Journal of Theoretical Biology 423 (2017) 31–40

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/jtbi

Maximum likelihood estimates of pairwise rearrangement distances

Stuart Serdoz[a], Attila Egri-Nagy[a,b], Jeremy Sumner[c], Barbara R. Holland[c], Peter D. Jarvis[c], Mark M. Tanaka[d,e], Andrew R. Francis[a,*]

[a] *Centre for Research in Mathematics, Western Sydney University, Australia*
[b] *Akita International University, Japan*
[c] *School of Physical Sciences, University of Tasmania, Australia*
[d] *School of Biotechnology and Biomolecular Sciences, University of New South Wales, Australia*
[e] *Evolution & Ecology Research Centre, University of New South Wales, Australia*

ARTICLE INFO

Article history:
Received 7 January 2017
Revised 12 April 2017
Accepted 13 April 2017
Available online 20 April 2017

Keywords:
Genome rearrangement
Inversion
Maximum likelihood
Phylogeny
Algebraic biology
Group theory
Coset

ABSTRACT

Accurate estimation of evolutionary distances between taxa is important for many phylogenetic reconstruction methods. Distances can be estimated using a range of different evolutionary models, from single nucleotide polymorphisms to large-scale genome rearrangements. Corresponding corrections for genome rearrangement distances fall into 3 categories: Empirical computational studies, Bayesian/MCMC approaches, and combinatorial approaches. Here, we introduce a maximum likelihood estimator for the inversion distance between a pair of genomes, using a group-theoretic approach to modelling inversions introduced recently. This MLE functions as a corrected distance: in particular, we show that because of the way sequences of inversions interact with each other, it is quite possible for minimal distance and MLE distance to differently order the distances of two genomes from a third. The second aspect tackles the problem of accounting for the symmetries of circular arrangements. While, generally, a frame of reference is locked, and all computation made accordingly, this work incorporates the action of the dihedral group so that distance estimates are free from any *a priori* frame of reference. The philosophy of accounting for symmetries can be applied to any existing correction method, for which examples are offered.

1. Introduction

Estimates of evolutionary distance between pairs of taxa are key ingredients for reconstructing phylogenies, but are difficult to obtain reliably (Felsenstein, 2004; Gascuel, 2005). This is especially true for evolutionary models in which events can interact with each other in a way that affects inference. One estimate of distance between two genomes is the *minimal* distance which is model-specific and represents an assumption of parsimony in evolutionary paths through genome space (see Fertin, 2009 for a discussion of rearrangement models in this context). In fact, for most models, there are infinitely many possible evolutionary paths between any two genomes, and the minimal distance is simply the length of one of the shortest of these; by definition the minimal distance can only underestimate the true number of evolutionary events.

The problems with using a minimal distance are well documented, especially when time periods are long and the space of obtainable genomes becomes saturated. Given enough time, all evolutionary endpoints become equally likely, and any signal of actual evolutionary time is lost. In some models, metrics have been developed to account for multiple changes; the most well-known perhaps being the Jukes–Cantor correction for models of single nucleotide substitution (Jukes and Cantor, 1969). This method requires all events to be *independent* (a common assumption with nucleotide substitution), but such independence does not hold for most genome rearrangement models (such as inversion) and so alternative approaches are needed.

Given pairwise distances obtained from a phylogenetic tree, Buneman (1971) demonstrated that the recovered tree is unique, a fact which also follows from the 4-point condition (Buneman, 1974). Furthermore, Warnow (1996) and Atteson (1999) suggest that if the true evolutionary distance inference is sufficiently accurate, even polynomial time reconstruction algorithms, such as Neighbor Joining (Saitou and Nei, 1987), will return the correct

[*] Corresponding author.
E-mail addresses: 16115907@student.westernsydney.edu.au (S. Serdoz), egri-nagy@aiu.ac.jp (A. Egri-Nagy), jeremy.sumner@utas.edu.au (J. Sumner), barbara.holland@utas.edu.au (B.R. Holland), peter.jarvis@utas.edu.au (P.D. Jarvis), m.tanaka@unsw.edu.au (M.M. Tanaka), a.francis@westernsydney.edu.au (A.R. Francis).

http://dx.doi.org/10.1016/j.jtbi.2017.04.015
0022-5193/© 2017 Elsevier Ltd. All rights reserved.

phylogeny. Recent work by Gascuel and Steel (2015) places the results of Atteson et al. in a statistical framework.

Some studies attempt to find a relationship between true distance and minimal distance (or some other available measure such as breakpoint distance), and use this to produce an estimate of true distance as a function of minimal distance. For instance, Wang and Warnow (2001) introduced an estimator of true evolutionary distance called *IEBP* (inverting the expected breakpoint distance). The method operates under the generalised Nadeau–Taylor model (Nadeau and Taylor, 1984) and provides a robust polynomial time algorithm to estimate true evolutionary distance. Similarly, the *EDE* (empirically derived estimator) of Moret et al. (2001) samples the relationship between inversion distance and true evolutionary distance before providing a fit. Applications of IEBP and EDE can be seen in Wang (2002).

While a useful correction, such estimates are based on just one factor – the minimal distance – and can't account for underlying structure of the genome space (in our framework, the Cayley graph of the group). The key point being that not all elements of equal minimal distance are equally likely.

