Historic inference from non-recombinant genetic data

Vladimir Voevodsky

Started November 15, 2004

1 Introduction

This text contains a preliminary description of the algorithms for generating sample non-recombinant genealogies from demographic parameters (birth and death rates), for reconstructing these genealogies from their genomic images and for reconstructing the demographic parameters from the genealogies. In addition to the usual data structures these algorithms operate with objects of the following two custom classes:

Balanced Trees Mathematically speaking, a balanced tree is a rooted tree[1] together with a map from the set of edges to positive real numbers (we call these numbers the "lengths" of the edges) such that the length of all paths from the root to leaves are the same. This length is called the depth of the tree. For instance, a balanced tree with two leaves is determined by two non-negative numbers l_0 and l_1 where l_0 is the length of the edge going from the root to the internal point of the tree and l_1 the length of either of the two edges going from the internal point to the leaves. The depth of such a tree is $l_0 + l_1$.

For a vertex of a balanced tree its depth is defined as the length of any path leading from this vertex to a leaf. The depth of a leaf is zero and the depth of the root is the depth of the tree. A balanced tree can be thought of as a genealogy where edges represent individuals, vertices represent the birth (division) events and the depth of a vertex is the time from the present to the corresponding event. In our context edges will often represent not the actual individuals but the "virtual" ones corresponding to sequences of individuals descended from each other in such a way that at each division point the is only one descendant.

Enriched Balanced Trees An enriched tree is just a balanced tree together with a map from its vertices to the set of sequences of 0's and 1's of a fixed length L. In the genealogical interpretation the value of this map on a vertex is the genome of the individual corresponding to the incoming edge at the moment of the division event. Clearly, an enrichment is completely determined by the genome of the initial individual together with the data given for each vertex and showing how the genome of this vertex differs from the genome of the previous one. We will use this later way of describing enrichment since in our context it is much more efficient computationally.

There are the following main modules:

Direct Demographic This module contains methods for generating balanced trees as "reduced trajectories" of Markov birth and death processes. It takes as an input the time of the initial

[1] The root of a rooted tree is not considered to be a vertex. In particular we say that one point is a rooted tree with zero vertices.

moment T, the initial number of individuals N_0 and the birth $b(t)$ and death $d(t)$ rates on the interval $[T, 0]$ and outputs balanced trees which correspond to the genealogies of the present day descendants of the initial population. The number $\tilde{N_0}$ of the balanced trees it generates is the number of members of the initial population who happened to have at least one living descendant at time zero. The depth of each tree is T.

Direct Genomic This module contains methods for producing enrichments of existing balanced trees. It takes as an input a balanced tree, a natural number L representing the length of the genomes and mutation rate m on the interval $[T, 0]$. At the moment the mutation rate is assumed to be constant. For simplicity the genome of the initial individual is always taken to be constantly zero (NB: the inverse modules discussed below do not know this). It outputs an enriched tree based on the original balanced tree.

Intermediate Processing This module contains methods for modifying enriched trees. It includes adding noise of different kinds and restricting the tree to a subset of the present day population.

Inverse Genealogical This module contains an algorithm for producing an approximate reconstruction of the genealogy of a population in the form of a balanced tree from the genetic distances between the members of the present day population. It takes as an input a number N corresponding to the population size, a function $gd(i, j)$ defined for $1 \leq i, j \leq N$ corresponding to the genetic distances between the members of the population and the numbers L and m corresponding to the genome length and to the per-site mutation rate. It outputs a balanced tree with N leaves.

Inverse Demographic This module contains an algorithm for producing an approximate reconstruction of the demographic parameters (the birth and the death rates $b(t)$, $d(t)$) from the genealogy of a subset of the present day population. It takes as an input natural numbers $N_s \leq N_{total}$ corresponding to the size of the sample and the total size of the present day population, a balanced tree with N_s leaves corresponding to the genealogy of the sample, a real number T corresponding to the length of the considered time interval and non-negative numbers $b_{min}, b_{max}, d_{min}, d_{max}$ which determined the minimal and maximal allowed birth and death rates. It outputs a natural number N_0 corresponding to the size of the population at time T and functions $d(t)$, $b(t)$ on $[-T, 0]$.

2 Marked, balanced and enriched trees

We define a marked tree as a rooted tree together with a map λ from its set of edges to the set $\mathbf{R}_{>0}$ of positive real numbers. We call $\lambda(\gamma)$ the length of edge γ. One defines a notion an isomorphism between two marked trees in the obvious way. For $N \geq 0$, set MT_N to be the set of isomorphism classes of marked trees with N leaves. We have $MT_0 = pt$, $MT_1 = \mathbf{R}_{>0}$, $MT_2 = \mathbf{R}_{>0} \times Symm^2(\mathbf{R}_{>0})$ etc. For two vertices in a marked tree define their distance as the sum of lengths of the edges forming the shortest (unoriented) path from one vertex to another. Let $MT_{N,T}$ be the subset inn MT_N which consists of marked trees such that the distance from any leaf to the root is $\leq T$.

A marked tree is called balanced if all the distances from the root to the leaves are equal. Since any vertex in a tree can be connected to the root by a unique oriented path this condition implies that for any vertex all the distances from it to any of the leaves lying under it are equal. We call the corresponding number the depth of a vertex. The depth of a leaf is zero. The depth of the root is called the depth of the tree. Let $BT_{N,T}$ be the subset of balanced trees of depth T in MT_N.

For any marked tree Γ in $MT_{N,T}$ define a balanced tree $Cn(\Gamma)$ as follows. Intuitively $B_T(\Gamma)$ is obtained by removing all the branches of our tree which are shorter than T where the length is counted from the root.

More formally, let us say that an edge is T-lucky if there is a leaf under its ending point which has the distance at least T from the root. The union of all (closed) T-lucky edges forms a tree $Cl(\Gamma)$ whose branching points correspond to vertices of Γ which have at least two outgoing T-lucky edges. The set of leaves of $Cl(\Gamma)$ is the subset of the set of leaves of Γ which consists of leaves lying at the distance T from the root. This construction defines a map $Cl : MT_N \to \coprod_{M \leq N} BT_{M,T}$.

3 Direct demographic module

Consider hypothetical creatures which we call singletons. A singleton is born from one parent, lives for some time and then dies or divides into two new singletons. We consider the history of a population of singletons on an interval $[T, 0]$. Let $d(t_1, t_2)$ (resp. $b(t_1, t_2)$) be the probability that a singleton which is alive at time t_1 will die (resp. will divide at least once) during the time interval $[t_1, t_2]$. Assume that for each $t \in [T, 0)$ the limits

$$d(t) = \lim_{\Delta t \to 0} d(t, t + \Delta t)/\Delta t$$

$$b(t) = \lim_{\Delta t \to 0} b(t, t + \Delta t)/\Delta t$$

exist. We will call $d(t)$ and $b(t)$ the death and birth rate in our population at time t. The story of a given singleton (alive at time T) and its descendants on the interval $[T, 0]$ can be encoded by a marked tree. The edges of this tree correspond to the individuals involved, the ramification points to the division events, the leaves which lie at distance $< T$ from the root to the death events and the leaves which lie at the distance T from the root to the descendants who were alive at time 0. We call it the tree of descendants of a singleton.

Formally speaking, our "demographic parameters" $d(t)$ and $b(t)$ define a probability measure $\mu = \mu(d(t), b(t))$ on the space $MT_{N,T}$ such that for a good enough subset U in $MT_{N,T}$ the number $\mu(U)$ is the probability that the tree of descendants of a singleton in a population developing according to $d(t)$ and $b(t)$, lies in U. Instead of defining this measure analytically we will describe a probabilistic algorithm which generates marked trees according to this measure. Note that such an algorithm uniquely defines the measure.

For convenience we will assume that for all $t \in [0, T]$ one has $b(t) > 0$. Set

$$B(t_1, t_2) = \int_{t_1}^{t_2} b(t)$$

$$D(t_1,t_2)=\int_{t_1}^{t_2} d(d)$$

We call B and D the cumulative birth and death rates. Consider a singleton alive at tie t_1. One can easily see that the probability that nothing (neither death nor division) happens to it on the interval $[t_1,t_2]$ equals $exp(-B(t_1,t_2)-D(t_1,t_2))$. Correspondingly the probability that something happens is $1-exp(-B(t_1,t_2)-D(t_1,t_2))$. From the definition of $b(t)$ and $d(t)$ it follows that if we know that an event happened at time t then this event is a division with the probability $b(t)/(b(t)+d(t))$ and a death with the probability $d(t)/(b(t)+d(t))$.

To generate the descendant tree of an individual it is clearly sufficient to be able to produce for any initial moment t_1, the moment t of the next event and then use the previous remark to determined whether this event is a death or a division. Then one uses the obvious recursive construction and either repeats the procedure for the descendants or returns back to the ancestor to follow the next branch.

In practice a computer program can only produce random numbers in the interval $[0,1]$ according to the standard measure $\xi:[x_0,x_1]\mapsto x_1-x_0$. We need to produce a random number t starting with t_1 according to the measure m determined by

$$m([t_1,t_2])=1-exp(-B(t_1,t_2)-D(t_1,t_2)).$$

One can easily that that to get t we need to find a function $\phi:[0,1]\to[t_1,\infty)$ such that $\phi_*(\xi)=m$ and then set $t=\phi(x)$ where x is a "standard" random number in $[0,1]$. Let us look for ϕ in the class of increasing functions. For such a function the condition $\phi_*(\xi)=m$ means that

$$\phi^{-1}(t)=1-exp(-B(t_1,t)-D(t_1,t)).$$

i.e.

$$\phi=(1-exp(-B(t_1,-)-D(t_1,-)))^{-1}.$$

The positivity condition which we imposed on $b(t)$ guarantees that the function on the right is an increasing one and therefore it has a well defined inverse which is also an increasing function. Since inverses of increasing functions can be easily computed numerically this completes our description of the algorithm for producing the descendant trees from $d(t)$ and $b(t)$.

Let an ordered marked tree be a marked tree together with an ordering of its leaves.