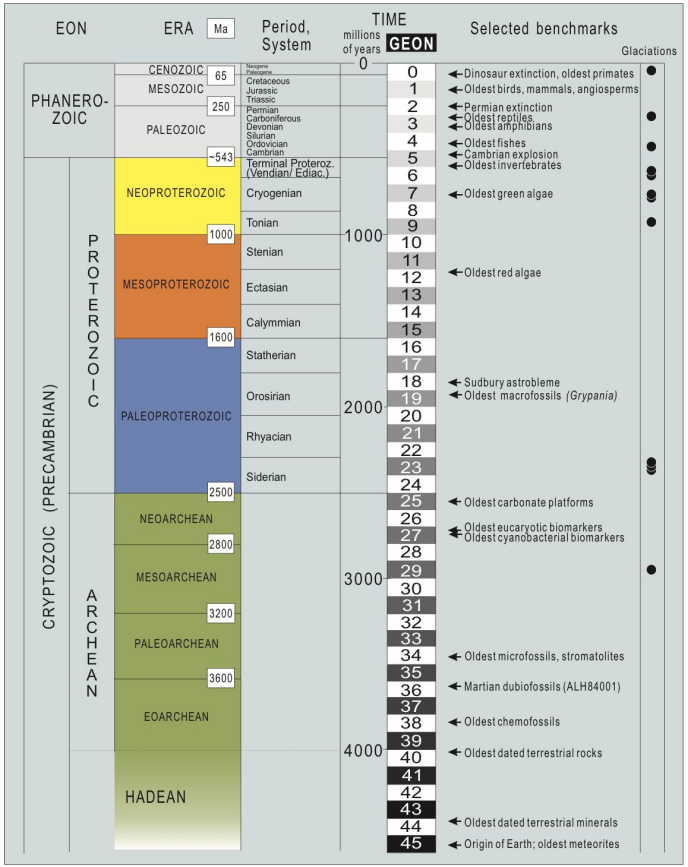
Bio 208FS: Getting started with Python

Paul Magwene

Learning Goals

- Computational Goals
  - Learn how to use Jupyter Notebooks
  - Learn about core data types in Python
  - Learn how to carry out simple numerical calculations in Python
  - Learn about variables
  - Learn about how to create and manipulate Python lists
  - First introduction to Numpy arrays
  - Create some simple visualizations
- Biological Goals
  - How old is life on earth?
  - What are the 3 major domains of life? What are models for the phylogenetics between these domains? What does LUCA mean?
  - What are some of the major differences in genome structure between the domains?
  - What is the endosymbiotic theory for the origin of eukaryotes?

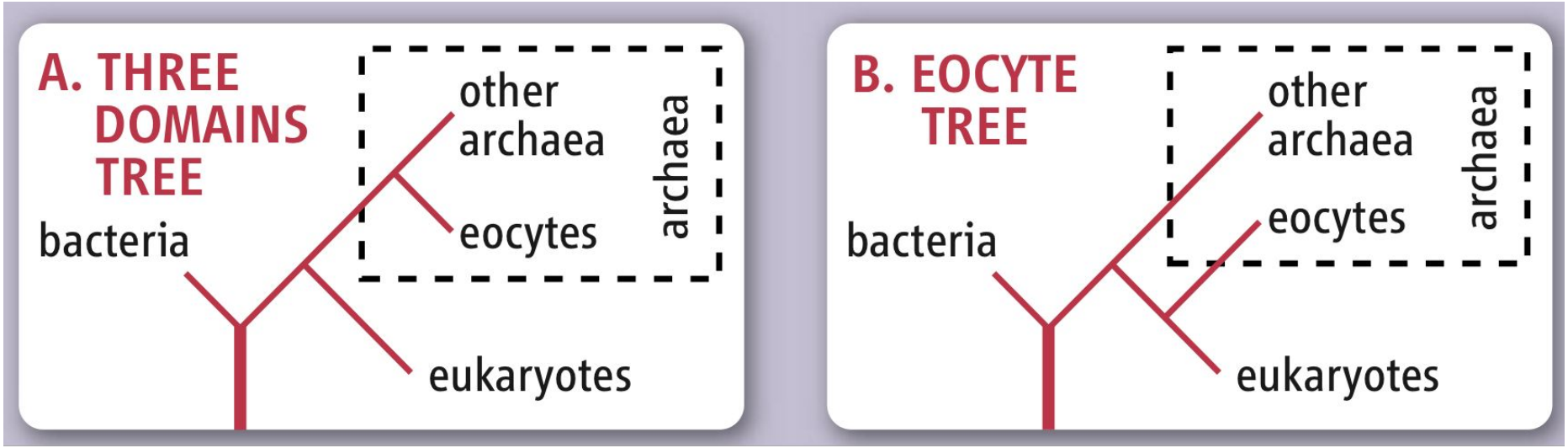
Geological overview of life on earth



Geologic scale - geon scale

(after Hofmann, 1990, 1992; 2000 - Geolog, v. 29, pt. 1, p. 18)
www.eps.mcgill.ca/~hofmann/geonscale.html

Three Major Domains in the Tree of Life



from Zimmer 2009, Science

Endosymbiotic theory on the origin of eukaryotes

FIGURE 4.02
Symbiosis with Respiring Bacteria Gives Rise to the Primitive Eukaryote

The ancestor to the eukaryote, or "urkaryote," engulfs a respiring bacterium by surrounding it with an infolding of the cell membrane. Consequently, there is now a double membrane around the newly enveloped bacterium. The symbiont, now called a "mitochondrion," divides by fission like a bacterium and provides energy for the primitive eukaryote. The mitochondrion develops infoldings of the inner membrane that increase its energy producing capacity.

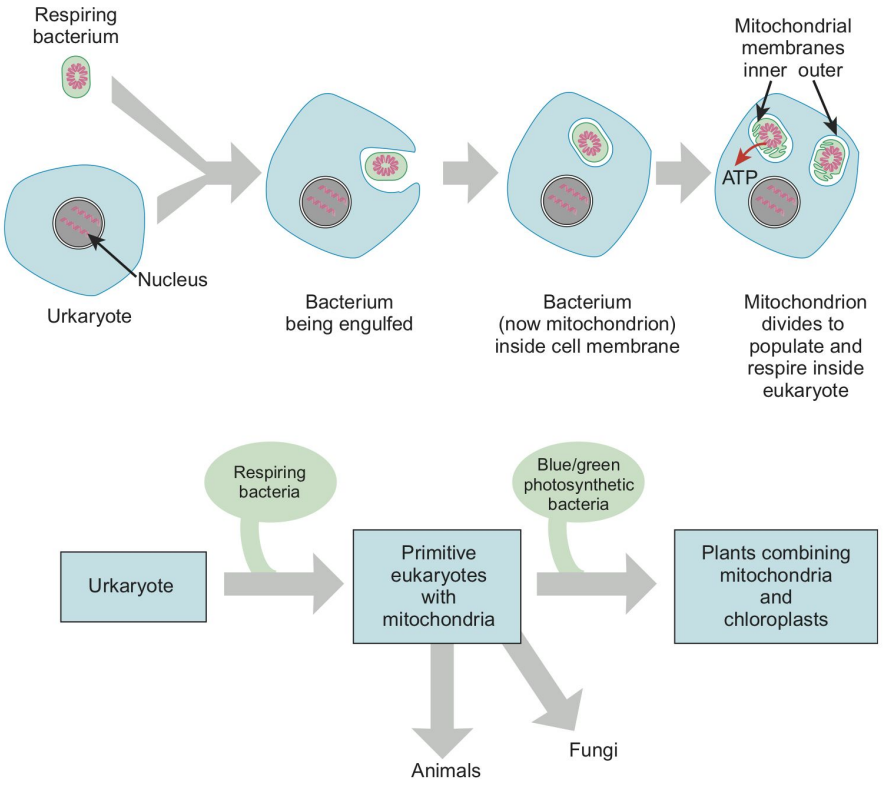


Image from Clark et al. 2019, Molecular Biology (Chapt 4)

Genome similarities and differences among the domains of life

- Similarities
  - DNA is the universal information molecule for life on earth
  - Homologous protein machinery responsible for DNA replication, transcription, and translation
  - Similar genetic codes
- Differences
  - Eubacteria and Archaea
    - Circular chromosomes
    - Genomes compacted by supercoiling
  - Eukaryotes
    - Linear chromosomes (except for mitochondria which are circular)
    - Genomes contained within a nucleus
    - Genomes compacted by chromatin

How big are genomes?

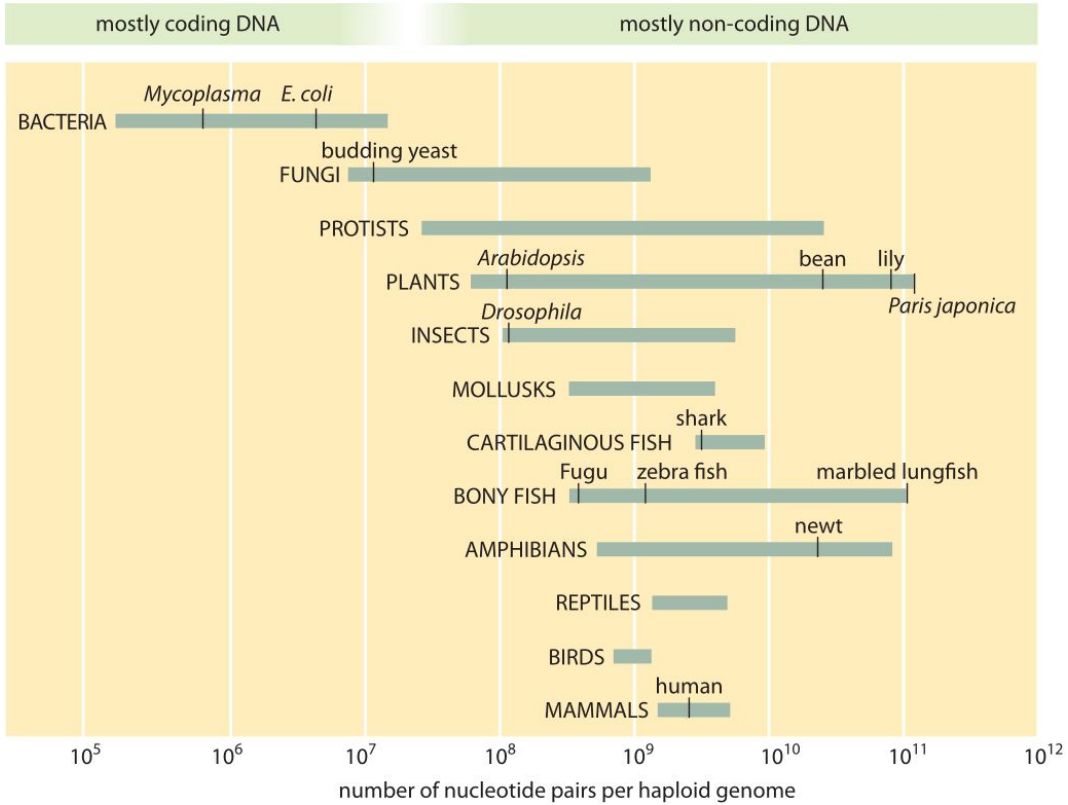


Image from Milo et al, Cell Biology by the Numbers

Jupyter Hands-on Session

Jupyter notebook

- Starting Jupyter from the Anaconda Navigator
- Naming a notebook
- Code cells
- Markdown cells
- Jupyter notebook execution model
  - When submitting a notebook for an assignment (or to share with someone else), always test your code works as intended by restarting the kernel and running from the top down
- Saving and sharing notebooks
- Efficiency tip
  - learn some of the Jupyter keyboard shortcuts!

Python data types

- Numerical types
  - Integers (ints)
  - Floating point values (floats)
  - Complex numbers (complex)
  - Support arithmetic operations
- Booleans
  - True and False value
  - Supports logical operations
- Strings
  - Represent textual data
- Use the `type` function to learn an objects type

Core arithmetic operations in Python

- Addition and subtraction
- Multiplication and division
- Exponentiation
  - Scientific notation for representing very large or small floating point values
- Use parentheses to ensure correct precedence and disambiguate potentially confusing statements
- Numerical comparisons return Boolean values
- Keep in mind limits of floating point representations when making numerical comparisons

Solve these simple numerical problems

- What is the average genome size of the organisms listed to the right?
- Which of these genomes has the highest density of genes?
- If human chromosomes were uniform in size (they're not) and genes were uniformly distributed across chromosomes (they're not!) how many genes per chromosome would you expect, and what would the average size of each gene be?

organism	genome size (base pairs)	protein coding genes	number of chromosomes
model organisms			
model bacteria *E. coli*	4.6 Mbp	4,300	1
budding yeast *S. cerevisiae*	12 Mbp	6,600	16
fission yeast *S. pombe*	13 Mbp	4,800	3
amoeba *D. discoideum*	34 Mbp	13,000	6
nematode *C. elegans*	100 Mbp	20,000	12 (2n)
fruit fly *D. melanogaster*	140 Mbp	14,000	8 (2n)
model plant *A. thaliana*	140 Mbp	27,000	10 (2n)
moss *P. patens*	510 Mbp	28,000	27
mouse *M. musculus*	2.8 Gbp	20,000	40 (2n)
human *H. sapiens*	3.2 Gbp	21,000	46 (2n)

Additional functionality can be found in libraries

- Import the Python `math` library
- Read the documentation for the `math` library
- Explore the following functions and constants
  - math.sqrt
  - math.exp
  - math.log, math.log10, math.log2
  - math.ceil / math.floor
  - math.isclose
  - math.remainder
  - math.pi
  - math.sin, math.cos, math.tan

Variables

- A variable is a name or handle associated with an object we can compute on
- Variables are an important tool for abstraction, in the same way variables are used in mathematics
- Variable naming rules in Python

Lists

- In Python, lists are ordered collections of heterogeneous object types
- Lists have a length
- Indexing is the operation of accessing elements of a list
  - Python uses "zero indexing" (the first element of a list, `L`, is `L[0]`)
  - Indexing single elements
  - Indexing ranges of elements (slicing)
  - Indexing from the end of a list
- The items in a list can be deleted or changed
- Lists can be concatenated, multiplied (concatenated with self)
- Other useful list operations
  - Testing whether an item is in/not in a list
  - Other useful list functions: max, min, sum(for numerical lists)

Numpy arrays

- Numpy is the de facto standard numerical computing library for Python
- Convenient to import the numpy library using a shortcut
  - `import numpy as np`
- Numpy arrays are ordered collections of homogeneous objects
- Numpy arrays are very convenient for numerical computations on sets of numbers
  - basic arithmetic operations and many function work element-wise for numpy arrays

Matplotlib

- Matplotlib is the de facto standard plotting library for Python
- Powerful library for creating many different plot types
- Drawing matplotlib plots in Jupyter notebooks using the `%matplotlib inline` invocation (notebook "magics")
- The pyplot submodule provides a convenient interface for some of the most common plot types
  - `from matplotlib import pyplot`
- We're just going to look at two plot types today
  - Histograms -- `pyplot.hist`
  - Scatterplots -- `pyplot.scatter`

Create these plots

- Draw a histogram of genome sizes for the table on the right
  - Recreate the histogram using log10(genome size)
- Compare the above histograms. Is one more useful than the other? Why?
- Draw a scatter plot relating genome size (x-axis) to the number of protein coding genes (y-axis)
  - Recreate the same plot, but with log10(genome size) as the x-axis
- When and why might log scaling be useful?

For a larger version see Cell Biology by the Numbers

organism	genome size (base pairs)	protein coding genes	number of chromosomes
model organisms			
model bacteria *E. coli*	4.6 Mbp	4,300	1
budding yeast *S. cerevisiae*	12 Mbp	6,600	16
fission yeast *S. pombe*	13 Mbp	4,800	3
amoeba *D. discoideum*	34 Mbp	13,000	6
nematode *C. elegans*	100 Mbp	20,000	12 (2n)
fruit fly *D. melanogaster*	140 Mbp	14,000	8 (2n)
model plant *A. thaliana*	140 Mbp	27,000	10 (2n)
moss *P. patens*	510 Mbp	28,000	27
mouse *M. musculus*	2.8 Gbp	20,000	40 (2n)
human *H. sapiens*	3.2 Gbp	21,000	46 (2n)
viruses			
hepatitis D virus (smallest known animal RNA virus)	1.7 Kb	1	ssRNA
HIV-1	9.7 kbp	9	2 ssRNA (2n)
influenza A	14 kbp	11	8 ssRNA
bacteriophage λ	49 kbp	66	1 dsDNA
Pandoravirus salinus (largest known viral genome)	2.8 Mbp	2500	1 dsDNA
organelles			
mitochondria - *H. sapiens*	16.8 kbp	13 (+22 tRNA +2 rRNA)	1
mitochondria – *S. cerevisiae*	86 kbp	8	1
chloroplast – *A. thaliana*	150 kbp	100	1
bacteria			
C. ruddii (smallest genome of an endosymbiont bacteria)	160 kbp	182	1
M. genitalium (smallest genome of a free living bacteria)	580 kbp	470	1
H. pylori	1.7 Mbp	1,600	1
Cyanobacteria S. elongatus	2.7 Mbp	3,000	1
methicillin-resistant *S. aureus* (MRSA)	2.9 Mbp	2,700	1
B. subtilis	4.3 Mbp	4,100	1
S. cellulosum (largest known bacterial genome)	13 Mbp	9,400	1
archaea			
Nanoarchaeum equitans (smallest parasitic archaeal genome)	490 kbp	550	1
Thermoplasma acidophilum (flourishes in pH<1)	1.6 Mbp	1,500	1
Methanocaldococcus (Methanococcus) jannaschii (from ocean bottom hydrothermal vents; pressure >200 atm)	1.7 Mbp	1,700	1
Pyrococcus furiosus (optimal temp 100°C)	1.9 Mbp	2,000	1
eukaryotes - multicellular			
pufferfish *Fugu rubripes* (smallest known vertebrate genome)	400 Mbp	19,000	22
poplar *P. trichocarpa* (first tree genome sequenced)	500 Mbp	46,000	19
corn *Z. mays*	2.3 Gbp	33,000	20 (2n)
dog *C. familiaris*	2.4 Gbp	19,000	40
chimpanzee *P. troglodytes*	3.3 Gbp	19,000	48 (2n)
wheat *T. aestivum* (hexaploid)	16.8 Gbp	95,000	42 (2n=6x)
marbled lungfish *P. aethiopicus* (largest known animal genome)	130 Gbp	unknown	34 (2n)
herb plant *Paris japonica* (largest known genome)	150 Gbp	unknown	40 (2n)