Pairwise-Sequence Alignment BIFX-550

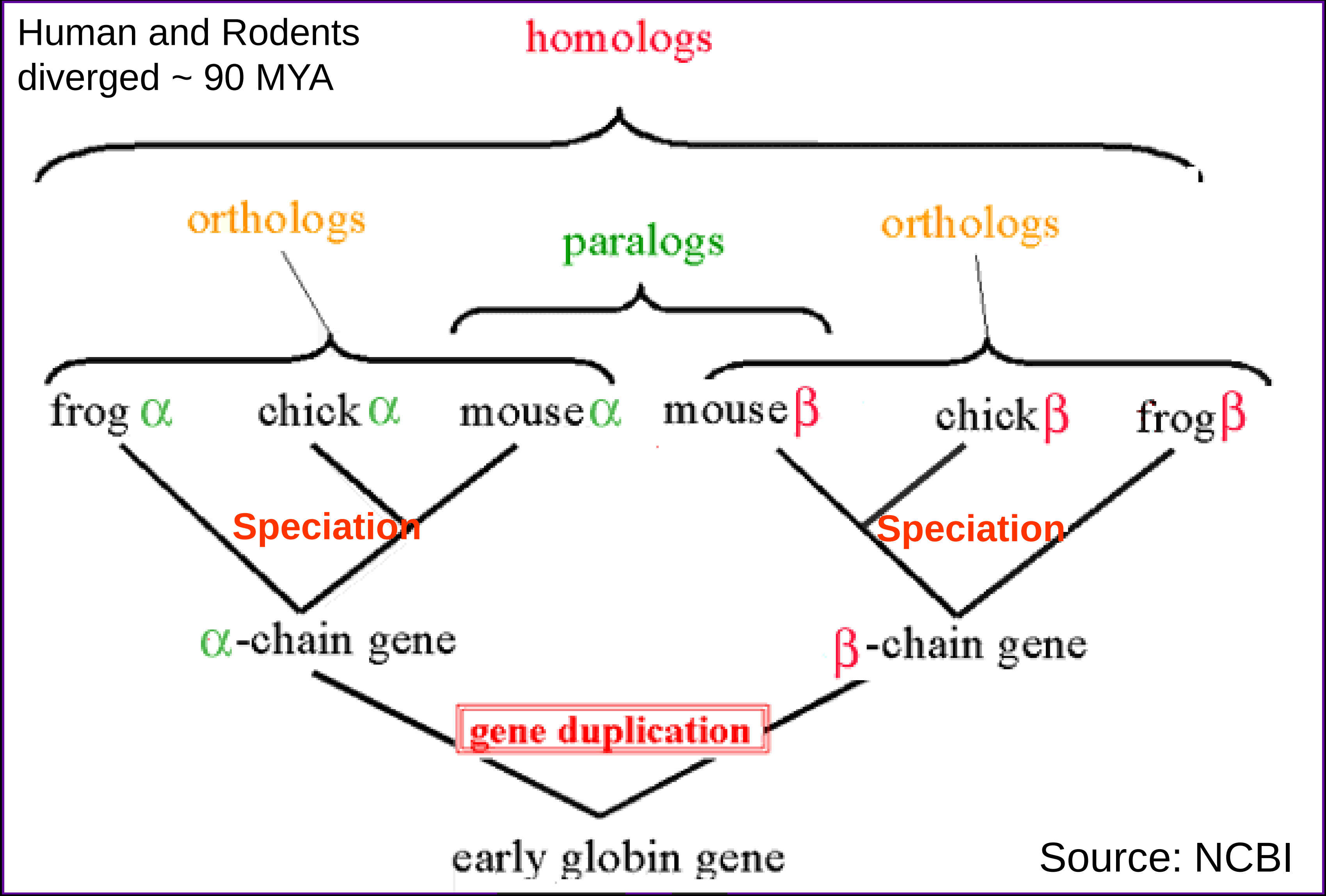
S. *Ravi*chandran, Ph.D.

Agenda

- **Please make sure you have created your galaxy account; also check to make you can login**
- Homology
  - Orthologs, Paralogs, Xenologs
- Scoring Matrices
  - PAM, BLOSUM
- Dynamic Programming
  - Global and Local Alignment
- Pairwise Alignment of DNA/Protein using NCBI Server

- Relatedness (homology) among proteins/DNAs
  - Common function?
  - Homology (common ancestor)
    - When two sequences (proteins/genes) are highly similar, they might be homologous
    - Converse is not true (lack of similarity != No Homology)
  - What is homology?

Source from NCBI but modified by Dr. S. Ravichandran



Homology

- Homology: implies evolutionary relationship
- Common ancestor
- Not measured in degrees
- Means either 2 genes/sequences are related or not
- Publications
  - Walter Fitch and Eugene Koonin

Example of Sequence Alignment

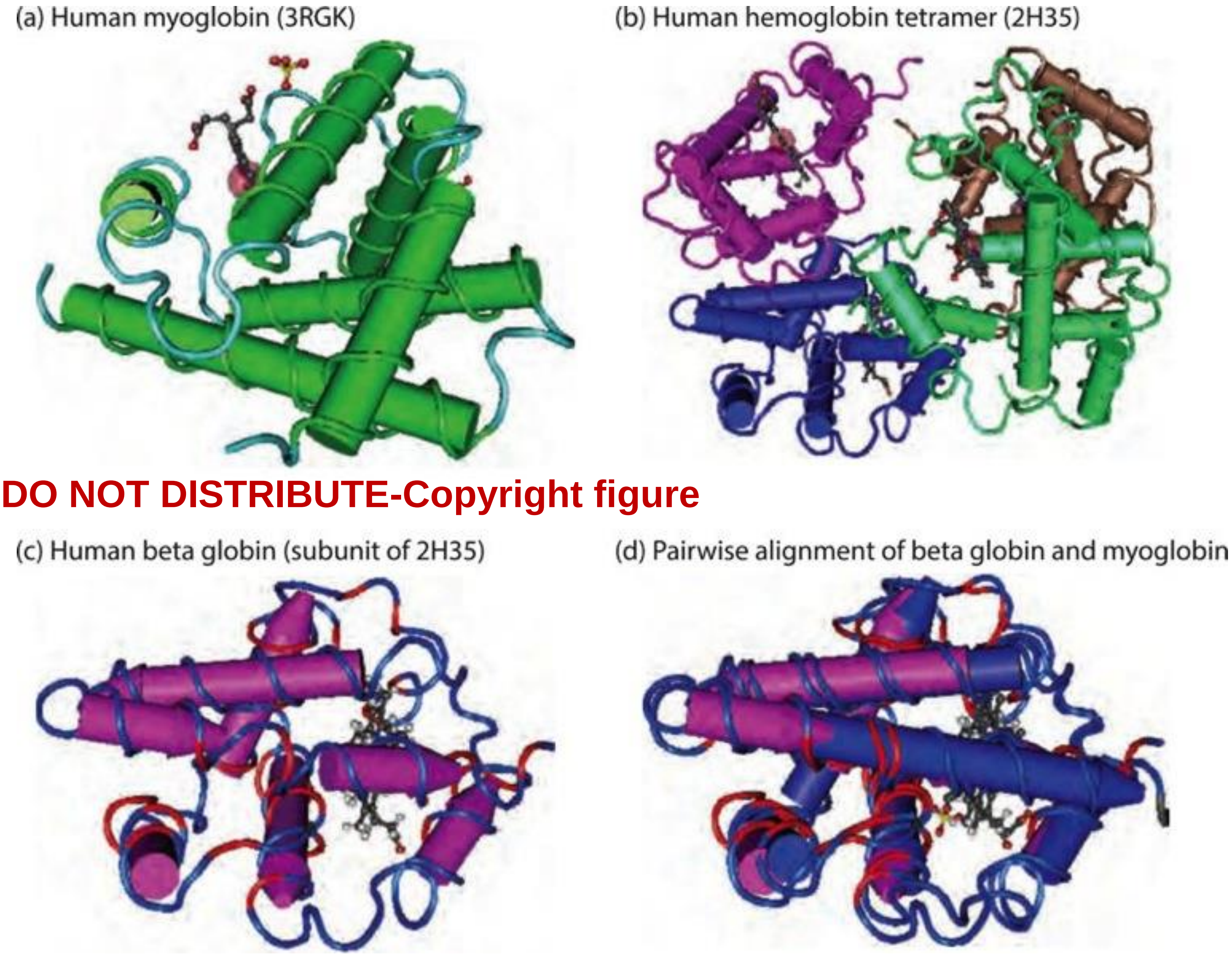
Query: h-HBB
Subjct: h-Mb

```
Query      4    LTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  61
NP_005359  3    .SDG.WQL.LNV....EA.IPGH.Q.V.I..FKGH.E.LEK.DK.KH.KSE.EMKASEDL  62

Query      62   KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK  121
NP_005359  63   .K..AT..T.LGGI.KKKGHHEAEIKP.AQS.AT.HKIPVKYLEFISECIIQ..QSKHPG  122

Query      122  EFTPPVQAAYQKVVAGVANALAHKY  146
NP_005359  123  D.GADA.G.MN.ALELFRKDM.SN.  147
```

Score	Expect	Method	Identities	Positives	Gaps
43.1 bits(94)	1e-09	Compositional matrix adjust.	37/145(26%)	43/145(29%)	2/145(1%)

??? Gap Similar Identical Score

Homologous proteins example



(a) Human myoglobin (3RGK)

(b) Human hemoglobin tetramer (2H35)

(c) Human beta globin (subunit of 2H35)

(d) Pairwise alignment of beta globin and myoglobin

Figure 3.1 Bioinformatics and Functional Genomics, (3rd Ed.) by Jonathan Pevsner

Very limited sequence similarity

How to find out whether two proteins are related?

Sequence Alignment

Comparing 3D is one way
Not all 3D information is available
So, Sequence Relationship is the common approach

What sequences to use for alignments?

- Why protein (not DNA?) sequences?
  - – Protein aa: 20 letters
  - – DNA/RNA nt: 4 letters

CCU
CCC
CCA
CCG

Proline

- More information (?) in the protein sequence
  - – Models using proteins can look back (identify ancestors) 1BYA (*Prof. Pearson several papers*); DNA (600 MYA)
    - glutathione transferases

How to identify true(?) hits?

```
Query  303  SDVICQSEPDDSFPSSGSVS---LYEVERCQQLSATILTDHQYLERTPLCAILKQKAPQQ  359
               +CQSE +DSF +  S     LYEVERCQQLSATILTDHQYLE+TPLCAILKQ APQQ
Sbjct  301  FGGVCQSEQEDSFSNISSSGSVSLYEVERCQQLSATILTDHQYLEKTPLCAILKQNAPQQ  360
```

Query: human POT1
Sbjct : Bos mutus POT1 (**Wild Yak**)
Note only part of the sequence alignments are shown

Trichinella spiralis is a nematode parasite, occurring in rodents, pigs, horses, bears, and humans, and is responsible for the disease trichinosis.

Query: human POT1

```
conserved hypothetical protein [Trichinella spiralis]
Sequence ID: ref|XP_003378812.1|Length: 382Number of Matches: 1

Range 1: 238 to 341GenPeptGraphicsNext MatchPrevious Match
Alignment statistics for match #1
Score           Expect          Method          Identities      Positives       Gaps
56.6 bits(135)  2e-05           Compositional matrix adjust.    34/111(31%)     57/111(51%)     9/111(8%)
```

```
Query  47   SFLLKVWDGTR--TPFPSWRVLIQDLVLEGDLSHIHRLQNLTIDILVYDNHVHVARSLKV  104
             ++L+VWDG+   T F    V I     + +LS   + +N   D+ +YD H  VA++LK
Sbjct  238  GWILRVWDGSSPATSFKLDSVNIDGFTADEELSL--KAENFAADVFLYDEHCTVAKALKP  295

Query  105  GSFLRIYSLHTKLQSMNSENQTMLSLEFHLHGGTSYGRGIRVLPESNSDVD  155
            G F+ +Y+LH         N       +F +H G SYGR ++++   +  V+
Sbjct  296  GDFVILYNLHLYYPYGGRSN-----CQFTMHSGNSYGRRVQLISADDELVN  341
```

?????

Can’t we manually align sequences?

- Works when
  - – We have closely related sequences
  - – Smaller number of sequences

```
Query      4    LTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  61
NP_005359  3    .SDG.WQL.LNV....EA.IPGH.Q.V.I..FKGH.E.LEK.DK.KH.KSE.EMKASEDL  62

Query      62   KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK  121
NP_005359  63   .K..AT..T.LGGI.KKKGHHEAEIKP.AQS.AT.HKIPVKYLEFISECIIQ..QSKHPG  122

Query      122  EFTPPVQAAYQKVVAGVANALAHKY  146
NP_005359  123  D.GADA.G.MN.ALELFRKDM.SN.  147
```

- Not work when
  - – We have to align query against a DB (modest size of 100s)
    - Usual situation

- We need algorithms for alignment/evaluation
  - – Math/Statistics

Goal of Sequence Alignment

- To extract the information whether the two sequences have similar aa/nt in proper order and to access whether they are homologous
- Gaps indicate what?
  - To capture the evolution of the sequence
  - In that process allowing for Insertions, Deletions and substitutions
- To access the alignments, we need to score each residue alignment
  - Identical, similar or gap (creation & extension penalty)
  - %identity/similarity etc.

Simple Match/Mismatch Scoring Matrix

Note

A → A +2

T → T +2

(Not the same for amino acids)

Equal probability so scores are same

Simple scoring matrix

Match: +2; Mismatch: -3

25% Probability of each NT occurrence

	A	T	G	C
A	+2	-3	-3	-3
T	-3	+2	-3	-3
G	-3	-3	+2	-3
C	-3	-3	-3	+2

Same points for mismatches: -3

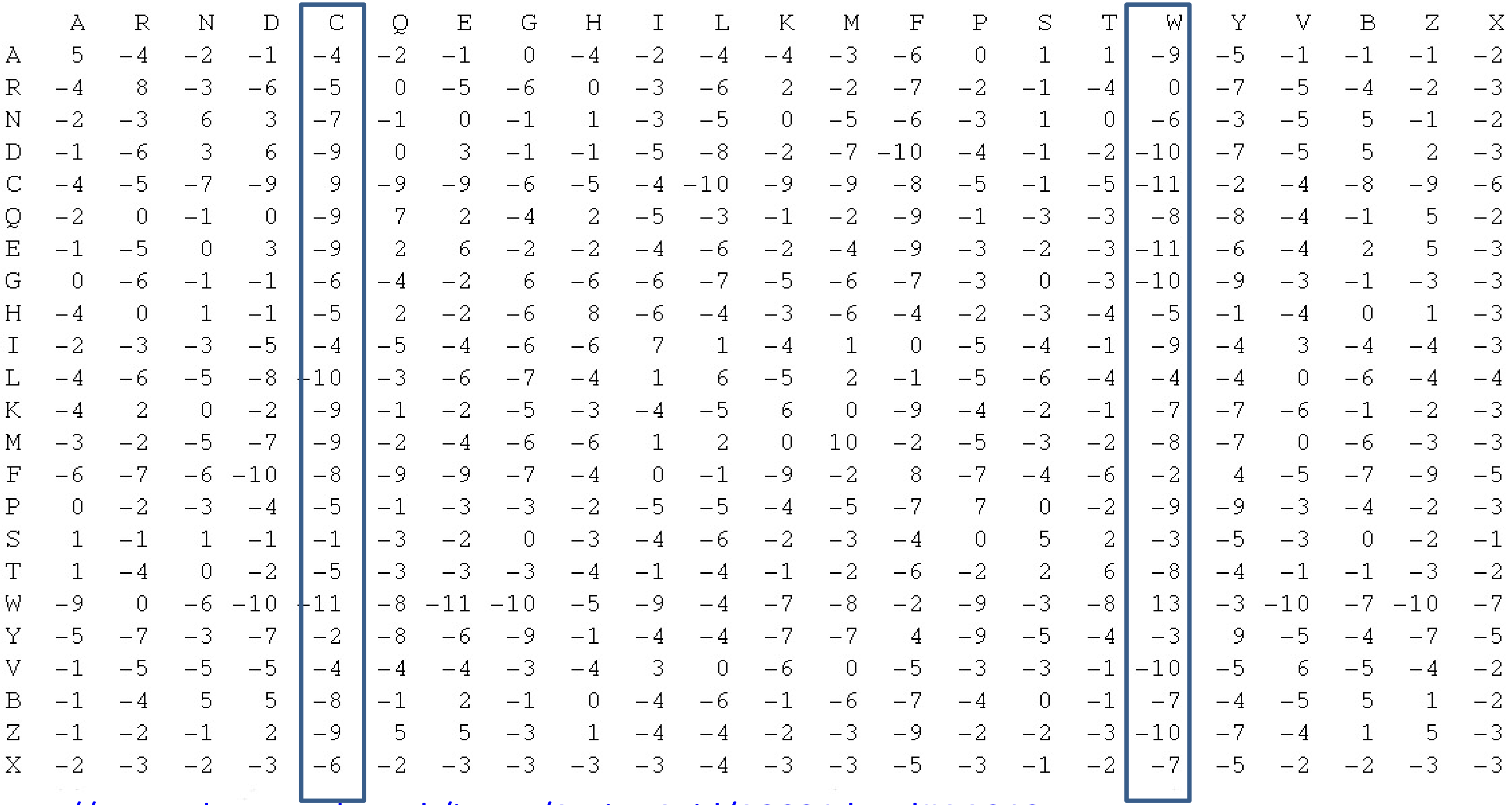
Negative Score

Common same substitution; A → A: 5
Less common; W → 13

PAM70

YW:7

WW:13



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X
A	5	-4	-2	-1	-4	-2	-1	0	-4	-2	-4	-4	-3	-6	0	1	1	-9	-5	-1	-1	-1	-2
R	-4	8	-3	-6	-5	0	-5	-6	0	-3	-6	2	-2	-7	-2	-1	-4	0	-7	-5	-4	-2	-3
N	-2	-3	6	3	-7	-1	0	-1	1	-3	-5	0	-5	-6	-3	1	0	-6	-3	-5	5	-1	-2
D	-1	-6	3	6	-9	0	3	-1	-1	-5	-8	-2	-7	-10	-4	-1	-2	-10	-7	-5	5	2	-3
C	-4	-5	-7	-9	9	-9	-9	-6	-5	-4	-10	-9	-9	-8	-5	-1	-5	-11	-2	-4	-8	-9	-6
Q	-2	0	-1	0	-9	7	2	-4	2	-5	-3	-1	-2	-9	-1	-3	-3	-8	-8	-4	-1	5	-2
E	-1	-5	0	3	-9	2	6	-2	-2	-4	-6	-2	-4	-9	-3	-2	-3	-11	-6	-4	2	5	-3
G	0	-6	-1	-1	-6	-4	-2	6	-6	-6	-7	-5	-6	-7	-3	0	-3	-10	-9	-3	-1	-3	-3
H	-4	0	1	-1	-5	2	-2	-6	8	-6	-4	-3	-6	-4	-2	-3	-4	-5	-1	-4	0	1	-3
I	-2	-3	-3	-5	-4	-5	-4	-6	-6	7	1	-4	1	0	-5	-4	-1	-9	-4	3	-4	-4	-3
L	-4	-6	-5	-8	-10	-3	-6	-7	-4	1	6	-5	2	-1	-5	-6	-4	-4	-4	0	-6	-4	-4
K	-4	2	0	-2	-9	-1	-2	-5	-3	-4	-5	6	0	-9	-4	-2	-1	-7	-7	-6	-1	-2	-3
M	-3	-2	-5	-7	-9	-2	-4	-6	-6	1	2	0	10	-2	-5	-3	-2	-8	-7	0	-6	-3	-3
F	-6	-7	-6	-10	-8	-9	-9	-7	-4	0	-1	-9	-2	8	-7	-4	-6	-2	4	-5	-7	-9	-5
P	0	-2	-3	-4	-5	-1	-3	-3	-2	-5	-5	-4	-5	-7	7	0	-2	-9	-9	-3	-4	-2	-3
S	1	-1	1	-1	-1	-3	-2	0	-3	-4	-6	-2	-3	-4	0	5	2	-3	-5	-3	0	-2	-1
T	1	-4	0	-2	-5	-3	-3	-3	-4	-1	-4	-1	-2	-6	-2	2	6	-8	-4	-1	-1	-3	-2
W	-9	0	-6	-10	-11	-8	-11	-10	-5	-9	-4	-7	-8	-2	-9	-3	-8	13	-3	-10	-7	-10	-7
Y	-5	-7	-3	-7	-2	-8	-6	-9	-1	-4	-4	-7	-7	4	-9	-5	-4	-3	9	-5	-4	-7	-5
V	-1	-5	-5	-5	-4	-4	-4	-3	-4	3	0	-6	0	-5	-3	-3	-1	-10	-5	6	-5	-4	-2
B	-1	-4	5	5	-8	-1	2	-1	0	-4	-6	-1	-6	-7	-4	0	-1	-7	-4	-5	5	1	-2
Z	-1	-2	-1	2	-9	5	5	-3	1	-4	-4	-2	-3	-9	-2	-2	-3	-10	-7	-4	1	5	-3
X	-2	-3	-2	-3	-6	-2	-3	-3	-3	-3	-4	-3	-3	-5	-3	-1	-2	-7	-5	-2	-2	-3	-3

