

Paleoecology and coalescence: phylogeographic analysis of hypotheses from the fossil record

Mitchell B. Cruzan and Alan R. Templeton

The distribution of genetic variation among populations is determined by the contemporary and historical processes of genetic drift, gene flow and migration. Typically, population genetic data have been interpreted in the context of models that only consider the effects of contemporary processes¹; however, in recent years, methods for examining the influence of historical patterns of migration and dispersal on gene distributions have become available²⁻⁵. These approaches differ from traditional analyses of gene frequencies by integrating genealogical and distributional information to make inferences about historical patterns of gene flow¹. They have also allowed more informed interpretations of intraspecific geographic variation than was previously possible. Although these approaches are becoming more routine for the assessment of historical processes in a variety of animal groups, their

application to the analysis of genetic variation within plant taxa has been considerably less frequent^{6,7}. Here, our purpose is to review model systems and quantitative methods for phylogeographic analyses. Specifically, we focus on: (1) features of organisms that make them more likely to provide detailed information on biogeographic processes; (2) the advantages of integrating information from fossil-pollen databases for the corroboration of historical processes and for hypothesis development; and (3) the application of quantitative approaches for testing specific phylogeographic hypotheses. Our goal is to increase awareness of the potential for insights into historical processes that can be gained by integrating paleoecological information with data on the contemporary distributions of cytoplasmic genetic variation.

The fossil pollen database

The efforts of palynologists over the past century have produced substantial pollen databases for several continental areas⁸. Paleoecologists have used pollen-abundance profiles from pond or lake sediment cores to reconstruct the composition of historical vegetation and to infer patterns of migration for a variety of plant species in Europe and North America^{9,10}. These studies have contributed considerably to our understanding of (1) vegetation responses to climate change¹¹, (2) rates and patterns of dispersal of invading species¹², (3) the effects of physical

The application of principles from coalescence theory to genealogical relationships within species can provide insights into the process of diversification and the influence of biogeography on distributional patterns. There are several features that make some organisms more suitable for detailed studies of historical processes; in particular, limited dispersal, which serves to conserve the patterns of genetic variation that developed during colonization. We describe the potential benefits of studies that integrate analyses of genetic variation with information from the fossil pollen record and present recent examples of the application of quantitative methods of phylogeographic analysis.

Mitchell Cruzan is at the Dept of Ecology and Evolutionary Biology, University of Tennessee, 569 Dabney, Knoxville, TN 37996, USA (cruzan@utk.edu) and Alan Templeton is at the Dept of Biology, Washington University, St Louis, MO 63130-4899, USA (temple_a@biology.wustl.edu).

and biological factors on patterns of colonization¹², and (4) the composition and distribution of paleo-plant communities⁸.

Although palynological studies have produced a wealth of information, this approach also has its limitations. For example, pollen-abundance profiles are typically available only from wind-pollinated trees, thus little is known about the migration patterns of most plant and animal species^{7,13,14}. For pollen types that can be identified, species and varieties cannot always be distinguished from others in the same group (e.g. *Quercus*, the oaks), thus dispersal patterns of individual taxa might be obscured by pollen deposition from plants with a range of ecological requirements¹². Another limitation is the availability of appropriate sample sites and of the inferential power of the data obtained. Typically the distribution of pollen-core sites is sparse and, in some cases, sam-

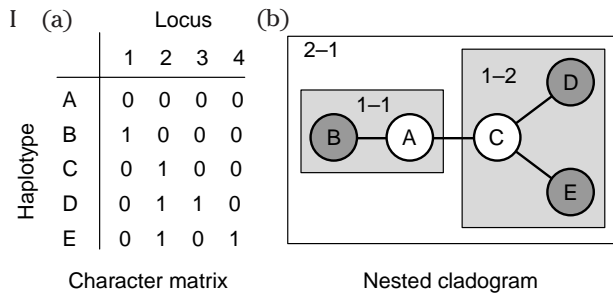
ples are completely lacking from a geographic region^{15,16}. Furthermore, the density of individuals might be too low, or a population too distant from a sample site, to be detected, thus leading to errors in estimates of the initial arrival time of a species to a region, and to mistaken inferences about the location of refugial populations and patterns of migration¹². In spite of these limitations, the analysis of pollen deposition has produced a broad sketch of postglacial vegetation dynamics that serves as an informative template for detailed phylogeographic analyses of the migration and colonization patterns of individual species.

Postglacial changes in species distributions

Changing climates and their associated glacial cycles over the past 2.4 million years have periodically fragmented many species into widely separated refugia^{8,17}. Restriction of distributions to small refugia during glacial episodes and resulting constraints on population size might lead to the loss of allelic variation¹⁸. This is particularly true for cytoplasmic variants, because the effective population size is between half and a quarter that of a diploid locus¹⁹. Consequently, populations derived from separate refugia are often characterized by different cytoplasmic haplotypes^{4,5,7,14,16,20}. At the end of the last glaciation (18 000 BP), the warming climate and the retreat of the glaciers led to the rapid migration of species out of refugial areas as they spread into previously unavailable or unsuitable habitats²¹.

Box 1. One-step haplotype phylogenies

Intraspecific sequence variation can be used to infer historical relationships among haplotypes. The hypothetical unrooted phylogeny for five haplotypes (A–E) depicted in Fig. 1 was constructed using information on allelic states at four completely linked loci (1–4). Each step (i.e. the connection between two haplotypes) in the phylogeny represents the gain or loss of a single mutation as indicated in Fig. 1a. In some instances, intermediate haplotypes might be missing from a sample (e.g. if haplotype A were missing in this example the two most closely related haplotypes, B and C, would be separated by two mutations), in which case the positions of the missing haplotypes are designated with zeros⁵. Closed loops are possible using this approach (i.e. if a haplotype is an equal number of steps from two other haplotypes), but these can often be resolved to a single connection using the geographic locations and the population frequencies of haplotypes^{5,27}. Ambiguities can also be resolved by considering the distribution of homoplasies under a model of intraspecific DNA evolution⁴⁸. The structure of the resulting cladogram indicates the hypothesized phylogenetic relationships among haplotypes. Interior haplotypes (open circles) are assumed to be ancestral to the more recently derived tip haplotypes (closed circles).



Trends in Ecology & Evolution

Once a phylogeny is adequately resolved, the relationships among haplotypes and clades are used to place them into groups for the analysis of geographic association. Haplotypes (zero-step clades) are first grouped into one-step clades starting from the tips of each branch (clades 1–1 and 1–2)⁵. Grouping proceeds by increasing the level of nesting (2-step groups and 3-step groups, etc.) until the final nesting level includes the entire phylogeny (level 2–1). The resulting nested cladogram is used to analyze the patterns of geographic association for each clade at each nesting level.

Ecological and genetic processes occurring during colonization can be substantially different than processes associated with the maintenance of populations¹⁸. In particular, empirical estimates of local dispersal rates for many species of trees are not adequate to explain the postglacial migration rates observed in the pollen record (known as ‘Reid’s paradox’²²). The hypothesis that rare dispersal events were responsible for rapid rates of range expansion was supported by theoretical analyses, which showed that an adequate fit to observed patterns of migration could be obtained if dispersal is assumed to follow a leptokurtic (‘fat tailed’) distribution^{12,22}. Analysis of these dispersal functions postulated the establishment of rare ‘pioneers’ ahead of the advancing wave of migration. These processes produce fragmented advancing fronts, thus new populations can be established as a result of dispersal from pioneer populations, as well as from populations that are part of the continuous distribution. However, note that the fragmented leading edge expected under these dispersal models would still appear as a contiguous migration front in most analyses of the pollen record, because of the threshold values of pollen numbers used to indicate the presence of each species⁸ (but see Ref. 23).

The patterns of cytoplasmic variation found in a diversity of plant and animal species appear to be consistent with expectations for the leptokurtic- or stochastic-dispersal modes. The distribution of observed cytoplasmic variation

has two features that suggest that migration proceeded through the establishment of pioneer populations ahead of the migration front.

First, reduced allele diversity across regions of recent range expansion, which are consistent with a history of repeated population bottlenecks, have been found for several species^{6,7}. For example, colonization of northern Europe by oaks (*Quercus*)¹⁴, alders (*Alnus*)¹⁴, common beech (*Fagus*)¹⁴, ragwort (*Senecio*)²⁴, pond turtles (*Emys orbicularis*)²⁵, newts (*Triturus*)¹⁴, and grasshoppers (*Chorthippus*)¹⁴ appears to have involved only a subset of the haplotypes characteristic of the southern regions. Similar patterns of loss of haplotype variation in previously glaciated regions have been documented for North American colonization events⁶.

Second, when fine-scale mapping of cpDNA haplotypes of oak species was conducted, a mosaic of haplotypes was revealed²⁶ with average patch diameters between 15 and 30 km. Such a pattern is probably the result of ‘jump’ dispersal events leading to the establishment of pioneer populations, which in turn expanded via diffusion dispersal to form a continuous distribution. The resulting mosaic pattern of haplotypes is striking and strongly suggests that stochastic dispersal processes were prevalent during the range expansion of oaks. A stochastic dispersal pattern is also supported by the fine-scale analysis of fossil pollen sites, where a fragmented leading edge of the migration front was detected for American beech^{8,23}.

Model systems for phylogeographic analyses

There are several features that render some organisms more amenable to phylogeographic study. Limited dispersal facilitates the successful inference of historical patterns of migration because it preserves the patterns of genetic variation that were created during the establishment of current distributions^{5,27}. For example, if rates of recurrent dispersal were excessively high, traces of distributional change would quickly become obscured. Once a region has been colonized the stability of phylogeographic patterns will depend on the longevity and vagility of individuals¹⁸. Thus, high levels of philopatry will tend to preserve genetic patterns that developed during colonization of a region.

Maintenance of phylogeographic patterns can be even more accentuated in some taxa (e.g. trees^{18,26} and turtles²⁵), which have high levels of philopatry as well as long life spans. The combination of restricted seed dispersal in plants²⁸ or sex-biased dispersal, which occurs in many animals²⁹, and the maternal inheritance of cytoplasmic genomes [chloroplast DNA (cpDNA) in most Angiosperms and mitochondrial DNA (mtDNA) in animals and many Gymnosperms]^{4,30} facilitates the conservation of geographic patterns of cytoplasmic genetic variation that developed as a result of historical migration. Furthermore, the use of a non-recombining cytoplasmic genome allows the reconstruction of matriarchic genealogical relationships, because new mutations are consistently transmitted along maternal lineages and cannot be exchanged among lines. Thus, with high levels of philopatry, patterns of variation in maternally inherited cytoplasmic haplotypes should reliably reflect the patterns of dispersal that occurred during postglacial recolonization of temperate habitats. By integrating information on phylogenetic relationships and geographic distributions of cytoplasmic haplotypes, the probable historical processes that led to contemporary distributions can be reconstructed.

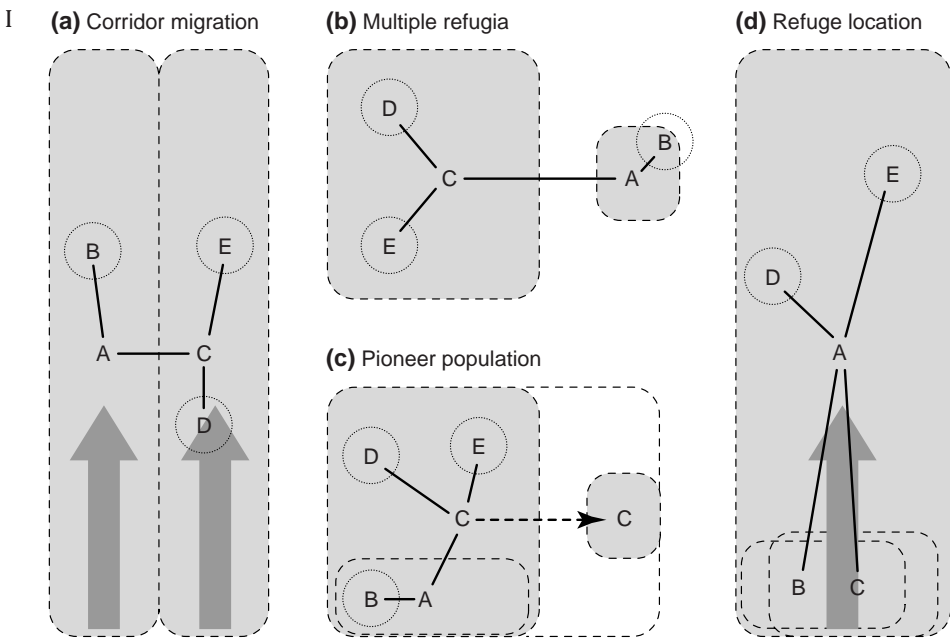
Box 3. Mining the fossil pollen database for phylogeographic hypotheses

Four examples of haplotype distribution patterns and their phylogeographic interpretation are illustrated here. In each case the phylogeny is the same as depicted in Fig. 1 of Box 1 and the distribution of the species is indicated by the shaded area.

Corridor migration: probable paths of migration during and after the last glaciation have been hypothesized for several plant and animal species^{7,8,17}. Northward migration from isolated locations would be expected to result in distinct regions, each of which is defined by one or more haplotypes that is derived from a single refugium (Fig. 1a). Haplotypes occurring within each region should all be more closely related to each other than to haplotypes in different regions, thus producing a conclusion of 'past fragmentation' from a nested-clade analysis^{5,27}. Further evidence for distinct corridors would be clade distributions that are elongated along the probable migration axis and might be bounded by geographic barriers.

Long-distance dispersal events: cases in the pollen record, where pioneer populations appear to have been established well ahead of the migrating front, have been hypothesized to be the consequence of rare long-distance dispersal events⁸. Alternatively, these disjunct populations might have been derived from previously unknown refugia¹⁷. If the latter were the case, then populations in the disjunct and primary ranges should display more distant haplotype relationships (past fragmentation; Fig. 1b). However, if the disjunct population shared haplotypes with populations from the primary range then the long-distance dispersal hypothesis would be supported (Fig. 1c).

Location of refugia: with leptokurtic or stochastic dispersal, refugial regions would be apparent from the presence of clusters of closely related haplotypes, one or a few of which are distributed across the expanded range. Nested-clade analysis should reveal a set of interior haplotypes with overlapping and restricted distributions, and with large displacements from the clade center (Fig. 1d). Identification of refugia using qualitative criteria, similar to the ones described for stochastic dispersal, have been made in several recent studies^{15,16,34,41}.



Trends in Ecology & Evolution

disjunct population could have originated either from long distance dispersal or via migration from a cryptic refugium (Box 3)^{8,17}. As a case in point, the occurrence of Scots Pine (*Pinus sylvestris*) in Scotland was originally interpreted as an example of long distance dispersal³⁹. However, recent analysis of mtDNA variation indicates that Scottish populations are probably derived from a separate refugium⁴⁰. There are numerous similar opportunities in the paleoecological literature where an analysis of cytoplasmic genetic variation could provide additional information that might help discriminate among alternative hypotheses.

The other major area where phylogeography can provide information not easily gleaned from analyses of sediment cores is the examination of patterns of distributional change in species that are not well represented in the fossil record. The analysis of cytoplasmic genetic variation in a larger variety of species would provide a more comprehensive view of the composition of historical communities and a clearer picture of the migration and distributional changes for species that represent a diversity of trophic levels and life history characteristics. For exam-

be conducted for any one-step phylogeny by comparing the levels of dispersion for tip versus interior haplotypes and clades^{5,33}. Interior haplotypes are expected to have higher levels of dispersion because they are older than their corresponding tip haplotypes and have had more time to become geographically widespread²⁷. Once it has been determined that the distribution of haplotypes is geographically structured, the one-step phylogeny can be organized into nested cladograms for the analysis of geographic associations (Box 2)^{5,27}.

Statistical analyses of haplotype distributions can discriminate among a broad range of historical processes and the benefit of these investigations can be enhanced further by integrating phylogeographic patterns with paleoecological information (Box 3). In many cases, phylogeographic studies allow hypotheses that cannot be addressed using traditional paleoecological approaches to be tested. For example, a

phylogeographic analyses for a variety of plant and animal taxa have provided evidence of three primary Pleistocene refugia for the contemporary European biota^{7,14,41}. For species with sparse or nonexistent fossil records, historical changes in associated vegetation have been used to infer the distribution of refugia and patterns of range expansion [e.g. black bears (*Ursus americanus*)^{14,16}, several Australian lizards²⁰, and bark beetles (*Ips typographus*)⁴²].

In several cases the presence or importance of Pleistocene refugia has been questioned^{16,41}. This is particularly true for arctic environments where the fossil evidence is sparse^{8,17}. It has been suggested that some coastal regions remained ice free because they were on the lee side of mountain ranges (coastal refugia) or that mountain peaks protruded above the glaciers (nunataks), potentially providing sufficiently moderate climates to allow the persistence of some arctic species¹⁷. Although candidate species, which

might have survived in such environments, have been proposed¹⁷, there have been few tests for the existence of these refugia^{6,15,16,43}. In cases of known refugia, phylogeographic analyses of species with poor fossil records would provide more accurate information on refugial communities and their environments. The compilation of phylogeographic information from a set of species from the same region would serve to enhance our understanding of the distribution of refugia, the environmental conditions present in each refuge, and the severity of population bottlenecks that species might have suffered during the last glaciation.

Future prospects

For many studies in comparative and population biology, a knowledge of historical patterns of migration and dispersal can provide insights into contemporary distributions of genetic and phenotypic variation. Phylogeographic analyses can help determine the origins of hybrid zones^{33,44} and geographic variation within species^{4,6,7,14}. A knowledge of the distribution of variation in cytoplasmic markers will allow more informed assessments of genetic structure for nuclear loci. For example, estimating gene flow across a set of populations from separate Pleistocene refugia might produce misleading results because populations on either side of the suture zone might not have had time to approach equilibrium levels of genetic exchange⁴⁵. Knowledge of population history might also provide useful information for management decisions under various conservation scenarios^{46,47}. Phylogeographic analyses can provide insights into the historical processes responsible for endemic distributions, the origins of invasive species and the identification of regions that have retained substantial levels of genetic variation.

The strength of quantitative methods of phylogeographic analysis lies in their objective approach to the assessment of historical patterns of migration and dispersal. Studies of cytoplasmic genetic variation can be further enhanced by integrating paleoecological data on the location and ages of fossils. In this respect the fossil pollen record is particularly useful because it provides detailed information on the historical distributions of plant taxa. Limited dispersal and their utility as indicators of associated animal taxa make many plant species particularly good candidates for detailed phylogeographic analyses. Further studies of these and other highly philopatric taxa will continue to provide valuable information on broad scale biogeographic patterns and on the ecological and genetic processes associated with historical environments.

Acknowledgements

We would like to thank R. Baucom, J. Estill, A. Morris, C. Murren and two anonymous reviewers for their comments on earlier drafts of this article.

References

- Neigel, J.E. (1997) A comparison of alternative strategies for estimating gene flow from genetic markers. *Ann. Rev. Ecol. Syst.* 28, 105–128
- Crandall, K. and Templeton, A. (1993) Empirical tests of some predictions from coalescent theory with applications to intraspecific phylogeny reconstruction. *Genetics* 134, 959–969
- Hudson, R.R. (1990) Gene genealogies and the coalescent process. *Oxford Surv. Evol. Biol.* 7, 1–44
- Avise, J.C. (1994) *Molecular Markers, Natural History, and Evolution*, Chapman & Hall
- Templeton, A.R. (1998) Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. *Molec. Ecol.* 7, 381–397
- Soltis, D.F. *et al.* (1997) Chloroplast DNA intraspecific phylogeography of plants from the Pacific Northwest of North America. *Plant Syst. Evol.* 206, 353–373
- Taberlet, P. *et al.* (1998) Comparative phylogeography and postglacial colonization routes in Europe. *Molec. Ecol.* 7, 453–464
- Delcourt, H.R. and Delcourt, P.A. (1991) *Quaternary Ecology: A Paleoeological Perspective*, Chapman & Hall
- Huntley, B. and Webb, T., III, eds (1988) *Vegetation History*, Kluwer Academic Publishers
- Davis, M.B. (1983) Quaternary history of deciduous forests of Eastern North America and Europe. *Ann. Mo. Bot. Gard.* 70, 550–563
- Whitlock, C. and Bartlein, P.J. (1997) Vegetation and climate change in northwest North America during the past 125 kyr. *Nature* 388, 57–61
- Bennett, K.D. (1986) The rate of spread and population increase of forest trees during the postglacial. *Philos. Trans. R. Soc. London Ser. B* 314, 523–531
- Cain, M.L. *et al.* (1998) Seed dispersal and the holocene migration of woodland herbs. *Ecol. Monogr.* 68, 325–347
- Hewitt, G.M. (1999) Post-glacial re-colonization of European biota. *Biol. J. Linn. Soc.* 68, 1–2
- Tremblay, N.O. and Schoen, D.J. (1999) Molecular phylogeography of *Dryas integrifolia*: glacial refugia and postglacial recolonization. *Molec. Ecol.* 8, 1187–1198
- Byun, A.S. *et al.* (1999) Coastal refugia and postglacial recolonization routes: a reply to Demboski, Stone and Cook. *Evolution* 53, 2013–2015
- Pielou, E.C. (1991) *After the Ice Age*, The University of Chicago Press
- Hewitt, G.M. (1996) Some genetic consequences of ice ages, and their role in divergence and speciation. *Biol. J. Linn. Soc.* 58, 247–276
- Petit, R.J. *et al.* (1993) Finite island model for organelle and nuclear genes in plants. *Heredity* 71, 630–641
- Schneider, C. and Moritz, C. (1999) Rainforest refugia and evolution in Australia's wet tropics. *Proc. R. Soc. London B* 266, 191–196
- Webb, T. and Bartlein, P.J. (1992) Global changes during the last 3 billion years: climatic controls and biotic responses. *Ann. Rev. Ecol. Syst.* 23, 141–173
- Clark, J.S. *et al.* (1998) Reid's paradox of rapid plant migration. *BioScience* 48, 13–24
- Webb, S.L. (1987) Beech range extension and vegetation history: pollen stratigraphy of two Wisconsin lakes. *Ecology* 68, 1993–2005
- Comes, H. and Abbott, R. (1998) The relative importance of historical events and gene flow on the population structure of a Mediterranean ragwort, *Senecio gallicus* (Asteraceae). *Evolution* 52, 355–367
- Lenk, P. *et al.* (1999) Mitochondrial phylogeography of the European pond turtle, *Emys orbicularis* (Linnaeus 1758). *Mol. Ecol.* 8, 1911–1922
- Petit, R.J. *et al.* (1997) Chloroplast DNA footprints of postglacial recolonization by oaks. *Proc. Natl. Acad. Sci. U. S. A.* 94, 9996–10001
- Crandall, K.A. and Templeton, A.R. (1996) Applications of intraspecific phylogenies. In *New Uses for New Phylogenies* (Harvey, P.H. *et al.*, eds), pp. 81–102, Oxford University Press
- Westoby, M. *et al.* (1997) Comparative ecology of seed size and dispersal. In *Plant Life Histories: Ecology, Phylogeny, and Evolution* (Silvertown, J. *et al.*, eds), pp. 143–162, Cambridge University Press
- Swingland, I.R. and Greenwood, P.J. (1987) *The Ecology of Animal Movement*, Clarendon Press
- Birky, C.W.J. (1976) The inheritance of genes in mitochondria and chloroplasts. *BioScience* 26, 26–32
- Schaal, B.A. *et al.* (1998) Phylogeographic studies in plants: problems and prospects. *Mol. Ecol.* 7, 465–474
- Dumolin-Lapègue, S. *et al.* (1997) An enlarged set of consensus primers for the study of organelle DNA in plants. *Mol. Ecol.* 6, 393–397
- Maskas, S.D. and Cruzan, M.B. (2000) Patterns of intraspecific diversification in the *Piriqueta caroliniana* complex in eastern North America and the Bahamas. *Evolution* 54, 815–827
- Dumolin-Lapègue, S. *et al.* (1997) Phylogeographic structure of white oaks throughout the European continent. *Genetics* 146, 1475–1487
- Wood, N. and Bidwell, J. (1996) Genetic screening and testing by induced heteroduplex formation. *Electrophoresis* 17, 247–254
- Harding, R. (1996) New phylogenies: an introductory look at the coalescent. In *New Uses for New Phylogenies* (Harvey, P.H. *et al.*, eds), Oxford University Press
- Excoffier, L. *et al.* (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA. *Genetics* 131, 479–491

- 38 Turner, T.F. *et al.* (2000) Nested cladistic analysis indicates population fragmentation shapes genetic diversity in a freshwater mussel. *Genetics* 154, 777–785
- 39 Birks, H.J.B. (1989) Holocene isochrone maps and patterns of tree-spreading in the British Isles. *J. Biogeog.* 16, 503–540
- 40 Sinclair, W.T. *et al.* (1999) The postglacial history of Scots pine (*Pinus sylvestris* L.) in western Europe: evidence from mitochondrial DNA variation. *Mol. Ecol.* 8, 83–88
- 41 Willis, K.J. and Whittaker, R.J. (2000) Paleoeecology - The refugial debate. *Science* 287, 1406–1407
- 42 Stauffer, C. *et al.* (1999) Phylogeography and postglacial colonization routes of *Ips typographus* L. (Coleoptera, Scolytidae). *Mol. Ecol.* 8, 763–773
- 43 Holder, K. *et al.* (1999) A test of the glacial refugium hypothesis using patterns of mitochondrial and nuclear DNA sequence variation in rock ptarmigan (*Lagopus mutus*). *Evolution* 53, 1936–1950
- 44 Hewitt, G.M. (1993) After the ice: *Parallelus* meets *Erythropus* in the Pyrenees. In *Hybrid Zones and the Evolutionary Process* (Harrison, R.G., ed), pp. 140–164, Oxford University Press
- 45 Whitlock, M.C. and McCauley, D.E. (1999) Indirect measures of gene flow and migration: $F_{ST} < 1/(4Nm + 1)$. *Heredity* 82, 117–125
- 46 Moritz, C. and Faith, D.P. (1998) Comparative phylogeography and the identification of genetically divergent areas for conservation. *Mol. Ecol.* 7, 419–429
- 47 Templeton, A.R. and Georgiadis, N.J. (1996) A landscape approach to conservation genetics: conserving evolutionary processes in the African Bovidae. In *Conservation Genetics: Case Histories from Nature* (Avice, J.C. and Hamrick, J.L., eds), pp. 398–430, Chapman & Hall
- 48 Templeton, A.R. *et al.* (1992) A cladistic analysis of phenotypic associations and haplotypes inferred from restriction endonuclease mapping and sequence data. III. Cladogram estimation. *Genetics* 132, 619–633
- 49 Posada, D. *et al.* (2000) A program for the cladistic nested analysis of the geographical distribution of genetic haplotypes. *Mol. Ecol.* 9, 487

Statistical methods for detecting molecular adaptation

Ziheng Yang and Joseph P. Bielawski

It has been proved remarkably difficult to get compelling evidence for changes in enzymes brought about by selection, not to speak of adaptive changes¹.

Although Darwin's theory of evolution by natural selection is generally accepted by biologists for morphological traits (including behavioural and physiological), the importance of natural selection in molecular evolution has long been a matter of debate. The neutral theory² maintains that most observed molecular variation – both polymorphism within species and divergence between species – is due to random fixation of selectively neutral mutations. Well established cases of molecular adaptation have been rare³. Several tests of neutrality have been developed and applied to real data, and although they are powerful enough to reject strict neutrality in many genes, they rarely provide unequivocal evidence for positive darwinian selection.

Most convincing cases of adaptive molecular evolution have been identified through comparison of synonymous (silent; d_S) and nonsynonymous (amino acid-changing; d_N) substitution rates in protein-coding DNA sequences, thus providing fascinating case studies of natural selection in action on the protein molecule. Selected examples are listed in Table 1; see Hughes⁴ for detailed descriptions of many case studies. Here, we summarize recent methodological developments that improve the power to detect adaptive molecular evolution, and examine their strengths

The past few years have seen the development of powerful statistical methods for detecting adaptive molecular evolution. These methods compare synonymous and nonsynonymous substitution rates in protein-coding genes, and regard a nonsynonymous rate elevated above the synonymous rate as evidence for darwinian selection. Numerous cases of molecular adaptation are being identified in various systems from viruses to humans. Although previous analyses averaging rates over sites and time have little power, recent methods designed to detect positive selection at individual sites and lineages have been successful. Here, we summarize recent statistical methods for detecting molecular adaptation, and discuss their limitations and possible improvements.

Ziheng Yang and Joseph Bielawski are at the Galton Laboratory, Dept of Biology, University College London, 4 Stephenson Way, London, UK NW1 2HE (z.yang@ucl.ac.uk; j.bielawski@ucl.ac.uk).

and weaknesses, so that they can be used to detect more cases of molecular adaptation.

Measuring selection using the nonsynonymous/synonymous (d_N/d_S) rate ratio

Traditionally, synonymous and nonsynonymous substitution rates (Box 1) are defined in the context of comparing two DNA sequences, with d_S and d_N as the numbers of synonymous and nonsynonymous substitutions per site, respectively⁵. Thus, the ratio $\omega = d_N/d_S$ measures the difference between the two rates and is most easily understood from a mathematical description of a codon substitution model (Box 2). If an amino acid change is neutral, it will be fixed at the same rate as a synonymous mutation, with $\omega = 1$. If the amino acid change is deleterious, purifying selection (Box 1) will reduce its fixation rate, thus $\omega < 1$. Only when the amino acid change offers a selective advantage is it fixed at a higher rate than a synonymous mutation, with $\omega > 1$. Therefore, an ω ratio significantly higher than one is convincing evidence for diversifying selection.

The codon-based analysis (Box 2) cannot infer whether synonymous substitutions are driven by mutation or selection, but it does not assume that synonymous substitutions are neutral. For example, highly biased codon usage can be caused by both mutational bias and selection (e.g. for translational efficiency⁶), and can greatly affect synonymous substitution rates. However, by employing parameters π_j for the frequency of codon j in the model (Box 2), estimation of