



# ***Plant Bacteriology***

## **Bacterial Phylogeny**

Compiled by N. Hassanzadeh

Version 4.25

January 1, 2025

Website Address:

<http://www.phytobacteriology.com>



# Contact address

---

- Department of Plant Protection, Faculty of Agricultural Sciences and Food Industries, Science & Research Branch, Islamic Azad University, Tehran-Iran.
- P.O. Box: 14155/775, Postal Code: 1477893855
- Branch website: [www.srbiau.ac.ir](http://www.srbiau.ac.ir)
- **e-mail addresses:**
- [hasanzadehr@srbiau.ac.ir](mailto:hasanzadehr@srbiau.ac.ir)
- [hasanzadehr@yahoo.com](mailto:hasanzadehr@yahoo.com)



# Table of Contents

---

- Books on plant bacterial phylogeny
- Proceedings/Reviews/Monographs/Book chapters/PowerPoints/PDF files
- **What is phylogeny?**
- Phylogenetic relationships
- **A brief history of origin of life**
- Evolution of the earth and earliest life forms
- Primitive organisms and metabolic strategies
- Evolutionary history of life bacteria
- **Primitive organisms and molecular coding**
- 1. **Pre-RNA World : PNA/TNA/GNA world**
- RNA World
- DNA/Protein World
- Proto-cell world
- 2. **The modern cell: DNA  $\Rightarrow$  RNA  $\Rightarrow$  Protein**
- **Phylogenetic taxonomy**
- Brief history of molecular phylogenetics
- Taxonomy **vs** phylogeny
- Prokaryotic phylogeny
- Critical issues in bacterial/prokaryotic phylogeny
- 16S rRNA-based trees



# Table of Contents

---

- Microarray technology- A modern method for detection and hierarchical studies
- Chemical and Molecular Approaches in Bacterial Phylogeny
- A natural system of classification- **History of descent**
- Five-kingdom Classification
- Three domains Woesean Universal Tree of Life
  - Bacteria*** (Eubacteria)
  - Archaea*** (Archaeobacteria)
  - Eukarya*** (Eukaryotes)
- **LUCA:** Last Universal Common Ancestor
- Characteristics of the domains:
  1. Archaea
  2. Eukarya
  3. Bacteria
- The 5 major classes of proteobacteria in domain Bacteria
- **Some selected plant diseases caused by Proteobacteria**





# Table of Contents

---

- Endosymbiosis theory for eukaryote origin(Mitochondria and chloroplasts endosymbiotic theory)
- Three alternate hypotheses of eukaryotic and prokaryotic evolution
- Web and Network Model
- Other models based on 16S rRNA sequences: Independent analyses that either confirm or refute the rRNA (Woesean tree)
  - The analysis of Leart *et al.*,2003
  - Gupta's indel analysis,1998
  - Brochier and Philippe,2002
  - Cavalier-Smith megaclassification,2002
  - Arthur L. Koch,2003 argues the first cells: Gram-positive or Gram-negative?
  - Rivera & Lake Circle life tree,2004
  - Lake and colleague's two domains Eocyte hypothesis, 1984
- Two domains universal tree of life: update of Woesean Universal Tree of Life, based on 16S rRNA sequences:
  - Bacteria**
  - Arkarya** (a new name proposed for the clade grouping Archaea and Eukarya)
  - Ruggiero *et al.*,2015



# Table of Contents

---

- **Major topics in practical phylogeny**
- Mutation rate (DNA or protein mutation)
- Molecular chronometers - An evolutionary clocks
- Homoplasy and long branches
- Gene trees **vs.** species trees
- Assessing sequence quality: Chromas, BioEdit,...
- Sequence alignments
- BLAST
- **Phylogenetic methods can be divided into three general categories:**
  - Maximum Parsimony
  - Maximum likelihood
  - Maximum distance
- **Distance based tree reconstruction:**
  - UPGMA algorithm
  - Neighbor-joining algorithm
  - Bootstrapping algorithm
- Evaluation of tree reproducibility
- Interpreting phylogenetic tress



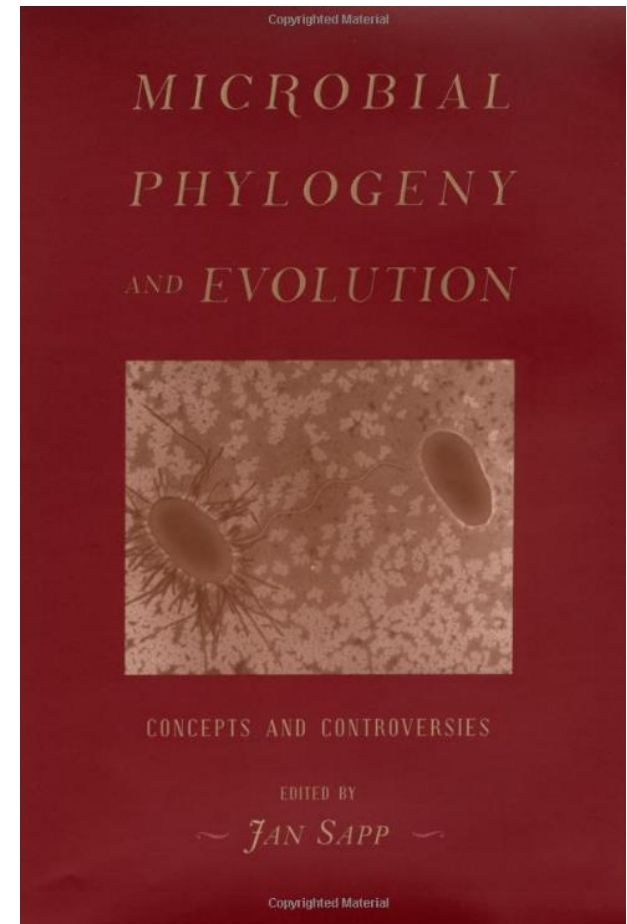
# Table of Contents

---

- Examples of Phylogenetic analyses
- *Acidovorax*
- Coryneforms
- *Erwinia*
- *Pseudomonas*
- Rhizobia
- *Xanthomonas*
- *Xylella fastidiosa*
- Mollicutes (*Spiroplasma*, *Phytoplasma*,..)
- Glossary of general terms
- Selected References

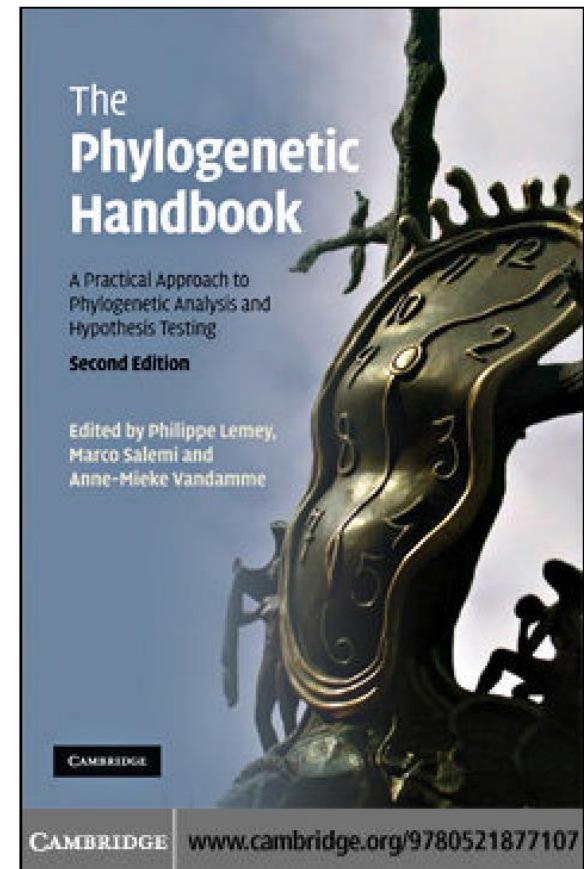
# Microbial Phylogeny and Evolution

- Microbial Phylogeny and Evolution
- Jan Sapp (ed.)
- 2005
- Oxford University Press
- 326 pp.



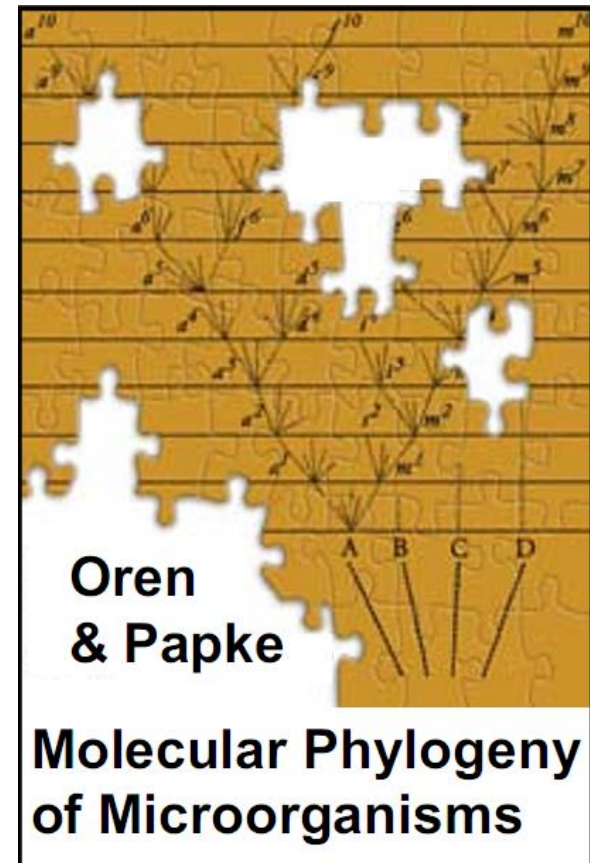
# The Phylogenetic Handbook-A Practical Approach to Phylogenetic Analysis and Hypothesis Testing

- The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing.
- Philippe Lemey, M. Salemi and A.M. Vandamme (Eds.)
- Cambridge University Press
- Second edition, 2009
- 723 pp.



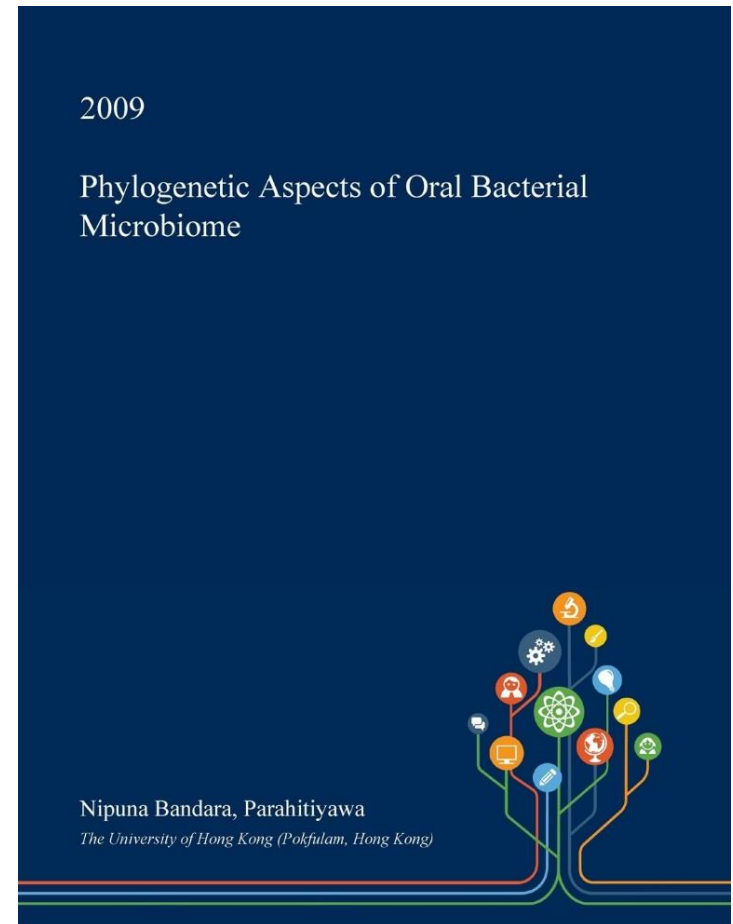
# Molecular Phylogeny of Microorganisms

- Molecular Phylogeny of Microorganisms
- Editor: Aharon Oren and R. Thane Papke
- Publisher: Caister Academic Press
- 2010
- c. 220 pages.



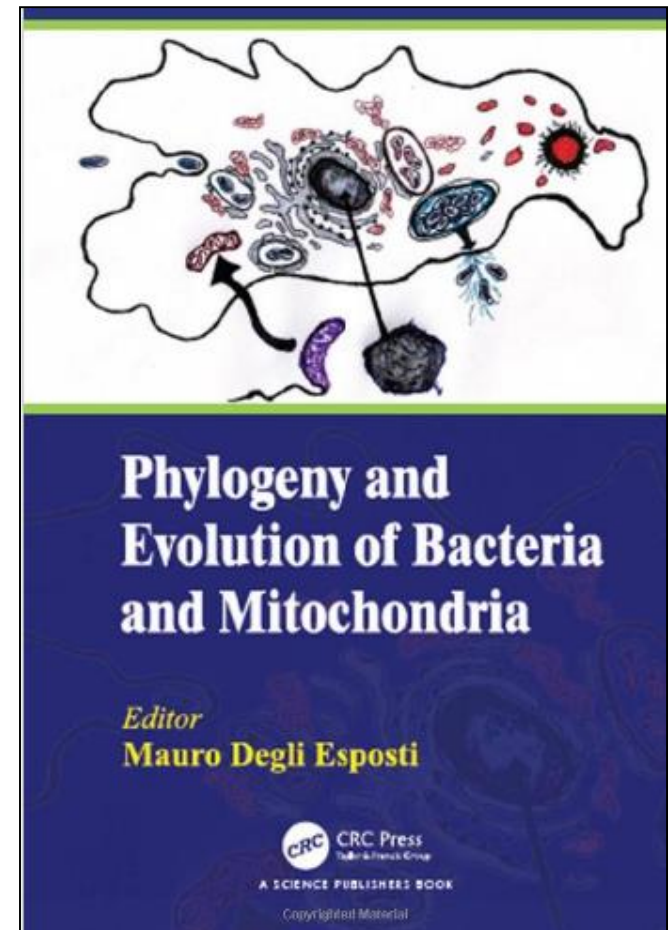
# Phylogenetic Aspects of Oral Bacterial Microbiome

- Phylogenetic Aspects of Oral Bacterial Microbiome
- Nipuna Bandara, Parahitiyawa
- Dissertation
- Open Dissertation Press
- 2009



# Phylogeny and Evolution of Bacteria and Mitochondria

- Phylogeny and Evolution of Bacteria and Mitochondria
- Editor: **Mauro Degli Esposti**
- CRC Press
- 2018
- 236 pages.

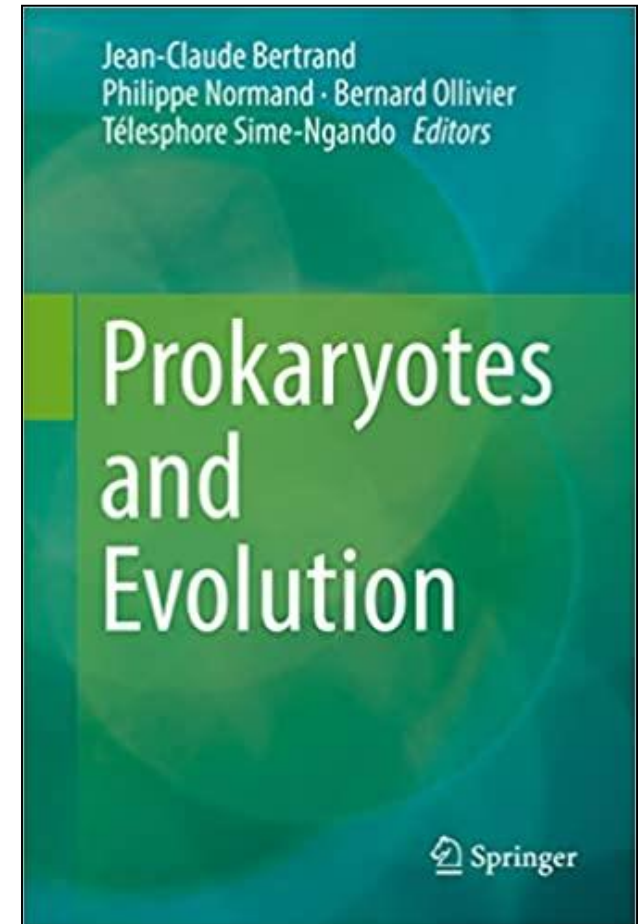






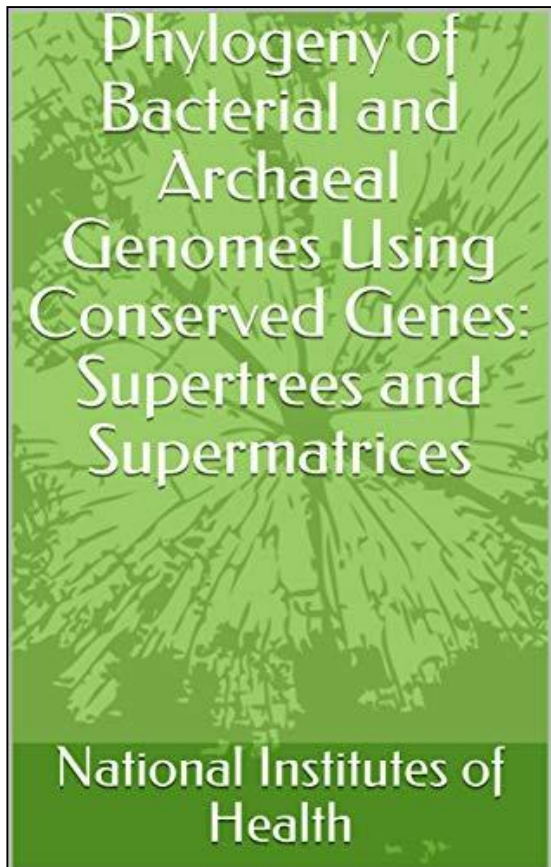
# Prokaryotes and Evolution

- Prokaryotes and Evolution
- Jean-Claude Bertrand, Philippe Normand, Bernard Ollivier, Télecphore Sime-Ngando (Editors)
- Springer
- 2019
- 405 pages.



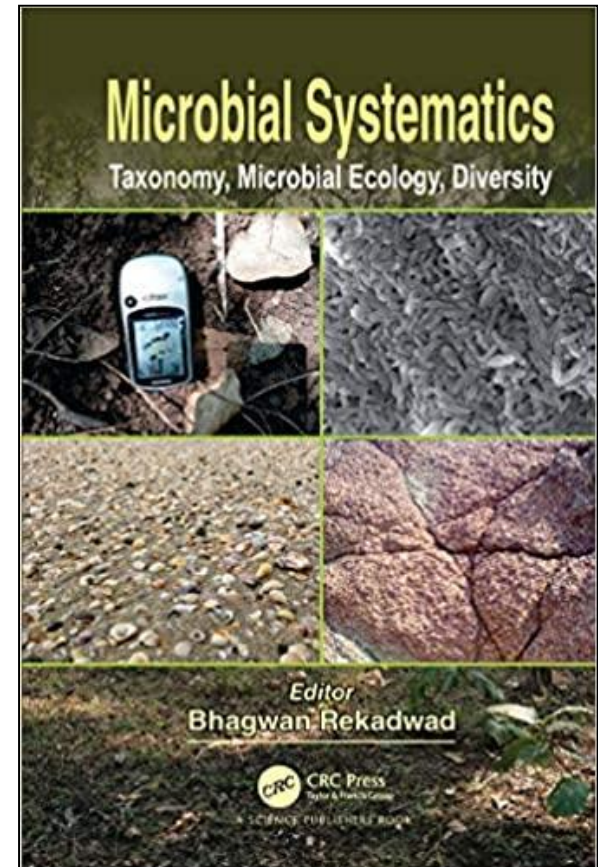
# Phylogeny of Bacterial and Archaeal Genomes Using Conserved Genes: Supertrees and Supermatrices

- Phylogeny of Bacterial and Archaeal Genomes Using Conserved Genes: Supertrees and Supermatrices
- National Institutes of Health (Author)
- 2020
- 36 pp.



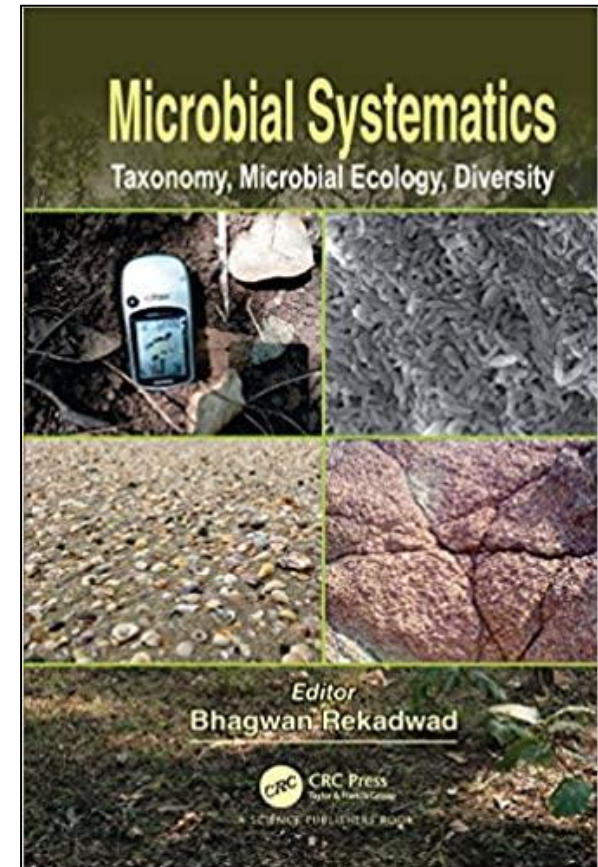
# Microbial Systematics: Taxonomy, Microbial Ecology, Diversity

- **Microbial Systematics: Taxonomy, Microbial Ecology, Diversity**
- **by Bhagwan Rekadwad**
- **CRC Press**
- **2020**
- **218 pp.**



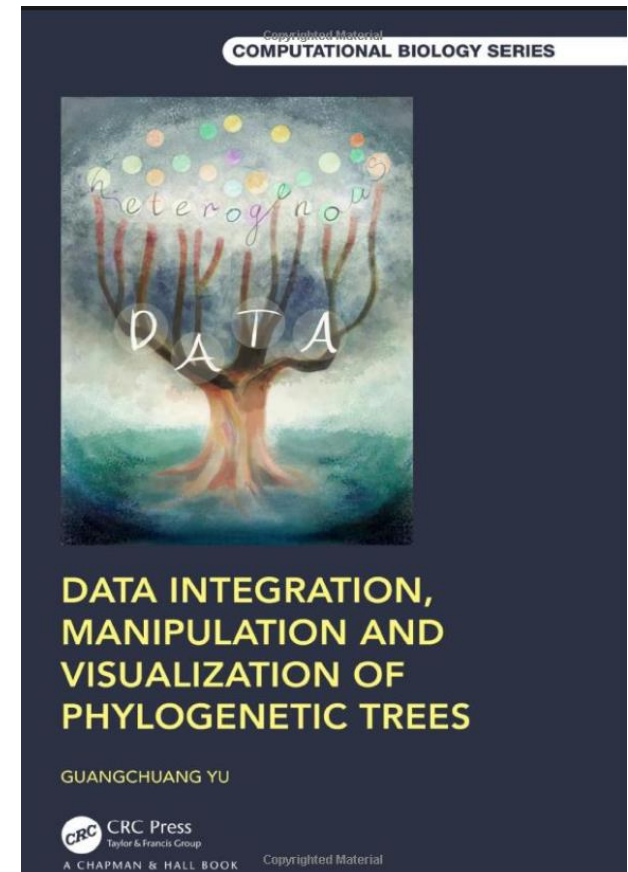
# Classification of 16S rRNA reads is improved using a niche-specific database constructed by near-full length sequencing

- **Classification of 16S rRNA reads is improved using a niche-specific database constructed by near-full length sequencing.**
- **by Bhagwan Rekadwad**
- **Publication date: 2020**
- **218 pp.**



# Data Integration, Manipulation and Visualization of Phylogenetic Trees

- **Data Integration, Manipulation and Visualization of Phylogenetic Trees**
- by **Guangchuang Yu**
- Publication date: 2020
- Chapman and Hall/CRC.
- 255 pp.







# PowerPoints/PDF files

---

- **Buchanan, R.E.1918. Bacterial Phylogeny as Indicated by Modern Types.** The American Naturalist 52, no. 616/617 (Apr. - May, 1918): 233-246.
- **Diversity\_1. pdf.2004. Microbial Diversity.**
- **Duffy, B, C. Pelludat, F. Pasquier, B. Frey, J.E. Frey. 2008. Diagnostic Microarrays.** COST 873 Training Workshop, CSL, York UK.
- **Eisen *et al.*,2004. Phylogenomics: A Genome Level Approach to Assembling the Bacterial Branches of the Tree of Life.** <http://www.tigr.org/tol>. 24 slides.
- **Hoekstra, 2005. Chap13. Phylogeny and systematics.** Pages 303-330. pdf.
- **Holmes I. Phylogenetic trees.** BioE131/231. 20 slides.
- **Holmes I. Alignment basics.** 27 slides.
- **Jones, J.B. 2006. Lecture 1 phylobacteriology.** 19 pages.
- **Karlm.2004. Taxonomy.**
- **Luskin, C. Problems with the Natural Chemical "Origin of Life".** As found on the IDEA Center website at [www.ideacenter.org](http://www.ideacenter.org).
- **Ong, Han Chuan. Ocean 339D Phylogenetic Lecture.** 23 slides. [hanong@u.washington.edu](mailto:hanong@u.washington.edu).
- **Parkinson, N. Identifying Relatedness Between Bacterial Plant Pathogensm.** [Parkinson\\_MolID\\_CSL\\_1.pdf](#).
- **Parks, D.H., M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, P-Alain Chaumeil, and P. Hugenholtz.2018. A proposal for a standardized bacterial taxonomy based on genome phylogeny.** bioRxiv preprint.



# PowerPoints/PDF files

---

- **Pun ,P. 2011. The Three Domains of Life: A Challenge to the concept of the Universal Cellular Ancestor? 50 slides.**
- **Spiegel, N. 2007. Searching Sequence Databases. Wiley Publishing. 33 slides.**
- **Stead, D. Classification and Nomenclature of Plant Pathogenic Bacteria - A Review (m\_stead\_CSL\_1. pdf).**
- **Wilke,T. CS 177. Phylogenetics I. 44 slides.**
- **De La Fuente, Leonardo.2009. Introducción. Auburn University. 82 slides.**
- **Ehdieh Khaledian, Kelly A. Brayton and Shira L. 2020. A Systematic Approach to Bacterial Phylogeny Using Order Level Sampling and Identification of HGT Using Network Science. Microorganisms 8 (2), 312.**
- **Wang, J. 2022. Methods and Applications in Molecular Phylogenetics. Bioinformatics 21 (10), 2329-2335.**
- **Kahn, A.K. and, Almeida, R. P. P. 2022. Phylogenetics of Historical Host Switches in a Bacterial Plant Pathogen. Applied and Environmental Microbiology Vol. 88, No. 7.**



# The origin of life

## Understanding the origins of life on earth

---

- The origin of life on Earth is a relatively poorly understood area of science.
- Complex organic molecules arose from the “primordial soup” which would eventually make possible the abundant variety of organisms, tissues, cellular structures and biological processes that exist today.

**Primordial:** having existed from the beginning; in an earliest or original stage or state.





# What is Phylogeny?

## Systematics or phylogeny

---

- The study of the evolutionary history of organisms.
- To identify all species of life on Earth.



# Phylogeny

## Common ancestor

---

- Biologists estimate that there are about 5 to 100 million species of organisms living on Earth today.
- All organisms evolved from common ancestor:
  1. Similar plasma membrane;
  2. Use ATP for energy;
  3. DNA is genetic storage.

# Phylogeny

## Tree of Life

### Horizontal gene transfer

---

- Evidence from morphological, biochemical, and gene sequence data suggests that:
  1. All organisms on earth are genetically related, and
  2. The genealogical relationships of living things can be represented by a vast evolutionary tree, the Tree of Life.

# Phylogeny

## Tree of Life

### Horizontal gene transfer

---

- The Tree of Life then represents the phylogeny of organisms i.e. the history of organismal lineages as they change through time.
- It implies that:
  1. different species arise from previous forms via descent, and
  2. that all organisms, from the smallest microbe to the largest plants and vertebrates, are connected by the passage of genes along the branches of the phylogenetic tree that links all of Life.

# Phylogeny

## Common ancestor

### Phylogenetic modeling concepts

---

1. Phylogenetic modeling concepts are constantly changing.
2. It is one of the most dynamic fields of study in all biology.
3. Over the last several decades, new research has challenged scientists' ideas about how organisms are related.
4. Many phylogenetic trees are models of the evolutionary relationship among species.

# Phylogeny

## Common ancestor

### Classical phylogenetic modeling concepts

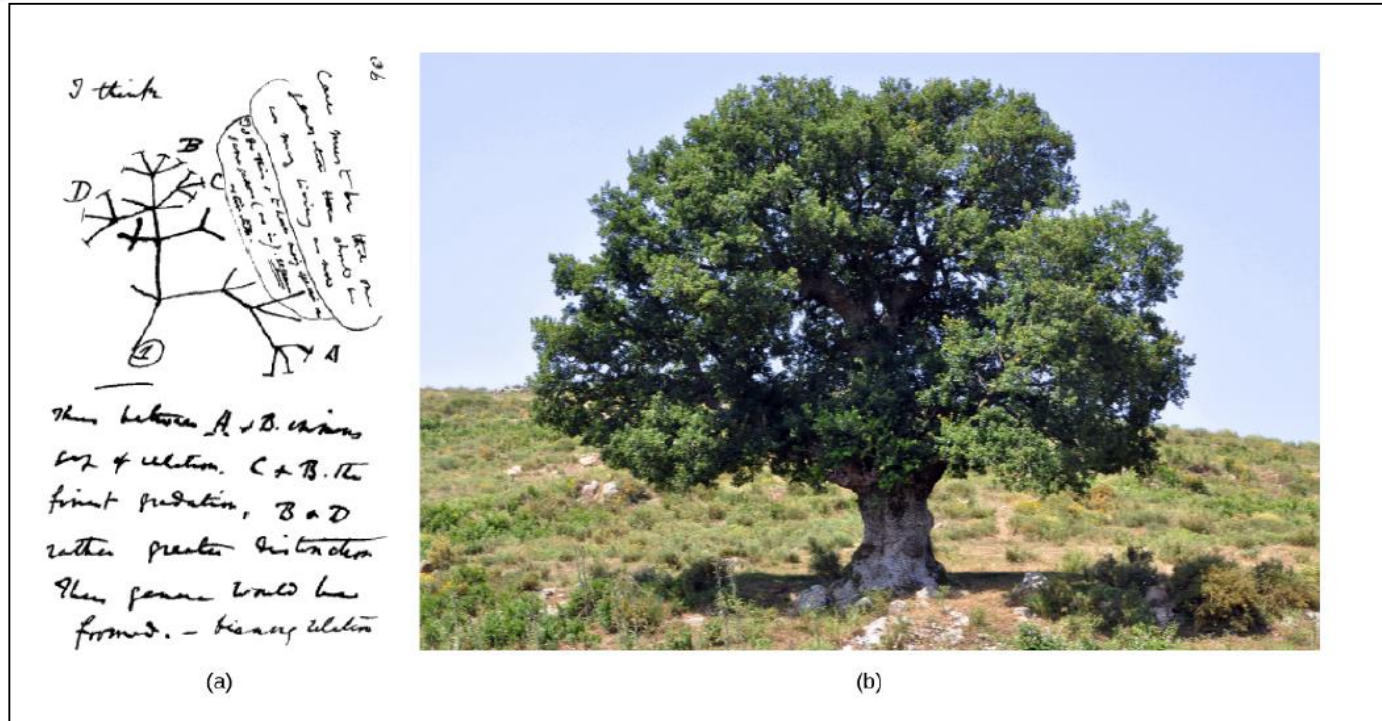
---

- The phylogenetic tree concept with a single trunk representing a common ancestor, with the branches representing:
  1. the divergence of species from this ancestor,
  2. fits well with the structure of many common trees, such as the oak.

# Phylogeny

## Tree of Life

### Classical phylogenetic modeling concepts



The (a) concept of the “tree of life” dates to an 1837 Charles Darwin sketch. Like an (b) oak tree, the “tree of life” has a single trunk and many branches.

# Phylogeny

## Tree of Life

### Modern phylogenetic tree of life

---

- Classical thinking about **prokaryotic** evolution, included in the **classic tree model**, is that species **evolve clonally**.
- Scientists did not consider the **concept of genes transferring between unrelated species** as a possibility until **relatively recently**.
- Horizontal gene transfer (HGT), or lateral gene transfer, is the **transfer of genes between unrelated species**.



# Modern phylogenetic tree of life

## Horizontal Gene Transfer

### Origin of eukaryotes

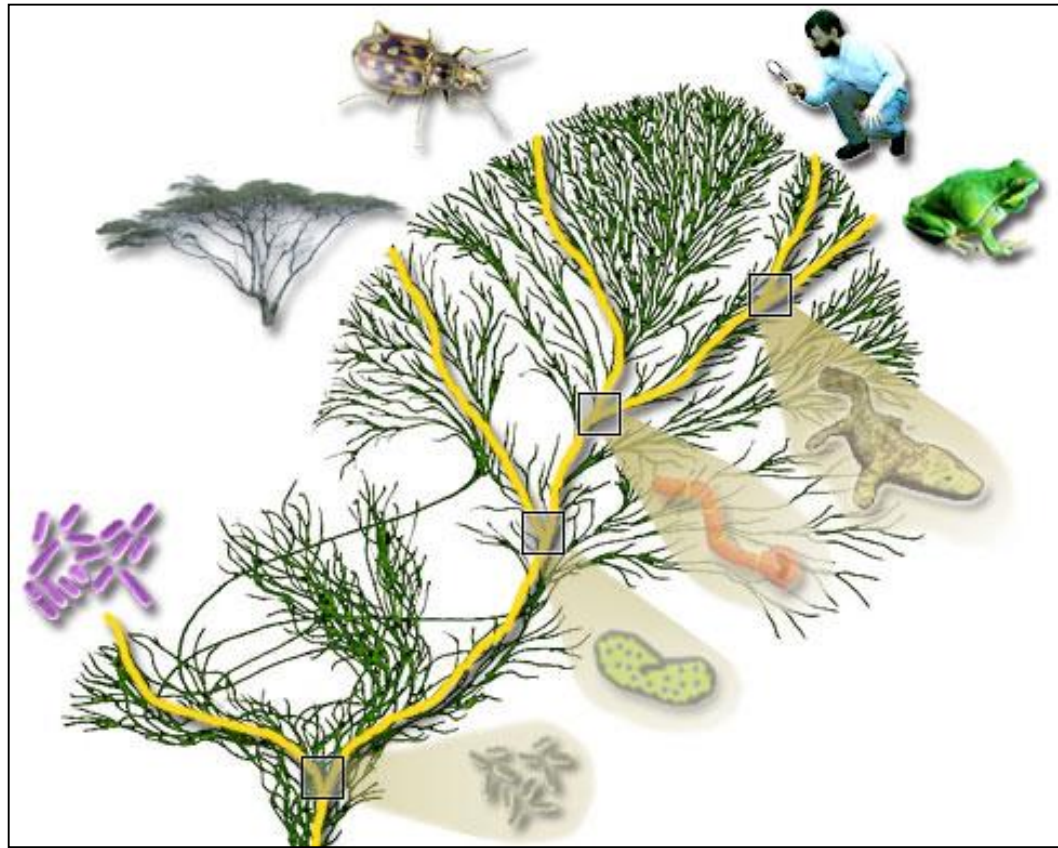
---

1. Although it is likely that single celled Eukaryotes were also present on Earth from the very beginning,
  2. there is also considerable evidence that Archae, Bacteria, and Viruses transferred genes to these single celled Eukaryotes, thus trigger multi-cellularity (Joseph 2009b,c).
- Thus we see that the genomes of modern day eukaryotic species, including humans, contain highly conserved genes were acquired from Archae and Bacteria.

# Phylogeny

## Tree of Life

### Modern phylogenetic tree of life



All organisms are connected by the passage of genes along the branches of the phylogenetic Tree of Life.

# Phylogeny

## Modern phylogenetic tree of life

### Prokaryotic and Eukaryotic HGT Mechanisms Summary



	Mechanism	Mode of Transmission	Example
<b>Prokaryotes</b>	transformation	DNA uptake	many prokaryotes
	transduction	bacteriophage (virus)	bacteria
	conjugation	pilus	many prokaryotes
	gene transfer	agents phage-like particles	particles purple non-sulfur bacteria
<b>Eukaryotes</b>			
	from food organisms	unknown	aphid
	jumping genes	transposons	rice and millet plants
	epiphytes/parasites	unknown	yew tree fungi
	from viral infections		

# Phylogeny

## Modern phylogenetic tree of life

### HGT occurs in prokaryotes

---

- HGT is an ever-present phenomenon, with many evolutionists postulating a major role for this process in evolution, thus complicating the simple tree model.
- Genes pass between species which are only distantly related using standard phylogeny, thus adding a layer of complexity to understanding phylogenetic relationships.
- The various ways that HGT occurs in prokaryotes is important to understanding phylogenies.

# Phylogeny

## Modern phylogenetic tree of life

### HGT occurs in prokaryotes

- HGT mechanisms are quite common in the Bacteria and Archaea domains, thus significantly changing the way scientists view their evolution.
- The majority of evolutionary models, such as in the Endosymbiont Theory, propose that eukaryotes descended from multiple prokaryotes, which makes HGT all the more important to understanding the phylogenetic relationships of all extant and extinct species.

# Phylogeny

## Modern phylogenetic tree of life

### HGT occurs in prokaryotes

- The **Endosymbiont Theory** purports that the eukaryotes' mitochondria and the green plants' chloroplasts and flagellates originated as **free-living prokaryotes** that **invaded primitive eukaryotic cells** and become established as permanent symbionts in the cytoplasm.

For more information about **Endosymbiont Theory**, see Slides 220 and above.

# Phylogeny

## Modern phylogenetic tree of life

### HGT occurs in prokaryotes

- Microbiology students are well aware that genes transfer among common bacteria.
- These gene transfers between species are the major mechanism whereby bacteria acquire resistance to antibiotics.
- Classically, scientists believe that three different mechanisms drive such transfers.
  1. **Transformation**: bacteria takes up naked DNA
  2. **Transduction**: a virus transfers the genes
  3. **Conjugation**: a hollow tube, or pilus transfers genes between organisms.

# Modern phylogenetic tree of life

## HGT occurs in eukaryotes

### Origin of eukaryotes

---

- Although HGT mechanisms are quite common in the Bacteria and Archaea domains, but some do not view HGT as important to eukaryotic evolution.
- **HGT does occur in Eukarya domain as well.**
- Genes transferred to the eukaryotic genome by prokaryotes and Viruses, include:
- exons, introns, transposable elements, informational and operational genes, RNA, ribozomes, mitochondria, and the core genetic machinery for translating, expressing, and repeatedly duplicating genes and the entire genome.

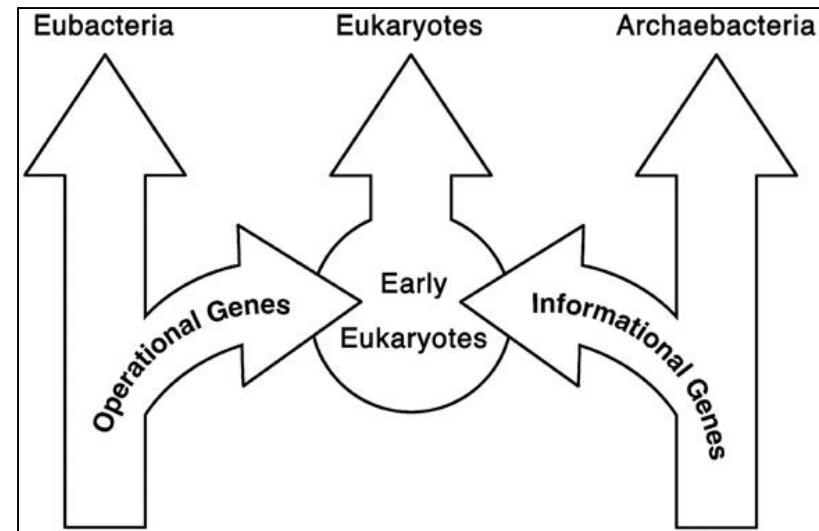


# Modern phylogenetic tree of life

## Horizontal Gene Transfer

### Origin of eukaryotes

- Almost all scientists will agree that modern day life can trace its genetic ancestry to the first life forms to appear on Earth.
- These first Earthlings (Archae, Bacteria, and their viral genetic luggage/baggage) contained the genes and genetic information for:
  1. altering the environment,
  2. the "evolution" of multicellular Eukaryotes, and
  3. the metamorphosis of all subsequent species (Joseph 2009b,c).



### Illustrate how prokaryotes and eukaryotes transfer genes horizontally

Metamorphosis is a process by which animals undergo extreme, rapid physical changes some time after birth.

# Modern phylogenetic tree of life

## Horizontal Gene Transfer

### Origin of eukaryotes from prokaryotes rather Archaea

- As a consequence of this modern DNA analysis, the idea that **eukaryotes evolved directly from Archaea** has **fallen out of favor**.
- While eukaryotes share many features that are absent in bacteria, such as the **TATA box** (located in many genes' promoter region), the discovery that some eukaryotic genes were more homologous with **bacterial DNA than Archaea DNA** made this idea less tenable.
- Furthermore, scientists have proposed **genome fusion from Archaea and Bacteria by endosymbiosis** as the ultimate event in **eukaryotic evolution**.

# Modern phylogenetic tree of life

HGT occurs in eukaryotes

Origin of eukaryotes

---

- Not all of these genes have been expressed, whereas yet other were silenced or activated in response to specific environmental signals, thereby giving rise to new species (Joseph 2000, 2009b,c).

# Phylogeny

## Tree of Life

### Major branches of tree of life

---

- The Tree of Life on planet Earth begins about 3.7 billion years ago.
- There are three major branches:
  1. The Bacteria;
  2. The Archaea, and
  3. The Eukaryota.
- The Bacteria are common prokaryotes living in virtually all environments.
- They include:
  1. The human gut commensal *Escherichia coli*,
  2. Soil bacteria like *Bacillus subtilis*,
  3. Pathogens like *Salmonella*, *Agrobacterium*.

A billion is 1 000 000 000 (a thousand million or more rarely milliard).



# A Brief History of Origin of Life

---

1. **Evolution of the Earth and Earliest Life Forms**
2. **Primitive Organisms and Metabolic Strategies**



# A Brief History of Origin of Life

---

## **A. Evolution of the Earth and Earliest Life Forms:**

- Origin of the earth;
- Evidence for microbial life on the early earth;
- Conditions on early earth;
- Origin of life.

## **B. Primitive Organisms and Metabolic Strategies:**

- Metabolism of primitive organisms;
- Further metabolic evolution and photosynthesis: oxygenation of the atmosphere.

## **C. Primitive Organisms and Molecular Coding:**

- From RNA world to DNA/protein world.

# The origin of life

## Intelligent life

**Hypothetical birth date of 13.6 bya for the beginning of life**

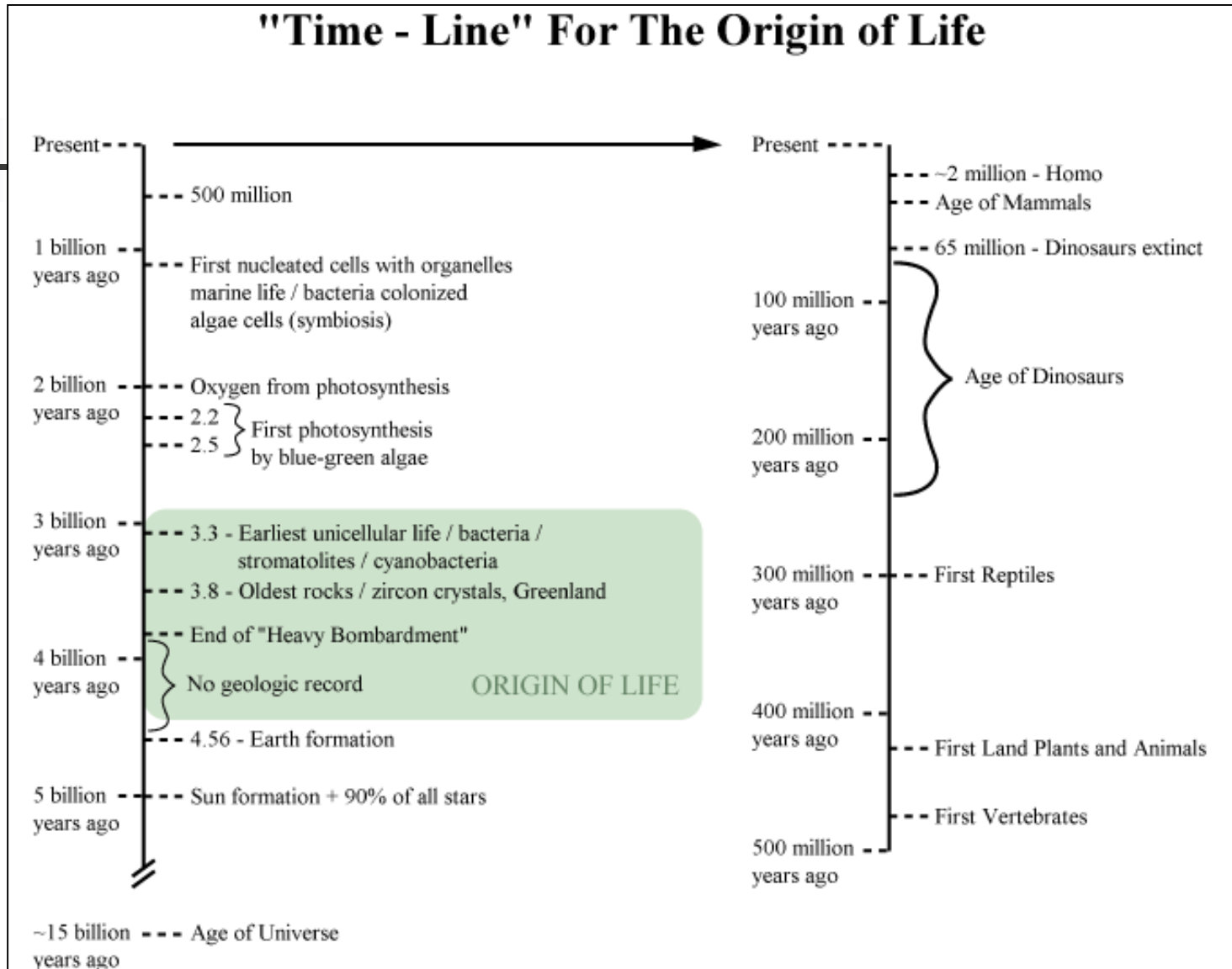
---

- If **we ignore the reality of an infinite universe**, and pick a **hypothetical birth date of 13.6 bya for the beginning of life**, and using the evolution of life on Earth as an example, then it could also be predicted that sentient, **intelligent life would have evolved on numerous Earth-life planets by 9 bya**.
- This could mean that the genetic template for the evolution of all manner of life, including those similar to **humans**, would have been established **almost 5 billion years before Earth became Earth**.

# History of life on earth

15 billion years ago age of Universe

## "Time - Line" For The Origin of Life





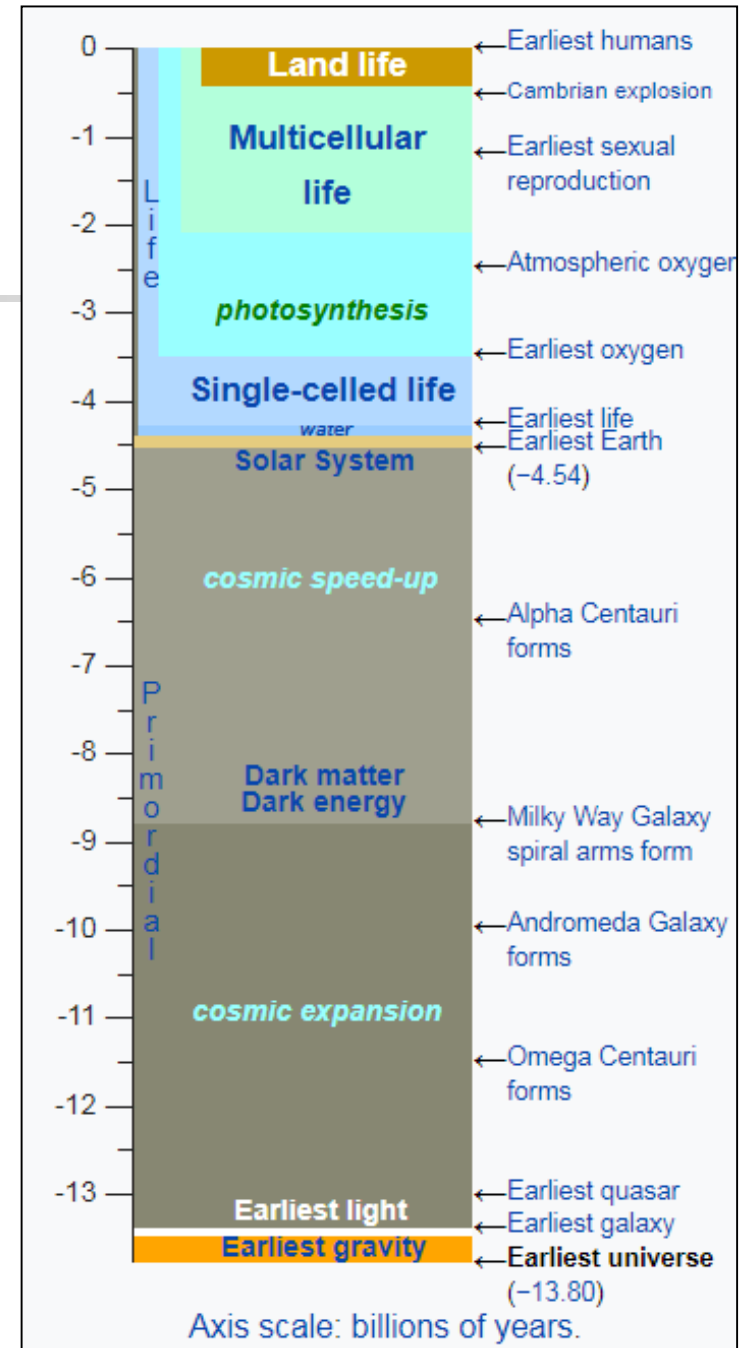
# Age of the universe

## Chronology of the universe

### Nature timeline

- Detailed measurements of the expansion rate of the universe place this moment at approximately 13.8 billion years ago, which is thus considered the **age of the universe**.
- After the initial expansion, the universe cooled sufficiently to allow the formation of subatomic particles, and later simple atoms.
- Giant clouds of these primordial elements later coalesced through gravity in halos of dark matter, eventually forming the stars and galaxies visible today.

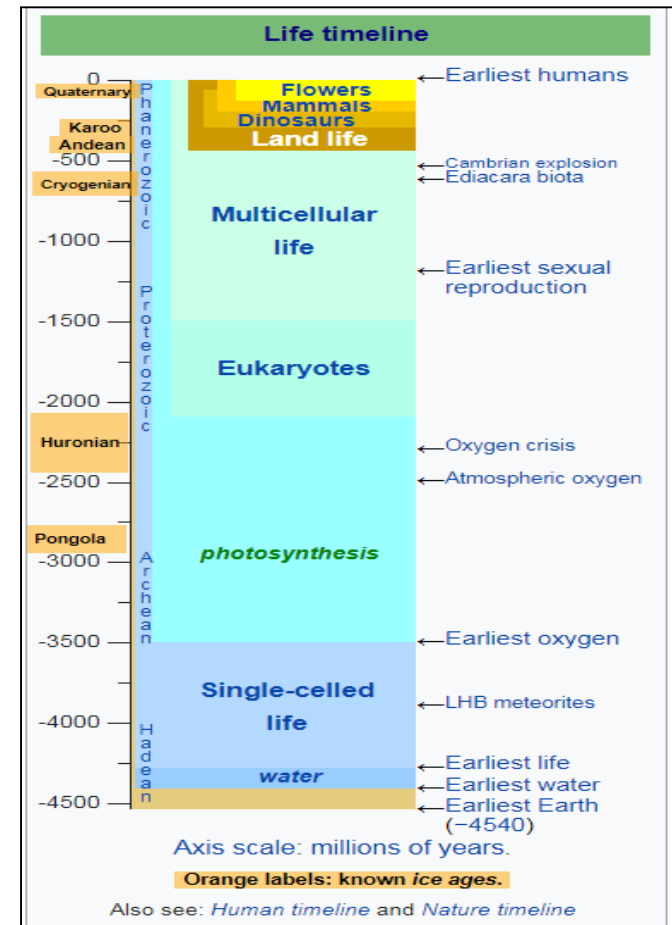
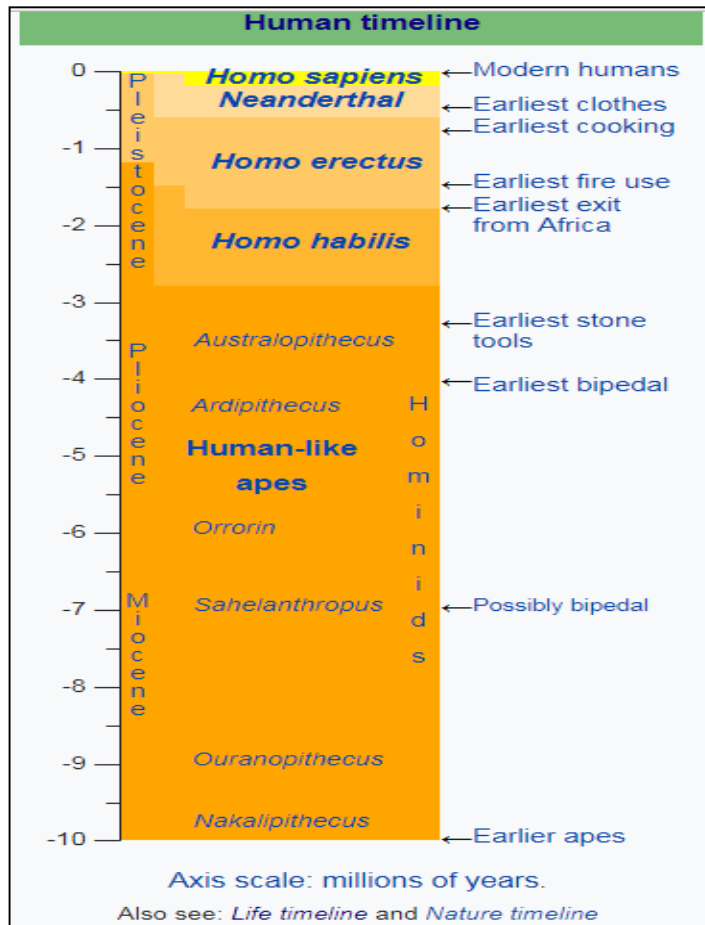
Wikipedia, 2017



# Age of the universe

## Nature timeline

## Life and human timeline





# The origin of life

## Origin of life vs. evolution of life

---

- The "origin of life" (OOL) is best described as the chemical and physical processes that brought into existence the first self-replicating molecule.
- It differs from the "evolution of life" because Darwinian evolution employs mutation and natural selection to change organisms, which requires reproduction.
- Since there was no reproduction before the first life, no "mutation - selection" mechanism was operating to build complexity.
- Hence, OOL theories cannot rely upon natural selection to increase complexity and must create the first life using only the laws of chemistry and physics.

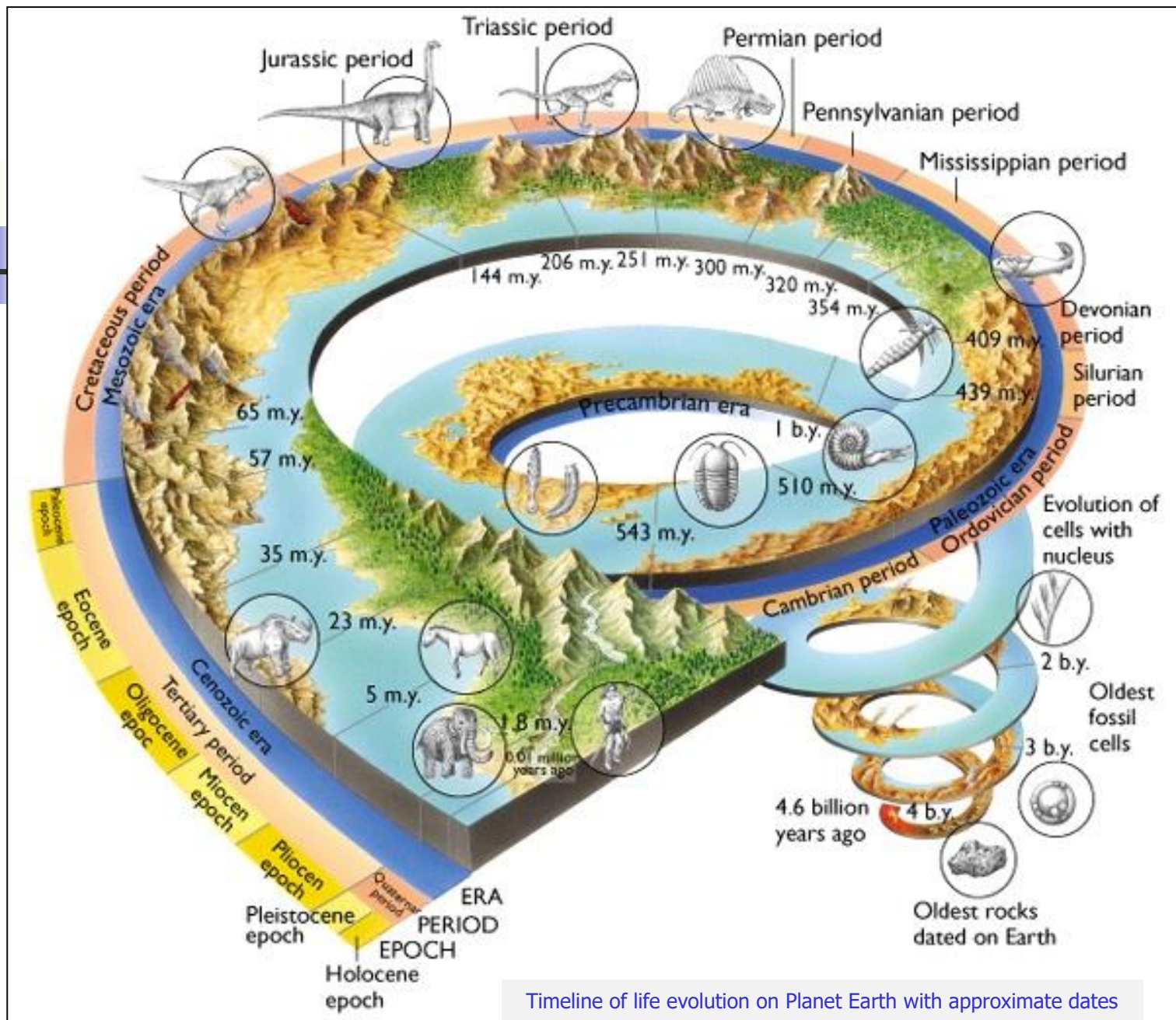


# Origin of the earth/life

## Origin of life (OOL)

---

- **Bacteria** lived as early as 3.5 billion years ago.
- The evolutionary history of life, spanning a period of more than 3.5 billion years (Giga annum or Ga).
- Given that mainstream scientists believe:
- **Earth** is about 4.54 billion years old, and that the
- **Earth's crust** did not solidify until 4 billion years ago.
- There may be as few as 200 million years allowed for the OOL.
- That may seem like a long time, but it only represents about 1/22 of the earth's total history.



Timeline of life evolution on Planet Earth with approximate dates



# Evolutionary history of life

## Geologic Time Scale

---

- **Earth:** 4.5 billion years old
- **Life:** 4 billion years
- **Vertebrates:** 500 million
- **Mammals:** 180 million
- **Man:** 3 million
- **Fire:** 500,000 years?
- **Writing:** 5,000 years.

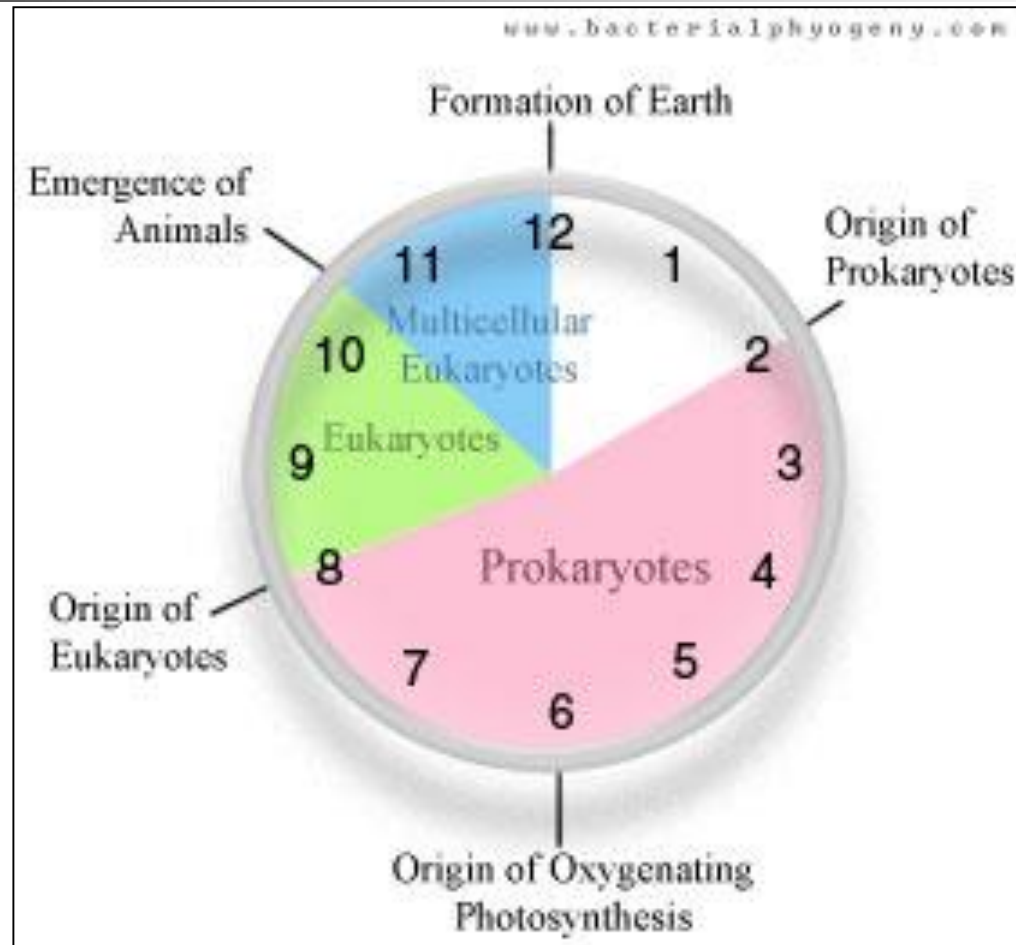
### 12 hour clock:

- **2:40 AM life began**
- 8:48 PM Cambrian explosion
- 9:20 PM vertebrates arise
- 11:02 PM mammals arise
- **11:59:02 PM man arise**
- Last 10 seconds – fire
- Last 100 msec – writing
- Last nanosec – cell phones!
- Most of the history of life.

- Most of the history of life was dominated by blue-green algae (90% of 4 billion years).
- Then sexual reproduction arose as an out come of the Cambrian Epoch (last 10%).
- This introduced biological uncertainty;
- Rapid rates of formation of new species.

# Evolutionary history of life

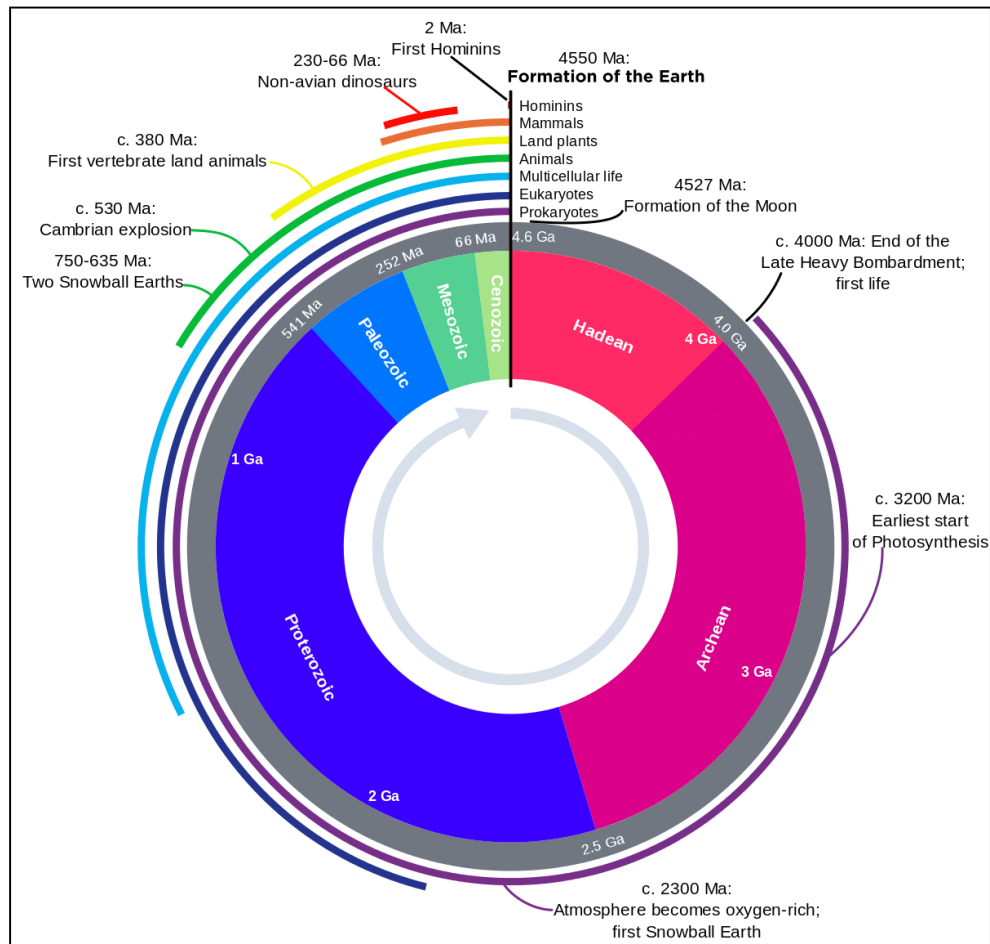
## Bacteria





# Evolutionary history of life

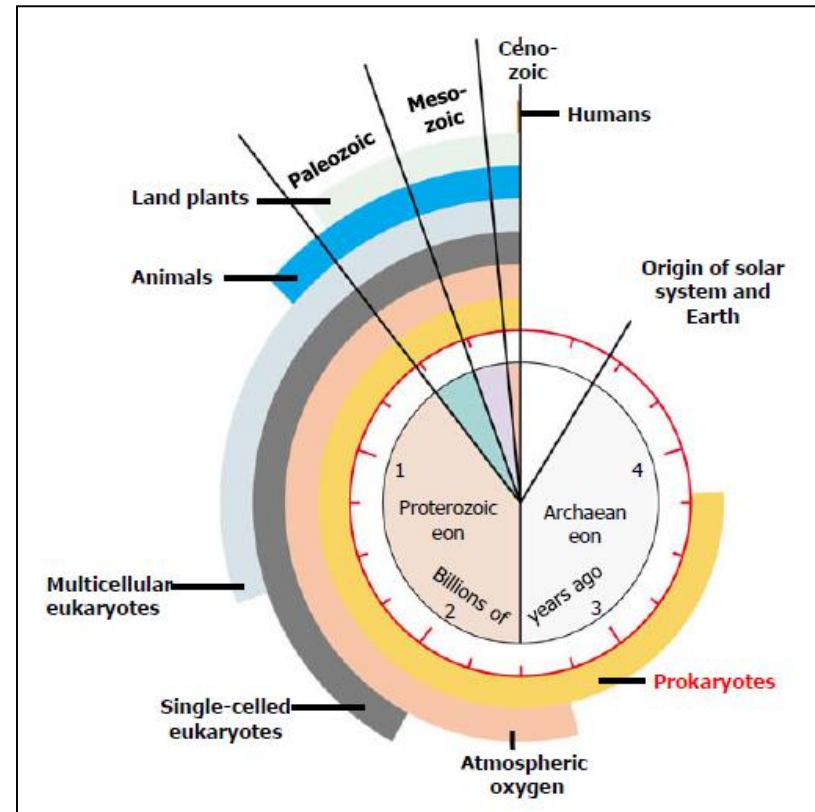
Geologic time represented in a diagram called a geological clock, showing the relative lengths of the eons of Earth's history and noting major events





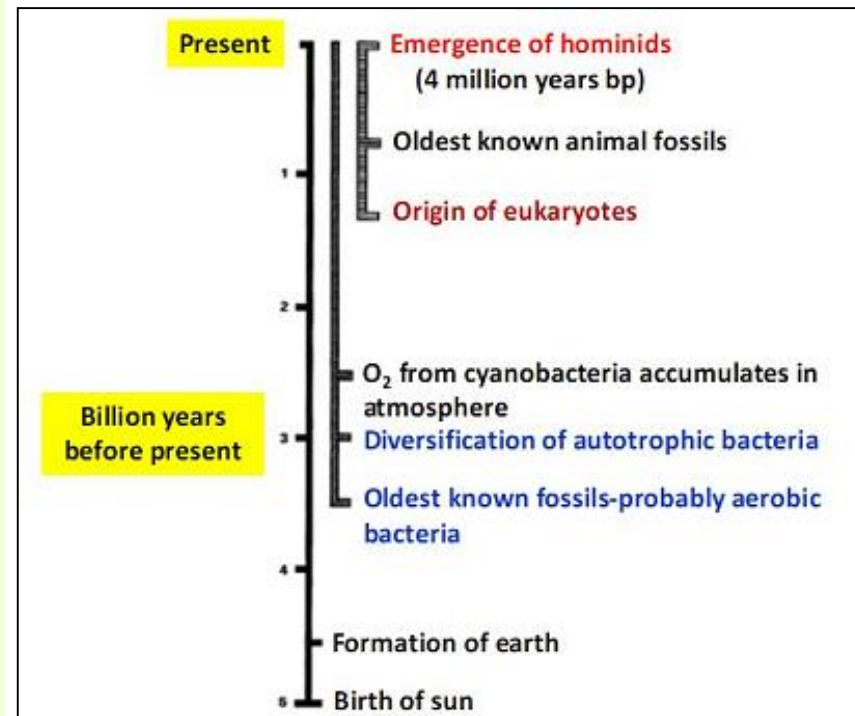
# Evolutionary history of life

- A clock analogy tracks the origin of the Earth to the present day.
- Also shows some major events in the history of Earth and its life.



# Evolution of life on earth

- Before Present (BP) years is a time scale used mainly in geology, and other scientific disciplines to specify when events in the past occurred.
- Because the "present" time changes, standard practice is to use 1 January 1950 as the origin of the age scale, reflecting the fact that radiocarbon dating became practicable in the 1950s.

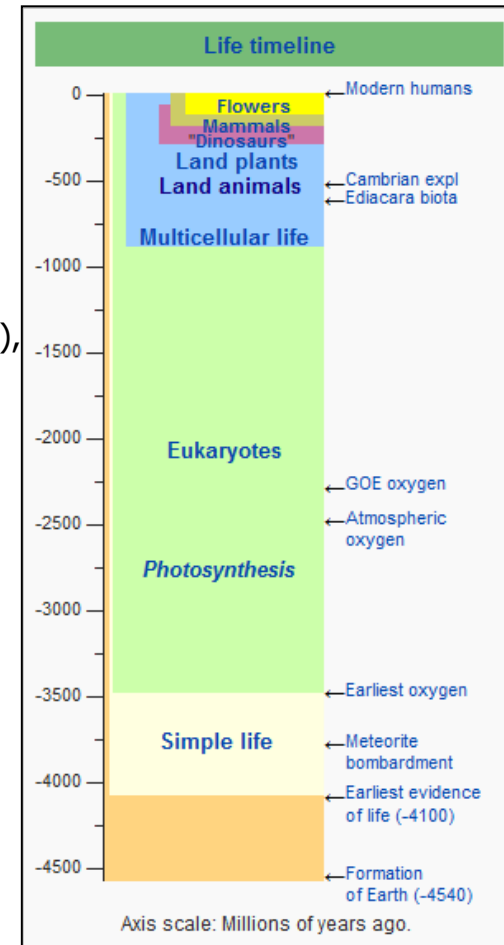


An **autotroph** is an organism that can produce its own food using light, water, carbon dioxide, or other chemicals. Because **autotrophs** produce their own food, they are sometimes called **producers**.

# Evolution of life on earth

## Basic timeline

- The **basic timeline** is a 4.5 billion year old earth with (very approximate) dates:
- **3.8 billion years of simple cells (prokaryotes),**
- 3 billion years of **photosynthesis,**
- **2 billion years of complex cells (eukaryotes),**
- 1 billion years of **multicellular life,**
- **600 million years of simple animals,**
- 570 million years of **arthropods** (ancestors of insects, arachnids and crustaceans),
- 550 million years of complex animals,
- 500 million years of **fish** and proto-amphibians,
- **475 million years of land plants,**
- **400 million years of insects and seeds,**
- 360 million years of **amphibians,**
- 300 million years of **reptiles,**
- 200 million years of **mammals,**
- 150 million years of **birds,**
- **130 million years of flowers,**
- 65 million years since the non-avian **dinosaurs died out,**
- 2.5 million years since the appearance of the **genus Homo,**
- **200,000 years since humans started looking like they do today,**
- **25,000 years since Neanderthals died out.**



# History of life on earth

Millions of years before present	Geological/fossil record [abstracted from <i>Encyclopaedia Britannica</i> , 1986]
about 4,600	<b>Planet earth formed</b>
3,500-3,400	<b>Microbial life present</b> , evidenced by <b>stromatolites</b> (sedimentary structures known to be formed by microbial communities) in some Western Australian deposits
2,800	<b>Cyanobacteria</b> (formerly called blue-green algae are relatively simple, primitive life forms closely related to bacteria) <b>capable of oxygen-evolving photosynthesis</b> (based on carbon dating of organic matter from this period). They would have been preceded by bacteria that perform anaerobic photosynthesis.
2,000-1,800	<b>Oxygen begins to accumulate in the atmosphere</b>
1,400	<b>Microbial assemblages of relatively large unicells</b> (25-200 micrometres) found in marine siltstones and shales, indicating the <b>presence of eukaryotic (nucleate) organisms</b> . These fossils have been interpreted as cysts of planktonic algae. [ <b>Eukaryotes are thought to have originated about 2,000 million years ago</b> ]
800-700	<b>Rock deposits containing about 20 different taxa of eukaryotes</b> , including probable protozoa and filamentous green algae
640	<b>Oxygen reaches 3% of present atmospheric level</b>
650-570	The oldest fossils of <b>multicellular animals</b> , including primitive arthropods
570 onwards	The first evidence of plentiful living things in the rock record
400 onwards	<b>Development of the land flora</b>
100	<b>Mammals, flowering plants, social insects appear</b>



# The origin of modern human

## Hominid and hominin – what's the difference?

---

- **Hominid** – the group consisting of all modern and extinct Great Apes (that is, modern humans, chimpanzees, gorillas and orang-utans plus all their immediate ancestors).
- **Hominin** – the group consisting of modern humans, extinct human species and all our immediate ancestors (including members of the genera *Homo*, *Australopithecus*, *Paranthropus* and *Ardipithecus*).
- See more at:  
<http://australianmuseum.net.au/hominid-and-hominin-whats-the-difference#sthash.ScE7IWfW.dpuf>



# The origin of modern human

## Hominid and hominin

4.4 million years:	Appearance of Ardipithecus, an early Hominin Genus.
4 million years:	North and South America joined at the Isthmus of Panama. Animals and plants cross the new land bridge.
3.9 million years:	Appearance of Australopithecus, Genus of Hominids.
3.7 million years:	Australopithecus Hominids inhabited Eastern and Northern Africa.
2.7 million years:	Evolution of Paranthropus (extinct hominins).
2.4 million years:	Homo Habilis appeared.
2 million years:	Tool-making Humanoids emerged. Beginning of the Stone Age, lasted several million years.
1.7 million years:	Homo Erectus first moved out of Africa.
1.2 million years:	Evolution of Homo antecessor. The last members of Paranthropus died out.
700,000 years:	Human and Neanderthal lineages started to diverge genetically.
600,000 years:	Evolution of Homo Heidelbergensis.
530,000 years:	Development of speech in Homo Heidelbergensis.
400,000 years:	Hominids hunted with wooden spears and used stone cutting tools.
370,000 years:	Human ancestors and Neanderthals were fully separate populations.
350,000 years:	Evolution of Neanderthals.
300,000 years:	Hominids used controlled fires. Neanderthal man spread through Europe
200,000 years:	Anatomically modern humans appeared in Africa.
105,000 years:	Stone age humans foraged for grass seeds such as sorghum.
80,000 years:	Non-African humans interbred with Neanderthals.
60,000 years:	Oldest male ancestor of modern humans.
40,000 years:	Cro-Magnon man appeared in Europe.
30,000 years:	Neanderthals disappeared from fossil record.
15,000 years:	Bering land bridge between Alaska and Siberia allowed human migration to America.



# Evolutionary history of life

## Prokaryotic phylogeny

---

- **Bacteria** represent the **oldest form of life**.
- The **evolution of bacteria** over **at least 3.5 billion years spans** and occurred in step with its geochemical development.
- Prokaryotic evolution has the main role in the origin of the eukaryotic cell:
  1. Responsible for **creating oxygen atmosphere**.
  2. Plays important role in **genetic diversity**.
  3. **Transfer of genes** via **viruses, plasmids, other DNA fragments**.
  4. **Rapid generation time** is an **alternative evolutionary strategy**.



# Evolutionary history of life

## Prokaryotic phylogeny

---

- The **Bacteria** make up the vast majority of **prokaryotes**.
- Hence, **discerning**(detect or distinguish) the **evolutionary relationships among them** constitutes a **major part of understanding prokaryotic phylogeny**.





# Evolutionary history of life

## Bacteria

---

- Prokaryotic organisms were the sole inhabitants of this planet for the first 2-2.5 billion years.
- To understand such fundamental questions as:
  1. The nature and origin of the first cell,
  2. Origin of different types of metabolism,
  3. Information transfer processes,
  4. Photosynthesis, origin of the eukaryotic cells,
  5. Evolution of disease-causing as well as
  6. Beneficial microbes, a sound understanding of the bacterial (prokaryotic) evolution is essential.



# Evolutionary history of life

## Bacteria

---

- The analyses of genome sequence data using new approaches are providing valuable insights in understanding some of these most ancient and important aspects of the evolutionary history of life.

# Fossil record

## Layers of a bacterial mat

- Fossilized mats 2.5 billion years old mark a time when photosynthetic prokaryotes were producing enough O<sub>2</sub> to make the atmosphere aerobic.



Layers of a bacterial mat

# Fossil record

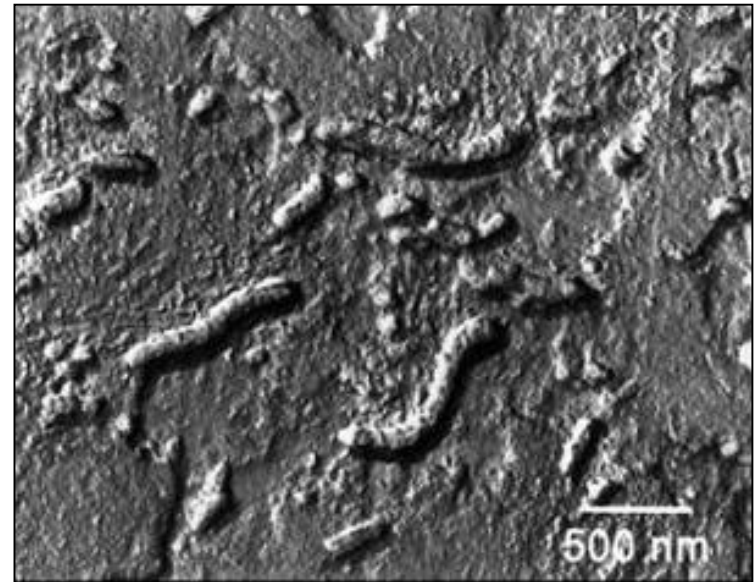
## Fossilized prokaryote and a living bacterium



# Nanobacteria

## The smallest cell-walled organisms on earth

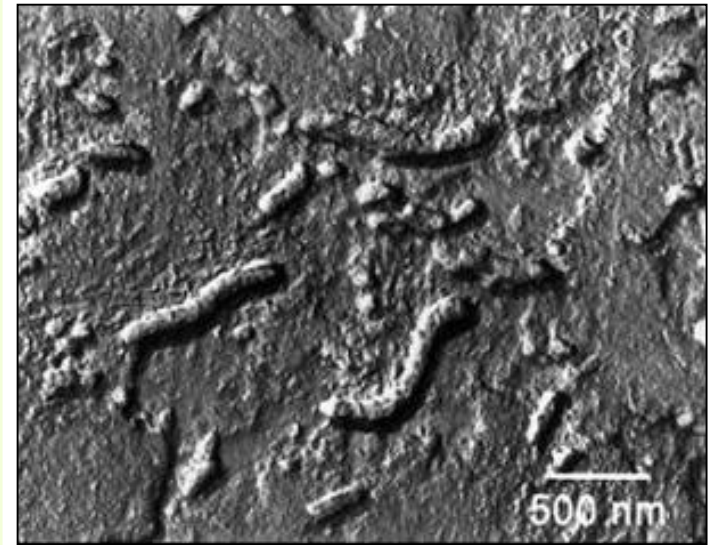
- **Nanobacteria** (singular nanobacterium) or **nanobes** (sometimes used as distinct terms, they are often used interchangeably) are **nano-sized bacteria** found in organisms (even human blood) and rocks.
- Nanobacteria might have a potential role in forming kidney stones.
- Smallest cell-walled organisms on earth, smaller than 300nm (1/10 the size of bacteria).



# Nanobacteria

## The smallest cell-walled organisms on earth

- Some questioning whether or not an organism of this size has enough room to house necessary cell components such as DNA, RNA, and plasmids.
- Nanobe studies challenge our perception of life.
- Microbes have already expanded our understanding of the harsh conditions that can support life.
- So, if nanobes do exist as living biota, they will broaden our perspective on the scale of life.





# A Brief History of Origin of Life

---

## 3. Primitive Organisms and Molecular Coding: RNA life

The "RNA World" is essentially a hypothetical stage of life between the first replicating molecule and the highly complicated DNA/protein world.

The modern cell is: DNA → RNA → Protein

# A Brief History of Origin of Life

## Coherent pathway

- A major new hypothesis outlines a **coherent (consistent) pathway** that:
  1. starts from no more than rocks, water and carbon dioxide, and
  2. leads to the emergence of the strange bioenergetic properties of living cells.

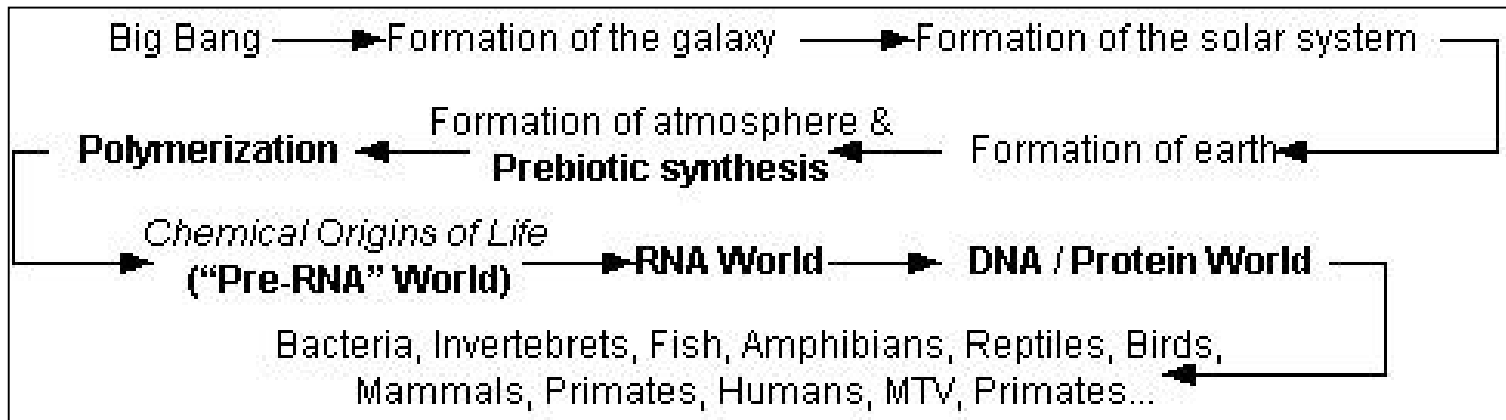


These considerations could also explain the deep divergence between bacteria and archaea (single celled microorganisms).



# A Brief History of Origin of Life

- According to **Stanley Miller**, famous origin of life researcher, the chain of events looked something like this:

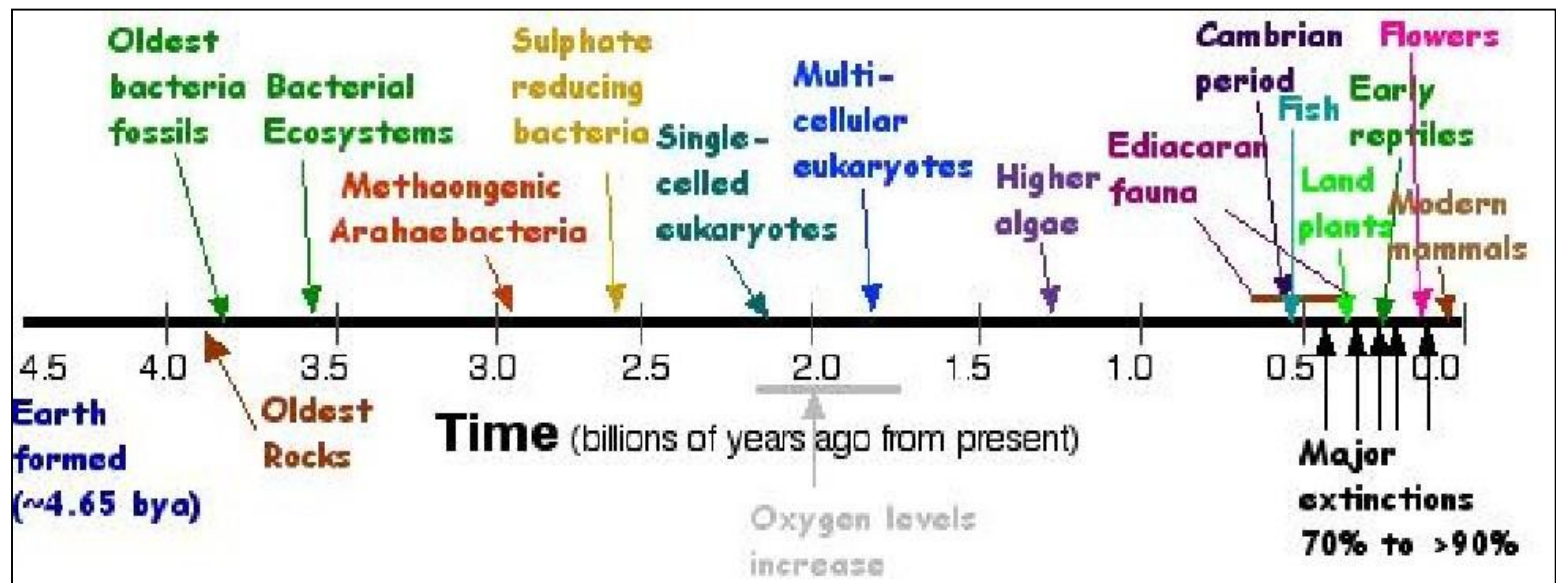


The prebiotic synthesis of organic compounds as a step toward the origin of life," **S. L. Miller**, *Major Events in the History of Life* (London: Jones and Bartlett Publishers, 1992).

# A Brief History of Origin of Life

## Earth History Timeline Major Events

- Timeline of Earth's History Recent History of Life on Earth – 600 millions years ago to the Present.





# A Brief History of Origin of Life

## Steps for cell formation

---

1. Pre-Biotic Synthesis
2. Polymerization
3. Pre-RNA World: Getting A Sufficient Self-Replicating Molecule
4. RNA World
5. DNA/Protein World
6. Making Proto-cells (first cells).

After seeing difficulties faced by the origin of life, perhaps this is why over 20 years ago, the noted scientist who discovered the structure of DNA, **Francis Crick**, said:

The origin of life appears to be almost a miracle, so many are the conditions which would have had to be satisfied to get it going."



# A Brief History of Origin of Life

---

## 1. Pre-Biotic Synthesis:

- **Collection of chemicals.** Sufficient quantities of chemicals thought to be necessary for life's natural origin were formed.

## 2. Polymerization:

- The process by which "**monomers**" (simple organic molecules such as amino acids, sugars, lipids, simple carbohydrates, nucleic acids) form covalent bonds with one another to produce "**polymers**" (complex organic molecules).





# A Brief History of Origin of Life

---

## 3. Pre-RNA World:

- A sufficient self-replicating molecule.
- Since molecules like RNA or DNA are too complex to be existed earlier, so there must have been some other more simple precursor to RNA or DNA.
- It has been hypothesized that the earliest life on Earth may have used PNA (peptide nucleic acid) as a genetic material due to its extreme robustness(resist to change), and later transitioned to a DNA/RNA-based system.

Prebiotic RNA had two properties not evident today: a capacity to replicate without the help of proteins and an ability to catalyze every step of protein synthesis.



# A Brief History of Origin of Life

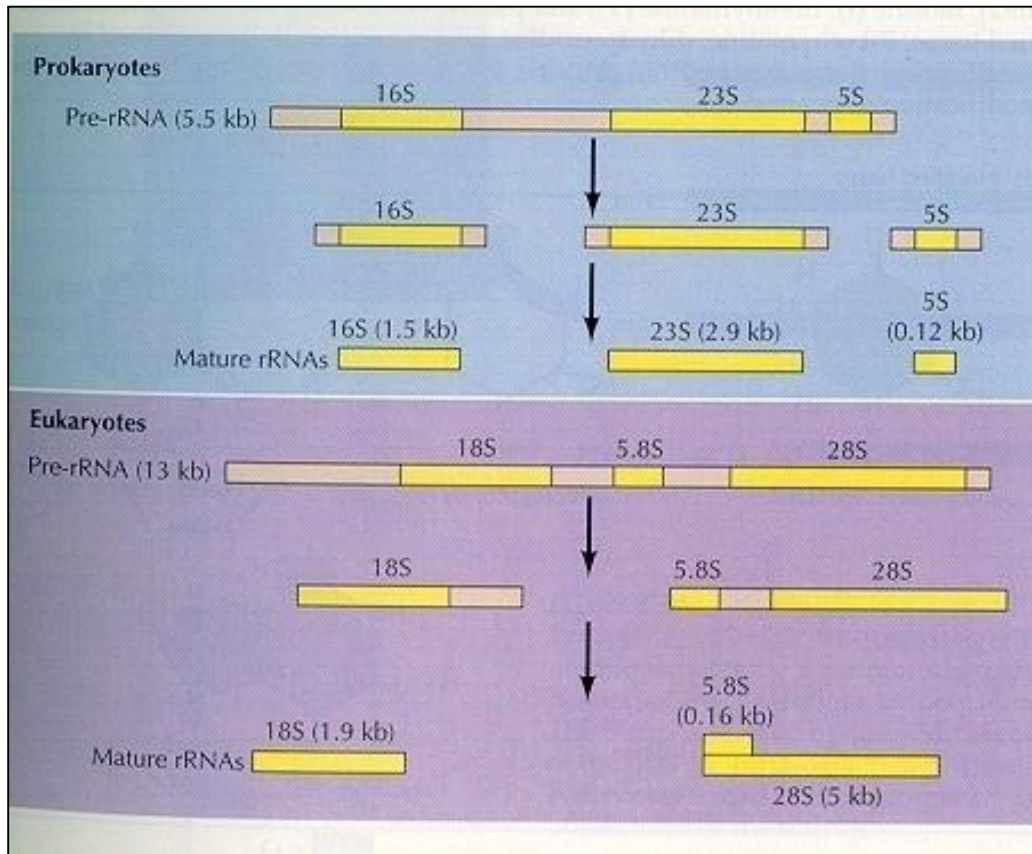
---

## 4. RNA World:

- Some time after the first “self-replicating” molecule (pre-RNA) formed, according to the story, RNA came along.
- Today, RNA is a genetic molecule in all cells, similar to DNA, but more versatile within the cell.
- The “RNA World” is essentially a hypothetical stage of life between:
  1. The first replicating molecule, and
  2. The highly complicated DNA-protein-based life.

# Pre-rRNA

**Prokaryotic cells contain three rRNAs (16S, 23S and 5S), which are formed from cleavage of a pre-rRNA transcript**



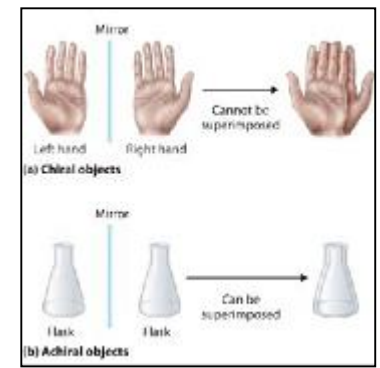
*Figure 6.37*

## **Processing of ribosomal RNAs**

Prokaryotic cells contain three rRNAs (16S, 23S, and 5S), which are formed by cleavage of a pre-rRNA transcript. Eukaryotic cells (e.g., human cells) contain four rRNAs. One of these (5S rRNA) is transcribed from a separate gene; the other three (18S, 28S, and 5.8S) are derived from a common pre-rRNA. Following cleavage, the 5.8S rRNA (which is unique to eukaryotes) becomes hydrogen-bonded to 28S rRNA.

# Pre-RNA World

## Where did RNA come from?



- It has been assumed that there was a **much simpler informational macromolecule than RNA**.
- It has been dubbed preRNA (or pre-RNA).
- This molecule may have been achiral and may have used **bases other than AUGC**.
- **An example** of an alternative backbone is **PNA (peptide nucleic acid)**.

**Achiral**: a type of molecule that has a nonsuperposable mirror image. **Chiral** means mirror image not the same.





# Pre-RNA World

## PNA/TNA/GNA

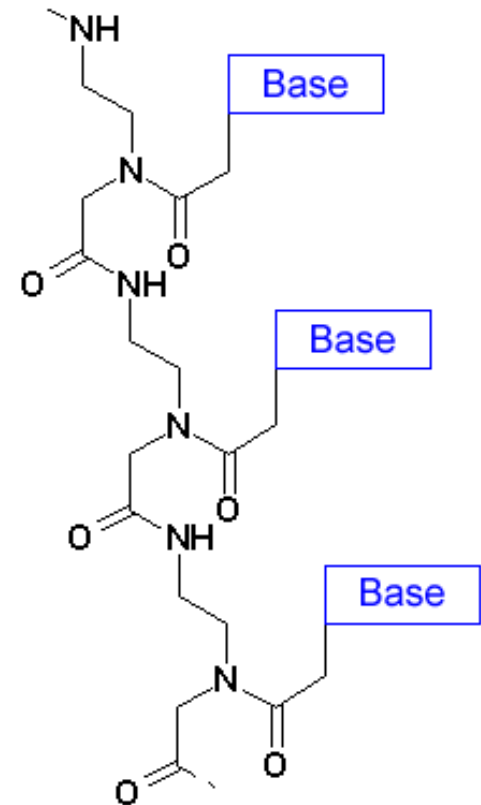
---

- PNA is more stable than RNA and appears to be more readily synthesized in prebiotic conditions, especially where the synthesis of ribose and adding phosphate groups are problematic.
- Two more starting molecules(ancestors of DNA) are:
  1. Threose nucleic acid (TNA World)has also been proposed as a starting point, as has
  2. Glycol nucleic acid (GNA World).

# PNA structures

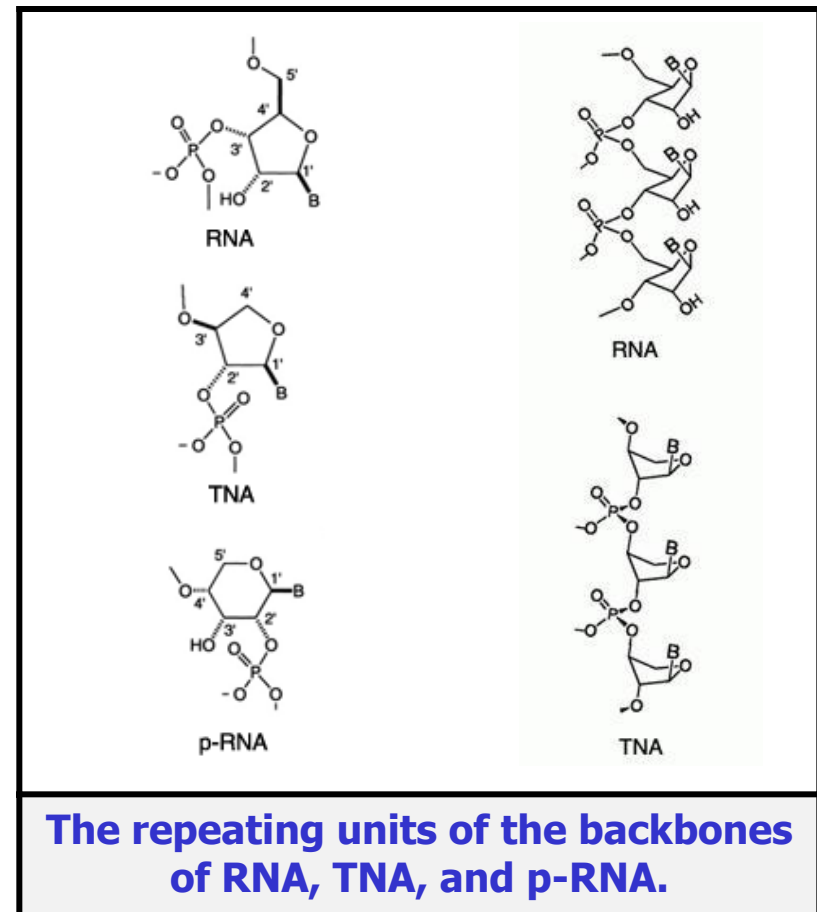
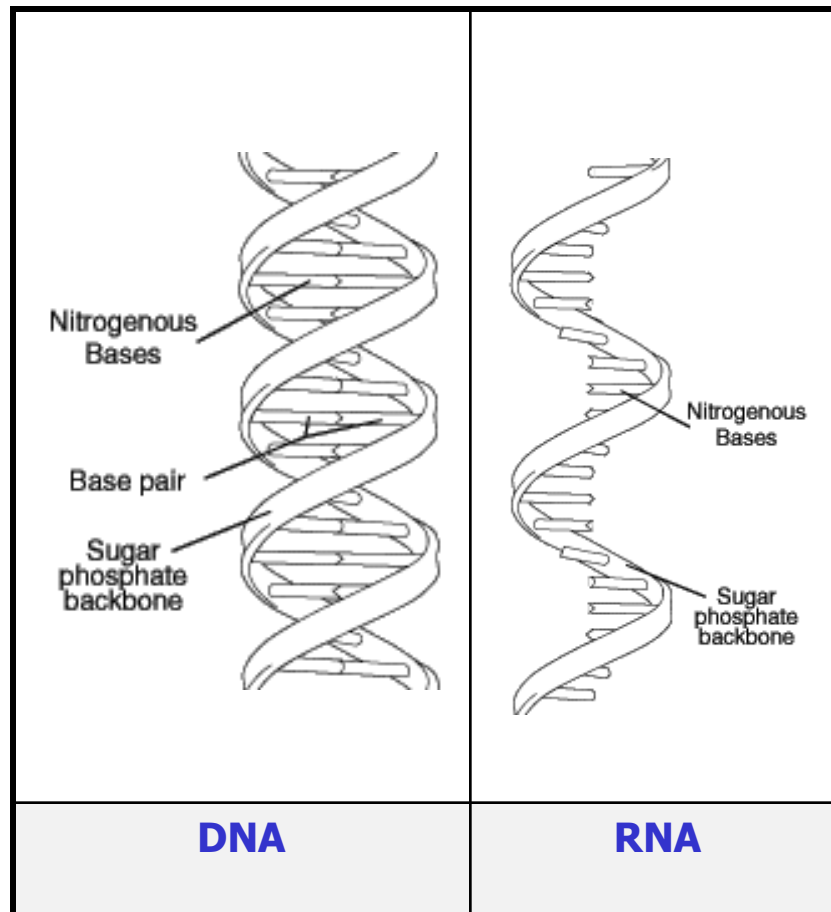
## Peptide nucleic acid

- **PNA is peptide nucleic acid**, a chemical similar to DNA or RNA but differing in the composition of its "backbone".
- DNA and RNA have a ribose sugar backbone, whereas PNA's backbone is composed of repeating **N-(2-aminoethyl)-glycine units** linked by peptide bonds.
- Backbone of PNA contains no charged phosphate groups.
- The various **purine and pyrimidine bases** are linked to the backbone by methylene carbonyl bonds.
- PNAs are depicted like peptides, with the
  1. **N-terminus** at the first (left) position, and
  2. The **C-terminus** at the right.



# Possible ancestors of DNA:

## PNA, p-RNA, and TNA



# RNA world

## Era of nucleic acid life

### The RNA world hypothesis

---

- The RNA world hypothesis proposes that RNA was the first life-form on earth, later developing a cell membrane around it and becoming the first prokaryotic cell.
- All life on earth appears to share the same origins.
- There is considerable evidence that there was a period of time on Earth called the RNA world.
- In this world life existed as RNA as both phenotype and genotype.

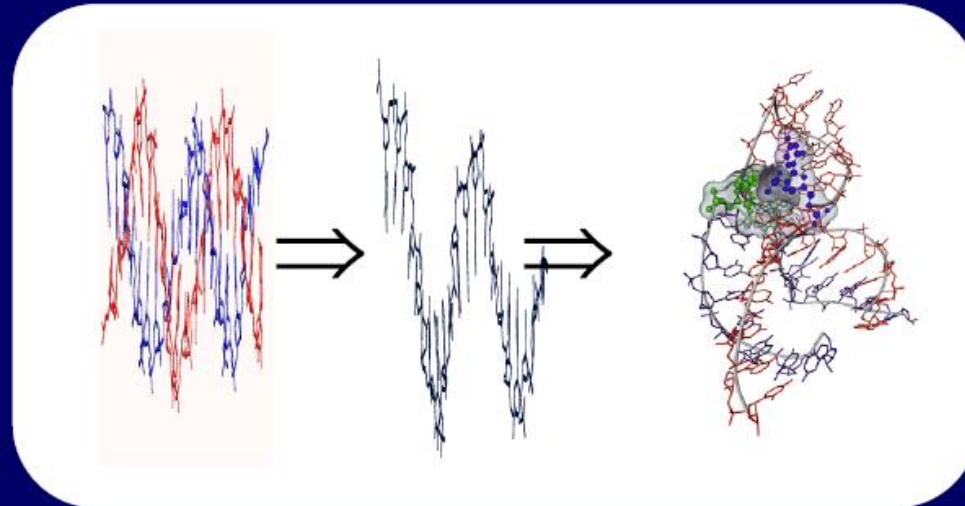
Carl Woese was also the originator of the RNA world hypothesis in 1977, although not by that name.

# RNA world

## Pre-RNA

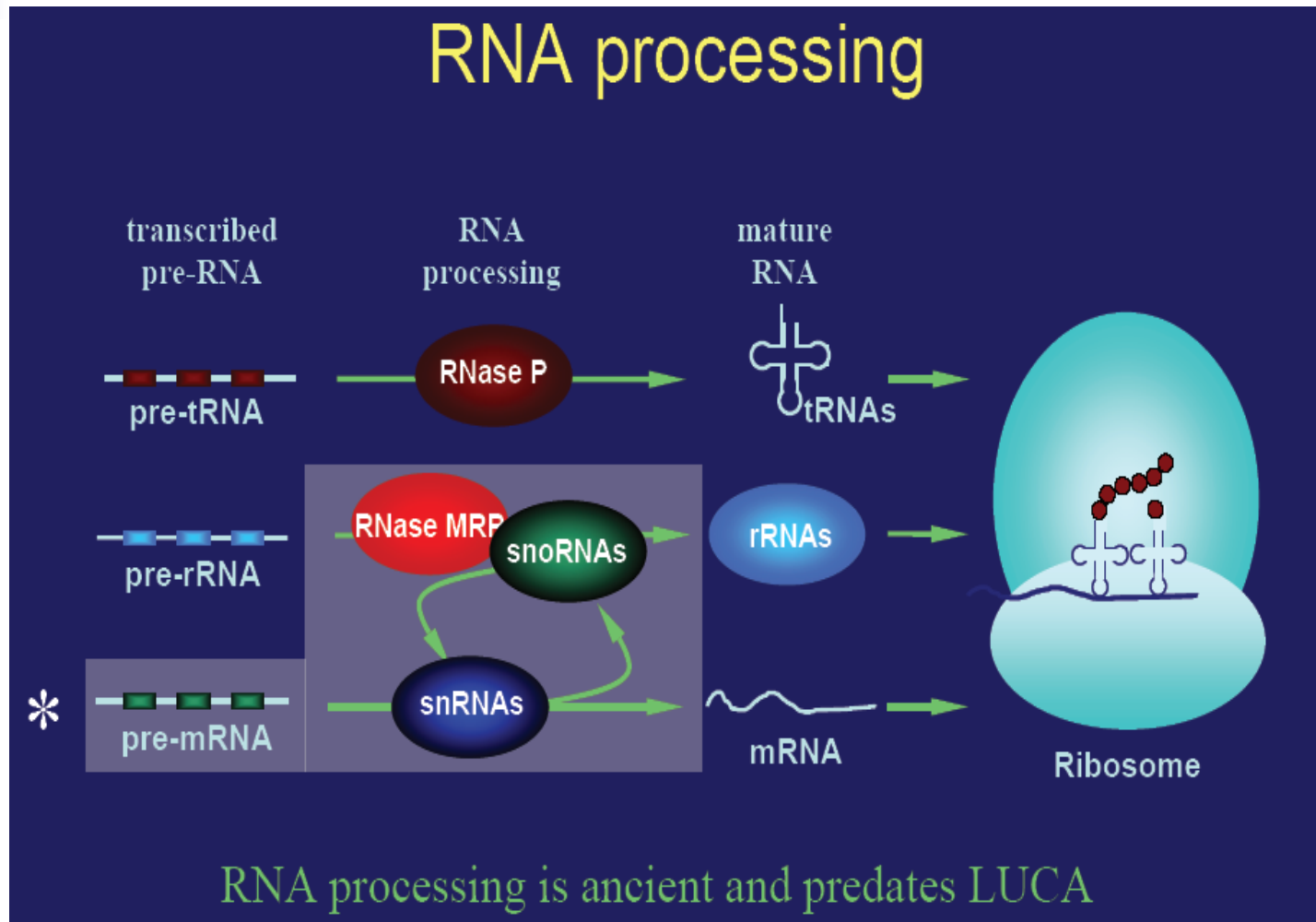
The RNA world:

RNA      RNA  
RNA  $\Rightarrow$  Pre-RNA  $\Rightarrow$  RNA



# RNA processing

## Pre-RNA





# RNA world

## Pre-RNA

- The prebiotic RNA had two properties not evident today:
  1. A capacity to replicate without the help of proteins, and
  2. An ability to catalyze every step of protein synthesis.

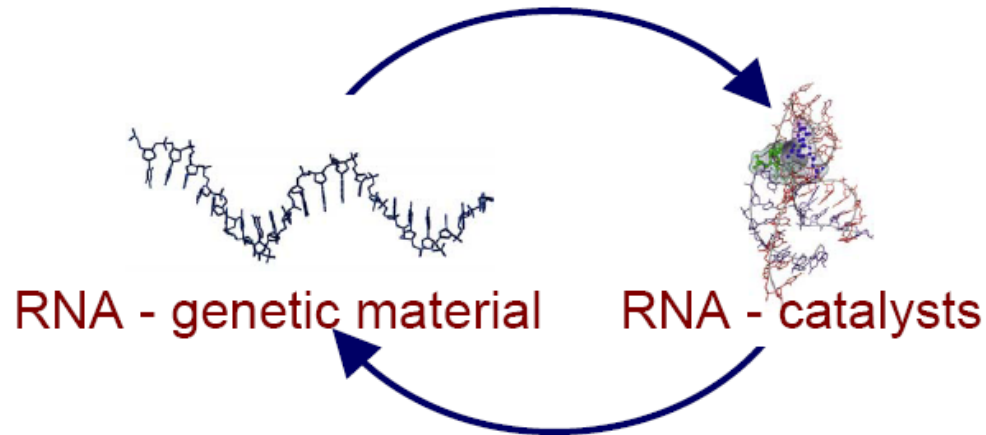
### The RNA world hypothesis:

That there was a period in the evolution of life where RNA was both biological catalyst and genetic material.

The RNA world hypothesis holds that in the primordial soup (or sandwich), there existed free-floating nucleotides. These nucleotides regularly formed bonds with one another, which often broke because the change in energy was so low.

# RNA world

**Q.1. Which came first? RNA-genetic material or RNA catalysts?**



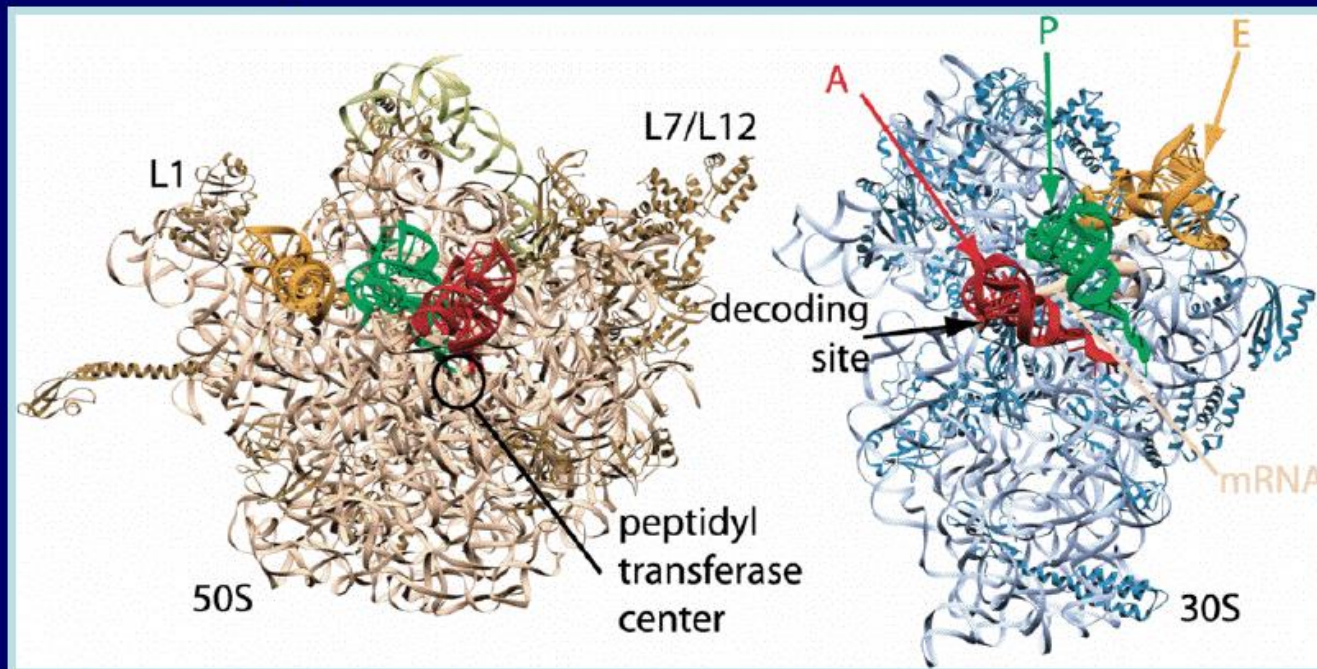


# Protein synthesis

## Catalytic RNA

### The ribosome

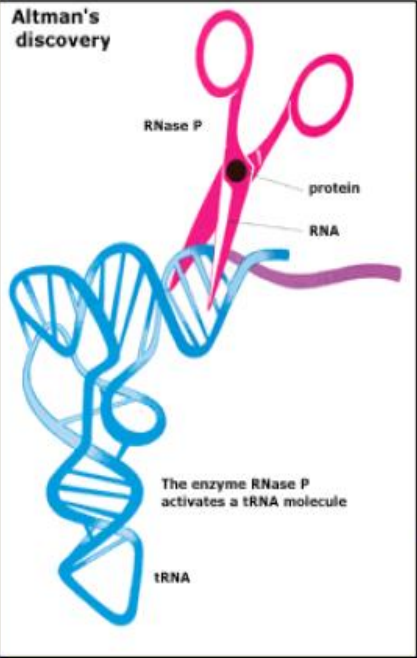
Proteins are synthesised by catalytic RNA: the ribosome



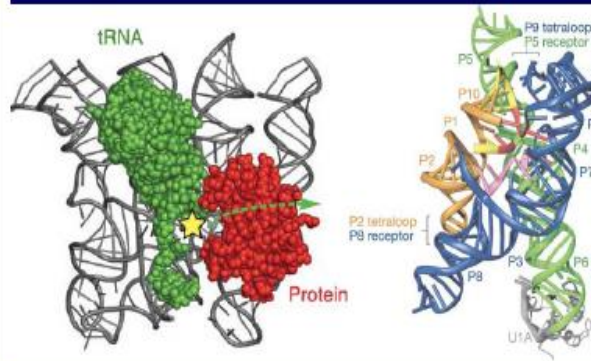
# Catalytic RNA

## Catalytic RNA

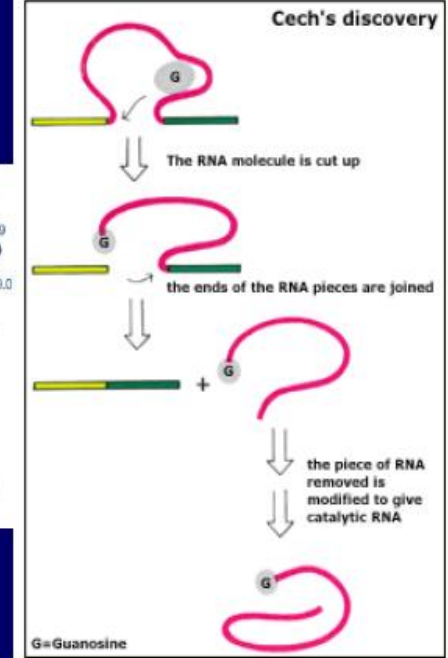
Altman's discovery



RNase P



Cech's discovery



Group I introns



# Modern RNA genomes

---

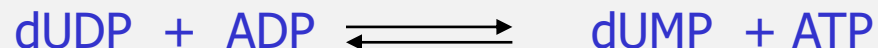
- The RNA which is the genetic material of some viruses i.e. TMV.
- Plant viroid RNAs ( $\approx$  400 nt, catalytic RNA, Code no protein).



# The reasons for RNA world

---

1. RNA has a template structure.
2. RNA has catalytic properties.
3. RNA appears in various presumably ancient cellular processes (i.e. ribosome, primer for DNA, etc.).
4. Ribonucleotides are components of many coenzymes (e.g. CoA, NADH, etc.).
5. The biosynthesis of histidine is uses ATP and PRPP.
6. The biosynthesis of deoxynucleotides is from ribonucleotide diphosphates.
7. The biosynthesis of dTMP is from dUMP (Thymidylate synthase (TS) is the enzyme that catalyzes the transformation of deoxyuridine monophosphate (dUMP) into deoxythymidine monophosphate (dTMP) in cells).





# RNA world

## The chief problem facing an RNA world

---

- The chief problem facing an RNA world is that RNA cannot perform all of the functions of DNA adequately to allow for replication and transcription of proteins.
- OOL theorist Leslie Orgel notes that an "RNA World" could only form the basis for life, if prebiotic RNA had two properties not evident today:
  1. A capacity to replicate without the help of proteins and
  2. An ability to catalyze every step of protein synthesis.
- The RNA world is thus a hypothetical system behind which there is little positive evidence, and much materialist philosophy.



# A Brief History of Origin of Life

## 5. DNA/Protein World

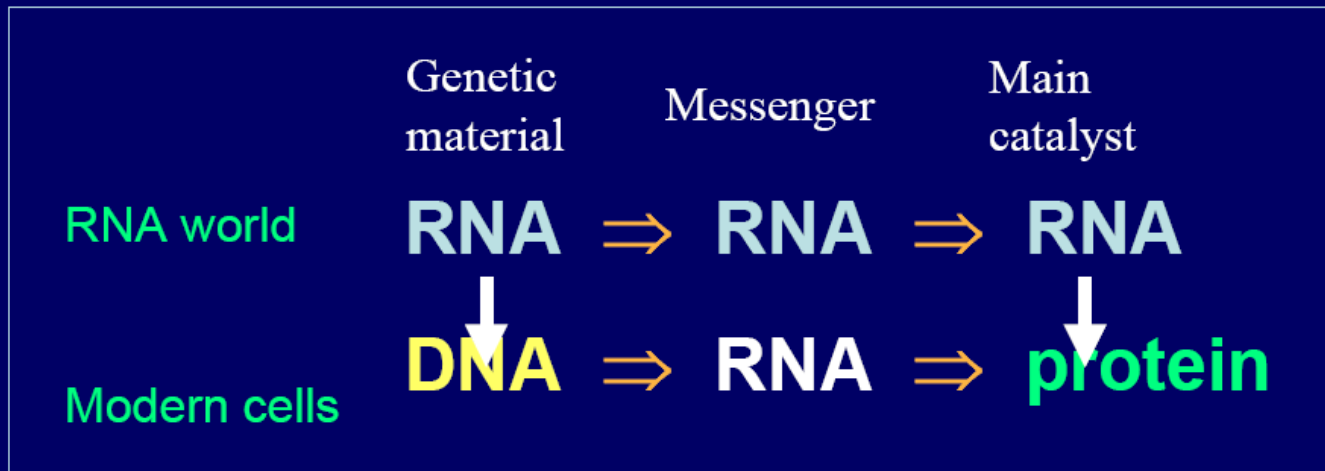
---

- Since the RNA in the RNA world is alive, it is assumed that RNA evolved into DNA through some sort of genetic takeover event.
- In other words, RNA enzymes made DNA, which replaced it in the genome.
- Proteins were added into the mix at some point.

# RNA

## DNA formation

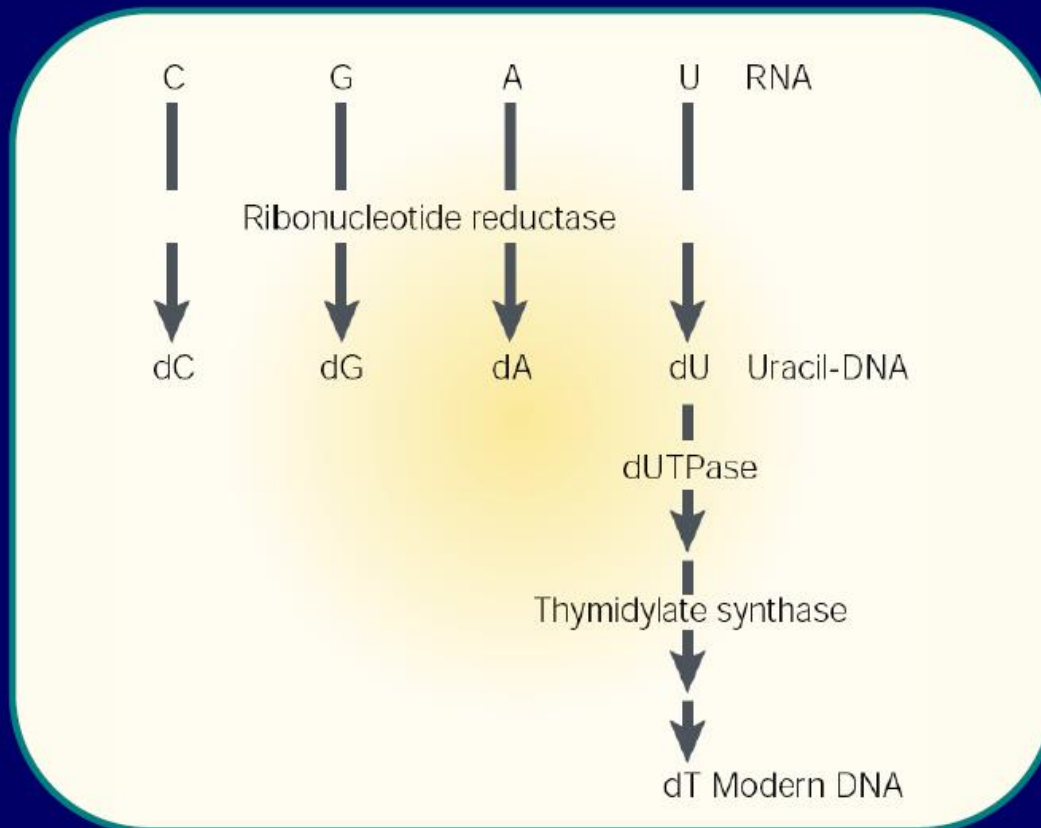
RNA: jack of all trades,  
master of none.



- These transitions are expected to have occurred because DNA is superior to RNA as an information storage molecule, and proteins are superior to RNA as a biological catalyst.
- RNA still carrying out these roles may in some cases be 'relics' from an earlier period in the evolution of life.

# Origin of DNA

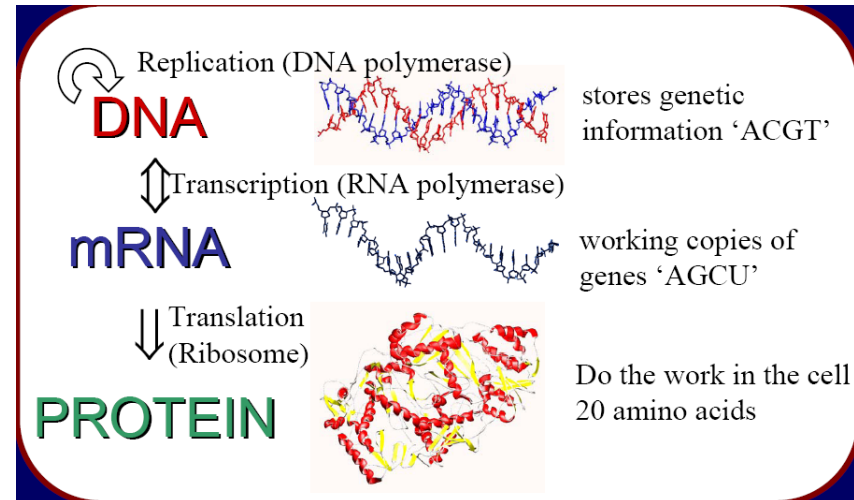
## Origins of DNA in two steps





# DNA/Protein World

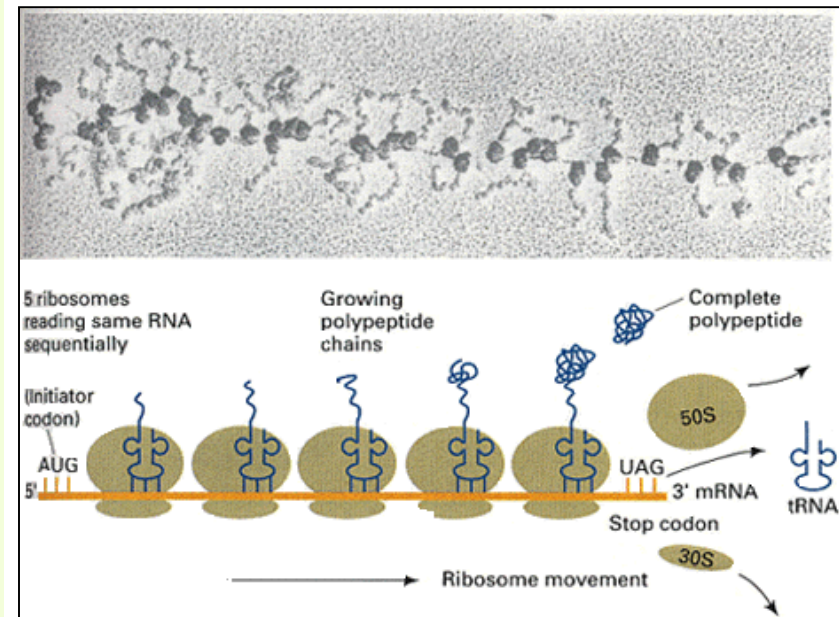
- The transcription - translation process is the means by which the information in the DNA code creates protein (protein synthesis).



# Protein synthesis

## Ribosomal function

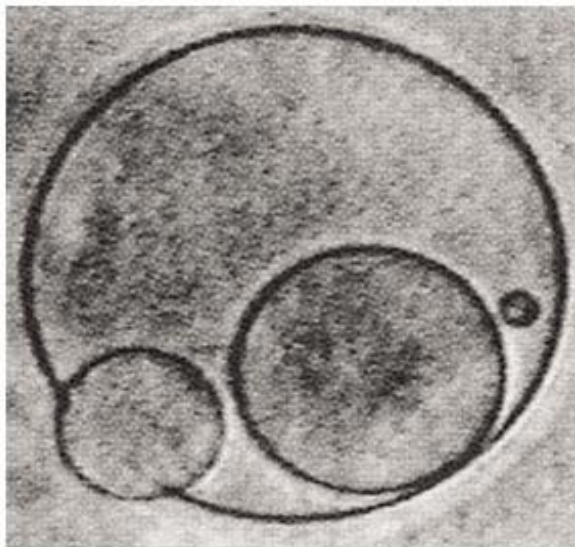
- During protein synthesis a ribosome moves along an mRNA molecule, reading the codon and adding the correct amino acid (from the corresponding aminoacyl tRNA) to the growing protein.
- When a stop codon is reached, translation ceases, and the mRNA and protein are released.



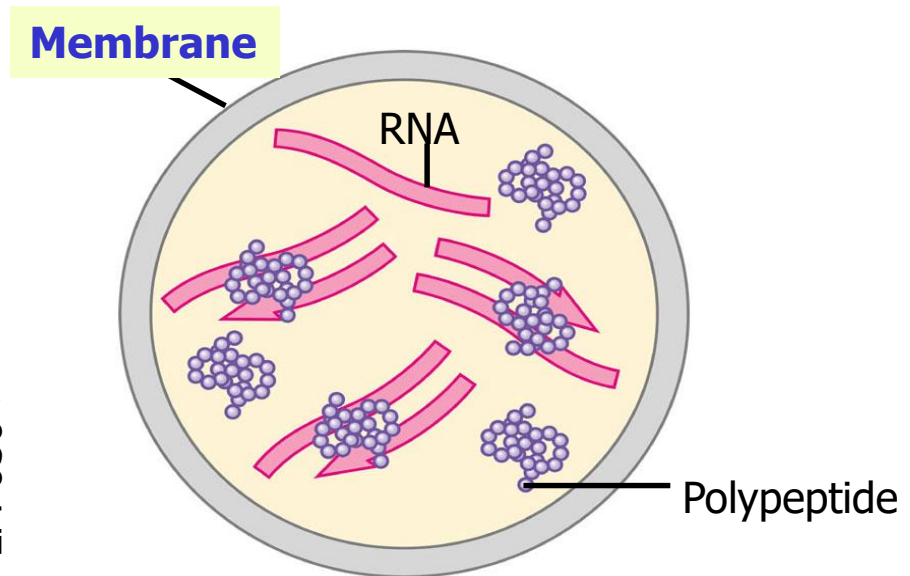
# Membranes

## Functions

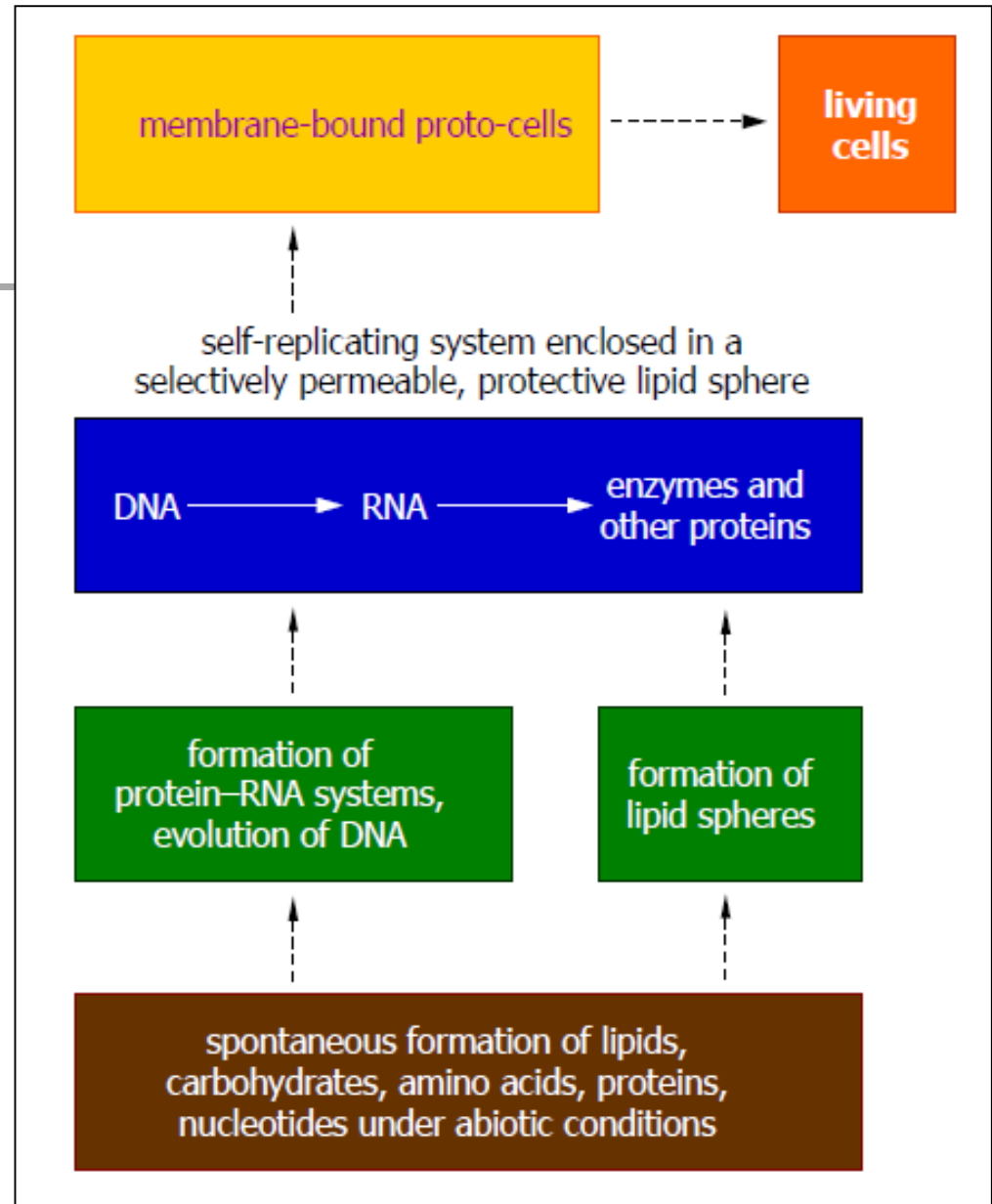
- Membranes may have separated **various aggregates of self-replicating molecules** which could be acted on by natural selection.



LM 650x



## The functions of membranes.





# A Brief History of Origin of Life

## 6. Making Proto-cells

---

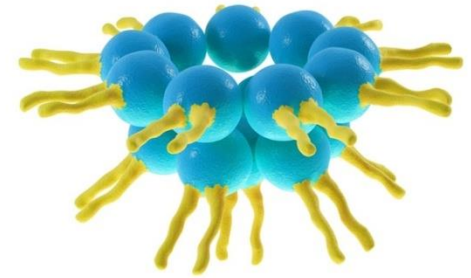
- **Protocells: Both the past and the future of biology**
- Protocells were those **primordial** (original primitive), **chemical objects** that proved **capable of the evolutionary adaptations needed to produce biological cells.**

The **biological cell** is an extremely advanced microscopic entity.

**All biological cells contain macromolecules.**

There are **three major groups of macromolecules**,  
**polysaccharides, proteins and nucleic acids.**

# Protocells



- Early protocells are assumed to be spherical, their shape being determined by the same physical forces that form oil droplets, mainly surface tension.
- As with other similar structures, such as bubbles, the spherical shape arises from minimization of surface energy and surface area.
- This is a spherical membraneless microdroplet which can spontaneously arise from weak organic solutions.

# Protocells

## Q.2. Which came first? RNA or DNA?

- Which came first?
- DNA needs enzymes (DNA polymerase and associated enzymes) to replicate, but the enzymes are encoded by DNA.
- DNA needs protection of the cell wall, but the cell wall is also encoded by the DNA.
- The answer is that neither came first for all are required in DNA-based life.



These fundamental components form an irreducibly complex system in which all components must have been present from the start.



# Protocells

## Q.2. Which came first? RNA or DNA?

---

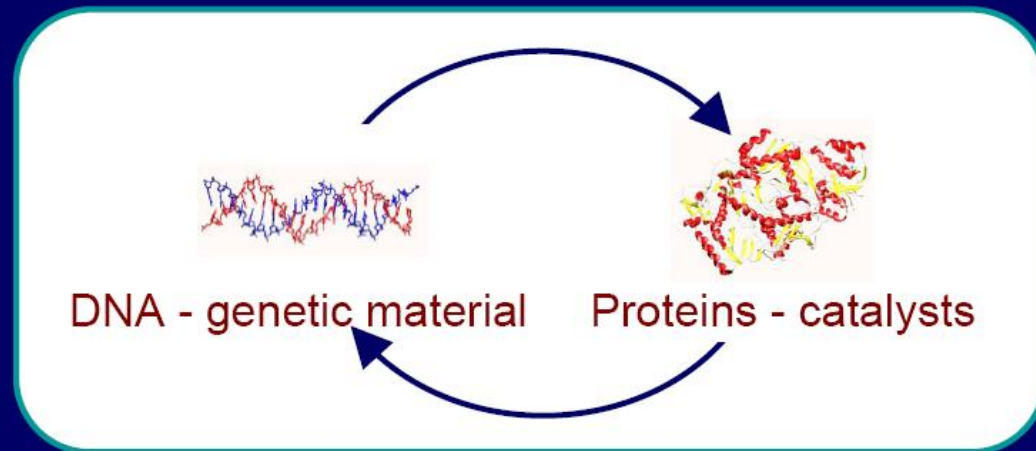
- Protocells were a simple structure that spontaneously arose and acted as a vehicle for the evolution of life on Earth.
- Protocells are thought to have facilitated the reproduction of RNA and therefore the exchange of genetic information at a time before the advent of DNA and proteins (**the RNA world hypothesis**).
- This author agrees, that RNA appeared before DNA.



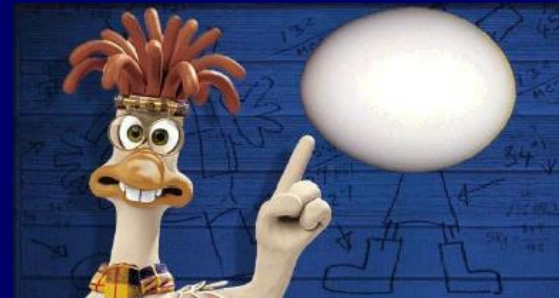
# Protocells and Biological Cells

## Q 3. Which came first-DNA or Protein?

Q: Which came first -  
DNA or Protein?

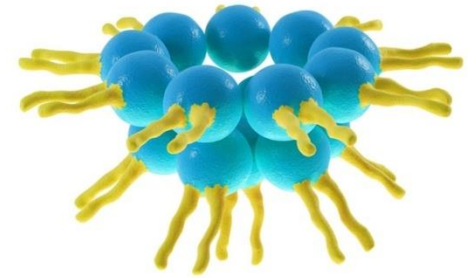


**DNA** ⇒ **RNA** ⇒ **protein**



# Protocells

## Definition



- Protocells are spherical **membraneless** microdroplet structures which are formed from the **aggregation of abiotic (non-living) components**.
- Protocells can spontaneously **arise from weak organic solutions**.
- Despite this, they display certain characteristics akin (of similar) to living cells.
- Protocells are basically:
  1. **self-organized,**
  2. **endogenously ordered,**
  3. **spherical collection of lipids.**



# Protocells

## Differences between protocells and cells

---

- There are many differences between protocells and biological cells.
- Biological cells generally have three features:
  1. **A stable and semi-permeable membrane** which encapsulates cell components
  2. **Genetic material** which can be passed on in cell formation and which controls cellular behavior and function
  3. **Energy generation** *via* metabolic pathways which enables growth, self-maintenance, and reproduction.



# Protocells

## Differences between protocells and cells

---

- Protocells display **certain characteristics** in common with cells. E.g.
- In the case of **membrane transport**, **modern cells use complex protein machineries**.
- Whereas, protocells may have **achieved membrane transport** (which is crucial for the exchange of content) passively **via processes such as osmosis**.
- In this way protocells could have exchanged ions and small molecules with their surrounding environment.



# Protocells

## Differences between protocells and bacteria

---

- There are many differences between protocells and bacteria, the simplest extant forms of independent cellular life.
  1. Differences in morphology
  2. Macromolecular chemistry
  3. Phospholipids
  4. RNA
  5. Bases other than adenine
  6. Genetics and data processing.



# Phylogenetic Taxonomy

---

- **To get accurate phylogeny, must decide which characteristics give best insight.**
- **DNA and RNA sequencing techniques are considered to give the most meaningful phylogenies.**



# Brief history of molecular phylogenetics

---

- 1900s
- Immunochemical studies: Cross-reactions stronger for closely related organisms.
- Nuttall (1902) - apes are closest relatives to humans.
- 1960s -1970s
- Protein sequencing methods, electrophoresis, DNA hybridization and PCR contributed to a boom in molecular phylogeny.
- Late 1970s to present
- Discoveries using molecular phylogeny:
  - Endosymbiosis - Margulis, 1978
  - Divergence of phyla and kingdom - Woese, 1987.
  - Many Tree of Life projects completed or underway.



# Classification, Taxonomy and Phylogeny

---

- **Key definitions to match up and learn!**
- Taxonomy: The study of principles of classification.
- Classification: The process of sorting living things into groups.
- Phylogeny: The study of evolutionary relationships between organisms.





# Classification, Taxonomy and Phylogeny

---

- **Species** (from the Latin: kind): A group whose members **posses similar anatomical characteristics** and have the ability to interbreed.
- **Speciation**: **The evolution of a new species.**
- **Taxonomy**: The branch of science concerned with **naming and classifying the diverse forms of life.**
- **Phylogeny**: **the sequence of events involved in the evolutionary development of a species or taxonomic group of organisms.**



# Speciation

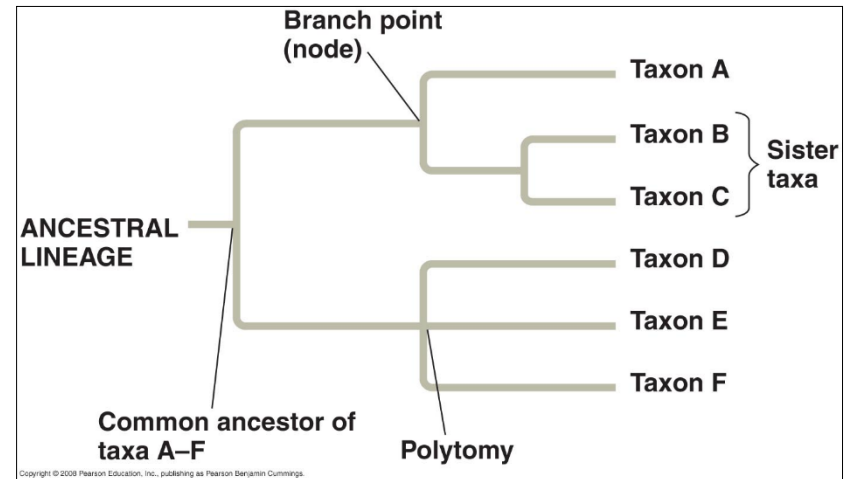
---

- A natural process usually resulting in an increase in the number of species in a particular group.
- Speciation is not a single process but an array of processes and it may be reticulate or non-reticulate.

**Reticulate speciation:** The evolution of a new species through a hybridization event involving two different species. A species evolving from reticulate speciation has two ancestral species.

# Taxa

- A taxon is any group of species designated by name. Example taxa include: kingdoms, classes, etc.
- Every node should give rise to two lineages.
- If more than two lineages are shown, it indicates an unresolved pattern of divergence or polytomy.



**Sister taxa** are groups or organisms that share an immediate common ancestor. A **polytomy** shows the simultaneous speciation of three or more species.



# Taxonomy **vs** Phylogeny

---

- Taxonomy is traditionally phenotypic.
- Phylogeny is mainly genetic.
- Some call the phylogenetic classification as genotypic classification, since it is based on actual differences among cells.



# Phylogeny **vs** systematics

---

- Phylogeny refers to the history of a species, to its relationships to other species (in Greek *phyl* - refers to tribe; *gen* - refers to origin or descent).
- Systematics refers to the methods used to discover that history (in Greek *systematos* refers to a complex whole put together).



# Traditional systematics **vs.** phylogenetic systematics

---

- Taxonomists tend to fall into two schools:
  1. Traditional systematics
  2. Phylogenetic or cladistic systematics
- Modern phylogenetic methods are making many changes in traditional views of the Tree of Life.
- Since the 1970s, phylogenetic systematics has been gradually replacing traditional systematics.
- The student must understand both systems.



# Phylogenetic or cladistic systematics

## The Goal

---

- The goal of phylogenetic or cladistic systematics is to define monophyletic taxa (clades).
- A typical goal of systematics (and paleontology) is the construction of phylogenies.
- Cladistics is especially significant in paleontology, as it points out gaps in the fossil evidence.
- A phylogeny thus can be a description of the macroevolutionary history of a species or of more than one species.

Clade (clades) defined as a single complete branch of the Tree of Life.  
Group of closely related organisms with most features in common.



# Macroevolution vs. Microevolution

---

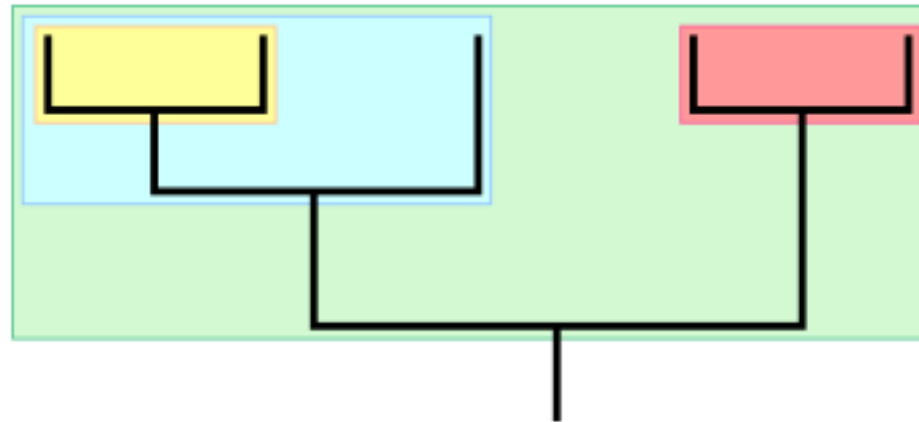
- Microevolution is evolution that occurs below the level of species.
- Macroevolution is evolution that occurs above the level of the species.
  1. Macroevolution is the origin of taxonomic groups higher than the species level.
  2. Macroevolutionary change is substantial enough that we view its products as new genera, new families, or even new phyla.



# Phylogenetic or cladistic systematics

## Definition of a clade

- A **clade** is any taxon that consists of **all the evolutionary descendants of a common ancestor**.
- Each different colored rectangle is a true clade.



Clade (clades) defined as a single complete branch of the Tree of Life.  
Group of closely related organisms with most features in common.



# Cladistic classification

---

- Millions of years ago, a single cell started an evolution that gave rise to the tree of life and its three main domains: Archaea, Bacteria and Eukaryota.
- Each branch is an example of a clade. A clade represents a group that includes a common ancestor and all descendants.
- Cladistic is a modern form of taxonomy that places organisms on a branched diagram called a cladogram (like a family tree) based on traits such as DNA similarities and phylogeny.



# Cladistic classification

## What is a cladogram?

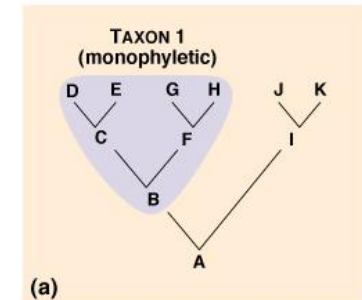
---

- A cladogram is a branching diagram which shows the evolutionary relationship among a group of clades.
- A **clade** is a group of organisms, comprised of all the evolutionary descendants of a common ancestor.
- A cladogram **does not depict the amount of evolutionary change in the group, nor does it indicate the evolutionary time or the genetic distance.**
- Each branch of the cladogram ends with a clade.
- It starts from a last common ancestor.
- **Cladograms are usually formed based on the morphological characters.**

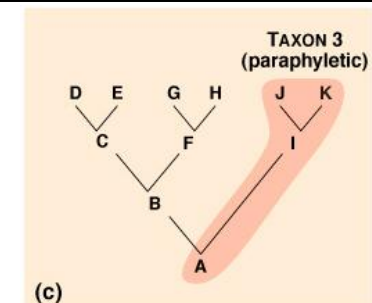
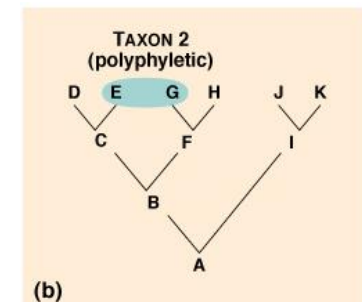
# Cladistic classification

## Monophyletic, paraphyletic, and polyphyletic trees

- **Traditionally:**
- A **monophyletic taxon** is understood to be one that includes a group of organisms descended from a single ancestor [as in (a)].
- A **polyphyletic taxon** is composed of unrelated organisms descended from more than one ancestor [as in (b)].
- One type of monophyletic taxon is a **paraphyletic taxon**, which includes an ancestor and a group of organisms descended from it [as in (c)].



©1999 Addison Wesley Longman, Inc.





# Cladograms **vs** Phylogenetic Trees

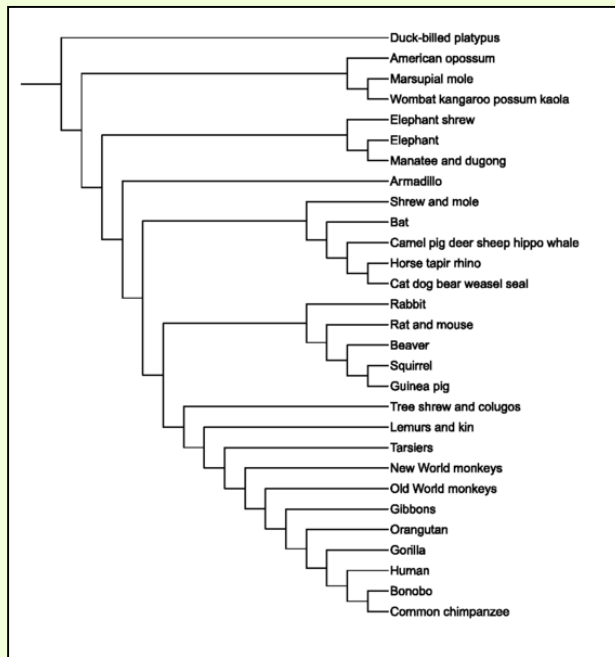
## Evolutionary time or genetic distance

---

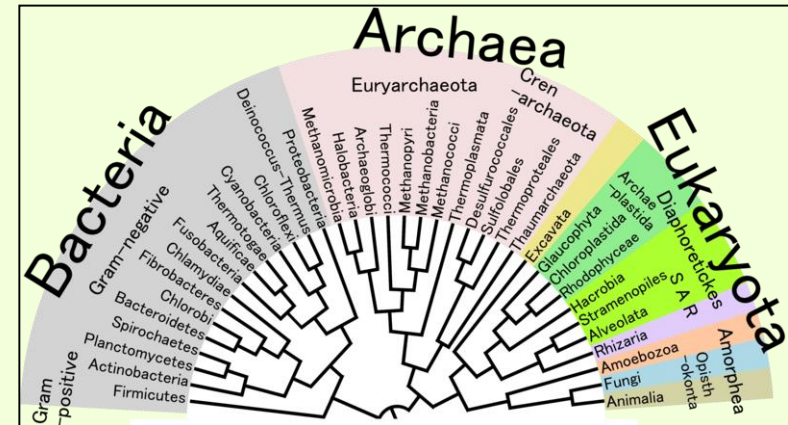
- **Cladogram:** Cladogram does not represent the evolutionary time or the genetic distance.
- **Phylogenetic Tree:** Phylogenetic tree represents the evolutionary time and the genetic distance between the group of organisms.

# Cladograms vs Phylogenetic Trees

- Cladograms are usually formed based on the morphological characters.

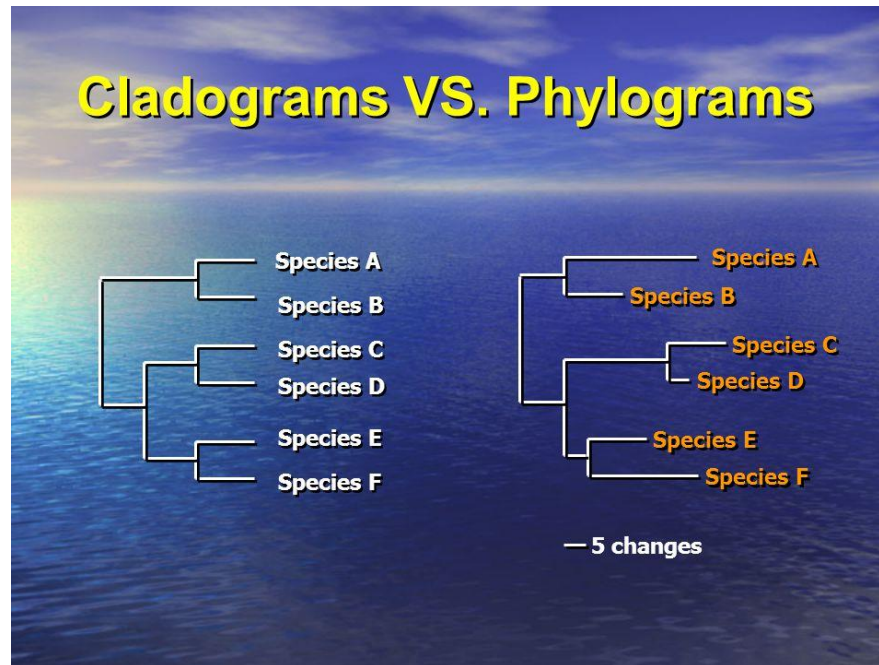


- Several characteristics like external morphology, internal anatomy, biochemical pathways, behavior, DNA and protein sequences, as well as the evidence of fossils have to be used.



# Cladograms **vs** Phylogenetic Trees

- **Cladogram** – is not an evolutionary tree. Therefore, it doesn't show evolutionary relationships.
- **Phylogram** – Phylogenetic tree is an evolutionary tree. It shows evolutionary relationships.





# **Critical issues in:** **Bacterial/Prokaryotic phylogeny**

---

## **Molecular Phylogeny**





# Problems with bacterial phylogeny

---

- To understand bacterial phylogeny, it is essential that the following two critical issues be resolved:
  1. Development of well-defined (molecular) criteria for identifying the main groups within Bacteria.
  2. To understand how the different main groups are related to each other and how they branched off from a common ancestor.
- These issues are not resolved at present.



# Problems with bacterial phylogeny

## Critical issues in Bacterial/Prokaryotic phylogeny

---

- How Archaea and Bacteria are related to each other?
- To delineate the branching order and hierarchical relationships among the major groups/taxa within **Bacteria**.
- Criteria for the higher taxonomic ranks within **Bacteria**.
- Evolutionary relationships among **photosynthetic bacteria**.
- Assessment of the extent of **lateral gene transfer (LGT)** and its impact on **Bacterial phylogeny**.
- Implications of the prokaryotic evolution on the origin of the eukaryotic cell.



# Problems with bacterial phylogeny

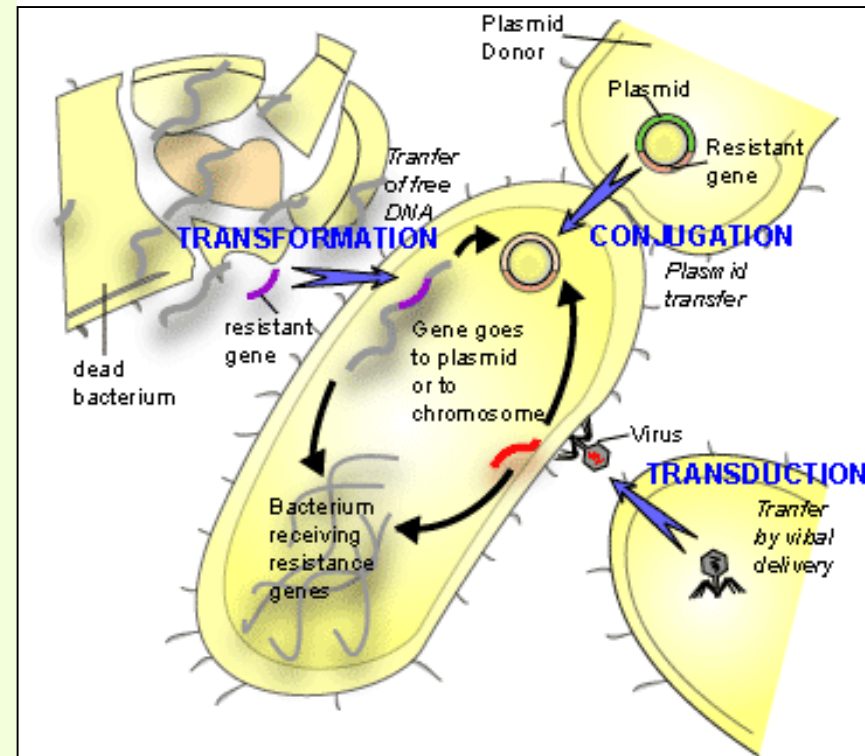
## Critical issues in Bacterial/Prokaryotic phylogeny

---

- Lateral or Horizontal Gene Transfer (LGT/HGT) influence on:
  - Evolutionary relationships
  - The relationship of Archaea to Bacteria
  - The origin of eukaryotes.
- If **organism type A** and **organism type B** carry the **same gene for a protein**, it may not be because they both belong to the same taxonomic group, but because **one of them acquired that gene (by infection or passive uptake)** from a **third type of organism**, **which is not ancestral to them**.

# Lateral or Horizontal Gene Transfer (HGT)

- Lateral or horizontal gene transfer (LGT or HGT) is a process whereby genetic material contained in small packets of DNA can be transferred between individual bacteria.
- There are three possible mechanisms of HGT.
- These are:
  1. Transduction,
  2. Transformation, or
  3. Conjugation.



Horizontal gene transfer: Incorporation of a **foreign gene** acquired from an **unrelated species** into the genome of another organism.



# Problems with bacterial phylogeny

## Critical issues in Bacterial/Prokaryotic phylogeny

---

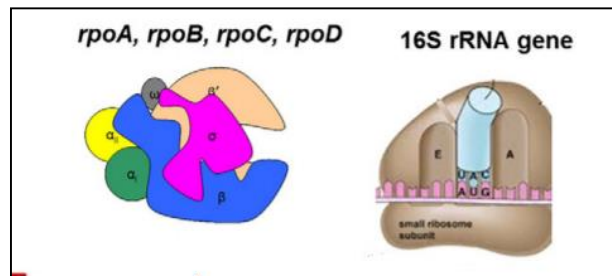
- Microscopic and molecular studies show that **<1% of the microbes in most environments have been grown in pure culture.**
- True in terms of #s and phylogenetic diversity.
- This means we know little about their biology.

# Problems with bacterial phylogeny

## Critical issues in Bacterial/Prokaryotic phylogeny

### 16S rDNA

- Phenotype not very useful for bacterial phylogeny.
- Most molecular studies based on 16s rRNA sequence analysis (rRNA Tree).
- Studies of other genes do not always agree with rRNA, especially for deep branches.



# Alternative genes

## Comparison of 16S rRNA, recA, gyrB, rpoB genes

### 16S rRNA

- Among these molecular markers, 16S rRNA, an ~1500 base pair gene coding for a catalytic RNA that is part of the 30S ribosomal subunit, has desirable properties that allowed it to become the most commonly used such marker.
- Foremost, the functional constancy of this gene assures it is a valid molecular chronometer, which is essential for a precise assessment of phylogenetic relatedness of organisms.
- It is present in all prokaryotic cells and has conserved and variable sequence regions evolving at very different rates, critical for the concurrent universal amplification and measurement of both close and distant phylogenetic relationships.

# Problems with bacterial phylogeny

## Critical issues in Bacterial/Prokaryotic phylogeny

### Limitation of 16S rDNA amplification

---

- Until today, analysis of 16S ribosomal RNA (rRNA) sequences has been the de-facto gold standard for the assessment of phylogenetic relationships among prokaryotes.
- Unfortunately, only a few genes in prokaryotic genomes qualify as universal phylogenetic markers and almost all of them have a lower information content than the 16S rRNA gene.
- The branching order of the individual phyla is not well-resolved in 16S rRNA-based trees.



# Problems with bacterial phylogeny

## Critical issues in Bacterial/Prokaryotic phylogeny

### Limitation of 16S rDNA amplification

- In this work, genomic analyses evidenced the presence of multiple and heterogeneous rRNA operons (*rrn*) within individual genomes of *Azospirillum* strains.
- Intra-genomic heterogeneity of 16S rRNA genes was higher in *A. lipoferum* 4B and led to ambiguities while trying to detect its closest relatives within the genus.



# Phylogenetic Anchors

## The limits 16S-23S rRNA gene ITS region

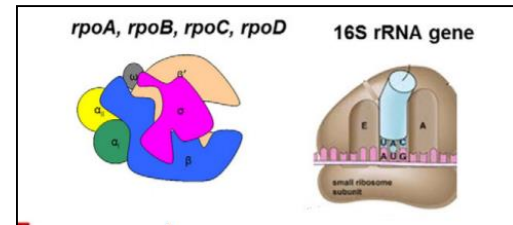
---

- In the search for alternative genetic markers, some authors have turned their attention to the 16S-23S rRNA internal transcribed spacer for a source of inter-species genetic variability in bacteria.
- However, it may suffer from the same limitations than 16S rRNA (i.e. multiple heterogeneous copies).

# Alternative genes

## Comparison of 16S rRNA, *recA*, *gyrB*, *rpoB* genes

- Molecular techniques in a comparative analysis of housekeeping genes such as *oprI*, *rpoD*, *gyrA*, *gyrB*, etc. but also 16S rRNA.
- A housekeeping gene is a gene that codes for proteins needed all the time.
- These could include:
- Heat-shock proteins such as:
- *dnaK* (heat shock protein 70, molecular chaperone DnaK);
- *gyrB* (DNA gyrase subunit B); and
- *rpoD* (RNA polymerase sigma-70 factor).



# Specific genes

## Comparison of 16S rRNA, recA, gyrB, rpoB genes

### rpoB

---

- Compared to the 16S rRNA gene sequences, variable regions were scattered along the whole fragment sequence, indicating that this fragment of the rpoB gene is more polymorphic.
- However, the comparison of rpoB sequences for species based identification has yet not been explored completely.

# Specific genes

## Comparison of 16S rRNA, recA, gyrB, rpoB genes **gyrA and gyrB sequencing**

- Among the DNA metabolic enzymes altering its topology, type II DNA topoisomerases/DNA gyrase is essential and ubiquitous.
- DNA gyrase is encoded by both gyrB and gyrA which belongs to the single gene family.
- The presence of highly conserved motifs in these gene sequences provides a useful tool for the designing of universal primers for the study of bacterial identification and diversity.
- As higher genetic variation is observed among the protein coding genes, they can be used for the identification and classification of closely related taxa.

# RibAlign:

**A software tool and database for eubacterial phylogeny based on concatenated ribosomal protein subunits**

---

- Emphasis has been placed on methods that are based on **multiple genes or even entire genomes**.
- The concatenation of **ribosomal protein sequences is one method which has been ascribed an improved resolution**.
- Since **there is neither a comprehensive database for ribosomal protein sequences nor a tool that assists in sequence retrieval and generation of respective input files for phylogenetic reconstruction programs**, **RibAlign has been developed to fill this gap**.



# Microarray technology

**Modern method for detection and hierarchical studies**

---

- DNA microarrays (which often also are called DNA or gene chips) offer the latest technological advancement for multi-gene detection and diagnostics.
- DNA microarrays were first described by Schena *et al.* (1995) for the simultaneous analyses of large-scale gene expressions by a large number of genes.
- Some microarray experiments can contain up to 30,000 target spots.
- Usually chemically synthesized oligonucleotides 20-70 nucleotides in length, can be attached to a slide and the genes they represent can all be analyzed in a single experiment.



# DNA microarrays

## DNA or gene chips

---

- DNA microarray protocols normally rely on the principle of nucleic acid hybridization, with hundreds to thousands of probes arrayed as spots *en miniature* onto a solid support.
- The solid supports themselves are usually glass microscope slides, but can also be silicon chips or nylon membranes (chemically inert).
- The spots themselves can be DNA, cDNA, or oligonucleotides.





# Designing a Microarray Experiment

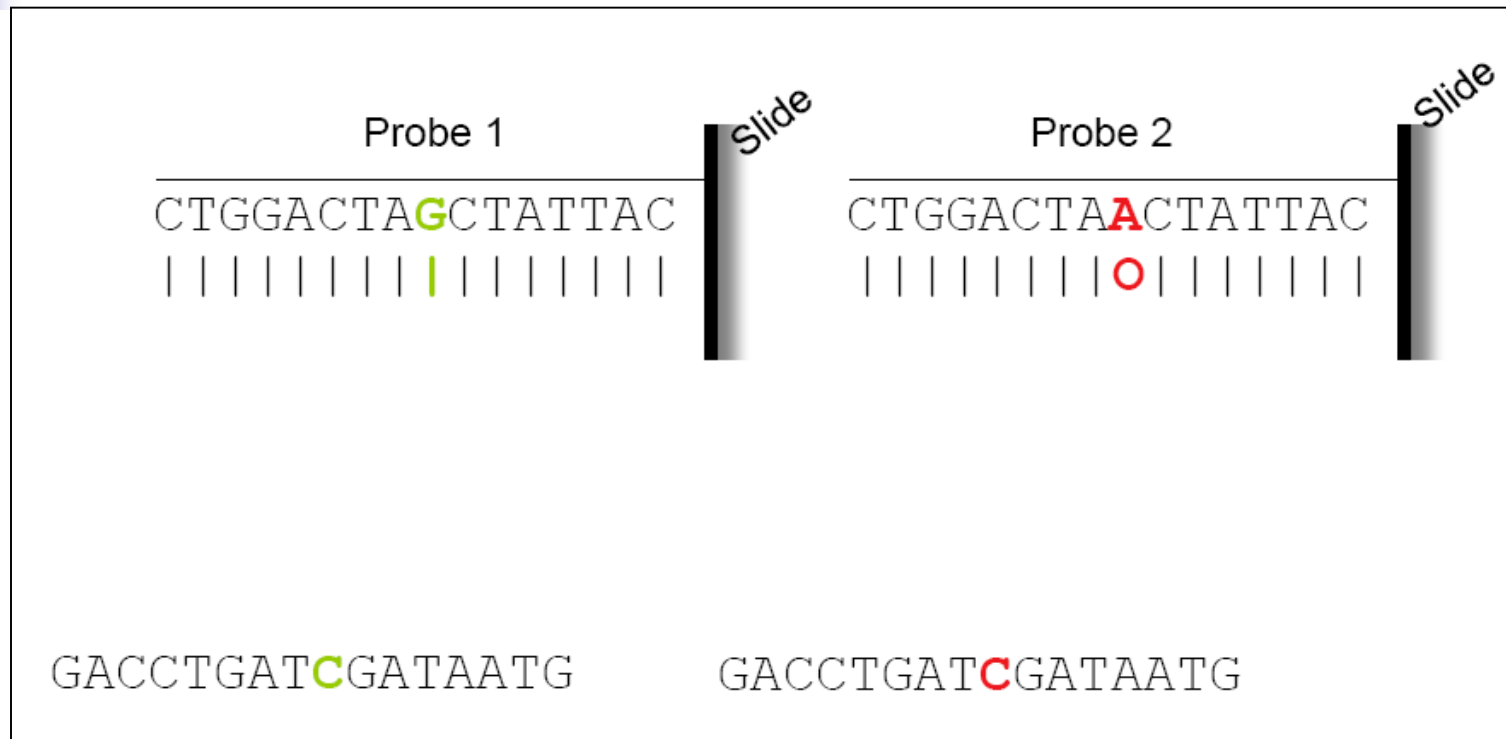
## The basic steps

---

- Spot oligos on to a specially coated slide using a robot (can be stored for several months).
- Extract sample DNA (same as with other PCR-based methods).
- Run standard PCR to amplify the probe target sequence(s) using fluorescent labels to mark the amplicon ends.
- Hybridize the PCR products with the microarray.
- Observe results using a fluorescent reader.

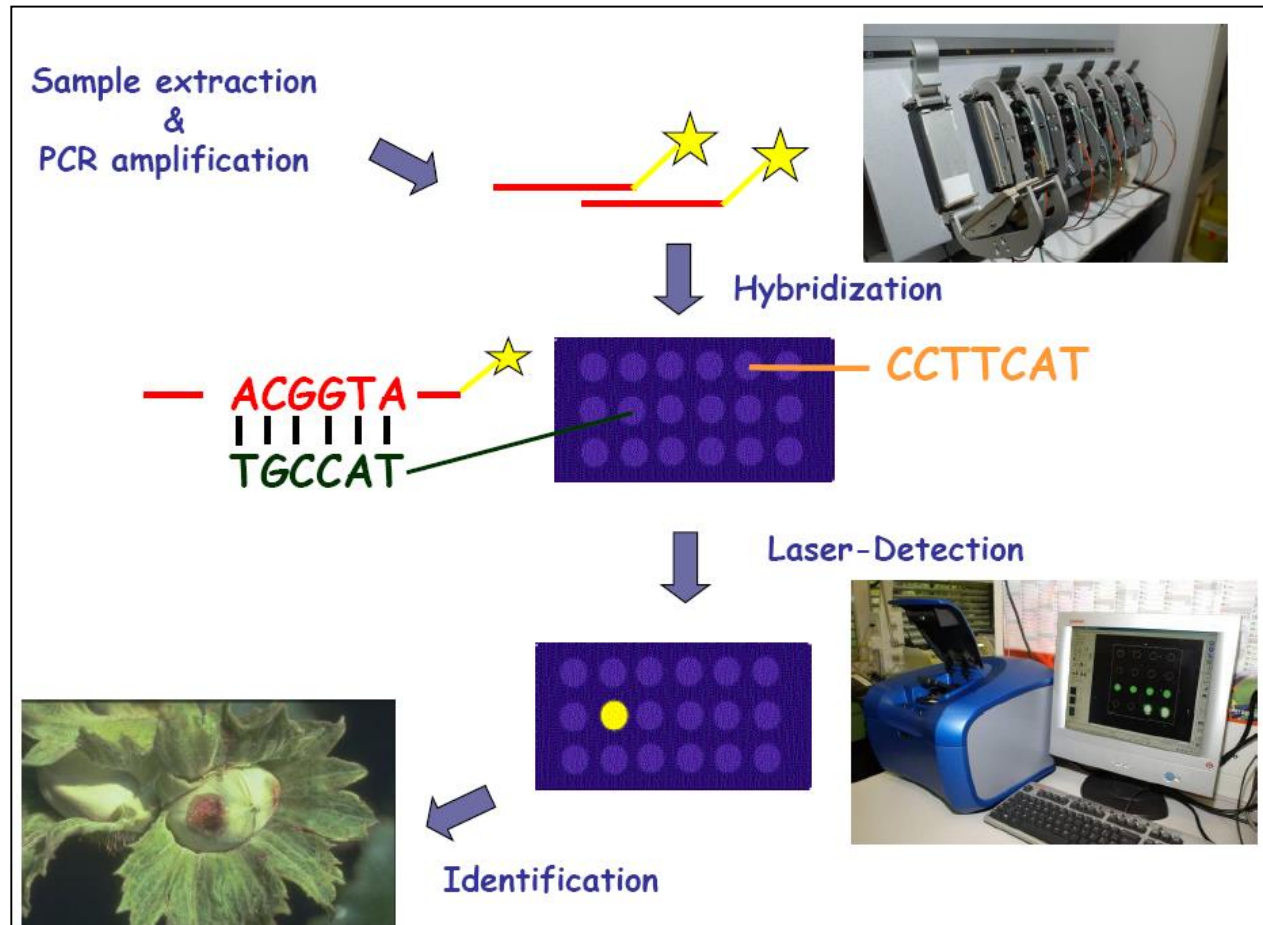
# DNA microarrays

## DNA hybridization principle

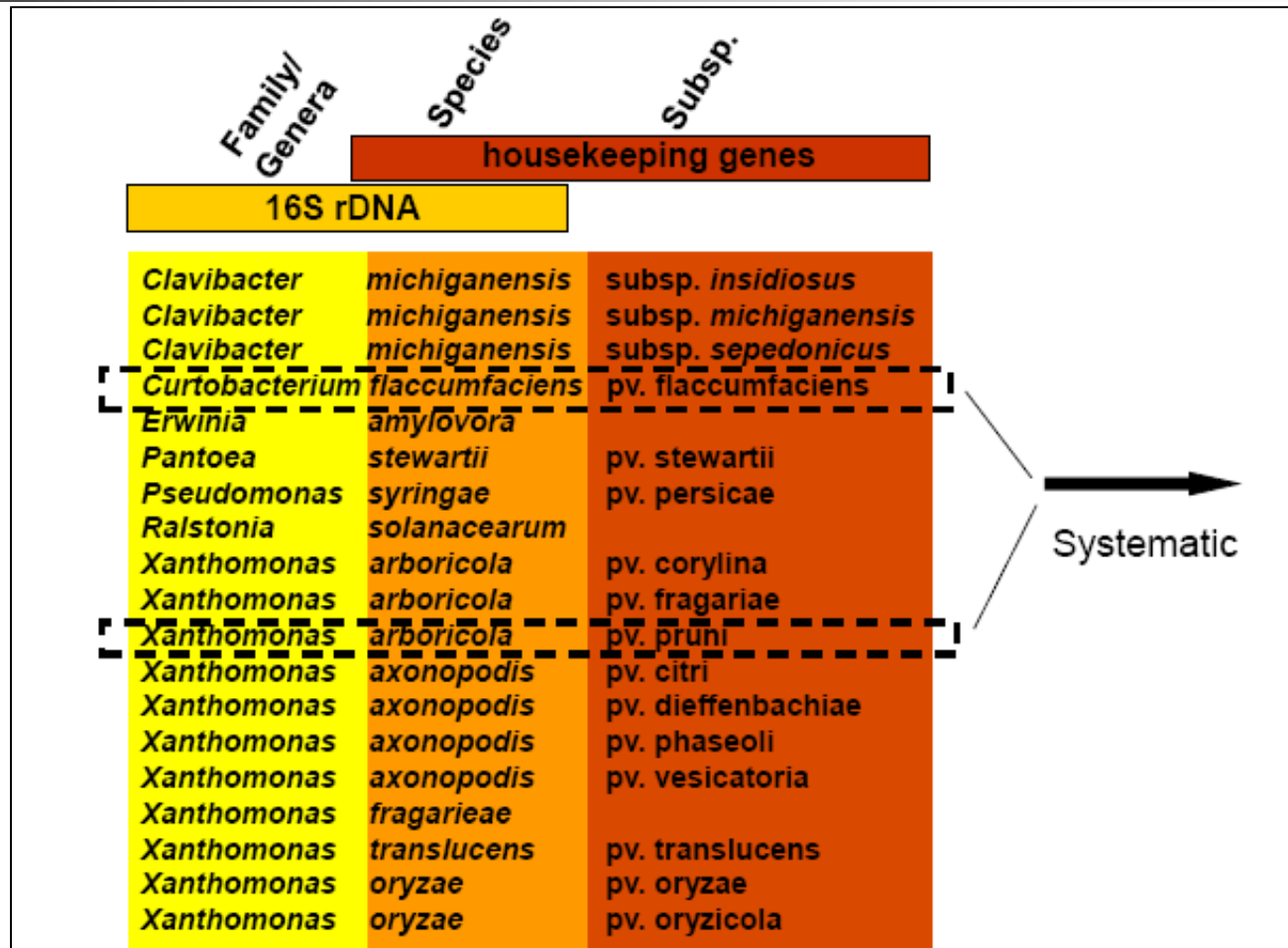


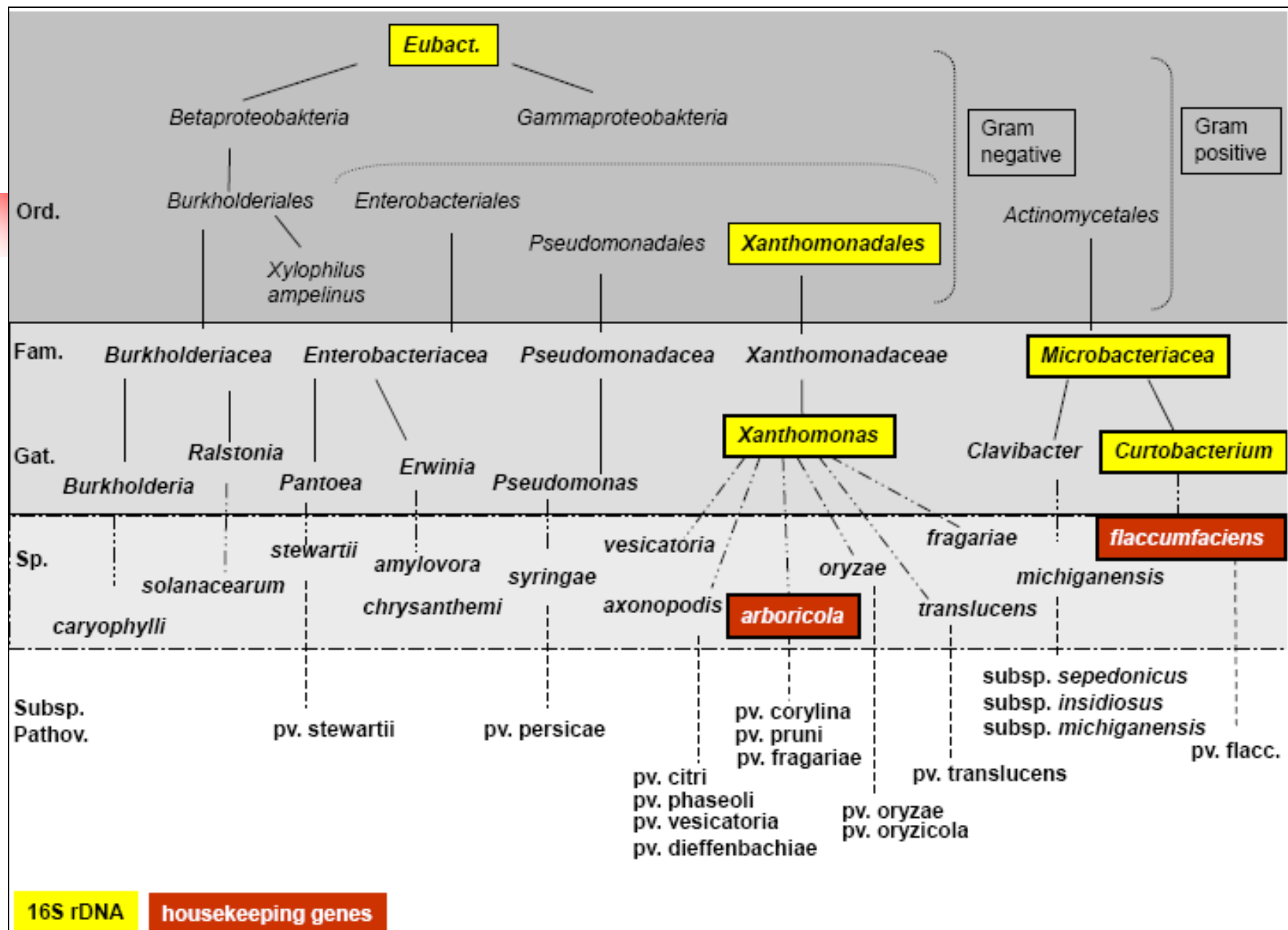
# DNA microarrays

## DNA hybridisation principle

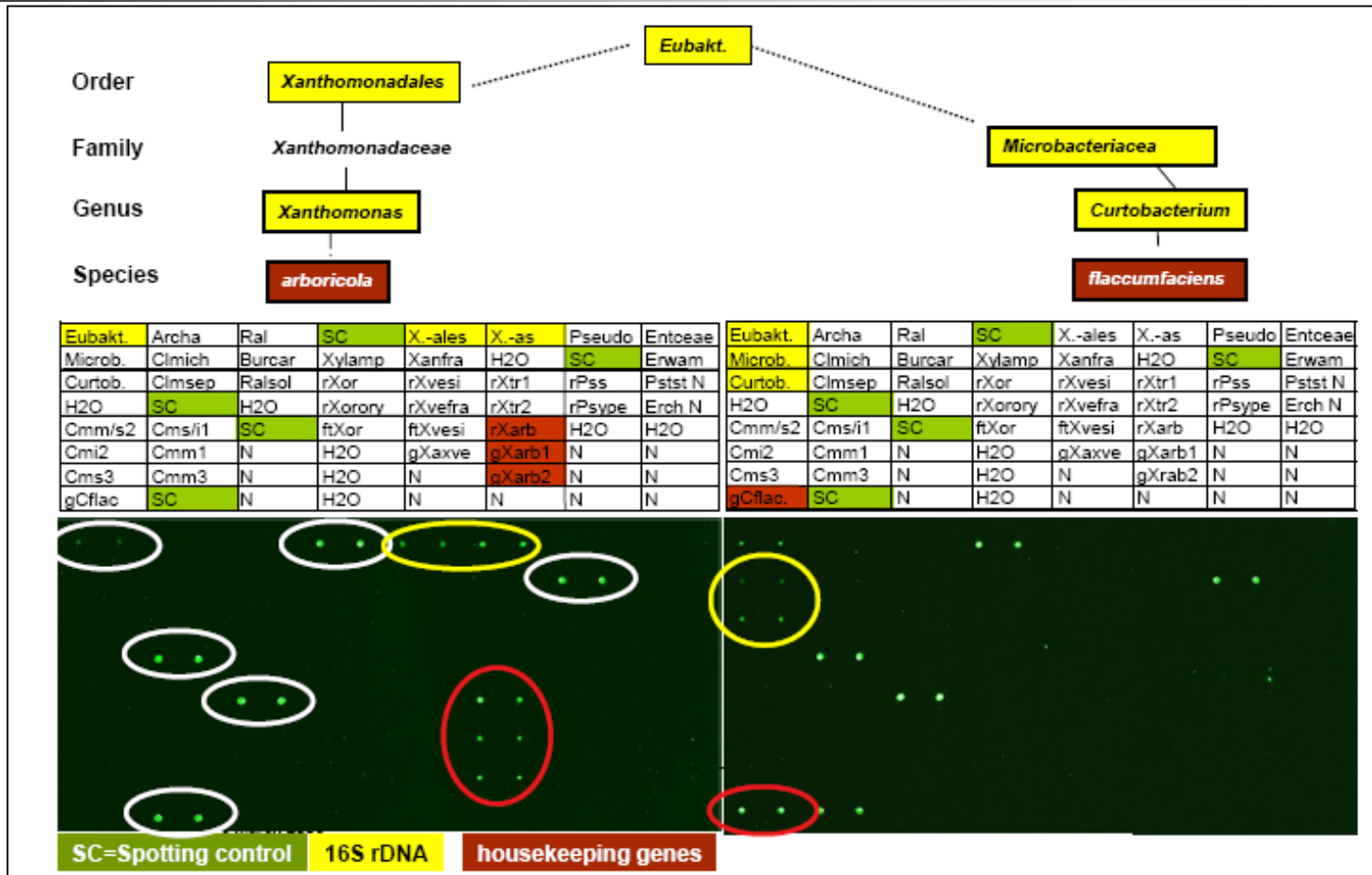


# Intelligent chip with hierarchical design



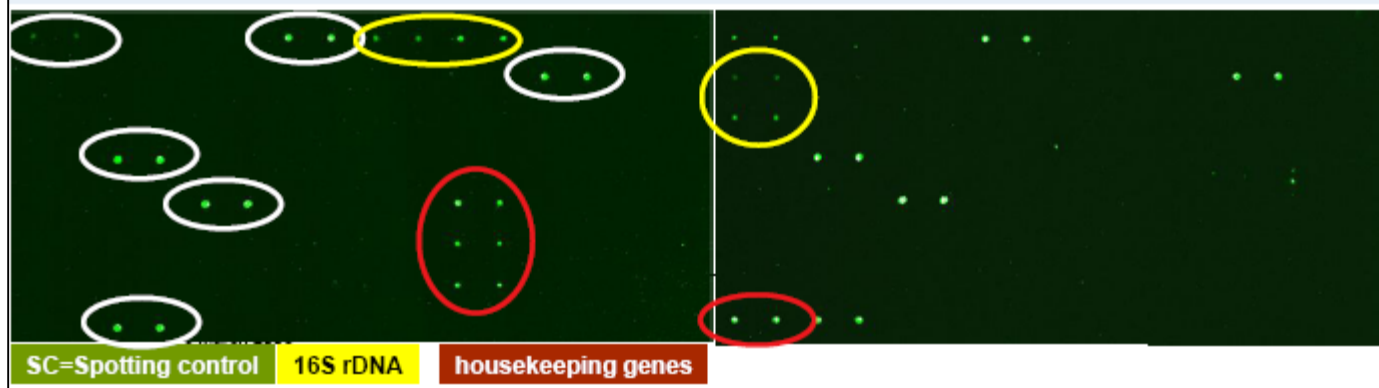


# Intelligent chip with hierarchical design



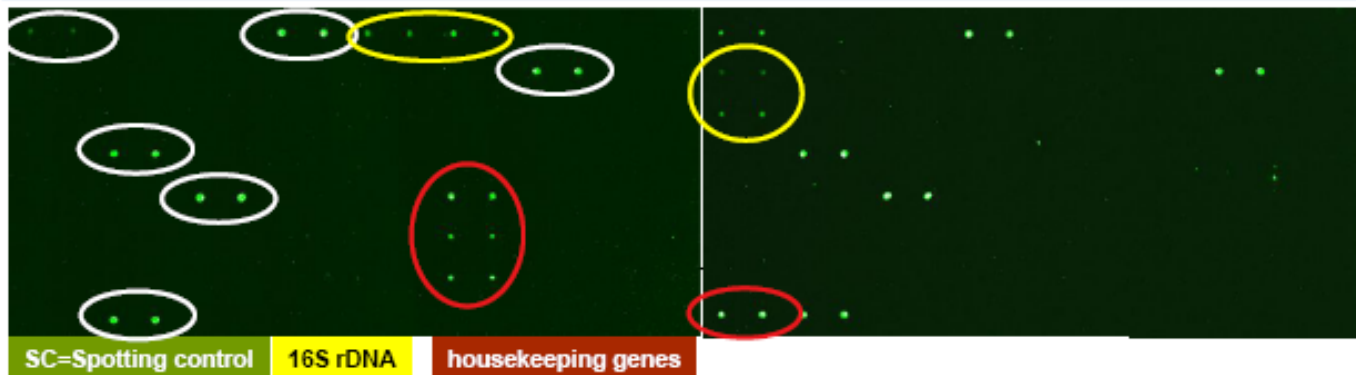
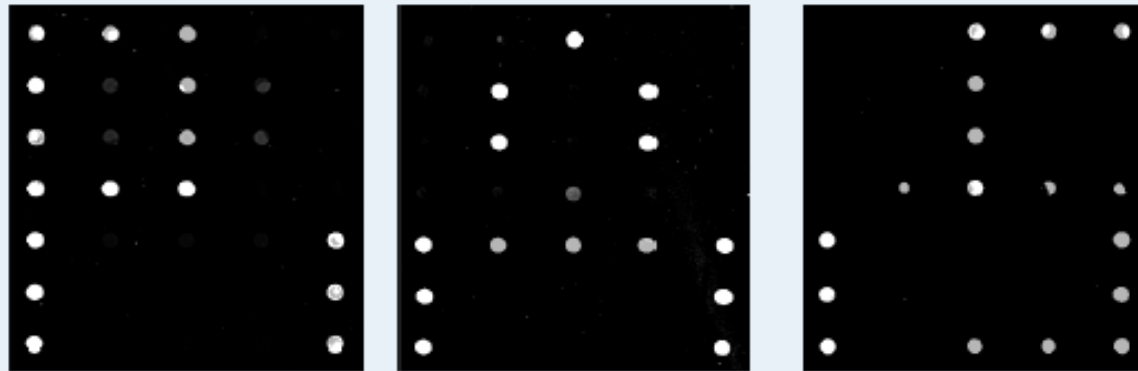
# Intelligent chip with hierarchical design

This example shows 2 separate samples, but you can also detect both bacteria in one sample on one slide.



# Intelligent chip with hierarchical design

Simplify analysis by placement of spots





# Intelligent chip with hierarchical design

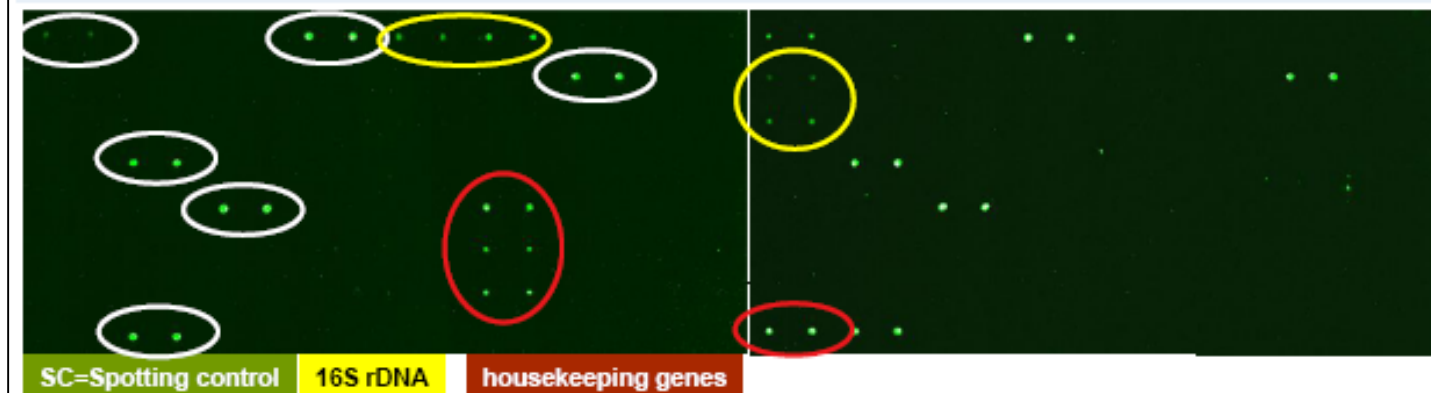
## Advantages:

Simple to interpret as + or –

contrasts with difficulty of interpreting transcriptomic microarrays for *intensity* of spots

Low cross-hybridisation (very specific)

Single multiplex PCR reaction (5 genes) to reach species level (subspecies for some target bacteria)



# Intelligent chip with hierarchical design

## Failures:

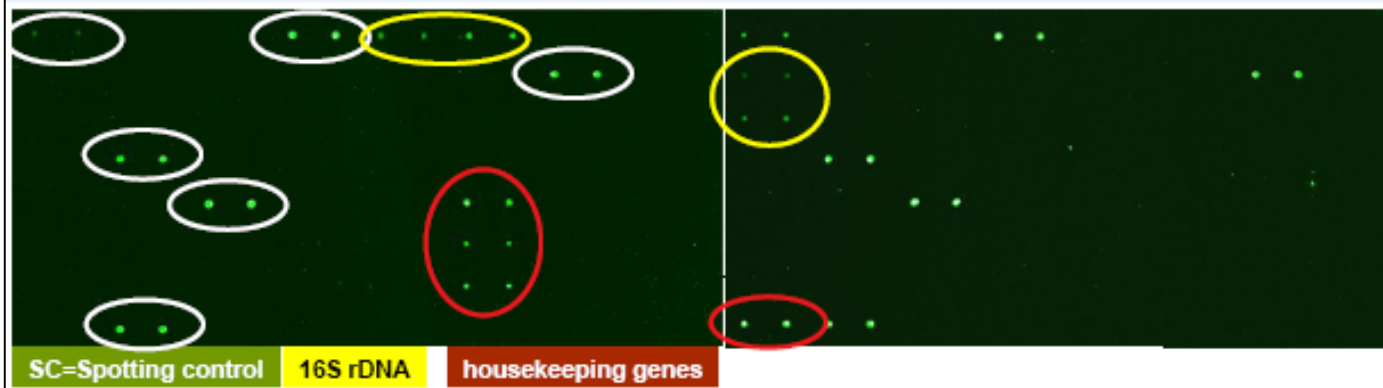
Some bacteria could not be differentiated to subspecies

Main problem was the *Xanthomonas* group

Unfortunately this is a main group in the quarantine list

Better target genes? *Maybe but a published Xanth chip has 4 gene targets just for that group. CSL advances??*

Adding more genes defeats purpose of a single PCR step





# Intelligent chip with hierarchical design

---

## Outlook:

INRA Angers (F) – small, low target, simple chips

CSL (UK) chips

PRI (NL) Padlock Probe based chips (higher specificity, quantitative option)

Genome Chips - Random Design

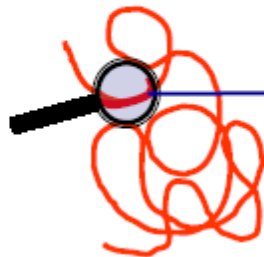
# Intelligent chip with hierarchical design



## Gene vs. Genome Principle

### Gene (intelligent design)

- specific sequence info,  
small number of genes

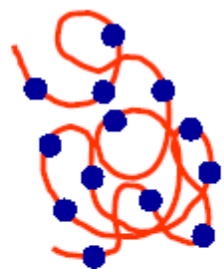


	166	190	200	210	22
AY332664	166	CCATCTAGGTC	CAATTA	CACTTCATTAC	AGAAAT
AY332669	166	CCATCTAGGTC	CAATTA	CACTTCATTAC	AGAAAT
AY332671	166	CAATCTAGGTC	CAATTA	CACTTCATTAC	AGAAAT
AY332682	166	CAATCTAGGTC	CAATTA	CACTTCATTAC	AGAAAT
AY332685	166	CAATCTAGGTC	CAATTA	CACTTCATTAC	AGAAAT
AY332684	166	CAATCTAGGTC	CAATTA	CACTTCATTAC	AGAAAT
AY332686	166	CAATCTAGGTC	CAATTA	CACTTCATTAC	AGAAAT

Co-dominant markers  
(16S, *gyrB*, etc.)

### → Genome (random design)

- Presence/Absence  
Info from total Genomes
- e.g., Potential SNPs



Org1: 1000101101001  
Org2: 1001001101100  
Org3: 1000010101101  
Org4: 0100110101111

Dominant markers

SNPs: ddA / ddG / ddT / ddC

Co-dominant markers

## Genome Chip Design

### Computer Simulation

[illegible]

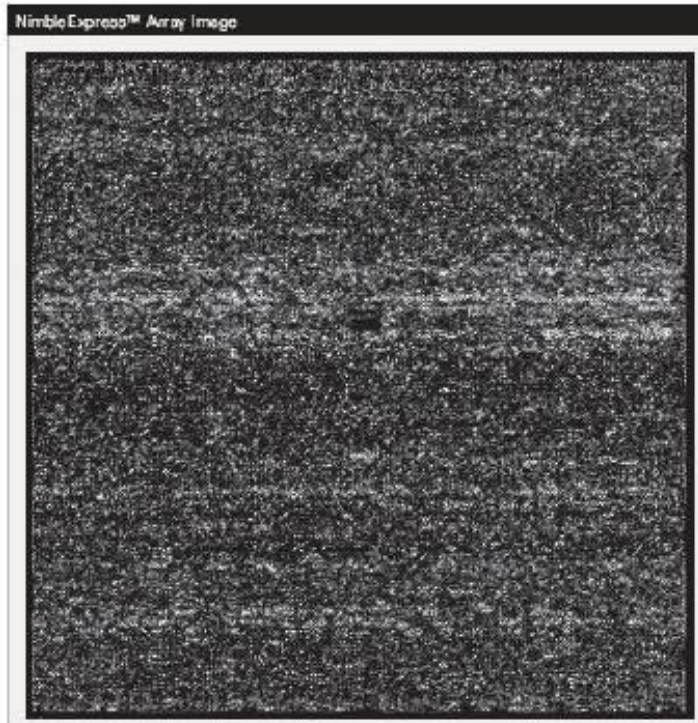
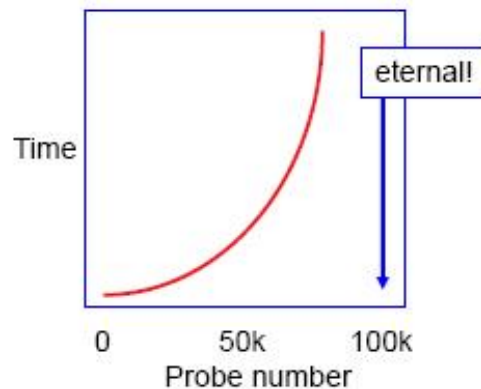
# Nimblegen-High Density Microarray

## 385'000 Feature Chip

### Nimblegen - High Density Microarray

#### • 385'000 Feature Chip

- 4-times redundant
- Design of 95'000 Probes (13mer)
  - Modification of Programme
  - 17 computers over weekend to interpret



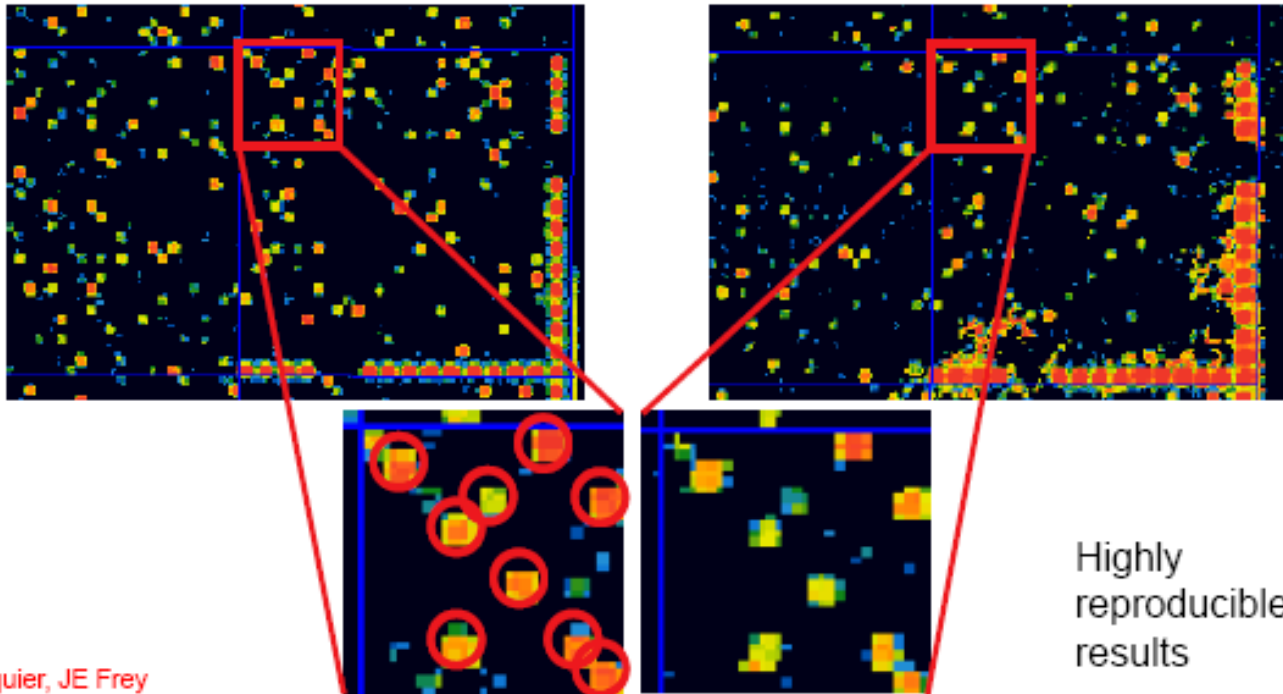
F. Pasquier, JE Frey

# Nimblegen-High Density Microarray

## 385'000 Feature Chip

### Genome Chip

385'000 feature chip: Comparison of 2 *E. coli* hybridisations:



F. Pasquier, JE Frey







# Chemical and Molecular Approaches in Bacterial Phylogeny

---

## Chemical:

- **Cell wall composition**
- **Membrane lipid signatures**
- **Electrophoretic comparison of proteins**

## Molecular:

- **Nucleic acid basic composition**
- **Nucleic acid hybridization**
- **Gene sequence comparisons**



# Molecular Approaches in Bacterial Phylogeny

---

- **Nucleic acid basic composition**
- **Nucleic acid hybridization**
- **Gene sequence comparisons**



# Nucleic acid basic composition

---

- DNA base composition indicates relatedness of organisms.
- Base composition is usually expressed as GC content.
- If the GC content differs by a small percentage the organisms are not closely related.
- The GC content itself does not always mean that organisms are related.
- For example, humans and *Bacillus* have similar GC contents but are very different organisms.



# Nucleic acid base composition

---

$$\text{Mol\% (G + C)} = \frac{\text{G} + \text{C}}{\text{G} + \text{C} + \text{A} + \text{T}} \times 100\%$$

- Determined from melting temp (thermal denaturation temperature,  $T_m$ )

Using the data:

Closely related organisms should have similar G+C ratio.



# Nucleic acid hybridization

## Method

---

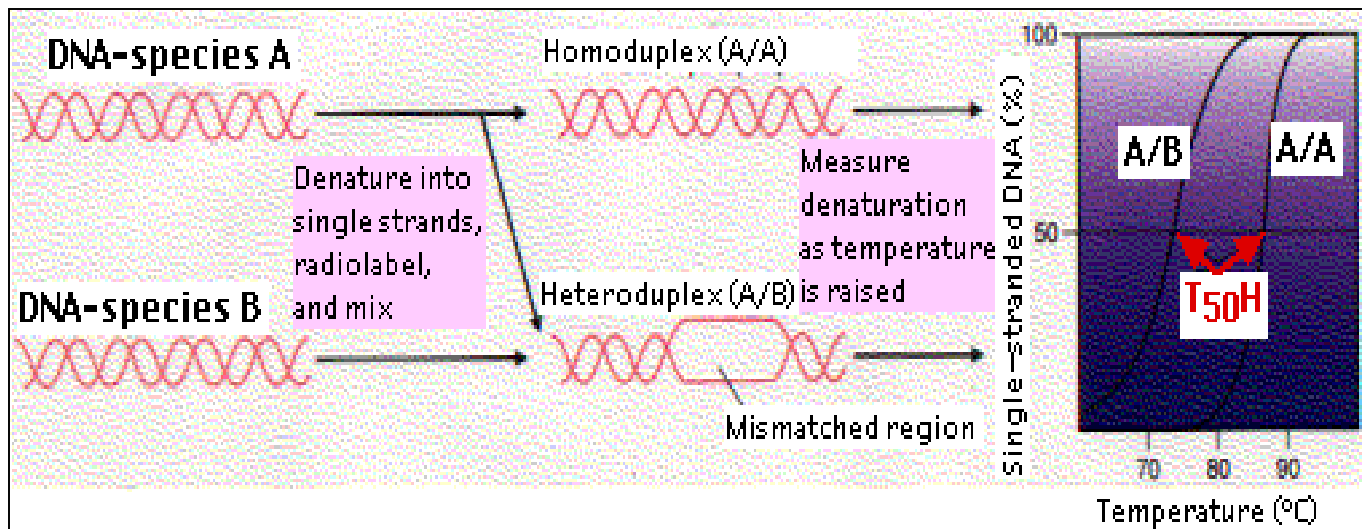
- **Two organisms:** grow one in  $[3H]$  thymine, the other one without it.
- Harvest and isolate DNA.
- Denature DNA from one organism (heating) and bind it to a filter membrane.
- Add denatured DNA from the other organism. Strands w/ complementary bases will reassociate to form dsDNA.
- Wash and add **S1 nuclease to remove any single stranded DNA.**
- **Expose to X-ray film.**
- If **closely related** they would anneal (bind) **if conditions are right** ( $60-70^{\circ}C$ ).
- You can get binding using lower temperatures ( $35-55^{\circ}C$ ) but this is just background!

Homology above 70% - same species  
Homology above 20% - same genus

# Nucleic acid hybridization

## DNA/DNA hybridization

- DNA hybridization can measure how similar the DNA of different species is—more similar DNA hybrids “melt” at higher temperatures
- The sensitivity of DNA-DNA hybridization declines rapidly as the organisms become more diverged, limiting the method to characterization of closely related strains, species and genera.



# DNA-DNA hybridization

## *Acidovorax*

- Native DNA of two *Acidovorax valerianellae* causal agent of lamb's lettuce strains, CFBP 4730T and CFBP 4723, was labelled with tritiated nucleotides (<sup>3</sup>H nucleotides) by nick-translation.
- The S1 nuclease/trichloroacetic acid method was used as indicated by Gardan *et al.*, 2000.
- The reassociation temperature was 70° C.
- Levels of DNA relatedness among *Acidovorax valerianellae* and related strains hybridization was determined at 70° C.
- ND, Not determined.

Source of unlabelled DNA	Relative binding with labelled DNA from:	
	CFBP 4730 <sup>T</sup>	CFBP 4723
<i>A. valerianellae</i> sp. nov.		
CFBP 4730 <sup>T</sup>	100	91
CFBP 4720	100	100
CFBP 4721	84	100
CFBP 4723	100	100
CFBP 4725	100	98
CFBP 4726	95	99
CFBP 4728	89	88
CFBP 4731	100	100
CFBP 4732	100	93
CFBP 4733	92	89
CFBP 4734	100	100
<i>A. anthurii</i> CFBP 3232 <sup>T</sup>	24	ND
<i>A. avenae</i> subsp. <i>avenae</i>		
CFBP 2425 <sup>T</sup>	19	ND
CFBP 1201	23	ND
<i>A. avenae</i> subsp. <i>cattleyae</i> CFBP 2423 <sup>T</sup>	35	ND
<i>A. avenae</i> subsp. <i>citrulli</i> CFBP 4459 <sup>T</sup>	29	ND
<i>A. konjaci</i> CFBP 4460 <sup>T</sup>	15	ND



# Gene sequence comparisons

## Small-subunit ribosomal RNA (SSU rRNA)

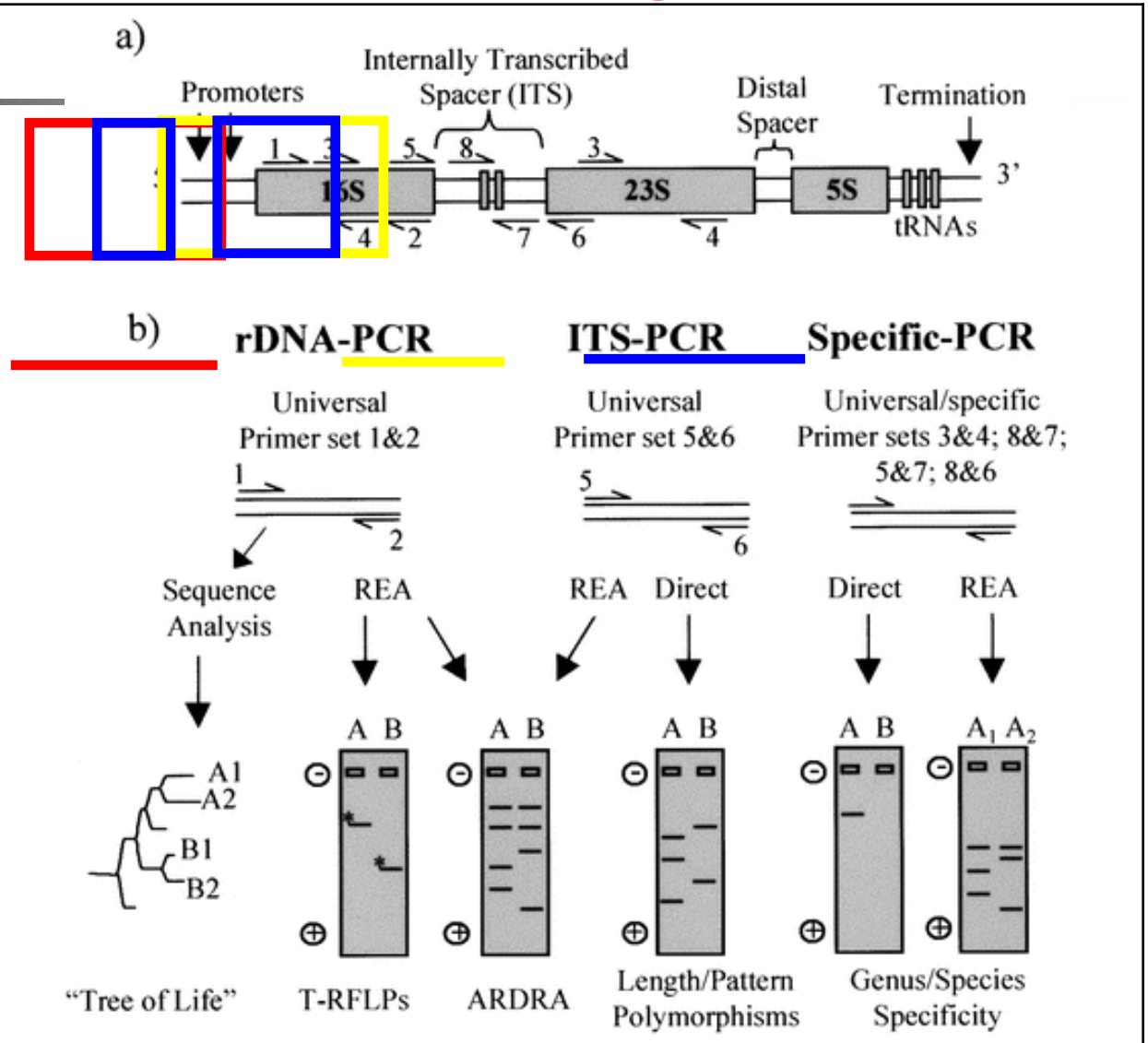
---

- DNA sequencing has provided a new approach for studying evolutionary relationships, since:
  1. All organisms have a genome.
  2. The genes that code for vital cellular functions are conserved to a remarkable degree through evolutionary time.
  3. Even these genes accumulate random changes with time (usually in the regions that are not vital for function).
- In this respect the gene changes are rather like the scars on a boxer's face - a record of the accumulated impact of time.
- So, by comparing the genes that code for vital functions of all living organisms, it should be possible to assess the relatedness of different organisms.



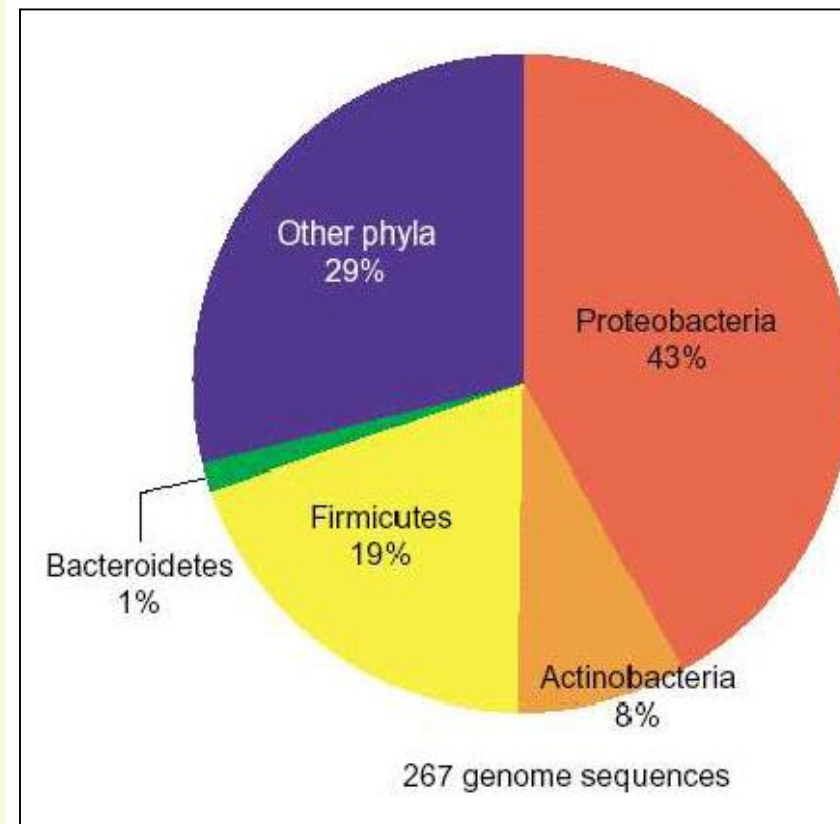
# Gene sequence comparisons

## PCR of bacterial ribosomal genes



# Biased sampling of bacterial genomes

- A phylum of bacteria comprised of three classes:
  1. **Bacteroides,**
  2. **Flavobacteria, and**
  3. **Sphingobacteria.**
- These **gram-negative bacteria** found primarily in the intestinal tracts and mucous membranes of warm-blooded animals.





# DNA sequencing

## Small-subunit ribosomal RNA (SSU rRNA)

---

- The gene most commonly used for this codes for the RNA in the **small subunit (SSU)** of the **ribosome**.
- Some regions of this **SSU rRNA** (also termed **16S rRNA**) are **highly conserved in all organisms**, whereas
- Other regions are **more variable**.

# Phylogenetic Trees

## Phylogenetic resolution

**Highly conserved sequences contain too little information to resolve close relationships**

**Phylogenetic resolution**

- Use variable regions to compare closely related sequences
  - Badly conserved sequences contain too much noise to resolve distant relationships
- Use conserved regions to compare distantly related sequences
  - Highly conserved sequences contain too little information to resolve close relationships

**Hi-res**

```
ACCTTGAC-----ACAGACTAGCGTGGCGACTCGATCAT-----TCCAAATCTAGGGGAATG-CCGAAAC
TCCGTTAACCTACGGCCTTATCGGGGACATTGAAAT--ATAACGA-----CCTCCTATAAGACGTGTG-GCAAGCT
ATCTTTAGACGA-----AATG-----CCCTATAGATCGGGCCACAAAGAC-----GTCTG--AGTGTTCGT
TCTTTTAACT-----GACGTAAACCATCACCGACTGCAATGAAGAGCG-----ACCCGTGGAGGTCCCTATA-ACGA--
ACCTCAAGA-----CCCACTAAGTGGGGGGCA-----TAG-----CCGCAACGGGGTGGGCGTGTCTAA
CCTCCAGGCGA-----CACT-----TATTGTTGCA-----CCTCAATACCGGCTTATCTACTACCAAGGGGCG--CATGGCCGT
TCTGTGCGCGGTTTACCAGACTGGAACATAAGTAAAGAGGACATATAAGATCT-----GCGAGTCCGTCGGAC
TAAAGAT-----GTCCCTAACGACATCGACTATTAACT-----TTCAATTTGACGTGAACA-GCCGGC
TACCGCAGGTGG-----GATGACACAATTG-----AGCTGTTAGCTGTGGGCCAATACCGGCTTTGGGGGTGGTTCACTCATC
GCCCGTACACAA-----GATTGTCGGCACTGTTGCAAAAGTACCGCGCTAAACCGTGTCCCTACTTCC--GGCT--CGTGTATC
```

**Lo-res**

```
IIITEFMTYGHLLDYLRECHQEVHAWVLLYNATQISSAMEYLEKGNPIHRLAARNCLVGEHHLVAVADPGLSRLMTGDTTAAHAGARF
IIITEFMTYGHLLDYLRECHQEVHAWVLLYNATQISSAMEYLEKGNPIHRLAARNCLVGEHHLVAVADPGLSRLMTGDTTAAHAGARF
IIITEFMTYGHLLDYLRECHQEVHAWVLLYNATQISSAMEYLEKGNPIHRLAARNCLVGEHHLVAVADPGLSRLMTGDTTAAHAGARF
IIITEFMTYGHLLDYLRECHQEVHAWVLLYNATQISSAMEYLEKGNPIHRLAARNCLVGEHHLVAVADPGLSRLMTGDTTAAHAGARF
IIITEFMTYGHLLDYLRECHQEVSAVLLYNATQISSAMEYLEKGNPIHRLAARNCLVGEHHLVAVADPGLSRLMTGDTTAAHAGARF
IIITEFMTYGHLLDYLRECHQEVSAVLLYNATQISSAMEYLEKGNPIHRLAARNCLVGEHHLVAVADPGLSRLMTGDTTAAHAGARF
IIITEFMTYGHLLDYLRECHQEVHAWVLLYNATQISSAMEYLEKGNPIHRLAARNCLVGEHHLVAVADPGLSRLMTGDTTAAHAGARF
```



# 16S ribosomal RNA

## Comparisons of the sequence

---

- The nucleotide base sequence of the gene which codes for 16S ribosomal RNA is becoming an important standard for the definition of bacterial species.
- Comparisons of the sequence between different species suggest the degree to which they are related to each other.
- Differences in the DNA base sequences between different organisms can be determined quantitatively, such that a phylogenetic tree can be constructed to illustrate probable evolutionary relatedness between the organisms.



# 16S ribosomal RNA

## Signature sequences

---

- Specific base sequences in the rRNA known as **signature sequences** were commonly found in particular groups of organisms.
- Signatures are generally found in **defined regions** of the **16S rRNA molecule**, but are only readily apparent when the computer scans sequence alignments.
- They allow for placing unknown organisms in the correct major **phylogenetic group**, and can be useful for **constructing genus and species-specific nucleic acid probes** which are used **exclusively for identification purposes in microbial ecology and diagnostics**.



# 16S ribosomal RNA

## Signature sequences

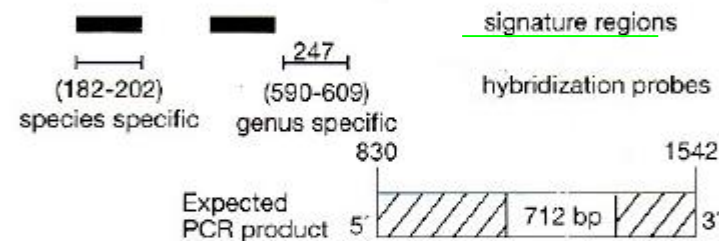
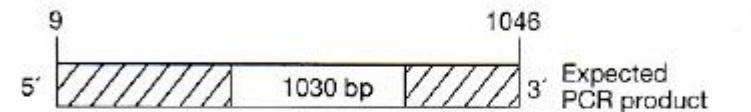
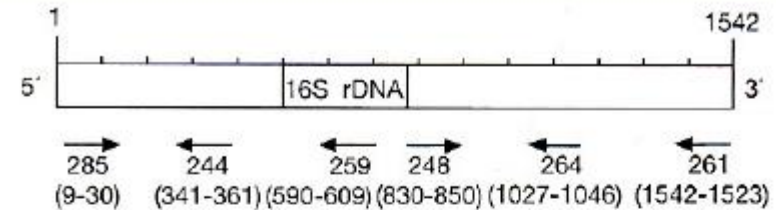
---

- Highly conserved organisms are classified as:
  1. Separate species if their sequences show less than 98% homology, and
  2. Different genera if their sequences show less than 93% identity.



# Mycobacterium speciation using 16S rRNA gene

- Species specific vs genus specific regions of 16S rRNA gene
- Examine sequence alignment



285: 5' GAGAGTTTGATCCTGGCTCAG 3'  
 244: 5' CCCACTGCTGCCTCCCGTAG 3'  
 259: 5' TTTCACGAACAACGCGACAA 3'  
 248: 5' GTGTGGGTTTCCTTCCTTGG 3'  
 264: 5' TGCACACAGGCCACAAGGGA 3'  
 261: 5' AAGGAGGTGATCCAGCCGCA 3'

## species-specific

M.tb complex: 5' ACCACAAGACATGCATCCCG 3'  
 M.avium: 5' ACCAGAAGACATGCGTCTTG 3'  
 M.intracellulare: 5' ACCTAAAGACATGCGCCTAA 3'  
 M.leprae: 5' ATAGGACTTCAAGGCGCATG 3'  
 M.genavense: 5' CCACAAAACATGCGTTCCGTG 3'





# 16S ribosomal RNA

## Sequence methodology

---

- Today, 16S rRNA sequences are more readily obtained by amplifying nearly full length genes with the polymerase chain reaction (PCR) and "universal primers" specific for conserved regions of the 16S rDNA sequence.
- The reaction product can be sequenced directly or cloned into a plasmid vector and then sequenced.
- In current methods, the genes for rRNA, rather than RNA itself are sequenced.



# 16S ribosomal RNA

## Sequence methodology

---

- Since thousands of full and partial 16S sequences are available through the Web, **classifying an unknown bacterium** is readily accomplished using one of the many comparison and search algorithms available online (e.g. Blastn at <http://www.ncbi.nlm.nih.gov>).
- It usually takes **about a day or two to obtain sequences for an unknown organism** if the equipment and technical expertise is in place, **versus several days to weeks** using conventional phenotypic testing.

An **algorithm** is a step by step procedure to solve logical and mathematical problems. **There are several algorithms used to infer phylogenetic trees**, but the most widely-used algorithms fall into three main categories: **Distance algorithms, Maximum parsimony algorithms and Likelihood algorithms.**

# 16S ribosomal RNA

## Sequence comparisons

- When only **four species** were compared with each other, **a relatively short segment stood out** as appearing to be **"frame-shifted"** when comparing *Pseudomonas fluorescens* with a group of **three enterics**.
- This situation is shown as follows with the **nucleotide bases of the segment in question shown in red**.

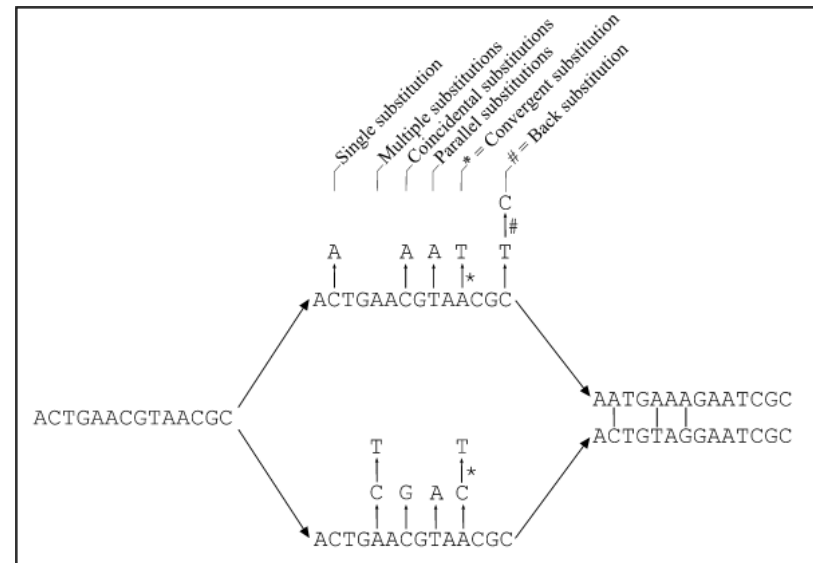
<i>Pseudomonas fluorescens</i>	...gctaataccgcata <b>acgtcctacg</b> ggagaaagcagggg...
Our new organism, shown below as "AH"	...gctaataccgcata <b>acgtcgcaag</b> accaaagcggggg...
<i>Buchivia aquatica</i>	...gctaataccgcgta <b>acgtcgaaag</b> accaaagcggggg...
<i>Edwardsiella tarda</i>	...gctaataccgcata <b>acgtcgcaag</b> accaaagtggggg...

Databases of various gene sequences are found on the web. **Genbank's database** was used as the source of the above sequences.

# Sequence comparisons

## Comparison of two homologous DNA sequences

- Two homologous DNA sequences which descended from an ancestral sequence and accumulated mutations since their divergence from each other.
- Note that although 12 mutations have accumulated, differences can be detected at only three nucleotide sites.
- (from Fundamentals of Molecular Evolution, Wen-Hsiung Li and Dan Graur, 1991).



# 16S ribosomal RNA

## Sequence comparisons

- When a 1308-base stretch of that part of the chromosome which codes for 16S ribosomal RNA was lined up and analyzed to find the extent to which the above four organisms differed from each other, the percent difference between any two organisms was determined, and the results are summarized as follows:

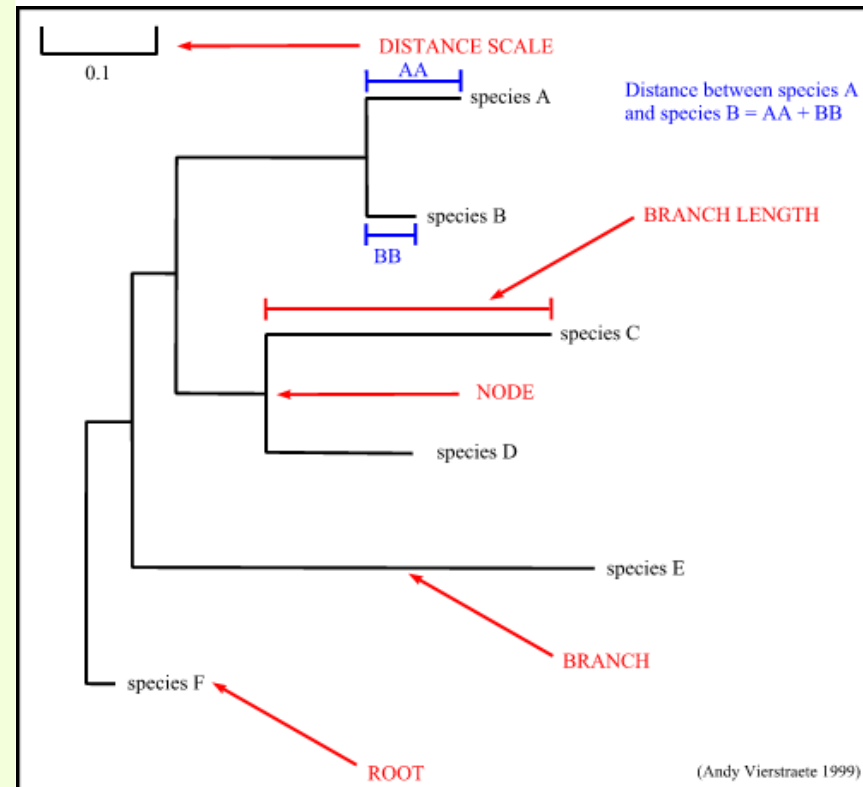
PF	PF			
AH	14.8*	AH		
BA	14.5	3.2	BA	
ET	14.9	4.3	5.0	ET

\* An example: The same bases appear in the same sequence, position by position, for each of the two organisms except for 14.8% of the time.

# Construction of a phylogenetic tree

## Terminology

- **Tips** (sometimes called **leaves** or **terminal nodes** or **nodes**): represents a taxonomic unit. This can be a taxon (an existing species) or an ancestor (unknown species: represents the ancestor of 2 or more species).
- **Branch:** defines the relationship between the taxa in terms of descent and ancestry.
- **Topology:** is the branching pattern.
- **branch length:** often represents the number of changes that have occurred in that branch.
- **Root:** is the common ancestor of all taxa.
- **Distance scale:** scale which represents the number of differences between sequences (e.g. 0.1 means 10% differences between two sequences).



Distance between species A and species B = AA + BB.



# Construction of a phylogenetic tree

## The scale bar

---

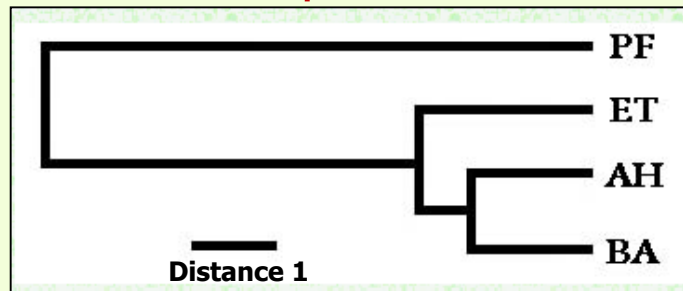
- The horizontal lines are branches and represent evolutionary lineages changing over time.
- The longer the branch in the horizontal dimension, the larger the amount of change.
- The bar at the bottom of the figure provides a scale for genetic change.
- The bar number '0.05' shows the length of branch that represents an amount genetic change of 0.05.
- 1. The units of branch length are usually nucleotide substitutions per site – that is the number of changes or 'substitutions' divided by the length of the sequence (although they may be given as % change, i.e., the number of changes per 100 nucleotide sites).

# 16S rRNA sequence comparison

## Construction of a phylogenetic tree

### The scale bar

- The results of "cluster analyses", such as the UPGMA method, are often referred to as "dendrograms".
- A scale bar usually indicates distances.
- The scale bar represents the percentage of dissimilarity (distance) between two aligned sequences.
- The scale bar indicates the number of changes per nucleotide per unit branch length.
- The bar at the bottom signifies approximately 1% base difference.
- Scale bar indicates 1% sequence dissimilarity (one substitution per 100 nt).





# 16S rRNA sequence comparison

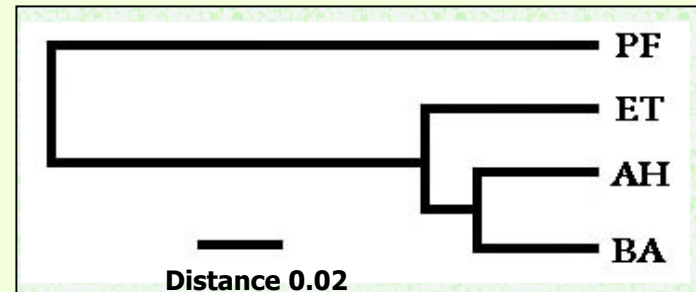
## Construction of a phylogenetic tree

### The scale bar

- The scale bar 0.1 means 0.1 nucleotide substitutions per site (0.1 change per nucleotide=10% differences between two sequences).
- The actual value will depend on the branch lengths in the tree.
- The scale bar=0.02 represents 0.02% nucleotide substitutions per nucleotide. i.e. 2% differences between two sequences).
- The scale bar=0.022 represents an estimated 22 base substitutions per 1000 nt positions according to the Kimura index.

Human **ATG****T**TGACTC  
Mouse **ATG****C**TGACTC

There is one site that is different between the two sequences, and we could say that based upon this tiny sample there are  $1/10 = 0.1$  substitutions per site.

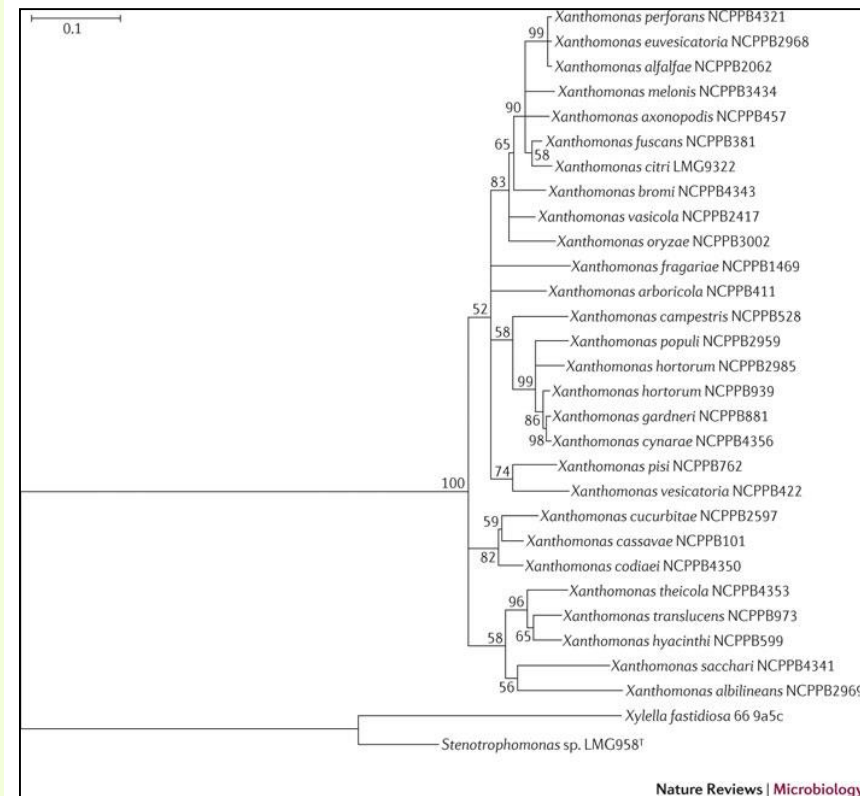


# 16S rRNA sequence comparison

## Construction of a phylogenetic tree

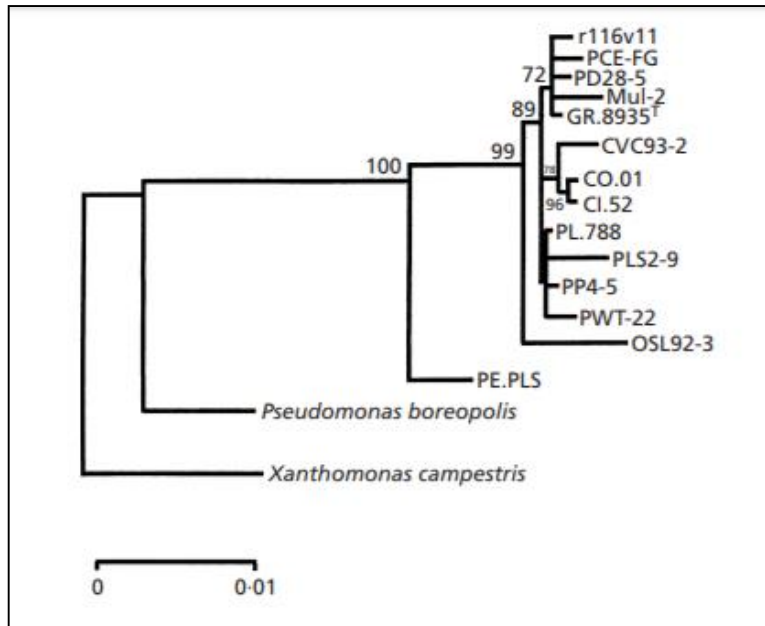
### The scale bar

- This neighbour-joining tree is based on the DNA gyrase subunit B (*gyrB*) gene sequence of *Xanthomonas* spp., *Xylella fastidiosa* and a *Stenotrophomonas* sp.
- Bootstrap values (for 1,000 replicates) are given at the nodes, and branches with <50% bootstrap support were collapsed to better reveal the phylogenetic structure.
- The scale bar corresponds to 0.1 change per nucleotide.

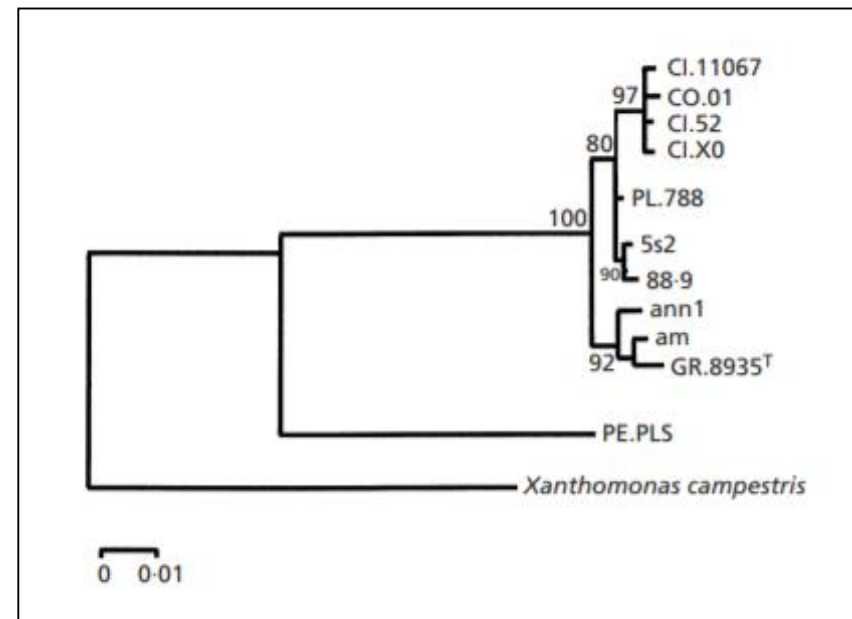


Nature Reviews | Microbiology

# Phylogenetic relationships of *Xylella fastidiosa* strains from different hosts, based on 16S rDNA and 16S-23S intergenic spacer sequences



Phylogenetic tree constructed using the neighbor joining method, based on 16S rDNA sequence data for *Xylella fastidiosa* and *Pseudomonas boreopolis*, with *Xanthomonas campestris* as the outgroup. Gaps and missing information excluded from the analysis. The numbers above the branches are bootstrap values obtained for 1000 replications (expressed as percentages; only values greater than 70% are shown). Bar, 1% sequence divergence.



Phylogenetic tree constructed using the neighbor joining method, based on 16S-23S intergenic spacer sequence data for *Xylella fastidiosa*, with *Xanthomonas campestris* as the outgroup. Gaps and missing information were excluded from the analysis. The numbers above the branches are bootstrap values obtained for 1000 replications (expressed as percentages; only values greater than 70% are shown). Bar, 1% sequence divergence.



# **Classification systems**

## **History of classification systems**

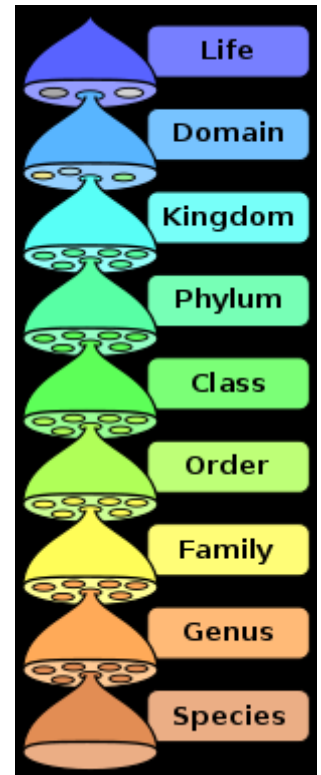
---

**From traditional to natural classifications**  
**Two-kingdom to six-kingdom systems**

# Kingdom

## Definition of the rank kingdom

- In biology, **kingdom** (Latin: regnum, pl. **regna**) is a taxonomic rank, which is either the **highest rank** or in the more recent three-domain system (**Woese three-domain system**) the rank, **below domain**.
- Kingdoms are divided into smaller groups called **phyla** (in zoology) or **divisions** in botany.
- The complete sequence of ranks is: **life, domain, kingdom, phylum, class, order, family, genus and species**.



**Domains** - placed above the phylum and kingdom levels of classification.

# Bacterial nomenclature

## The primary objective of Code of Nomenclature of Bacteria(now Prokaryotes)

- The Bacteriological Code governs names of prokaryotes in the ranks of:
- Class, Subclass, Order, Suborder, Family, Subfamily, Tribe, Subtribe, Genus, Subgenus, Species and Subspecies.
- Taxa above the rank of Class (Phylum, Kingdom, Division and Domain) are not covered by the Code.

**Domain:** The highest of taxonomic rank ('80s)

Kingdom (not used by most bacteriologists), 1969

Phylum or division of the kingdom

**Class**

Order

Family (related genera)

Genus (related species) plural: Genera

Species (related strains) both singular & plural

Subspecies

# The Standards

## Pathovar system of nomenclature

### The preferred names of infrasubspecific subdivisions

<ul style="list-style-type: none"><li>■ <b>Domain:</b> The highest of taxonomic rank ('80s)</li><li>■ <b>Kingdom</b> (Not used by most bacteriologists), 1969</li><li>■ <b>Phylum or division</b> of the kingdom</li><li>■ <b>Class</b></li></ul>	Names not covered by Code
<ul style="list-style-type: none"><li>■ <b>Order</b></li><li>■ <b>Family</b>(related genera)</li><li>■ <b>Genus</b>(related species) plural: Genera</li><li>■ <b>Species</b>(related strains) both singular &amp; plural</li><li>■ <b>Subspecies</b></li></ul>	Names covered by Code
<ul style="list-style-type: none"><li>■ <b>Biovar</b> (usual abbreviation: <b>bv.</b>),</li><li>■ <b>chemoform</b>, <b>chemovar</b>,</li><li>■ <b>cultivar</b>(usual abbreviation: <b>cv.</b>),</li><li>■ <b>morphovar</b>,</li><li>■ <b>pathovar</b> (usual abbreviation: <b>pv.</b>),</li><li>■ <b>phagovar</b>,</li><li>■ <b>serovar</b>.</li></ul>	Names not covered by Standards

# Domain Bacteria

## Bacterial phylum

The bacterial phyla are the major lineages (phyla or divisions) of the domain Bacteria

- "Abditibacteriota"
- "Acidobacteria"
- "Actinobacteria"
- "Candidatus Aminicenantes"
- "Aquificae"
- "Armatimonadetes"
- "Bacteroidetes"
- "Balneolaeota"
- "Caldiserica"
- "Calditrichaeota"
- "Chlamydiae"
- "Chlorobi"
- "Chloroflexi"
- "Chrysiogenetes"
- "Candidatus Cloacimonetes"
- "Coprothermobacterota"
- "Candidatus Cryoserica"
- "Cyanobacteria"
- "Deferribacteres"
- "Deinococcus-Thermus"
- "Candidatus Dependitiae"
- "Dictyoglomi"
- "Elusimicrobia"
- "Candidatus Eremiobacteraeota"
- "Candidatus Fermentibacteria"
- "Fibrobacteres"
- "Firmicutes"
- "Fusobacteria"

- "Fusobacteria"
- "Gemmatimonadetes"
- "Candidatus Goldbacteria"
- "Candidatus Kapabacteria"
- "Kiritimatiellaeota"
- "Candidatus Krumholzibacteriota"
- "Lentisphaerae"
- "Candidatus Margulisbacteria"
- "Candidatus Mcinerneyibacteriota"
- "Candidatus Melainabacteria"
- "Candidatus Microgenomates"
- "Nitrospirae"
- "Nitrospirae"
- "Candidatus Omnitrifica"
- "Candidatus Parcubacteria"
- "Candidatus Parcunitrobacteria"
- "Candidatus Peregrinibacteria"
- "Planctomycetes"
- "Proteobacteria"
- "Rhodothermaeota"
- "Spirochaetes"
- "Candidatus Sumerlaeota"
- "Synergistetes"
- "Tenericutes"
- "Thermodesulfobacteria"
- "Thermomicrobia"
- "Thermotogae"
- "Verrucomicrobia"





# Classification systems

## Based on Kingdoms

---

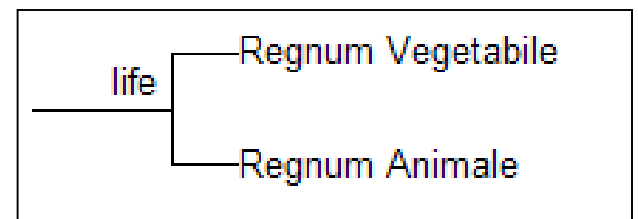
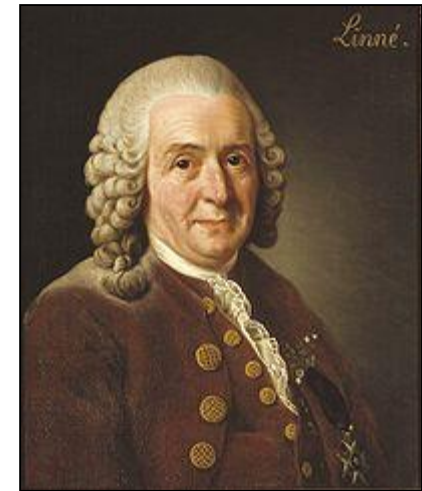
- Historically, the number of kingdoms in widely accepted classifications has grown from **two to six**:
  1. Two-kingdoms
  2. Three-kingdoms
  3. Four-kingdoms
  4. Five-kingdoms
  5. Six-kingdoms
    - 5.1. Cavalier-Smith's six kingdoms
- However, **phylogenetic research** from about 2000 onwards does not support any of the traditional systems.

# Traditional system of classification

## Two kingdoms

Proposed by C. Linnaeus, 1735

- A traditional (artificial but not a natural one) system of classification developed by Carl Linnaeus (1707-1778).
- Originally there were only two kingdoms:
  1. Plants
  2. Animals
- The invention of the microscope led to the discovery of new organisms.



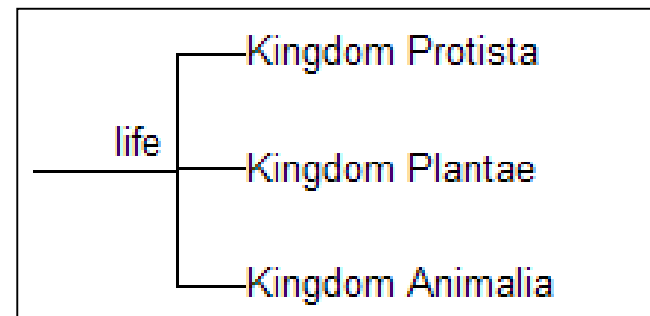
# Traditional system of classification

## Three kingdoms

Proposed by E. Haeckel, 1866

- In 1866, following earlier proposals by Richard Owen, John Hogg and Ernst Haeckel proposed a third kingdom of life, the protists.
- Haeckel revised the content of this kingdom a number of times before settling on a division based on whether organisms were:
  1. Unicellular (Protista), or
  2. Multicellular (animals and plants).

Phytoplanktons, also known as microalgae microscopic are marine algae. Some phytoplankton are bacteria, some are protists, and most are single-celled plants. These plants produce oxygen as a byproduct of photosynthesis. Phytoplankton produce at least 50% of the Earth's oxygen.

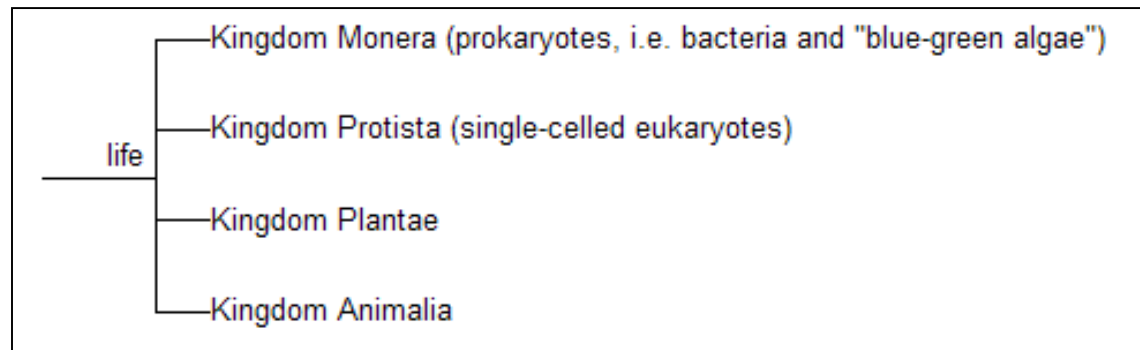


# Traditional system of classification

## Four kingdoms

Proposed by H. F. Copeland, 1938

- The development of microscopy, and the electron microscope in particular, revealed an important distinction between those unicellular organisms whose cells do not have a distinct nucleus, prokaryotes, and those unicellular and multicellular organisms whose cells do have a distinct nucleus, eukaryotes.
- In 1938, Herbert F. Copeland proposed a four-kingdom classification, moving the two prokaryotic groups, bacteria and "blue-green algae", into a separate Kingdom Monera.





# Natural system of classification

## History of descent

---

- When the natures of objects are defined by a **common history** then there is a **natural way** to classify them.
- Organisms are similar because of their **common ancestry**.
- When the natures of objects are defined by a **common history** then there is a **natural way** to classify them.
- For most objects, **their natures are largely independent of their histories**;
- But organisms are products of their **genetic history**.



# Natural system of classification

## History of descent

---

- In 1946, the great microbiologist C.B. van Niel published a thoughtful essay on 'The classification and natural relationships of bacteria' in which he reviewed the history of earlier works.
- He emphasized that even if we knew the phylogenetic relations among bacteria, a classification based on such relations would not necessarily be the best or most efficient for determinative purposes.

# The first natural system of classification

## Five kingdoms

Proposed by R. Whittaker, 1969



Robert H. Whittaker  
(1920-1980)

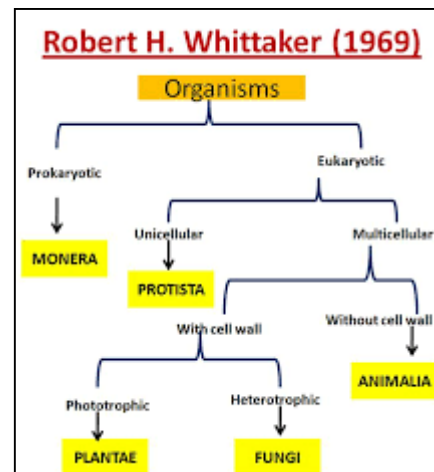
- By 1969, Robert Whittaker proposed that fungi, which were formerly classified as plants.
- This five-kingdom system, 1969 has become a popular standard and with some refinement is still used in many works and forms the basis for new multi-kingdom systems.
- R. Whittaker classified organisms based on:
  1. Cell type
  2. Level of organization
  3. Mode of nutrition

# Natural system of classification

## Five-kingdoms

1. **Plantae:** Plants
2. **Anamalia:** Animals
3. **Fungi:** Molds and yeasts
4. **Protista:** Protozoans, algae, none of the above
5. **Monera:** (**Prokaryotae**) prokaryotes; eubacteria, eocytes?

Cyanobacteria are one of the phyla of the Kingdom **Protista**.







# Natural system of classification

## Demerits of Five Kingdom approach

---

- The Five Kingdom approach is attractive in its simplicity, but has significant problems:
  1. One of these concerns the protists - a wide range of disparate organisms such as amoebae, slime moulds, ciliates, algae, etc. that are grouped together as a kingdom with little justification.
  2. Another problem stems from the recognition in the 1980s that some bacterium-like organisms (first given the name archaebacteria, and now called archaea) are so different from the true bacteria that they can be separated as a group.
- They are prokaryotes, and they look like bacteria, but in terms of cellular biochemistry and genetics the archaea differ from both eukaryotes and bacteria.

# The second natural classification scheme

## Six kingdoms

**Proposed by Woese *et al.*, 1977**

---

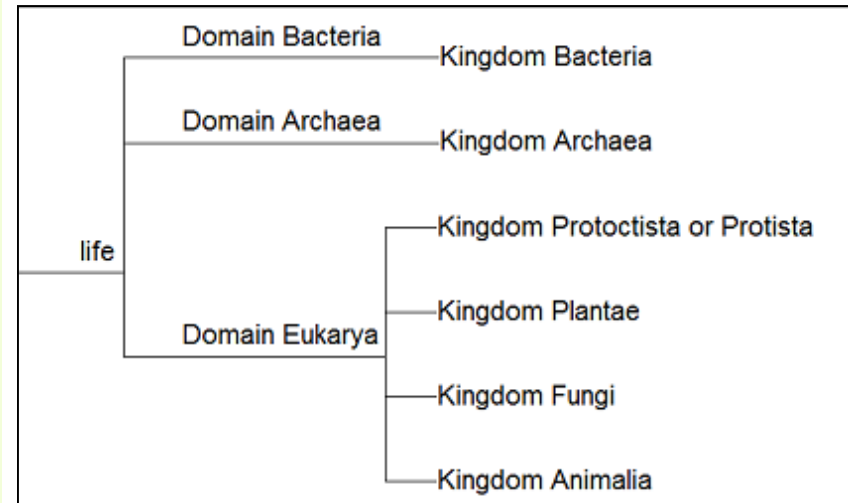
- From 1971 to 1985, Carl Woese and colleagues generated oligonucleotide catalogs of 16S/18S rRNAs from more than 400 organisms.
- Carl Woese and colleagues, studying ribosomal RNA RNA gene sequences, suggest that procaryotes divided into two distinct lineages early in the earth's evolution.
- Six-kingdom system - differs from five-kingdom system by dividing procaryotes into bacteria and archaea.

# The second natural classification scheme

## Six kingdoms

**Proposed by Woese *et al.*, 1977**

1. Kingdom **Eubacteria**
2. Kingdom **Archaebacteria**
3. Kingdom **Protocista**
4. Kingdom **Plantae**
5. Kingdom **Fungi**
6. Kingdom **Animalia**



Based on this work, they concluded that the **Archaea** are more closely related to humans than to bacteria.

**Kingdom Animalia or animals**

**Examples:**

**Arthropoda** – includes insects, arachnids, and crustaceans

**Chordata** – includes vertebrates and, as such, **human beings**.

# The Third natural classification scheme

## Six kingdoms

Proposed by T. Cavalier-Smith, 1998

---

- In 1981, Cavalier-Smith's proposed the division of all organisms into **eight kingdoms**.
- By 1998, Cavalier-Smith had reduced the **total number of kingdoms from eight to six**:
  - **Animalia, Protozoa, Fungi, Plantae** (including red and green algae), **Chromista and Bacteria**.
- In 2015, Cavalier-Smith and his collaborators once again revised the classification (Ruggiero *et al.*, 2015).
- In this scheme they reintroduced the **division of prokaryotes into two kingdoms**:
  1. **Bacteria (=Eubacteria), and**
  2. **Archaea (=Archeobacteria).**

# The Third natural classification scheme

## Six kingdoms

Proposed by T. Cavalier-Smith, 1998

---

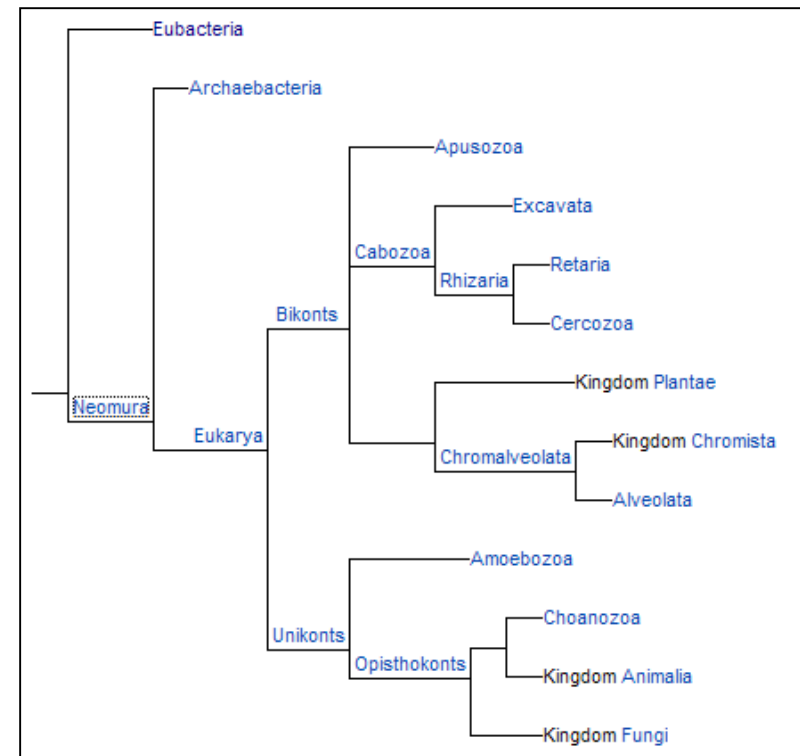
- Thomas Cavalier-Smith, 1998 has published a six-kingdom model on the evolution and classification of life, particularly protists.
  1. Animalia
  2. Protozoa
  3. Fungi
  4. Plantae (including red and green algae),
  5. Chromista
  6. Bacteria
- This was revised in subsequent papers.
- In total, his views have been influential but controversial, and not always widely accepted.

# The Third natural classification scheme

## Six kingdoms

A revised six-kingdom system proposed by T. Cavalier-Smith, 1998

- Cavalier-Smith does not accept the importance of the fundamental eubacteria-archaebacteria divide put forward by Woese and others and supported by recent research.
- His Kingdom Bacteria includes the Archaeobacteria as part of a subkingdom along with a group of eubacteria (Posibacteria).
- Nor does he accept the requirement for groups to be monophyletic.



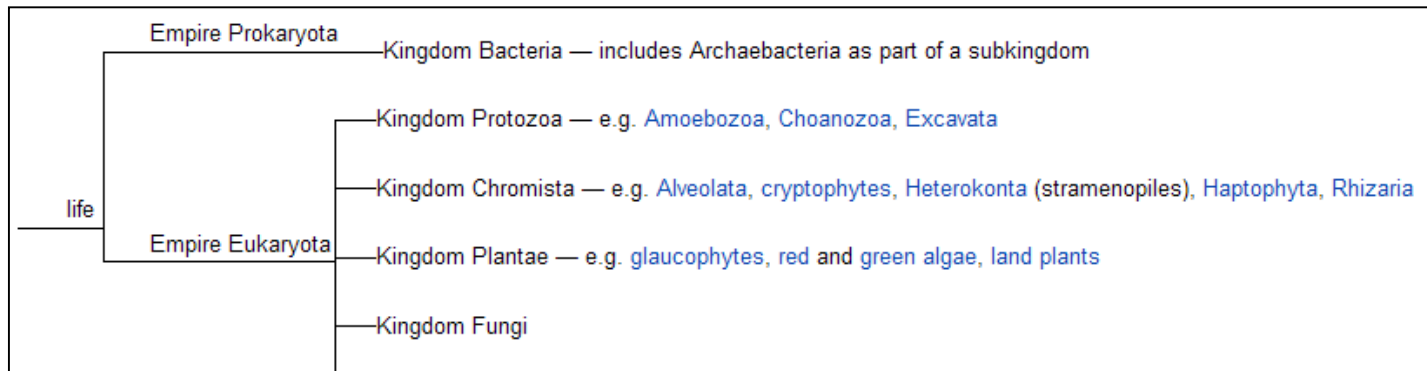
By September 2003, Cavalier-Smith's tree of life looked like above.

# The Third natural classification scheme

## Six kingdoms

Proposed by T. Cavalier-Smith, 2004 & 2009

- The version published in 2009 is shown below.
- Compared to the version he published in 2004 the alveolates and the rhizarians have been moved from Kingdom Protozoa to Kingdom Chromista.
- His Kingdom Protozoa includes the ancestors of Animalia and Fungi.
- Thus the diagram below does not represent an evolutionary tree.



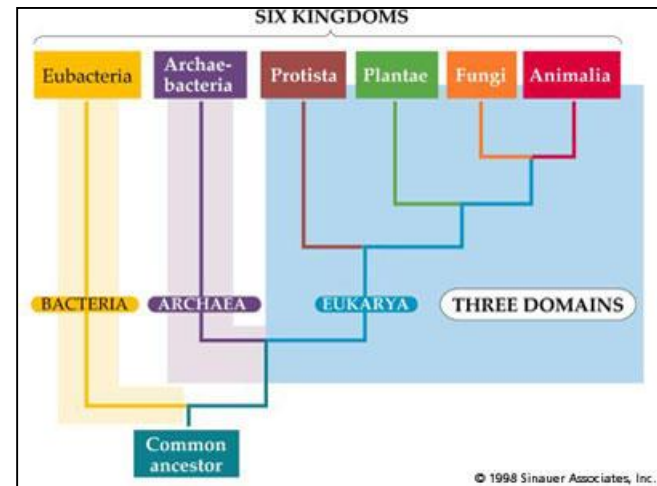
# The Six Kingdoms

The six kingdoms of living things are divided into two major groups, Prokaryotes and Eukaryotes

## Cavalier-Smith megaclassification of prokaryotes(life):

- Currently, textbooks from the United States use a **system of six kingdoms**. They classify organisms into **three domains** and into **six Kingdoms of life**.
- The kingdoms are further divided into **two prokaryote kingdoms** and **four eukaryote kingdoms**:

1. Plants
2. Animals
3. Archaeobacteria
4. Eubacteria
5. Fungi
6. Protists





# Summary of the sequence from the two-kingdom system up to Cavalier-Smith's six-kingdom system

Linnaeus 1735	Haeckel 1866	Chatton 1925	Copeland 1938	Whittaker 1969	Woese et al. 1977	Woese et al. 1990	Cavalier-Smith 2004
2 kingdoms	3 kingdoms	2 empires	4 kingdoms	5 kingdoms	6 kingdoms	3 domains	6 kingdom
(not treated)	Protista	Prokaryota	Mychota	Monera	Eubacteria	Bacteria	Bacteria
Vegetabilia	Plantae	Euokaryota	Protoctista	Protista	Archaeobacteria	Archaea	Protozoa
Animalia	Animalia		Plantae	Fungi	Protista	Eukarya	Chromista
			Animalia	Plantae	Fungi		Fungi
				Animalia	Plantae		Plantae
					Animalia		Animalia

# Summary of the sequence from the two-kingdom system up to Cavalier-Smith's six-kingdom system

Linnaeus 1735	Haeckel 1866	Chatoon 1925	Copeland 1938	Whittaker 1969	Woese <i>et al.</i> 1977	Woese <i>et al.</i> 1990	Cavalier-Smith 1993	Cavalier-Smith 1998
2 kingdoms	3 kingdoms	2 empires	4 kingdoms	5 kingdoms	6 kingdoms	3 domains	8 kingdoms	6 kingdoms
(not treated)	Protista	Prokaryota	Monera	Monera	Eubacteria	Bacteria	Eubacteria	Bacteria
					Archaeobacteria	Archaea	Archaeobacteria	
		Eukaryota	Protoctista	Protista	Protista	Eukarya	Archezoa	Protozoa
							Protozoa	
							Chromista	Chromista
							Plantae	Plantae
Vegetabilia	Plantae		Plantae	Plantae	Plantae		Fungi	Fungi
Animalia	Animalia		Animalia	Animalia	Animalia		Animalia	Animalia



# Woesian tree of life, 1977

---

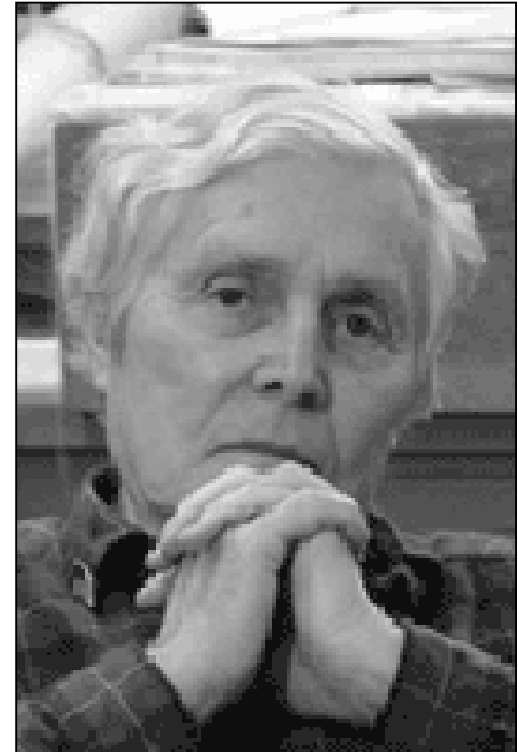
## **Domain Concept**

**Using ribosomal RNA sequence as an evolutionary measure**

# Carl Richard Woese

The famous American microbiologist and physicist  
**Discovered Life's 'Third Domain'**

- Carl Richard Woese (pronounced woes) born 15 July, 1928, died aged 84. December 30, 2012.
- Woese is famous for defining the Archaea (a new domain or kingdom of life) in 1977 by phylogenetic taxonomy of 16S ribosomal RNA, a technique pioneered by Woese and which is now standard practice.
- He was also the originator of the RNA world hypothesis in 1977, although not by that name.

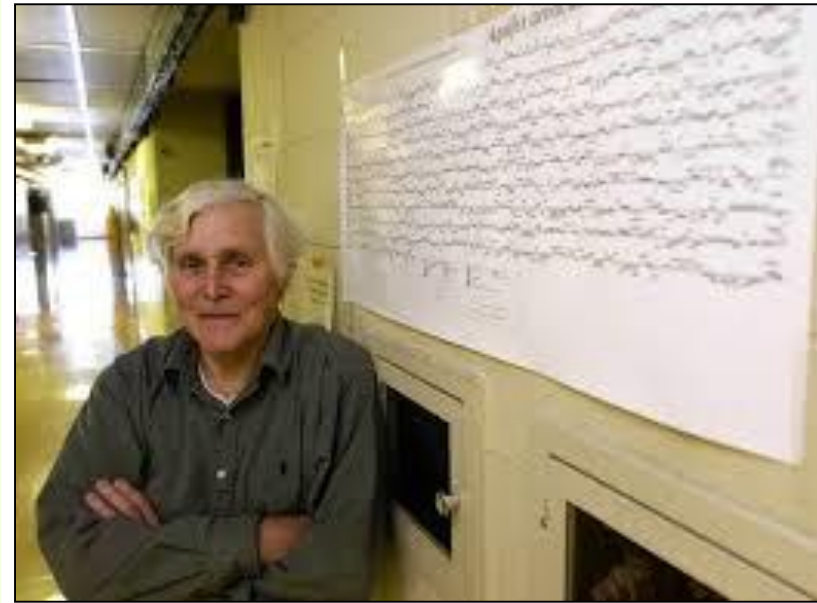


B.A. (Math and Physics), Amherst College, 1950  
Ph.D. (Biophysics), Yale University, 1953  
Postdoctoral (Biophysics), Yale University, 1953-1960  
Biophysicist, General Electric Research Laboratory,  
1960-1963.

# Carl Richard Woese

The famous American microbiologist and physicist  
**Discovered Life's Third Domain(Archaea)**

- He revolutionized the world of evolutionary biology when he announced his discovery of a life form so different from other organisms that it amounted to an entirely new category.
- Dr. Woese received many honors and awards, including:
  1. A MacArthur Foundation "Genius" grant in 1984,
  2. The National Medal of Science in 2000, and
  3. The Crafoord Prize in Biosciences from the Royal Swedish Academy of Sciences in 2003.



B.A. (Math and Physics), Amherst College, 1950  
Ph.D. (Biophysics), Yale University, 1953  
Postdoctoral (Biophysics), Yale University, 1953-1960  
Biophysicist, General Electric Research Laboratory, 1960-1963.



# Archaeobacteria (Archaea)

## The third domain

---

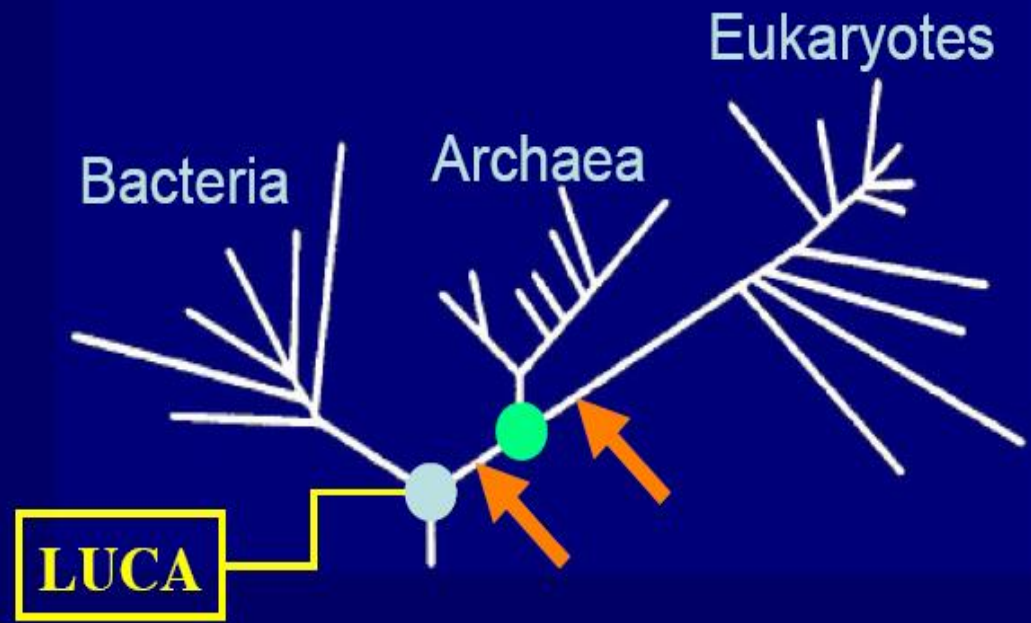
- Microscopic characteristics have classified the living world into the **two primary domains** of:
  1. Eukaryotes (Eukarya), and
  2. Prokaryotes (Bacteria).
- **Woese and coworkers** proposed a **third domain of life** based on the studies of a heretofore poorly known group of prokaryotes, the
  3. Archaeobacteria (Archaea).
- From the identification of signature sequences on the **16S ribosomal RNA**, which are distinctive in **eukaryotes, prokaryotes and archaeobacteria**, the **third domain Archaea** was proposed (1977 and 1978).

# Woesian tree of life

## The first phylogenetic tree

### Woe Is the Tree of Life

Carl Woese -  
first phylogenetic  
tree of life



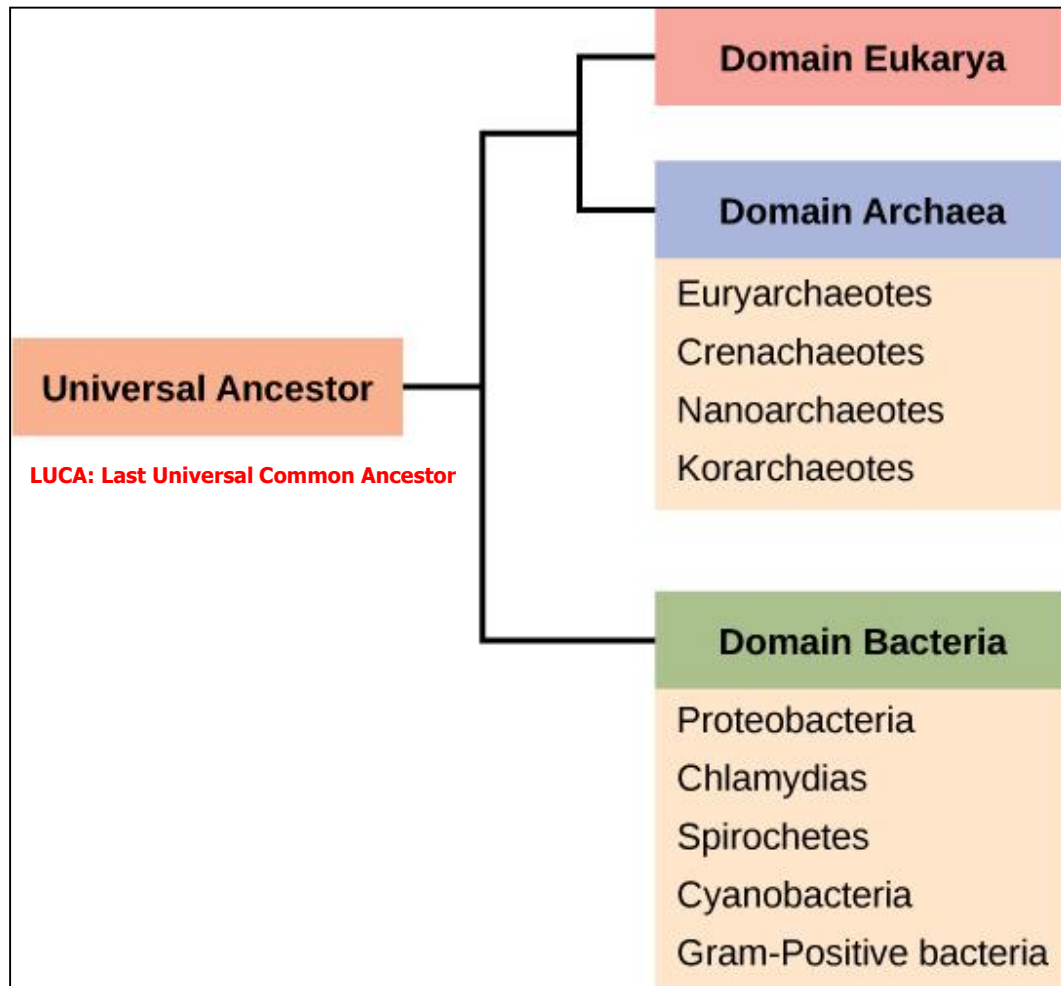
**LUCA: Last Universal Common Ancestor**

### Conclusions:

1. LUCA was bacterial-like (A prokaryote)
2. Eukaryotes evolved from Archaea

# Evolutionary relationships among the three domains

**Based on their ribosomal RNA differences**







# Last Universal Common Ancestor

## RNA & LUCA

---

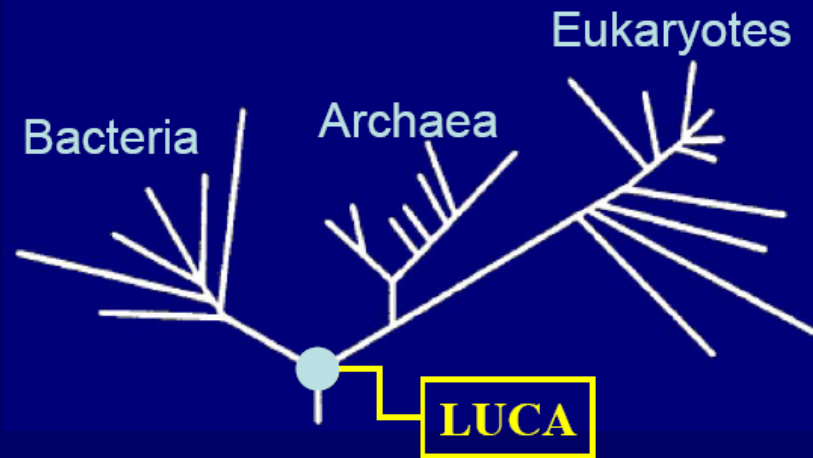
- There are various hypotheses as to the origin of prokaryotic and eukaryotic cells.
- Because all cells are similar in nature, it is generally thought that all cells came from a common ancestor cell termed the last universal common ancestor (LUCA).
- LUCA eventually evolved into three different cell types, each representing a domain.
- The three domains are the *Archaea*, the *Bacteria*, and the *Eukarya*.

# Last Universal Common Ancestor

## RNA & LUCA

### RNA & LUCA

- The RNA world precedes LUCA.
- It is possible that some modern RNAs have their origins in the RNA world.
- If we can determine which RNAs are likely to date from this early period, we can build up a picture of the RNA world
- Anything we can establish about the RNA world from 'relics' also tells us about LUCA.



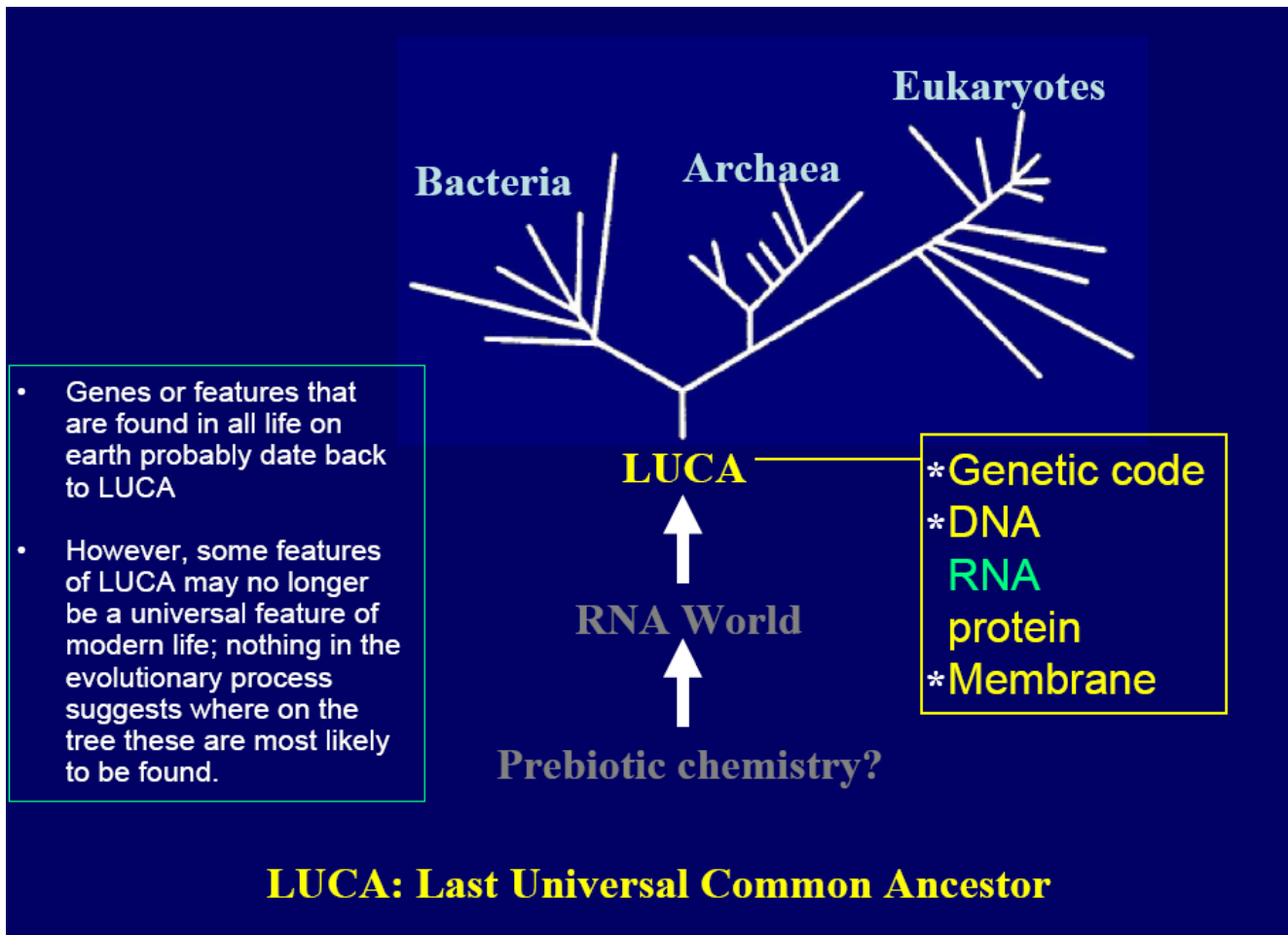
**RNA World**

**Prebiotic chemistry**

**Relic:** "remains, leave behind, or abandon."

# Last Universal Common Ancestor

## LUCA





# Evolution from a common ancestor

## Biological features of the LUCA

---

LUCA was probably RNA-rich

Majority of RNA world relics appear to be preserved in eukaryotes

Available evidence suggests LUCA was not a thermophile

The prokaryote lineages appear streamlined, and this likely reflects lifestyle



# Three-Domain Classification

Universal tree of life, based on 16S rRNA sequences

---

- Woese recognized the full potential of rRNA sequences as a measure of phylogenetic relatedness.
- He initially used an RNA sequencing method that determined about 1/4 of the nucleotides in the 16S rRNA (the best technology available at the time).
- He reasoned that all organisms had to have 16S rRNA, and since it was used by all organisms to make all proteins, the sequence would be highly conserved.
- Over the next decade he soon developed a huge library of 16S rDNA sequences, which could be compared with one another to produce what has since been called the universal tree of life.



# rRNA trees

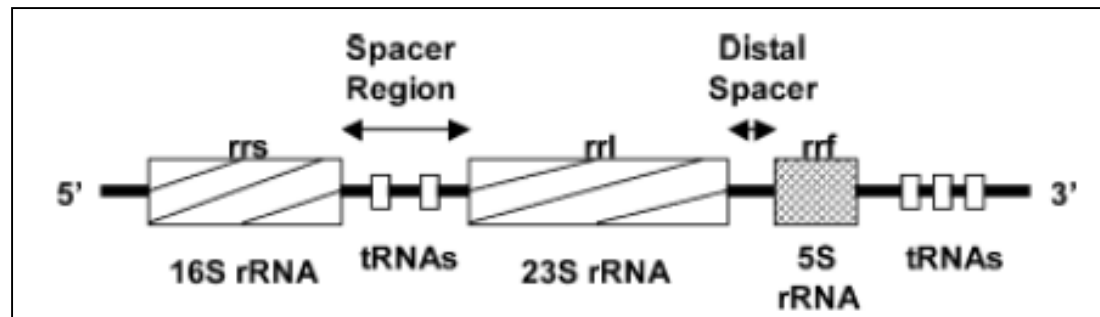
## Tree of Life

---

- Trees of small subunit ribosomal RNA (rRNA trees), which are sometimes called the tree of life (sometimes even called the Tree of Life, capitalized as if it warrants religious reverence(emotion)).

# Ribosomal RNA operon(rrn)

- The **rrn locus** consisted of a **16S rRNA gene (rrs)**, followed by an **intergenic transcribed spacer (ITS)** containing two genes of tRNA<sup>Ile</sup> and tRNA<sup>Ala</sup>, a **23S rRNA gene (rrl)**, an ITS devoid of tRNA genes and a **5S rRNA gene (rrf)**.
- The internally transcribed spacer region (ITS) between the **16S and 23S rRNA genes** appears to be **more variable than 16S and 23S rRNA genes**.



Schematic diagram of a typical ribosomal RNA operon.

# Ribosomal RNA genes and their sequences

## Ribosomal RNAs in Prokaryotes

- The name is based on the rate that the molecule sediments (sinks) in water.
- Bigger molecules sediment faster than small ones.
  1. The 5S rRNA is too small, contains limited info.
  2. 23S rRNA is too large, too difficult to manage
  3. 16S rRNA has the right size for studies.

Name	Size (nucleotides)	Location
16S	1500	Small subunit of ribosome
5S	120	Large subunit of ribosome
23S	2900	Large subunit of ribosome





# rRNAs

## Molecular chronometers

---

- rRNA has revolutionised bacteriology by providing sequences that are unique to species, genera, etc.
- Signature sequences allow unequivocal assignment of an unknown organism to a clade irrespective of other genes or properties which could have derived from gene transfer.
- Ribosomal evolution is very slow.
- Ribosomal genes are proven to be highly correlated to phylogeny - taxonomic evolution.



# rRNAs

## Molecular chronometers

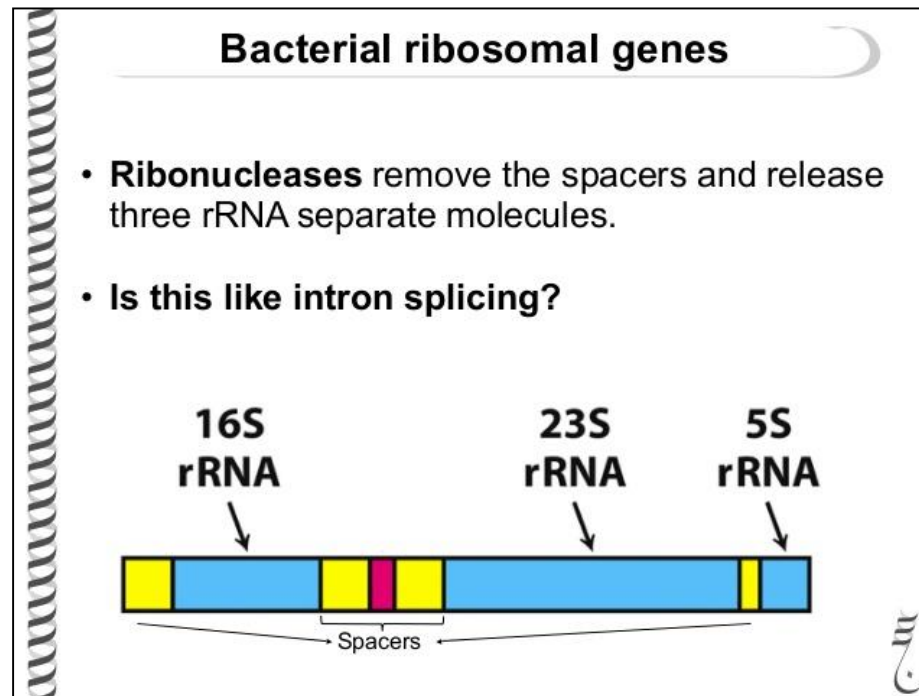
---

- Ribosomal genes produce the ribosomes consisting of subunits made up of proteins and rRNA (coded by rDNA).
- However inferences about other genetic properties based on the inter-relatedness based on rRNA are still problematic.

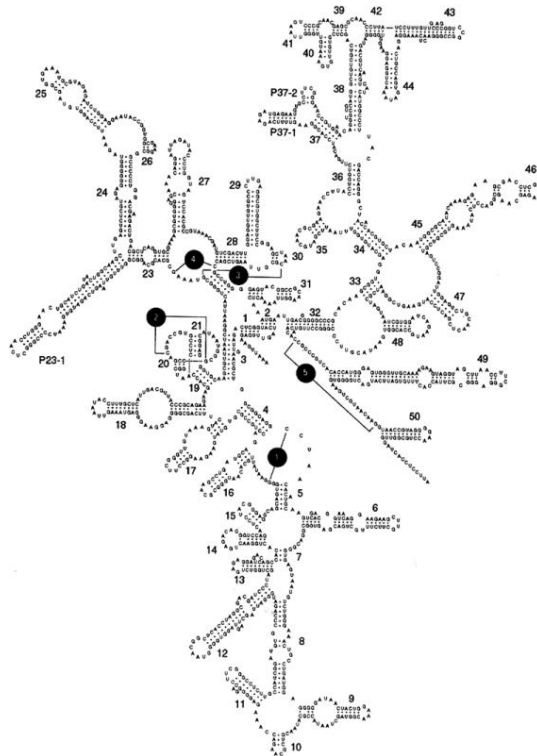
# rRNA 16S and 23S genes

## Two Molecular chronometers

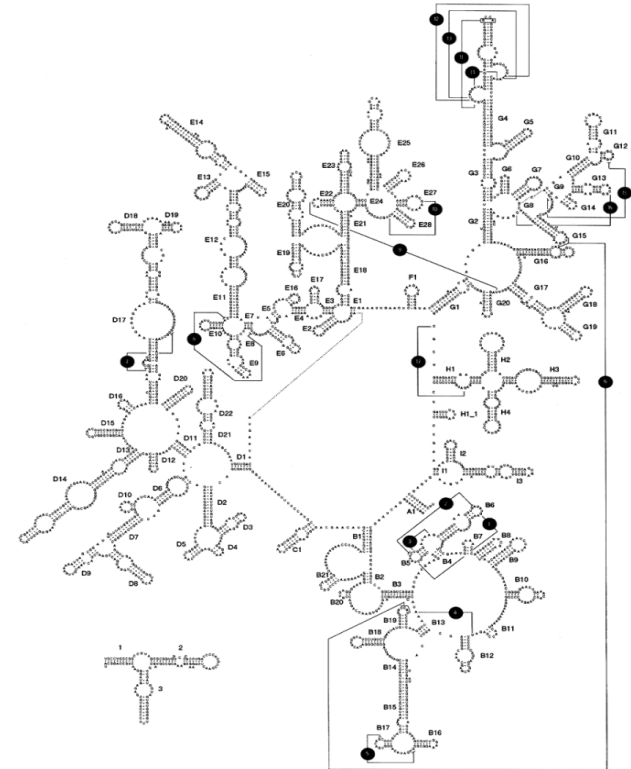
- The **rRNA 16S** and **23S genes** are the most widely used **molecular chronometers** for inferring microbial phylogeny and have been instrumental in developing a comprehensive view of **microbial phylogeny and systematics**.



# Structure of 5 ,16 & 23S rRNA molecules



**16S rRNA molecular structure**

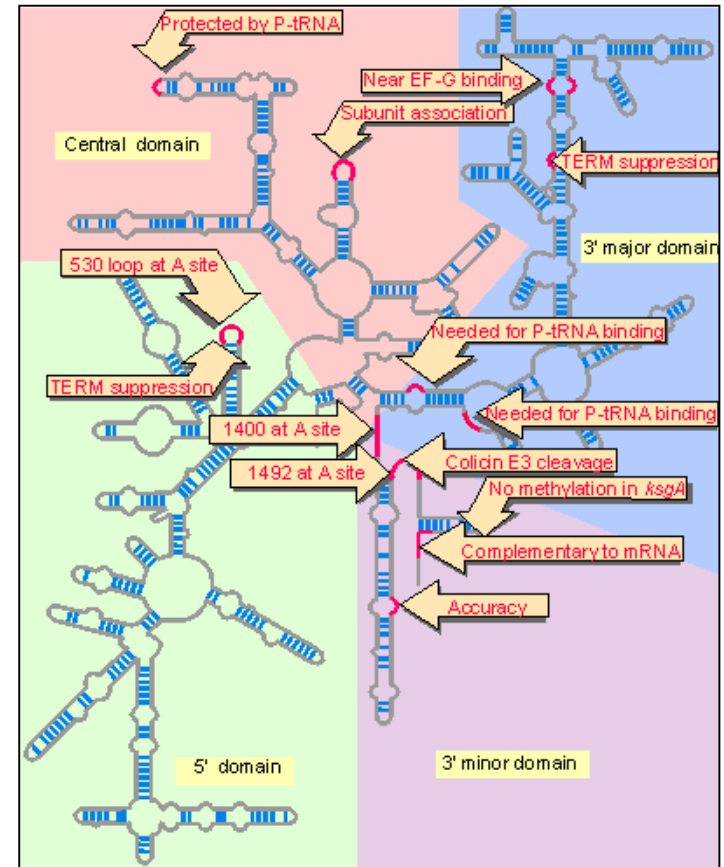


**Structure of 5 & 23S rRNA molecules**

# Structure of 16S rRNA

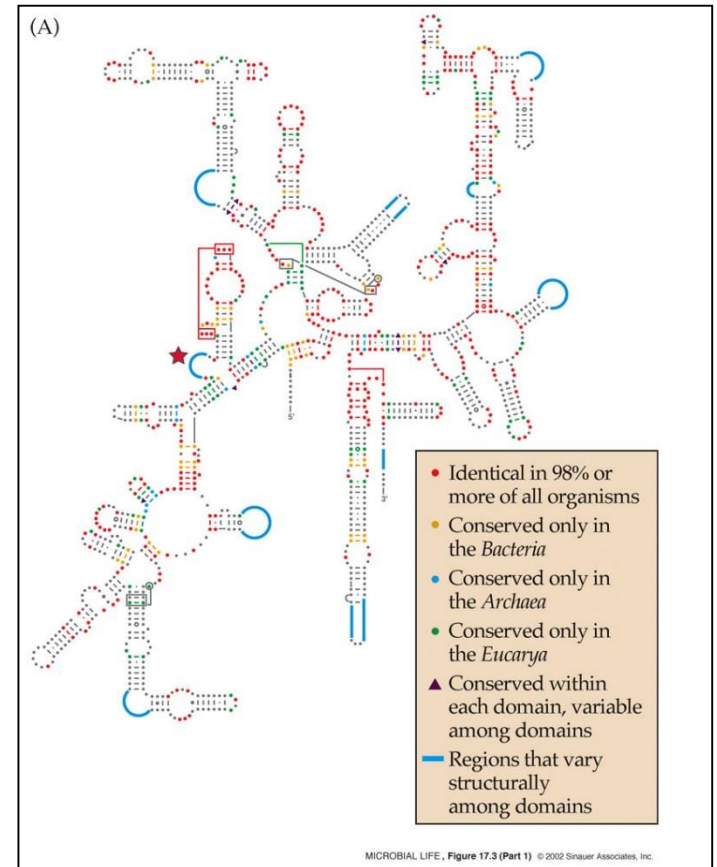
The large colored blocks indicate the four domains of the rRNA

- Some sites in 16S rRNA are protected from chemical probes when 50S subunits join 30S subunits or when aminoacyl-tRNA binds to the **A site**.
- Others are the sites of mutations that affect protein synthesis.
- **TERM suppression sites** may affect termination at some or several termination codons.



# Structure of 16S rRNA

- 16s rRNA is present in the small subunit of **prokaryotic** ribosomes as well as **mitochondrial ribosomes** in **eukaryotes**.





# 16S ribosomal RNA

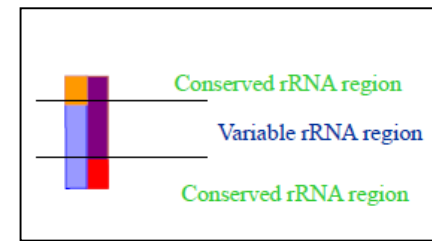
## Gold standard

---

- Analysis of 16S ribosomal RNA (rRNA) sequences has been the de-facto gold standard for the assessment of phylogenetic relationships among prokaryotes.
- Although phylogenetic information content of the 23S rRNA molecule is greater than that of the 16S rRNA molecule, the number of currently available complete 23S rRNA sequences is rather poor in comparison to those of the 16S rRNA.
- Therefore, 16S rRNA approach remains the "gold standard" for elucidating bacterial phylogeny.

# 16S ribosomal DNA

**A set of 16S rDNA PCR primers for exploring bacterial diversity**

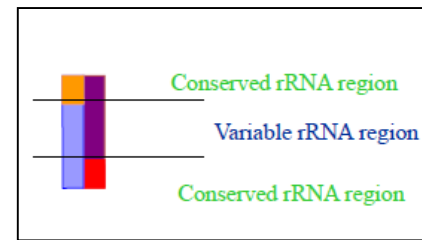


- Most of these new methods are based on **sequences of the 16S rRNA gene**, a gene encoding a molecule of RNA used in **bacterial and archaeal ribosomes**.
- The 16S rRNA gene is approximately **1500 bases in length** and contains regions that are:
  1. **Highly 'conserved'** (i.e., have the **same sequence in all bacteria and archaea**), and
  2. **Highly 'variable'** (i.e., have sequences that are **unique at the genus or species level**).
- Thus the **conserved regions** of the gene can be used to bind **primers for PCR and sequencing**, and the **variable regions** to determine the identity of the organism.

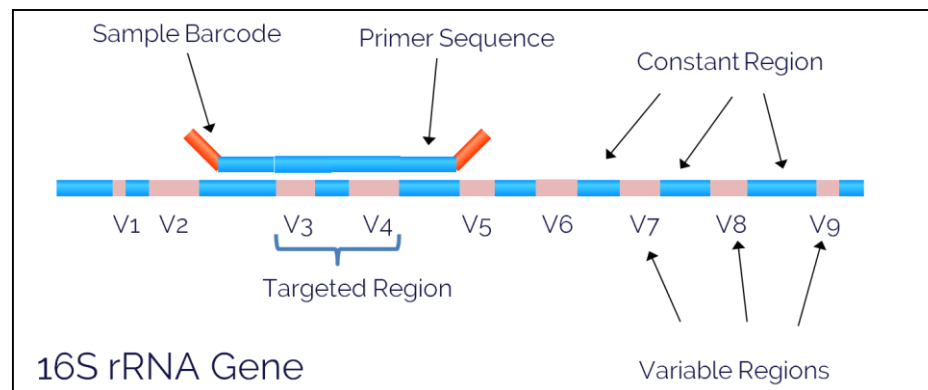


# 16S ribosomal DNA

**A set of 16S rDNA PCR primers for exploring bacterial diversity**



- Conveniently, the 16S rRNA gene consists of both **conserved and variable regions**.
- While the **conserved region** makes universal amplification possible,
- sequencing the variable regions allows discrimination between specific different microorganisms such as **bacteria, archaea and microbial eukarya**.



# 16S/18S Ribosomal RNA

## A visual comparison

- A second group(**eukaryotes**) is defined by the **18S rRNAs** of the **eukaryotic cytoplasm-animal, plant, fungal, and slime mold**(unpublished data)(woese and Fox,1997).
- The extraordinary conservation of rRNA genes can be seen in these fragments of the small subunit rRNA gene sequences from organisms spanning the known diversity of life:

```
human...GTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGCTGCAGTTAAAAAG...
yeast...GTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCAGTTAAAAAG...
corn...GTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCAGTTAAAAAG...
Escherichia coli...GTGCCAGCAGCCGCGGTAATACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCG...
Anacystis nidulans...GTGCCAGCAGCCGCGGTAATACGGGAGAGGCAAGCGTTATCCGGAATTATTGGGCGTAAAGCG...
Thermotoga maritima...GTGCCAGCAGCCGCGGTAATACGTAGGGGGCAAGCGTTACCCGATTACTGGGCGTAAAGGG...
Methanococcus vannieli...GTGCCAGCAGCCGCGGTAATACCGACGGCCCGAGTGGTAGCCACTCTTATTGGGCCTAAAGCG...
Thermococcus celer...GTGGCAGCCGCCGCGGTAATACCGGCGGCCCGAGTGGTGGCCGCTATTATTGGGCCTAAAGCG...
Sulfolobus sulfotaricus...GTGTCAGCCGCCGCGGTAATACCAGCTCCGCGAGTGGTGGGGTGATTACTGGGCCTAAAGCG...
```



# Three-Domain Classification

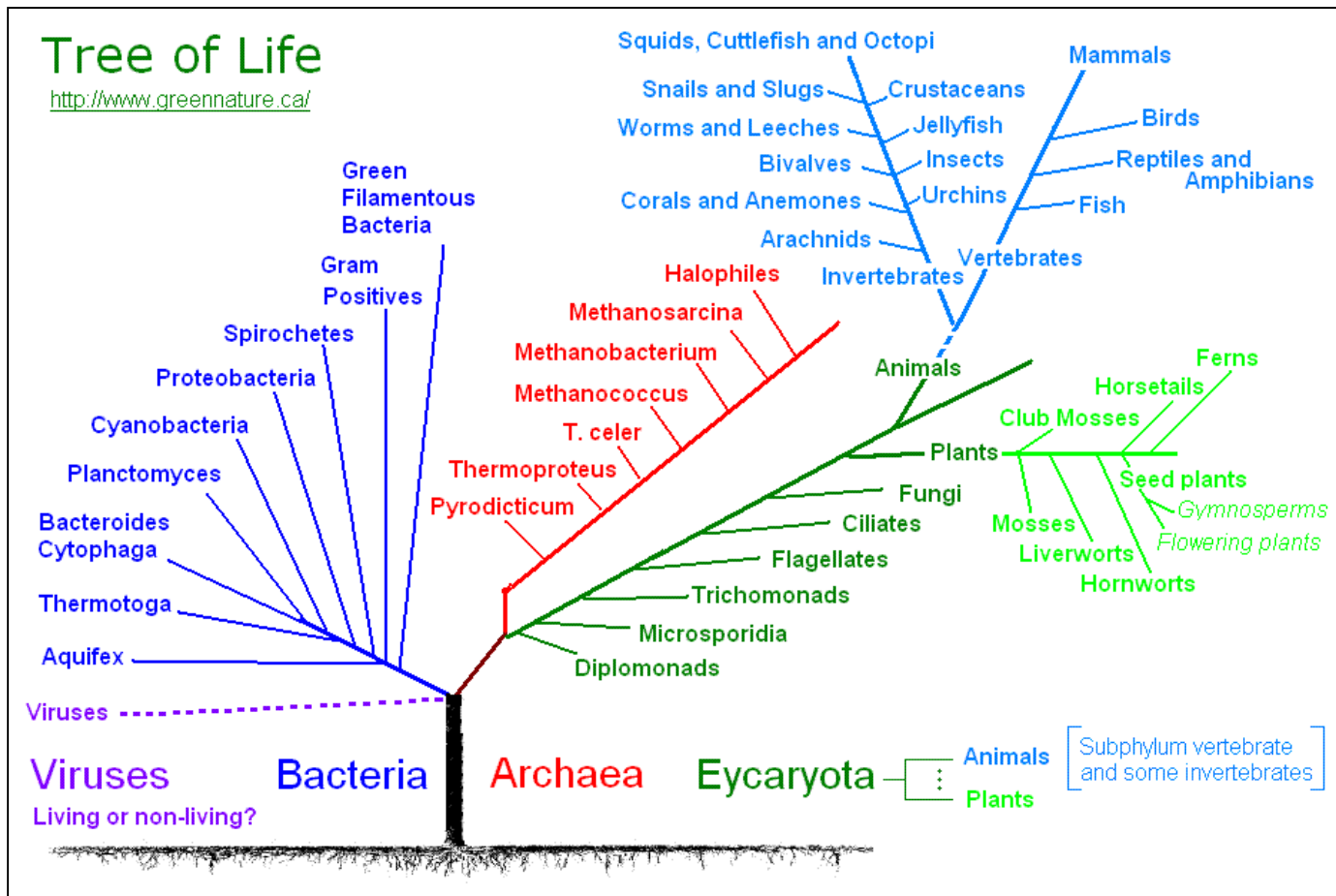
**Universal tree of life, based on 16S rRNA sequences**

---

- C. Woese had done gene sequencing to estimate phylogenetic or evolutionary relationship.
- Genes employed are rRNA.
- With his data he constructed universal tree of life or Woesian tree of life.
- According to him:
  1. Archaea are ancient most bacteria, and
  2. Eubacteria are present day or evolved bacteria.

# Three-Domain Classification

Universal tree of life, based on 16S rRNA sequences

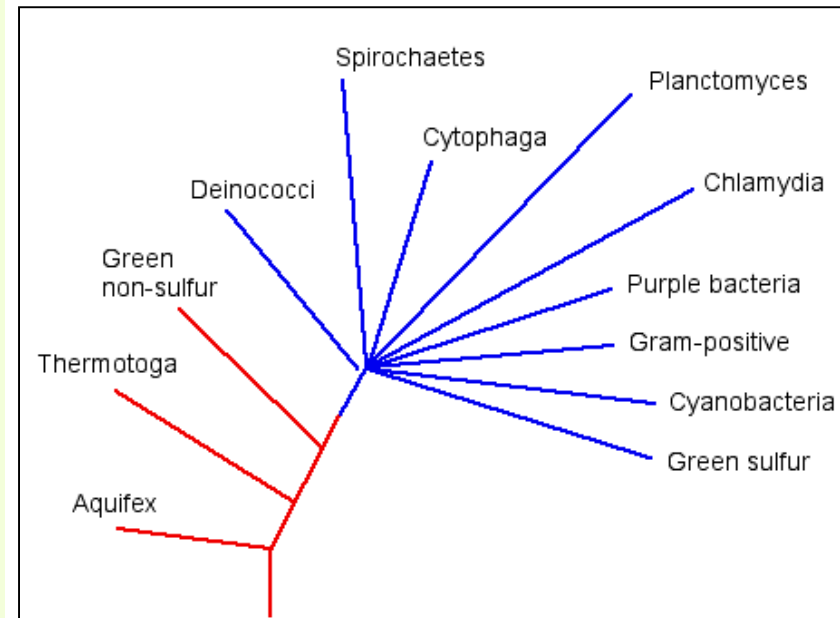


# Three-Domain Classification

## 1. Domain Bacteria

consist of approximately 12 distinct groups

- Most of these groups appear to have radiated from the same point.
- These are called the "main radiation" groups.
- A few branches are deeper and earlier, and appear to represent more primitive bacterial groups.



Purple bacteria or purple photosynthetic bacteria are proteobacteria that are phototrophic, that is, capable of producing their own food via photosynthesis.

# Domain Bacteria

## Proteobacteria

### Five classes based upon rRNA data

---

- The 5 major classes of proteobacteria:
  1. **Alphaproteobacteria**: Oligotrophic forms including the purple nonsulfur photosynthesizers.
  2. **Betaproteobacteria**: Metabolically similar to alphaproteobacteria.
  3. **Gammaproteobacteria**: Diverse methods of energy metabolism.
  4. **Deltaproteobacteria**: Includes predators and the fruiting myxobacteria.
  5. **Epsilonproteobacteria**: Contains some human pathogens (*Helicobacter* spp. in the stomach, *Campylobacter* spp. in the duodenum).

# Domain Bacteria

## Proteobacteria

### Five classes based upon rRNA data

---

- The **Proteobacteria** account for **more than 40%** of all **validly published prokaryotic genera** and encompass a major proportion of the **traditional Gram-negative bacteria**.
- All cultivable Gram-negative plant pathogenic prokaryotes occur within the **alpha, beta and gamma subdivisions** of the **phylum Proteobacteria** based on DNA sequencing.
- **All species** contain:
  1. **Peptidoglycan**, and
  2. **an outer membrane containing lipopolysaccharide**.

# Domain Bacteria

## Proteobacteria

### Five classes based upon rRNA data

---

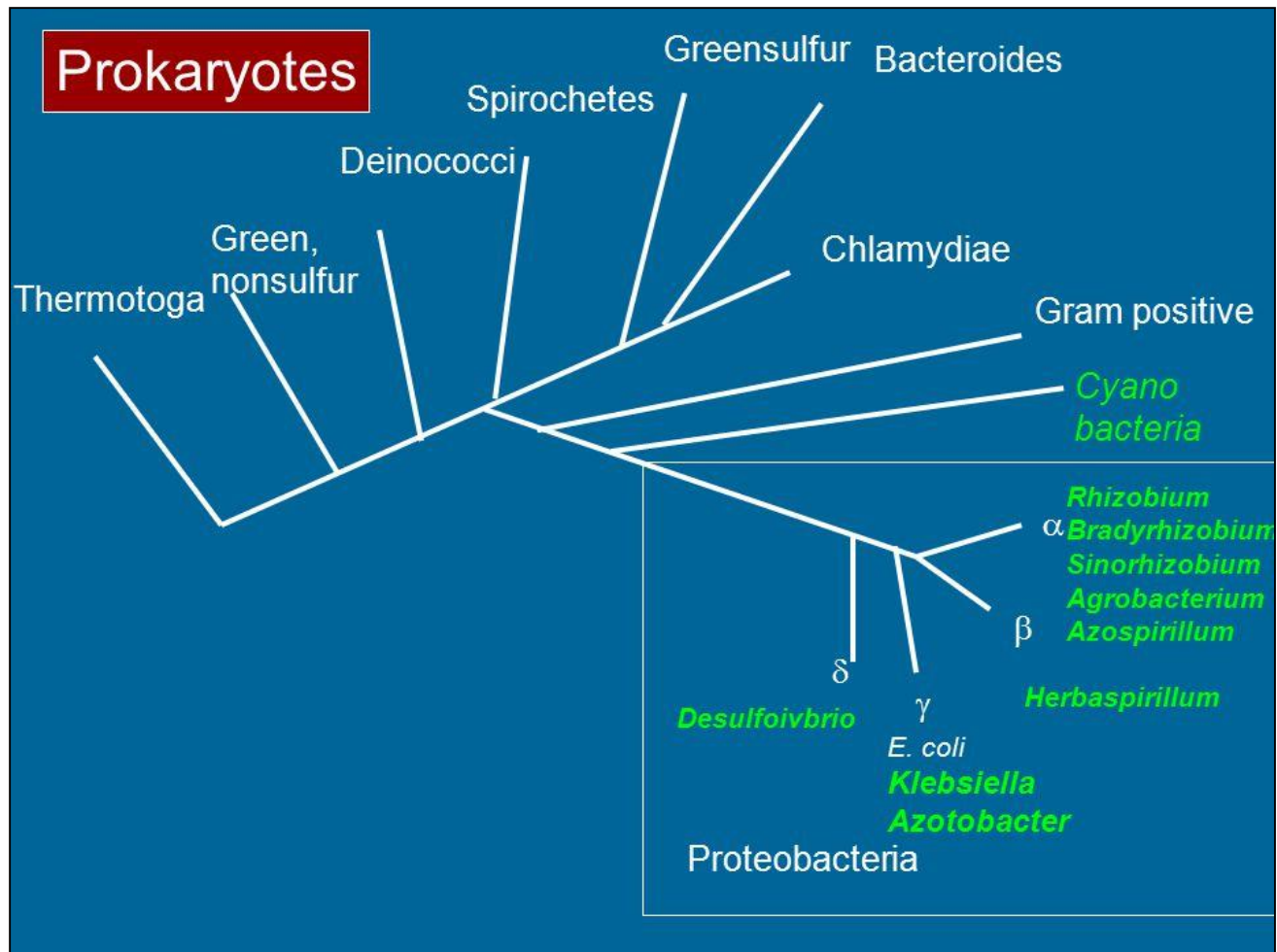
- Within the domain Bacteria, the phylum Proteobacteria constitutes at present the largest and phenotypically most diverse phylogenetic lineage.
- In 2002, the Proteobacteria consist of more than 460 genera and more than 1600 species, scattered over 5 major phylogenetic lines of descent known as the classes:
  1. Alphaproteobacteria
  2. Betaproteobacteria
  3. Gammaproteobacteria
  4. Deltaproteobacteria
  5. Epsilonproteobacteria



# Domain Bacteria

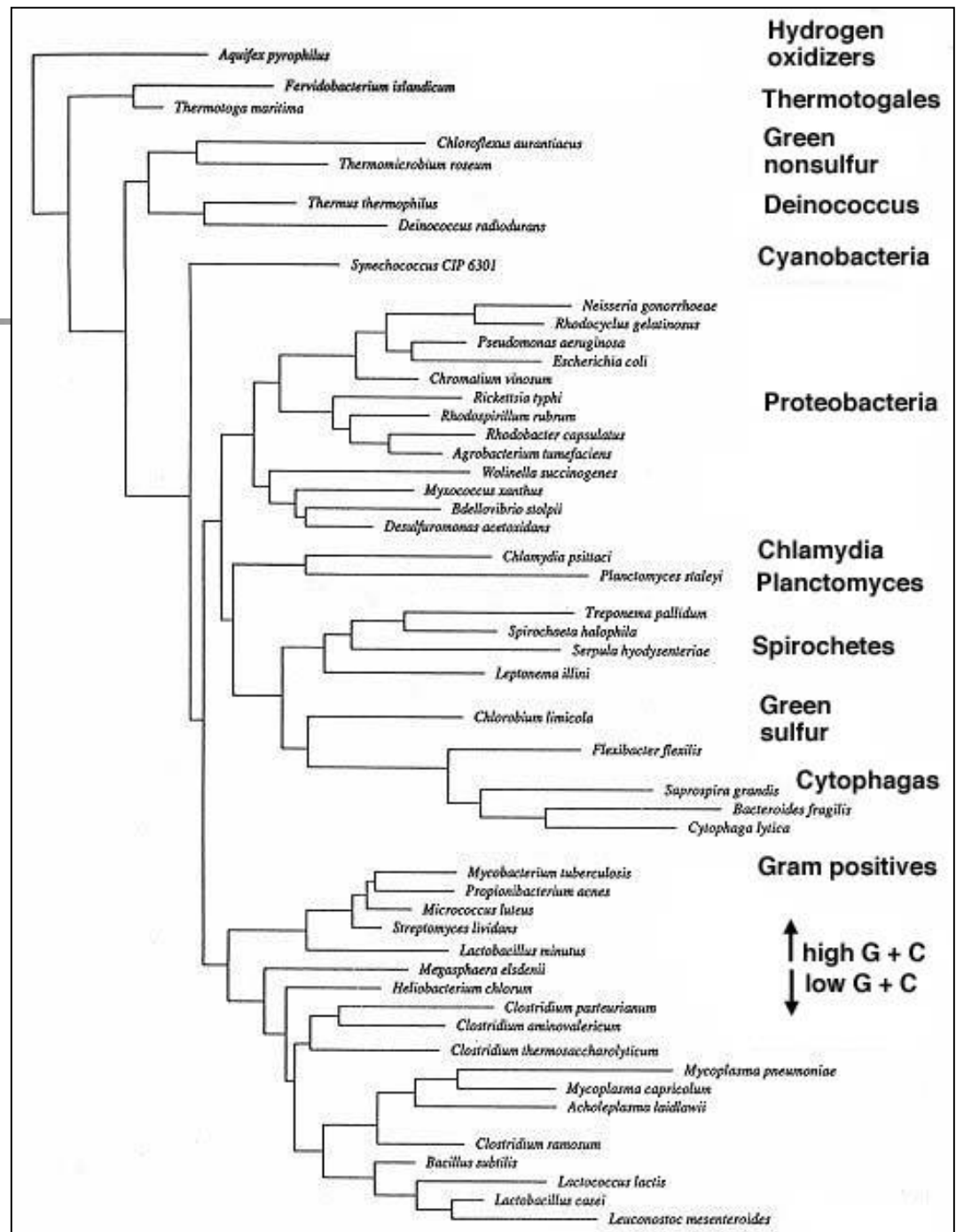
## Proteobacteria

Five classes based upon rRNA data



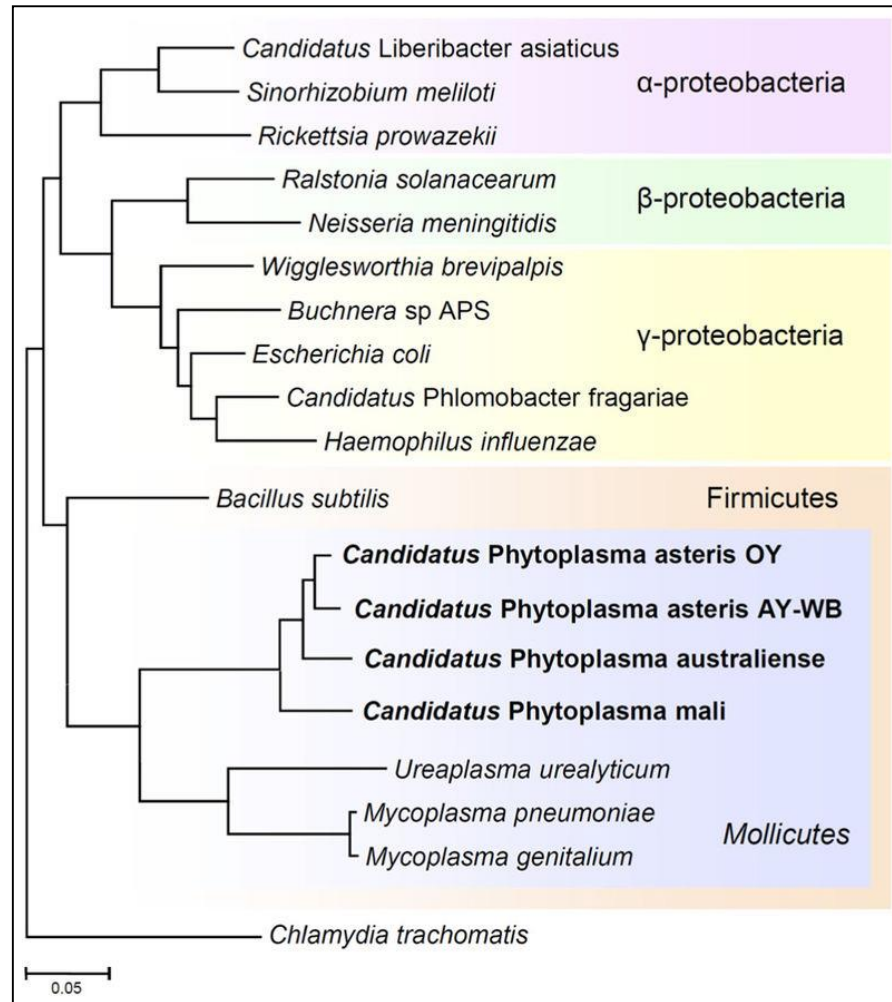
# Domain Bacteria

## Major Groups within the Bacteria.



# Proteobacteria

## Phylogenetic position of Mollicutes among bacteria, using 16S rRNA sequences





# LATEST TREE OF LIFE

## Based on comparative genomics

40-80  
(bootstrap support)

0.1  
(substitutions / site)

**γ-proteobacteria**

**Gram -**

**β-proteobacteria**

**α-proteobacteria**

**ε-proteobacteria**

δ-proteobacteria  
Acidobacteria  
Cyanobacteria  
Deinococcales  
Chloroflexi  
Aquificae  
Thermotoga  
Fusobacteria  
Chloroflexi

**ARCHAEBACTERIA**

**EUKARYOTES**

Nanoarchaeota  
Crenarchaeota  
Kinetoplastida  
Chromalveolata  
Plantae  
Amoebozoa  
Fungi

Metazoa

**PLANTS**

**FUNGI**

**NEMATODES**

**MOLLUSCS**

**Gram +**

**BACTERIA**

**Firmicutes**

**Clavibacter**

Planctomycetes  
Spirochaetes  
Actinobacteria  
Fibrobacteres  
Chlorobi  
Bacteroidetes

# Domain Bacteria

## Comparing three systems of Proteobacteria classification

Classification		
1 <sup>a</sup>	2 <sup>b</sup>	3 <sup>c</sup>
Class Proteobacteria	Phylum Proteobacteria	Division Proteobacteria
		Subdivision Rhodobacteria
Subclass alpha	Class "Alphaproteobacteria" <sup>d</sup>	Class Alphabacteria
Subclass beta	Class "Betaproteobacteria"	Class Chromatibacteria
Subclass gamma	Class "Gammaproteobacteria"	
		Subdivision Thiobacteria
Subclass delta	Class "Deltaproteobacteria"	Class Deltabacteria
Subclass epsilon	Class "Epsilonproteobacteria"	Class Epsilobacteria

<sup>a</sup>From Stackebrandt et al. (1988b).  
<sup>b</sup>From *Bergey's Manual of Systematic Bacteriology* (Garrrity, 2001).  
<sup>c</sup>From Cavalier-Smith (2002).  
<sup>d</sup>Quotation marks are used for names that have not yet been validated.

In a recently revised megaclassification of the prokaryotes, Cavalier-Smith (2002) proposes a new classification and nomenclature for the five major subgroups of the Proteobacteria.

# Domain Bacteria

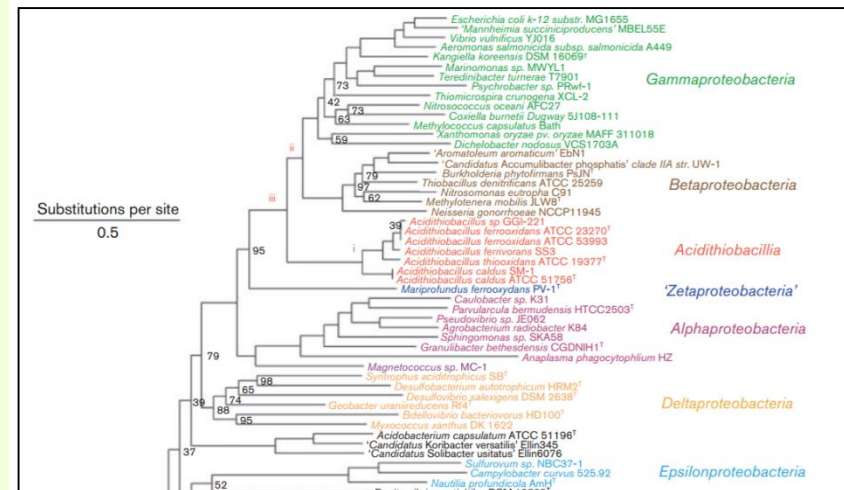
## Comparing three systems of Proteobacteria classification

- The phylum Proteobacteria has its taxonomic origin as the 'purple bacteria', defined as four bacterial groups (**alpha, beta, gamma and delta**), which were classified by their 16S rRNA gene sequence structures (Woese, 1987).
- The phylum was formally established, also using phylogenetic analysis of 16S rRNA gene sequences, by Garrity *et al.*, 2005a, with **five constituent classes** containing all known Gram-negative bacteria:
- **Alphaproteobacteria, Betaproteobacteria, Gammaproteobacteria, Deltaproteobacteria and Epsilonproteobacteria.**

# Domain: Bacteria

## Phylum: Proteobacteria

- A new class (a sixth class) within the phylum Proteobacteria, **Acidithiobacillia** classis nov., was proposed by Williams and Kelly, 2013 and replaced by the '**Zetaproteobacteria**', a sixth class was proposed earlier by Emerson *et al.*, 2007 and McAllister *et al.*, 2011).



**Zetaproteobacteria** was excluded by Williams and Kelly, 2013.

# Domain: Bacteria

## Phylum: Proteobacteria

### Sequence of some representative rRNA-targeted oligonucleotide probes

Probe	Position	Probe sequence (5' → 3')	Specificity
ALF1b	16S rRNA 19–35	CGTTCG(C/T)TCTGAGCCAG	“Alphaproteobacteria,” but not exclusive
BET42a	23S rRNA 1027–1043	GCCTTCCCACATTCGTTT	“Betaproteobacteria”
GAM42a	23S rRNA 1027–1043	GCCTTCCCACATTCGTTT	“Gammaproteobacteria,” but not the deeply branching taxa
Delta 385	16S rDNA 385–402	CGGCGT(C/T)GCTGCGTCAGG	“Deltaproteobacteria” sulfate-reducers, but not exclusive

### Probes for fluorescent in-situ hybridization

Specific 16S rRNA sequence signatures for the various classes of the Proteobacteria have been described and used for the construction of DNA probes. Such probes were extensively applied for the detection and visualization of Proteobacteria.



**Domain:  
Bacteria  
Phylum:  
Proteobacteria**

**Some selected key  
genera, general  
characteristics, and  
differentiating  
features of the five  
classes of the  
Proteobacteria.**

	Proteobacterial class				
	Alpha	Beta	Gamma	Delta	Epsilon
Important genera	<i>Acetobacter</i> <i>Agrobacterium</i> <sup>a</sup> <i>Bartonella</i> <sup>a</sup> <i>Bradyrhizobium</i> <i>Brucella</i> <sup>a</sup> <i>Caulobacter</i> <sup>a</sup> <i>Ehrlichia</i> <i>Gluconobacter</i> <i>Hyphomicrobium</i> <i>Mesorhizobium</i> <sup>a</sup> <i>Methylobacterium</i> <sup>b</sup> <i>Nitrobacter</i> <i>Rhizobium</i> <i>Rhodobacter</i> <sup>b</sup> <i>Rhodospirillum</i> <i>Sinorhizobium</i> <sup>a</sup> <i>Sphingomonas</i> <sup>b</sup> <i>Rickettsia</i> <sup>a,b</sup> <i>Wolbachia</i> <sup>b</sup>	<i>Alcaligenes</i> <i>Bordetella</i> <sup>a,b</sup> <i>Burkholderia</i> <sup>b</sup> <i>Comamonas</i> <i>Neisseria</i> <sup>a,b</sup> <i>Nitrosomonas</i> <sup>b</sup> <i>Ralstonia</i> <sup>b</sup> <i>Rhodocyclus</i> <i>Sphaerotilus</i> <i>Spirillum</i> <i>Thiobacillus</i>	<i>Actinobacillus</i> <sup>b</sup> <i>Azotobacter</i> <i>Buchnera</i> <sup>a</sup> <i>Chromatium</i> <i>Coxiella</i> <sup>b</sup> <i>Erwinia</i> <sup>b</sup> <i>Escherichia</i> <sup>a,b</sup> <i>Francisella</i> <sup>b</sup> <i>Haemophilus</i> <sup>a,b</sup> <i>Legionella</i> <sup>b</sup> <i>Methylococcus</i> <sup>b</sup> <i>Pasteurella</i> <sup>a</sup> <i>Pectobacterium</i> <i>Pseudomonas</i> <sup>a,b</sup> <i>Salmonella</i> <sup>a,b</sup> <i>Shewanella</i> <sup>b</sup> <i>Shigella</i> <sup>a,b</sup> <i>Stenotrophomonas</i> <i>Vibrio</i> <sup>a,b</sup> <i>Xanthomonas</i> <sup>a,b</sup> <i>Xylella</i> <sup>a,b</sup> <i>Yersinia</i> <sup>a,b</sup>	<i>Bdellovibrio</i> <i>Chondromyces</i> <i>Desulfobacter</i> <i>Desulfovibrio</i> <sup>b</sup>  <i>Geobacter</i> <sup>b</sup> <i>Myxococcus</i> <sup>b</sup> <i>Polyangium</i> <i>Syntrophus</i>	<i>Campylobacter</i> <sup>a</sup> <i>Helicobacter</i> <sup>a</sup> <i>Sulfurospirillum</i> <i>Wolinella</i>
Number of genera/ number of species <sup>c</sup>	140/425	76/225	181/755	57/165	6/49
Major ubiquinone type <sup>d</sup>	Q-10	Q-8	Q-8, Q-9, or Q-10 to Q-14	—	—
Major menaquinone type <sup>d</sup>	Some contain also MK-9 or MK-10	Some contain also MK-8	Some contain also MK-8 or MK-7	MK-6, MK-6(H <sub>2</sub> ), MK-7, MK-7(H <sub>2</sub> ) or MK-8 <sup>e</sup>	MK-6, methyl-substituted MK-6 <sup>f</sup>
Characteristic polyamines <sup>g</sup>	Most contain a triamine (sym-homospermidine or spermidine)	2-Hydroxy-putrescine	Spermidine and/or putrescine or cadaverine; or 1,3-diaminopropane	Most contain a triamine (sym-homospermidine or spermidine)	Spermidine

Symbols and abbreviations: —, absent; DMK, demethylmenaquinone; and MK-6(H<sub>2</sub>), hydrogenated menaquinone-6.

The genome of at least one representative strain has been sequenced (as of mid 2002).

Sequencing of the genome of at least one representative strain is in progress (as of mid 2002; see, e.g., <http://www.tigr.org/> or <http://www.ncbi.nlm.nih.gov>).

Only validly published names (situation as of mid 2002).

Collins and Jones (1981), Hiraishi et al. (1984), <http://www.wdcm.nig.ac.jp/cgi-bin/search.cgi>, and H.J. Busse, personal communication.

Collins and Widdel (1986).

Moss et al., (1990).

Auling (1992), Busse and Auling (1988), and Hamana and Matsuzaki (1993).

# Domain: Bacteria

## Phylum: Proteobacteria

Some selected  
plant diseases  
caused by  
Proteobacteria.

Proteobacterial class and species	Family *	Disease (symptoms)
"Alphaproteobacteria"		
<i>Agrobacterium rhizogenes</i>	Rhizobiaceae	Hairy root
<i>Agrobacterium tumefaciens</i>	Rhizobiaceae	Crown gall
" <i>Candidatus Liberibacter asiaticus</i> "	in cluster of Rhizobiaceae, Bartonellaceae, etc.	Greening disease on citrus (a phloem-restricted disease)
"Betaproteobacteria"		
<i>Acidovorax anthurtii</i>	Comamonadaceae	Leaf-spot on <i>Anthurium</i>
<i>Burkholderia cepacia</i>	"Burkholderiaceae"	Soft rot (sour skin on onion)
<i>Burkholderia glumae</i>	"Burkholderiaceae"	Sheath necrosis on rice
<i>Ralstonia solanacearum</i>	"Ralstoniaceae"	Moko disease on banana (vascular wilt)
<i>Xylophilus ampelinus</i>	Comamonadaceae	Necrosis and canker on grapevine
"Gammaproteobacteria"		
<i>Brenneria (Erwinia) salicis</i>	Enterobacteriaceae	Watermark disease on willow
<i>Brenneria nigrifluens</i>	Enterobacteriaceae	Bark canker on Persian walnut ( <i>Juglans regia</i> )
<i>Erwinia amylovora</i>	Enterobacteriaceae	Fire blight on pome fruit (vascular wilt)
<i>Erwinia stewartii</i>	Enterobacteriaceae	Stewart's wilt on corn (vascular wilt)
<i>Pectobacterium (Erwinia) carotovorum</i>	Enterobacteriaceae	Soft rot
<i>Pseudomonas agarici</i>	Pseudomonadaceae	Spots on mushrooms
<i>Pseudomonas marginalis</i>	Pseudomonadaceae	Soft rot (pink eye) on potato
<i>Pseudomonas savastanoi</i>	Pseudomonadaceae	Galls on olive trees
<i>Pseudomonas syringae</i>	Pseudomonadaceae	Wildfire on tobacco, haloblight on beans, spots on tomato and pepper (blights and spots)
<i>Pseudomonas syringae</i>	Pseudomonadaceae	Canker on stone fruit
<i>Xanthomonas campestris</i>	"Xanthomonadaceae"	Black rot on crucifers (vascular wilt)
<i>Xanthomonas citri</i>	"Xanthomonadaceae"	Canker on citrus
<i>Xanthomonas oryzae</i>	"Xanthomonadaceae"	Blight on rice
<i>Xanthomonas populi</i>	"Xanthomonadaceae"	Canker on poplar trees
<i>Xanthomonas translucens</i>	"Xanthomonadaceae"	Blight on cereals
<i>Xanthomonas vesicatoria</i>	"Xanthomonadaceae"	Spots on tomato and pepper
<i>Xylella fastidiosa</i>	"Xanthomonadaceae"	Pierce's disease (e.g., on grapevine)

\* According to *Bergey's Manual of Systematic Bacteriology* (Garrrity and Holt, 2001). See also Fig. 1. Quotation marks are used for names which have not yet been validated (as of mid 2002).



# Alphaproteobacteria

## Purple sulfur bacteria

---

- 5/6 genera contain plant pathogens.
  1. *Acetobacter* and *Gluconobacter* in *Acetobacteriaceae*;
  2. *Sphingomonas*
  3. *Agrobacterium*, and
  4. *Candidatus Liberibacter*.



# Betaproteobacteria

## Purple non-sulfur bacteria

---

- Six genera contain pathogens and these represent 4 of the 5 families in the *Burkholderiales*.
- *Acidovorax* in *Comamonadaceae*
- *Burkholderia* in *Burkholderiaceae*
- *Ralstonia* in *Ralstoniaceae*
- *Herbaspirillum* and *Janthinobacterium* in *Oxalobacteriaceae*
- *Xylophilus* (family not certain).



# Gammaproteobacteria

---

- Three main families:
- *Enterobacteriaceae* - 10 genera containing plant pathogens. e.g. *Erwinia*.
- *Pseudomonodaceae* - 1 genus (*Pseudomonas*).
- *Xanthomonodaceae* - 2 genera (*Xanthomonas* and *Xylella*).



# Three-Domain Classification

## 2. Domain Archaea

---

- Most of the archaea are methanogens and extremophilic in origin.
- They reside in extremely hostile conditions.

Hostility	Name
40-85°C	Thermophile
>85°C	Hyperthermophiles
20-40°C	Mesophiles
<20°C	Psychrophiles
15% of NaCl	Halophiles
pH>7	Alkaliphiles
pH<7	Acidophiles



# Three-Domain Classification

## Domain Archaea

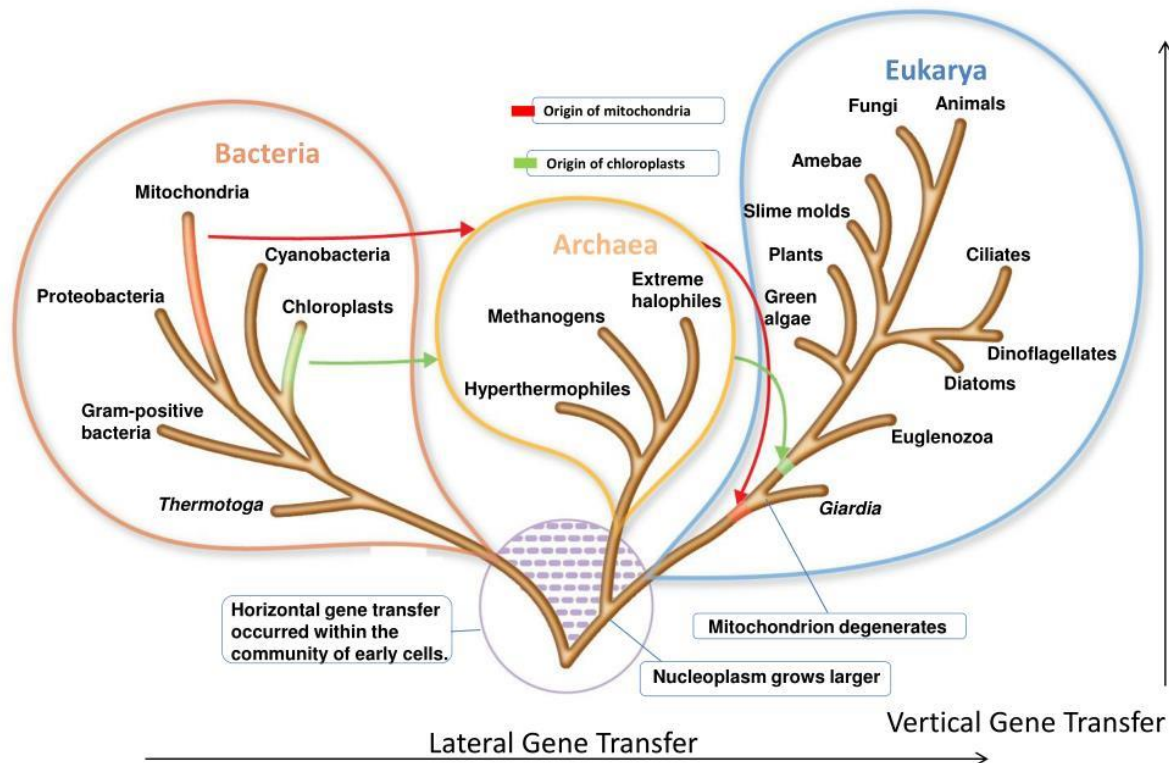
---

- Can archaea be cultured?
- Culturing methanogenic archaea is fastidious, expensive, and requires an external source of hydrogen and carbon dioxide.
- Until now, these microorganisms have only been cultivated under strictly anaerobic conditions.
- Note: Aerobic halophilic archaea are pretty easy to grow in standard labs.

# Three-Domain Classification

## 3. Domain Eukarya

Figure 10.1 The Three-Domain System.



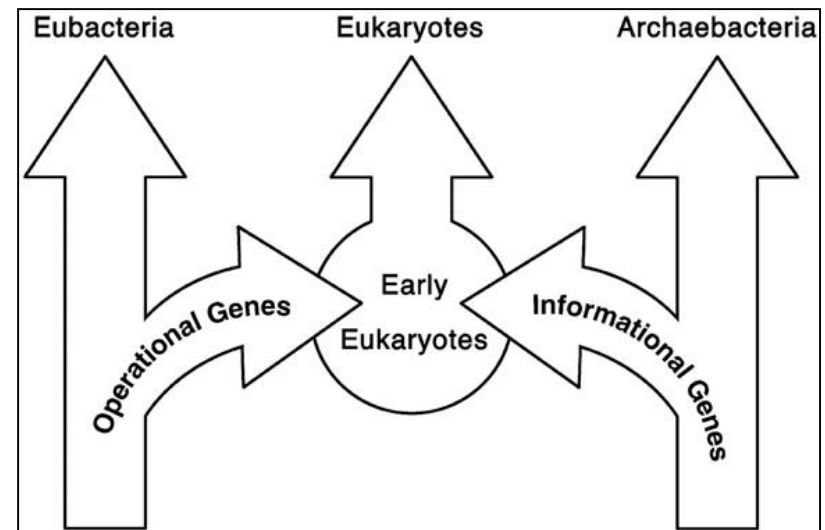


# Endosymbiosis theory for eukaryote origin

## Mitochondria and chloroplasts

### endosymbiotic theory

1. When data from mitochondrial and chloroplast rRNA are placed in the universal tree of life, they appear along with the Bacteria.
2. Mitochondria probably arose from a group of bacteria that includes the modern genera *Agrobacterium*, *Rhizobium*, and the rickettsias.
3. Chloroplasts share a common ancestor with the cyanobacteria.



Informational genes involve central processes of gene expression (protein synthesis); they tend to be transferred vertically. Operational genes (those involved in housekeeping) involve metabolic processes that function independently of other components. They are more likely to be transferred horizontally.



# Endosymbiosis theory for eukaryote origin

## Mitochondria and chloroplasts

### endosymbiotic theory

---

- Although it is likely that single celled Eukaryotes were also present on Earth from the very beginning, there is also considerable evidence that Archaea, Bacteria, and Viruses transferred genes to these single celled Eukaryotes, thus trigger multi-cellularity (Joseph 2009b,c).
- Thus we see that the genomes of modern day eukaryotic species, including humans, contain highly conserved genes were acquired from Archaea and Bacteria.



# Endosymbiosis theory for eukaryote origin

## Mitochondria and chloroplasts

### endosymbiotic theory

---

- However, not all of these genes have been expressed, whereas yet other were silenced or activated in response to specific environmental signals, thereby giving rise to new species (Joseph 2000, 2009b,c).
- Genes transferred to the eukaryotic genome by prokaryotes and Viruses, include exons, introns, transposable elements, informational and operational genes, RNA, ribozomes, mitochondria, and the core genetic machinery for translating, expressing, and repeatedly duplicating genes and the entire genome.

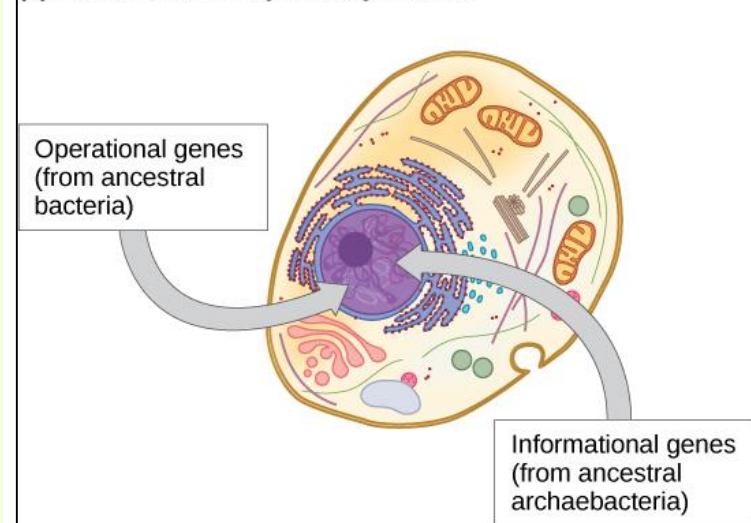
# Endosymbiosis theory for eukaryote origin

## Mitochondria and chloroplasts

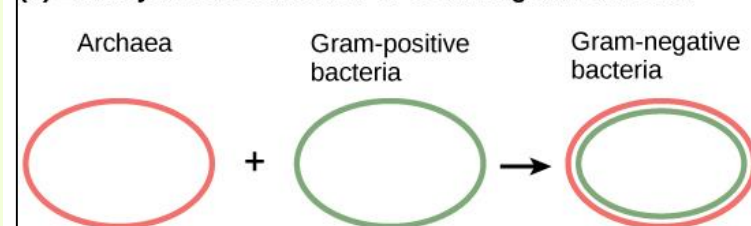
### endosymbiotic theory

- The theory that mitochondria and chloroplasts are endosymbiotic in origin is now widely accepted.
- More controversial is the proposal that:
  - a) the eukaryotic nucleus resulted from the fusion of archaeal and bacterial genomes; and that
  - b) Gram-negative bacteria, which have two membranes, resulted from the fusion of Archaea and Gram-positive bacteria, each of which has a single membrane.

(a) Genome fusion by endosymbiosis

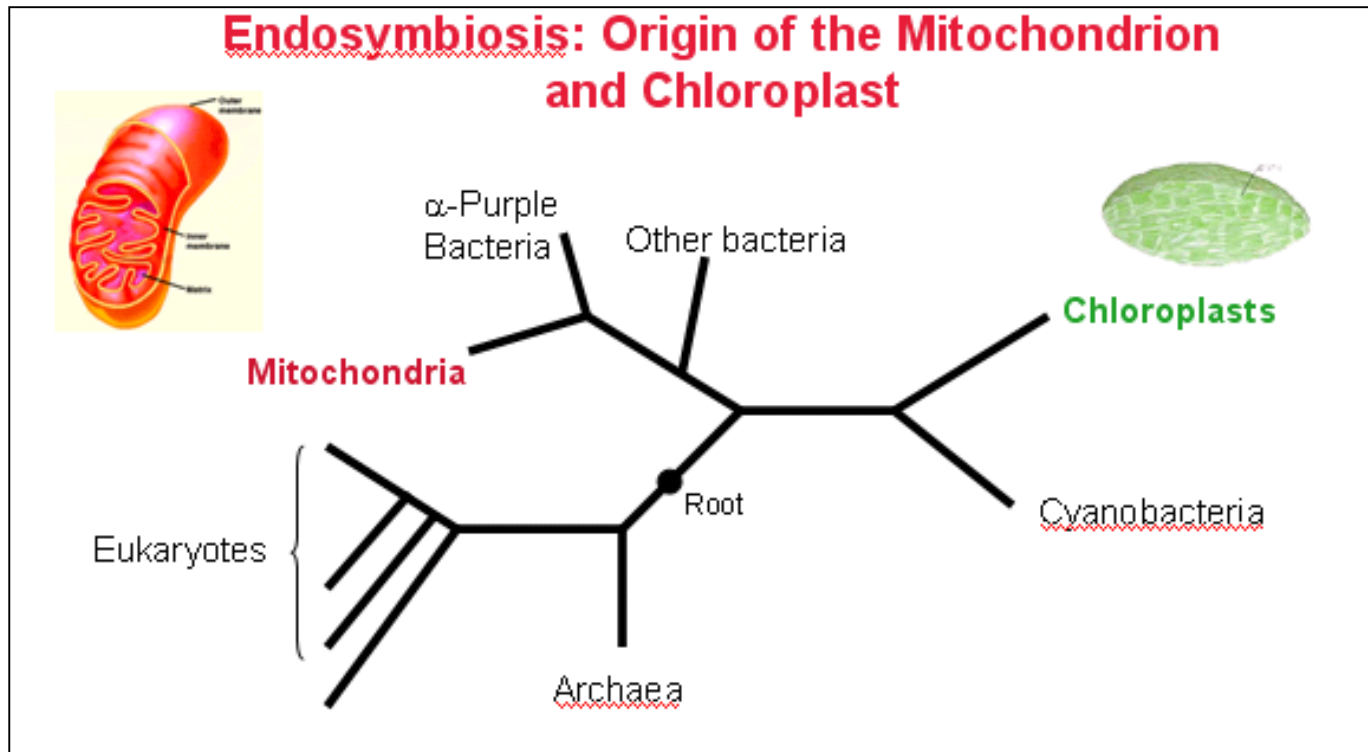


(b) Endosymbiotic formation of Gram-negative bacteria



# Endosymbiosis theory for eukaryote origin

## Endosymbiosis

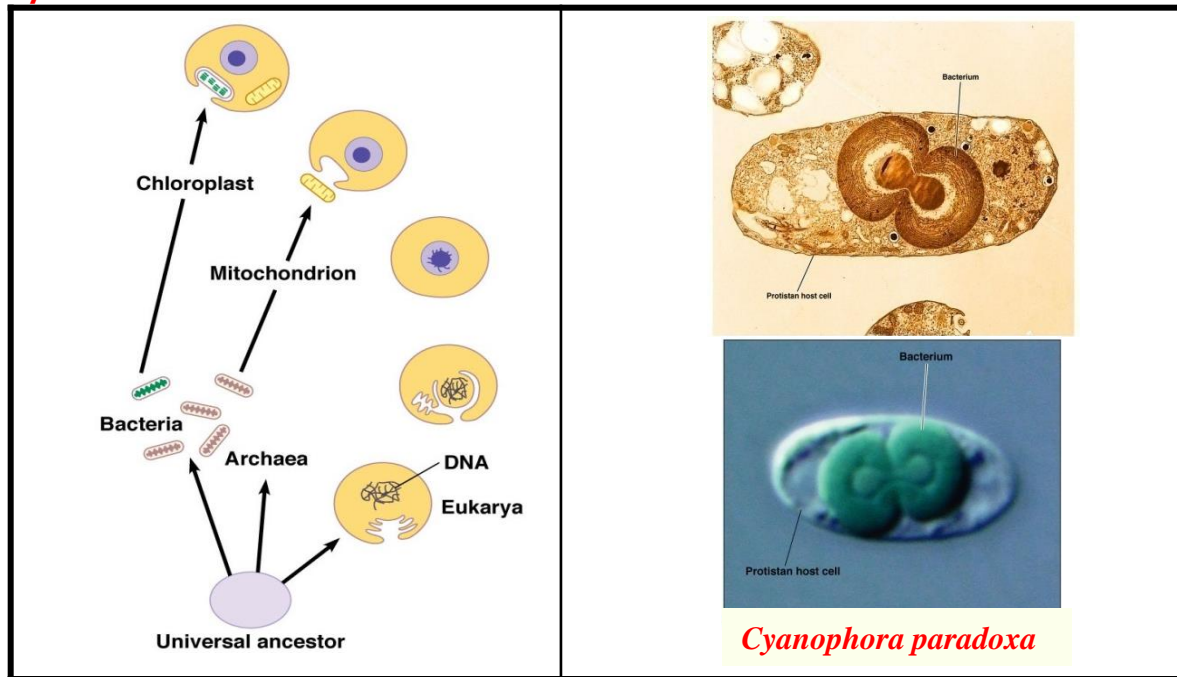


Mitochondria and chloroplasts are derived from the  $\alpha$ -purple bacteria and the cyanobacteria, respectively, via separate endosymbiotic events.

# Endosymbiotic Theory

There is compelling evidence that mitochondria and chloroplasts were once primitive bacterial cells. This evidence is described in the endosymbiotic theory

- Archaea invaded by bacteria capable of cellular respiration (mitochondria) and capable of photosynthesis.
- Bacteria take up permanent residence and become organelles of eukaryotes.



# Endosymbiosis theory for eukaryote origin

## *Cyanophora paradoxa*

- The **glaucophytes** are of interest to biologists studying the development of chloroplasts because some studies suggest they may be similar to the **original algal type** that led to **green plants and red algae**.
- The chloroplasts of glaucophytes are known as 'cyanelles' or 'cyanoplasts'.
- Unlike the **chloroplasts in other organisms**, they have a **peptidoglycan layer**, believed to be a relic of the endosymbiotic origin of plastids from **cyanobacteria**.
- *C. paradoxa* has two cyanelles or chloroplasts where
  1. **nitrogen fixation occurs** alongside the
  2. **primary function of photosynthesis**.

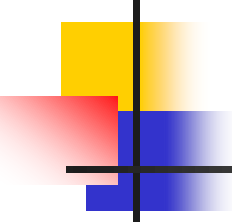
Plastid- A major **double-membrane organelle** found, among others, in the cells of **plants and algae**.

# Endosymbiosis theory for eukaryote origin

## Endosymbiosis

- Evidence that mitochondria and plastids (e.g. chloroplasts) arose from bacteria is as follow:
  1. New mitochondria and chloroplasts are formed only through a process similar to binary fission.
  2. Both mitochondria and plastids contain single circular DNA that is different from that of the cell nucleus and that is similar to that of bacteria (both in their size and structure).
  3. The genomes, including the specific genes, are basically similar between mitochondria and the Rickettsial bacteria.
  4. Mitochondria have several enzymes and transport systems similar to those of bacteria.
  5. These organelles' ribosomes are like those found in bacteria (70S).





## Origin of diderm (Gram-negative) bacteria: antibiotic selection pressure rather than endosymbiosis likely led to the evolution of bacterial cells with two membranes

---

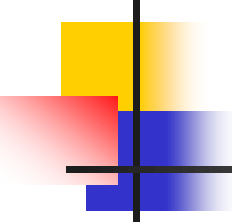
- The **prokaryotic organisms** can be divided into **two main groups** depending upon whether their cell envelopes contain one membrane (**monoderms**) or two membranes (**diderms**).
- It is important to understand how these and other variations that are observed in the cell envelopes of prokaryotic organisms have originated.
- In 2009, **James Lake** proposed that **cells with two membranes (primarily Gram-negative bacteria)** originated from an **ancient endosymbiotic event** involving an Actinobacteria and a Clostridia (Lake 2009).



## Origin of diderm (Gram-negative) bacteria: antibiotic selection pressure rather than endosymbiosis likely led to the evolution of bacterial cells with two membranes

---

- Some bacterial phyla, such as **Deinococcus-Thermus**, which **lack lipopolysaccharide (LPS)** and yet contain some characteristics of the **diderm bacteria**, are postulated as **evolutionary intermediates (simple diderms)** in the transition between the **monoderm bacterial taxa** and the **bacterial groups that have the archetypal LPS-containing outer cell membrane found in Gram-negative bacteria**.
- It is possible to **distinguish the two stages in the evolution of diderm-LPS cells** (viz. **monoderm bacteria → simple diderms lacking LPS → LPS containing archetypal diderm bacteria**) by means of conserved inserts in the Hsp70 and Hsp60 proteins.



## Origin of diderm (Gram-negative) bacteria: antibiotic selection pressure rather than endosymbiosis likely led to the evolution of bacterial cells with two membranes

---

- There is no reliable evidence to support the endosymbiotic origin of double membrane bacteria.
- In contrast, many observations suggest that antibiotic selection pressure was an important selective force in prokaryotic evolution and that it likely played a central role in the evolution of diderm (Gram-negative) bacteria.

# Endosymbiosis theory for eukaryote origin

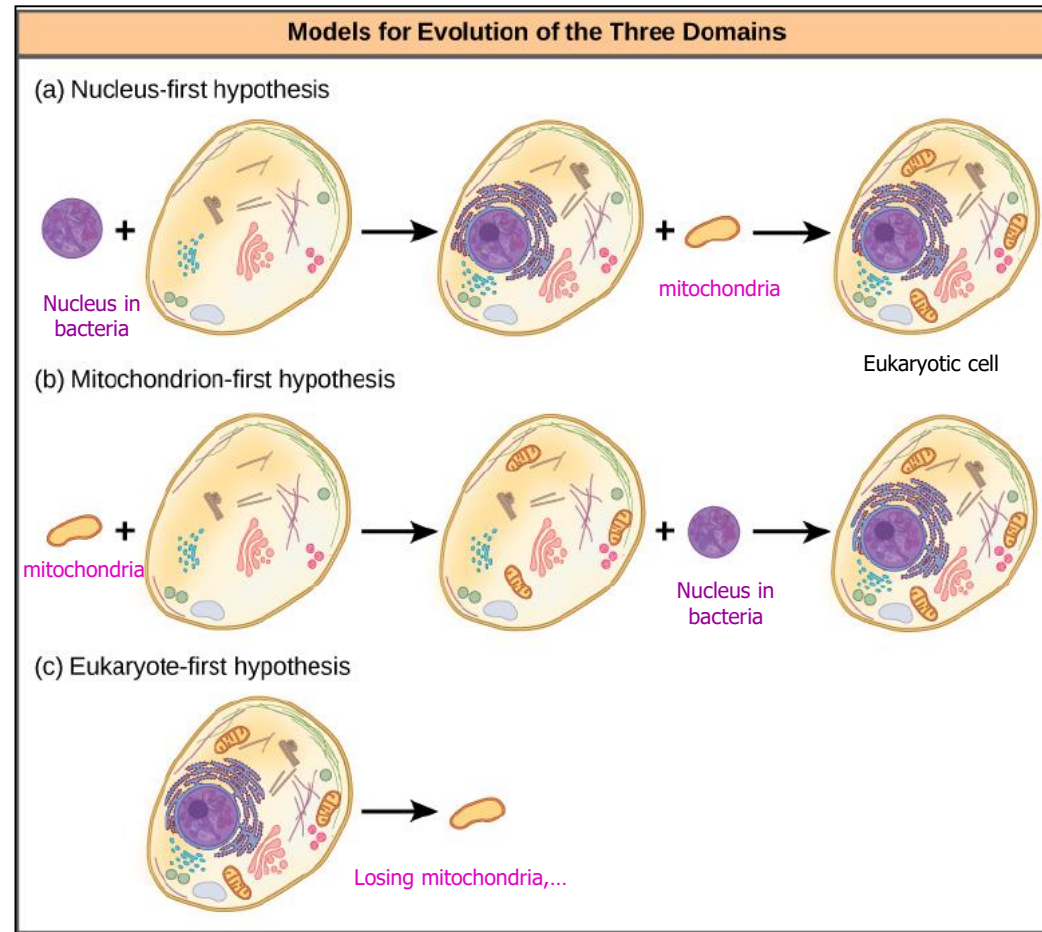
## How did the eukaryotic cell evolve?

- a) **The nucleus-first hypothesis** proposes that the **nucleus evolved in prokaryotes first**, followed by a **later fusion of the new eukaryote with bacteria that became mitochondria**.
- b) **The mitochondria first hypothesis** proposes that mitochondria were first established in a **prokaryotic host**, which subsequently acquired a nucleus, by fusion or other mechanisms, to become the **first eukaryotic cell**.
- c) **The eukaryote-first hypothesis** proposes that **prokaryotes actually evolved from eukaryotes by losing genes and complexity**.
- All of these hypotheses are testable. Only time and more experimentation will determine which hypothesis data best supports.

# Endosymbiosis theory for eukaryote origin

## Three alternate hypotheses of eukaryotic and prokaryotic evolution

- a) The nucleus-first hypothesis- nucleus evolved in prokaryotes first, followed by a later fusion of the new eukaryote with bacteria that became mitochondria.
- b) The mitochondrion-first hypothesis- mitochondria were first established in a prokaryotic host, which subsequently acquired a nucleus, by fusion or other mechanisms, to become the first eukaryotic cell.
- c) The eukaryote-first hypothesis proposes that prokaryotes actually evolved from eukaryotes by losing genes and complexity.



# Web and Network Model

W. Ford Doolittle

Web of life

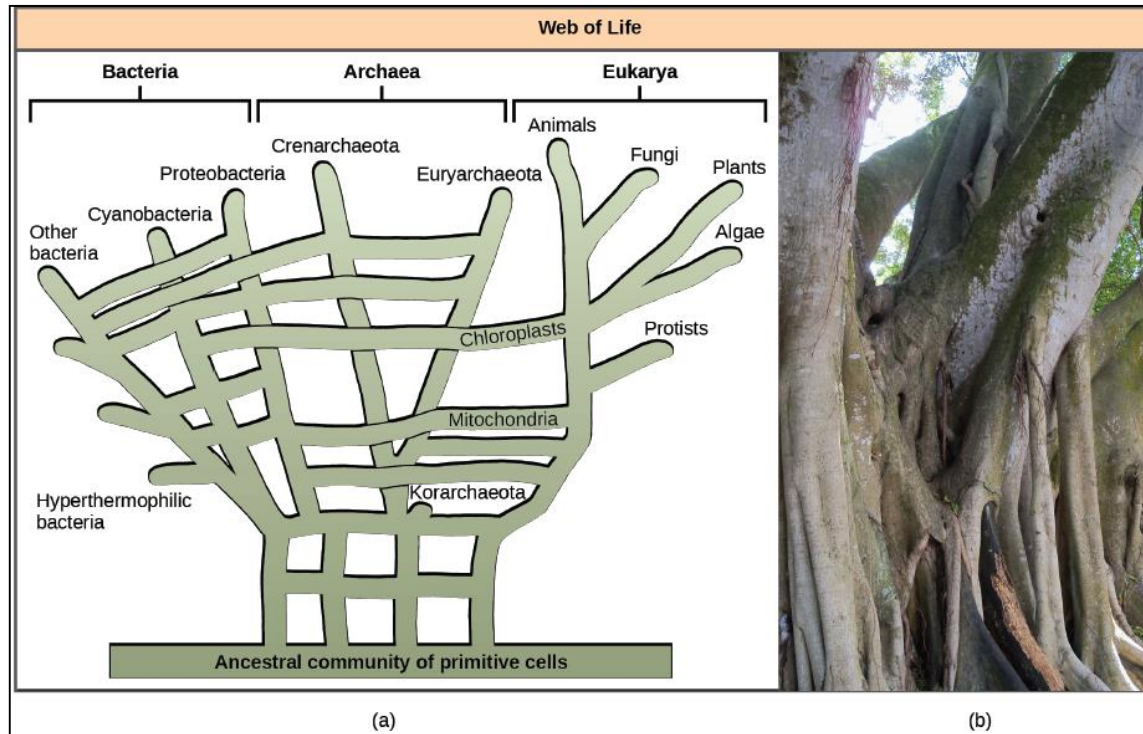
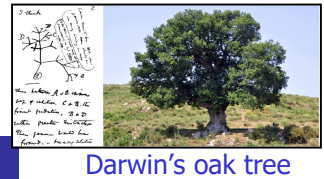


---

- In 1999, W. Ford Doolittle proposed a phylogenetic model that resembles a web or a network more than a tree.
- The hypothesis is that eukaryotes evolved not from a single prokaryotic ancestor, but from a pool of many species that were sharing genes by HGT mechanisms.
  - a) some individual prokaryotes were responsible for transferring the bacteria that caused mitochondrial development to the new eukaryotes; whereas, other species transferred the bacteria that gave rise to chloroplasts.
  - b) Scientists often call this model the “**web of life.**”

# Web and Network Model

## Web of life



- (a) phylogenetic model resembles a web or a network more than a tree proposed by W. Ford Doolittle, 1999. The hypothesis is that eukaryotes evolved not from a single prokaryotic ancestor, but from a pool of many species that were sharing genes by HGT mechanisms. Connections between branches occur by horizontal gene transfer.
- (b) Visually, this concept is better represented by the multi-trunked Ficus than by an oak's single trunk similar to Darwin's tree.

# Independent analyses that either confirm or refute the rRNA (Woesean tree)

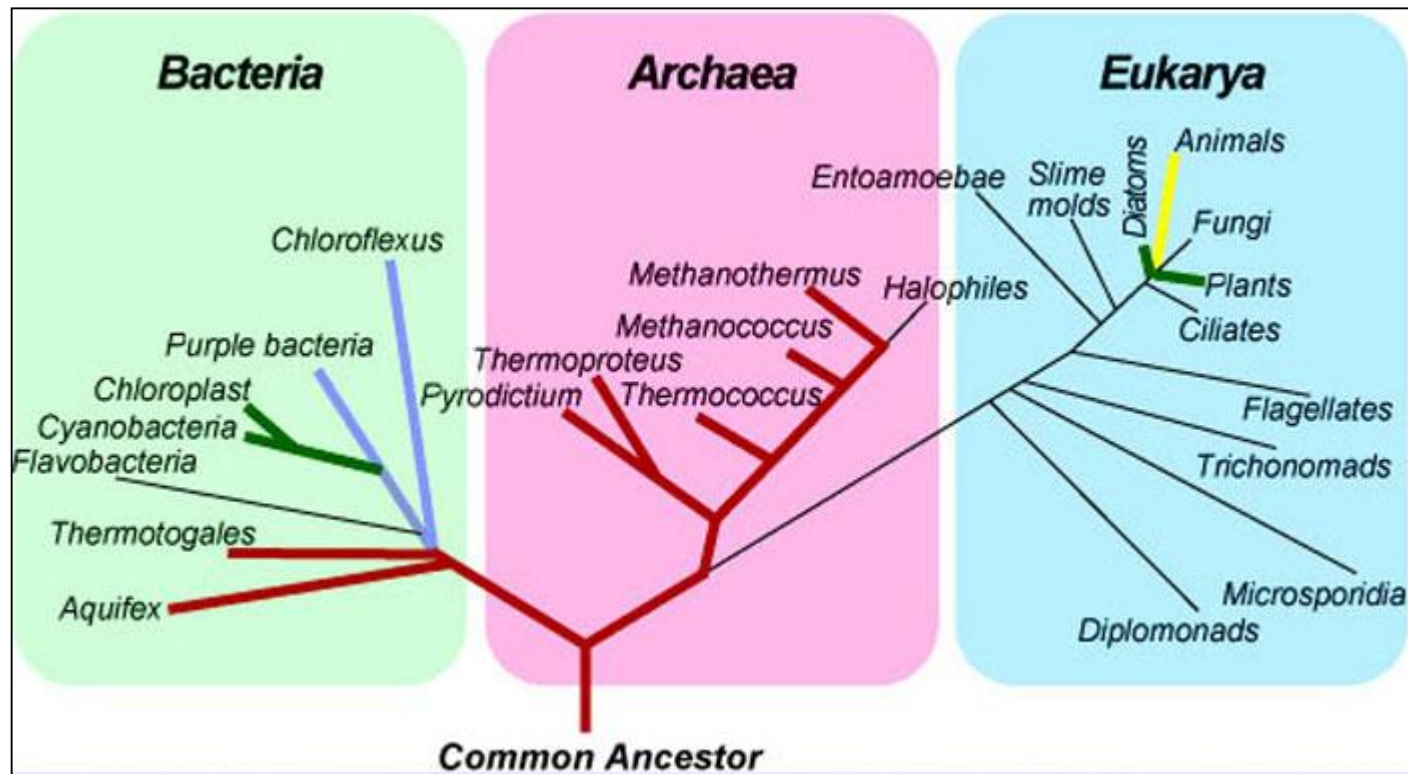


- Brochier and Philippe, 2002
- Leart *et al.*, 2003
- Gupta's indel analysis, 1998
- Cavalier-Smith analysis, 2002
- Arthur L. Koch, 2003
- Rivera and Lake analysis, 2004
- Lake and colleague's Eocyte hypothesis, 1984
- Rivera and James analysis, 2004
- The new tree of life by Hug *et al.*, 2016
- Ruggiero *et al.*, 2015



# Three-Domain Classification

## Phylogenetic position of Mollicutes among bacteria, using 16S/18S rRNA sequences

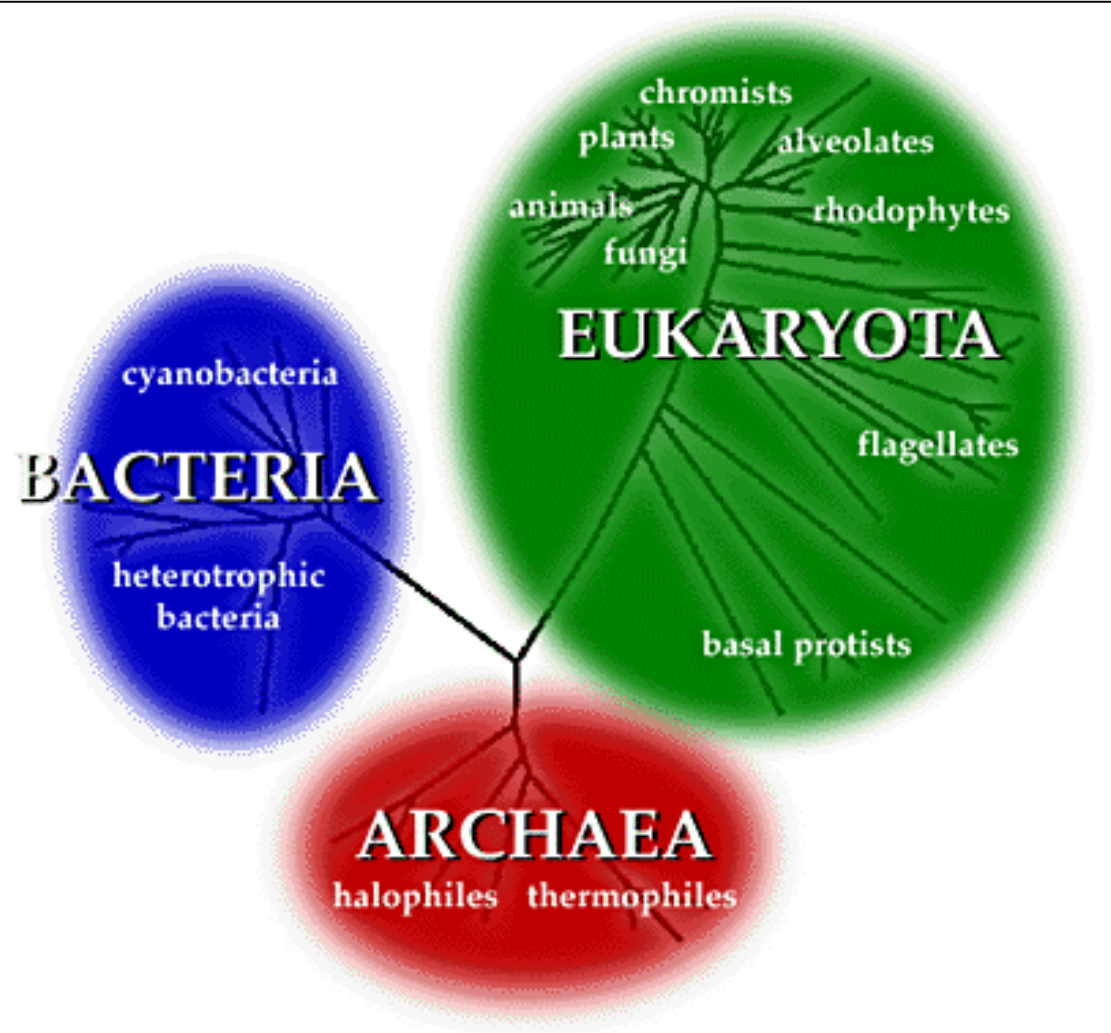


Cyanobacteria are relatives of the bacteria but not eukaryotes. Because they are photosynthetic and aquatic, cyanobacteria are often called "blue-green algae".  
**Archaea are called 'extremeophiles'.**

# Woesian tree of life

## Three-Domain Classification

### Phylogenetic Relationships



# Woesian tree of life

## Three-Domain Classification

### Phylogenetic Relationships

- Archaea are so named because they are believed to be the least evolved forms of life on Earth (archae meaning ancient).
- The ability of some archaea to live in environmental conditions similar to the early Earth gives an indication of the ancient heritage of the domain.
- The early Earth was hot, with a lot of extremely active volcanoes and an atmosphere composed mostly of nitrogen, methane, ammonia, carbon dioxide, and water.
- There was little if any oxygen in the atmosphere.
- Archaea and some bacteria evolved in these conditions, and are able to live in similar harsh conditions today.
- Many scientists now suspect that those two groups diverged from a common ancestor relatively soon after life began.

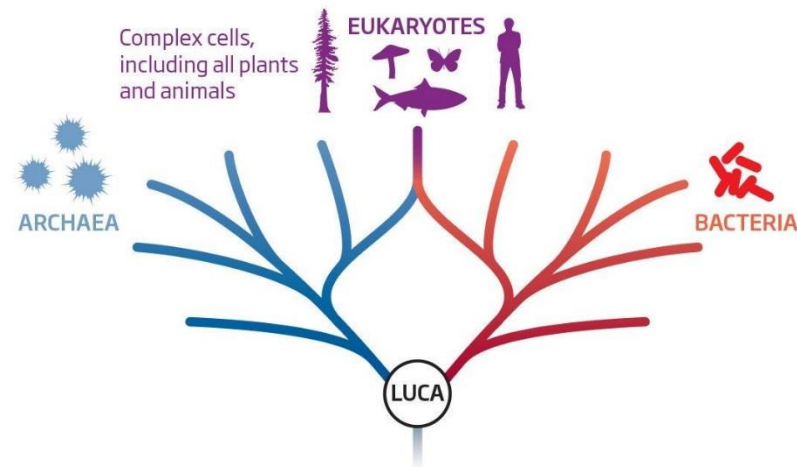
# Woeseian tree

Evolutionary relationships among the three domains  
**Based on their ribosomal RNA differences**

- The diagram models the pattern of **ribosomal RNA sequence diversification**, and presumably of the change in the basal genetic machinery of life.

## Meet your maker

We're getting closer to understanding what the last universal common ancestor of all life on Earth, LUCA, was like and where it lived



LUCA emerged around 3.8 billion years ago and gave rise to two kinds of simple cells: **bacteria and archaea**. By looking **for genes common to almost all cells living today**, previous studies have **identified around 100 genes almost certainly present in LUCA**.



# Woesian tree

---

- Data from other labs to confirm or refute what he was finding were hard to come by.
- He preferred to be in the lab sequencing the rRNA for a new organism rather than socializing with fellow scientists and lobbying for them to support his interpretation of the data.



# Woesian tree

## Challenged by other sequence analyses

---

- The three domain paradigm was challenged by:
  1. Other sequence analyses, and
  2. The morphological characterization of cellular envelop of gram negative and gram-positive bacteria.
- The former (gram negative) are surrounded by an external and an internal membrane (diderm) and while the latter (gram positive), one membrane (monoderm).

# Brochier and Philippe,2002

## The first emerging bacterial group- A non-hyperthermophilic ancestor for bacteria

- The first phyla that emerge in the tree of life based on ribosomal RNA (rRNA) sequences are hyperthermophilic, which led to the hypothesis that the universal ancestor, and possibly the original living organism, was hyperthermophilic.
- Here we reanalyse the bacterial phylogeny based on rRNA using a more reliable approach, and find that hyperthermophilic bacteria (such as Aquificales and Thermotogales) do not emerge first, suggesting that the bacteria had a non-hyperthermophilic ancestor.
- It seems that Planctomycetales, a phylum with numerous peculiarities, could be the first emerging bacterial group.

# Planctomycetes

**The first emerging bacterial group- A non-hyperthermophilic ancestor for bacteria**

---

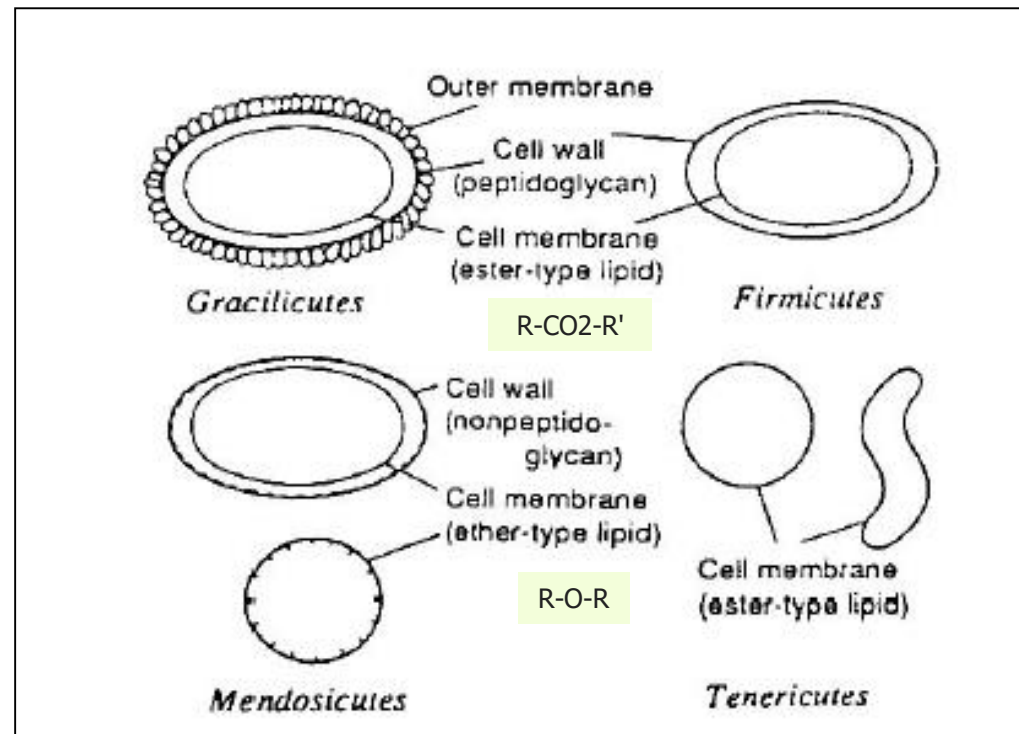
- Planctomycetes are a phylum of aquatic bacteria.
- They don't have nucleus and reproduce by budding.
- Cavailier-Smith has postulated that the Planctomycetes are within the clade Planctobacteria in the larger clade Gracilicutes.
- The organisms belonging to this group lack murein (peptidoglycan) in their cell wall.
- Instead their walls are made up of glycoprotein rich in glutamate.
- Planctomycetes have internal structures that are more complex than would be typically expected in prokaryotes.



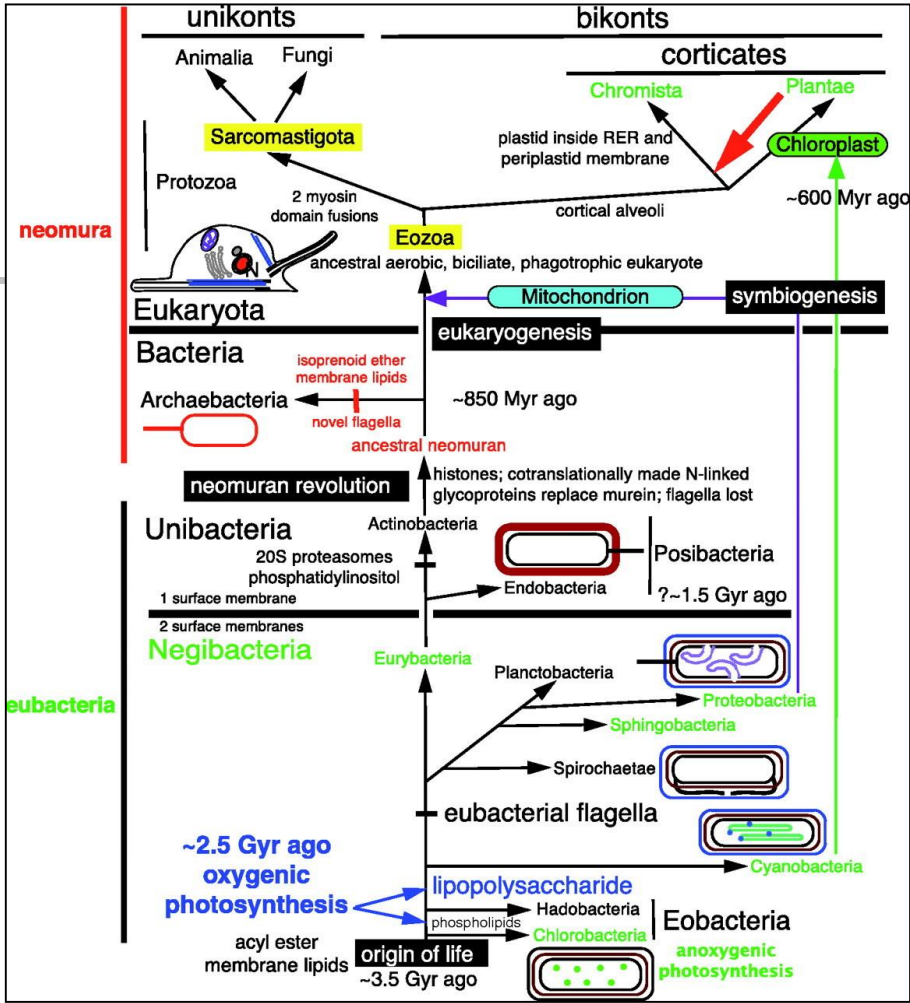
# Four main bacterial cell wall

## Gracilicutes, Firmicutes, Tenericutes, Mendosicutes

- Cellular envelop in **Gram negative bacteria** are surrounded by two layers: an external and an internal membrane (diderm) while **Gram positive bacteria** have one membrane (monoderm).



