

Chapter 2: Sequence Alignment

2.2 Scoring Matrices

Prof. Yechiam Yemini (YY)
Computer Science Department
Columbia University

COMS4761-2007

Overview

- PAM: scoring based on evolutionary statistics
- Elementary Markovian models of evolution
- BLOSUM: tuning scoring to evolutionary conservation

Recommended brief tutorial on sequence alignment techniques:
<http://www.psc.edu/biomed/training/tutorials/sequence/db/>

COMS4761-2007

2

Introduction To Log-Likelihood Estimation

- Driving question: how likely is an alignment (X,Y)?
 - $X=\dots a \dots \quad Y=\dots b \dots$
 - How likely is evolution to substitute (a,b)?
- Define
 - $P(a)$ = probability of a symbol "a"
 - $P(a,b)$ = probability of evolutionary substitution (a,b)
- Likelihood ratio: $L(a,b)=p(a,b)/p(a)p(b)$
 - Compares evolutionary substitution against a random one
- Log-odds discrimination: $s(a,b)=10^* \log [p(a,b)/p(a)p(b)]$
 - Consider only orders of magnitude discrimination

COMS4761-2007

3

Log-Odds Scoring

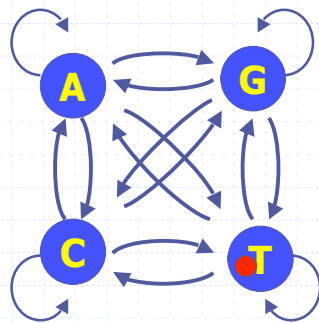
- Scoring Matrix: $s(a,b)=10^* \log [p(a,b)/p(a)p(b)]$
- The score of an alignment (X,Y) is given by:
 - $S(X,Y)=10^* \log [\prod p(X_i, Y_i)/p(X_i)p(Y_i)]=\sum s(X_i, Y_i)$
 - $S(X,Y)$ discriminates between random and evolutionary likely alignment
- DP maximizes the log-odds score of an alignment
- But, how do we acquire the probabilities $p(a), p(a,b)$?

COMS4761-2007

4

Digression: Markovian Models

- How do we model statistical sequence changes?
- Start with a finite state machine (FSM) (regular grammar)
- An FSM path generates unique sequence of changes
- Example: $S(t)$ = state at time t



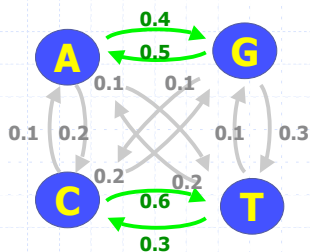
$S(0)=T$
 $S(1)=A$
 $S(2)=G$
 $S(3)=C$
 $S(4)=G$

COMS4761-2007

5

Markov Chain: FSM + Transition Probabilities

- Finite, homogeneous Markov Chain assumptions:
 - Markovian: transition probabilities depend on current state only
 $P[S_{t+1}=j | S_t=i, S_{t-1}=i_{t-1}, \dots, S_0=i_0] = P[S_{t+1}=j | S_t=i]$
 - Homogeneity: transition probabilities are time independent
 $P[S_{t+1}=j | S_t=i]=a(i,j)$
- Notes:
 - Transition matrix is stochastic: $a(i,j) \geq 0$, $\sum_j a(i,j)=1$ ($A \geq 0$, $A\mathbf{1}=\mathbf{1}$)
 - Diagonal elements of A are given by: $a(i,i)=1-\sum_{k \neq i} a(i,k)$



Transition Matrix

	A	G	C	T
A	0.3	0.4	0.2	0.1
G	0.5	0.1	0.1	0.3
C	0.1	0.2	0.1	0.6
T	0.2	0.1	0.3	0.5

COMS4761-2007

6

Computing Evolution Probabilities

- Suppose the initial symbol probability is p
 - $\underline{\pi}^1 = (\pi^1(A), \pi^1(G), \pi^1(C), \pi^1(T))$
- Evolution:
 - After one generation: $\pi^2(y) = \sum_x a(x,y) \pi^1(x) \rightarrow \underline{\pi}^2 = \underline{\pi}^1 \mathbf{A}$
 - After n generations: $\underline{\pi}^{n+1} = \underline{\pi}^1 \mathbf{A}^n$
- The transition matrix for n generations: \mathbf{A}^n

	A	G	C	T
A	0.3	0.4	0.2	0.1
G	0.5	0.1	0.1	0.3
C	0.1	0.2	0.1	0.6
T	0.2	0.1	0.3	0.5

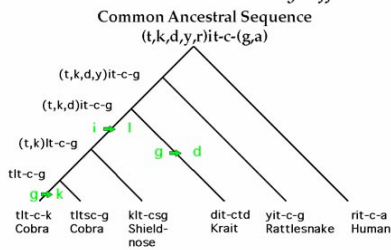
COMS4761-2007

7

Scoring Matrices: PAM [Dayhoff, 1978]

- PAM: Percent Accepted Mutations
- Key idea: estimate probabilities of evolutionary change
 - Model evolution as a Markov chain of substitutions;
 - PAM = unit of evolutionary time $\sim 10^7$ years
 - Evolution performs Markovian substitutions at each PAM
 - Consider protein families with high (99%) sequence similarity
- Dayhoff carefully selected and studied 71 protein families
 - Assure that only one generation of changes is reflected
 - Eliminate noisy estimates

Amino Acid Substitutions: Dayhoff Model



COMS

8

PAM Computations

Data: sequences of a protein family

Build phylogenetic tree

Count frequencies

$f(a,b)$ = frequency of (a,b) substitution;
 $p(a)$ = frequency of "a" [$p(a) = \sum_x f(a,x)$]

$s(a,b) = 10 * \log[f(a,b)/p(a)p(b)]$

COMS4761-2007

9

Scoring Matrices: PAM1 → PAM60 → PAM250

■ Assume evolution is Markovian

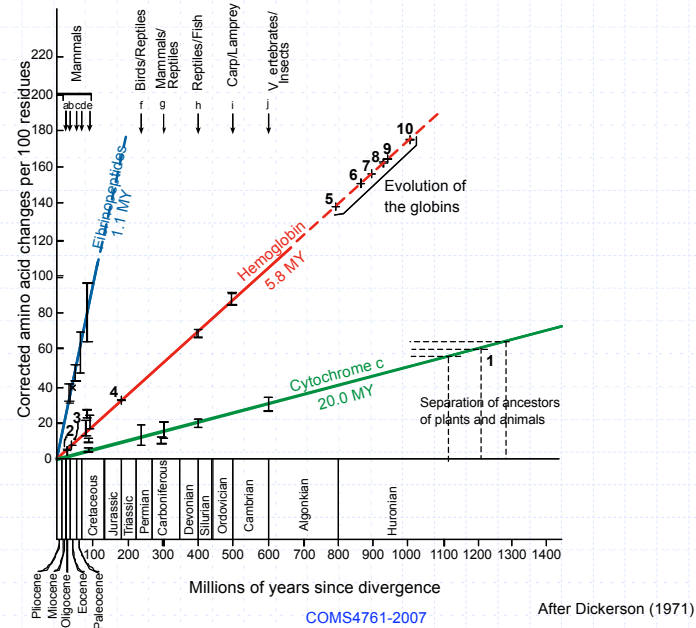
- Define $PAM_n(a,b)$ = probability of changing a to b through n mutations
- n = evolutionary time measured in PAM units
- $PAM_n = (PAM_1)^n$

■ Compute PAM_n for n=1,60,80,120, 250

COMS4761-2007

10

Proteins Evolve At Different Rates (T. Speed Slides)



Proteins Evolve At Different Rates

Protein	^a PAMs/100 residues/10 ⁸ years	Theoretical lookback time ^b
Pseudogenes	400	45 ^c
Fibrinopeptides	90	200 ^c
Lactalbumins	27	670 ^c
Lysozymes	24	750 ^c
Ribonucleases	21	850 ^c
Hemoglobins	12	1.5 ^d
Acid proteases	8	2.3 ^d
Triosephosphate isomerase	3	6 ^d
Phosphoglyceraldehyde dehydrogenase	2	9 ^d
Glutamate dehydrogenase	1	18 ^d

^aPAMs, Accepted point mutations. ^bUseful lookback time = 360 PAMs. ^cMillion years. ^dBillion years.
From Doolittle 1986

COMS4761-2007

14

BLOSUM: Henikoff & Henikoff 1992

- BLOSUM: Block Substitution Matrices
- Motivation: PAM use of matrix power can result in large errors
- Key idea: consider conserved patterns (blocks) of a large sample of proteins
 - Classify protein families (over 500 families)
 - Family has characteristic patterns (signatures) that are conserved
 - Use these conserved patterns to compute substitution probability
- $P(a,b)$ = probability of (a,b) substitution; $P(a)$ = probability of "a"

```

Bpi Bovine  npGivaRItqkgLdyacqggvltlQkele
Bpi Human  npGvvvRIsqkgLdyasqggtaalQkelk
Cept Human  eaGivcRItkpaLlVlnhetakviQtafq
Lbp Human   npGlvaRItDKGLqyaaqegllalQsell
Lbp Rabbit  npGlitRItDKGLeyaaregllalQrkll
    
```

COMS4761-2007

15

Scoring Matrices (e.g., BLOSUM)

- BLOSUM_x=based on patterns that are x% similar
- The level of x% can provide different performance in identifying similarity
- BLOSUM62 provides good scoring (used as default)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*	
A	5	-2	-1	-2	-1	-1	-1	0	-2	-1	-2	-1	-1	-3	-1	1	0	-3	-2	0	-2	-1	-1	-5	
R	2	7	-1	-2	-4	1	0	-3	0	4	3	-2	-3	-3	-1	-1	-3	-1	-3	-1	-3	-1	0	-1	-5
N	-1	-1	7	-2	-2	0	0	0	1	-3	4	0	-2	-4	-2	1	0	-4	-2	-3	4	0	-1	-5	
D	-2	-2	2	8	-4	0	2	-1	-1	-4	-4	-1	-4	-5	-1	0	-1	-5	-3	-4	5	1	-1	-5	
C	-1	-4	-2	-4	13	-3	-3	-3	-3	-2	-2	-3	-2	-2	-4	-1	-1	-5	-3	-1	-3	-3	-2	-5	
Q	-1	1	0	0	-3	7	2	-2	1	-3	-2	2	0	4	-1	0	-1	-1	-1	-3	0	4	-1	-5	
E	-1	0	0	2	-3	2	6	-3	0	4	-3	1	-2	-3	-1	-1	-1	-3	-2	-3	1	5	-1	-5	
G	0	-3	0	-1	-3	-2	-3	8	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4	-1	-2	-2	-5	
H	-2	0	-1	-1	-3	1	0	-2	10	-4	-3	0	-1	-1	-2	-1	-2	-3	-2	-4	0	0	-1	-5	
I	-1	-4	-3	-4	-2	-3	-4	-4	4	5	2	-3	2	0	-3	-3	-1	-3	-1	4	4	-3	-1	-5	
L	-2	-3	-4	-4	-2	-3	-4	-3	2	5	-3	3	1	-4	-3	-1	-2	-1	-4	-3	-1	-4	-3	-5	
K	-1	3	0	-1	-3	2	1	-2	0	-3	-3	6	-2	-4	-1	0	-1	-3	-2	-3	0	1	-1	-5	
M	-1	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	7	0	-3	-2	-1	-1	0	1	-3	-1	-1	-5	
F	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	1	4	0	8	-4	-3	-2	1	4	-1	4	-4	-2	-5	
P	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	10	-1	-1	-4	-3	-3	-2	-1	-2	-5	
S	1	-1	1	0	-1	0	-1	0	-1	0	-1	-3	0	-2	-3	-1	5	2	-4	-2	0	0	-1	-5	
T	0	-1	0	-1	-1	-1	-2	-2	-1	-1	-1	-2	-1	2	5	-3	-2	0	0	-1	0	-1	0	-5	
W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-2	-3	-1	1	-4	-4	-3	15	-2	-3	-5	-2	-3	-5		
Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-2	0	4	-3	-2	-2	2	8	-1	-3	-2	-1	-5		
V	0	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	0	-3	-1	5	-4	-3	-1		
B	-2	-1	4	5	-3	0	1	-1	0	-4	4	0	-3	-4	-2	0	0	-5	-3	-4	5	2	-1		
Z	-1	0	0	1	-3	4	5	-2	0	-3	-3	-1	-4	-1	0	-1	-2	-2	-3	2	5	-1	-5		
X	-1	-1	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-1	0	-3	-1	-1	-1	-1	-1	-5		
*	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	1		

COMS4761-2007

16