PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Proteins and Molecular Palaeontology [and Discussion]

R. P. Ambler, M. Daniel, G. A. Dover, B. Halstead and J. P. Thorpe

Phil. Trans. R. Soc. Lond. B 1991 **333**, 381-389
doi: 10.1098/rstb.1991.0088

Proteins and molecular palaeontology

R. P. AMBLER AND M. DANIEL

Institute of Cell and Molecular Biology, Division of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JR, U.K.

SUMMARY

Protein taxonomy has existed as a concept at least since 1958, but despite the efforts of the past 30 years, comparative studies of protein sequence, structure and distribution have not revolutionized any areas of systematics.

The most interesting results of single gene phylogenies have been the anomalies, such as insulin in hystricomorphs or cytochrome *c* in the rattlesnake. Is it likely that protein sequence information can be obtained in sufficient quality and quantity from ancient material as to change this finding? The paper will assess possibilities and the likely limitations of chemical studies of ancient protein material.

1. INTRODUCTION†

Thirty years ago we eagerly awaited the publication and arrival in Britain of C. B. Anfinsen's *The molecular basis of evolution* (1959). By this time, the basic idea that nucleic acid is information written in a four-letter alphabet coding for proteins written in a twenty-letter alphabet was established (although not universally accepted), but the nature of the code and the mechanisms of synthesis were still unknown. Sanger (1956) has shown how the amino acid sequences of proteins could be determined, and the first information on sequence differences between homologous proteins (those performing the same function in different organisms) was becoming available. Anfinsen suggested several possible uses for comparative amino acid sequence information, including the delineation of the minimum structure essential for biological function, clues as to the rate at which successful mutations have occurred during evolutionary time, and as 'an additional basis for establishing phylogenetic relationships'.

The species specificity of proteins was recognized before any detailed knowledge was available about their structure, and the 'prehistory of protein molecular evolution' is represented by G. H. F. Nuttall's immunological experiments before 1904 with unfractionated human serum, and the studies by E. T. Reichart and A. P. Brown (1909) on the morphology of haemoglobin crystals from a wide range of vertebrates. By 1952, G. Wald could say that 'It seems to be true that every species of organism makes at least a few specific proteins'.

This last purpose fascinated many people during the subsequent 30 years, for much of which protein sequencing provided the only way of getting detailed comparative information about genetic fine structure. Comparative sequence studies have played an important role in identifying and elucidating many of the complexities of genomic structure, but have had surprisingly little effect on establishing phylogenetic relationships, or of disturbing relationships derived from the classical methods of palaeontology and comparative morphology. A major exception is for the prokaryotes, for which information derived from sequences of ribosomal RNA (Woese 1987) is widely accepted as providing, at last, a natural classification for organisms that have a very sparse fossil record and little useful morphology. Similar studies with 18*S* rRNA have also begun to perturb metazoan phylogeny (Patterson 1990).

Proteins are made up from 20 universal amino acids joined through α-peptide linkage to form linear molecules containing up to about 1000 amino acid residues, although most types of protein are not as long as this. The linear order of the amino acids specifies the folded three-dimensional structure, although the folding rules are not yet understood. As there are no restrictions on what sequences can be synthesized, the number of different possible proteins is very large: thus there are $20^{100} = 10^{130}$ possible 100-residue proteins. A typical cell produces at least 10^4 different proteins, and as proteins are species specific, and there are perhaps 10^7 different species of organisms now living (and many more extinct or not yet in existence), the protein repertoire is still largely untapped.

The earliest amino acid sequences determined were perforce of small and abundant proteins from readily available sources. Thus the six earliest insulin sequences determined were from 'slaughterhouse' animals used by the pharmaceutical industry for preparing the hormone for clinical purposes. Insulin was not an attractive system for wider comparative sequence studies because its occurrence is limited to vertebrates

† The introduction will refer for the development of the subject of 'molecular evolution' mainly to symposium articles rather than to the definitive journal papers.

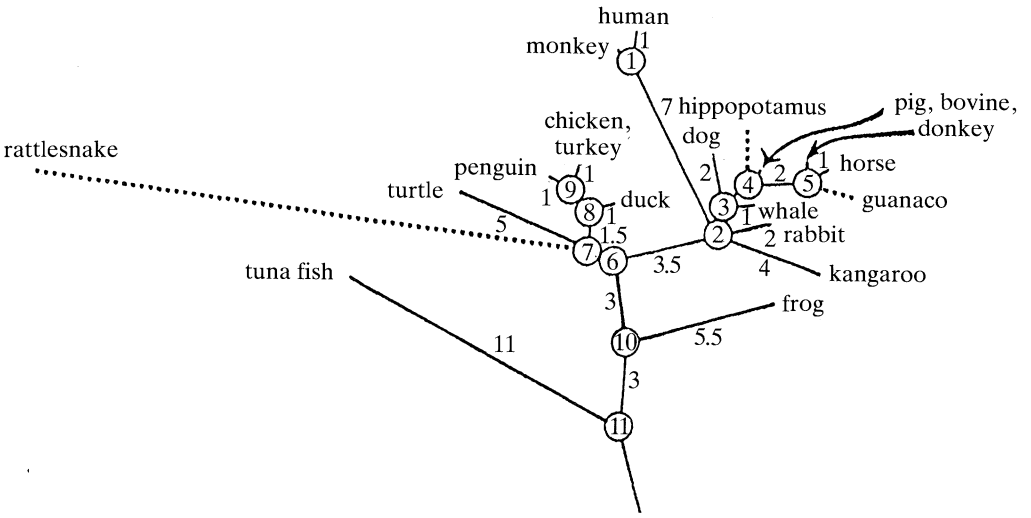

Figure 1. Part of a phylogenetic tree of cytochrome *c*, from Dayhoff & Eck (1968). The numbers of inferred amino acid changes per hundred residues are shown on the tree. These authors note: 'Most of the major branches fall clearly into the topology shown. However, the evidence is weak for exact attachment of the reptile branch. Rattlesnake cytochrome *c* has undergone a large number of changes since its divergence from the ancestral line. Several of the changes happen to be uniquely shared with the primates. Coincidence is the only explanation we can think of for these identities. We have drawn the rattlesnake connection from biological considerations.'

