Measurably evolving populations

Alexei J. Drummond[1,2], Oliver G. Pybus[2], Andrew Rambaut[2], Roald Forsberg[1] and Allen G. Rodrigo[3]

[1]Department of Statistics, University of Oxford, South Parks Road, Oxford, UK, OX1 3TG
[2]Department of Zoology, University of Oxford, South Parks Road, Oxford, UK, OX1 3PS
[3]School of Biological Sciences and Allan Wilson Centre for Molecular Ecology and Evolution, University of Auckland, Private Bag 92019, Auckland, New Zealand

The availability of nucleotide and amino acid sequences sampled at different points in time has fostered the development of new statistical methods that exploit this temporal dimension. Such sequences enable us to observe evolution in action and to estimate the rate and magnitude of evolutionary processes through time. Populations for which such studies are possible – measurably evolving populations (MEPs) – are characterized by sufficiently long or numerous sampled sequences and a fast mutation rate relative to the available range of sequence sampling times. The impact of sequences sampled through time has been most apparent in the disciplines of RNA viral evolution and ancient DNA, where they enable us to estimate divergence times without paleontological calibrations, and to analyze temporal changes in population size, population structure and substitution rates. Thus, MEPs could increase our understanding of evolutionary processes in diverse organisms, from viruses to vertebrates.

Microevolutionary processes – those that can produce genetic changes in populations observable on a 'human timescale' – have been intensively studied, at least since the pioneering theoretical work of the population geneticists R.A. Fisher, J.B.S. Haldane and Sewall Wright. Their models, describing the action of genetic drift, selection and migration on populations of reproducing individuals, have been complemented by numerous studies of natural and experimental populations. In the fields of evolution and ecology, population genetics has been central to the investigation of molecular variation within and between populations, and to the ongoing development of statistical methods based on microevolutionary principles [1,2].

There has also been considerable progress in our understanding of molecular phylogenetics, in which evolutionary trees spanning much longer periods of time are inferred from DNA or protein sequence variation [3–5]. Whereas the study of phylogenetics is implicitly based on the principles of microevolution, the timescales are such that the modeling of individuals in populations is not generally considered. Although both fields owe much to the neutral theory of molecular evolution [6], the unification of population genetics and phylogenetics has only recently begun (reviewed in [7]), largely driven by the advent of genealogy-based population genetics (coalescent theory) [8–10].

Research in both of these fields has generally assumed that no significant accumulation of mutations occurs during the time over which molecular sequences are sampled. Therefore, mutations have typically been considered as strictly historical events (i.e. as changes on ancestral lineages that pre-date the sampling period). However, there is an inherent limit to how much can be understood from single snapshots of genetic diversity. Here, we consider populations from which molecular sequences are sampled over hundreds or thousands of generations, so that mutations accumulate during the sampling period and cannot be ignored when modeling the microevolutionary process. We call sequences sampled over long periods heterochronous, whereas, those sampled at effectively the same time, we term isochronous (Box 1). We define measurably evolving populations (MEPs) as populations from which molecular sequences can be taken at different points in time, among which there are a statistically significant number of genetic differences. MEPs are characterized by either a high mutation rate, or a wide range of sequence sampling times (Table 1). However, detecting significant evolutionary change in populations that lack these properties might be feasible if long or numerous sequences are available. Therefore, our definition is an operational one, depending both on the availability of experimental resources and the properties of the population and genomic region under investigation. Our definition leads us to recognize two primary sources of MEPs, given the limitations of today's technology, namely: (1) RNA viruses; and (2) well characterized vertebrate subfossil material from which ancient DNA can be reliably amplified. Here, we review the use of MEPs in diverse fields, including phylogeography, medical genetics and molecular evolutionary theory, and outline the role of the MEP concept in future exploration of evolutionary process.

Theoretical concepts

Most statistical methods in phylogenetics and population genetics have been developed to analyze isochronous sequences; for example, tree-building methods, such as UPGMA, phylogenetic tests of the molecular clock hypothesis [3] and population genetic methods based on coalescent theory [8–11]. However, such approaches should not be directly applied to MEPs for two important reasons: (1) methods that fail to incorporate the temporal

Corresponding author: Alexei J. Drummond (alexei.drummond@zoo.ox.ac.uk).

http://tree.trends.com 0169-5347/$ - see front matter © 2003 Elsevier Ltd. All rights reserved. doi:10.1016/S0169-5347(03)00216-7

Box 1. Power to detect mutations between heterochronous sequences

Consider a genetic locus (X) that is evolving at a constant rate through time. On each of two occasions, an individual is randomly sampled from the study population at locus X. The expected difference in the number of mutations that have accumulated along the two lineages descended from the most recent common ancestor (MRCA) of the sequences is $\delta = \mu t$, where t is sampling interval, and μ is the number of mutations per locus per unit time. Although the value of δ can be estimated, there is no guarantee that any estimate of δ will be significantly different from zero. The power of any statistical test of the null hypothesis $\delta = 0$ depends on the observed number of mutations along both lineages (x and y in Fig. Ia); these values in turn depend on the true values of δ and Θ, where Θ is the expected number of mutations accumulated by both lineages up to the first sampling occasion. If H_0 is the null hypothesis that $\delta = 0$, the power of a test of H_0 is given by Eqn I:

$$\text{Power} = \Pr(\text{Reject } H_0 | \delta, \Theta)$$

$$= \sum_{x=0}^{\infty} \sum_{y=0}^{\infty} \Pr(\text{Reject } H_0 | x, y, \delta, \Theta) \Pr(x|\delta, \Theta) \Pr(y|\Theta) \qquad \text{[Eqn I]}$$

Assuming a simple Poisson model, and using likelihood ratio tests, the power to detect values of δ statistically different from zero is plotted in Fig. Ib. For any given value of Θ, power varies as a function of δ, with lower values of δ giving lower power. Similarly, if Θ is high, there is little power to infer a non-zero δ. This illustrates why, for instance, mitochondrial sequences obtained only a few years apart can effectively be treated as isochronous: the relative values of δ and Θ mean that δ is statistically indistinguishable from zero. Therefore, not all populations enable us to distinguish between isochronous and heterochronous sequence samples. Those that admit such distinctions can be said to evolve measurably; that is, our estimate of δ enables us to discriminate between two sequences sampled at different times. Consequently, although all populations evolve, only some evolve measurably.

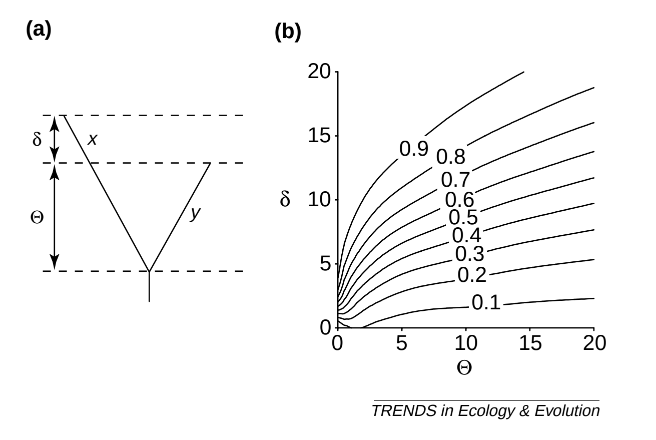


Fig. I.

structure in heterochronous sequences will suffer from unpredictable biases in statistical inference and hypothesis testing; and (2) correctly incorporating the temporal structure inherent in heterochronous sequences increases the statistical power of methods used to infer evolutionary processes, particularly when those processes change through time. Unfortunately, neither of these issues can be resolved by simply splitting heterochronous data into subsets of approximately isochronous sequences that are subsequently treated as independent observations, because individuals from subsequent generations share genetic history with individuals in the recent past, even in rapidly evolving virus populations. Therefore, the analysis of MEPs requires that the temporal structure of the data be fully integrated into the phylogenetic and population genetic models used (Boxes 2–4). Inferences based on these models are commonly obtained using maximum likelihood (ML) [12–15] or Bayesian statistical frameworks [16,17].

The concept of the molecular clock [18] is fundamental to the investigation of MEPs, because it connects the temporal information contained in the sampling times to the genetic similarities embedded in the sequences. Although the assumption of a constant and universal rate of molecular evolution is a powerful tool for understanding molecular variation, it is obviously a simplification. For example, in a recent comprehensive study of heterochronous RNA virus sequences that used the method described in Box 2, only seven out of 50 virus species did not reject the hypothesis of a constant-rate molecular clock [19]. Yet in the same study, simulations showed that, if the variation in evolutionary rate is located

Table 1. Measurably evolving population data sets[a]

Organism	Data set	Locus	Sampling interval, τ (years)	Sequence length (□)	Estimated mutation rate, μ (site^{-1} y^{-1})	δ_{min} (= $\mu\tau$L)	Sequences	Refs
Human influenza A	H3N2	*HA1*	13	987	5.7×10^{-3}	73.1	254	[66]
HIV-1	p11	*env*	8.33	660	1.02×10^{-2}	56.2	39	[14]
Dengue-4	–	*E*	38	1485	7.91×10^{-4}	44.6	17	[15]
PRRSV (European)	–	*ORF3*	8	747	5.8×10^{-3}	34.7	44	[26]
HIV-1	p1	*env*	6.42	660	4.53×10^{-3}	19.2	77	[14]
HIV-1	–	*envV3*	7	300	6.7×10^{-3}	14.1	13	[49]
HIV-1	Pre–treatment	*env*	0.59	660	2.27×10^{-2}	8.8	28	[17]
HIV-1	–	*p17gag*	7	430	2.7×10^{-3}	8.1	13	[49]
Brown bear	30 + 17 modern	*HVR1*	~59000	195	4.3×10^{-7}	4.9	47	[b]
Hepatitis C	–	*E1*	17	297	7.89×10^{-4}	4.0	20	[50]
Adelie penguin	–	*HVR1*	~6500	326	9.3×10^{-7}	2.0	100	[30]
Hepatitis C	–	*NS5*	17	219	4.98×10^{-4}	1.9	23	[50]

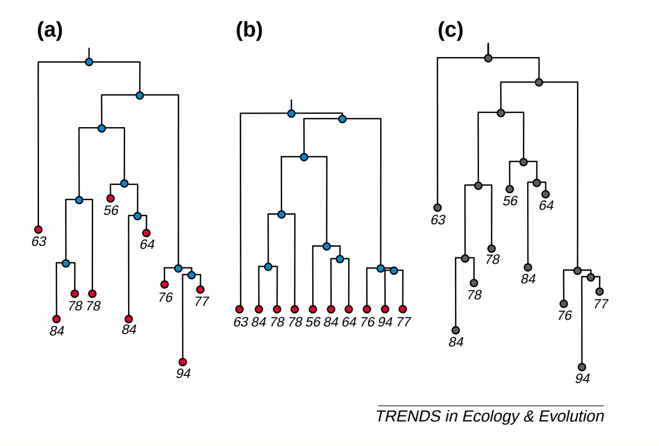
[a] The table is ordered by δ_{min}, which represents a very conservative lower bound on the expected number of mutations over the sampling interval, assuming only a single lineage spans the interval. In conjunction with an estimate of Θ and the power analysis in Box 1, δ_{min} can be used to obtain an approximate bound on the power to reject the null hypothesis of no mutations over the sampling interval. For some of these data sets (e.g. Dengue-4) the total age of the tree is less than twice the sampling interval, so that most of the evolution of the sequences sampled has actually occurred since the first sample was taken.

[b] A.Drummond, PhD thesis, University of Auckland, 2002.

Box 2. The molecular phylogenetics of heterochronous sequences

Sequences sampled at different points in time (heterochronous data) can be used to estimate both the ancestral relationships of those sequences and the rate at which they evolve [15]. The ancestry is represented by a genealogy (*G*), comprising a tree topology, and the unknown times of the internal nodes (Fig. Ia). The expected number of mutations on each branch is equal to the time represented by that branch multiplied by the overall mutation rate, μ. Given *G*, μ and a suitable model of evolution (denoted Φ), standard phylogenetic algorithms [3] are used to calculate the likelihood of the sequence data *D*, $\Pr\{D|G,\Phi,\mu\}$. The unknown ancestral aspects of the genealogy (the topology and ancestral node times) together with the mutation rate and the other model parameters can be estimated using either maximum likelihood (ML) or Bayesian methods. In practice, the tree topology component is often fixed to decrease the computation time.

μ scales the branch lengths of the tree from units of calendar time (months or years) into units of genetic distance (expected number of mutations per site). As the mutation rate gets smaller, or the times of the internal nodes of the genealogy get older relative to the range of sampling times, the sampling times of the sequences become increasingly unimportant and the isochronous model (Fig. Ib) becomes a reasonable approximation. If the ML estimate of μ does not fit the data significantly better than a model where $\mu = 0$ (as assessed using a likelihood ratio test), then we cannot reject the hypothesis that the data are isochronous and therefore the sampled population is not measurably evolving [15]. A further comparison can be made between the molecular clock models (Fig. Ia,b) and a general model that only estimates the product of time and mutation rate for each branch, thereby ignoring the time information (Fig. Ic) [3]. This latter model does not assume that the mutation rate is constant; thus, if the heterochronous model is a significantly worse fit to the data than is the (unconstrained) general model, then we must reject the assumption that μ is constant. This represents an extension of the original molecular clock test [3] to accommodate heterochronous data.



TRENDS in Ecology & Evolution

Fig. I.

in only a few lineages, then a single rate estimated using the molecular clock assumption is an accurate measure of the average rate and can be used as a reliable timescale. However, if rate variation among lineages is pervasive [16,20–22] or the rate of evolution changes through time [12,23], then this variation cannot be ignored and must be incorporated into the models used. Heterochronous sequences can provide sufficient statistical power to test directly hypotheses regarding the tempo and mode of molecular evolution [12,15,24–26]. By contrast, isochronous sequences alone do not enable mutation rate and time to be separately estimated, and therefore require external calibration dates before such hypotheses can be investigated.

Measuring evolution with ancient DNA

The amplification of mitochondrial DNA (mtDNA) from subfossil bone material at least 60 000 years old [27,28] has become increasingly reliable, provided that the cold conditions that are a requirement for long-term DNA survival are met [27,29,30]. With the success of mtDNA recovery from late Pleistocene subfossil material, researchers are now embarking on the recovery of DNA sequences from similarly aged nuclear loci [31]. This increasing abundance of ancient DNA sequences provides an opportunity to apply the MEP concept to organisms with much lower mutation rates than those found in viruses. A recent survey of >5000 mammalian nuclear genes found an average substitution rate of 2.2×10^{-9} substitutions site^{-1} y^{-1} [32]; about five million times slower than HIV-1 envelope gene (*env*) substitution rates. However, the highly variable region 1 (HVR1), located in the control region of the mitochondrial genome, exhibits some of the fastest rates of evolution in vertebrates, $\sim 10^{-7}$–10^{-6} mutations site^{-1} y^{-1} [30,33,34], two orders of magnitude faster than the average mammalian nuclear gene. Ancient mtDNA sequences are beginning to shed light on the rate of the molecular clock in individual species and, in doing so, provide independent internal calibrations for the ages of species and populations that have diverged in the recent past.

Recently, mtDNA recovered from well preserved bones underlying Adelie penguin *Pygoscelis adeliae* colonies across the Antarctica were used to estimate the rate of evolution of their HVR1 [30] using the methods described in Box 4. The high estimated rate of 0.4–1.4 mutations site^{-1} million y^{-1} is consistent with pedigree-based estimates of the mutation rates of HVR1 in humans [33–35] and provides an internal calibration for the age of the most recent common ancestor of modern Adelie penguins, to only 40 000–120 000 years ago. In a different analysis, 30 ancient mtDNA sequences of brown bear *Ursus arctos*, ranging from 10 000 to >59 000 years old, were used to illuminate the temporal and spatial movements of brown bear in the Beringian peninsula during the geologically and climatically tumultuous period of the late Pleistocene [27]. These data demonstrate that, by sampling extinct lineages and ancestral diversity, heterochronous sequences can provide historical information that is otherwise unattainable from modern sequence genealogies. Likewise, recent studies of ancient mtDNA from the extinct cave bear *Ursus spelaeus* have attempted to reconstruct temporal and spatial trends in Pleistocene phylogeography [28,36]. These examples demonstrate that DNA sequences from ancient sources can provide insights into the rate of evolution and the timings of specific past ecological and phylogeographical events. Conversely, independent estimates of evolutionary rate might suggest

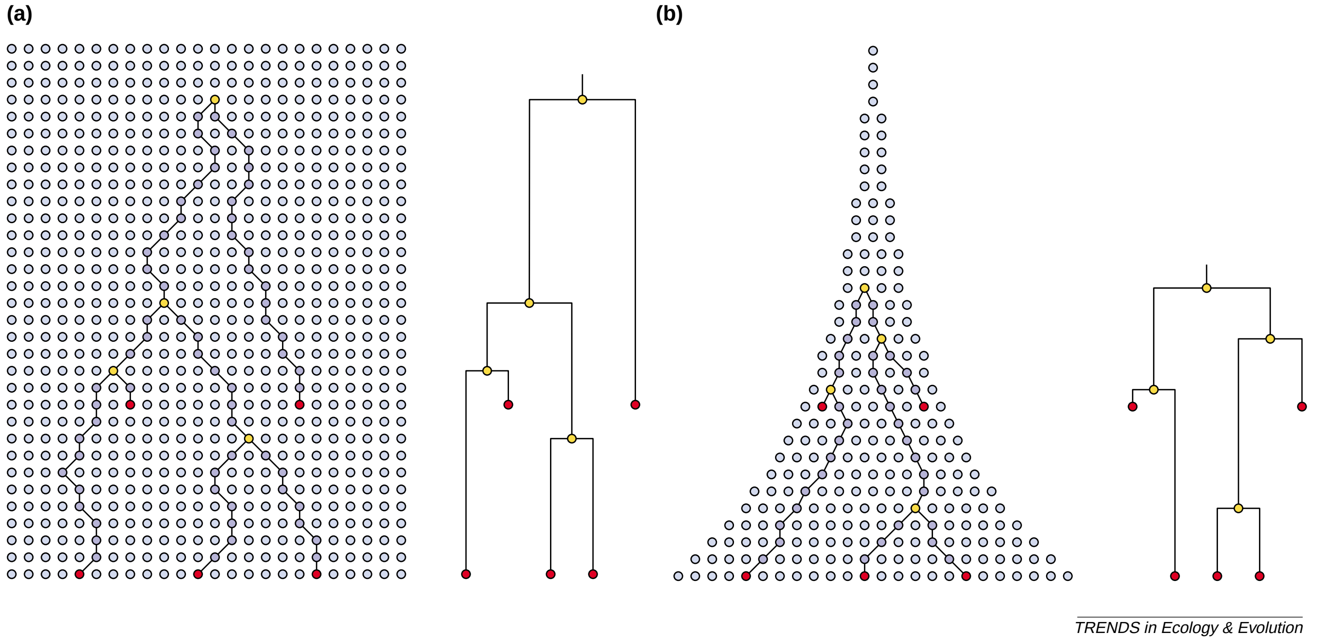
Box 3. The coalescent and measurably evolving populations

Molecular sequences, whether sampled simultaneously or serially through time, can be used to reconstruct the demographic history of natural populations. This approach is based on a population genetic model called the coalescent, introduced by Kingman [8] and generalized by Griffiths and Tavaré [9]. Although these papers consider populations sampled at one time point, the coalescent model has been subsequently extended to measurably evolving populations by Rodrigo and Felsenstein [46].

The coalescent describes the relationship between the demographic history of a large population and the shared ancestry of individuals randomly sampled from it, as represented by a genealogical tree. This tree, in turn, determines the pattern of genetic diversity seen in sampled sequences (Box 2). Figure I illustrates the shared ancestry of individuals sampled from constant-sized (a) and exponentially growing populations (b), respectively. Moving back in time from the present, we follow the number of lineages in the genealogy in each generation. This value decreases when two lineages share a common ancestor (a coalescence event), and increases when sampled individuals are encountered (a sampling event). Because the probability that a coalescence event occurs at a particular time is inversely proportional to the population size at that time, the pattern of observed coalescence and sampling events can be used to estimate the demographic history of the population.

The coalescent model provides a probability distribution of times between the coalescence events in the sample genealogy. This distribution depends on a demographic function (e.g. constant size, exponential growth or logistic growth) that describes population-size change, and the likelihood of this function can be calculated given a specified genealogy. Hence, the demographic function can be estimated from gene sequences by combining the phylogenetic and coalescent likelihood functions, as described in Box 4.

The coalescent is a variable process, so it often produces estimates with large confidence limits. Statistical power is increased by the use of sequences from multiple unlinked loci (as they represent multiple independent runs of the coalescent process), or by the use of sequences sampled at different times. The greater the spread of sampling times of sequences, the more precise the estimates of population size and substitution rate will be [23]. Finally, heterochronous sequences contain information about mutation rate (Box 2), so the demographic function can be estimated in calendar time units (generations or years).



TRENDS in Ecology & Evolution

Fig. I.

that heterochronous samples taken from a population should be measurably evolving. This expectation has recently been used to challenge the validity of claims of successful isolation of DNA from exceptionally ancient sources of bacteria [37,38].

Previously, phylogenetic calibrations of the rate of evolution within species have not been possible because of the difficulty in assigning fossils to specific lineages at this taxonomic level. As a result, most estimates of the rate of molecular evolution have been at the level of genera or higher. Independent estimates based on ancient DNA will not only give us a more detailed picture of variation in evolutionary rates across species and populations, but will also help test the validity of recent methods designed to estimate the variation in rates across lineages [16,20–22].

Measuring the evolution of RNA viruses

As a group, RNA viruses encompass such well known pathogens as HIV, influenza and foot and mouth disease, and are characterized by populations that continuously generate huge numbers of mutations owing to their large numbers, very short generation times and the error-prone nature of their replication machinery [39,40]. Some of these mutations are carried to fixation by random genetic drift or by the strong directional selection exerted by host immune responses, resulting in a very fast rate of substitution of the order of 10^{-3} substitutions site^{-1} y^{-1} [19]. This rate is a million-fold greater than that observed in eukaryotes [41], so that samples of RNA viruses showing measurable evolution can be obtained from short sequences (~300 nucleotides) sampled over short time intervals (~1 year). As a consequence of this wealth of

Box 4. Statistical inference of MEPs using Bayesian MCMC

Often, we wish to estimate parameters such as the mutation rate (μ) and are uninterested in the actual genealogy (G) of the sampled sequences. Suppose we wish to estimate two fundamental population genetic parameters: μ, and (time-scaled) effective population size (θ). The joint probability density of μ and θ is (Eqn I):

$$P(\mu, \theta|D) = \int_{G,\Phi} \Pr\{D|G, \Phi, \mu\} f(G|\theta) h(\theta, \Phi, \mu) \qquad \text{[Eqn I]}$$

$\Pr\{D|G, \Phi, \mu\}$ is the phylogenetic likelihood described in Box 2, $f(G|\Phi)$ is the coalescent distribution described in Box 3, and $h(\theta,\Phi,\mu)$ is the prior distribution of the parameters involved. The integral in Eqn 1 can be approximated using Metropolis-Hastings Markov chain Monte Carlo (MCMC) methods [17]. The MCMC algorithm returns a representative sample of parameter values, given the sequence data and their sampling times. Highly probable genealogies contribute the most to the overall parameter estimates. Thus, the error bars on our estimated parameters take into account our uncertainty in G. In the above example, the genealogy and other model parameters (Φ) are treated as nuisance parameters.

MCMC is typically implemented as a random walk over the space of all possible genealogy and parameter combinations. After an initial period known as burn-in (Fig. Ia), the number of times the algorithm visits any particular tree is proportional to the probability of that tree given the data. At each step, MCMC proceeds by proposing a new set of parameter values (of which the tree topology is one) and then either accepting or rejecting the newly proposed state using the Metropolis–Hastings criterion [64,65]. Figure Ib shows a plot of the parameters μ and θ in the early stages of an MCMC chain, and Fig. Ic shows the marginal posterior density of a complete run after burn-in is removed. In theory, it is easy to implement an MCMC algorithm for complex models as long as the model can be simulated. However, there is a price for this simplicity: in practice, the Markov chain random walk can get stuck in suboptimal minima because of inefficiencies in the design of the random walk moves. Theory tells us that it will get out eventually, but the samples that we actually obtain from a run of finite length on a particular data set might be misleading. For this reason, it is necessary to make careful checks of MCMC output [17].

MCMC has become popular because it enables us to use realistic models on reasonably large data sets and it also lends itself to Bayesian inference, in which prior information can be incorporated into the analysis [5].

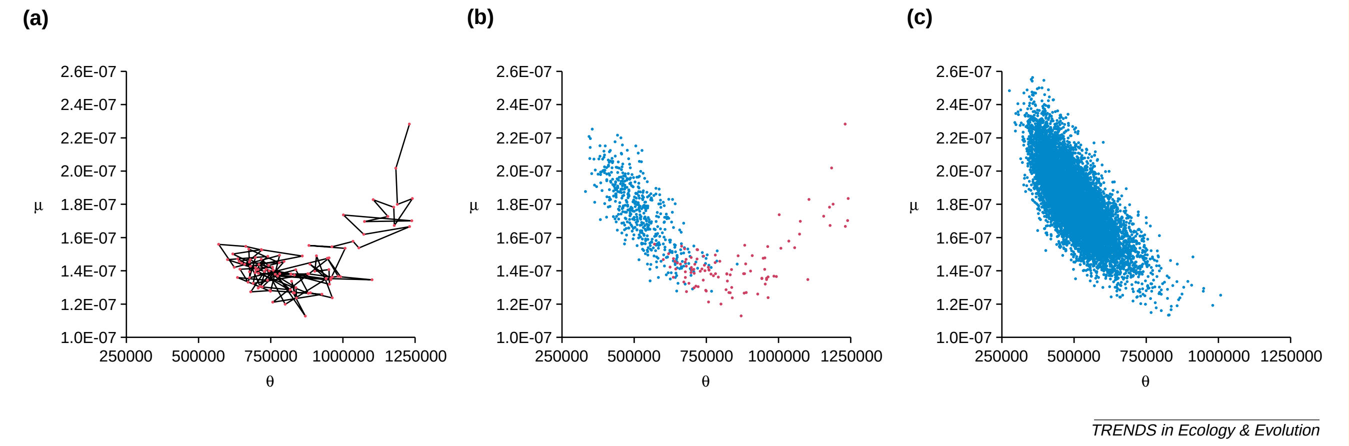


Fig. I.

measurably evolving data sets (Table 1), studies of the molecular evolution of RNA viruses have sparked the earliest use and development of methods for the analysis of heterochronous sequences [42–44].

A further consequence of the high mutation rate in RNA viruses is that the population of viruses within a single host can undergo evolutionary changes over the course of an infection, making the intrahost population both of practical medical interest as well as a source of insight into molecular evolutionary processes. It therefore makes sense to distinguish between studies that are performed on the larger interhost population and those on the smaller intrahost population scale. The dynamics and evolution of intrahost populations are naturally of great interest in long-term chronic infections, as caused by HIV and the hepatitis B and C viruses. For HIV-1, studies using heterochronous intrahost samples have been used to answer questions about the correlation between viral population dynamics and evolution. As a recent example of this, the difference in genetic divergence between heterochronous samples of latent (non-replicating) and actively replicating viruses has been used to identify reservoirs of dormant HIV viruses [45]. The parameters for this simulation-based work were taken from a seminal intrahost study, in which Shankarappa *et al.* [24] studied the rate of substitution and genetic diversity in the C2–H5 region of the *env* gene using heterochronous sequences from nine different patients, and showed how these values changed through disease progression in concert with immunological parameters.

More specific models of rate change in viral populations have been implemented by Drummond *et al.* [12] and Seo *et al.* [23]. The first of these implemented a model that allows for a temporal change in the substitution rate of the whole intrahost viral population. When applied to heterochronous intra-patient HIV-1 sequences from a patient undergoing antiviral therapy, this method showed a significant decrease in the rate of substitution upon the onset of antiviral therapy, implying that the antiviral drugs used were effective in significantly slowing the rate of HIV replication.

The above result was based on a single estimated genealogy, but was later confirmed by a full population genetic model that incorporates genealogical uncertainty

Table 2. Software packages available for the analysis of MEPs

Program	Availability	Methods[a]	Refs
BEAST	http://evolve.zoo.ox.ac.uk/beast/	Bayesian MCMC estimation of rates, dates, trees and demographic parameters from isochronous or heterochronous molecular sequences	
GENIE	http://evolve.zoo.ox.ac.uk/software/	Parametric and non-parametric estimation of demographic history from heterochronous sequence trees	[67]
MEPI	http://www.cebl.auckland.ac.nz/mepi/	Bayesian MCMC estimation of rates, dates, trees and demographic parameters from isochronous or heterochronous molecular sequences	[17]
MULTIDIVTIME	Request from author: thorne@statgen.ncsu.edu	Multilocus MCMC method for co-estimation of rates and divergence times using ML trees	[16]
PAML	http://abacus.gene.ucl.ac.uk/software/paml.html	ML phylogeny-based estimation of rates and dates, including ancestral state reconstruction and codon models	[68]
R8S	http://ginger.ucdavis.edu/r8s/	Parametric, semi-parametric and non-parametric rate smoothing and divergence time estimation of a tree, including those generated from isochronous or heterochronous data	[69]
RHINO	http://evolve.zoo.ox.ac.uk/software/	ML phylogeny-based estimation of rates and dates, including models of local clocks	
TIPDATE	http://evolve.zoo.ox.ac.uk/software/	ML estimation of rate, model parameters and divergence times	[15]

[a]Abbreviations: MCMC, Markov chain Monte Carlo; MEP, measurably evolving populations; ML, maximum likelihood.

[17]. This analysis is an example of the application of coalescent theory [8,9,46] (Boxes 3,4) to the study of intra-patient HIV-1 evolution. Coalescent models incorporating heterochronous sequences have also provided genetic estimates of viral generation time that correspond well with estimates from models of viral and immune system dynamics [14,47,48]. Furthermore, an interesting correlation between substitution rate and population size in intra-patient populations has been found, hinting at the importance of negative selection in shaping HIV dynamics [14].

The dynamics of intra- and interhost populations have been linked in a study by Leitner and Albert [49]. Heterochronous sequences were sampled from several HIV-1-infected individuals connected by a known transmission history. By using the time information in the data to calibrate a molecular clock, the authors compared the time of the most recent common ancestor of different intra-host populations with the known times of transmission. This comparison showed that two viruses transmitted simultaneously from a single donor could be substantially genetically diverged, indicating that transmitted viruses are drawn from a large population of viruses.

At the interhost population level, a typical source of heterochronous virus sequences are molecular epidemiological studies conducted over many years in an effort to monitor trends in the distribution of viral genetic diversity. The temporal information in heterochronous sequences enables phylogenies to be calibrated in a calendar timescale from which the timing of epidemiological events can be inferred and, in conjunction with coalescent theory (Box 3), enables past epidemiological dynamics to be estimated (e.g. [50,51]). As illustrated below, this is of particular importance in the study of emerging viral diseases, for which genetically estimated dates of evolutionary events can be compared with relevant epidemiological information to suggest the sources and transmission routes of new pathogens.

Influenza A virus is a continuously emerging disease in humans owing to recurrent transfer of new antigenic variation from avian influenza viruses [52]. The dramatic human influenza A outbreak known as the Spanish influenza pandemic killed >30 million people between 1918 and 1919. To understand the order and timing of species transmission events leading to this pandemic, Gorman *et al.* [53] used heterochronous samples of viruses to show that the virus responsible for the pandemic was a newly introduced avian virus that entered the human population shortly before the initiation of the pandemic in 1918. These findings have been supported by sequences of influenza genes obtained from tissue of victims of the 1918 pandemic, preserved in formalin and permafrost [54–56]. In conjunction with the fatal outcome of recent transmissions of avian viruses to humans, they highlight the constant risk of new pandemics arising from close contact between humans and birds [57].

Another example of the use of temporal information in studying emerging diseases comes from the porcine reproductive and respiratory syndrome virus (PRRSV), an agriculturally important virus that appeared in Western Europe around 1990. Owing to the absence of antibodies against PRRSV in historical blood samples from pigs, it was originally believed that the emergence of this disease resulted from a single species transfer from an unknown host to the pig population immediately before the onset of the epidemic. However, by using heterochronous sequences to date a PRRSV phylogeny, it was shown that the virus must have been transferred to pigs long before the onset of the present epidemic [26]. Together with the structure and timing of geographical groupings in the phylogeny, this finding prompted the search for a PRRSV reservoir in Eastern European pig populations. Here, great PRRSV diversity was discovered, indicating that these populations are indeed the most probable source of the transmission to Western Europe [58].

This discussion represents only a few of the biological questions concerning RNA virus evolution that have been addressed and raised using heterochronous data. The widespread accessibility of this type of data continues to stimulate methodological developments and research into MEPs.

Prospects

Measurably evolving populations provide an opportunity to ask questions about population dynamics and molecular evolution that are otherwise inaccessible using isochronous sequences. All populations accumulate mutations over time, but whether we treat a population as a MEP will depend on the amount of temporally related information in the data obtained. Given knowledge about the biological properties of a population, the MEP concept can guide us in designing sampling strategies suitable for a specific analysis [23]. Furthermore, we can assess whether the MEP concept is of importance for a particular population (Box 1) and, if not, what changes in methodology or understanding might make it so.

A solid theoretical basis for developments in this area, based on coalescent theory and likelihood models of molecular evolution, has been put in place in recent years (Boxes 2–4). However, the methods described here are still limited by several simplifying assumptions. Substantial population subdivision, recombination or selection can adversely affect the analysis of heterochronous sequences, as will biases arising from the non-random sampling of individuals from the study population. Extensions of the current ML and Markov chain Monte Carlo (MCMC) inference frameworks that incorporate recombination, selection and migration are needed. These processes fall squarely within the purview of population genetics and are already understood in the context of isochronous sequences [9,59–63]. The incorporation of these effects into the MEP framework should greatly improve and extend the range of analyses possible. The methods outlined here, and their descendents, will thus assist in answering fundamental questions about the tempo and mode of molecular evolution.

We have attempted to outline the present state of and the future prospects for the study of MEPs. Because MEPs potentially span the continuum of population genetic and phylogenetic timescales, they beg the question of whether the evolutionary process is indeed the same over short and long time frames. Whether the microevolutionary process can be successfully extended to macroevolutionary timescales remains central but unanswered. Additionally, we add the question of whether accurate detection and tracking of transiently selected mutations is possible with MEPs. We hope that continued development and application of the MEP concept will lead to answers to these and other fundamental evolutionary questions.

Several software packages now exist for analysing MEPs (Table 2). Links to these and other resources are available from http://evolve.zoo.ox.ac.uk/mep/.

Acknowledgements

This work was funded by EPSRC and MRC (A.J.D., R.F.), grant 21–02–0206 from the Danish Natural Science Research Council (R.F.), The Wellcome Trust (O.G.P.) and The Royal Society (A.R.). The authors acknowledge helpful comments from Geoff Nicholls and two anonymous referees.

References

1 Berthier, P. *et al.* (2002) Likelihood-based estimation of the effective population size using temporal changes in allele frequencies: a genealogical approach. *Genetics* 160, 741–751

2 Wang, J. and Whitlock, M.C. (2003) Estimating effective population size and migration rates from genetic samples over space and time. *Genetics* 163, 429–446

3 Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17, 368–376

4 Swofford, D.L. *et al.* (1996) Phylogenetic inference. In *Molecular Systematics* (Mable, B.K., ed.), pp. 407–514, Sinauer Associates

5 Huelsenbeck, J.P. *et al.* (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294, 2310–2314

6 Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*, Cambridge University Press

7 Arbogast, B.S. *et al.* (2002) Estimating divergence times from molecular data on phylogenetic and population genetic timescales. *Annu. Rev. Ecol. Syst.* 33, 707–740

8 Kingman, J.F.C. (1982) The coalescent. *Stoch. Proc. Appl.* 13, 235–248

9 Griffiths, R.C. and Tavare, S. (1994) Sampling theory for neutral alleles in a varying environment. *Philos. Trans. R. Soc. Lond. Ser. B* 344, 403–410

10 Kuhner, M.K. *et al.* (1995) Estimating effective population size and mutation rate from sequence data using Metropolis–Hastings sampling. *Genetics* 140, 1421–1430

11 Hudson, R.R. (1990) Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.* 7, 1–14

12 Drummond, A. *et al.* (2001) The inference of stepwise changes in substitution rates using serial sequence samples. *Mol. Biol. Evol.* 18, 1365–1371

13 Pybus, O.G. *et al.* (2000) An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* 155, 1429–1437

14 Seo, T.K. *et al.* (2002) Estimation of effective population size of HIV-1 within a host: a pseudomaximum-likelihood approach. *Genetics* 160, 1283–1293

15 Rambaut, A. (2000) Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* 16, 395–399

16 Thorne, J.L. and Kishino, H. (2002) Divergence time and evolutionary rate estimation with multilocus data. *Syst. Biol.* 51, 689–702

17 Drummond, A.J. *et al.* (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* 161, 1307–1320

18 Kimura, M. (1987) Molecular evolutionary clock and the neutral theory. *J. Mol. Biol.* 26, 24–33

19 Jenkins, G.M. *et al.* (2002) Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *J. Mol. Evol.* 54, 156–165

20 Huelsenbeck, J.P. *et al.* (2000) A compound poisson process for relaxing the molecular clock. *Genetics* 154, 1879–1892

21 Kishino, H. *et al.* (2001) Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol. Biol. Evol.* 18, 352–361

22 Sanderson, M.J. (2002) Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol. Biol. Evol.* 19, 101–109

23 Seo, T.K. *et al.* (2002) A viral sampling design for testing the molecular clock and for estimating evolutionary rates and divergence times. *Bioinformatics* 18, 115–123

24 Shankarappa, R. *et al.* (1999) Consistent viral evolutionary changes associated with the progression of Human Immunodeficiency Virus Type 1 infection. *J. Virol.* 73, 10489–10502

25 Korber, B. *et al.* (2000) Timing the ancestor of the HIV-1 pandemic strains. *Science* 288, 1789–1796

26 Forsberg, R. *et al.* (2001) A molecular clock dates the common ancestor of European-type porcine reproductive and respiratory syndrome virus at more than 10 years before the emergence of disease. *Virology* 289, 174–179

27 Barnes, I. *et al.* (2002) Dynamics of Pleistocene population extinctions in Beringian brown bears. *Science* 295, 2267–2270

28 Orlando, L. *et al.* (2002) Ancient DNA and the population genetics of cave bears (*Ursus spelaeus*) through space and time. *Mol. Biol. Evol.* 19, 1920–1933

29 Willerslev, E. *et al.* (2003) Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science* 300, 791–795

30 Lambert, D.M. *et al.* (2002) Rates of evolution in ancient DNA from Adelie penguins. *Science* 295, 2270–2273

31 Greenwood, A.D. *et al.* (1999) Nuclear DNA sequences from late Pleistocene megafauna. *Mol. Biol. Evol.* 16, 1466–1473

32 Kumar, S. and Subramanian, S. (2002) Mutation rates in mammalian genomes. *Proc. Natl. Acad. Sci. U. S. A.* 99, 803–808

33 Sigurgardottir, S. *et al.* (2000) The mutation rate in the human mtDNA control region. *Am. J. Hum. Genet.* 66, 1599–1609

34 Heyer, E. *et al.* (2001) Phylogenetic and familial estimates of mitochondrial substitution rates: study of control region mutations in deep-rooting pedigrees. *Am. J. Hum. Genet.* 69, 1113–1126

35 Parsons, T.J. *et al.* (1997) A high observed substitution rate in the human mitochondrial DNA control region. *Nat. Genet.* 15, 363–368

36 Hofreiter, M. *et al.* (2002) Ancient DNA analyses reveal high mitochondrial DNA sequence diversity and parallel morphological evolution of late pleistocene cave bears. *Mol. Biol. Evol.* 19, 1244–1250

37 Graur, D. and Pupko, T. (2001) The Permian bacterium that isn't. *Mol. Biol. Evol.* 18, 1143–1146

38 Nickle, D.C. *et al.* (2002) Curiously modern DNA for a '250 million-year-old' bacterium. *J. Mol. Evol.* 54, 134–137

39 Reanney, D.C. (1982) The evolution of RNA viruses. *Annu. Rev. Microbiol.* 36, 47–73

40 Holland, J. *et al.* (1982) Rapid evolution of RNA genomes. *Science* 215, 1577–1585

41 Drake, J.W. and Holland, J.J. (1999) Mutation rates among RNA viruses. *Proc. Natl. Acad. Sci. U. S. A.* 96, 13910–13913

42 Buonagurio, D.A. *et al.* (1986) Evolution of human influenza A viruses over 50 years: rapid, uniform rate of change in NS gene. *Science* 232, 980–982

43 Saitou, N. and Nei, M. (1986) Polymorphism and evolution of influenza A virus genes. *Mol. Biol. Evol.* 3, 57–74

44 Li, W-H. *et al.* (1988) Rates and dates of divergence between AIDS virus nucleotide sequences. *Mol. Biol. Evol.* 5, 313–330

45 Nickle, D.C. *et al.* (2003) Evolutionary indicators of human immunodeficiency virus type 1 reservoirs and compartments. *J. Virol.* 77, 5540–5546

46 Rodrigo, A.G. and Felsenstein, J. (1999) Coalescent approaches to HIV-1 population genetics. In *The Evolution of HIV* (Crandall, K., ed.), pp. 233–272, Johns Hopkins University Press

47 Rodrigo, A.G. *et al.* (1999) Coalescent estimates of HIV-1 generation time *in vivo*. *Proc. Natl. Acad. Sci. U. S. A.* 96, 2187–2191

48 Fu, Y.X. (2001) Estimating mutation rate and generation time from longitudinal samples of DNA sequences. *Mol. Biol. Evol.* 18, 620–626

49 Leitner, T. and Albert, J. (1999) The molecular clock of HIV-1 unveiled through analysis of a known transmission history. *Proc. Natl. Acad. Sci. U. S. A.* 96, 10752–10757

50 Pybus, O.G. *et al.* (2001) The epidemic behavior of the hepatitis C virus. *Science* 292, 2323–2325

51 Twiddy, S.S. *et al.* Comparative population dynamics of the mosquito-borne flaviviruses. *Infect. Genet. Evol.* (in press)

52 Webby, R.J. and Webster, R.G. (2001) Emergence of influenza A viruses. *Philos. Trans. R. Soc. Lond. Ser. B* 356, 1817–1828

53 Gorman, O.T. *et al.* (1990) Evolution of influenza A virus PB2 genes: implications for evolution of the ribonucleoprotein complex and origin of human influenza A virus. *J. Virol.* 64, 4893–4902

54 Reid, A.H. *et al.* (2000) Characterization of the 1918 'Spanish' influenza virus neuraminidase gene. *Proc. Natl. Acad. Sci. U. S. A.* 97, 6785–6790

55 Reid, A.H. *et al.* (1999) Origin and evolution of the 1918 'Spanish' influenza virus hemagglutinin gene. *Proc. Natl. Acad. Sci. U. S. A.* 96, 1651–1656

56 Reid, A.H. *et al.* (2002) Characterization of the 1918 'Spanish' influenza virus matrix gene segment. *J. Virol.* 76, 10717–10723

57 de Jong, J.C. *et al.* (1997) A pandemic warning? *Nature* 389, 554

58 Stadejek, T. *et al.* (2002) Identification of radically different variants of porcine reproductive and respiratory syndrome virus in Eastern Europe: towards a common ancestor for European and American viruses. *J. Gen. Virol.* 83, 1861–1873

59 Neuhauser, C. and Krone, S.M. (1997) The genealogy of samples in models with selection. *Genetics* 145, 519–534

60 Bahlo, M. and Griffiths, R.C. (2000) Inference from gene trees in a subdivided population. *Theor. Popul. Biol.* 57, 79–95

61 Beerli, P. and Felsenstein, J. (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proc. Natl. Acad. Sci. U. S. A.* 98, 4563–4568

62 Fearnhead, P. and Donnelly, P. (2001) Estimating recombination rates from population genetic data. *Genetics* 159, 1299–1318

63 Kuhner, M.K. *et al.* (2000) Maximum likelihood estimation of recombination rates from population data. *Genetics* 156, 1393–1401

64 Metropolis, N. *et al.* (1953) Equations of state calculations by fast computing machines. *J. Chem. Phys.* 21, 1087–1091

65 Hastings, W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57, 97–109

66 Fitch, W.M. *et al.* (1997) Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc. Natl. Acad. Sci. U. S. A.* 94, 7712–7718

67 Pybus, O.G. and Rambaut, A. (2002) GENIE: estimating demographic history from molecular phylogenies. *Bioinformatics* 18, 1404–1405

68 Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* 13, 555–556

69 Sanderson, M.J. (2003) r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19, 301–302