PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Nucleic Acid and Protein Clocks [and Discussion]

Bruce Runnegar, T. A. Brown and G. A. Dover

Phil. Trans. R. Soc. Lond. B 1991 **333**, 391-397
doi: 10.1098/rstb.1991.0089

Nucleic acid and protein clocks

BRUCE RUNNEGAR

Institute of Geophysics and Planetary Physics, University of California, Los Angeles, California 90024, U.S.A.

SUMMARY

The use of pairwise comparisons of correctly aligned DNA and protein sequences for the measurement of time in historical biology remains a contentious matter. However, the limited success of some molecular evolutionary clocks provides a stimulus to attempt to improve their resolution by the judicious selection of sequences for ease of alignment, commonality of function, taxonomic breadth and appropriate rates of evolution. Existing algorithms for correcting observed distances for superimposed nucleotide substitutions or amino acid replacements appear adequate for the task, given the noise that results from the inherent variability of the process. Some possible approaches are illustrated through the use of gene and protein sequences of the large subunit of ribulose 1,5-bisphosphate carboxylase/oxygenase: an enzyme that is demonstrably homologous from purple bacteria to flowering plants.

1. INTRODUCTION

It is common knowledge that the observed similarity between any pair of correctly aligned protein or nucleic acid sequences diminishes roughly in proportion to the evolutionary distance which separates the pair. It was this apparently regular decline in pairwise similarity – first observed in aligned protein sequences – which led Zuckerkandl and Pauling to suggest that the rate of sequence evolution might be sufficiently constant for any given protein to function as a 'molecular evolutionary clock' (Zuckerkandl 1987). This fruitful idea has been applied to a variety of questions in historical biology, but it has also been roundly criticized, both for the unsupportable assumptions upon which it is based and for the way in which it performs in cases where molecular dates may be compared with those derived directly from the fossil record. There are also internal tests of consistency which suggest that the molecular evolutionary clock is frequently imprecise and unreliable.

Despite these difficulties, the molecular clock remains a viable hypothesis because it works in a limited way in some contexts and it is consistent with the neutral theory of molecular evolution (Kimura 1983). However, as neither the utility of the molecular clock nor the general applicability of the neutral theory is well established, it is useful to consider whether molecular clocks have any practical value at the present time and if they might be improved in future. For example, it may be useful to attempt to calibrate molecular clocks that are tailored to particular organisms over specified timescales in much the same way that the empirically-determined $^{87}Sr/^{86}Sr$ curve may be used to date Cainozoic marine sediments (Hess *et al.* 1986). Such 'developed' clocks would be analogous to other measures of time in geology, most of which require both the empirical determination of the rate of a process and then the correction of that rate for secular variation by means of independent observations from another system. The ^{14}C timescale for the Holocene, which has been calibrated for changes in $^{14}C/C$ over time (Bard *et al.* 1990), is an excellent example of a sophisticated clock of this kind.

2. BACKGROUND

(*a*) *How good are molecular clocks?*

Wilson & Sarich's (1969) aggressive use of limited sequence data (supported by indirect measures of genomic similarity) to estimate that the divergence between man and ape occurred as recently as about 5 million years (Ma) ago moved the molecular clock hypothesis into mainstream science. Although their date of approximately 5 Ma is now widely regarded as being approximately correct, similar kinds of comparisons made using other protein sequences or other organisms soon revealed that rates of evolution of a single protein may vary greatly in different lineages of a single clade (Ohta & Kimura 1971; Jukes & Holmquist 1972). These rate variations are most obvious in comparisons of the products of a single radiation or 'star' phylogeny (Gillespie 1984). In such cases it is fairly easy to recognize lines in which the amount of accumulated change is greater or less than average. The extant members of such lines are known as 'fast-clock' or 'slow-clock' organisms. Gillespie (1986*a*) gave an extreme example; the amino acid sequences of cytochrome oxidase subunit II from mouse and cow differ by 9.3% whereas the human protein is 27% and 29% different from its cow and mouse counterparts, respectively. Judged solely on this basis, primates are (or were) 'fast-clock' organisms compared with rodents and artiodactyls; overall, primates generally exhibit decreased rates of DNA sequence evolution (Britten 1986).

Phil. Trans. R. Soc. Lond. B (1991) **333**, 391–397
Printed in Great Britain

It is much more difficult to determine whether there have been significant changes in the rate of evolution within a lineage if each end member of a monophyletic clade has accumulated an average number of nucleotide substitutions or amino acid replacements in the time that has elapsed since the origin of the clade. Gillespie (1984) used statistical methods to infer that the rate of molecular evolution is typically episodic rather than constant in that short periods of rapid change alternate with long periods of quiescence – a model that supports the idea that molecular evolution speeds up during periods of adaptive radiation (Goodman 1981). However, if the latter were true it should be relatively easy to use sequence comparisons to find the initial branch order of classical star trees such as the radiation of modern mammals (Benton 1990). The fact that initial branch order is notoriously difficult to resolve is a clear indication that the molecules were not aware that the radiation was taking place.

There is little doubt that molecular evolution normally proceeds at an non-constant and irregular rate both within and between lineages. Provided that the number of nucleotide or amino acid substitutions can be accurately estimated for each branch of a phylogenetic tree (see below), it is possible to use a relative rate test (Sarich & Wilson 1973; Scherer 1989) to determine the potential value of a set of aligned sequences for the measurement of time. Variability may also be quantified by the statistic R_t, the ratio of the variance in the number of substitutions in a lineage to the mean number of substitutions (Gillespie 1984, 1986*a*, *b*). Ohta & Kimura (1971) showed that R_t may be significantly greater than one, the expected value for R_t if molecular evolution approximated a simple Poisson process. Gillespie (1986*a*) obtained an extreme value of $R_t = 35.5$ for the cow–mouse–human comparisons mentioned above; for many proteins, R_t lies between about 1 and 5.

