Mapping Strongly Discordant Regions on the Genome Using Hidden Markov Models

Ali Osman Berk Şapcı[1], Shayesteh Arasti[2], Siavash Mirarab[3,*]

{[1]: Bioinformatics and Systems Biology Graduate Program, [2]: Department of Computer Science and Engineering, [3]: Department of Electrical and Computer Engineering}

Motivation

- Finding strongly discordant regions under **incomplete lineage sorting (ILS)**: inferring accurate phylogenies, detecting regions driven by non-ILS events...

What causes non-ILS discordances?

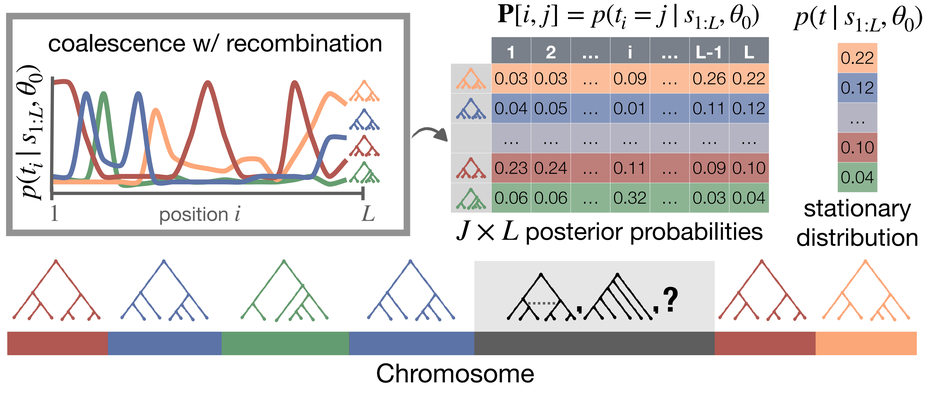
Real biological causes:
- Hybridization and introgression
- Suppression of recombination
- Selection and selective sweep

Artefacts:
- Alignment and assembly error
- Paralogy mistaken as homology
- Gene tree bias and noise

- Methods for detecting changes in locus tree and non-tree like scenarios: CoalHMM [1] , Phylonet-HMM [2], ERICA [7], etc.
- Not scalable, only limited to certain scenarios and a small number of taxa

Problem

- $s_{1:L}$: a sequence of length L, t_i: a topology at index i, resp.
- **Goal**: segmentation into binary states $(y_1, \dots y_L)$, $y_i \in \{0,1\}$:
  - 0: **null state**, generated under coalescence w/ recombination
  - 1: **alternative state**, exhibiting strong non-ILS discordance



- **Approach:** Use the posterior probabilities, **P**, to detect outlier segments!
- **Challenge:** Computing $p(t_i \mid s_{1:L}, \theta_0)$ is difficult even w/ Markovian assumptions
  - Quadratic in the number of topologies, J, which is $O(m!!)$ for m taxa

Summarizing the topology distribution across loci:

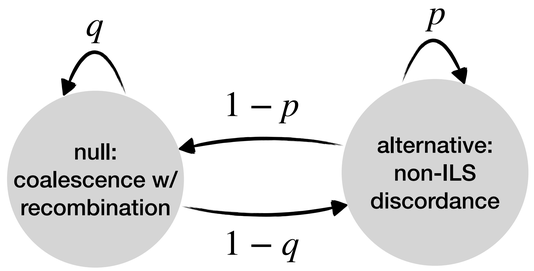
Idea: Given a species tree T and a target branch e: compute a summary statistic $x = g_e(t)$ without explicitly enumerating all the topologies s.t.

$g_e(t_i) \mid y_i = 0 \sim p(x_i \mid T) = p(g_e(t) \mid s_{1:L}, \theta_0)$ and $g_e(t_i) \mid y_i = 1 \sim p(x_i \mid \theta_1)$.

Assumptions: The stationary distribution exists and $p(t \mid T)$ approximates it.

Hidden Markov model formulation

An HMM w/ standard initial and transition distributions with Bayesian priors on parameters:



$$(q, 1-q) \sim \mathrm{Dir}(\beta + \kappa + \rho, \beta)$$
$$(p, 1-p) \sim \mathrm{Dir}(\beta + \rho, \kappa + \beta)$$

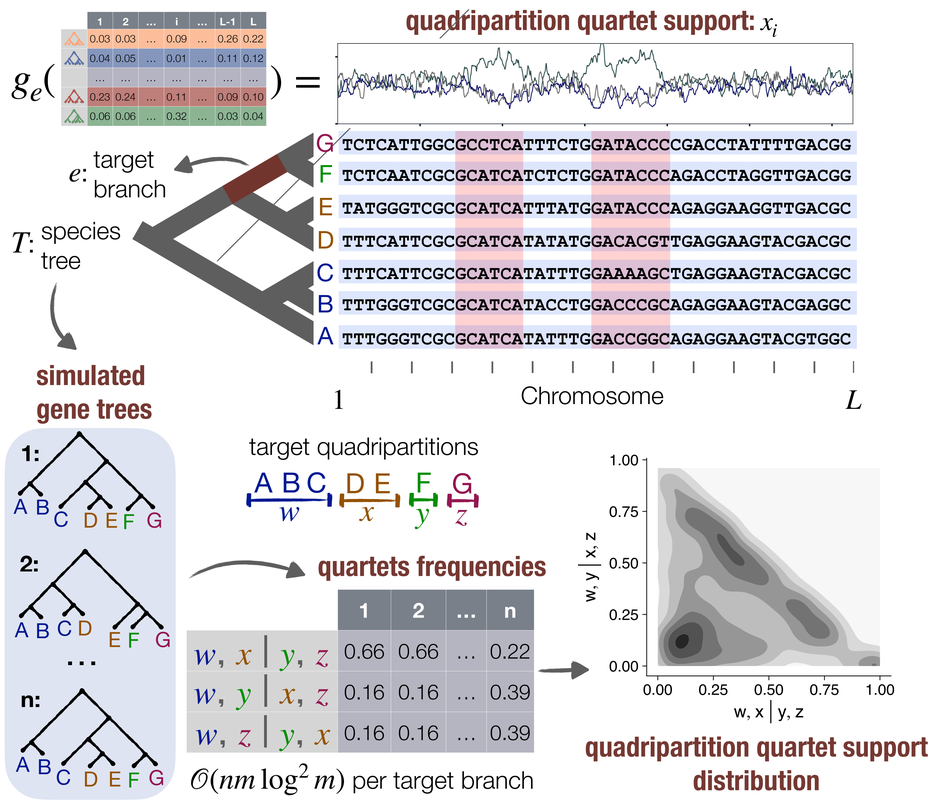
- **Initial distribution:** Categorical dist. with an uninformative Dirichlet conjugate prior
- **Transition matrix:** $\pi = \begin{pmatrix} q & 1-q \\ 1-p & p \end{pmatrix}$ reflects our assumptions with flexible & robust priors

Transition matrix hyperparameters

- **Challenges:**
  - noisy observations, high sampling rate
  - slightly off $p(x_i \mid T)$ (e.g., varying rates)
- **stickiness** (κ) favors continuous segments, reduces transitions btw. different states
- **sparsity** (ρ) boosts transitions to the null state and makes states imbalanced

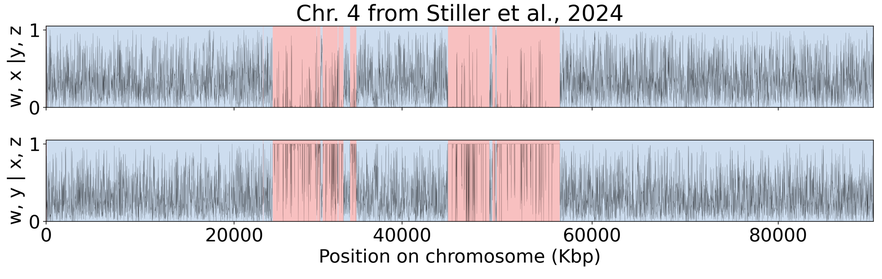
Quadripartition quartet support emissions

- $g_e(t_i) = x_i$: bivariate QQSs (w.r.t. e) for the gene tree sampled at locus i
- Obtain $p(x \mid T)$ for the null state via gene tree simulations using SimPhy [3]
- Discretize QQSs into quantile bins as the family of distribution is unknown
- Infer $p(x \mid \theta_1) = \mathrm{Cat}(x \mid \theta)$ using the same bins and a Dirichlet conjugate prior



Detecting a case of recombination suppression

- An avian dataset w/ 363 taxa [5], documented recombination suppression [4]
- Two discordant segments mapped to chicken chr. 4 (6262 gene trees sampled)



Discussion

- Using an appropriate summary statistic (e.g. QQS) as a proxy for posterior probabilities
- Detecting discordant regions efficiently
- Priors are robust against noisy emissions w/ high variance, and generalize well (no tuning)
- **Future work:** using CASTER [6] site scores, distinguishing between causes of discordances (i.e., multiple alternative states)

References: