LSM2241
Introduction to Molecular Evolution and Phylogenetic Inference

Greg Tucker-Kellogg
dbsgtk@nus.edu.sg

30 September and 7 October 2015

Outline

Background

Molecular Evolution and the tree of life

Inferring evolutionary history

Phylogenies and phylogenetics trees

Nucleotide substitution models

The process of inferring phylogeny from sequence data

Distance-based methods

Character-based methods

Lateral (horizonal) gene transfer

Roundup and next time

Topic

Background

Molecular Evolution and the tree of life

Inferring evolutionary history

Phylogenies and phylogenetics trees

Nucleotide substitution models

The process of inferring phylogeny from sequence data

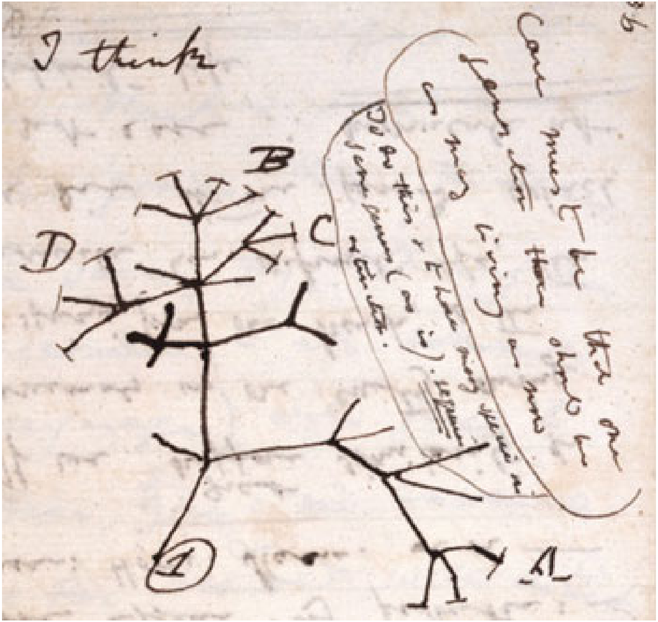
Distance-based methods

Character-based methods

Lateral (horizonal) gene transfer

Roundup and next time

Darwin’s notebooks (1837)


I think
B
C
D
1
A
Case must be that one generation
then should be as many
living as now
To do this & to have many
species in same genus (as is)
requires extinction.

Darwin's Universal Tree of Life

"The affinities of all the beings of the same class have sometimes been represented by a great tree. I believe this simile largely speaks the truth. The green and budding twigs may represent existing species; and those produced during each former year may represent the long succession of extinct species."

CHARLES DARWIN **ORIGIN OF SPECIES, 6TH EDITION** 1872

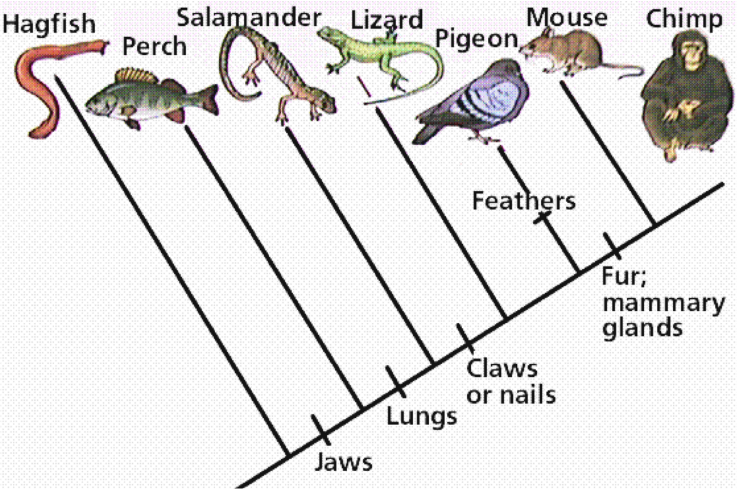
The tree of life

- The idea of the tree of life predates Darwin by several millennia
- Evolution by means of natural selection (Darwin and Wallace) provided a scientific basis for understand the history of life by “Descent with Modification”
- The **Modern Synthesis** in the 1930s brought together natural selection and genetics with mathematical precision

however…

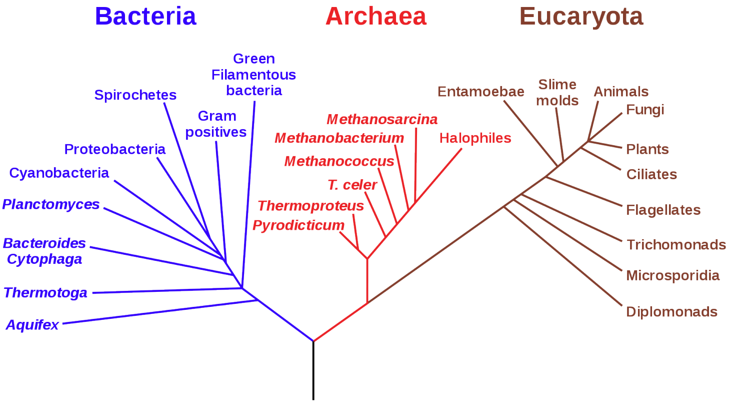
Only in the 1960s was *molecular* information used to determine phylogenies

Phylogenetic trees are built from observable characters



A tree showing anatomical features of animals

A landmark tree based entirely on ribosomal RNA sequence conservation



Carl Woese’s Phylogenetic tree, built in 1977, following Woese and Fox 1977

What's so special about the tree by Woese?

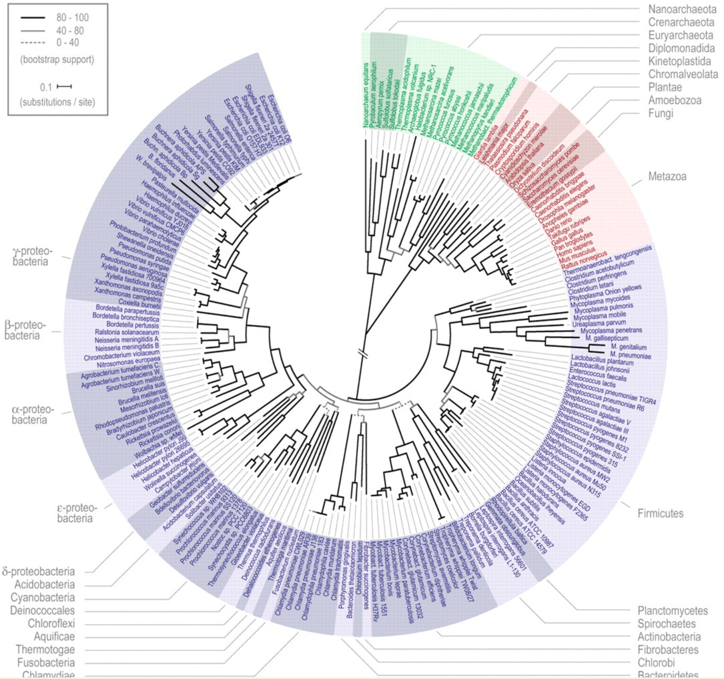
- Woese's work assigned *three* domains of life, rather than just prokaryotes and eukaryotes
- Two domains of prokaryotes
  - One is *Bacteria*
  - One is *Archaea*
- The archaea in this tree are more like eukaryotes than like bacteria
- Last Universal Commmon Ancestor (LUCA)

Woese's tree

Built from phylogenetic analysis of 16S rRNA

What does the tree look like today?

- Well, what did it look like in 2006? (Ciccarelli et al. 2006)
- Trees are now built based on comparison of whole genomes!


80 - 100
40 - 80
0 - 40
(bootstrap support)
0.1
(substitutions / site)
Nanoarchaeota
Crenarchaeota
Euryarchaeota
Diplomonadida
Kinetoplastida
Chromalveolata
Plantae
Amoebozoa
Fungi
Metazoa
Firmicutes
Planctomycetes
Spirochaetes
Actinobacteria
Fibrobacteres
Chlorobi
Bacteroidetes
Chlamydiae
Fusobacteria
Thermotogae
Aquificae
Chloroflexi
Deinococcales
Cyanobacteria
Acidobacteria
δ-proteobacteria
ε-proteobacteria
α-proteobacteria
β-proteobacteria
γ-proteobacteria
Nanoarchaeum equitans
Pyrobaculum aerophilum
Aeropyrum pernix
Sulfolobus solfataricus
Sulfolobus tokodaii
Thermoplasma acidophilum
Thermoplasma volcanium
Archaeoglobus fulgidus
Halobacterium sp. NRC-1
Methanosarcina mazei
Methanosarcina acetivorans
Pyrococcus furiosus
Pyrococcus abyssi
Pyrococcus horikoshii
Methanococcus jannaschii
Methanococcus maripaludis
Methanopyrus kandleri
Methanobact. thermautotrophicum
Giardia lamblia
Leishmania major
Thalassiosira pseudonana
Plasmodium falciparum
Cryptosporidium hominis
Cyanidioschyzon merolae
Arabidopsis thaliana
Oryza sativa
Dictyostelium discoideum
Schizosaccharomyces pombe
Saccharomyces cerevisiae
Eremothecium gossypii
Caenorhabditis briggsae
Caenorhabditis elegans
Drosophila melanogaster
Anopheles gambiae
Danio rerio
Takifugu rubripes
Gallus gallus
Pan troglodytes
Homo sapiens
Mus musculus
Rattus norvegicus
Thermoanaerobact. tengcongensis
Clostridium acetobutylicum
Clostridium perfringens
Clostridium tetani
Phytoplasma Onion yellows
Mycoplasma mycoides
Mycoplasma pulmonis
Mycoplasma mobile
Ureaplasma parvum
Mycoplasma penetrans
M. gallisepticum
M. genitalium
M. pneumoniae
Lactobacillus plantarum
Lactobacillus johnsonii
Enterococcus faecalis
Lactococcus lactis
Streptococcus pneumoniae TIGR4
Streptococcus pneumoniae R6
Streptococcus mutans
Streptococcus agalactiae V
Streptococcus agalactiae III
Streptococcus pyogenes M1
Streptococcus pyogenes 8232
Streptococcus pyogenes SSI-1
Streptococcus pyogenes 315
Staphylococcus epidermidis
Staphylococcus aureus MW2
Staphylococcus aureus Mu50
Staphylococcus aureus N315
Listeria innocua
Listeria monocytogenes EGD
Listeria monocytogenes F2365
Oceanobacillus iheyensis
Bacillus halodurans
Bacillus subtilis
Bacillus anthracis
Bacillus cereus ATCC 14579
Bacillus cereus ATCC 10987
Rhodopirellula baltica
Gemmata obscuriglobus
Leptospira interrogans 56601
Leptospira interrogans L1-130
Borrelia burgdorferi
Treponema denticola
Treponema pallidum
Bifidobacterium longum
Tropheryma whipplei TW08/27
Tropheryma whipplei Twist
Streptomyces coelicolor
Streptomyces avermitilis
Corynebacterium diphtheriae
Corynebacterium efficiens
Corynebact. glutamicum 13032
Corynebacterium glutamicum
Mycobacterium paratuberculosis
Mycobacterium leprae
Mycobacterium bovis
Mycobact. tuberculosis 1551
Mycobact. tuberculosis H37Rv
Fibrobacter succinogenes
Chlorobium tepidum
Bacteroides thetaiotaomicron
Porphyromonas gingivalis
Chlamydia trachomatis
Chlamydia muridarum
Chlamydophila caviae
Chlamydophila pneumoniae 183
Chlamydia pneumoniae J138
Chlamydia pneumoniae AR39
Chlamydia pneumoniae CWL029
Fusobacterium nucleatum
Thermotoga maritima
Aquifex aeolicus
Dehalococcoides ethenogenes
Deinococcus radiodurans
Thermus thermophilus
Gloeobacter violaceus
Thermosynechococcus elongatus
Synechocystis sp. PCC6803
Nostoc sp. PCC 7120
Prochlorococcus marinus 1378
Prochlorococcus marinus SS120
Prochlorococcus marinus 9313
Synechococcus sp. WH8102
Solibacter usitatus
Acidobacterium capsulatum
Desulfovibrio vulgaris
Bdellovibrio bacteriovorus
Geobacter sulfurreducens
Campylobacter jejuni
Wolinella succinogenes
Helicobacter hepaticus
Helicobacter pylori 26695
Helicobacter pylori J99
Wolbachia sp. wMel
Rickettsia conorii
Rickettsia prowazekii
Caulobacter crescentus
Bradyrhizobium japonicum
Rhodopseudomonas palustris
Mesorhizobium loti
Brucella melitensis
Brucella suis
Sinorhizobium meliloti
Agrobacterium tumefaciens W.
Agrobacterium tumefaciens C.
Nitrosomonas europaea
Chromobacterium violaceum
Neisseria meningitidis B
Neisseria meningitidis A
Ralstonia solanacearum
Bordetella pertussis
Bordetella bronchiseptica
Bordetella parapertussis
Coxiella burnetii
Xanthomonas campestris
Xanthomonas axonopodis
Xylella fastidiosa 9a5c
Xylella fastidiosa 700964
Pseudomonas aeruginosa
Pseudomonas syringae
Pseudomonas putida
Shewanella oneidensis
Photobacterium profundum
Vibrio cholerae
Vibrio parahaemolyticus
Vibrio vulnificus CMCP6
Vibrio vulnificus YJ016
Haemophilus ducreyi
Haemophilus influenzae
Pasteurella multocida
W. brevipalpis
B. floridanus
Buchn. aphidicola Bp
Buchnera aphidicola APS
Buchnera aphidicola Sg
Photorhabdus luminescens
Yersinia pestis Medievalis
Yersinia pestis KIM
Yersinia pestis CO92
Salmonella typhimurium
Salmonella enterica
Salmonella typhi
Escherichia coli O157:H7
Escherichia coli EDL933
Shigella flexneri 2a 301
Shigella flexneri 2a 2457T
Escherichia coli K12
Escherichia coli O6

Topic

Background

Molecular Evolution and the tree of life

Inferring evolutionary history

Phylogenies and phylogenetics trees

Nucleotide substitution models

The process of inferring phylogeny from sequence data

Distance-based methods

Character-based methods

Lateral (horizonal) gene transfer

Roundup and next time

What is Molecular Evolution?

- The processes of evolution as it occurs among DNA, RNA, and proteins
- The science of studying processes in evolution at a molecular level
- Inferring evolutionary history from molecular data

What does the “molecular” in molecular evolution offer?

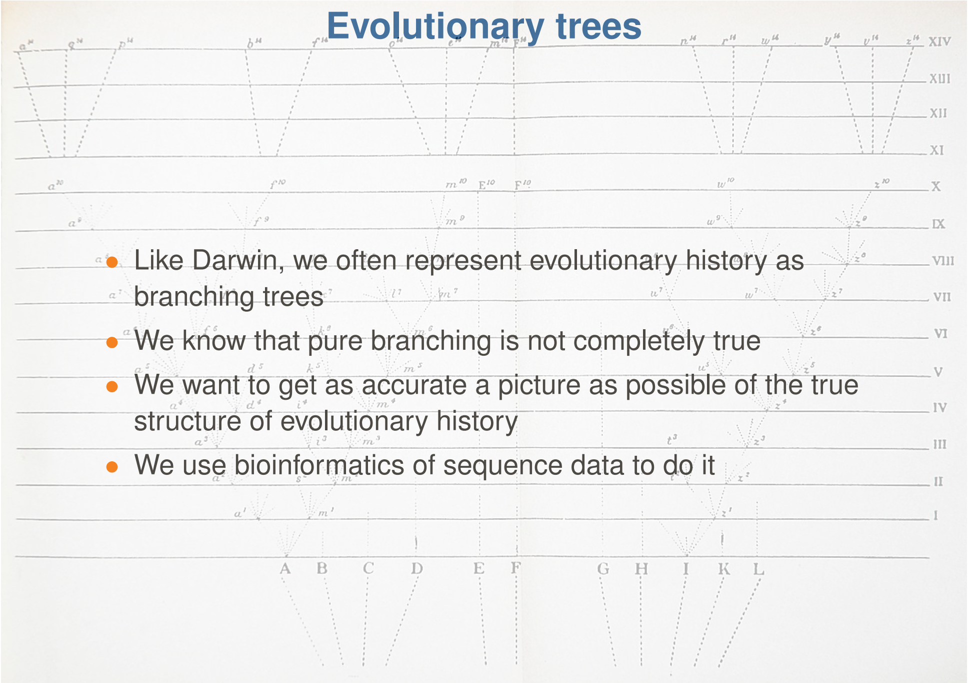
- New insights into early evolution
- Evolution at the level of genes
  - Gene duplication
  - New insights into emergence of gene function
- Understanding of horizontal gene transfer
- Many others!

Evolution at the level of DNA sequence

1. Random mutations are introduced into genetic material, and passed to the next generation
2. favorable mutations accumulate, and deleterious mutations are eliminated, by selection

How can we tell these processes apart, and how can we identify the branching patterns in evolutionary trees?

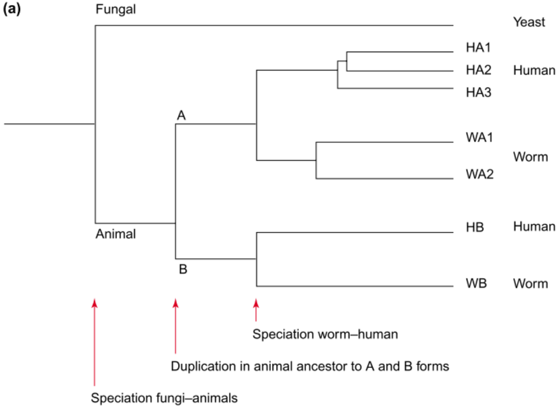
Evolutionary trees



- Like Darwin, we often represent evolutionary history as branching trees
- We know that pure branching is not completely true
- We want to get as accurate a picture as possible of the true structure of evolutionary history
- We use bioinformatics of sequence data to do it

What are the precise relationships between genes on this history?

The scenario posed by Sonnhammer and Koonin 2002



Categories of orthologs and paralogs

In-paralogs Paralogs in which a comparison is describing genes related by duplication more recently than a speciation event

Out-paralogs When a speciation under consideration happened more recently than the duplication event giving rise to paralogs

Co-orthologs When speciation gives rise to orthologs, which is then followed by gene duplication, the groups of paralogs in later species are related to each other as co-orthologs.

Characters

- A *character* is a heritable changes in feature of an organism
  - morphology
  - DNA sequence
  - etc.
- The more similar characters there are between two organisms (or sequences) the closer the inferred relationship
- However, characters can be either *unique* or *non-unique* with respect to evolutionary history

Unique characters

- *Unique characters* evolved only once and are thus good markers of evolutionary history
- *Hair* is a unique character
    - If an organism has hair it is likely to be a mammal

Non-unique characters and convergent evolution

- Some traits emerge independently in a process of *convergent evolution*.
- The (non-unique) trait alone does not infer phylogeny
- Examples
  - wings
  - eyesight
  - tails
- Similarity through convergent evolution is called *homoplasy*

Topic

Background

Molecular Evolution and the tree of life

Inferring evolutionary history

Phylogenies and phylogenetics trees

Nucleotide substitution models

The process of inferring phylogeny from sequence data

Distance-based methods

Character-based methods

Lateral (horizonal) gene transfer

Roundup and next time

Evolutionary processes from ancestor to current species

Mutations in the genome create new alleles and change *allele frequency*. Changes that give an advantage in selection may become fixed

Forces at work

- Negative selection
- drift
- Positive selection
- genetic expansion (duplication)
- Genetic loss (deletions)
- *Horizonal gene transfer*

Purifying (negative) selection

- Random mutations in DNA are created
- If they occur at functionally important sites, these mutations will have *negative consequences* for fitness.
- *Therefore*, sequences at these sites will be conserved, and will change less than neutral changes

Negative selection

The natural selection *against* deleterious mutations, preserving functionally important sites

Positive selection

Definition (Positive Selection)

Darwinian selection and fixing of advantageous mutations

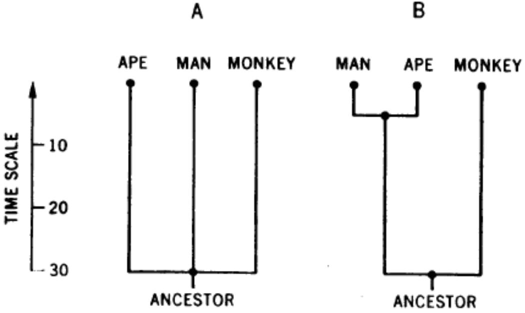
- Also called "molecular adaptation"
- Observable by ratio of

$$\frac{\text{nonsynonymous mutation rate}}{\text{synonymous mutation rate}}$$

Drift, or selectively *neutral* change

- In the 1960s, Kimura proposed that most mutations were selectively *neutral*, and thus most fixed change was from genetic drift, rather than natural selection.
- This is called *The Neutral Theory*.
- The Neutral Theory *is* compatible with the natural selection, but hypothesizes that *most* observed changes are neutral.
- The neutral theory (as theory) has not held up to formal tests of predictions of ancient gene frequencies.
- The neutral theory gives a rationale for a *molecular clock*.

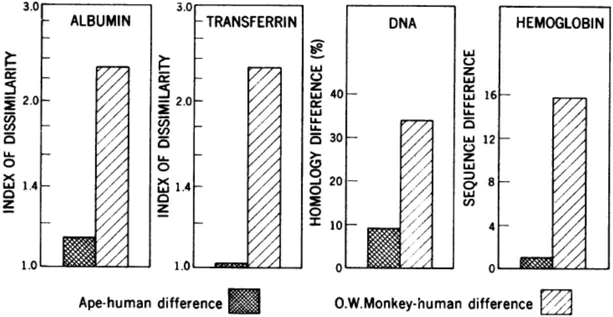
An example: alternative hypotheses for homonid evolution (1969)



Different hypthesis for divergence in the phylogeny of humans, great apes, and old world monkeys, from Wilson and Sarich 1969

Resolving the hypotheses using immunological affinity and DNA hybridization

Differences between human, ape and old world monkey by either microcomplement fixation (protein) or DNA hybridization (1969).

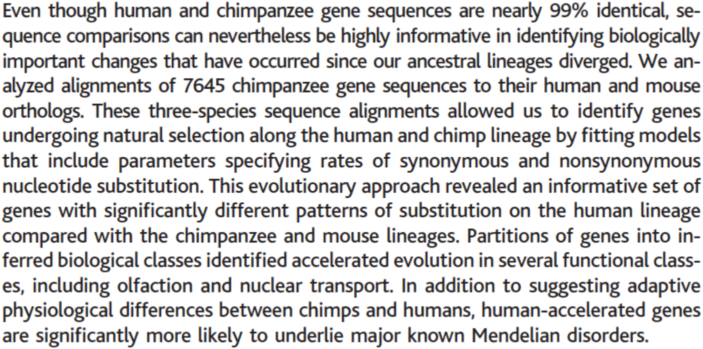


Synonymous versus non-synonymous mutations

- The degeneracy of the genetic code means some nucleotide mutations will not lead to protein change.
- Transition/transversion occur at unequal rates
- Positive selection protein sequences can be inferred from elevated rate of nonsynonymous mutations over a neutral model

Our example again (revisited in 2003)

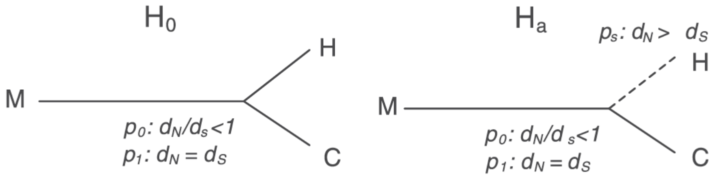
The abstract of Clark et al. 2003



Even though human and chimpanzee gene sequences are nearly 99% identical, sequence comparisons can nevertheless be highly informative in identifying biologically important changes that have occurred since our ancestral lineages diverged. We analyzed alignments of 7645 chimpanzee gene sequences to their human and mouse orthologs. These three-species sequence alignments allowed us to identify genes undergoing natural selection along the human and chimp lineage by fitting models that include parameters specifying rates of synonymous and nonsynonymous nucleotide substitution. This evolutionary approach revealed an informative set of genes with significantly different patterns of substitution on the human lineage compared with the chimpanzee and mouse lineages. Partitions of genes into inferred biological classes identified accelerated evolution in several functional classes, including olfaction and nuclear transport. In addition to suggesting adaptive physiological differences between chimps and humans, human-accelerated genes are significantly more likely to underlie major known Mendelian disorders.

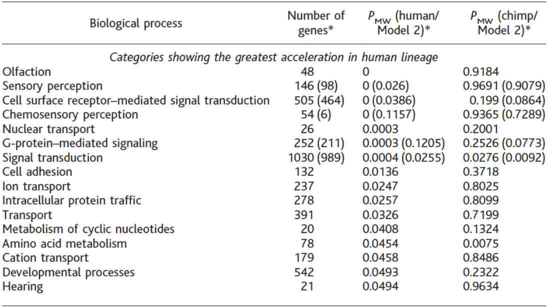
Two alternative models of molecular change

The rates d_S and d_N are the estimated rates of synonymous and nonsynonymous mutations



Some kinds of genes have been subject to positive selection in the human lineage from common ancestor with chimp

Biological processes (Gene Ontology) most significantly associated with positive selection in human lineage.



Biological process	Number of genes*	P_{MW} (human/ Model 2)*	P_{MW} (chimp/ Model 2)*
Categories showing the greatest acceleration in human lineage			
Olfaction	48	0	0.9184
Sensory perception	146 (98)	0 (0.026)	0.9691 (0.9079)
Cell surface receptor–mediated signal transduction	505 (464)	0 (0.0386)	0.199 (0.0864)
Chemosensory perception	54 (6)	0 (0.1157)	0.9365 (0.7289)
Nuclear transport	26	0.0003	0.2001
G-protein–mediated signaling	252 (211)	0.0003 (0.1205)	0.2526 (0.0773)
Signal transduction	1030 (989)	0.0004 (0.0255)	0.0276 (0.0092)
Cell adhesion	132	0.0136	0.3718
Ion transport	237	0.0247	0.8025
Intracellular protein traffic	278	0.0257	0.8099
Transport	391	0.0326	0.7199
Metabolism of cyclic nucleotides	20	0.0408	0.1324
Amino acid metabolism	78	0.0454	0.0075
Cation transport	179	0.0458	0.8486
Developmental processes	542	0.0493	0.2322
Hearing	21	0.0494	0.9634

Topic

Background

Molecular Evolution and the tree of life

Inferring evolutionary history

Phylogenies and phylogenetics trees

Nucleotide substitution models

The process of inferring phylogeny from sequence data

Distance-based methods

Character-based methods

Lateral (horizonal) gene transfer

Roundup and next time

Some definitions

Phylogenetics the study of evolutionary relations using genetic data, quantitative traits, characters, and morphological data matrices

Molecular Phylogenetics the use of molecular data such as DNA and protein sequences to infer phylogenies

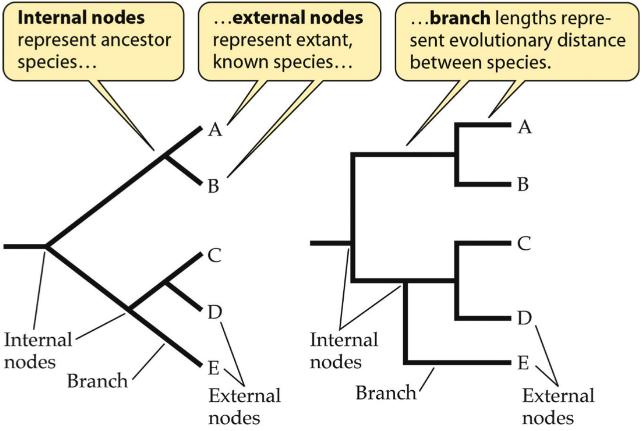
Phylogenetic tree A branching diagram (a tree) showing inferred evolutionary relationships (phylogenies)

Why phylogenetic analysis?

- Phylogenetic trees are the key structure to empower clear thinking about the relationships between species, and between sequences.
- The challenge of phylogenetics is that we generally only have genetic data on sequences that exist today, and we infer ancestral sequences

What do phylogenetic trees look like?

Two different display formats to show relatedness



Internal nodes represent ancestor species…

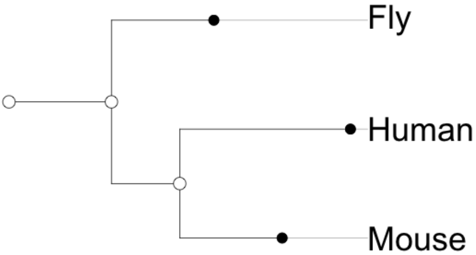
…**external nodes** represent extant, known species…

…**branch** lengths represent evolutionary distance between species.

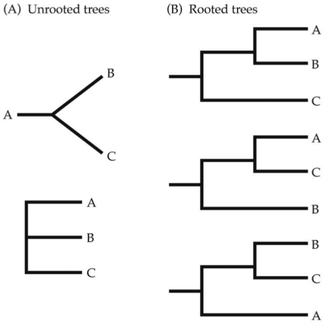
MICROBIAL LIFE, **Figure 17.4** © 2002 Sinauer Associates, Inc.

A phylogenetic tree with branch lengths

Branch lengths may be used to show some measure of evolutionary distance



Unrooted and rooted trees



MICROBIAL LIFE, Figure 17.5 © 2002 Sinauer Associates, Inc.

The roots of a tree

- if a tree is rooted, the nodes represent inferred *most recent common acncestor*
- sometimes, the lengths of branches on a rooted tree can be meaningful
- Unrooted trees often *look* rooted because of the way they are displayed.

A few more terms

Bifurcating trees two descendents arise from every interior node in a phylogenetic tree

Dendrogram A general term for the diagram of a phylogenetic tree. Also used for non-phylogenetic trees in other forms of data analysis

Phylogram A tree where branch lengths represent amount of character change (used in parsimony)

Chronogram A tree where branch lengths represent evolutionary time

Topic

Background

Molecular Evolution and the tree of life

Inferring evolutionary history

Phylogenies and phylogenetics trees

Nucleotide substitution models

The process of inferring phylogeny from sequence data

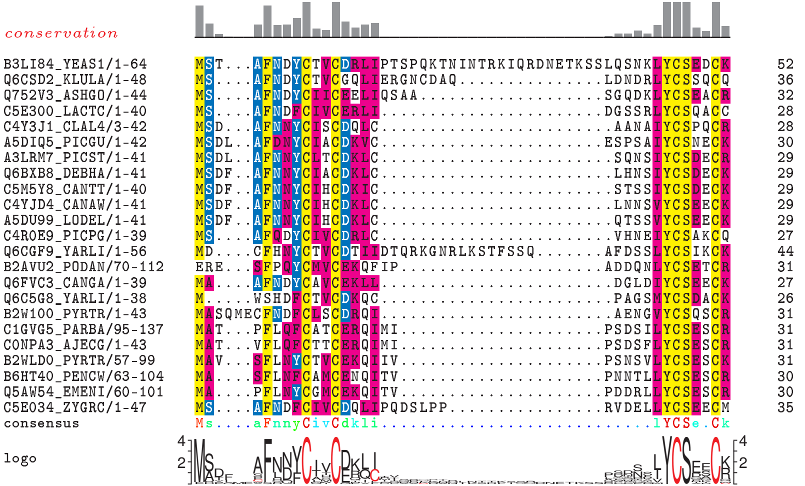
Distance-based methods

Character-based methods

Lateral (horizonal) gene transfer

Roundup and next time

Multiple Sequence Alignment aligns *characters subject to selection*

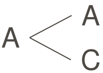


```
conservation

B3LI84_YEAS1/1-64     MST...AFNDYCTVCDRLIPTSPQKTNINTRKIQRDNETKSSLQSNKLYCSEDCK   52
Q6CSD2_KLULA/1-48     MS....AFNDYCTVCGQLIERGNCDAQ...............LDNDRLYCSSQCQ   36
Q752V3_ASHGO/1-44     MS....AFNDYCIICEELIQSAA...................SGQDKLYCSEACR   32
C5E300_LACTC/1-40     MS....AFNDFCIVCERLI.......................DGSSRLYCSQACC   28
C4Y3J1_CLAL4/3-42     MSD...AFNNYCISCDQLC........................AANAIYCSPQCR   28
A5DIQ5_PICGU/1-42     MSDL..AFDNYCIACDKVC.......................ESPSAIYCSNECK   30
A3LRM7_PICST/1-41     MSDL..AFNNYCLTCDKLC........................SQNSIYCSDECR   29
Q6BXB8_DEBHA/1-41     MSDF..AFNNYCIACDKLC........................LHNSIYCSDECK   29
C5M5Y8_CANTT/1-40     MSDF..AFNNYCIHCDKIC........................STSSIYCSDECK   29
C4YJD4_CANAW/1-41     MSDF..AFNNYCIHCDKLC........................LNNSVYCSEECK   29
A5DU99_LODEL/1-41     MSDF..AFNNYCIHCDKLC........................QTSSVYCSEECK   29
C4R0E9_PICPG/1-39     MS....AFQDYCIVCDRLC........................VHNEIYCSAKCQ   27
Q6CGF9_YARLI/1-56     MD....CFHNYCTVCDTIIDTQRKGNRLKSTFSSQ.......AFDSSLYCSIKCK   44
B2AVU2_PODAN/70-112   ERE...SFPQYCMVCEKQFIP.....................ADDQNLYCSETCR   31
Q6FVC3_CANGA/1-39     MA....AFNDYCAVCEKLL........................DGLDIYCSEECK   27
Q6C5G8_YARLI/1-38     M.....WSHDFCTVCDKQC........................PAGSMYCSDACK   26
B2W100_PYRTR/1-43     MASQMECFNDFCLSCDRQI........................AENGVYCSQSCR   31
C1GVG5_PARBA/95-137   MAT...PFLQFCATCERQIMI.....................PSDSILYCSESCR   31
C0NPA3_AJECG/1-43     MAT...VFLQFCTTCERQIMI.....................PSDSFLYCSESCR   31
B2WLD0_PYRTR/57-99    MAV...SFLNYCTVCEKQIIV.....................PSNSVLYCSESCK   31
B6HT40_PENCW/63-104   MA....SFLNFCAMCENQITV.....................PNNTLLYCSESCR   30
Q5AW54_EMENI/60-101   MA....PFLNYCGMCEKQITV.....................PDDRLLYCSESCR   30
C5E034_ZYGRC/1-47     MS....AFNDFCIVCDQLIPQDSLPP................RVDELLYCSEECM   35
consensus             Ms....aFnnyCivCdkli............................lYCSe.Ck
logo
```

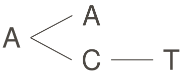
Substitutions $\neq$ evolutionary distance

Single



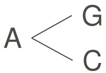
One change,
one difference

Multiple



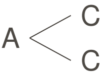
Two changes,
one difference

Coincidental



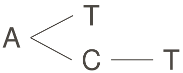
Two changes,
one difference

Parallel



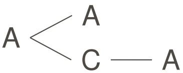
Two changes,
no difference

Convergent



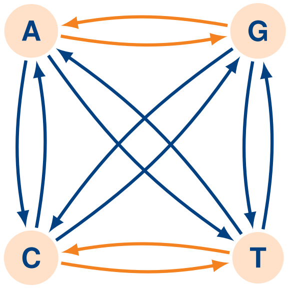
Three changes,
no difference

Back



Two changes,
no difference

Nucleotide transitions and transversions



Purines and pyrimidines

- Transitions substitute purine for purine or pyrimidine for pyrimidine
- Transversions substitute purine for pyrimidine or vice versa
- Transitions are more frequent than transversions

A second look at the genetic code

TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys
TTC	Phe	TCC	Ser	TAC	Tyr	TGC	Cys
TTA	Leu	TCA	Ser	TAA	Stp	TGA	Stp
TTG	Leu	TCG	Ser	TAG	Stp	TGG	Trp
CTT	Leu	CCT	Pro	CAT	His	CGT	Arg
CTC	Leu	CCC	Pro	CAC	His	CGC	Arg
CTA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CTG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser
ATC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
ATA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
ATG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly
GTC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GTA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GTG	Val	GCG	Ala	GAG	Glu	GGG	Gly

Topic

Background

Molecular Evolution and the tree of life

Inferring evolutionary history

Phylogenies and phylogenetics trees

Nucleotide substitution models

The process of inferring phylogeny from sequence data

Distance-based methods

Character-based methods

Lateral (horizonal) gene transfer

Roundup and next time

The basic steps

1. Decide what sequences to include in the tree
2. Align the sequences
3. Estimate the tree
4. Root the tree
5. Draw the tree

Deciding what sequences to include

- This is the task of deciding the tips of the tree
- It is often the most time-consuming
- Think about the question you are trying to answer
- How to do it?
    - The most common approach is one you know well: BLAST

Aligning the sequences

- you can use any of the standard MSA programs like CLUSTAL, MUSCLE, or T-Coffee
- You may need to edit the alignments by hand if you know something about your homologies
- All of the caveats of MSA apply here

Remember the purpose of MSA

We use MSA to put homologous sequences in the same column of an alignment

Determining the model of nucleotide evolution

- The *Jukes-Cantor* model assumes equal transition and transversion frequencies
- This is the simplest possible model
- MEGA and other tools let you choose an appropriate model based on nucleotide sequence alignments

How to estimate a tree

There are too many trees to enumerate all of them (10^{195} rooted, labeled trees for 100 species) , so we must dramatically restrict our search

We generally categorize methods of tree construction into two forms:

Distance-based methods After alignment, aligned sequences are converted into *distances* representing pairwise differences between sequences.

Character-based methods After alignment, each position in the alignment is compared. Homologous positions must be aligned!

How to root a tree

Outgroup place root between distant sequence (chosen by you) and the rest of the tree

Midpoint place root at midpoint of longest path between two leaves

Gene duplication place root between paralogous gene copies

Assume a molecular clock see UPGMA

Topic

Background

Molecular Evolution and the tree of life

Inferring evolutionary history

Phylogenies and phylogenetics trees

Nucleotide substitution models

The process of inferring phylogeny from sequence data

Distance-based methods

Character-based methods

Lateral (horizonal) gene transfer

Roundup and next time

What distance-based methods share

- When aligning sequence S_i with sequence S_j
- Calculate a *dissimilarity* matrix that has some measure of *distance* between sequences. We call this D_{ij}.
- For D_{ij}, we often use the % mismatch between S_i and S_j at aligned positions
- Other forms of distance are possible

Forms of distance

- *Metric* distances obey the triangle inequality:

$$D_{ik} \leq D_{ij} + D_{jk}$$

- A uniform molecular clock assumes an *ultrametric* distance

$$D_{ik} \leq \max\left(D_{ij}, D_{jk}\right)$$

Distance corrections

- Differences seen in alignments are often adjusted under different models of evolution.
- For example, the Jukes-Cantor distance is used to describe nucleotide substitutions, assuming equal transitions and transversions

$$D_{xy} = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right)$$

where D_{xy} is the distance between sequences x and y in terms of number of changes per site, and p is the *observed* proportion of sites that differ between them

Two distance-based methods

Neighbor Joining

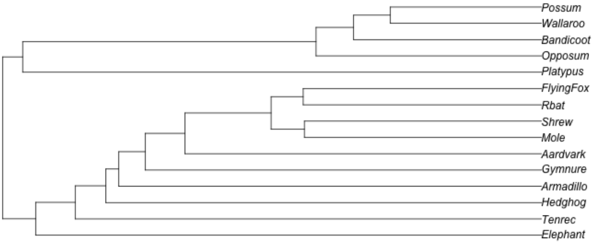
- gives an unrooted tree
- used to build guide trees in MSA

Unweighted Pair Group Mean Average (UPGMA)

- gives a rooted tree
- assumes a molecular clock

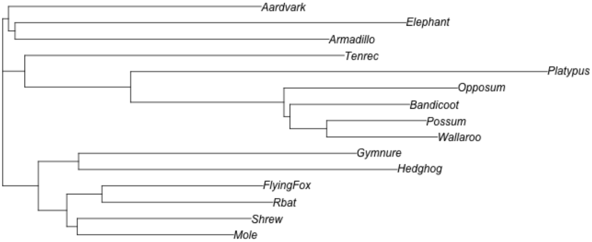
A phylogenetic tree of contemporary sequences using a molecular clock

A uniform clock (ultrametric distances) leads to identical distances from the root of a tree to all the leaves



Non-uniform speeds (no molecular clock)

a tree that does not assume a uniform clock will have different distances from leaves to the root.



Neighbor Joining (what it's for)

- Guaranteed to produce a correct tree if distances are additive
- May work well, even if distances are not additive
- A global measure, keeping total branch length to a minimum
- Leads to an unrooted tree (all additive trees are unrooted)

Key concept

divergence is a measure of total branch lengths in the neighbor joining process

Neighbor Joining: how it works

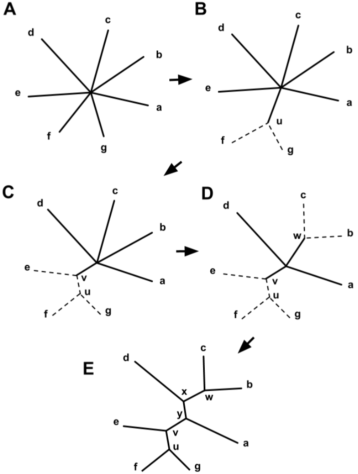
- Calculate a distance matrix between sequences
  - Represent the starting structure as a *star tree*
  - This tree is missing internal nodes
- Using the distance matrix, calculate the net divergence Q_{ij} of each taxon pair among the r sequences

$$Q_{ij} = (r-2)D_{ij} - \sum_{k=1}^{r} D_{ij} - \sum_{k=1}^{r} D_{jk}$$

- The pair with the lowest net divergence are joined
- Reduce the size of the matrix and repeat until done

Neighbor joining: illustration

From Wikipedia, under CCA license



Neighbor Joining: Results

- Fast
- unrooted
- Generally best for fairly closely related sequences. MEGA suggests use of Jukes-Cantor distances to decide if NJ is a good idea

Unweighted Pair Group Mean Average (UPGMA)

How

- pair with smallest distance is joined, with node halfway between them
- These become a cluster
- New distance matrix is calculated with the cluster replacing the individual entries
- Repeat until done

Results

- Ultrametric, assumes molecular clock
- Produces rooted tree, with all branches ending at the same time point
- Example from Dr Richard J. Edwards at University of Southampton

Topic

Background

Molecular Evolution and the tree of life

Inferring evolutionary history

Phylogenies and phylogenetics trees

Nucleotide substitution models

The process of inferring phylogeny from sequence data

Distance-based methods

Character-based methods

Lateral (horizonal) gene transfer

Roundup and next time

The main approaches

Character-based methods look at each column in the alignment

Parsimony seeks the minimum number of changes. Parsimony often results in several equivalent trees

Maximum Likelihood Under a model of evolution, what tree maximizes the likelihood of observing the data? How likely is it?

Bayesian analysis Given the data, which tree (or trees) have the highest likelihood?

Parsimony

- You have seen this before in PAM matrix construction
- Parsimony seeks the tree that requires the least amount of evolution (the most parsimonious set of changes)
- Parsimony trees are found (with no guarantee of optimality) by a variety of computer science algorithms
  - heuristic methods
  - branch and bound
  - nearest neighbor interchange

Parsimony and long branch attaction

What happens when a character evolves rapidly?

- in DNA, with only four letters, a rapidly evolving site might be inferred to be unchanges, when in reality it will have changed twice (e.g., A -> C -> A).
- This can lead to statistically inconsistent results
- Maximum Parsimony is particularly prone to this behavior
- The more data, the more likely to be wrong!
- Various detection and solution methods, all imperfect

Maximum Likelihood

Given

- some data (an alignment)
- a model (a model of evolutionary change)
- A hypothesis (a tree)

Maximum likelihood . . .

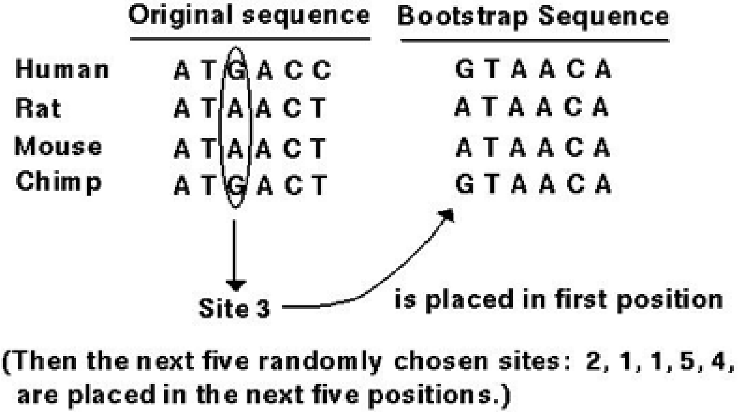
selects the hypothesis (the tree) that maximizes the probability of observing the data under the model

How reliable is a tree?

- For parsimony and maximum likelihood methods, we use *resampling* methods to estimate the reliability of a tree.
- The most common resample methods are *bootstrap* and *jacknife*
- In a bootstrap sample, every character is replaced at random by another character
- A bootstrap estimate looks at >1000 such samples

Bootstrapping

graphical description of bootstrap, from UBC Zoology department



	Original sequence	Bootstrap Sequence
Human	A T G A C C	G T A A C A
Rat	A T A A C T	A T A A C A
Mouse	A T A A C T	A T A A C A
Chimp	A T G A C T	G T A A C A

Site 3 is placed in first position

(Then the next five randomly chosen sites: 2, 1, 1, 5, 4, are placed in the next five positions.)

Rooting a tree

Outgroup approach

- In this, we know the answer: a relation that is *not* in the group of interest
- Too close, we may not be sure it is an outgroup
- Two far, it may not be useful for rooting the remaining relationships

Molecular clock approach

- If we assume an equivalent rate of change on all characters, this forms a molecular clock
- Somewhere on the tree, there should be a point where everything is equidistant to the tips
- The PAM1 matrix construction used this model

Topic

Background

Molecular Evolution and the tree of life

Inferring evolutionary history

Phylogenies and phylogenetics trees

Nucleotide substitution models

The process of inferring phylogeny from sequence data

Distance-based methods

Character-based methods

Lateral (horizonal) gene transfer

Roundup and next time

What does the tree of life represent?

- We'd like it to represent evolutionary history
- But in some cases, there appear to be exceptions to the simple branching of trees
- Some plants species combine the genomes of parental species (a form of species formation by hybridization)
- Mitochondria and plastids orginate from symbiosis
- *Horizontal transfer* of genetic material connects branches of the tree of life

What is it

Definition (Horizontal Gene transfer)

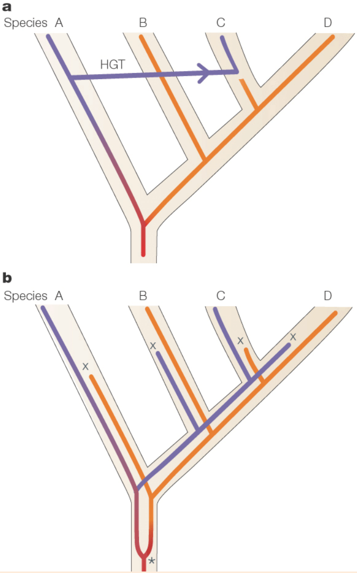
Horizontal Gene Transfer the transfer of genetic material between organisms through any means other than from parent to offspring (vertical gene transfer)

Lateral Gene Transfer Horizontal Gene Transfer

How does it occur?

1. Bacterial mechanisms (F-pilus)
2. Viral payload
3. Membrane shock

Appearance of genes characteristic of unrelated group

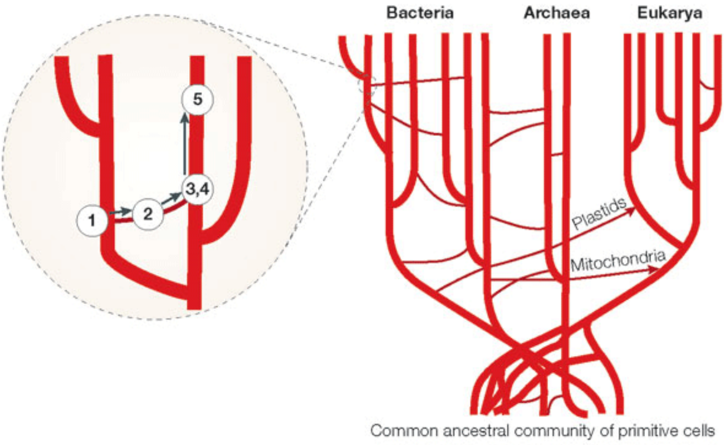


Two models

1. Horizonal Gene Transfer
2. Duplication followed by gene loss

Gogarten and Townsend 2005

A tree of life with horizontal gene transfer



Copyright © 2005 Nature Publishing Group
Nature Reviews | Microbiology

the inset shows a model of five required steps for HGT, from Smets and Barkay 2005

Steps for HGT

1. Preparation of genetic material for transfer
2. Transfer between donor and recipient organism
3. Entry into recipient organism
4. Establishment of the genetic material’s ability to replicate in the recipient
5. Stable inheritance

The extent of HGT is controversial

- There is no doubt it occurs, and it is important
- It complicates interpretation of the tree of life

Topic

Background

Molecular Evolution and the tree of life

Inferring evolutionary history

Phylogenies and phylogenetics trees

Nucleotide substitution models

The process of inferring phylogeny from sequence data

Distance-based methods

Character-based methods

Lateral (horizonal) gene transfer

Roundup and next time

What have we learned?

- Phylogenetic tree estimations are used to infer evolutinary history under a model of branching trees
- Inferring correct trees is challenging in many cases
- Rooting a tree can be done by different methods, depending on assumptions and tools such as outgroups
- A molecular clock assumption can give a rooted tree without an outgroup, but with often unjustified assumptions

Next time

- We move from sequence to structure
- What are the structures that define proteins, and how do we study the information in them?
- What are the resources and databases to study structure?

Bibliography

Ciccarelli, Francesca D et al. (2006). “Toward automatic reconstruction of a highly resolved tree of life.” In: *Science (New York, N.Y.)* 311.5765, pp. 1283–7. DOI: 10.1126/science.1123061 (cit. on p. 10).

Clark, Andrew G et al. (2003). “Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios.” In: *Science (New York, N.Y.)* 302.5652, pp. 1960–3. DOI: 10.1126/science.1088821 (cit. on p. 30).

Gogarten, J Peter and Jeffrey P Townsend (2005). “Horizontal gene transfer, genome innovation and evolution.” In: *Nature reviews. Microbiology* 3.9, pp. 679–87. DOI: 10.1038/nrmicro1204 (cit. on p. 75).

Smets, BF and Tamar Barkay (2005). “Horizontal gene transfer: perspectives at a crossroads of scientific disciplines”. In: *Nature Reviews Microbiology* 3.September, pp. 675–678 (cit. on p. 76).

Sonnhammer, Erik L L and Eugene V Koonin (2002). “Orthology, paralogy and proposed classification for paralog subtypes”. In: *Trends in Genetics* 18.12, pp. 619–620. DOI: 10.1038/nbt749 (cit. on p. 17).

Wilson, AC and VM Sarich (1969). “A molecular time scale for human evolution”. In: *Proceedings of the National . . .* (Cit. on p. 27).

Woese, C R and G E Fox (1977). “Phylogenetic structure of the prokaryotic domain: the primary kingdoms.” In: *Proceedings of the National Academy of Sciences of the United States of America* 74.11, pp. 5088–5090. DOI: