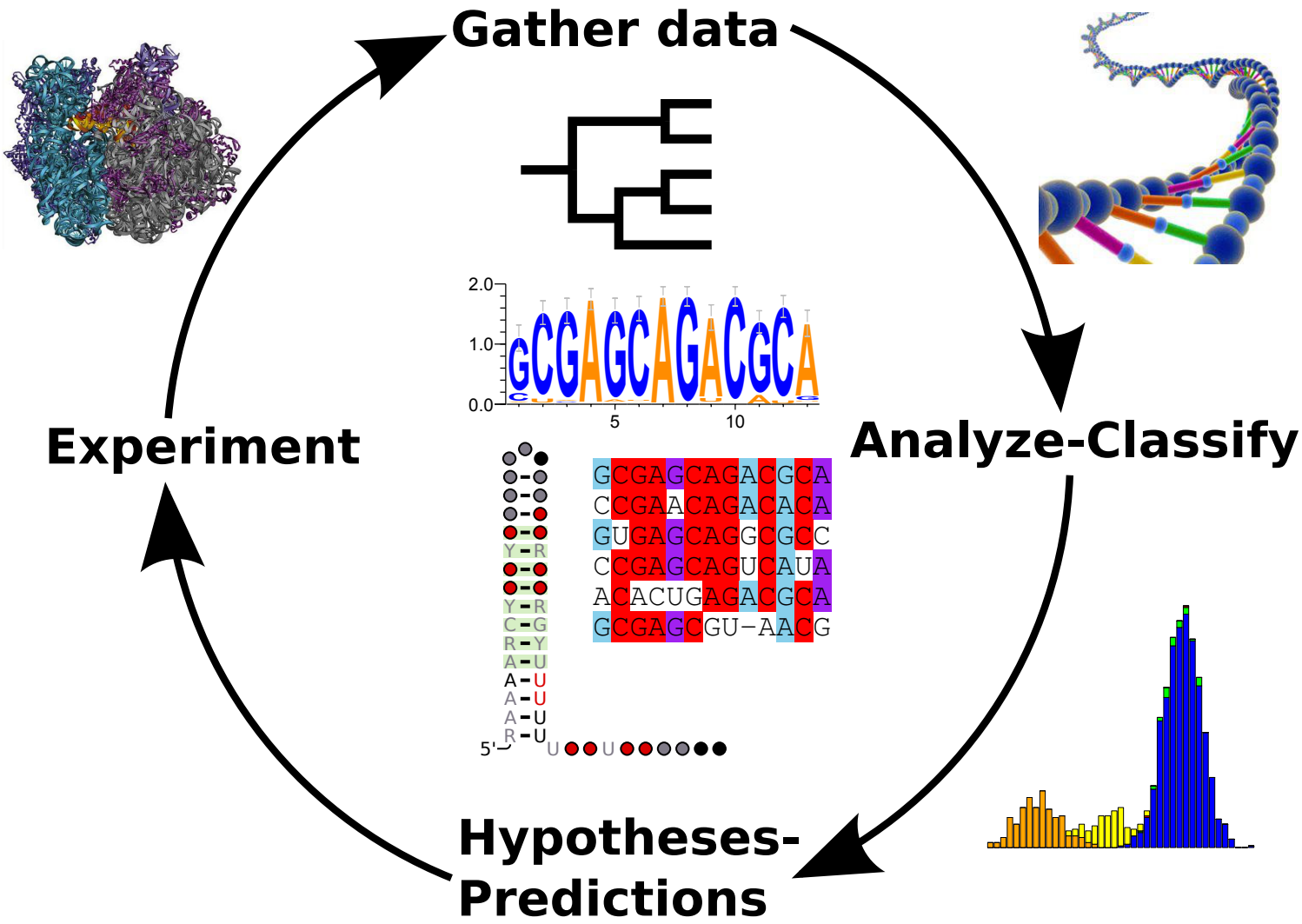
Metagenomics needs computers!

Paul Gardner

September 28, 2013

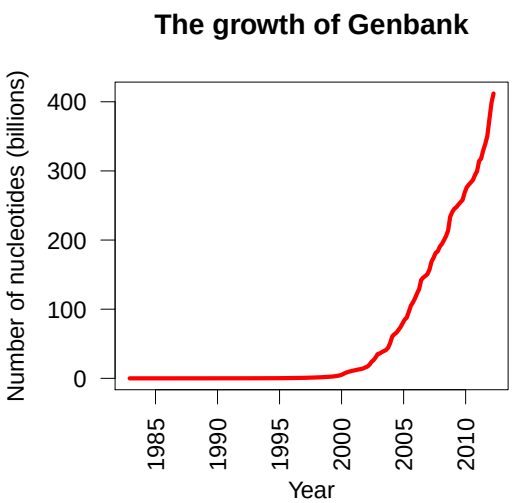
What is computational biology?

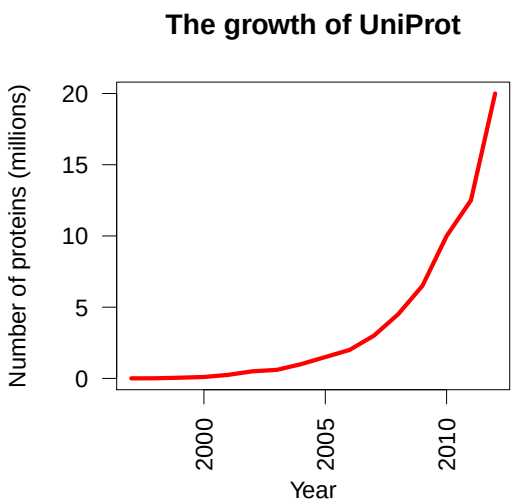
How did the Geeks get so important?

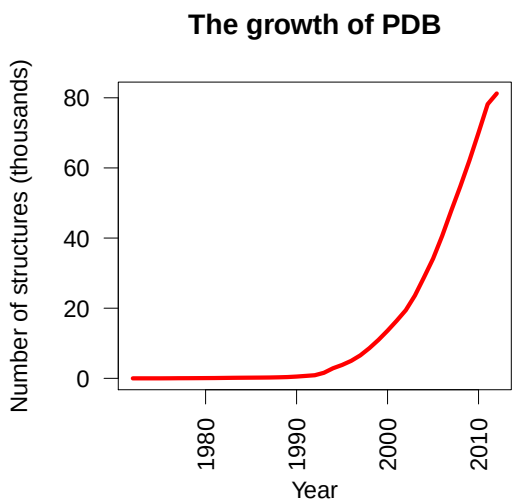


The data deluge and the rise of computational biology

- There is a deluge of data being generated by new techniques in sequencing and structure determination
- Computational tools are the only way to store and analyse the amount of data that now drives a lot of biological discoveries







