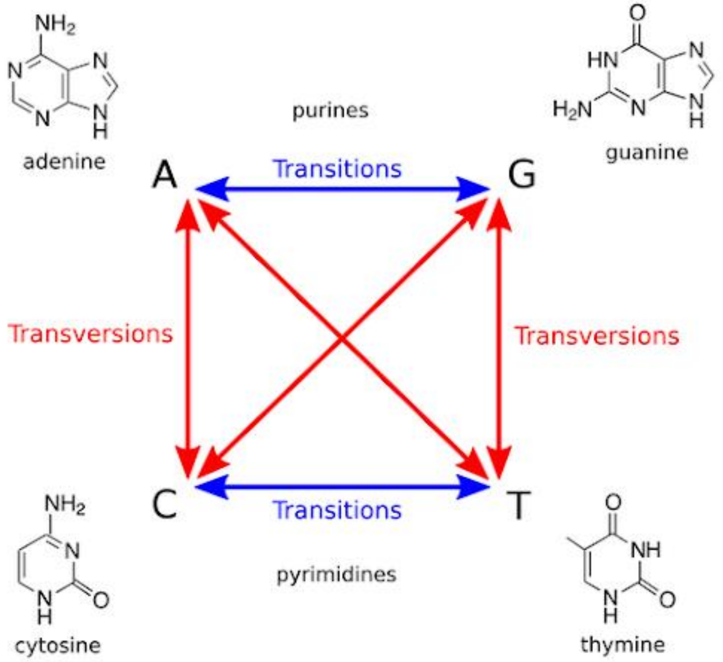
Comparison of sequences

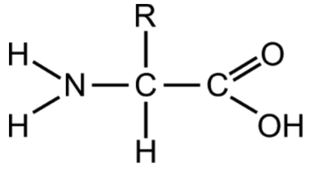
INDRAPRASTHA INSTITUTE *of* INFORMATION TECHNOLOGY **DELHI**

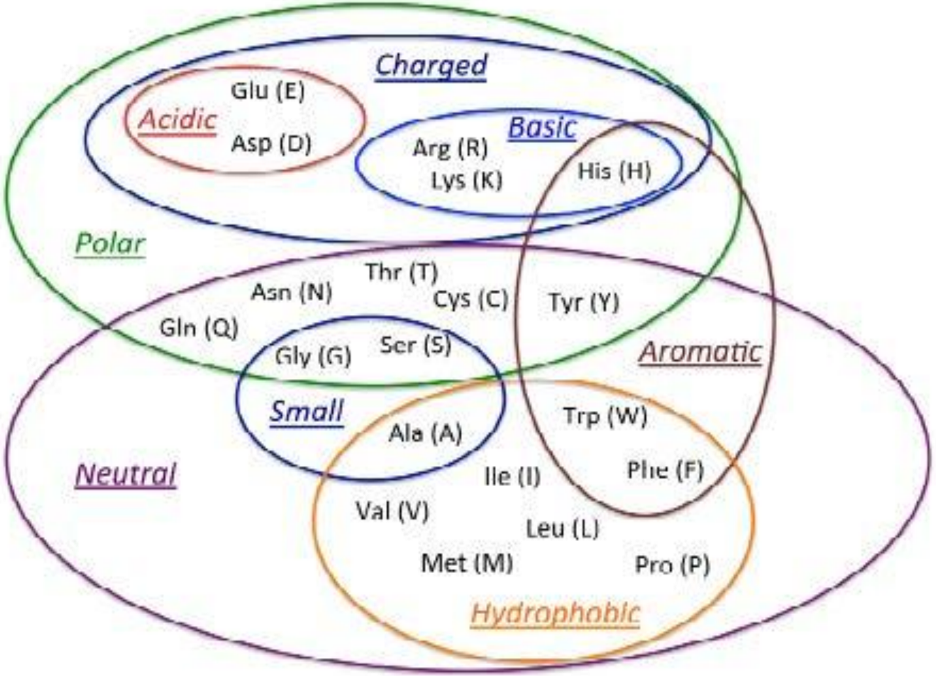
Dr. Jaspreet Kaur Dhanjal
Assistant Professor, Center for Computational Biology
Email ID: jaspreet@iiitd.ac.in

August 22, 2025

Scoring mismatches







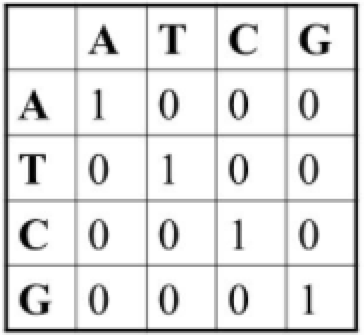
Substitution Matrices

Scoring schemes or Weight matrices

- Substitution matrices are used to score substitution events in alignments
- Particularly important in protein sequence alignments but relevant to DNA sequences as well
- Each scoring matrix represents a particular theory of evolution
- All algorithms to compare protein sequences rely on some scheme to score the equivalence of each of the 210 possible pairs
- 190 different pairs + 20 identical pairs
- Higher scores for identical/similar amino acids (e.g. A,A or I, L)
- Lower scores to different character (e.g. I, D)

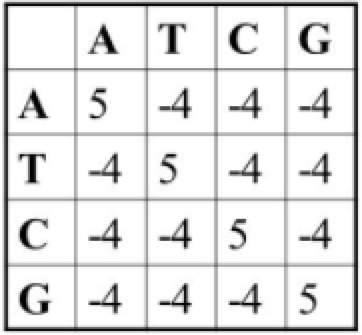
Identity scoring matrix

- Simplest Scoring scheme
- Score 1 for identical pairs
- Score 0 for non-identical pairs
- Unable to detect similarity
- Percent identity



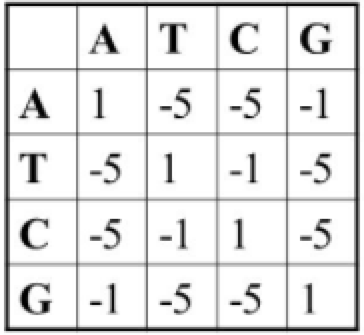
	A	T	C	G
A	1	0	0	0
T	0	1	0	0
C	0	0	1	0
G	0	0	0	1

Identity



	A	T	C	G
A	5	-4	-4	-4
T	-4	5	-4	-4
C	-4	-4	5	-4
G	-4	-4	-4	5

BLAST

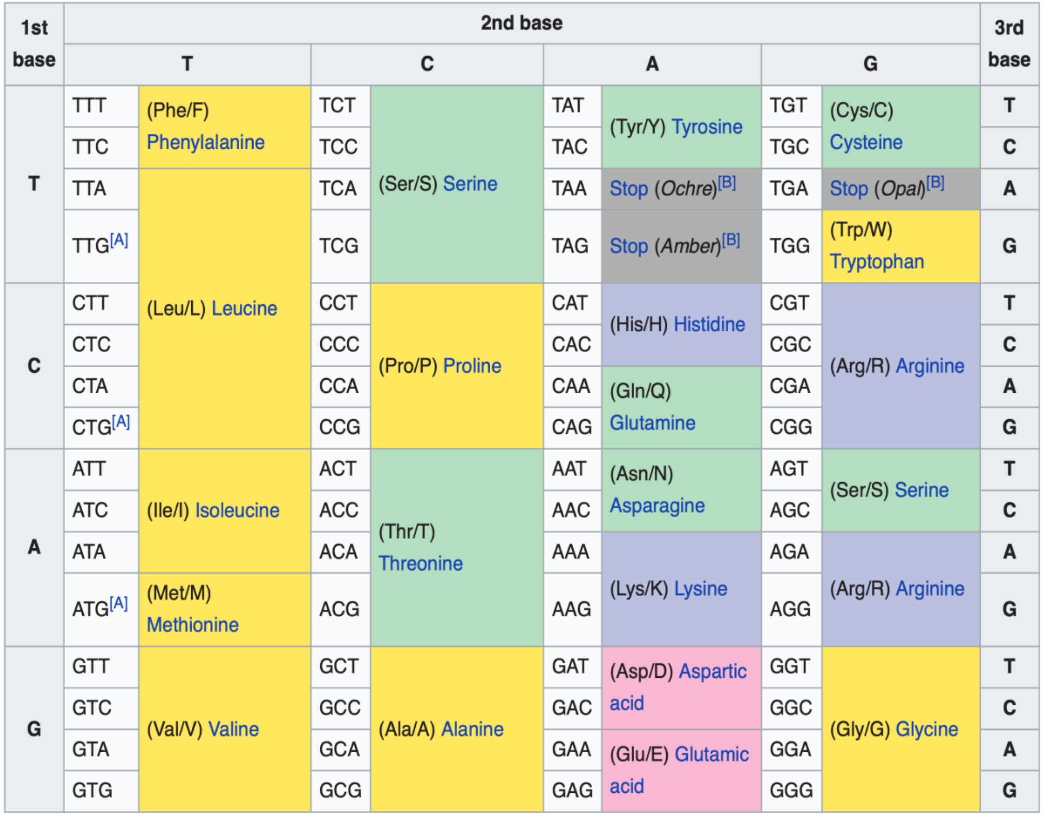


	A	T	C	G
A	1	-5	-5	-1
T	-5	1	-1	-5
C	-5	-1	1	-5
G	-1	-5	-5	1

Transition/Transversion

Genetic code scoring scheme

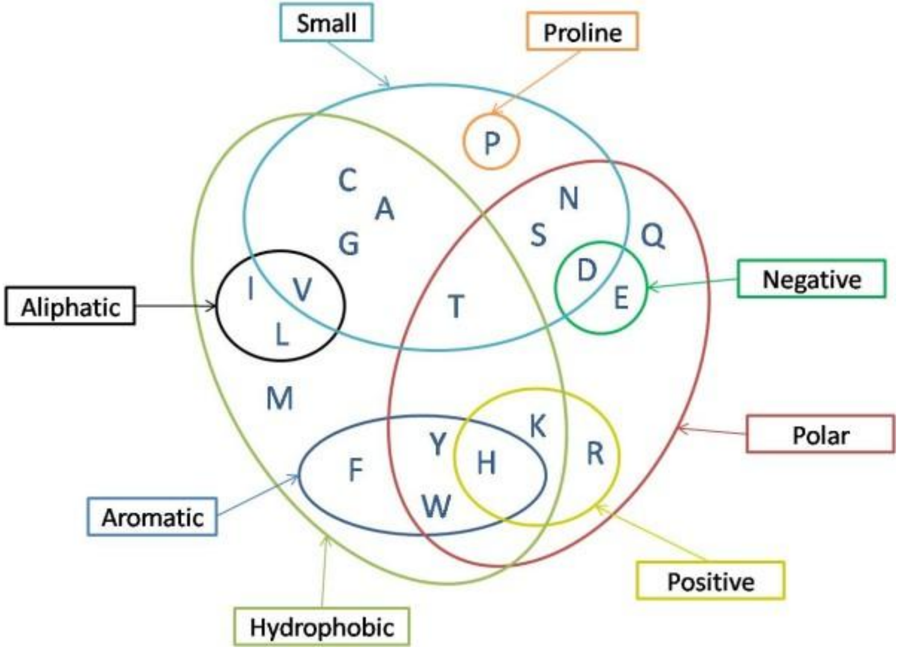
- Introduced by Fitch 1966
- Nucleotide Base change required (0,1,2,3) to interconvert the codons for the two amino acids
- Used both in the construction of phylogenetic trees and in the determination of homology between protein sequences having similar three dimensional structures
- Rarely used today



1st base	2nd base								3rd base
	T		C		A		G		
T	TTT	(Phe/F) Phenylalanine	TCT	(Ser/S) Serine	TAT	(Tyr/Y) Tyrosine	TGT	(Cys/C) Cysteine	T
	TTC		TCC		TAC		TGC		C
	TTA	(Leu/L) Leucine	TCA		TAA	Stop (*Ochre*)[B]	TGA	Stop (*Opal*)[B]	A
	TTG[A]		TCG		TAG	Stop (*Amber*)[B]	TGG	(Trp/W) Tryptophan	G
C	CTT		CCT	(Pro/P) Proline	CAT	(His/H) Histidine	CGT	(Arg/R) Arginine	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	(Gln/Q) Glutamine	CGA		A
	CTG[A]		CCG		CAG		CGG		G
A	ATT	(Ile/I) Isoleucine	ACT	(Thr/T) Threonine	AAT	(Asn/N) Asparagine	AGT	(Ser/S) Serine	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine	A
	ATG[A]	(Met/M) Methionine	ACG		AAG		AGG		G
G	GTT	(Val/V) Valine	GCT	(Ala/A) Alanine	GAT	(Asp/D) Aspartic acid	GGT	(Gly/G) Glycine	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	(Glu/E) Glutamic acid	GGA		A
	GTG		GCG		GAG		GGG		G

Chemical Similarity Scoring

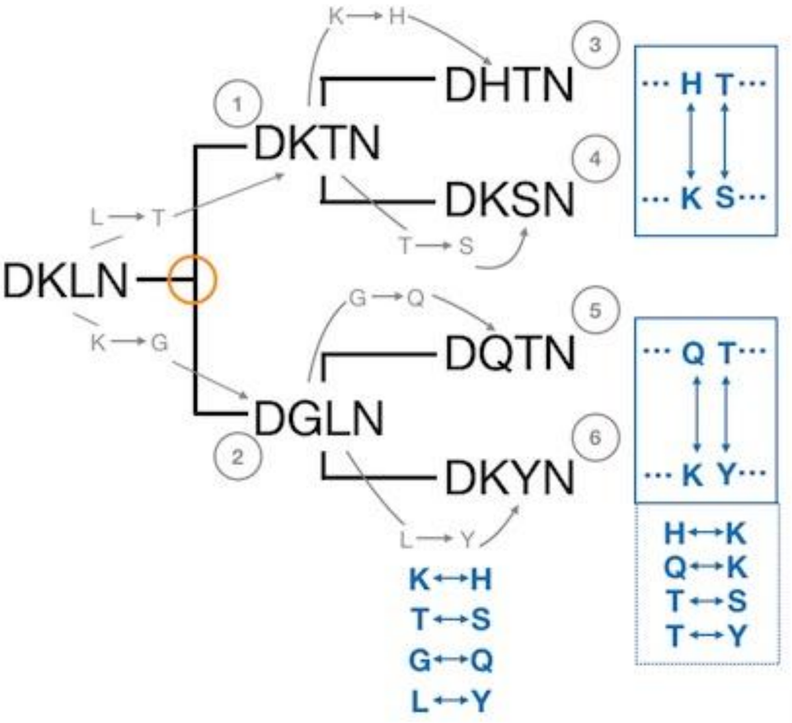
- Introduced by MacLachlan 1972
- Greater weight to the alignment of amino acids with similar physico-chemical properties
- Amino acids are classified on the basis of polar or non-polar character, size, shape and charge
- Score 0 for opposite (e.g. E & F) and 6 for identical character (e.g. F & F)

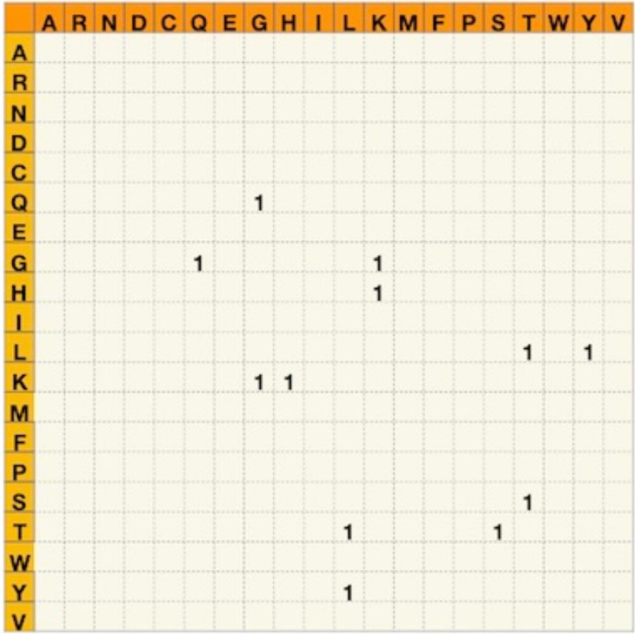


PAM substitution matrices

- Dayhoff, Schwarz and Orcutt 1978 constructed the PAM (Point Accepted Mutations) matrices
- Took 71 protein families - where the sequences differed by no more than 15% of residues (i.e. 85% identical)
- Obtained frequencies for residue X being substituted by residue Y over time period Z
- Ignores evolutionary direction
- Based on 1572 residue changes
- They defined a substitution matrix as 1 PAM (point accepted mutation) if the expected number of substitutions was 1% of the sequence length
- Or, The PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence.
- To increase the distance, they multiplied the PAM1 matrix
- PAM250 is one of the most commonly used PAM matrix

PAM





	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A																				
R																				
N																				
D																				
C																				
Q								1												
E																				
G						1						1								
H												1								
I																				
L																	1		1	
K								1	1											
M																				
F																				
P																				
S																	1			
T											1					1				
W																				
Y											1									
V																				

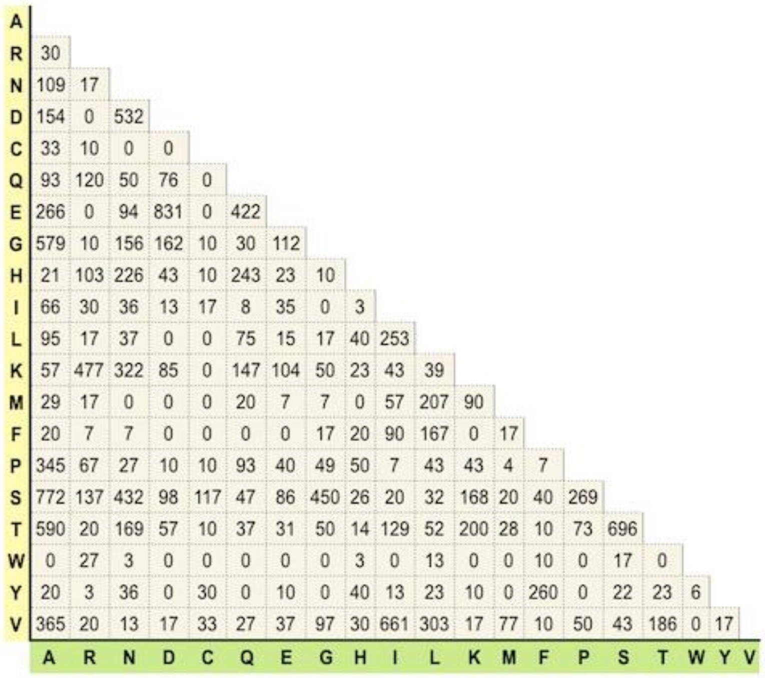
https://bioinformaticshome.com/bioinformatics_tutorials/sequence_alignment/substitution_matrices_page2.html

PAM

A G L L

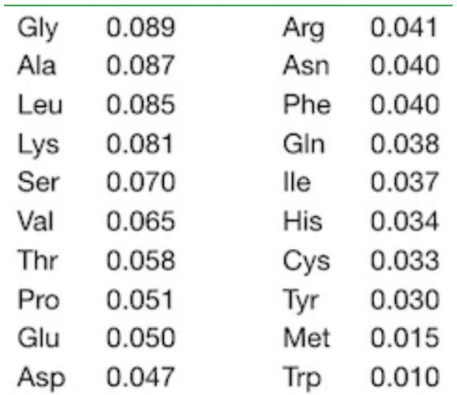
A G A V

Amino acids:	A	G	L	V
Changes:	1	0	2	1
Frequency of occurrence:	3	2	2	1
Relative mutability:	0.33	0	1	1



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	50	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	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

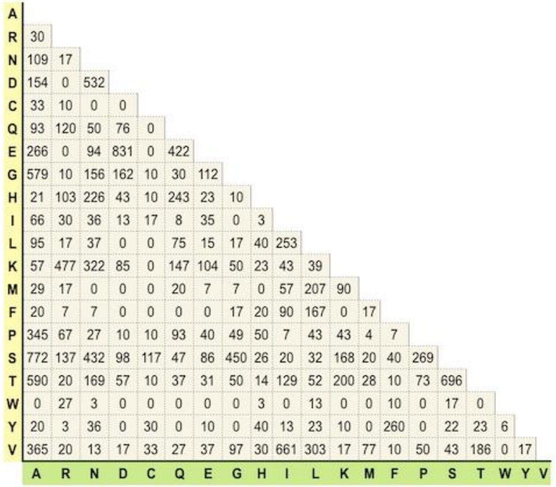
The total count of 1,572 accepted point mutations from 71 evolutionary trees. The displayed counts are original counts times 10.



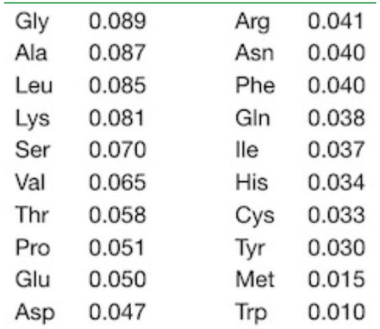
Gly	0.089	Arg	0.041
Ala	0.087	Asn	0.040
Leu	0.085	Phe	0.040
Lys	0.081	Gln	0.038
Ser	0.070	Ile	0.037
Val	0.065	His	0.034
Thr	0.058	Cys	0.033
Pro	0.051	Tyr	0.030
Glu	0.050	Met	0.015
Asp	0.047	Trp	0.010

Normalized Frequencies of the Amino Acids in the Accepted Point Mutation Data

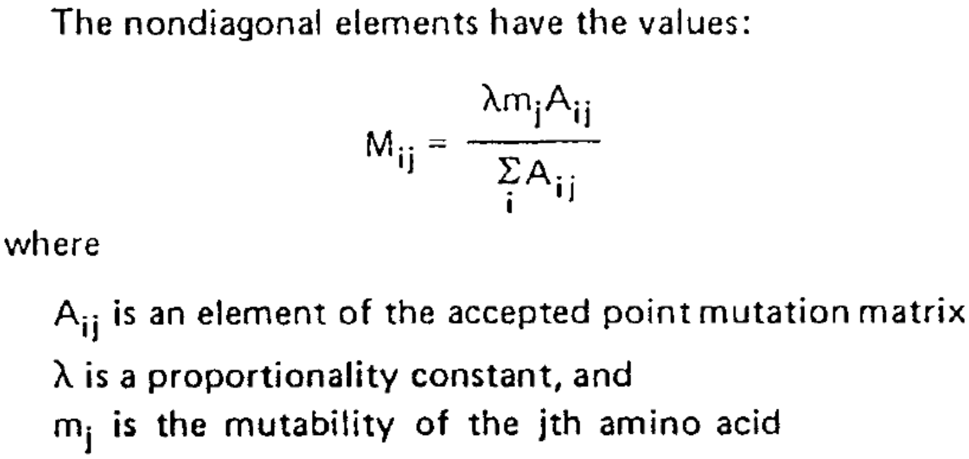
PAM



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
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	50	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	



Gly	0.089	Arg	0.041
Ala	0.087	Asn	0.040
Leu	0.085	Phe	0.040
Lys	0.081	Gln	0.038
Ser	0.070	Ile	0.037
Val	0.065	His	0.034
Thr	0.058	Cys	0.033
Pro	0.051	Tyr	0.030
Glu	0.050	Met	0.015
Asp	0.047	Trp	0.010



The nondiagonal elements have the values:

$$M_{ij} = \frac{\lambda m_j A_{ij}}{\sum_i A_{ij}}$$

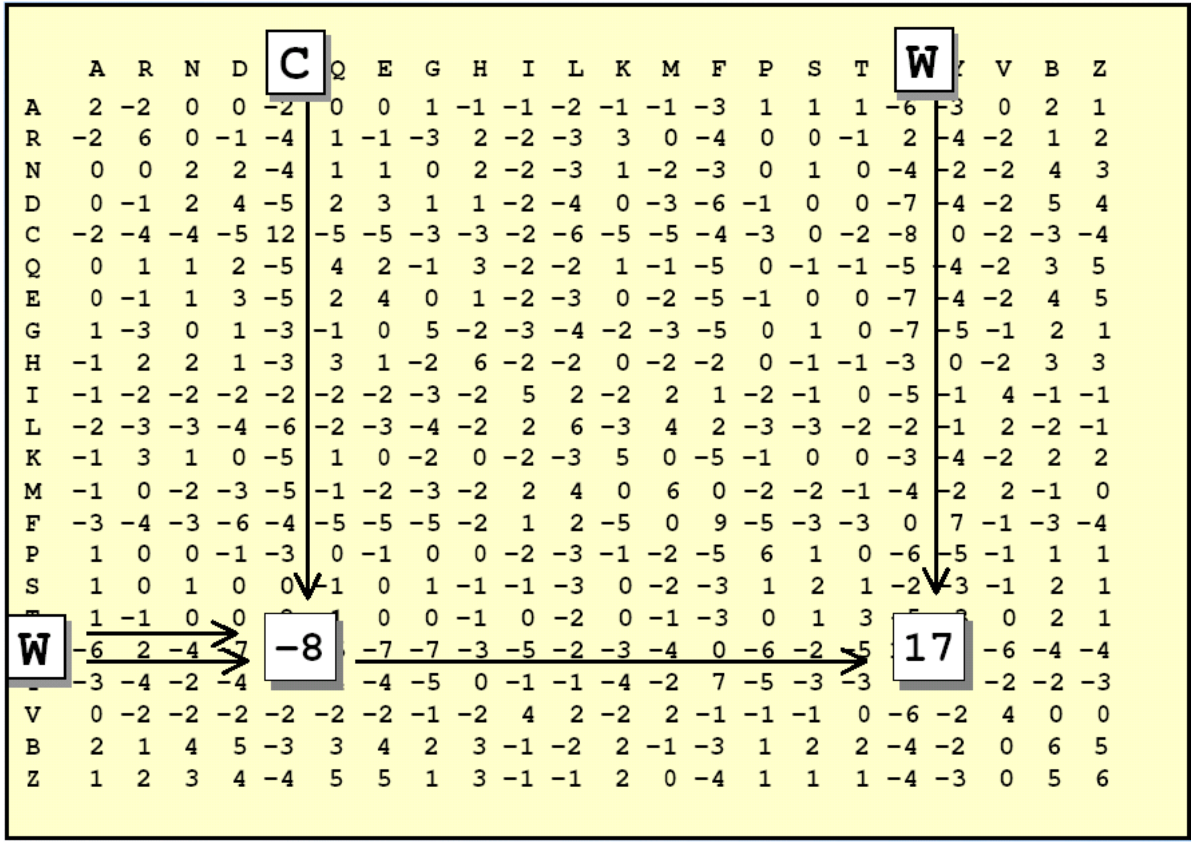
where

A_{ij} is an element of the accepted point mutation matrix

λ is a proportionality constant, and

m_j is the mutability of the jth amino acid

PAM 250

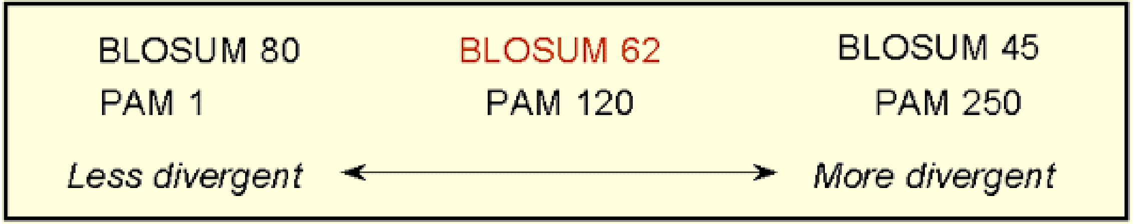


	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z
A	2	−2	0	0	−2	0	0	1	−1	−1	−2	−1	−1	−3	1	1	1	−6	−3	0	2	1
R	−2	6	0	−1	−4	1	−1	−3	2	−2	−3	3	0	−4	0	0	−1	2	−4	−2	1	2
N	0	0	2	2	−4	1	1	0	2	−2	−3	1	−2	−3	0	1	0	−4	−2	−2	4	3
D	0	−1	2	4	−5	2	3	1	1	−2	−4	0	−3	−6	−1	0	0	−7	−4	−2	5	4
C	−2	−4	−4	−5	12	−5	−5	−3	−3	−2	−6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

BLOSUM matrices

- Henikoff 1991
- Aligned ungapped regions from protein families from the BLOCKS database
- The BLOCKS database contain short protein sequences of high similarity clustered together
- Sequences were clustered whenever the % identify exceeded some percentage level
- Calculated the frequency of any two residues being aligned in one cluster also being aligned in another
- Resulted in the fraction of observed substitutions between any two residues over all observed substitutions
- The resulting matrices are numbered inversely from the PAM matrices so the BLOSUM50 matrix was based on clusters of sequence over 50% identity, and BLOSUM62 where the clusters were at least 62% identical

BLOSUM 62



BLOSUM 80 PAM 1

BLOSUM 62 PAM 120

BLOSUM 45 PAM 250

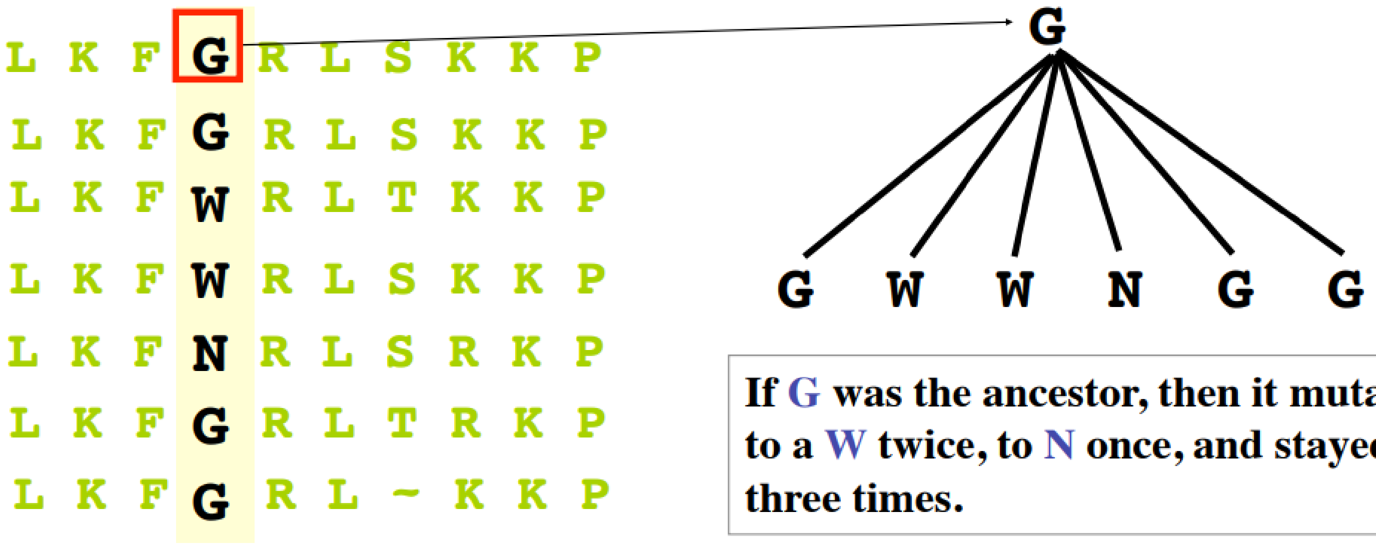
Less divergent ←→ *More divergent*

Ala	4																			
Arg	−1	5																		
Asn	−2	0	6																	
Asp	−2	−2	1	6																
Cys	0	−3	−3	−3	9															
Gln	−1	1	0	0	−3	5														
Glu	−1	0	0	2	−4	2	5													
Gly	0	−2	0	−1	−3	−2	−2	6												
His	−2	0	1	−1	−3	0	0	−2	8											
Ile	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4
	Ala	**Arg**	**Asn**	**Asp**	**Cys**	**Gln**	**Glu**	**Gly**	**His**	**Ile**	**Leu**	**Lys**	**Met**	**Phe**	**Pro**	**Ser**	**Thr**	**Trp**	**Tyr**	**Val**

Probability ratios are expressed as log odds

Calculation of log odd ratios

- Counting mutations without knowing ancestral sequences
- Assume any of the characters could be the ancestral one. Assume equal distance to the ancestor from each taxon.



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

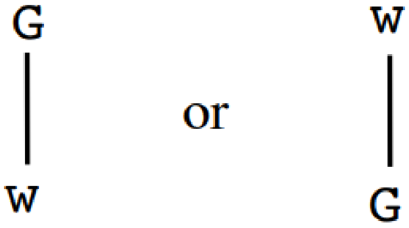
L K F G R L T R K P

L K F G R L ~ K K P

If G was the ancestor, then it mutated to a W twice, to N once, and stayed G three times.

Calculation of log odd ratios

- Substitution matrices are symmetrical
- Since we don't know which sequence came first, we don't know whether



G

W

or

W

G

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

- P(G-->W) and P(W-->G) are the same number.

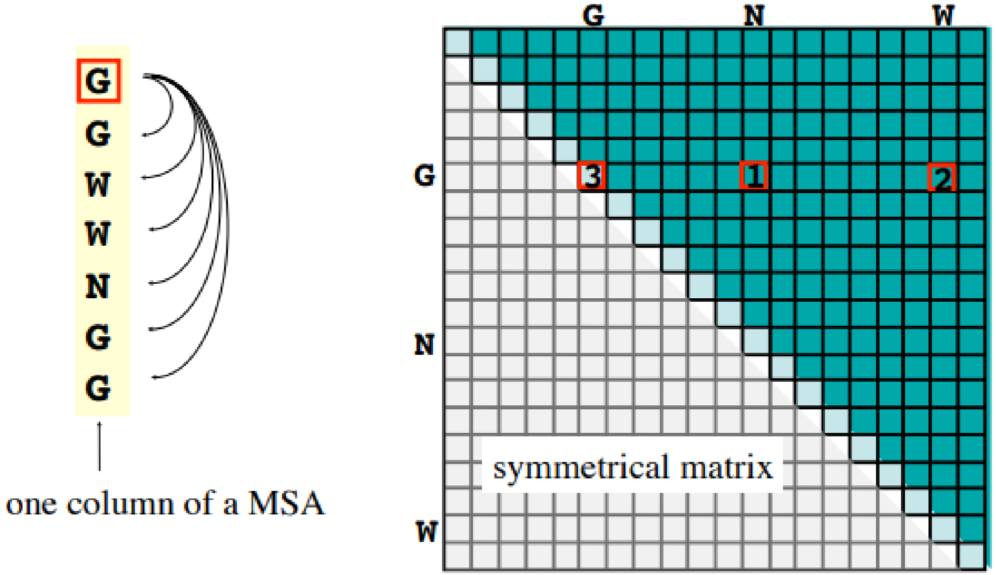
(That's why we only show the upper triangle)

Calculation of log odd ratios

Q: What is the probability of amino acid X mutating to amino acid Z?

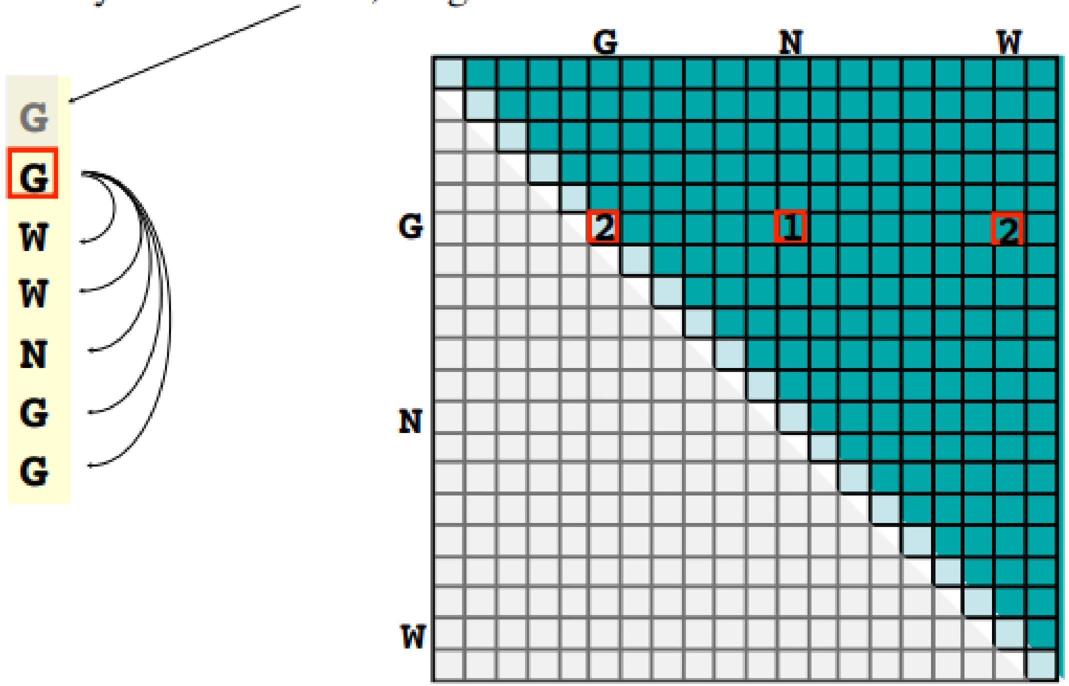
Summing the substitution counts

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



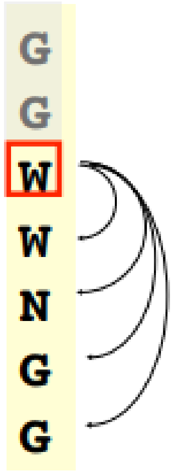
Calculation of log odd ratios

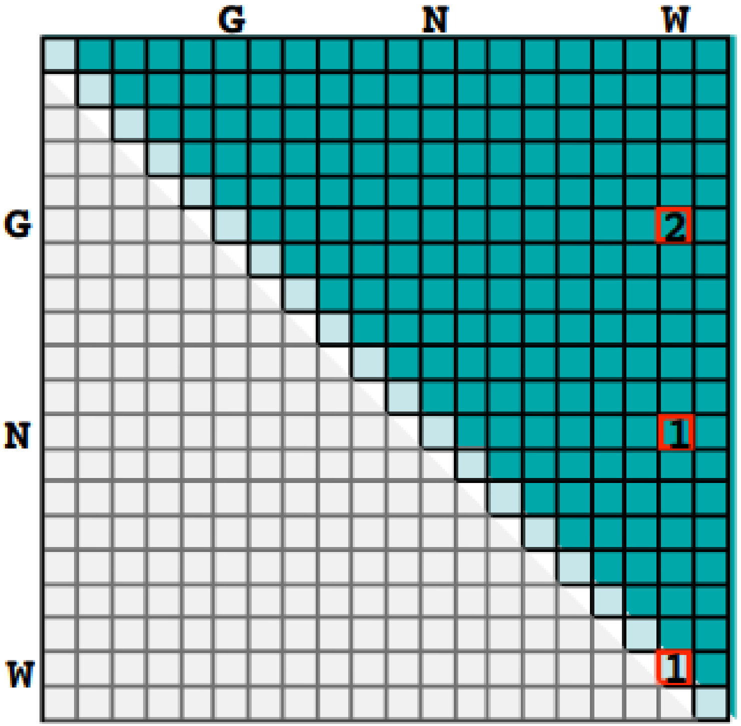
Next possible ancestor, G again.



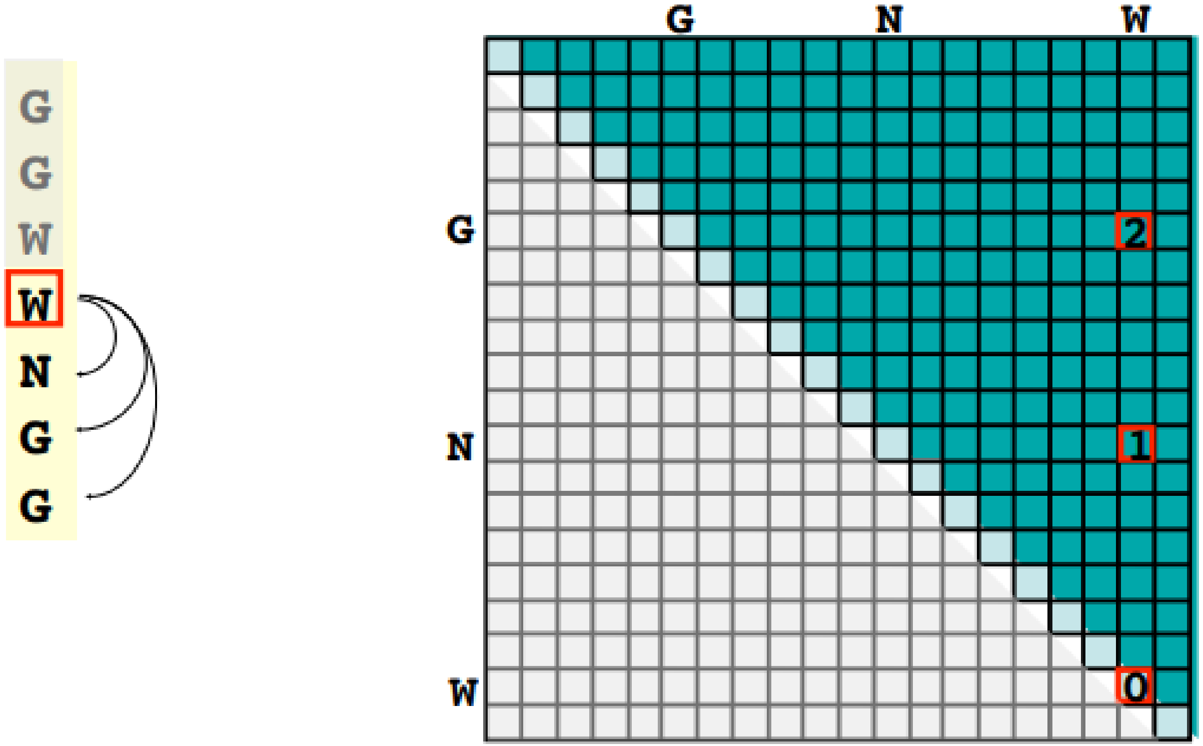
We already counted this G, so ignore it.

Calculation of log odd ratios

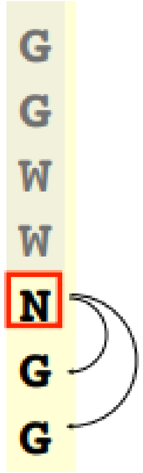


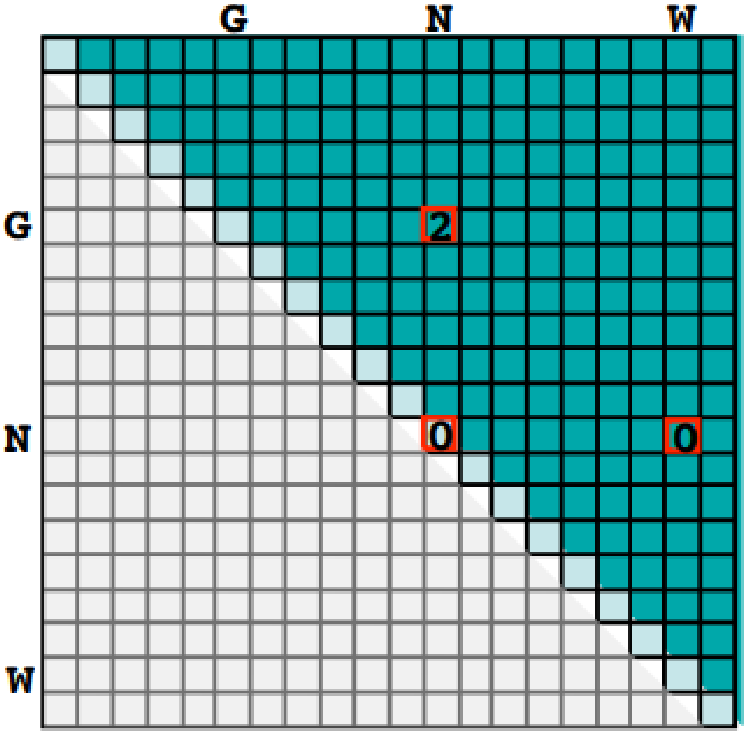


Calculation of log odd ratios



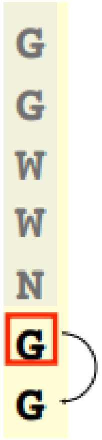
Calculation of log odd ratios



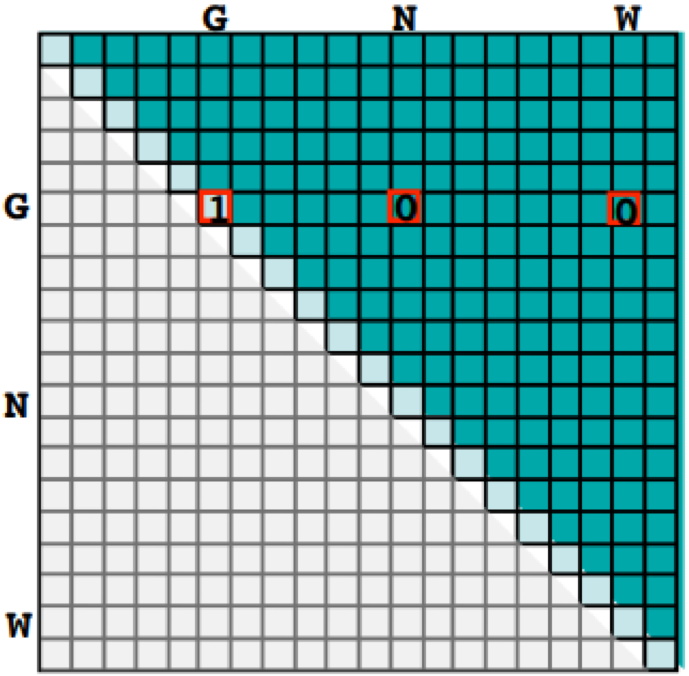


Calculation of log odd ratios

Next...G again



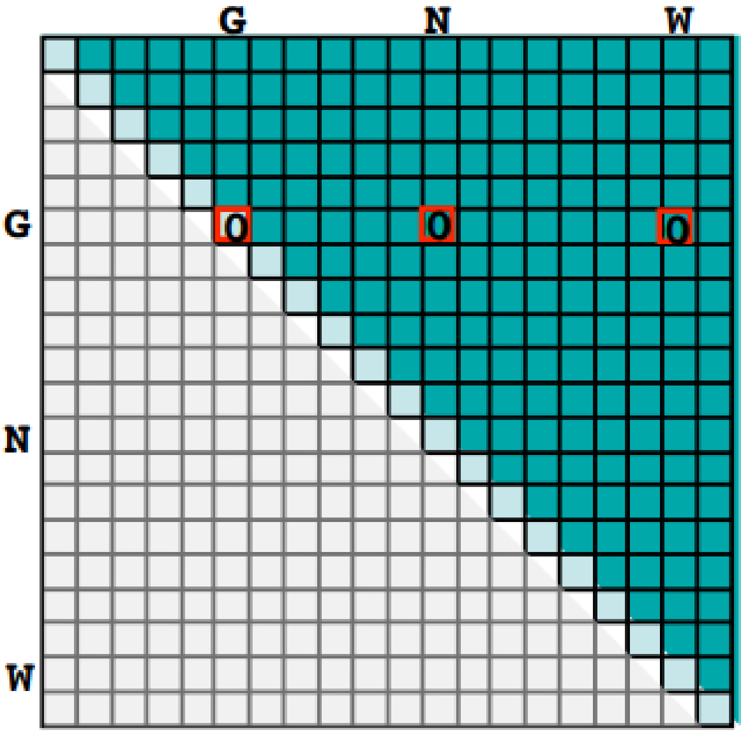
Counting **G** as the ancestor many times as it appears recognizes the increased likelihood that **G** (the most frequent aa at this position) is the true ancestor.



Calculation of log odd ratios

G
G
W
W
N
G
G

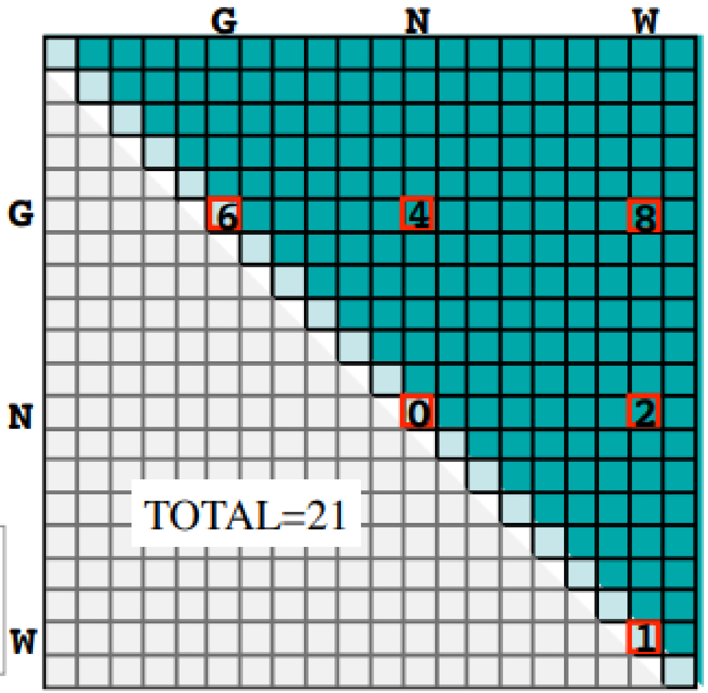
(no counts for last seq.)



Calculation of log odd ratios

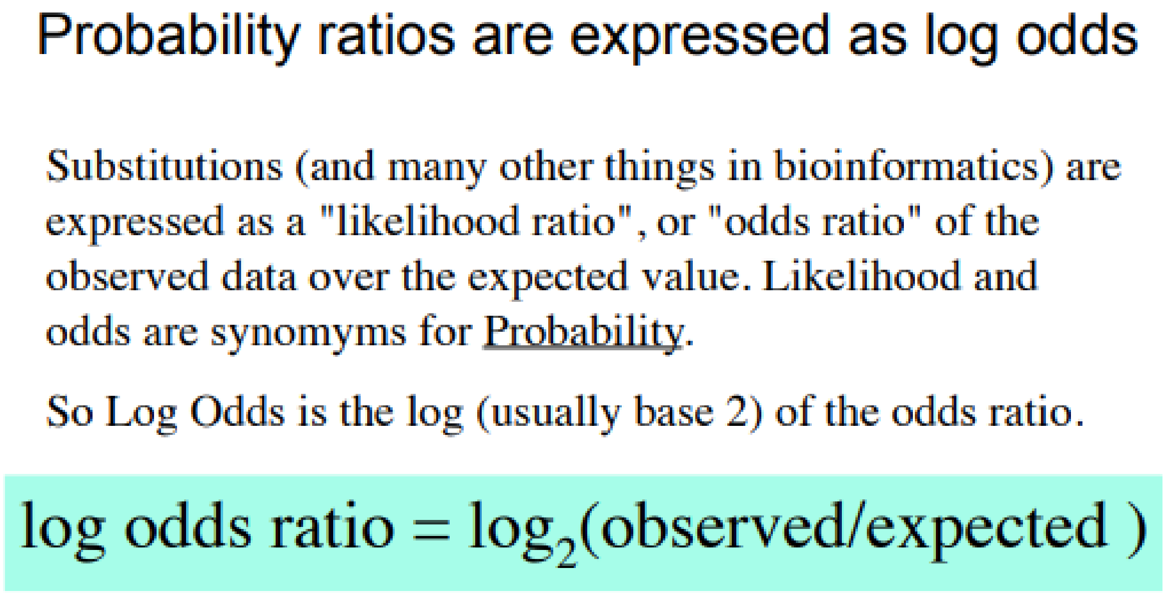
Go to next column. Continue summing.

```
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...

Calculation of log odd ratios



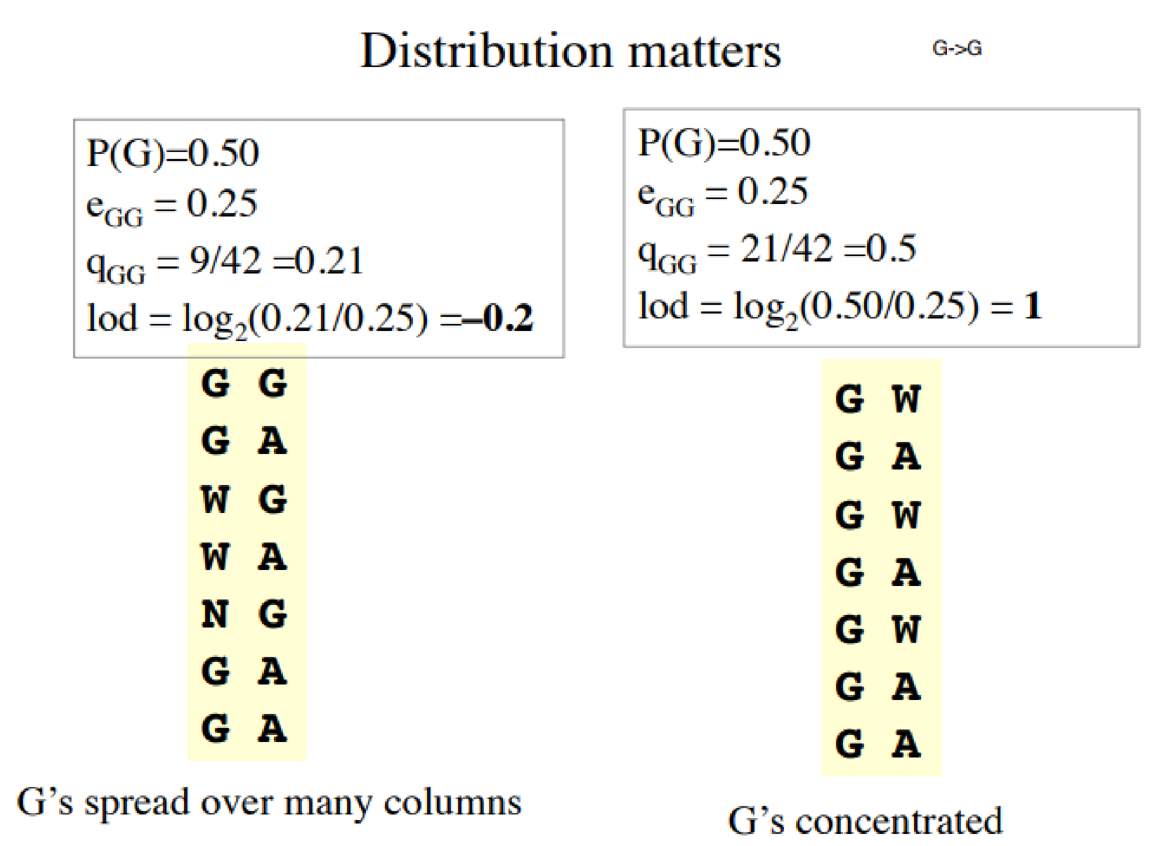
Probability ratios are expressed as log odds

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 synomyms for Probability.

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

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

Calculation of log odd ratios



Distribution matters

G->G

P(G)=0.50

$e_{GG} = 0.25$

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

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

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

G’s spread over many columns

P(G)=0.50

$e_{GG} = 0.25$

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

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

```
G W
G A
G W
G A
G W
G A
G A
```

G’s concentrated

Calculation of log odd ratios

Distribution matters

G->W

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

$e_{GW} = 0.07$

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

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

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

G and W seen together more often than expected.

$P(G)=0.50$, $P(W)=0.21$

$e_{GW} = 0.50 \text{ x } 0.21 = 0.105$

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

$\text{lod} = \log_2(0.07/0.105 = -0.58$

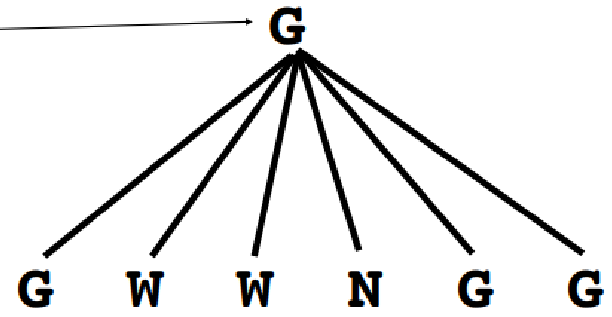
G	W
G	A
G	W
G	A
G	W
G	A
A	G

G’s and W’s not seen together.

Calculation of log odd ratios

- Counting mutations without knowing ancestral sequences
- Assume any of the characters could be the ancestral one. Assume equal distance to the ancestor from each taxon.

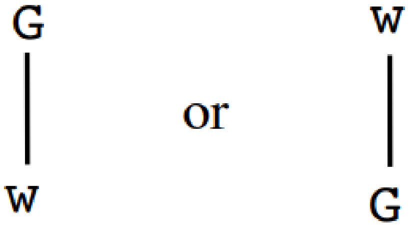
```
L K F G R L S K K P
L K F G R L S K K P
L K F W R L T K K P
L K F W R L S K K P
L K F N R L S R K P
L K F G R L T R K P
L K F G R L ~ K K P
```



If G was the ancestor, then it mutated to a W twice, to N once, and stayed G three times.

Calculation of log odd ratios

- Substitution matrices are symmetrical
- Since we don't know which sequence came first, we don't know whether



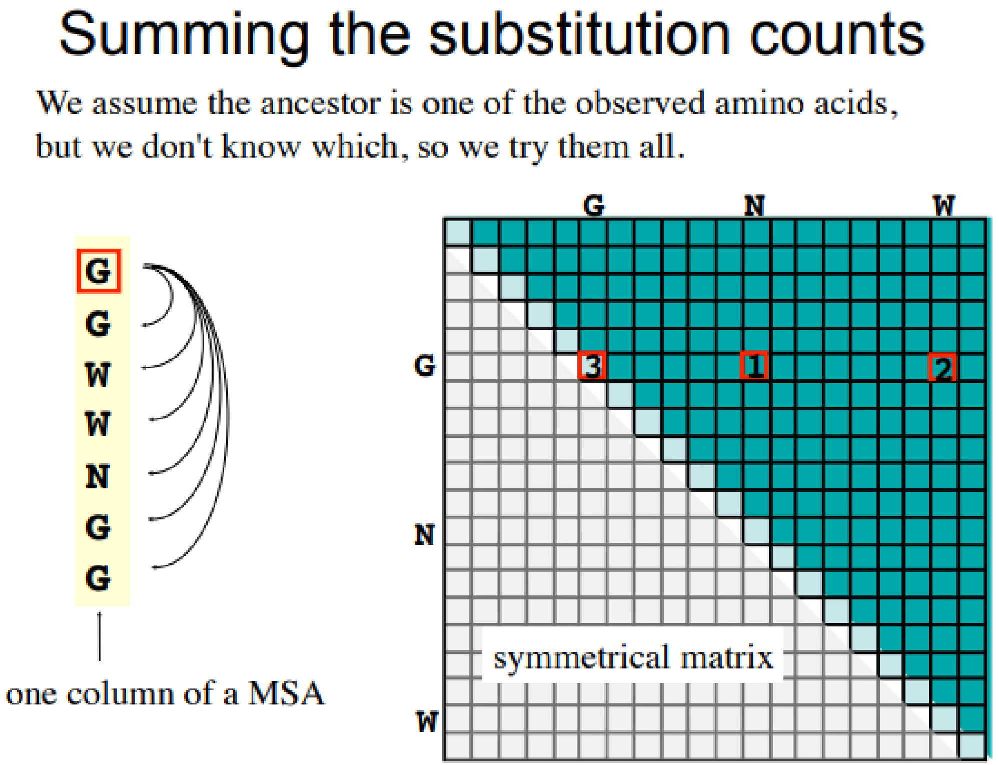
...is correct. So we count this as one mutation of each type.

- P(G-->W) and P(W-->G) are the same number.

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

Calculation of log odd ratios

Q: What is the probability of amino acid X mutating to amino acid Z?



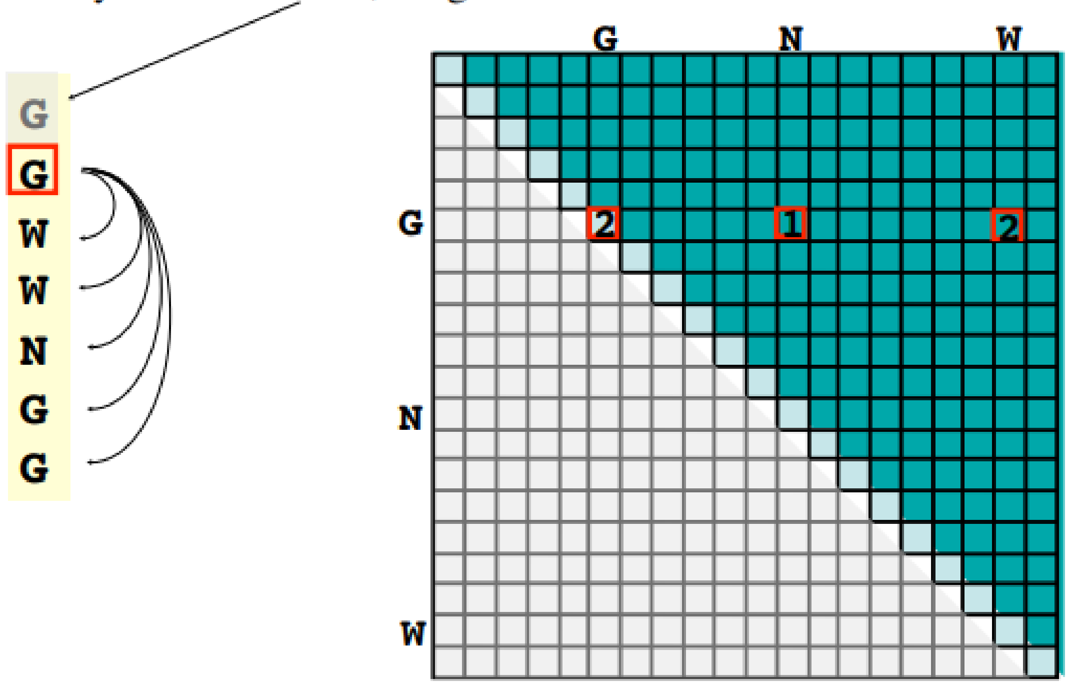
Summing the substitution counts

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

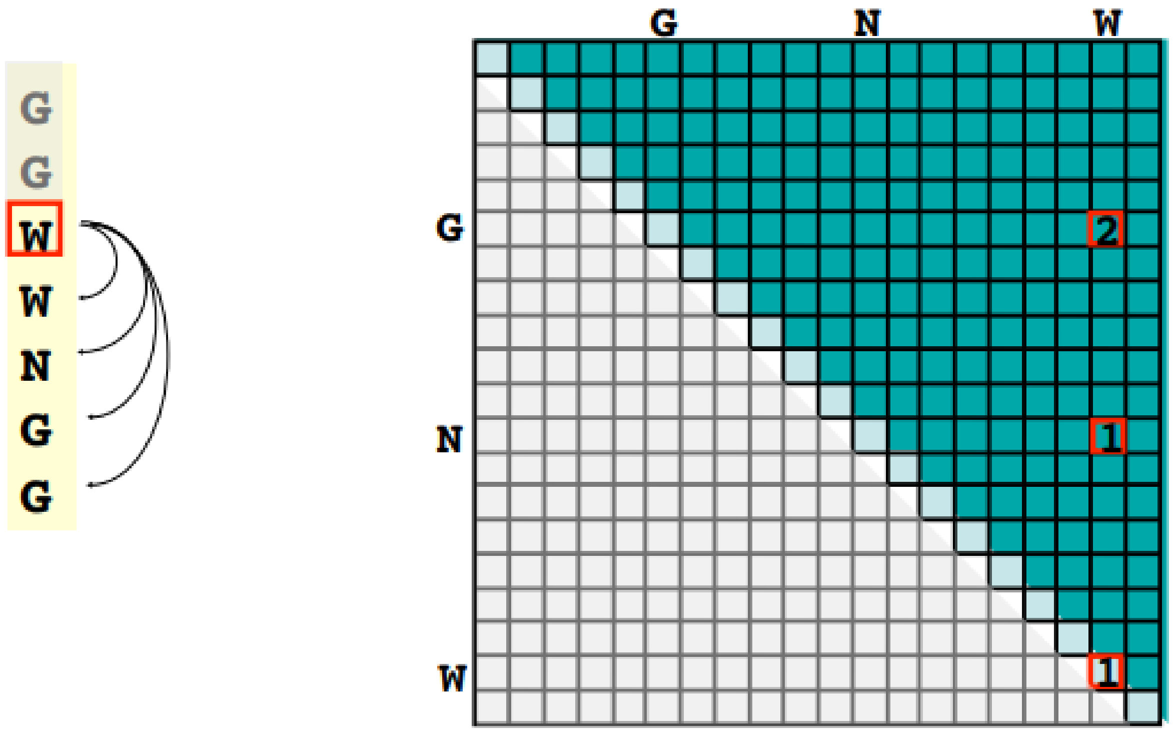
Calculation of log odd ratios

Next possible ancestor, G again.

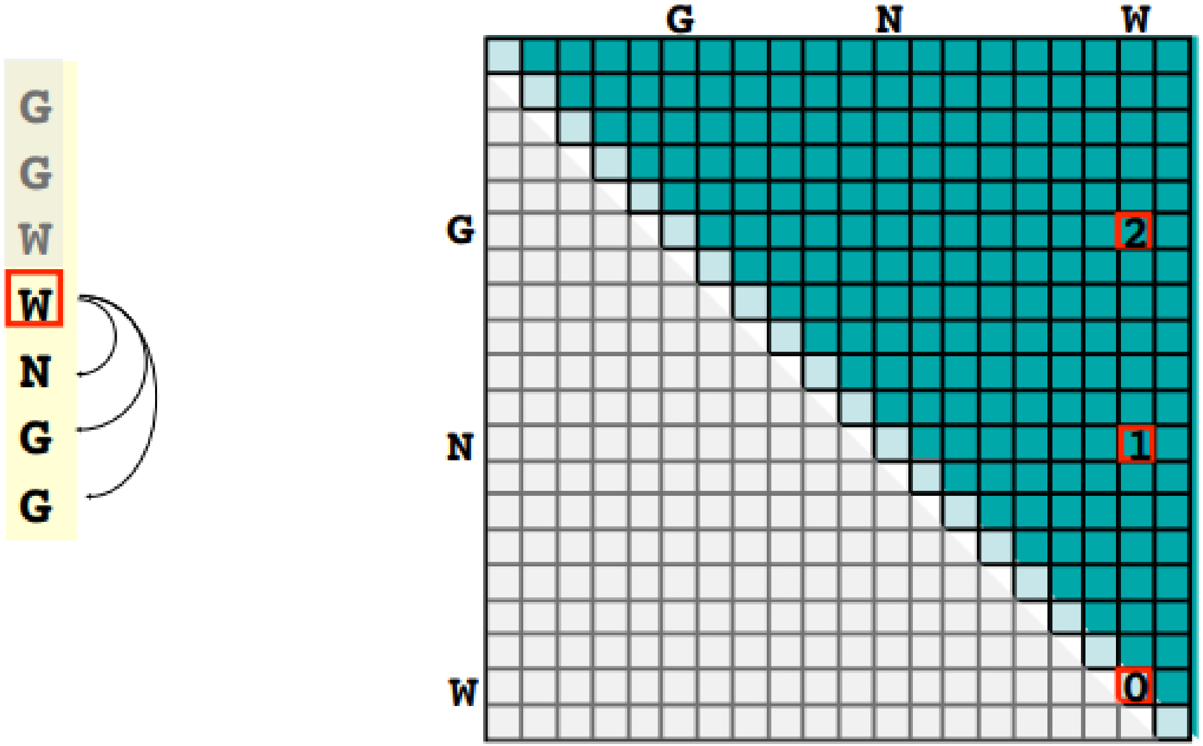
We already counted this G, so ignore it.



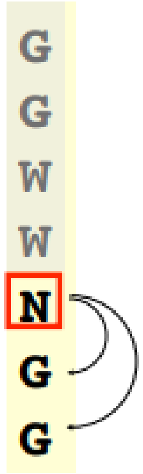
Calculation of log odd ratios

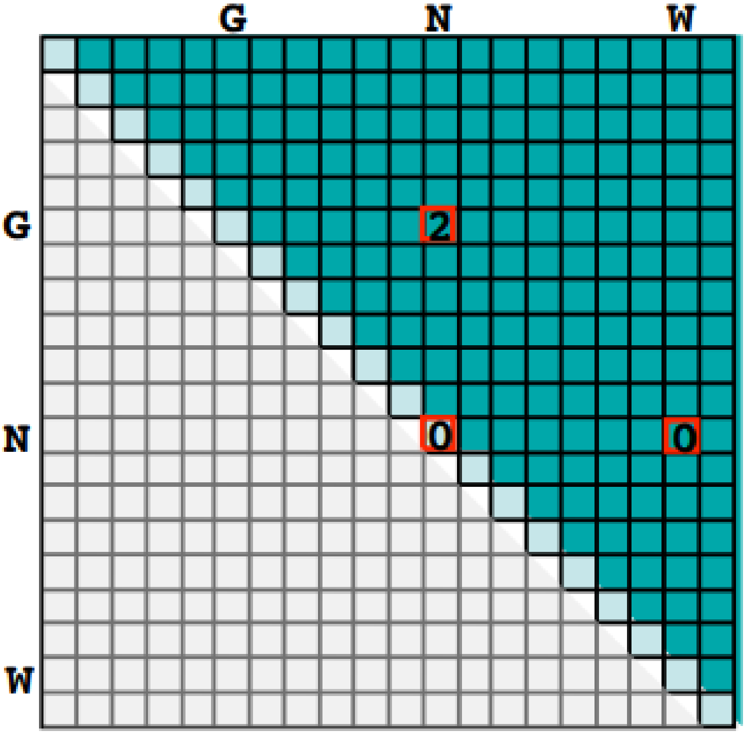


Calculation of log odd ratios



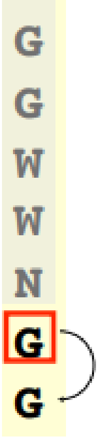
Calculation of log odd ratios



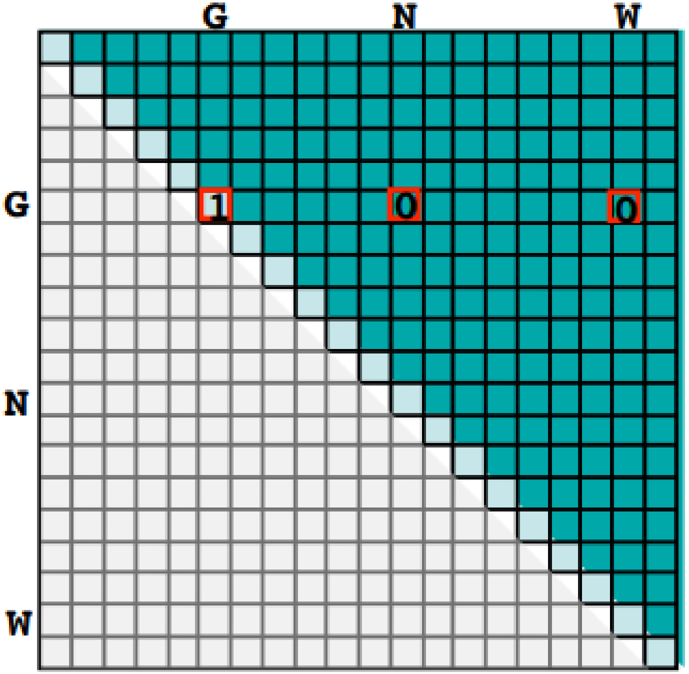


Calculation of log odd ratios

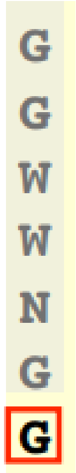
Next...G again



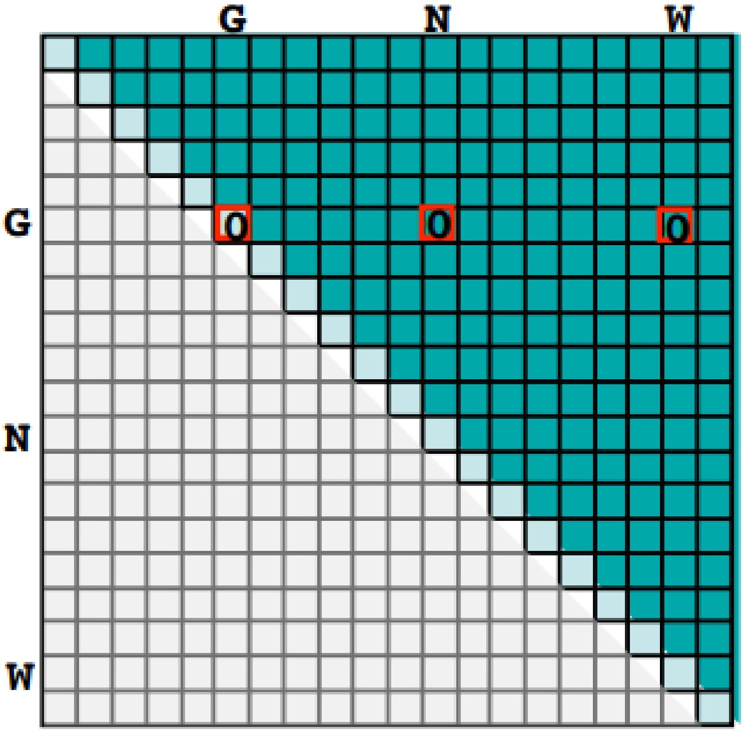
Counting **G** as the ancestor many times as it appears recognizes the increased likelihood that **G** (the most frequent aa at this position) is the true ancestor.



Calculation of log odd ratios



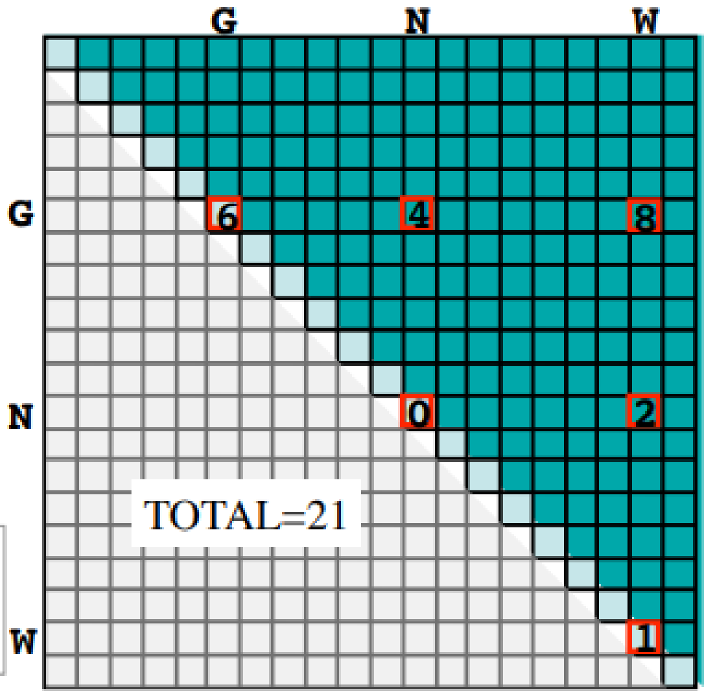
(no counts for last seq.)



Calculation of log odd ratios

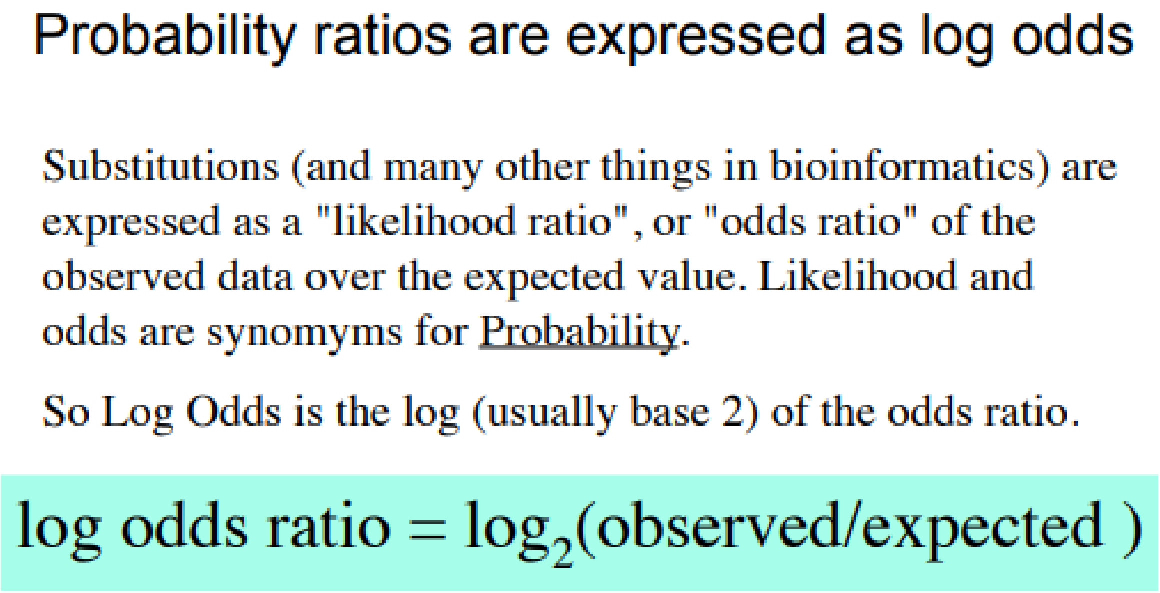
Go to next column. Continue summing.

```
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...

Calculation of log odd ratios



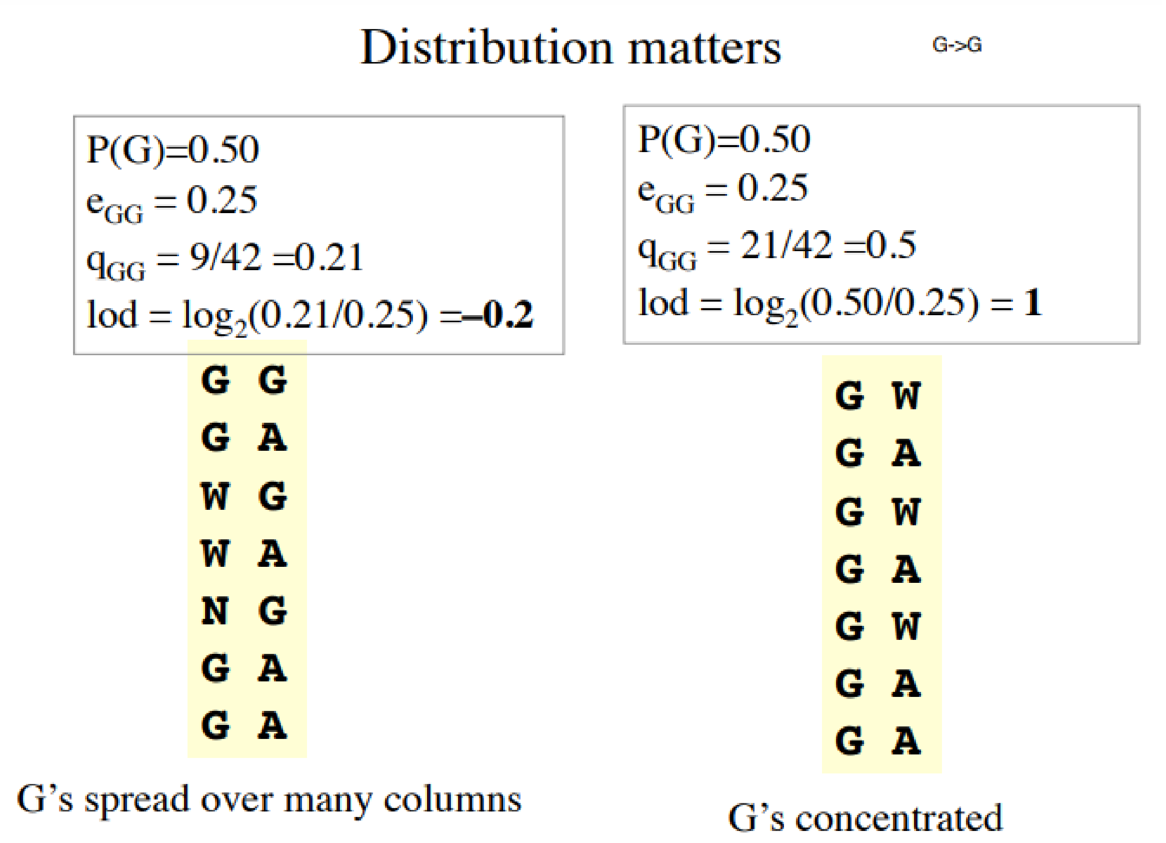
Probability ratios are expressed as log odds

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 synomyms for Probability.

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

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

Calculation of log odd ratios



Distribution matters

G->G

P(G)=0.50

$e_{GG} = 0.25$

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

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

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

G’s spread over many columns

P(G)=0.50

$e_{GG} = 0.25$

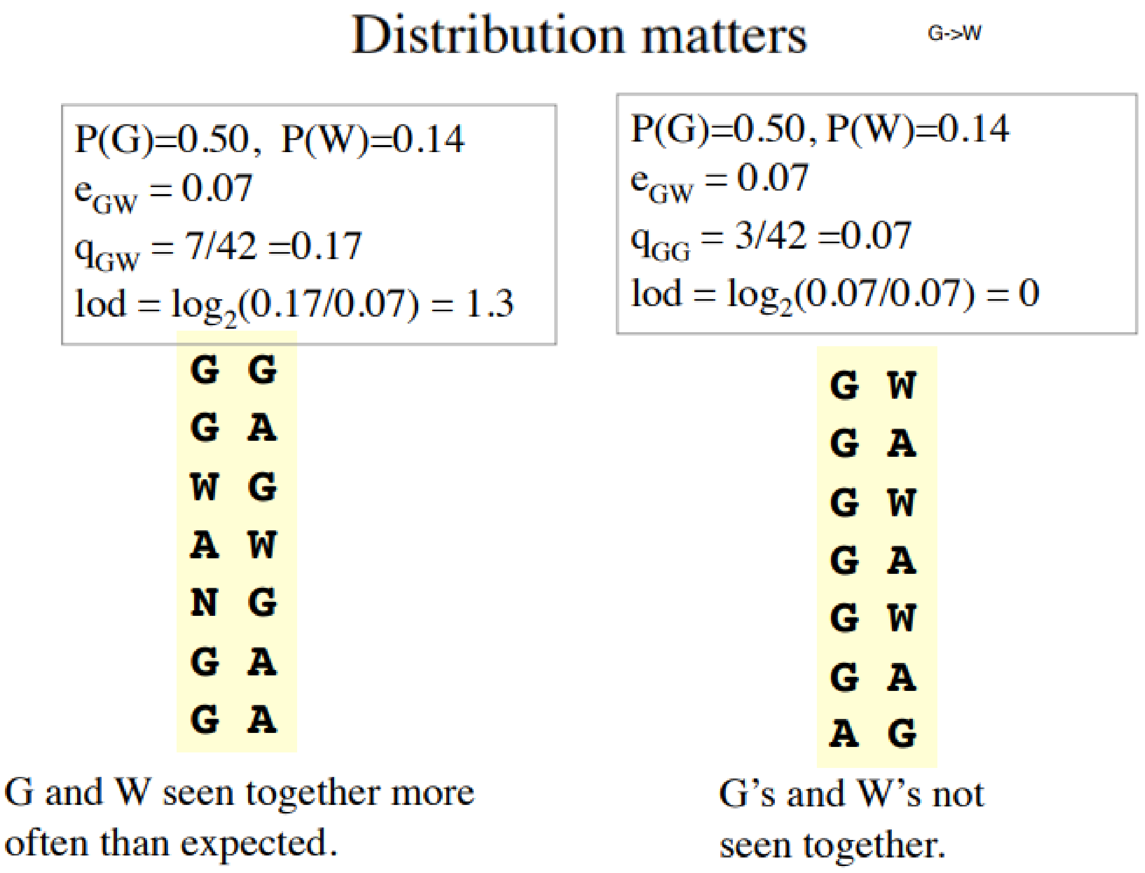
$q_{GG} = 21/42 = 0.5$

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

```
G W
G A
G W
G A
G W
G A
G A
```

G’s concentrated

Calculation of log odd ratios



Distribution matters

G->W

$P(G)=0.50$, $P(W)=0.14$
$e_{GW} = 0.07$
$q_{GW} = 7/42 = 0.17$
lod $= \log_2(0.17/0.07) = 1.3$

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

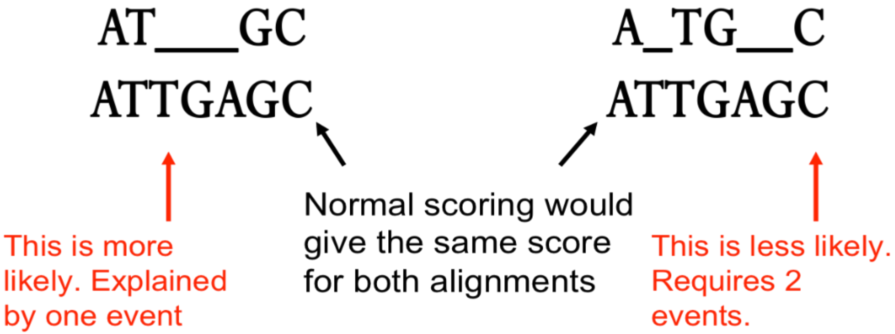
G and W seen together more often than expected.

$P(G)=0.50$, $P(W)=0.14$
$e_{GW} = 0.07$
$q_{GG} = 3/42 = 0.07$
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.

GAP penalty



AT___GC
ATTGAGC

A_TG__C
ATTGAGC

This is more likely. Explained by one event

Normal scoring would give the same score for both alignments

This is less likely. Requires 2 events.

Linear gap penalty score: $\gamma(g) = -\,gd$

Affine gap penalty score: $\gamma(g) = -d - (g - 1)\,e$

$\gamma(g)$ = **gap penalty score of a gap of length g**

d = **gap opening penalty**
e = **gap extension penalty**
g = **gap length**

Scoring insertions or deletions using GAP penalty

Affine gap penalty score: $\gamma(g) = -d - (g - 1)\, e$

Match → 1 and Mismatch → 0

$\gamma(g)$ = gap penalty score of a gap of length g

d = gap opening penalty → -3

e = gap extension penalty → -0.1

g = gap length

Total Score: 4

```
T A T G T G C G T A T A
  | | | |
  A T G T T A T A C
```

Total Score: 8 + (-3.2) = 4.8

```
T A T G T G C G T A T A
  | | | |       | | | |
  A T G T - - - T A T A C
```