The evolution of base composition and phylogenetic inference

Arne Ø. Mooers and Edward C. Holmes

Base composition varies at all levels of the phylogenetic hierarchy and throughout the genome, and can be caused by active selection or passive mutation pressure. This variation can make reconstructing trees difficult. However, recent tree-based analyses have shed light on the forces responsible for the evolution of base composition, forces that might be very general. More explicit tree-based work is encouraged.

Arne Mooers is at the Institute for Biodiversity and Ecosystem Dynamics, and the Zoological Museum, University of Amsterdam, Amsterdam 1090GT, The Netherlands (mooers@bio.uva.nl); and Edward Holmes is at the Dept of Zoology, University of Oxford, South Parks Road, Oxford, UK OX1 3PS (edward.holmes@zoology.oxford.ac.uk).

The observation, made 50 years ago, that nuclear material contains equal amounts of the bases adenine (A) and thymine (T), and equal amounts of guanine (G) and cytosine (C), was crucial in helping to elucidate the structure of DNA. Although it could be so, the proportion of A+T in a genome rarely equals that of G+C and different organisms exhibit different patterns of base composition variation (Box 1). In parallel with discussions about the importance of these observations, in the early 1990s, phylogeneticists discovered that the variation in GC content among organisms could wreak havoc on attempts to reconstruct evolutionary history. This was because the tree-building techniques then in use often grouped together unrelated species with similar GC content. Ideally, these observations might have sparked more explicit tree-based analyses regarding the evolution of GC content in a perfect marriage of microevolutionary mechanisms and macroevolutionary comparative biology. Such a union has now begun. Here, we look at some of the general patterns of GC content, in light of the various theories for its evolution, and highlight some of the impacts these studies have had on phylogenetic inference.

General patterns of base composition variation

At the level of the entire genome, GC content varies greatly within and among major groups of organisms. Surprisingly, the best systematic compilation is still Sueoka's initial survey[1]. Bacterial species have the widest range of GC content, from 25% in *Mycoplasma capricolum* to over 75% in the gram-positive actinobacterium *Micrococcus luteus*[2], although protist and algal genomes are nearly as variable. The range in total GC content diminishes within both plants and animals, and vertebrate GC content ranges from 37% in the tuna *Thunnus thunnus* to 50% in the lamprey *Lampetra fluviatilis*[3]. *Lampetra fluviatilis* is in fact an outlier; vertebrate genomes tend to be GC poor[3], with birds and mammals having increased global GC content relative to poikilothermic vertebrates[4]. Total GC content also varies taxonomically within mitochondrial genomes, although all mitochondrial genomes sequenced to date are GC poor[5]: it is lowest in insects and nematodes (range 15–35%), somewhat higher in molluscs (range 29–40%) and highest in birds, mammals (particularly primates) and teleost fish (range 32–46%)[5].

GC content can also be highly variable (and highly structured) within genomes. Bacterial DNA seems reasonably uniform in GC content, although often with a bias as to which bases (G versus C and A versus T) are found on the leading and lagging strands – a skew (Box 1) that is often correlated with distance from the origin of replication, but not with overall GC content[6]. Mitochondrial genomes are also relatively uniform and exhibit (taxon-dependent) skews, although in this case, the amount of skew in the coding strand is strongly correlated with overall GC content[5]. Uniform GC content has been reported for the nuclear genomes of fishes and *Xenopus*[7], but this pattern seems not to extend to poikilotherm vertebrates in general[8]. By contrast, the nuclear genomes of yeast, plants, trypanosomes, birds and most mammalian orders contain isochores (Box 1)[7]. Within these taxa, isochore structure is variable; for example, pangolins, myomorph rodents, fruit bats and some insectivores seem to be lacking the most GC-rich isochores[7]. In both mammals[9] and plants[10], protein-coding genes seem to be concentrated in particular isochores; for example, in mammals they are most abundant (and perhaps also shorter) in the most GC-rich (H3) isochores.

Mutation pressure?…

Two main evolutionary processes have been invoked to explain why patterns of base composition vary within and among species: biases in the process of mutation (Box 1), such that the rates of change from G•C ↔ A•T are not constant in time or space[11]; and natural selection, either on overall GC content[4,7,12] or on specific patterns of codon usage (Box 1)[13,14].

A variety of genome studies have revealed that at least some of the variation in base composition can be pinned down to biases in the underlying process of mutation, particularly because similar GC contents are often found in regions with different functional constraints, such as exons and introns. The presence of isochores also means that common base compositions extend to large stretches of noncoding DNA. In mammals, there is an even more fundamental link between isochore base composition and the process of DNA replication (when most mutations are likely to occur). AT-rich regions, which produce Giemsa-dark G chromosome bands, appear to replicate late in the cell cycle, whereas GC-rich sequences are mainly located within the Giemsa-pale T bands that have earlier times of replication[15]. However, it is unclear exactly how replication times might affect mutational biases.

In addition, differences in rates of recombination along the genome, which seem to be correlated with GC content in mammals, might also affect patterns of mutation[16]. Finally, because mutational patterns differ among replication strands, mutational bias is also the most probable

 0169-5347/00/$ – see front matter PII: S0169-5347(00)01934-0

Box 1. Glossary

Codon usage: refers to the relative frequency with which synonymous codons are used to code for amino acids; some codons might be used more than others. GC content and codon usage are logically linked when synonymous codons differ in the proportion of G+C.
GC content: the proportion of a DNA strand, RNA strand, gene, gene region, chromosome or genome that is guanine (G) or cytosine (C) rather than adenine (A) or thymine (T) [or for RNA, uracil (U)]. If all nucleotides are used equally, the GC content would be 50%; this is often assumed when estimating phylogenetic trees.
Isochore: long DNA fragments (generally >200 kb) that show uniform base composition. Vertebrate isochores can be partitioned into several categories based on GC content. GC-poor isochores are designated L(ight), whereas H(eavy) isochores are GC-richer and come in three classes (H1, H2 and H3) indicating increasing GC content.
Mutation pressure: if we only consider two types of mutations, those that convert the bonded nucleotide pairs A•T (or T•A) to G•C (or C•G), and those that do the reverse, then mutation pressure (μ_D, where D denotes 'directional') is the relative rate per base pair of the former type. Mutation pressure $\mu_D > 0.5$ indicates a preference for GC substitutions and $\mu_D < 0.5$ indicates a preference for AT substitutions. At equilibrium the observed GC content should equal μ_D (Ref. 11).
Skew: the relative amount of G versus C or A versus T along a single strand of DNA, measured with a sliding window (running average over a number of consecutive positions, often 100 000–300 000 bp). GC skew = (G−C)/(G+C), where G and C are the frequencies of the two bases in the window measured.
Stationarity: the property of having the substitution probability matrices (representing the model of character change) remain constant throughout the tree. GC content evolution (for example, owing to changes in μ_D and/or changes in selection pressure) represents one important departure from stationarity.

explanation for the skew in GC content in mitochondria and along bacterial genomes (although it is possible that lateral gene transfer might also cause some genome heterogeneity through the importation of DNA with a 'foreign' base composition).

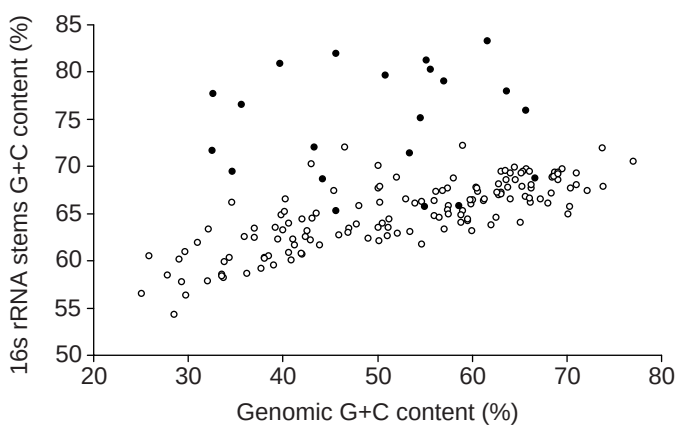

Fig. 1. The relationship between genomic GC content and GC content of the stem (double-stranded) portions of 16s rRNA for bacterial genera. Closed circles represent thermophilic genera (optimal growth temperature for 21 genera ranges from 47°C to 105°C) and open circles represent mesophilic genera (optimal growth temperature for 144 genera ranges from 17°C to 41°C). Bearing in mind that there is no relationship between genomic GC content and optimal growth temperature for either thermophiles or mesophiles, or across all bacteria[19], this graph highlights that: (1) thermophilic bacteria generally do not have higher genomic GC content than mesophilic bacteria; (2) rRNA GC content seems highly correlated with genomic GC content for mesophiles but not for thermophiles, thus suggesting a decoupling of rRNA GC and genomic GC evolution in thermophiles; and (3) rRNA GC content is much higher for thermophiles than for mesophiles, thus suggesting a selective link between temperature stress and the need for thermostable double-stranded structures in rRNA. *Modified, with permission, from Ref. 19.*

…or natural selection?

Despite observations consistent with the mutation–pressure hypothesis, the underlying genetic mechanisms, such as differences in the proteins used in DNA replication and repair, remain unknown and merit further study. Moreover, selective explanations for changes in GC content have also been proposed. For example, early work suggested that UV exposure might select for increased global GC content in bacteria[17]. A more recent survey of eight genera of bacteria suggests that strains of nitrogen (N)-fixing aerobes have higher total GC content than congeners that do not fix N, and that this might be linked to the use of more N in GC pairings[18].

The most active neutralist–selectionist debate has concerned patterns of GC evolution in vertebrates. The major selective scenario for isochore structure is based on the observation of a shared presence of GC-rich isochores, replete with high densities of protein-coding genes, in homeothermic birds and mammals. GC-rich DNA is supposed to produce a more heat-stable helix (and more stable mRNA transcripts) and thus be selectively advantageous in animals with high body temperatures[7]. This scenario can accommodate the correlated GC contents of entire isochores if the selection pressure applies to larger stretches of DNA. However, the selective advantages accrued are likely to be small. Consider also that, contrary to early suggestions, a general association between nuclear GC content and optimal temperature is not seen in bacteria[19] (Fig. 1). Combined with the observation that GC-rich *Drosophila* live in colder climes than their GC-poor relatives[20], and that cold-blooded turtles and crocodiles possess GC-rich isochores[8], one suspects that the basic premises behind the temperature hypothesis might need revision. That said, although the GC content in protein-coding genes seems unconnected with optimal temperature in bacteria, the rRNAs of thermophilic bacteria do have higher GC content than nonthermophiles (Fig. 1), thus suggesting that selection for thermal stability remains a possibility in some genes. Intriguingly, GC-rich codons also code for thermally stable, hydropathic amino acids[2]. If the genomes of different taxonomic groups deal with temperature stresses (or other assaults on integrity) differently, there might still be a link between GC content and external selective forces.

Although the power of natural selection is uncertain at the level of whole genomes, it is increasingly clear that it can make the most intricate of alterations to patterns of codon usage. In the best studied cases of *Escherichia coli*, yeast and *Drosophila*, selection has been shown to favour a set of 'preferred codons', which maximize expression levels, perhaps by increasing translational efficiency or accuracy[14,21]. Genes expressed most often are expected to be under strongest selection and, indeed, show the greatest degree of codon bias, as well as a reduced rate of synonymous substitution (because there is a reduction in the number of potentially neutral silent changes). Synonymous sites within genes expressed at lower levels will drift, such that codon usage reflects more of the underlying mutation pressure. In *Drosophila*, there is also evidence that codon bias is stronger in regions with low rates of recombination and in shorter genes, perhaps because preferred codons have a greater functional effect or because there is less interference among positions[22].

In contrast to *E. coli*, yeast and *Drosophila*, mutation pressure is the most commonly cited explanation for codon bias in mammals. Here, there is little evidence that synonymous codons differ in fitness and instead codon choice seems to reflect the mutation pressure of the local isochore,

although levels of gene expression and tRNA abundance are difficult to quantify in these species and thus selection has been proposed[12]. Why might natural selection act with less potency on codon choice in mammals? Population-genetic theory tells us that the ability of natural selection to shape Nesubstitution dynamics depends on the product of the effective population size (N_e) and the selective coefficient, s. When $N_e s \ll 1$, which might be the case for synonymous sites in mammalian species with relatively small population sizes, genetic drift will be the major force controlling substitution patterns and thus will prevent selection from efficiently fixing preferred codons. Such a process might also explain why less codon bias and higher rates of synonymous change have been observed in *Drosophila* species with restricted distributions and thus small N_e (Ref. 23).

Another way in which the effects of natural selection and mutation pressure can be teased apart is by analysing the types and frequency of nucleotide changes that are polymorphic within species. Although a variety of methods exist that consider polymorphism data[12,14], one popular approach is to compare those mutations segregating within species with the substitutions that have become fixed between them. Under neutral theory, patterns of genetic variation should be correlated within and between species. For example, mutational pressure should have the same effect on base composition whether the comparisons are conducted within or between species. Things are different under natural selection. Selectively favoured changes will be polymorphic for less time than neutral ones because they reach fixation faster, and thus such sites will show low levels of polymorphism relative to fixed differences between species[24]. Conversely, although slightly deleterious changes might segregate within populations, they are unlikely to reach fixation and thus will show higher ratios of polymorphism to fixed differences. Patterns of codon usage in *Drosophila* have been analysed using this approach. *Drosophila simulans* exhibits a low ratio of polymorphic to fixed synonymous differences, as expected if there are selectively advantageous preferred codons, although no significant difference was found in *Drosophila melanogaster*[14].

A phylogenetic perspective

Whatever the mechanistic basis, it is evident that patterns of base composition can change through time. As such, the nature and dynamics of this process can be reconstructed using phylogenetic methods. For example, many intracellular bacteria exhibit low GC contents and simple parsimony mapping revealed a major reduction in GC content soon after *Buchnera* bacteria species set up a symbiotic relationship with aphids. This was interpreted as being the result of strong mutation pressure in these species, which also have low values for N_e, rather than as a result of codon selection[25].

Unfortunately, such explicitly phylogenetic studies of GC dynamics are rare, perhaps because of concerns that changes in base composition can affect the accuracy with which we reconstruct trees in the first place. Since the early work of Saccone *et al.*, who recognized that a nonstationary substitution process could have a major bearing on estimates of genetic distance and substitution rates[26], the potentially deleterious effects of comparing genes with different base compositions have become ever more apparent (for a recent example see Ref. 27). The first tree-building methods specifically designed to tackle the issue of varying base compositions, by incorporating the most general of substitution models, were the LogDet transformation[28] (Box 2 and Fig. 2) and a closely related method that uses paralinear distances[29]. Although both techniques appeared to do the job well, they were still limited because another crucial determinant of phylogenetic accuracy – the extent of among site rate variation – was absent. However, new techniques can take into account the number of invariable sites, therefore improving their performance[30].

Another approach, that of Galtier and Gouy[31], also offers the potential to recover evolutionary trees in the face of differing base compositions and variable rates of substitution among sites. In this method, the equilibrium sequence GC content (the observed GC content expected after long periods of evolution) is allowed to vary from branch to branch in the phylogenetic tree, rather than assuming that this parameter is fixed among lineages and through time. This makes the model nonreversible (trees must have a direction, indicating how GC content varies through time) and, therefore, the ancestral GC content at the root must be estimated. Herein lies the beauty of the approach: not only can it be used for estimating phylogenetic trees but also for reconstructing the patterns of GC evolution. Therefore, it allows both phylogenetic patterns and evolutionary processes to be inferred from the same data. The model used is parameter rich (Box 2), but

Box 2. Correcting for GC content when building trees

LogDet (Ref. 28) is a transformation of observed evolutionary distances (analagous to a normalizing square-root arcsine transformation of proportions) that conforms to the basic assumptions of tree-building algorithms under a broad class of substitution models. Importantly, the transformation (based on taking the determinant of the matrix representing the proportions of every type of substitution between pairs of aligned sequences) produces additive distances, which are robust in the face of directional tendencies in different lineages towards different GC contents.

Although the LogDet transformation is robust to nonstationarity, it is also possible to construct explicit substitution models where the substitution process can be different in different lineages[31]. This incurs the cost of estimating separate parameters for separate lineages. The simplest model that incorporates a changing GC content will estimate $4n-3$ parameters when estimating a tree with n species [$=2(2n-2)+1$, representing the GC contents and branch lengths in the rooted tree, plus the GC content of the root]. By contrast, the most sophisticated time-reversible model that does not incorporate changing GC content would only estimate $2n+6$ [$=(2n-3)+9$] parameters. If among-site variation in rates of substitution is also incorporated, all models increase in complexity by the same amount. Because the total amount of data is constant, the tradeoff in all such cases is realism of the process versus increasing error because additional parameters are estimated with fewer observations.

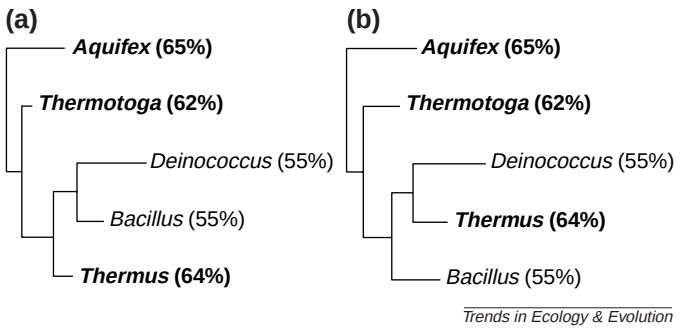

Fig. 2. The effect of GC content variation on phylogenetic reconstruction. (a) The recovered tree for 16s rRNA for three thermophilic (high GC content) and two mesophilic (lower GC content) bacteria genera, using both maximum parsimony and minimum distance methods. The total GC content of the aligned sequences are in brackets. If only variable positions are considered instead, the differences between thermophiles and mesophiles become more extreme. The two mesophiles, with similar GC content, group together. (b) The recovered tree when the log-det transformation is applied to the data. This tree, with the mesophiles separated by *Thermus*, is more in line with current views on bacterial relationships. *Figure redrawn, with permission, from T.M. Embley et al. (http://www.dbbm.fiocruz.br/james/index.html).*

simulation work suggests that meaningful estimates of ancestral GC content are possible with sequences as short as 500 bp. Indeed, for a series of trees with ten species, the approach correctly recovered ancestral GC content even in the face of global pressure to increase it through time, such that all descendent sequences had higher values than the root; or to decrease it, such that all descendent sequences had lower values than the root[31].

Application of this method to real sequence data has generated some striking results. For example, the ancestral GC content at the third position of 27 mammalian genes was found, in most cases, to more closely mirror that of *Homo sapiens* than that of murid rodents[32]. This was interpreted as evidence for an evolutionary shift towards the GC poorer, more homogeneous pattern along the murid lineage, perhaps associated with poorer DNA repair mechanisms in this lineage[33]. Similar events along other lineages of eutherian mammals might have a bearing on our attempts to accurately reconstruct the phylogenetic relationships of these animals, as will doubtless be true for many other systems.

In a second, perhaps more dramatic, application, the GC content of the rRNA sequences of extant life's most recent common ancestor was estimated to be GC poor and thus unlikely to represent a thermophile[34], thus questioning the commonly held view that life originated in a hot, auxotrophic environment. Accounting for changes in base composition among lineages is an important way to improve historical inference.

In spite of these methodological developments, it is evident that the effects of GC variation on phylogenetic inference need to be explored further. For example, the much cited simulation study of Hillis[35] showed that, under a particular model of DNA substitution, the problems associated with combining sequences with different rates of substitution ('long-branch attraction') could be partially offset by dense taxon sampling, thus dispersing homoplasic changes across the tree, particularly with large sequence data sets. Understandably, this has stimulated brute-force phylogenetic studies, incorporating vast numbers of taxa and with many genes combined (e.g. see Ref. 36 on angiosperms). However, what is unclear is how much is compromised if the genes, or even individual lineages, differ greatly in base composition, such that a single substitution model cannot be applied equally. Furthermore, the phylogenetic effects of GC differences do not just apply to genomes, DNA strands, isochores or genes, but also to neighbouring nucleotide sites[37], thus suggesting that even the most sophisticated of current tree-building methods might still be too simplistic for some aspects of molecular evolution. More generally, it is crucial that we appreciate the potential for any changing substitution process to undermine phylogenetic accuracy: we can expect lineage-specific codon preferences, changes in which bases, amino acids or which sites that can be substituted, as well as their interactions, to lead to problems[38,39]. Bias owing to parallel changes in codon bias must also be confronted[40]. We envisage that the realization that evolution is often a directional process will have a major impact on the phylogenetic studies of the future.

The causes and consequences of changes in base composition are increasingly recognized as being of major importance in molecular evolution. Not only do they provide a valuable window on fundamental evolutionary processes, but also they greatly affect our ability to accurately reconstruct phylogenetic history. The goal for the future will be the development of analytical methods that are both sensitive enough to reveal the causes of even the most minor shifts in GC content within and among lineages, and sufficiently robust to accurately retrace evolutionary history in the face of major alterations in the substitution process.

Acknowledgements

We thank N. Galtier, P. Lockhart, J. McInerney, P. Sharp and N. Sueoka for information and discussion, and R. Jenner, A. Sand and M. Steel for comments on this article. We also thank the Dutch Science Foundation (NWO) and The Royal Society for grant support.

References

1 Sueoka, N. (1964) Compositional variation and heterogeneity of nucleic acids and protein in bacteria. In *The Bacteria* (Gunsalus, I.C. and Stanier, R.Y., eds), pp. 419–443, Academic Press
2 Li, W-H. (1997) *Molecular Evolution*, Sinauer
3 Jabbari, K. *et al.* (1997) Evolutionary changes in CpG and methylation levels in the genome of vertebrates. *Gene* 205, 109–118
4 Bernardi, G. *et al.* (1997) The major compositional transitions in the vertebrate genome. *J. Mol. Evol.* 44, S44–S51
5 Saccone, C. *et al.* (1999) Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. *Gene* 238, 195–209
6 McLean, M.J. *et al.* (1998) Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *J. Mol. Evol.* 47, 691–696
7 Bernardi, G. (1995) The human genome: organization and evolutionary history. *Annu. Rev. Genet.* 29, 445–476
8 Hughes, S. *et al.* (1999) Warm-blooded isochore structure in Nile crocodile and turtle. *Mol. Biol. Evol.* 16, 1521–1527
9 Gardiner, K. (1996) Base composition and gene distribution: critical patterns in mammalian genome organization. *Trends Genet.* 12, 519–524
10 Carels, N. *et al.* (1998) Compositional properties of homologous coding sequences from plants. *J. Mol. Evol.* 46, 45–53
11 Sueoka, N. (1988) Directional mutation pressure and neutral molecular evolution. *Proc. Natl. Acad. Sci. U. S. A.* 85, 2653–2657
12 Eyre-Walker, A. (1999) Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. *Genetics* 152, 675–683
13 Ikemura, T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* 2, 13–34
14 Akashi, H. *et al.* (1998) Mutation pressure, natural selection, and the evolution of base composition in *Drosophila. Genetica* 102/103, 49–60
15 Tenzen, T. *et al.* (1997) Precise switching of DNA replication timing in the GC content transition area in the human major histocompatibility complex. *Mol. Cell. Biol.* 17, 4043–4050
16 Eyre-Walker, A.C. (1993) Recombination and mammalian genome evolution. *Proc. R. Soc. London Ser. B* 252, 237–243
17 Singer, C.E. and Ames, B.N. (1970) Sunlight ultraviolet and bacterial DNA base ratios. *Science* 170, 822–826
18 McEwan, C.E.A. *et al.* (1998) Nitrogen-fixing aerobic bacteria have higher genomic GC content than non-fixing species within the same genus. *Hereditas* 128, 173–178
19 Galtier, N. and Lobry, J.R. (1997) Relationships between genomic G+C content, RNA secondary structures, and optimal group temperature in prokaryotes. *J. Mol. Evol.* 44, 632–636
20 Rodriguez-Trelles, F. *et al.* (1999) Switch in codon bias and increased rates of amino acid substitution in the *Drosophila saltans* species group. *Genetics* 153, 339–350
21 Sharp, P.M. and Matassi, G. (1994) Codon usage and genome evolution. *Curr. Opin. Genet. Dev.* 4, 851–860
22 Comeron, J.M. *et al.* (1999) Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila. Genetics* 151, 239–249
23 Llopart, A. and Aguade, M. (1999) Synonymous rates at the RpII215 gene of *Drosophila*: variation among species and across the coding region. *Genetics* 152, 269–280
24 McDonald, J.H. and Kreitman, M. (1991). Adaptive evolution at the *Adh* locus in *Drosophila. Nature* 351, 652–654
25 Clark, M.A. *et al.* (1999) Sequence evolution in bacterial endosymbionts having extreme base compositions. *Mol. Biol. Evol.* 16, 1586–1598
26 Lanave, C. *et al.* (1984) A new method for calculating evolutionary substitution rates. *J. Mol. Evol.* 20, 86–93
27 Foster, P.G and Hickey, D.A. (1999) Compositional bias may affect both DNA-based and protein-based phylogenetic reconstruction. *J. Mol. Evol.* 48, 284–290
28 Lockhart, P.J. *et al.* (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol.* 11, 605–612
29 Lake, J.A. (1994) Reconstructing evolutionary trees from DNA and protein sequences – paralinear distances. *Proc. Natl. Acad. Sci. U. S. A.* 91, 1455–1459

30 Waddell, P.J. *et al.* (1999) Using novel phylogenetic methods to evaluate mammalian mtDNA, including amino acid invariant sites LogDet plus site stripping, to detect internal conflicts in the data, with special reference to the positions of hedgehog, armadillo and elephant. *Syst. Biol.* 48, 31–53

31 Galtier, N. and Gouy, M. (1998) Inferring pattern and process: maximum likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol.* 15, 871–879

32 Galtier, N. and Mouchiroud, D. (1998) Isochore evolution in mammals: a human-like ancestral structure. *Genetics* 150, 1577–1584

33 Mouchiroud, D. *et al.* (1997) Impact of changes in GC content on the silent molecular clock in murids. *Gene* 205, 317–322

34 Galtier, N. *et al.* (1999) A nonhyperthermophilic common ancestor to extant life forms. *Science* 283, 220–221

35 Hillis, D.M. (1996) Inferring complex phylogenies. *Nature* 383, 130–131

36 Soltis, P.S. *et al.* (1999) Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature* 402, 402–404

37 Karlin, S. and Mrazek, J. (1997) Compositional differences within and between eukaryotic genomes. *Proc. Natl. Acad. Sci. U. S. A.* 94, 10227–10232

38 Lockhart, P.J. *et al.* (1998) A covariate model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. *Mol. Biol. Evol.* 15, 1183–1188

39 Steel, M. *et al.* (2000) Invariable site models and their use in phylogeny reconstruction. *Syst. Biol.* 49, 225–232

40 Chang, B.S.W. and Campbell, D.L. Bias in phylogenetic reconstruction of vertebrate rhodopsin sequences. *Mol. Biol. Evol.* (in press)

Arbuscular mycorrhizal fungi, Collembola and plant growth

Alan Gange

Arbuscular mycorrhizal fungi are ubiquitous in field soils, as are mycophagous animals such as Collembola. It has been suggested that these animals reduce the functioning of the mycorrhiza and are thus detrimental to plant growth. However, recent choice experiments suggest that Collembola preferentially feed on nonmycorrhizal fungi in the rhizosphere. If these preferences also occur in field soils, then Collembola might indirectly benefit plants through an enhancement of mycorrhizal functioning and indirect multitrophic links to foliar-feeding insect herbivores.

Alan Gange is at the School of Biological Sciences, Royal Holloway University of London, Egham, Surrey, UK TW20 0EX (a.gange@rhbnc.ac.uk).

There is an intriguing paradox in arbuscular mycorrhizal (AM) fungal ecology. Although there are numerous laboratory studies that have shown a variety of benefits to plants in forming a mycorrhizal association, there have been far fewer occasions when these benefits have also been demonstrated in natural situations[1]. Positive effects of AM fungi on plants include enhanced nutrient uptake, protection against pests and pathogens, and relief of drought stress[2]. Recently, it has been suggested that the response of any plant to AM colonization lies along a continuum from positive to negative[3]. Therefore, the question naturally arises as to whether laboratory studies at the positive end of this continuum are realistic mimics of field situations, which appear to lie in the null (no response) area. There is an urgent need to understand this problem, so that the ecology of mycorrhizas might be better described and future management of the symbiosis improved. In laboratory studies, many factors can be controlled, and some or all of these might be responsible for reducing the efficacy of AM fungi in field conditions. These include soil nutrient levels, plant stress factors (e.g. drought), plant diseases, and herbivorous and mycophagous animals[4]. An important group in the latter category is the Collembola (springtails).

Collembola are abundant microarthropods in virtually all soils, feeding on a range of materials, including fungi, bacteria, lichens, decomposing vegetation and detritus. The feeding ecology of most species is poorly known[5], but there appears to be a preference for fungal hyphae over other food types. By consuming dead vegetation and hyphae, these animals can play an important role in decomposition processes[6]. In many cases, Collembola can enhance the decomposition process because hyphal grazing stimulates growth and respiration of the fungi[7]. The fact that most of the subterranean species feed (at least in part) on fungi has led to their being regarded as important regulators of the mycorrhizal symbiosis. Reviews of AM–soil fauna interactions suggest that Collembola have the potential to restrict mycorrhizal functioning in the field[8,9], but null and stimulative effects of their feeding have also been recorded. Other authors question their importance[4], thus it is timely to ask whether Collembola are responsible for the disruption of AM associations, and whether they are one reason for the failure of field experiments to match results obtained under controlled conditions.

Laboratory studies

It is apparent that one species, *Folsomia candida*, has been used in almost 50% of laboratory pot trials (Fig 1a). The reason is simple: *F. candida* is exceptionally easy to culture, unlike many other species. However, to quote from a recent review[10], using this species as representative of all Collembola 'is about as ecologically sound as choosing a mole as a "typical" mammal'. It is interesting that the frequency distribution of AM fungal species used shows a less skewed distribution (Fig. 1b), but this also represents species that are generally amenable to pot culture. There have been remarkably few attempts to recreate a field situation in the laboratory by using co-occurring species of Collembola and AM fungi from one field site (but see Ref. 11).

Collembola densities in pot trials are usually given in numbers of individuals per dm^3, but these have been converted to numbers per m^2 for ease of comparison with known field densities (Table 1). With few exceptions[12], the density of animals used has been at, or below, that normally encountered in comparable field situations[21]. A feature of this summary is that only the two early studies recorded a negative effect on plant growth resulting from collembolan grazing on the mycorrhiza[18,19]. In these laboratory trials, there are three

 0169-5347/00/$ – see front matter PII: S0169-5347(00)01940-6