The Penny Ante: Mathematical biology or biological mathematics?

Mike Steel
Allan Wilson Centre for Molecular Ecology and Evolution, and Department of Mathematics and Statistics, Canterbury University, Private Bag 4800, Christchurch

One of the great joys in mathematics is when a complex problem that has no right to possess any reasonable solution can be solved by some elegant and exact formula. When the problem has some bearing on a real scientific question, the discovery is all the sweeter. Many examples of this can be found in different fields of science. In evolutionary biology, one of my favourites is the 'bichromatic binary tree (BBT) theorem'. This deals with a question that arose when biologists began constructing phylogenetic trees using an 'Occam's Razor' approach of finding the simplest explanation for the data – the maximum parsimony tree. The question was to find how many binary phylogenetic trees have a given parsimony score for a given assignment of states to the species? Although there are many elegant formulae for counting trees – some of which date back to mathematicians working around the time of Charles Darwin – we have no reason to suspect that this parsimony counting problem should have any nice, easy-to-use solution.

A mathematician would probably not have even tried to search for a formula – but legend has it that David Penny, during some of the less riveting moments of the Los Angeles Olympics in 1984, began counting various cases and managed to formulate a a shortlist of possible formulae for this problem via a combination of trial and error, inspired guesswork and determination. Some of these formulae failed, but eventually, one of them seemed to work. Some mathematicians got interested. However, it took a full year, a team effort of four mathematicians, and a large computer to tackle some very complex algebra and eventually confirm that this remarkable formula was actually correct in all cases (Carter *et al.* 1990). Since then a more direct proof has been found, and it tells us a lot about the structure of most-parsimonious trees (for more details, see Semple & Steel (2003)). Today, Penny's formula remains the most remarkable closed-form expression for any class of phylogenetic trees in evolutionary biology.

The story of the BBT theorem contains an interesting moral. The textbook image of 'mathematical biology' is of the biologist who comes to a mathematician with a problem. The mathematician goes away and, after many cups of coffee and lots of chalk-dust on the cardigan sleeves, solves it and presents it to the appreciative biologist. Of course, this scenario rarely happens – often, the mathematician will decide that the original problem is too tricky, so will change some of the rules (assumptions) and turn it into a different problem that can be solved, all the while secretly hoping not to be accused of 'cheating' too much. But even if, after much work, the original problem can be solved, the mathematician knows there is still a good chance the biologist will not really appreciate the beautiful mathematical calculations, perhaps referring to them as mere 'chicken scratchings'. To add further insult, the biologist is likely to mutter something like: 'Actually, I think the assumptions I described to you last week are probably wrong – the process is much more complex, you know!' In the story of the BBT theorem, the process of interaction is different again: in this case, the biologist came to the mathematician and said 'Here is the solution; now prove it for me!'

Being presented with a claimed mathematical solution to a problem by a biologist is rare; usually, the interaction between mathematicians and biologists involves a more to-and-fro approach, discussing what might be the case, trying to identify what properties of the data are responsible for different signals, attempting to formalise the problem in a mathematically precise, tractable model, and distilling the problem down to a few well-chosen questions for further study.

The remarkable collaboration between Mike Hendy and David Penny has been characterised by this type of interaction, and has delivered several highlights. An early gem was the development of a branch-and-bound algorithm that allowed optimal most-parsimonious phylogenetic trees to be constructed

Mike Steel is a Principal Investigator at the Allan Wilson Centre for Molecular Ecology and Evolution and is Professor of Mathematics and Statistics and Director of the Biomathematics Research Centre at the University of Canterbury, Christchurch.

He had a brief career as a newspaper journalist following his MSc in mathematics, and was supervised for his PhD thesis by David Penny and Mike Hendy. He graduated from Massey University in 1989 before taking up a postdoctoral fellowship in Germany. He has published 150 research papers on applications of mathematics in evolutionary biology, including 23 joint papers and book chapters with David Penny. He enjoys finding 'field trip' excuses to head into the Southern Alps and may be contacted at M.Steel@math.canterbury.ac.nz

on more than just a handful of taxa. Apart from its direct usefulness in data analysis, it also set the stage for their bold paper in *Nature* (Penny *et al.* 1982) that provided the first formal test of the theory of evolution in a Popperian framework, requiring, once again, new mathematical input to determine how similar one would expect two unrelated trees to be by chance alone. A second and more recent highlight of the Hendy–Penny collaboration has been the elegant and useful Hadamard representation of symmetric models of DNA substitution (Hendy & Penny 1989).

It was particularly exciting to be doing a PhD under their supervision at the time they were making this second discovery. More recently, I have been fortunate to find myself on the receiving end of the 'Here is the solution; now prove it!' approach of David Penny. Often this has been presented as a challenge or bet, with an appropriate prize (a bottle of rare single malt whisky) as reward for any success. This tradition of the 'Penny Ante' continues to this day at our annual international phylogenetics meetings (see University of Canterbury 2008), with problems and challenges posted online to entice the participants. More than one paper has resulted from this wager; and any funding body would surely be impressed at the research output per dollar spent on the whisky prize.

The offer of a prize or challenge may seem quaintly amusing or an outright gimmick, but for many mathematicians, it is precisely the incentive that will make them pick up a pen and start working hard. Inspired by David Penny's constant pestering about finding the limits of phylogenetic accuracy in sequence data – a theme that we nicknamed 'the war on error' during the Bush years – several mathematicians began to establish new results that explicitly quantified phylogenetic signal. Yet an outstanding problem resisted all attempts to solve; it was only after we offered a $100 prize that some very smart mathematicians at UC Berkeley finally cracked the problem in 2006, and we now have a much more precise idea of how much DNA sequence data would be needed to construct a very large tree (Daskalakis *et al*. 2006). The prize was, of course, symbolic (but it had to be presented in person) – we estimate it would have cost at least 300 times this amount if they had charged us for their time.

Joel Cohen wrote an essay in which he claimed: 'biology is mathematics' new physics, only better' (Cohen 2004). While certain areas of modern physics, such as string theory, have become theory-rich but data-poor (though the CERN large hadron collider may possibly reverse this trend), molecular evolutionary biology, by contrast, is overwhelmed with genomic data but the analysis is still trying hard to catch up. Moreover, unlike string theory, biology has a single and simple unified theory (evolution) that provides a framework for understanding what these data tell us about species, populations, and their origin and development.

The physicist Eugene Wigner wrote of the 'unreasonable effectiveness of mathematics' (Wigner 1960) in areas of science like physics and astronomy. In biology, the connection is less transparent and has taken longer to flourish; later, the Italian-born mathematician Gian-Carlo Rota (1986) made this scathing remark:

> *The lack of real contact between mathematics and biology is either a tragedy, a scandal or a challenge, it is hard to decide which.*

But in the two decades since Rota's comment, some spectacular advances have been made in many areas of computational, statistical and mathematical biology, including the rise of complete new disciplines such as bioinformatics. The type of mathematics being applied is often quite different from that employed in the physical sciences, where calculus, differential equations, and dynamical systems dominate. This part of traditional mathematics is sometimes referred to by its detractors as 'steam engine mathematics', though it is still very useful in various areas of mathematical biology – from modelling the propagation of electrical impulses along nerve fibres to the spread of an epidemic through a large population.

However, the mathematics required in molecular evolution has mostly involved other fields – combinatorics, probability, algorithmic theory, Markov Chain Monte Carlo methodology, and so forth. These have allowed the study and analysis of evolutionary relationships on a scale undreamed of twenty years ago. Indeed, it is remarkable how many of the controversies in evolutionary biology at that time (for instance, the chimp/human/gorilla trichotomy or the recent radiation of modern humans from Africa) have been largely resolved. In their place, a new suite of questions have arisen. Developing the theory to solve them will require the efforts of the next generation of mathematicians, statisticians and computer scientists, working closely with biologists – and listening for those occasions when a biologist has a hunch what the answer is and lays down the ante: 'just prove it!'

References

Carter, M.; Hendy, M.; Penny, D.; Szekely, L.A.; Wormald, N.C. 1990. On the distribution of length of evolutionary trees. *SIAM Journal on Discrete Mathematics 3*: 38–47.

Cohen J. E. 2004. Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better. *PLoS Biology 2(12)*: e439.

Daskalakis, C.; Mossel, E.; Roch, S. 2006. Optimal phylogenetic reconstruction. *Proceedings of the Thirty-eighth Annual ACM Symposium on Theory of Computing (STOC 2006)*: 159–168.

Hendy, M.D.; Penny, D. 1989. A framework for the quantitative study of evolutionary trees. *Systematic Zoology 38*: 297–309.

Penny, D.; Foulds, L.R.; Hendy, M.D. 1982. Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. *Nature 297*: 197–200.

Rota, G-C. 1986. Discrete thoughts. In: Kac, M., Rota G-C., Schwartz J.T. (Eds) *Discrete Thoughts: Essays on Mathematics, Science, and Philosophy.* Boston, Birkhauser.

Semple, C.; Steel, M. 2003. *Phylogenetics*. Oxford, Oxford University Press. 239 p.

University of Canterbury. 2008. Penny Ante. http://www.math.canterbury.ac.nz/bio/events/kaikoura09/penny.shtml

Wigner, E. 1960. The unreasonable effectiveness of mathematics in the natural sciences. *Communications in Pure and Applied Mathematics 13*: 1–14.