

Doom01: biological mathematics in evolutionary processes

David Penny and Susan Holmes

The fifth annual workshop on the mathematics and biology of evolutionary trees was held in Whakapapa Village, in the volcanic centre of the North Island of New Zealand from 12 to 16 February 2001.

The active volcano Mt Ngauruhoe (pronounced *Naru ho ee* – Mt Doom in the new film, *Lord of the Rings*) was the backdrop for this unique workshop, which stimulated dynamic interaction between mathematicians and biologists.

Phylogenetics is conceptually one of the most difficult areas in biology, inferring events from millions to billions of years ago will always be hard. In theoretical terms, phylogenetics has still only explored a small part of the available mathematical and statistical theory, and there was an interesting difference in perspective between the different subject areas, with biologists interested in getting their trees right and the theoreticians also wanting the same rigor for their epistemological taxonomy. Presentations at this meeting varied from the methodological perspectives with basic mathematics, graph theory, algorithms, probability and statistics, to the presentation of new data stimulating new questions and/or analytical techniques (abstracts are available at <http://imbs.massey.ac.nz/Doom.htm>).

A need for new data types

There is a challenge for biologists to find new data types, because theorems are now available on the minimum possible number of perfect characters necessary to define a tree uniquely. Mike Steel (University of Canterbury, New Zealand) showed that, for binary characters, this is $t=3$ (t is the number of taxa). However, when the number of character states is unlimited, five characters are sufficient. These theorems reopen the phylogeny problem after 10–20 years of focussing on primary sequence data. This suggests that biologists should be more open to new types of characters, and gene order and SINES (small interspersed nuclear elements) are already candidates. David Bryant [Centre de Recherche Mathématiques (CRM), Montreal, Canada] and Nadia El-Mabrouk (CRM) described the good progress that has been made for analyzing gene order data.

Stretching the paradigms

A formal description of the 'geometrical space' of trees was provided by Susan Holmes (Stanford University, CA, USA), and its geometric properties should aid in identifying confidence sets for trees and unifying different consensus methods in a common framework. Tom Hagedorn (CRM) reported invariants for more general models of evolution. In the past, invariants have not been that useful in practice (possibly because of their high variances), but they do help us to understand the basis of the different mutation models. Searching for distant RNA molecules is aided by combining primary and secondary structure information (Paul Gardner, Massey University, New Zealand), and decomposing LogDet distances into contributions from each amino acid allows inferences about changes in substitution process in the tree to be made.

Other speakers questioned why we often limit theoretical phylogenetics to only a subset of useful statistics and mathematics. There has been a strong preoccupation with consistency of various methods, but now it is necessary to consider that sequences are of fixed, rather than of infinite length. Questions arise about the 'information' available in sequences, and the identification of how much of this information that we actually use. This leads to questions about the statistical sufficiency of the estimators and identifiability of the underlying parameters.

Many examples of complex (nonbinary) patterns of evolution were described. Those discussed included recombination (especially in RNA viruses), mass spawning and hybridization (illustrated by Madeleine van Oppen's study of reef corals, and by Lynne van Herwerden's study on reef fish), gene conversion, gene duplication and loss, and multiple mutations in populations; Russell Gray (University of Auckland, New Zealand) demonstrated the case of borrowing words between different Pacific languages. Much of what is labeled 'junk DNA' is of interest for analyzing microevolution at the population level and for understanding how much of various viral leftovers still persist within the genome. It might be that viral leftovers possibly produce as much information as the actual genes themselves.

New methodologies

Biologists are beginning to use more complicated statistical methods, such as those shown by Chris Simon's (Victoria University, Wellington, New Zealand) Bayesian analysis of the New Zealand *Cicada* data, Andy Rambaut's (University of Oxford, UK) nonparametric analysis of HIV data, Alex Grossman's (Centre National de Recherches Scientifiques, Versailles, France) use of rank correlation methods to test LogDet distances and Lars Jermin's (Australian Genomic Information Center, Sydney, Australia) use of model averaging.

Networks instead of trees

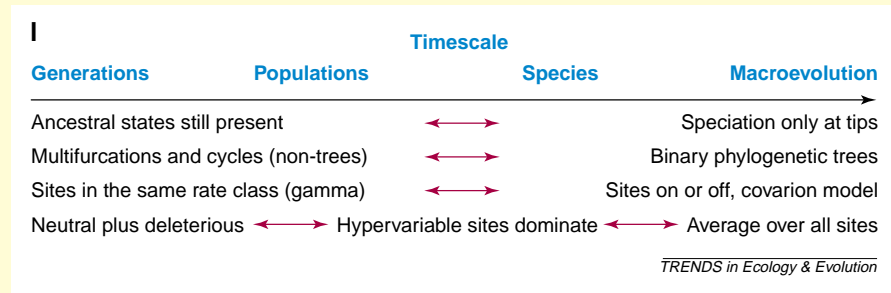
We seem much less wedded to the binary tree paradigm and there were many examples of networks¹ used instead of simple binary trees given at the meeting, including split decomposition analyses of languages and buttercups (Peter Lockart, Massey University), and Barbara Holland (Massey University) used median networks for the Adélie penguins *Pygoscelis adeliae*. In cases such as Y-chromosome studies in Polynesians (Matt Hurles, University of Cambridge, UK), additional data resolves most cycles in the graphs. However, in other data, the resolving of trees might be hampered by repeated gene conversion events, or by the fact that there is never enough data available to resolve the many parallel mutations. Progress was reported by Charles Semple (University of Canterbury) on identifying the minimum number of hybridizations required to untangle a complex network. There were also advances reported for models of speciation, especially the tip growth and uniform models. With a more accurate description of evolutionary patterns, we can expect improvement in the formal testing of models of speciation. There were several data analyses that used the well-tested multivariate techniques such as multidimensional scaling and discriminant analysis to give an overall picture of the data before the modelling phase.

The timescale problem

Are all our methods equally applicable for timescales from generations, to populations, to species, to long-range macroevolution?

Box 1. Effects of timescale

At shorter times, ancestral character states (haplotypes) are still present in the population, leading to multifurcations (rather than binary trees) (Fig. 1). For a variety of reasons there may be cycles in the graph, leading to non-tree models. Over the shorter term, and for the simplest cases, sites in a sequence are expected to remain in the same rate class. However, over the longer term, there are changes in the 3D structure of macromolecules^a.



Reference

^a Lesk, A.M. (2001) *Introduction to Protein Architecture: The Structural Biology of Proteins*, Oxford University Press

The question of multiscale analyses has made its way through computational mathematics, physics, astronomy, statistics and now into biology. Whether we are interested in different genes, populations or species, we use the same techniques – for example, binary trees with just the tips (leaves) can be identified as either existing haplotypes in a population, or as separate species (Box 1). Is it valid to use the same form of analysis for such different timescales? A population will have internal nodes that represent ancestral states still present in the population. What does this mean for current coalescence methods that only allow tip growth? Similarly, we tend to assume that there is just one ‘rate’ of evolution for timescales from genealogies of individuals, to evolutionary trees of families and orders. It appears that different timescales lead to different ‘rates’ of evolution (Box 1). Each measurement is valid (it is just that they are different) and there is an effect of timescale on the shape of the tree².

Interesting new data sets are a stimulus to improved forms of analysis; given the many new data sets that were published in 2000, over-stimulation could well have resulted. The origins and early evolution of HIV in humans was aided by a new data set from the Congo, relevant to the discussion of some of the more creative explanations for the appearance of AIDS in Africa. Dorit Liebers (Max Planck Institute for Evolutionary Anthropology,

Leipzig, Germany) showed how the classical textbook example of ring speciation, the gulls of the Arctic, has been complemented by an impressive collection of new data sets for gulls around the northern oceans, allowing a clever testing strategy for the ring speciation model to be implemented. A new data set on Adélie penguins has been collected from the Antarctic, which includes many ancient DNA sequences that Barbara Holland demonstrated to be a major challenge for classical tree-building programs.

Rooting a tree

All the new data showed that the problem of rooting a tree is often the most difficult aspect to get correct, and that the assumptions made are overly simple and can be quite misleading. Three papers reported new mitochondrial genomes (including the first from extinct organisms) that should help resolve some of the confusion pertaining to fairly weak signals from far away times. Going even further back in time to the origin of chloroplasts and photosynthesis, the problem of the root became even harder. Tony Larkum (University of Sydney, Australia) pointed out how slight differences in procedure or assumptions change the root, even if the underlying phylogeny is stable. Perhaps a common feature from each of the major new data sets (from populations to ancient divergences) is that they are still raising challenges for the theoreticians.

The data is fine, it is theory that is inadequate.

There was certainly no complacency on the issue of inadequate theory, and it shows the advantage of a workshop that includes both biologists and mathematicians. Bridging these subjects are the programmers who have a foot in both camps and, at the meeting, examples of the results at this interface between idealism and reality were described. In the week that *Nature* and *Science* published the public and private assemblies of the human genome, Daniel Huson (Celera Genomics, Rockville, MD, USA) described the elegant graph-theory techniques that were used in his combined assembly of the human genome. Almost as elegant, and equally fundamental, was Allen Rodrigo (University of Auckland) and Alexei Drummond’s (University of Auckland) analysis of HIV sequences sampled at different times from the same host. This is evolution in action and, even apart from its biomedical importance, it illustrates the importance of RNA viruses for evolution.

In the best traditions of workshops, the formal presentations stimulated further discussion, ideas were developed, theorems were proposed and new collaborations were arranged. There was the usual outbreak of MSV (‘Mike Steel virus’), which causes temporary loss of sanity, leading victims to impossible feats of physical activity, exploits that they would normally have the sense not to do. The annual tradition is being maintained and Allen Rodrigo is the organizer for the 2002 workshop.

Acknowledgements

Peter Lockhart deserves especial credit as organizer. Primary financial support was from the New Zealand Marsden Fund.

References

- 1 Strimmer, K. and Moulton, V. (2000) Maximum Likelihood for networks using directed graphical models. *Mol. Biol. Evol.* 17, 875–881
- 2 Steel, M.A. and Penny, D. (2000) Parsimony, likelihood and the role of models in molecular phylogenetics. *Mol. Biol. Evol.* 17, 839–850

David Penny*

Molecular BioSciences, Massey University, Palmerston North, New Zealand.

*e-mail d.penny@massey.ac.nz

Susan Holmes

Dept of Statistics, Stanford University, CA 94305, USA.