

Calculating the marginal likelihood of a local phylogeny

Thomas Mailund

February 13, 2007

I describe how to compute the marginal likelihood for local phylogenies in Blossoc. I then briefly describe how we with this can test hypotheses such as the number of risk-affecting mutations on a tree and I describe a possible extension of the setting to model the phenotype being determined by the genotypes (rather than considering each chromosome independent), integrating over the disease model.

Introduction

In Blossoc, Mailund et al.¹, we perform association mapping by building local phylogenies for each marker in a dataset, using neighbouring SNPs, assuming an infinite sites model of mutations, and using perfect-phylogeny algorithms to efficiently compute these trees.

Once a local phylogeny is constructed, it is scored according to how well the tree helps explain the phenotype. Each tree, T , defines a set of models for the data in the following way: for each subset of nodes in the tree, $\{n_i\}$, we can assign a "risk factor", θ_i , such that any leaf has risk θ_i of being affected and $(1 - \theta_i)$ of being unaffected, if n_i is the closest ancestor to the leaf in the set $\{n_i\}$. The biological interpretation of this is, that on the edge leading to node n_i , there has occurred a mutation that affects the disease risk of any descendants of n_i . In each such set $\{n_i\}$ we include the root of the tree, modelling the wild-type risk. In the following, we denote by S_i the set of leaves with n_i as their closest ancestor from $\{n_i\}$. These sets have, of course, the properties $S_i \cap S_j = \emptyset$ for $i \neq j$ and $\cup_i S_i$ is the set of leaves of T .

Assuming independence between leaves, conditional on the node set $\{n_i\}$ and risks, the likelihood of $\Theta = (\{n_i\}, \{\theta_i\})$ simply becomes

$$L(\Theta, T) = \prod_i \theta_i^{A_i} (1 - \theta_i)^{U_i} \quad (1)$$

where A_i denotes the number of affected individuals in S_i and U_i denotes the number of unaffected individuals in S_i .

Ideally, we wish to consider the set of nodes, $\{n_i\}$ and the risk-parameters, $\{\theta_i\}$, nuisance parameters and score just the tree¹

$$L(T) = \int L(\Theta, T) P(\Theta | T) d\Theta \quad (2)$$

¹For initial scans of the data, we are only interested in locating high-scoring loci, the actual parameters used for scoring the trees are not of interest. When the high-scoring trees are later analysed, these parameters becomes, of course, of interest.

From this we could compare the tree model with the null model

$$L_0 = \int \theta^A (1 - \theta)^U d\theta \quad (3)$$

where A is the total number of affected individuals and U the total number of unaffected individuals to test for evidence of association.

Rather than doing this, however, we have calculated the maximal joint likelihood from (1) penalised with different cost functions from model selection criteria such as Akaike's Information Criteria (AIC) and Bayesian Information Criteria (BIC). In the following, I describe how the desired marginal likelihood, (2), can be computed.

Computing the marginal likelihood

Our goal is to compute the likelihood of the local phylogeny, T , $L(T) = P(D|T)$ where D is the observed data, in this case the phenotype status on the leaves of T . This likelihood is given as

$$L(T) = P(D|T) = \int P(D, \Theta | T) d\Theta = \int P(D|\Theta, T) P(\Theta|T) d\Theta \quad (4)$$

where $P(D|\Theta, T)$ is just the joint likelihood $L(\Theta, T)$ from (2), giving us

$$L(T) = \int P(\Theta|T) \prod_i \theta_i^{A_i} (1 - \theta_i)^{U_i} d\Theta \quad (5)$$

If we now assume that risk-factors from individual n_i s are mutually independent, independent of T and independent of the set of the set $\{n_i\}$, with the same prior $\pi(\theta)$, this simplifies to

$$L(T) = \int P(\Theta|T) \prod_i \theta_i^{A_i} (1 - \theta_i)^{U_i} d\Theta \quad (6)$$

$$= \int P(\{n_i\}|T) P(\{\theta_i\}|T) \prod_i \theta_i^{A_i} (1 - \theta_i)^{U_i} d\Theta \quad (7)$$

$$= \sum_{\{n_i\}} P(\{n_i\}|T) \prod_i \int \pi(\theta_i) \theta_i^{A_i} (1 - \theta_i)^{U_i} d\theta_i \quad (8)$$

where, if we as in Waldron et al.² choose a beta distribution for $\pi(\theta)$ the entire integral just becomes a Beta function, i.e. if $\pi(\theta) = \frac{\theta^{\alpha-1} (1-\theta)^{\beta-1}}{B(\alpha, \beta)}$ the integral becomes $B(\alpha + A_i, \beta + U_i) / B(\alpha, \beta)$ and we obtain

$$L(T) = \sum_{\{n_i\}} P(\{n_i\}|T) \prod_i \frac{B(\alpha + A_i, \beta + U_i)}{B(\alpha, \beta)} \quad (9)$$

As for the prior on sets of nodes we assign risk-factors to, we can choose an uninformative prior and let $P(\{n_i\} | T) = 1/|\mathcal{N}|$ where \mathcal{N} denotes the number of such sub-sets of nodes $\mathcal{N} = 2^n$ where n is the number of nodes in T excluding the root (which will always be part of $\{n_i\}$). Alternatively, and probably more realistically, we can put a prior on the size of the set of nodes, $\psi(|\{n_i\}|)$ and set

$$P(\{n_i\} | T) = \frac{\psi(|\{n_i\}|)}{\binom{n}{|\{n_i\}|}}. \quad (10)$$

In practical terms, this could restrict the set of likely models to those with, say, at most three mutations on the tree, with highest probability for one node (which would by necessity be the root, so this would essentially be the null model) and lowest for four nodes (the root plus three “mutant” nodes). If $\psi(m) = 0$ for $m > 4$ (or some other small constant) this will also significantly speed up calculating the sum in (9).

We might also want to put a prior on the tree-depth of the nodes in $\{n_i\}$, since by the way trees are constructed in Blossoc, we tend to trust the topology of the tree nearer to the root more than the topology further from the root. But this is probably dealt with more cleanly by simply restricting the depth of the inferred phylogenies.

Model testing

With this setup, it also becomes straight forward to test different models against each other: testing the null-model against the tree model is obviously just considering the Bayes factor $L(T)/L_0$, but just as trivially we can from (9) compute the likelihood for the tree restricted to certain number of mutations and e.g. test the hypothesis “two mutations” against “one mutation” as the Bayes factor of $L(T, |\{n_i\}| = 3)$ vs. $L(T, |\{n_i\}| = 2)$.

Extension to genotypes

In the above, the disease status of each leaf has been considered independent of other leaves, conditional on Θ . This, however, is not realistic when considering diploid individuals: here each individual has two chromosomes—and thus two leaves in the tree—and the phenotype is associated with the *individual*, not his individual chromosomes. So, to more accurately model the disease status, we should instead associate a risk factor, θ_{ij} to each *genotype*, ij where genotype ij is defined to be the leaves i, j where n_i is the closest ancestor of i in the set $\{n_i\}$ and n_j is the closest ancestor of j in the set $\{n_i\}$.

If we, for ease of computation, assume independent risk factors, we can again just follow the ideas in Waldron et al.² and (9) and get

$$L(T) = \sum_{\{n_i\}} P(\{n_i\} | T) \left(\prod_i \frac{B(\alpha + A_{ii}, \beta + U_{ii})}{B(\alpha, \beta)} \right) \left(\prod_i \prod_{j>i} \frac{B(\alpha + A_{ij}, \beta + U_{ij})}{B(\alpha, \beta)} \right) \quad (11)$$

where now A_{ij} and U_{ij} denotes the number of affected and unaffected individuals, respectively, with genotype ij .

References

1. T. Mailund, S. Besenbacher, and M. Schierup. Whole genome association mapping by incompatibilities and local perfect phylogenies. *BMC Bioinformatics*, 7(454), 2006.
2. E. R. B. Waldron, J. C. Whittaker, and D. J. Balding. Fine mapping of disease genes via haplotype clustering. *Genet Epidemiol*, 30(2):170–179, 2006.