<table>
<thead>
<tr>
<th>Introns early</th>
<th>Introns late</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Self splicing RNA are an example for catalytic RNA that could have been present in RNA world.</td>
<td>• Mapping individual introns onto organismal evolutionary history shows that many introns inserted into the sites where they are found presently more recently.</td>
</tr>
<tr>
<td>• There is little reason to assume that the RNA world was not plagued by self-splicing parasites.</td>
<td>• Intron preference for linker region could be the result of selection (they do less harm here than in tightly packed domains.</td>
</tr>
<tr>
<td>• Neighboring Introns are more frequently in same phase than expected by chance</td>
<td>• Exon shuffling in the maturation of the adaptive immune system is a modern trait of vertebrates.</td>
</tr>
<tr>
<td>• Spliceosomal introns are present in deep branching eukaryotes</td>
<td>• Even if introns are ancient, this does not prove that they played a role in assembling the now existing protein families.</td>
</tr>
<tr>
<td>• Introns frequently are found in linker regions</td>
<td></td>
</tr>
<tr>
<td>• Exon shuffling can create a large number of different catalytic sites (see the maturation of the immune system)</td>
<td></td>
</tr>
</tbody>
</table>
ATPase dataset from last Friday

Alignment

clustal vs muscle

Conserved part are aligned reproducibly
ATPase dataset from last Friday

Alignment

clustal  vs  muscle

The alignment of the less conserved parts is questionable. Using the progressive alignment approach on these sequences can cause problems in downstream analyses.
ATPase dataset from last Friday

Alignment PRANK vs muscle

Conserved part are aligned reproducibly; PRANK has more gaps, and more matches in pairwise alignments.
ATPase dataset from last Friday

Seaview -> muscle alignment -> GBLOCKS

The LAGLIDADG motifs of the homing endonuclease domains are part of conserved blocks
ATPase dataset from last Friday

Sieview: mase format saves site groups

Selected sites and sequences can be analyzed separately.
Tree from Extein
Muscle

Tree from Intein
PRANK

Tree from Intein
Muscle

[Alpha: 0.244036]

[Alpha: 1.195411]
ATPase dataset from last Friday
Reliably aligned positions determined with *guidance*

MSA color-coded by GUIDANCE scores
ATPase dataset from last Friday
Reliably aligned columns determined with GBLOCKS

Problem: If Glocks is used, one looses resolution in subgroups!
Terminology Review

- Branches, splits, bipartitions
- In a rooted tree: clades (for unrooted trees sometimes the term clann is used)
- Mono-, Para-, polyphyletic groups, cladists and a natural taxonomy

The term cladogram refers to a strictly bifurcating diagram, where each clade is defined by a common ancestor that only gives rise to members of this clade. I.e., a clade is monophyletic (derived from one ancestor) as opposed to polyphyletic (derived from many ancestors). (Note: you do need to know where the root is!)

A clade is recognized and defined by shared derived characters (= synapomorphies). Shared primitive characters (= sympleisiomorphies, alternative spelling is symplesiomorphies) do not define a clade, but a paraphyletic group. Homoplasies define polyphyletic groups (see in class example drawing ala Hennig).

To use these any terms you need to have polarized characters; for most molecular characters you don't know which state is primitive and which is derived (exceptions:....).
homology

Two sequences are homologous, if there existed an ancestral molecule in the past that is ancestral to both of the sequences

Types of Homology

**Orthologs**: “deepest” bifurcation in molecular tree reflects speciation. These are the molecules people interested in the taxonomic classification of organisms want to study.

**Paralogs**: “deepest” bifurcation in molecular tree reflects gene duplication. The study of paralogs and their distribution in genomes provides clues on the way genomes evolved. Gen and genome duplication have emerged as the most important pathway to molecular innovation, including the evolution of developmental pathways.

**Xenologs**: gene was obtained by organism through horizontal transfer. The classic example for Xenologs are antibiotic resistance genes, but the history of many other molecules also fits into this category: inteins, selfsplicing introns, transposable elements, ion pumps, other transporters,

**Synologs**: genes ended up in one organism through fusion of lineages. The paradigm are genes that were transferred into the eukaryotic cell together with the endosymbionts that evolved into mitochondria and plastids (the -logs are often spelled with "ue" like in orthologues)

see Fitch's article in TIG 2000 for more discussion.
What is in a tree?

Trees form molecular data are usually calculated as unrooted trees (at least they should be - if they are not this is usually a mistake).
To root a tree you either can assume a molecular clock (substitutions occur at a constant rate, again this assumption is usually not warranted and needs to be tested),
or you can use an outgroup (i.e. something that you know forms the deepest branch).
   For example, to root a phylogeny of birds, you could use the homologous characters from a reptile as outgroup; to find the root in a tree depicting the relations between different human mitochondria, you could use the mitochondria from chimpanzees or from Neanderthals as an outgroup; to root a phylogeny of alpha hemoglobins you could use a beta hemoglobin sequence, or a myoglobin sequence as outgroup.

Trees have a branching pattern (also called the topology), and branch lengths.

Often the branch lengths are ignored in depicting trees (these trees often are referred to as cladograms - note that cladograms should be considered rooted).
You can swap branches attached to a node, and in an unrooted you can depict the tree as rooted in any branch you like without changing the tree.
Test: Which of these trees is different?

More tests here
Phylogenetic Reconstruction – Why?

• A) Systematic classification of organisms
  e.g.: Who were the first angiosperms? (i.e. where are the first angiosperms located relative to present day angiosperms?) Where in the tree of life is the last common ancestor located?

B) Evolution of molecules
  e.g.: domain shuffling, reassignment of function, gene duplications, horizontal gene transfer, drug targets, detection of genes that drive evolution of a species/population (e.g. influenza virus)

C) Identification of organisms
  e.g., phylotyping in microbiome samples, origin of genes and viruses (e.g., HIV virus, recent ebola outbreak)
Phylogenetic analysis is an inference of evolutionary relationships between organisms. Phylogenetics tries to answer the question “How did groups of organisms come into existence?”

Those relationships are usually represented by tree-like diagrams.

Note: the equation of biological evolution with a tree-like process has limited validity at best.

Steps of the phylogenetic analysis

1. Compilation of sequence dataset
2. Alignment
3. Determination of substitution model
4. Tree building
5. Tree evaluation
Phylogenetic reconstruction - How

Distance analyses
  calculate pairwise distances
  (different distance measures, correction for multiple hits, correction for codon bias)

make distance matrix (table of pairwise corrected distances)

calculate tree from distance matrix

  i) using optimality criterion
     (e.g.: smallest error between distance matrix and distances in tree, or use
  ii) algorithmic approaches (UPGMA or neighbor joining)
Phylogenetic reconstruction - How

Parsimony analyses
find that tree that explains sequence data with minimum number of substitutions
(tree includes hypothesis of sequence at each of the nodes)

Maximum Likelihood analyses
given a model for sequence evolution, find the tree that has the highest probability under this model.
This approach can also be used to successively refine the model.

Bayesian statistics use ML analyses to calculate posterior probabilities for trees, clades and evolutionary parameters. Especially MCMC approaches have become very popular in the last year, because they allow to estimate evolutionary parameters (e.g., which site in a virus protein is under positive selection), without assuming that one actually knows the "true" phylogeny.
Else:
spectral analyses, like evolutionary parsimony, look only at patterns of substitutions,

Another way to categorize methods of phylogenetic reconstruction is to ask if they are using

an **optimality criterion** (e.g.: smallest error between distance matrix and distances in tree, least number of steps, highest probability), or

**algorithmic approaches** (UPGMA or neighbor joining)

Packages and programs available: PHYLIP, phylml, MrBayes, Tree-Puzzle, PAUP*, clustalw, raxml, PhyloGenie, HyPhy
Bootstrap?

• See [here](#)
Elliot Sober’s Gremlins

Observation: Loud noise in the attic

Hypothesis: gremlins in the attic playing bowling

Likelihood = $P(\text{noise} | \text{gremlins in the attic})$
Bayes’ Theorem

\[ P(\text{model}|\text{data}, I) = \frac{P(\text{model}, I) \cdot P(\text{data}|\text{model}, I)}{P(\text{data}, I)} \]

- **Posterior Probability**: represents the degree to which we believe a given model accurately describes the situation given the available data and all of our prior information I.

- **Prior Probability**: describes the degree to which we believe the model accurately describes reality based on all of our prior information.

- **Likelihood (P(data|model, I))**: describes how well the model predicts the data.

- **Normalizing constant (P(data,I))**: ensures the probability distribution sums to 1.
Alternative Approaches to Estimate Posterior Probabilities

Bayesian Posterior Probability Mapping with MrBayes

(Huelsenbeck and Ronquist, 2001)

Problem: Strimmer’s formula

\[ p_i = \frac{L_i}{L_1 + L_2 + L_3} \]

only considers 3 trees (those that maximize the likelihood for the three topologies)

Solution:

Exploration of the tree space by sampling trees using a biased random walk
(Implemented in MrBayes program)

Trees with higher likelihoods will be sampled more often

\[ p_i \approx \frac{N_i}{N_{\text{total}}} \]

, where \( N_i \) - number of sampled trees of topology \( i \), \( i=1,2,3 \)

\( N_{\text{total}} \) – total number of sampled trees (has to be large)
Illustration of a biased random walk

Image generated with Paul Lewis's MCRobot
Likelihood estimates do not take prior information into consideration:
e.g., if the result of three coin tosses is 3 times head, then the likelihood estimate for the frequency of having a head is 1 (3 out of 3 events) and the estimate for the frequency of having a head is zero.

$$P(A,B) = P(A)$$  The probability that both events (A and B) occur

$$P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$  Both sides expressed as conditional probability

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

If A is the model and B is the data, then

$P(B|A)$ is the likelihood of model A

$P(A|B)$ is the posterior probability of the model given the data.

$P(A)$ is the considered the prior probability of the model.

$P(B)$ often is treated as a normalizing constant.