

The parsimony assumption in distance based methods

Stuart Serdoz

University of Western Sydney

16115907@student.uws.edu.au

November 7, 2013

- 1 Parsimony in distance based methods
- 2 Random walks on groups
- 3 Mixing time

Distance based methods

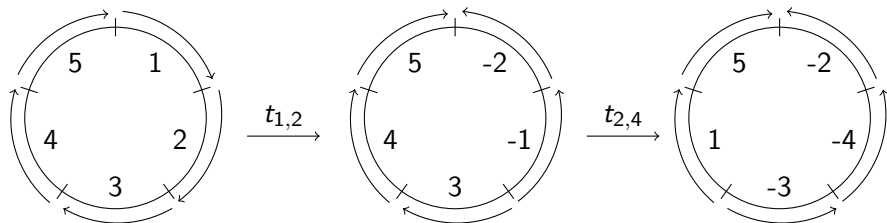
- Distance methods rely on constructing a matrix of pairwise distances between taxa.
- Pairwise distances represent geodesic distances (not number of evolutionary events).
- Main criticisms based upon the idea that geodesic distances do not reflect the real evolutionary history.

so lets do something about it.

How to we measure the strength of the parsimony assumption?

Construction of group models

- Group model not about forcing current algebraic model onto biological situation.
- Circular bacterial genome modelled as a signed permutation.
- Inversions most common large scale rearrangement.
- 'Legal' inversions are taken as the generators of the group. The evolution of a particular genome can be seen as a path through the Cayley graph.



Random walks on finite groups

- Consider a finite group G with symmetric generating set S with identity included in S .
- A random walk involves starting from the identity, then randomly selecting a generator (inversion) from S and multiplying by that generator.
- k step walk simulates k random (legal) inversions.
- Calculate the geodesic distance of endpoint from identity, and compare to k .

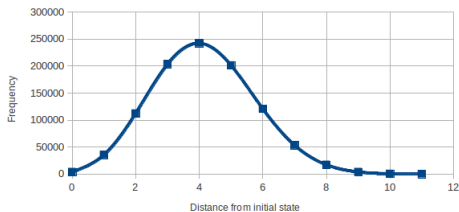
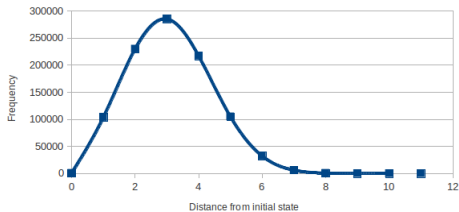


Figure: Left: 10 step walks

Right: 20 step walks

What does this mean for the assumption of parsimony?

- Parsimony relationship \implies geodesic length = evolutionary path length.
- For short paths, parsimony good.
- For long paths, parsimony bad.

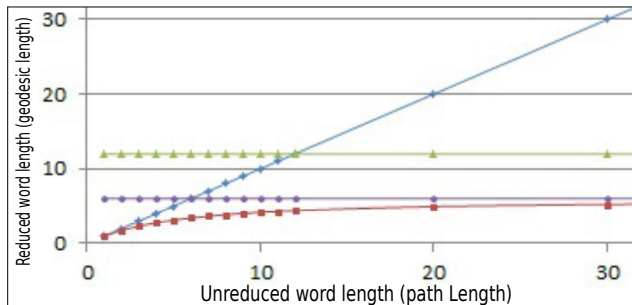
Is there a predictable relationship between the path length (# of inversions) and the expected geodesic distance?

For how long is the assumption of parsimony good?

When does the parsimony assumption become nonsense?

Simulation of the inversion process

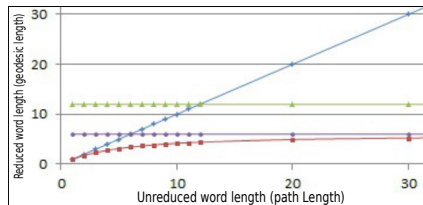
Consider the average geodesic length of random walks of many different path lengths.



- Note the diameter of the group (green line) represents a hard upper bound.
- Average geodesic distance seems to approach half the diameter.

Asymptotic behaviour: the mixing time

- The mixing time of a Markov chain is a measure of how long it takes to chain to reach a distribution *close* to stationary.
- The mixing time represents a hard upper bound on the parsimony assumption.
- If we consider a path length longer than the mixing time, it implies that every genome is equally likely- might as well throw darts.
- Mixing time represents a point on the horizontal axis.



So when is parsimony good?

- Three different stages of behaviour.
 - Early (linear) stage \implies Parsimony good
 - Late (asymptotic) stage \implies Parsimony bad
 - Intermediate stage \implies ???
- Where do we define the transition from linear to intermediate?
- Is it at this point where parsimony assumption far from expected distance?
- In our small example, linear stage very short. Real world examples (80 regions) linear stage much longer.

Note that I don't hate the parsimony assumption! We would just like to clarify where and when it can be used.

After observing the expected distance approaching half the diameter, we went digging.

Theorem (Asymptotic expected distance)

Consider a path γ derived from a random walk on a finite group G generated by a set S of involutions. The limiting behaviour of the geodesic distance D_γ is given by

$$\lim_{n \rightarrow \infty} E(D_\gamma) = \frac{d_G}{2}$$

where d_G is the diameter of the group.

- Is there some closed form for the expected distance as a function of path length? (not hopeful)
- Define where parsimony fails by considering some sort of distance between the expected distance function, and the diagonal representing the unreduced word length.
- Perhaps improve bounds on the mixing time by incorporating more aggressive assumptions based off algebraic structure of our inversion groups.