Molecular Phylogenies Without Aligned Sequences

Peter F. Stadler

Bioinformatics Group, Dept. of Computer Science & Interdisciplinary Center for Bioinformatics, **University of Leipzig**

Institute for Theoretical Chemistry, Univ. of Vienna (external faculty)
The Santa Fe Institute (external faculty)

Bled, February 2005

What can be used?

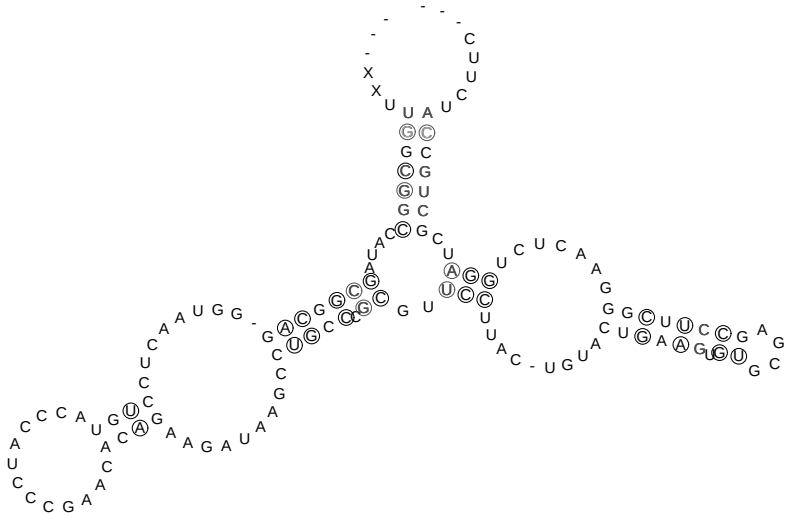
- "Molecular Morphology" of rRNAs
  (Caetano-Anollès, Misof)
- (Mitochondrial) Genome Structure
  (Sankoff, Boore, Warnow, . . .)
- Gene Content (Fitz-Gibbons, Snel)
- Repetitive Elements
- Presence/Absence of Phylogenetic Footprints
  (Prohaska et al)
- Structure of Metabolic Networks
- ???

I. Phylogenetic Usage of RNA Structures

- SSU rRNAs are the most commonly used class of sequences in molecular phylogenetics
- LSU rRNAs are also regularly used in molecular phylogenetics
- The combined tRNA complement has been shown to contain phylogenetic information comparable to rRNAs
- Recent work indicates that other structured non-coding RNAs are phylogenetically informative
- RNA secondary structure is a valuable source of information
- EST sequencing covers a signficant number of ncRNAs, many of which are structured

Evolution of Structured RNA

- Sequences in stems evolve highly correlated

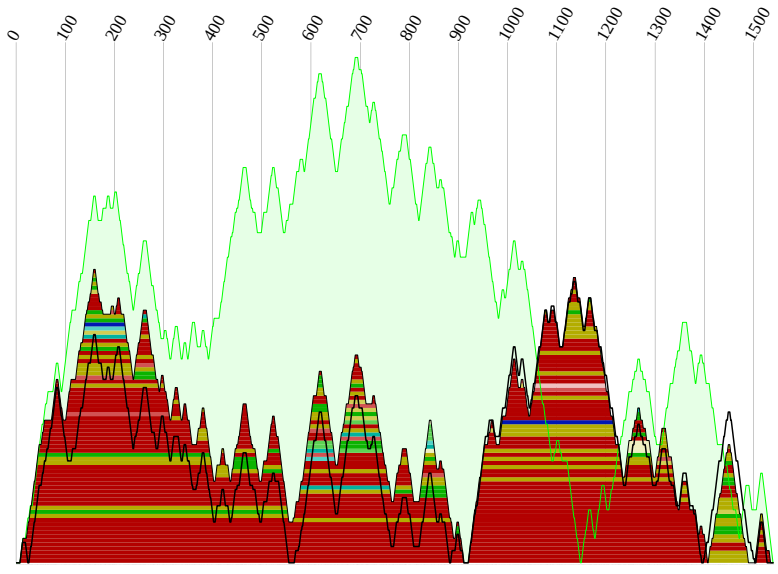


Sequence variation in stem regions of nine 5S rRNAs from Methanomicrobiales.
Almost all substitutions in stems preserve the structure.

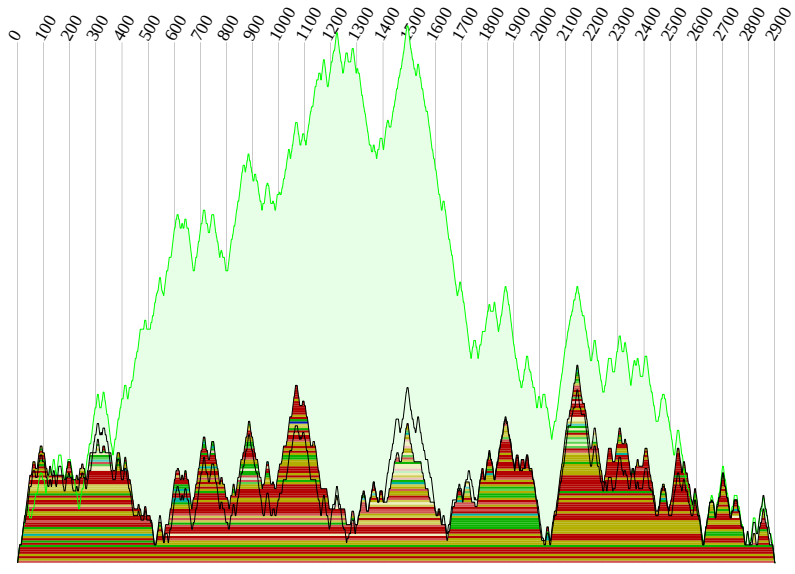
- Consequences:
  - All methods that assume independent positions drastically overestimate statistical confidence.
  - Incorrect substitution model may result in incorrect phylogenies
- Remedy: Use substitution models based on conserved RNA structure
  (E.g. Schöninger & von Haeseler, Knudsen, Higgs *et al.*)
  SOFTWARE: `phase package` (Paul Higg's lab, Hamilton, CND)

