Spring 2024

BIOS 477/877

Bioinformatics and Molecular Evolution

Lecture 8

1

TODAY'S TOPICS

- Gap Penalty
- Scoring (Substitution) Matrix
- Assignment 4

2

Alignment Strategy

- **Protein alignment is easier** than DNA alignment
  - → DNA has only 4 nucleotide types (they can match just by chance more easily)
  - → Protein sequences evolve more slowly than DNA sequences (genetic code is redundant; nonsynonymous substitutions are less frequent than synonymous substitutions)
- If DNA sequences are from coding regions:
  - → Translate them, and **align at the protein level** first
  - → This ensures gaps inserted between codons (prevents insertion of frame-shifting gaps)
- Do not blindly rely on the default parameter set
  - → Try various **scoring matrices, gap penalties, *etc.***

3

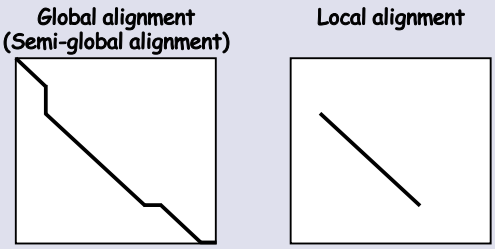
DNA alignment at protein level

1. Translate DNA sequences to amino acid sequences
2. **Align them at the protein level**
3. Reverse translate the protein alignment to DNA alignment

- **TranslatorX** http://translatorx.co.uk/
- **RevTrans 2.0** https://services.healthtech.dtu.dk/services/RevTrans-2.0/
- **PAL2NAL** http://www.bork.embl.de/pal2nal/
- **tranalign** https://www.bioinformatics.nl/cgi-bin/emboss/tranalign
  (included in EMBOSS servers; see the course/Link page.)

4

Global vs. local alignments



$s(a_i,b_j) = 2$ where $a_i=b_j$
$s(a_i,b_j) = -1$ where $a_i \neq b_j$
$w = -2$
(w=0 for free end gap)

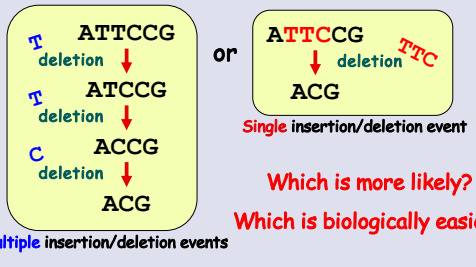
Match/Mismatch scores

Gap penalty

5

Indel Evolution and Gap Penalty

- A gap of length $k \neq k$ gaps of length 1



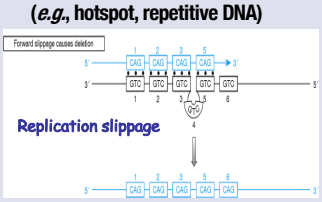
Which is more likely?

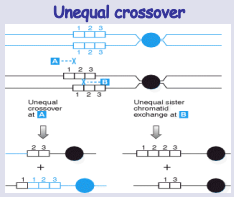
Which is biologically easier?

6

Indel Evolution and Gap Penalty

- Indel mutations are often strongly deleterious
- Indel events are rare (less common than point mutations)
- Multi-residue indels are not uncommon
  (*e.g.*, hotspot, repetitive DNA)





From Human Molecular Genetics 2 (available in NCBI Bookshelf)

7

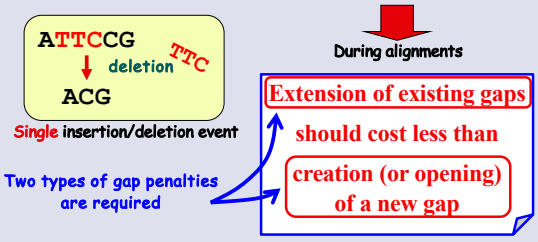
Indel Evolution and Gap Penalty

- Indel mutations are often strongly deleterious
- Indel events are rare (less common than point mutations)
- Multi-residue indels are not uncommon
- Fewest number of unlikely events
  → most likely evolutionary hypothesis
  Maximum parsimony

8

Indel Evolution and Gap Penalty

➢ Fewer, but longer, indel event is more likely than too many small indels



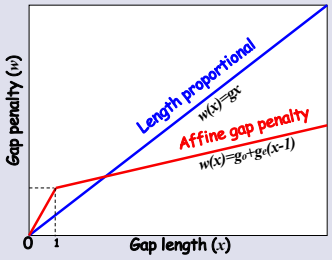
9

Gap Penalty Functions

- Linear (length-proportional) gap penalty:
  $w(x) = gx$
  g: gap penalty
  x: length of a gap
- Affine gap penalty:
  $w(x) = \begin{cases} g_o + g_e (x - 1) & \text{when } x > 0 \\ 0 & \text{when } x = 0 \end{cases}$
  g_o: gap opening penalty
  g_e: gap extension penalty [usually $g_o > g_e$]
  x: length of a gap

10

Gap Penalty Functions



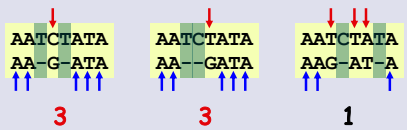
Affine gap penalty:
Extension of existing gaps costs less than creation/opening of new gaps

(g_o: gap opening penalty, g_e: gap extension penalty)

11

Simple Alignments

➢ Varied length & gaps considered



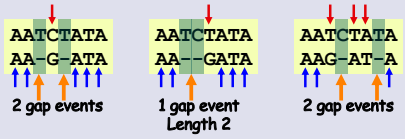
- Alignment Score =
  (match score) x (the number of matched pairs) +
  (mismatch score) x (the number of mismatched pairs) +
  (gap penalty) x (the number of gaps)

If match score = 1, mismatch score = 0, gap penalty = -1
& if using linear gap penalty,

12

Simple Alignments

➢ Varied length & gaps considered



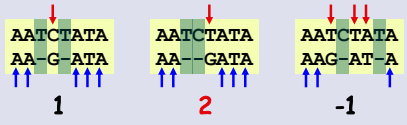
• Alignment Score =
(match score) x (the number of matched pairs) +
(mismatch score) x (the number of mismatched pairs) +
∑[(for each gap event) → {(gap opening penalty) +
(gap extension penalty) x (gap length - 1)}]

If match score = 1, mismatch score = 0,
& if using affine gap penalty
gap opening penalty = -2, gap extension penalty = -1

13

Simple Alignments

➢ Varied length & gaps considered



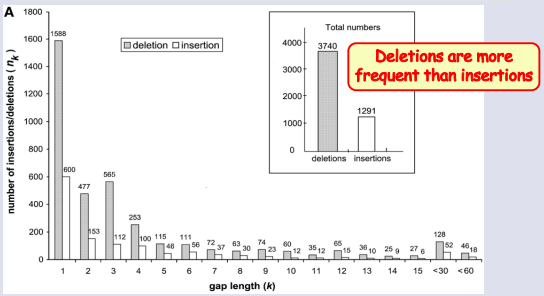
• Alignment Score =
(match score) x (the number of matched pairs) +
(mismatch score) x (the number of mismatched pairs) +
∑[(for each gap event) → {(gap opening penalty) +
(gap extension penalty) x (gap length - 1)}]

If match score = 1, mismatch score = 0,
& if using affine gap penalty
gap opening penalty = -2, gap extension penalty = -1

14

Empirical Indel Distribution: DNA

Based on the comparisons of >1700 processed pseudogenes against their functional homologues in the human genome



Zhang and Gerstein (2003)

15

Empirical Indel Distribution: DNA

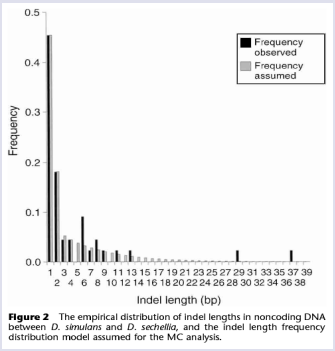


Figure 2 The empirical distribution of indel lengths in noncoding DNA between *D. simulans* and *D. sechellia*, and the indel length frequency distribution model assumed for the MC analysis.

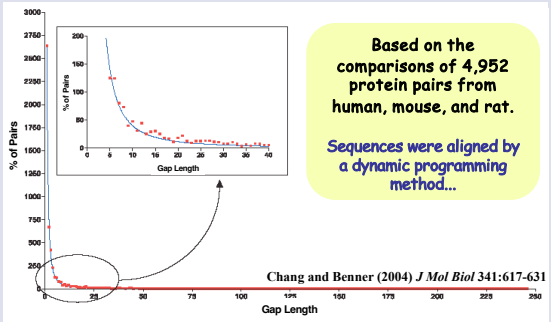
Based on the comparisons of 23 noncoding region sequences between *Drosophila simulans* and *D. sechellia*

Keightley and Johnson (2004) *Genome Res* 14:442-450

This distribution was later used in MCALIGN2:
Wang *et al.* (2006) *BMC Bioinformatics* 7: 292.

16

Empirical Indel Distribution: Protein



Based on the comparisons of 4,952 protein pairs from human, mouse, and rat.

Sequences were aligned by a dynamic programming method...

Chang and Benner (2004) *J Mol Biol* 341:617-631

17

Gap Penalty Function (more realistic)

➢ Empirical indel size distributions (both for DNA and proteins) can be described by a power law:
$f_k = Ck^{-b}$ [k: indel size, b: the power parameter]

→ Corresponding gap penalty function
$w = a + b \ln(k)$
a: gap opening penalty
b: gap extension penalty

• Gap extension penalty is proportional to the logarithm of gap length k (logarithmic gap penalty system)

→ increases more slowly with gap length than in the affine gap penalty system (easier long gaps)

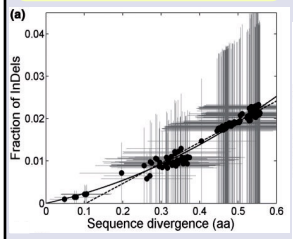
(*e.g.*, Cartwright 2006; Cartwright 2009; Loewenthal *et al.* 2021)

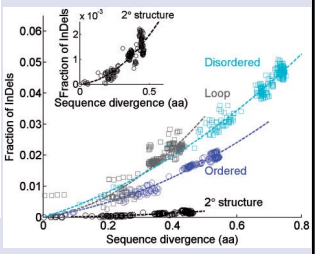
18

Empirical Indel Distribution: Protein

Based on the comparisons of 1,310 orthologous families from 22 fungal species.

Loop regions have more indels compared to the regions with secondary structures

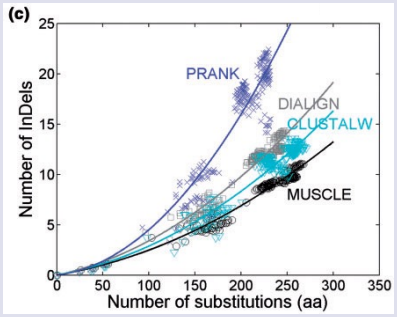




Toth-Petroczy and Tawfik (2013) *Mol Biol Evol* 30: 761-771

19

Empirical Indel Distribution: Protein



Indel rates/lengths are affected by different alignment methods!

Toth-Petroczy and Tawfik (2013) *Mol Biol Evol* 30: 761-771

20

Scoring (Substitution) Matrix: DNA

➢ Match/mismatch scores can be expressed in a matrix format

i \ j	A	T	C	G
A	1	0	0	0
T	0	1	0	0
C	0	0	1	0
G	0	0	0	1

DNA Identity Matrix

Match score = 1
Mismatch score = 0

$s(i, j)$: the similarity score between nucleotides i and j
$w(x)$: gap penalty

21

Scoring (Substitution) Matrix: DNA

➢ Match/mismatch scores can be expressed in a matrix format

	A	T	C	G
A	1	?	?	?
T	?	1	?	?
C	?	?	1	?
G	?	?	?	1

Transition/Transversion Matrix

Match score = 1
Mismatch score:
{ transition = ?
{ transversion = ?

22

DNA Substitution Types

Pyrimidines

Cytosine (C)

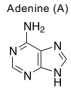
Uracil (U)

Thymine (T)

T ⟷ C
A ⟷ G

Purines

Adenine (A)



Guanine (G)

Transition substitutions: between pyrimidines, between purines
Transversion substitutions: between pyrimidines and purines

23

DNA Substitution Types

Pyrimidines

Cytosine (C)

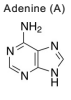
Uracil (U)

Thymine (T)

T ⟷ C
A ⟷ G

Purines

Adenine (A)



Guanine (G)

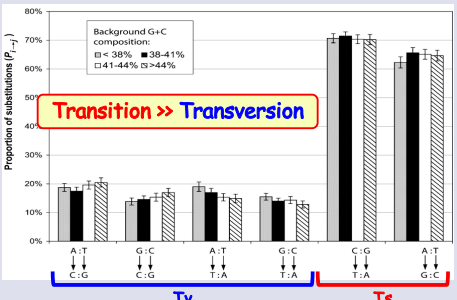
⟷ Transition ? ⟷ Transversion

Which substitutions happen more often?

24

DNA Substitution Types

Based on the comparisons of >1700 processed pseudogenes against their functional homologues in the human genome



Zhang and Gerstein (2003)

26

Scoring (Substitution) Matrix: DNA

- Match/mismatch scores can be expressed in a matrix format

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

Transition/Transversion Matrix

Match score = 1

Mismatch score:

- transition = -1 (more allowed → smaller penalty)
- transversion = -5 (fewer allowed → larger penalty)

27

Amino acid substitution matrices

- Identity matrix
- Genetic code matrix
- Matrices based on AA properties
- Matrices based on empirical data
  - Dayhoff matrices (PAM120 *etc.*)
  - BLOSUM matrices (BLOSUM62 *etc.*)
  - Gonnet matrices (Gonnet 250 *etc.*)
  - JTT matrices
  - and more …

AAindex (566 indices + ~150 matrices): https://www.genome.jp/dbget/aaindex.html

28

Protein identity matrix

	C	S	T	P	A	G	⋯
C	1	0	0	0	0	0	
S	0	1	0	0	0	0	
T	0	0	1	0	0	0	
P	0	0	0	1	0	0	
A	0	0	0	0	1	0	
G	0	0	0	0	0	1	
⋮							

Match score = 1
Mismatch score = 0

(20 x 20)

	C	S	T	P	A	G	⋯
C	6	-1	-1	-1	-1	-1	
S	-1	6	-1	-1	-1	-1	
T	-1	-1	6	-1	-1	-1	
P	-1	-1	-1	6	-1	-1	
A	-1	-1	-1	-1	6	-1	
G	-1	-1	-1	-1	-1	6	
⋮							

Match score = +6
Mismatch score = -1

29

Genetic code matrix

(A simpler version)

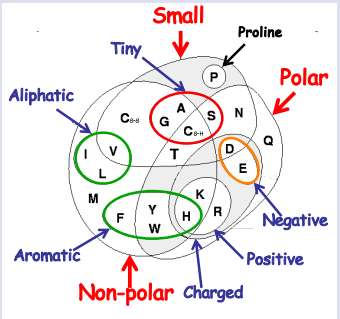
	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z
A	3.0	2.0	1.0	2.0	2.0	1.0	2.0	1.0	1.0	1.0	1.0	1.0	2.0	1.0	1.0	2.0	2.0	2.0	1.0	1.0	2.0	A
B		3.0	1.0	3.0	2.0	1.0	2.0	2.0	2.0	2.0	1.0	1.0	3.0	1.0	2.0	1.0	2.0	2.0	2.0	0.0	2.0	B
C			3.0	1.0	0.0	2.0	2.0	1.0	1.0	0.0	1.0	0.0	1.0	1.0	0.0	2.0	2.0	1.0	1.0	2.0	0.0	C
D				3.0	2.0	1.0	2.0	2.0	1.0	1.0	1.0	0.0	2.0	1.0	1.0	1.0	1.0	1.0	2.0	0.0	2.0	D
E					3.0	0.0	2.0	1.0	1.0	2.0	1.0	1.0	1.0	1.0	2.0	1.0	1.0	1.0	2.0	1.0	3.0	E
F						3.0	1.0	1.0	2.0	0.0	2.0	1.0	1.0	1.0	0.0	1.0	2.0	1.0	2.0	1.0	2.0	0.0	F
G							3.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	2.0	2.0	1.0	2.0	2.0	2.0	G
H								3.0	1.0	1.0	2.0	0.0	2.0	2.0	2.0	2.0	2.0	1.0	1.0	1.0	0.0	H
I									3.0	2.0	2.0	2.0	2.0	1.0	1.0	2.0	2.0	2.0	2.0	0.0	1.0	I
K										3.0	1.0	2.0	2.0	1.0	2.0	2.0	2.0	1.0	2.0	1.0	2.0	K
L											3.0	2.0	1.0	2.0	2.0	2.0	2.0	1.0	2.0	2.0	1.0	L
M												3.0	1.0	1.0	1.0	2.0	1.0	2.0	2.0	1.0	0.0	M
N													3.0	1.0	1.0	1.0	2.0	2.0	1.0	0.0	2.0	N
P														3.0	2.0	2.0	2.0	2.0	1.0	1.0	1.0	P
Q															3.0	2.0	1.0	1.0	1.0	1.0	3.0	Q
R																3.0	2.0	2.0	1.0	2.0	1.0	R
S																	3.0	2.0	1.0	2.0	1.0	S
T																		3.0	1.0	1.0	1.0	T
V																			3.0	1.0	1.0	V
W																				3.0	1.0	W
Y																					3.0	Y
Z																						3.0	Z

3.0: Identical (no change required)
2.0: One nucleotide change required
1.0: Two nucleotide changes required
0.0: Three nucleotide changes required

(B: D/Asp or N/Asn, Z: E/Glu or Q/Gln)

30

Amino acid physico-chemical properties

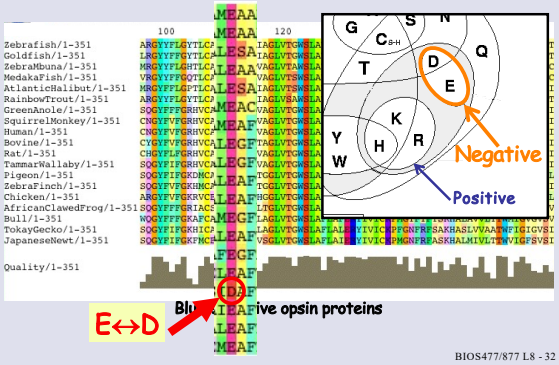


Major factors in Protein folding Protein functions

- Substitutions between similar amino acids are more common
- In functionally or structurally important regions, substitutions are limited between similar amino acids

31

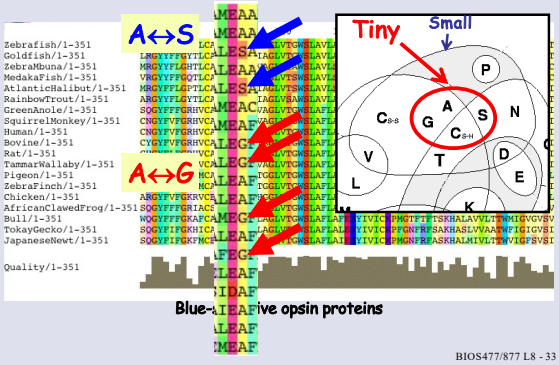
Amino acid substitutions



BIOS477/877 L8 - 32

32

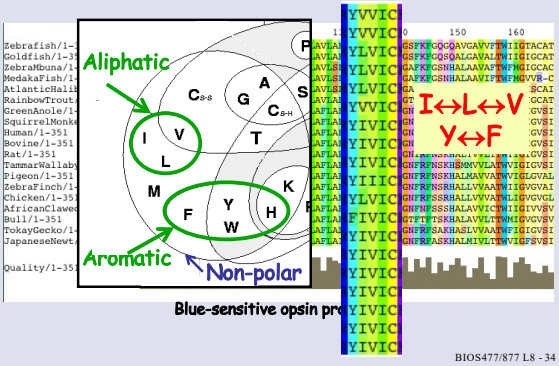
Amino acid substitutions



BIOS477/877 L8 - 33

33

Amino acid substitutions



BIOS477/877 L8 - 34

34

Amino acid substitution matrices

- Matrices based on various amino acid properties (hydrophobicity, charge, electronegativity, size, *etc.*)
  - → Biologically meaningful matrix can be obtained by combining all of these matrices (including genetic code matrix). Not easy!
- Matrices based on empirical data
  - → Alignments show the results of experiments done by the Nature
  - → Capture the relative substitutability of amino acid pairs in the context of evolution
  - → The model of protein evolution

BIOS477/877 L8 - 35

35

Amino acid substitution matrices

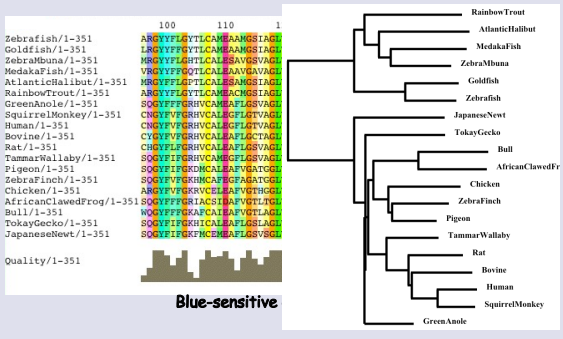
- Identity matrix
- Genetic code matrix
- Matrices based on AA properties
- Matrices based on empirical data
  - Dayhoff matrices (PAM120 *etc.*)
  - BLOSUM matrices (BLOSUM62 *etc.*)
  - Gonnet matrices (Gonnet 250 *etc.*)
  - JTT matrices
  - and more …

AAindex (566 indices + ~150 matrices): https://www.genome.jp/dbget/aaindex.html

BIOS477/877 L8 - 36

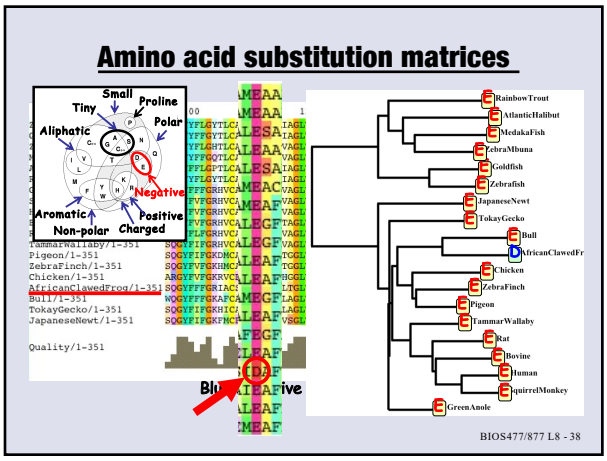
36

Amino acid substitution matrices

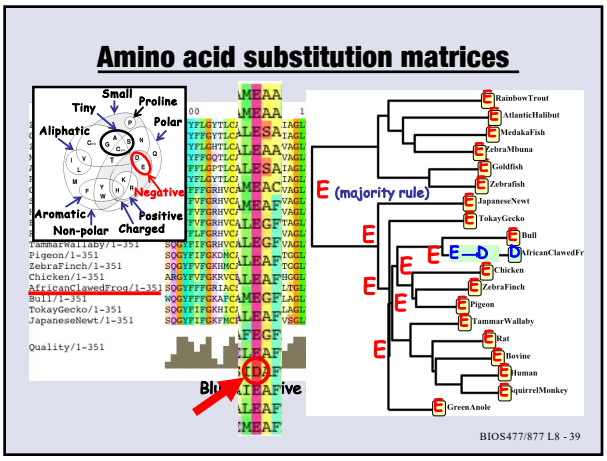


BIOS477/877 L8 - 37

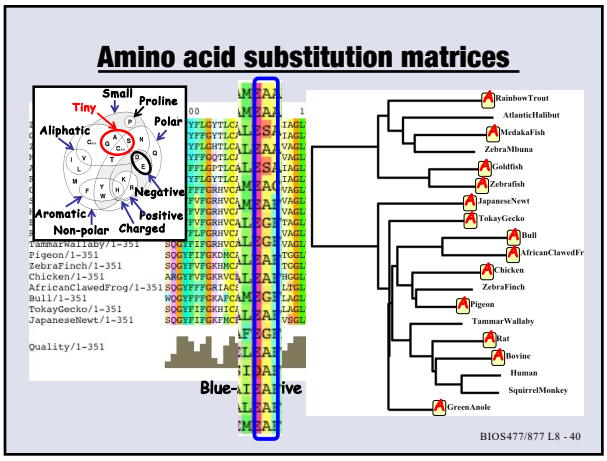
37



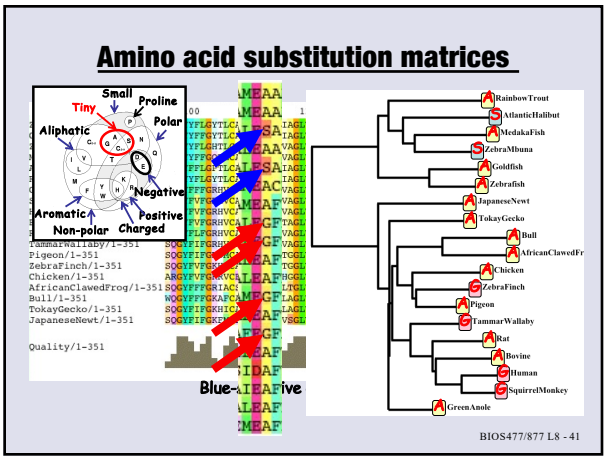
38



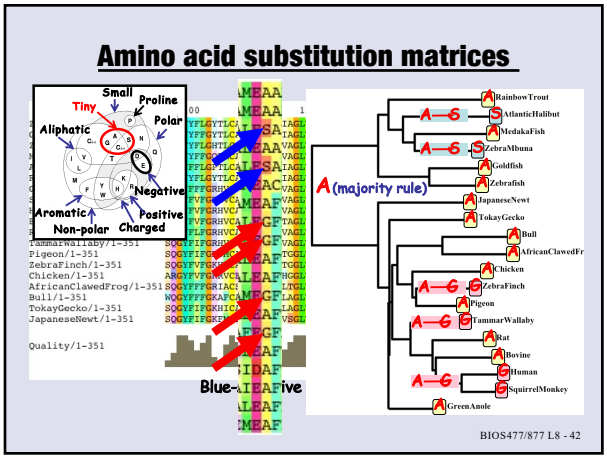
39



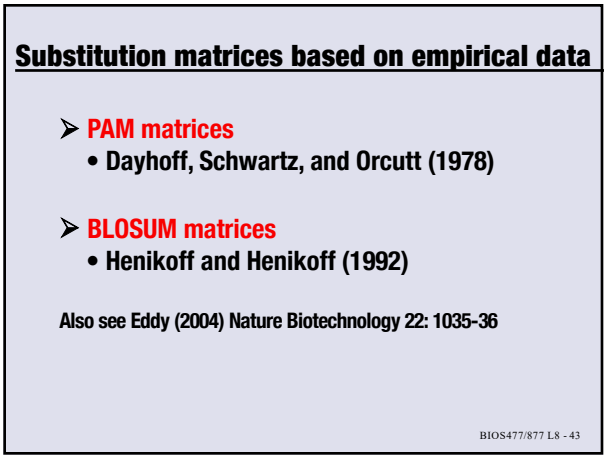
40



41



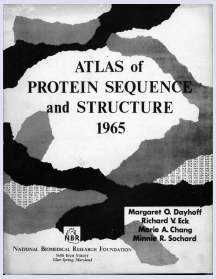
42



43

Margaret O. Dayhoff
(1925-1983)

Founder of
the field of Bioinformatics
The first woman in the field



Collection of all known protein sequences
1st *Atlas* contained 65 proteins

Developed into PIR (Protein Information Resource), a brain-child of Dayhoff

Dayhoff developed a single letter code for the amino acids

Read the Smithsonian website; also in Strasser (2010)

44

PAM matrices (Dayhoff *et al.* 1978)

- Accepted point mutations (point accepted mutations, percent accepted mutations)
  - → accepted by selection: no (or very weak) deleterious effect, maintaining the function
- Based on 1,572 changes in 71 groups of closely related proteins (34 protein families)
  - → at least 85% identical
  - → no ambiguity in alignments, no gap
  - → most likely observed substitutions do not affect protein functions (accepted by selection, close to neutral)
  - → successive (multiple) substitutions at one site are minimal (no hidden substitution)

45