http://www.sbcs.qmul.ac.uk/iupac/AminoAcid/A2021.html#AA212

Symmetric matrix
Mismatches depend on the type of AA

PAM1

PAM1	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	0.9890	0.0004	0.0003	0.0004	0.0003	0.0004	0.0008	0.0011	0.0001	0.0002	0.0005	0.0005	0.0003	0.0001	0.0006	0.0023	0.0011	0.0000	0.0001	0.0015
R	0.0005	0.9907	0.0004	0.0002	0.0001	0.0011	0.0005	0.0004	0.0004	0.0001	0.0004	0.0033	0.0001	0.0000	0.0002	0.0005	0.0005	0.0001	0.0002	0.0002
N	0.0005	0.0005	0.9888	0.0021	0.0001	0.0007	0.0006	0.0009	0.0007	0.0002	0.0002	0.0013	0.0001	0.0001	0.0002	0.0017	0.0011	0.0000	0.0002	0.0001
D	0.0006	0.0002	0.0018	0.9905	0.0000	0.0005	0.0030	0.0007	0.0003	0.0000	0.0000	0.0005	0.0000	0.0000	0.0002	0.0008	0.0006	0.0000	0.0001	0.0000
C	0.0012	0.0002	0.0002	0.0000	0.9946	0.0001	0.0000	0.0002	0.0001	0.0002	0.0003	0.0000	0.0002	0.0003	0.0000	0.0009	0.0004	0.0001	0.0003	0.0008
Q	0.0009	0.0016	0.0008	0.0007	0.0000	0.9856	0.0028	0.0004	0.0009	0.0002	0.0008	0.0022	0.0004	0.0001	0.0005	0.0009	0.0008	0.0001	0.0001	0.0003
E	0.0011	0.0004	0.0005	0.0028	0.0000	0.0018	0.9890	0.0003	0.0003	0.0001	0.0002	0.0015	0.0001	0.0000	0.0003	0.0007	0.0005	0.0000	0.0001	0.0004
G	0.0012	0.0003	0.0006	0.0005	0.0001	0.0002	0.0002	0.9952	0.0001	0.0000	0.0001	0.0002	0.0000	0.0000	0.0001	0.0008	0.0002	0.0000	0.0000	0.0001
H	0.0005	0.0008	0.0013	0.0006	0.0001	0.0014	0.0007	0.0003	0.9895	0.0002	0.0003	0.0008	0.0002	0.0004	0.0002	0.0006	0.0007	0.0001	0.0013	0.0002
I	0.0002	0.0001	0.0001	0.0000	0.0001	0.0001	0.0001	0.0000	0.0001	0.9878	0.0035	0.0002	0.0010	0.0005	0.0001	0.0001	0.0006	0.0000	0.0001	0.0051
L	0.0005	0.0002	0.0001	0.0000	0.0001	0.0003	0.0001	0.0001	0.0001	0.0022	0.9919	0.0002	0.0012	0.0010	0.0002	0.0002	0.0002	0.0001	0.0002	0.0014
K	0.0006	0.0030	0.0010	0.0005	0.0000	0.0014	0.0015	0.0003	0.0003	0.0002	0.0003	0.9883	0.0002	0.0000	0.0003	0.0007	0.0009	0.0000	0.0001	0.0003
M	0.0009	0.0002	0.0001	0.0000	0.0001	0.0006	0.0003	0.0001	0.0002	0.0026	0.0048	0.0005	0.9859	0.0009	0.0000	0.0004	0.0007	0.0001	0.0002	0.0012
F	0.0002	0.0000	0.0001	0.0000	0.0001	0.0001	0.0000	0.0000	0.0002	0.0007	0.0022	0.0001	0.0005	0.9923	0.0001	0.0001	0.0002	0.0003	0.0022	0.0005
P	0.0010	0.0003	0.0002	0.0003	0.0000	0.0004	0.0004	0.0002	0.0001	0.0001	0.0004	0.0004	0.0000	0.0000	0.9943	0.0008	0.0007	0.0000	0.0001	0.0002
S	0.0029	0.0005	0.0013	0.0007	0.0003	0.0005	0.0007	0.0010	0.0002	0.0001	0.0003	0.0006	0.0002	0.0001	0.0006	0.9862	0.0033	0.0000	0.0002	0.0003
T	0.0014	0.0005	0.0008	0.0005	0.0001	0.0005	0.0005	0.0002	0.0002	0.0005	0.0004	0.0009	0.0003	0.0001	0.0005	0.0032	0.9879	0.0000	0.0001	0.0014
W	0.0001	0.0004	0.0001	0.0000	0.0001	0.0001	0.0001	0.0002	0.0001	0.0002	0.0005	0.0001	0.0001	0.0010	0.0000	0.0002	0.0001	0.9956	0.0010	0.0001
Y	0.0002	0.0003	0.0003	0.0001	0.0001	0.0001	0.0001	0.0001	0.0009	0.0002	0.0005	0.0002	0.0001	0.0028	0.0001	0.0004	0.0002	0.0004	0.9924	0.0004
V	0.0017	0.0002	0.0001	0.0000	0.0002	0.0002	0.0003	0.0001	0.0001	0.0042	0.0019	0.0002	0.0004	0.0003	0.0002	0.0002	0.0012	0.0000	0.0002	0.9884

PAM250

PAM250	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	0.1350	0.0460	0.0425	0.0501	0.0212	0.0359	0.0580	0.0827	0.0194	0.0480	0.0717	0.0535	0.0195	0.0239	0.0484	0.0774	0.0707	0.0057	0.0195	0.0708
R	0.0677	0.1583	0.0485	0.0496	0.0114	0.0530	0.0645	0.0583	0.0271	0.0330	0.0566	0.1097	0.0154	0.0193	0.0367	0.0580	0.0587	0.0090	0.0215	0.0437
N	0.0733	0.0568	0.1092	0.0881	0.0124	0.0442	0.0722	0.0801	0.0310	0.0304	0.0473	0.0713	0.0138	0.0202	0.0366	0.0741	0.0687	0.0057	0.0235	0.0413
D	0.0733	0.0493	0.0747	0.1593	0.0091	0.0468	0.1095	0.0752	0.0259	0.0239	0.0374	0.0662	0.0114	0.0143	0.0385	0.0671	0.0616	0.0040	0.0173	0.0352
C	0.0884	0.0323	0.0299	0.0260	0.2660	0.0215	0.0293	0.0466	0.0173	0.0446	0.0673	0.0311	0.0185	0.0340	0.0225	0.0619	0.0544	0.0103	0.0293	0.0689
Q	0.0748	0.0751	0.0534	0.0666	0.0107	0.0705	0.0863	0.0585	0.0309	0.0375	0.0653	0.0840	0.0183	0.0224	0.0434	0.0625	0.0621	0.0070	0.0220	0.0486
E	0.0781	0.0589	0.0563	0.1007	0.0094	0.0558	0.1334	0.0608	0.0256	0.0314	0.0492	0.0774	0.0144	0.0165	0.0410	0.0636	0.0605	0.0048	0.0175	0.0446
G	0.0884	0.0424	0.0496	0.0549	0.0119	0.0300	0.0483	0.3387	0.0170	0.0208	0.0339	0.0454	0.0103	0.0124	0.0318	0.0657	0.0478	0.0052	0.0130	0.0326
H	0.0647	0.0616	0.0601	0.0591	0.0138	0.0496	0.0635	0.0530	0.0946	0.0353	0.0612	0.0667	0.0170	0.0399	0.0353	0.0578	0.0580	0.0108	0.0539	0.0440
I	0.0649	0.0304	0.0239	0.0221	0.0145	0.0243	0.0315	0.0264	0.0143	0.1460	0.1779	0.0358	0.0403	0.0510	0.0251	0.0394	0.0542	0.0086	0.0276	0.1415
L	0.0597	0.0320	0.0228	0.0213	0.0134	0.0261	0.0305	0.0264	0.0153	0.1094	0.2390	0.0359	0.0435	0.0649	0.0271	0.0368	0.0462	0.0110	0.0327	0.1060
K	0.0714	0.0995	0.0552	0.0604	0.0100	0.0538	0.0768	0.0567	0.0267	0.0354	0.0576	0.1240	0.0165	0.0191	0.0396	0.0616	0.0635	0.0059	0.0200	0.0464
M	0.0671	0.0361	0.0275	0.0268	0.0153	0.0303	0.0370	0.0332	0.0175	0.1027	0.1798	0.0425	0.0608	0.0583	0.0259	0.0439	0.0535	0.0104	0.0311	0.1004
F	0.0461	0.0253	0.0225	0.0188	0.0157	0.0208	0.0237	0.0224	0.0231	0.0726	0.1501	0.0276	0.0326	0.2041	0.0191	0.0315	0.0372	0.0301	0.1054	0.0712
P	0.0834	0.0429	0.0366	0.0453	0.0093	0.0359	0.0525	0.0512	0.0182	0.0320	0.0561	0.0510	0.0130	0.0171	0.2614	0.0656	0.0632	0.0041	0.0160	0.0454
S	0.1006	0.0512	0.0558	0.0597	0.0193	0.0390	0.0614	0.0799	0.0225	0.0379	0.0575	0.0600	0.0166	0.0213	0.0495	0.0997	0.0862	0.0062	0.0213	0.0545
T	0.0899	0.0507	0.0507	0.0536	0.0166	0.0379	0.0572	0.0569	0.0221	0.0510	0.0707	0.0605	0.0198	0.0246	0.0467	0.0844	0.1101	0.0058	0.0211	0.0698
W	0.0342	0.0366	0.0197	0.0164	0.0149	0.0201	0.0215	0.0291	0.0194	0.0381	0.0793	0.0264	0.0181	0.0937	0.0144	0.0285	0.0274	0.3398	0.0843	0.0382
Y	0.0468	0.0351	0.0326	0.0284	0.0169	0.0253	0.0312	0.0292	0.0387	0.0489	0.0942	0.0360	0.0217	0.1312	0.0222	0.0393	0.0397	0.0336	0.1955	0.0535
V	0.0803	0.0337	0.0271	0.0273	0.0187	0.0264	0.0376	0.0346	0.0149	0.1185	0.1443	0.0394	0.0330	0.0419	0.0299	0.0475	0.0622	0.0072	0.0253	0.1501

How to extract GONNET matrices using R?

```
#install.packages(“TKF”)
library(TKF)
data("GONNET")
PAM1     <- PAMn(GONNET,1)
round(PAM1[,1:20],3)
PAM250 <- PAMn(GONNET, 250)
round(PAM250[,1:20],3)
```

GONNET is an extension of PAM matrices

Substitution Matrices

To create an alignment and to identify homologous sequences, we need a **score**. What score should we assign?

Model/Probability Model

- We need Probability to get through this part
- Why?
  - Aligning two sequences; What is the prob. of this alignment compared to other alignments?
  - Random sequence model or Null Model (base model to compare with anything)

- $x_1 \ldots x_q$;
- Null prob $q_{x1} q_{x2} \ldots q_{xn}$ = $\prod_{i=1}^{n} q_{x_i}$

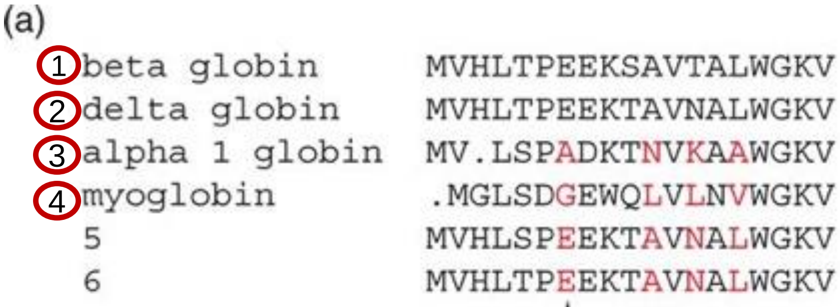
Dayhoff Matrix in 7 Steps

- 1978
- Step 1 of 7:
  - What mutations are accepted in closely related sequences
    - Model Accepted Point Mutation
      - Easier name: Point Accepted Mutation (PAM)
  - Collected closely (85% or >) sequences
    - Ungapped MSA
  - Used phylogenetic trees rather than comparing two sequences directly

Step 1 of 7

PAUP was used for Phylogenetic Analysis

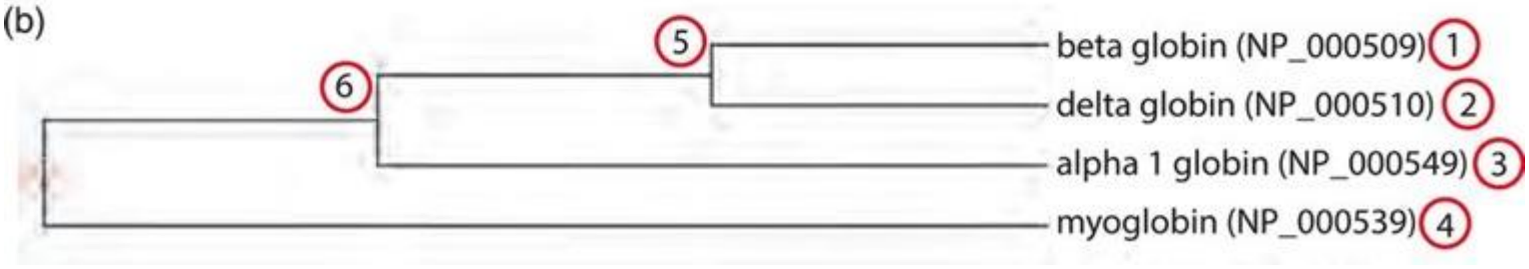
Sequence Alignment of Human globins/myoglobins



(a)

```
1 beta globin      MVHLTPEEKSAVTALWGKV
2 delta globin     MVHLTPEEKTAVNALWGKV
3 alpha 1 globin   MV.LSPADKTNVKAAWGKV
4 myoglobin        .MGLSDGEWQLVLNVWGKV
  5                MVHLSPEEKTAVNALWGKV
  6                MVHLTPEEKTAVNALWGKV
```

Ancestral is E and has evolved to become G or A



(b)
5
6
beta globin (NP_000509) 1
delta globin (NP_000510) 2
alpha 1 globin (NP_000549) 3
myoglobin (NP_000539) 4

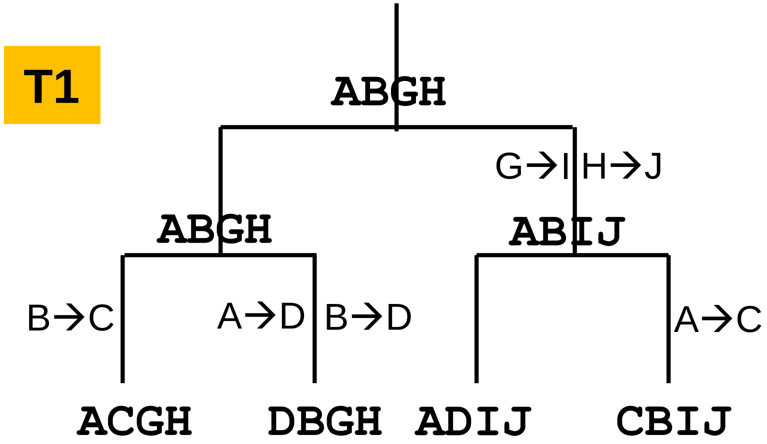
Figure 3.7 Bioinformatics and Functional Genomics, (3^{rd} Ed.) by Jonathan Pevsner

Example

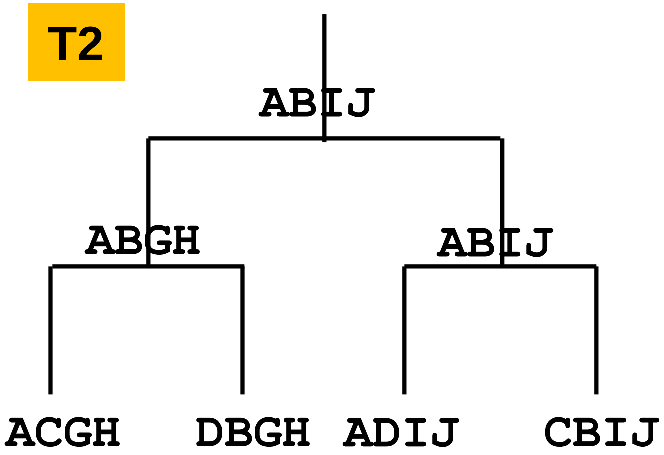
ACGH
DBGH
ADIJ
CBIJ

UNGAPPED

T1

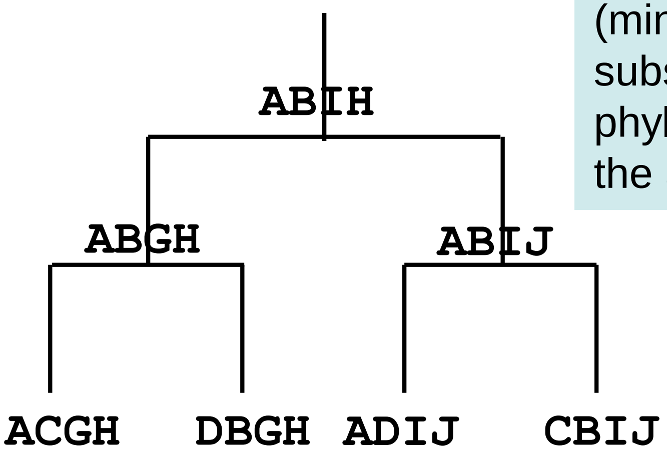


T2

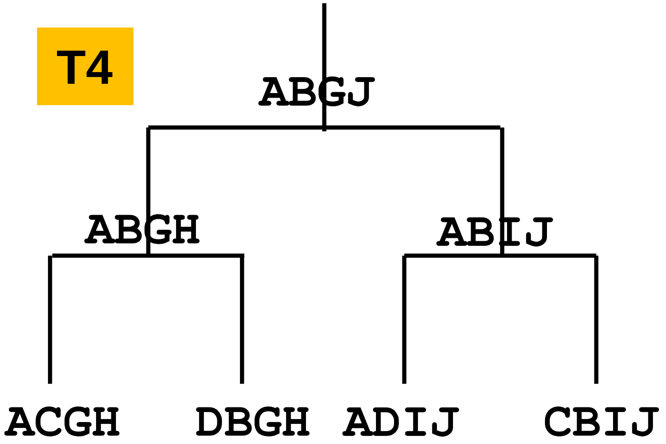


Four Parsimonious (minimum # of substitutions) phylogenetic trees for the alignment

T3



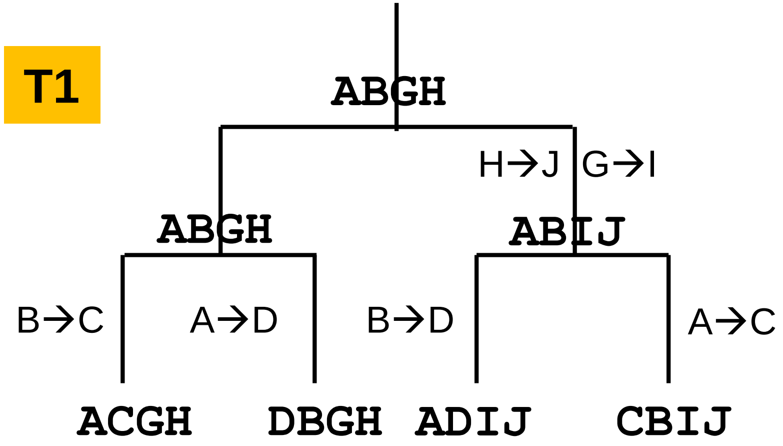
T4



Ref: Didier Gonze, Borodovsky & Ekisheva (2007)

Alignments → Trees

- Ungapped alignments → Trees (T1-T4)
- Each tree produces 6 alignments



A B G H
A B G H

A B G H
A B I J

A B G H
A C G H

A B G H
D B G H

A B I J
A D I J

A B I J
C B I J

Matrix of **A(i,j)**

APM counts

	A	B	C	D	G	H	I	J
A	🟥	0	4	4	0	0	0	0
B	0	🟥	4	4	0	0	0	0
C	4	4	🟥	0	0	0	0	0
D	4	4	0	🟥	0	0	0	0
G	0	0	0	0	🟥	0	4	0
H	0	0	0	0	0	🟥	0	4
I	0	0	0	0	4	0	🟥	0
J	0	0	0	0	0	4	0	🟥
Total	**8**	**8**	**8**	**8**	**4**	**4**	**4**	**4**

Step1-Outcome

- What amino acid substitutions are likely and which ones are unlikely
  - C and W had shown to be sparsely substituted
  - N, S are commonly substituted from Dayhoff’s data

- Relative mutability, m_j (given the short evolutionary period)
- **Note the mutation rate is different for diff. AA**

AA	A	B	I	H	G	J	C	D
Changes	8	8	4	4	4	4	8	8
Freq. of occurrence	40	40	24	24	24	24	8	8
Relative mutability m_j	0.2	0.2	0.167	0.167	0.167	0.167	1	1

$$m_j = \frac{\text{number of changes of } j}{\text{number of occurances of } j}$$

↓

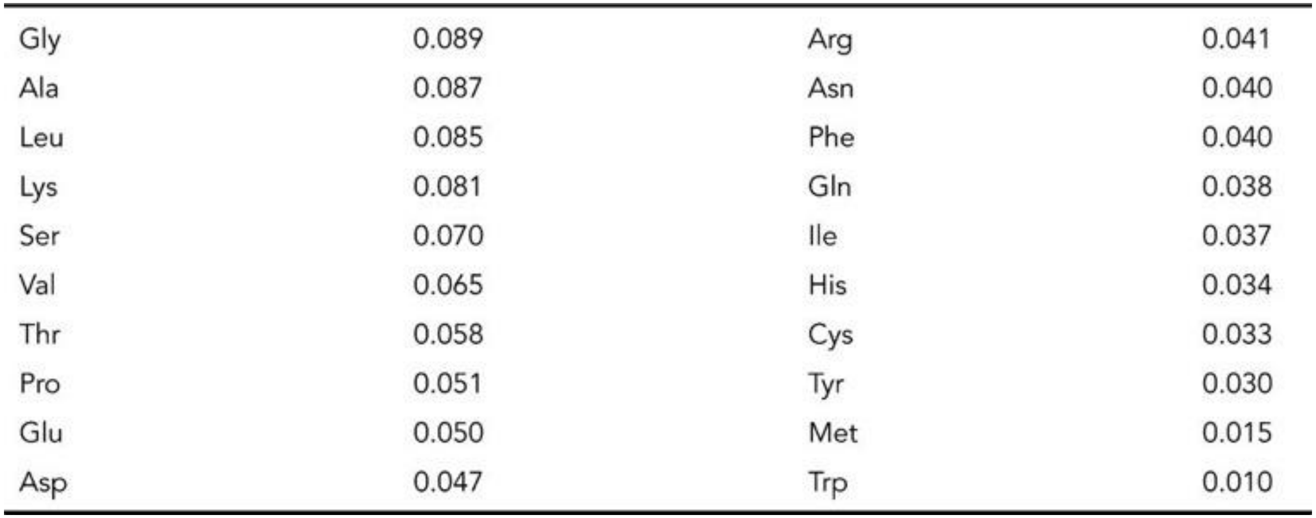
Number of times the residue occurs in the alignment (gleaned from the tree)

```
ABGH    ABGH     ABGH
ABGH    ABIJ     ACGH

ADIJ  ABIJ     ABIJ
CBIJ  ADIJ     CBIJ
```

Steps 2 and 3 (of 7): Frequency of Occurrence & Relative Mutability

f_i

Table 3.1

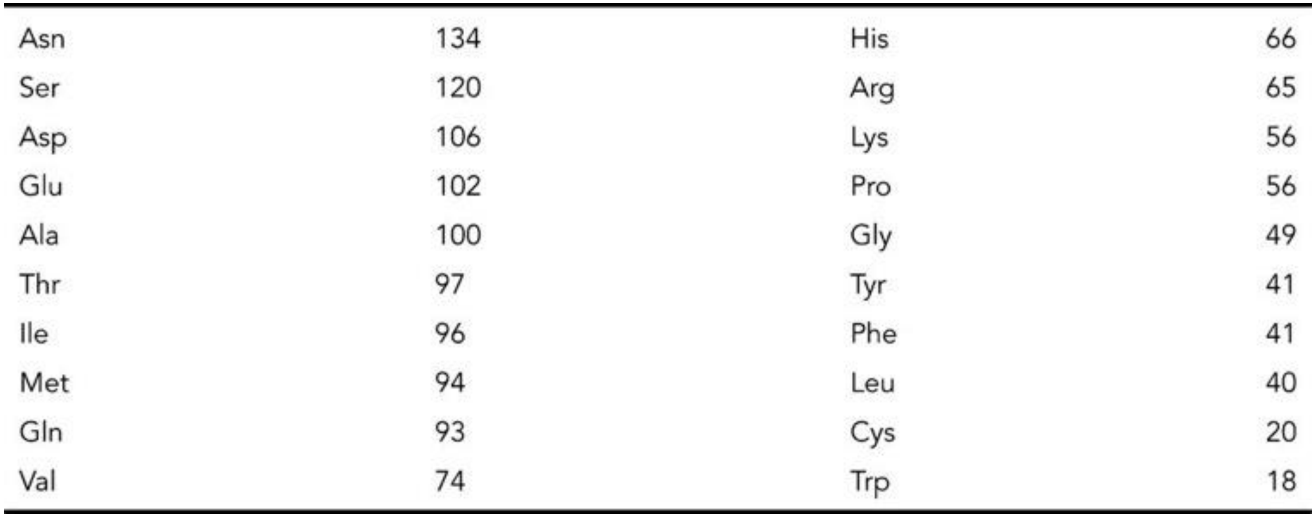


Gly	0.089	Arg	0.041
Ala	0.087	Asn	0.040
Leu	0.085	Phe	0.040
Lys	0.081	Gln	0.038
Ser	0.070	Ile	0.037
Val	0.065	His	0.034
Thr	0.058	Cys	0.033
Pro	0.051	Tyr	0.030
Glu	0.050	Met	0.015
Asp	0.047	Trp	0.010

Normalized Frequency Sum to one

If the freq are 1/20 they all would be 0.05

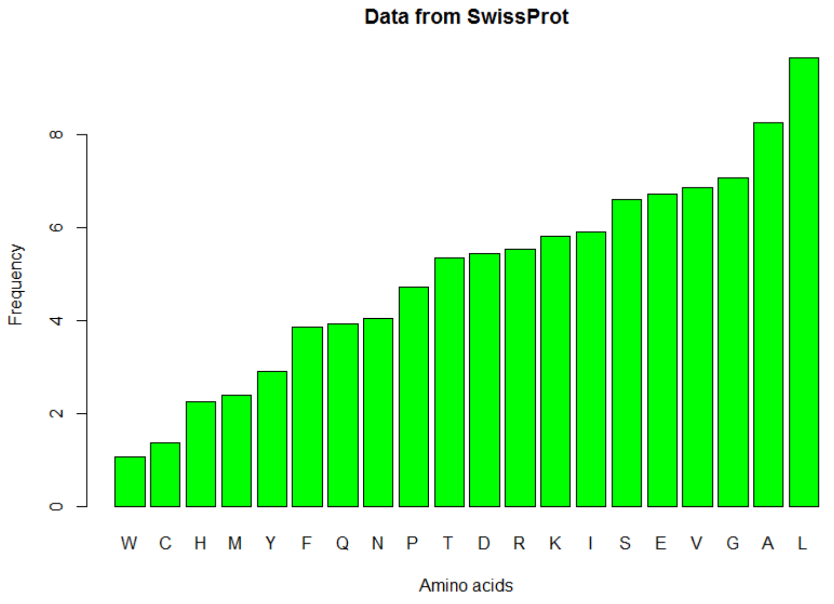
Table 3.2

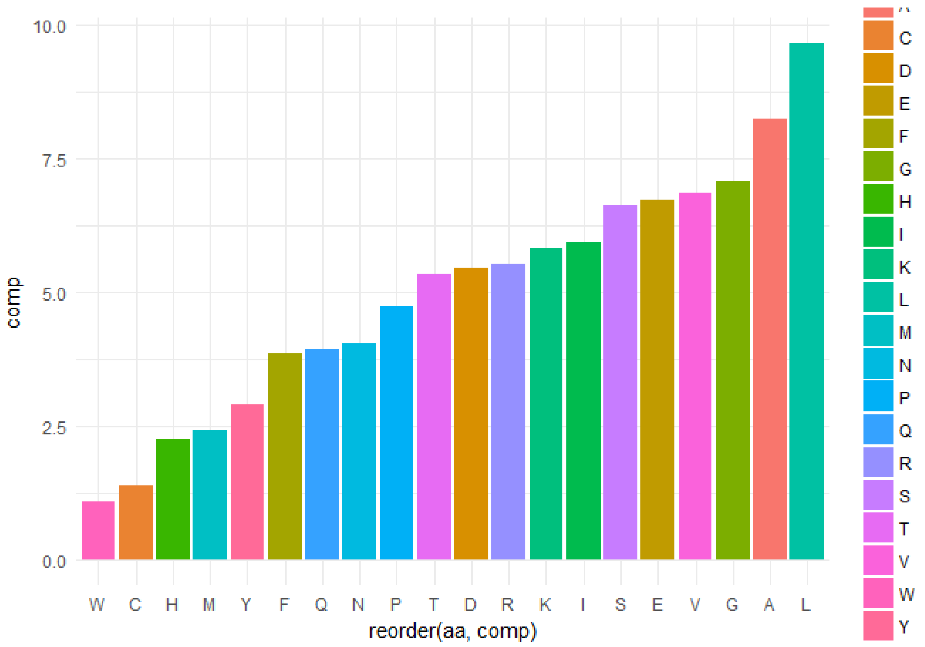


Asn	134	His	66
Ser	120	Arg	65
Asp	106	Lys	56
Glu	102	Pro	56
Ala	100	Gly	49
Thr	97	Tyr	41
Ile	96	Phe	41
Met	94	Leu	40
Gln	93	Cys	20
Val	74	Trp	18

Relative Mutability

Ala (A) 8.25	Gln (Q) 3.93	Leu (L) 9.65	Ser (S) 6.62
Arg (R) 5.53	Glu (E) 6.73	Lys (K) 5.81	Thr (T) 5.35
Asn (N) 4.05	Gly (G) 7.07	Met (M) 2.41	Trp (W) 1.09
Asp (D) 5.46	His (H) 2.27	Phe (F) 3.86	Tyr (Y) 2.91
Cys (C) 1.38	Ile (I) 5.92	Pro (P) 4.73	Val (V) 6.86





Effective frequencies (f_j) for more than 1 block

Amino acid	Gly	Ala	Leu	Lys	Ser	Val	Thr
Frequency f	0.089	0.087	0.085	0.081	0.070	0.065	0.058

Amino acid	Pro	Glu	Asp	Arg	Asn	Phe	Gln
Frequency f	0.051	0.050	0.047	0.041	0.040	0.040	0.038

Amino acid	Ile	His	Cys	Tyr	Met	Trp
Frequency f	0.037	0.034	0.033	0.030	0.015	0.010

Effective frequency of the 20 amino acids determined for the original alignment data(70 blocks) ***(Dayhoff et al., 1978)***

$$f_j = k \sum_b q_j^{(b)} N^{(b)}$$

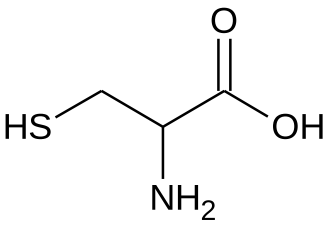
b: blocks
$q_j^{(b)}$ is the observed frequency of amino acid j in block b
$N^{(b)}$ is the number of substitutions in a tree built for b

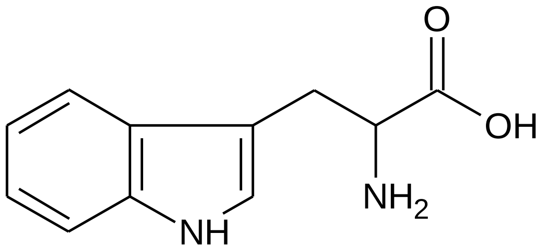
K is chosen such that sum of f_j = 1

Relative Mutability

- Different for different AAs
  - W and C are less mutable
    - Why?
  - N,S,D,E are more mutable
    - Why

AA	m_i	AA	m_i
N	**134**	H	66
S	**120**	R	65
D	**106**	K	56
E	**102**	P	56
A	100	G	49
T	97	Y	41
I	96	F	41
M	94	L	40
N	93	**C**	**20**
V	74	**W**	**18**





Alanine had been arbitrarily set to 100 (Dayhoff, 1978)

$$f(j) = \frac{n(j)}{N}$$

Frequency of jth amino acid

Entry, A(i,j) will contain the # of times j is mutated to i

$$m(j) = \frac{\sum_{i=1, i \neq j}^{20} A(i,j)}{n(j)}$$

Mutability of jth amino acid
A(i,j) is the count of j → i

$$\frac{1}{Nf(j)} = \frac{1}{n(j)} = \frac{m(j)}{\sum_{i=1, i \neq j}^{20} A(i,j)}$$

The above equation can be rewritten as

Note the above calculation ignores self mutation (A → A etc.)

$$f(j) = \frac{n(j)}{N}$$

$$m(j) = \frac{\sum_{i=1, i\neq j}^{20} A(i,j)}{n(j)}$$

$$\frac{1}{Nf(j)} = \frac{1}{n(j)} = \frac{m(j)}{\sum_{i=1, i\neq j}^{20} A(i,j)}$$

Goal is to compute probability matrix

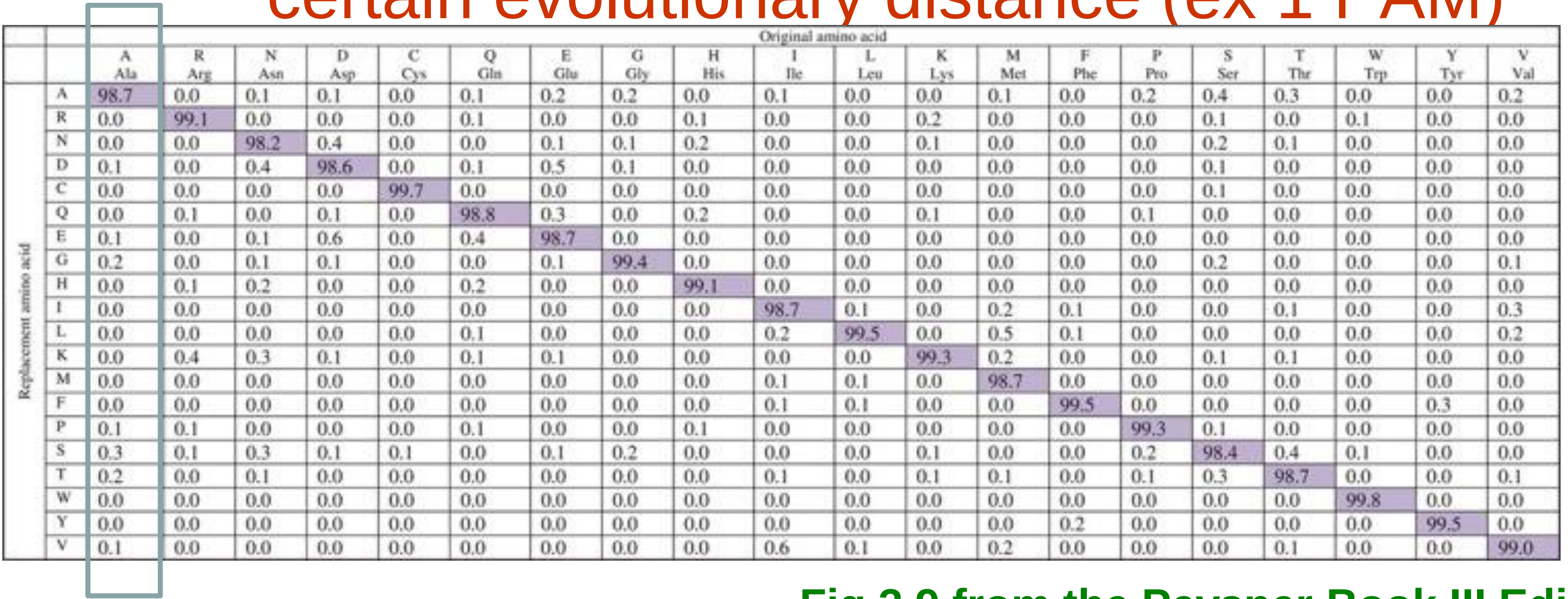
M(i,j) is the probability of the aa in the column j having been substituted by an aa in row i over an evolutionary distance. Note this only includes non-diagonal entries

$$M(i,j) = \lambda A(i,j) \frac{m(j)}{\sum_{i=1, i\neq j}^{20} A(i,j)} = \frac{\lambda A(i,j)}{Nf(j)}$$

λ is a constant

Equation only computes the non-diagonal and the diagonal entry is just one minus of that quantity

Step 4 (of 7): Mutation Probability Matrix (M_{ij}) for a certain evolutionary distance (ex 1 PAM)



		Original amino acid																			
		A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
Replacement amino acid	A	98.7	0.0	0.1	0.1	0.0	0.1	0.2	0.2	0.0	0.1	0.0	0.0	0.1	0.0	0.2	0.4	0.3	0.0	0.0	0.2
	R	0.0	99.1	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.2	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0
	N	0.0	0.0	98.2	0.4	0.0	0.0	0.1	0.1	0.2	0.0	0.0	0.1	0.0	0.0	0.0	0.2	0.1	0.0	0.0	0.0
	D	0.1	0.0	0.4	98.6	0.0	0.1	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
	C	0.0	0.0	0.0	0.0	99.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
	Q	0.0	0.1	0.0	0.1	0.0	98.8	0.3	0.0	0.2	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
	E	0.1	0.0	0.1	0.6	0.0	0.4	98.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	G	0.2	0.0	0.1	0.1	0.0	0.0	0.1	99.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.1
	H	0.0	0.1	0.2	0.0	0.0	0.2	0.0	0.0	99.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	I	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	98.7	0.1	0.0	0.2	0.1	0.0	0.0	0.1	0.0	0.0	0.3
	L	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.2	99.5	0.0	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.2
	K	0.0	0.4	0.3	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0	99.3	0.2	0.0	0.0	0.1	0.1	0.0	0.0	0.0
	M	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	98.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	F	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	99.5	0.0	0.0	0.0	0.0	0.3	0.0
	P	0.1	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	99.3	0.1	0.0	0.0	0.0	0.0
	S	0.3	0.1	0.3	0.1	0.1	0.0	0.1	0.2	0.0	0.0	0.0	0.1	0.0	0.0	0.2	98.4	0.4	0.1	0.0	0.0
	T	0.2	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.1	0.0	0.1	0.3	98.7	0.0	0.0	0.1
	W	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	99.8	0.0	0.0
	Y	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	99.5	0.0
	V	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.1	0.0	0.2	0.0	0.0	0.0	0.1	0.0	0.0	99.0

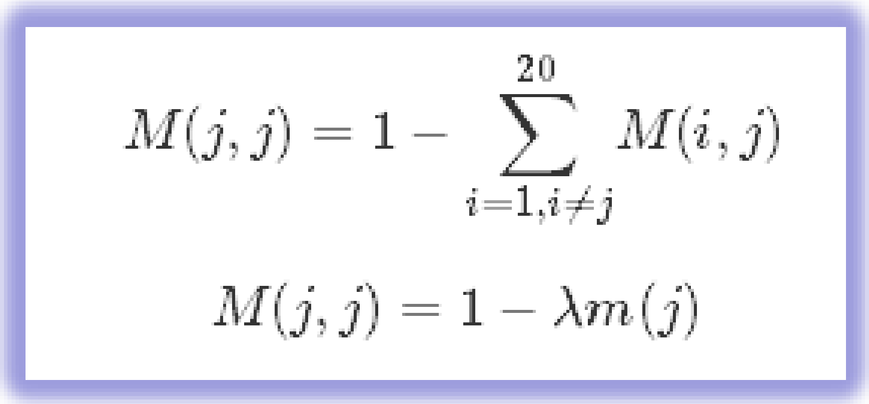
A→V
!=
V→A

100% or 1 (probability values)

Fig 3.9 from the Pevsner Book III Edition

Each entry (i,j) shows the probability of an amino acid (j; columns) to be replaced by another amino acid (i, row) over an evolutionary distance of 1 PAM

What is 1 PAM? 1% of amino acids have changed in the sequences from which this data is derived (Note the time of evolution could be different for different sequences)



$$M(j,j) = 1 - \sum_{i=1, i \neq j}^{20} M(i,j)$$

$$M(j,j) = 1 - \lambda m(j)$$

λ is the same constant (can be derived with little algebra; not showing the steps)

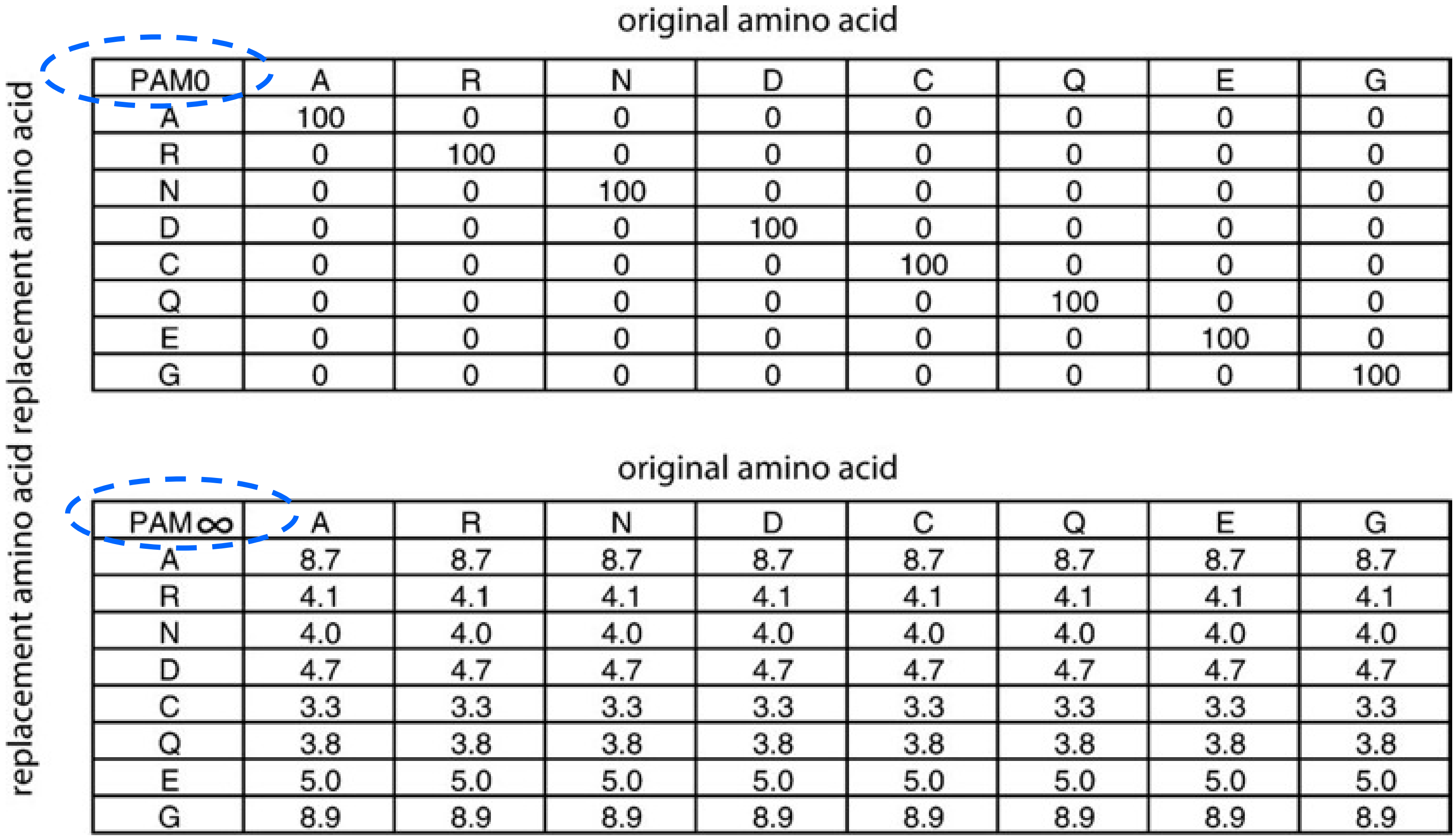
Please visit, https://en.wikipedia.org/wiki/Point_accepted_mutation for a nice introduction

Entries of off-diagonal mutation probability matrix

f(j) M(i,j) = f(i) M(j,i) = (λ/N) A(j,i) = (λ/N) A(i,j)

Other PAM matrices

- What is PAM1 depend on
  - Sequence alignments that are closer
  - Also depend on the sequences that are considered
- Let us consider the extreme case of PAM0
  - Only diagonal



original amino acid

replacement amino acid

PAM0	A	R	N	D	C	Q	E	G
A	100	0	0	0	0	0	0	0
R	0	100	0	0	0	0	0	0
N	0	0	100	0	0	0	0	0
D	0	0	0	100	0	0	0	0
C	0	0	0	0	100	0	0	0
Q	0	0	0	0	0	100	0	0
E	0	0	0	0	0	0	100	0
G	0	0	0	0	0	0	0	100

original amino acid

replacement amino acid

PAM∞	A	R	N	D	C	Q	E	G
A	8.7	8.7	8.7	8.7	8.7	8.7	8.7	8.7
R	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1
N	4.0	4.0	4.0	4.0	4.0	4.0	4.0	4.0
D	4.7	4.7	4.7	4.7	4.7	4.7	4.7	4.7
C	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3
Q	3.8	3.8	3.8	3.8	3.8	3.8	3.8	3.8
E	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
G	8.9	8.9	8.9	8.9	8.9	8.9	8.9	8.9

Fig 3.12 from the Pevsner Book III Edition
PLEASE DO NOT DISTRIBUTE-Copyright figure

So far, we have a probability matrix, but we want to score alignments (ie how is this different from random alignments)

To get a scoring matrix, we need to convert the probability matrix into odds matrix

Rij is the Relatedness Odds Matrix

Final steps (6 and 7)

M_{ij} is the probability that the original aa (j) will be substituted by (i)

$$R_{ij} = \frac{M(i,j)f(j)}{f(i)f(j)} = \frac{M(i,j)}{f(i)}$$

R_{ij} is the relatedness odds ratio

Table 3.1 provides Normalized freq. (f_i)s

Think back on conditional prob.; note fjs are normalized frequency

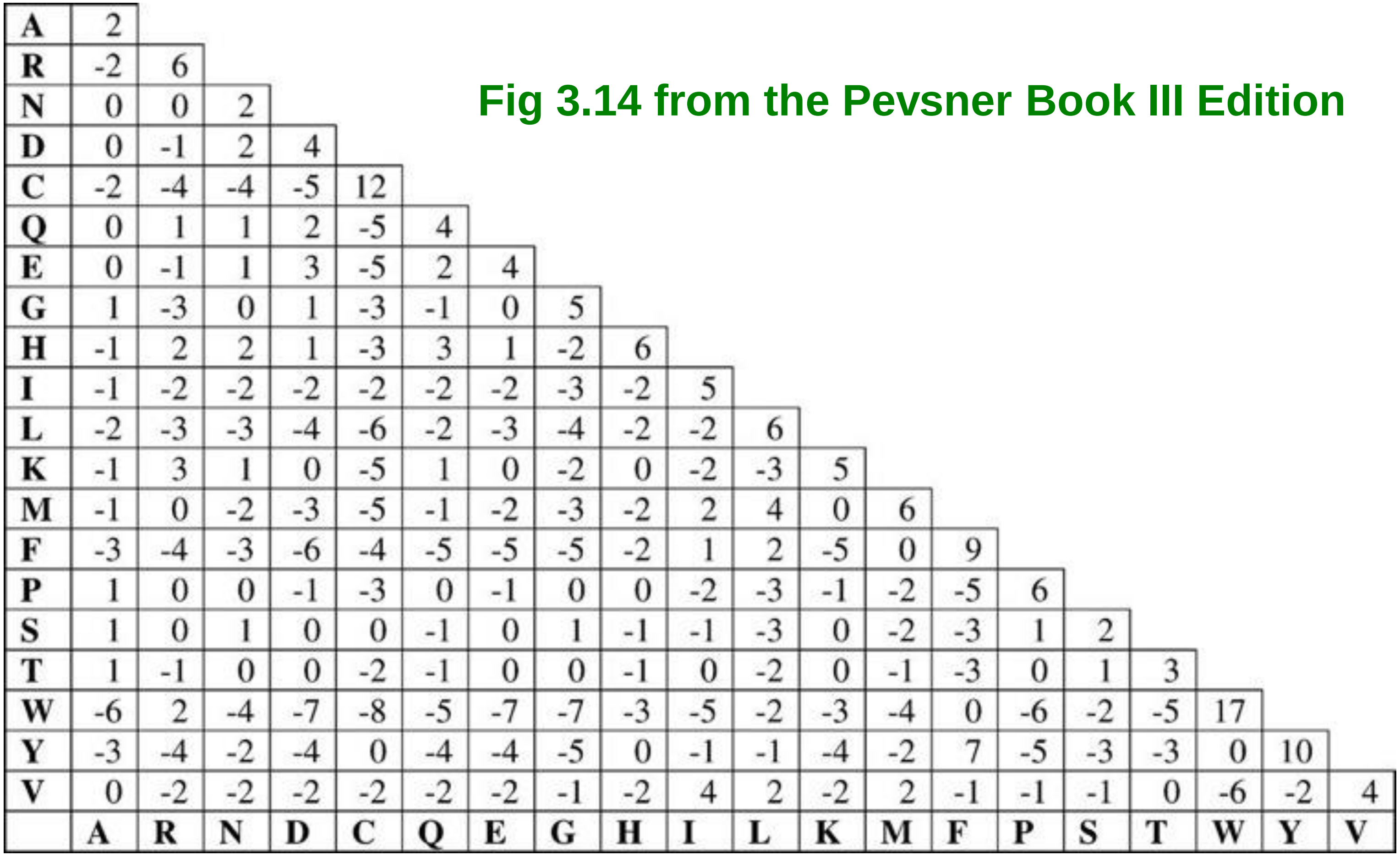
$$\text{Probability of an authentic alignment} = \frac{\text{P(aligned | authentic)}}{\text{P(aligned | random)}}$$

Log odds score

$$S_{ij} = 10 \times log_{10}[\frac{M(i,j)}{f(i)}]$$

Unlike M_{ij}, S_{ij} are symmetric.

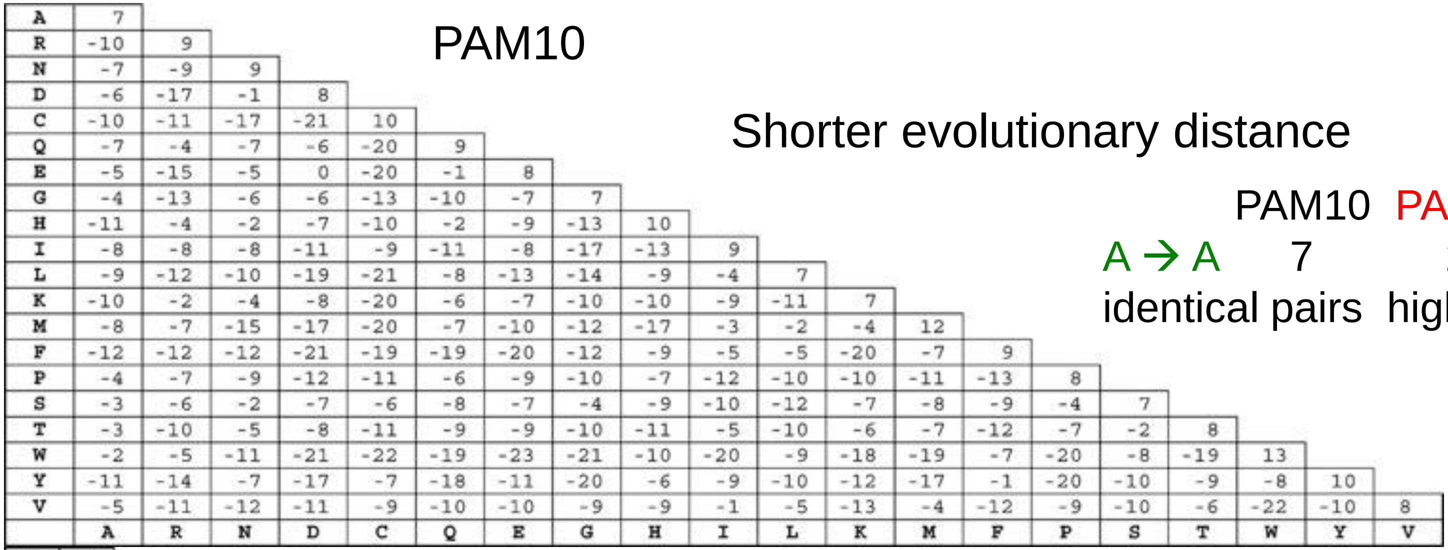
PAM250



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

Fig 3.14 from the Pevsner Book III Edition

PAM10



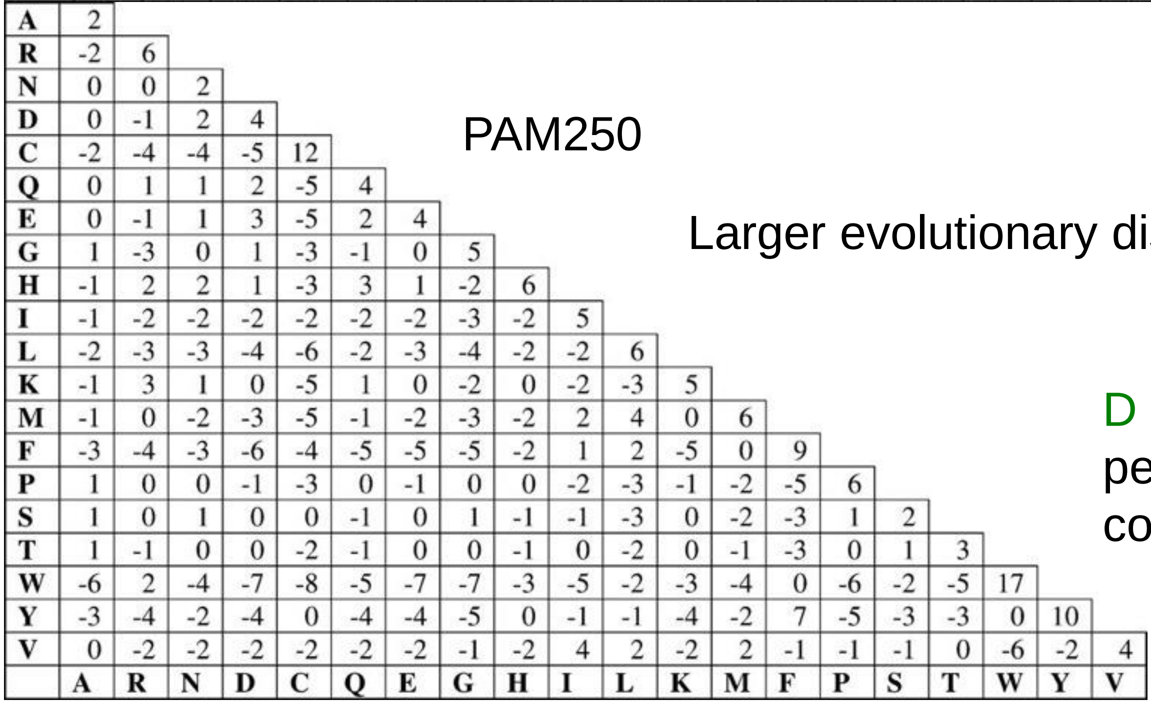
A	7																			
R	-10	9																		
N	-7	-9	9																	
D	-6	-17	-1	8																
C	-10	-11	-17	-21	10															
Q	-7	-4	-7	-6	-20	9														
E	-5	-15	-5	0	-20	-1	8													
G	-4	-13	-6	-6	-13	-10	-7	7												
H	-11	-4	-2	-7	-10	-2	-9	-13	10											
I	-8	-8	-8	-11	-9	-11	-8	-17	-13	9										
L	-9	-12	-10	-19	-21	-8	-13	-14	-9	-4	7									
K	-10	-2	-4	-8	-20	-6	-7	-10	-10	-9	-11	7								
M	-8	-7	-15	-17	-20	-7	-10	-12	-17	-3	-2	-4	12							
F	-12	-12	-12	-21	-19	-19	-20	-12	-9	-5	-5	-20	-7	9						
P	-4	-7	-9	-12	-11	-6	-9	-10	-7	-12	-10	-10	-11	-13	8					
S	-3	-6	-2	-7	-6	-8	-7	-4	-9	-10	-12	-7	-8	-9	-4	7				
T	-3	-10	-5	-8	-11	-9	-9	-10	-11	-5	-10	-6	-7	-12	-7	-2	8			
W	-2	-5	-11	-21	-22	-19	-23	-21	-10	-20	-9	-18	-19	-7	-20	-8	-19	13		
Y	-11	-14	-7	-17	-7	-18	-11	-20	-6	-9	-10	-12	-17	-1	-20	-10	-9	-8	10	
V	-5	-11	-12	-11	-9	-10	-10	-9	-9	-1	-5	-13	-4	-12	-9	-10	-6	-22	-10	8
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

Shorter evolutionary distance

	PAM10	PAM250
A → A	7	2

identical pairs high score in 10 (close)

PAM250



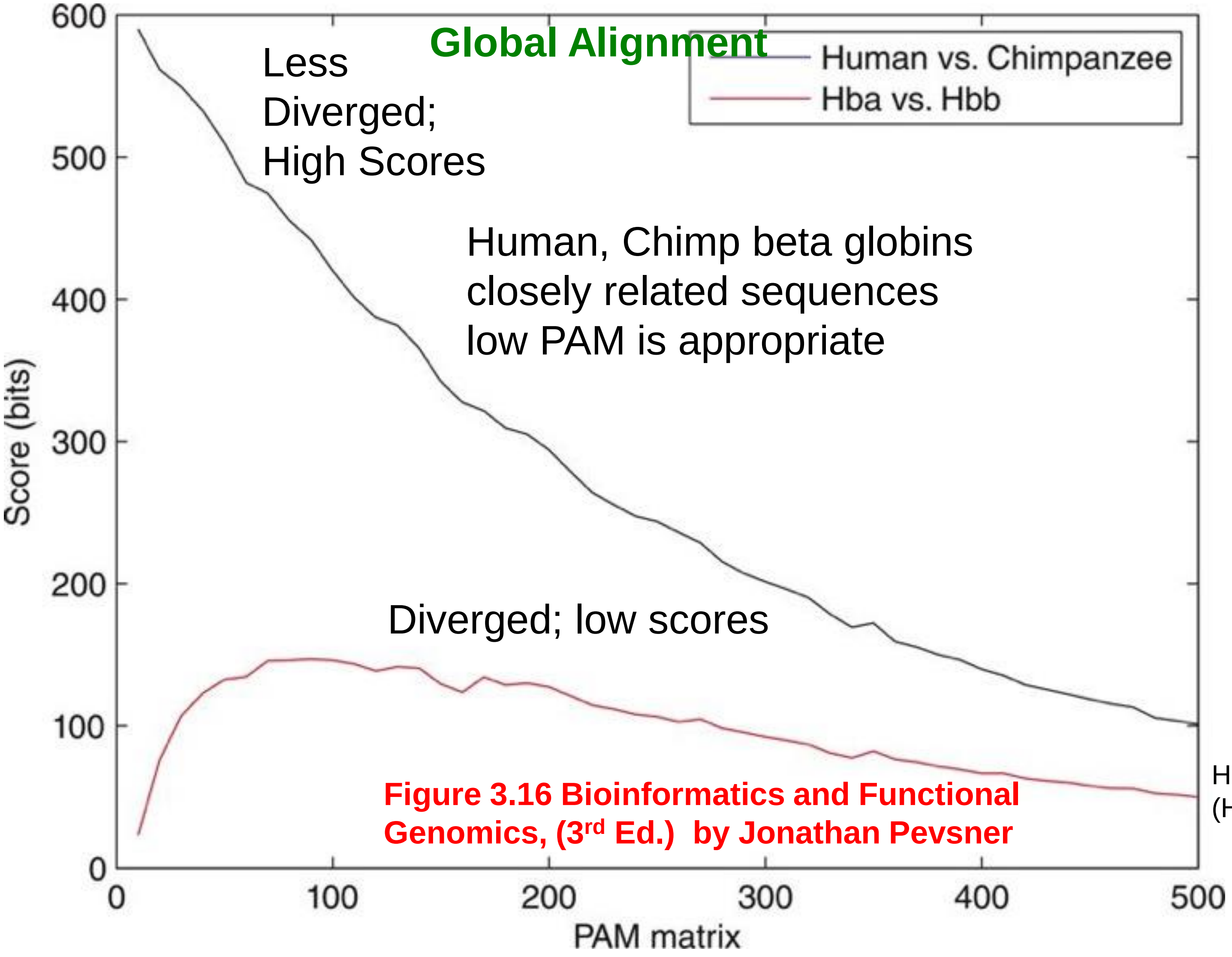
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	12															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L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

Larger evolutionary distance

	PAM10	PAM250	
D → R	-17	-1	mismatch

penalty is higher in PAM10 compared to PAM250

Appropriate PAM matrix?