Basic Requirement: Good Structural Models

- Consensus structures of small datasets: `RNAalifold`



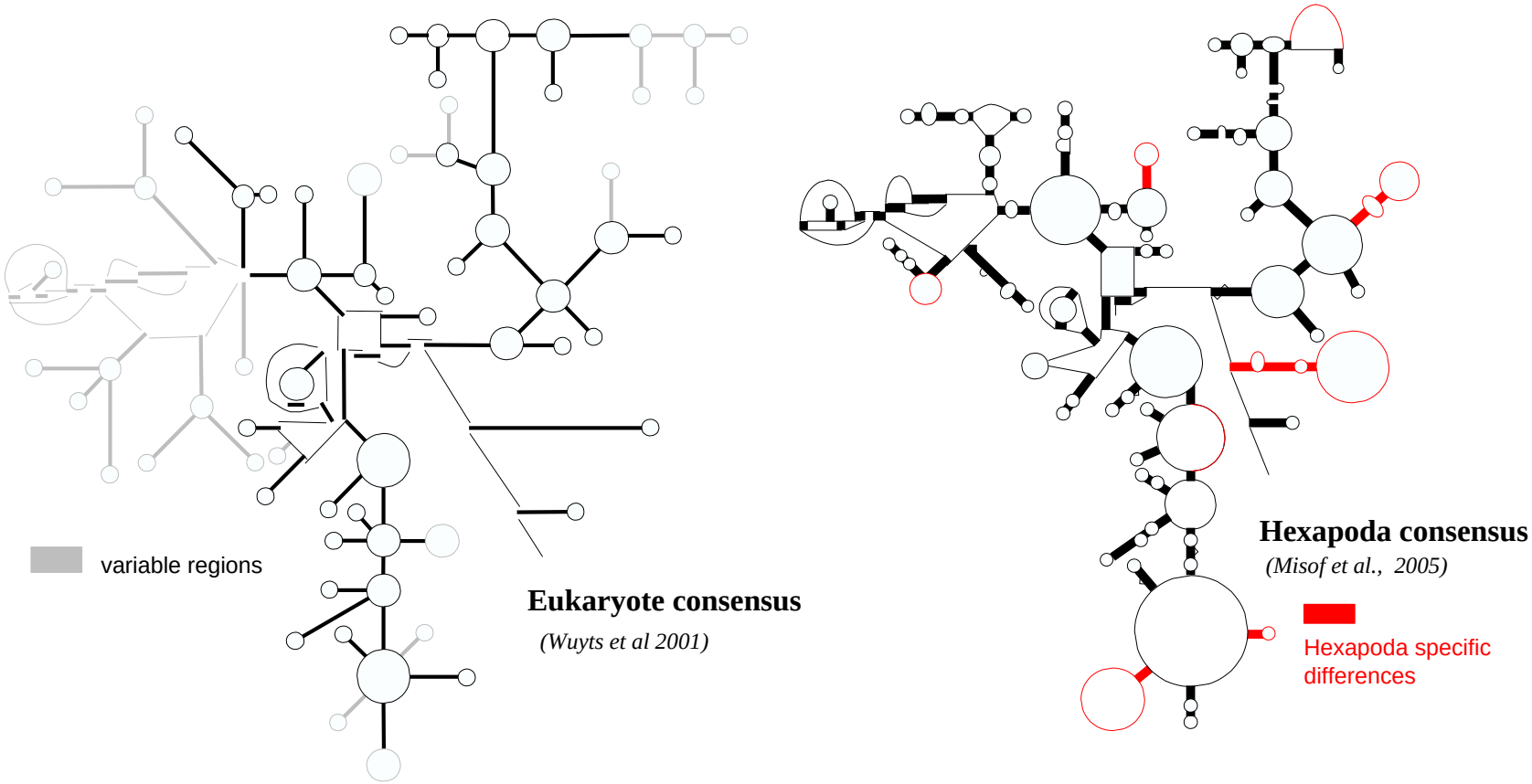
16S rRNA



23S rRNA

Mountain representation of the secondary structure of *E. coli* rRNAs computed using `RNAalifold` from alignments of 5 prokaryotic sequences (16S: *A. globiformis*, *Anabaena.sp*, *A.tumefaciens*, *B.japonicum*, *E.coli*; 23S: *B.subtilis*, *T.thermoph.*, *Pir.marina*, *Rb.sphaero*, *E.coli*)
Green line: predicted single structure;
black line: published consensus structure;
solid colored area: `RNAalifold` prediction

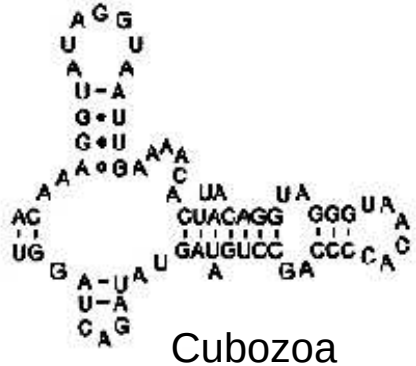
RNA Secondary Structure is Phylogenetically Informative



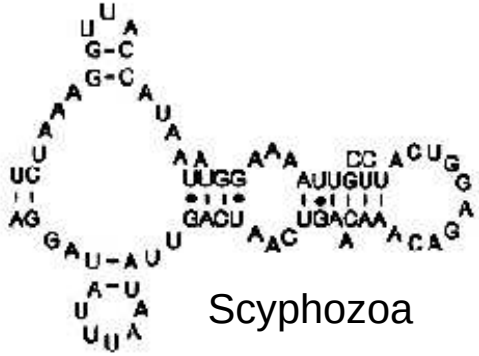
Structural evolution of SSU rRNA

Secondary Structure is Informative at Ancient Nodes

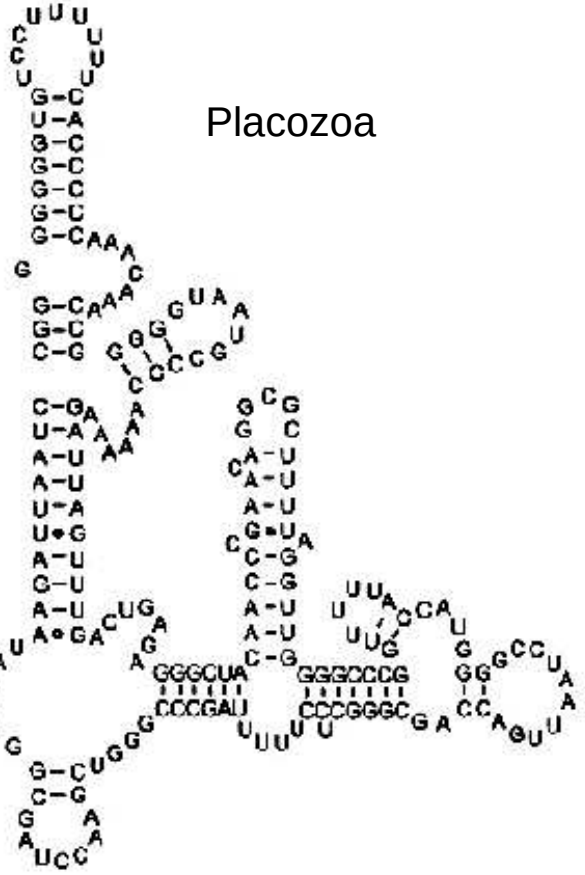
mitochondrial LSU rRNA fragment
(Endler & Schierwater 2003)



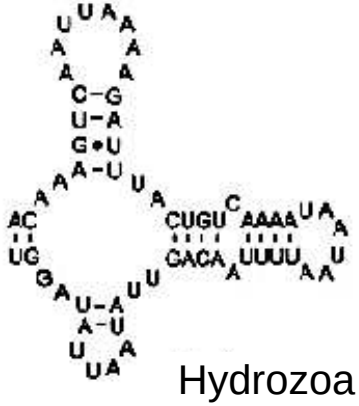
Cubozoa



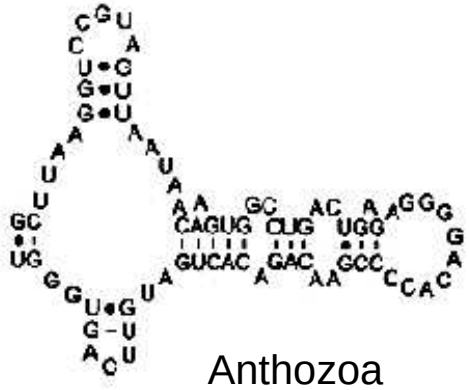
Scyphozoa



Placozoa



Hydrozoa



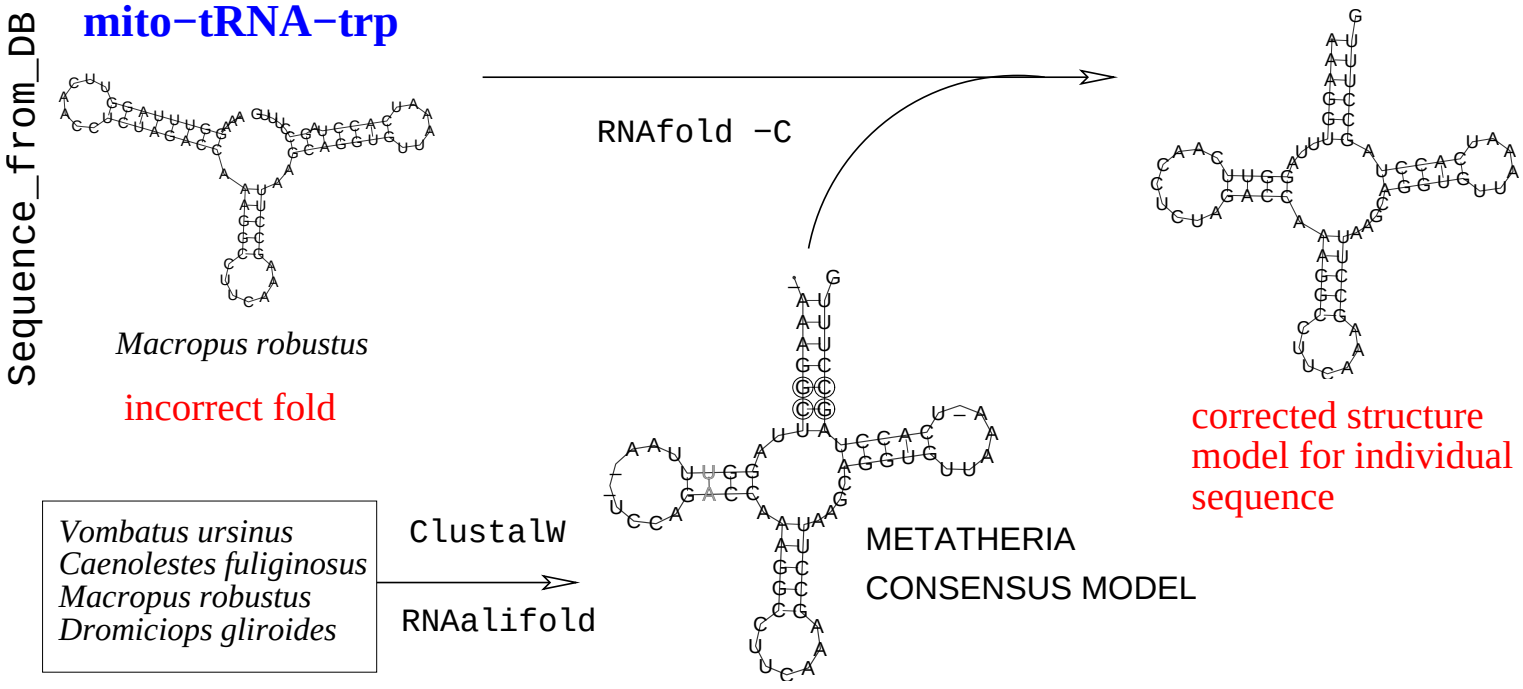
Anthozoa

Research Plan

- Develop Methods for Inferring Phylogenies *directly* from secondary structure models
  RNA secondary structure evolves much slower than sequence, hence it is ideal for resolving ancient nodes.
- Evaluate RNA secondary structure comparison tools such as `RNAforrester`, `RNAdistance`, `MARNA`, etc. for their use in phylogenetics
- Develop methods that can distinguish between uninformative variations and phylogenetically relevant structural differences.