(despite occasional reports of it from more distant sources (LeRoith *et al.* 1981)). A more promising protein proved to be cytochrome *c*, with sequence evidence from only a few amino acids around the haem attachment site (Tuppy 1958) of animal, yeast and bacterial proteins encouraging Anfinsen (1959) to believe 'that certain units of the universal gene pool may be extremely ancient'.

2. VALIDATION OF PROTEIN PHYLOGENIES

Crick said in 1958 'Biologists should realize that before long we shall have a subject which might be called "protein taxonomy" – the study of the amino acid sequences of the proteins of an organism, and the comparison of them between species'. The next phase was the determination of amino acid sequences with the direct intention of assessing interspecies differences, and in interpreting these differences in evolutionary terms. The report of a meeting at Rutgers in 1964, published as the influential 'Evolving genes and protein' (Bryson & Vogel 1965), produced papers that began the discussion of the different rates of protein evolution and of conserved residues in sequences. At about the same time the first phylogenetic tree derived from the comparison of protein sequences appeared, for the small and rapidly evolving fibrinopeptides (Doolittle & Blombäck 1964).

The primary test system proved to be cytochrome *c*, as it could be isolated in adequate quantities (1–10 μmol) from almost any reasonably abundant eukaryote, seemed to be genetically simple, and appeared to have been evolving so slowly that the proteins from organisms as distinct as yeast and man were clearly functionally and structurally homologous. In contrast, the globins were genetically complex, with multiple forms present in the same organism, and appeared to be evolving considerably faster. The globins are widely distributed in eukaryotes, and are even known in bacteria (Wakabayashi *et al.* 1986), but their occurrence in invertebrates, plants and lower eukaryotes is very erratic.

There are now mitochondrial cytochrome *c* sequences recorded for the protein from about 90 different species, including protozoa, yeasts, algae and a few invertebrates as well as the 36 vertebrates and a similar number of high plants. Phylogenetic trees have been derived from these data as they have become available, beginning with Fitch & Margoliash (1967) and Dayhoff & Eck (1968), and these clearly show the general congruence of information from sequences of a single homologous gene and relationships derived from the traditional methods of the fossil record and morphology (figure 1).

3. ANOMALIES IN PROTEIN EVOLUTION

The early insulin sequences suggested that this protein was very resistant to evolutionary change, as the only differences between cattle, horse, pig and whale insulins were in a single three-residue segment of the molecule. However, L. F. Smith (1966, 1972) obtained contrasting results when he studied the insulins of the South American hystricomorph rodents, the coypu, chinchilla and guinea-pig, which he did because guinea-pig insulin had already been reported by Moloney & Coval (1955) to be immunologically distinctive. Smith (1966) showed that the guinea-pig insulin differed from the pig hormone in 17 of 51 positions, in contrast to the mere four positions in which pig and rat differ. Smith (1972) went on to show that coypu insulin was also very different, but that from chinchilla was similar to pig and other familiar mammals. The cytochrome *c* of guinea-pig is identical in sequence to those of rat and mouse (Carlson *et al.* 1977), and Beintema & Lenstra (1982) have shown that hystricomorph pancreatic ribonucleases have

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

evolved quite normally. Blundell & Wood (1975) have explained the guinea-pig insulin anomaly by showing that it and the coypu protein do not form a zinc-binding and protease-resistant hexamer. Thus once the structural requirement for zinc binding had been lost from the guinea-pig line, selection no longer acted to conserve the subunit contacts, and rapid change occurred.

A striking anomaly in the cytochrome *c* data is the rattlesnake sequence, originally reported by Emil Smith and O. P. Bahl (Bahl & Smith 1965), which they noted as particularly resembling the human protein in several positions. Fitch & Margoliash (1967) remarked that the turtle sequence grouped with the birds rather than the snake, and Dayhoff & Eck (1968) showed the snake on a long branch from its 'biological' position on the bird–turtle node (figure 1), and commented on the apparent parallel evolution of primates and rattlesnake. The snake did not appear in Dayhoff's subsequent trees (see Dayhoff 1969), and Fitch (1973) considered the possibility that the published sequence might be wrong. We have re-investigated the sequence (Ambler & Daniel 1991), and believe that the correct sequence differs in nine places from that used for evolutionary theorizing since 1965. Nevertheless, the sequence still resembles that of human cytochrome *c* more than that of any other protein that we know, and we believe that this is an example of convergent evolution accompanied by accelerated change in the line connecting the rattlesnake to the ancestral vertebrates.

A case of molecular convergence, but in a situation where a functional explanation can be put forward, has been found for the lysozymes of ruminants and leaf-eating primates (Stewart *et al.* 1987), both of which ingest large amounts of bacteria into their stomachs and which could be expected to benefit from a potent bacteriolytic enzyme in their digestive tract. The lysozymes of the cow and a colobine monkey (*Presbytis entellus*) share both some enzymic properties (such as functioning at low pH and resistance to pepsin degradation) and several uniquely shared amino acid residues not present in other members of their respective lines.

4. PROTEIN EVOLUTION: SELECTION OR DRIFT?

Two theoretical questions have concerned molecular evolutionary studies for the last 25 years. The first is the concept of the molecular clock, the suggestion by Zuckerkandl & Pauling (1962) that sequences might be evolving at a constant rate. The implications of this possibility have been developed by Wilson *et al.* (1977), and attempts made to show the universality of a rate of sequence change through 3×10^9 years (Ochman & Wilson 1987). The second question has been the relative importance of selection and drift for evolutionary change in proteins and nucleic acids. The fixation of neutral mutations was proposed by Kimura (1968) to explain the apparently high rate of observed nucleotide substitutions, and King & Jukes (1969) in their classic article 'Non-Darwinian evolution' assessed the question in the light of what was then known about genome structure.

In more recent years, some of the apparent paradoxes have been resolved and the evidence suggests that selection is acting on most amino acid substitutions, although only to a far lesser extent on 'silent' third-base nucleotide changes. For instance it was initially claimed that most mitochondrial cytochromes *c* were functionally indistinguishable (Margoliash *et al.* 1972). Subsequently, alterations to assay conditions enabled kinetic differences to be detected between cytochromes *c* that differ from each other in only one or two amino acid residues (Margoliash *et al.* 1976), although some question remains about the physiological significance of the kinetic differences (Kamen *et al.* 1978). Two distinct forms of the enzyme alcohol dehydrogenase occur in natural populations of *Drosophila melanogaster*, and in each population the relative proportions of each allele vary with latitude and habitat. The forms differ by a single amino acid substitution (Retzios & Thatcher 1978), but DNA sequencing of 11 genes from five different natural populations (Kreitman 1983) detected 42 silent nucleotide polymorphisms but only the same single amino acid substitution

5. PHYLOGENETIC ACHIEVEMENTS WITH MODERN PROTEIN SEQUENCES

There have been at least two major systematic attempts to use protein sequences to solve specific taxonomic problems. The first was that by Boulter to use mitochondrial cytochrome *c* (Boulter *et al.* 1972) and chloroplast plastocyanin (Boulter *et al.* 1979) sequences to deduce flowering plant phylogeny. Cronquist (1976) reviewed the progress of the work from the viewpoint of a classical taxonomist, and concluded that the 'computer based trees for cytochrome *c* of a limited number of kinds of angiosperms that have so far been produced are out of harmony with all previous phylogenetic systems, and conclusions that have been based on them are seriously discordant with the fossil record'. Syvanen *et al.* (1989) agree with Boulter that for cytochrome *c*, 'the classical plant taxonomic ambiguities extend to the molecular level'.

A second major effort has been the use of myoglobin sequences by Joysey and his associates (Romero-Herrera *et al.* 1978, 1982; Joysey 1988).

It has seemed that both studies have been on the verge of making major contributions but have lacked both the quantity of data and adequate mathematical techniques to convince sceptics in cases where molecular and classical phylogenies differ. Maybe all that is needed to put Humpty-Dumpty together again is more horses and more men. DNA methodology now makes it easy (if expensive) to get sequences of longer genes with more appropriate functions from a wider variety of organisms. But nevertheless all reconstructions of phylogenies from modern sequences depend on extrapolation to try to deduce what past sequences may have been like, so lack the reinforcement that would be given by any evidence about what past sequences really were.

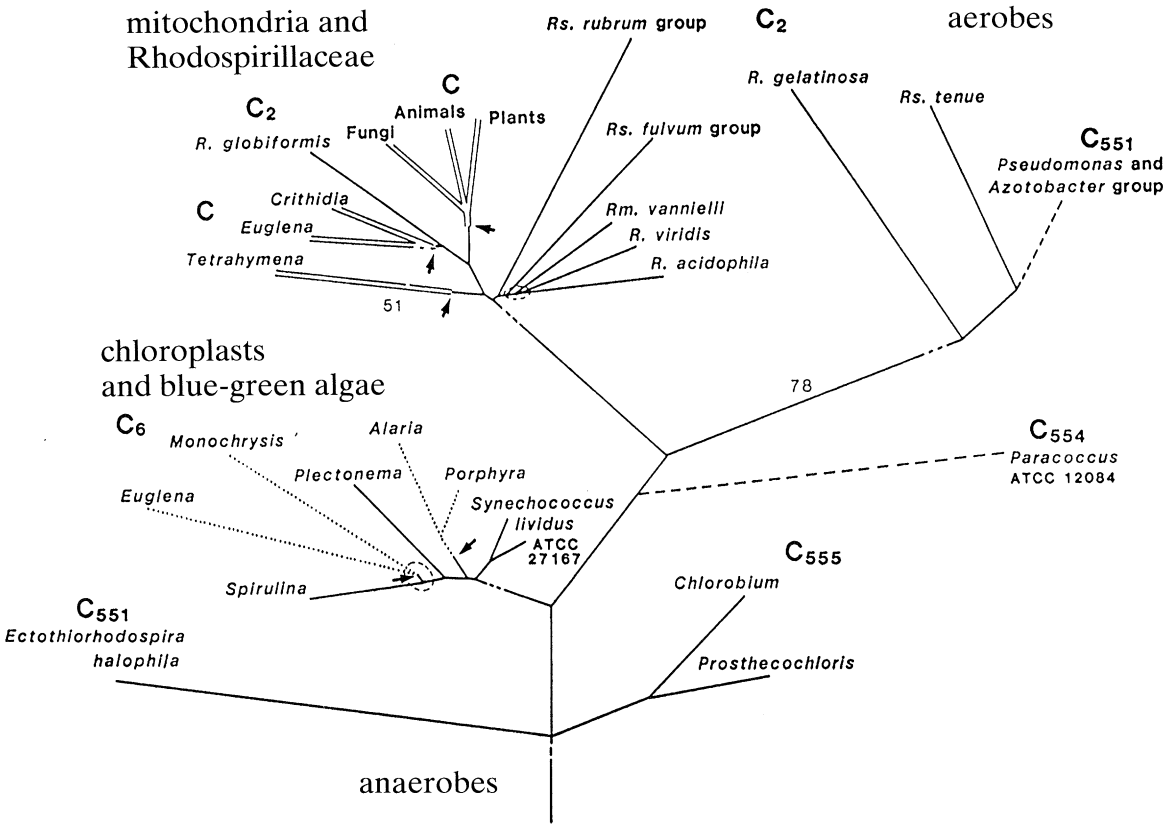

Figure 2. Evolutionary tree derived from *c*-type cytochromes from mitochondria, chloroplasts and bacteria, from Dayhoff (1983). The arrows mark supposed occurrences of symbiotic assimilation of bacteria in the formation of organelles.

6. WHAT PROBLEMS ARE SUITABLE FOR SOLVING BY SEQUENCE PHYLOGENIES?

Sequence approaches have not yet caused much disturbance to relationships deduced through the classical methods of palaeontology and comparative morphology, so there are good arguments for concentrating efforts and resources on areas in which molecular methods are likely to be able to make unique contributions.

The general equivalence of single-gene trees (figure 1) with those deduced from fossil evidence has justified attempts to use sequence methods to try to reconstruct what happened in the very remote past. Results from proteins provide convincing evidence for the theory of the origin of the eukaryotic organelles (Margulis 1970) from prokaryotic symbionts. Sequences of cytochromes *c* (Ambler & Bartsch 1975; Ambler *et al.* 1976), plastocyanins and ferredoxins showed great similarity between proteins from organelles and from bacteria. Dayhoff (1983) interpreted sequence evidence as showing that mitochondria and chloroplasts are derived from organisms related to specific lines of modern bacteria (figure 2). The results from protein studies are in general, but not total, agreement with those from 16*S* rRNA sequences (Woese 1987). The significance of the discordances and the importance of lateral gene transfer in molecular evolution, remain matters for controversy (Ambler *et al.* 1979; Woese *et al.* 1980; Ambler 1985).

Attempts to explore questions such as the origin of the vertebrates and the relationship of invertebrate groups were limited until recently by the difficulty in getting sufficient material to isolate proteins from such interesting organisms as tardigrades or ascidians. DNA technology has removed this limitation, and if resources were available much worthwhile work could be done with such organisms.

Another area where sequence methods can be used with good effect is the study of human evolution. Protein studies have not yet been conclusive in elucidating primate relationships, primarily because of the small amount of sequence difference between human and ape genes (King & Wilson 1975). However, the very large amount of information now available about the human genome, together with the information on polymorphisms derived from population studies, means that there is a firm base to which any information on other primates can be linked. This data base also enables any information that can be derived from human material from archaeological contexts to be interpreted informatively. A recent example was the report of the HLA typing of 7500-year-old human remains by DNA techniques (Lawlor *et al.* 1991).

7. ANCIENT PROTEIN MATERIAL

As Abelson (1956) showed, amino acids and polypeptides are a constituent of many but not all fossils. Protein residues occur in ancient bone (Armstrong *et al.* 1983; Ascenzi *et al.* 1985), shell (Abelson 1956;

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Weiner *et al.* 1976) or on stone implements (Loy 1983). In these contexts, the surviving material is in close association with an inorganic matrix, and such association may be essential for long-term survival. The most convincing evidence for the survival of identifiable proteins comes from the immunological work of Lowenstein (1981, this symposium) and others, although positive results by these methods might be given by material that had suffered considerable degradation.

8. PROTEIN DEGRADATION WITH TIME

Abelson (1956) was the pioneer investigator of amino acid survival in fossil material. He showed that amino acids and peptides could be isolated from ancient bone and shell, but that the composition did change with time, with the less stable amino acids such as serine, threonine and tyrosine being absent from older fossils. He showed that the rate of loss of the stable amino acid alanine from shell samples over millions of years was consistent with its degradation in short times at high temperatures in laboratory experiments.

Keilin & Wang (1947) examined samples of liquid blood that had been stored in sealed ampoules in the dark at room temperature for periods of up to 44 years. They examined the functional properties of the haemoglobin and measured the enzymic activity of four endoerythrocytic enzymes, and found that, for the samples that had remained free of microbial contamination, these parameters were within the limits expected for fresh blood of the species concerned. They were also able to crystallize haemoglobin from a sample of guinea pig that had been collected and preserved aseptically. Further ampoules of these bloods (which were part of Nuttall's 1904 studies) are believed to survive, and so could now be examined by modern methods after 90 years storage.

Sensabaugh *et al.* (1971) examined the proteins from a sample of blood that had been dried as a thin film on a plastic sheet and then stored in the dark and under desiccation for eight years. They showed that covalent aggregation of protein had occurred as had some degradation, and that whereas activity remained for most of the expected enzymes, the best recoveries were less than 40 %. Electrophoretically, the protein pattern was less distinct and more anodal than for fresh blood. They suggested this change had been caused by reaction of lysine side chains, but it would also be an effect of progressive deamidation.

As an extension of Abelson's work with alanine, several people have tried to study the time degradation of proteins by accelerating decay with enhanced temperature. Totten *et al.* (1972) measured the amino acid composition of hydrolysates of fractions from powdered oyster shell that had been heated dry at 130 °C for periods of from one to eight weeks. From unheated samples, the protein fraction (defined as material insoluble in 2 M-HCl) yielded predominantly aspartic acid and glycine, but the amount and relative proportion of these two amino acids was very much lower in all the heated samples. Within the first week of heating, 99 % of the glycine and aspartic acid, and from 50 to 80 % of the other amino acids, had been lost from the insoluble protein fraction. The amino acid profile of the heated protein appeared to match that of fossil oyster shell, and the authors interpreted the results as indicating great differences in the stability of individual proteins in the shell matrix.

9. HOW DOES PROTEIN DETERIORATE WITH TIME?

When an organism dies, most of its proteins and other biomolecules will rapidly be broken down. The initial stages will be mediated by the autolytic processes of the organism's own enzymes, and the process will then be accelerated by the participation of other organisms, from hyaenas to bacteria. Only if the organism dies under especially favourable circumstances is it likely that proteins will survive to degrade through non-biological processes. Such circumstances will include rapid drying and entombment in a non-oxidative medium; proteins in bone and shell may be naturally protected. Whereas very ancient protein material does survive, there are thermodynamic limitations to the survival of any molecules over long periods. Claims for the existence of deep sea organisms capable of life and growth at 250 °C (Baross & Deming 1983) have been convincingly debunked by a consideration of the half-lifes in solution of biological molecules at this temperature (White 1984). The same molecules will have half-lifes for decay at the surface temperature at which they are surviving in ancient human or animal remains. It is possible that the close association of protein with a mineral matrix may retard its decay (Pinck & Allison 1951). Nevertheless, even in the absence of micro-organisms and enzymes, proteins will continue to decay, and in the driest terrestrial environments there may well be enough water present to allow hydrolytic degradation to occur.

The ways in which the covalent structure of a protein can alter include: (i) hydrolysis of peptide bonds; (ii) modification of amino acid side and terminal groups which may include the cross-linking of peptide chains; (iii) racemization of chiral centres.

The stability of peptide bonds is largely a function of the amino acid residues that form the bond, with great differences between the stability of (say) –Ile–Ile– and –Asp–Pro–. In the laboratory, all but the most labile bonds are stable for days under conditions as severe as pH 2 and 60 °C. An enzymic digest of a protein sample stored as a freeze-dried powder for 30 years at room temperature showed no evidence for peptide bond breakage during storage, although some amide groups (probably in –Asn–Gly– sequences) seemed to have hydrolysed (R. P. Amber, unpublished results).†

The side chain amide groups of asparagine and glutamine may also be labile. In a set of 64 synthetic pentapeptides the amide half-lives at pH 7.5 and 37 °C varied from six days to nearly ten years (table 1;

† The protein was *Salmonella* flagellin, prepared in 1960 and examined recently for the distribution of ε-*N*-methyl lysine residues (Ambler & Rees 1958), possible now that the gene sequence of the protein has been determined (Joys 1985).

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Table 1. *Deamidation half-times for some synthetic peptides (from Robinson 1974) in pH 7.5 phosphate buffer at 37 °C*

Peptide	$t_{\frac{1}{2}}$ days
Gly–Ser–Asn–His–Gly	6
Gly–Ser–Asn–Ala–Gly	52
Gly–Leu–Asn–Ala–Gly	217
Gly–Leu–Gln–Ala–Gly	663
Gly–Ile–Gln–Ala–Gly	1087
Gly–Thr–Gln–Ala–Gly	3409

Robinson 1974). Many of the side chains are susceptible to oxidation, particularly the sulphur amino acids methionine and cysteine. The aromatic amino acids tryptophan and tyrosine are vulnerable to both radiation and to oxidation, and the hydroxyamino acids threonine and serine (and hydroxyproline) are more susceptible to hydrolytic damage than the other protein amino acids. Peptide chains may also be cross-linked through reactive side chains, or covalently linked to non-protein material.

At temperatures and pressures near those that now occur on the surface of the earth, amino acids in or derived from proteins appear to undergo complete racemization in timespans of around a million years (Bada, this symposium). Racemization can occur to residues in proteins without the peptide bond breaking (Smith & Evans 1980), and such change would ensure the denaturation of the protein.

10. THE APPARENT STATE OF PROTEIN IN ANCIENT SAMPLES

Collagen is the predominant matrix protein in fresh bone, and has a highly characteristic amino acid composition, very rich in glycine and proline, and also containing two unusual amino acids, hydroxyproline and hydroxylysine formed by postsynthetic modification of two of the 'twenty' amino acids. Hydroxyproline is present in particularly large amounts. Armstrong *et al.* (1981) have measured the amino acid composition of the protein residue of a large number of fossil bones. Only the Pleistocene material showed any compositional similarity to modern collagen, and the occurrence in it of hydroxyproline and hydroxylysine was very erratic. It is difficult to imagine being able to obtain useful amino acid sequence information from a collagen degradation product in which many of the hydroxyproline residues had been altered.

Weiner *et al.* (1976) made a careful study of the protein residue in a Cretaceous mollusc shell. They obtained evidence that the microarchitecture of the fossil shell remained intact, that discrete high molecular mass protein components remained, and that only a low level of alloisoleucine had formed, suggesting that the protein had been preserved in an anhydrous environment within the shell. They also found the unstable amino acids serine, methionine and tyrosine to have survived very well in the fossil protein, and cyst(e)ine may also have been present.

Loy (1983, 1987) used a variety of methods to identify specific proteins in dried residues remaining on stone implements from archaeological contexts dating back to 100000 years before present (BP). He reports evidence that the residues are dried blood derived from the use or the manufacture of the implements, and uses techniques that would require intact protein molecules to identify the species of origin of the blood. The methods include crystallization and examination of the crystal habit of haemoglobin, the pioneer method of Reichart & Brown (1909) and the isoelectric focusing pattern of blood proteins. Loy & Wood (1989) have also succeeded in crystallizing haemoglobin isolated from bone, and have convinced themselves that some of the blood on a slab at Çayönü Tepesi, Turkey, was derived from the now extinct *Bos primigenius* (and see Hammond 1990). Loy and his associates have used the same techniques, supplemented by immunological methods, to identify human blood on implements from Barda Balka, Iraq, dated to about 100000 years BR (Loy 1987; Nuttall 1990).

Ascenzi *et al.* (1985) have shown by immunological methods that haemoglobin can be detected in samples of exhumed human bone as old as 4000 years, with the amount that survives decreasing with age. An unfortunate typographical omission from their paper† means that the amount of haemoglobin recoverable from recent bone is given as up to 30 g haemoglobin per 100 g bone powder, instead of the 30 μg they had intended. Their finding makes the reported success of Loy & Wood (1989) in crystallizing haemoglobin from 8000-year-old auroch bone somewhat surprising.

11. POTENTIAL SYSTEMS FOR PROTEIN PALAEONTOLOGY

The experience of biochemists in using the sequences of modern proteins to try to elucidate phylogeny has some lessons for bimolecular palaeontology. The requirements for a suitable system include: (i) size of gene product; (ii) genetic simplicity (paralogy/orthology/multiple genes); (iii) abundance (level of expression); (iv) distribution; (v) ease of isolation; (vi) availability of information about homologous proteins; (vii) rate of evolutionary change; (viii) nature of sequence; (ix) understanding of function. In addition, for ancient proteins, a further criterion will apply: (x) long-term stability.

Factor (vii) has limited many studies. Thus cytochrome *c* has been evolving too slowly to have been appropriate for elucidating flowering plant evolution (Boulter *et al.* 1972). No protein yet discovered would seem suitable for the study of hominid or great ape relationships,‡ differences having to be searched for in non-coding repetitive DNA or in the third bases of structural genes.

The size (factor (i)), multiplicity of genes (factor (ii)) and the dullness of the semi-repetitive sequence (factor (viii)), plus the questionable stability of hydroxyproline (factor (X)) makes collagen seem far from ideal as a

† Personal communication from M. Brunori, and correction in Lendaro *et al.* (1991).

‡ After this article was written, Ruvolo *et al.* (1991) reported that they have resolved the human–chimpanzee–gorilla trichotomy through sequence characterization of the mitochondrial cytochrome oxidase subunit II gene.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

protein to study. The shell proteins suffer from our ignorance of the molecular genetics (factor (ii)) of the organisms that produce them and the lack of information yet about factors (vi) and (vii), although their sequestration in the inorganic matrix appears to confer exceptional survival (factor (x)) on some of them (Weiner *et al.* 1976).

The techniques for comparing present-day sequences have required that the information should be virtually complete for every member of the set, and no one has yet attempted to construct a phylogeny from a set of different incomplete fragments. Nevertheless it is likely that with ancient proteins the sequence for each old species will contain unstable regions which will not be recoverable, and such gaps will make analysis and reconstruction of trees unsatisfactory.

12. CONCLUSIONS

For the past 15 years it has been practicable to sequence DNA faster and at least as accurately as proteins, although protein sequencing methodology has also made great advances during this period. In addition, as we will have heard in the second half of this symposium, the polymerase chain reaction has revolutionized the prospects of obtaining sequence information from damaged DNA (Pääbo *et al.* 1990), although there appears no likelihood of devising an equivalent technique for averaging and sequencing damaged proteins. DNA may well survive for millions of years only in very special circumstances, so protein techniques may be advantageous for very ancient samples if we can better understand the factors responsible for time-dependent protein breakdown, and we determine the types of locations where protein can survive in a state of exceptional preservation.

Nevertheless, the problems in interpreting comparisons between homologous structural genes remain the same whether they have been deduced from experiments at the amino acid or the nucleotide level, and the lessons learnt from the past 35 years of protein sequencing and discussed here remain valid. In particular, most of the criteria noted in §11 for the choice of potential systems for protein palaeontology are true whatever methodological level is used for the sequence determination, and the likely interpretational problems should be considered before major programmes are started.

The study of ancient proteins will continue to be useful and exciting even if the possibilities for the retrieval of sequence information are limited, and progress will certainly require a better understanding of how proteins deteriorate with time.

REFERENCES

Abelson, P. H. 1956 Palaeobiochemistry. *Scient. Am.* **195**, 83–92.

Ambler, R. P. 1985 Protein sequencing and taxonomy. In *Computer-assisted bacterial systematics* (ed. M. Goodfellow, D. Jones & F. G. Priest), pp. 307–335. London: Academic Press.

Ambler, R. P. & Bartch, R. G. 1975 Amino acid sequence similarity between cytochromes *f* from a blue-green bacterium and algal chloroplasts. *Nature, Lond.* **253**, 285–288.

Ambler, R. P. & Daniel, M. 1991 Rattlesnake cytochrome *c*: a re-appraisal of the reported amino acid sequence. *Biochem. J.* **274**, 825–831.

Ambler, R. P. & Rees, M. W. 1958 ε-Methyl-lysine in bacterial flagellar protein. *Nature, Lond.* **184**, 56–57.

Ambler, R. P., Meyer, T. E. & Kamen, M. D. 1976 Primary structure determination of two cytochromes c_2: close similarity to functionally unrelated mitochondrial cytochrome *c*. *Proc. natn. Acad. Sci. U.S.A.* **73**, 472–475.

Ambler, R. P., Meyer, T. E. & Kamen, M. D. 1979*a* Anomalies in amino acid sequences of small cytochromes *c* and cytochromes *c′* from two species of purple photosynthetic bacteria. *Nature. Lond.* **278**, 661–662.

Ambler, R. P., Daniel, M., Hermoso, J., Meyer, T. E., Bartsch, R. G. & Kamen, M. D. 1979*b* Cytochrome c_2 sequence variation among the recognized species of purple nonsulphur photosynthetic bacteria. *Nature. Lond.* **278**, 659–660.

Anfinsen, C. B. 1959 *The molecular basis of evolution.* New York: Wiley.

Armstrong, W. G., Halstead, L. B., Reed, F. B. & Wood, L. 1983 Fossil proteins in vertebrate calcified tissues. *Phil. Trans. R. Soc. Lond.* B **301**, 301–343.

Ascenzi, A., Brunori, M., Citro, G. & Zito, R. 1985 Immunological detection of hemoglobin in bones of ancient Roman times and of Iron and Eneolithic Ages. *Proc. natn. Acad. Sci. U.S.A.* **82**, 7170–7172.

Bahl, O. P. & Smith, E. L. 1965 Amino acid sequence of rattlesnake heart cytochrome *c*. *J. biol. Chem.* **240**, 3585–3593.

Baross, J. A. & Deming, J. W. 1983 Growth of 'black smoker' bacteria at temperatures of at least 250 °C. *Nature, Lond.* **303**, 423–426.

Beintema, J. J. & Lenstra, J. A. 1982 Evolution of pancreatic ribonucleases. In *Macromolecular sequences in systematic and evolutionary biology* (ed. M. Goodman), pp. 43–73. New York: Plenum

Blundell, T. L. & Wood, S. P. 1975 Is the evolution of insulin Darwinian or due to selectively neutral mutation? *Nature, Lond.* **257**, 197–203.

Boulter, D., Peacock, D., Guise, A., Gleaves, J. T. & Estabrook, G. 1979 Relationships between the partial amino acid sequences of plastocyanin from members of ten families of flowering plants. *Phytochemistry* **18**, 603–608.

Boulter, D., Ramshaw, J. A. M., Thompson, E. W., Richardson, M. & Brown, R. H. 1972 A phylogeny of higher plants based on the amino acid sequences of cytochrome *c*, and its biological implications. *Proc. R. Soc. Lond.* B **181**, 441–455.

Bryson, V. & Vogel, H. J. 1965 *Evolving genes and proteins.* New York: Academic Press.

Carlson, S. S., Mross, G. A., Wilson, A. C., Mead, R. T., Wolin, D., Bowen, S. F., Foley, N. T., Muijsens, A. O. & Margoliash, E. 1977 Primary structure of mouse, rat and guinea-pig cytochromes *c*. *Biochemistry* **16**, 1437–1442.

Crick, F. H. C. 1958 On protein synthesis. *Symp. Soc. exp. Biol.* **12**, 138–163.

Cronquist, A. 1976 The taxonomic significance of the structure of plant proteins: a classical taxonomist's view. *Brittonia* **28**, 1–27.

Dayhoff, M. O. 1969 *Atlas of Protein Sequence and Structure*, Vol. 4. Silver Spring, Maryland: National Biomedical Research Foundation.

Dayhoff, M. O. 1983 Evolutionary connections of biological kingdoms based on protein and nucleic acid sequence evidence. *Precambrian Res.* **20**, 299–318.

Dayhoff, M. O. & Eck, R. V. 1968 *Atlas of Protein Sequence*

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

and Structure 1967–68. Silver Spring, Maryland: National Biomedical Research Foundation.

Doolittle, R. F. & Blombäck, B. 1964 Amino acid sequence investigations of fibrinopeptides from various mammals: evolutionary implications. *Nature, Lond.* **202**, 147–152.

Fitch, W. M. 1973 Aspects of molecular evolution. *A. Rev. Genet.* **7**, 343–380.

Fitch, W. M. & Margoliash, E. (1967) Construction of phylogenetic trees. *Science, Wash.* **155**, 279–284.

Hammond, N. 1990 Human bloodstains found on neolithic sacrificial altar. *The Times, Lond.* 18 January.

Joys, T. M. 1985 The covalent structure of the phase-1 flagellar filament protein of *Salmonella typhimurium* and its comparison with other flagellins. *J. biol. Chem.* **260**, 15758–15761.

Joysey, K. A. 1988 The use of amino acid sequences in phylogenetic analysis. In *Molecular evolution and the fossil record* (ed. T. W. Broadhead), pp. 34–48. Knoxville, Tennessee: Palaeontological Society of America.

Kamen, M. D., Errede, B. J. & Meyer, T. E. 1978 Comparative studies of cytochrome *c*. In *Evolution of protein molecules* (ed. H. Matsubara & T. Yamanaka), pp. 378–385. Tokyo: Japan Scientific Societies Press.

Keilin, D. & Wang, Y. L. 1947 Stability of haemoglobin and of certain endoerythrocytic enzymes *in vitro*. *Biochem. J.* **41**, 491–500.

Kimura, M. 1968 Evolutionary rate at the molecular level. *Nature, Lond.* **217**, 624–626.

King, J. L. & Jukes, T. H. 1969 Non-Darwinian evolution. *Science, Wash.* **164**, 788–798.

King, M.-C. & Wilson, A. C. 1975 Evolution at two levels in humans and chimpanzees. *Science, Wash.* **188**, 107–116.

Kreitman, M. 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature, Lond.* **304**, 412–419.

Lawlor, D. A., Dickel, C. D., Hauswirth, W. W. & Parham, P. 1991 Ancient *HLA* genes from 7,500-year-old archaeological remains, *Nature, Lond.* **349**, 785–788.

Lendaro, E., Ippoliti, R., Bellelli, A., Brunori, M., Zito, R., Citro, G. & Ascenzi, A. 1991 On the problem of immunological detection of antigens in skeletal remains. *J. Phys. Anthrop.* (*In the press.*)

LeRoith, D., Shiloach, J., Roth, J. & Lesniak, M. A. 1981 Insulin or a closely related molecule is native to *Escherichia coli*. *J. biol. Chem.* **256**, 6533–6536.

Lowenstein, J. M. 1981 Immunological reactions from fossil material. *Phil. Trans. R. Soc. Lond.* B **292**, 143–149.

Loy, T. H. 1983 Prehistoric blood residues: detection on tool surfaces and identification of species of origin. *Science, Wash.* **220**, 1269–1270.

Loy, T. H. 1987 Recent advances in blood residue analysis. In *Archaeometry: further Australasian studies* (ed. W. R. Ambrose & J. M. J. Mummery), pp. 57–65.

Loy, T. H. & Wood, A. R. 1989 Blood residue analysis at Çayönü Tepesi, Turkey. *J. field Archaeol.* **16**, 451–460.

Margoliash, E., Ferguson-Miller, S., Brautigan, D. L. & Chaviano, A. H. 1976 Functional basis for evolutionary change in cytochrome *c* structure. In *Structure-function relationships of proteins* (ed. R. Markham & R. W. Horne), pp. 145–165. Amsterdam: Elsevier.

Margoliash, E., Fitch, W. M., Markowitz, E. & Dickerson, R. E. 1972 Functional limits of cytochrome *c* variability. In *Oxidation–reduction enzymes* (ed. A. Akeson & A. Ehrenburg), pp. 5–17. Oxford: Pergamon.

Margulis, L. 1970 *Origin of eukaryotic cells*. New Haven, Connecticut: Yale University press.

Moloney, P. J. & Coval, M. 1955 Antigenicity of insulin: diabetes induced by specific antibodies. *Biochem. J.* **59**, 179–185.

Nuttall, G. H. F. 1904 Blood immunity and blood relationship. London: Cambridge University Press.

Nuttall, N. 1990 Neanderthal man may reveal his bloodline. *The Times, Lond.* 28 December.

Ochman, H. & Wilson, A. C. 1987 Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. *J. molec. Evol.* **26**, 74–86.

Pääbo, S., Irwin, D. M. & Wilson, A. C. 1990 DNA damage promotes jumping between templates during enzymatic amplification. *J. biol. Chem.* **265**, 4718–4721.

Patterson, C. 1990 Metazoan phylogeny: reassessing relationships. *Nature, Lond.* **344**, 199–200.

Pinck, L. A. & Allison, F. E. 1951 Resistance of a protein-Montmorillonite complex to decomposition by soil microorganisms, *Science, Wash.* **114**, 130–131.

Reichart, E. T. & Brown, A. P. 1909 *The crystallography of haemoglobins*. Publication No. 16. Washington D.C.: Carnegie Institute.

Retzios, A. D. & Thatcher, D. R. 1978 Chemical basis of the electrophoretic variation at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Biochimie* **61**, 701–704.

Robinson, A. B. 1974 Evolution and the distribution of glutaminyl and asparaginyl residues in proteins. *Proc. natn. Acad. Sci. U.S.A.* **7**, 885–888.

Romero-Herrera, A. E., Lehmann, H., Joysey, K. A. & Friday, A. E. 1978 On the evolution of myoglobin. *Phil. Trans. R. Soc.* B **283**, 61–163.

Romero-Herrera, A. E., Lieska, N., Friday, A. E. & Joysey, K. A. 1982 The primary structure of carp myoglobin in the context of molecular evolution. *Phil. Trans. R. Soc.* B **297**, 1–25.

Ruvolo, M., Disotell, T. R., Allard, M. W., Brown, W. M. & Honeycutt, R. L. 1991 Resolution of the African hominoid trichotomy by the use of a mitochondrial gene sequence. *Proc. natn. Acad. Sci. U.S.A.* **88**, 1570–1574.

Sanger, F. 1956 The structure of insulin. In *Currents in biochemical research* (ed. D. E. Green). New York: Interscience.

Sensabaugh, G. F., Wilson, A. C. & Kirk, P. L. 1971 Protein stability in preserved biological remains, I & II. *Int. J. Biochem.* **2**, 545–557, 558–568.

Smith, G. G. & Evans, R. C. 1980 The effect of structure and conditions on the rate of racemization of free and bound amino acids. In *Biogeochemistry of amino acids* (ed. P. E. Hare, T. C. Hoering & K. King), pp. 257–282. New York: Wiley.

Smith, L. F. 1966 Species variation in the amino acid sequence of insulin. *Am. J. Med.* **40**, 662–666.

Smith, L. F. 1972 Amino acid sequences of insulins. *Diabetes, Suppl. 2* **21**, 457–460.

Stewart, C.-B., Schilling, J. W. & Wilson, A. C. 1987 Adaptive evolution in the stomach of foregut fermenters. *Nature, Lond.* **330**, 401–404.

Syvanen, M., Hartman, H. & Stevens, M. 1989 Classical plant taxonomic ambiguities extend to the molecular level. *J. molec. Evol.* **28**, 536–544.

Totten, D. K., Davidson, F. D. & Wyckoff, R. W. G. 1972 Amino acid composition of heated oyster shells. *Proc. natn. Acad. Sci. U.S.A.* **69**, 784–785.

Tuppy, H. 1958 Ür die Artspezifitäder Proteinstruktur. In *Symposium on protein structure* (ed. A. Neuberger), pp. 66–76. London: Methuen.

Wakabayashi, S., Matsubara, H. & Webster, D. A. 1986 Primary sequence of a dimeric bacterial haemoglobin from *Vitreoscilla*. *Nature, Lond.* **322**, 481–483.

Wald, G. 1952 Biochemical evolution. In *Trends in physiology and biochemistry* (ed. E. S. G. Barron), pp. 337–376. New York: Academic Press.

Weiner, S., Lowenstam, H. A. & Hood, L. 1976 Character-

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

ization of 80-million-year-old mollusk shell proteins. *Proc. natn. Acad. Sci. U.S.A.* **73**, 2541–2545.

White, R. H. 1984 Hydrolytic stability of biomolecules at high temperatures and its implication for life at 250 °C. *Nature, Lond.* **310**, 430–432.

Wilson, A. C., Carlson, S. C. & White, T. J. 1977 Biochemical evolution. *A. Rev. Biochem.* **46**, 573–639.

Woese, C. R. 1987 Bacterial evolution. *Microbiol. Rev.* **51**, 221–271.

Woese, C. R., Gibson, J. & Fox, G. E. 1980 Do genealogical patterns in purple photosynthetic bacteria reflect interspecific gene transfer? *Nature, Lond.* **283**, 210–212.

Zuckerkandl, E. & Pauling, L. 1962 Molecular disease, evolution and genic heterogeneity. In *Horizons in biochemistry* (ed. M. Kasha & B. Pullman), pp. 189–225. New York: Academic.

Discussion

G. A. Dover (*Department of Genetics, University of Cambridge, U.K.*). Internally repetitious genes and their corresponding proteins, such as collagen, are often not as 'dull' as is generally supposed. It is clear from the studies of Howard Green and colleagues on the involucrin gene in primates and of Alec Jeffreys and colleagues on loci of minisatellite DNA in humans, that the distribution of mutations among repeats in an assay can be mapped precisely using modern techniques, and that differences between assays in such distribution patterns are highly informative for monitoring lineages of assays, resolving phylogenies of closely related species and for understanding the genomic mechanisms, such as slippage and unequal crossing over, responsible for the diffusion of mutations in repeated assays. For short reviews of these studies and of the primary references see Dover (1990*a*, *b*) The message is: don't throw away the 'dull' repeats they could be useful!

References

Dover, G. A. 1990*a* Remodelling the involucrin genes and why we are not chimps. *Bioassays* **12**, 241–243.

Dover, G. A. 1990*b* DNA fingerprinting: mapping 'frozen accidents'. *Nature, Lond.* **344**, 812–813.

B. Halstead (*Department of Geology, Imperial College, London, U.K.*). It may well be the case, as Professor Ambler insists, that comparative studies of protein primary structures have not revolutionized any area of systematics, but what has emerged is that molecules, just like for example limbs, do evolve. Anomalies such as the sequence similarities of myoglobin between burrowing mammals like the mole and the viscacha rodent and between diving animals such as seal, whale and penguin, reflect 'hypoxic' substitutions, relating to environments of oxygen depletion. Likewise the similarities between birds and mammals relate to their high metabolic rates not any genetic relationship (Romero-Herrera *et al.* 1978; Joysey 1988).

I would like to speak to the defence of collagen. This structural protein may well appear unsatisfactory and dull and may not be ideal but it does have one great advantage, by virtue of its intimate association with hydroxyapatite in bones and teeth, it is in fact preserved in the fossil record. For this reason it has been our target molecule and our ideal 'experimental animal' enabling us to ascertain levels of contemporary (ancient) and present-day (modern) contamination. Furthermore it has been possible to obtain insights into processes of protein diagenesis and ageing. Professor W. G. Armstrong has succeeded in achieving peptide analyses of fossil collagen thus confirming the survival of intact proteins (Armstrong *et al.* 1983).

J. P. Thorpe (*Department of Environmental and Evolutionary Biology, University of Liverpool, Port Erin Marine Laboratory, Isle of Man, U.K.*). In relation to Professor Ambler's talk and to other discussion of the systematic uses of data on the divergence of protein molecules I would like to make some general points, most of which are also relevant to similar work on nucleic acids. Firstly the comparison of the divergence of homologous molecules from various taxa is generally based upon the assumption that this divergence is related to evolutionary time, the so called 'molecular clock' hypothesis. Although controversial it is now widely accepted that the structures of molecules do diverge with time, the controversy largely concerns the extent to which the rate of divergence may or may not be stochastically linear with time or may or may not vary between evolutionary lineages. Even if true requirement for stochastic linearity also makes the assumption that all substitutions are selectively neutral and that effective population sizes are roughly similar and relatively constant. Obviously, if it occurs to any significant extent, selection is likely to alter the rate of molecular evolution and also, even under neutralist expectations, constant or episodic small population sizes (i.e. population crashes or 'bottlenecks') will significantly increase the probability of random mutations going to completion within the population and thus will serve to increase the apparent rate of molecular change. Thus the aberrant rattlesnake sequence data mentioned by Professor Ambler could be explained either by rapid evolution in response to a particularly stressful environment (e.g. extreme desert conditions) or by population crashes or small effective population sizes.

The second point concerns the emphasis by Professor Ambler on the advantages of comparatively small protein molecules. As he points out these are, of course technically easier to work with. However, they are also likely to yield less valuable data because the 'sampling errors' expected from stochastic variation along will be relatively far greater in small proteins (e.g. 100–200 units) than much larger molecules like collagen, for example, with a few thousand units. it is also of note that because empirical data suggest that different proteins evolve at very different rates it is critical that suitable proteins are chosen if divergence data are to be of maximum use. Rapidly evolving proteins over prolonged timescales will have diverged so greatly that the high probabilities of back mutation and of multiple substitution at single sites make evolutionary interpretation both difficult and controversial. In practice, levels of substitution of over about 30% are probably best avoided. Conversely, proteins evolving too slowly will result in few substitution so that between lineages possible differences in levels of divergence may be masked by stochastic errors.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES