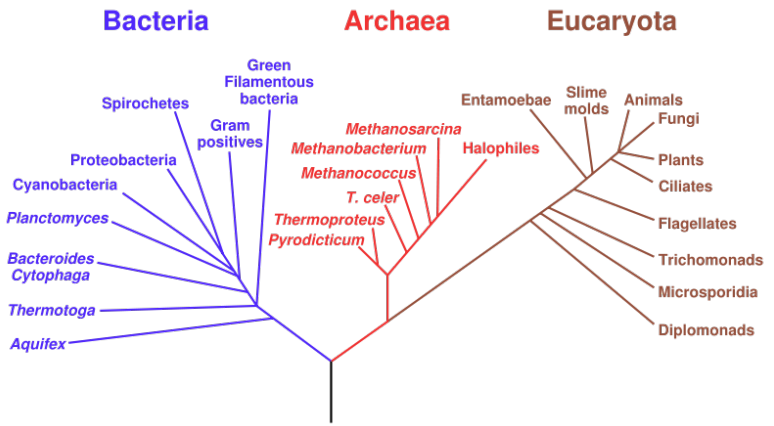
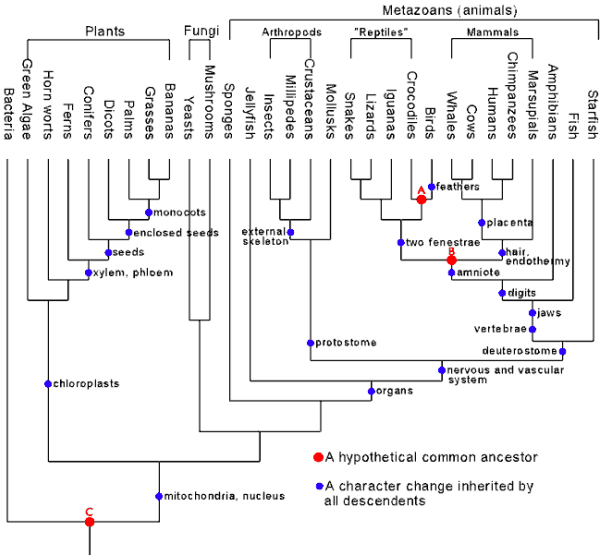
Mutations

Tree Of Life



Tree Of Life



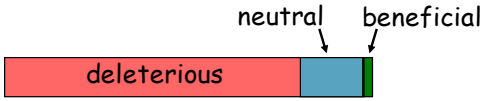
Mutations vs. Substitutions

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

Which rate is greater?

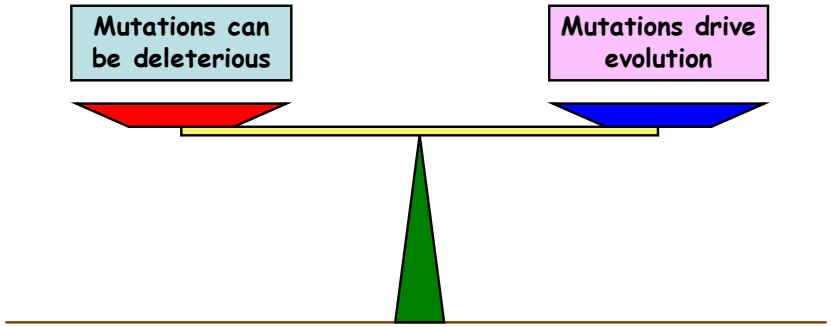
Replicative proofreading and DNA repair constrain the mutation rate

Selectionist Evolution

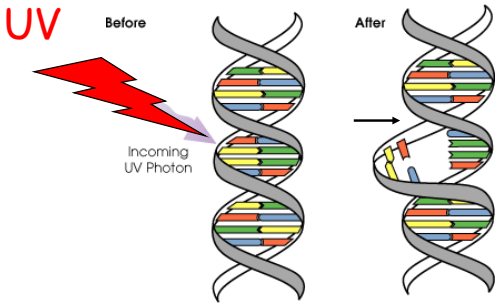


- Most mutations are deleterious; removed via negative selection
- Advantageous mutations positively selected
- Variability arises via selection

Why Are Mutations Important?



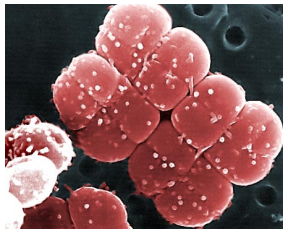
UV Damage to DNA



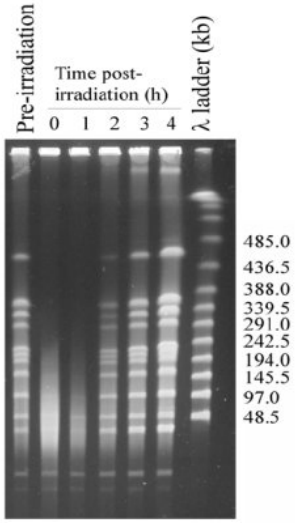
Thymine dimers

What happens if damage is not repaired?

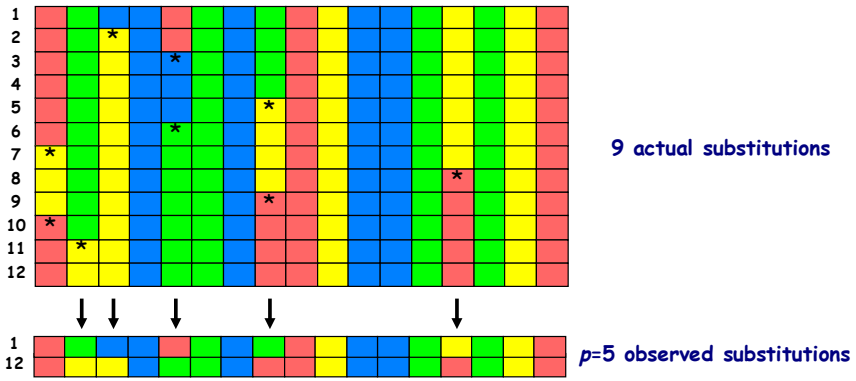
Radiation Resistant



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

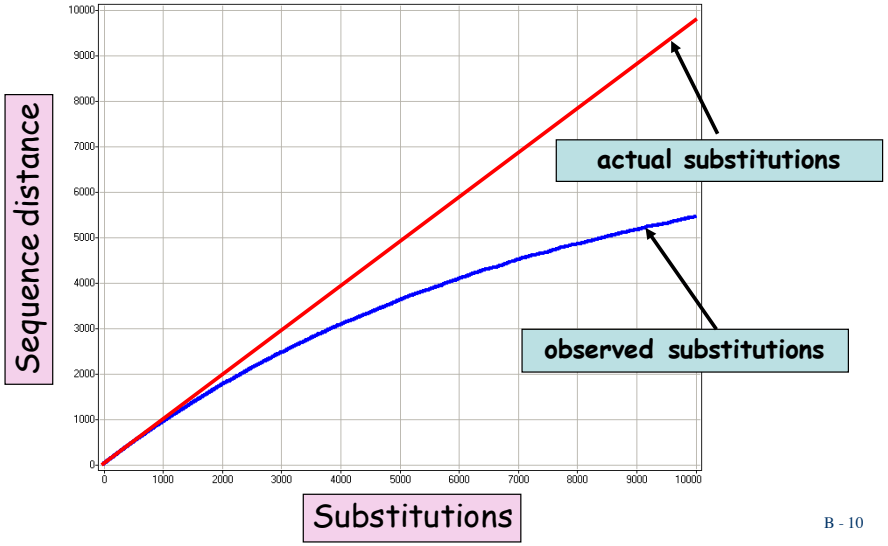


A Sequence Mutating at Random



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

Simulating Random Mutations

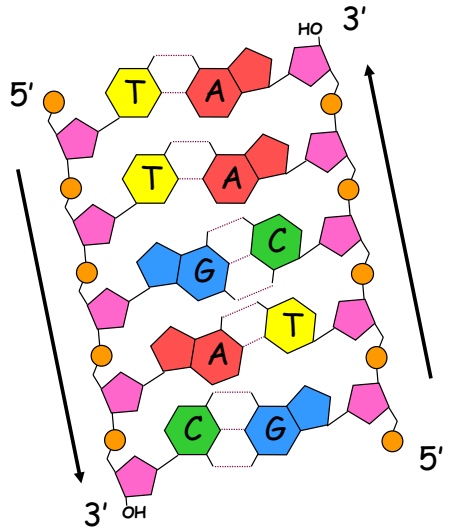


Measuring Sequence Divergence: Why Do We Care?

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

* Comparative genomics is an important field. Determining not only how many substitutions exist between two sequences but how similar two sequences are.

DNA Structure



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

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

Not All Base Substitions 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*)

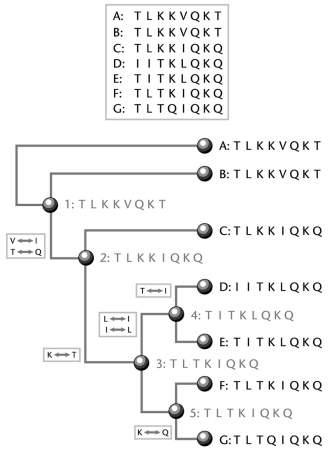
Transition rate ~2x transversion rate

Substition Rates Differ Across Genomes

Alignment of 3,165 human-mouse pairs

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 amino acid point accepted mutations



Original PAM Substitution Matrix

Original amino acid

Replacement amino acid	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 *i* was replaced with residue *j*

Deriving PAM Matrices

For each amino acid, calculate its *relative mutability*, i.e., the likelihood that the amino acid will mutate:

$$m_j = \frac{\text{\# times amino acid } j \text{ mutated}}{\text{total occurrences of amino acid } j}$$

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 × proportion of all substitutions to j by changing to i

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

PAM1 Mutation Probability Matrix

Original amino acid

Replacement amino acid	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

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 amino acid 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

PAM Matrix

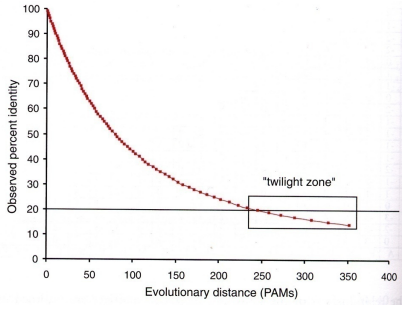
		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	Cys	12																			
S	Ser	0	2																		
T	Thr	-2	1	3																	
P	Pro	-3	1	0	6																
A	Ala	-2	1	1	1	2															
G	Gly	-3	1	0	-1	1	5														
N	Asn	-4	1	0	-1	0	0	2													
D	Asp	-5	0	0	-1	0	1	2	4												
E	Glu	-5	0	0	-1	0	0	1	3	4											
Q	Gln	-5	-1	-1	0	0	-1	1	2	2	4										
H	His	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R	Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K	Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M	Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I	Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L	Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V	Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F	Phe	-4	-3	-3	-5	-5	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y	Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W	Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17

Using PAM Scoring Matrices

PAM1: 1% difference (99% identity)

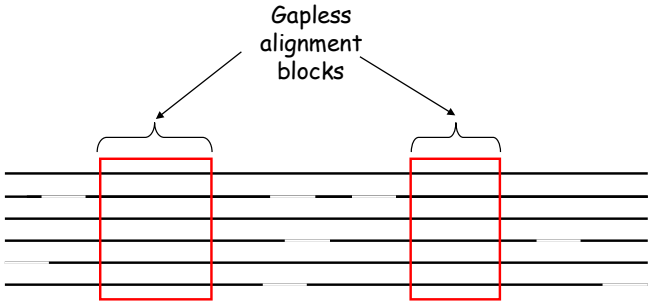
Can "evolve" the mutation probability matrix by multiplying it by itself, then take log odds ratio

(PAM*n* = PAM matrix multiplied by itself *n* times)



BLOSUM = BLOCKS Substitution Matrix

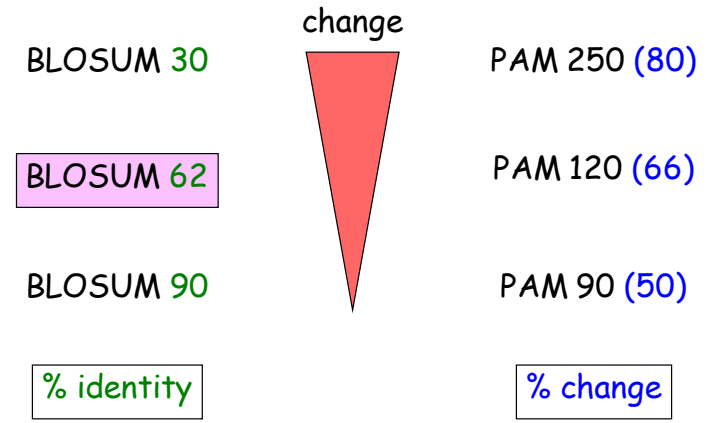
- Like PAM, empirical proteins substitution matrices, use log odds ratio to calculate substitution scores
- Large database: local alignments of conserved regions of distantly related proteins



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 and BLOSUM

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 analyses involving sequence comparison
- The choice of matrix can strongly influence the outcome of the analysis