

Bioinformatics 1: lecture 5

Follow-up of Lecture 4?

Substitution matrices

Multiple sequence alignment

A teacher's dilemma

<u>To understand...</u>	<u>You first need to know...</u>
Multiple sequence alignment	Substitution matrices
Substitution matrices	Phylogenetic trees
Phylogenetic trees	Multiple sequence alignment

Substitution matrices

- Used to score aligned positions, usually of amino acids.
 - Expressed as the *log-likelihood ratio of mutation* (or *log-odds ratio*)
 - Derived from multiple sequence alignments
-

Two commonly used matrices: PAM and BLOSUM

- PAM = **percent accepted mutations** (Dayhoff)
- BLOSUM = **Blocks substitution matrix** (Henikoff)

PAM

M Dayhoff, 1978

- Evolutionary time is measured in Percent Accepted Mutations, or PAMs

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	12																				C
S	0	2																			S
T	-2	1	3																		T
P	-3	1	0	6																	P
A	-2	1	1	1	2																A
G	-3	1	0	-1	1	5															G
N	-4	1	0	-1	0	0	2														N
D	-5	0	0	-1	0	1	2	4													D
E	-5	0	0	-1	0	0	1	3	4												E
Q	-5	-1	-1	0	0	-1	1	2	2	4											Q
H	-3	-1	-1	0	-1	-2	2	1	1	3	6										H
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6									R
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5								K
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6							M
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5						I
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6					L
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4				V
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9			F
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10		Y
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17	W
C		S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

- One PAM of evolution means 1% of the residues/bases have changed, averaged over all 20 amino acids.
- To get the relative frequency of each type of mutation, we count the times it was observed in a database of multiple sequence alignments.
- Based on global alignments
- Assumes a Markov model for evolution.

BLOSUM

Henikoff & Henikoff, 1992

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

- Based on database of ungapped local alignments (BLOCKS)

- Alignments have lower similarity than PAM alignments.
- BLOSUM number indicates the percent identity level of sequences in the alignment. For example, for BLOSUM62 sequences with approximately 62% identity were counted.
- Some BLOCKS represent functional units, providing validation of the alignment.

Multiple Sequence Alignment

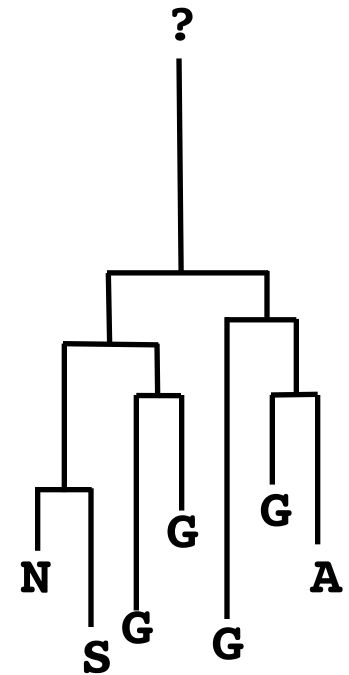
QUERY	1	KVEQAVETEPEPELR	---	QQTE	---	WQSGQRWELALGRFWDYLRWVQT	42
114042	19	KVEQPVEPETEPDVR	---	QQAE	---	WQSGQPWELALGRFWDYLRWVQT	60
178853	19	KVEQAVETEPEPELR	---	QQTE	---	WQSGQRWELALGRFWDYLRWVQT	60
4557325	19	KVEQAVETEPEPELR	---	QQTE	---	WQSGQRWELALGRFWDYLRWVQT	60
114040	19	KVEQPVEPETEPPELR	---	QQAE	---	GQSGQPWELALGRFWDYLRWVQT	60
1942471	1	KVEQAVETEPEPELR	---	QQTE	---	WQSGQRWELALGRFWDYLRWVQT	42
1263123	19	KVEQAVETEPEPELR	---	QQTE	---	WQSGQRWELALGRFWDYLRWVQT	60
1942472	1	KVEQAVETEPEPELR	---	QQTE	---	WQSGQRWELALGRFWDYLRWVQT	42
178849	19	KVEQAVETEPEPELR	---	QQTE	---	WQSGQRWELALGRFWDYLRWVQT	60
364011	19	KVKQAVETEPEPELR	---	QQTE	---	WQSGQRWELALGRFWDYLRWVQT	60
309109	19	-----EGEPEVT	---	DQLE	---	WQSNQPWEQALNRFWDYLRWVQT	52
114041	19	-----EGEPEVT	---	DQLE	---	WQSNQPWEQALNRFWDYLRWVQT	52
225946	19	-----ETEQEVEVP	---	EQAR	---	WKAGQPWELALGRFWDYLRWVQS	54
114038	19	-----DVEPEVEVR	---	EPAV	---	WQSGQPWELALSRFWDYLRWVQT	54
3915605	5	---EPELERELEPKVQ	---	QELEPEAG	---	WQTGQPWEAALARFWDYLRWVQT	48
114044	19	-----QTEQEVEVP	---	EQAR	---	WKAGQPWELALGRFWDYLRWVQS	54
2388609	21	-----EPGPPPEVHVW	---	EEPKE	---	WQGSQPWEQALGRFWDYLRWVQS	59
461527	21	-----EPGPPPEVHVW	---	EESKE	---	WQGSQPWEQALGRFWDYLRWVQS	59
1703338	19	-----EGELEVT	---	DQLP	---	GQSDQPWEQALNRFWDYLRWVQT	52
202959	43	-----EGELEVT	---	DQLP	---	GQSDQPWEQALNRFWDYLRWVQT	76
295916	19	-----EGELEVT	---	DQLP	---	GQSDQPWEQALNRFWDYLRWVQT	52
913986	19	-----EGELEVT	---	DQLP	---	GQSDQPWEQALNRFWDYLRWVQT	52
71796	19	-----EGELEVT	---	DQLP	---	GQSDQPWEQALNRFWDYLRWVQT	52
416629	21	---EGELGPEEPLTT	---	QQPR	---	GKDSQPWEQALGRFWDYLRWVQT	59
2119392	21	---EGELGPEEPLTT	---	QQPR	---	GKDSQPWEQALGRFWDYLRWVQT	59
483174	3	-----QQELE	---	PEAG	---	WQTGQPWEAALARFWDYLRWVQT	34
192005	1	-----	---	DQLE	---	WQSNQPWEQALNRFWDYLRWVQT	27
3891444	1	-----	---	---	---	SGQRWELALGRFWDYLRWVQT	21
230118	1	-----	---	---	---	GQRWELALGRFWDYLRWVQT	20
230119	1	-----	---	---	---	GQRWELALGRFWDYLRWVQT	20
230129	1	-----	---	---	---	GQRWELALGRFWDYLRWVQT	20

A multiple sequence alignment is made using many pairwise sequence alignments

Columns in a MSA have a common evolutionary history

/bach1/server/isites/tmp/junk

QUERY	1	KVEQAVETEPEPELR	---	QQTE	---	WQSGQRWELAI	G	FWDYLRWVQT	42
114042	19	KVEQAVETEPDVR	---	QQAE	---	WQSGQPWELAI	G	FWDYLRWVQT	60
178853	19	KVEQAVETEPEPELR	---	QQTE	---	WQSGQRWELAI	G	FWDYLRWVQT	60
4557325	19	KVEQAVETEPEPELR	---	QQTE	---	WQSGQRWELAI	G	FWDYLRWVQT	60
114040	19	KVEQAVETEPDVR	---	QQAE	---	GQSGQPWELAI	G	FWDYLRWVQT	60
1942471	1	KVEQAVETEPEPELR	---	QQTE	---	WQSGQRWELAI	G	FWDYLRWVQT	42
1263123	19	KVEQAVETEPEPELR	---	QQTE	---	WQSGQRWELAI	G	FWDYLRWVQT	60
1942472	1	KVEQAVETEPEPELR	---	QQTE	---	WQSGQRWELAI	G	FWDYLRWVQT	42
178849	19	KVEQAVETEPEPELR	---	QQTE	---	WQSGQRWELAI	G	FWDYLRWVQT	60
364011	19	KVKQAVETEPEPELR	---	QQTE	---	WQSGQRWELAI	G	FWDYLRWVQT	60
309109	19	---	EGEPEVT	DQLE	---	WQSNQPWEQAI	N	FWDYLRWVQT	52
114041	19	---	EGEPEVT	DQLE	---	WQSNQPWEQAI	N	FWDYLRWVQT	52
225946	19	---	ETEQEVEVP	EQAR	---	WKAGQPWELAI	G	FWDYLRWVQS	54
114038	19	---	DVEPEVEVR	EPAV	---	WQSGQPWELAI	S	FWDYLRWVQT	54
3915605	5	---	EPELERELEPKVQ	QELE	PEAG	WQTGQPWEAAI	A	FWDYLRWVQT	48
114044	19	---	QTEQEVEVP	EQAR	---	WKAGQPWELAI	G	FWDYLRWVQS	54
2388609	21	---	EPGPPPEVHVWV	ECPK	---	WQGSQPWEQAI	G	FWDYLRWVQS	59
461527	21	---	EPGPPPEVHVWV	EESK	---	WQGSQPWEQAI	G	FWDYLRWVQS	59
1703338	19	---	EGELEVT	DQLP	---	GQSDQPWEQAI	N	FWDYLRWVQT	52
202959	43	---	EGELEVT	DQLP	---	GQSDQPWEQAI	N	FWDYLRWVQT	76
295916	19	---	EGELEVT	DQLP	---	GQSDQPWEQAI	N	FWDYLRWVQT	52
913986	19	---	EGELEVT	DQLP	---	GQSDQPWEQAI	N	FWDYLRWVQT	52
71796	19	---	EGELEVT	DQLP	---	GQSDQPWEQAI	N	FWDYLRWVQT	52
416629	21	---	EGELGPE	EPLTT	---	GKDSQPWEQAI	G	FWDYLRWVQT	59
2119392	21	---	EGELGPE	EPLTT	---	GKDSQPWEQAI	G	FWDYLRWVQT	59
483174	3	---	QQELE	PEAG	---	WQTGQPWEAAI	A	FWDYLRWVQT	34
192005	1	---	---	DQLE	---	WQSNQPWEQAI	N	FWDYLRWVQT	27
3891444	1	---	---	---	---	SGQRWELAI	G	FWDYLRWVQT	21
230118	1	---	---	---	---	GQRWELAI	G	FWDYLRWVQT	20
230119	1	---	---	---	---	GQRWELAI	G	FWDYLRWVQT	20
230129	1	---	---	---	---	GQRWELAI	G	FWDYLRWVQT	20

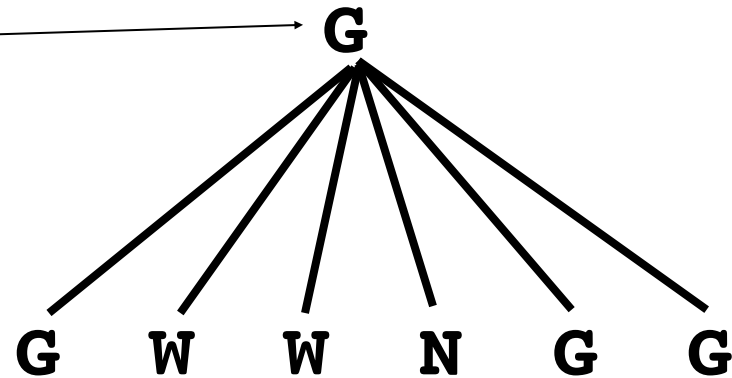


By aligning the sequences, we assert that the aligned residues in each column had a common ancestor.

Counting mutations without knowing ancestral sequences

Assume *any* of the sequences could be the ancestral one.

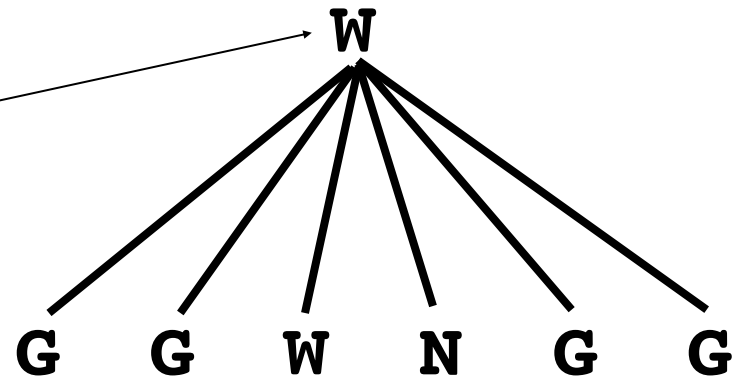
L	K	F	G	R	L	S	K	K	P
L	K	F	G	R	L	S	K	K	P
L	K	F	W	R	L	T	K	K	P
L	K	F	W	R	L	S	K	K	P
L	K	F	N	R	L	S	R	K	P
L	K	F	G	R	L	T	R	K	P
L	K	F	G	R	L	~	K	K	P



If the first sequence was the ancestor, then it mutated to a **W** twice, to **N** once, and conserved **G** three times.

Or, instead of G we could have picked W as the ancestor...

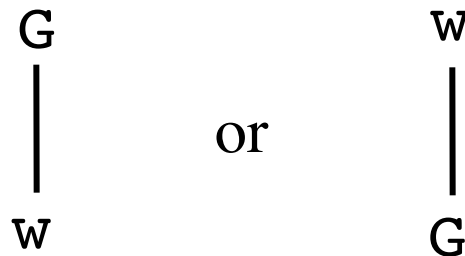
L	K	F	G	R	L	S	K	K	P
L	K	F	G	R	L	S	K	K	P
L	K	F	W	R	L	T	K	K	P
L	K	F	W	R	L	S	K	K	P
L	K	F	N	R	L	S	R	K	P
L	K	F	G	R	L	T	R	K	P
L	K	F	G	R	L	~	K	K	P



W was the ancestor, then it mutated to a **G** four times, to **N** once, and was conserved once.

Substitution matrices are symmetrical

Since we don't know which sequence came first, we don't know whether



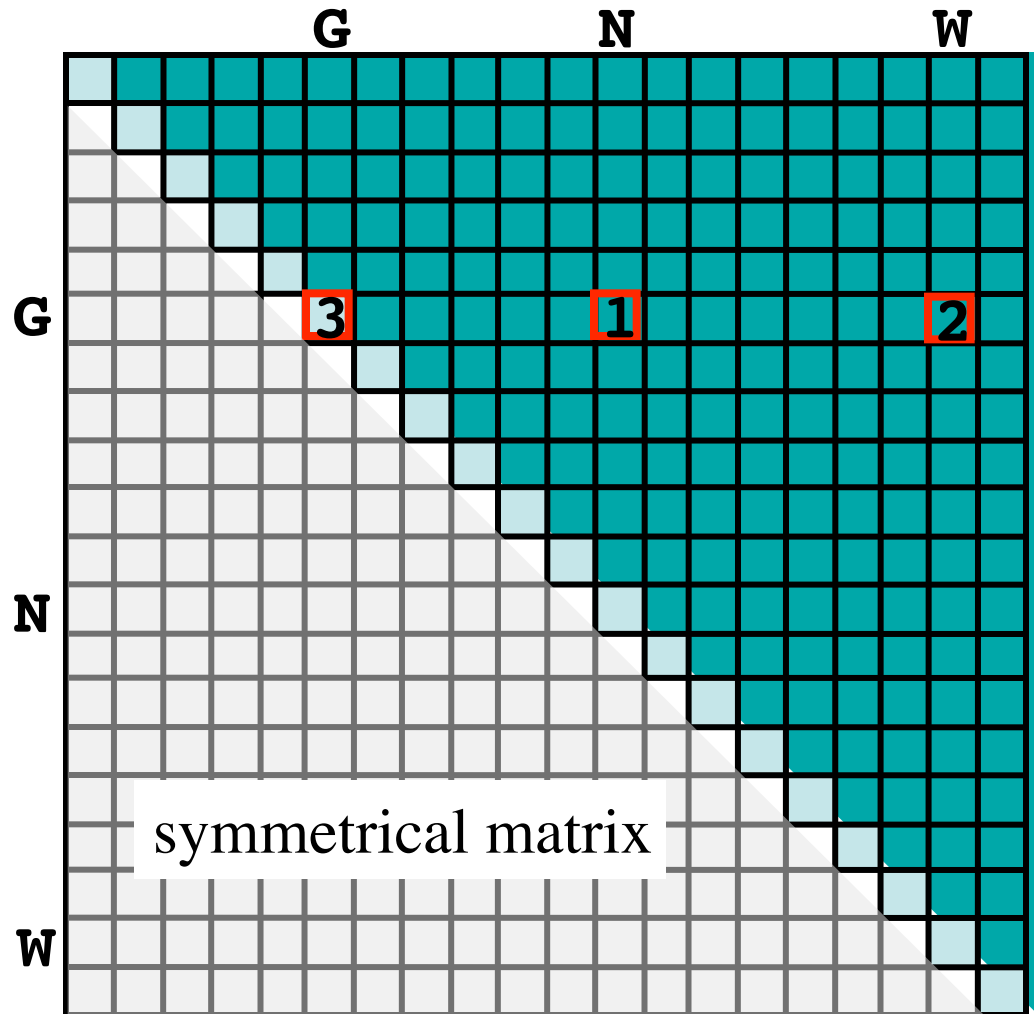
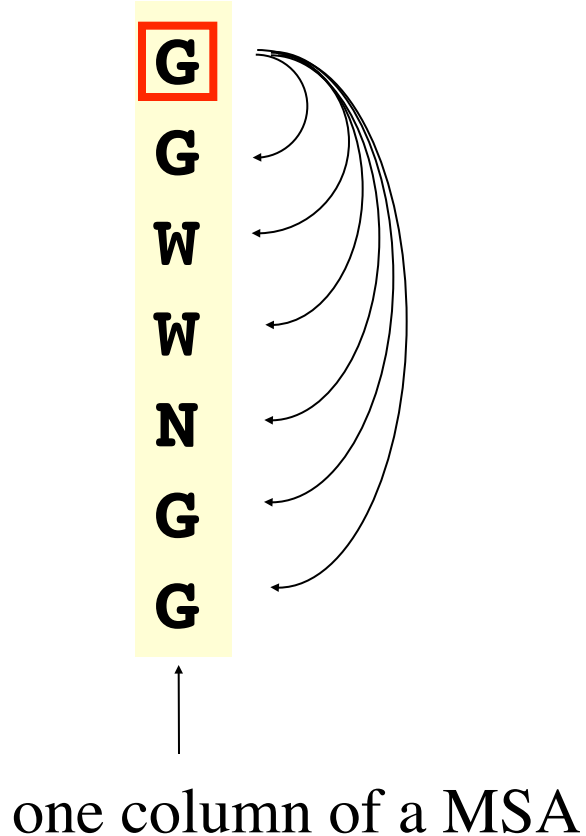
...is correct. So we count this as one mutation of each type.

