6.096 Lecture 9

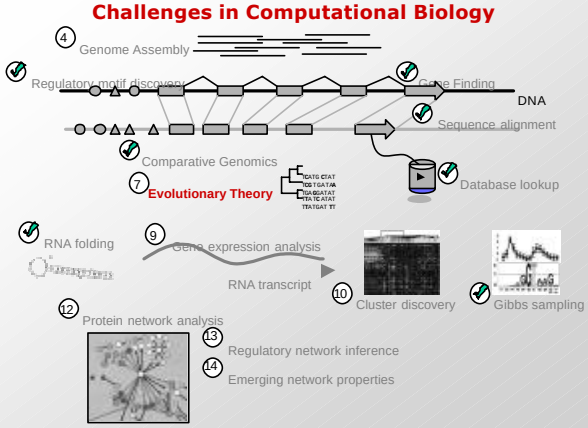
Evolution

Evolutionary Change – Phylogenetic Trees – Genome Evolution

Evolutionary change

Lecture 1 - Introduction
Lecture 2 - Hashing / BLAST
Lecture 3 - Combinatorial Motif Finding
Lecture 4 - Statistical Motif Finding
Lecture 5 - Sequence alignment and Dynamic Programming
Lecture 6 - RNA structure and Context Free Grammars
Lecture 7 - Gene finding and Hidden Markov Models
Lecture 8 - HMMs algorithms and Dynamic Programming
Lecture 9 - Evolutionary change, Phylogenetic trees

Challenges in Computational Biology



6.891 – Computational Evolutionary Biology

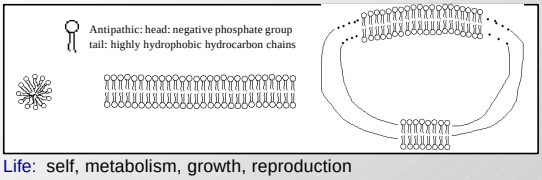
- Prof. Robert C. Berwick
- Graduate course
- 12 units
- Taught this coming fall

Overview

Early evolution
The last 3.5 billion years
Phylogenetic trees
UPGMA
Neighbor Joining
Parsimony
Rapid evolution

The first life form

- Primordial soup
  - Molecules of all sizes and shapes are floating
  - Nucleotides of all forms, probably amino acids too
  - Throw in some antipathic phospholipids: 3 stable structures

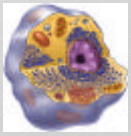


- Life: self, metabolism, growth, reproduction
- Self: must not dilute, keep useful molecules near each other
- Metabolism: take things from the outside and make them yours
- Growth: Stages of life, ever-changing, dynamic, directionality
- Self-replication: self-advantageous, improving on previous form

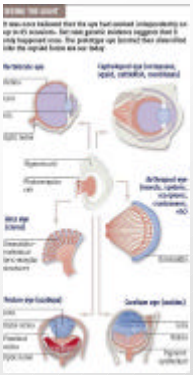
Great inventions happened once

- Life's great inventions
  1. Chemicals: molecules of life present in the soup
  2. Membrane: separate self from others
  3. Polymers: make complex structures from simple components
  4. Self-replication: molecules that can make more of themselves
  5. Catalysts: molecules that favor particular reactions
  6. Specialized polymers: DNA, RNA, Proteins, ribosome, tRNA
  7. Molecule modifications: splicing, editing, protein modifications

- From prokaryotes to complex life forms
  1. Subcellular: Nucleus, organelles, ER, golgi, lysosome
  2. Multicellular: Interactions, communication, symbiosis
  3. Differenciation: Cells specialize function, cell fate
  4. Body plan: high-level control of complex interactions
  5. AP credit: learning, language, writing, typing, squash



Top 10 greatest inventions



1. Multi-cellularity
2. **The eye**
3. The brain
4. Language
5. Photosynthesis
6. Sex
7. Death
8. Parasitism
9. Superorganisms
10. Symbiosis

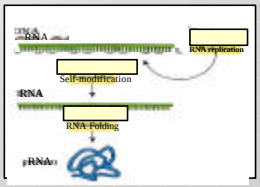
New Scientist, April 9, 2005

Courtesy of New Scientist. Used with permission.

http://www.newscientist.com/channel/life/mg18624941.700

RNA World

- One compelling snapshot in the early stages of life on Earth
- RNA can catalyze enzymatic reactions
  - 2ndary fold can act like a protein-like helper to reactions
  - Proteins are more efficient, but arose later
- RNA can pass on inherited information
  - By complementarity, RNA can transfer information to progeny
  - RNA can be reverse-transcribed to DNA (today)
  - RNA polymerization can be catalyzed by a ribozyme in a non-specific manner, replicating any RNA by complementarity
  - DNA is more stable, but arose later
- RNA World is possible
- RNA invents successors
  - RNA invents protein
    - Ribosome core is RNA
    - Translation from RNA template
  - RNA and protein invent DNA
    - Stable, protected, specialized structure, no catalysis
    - Proteins do: RNA→DNA reverse transcription
    - Proteins do: DNA→DNA replication
    - Proteins do: DNA→RNA transcription



Overview

Early evolution

The last 3.5 billion years

Phylogenetic trees

UPGMA

Neighbor Joining

Parsimony

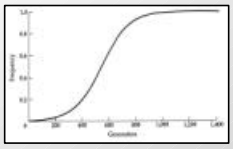
Rapid evolution

Evolutionary Mechanisms

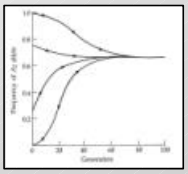
- Types of mutations
  - Single substitution: A to C, G or T, etc.
  - Deletion: 1 bp ... chromosomes (aneuploidy)
  - Duplication: as above (often at tandem repeats)
  - Inversion: ABCDEFG to ABedcFG
  - Translocation: ABCD & WXYZ to ABYZ & WXCD
  - Insertion: ABCD to ABινσερτCD
  - Recombination: ABCDEFGH → ABcDEFGH
    Abc DEf GH → ABCDEf GH

Selective pressure

- **Directional selection**
  - One allele is useful. Homozygote most fit
  - Fitness of heterozygote is the mean of the fitness of the two homozygotes
    AA = 1; Aa = 1 + s; aa = 1 + 2s
  - Always increase frequency of one allele at expense of the other



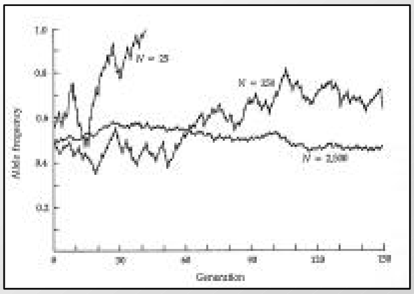
- **Stabilizing selection**
  - *Heterozygote most fit*
  - heterozygote has highest fitness
    AA = 1, Aa = 1 + s; aa = 1 + t
    where 0 < t < s
  - reach equilibrium where two alleles coexist



H&C 1997 p. 229

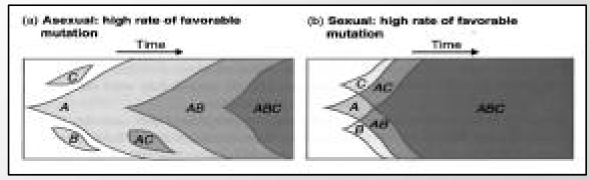
Genetic Drift

- The larger the effective population size, the smaller the drift
- With small populations, alleles can appear and disappear without selective pressures



Allele fixation

- Motivating sexual reproduction
  - Genetic exchange is facilitated. Exchange of alleles more frequent.
  - Force interbreeding at a cost: 50% less chances of reproduction

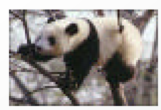


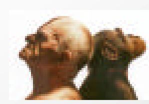
from Crow & Kimura 1970
Clark & Hartl 1997 p. 182

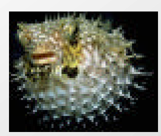
Overview

Early evolution
The last 3.5 billion years
Phylogenetic trees
UPGMA
Neighbor Joining
Parsimony
Rapid evolution

Open questions (?)







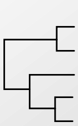
- Panda
  - Bear or raccoon?
- Out of Africa
  - mitochondrial evolution story?
- Human evolution
  - Did we ever meet Neanderthal?
- Primate evolution
  - Are we chimp-like or gorilla-like?
- Vertebrate evolution
  - How did complex body plans arise?
- Recent evolution
  - What genes are under selection?

Inferring Phylogenies

Trees can be inferred by several criteria:

- Morphology of the organisms
- Sequence comparison

Example:



Kangaroo:	ACAGTGACGCCCCAAACGT
Elephant:	ACAGTGACGCTACAAACGT
Dog:	CCTGTGACGTAACAAACGA
Mouse:	CCTGTGACGTAGCAAACGA
Human:	CCTGTGACGTAGCAAACGA

Traits – as many as we have letters in DNA

```
YAL042W      -MKRSTLLSLDAFAKTEEDVRVRTRAGGLITLSCILTTLFLLVNEWGQFNSVVTRPQLVV
candida586   MSSRPKLLSFDAFAKTVEDARIKTTSGGIITLICILITLVLIRNEYVDYTTIITRPELVV
cdub17784    MSSRPKLLSFDAFAKTVEDARIKTTSGGIITLICILITLVLIRNEYVDYTTIITRPELVV
cgla72177    -MKKSTLLSFDAFAKTEEDVRIRTRSGGFITLGCLVVTLMLLLSEWRDFNSVVTRPELVI
cgui48535    -MPQPKLLSFDAFAKTVEDARVRTPAGGIITLICVIVVLYLIRNEYLEYTSIINRPELVV
clus15345    MSSRPRLLSLDAFAKTVEDARVKTASGGVITLVCVLIVLFLIRNEYSDYMLVVVRPELVV
ctro67868    MSSRPKLLSFDAFAKTVEDARIKTASGGIITLICVLITLILIRNEYIDYTTIITRPELVV
klac20931    -MKKSPLLSIDAFGKTEEDVRVRTRTGGLITVSCIIITMLLLVSEWKQFSTIVTRPDLVV
                :. ***:***.** **.*::* :**.**: *:: .: *: .*: ::  :: **:**:

YAL042W      DRDRHAKLELNMDVTFPSMPCDLVNLDIMDDSGEMQLDILDAGFTMSRLNSEG------R
candida586   DRDINKQLDINLDISFINLPCDLISIDLLDVTGDLSLNIIDSGLKKIRLLKNKQGDVIVN
cdub17784    DRDINKQLDINLDISFINLPCDLISIDLLDVTGDLSLNIIDSGLKKIRLLKNKQGDVIVN
cgla72177    DRDRSLRLDLNLDITFPSMPCELLTLDIMDDSGEVQLDIMNAGFEKTRLSKEG------K
cgui48535    DRDINKKLEINLDISFPDIPCDVLTMDILDVSGDLQVDLLLSGFEKFRLLKDG------L
clus15345    NRDVNRQLDINLDITFPDVPCGVMSLDILDMTGDLHLDIVESGFEMFRVLPLG------E
ctro67868    DRDINKQLDINLDISFINLPCDLISVDLLDVTGDQQLDIIDSGLKKVRLLKNKQGDVIIN
klac20931    DRDRHLKLDLNLDVTFPSMPCNVLNLDILDDSGEFQINLLDSGFTKIRISPEG------K
             :**   :*::*:*::* .:** ::.:*::* :*:  :::: :*:   *:

YAL042W      PVGDATELHVGGNGDGTAPV--NND---PNY-CGPCYGAKDQSQN-ENLAQEEKVCCQDC
candida586   EIEDDEPAFNNDIELSDLAKGLPEGSDENAY-CGSCYGALPQDK--------KQFCCNDC
cdub17784    EIEDDEPAFNNDIELTDLAKGLPEGSDENAY-CGSCYGALPQDK--------KQFCCNDC
cgla72177    VLGTA-DMKIGEAAKKDKEA--QLAKLGANY-CGNCYGARDQGKNNDDTPRDQWVCCQTC
cgui48535    EIRDESPVMSSAGELEERAR----GRAPDGL-CGSCYGALPQDEN-------LDYCCNDC
clus15345    EISDDLPLLSGAKKFEDVCGPLTEDEISRGVPCGPCYGAVDQTD--------NKRCCNTC
ctro67868    EIEDDKPALNSDVSLKELAKGLPEGSDQNAY-CGPCYGALPQDK--------KQFCCNDC
klac20931    ELSKE-KFQVGDKS--SKQS--FNE---EGY-CGPCYGALDQSKN-DELPQDQKVCCQTC
              :        .                    ** ****  * .            **: *

YAL042W      DAVRSAYLEAGWAFFDGKNIEQCEREGYVSKINEHLN--EGCRIKGSAQINRIQGNLHFA
candida586   NTVRRAYAEKHWSFYDGENIEQCEKEGYVGRLRERINNNEGCRIKGTTKINRVSGTMDFA
cdub17784    NTVRRAYAEKHWSFYDGENIEQCEKEGYVARLRERINNNEGCRIKGTTKINRVSGTMDFA
cgla72177    DDVRQAYFEKNWAFFDGKDIEQCEREGYVQKIADQLQ--EGCRVSGSAQLNRIDGNLHFA
cgui48535    ETVRLAYAQKAWGFFDGENIEQCEREGYVARLNEKINNFEGCRIKGTGKINRISGNLHFA
clus15345    EAVRMAYAVQEWGFFDGSNIEQCEREGYVEKMVSRINNNEGCRIKGSAKINRISGNLHFA
ctro67868    NTVRRAYAEKQWQFFDGENIEQCEKEGYVKRLRERINNNEGCRIKGSTKINRVSGTMDFA
klac20931    DDVRAAYGQKGWAFKDGKGVEQCEREGYVESINARIH--EGCRVQGRAQLNRIQGTIHFG
             : ** **    * * **..:****:****  :  :::  ****:.*  ::**:.*.:.*.
```

Modeling Nucleotide Evolution

During infinitesimal time Δt, there is not enough time for two substitutions to happen on the same nucleotide

So we can estimate $P(x \mid y, \Delta t)$, for $x, y \in \{A, C, G, T\}$

Then let

$$S(\Delta t) = \begin{bmatrix} P(A|A, \Delta t) & \ldots & P(A|T, \Delta t) \\ \ldots & & \ldots \\ P(T|A, \Delta t) & \ldots & P(T|T, \Delta t) \end{bmatrix}$$

Modeling Nucleotide Evolution

Reasonable assumption: multiplicative (implying a stationary Markov process)

$S(t+t') = S(t)S(t')$

That is, $P(x \mid y, t+t') = \Sigma_z P(x \mid z, t)\, P(z \mid y, t')$

Jukes-Cantor: constant rate of evolution

For short time ε, $S(\varepsilon) = \begin{bmatrix} 1 - 3\alpha\varepsilon & \alpha\varepsilon & \alpha\varepsilon & \alpha\varepsilon \\ \alpha\varepsilon & 1 - 3\alpha\varepsilon & \alpha\varepsilon & \alpha\varepsilon \\ \alpha\varepsilon & \alpha\varepsilon & 1 - 3\alpha\varepsilon & \alpha\varepsilon \\ \alpha\varepsilon & \alpha\varepsilon & \alpha\varepsilon & 1 - 3\alpha\varepsilon \end{bmatrix}$

Modeling Nucleotide Evolution

Jukes-Cantor:

For longer times,

$$S(t) = \begin{bmatrix} r(t) & s(t) & s(t) & s(t) \\ s(t) & r(t) & s(t) & s(t) \\ s(t) & s(t) & r(t) & s(t) \\ s(t) & s(t) & s(t) & r(t) \end{bmatrix}$$

Where we can derive:

$$r(t) = \tfrac{1}{4}\,(1 + 3\, e^{-4\alpha t})$$
$$s(t) = \tfrac{1}{4}\,(1 - e^{-4\alpha t})$$

Modeling Nucleotide Evolution

Kimura:

Transitions: A/G, C/T
Transversions: A/T, A/C, G/T, C/G

Transitions (rate α) are much more likely than transversions (rate β)

$$S(t) = \begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{bmatrix} r(t) & s(t) & u(t) & u(t) \\ s(t) & r(t) & u(t) & u(t) \\ u(t) & u(t) & r(t) & s(t) \\ u(t) & u(t) & s(t) & r(t) \end{bmatrix} \end{array}$$

Where

$$s(t) = \tfrac{1}{4}\,(1 - e^{-4\beta t})$$
$$u(t) = \tfrac{1}{4}\,(1 + e^{-4\beta t} - e^{-2(\alpha+\beta)t})$$
$$r(t) = 1 - 2s(t) - u(t)$$

Phylogeny and sequence comparison

Basic principles:

- Degree of sequence difference is proportional to length of independent sequence evolution
- Only use positions where alignment is pretty certain – avoid areas with (too many) gaps

Distance between two sequences

Given (portion of) sequences x^i, x^j,

Define

d_{ij} = distance between the two sequences

One possible definition:

d_{ij} = fraction f of sites u where $x^i[u] \neq x^j[u]$

Better model (Jukes-Cantor):

$d_{ij} = -\tfrac{3}{4} \log(1 - 4f / 3)$

Overview

Early evolution
The last 3.5 billion years
Phylogenetic trees
UPGMA
Neighbor Joining
Parsimony
Rapid evolution

A simple clustering method for building tree

UPGMA (unweighted pair group method using arithmetic averages)

Given two disjoint clusters C_i, C_j of sequences,

$$d_{ij} = \frac{1}{|C_i| \times |C_j|} \sum_{\{p \in Ci,\, q \in Cj\}} d_{pq}$$

Note that if $C_k = C_i \cup C_j$, then distance to another cluster C_l is:

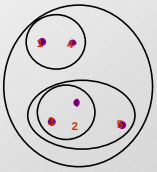
$$d_{kl} = \frac{d_{il}\,|C_i| + d_{jl}\,|C_j|}{|C_i| + |C_j|}$$

Algorithm: UPGMA

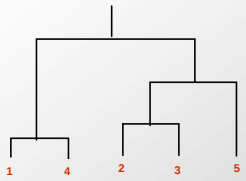
Initialization:
Assign each x_i into its own cluster C_i
Define one leaf per sequence, height 0

Iteration:
Find two clusters C_i, C_j s.t. d_{ij} is min
Let $C_k = C_i \cup C_j$
Define node connecting C_i, C_j,
& place it at height $d_{ij}/2$
Delete C_i, C_j

Termination:
When two clusters i, j remain,
place root at height $d_{ij}/2$



Ultrametric Distances & UPGMA



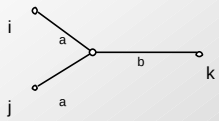
UPGMA is guaranteed to build the correct tree if distance is ultrametric

Proof:

1. The tree topology is unique, given that the tree is binary
2. UPGMA constructs a tree obeying the pairwise distances

Ultrametric distances

- For all points i, j, k
  - two distances are equal and third is smaller
    d(i,j) <= d(i,k) = d(j,k)
    a+a <= a+b = a+b



where a <= b

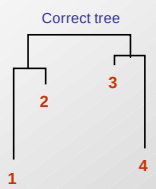
- Result:
  - All paths from labels are equidistant to the root
  - Rooted tree with uniform rates of evolution

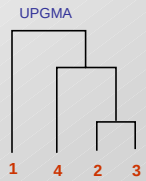
Weakness of UPGMA

Molecular clock assumption:
implies time is constant for all species

However, certain species (e.g., mouse, rat) evolve much faster

Example where UPGMA messes up:



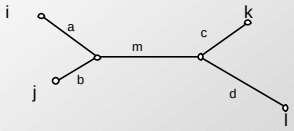


Overview

Early evolution
The last 3.5 billion years
Phylogenetic trees
UPGMA
Neighbor Joining
Parsimony
Rapid evolution

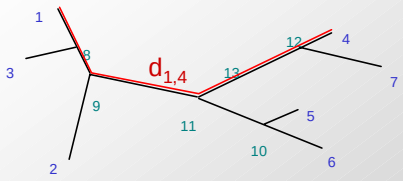
Additive trees

- All distances satisfy the four-point condition
  - For all i,j,k,l:
    - d(i,j) + d(k,l) <= d(i,k) + d(j,l) = d(i,l) + d(j,k)
    - (a+b)+(c+d) <= (a+m+c)+(b+m+d) = (a+m+d)+(b+m+c)



- Result:
  - All pairwise distances obtained by traversing a tree

Additive Distances



Given a tree, a distance measure is **additive** if the distance between any pair of leaves is the sum of lengths of edges connecting them

Given a tree T & additive distances d_{ij}, can uniquely reconstruct edge lengths:

- Find two neighboring leaves i, j, with common parent k
- Place parent node k at distance $d_{km} = ½ (d_{im} + d_{jm} – d_{ij})$ from any node m

Neighbor-Joining

- Guaranteed to produce the correct tree if distance is additive
- May produce a good tree even when distance is not additive

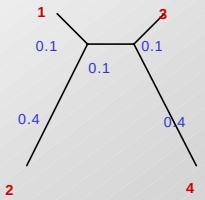
Step 1: Finding neighboring leaves

Define

$D_{ij} = d_{ij} – (r_i + r_j)$

Where

$$r_i = \frac{1}{|L| - 2} \Sigma_k d_{ik}$$



Claim: The above "magic trick" ensures that D_{ij} is minimal **iff** i, j are neighbors
Proof: Beyond the scope of this lecture (Durbin book, p. 189)

Algorithm: Neighbor-joining

Initialization:
Define T to be the set of leaf nodes, one per sequence
Let L = T

Iteration:
Pick i, j s.t. D_{ij} is minimal
Define a new node k, and set $d_{km} = ½ (d_{im} + d_{jm} – d_{ij})$ for all m ∈ L

Add k to T, with edges of lengths $d_{ik} = ½ (d_{ij} + r_i – r_j)$
Remove i, j from L;
Add k to L

Termination:
When L consists of two nodes, i, j, and the edge between them of length d_{ij}

Overview

Early evolution
The last 3.5 billion years
Phylogenetic trees
UPGMA
Neighbor Joining
Parsimony
Rapid evolution

Parsimony

- One of the most popular methods

Idea:

Find the tree that explains the observed sequences with a minimal number of substitutions

Two computational sub-problems:

1. Find the parsimony cost of a given tree (easy)
2. Search through all tree topologies (hard)

Parsimony Scoring

Given a tree, and an alignment column

Label internal nodes to minimize the number of required substitutions

Initialization:

Set cost C = 0; k = 2N – 1

Iteration:

If k is a leaf, set $R_k = \{ x^k[u] \}$

If k is not a leaf,

Let i, j be the daughter nodes;

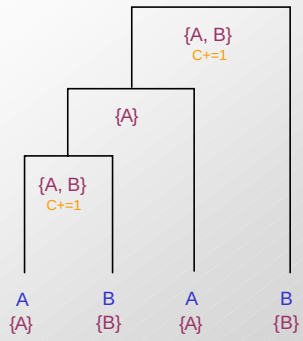
Set $R_k = R_i \cap R_j$ if intersection is nonempty

Set $R_k = R_i \cup R_j$, and C += 1, if intersection is empty

Termination:

Minimal cost of tree for column u, = C

Example



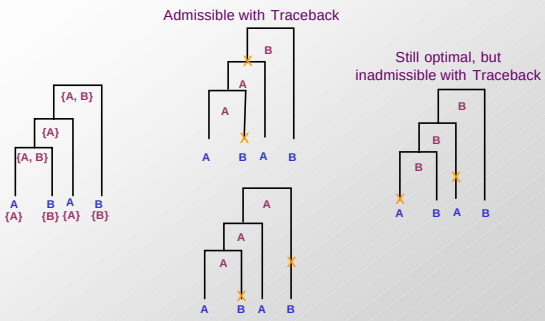
Traceback to find ancestral nucleotides

Traceback:

1. Choose an arbitrary nucleotide from R_{2N-1} for the root
2. Having chosen nucleotide r for parent k,
   If $r \in R_i$ choose r for daughter i
   Else, choose arbitrary nucleotide from R_i

Easy to see that this traceback produces some assignment of cost C

Example



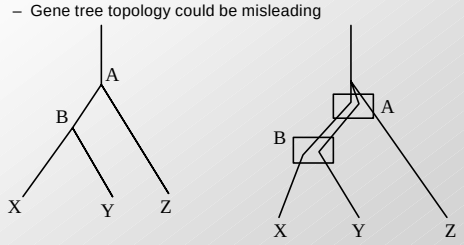
Bootstrapping to get the best trees

Main outline of algorithm

1. Select random columns from a multiple alignment – one column can then appear several times
2. Build a phylogenetic tree based on the random sample from (1)
3. Repeat (1), (2) many (say, 1000) times
4. Output the tree that is constructed most frequently

Gene vs. Species evolution

- Genes can start diverging before species separate
  - Genetic polymorphism within population could exist
  - After divergence, forms evolve differently in each species
  - Gene divergence could predate species diverge
  - Gene tree topology could be misleading



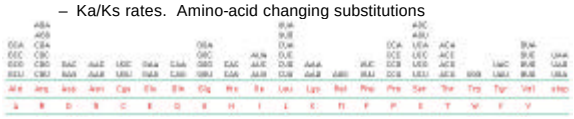
- Solution: Use multiple genes to infer a species tree

Overview

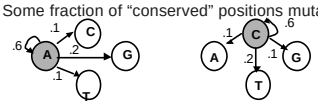
Early evolution
The last 3.5 billion years
Phylogenetic trees
UPGMA
Neighbor Joining
Parsimony
Rapid evolution

Measuring evolutionary rates

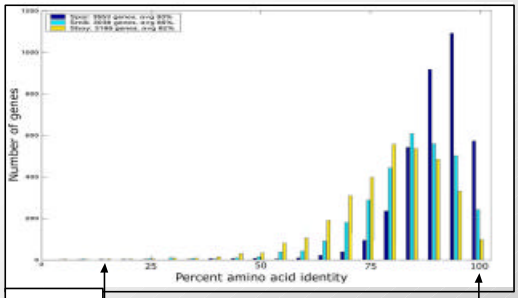
- Nucleotide divergence
  - Uniform rate. Overall percent identity.
- Transitions and transversions
  - Two-parameter model. A-G, C-T more frequent.
- Synonymous and non-synonymous substitutions
  - Ka/Ks rates. Amino-acid changing substitutions



- $N_{substitutions} > N_{mutations}$
  - Some fraction of "conserved" positions mutated twice



