Alignment

Lecture 6
Sept 12, 2016

Announcements

- Terminate your instance?

BLAST

Stats

$$E = Kmne^{-\lambda S}$$

BLAST

Stats

$$p = 1 - e^{-E}$$

BLAST

Is my p-value significant?

	H_o true	H_o false
Reject H_o	**Type 1 error (false pos)**	**Correct!**
Accept H_o	Correct!	Type 2 error (false neg)

BLAST null: There is no match between query and database entry

BLAST

Multiple testing correction

http://www.r-fiddle.org/#/

Finding Data

Read data

- http://www.ebi.ac.uk/ena
- http://www.ncbi.nlm.nih.gov/sra
- http://metagenomics.anl.gov/?page=MetagenomeSelect

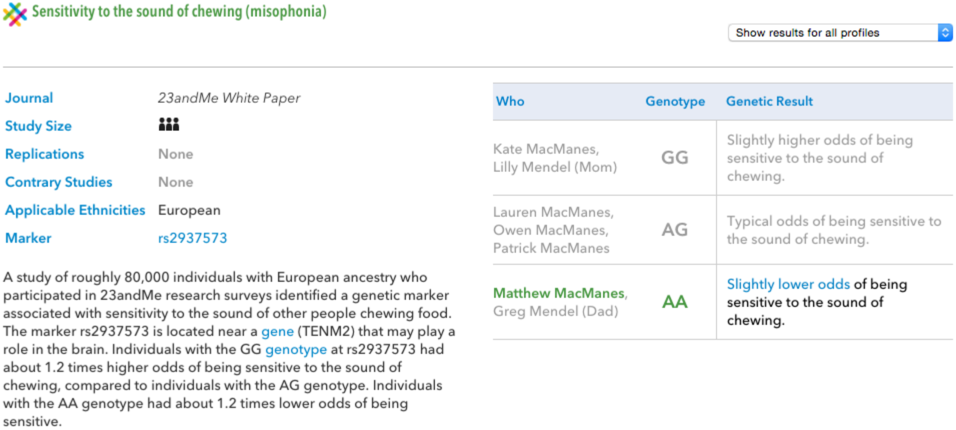
Assembly (and other) Data

- http://useast.ensembl.org/info/data/ftp/index.html
- http://www.ncbi.nlm.nih.gov/genome/
- http://datadryad.org/
- http://figshare.com/

Finding Data

Human Stuff

- http://www.ncbi.nlm.nih.gov/clinvar/
- http://www.ncbi.nlm.nih.gov/omim
- http://snpedia.com/index.php/SNPedia



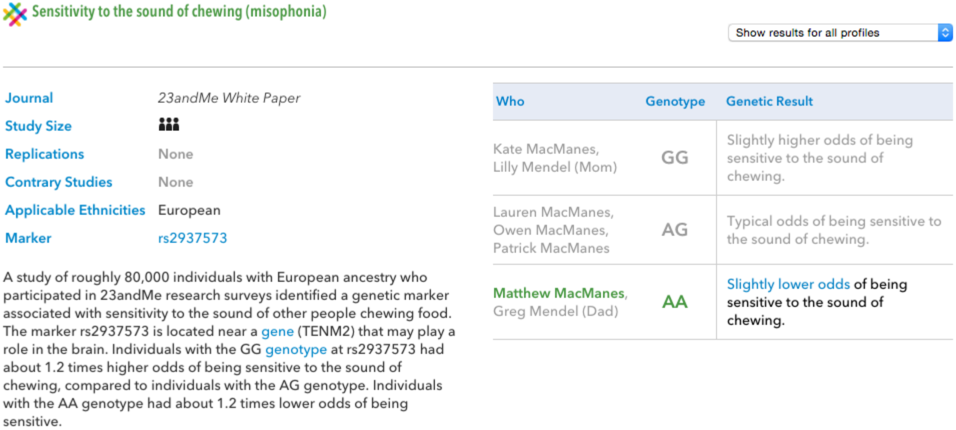
Sensitivity to the sound of chewing (misophonia)

Show results for all profiles

Journal	*23andMe White Paper*
Study Size	👤👤👤
Replications	None
Contrary Studies	None
Applicable Ethnicities	European
Marker	rs2937573

A study of roughly 80,000 individuals with European ancestry who participated in 23andMe research surveys identified a genetic marker associated with sensitivity to the sound of other people chewing food. The marker rs2937573 is located near a gene (TENM2) that may play a role in the brain. Individuals with the GG genotype at rs2937573 had about 1.2 times higher odds of being sensitive to the sound of chewing, compared to individuals with the AG genotype. Individuals with the AA genotype had about 1.2 times lower odds of being sensitive.

Who	Genotype	Genetic Result
Kate MacManes, Lilly Mendel (Mom)	GG	Slightly higher odds of being sensitive to the sound of chewing.
Lauren MacManes, Owen MacManes, Patrick MacManes	AG	Typical odds of being sensitive to the sound of chewing.
Matthew MacManes, Greg Mendel (Dad)	AA	Slightly lower odds of being sensitive to the sound of chewing.



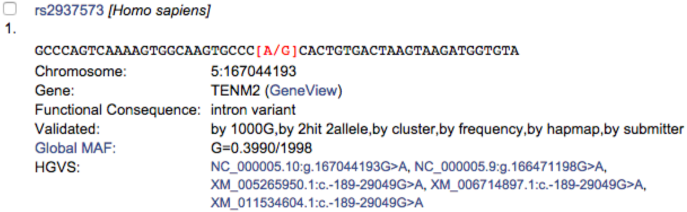
Sensitivity to the sound of chewing (misophonia)

Show results for all profiles

Journal	*23andMe White Paper*
Study Size	👤👤👤
Replications	None
Contrary Studies	None
Applicable Ethnicities	European
Marker	rs2937573

A study of roughly 80,000 individuals with European ancestry who participated in 23andMe research surveys identified a genetic marker associated with sensitivity to the sound of other people chewing food. The marker rs2937573 is located near a gene (TENM2) that may play a role in the brain. Individuals with the GG genotype at rs2937573 had about 1.2 times higher odds of being sensitive to the sound of chewing, compared to individuals with the AG genotype. Individuals with the AA genotype had about 1.2 times lower odds of being sensitive.

Who	Genotype	Genetic Result
Kate MacManes, Lilly Mendel (Mom)	GG	Slightly higher odds of being sensitive to the sound of chewing.
Lauren MacManes, Owen MacManes, Patrick MacManes	AG	Typical odds of being sensitive to the sound of chewing.
Matthew MacManes, Greg Mendel (Dad)	AA	Slightly lower odds of being sensitive to the sound of chewing.

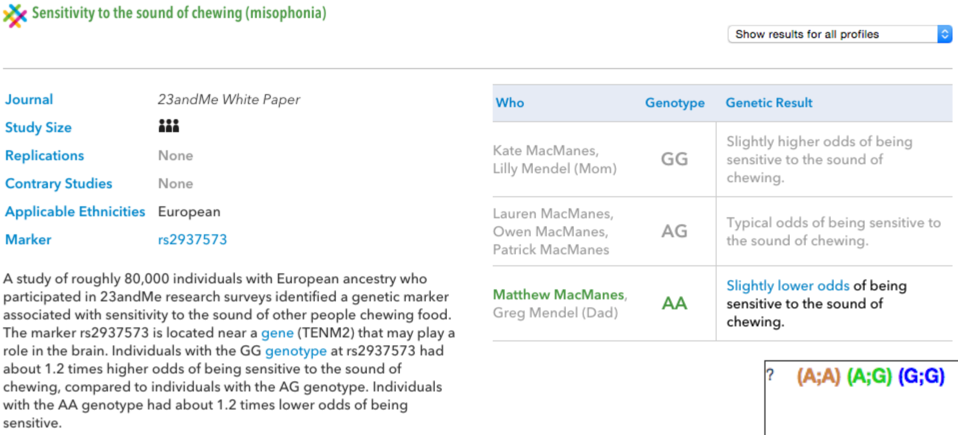


□ rs2937573 *[Homo sapiens]*

1.

```
GCCCAGTCAAAAGTGGCAAGTGCCC[A/G]CACTGTGACTAAGTAAGATGGTGTA
```

Chromosome:	5:167044193
Gene:	TENM2 (GeneView)
Functional Consequence:	intron variant
Validated:	by 1000G,by 2hit 2allele,by cluster,by frequency,by hapmap,by submitter
Global MAF:	G=0.3990/1998
HGVS:	NC_000005.10:g.167044193G>A, NC_000005.9:g.166471198G>A, XM_005265950.1:c.-189-29049G>A, XM_006714897.1:c.-189-29049G>A, XM_011534604.1:c.-189-29049G>A



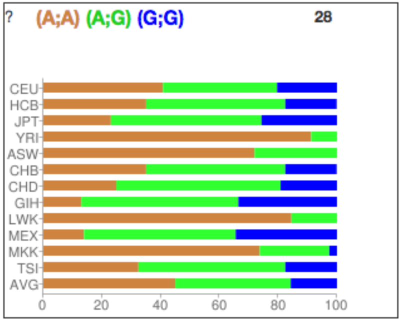
Sensitivity to the sound of chewing (misophonia)

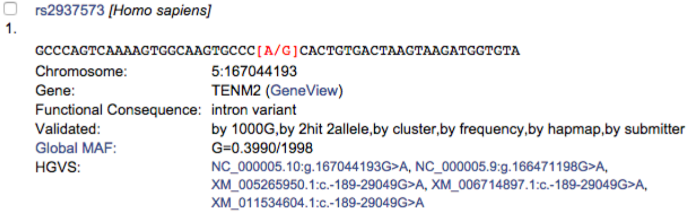
Show results for all profiles

Journal	*23andMe White Paper*
Study Size	👤👤👤
Replications	None
Contrary Studies	None
Applicable Ethnicities	European
Marker	rs2937573

Who	Genotype	Genetic Result
Kate MacManes, Lilly Mendel (Mom)	GG	Slightly higher odds of being sensitive to the sound of chewing.
Lauren MacManes, Owen MacManes, Patrick MacManes	AG	Typical odds of being sensitive to the sound of chewing.
Matthew MacManes, Greg Mendel (Dad)	AA	Slightly lower odds of being sensitive to the sound of chewing.

A study of roughly 80,000 individuals with European ancestry who participated in 23andMe research surveys identified a genetic marker associated with sensitivity to the sound of other people chewing food. The marker rs2937573 is located near a gene (TENM2) that may play a role in the brain. Individuals with the GG genotype at rs2937573 had about 1.2 times higher odds of being sensitive to the sound of chewing, compared to individuals with the AG genotype. Individuals with the AA genotype had about 1.2 times lower odds of being sensitive.





□ 1. rs2937573 *[Homo sapiens]*

```
GCCCAGTCAAAAGTGGCAAGTGCCC[A/G]CACTGTGACTAAGTAAGATGGTGTA
```

Chromosome:	5:167044193
Gene:	TENM2 (GeneView)
Functional Consequence:	intron variant
Validated:	by 1000G,by 2hit 2allele,by cluster,by frequency,by hapmap,by submitter
Global MAF:	G=0.3990/1998
HGVS:	NC_000005.10:g.167044193G>A, NC_000005.9:g.166471198G>A, XM_005265950.1:c.-189-29049G>A, XM_006714897.1:c.-189-29049G>A, XM_011534604.1:c.-189-29049G>A

Pairwise Alignment

What is alignment?

Pairwise Alignment

What is aligner actually doing?

Scoring Matrices

DAYHOFF'S MATRIX

Dayhoff Matrix

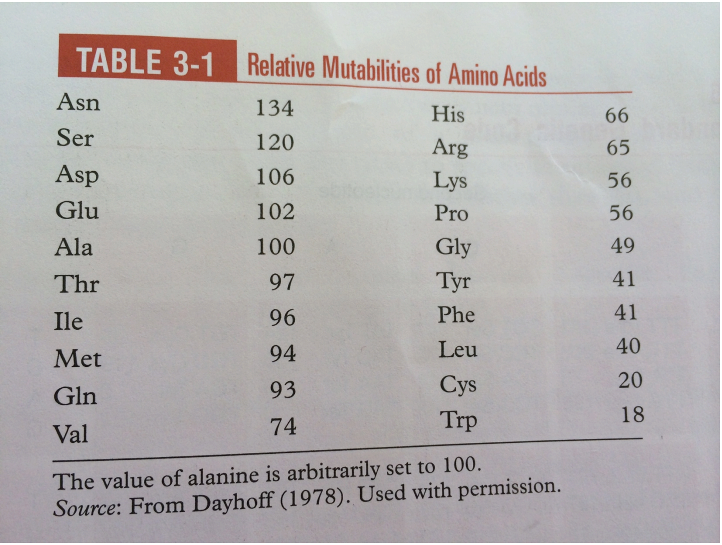


TABLE 3-1 Relative Mutabilities of Amino Acids

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

The value of alanine is arbitrarily set to 100.
Source: From Dayhoff (1978). Used with permission.

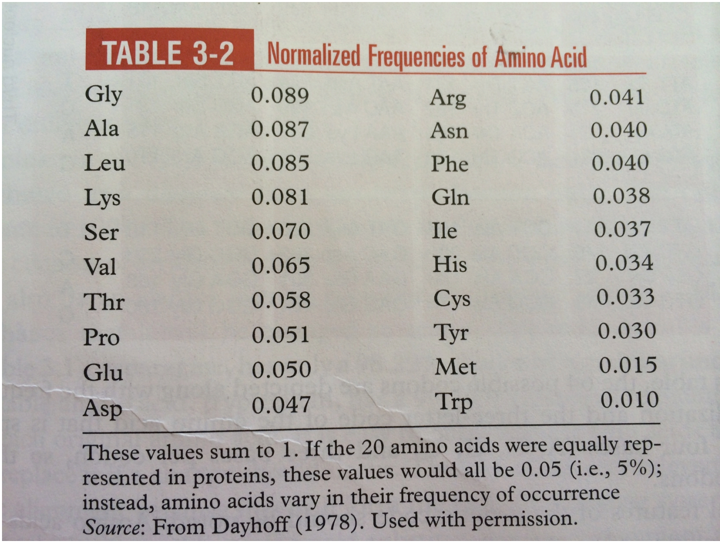
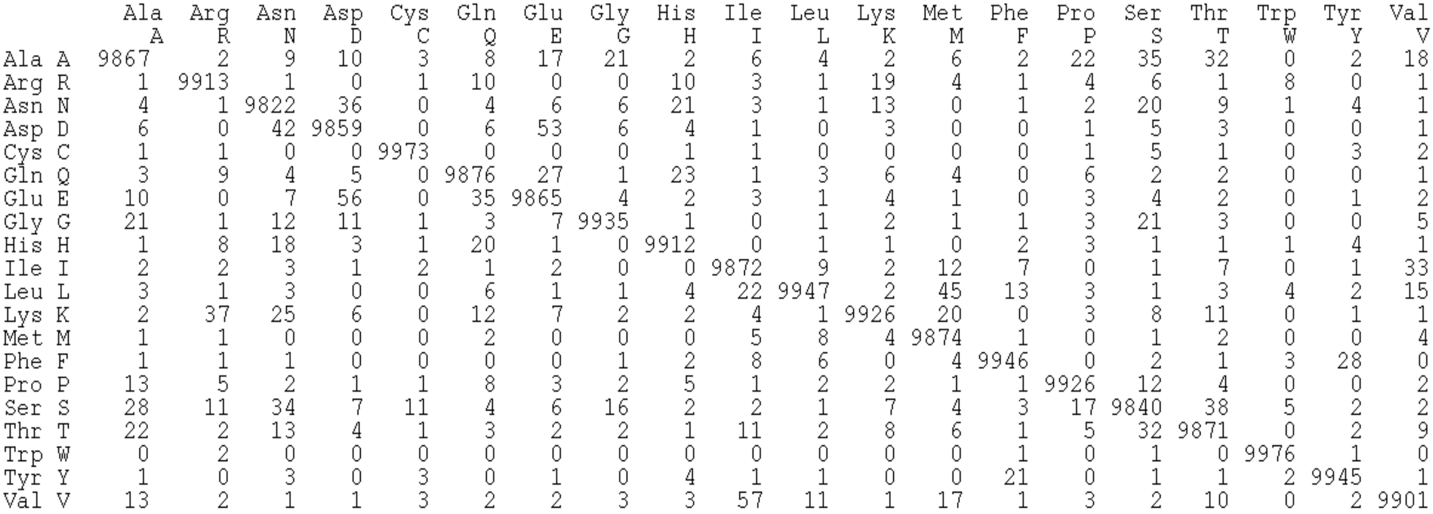


TABLE 3-2 Normalized Frequencies of Amino Acid

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

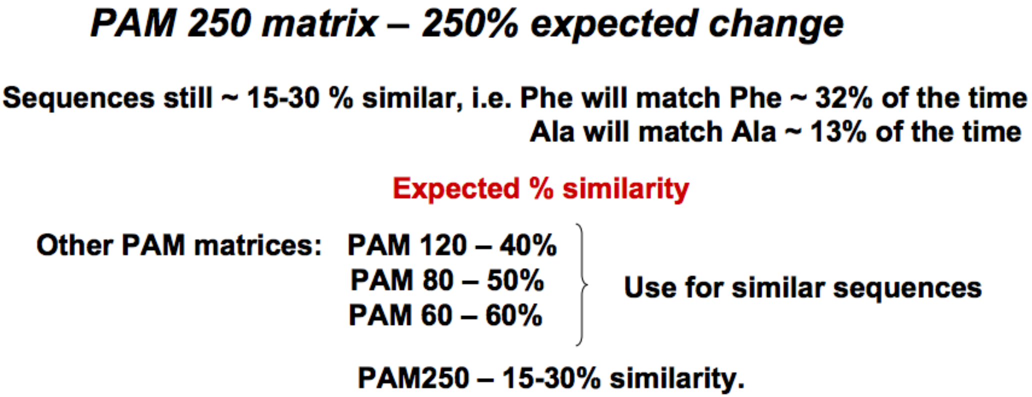
These values sum to 1. If the 20 amino acids were equally represented in proteins, these values would all be 0.05 (i.e., 5%); instead, amino acids vary in their frequency of occurrence
Source: From Dayhoff (1978). Used with permission.

PAM1 MATRIX



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

OTHER PAM MATRIX



PAM 250 matrix – 250% expected change

Sequences still ~ 15-30 % similar, i.e. Phe will match Phe ~ 32% of the time
Ala will match Ala ~ 13% of the time

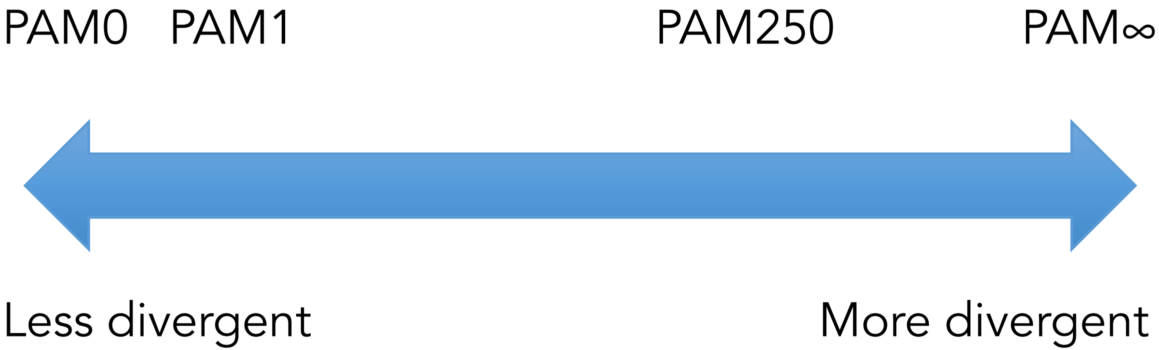
Expected % similarity

Other PAM matrices: PAM 120 – 40%
PAM 80 – 50%
PAM 60 – 60%

Use for similar sequences

PAM250 – 15-30% similarity.

Pam versus Divergence

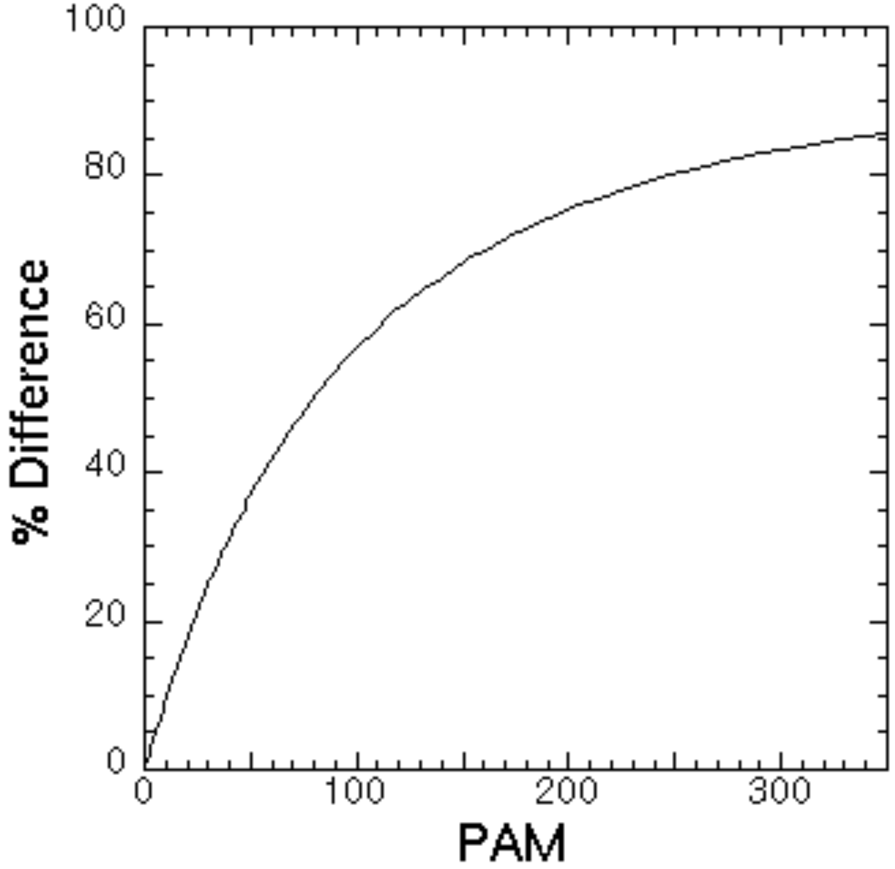


PAM0 PAM1 PAM250 PAM∞

Less divergent More divergent

http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp

PAM versus Divergence



PAM250

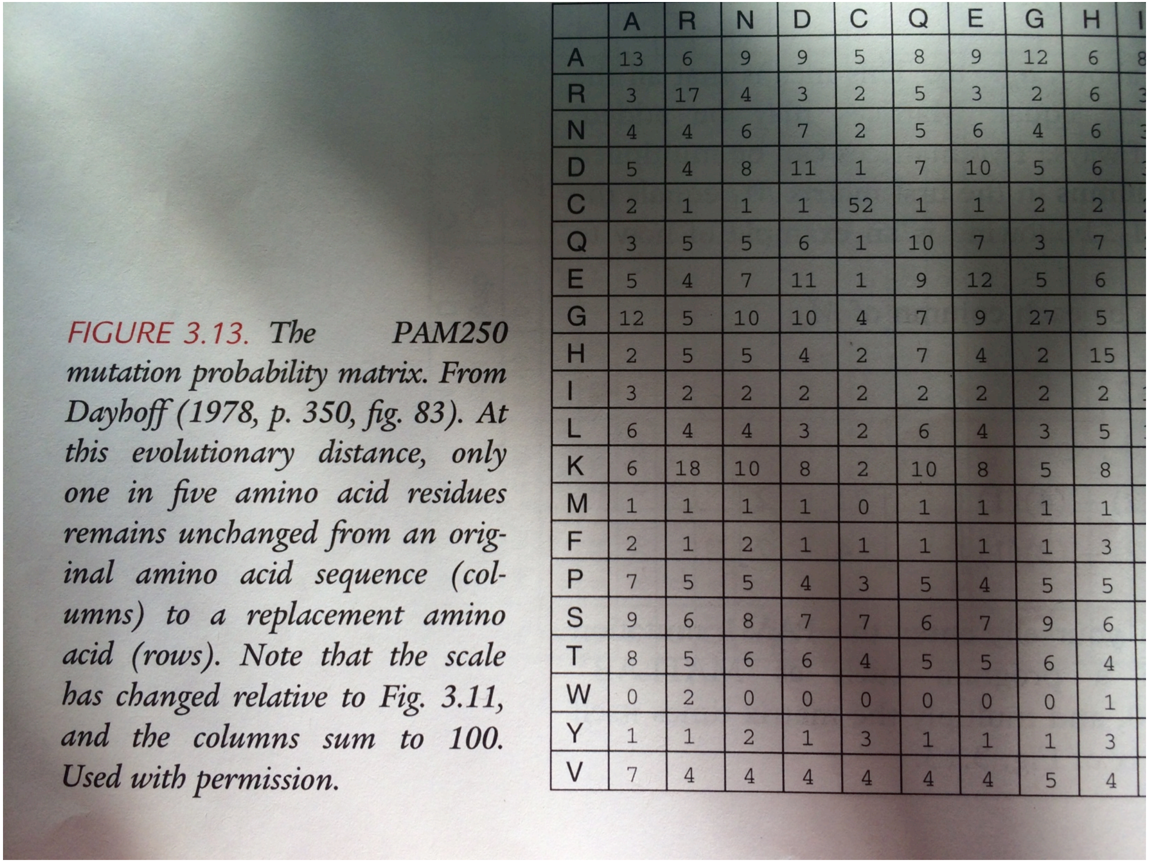
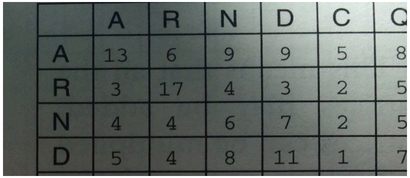


FIGURE 3.13. The PAM250 mutation probability matrix. From Dayhoff (1978, p. 350, fig. 83). At this evolutionary distance, only one in five amino acid residues remains unchanged from an original amino acid sequence (columns) to a replacement amino acid (rows). Note that the scale has changed relative to Fig. 3.11, and the columns sum to 100. Used with permission.

	A	R	N	D	C	Q	E	G	H	I
A	13	6	9	9	5	8	9	12	6	8
R	3	17	4	3	2	5	3	2	6	
N	4	4	6	7	2	5	6	4	6	
D	5	4	8	11	1	7	10	5	6	
C	2	1	1	1	52	1	1	2	2	
Q	3	5	5	6	1	10	7	3	7	
E	5	4	7	11	1	9	12	5	6	
G	12	5	10	10	4	7	9	27	5	
H	2	5	5	4	2	7	4	2	15	
I	3	2	2	2	2	2	2	2	2	
L	6	4	4	3	2	6	4	3	5	
K	6	18	10	8	2	10	8	5	8	
M	1	1	1	1	0	1	1	1	1	
F	2	1	2	1	1	1	1	1	3	
P	7	5	5	4	3	5	4	5	5	
S	9	6	8	7	7	6	7	9	6	
T	8	5	6	6	4	5	5	6	4	
W	0	2	0	0	0	0	0	0	1	
Y	1	1	2	1	3	1	1	1	3	
V	7	4	4	4	4	4	4	5	4	

FROM MUTATIONAL PROBABILITY TO SCORING MATRICES

$$s_{i,j} = 10 * \log_{10}\left(\frac{q_{i,j}}{p_i}\right)$$



	A	R	N	D	C	Q
A	13	6	9	9	5	8
R	3	17	4	3	2	5
N	4	4	6	7	2	5
D	5	4	8	11	1	7

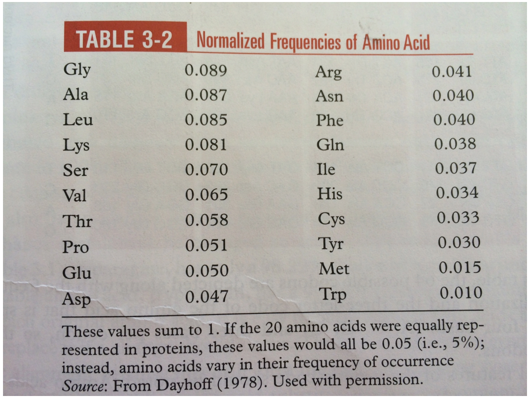


TABLE 3-2 Normalized Frequencies of Amino Acid

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

These values sum to 1. If the 20 amino acids were equally represented in proteins, these values would all be 0.05 (i.e., 5%); instead, amino acids vary in their frequency of occurrence

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

What do these scores mean?

ALIGNMENT – THINK BLAST

Q	ANCQE	versus	ANCQE
D	ANCGE		ANCHE

Blosum Matrix