G-->W and W-->G. In the end the 20x20 matrix will have the same number for elements (i,j) and (j,i).

(That's why we only show the upper triangle)

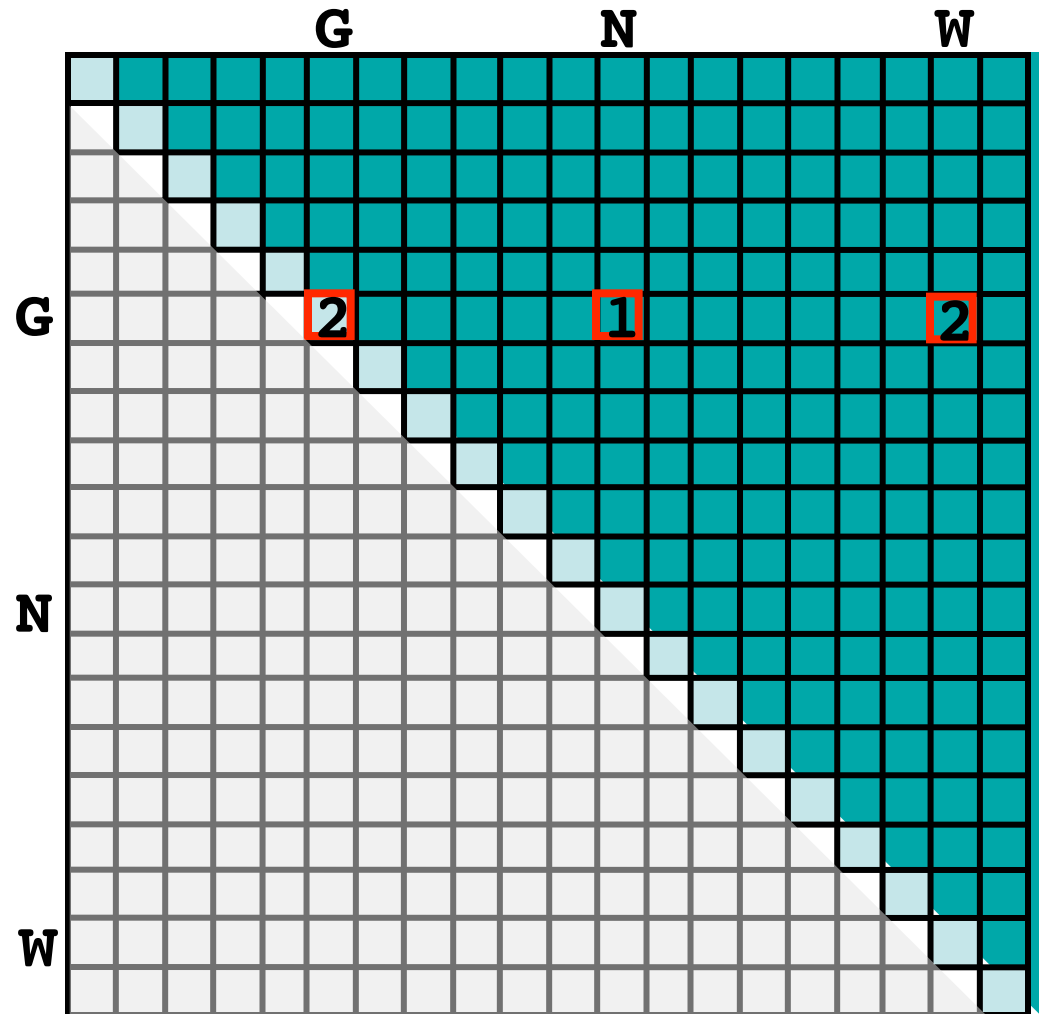
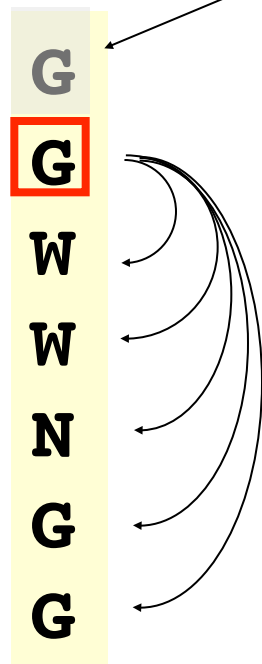
Summing the substitution

We assume the ancestor is one of the observed amino acids, but we don't know which, so we try them all.

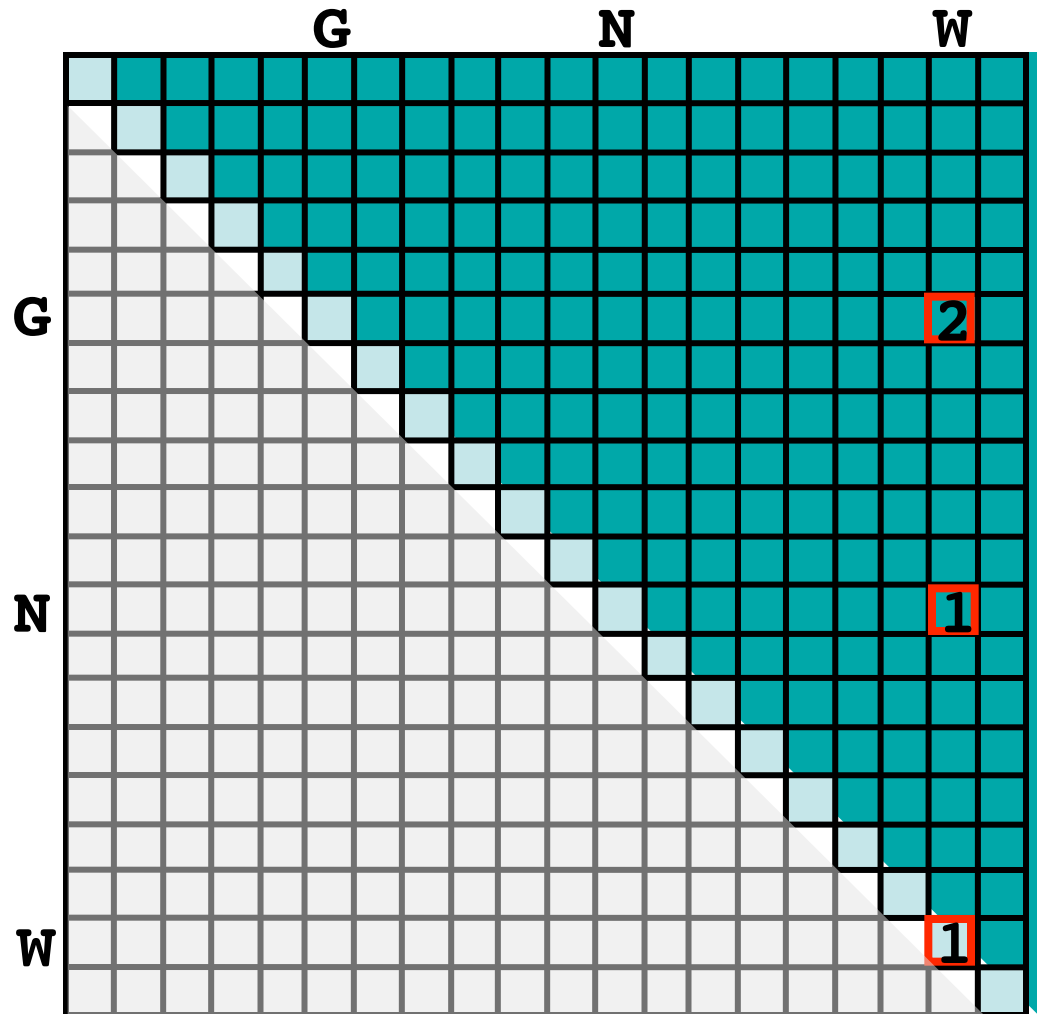
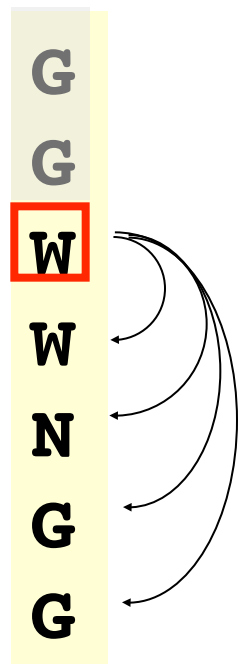


Next possible ancestor, G again.

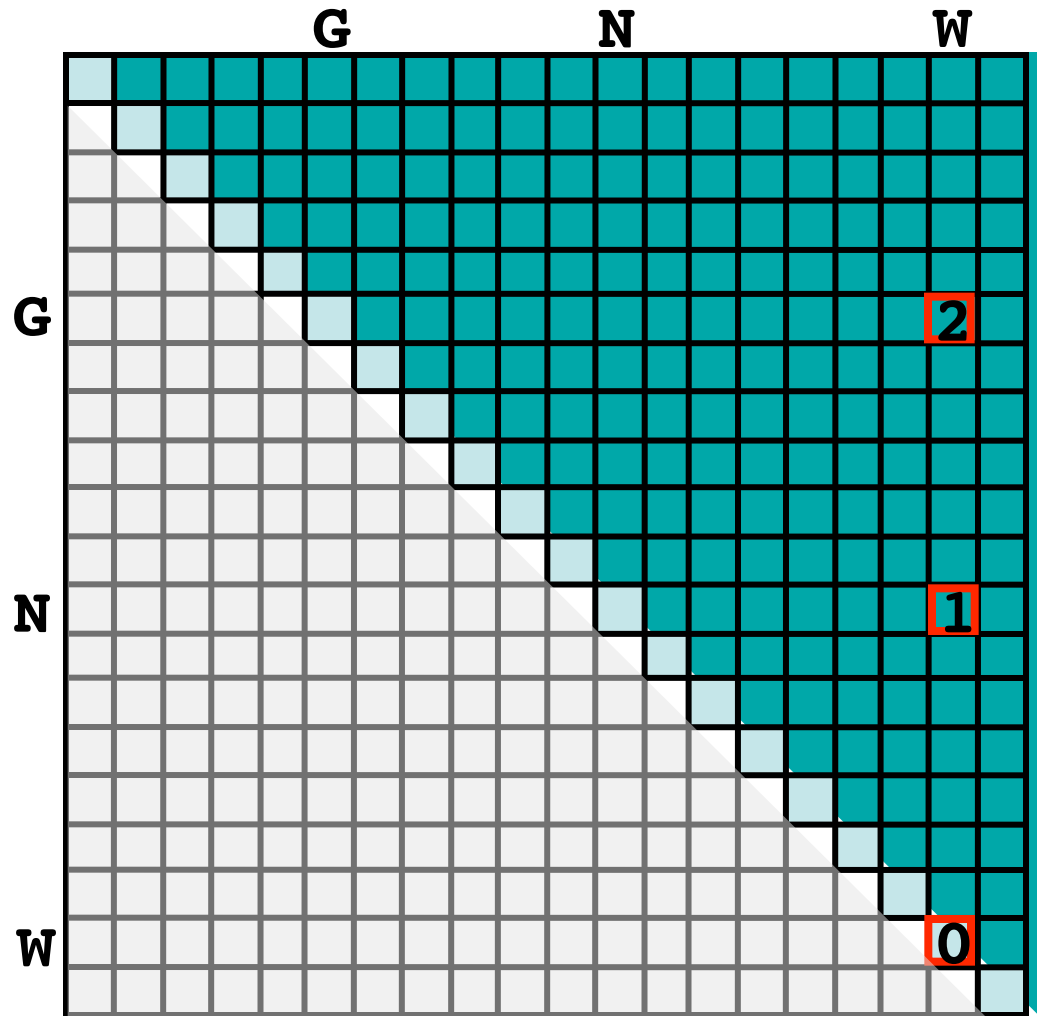
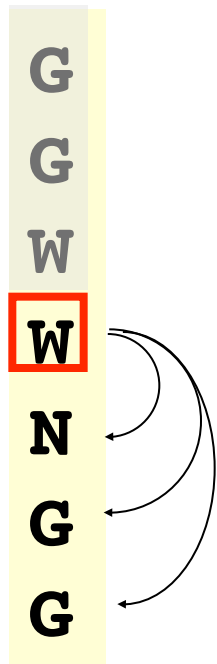
We already counted this residue against all others, so be blank it out.



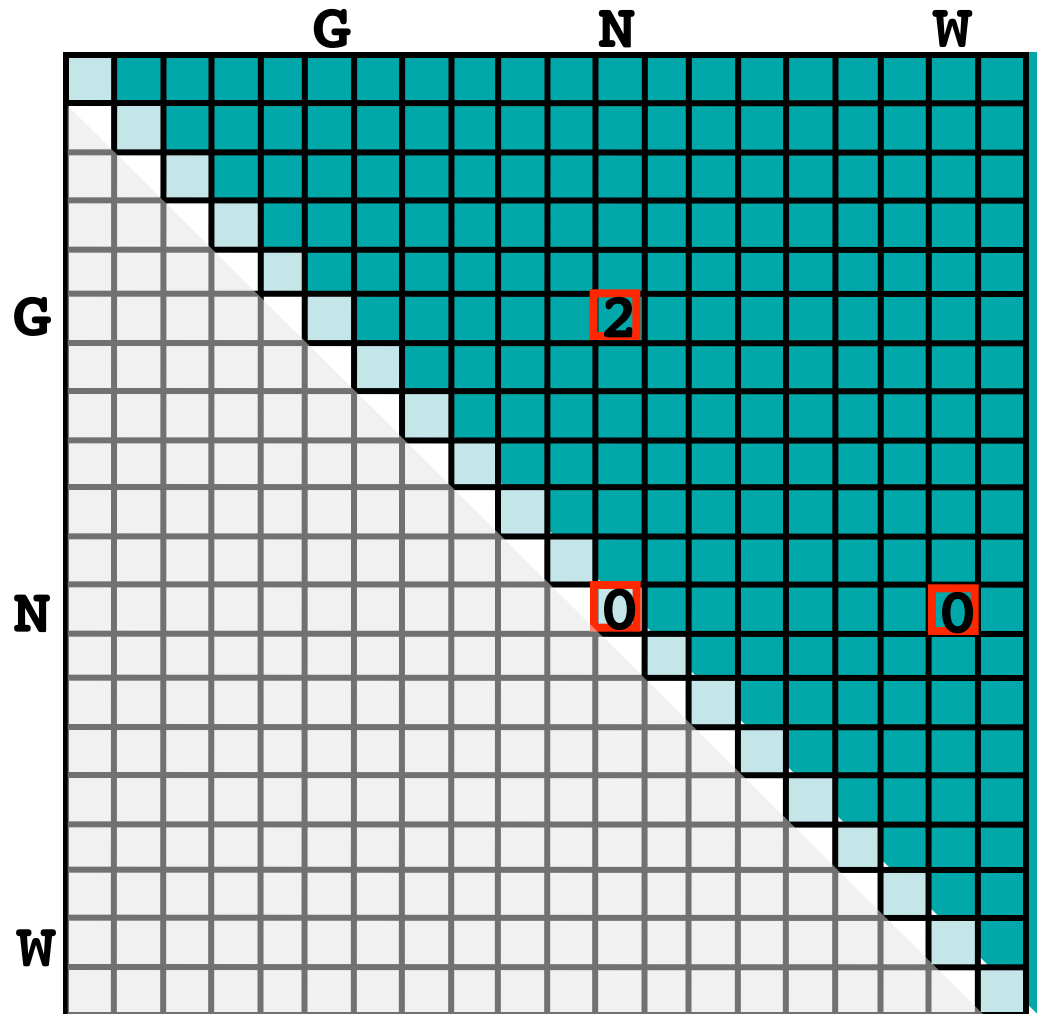
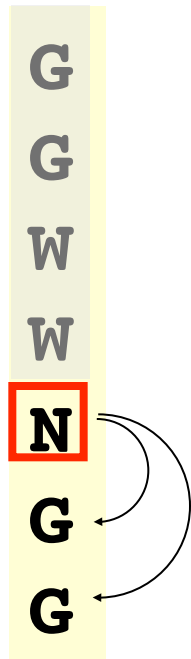
Next...W



Next...W again

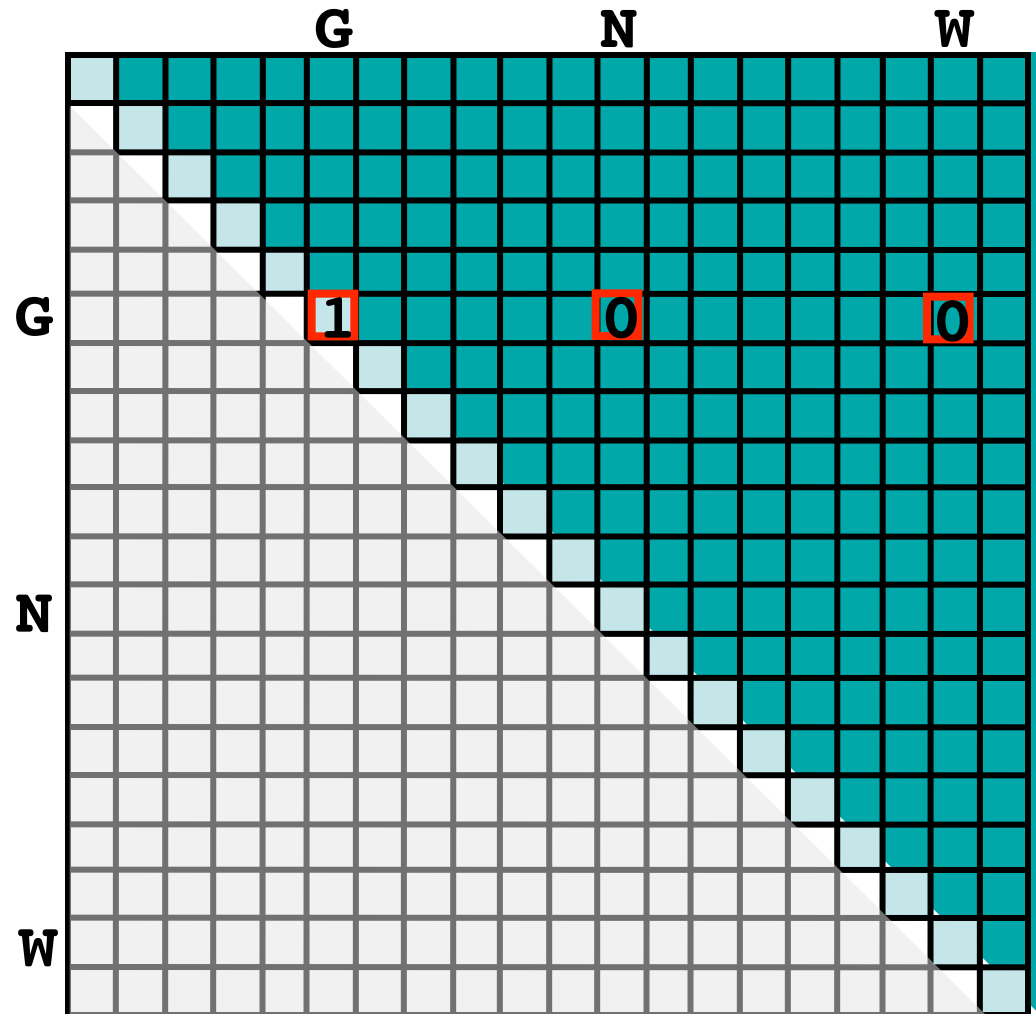


Next...N



Next...G again

G
G
W
W
N
G
G

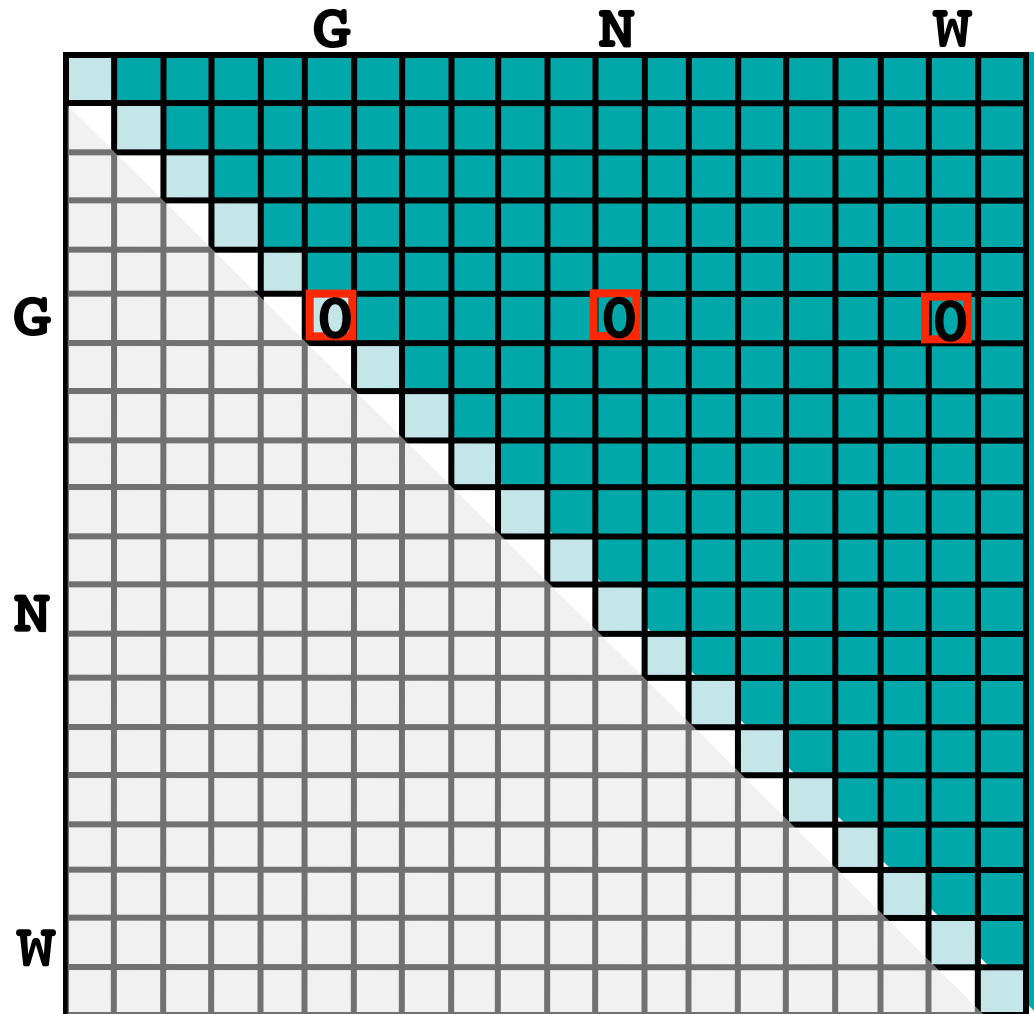


Counting **G** as the ancestor many times as it appears recognizes the increased likelihood that **G** (the most frequent aa at this position) is the true ancestor.

Last...G again.

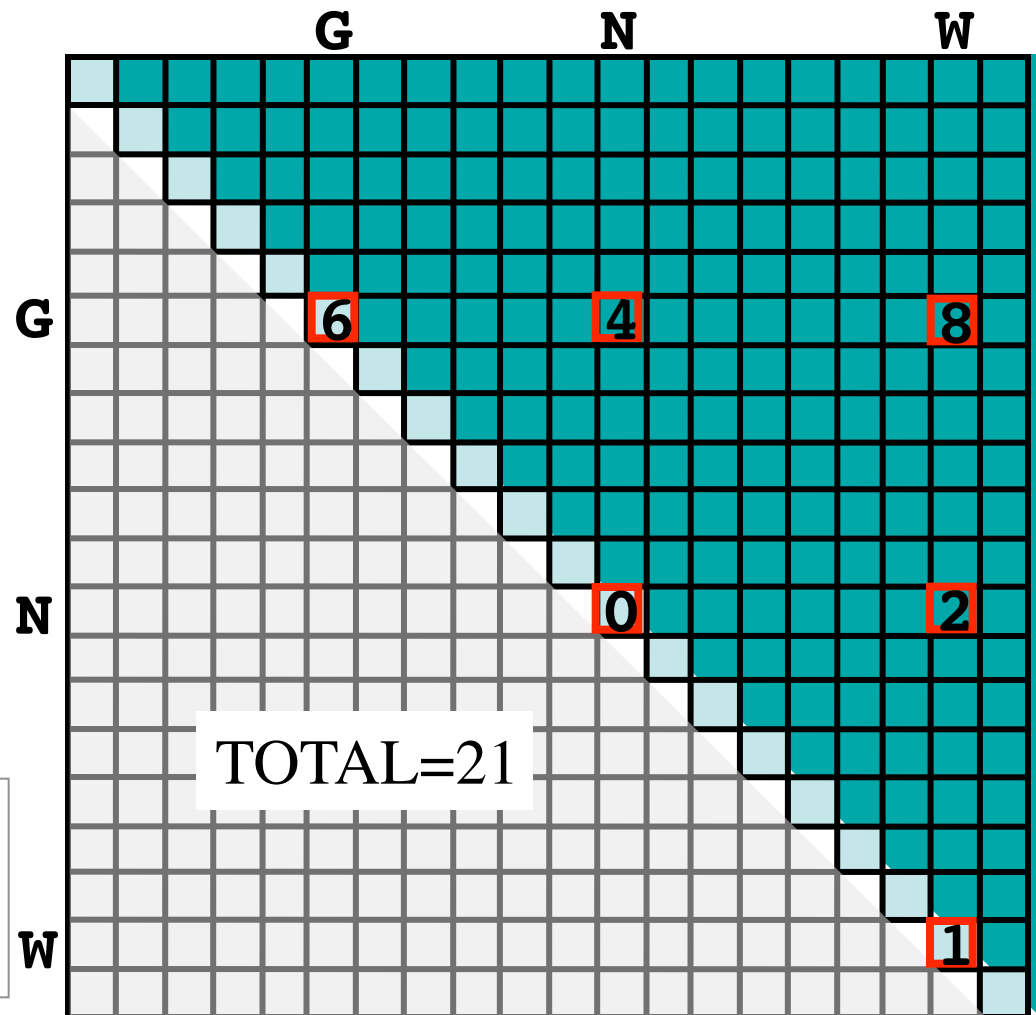
G
G
W
W
N
G
G

(no counts for last seq.)



Go to next column. Continue summing.

G P
G P
W I
W N
N P
G P
G A



Continue doing this for every column in every multiple sequence alignment...

log odds

Substitutions (and many other things in bioinformatics) are expressed as a "likelihood ratio", or "odds ratio" of the observed data over the expected value. Likelihood and odds are synonyms for Probability.

So Log Odds is the log (usually base 2) of the odds ratio.

$$\text{log odds ratio} = \log_2(\text{observed/expected})$$

Getting log-odds from counts

$$P(G) = 4/7 = 0.57$$

Observed probability of G->G
 $q_{GG} = P(G \rightarrow G) = 6/21 = 0.29$

Expected probability of G->G,
 $e_{GG} = 0.57 * 0.57 = 0.33$

$$\text{odds ratio} = q_{GG}/e_{GG} = 0.29/0.33$$

$$\text{log odds ratio} = \log_2(q_{GG}/e_{GG})$$

If the 'lod' is < 0., then the mutation is less likely than expected by chance. If it is > 0., it is more likely.

Different observations, same expectation

$$P(G)=0.50$$

$$e_{GG} = 0.25$$

$$q_{GG} = 9/42 = 0.21$$

$$\text{lod} = \log_2(0.21/0.25) = -0.2$$

G	G
G	A
W	G
W	A
N	G
G	A
G	A

G's spread over many columns

$$P(G)=0.50$$

$$e_{GG} = 0.25$$

$$q_{GG} = 21/42 = 0.5$$

$$\text{lod} = \log_2(0.50/0.25) = 1$$

G	W
G	A
G	W
G	A
G	W
G	A
G	A

G's concentrated

Different observations, same expectation

$$P(G)=0.50, P(W)=0.14$$

$$e_{GW} = 0.07$$

$$q_{GW} = 7/42 = 0.17$$

$$\text{lod} = \log_2(0.17/0.07) = 1.3$$

G	G
G	A
W	G
A	W
N	G
G	A
G	A

G and W seen together more often than expected.

$$P(G)=0.50, P(W)=0.14$$

$$e_{GW} = 0.07$$

$$q_{GG} = 3/42 = 0.07$$

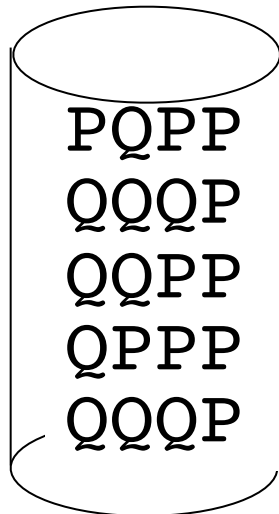
$$\text{lod} = \log_2(0.07/0.07) = 0$$

G	W
G	A
G	W
G	A
G	W
G	A
A	G

G's and W's not seen together.

In class exercise:

Get the substitution value for P->Q



sequence
alignment
database.

	P	Q
P		
Q		

substitution
counts

$$P(P) = \underline{\hspace{2cm}}, P(Q) = \underline{\hspace{2cm}}$$

$$e_{PQ} = \underline{\hspace{2cm}}$$

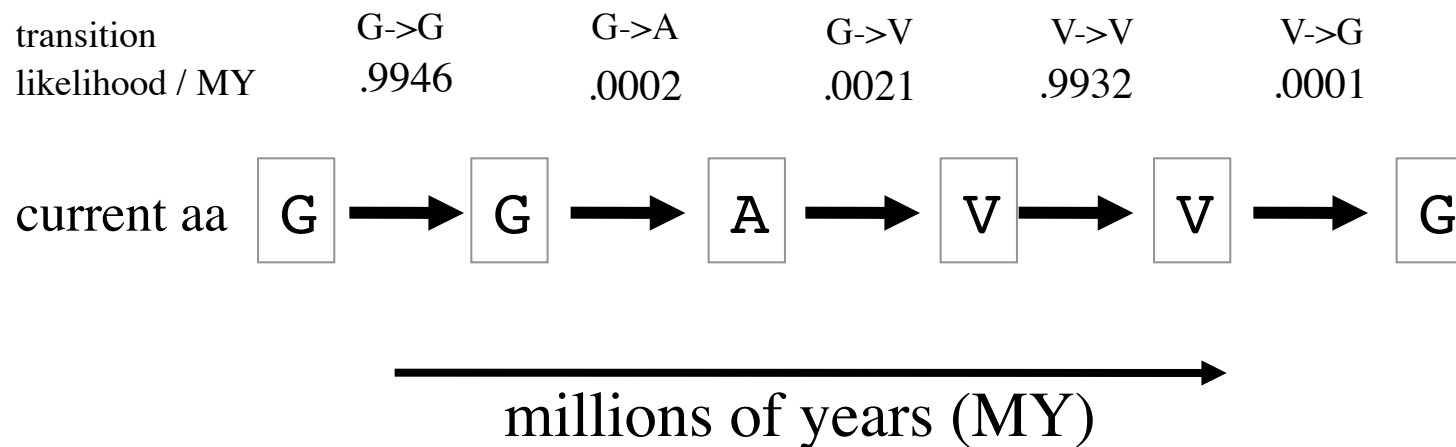
$$q_{PQ} = \underline{\hspace{1cm}} / \underline{\hspace{1cm}} = \underline{\hspace{2cm}}$$

$$\text{lod} = \log_2(q_{PQ}/e_{PQ}) = \underline{\hspace{2cm}}$$

expected (e), versus
observed (q) for P->Q

Markovian evolution and PAM

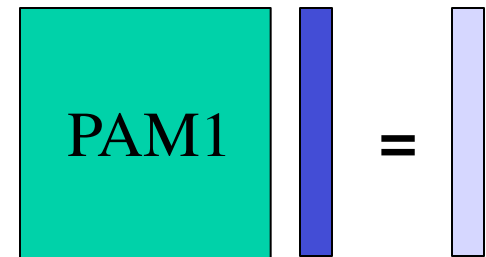
A **Markov process** is one where the likelihood of the next "state" depends only on the current state. The inference that **evolution is Markovian** assumes that base changes (or amino acid changes) occur at a constant rate and depend only on the *identity of the current base (or amino acid)*.



Markovian evolution is an extrapolation

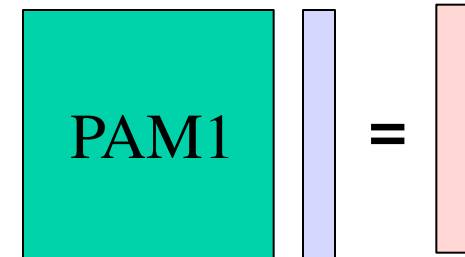
Start with all G's. Wait 1 million years. Where do they go?

Using PAM1, we expect them to mutate to about 0.0002 A, 0.0007 P, 0.9946 G, etc



Wait another million years.

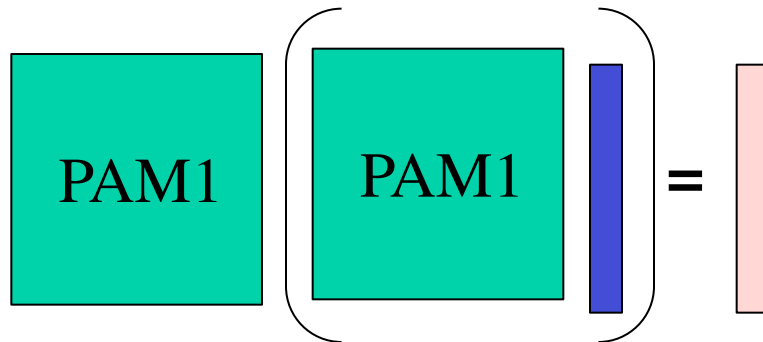
The new A's mutate according to PAM1 for A's, P's mutate according to PAM1 for P's, etc.



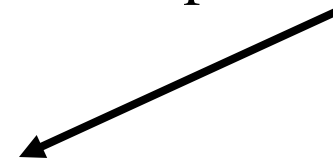
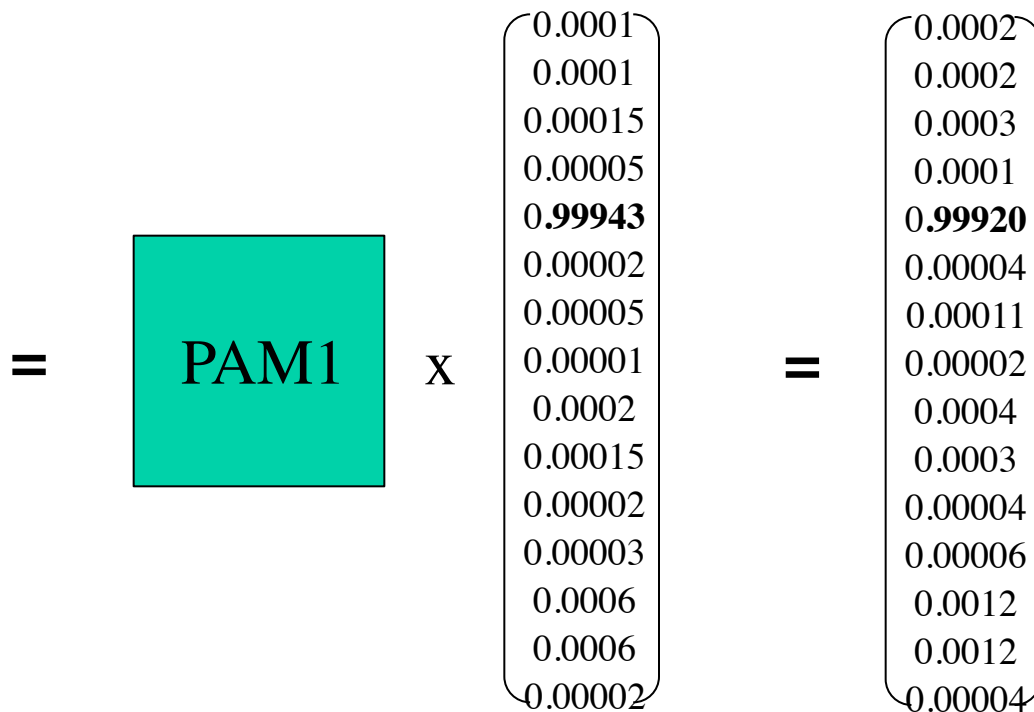
Wait another million, etc , etc etc.

What is the final distribution of amino acids at the positions that were once G's?

Matrix multiplication

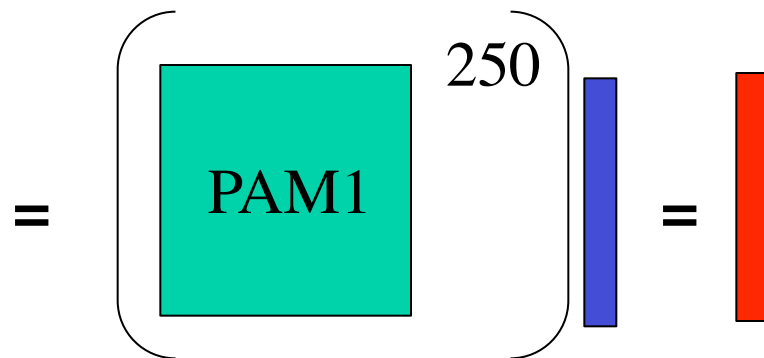
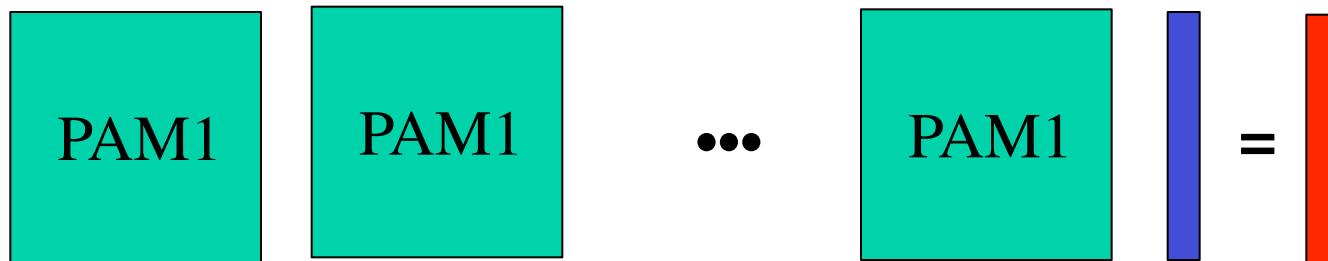


After 2MY each amino acid has mutated again according to the PAM1 probabilities.



etc.

“PAM250” = PAM²⁵⁰



Differences between PAM and BLOSUM

PAM

- PAM matrices are based on *global alignments* of closely related proteins.
- The PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence.
- Other PAM matrices are extrapolated from PAM1 using an assumed Markov chain.

BLOSUM

- BLOSUM matrices are based on *local alignments*.
- BLOSUM 62 is a matrix calculated from comparisons of sequences with approx 62% identity.
- All BLOSUM matrices are based on observed alignments; they are not extrapolated from comparisons of closely related proteins.
- BLOSUM 62 is the default matrix in BLAST (the database search program). It is tailored for comparisons of moderately distant proteins. Alignment of distant relatives may be more accurate with a different matrix.

PAM250

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	12																				C
S	0	2																			S
T	-2	1	3																		T
P	-3	1	0	6																	P
A	-2	1	1	1	2																A
G	-3	1	0	-1	1	5															G
N	-4	1	0	-1	0	0	2														N
D	-5	0	0	-1	0	1	2	4													D
E	-5	0	0	-1	0	0	1	3	4												E
Q	-5	-1	-1	0	0	-1	1	2	2	4											Q
H	-3	-1	-1	0	-1	-2	2	1	1	3	6										H
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6									R
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5								K
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6							M
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5						I
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6					L
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4				V
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9			F
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10		Y
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

BLOSUM62

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

In class exercise:
Which substitution matrix favors...

	PAM250	BLOSUM62
conservation of polar residues	<input type="checkbox"/>	<input type="checkbox"/>
conservation of non-polar residues	<input type="checkbox"/>	<input type="checkbox"/>
conservation of C, Y, or W	<input type="checkbox"/>	<input type="checkbox"/>
polar-to-nonpolar mutations	<input type="checkbox"/>	<input type="checkbox"/>
polar-to-polar mutations	<input type="checkbox"/>	<input type="checkbox"/>

Protein versus DNA alignments

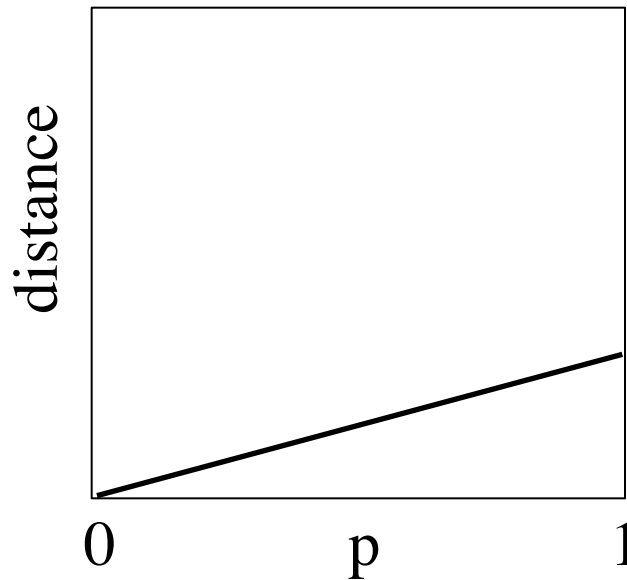
Are protein alignment better?

- Protein alphabet = 20, DNA alphabet = 4.
 - Protein alignment is more informative
 - Less chance of homoplasy with proteins.
 - Homology detectable at greater edit distance
 - Protein alignment more informative
- Better Gold Standard alignments are available for proteins.
 - Better statistics from G.S. alignments.
- On the other hand, DNA alignments are more sensitive to short evolutionary distances.

DNA evolutionary models: P-distance

What is the relationship between time and the %identity?

$$p = \frac{D}{L}$$



p is a good measure of time only when p is small.

DNA evolutionary models: Poisson correction

Corrects for multiple mutations at the same site. Unobserved mutations.

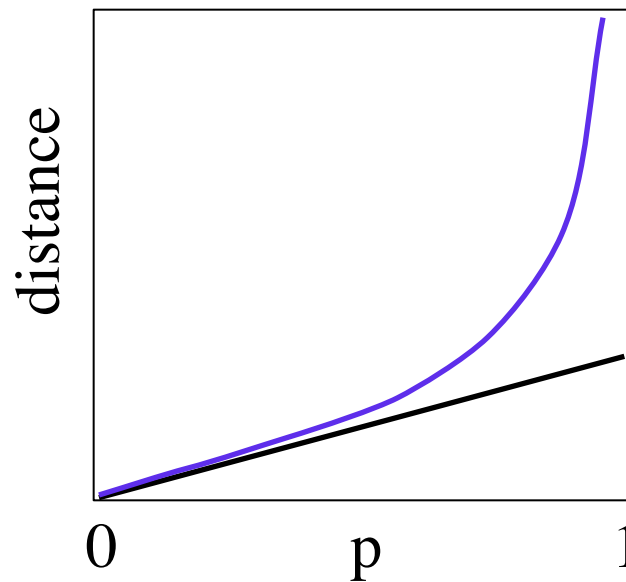
$$p = \frac{D}{L}$$

The fraction unchanged decays according to the Poisson function. In the time t since the common ancestor, $2rt$ mutations have occurred, where r is the mutation rate ($r = \text{genetic drift} * \text{selection pressure}$)

$$1-p = e^{-2rt}$$

$$d_P = 2rt$$

$$d_P = -\ln(1-p)$$



Poisson correction assumes p goes to 1 at $t=\infty$. Where should it really go?

DNA evolutionary models: Jukes-Cantor

What is the relationship between true evolutionary distance and *p-distance*?

Prob(mutation in one unit of time) = α $\alpha \ll 1$.

Prob(no mutation) = $1-3\alpha$

At time t , fraction identical is $q(t)$.

Fraction non-identical is $p(t)$.

$$p(t) + q(t) = 1$$

In time $t+1$, each of $q(t)$ positions stays same with prob = $1-3\alpha$.

	A	C	G	T
A	$1-3\alpha$	α	α	α
C	α	$1-3\alpha$	α	α
G	α	α	$1-3\alpha$	α
T	α	α	α	$1-3\alpha$

Prob that **both** sequences do not mutate = $(1-3\alpha)^2 = (1-6\alpha+9\alpha^2) \approx (1-6\alpha)$.

(Since $\alpha \ll 1$, we can safely neglect α^2 .)

Prob that a mismatch mutates back to an identity = $2\alpha p(t)$

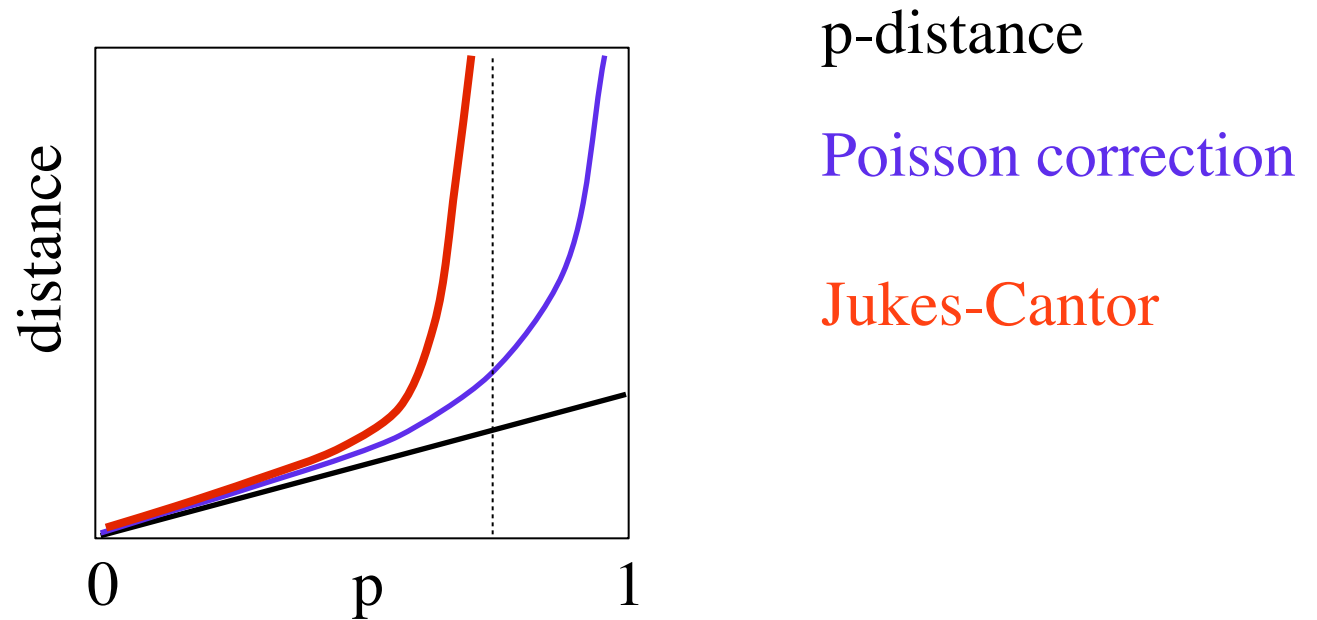
$$q(t+1) = q(t)(1-6\alpha) + 2\alpha(1-q(t))$$

$$d q(t)/dt \approx q(t+1) - q(t) = 2\alpha - 8\alpha q(t)$$

$$\text{Integrating: } q(t) = (1/4)(1 + 3\exp(-8\alpha t))$$

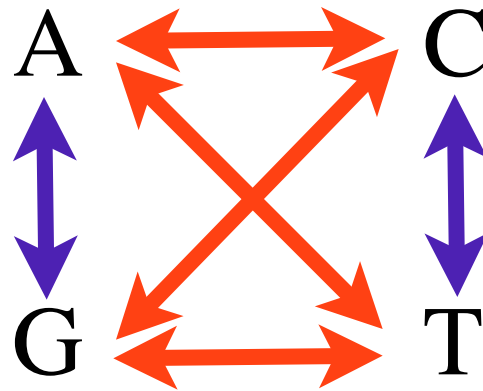
Solving for $d_{JC} = 6\alpha t = -(3/4)\ln(1 - (4/3)p)$, where p is the *p-distance*.

DNA evolutionary models



In Jukes-Cantor, p limits to $p=0.75$ at infinite evolutionary distance.

Transitions/transversions



In DNA replication, errors can be **transitions** (purine for purine, pyrimidine for pyrimidine) or **transversions** (purine for pyrimidine & vice versa)

$R = \text{transitions/transversions}$.

R would be $1/2$ if all mutations were equally likely. In DNA alignments, R is observed to be about 4. Transitions are greatly favored over transversions.

Jukes-Cantor with correction for transitions/ transversions

(Kimura 2-parameter model, d_{K2P})

	A	C	G	T
A	$1-2\beta-\alpha$	β	α	β
C	β	$1-2\beta-\alpha$	β	α
G	α	β	$1-2\beta-\alpha$	β
T	β	α	β	$1-2\beta-\alpha$

Split changes (D) into the two types, transition (P) and transversion (Q)

$$p\text{-distance} = D/L = P + Q$$

$$P = \text{transitions}/L, \quad Q = \text{transversions}/L$$

The the corrected evolutionary distance is...

$$d_{K2P} = -(1/2)\ln(1-2P-Q) - (1/4)\ln(1-2Q)$$

Further corrections are possible

	A	C	G	T
A	Green	Orange	Orange	Orange
C	Grey	Green	Orange	Orange
G	Grey	Grey	Green	Orange
T	Grey	Grey	Grey	Green

A nucleotide substitution matrix?

Additional corrections for:

- Sequence position (gamma)
- Isochores (GC-rich, AT-rich regions)
- ??

Things to remember

- Amino acid substitution matrices come from reliable alignments
 - expressed as log-odds, “lods”
 - log of ratio of observed vs expected
 - Markovian evolution
- DNA, p-distance
 - Poisson
 - Jukes-Cantor
 - transitions/transversions. Kimura.