Fast and slow evolving genes



YBR184W

13% aa identity. Ka/Ks 7-fold higher. 120bp deletion

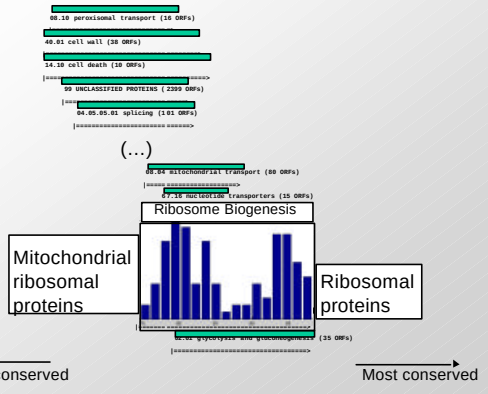
Spore-specific. Cell wall. 3 conserved domains.

MatA2

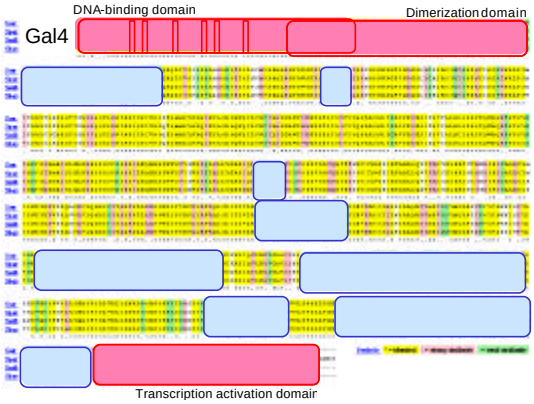
100% aa id. 100% nucleotide id. Mat-Alpha2 counterpart.

Role in mRNA complementation? Mating type switching?

Mutation rates by functional classification

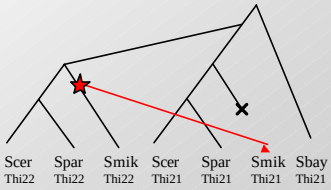


Protein domain evolution



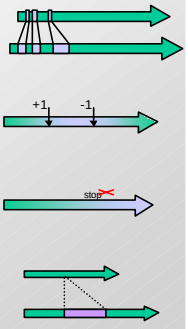
Gene loss / Gene conversion

- Observe positions of paralogs in *sensu stricto* to identify recently lost duplicates
  - Two copies in *S. bayanus*, one copy in *S. cerevisiae*. Recently lost in *S. cerevisiae* lineage
  - One copy in each genome, different chromosomes. Recently lost independently in both genomes
- Observe rates of change for both paralogs



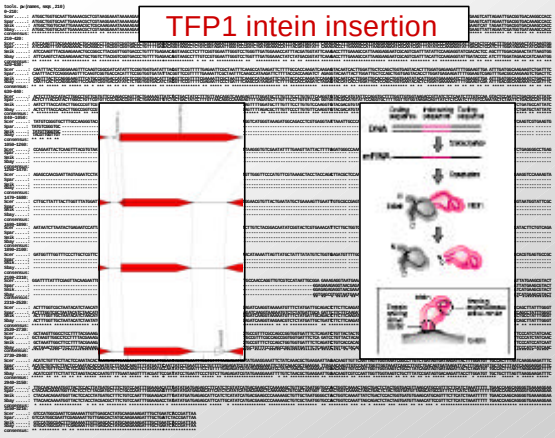
Rapid protein change

- Protein domain creation
  - Q/N stretches
  - Protein-protein interaction
- Compensatory frame-shifts
  - Explore new reading frames
  - RNA editing signals
- Stop-codon variation
  - Gain enables rapid change
  - Loss explores new diversity
  - Read-through is regulated
- Intein gain
  - Recent, present in *S.cerevisiae* only



Evolutionary shortcuts apparent in recent evolution

TFP1 intein insertion



Overview

Early evolution
The last 3.5 billion years
Phylogenetic trees
UPGMA
Neighbor Joining
Parsimony
Rapid evolution