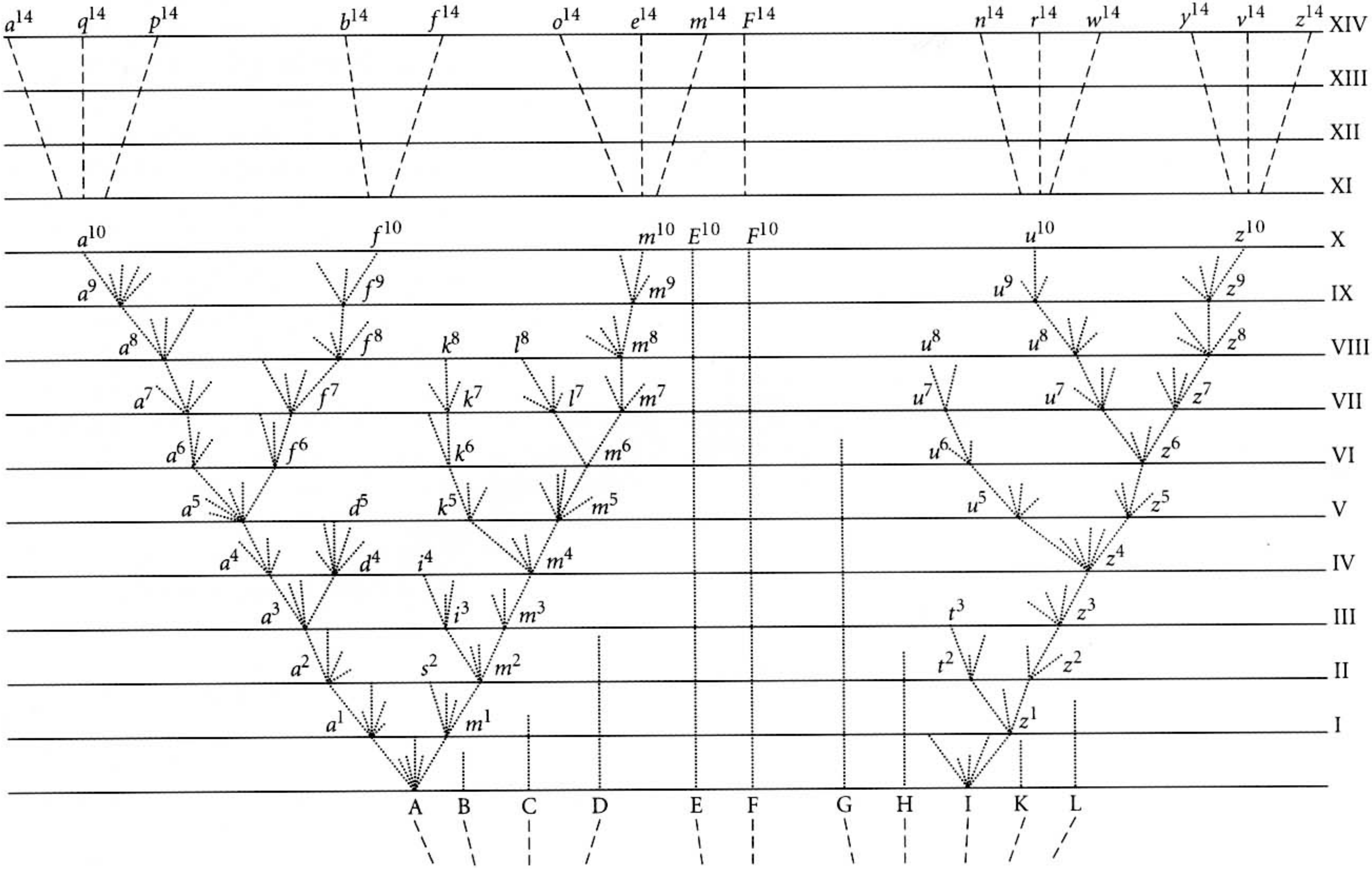
Week 3

Alignment, homology, evolution

Charles Darwin (1809-1882)

- Avid Naturalist from youth
- Bad Student
- Quit Med School
- Tried Clergy
- Connections got him the Naturalist job on the Beagle

The only figure in the Origin of Species



Chapter 1: Variation under Domestication

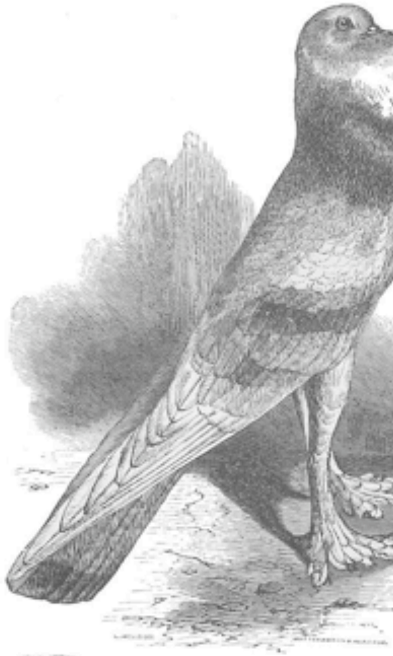


Fig. 18.—English Pouter.

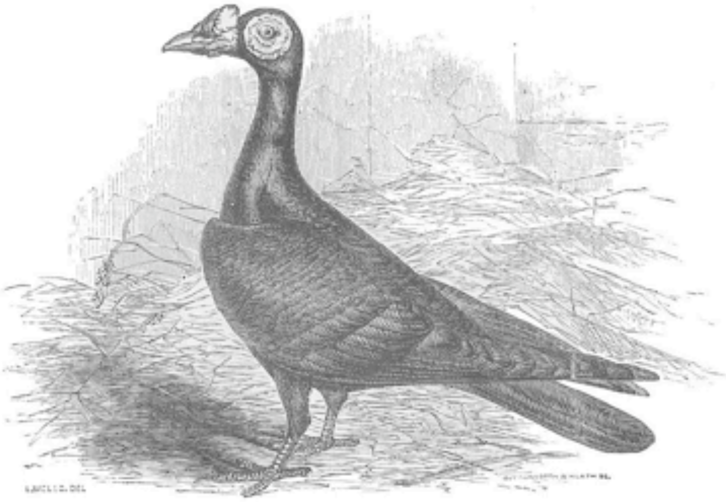


Fig. 19.—English Carrier.

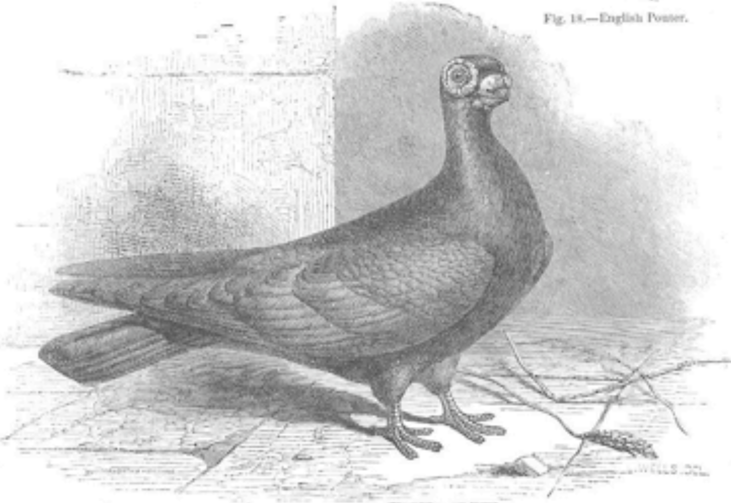


Fig. 20.—English Barb.

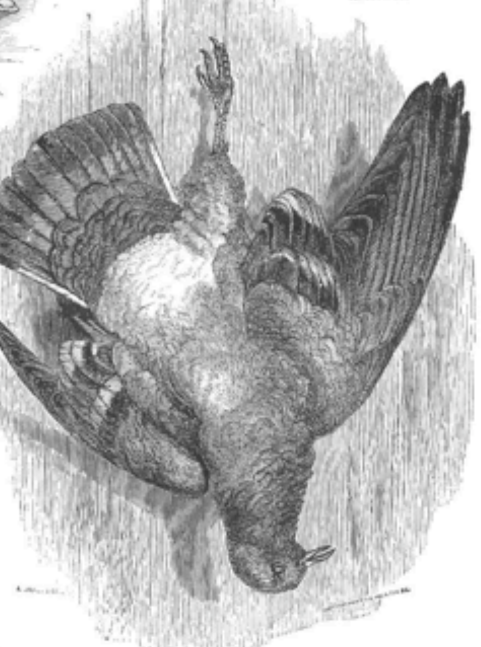


Fig. 17.—The Rock Pigeon, or Columba livia.*
The parent-form of all domesticated Pigeons.

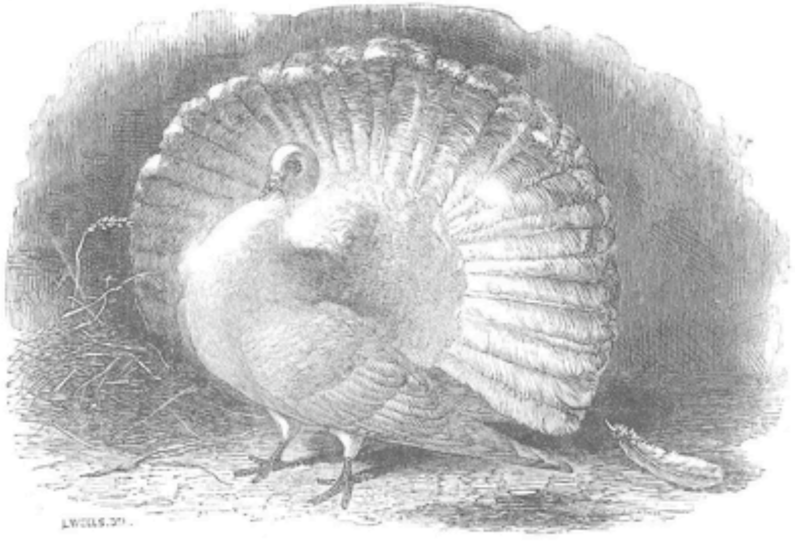


Fig. 21.—English Fantail.

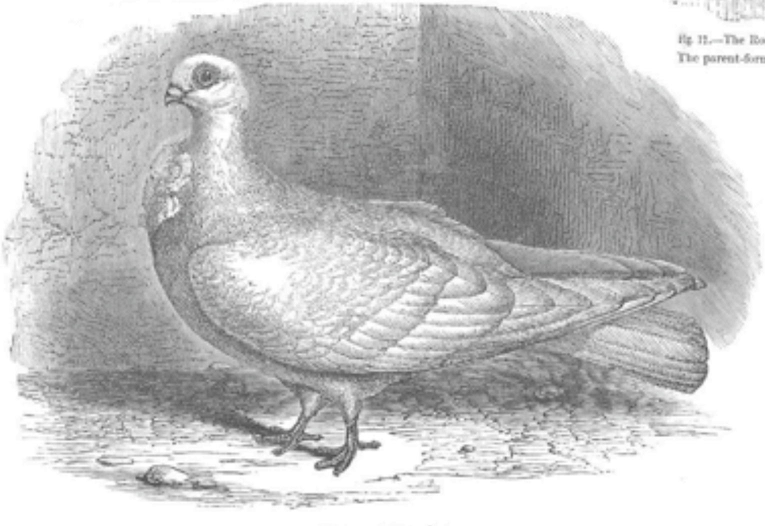


Fig. 22.—African Owl.

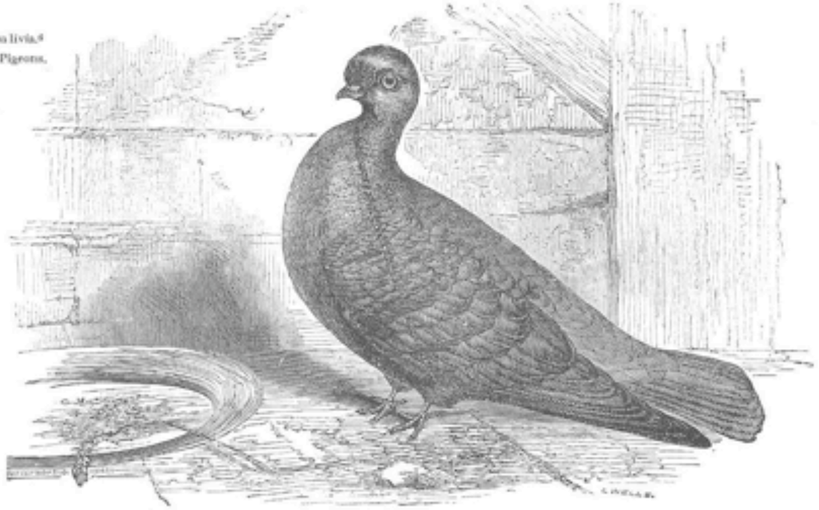


Fig. 23.—Short-faced English Tumbler.

Domesticated *Brassica oleracea* (Cabbages)

Cauliflower

Kohlrabi

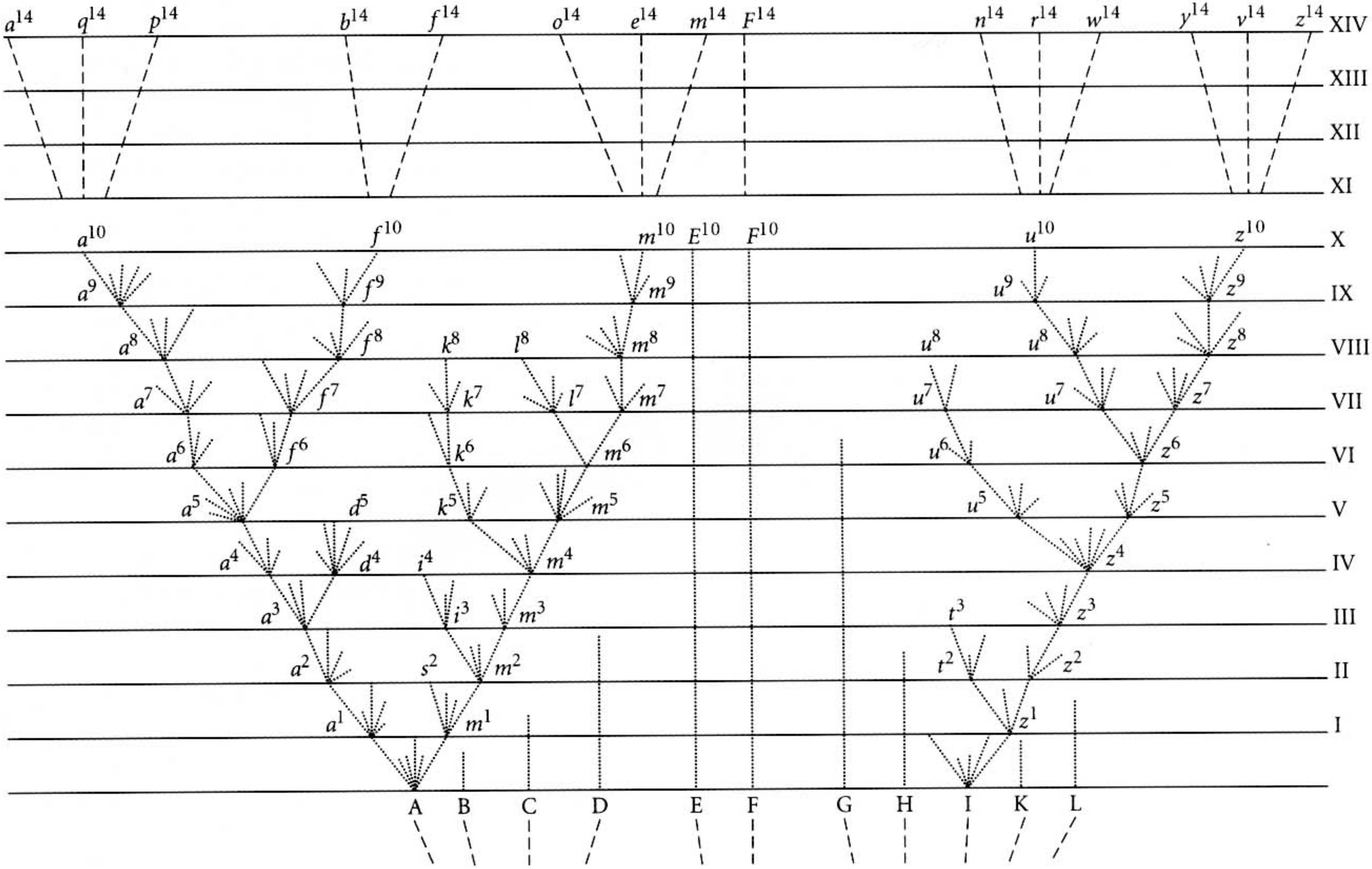
Broccoli

Turnips

Kale

Brussel Sprouts

The only figure in the Origin of Species



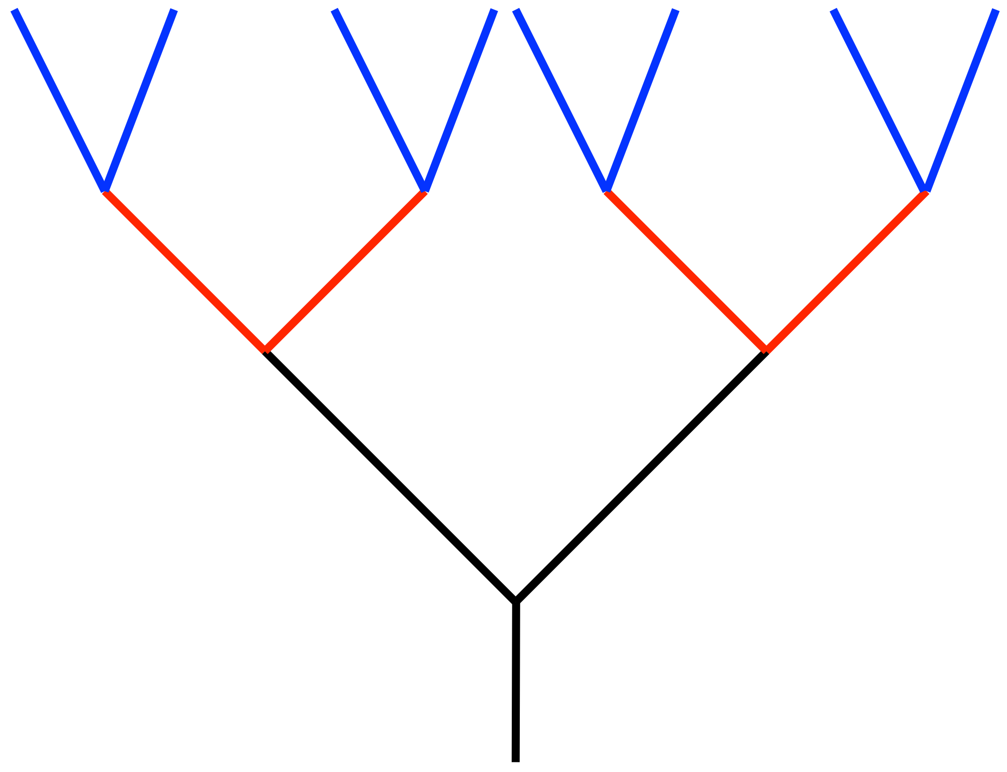
a^{14} q^{14} p^{14} b^{14} f^{14} o^{14} e^{14} m^{14} F^{14} n^{14} r^{14} w^{14} y^{14} v^{14} z^{14}
XIV
XIII
XII
XI
a^{10} f^{10} m^{10} E^{10} F^{10} u^{10} z^{10}
X
a^9 f^9 m^9 u^9 z^9
IX
a^8 f^8 k^8 l^8 m^8 u^8 u^8 z^8
VIII
a^7 f^7 k^7 l^7 m^7 u^7 u^7 z^7
VII
a^6 f^6 k^6 m^6 u^6 z^6
VI
a^5 d^5 k^5 m^5 u^5 z^5
V
a^4 d^4 i^4 m^4 z^4
IV
a^3 i^3 m^3 t^3 z^3
III
a^2 s^2 m^2 t^2 z^2
II
a^1 m^1 z^1
I
A B C D E F G H I K L

From Speciation to Phylogenies

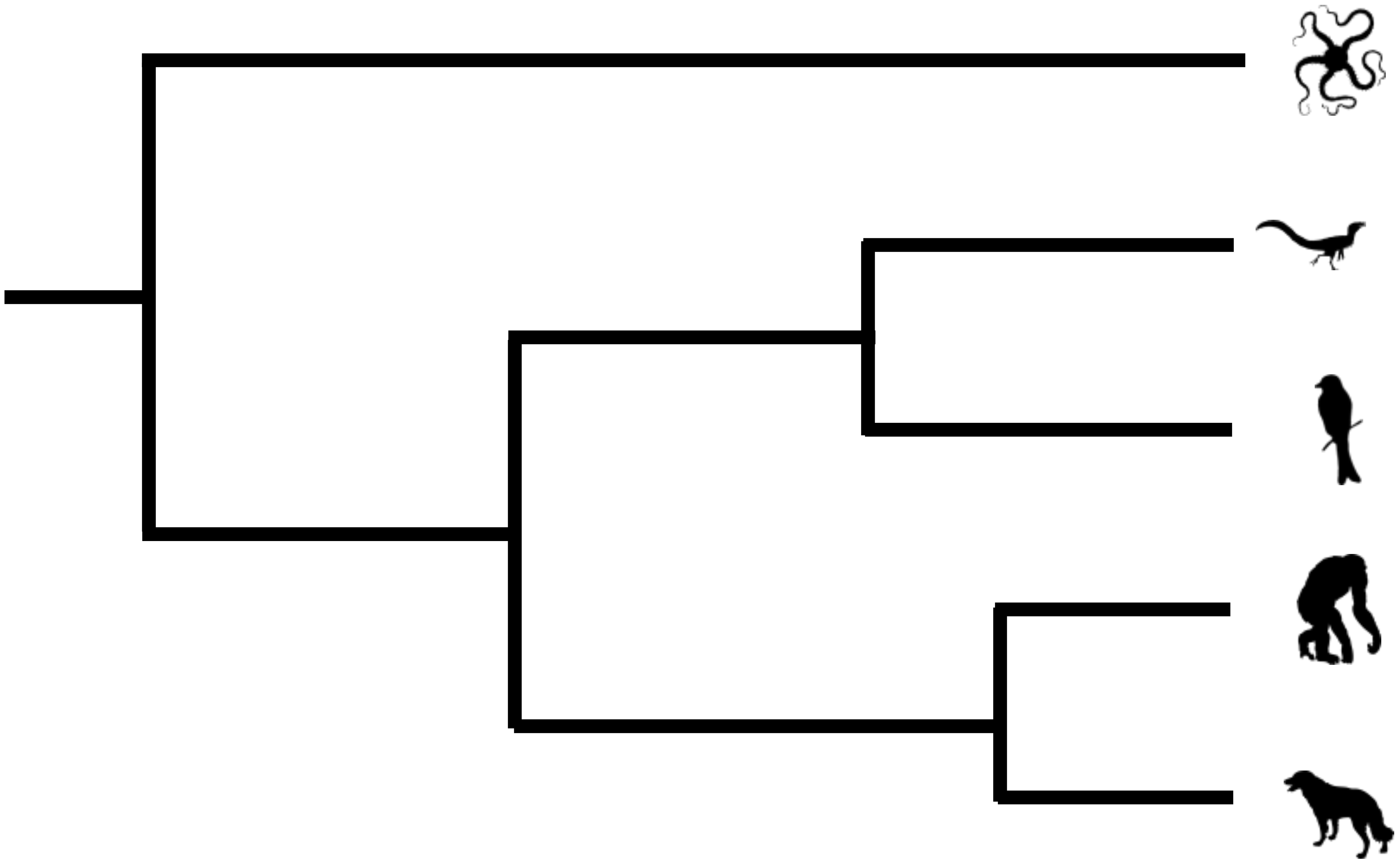
Many rounds of speciation define a unique phylogenetic tree of relationships

Sequential speciation creates a hierarchy

Speciation is lineage splitting

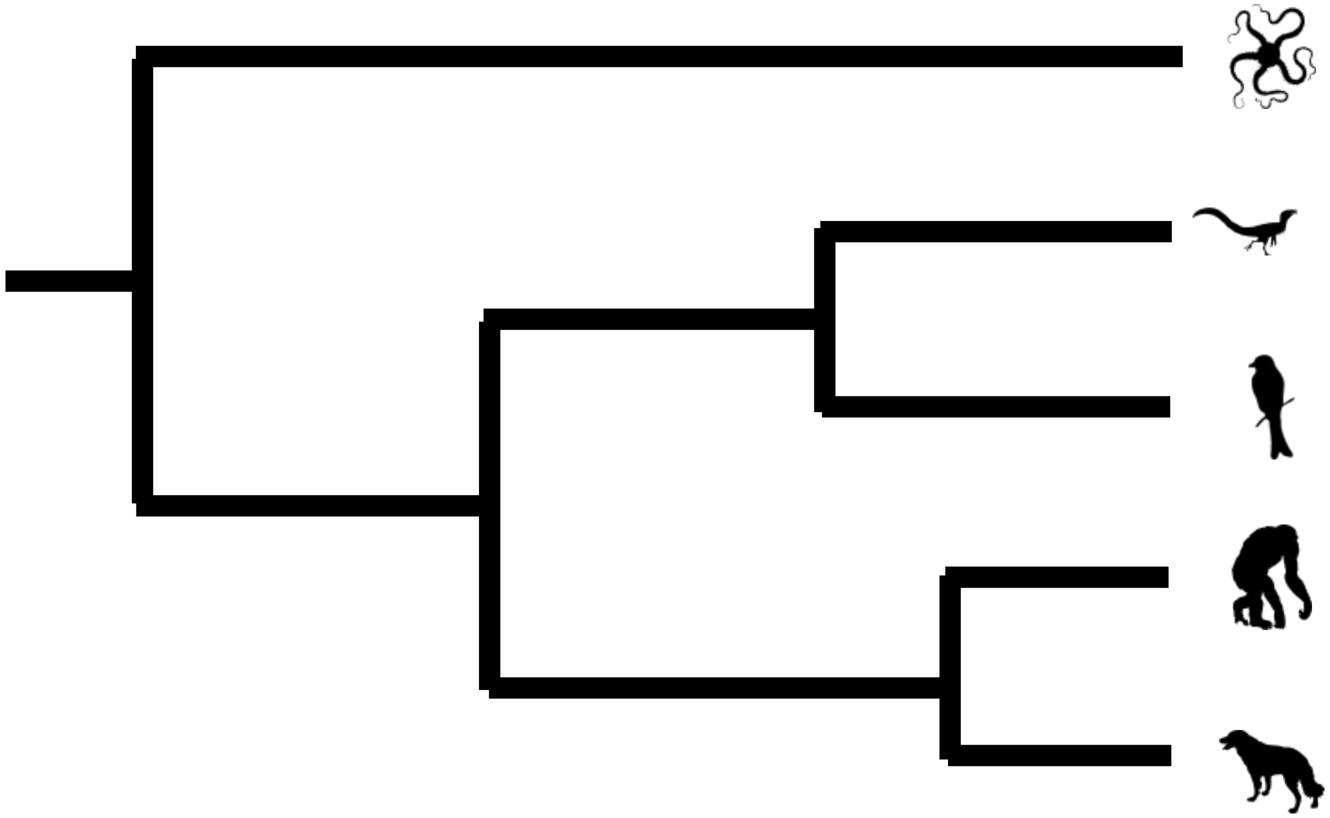


Phylogenies



a.k.a. evolutionary trees

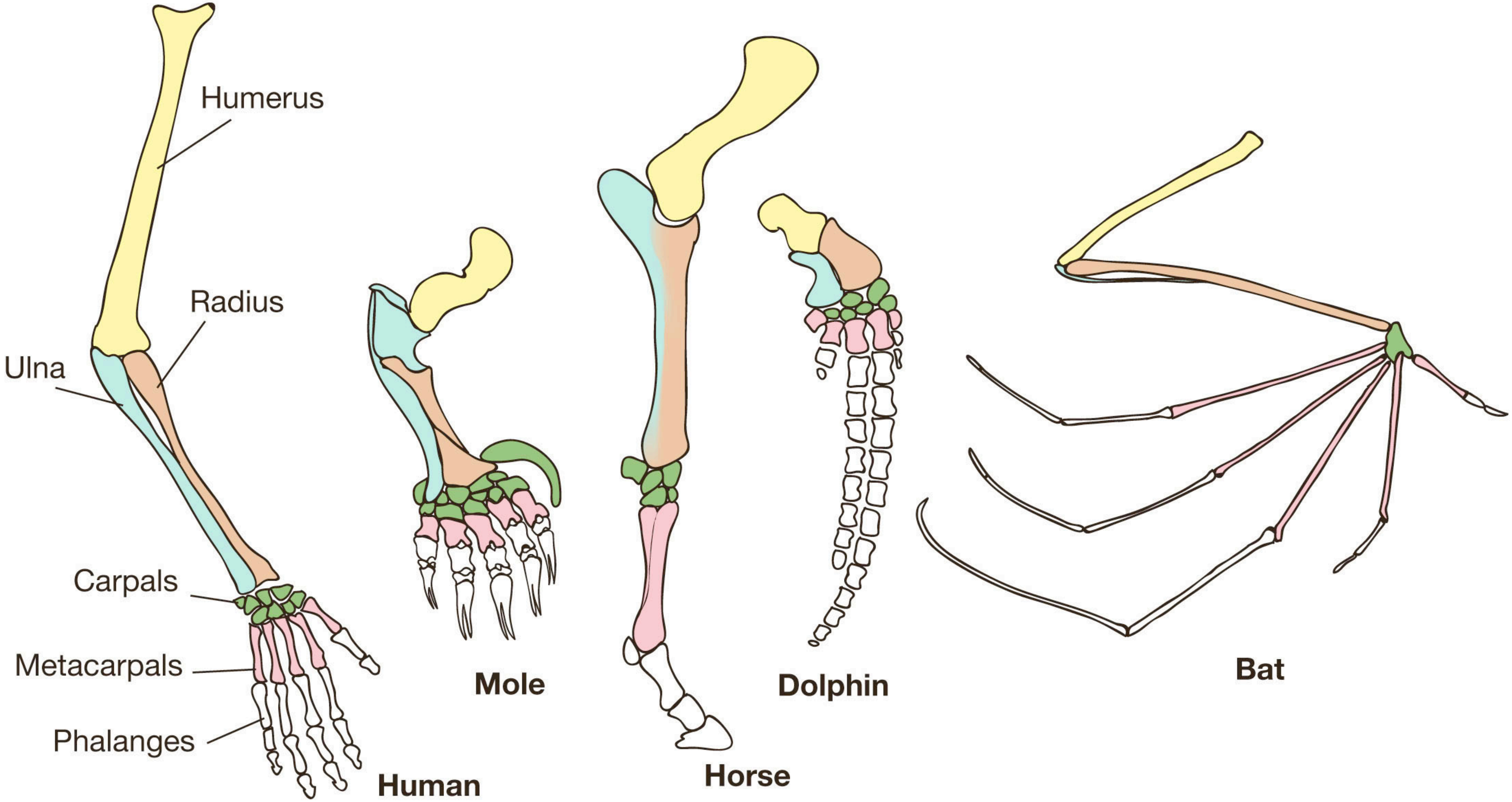
Phylogenies



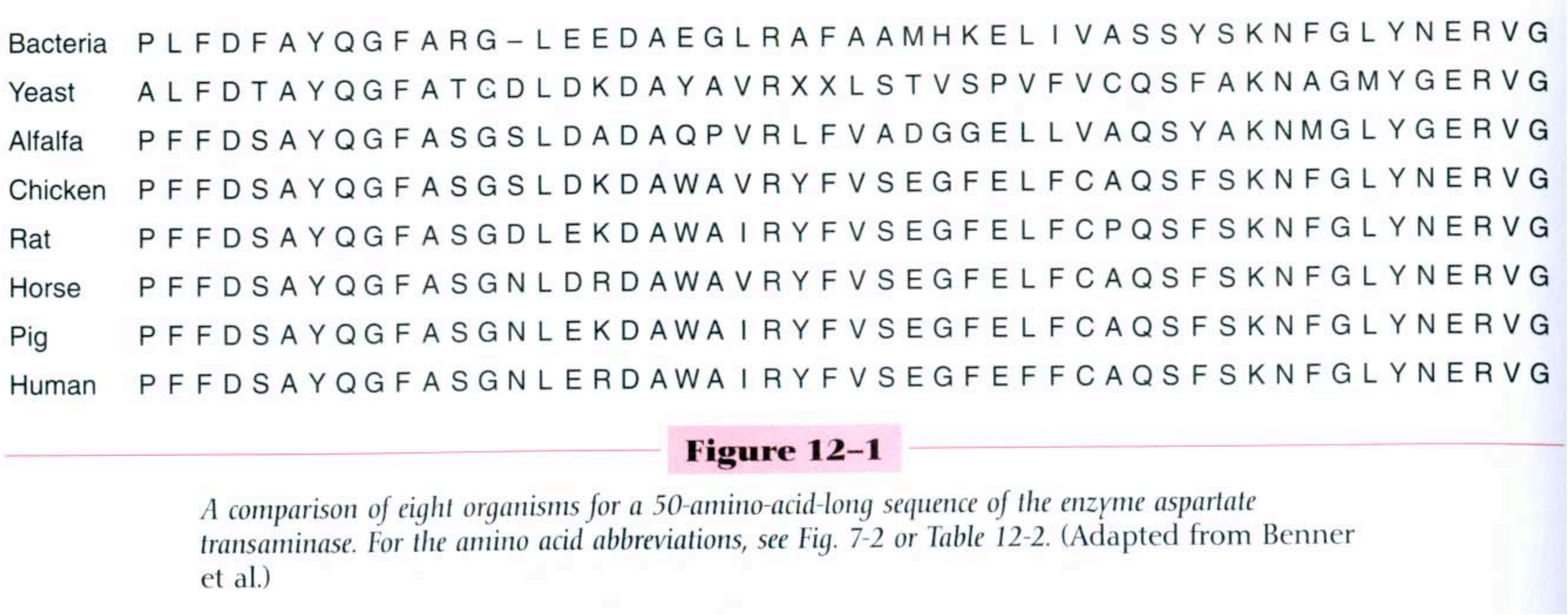
Genome Sequence

```
gtcgagttat cccaacctgg
agtatactga tagcttccac
caaccaggtg ctctactagg
gagtcctggg cacagcccta
atatggtgag ctcatacgat
```

Homology:
similarity of structure due to descent from a common ancestor



Molecular Homology of Amino Acid positions in Proteins



Bacteria	P L F D F A Y Q G F A R G – L E E D A E G L R A F A A M H K E L I V A S S Y S K N F G L Y N E R V G
Yeast	A L F D T A Y Q G F A T G D L D K D A Y A V R X X L S T V S P V F V C Q S F A K N A G M Y G E R V G
Alfalfa	P F F D S A Y Q G F A S G S L D A D A Q P V R L F V A D G G E L L V A Q S Y A K N M G L Y G E R V G
Chicken	P F F D S A Y Q G F A S G S L D K D A W A V R Y F V S E G F E L F C A Q S F S K N F G L Y N E R V G
Rat	P F F D S A Y Q G F A S G D L E K D A W A I R Y F V S E G F E L F C P Q S F S K N F G L Y N E R V G
Horse	P F F D S A Y Q G F A S G N L D R D A W A V R Y F V S E G F E L F C A Q S F S K N F G L Y N E R V G
Pig	P F F D S A Y Q G F A S G N L E K D A W A I R Y F V S E G F E L F C A Q S F S K N F G L Y N E R V G
Human	P F F D S A Y Q G F A S G N L E R D A W A I R Y F V S E G F E F F C A Q S F S K N F G L Y N E R V G

Figure 12–1

A comparison of eight organisms for a 50-amino-acid-long sequence of the enzyme aspartate transaminase. For the amino acid abbreviations, see Fig. 7-2 or Table 12-2. (Adapted from Benner et al.)

Deep homology of genetic code

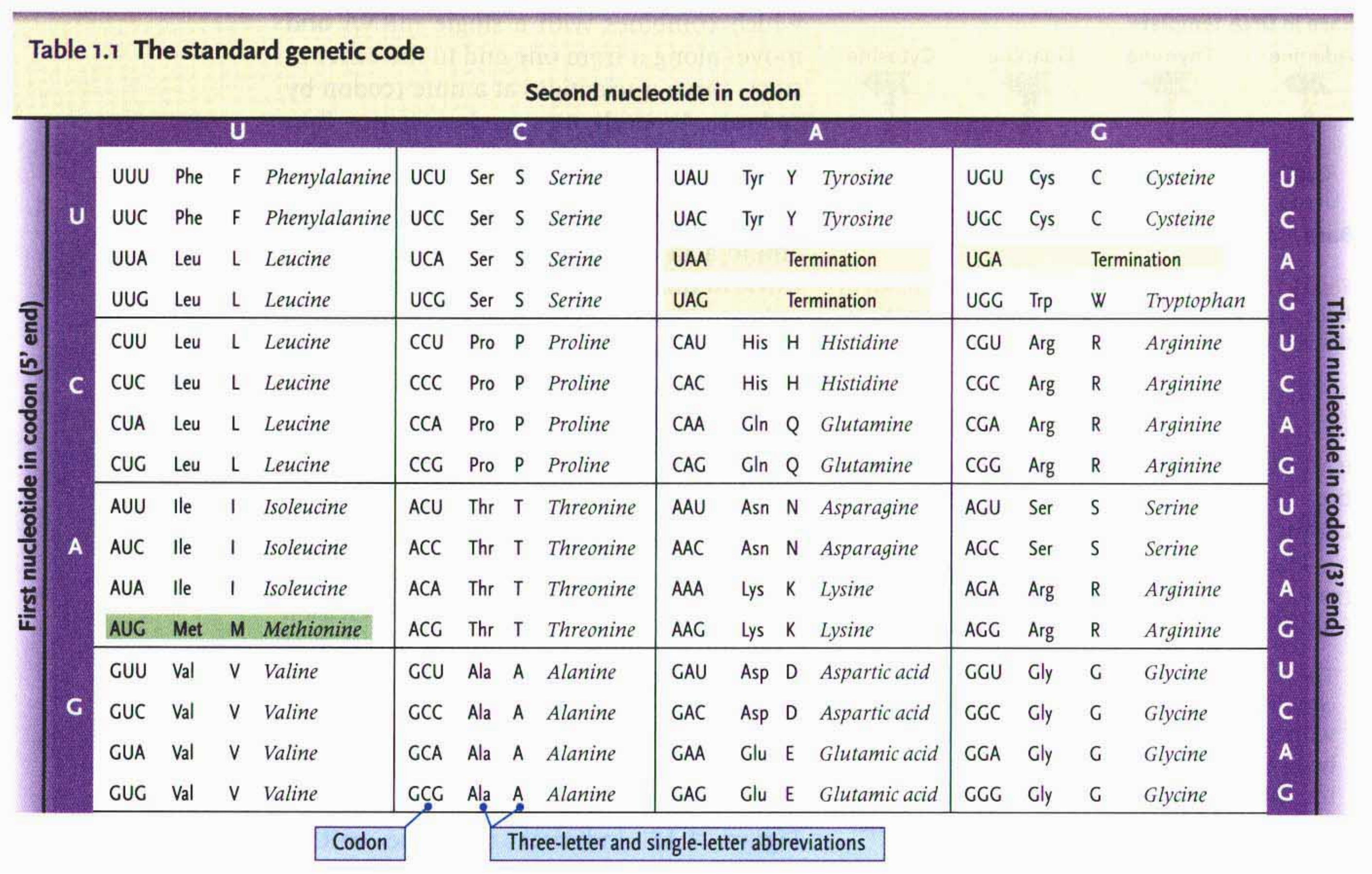


Table 1.1 The standard genetic code

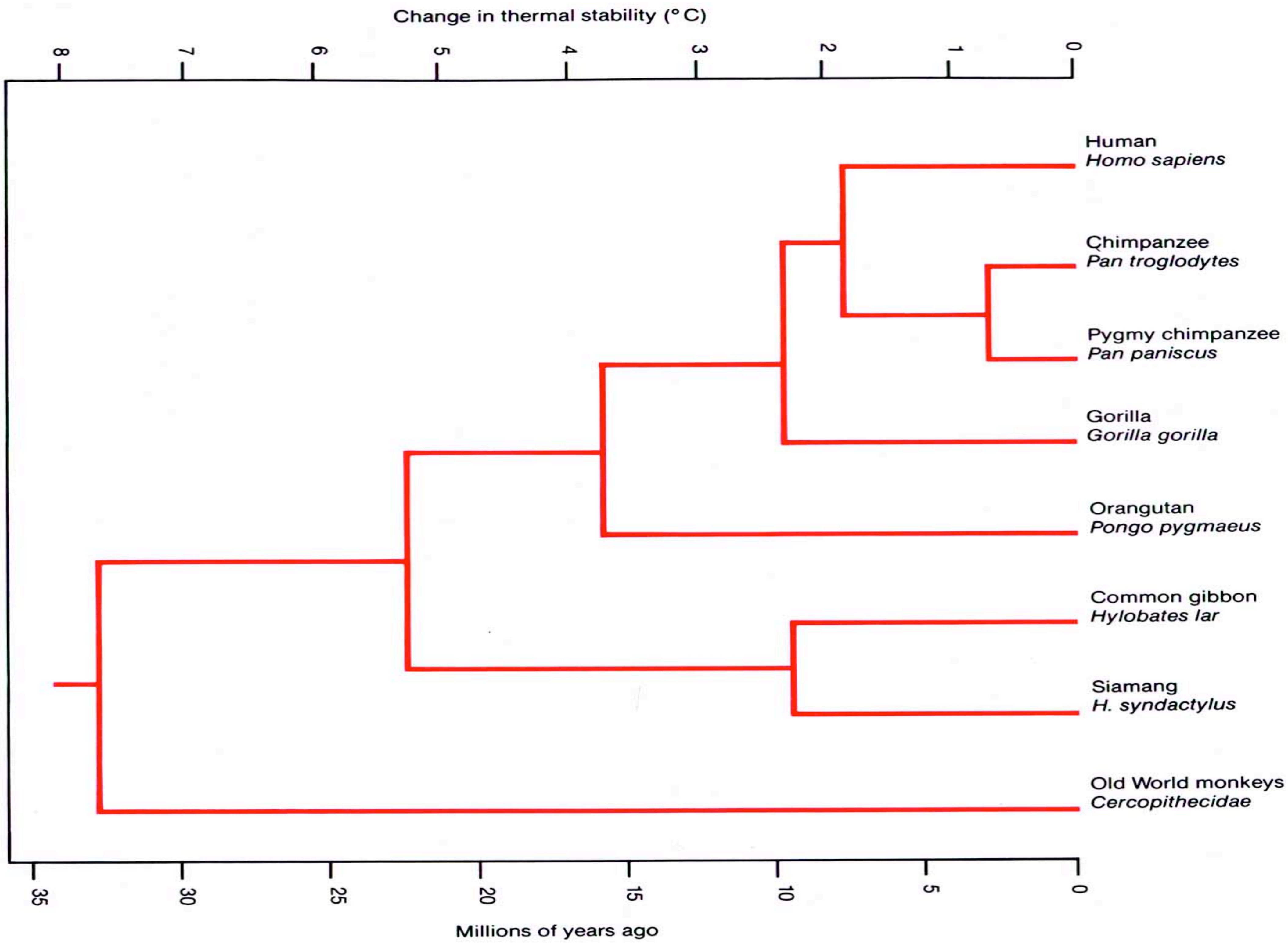
Second nucleotide in codon

First nucleotide in codon (5' end)	U				C				A				G				Third nucleotide in codon (3' end)
U	UUU	Phe	F	*Phenylalanine*	UCU	Ser	S	*Serine*	UAU	Tyr	Y	*Tyrosine*	UGU	Cys	C	*Cysteine*	U
	UUC	Phe	F	*Phenylalanine*	UCC	Ser	S	*Serine*	UAC	Tyr	Y	*Tyrosine*	UGC	Cys	C	*Cysteine*	C
	UUA	Leu	L	*Leucine*	UCA	Ser	S	*Serine*	UAA	Termination			UGA	Termination			A
	UUG	Leu	L	*Leucine*	UCG	Ser	S	*Serine*	UAG	Termination			UGG	Trp	W	*Tryptophan*	G
C	CUU	Leu	L	*Leucine*	CCU	Pro	P	*Proline*	CAU	His	H	*Histidine*	CGU	Arg	R	*Arginine*	U
	CUC	Leu	L	*Leucine*	CCC	Pro	P	*Proline*	CAC	His	H	*Histidine*	CGC	Arg	R	*Arginine*	C
	CUA	Leu	L	*Leucine*	CCA	Pro	P	*Proline*	CAA	Gln	Q	*Glutamine*	CGA	Arg	R	*Arginine*	A
	CUG	Leu	L	*Leucine*	CCG	Pro	P	*Proline*	CAG	Gln	Q	*Glutamine*	CGG	Arg	R	*Arginine*	G
A	AUU	Ile	I	*Isoleucine*	ACU	Thr	T	*Threonine*	AAU	Asn	N	*Asparagine*	AGU	Ser	S	*Serine*	U
	AUC	Ile	I	*Isoleucine*	ACC	Thr	T	*Threonine*	AAC	Asn	N	*Asparagine*	AGC	Ser	S	*Serine*	C
	AUA	Ile	I	*Isoleucine*	ACA	Thr	T	*Threonine*	AAA	Lys	K	*Lysine*	AGA	Arg	R	*Arginine*	A
	AUG	Met	M	*Methionine*	ACG	Thr	T	*Threonine*	AAG	Lys	K	*Lysine*	AGG	Arg	R	*Arginine*	G
G	GUU	Val	V	*Valine*	GCU	Ala	A	*Alanine*	GAU	Asp	D	*Aspartic acid*	GGU	Gly	G	*Glycine*	U
	GUC	Val	V	*Valine*	GCC	Ala	A	*Alanine*	GAC	Asp	D	*Aspartic acid*	GGC	Gly	G	*Glycine*	C
	GUA	Val	V	*Valine*	GCA	Ala	A	*Alanine*	GAA	Glu	E	*Glutamic acid*	GGA	Gly	G	*Glycine*	A
	GUG	Val	V	*Valine*	GCG	Ala	A	*Alanine*	GAG	Glu	E	*Glutamic acid*	GGG	Gly	G	*Glycine*	G

Codon

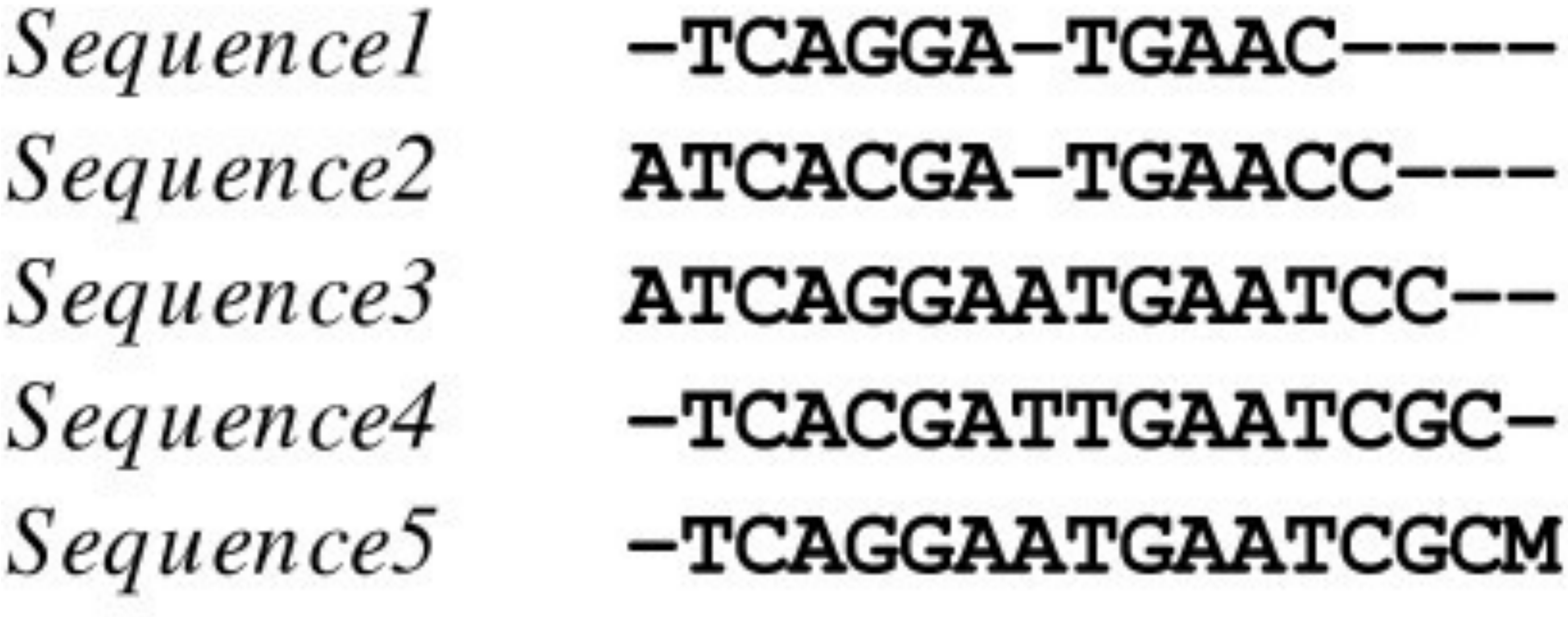
Three-letter and single-letter abbreviations

Phylogeny of Primates based on DNA-DNA Hybridization



Sibley and Ahlquist 1984. Mol. Evol. 20:2-15

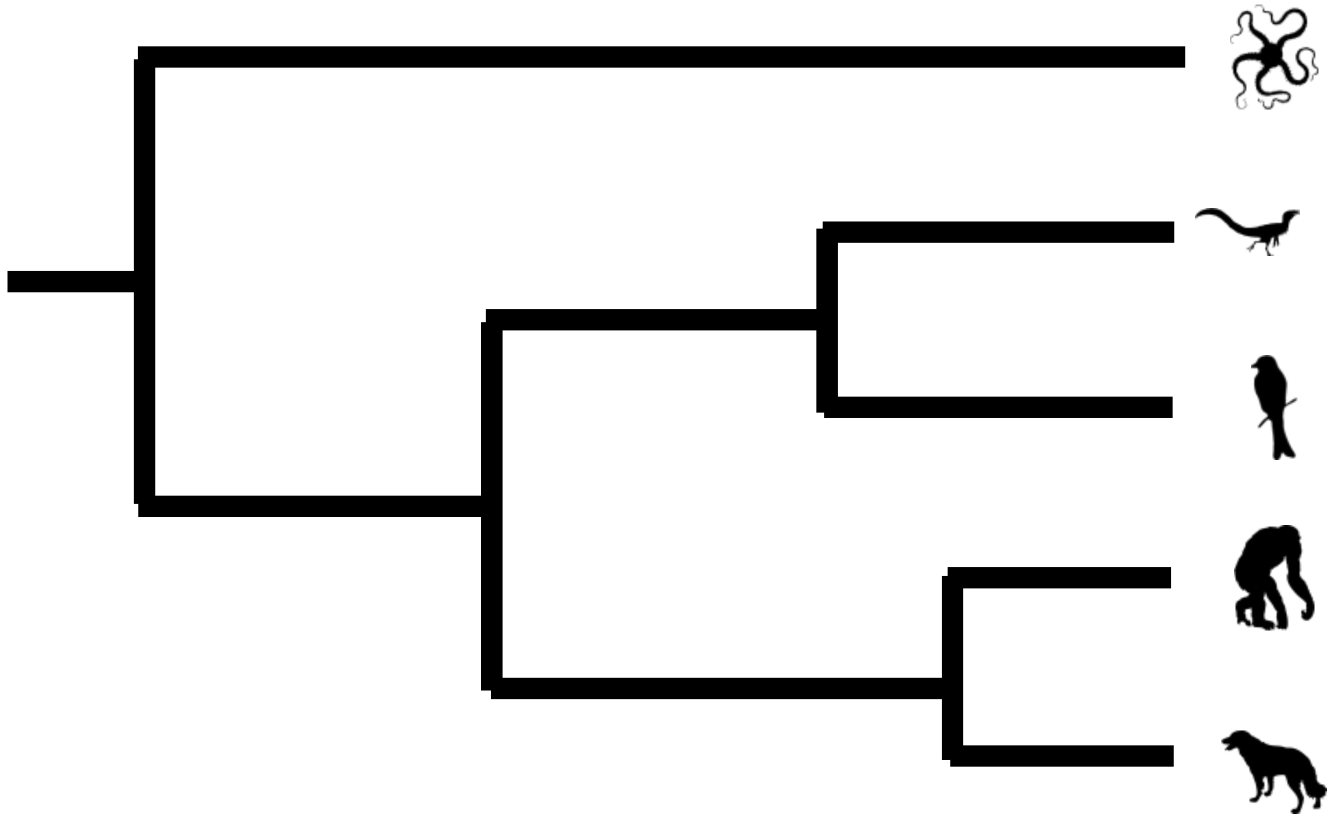
So want to focus on homologous traits / characters



```
Sequence1    -TCAGGA-TGAAC----
Sequence2    ATCACGA-TGAACC---
Sequence3    ATCAGGAATGAATCC--
Sequence4    -TCACGATTGAATCGC-
Sequence5    -TCAGGAATGAATCGCM
```

In genomes this means ALIGNMENT will be critical- which bases are the same bases?

Phylogenies

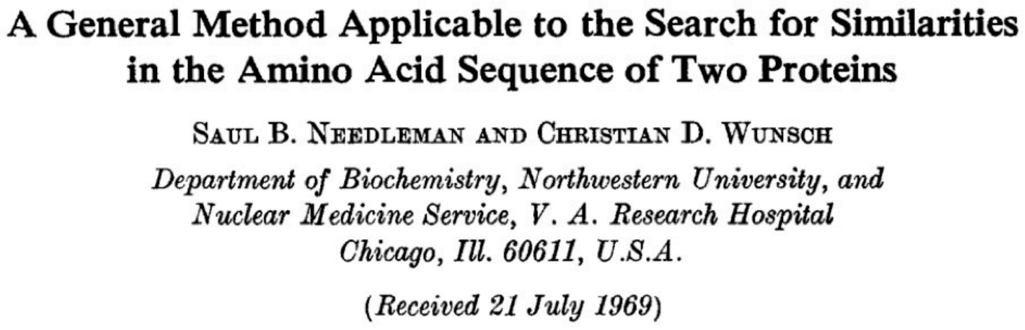


Genome Sequence

```
gtcgagttat cccaacctgg
agtatactga tagcttccac
caaccaggtg ctctactagg
gagtcctggg cacagcccta
atatggtgag ctcatacgat
```

Sequence Alignment

Key algorithms introduced starting in 1970

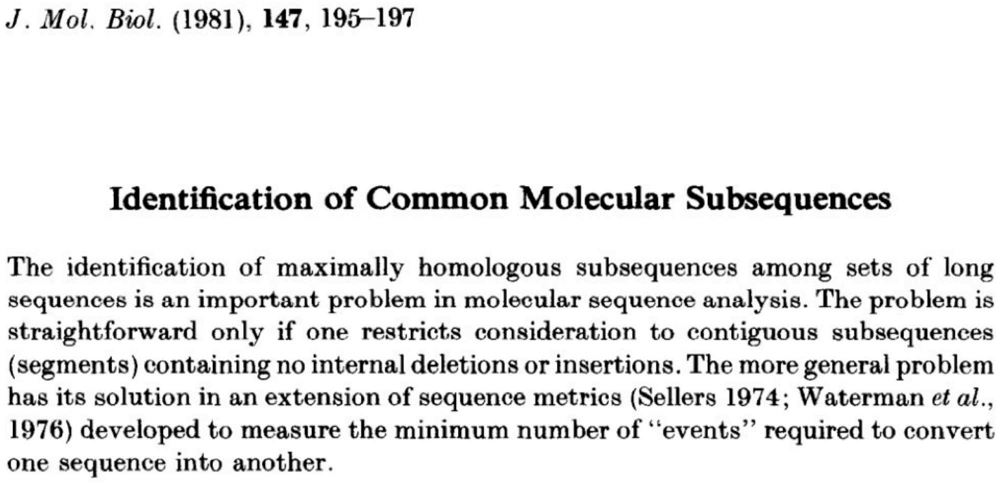


A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins

SAUL B. NEEDLEMAN AND CHRISTIAN D. WUNSCH

Department of Biochemistry, Northwestern University, and
Nuclear Medicine Service, V. A. Research Hospital
Chicago, Ill. 60611, U.S.A.

(*Received 21 July 1969*)



J. Mol. Biol. (1981), **147**, 195–197

Identification of Common Molecular Subsequences

The identification of maximally homologous subsequences among sets of long sequences is an important problem in molecular sequence analysis. The problem is straightforward only if one restricts consideration to contiguous subsequences (segments) containing no internal deletions or insertions. The more general problem has its solution in an extension of sequence metrics (Sellers 1974; Waterman *et al.*, 1976) developed to measure the minimum number of "events" required to convert one sequence into another.

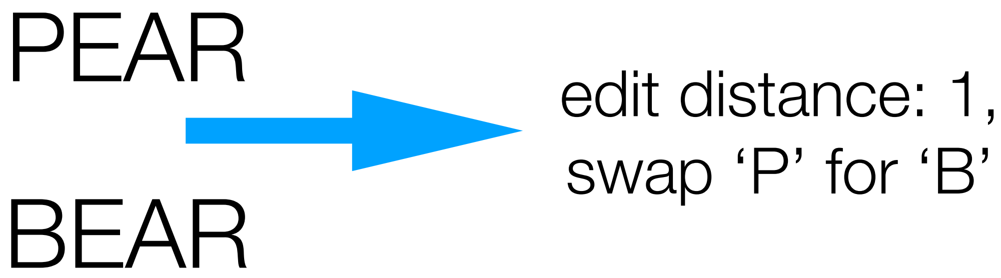
Global: Needleman-Wunsch algorithm

Local: Smith-Waterman algorithm

Sequence Alignment

key idea— edit distance

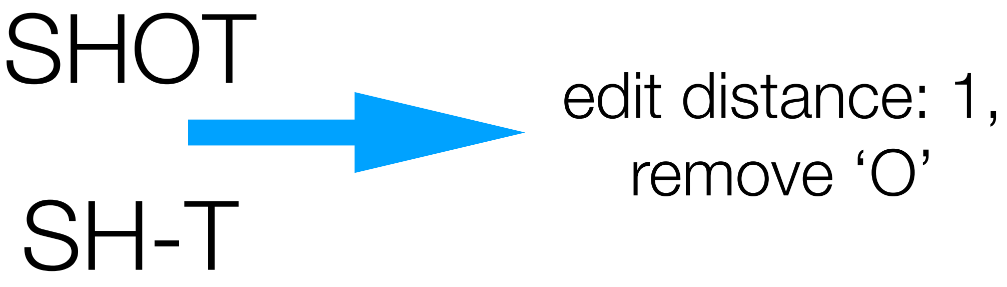
how many changes to make two words / sets the same?



PEAR

BEAR

edit distance: 1,
swap ‘P’ for ‘B’



SHOT

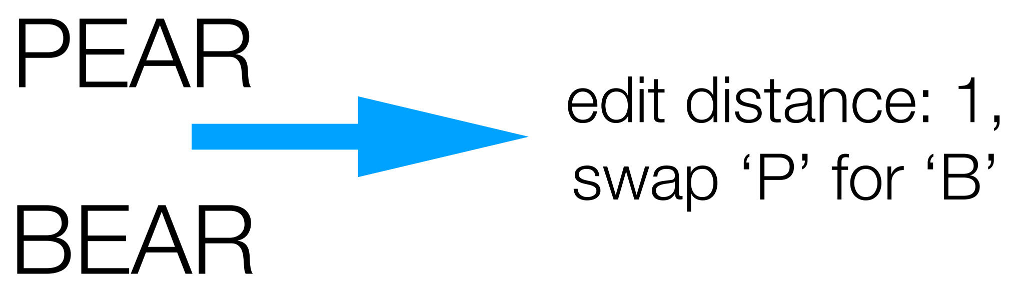
SH-T

edit distance: 1,
remove ‘O’

Sequence Alignment

key idea— edit distance

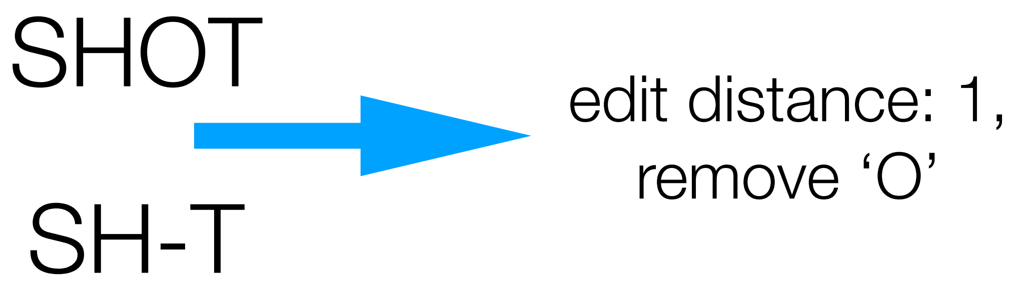
how many changes to make two words / sets the same?



PEAR

BEAR

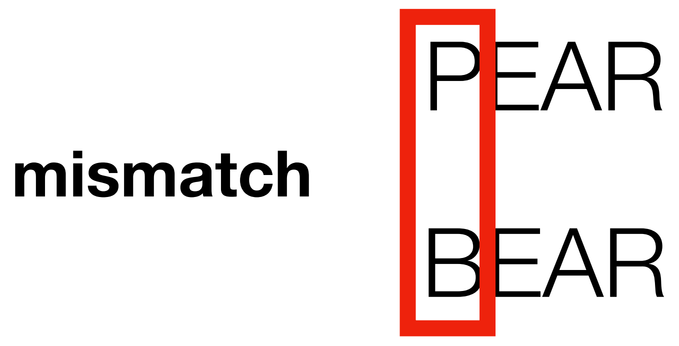
edit distance: 1,
swap ‘P’ for ‘B’



SHOT

SH-T

edit distance: 1,
remove ‘O’



mismatch

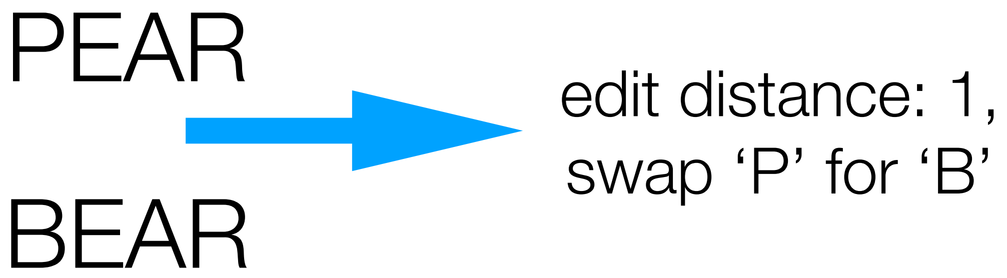
PEAR

BEAR

Sequence Alignment

key idea— edit distance

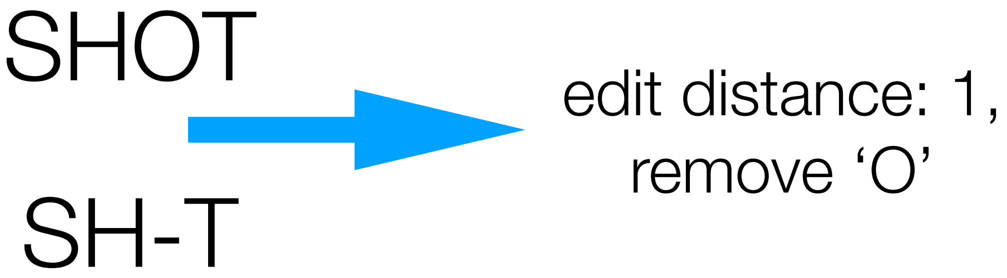
how many changes to make two words / sets the same?



PEAR

BEAR

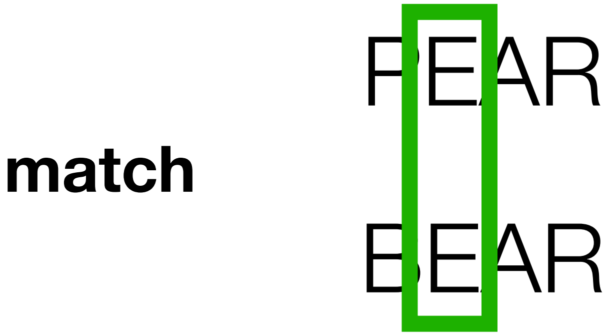
edit distance: 1,
swap ‘P’ for ‘B’



SHOT

SH-T

edit distance: 1,
remove ‘O’



match

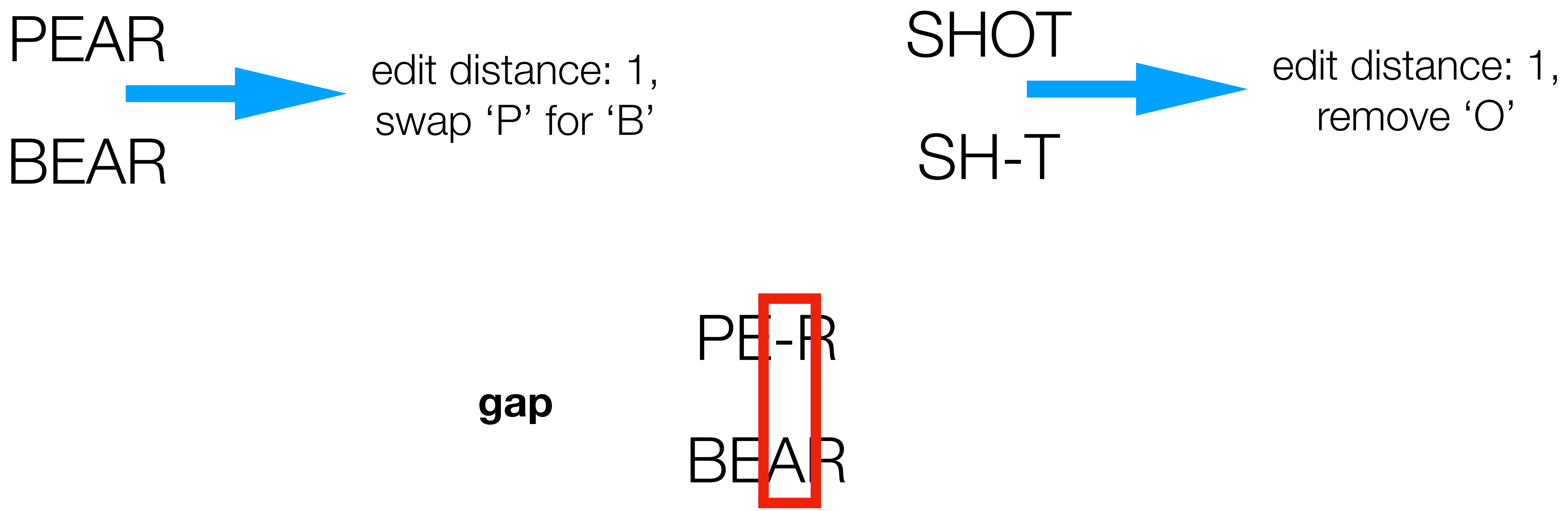
PEAR

BEAR

Sequence Alignment

key idea— edit distance

how many changes to make two words / sets the same?



PEAR

BEAR

edit distance: 1,
swap ‘P’ for ‘B’

SHOT

SH-T

edit distance: 1,
remove ‘O’

gap

PE-R

BEAR

Sequence Alignment

key idea— edit distance

how many changes to make two words / sets the same?

PEAR

BEAR

edit distance: 1,
swap ‘P’ for ‘B’

SHOT

SH-T

edit distance: 1,
remove ‘O’

What about DNA?

```
TGTTACGG
GGTTGACTA
```

```
TG-TT-ACGG
-GGTTGACTA
```

edit distance: 5

```
TGTT-ACGG
GGTTGACTA
```

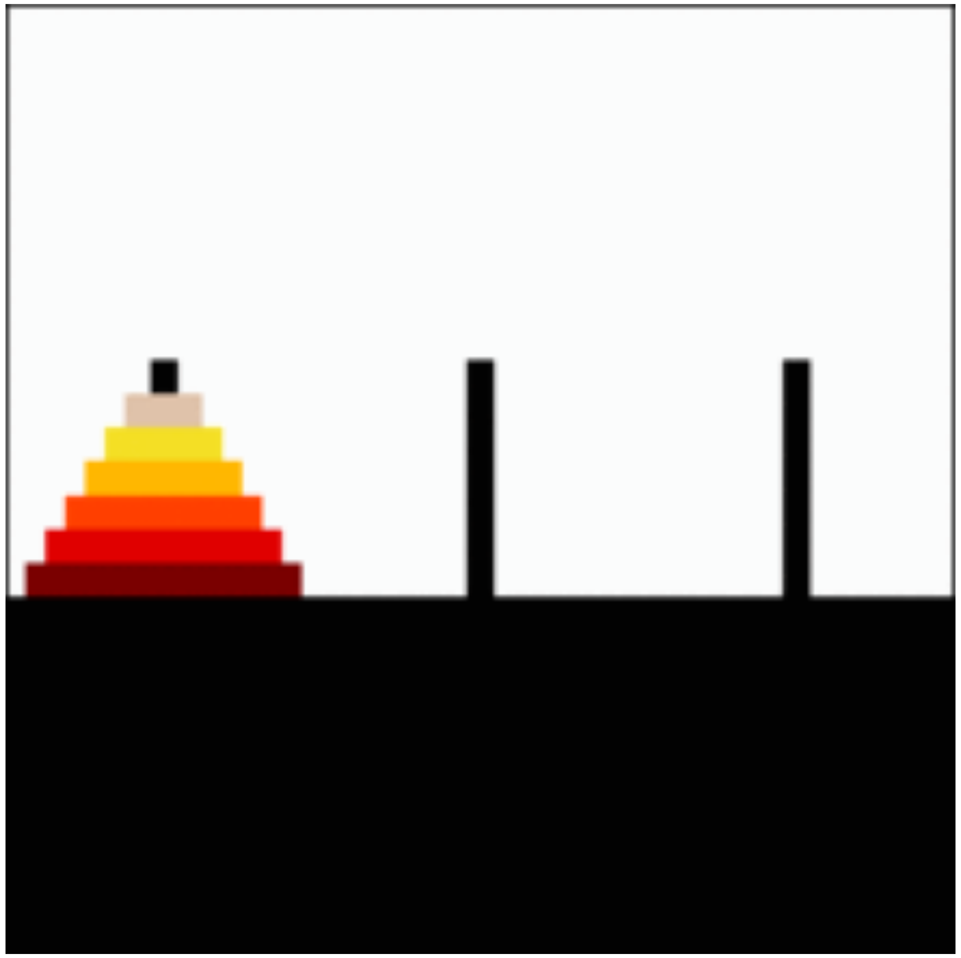
edit distance: 4

Sequence Alignment

Let’s align the these two sequences

Dynamic Programming

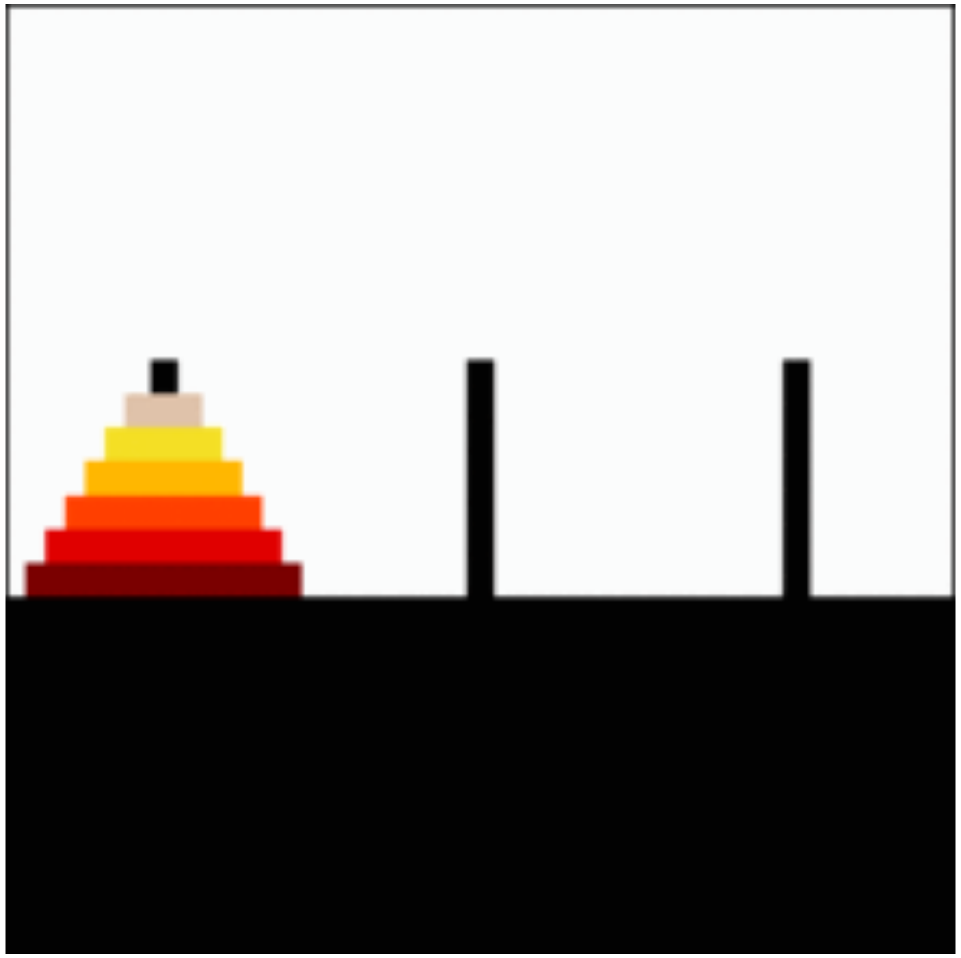
Tower of Hanoi Game



Goal: Get all disks in size order to last peg
Rules: One disk moves per turn. Smaller has to sit on larger

Dynamic Programming

Tower of Hanoi Game



Can break up **big** problem into simple **smaller** problem

Dynamic Programming

Fibonacci sequence

Originally meant to model the growth in rabbit population sizes!

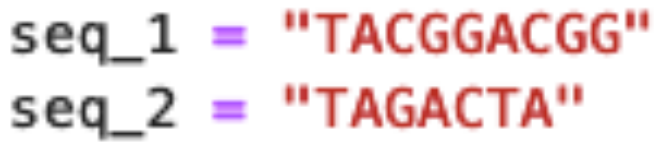
$$F_0 = 1; F_1 = 1$$

$$F_i = F_{i-1} + F_{i-2}$$

F(0)	F(1)	F(2)	F(3)	F(4)	F(5)	F(6)
1	1	2	3	5	8	13

Sequence Alignment

So want to find minimum edit distance
how do we search for it?



```
seq_1 = "TACGGACGG"
seq_2 = "TAGACTA"
```

Next key idea

treat sequences as matrix

		T	A	C	G	G	A	C	G	G
T										
A										
G										
A										
C										
T										
A										

Sequence Alignment

Global Alignment
Needleman-Wunsch

Assume scores as:

match = 4
mismatch = -3
gap = -2

1. initialize

Seq 1

Seq 2

		T	A	C	G	G	A	C	G	G
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
T	-2									
A	-4									
G	-6									
A	-8									
C	-10									
T	-12									
A	-14									

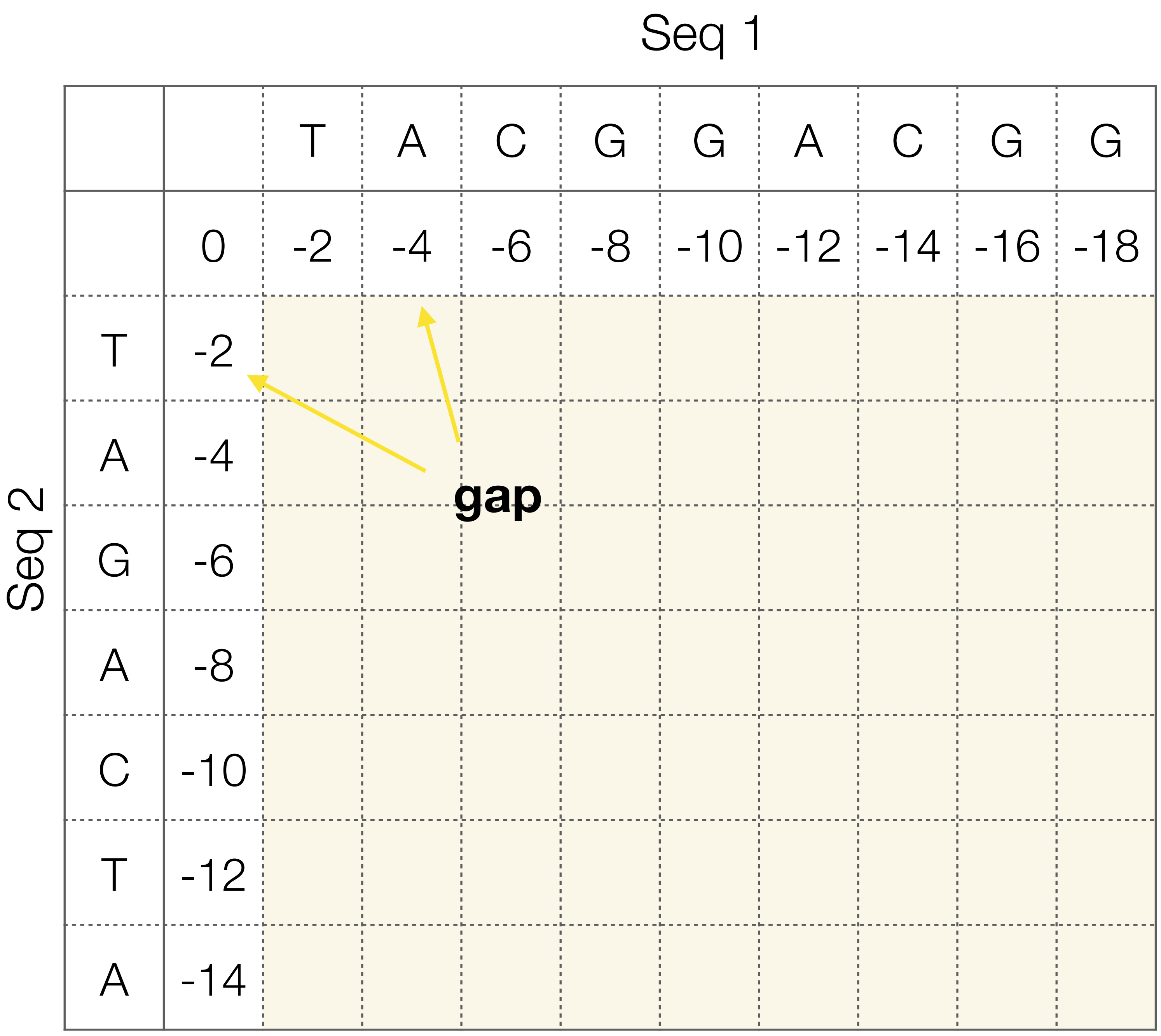
Sequence Alignment

Global Alignment
Needleman-Wunsch

Assume scores as:

match = 4
mismatch = -3
gap = -2

1. initialize



Seq 1

Seq 2

		T	A	C	G	G	A	C	G	G
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
T	-2									
A	-4									
G	-6									
A	-8									
C	-10									
T	-12									
A	-14									

Sequence Alignment

Global Alignment
Needleman-Wunsch

Assume scores as:

match = 4
mismatch = -3
gap = -2

1. initialize
2. fill in table

consider value of x

3 ways to get there:

1) from previous (mis)match
2) gap in seq1
3) gap in seq2

Seq 1

Seq 2

		T	A	C	G	G	A	C	G	G
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
T	-2	x								
A	-4									
G	-6									
A	-8									
C	-10									
T	-12									
A	-14									

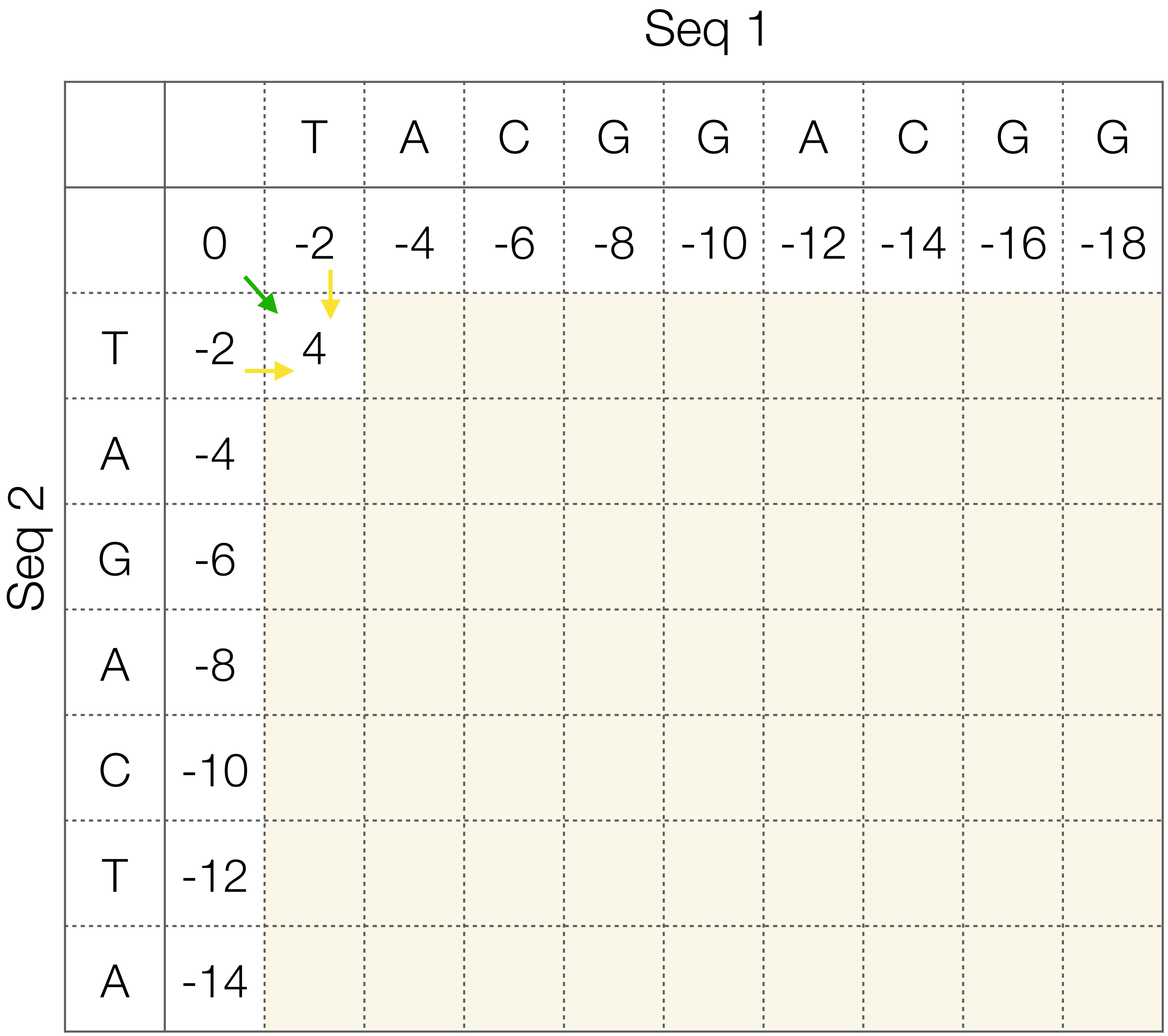
Sequence Alignment

Global Alignment
Needleman-Wunsch

Assume scores as:

match = 4
mismatch = -3
gap = -2

x_i,j = max(x_i-1,j + gap,
x_i,j-1 + gap,
x_i-1,j-1 + (mis)match)



Seq 1

Seq 2

		T	A	C	G	G	A	C	G	G
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
T	-2	4								
A	-4									
G	-6									
A	-8									
C	-10									
T	-12									
A	-14									

Sequence Alignment

Global Alignment
Needleman-Wunsch

Assume scores as:

match = 4
mismatch = -3
gap = -2

x_i,j = max(x_i-1,j + gap,
x_i,j-1 + gap,
x_i-1,j-1 + (mis)match)

Seq 1

Seq 2

		T	A	C	G	G	A	C	G	G
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
T	-2	4	2							
A	-4									
G	-6									
A	-8									
C	-10									
T	-12									
A	-14									

Sequence Alignment

Global Alignment
Needleman-Wunsch

Assume scores as:

match = 4
mismatch = -3
gap = -2

x_i,j = max(x_i-1,j + gap,
x_i,j-1 + gap,
x_i-1,j-1 + (mis)match)

Seq 1

Seq 2

		T	A	C	G	G	A	C	G	G
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
T	-2	4	2	0	-2	-4				
A	-4	2	8	6	4	2				
G	-6	0	6	5	10	8				
A	-8									
C	-10									
T	-12									
A	-14									

Sequence Alignment

Global Alignment
Needleman-Wunsch

Assume scores as:

match = 4
mismatch = -3
gap = -2

x_i,j = max(x_i-1,j + gap,
x_i,j-1 + gap,
x_i-1,j-1 + (mis)match)

1. initialize
2. fill in table
3. traceback

Seq 1

Seq 2

		T	A	C	G	G	A	C	G	G
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
T	-2	4	2	0	-2	-4	-6	-8	-10	-12
A	-4	2	8	6	4	2	0	-2	-4	-6
G	-6	0	6	5	10	8	6	4	2	0
A	-8	-2	4	3	8	7	12	10	8	6
C	-10	-4	2	8	6	5	10	16	14	12
T	-12	-6	0	6	5	3	8	14	13	11
A	-14	-8	-2	4	3	2	7	12	11	10

Sequence Alignment

Global Alignment
Needleman-Wunsch

traceback step

start at bottom right
work way back up
follow max score path

can read off alignment
backwards

Seq 1	**G**
Seq 2	**A**

Seq 1

Seq 2

		T	A	C	G	G	A	C	G	G
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
T	-2	4	2	0	-2	-4	-6	-8	-10	-12
A	-4	2	8	6	4	2	0	-2	-4	-6
G	-6	0	6	5	10	8	6	4	2	0
A	-8	-2	4	3	8	7	12	10	8	6
C	-10	-4	2	8	6	5	10	16	14	12
T	-12	-6	0	6	5	3	8	14	13	11
A	-14	-8	-2	4	3	2	7	12	11	10

Sequence Alignment

Global Alignment
Needleman-Wunsch

traceback step

start at bottom right
work way back up
follow max score path

can read off alignment
backwards

Seq 1 **GG**
Seq 2 **TA**

Seq 1

Seq 2

		T	A	C	G	G	A	C	G	G
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
T	-2	4	2	0	-2	-4	-6	-8	-10	-12
A	-4	2	8	6	4	2	0	-2	-4	-6
G	-6	0	6	5	10	8	6	4	2	0
A	-8	-2	4	3	8	7	12	10	8	6
C	-10	-4	2	8	6	5	10	16	14	12
T	-12	-6	0	6	5	3	8	14	13	11
A	-14	-8	-2	4	3	2	7	12	11	10

Sequence Alignment

Global Alignment
Needleman-Wunsch

traceback step

start at bottom right
work way back up
follow max score path

can read off alignment
backwards

Seq 1 **TACGGACGG**
Seq 2 **TA--GACTA**

Seq 1

Seq 2

		T	A	C	G	G	A	C	G	G
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
T	-2	4	2	0	-2	-4	-6	-8	-10	-12
A	-4	2	8	6	4	2	0	-2	-4	-6
G	-6	0	6	5	10	8	6	4	2	0
A	-8	-2	4	3	8	7	12	10	8	6
C	-10	-4	2	8	6	5	10	16	14	12
T	-12	-6	0	6	5	3	8	14	13	11
A	-14	-8	-2	4	3	2	7	12	11	10

Sequence Alignment

Local Alignment
Smith-Waterman

No negative scores

traceback starts at highest score

		T	A	C	G	G	A	C	G	G
	0	0	0	0	0	0	0	0	0	0
T	0	4	2	0	0	0	0	0	0	0
A	0	2	8	6	4	2	0	0	0	0
G	0	0	6	5	10	8	6	4	2	0
A	0	0	4	3	8	7	12	10	8	6
C	0	0	2	8	6	5	10	16	14	12
T	0	0	0	6	5	3	8	14	13	11
A	0	0	0	4	3	2	7	12	11	10

Sequence Alignment

Local Alignment
Smith-Waterman

No negative scores

traceback starts at highest score

terminates at first zero

```
TACGGAC
TA__GAC
```

		T	A	C	G	G	A	C	G	G
	0	0	0	0	0	0	0	0	0	0
T	0	4	2	0	0	0	0	0	0	0
A	0	2	8	6	4	2	0	0	0	0
G	0	0	6	5	10	8	6	4	2	0
A	0	0	4	3	8	7	12	10	8	6
C	0	0	2	8	6	5	10	16	14	12
T	0	0	0	6	5	3	8	14	13	11
A	0	0	0	4	3	2	7	12	11	10

Sequence Alignment

```
Seq1  TACGGACGG
Seq2  TAGACTA
```

Local Alignment
Smith-Waterman

```
TACGGAC
TA__GAC
```

Global Alignment
Needleman-Wunsch

```
TACGGACGG
TA__GACTA
```

in this case very similar, but not generally true
use local alignment when you want short, subset matches
global when whole sequence alignment wanted