biology letters
Molecular evolution

Biol. Lett. (2009) **5**, 391–393
doi:10.1098/rsbl.2009.0227
Published online 22 April 2009

Introduction

Putting the 'bio' into bioinformatics

Lindell Bromham*

Centre for Macroevolution and Macroecology, Botany and Zoology, School of Biology, Australian National University, Canberra, Australian Capital Territory 0200, Australia
**lindell.bromham@anu.edu.au*

Bioinformatic analyses have grown rapidly in sophistication and efficiency to accommodate the vast increase in available data. One of the major challenges has been to incorporate the growing appreciation of the complexity of molecular evolution into new analytical methods. As the reliance on molecular data in biology and medicine increases, we need to be confident that these methods adequately reflect the underlying processes of genome change. This special issue focuses on the way that patterns and processes of molecular evolution are influenced by features of populations of whole organisms, such as selection pressure, population size and life history. The advantage of this approach to molecular evolution is that it views genomic change not simply as a biochemical or stochastic process, but as the result of a complex series of interactions that shape the kinds of genomic changes that can and do happen.

Keywords: mutation; substitution; molecular evolution; relaxed clocks; barcoding; alignment

1. THE VALUE OF WHOLE ORGANISM APPROACHES TO MOLECULAR EVOLUTION

When I began my PhD, GenBank came in the post on several CDs (and a colleague first received GenBank on a DAT tape, which he printed out to look at all the sequences). Now, with GenBank holding 85 billion bases of sequence from nearly one-third of a million species, the amount of DNA sequence data is virtually unlimited (or soon will be). Analysis methods, and computational capacity, have had to grow rapidly to accommodate the vast amount of data. But as the analytical techniques get more sophisticated, they also tend to incorporate more assumptions about the evolutionary processes that produced the data. So progress in bioinformatics relies not only on advances in laboratory and computational techniques, but also on increased understanding of the patterns and processes of molecular evolution.

It is critical that the advances in computation do not come at the expense of biological veracity. For example, the increasing size of sequence datasets has brought a growing reliance on automatic alignment programs that, although constantly improving in sophistication, sometimes result in biologically unrealistic arrangements of some sequences, due to the complexity of patterns of sequence change. Some researchers claim that it is simply too time consuming to inspect alignments to detect these errors, yet if these poorly aligned regions are included in an analysis, any inference drawn from them is spurious. Just as we would be reluctant to accept sloppy laboratory techniques for the sake of expedience, we should be equally unhappy about cutting corners on the analysis. If our methods do not reflect real biological processes, we risk leading ourselves up a garden path of our own making.

Understanding patterns of genome change requires an evolutionary perspective that regards the genome as part of a whole organism. This issue includes papers that take an evolutionary approach to measuring or estimating the mutation rate, detecting and explaining differences in the rate of molecular evolution across the genome and between species, exploring how the interplay between positive selection, negative selection and drift creates complex patterns of molecular evolution, and developing realistic evolutionary models for analysing DNA to uncover evolutionary history and contemporary patterns of biodiversity. These papers share a common theme: that we need to combine molecular evolutionary theory with empirical measurement of the patterns and rates of molecular evolution to fully appreciate the complexity of the information contained in the genome.

2. MEASURING AND EXPLAINING DIFFERENCES IN MUTATION RATE

Mutation rate is a key parameter for many analyses of molecular data, yet reliable estimates are available for very few species. Consequently, analyses often borrow estimates made on distantly related species, despite the fact that mutation rate can vary substantially between lineages. For example, mutation rate is a critical parameter in epidemiological modelling based on sequence data. Some of the highest known mutation rate estimates have been made on RNA viruses such as HIV. But Sanjuan *et al.* (2009) obtained experimental estimates of the mutation rate from a plant RNA virus that are much lower than the values cited for RNA viruses of animals and bacteria. They suggest that the host environment has a major effect on viral mutation rates, and so that plant RNA viruses may be under different selection pressures from animal viruses.

Experimental approaches such as these are only possible for short-lived organisms that can be grown in the laboratory over many generations. But Hendy *et al.* (2009) demonstrate how simultaneous sampling of parents and offspring from a population can be used to estimate mutation rates in longer lived organisms (in this case penguins). By modelling the fate of mitochondrial mutations passed from mother to offspring, through the sampling bottleneck of gametogenesis, they obtain estimates of the mitochondrial mutation rate similar to those obtained through ancient DNA sampling, suggesting that this method may provide a more widely applicable means of getting mutation rate estimates.

One contribution of 11 to a Special Feature on 'Whole organism perspectives on understanding molecular evolution'.

Received 24 March 2009
Accepted 27 March 2009

Developing reliable analytical methods requires an appreciation of the mutational mechanisms that underlie genome change, and how they might vary across the genome and between species. For example, one of the most pervasive influences on rates of molecular evolution is the number of times the sequence is replicated per unit time (see Bromham 2009). But Elango *et al.* (2009) demonstrate that, in some primate species, rates of divergence at CG dinucleotides are more 'clock-like' than at other sites in the genome. They infer that mutations at these CpG sites, which are a byproduct of methylation, are replication independent so do not show the same patterns of lineage-specific rate variation as other sites in the genome. This insight could provide the basis of more accurate estimates of divergence rates and dates between different lineages.

3. SELECTION SHAPES RATES OF MOLECULAR EVOLUTION

Rates of substitution are determined by the balance between selection and drift, which can vary across the genome and between populations and lineages (see Bromham 2009). Untangling the effects of these factors requires analysis of sequences within an evolutionary framework, and an appreciation of whole organism influences on genome change. For example, Mank & Ellegren (2009) explain how increased substitution rate in genes with sex-biased expression could be interpreted as the signal of increased positive selection due to sexual selection accelerating changes to genes involved with secondary sexual characteristics. But alternatively, this pattern could be explained by decreased selective pressure on sex-biased genes, increasing the number of nearly neutral substitutions. Such a decrease in selection pressure could be due to the 'dispensability' of sex-biased genes, leading to evolutionary flexibility in gene expression patterns, or decreased pleiotropy, which reduces selective constraint on sequences. So, understanding the way that genes interact within the organism will be critical to resolving hypotheses concerning the role of selection in shaping rates of change.

Selection also plays a role in shaping genome-wide rates of molecular change. Galtier *et al.* (2009) examine the mitochondrial ageing hypothesis, which suggests that metabolism produces reactive oxygen species that degrade the mitochondrial genome, reducing the efficacy of DNA protection and repair mechanisms, thus causing an increase in the mutation rate. However, studies have failed to find a direct correlation between metabolic rate and rates of molecular evolution beyond that expected through covariation with life-history traits such as body size, generation time and longevity. Galtier *et al.* (2009) promote an alternative selection-based explanation: large, long-lived animals have undergone selection to reduce mitochondrial mutation rates in order to limit the incidence of harmful somatic mutations.

The effectiveness of selection in shaping mutation rates, and in governing the fate of mutations, is moderated by effective population size (N_e). Since N_e is affected by species' behavioural, ecological and life-history traits, it can vary substantially between lineages and over time (Bromham 2009; Woolfit 2009). This is important because the effects of population size changes on observed patterns of DNA change may otherwise be falsely interpreted as the signature of selection (Otto 2000). For example, population expansion leads to an increased number of low-frequency alleles, which could mimic the effects of a recent selective sweep, and population contraction should reduce the relative proportion of low-frequency alleles, which could be erroneously interpreted as a sign of balancing selection.

4. DO OUR ASSUMPTIONS MATCH OUR DATA?

The development of new bioinformatic methods has commonly involved an increase in complexity, through relaxation of constraints and addition of parameters. For example, models that assumed all substitutions were equally probable were progressively adapted to allow for observed biases, such as transitions being more common than transversions, some types of transversions being more common than others and the influence of base composition on substitution frequencies. Now, many researchers begin their analysis with a formal test of the best available substitution model for their dataset. Often, these tests return the most parameter-rich model tested, raising the possibility that a model with even more parameters would provide an even better fit to the data. In other words, selecting the best available model is not the same as selecting an appropriate model. Gatesy (2007) compared this with an overweight man shopping for underwear in the petite women's section of a department store: the best fit may simply not fit well enough.

Consider progress in the development of molecular dating methods for inferring evolutionary time scales from DNA sequences. Early studies were based on uniform rates in all lineages, until the extent of rate variation became apparent, for example, even among four closely related species, Elango *et al.* (2009) found substantial rate variation between species, between chromosomes, between regions of the genome (depending on whether they contained repetitive sequences or not) and between sites in a sequence. But we can rarely estimate rates directly, due to the relative paucity of independent calibrating information. So an ideal molecular dating method would predict rate changes on any branch of a phylogeny without requiring many (or any) external calibrations. This is what 'relaxed clock' methods aim to do, using a model of rate change to select the most likely set of rate changes along a phylogeny.

Relaxed clocks seem intuitively more biologically realistic than earlier dating methods. But solving a phylogeny when you allow all branches to vary in rate requires making some hefty assumptions about what kind of rate changes are most probable (see Welch & Bromham 2005). Because we do not yet fully understand how and why rates change, these models are necessarily based on best guesses and tractable statistical models. Many widely used relaxed clock models assume autocorrelated rates, such that each branch inherits its rate from its ancestral lineage.

Ho (2009) finds that while the assumption of rate autocorrelation is intuitively reasonable, given that both mutation rate and substitution rate are influenced by characteristics that are likely to be similar among close relatives, autocorrelated models currently lack empirical support. So while new molecular dating methods seem to capture more of the complexity of rate variation between lineages, until we have a better grasp of the patterns and causes of rate variation among lineages, we cannot know whether these models are adequate enough to give us reliable date estimates wherever they are applied.

5. DEVELOPING EVOLUTIONARY MODELS FOR DNA ANALYSIS

As more and more fields of biology and medicine come to rely on DNA analysis, there is an imperative to develop analytical methods that are based on an understanding of evolutionary principles. For example, DNA analyses are taking an increasingly important role in biodiversity studies as a way of assessing and monitoring species diversity. 'DNA barcoding' is critical for documenting microbial diversity, since the majority of bacteria cannot be cultured and identified by traditional laboratory-based observations. The growing field of metagenomics, increasingly applied in environmental monitoring (e.g. comparing microbial diversity in contaminated soil to unaffected habitats) and medicine (e.g. comparing gut floras in healthy and diseased individuals), relies on using computational analysis to assign DNA sequences from an environmental or medical sample to 'species' or functional types.

DNA barcoding typically uses threshold measures of sequence difference to define separate species (e.g. 3% difference at a given locus), but genetic distance may fail to capture realistic units of diversity, because the tempo and mode of evolution vary between lineages and across the genome (Barraclough *et al.* 2009). For example, Ettema & Andersson (2009) show that 'housekeeping genes' used in DNA barcoding, which perform basic metabolic or information roles, have very different patterns of evolution to the genes responsible for bacterial adaptation and diversification. Adaptive genes tend to be clustered into highly 'evolvable' regions, which may facilitate rapid adaptation to new niches. So the locus chosen or the barcoding method used may give different answers due to variation in the underlying evolutionary dynamics of genome change. To avoid reliance on arbitrary values of sequence divergence, Barraclough *et al.* (2009) demonstrate an alternative approach based on an evolutionary model of population divergence. A modelling approach makes assumptions explicit, which should facilitate improvement of the technique as our understanding of bacterial evolution increases.

6. UNDERSTANDING THE TEMPO AND MODE OF MOLECULAR EVOLUTION IS CRITICAL TO DNA ANALYSIS

The assumptions underlying new analytical methods can have an overwhelming effect on the results (Welch & Bromham 2005). While new, evermore complex models seem to be an improvement on previous, simpler methods, we cannot have confidence that the results we are getting are more reliable until we have an appreciation of how adequately the assumptions reflect the data. So there is much work to be done.

One way forward is to use empirical studies of molecular evolution to help us to hone the assumptions underlying the computational methods. This special issue of Biology Letters contains a diversity of approaches to taking a whole-organism view of the forces that shape molecular evolution, revealing patterns more rich and complex than previously imagined. These empirical results will inform the development of new and better bioinformatics analyses.

Thanks to Meg Woolfit, Simon Ho, Matt Phillips, Rob Lanfear and Katie Byron for their valuable input and helpful discussions. I am extremely grateful to Fiona Pring for her admirable coordination of this Special Feature.

Barraclough, T. G., Hughes, M., Ashford-Hodges, N. & Fujisawa, T. 2009 Inferring evolutionarily significant units of bacterial diversity from broad environmental surveys of single-locus data. *Biol. Lett.* **5**, 425–428. (doi:10.1098/rsbl.2009.0091)

Bromham, L. 2009 Why do species vary in their rate of molecular evolution? *Biol. Lett.* **5**, 401–404. (doi:10.1098/rsbl.2009.0136)

Elango, N., Lee, J., Peng, Z., Loh, Y.-H. E. & Yi, S. V. 2009 Evolutionary rate variation in Old World monkeys. *Biol. Lett.* **5**, 405–408. (doi:10.1098/rsbl.2008.0712)

Ettema, T. J. G. & Andersson, S. G. E. 2009 The α-proteobacteria: the Darwin finches in the bacterial world. *Biol. Lett.* **5**, 429–432. (doi:10.1098/rsbl.2008.0793)

Galtier, N., Jobson, R. W., Nabholz, B., Glémin, S. & Blier, P. U. 2009 Mitochondrial whims: metabolic rate, longevity and the rate of molecular evolution. *Biol. Lett.* **5**, 413–416. (doi:10.1098/rsbl.2008.0662)

Gatesy, J. 2007 A tenth crucial question regarding model use in phylogenetics. *Trends Ecol. Evol.* **22**, 509–510. (doi:10.1016/j.tree.2007.08.002)

Hendy, M. D., Woodhams, M. D. & Dodd, A. 2009 Modelling mitochondrial site polymorphisms to infer the number of segregating units and mutation rate. *Biol. Lett.* **5**, 397–400. (doi:10.1098/rsbl.2009.0104)

Ho, S. Y. W. 2009 An examination of phylogenetic models of substitution rate variation among lineages. *Biol. Lett.* **5**, 421–424. (doi:10.1098/rsbl.2008.0729)

Mank, J. E. & Ellegren, H. 2009 Are sex-biased genes more dispensable? *Biol. Lett.* **5**, 409–412. (doi:10.1098/rsbl.2008.0732)

Otto, S. P. 2000 Detecting the form of selection from DNA sequence data. *Trends Genet.* **16**, 526–529. (doi:10.1016/S0168-9525(00)02141-7)

Sanjuan, R., Agudelo-Romero, P. & Elena, S. F. 2009 Upper limit mutation rate estimation for a plant RNA virus. *Biol. Lett.* **5**, 394–396. (doi:10.1098/rsbl.2008.0762)

Welch, J. J. & Bromham, L. 2005 Molecular dating when rates vary. *Trends Ecol. Evol.* **20**, 320–327. (doi:10.1016/j.tree.2005.02.007)

Woolfit, M. 2009 Effective population size and the rate and pattern of nucleotide substitutions. *Biol. Lett.* **5**, 417–420. (doi:10.1098/rsbl.2009.0155)