		
1	W G V L	V	3	-2	3	4	0	4	-1	3	-1	4	4	1	1	1	-2	1	2	6	-6	-
2	L L S P	L	2	-2	-2	-1	3	0	-1	3	-1	6	5	-1	3	0	-1	3	1	4	1	-
3	V V V V	V	2	2	-2	-2	2	2	-3	11	-2	8	6	-2	1	-2	-2	0	2	15	-9	-
4	K E A T	A	6	-2	5	6	-5	4	1	0	5	-2	0	3	3	3	1	3	6	0	-6	-
5	A P L P	P	6	-1	0	1	-2	2	0	1	0	2	2	0	8	2	0	2	2	3	-5	-
6	G G G G	G	7	1	7	5	-6	15	-1	-3	0	-4	-3	4	3	6	1	6	2	-1	-6	-
7	S S Q E	D	4	-1	7	7	-6	7	2	-2	2	-3	-2	4	3	6	1	6	2	-1	-6	-
8	S S T P	S	4	4	2	2	-4	4	-1	0	2	-3	-2	2	7	0	1	10	6	0	-2	-

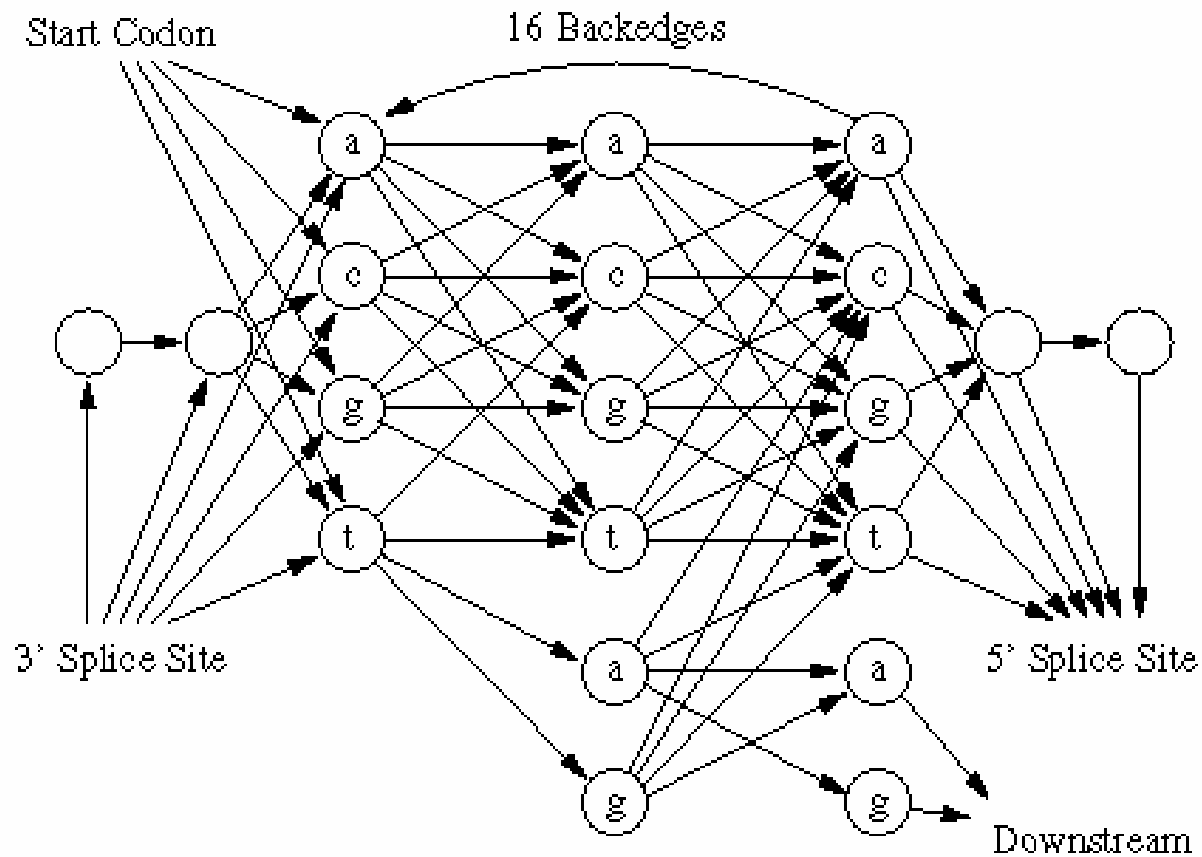
$$\mathbf{VLVAPGDS = 6+6+15+6+8+15+7+10=66}$$

$$\mathbf{LVLGPGA = 4+4+8+4+8+15-3+4= 44}$$

Profiles & PSSMs are Useful

- Helped identify active site of HIV protease
- Helped identify SH2/SH3 class of STP's
- Helped identify important GTP oncoproteins
- Helped identify hidden leucine zipper in HGA
- Regularly used to predict T-cell epitopes
- *Used in PSI-BLAST and PHI-BLAST*

Hidden Markov Models



Hidden Markov Models in Bioinformatics

- **Used extensively in Prokaryotic and Eukaryotic Gene Prediction**
- **Used to create Sequence Profiles and to classify sequences into families**
- **Used in Multiple Sequence Alignment**
- **Used in detecting and modeling protein domains or modules**
- **Used in recognizing protein topology**

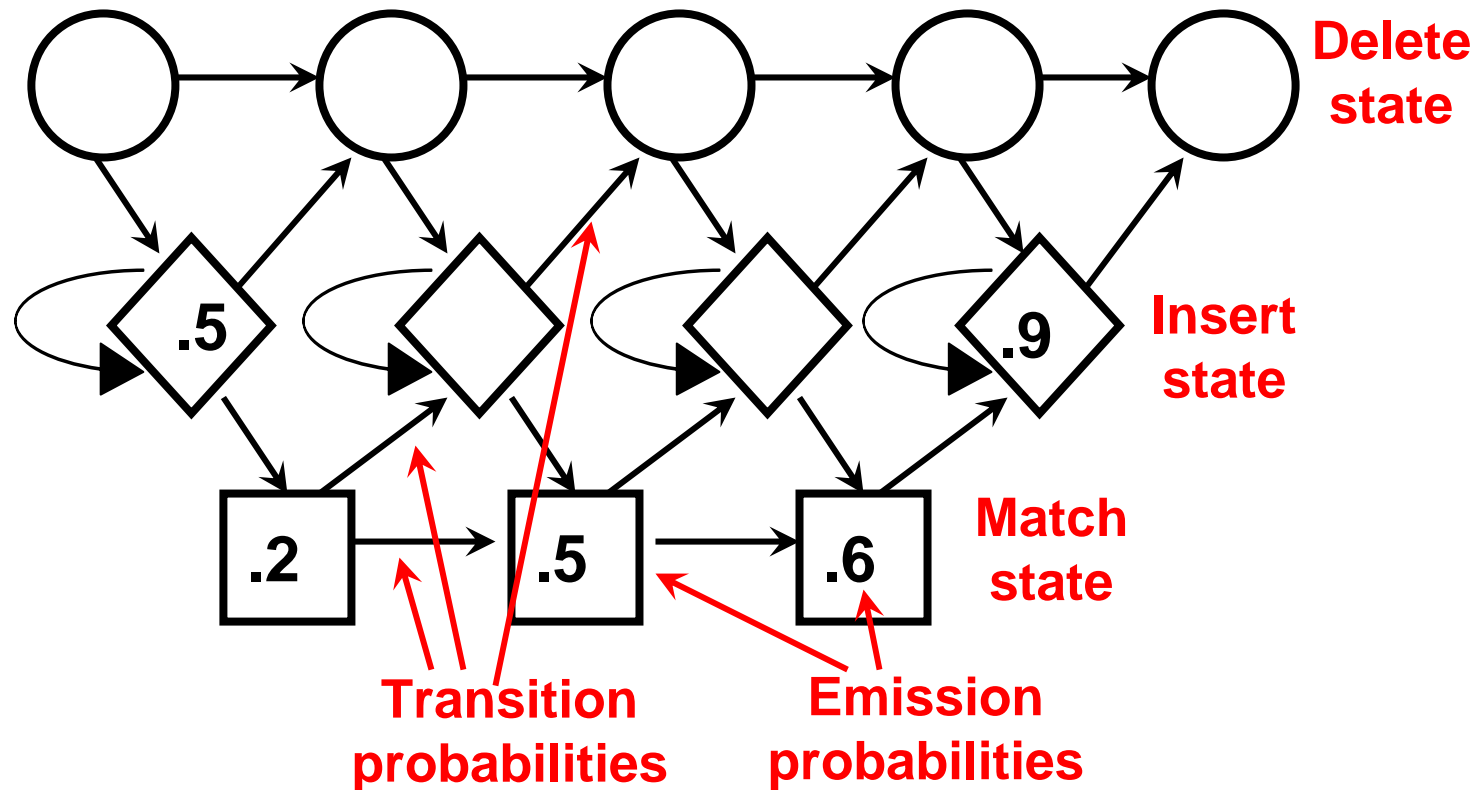
Hidden Markov Models

- **Flexible statistical or probabilistic model**
- **A more systematic approach to estimating model parameters or model probabilities**
- **Excellent for creating generalized weight matrices and for dealing with “gaps”**
- **Generally more powerful than profiles, PSSMs or dynamic programming**

Hidden Markov Models

- **Markov Model is a chain of events or states**
- **Each state has a set of emission probabilities for occupying that state**
- **MSA creates a Markov model of emission and transition probabilities**
- **Typically have a “Topology” which assumes a sequence of events is a multiplicative product of individual probabilities (independent, 1st order)**

Hidden Markov Topology



Making a Markov Model

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

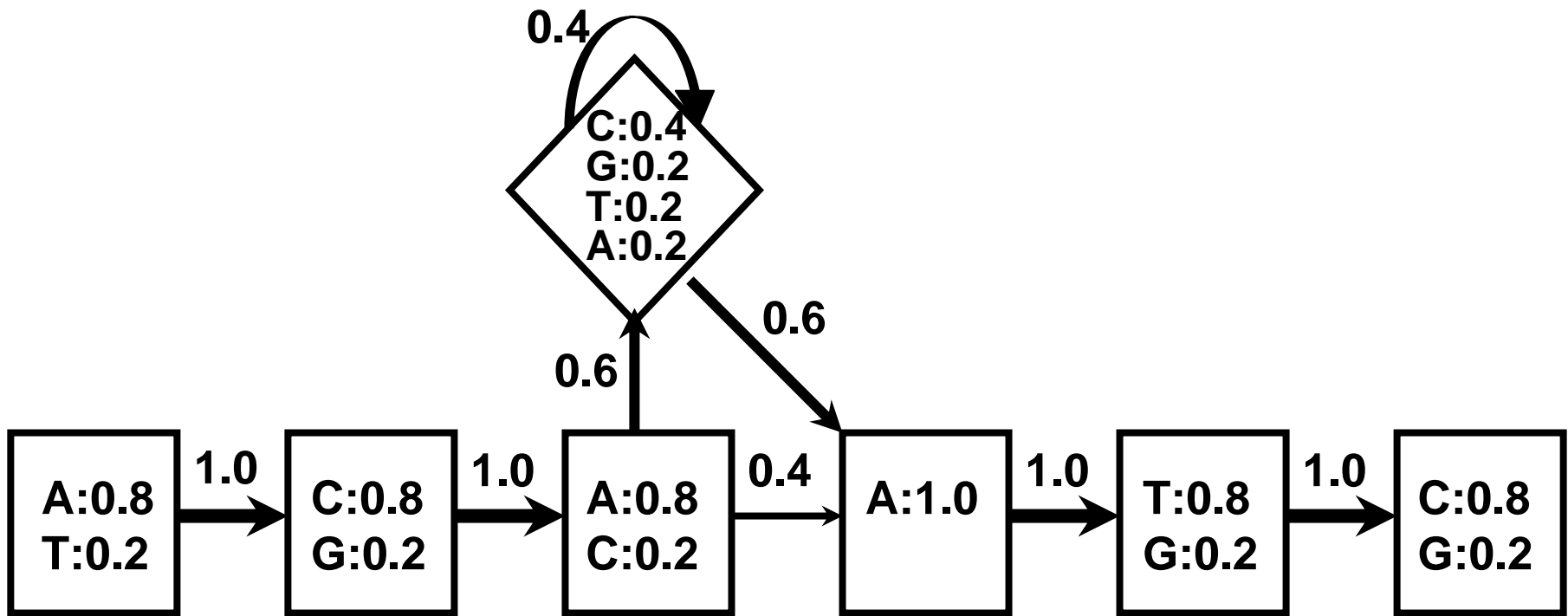
[AT] [CG] [AC] [ACGT-] (3) A [TG] [GC]

~3600 possible valid sequences

Making a Markov Model

			$\Delta=.4$	$\Delta=.6$	$\Delta=.6$			
	$p(C)=.8$ $p(G)=.2$		$p(A)=.2$ $p(T)=.2$	$p(C)=.4$ $p(G)=.2$		$p(T)=.8$ $p(G)=.2$		
A	C	A	-	-	-	A	T	G
T	C	A	A	C	T	A	T	C
A	C	A	C	-	-	A	G	C
A	G	A	-	-	-	A	T	C
A	C	C	G	-	-	A	T	C
	$p(A)=.8$ $p(T)=.2$	$p(A)=.8$ $p(C)=.2$				$p(A)=1$	$p(C)=.8$ $p(G)=.2$	

Making a Markov Model



$$P(\text{ACAC--ATC}) = 0.8 \times 1.0 \times 0.8 \times 1.0 \times 0.8 \times 1.0 \times 0.6 \times 0.4 \\ \times 0.6 \times 1.0 \times 1.0 \times 0.8 \times 1.0 \times 0.8 = 0.0047$$

Log-Odds (LOD)

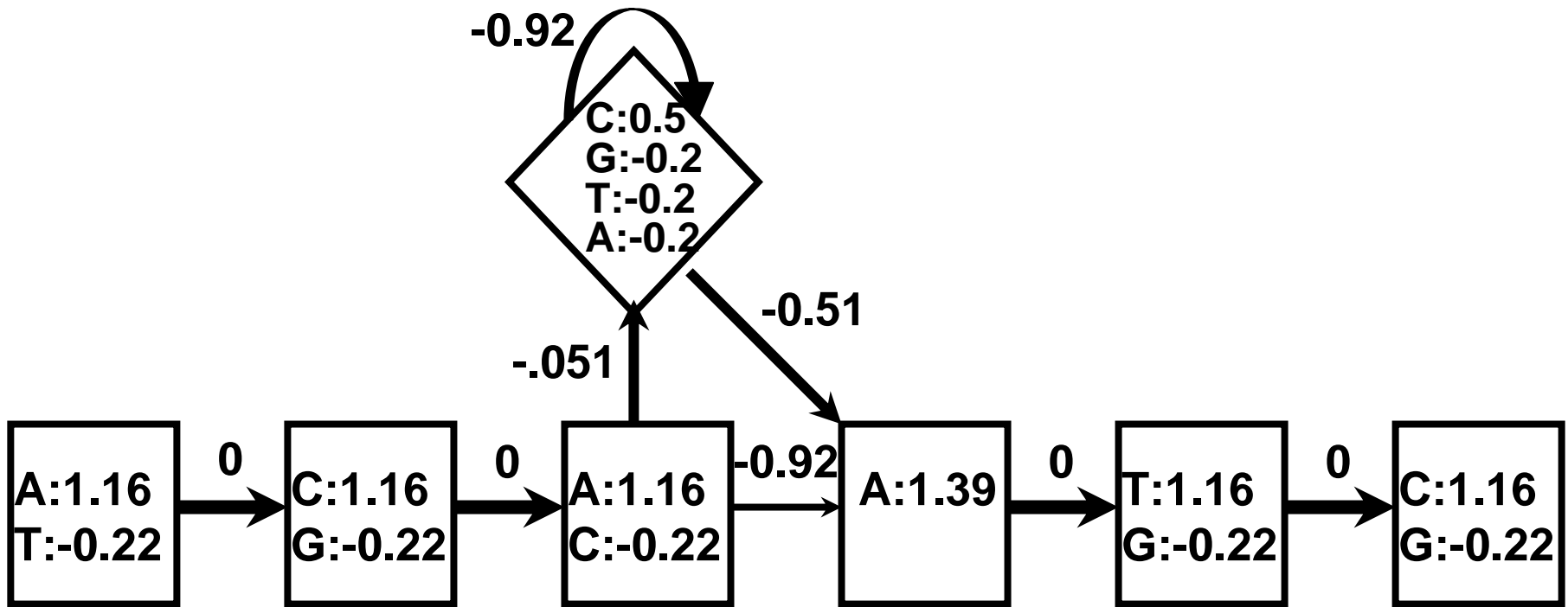
Def'n - LOD is the logarithm of the probability of an event divided by the probability of a null model

For DNA: $\text{LOD}(S) = \log \frac{P(S)}{0.25^L} = \log P(S) - L \log 0.25$

For protein: $\text{LOD}(S) = \log \frac{P(S)}{0.05^L} = \log P(S) - L \log 0.05$

S = sequence, L = length

Making a LOD Markov Model



$$\text{LOD(ACAC--ATC)} = 1.16 + 0 + 1.16 + 0 + 1.16 - 0.51 + 0.5 - 0.51 + 1.39 + 0 + 1.16 + 0 + 1.16 = 6.64$$

Other Sequences...

- $P(\text{ACA---ATG}) = 0.0033$ (LOD = 4.9)
- $P(\text{TCAACTATC}) = 0.000075$ (LOD = 3.0)
- $P(\text{ACAC--AGC}) = 0.0012$ (LOD = 5.3)
- $P(\text{AGA---ATC}) = 0.0033$ (LOD = 4.9)
- $P(\text{ACCG--ATC}) = 0.00059$ (LOD = 4.6)
- $P(\text{TGCT--AGG}) = 0.000023$ (LOD = -0.97) **Worst**
- $P(\text{ACAC--ATG}) = 0.0047$ (LOD = 6.7) **Best**

HMM Issues

- How to find the “optimal sequence” or score a new sequence?
- **Answer: Use Dynamic Programming (called the Viterbi algorithm) to find the optimal path**
- How to deal with sparse data?
- **Answer: Use Pseudocounts (i.e. add fake data that reflects natural substitution patterns or known frequencies)**

HMM Order & Conditional Probability

Order

1st $P(\text{ACTGTC}) = p(\text{A}) \times p(\text{C}) \times p(\text{T}) \times p(\text{G}) \times p(\text{T}) \dots$

2nd $P(\text{ACTGTC}) = p(\text{A}) \times p(\text{C}|\text{A}) \times p(\text{T}|\text{C}) \times p(\text{G}|\text{T}) \dots$

3rd $P(\text{ACTGCG}) = p(\text{A}) \times p(\text{C}|\text{A}) \times p(\text{T}|\text{AC}) \times p(\text{G}|\text{CT}) \dots$



$$P(\text{T}|\text{AC}) = \frac{\#(\text{ACT})}{\#(\text{ACT}) + \#(\text{ACA}) + \#(\text{ACG}) + \#(\text{ACC})}$$

Probability of T given AC

Conclusion

- **Mathematician ---> Bioinformatician**
- **Bioinformatics requires mathematics**
- **Bioinformatics makes use of a wide range of mathematical and statistical techniques**
- **It is important to be aware of these methods and not to be afraid of using them or working with programs that use them**
- **Most of these techniques are not “hard”**