Expected Distance

Stuart Serdoz

University of Western Sydney

16115907@student.uws.edu.au

November 6, 2014

Overview

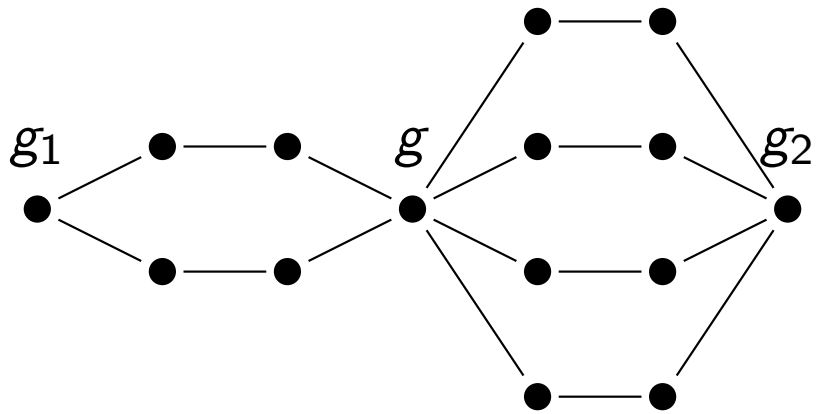
Distance based methods

- Distance methods rely on constructing a matrix of pairwise distances between taxa.
- Pairwise distances are minimal distances between the two taxa
- Criticisms exist based upon geodesic distances not reflecting the real inversion history.

- Our example in mind:
  - circular bacterial genomes, changing under inversion
  - modelled as a group action, genomes correspond to group elements
  - genome space seen as Cayley graph, inversions the generators defining the edges.

Criticism 1: Intervals - Number of geodesics

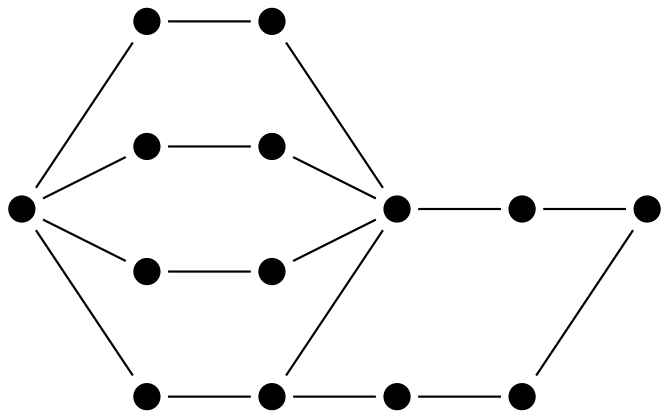
Assuming equal length paths are equally likely $\Longrightarrow$ minimal distance does not encode enough information.
The likelihood of reaching genome G is not just a function of the geodesic distance.



Miklos and Darling 2009 - Offered strategies to estimate the number of minimal paths with the intent of improving the application of minimal distance.
It was one of the first methods to attempt to include the structure of the group, and dealt with situations like above.

Criticism 2: Paths of any length

MD2009 acknowledged criticisms of minimal distance, and addressed some. But what about non-minimal paths?



We expect the likelihood of a extremely long path to be small (possibly negligible) however the number of paths jumps exponentially (at least) as the path length increases.

Perhaps in larger numbers, the longer paths aren't so negligible.

Expected Distance

- Let the r.v. I be the number of steps taken, and the r.v. X be the group element labelling the endpoint.
- The expected distance represents the expected number of steps leading from the identity to g

$$E(I|X = g) = \sum_{i \geq 0} i\, p(i|g).$$

- $p(i|g)$ hard to interpret/find. Hence by Bayes' thm.

$$p(i|g) = \frac{p(i,g)}{p(g)} = \frac{p(g|i)p(i)}{\sum_{j \geq 0} p(g|j)p(j).}$$

Expected Distance

$$E(I|X = g) = \frac{\sum_{i \geq 0} ip(g|i)p(i)}{\sum_{j \geq 0} p(g|j)p(j).}$$

Now to deal with the components

- $p(g|i)$ - The probability of reaching g after i steps is addressed with help from "paths of equal length are equally probable".

$$p(g|i) = \frac{\rho_i(g)}{n^i}$$

  where $\rho_i(g)$ is the number of length i paths ending at g.
- $p(i)$ - The probability of i steps occuring between observations is a bit of a problem. Rate of inversion as well as time between observations seems to be at play here.

p(i)

Based on the assumption that the expected number of steps in a given time interval is proportional to the length of time, for a time period of fixed length the distribution is Poisson in rT.

However, the time (T) until observation itself varies according to an exponential distribution. Hence

$$T \sim Exponential(\theta) \qquad \text{and} \qquad I|T \sim Poisson(T) \tag{1}$$

Hence some algebra shows us that $I \sim Geometric(\theta)$.

Expected Distance expression

Installing the resulting PMF yields

$$E(I|G=g) = \frac{\sum_{i\leq 0} i\,\rho_i(g)(\frac{\beta}{n})^i}{\sum_{i\leq 0} \rho_i(g)(\frac{\beta}{n})^i}$$

β is a parameter which controlls the assumed inversion rate/length of time.

Small β ($\beta \to 0$) corresponds to a low inversion rate/short timeframe.

Larger β ($\beta \to 1$) corresponds to faster inversion rates/longer timeframe.

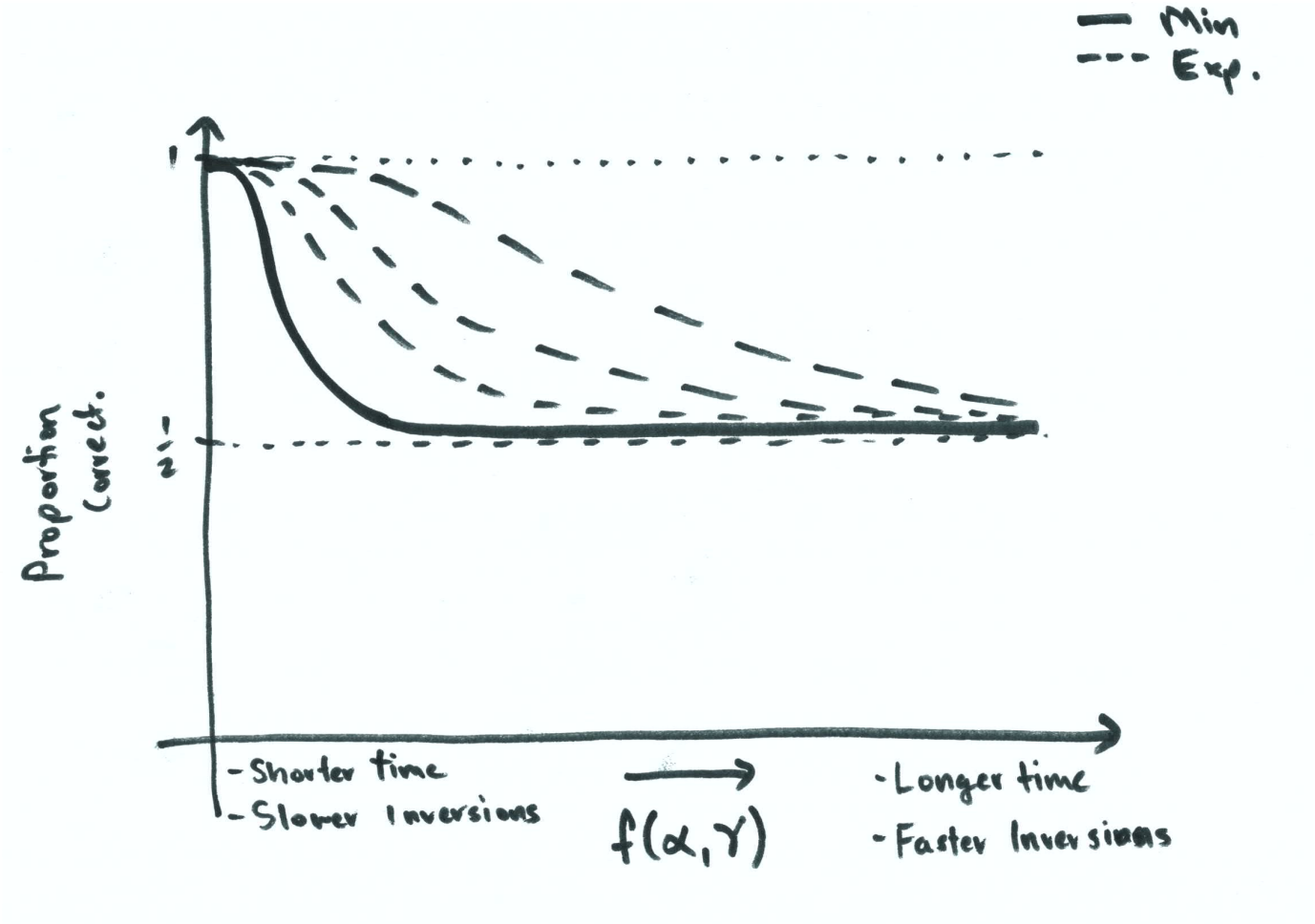
Simulation and Examples

The aim of expected distance was to reflect the true evolutionary path length better than the minimal distance and to compare.

Data simulation is provided by a branching process developed by Sangeeta. The input parameters are the bifurcation rate (α), and inversion rate (γ).

1. Simulate a 3 taxa tree.
2. Construct pairwise distance matrices using both minimal and expected distance.
3. Use Fitch-Margoliash to estimate the phylogeny.
4. Compare topologies of both estimated phylogenies with the true phylogeny.

Simulation and examples



ρ Matrix in S_3

The ρ matrix is where the computation gets tricky.

g	$\rho_0(g)$	$\rho_1(g)$	$\rho_2(g)$	$\rho_3(g)$	$\rho_4(g)$	$\rho_5(g)$	$\rho_6(g)$	$\rho_7(g)$
()	1	0	2	0	6	0	22	0
(2,3)	0	1	0	3	0	11	0	43
(1,2)	0	1	0	3	0	11	0	43
(1,2,3)	0	0	1	0	5	0	21	0
(1,3,2)	0	0	1	0	5	0	21	0
(1,3)	0	0	0	2	0	10	0	42

Algorithms are getting more efficient. Currently working on genomes with 9 regions ($|G| \approx 350000$). Working on ways to introduce more algebraic structure to speed up computation.

Saturation point

Naturally saturation is expected in extreme lengths.

$E(I|g = G)$ relies on the ratio $\frac{\rho_i(g)}{n^i}$.

But $\lim_{i\to\infty} \frac{\rho_i(g)}{n^i} = \frac{1}{|G|}$.

Perhaps it makes sense at some point (once sufficiently mixed) to approximate all $\frac{\rho_i(g)}{n^i}$.

This point may be moot with small β (short time).

Thank you for listening!