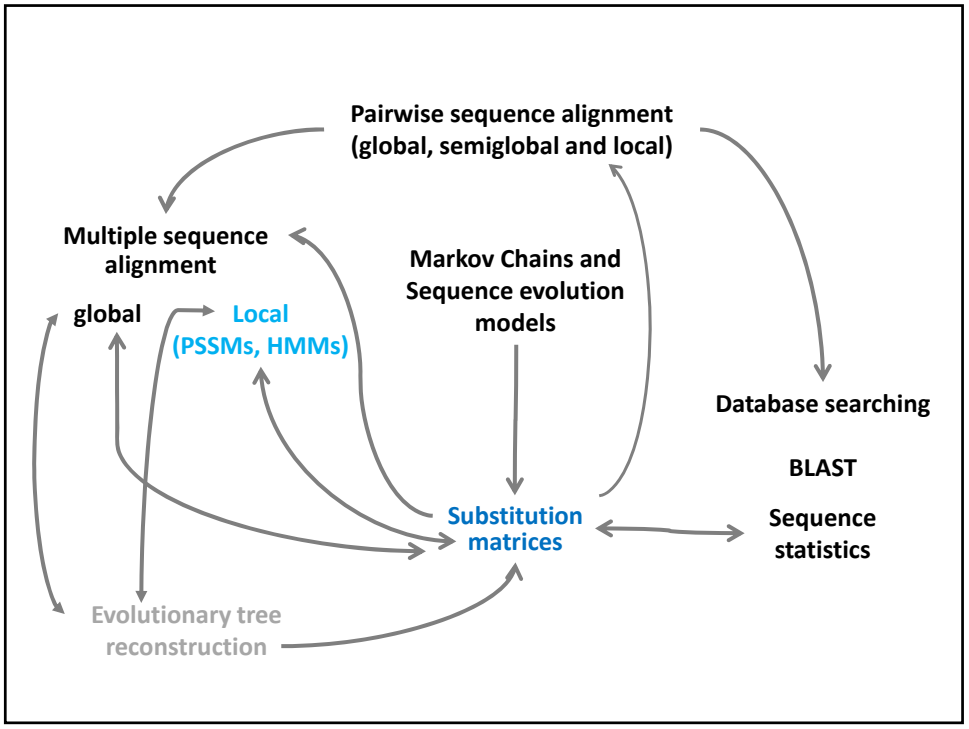


Scoring an alignment with a substitution matrix

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							
G							
H							
I							
L							
K							
M							
F							
P							
S							
T							
W							
Y							
V							

```
fly       KVINDNFEIV EGLMTTVHAT
human     KVIHDNFGIV EGLMTTVHAI
          5441666244 5645554841
```

(The blue numbers are negative.)

$$S^N[x,y] = c \log \frac{q^N_{xy}}{p_x p_y}$$

Frequency of x aligned with y in sequences with divergence *N*

Frequency of x aligned with y in "random" sequences

Scaling constant

			11
		2	7
	-3	-1	4
W	Y	V	

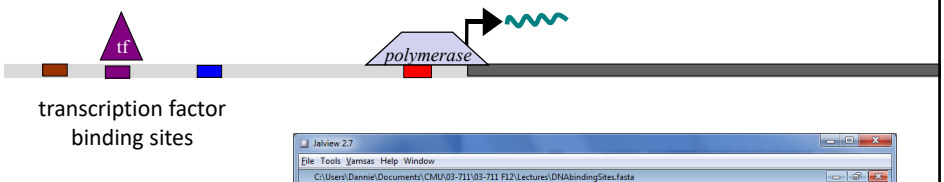
Blosum62 scoring matrix

Position Specific Scoring Matrices (PSSMs)

