Strong negative and positive selection can obscure ancestral signal in phylogenetic analysis

Alexander Tchourbanov

University of Wyoming, Laramie

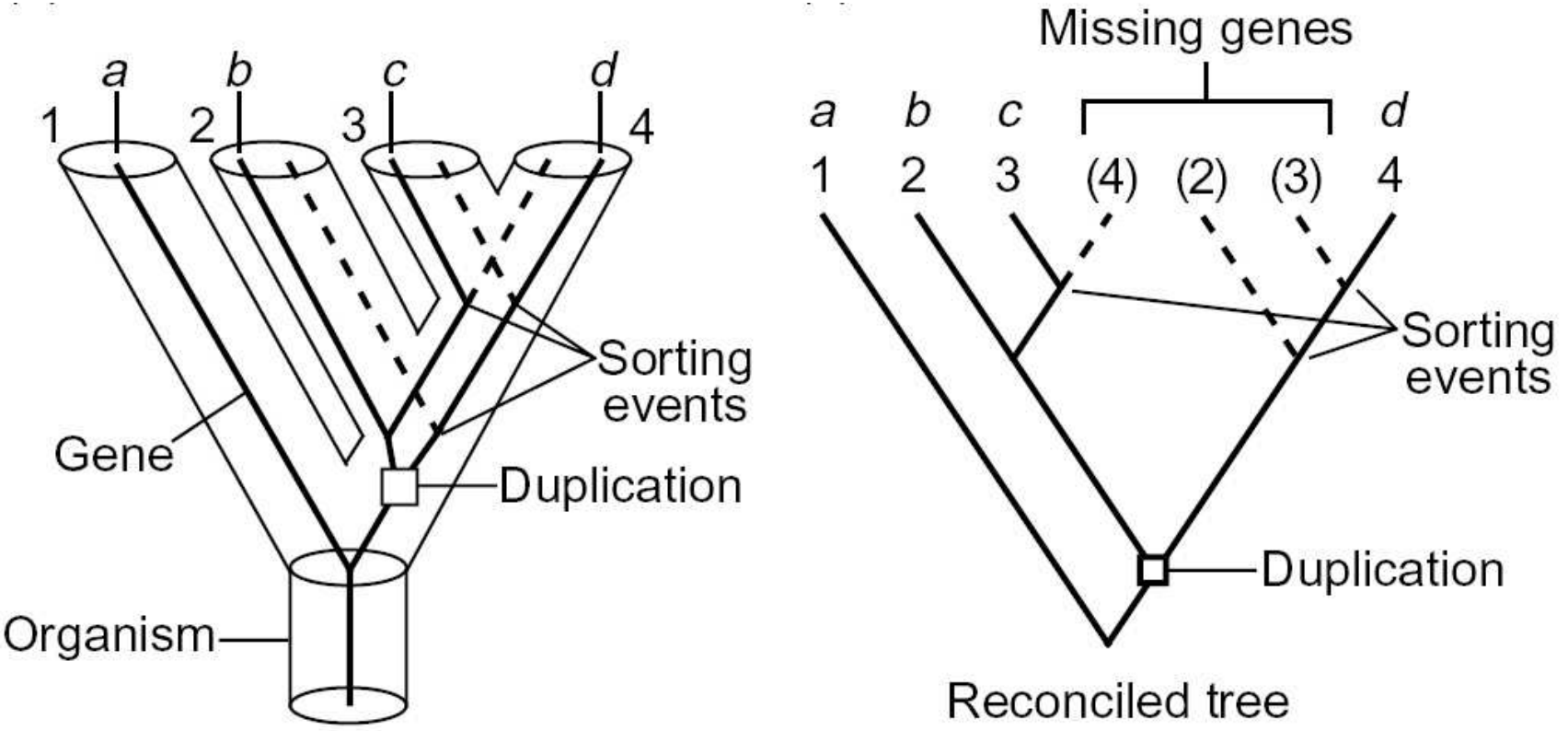
KaKsiTaxi project

- Genomes evolve through the consequence of duplication events followed by evolutionary adjustments.
- Homologous genes are classified as follows:
  - *Orthology*, i.e. homology through speciation
  - *Paralogy*, i.e. homology through replication within organism
- We are interested in effect of Ka/Ks ratio, characterizing the process of positive or negative selection, on fidelity of phylogenetic reconstruction.

Ka/Ks ratio

- In order to detect positive selection between gene sequences we compare, ratio of non-synonymous (Ka) to synonymous substitutions (Ks) is used:
  - $\frac{Ka}{Ks} < 1$ - purifying selection
  - $\frac{Ka}{Ks} = 1$ - neutral drift
  - $\frac{Ka}{Ks} > 1$ - positive selection
- Ks is ancestral component of phylogenetic signal and it saturates over time.
- Ka characterizes functional component of phylogenetic signal.

Reconciled trees



- Gene and taxonomy trees frequently mismatch.
- We postulate number of gene gains and losses to embed the gene tree into the specie tree [1].

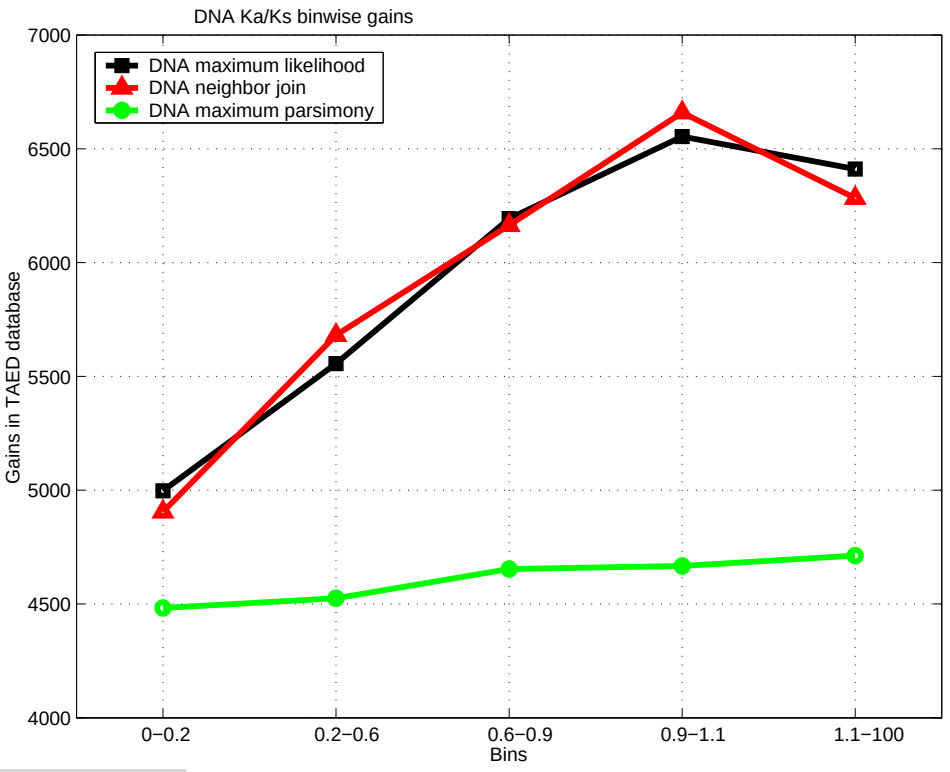
KaKsiTaxi framework

- Run `MUSCLE` to realign gene family sequences obtained from TAED database and then get rid of gaps.
- Using modified Nei-Gojobori technique we find position-specific $\frac{Ka}{Ks}$ ratios.
- Form the five bins according to $\frac{Ka}{Ks}$ ratio and check if they are have more than 3 sequences of more than 9 codons.
- For each bin we run `Phylip` tools:
  - Parsimony (both DNA/Protein)
  - Neighbor Join (both DNA/Protein)
  - Maximum likelihood (both DNA/Protein)

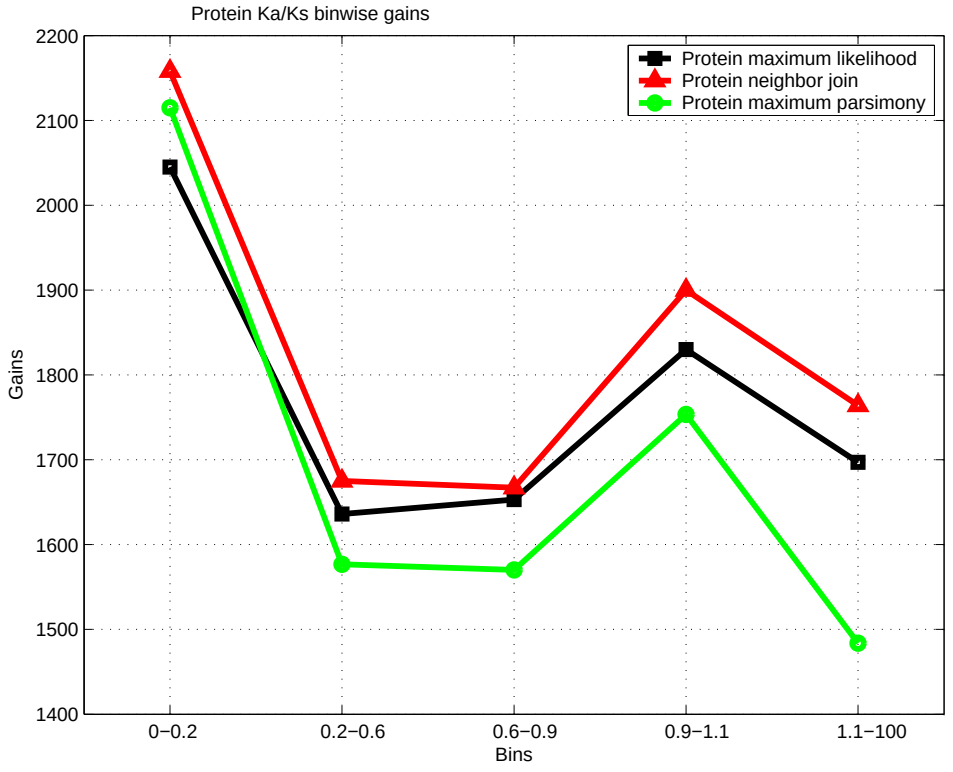
KaKsiTaxi framework (1)

- We collect resulting trees. Special weighting scheme is used to process equiprobable trees from maximum Parsimony method.
- We transform the trees so that we get rid of branches and set posterior to 1.0.
- Run `SoftParseMap` on precompiled index file and collect gains and losses as reconciled with recent NCBI taxonomy file.
- Output cumulative statistics.

Resulting gains

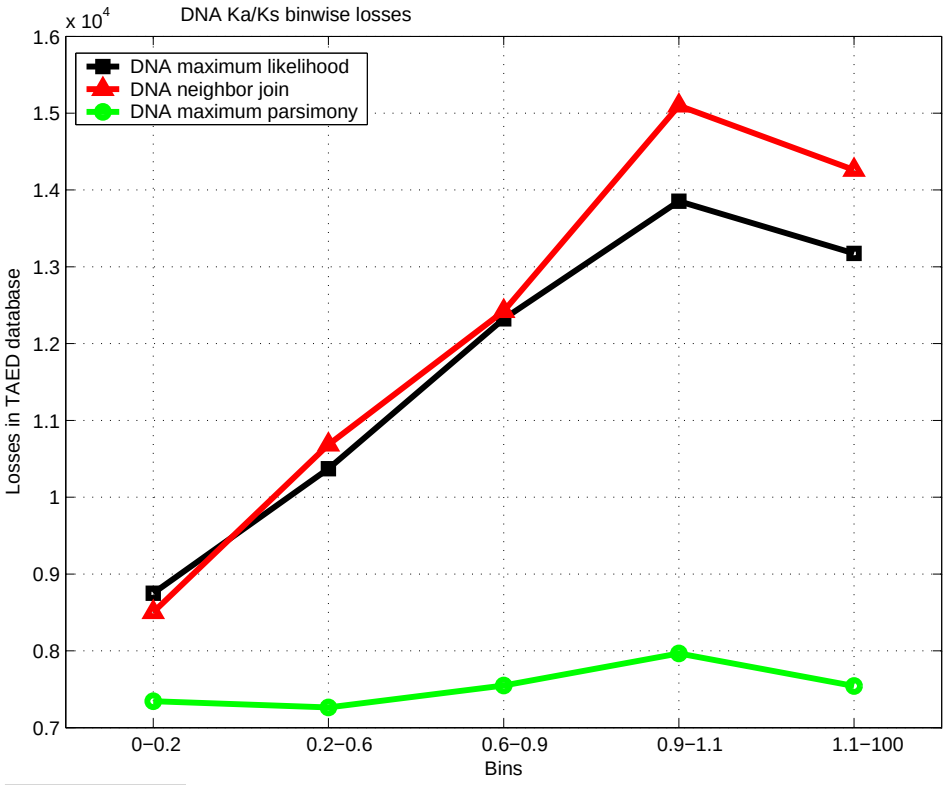


DNA gains

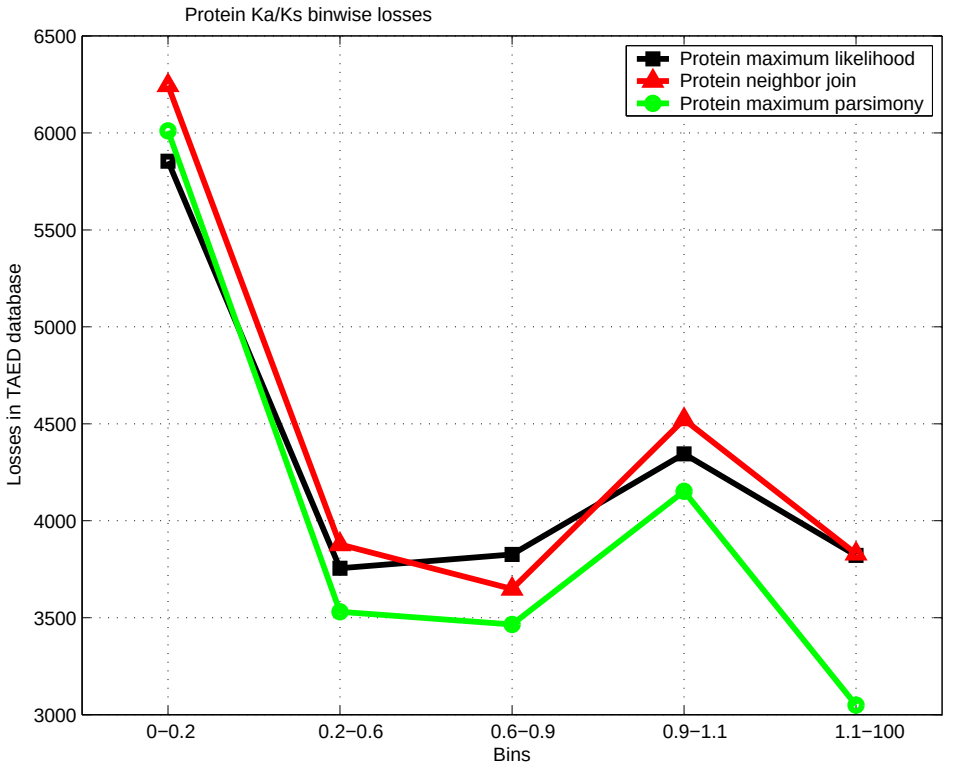


Protein gains

Resulting losses



DNA losses



Protein losses

References

[1] P. Roderic and M. Charleston, *Trees within trees: Phylogeny and historical associations*, Tree **13** (1998), no. 9.