Phylodynamics: how Genetics and Mathematics are changing our understanding of infectious diseases.

Luiz Max F. de Carvalho [lm.carvalho@ed.ac.uk]

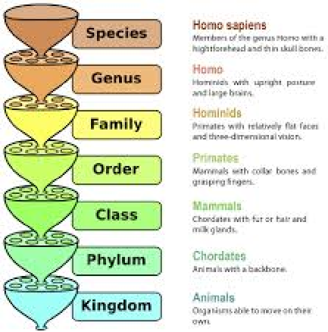
Rambaut group, Institute of Evolutionary Biology, University of Edinburgh, UK.
XXX Brazilian Mathematics Colloquium – IMPA, Rio de Janeiro, Brazil.

July 29, 2015

Plan for today

- A brief overview of how we came to using phylogenetic trees and mathematical models;
- The awesome things we can do with them;
- An interesting combinatorics problem on how to build better trees.

In the beginning there were hierarchies...


Species
Genus
Family
Order
Class
Phylum
Kingdom
Homo sapiens
Members of the genus Homo with a high forehead and thin skull bones.
Homo
Hominids with upright posture and large brains.
Hominids
Primates with relatively flat faces and three-dimensional vision.
Primates
Mammals with collar bones and grasping fingers.
Mammals
Chordates with fur or hair and milk glands.
Chordates
Animals with a backbone.
Animals
Organisms able to move on their own.

Linnaean system of classification ($\sim$ 1735)

The plate tectonics of Biology...

Haeckel (1866)

- Darwin did not know how genetic variation arose or was transmitted;
- Mendelian geneticists believed in evolutionary "jumps" and rejected natural selection;
- Contrary to what is usually believed, Lamarckian thought was important to maintain gradualism;
- Haldane, Fisher and Wright reconciled population genetics and selection[a]

[a] This is not the whole story! Check the preface in Mayr and Provine (1998) for a much better account.

I’ve never done anything useful...

Geoffrey Harold Hardy (1877-1947)

- “I have never done anything ’useful’. No discovery of mine has made, or is likely to make, directly or indirectly, for good or ill, the least difference to the amenity of the world.”[a]
- Oh yeah?

I’ve never done anything useful...

Geoffrey Harold Hardy (1877-1947)

- “I have never done anything ’useful’. No discovery of mine has made, or is likely to make, directly or indirectly, for good or ill, the least difference to the amenity of the world.”[a]
- Oh yeah?
-

$$
\begin{aligned}
A(p); a(q)& \\
p + q &= 1 \\
p^2 + 2pq + q^2 &= 1
\end{aligned}
$$

I've never done anything useful...

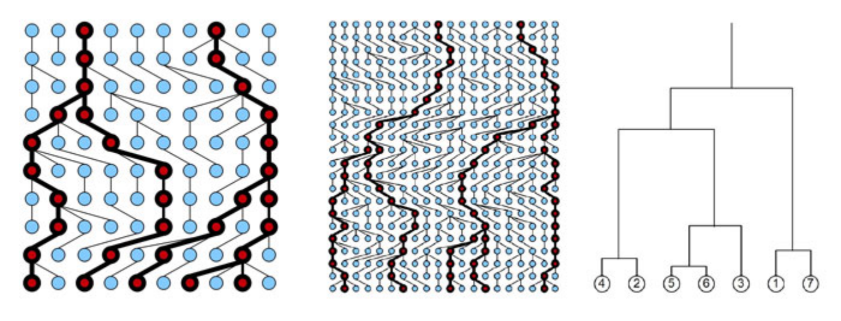
Geoffrey Harold Hardy (1877-1947)

- "I have never done anything 'useful'. No discovery of mine has made, or is likely to make, directly or indirectly, for good or ill, the least difference to the amenity of the world."[a]
- Oh yeah?
- $$\begin{aligned} A(p);\, a(q)& \\ p+q &= 1 \\ p^2+2pq+q^2 &= 1 \end{aligned}$$
- Stability!

[a] Titchmarsh (1950)

Population Genetics

- Hardy-Weinberg is nice and useful, but describes an ideal setting with no mutation, selection or population structure.
- In the 1980s, John Kingman starts a revolution...



Kingman's coalescent (1982)[1]

[1] Figure from http://culturemath.ens.fr/content/genealogie-de-populations-le-coalescent-de-kingman-2101

Population Genetics

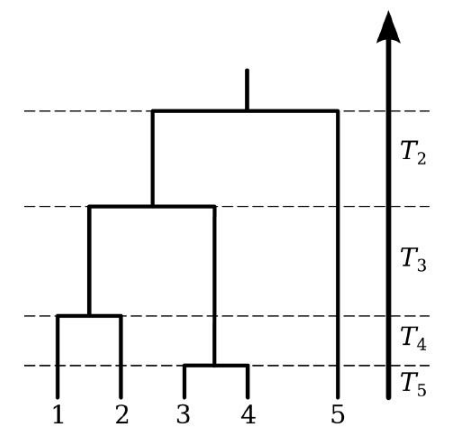


Figure 4 from Volz et al. (2013)

Let T_n denote the time for n lineages to *coalesce*, i.e., merge into one ancestral lineage, in a population of size N. Then:

$$Pr(T_n = t) = \lambda_n e^{-\lambda_n t}$$

$$\lambda_n = \binom{n}{2}\frac{1}{N}$$

Let T_{mrca} denote the age of the most recent common ancestor:

$$\begin{aligned}\mathbb{E}[T_{\text{mrca}}] &= \mathbb{E}[T_n] + \mathbb{E}[T_{n-1}] + \ldots + \mathbb{E}[T_2] \\ &= 1/\lambda_n + 1/\lambda_{n-1} + \ldots + 1/\lambda_2 \\ &= 2N(1 - \frac{1}{n})\end{aligned}$$

Bridging Disease Ecology and Genetics

if the tree is a reasonable representation of ancestry, then we can use it as proxy to hidden/unobservable population processes. One such process is infection:

$$\frac{dS}{dt} = -\beta IS$$
$$\frac{dI}{dt} = \beta IS - \gamma I$$
$$\frac{dR}{dt} = \gamma I.$$

Here is an idea[2]:

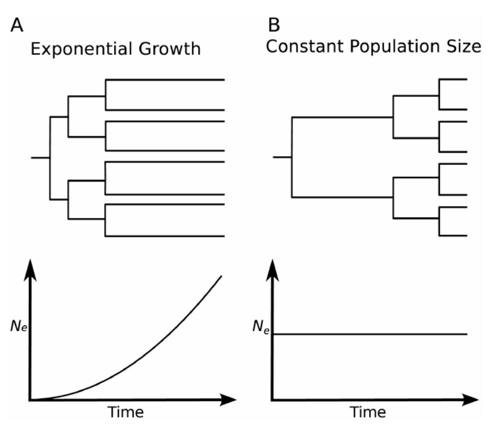
$$\lambda_n(t) = \binom{n}{2} \frac{2\beta S(t)}{I(t)}$$

Some more glossing over the details yields:

$$\lambda_n = \binom{n}{2} \frac{2\gamma}{I}$$

[2] Equations and ideas taken from Volz et al. (2013)

Bridging Disease Ecology and Genetics



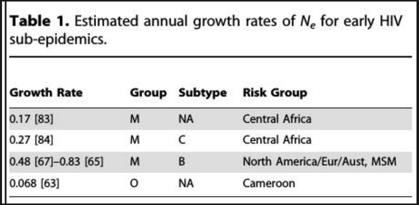
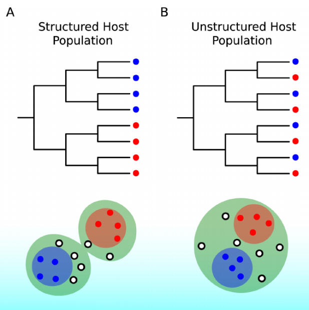


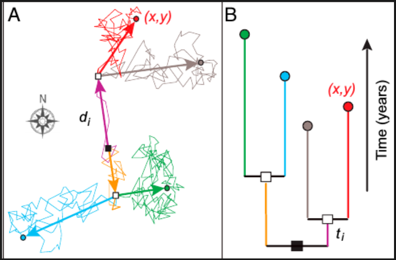
Table 1. Estimated annual growth rates of N_e for early HIV sub-epidemics.

Growth Rate	Group	Subtype	Risk Group
0.17 [83]	M	NA	Central Africa
0.27 [84]	M	C	Central Africa
0.48 [67]–0.83 [65]	M	B	North America/Eur/Aust, MSM
0.068 [63]	O	NA	Cameroon

Taken from Volz et al. (2013)

Bridging Disease Ecology and Genetics II





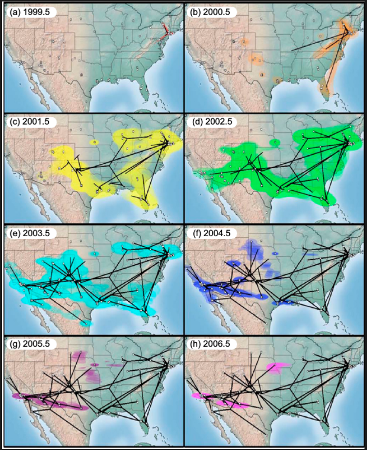
Let $\mathbf{X}(t)$ be the state at the t. We are interested in the likelihood of ending up in $\mathbf{X}(t)$ if we started at $\mathbf{X}(s)$. Here's an idea (Lemey et al., 2010):

$$p(\mathbf{X}(t)|\mathbf{X}(s)) \sim MVN(\mathbf{X}(s), \mathbf{P}^{-1} \times (t-s))$$

where $\mathbf{P}$ is the infinitesimal precision matrix of the diffusion process. Note that:
(i) the diffusion matrix depends only on time differences and;
(ii) analytical solutions are available through clever traversing of the tree (Pybus et al., 2012).

Bridging Disease Ecology and Genetics II

“...phylogenies reconstructed from spatial epidemics are branching structures that record the correlated histories of transmission among sampled infections...” (Pybus et al., 2012)



The spread of West Nile virus (WNV) in the US (Pybus et al., 2012)

Summary so far...

- By conditioning on the inferred ancestry of DNA/RNA sequences, we can use phylogenetic trees as correlation structures;
- It is possible to obtain insight into temporal (dynamic) and spatial features of populations from genetic data;
- Phylogenies are the centre of it all: almost everything we do is conditional on the tree.

Summary so far...

- By conditioning on the inferred ancestry of DNA/RNA sequences, we can use phylogenetic trees as correlation structures;
- It is possible to obtain insight into temporal (dynamic) and spatial features of populations from genetic data;
- Phylogenies are the centre of it all: almost everything we do is conditional on the tree.
- This suggests we should also pay close attention to tree estimation... Bayesian and maximum likelihood phylogenetic methods rely on stochastic tree search. There is great interest in making the traversal of tree-space more efficient.

An open problem – preliminaries

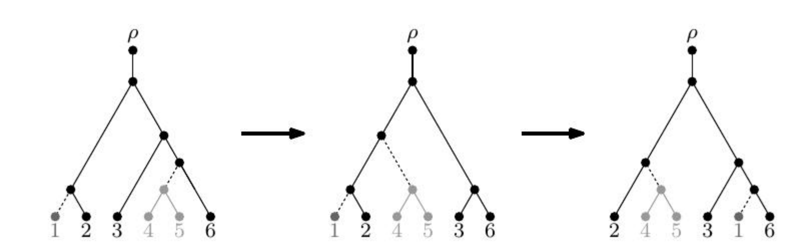
Defn. 1: An (unrooted) phylogenetic X-tree is a tree T on a tip (leaf) set X with all internal nodes of degree of at least 3. If the degree of all internal nodes is exactly 3 then we say we have a *binary* X-tree.

- Every unrooted (rooted) binary X-tree has $2n - 3$ $(2n - 2)$ edges (also called branches), with $n = |X|$;
- There exist $\frac{(2n-3)!}{2^{n-2}(n-2)!} = (2n - 3)!!$ rooted X-trees on n tips/leafs. For $n = 53$, there are roughly as many trees as there are particles in the observable universe!

See Steel (2014) for a gentle introduction to phylogenetic trees, specially targeted at mathematicians.

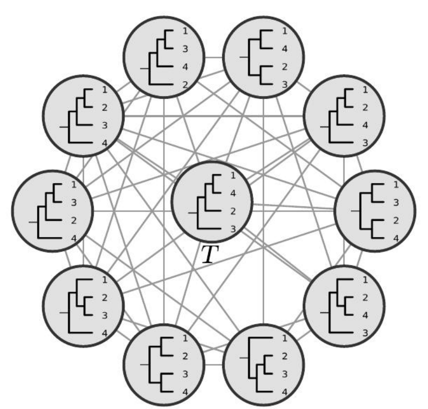
Tree surgery...

Let $Y \subset X$. Then call X' the Y-tree that has internal nodes compatible with those in T, i.e., X' is a *subtree* of X An Subtree-prune and re-graft (SPR) operation on a tree T detaches X' and re-grafts it onto another node in the tree.



Two consecutive (rooted) SPR operations on a 6-leaf tree (Whidden and Matsen, 2015)

Graphs about graphs...



The neighbourhood of T in the SPR graph (Whidden and Matsen, 2015)

The diameter is $n - \Theta(\sqrt{n})$ and the average degree is $2(n-3)(2n-7)$.

Ricci-Ollivier curvature (Whidden and Matsen, 2015)

When performing a random-walk in a metric space, it is convenient to rigorously define a measure of "curvature", that is, how "difficult" it is to go from x to y (Ollivier, 2009). Whidden and Matsen (2015) consider two points x and y on the SPR graph (G). Let m_x and m_y be the probability masses of the positions x and y on G after some (finite) time. Then:

$$W_1(m_x, m_y) := \min_{\xi \in \Pi(m_x, m_y)} \sum_{\{z,w\} \subset V} d(z,w)\xi(z,w) \tag{1}$$

which loosely measures how much "work" is involved in moving m_x to m_y along G. The so-called coarse Ricci-Ollivier curvature of x and y is then:

$$\kappa(m; x, y) := 1 - \frac{W_1(m_x, m_y)}{d(x,y)}. \tag{2}$$

Here, for two trees T and S, $d(T,S)$ is how many SPR operations are needed to transform T in S (vice-versa).

A conjecture

Whidden and Matsen (2015) prove a bunch of useful results about the SPR graph. For instance, they show that for two trees T and S:

$$\frac{-2}{d(T,S)} \leq \kappa(T,S) \leq \frac{2}{d(T,S)} \tag{3}$$

Moreover, if T and S are adjacent (with $|T| = |S| = n$), the the maximum curvature is

$$\frac{6n-17}{3n^2-13n+14} \tag{4}$$

Conjecture (Whidden and Matsen, 2015): Let k_n be the maximum curvature between two trees with n-leaves. Then:

- $k_n \leq \frac{2}{D(n)-1}$, *and*
- $\lim_{n\to\infty} k_n = \frac{2}{D(n)-1}$.

where $D(n)$ is the diameter of the SPR graph on n-leaves trees.

My two pence

- It is important to prove not only this conjecture, but also to study variants on different spaces resulting from other transformations;
- In the context of phylodynamics, height-restricted SPRs are arguably a more important transformation;
- Recursive application of the techniques used to prove the curvature for adjacent trees could be extended to k-radius neighbourhoods;
- A more productive approach could be to relate what we know for the graph (e.g. average degree, diameter, etc) to its local properties, taking advantage of more heavy graph-theoretic methods/results.

Wrap-up

- Maths was fundamental to advance our understanding of population genetics;

Wrap-up

- Maths was fundamental to advance our understanding of population genetics;
- Maths plays a central role in Phylodynamics, be it on developing models or figuring out how to fit them;

Wrap-up

- Maths was fundamental to advance our understanding of population genetics;
- Maths plays a central role in Phylodynamics, be it on developing models or figuring out how to fit them;
- There is a lot of cool maths to be done in:

Wrap-up

- Maths was fundamental to advance our understanding of population genetics;
- Maths plays a central role in Phylodynamics, be it on developing models or figuring out how to fit them;
- There is a lot of cool maths to be done in:
    - Lie algebras of Markov models of DNA evolution;
    - Further merging ODE-based models and phylogenetics;
    - Combinatorics on the space of phylogenetics trees ~~and networks~~!

Wrap-up

- Maths was fundamental to advance our understanding of population genetics;
- Maths plays a central role in Phylodynamics, be it on developing models or figuring out how to fit them;
- There is a lot of cool maths to be done in:
  - Lie algebras of Markov models of DNA evolution;
  - Further merging ODE-based models and phylogenetics;
  - Combinatorics on the space of phylogenetics trees ~~and networks~~!

- Maths rocks! Q.E.D

Thank you!

- Thanks a bunch for watching!
- I am grateful to the organising committee for the invitation;
- Special thanks to Andrew Rambaut (Edinburgh), Mike Steel (Cantenbury, NZ), Patrice Showers Corneli (Utah) and Erick Matsen (Hutchson)[3].

[3] Mathematicians are underlined.

References

Lemey, P., Rambaut, A., Welch, J. J., and Suchard, M. A. (2010). Phylogeography takes a relaxed random walk in continuous space and time. *Molecular biology and evolution*, 27(8):1877–1885.

Mayr, E. and Provine, W. B. (1998). *The evolutionary synthesis: perspectives on the unification of biology*. Harvard University Press.

Ollivier, Y. (2009). Ricci curvature of markov chains on metric spaces. *Journal of Functional Analysis*, 256(3):810–864.

Pybus, O. G., Suchard, M. A., Lemey, P., Bernardin, F. J., Rambaut, A., Crawford, F. W., Gray, R. R., Arinaminpathy, N., Stramer, S. L., Busch, M. P., et al. (2012). Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proceedings of the National Academy of Sciences*, 109(37):15066–15071.

Steel, M. (2014). Tracing evolutionary links between species. *The American Mathematical Monthly*, 121(9):771–792.

Titchmarsh, E. C. (1950). Godfrey Harold Hardy. *J. London Math. Soc.*, 25(2):81–138.

Volz, E. M., Koelle, K., and Bedford, T. (2013). Viral phylodynamics. *PLoS Comput. Biol.*, 9(3):e1002947.

Whidden, C. and Matsen, F. (2015). Ricci-ollivier curvature of the rooted phylogenetic subtree-prune-regraft graph. *arXiv preprint arXiv:1504.00304*.

Useful links

http://i.imgur.com/T5XLPLB.jpg
http://www.math.canterbury.ac.nz/~m.steel/files/presentations/winthrop1_to_4.pdf
http://www.math.canterbury.ac.nz/~m.steel/files/presentations/winthrop5_to_10.pdf
http://www.phylobabble.org/t/the-interface-between-mathematics-statistics-and-phylogenetics-topics-for-a-talk/537