CS 68: BIOINFORMATICS

Prof. Sara Mathieson
Swarthmore College
Spring 2018

Outline: Apr 9

- Recap parameter estimation for HMM
- Baum-Welch algorithm (EM for HMM)
- Motivation for Principal Components Analysis
- Overview of human evolution

- Office hours TODAY 3-5pm
- Check-in for Lab 8 this week in lab
- Email me if you would like to meet for your project
- Optional PCA lab released on Thursday

Recap: Parameter Estimation for HMMs

HMM informal quiz: discuss with a partner

1) The value of x_i only depends on _____.

2) The value of z_i only depends on _____.

3) What symbol would you use to denote the transition probability from state 2 to state 0?

4) What symbol would you use to denote the probability of emitting a 1 from state 3?

5) What symbol would you use to denote the probability of starting in state 2?

6) What is the runtime of the Viterbi, forward, backward algorithms?

7) What is the runtime of obtaining the posterior decoding and posterior mean?

HMM informal quiz: discuss with a partner

1) The value of x_i only depends on $\boxed{z_i}$.

2) The value of z_i only depends on _____.

3) What symbol would you use to denote the transition probability from state 2 to state 0?

4) What symbol would you use to denote the probability of emitting a 1 from state 3?

5) What symbol would you use to denote the probability of starting in state 2?

6) What is the runtime of the Viterbi, forward, backward algorithms?

7) What is the runtime of obtaining the posterior decoding and posterior mean?

HMM informal quiz: discuss with a partner

1) The value of x_i only depends on $\boxed{z_i}$.
2) The value of z_i only depends on $\boxed{z_{i-1}}$.
3) What symbol would you use to denote the transition probability from state 2 to state 0?
4) What symbol would you use to denote the probability of emitting a 1 from state 3?
5) What symbol would you use to denote the probability of starting in state 2?
6) What is the runtime of the Viterbi, forward, backward algorithms?
7) What is the runtime of obtaining the posterior decoding and posterior mean?

HMM informal quiz: discuss with a partner

1) The value of x_i only depends on $\boxed{z_i}$.
2) The value of z_i only depends on $\boxed{z_{i-1}}$.
3) What symbol would you use to denote the transition probability from state 2 to state 0? $\boxed{a_{2,0}}$
4) What symbol would you use to denote the probability of emitting a 1 from state 3?
5) What symbol would you use to denote the probability of starting in state 2?
6) What is the runtime of the Viterbi, forward, backward algorithms?
7) What is the runtime of obtaining the posterior decoding and posterior mean?

HMM informal quiz: discuss with a partner

1) The value of x_i only depends on $\boxed{z_i}$.
2) The value of z_i only depends on $\boxed{z_{i-1}}$.
3) What symbol would you use to denote the transition probability from state 2 to state 0? $\boxed{a_{2,0}}$
4) What symbol would you use to denote the probability of emitting a 1 from state 3? $\boxed{e_3(1)}$
5) What symbol would you use to denote the probability of starting in state 2?
6) What is the runtime of the Viterbi, forward, backward algorithms?
7) What is the runtime of obtaining the posterior decoding and posterior mean?

HMM informal quiz: discuss with a partner

1) The value of x_i only depends on $\boxed{z_i}$.

2) The value of z_i only depends on $\boxed{z_{i-1}}$.

3) What symbol would you use to denote the transition probability from state 2 to state 0? $\boxed{a_{2,0}}$

4) What symbol would you use to denote the probability of emitting a 1 from state 3? $\boxed{e_3(1)}$

5) What symbol would you use to denote the probability of starting in state 2? $\boxed{\pi_2}$

6) What is the runtime of the Viterbi, forward, backward algorithms?

7) What is the runtime of obtaining the posterior decoding and posterior mean?

HMM informal quiz: discuss with a partner

1) The value of x_i only depends on $\boxed{z_i}$.

2) The value of z_i only depends on $\boxed{z_{i-1}}$.

3) What symbol would you use to denote the transition probability from state 2 to state 0? $\boxed{a_{2,0}}$

4) What symbol would you use to denote the probability of emitting a 1 from state 3? $\boxed{e_3(1)}$

5) What symbol would you use to denote the probability of starting in state 2? $\boxed{\pi_2}$

6) What is the runtime of the Viterbi, forward, backward algorithms?

$O(K^2L)$, where K=number of hidden states, L=length of sequence

7) What is the runtime of obtaining the posterior decoding and posterior mean?

HMM informal quiz: discuss with a partner

1) The value of x_i only depends on z_i.

2) The value of z_i only depends on z_{i-1}.

3) What symbol would you use to denote the transition probability from state 2 to state 0? $a_{2,0}$

4) What symbol would you use to denote the probability of emitting a 1 from state 3? $e_3(1)$

5) What symbol would you use to denote the probability of starting in state 2? π_2

6) What is the runtime of the Viterbi, forward, backward algorithms?

$O(K^2L)$, where K=number of hidden states, L=length of sequence

7) What is the runtime of obtaining the posterior decoding and posterior mean?

$O(KL)$, need to compute posterior probability for all K states, along the sequence

$$P(z_i = k|\vec{x}) = \frac{f_k(i) \cdot b_k(i)}{P(\vec{x})}$$

Backward Algorithm Recap

$$b_k(i) = P(x_{i+1} \cdots x_L \mid z_i = k)$$

initialization

$$b_k(L) = 1$$

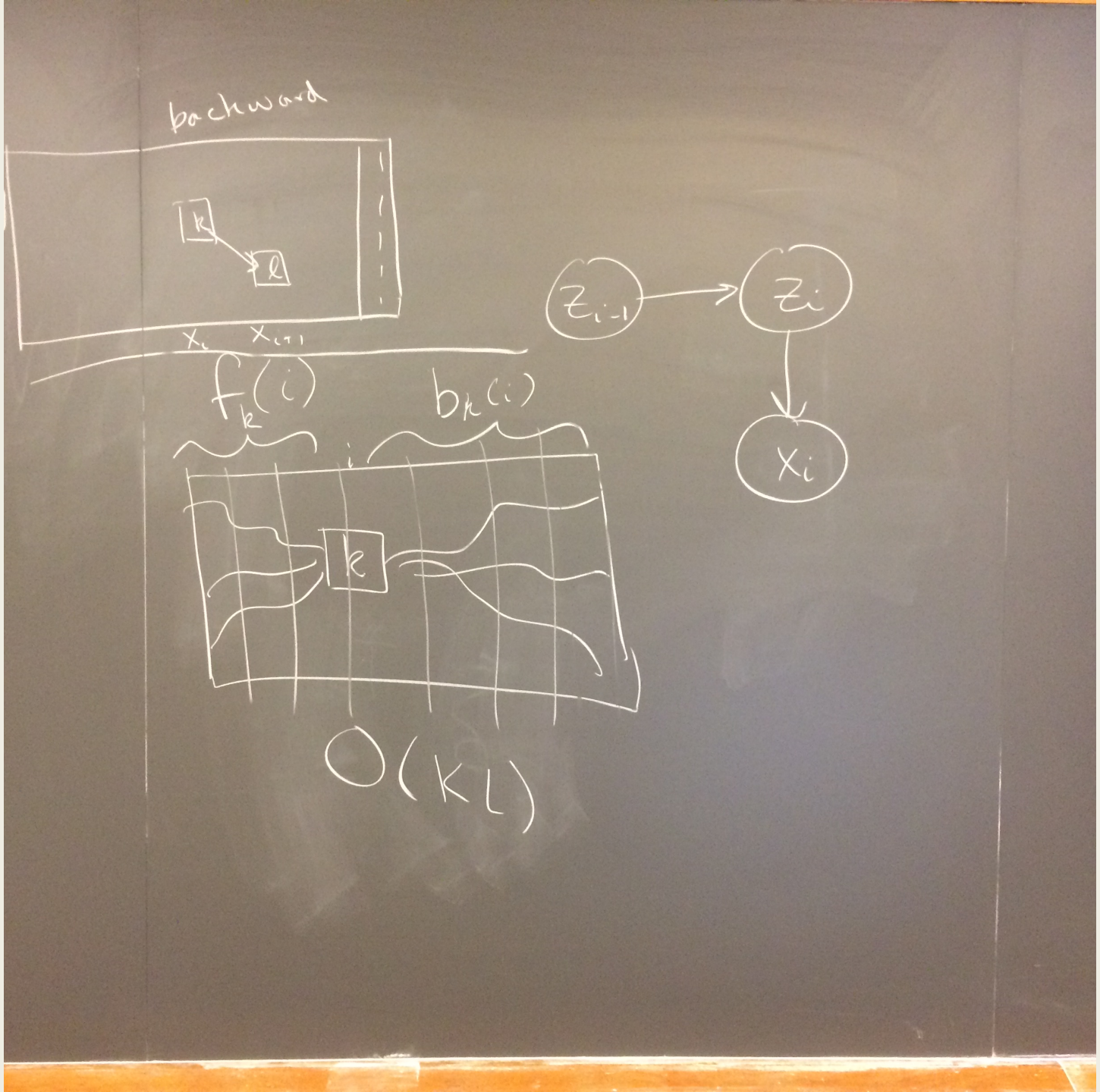
already saw x_L & the rest of $\vec{x}$

recursion

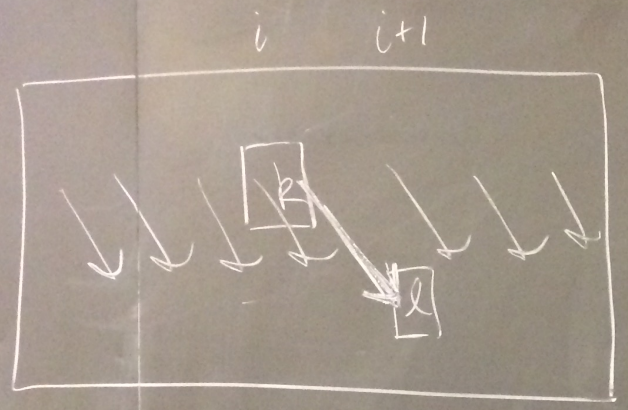
$$b_k(i) = \sum_{\ell} a_{k\ell} e_\ell(x_{i+1}) b_\ell(i+1)$$

order

$i = L-1, L-2, \ldots, 2, 1$

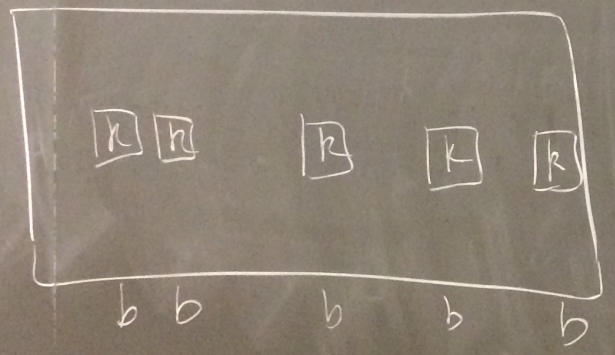

backward
k
l
x_i x_{i+1}
$f_k(i)$
$b_k(i)$
i
k
z_{i-1}
z_i
x_i
$O(KL)$

$A_{kl} =$



$$P(B|A) = \frac{P(B, A)}{P(A)}$$

$E_k(b) =$



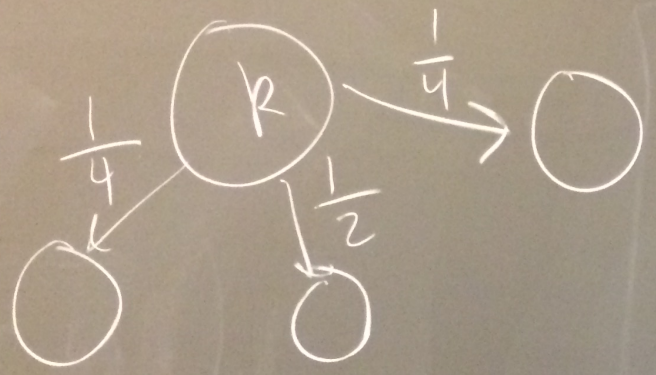
$$\boxed{P(\vec{x}) = \sum_k f_k(L)}$$

likelihood

want high!

end of seq

Markov Chain



outgoing → sum to 1

Parameter estimation for HMMs

$$A_{kl} = \#\text{ expected number of transitions between state } k \text{ and } l$$

$$E_k(b) = \#\text{ expected number of emissions of } b \text{ from state } k$$

$$a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}}$$

$$e_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b')}$$

Parameter estimation for HMMs

$$A_{kl} = \#\text{ expected number of transitions between state } k \text{ and } l$$

$$E_k(b) = \#\text{ expected number of emissions of } b \text{ from state } k$$

- For either Case 1 (state sequence known) or Case 2 (state sequence unknown), we need a way of computing the expected number of times we use a specific transition or emission
- For Case 1, we can observe these counts directly
- For Case 2, we will obtain these expected counts by adding up the probability of using each transition and emission across the entire sequence

$$a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}}$$

$$e_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b')}$$

What about the start state?

1) Count how many times we observe each state throughout the sequence, then normalize

2) See how many times (or how many expected times) we actually start in each state (this is assuming many independent sequences)

Case 2: state sequence unknown

- Compute the expected transition and emission counts using the forward and backward probabilities:

$$P(z_i = k, z_{i+1} = l | \vec{x}) = \frac{f_k(i) \cdot a_{kl} \cdot e_l(x_{i+1}) \cdot b_l(i+1)}{P(\vec{x})}$$

$$A_{kl} = \sum_{i=1}^{L-1} P(z_i = k, z_{i+1} = l | \vec{x})$$

Expected transition counts

Case 2: state sequence unknown

- Compute the expected transition and emission counts using the forward and backward probabilities:

$$P(z_i = k, z_{i+1} = l | \vec{x}) = \frac{f_k(i) \cdot a_{kl} \cdot e_l(x_{i+1}) \cdot b_l(i+1)}{P(\vec{x})}$$

$$A_{kl} = \sum_{i=1}^{L-1} P(z_i = k, z_{i+1} = l | \vec{x})$$

Expected transition counts

Expected emission counts

$$E_k(b) = \sum_{\{i | x_i = b\}} \frac{f_k(i) \cdot b_k(i)}{P(\vec{x})}$$

Case 2: state sequence unknown

- Compute the expected transition and emission counts using the forward and backward probabilities:

$$P(z_i = k, z_{i+1} = l|\vec{x}) = \frac{f_k(i) \cdot a_{kl} \cdot e_l(x_{i+1}) \cdot b_l(i+1)}{P(\vec{x})}$$

$$A_{kl} = \sum_{i=1}^{L-1} P(z_i = k, z_{i+1} = l|\vec{x})$$

Expected transition counts

Expected emission counts

$$E_k(b) = \sum_{\{i|x_i=b\}} \frac{f_k(i) \cdot b_k(i)}{P(\vec{x})}$$

- We still need to normalize as before:

$$a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}}$$

$$e_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b')}$$

Baum-Welch algorithm

Baum-Welch Algorithm (1960's)

Initialization

Choose model parameters (***a*** & ***e*** matrices, $\boldsymbol{\pi}$ vector) arbitrarily

If we have some prior knowledge of the problem, use that to initialize parameters

Baum-Welch Algorithm (1960's)

Initialization

Choose model parameters (***a*** & ***e*** matrices, $\boldsymbol{\pi}$ vector) arbitrarily

If we have some prior knowledge of the problem, use that to initialize parameters

Iteration

E-step

- Run forward algorithm to get $f_k(i)$ p. 59
- Run backward algorithm to get $b_k(i)$ p. 60
- Compute expected transition/emission counts A_{kl}, $E_k(b)$ Equations 3.20 & 3.21

Baum-Welch Algorithm (1960's)

Initialization

Choose model parameters (***a*** & ***e*** matrices, $\boldsymbol{\pi}$ vector) arbitrarily

If we have some prior knowledge of the problem, use that to initialize parameters

Iteration

E-step

- Run forward algorithm to get $f_k(i)$ p. 59
- Run backward algorithm to get $b_k(i)$ p. 60
- Compute expected transition/emission counts A_{kl}, $E_k(b)$ Equations 3.20 & 3.21

M-step

- Maximize likelihood $P(\boldsymbol{x})$, given expected counts
- Transition and emission: →

$$a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}} \qquad e_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b')}$$

Equation 3.18

- Initial probability update for Lab 8:

$$\pi_k = \frac{f_k(1) \cdot b_k(1)}{P(\vec{x})}$$

Baum-Welch Algorithm (1960's)

Initialization

Choose model parameters (***a*** & ***e*** matrices, $\boldsymbol{\pi}$ vector) arbitrarily

If we have some prior knowledge of the problem, use that to initialize parameters

Iteration

E-step

- Run forward algorithm to get $f_k(i)$ p. 59
- Run backward algorithm to get $b_k(i)$ p. 60
- Compute expected transition/emission counts A_{kl}, $E_k(b)$ Equations 3.20 & 3.21

M-step

- Maximize likelihood $P(\boldsymbol{x})$, given expected counts
- Transition and emission: →

$$a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}} \qquad e_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b')}$$

Equation 3.18

- Initial probability update for Lab 8:

$$\pi_k = \frac{f_k(1) \cdot b_k(1)}{P(\vec{x})}$$

Termination

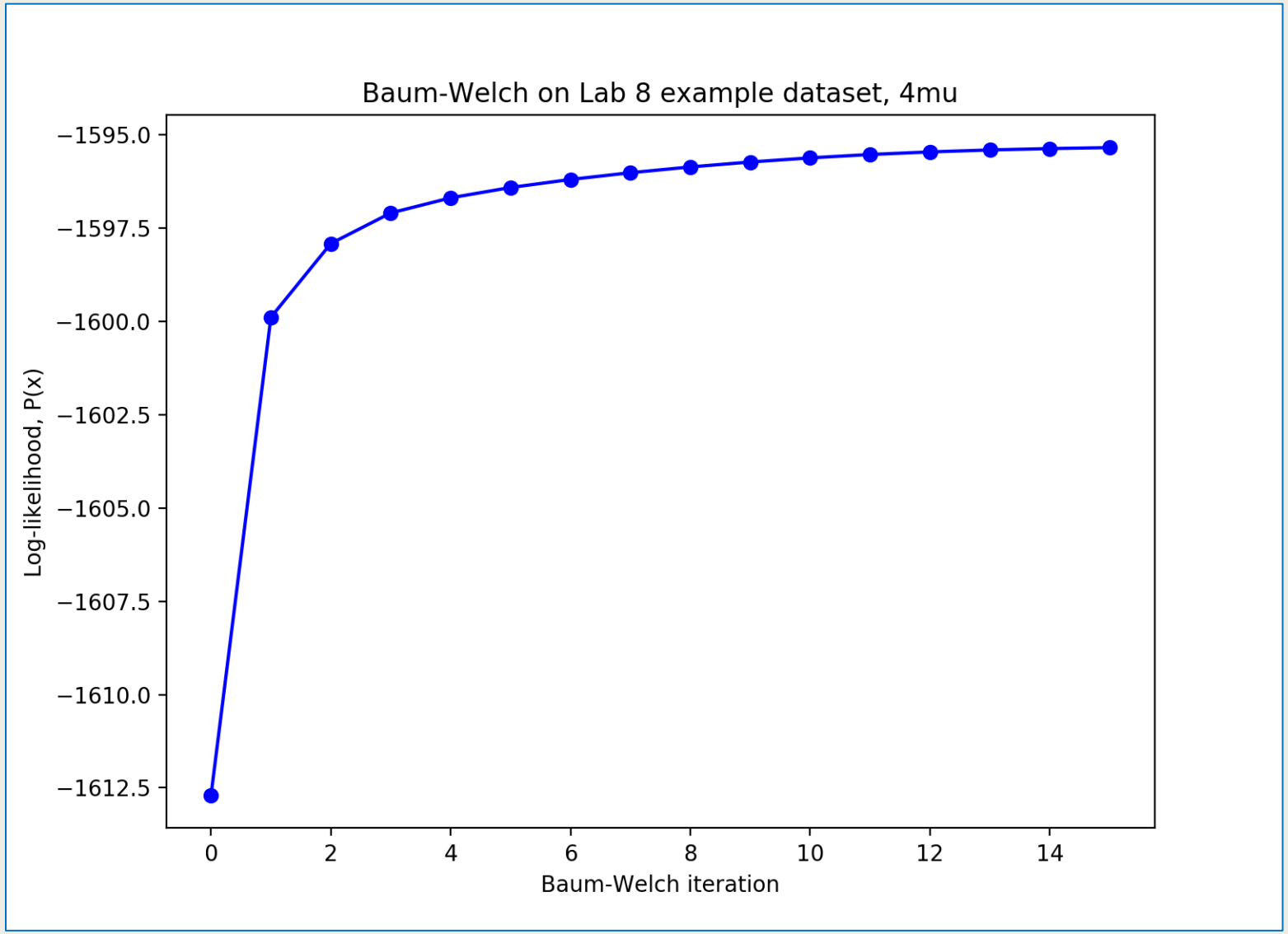
Stop when:

