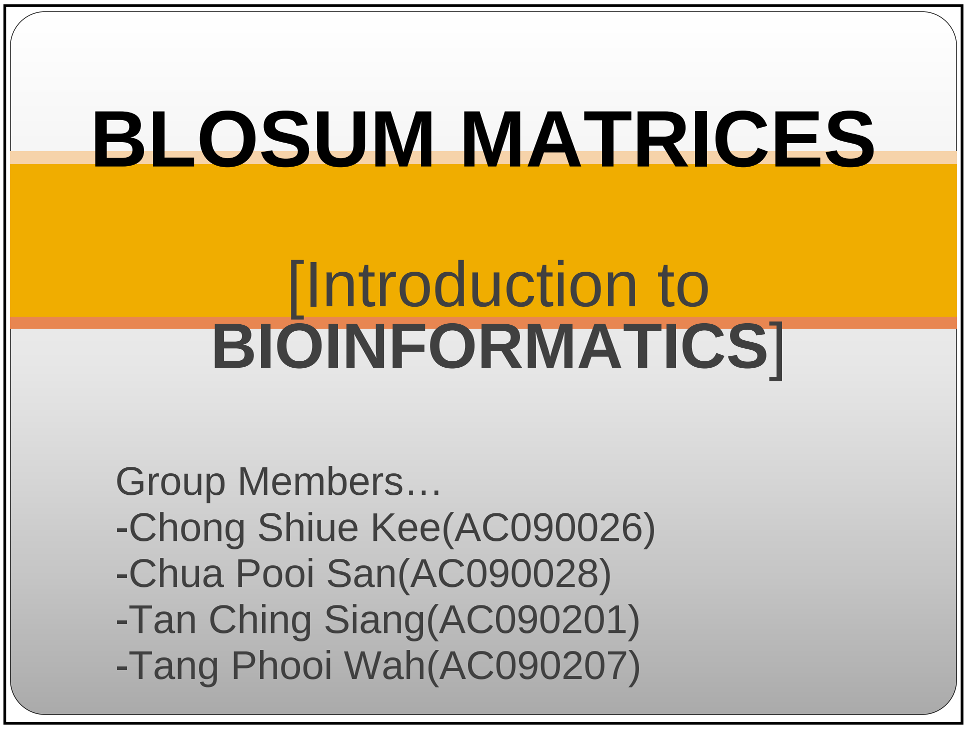
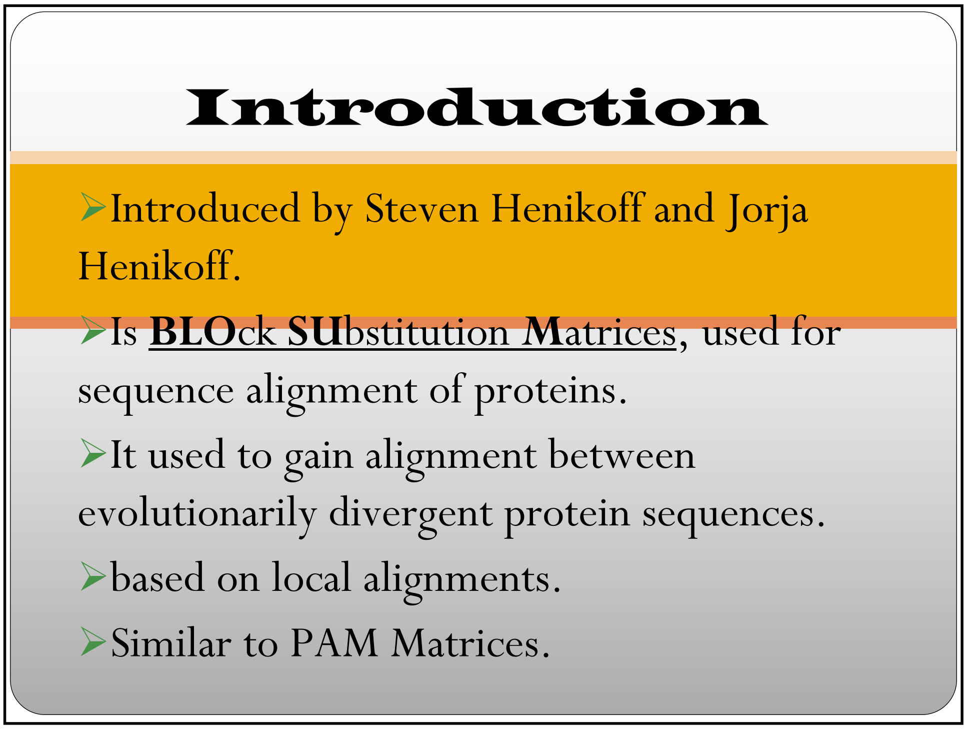


BLOSUM MATRICES

[Introduction to **BIOINFORMATICS**]

Group Members…
-Chong Shiue Kee(AC090026)
-Chua Pooi San(AC090028)
-Tan Ching Siang(AC090201)
-Tang Phooi Wah(AC090207)



Introduction

➢Introduced by Steven Henikoff and Jorja Henikoff.

➢Is **BLO**ck **SU**bstitution **M**atrices, used for sequence alignment of proteins.

➢It used to gain alignment between evolutionarily divergent protein sequences.

➢based on local alignments.

➢Similar to PAM Matrices.

Steven Henikoff & Jorja Henikoff

- Steven was born and raised in Chicago(1950)
- He has 2 sister, he is the youngest among 3 children.
- His father manufactures and sold plastic furniture covers.
- Invented several widely used biotech tool such as techniques, designed with the help of his wife Jorja, for deciphering the function of protein sequences by using the power of computers.

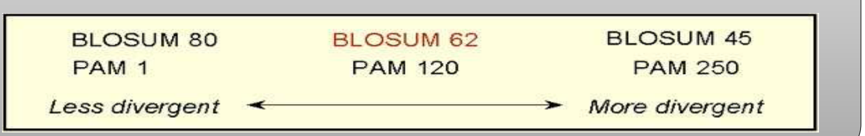
Relationship

PAM Matrix

- To compare the **closely related sequences**, PAM matrices with **lower numbers** are created.
- To compare the **distantly related proteins**, PAM matrices with **high numbers** are created.

BLOSUM Matrix

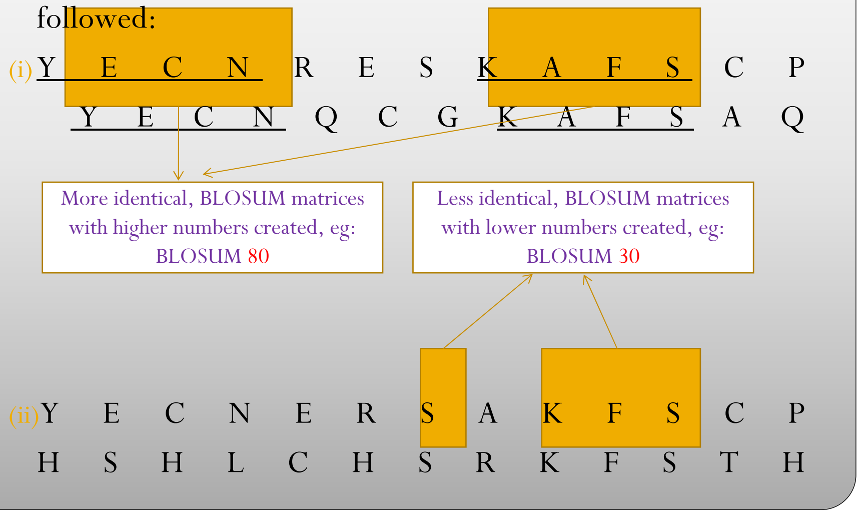
- To compare the **closely related sequences**, BLOSUM matrices with **higher numbers** are created.
- To compare the **distantly related proteins**, BLOSUM matrices with **low numbers** are created.



BLOSUM 80	BLOSUM 62	BLOSUM 45
PAM 1	PAM 120	PAM 250
Less divergent	← →	*More divergent*

BLOSUM Matrices with higher and lower numbers

Assume that the following 2 sequence were aligned as followed:



(i) Y E C N R E S K A F S C P

Y E C N Q C G K A F S A Q

More identical, BLOSUM matrices with higher numbers created, eg: BLOSUM 80

Less identical, BLOSUM matrices with lower numbers created, eg: BLOSUM 30

(ii) Y E C N E R S A K F S C P

H S H L C H S R K F S T H

Differences of PAM & BLOSUM

PAM Matrices

- based on global alignments of closely related proteins.
- PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence.

BLOSUM Matrices

- based on local alignments.
- BLOSUM 62 is a matrix calculated from comparisons of sequences with no more than 62% identical.

Cont…

PAM Matrices

- Other PAM matrices are extrapolated from PAM1.
- Higher numbers in matrices naming scheme denote larger evolutionary distance.

BLOSUM Matrices

- based on observed alignments; they are not extrapolated from comparisons of closely related proteins.
- Larger numbers in matrices naming scheme denote higher sequence similarity and therefore smaller evolutionary distance.

BLOSUM Matrix not extrapolated from another BLOSUM Matrix

- In PAM, PAM 2 is extrapolated from PAM 1.

  eg: PAM 2 = PAM 1 X PAM 1

  PAM 3 = PAM 2 X PAM 1

- But not for BLOSUM Matrices. Eg:

  ~~BLOSUM 20 = BLOSUM 19 X BLOSUM 1~~

  ~~BLOSUM 62 = BLOSUM 61 X BLOSUM 1~~

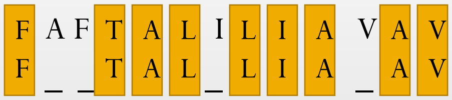
- Each BLOSUM Matrices exist using different alignment database.

Meaning..

- Global alignments which attempts to align every residues in every sequence are most useful when the sequences in the query set are similar and of roughly equal side.
- Local alignments are more useful for dissimilar sequences that are suspected to contain regions of similarility or similar sequence motifs within their larger sequence context.

Global Alignment and Local Alignment

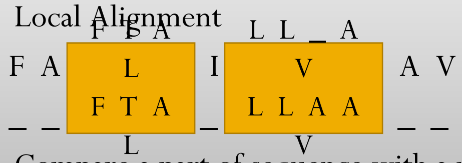
- Global Alignment



```
F A F T A L I L I A V A V
F _ _ T A L _ L I A _ A V
```

- Compare one by one

- Local Alignment



- Compare a part of sequence with a part of sequence

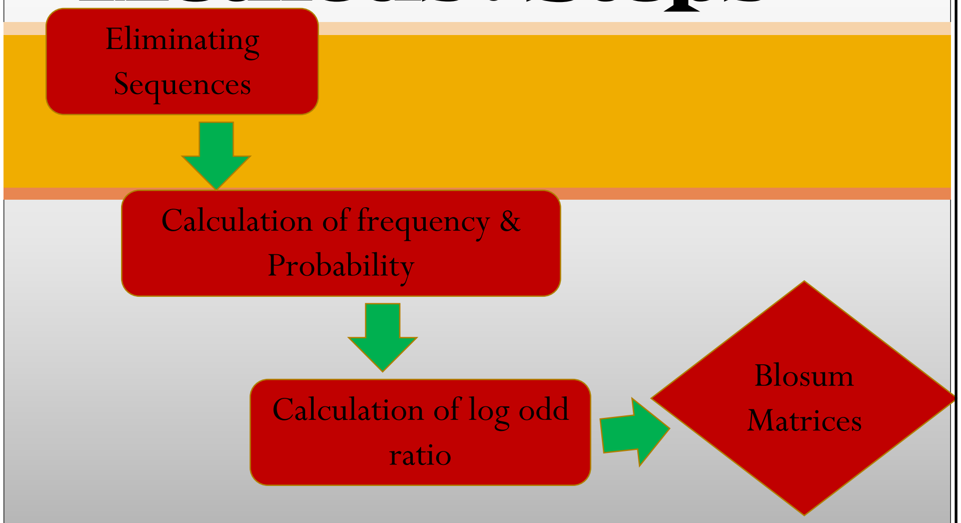
Contents

Blosum Matrices is obtained by using

→blocks of similar amino acid sequences as data.

→then, applying statistical methods to the data to obtain the similarity scores.

Statistical Methods : Steps

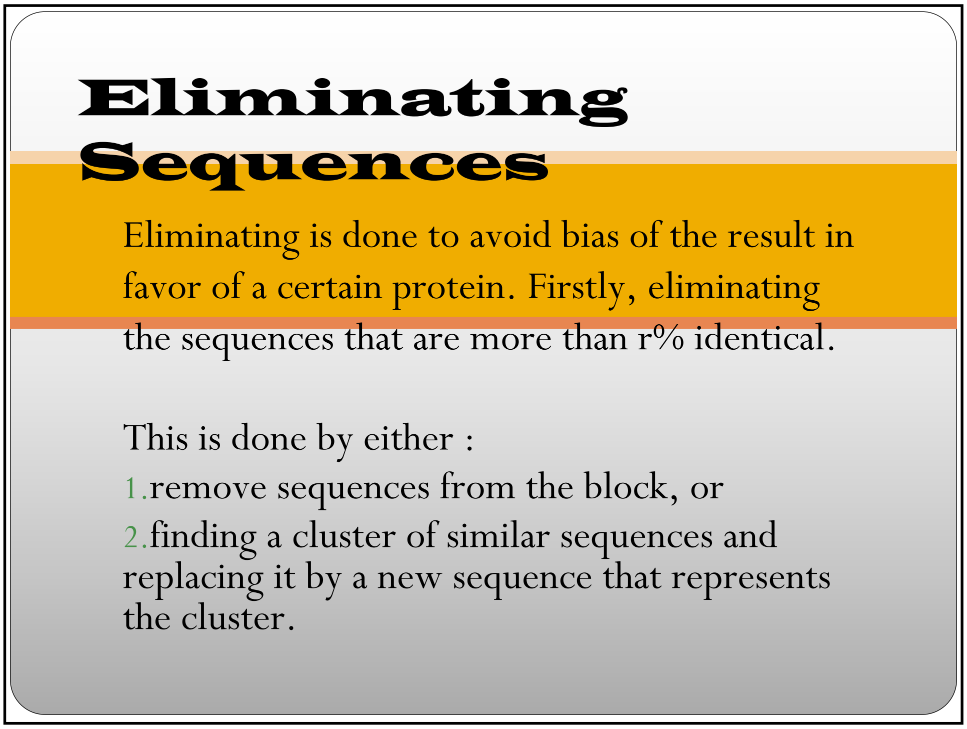


Eliminating Sequences

Calculation of frequency & Probability

Calculation of log odd ratio

Blosum Matrices

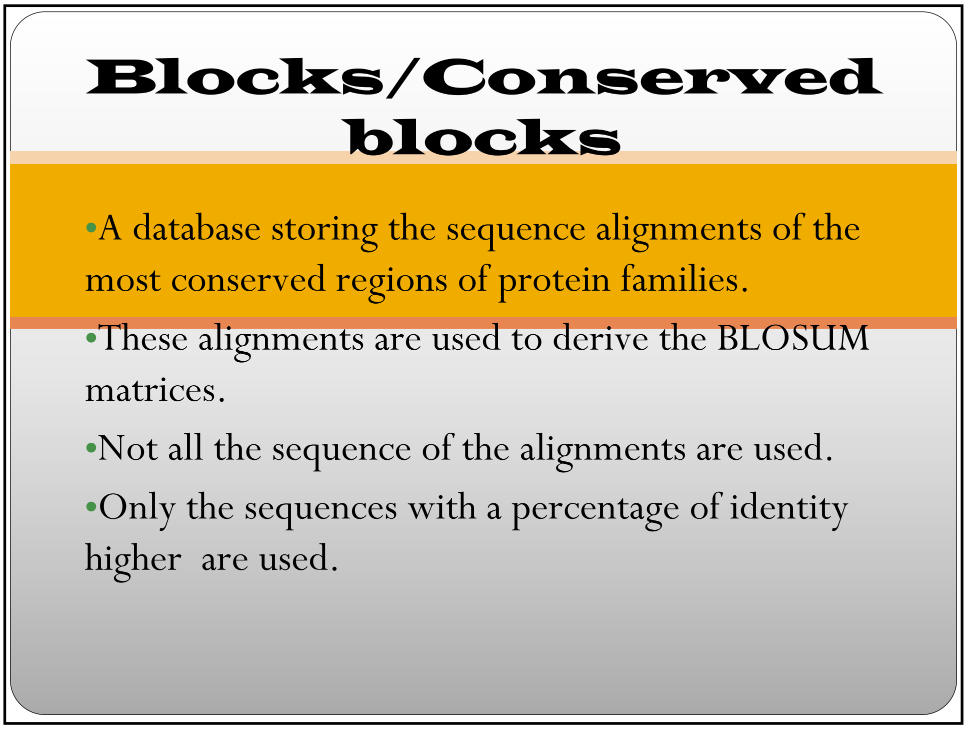


Eliminating Sequences

Eliminating is done to avoid bias of the result in favor of a certain protein. Firstly, eliminating the sequences that are more than r% identical.

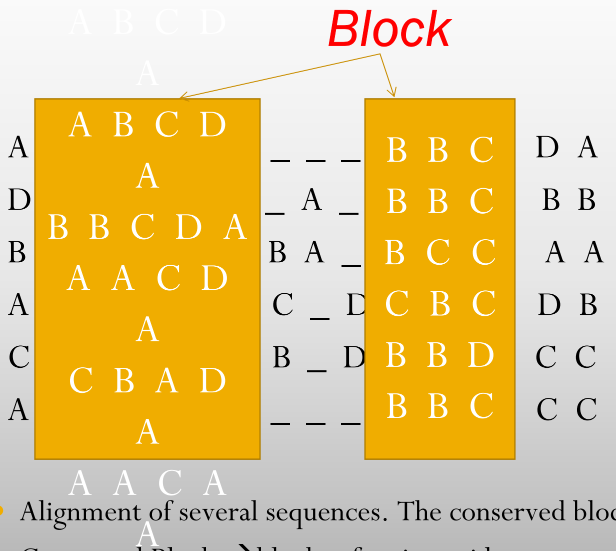
This is done by either :

1. remove sequences from the block, or
2. finding a cluster of similar sequences and replacing it by a new sequence that represents the cluster.



Blocks/Conserved blocks

- A database storing the sequence alignments of the most conserved regions of protein families.
- These alignments are used to derive the BLOSUM matrices.
- Not all the sequence of the alignments are used.
- Only the sequences with a percentage of identity higher are used.



A A B C D A _ _ _ B B C D A

D A B C D A _ A _ B B C B B

B B B C D A B A _ B C C A A

A A A C D A C _ D C B C D B

C C B A D A B _ D B B D C C

A A A C A A _ _ _ B B C C C

- Alignment of several sequences. The conserved blocks are marked
- Conserved Blocks→blocks of amino acid sequence with small change between them.

Reasons of using conserved blocks

Reason 1:

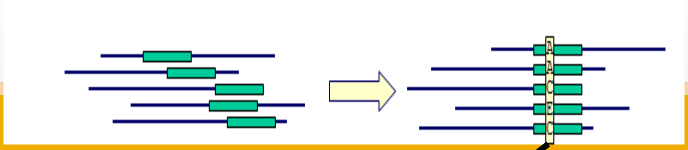
Easier to construct an alignment with more similar sequences from multiple alignment.

Reason 2:

Measure the probability of one amino acid to change into another.

Reason 3:

Restrict our examination by conservation of regions inside protein families.



Sequence	Position			
Seq1	B	A	B	A
Seq2	A	A	A	C
Seq3	A	A	C	C
Seq4	A	A	B	A
Seq5	A	A	C	C
Seq6	A	A	B	C

Calculating Frequency & Probability

By using the block, counting the pairs of amino acids in each column of the multiple alignment.

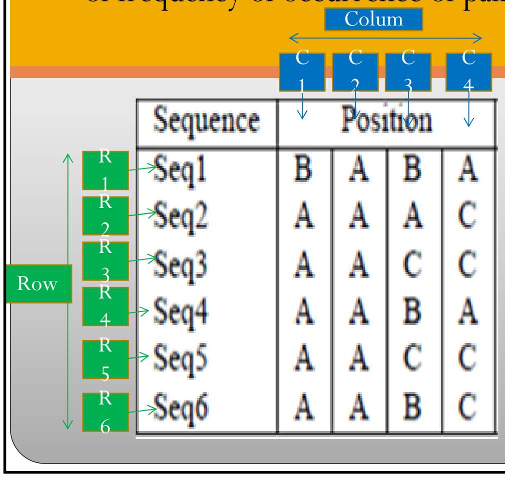
Sequence	Position			
Seq1	B	A	B	A
Seq2	A	A	A	C
Seq3	A	A	C	C
Seq4	A	A	B	A
Seq5	A	A	C	C
Seq6	A	A	B	C

Pair	Frequency of occurrence
AA	26
AB	8
AC	10
BB	3
BC	6
CC	7

TOTAL PAIRS = 26 + 8 + 10 + 3 + 6 + 7 = 60

Example of calculating of the frequency of occurrence for each pairs

Now, we take one example to show that how is the calculation of frequency of occurrence of pairs, (ex: AA pairs).



Sequence	Position			
Seq1	B	A	B	A
Seq2	A	A	A	C
Seq3	A	A	C	C
Seq4	A	A	B	A
Seq5	A	A	C	C
Seq6	A	A	B	C

From the LEFT column to RIGHT column, there is only 5 A's, 6A's, 1A's and 2A's respectively. Pairs them up and total it.

*** AB or BA is calculate as 1 pair if each of them placed at the same position.

Ex: B(C1R1),A(C1R2)

=BA/AB

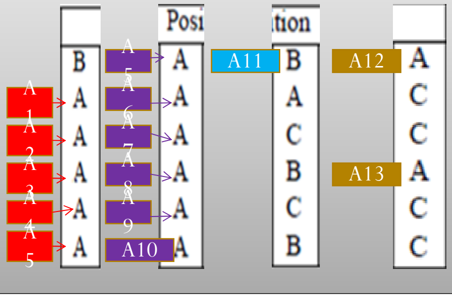
=**<u>1 pair</u>**

Example..Here is one of the example of showing you how to do the calculation

From this sample, the acid amino we used is only A, B and C.

So, the pairs we can gain from the sample is AA, AB, AC, BB, BC and CC. We now calculate how many of AA pairs can be obtained from this sample:

Step 1:

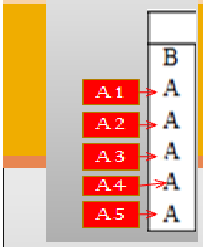


	Posi	tion	
B	A	B	A
A	A	A	C
A	A	C	C
A	A	B	A
A	A	C	C
A	A	B	C

To make it more simple and easy to understand, let us named each of the A from differ position with differ name.

Example : Calculation of frequency (1)

Step 2 :

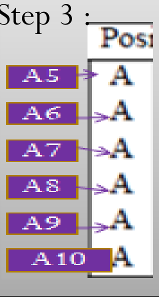


To calculate the number of pairs from Column 1(C1)
= n(C1 R2&R3 + C1 R2&R4 + C1 R2&R5 + C1 R2&R6 + C1 R3&R4 + C1 R3&R5 + C1 R3&R6 + C1 R4&R5 + C1 R4&R6 + C1 R5&R6)

=n(A1A2 + A1A3 + A1A4 + A1A5 + A2A3 + A2A4 + A2A5 + A3A4 + A3A5 + A4A5)

=**<u>10 pairs</u>**

Step 3 :



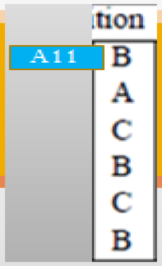
To calculate the number of pairs from Column 2(C2)
= n(C2 R1&R2 + C2 R1&R3 + C2 R1&R4 + C2 R1&R5 + C2 R1&R6 + C2 R2&R3 + C2 R2&R4 + C2 R2&R5 + C2 R2&R6 + C2 R3&R4 + C2 R3&R5 + C2 R3&R6 + C2 R4&R5 + C2 R4&R6 + C2 R5&R6)

=n(A5A6 + A5A7 + A5A8 + A5A9 + A5A10 + A6A7 + A6A8 + A6A9 + A6A10 + A7A8 + A7A9 + A7A10 + A8A9 + A8A10 + A9A10)

=**<u>15pairs</u>**

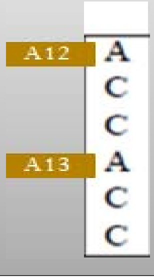
Example : Calculation of frequency(2)

Step 4 :



There is no any pair form column 3 due to there is only a A in this column.

Step 5:



To calculate the number of pairs from Column 4(C4)
= n(C4 R1&R4)
=n(A12A13)
=**<u>1 pairs</u>**

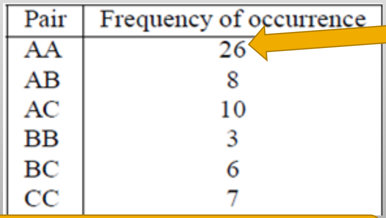
Example: Calculation of frequency (3)

Step 6 : Sum up all the total pairs from each columns .

Total (frequency)of AA pairs = C1 + C2 + C3 + C4

= 10 + 15 + 0 + 1

= 26 pairs



Pair	Frequency of occurrence
AA	26
AB	8
AC	10
BB	3
BC	6
CC	7

Total frequency of pairs = 60

✓There is the same method in calculating the frequencies of AB, AC, BB,BC,CC pairs.

***Remember : AB or reverse of it(BA), in the condition A and B still in the same position, is consider as one pair.

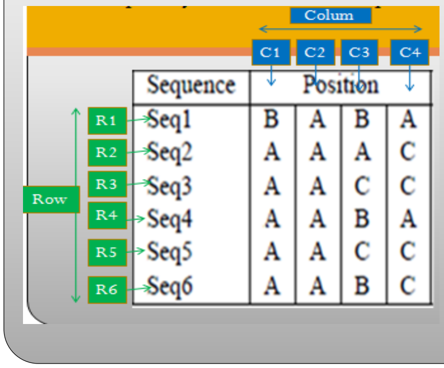
- You can also used the formula to calculate total number of substitution instead of calculating it manually
- Total column contribute Formula : d(d-1)/2
- d = column of depth → how many residues in the column
- Ex: 6 residues, result = 6(5)/2 = 15
- For 4 column= 4*15 = 60

Example : Calculation of frequency by using formula.

Formula : d(d-1)/2

d is column of depth, means of how many of residue in one column.

A, B, C is the residue of the column.


Colum
C1
C2
C3
C4
Row
R1
R2
R3
R4
R5
R6

Sequence	Position			
Seq1	B	A	B	A
Seq2	A	A	A	C
Seq3	A	A	C	C
Seq4	A	A	B	A
Seq5	A	A	C	C
Seq6	A	A	B	C

In each column, there is 6 residues (A, B and C), so means of d=6per column.

By using the formula,

= d(d-1)/2

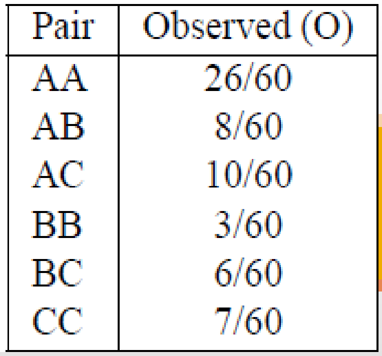
=6(5)/2

=15

Since it is getting 4 columns in this sample, so we times 4 to gain the total the frequency of all pairs.

=15 * 4

=**60**

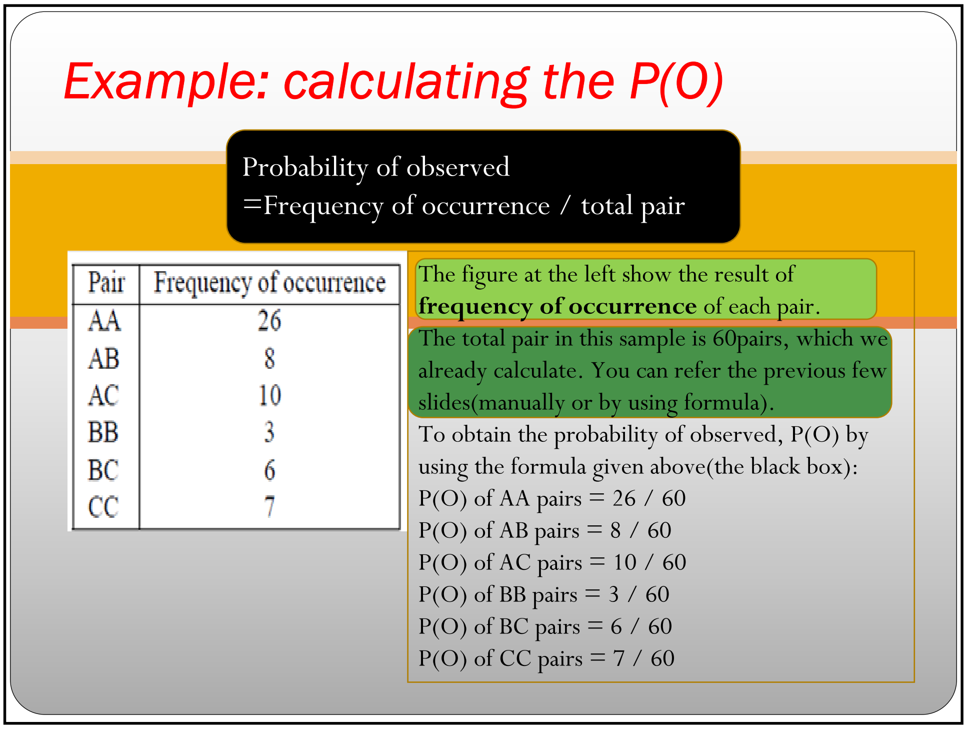


Pair	Observed (O)
AA	26/60
AB	8/60
AC	10/60
BB	3/60
BC	6/60
CC	7/60

Probability of observed

=Frequency of occurrence / total pair

Eg:AA pairs(O)= 26/60



Example: calculating the P(O)

Probability of observed
=Frequency of occurrence / total pair

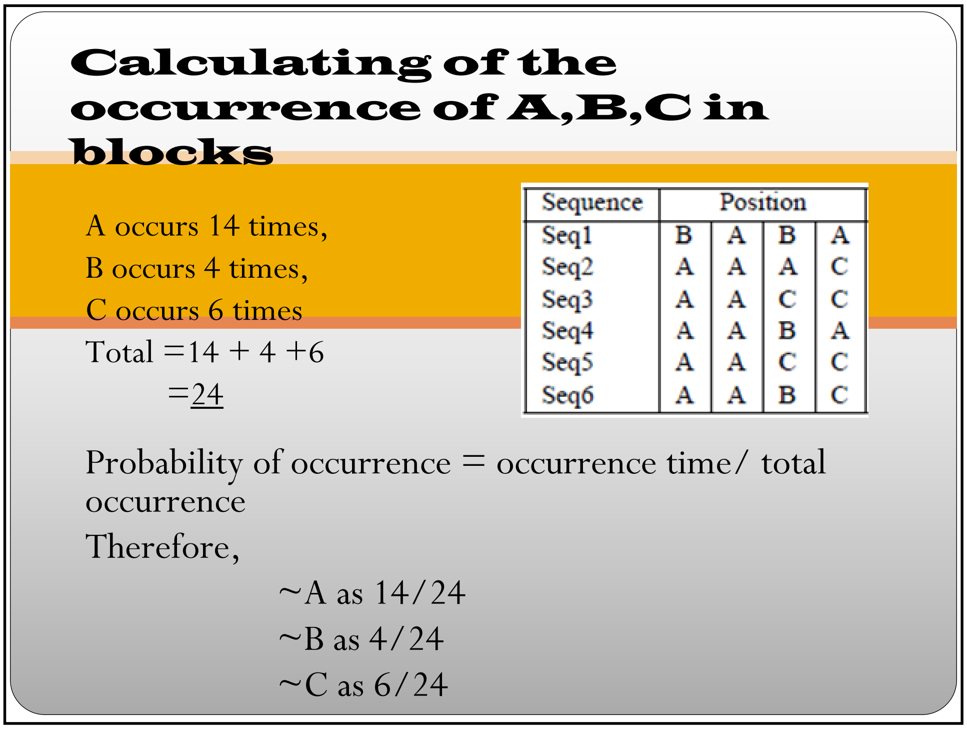
Pair	Frequency of occurrence
AA	26
AB	8
AC	10
BB	3
BC	6
CC	7

The figure at the left show the result of **frequency of occurrence** of each pair.

The total pair in this sample is 60pairs, which we already calculate. You can refer the previous few slides(manually or by using formula).

To obtain the probability of observed, P(O) by using the formula given above(the black box):

P(O) of AA pairs = 26 / 60
P(O) of AB pairs = 8 / 60
P(O) of AC pairs = 10 / 60
P(O) of BB pairs = 3 / 60
P(O) of BC pairs = 6 / 60
P(O) of CC pairs = 7 / 60



Calculating of the occurrence of A,B,C in blocks

A occurs 14 times,
B occurs 4 times,
C occurs 6 times
Total =14 + 4 +6
=24

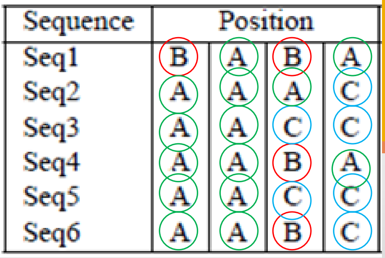
Sequence	Position			
Seq1	B	A	B	A
Seq2	A	A	A	C
Seq3	A	A	C	C
Seq4	A	A	B	A
Seq5	A	A	C	C
Seq6	A	A	B	C

Probability of occurrence = occurrence time/ total occurrence

Therefore,

~A as 14/24
~B as 4/24
~C as 6/24

Example :Calculation the occurrence of A, B and C in the sample



Sequence	Position			
Seq1	B	A	B	A
Seq2	A	A	A	C
Seq3	A	A	C	C
Seq4	A	A	B	A
Seq5	A	A	C	C
Seq6	A	A	B	C

Occurrence of A (total up the **green** circle in the left figure) = **14**

Occurrence of B (total up the **red** circle in the left figure) = **4**

Occurrence of C (total up the **blue** circle in the left figure) = **6**

Therefore, total occurrence of A, B and C in the sample is

14+4+6 = 24

Example : Calculation of probability of occurrence.

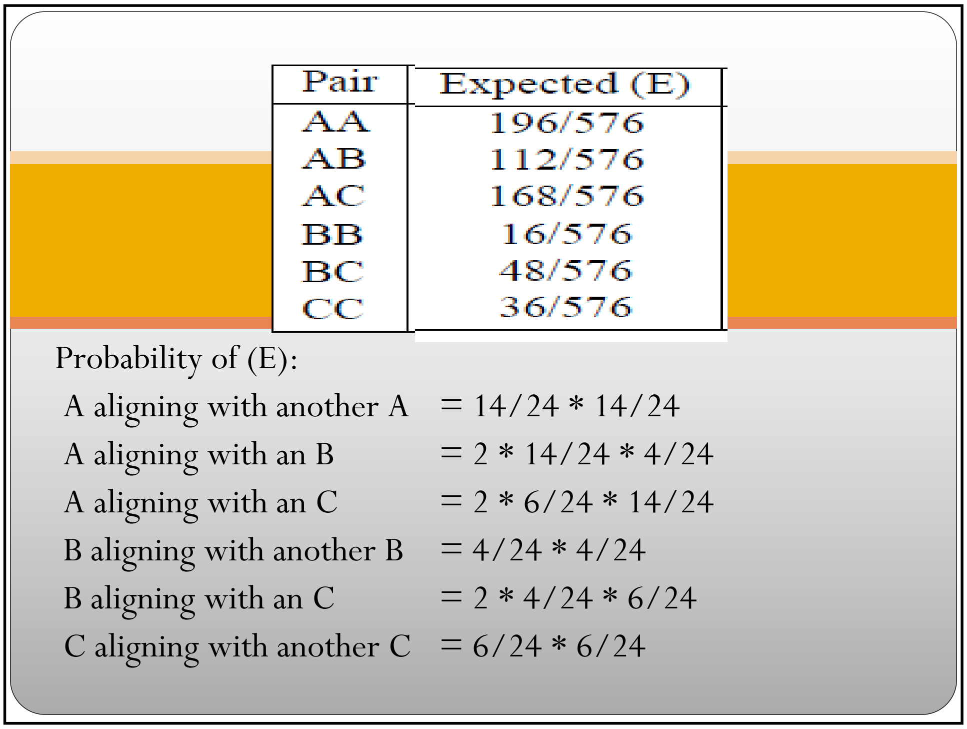
To gain the probability of occurrence of each, used this formula :