Global Alignment

Less
Diverged;
High Scores

Human, Chimp beta globins
closely related sequences
low PAM is appropriate

Diverged; low scores

Figure 3.16 Bioinformatics and Functional Genomics, (3rd Ed.) by Jonathan Pevsner

HBA, HBB diverged sequences
(HIGH PAM is appropriate)

BLOcks of amino acid SUbstitution Matrices

More on this in BLAST class

- BLOSUM
  - Based on BLOCKS database
  - Henikoff and Henikoff, Karlin and Altschul, Others
  - Considered local MSA of distantly related proteins
  - Scoring scheme similar to PAM
    - Log-odds ratio using base 2 log

$$S_{ij} = 2 * \log_2 \left[\frac{M_{ij}}{f_i} \right]$$

$$s(a,b) = \frac{1}{\lambda} \log\left(\frac{p_{ab}}{f_a f_b} \right)$$

General form for scoring matrices according to Dr. Altschul

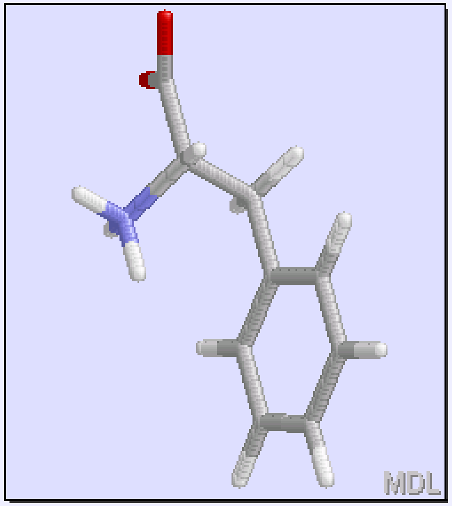
BLOSUM Summary

- BLOSUM62 is “standard” (better performances than PAM)
- Nature Biotechnology: http://www.nature.com/nbt/journal/v22/n8/abs/nbt0804-1035.html

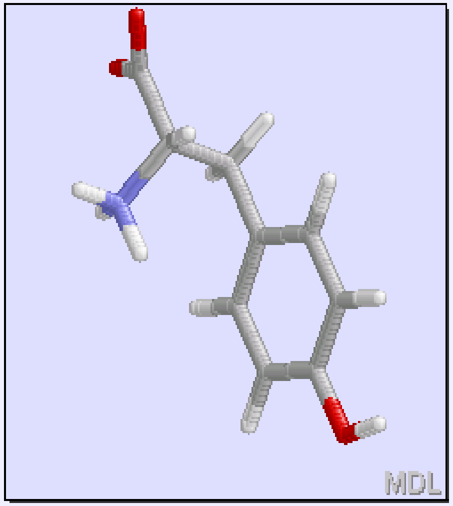
Example of BLOSUM62

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

L-phenylalanine (F)



L-tyrosine (Y)



Comm … ow weights

Rar … igh weights

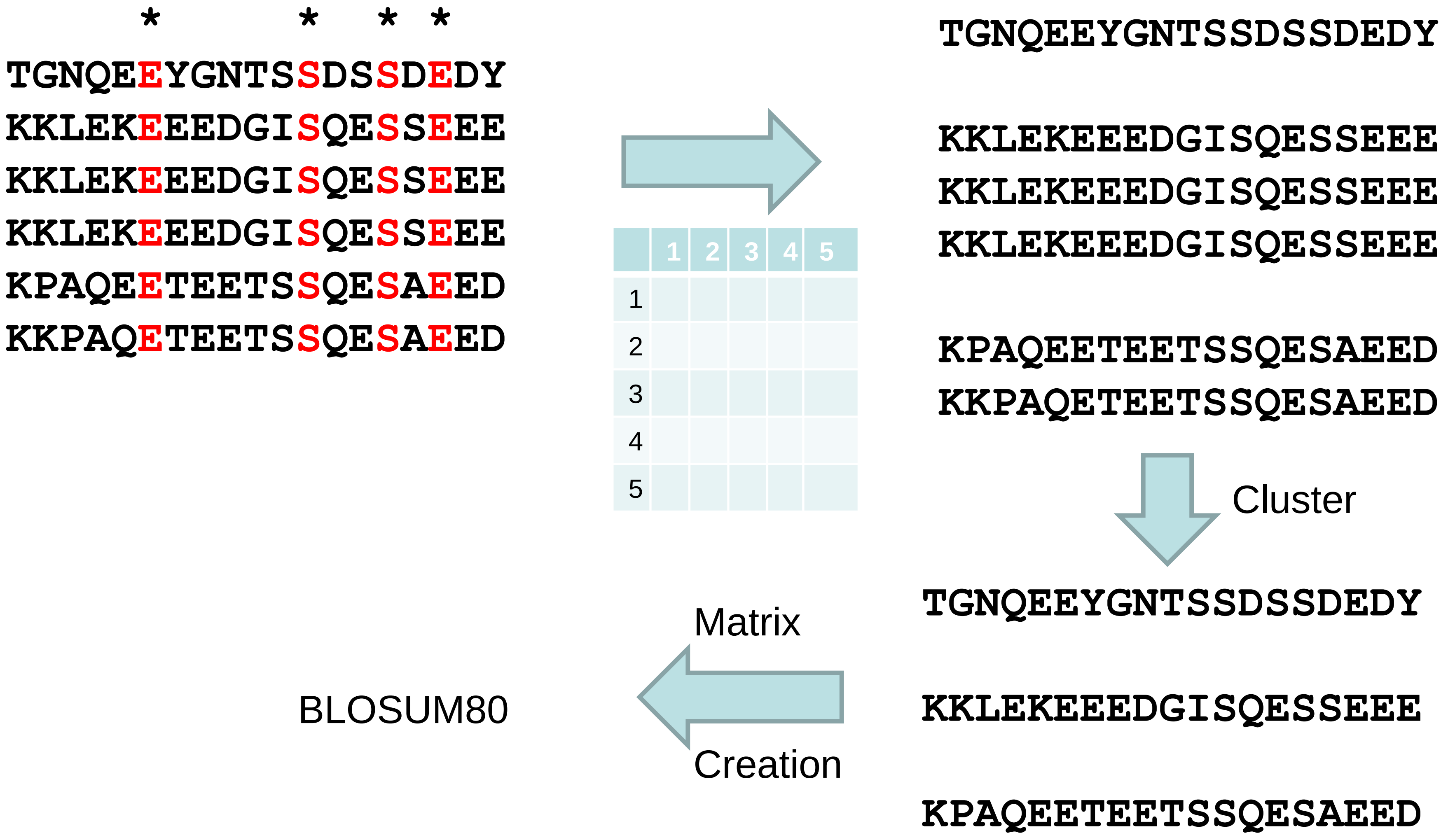
Negative for less likely substitutions

Positive for more likely substitutions

BLOSUM80

- Using sequences that share no more than 80% identity
- Sequences that are more than 80% identity are clustered and represented by a single sequence
- Why Clustering?
  - Reduces overrepresentation and bias
- BLOSUMn
  - Lower “n”s will help us identify more distantly related sequences
  - Higher “n”s will help us identify less diverged sequences

BLOSUM80



```
     *     *  * *
TGNQEEYGNTSSDSSDEDY
KKLEKEEEDGISQESSEEE
KKLEKEEEDGISQESSEEE
KKLEKEEEDGISQESSEEE
KPAQEETEETSSQESAEED
KKPAQETEETSSQESAEED
```

	1	2	3	4	5
1					
2					
3					
4					
5					

```
TGNQEEYGNTSSDSSDEDY

KKLEKEEEDGISQESSEEE
KKLEKEEEDGISQESSEEE
KKLEKEEEDGISQESSEEE

KPAQEETEETSSQESAEED
KKPAQETEETSSQESAEED
```

Cluster

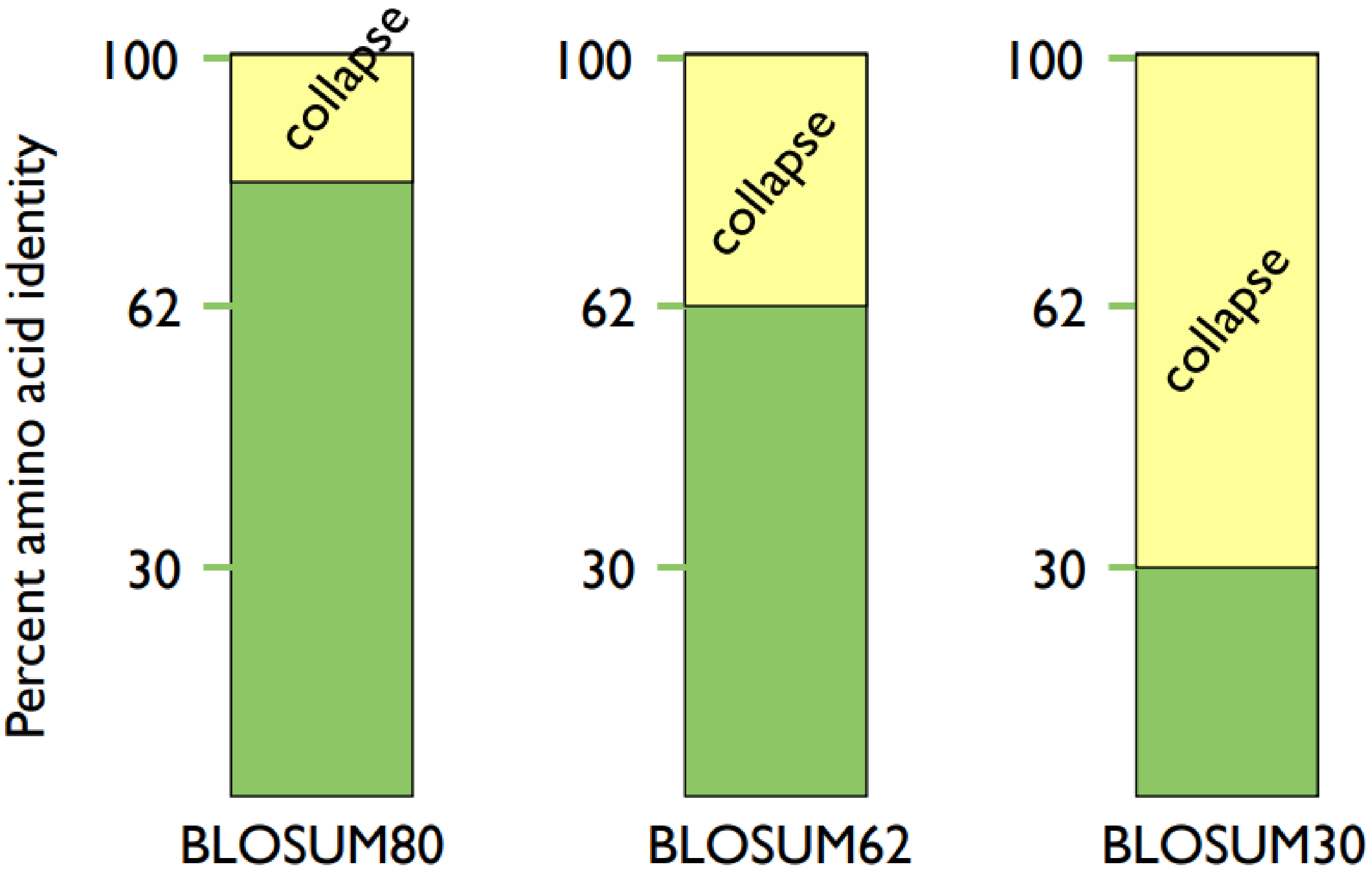
```
TGNQEEYGNTSSDSSDEDY

KKLEKEEEDGISQESSEEE

KPAQEETEETSSQESAEED
```

Matrix Creation

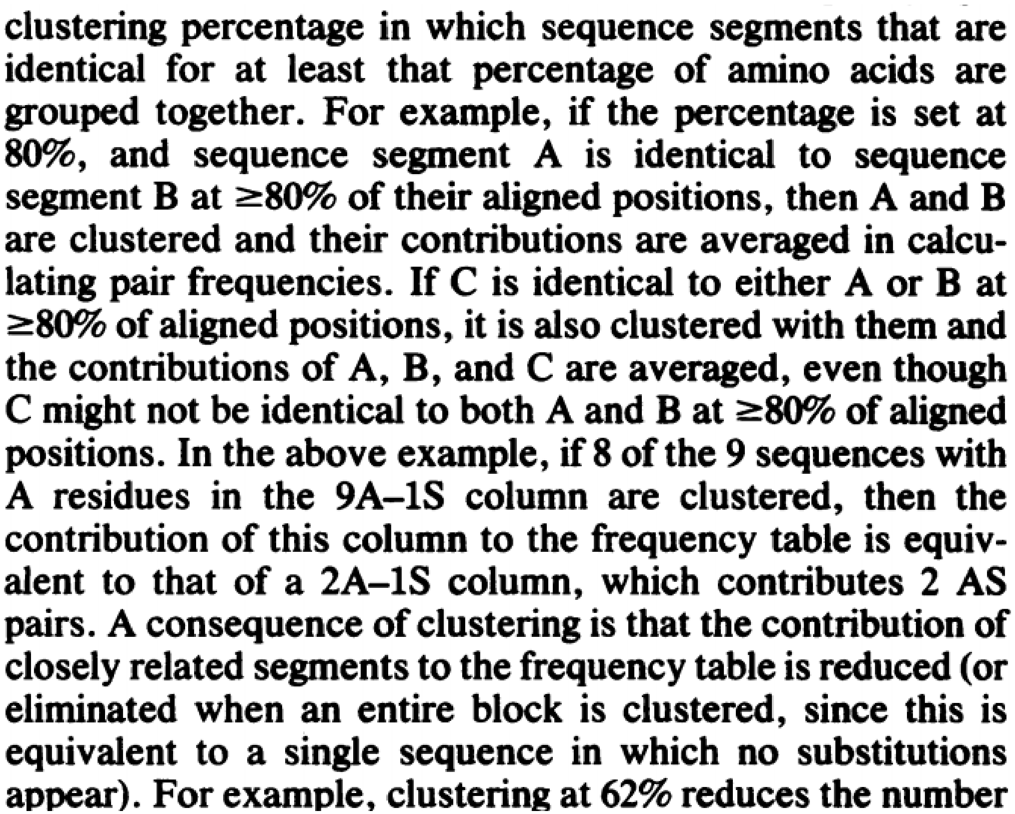
BLOSUM80

Figure from
Suprakash Datta Class
Slides

BLOSUM Matrices



Proc. Natl. Acad. Sci. USA Vol. 89, pp. 10915-10919, November 1992 Biochemistry

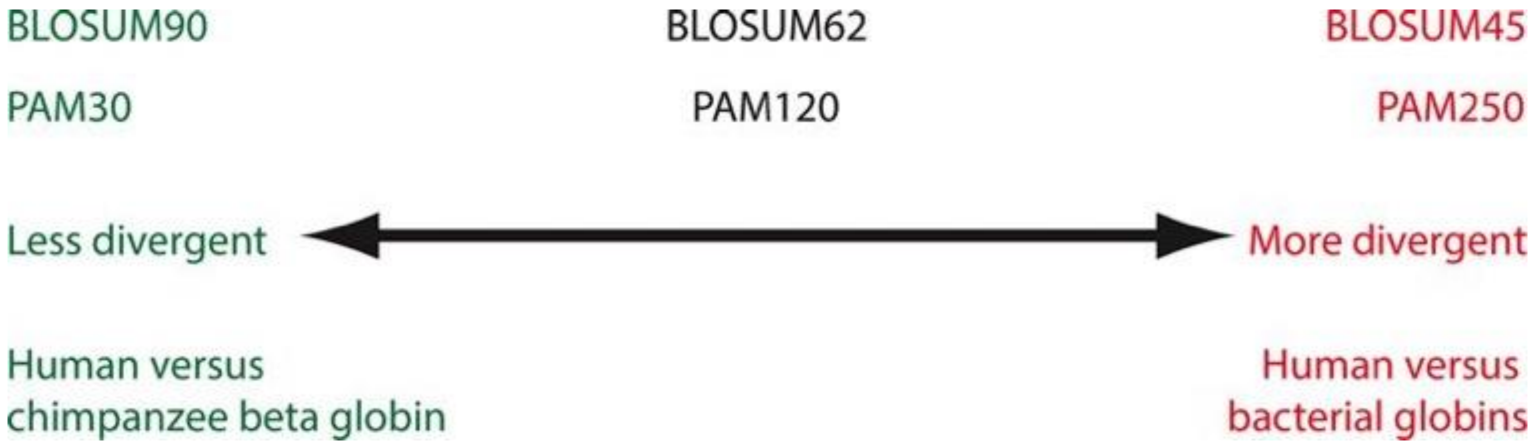


clustering percentage in which sequence segments that are identical for at least that percentage of amino acids are grouped together. For example, if the percentage is set at 80%, and sequence segment A is identical to sequence segment B at $\geq$80% of their aligned positions, then A and B are clustered and their contributions are averaged in calculating pair frequencies. If C is identical to either A or B at $\geq$80% of aligned positions, it is also clustered with them and the contributions of A, B, and C are averaged, even though C might not be identical to both A and B at $\geq$80% of aligned positions. In the above example, if 8 of the 9 sequences with A residues in the 9A–1S column are clustered, then the contribution of this column to the frequency table is equivalent to that of a 2A–1S column, which contributes 2 AS pairs. A consequence of clustering is that the contribution of closely related segments to the frequency table is reduced (or eliminated when an entire block is clustered, since this is equivalent to a single sequence in which no substitutions appear). For example, clustering at 62% reduces the number

Paper available from Class Reference folder

Matrix Comparison

Figure 3.17 Bioinformatics and Functional Genomics, (3rd Ed.) by Jonathan Pevsner
PLEASE DO NOT DISTRIBUTE-Copyright figure



Short alignments;
Highly similar

Longer weaker local alignments

PAM & BLOSUM Matrices

- PAM (Dayhoff et al 1988)
  - PAM1 is the matrix obtained by comparing sequences differ by no more than 1%
  - Higher PAMX are extrapolated from PAM1)
  - PAM250: Observed difference(80%) Evolutionary distance (250)
- Limitation: Matrices are derived from alignments of sequence that are 85% identity
  - Difficult to use in Twilight Zone

$PAM_2 = PAM_1 * PAM_1 = (PAM_1)^2$
$PAM_{250} = (PAM_1)^{250}$

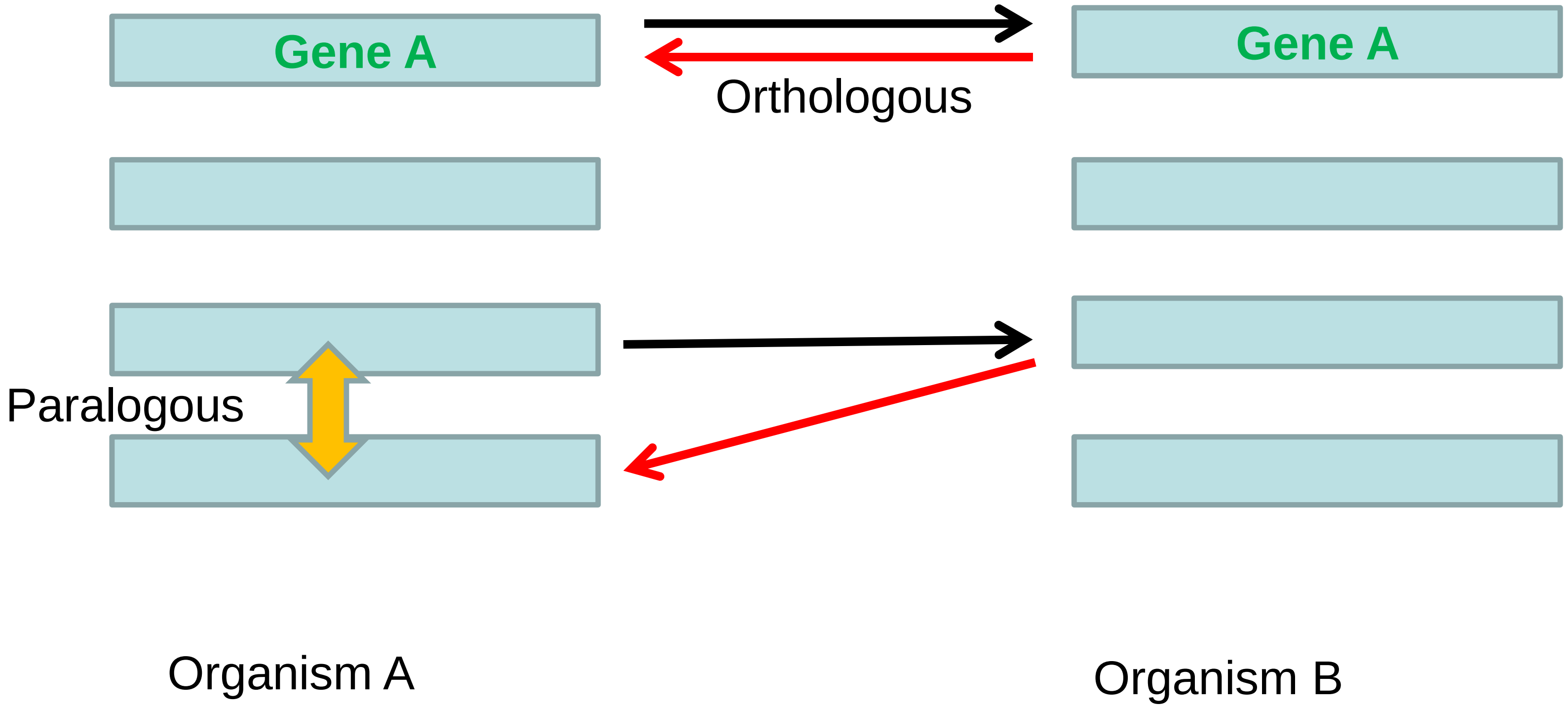
- Blosum (BLOcks SUbstitution Matrix) Heinkoff & Heinkoff (1992)
  - Derived from BLOCKS database
  - Distant relationships explained better than PAM
  - Blosum62 obtained from sequence BLOCKS clustered at >=62% identity
  - BlosumX are observed from actual alignments not extrapolated
  - Sequences are very similar use higher Blosum (low PAM)

Matrix Multiplication

We introduced the concept of homology, here let us learn how to identify the homologous sequences

Reciprocal Best Hits Concept to Deduce Homology

BLAST or any other alignment software



Questions for pondering

- PAM40
  - Highly related proteins
- BLOSUM80
  - Not ideal for scoring highly related sequences
  - Why?
  - Matrix is built based on sequences that share upto 80% identity

What matrix for what?

BLOSUMx	Good for	% Similarity
x = 90	Short and highly similar sequences	70-90
x = 80	Often good for identifying family members	50-60
x = 62	Most effective for a variety of range of similar sequences; default in NCBI BLAST	30-40
x = 30	For diverged; weak long alignments	<30

Based on Dr. Andy Baxevanis lectures

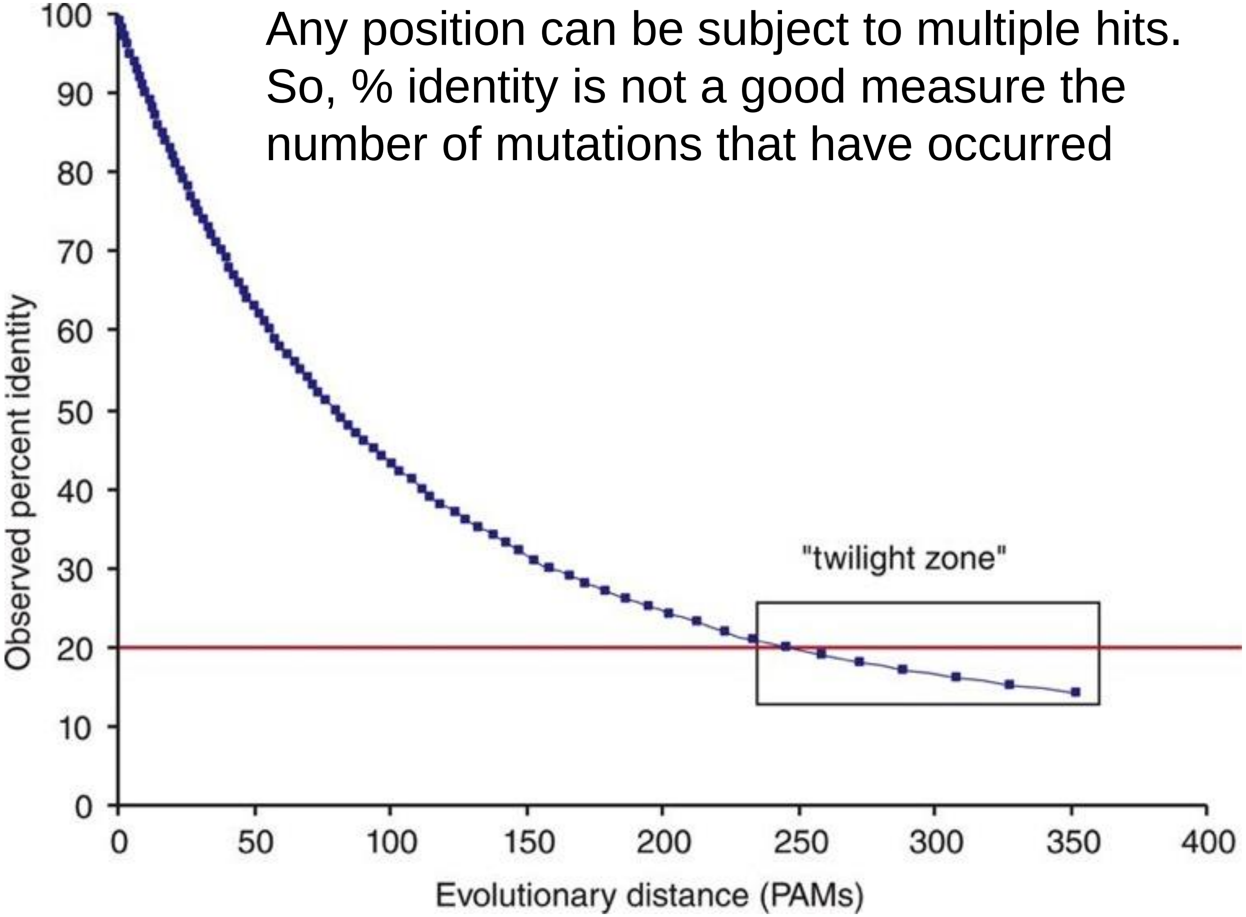
Divergence and Twilight zone

Take two sequences (100 aa in length), fix one and introduce mutations into the other.

Plot % identity vs PAM

PAM250
250 hits/100 aligned aa

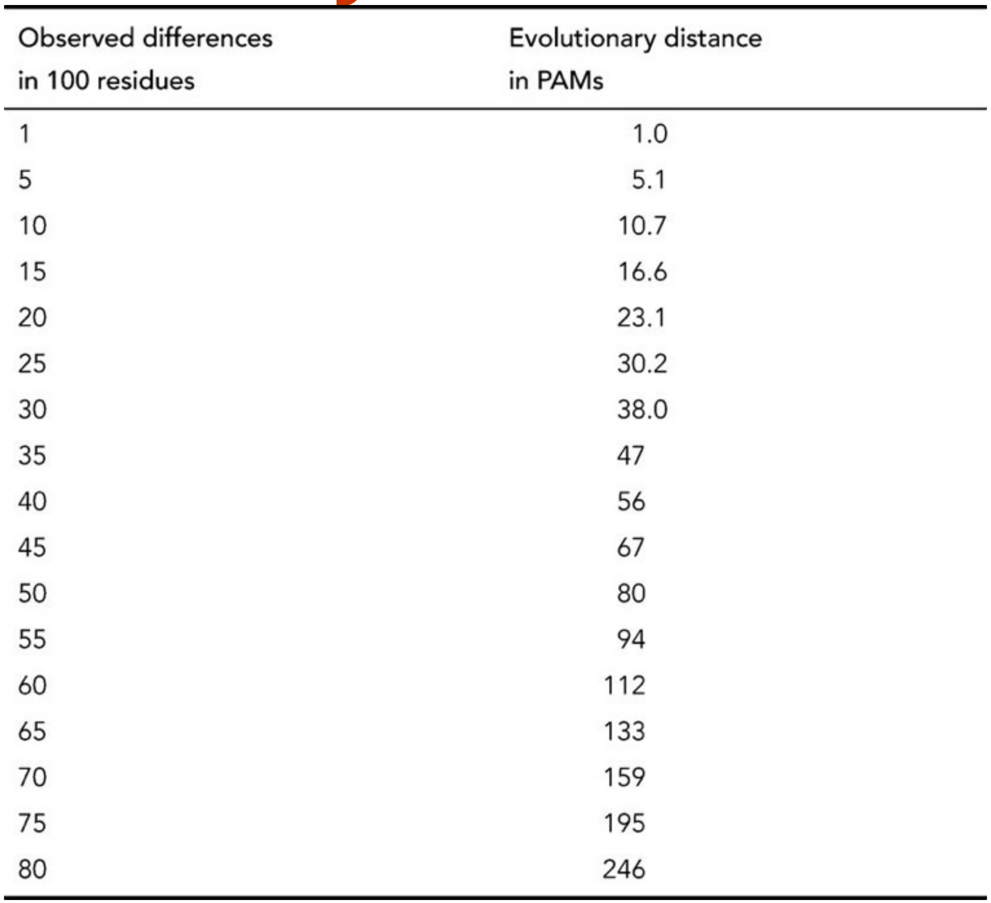
Hit: a change in aa that occurs by mutation



Any position can be subject to multiple hits. So, % identity is not a good measure the number of mutations that have occurred

Figure 3.19 Bioinformatics and Functional Genomics, (3rd Ed.) by Jonathan Pevsner PLEASE DO NOT DISTRIBUTE-Copyright figure

Observed differences and evolutionary distance



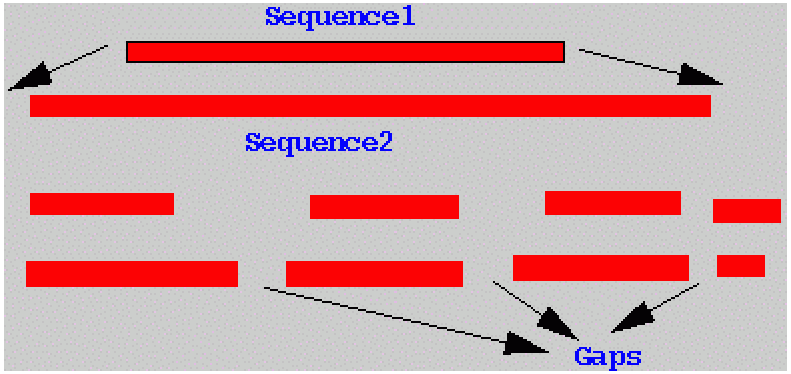
Observed differences in 100 residues	Evolutionary distance in PAMs
1	1.0
5	5.1
10	10.7
15	16.6
20	23.1
25	30.2
30	38.0
35	47
40	56
45	67
50	80
55	94
60	112
65	133
70	159
75	195
80	246

Table 3.3 from Bioinformatics and Functional Genomics, (3rd Ed.) by Jonathan Pevsner
PLEASE DO NOT DISTRIBUTE-Copyright figure

Having solved the scoring, let us tackle how to do the alignment?

Types of Alignments

- Global Alignment
  - – Sequence alignment over the whole range
- Local Alignment
  - – Identifying local regions (islands) by introducing gaps



Types of pair-wise alignments

- Brute Force
  - Creating all possible subsets to identify best alignment
  - Seq A: Length M
  - Seq B: Length N
  - roughly $2^{(M+N)}$ total comparisons
  - Not an ideal method

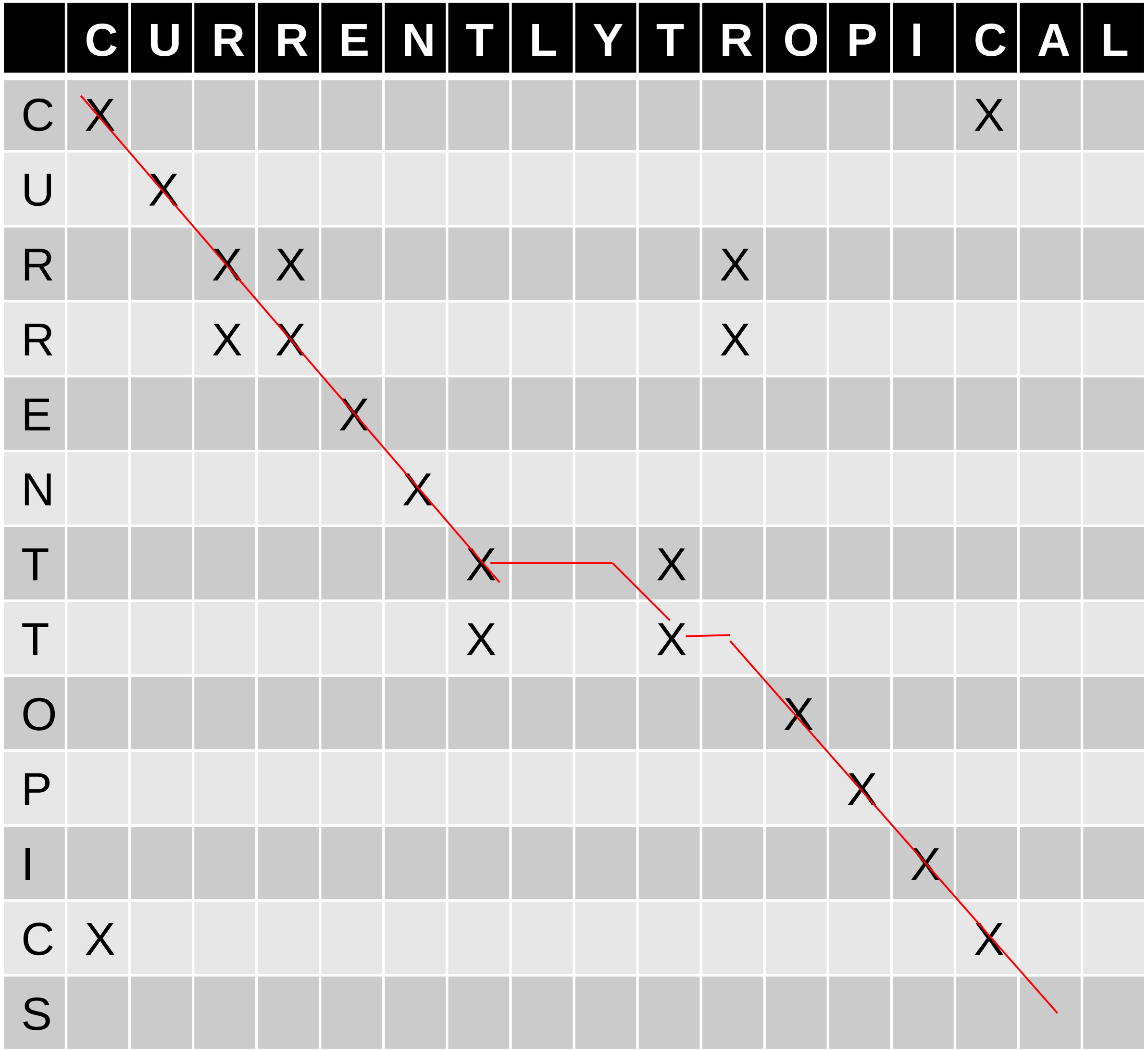
Types of pair-wise alignments

- Dot-matrix
  - Identifying all possible matches between two sequences
  - Sequence#1: CURRENTLYTROPICAL
  - Sequence#2: CURRENTTOPICS

http://www.srmuniv.ac.in/sites/default/files/files/5(6).pdf

Types of pair-wise alignments

- Dot-matrix



	C	U	R	R	E	N	T	L	Y	T	R	O	P	I	C	A	L
C	X														X		
U		X															
R			X	X							X						
R			X	X							X						
E					X												
N						X											
T							X			X							
T							X			X							
O												X					
P													X				
I														X			
C	X														X		
S																	

Connect dots across the diagonal using either horizontal (x) or vertical lines (y)

CURRENTLYTROPICAL
CURRENT—-T-OPICS

Types of pair-wise alignments

- Dot-matrix:
  - Pros: Easy to understand and useful to identify repeats, palindrome etc
  - Cons: Time consuming if more than one pairwise alignment have to be carried out
- Dynamic Programming
  - Compares each character in such a way to maximize the number of matches (identical or similar)

Types of pair-wise alignments

- Dynamic Programming (DP)
  - Global Dynamic Programming
  - Local Dynamic Programming
- DP Algorithms
  - Needleman and Wunsch (Global)
  - Smith Waterman (Local)

Global DP

- Generates an alignment for 2 sequences that maximizes the matches and minimizes the # of gaps
- End-to-end alignment
- Linear gap penalty Substitution/Mismatch
  - Ex just one value, -5
- Best when sequences are similar

MM: MisMatch
Sub: Substitution

Dynamic Programming

- No Gap penalty
  - Scoring System: Match +8; sub/MM = -5

```
CGGGGGGAACT
CGGGGGGATCT
```

10*8 -5 = 75

- Linear gap penalty
  - M: +8; Sub/MM: -5; Gap = -3

```
C---GGGAACT
CGGGGGGATCT
```

- Affine Gap penalty:
  - M=+8; Sub/MM = -5; Gap Open = -3, Gap Ext = -1

Deduction for a gap = G + Ln (Note G > L)
G: Gap Opening Penalty
L: gap-extension penalty
n: Length of a gap

```
C---GGGAACT
CGGGGGGATCT
```

Global Alignment Reference

Needleman, S.B. and Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 48(3):443-53(1970).

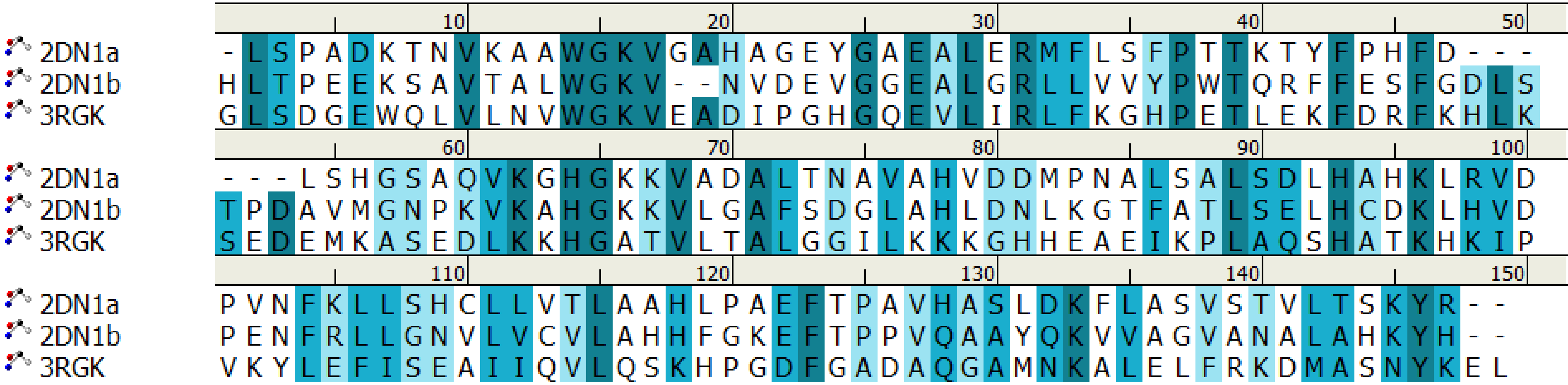
Local DP

Smith, T.F. and Waterman, M.S.
Identification of common molecular subsequences. J Mol Biol. 147(1):195-7 (1981)

In many cases, we are inherently looking for local alignments.
This is true given the fact that most proteins are modular.
BLAST default matrix is BLOSUM62

Sequence comparison of Homologous sequences

Globin family; 2DN1a/2DN1b hemoglobin-alpha/beta; 3RGK is Myoglobin



```
                         10                  20                  30                  40                  50
2DN1a  - L S P A D K T N V K A A W G K V G A H A G E Y G A E A L E R M F L S F P T T K T Y F P H F D - - -
2DN1b  H L T P E E K S A V T A L W G K V - - N V D E V G G E A L G R L L V V Y P W T Q R F F E S F G D L S
3RGK   G L S D G E W Q L V L N V W G K V E A D I P G H G Q E V L I R L F K G H P E T L E K F D R F K H L K
                         60                  70                  80                  90                 100
2DN1a  - - - L S H G S A Q V K G H G K K V A D A L T N A V A H V D D M P N A L S A L S D L H A H K L R V D
2DN1b  T P D A V M G N P K V K A H G K K V L G A F S D G L A H L D N L K G T F A T L S E L H C D K L H V D
3RGK   S E D E M K A S E D L K K H G A T V L T A L G G I L K K K G H H E A E I K P L A Q S H A T K H K I P
                        110                 120                 130                 140                 150
2DN1a  P V N F K L L S H C L L V T L A A H L P A E F T P A V H A S L D K F L A S V S T V L T S K Y R - -
2DN1b  P E N F R L L G N V L V C V L A H H F G K E F T P P V Q A A Y Q K V V A G V A N A L A H K Y H - -
3RGK   V K Y L E F I S E A I I Q V L Q S K H P G D F G A D A Q G A M N K A L E L F R K D M A S N Y K E L
```

Sequence similarity 37.6% ; Identity: 17.4%

Local DP

- Alignment that maximizes regions of similarity
- Not necessarily end-to-end
- Uses affine gap penalty
  - https://en.Wikipedia.org/wiki/Gap_penalty
- Often uses, a one time gap penalty for each stretch of gaps plus a gap extension penalty as a function of length of gap

```
Query  57   FTALCQKLKIPDHVRERAWLTWEKVSSVDGVLGGYIQKKKELWGICIFIAAV--------  108
            F  LC +L + +  R  AW ++  +S    + G  +      W  C    A
Sbjct  49   FDELCSRLNMDEAARAEAWDSYRSMSESYTLEGNDLH-----WLACALYVACRKSVPTVS  103

Query  109  --DLDEMSFTFTELQKNIEISVHKFFNLLKEIDTSTKVD----NAMSRLLKKYDVLFALF  162
               ++    + T + K  E S+ +FFN +K+ +    +         RL + + V   +F
Sbjct  104  KGTVEGNYVSLTRILKCSEQSLIEFFNKMKKWEDMANLPPHFRERTERLERNFTVSAVIF  163

Query  163  SKLERTCELIY---------LTQPSSSISTEINSALVLKVSWITFLLAKGEVLQMEDDLV  213
             K E   + I+           +           + +    W+ F+ AKG    + DDLV
Sbjct  164  KKYEPIFQDIFKYPQEEQPRQQRGRKQRRQPCTVSEIFHFCWVLFIYAKGNFPMISDDLV  223

Query  214  ISFQLMLCVLDYFIKLSPPMLLKEPYKTAVIPINGSPRTPRRGQNRSARIAKQLENDTRI  273
             S+ L+LC LD     +    L+   +  ++  N    +    ++  A+ +K   +   I
Sbjct  224  NSYHLLLCALDLVYGNA----LQCSNRKELVNCNFKGLS----EDFHAKDSKPSSDPPCI  275

Query  274  IEVLCKEHECNIDEVKNVYFKNFIPFMNSL--------------GLVTSNGLPE-VENLS  318
            IE LC  H+  + E K +    + P++  L              G +      E  + ++
Sbjct  276  IEKLCSLHDGLVLEAKGIKEHFWKPYIRKLYEKKLLKGKEENLTGFLEPGNFGESFKAIN  335

Query  319  KRYEEIYLKNKDLDARLFLDHDKTLQTDSID-----------------------SFETQR  355
            K YEE  L   +LD R+FL  D   +  ++                         F+  +
Sbjct  336  KAYEEYVLSVGNLDERIFLGEDAEEEIGTLSRCLNAGSGTETAERVQMKNILQQHFDKSK  395

Query  356  TPRKSNLDEEVNVI----PPHTPVRTVMNTIQQLMMILNSASDQPSENLISYFNNCTVNP  411
              R S     V  I    P  TPV T  +++ +L  +L    + PSE L      C+ +P
Sbjct  396  ALRISTPLTGVRYIKENSPCVTPVSTATHSLSRLHTMLTGLRNAPSEKLEQILRTCSRDP  455

Query  412  KESILKRVKDIGYIFKEKFA--KAVGQGCVEIGSQRYKLGVRLYYRVMESMLKSEEERLS  469
             ++I  R+K++  I+ + F   +       EI S+ ++    LYY+V+ES+++ E++RL
Sbjct  456  TQAIANRLKEMFEIYSQHFQPDEDFSNCAKEIASKHFRFAEMLYYKVLESVIEQEQKRLG  515

Query  470  IQNFSKLLNDNIFHMSLLACALEVVMATYSRSTSQNLDSGTDLSFPWILNVLNLKAFDFY  529
              + S +L  + FH SLLAC LEVV  +Y              +FP+I  +  +  + FY
Sbjct  516  DMDLSGILEQDAFHRSLLACCLEVVTFSYKPPG----------NFPFITEIFEVPLYHFY  565

Query  530  KVIESFIKAEGNLTREMIKHLERCEHRIMESLAWLSDSPLFDLIKQSKDREGPTDHLESA  589
            KVIE FI+AE  L RE++KHL + E +I++ LAW  +SPL++ I+ +++R  PT   E
Sbjct  566  KVIEVFIRAEDGLCREVVKHLNQIEEQILDHLAWKPESPLWEKIRDNENRV-PTCE-EVM  623

Query  590  CPLNLPLQNNHTAADMYLSPVRSPKKKGSTTRVNSTANAETQATSAFQTQKPLKSTSLSL  649
             P NL   +    A   L+P R  + +  T  +  +  + T     + +  P  +T   L
Sbjct  624  PPQNLERADEICIAGSPLTPRRVTEVRADTGGLGRSITSPTTLYDRY-SSPPASTTRRRL  682

Query  650  F  650
            F
Sbjct  683  F  683
```

RB1 Blast search with RefSeq and excluding XM_ and Environmental Seqs Default Parameters

Query: RB1_human
Sbjct: Retinoblastoma-related protein 2 [Macaca mulatta]

```
Score            Expect    Method
180 bits(457)    3e-43     Compositional
                           matrix adjust

.
Identities       Positives        Gaps
163/661(25%)     281/661(42%)     93/661(14%)
```

Scoring matrix that gives a score for aligning two characters (total score taking into account INDELs)

How to generate the alignments?
Algorithm??

Dynamic Programming (DP)

- Richard Bellman, 1950s by mathematician
  - RAND Corp. on optimal decision processes
    - Funding for the project came from Navy/Military

- A name that he choose to hide his project from the then US Secretary of Defense Charles Wilson,
  - a man not friendly to either basic or mathematics research.

Dynamic Algorithm

- Goal: To find an optimal alignment that maximizes the score of two sequences
- Can’t we manually align them?
- The possible alignments are
- different alignments for two sequences of length N
  - – N=100; 10^{58} Sequence alignments! .

$$\frac{2^{2N}}{\sqrt{2\pi N}}$$

Using Global Alignment as an example

Dynamic Programming

- Notation
- Two sequences *x* and *y*.
  - – Length M and N respectively
- x_i is the i th residue in x
- y_j is the j th residue in y
- Scoring Matrix
  - – Linear Gap penalty $\gamma = -6$
    - Penalty increases with # of gaps

	A	G	C	T
A	+5	−2	−2	−2
G	−2	+5	−2	−2
C	−2	−2	+5	−2
T	−2	−2	−2	+5

What is an optimal alignment?

Seq. x of length M
Seq. y of length N

- Recursive definition of alignment
- How can an alignment end?
- 3 possibilities

x_M
y_N

x_M
y_N--

x_M--
y_N

x_M
One residue

x_M
y_N-

x_M is aligned to a gap and Y_N had already appeared earlier in the sequence alignment

- The optimal alignment is the one with the highest score from the above 3 cases

What is an optimal alignment? How can an alignment end?

Seq. x of length M = 18
Seq. y of length N = 8

Three possibilities

```
CAGCACTTGGATTCTCGG
CAGC-----G-T----GG
```

x_M
y_N

```
CAGCA-CTTGGATTCTCGG
---CAGCGTGG--------
```

x_M
y_N -

x_M
y_N -

x_M is aligned to a gap and Y_N had already appeared earlier in the sequence alignment

```
CAGCA-CTTGGATTCTCGG-
---CAGCGTG---------G
```

x_M -
y_N

Dynamic Programming

- Let us look at the scoring schemes for the previous three cases
- Bigger alignments are made up of optimal sub-alignments
- You can do this recursively
- S(i,j) is the alignment score of the sequence prefix, $\mathbf{x_1 \ldots x_i}$ with $\mathbf{y_1 \ldots y_j}$

Dynamic Programming

- $S(M,N) = \mathbf{S(x_{M,}y_{N})} + \mathbf{S(M\text{-}1, N\text{-}1)}$

x_M
y_N

$1..x_{M-1}$
$1..y_{N-1}$

- $S(M\text{-}1,N) = \gamma + S(M\text{-}1,N)$

γ Gap penalty

- $S(N\text{-}1,M) = \gamma + S(N\text{-}1,M)$

Dynamic Programming

- To calculate
- S(M,N)
    - ❑ S(M-1,N-1), S(M,N-1), S(M-1,N)
- In turn for S(M-1,N-1), we need
    - ❑ S(M-2,N-2),S(M-1,N-2),S(M-2,N-1)
- S(M,N-1)
    - ❑ S(M-1,N-2), S(M,N-2),S(M-1,N-1)
- S(M-1,N)
    - ❑ S(M-2,N-1),S(M-1,N-1),S(M-2,N)

We keep going back to building smaller and smaller pieces until we reach S(0,0)

	A	G	C	T
A	+5	−2	−2	−2
G	−2	+5	−2	−2
C	−2	−2	+5	−2
T	−2	−2	−2	+5

Recursive Definition of all scores of S(i,j)

$$S(i,j) = \max \begin{cases} S(i-1, j-1) + \sigma(x_i, y_j) \\ S(i-1, j) + \gamma \\ S(i, j-1) + \gamma \end{cases}$$

$$\gamma = -6$$

Once the matrix is filled-in. We can start at the bottom cell and ask, how we could have gotten here?

No gap with gap alignment is allowed

Possibility of more than one best alignments

Build a matrix of dimensions m+1 by n+1

First we fill in the boundary conditions
S(0,0) = 0, Fill First Row/Column

+5 Match; -2 Mis-Match and **-6 for INDELs**

j → Sequence y

i ↓ Sequence x

		0	1	2	3	4	5	6	7	8 = N
			T	G	C	T	C	G	T	A
0										
1	T									
2	T									
3	C									
4	A									
5	T									
6 = M	A									

Based on Eddy, S. Talk and papers

TGCTCGTA

+5 Match; -2 Mis-Match and -6 for INDELs

j ⟶ Sequence y

i

Sequence x

		0	1	2	3	4	5	6	7	8 = N
			T	G	C	T	C	G	T	A
0		0	← -6	← -12	← -18	← -24	← -30	← -36	← -42	← -48
1	T									
2	T									
3	C									
4	A									
5	T									
6 = M	A									

Based on Eddy, S. Talk and papers

Example of Local Alignment

TTCATA

+5 Match; -2 Mis-Match and **-6 for INDELs**

j → Sequence y

i ↓ Sequence x

		0	1	2	3	4	5	6	7	8 = N
			T	G	C	T	C	G	T	A
0		0	← -6	← -12	← -18	← -24	← -30	← -36	← -42	← -48
1	T	↑ -6								
2	T	↑ -12								
3	C	↑ -18								
4	A	↑ -24								
5	T	↑ -30								
6 = M	A	↑ -36								

Based on Eddy, S. Talk and papers

+5 Match; -2 Mis-Match and **-6 for INDELs**

j → Sequence y

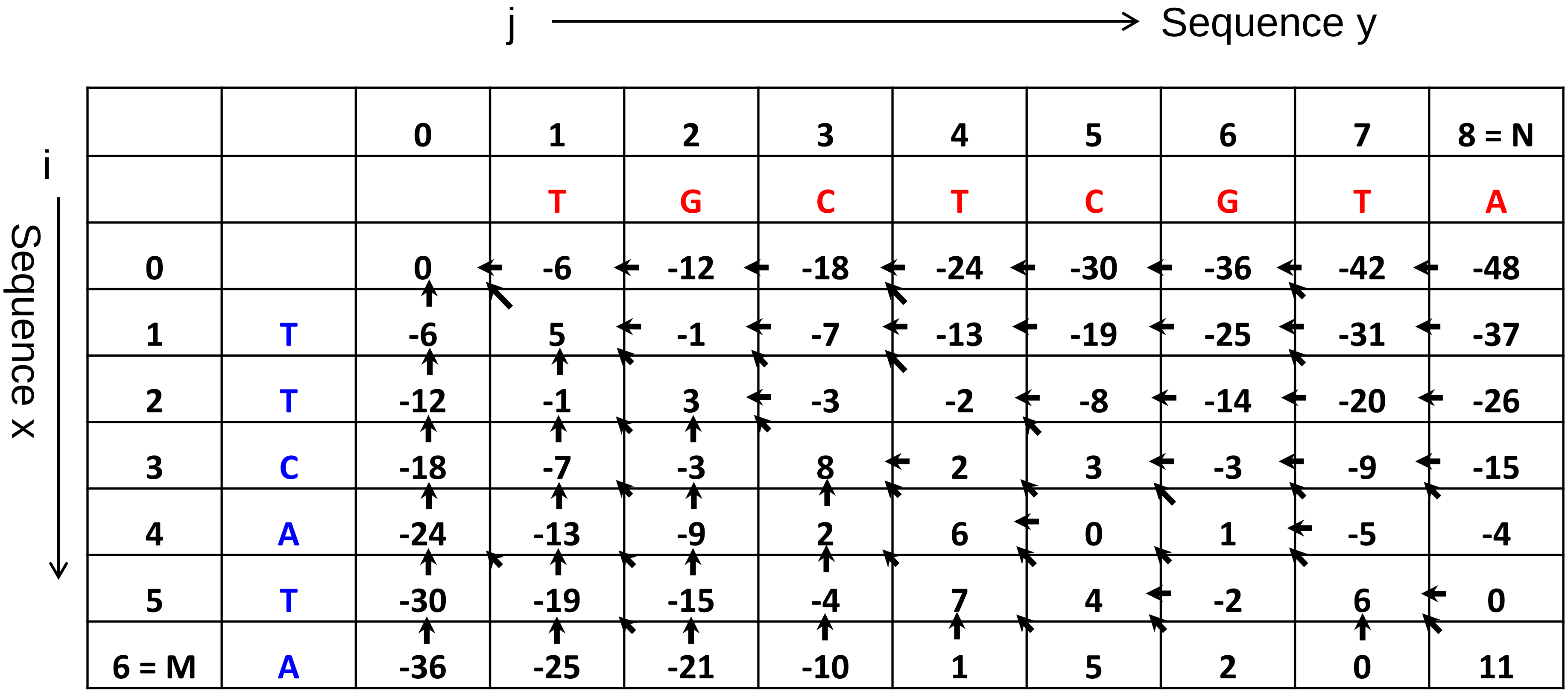
i

Sequence x

		0	1	2	3	4	5	6	7	8 = N
			T	G	C	T	C	G	T	A
0		0	← -6	← -12	← -18	← -24	← -30	← -36	← -42	← -48
1	T	↑ -6	↖ 5	← -1	← -7	↖ ← -13	← -19	← -25	↖ ← -31	← -37
2	T	↑ -12								
3	C	↑ -18								
4	A	↑ -24								
5	T	↑ -30								
6 = M	A	↑ -36								

Based on Eddy, S. Talk and papers

+5 Match; -2 Mis-Match and **-6 for INDELs**



		0	1	2	3	4	5	6	7	8 = N
			T	G	C	T	C	G	T	A
0		0	-6	-12	-18	-24	-30	-36	-42	-48
1	T	-6	5	-1	-7	-13	-19	-25	-31	-37
2	T	-12	-1	3	-3	-2	-8	-14	-20	-26
3	C	-18	-7	-3	8	2	3	-3	-9	-15
4	A	-24	-13	-9	2	6	0	1	-5	-4
5	T	-30	-19	-15	-4	7	4	-2	6	0
6 = M	A	-36	-25	-21	-10	1	5	2	0	11

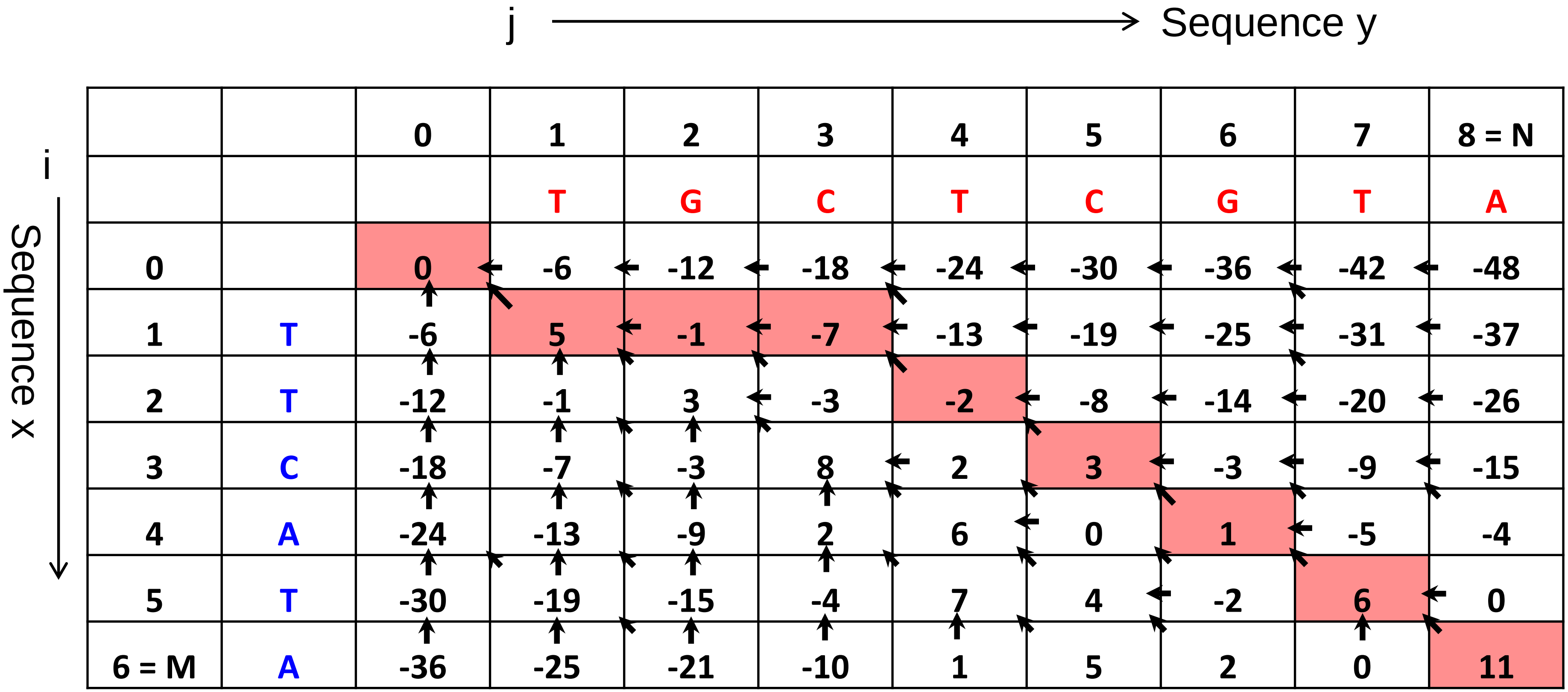
Based on Eddy, S. Talk and papers

TGCTCGTA

T--TCATA

56655255 = 11

+5 Match; -2 Mis-Match and -6 for INDELs


j
Sequence y
i
Sequence x

		0	1	2	3	4	5	6	7	8 = N
			T	G	C	T	C	G	T	A
0		0	-6	-12	-18	-24	-30	-36	-42	-48
1	T	-6	5	-1	-7	-13	-19	-25	-31	-37
2	T	-12	-1	3	-3	-2	-8	-14	-20	-26
3	C	-18	-7	-3	8	2	3	-3	-9	-15
4	A	-24	-13	-9	2	6	0	1	-5	-4
5	T	-30	-19	-15	-4	7	4	-2	6	0
6 = M	A	-36	-25	-21	-10	1	5	2	0	11

Based on Eddy, S. Talk and papers

Global Alignment

- Scoring Scheme:
  - Match: 1; MisMatch = 0; Gap Penalty = 0

Dynamic Programming

- Guaranteed to give you the optimal alignment
  - Biologically meaningful or not is not the algorithm’s problem
  - Scoring scheme should address that question
- Dyn. Prog. Algorithm can also align two random sequences (**Example that we worked on today**)
  - Statistical theory should be able to address this question using the alignment scores for this point

Local Alignment Algorithm and Extensions

- Similar algorithm as we discussed today
  - Some differences (m x n rather than (m+1) x (n+1)
- What is being used today?
  - A modified version
  - Why?
    - Time
    - Most problems
      - Query search DB
      - Align two sequences of length m and n is roughly m*n
      - Or query (length m) against a DB (N) is m * N
    - Big-oh notation (O(mn) for Needleman-Wunch)

Query length: m
DB Length: N

Types of pair-wise alignments

- Last one in this category
- Word methods
  - Heuristic; Will not produce the best alignment always but very fast; commonly used
  - More on this in future classes

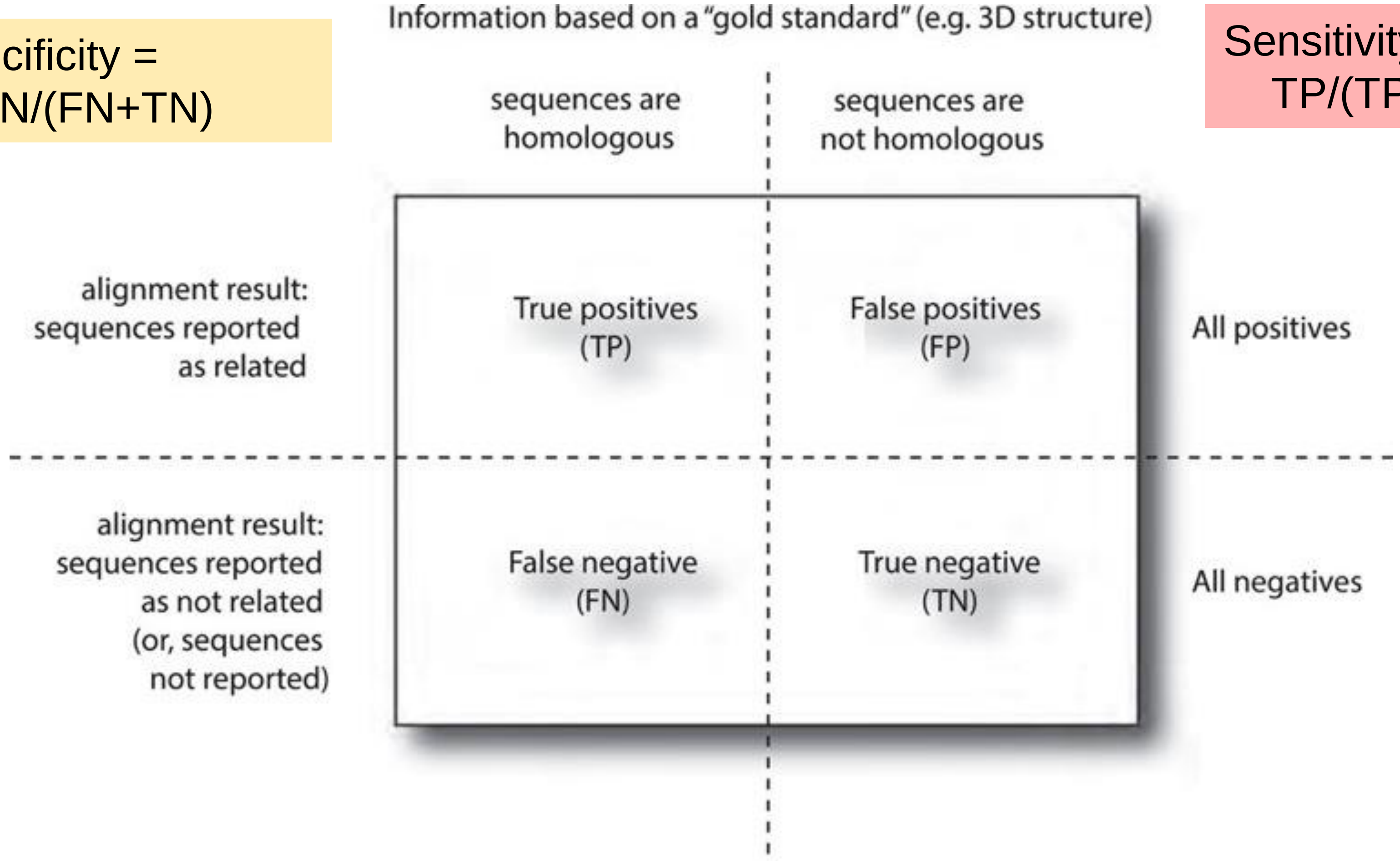
Beyond Simple DP

- FASTA
  - William Pearson (Univ. of Virginia)
- BLAST
  - Next class

How are we doing? Comparing to Gold(??) standard

Specificity = TN/(FN+TN)

Sensitivity = TP/(TP+FN)



Information based on a “gold standard” (e.g. 3D structure)

	sequences are homologous	sequences are not homologous	
alignment result: sequences reported as related	True positives (TP)	False positives (FP)	All positives
alignment result: sequences reported as not related (or, sequences not reported)	False negative (FN)	True negative (TN)	All negatives

Figure 3.26 from Bioinformatics and Functional Genomics, (3rd Ed.) by Jonathan Pevsner
PLEASE DO NOT DISTRIBUTE-Copyright figure

Central Limit Theorem (CLT)

"Central Limit Theorem states that the distribution of the sum (or average) of a large number of independent, identically distributed variables will be approximately normal, regardless of the underlying distribution."
http://www.math.uah.edu/stat/sample/CLT.html

Sample Mean Distribution will become increasingly close to a normal distribution as the sample size increases, regardless of the population distribution

Simple Random Sample: draws uniformly at random without replacement from the population

The Central Limit Theorem (CLT)

X_1 = [150, 140, 130, 121.5, 141.9]

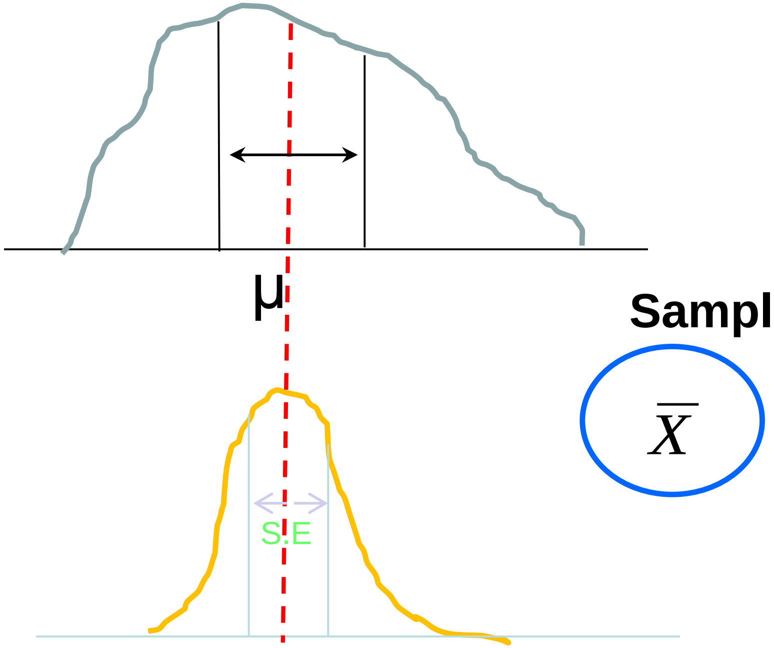
Sample size: n

Sample means: (center)

$$\bar{X}_1, \bar{X}_2, \bar{X}_3 \ldots$$

? Distribution of $\bar{X}$

- **They have a mean of** μ
- **Have SE** $\dfrac{\text{sd}}{\sqrt{n}}$



$$Mean_{\bar{X}} = \mu$$

$n \to \infty$ **Sampling → Normal approaches**

Parameter vs Statistic

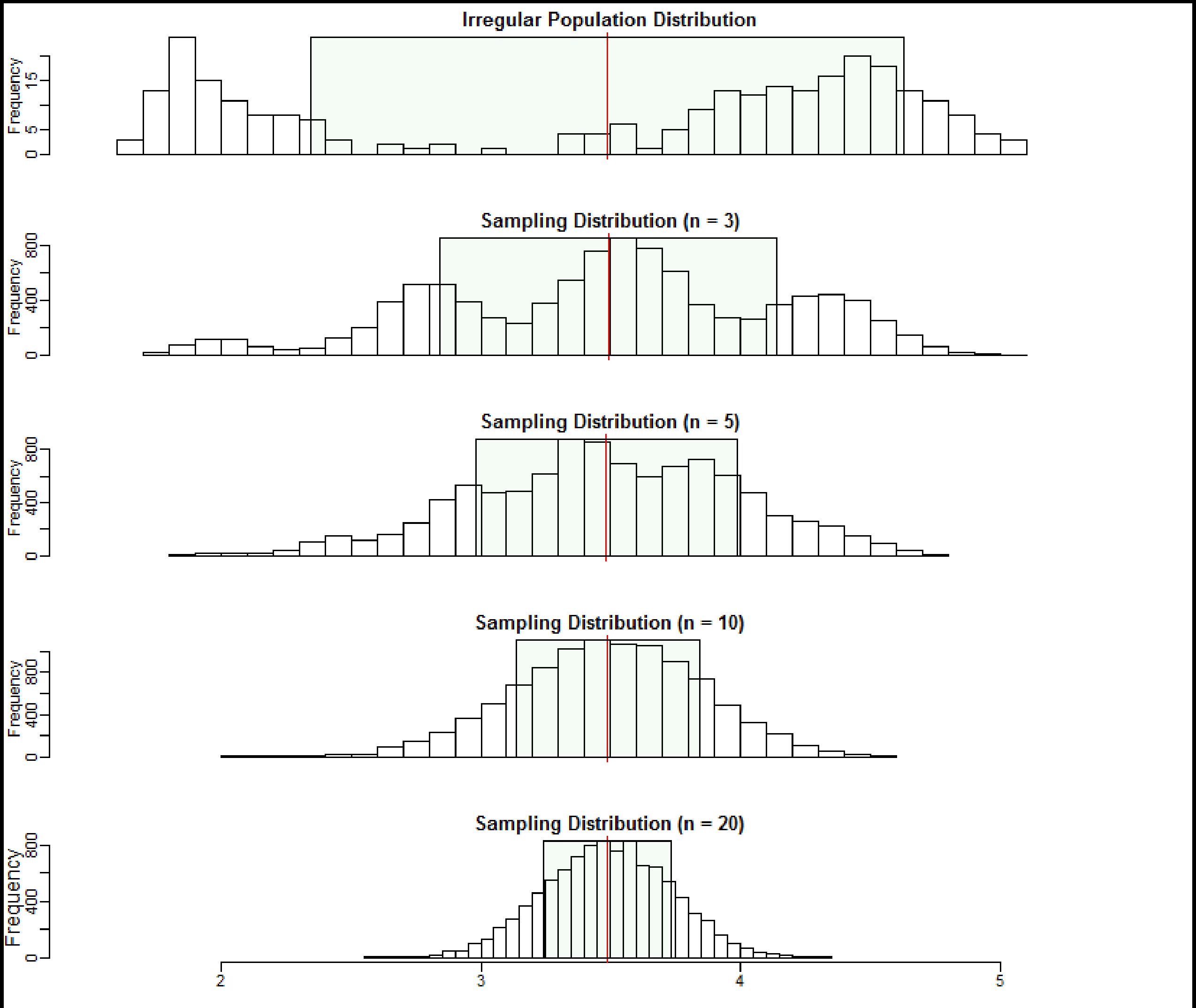
- Parameter
  - – Numerical descriptive measure, random, one that describes the population

- Statistic
  - – Numerical descriptive measure, random, one that describes the sample

Central Limit Theorem in Figures

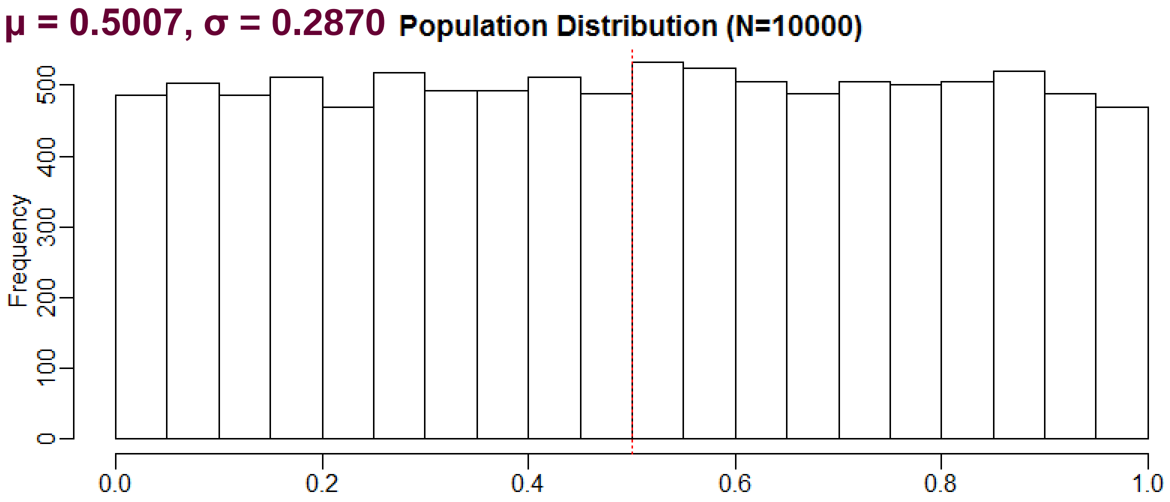
Number of Samples = 10,000

1SD Window shown

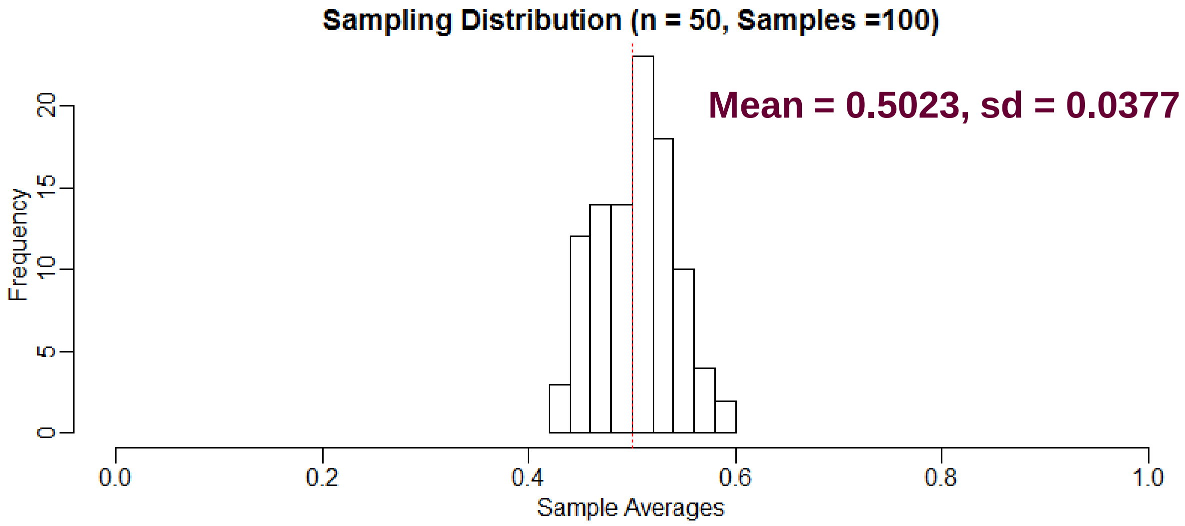


Sampling Distribution & Population Distribution Shape

Uniform



Sample
SD = sd

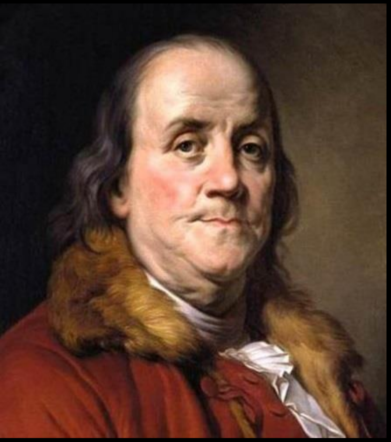


Normal

$$SE = \frac{\text{sd}}{\sqrt{n}}$$

0.00534

Hypothesis Testing & US Judicial System



"that it is better [one hundred] guilty Persons should escape than that one innocent Person should suffer."

Jury	Person	
	Innocent	Guilty
Not Guilty	✓	✗
Guilty	✗	✓

Hypothesis Testing & US Judicial System

US	Population	
	$\mu = \mu_0$	$\mu \neq \mu_0$
Fail to Reject	✓	✖
Reject	✖	✓

Hypothesis Testing & US Judicial System

US	Population	
	$\mu = \mu_0$	$\mu \neq \mu_0$
Fail to Reject	✓	Type-II
Reject	Type-I	✓

Hypothesis Testing & US Judicial System

US	Population	
	$\mu = \mu_0$	$\mu \neq \mu_0$
Fail to Reject	✓	Type-II = β
Reject	Type-I = α	✓

$1 - \beta$ = Power

Goal is to make α and β smaller

Hypothesis testing by example

- H_0: μ_0 = 15.5 mm Hg
- H_A: $\mu_0 \neq$ 15.5 mm Hg

Sample	Population
X = 16.5 mm Hg	μ_0 = 15.5 mm Hg
n = 49	σ = 2.6 mm Hg

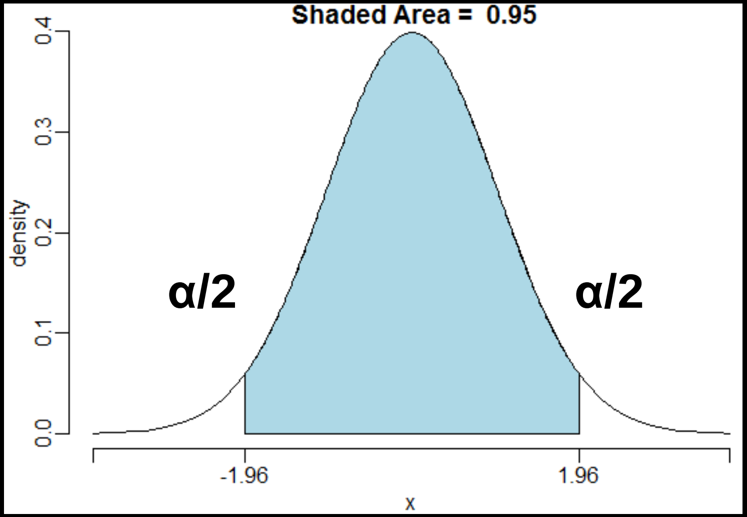
– Establish what value(s) we will accept to be different from the NULL distribution (α level = 0.05)

– Assuming CLT, we can perform an one sample Z-test

Hypothesis testing by example

Alpha values lead us to Critical Values

Values that represent Z values that correspond to α/2 = ±2.5% are called critical values (± 1.96)



P-value Definition

- Given a H_0, H_A and a Test Statistic T, the p-value can be defined as

"*the probability, computed assuming that H_0 is true, that the test statistic would take a value as extreme or more extreme than that actually observed*"

Moore, D.S. (2007) The Basic Practice of Statistics

Statistical Significance

When we align two proteins, we get a score. We often use Hypothesis Testing to ascertain whether this could have resulted by chance

- Hypothesis Testing
  - Null Hypothesis (H_0):
    - Two sequences are not related
  - Alternate Hypothesis (H_A)
    - Yes, they are evolutionarily related
  - Cutoff (usually 0.5 but can be different)
- Generate sample random alignments of the same composition as one of query sequences (One approach)

Statistical Significance

$$Z = \frac{x - \mu}{s}$$

- Sample mean and deviation
- Beta globin to myoglobin
  - Scramble myoglobin 1000 times
  - Compare the random sequence with beta globin
  - Compare the real score with the distribution (Gaussian?)
- How are we doling (real score) compared to the random sample alignments
- Hypothesis testing
- P-value

What if the distribution is not Gaussian or bell-shaped?

- Global (not Local alignment)
  - Not Gaussian
  - So, normal z-scores will be wrong
  - Refer to publications in the book
- Local alignment
  - Distribution is normal
- Normally Probability is not used, but a related value called E (more in this next class)

Bonferroni correction for multiple comparison

- We usually compare query to a DB. So, there is a chance of identifying an accidental high scoring alignment(s)
- For multiple comparisons, people often use Bonferroni correction
  - Use stringent cut-off
  - Cut-off/# of searches = $0.05/(10^{6}) \sim 10^{-8}$

Information Theory based approach by D. Altschul

- How to identify real from random alignments
- H = Relative Entropy (expected Substitution Score/residue)
- q_{ij} are target frequency $$H = \sum_{i,j} q_{i,j} s_{i,j} = \sum_{i,j} q_{i,j} \log_2 \frac{q_{ij}}{p_i p_j}$$
- P_i or P_j are background frequencies

- PAM250: H = 0.36 bits
- PAM10: H = 3.43 bits

H: Information Content of the target and background distributions for a particular scoring matrix

$$H = \sum_{i,j} q_{i,j} s_{i,j} = \sum_{i,j} q_{i,j} \log_2 \frac{q_{ij}}{p_i p_j}$$

H is the sum of all q_{ij} and S_{ij}

- Altschul estimated 30 bits of information are required to identify an authentic alignment (i.e. DB space = 2^30 = 1B)
- This means you need the DB size to be 1B to raise above the background noise.
- If you know this you can calculate what alignment length I should have to get meaningful results

H: Information Content of the target and background distributions for a particular scoring matrix

$$H = \sum_{i,j} q_{i,j} s_{i,j} = \sum_{i,j} q_{i,j} \log_2 \frac{q_{ij}}{p_i p_j}$$

H is the sum of all q_{ij} and S_{ij}

- PAM10, H = 3.43, you need an alignment with at least 9 residues

9 * 3.43 = 30.87 ~ 31

- PAM250, H = 0.36, at least 83 aa residues are needed to distinguish an authentic alignment

83 * 0.36 = 29.88 ~ 30

Rel entropy
= -H + I
= -H + 4.3

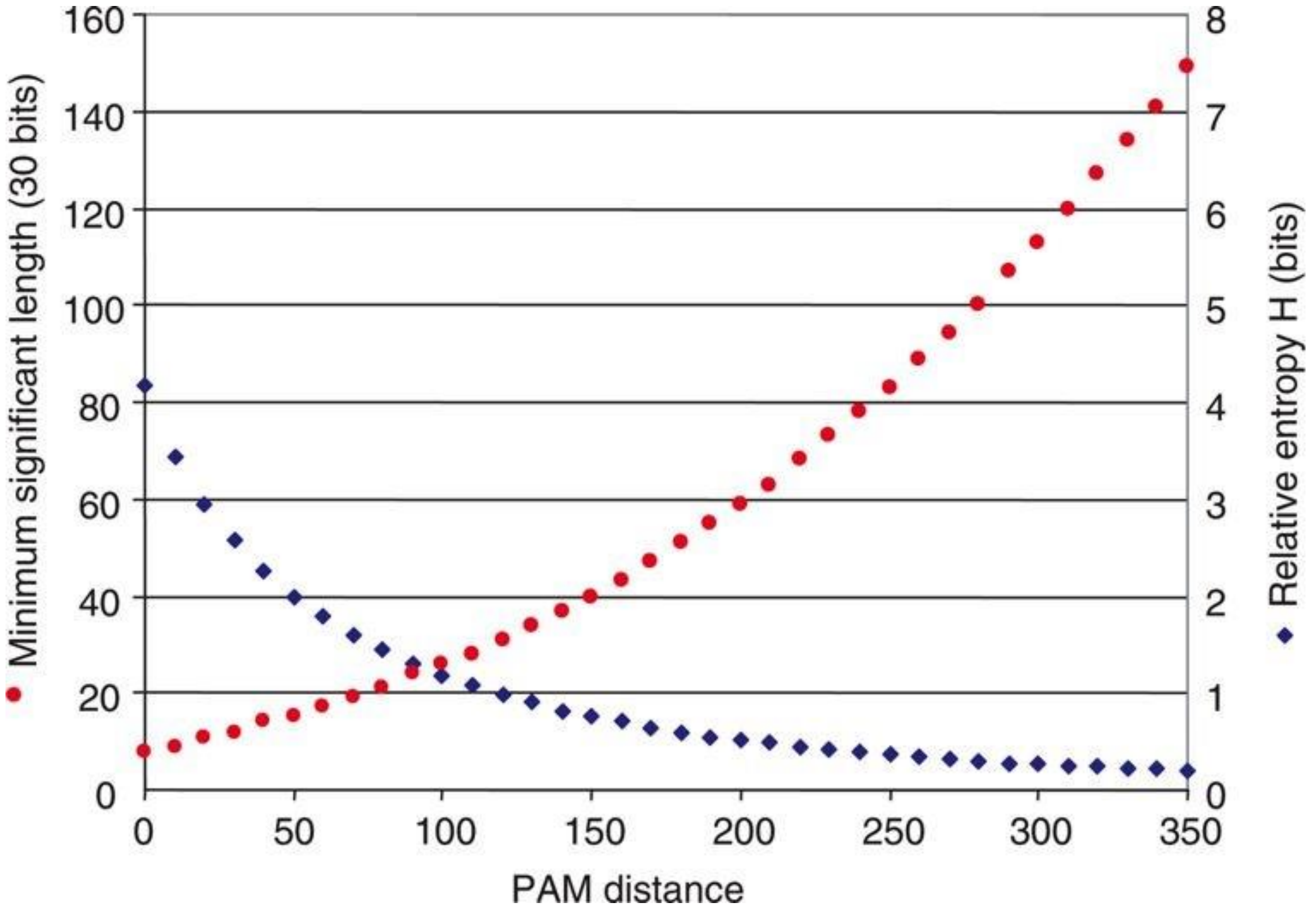


Figure 3.26 from Bioinformatics and Functional Genomics, (3rd Ed.) by Jonathan Pevsner
PLEASE DO NOT DISTRIBUTE-Copyright figure

Typos

- Page number 87/88

```
pam250 <- pam^250
```

  - Matrix multiplication

Computer Lab

- Problems/Computer Lab
  - 3-2, 3-4,3-6, 3-7

Thanks

ravichandran@hood.edu