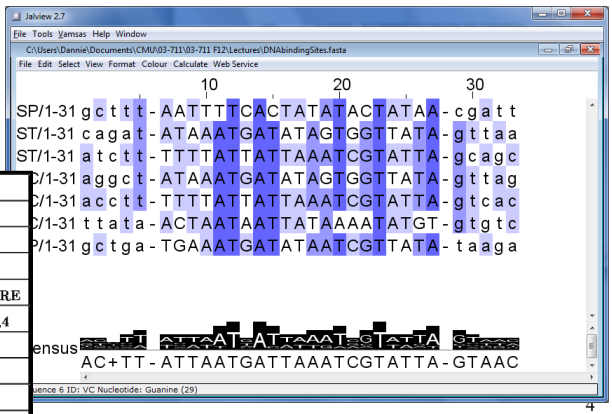
compare a sequence with a known pattern

Conserved patterns in biological sequences

Example: Transcription factor binding sites



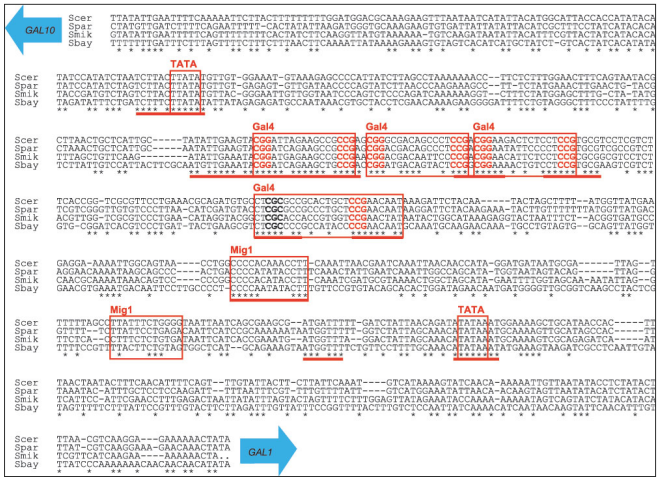
transcription factor binding sites



	Rap1
	Rpn4
	Gcn4
	HSE
	Mig1/STRE
	Hap2,3,4
	Cbf1
	MCB
	Lys14
	Leu3

Conserved patterns in biological sequences

Example: Transcription factor binding sites



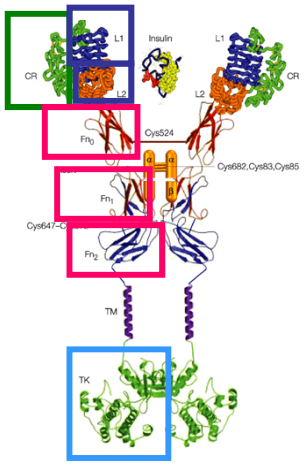
5

Kellis *et al, Nature*, 03

Conserved patterns in biological sequences

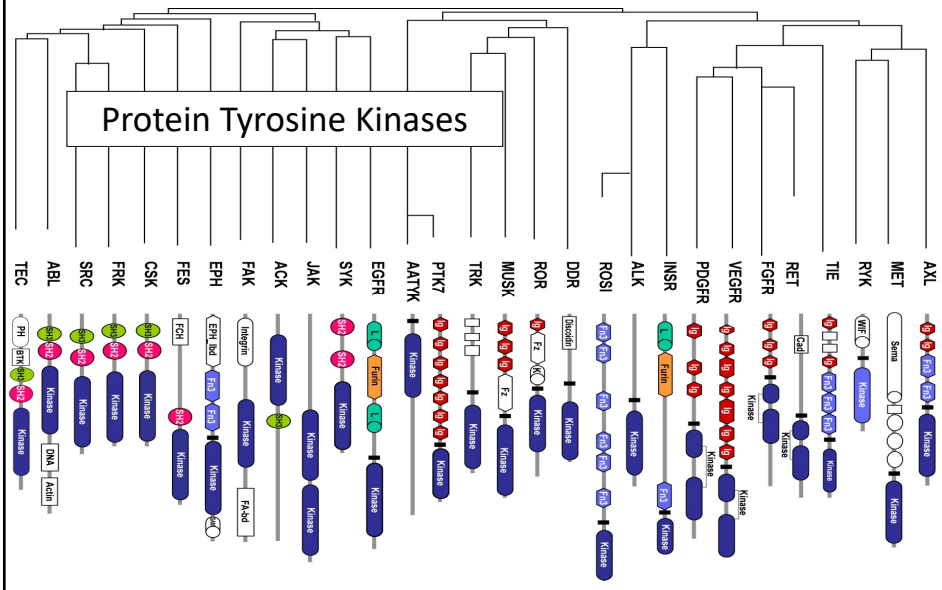
Example: Protein domains

- Fold independently
- Carry out specific functions
- Found in diverse contexts
- Conserved in evolution



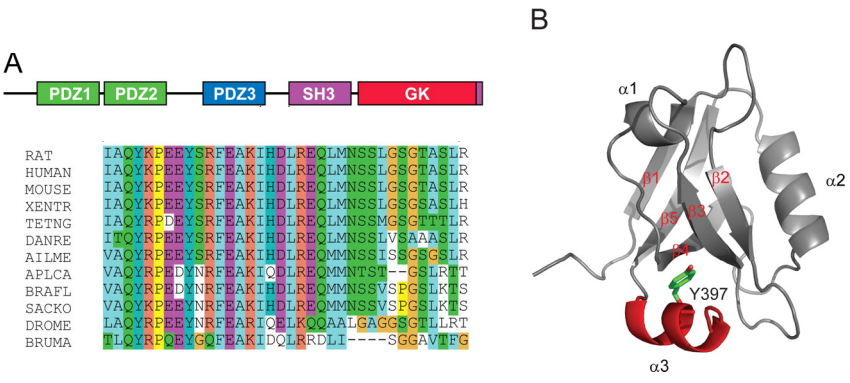
Insulin receptor

6



Adapted from Robinson *et al*., 2000

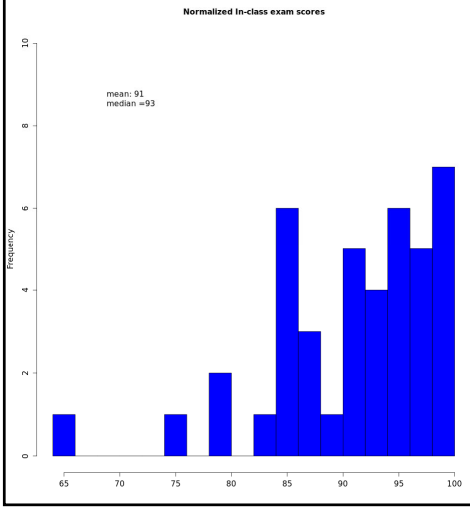
Domain architecture of PSD-95 and crystal structure of PDZ3.

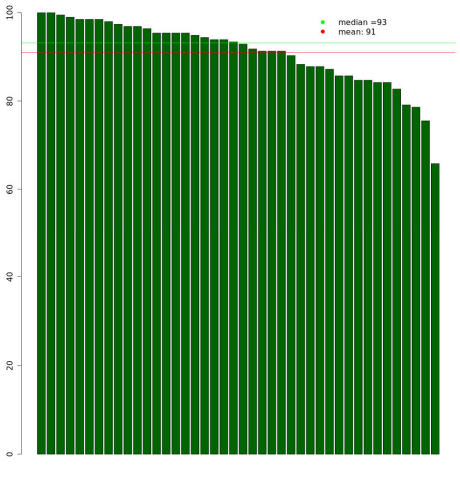


Zhang J et al. J. Biol. Chem. 2011;286:41776-41785

©2011 by American Society for Biochemistry and Molecular Biology

Note: The maximum possible score on this exam was 98. In the plots presented here, the scores have been rescaled to have a maximum of 100 points.