To explain the values of R_t obtained from sequence comparisons, Gillespie (1984, 1986*b*) has suggested that molecular evolution follows a doubly stochastic Poisson process: a Poisson process with a rate that changes randomly on a time scale that is short (say 10^3–10^4 years) compared with the length of the lineages under study (10^7–10^8 years). A bottleneck in a population would normally reduce allelic variability but would neither increase nor reduce the amount of accumulated change in the sequence of any particular protein or gene. At higher taxonomic levels the situation may be different because the survival of a single dominant species from a group of related species need not involve any significant reduction in population size (perhaps the reverse). If the surviving species had already accumulated an unusual (non-average) amount of change, its random selection from a pool of species could be regarded as a second lottery; hence the 'doubly stochastic process'.

Given these problems, how well do molecular clocks work in practice? To a certain extent, their utility depends upon the nature of the question being asked. For example, pairwise distances obtained from the amino acid sequences of Cu/Zn superoxide dismutase from a variety of organisms were more than sufficient to reject the hypothesis that a gene for this protein had been transferred recently from the ponyfish to its endosymbiont, *Photobacterium* (Leunissen & de Jong 1986), despite the fact that the rate of evolution of superoxide dismutase in eukaryotes is notoriously variable (Lee *et al.* 1985; Ayala 1986). Similarly, Ruvolo *et al.* (1991) have successfully used the mitochondrial gene (cytochrome oxidase subunit II) which Gillespie (1986*a*) found to be so unreliable in the mouse–cow–human comparisons to tease apart and perhaps date the nodes of the human–chimp–gorilla trichotomy; a relationship that has been difficult to resolve using other methods.

Several ingenious tests of the molecular evolutionary clock have yielded both encouraging and discouraging results. For example, it is known from direct observation that the NS gene of human influenza A virus has evolved very rapidly with clock-like regularity over the past 50 years (Buonaguiro *et al.* 1986; Gojobori *et al.* 1990). Over longer timescales the results have generally been disappointing (see, for example, Romero-Herrera *et al.* (1978); Martin & Dowd (1988)) although not bad enough to reject the whole concept (Ayala 1986).

The purpose of this brief discussion is not to defend the molecular evolutionary clock, but rather to suggest how it might be possible to improve its resolution. There is clearly some temporal signal in sequence comparisons as the determination of branch order necessarily generates a relative time scale. The problem is to isolate the temporal and genealogical components of the signal and, so far as is possible, to analyse each independently.

(*b*) *Ways to improve the molecular evolutionary clock*

Although indirect measures of sequence similarity (immunological distance, restriction-fragment-length polymorphisms, DNA–DNA hybridization, etc.) are frequently used to estimate time in historical biology, the purest signal must ultimately come from complete nucleic acid and protein sequences. However, it is not so obvious that nucleic acid sequences are better than polypeptide sequences (especially for distant comparisons) because the loss of information caused by the degeneracy of the genetic code in going from DNA to protein is compensated for by the fact that there are 20 different amino acids but only four types of nucleotide. If it is also appreciated that transitions (purine–purine or pyrimidine–pyrimidine interchanges) greatly outnumber transversions (purine–pyrimidine interchanges) in most evolving DNA sequences (DeSalle *et al.* 1987), that two superimposed transitions (e.g. A→G→A or C→T→C) leave no evidence of either event (Brown *et al.* 1979), and that a transitional change in the third position of a codon is almost invariably silent, it becomes clear that correcting for superimposed nucleotide substitutions (or consequent amino acid replacements) becomes a matter of paramount importance for the calculation of evolutionary rates.

Many of the other potential sources of error may be

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

minimized through a judicious choice of data. Difficulties that might result from misaligned sequences can be eliminated if the alignments are based upon crystallographically determined three-dimensional structures (e.g. Bashford *et al.* 1987) or if the molecule exhibits only trivial differences in length over a wide taxonomic range (e.g. the large subunit of ribulose 1,5-bisphosphate carboxylase/oxygenase (RuBisCO), discussed below). Furthermore, it makes sense to edit out regions of doubtful homology and variable function from distantly related sequences. For example, if each of the bends that connects the seven lengths of α-helix in vertebrate globins is removed, one is left with 112 homologous amino acids, no gaps and a commonality of secondary structure that might be expected to even out rates of change.

We should perhaps avoid the complex enzymes that have been traditionally used for clock purposes and, instead, focus attention on those molecules that have long lengths of homogeneous secondary structure (e.g. collagens and myosins) in which the nature of any particular amino acid residue seems of little importance. It is, of course, necessary to choose molecules for which the genes are well known in order to avoid inadvertent paralogous comparisons, and to select those which are known to have an evolutionary rate appropriate for the problem under consideration.

3. RATE OF EVOLUTION OF RuBisCO

RuBisCO is a crucial enzyme for CO_2 fixation in autotrophic bacteria and the chloroplasts of eukaryotes. Because RuBisCO discriminates between the light and heavy isotopes of carbon, it is responsible for the light isotopic signature of most organic matter, and, on that basis, has been inferred to have been in existence for at least the last 3500 Ma (Schidlowski 1988).

In plant chloroplasts, RuBisCO is a complex of eight large and eight small subunits. The small subunits are coded by nuclear DNA whereas the single-copy gene for the large subunit is located in the chloroplast genome. Numerous partial amino acid sequences of the small subunit have been used to explore relationships within higher plants (Martin & Dowd 1988); the rate of evolution of the large subunit is much slower and is therefore generally unsuitable for that purpose. However, in a palaeontological context the large subunit is of great interest because of its demonstrable antiquity, the fact that it is a vital component of organisms with an extensive and good fossil record (Cyanobacteria, eukaryotic algae, land plants), and the fact that a substantial part of the DNA sequence of the large subunit gene has been reported from a magnolian leaf of Miocene age (Golenberg *et al.* 1990).

```
LATAHENSIS      CCCGTTGCTGGGGAGGAAAATCAATATATTGCTTATGTAGCTTACCCTTTAGACCTTTTT
MACROPHYLLA     CCCGTTCCTGGGGAGGAAAGTCAATTTATTGCTTATGTAGCTTACCCTTTAGACCTTTTT
TULIPIFERA      CCCGTTGCTGGGGAGACCGATCAATATATTGTTTATGTAGCTTACCCTTTAGACCTTTTT
PERSEA          CCCGTTGCTGGGGAGGAAAGTCAATTTATTGCCTATGTAGCTTACCCTTTAGACCTTTTT
BIDENTIS        CCTGTTCCTGGAGAAGACAATCAATATATTGCTTATGTAGCTTACCCATTAGACCTTTTT
PRINGLEI        CCTGTTCCTGGAGAAGACAATCAATATATTGCTTATGTAGCTTACCCATTAGACCTTTTT
COTTON          CCCGTTCCTGGAGAAGAAGATCAATATATATGTTATGTAGCTTACCCTTTAGACCTTTTT
ATRIPLEX        CCTGTTGCTGGAGAAGAAAATCAATATATTTGTTATGTAGCATATCCCTTAGACCTTTTT
PETUNIA         CGTGTTATTGGAGAAAGAGATCAATTTATTGCTTATGTAGCTTACCCTTTAGACCTTTTT
TOBACCO         CGTGTTGTTGGAGAAAAAGATCAATATATTGCTTATGTAGCTTACCCTTTAGACCTTTTT
PEA             CCTGTTCCTGGAGAAGATAATCAATTTATTGCTTATGTAGCTTATCCCTTAGACCTTTTT
ALFALFA         CCTGTTGCTGGAGAAGAGACTCAATTTATTGCTTATGTAGCTTATCCCTTAGACCTTTTT
AMARANTHUS      CCCGTTGCTGGAGAAGAAAATCAATATATTTGTTATGTAGCGTATCCTTTAGACCTTTTT
NEURACHINE      CCCGTTCCTGGGGAGGAAGATCAATATATCTGTTATGTAGCTTATCCATTAGACCTATTT
RICE            CCCGTTGTTGGGGAGGATAATCAATATATCGCTTATGTAGCTTATCCATTAGACCTATTT
WHEAT           CCTGTTGCTGGGGAAGACAGCCAATGGATCTGTTATGTAGCTTATCCATTAGACCTATTT
LIVERWORT       CCTGTTCCTGGAGAAGAAAATCAATATATTGCTTATGTAGCTTATCCTTTAGATTTATTT
CHLAMYDOMONAS   CCAGTTCCGGGTGAAGACAACCAATACATTGCTTACGTAGCTTACCCAATCGACTTATTC
EUGLENA         CCTGTTCCTGGTGAAAGTAACCAATATATTGCTTACGTGGCATATCCTATTGATCTTTTT
ANABAENA        CCAGGTCCCGGCGAAGACAACCAATCCATTGCCTACATCGCTTATCCTTTGGATCTGTTT
SYNECHOCOCCUS   CCGGTGCAAGGCGAAGAGAACTCCTACTTTGCGTTCATCGCTTACCCGCTCGACCTGTTT
CHROMATIUM      GACGTGCCGGGCGACGACACCTGCTTCTATGCCTTCATCGCCTATCCGATCGATCTGTTC
```

Figure 1. Samples of the set of aligned *rbc*L sequences analysed. Each sequence consists of 20 codons. Fully conserved nucleotides are indicated by asterisks. Note that almost all conserved sites are in the first or second positions of the codons. The nucleotides correspond to positions 445–504 as numbered in Golenberg *et al.* (1990).

RuBisCO DNA (*rbc*L gene) and protein sequences were either downloaded from GENBANK and EMBL databases using GCG software or entered directly from published sources. The 711 nucleotides (237 codons) corresponding to two lengths of DNA reported from the Miocene fossil, *Magnolia latahensis*, were assembled for the following species: *Magnolia macrophylla*, *Liriodendron tulipifera* (tulip poplar), *Persea americana* (Golenberg *et al.* 1990); *Flaveria bidentis*, *Flaveria pringlei*, *Neurachne tenuifolia*, *Atriplex rosea* (Hudson *et al.* 1990); *Gossypium hirsutum* (cotton; Gulov *et al.* 1990); *Petunia hybrida* (Aldrich *et al.* 1986*a*); *Pisum sativum* (pea; Zurawski *et al.* 1986); *Medicago sativa* (alfalfa; Aldrich *et al.* 1986*b*); *Oryza sativa* (rice; Nishizawa *et al.* 1987); *Triticum aestivum* (wheat; Terachi *et al.* 1987); *Amaranthus hypochondriacus* (Michalowski *et al.* 1990); *Chlamyodomonas reinhardtii* (Goldschmidt-Clermont & Rahire 1986); *Marchantia polymorpha* (liverwort; Fukuzawa *et al.* 1988); *Euglena gracilis* (Gingrich & Hallick 1985); *Anabaena* sp. (Curtis & Haselkorn 1983); *Synechococcus* sp. (Reichelt & Delaney 1983) and *Chromatium vinosum* (Viale *et al.* 1989). Amino acid sequences comprising 438 contiguous residues from each of these species except *M. latahensis*, *P. americana* and *F. pringlei* were analysed as well.

For the DNA sequences, pairwise differences and ratios of transversions to transitions were obtained for: (i) all 711 nucleotides; (ii) the 367 nucleotides (52%) not conserved in all 22 aligned sequences; (iii) the 230 nucleotides (32%) occurring in the third position of the 237 codons and not conserved in all 22 sequences. It was then possible to attempt to correct for superimposed substitutions in the following way.

Fitch (1986) suggested that the presence of a substantial number of invariable sites in any DNA sequence will result in an under-estimation of the amount of change that has actually taken place. As the protein-coding regions of *rbc*L genes have many nucleotides that have been conserved from bacterium to angiosperm (figure 1) over a period of at least 1500 Ma, it is reasonable to conclude that these sites are 'invariable'. Fitch (1986) and Shoemaker & Fitch (1989) have provided a method of assessing the importance of invariable sites and it is therefore possible to use this method to test the *rbc*L sequences for their amount of 'invariability'.

The ratio of transversions to transitions (d_v/d_s) is plotted against the fraction of nucleotides (d) that are observed to differ between each pair of sequences. The points obtained are then compared with theoretical curves that represent the proportion of transitional

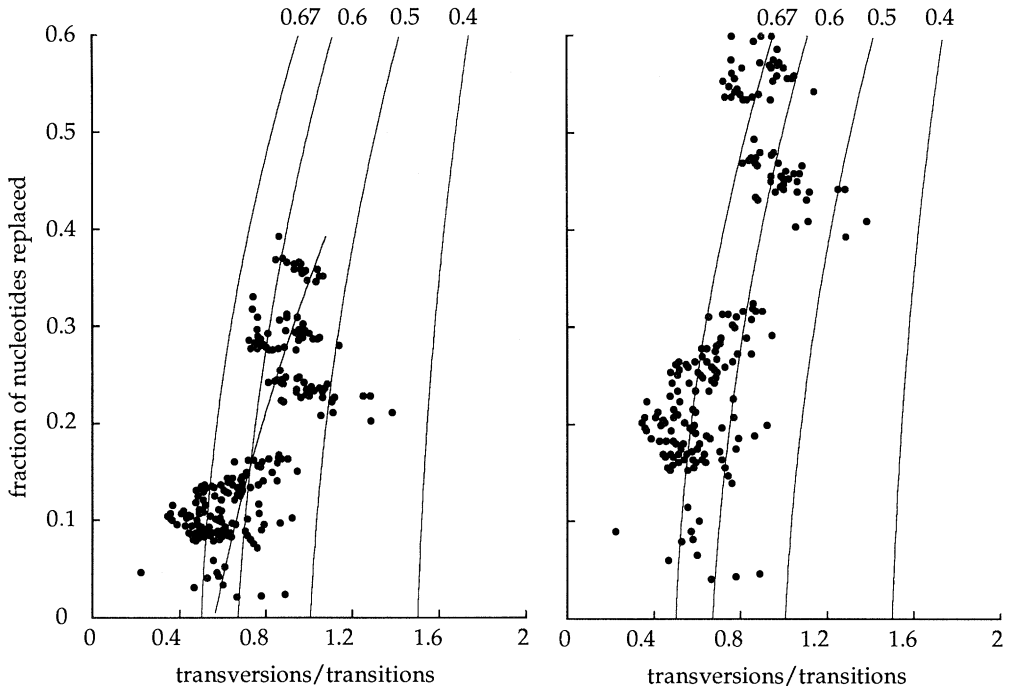

Figure 2. Fitch's (1986) graphical method of assessing the effect of invariable sites on the estimated distance between pairs of aligned sequences applied to the *rbc*L data. The four curves labelled 0.67, 0.6, 0.5, 0.4 are theoretical estimates of s, the fraction of substitutions that are transitions, and the fifth curve is a second order polynomial fitted to the plotted data (*a*). It can be seen by inspection that multiplying each value of d (the fraction of nucleotides replaced) by two will move the points so that they lie on or near the curve $s = 0.67$. The same effect is achieved by removing all 344 fully conserved nucleotides (48%) from the aligned sequences and replotting (*b*). This suggests that the presence of invariable sites significantly reduces estimates of evolutionary distance.

substitutions (s) and the rate of nucleotide substitutions per site (r). Only the former curves are used here (figure 2*a*, *b*) because it is sufficient to see the relation between the data points and s to determine the fraction of invariable sites (Fitch 1986). It can be seen in figure 2*a* that if each value of d is multiplied by two, the points will approach a distribution described by the theoretical curve $s = 0.67$. In other words, two thirds of all substitutions have been transitions in spite of the fact that the observed ratio of transversions to transitions approaches 1.0 for distant comparisons.

Much of the cause of the dispersion of the observations from the theoretical values for s is considered to be the effect of the presence of invariable sites. If the 344 fully conserved nucleotides are removed from the aligned sequences, and the new values of d against d_v/d_s are plotted, the points fall, as expected, close to the curve $s = 0.67$ (figure 2*b*). Thus the correction of d by the removal of the observed 48% of invariable sites should result in more realistic measures of the true pairwise distances. A similar result (not shown) is obtained if only the variable third position nucleotides are considered, but as this procedure reduces the sequences from 367 to 230 nucleotides in length, it probably also increases the noise level of the data.

The values for d obtained by all three methods were then corrected for superimposed substitutions using the formula of Kimura (1980). Plots of rates obtained from comparisons made from all 711 sites and from the 367 variable sites are compared in figure 3 with rates from the aligned amino acid sequences scored by means of the Dayhoff mutation-probability matrix. The calculated similarities were transformed into measures of evolutionary distance, D, by means of the relation

$$D = -\ln[(S_{\text{real}} - S_{\text{rand}})/(S_{\text{ident}} - S_{\text{rand}})] \times 100, \qquad (1)$$

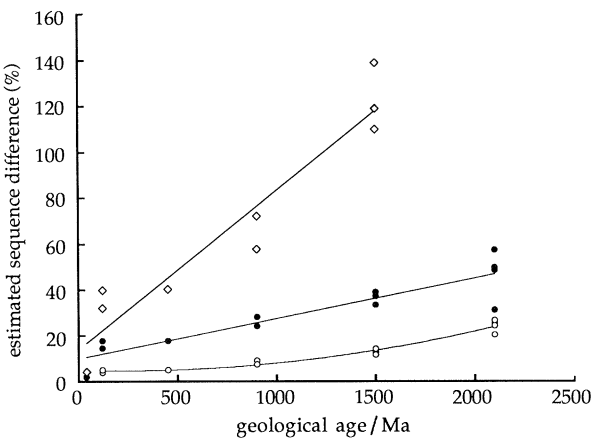

Figure 3. Rate of evolution of the large subunit of RuBisCO interpreted by using three different methods. The open circles are based upon distance scores obtained from protein sequences of the complete molecule using the Dayhoff mutation-probability-matrix as developed by Feng *et al.* (1985). The distances represented by the filled circles were calculated from the 711 nucleotides that are equivalent to the *M. latahensis* sequences, using Kimura's (1980) formula, and the lozenges are points obtained in the same way from the 367 variable nucleotides in the 711-nucleotide sequences. Divergence times are mostly based upon the fossil record as explained in the text. The lines fitted to the data were calculated by least squares regression.

where S_{real} is the observed similarity score, S_{rand} is the mean of the similarity scores obtained when sequences of the same length and composition are scrambled and aligned, and S_{ident} is the score obtained when each real sequence is aligned with itself (Feng *et al.* 1985).

The minimum divergence times, T_d, were estimated as follows:

1. The Miocene magnolian leaf, *Magnolia latahensis*, is 17–20 Ma old (Golenberg *et al.* 1990). If *M. latahensis* were the direct ancestor of the extant species *M. macrophylla*, the corrected difference between the two *rbc*L sequences ought to be equivalent to the difference expected between *M. macrophylla* and any other living species from which it diverged some 8.5–10 Ma before present (BP). T_d is halved for comparisons between *Magnolia latahensis* and living plants because *M. latahensis* RuBisCO stopped evolving 17–20 Ma BP. If the species were extant it should have accumulated additional changes in the last 17–20 Ma. In other words, the one-way distance from *M. latahensis* to a living species is equal to or greater than the two-way distance between any two contemporary species that diverged $(17–20)/2 = 8.5–10$ Ma BP. As *M. latahensis* is probably not the immediate ancestor of *M. macrophylla*, the effective value of T_d for this pair of species is $\geqslant 10$ Ma. For example, if the last common ancestor of *M. latahensis* and *M. macrophylla* lived 25 Ma BP, the effective value of T_d would be $10+5 = 15$ Ma.

2. The separation of monocot and dicot angiosperms had probably taken place before the Aptian stage of the Early Cretaceous (Taylor & Hickey 1990; 125 Ma BP, Harland *et al.* 1989); this value is used in figure 3 even though analyses of chloroplast genes (Wolfe *et al.* 1989) and glyceraldehyde-3-phosphate dehydrogenase cDNA sequences (Martin *et al.* 1989) have suggested earlier dates for the monocot–dicot divergence (200 Ma and more than 300 Ma, respectively).

3. The derivation of land plants from charophycean green algae is strongly supported by recent molecular data (Delwiche *et al.* 1989; Devereux *et al.* 1990; Manhart & Palmer 1990). On cladistic grounds, the liverworts appear to be the sister group of the clade containing the other bryophytes (hornworts and mosses) and the vascular plants (Mishler & Churchill 1985), so the split between *Marchantia polymorpha* and the angiosperms must predate the earliest vascular plants; the age of the base of the Silurian (440 Ma BP; Harland *et al.* 1989) provides a minimum date for this divergence.

4. *Chlamydomonas* and *Euglena* may be more or less equidistant from the Charophyceae (Devereux *et al.* 1990) and were assumed to be so; their time of divergence from the line leading to the higher plants is taken (without any direct evidence) to be approximately 900 Ma.

5. A minimum age for the endosymbiotic event that gave rise to the chloroplasts of green algae and higher plants is given by the occurrence of a megascopic 'alga' (*Grypania spiralis*) in rocks that are thought to be about 1300–1400 Ma old in North America, China and India (Walter *et al.* 1990; Runnegar 1991). Somewhat older dates for the origin of chloroplasts ($\geqslant 1500$ Ma, figure 3) are suggested by the occurrence of spheroidal microfossils that appear to be the remains of eukaryotic phytoplankton in strata of middle Proterozoic age (see Peat *et al.* 1978).

6. Well preserved, free-living Cyanobacteria are found in Canadian and Australian cherts that are $\geqslant 2000$ Ma old (see Hofmann 1976; Knoll *et al.* 1988); a tentative age $\geqslant 2100$ Ma is used for the separation of the Cyanobacteria from the photosynthetic purple bacteria that are the ancestors of *Chromatium* (figure 3).

4. DISCUSSION

Two things are immediately obvious from figure 3: (i) to a first approximation, the rate of evolution of the large subunit of RuBisCO has been constant for at least 2000 Ma; (ii) relatively simple corrections for superimposed substitutions in DNA sequences and superimposed amino acid replacements in equivalent protein sequences give a similar result. These results are no different in quality from those obtained previously from other molecules (e.g. cytochromes *c* and haemoglobins) except for the fact that RuBisCO evolution can be calibrated from the fossil record, albeit approximately, over a much longer timescale. The slow rate of evolution ($\sim 1\%$ change in amino acid sequence per 100 Ma) is characteristic of both chloroplast and plant mitochondrial DNA (Zurawski & Clegg 1987; Palmer & Herbon 1988). It may also be indicative of a phenomenon which J. William Schopf refers to as 'hypobradytely': an exceptionally slow rate of morphological evolution as evidenced by the fact that cyanobacterial fossils of Proterozoic age may be referred to modern lower taxa.

The rate of evolution of *rbc*L genes has increased significantly in most angiosperms; the corrected sequence difference between monocots and dicots is 14.6% (711 sites) compared with 17.8% for the much deeper angiosperm–liverwort comparison. The 2% difference between the Miocene magnolia and living *M. macrophylla* over the equivalent of, say, 15–20 Ma (see above) implies that the 15% difference between monocots and dicots is a result of some 120–150 Ma of independent evolution, as expected from the fossil record. However, not all angiosperms exhibit elevated rates; considerably lower rates have been obtained from two partial sequences and restriction-fragment-length polymorphisms of palm *rbc*L genes (Wilson *et al.* 1990).

As with other tests of the molecular evolutionary clock, this one is not conclusive. However, it suggests that slowly evolving, highly conserved sequences may be used to date distant divergences of organisms that lack a fossil record. The challenge is to select sequences that may be used in a uniform way and be calibrated by appropriate events in the history of life that are also visible in the fossil record. It is conceivable that one composite sequence, chosen for ease of alignment, taxonomic breadth and regions of slow, intermediate and fast evolutionary rate could ultimately become a standard for the measurement of time through molecular methods.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

The calculations were done with programs supplied by R. F. Doolittle and transported to the Macintosh by Peter Markiewicz.

5. REFERENCES

Aldrich, J., Cherney, B., Merlin, E. & Palmer, J. 1986*a* Sequence of the *rbc*L gene for the large subunit of ribulose bisphosphate carboxylase-oxygenase from petunia. *Nucl. Acids Res.* **14**, 9534.

Ayala, F. J. 1986 On the virtues and pitfalls of the molecular evolutionary cock. *J. Hered.* **77**, 226–235.

Bard, E., Hamelin, B., Fairbanks, R. G. & Zindler, A. 1990 Calibration of the ^{14}C timescale over the past 30,000 years using mass spectrometric U-Th ages from Barbados corals. *Nature, Lond.* **345**, 405–410.

Bashford, D., Chothia, C. & Lesk, A. M. 1987 Determinants of a protein fold. Unique features of the globin amino acid sequences. *J. molec. Biol.* **196**, 199–216.

Benton, M. J. 1990 Phylogeny of the major tetrapod groups: morphological data and divergence dates. *J. molec. Evol.* **30**, 409–424.

Britten, R. J. 1986 Rates of DNA sequence evolution differ between taxonomic groups. *Science, Wash.* **231**, 1393–1398.

Brown, W. M., George, M. & Wilson, A. C. 1979 Rapid evolution of mitochondrial DNA. *Proc. natn. Acad. Sci. U.S.A.* **76**, 1967–1971.

Buonagurio, D. A., Nakada, S., Parvin, J. D., Krystal, M., Palese, P. & Fitch, W. M. 1986 Evolution of human influenza A virus over 50 years: rapid, uniform rate of change in NS gene. *Science, Wash.* **232**, 980–982.

Curtis, S. E. & Haselkorn, R. 1983 Isolation and sequence of the gene for the large subunit of ribulose-1,5-bisphosphate carboxylase from the cyanobacterium *Arabaena* 7/20. *Proc. natn. Acad. Sci. U.S.A.* **80**, 1835–1839.

Delwiche, C. F., Graham, L. E. & Thomson, N. 1989 Lignin-like compounds and sporopollenin in *Coleochaete*, an algal model for land plant ancestry. *Science, Wash.* **245**, 399–401.

DeSalle, R., Freedman, T., Prager, E. M. & Wilson, A. C. 1987 Tempo and mode of sequence evolution in mitochondrial DNA of Hawaiian *Drosophila*. *J. molec. Evol.* **26**, 157–164.

Devereux, R., Loeblich, A. R. & Fox, G. E. 1990 Higher plant origins and the phylogeny of green algae. *J. molec. Evol.* **31**, 18–24.

Feng, D. F., Johnson, M. S. & Doolittle, R. F. 1985 Aligning amino acid sequences: comparison of commonly used methods. *J. molec. Evol.* **21**, 112–125.

Fitch, W. M. 1986 The estimate of total nucleotide substitutions from pairwise differences is biased. *Phil. Trans. R. Soc. Lond.* B **312**, 317–324.

Fukuzawa, H., Kohchi, T., Sano, T., Shirai, H., Umesono, K., Inokuchi, H., Ozeki, H. & Ohyama, K. 1988 Structure and organization of *Marchantia polymorpha* chloroplast genome. III. Gene organization of the large single copy region from *rbc*L to *trnI*(*CAU*). *J. molec Biol.* **203**, 333–351.

Gillespie, J. H. 1984 The molecular clock may be an episodic clock. *Proc. natn. Acad. Sci. U.S.A.* **81**, 8009–8013.

Gillespie, J. H. 1986*a* Variability of evolutionary rates of DNA. *Genetics* **113**, 1077–1091.

Gillespie, J. H. 1986*b* Natural selection and the molecular clock. *Molec. Biol. Evol.* **3**, 138–155.

Gingrich, J. C. & Hallick, R. B. 1985 The *Euglena gracilis* chloroplast ribulose-1,5-bisphosphate carboxylase gene. *J. biol. Chem.* **260**, 16156–16161.

Goldschmidt-Clermont, M. & Rahire, M. 1986 Sequence, evolution and differential expression of two genes encoding variant small subunits of ribulose bisphosphate carboxylase/oxygenase in *Chlamydomonas reinhardtii*. *J. molec. Biol.* **191**, 421–432.

Golenberg, E. M., Giannasi, D. E., Clegg, M. T., Smiley, C. J., Durbin, M., Henderson, D. & Zurawski, G. 1990 Chloroplast DNA sequence from a Miocene *Magnolia* species. *Nature, Lond.* **344**, 656–658.

Goodman, M. 1981 Decoding the pattern of protein evolution. *Prog. Biophys. molec. Biol.* **37**, 105–164.

Gulov, M. K., Ulmasov, T. N., Aliev, K. A., Andrianov, V. M. & Piruzian, E. S. 1990 Nucleotide sequence of the chloroplast *rbc*L gene from cotton *Gossypium hirsutum*. *Nucl. Acids Res.* **18**, 185.

Harland, W. B., Armstrong, R. L., Cox, A. V., Craig, L. E., Smith, A. G. & Smith, D. G. 1989 *A geological time scale 1989*. (263 pages.) Cambridge University Press.

Hess, J., Bender, M. L. & Schilling, J. 1986 Evolution of the ratio of strontium-87 to strontium-86 in seawater from Cretaceous to Present. *Science, Wash.* **231**, 979–984.

Hofmann, H. J. 1976 Precambrian microflora, Belcher Islands, Canada: significance and systematics. *J. Paleont.* **50**, 1040–1073.

Hudson, G. S., Mahon, J. D., Anderson, P. A., Gibbs, M. J., Badger, M. R., Andrews, T. J. & Whitfield, P. R. 1990 Comparison of *rbc*L genes for the large subunit of ribulose bisphosphate carboxylase from closely related C_3 and C_4 plant species. *J. biol. Chem.* **265**, 808–814.

Jukes, T. H. & Holmquist, R. 1972 Evolutionary clock: nonconstancy of rate in different species. *Science, Wash.* **177**, 530–532.

Kimura, M. 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. molec. Evol.* **16**, 111–120.

Kimura, M. 1983 *The neutral theory of molecular evolution*. (367 pages.) New York: Cambridge University Press.

Knoll, A. H., Strother, P. K. & Rossi, S. 1988 Distribution and diagenesis of microfossils from the lower Proterozoic Duck Creek Dolomite, Western Australia. *Precamb. Res.* **38**, 257–279.

Lee, Y. M., Friedman, D. J. & Ayala, F. J. 1985 Superoxide dismutase: an evolutionary puzzle. *Proc. natn. Acad. Sci. U.S.A.* **82**, 824–828.

Leunissen, J. A. M. & de Jong, W. W. 1986 Copper/zinc superoxide dismutase: how likely is gene transfer from ponyfish to *Photobacterium leiognathi*? *J. molec. Evol.* **23**, 250–258.

Manhart, J. R. & Palmer, J. D. 1990 The gain of two chloroplast tRNA introns marks the green algal ancestors of land plants. *Nature, Lond.* **345**, 268–270.

Martin, P. G. & Dowd, J. M. 1988 A molecular evolutionary clock for angiosperms. *Taxon* **37**, 364–377.

Martin, W., Gierl, A. & Saedler, H. 1989 Molecular evidence for pre-Cretaceous angiosperm origins. *Nature, Lond.* **339**, 46–48.

Michalowski, C. B., Bohnert, H. J., Klessig, D. F. & Berry, J. O. 1990 Nucleotide sequence of *rbc*L from *Amaranthus hypochondriacus* chloroplasts. *Nucl. Acids Res.* **18**, 2187.

Mishler, B. D. & Churchill, S. P. 1985 Transition to a land flora: phylogenetic relationships of the green algae and bryophytes. *Cladistics* **1**, 305–328.

Nishizawa, Y. & Hirai, A. 1987 Nucleotide sequence and expression of the gene for the large subunit of rice ribulose 1,5-bisphosphate carboxylase. *Jap. J. Genet.* **62**, 389–395.

Ohta, T. & Kimura, M. 1971 On the constancy of the evolutionary rate of cistrons. *J. molec. Evol.* **1**, 18–25.

Palmer, J. D. & Herbon, L. A. 1988 Plant mitochondrial

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

DNA evolves rapidly in structure, but slowly in sequence. *J. molec. Evol.* **28**, 87–97.
Peat, C. J., Muir, M. D., Plumb, K. A., McKirdy, D. M. & Norvick, M. S. 1978 Proterozoic microfossils from the Roper Group, Northern Territory, Australia. *BMR J. Geol. Geophys.* **3**, 1–17.
Reichelt, B. Y. & Delaney, S. F. 1983 The nucleotide sequence for the large subunit of ribulose 1,5-bisphosphate carboxylase from a unicellular cyanobacterium, *Synechococcus* PCC6310. *DNA* **2**, 121–129.
Romero-Herrera, A. E., Lehmann, H., Joysey, K. A. & Friday, A. E. 1978 On the evolution of myoglobin. *Phil. Trans. R. Soc. Lond.* B **283**, 61–163.
Runnegar, B. 1991 Precambrian oxygen levels estimated from the biochemistry and physiology of early eukaryotes. *Global Planet. Change.* (In the press.)
Ruvolo, M., Disotell, T. R., Allard, M. W., Brown, W. M. & Honeycutt, R. L. 1991 Resolution of the African hominoid trichotomy by use of a mitochondrial gene sequence. *Proc. natn. Acad. Sci. U.S.A.* **88**, 1570–1574.
Sarich, V. M. & Wilson, A. C. 1973 Generation time and genomic evolution in primates. *Science, Wash.* **179**, 1144–1147.
Scherer, S. 1989 The relative-rate test of the molecular clock hypothesis: a note of caution. *Molec. Biol. Evol.* **6**, 436–441.
Schidlowski, M. 1988 A 3,800-million-year isotopic record of life from carbon in sedimentary rocks. *Nature, Lond.* **333**, 313–318.
Shoemaker, J. S. & Fitch, W. M. 1989 Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. *Molec. Biol. Evol.* **6**, 270–289.
Taylor, D. W. & Hickey, L. J. 1990 An Aptian plant with attached leaves and flowers: implications for angiosperm origins. *Science, Wash.* **247**, 702–704.
Terachi, T., Ogihara, Y. & Tsunewaki, K. 1987 The molecular basis of genetic diversity among cytoplasms of *Triticum* and *Aegilops*. VI. Complete nucleotide sequences of the *rbc*L genes encoding H- and L-type rubisco large subunits in common wheat and *Ae. crassa*. *Jap. J. Genet.* **62**, 375–387.
Viale, A. M., Kobayashi, H. & Akazawa, T. 1989 Expressed genes for plant-type ribulose 1,5-bisphosphate carboxylase/oxygenase in the photosynthetic bacterium *Chromatium vinosum*, which possesses two complete sets of the genes. *J. Bact.* **171**, 2391–2400.
Walter, M. R., Rulin, D. & Horodyski, R. J. 1990 Coiled carbonaceous megafossils from the middle Proterozoic of Jixian (Tianjin) and Montana. *Am. J. Sci.* **290**-A, 133–148.
Wilson, A. C. & Sarich, V. M. 1969 A molecular time scale for human evolution. *Proc. natn. Acad. Sci. U.S.A.* **63**, 1088–1093.
Wilson, M. A., Gaut, B. & Clegg, M. T. 1990 Chloroplast DNA evolves slowly in the palm family (Arecaceae). *Molec. Biol. Evol.* **7**, 303–314.
Wolfe, K. H., Gouy, M., Yang, Y., Sharp, P. M. & Li, W. 1989 Date of the monocot-dicot divergence estimated from chloroplast sequence data. *Proc. natn. Acad. Sci. U.S.A.* **86**, 6201–6205.
Zuckerkandl, E. 1987 On the molecular evolutionary clock. *J. molec. Evol.* **26**, 34–46.
Zurawski, G. & Clegg, M. T. 1987 Evolution of higher-plant chloroplast DNA-encoded genes: implications for structure-function and phylogenetic studies. *A. Rev. Pl. Physiol.* **38**, 391–418.
Zurawski, G., Whitfield, P. R. & Bottomley, W. 1986 Sequence of the gene for the large subunit of ribulose 1,5-bisphosphate carboxylase from pea chloroplasts. *Nucl. Acids Res.* **14**, 3975.

Discussion

T. A. Brown (*Department of Biochemistry and Applied Molecular Biology, UMIST, Manchester, U.K.*). The comparisons that Professor Runnegar showed of RuBisCO sequences from different species indicated that not all of the nucleotide conservation can be explained by amino acid conservation in the proteins themselves. Is this a correct interpretation and if so is there an explanation?

B. Runnegar. The interpretation is correct. For RuBisCO there is no clear explanation but in other proteins conservation occurs in the intraboundary regions. There may be other phenomena like this.

G. A. Dover (*Department of Genetics, University of Cambridge, U.K.*). I agree that the concept of a 'molecular clock' of diverging proteins and DNA has proved useful for sorting deep divisions between taxa on an evolutionary timescale. Some degree of caution is required, however, for more closely related taxa in that as the time separating taxa becomes shorter than the confounding effects of external selection and internal genomic mechanisms of homogenization become greater. There are some well known examples of the effects of the latter which lead to bizarre groupings of, for example, rice and human rDNA sequences to the exclusion of *Drosophila*. Such 'anomalities' can often be ascribed to the combined actions of a mechanism like slippage, accidentally generating similar short repetitive motifs, and gene conversion, involved with intergenic crosstalk. Such eventualities appear as examples of 'homoplasy' in phylogenies, i.e. the 'illogical' sharing of character states between taxa on different branches of a gene. For a comprehensive review of extensive problems with 'homoplasy' and so-called 'evolutionary noise' see Systma (1990) and for a discussion of DNA flux and the molecular clock see Dover (1987) and Hancock & Dover (1991).

We need to have detailed knowledge of the internal dynamics of all molecular probe before interpreting the results of their use in sorting taxonomic relationships.

References

Dover, G. A. 1987 DNA turnover and the molecular clock. *J. molec. Evol.* **26**, 47–58.
Hancock, J. M. & Dover, G. A. 1991 'Compensatory slippage' in the evolution of ribosomal RNA genes. *Nucl. Acids Res.* **18**, 5949–5954.
Systma, K. J. 1990 DNA and morphology: inference of plant phylogeny. *Trends Ecol. Evol.* **5**, 104–110.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES