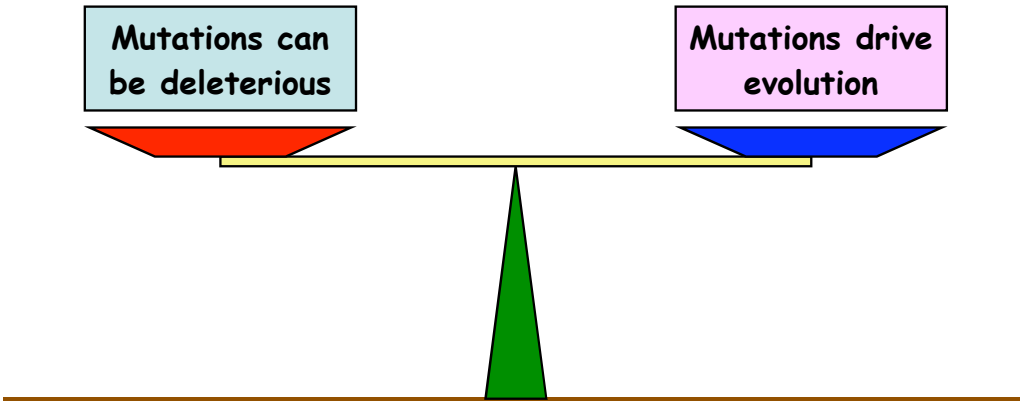
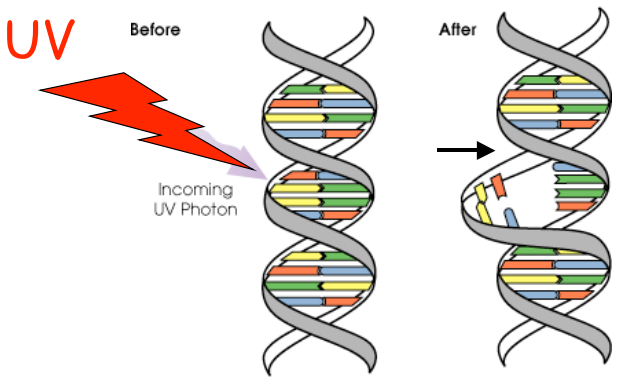
Mutations, the molecular clock, and models of sequence evolution

Why are mutations important?



Replicative proofreading and DNA repair constrain mutation rate

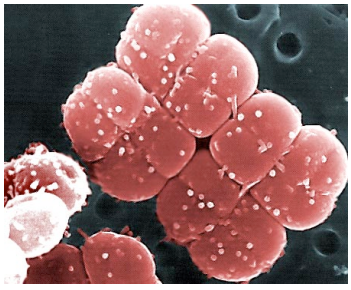
UV damage to DNA

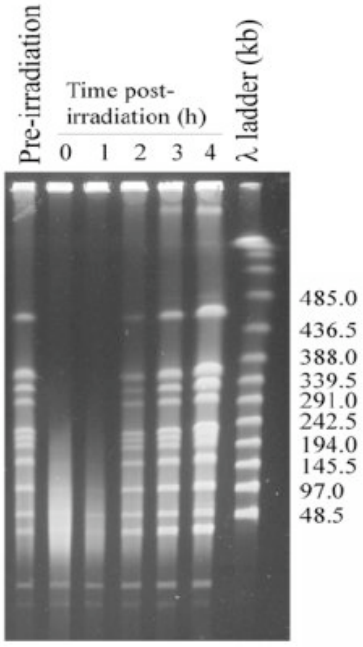


Thymine dimers

What happens if damage is not repaired?

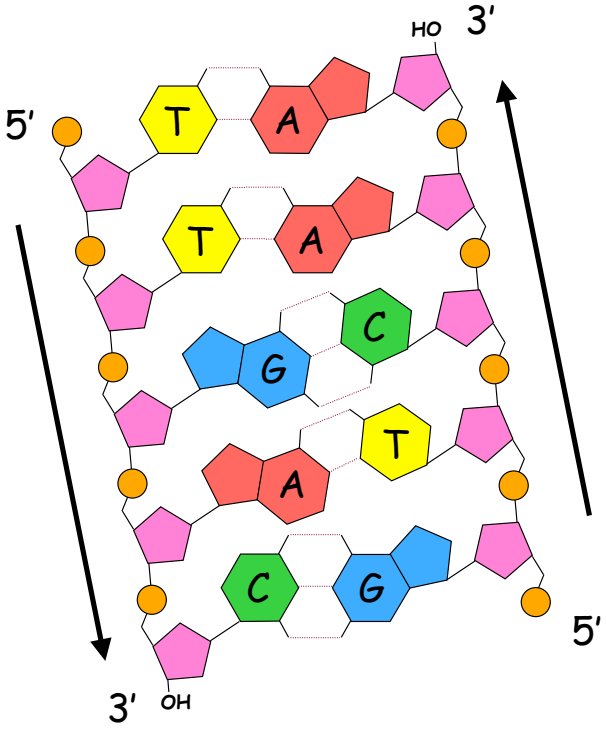
Deinococcus radiodurans is amazingly resistant to ionizing radiation





- 10 Gray will kill a human
- 60 Gray will kill an *E. coli* culture
- *Deinococcus* can survive 5000 Gray

DNA Structure



Information polarity
Strands complementary

G-C: 3 hydrogen bonds
A-T: 2 hydrogen bonds

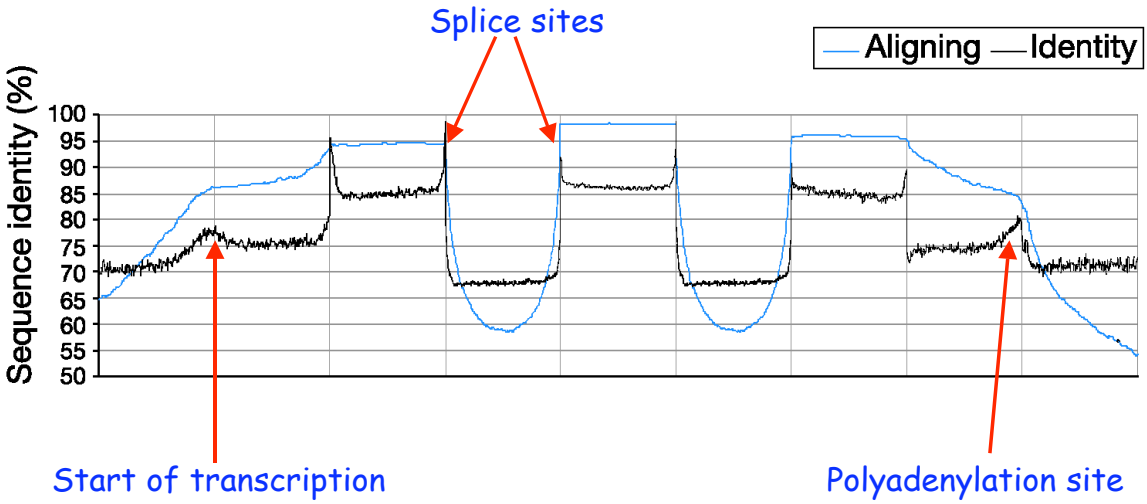
Two base types:
- Purines (A, G)
- Pyrimidines (T, C)

Not all base substitutions are created equal

- Transitions
  - Purine to purine (A → G or G → A)
  - Pyrimidine to pyrimidine (C → T or T → C)
- Transversions
  - Purine to pyrimidine (A → C or T; G → C or T)
  - Pyrimidine to purine (C → A or G; T → A or G)

Transition rate ~2x transversion rate

Substitution rates differ across genomes



Alignment of 3,165 human-mouse pairs

Mutations vs. Substitutions

- Mutations are changes in DNA
- Substitutions are mutations that evolution has tolerated

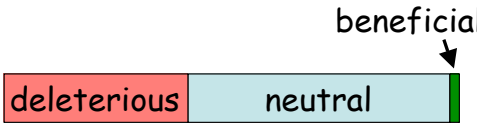
Which rate is greater?

How are mutations inherited?

Are all mutations bad?

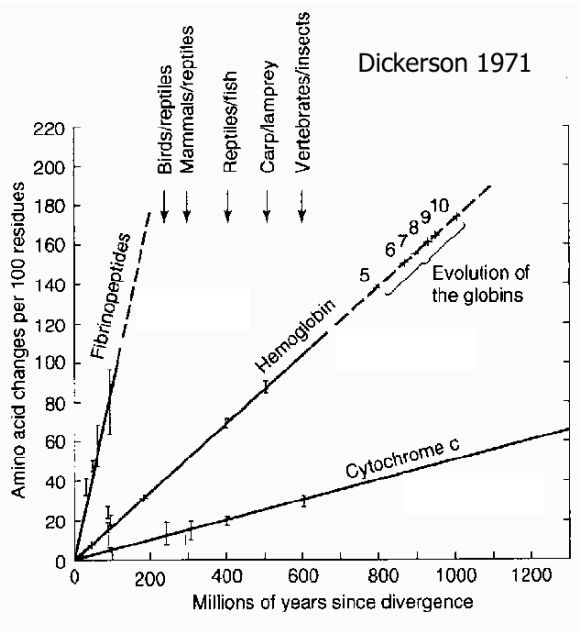
Selectionist vs. Neutralist Positions

- Most mutations are deleterious; removed via negative selection
- Advantageous mutations positively selected
- Variability arises via selection



- Some mutations are deleterious, many mutations neutral
- Neutral alleles do not alter fitness
- Most variability arises from genetic drift

What is the rate of mutations?



Rate of substitution constant: implies that there is a molecular clock

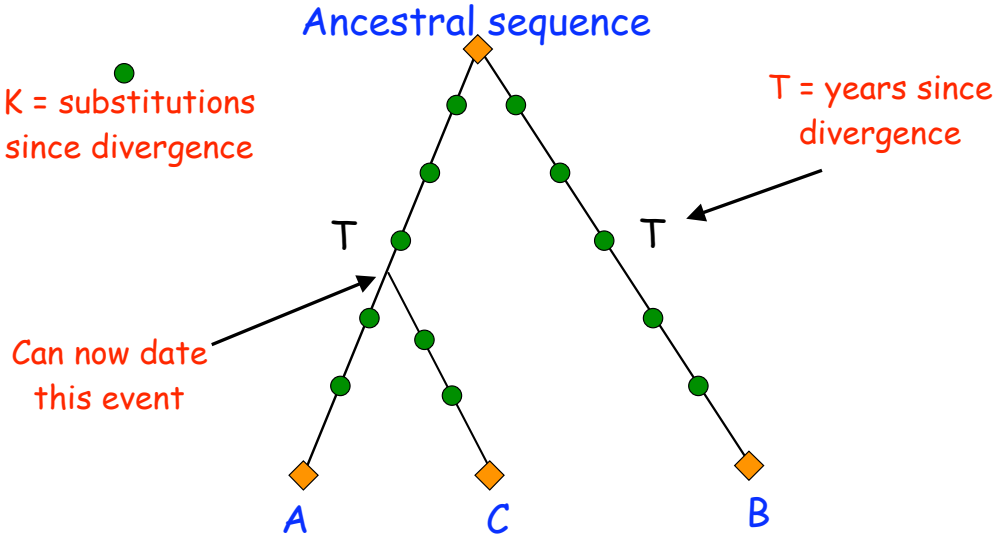
Rates proportional to amount of functionally constrained sequence

Why care about a molecular clock?

(1) The clock has important implications for our understanding of the mechanisms of molecular evolution.

(2) The clock can help establish a time scale for evolution.

Dating evolutionary events with a molecular clock



sub. rate = K/2T

What are the assumptions?

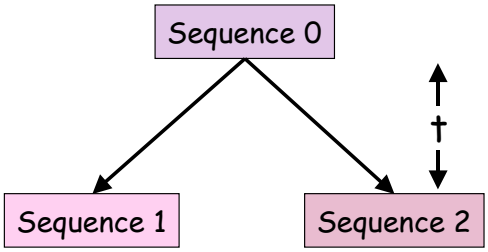
Properties of the molecular clock

- Clock is erratic
- Clock calibrations require geological times
- Many caveats - varying generation times, different mutation rates, changes in gene function, natural selection
- Is the molecular clock hypothesis even useful at all?

Measuring sequence divergence: Why do we care?

- Use in sequence alignments and homology searches of databases
- Inferring phylogenetic relationships
- Dating divergence, correlating with fossil record

How do you measure how different two homologous DNA sequences are?

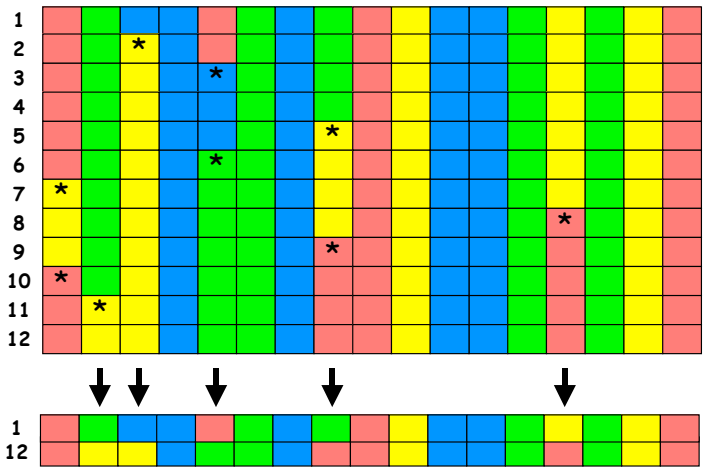



```
Seq1  ACCATGGAATTTTATACCCT
      | |  | | | |  | | |  | | |  | | | |
Seq2  ACTATGGGATTGTATCCCCT
```

p distance = # differences / aligned length

p distance = 4/20 = 0.2

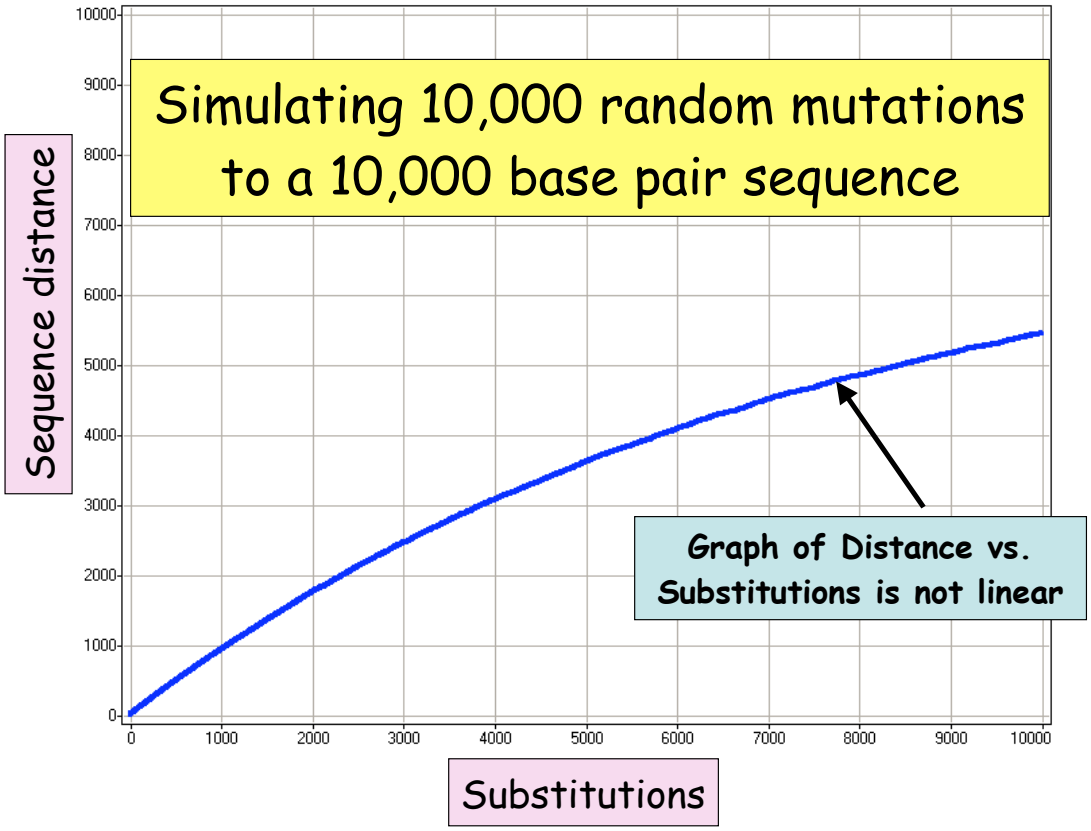
A sequence mutating at random



9 substitutions

5 pairwise changes

Multiple substitutions at one site can cause underestimation of number of substitutions



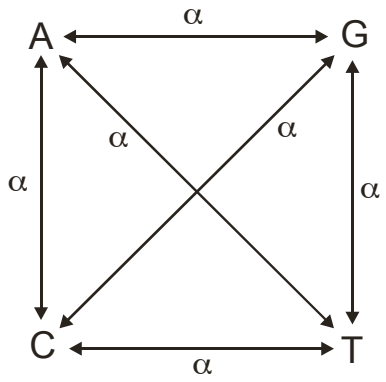
Wouldn't it be great to be able to correct for multiple substitutions?

True # subs (K) = CF x p distance

What probabilities does this correction factor need to consider?

What is a model of nucleotide sequence evolution?

Theoretical expression of nucleotide composition and likelihood of each possible base substitution



Base frequencies equal, all substitutions equally likely

Jukes Cantor Correction

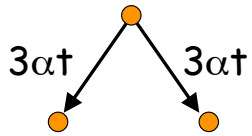
Step 1 - Define rate matrix

$$Q = \begin{pmatrix} & [A] & [C] & [G] & [T] \\ [A] & - & \alpha & \alpha & \alpha \\ [C] & \alpha & - & \alpha & \alpha \\ [G] & \alpha & \alpha & - & \alpha \\ [T] & \alpha & \alpha & \alpha & - \end{pmatrix}$$

"instantaneous rate matrix"
Q = rate of substitution per site

- For any nt, # subs/time = 3α
- In time t, there will be $3\alpha t$ subs
- **Wait! We don't know α or t !...**

…But we do know relationship between K, α, and t



subs = K = 2(3αt)

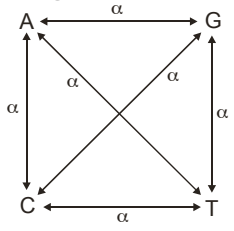
K = Correction factor x p distance

Can we express p distance in terms of α and t ?

Jukes Cantor Correction

Step 2 - Derive $P_{nt(t+1)}$ in terms of $P_{nt(t)}$ and α

(Rate of change to another nt = α)



$P_{A(0)} = 1$

$P_{A(1)} = P_{A(0)} - 3\alpha = 1 - 3\alpha$

$P_{A(2)} = (1-3\alpha)\, P_{A(1)} + (1-P_{A(1)})\, \alpha$

= prob. of staying A x prob. stayed A 1st time + prob. A changed first time x prob. reverted to A

$$P_{A(t+1)} = (1-3\alpha)\, P_{A(t)} + (1-P_{A(t)})\, \alpha$$

Jukes Cantor Correction

Step 3 - Derive probabilities of nt staying same or changing for time t

$$P_{A(t+1)} = (1-3\alpha)\, P_{A(t)} + (1-P_{A(t)})\, \alpha$$

Probability nt stays same → $P_{ii(t)} = 1/4 + 3/4e^{-4\alpha t}$

Probability nt changes → $P_{ij(t)} = 1/4 - 1/4e^{-4\alpha t}$

Jukes Cantor Correction

Step 4 - compute probability that two homologous sequences differ at a given position

p = 1 – prob. that they are identical

p = 1 – (prob. of both staying the same + prob. of both changing to the same thing)

$$p = 1 - \{ (P_{AA(t)})^2 + (P_{AT(t)})^2 + (P_{AC(t)})^2 + (P_{AG(t)})^2 \}$$

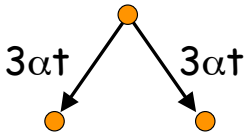
$$p = 3/4(1- e^{-8\alpha t})$$

Jukes Cantor Correction

Step 5 - calculate number of subs in terms of proportion of sites that differ

$p = 3/4(1- e^{-8\alpha t})$

$8\alpha t = -\ln(1- 4/3p)$



Number subs = $K = 2(3\alpha t)$

$K = -3/4 \ln(1-4/3p)$

For p=0.25, K=0.304

K = Correction factor x p distance

Do we need a more complex nucleotide substitution model ?

- Different nucleotide frequencies
- Different transition vs. transversion rates
- Different substitution rates
- Different rates of change among nt positions
- Position-specific changes within codons
- Various curve fitting corrections

What about substitutions between protein sequences?

- Model of DNA sequence evolution: 4x4 matrix
- What size matrix needed for all amino acids?

- p distance = # differences / length

- Theoretical correction for single rate of amino acid change: $K = -19/20 \ln(1-20/19p)$****

But it's more complicated to model protein sequence evolution

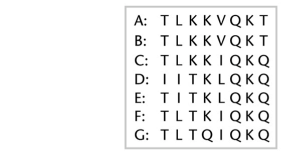
- Substitution paths between amino acids not a uniform length

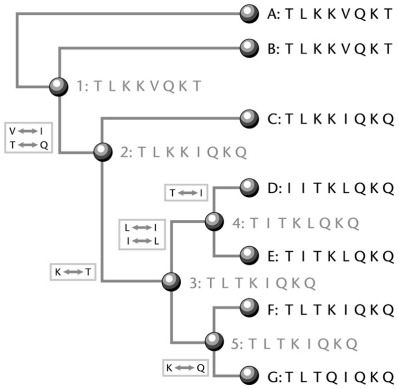
- Amino acid changes have unpredictable effects on protein function

- Solution: use empirical data on amino acid substitutions

The PAM model of protein sequence evolution

- Empirical data-based substitution matrix
- Global alignments of 71 families of closely related proteins.
- Constructed hypothetical evolutionary trees
- Built matrix of 1572 a.a. point accepted mutations





Original PAM substitution matrix

Original amino acid j

Replacement amino acid i

	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A																				
R	30																			
N	109	17																		
D	154	0	532																	
C	33	10	0	0																
Q	93	120	50	76	0															
E	266	0	94	831	0	422														
G	579	10	156	162	10	30	112													
H	21	103	226	43	10	243	23	10												
I	66	30	36	13	17	8	35	0	3											
L	95	17	37	0	0	75	15	17	40	253										
K	57	477	322	85	0	147	104	60	23	43	39									
M	29	17	0	0	0	20	7	7	0	57	207	90								
F	20	7	7	0	0	0	0	17	20	90	167	0	17							
P	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
S	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
T	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
W	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
Y	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
V	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	
	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val

Dayhoff, 1978

Count number of times residue b was replaced with residue a

$= A_{i,j}$

Deriving PAM matrices

For each amino acid, calculate relative mutabilities:

$$m_j = \frac{\text{\# times a.a. j mutated}}{\text{total occurrences of a.a.}}$$

Likelihood a.a. will mutate

TABLE 3-1 Relative Mutabilities of Amino Acids

The value of alanine is arbitrarily set to 100.

Asn	134	His	66
Ser	120	Arg	65
Asp	106	Lys	56
Glu	102	Pro	56
Ala	100	Gly	49
Thr	97	Tyr	41
Ile	96	Phe	41
Met	94	Leu	40
Gln	93	Cys	20
Val	74	Trp	18

Source: From Dayhoff (1978). Used with permission.

Deriving PAM matrices

Calculate mutation probabilities for each possible substitution

$M_{i,j}$ = relative mutability x
proportion of all subs of j represented by change to i

$$M_{i,j} = \frac{m_j \times A_{i,j}}{\sum_i A_{i,j}}$$

$M_{j,j}$ = 1- m_j = probability of j staying same

PAM1 mutation probability matrix

Original amino acid **j**

Replacement amino acid **i**

	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I	2	2	3	1	2	1	2	0	0	9872	9	2	21	7	0	1	7	0	1	33
L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

Dayhoff, 1978

Probabilities normalized to 1 a.a. change per 100 residues

Deriving PAM matrices

Calculate log odds ratio to convert mutation probability to substitution score

$$S_{i,j} = 10 \times \log_{10}\left(\frac{(M_{i,j})}{f_i}\right)$$

Mutation probability
(Prob. substitution from j to i is an accepted mutation)

Frequency of residue i
(Probability of a.a. i occurring by chance)

Deriving PAM matrices

Scoring in log odds ratio:

-Allows addition of scores for residues in alignments

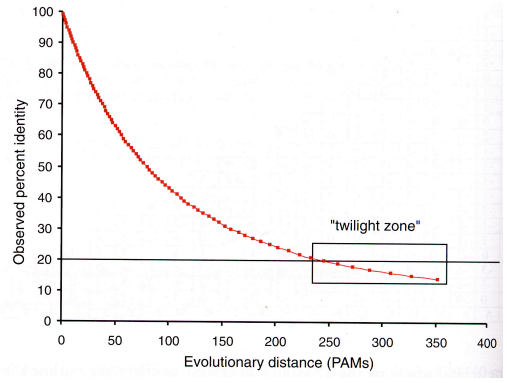
Interpretation of score:

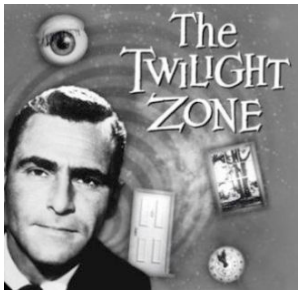
- Positive: non-random (accepted mutation) favored
- Negative: random model favored

Using PAM scoring matrices

PAM1 - 1% difference (99% identity)

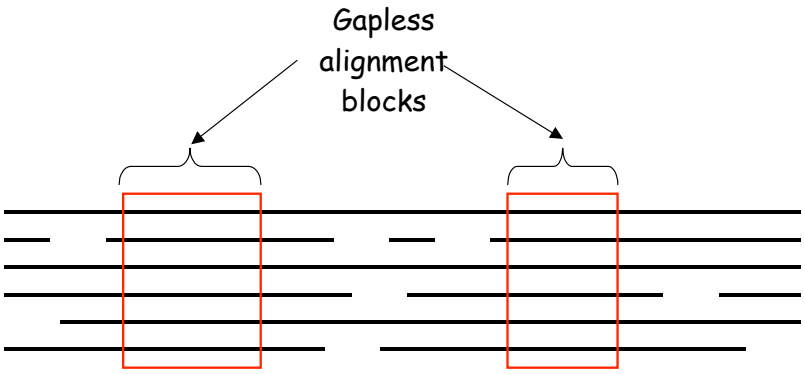
Can "evolve" the mutation probability matrix by multiplying it by itself, then take log odds ratio (PAMn = PAM matrix multiplied n times)





BLOSUM = BLOCKS substitution matrix

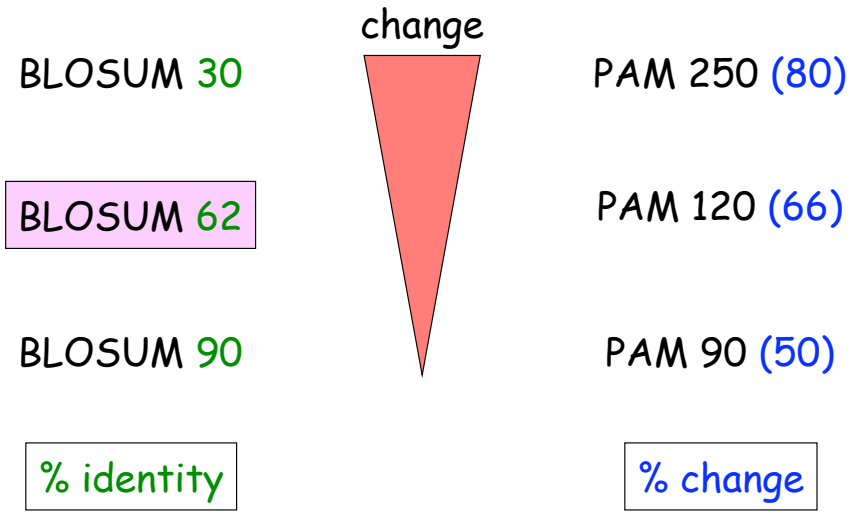
- Like PAM, empirical proteins substitution matrices, use log odds ratio to calculate sub. scores
- Large database: local alignments of conserved regions of distantly related proteins



BLOSUM uses clustering to reduce sequence bias

- Cluster the most similar sequences together
- Reduce weight of contribution of clustered sequences
- BLOSUM number refers to clustering threshold used (e.g. 62% for BLOSUM 62 matrix)

BLOSUM and PAM substitution matrices



BLAST algorithm uses BLOSUM 62 matrix

PAM

- Smaller set of closely related proteins - short evolutionary period
- Use global alignment
- More divergent matrices extrapolated
- Errors arise from extrapolation

BLOSUM

- Larger set of more divergent proteins-longer evolutionary period
- Use local alignment
- Each matrix calculated separately
- Clustering to avoid bias
- Errors arise from alignment errors

Importance of scoring matrices

- Scoring matrices appear in all analysis involving sequence comparison.

- The choice of matrix can strongly influence the outcome of the analysis.