We need giant computers

We need giant computers

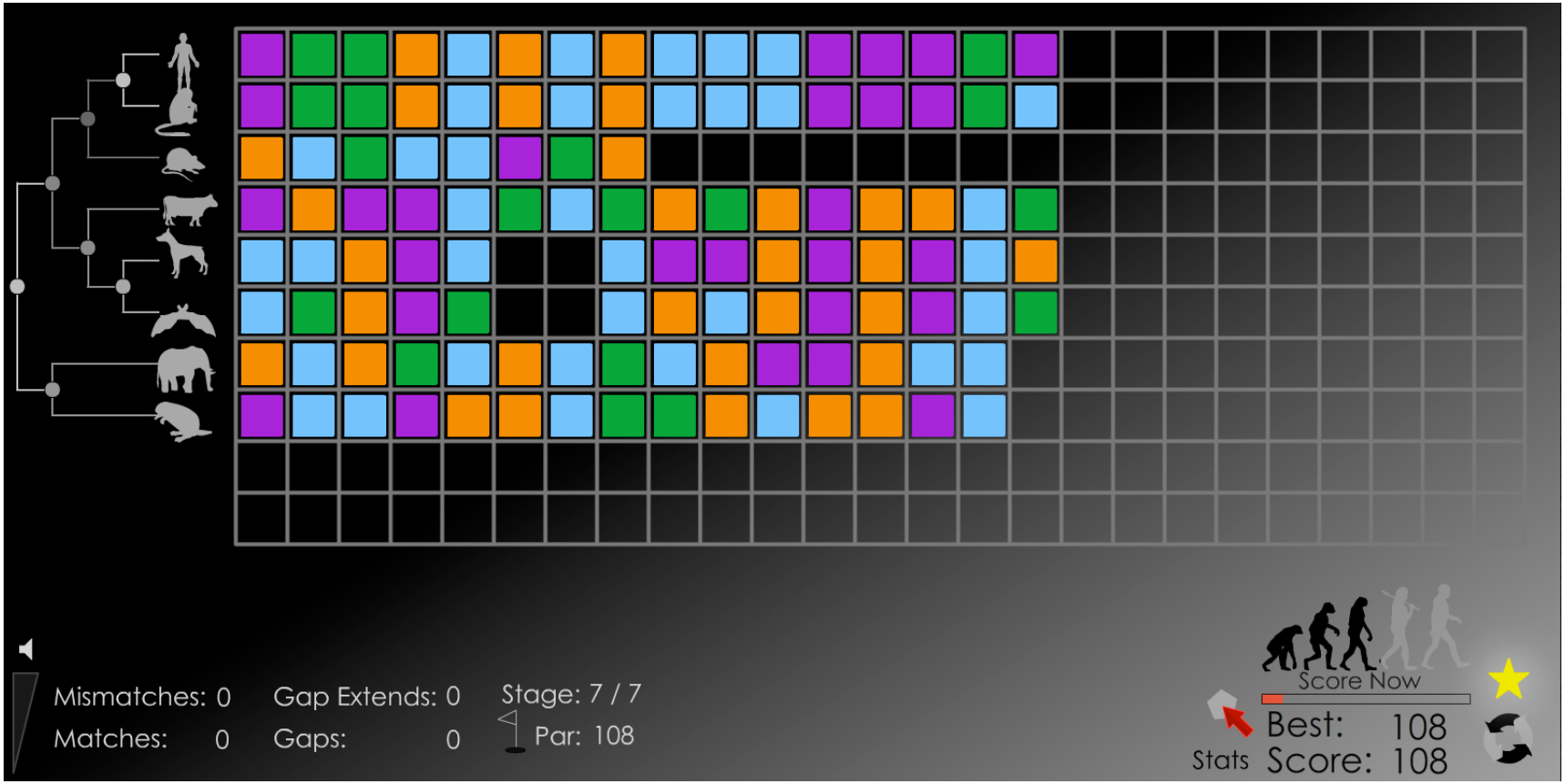
- UC Biology cluster currently runs a 152 fast CPU cluster
- Access to BlueGene computer facility, $\approx$ 1,000 CPUs

New ways to build "computers" ...

- Who likes gaming?

Games with a purpose

- Games with purpose
  - "We spend 3 billion hours a week as a planet playing videogames." – Jane McGonigal
- Phylo, EteRNA, Fold.it

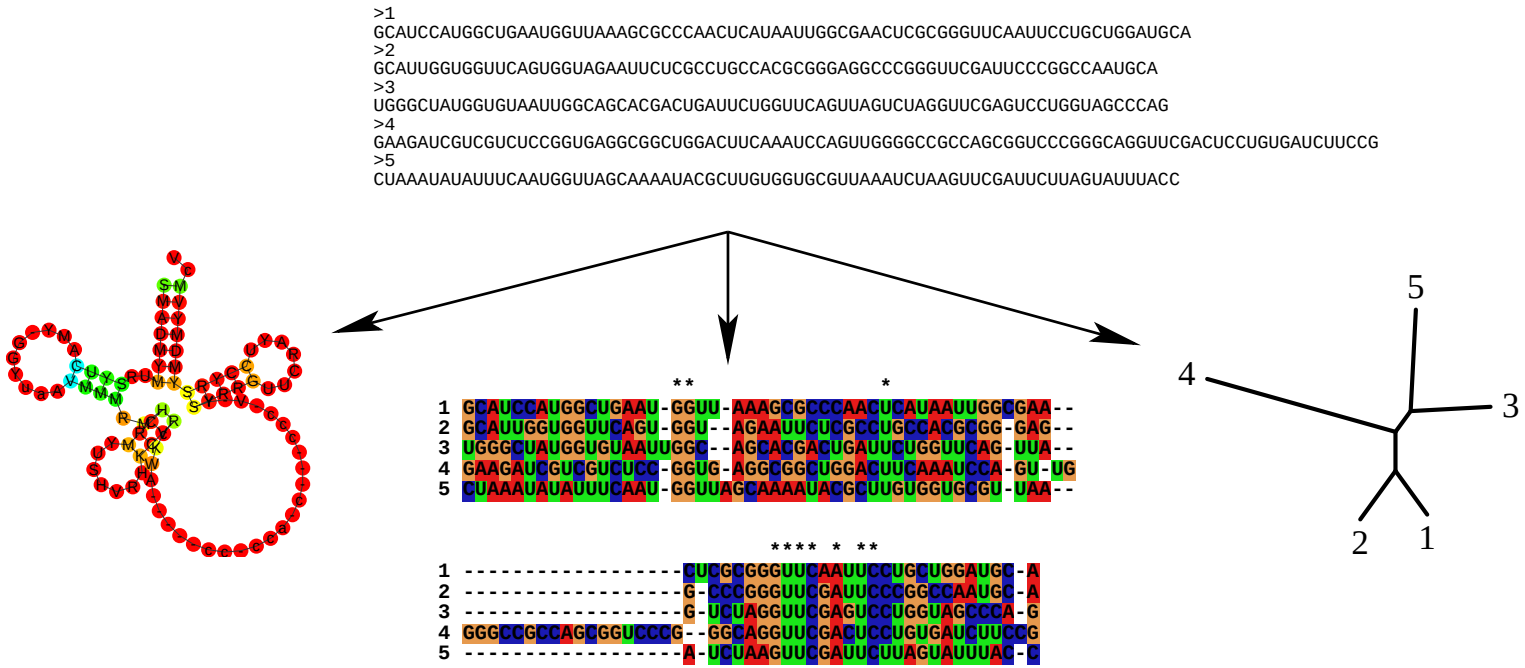


Where can I find biological data?

- National Center for Biotechnology Information
  - PubMed
  - Genbank
  - Gene Expression Omnibus
  - dbSNP
  - RefSeq
  - Entrez Gene
  - OMIM
  - Taxonomy
- European Bioinformatics Institute
  - EMBL/ENA
  - ENSEMBL
  - InterPro
  - Reactome
  - UniProt
  - PRIDE
  - ArrayExpress
  - Gene Ontology
- The Ribosomal DB Project
- KEGG
- Pfam and Rfam
- UCSC genome browsers
- PDB
- GIYF and Wikipedia

What can computational biology do for me?

- Sequence alignment
- Phylogenetic analysis
- Analysis of gene expression
- Infer regulatory networks
- RNA and Protein structure prediction
- Gene prediction and annotation



How can we measure evolutionary relatedness?

- Is a sponge more related to a jellyfish or the custard apple?
- How can we test this?

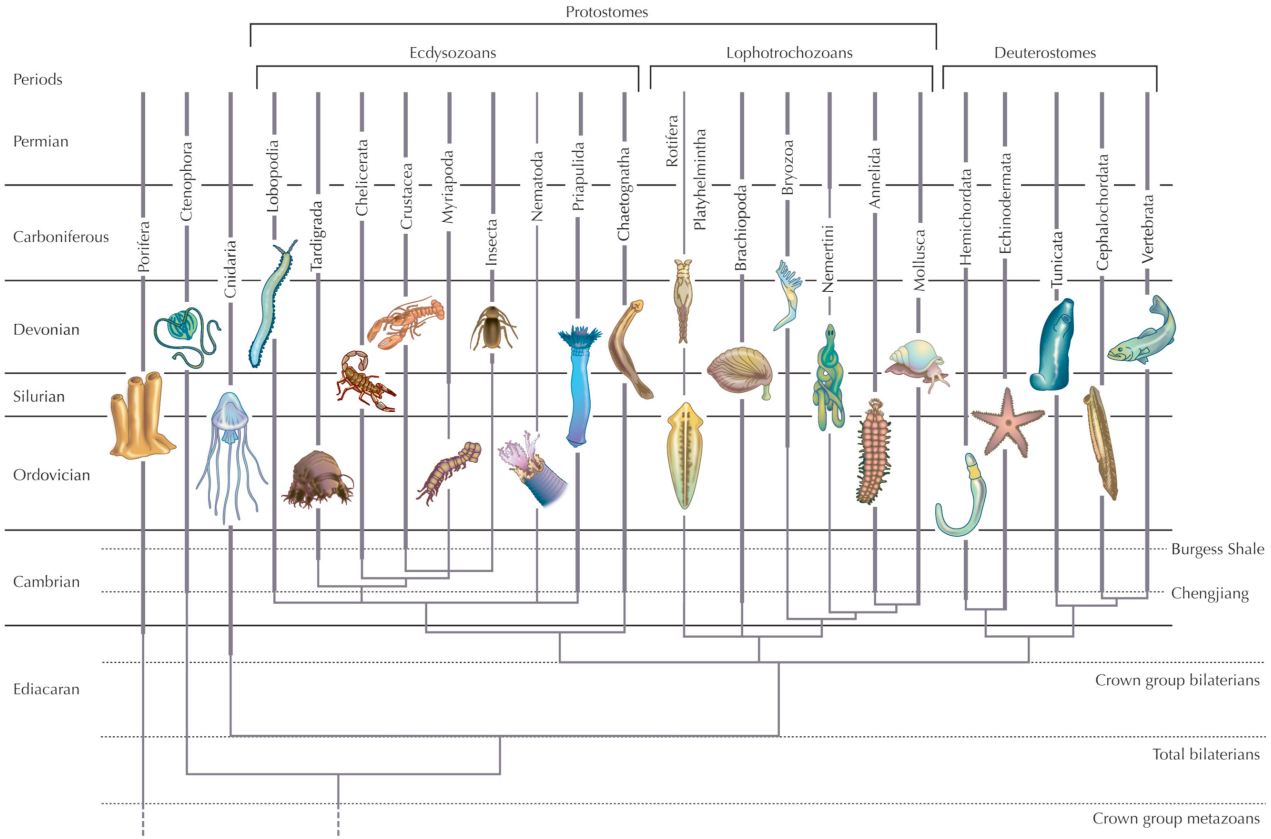


FIGURE 10.16. The fossil record and metazoan phylogeny. *Dark lines* represent the temporal range of phyla from their first appearance in the fossil record to the present. Extrapolation into the Ediacaran is based on molecular clock data. Some relationships, particularly among the arthropods, remain controversial (see Fig. 10.17).

10.16, adapted from original drawing by Susan Butts, based on design by Matthew Wills

Barton *et al.* (2007) Evolution. *Cold Spring Harbor Laboratory Press*

The homology search problem

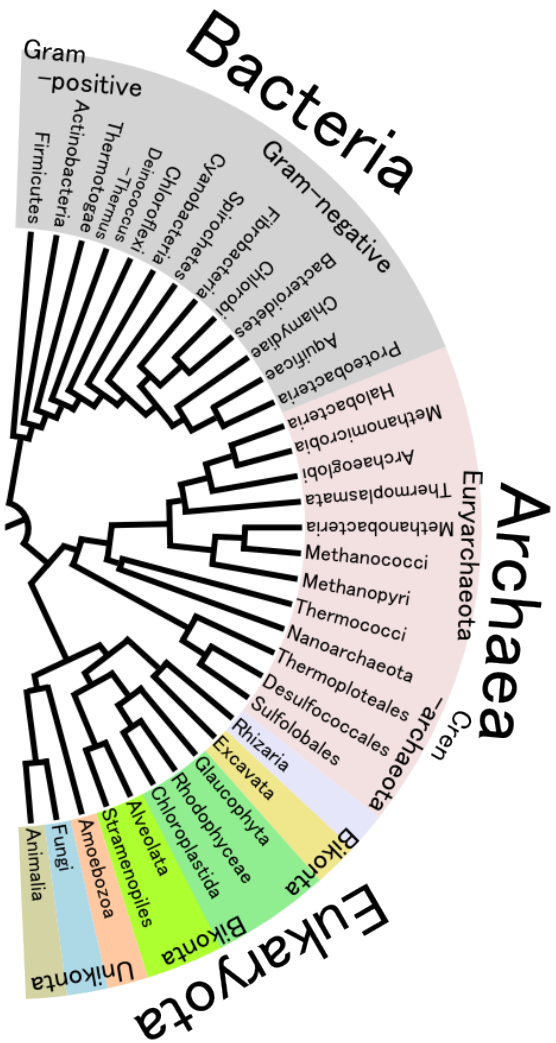


Image source: www.wikimedia.org

- How can we compare DNA sequences?

```
H.pylori          ------GVDA NALHRPKRFF GAARNIEEGG SLTIIATALI ETGSRMDEVI ------FEEF
D.radiodurans     ------GVDS TALYPPKRFL GAARNIEEGG SLTIIATAMV ETGSTGDTVI ------FEEF
M.tuberculosis    ------GVDS TALYPPKRFL GAARNIEEGG SLTIIATAMV ETGSTGDTVI ------FEEF
C.pneumoniae      ------GVDA SALHKPKRFF GAARNIEGGG SLTILATALI DTGSRMDEVI ------FEEF
F.nucleatum       ------GIDP TALYHPKNFF GAARNIKDGG SLTIIATILV DTGSKMDEVI ------YEEF
S.enterica        ------GVDA NALHRPKRFF GAARNVEEGG SLTIIATALI DTGSKMDEVI ------YEEF
B.thetaiotaomicron------GVDA NALHKPKRFF GAARNIENGG SLTIIATALI DTGSKMDEVI ------FEEF
L.interrogans     ------GVDS NALHKPKRFF GAARNIEEGG SLTIIATALI DTGSKMDEVI ------FEEF
P.marinus         GYQPTLGTDV GELQER----- -ITSTLE--G SITSIQAVYV PADDLTDPAP ------ATTF
U.parvum          GYQSTLESDV THIQNR----- -LFRNKN--G SITSFQTIFL PMDDLSDPSA ------VAVF
B.subtilis        ------GIDP AAFHRPKRFF GAARNIEEGG SLTILATALV DTGSRMDDVI ------YEEF
C.difficile       ------GIDP GALHGPKKFF GAARNIRQGG SLTILGTALV ETGSRMDDVI ------FEEF
S.griseus         ------GVDS TALYPPKRFF GAARNIEDGG SLTILATALV ETGSRMDEVI ------FEEF
F.nodosum         ------GVDP AALYKPKHFF GAARNTREGG SLTIIATALI ETGSKMDEVI ------FEEF
M.infernorum      ------GVDA NALQKPRRFF ATARNLEEGG SVTIIATALI DTGSKMDDVI ------FEEF
T.yellowstonii    ------GLEA TALQKPKRFF GTARNIEEGG SLTIIATALV ETGSRMDDVI ------FEEF
E.coli            ------GVDA NALHRPKRFF GAARNVEEGG SLTIIATALI DTGSKMDEVI ------YEEF
H.volcanii        GYPAYLAARL SEFYERAGYF TTVNGEE--G SVSVIGAVSP PGGDFSEPVT QNTLRIVKTF
```

Metagenomics

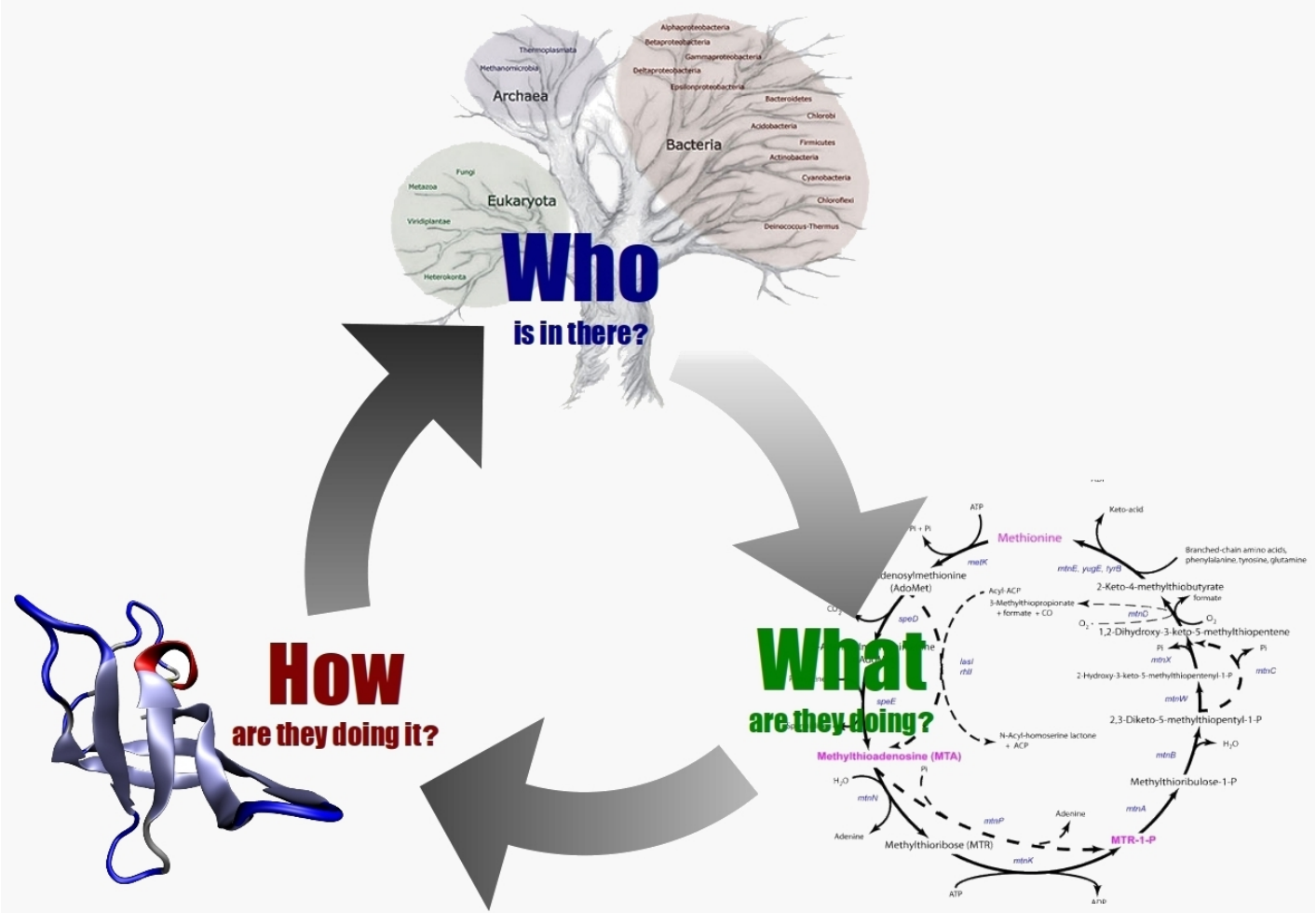


Image source: Center for Biological Sequence Analysis at the Technical University of Denmark

Metagenomics

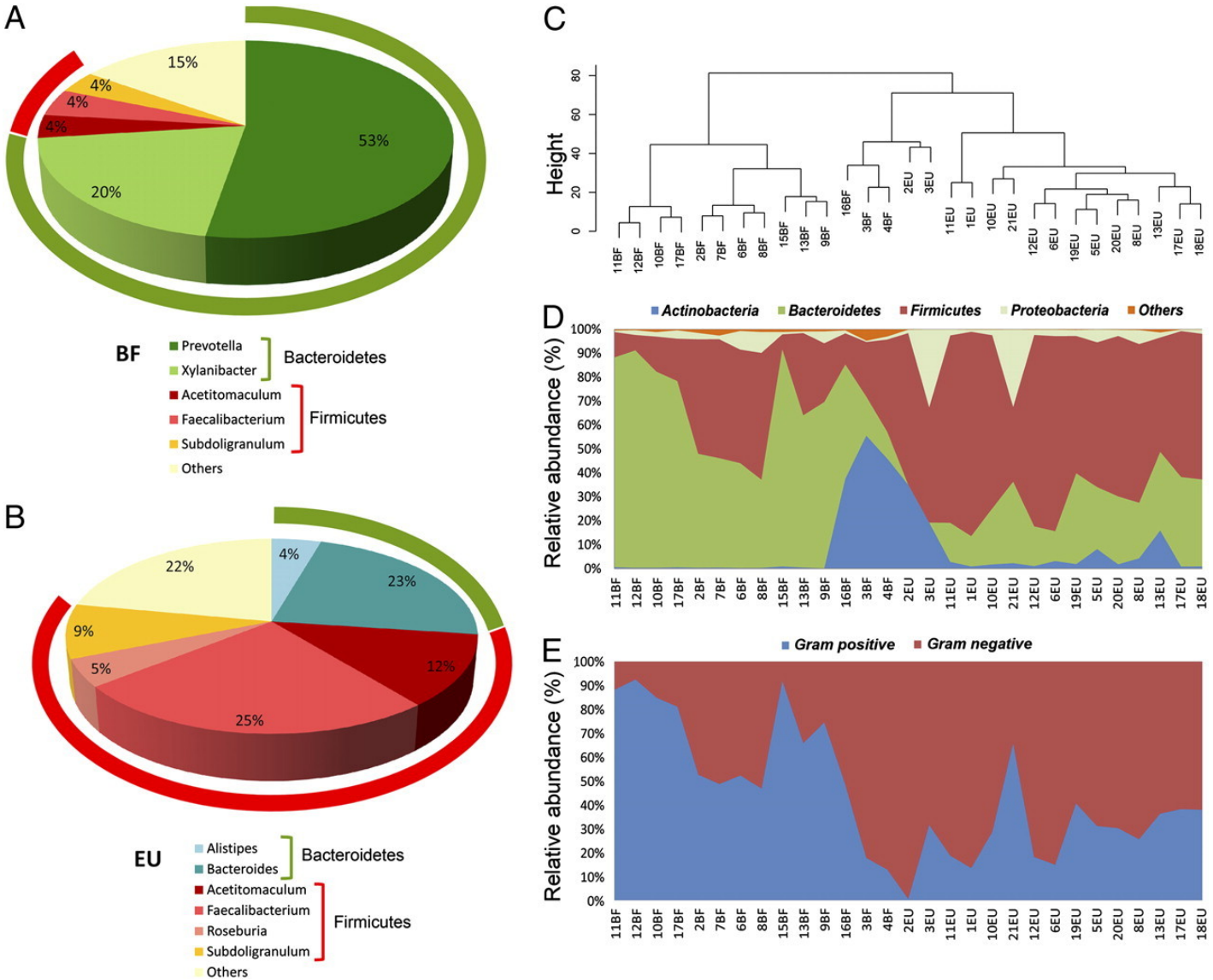


Image source: Filippo *et al.* (2010) *PNAS.*

Computer lab

- biolc100 the password is saturday28 (additional usernames: biolc101-biolc131)
- data is all available from http://tinyurl.com/uckatoa
- $-36^o\ 53'\ 9''\ S,\ 175^o\ 49'\ 21''\ E$ Hot Water Beach
- $-36^o\ 47'\ 37.66''\ S, 175^o\ 52'\ 37.82''\ E$ Mutton bird burrow
- $-36^o\ 52'\ 29''\ S, 174^o\ 46'\ 26''\ E$ Epsom Girls High