Probability of occurrence = occurrence time/ total occurrence

Probability occurrence of A = 14/24

Probability occurrence of B = 4/24

Probability occurrence of C = 6/24



Pair	Expected (E)
AA	196/576
AB	112/576
AC	168/576
BB	16/576
BC	48/576
CC	36/576

Probability of (E):

A aligning with another A = 14/24 * 14/24

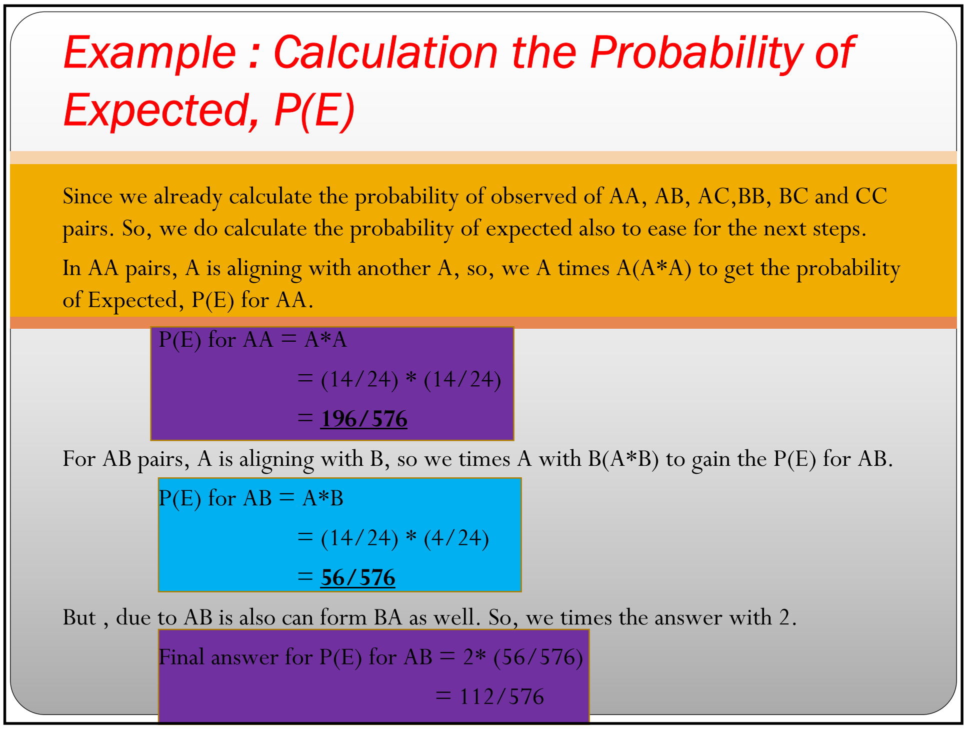
A aligning with an B = 2 * 14/24 * 4/24

A aligning with an C = 2 * 6/24 * 14/24

B aligning with another B = 4/24 * 4/24

B aligning with an C = 2 * 4/24 * 6/24

C aligning with another C = 6/24 * 6/24



Example : Calculation the Probability of Expected, P(E)

Since we already calculate the probability of observed of AA, AB, AC,BB, BC and CC pairs. So, we do calculate the probability of expected also to ease for the next steps.

In AA pairs, A is aligning with another A, so, we A times A(A*A) to get the probability of Expected, P(E) for AA.

P(E) for AA = A*A
= (14/24) * (14/24)
= **196/576**

For AB pairs, A is aligning with B, so we times A with B(A*B) to gain the P(E) for AB.

P(E) for AB = A*B
= (14/24) * (4/24)
= **56/576**

But , due to AB is also can form BA as well. So, we times the answer with 2.

Final answer for P(E) for AB = 2* (56/576)
= 112/576

Example : Calculation the Probability of Expected, P(E)(2)

For AC pairs, A is aligning with C, so we times A with C(A*C) to gain the P(E) for AC.

P(E) for AC = A*C
= (14/24) * (6/24)
= **<u>84/576</u>**

But , due to AC is also can form CA as well. So, we times the answer with 2.

Final answer for P(E) for AC= 2* (84/576)
= **<u>168/576</u>**

In BB pairs, B is aligning with another B, so, we B times B(B*B) to get the probability of Expected, P(E) for BB.

P(E) for BB = B*B
= (4/24) * (4/24)
= **<u>16/576</u>**

Example : Calculation the Probability of Expected, P(E)(3)

For BC pairs, B is aligning with C, so we times B with C(B*C) to gain the P(E) for BC.

P(E) for BC = B*C
= (4/24) * (6/24)
= **<u>24/576</u>**

But , due to BC is also can form CB as well. So, we times the answer with 2.

Final answer for P(E) for BC= 2* (24/576)
= **<u>48/576</u>**

In CC pairs, C is aligning with another C, so, we C times C(C*C) to get the probability of Expected, P(E) for CC.

P(E) for CC = C*C
= (6/24) * (6/24)
= **<u>36/576</u>**

log odd ratio

It gives the ratio of the occurrence each amino acid combination in the observed data to the expected value of occurrence of the pair.

It is rounded off and used in the substitution matrix.

* Value stored for Blosum= 2log odd ratio rounded to nearest integer

log odd ratio = $2\log_2$ (O/E)

Example: Calculation of the log odd ratio(1).

log odd ratio = $2\log_2$ (O/E)

The purpose of calculate the log odd ratio is for us to get the substitution matrices(BLOSUM Matric). Using log is in the purpose to minimize the value.

Here are the calculation by using log odd ratio:

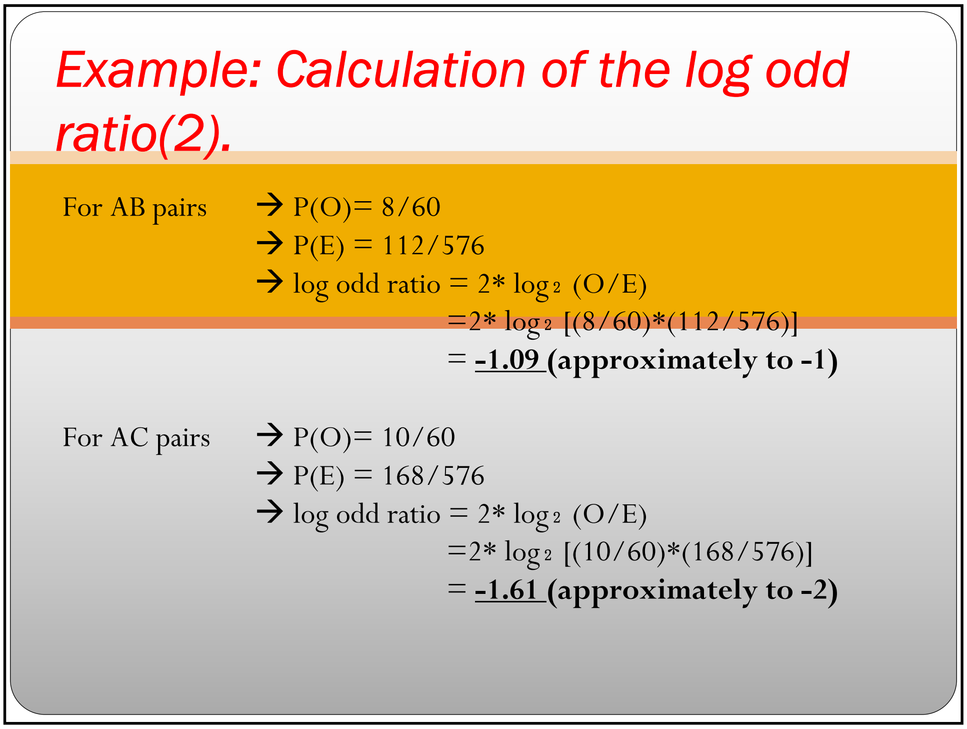
For AA pairs → P(O)= 26/60

→ P(E) = 196/576

→ log odd ratio = 2* $\log_2$ (O/E)

=2* $\log_2$ [(26/60)*(196/576)]

= **0.70 (approximately to 1)**



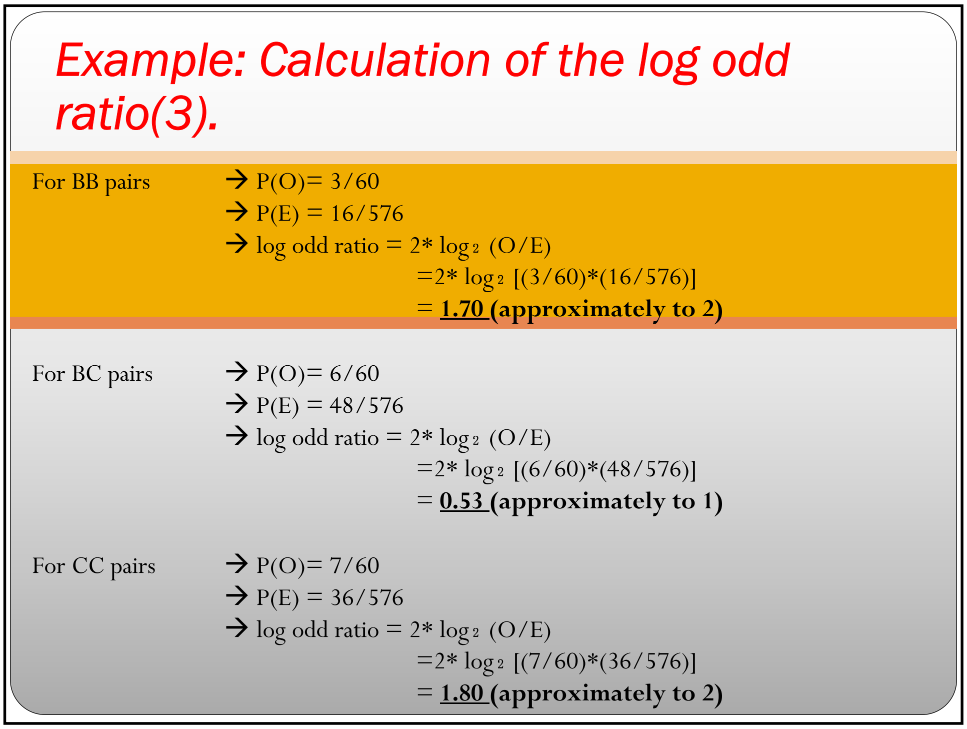
Example: Calculation of the log odd ratio(2).

For AB pairs

→ P(O)= 8/60
→ P(E) = 112/576
→ log odd ratio = 2* $\log_2$ (O/E)
=2* $\log_2$ [(8/60)*(112/576)]
= **-1.09 (approximately to -1)**

For AC pairs

→ P(O)= 10/60
→ P(E) = 168/576
→ log odd ratio = 2* $\log_2$ (O/E)
=2* $\log_2$ [(10/60)*(168/576)]
= **-1.61 (approximately to -2)**



Example: Calculation of the log odd ratio(3).

For BB pairs

→ P(O)= 3/60
→ P(E) = 16/576
→ log odd ratio = 2* $\log_2$ (O/E)
=2* $\log_2$ [(3/60)*(16/576)]
= **1.70 (approximately to 2)**

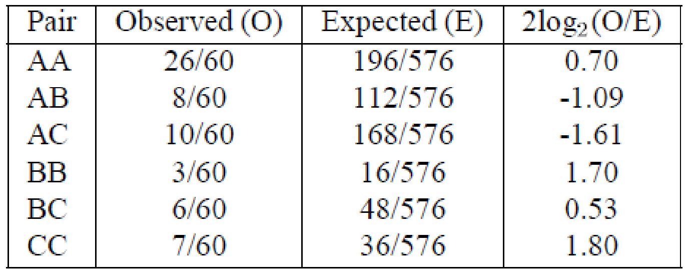
For BC pairs

→ P(O)= 6/60
→ P(E) = 48/576
→ log odd ratio = 2* $\log_2$ (O/E)
=2* $\log_2$ [(6/60)*(48/576)]
= **0.53 (approximately to 1)**

For CC pairs

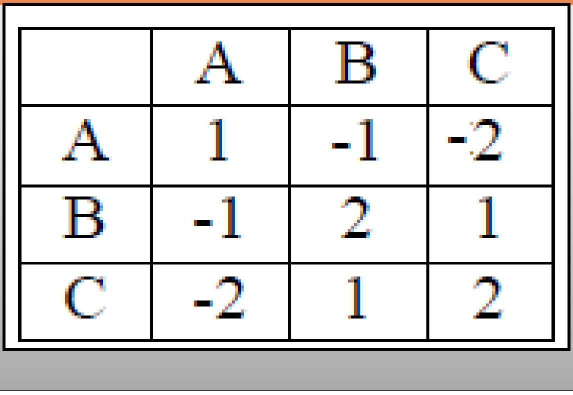
→ P(O)= 7/60
→ P(E) = 36/576
→ log odd ratio = 2* $\log_2$ (O/E)
=2* $\log_2$ [(7/60)*(36/576)]
= **1.80 (approximately to 2)**

Overall



Pair	Observed (O)	Expected (E)	$2\log_2(O/E)$
AA	26/60	196/576	0.70
AB	8/60	112/576	-1.09
AC	10/60	168/576	-1.61
BB	3/60	16/576	1.70
BC	6/60	48/576	0.53
CC	7/60	36/576	1.80

The odds for relatedness are calculated from log odd ratio, which are then rounded off to get the substitution matrices→ **BLOSUM** matrices. (looks as follows)



	A	B	C
A	1	-1	-2
B	-1	2	1
C	-2	1	2

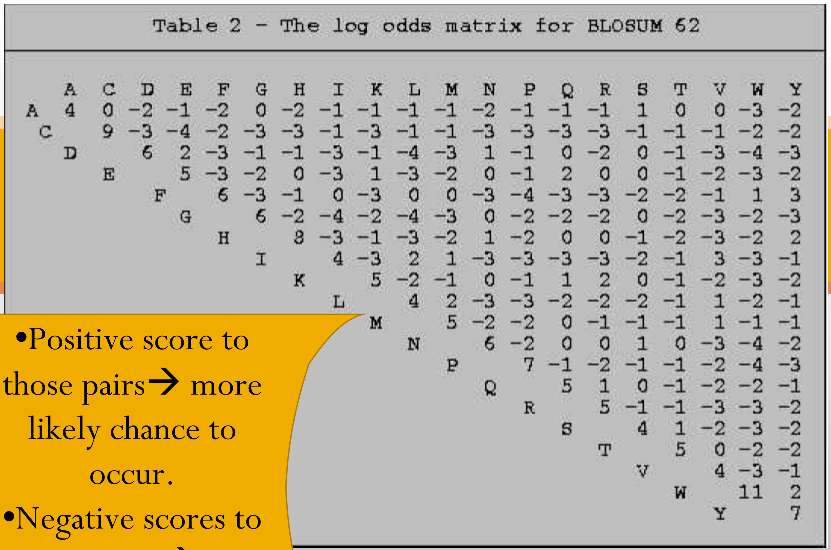
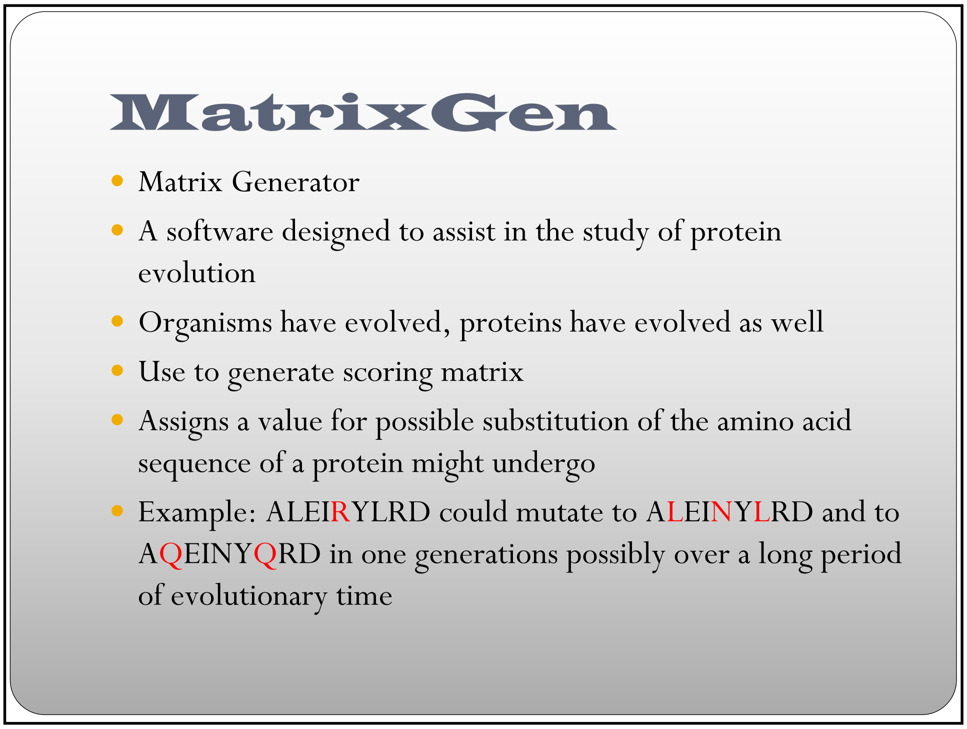


Table 2 - The log odds matrix for BLOSUM 62

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

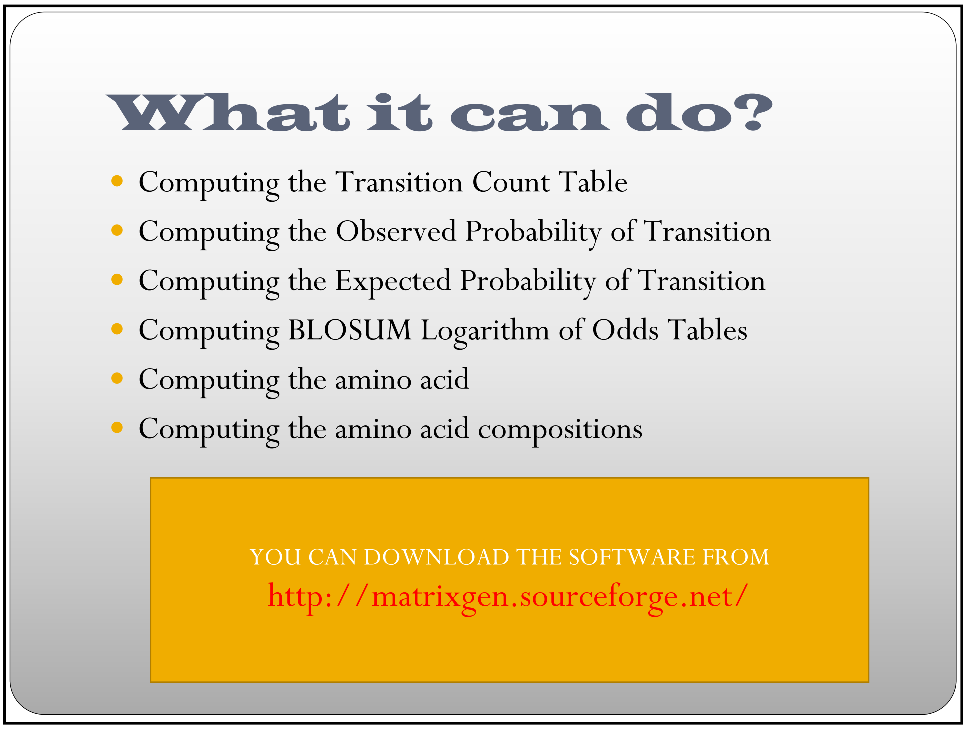
- Positive score to those pairs→ more likely chance to occur.
- Negative scores to those pairs → less likely to occur.

What Software we Use?



MatrixGen

- Matrix Generator
- A software designed to assist in the study of protein evolution
- Organisms have evolved, proteins have evolved as well
- Use to generate scoring matrix
- Assigns a value for possible substitution of the amino acid sequence of a protein might undergo
- Example: ALEIRYLRD could mutate to ALEINYLRD and to AQEINYQRD in one generations possibly over a long period of evolutionary time



What it can do?

- Computing the Transition Count Table
- Computing the Observed Probability of Transition
- Computing the Expected Probability of Transition
- Computing BLOSUM Logarithm of Odds Tables
- Computing the amino acid
- Computing the amino acid compositions

YOU CAN DOWNLOAD THE SOFTWARE FROM
http://matrixgen.sourceforge.net/

Explanation

- **For computing the amino acid**

→It provides the raw count of each amino acid in the sequence

→It will only display data along the diagonal of the results grid

→Ex: The data in Row A Column A represents the number of times the amino acid Alanine appears in the alignment

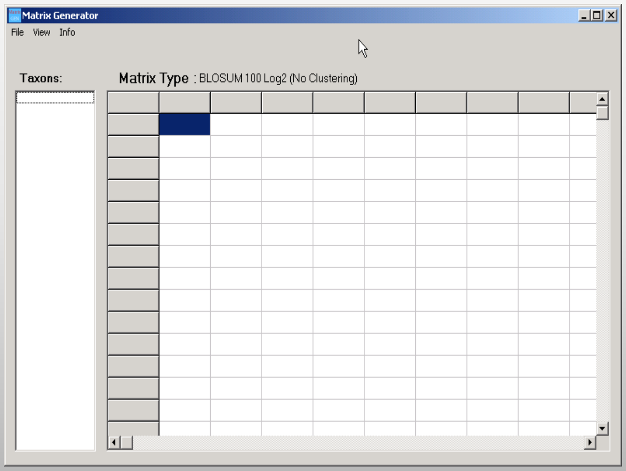
- **For computing the amino acids composition**

→It is identical to the amino acid count view with one exception

→The composition view provides the percentage of the alignment that consists of a given residue

→ This percent is opposed to the raw number given in the amino acid counts view

User Interface Overview

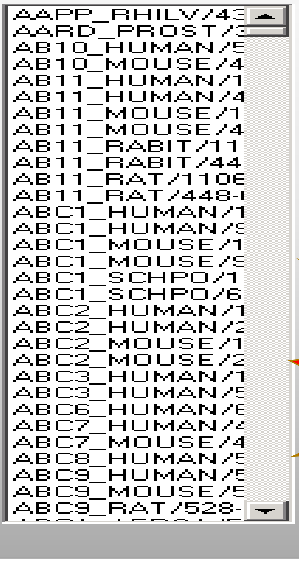


Matrix Type : BLOSUM 100 Log2 (No Clustering)

Figure 1 – Main Screen

- The major components of this screen are:
  - → Taxon Box
  - → Result Grid
  - → Matrix Type Label
  - → Menu Bar

Taxon Box

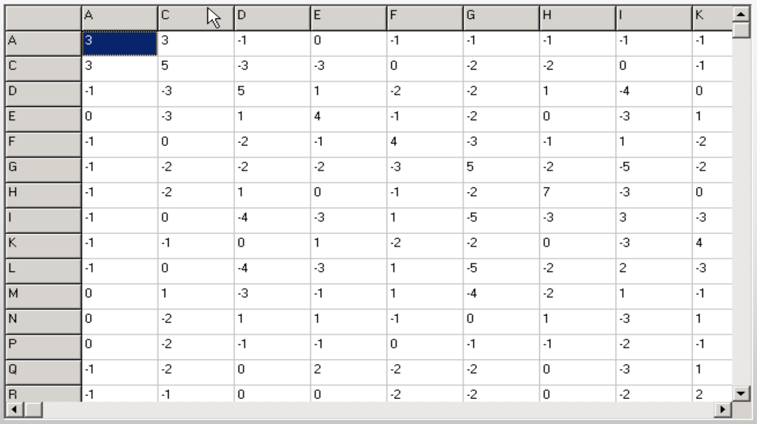


```
AAPP_RHILV/43
AARD_PROST/3
AB10_HUMAN/5
AB10_MOUSE/4
AB11_HUMAN/1
AB11_HUMAN/4
AB11_MOUSE/1
AB11_MOUSE/4
AB11_RABIT/11
AB11_RABIT/44
AB11_RAT/110E
AB11_RAT/448-
ABC1_HUMAN/1
ABC1_HUMAN/9
ABC1_MOUSE/1
ABC1_MOUSE/9
ABC1_SCHPO/1
ABC1_SCHPO/6
ABC2_HUMAN/1
ABC2_HUMAN/2
ABC2_MOUSE/1
ABC2_MOUSE/2
ABC3_HUMAN/1
ABC3_HUMAN/5
ABC6_HUMAN/6
ABC7_HUMAN/4
ABC7_MOUSE/4
ABC8_HUMAN/5
ABC9_HUMAN/5
ABC9_MOUSE/5
ABC9_RAT/528-
```

The taxon box lists the taxons contributing to the entries in the results grid.

Each rows inside the taxon box represent the amino acid sequences(From alignment file)

Result Grid



	A	C	D	E	F	G	H	I	K
A	3	3	-1	0	-1	-1	-1	-1	-1
C	3	5	-3	-3	0	-2	-2	0	-1
D	-1	-3	5	1	-2	-2	1	-4	0
E	0	-3	1	4	-1	-2	0	-3	1
F	-1	0	-2	-1	4	-3	-1	1	-2
G	-1	-2	-2	-2	-3	5	-2	-5	-2
H	-1	-2	1	0	-1	-2	7	-3	0
I	-1	0	-4	-3	1	-5	-3	3	-3
K	-1	-1	0	1	-2	-2	0	-3	4
L	-1	0	-4	-3	1	-5	-2	2	-3
M	0	1	-3	-1	1	-4	-2	1	-1
N	0	-2	1	1	-1	0	1	-3	1
P	0	-2	-1	-1	0	-1	-1	-2	-1
Q	-1	-2	0	2	-2	-2	0	-3	1
R	-1	-1	0	0	-2	-2	0	-2	2

This is where you can get the transition scores from one a.a to another. After processing the alignment file, the result shows like that.

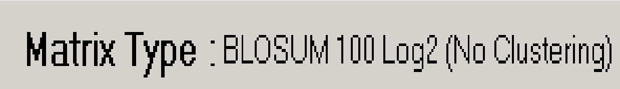
- Letters along the top and down the left hand side of the result grid are single letter for Amino acids.

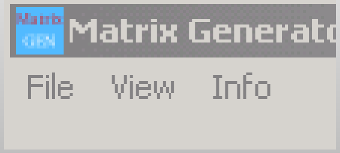
Symbol	Amino Acid
A	Alanine
V	Valine
I	Isoleucine
L	Leucine
M	Methionine
F	Phenylalnine
Y	Tyrosine
W	Tryptophan
K	Lysine
R	Arganine
H	Histidine
D	Aspartate
E	Glutamate
S	Serine
T	Threonine
N	Asparagine
O	Glutamine
C	Cytosine
P	Proline
G	Glysine

- Each of the entries in the results grid has a number
- This number represents the likelihood of transition from one a.a to another
- For example: Column D Row I → -4
- It means that the transition of Aspartate(D) to Isoleucine(I) (vice versa) were more rarely observed than others mutation
- sometimes, you might end up with non-numeric entry in the result grid
- Particular transition was never observed in the sample
- “i” is inserted in the entry
- This symbol is used by PAUP to indicate that a specific transition is restricted

Matrix Type Label

- Describing the calculations used to generate the results


Matrix Type : BLOSUM 100 Log2 (No Clustering)

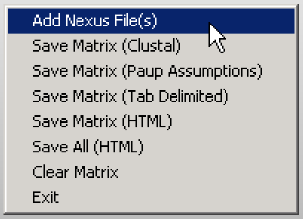


Menu Bar-initiating most tasks in Matrix Gen.

How to Start?

- Adding a file is the step required for MatrixGen to create a scoring matrix
- You can use one or more files
- To add single file:

→ click on File > Add Nexus File(s)

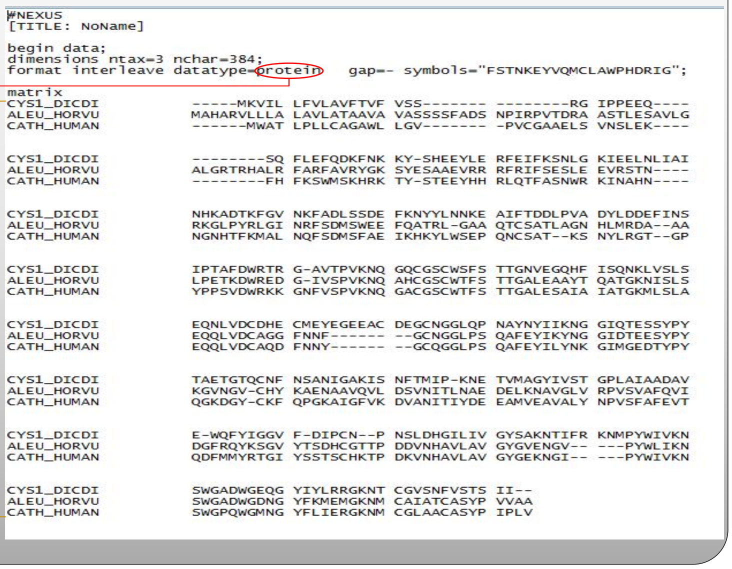


REMEMBER:MatrixGen only read Nexus File(.pau)
You can use **seqVerter** to change others format of file to Nexus

Example of nexus file(Format)

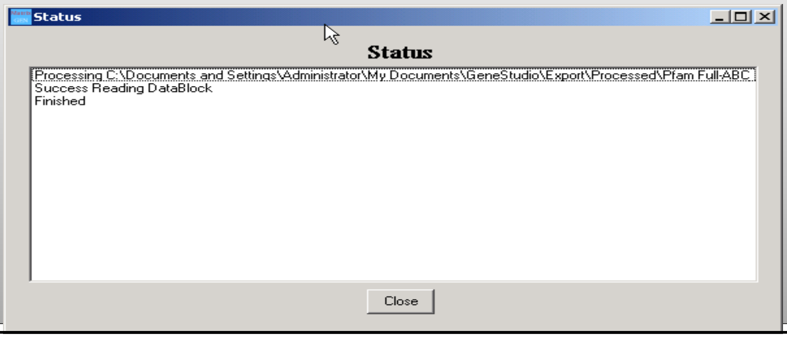
The datatype is protein. How we know?

Amino acid sequences (protein)



```
#NEXUS
[TITLE: NoName]

begin data;
dimensions ntax=3 nchar=384;
format interleave datatype=protein   gap=- symbols="FSTNKEYVQMCLAWPHDRIG";
matrix
CYS1_DICDI          -----MKVIL LFVLAVFTVF VSS------- --------RG IPPEEQ----
ALEU_HORVU          MAHARVLLLA LAVLATAAVA VASSSSFADS NPIRPVTDRA ASTLESAVLG
CATH_HUMAN          ------MWAT LPLLCAGAWL LGV------- -PVCGAAELS VNSLEK----

CYS1_DICDI          --------SQ FLEFQDKFNK KY-SHEEYLE RFEIFKSNLG KIEELNLIAI
ALEU_HORVU          ALGRTRHALR FARFAVRYGK SYESAAEVRR RFRIFSESLE EVRSTN----
CATH_HUMAN          --------FH FKSWMSKHRK TY-STEEYHH RLQTFASNWR KINAHN----

CYS1_DICDI          NHKADTKFGV NKFADLSSDE FKNYYLNNKE AIFTDDLPVA DYLDDEFINS
ALEU_HORVU          RKGLPYRLGI NRFSDMSWEE FQATRL-GAA QTCSATLAGN HLMRDA--AA
CATH_HUMAN          NGNHTFKMAL NQFSDMSFAE IKHKYLWSEP QNCSAT--KS NYLRGT--GP

CYS1_DICDI          IPTAFDWRTR G-AVTPVKNQ GQCGSCWSFS TTGNVEGQHF ISQNKLVSLS
ALEU_HORVU          LPETKDWRED G-IVSPVKNQ AHCGSCWTFS TTGALEAAYT QATGKNISLS
CATH_HUMAN          YPPSVDWRKK GNFVSPVKNQ GACGSCWTFS TTGALESAIA IATGKMLSLA

CYS1_DICDI          EQNLVDCDHE CMEYEGEEAC DEGCNGGLQP NAYNYIIKNG GIQTESSYPY
ALEU_HORVU          EQQLVDCAGG FNNF------ --GCNGGLPS QAFEYIKYNG GIDTEESYPY
CATH_HUMAN          EQQLVDCAQD FNNY------ --GCQGGLPS QAFEYILYNK GIMGEDTYPY

CYS1_DICDI          TAETGTQCNF NSANIGAKIS NFTMIP-KNE TVMAGYIVST GPLAIAADAV
ALEU_HORVU          KGVNGV-CHY KAENAAVQVL DSVNITLNAE DELKNAVGLV RPVSVAFQVI
CATH_HUMAN          QGKDGY-CKF QPGKAIGFVK DVANITIYDE EAMVEAVALY NPVSFAFEVT

CYS1_DICDI          E-WQFYIGGV F-DIPCN--P NSLDHGILIV GYSAKNTIFR KNMPYWIVKN
ALEU_HORVU          DGFRQYKSGV YTSDHCGTTP DDVNHAVLAV GYGVENGV-- ---PYWLIKN
CATH_HUMAN          QDFMMYRTGI YSSTSCHKTP DKVNHAVLAV GYGEKNGI-- ---PYWIVKN

CYS1_DICDI          SWGADWGEQG YIYLRRGKNT CGVSNFVSTS II--
ALEU_HORVU          SWGADWGDNG YFKMEMGKNM CAIATCASYP VVAA
CATH_HUMAN          SWGPQWGMNG YFLIERGKNM CGLAACASYP IPLV
```

- After you select and open the file, MatrixGen will take some times to produce the matrix
- A status screen will appear
- It shows the processing of the file
- If any errors occur, a message displaying the problem will appear in the box
- After the file has been successfully processed, you will see the message "Success Reading DataBlock



Status

```
Processing C:\Documents and Settings\Administrator\My Documents\GeneStudio\Export\Processed\Pfam Full-ABC
Success Reading DataBlock
Finished
```

Saving of Data

- Not an alignment program
- To be a useful tool for creating accurate alignments, it must produce results that can be used by programs that generate alignments

PAUP assumptions block format

- It is used by PAUP (Phylogenetic analysis using parsimony)
- PAUP is used to generate phylogenetic trees
- To save:File > Save Matrix(PAUP Assumptions)

Clustal format

- Clustal is a free program used for aligning sequences
- To save:File > Save Matrix(Clustal)

HTML format

- Format of internet
- Allow you to share the results with the world
- There are two options:
- Save Matrix(HTML) → save the matrix currently being displayed as HTML matrix
- Save All(HTML) → save all MatrixGen matrices into one HTML file(does not matter you are viewing which matrix)

Tab Delimited

- Save matrix as text file with tab separating each data field
- Can be opened in spreadsheets, such as Excel
- Build graph

Cont…

- MatrixGen not an alignment software → can't show us the alignment sequences(will not show you which amino acids align to which amino acids.)
- It only calculates the matrix!!!
- FOR SAVING FILES(important!!!)

→ For example, if you save the file as HTML format as name result(result.html).

→ After that, even though you delete the file and save it as another format(ex: Tab delimited) with the same name(result), the file can be saved but it still in HTML format

→ So, when you want to save a file with different format, you need to save it in another name

→ In this case, if you save the tab delimited file with the name result2, the file will be saved in result2.txt

Remember:
For second file, you need to save it in another name for different format even though you already delete the first file(Don't save the 2nd file with the same name as 1st file especially save in different format)

Conclusion

BLOSUM Matrices are used to find the probability of the substitution of an amino acids with another amino acids in the block of the similar sequences. Performance of the characteristic of protein are remain unchanged after the substitution.