Problem 4(a)

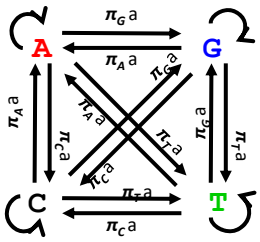
	A	G	C	T
A	0.59	0.16	0.14	0.11
G	0.09	0.66	0.14	0.11
C	0.09	0.16	0.64	0.11
T	0.09	0.16	0.14	0.61

This model has 4 free parameters:

- Rate: a
- Stationary distribution:
  - $p(A)=\pi_A$
  - $p(G)=\pi_G$
  - $p(C)=\pi_C$

Note that $p(T)=1-(\pi_A+\pi_G+\pi_C)$, so only 3 parameters are required for the stationary distribution.

Felsenstein (1981)



$p(A) \neq p(G) \neq p(C) \neq p(T)$

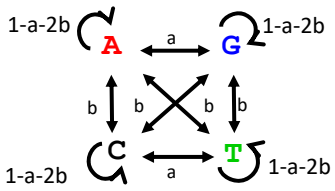
$p(A)=\pi_A$
$p(G)=\pi_G$
$p(C)=\pi_C$
$p(T)=\pi_T$

10

Problem 4(b)

Kimura 2 parameter model (K2P) (1980)

	A	G	C	T
A	0.6	0.2	0.1	0.1
G	0.2	0.6	0.1	0.1
C	0.1	0.1	0.6	0.2
T	0.1	0.1	0.2	0.6



p(A)=0.25
p(G)=0.25
p(C)=0.25
p(T)=0.25

This model has 2 free parameters:

- Rates: a, b

11

Problem 4(c)

	A	G	C	T
A	0.3	0.1	0.2	0.2
G	0.1	0.3	0.2	0.2
C	0.2	0.2	0.3	0.1
T	0.2	0.2	0.1	0.3

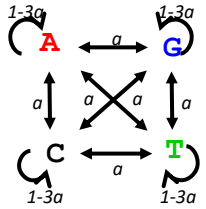
This is not a Markov chain transition probability matrix.
The rows do not sum to 1.

12

Problem 4(d)

Jukes-Cantor model (1969)

	A	G	C	T
A	0.55	0.15	0.15	0.15
G	0.15	0.55	0.15	0.15
C	0.15	0.15	0.55	0.15
T	0.15	0.15	0.15	0.55



*p(**A**)=0.25*
*p(**G**)=0.25*
*p(**C**)=0.25*
*p(**T**)=0.25*

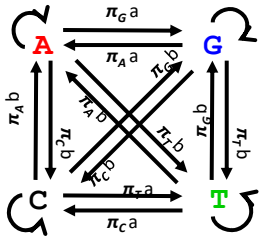
This model has 1 free parameter:
- Rates: a

13

Hasegawa, Kishino & Yano (HKY) (1985)

This model has 5 free parameters:
- Rates: a, b
- Stationary distribution:
  - *p(**A**)=π_A*
  - *p(**G**)=π_G*
  - *p(**C**)=π_C*

*Note that p(**T**)=**1- (π_A +π_G+π_C),** so only 3 parameters are required for the stationary distribution.*



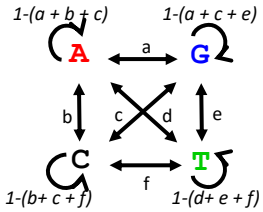
*p(**A**)≠p(**G**) ≠ p(**C**) ≠ p(**T**)*

14

General Time Reversible model

This model has 9 free parameters:

- Rates: a, b, c, d, e, f
- Stationary distribution:
  - $p(A)=\pi_A$
  - $p(G)=\pi_G$
  - $p(C)=\pi_C$



Note that $p(T)=1-(\pi_A+\pi_G+\pi_C)$, so only 3 parameters are required for the stationary distribution.

$p(A)\neq p(G) \neq p(C) \neq p(T)$

15