Pipeline for Structure Annotation

Automatic structure annotation of known RNA sequences within undisputed monophyletic groups

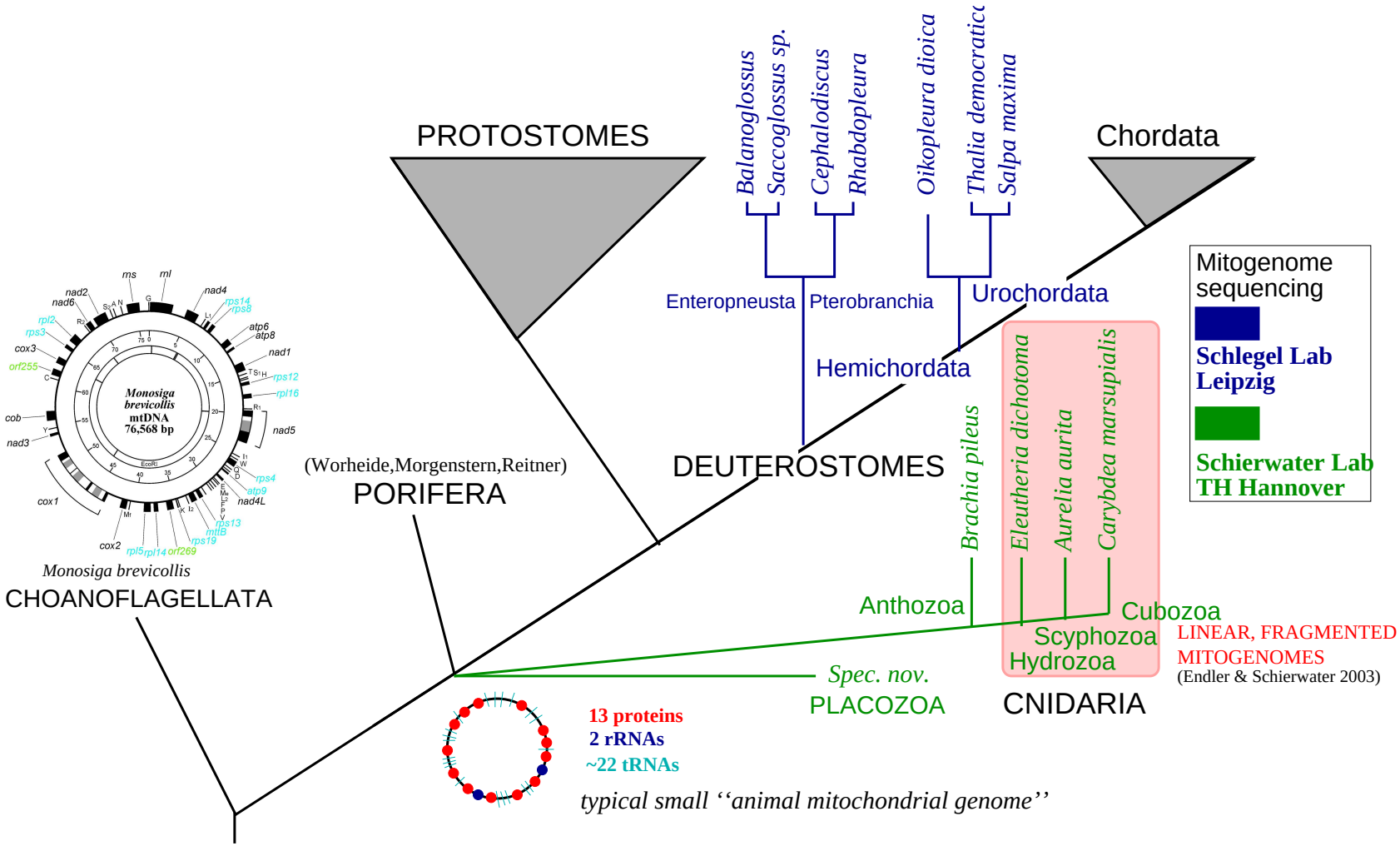


⇒ Roman

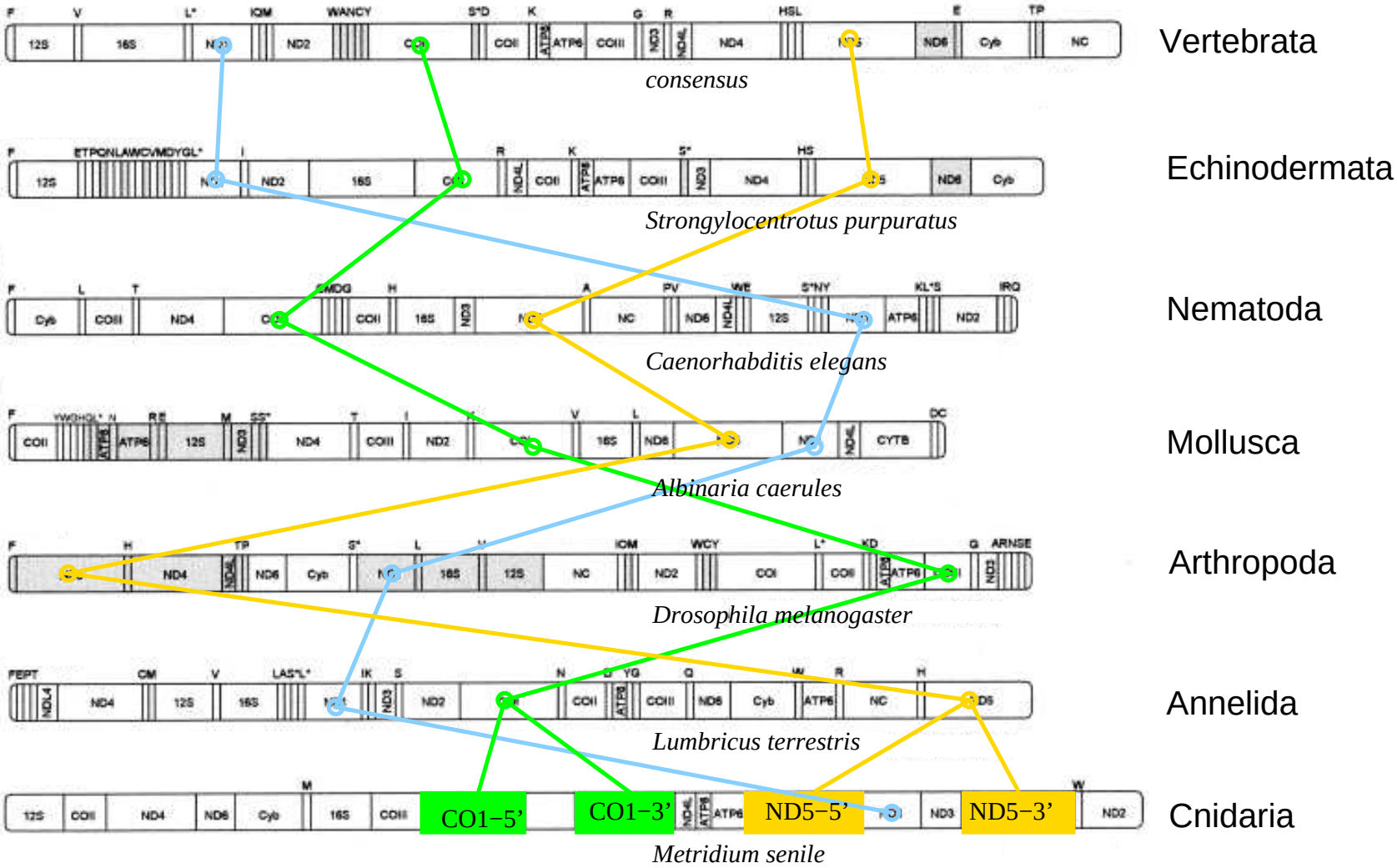
The Plan: Structure-Based Phylogenies

- Structure-based alignments and reliable structure models are the basis to study the evolution of structure itself
- Distance measures for aligned structures are readily available:
  tree alignment distances (Giegerich), tree edit distances (Sankoff, Backofen, `Vienna RNA Package`), profile distances (`Vienna RNA Package`)
  $\Rightarrow$ Distance-based phylogenies
- Structure alignments also allow the development of parsimony based methods
- Reconstruct most basal nodes of the metazoan tree based on secondary structure information

II. Mitochondrial Genome Structure



Rearrangements of Mitochondrial Genomes

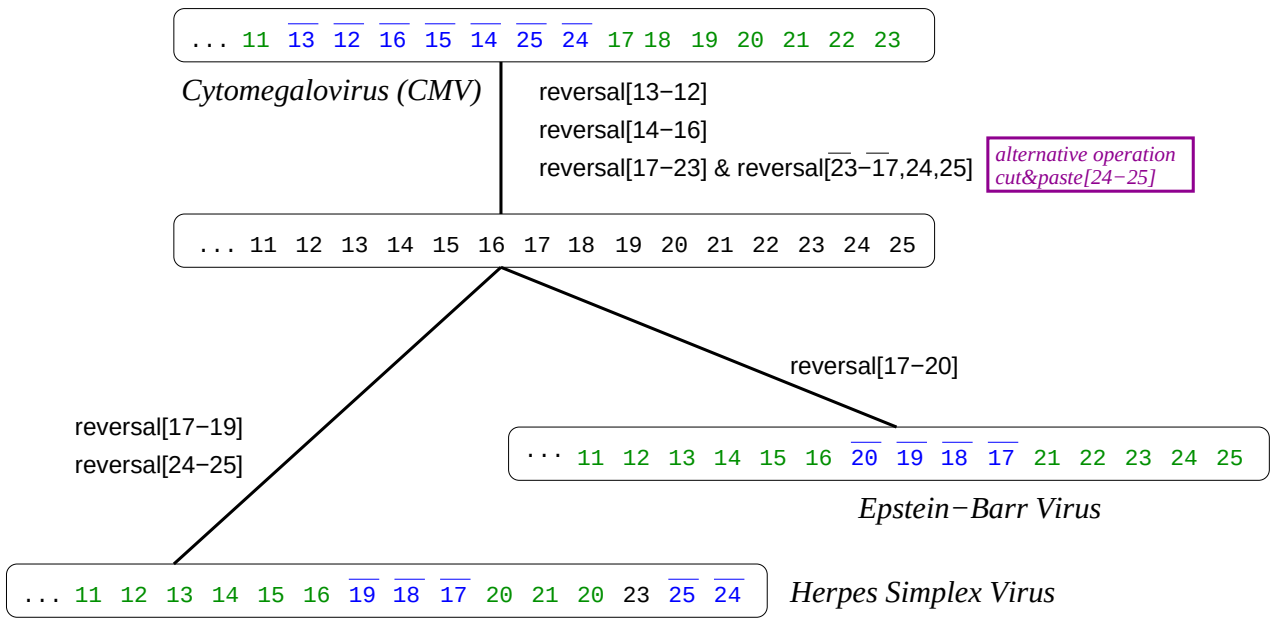


Number of tRNAs seems to vary; variation in protein content possible

The Median Problem

Multiple Genome Rearrangement Problem:
Find a phylogenetic tree T that minimizes the total number of reversal operations necessary to explain the observed genome orders

Already NP complete for 3 species: the Median Problem



adapted from Bourque & Pevzner (2002)

Middendorf & Merkle's Revoluzer Approach

Conserved interval $[a, b]$ in a set of permutations (Bergeron & Stoye 2003)

- either a preceeds b or $\overline{b}$ preceeds $\overline{a}$ in each step
- the unsigned elements between a and b are the same in every step.

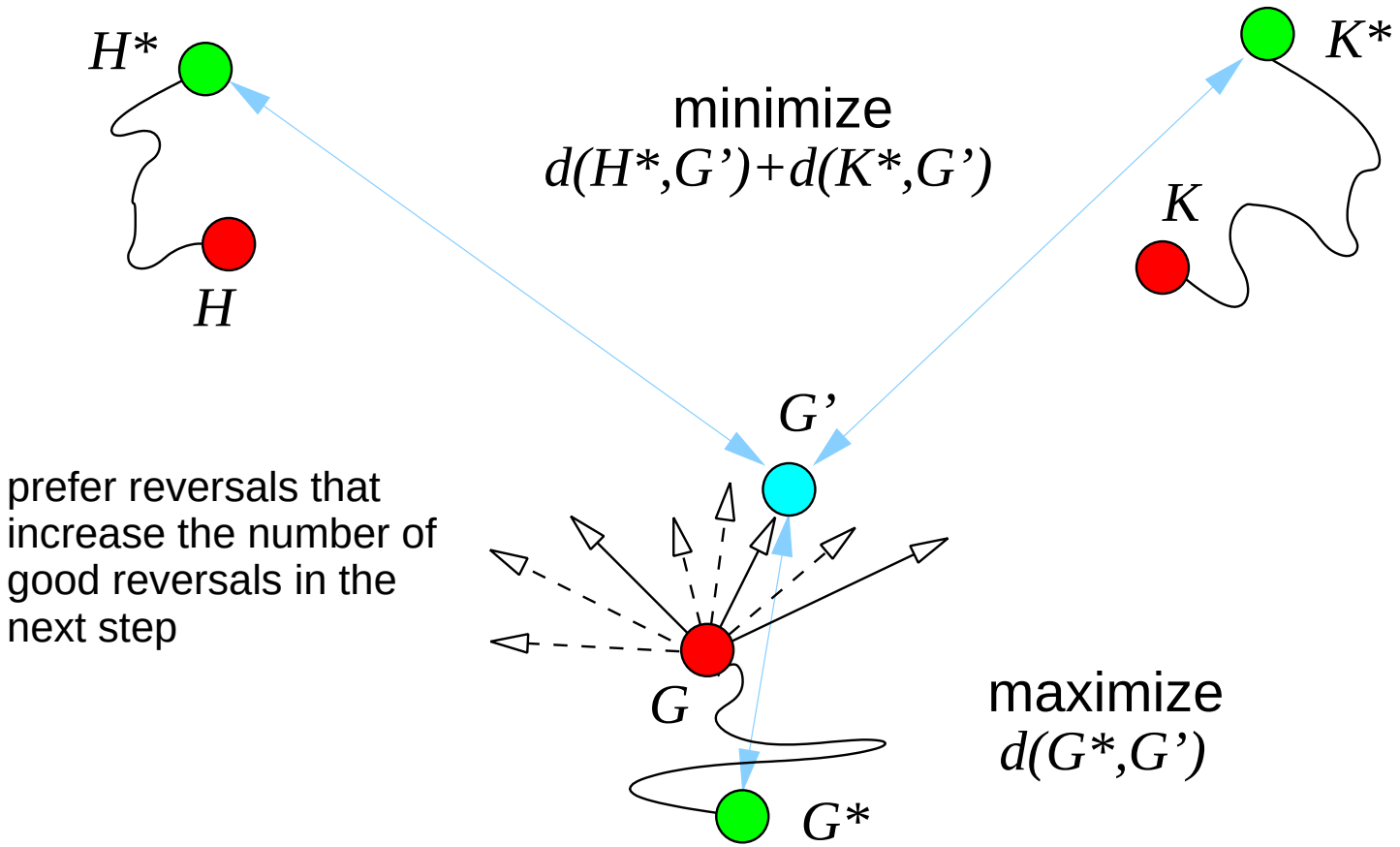
EXAMPLE. From silkworm to locust in 6 steps:

1	2	3	4	5	6	7	8	9	10	11	12	14	13	15	16	17
1	2	3	4	-14	-12	-11	-10	-9	-8	-7	-6	-5	13	15	16	17
1	2	3	4	-14	5	6	7	8	9	10	11	12	13	15	16	17
1	2	3	4	-13	-12	-11	-10	-9	-8	-7	-6	-5	14	15	16	17
1	2	3	5	6	7	8	9	10	11	12	13	-4	14	15	16	17
1	2	3	5	4	-13	-12	-11	-10	-9	-8	-7	-6	14	15	16	17
1	2	3	5	4	6	7	8	9	10	11	12	13	14	15	16	17

1	2	3	4	5	6	7	8	9	10	11	12	14	13	15	16	17
1	2	3	-4	5	6	7	8	9	10	11	12	14	13	15	16	17
1	2	3	-4	-5	6	7	8	9	10	11	12	14	13	15	16	17
1	2	3	5	4	6	7	8	9	10	11	12	14	13	15	16	17
1	2	3	5	4	6	7	8	9	10	11	12	-14	13	15	16	17
1	2	3	5	4	6	7	8	9	10	11	12	-14	-13	15	16	17
1	2	3	5	4	6	7	8	9	10	11	12	13	14	15	16	17

Cycles of elementary intervals (Hannenhalli & Pevzner, 1995)
Nearly all sorting and neutral reverals are on the same cycle.
Use only preserving reversals on cycles $\Rightarrow$ only $\mathcal{O}(n^2)$ candidates

Solving the Median Problem with `Revoluzer`

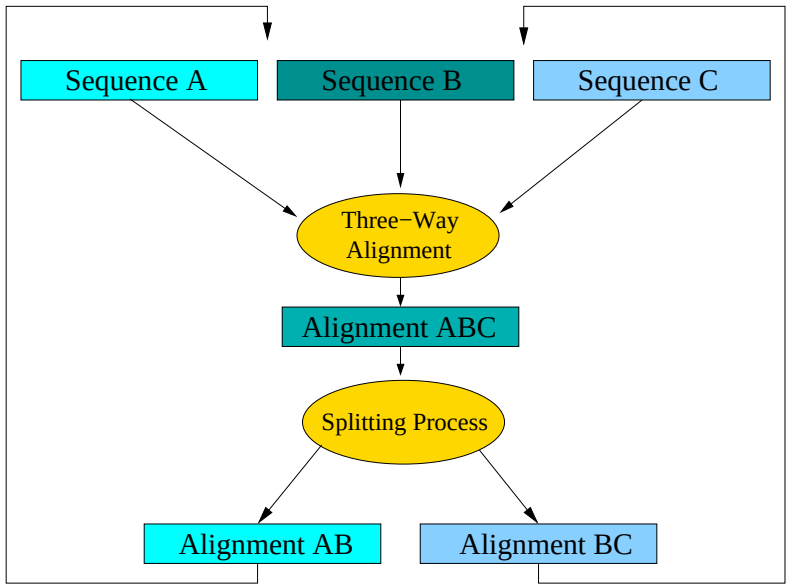


Iterative procedure terminates when $G = H = K$.

If stuck, retracts last step(s) and uses different candidate G'.

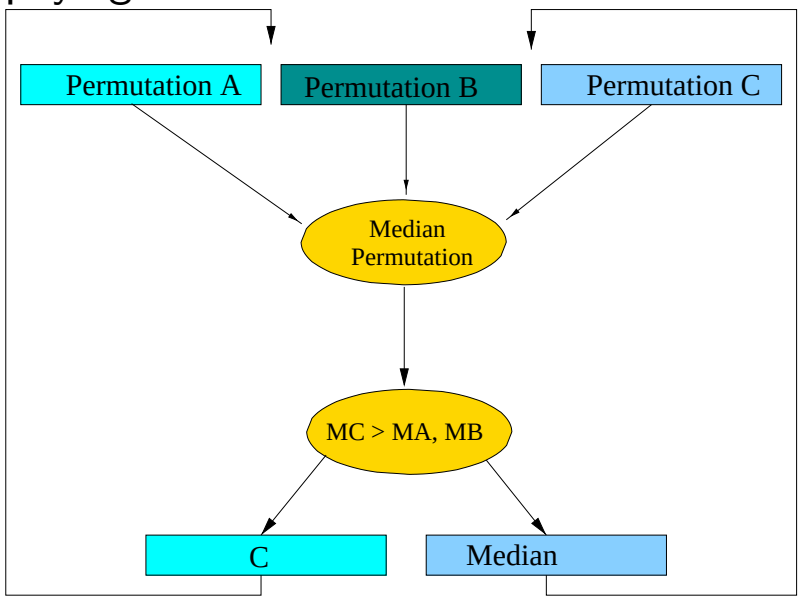
From Medians to Phylogenetic Networks

Generalization of progressive multiple sequence alignment:



Combination of Bryant's `nnet` algorithm with three-way sequence aligments.
Matthias Kruspe's talk

Proposed algorithm for reconstructing phylogenies:

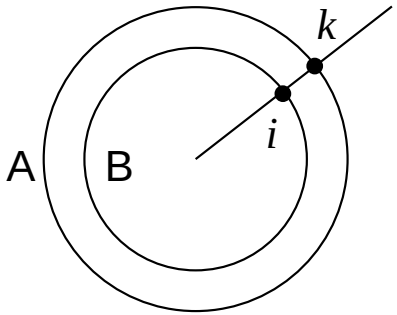


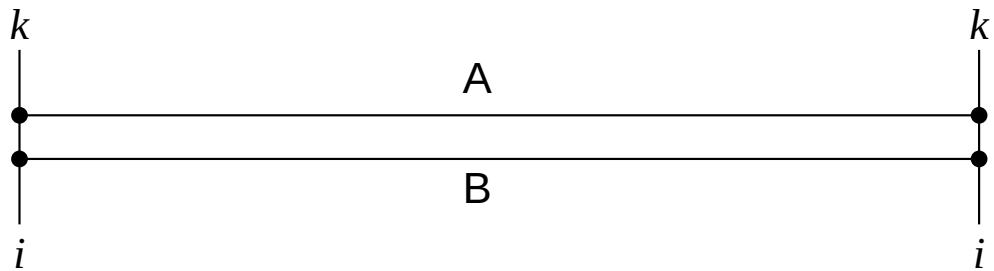
Combination of Bryant's `nnet` algorithm with solution of median problems by `revoluzer`.

- **Advantage:** Self-correcting, i.e., mistakes in early steps are not frozen immediately.
- Detailed comparison with "classical" sequence-based analysis
- Combined analysis: use sequence data to determine supported splits predicted from rearrangement data and *vice versa*.

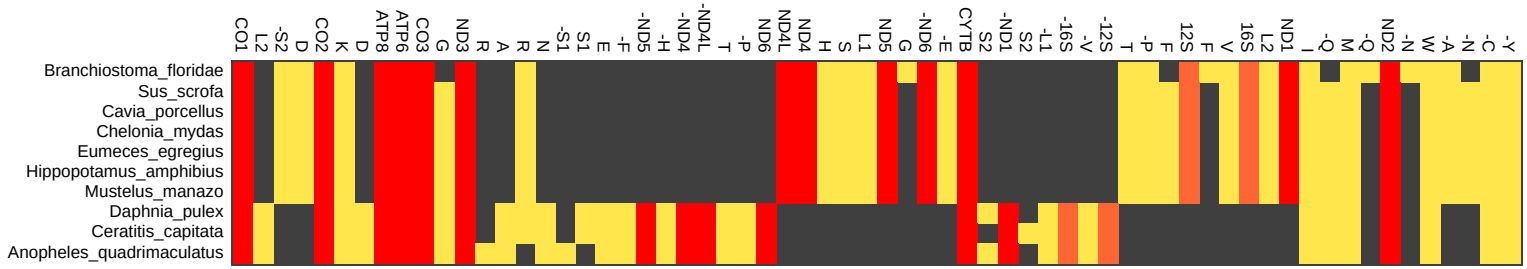
Differences in Content: Circular List Alignments

- Regard mitogenomes as circularly ordered lists of genes (instead of permutations)
- Circular alignments can be reduced to string alignments





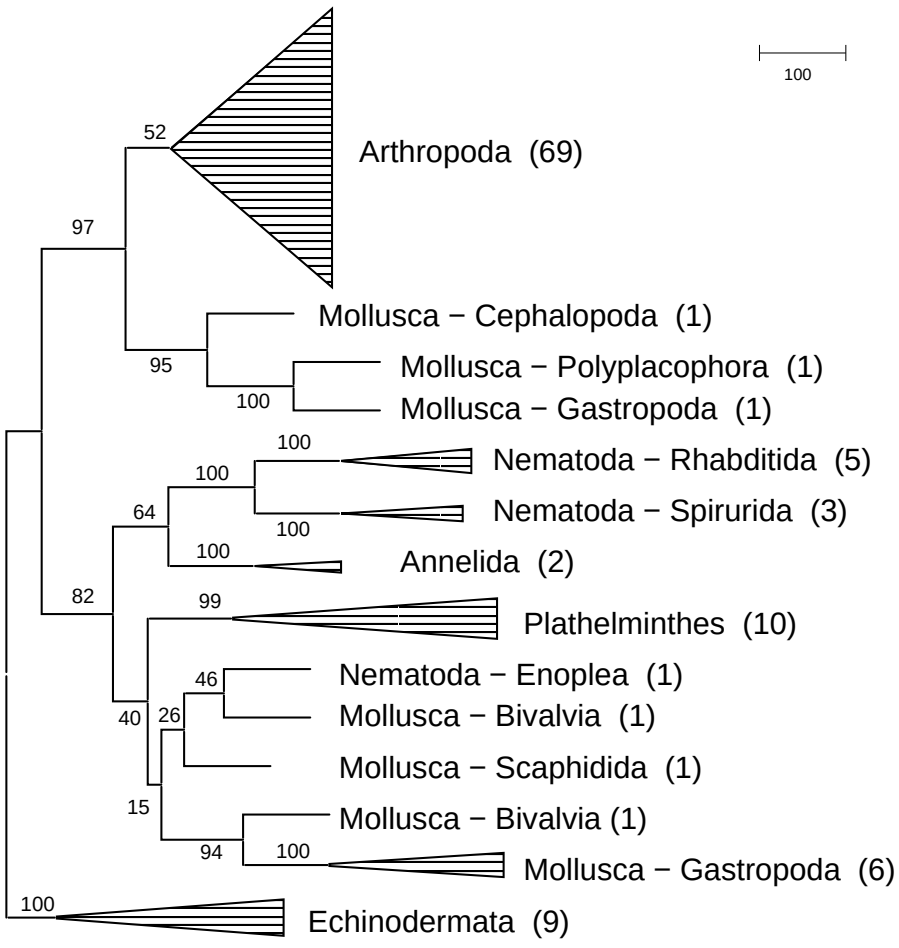
- Convenient representation of results



List alignments: Fried *et al.* J. Chem. Inf. Comput. Sci. 44: 332-338 (2004)

Generalization to circular lists: Fritzsch *et al.*, submitted (2004)

First Results and Difficulties



Features and Results:

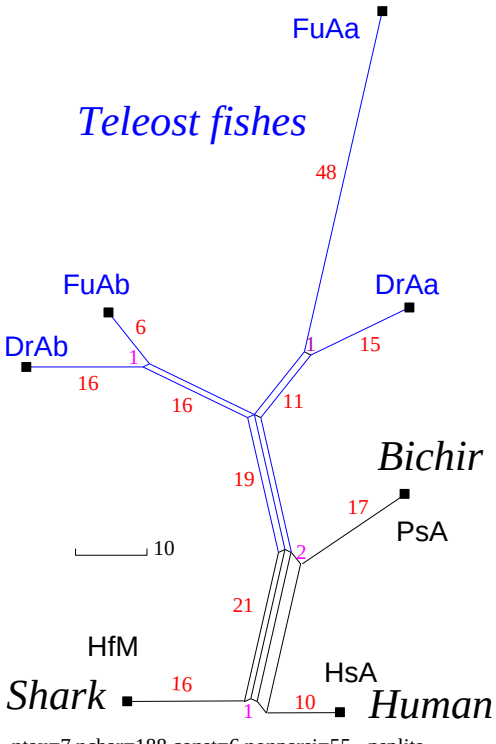
- easily deals with insertions/deletions and missing data
- allows reconstruction of ancestral states by means of "normal" parsimony algorithms
- plausible trees: e.g protostome/deuterostome split and protostome tree shown here

Problems:

- does not exactly recover reversal distances
- high computational cost for exact circular alignments
- cost model for size and content-dependent indels not yet optimized

III. Footprints and Phylogeny

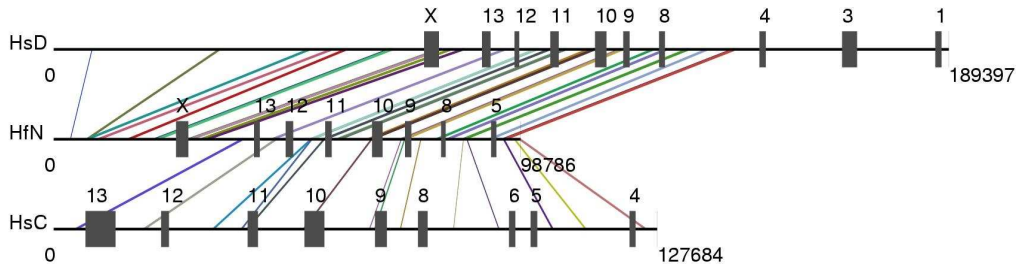
Footprints contains phylogenetic information

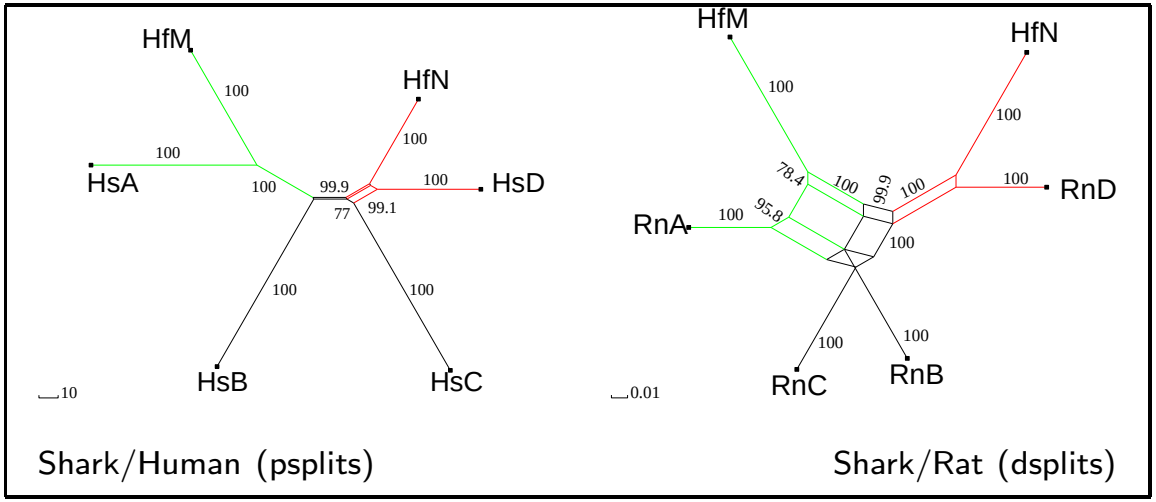


ntax=7 nchar=188 const=6 nonparsi=55 −psplits

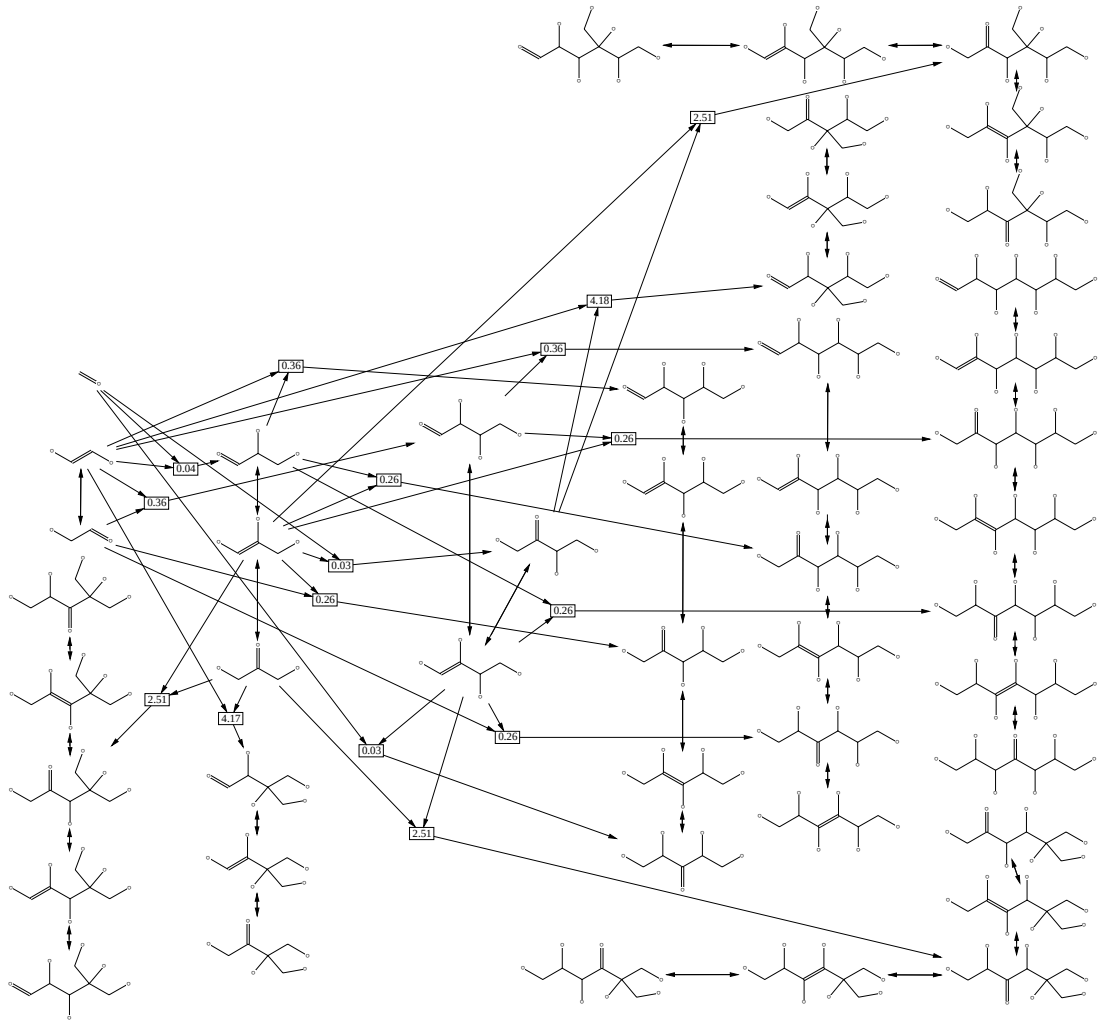
Application:

Identity of the shark *Hox-N* cluster

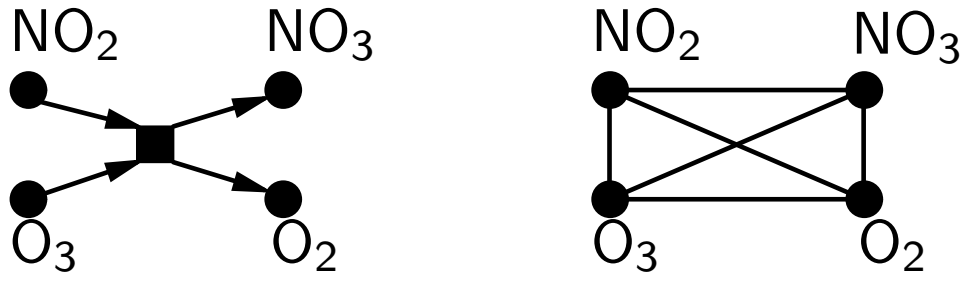




IV. Metabolic Network Structure



Representation of Chemical Reaction Networks



Representations of the reaction $NO_2 + O_3 \rightarrow NO_3 + O_2$ in hypergraph form drawn as the equivalent directed bipartite graph (l.h.s) and as part of a substrate graph (r.h.s).

Set X of reactants. Reaction $E^- \rightarrow E^+$, $E^{\pm} \subseteq X$.
stoichiometric coefficients $n^+_{x,E}$, $n^-_{x,E}$.
stoichiometric matrix $\mathbf{S}$ with entries $\mathbf{S}_{xE} = n^+_{x,E} - n^-_{x,E}$
$\Rightarrow$ Metabolic Flux Analysis

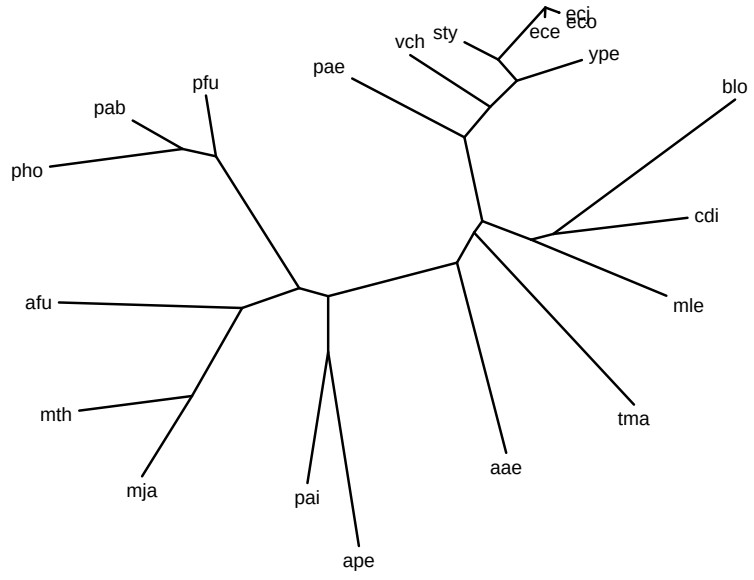
Algebra of Reaction Networks

- Support $\mathrm{supp}\mathcal{E} = \bigcup\{E | E \in \mathcal{E}\}$
- CleanUp $\lfloor \mathfrak{M} \rfloor = (\mathrm{supp}\mathcal{E}, \mathcal{E})$
- Restriction $\mathcal{E}[A] = \left\{E \in \mathcal{E} | A \subseteq (E^+ \cup E^-)\right\}$
  $\mathfrak{M}[A] = \lfloor\, (A, \mathcal{E}[A])\, \rfloor$
  $\mathfrak{M}[\mathcal{E}] = \mathfrak{M}[\mathrm{supp}\mathcal{E}]$
- Union $\mathfrak{M} = \mathfrak{M}' \cup \mathfrak{M}'' = (X' \cup X'', \mathcal{E}' \cup \mathcal{E}'')$
- Intersection $\mathfrak{M} = \mathfrak{M}' \cap \mathfrak{M}'' = \lfloor\, (X' \cap X'', \mathcal{E}' \cap \mathcal{E}'')\, \rfloor$
- Difference $\mathfrak{M} = \mathfrak{M}' \setminus \mathfrak{M}'' = \lfloor\, (\mathrm{supp}(\mathcal{E}' \setminus \mathcal{E}''), \mathcal{E}' \setminus \mathcal{E}'')\, \rfloor$
- Strict Difference $\mathfrak{M} = \mathfrak{M}' \backslash\backslash \mathfrak{M}'' = \lfloor\, (X' \setminus X'', (\mathcal{E}' \setminus \mathcal{E}'')[X' \setminus X''])\, \rfloor$
- Symmetric Difference $\mathfrak{M} = \mathfrak{M}' \triangle \mathfrak{M}'' = \lfloor (\mathfrak{M}' \cup \mathfrak{M}'') \setminus (\mathfrak{M}' \cap \mathfrak{M}'') \rfloor$
- Symmetric Strict Difference $\mathfrak{M} = \mathfrak{M}' \diamondsuit \mathfrak{M}'' \lfloor (\mathfrak{M}' \cup \mathfrak{M}'') \backslash\backslash (\mathfrak{M}' \cap \mathfrak{M}'') \rfloor$

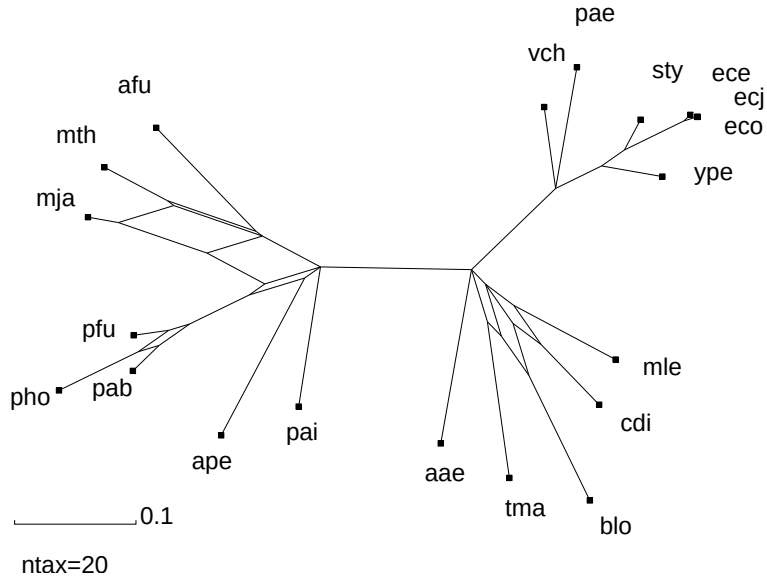
Distance of Networks

$$d(\mathfrak{N}', \mathfrak{N}'') = \frac{\|\mathfrak{N}' \triangle \mathfrak{N}''\|}{\|\mathfrak{N}'\| + \|\mathfrak{N}''\| - \|\mathfrak{N}' \cap \mathfrak{N}''\|}$$

Alternatively, use strong symmetric difference.



Fitch algorithm

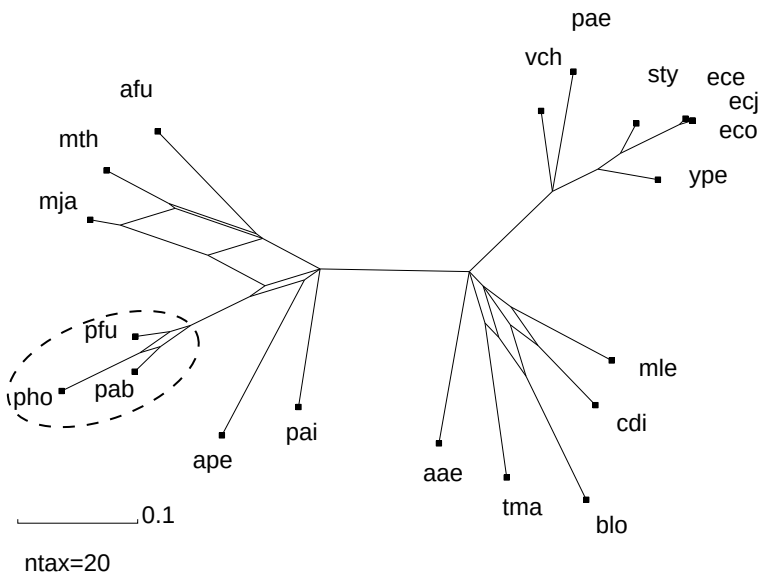


Splits decomposition with Fitch-Margoliash Power 2 distance

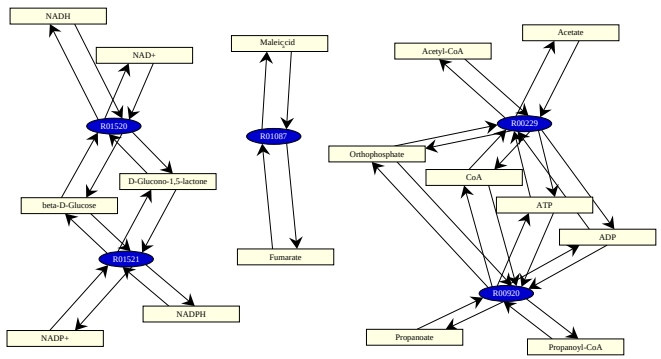
Metabolic Innovations

For a split $\sigma = \{U, \bar{U}\}$ in the tree define

$$\mathfrak{D}(\sigma) = \left(\bigcup_{k \in U} \mathfrak{M}_k \right) \setminus \left(\bigcup_{k \in \bar{U}} \mathfrak{M}_k \right)$$

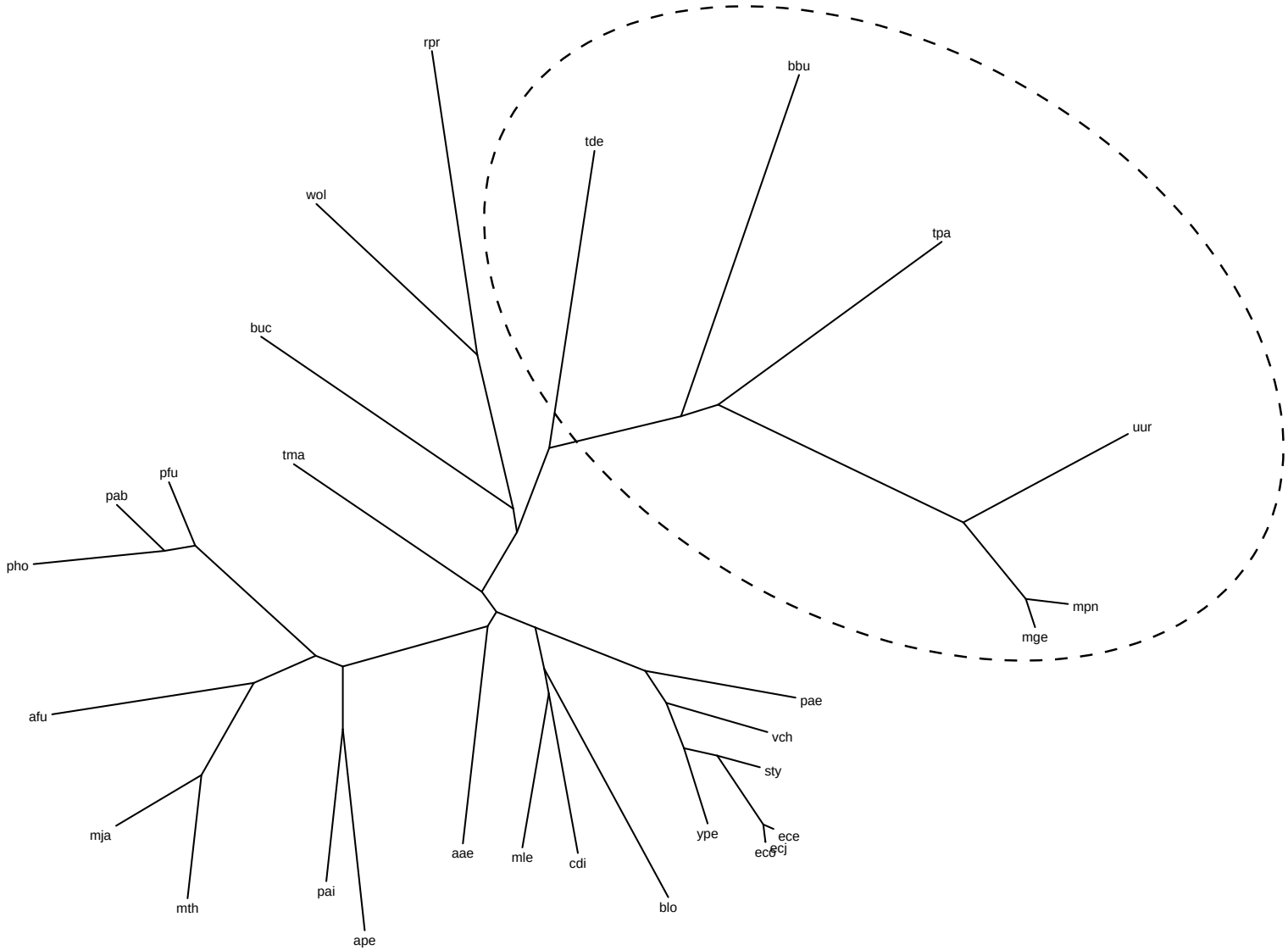


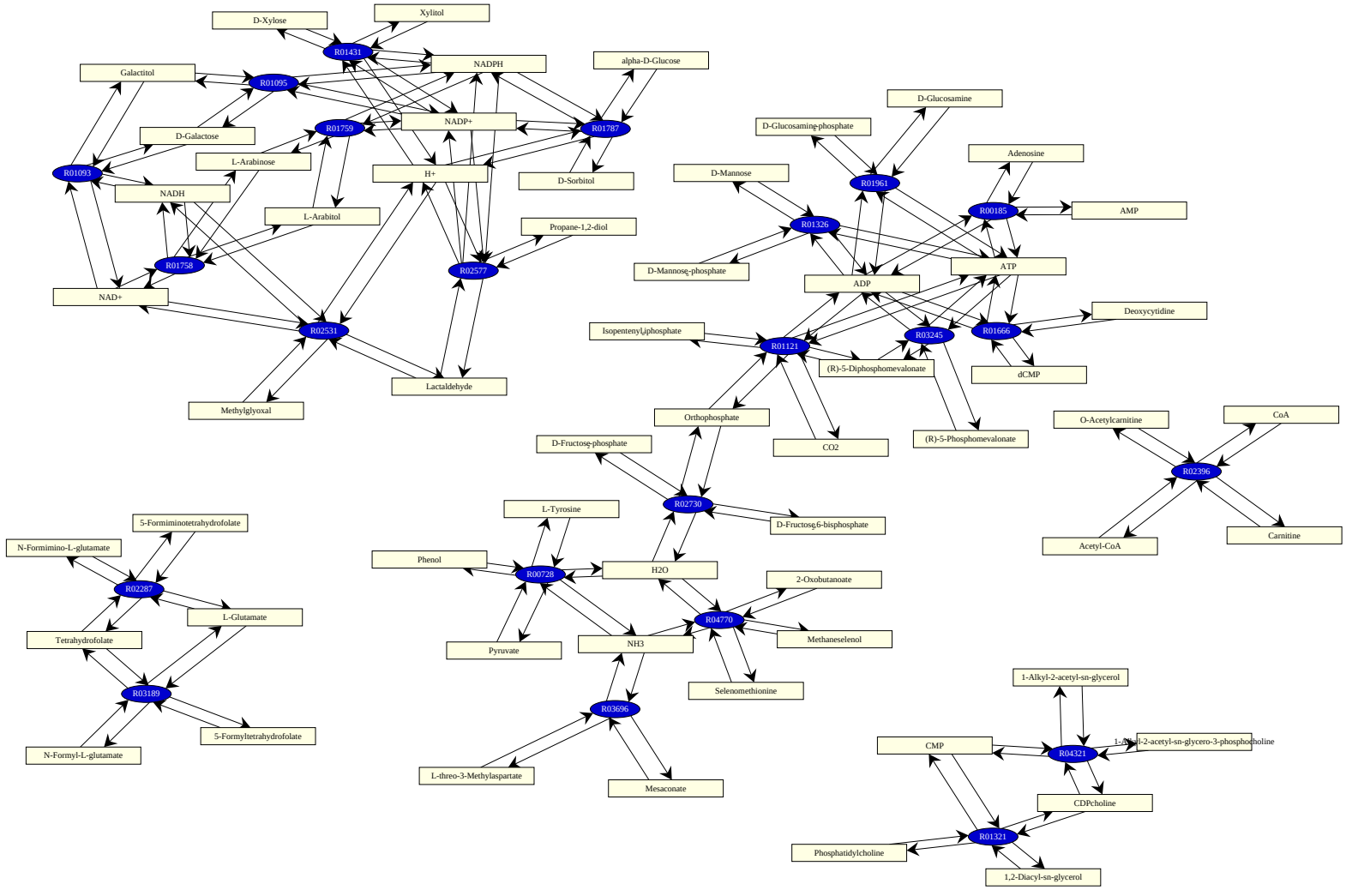
Pyrococcus spp. clade



differential network

Intracellular pathogens





Mollicuts, such as *Mycoplasma*, and *Spirochets*, such as *Treponema*, are grouped together.
These pathogens possess a minimal gene-sets. $\Rightarrow$ metabolic networks are highly optimized and host-dependent.
This set of organism exhibits specific reactions that are absent in the remaining organisms of the phylogeny.

Thanks

- Ivo L. Hofacker, Roman Stocsics (RNA morphologies)
- Matthias Kruspe (Alignments)
- Sonja Prohaska, Claudia Fried, Günter P. Wagner (Footprint evolution)
- Guido Fritzsch, Martin Schlegel, Daniel Merkle, Martin Middendorf (Mitochondrial Genomes)
- Christian V. Forst, Ivo Hofacker, Christoph Flamm (Metabolic Networks)