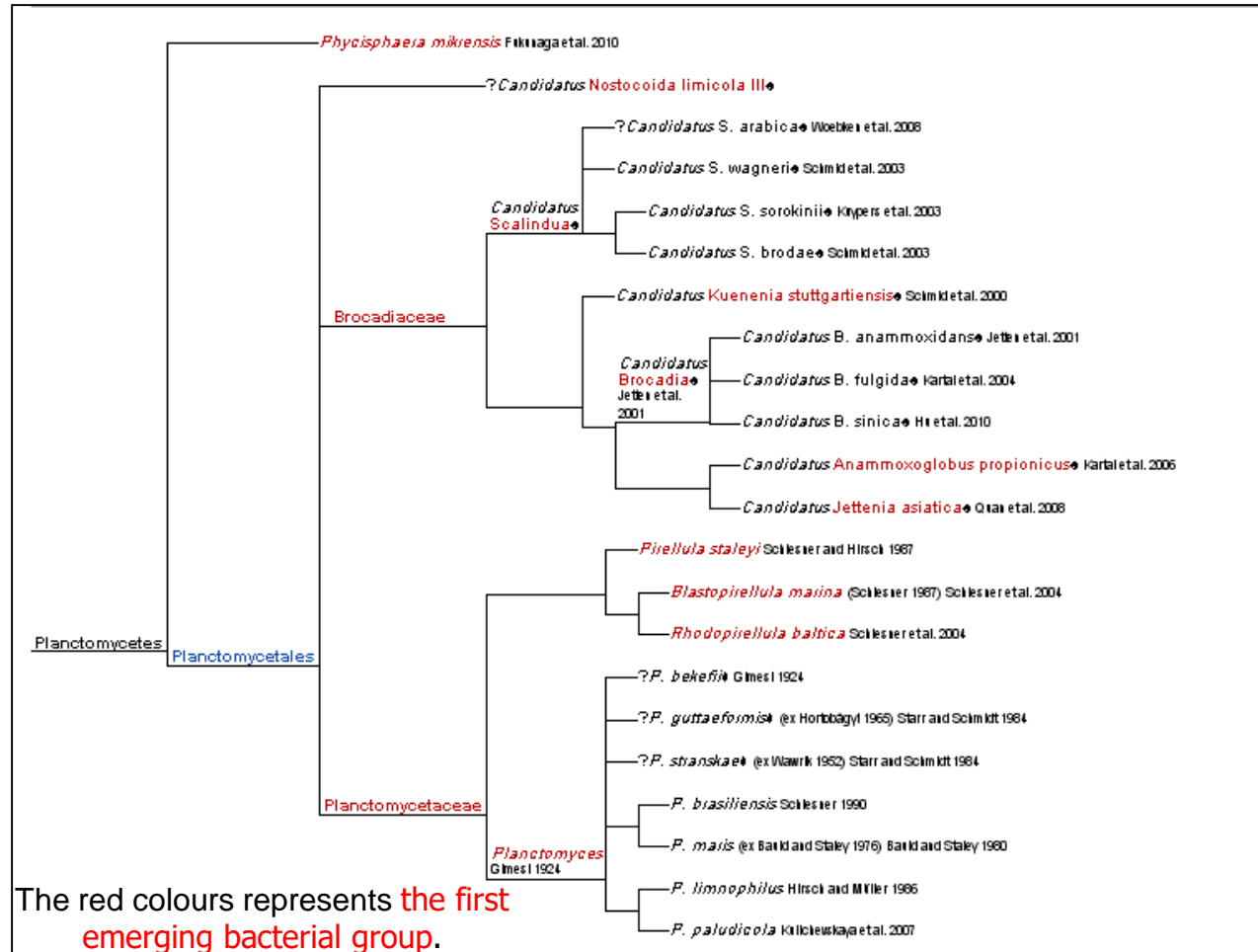
- **Intracellular coevolutionary theory:**
- The last common ancestor of eukaryotes was a sexual phagotrophic protozoan with mitochondria.
- The eukaryotic cytoskeleton and endomembrane system originated through cooperatively enabling the evolution of phagotrophy.
- Eukaryotes plus their archaeobacterial sisters form the clade Neomura.



Bikont is a eukaryotic cell with two flagella; thought to be the ancestor of all plants while unikont is a eukaryotic cell with a single flagellum; thought to be the ancestor of all animals.

# Planctomycetes

The first emerging bacterial group- A non-hyperthermophilic ancestor for bacteria





# Protein sequences

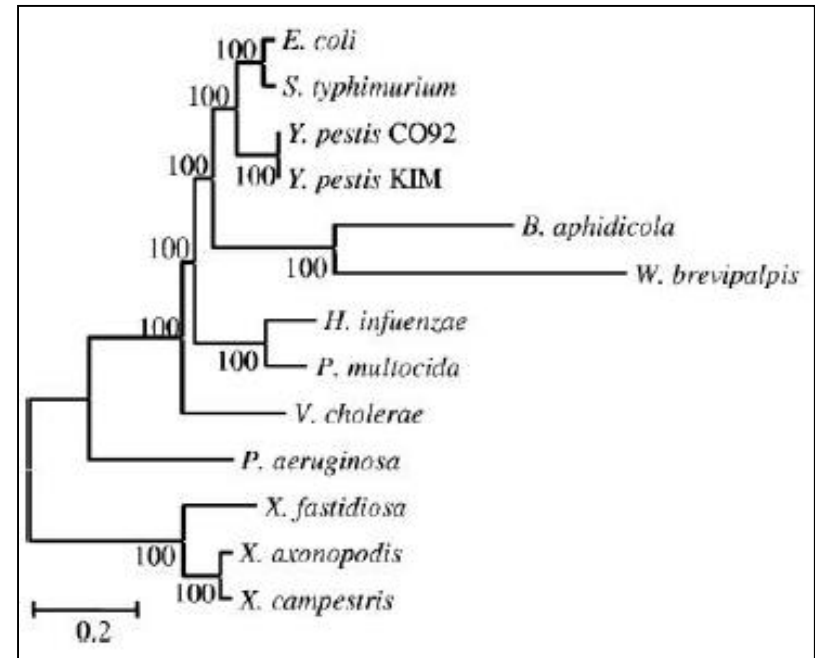
---

- Sequence analyses based on functional proteins across the three domains also suggest each of the three domains as independent monophyletic lineage representing:
  - Ribosomal,
  - Metabolic,
  - Biosynthetic proteins,
  - Replicational,
  - Transcriptional, and
  - Translational machineries.

# Protein analysis of Leart *et al.*, 2003

## A consistent result

- Neighbor-joining tree based on the concatenation of 205 proteins (Lerat *et al.*, 2003).
- The topology agrees with the rRNA tree of Woese.



# Radhey S. Gupta

Department of Biochemistry and Biomedical

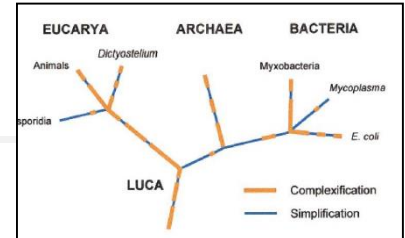
**Current research interest**

- Prof. S. Gupta currently focus entirely on comparative genomic studies to understand microbial phylogeny.



# Gupta's indel analysis, 1998

## Summary



- They concluded:
- Gram-positive bacteria arose first, and that both Archaea and Gram-negative bacteria arose from Gram-positive bacteria in response to antibiotic selection pressure.
- Gupta's phylogenetic tree for bacteria corroborates the standard 16S rRNA tree.
- However, the Woese group has presented convincing evidence from the 16S rRNA sequences to show that Archaea and Eukarya separated from a prokaryotic precursor and are not derivatives of the Bacteria as Gupta believes (pertinent conflict).

**Overall view: Gram-positive ==> Gram-negative**



# Gupta's indel analysis, 1998

## Bacterial main groups

The various main bacterial groups have branched off from a common ancestor in the following order (Gupta & Griffiths, 2002):

- **Low G+C Gram-positive ==> High G+C Gram-positive ==> Clostridium-Fusobacteria-Thermotoga ==> Deinococcus-Thermus-Green nonsulfur bacteria ==> (Gram-negative) Cyanobacteria ==> Spirochetes ==> Chlamydia-Cytophaga-Bacteroides-Green sulfur bacteria ==> Aquifex ==> Proteobacteria-1 (epsilon and delta) ==> Proteobacteria-2 (alpha) ==> Proteobacteria-3 (beta) and ==> Proteobacteria-4 (gamma).**

**Overall view: Low G+C Gram positive ==> High G+C Gram positive ==> Gram-negative**





# Signature approach for determining bacterial phylogeny

## Gupta's indel analysis

---

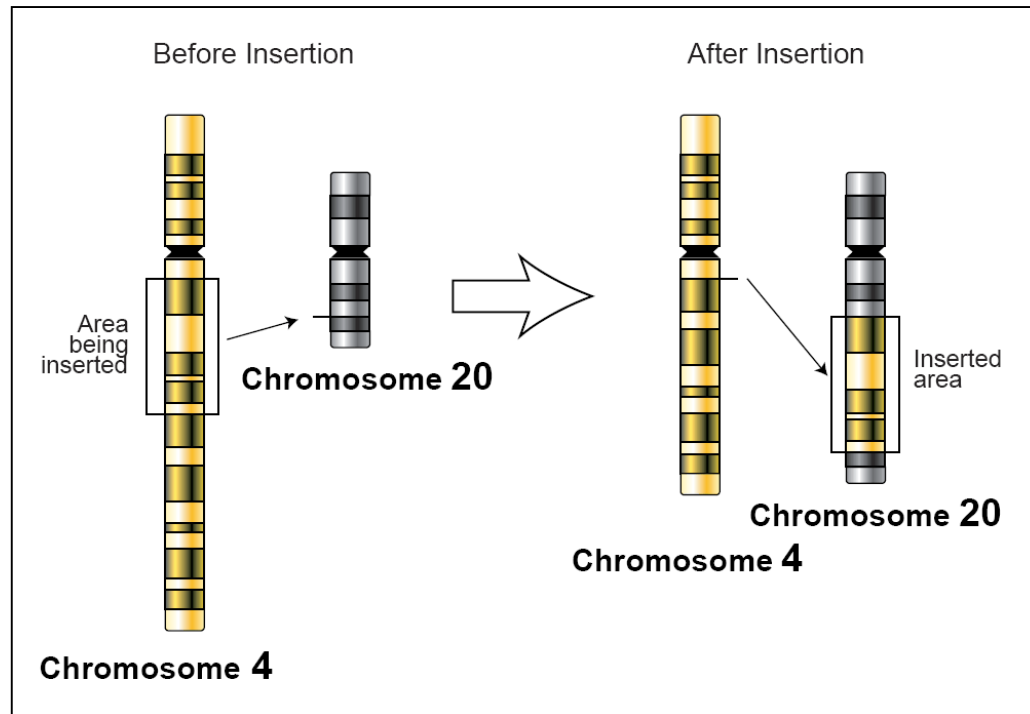
- Gupta *et al.*, 1998-2002, analyzed the completed, published sequences of many genomes, both bacterial and archaeal.
- The scheme was based on “signature” genomic insertions or deletions.
- Differences of ‘significance’ they called ‘indels’ (insertions/deletions).

**Indel:** An insertion or deletion in protein sequences that is flanked on both sides by conserved regions to ensure that it provides a reliable genetic/evolutionary markers; Based upon the presence or absence of the indel in outgroup species, it is possible to infer whether the indel represents an insert or a deletion in the gene/protein sequences.

# Gupta's indel analysis

## Chromosomal insertion

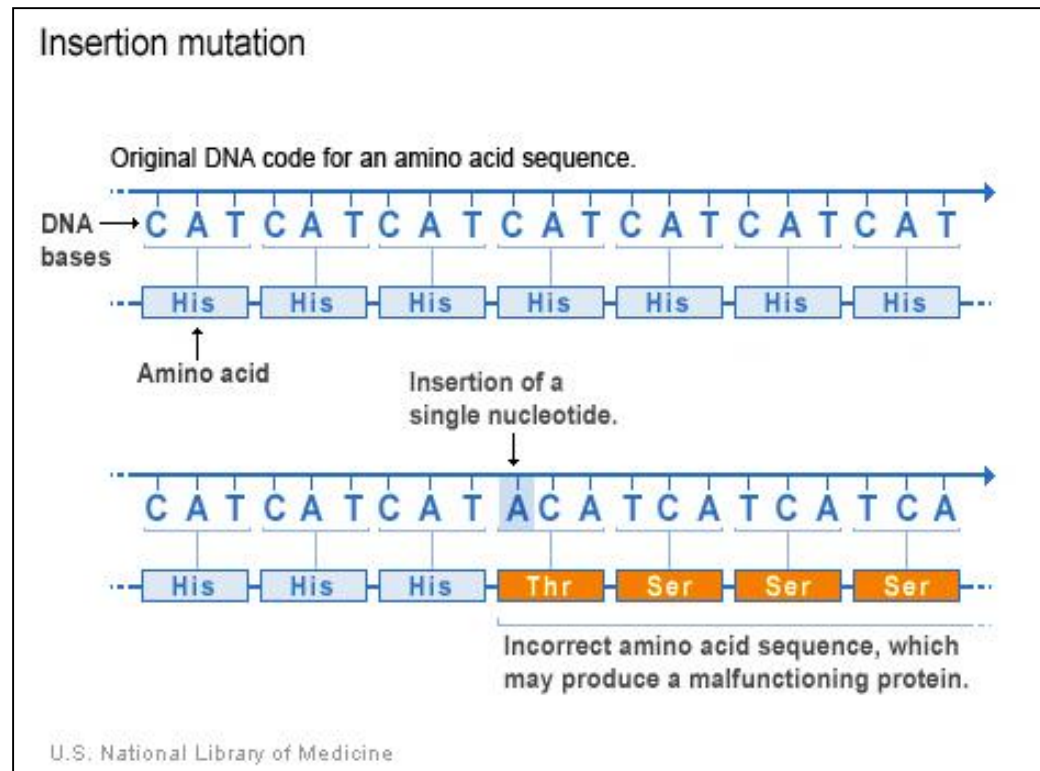
- In genetics, an **insertion** (also called an **insertion mutation**) is the **addition of one or more nucleotide base pairs into a DNA sequence.**



# Gupta's indel analysis

## Chromosomal insertion

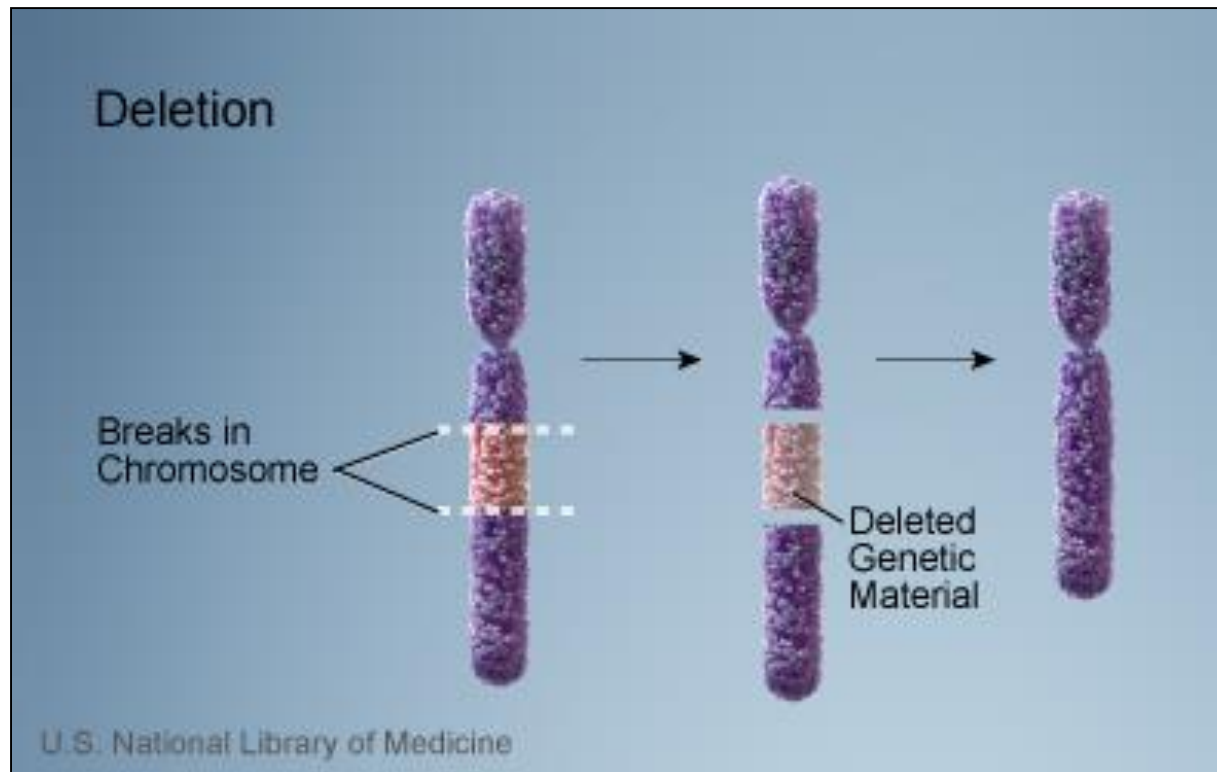
- In this example, **one nucleotide (adenine)** is added in the DNA code, **changing the amino acid sequence** that follows.



# Gupta's indel analysis

## Chromosomal deletion

- A deletion occurs when a chromosome breaks and some genetic material is lost.





# Signature approach for determining bacterial phylogeny

## Gupta's indel analysis

---

- Because the **smallest indel** in a protein sequence requires the **addition or deletion of 3 in-frame nucleotides** in a gene sequence, the **conserved indels** represent **Rare Genetic Changes** that are unlikely to occur by chance in different species.
- Hence, they provide useful molecular markers for evolutionary studies.



# Gupta's indel analysis

---

- Gupta's indel analysis is a very interesting **alternative** to "simple" sequence (**Woese**) analysis:
  - It produces an interesting, almost **linear tree topology**.
  - The branching order is not quite that of the **rRNA tree**, but the major groups seem to be consistent.
- Note that the **evolution of Archaea from Bacteria** or **Archaea-Bacteria separation** took place at a very early in prokaryotic evolution.



# Gupta's indel analysis

---

- Based upon conserved indels in protein sequences most of the prokaryotic phyla that were previously identified solely on the basis of branching in the 16S rRNA tree, can now be identified in clear molecular terms, enabling further genetic and biochemical studies on them."

## Gupta's indel analysis

## Sequenced bacterial genome

### Proteobacteria ( $\gamma$ -subdivision)

*Escherichia coli* K12  
*Escherichia coli* O157:H7  
*Escherichia coli* O157:H7 EDL933  
*Escherichia coli* CFT073  
*Buchnera* sp. APS  
*Buchnera aphidicola*  
*Buchnera aphidicola* Sg  
*Pasteurella mutocida*  
*Pseudomonas aeruginosa*  
*Pseudomonas putida* KT 2400  
*Pseudomonas syringae*  
*Vibrio cholerae*  
*Vibrio parahaemolyticus*  
*Vibrio vulnificus*  
*Xylella fastidiosa*  
*Xylella fastidiosa* Temecula  
*Haemophilus influenzae*  
*Yersinia pestis* C092  
*Yersinia pestis* KIM  
*Salmonella typhimurium* LT2  
*Salmonella typhi*  
*Xanthomonas citri*  
*Xanthomonas campestris*  
*Xylella fastidiosa*  
*Shewanella oneidensis*  
*Shigella flexneri* 2a  
*Wiggelsworthia brevipalpis*  
*Coxiella burnetii*

### Proteobacteria ( $\alpha$ -subdivision)

*Rickettsia prowazekii*  
*Caulobacter crescentus*  
*Mesorhizobium loti*  
*Bradyrhizobium japonicum*  
*Agrobacterium tumefaciens*-Dupont  
*Agrobacterium tumefaciens*-Cereon  
*Rickettsia conorii*  
*Sinorhizobium loti*  
*Brucella melitensis*  
*Brucella suis*  
*Rhodopseudomonas palustris*

### Proteobacteria ( $\beta$ -subdivision)

*Neisseria meningitidis* MC58  
*Neisseria meningitidis* Z2491  
*Ralstonia solanacearum*

### Proteobacteria ( $\delta$ , $\epsilon$ -subdivision)

*Helicobacter pylori* 26695  
*Helicobacter pylori* J99  
*Campylobacter jejuni*

### Aquifex

*Aquifex aeolicus*

### Chlamydia-CFBG

*Chlamydia trachomatis*  
*Chlamydia muridarum*  
*Chlamydomphila pneumoniae* CWL029  
*Chlamydomphila pneumoniae* J138  
*Chlamydomphila pneumoniae* AR39  
*Chlorobium tepidum*  
*Bacteroides thetaiotamicron*

### Spirochetes

*Borrelia burgdorferi*  
*Treponema pallidum*  
*Leptospira interrogans*

### Cyanobacteria

*Synechocystis* sp. PCC6803  
*Nostoc* sp. PCC7120  
*Thermosynechococcus elongatus*

### Clostridia-Thermotoga

*Thermotoga maritima*  
*Clostridium acetobutylicum*  
*Clostridium perfringens*  
*Clostridium tetani* E88  
*Fusobacterium nucleatum*  
*Thermoanaerobacter tengcongensis*

### Deinococcus-Thermus

*Deinococcus radiodurans*

### Actinobacteria

*Mycobacterium tuberculosis* H37  
*Mycobacterium tuberculosis* 1551  
*Mycobacterium leprae*  
*Corynebacterium glutamicum*  
*Corynebacterium efficiens*  
*Streptomyces coelicolor*  
*Bifidobacterium longum*  
*Tropheryma whippelii* Twist  
*Tropheryma whippelii* TW08/27

### Firmicutes

*Bacillus subtilis*  
*Bacillus halodurans*  
*Bacillus anthracis*  
*Oceanobacillus iheyensis*  
*Staphylococcus aureus* N315  
*Staphylococcus aureus* MW2  
*Staphylococcus epidermidis*  
*Staphylococcus aureus* Mu50  
*Streptococcus pyogenes*  
*Streptococcus pyogenes* S315  
*Streptococcus pyogenes* S8232  
*Streptococcus pneumoniae* R6  
*Streptococcus pneumoniae* TIGR4  
*Streptococcus agalactiae* 2603  
*Streptococcus agalactiae* NEM316  
*Streptococcus mutans* UA159  
*Mycoplasma genitalium*  
*Mycoplasma pneumoniae*  
*Mycoplasma pulmonis*  
*Mycoplasma penetrans*  
*Ureaplasma urealyticum*  
*Lactococcus lactis*  
*Lactobacillus plantarum*  
*Listeria innocua*  
*Listeria monocytogenes*

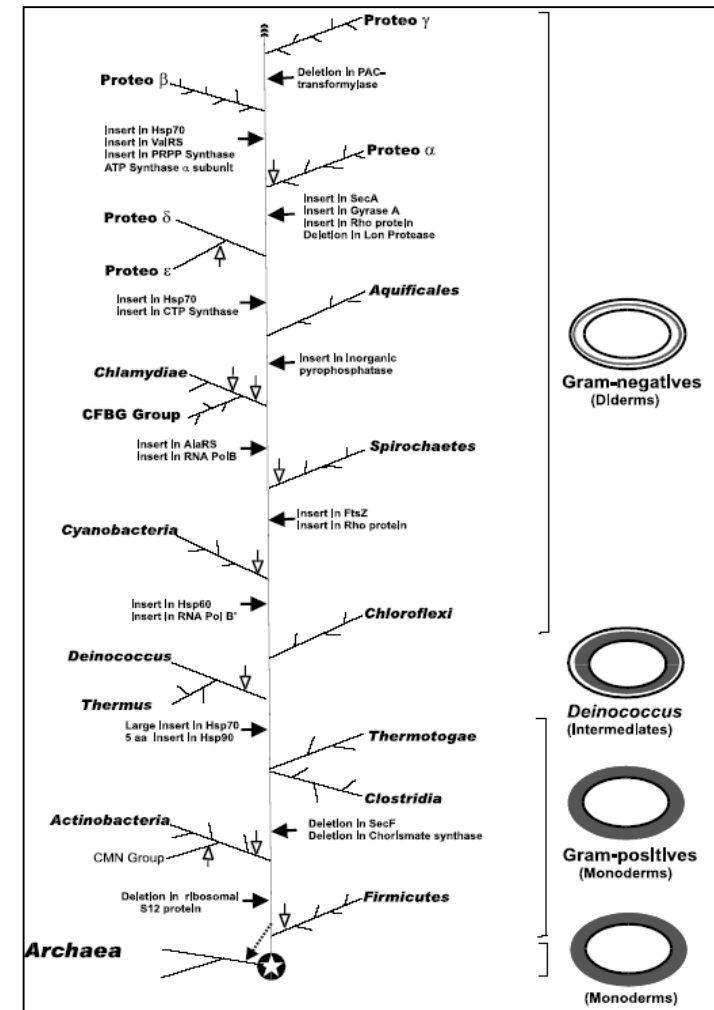




# Evolutionary model based on signature sequences indicating the branching order of the main bacterial groups

The predictions of the indel model are strongly supported by analyses of the genome sequence data thus strongly supporting this model

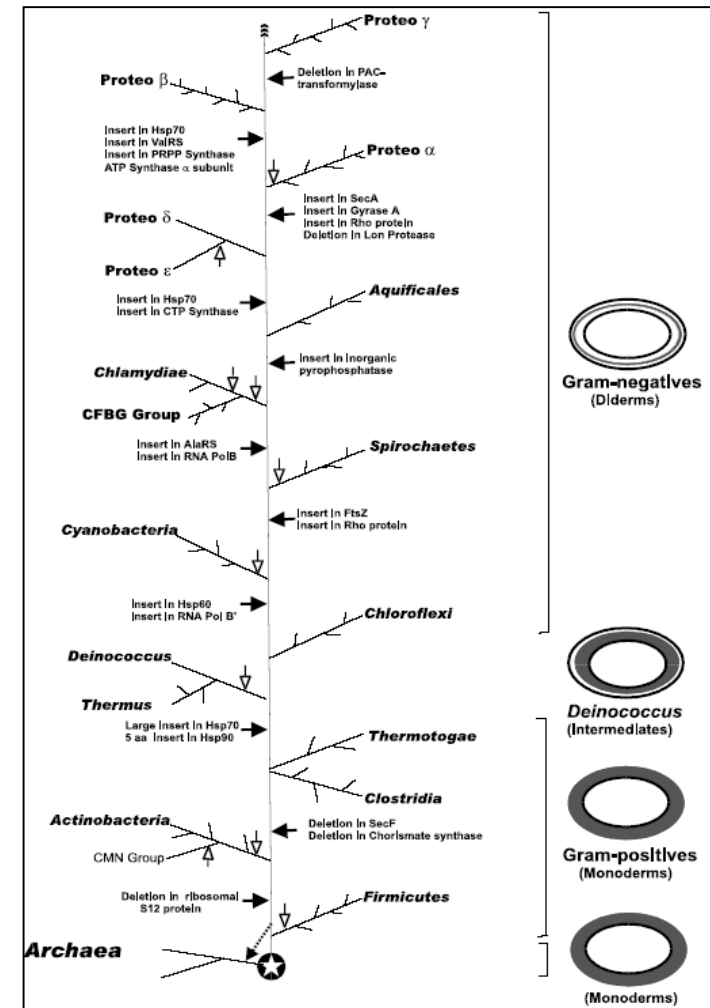
- **The filled arrows** depict the stages at which the **different main-line signatures** indicated in **previous slide** have been introduced.
- These signatures are expected to be present in **bacterial groups that have diverged at a later time** (i.e., those lying above the indicated insertion points), but they should be **absent** in the **earlier branching groups**.



# Evolutionary model based on signature sequences indicating the branching order of the main bacterial groups

The predictions of the indel model are strongly supported by analyses of the genome sequence data thus strongly supporting this model

- **The unfilled arrows** denote the positions of many group-specific signatures (not shown here).
- **The dotted arrow** at the bottom indicates the possible derivation of Archaea from Gram-positive bacteria.
- The cell structures of different groups of bacteria are indicated on the right.



**Predicted versus  
observed  
distribution of  
indels in 100  
bacterial  
genomes.**

Gupta, 2005

Protein	Signature Description	No. Genomes with Protein	No. Genomes with Indels Expected/ Found	No. Genomes Lacking the Indel Expected/ Found	Exceptions Observed
Rib. S12 protein	13 aa <i>Firmicute</i> insert	100	25/25	75/75	0
Hsp70/DnaK	21–23 aa G+/G- insert	100	60/60	40/40	0
Hsp90	5 aa G+/G- insert	52	11/11	41/41	0
Chorismate Synthase	15–17 aa deletion after <i>Actinobacteria</i>	89	29/29	60/60	0 <sup>a</sup>
SecF protein	3–4 aa deletion after <i>Actinobacteria</i>	81	15/17	56/54	2 <sup>b</sup>
Hsp60/GroEL	1 aa insert after <i>Deinococcus</i>	98	65/66	33/32	1 <sup>c</sup>
RNA Polymerase $\beta^L$ - subunit	>150 aa after <i>Deinococcus</i>	100	59/59	41/41	0
FtsZ protein	1 aa insert after cyanobacteria	91	51/51	40/40	0
Rho p1	ore spirochetes	83	56/57	27/26	1 <sup>d</sup>
Ala-tR RADHEY S. GUPTA	chets	100	53/53	47/47	0
RNA Polymerase	20–120 aa insert after	100	53/53	47/47	0
$\beta$ - subunit	spirochetes				
Inorganic pyro- phosphatase	2 aa insert common to <i>Aquifex</i> and proteo.	71	45/45	26/26	0
Hsp70/DnaK	2 aa Proteo insert	100	45/45	55/55	0
CTP Synthetase	10 aa Proteo Indel	92	45/45	47/47	0
Lon protease	1 aa deletion in $\alpha\beta\gamma$ - proteobacteria	70	41/43	29/27	2 <sup>e</sup>
Rho Protein	3 aa $\alpha\beta\gamma$ -Proteo indel	83	42/43	41/40	1 <sup>f</sup>
DNA Gyrase	26–34 aa insert in $\alpha\beta\gamma$ - proteobacteria	100	42/42	58/58	0
A subunit					
SecA protein	7 aa $\alpha\beta\gamma$ -Proteo indel	100	42/42	58/58	0
HSP70/DnaK	4 aa $\beta\gamma$ -Proteo insert	100	31/34	69/66	3 <sup>g</sup>
ATP Synthase	11 aa insert in $\beta\gamma$ - proteobacteria	92	31/32	61/60	1 <sup>h</sup>
$\alpha$ -subunit					
Val-tRNA Synth.	37 aa $\beta\gamma$ -Proteo insert	100	31/31	69/69	0
PRPP synthetase	1 aa $\beta\gamma$ -Proteo insert	94	31/31	63/63	0
PAC- formyltransferase	2 aa $\gamma$ -Proteo deletion	83	55/55	28/28	0



# Cavalier-Smith megaclassification, 2002

---

## Regnum concept

Noun. **regnum** (plural **regnums** or **regna**) (biology, taxonomy) A rank in the classification of organisms, also known as **kingdom**.

# Tomas Cavalier-Smith

## Professor of Evolutionary Biology

- Professorial Fellow (born 21 October 1942), is a Professor of Evolutionary Biology in the Department of Zoology, at the University of Oxford.
- He was presented with the International Prize for Biology (a prize of 10 million yen) in 2004.
- He worked out on cell and genome evolution:
  1. large scale phylogeny and the tree of life;
  2. origins of eukaryotes, animals, plants.

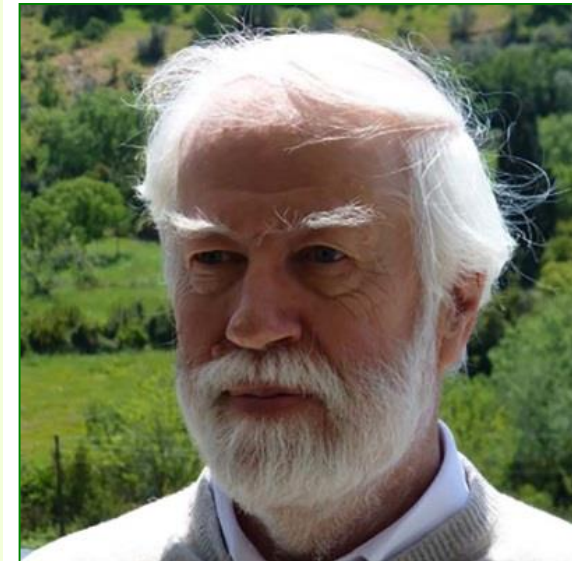


He also won the 2007 Frink Medal of the Zoological Society of London.

# Tomas Cavalier-Smith

## Professor of Evolutionary Biology

- He states I especially like synthesizing very diverse information into simple explanations and attacking wrong ideas.
- My laboratory focuses on the **evolution**, **ecology**, and **biogeography** of amoeboid and flagellate free-living Protozoa using cell culturing, **DNA sequencing** (genes and genomes), **phylogenetic analysis**, **bioinformatics**, and **light and electron microscopy**.
- But my **theoretical** interests are much **wider**, ranging from the **origin of cells**, and their diversification to make the major bacterial and eukaryotic groups.







# Tomas Cavalier-Smith

## Professor of Evolutionary Biology

---

- Prof. Cavalier-Smith of Oxford University has produced a large body of work which is well regarded. Still, he is controversial in a way that is a bit difficult to describe.
- The issue may be one of writing style.
- Cavalier-Smith has a tendency to make pronouncements where others would use declarative sentences, to use declarative sentences where others would express an opinion, and to express opinions where angels would fear to tread.
- In addition, he can sound arrogant, reactionary, and even perverse. On the other [hand], he has a long history of being right when everyone else was wrong.





# Tomas Cavalier-Smith

## Professor of Evolutionary Biology

---

- This makes for very long, very complex papers and causes all manner of dark murmuring, tearing of hair, and gnashing of teeth among those tasked with trying to explain his views of early life.

# Cavalier-Smith megaclassification

## Uprooting and replanting the tree of life

### Two Empires: Prokaryota and Eukaryota and six kingdoms

---

- There is the long-winded, vocabulary-rich analysis of Cavalier-Smith (2002), which is also very interesting.
- Cavalier-Smith basically concludes that double-membraned Gram-negative bacteria (he calls them "Negibacteria") lie near the root of the bacterial tree (3700 Mya), and that the Archaea and Eucarya are relatively recent (850 Mya) emergents from a line that also gave rise to the modern Gram-positive bacteria and actinobacteria.



# Cavalier-Smith megaclassification

## Origin and evolution of life

---

- According to **Woese classification**, there are **three branches** to the tree of life:
  1. **Bacteria**,
  2. **Eukaryotes**, and
  3. **Archaeobacteria**.
- **Bacteria** evolved **3500-3850 million years ago**.
- **Archaeobacteria** were also believed to be **ancient** because of their unusual cell structure.
- But **Prof. Cavalier-Smith** argues.

# Summary of the sequence from the two-kingdom system up to Cavalier-Smith's six-kingdom system

## From phenetic towards a phylogenetic Classification

Linnaeus 1735	Haeckel 1866	Chatoon 1925	Copeland 1938	Whittaker 1969	Woese <i>et al.</i> 1977	Woese <i>et al.</i> 1990 (Revised)	Cavalier-Smith 1993	Cavalier-Smith 1998 (Revised)
2 kingdoms	3 kingdoms	2 empires	4 kingdoms	5 kingdoms	6 kingdoms	3 domains	8 kingdoms	6 kingdoms
(not treated)	Protista	Prokaryota	Monera	Monera	Eubacteria	Bacteria	Eubacteria	Bacteria
					Archaeobacteria	Archaea	Archaeobacteria	
		Eukaryota	Protoctista	Protista	Protista	Eukarya	Archezoa	Protozoa
							Protozoa	
							Chromista	Chromista
							Plantae	Plantae
Vegetabilia	Plantae		Plantae	Plantae	Plantae		Fungi	Fungi
Animalia	Animalia		Animalia	Animalia	Animalia		Animalia	Animalia

# Cavalier-Smith megaclassification

## Cavalier-Smith's six-kingdom schema

### Two Empires: Prokaryota and Eukaryota and six kingdoms

---

- In 1981, Cavalier-Smith's proposed the division of all organisms into **eight kingdoms**.
- Bacteria, Eufungi, Ciliofungi, Animalia, Biliphyta, Viridiplantae, Cryptophyta, and Euglenozoa.
- By 1998, Cavalier-Smith had reduced the **total number of kingdoms from eight to six**:
- **Animalia, Protozoa, Fungi, Plantae (including red and green algae), Chromista and Bacteria.**
- In 2015, Cavalier-Smith and his collaborators once again revised the classification. In this scheme they reintroduced the **division of prokaryotes into two kingdoms**:
  1. **Bacteria (=Eubacteria) and**
  2. **Archaea (=Archebacteria).**



# Cavalier-Smith megaclassification

## Origin and evolution of life

---

- His research shows that **archaebacteria** and **eukaryotes** should be placed together in one big group called **neomura**, which means **new walls**.
- These organisms have a **common ancestor** that evolved **850 million years ago** to contain a substance called **glycoprotein in its membrane**, which gave it greater **fluidity than the rigid cell walls of ordinary bacteria**.
- The unusual cell structure of **archaebacteria** can be explained as relatively **recent adaptations to life in extreme environments** such as **boiling water and hot acid**.



# Cavalier-Smith megaclassification

## Origin and evolution of life

---

- The **neomuran ancestor** has been identified as an **actinobacterium (G+ve)**, which is related to the **bacteria** that cause **tuberculosis and leprosy**.
- It is intriguing to think that **we are more closely related to tuberculosis bacteria than they are to *E. coli* (G-ve)**, says Prof. Cavalier-Smith.



# Cavalier-Smith megaclassification

## Origin and evolution of life

---

- All eukaryotes have a complex endoskeleton (the cytoskeleton) of microtubules and actin filaments that use attached molecular motors to mediate chromosome segregation and cell division, respectively.
- By contrast, bacteria have an exoskeleton (cell wall) important for DNA segregation and cell division.
- There has been much discussion of how these and other profound differences between bacteria and eukaryotes have arisen.





# Cavalier-Smith megaclassification

## Bacterial origins of Life through two big bangs

---

- For most of the history of life, immensely long periods of relative stasis have followed **two explosive radiations or biological big bangs**, each stimulated by revolutionary innovations in cell biology:
  1. The origin about **3700 My ago** of the **first eubacterial cell** with **peptidoglycan walls** and **photosynthesis**(Cavalier-Smith,2001).
  2. The origin about **850 My ago** of the **ancestral neomuran cell**, when **N-linked glycoproteins** replaced **peptidoglycan** and the **pre-eukaryote neomurans** evolved phagotrophy, internal skeletons and the endomembrane system.

Phagotrophy in the origins of photosynthesis in eukaryotes.



# Cavalier-Smith megaclassification

## Neomuran revolution and bacterial origins of Life at two-stage process

---

- The **ancestors of eukaryotes**, the **stem Neomura**, are shared with **archaebacteria** and evolved during the **neomuran revolution**, in which:
  1. N-linked glycoproteins replaced murein peptidoglycan and 18 other suites of characters changed radically through adaptation of an ancestral actinobacterium to thermophily.
  2. In the next phase, **archaebacteria** and **eukaryotes** diverged dramatically.

# Cavalier-Smith megaclassification

## Neomuran revolution and bacterial origins of Life at two-stage process

---

- Archaeobacteria retained the wall and therefore their general bacterial cell and genetic organization, but became adapted to even hotter and more acidic environments by substituting prenyl ether lipids for the ancestral acyl esters and making new acid resistant flagellar shafts.

# Cavalier-Smith megaclassification

## Neomuran revolution and bacterial origins of Life at two-stage process

- At the same time, **eukaryotes** converted the **glycoprotein wall** into a **flexible surface coat** and evolved rudimentary phagotrophy for the first time in the history of life.
- This triggered a massive reorganization of their cell and chromosomal structure and enabled an alpha-proteobacterium to be enslaved and converted into a protomitochondrion to form the **first aerobic eukaryote and protozoan**, around 850 My ago.
- Substantially later, a **cyanobacterium** (photosynthetic gram negative bacterium) was enslaved by the common ancestor of the **plant kingdom** to form the **first chloroplast**.



# Cavalier-Smith Bacterial megaclassification

Two Empires: Prokaryota and Eukaryota and six kingdoms

**Negibacteria as a root of the universal tree**

---

- Prokaryotes constitute a single kingdom, **Bacteria**.
- Bacteria is divided into two new subkingdoms:
  1. **Negibacteria**(G-ve bacteria), with two bounding membranes.
  2. **Unibacteria**(G+ve bacteria), with one bounding membranes comprising the new phyla **Archaeobacteria** and **Posibacteria**.
- Other new bacterial taxa are established in a revised higher-level classification that recognizes only **eight phyla** and **29 classes**.

# Revised classification of kingdom Bacteria and its eight phyla (divisions)

Taxon	Etymology	Description	Type
<b>Subkingdom 1. NEGIBACTERIA*</b> (Cavalier-Smith, 1987b) subregnum nov.	Contraction from L. <i>negativus</i> negative, since most stain Gram-negative	Cell bounded by two concentric lipid bilayers, the cytoplasmic membrane and an outer membrane bearing porins; ancestrally with peptidoglycan and lipoprotein between the membranes; SRP lacks helices 1-4 and 19p; protein secretion predominantly post-translational	Order Enterobacterales
<b>Infrakingdom 1. Eobacteria</b> (Cavalier-Smith, 1992a) infraregnum nov.	Gr. <i>eos</i> dawn, because the absence of lipopolysaccharide suggests they may be the earliest negibacteria	No lipopolysaccharide or sphingolipids; peptidoglycan with ornithine, not diaminopimelic acid; usually thermophilic; flagella absent; gas vesicles absent	Order Chloroflexales
Division 1. Eobacteria (Cavalier-Smith, 1992a) divisio nov.	As for infrakingdom above	As for infrakingdom above	Order Chloroflexales
Class 1. Chlorobacteria (Cavalier-Smith, 1992a) classis nov.	Gr. <i>chloros</i> yellow green, from the colour of the photosynthetic species	Filamentous green bacteria, with bacteriochlorophyll <i>a</i> and usually chlorosomes, gliding green non-sulphur photosynthetic bacteria, with pheophytin quinone type-2 reaction centres, with or without chlorosomes ( <i>Chloroflexus</i> , <i>Heliothrix</i> , <i>Roseiflexus</i> , <i>Oscillochloris</i> ), and their colourless relatives, e.g. <i>Thermomicrobium</i> , <i>Herpetosiphon</i> , <i>Thermotoga</i> , <i>Dhalococcus</i> (a halorespirer)	Order Chloroflexales
Class 2. Hadobacteria (Cavalier-Smith, 1992a; emend. 1998) classis nov.	Gr. <i>hadēs</i> hell, because they can resist extremes of heat or radiation	Heterotrophic thermophiles or highly radiation-resistant bacteria with thick murein layer; with semi-crystalline S-layer, e.g. <i>Deltaproteus</i> , <i>Thermus</i> , <i>Methanothermobacter</i> ; more closely related to each other on rRNA trees than to Chlorobacteria	Order Thermales
<b>Infrakingdom 2. Glycobacteria*</b> (Cavalier-Smith, 1998) infraregnum nov.	Gr. <i>glykys</i> sweet, because they have surface lipopolysaccharide	Outer membrane with lipopolysaccharide or lipooligosaccharide; peptidoglycan with diaminopimelic acid or ornithine; gas vesicles widespread	Order Enterobacterales
Division 1. Cyanobacteria (Stanier 1974) nom. rev. (ex Stanier & Cohen-Bazire, 1977 as class)	Gr. <i>kyanos</i> blue-green, because of their common colour and the traditional name Cyanophyceae or blue-green algae	Oxygenic photosynthesis with chlorophyll <i>a</i> ; flagella absent; often glide; ancestrally with phycobilisomes, sometimes lost	Order Chroococcales
Subdivision 1. Gloeobacteria subdivisio nov.	From <i>Gloeobacter</i> , the only known genus	Without thylakoids	Order Gloeobacterales
Class 1. Gloeobacteria (Cavalier-Smith, 1998) classis nov.	As for subdivision above	As for subdivision above	Order Gloeobacterales
Order 1. Gloeobacterales ord. nov.	As for subdivision above	Having phycobilisomes but no thylakoids	Genus <i>Gloeobacter</i>
Subdivision 2. Phycobacteria (Cavalier-Smith, 1998) subdivisio nov.	Gr. <i>phukos</i> seaweed, because all the traditional blue-green algae and the prochlorophytes are included	With thylakoids; gliding motility by slime secretion; classical Cyanophyceae and prochlorophytes. The five traditional cyanobacterial orders, already valid under the Code of Botanical Nomenclature, are here also formally validated under the Bacteriological (= Prokaryotic) Code	Order Chroococcales
Class 1. Chroobacteria classis nov.	From the genus <i>Chroococcus</i>	Unicellular, palmelloid, colonial or with filaments lacking heterocysts	Order Chroococcales
Order 1. Chroococcales ord. nov.	As for class above	Unicellular and colonial (non-filamentous) cyanobacteria (with phycobilisomes and prochlorophytes with chlorophyll <i>b</i> instead	Genus <i>Chroococcus</i>

# Summarized Table:

Two Empires: Prokaryota and Eukaryota and six kingdoms

## Regnum (Kingdom) Bacteria

### 1. Subkingdom(subregnum): **Negibacteria** (G-ve bacteria)

1. Infrakingdom(subregnum) Eobacteria

2. Infrakingdom(subregnum) Glycobacteria

Superdivision Exoflagellate

Division 1. Planctobacteria

Division 2. Proteobacteria(most G-ve phytobacteria)

### 2. Subkingdom(subregnum): **Unibacteria** (G+ve bacteria)

Division 1. Posibacteria

Subdivision 1. Endobacteria

Class 1. Togobacteria

Class 2. Teichobacteria e.g. Bacillales

Class 3. Mollicutes

Subdivision 2. Actinobacteria

Class 1. Arthrobacteria

Class 2. Arabobacteria

Class 3. Streptomyces e.g. Coryneforms

Division 2. Archebacteria



# Cavalier-Smith megaclassification

## Characters used in megaclassification scheme

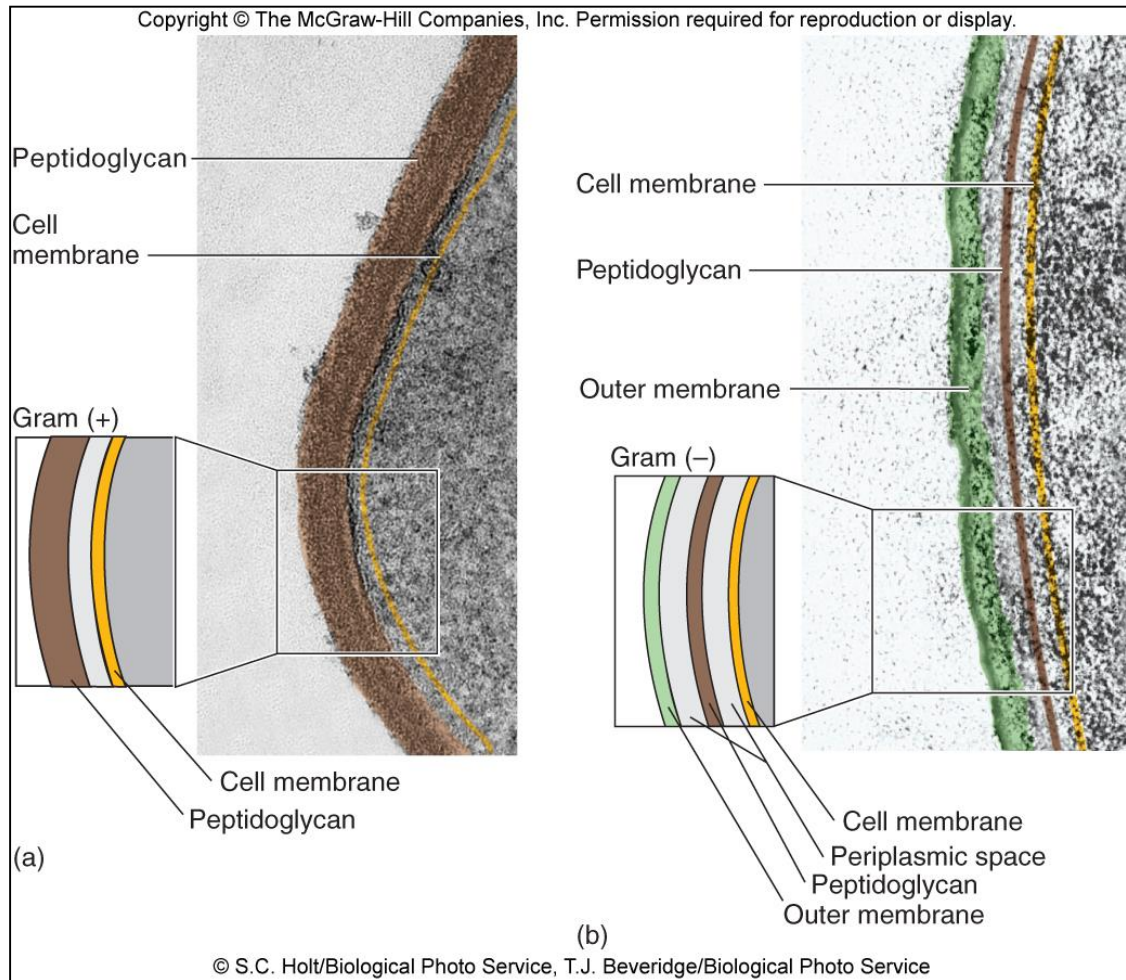
---

- The classification takes into account many phenotypic characteristics, and is not sequence-based.
- These include:
  1. Morphological,
  2. Palaeontological(the study of fossils), and
  3. Molecular data.
- These are integrated into a unified picture of large-scale bacterial cell evolution despite occasional lateral gene transfers.



# Two main bacterial cell wall

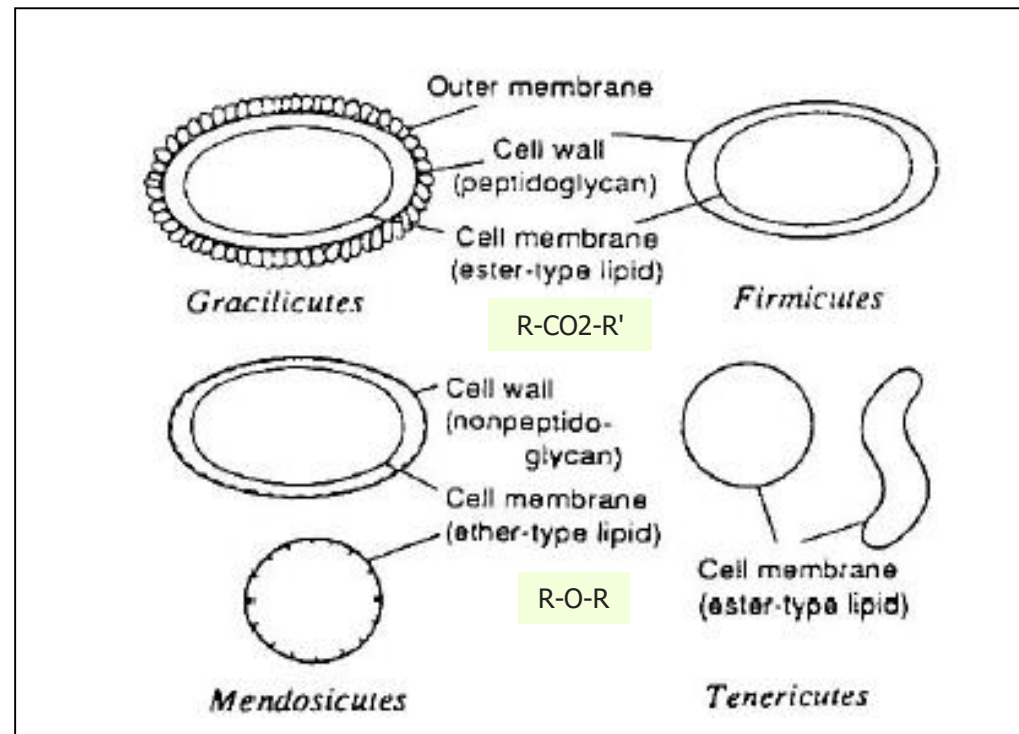
## Gracilicutes and Firmicutes



# Four main bacterial cell wall

## Gracilicutes, Firmicutes, Tendericutes, Mendosicutes

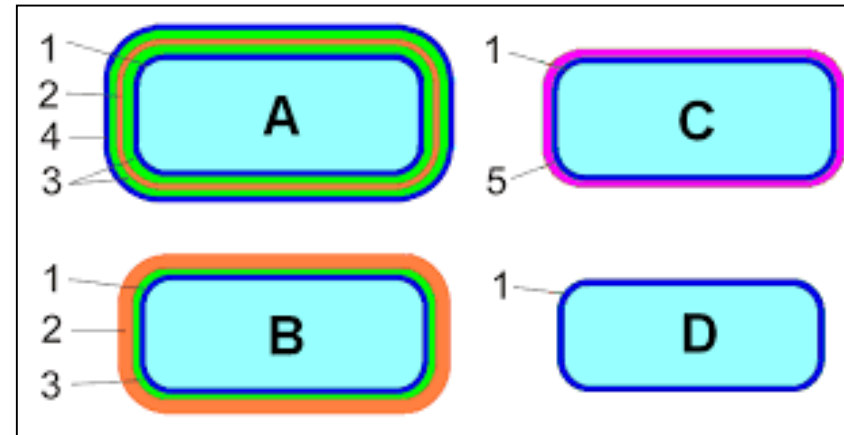
- Cellular envelop in **Gram negative bacteria** are surrounded by two layers: an external and an internal membrane (diderm) while **Gram positive bacteria** have one membrane (monoderm).



# Four main bacterial cell wall

**Gracilicutes, Firmicutes, Tenericutes, Mendosicutes**

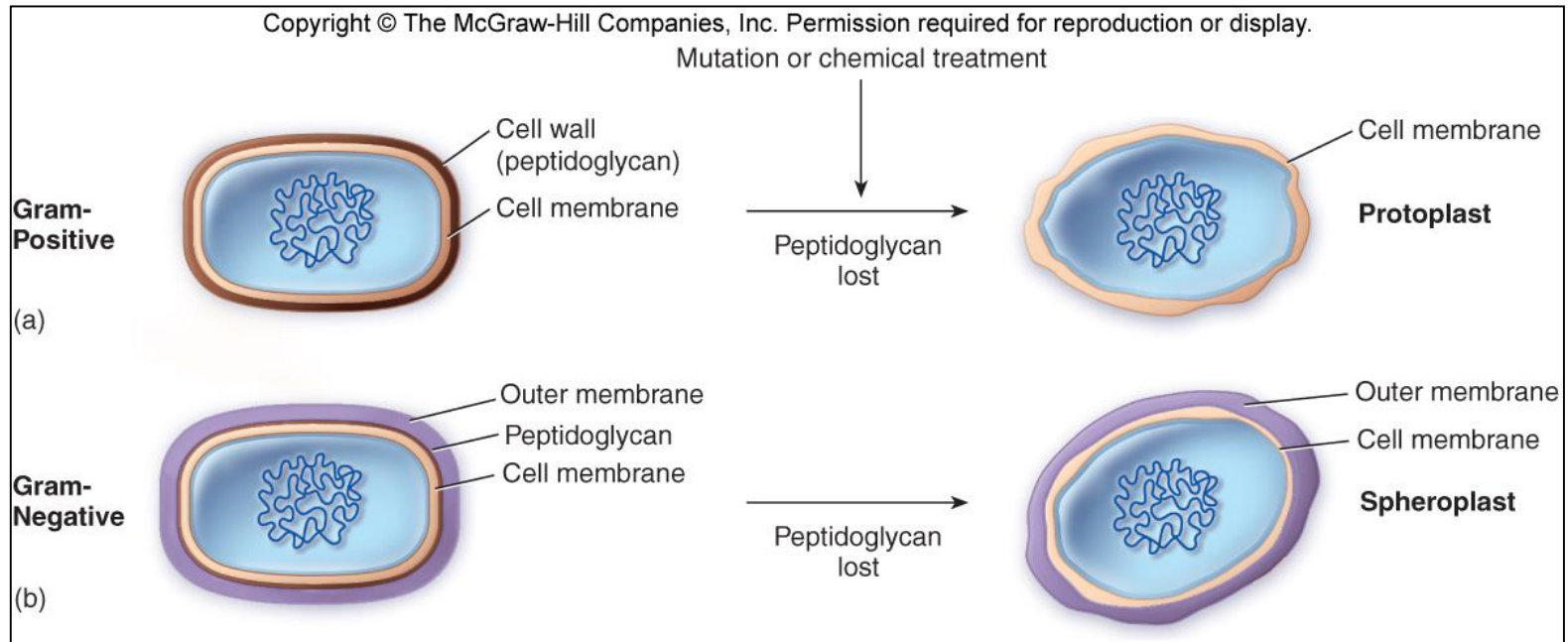
1. **Gracilicutes** (Gram negative);
2. **Firmicutes** (Gram positive);
3. **Tenericutes** (lack a cell wall, more soft. E.g. **phytoplasma**);
4. **Mendosicutes** (with no peptidoglycan in cell wall. E.g. **archaea**).



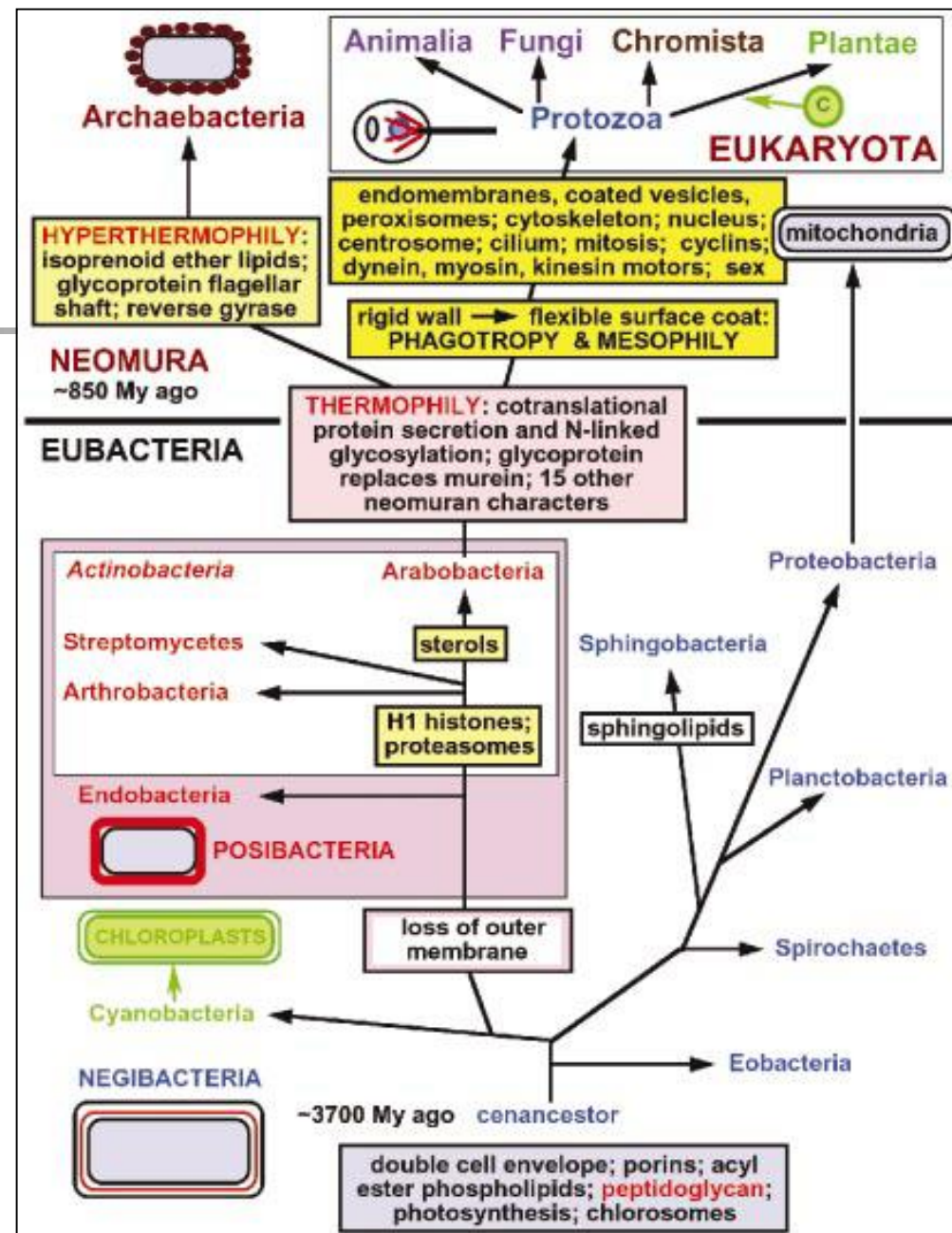
# Type of L-form bacteria

**Class I: spheroplasts (with outer membrane can revert)**

**Class II protoplasts (without outer membrane cannot revert)**



The bacterial origins  
of eukaryotes as a  
two-stage process.  
This paper very  
strongly supports  
actinobacterial origin  
of neomura.

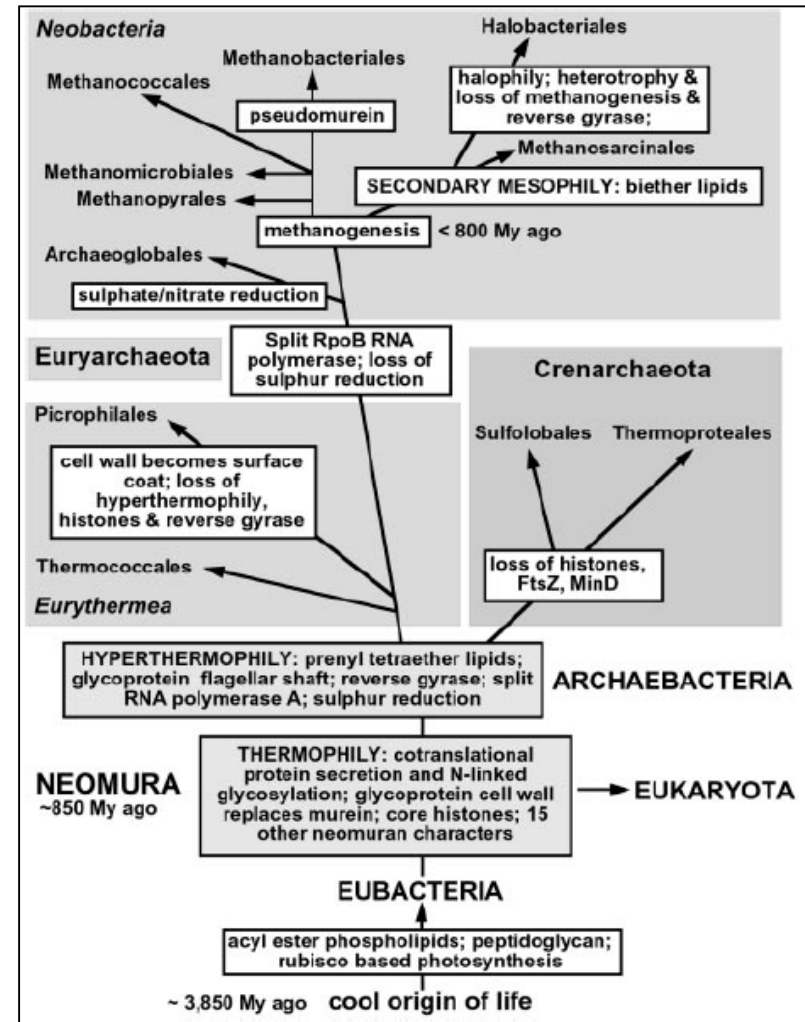




# The bacterial origins of Archaeobacteria as a two-stage process

- Archaeobacteria originated by two successive revolutions in cell biology:
  1. A neomuran phase shared with their eukaryote Sisters.
  2. Followed shortly by a uniquely archaeobacterial one.
- Bacterial DNA does not have histones.
- Histone proteins are among the most highly conserved proteins in eukaryotes, emphasizing the important role they play in DNA winding and gene regulation.

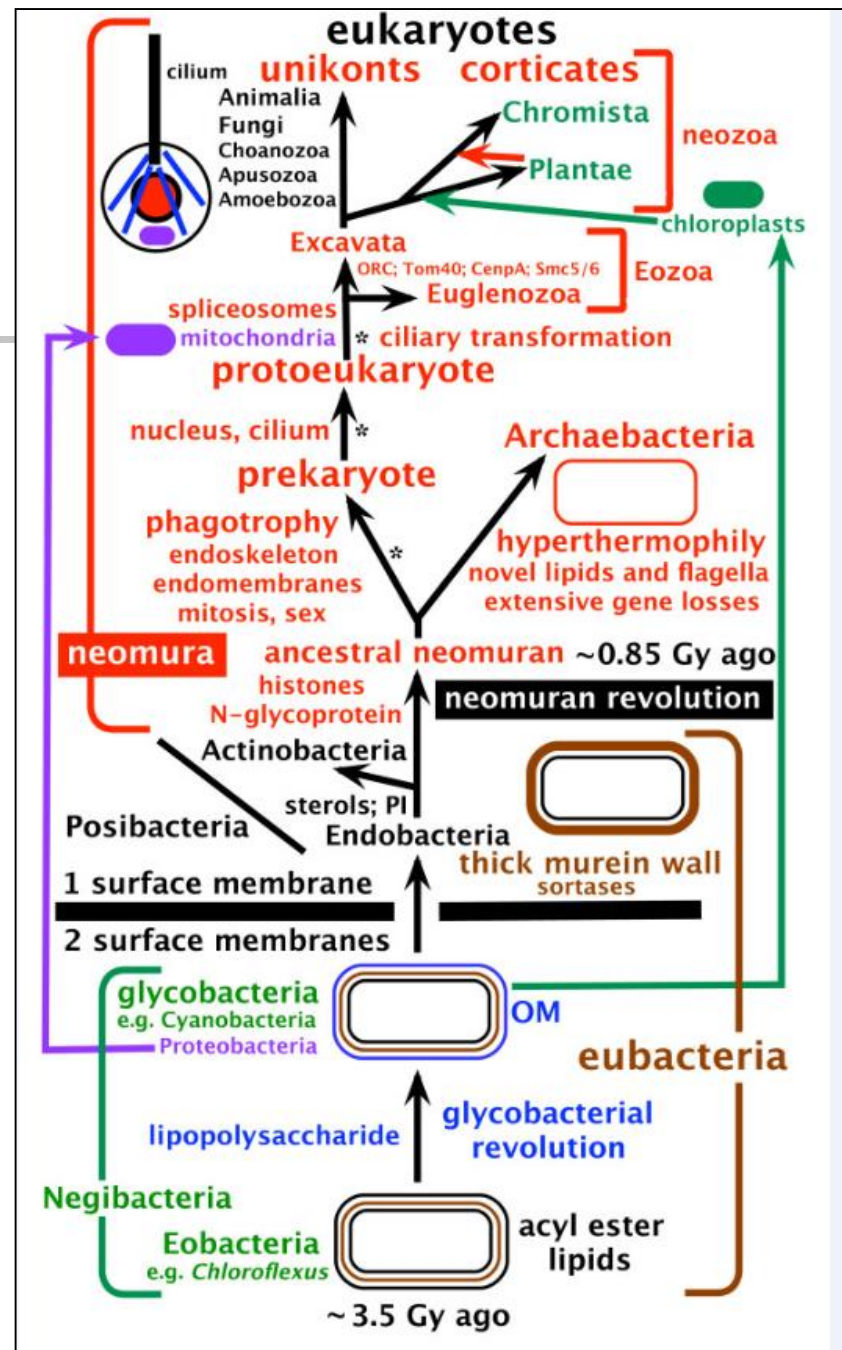
Cavalier-Smith, 2001



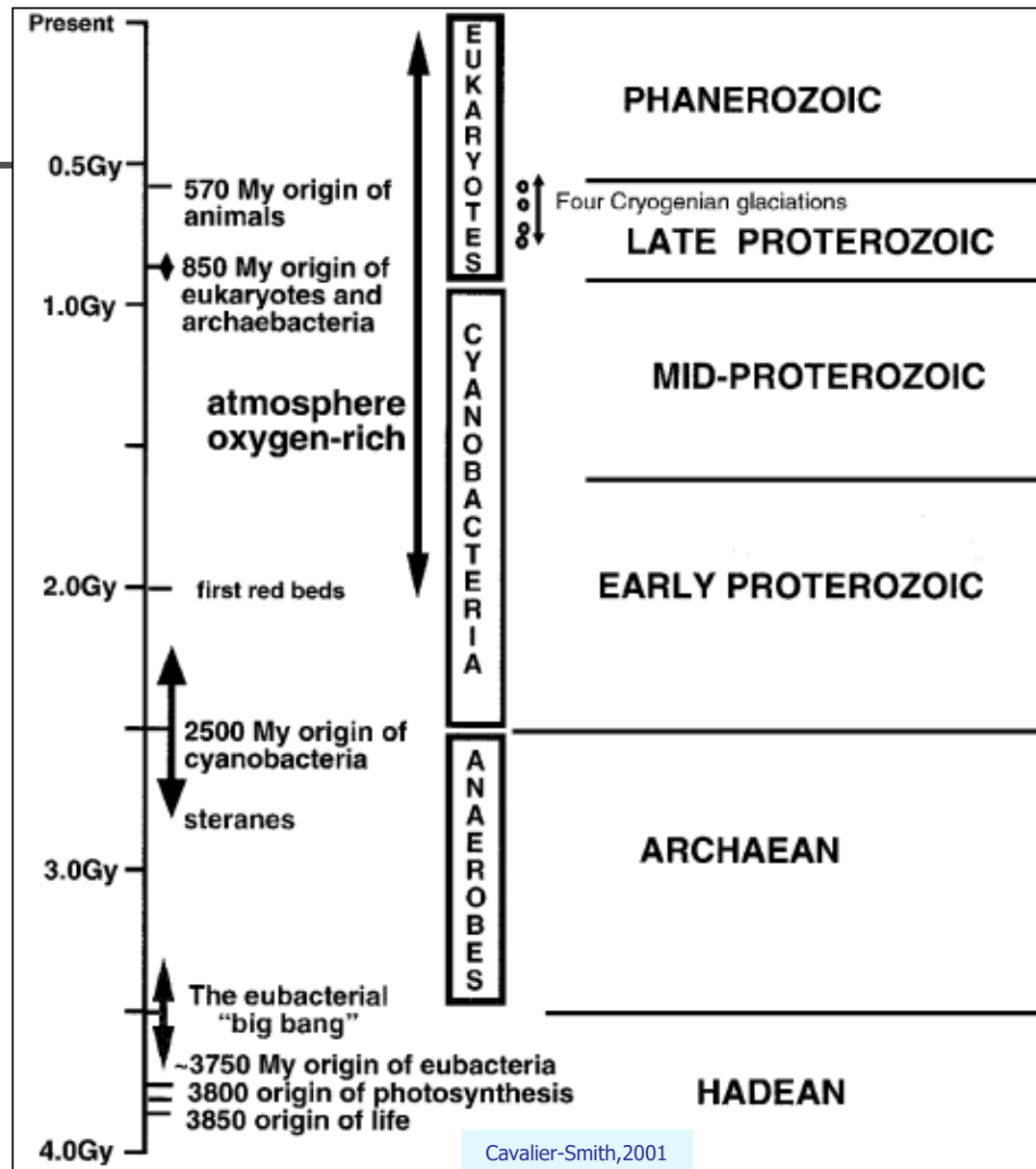
**Tree of life and major steps in cell evolution after Cavalier-Smith, ca 2010, before his 2015 revision.**

Cavalier-Smith T. 2014. *The neomuran revolution and phagotrophic* (not comparable) origin of eukaryotes in the light of intracellular coevolution and a revised tree of life. In: *The origin and evolution of eukaryotes*. Keeling PJ, Koonin EV, editors. Cold Spring Harb Perspect Biol.

Wikipedia, 2017



# Major features of the fossil record interpreted in the light of cell and molecular biology







# Is Cavalier scheme inconsistent with Gupta's?

- The Cavalier scheme:
- This scheme is not totally inconsistent with Gupta's if you change every "insertion" to a "deletion" and vice-versa and run the evolution from right to left instead of left to right.

**Gupta's view:** Gram-positive → Gram-negative

**Cavalier's view:** Gram-positive ← Gram-negative

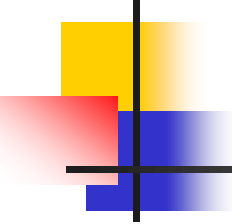
- But many evidences(next slide)indicate it is not correct to state that simply by changing the various inserts to a deletion, the wo schemes become very similar.



# Is Cavalier scheme inconsistent with Gupta's?

---

1. In addition to whether the first cell was Gram-positive or Gram-negative, there are important differences between evolutionary schemes of Cavalier-Smith and Gupta.
2. Cavalier-Smith does not place the root within the Gram-negative in the Gammaproteobacteria, which would be required if the branching order of various groups was simply reversed in the two case.
3. Another important difference between Cavalier-Smith's scheme and Gupta is that according to Cavalier-Smith the Archaea have evolved very recently, which is again not supported by Gupta scheme.



# Arthur L. Koch, 2003 argues

## The first cells: Gram-positive or Gram-negative?

---

- At some point in the evolution of life, the **domain Bacteria** arose from **prokaryotic progenitors** (an originator of a line of descent/ a direct ancestor).
- The cell that gave rise to the **first bacterium** has been given the name (among several other names) '**last universal ancestor (LUA)**'.
- This cell had an **extensive, well-developed suite of biochemical strategies** that increased its ability to grow.



# Arthur L. Koch, 2003 argues

## The first cells: Gram-positive or Gram-negative?

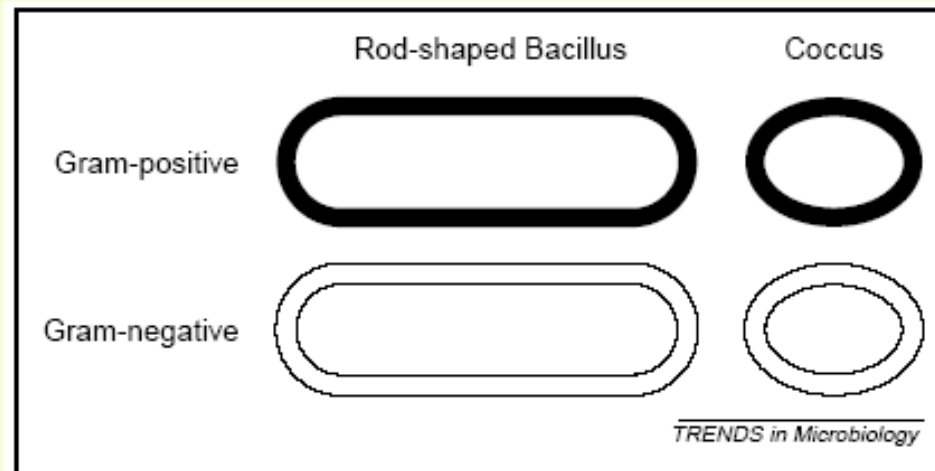
---

- The **first bacterium** is thought to have acquired:
  1. a covering, called a **sacculus** (a small sac), or
  2. **exoskeleton**, that made it stress-resistant.
- This protected it from rupturing as a result of turgor pressure stress arising from the success of its metabolic abilities.
  1. So what were the properties of this cell's wall?
  2. Was it Gram-positive or Gram-negative?
  3. And was it a coccus or a rod?

# Origin of first bacterium from first cell

## Arthur L. Koch, 2003

- Four possibilities for the wall of the first bacterium.
- These four types represent a majority of organisms.
- There are other shapes (curved, spiral and tapered) but these are probably less likely than the initial bacterial form.





# Origin of first bacterium from first cell

## Koch conclusions

---

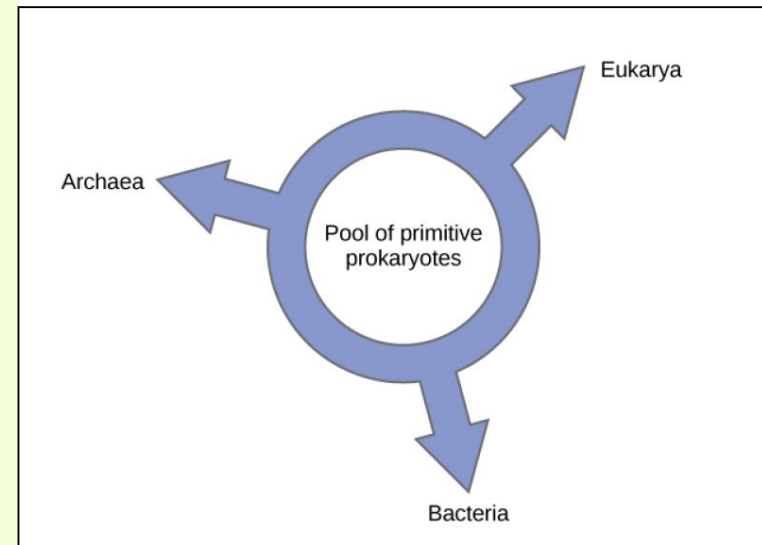
- The **coccus** is the **simplest of possible cell shapes** and the **growth of cocci** is that of **rod shaped organisms**.
- According ideas of **Woese, Seifert and Fox, Vicente's group** and **Gupta's group** the **first cell** should be **rod-shaped**.
- Because of cell wall composition and strategy for growth in **Gram-positive which is much simpler than that of Gram-negative cells**, it was postulated the **first cell** to be **Gram-positive, rod-shaped organism**.

# Circle life tree

## Ring of life

**Astrobiologist Mary Rivera and Molecular biologist James Lake, 2004**

- This is a phylogenetic model where **all three domains of life evolved from a pool of primitive prokaryotes**.
- According to Lake, this structure is the best fit for data from extensive DNA analyses performed in his laboratory, and that the ring model is the only one that adequately takes HGT and genomic fusion into account.
- However, other phylogeneticists remain highly skeptical of this model.

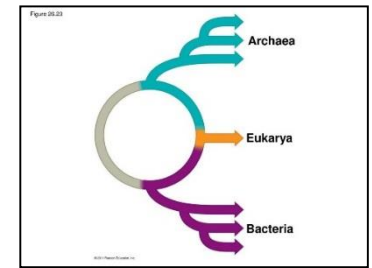


According to the “ring of life” phylogenetic model, the three domains of life evolved from a pool of primitive prokaryotes.

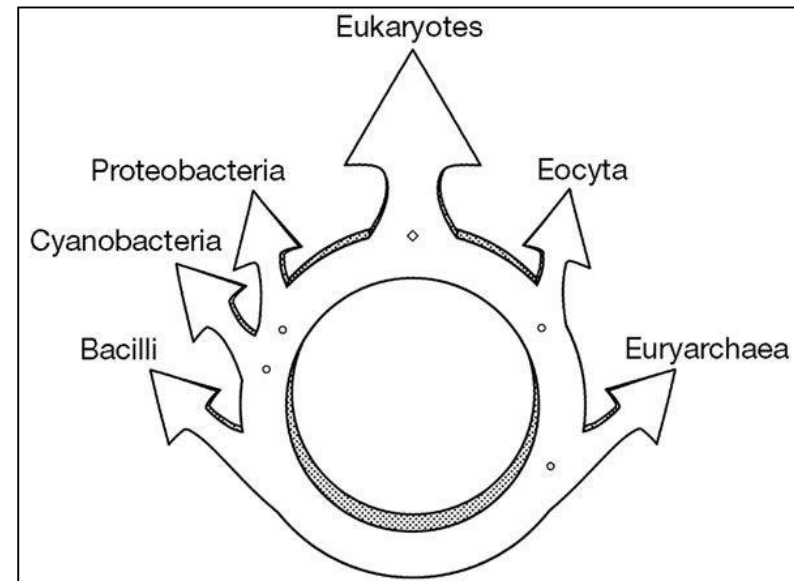
# Circle life tree

## Ring of life

Astrobiologist Mary Rivera and Molecular biologist James Lake, 2004



- They explained that the Ring of Life structure is a result of a **single fusion event between two prokaryotic genomes at the base of the eukaryotic tree**, probably between the ancestors of a photosynthetic bacterium and an archaeon.
- A recent paper, based on an analysis that supposedly takes **horizontal gene transfer** into account, suggests that the **tree is not a tree at all, but a circle**.



In this scenario, **eukaryotes are not ancient**: they are a more recent group than either of the two prokaryotic groups.

This model for the origin of eukaryotes is very different to Woese's tree. The **Archaea**, shown **on the bottom right**, includes the **Euryarchaea**, the **Eocyta** and the **informational eukaryotic ancestor**.



# Circle life tree

## Ring of life

**Vertical transfer(tree life) vs. horizontal transfer(ring life)**

---

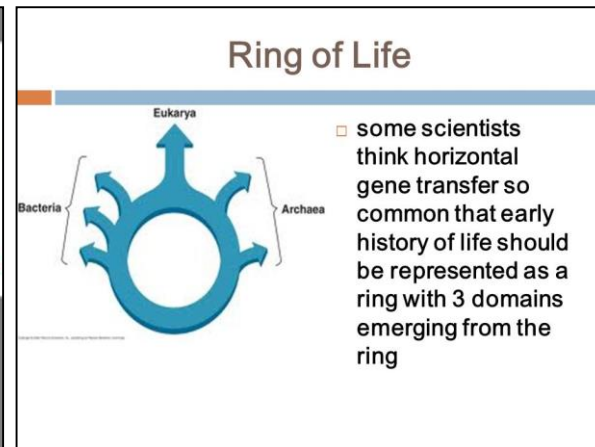
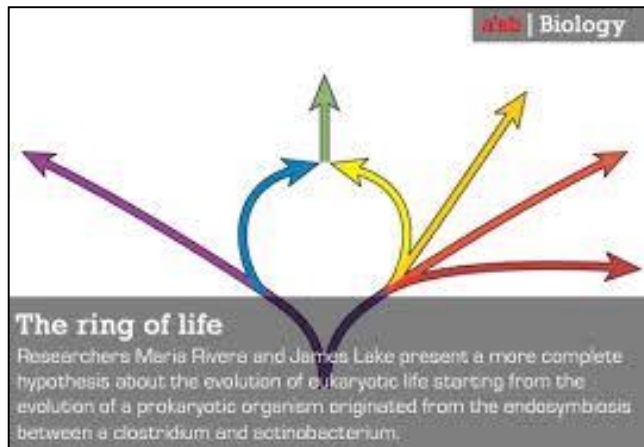
- **Our bacterial parentage: the union of Archaea and Eubacteria**
  1. Vertical transfer of genes producing a tree, with each new production becoming a new branch.
  2. Horizontal gene transfer would produce a genuine (original) circle, or ring, in which two organisms fuse genomes to produce a new organism.
- The most recent version of ring of life scenario is that eukaryogenesis (evolution of eukaryotic life) was triggered by the engulfment of an alpha-proteobacterium by a wall-less giant archaeon capable of phagocytosis.
- The fusion of two genomes may have produced the eukaryotes.

# Circle life tree

## Ring of life

Rivera & James, 2004

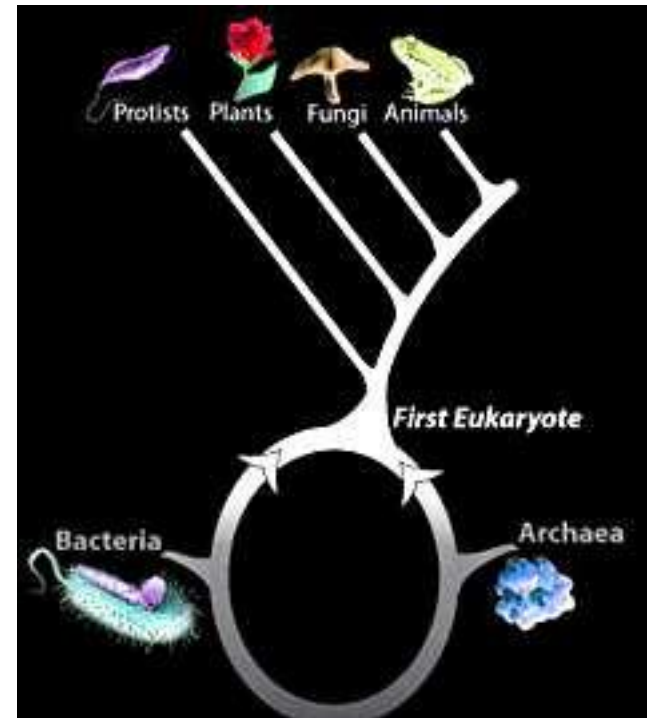
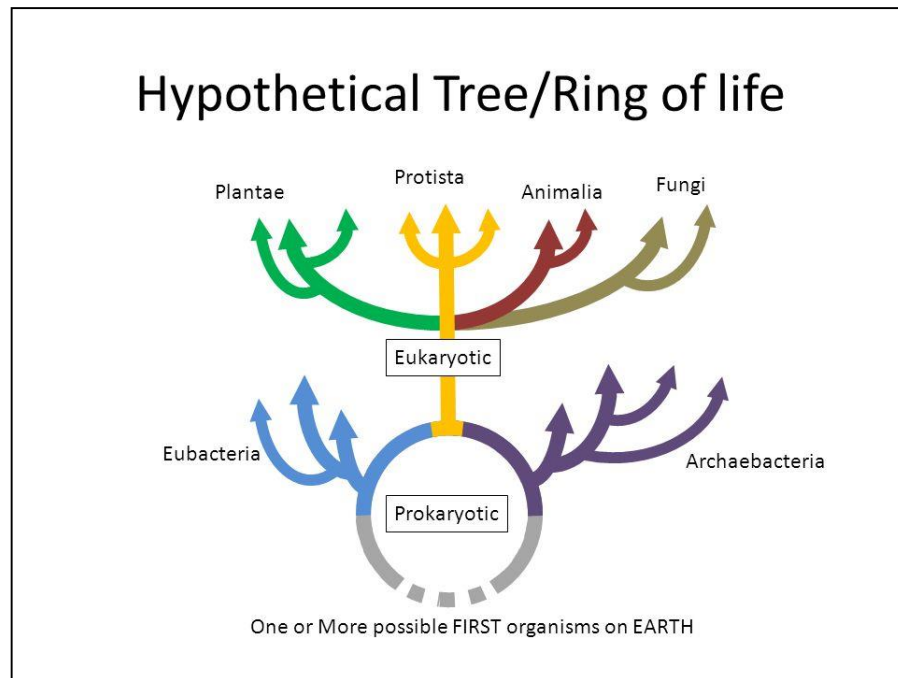
- The **eukaryotes plus the two eukaryotic root organisms** (the operational and informational ancestors) comprise the **eukaryotic domain**.
- **Ancestors** defining major groups in the **prokaryotic domain** are indicated by **small circles on the ring**.
- The **Archaea**, shown **on the bottom right**, includes the **Euryarchaea**, the **Eocyta** and the **informational eukaryotic ancestor**.



# Circle life tree

## Ring of life

Rivera & James, 2004



It's not a tree; it's actually a ring of life. A ring explains the data far better. **One Ring to Rule Them All**. At least 2 billion years ago, ancestors of these **two diverse prokaryotic groups (archaea and bacteria)** fused their genomes to form the first eukaryote, and in the processes **two different branches of the tree of life were fused to form the ring of life.** "The ring will lead to a better understanding of eukaryotes."

# Eocyte hypothesis

## Two-domain tree theory (the eocyte tree)

### Two-domains trees vs. three-domains tree

---

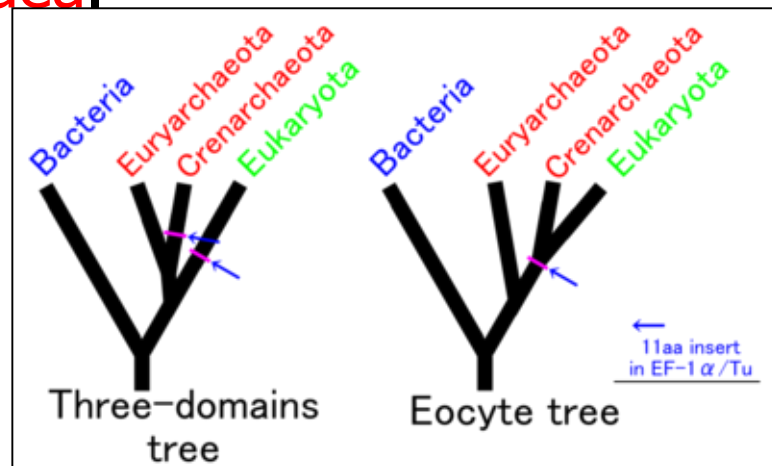
- Two-domains trees, which was first proposed by James Lake and colleagues in 1984 based upon ribosome structure.
- The three-domains and two-domains trees – **competing hypotheses for the origin of eukaryotes(eukaryogenesis-The evolution of eukaryotic life).**

# Eocyte hypothesis

## Two-domain tree theory (the eocyte tree)

### Eukarya branched within Archaea

- The **eocyte hypothesis** is a hypothesis proposed in the 1980s by **James Lake** that **eukarya** evolved from a subgroup of **Archaea** called as **Eocytes**.
- In taxonomy, the **Crenarchaeota** (also known as **Crenarchaea** or **eocytes**) are **a phylum or a kingdom of the Archaea**.

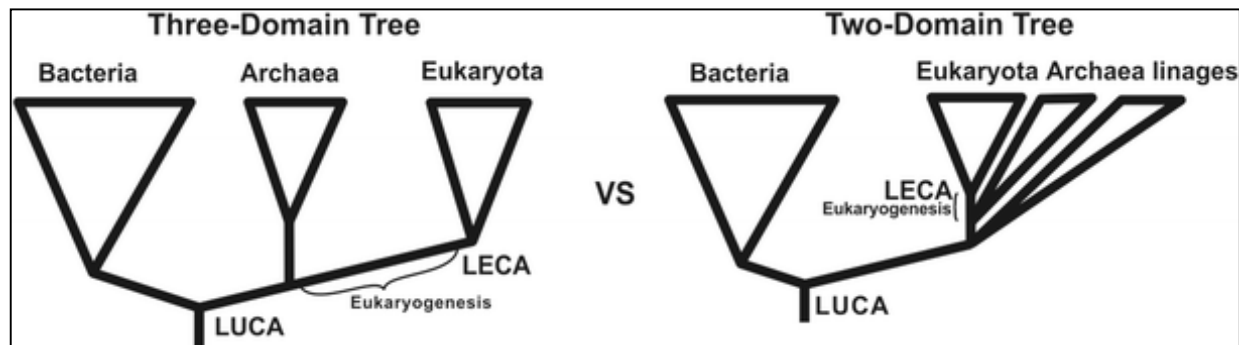


# Eocyte hypothesis

## Two-domains/eocyte tree

### Eukaryogenesis-The evolution of eukaryotic life

- The Eocyte hypothesis:
  1. The **Bacteria and Archaea** can still be considered **distinct primary domains**, but
  2. The **eukaryotes** originate from **within the domain Archaea**.
- In other words, in the 'two-domains/eocyte tree', the **eukaryotic lineage has an archaeal parent**.

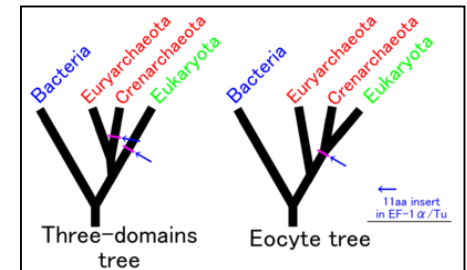


# Eocyte hypothesis

## Two-domains/eocyte tree

### Last eukaryotic common ancestor (LECA)

- A phylogenomic investigation of 28 vertically inherited last eukaryotic common ancestor (LECA) clades supported eukaryotes either
  1. branching deep within Archaea or close to the root of Archaea,
  2. but separate from:
    - *Crechanareota* and
    - *Euryarchaeota* (Rochette *et al.*, 2014).

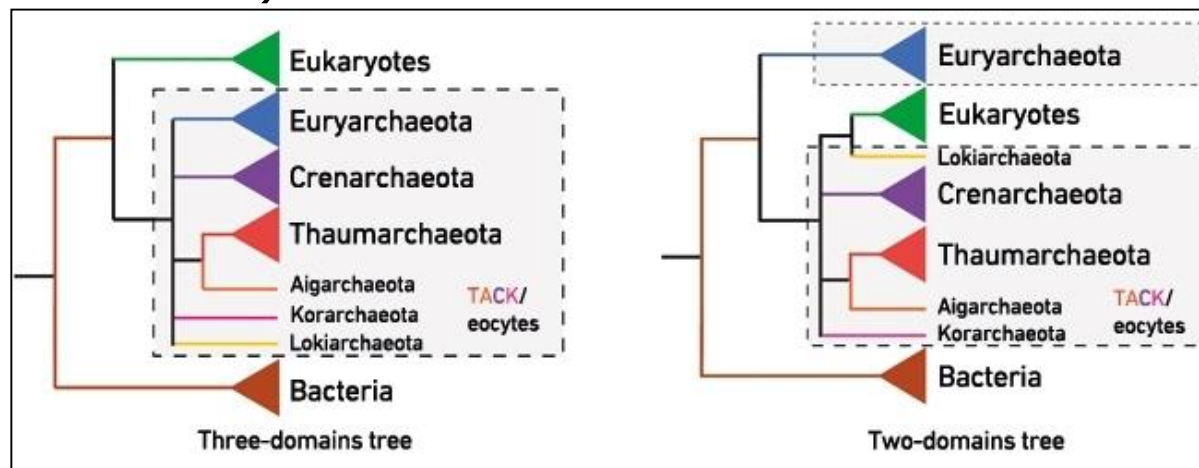


# Eocyte hypothesis

## Two-domains/eocyte tree

### Two-domains trees vs. three-domains tree

- The iconic three-domains tree (Woese's universal tree) appears in most textbooks and divides cellular life into three separate major groups or 'domains': the
- bacteria, the archaea and the eukaryotes.
- In this tree the eukaryotes are held to have originated from a common prokaryotic ancestor shared with the archaea (enclosed in the **shaded box**).



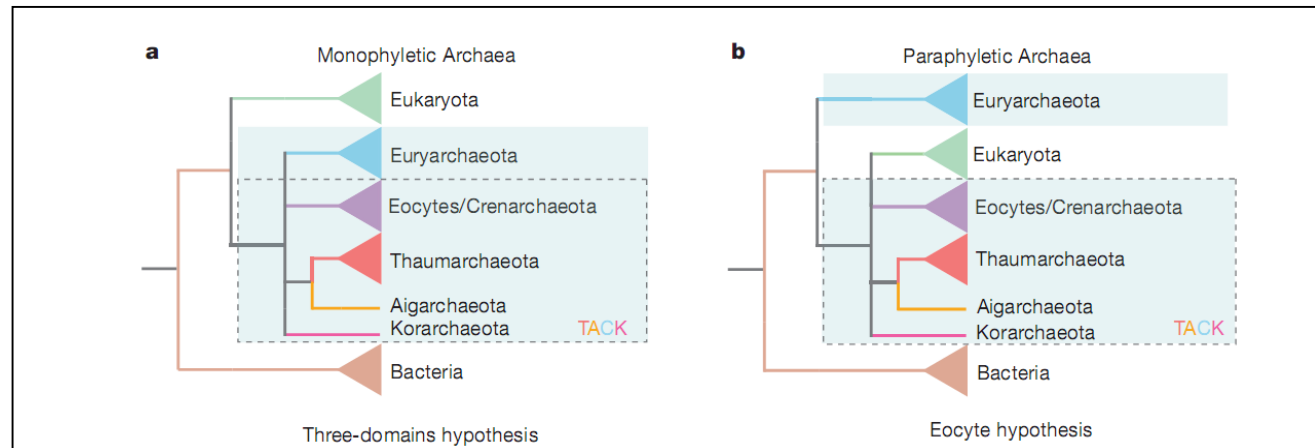


# Eocyte hypothesis

## Two-domains/eocyte tree

### Two-domains trees vs. three-domains tree

- Competing hypotheses for the origin of the eukaryotic host cell.
  - a) In this tree the Archaea and Eukaryota are most closely related to each other because they share a common ancestor that is not shared with Bacteria.
  - b) The rooted eocyte tree recovers the host-cell lineage nested within the archaea as a sister group to the eocytes (which Woese et al. called the Crenarchaeota); this implies that, on the basis of the small set of core genes, there are only two primary domains of life—the Bacteria and the Archaea.

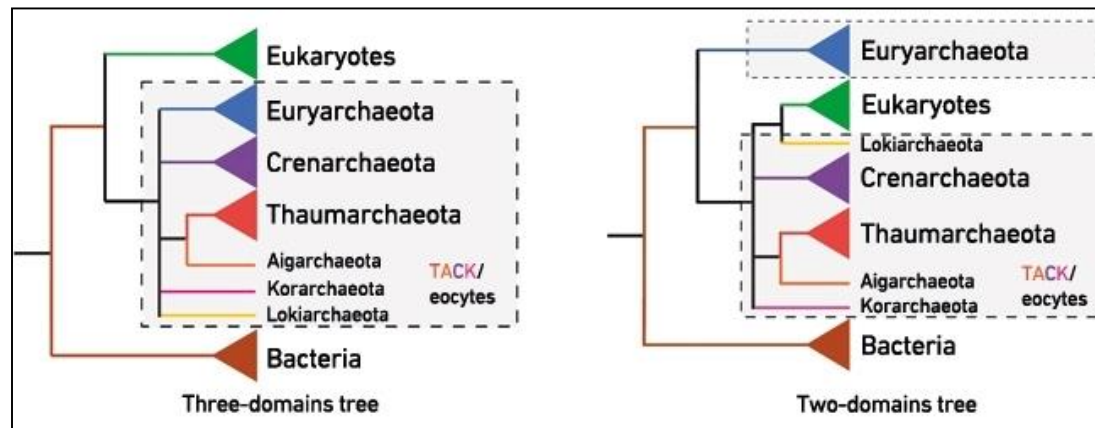


# Eocyte hypothesis

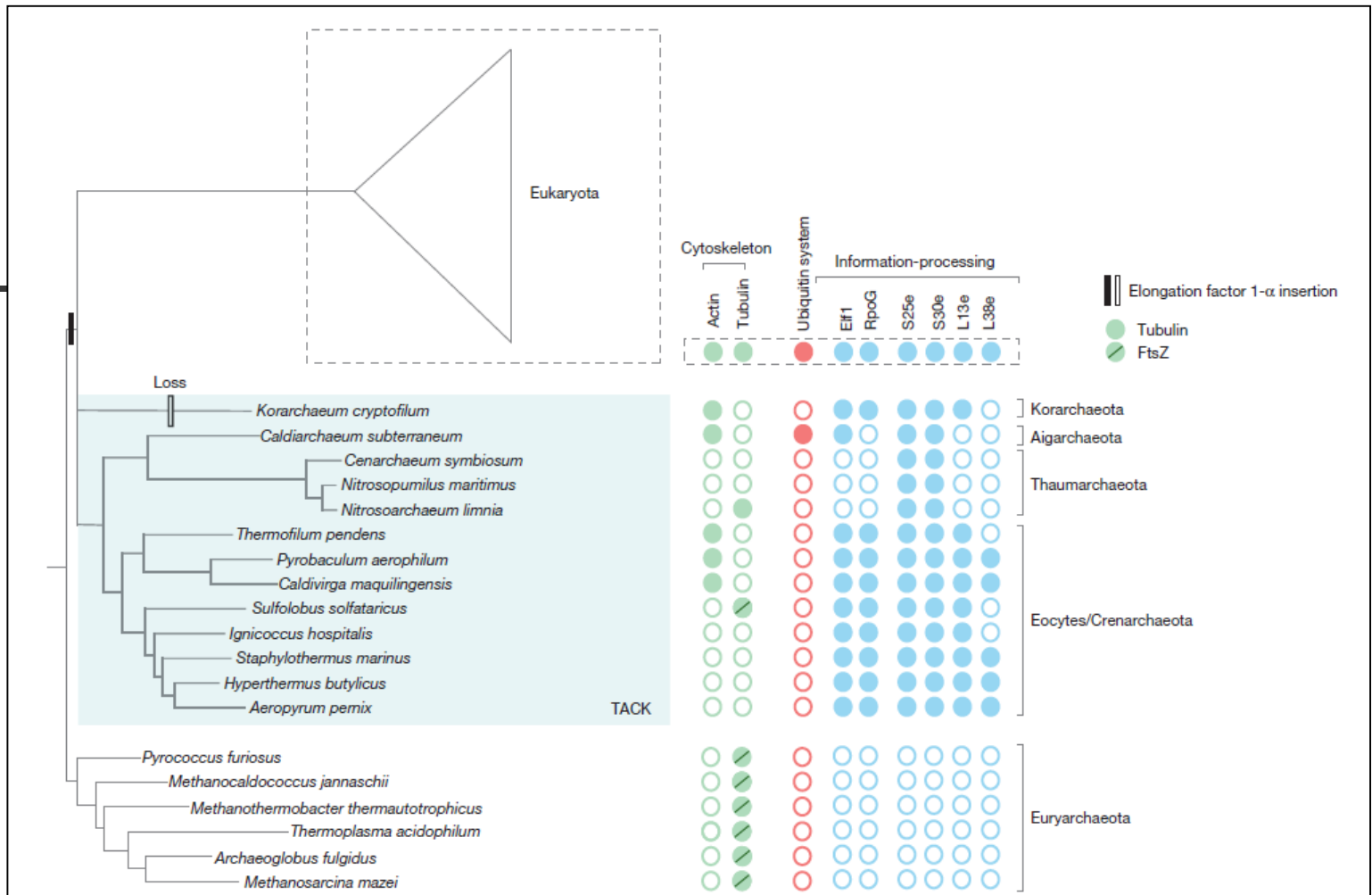
## A different universal tree

### Two-domains trees vs. three-domains tree

- By contrast, the two-domains/eocyte tree recovers eukaryotes nested inside the archaea with the newly discovered *lokiarchaeota* currently thought to be the closest archaeal relatives of the eukaryotes.
- In the two-domains/eocyte tree the eukaryotic lineage had an ancestor that was already an archaea.
- The genomic and cellular features of these lineages could potentially illuminate important stages in the evolution of eukaryotic cells like our own.



TACK aracheae includes:  
the *thaumarchaeota*, *aigarchaeota*, *crenarchaeota* and *korarchaeota*.



## Archaeal links in the origin of eukaryotes.

A schematic tree depicting the **relationships between Archaea and the eukaryotic nuclear lineage**, consistent with recent analyses of core genes using new methods and rooted using the **Bacteria as the outgroup**.

# Eocyte hypothesis

## A different universal tree

### Two-domains trees vs. three-domains tree

---

- The “eocyte” scenario is supported by phylogenetic analyses of universal proteins that use sophisticated methods for tree reconstruction, which are thought to be very efficient at identifying weak phylogenetic signals.
- However, these data are controversial, because most universal proteins are small (e.g., ribosomal proteins) and very divergent between Bacteria and Archaea/Eukarya, which makes archaeal/eukaryal relationships difficult to resolve.

# The new tree of life and ongoing debate

## The new tree of life by Hug et al., 2016

### The first comprehensive phylogenomic tree

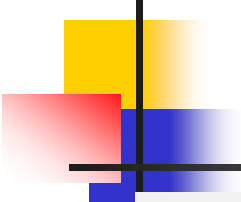
- The tree of life proposed by Hug *et al.*, 2016 is the first comprehensive phylogenomic tree since the advent of genome-resolved metagenomic sequencing and analysis methods.
- One representative high-quality or complete genome per genus (3083 organisms, out of which 1011 organisms are novel) was used for phylogenomic reconstruction of this tree.
- The SSU rRNA gene-based phylogeny largely agrees with concatenated 16 ribosomal protein-based phylogeny.
- However,
  1. the former one(SSU rRNA genes) shows a three-domain topology,
  2. while the latter one(16 ribosomal proteins) shows a two-domain topology, placing Eukarya sibling to *Lokiarchaeota*, a proposed phylum of the domain Archaea.



# Characteristics of the SSU(small subunit) rRNA for exemplary species

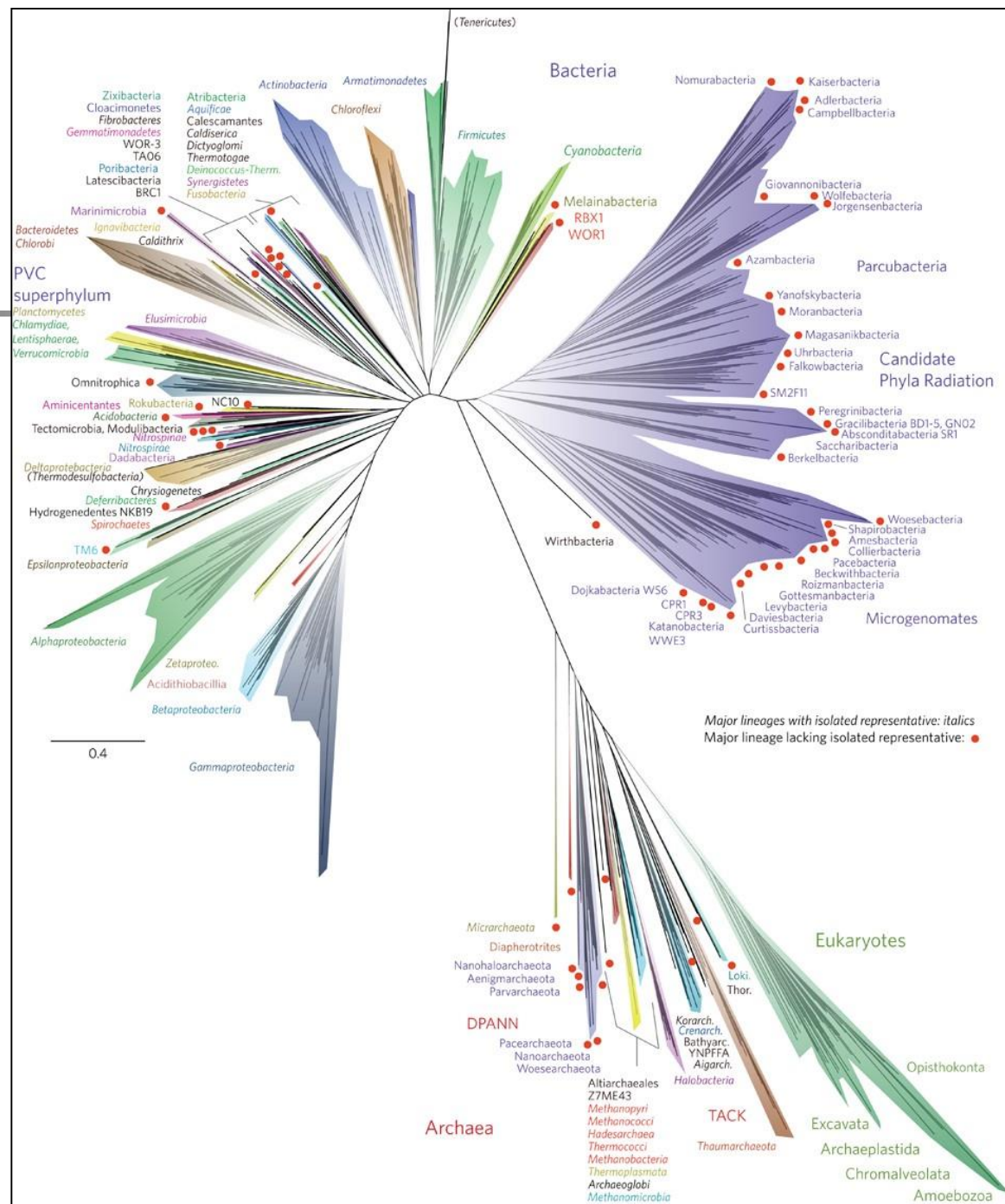
- Small subunit ribosomal ribonucleic acid (SSU rRNA) is the smallest of the two major RNA components of the ribosome.
- Associated with a number of ribosomal proteins, the SSU rRNA forms the small subunit of the ribosome.
- It is encoded by the SSU-rDNA.

Type	SSU rRNA size	Species	Length
Archaeal (Prokaryotic)	16S	<i>Halobacterium salinarum</i>	1,473 nt
Plastid	16S	<i>Arabidopsis thaliana</i>	1,491 nt
Bacterial (Prokaryotic)	16S	<i>Escherichia coli</i>	1,541 nt
Eukaryotic	18S	<i>Homo sapiens</i>	1,969 nt
Mitochondrial	12S	<i>Homo sapiens</i>	954 nt



The first  
comprehensive  
phylogenomic tree  
with three domains:  
Bacteria  
Archaea  
Eukaryotes.

A current view of the  
tree of life,  
encompassing the  
total diversity  
represented by  
sequenced genomes.  
The new tree of life  
by Hug *et al.*, 2016.

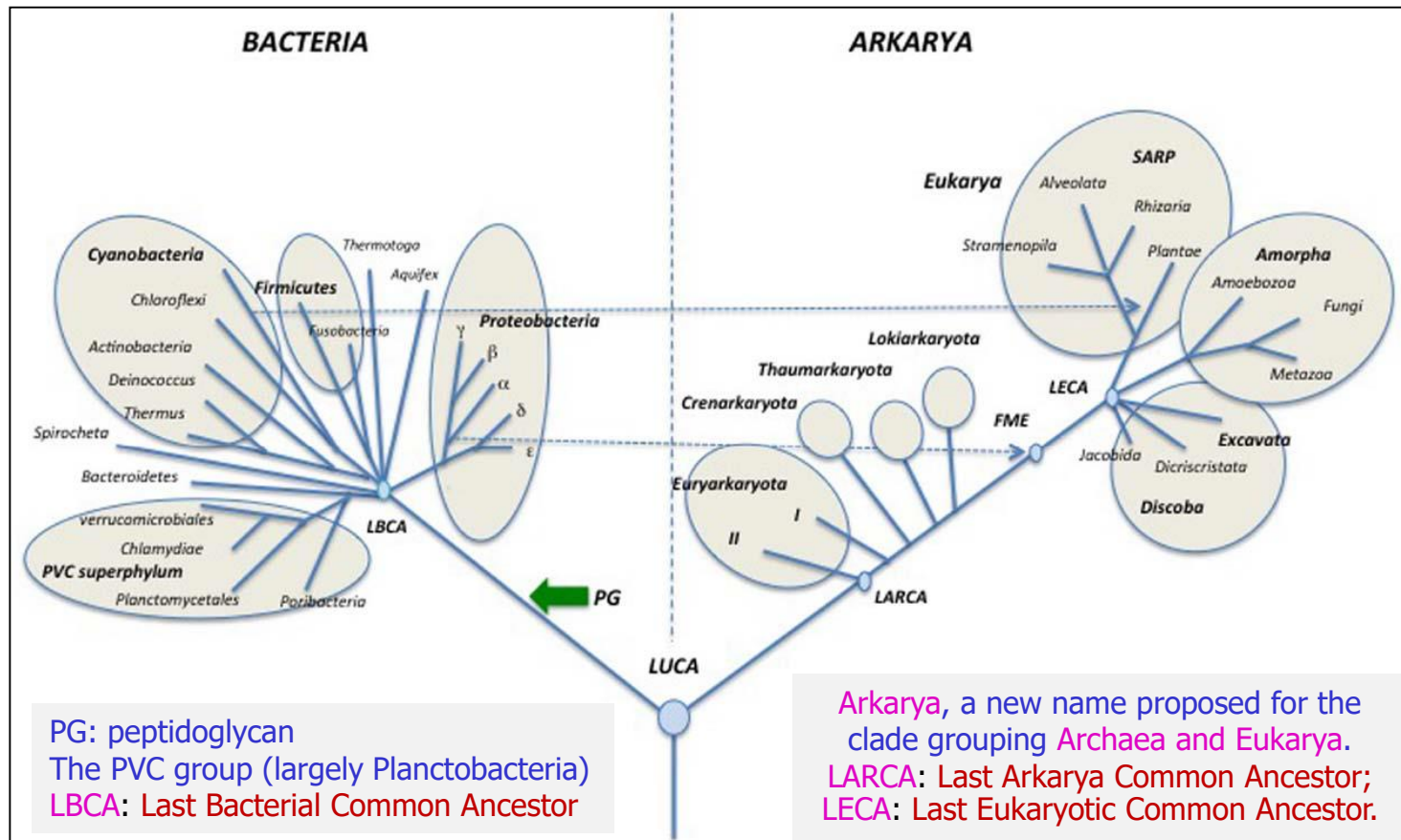




# The universal tree of life: an update

Universal tree of life, based on 16S rRNA sequences

Two domains: **Bacteria** and **Arkarya**

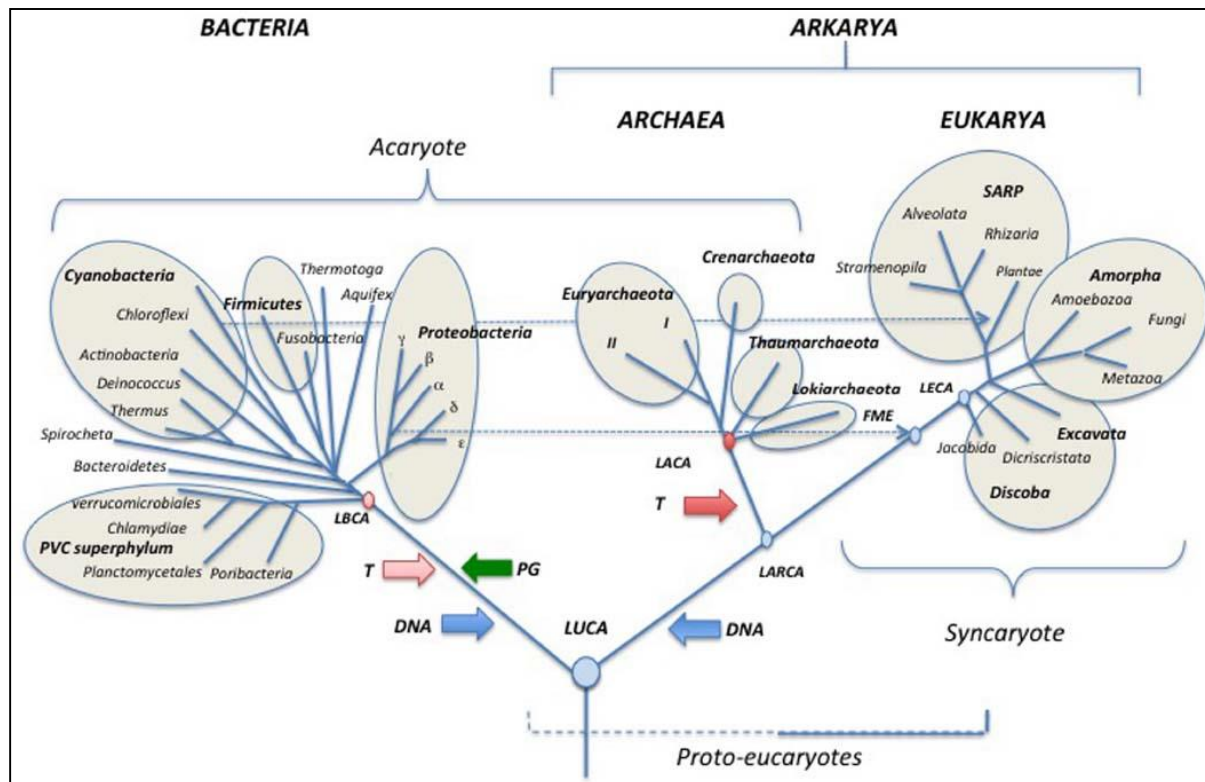




# The universal tree of life: an update

## Bacteria and Arkarya

### LBCA, LACA, LARCA and LECA



**Schematic universal tree updated from (Woese *et al.*, 1990).**

DNA (blue arrows) introduction of DNA; T (pink and red arrows) thermoreduction. **LBCA**: Last Bacterial Common Ancestor, pink circle: thermophilic LBCA; **LACA**: Last Archaeal Common Ancestor, red circle, hyperthermophilic LACA. **LARCA**: Last Arkarya Common Ancestor; **FME**: First Mitochondriate Eukarya; **LECA**: Last Eukaryotic Common Ancestor; blue circles, mesophilic ancestors. **SARP**: Stramenopila, Alveolata, Rhizobia,



# Ruggiero *et al.*, 2015

## A Higher Level Classification of All Living Organisms

---

### Superkingdom concept

Noun. **regnum** (plural **regnums** or **regna**) (biology, taxonomy) A rank in the classification of organisms, also known as **kingdom**.

Ruggiero, M.A., D.P. Gordon, T.M. Orrell, N. Bailly, T. Bourgoïn, R.C. Brusca, T. Cavalier-Smith, M.D. Guiry and P.M. Kirk. 2015. **Correction: A Higher Level Classification of All Living Organisms**. PLoS ONE 10(6): e0130114. doi:10.1371/journal.pone.0130114.

# Ruggiero *et al.*, 2015

## Two superkingdoms: Prokaryota and Eukaryota and seven kingdoms

- We are proposing a two-superkingdom (Prokaryota and Eukaryota), seven-kingdom classification that is a practical extension of Cavalier-Smith's six-kingdom schema(1998).
- Our schema includes:
- **The prokaryotic kingdoms:**
  1. Archaea (Archaeobacteria), and
  2. Bacteria (Eubacteria), and
- **The eukaryotic kingdoms:**
  1. Protozoa,
  2. Chromista,
  3. Fungi,
  4. Plantae, and
  5. Animalia.

**Chromista** so-called "crown eukaryotes", includes not only plants, animals, and fungi, but also Alveolates and possibly the red algae.

Cavalier-Smith in his megaclassification(1998) proposed two Empires: Prokaryota and Eukaryota and six kingdoms: Bacteria, Protozoa, Chromista, Plantae, Fungi and Animalia.

# Ruggiero *et al.*, 2015

## Two superkingdoms: Prokaryota and Eukaryota and seven kingdoms

Linnaeus 1735	Haeckel 1866	Chatoon 1925	Copeland 1938	Whittaker 1969	Woese et al. 1977	Woese et al. 1990	Cavalier- Smith, 1993	Cavalier- Smith, 1998	Ruggiero <i>et al.</i> 2015
2 kingdoms	3 kingdoms	2 empires	4 kingdoms	5 kingdoms	6 kingdoms	3 domains	8 kingdoms	6 kingdoms	7 kingdoms
(not treated)	Protista	Prokaryota	Mychota	Monera	Bacteria	Eubacteria	Eubacteria	Bacteria	Bacteria
					Archaeobacteria	Archaea	Archaeobacteria		Archaea
		Euokaryota	Protoctista	Protista	Protista	Eukarya	Archezoa	Protozoa	Protozoa
							Protozoa		
Vegetabilia	Plantae		Plantae	Plantae	Plantae		Chromista	Chromista	Chromista
				Fungi	Fungi		Plantae	Plantae	Plantae
							Fungi	Fungi	Fungi
Animalia	Animalia		Animalia	Animalia	Animalia		Animalia	Animalia	Animalia

# Ruggiero *et al.*, 2015

## A Higher Level Classification of All Living Organisms

List of ranks used in the hierarchy with the number of taxa per rank

Rank	Number of Taxa
Superkingdom	2
<b>Kingdom</b>	<b>7</b>
Subkingdom	11
Infrakingdom	8
Superphylum	6
<b>Phylum</b>	<b>96</b>
Subphylum	60
Infraphylum	4
Superclass	12
<b>Class</b>	<b>351</b>
Subclass	145
Infraclass	23
Superorder	52
<b>Order</b>	<b>1,467</b>

Main ranks are in bold type; unnamed taxa are not counted.

# Ruggiero *et al.*, 2015

## A Higher Level Classification of All Living Organisms **Prokaryota**

---

- The higher classification of prokaryotes is still somewhat unsettled.
- Woese and Fox (1997) treated **Archaeobacteria** (Archaea) and **Eubacteria** (Bacteria) as separate kingdoms.
- Margulis and Schwartz (2001) recognized the **superkingdom Prokarya**, containing one kingdom **Bacteria** that included a **subkingdom Archaea**.
- Cavalier-Smith (1998 and 2014) also treated **Archaeobacteria** and **Eubacteria** as prokaryote subkingdoms.

# Ruggiero *et al.*, 2015

## A Higher Level Classification of All Living Organisms

### Prokaryota

---

- As no prokaryote names above the ranks of class are covered by ICNB rules, there is no official higher classification of prokaryotes (Parte, 2014) and any attempt at such is necessarily difficult.
- We have chosen to adopt the classification in current use by the Catalogue of Life (CoL's database):  
<http://www.catalogueoflife.org/col/>
- It is derived from the TOBA (Taxonomic Outline of Bacteria and Archaea) and recognizes Bacteria and Archaea as equivalent in rank to the eukaryote kingdoms.

# Ruggiero *et al.*, 2015

## A Higher Level Classification of All Living Organisms

### Prokaryota

- We treat them as *de facto* (accepted) kingdoms until there is a better resolution of their status.
- The number of **negibacterial** “phyla” currently **recognized** (LPSN, 2014) is **probably excessive compared with eukaryotes** and mainly reflects uncertainty about the true relationships of many small phyla, probably exaggerating the significance of their biological disparity.
- Greater use of multigene trees rather than over reliance on rRNA gene trees alone may eventually allow further simplification by grouping them into fewer phyla, possibly only about half the present number (Margulis and Schwartz, 2001).



# Catalogue of Life

## Species 2000 CheckList

<http://www.catalogueoflife.org/col/>



The screenshot shows the homepage of the Species 2000 & ITIS Catalogue of Life. At the top, it says "Species 2000" in a stylized font, followed by "ITIS" and "Catalogue of Life: 30th April 2017" with the tagline "indexing the world's known species". Below this is a horizontal bar with language options: English, French, Spanish, Chinese, Russian, Portuguese, Dutch, German, Polish, Lithuanian, Thai, and Vietnamese. On the left, there is a sidebar with "Browse", "Search", and "Info" links. The main content area is titled "Search the Catalogue of Life - updated edition around the year" and contains a search form with a text input field, checkboxes for "Show extinct taxa (+)" and "Match whole words only", and a "Search" button. At the bottom, a disclaimer states: "Annual Checklist Interface v1.9 r2126ab0 developed by Naturalis Biodiversity Center. Please note, this site uses cookies. If you continue to use the site we will assume that you agree with this."

Roskov Y, Kunze T, Orrell T, Abucay L, Paglinawan L, Culham A, et al., editors. *Species 2000 & ITIS Catalogue of Life, 2014 Annual Checklist* [DVD]. 2014; Naturalis, Leiden, the Netherlands: Species 2000.

# Catalogue of Life

## Species 2000 CheckList

<http://www.catalogueoflife.org/col/>



Species 2000  
Catalogue of Life: 30th April 2017  
indexing the world's known species

English French Spanish Chinese Russian Portuguese Dutch German Polish Lithuanian Thai Vietnamese

Browse Search Info

Search all names - Results for "xylella"

Records found: 4

Records per page: 20 Update

[Export search results](#) | [New search](#)

Name	Rank	Name status	Group	Source database
<a href="#">Xylella</a>	Genus		Bacteria	
<a href="#">Xylella fastidiosa</a> Wells et al., 1987	Species	accepted name	Bacteria	
<a href="#">Xylella fastidiosa fastidiosa</a> Wells et al., 1987	Intraspecific taxon	accepted name	Bacteria	
<a href="#">Xylella fastidiosa multiplex</a> Schaad et al., 2009	Intraspecific taxon	accepted name	Bacteria	

[Export search results](#) | [New search](#)

Annual Checklist Interface v1.9 r2126ab0 developed by Naturalis Biodiversity Center. Please note, this site uses [cookies](#). If you continue to use the site we will assume that you agree with this.

Beyond its immediate use as a management tool for the [CoL](#) and [ITIS](#) (Integrated Taxonomic Information System), it is immediately valuable as a reference for taxonomic and biodiversity research, as a tool for societal communication, and as a classificatory "backbone" for biodiversity databases, museum collections, libraries, and textbooks. Such a modern comprehensive hierarchy has not previously existed at this level of specificity.

## Subkingdom: Posibacteria

SUBKINGDOM POSIBACTERIA (Continued)

PLOS ONE | DOI:10.1371/journal.pone.0119248 April 29, 2015 13/90

**PLOS ONE** A Higher Level Classification of All Living Organisms

Table 2. (Continued)

Phylum Actinobacteria	Class Actinobacteria	Order Acidimicrobiales
		Order Actinomycetales
		Order Bifidobacteriales
		Order Coriobacteriales
		Order Euzetiales
		Order Gemmatimonadetes
		Order Nitrospirales
		Order Rubrobacteriales
		Order Solirubrobacteriales
		Order Thermococcales
Phylum Chloroflexi [n Chlorobacteria]		
Class Anaerolineae	Order Anaerolineales	
Class Caldilineae	Order Caldilineales	
Class Chloroflexia	Order Chloroflexiales	
	Order Herpetosiphonales	
Class Dehalococcoidia	Order Dehalococcoidales	
Class Kilonobacteria	Order Kilonobacteriales	
	Order Thermogemmatimonadetes	
Class Thermomicrobia	Order Spirochaetales	
	Order Thermomicrobiales	
Phylum Firmicutes	Class Bacilli	Order Bacillales
		Order Lactobacillales
	Class Clostridia	Order Clostridiales
		Order Helicobacteriales
		Order Natronaerobiales
		Order Thermomicrobiales
	Class Erysipelotrichia	Order Erysipelotrichales
	Class Negativicutes	Order Selenomonadales
	Class Thermolithobacteria	Order Thermolithobacteriales
Phylum Tenericutes	Class Mollicutes	Order Achaeoplasmales
		Order Anaeroplasmatales
		Order Entomoplasmatales
		Order Haloplasmatales
		Order Mycoplasmales

(Continued)

Unibacteria, comprising Archaeobacteria and Posibacteria. It was not recognized in this scheme.

## Subkingdom: Negibacteria

SUBKINGDOM NEGIBACTERIA

PLOS ONE | DOI:10.1371/journal.pone.0119248 April 29, 2015 10/90

**PLOS ONE** A Higher Level Classification of All Living Organisms

Table 2. (Continued)

Phylum Acidobacteria	Class N.N. (Bryobacter)	
	Class Acidobacteria	Order Acidobacteriales
	Class Holophagae	Order Acanthopileobacteriales
		Order Holophagales
Phylum Aquificae	Class Aquificae	Order Aquificales
Phylum Amnatmonadetes		

(Continued)

	Class Amnatmonadetes	Order Amnatmonadales
	Class Oribacteriales	Order Oribacteriales
	Class Fimbriomonadales	Order Fimbriomonadales
Phylum Bacteroidetes	Class Bacteroidia	Order Bacteroidales
	Class Cytophagia	Order Cytophagales
	Class Flavobacteriia	Order Flavobacteriales
	Class Sphingobacteriia	Order Sphingobacteriales
Phylum Caldiseptica	Class Caldiseptica	Order Caldisepticales
Phylum Chlamydiae	Class Chlamydiae	Order Chlamydiales
Phylum Chlorobi	Class Chlorobia	Order Chlorobiales
	Class Ignimbacteriia	Order Ignimbacteriales
Phylum Chrysiogenetes	Class Chrysiogenetes	Order Chrysiogenales
Phylum Cyanobacteria [n Cyanophyta]	Class Cyanophyceae [n Phycobacteria]	Order Chroococcales
		Order Nostocales
		Order Oscillatoriales
		Order Pseudonatales
		Order Synechococcales
	Class Gloeobacteria [n Gloeobacteriophyceae]	Order Gloeobacteriales
Phylum Deferibacteres	Class Deferibacteres	Order Deferibacteriales
Phylum Deinococcus-Thermus [n Halobacteria]	Class Deinococcus	Order Deinococcales
		Order Thermodesulfobacteriales
Phylum Dictyoglomi	Class Dictyoglomi	Order Dictyoglomales
Phylum Elusimicrobia	Class Elusimicrobia	

(Continued)

PLOS ONE | DOI:10.1371/journal.pone.0119248 April 29, 2015 11/90

Ruggiero *et al.*, 2015



# **Major Topics**

## **In Practical Phylogeny**

---

**Theoretically Discussed Topics**



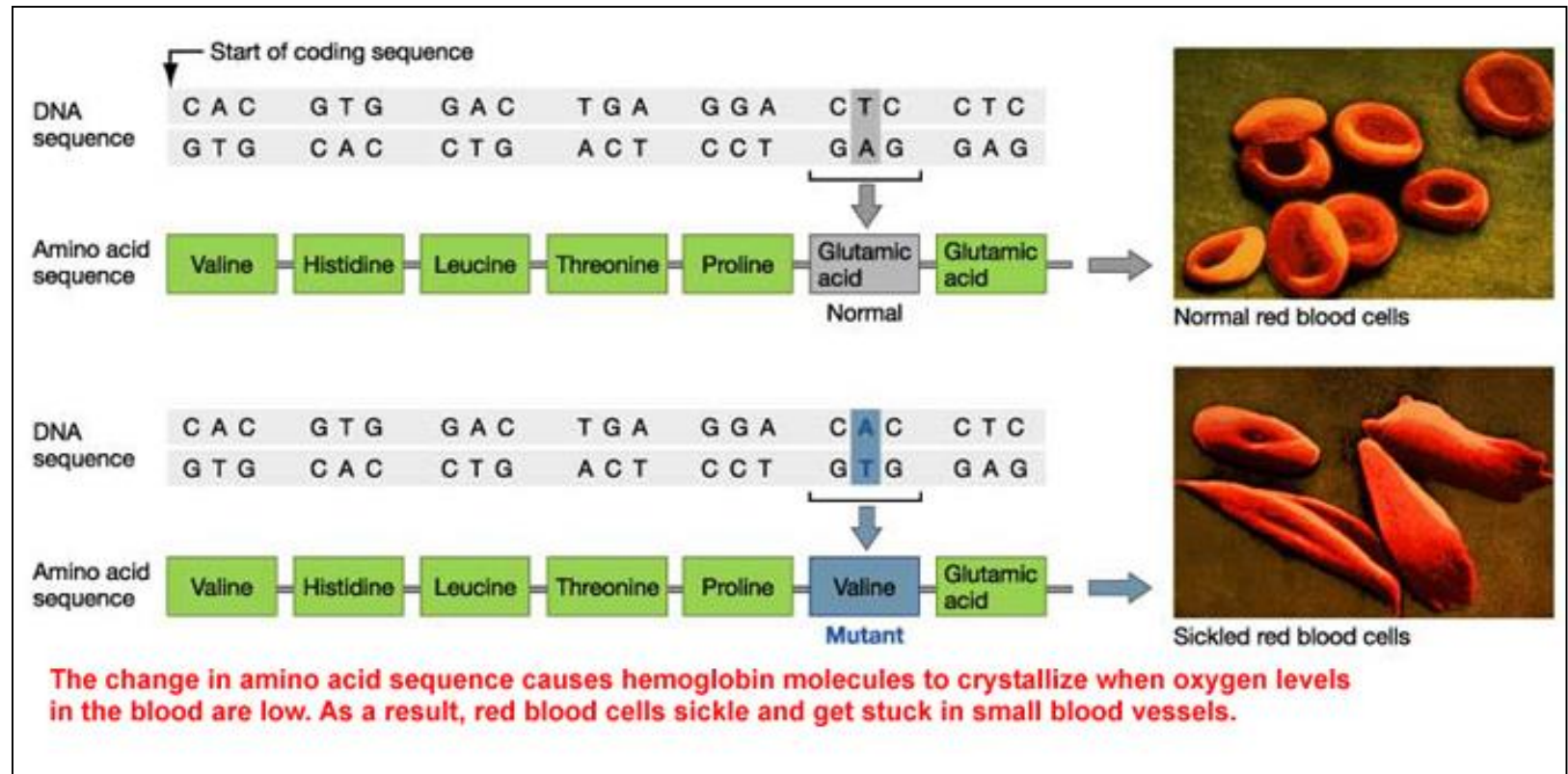
# 1. Mutation Rate

---

- Rate of accumulation of mutations happens at different rates in different genes.
- This happens because the gene products (RNA or protein) differ in how many changes they can tolerate and still function.
- DNA regions evolving at a very slow rate do not contain much phylogenetic information, because the sequences will not differ much, if at all, between taxa.
- If DNA regions are evolving very rapidly, then there will be so many parallelisms and mutations that all information will be lost (obliterated by too much evolution).

# Mutation

**Change in primary amino acid sequence = defective protein/Sickle cell**





## 2. Proteins revolution rate

- However, it has been found that different rates of DNA base replacements due to accumulation of mutations over time (which result in amino acid replacements) exist for different genes, species, etc.

### Rates of amino acid replacement in different proteins

Protein	Rate (mean replacements per site per 10 <sup>9</sup> years)
Fibrinopeptides	8.3
Insulin C	2.4
Ribonuclease	2.1
Haemoglobins	1.0
Cytochrome C	0.3
Histone H4	0.01



# Proteins revolution rate

---

- Accumulation of **mutations over time** result in **amino acid replacements** which exist for different genes, species, etc.
- The initial proposal saw the clock as a Poisson process with a constant rate.
- It is now known to be **more complex** - differences in rates occur for:
  - Different sites in a molecule
  - Different genes
  - Different regions of genomes, and
  - Different different taxonomic groups for the same gene.
- There seems to be no clear evidence for a universal molecular clock.





# 3. Molecular chronometers

## An evolutionary clocks

---

- An **evolutionary chronometer** is a characteristic that is a measure of evolutionary change.
- **Changes are:**
  - Neutral
  - Occur randomly
  - Increase over time
- Thus, **sequenced informational macromolecules** are the most useful chronometers in molecular biology.
- **Sequences change very slowly over evolutionary time.**
- **Choosing the Right Chronometer**
- **Ribosomal RNAs as Evolutionary Chronometers**



# Molecular clock

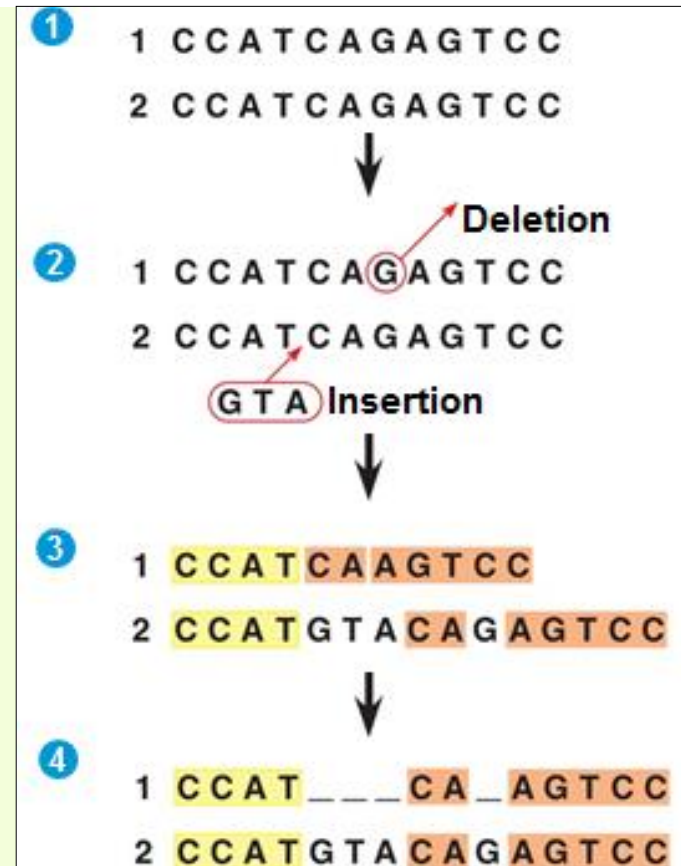
---

- The idea of a **molecular clock** was initially suggested by Zuckerkandl and Pauling in 1962.
- They noted that rates of amino acid replacements in **animal hemoglobins** were roughly proportional to time - as judged **against the fossil record**.
- However, it has been found that different rates of DNA base replacements (**Proteins revolution rate**).

# Molecular clock

**Homologous structures are coded by genes with a common origin**

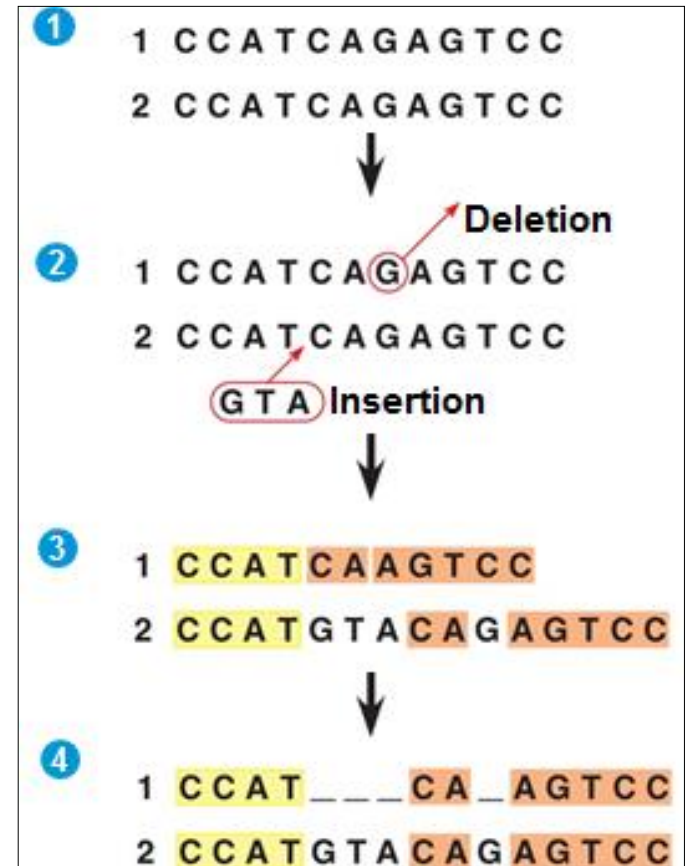
- These genes may mutate but they still retain some common and ancestral DNA sequences.
- Genomic sequencing, computer software and systematics are able to identify these molecular homologies.
- The more closely related two organisms are, the more their DNA sequences will be alike.
- The colored boxes represent DNA homologies.



# Molecular clock

**Homologous structures are coded by genes with a common origin**

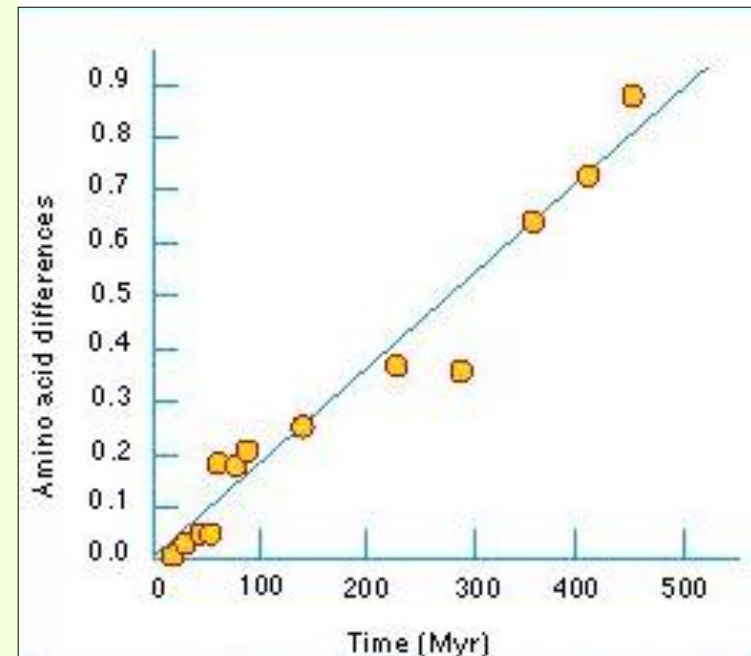
- The molecular clock hypothesis states: **Among closely related species**, a given gene usually evolves at reasonably constant rate.
- These **mutation events** can be used to **predict times of evolutionary divergence**.
- Therefore, the protein encoded by the gene accumulates amino acid replacements at a relatively constant rate.



# Molecular clock

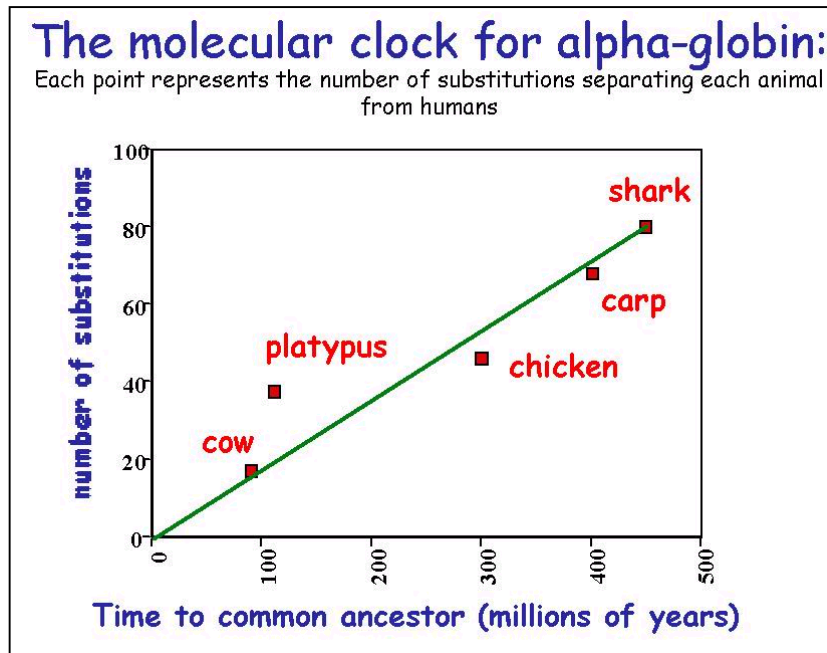
**Homologous structures are coded by genes with a common origin**

- The amino acid replacement for hemoglobin has occurred at a relatively constant rate over 500 million years.
- The slope of the line represents the average rate of change in the amino acid sequence of the molecular clock.
- Different genes evolve at different rates and there are many other factors that can affect the rate.



# Molecular clock

- The rates of **amino acid replacements** in animal hemoglobins were roughly proportional to time.





# Phylogenetics

## Molecular clock

---

- Based on **molecular clock** (Substitutions occur with time).
- **Phylogenies reflect evolutionary history.**
- Development of DNA sequencing technologies.
- Development of programs which compare sequences, produce matrices and construct phylogenies.



# Phylogenetic Trees

## Molecular clock

---

- Branches, clades and lineages reflect evolutionary history and relatedness.
- Can use databases for reference sets.
- Based on alignment, takes account of position.
- Remarkably, 16S sequencing can identify the majority of bacteria to species/genus level.





## 4. Saturation

---

- **Saturation** is due to **multiple changes** at the **same site** subsequent to lineage splitting.
- Most data will contain some fast evolving sites which are potentially saturated (**e.g.** in proteins, often **DNA base position 3 in the genetic code**).
- In severe cases the data becomes essentially random and all information about relationships can be lost.



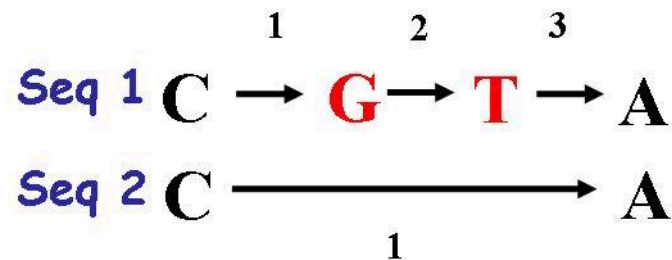
# Multiple changes at the same site

Multiple changes at a single site -  
hidden changes

Seq 1 AGCGAG

Seq 2 GCGGAC

**Number of changes**

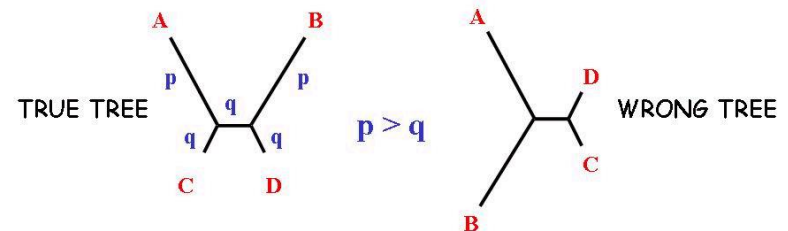


# Fast evolving sites

- Most data will contain some fast evolving sites which are potentially saturated (e.g. in proteins, often DNA base position 3 in the genetic code).

Unequal rates in different lineages may cause problems for phylogenetic analysis

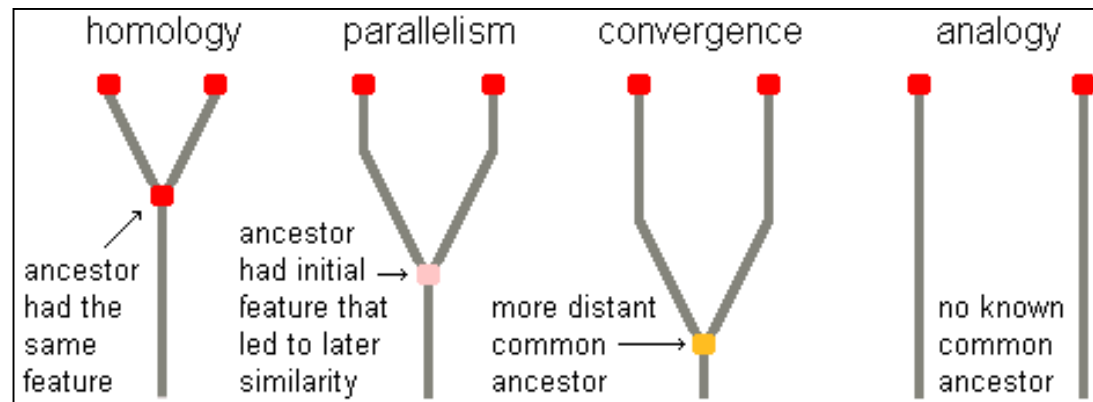
- Felsenstein (1978) made a simple model phylogeny including four taxa and a mixture of short and long branches



- All methods are susceptible to “long branch” problems
- Methods which do not assume that **all sites change at the same rate** are generally better at recovering the true tree

## 5. Homoplasy

- Homoplasy is similarity that is not homologous (not due to common ancestry).
- Homoplasy is the result of independent evolution:
  1. Convergence,
  2. Parallelism, and
  3. Reversal.





# Homoplasy vs. Homology

---

- **Homology:** Common ancestry of two or more character states. i.e. similarity of a trait in two or more species indicates descent from a common ancestor.
- **Homoplasy:** A collection of phenomena that leads to similarities in character states for reasons other than inheritance from a common ancestor (e.g. convergence, parallelism, reversal).
- The commonest cause of homoplasy in morphological traits is convergence, in DNA sequences mutation.
- Homoplasy is huge problem in morphology data sets! But in molecular data sets, too!

# Homoplasy

- Homoplasy can provide misleading evidence of phylogenetic relationships (if mistakenly interpreted as homology).

## Homoplasy - misleading evidence of phylogeny

- If misinterpreted as homology, the absence of tails would be evidence for a wrong tree: grouping humans with frogs and lizards with dogs





# Convergent revolution

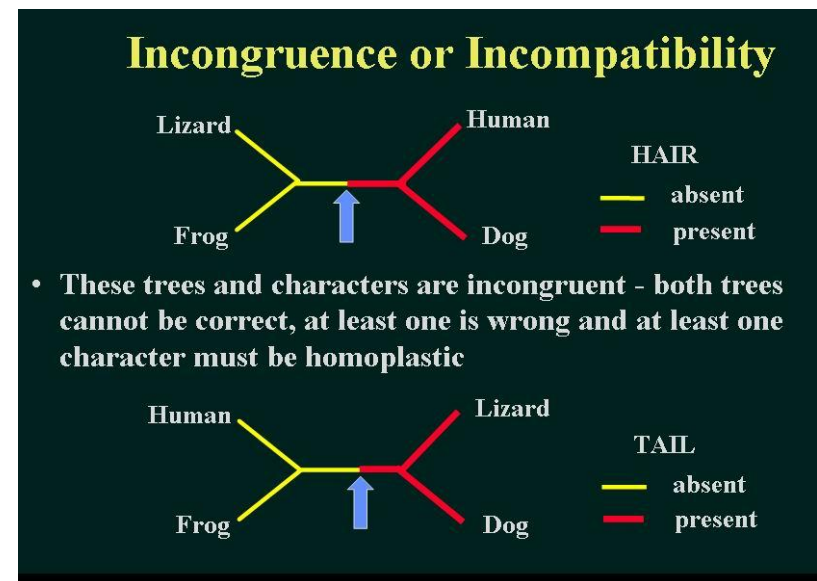
---

- Evolution of similarities in unrelated **groups of organisms**.
- Adaptation for similar function may lead to novel characteristics (**homoplasies**), which are similar, although they are **not inherited** from a common ancestor.
- In **some cases**, **such similarities may be superficial**, as in the wings of **birds, bats, and insects**.
- **In others**, **similarities can be so striking** that it is difficult to determine that the traits arose independently and then later converged upon their current form.

# Homoplasy

## Incongruence or Incompatible

- Incongruence and therefore homoplasy can be common in molecular sequence data.
- There are a limited number of alternative character states (e.g. only A, G, C and T in DNA).
- Rates of evolution are sometimes high.





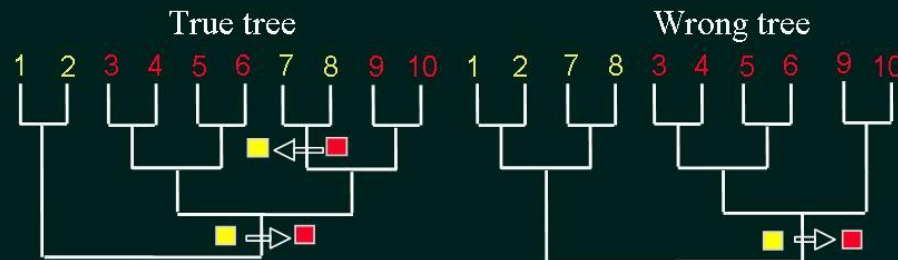
# Homoplasy

## Independent evolution (reversal)

- Homoplasy is similarity that is not homologous.
- It is the result of independent evolution (convergence, parallelism, reversal).

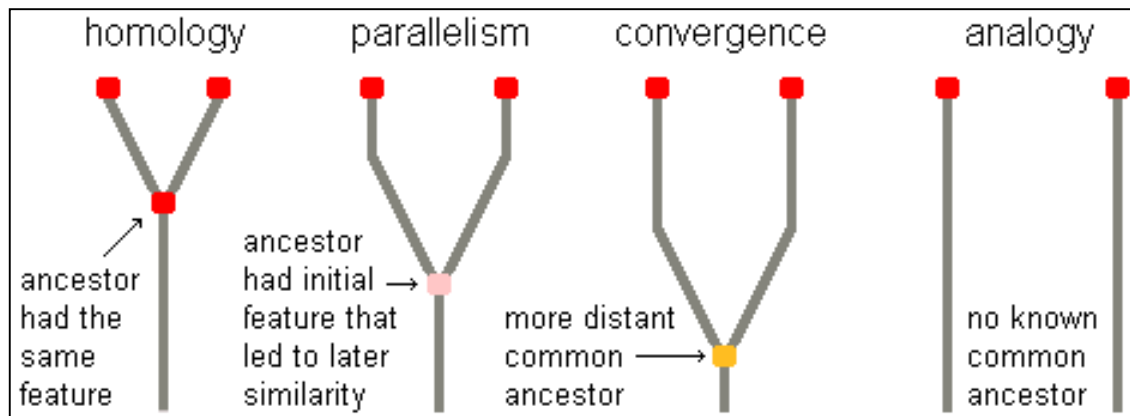
### Homoplasy - reversal

- Reversals are evolutionary changes back to an ancestral condition
- As with any homoplasy, reversals can provide misleading evidence of relationships



# Homoplasy and Long Branches

- Sequence data are unambiguous, but you can't detect **convergence** or **parallelism** by looking at the sequences, you have to have a phylogeny.
- For instance, at the same site, there may be two different transitions from **A** to **T**, but you **can't distinguish** them from the data.



# Molecular data and homoplasy

- Gene sequences represent character data.
- Characters are positions in the sequence (not all workers agree; some say one gene is one character).
- Character states are the nucleotides in the sequence (or amino acids in the case of proteins).

	260	*	280	*	300	*	320	
0841r :	CCCTTCAATTTTATT		---	---	AGAGTTT	AGGAGAAAT	AAGTATGTG	: 272
0992r :	CCCTCCAATTTTATTAG	CTTGCCCTACTCCCTTT	GGG	CACAGAGTTT	AGGAGAAAT	AAGTATGTG	G	: 213
3803r :	CCCTCCAATTTTATTAG	CTTGCCCTACTCCCTTT	GGG	CACAGAGTTT	AGGAGAAAT	AAGTATGTG	G	: 305
4062r :	CCCTCCAATTTTATTAG	CTTGCCCTACTCCCTTT	GGG	AACAGAGTTT	AGGAGAAAT	AAGTATGTG	G	: 319
3802r :	CCCTCCAATTTTATTAG	CTTGCCCTACTCCCTTT	GGG	CACAGAGTTT	AGGAGAAAT	AAGTATGTG	G	: 282
ph2f :	CCCTCCAATTTTATTAG	CTTGCCCTACTCCCTTT	GGG	CACAGAGTTT	AGGAGAAAT	AAGTATGTG	G	: 306
	CCTcCAATTTTATTAg ttgcctactcctttggg acAGAGTTTtagGAGAAATAAGTATGTG							

## Problems:

The probability that two nucleotides are the same just by chance mutation is 25%  
 what to do with insertions or deletions (which may themselves be characters)  
 homoplasy in sequences may cause alignment errors.



## 6. Gene Trees **vs.** Species Trees

---

- A **gene tree** is a phylogeny based on a single gene; it is the evolutionary history of that gene.
- A **species tree** (also called organismal phylogeny) is the “**true phylogeny**” of the group of taxa, or the evolutionary history of the group.
- **Gene trees and species trees** are often different, and gene trees are often different from one another.



# Phylogenetic Methods

## Analyses

---

- Understanding Tree
- Alignments
- Distances
- Clustering Methods
- Bootstrapping
- Likelihood Methods
- Parsimony



# Sequence Alignment

## Analyses

---

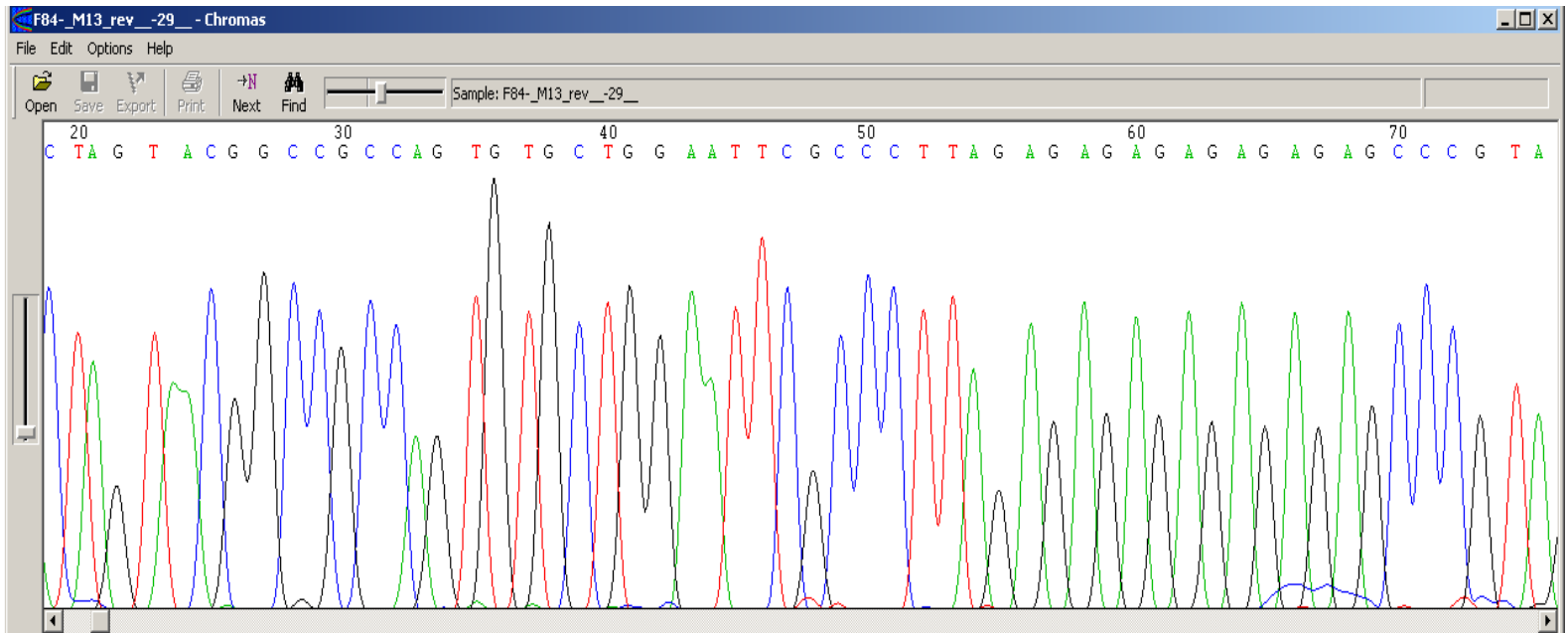
1. Choosing the sequence type
2. Alignment of sequence data
3. Search for the best tree
4. Evaluation of tree reproducibility

# Choosing the sequence type

## Assessing sequence quality

### Chromas

- Assess sequence quality, make corrections into the sequence



# Choosing the sequence type

## Assessing sequence quality

### Chromas

---

- Reverse and compliment the sequence
- Export sequences in plain text in Fasta, EMBL, GenBank or GCG format
- Copy the sequences in plain text or Fasta format into other software applications



# Choosing the sequence type

## Assessing sequence quality

**Bioedit**



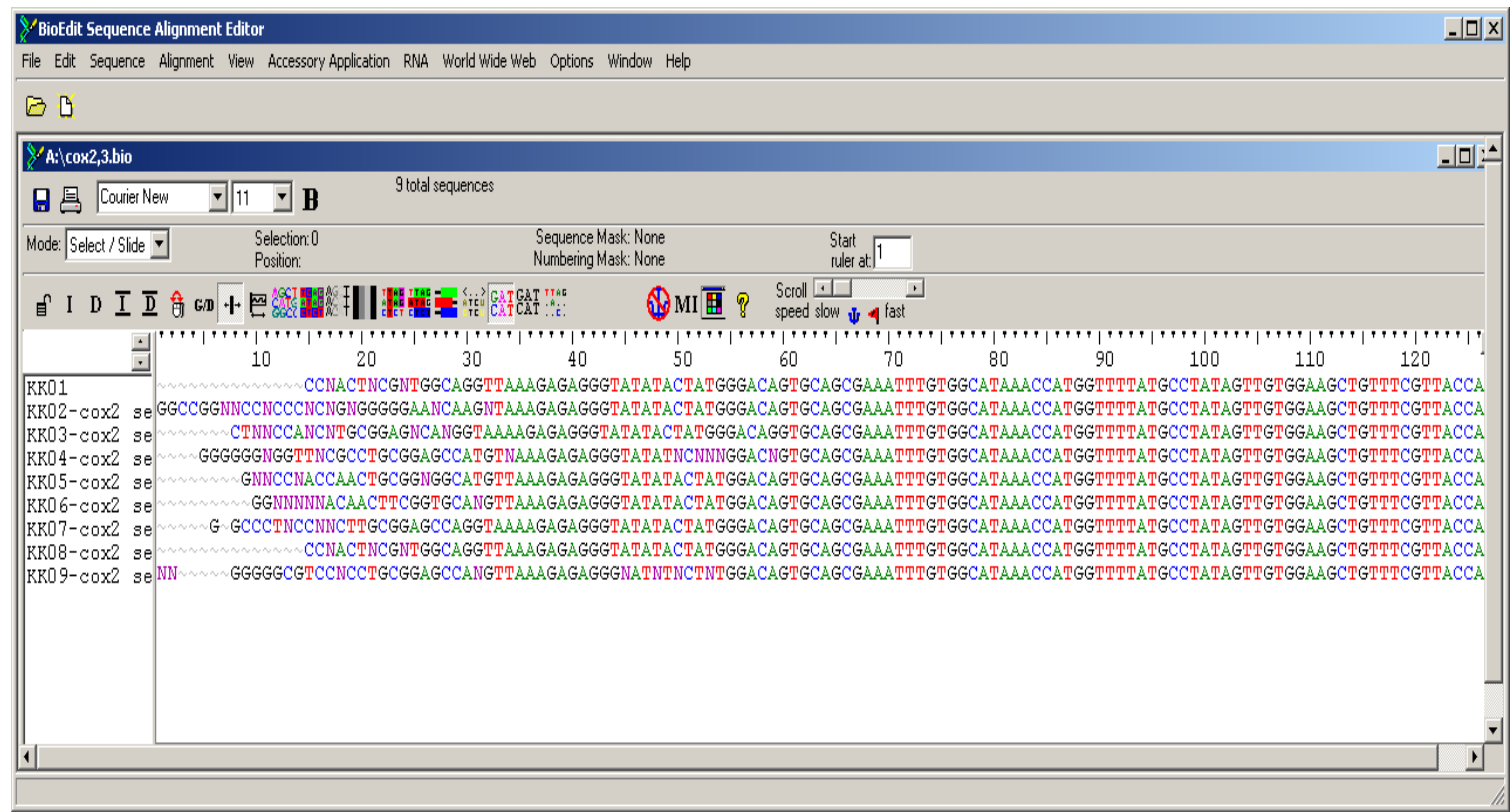
---

- Joining different parts of a sequence together (consensus sequence)
- Sequence alignments (manual vs. ClustalW)
- Alignments up to 20.000 sequences
- Export in GenBank, Fasta, or PHYLIP format

# Choosing the sequence type

## Assessing sequence quality

### Bioedit





# Sequence Alignment

Used in phylogenetic reconstruction

---

- An **alignment** is an hypothesis of positional **homology** between **bases/Amino Acids**.
  1. **Structural alignment**: establishing similarities in the 3D structure of protein molecules.
  2. **Sequence alignment**, in bioinformatics, arranging the sequences of DNA, RNA, or protein to identify similarities.
  3. **Alignment program**, software used in sequence alignment Engineering.



# Multiple Sequence Alignment vs. Pairwise Sequence Alignment

---

- **Pairwise Sequence Alignment:**
  - It is used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between **two biological sequences** (protein or nucleic acid).
- **Multiple sequence alignment:**
  - By contrast, **multiple sequence alignment** (MSA) is the alignment of **three or more biological sequences of similar length**.
  - From the output of MSA applications, homology can be inferred and the evolutionary relationship between the sequences studied.



# Alignment programs

## Used in phylogenetic reconstruction

---

- Alignment program, software used in sequence alignment Engineering. e.g. CLUSTAL, MUSCLE, MAFFT and other programs should all do a fine job of aligning 16S rRNA (or rDNA, the rRNA gene) especially within one family or genus of organisms.
- ClustalW2 is a general purpose DNA or protein multiple sequence alignment program for three or more sequences.
- For the alignment of two sequences please instead use our pairwise sequence alignment tools. E.g. EMBOSS Needle, PromoterWise; etc.



# Alignment programs

## Used in phylogenetic reconstruction

---

- Finding similar nucleotide composition for further analysis
- **Manually: can take weeks**
- ClustalW
- Check the alignment made by ClustalW
- You may have to go back to **Chromas** to check the sequences once again.



# Alignment programs

Used in phylogenetic reconstruction

---

- Finding similar nucleotide composition for further analysis
- Manually: can take weeks
- ClustalW
- Check the alignment made by ClustalW
- You may have to go back to Chromas to check the sequences once again.



# Alignment programs

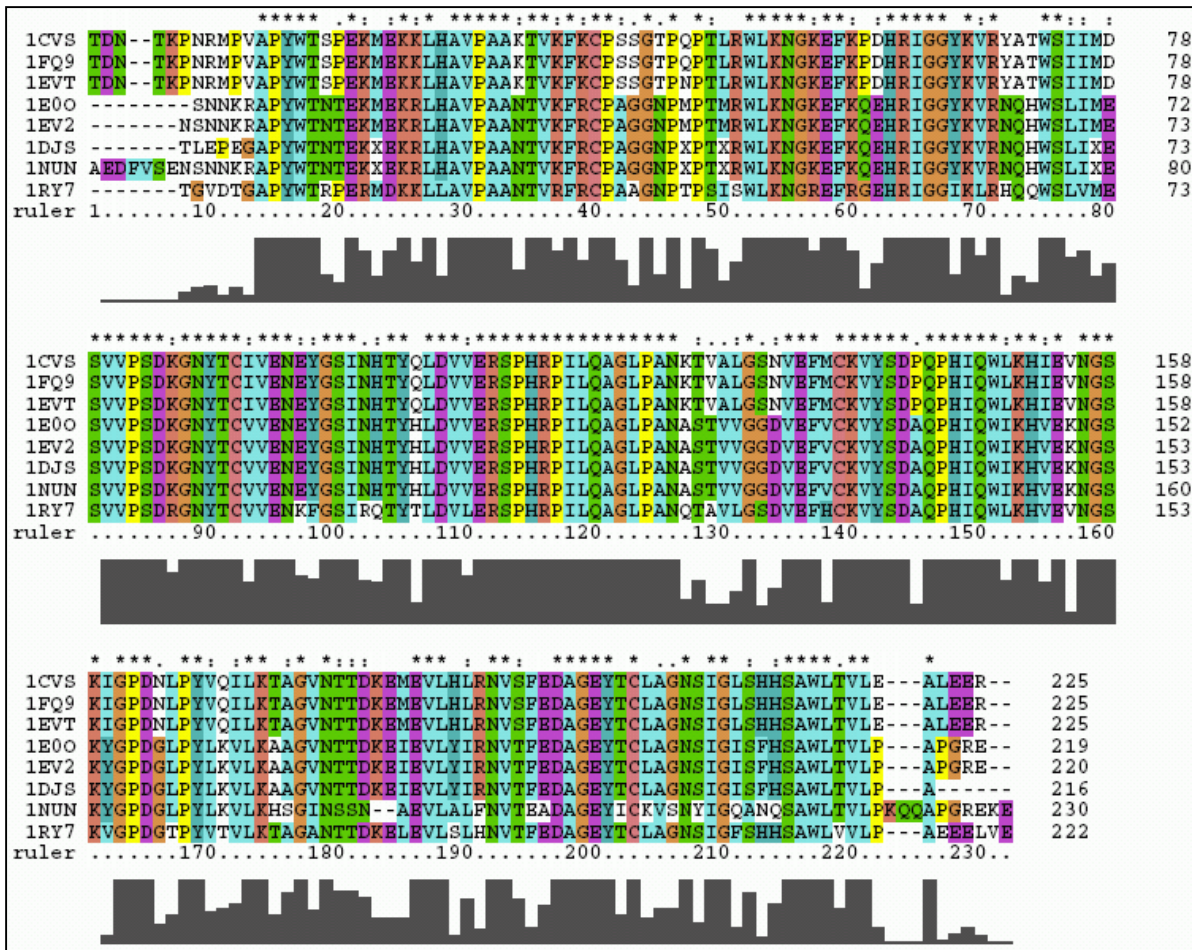
## Used in phylogenetic reconstruction

---

- If you are aligning protein-coding sequences, please note that **CLUSTALW** will not respect the codon positions and may insert alignment gaps within codons.
- For aligning **cDNA** or sequence data containing **codons**, we recommend that you **align the translated protein sequences** (see Aligning coding sequences via protein sequences).



# Alignment basics



# Multiple Alignment Results

- The homologous portions of each alignment were taken for tree-building.

```

gi|16271976_128895-128968      ---TCGGCGAGATAGGATTTGAACCTACGACCCACTGGTCCCAAAACCAGT 47
gi|15891923_1167246-1167319    ---TCGGGGCGGAGAGATTGGAACCTCCCGACCCCTCTGGTCCCAAAACCAGA 47
gi|16077068_96057-96133        TGGTCGGGAAGACAGGATTGGAACCTGCGACCCCATGGTCCCAAAACCATG 50
gi|16124256_1030929-1031005    TGGTCGGAGTGGCAGGATTTGAACCTGCGACCCCTGCGTCCCGAACGCAG 50
gi|7525012_66490-66563        ---TAGGGATGACAGGATTTGAACCCGTGACATTTTGTACCCAAAACAAA 47
gi|11466763_64229-64303        ---TAGGGATGACAGGATTTGAACCTGTGACATTTTGTACCCAAAACAAA 47
gi|11465652_4889-4962          ---TCGGGATAGCAGGATTTGAACCTGCGACATCCTGCTCCCAAGCAGG 47
gi|13449290_104457-104531      ---TCAAGGTGACAGGATTGGAACCTATGGCCCTCTGTACCCAAAACAGA 47
arabidopsis                    -----AAGGTGGCAGGATTGGAACCTATGGCCCTCTGTACCCGAAACAGA 45
gi|11465620_34675-34746        ---TCAAGATGGACAGATTTGAACTGACATTCCCTTGCACCCAAAGCAAG 47
gi|6226515_731-802            ---TCAGATAGGATAGACTCGAACTAAGTCTTTCTCCCAAGGAAG 47
gi|17981852_15956-16024        --ATCAGAGAAAAAGTCTTTAACTCCACCA---TTAGCACCCAAAGCTAA 45
humantrnapro                   ---TCAGAGAAAAAGTACTTGACTTTACCA---TCAGCGCCCAAGCTAA 44
gi|5835233_9903-9963           ---CAAGAGAAAAAGAAATTT--CTTTTTCA---TTAATCCCCAAAATTAA 42
gi|5834884_1-55                -----TCAGTAATAATATCT---TAGCAACCCAAATGCTA 32
gi|5834953_15350-15416        ---TCAAGAAGAAGGAGCTACTCCCCACCA---CCAGCACCCAAAGCTGG 44

```

\*\*\* \*\*

# Alignment

Alignment can be easy or difficult to detect, depending on the situation

An alignment involves hypotheses of positional homology between bases or amino acids

```
<-----(------HELIX 19-----)
<-----(22222222-000000-111111-00000-111111-0000-22222222
Thermus ruber      UCCGAUGC-UAAAGA-CCGAAG=CUCAA=CUUCGG=GGGU=GCGUUGGA
Th. thermophilus  UCCCAUGU-GAAAGA-CCACGG=CUCAA=CCGUGG=GGGA=GCGUGGGA
E.coli            UCAGAUGU-GAAAUC-CCC GGG=CUCAA=CCUGGG=AACU=GCAUCUGA
Ancyst.nidulans   UCUGUUGU-CAAAGC-GUGGGG=CUCAA=CCUCAU=ACAG=GCAAUGGA
B.subtilis        UCUGAUGU-GAAAGC-CCCCGG=CUCAA=CCGGGG=AGGG=UCAUUGGA
Chl.aurantiacus   UCGGCGCU-GAAAGC-GCCCCG=CUUAA=CGGGGC=GAGG=CGCGCCGA
match             **          ***          * * * * *
```

Alignment of 16S rRNA sequences from different bacteria

# Multiple Sequence Alignment

- The typical method uses 16S rRNA sequences: part of the 30S subunit.
- Easy to sequence.

Alignment can be easy or difficult

```
GCGGCCCA TCAGGTAGTT GGTGG
GCGGCCCA TCAGGTAGTT GGTGG
GCGTTCCA TCAGCTGGTT GGTGG
GCGTCCA TCAGCTAGTT GGTGG
GCGGCGCA TTAGCTAGTT GGTGA
*****
```

Easy

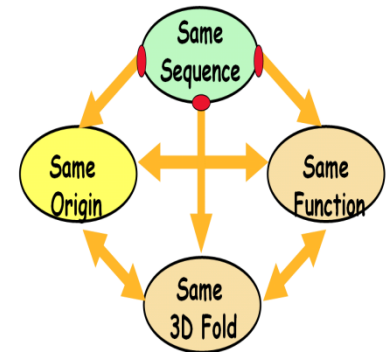
```
TTGACATG CCGGGG---A AACCG
TTGACATG CCGGTG--GT AAGCC
TTGACATG -CTAGG---A ACGCG
TTGACATG -CTAGGGAAC ACGCG
TTGACATC -CTCTG---A ACGCG
*****
```

Difficult due  
to insertions  
or deletions  
(indels)

# Sequence Similarity

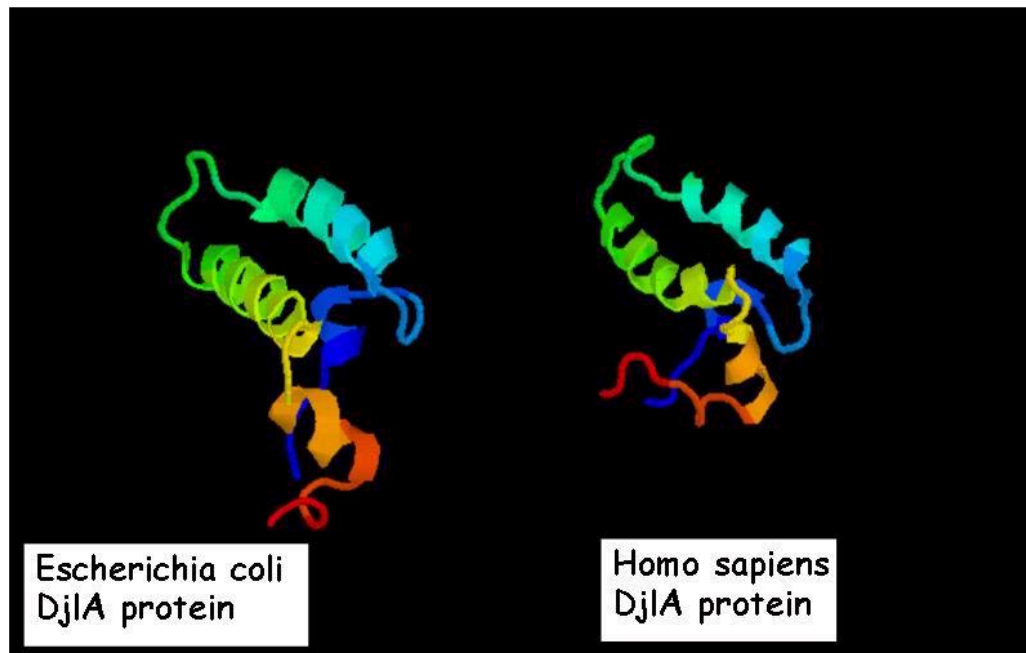
## Protein sequences and DNA sequences

- Two protein sequences with more than 25 % identity (over 100 amino acids ) are homologues.
- Two DNA sequences with more than 70 % identity (over 100 nucleotides) are homologues.
- Homologous sequences have:
  - A common ancestor (proteins and DNA)
  - A similar 3D structure (proteins)
  - Often a similar function (proteins)



# Protein alignment

Protein Alignment may be guided  
by Tertiary Structure Interactions





# Why do we need alignment?

---

- To **predict function** of proteins or RNAs
  - Complication: function evolves!
- To **predict structure** of proteins or RNAs
  - a.k.a. "Homology Modelling"
  - General ("X and Y have the same fold")
  - Specific (comparative modeling)
- To **identify conserved elements**
  - critical residues in proteins (active sites, binding pockets)
  - functional domains in proteins
  - protein-coding genes in genomes ("Comparative genomics")
- To study **molecular evolution**

In essence, "alignment" is the basic operation of **comparing sequences** to see if & how they are **related**.





# How to align

---

- Phylogenetic tree was developed by comparing molecular sequences:
- Align and compare homologous sequences.
- Number of positions that differ can be determined:(calculate a measure of difference between the sequences) = evolutionary distance (ED)
- Examine all possible branching arrangements and arrange to best fit the data.
- Organisms are clustered together based on similarity of sequences.





# Multiple Sequence Alignment Methods

## How to make alignments?

---

- **Visual inspection**
  - dotplots
- **Manual editing**
  - alignment editors
- **Automated methods**
  - scoring schemes
  - dynamic programming algorithms



# How to align

---

- Alignments can be global or local.
- **BLAST** calculates local alignments, for databank searches and to find pairwise similarities local alignments are preferred.



# BLAST

## Basic Local Alignment Search Tool

---

- BLAST is a tool for comparing one sequence with all the other sequences in a database.
- BLAST can compare:
  - DNA sequences
  - Protein sequences
- BLAST is more accurate for comparing protein sequences than for comparing DNA sequences.



# BLAST

## Basic Local Alignment Search Tool

---

- BLAST makes local alignments
  - It only aligns what can be aligned
  - It ignores the rest
- BLAST is very fast
  - You need only a few minutes to search Swiss-Prot on a standard PC
- Many BLAST flavors are available for a variety of tasks.



# BLAST

## Basic Local Alignment Search Tool

---

1. **BLASTing a Protein Sequence**
2. **BLASTing DNA Sequences**



# BLASTing a Protein Sequence

## **blastp & blastn**

### Choosing the right BLAST flavor for proteins

#### *What you want*

#### *The right flavor*

I want to find something about the function of my protein.

**blastp**, to compare your protein with other proteins contained in databases.

I want to discover new genes encoding simple proteins

**tblastn**, to compare your protein with DNA sequences translated into their six possible reading frames (3 on each strand).

# BLASTing a Protein Sequence

## Heat shock Protein (HSP90)

**blastp**

- With the HSP90 sequence in hand we used **Blastp** to find homologous sequences
- We were surprised to find a lot of homologous sequences across many species like Humans, Chicken, Pig, Mouse, Horse, Fish, Coral, fruit fly, mosquito, nematode, & even crops like rice, maize & tobacco.
- The first 100 matches had e-values ranging from 0 to e-153, so they were *\*very\** strong matches indicating a high degree of conservation of the protein through evolution.

ID	Name	Score	Evalue
304882	heat shock 90kDa protein 1, alpha [Homo sapiens] N...	1247	0.0
352285	heat shock protein 1, alpha [Mus musculus] NP_0346...	825	0.0
761972	heat shock protein 86 [Rattus norvegicus] NP_78693...	825	0.0
341493	heat shock protein 90A [Cricetulus griseus] AAA369...	817	0.0
609431	heat shock protein 90 - chicken	816	0.0
609432	heat shock protein 84 - mouse	745	0.0
449511	(Q9W6K6) Heat shock protein hsp90 beta [Salmo sala...	731	0.0
459017	heat shock protein hsp90 [Oncorhynchus tshawytscha...	730	0.0
446434	heat shock protein hsp90beta [Danio rerio] AAC2156...	729	0.0
361999	heat shock protein 90 [Rattus sp.] AAB23369.1 [S45...	724	0.0
460597	heat shock protein 90 [Pleurodeles waltl] AAA92343...	719	0.0
738604	90-kDa heat shock protein [Bombyx mori] BAB41209.1...	712	0.0
146263	Heat shock protein 83 CG1242-PA [Drosophila melano...	669	0.0
755572	heat shock protein 90 [Dendronephthya klunzingeri]...	662	0.0
226533	(P33126) Heat shock protein 82 [Oryza sativa (Rice)]	612	e-174
1888761	heat shock protein 82 - common tobacco (fragment)	612	e-174
252633	heat shock protein [Arabidopsis thaliana] CAA72513...	600	e-170
236351	(Q9XGF1) HSP80-2 [Triticum aestivum (Wheat)]	598	e-169
283559	(Q08277) Heat shock protein 82 [Zea mays (Maize)]	593	e-168
152674	heat shock protein 86 [Plasmodium falciparum] AAA6...	591	e-167
1899880	(Q8LLI6) Heat shock protein Hsp90 [Achlya ambisex...	579	e-164
245912	heat shock protein 90 [Lycopersicon esculentum] AA...	544	e-153



# Running blast

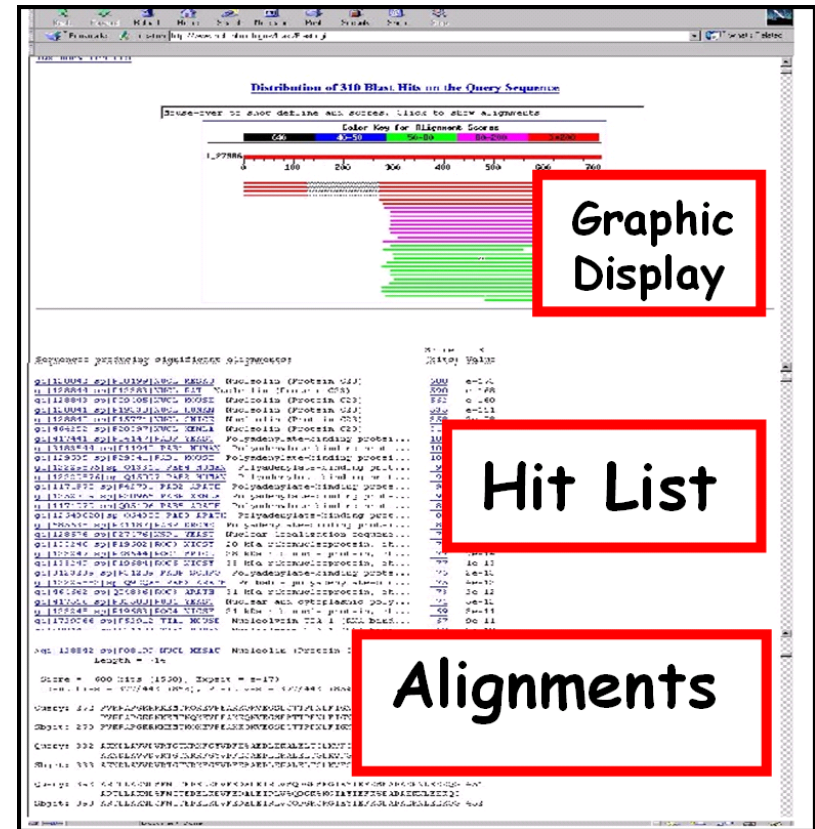
---

- Choose one of the public servers
  - NCBI [www.ncbi.nlm.nih.gov/blast](http://www.ncbi.nlm.nih.gov/blast)
  - EBI [www.ebi.ac.uk/blast](http://www.ebi.ac.uk/blast)
  - EMBNet [www.expasy.ch/blast](http://www.expasy.ch/blast)
- Select a database to search:
  - NR to find any protein sequence
  - Swiss-Prot to find proteins with known functions
  - PDB to find proteins with known structures
- Cut and paste your sequence
- Click the **BLAST** button



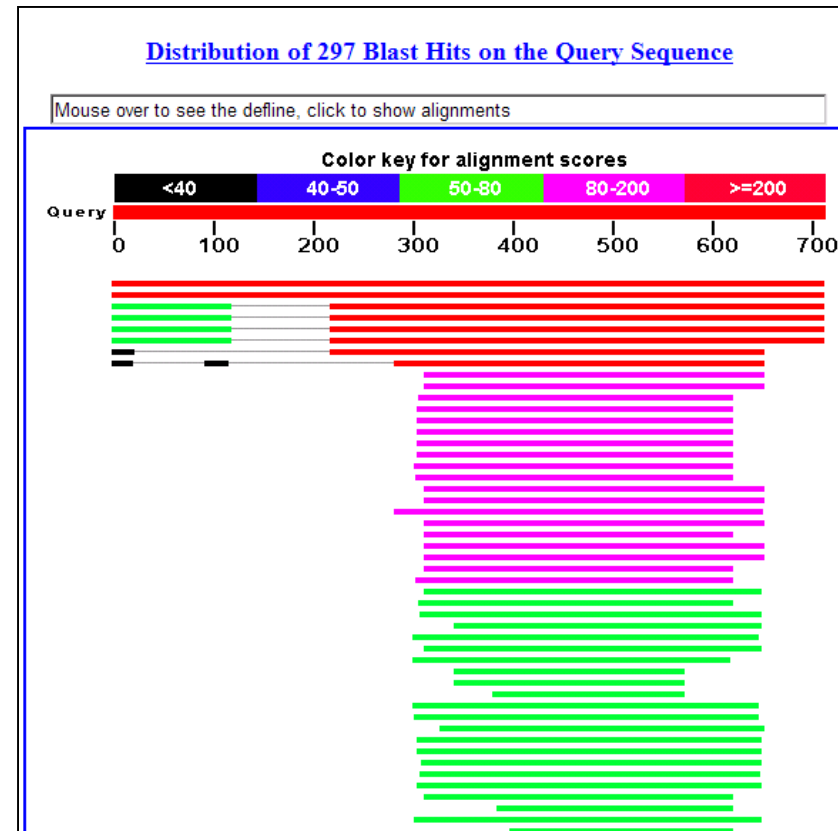
# Reading BLAST Output

- Graphic Display
  - Overview of the alignments
- Hit List
  - Gives the score of each match
- Alignments
  - Details of each alignment



# The Graphic Display

- The Horizontal Axis (0-700) corresponds to your protein (query).
- Color codes indicate that match's quality
  - Red: very good
  - Green: acceptable
  - Black: bad
- Thin lines join independent matches on the same sequence.



# The Hit List

- Sequence accession number
  - Depends on the database
- Description
  - Taken from the database
- Bit score
  - **High** bit score = **good** match
- E-Value
  - **Low** E-value = **good** match
- Links
  - Genome
  - Uniref, database of transcripts

Distance tree of results <a href="#">NEW</a> <a href="#">Related Structures</a>		Score (Bits)	E Value	
Sequences producing significant alignments:				
<a href="#">ref XP_516145.2 </a>	PREDICTED: hypothetical protein [Pan troglodyte	<a href="#">803</a>	0.0	<a href="#">G</a>
<a href="#">ref XP_001116949.1 </a>	PREDICTED: similar to nucleolin [Macaca mula	<a href="#">793</a>	0.0	<a href="#">UG</a>
<a href="#">sp Q4R4J7 NUCL_MACFA</a>	Nucleolin >dbj BAE00345.1  unnamed prote...	<a href="#">746</a>	0.0	
<a href="#">ref NP_005372.2 </a>	nucleolin [Homo sapiens] >sp P19338 NUCL_HUM...	<a href="#">744</a>	0.0	<a href="#">UG</a>
<a href="#">sp Q5RF26 NUCL_FONEY</a>	Nucleolin >emb CAH89631.1  hypothetical ...	<a href="#">739</a>	0.0	
<a href="#">gb AAAS9954.1 </a>	nucleolin	<a href="#">736</a>	0.0	<a href="#">G</a>
<a href="#">dbj BAC03738.1 </a>	unnamed protein product [Homo sapiens]	<a href="#">712</a>	0.0	<a href="#">UG</a>
<a href="#">ref XP_614626.2 </a>	PREDICTED: similar to nucleolin-related prot...	<a href="#">702</a>	0.0	
<a href="#">ref NP_072143.1 </a>	nucleolin-related protein [Rattus norvegicus...	<a href="#">701</a>	0.0	<a href="#">UG</a>
<a href="#">ref XP_850477.1 </a>	PREDICTED: similar to nucleolin-related prot...	<a href="#">681</a>	0.0	<a href="#">G</a>
<a href="#">ref XP_861643.1 </a>	PREDICTED: similar to nucleolin-related prot...	<a href="#">678</a>	0.0	<a href="#">G</a>
<a href="#">ref XP_861613.1 </a>	PREDICTED: similar to nucleolin-related prot...	<a href="#">678</a>	0.0	<a href="#">G</a>
<a href="#">sp P08199 NUCL_MESAU</a>	Nucleolin (Protein C23)	<a href="#">654</a>	0.0	
<a href="#">ref NP_036881.1 </a>	nucleolin [Rattus norvegicus] >sp P13383 NUC...	<a href="#">643</a>	0.0	<a href="#">UG</a>
<a href="#">gb AAH85751.1 </a>	Nucleolin [Rattus norvegicus]	<a href="#">642</a>	0.0	<a href="#">UG</a>
<a href="#">ref XP_861582.1 </a>	PREDICTED: similar to nucleolin-related prot...	<a href="#">642</a>	0.0	<a href="#">G</a>
<a href="#">gb AAA36966.1 </a>	nucleolin, C23	<a href="#">641</a>	0.0	
<a href="#">pir JH0148</a>	nucleolin - rat	<a href="#">639</a>	0.0	
<a href="#">dbj BAC27474.1 </a>	unnamed protein product [Mus musculus]	<a href="#">637</a>	0.0	<a href="#">UG</a>
<a href="#">gb AAH05460.1 </a>	Nucleolin [Mus musculus]	<a href="#">632</a>	2e-179	<a href="#">UG</a>
<a href="#">ref NP_035010.3 </a>	nucleolin [Mus musculus] >sp P09405 NUCL_MOU...	<a href="#">632</a>	2e-179	<a href="#">G</a>
<a href="#">dbj BAE38940.1 </a>	unnamed protein product [Mus musculus]	<a href="#">631</a>	4e-179	<a href="#">UG</a>
<a href="#">dbj BAE36484.1 </a>	unnamed protein product [Mus musculus]	<a href="#">631</a>	4e-179	<a href="#">UG</a>
<a href="#">dbj BAE40448.1 </a>	unnamed protein product [Mus musculus] >dbj B...	<a href="#">631</a>	5e-179	<a href="#">UG</a>
<a href="#">dbj BAC26311.1 </a>	unnamed protein product [Mus musculus]	<a href="#">628</a>	3e-178	<a href="#">UG</a>

# Partial 16S rDNA sequence alignment

## *Xanthomonas* and *Stenotrophomonas* spp.

- Partial 16S rDNA sequence alignment of 13 *Xanthomonas*- and *Stenotrophomonas* type-strains and seven *X. translucens* pv. *graminis* (X.t.g.) isolates.
- Shading indicates sequence differences to the X.t.g. type-strain.
- Bars mark the diagnostic PCR primer site characteristic for the X.t.g. group.
- Numbers on top denote position in the *E. coli* reference sequence.

	58	105
<i>Stenotrophomonas maltophilia</i>	CAAGTCGAACGGCAGCAG-GAGAGCTTGCTCT-CTGGGTGGCGAGTGG	
<i>X. bromi</i>	CAAGTCGMRCGGCAGCACAGTAAGARCTTKCTCTTATGGGTGGCGAGTGG	
<i>X. cassavae</i>	CAAGTCGAACGGCAGCACAGTAAGAGCTTGCTCTTATGGGTGGCGAGTGG	
<i>X. oryzae</i> pv. <i>oryzae</i>	CAAGTCGAACGGCAGCACAGTAAGAGCTTGCTCTTATGGGTGGCGAGTGG	
<i>X. campestris</i> pv. <i>campestris</i>	CAAGTCGAACGGCAGCACAGTAAGAGCTTGCTCTTATGGGTGGCGAGTGG	
<i>X. sacchari</i>	CAAGTCGAMCGGCAGCACAG-GAGAGCTTGCTCT-CTGGGTGGCGAGTGG	
<i>X. albilineans</i>	CAAGTCGAACGGCAGCACAGTGGTAGCAATACC--ATGGGTGGCGAGTGG	
<i>X. hyacinthi</i>	CAAGTCGAACGGCAGCACAGTGGTAGCAATGCC--ATGGGTGGCGAGTGG	
<i>X. melonis</i>	CAAGTCGAACGGCAGCACAGTGGTAGCAATACC--ATGGGTGGCGAGTGG	
<i>X. translucens</i> pv. <i>translucens</i>	CAAGTCGAACGGCAGCACAGTGGTAGCAATACC--ATGGGTGGCGAGTGG	
<i>X. translucens</i> pv. <i>poae</i>	CAAGTCGAACGGCAGCACAGTGGTAGCAATACC--ATGGGTGGCGAGTGG	
<i>X. translucens</i> pv. <i>arrhenatheri</i>	CAAGTCGAACGGCAGCACAGTGGTAGCAATACC--ATGGGTGGCGAGTGG	
X.t.g. 25	CAAGTCGAACGGCAGCACAGTGGTAGCAATACC--ATGGGTGGCGAGTGG	
<i>X. translucens</i> pv. <i>graminis</i>	CAAGTCGAACGGCAGCACAGTGGTAGCAATACC--ATGGGTGGCGAGTGG	
X.t.g. 3	CAAGTCGAACGGCAGCACAGTGGTAGCAATACC--ATGGGTGGCGAGTGG	
X.t.g. 10	CAAGTCGAACGGCAGCACAGTGGTAGCAATACC--ATGGGTGGCGAGTGG	
X.t.g. 12	CAAGTCGAACGGCAGCACAGTGGTAGCAATACC--ATGGGTGGCGAGTGG	
X.t.g. 21	CAAGTCGAACGGCAGCACAGTGGTAGCAATACC--ATGGGTGGCGAGTGG	
X.t.g. 23	CAAGTCGAACGGCAGCACAGTGGTAGCAATACC--ATGGGTGGCGAGTGG	
X.t.g. 29	CAAGTCGAACGGCAGCACAGTGGTAGCAATACC--ATGGGTGGCGAGTGG	

# Sequencing and alignment of partial 16S rRNA region

## Phylogenetic tree

---

- Each bacterial sequence was subjected to software analysis ([www.ebi.ac.uk](http://www.ebi.ac.uk) and <http://itol.embl.de/>) to draw phylogenetic tree.

# Sequencing and alignment of partial 16S rRNA region

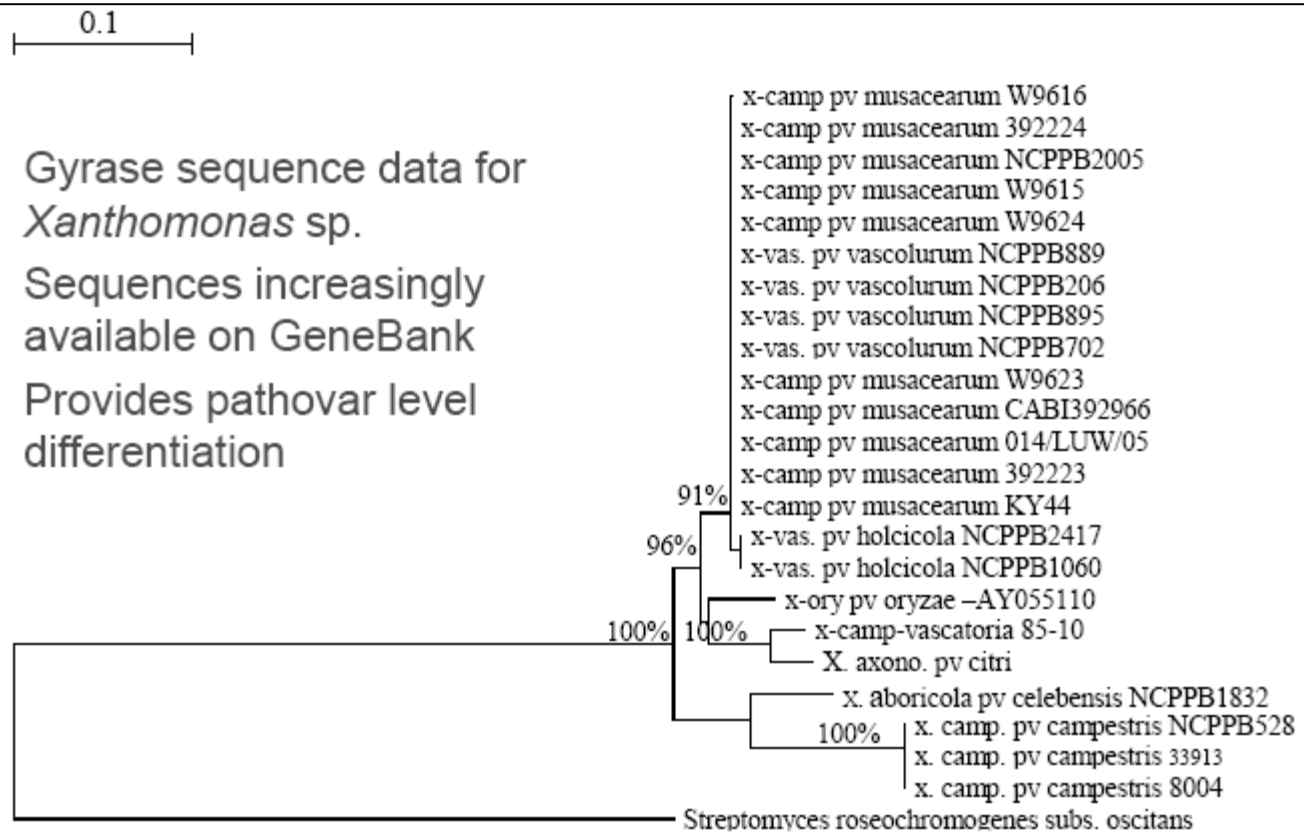
## Comparison of the 16SrRNA sequences

- Comparisons of the sequence between different species suggest the degree to which they are related to each other.
- Differences in the DNA base sequences between different organisms can be determined quantitatively, such that a phylogenetic tree can be constructed to illustrate probable evolutionary relatedness between the organisms.
- As the 16SrRNA is so highly conserved organisms are classified as separate species if:
- their sequences show less than 98% homology, and
- are classified as different genera if their sequences show less than 93% identity.

# Sequence alignment

## Gyrase sequence data for *Xanthomonas* sp.

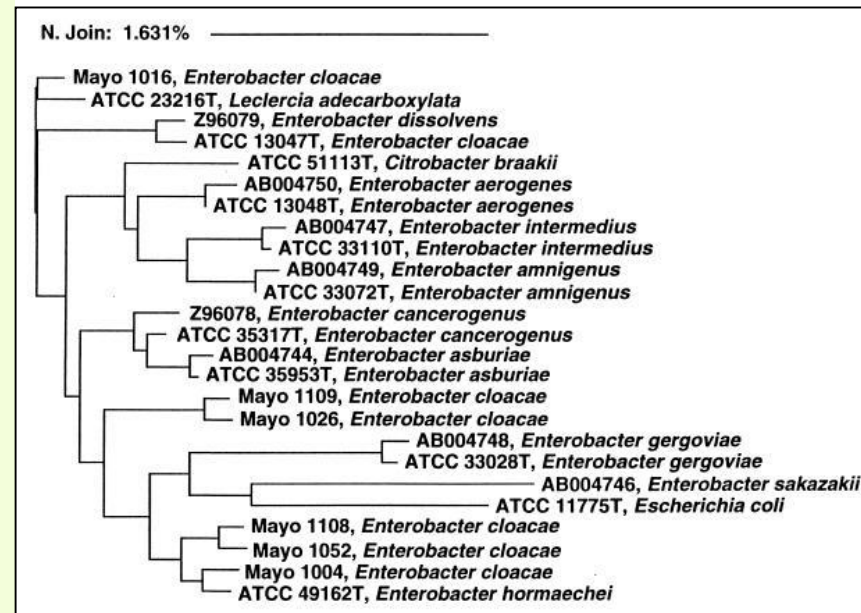
- Gyrase sequence data for *Xanthomonas* sp.
- Sequences increasingly available on GeneBank
- Provides pathovar level differentiation



# Identification of *Enterobacter* spp.

## Based on sequence analysis of different regions of the 16S rRNA gene

- Neighbor-joining analysis of DNA sequences from several *Enterobacter* spp.
- Phylogenetic analysis was based on full 16S rRNA gene sequences, and the scale reflects relative phylogenetic distance.
- Isolates with names beginning with Mayo were evaluated in this study.
- Isolates with names beginning with accession numbers were retrieved from GenBank.
- The remaining isolates, whose names begin with ATCC numbers, were type strains stored in the MicroSeq database.







# The E-Values

---

- E-value means **expectation value**.
- The E-value is the measure most commonly **used for estimating sequence similarity**.
- How many times is a match at least as good expected to happen by chance?
  - This estimate is based on the similarity measure.
- If a match is highly unexpected, it probably results from something other than chance
  - Common origin is the most likely explanation.
  - This is how homology is inferred.



# Which Value for Your E-Values ?

---

- Low E-value  $\Leftrightarrow$  good hit
  - 1 = bad e-Value
  - $10^{e-3}$  = borderline E-value
  - $10^{e-4}$  = good E-value
  - $10^{e-10}$  = very good E-value
- E-values lower than  $10^{e-4}$  indicate possible homology.
- E-values higher than  $10^{e-4}$  require extra evidence to support homology.



# Why Use E-Values?

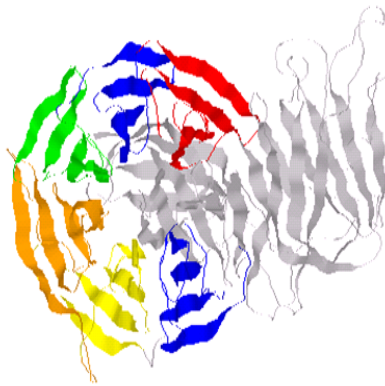
---

- **E-values** make it possible to compare alignment of different lengths.
- E-values are used by most sequence comparison programs:
  - PSI-BLAST
  - Domain Search
  - FASTA
- E-values always have the same meaning
  - You can compare the output of different programs

# Structural Analysis with BLAST

## *What you need*

### Predicting a Protein 3D structure



## *The BLAST way*

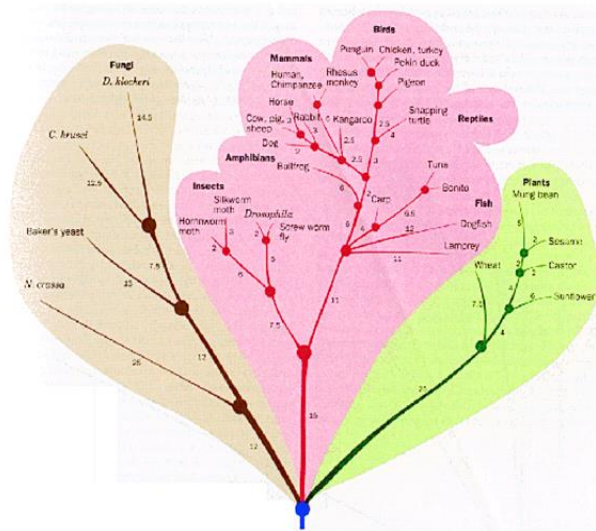
Use blastp to BLAST your protein against PDB (the database of protein structure). If you get a good hit, (more than 25% identity), then you know that your protein and this good hit have a similar 3D structure.

The complicated alternative is to do homology modeling, Xray or NMR analysis of your protein.

# Gathering Members of a Protein Family

## What you need

### Finding a protein family members



## The BLAST way

Use blastp (or its more powerful cousin Psi-BLAST) and run it on NR the non-redundant protein family. Once you have all the members of the family, you can make a multiple sequence alignment (see Chapter 11) and draw a phylogenetic tree.

The Complicated alternative is to use PCR for Cloning your sequences



# BLASTing DNA Sequence

---

- The BLAST program you need depends on your DNA sequence:
  - Coding DNA
  - Non Coding DNA
- BLASTing DNA sequences is less accurate than BLASTing protein sequences.
- If your sequence is coding, **blastx** and **tblastx** will translate it for you on its 6 possible reading frames.



# Asking the Right Question with BLAST

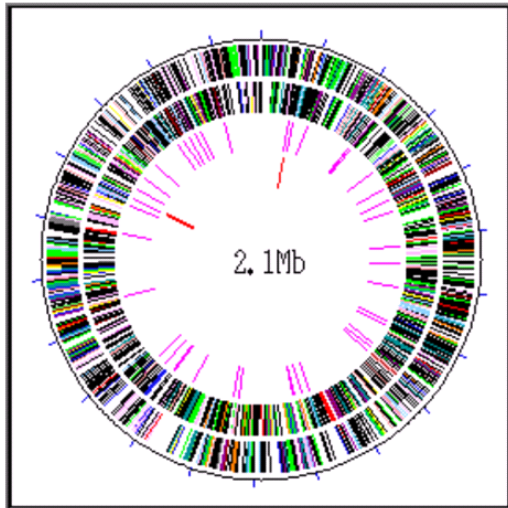
## Choosing the right flavor of BLAST for DNA

<i>Question</i>	<i>Answer</i>
Am I interested in non-coding DNA?	Yes: use <b>blastn</b> . Never forget that blastn is only for closely related DNA sequences (more than 70 percent identical)
Do I want to discover new Proteins?	Yes: use <b>tblastx</b> .
Do I want to discover proteins encoded in my query DNA sequence?	Yes: use <b>blastx</b>
Am I unsure of the quality of my DNA?	Yes: use <b>blastx</b> if you suspect your DNA sequence is coding for a protein but that it may contain sequencing errors.

# Gene-Hunting with BLAST

## *What you need*

### Finding genes in a genome



## *The BLAST way*

Cut your genome sequence in little (2-5kb) overlapping sequences. Use blastx to BLAST each piece of genome against NR (the Non Redundant Protein database). This works better if you have no introns (bacteria).

**The complicated alternative is to run a gene prediction software.**





## 4. Phylogenetic Trees

---

- In phylogenetic studies, the most convenient way of presenting evolutionary relationships among a group of organisms is the phylogenetic tree.



# Phylogenetic Trees

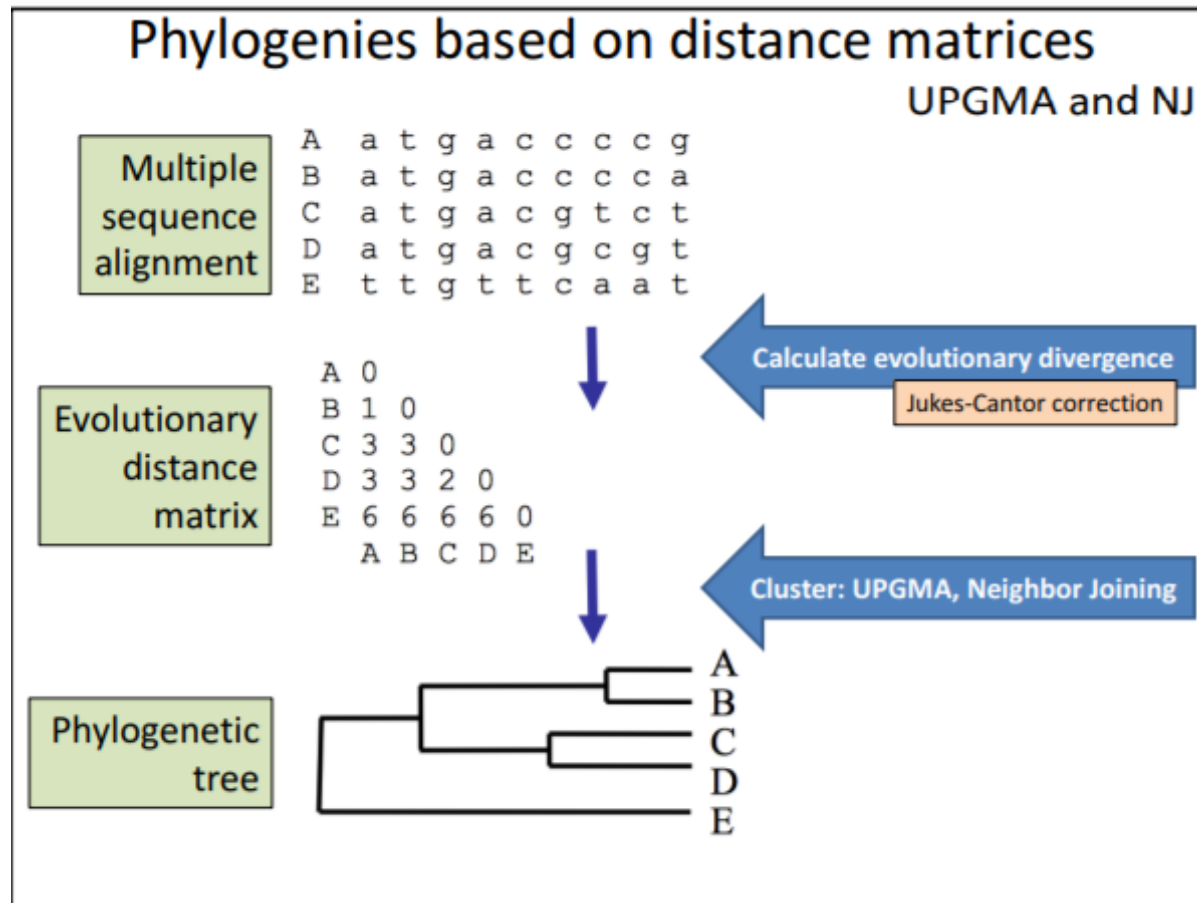
## How to construct a tree with UPGMA?

---

- Prepare a distance matrix
- Repeat step 1 and step 2 until there are only two clusters
- Step 1:
  - Cluster a pair of leaves (taxa) by shortest distance
- Step 2:
  - Recalculate a new average distance with the new cluster and other taxa, and make a new distance matrix

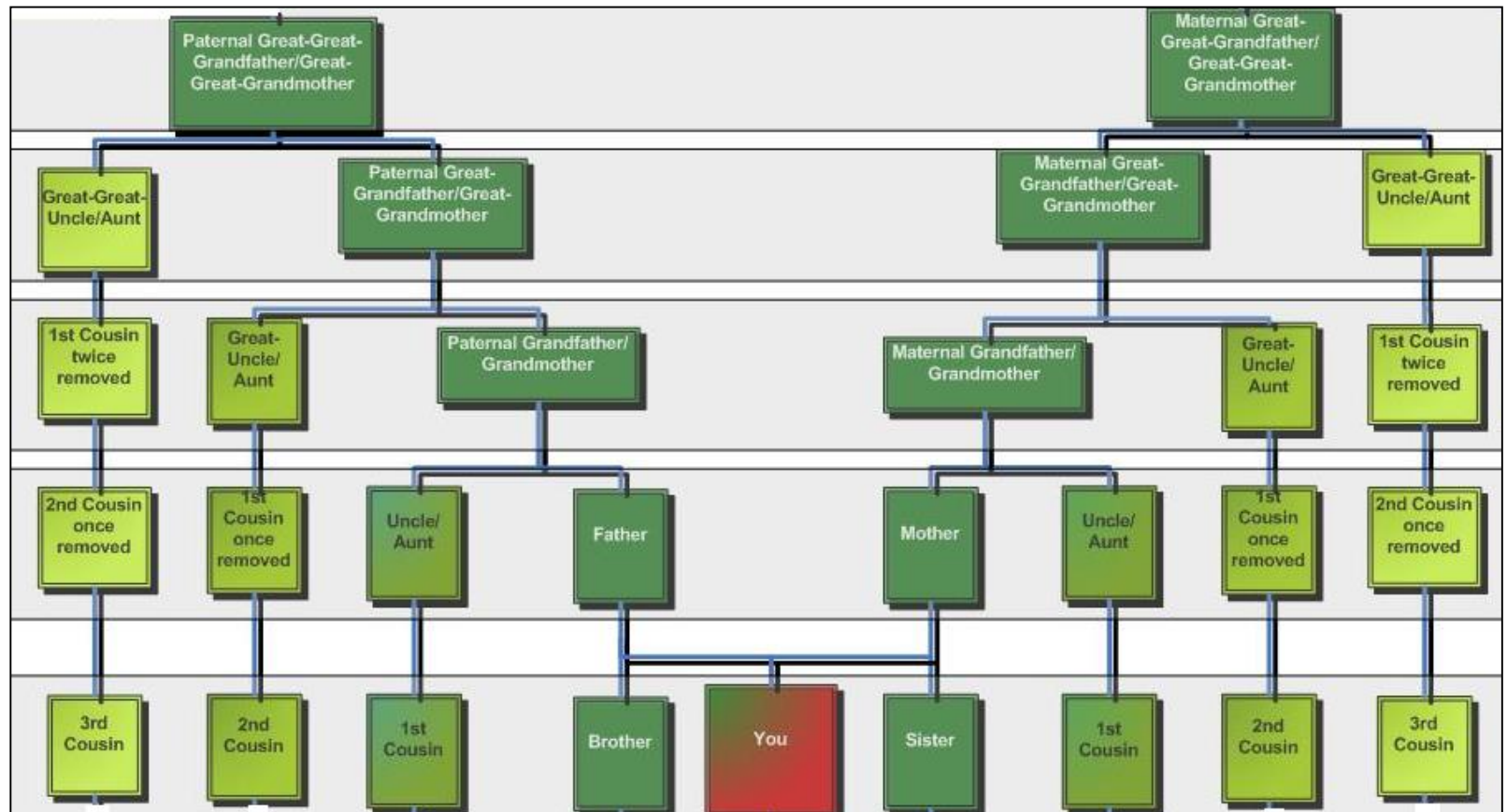
# Phylogenetic Trees

Alignment and drawing the tree based on distance matrices(UPGMA and NJ)



# Phylogenetic Trees

Phylogenies explain genealogical relationships





# Phylogenetic Trees

## Tree Terminology

---

- **Leaves(taxa)**: current organisms, species, or genomic sequence.
- **Node**: A branch point in a tree (a presumed ancestral OTU).
- **Branch**: Relationship between organisms, species, or genomic sequence. Defines the relationship between the taxa in terms of descent and ancestry.
- **Topology**: The branching patterns of the tree.
- **Branch length** (scaled trees only): Represents the number of changes that have occurred in the branch. Evolutionary time.
- **Root**: The common ancestor of all taxa. Origin of evolution.
- **Clade**: A group of two or more taxa or DNA sequences that includes both their common ancestor and all their descendents.
- **Operational Taxonomic Unit (OTU)**: Taxonomic level of sampling selected by the user to be used in a study, such as individuals, populations, species, genera, or bacterial strains.



# Phylogenetic Trees

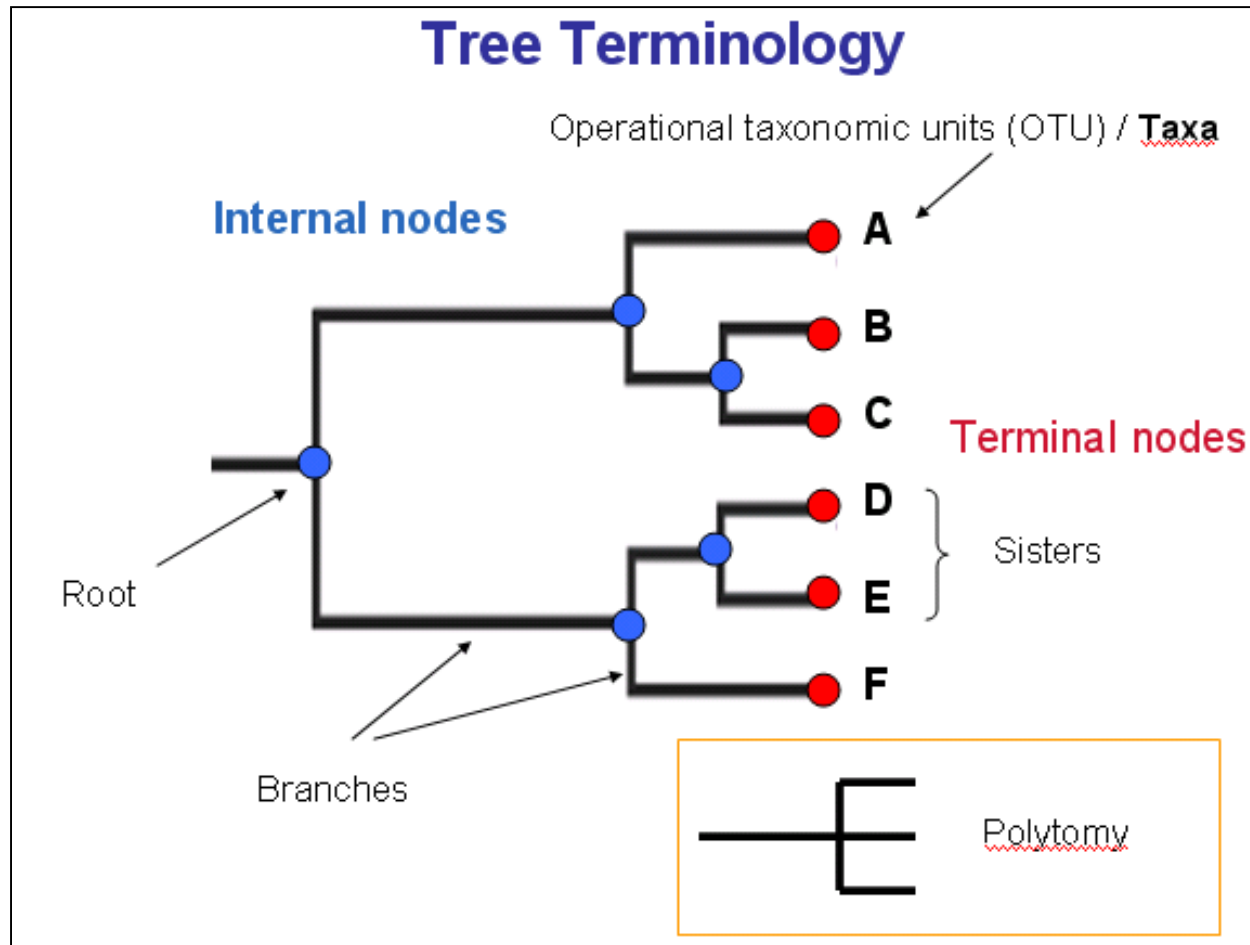
**Phylogenies explain genealogical relationships**

---

1. **Topology** (branching order)
2. **Branch lengths** (indication of genetic change)
3. **Nodes**
  - i. **Tips** (sampled sequences known as taxa)
  - ii. **Internal nodes** (hypothetical ancestors)
  - iii. **Root** (oldest point on the tree)
4. **Confidence** (bootstraps/probabilities)

# Phylogenetic Trees

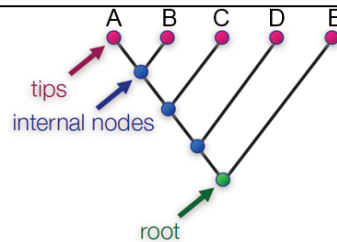
## Tree Terminology



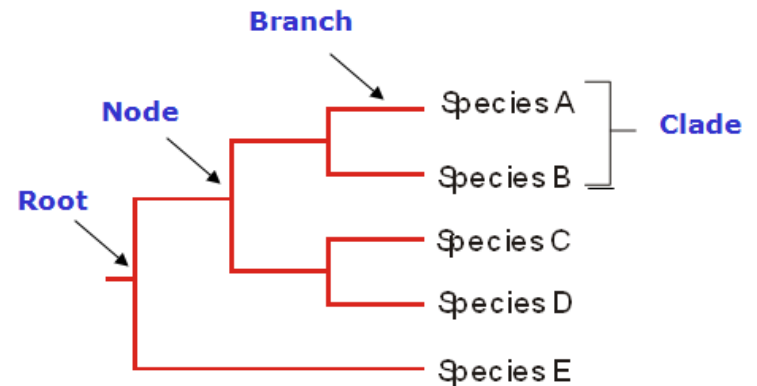
# Phylogenetic Trees

## Topology

### Three types of nodes



- Nodes occur at the ends of branches
- There are three types of nodes:
  - Tips** (sampled sequences known as taxa)
  - Internal nodes** (hypothetical ancestors)
  - Root** (oldest point on the tree)



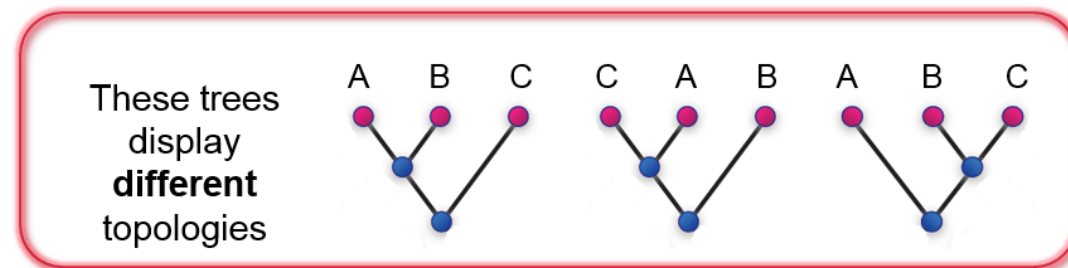
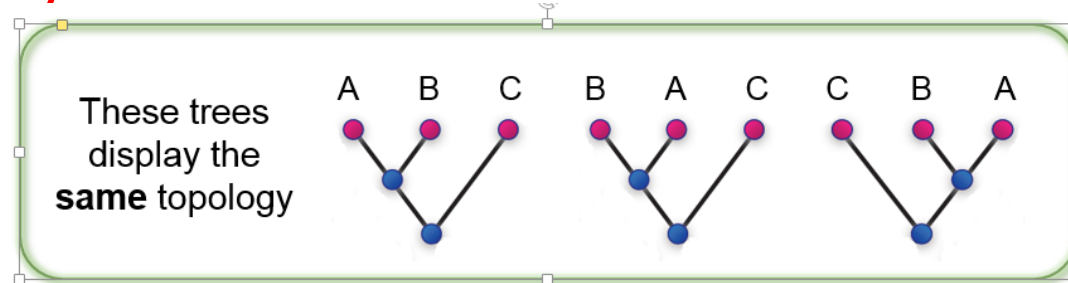


# Phylogenetic Trees

## Topology

## Branching Order

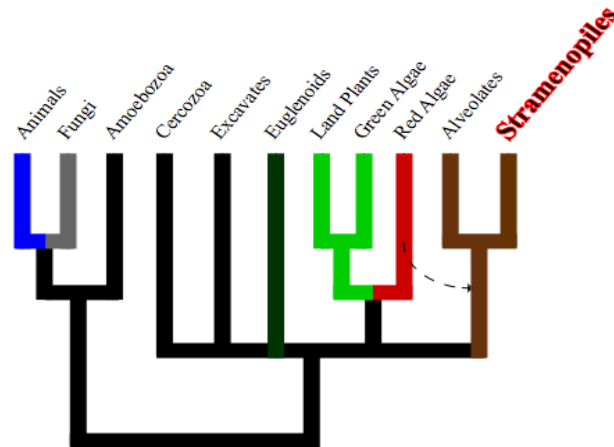
- The topology describes the branching structure of the tree, which indicate patterns of relatedness.
- That is, it shows which **species share more common ancestry than which others.**



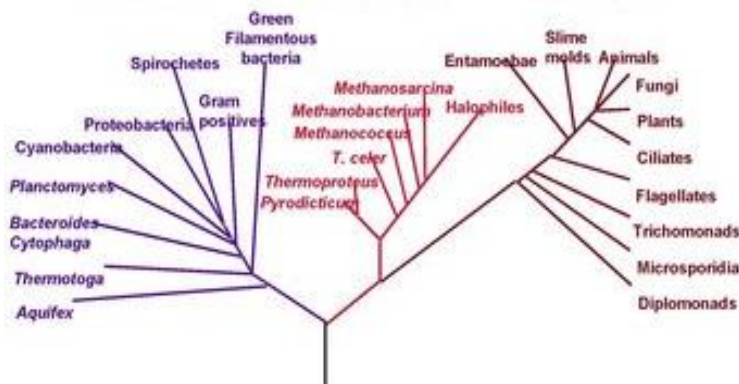
# Phylogenetic Trees

## Topology

Trees can be represented in several forms



Rectangular cladogram



Slanted cladogram



# Phylogenetic Trees

Trees can be unrooted or rooted

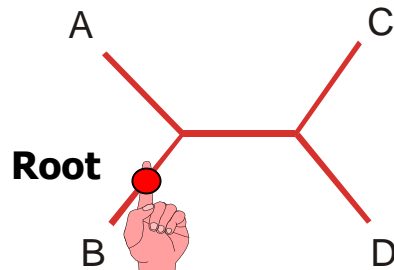
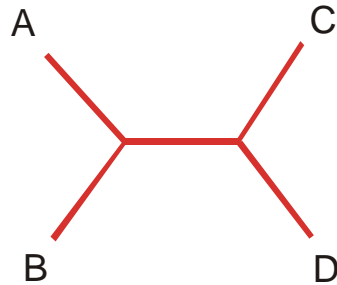
---

- Rooted trees: Has a root that denotes common ancestry.
- Unrooted trees: Only specifies the degree of kinship among taxa but not the evolutionary path.

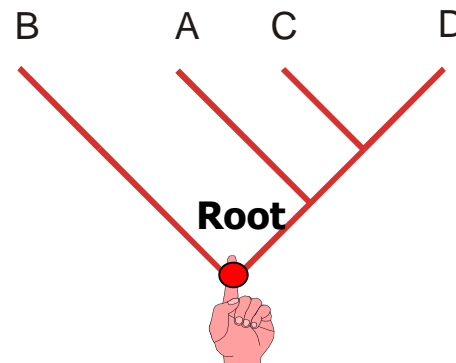
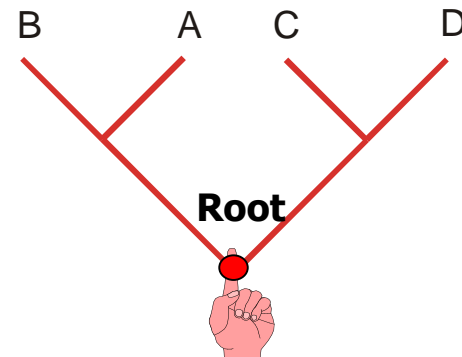
# Phylogenetic Trees

Trees can be unrooted or rooted

Unrooted tree



Rooted tree

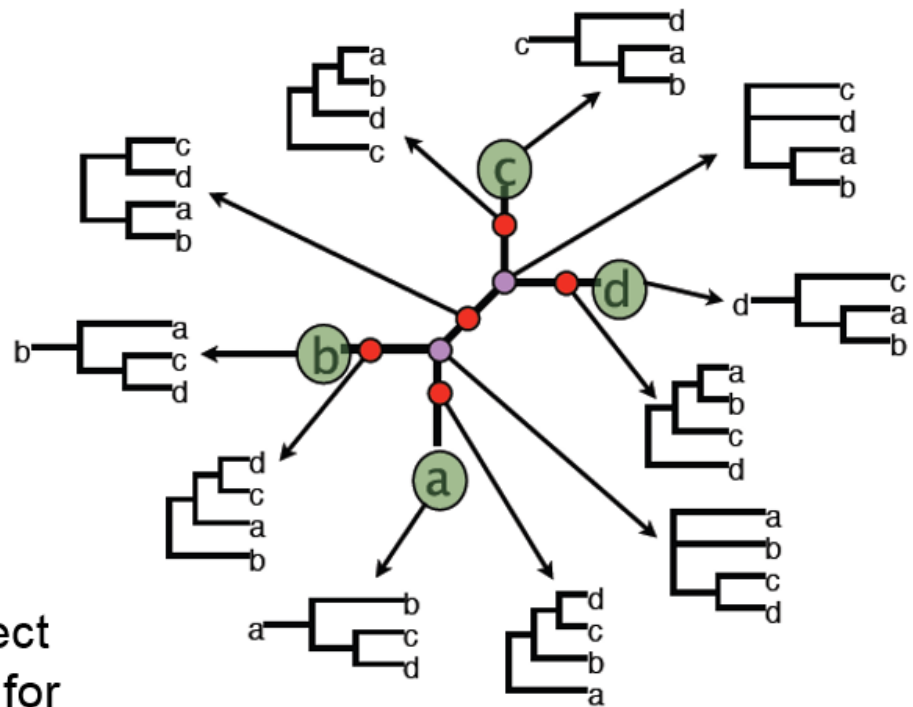


# Phylogenetic Trees

**There are multiple rooted tree topologies for any given unrooted tree**

Unrooted trees can be rooted on their:

- **branches**
  - **interior nodes**
  - **terminal nodes**
- } \*
- Most tree-building methods produce unrooted trees
  - Identifying the correct root is often critical for interpretation!



*Figure Aiden Budd*

# Phylogenetic Trees

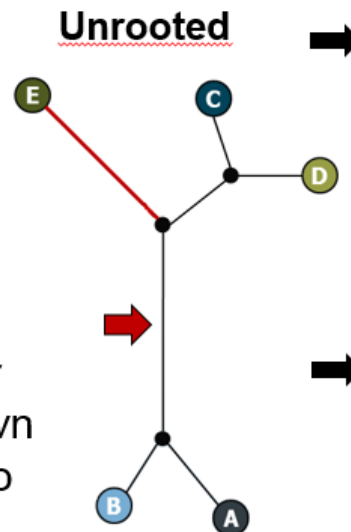
## How to root a tree

- **Midpoint rooting**

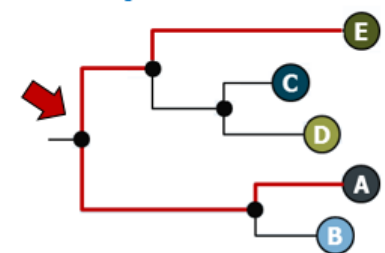
- Assume constant evolutionary rate
- Often not the case!

- **Outgroup rooting**

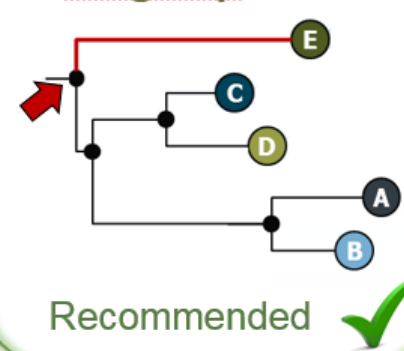
- The **outgroup** is one or more taxa that are known to have diverged prior to the group being studied
- The node where the **outgroup** lineage joins the other taxa is the root



### Midpoint rooted



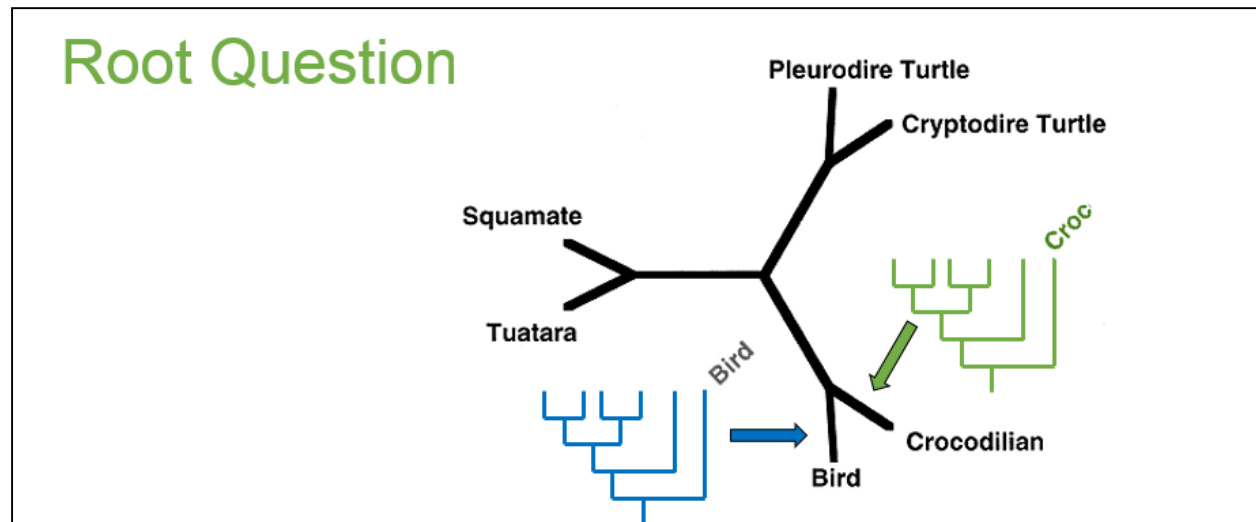
### Outgroup rooted



# Phylogenetic Trees

## Root Question

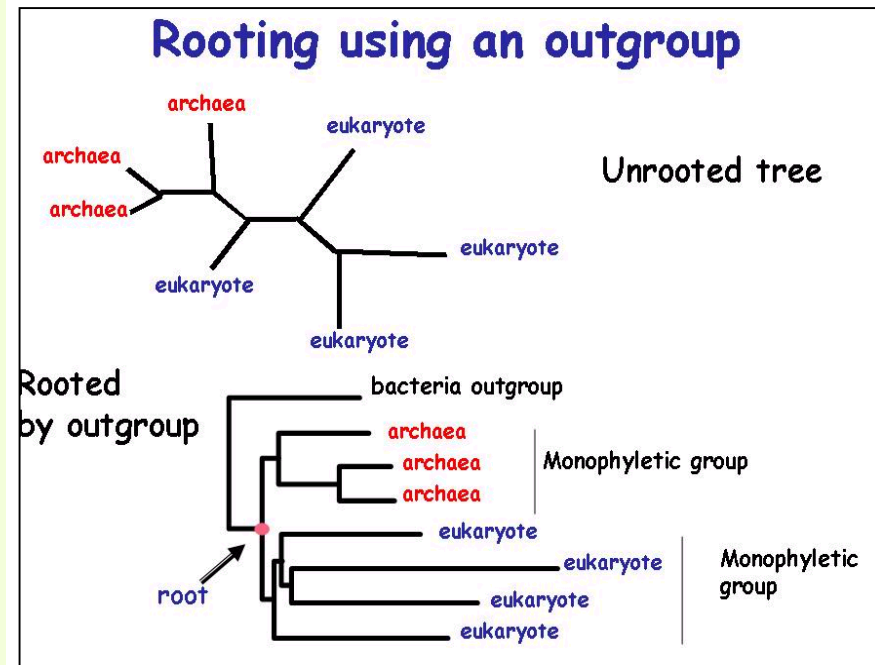
- This tree shows a **cladogram** i.e. the branch lengths do not indicate genetic change.
- Indicate any root positions where **bird and crocodile are not sister taxa** (each other's closest relatives).



# Phylogenetic Trees

## Rooting via outgroups

- In cladistics or phylogenetics, an **outgroup** is a group of organisms that serve as a **reference group** when determining the evolutionary relationship among three or more monophyletic groups of organisms.
- The outgroup is used as a point of comparison for the ingroup.
- Trees are rooted by the choice of outgroup.



The red circle represents the root of tree.  
Monophyletic groups (clades): Contain species which are more closely related to each other than to any outside of the group.





# Phylogenetic Trees

## Possible evolutionary trees

- As the number of taxa increases, the number of possible trees explodes.

Number of taxa	Number of possible binary trees
3	1
4	15
10	34 459 425
20	8 200 794 532 637 891 559 375
500	$1.0084917894 \times 10^{1290}$



# Phylogenetic Trees

## Possible evolutionary trees

Taxa ( $n$ )	rooted $(2n-3)!/(2n-2(n-2)!)$	unrooted $(2n-5)!/(2n-3(n-3)!)$
2	1	1
3	3	1
4	15	3
5	105	15
6	954	105
7	10,395	954
8	135,135	10,395
9	2,027,025	135,135
10	34,459,425	2,027,025



# Phylogenetic Trees

## How many trees can we build?

**Table 6.1.1** Number of Unrooted and Rooted Trees for 2 to 10 Sequences

No. of sequences	No. of unrooted trees	No. of rooted trees
2	1	1
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10,395
8	10,395	135,135
9	135,135	2,027,025
10	2,027,025	34,459,425

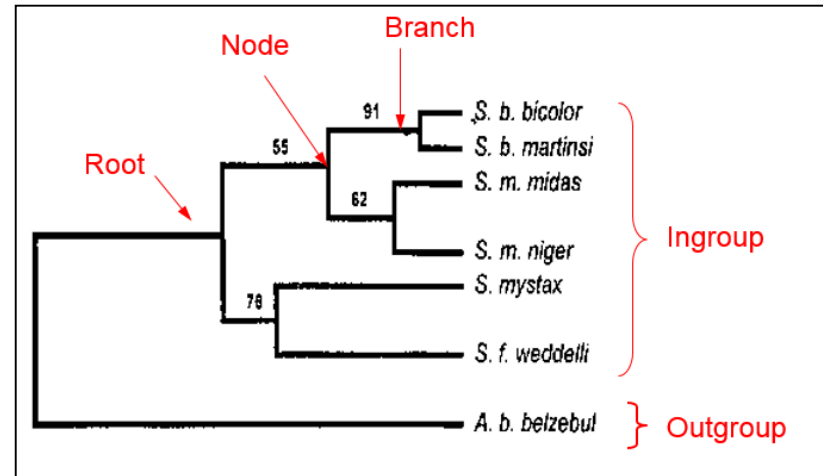
20 sequences = 8,200,794,532,637,891,559,000 possible trees.  
For high number of sequences (typically >15) no guarantee to find best tree.

# Phylogenetic Trees

## Ingroup vs. outgroup

### Choice of outgroup

- The **outgroup** should be a taxon known to be less closely related to the rest of the taxa (**ingroups**).
- The best outgroups satisfy two characteristics:
  1. They must not be members of the ingroup.
  2. They must be related to the ingroup, close enough for meaningful comparisons to the ingroup.





# Interpreting phylogenetic trees

## Phylogenetic interpretation skill set

---

1. Tree-thinking skills
  - relatedness, confidence, **homology**
2. Knowledge of phylogenetic methods and their limitations
3. Knowledge of biological processes affecting sequence evolution
  - **gene duplication**, recombination, horizontal gene transfer, population genetic processes, and many **more!**
4. Knowledge of the data you wish to interpret

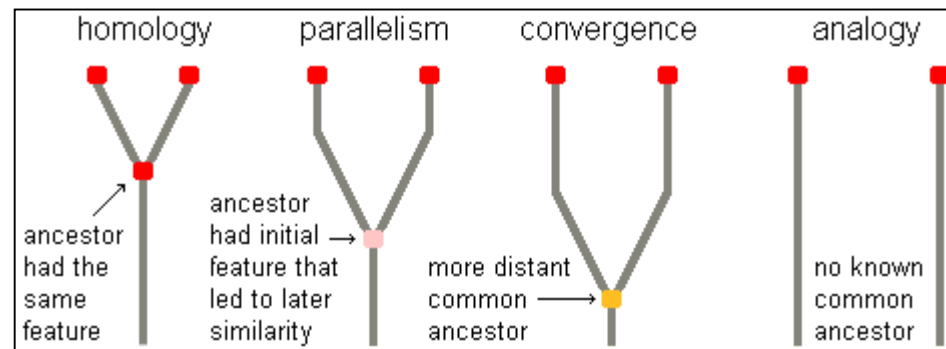
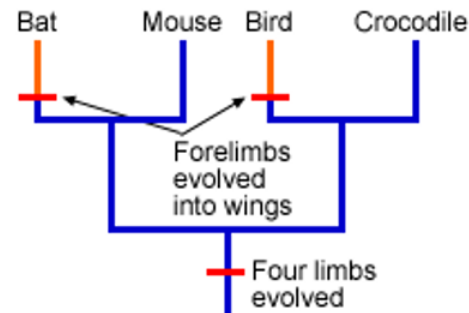
# Interpreting phylogenetic trees

## Phylogenetic interpretation skill set

**Homology is similarity due to shared ancestry**

Example: limbs and wings

- Limbs **are** homologous they share a common ancestor
- Wings are **not** homologous they are an analogous as they have evolved similarity independently



Homology, parallelism, convergence and analogy



# Interpreting phylogenetic trees

---

- It is very important to understand what phylogenetic trees do, and do not, mean.
- The trees provide two kinds of information:
  1. Branching order
  2. Branch length



# Phylogenetic Trees

## Scaled trees and Unscaled trees

---

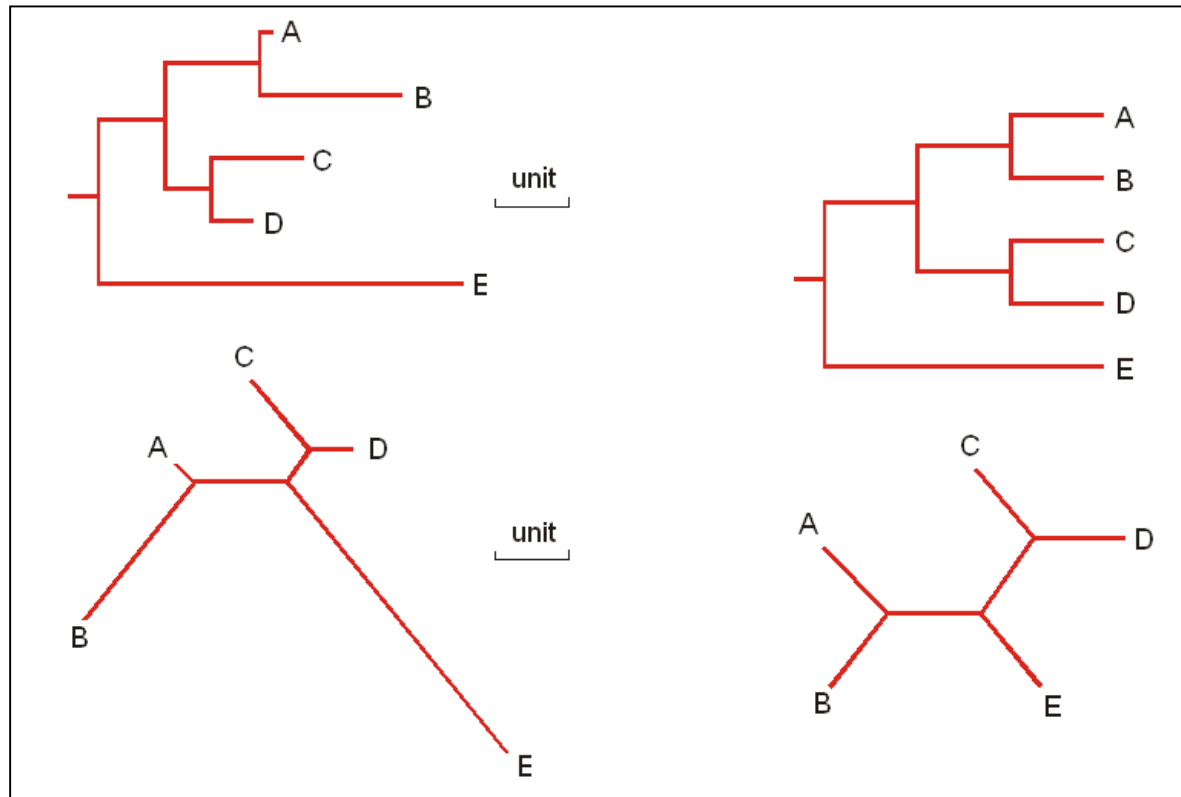
- **Scaled trees:** Branch lengths are **proportional** to the number of nucleotide/amino acid changes that occurred on that branch (usually a scale is included).
- **Unscaled trees:** Branch lengths are **not proportional** to the number of nucleotide/amino acid changes (usually used to illustrate evolutionary relationships only).



# Phylogenetic Trees

## Scaled trees and Unscaled trees

- Trees can be or unscaled (with or without branch lengths):



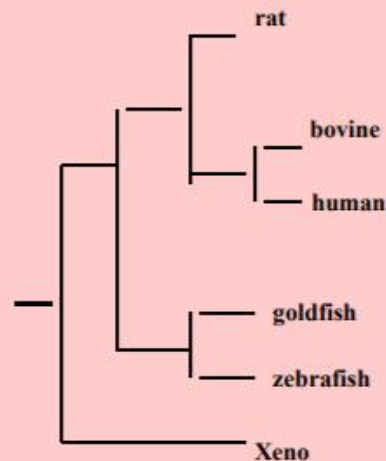
# Phylogenetic concepts

## Interpreting a Phylogeny

### How to read the tree

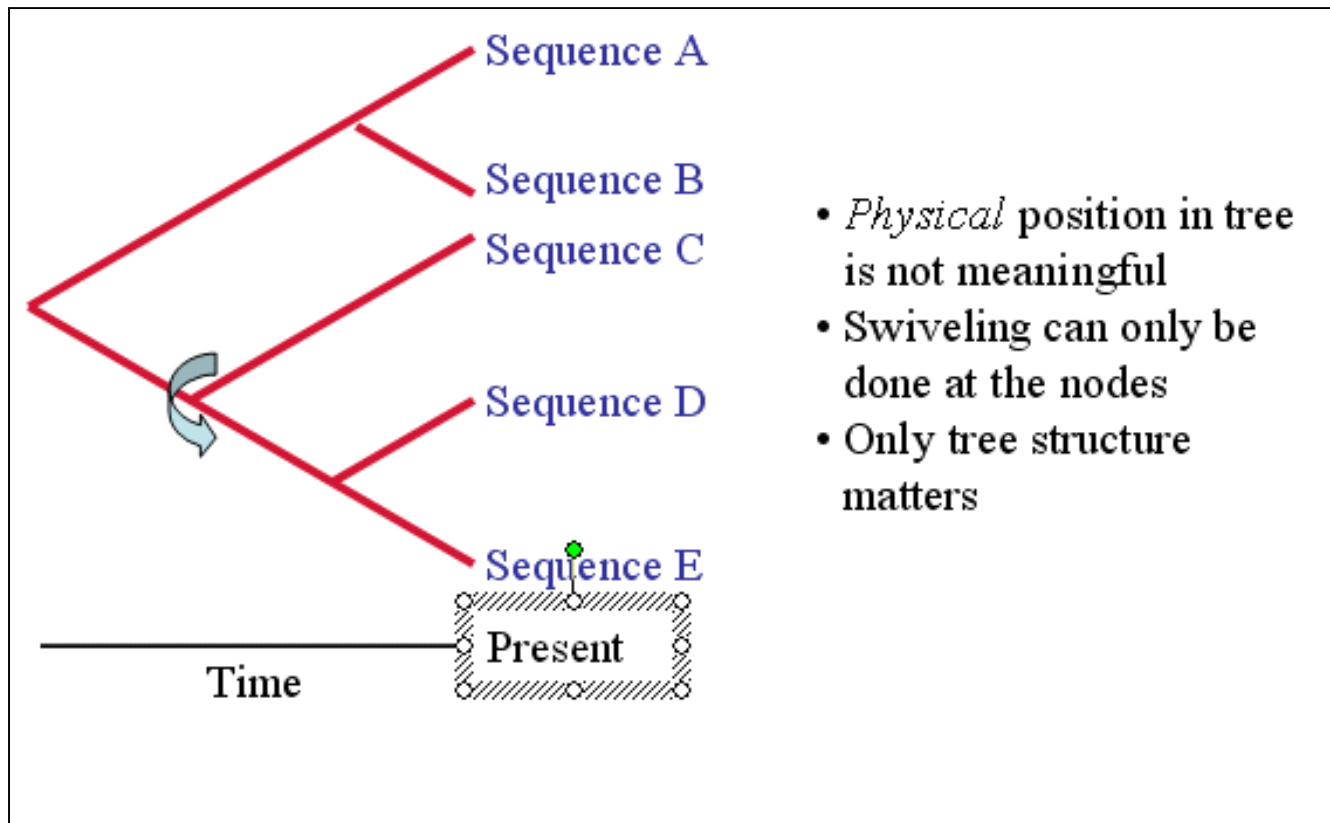
How to read the tree?

Start at the base and follow  
The progression of the branch  
points (nodes)

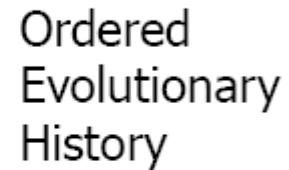


# Phylogenetic concepts

## Interpreting a Phylogeny



Added branch length=  
% sequence identity



Branch length = Sequence divergence.  
Also comparative rate of substitution/evolution



# Phylogenetic Trees

## Scale bar and Branch length

Scale bars, or branch lengths correspond to "the mean number of nucleotide substitutions per site" on the respective branch

---

- Most of the time, the numbers at the nodes of a tree are the percent values supporting the nodes.
- For example, when 90% is placed at the node of a clade (cluster, or group), it means that 90% of the tested tree replicates (or approaches) support the presence of this clade.
- A higher number means better statistical support to that particular clade (therefore, is better).

# Phylogenetic Trees

## Scale bar and Branch length

**Scale bars, or branch lengths correspond to "the mean number of nucleotide substitutions per site" on the respective branch**

- There are several methods to get these values.
- One popular method is the bootstrapping analysis, in which replicates (e.g., 1000 replicates) of a dataset are analyzed to get "bootstrapping supporting values/proportions".
- Bootstrapping analysis can be used in all analysis, such as:
  1. maximum likelihood (ML),
  2. minimum distance (MD), and
  3. maximum parsimony (MP).

# Phylogenetic Trees

## Scale bar and Branch length

Scale bars, or branch lengths correspond to "the mean number of nucleotide substitutions per site" on the respective branch

---

- Sometimes, people may also label their trees with branch lengths.
- By the way, almost all trees also have a single scale bar representing the amount of substitutions (nt or aa).

# Phylogenetic Trees

## Scale bar and Branch length

Scale bars, or branch lengths correspond to "the mean number of nucleotide substitutions per site" on the respective branch

- Branch lengths indicate genetic change i.e. the **longer the branch, the more genetic change** (or divergence) has occurred.
- Typically we measure the extent of genetic change by estimating the **average number of nucleotide or protein substitutions per site**.

Human	ATG <b>T</b> GACTC
Mouse	ATG <b>C</b> TGACTC

### Simple sequence alignment

There is one site that is different between the two sequences, and we could say that based upon this tiny sample there are  $1/10 = 0.1$  substitutions per site.

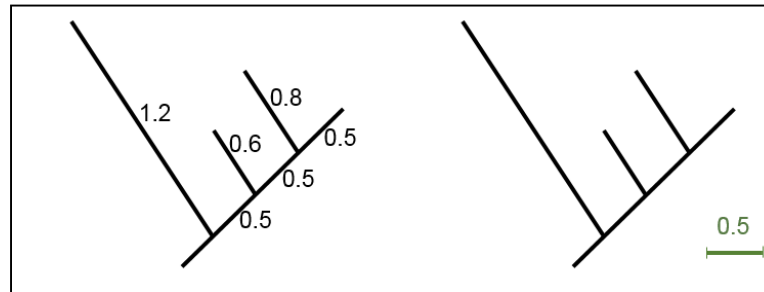


# Phylogenetic Trees

## Scale bar and Branch length

Scale bars, or branch lengths correspond to "the mean number of nucleotide substitutions per site" on the respective branch

- Branch lengths indicate genetic change i.e. the **longer the branch, the more genetic change** (or divergence) has occurred.
- Typically we measure the extent of genetic change by estimating the **average number of nucleotide or protein substitutions per site**.



Scale bars, or branch lengths

These are alternative representations of the same phylogeny.

# Phylogenetic Trees

## Scale bar and Branch length

Scale bars, or branch lengths correspond to "the mean number of nucleotide substitutions per site" on the respective branch

---

- **Scale bar:**
- A scale bar can represent branch lengths.
- The "scale bar" is a reference, basically a ruler, allowing someone viewing the tree to measure the lengths of the branches in the tree, and to compare different trees.
- Typically, the scale bar line represents an evolutionary distance of 0.10 or 0.05.
- The scale bar represents the number of substitution per 100 sites for unit branch length.

# Phylogenetic Trees

## Scale bar and Branch length

Scale bars, or branch lengths correspond to "the mean number of nucleotide substitutions per site" on the respective branch

- **Branch length:**
- Branch lengths indicate genetic change i.e. the longer the branch, the more genetic change (or divergence) has occurred.
- The units of branch length are usually nucleotide substitutions per site – that is the number of changes or 'substitutions' divided by the length of the sequence (although they may be given as % change, i.e., the number of changes per 100 nucleotide sites).
- In many phylogenetic tree schemes branch length contains no information at all.

# Algorithms for phylogenetic reconstruction

## Methods in Phylogenetic Reconstruction

---

- **Phylogenetic tree building (or inference) methods** such as distance, max. likelihood, and max. parsimony);
- **Post- phylogenetic informations** (such as molecular clocks and selection), and
- **Useful subsidiary statistical techniques** (such as bootstrapping and likelihood ratio test).



# Definitions

---

- **Maximum parsimony:** states that says when considering multiple explanations for an observation, one should first investigate the simplest explanation that is consistent with the facts.
- The principle that things should be kept as simple as possible.
- Try all possible trees and choose those that are simplest, those that imply the fewest changes in characters.
- **Character state:** The specific value taken by a character in a specific taxon.
- The best tree is the one with the fewest changes in character states and the least convergence.



# Definitions

---

- **Maximum likelihood:** states that when considering multiple phylogenetic hypotheses, one should take into account the one that reflects the most likely sequence of evolutionary events given certain rules about how DNA changes over time.
- The best tree is the one with the highest probability— the greatest likelihood.
- **Bayesian inference:** A statistical method that first establishes a basic expectation (the prior probability), and then estimates the likelihood of observing the data given that expectation (the posterior probability).



# Methods in Phylogenetic Reconstruction

## Methods in Phylogenetic Reconstruction

- ✓ Distance
- ✓ Maximum Parsimony
- ✓ Maximum Likelihood

Bayesian

\* All algorithms are calculated using available software, eg. PAUP, PHYLIP, McClade, Mr. Bayes etc.



# Phylogenetics

## Popular methods for inferring phylogenetic trees

- Once a DNA sequence is obtained for the **16S rRNA gene**, **several computer algorithms** can be used to estimate the evolutionary distance between the unknown sequence and all others present in a database (**e.g.** <http://rdp.cme.msu.edu/html>).
- After **aligning sequences** using programs like **ClustalW** or **ClustalX**, **phylogeny algorithms** are used to calculate relatedness.
- There are a lot of different methods for making a phylogeny. **The most common are:**
  1. **Distance matrix methods**
  2. **Maximum parsimony**
  3. **Maximum Likelihood**
- In the best circumstances **all three types of analyses** will give the **same phylogenetic relationships**.





# Phylogenetics

## Popular methods for inferring phylogenetic trees

---

1. Phylogenetic tree types
2. Distance Matrix method:
  - UPGMA
  - Neighbor joining
3. Character State method:
  - Maximum likelihood

# Comparison of the most popular phylogenetic methods

## Distance, Maximum parsimony and Maximum likelihood

- **Maximum parsimony** procedures search for tree topologies which require a minimum number of base changes to correlate with the sequence data.
- The **maximum likelihood** procedure is considered the most sophisticated method for developing a phylogenetic tree.
- It also **searches tree topologies** in ways that reflect how current sequences were most likely to have been generated.

	Character-based methods	Noncharacter-based methods
Methods based on an explicit model of evolution	Maximum-likelihood methods	Pairwise-distance methods
Methods not based on an explicit model of evolution	Maximum-parsimony methods	

Pairwise distance methods are non-character-based methods that make use of an explicit substitution model.

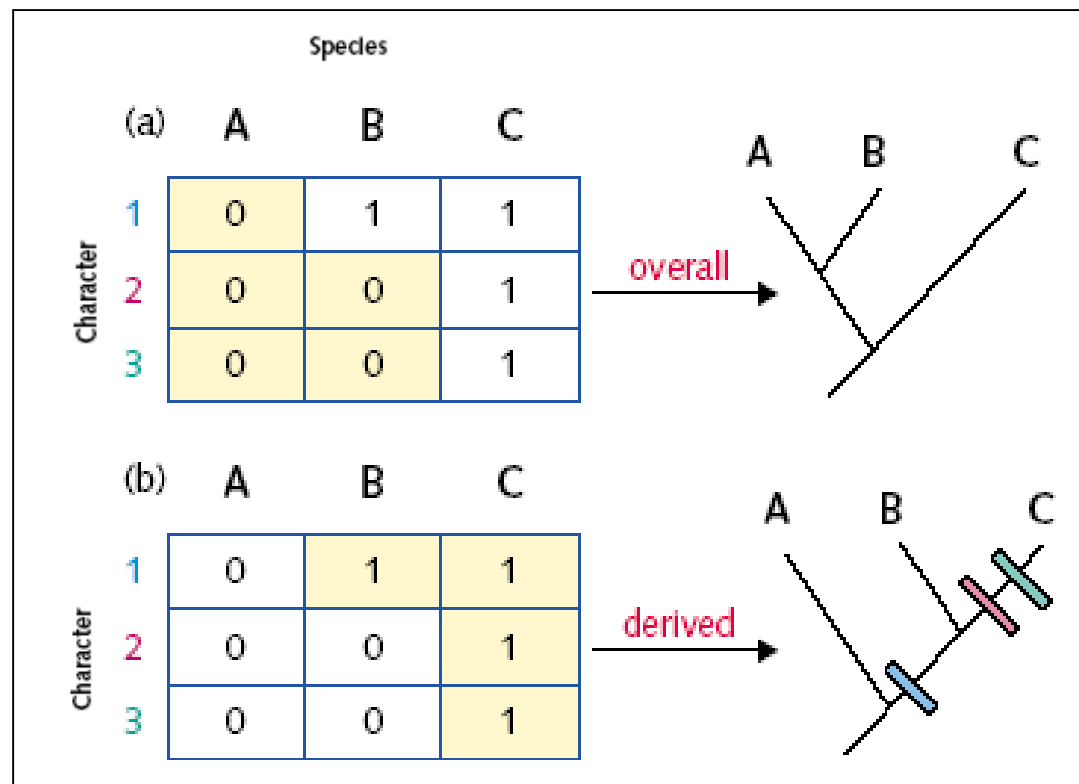
# Comparison of the most popular phylogenetic methods

**Distance, Maximum parsimony and Maximum likelihood**

## Comparison of Methods

Distance	Maximum parsimony	Maximum likelihood
Uses only pairwise distances	Uses only shared derived characters	Uses all data
Minimizes distance between nearest neighbors	Minimizes total distance	Maximizes tree likelihood given specific parameter values
Very fast	Slow	<b>Very</b> slow
Easily trapped in local optima	Assumptions fail when evolution is rapid	Highly dependent on assumed evolution model
Good for generating tentative tree, or choosing among multiple trees	Best option when tractable (<30 taxa, homoplasy rare)	Good for very small data sets and for testing trees built using other methods

# Overall or derived similarity





# Methods in Phylogenetic Reconstruction

## Distance matrices

---

- Calculate all the distance between leaves (taxa);
- Based on the distance, construct a tree;
- Good for continuous characters;
- Simple, finds only one tree
- Not very accurate.
- Fastest method:
  1. UPGMA
  2. Neighbor-joining.

# Distance-based phylogeny reconstruction Algorithms

## Distance algorithms

### Distance matrix methods

- A major family of phylogenetic methods has been the **distance matrix methods**.
- A phylogeny **tree is built based on the distance between the taxa** (the more similar ones should be evolutionary more related).
  1. **UPGMA algorithm.**
  2. **Neighbor-joining algorithm.**

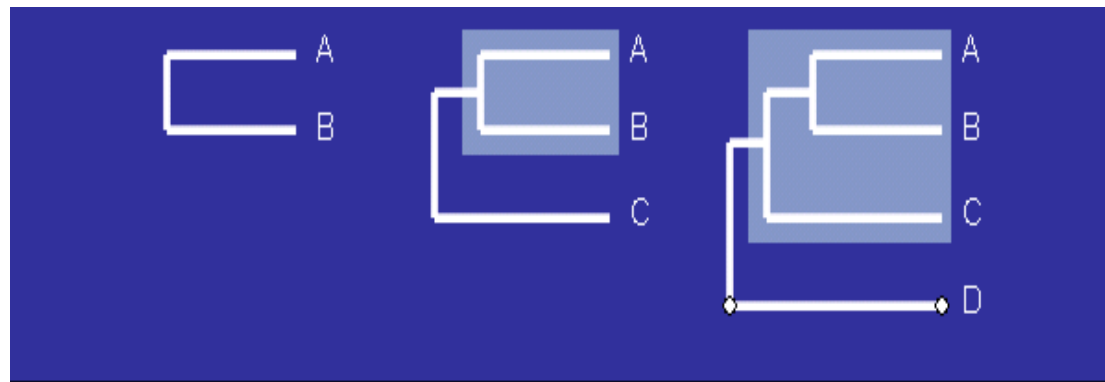
To construct a phylogeny you can use:

1. **the Neighbour-Joining tree building method**, and
2. the **Tamura-Nei model**.
  - For the **genetic distance model** select **Tamura-Nei** and for the **tree build method** select **Neighbor-Joining**.
  - To build a Neighbour-Joining tree you can use the Tamura-Nei model.

# Methods in Phylogenetic Reconstruction

## Distance matrices

- Using a sequence alignment, pairwise distances are calculated.
- Creates a distance matrix.
- A phylogenetic tree is calculated with clustering algorithms, using the distance matrix.

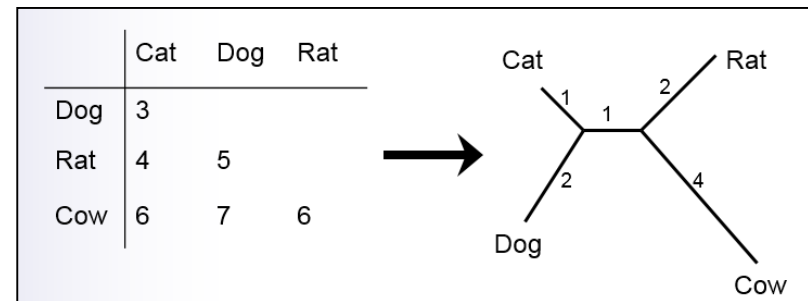


# Methods in Phylogenetic Reconstruction

## Distance matrices

- Distance Based Methods for estimating phylogenetic trees:
- There are many ways of building phylogenetic trees, one family of methods uses a distance matrix as a starting point.
- A distance matrix is a table that indicates pairwise dissimilarity, for instance...

	Cat	Dog	Rat	Cow
Cat	0	2	4	7
Dog	2	0	5	6
Rat	4	5	0	3
Cow	7	6	3	0





# Methods in Phylogenetic Reconstruction

## Distance matrices

- Distances can be derived from Multiple Sequence Alignments (MSAs).
- The most basic distance is just a count of the number of sites which differ between two sequences divided by the sequence length.
- These are sometimes known as p-distances.

Cat	ATTGCGGTA
Dog	ATCTGCGATA
Rat	ATTGCCGTTT
Cow	TTCGCTGTTT



	Cat	Dog	Rat	Cow
Cat	0	0.2	0.4	0.7
Dog	0.2	0	0.5	0.6
Rat	0.4	0.5	0	0.3
Cow	0.7	0.6	0.3	0



# Methods in Phylogenetic Reconstruction

## Maximum parsimony

---

- Finds the optimum tree by minimizing the number of evolutionary changes.
- No assumptions on the evolutionary pattern
- May oversimplify evolution.
- May produce several equally good trees.



# Methods in Phylogenetic Reconstruction

## Maximum parsimony and minimum evolution methods

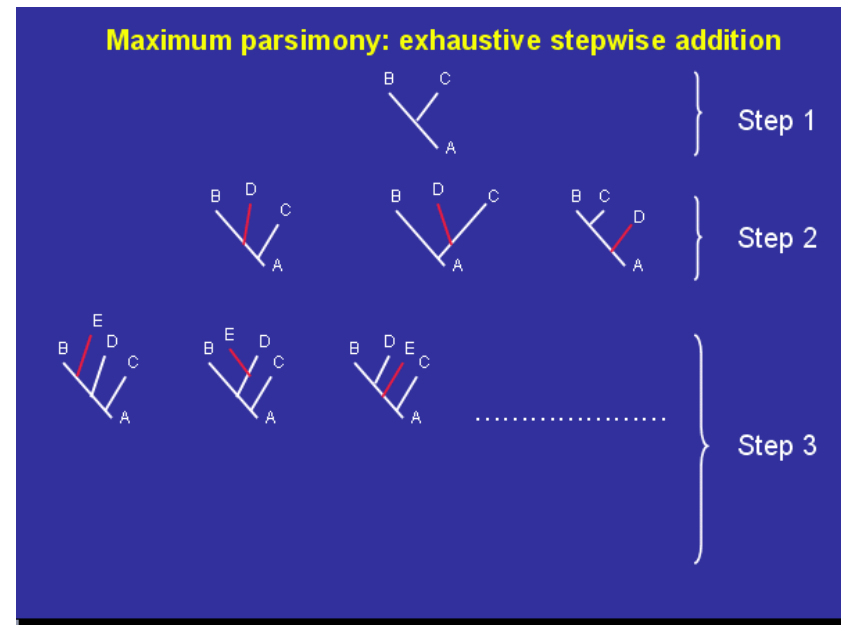
---

- Maximum parsimony and minimum evolution are methods that try to:
  1. minimize branch lengths by either minimizing distance (minimum evolution), or
  2. minimizing the number of mutations (maximum parsimony).
- The major problem with these methods is that they fail to take into account many factors of sequence evolution (e.g. reversals, convergence, and homoplasy).
- Thus, the deeper the divergence times that more likely these methods will lead to erroneous or poorly supported groupings.

# Methods in Phylogenetic Reconstruction

## Maximum parsimony

- The most parsimonious tree is the one that has the fewest evolutionary changes for all sequences to be derived from a common ancestor.
- Usually several equally parsimonious trees result from a single run.

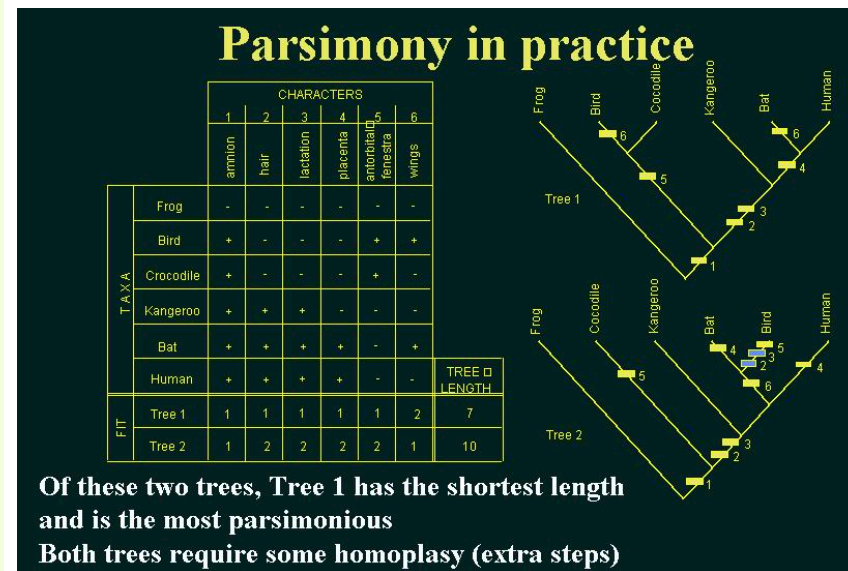


# Methods in Phylogenetic Reconstruction

## Maximum Parsimony

### Parsimony in practice

- Characters differ in their fit to different trees.
- Given a set of characters, such as aligned sequences, parsimony analysis works by determining the **fit** (number of steps) of each character on a given tree.
- The sum over all characters is called Tree Length.
- Most parsimonious trees (MPTs) have the **minimum tree length** needed to explain the observed distributions of all the characters.





# Methods in Phylogenetic Reconstruction

## Maximum Likelihood

---

- The best tree is found based on assumptions on evolution model.
- Nucleotide models more advanced at the moment than amino acid models.
- Programs require lot of capacity from the system.



# Methods in Phylogenetic Reconstruction

## Maximum Likelihood

---

- Creates all possible trees like **Maximum Parsimony method** but instead of retaining trees with shortest evolutionary steps.....
- Employs a model of evolution whereby different rates of transition/transversion ratio can be used.
- Each tree generated is calculated for the probability that it reflects each position of the sequence data.
- **Calculation is repeated for all nucleotide sites.**
- Finally, the tree with the best probability is shown as the **maximum likelihood tree** - **usually only a single tree remains.**
- It is a more realistic tree estimation because it does not assume equal transition-transversion ratio for all branches.

# Methods in Phylogenetic Reconstruction

## UPGMA algorithm

### UPGMA vs. NJ

---

- **NJ**(Neighbor joining) and **UPGMA** (Unweighted Pair Group Method with Arithmetic Mean) are clustering algorithms that can **make quick trees** but are not the most reliable, especially when dealing with **deeper divergence times**.
- These method are good to give you an idea about your data, but are **almost never acceptable for publication**.



# Methods in Phylogenetic Reconstruction

## UPGMA algorithm

### UPGMA vs. NJ

---

- The UPGMA method (Unweighted Pair Group Method with Arithmetic Mean) is the simplest method of tree construction.
- It assumes that evolution has occurred at a constant rate in the different lineages. This means that a root of tree is also estimated.
- Thus, UPGMA works by progressively clustering the most similar taxa until all the taxa form a rooted clock-like tree.
  1. UPGMA is consistent for clock-like distances, and
  2. NJ is inconsistent for any additive distances. Additive means distance between species.



# Other phylogeny algorithms

## "Neighbor-joining" (e.g. "neighbor" program)

---

- The **neighbor(neighbour)-joining method** builds a tree where the evolutionary rates are free to differ in different lineages.
- The **Neighbour Joining method** is a method for reconstructing phylogenetic trees, and computing the lengths of the branches of this tree.
- In each stage, the two nearest nodes of the tree (the term "**nearest nodes**" are chosen and defined as **neighbours** in our tree).

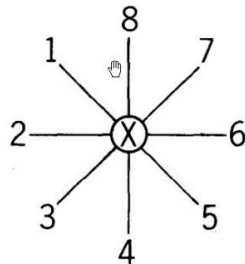
# Other phylogeny algorithms

## Neighbor-joining method

- NJ method not only provides **topology** but also provides final tree with **branched lengths**.
- **Join the closest neighbors** (OTUs with similar characters).

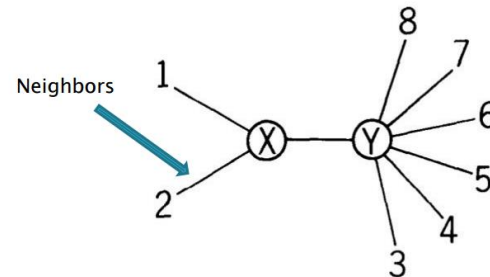
### Algorithm

- ▶ construct an unresolved tree with star topology



### Algorithm

- ▶ Join the closest neighbors ( OTUs with similar characters )





# Other phylogeny algorithms

## Bootstrapping

---

- **Confidence estimates (e.g. Bootstrap):**
- To evaluate the reliability of the inferred tree, the option of doing a **bootstrap analysis** is allowed.
- A **bootstrap value** is attached to each branch, and this value is a measure of confidence in this branch.
- A tree is constructed. This process is repeated 100 of times.
- **The maximum value is 100.**
- The number at internal branches show the bootstrap support (%).



# Bootstrapping

## Confidence or “faith”?

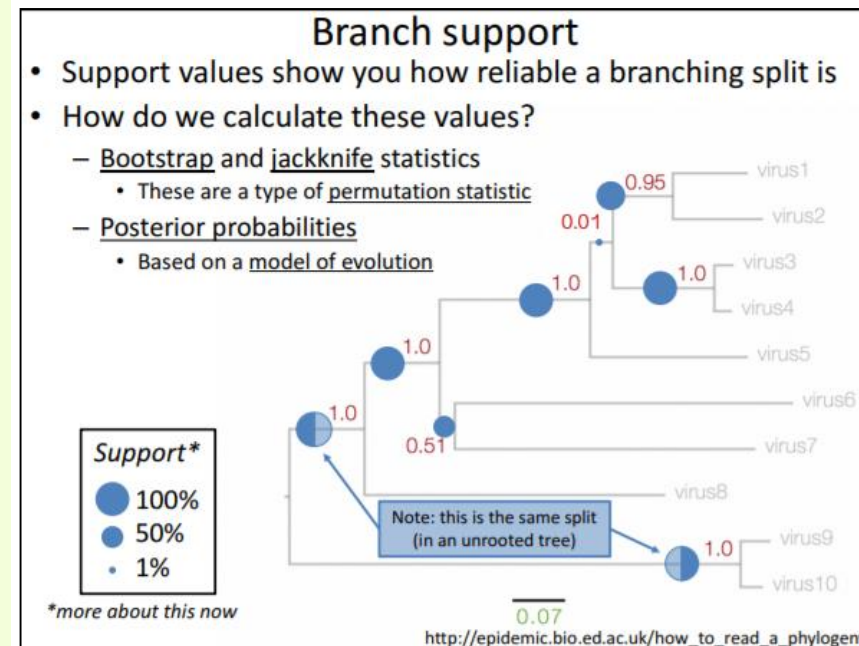
---

- Is the tree correct? How robust?
- Accuracy is difficult to judge (we almost never know the true phylogeny).
- Resampling methods: **bootstrap**, jackknife
- **Bootstrap**: Generates pseudoreplicates, random samples with replacements
- **“Bootstrap value”**= Frequency with which a group of sequences appear in bootstrap trees (**expressed as %**).
  1. High bootstrap values (>70%) indicate reliable trees.
  2. Lower percentages indicate that there is insufficient information in the sequences to be sure about the resulting tree.

# Construction of a phylogenetic tree

## The scale bar distance and bootstrap values

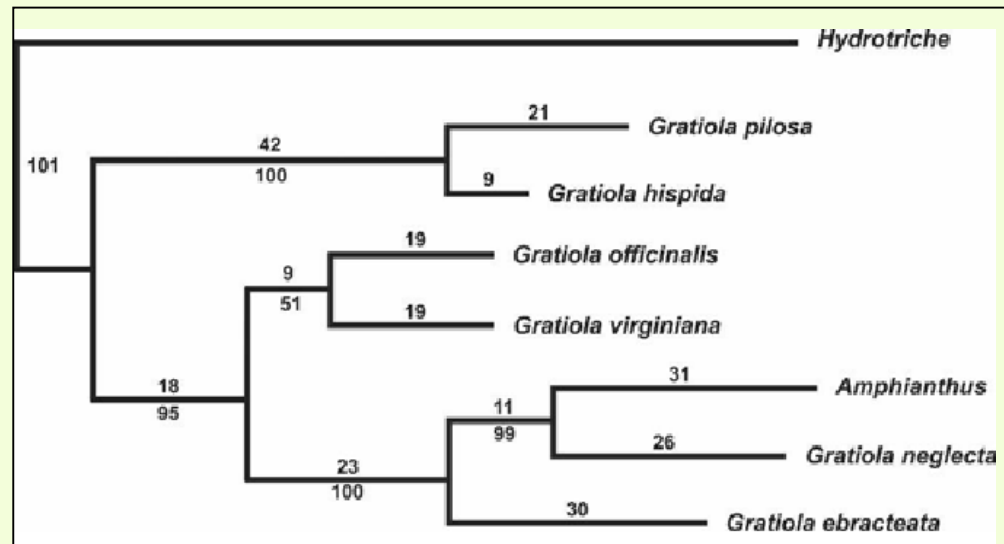
- The scales represents the number of differences between sequences:
- The scale bar at the bottom (0.7) shows the number of substitutions per position;
- The numbers in parenthesis show the number of species in the respective branches; and,
- The number at internal branches show the bootstrap support (%).



# Construction of a phylogenetic tree

## Branch length and bootstrap values

- Plant genus *Gratiola*.
- Numbers above branches are branch lengths;
- Numbers below branches are bootstrap values.

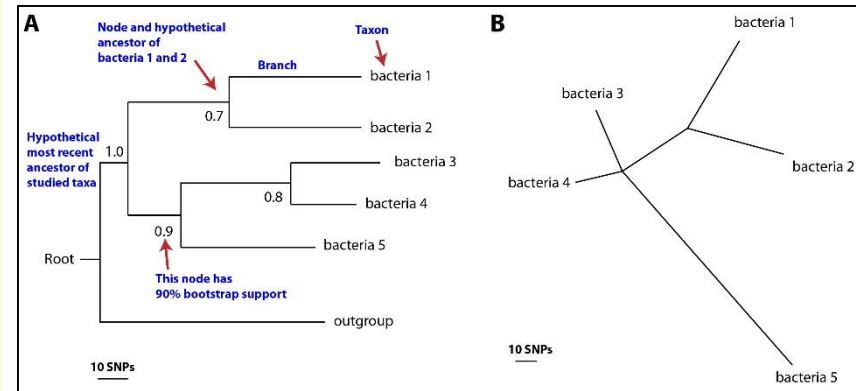


# Bootstrapping

## Bootstrapped tree

Values are in percentage

- The node separating bacterial strains 1 and 2 from strains 3, 4 and 5 is the most confident relationship in the tree with 100% bootstrap support.
- The most closely related two strains are bacteria 3 and bacteria 4, as shown by branch length in both the horizontal tree (A) and the radial tree (B).
- The least confident relationship in the tree is bacteria 1 and bacteria 2, which has 70% bootstrap support.



Phylogenetic trees based on DNA sequence are typically built using SNPs (single-nucleotide polymorphisms).

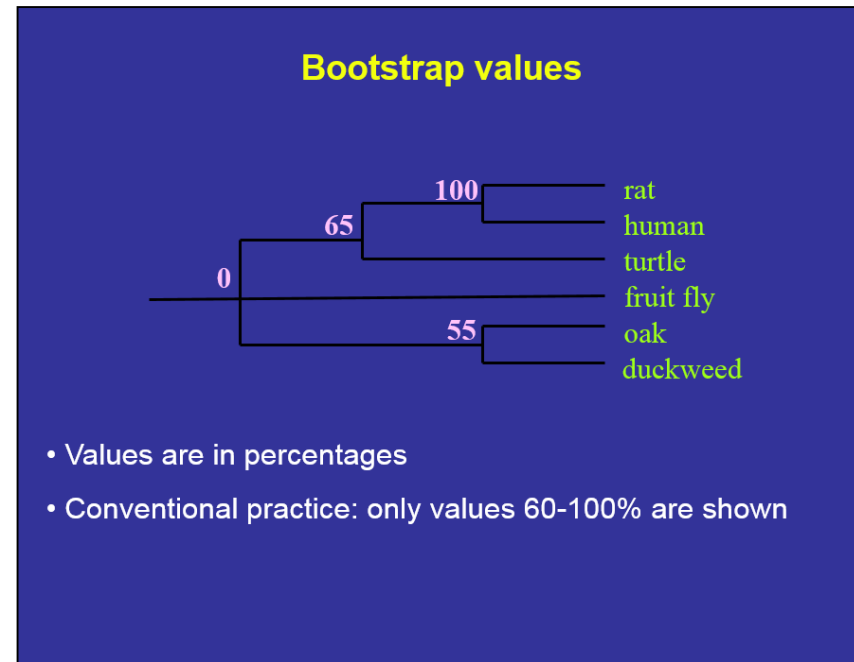


# Bootstrapping

## Bootstrapped tree

### Interpreting bootstrap values

- Any branch with 100% support is certain.
- This means that the species within it were always found together as a cluster.
- No other sequences belong to that cluster.

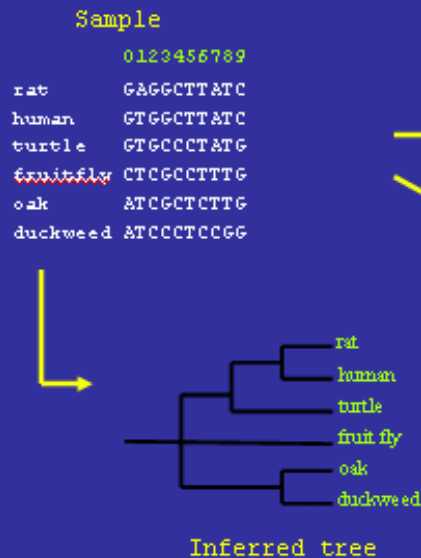


# Tree with bootstrap values

## Bootstrapping

### The Bootstrap

- Computational method to estimate the confidence level of a certain phylogenetic tree.



Pseudo sample 1

```
001122234556667
rat      GGAAGGGGCTTTTA
human    GGTGGGGCTTTTA
turtle   GGTGGGCCCTTTA
fruitfly CCTCCCGCCCTTT
oak       AATTCCCGCTTCCCT
duckweed AATTCCCGCTTCCCT
```

Pseudo sample 2

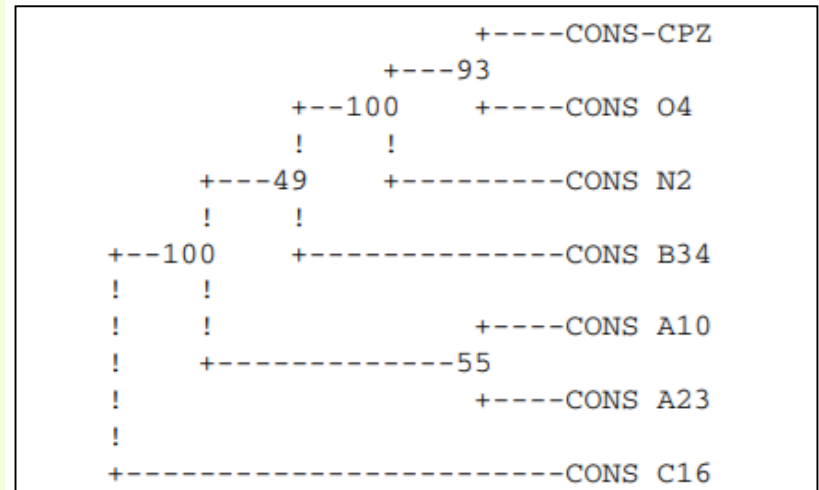
```
445556777888899
rat      CCTTTTAAATTTCC
human    CCTTTTAAATTTCC
turtle   CCCCCTAAATTTGG
fruitfly CCCCCTTTTTTTTGG
oak       CCTTCTTTTTTTTGG
duckweed CCTTCCCCGGGGGG
```

Many more replicates  
(between 100 - 1000)

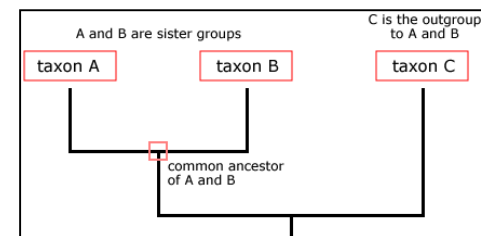
# Tree with bootstrap values

## Bootstrap values at the inner nodes

- Figure shows bootstrap values at the inner nodes. For example:
  - 93** means that the species **CONS-CPZ** and **CONS O4** were siblings(sister or brother from common parents) in 93% of the bootstrap replications;
  - 49** means that the sequences **CONS-CPZ**, **CONS O4**, **CONS N2** and **CONS B34** were grouped together in what is called a **monophyletic** (a **group** containing the most common ancestor of a given set of taxa and **all the descendants of that most recent common ancestor**) **clade** in 49% of the bootstrap replications.



A **monophyletic group** (also described as a **clade**) is a group of taxa that share a more recent common ancestor with each other than to any other taxa.





# Tree with bootstrap values

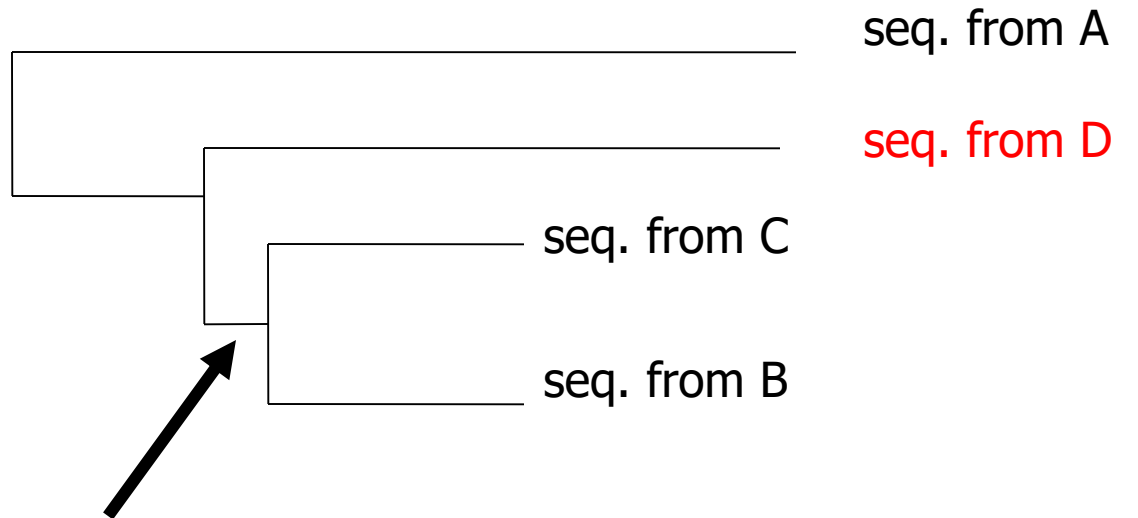
## Bootstrap values on branch length

---

- The 'branch lengths' are not true branch lengths, but rather reflect the % bootstrap values.
- Higher the bootstrap value, higher the confidence level of the clade in the phylogenetic tree.
- It tells you if 1000 times this tree is made using a particular data, this much is the confidence value (Bootstrap value).
  1. If you get 100 out of 100 (and your data is sufficiently large to support this), we are pretty damned sure that the observed branch is not due to a single extreme data point.
  2. If you get 50 out of 100, we cannot be as certain.

# Tree with bootstrap values

## Lack of resolution



e.g., 40% bootstrap support for bipartition (AD)(CB)

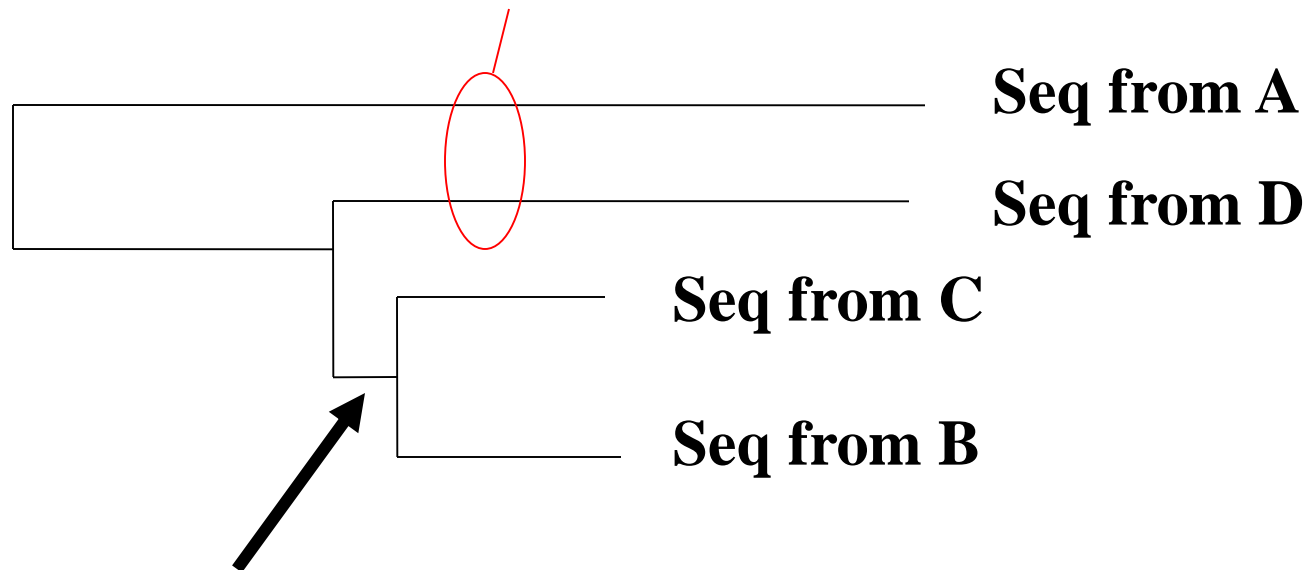
(typical >80%)

100 means that the node is well-supported.  
A lower bootstrap represents uncertainty of a node.

# Tree with bootstrap values

## Long branch attraction artifact (LBA)

the two longest branches join together



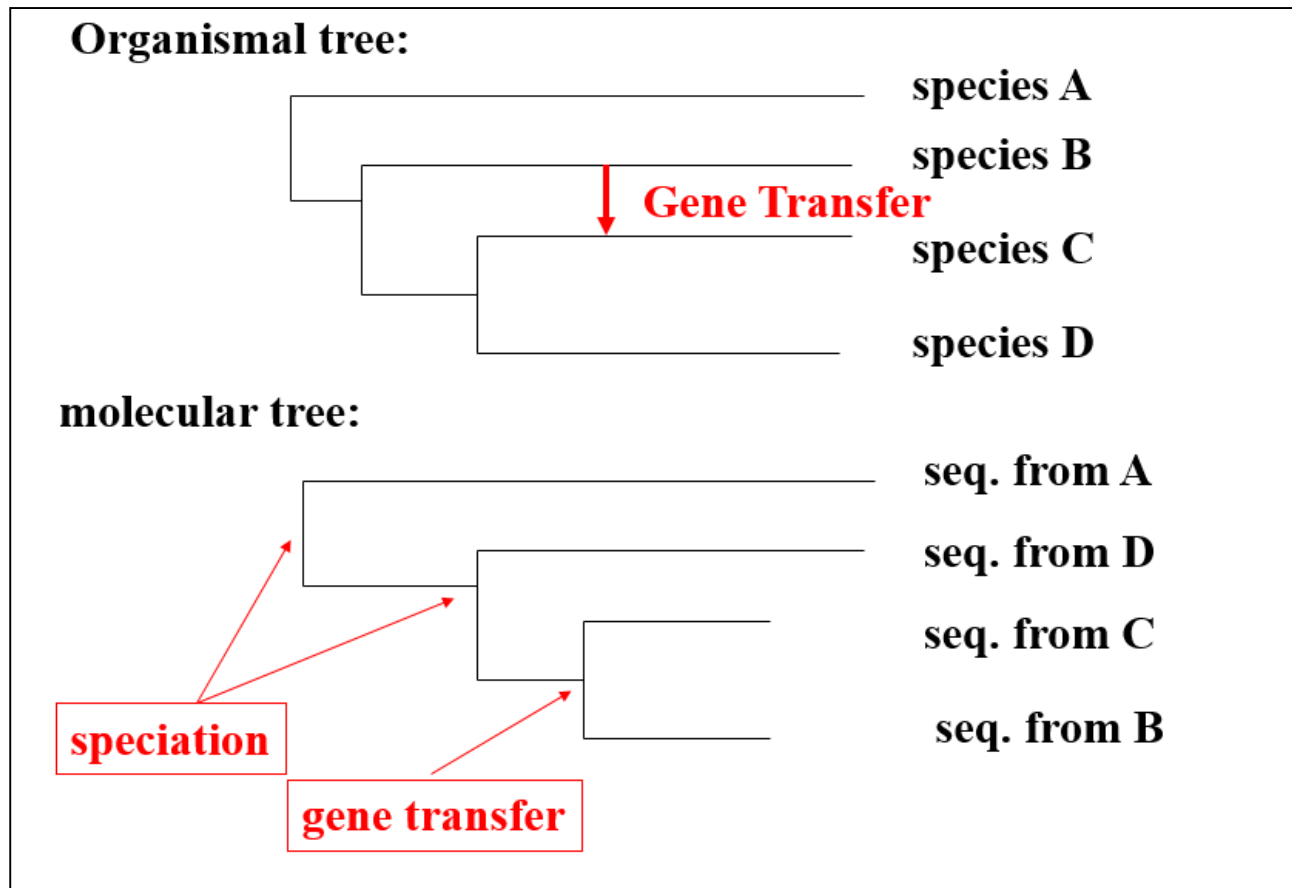
**Strong support, e.g., 100% bootstrap for (AD)(CB)**

There is consistent (100% bootstrap) support that taxa A and D are more closely related to each other than they are to C and B.

# Tree with bootstrap values

## Organismal tree and molecular tree

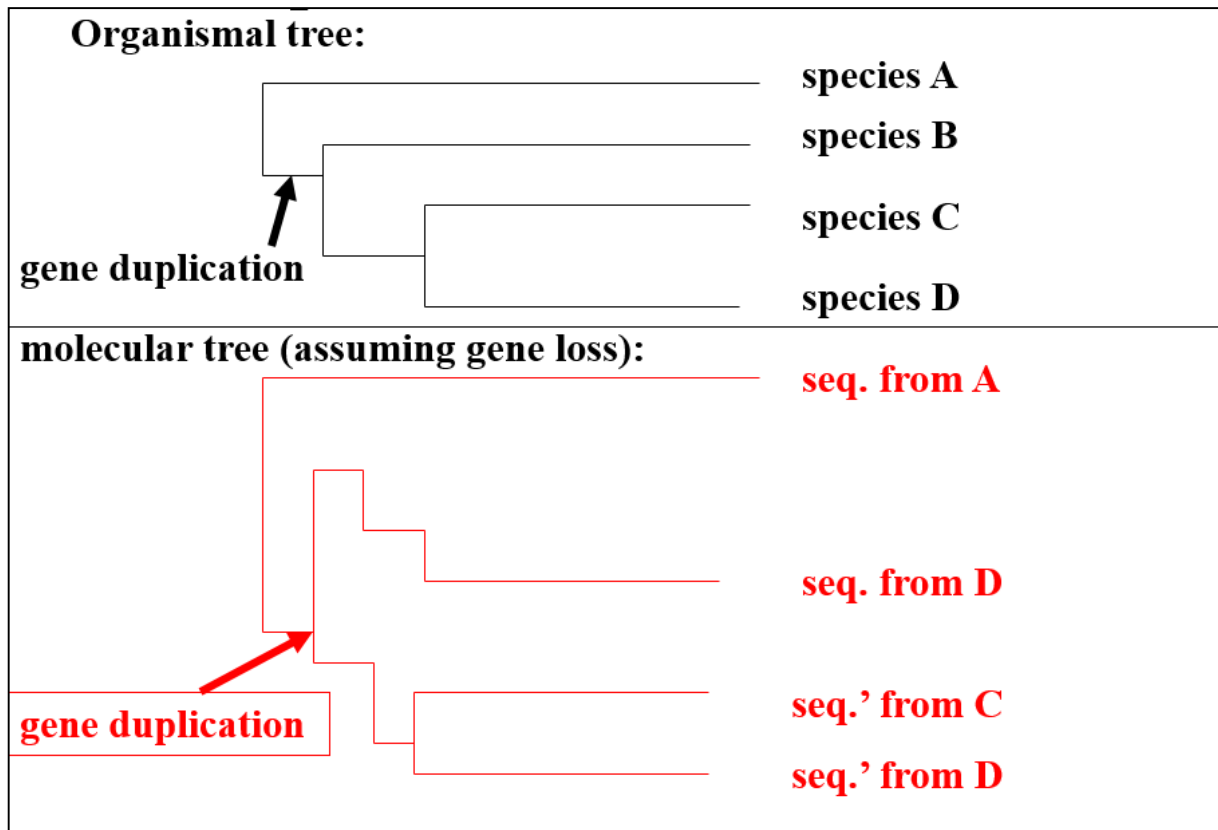
### Gene transfer and speciation



# Tree with bootstrap values

## Organismal tree and molecular tree

### Gene duplication

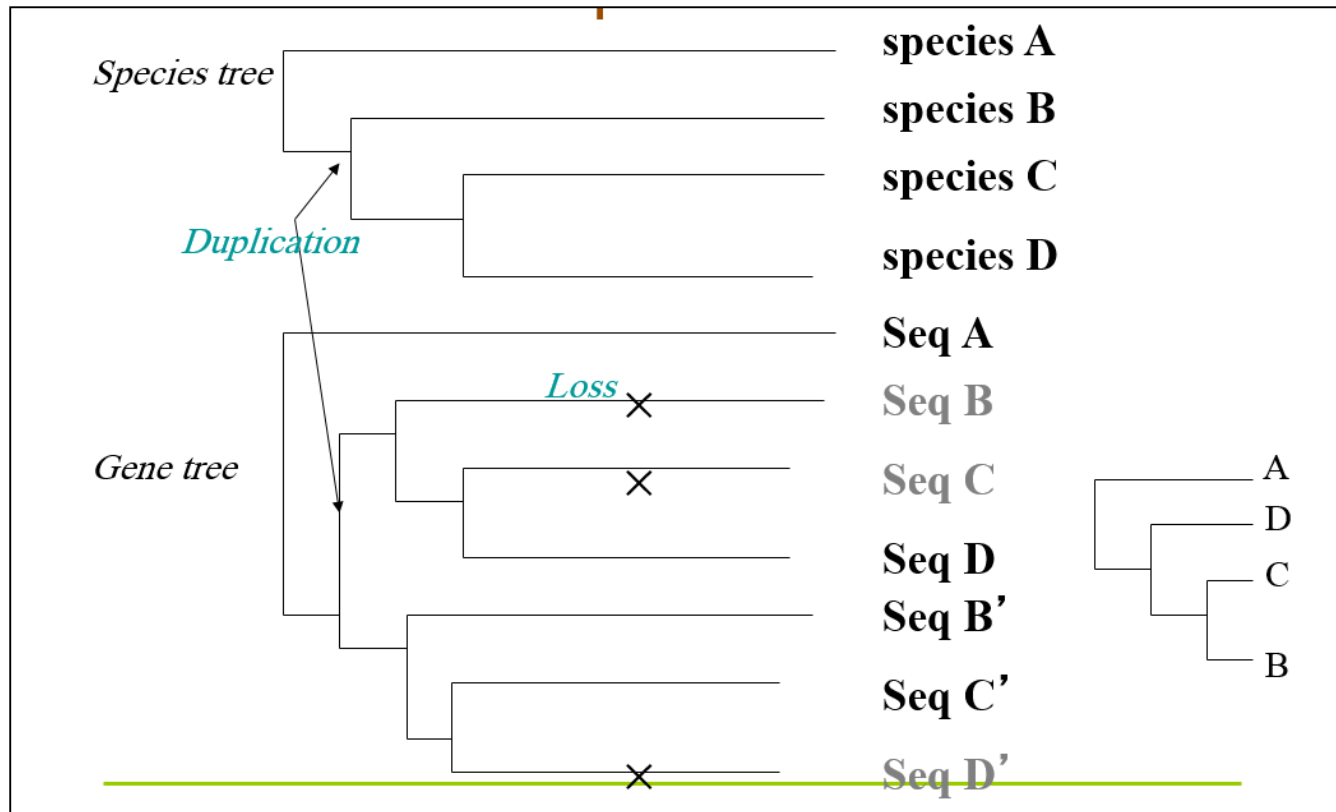




# Tree with bootstrap values

## Species tree vs. gene tree

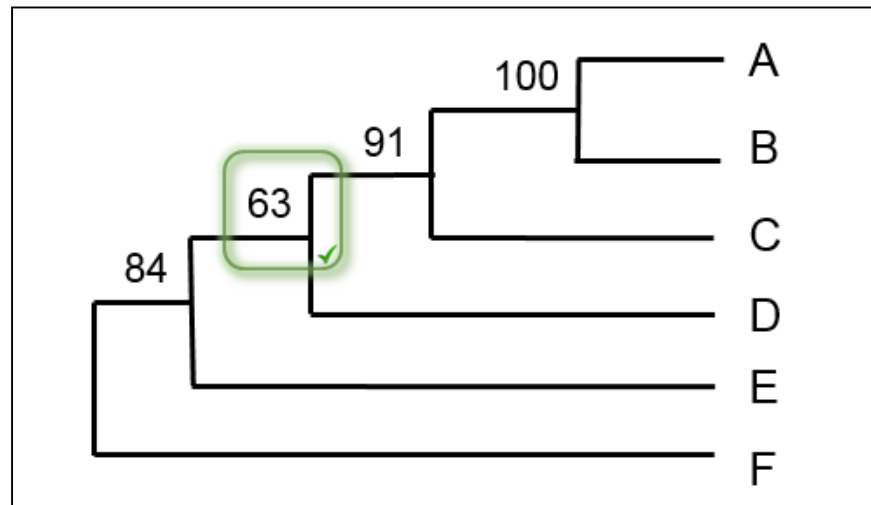
### Gene duplication and loss



# Tree with bootstrap values

## Confidence Question

- Which of the bootstrap values indicates our confidence in the grouping of A, B, C, and D together as a monophyletic group? Do you think we can be confident in this grouping?



Note: high bootstrap values do not always mean that we have confidence in a branch. False confidence can be generated under some phylogenetic methods.



# Tree with bootstrap values

## Cut-off method

---

- It might be said that high bootstrap proportions are a necessary, but not sufficient, condition for having high confidence in a group.
- The exact interpretation of the bootstrap proportion is elusive; higher is clearly better, but what is a reasonable cut-off?
- Some workers have concluded that bootstrap proportions are conservative measures of support, so a value of 70% might indicate strong support for a group.



# Analysing the aligned sequence matrix

---

- PHYLIP
- POY
- PAUP, GCG
- And many more... (274 software packages described at one website)





# General-purpose packages

---

- [PHYLIP](#)
- [PAUP\\*](#)
- [MEGA](#)
- [Phylo\\_win](#)
- [ARB](#)
- [DAMBE](#)
- [PAL](#)
- [Bionumerics](#)
- [Mesquite](#)
- [CIPRES](#)
- [PaupUp](#)



# Parsimony programs

---

- [PAUP\\*](#)
- [Hennig86](#)
- [MEGA](#)
- [RA](#)
- [Nona](#)
- [PHYLIP](#)
- [TurboTree](#)
- [CAFCA](#)
- [Phylo\\_win](#)
- [sog](#)
- [gmaes](#)
- [LVB](#)
- [GeneTree](#)
- [TAAR](#)
- [ARB](#)
- [DAMBE](#)
- [MALIGN](#)
- [POY](#)
- [Gambit](#)
- [TNT](#)
- [GelCompar II](#)
- [Bionumerics](#)
- [Network](#)
- [TCS](#)
- [GAPars](#)
- [PAUPRat](#)
- [Mesquite](#)
- [PAST](#)
- [FootPrinter](#)
- [BPAnalysis](#)
- [Simplot](#)
- [Parsimov](#)
- [NimbleTree](#)
- [PaupUp](#)



# Distance matrix methods

---

- [PHYLIP](#)
- [PAUP\\*](#)
- [MEGA](#)
- [MacT](#)
- [ODEN](#)
- [TREECON](#)
- [DISPAN](#)
- [RESTSITE](#)
- [NTSYSpc](#)
- [METREE](#)
- [TreeTree](#)
- [GDA](#)
- [Hadtree, Prepare and Trees](#)
- [GCG Wisconsin Package](#)
- [SeqPup](#)
- [PHYLTEST](#)
- [Lintre](#)
- [WET](#)
- [Phylo\\_win](#)
- [POPTREE](#)
- [Gambit](#)
- [gmaes](#)
- [DENDRON](#)
- [Fingerprinting II Informatix Software](#)
- [BIONJ](#)
- [TFPGA](#)
- [MVSP](#)
- [ARB](#)
- [Darwin](#)
- [T-REX](#)
- [sendbs](#)
- [nneighbor](#)
- [DAMBE](#)
- [weighbor](#)
- [DNASIS](#)
- [MINSPNET](#)
- [PAL](#)
- [Arlequin](#)
- [vCEBL](#)
- [HY-PHY](#)
- [Vanilla](#)
- [GelCompar II](#)
- [Bionumerics](#)
- [qclust](#)
- [TCS](#)
- [Populations](#)
- [Winboot](#)
- [SYN-TAX](#)
- [PTP](#)
- [SplitsTree](#)
- [FastME](#)
- [APE](#)
- [MacVector](#)
- [Discovery Studio Gene](#)
- [QuickTree](#)
- [Simplot](#)
- [ProfDist](#)
- [START](#)
- [STC](#)
- [NimbleTree](#)
- [CBCAnalyzer](#)
- [PaupUp](#)





# Computation of distances

---

- [PHYLIP](#)
- [PAUP\\*](#)
- [RAPDistance](#)
- [MULTICOMP](#)
- [Microsat](#)
- [DIPLOMO](#)
- [OSA](#)
- [DISPAN](#)
- [RESTSITE](#)
- [NTSYSpc](#)
- [TREE-PUZZLE](#)
- [Hadtrees, Prepare and Trees](#)
- [GCG Wisconsin Package](#)
- [AMP](#)
- [GCUA](#)
- [DERANGE2](#)
- [POPGENE](#)
- [TFPGA](#)
- [REAP](#)
- [MVSP](#)
- [RSTCALC](#)
- [Genetix](#)
- [DISTANCE](#)
- [Darwin](#)
- [sendbs](#)
- [K2WuLi](#)
- [GeneStrut](#)
- [Arlequin](#)
- [DAMBE](#)
- [DnaSP](#)
- [PAML](#)
- [puzzleboot](#)
- [PAL](#)
- [Vanilla](#)
- [GelCompar II](#)
- [Bionumerics](#)
- [qclust](#)
- [Populations](#)
- [Winboot](#)
- [ESTAT](#)
- [SYN-TAX](#)
- [Phylo\\_win](#)
- [Phyltools](#)
- [MSA](#)
- [APE](#)
- [YCDMA](#)
- [NSA](#)
- [T-REX](#)
- [LDDist](#)
- [DIVAGE](#)
- [Genepop](#)
- [START](#)
- [Swaap](#)
- [Swaap\\_PH](#)
- [GeneContent](#)
- [SPAGeDi](#)
- [CBCAnalyzer](#)
- [PaupUp](#)

# Maximum likelihood and Bayesian methods

- [PHYLIP](#)
- [PAUP\\*](#)
- [fastDNAmI](#)
- [MOLPHY](#)
- [PAML](#)
- [Spectrum](#)
- [SplitsTree](#)
- [PLATO](#)
- [TREE-PUZZLE](#)
- [Hadtrees, Prepare and Trees](#)
- [SeqPup](#)
- [Phylo\\_win](#)
- [PASSML](#)
- [ARB](#)
- [Darwin](#)
- [BAMBE](#)
- [DAMBE](#)
- [Modeltest](#)
- [TreeCons](#)
- [VeryfastDNAmI](#)
- [PAL](#)
- [dnarates](#)
- [TrExMI](#)
- [HY-PHY](#)
- [Vanilla](#)
- [DT-ModSel](#)
- [Bionumerics](#)
- [fastDNAmIRev](#)
- [RevDNArates](#)
- [rate-evolution](#)
- [MrBayes](#)
- [Hadtrees, Prepare and Trees](#)
- [CONSEL](#)
- [PAUPRat](#)

- [EDIBLE](#)
- [Mesquite](#)
- [PTP](#)
- [Treefinder](#)
- [MetaPIGA](#)
- [RAXML](#)
- [PHASE](#)
- [PHYML](#)
- [BEAST](#)
- [r8s-bootstrap](#)
- [MrBayes tree scanners](#)
- [MTgui](#)
- [MrModeltest](#)
- [BootPHYML](#)
- [p4](#)
- [Porn\\*](#)
- [SIMMAP](#)
- [Spectronet](#)
- [CIPRES](#)
- [Rhino](#)
- [IM](#)
- [ProtTest](#)
- [ModelGenerator](#)
- [Simplot](#)
- [MDIV](#)
- [MrAIC](#)
- [Modelfit](#)
- [IQPNNI](#)
- [PARAT](#)
- [ALIFRITZ](#)
- [PhyNav](#)
- [DPRML](#)
- [Continuous](#)
- [MultiPhyI](#)
- [NimbleTree](#)
- [PaupUp](#)



# Bootstrapping and other measures of support

---

- [PHYLIP](#)
- [PAUP\\*](#)
- [PARBOOT](#)
- [Random Cladistics](#)
- [AutoDecay](#)
- [TreeRot](#)
- [DNA Stacks](#)
- [OSA](#)
- [DISPAN](#)
- [TreeTree](#)
- [PHYLTEST](#)
- [Lintre](#)
- [sq](#)
- [POPTREE](#)
- [MEGA](#)
- [PICA](#)
- [ModelTest](#)
- [TAXEQ3](#)
- [TreeCons](#)
- [BAMBE](#)
- [DAMBE](#)
- [puzzleboot](#)
- [CodonBootstrap](#)
- [Gambit](#)
- [TrExMI](#)
- [PAL](#)
- [PHYCON](#)
- [MrBayes](#)
- [CONSEL](#)
- [Populations](#)
- [LVB](#)
- [EDIBLE](#)
- [Winboot](#)
- [Mesquite](#)
- [Phylo\\_win](#)
- [PAST](#)
- [Treefinder](#)
- [RAXML](#)
- [Phyltools](#)
- [PHASE](#)
- [PHYML](#)
- [BEAST](#)
- [r8s-bootstrap](#)
- [MrBayes tree scanners](#)
- [T-REX](#)
- [MTgui](#)
- [MrModeltest](#)
- [BootPHYML](#)
- [Porn\\*](#)
- [Discovery Studio Gene](#)
- [ProtTest](#)
- [ModelGenerator](#)
- [Simplot](#)
- [MCS](#)
- [Permute!](#)
- [ELW](#)
- [MultiPhyl](#)
- [GHOSTS](#)
- [PaupUp](#)



# Tree-based sequence alignment

---

- [TreeAlign](#)
- [ClustalW](#)
- [MALIGN](#)
- [GeneDoc](#)
- [GCG Wisconsin Package](#)
- [TAAR](#)
- [Ctree](#)
- [DAMBE](#)
- [POY](#)
- [ALIGN](#)
- [DNASIS](#)
- [FootPrinter](#)
- [ALIFRITZ](#)
- [T-Coffee](#)
- [ArboDraw](#)



# Tree plotting/drawing

---

- [PHYLIP](#)
- [PAUP\\*](#)
- [TreeTool](#)
- [TreeView](#)
- [NJplot](#)
- [DendroMaker](#)
- [Tree Draw Deck](#)
- [Phylodendron](#)
- [ARB](#)
- [unrooted](#)
- [DAMBE](#)
- [TREECON](#)
- [Mavric](#)
- [TreeExplorer](#)
- [TreeThief](#)
- [Bionumerics](#)
- [FORESTER](#)
- [MacClade](#)
- [MEGA](#)
- [Mesquite](#)
- [Phylogenetic Tree Drawing](#)
- [APE](#)
- [T-REX](#)
- [TreeJuxtaposer](#)
- [Spectronet](#)
- [TreeSetViz](#)
- [Drawtree server](#)
- [TreeGraph](#)
- [Bosque](#)
- [ArboDraw](#)
- [PaupU](#)

# Analyzing particular types of data

Here you will find lists of programs that analyze types of data other than molecular sequence data

## ■ RAPDs, RFLPs, or AFLPs

- [tfpga](#)
- [RAPDistance](#)
- [Fingerprinting II Informatix Software](#)
- [GelCompar II](#)
- [Bionumerics](#)
- [Winboot](#)
- [REAP](#)
- [RESTDITE](#)
- [MVSP](#)
- [DENDRON](#)
- [Phyltools](#)
- [Network](#)

## ■ Continuous quantitative characters

- [PHYLIP](#)
- [Mesquite](#)
- [ANCMML](#)
- [COMPARE](#)
- [CMAP](#)
- [PDAP](#)
- [ACAP](#)
- [Phylogenetic Independence](#)
- [APE](#)
- [CAIC](#)
- [TreeScan](#)
- [PHYLOGR](#)
- [Continuous](#)

## ■ Gene frequencies (aside from microsatellite loci)

- [PHYLIP](#)
- [DAMBE](#)
- [DISPAN](#)
- [GDA](#)
- [POPGENE](#)
- [YCDMA](#)
- [FSTAT](#)
- [Arlequin](#)
- [DnaSP](#)
- [APE](#)
- [DIVAGE](#)
- [GeneStrut](#)
- [POPTREE](#)
- [Genepop](#)
- [SPAGeDi](#)

## ■ Microsatellite data

- [RSTCALC](#)
- [POPTREE](#)
- [Microsat](#)
- [Populations](#)
- [MSA](#)
- [YCDMA](#)
- [Network](#)
- [IM](#)



# PHYLIP

## Phylogeny Inference Package

---

- PHYLIP (**Phy**logeny **I**nference **P**ackage) is available free in Windows/MacOS/Linux systems.
- Parsimony, distance matrix and likelihood methods (bootstrapping and consensus trees).
- Data can be molecular sequences, gene frequencies, restriction sites and fragments, distance matrices and discrete characters.



# PHYLIP

## Phylogeny Inference Package

---

- PHYLIP (**Phy**logeny **I**nference **P**ackage) includes programs to carry out **parsimony, distance matrix methods, maximum likelihood, and other methods** on a variety of types of data including:
- DNA and RNA sequences, protein sequences, restriction sites, **0/1 discrete characters data**, gene frequencies, continuous characters and distance matrices.
- It is the **most widely-distributed phylogeny package**, with over 20,000 registered users, some of them satisfied.
- **It competes with PAUP\*** to be the program responsible for the most published trees.



# PHYLIP

## Phylogeny Inference Package

```
14 5
10 13 8 8 8
Roscoff 0.041 0 0.125 0.042 0.417 0.042 0.333 0 0
0.042 0 0.083 0.083 0.083 0 0.042 0.292 0 0.292 0 0
0 0.042 0 0.125 0.625 0.042
0.042 0 0.125 0.042 0.500 0.125 0
0.083 0.042 0.375 0.208 0.208 0.083 0
0 0.071 0.060 0.238 0.095 0.536 0 0
0.036 0 0.071 0.012 0 0.012 0.357 0.083 0.298 0.012 0.107
0 0.012 0.036 0.083 0.143 0.655 0.048
0 0.048 0.071 0.869 0.012 0
0.024 0.167 0.274 0.333 0.190 0.012
0 0.200 0.100 0.050 0.650 0 0
0 0.200 0.050 0 0.050 0 0.300 0 0.100 0.150 0.100
0 0.050 0.100 0.900 0
0 0.050 0.100 0.850 0
0 0.250 0.350 0.200 0.200 0
0 0 0.083 0.125 0 0.667 0.042 0.042
0.041 0 0.041 0.042 0 0 0.167 0.042 0.542 0.042 0.083
0 0.041 0 0.250 0.667 0.042
0 0.042 0.083 0.833 0.042 0
0.042 0.375 0.500 0 0.042 0.042
0 0 0.050 0.450 0 0.500 0 0
0 0.200 0.150 0.100 0 0.500 0 0.050 0 0
0.050 0.050 0.700 0.150 0.050 0 0
0 0.100 0 0.900 0
0 0 0.300 0.550 0.150 0
0 0 0 0.375 0.062 0.562 0 0
0 0.125 0.062 0.250 0 0 0.438 0 0.125 0 0
0 0.062 0 0 0.938 0
0 0 0.188 0.188 0.562 0 0.062
0.062 0.062 0.125 0.438 0.312 0
0 0 0 0.300 0.700 0 0
0 0 0 0 0.200 0 0.600 0 0
0 0 0 0 1.000 0
0 0 0.400 0.600 0 0
0 0.400 0.400 0.100 0.100 0
0 0 0.273 0.136 0.091 0.500 0 0
0 0.091 0.045 0 0 0.409 0.091 0.318 0 0.045
0 0 0.182 0.091 0.727 0
0 0.091 0.136 0.773 0 0
0.045 0.045 0.364 0.455 0 0.045 0.045
Uto 0.050 0.050 0.100 0.250 0.050 0.050 0.450 0 0
0 0 0 0 0.100 0.200 0.050 0.650 0 0
0 0 0.150 0.150 0.650 0.050
0 0.200 0.150 0.650 0 0
0 0.250 0.250 0.400 0.100 0
0 0.125 0.188 0.062 0 0.625 0 0
0 0 0.062 0 0.062 0 0.312 0 0.438 0.125 0
0 0 0 0.062 0.938 0
0 0 0.250 0.750 0 0
0 0.312 0.438 0.125 0.125 0
0 0 0 0.143 0.143 0.714 0 0
0 0.143 0 0 0 0.571 0 0.143 0.143 0
0 0 0 0.286 0.714 0
0 0 0 0.857 0.143 0
0.071 0.286 0.429 0.214 0 0
0 0 0.100 0.350 0 0.550 0 0
0 0 0 0 0.250 0 0.750 0 0
0 0 0.200 0.150 0.650 0
0 0.300 0.100 0.600 0 0
0.100 0.200 0.300 0.400 0 0
0.062 0.125 0.250 0 0 0.562 0 0
0 0.187 0 0 0 0.438 0 0.375 0 0
0 0 0 1.000 0
0 0 0.062 0.938 0 0
0 0.062 0.500 0.375 0.062 0
0 0 0.300 0.150 0 0.550 0 0
0 0.200 0.100 0 0.650 0 0.050
```

# PHYLIP

## Phylogeny Inference Package

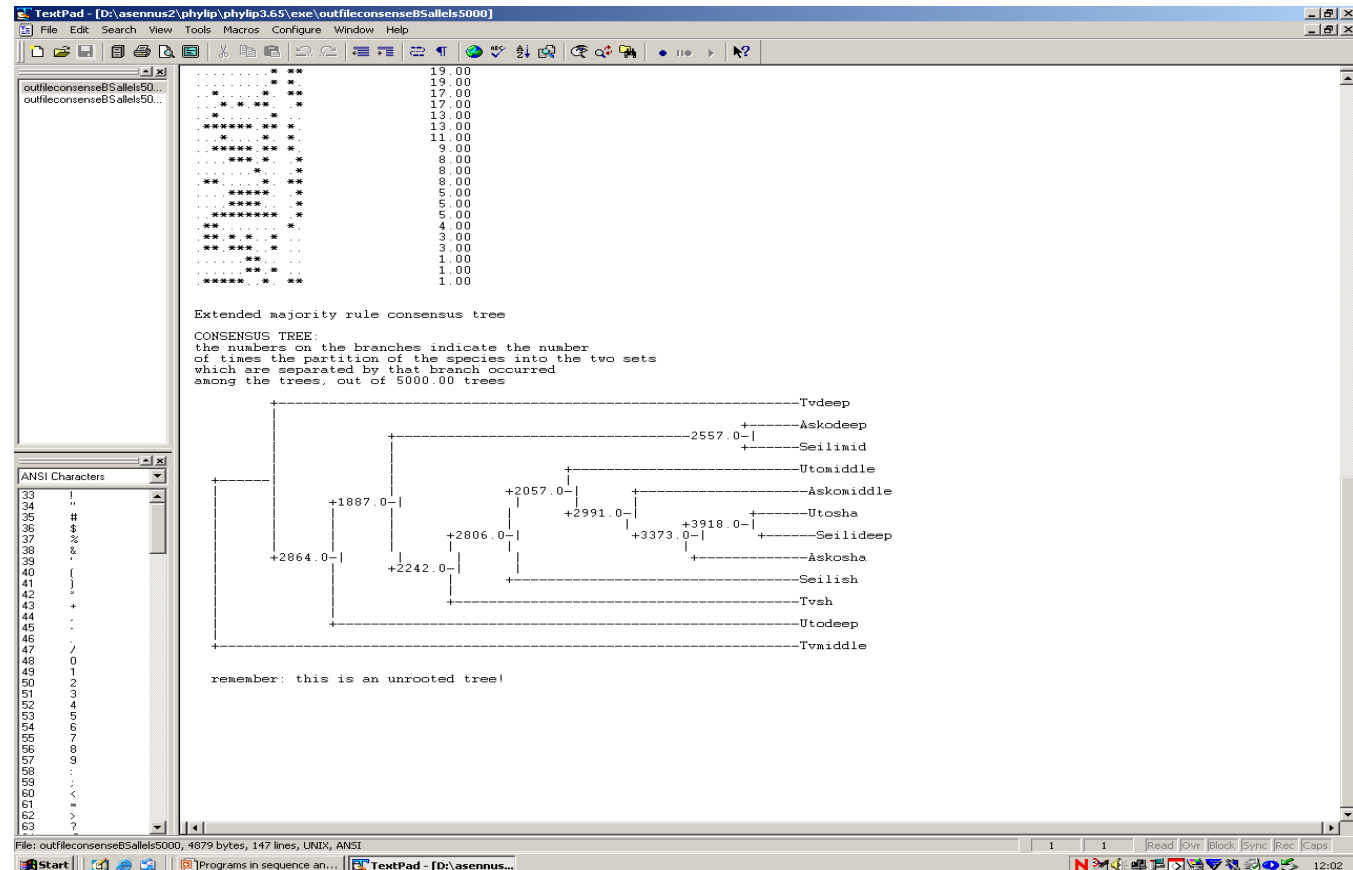
```
D:\asennus2\phylip\phylip3.65\exe\seqboot.exe

Bootstrapping algorithm, version 3.65

Settings for this run:
D      Sequence, Morph, Rest., Gene Freqs?      Molecular sequences
J      Bootstrap, Jackknife, Permute, Rewrite?   Bootstrap
%      Regular or altered sampling fraction?     regular
B      Block size for block-bootstrapping?      1 (regular bootstrap)
R      How many replicates?                     100
W      Read weights of characters?              No
C      Read categories of sites?                No
S      Write out data sets or just weights?     Data sets
I      Input sequences interleaved?             Yes
0      Terminal type (IBM PC, ANSI, none)?     IBM PC
1      Print out the data at start of run       No
2      Print indications of progress of run    Yes

Y to accept these or type the letter for one to change
```

# Phylogeny Inference Package



# MEGA

## Molecular Evolutionary Genetic Analysis

### MEGA 6

---

- MEGA is an integrated tool for conducting sequence alignment, inferring phylogenetic trees, estimating divergence times, mining online databases, estimating rates of molecular evolution, inferring ancestral sequences, and testing evolutionary hypotheses.
- MEGA is used by biologists in a large number of laboratories for:
  1. Reconstructing the evolutionary histories of species and
  2. Inferring the extent and nature of the selective forces shaping the evolution of genes and species.
  3. Clustal W is already built-in in MEGA 6.

# MEGA

## Molecular Evolutionary Genetic Analysis

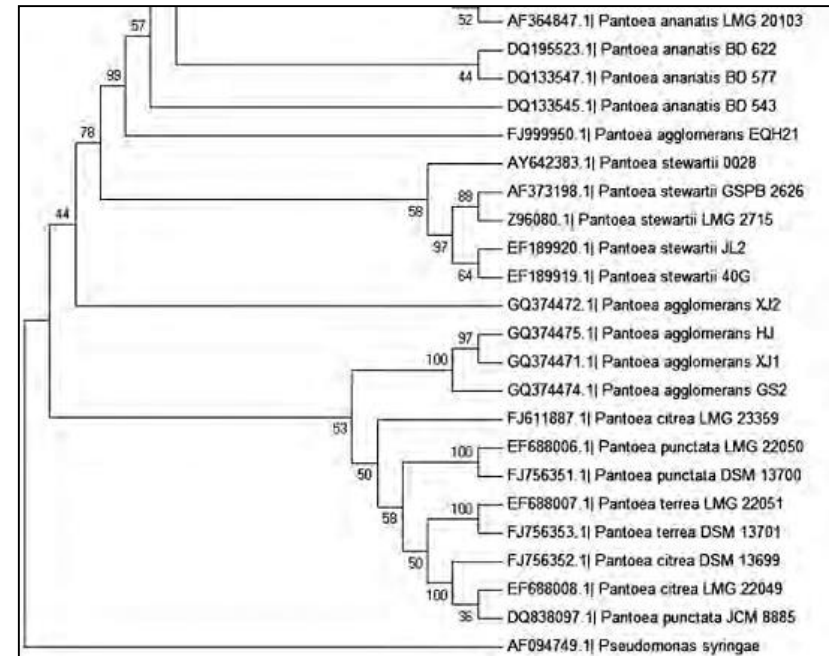
### MEGA 7



# PCR detection of *Pantoea* spp.

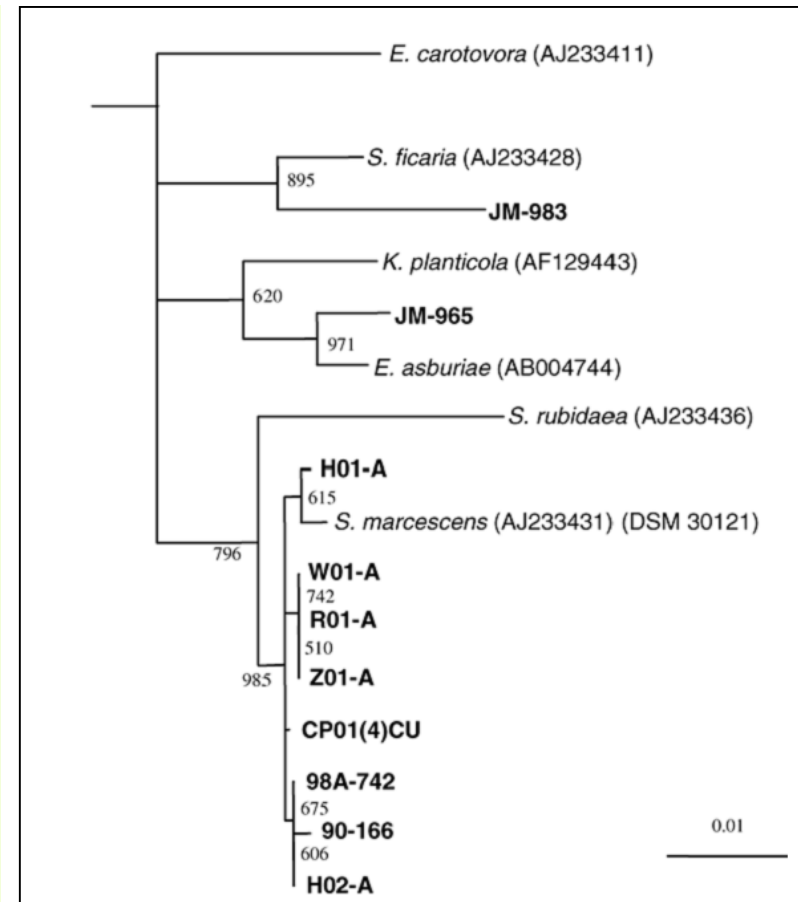
## Based on 16S rRNA sequences

- Dendrogram constructed by neighbor joining analysis of the 16S rRNA gene sequences from different *Pantoea* species and a *Pseudomonas syringae* strain sequence (AF094749) as an outgroup.
- The nucleotide sequences were analyzed using the BioEdit and Mega 4.0 software.
- Multiple sequence alignments were performed using the ClustalW program.
- Phylogenetic analysis was carried out by the neighbor joining algorithm implemented with Mega 4.0.
- Bootstrap values for phylogenetic comparisons were based on 1000 pseudoreplicates.



# PCR detection of *Serratia marcescens* Causing Cucurbit Yellow Vine Disease Based on 16S rRNA sequences

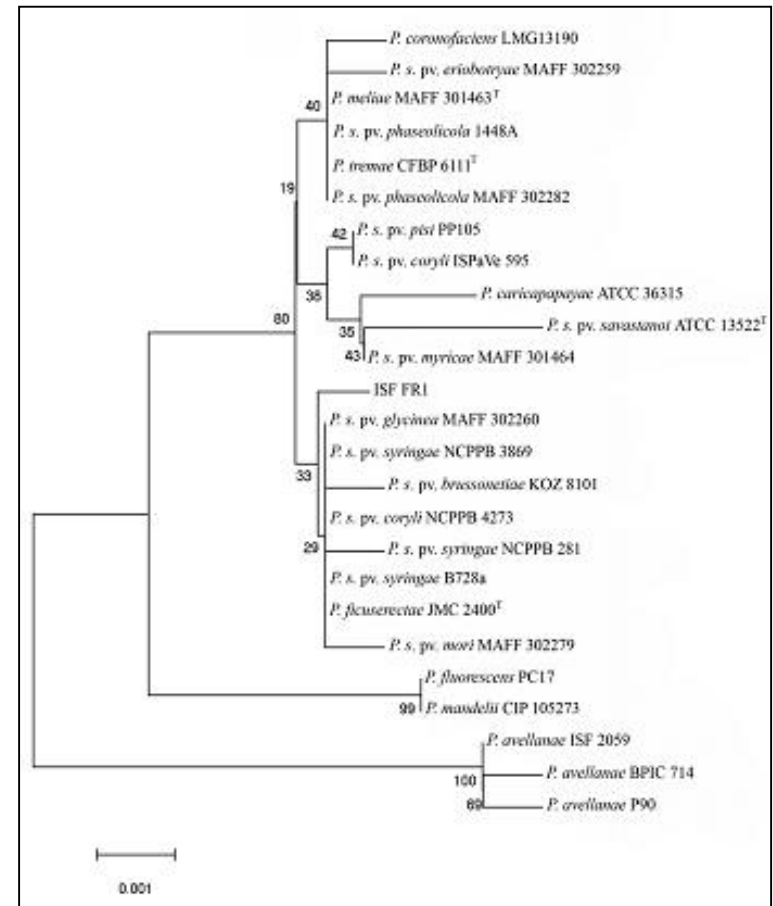
- Phylogenetic distance tree compiled from 16S rDNA sequence data using programs DNADIST and NEIGHBOR, with the endosymbiont of *Sitophilus oryzae* as outgroup.
- Branches with bootstrap values less than 500 were collapsed, and two branches (bootstrap values 510 and 606) were of relative lengths insufficient for resolution at the scale of this figure.
- Strains indicated in bold font were used in this study; the remainder are database reference strains (NCBI RefSeq 16S rRNA database).



# PCR detection of *P. syringae*

## Based on 16S rRNA sequences

- Dendrogram based on 16S rDNA gene sequences of endophytic *Pseudomonas syringae* ISF FR1, *P. syringae* pathovars and *Pseudomonas* spp. obtained with neighbor-joining algorithm.
- Multiple alignment of 16S rDNA sequences were performed using the ClustalW algorithm.
- Cluster analysis was conducted using MEGA, version 3.1 (Kumar *et al.*, 2004) software.
- The scale bar represents the number of substitutions in each sequence.
- Bootstrap values (1,000 replicates) are also shown.





# The calculation of association coefficients for two organisms

Phylogenetic relationships are not measured with a simple coefficient.

## The Calculation of Association Coefficients for Two Organisms

In this example, organisms A and B are compared in terms of the characters they do and do not share. The terms in the association coefficient equations are defined as follows:

		Organism B	
		1	0
Organism A	1	a	b
	0	c	d

a = number of characters coded as present (1) for both organisms

b and c = numbers of characters differing (1,0 or 0,1) between the two organisms

d = number of characters absent (0) in both organisms

Total number of characters compared =  $a + b + c + d$

The simple matching coefficient ( $S_{SM}$ ) =  $\frac{a + d}{a + b + c + d}$

The Jaccard coefficient ( $S_J$ ) =  $\frac{a}{a + b + c}$



# Examples of Phylogenetic analyses

---

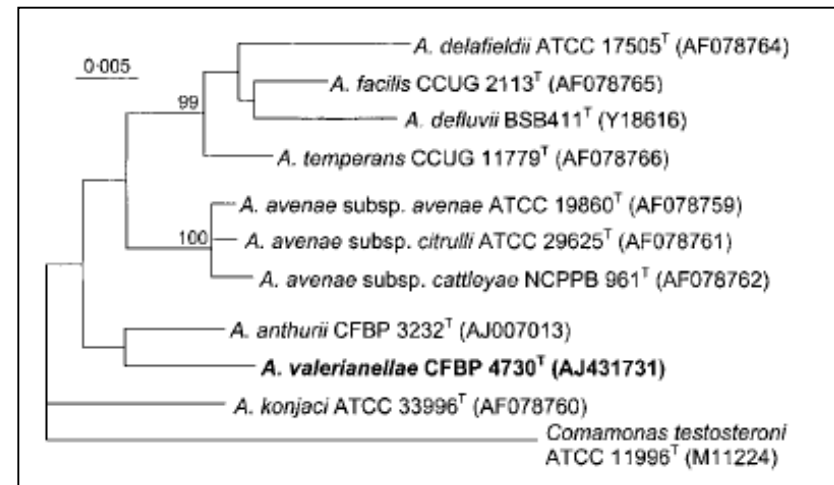
**Mainly based upon:**

- **16S rDNA sequences**
- **16S-23S rDNA intergenic spacer sequences(ITS)**

# Phylogenetic analysis of *Acidovorax* species based upon 16S rDNA sequences

## Neighbor(neighbour)-joining tree

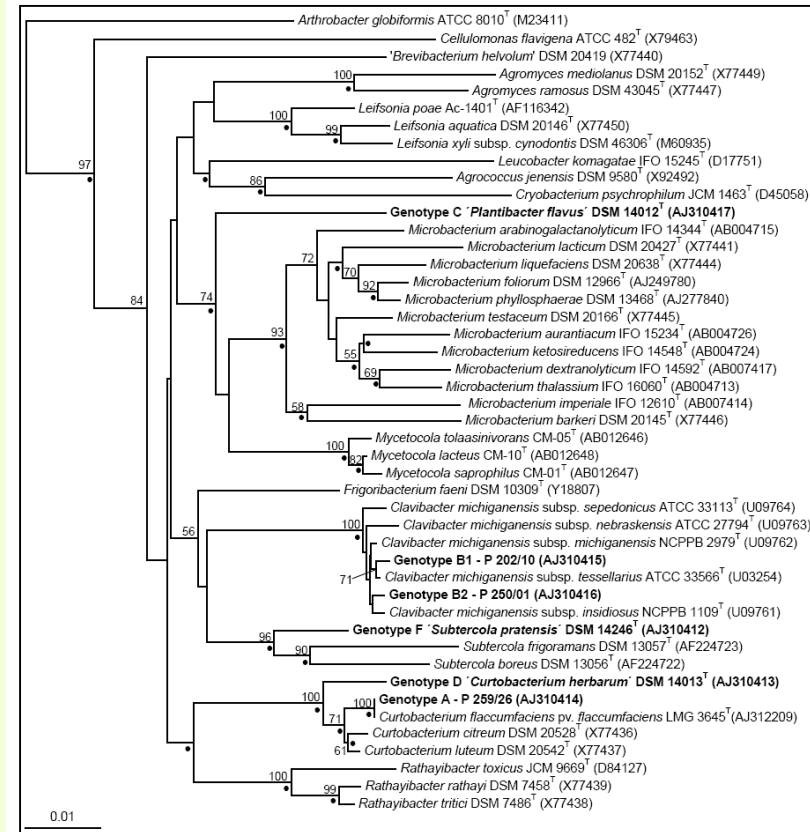
- Neighbor-joining tree obtained from 16S rRNA gene sequences.
- The scale bar represents 1 estimated base substitution per 200 nucleotide positions.
- Percentages refer to bootstrap values of 100 calculated trees.
- EMBL/GenBank accession numbers are shown in parentheses.
- An expanded version of this tree, showing more taxa, is available as supplementary material in IJSEM Online.



Taxon labels in bold indicates  
***A. valerianellae* strain CFBP.**

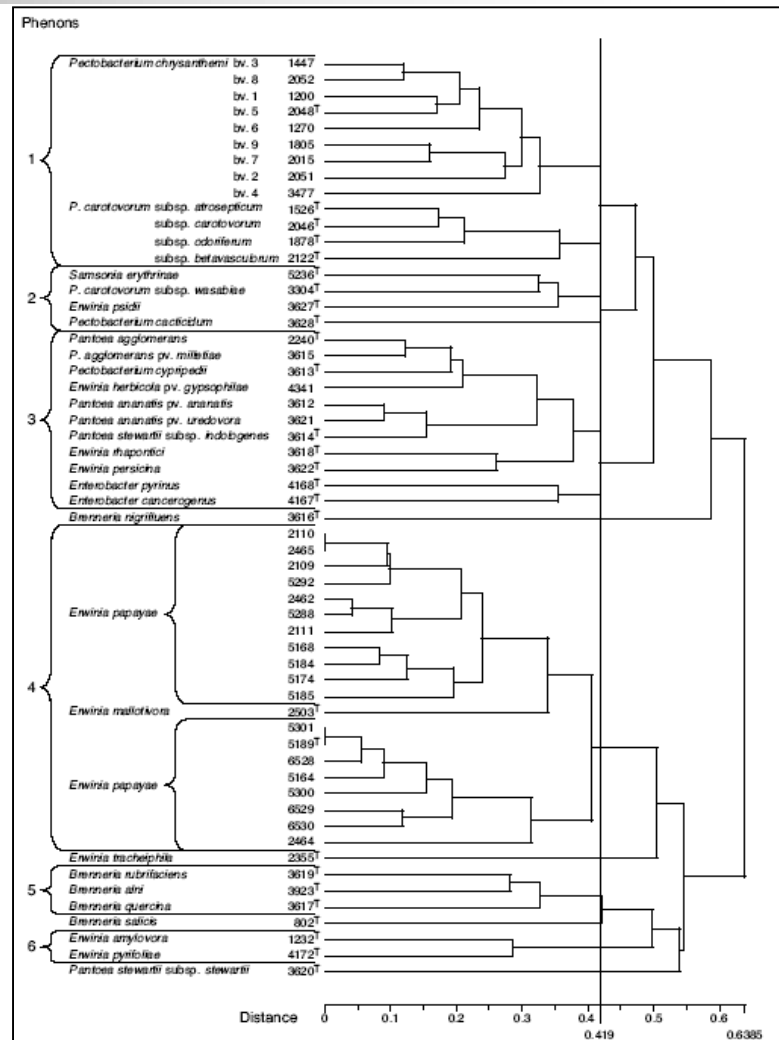
# Phylogenetic analysis of **coryneforms** based upon 16S rDNA sequences

- Phylogenetic tree showing the relationship of the isolated genotypes within the family *Microbacteriaceae*.
- The tree is based on a 1486 bp alignment of the **16S rDNA sequences** and was constructed using **Neighbor-Joining method** (Saitou & Nei, 1987).
- Dots indicate branches of the tree that were also formed using the **Maximum likelihood method** (Felsenstein, 1981).
- To estimate the root position of the tree, *Brevibacterium linens* was used as an outgroup.
- The values are the number of time that a branch appeared in 100 bootstrap replications.
- Strains characterized in this study are in bold characters.



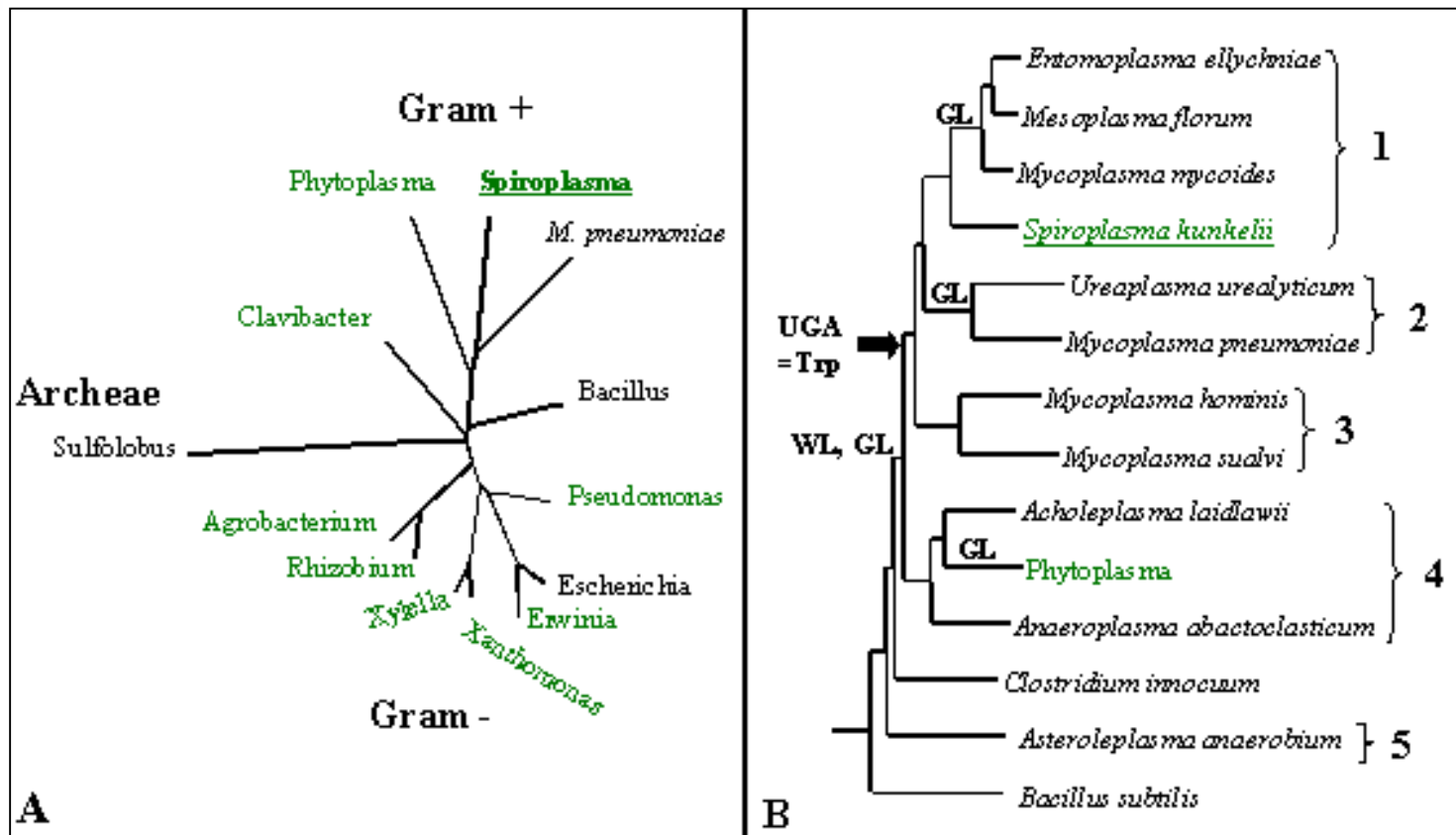
# Phylogenetic analysis of *Erwinia* species based upon 16S rDNA sequences

- Rooted tree, subset of a larger tree available as supplementary material, result of a **neighbor-joining bootstrap analysis** (1000 replications).
- **Bootstrap percentages** are indicated only for branches that were retrieved also by MP (strict consensus of 6 equally parsimonious trees) and ML at  $P < 0.01$ , therefore indicating robust clades. Or
- **ML analysis, in likelihood** was 6492 and 4370 trees examined.



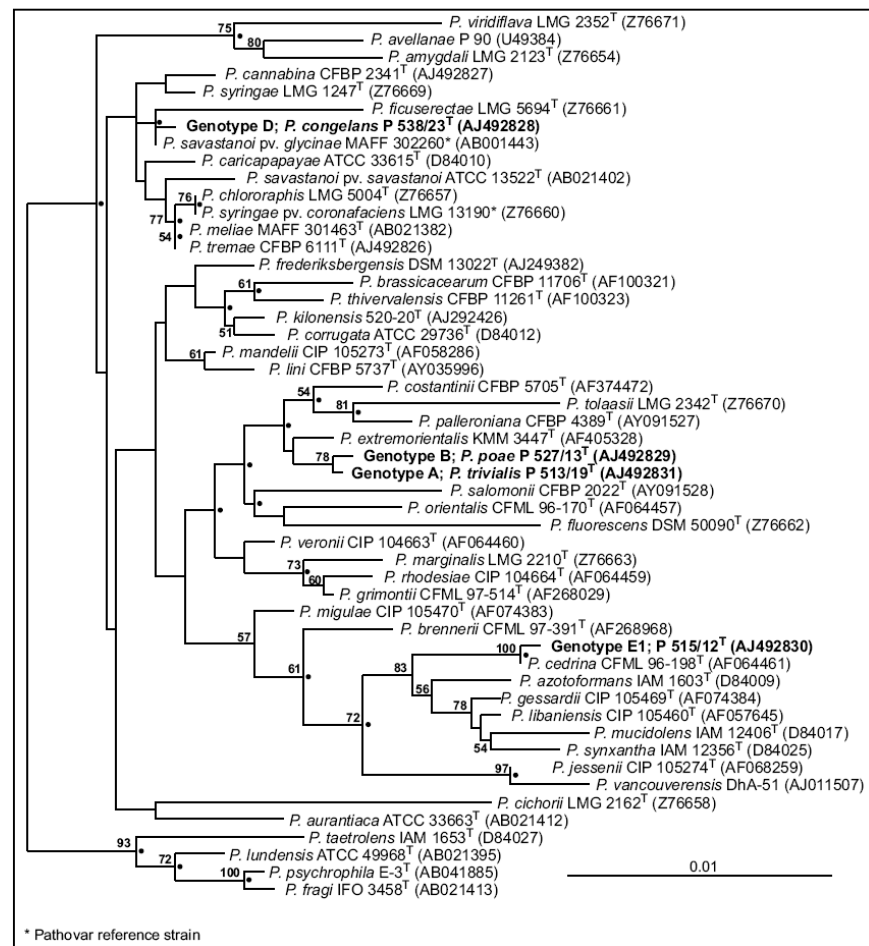
# Phylogenetic relationships of certain bacterial clades

## Five Phylogenetic groups in class Mollicutes



# Phylogenetic analysis of *Pseudomonas* species based upon 16S rDNA sequences

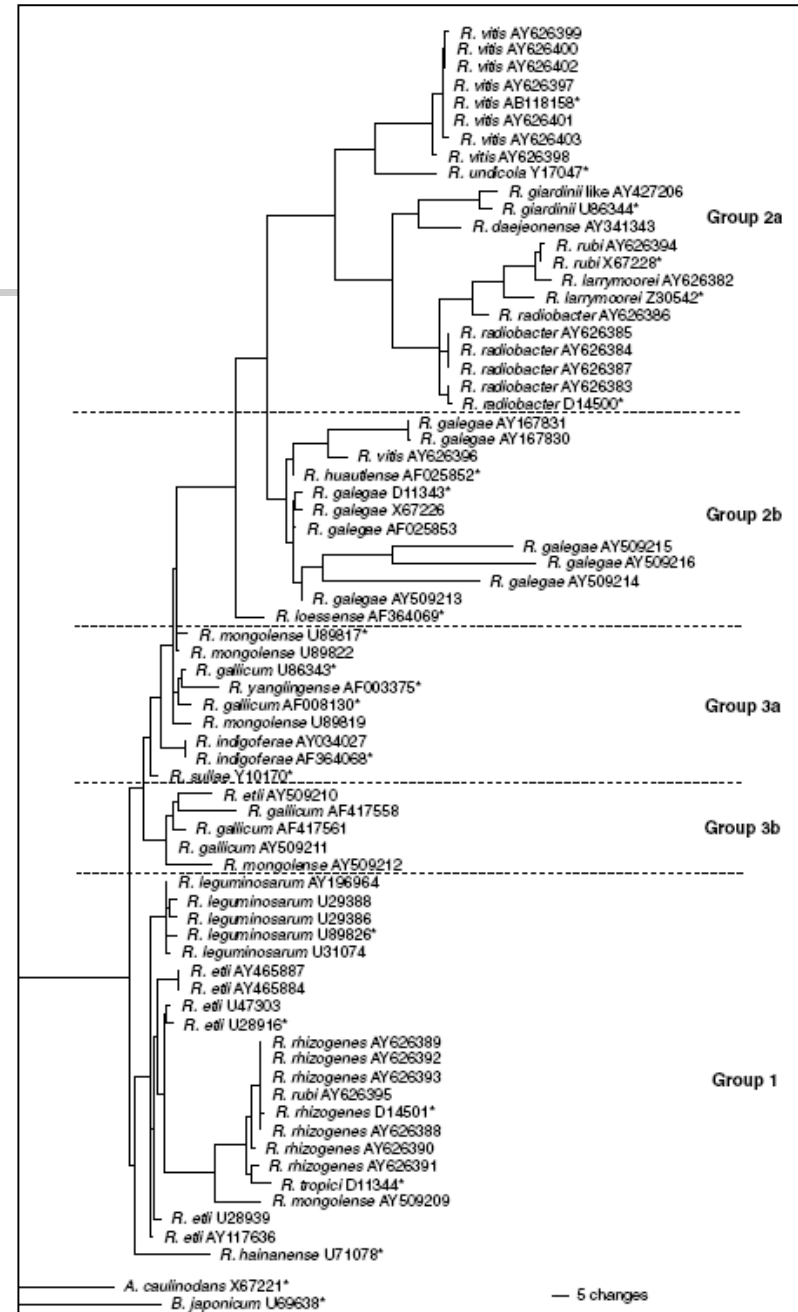
- The tree is based on a 1282 bp alignment of **16S rDNA sequences** and was constructed using the **neighbour joining method**.
- Dots indicate branches of the tree that were also formed using the **maximum-likelihood method**.
- To estimate the root position of the tree, *E. coli* (accession no. J01695) was used as an **outgroup**.
- The values are number of time that a branch appeared in 100 bootstrap replications.
- Strains characterized in this study are in bold.
- Bar, relative sequence divergence.



# Phylogenetic relationships of *Rhizobia* and related species based on 16S rDNA sequence analyses

- Inferred relationships of species in the genus *Rhizobium* using Maximum Likelihood.
- Sequences from type strains are marked \*.
- There is no significant internal division of the *Rhizobium* clade to suggest that it represents more than one genus.
- Plant pathogenic (*Agrobacterium*) species are distributed within the genus.

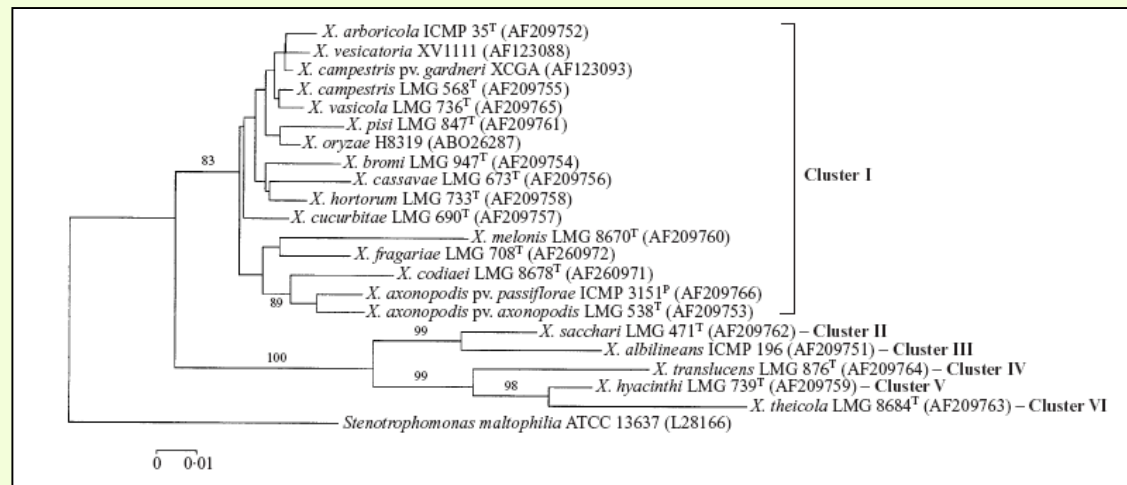
Young *et al.*, 2004





# Phylogenetic analysis of *Xanthomonas* species based upon 16S-23S rDNA ITS sequences

- ITS sequences were aligned using the **clustal w** program.
- Evolutionary distances were obtained by the **p-distance** method.
- Topology of the phylogenetic tree was assessed by the **neighbour-joining method** and **bootstrap values** were obtained from 2000 replicates using the **mega**.
- Bar, 0±01 changes per nucleotide.



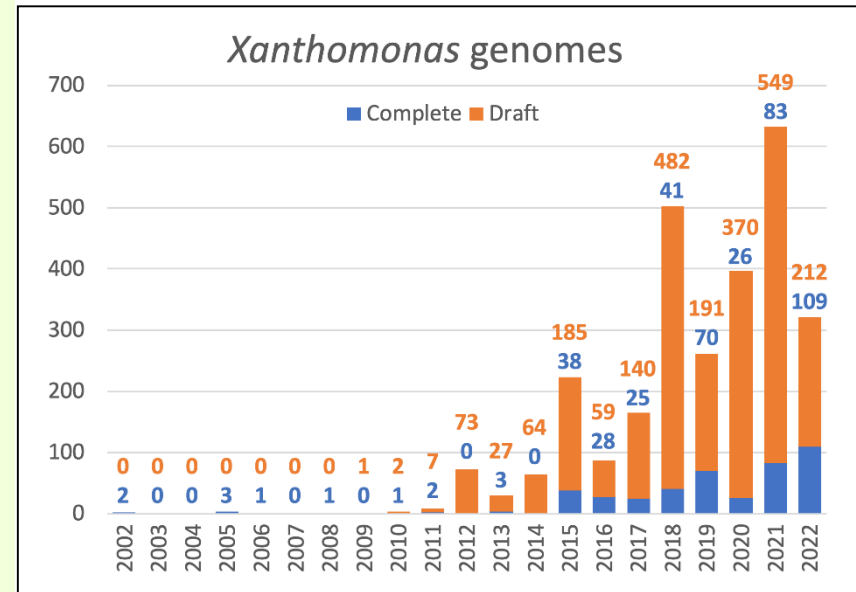
# Celebrating the 20th anniversary of the first *Xanthomonas* genome sequences



- Celebrating the 20th anniversary of the first *Xanthomonas* genome sequences– how genomics revolutionized taxonomy, provided:
  1. insight into the emergence of pathogenic bacteria,
  2. enabled new fundamental discoveries, and
  3. helped developing novel control measures – a perspective from the French network on *Xanthomonads*.

# Celebrating the 20th anniversary of the first *Xanthomonas* genome sequences

- NCBI *Xanthomonas* genome statistics (as of 13 July 2023).  
*Xanthomonas* genome assembly
- metadata were extracted from NCBI GenBank at <https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=338>.
- GenBank assembly levels 'Contig', 'Scaffold' and 'Chromosome' were considered together as Draft level.
- The complete list of genomes and relevant metadata are available in Supplementary Table S1.



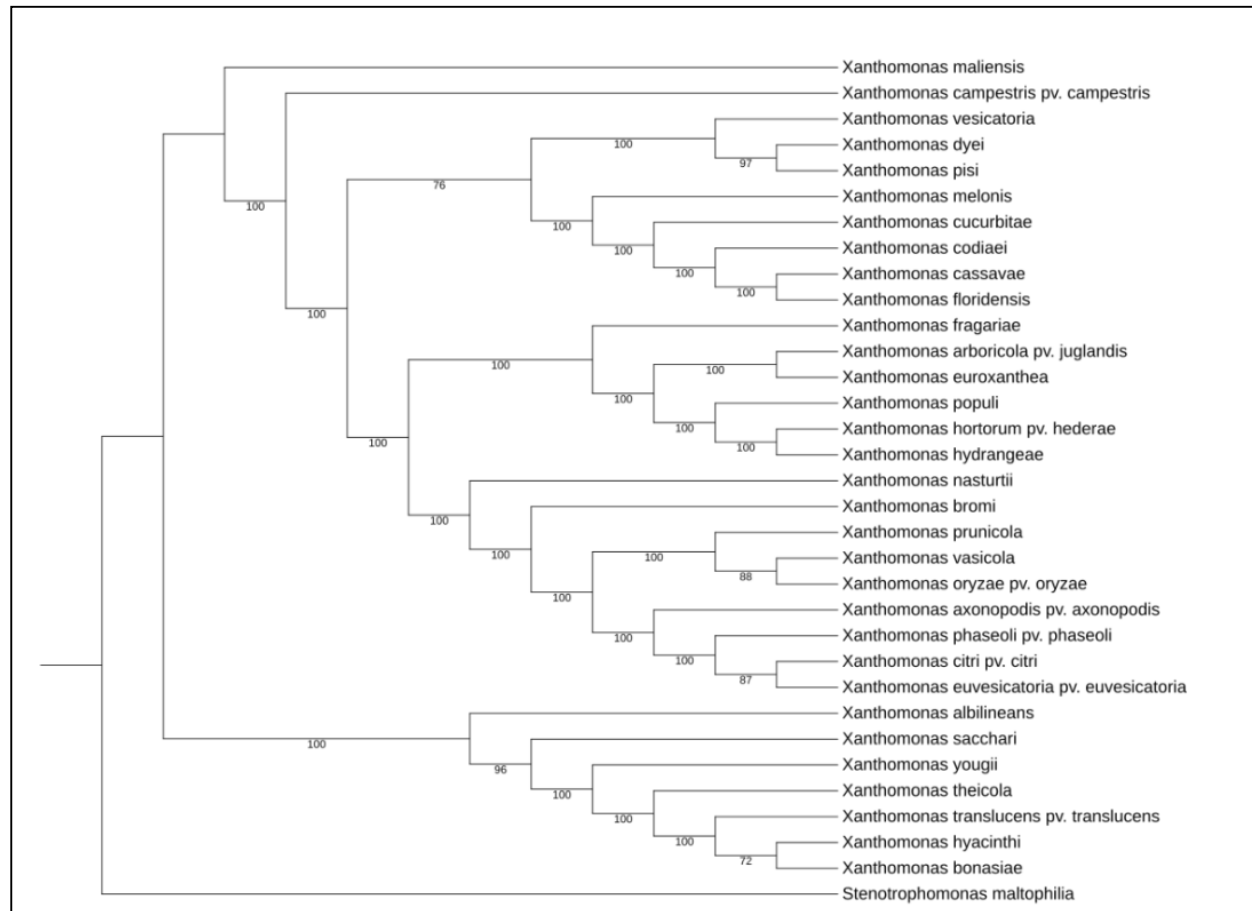
# Celebrating the 20th anniversary of the first *Xanthomonas* genome sequences

- Phylogenetic tree of the 32 valid species of *Xanthomonas* provided after TYGS analysis (Meier-Kolthoff *et al.*, 2022).
- Tree inferred with FastME 2.1.6.1 (Lefort *et al.*, 2015) from GBDP distances calculated from genome sequences retrieved from Genbank.
- The branch lengths are scaled in terms of GBDP distance formula d5.
- The numbers on branches are GBDP pseudo-bootstrap support values > 70% from 100 replications, with an average branch support of 97.2% (Farris, 1972).
- The Newick file was edited in iTOL (<https://itol.embl.de/>) and rooted on the outgroup *Stenotrophomonas maltophilia*.
- The complete list of genomes and GenBank Assembly accession numbers are available in Supplementary Table S2.

# Celebrating the 20th anniversary of the first *Xanthomonas* genome sequences

- Phylogenetic tree of the 32 valid species of *Xanthomonas* provided after TYGS analysis (Meier-Kolthoff *et al.*, 2022).
- Tree inferred with FastME 2.1.6.1 (Lefort *et al.*, 2015) from GBDP distances calculated from genome sequences retrieved from Genbank.
- The branch lengths are scaled in terms of GBDP distance formula d5.
- The numbers on branches are GBDP pseudo-bootstrap support values > 70% from 100 replications, with an average branch support of 97.2% (Farris, 1972).
- The Newick file was edited in iTOL (<https://itol.embl.de/>) and rooted on the outgroup *Stenotrophomonas maltophilia*.
- The complete list of genomes and GenBank Assembly accession numbers are available in Supplementary Table S2.

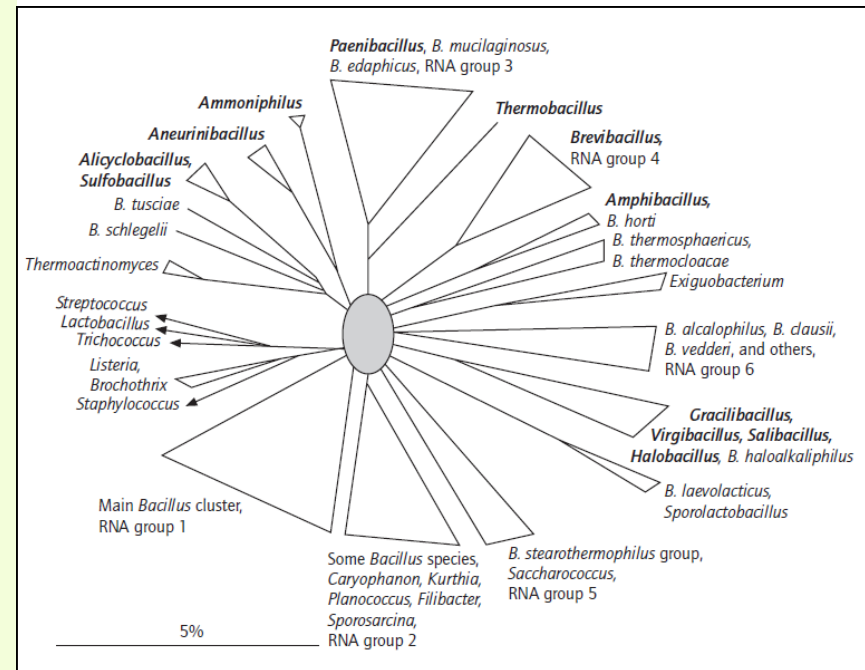
# Celebrating the 20th anniversary of the first *Xanthomonas* genome sequences



# Phylogeny

## *Bacillus* and novel genera originated from genus *Bacillus*

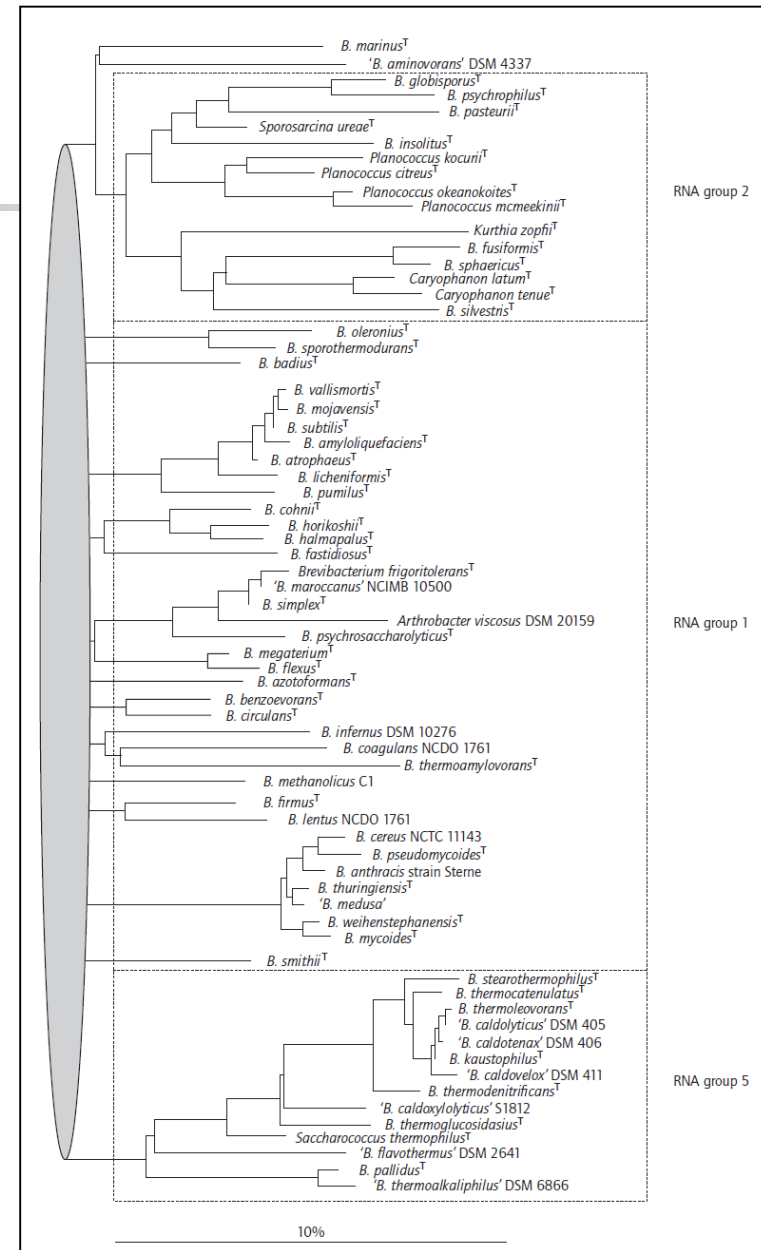
- Schematic outline of the phylogenetic diversity of 16S rDNA of aerobic, rod shaped and spore-forming, Gram-positive bacteria, classified as species of *Bacillus*, genera that originated from the dissection of *Bacillus*, and species that were affiliated to novel genera because of their distinct phylogenetic positions.
- *Bacillus* species were found to form clusters that have been named RNA groups 1 to 6.



# Phylogenetic relationships of **Bacilli** RNA groups

- Detailed neighbour-joining tree of **species of RNA groups 1, 2 and 5**.
- The dotted area indicates the uncertainty of the order at which the lineages diverge from each other.
- The area was chosen somewhat arbitrarily and may just as well cover more recent branching points.
- The bar indicates 10% nucleotide substitutions.
- **B**, *Bacillus*; **T**, type strain.

Berkeley *et al.*, 2002







# Glossary of general terms

---

- **Analogue:** An organ or structure that is similar in function to one in another kind of organism but is of dissimilar evolutionary origin.
- **Bioinformatics:** Bioinformatics have become an essential tool not only for basic research but also for applied research in biotechnology and biomedical sciences (Kamel, 2003).
- **Bioinformatics** is an emerging scientific discipline that uses information technology to organize, analyze, and distribute biological information in order to answer complex biological questions.
- **Bioinformatics** is an interdisciplinary research area, which may be broadly defined as the interface between biological and computational sciences (Singh and Kumar, 2001).
- **Bioinformatics** programs that used to process the interest sequence against those deposited in the database such as **Gene Runner version 3.05**, Basic Local Alignment Search Tools (BLASTn) and **Ribosomal Database Project (RDP)**.
- **Cenancestor:** An alternative term for the **Last Common Ancestor** of all life on Earth.



# Glossary of general terms

---

- **Clusters of Orthologous Groups of Proteins (COGs):** Phylogenetic classification of proteins encoded in complete genomes.
- **Dendrogram:** A branching diagram that shows the relative sequence similarity between many different proteins or genes to indicate the **phylogenetic relationships**; typically horizontal lines indicate the degree of differences in sequences, while vertical lines are used for clarity to separate branches.
- **Domain:** The highest taxonomic division in the classification of living organisms. The three domains are the **Archaea**, the **Bacteria** and the **Eucarya**. Domains are subdivided into kingdoms. While the three domain model is widely used in astrobiology, **some biologists** prefer other schemes such as the **Five-Kingdom system**.
- **Eubacteria:** An alternative name for the **domain bacteria** (or true bacteria).
- The **electropherogram** is a graphical representation of data received from a sequencing machine and is also known as a **trace**.
- **Gene flow:** Movement of genes (under examination) through specific process, from one population to another population geographically separated apart.
- **Genetic polymorphism:** The stable, long term existence of multiple **alleles** at a gene locus. Technically a locus is said to be polymorphic if the most common **homozygote** occurs at a frequency of less than **90% in the population**.



# Glossary of general terms

---

- **Homologous:** Diploid organisms that has inherited the same **allele** from both parents ie carries identical alleles at the corresponding sites on **homolgous** chromosomes.
- **Homology:** Similarity attributed to descent from a common ancestor.
- **Last Common Ancestor:** The last common ancestor of all organisms living today. **The root of the tree of life.**
- **Lateral Gene Transfer:** The transfer of genes between different species. Lateral gene transfer may have been widespread in the early stages of life on Earth and this complicates the **interpretation of the tree of life.**
- **LUCA:** Another term used for the **Last Common Ancestor** of all living organisms. Acronym for Last Universal Common Ancestor.
- **Monophyletic group:** Derived from a common ancestor. Taxa derived from and including a single founder species.
- **Orthologous/orthologue/orthology:** **Genes in different species** that are homologous (similar) because they are **derived from a common ancestral gene** (during speciation).
- **Open reading frame (ORF):** A DNA sequence lying between start and stop codons which is capable of transcription.



# Glossary of general terms

---

- **Paralogous/paralogue/paralogy:** Two genes from the same organism which are similar because they derive from a gene duplication.
- **Paraphyletic group:** Groups which have evolved from and include a single ancestral species (known or hypothetical) but which do not contain all the descendants of that ancestor.
- **Polymorphism:** The existence within a species or a population of different forms of individuals, ...
- **Polyphyletic group:** A group that does not include the common ancestor of the group. The common ancestor is placed in another group or a taxonomic group having origin in several different lines of descent.
- **Pre-RNA World:** A hypothetical early stage in the development of life which preceeded the RNA World and used some other genetic material in place of RNA or DNA.
- **RNA polymerase:** The basic structure of RNA polymerase consists of four polypeptides – two identical  $\alpha$  chains plus two other chains ( $\beta$  and  $\beta'$ ) that are related to one another but are not identical.



# Glossary of general terms

---

- **RNA World:** A hypothetical early stage in the development of life in which RNA molecules provided both the genome and the catalysts, roles which subsequently were taken over by DNA and proteins.
- **Ribotyping:** Restriction fragment length polymorphism analysis of rRNA genes that is used for differentiating between species or strains.
- **Tree of Life:** A phylogenetic tree covering all groups of life on Earth. The term is commonly used for the tree derived by molecular phylogeny using small sub-unit ribosomal RNA as pioneered by Carl Woese in the 1970s.



# Selected References

- **Abedone, S.T. Power point presentations and lecture notes.** The Ohio State University. [www.phage.org](http://www.phage.org). [abedon.1@osu.edu](mailto:abedon.1@osu.edu).
- **Barbieri, M. 1981. The Ribotype Theory on the Origin of Life.** Journal of Theoretical Biology, 91, 545-601.
- **Barbieri, M. 1985. The Semantic Theory of Evolution.** Harwood Academic Publishers, London and New York.
- **Barbieri, M. 1998. The Organic Codes. The basic mechanism of macroevolution.** Rivista di Biologia-Biology Forum, 91, 481-514.
- **Barbieri, M. 2001. The Organic Codes. The birth of semantic biology.** Pequod, Ancona. (new edition to be published by Cambridge University Press).
- **BIOLOGY 303: MICROBIOLOGY. LECTURE 12: Bacterial taxonomy and phylogeny II, 2004.**
- **Bergey's Manual of Determinative Bacteriology, 9th edition (1993).**
- **Bergey's Manual of Systematic Bacteriology:** This is a set of 4 volumes, The 1st edition appeared in installments, from 1984-1989. The first volume of the 2nd edition appeared in 2001, and Volume 2 is "coming soon".
- **Balch, W.E.; Magrum, L.J.; Fox, G.E.; Wolfe, C.R.; & Woese, C.R. (August 1977). An ancient divergence among the bacteria.** J. Mol. Evol. 9 (4): 305-11.



# Selected References

---

- Burki, F., Y. Inagaki, J. Brate, J. M. Archibald, P. J. Keeling, T. Cavalier-Smith, M. Sakaguchi, T. Hashimoto, A. Horak, S. Kumar, D. Klaveness, K. S. Jakobsen, J. Pawlonski, and K. Shalchian-Tabrizi. 2009. Large-scale phylogenomic analyses reveal that two enigmatic protist lineages, Telonemia and Centroheliozoa, are related to photosynthetic chromalveolates. *Genome. Biol. Evol.* 1(1): 231-238.
- Cairns-Smith, A.G. 1993. *Seven Clues to the Origin of Life: A Scientific Detective Story*. pg. 44-45 Cambridge University Press.
- Cavalier-Smith, T. 1998. A revised six-kingdom system of life. *Biological Reviews* 73 (03): 203-66.
- Cavalier-Smith, T. 2002. The neomuran origin of archaeobacteria. *Int.J.Sys.Env.Mic.* 52:7-76.
- Cavalier-Smith, T. 2004. Only six-kingdoms of life. *Proc. R. Soc. Lond. B* 271 (1545): 1251-62.
- Cavalier-Smith, T. 2006. Cell evolution and earth history: stasis and revolution. *Phil. Trans. Roy. Soc. Lond. B.* 361, 969-1006.
- Cavalier-Smith, T. 2006. Rooting the tree of life by transition analysis. *Biol. Direct* 1: 19.
- Cooper, G.M. 2000. *The cell: A Molecular Approach* 2d ed. Amer. Soc. Microbiol., Washington and Sinauer Assoc., Sunderland, MA.



# Selected References

---

- Cook, B.M.; D. Gareth Jones and B. Kaye (Eds.). 2006. **The Epidemiology of plant diseases**. Second Edition, Springer, 576 pp.
- Deacon, J. 2003. **The Microbial World Microorganisms and microbial activities**.
- Gao, B. and R. S. Gupta. 2007. **Phylogenomic analysis of proteins that are distinctive of *Archaea* and its main subgroups and the origin of methanogenesis**. BMC Genomics 8:86.
- Gupta, R. S. 1998. **Protein Phylogenies and Signature Sequences: A Reappraisal of Evolutionary Relationships Among Archaeobacteria, Eubacteria, and Eukaryotes**. Microbiol.Mol.Biol.Rev. 62: 1435-1491.
- Gupta, R. S. and E. Griffiths. 2002. **Critical Issues in Bacterial Phylogenies**. Theor.Popul.Biol. 61:423-434.
- Gupta, R. S. 2011. **Origin of diderm (Gram-negative) bacteria: antibiotic selection pressure rather than endosymbiosis likely led to the evolution of bacterial cells with two membranes**. Antonie van Leeuwenhoek.
- Hanno Teeling and Frank Oliver Gloeckner. 2006. **RibAlign: A software tool and database for eubacterial phylogeny based on concatenated ribosomal protein subunits**. BMC Bioinformatics.
- Kersters *et al.*, 2003. **Introduction to the Proteobacteria**.
- Luskin, C. 2004. **Problems with the Natural Chemical "Origin of Life"** . IDEA Center.





# Selected References

---

- Olsen, G.J. **Classification and Phylogeny**. Bacteriology at UW-Madison.
- Opperdoes, F. 1997. Construction of a distance tree using clustering with the Unweighted Pair Group Method with Arithmetic Mean (UPGMA).
- Parkinson, N. Identifying Relatedness Between Bacterial Plant Pathogens. Parkinson\_MolID\_CSL\_1.pdf.
- **Phylogeny Programs**. htm.
- **Unit Four - Phylogeny of Bacteria**
- **Plant Systematic Methodology**. htm. **Molecular Systematics**. Part II Taxonomy.
- Poole, A. 2006. **LUCA , the Last Universal Common Ancestor of all life**. Stockholm Univ., Sweden.
- Prescott *et al.*, 2005. **Microbiology- Prokaryotes: Bacterial Genetic Systems**.
- Ruggiero, M.A., D.P. Gordon, T.M. Orrell, N. Bailly, T. Bourgoin, R.C. Brusca, T. Cavalier-Smith, M.D. Guiry and P.M. Kirk. 2015. **Correction: A Higher Level Classification o f All Living Organisms**. PLoS ONE 10(6): e0130114. doi:10.1371/journal.pone.0130114.
- **RNA-DNA**.htm. 2001. **DNA, RNA, PNA; stepping backwards in time**. University of California **Tree of Life design and icons copyright© 1995-2004 Tree of Life Project**.
- **Taxon 1**.htm. **Classification and evolution of microbes**.



# Selected References

- **Taxonomy.** <http://bricker.tcnj.edu/micro/le7/ribo.gif>.
- **Van Niel, C.B. 1946. The classification and natural relationships of bacteria.** Cold Spring Harbor Symp 11, 285-301.
- **Woese, C.R., Fox, G.E., Zablen, L., Uchida, T., Bonen, L., Pechman, K., Lewis, B.J., Stahl, D. 1975. Conservation of primary structure in 16S ribosomal RNA.** Nature 6;254(5495): 83-86.
- **Woese, CR, GE, Fox (November 1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms.** Proc. Natl. Acad. Sci. U.S.A. 74 (11): 5088-90.
- **Woese CR, Fox GE. 1977. The concept of cellular evolution.** J. Mol. Evol. 1977 Sep 20;10(1):1-6.
- **Woese, CR, Magrum, LJ, GE Fox. 1978. Archaeobacteria.** J. Mol. Evol. 11:245-252.
- **Woese, CR, Gutell, R, Gupta R, HF. Noller. 1983. Detailed analysis of the higher-order structure of 16S-like ribosomal ribonucleic acids.** Microbiol. Rev. Dec; 47(4): 621-669.
- **Woese C, Kandler O, Wheelis M. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya.** Proc. Natl. Acad. Sci. USA 87 (12): 4576-9.
- **Woese, C.R. 2000. Interpreting the universal phylogenetic tree.** Proc. Natl. Acad. Sci. USA 97(15): 8392-6.
- **Woese, C.R. 2002. On the evolution of cells.** Proc. Natl. Acad. Sci. USA 99(13):8742-7.
- **Woese, C. R. 2004. A new biology for a new century.** MMBR. June 68(2):173-86.