BIOS 477/877
Bioinformatics and Molecular Evolution

Instructor: Etsuko Moriyama
(School of Biological Sciences)

Spring 2026 | Lecture 8

1

Today's topics

- Alignment strategy: a good practice
- Gap Penalty
- Scoring (Substitution) Matrix
- Graduate Only Assignment
- 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

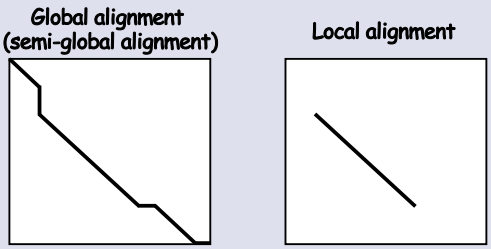
[Some useful tools]

- TranslatorX
- RevTrans 2.0
- PAL2NAL
- tranalign

(included in EMBOSS servers; see the course/Link page.)

4

Global vs. local alignments



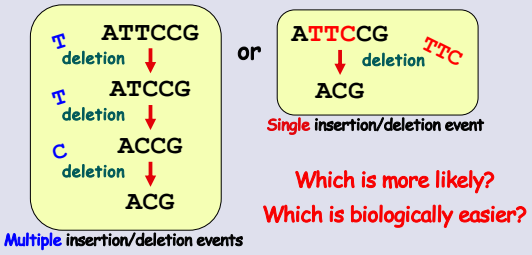
$s(a_i,b_j) = 2$ where $a_i=b_j$
$s(a_i,b_j) = -1$ where $a_i \neq b_j$
$w = -2$
(w=0 for free end gap)

Match/Mismatch scores
Gap penalty

5

Insertions and deletions: mechanisms

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

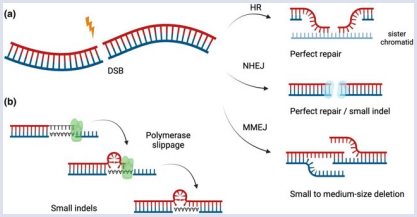


Which is more likely?
Which is biologically easier?

6

Insertions and deletions: mechanisms
(continued)

- 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)



DSB: double-stranded breaks
HR: homologous repair
NHEJ: non-homologous end joining
MMEJ: microhomology-mediated end joining

Redeling et al. (2024)

7

Insertions and deletions: mechanisms
(continued)

- 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

AATCTATA → AA-G-ATA (2 indels)
AATCTATA → AA--GATA (1 indel (more likely than 2 indel events))

8

Indel evolution and gap penalty

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

ATTCCG → deletion (TTC) → ACG

- Indels of multiple residues
- Single insertion/deletion event

During alignment:
Extension of existing gaps should cost less than creation (or opening) of a new gap

Two types of gap penalties are required

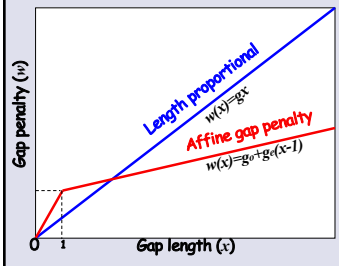
9

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



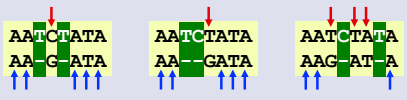
Affine gap penalty:
Extension of existing gaps costs less than creation/opening of new gaps

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

11

Simple alignments

- Varied length & gaps considered



Score= 3 3 1

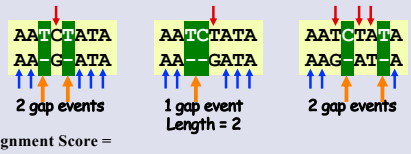
Alignment Score =
(match score) x (the number of matched pairs) +
(mismatch score) x (the number of mismatched pairs) +
(gap penalty) x (the number of gaps)

Use match score = 1, mismatch score = 0, gap penalty = -1
using linear gap penalty

12

Simple alignments (with affine gap penalty)

- Varied length & gaps considered



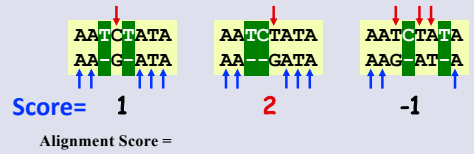
Alignment Score =
(match score) x (the number of matched pairs) +
(mismatch score) x (the number of mismatched pairs) +
∑[(for each gap event) → {(gap opening penalty) +
(gap extension penalty) x (gap length - 1)}]

Match score = 1, mismatch score = 0, & using affine gap penalty:
gap opening penalty = -2, gap extension penalty = -1

13

Simple alignments (with affine gap penalty)

- Varied length & gaps considered



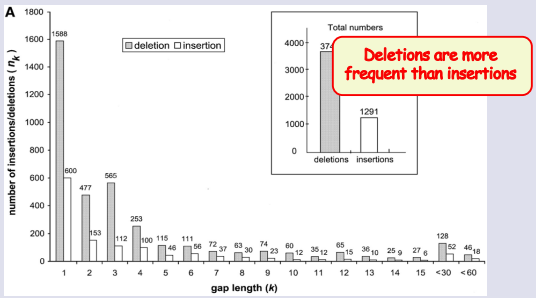
Alignment Score =
(match score) x (the number of matched pairs) +
(mismatch score) x (the number of mismatched pairs) +
∑[(for each gap event) → {(gap opening penalty) +
(gap extension penalty) x (gap length - 1)}]

Match score = 1, mismatch score = 0, & 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



Deletions are more frequent than insertions

Zhang and Gerstein (2003)

15

Empirical indel distribution: DNA

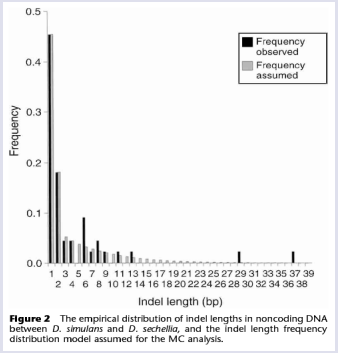


Figure 2 The empirical distribution of indel lengths in noncoding DNA between *D. simulans* and *D. sechellia*, and the indel length frequency distribution model assumed for the MC analysis.

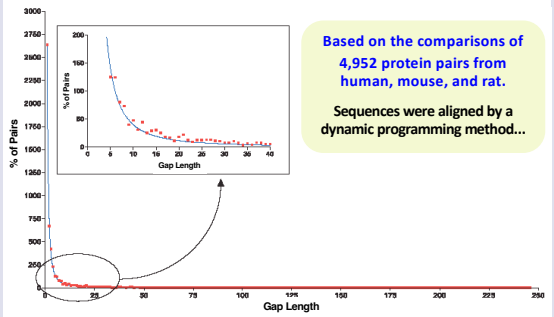
Based on the comparisons of 23 noncoding region sequences between *Drosophila simulans* and *D. sechellia*

Keightley and Johnson (2004)

This distribution was later used in MCALIGN2 (Wang *et al.* 2006)

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...

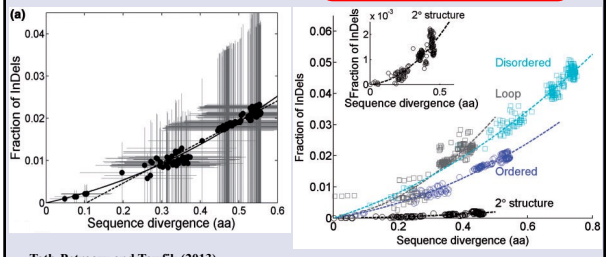
Chang and Benner (2004)

17

Empirical indel distribution: protein

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

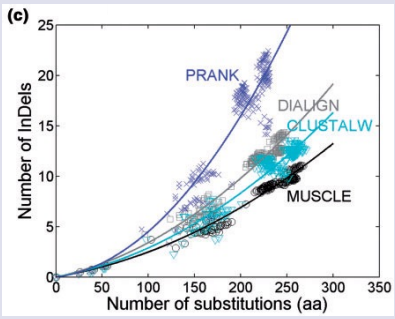
Loop regions have more indels compared to the regions with secondary structures



Toth-Petroczy and Tawfik (2013)

18

Empirical indel distribution: protein



Indel rates/lengths are affected by different alignment methods!

Toth-Petroczy and Tawfik (2013)

19

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

  Affine gap penalty: $w = a + bk$

  - 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 and 2009); Loewenthal *et al.* 2021; Wygoda *et al.* (2024)
Review: Redelings *et al.* (2024)

20

Scoring (substitution) matrix: DNA identity matrix

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

Match/mismatch scores can be expressed in a matrix format

Match score = 1
Mismatch score = 0

$s(i, j)$: the similarity score between nucleotides i and j

21

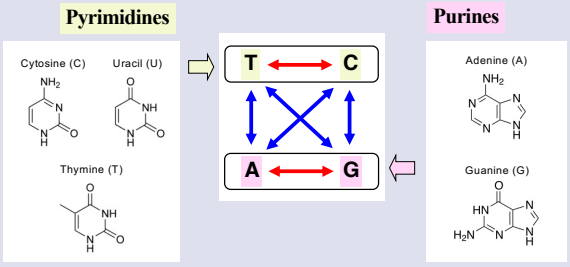
Scoring (substitution) matrix: Transition/transversion matrix

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

Match score = 1
Mismatch score:
{ Transition = ?
{ Transversion = ?

22

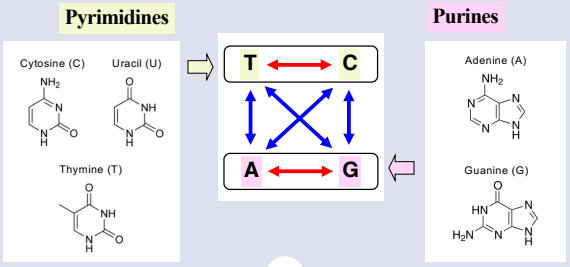
Types of DNA substitutions



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

23

Types of DNA substitutions



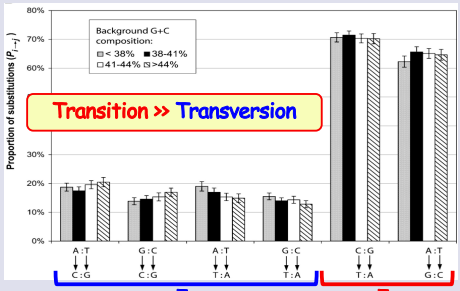
↔ Transition ? ↔ Transversion

Which substitutions happen more often?

24

Types of DNA substitutions

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



Zhang and Gerstein (2003)

26

Scoring (substitution) matrix: Transition/transversion matrix

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

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)

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)

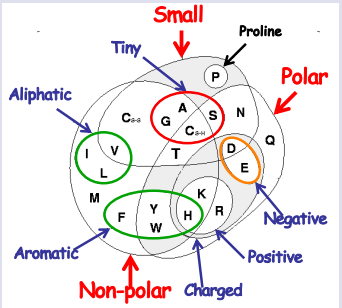
	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z
A	3.0	2.0	1.0	2.0	2.0	1.0	2.0	1.0	1.0	1.0	1.0	1.0	1.0	2.0	1.0	1.0	2.0	2.0	2.0	1.0	1.0	2.0
B		3.0	1.0	3.0	2.0	1.0	2.0	2.0	2.0	2.0	1.0	1.0	3.0	1.0	2.0	1.0	2.0	2.0	2.0	0.0	2.0	2.0
C			3.0	1.0	0.0	2.0	2.0	1.0	1.0	0.0	1.0	1.0	0.0	1.0	0.0	2.0	1.0	1.0	1.0	2.0	0.0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

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



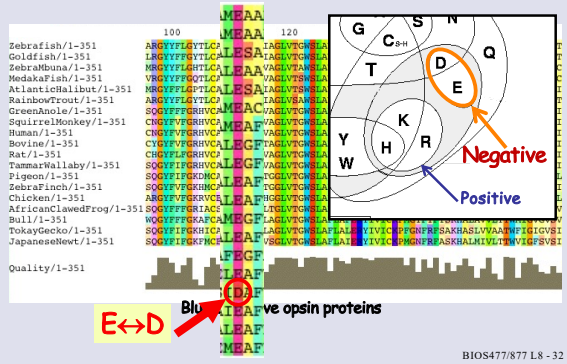
Physicochemical properties of amino acids:

- 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

Livingstone and Barton (1993)

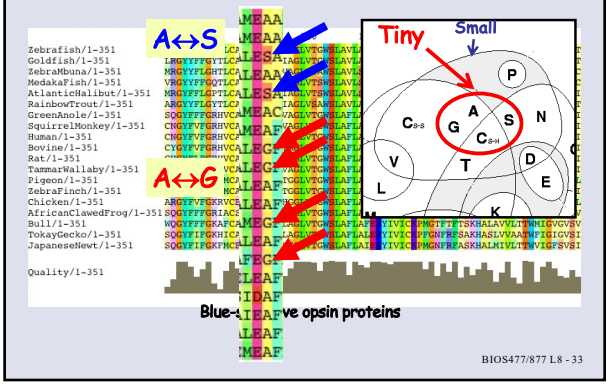
31

Amino acid substitutions



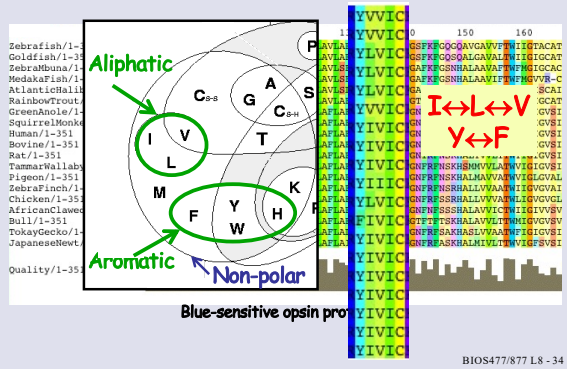
32

Amino acid substitutions



33

Amino acid substitutions



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 even genetic code matrix). But this is 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

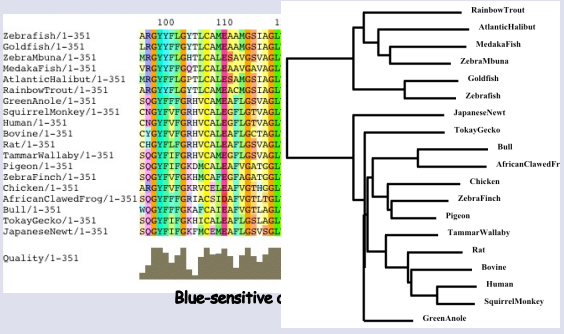
- Identity matrix
- Genetic code matrix
- Matrices based on AA properties
- Matrices based on empirical data
  - Dayhoff matrices (PAM120 etc.)
  - BLOSUM matrices (BLOSUM62 etc.)
  - Gonnet matrices (Gonnet 250 etc.)
  - JTT matrices
  - and more …

AAindex (566 indices + ~150 matrices)

BIOS477/877 L8 - 36

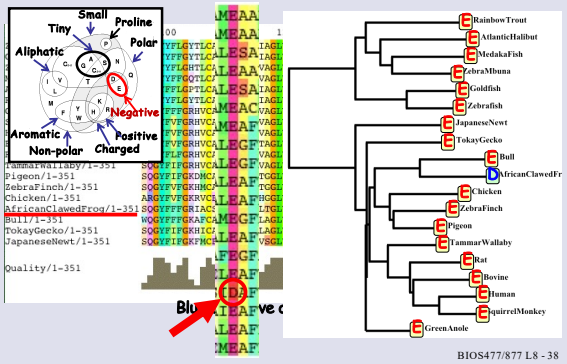
36

How to count amino acid substitutions



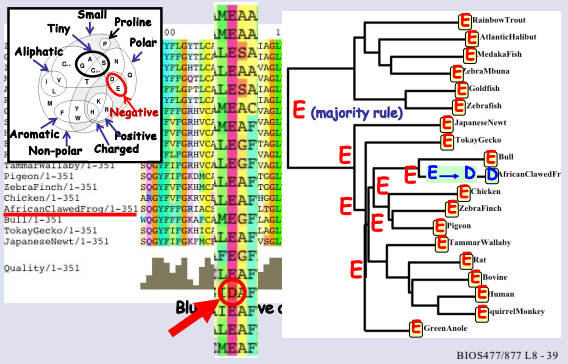
37

How to count amino acid substitutions



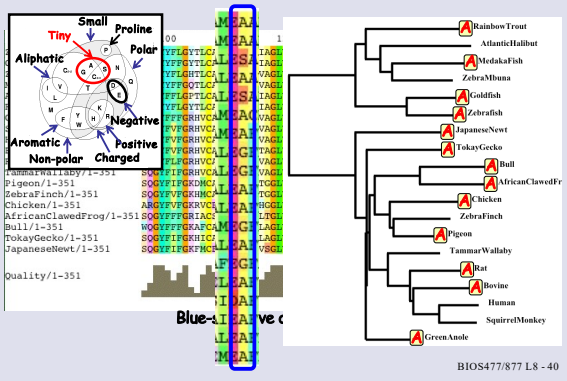
38

How to count amino acid substitutions



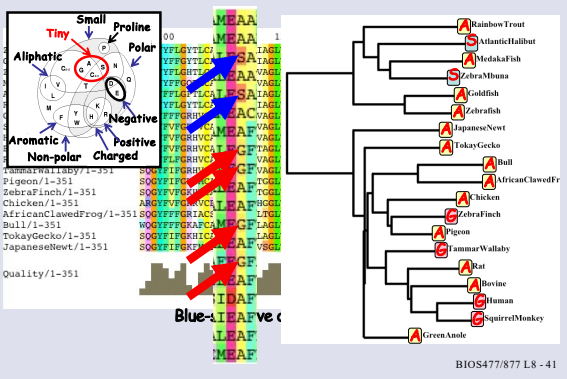
39

How to count amino acid substitutions



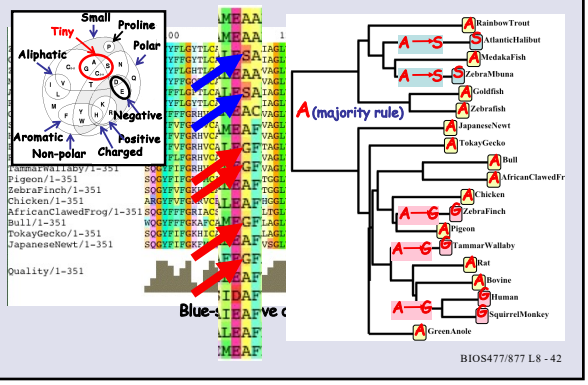
40

How to count amino acid substitutions



41

How to count amino acid substitutions



42

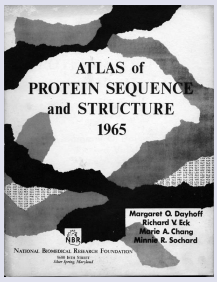
Amino acid substitution matrices based on empirical data

- **PAM matrices**
  Dayhoff, Schwartz, and Orcutt (1978)
- **BLOSUM matrices**
  Henikoff and Henikoff (1992)
  - Also see Eddy (2004)

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

Graduate Only Assignment

- **Report 1: March 8**
  - → Short proposal (2-3 pages)
- **Report 2: May 3**
  - → Final report (5-7 pages)
- Comparative analysis of bioinformatics methods
  - e.g., Multiple alignment methods:
    Clustal W vs. MUSCLE vs. MAFFT
    - → **Compare three methods**
    - → Choose appropriate input datasets (3 or more)
    - → Read the Guidelines!
- **Undergraduate students can do these assignments for bonus points (both reports need to be submitted)**

46