

# Last Universal Common Ancestor

”Nothing in Biology Makes Sense  
Except in the Light of Evolution”



Theodosius Dobzhansky (1900-1975)

**Dephney Mathebula**

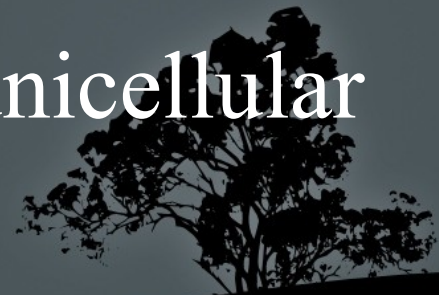
**Dieter Winkler**

**Hloniphile Sithole**



# INTRODUCTION

- LUCA Last Universal Common Ancestor
- Organism from which all organisms living on Earth descends.
- Have lived aprox. 3.5 to 3.8 billion years ago
- Named by Carl Woese, 1999
- hypothetical and controversial unicellular organism or single cell



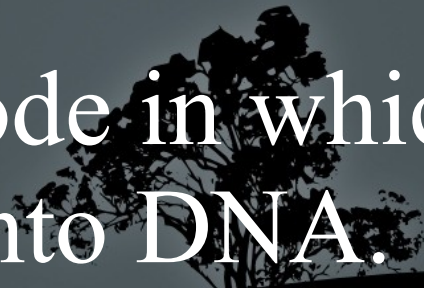
# Last Universal Common Ancestor

Question:

How do we know that all life evolved from a single cell?

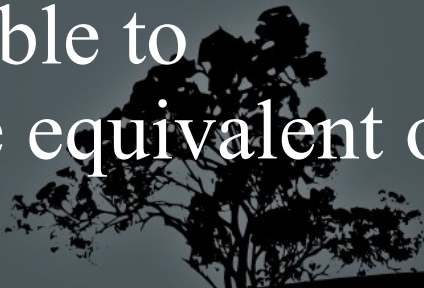
Answer:

It is written in the genetic code in which most genes are written into DNA.



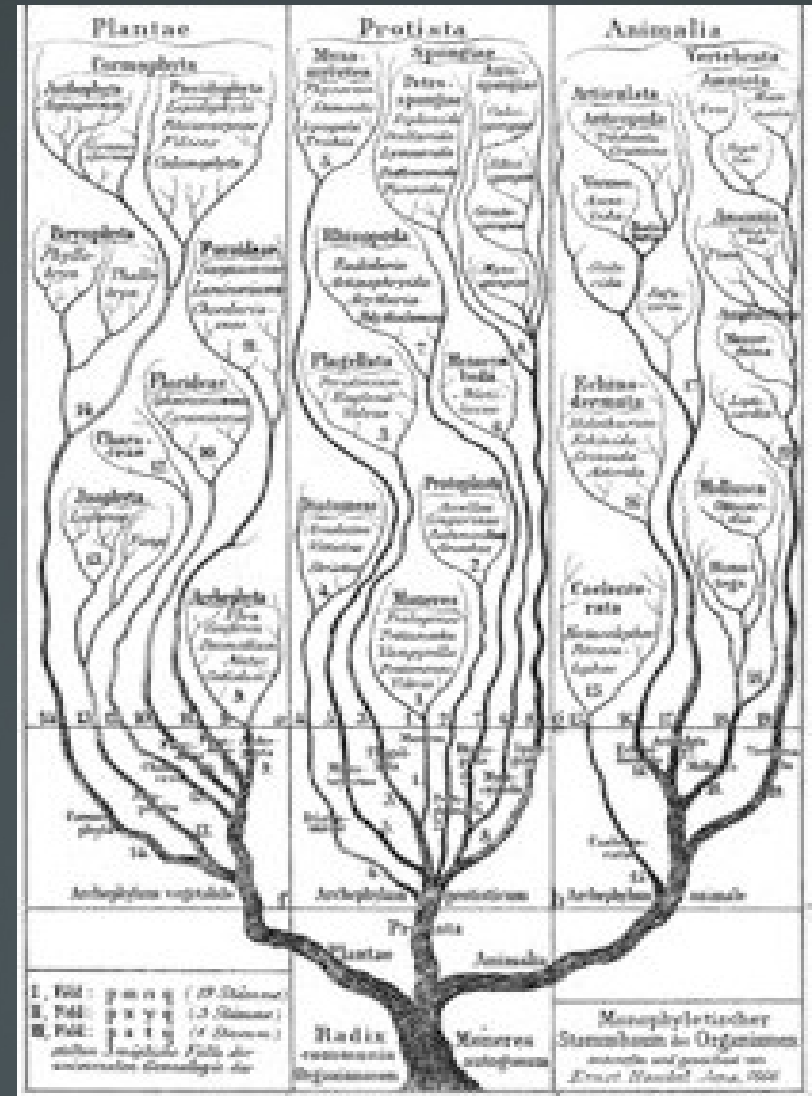
# Last Universal Common Ancestor

## The genetic code is universal for all life

- It tells us that everything is related.
  - All life regenerates itself by producing offspring
  - Over time small changes in due offspring result in small changes to the protein recipes.
  - But since the recipes are written in the same language (genetic code), it is possible to compare these recipes to build the equivalent of a family tree.
- 

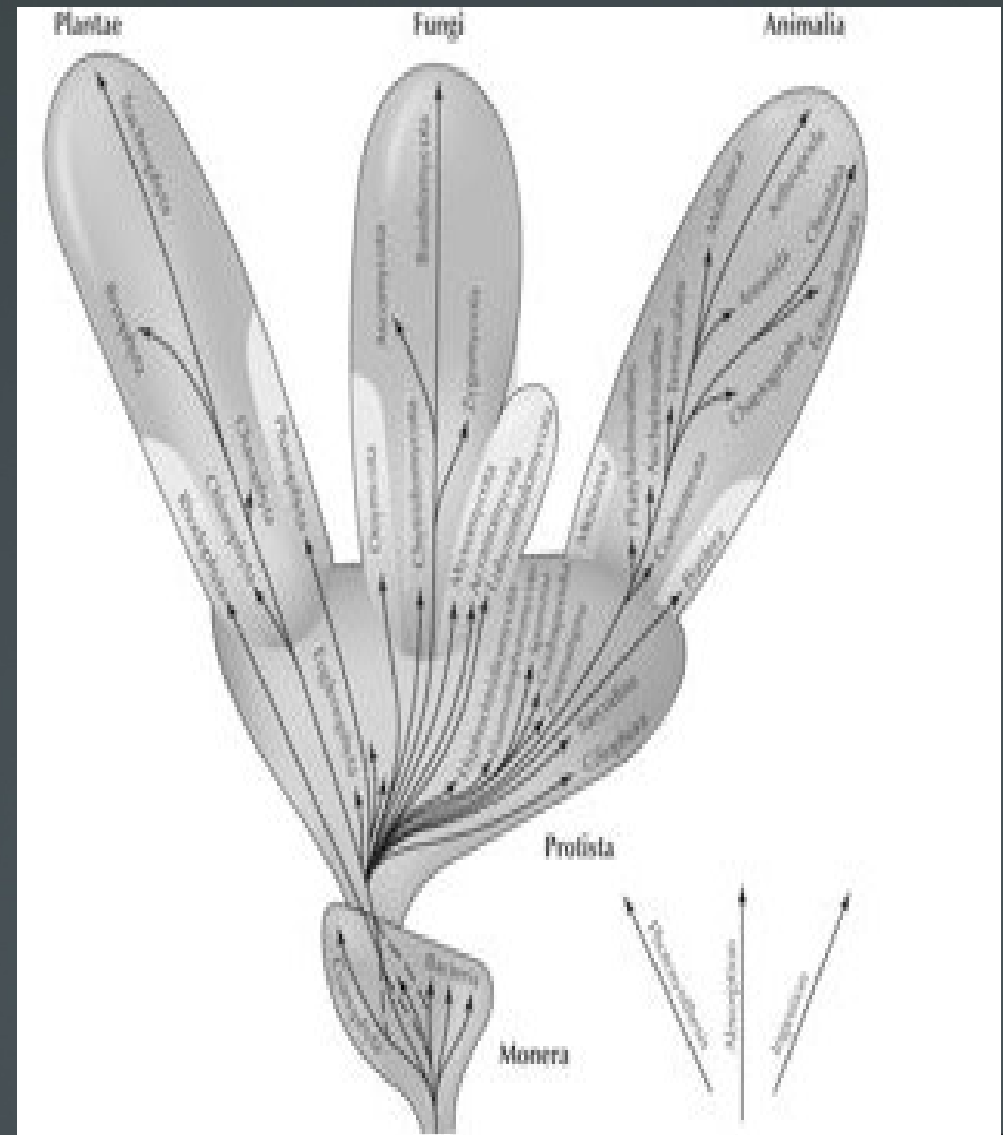
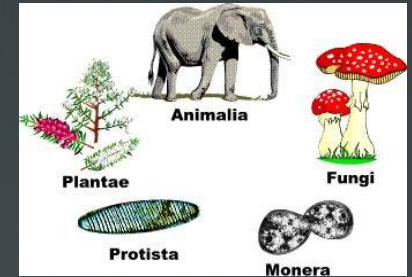
# TREE OF LIFE

- Ernst Haeckel's attempt to draw a monophyletic tree of life (published in 1866)
- Note that the tree is branched into three main lineages: plants, animals and protists.
- Bacteria (as Monera) ancestral to all other organisms



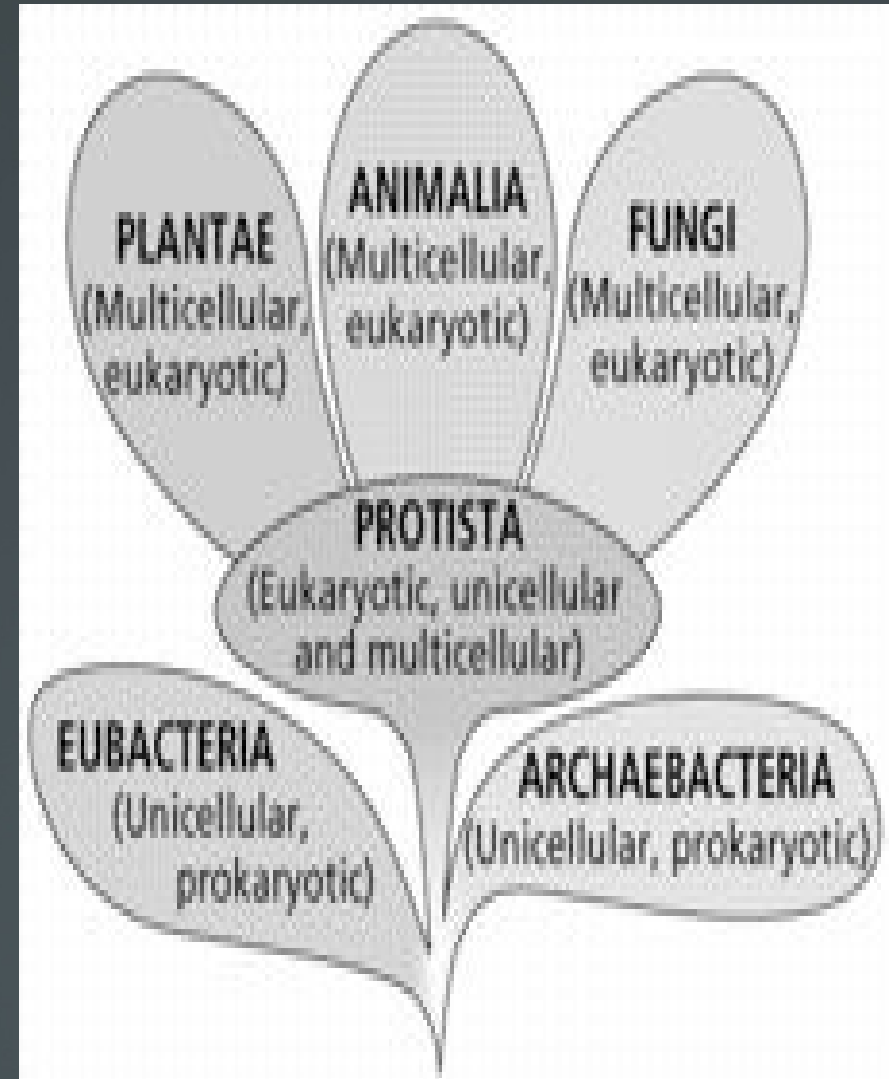
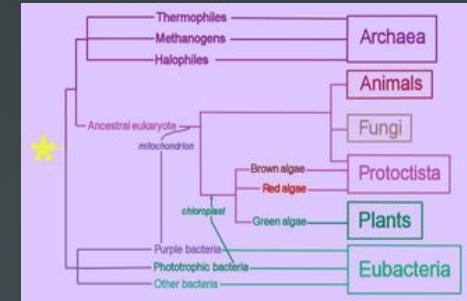
# TREE OF LIFE

- Whittaker's five-kingdom tree, 1969 based on observations:
  - Plantae
  - Animalia
  - Protista
  - Monera
  - Fungi



# TREE OF LIFE

- Six Kingdoms (Carl R Woese, 1977) based on molecular biology studies:
  - Plantae
  - Animalia
  - Protista
  - Fungi
  - Eubacteria
  - Archaeobacteria



# TREE OF LIFE

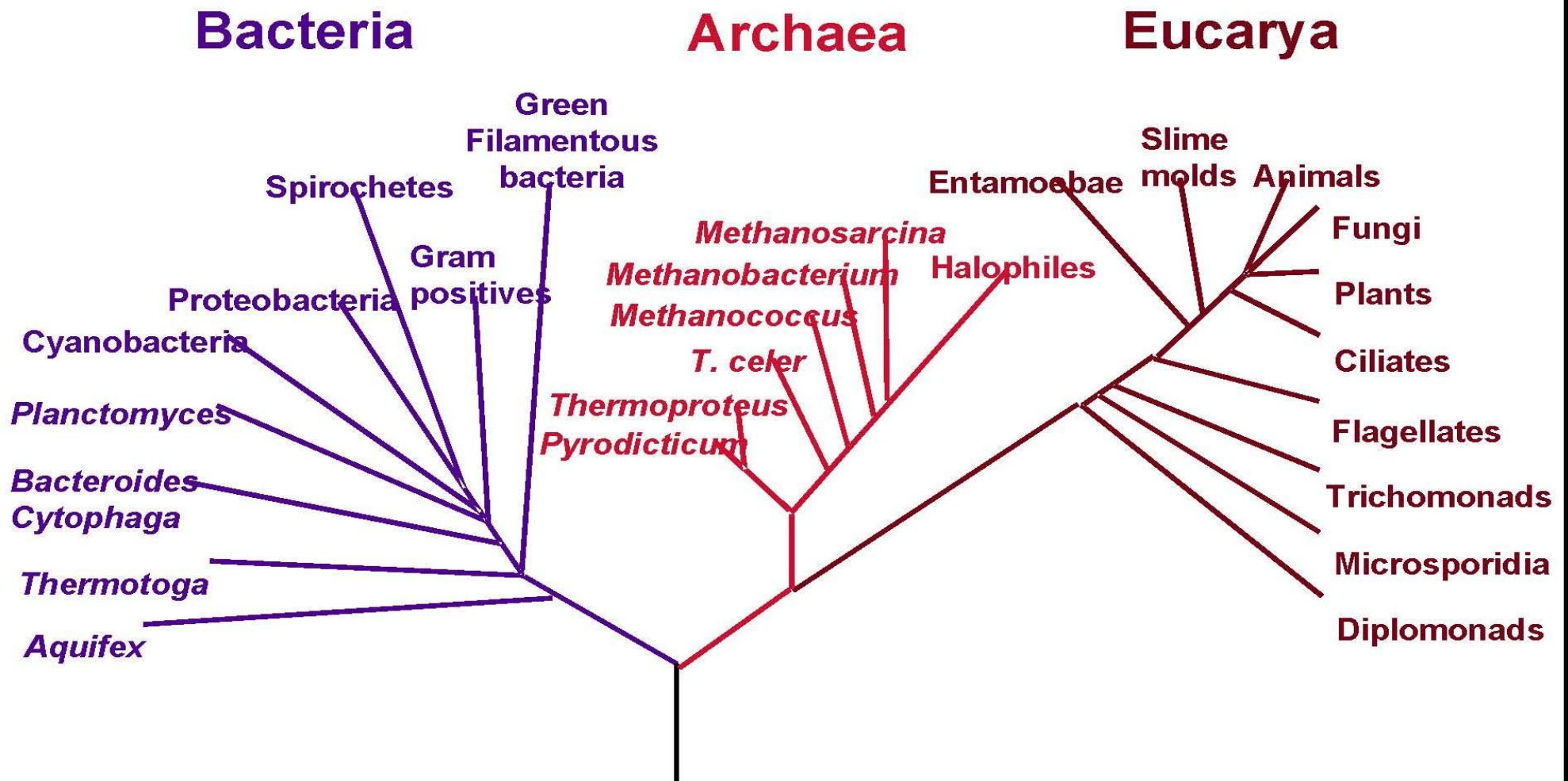
- Before the prototype tree of life, it was thought that life had two major branches:
  - prokarotes
  - eukarotes
- The Tree of life is now known to consist of three domains:
  - Archae
  - Bacteria
  - Eukaryota.



# TREE OF LIFE

Tree of Life is as follows:

## Phylogenetic Tree of Life




# THE PHYLOGENETIC TREE OF LIFE

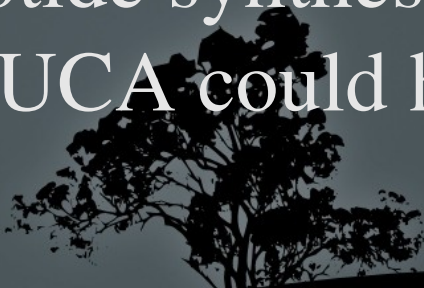
- The tree appears to be rootless
- Three domains diverge from a common point
- Indicates the presence of a LUCA
- This ancestor no longer exists as an identifiable species
- The archaea are unique in that they utilize their DNA in a way that more closely resembles eukaryotic cells than bacterial cells.



# Comparative genomics and modern phylogenetic approaches

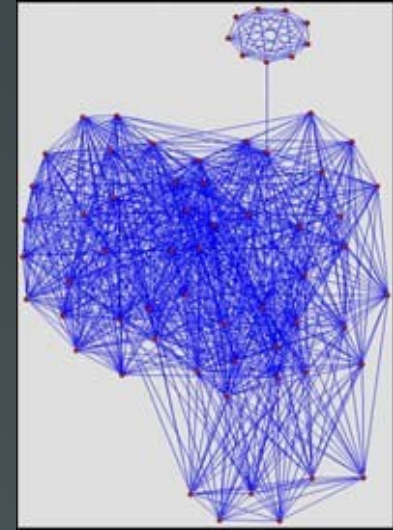
- Allow us to infer the gene content of LUCA of all known currently living cellular organisms
  - Most of the estimates produce:
    - a putative LUCA with 500-1000 protein-coding
    - a biochemically coherent metabolism
  - This estimate is not strongly sensitive to the topology of the Tree of Life
  - Identity of the genes that are placed in LUCA may depend on the position of the deep branches and the root of the tree.
- 

# Genomics

- Allows metabolic traits to be compared through the presence or absence of genes, and by sequence comparisons.
  - Simple comparison of the presence or absence of homologous genes does not take into account the problems of gene loss or acquisition by horizontal gene transfer.
  - It is criticised since it resulted in exclusion of *de novo* pathways for deoxyribonucleotide synthesis, leading to the conclusion that the LUCA could have had an RNA genome.
- 

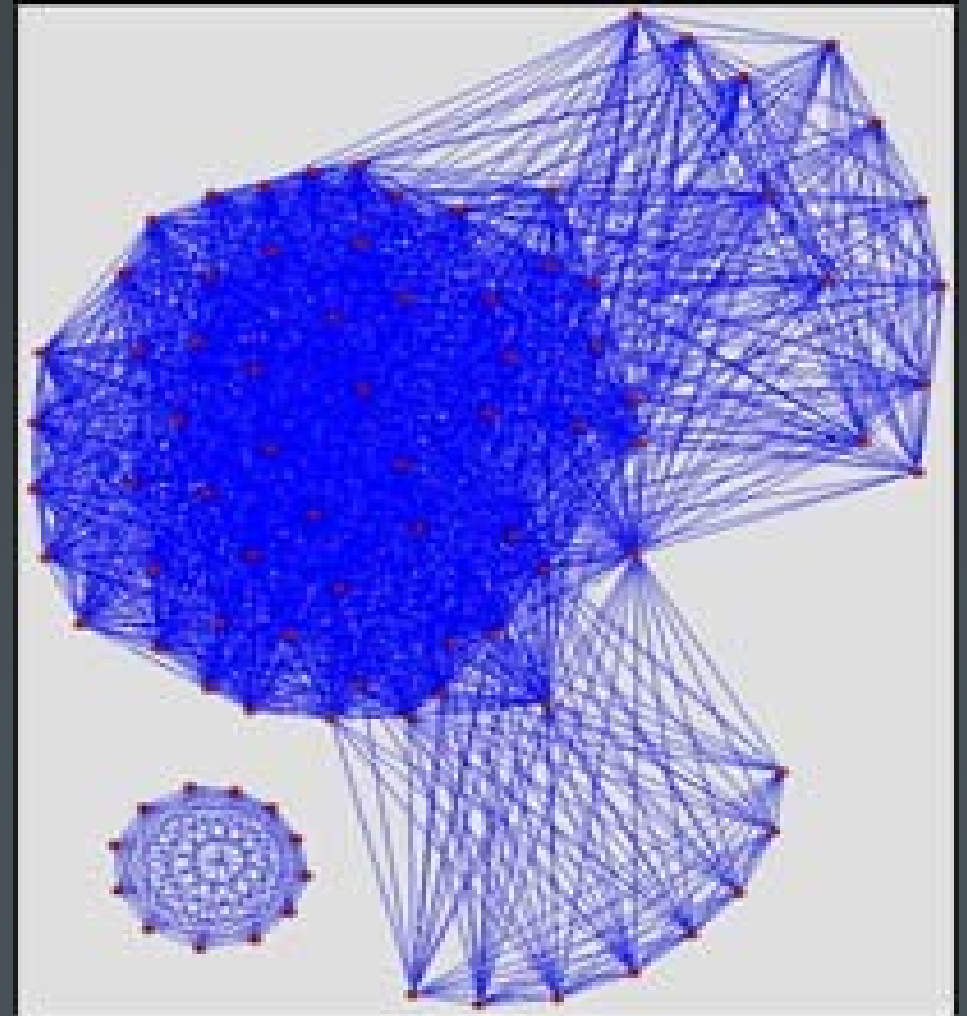
# Clusters of orthologous groups of proteins (COGs)