As an optimal estimate of true distance, we would like (very loosely) some sort of *expected* distance – a function of final arrangement – constructed as a weighted average of evolutionary paths, pushing the problem into the intersection of combinatorics and statistics. In this vein, Eriksen (2002) offered an approximation of the expected number of inversions to have occurred given n breakpoints. This was followed by a method of estimating the expected inversion distance by looking at the expected transposition distance (Eriksen and Hultman, 2004), and generalizations such as Eriksen (2005) and Dalevi and Eriksen (2008).

Given the sizes of the spaces involved, MCMC and Bayesian methods play an important role. York et al. (2002) use a Bayesian framework to estimate true distances for inversions. On the MCMC front, Miklós (2003) introduced a time continuous stochastic approach to genome rearrangements (modelled as a Poisson process), allowing reliable estimates of true distances. The key aim being to describe the posterior distribution of true evolutionary distance given two arrangements. There have been several generalizations to these methods: Durrett et al. (2004) include translocations as well as inversions; Larget et al. (2005) describe a Bayesian method for phylogeny inference and offer a comparison between their approach and a parsimony approach; and Miklós and Darling (2009) provide a method to estimate the *number* of minimal walks.

This paper describes a novel *maximum likelihood* approach to corrected rearrangement distances. We focus on models of genome rearrangement involving invertible operations, such as inversion and translocation, which can be described in group-theoretic terms, using the framework introduced in Egri-Nagy et al. (2014b) and Francis (2014). This algebraic framework treats genomes as the images of the actions of elements of a finite reflection group, and allows us to treat the genome as not fixed in space, but free to rotate in Euclidean space. Each genome is then considered to be a coset in the quotient of the main reflection group by the dihedral group.

The next section describes the general group-theoretic models of chromosome rearrangements on which this paper is based. The third section introduces the likelihood function under our model, and gives some basic examples of what these functions look like. Next, we compare the minimal distance to the MLE and give an example of how the resulting phylogenetic inference can give different results. We then consider properties of group elements that may characterise the likelihood function and hence the MLE of distance. The penultimate section describes what is required to account for dihedral symmetry, and illustrates the approach with some example phylogenies. We end with a discussion of some of the issues involved in using the MLE.

2. Group theoretic models of rearrangement

In this section we describe group-theoretic models of genome rearrangement, following the development in Egri-Nagy et al. (2014b). Such models allow events that change the underlying sequence in a reversible way, including for example inversion and translocation but not insertion or excision. The invertible rearrangements defined by the model then generate a *group*, and there is a one-to-one correspondence between the set of possible genome arrangements and the set of elements of this group.

This correspondence in practice requires two additional assumptions. First, we choose one genome as the reference genome, that will correspond to the group identity element. This is arbitrary, and is discussed in more detail below. Second, we assume there is no rotation of the genome in 3-dimensional space. We think of this as fixing a “frame of reference” for all genomes. This assumption is removed for calculating MLEs of evolutionary distances in ways described below, by taking a quotient by the dihedral group.

The genome space is then realized as a graph with genomes as vertices and allowable evolutionary events defining edges between them. This corresponds to a graph based on the group, called the *Cayley graph*, whose vertices are group elements and edges represent multiplication by the group generators. Thus the Cayley graph can be thought of as a map of the genome space, with vertices the possible genomes (group elements) and edges the possible rearrangement events (generators of the group) (Clark et al., 2016). The Cayley graph depends on both the group $\mathcal{G}$ and the generating set $\mathcal{S}$.

Given a choice of one arrangement as the reference genome G_0, every other genome arrangement can be obtained from G_0 by a sequence of rearrangements. Because each allowable rearrangement event defines a generator of the group, this sequence of rearrangements is a product of group generators, and therefore corresponds to a group element itself. Thus the reference genome G_0 corresponds to the identity element e of the group $\mathcal{G}$, and each other possible genome corresponds to a unique group element (remembering that for now we assume a fixed frame of reference). Note that there may be many sequences of events giving rise to the same genome, and these correspond to different walks through the Cayley graph.

A brief note on the language of paths and walks. In graph theory a *walk* through a graph is an alternating sequence of vertices and edges beginning at one vertex and ending at another. This may or may not involve traversing the same edge or vertex multiple times. A *path* on the other hand is a walk in which no vertices or edges are visited more than once. To avoid confusion we will use “walk” in the context of the Cayley graph, but it is worth noting that *minimal* walks between two group elements on the Cayley graph are all paths, in this sense. It is common, however, outside of graph theory, to use the expression “evolutionary path” between two organisms without the implication that no genome has been visited more than once (allowing, for instance, homoplasy or convergent evolution), and we will also use “path” in that context, where clear.

Returning to walks and distances on the Cayley graph, observe that the choice of reference genome is not important. For any two genomes G_1 and G_2 with corresponding group elements g_1 and g_2, there is a unique group element (namely $g_1^{-1}g_2$ when acting on the right) that transforms G_1 into G_2. As a result of the transitive group action (Babai, 1996), the group element is independent of the choice of reference genome. For instance if G_1 was chosen as the reference genome then the walk from G_1 to G_2 would still correspond to the group element $g_1^{-1}g_2$ (which in this case would be simply g_2, since here $g_1 = e$).