Abstract

The movement of cancer cells within a tumor has important implications for tumor growth and evolution, and for intratumor heterogeneity, which can impact our ability to tailor effective treatments to a cancer's genetic composition. While even small levels of migration within a tumor can have significant consequences, the rate at which cells move around is very difficult to estimate, either directly or indirectly. Single-cell sequencing, when coupled with precise information about the spatial location of those cells within a cancer, opens the possibility of applying the methods of phylogeographic reconstruction to the estimation of intratumor migration rates by cancer cells. I constructed an algorithm that employs a hidden Markov model to model phenotypes with different migration rates, and that uses expectation maximization to identify the likeliest locations for coalescent events in a given phylogeny. When applied to data from a real tumor, the algorithm again repeatably converged on the same parameter estimate. Application of this algorithm to further data sets, the integration of a ML framework for phylogenetic reconstruction, and further exploration of these results are warranted.

Resources used

All of the code used was written specifically for this project, with a few exceptions: the code for the omnibus test for non-uniformity of circular data (Sinz et al 2018); the code for the approximation for log ($n!$) developed by Srinivasa Ramanujan (Ramanujan 1988); and the code for generating a neighbor-joining phylogeny from CNV data, which was provided by Brian Haas.

Algorithm

My algorithm constructs a network of nodes and edges corresponding to a given phylogeny. It is assumed that the topology of the phylogeny and the amounts of time represented by each of its edges are correct. Tip nodes[1] correspond to cells in the present-day cancer. Internal nodes represent the putative common ancestors of those descendants. Edges represent lines of descent from one node (common ancestor) to the next.

All nodes have locations in (x, y) space as well as in time (t). Each edge crosses a distance (d), the distance between its ancestral and its descendant node, and is associated with an elapsed amount of time (β, defined as $t_{descendant} - t_{ancestor}$) between those nodes.

Each edge may be traversed in one of two phenotypes, epithelial or mesenchymal. These are modeled with a hidden Markov model. Migration is modeled as a process of Brownian diffusion, with the direction of each edge assumed to be drawn from a uniform distribution over the interval [0, 2π). When a particle is undergoing Brownian diffusion, the distance between its locations before and after a period of time β are distributed as a normal distribution with μ=0 and standard deviation $\psi\sqrt{\beta}$, where ψ^2 is the expected increase in the variance of a particle's position per unit time. ψ therefore is in units of $distance/\sqrt{time}$ (Lemmon and Lemmon 2008). The probability of a edge in state k crossing a distance d in an amount of time β is (Eq. 1; modified from Lemmon and Lemmon 2008):

$$b_k = \frac{\sqrt{2}d}{\psi_k^2\sqrt{\beta\pi}} exp\left\{\frac{-d^2}{2\beta\psi^2}\right\}$$

[1] In this report, I use the terminology that is customary in the phylogenetics literature rather than in the computer-science literature, e.g., "tips" rather than "leaves".

The algorithm is initialized by assigning the known locations of the present-day cancer cells to the tip nodes, and by assigning values for t to the internal nodes based on the phylogeny that has been passed to the algorithm. Assigning t to all nodes defines the elapsed time β_e corresponding to each edge e. (x, y)-coordinates are chosen randomly for each internal node from within the rectangular area defined by the maximum and minimum values seen for x and y among the tip nodes. These (x, y)-coordinates define the distance d_e crossed by each edge, allowing ψ_e to be calculated for each edge as $d_e/\sqrt{\beta_e}$.

Under the assumption that cells in the mesenchymal phenotype are less common than cells in the epithelial phenotype (A. Lambert, pers. comm.), $\psi_{epithelial}$ is initialized at the mean of the lowest 95% of the values for ψ_e, with $\psi_{mesenchymal}$ initialized at the mean of the highest 5%. Prior probabilities for the two phenotypes are, similarly, set at 0.95 and 0.05. Transition probabilities are initialized such that the phenotype of a lineage has only a probability of 0.025 of changing at any given node.

For optimization of the (x,y)-locations of the internal nodes, these nodes are arranged in random order. For each, five proposals are considered. The node can be left in its current location, or it can be moved by one step, positive or negative, along the x- or y-axis. The step size used is the absolute value of a number chosen randomly for each of the four proposals from a normal distribution with $\mu = 0$ and a standard deviation of $\sigma_s T_s$, where σ_s is an initial step size of 10μm and T_s is a "temperature" parameter that decreased from 1.0 to 0.01 over the course of 750 iterations of EM.

For each of these five proposals, d_e is calculated, given the proposed new position, for the three edges neighboring that node (or two in the case of the root node). Based on Eq. 1, b_k is calculated for these edges. One of the five proposals is selected with a probability proportional to (Eq. 2):

$$\left(\prod_e b_k\right)^{\frac{1}{exp\{T_p\}}}$$

where T_p is a second "temperature" parameter that went from 1.0 to -3.0 over the 750 iterations. The use of T_s was intended to make early movements toward optimum (x,y)-coordinates more rapid, while allowing fine-tuning later in the process. The use of T_p was intended to allow the algorithm to range more widely over sub-optimal areas in parameter space in the early iterations (in hopes of avoiding becoming trapped in a local maximum) but to constrain the algorithm from such wandering later on. Values for the start and end points of these parameters were chosen based on observations of the behavior of the model.

Once one of the five proposals has been selected and implemented for a node, the algorithm moves on to the next node. The maximum-likelihood (x,y)-locations for all nodes are therefore approached in a more simultaneous fashion than in the algorithm of Lemmon and Lemmon (2008), thereby reducing the influence that the random ordering of nodes can have on the final locations that are reached.

When all nodes have passed through the proposal process, the algorithm calculates the marginal probabilities of each edge being crossed in either the epithelial or the mesenchymal phenotype. This requires two traversals of the tree. In the first traversal, from tips to root, a probability $f_j(i)$ is calculated for each state j at each node i. In the second traversal, from root to tips, a probability $g_k(i\text{-}1)$ is calculated. The procedure was as follows:

At each tip node, values for $f_k(i)$ are calculated as (Eq. 3):

$$f_j(1) = a_{0j} b_j(1)$$

where a_{oj} is the prior probability for phenotype j and $b_j(1)$ is the probability (Eq. 1) of phenotype j having "emitted" the value of $d_e/\sqrt{\beta_e}$ for the edge e ancestral to this node, which is node 1.

At internal nodes, following the work of Felsenstein (1973, 1981), $f_k(i)$ incorporates values of $f_j(i-1)$ from both descendant nodes (Eq. 4):

$$f_k(i) = b_k(i) \sum_{j1} f_{j1}(i-1) a_{j1k} \sum_{j2} f_{j2}(i-1) a_{j2k}$$

At the root, the likelihood for the entire phylogeny is calculated as (Eq. 5):

$$L = \sum_k a_{k0}\, f_k(N)$$

where a_{oj} is the prior probability for phenotype j.

To start the second traversal, $g_j(N)$ at the root node is assigned the value of the prior probability a_{j0}. For each of its daughter nodes, $g_{ji}(N\text{-}1)$ is then calculated as (Eq. 6):

$$g_{j1}(i-1) = \sum_k \left(g_k(i) a_{j1k} b_k(i-1) \sum_{j2} f_{j2}(i-1) a_{j2k} \right)$$

where $g_k(i)$ is from the node ancestral to node i-1, and $f_{j2}(i-1)$ is from the node that is the *sibling* to node i-1, as calculated in the earlier traversal from tips to root.

$f_j(i)$ is the probability that, regardless of the paths taken from node i to the tips descended from it, the edge ancestral to i arrived at node i in state j. $g_{j1}(i-1)$ is the probability that, regardless of the paths taken from node i-1 to the tips *not* descended from node i, the edge ancestral to node i *left* node i-1 in state j. From these two probabilities, the algorithm calculates the marginal probability that the edge ancestral to node i was crossed in state j, regardless of what paths might have been taken before or after that point.

After both traversals are complete, for each state j, ψ_j is calculated as the weighted average of $d_e/\sqrt{\beta_e}$ over all edges; the weights are the marginal probabilities of having crossed edge e in state j. The marginal probabilities are also used to estimate new values for the transition probabilities a_{jk} and the prior probabilities for the two states.

The algorithm then, again, places the internal nodes in random order and goes through the proposal process to adjust the location of each one in (x,y)-space. Because sub-optimal proposals can be accepted, the algorithm does not converge on an exact final set of parameters. Rather, I allowed the algorithm to run until it appeared that the values for ψ_j remained more or less constant for a large number of iterations. (Generally, I ran a total of 750 iterations; the values for ψ_j usually neared stability around 400 iterations.) I accepted the estimates of ψ_j from the iteration with the highest L to be the best estimates of ψ_j.

Literature Cited

Felsenstein, J. 1973. Maximum-likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology* 22: 240-249.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* 17: 368-376.

Lemmon, A. R., & Lemmon, E. M. (2008). A likelihood framework for estimating phylogeographic history on a continuous landscape. Systematic Biology, 57(4), 544-561.

Ramanujan, S. The Lost Notebook and Other Unpublished Papers. 1988. *Narosa, New Delhi*.

Sinz, F., P. Berens, T. Wallis, M. Waskom, & Matthias K. (2018). https://github.com/circstat/pycircstat/blob/master/pycircstat/tests.py. Accessed December 12, 2018.