Module 4: Phylogenetics

5th June 2019

Bulk data

- A common approach is to treat the presence absence of mutations in samples as characters and apply standard phylogenetic methods.
- An alternative approach is to perform deconvolution of samples and reconstruct trees based on clones.
    - One way is to do deconvolution first.
    - Alternative is to jointly build the tree and do deconvolution.

Questions

- What issues does performing classical phylogenetics pose?
- What are the benefits of each deconvolution approach?

Single cell data

- Single cell data maps well onto phylogenetic problems.
  - Cells are species or individuals in a population.
- scWGS data is sparse in coverage (0.01x-1x)
  - Low probability of having reads covering SNVs
- CNVs are easy to detect but hard to model the evolution.
  - Main issue is convergence and overlapping events.
- Change-points associated with CNVs are a possible character.
  - Provided bins are small enough change-points should be distinct.

Discussion
What are the challenges of using change-points?

Probabilistic phylogenetic methods

Problem setup

- Observation matrix X with M rows corresponding to species and N columns corresponding to features (characters).
- We want to infer the evolutionary tree relating the species $\tau = (E, V)$.
- We may also have branch lengths Λ.
- Trees can be rooted or unrooted.

Question
Should we use rooted or unrooted trees for cancer phylogenetics?

Transition probabilities

- To define a probabilistic phylogenetic model we need to define the probability of moving from character i to j along a branch.
- The traditional approach is to define a a rate matrix Q where Q_{ij} is the instantaneous rate of transition from state i to j.
- The transition matrix is then $P = \exp(Qt)$ where t is the branch length.
- If there are no branch lengths we can define P directly.
- Given the transition probabilities we would then like to compute the probability of the data on the tree.

Notation

- x_v - value of the character at leaf node v
- y_v - value of the character at internal node v
- $L(\tau)$ - leaf nodes of τ
- $I(\tau)$ - internal nodes excluding root
- $\tau(v)$ - subtree rooted at node v
- $\rho(v)$ - parent of node v
- $\gamma(v)$ - set of children of v
- r - root node

Tree probability

- If we have all the internal node labels, **y**,the likelihood is given by

$$\begin{aligned} p(\mathbf{x}|\tau, P, \mathbf{y}) &= \prod_{v \in L(\tau)} P_{y_{\rho(v)} x_v} \prod_{v \in I(\tau)} P_{y_{\rho(v)} y_v} \\ &= \prod_{v \in \gamma(r)} P_{y_r y_v} p(\mathbf{x}|\tau(v), P, \mathbf{y}) \end{aligned}$$

- In general we do not known **y** so we would like to marginalise

$$p(\mathbf{x}|\tau, P) = \sum_{\mathbf{y}} p(\mathbf{x}|\tau, P, \mathbf{y})$$

Felsenstein pruning

- The marginalisation can be performed efficiently using a recursive algorithm similar to the FB.

$$\begin{aligned} \alpha_v(i) &= \prod_{u \in \gamma(v)} \sum_{y_u = j} P_{ij} \alpha_u(j) \\ \alpha_v(i) &= \begin{cases} 1 & i = x_v \\ 0 & i \neq x_v \end{cases} \end{aligned}$$

- The first line is the internal node recursion and the second line is the initial condition at the leafs.
- To apply this algorithm we start at the leafs at work backwards towards the root.
- The algorithm allows to evaluate the likelihood in $\mathcal{O}(|V||S|^2)$ as opposed to $\mathcal{O}(|V|^{|S|})$.

Bayesian phylogenetic models

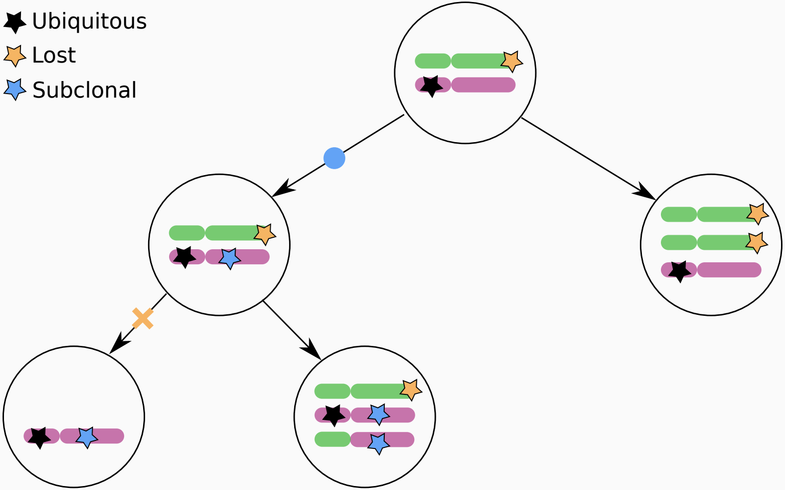
- Thus far we have a likelihood.
- To specify a Bayesian models we will need priors.
- These include $p(\tau)$ tree prior, $p(\Lambda)$ branch length prior (if applicable), $p(\theta)$ transition matrix parameter prior.
- Inference is generally hard for phylogenetic methods.
  - There are $\mathcal{O}(n!)$ trees for n species.
  - This is discrete state space so we cannot use gradient descent.
  - Brute force enumeration often the only approach.
- MCMC is not much harder than MAP!

Probabilistic model for mutation loss

Problem

- We consider the problem of building phylogenies from bulk sequence data.
- We will assume we have M samples from a patient with N SNVs.
- We will treat the samples as species and consider the presence/absence of SNVs as characters.
- We face two challenges:
    - Samples are mixtures of clones.
    - Mutations may be loss due to CNVs.

Mutation history



★ Ubiquitous

☆ Lost

☆ Subclonal

Model assumptions

- We make the following assumptions
  1. Mutations originate at most once on the tree.
  2. Mutations can be lost after they are acquired.
  3. Mutations evolve independently i.e. our tree probability decomposes as the product of mutations.
- We will assume that we cannot perfectly observe the presence/absence of mutations.
  - Our input data will then be probabilities.
- More exactly the probability an SNV is clonally present in a sample.
  - We need this because of sequence coverage and clonal mixtures.

Pre-processing

- Before constructing the tree we will need to compute the probability a mutation is clonally present.
- Let c_b denote the number of mutated copies, c_t total number of copies, t the tumour content and ϵ the sequencing error.
- The probability of observing a read with the mutation is

$$r = \begin{cases} \frac{c_b t}{2(1-t)+c_t t} & c_b > 0 \\ \epsilon & c_b = 0 \end{cases}$$

- Our allelic count data is modelled as Binomial and we obtain the probability of presence by summing all $c_b > 0$.
    - We use CNV data as in module 2 to inform the copy number.

Tree probability

- We use a modified version of the pruning algorithm.
    - We need the modification because the assumption of single origin creates dependencies in the tree.
- Let π_l be the probability a mutation is lost along a branch, $p(z_v = 0|\cdot)$ the probability a mutation is absent at node v and $p(z_v = 1|\cdot)$ the probability it is present.
- We will compute $Q(v, \tau)$, the probability the mutation is present at node v given all possible combinations of losses on the sub-tree rooted at v.

$$Q(v,\tau) \quad = \quad \begin{cases} \pi_l p(z_v = 0|\cdot) + (1-\pi_l) p(z_v = 1|\cdot) & v \in L(\tau) \\ \pi_l \prod_{u \in L(\tau(v))} p(z_u = 0|\cdot) + (1-\pi_l) \prod_{u \in \gamma(i)} Q(u,\tau) & v \notin L(\tau) \end{cases}$$

Tree probability

- Let w be the node where a mutation originates.
- Then we have

$$p(\mathbf{x}|\tau, w) \quad = \quad Q(w, \tau) \prod_{v \in L(\tau) \setminus L(w)} p(z_v = 0|\cdot)$$

- Now we do not know w, so we place a Uniform prior, $p(w)$, and marginalise.

$$p(\mathbf{x}|\tau) \quad = \quad \sum_{w \in V(\tau)} p(\mathbf{x}|\tau, w) p(w)$$

- Thus far we have focused on a single mutation. The probability for all mutations is

$$p(X|\tau) \quad = \quad \prod_{n=1}^{N} p(\mathbf{x}_n|\tau)$$

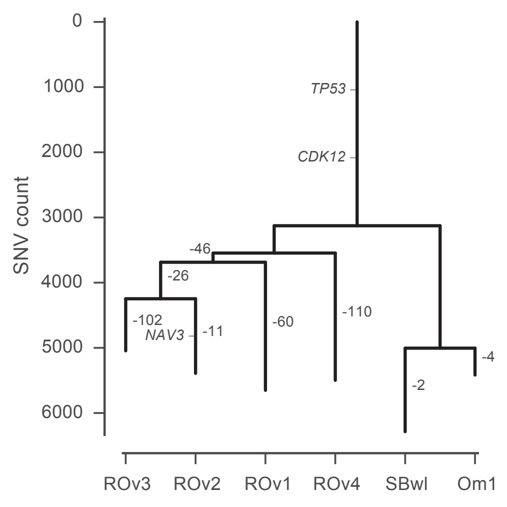
Inference

- We assume the number of samples is small.
    - In this case we can enumerate all trees and evaluate their probabilities to compute the MAP estimator.
- We also optimise the probability of loss on each tree.
- This leads to MAP estimators $\hat{\tau}, \hat{\pi}_l$.
- For more than 10 samples this approach is not practical.
    - We would then need to turn to MCMC methods.

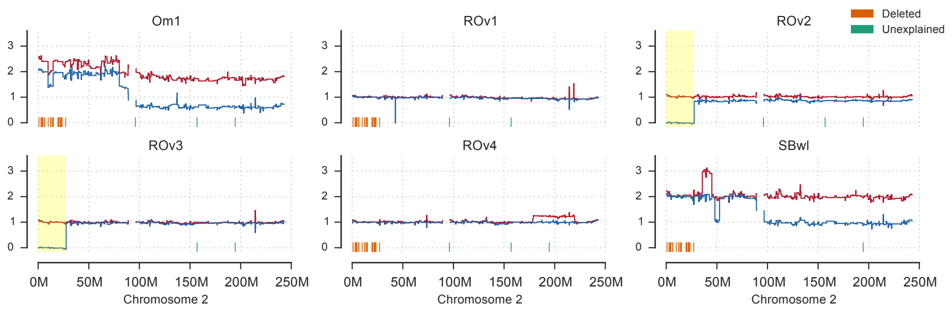
Inferring origin, presence and loss of mutations

- The pruning recursion can be modified by switching summation for maximisation to find the most probable labelling of the internal node.
    - This is the same as the relationship between FM and Viterbi for HMMs
- Using this approach we can label the presence absence of mutations at each node in the tree (subject to single origin constraint).
- Once we have the presence/absence labelling we can compute origin points and loss events.

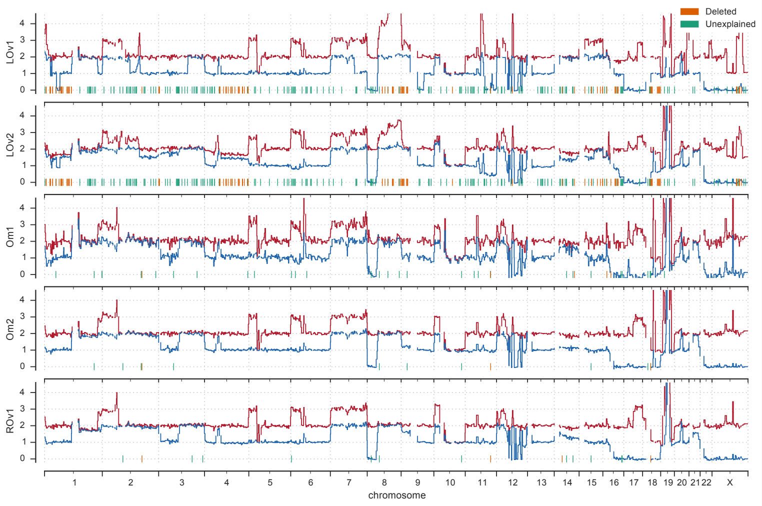
Model output



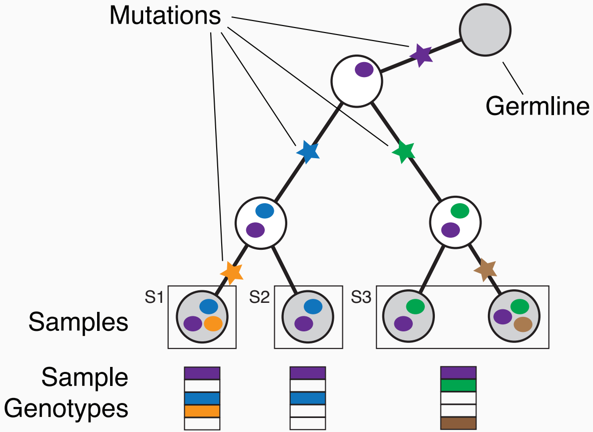
Model validation



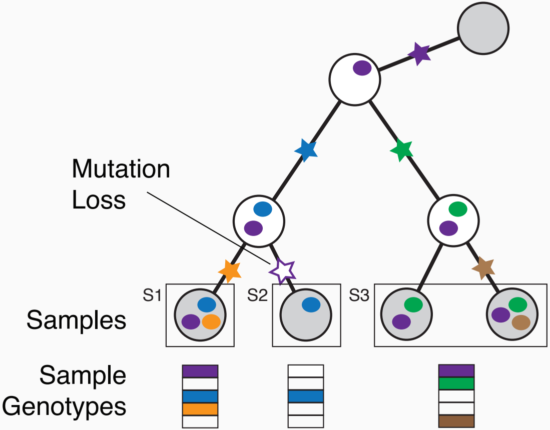
Model failures



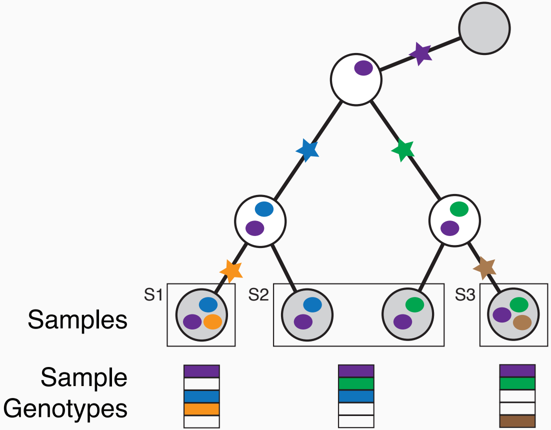
Why we fail



Why we fail



Why we fail


S1
S2
S3
Samples
Sample
Genotypes