- Likelihood does not improve by much (choose threshold)
- Maximum number of iterations exceeded 15 iterations for us

Note: in slides: i=1,2,...L

Python: i=0,1,...(L-1)

Log-likelihood improves with each Baum-Welch iteration



HMM examples

HMMs in practice

Example 1: part-of-speech tagging

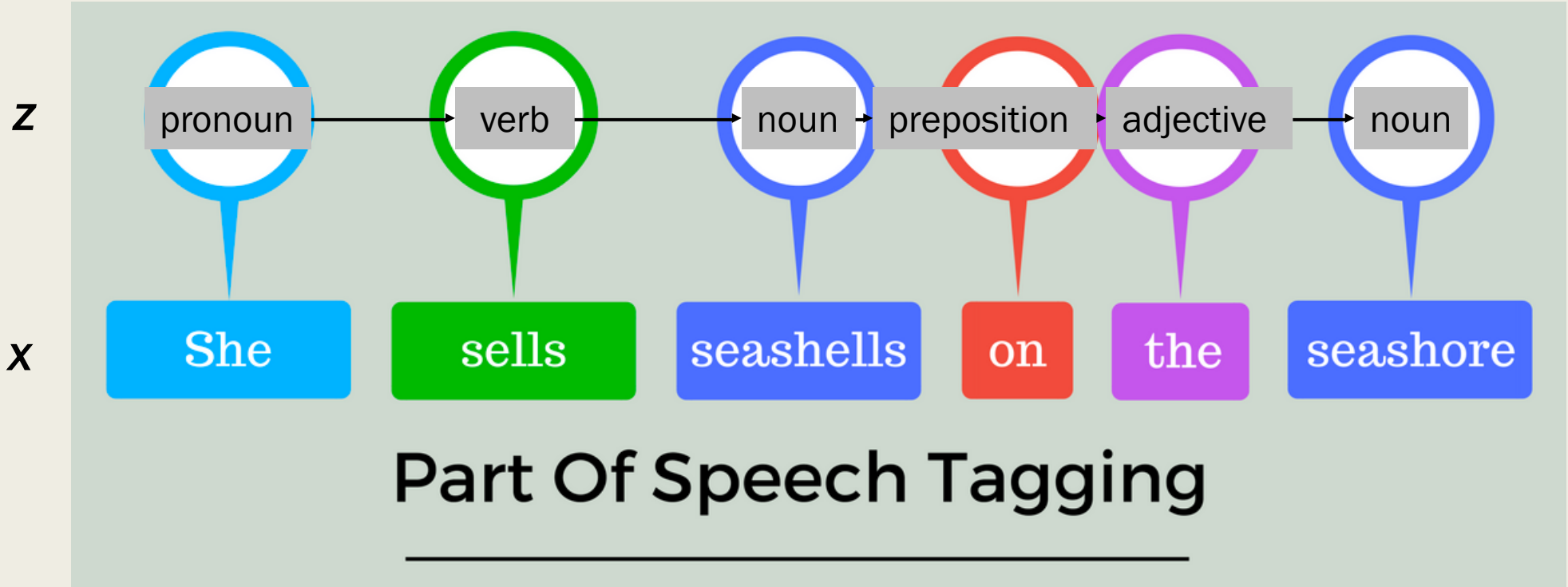


Image: modified from http://nlpforhackers.io/training-pos-tagger/

HMMs in practice

Example 2: multiple sequence alignment

HMM architecture that models matches, insertions, and deletions

Hidden Markov models of biological primary sequence information

(multiple sequence alignments/protein modeling/adaptive algorithms/sequence classification)

PIERRE BALDI*†, YVES CHAUVIN‡§, TIM HUNKAPILLER*¶, AND MARCELLA A. MCCLURE‖**

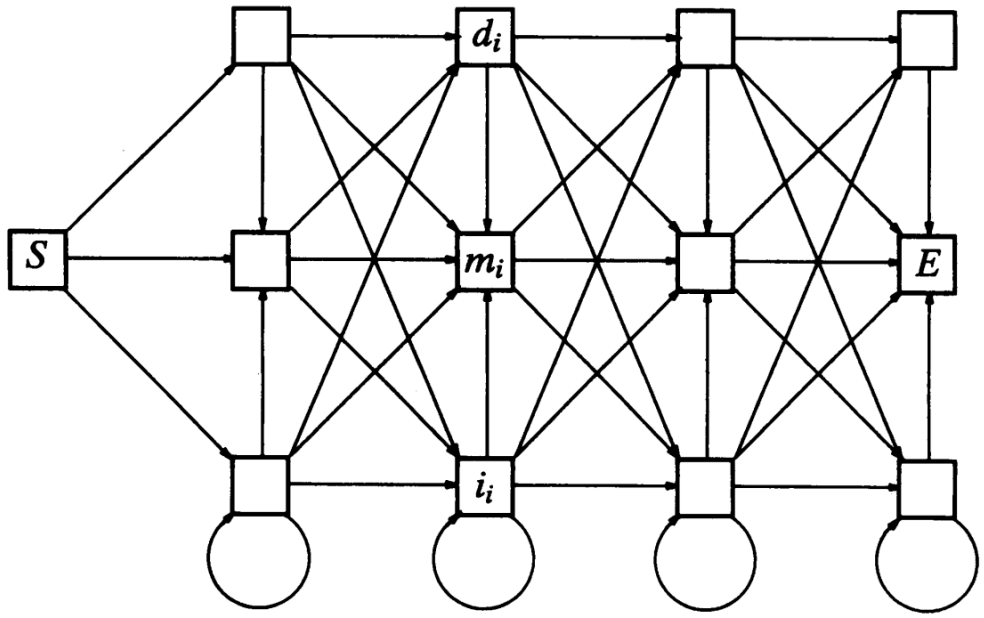


FIG. 1. HMM architecture. S and E are the start and end states. Sequence of main states m_i is the backbone. Side states d_i (resp. i_i) correspond to deletions (resp. insertions).

Multiple sequence alignment results on proteins

```
                                                 1         2                   3                   4
                                .............AAAAAAAAAAAAAAAAA.BBBBBBBBBBBBBBBBBBBBBBBBBBBBBCCCCCCC..
V HAHU      ------------------VLSPA-DKTNVKAAWGKVGAHAgEYGAE-A---L--E-R--MFL----SFPTTKTYFP
T HAOR      ------------------MLTDA-EKKEVTALWGKAAGHGeEYGAE-A---L--E-R--LFQ----AFPTTKTYFS
V HADK      ------------------VLSAA-DKTNVKGVFSKIGGHAeEYGAE-T---L--E-R--MFI----AYPQTKTYFP
V HBHU      ------------------VHLTP-EEKSAVTALWGKVNVD-EVGGE-A---L--G-R--LLV----VYPWTQRFFE
T HBOR      ------------------VHLSG-GEKSAVTNLWGKVNIN-ELGGE-A---L--G-R--LLV----VYPWTQRFFE
V HBDK      ------------------VHWTA-EEKQLITGLWGKVNVA-DCGAE-A---L--A-R--LLI----VYPWTQRFFA
V MYHU      ------------------GLSDG-EWQLVLNVWGKVEADIpGHGQE-V---L--I-R--LFK----GHPETLEKFD
V MYOR      ------------------GLSDG-EWQLVLKVWGKVEGDLpGHGQE-V---L--I-R--LFK----THPETLEKFD
V IGLOB     mkffavlalcivgaias-PLTA-DEASLVQSSWKAVSHN-EV--E-I---L--A-A--VFA----AYPDIQNKFS
V GPYL      ------------------GVLTDvQVALVKSSFEEFNANIpKN---t-hr--ffT-L--VLE----IAPGAKDLF-
V GPUGNI    ------------------ALTEK-QEALLKQSWEVLKQNIpAH---s-l-------R--LFAliieAAPESKYVF-
V GGZLB     ------------------ML-D----QQTINIIKATVPVLkEHGVT-Ittt-f--y-knLFA----KHPEVRPLFD
            ...........................................................*................*.....*.
            ...............................................................................I....
```

Motivation for Principal Components Analysis on Genetic Data: Human Evolution

Note: PCA can be applied to any species!
But PCA has bee particularly successful in humans

Sequence diversity so far…

- So far we have looked at measures of sequence diversity within a <u>single</u> population
- S (# segregating sites), π (pairwise heterozygosity), and the SFS (site frequency spectrum)
- What about measures of diversity <u>between</u> populations?

Principal component analysis (PCA)

- Transforms *n*-dimensional data so that the new first dimension explains as much of the **variation** as possible, the new second explains as much of the remaining variation as possible, and so on

- Typically, we look at the first few dimensions of the transformed data and use as a means of dimensionality reduction

- PCA is a linear transformation

- In genetics, we use PCA for:
  - *Data visualization*
  - *Infer qualitative relationships* *<u>between populations</u>*

Slide: Modified from Iain Mathieson

Principal component analysis (PCA)

- Transforms *n*-dimensional data so that the new first dimension explains as much of the **variation** as possible, the new second explains as much of the remaining variation as possible, and so on

- Typically, we look at the first few dimensions of the transformed data and use as a means of dimensionality reduction

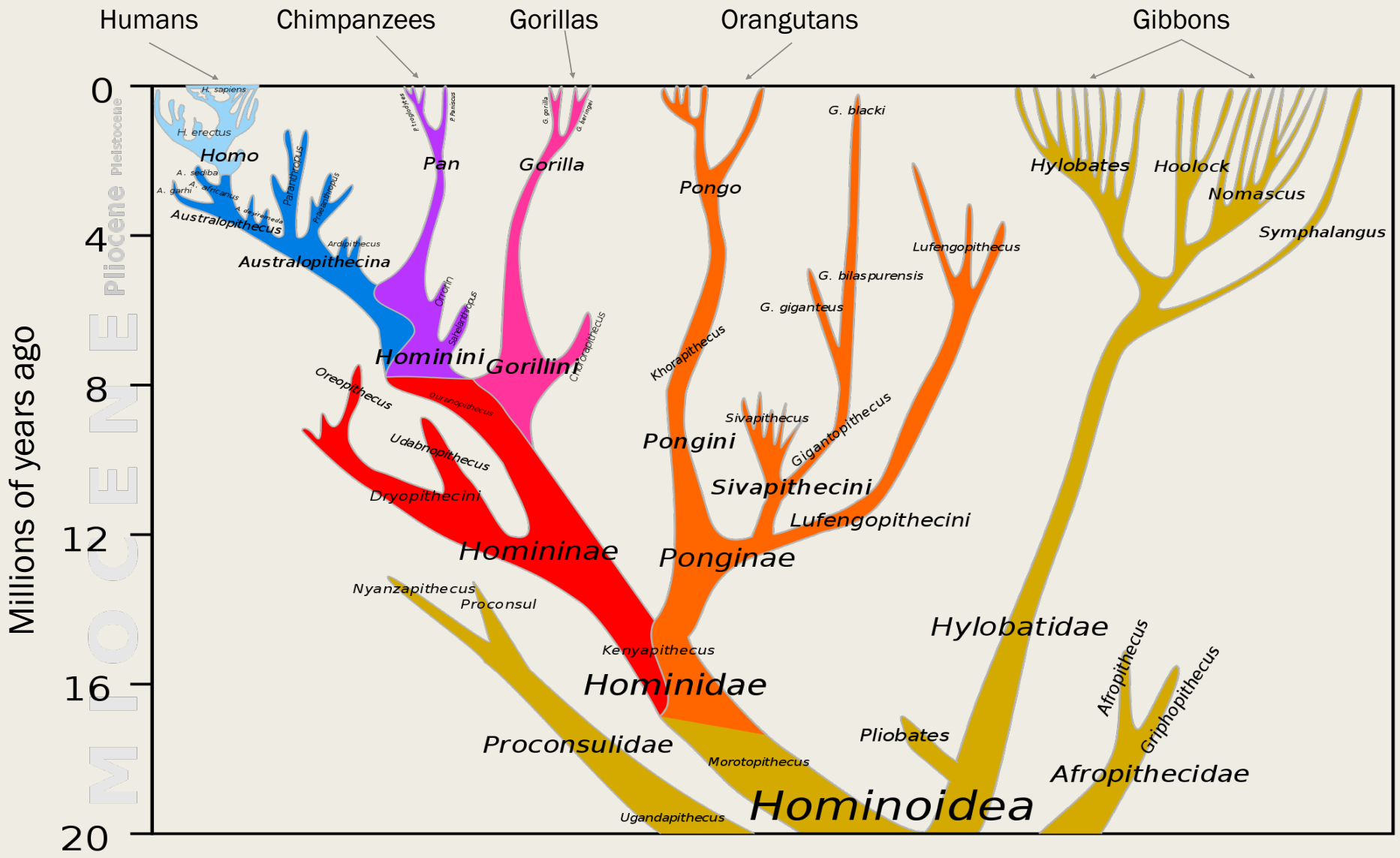
- PCA is a linear transformation

- In genetics, we use PCA for:
  - *Data visualization*
  - *Infer qualitative relationships <u>between populations</u>*

Next time: mathematical details
First: how did human populations arise?

Slide: Modified from Iain Mathieson

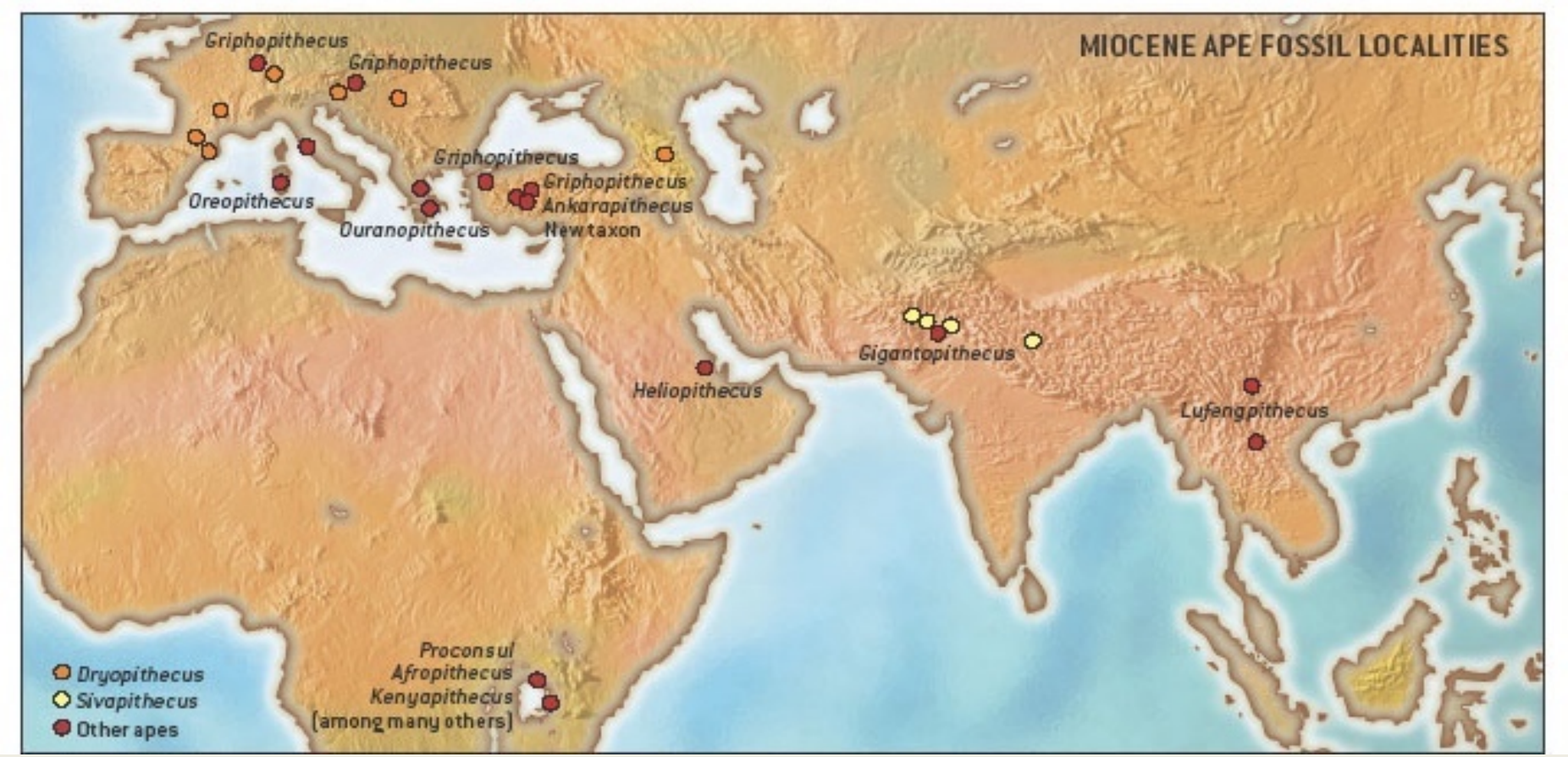
20 million years of Apes



Slide: Modified from Iain Mathieson

Wikipedia/Dbachmann

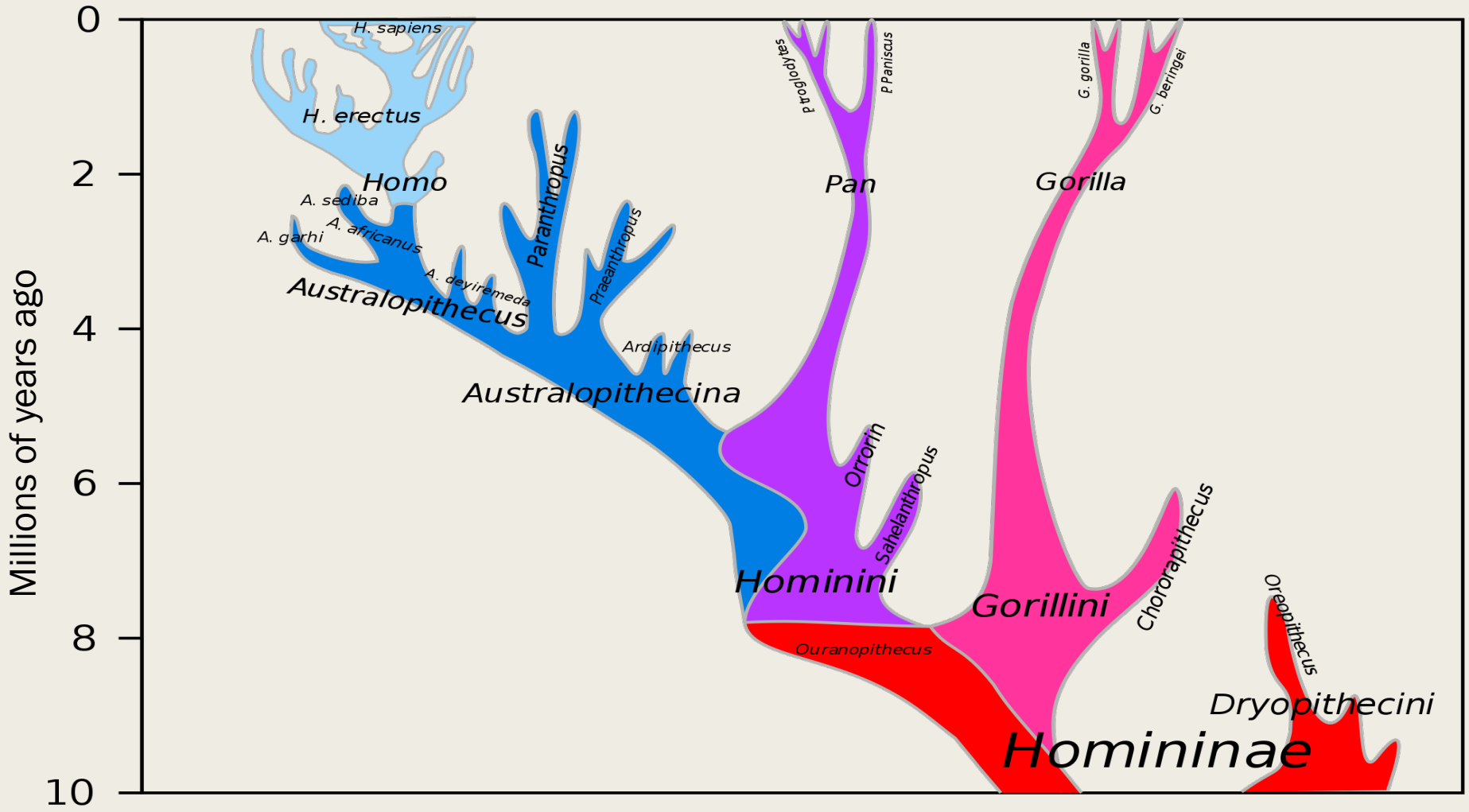
Planet of the (Miociene) Apes



Slide: Modified from Iain Mathieson

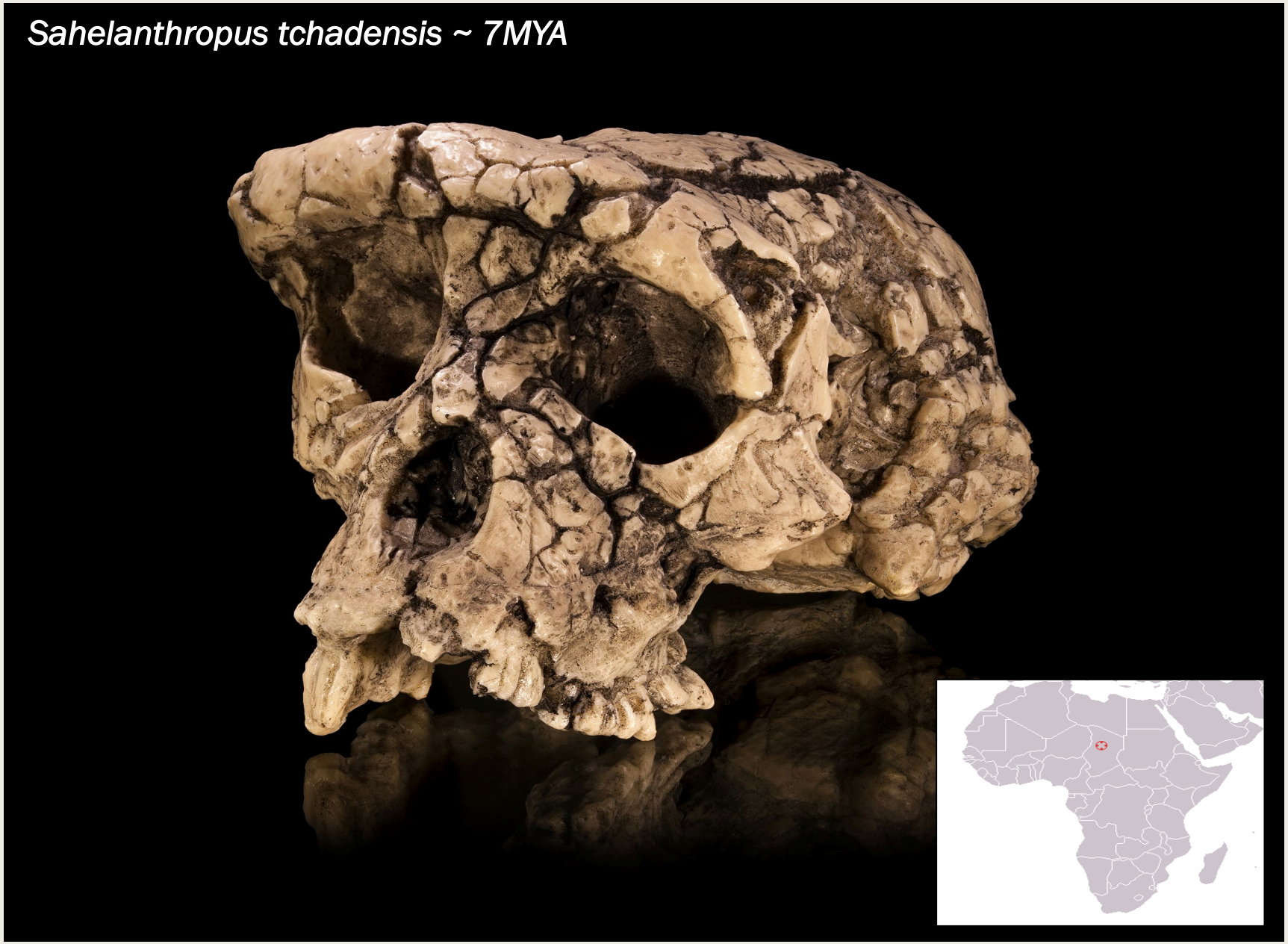
Begun & Gurche 2006: doi:10.1038/scientificamerican0606-4sp

Least Common Ancestor of Humans and Chimpanzees



Slide: Modified from Iain Mathieson

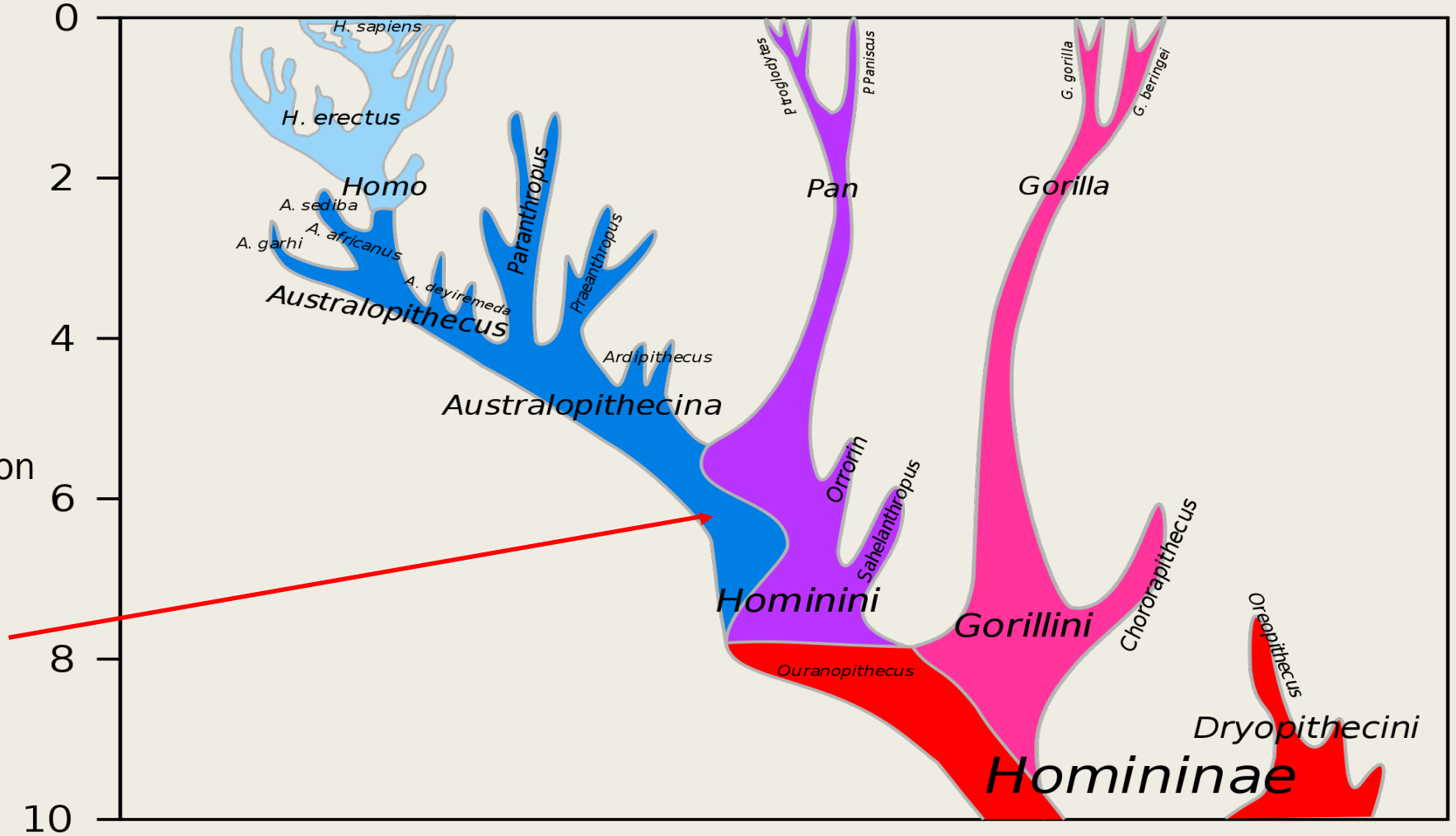
Wikipedia/Dbachmann



Slide: Modified from Iain Mathieson

Least Common Ancestor of Humans and Chimpanzees

Sahelanthropus on human lineage

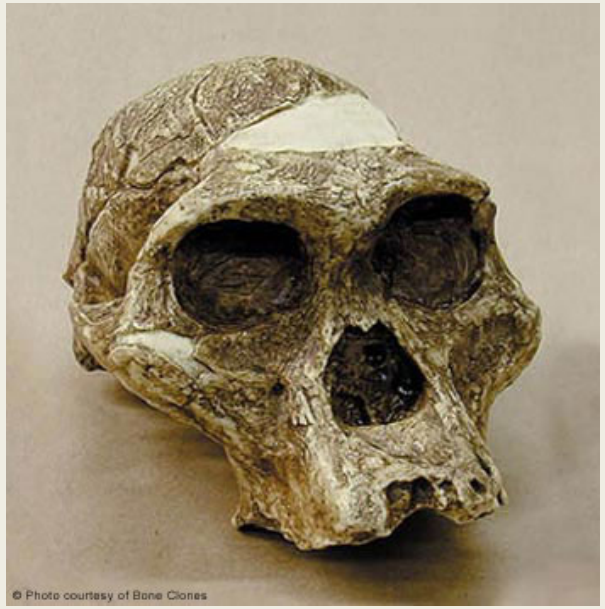


Slide: Modified from Iain Mathieson

Wikipedia/Dbachmann

Australopithecus

- Diverse group of African hominins 2-4 MYA



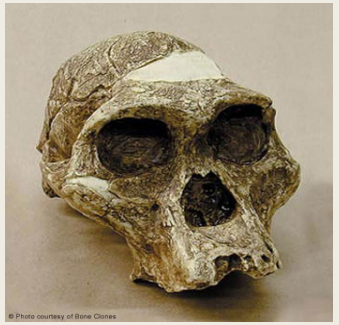
Australopithecus africanus

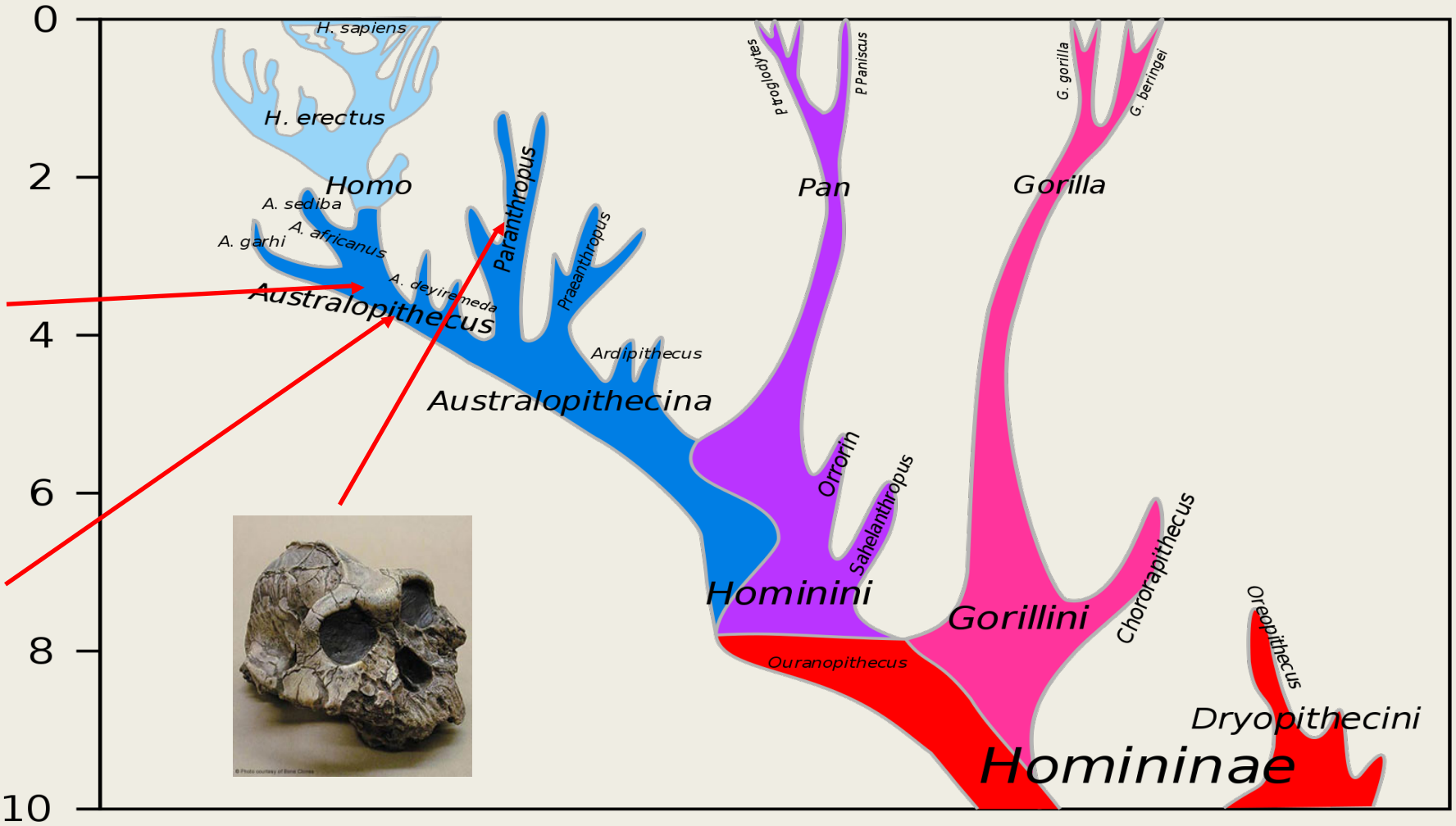
Australopithecus afarensis (Lucy)

Paranthropus boisei

Slide: Gil McVean

Further down on the human lineage



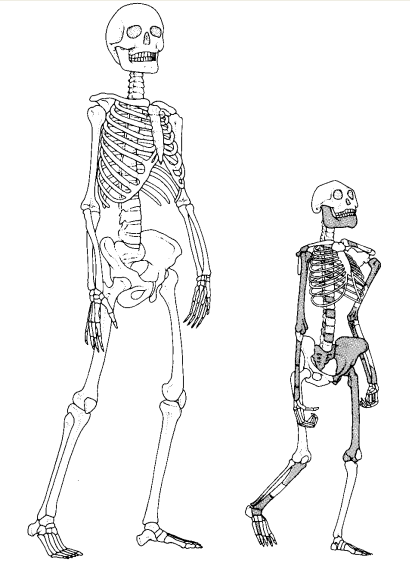


Slide: Modified from Iain Mathieson

Wikipedia/Dbachmann

Bipedalism by at least 4 million years ago

Lucy (*A. afarensis*)
Ethiopia, 3.2 MYA
Bipedal. Large teeth, small brain.



(Right) The skeleton of Lucy reconstructed at Kent State University, Ohio, in dental plaster contrasted with the skeleton of a modern human female of average height in walking position. The original parts of the australopithecine are grey and mirror images of known bones and parts based on other fossils are white. The cranium should also be partly black but has been left white for clarity. Lucy was only about 105 cm (3 ft 5 in) high but other *Australopithecus afarensis* individuals were up to 150 cm (4 ft 11 in) high. Notice Lucy's relatively long arms.

Laetoli footprints
Tanzania 3.6 MYA

Slide: Modified from Iain Mathieson

Tool use and big brains

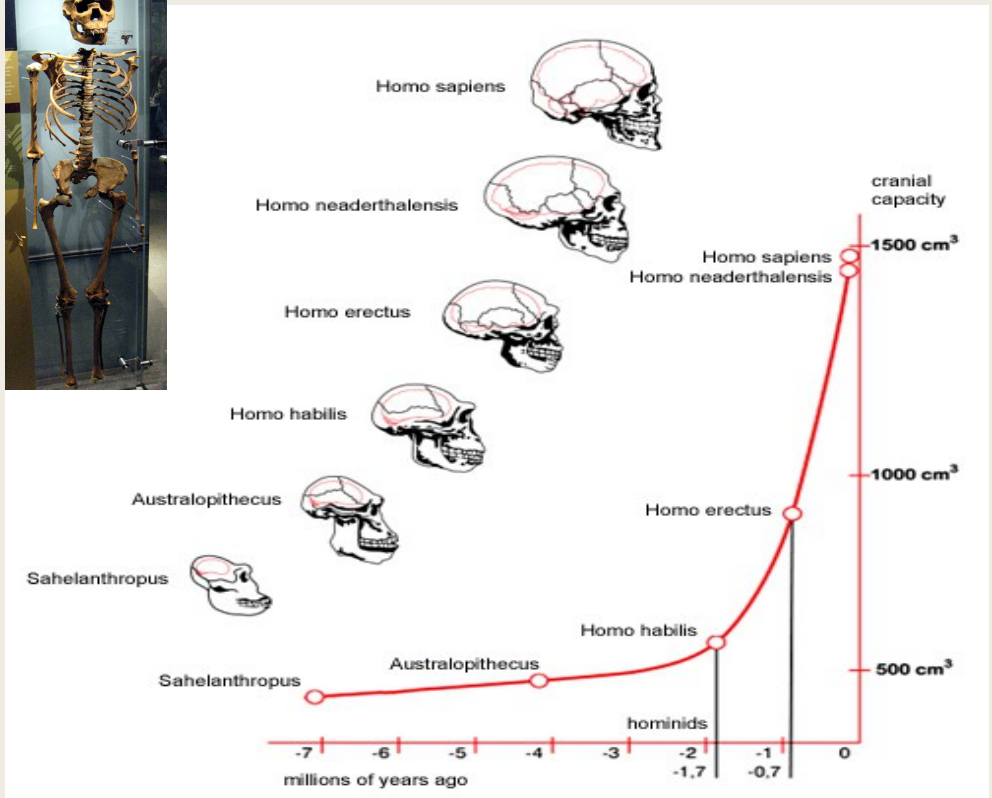
Oldowan stone tools from 2.5 MYA
(*Homo habilis*?)

OH24 1.8MYA

José-Manuel Benito Álvarez

Big brains starting around 2 MYA
(*Homo ergaster*? *Homo erectus*)

Turkana Boy 1.5MYA



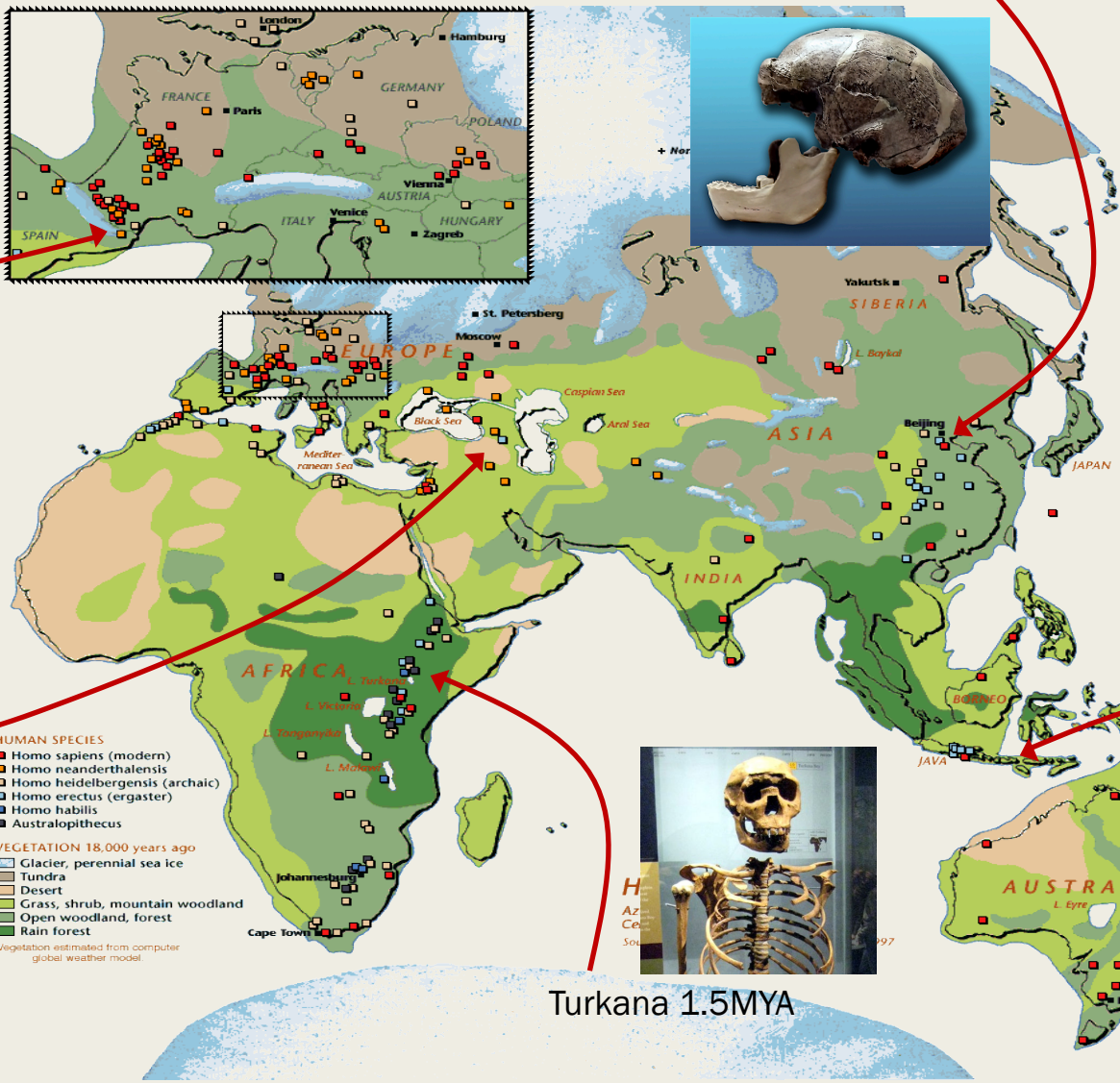
http://www.evoanth.net/

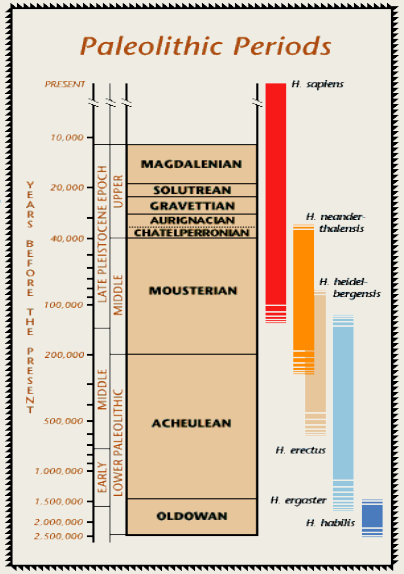
Slide: Modified from Iain Mathieson

Homo erectus: widely distributed from around 1.9 MYA to perhaps 150KYA

Tautavel 0.45MYA

Zhoukoudian 0.75MYA





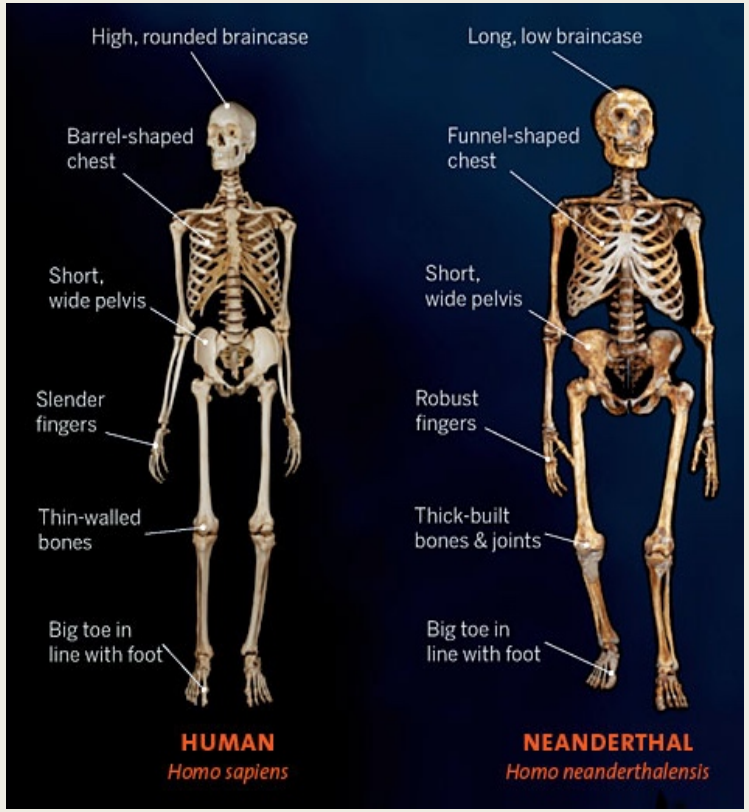
Dmanisi 1.8MYA

Turkana 1.5MYA

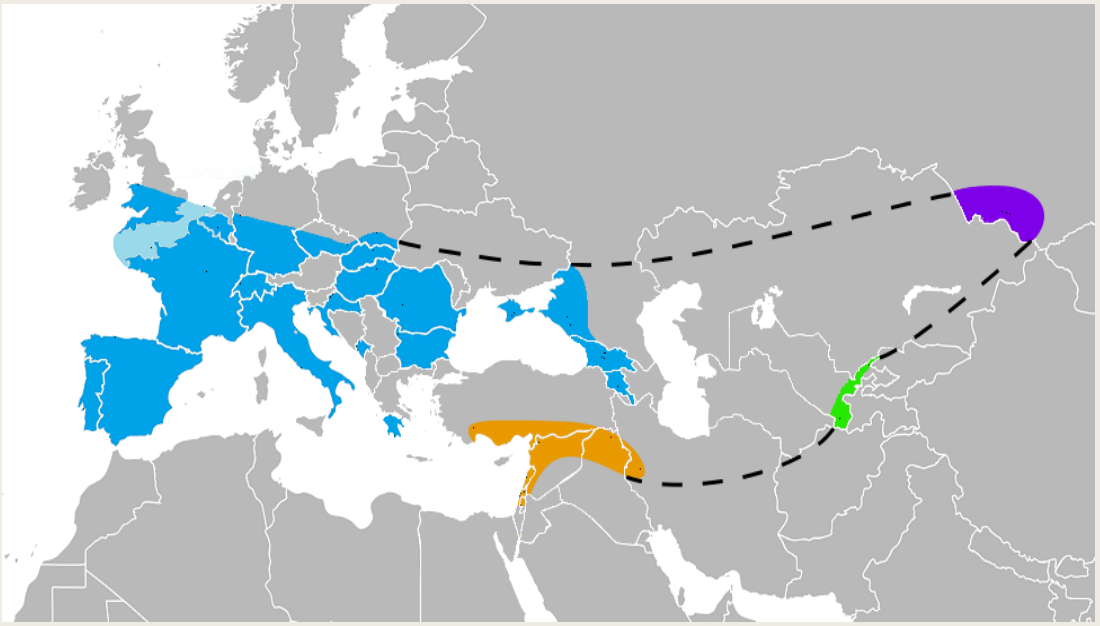
"Java Man" 0.7MYA

Slide: Modified from Iain Mathieson

Neanderthals



Smithsonian



Max-Planck Institute for Evolutionary Anthropology

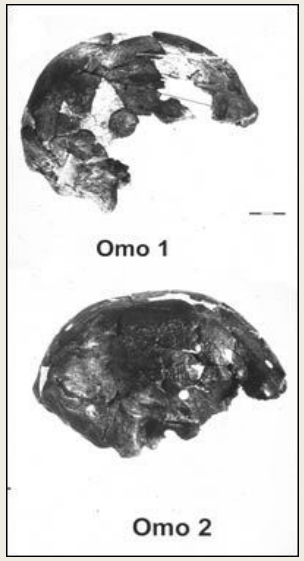
Wide distribution in Eurasia until ~40,000 years. Overlapping with modern humans.

Slide: Modified from Iain Mathieson

(Anatomically) Modern humans appear

Jebel Irhoud, Morocco, ~300 KYA

Quafzeh
Israel
80-120 KYA



Omo Kibish
Ethiopia
200 KYA

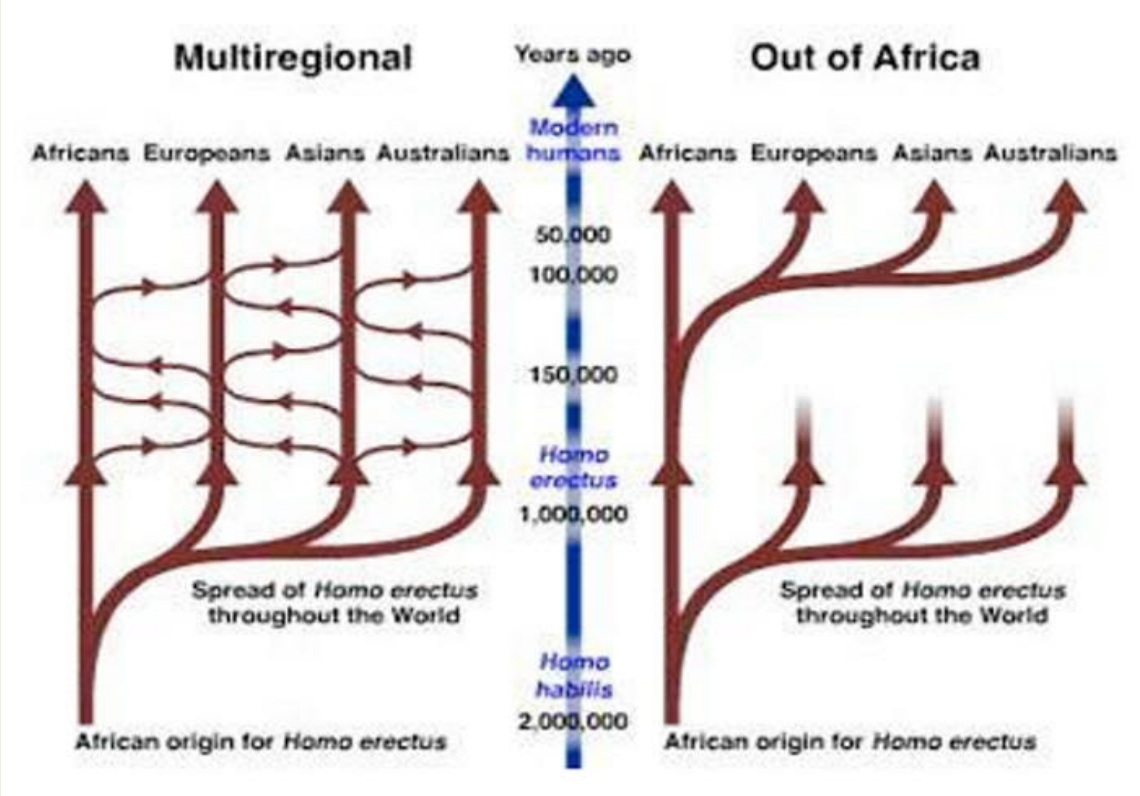
Ust'-Ishim, Siberia 45 KYA

Slide: Modified from Iain Mathieson

Figures: Smithsonian, Fu et al. 2014

Modern Humans

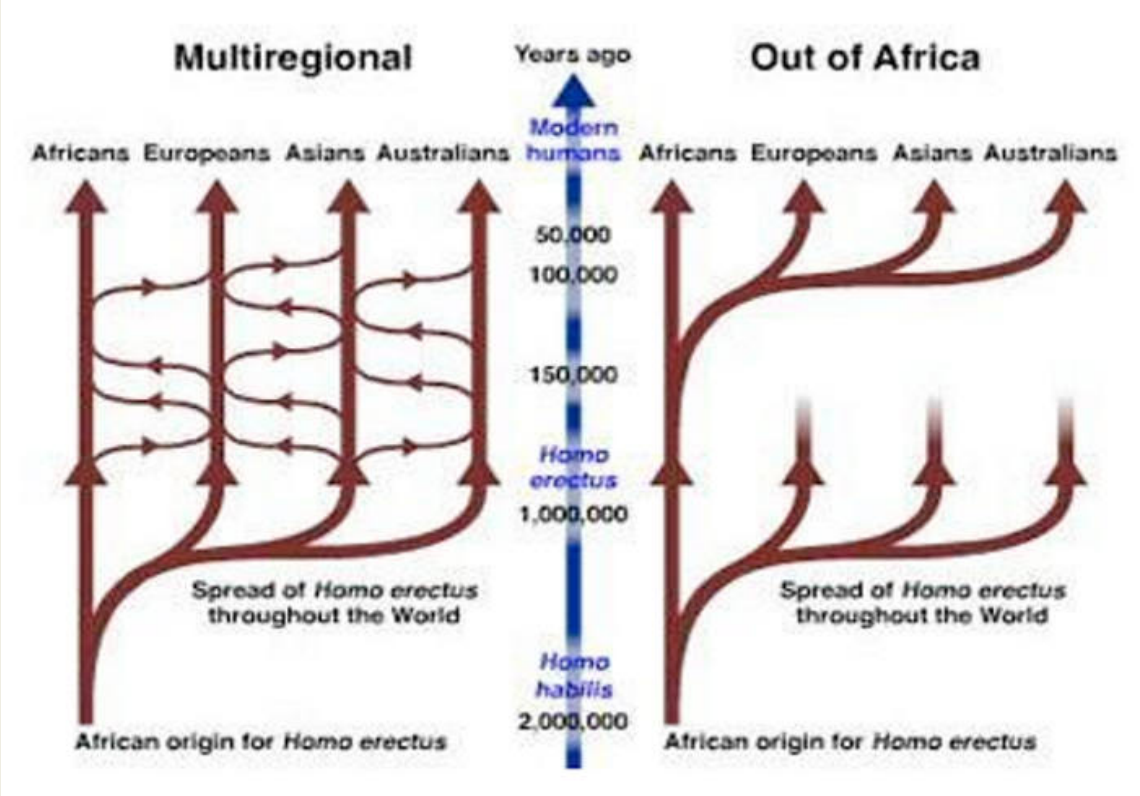
Multi-regionalism vs Recent African Origin



An old debate about human origins

Slide: Modified from Iain Mathieson

Multi-regionalism vs Recent African Origin

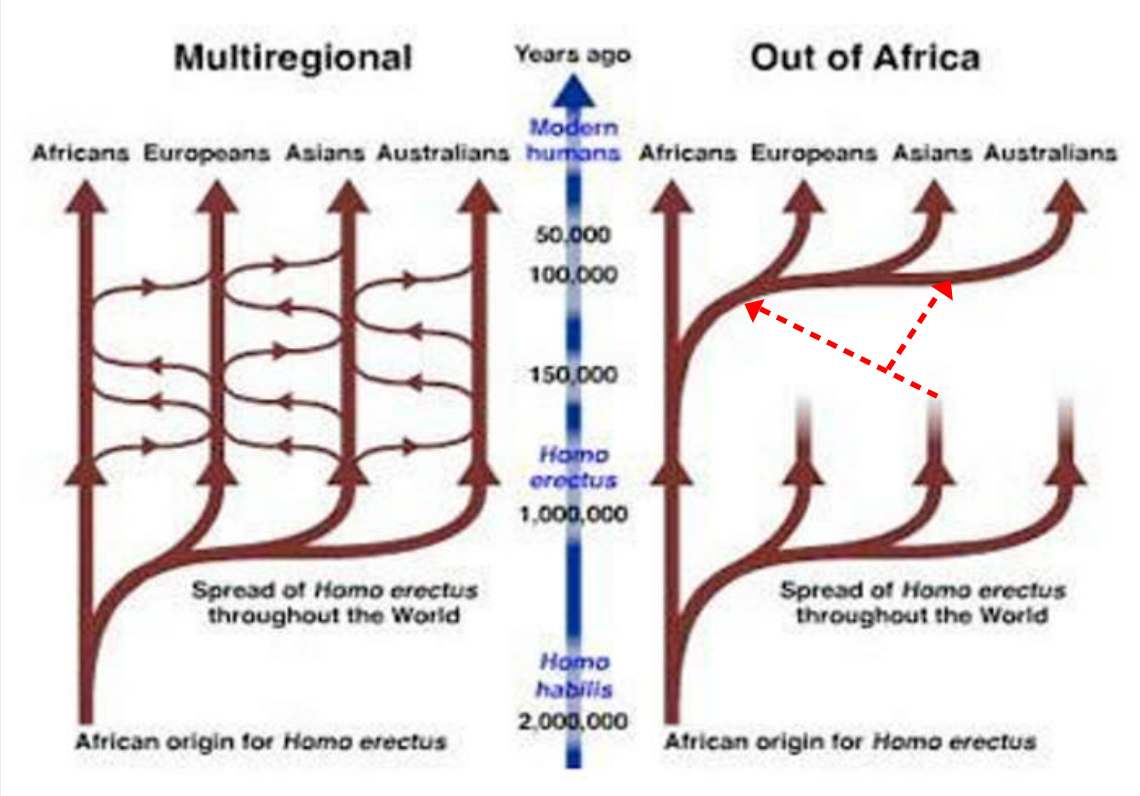


An old debate about human origins

Genetic data largely resolved this question in favor of RAO, starting in the late 1980's

Slide: Modified from Iain Mathieson

Multi-regionalism vs Recent African Origin



An old debate about human origins

Genetic data largely resolved this question in favor of RAO, starting in the late 1980's

Recent data , particularly ancient DNA slightly modifies this model.

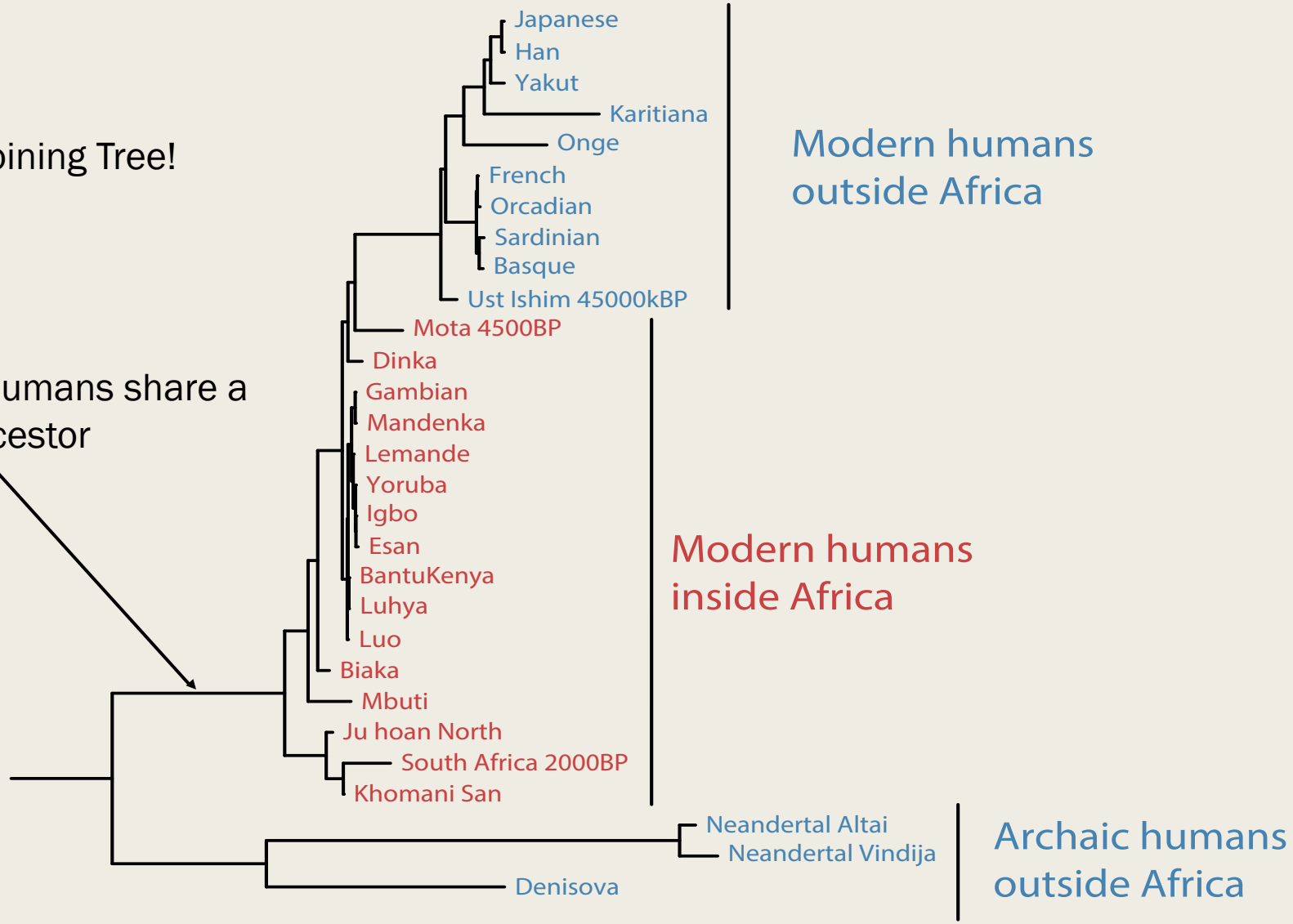
"Leaky replacement" – Svante Pääbo

Slide: Modified from Iain Mathieson

DNA data confirms RAO (Recent African Origin)

Neighbor Joining Tree!

All modern humans share a common ancestor



Skoglund & Mathieson 2018