- A model that invokes gene gains at the root
- It is followed by an occasional gene loss on the way to some present-day species.
- Seems to be a straightforward scenario for the clusters of orthologous groups of proteins or COGs in the vast majority of species.
- multiple losses occur in a large number of lineages which explain that gene losses are more common in evolution than gene gains



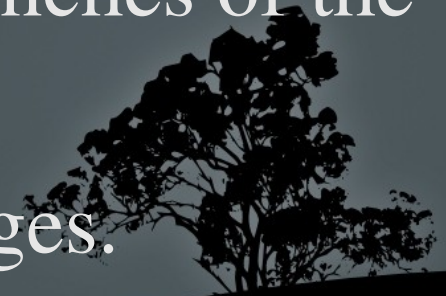
# Clusters of orthologous groups of proteins (COGs)

- COGs experienced more than 90 gene losses per one gene gain
- this scenario also means that:
  - no new genes emerged after LUCA
  - the LUCA genome contained the ancestor of every COG.



# Clusters of orthologous groups of proteins (COGs)

- A considerable fraction of phyletic vectors will have to be explained by some combinations of three factors:
  - The first emergence of this gene at particular node of the tree
  - Transfer of this gene between branches of the tree.
  - Losses of this gene in some lineages.



# The Inference of the LUCA Gene Set

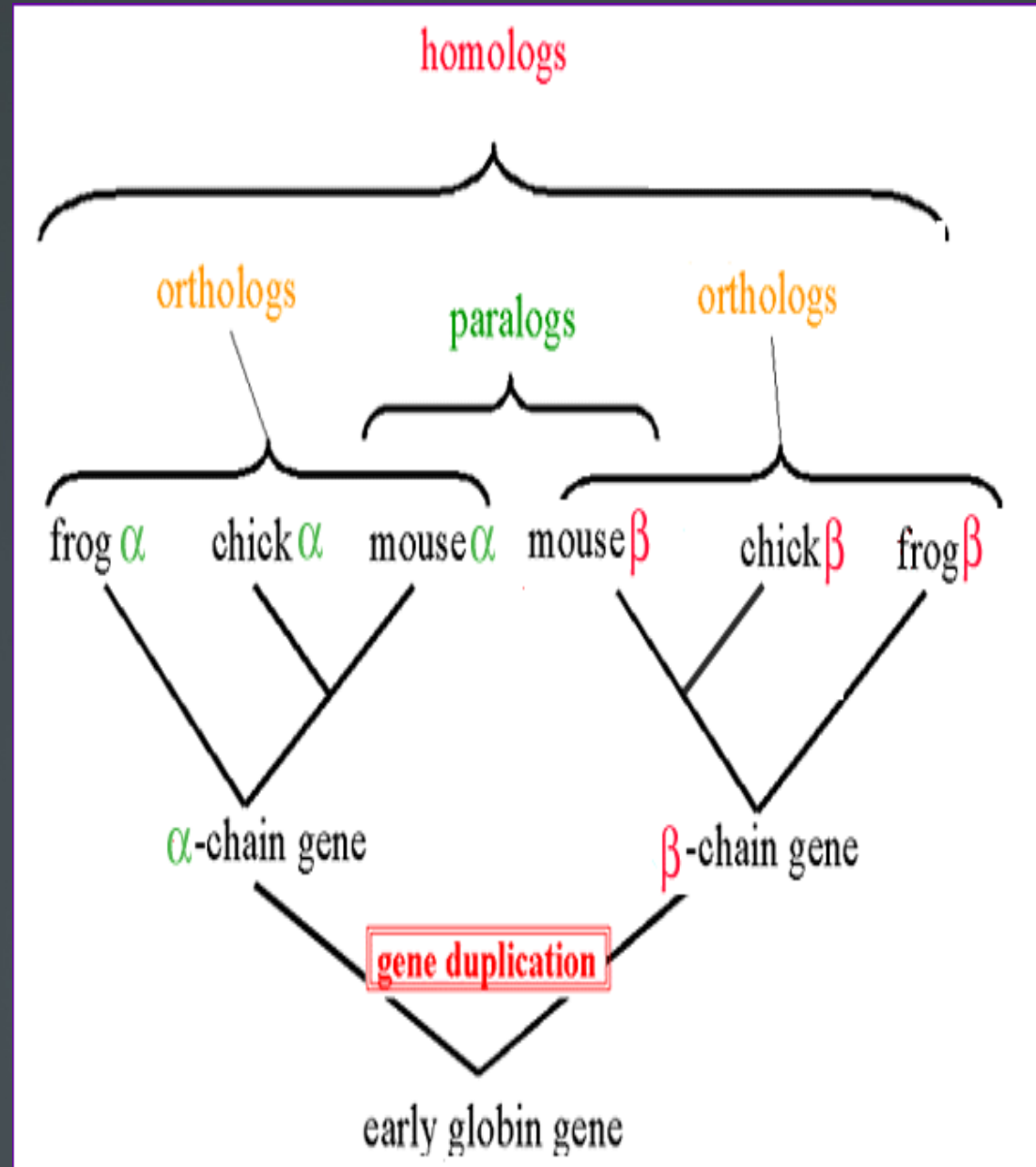
- It requires a rooted species tree.
- The position of the root in the Tree of Life is sometimes also contested.
- It was noted that the gene set of LUCA was sensitive to the changes to the tree topology.
- The origin of the genes can be placed no further back than the root of the cluster of branches.




# Difference between Orthology and Paralogy

## Homologous sequences:

- Orthologs and Paralogs are two types of homologous sequences.
- Orthology describes genes in different species that derive from a common ancestor.
- Orthologous genes may or may not have the same function.
- Paralogy describes homologous genes within a single species that diverged by gene duplication.



# Inference of Ancestral Gene Sets

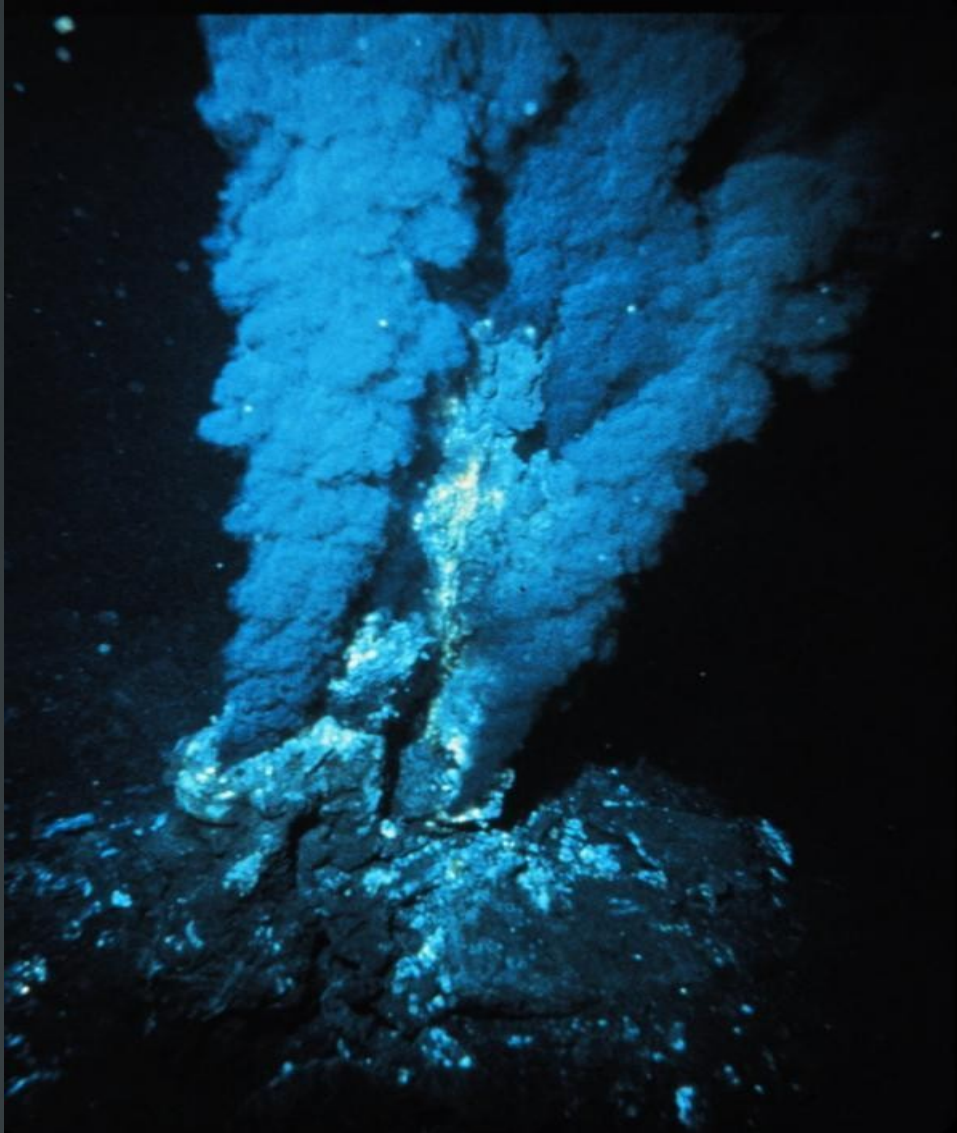
- There are two types of approaches to constructing such sets
    - First Approach
      - One way is to collect all homologs (orthologs and paralogs) using appropriate similarity search programs
      - to delineate all homologous families
      - to infer all homologous families
      - to build a gene tree for each family
      - to infer all duplication and speciation events in each gene tree based on the algorithmically defined comparison between this gene tree and the species tree
      - and to partition homologous families into orthologs and paralogs.
- 

# Inference of Ancestral Gene Sets

- The other type of approach uses the notion of symmetric best matches, sometimes also bidirectional best hits, which are pairs of genes in two genomes, one gene in each, that are one another's top-ranked matches in a database search, such as BLAST or FASTA.
- These pairs can be algorithmically processed to form clusters, representing the sets of most similar genes across genomes.



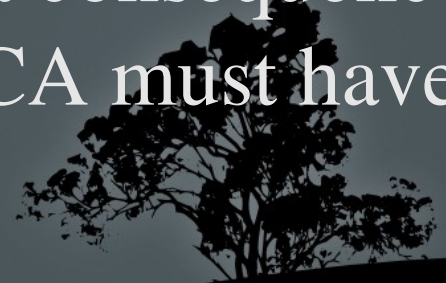
# Hypotheses about LUCA



- Cladograms based on genetic distance between living cells indicated that Archaea split early from the rest of life when LUCA was hypothesized.
- This was inferred from the fact that all known archaeans were highly resistant to environmental extremes such as high salinity, temperature or acidity, and led to suggestions that LUCA evolved in areas like the deep ocean vents.

# Hypotheses about LUCA

- The mesophilic LUCA could have had many features of the eukaryote genome, but its cytology is unknown.
- Since archaeans were discovered in less hostile environments many taxonomists now believe archaeans to be more closely related to eukaryotes than bacteria.
- The fusion hypothesis has important consequences for the LUCA — if correct, the LUCA must have been like bacteria and/or archaea.



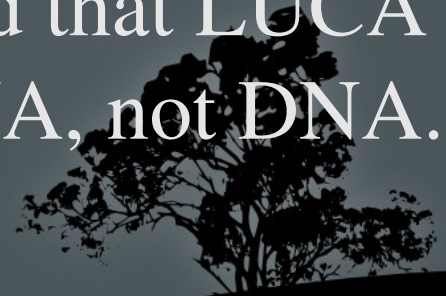
# Hypotheses about LUCA

- It is shown that the HAD (haloacid dehalogenase) superfamily contains 33 major families distributed across the three superkingdoms of life.
- Analysis of the phyletic patterns suggests that at least five distinct HAD proteins are traceable to the last universal common ancestor (LUCA) of all extant organisms.



# The Minimal Genome Project

- It entails the minimum number of genes required to make a cell.
- The most striking features of the minimal genome were:
  - A mere 256 genes
  - No biosynthetic machinery for making the building blocks of DNA.
- From this they tentatively concluded that LUCA stored its genetic information in RNA, not DNA.



# A wider problem with the minimal genome

Problem: genomes you begin with probably affect the final set of genes.




# A wider problem with the minimal genome

- The reasons are:
  - The number of genomes which must be compared before we are confident we aren't omitting something.
  - Not all cells might have been free-living cells such as parasites.
  - Gene losses
  - Genes may have spread so well that they sometimes appear to date back to the time of LUCA, whereas actually , they arose more recently.



# A wider problem with the minimal genome

- Gene gains can also take place and there are three sources of gene gains:
    - duplication of an existing gene followed by divergence.
    - *de nova* emergence of new (open reading frames) ORF from a new coding sequence or by recoding.
    - gain of a gene from another organism by horizontal gene transfer.
- 

# Biological roles of ncRNA as fossils

- Noncoding RNAs are a diverse group of genes involved in many cellular processes.
- The more conserved ncRNAs are thought to be molecular fossils or relics from LUCA and the RNA world.
- Molecular fossils, or 'relics' can be broadly defined as those parts of modern metabolism which have persisted from an earlier stage in evolution.



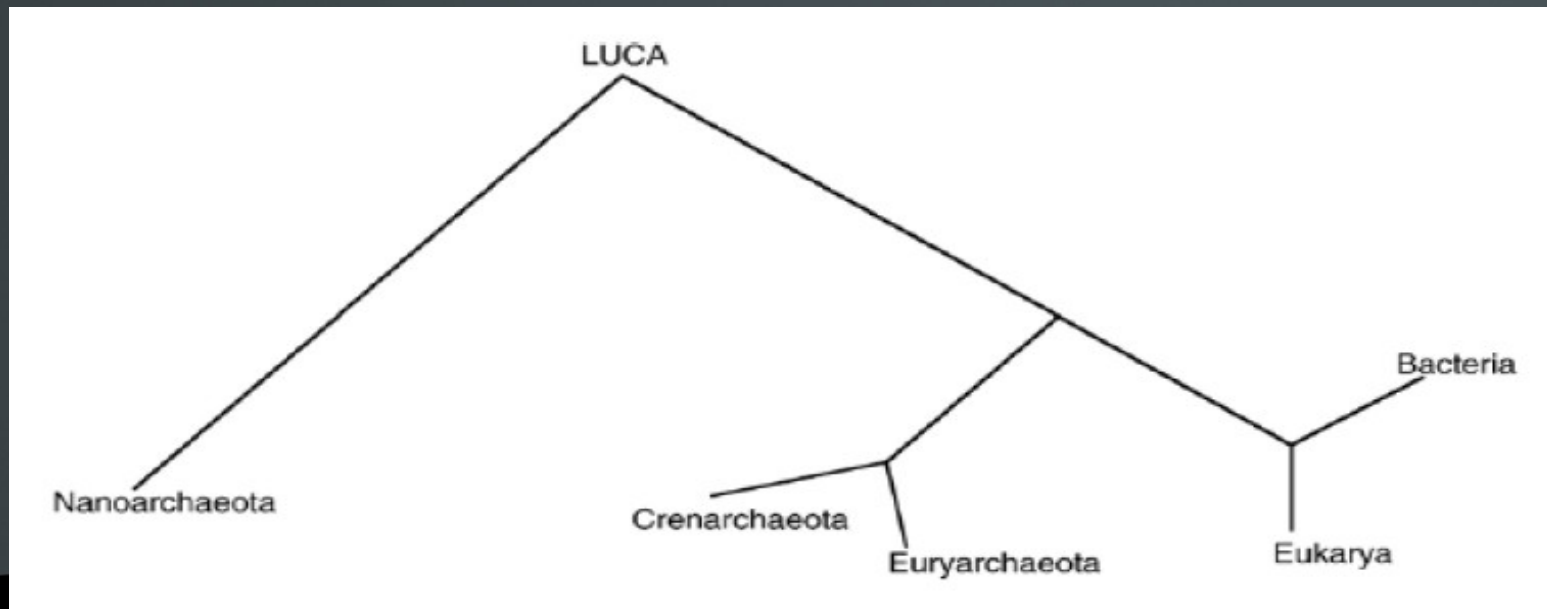
# Methods to reconstruct Roots

- A simple way of rooting the universal phylogenetic tree is based on the uniqueness and ancestrality of a trait.
- Uniqueness means that the trait in the phylogenetic tree being examined has been observed only once.
- Ancestrality means that the trait can certainly be attributed, to the universal ancestor of the phylogenetic tree being examined.
- This hypothesis seems to be confirmed by the those two properties of some traits possessed by *Nanoarchaeum equitans*



# Methods to reconstruct Roots

- characteristics of uniqueness and ancestrality favours the hypothesis of a trifurcation of the universal phylogenetic tree without there being an actual root.
- the genome of *N. equitans* might specifically reflect the rooting of the Tree of Life.




# Implications of the ability to reconstruct phylogenetic trees

- The further back in time an evolutionary divergence is studied, the greater the likelihood that any given gene in a genome has been transferred.
- It may be the case that all bacterial genes have been subject to horizontal gene transfer at some point in their evolutionary history.
  - This could undermine the utility of phylogenetic tree reconstruction for deep divergences



# Gene transfer contradicts the reconstruction of the LUCA


- The problem is a complex one:
    - Horizontal gene transfer has been demonstrated.
    - Limitations of the methods for building evolutionary trees can give false evidence for gene transfer.
    - Evolutionary information is not considered when examining genetic relationships and the data that have been used to argue for horizontal gene transfer are weak.
    - There is little consensus on the reliability of methods for detecting horizontal gene transfer.
    - The type of data required to demonstrate ancient HGT events.
- 

# Identification of Ancient Gene Transfer


- Biologists can readily identify genes in the eukaryote repertoire that have come in via the mitochondrion, a compartment in the eukaryote cell which is bacterial in origin.



# The Recent History of our Past

- There is now overwhelming evidence that we are part bacterium.
  - Evidence indicates gene swapping in human DNA.
    - Our bacterial ancestry comes in the form of mitochondria, tiny power plants housed in our cells.
    - Mitochondria were once full-blown bacteria that took up residence in and struck up a partnership with one of our distant single-celled ancestors.
    - Since then, much of the DNA from the original bacterium has been lost, but a much of it has ended up in the DNA of our nucleus.
- 

# The Recent History of our Past

- Mitochondrial Eve: DNA studies show human mitochondria can trace their lineage to a “mitochondrial Eve”, the matrilineal most recent common ancestor for all humans alive today, who lived approximately 150,000 years ago.
  - Y-chromosomal Adam: DNA studies show that all Y chromosomes in currently living men are descended from a “Y-chromosomal Adam”, the patrilineal most recent common ancestor for all humans alive today, who lived approximately 60,000 years ago.
- 

# Recommendations for further research

- Genome research suggests that not all genes transfer equally easily.
- The question is whether early evolution was more reliant on horizontal gene transfer than inheritance.
- One underestimates how many genes were originally in LUCA.
- If an RNA is older than LUCA, then LUCA had it too, even if that RNA is no longer universal.



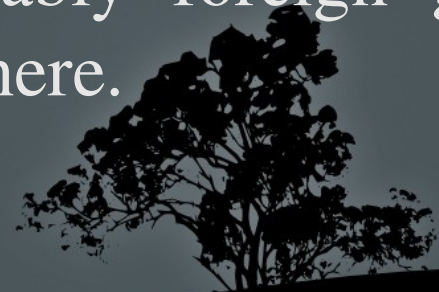
# Recommendations for further research

- Carl Woese, one of the key contributors in the bid to reconstruct the tree of life, inspired researchers by suggesting that:
  - there may be more than one LUCA.
  - LUCA was also into gene swapping, and on a much larger scale than what we observe in modern bacteria.
  - gene swapping was once more important than inheritance from parent to offspring, and that early archaea, bacteria and eukaryotes each emerged independently from a 'sea' of gene transfer.



# Recommendations for further research

- Two bacteria from the same species may reveal major differences.
  - Escherichia coli is a common gut bacterium that is part of our natural gut 'flora.'
  - the O157:H7 strain causes severe gastrointestinal ailments.
    - 1387 of the 5416 (26%) genes in O157:H7 are not in K-12.
    - 528 of K-12's 4405 (12%) genes are not in O157:H7.
  - Many of the O157:H7 genes are arguably 'foreign' genes that have been borrowed from elsewhere.



# Concluding Remarks

- It is unclear whether the LUCA was a single 'species' or whether there was extensive horizontal transfer between divergent life forms.
- The concentration of RNA relics within the nucleus suggests this organelle is more ancient than previously supposed.
- The jury is still not out as to how to reconstruct LUCA, and whether horizontal gene transfer will turn this task into a futile one.



# Mind boggling questions

- Were there three domains or two, with the third arising by fusion?
- Was LUCA prokaryote-like or eukaryote-like or even a mixture?
- Is the genetic code the only one possible?
- Was early evolution more reliant on horizontal gene transfer than inheritance?
- Are genes equally swappable?
- Is gene swapping as common across other branches of the tree?
- Was there one or more LUCAs?



But is it enough to save LUCA?




# References

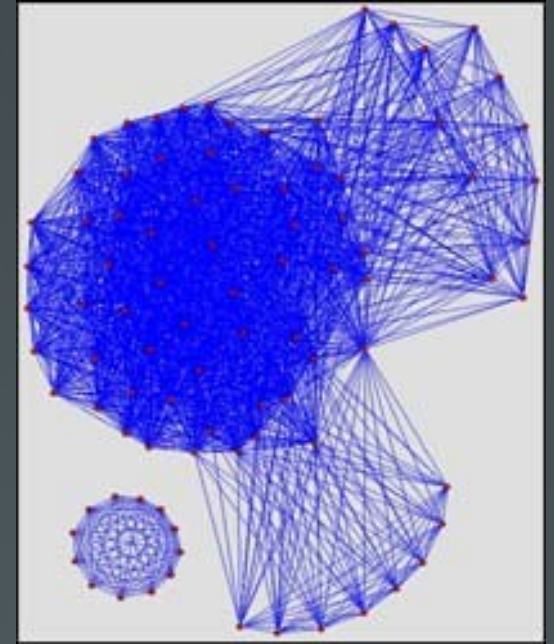
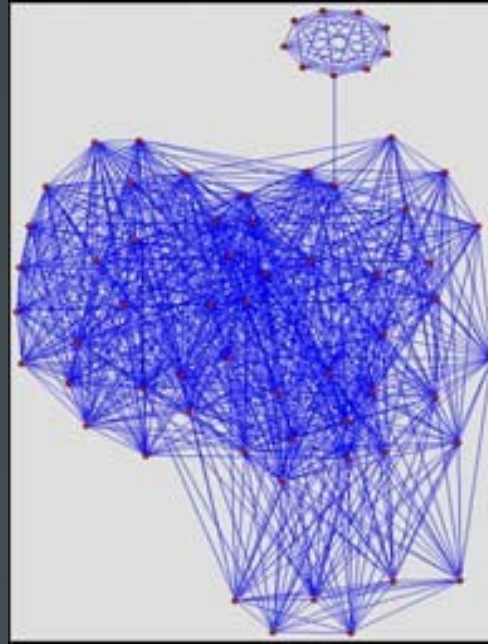
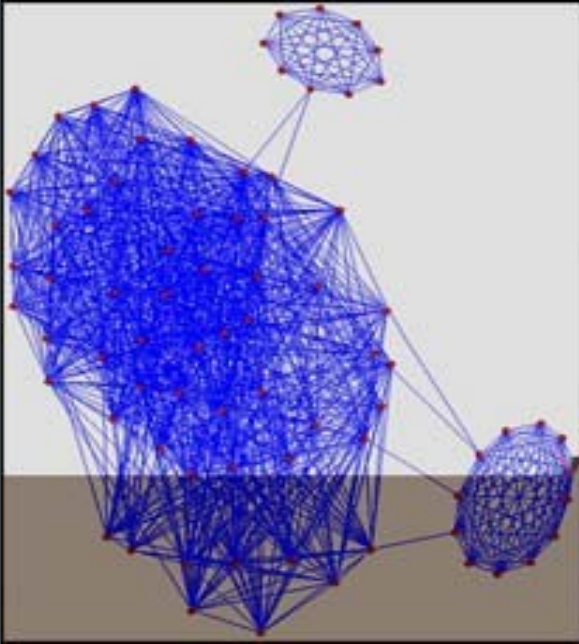
- [1]. Anthony M. Poole, An ActionBioscience.org original article(2002).
- [2]. Vivek Gowri-Shankar and Magnus Rattray, A Reversible Jump Method For Bayesian Phylogenetic Inference with a Nonhomogeneous Substitution model(2007).
- [3]. David Penny and Anthony Poole. The nature of the last universal common ancestor(2003).
- [4]. Jonh Whitfield , Born in a watery commune (2002).
- [5]. Mat WK, Xue H, Wong JT. (2008) The genomics of LUCA. Front Biosci. 13:5605-13.
- [6].Mushegian A. (2008) Gene content of LUCA, the last universal common ancestor. Front Biosci.13:4657-66.
- [7]. Ranea JA, Sillero A, Thornton JM, Orengo CA. (2006) Protein superfamily evolution and the last universal common ancestor (LUCA). J Mol Evol. 63(4):513-25. Epub 2006 Oct4.
- [8]. Koonin EV. (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat Rev Microbiol. 1(2):127-36.
- [9]. Penny D, Poole A. (1999) The nature of the last universal common ancestor. Curr Opin Genet Dev.;9(6):672-7.
- [10]. Hoenigsberg H. (2003) Evolution without speciation but with selection: LUCA, the Last Universal Common Ancestor in Gilbert's RNA world. Genet Mol Res. 2(4):366-75



# COG's

- Graph representing COG0616 with 89 proteins from 51 species. Two nodes (proteins)
  - are connected by an edge if and only if their distance vectors to other nodes (including themselves)
  - are positively correlated. Clustering shows that the 12-node connected component does not belong
  - to the COG, indicating that these 12 proteins are out-paralogs to the proteins in the main cluster,
  - leaving the COG with 77 proteins from 47 species.
- 

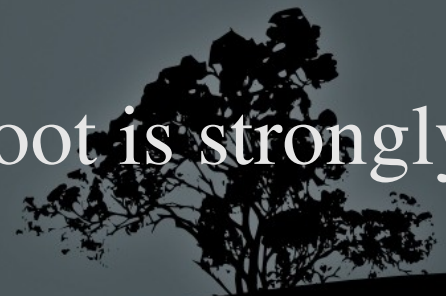
# COG's



# Methods to reconstruct Roots

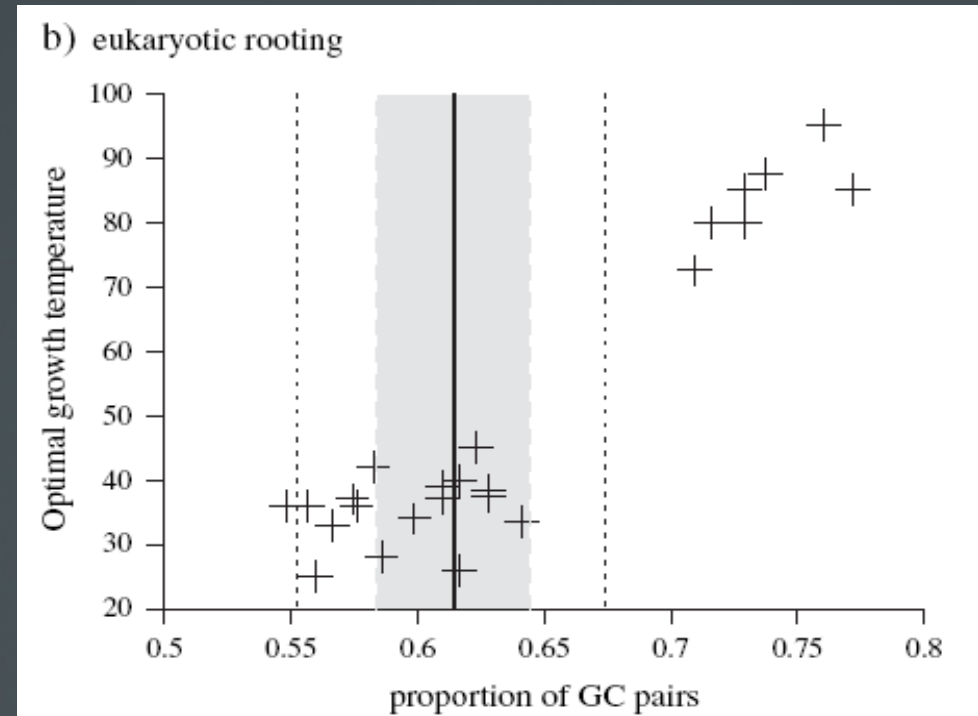
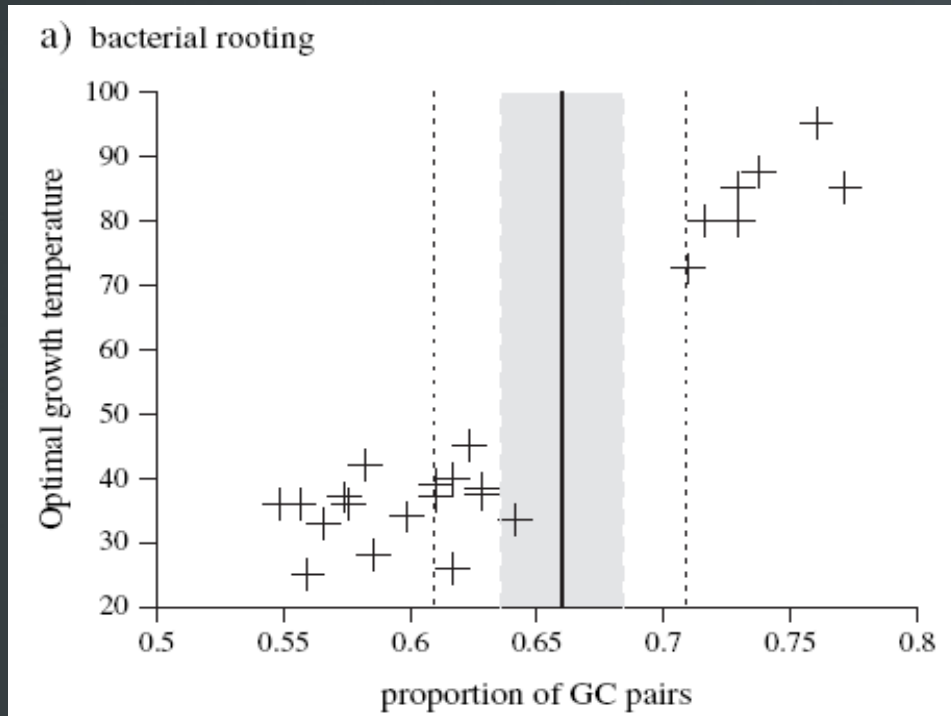
## GC Content and Thermophily:

- A previous study of the same genes from a similar set of species found robust evidence for a mesophilic LUCA by inference of the CG composition at the root of the tree.
- It is important to consider the CG pairs since they are most tightly bound than the alternative, AU pairs, and are therefore more stable at high temperatures.
- The helical GC composition at the root is strongly dependent on the root position.




# Methods to reconstruct Roots

- Because the monophyly of Archaea is uncertain and not supported by a data set used by [Vivek Gowri-Shankar and Magnus Rattray, 2007].
- only consider the bacterial and eukaryotic rootings.

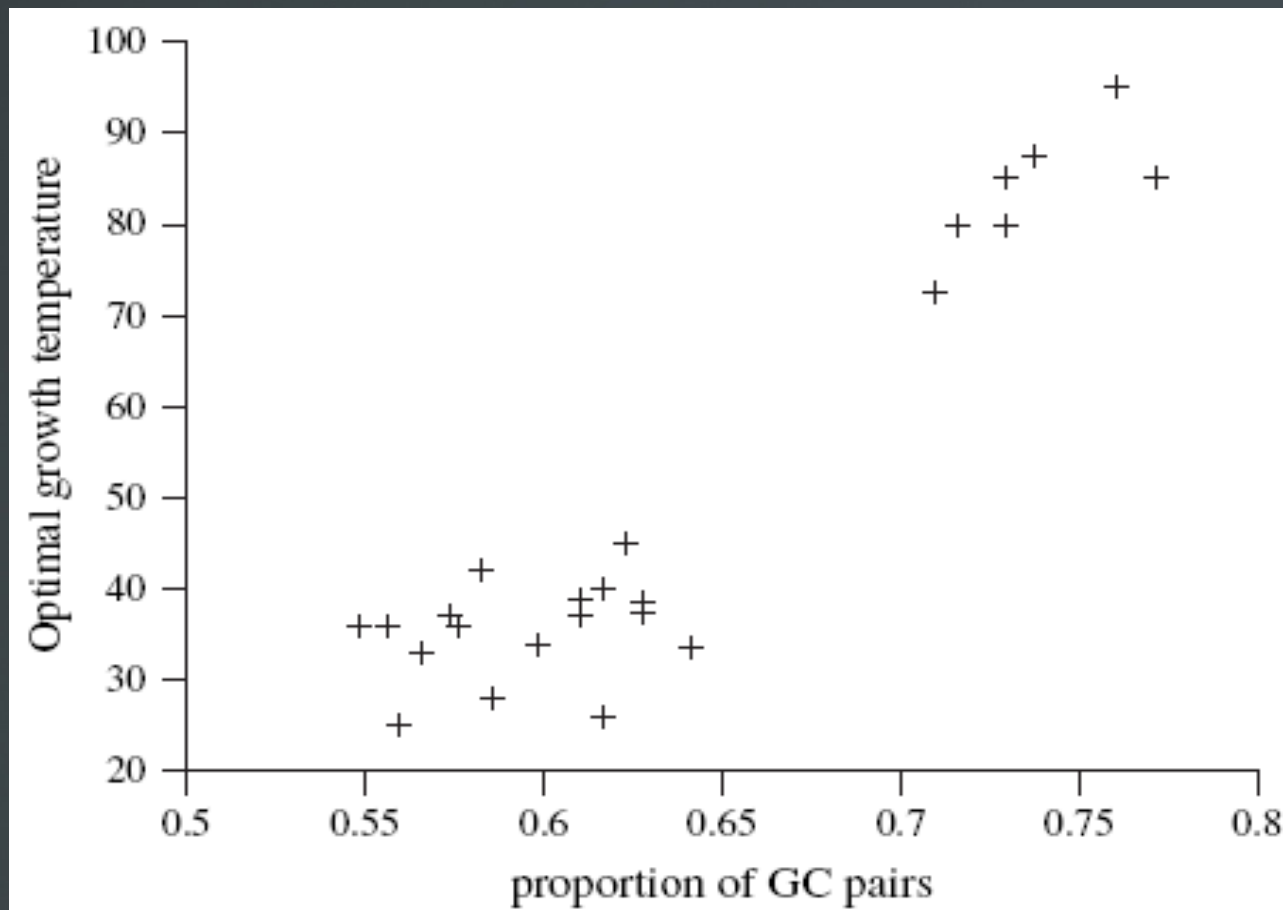


# Methods to reconstruct Roots

- The GC content of nuclear rRNA sequences is quite variable among prokaryotes.
  - The GC content of nuclear RNA genes is correlated with the optimal growth temperature (OGT) in prokaryotes.
  - The regression results were obtained using standard linear regression and therefore ignore the contradicting influence of the phylogenetic signal.
  - This model conflicted the fact that LUCA is hyperthermophilic.
- 

# Methods to reconstruct Roots

## GC Content and Thermophily:



[Vivek Gowri-Shankar and Magnus Rattray, 2007]

# Methods to reconstruct Roots

- Approach developed in the paper of [Vivek Gowri-Shankar and Magnus Rattray, 2007]
  - Firstly, the number of composition vectors is a parameter of the model which is allowed to vary by using Markov chain Monte Carlo (MCMC) methods for Bayesian parameter estimation. MCMC handles parameters.
  - Secondly, rooted trees are considered, and the root position is not constrained to internal trifurcating nodes. MCMC handles parameters.
  - Thirdly, it is shown that using a uniform prior for the allocation of composition vectors on the branches has some unexpected, and probably unwanted, side effect.