J. Molec. Evolution 1, 26–45 (1971)

The Structure of Cytochrome *c* and the Rates of Molecular Evolution*

RICHARD E. DICKERSON

Norman W. Church Laboratory of Chemical Biology
California Institute of Technology, Pasadena, California 91109

Received January 4, 1971

Summary. The x-ray structure analysis of ferricytochrome *c* shows the reasons for the evolutionary conservatism of hydrophobic and aromatic side chains, lysines, and glycines, which had been observed from comparisons of amino acid sequences from over 30 species. It also shows that the negative character of one portion of the molecular surface is conserved, even though individual acidic side chains are not, and that positive charges are localized around two hydrophobic "channels" leading from the interior to the surface.

The reason for the unusual evolutionary conservation of surface features in cytochromes *c* is probably the interaction of the molecule with two other large macromolecular complexes, its reductase and oxidase. This conservation of surface structure also explains the relatively slow rate of change of cytochrome *c* sequences in comparison with the globins and enzymes of similar size.

The rate of evolution of a protein is the rate of occurrence of mutations in the genome modified by the probability that a random change in amino acid sequence will be tolerable in a functioning protein. The observed rates of change in fibrinopeptides, the globins, cytochrome *c*, and several enzymes are interpreted in terms of the proteins' biological roles.

Key-Words: Cytochrome *c* — Evolutionary Rates and Molecular Structure.

Molecular evolution as a field of study really began in 1960, when Perutz and Kendrew observed from their x-ray analyses that the protein chains in hemoglobin and myoglobin were folded in an essentially identical manner. Subsequent amino acid sequence comparisons of myoglobins and hemoglobins, and x-ray analyses of other globins have shown that protein molecules contain a record of their evolutionary history that is fully as informative as that obtained from macroscopic anatomical features [5, 10, 23, 24].

The globins as tools for studying the process of molecular evolution have the advantages of easy accessibility and purification, reasonably small molecular weight, and wide distribution among vertebrates and many invertebrate species. They have the disadvantage of a multiplicity of chain types in the vertebrate hemoglobins which, although interesting in its own

* Contribution No. 4114 from the Norman W. Church Laboratory of Chemical Biology, California Institute of Technology, Pasadena, California 91109.

right, tends to confuse the central issue of molecular differentiation and evolution.

Cytochrome *c* has been studied in even greater detail by amino acid sequence methods. It is found in the mitochondria of every eukaryotic or nucleated cell, and hence is distributed throughout all living things above the level of bacteria and bluegreen algae. It is smaller than the globins—104 amino acids in vertebrates and a few more in other species. It is the easiest component of the mitochondrial respiratory chain to extract, and is reasonably stable when isolated. It is therefore a natural candidate for study, and the amino acid sequences have been determined by Margoliash, Smith, Boulter, and others for more than thirty-five species [3, 6, 18]. From these data it has become apparent that elaborate family trees can be constructed which reproduce faithfully (within present statistical limits of confidence) the evolutionary history of life as deduced from more traditional methods [8, 13–16]. Although the method has not yet been used to settle any of the major questions regarding the origins and interrelationships of phyla, it is clear that is has the potentialities for doing so.

Several investigators have observed that different proteins evolve in primary sequence at different rates, and have suggested in general terms that individual proteins must differ in the stringency of specifications for an operative molecule [5, 11, 15]. Cytochrome *c*, for example, has been observed to evolve significantly more slowly than the globins. It has been implied that to a first approximation at least, each protein has evolved at a roughly constant rate, so that it is meaningful to talk about a comparison of "the rates of evolution of hemoglobin and cytochrome *c*". This assumption has not been given a thorough test using all of the available paleontological evidence as to dates of branch points in the evolutionary tree. One of the purposes of this paper is to collect such data and examine the linearity assumption.

The second purpose of this paper is to correlate the observed rates of molecular evolution with biological function, in those proteins for which suitable data are available. In the absence of evidence to the contrary, we must assume that the rates of mutation of bases in the genomes for different proteins are not widely different, and that the difference in rates of sequence changes of hemoglobin and cytochrome *c* do not arise mainly from differences in mutation rates. What is observed is the result of mutations followed by natural selection or by some other process capable of establishing the mutations in a population. In general, the more stringent the restrictions on an operating molecule, the smaller the proportion of chance sequence alterations that will be acceptable, and the slower the primary sequence will evolve.

The structure of ferricytochrome *c* has recently been determined to a resolution of 2.8 Å [6]. With a knowledge of the biochemical role of cytochrome *c* and of its three-dimensional structure, it is possible to see *why* this protein has evolved more slowly than the globins. The relevant aspects of the cytochrome *c* structure are summarized in the following section. These results can then be generalized to enzymes and to subunit proteins, and lead to some general principles governing the rates of molecular evolution.

The Structure of Cytochrome *c*

The structure of horse heart ferricytochrome *c* has been determined by x-ray diffraction methods to a resolution of 2.8 Å, and the results have been extended to the closely similar proteins from bonito and tuna. A full report of this work is published in references [6] and [7]. The most pertinent aspects in considering the evolution of protein molecules are summarized below.

The 104 amino acids are wrapped around the heme group as shown in Fig. 1, leaving one edge of the heme exposed in the "heme crevice". The heme is attached covalently to the protein chain through cysteines 14 and 17, and the fifth and sixth of the octahedral coordination positions around the iron atom are filled by a ring nitrogen atom of histidine 18 and the sulfur of methionine 80.

The polypeptide chain is essentially a shell one layer thick. Residues 1–46 fold to build the right side of the molecule as seen in Fig. 1, residues 46–92 shape the left side, and the last residues 92–104 form an α-helical tail over the top to the right side once more. Most of the hydrophobic side chains point toward the heme and are packed closely around it. This accounts for the familiar observation that hydrophobic groups are generally conserved in cytochromes *c* from all species. In no species' cytochrome is a buried hydrophobic group in horse cytochrome ever found as a charged group or even as an uncharged but polar side chain. Strong evolutionary selection pressures evidently exist to weed out those DNA mutations which lead to radical changes in the interior of the molecule. Considerably less selection pressure exists in general against changes on the surface of cytochrome *c* and other protein molecules. This, of course, is quite understandable, and it would be surprising if it were otherwise.

The aromatic side chains—tyrosine, tryptophan and phenylalanine—are highly conserved in cytochromes *c* from different species. In most cases it is the aromaticity that seems to be important, but in specific instances the hydrogen-bonding capabilities of tyrosine and tryptophan are used. Tyrosine 67, tryptophan 59, and tyrosine 74 all occur near one another on the left side of the heme in Fig. 1. All three groups are invariant throughout evolution. The latter two are associated with a loop of chain (residues 57–74) which has been called the "left channel". Various chemical and intuitive

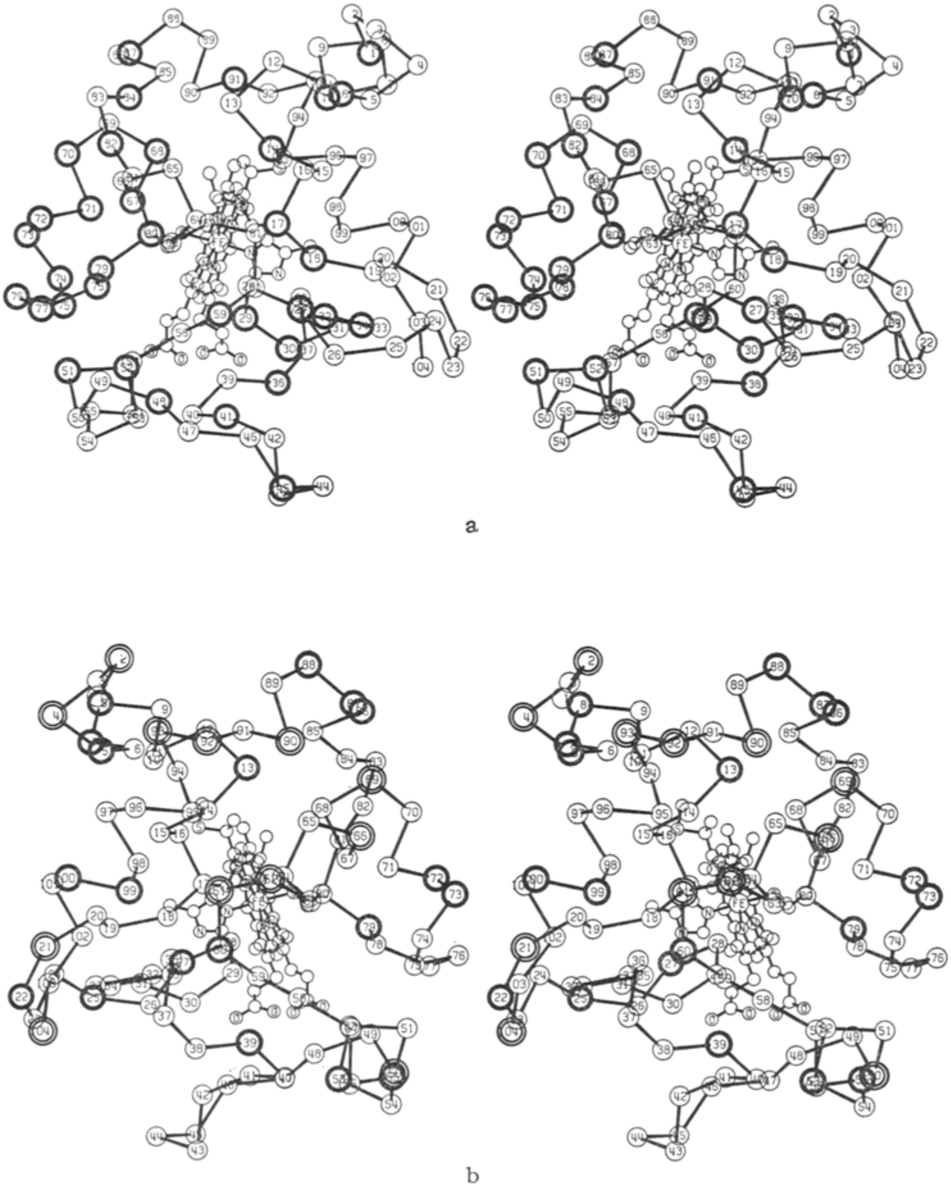

Fig. 1 a and b. Stereo α-carbon diagrams of horse ferricytochrome *c*, as shown by the 2.8 Å resolution x-ray analysis. Each amino acid residue is represented by a numbered circle at the position of its α-carbon atom, and peptide bonds are symbolized by straight lines connecting these atoms. The heme group and those side chains that interact with the heme are shown in their entirety. (a) Front view, with heavy circles marking those amino acids that are totally invariant among 29 species as listed in Reference [6]. (b) Back view, with the 19 lysines marked by heavy circles, and the 12 acidic groups (Asp and Glu) marked by double circles. Note the strong localization of charge, with negative charges at the top of this view and positive charges to either side lines of argument detailed in references [6] and [7] suggest that these aromatic groups, the "channel", and the left side of the molecule may all be involved in the reduction mechanism and in interaction with the cytochrome reductase complex. The totally invariant sequence 70–80, commented upon by Margoliash and others from sequence comparisons, is thus explained, if it is a key component of one of the redox mechanisms of the protein.

The aromatic side chains phenylalanine 10 and tyrosine 97 are associated with another structural feature, the "right channel", defined by loop 6–21 and α-helix 92–102 (Fig. 1). This feature and the heme crevice may be the sites of interaction with the cytochrome oxidase complex. Chemical modification of lysine 13 at the top of the crevice, for example, drastically reduces the reaction with oxidase. Phenylalanine 10 is invariant, but tyrosine 97 is sometimes found as phenylalanine. Its —OH group extends away from the molecular surface.

Two more aromatic groups are paired at the base of the heme crevice. The invariant tyrosine 48 is hydrogen bonded to the more deeply buried propionic acid chain of the heme. Phenylalanine 46 can also occur as tyrosine, and in bonito and tuna, this tyrosine appears to be hydrogen-bonded to the more exposed heme propionic group. The evolutionarily invariant phenylalanine 82 lies at the top of the heme crevice where it could interact with the oxidase complex. Only phenylalanine 36, buried in the surface at the back of the molecule, appears to be unimportant. It is the only aromatic residue which can be replaced by a non-aromatic one, although the replacement (isoleucine) is still bulky and hydrophobic. In summary, the selection pressures which insure the retention of aromatic side chains in cytochrome *c* appear to be the need for aromaticity, the existence of hydrogen bonds (in some cases), and in one case merely the need for a large nonpolar residue. The conservation of these residues and their distribution in the molecule suggest that they must play a role in the functioning of cytochrome *c*.

The charged side chains are all at the surface of the molecule, in a highly asymmetric and nonrandom arrangement. The nineteen lysines are distributed around the periphery of the right and left "channels", on either side of the molecule as seen from the back in Fig. 1b. The upper rear of the molecule, between these two positively charged zones, is totally devoid of basic groups. It is occupied instead by acidic side chains, creating a "negative patch". Since it is known that the interaction of cytochrome *c* with its oxidase involves electrostatic forces, with the positive charge being on the cytochrome *c* molecule, this nonrandom distribution of lysines is a strong clue that the right or left "channel" may be the binding site for the oxidase, as suggested earlier.

The two positive regions and the negative patch are conserved in the course of evolution in slightly different ways. Many of the lysine are totally invariant, or are replaced only by arginine in certain branches of the family tree. In contrast, *none* of the twelve acidic side chains is totally invariant, and only glutamic acid 90 remains acidic in every species. The effective selection pressure appears to be not for retention of negative charge at specific loci in the polypeptide chain, but rather for retention of overall negative character of the upper rear of the cytochrome molecule. The specific sites of negative charge vary from one species to another, but no species has fewer than six negative charges within this region of the molecular surface (horse has nine), and no species has more than five negative charges scattered anywhere else over the surface (horse has three). This is a dramatic illustration of the principle that natural selection operates on the *folded* protein molecule, and not on the sequence of amino acids.

Cytochrome *c*, in summary, is an unusually conservative molecule, as many people have pointed out in the past from sequence comparisons. It is strikingly conservative in its retention of hydrophobic groups, aromatic residues, glycine, lysine, and in the distribution of acidic groups on the molecular surface if not in the amino acid sequence. Although cytochrome *c* is as old or older than the hemoglobins, it does not appear to have changed as much. In contrast, if we look at other proteins such as the fibrinopeptides [17], we see that they have changed much faster than the globins. Different proteins apparently change at different rates, and we are now in a position to examine more carefully the matter of the relative rates of evolution of proteins and the reasons for them.

The Rates of Molecular Evolution

The proteins for which enough sequence information is available to make statistically meaningful comparisons are cytochrome *c*, hemoglobin, and the fibrinopeptides. A convenient tabulation of these and other sequences is to be found in reference [3], along with matrix tables which give the number of amino acid differences between sequences from different species. The species whose sequences are compared in this paper are listed in Table 1. All possible species comparisons between two classes of organisms have been made for each of the three proteins, and the average differences between classes are given in Table 2. As an example, the comparison of cytochrome *c* from man and rhesus monkey with seven other mammalian sequences leads to a mean difference between primates and other mammals of 10.1 residues out of 104, with an average deviation from the mean of 0.8 residues. For purposes of comparison between proteins, these figures have been normalized to a common base of 100 amino acids. They have then been corrected for the possibility that two or more amino acid changes have occurred at the same place, which would not be detectable from the comparison of two present-day sequence. If *m* is the total number of amino

Table 1. *Species whose sequences are compared in this paper*

A. Cytochrome *c*	B. Hemoglobin α	B. Hemoglobin β	C. Fibrinopeptides A and B
Man	Man	Man	Primates
Rhesus monkey	Rhesus monkey	Gorilla	Man
Horse	Gorilla	Spider monkey	Green monkey
Donkey	Horse	Rhesus monkey	Rhesus monkey
Pig (Cow, Sheep)	Cow	Horse	Drill
Dog	Sheep	Goat (A and C)	Rodents, Lagomorphs
Gray whale	Goat (A and B)	Cow (fetal and B)	Rat
Rabbit	Llama	Sheep (A, B and C)	Rabbit
Kangaroo	Rabbit	Barbary sheep (C)	Carnivores
Chicken (Turkey)	Pig	Llama	Dog
Penguin	Mouse	Rabbit	Fox
Pekin duck	Carp	Pig	Cat
Pigeon		Mouse	Perissodactyls
Rattlesnake			Horse
Snapping turtle			Donkey
Bullfrog		B. Other Globins	Zebra
Tuna			Artiodactyls
Dogfish		Human γ	Pig
Lamprey		Human δ	Llama
Screwworm fly		Sperm whale myoglobin	Vicuna
Fruit fly		Lamprey globin	Camel
Silkworm moth			Cow
Tobacco horn worm moth			European bison
Baker's yeast			Cape buffalo
Neurospora crassa			Water buffalo
Candida krusei			Sheep
Wheat			Goat
Mung bean			Pronghorn
Sunflower			Reindeer
Sesame seed			Mule deer
Castor			Muntjak
			Sika deer
			Red deer
			American elk
			Marsupials
			Kangaroo

acid changes which have occurred in a 100 residue chain, then the number, n, which will actually be seen, allowing for masking of some changes by subsequent changes at the same locus in one or the other of two divergent sequences, is[1]:

$$\frac{n}{100} = 1 - e^{-m/100}$$

[1] This simple expression ignores mutations which revert to the original amino acid. These, although invisible, technically should be included in the observed quantity n. The inferred quantity m will therefore be systematically slightly too small, but this first order correction is good enough to make the biochemical points that are the subject of this paper. More intricate corrections would add complexity but very little more chemical insight.

or:

$$\frac{m}{100} = -\ln\left(1 - \frac{n}{100}\right).$$

These corrected values are also listed in Table 2. By themselves, they measure the degree of difference between sequences in classes of organisms, and agree qualitatively well with our ideas of the degree of evolutionary relatedness of these classes. But they become much more informative if the dimension of time is added.

The question of dates for the branch points in molecular trees of evolution has not received the attention it deserves, principally because the people most concerned with sequences have not been biologists, and the zoologists who have the best command of information on dates have maintained a reserved skepticism toward the entire protein endeavor. The author can make no claim for expertise in this area, but has tried to adopt the best values he could find from reliable references. The dates in Table 3 hopefully will at least serve as a starting point for discussion by others.

The branch points used are diagrammed in Fig. 2. The dating of these branches is difficult. The branch point between mammals and reptiles, for example, is not the time at which the mammals first became dominant, or even the time at which clearly defined mammals first appeared. Rather,

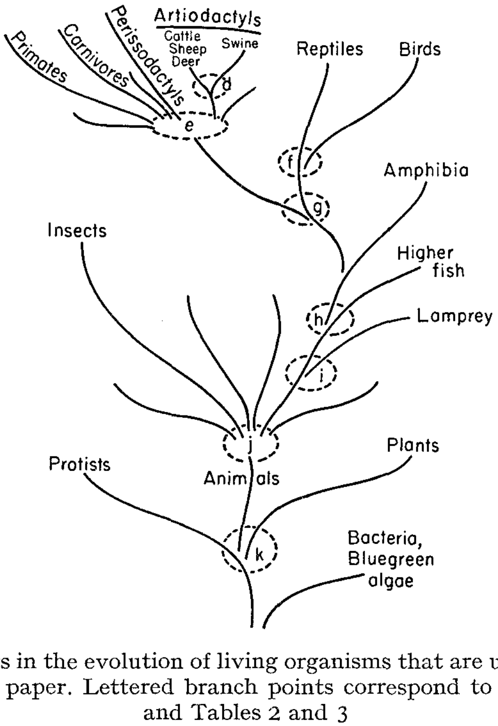

Fig. 2. Branch points in the evolution of living organisms that are used in the sequence comparisons in this paper. Lettered branch points correspond to the usage of Fig. 3 and Tables 2 and 3

Table 2. Amino

Branch point	Residues[a]	Number of comparisons
A. Cytochrome *c*		
Primates/Other mammals	104	14
Horse, donkey/Other nonprimate mammals	104	10
Birds/Reptiles	104	8
Mammals/Birds	104	35
Mammals/Reptiles	104	18
Fish/Other vertebrates	104	46
Vertebrates/Insects	108	74
Moths/Flies	108	4
Animals/Plants/Protists	112	180
B. Globins		
Human, gorilla β/Other primate β	146	4
Primate δ/Primate β	146	8
Primate β, δ/Other mammalian β	146	84
Between nonprimate orders, β	146	55
Human γ/Mammalian β, δ	146	20
Between orders of mammalian α	141	52
Mammalian α/Carp α	141	12
All α/All β, γ, δ	146	56
Lamprey globin/Hemoglobins	159	15
Myoglobins/Hemoglobins	151	30
Lamprey globin/Myoglobins	165	2
Hemoglobins/Insect globin	151	15
Myoglobins/Insect globin	157	2
Lamprey globin/Insect globin	160	1
C. Fibrinopeptides		
Deer/Cattle	42	36
Dog, Fox/Cat	42	2
Camel, Llama, Vicuna/Deer and Cattle	42	36
Swine/Other artiodactyls	42	15
Between orders of mammals	42	270

[a] When two chains with different lengths are compared in which the chain differences may be presumed to arise from deletions (as hemoglobin α and β), a deletion is classified as a 21st kind of amino acid in counting changes.

it is the time at which the reptilian ancestors of the mammals separated from those other reptiles which would ultimately give rise to the living reptiles. This means that the branch dates are generally earlier than those which might be assumed by a cursory inspection of the paleontological record. The principal source of dates for branch points has been Young [22] although Romer [19, 20] and other sources have been in general agreement. The dating of geological periods in that of Romer [19], Kulp [12], and Chapacid changes

Changes per molecule	Changes per 100 residues, *n*	Changes per 100 residues, corrected for multiple hits, *m*	Branch	Age (MY)	Point on graph
10.1 ± 0.8	9.7 ± 0.8	10.2 ± 0.9	e	90	
5.1 ± 1.3	4.9 ± 1.3	5.0 ± 1.3	e	90	
13.1 ± 5.4	12.7 ± 5.2	13.6 ± 6.0	f	240	
9.9 ± 1.4	9.5 ± 1.4	10.0 ± 1.5	g	300	
14.8 ± 4.3	14.2 ± 4.1	15.3 ± 4.8	g	300	
18.7 ± 2.1	18.0 ± 2.0	19.8 ± 2.4	f	400	
26.6 ± 2.6	25.6 ± 2.5	29.6 ± 3.3	j	600	
13.8 ± 0.9	13.2 ± 0.9	14.2 ± 1.0	—	?	
47.1 ± 2.0	45.2 ± 1.9	60.1 ± 3.5	—	?	1
7.0 ± 0.5	4.8 ± 0.3	5.0 ± 0.3	—	?	2
10.0 ± 1.0	6.8 ± 0.7	7.0 ± 0.7	—	?	3
27.0 ± 3.7	18.5 ± 2.5	20.4 ± 3.1	e	90	
31.2 ± 3.8	22.0 ± 2.6	24.8 ± 3.3	e	90	
39.7 ± 1.3	27.1 ± 0.9	31.6 ± 1.2	—	?	4
21.2 ± 3.4	15.0 ± 2.4	16.3 ± 2.8	e	90	
70.4 ± 1.4	49.8 ± 1.0	68.9 ± 2.0	h	400	
86.5 ± 1.7	58.1 ± 1.2	87.0 ± 2.9	i	500	
119.0 ± 3.6	75.0 ± 2.3	138.6 ± 9.2	—	?	5
117.8 ± 1.8	78.0 ± 1.2	151.4 ± 5.5	—	?	6
130.5 ± 1.5	79.1 ± 0.9	156.5 ± 4.3	—	?	7
121.1 ± 3.3	80.2 ± 2.2	161.9 ± 11.0	—	?	8
126.5 ± 0.5	80.5 ± 3.2	163.5 ± 16.0	—	?	9
132	82	174.2	—	?	10
13.0 ± 1.5	31.0 ± 3.6	37.1 ± 5.2	a	33	
14.0	33.3	40.5	b	45	
15.8 ± 1.0	37.6 ± 2.4	47.1 ± 3.8	c	50	
18.5 ± 2.6	44.0 ± 6.2	58.0 ± 11.1	d	60	
23.2 ± 3.1	55.2 ± 7.4	80.3 ± 16.4	e	90	

ter 1 of Young [22], which places the beginning of the Cambrian at 600 million years (MY) ago, rather than the shorter time scale used by Young in his later diagrams.

On the right of Table 2, the branches between classes of organisms as lettered in Fig. 2 are identified, and their ages from Table 3 are given. Fig. 3 shows these data in a plot of corrected amino acid differences between two lines of evolution per 100 residues, against the time in the past at which these lines diverged. The limits of error as measured by the average deviations from the mean are represented by the lengths of the vertical bars. The individual branch points are identified at the top of the graph. Paleonto-

Table 3. *Dating of branch points*

Branch	Period	Age (MY)	Reference
a) Divergence of deer and cattle	Middle Oligocene	33	Young [22], p. 747
b) Divergence of dog and cat families	Late Eocene	45	Young, p. 678
c) Divergence of camel, llama, vicuna from deer and cattle	Middle Eocene	50	Young, p. 747
d) Separation of swine from other artiodactyls	Early Eocene	60	Young, p. 747
e) Differentiation of various orders of mammals	Middle and late Cretaceous	90	Young, p. 574, Szalay [21]
f) Divergence of birds and reptiles	Via pseudosuchian reptiles of early Triassic, back to eosuchians of late Permian	240	Young, p. 402, 417, 510
g) Divergence of mammals and reptiles	Via therapsid reptiles of middle Permian and late Triassic, back to anapsids of Permian, ultimately to cotylosaurs of late Carboniferous	300	Young, p. 389, 539–541
h) Divergence of higher vertebrates and fish	Devonian	400	Young, p. 356
i) Divergence of carp and lamprey lines	Early Ordovician	500	Young, p. 177
j) Divergence of vertebrate and insect lines, radiation of invertebrate phyla	Early Cambrian	600	Young, Ch. III

logically datable points for cytochrome *c* extend as far back as 600 MY, during the separation of vertebrates from other animal phyla. The earliest point for hemoglobin is 500 MY, when lamprey and the higher fish diverged. All of the fibrinopeptide data involve orders of mammals. The first observation is that the data for all three proteins do fall on straight lines. Within the limits of error shown, the rate of change of amino acid sequence has remained essentially unchanged for each protein.

A second observation is that these average rates of change are quite different for the three proteins. Extending a definition proposed by Nolan and Margoliash (Reference [18], p. 739), we can define the *Unit Evolutionary Period* (UEP) as the time in millions of years for a one percent change in amino acid sequence to show up between two divergent lines of evolution. The UEP for cytochrome *c* is then 20.0 MY, that for hemoglobin is 5.8 MY, and that for the fibrinopeptides is only 1.1 MY.

What is observed in these figures is the rate of accumulation of changes in amino acids from mutations in DNA, after those changes which produce an unworkable molecule have been weeded out. Natural selection ensures

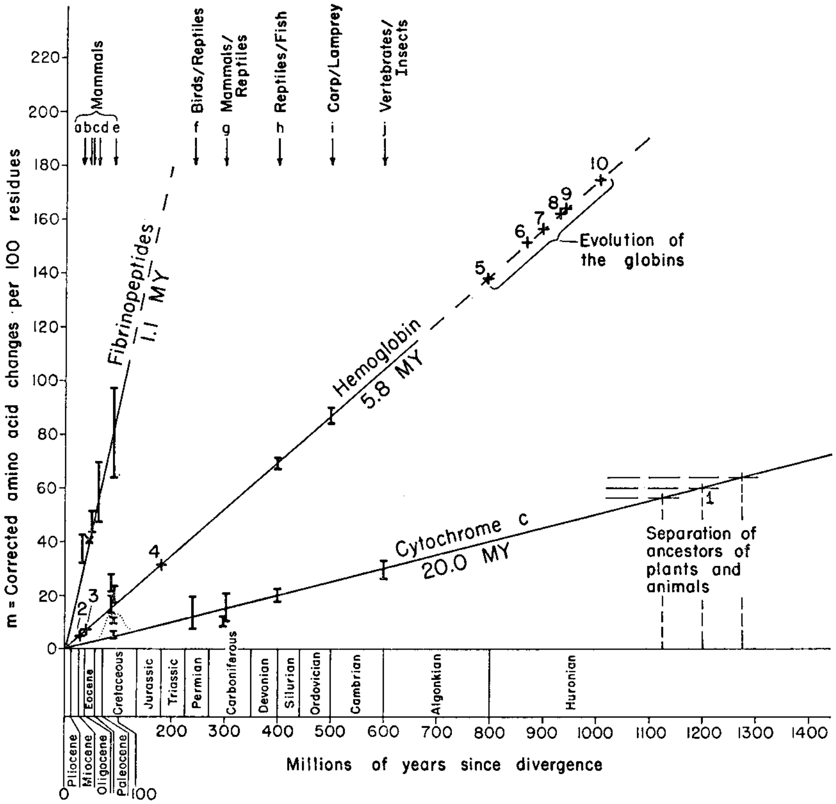

Fig. 3. The rates of macromolecular evolution in the fibrinopeptides, hemoglobin, and cytochrome *c*. The amino acid differences between divergent lines of evolution, corrected for multiple changes at the same locus, are from Table 2. The dating of branch points in evolution is from Table 3. Mean errors in amino acid differences are indicated by vertical bars. Comparisons for which no adequate time coordinate is available are indicated by numbered crosses. Point *1* represents a date of 1200 ± 75 MY (million years) for the separation of plants and animals, based on a linear extrapolation of the cytochrome curve. Points *2–10* refer to events in the development of the globin family, as detailed in Table 2. The δ/β separation is at point *3*, γ/β is at *4*, and α/β is at 500 MY (carp/lamprey). The *Unit Evolutionary Period* in MY is given by each curve. An earlier version of this figure has appeared in Ref. [5]

that the unviable molecules die out of the genetic record. The more finely specified a protein molecule must be for proper operation, the less likely it will be that a random change in amino acids is acceptable, and the longer the *Unit Evolutionary Period* of the protein. This is illustrated in Fig. 4. The fibrinopeptides have little known function after they are cut out of fibrinogen when it is converted to fibrin in a blood clot. Virtually any amino acid change that still permits the peptides to be removed may be acceptable, and the rate of change of these peptides may be close to the actual rate of mutation of DNA.

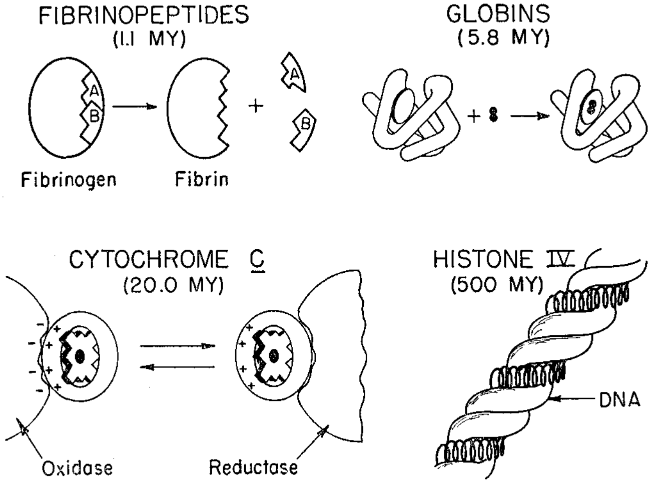

Fig. 4. The more complex the interactions of a protein with other molecules or macromolecules, the longer will be its *Unit Evolutionary Period*. The discarded fibrinopeptides have a UEP of slightly over 1 MY; the Histone IV bound to DNA within the nucleus has a UEP 500 times as long. The UEP for cytochrome *c* is longer than that for the globins primarily because cytochrome *c* interacts with other macromolecular complexes, whereas hemoglobin binds to O_2 and CO_2 in solution

Hemoglobin has a quite definite function: that of binding O_2 in solution at the lungs, and interacting with CO_2 at the tissues. The specifications for a hemoglobin molecule are more restrictive than for a fibrinopeptide, and the UEP is correspondingly almost six times as long. Cytochrome *c* interacts, not with small molecules, but with other macromolecules. It is involved with cytochrome oxidase and reductase, both of which are macromolecular complexes much larger than it is, and probably with the mitochondrial membrane as well. The x-ray analysis has shown that a large proportion of the surface of cytochrome *c* is structured and conserved throughout evolution—especially the distribution of charged groups and the locations of aromatic residues. The limits on a functioning cytochrome *c* molecule are evidently narrower than for hemoglobin, and its UEP is longer yet.

Histone IV is an interesting extension of these ideas. In the 102 amino acids of histone IV from calf thymus and pea seedling, only two changes were found [4]. The assumption of a provisional date from Fig. 3 of 1 000 to 1 200 MY for the divergence of plants and animals yields a statistically shaky but suggestive UEP of 500 MY for this protein. Again this is understandable in terms of the role of the protein. Histone IV binds to DNA in the nucleus, and is believed to have a control function in the expression of the information coded in the nucleic acids. It is part of the machinery of the archives of the cell, and it is hardly surprising that a protein so close to the genetic information storage system would be closely specified.

Limits of the Linearity Approximation

The selection pressures which shape evolution are far from constant throughout any conceivable time period. The only reason that protein molecules show the linearity of Fig. 3 is because they are so well adapted for their roles and so well shielded from the immediate effects of selection, which operates on populations of whole organisms. If we were to examine a protein during the period in which it was being used for a new function, we would find much more intense selection pressures in favor of a new structure, and consequently a more rapid change in sequence.

A good example of this is the comparison of lysozyme and α-lactalbumin. It is now generally accepted that α-lactalbumin, one component of a lactose synthesis system, evolved from the precursors of the lysozymes when milk-producing animals developed, or approximately 100–150 MY ago [5]. The UEP for lysozyme, as estimated from human and hen sequences in Table 4, is of the order of 5.3 MY. But α-lactalbumin is far too unlike lysozyme for the approximation of a linear rate of evolution to be valid. If linearity were assumed blindly, then the data in Table 4 would tell us that α-lactalbumin diverged from the lysozymes 390 MY ago in the early Devonian, which is phylogenetically absurd. A more reasonable explanation is that the *change* in pressures exerted on α-lactalbumin when a polysaccharide degrading enzyme began functioning in a polysaccharide synthesizing role led to an increase in the rate at which mutational changes were retained in the sequence rather than being weeded out as harmful to the *status quo*.

It is wrong, therefore, to try to compare two *different* proteins and assume a constant rate of evolution between them. The linear approximation is valid only when comparing members of a family of proteins, all of which are performing the same function in similar circumstances. The extrapolation of the cytochrome *c* data to the separation of plants and animals (Fig. 3, point *1*) is probably valid because the mitochondrial terminal oxidation chain of which cytochrome *c* is a member had evolved *prior* to this stage in the history of life. (It also agrees with the pre-Cambrian record of the earliest metazoan fossils [9].) In contrast, points 5–10 of Fig. 3, involving comparisons between monomeric and tetrameric hemoglobins from vertebrates and insects, and hemoglobin with myoglobin, are too early for the linear approximation to be valid. The oxygen transporting system probably evolved along with the metazoans, long after the evolution of the terminal oxidation chain and cytochromes. Hemoglobin and myoglobin, although closer in function, are analogous to lysozyme and α-lactalbumin—descendants of a common ancestor but now proteins playing different roles. If increased selection pressure causes more rapid changes, then the branch point for two diverging proteins will appear too ancient. The figure of 940 MY for the divergence of myoglobin and hemoglobin is therefore probably an exaggerated upper limit which is too great by at least 200 MY.

Table 4. *Fragmentary data from other proteins*

Species comparison	Residues	Number of comparisons	Changes per molecule	Changes per 100 residues, n	Changes per 100 residues, corrected for multiple hits, m	Branch	Age (MY)	UEP (MY)	Ref.
D. *Proinsulin polypeptide C*									
Man/Cow, pig	35	2	13.5	38.5	48.4	e	90	1.9	[2]
E. *Ribonuclease*									
Rat/Cow	127	1	44	34.6	42.5	e	90	2.1	[3]
F. *Lysozyme and α-lacto-globulin*									
Chicken/Human lysozyme	132	1	57	43.2	56.6	g	300	5.3	[5]
Human lysozyme/Bovine α-lactalbumin	132	1	82	62.1	73.6	—	?	?	[5]
G. *Trypsinogen* (incomplete sequence)									
Dogfish/Cow	123	1	38	31	37	h	400	10.8	[1]
H. *Insulin chains A and B*									
Chicken/Mammals	52	10	6.9 ± 1.1	13.3 ± 2.1	14.3 ± 2.4	g	300	21.0	[3]
Fish/Mammals, chicken	52	55	16.5 ± 2.0	31.7 ± 3.8	38.1 ± 5.4	h	400	10.5	[3]
I. *Glyceraldehyde-3-phosphate dehydrogenase*									
Pig/Lobster	334	1	93	27.8	32.6	j	600	18.4	[3]

Relative Rates of Macromolecular Evolution

The ideas of the previous section can be illustrated with enzymes where not enough sequences are yet known for statistically valid comparisons to be made, but where rough figures can be obtained. Some of these are listed in Table 4, in order of increasing *Unit Evolutionary Period.* The part of proinsulin that is cut out when insulin is produced, proinsulin C, has a UEP almost as short as the fibrinopeptides, and for the same reason. The hormonally active A and B chains of insulin, in contrast, evolve more slowly. The great difference between fish and higher vertebrate insulins which produces disagreement in insulin UEP figures is probably a reflection of the changing role of the hormone, of the type that has been mentioned for lysozyme and for hemoglobin.

The enzymes ribonuclease, lysozyme, trypsinogen, and glyceraldehyde-3-phosphate dehydrogenase illustrate a general correlation of increasing UEP with increasing size. The larger the protein molecule, the greater the ratio of interior to exterior amino acids. Aside from the active site, most of the surface of an enzyme is devoid of precise stereochemical function, and tolerant of substitutions of one amino acid for another of similar type. In the interior where side chains pack tightly together, in contrast, substitution of even a Val for a Leu is frequently intolerable. Interior hydrophobic residues tend to be more conservative than do surface polar groups. The UEP for a protein accordingly rises with molecular weight, other factors being equal, because more of the side chains are buried and subject to stringent size criteria. If for some reason a greater fraction of the surface is critical than in small monomeric enzymes, then the UEP will be correspondingly longer. Hemoglobin and GPDH can be thought of either as large, four-chain molecules with a large ratio of interior to exterior amino acids, or alternatively as aggregations of four smaller molecules in which

Table 5. *Charged groups in proteins whose molecular structures are known*

Protein	Amino acids	Charged groups[a]	Percent charged	UEP (MY)[b]
Cytochrome *c*	104	33	32	20.0
Ribonuclease	124	24	19	(2.1)
Lysozyme	129	27	21	(5.3)
Myoglobin	153	40	26	(5.8)
Papain	200	33	16	—
Chymotrypsin	241	32	13	(10.8) (from trypsinogen)
Subtilisin	275	28	10	—
Carboxypeptidase	307	53	17	—
Hemoglobin	574	82	14	5.8

[a] Lys, Arg, Asp, Glu.

[b] Figures in parentheses are approximations from inadequate data.

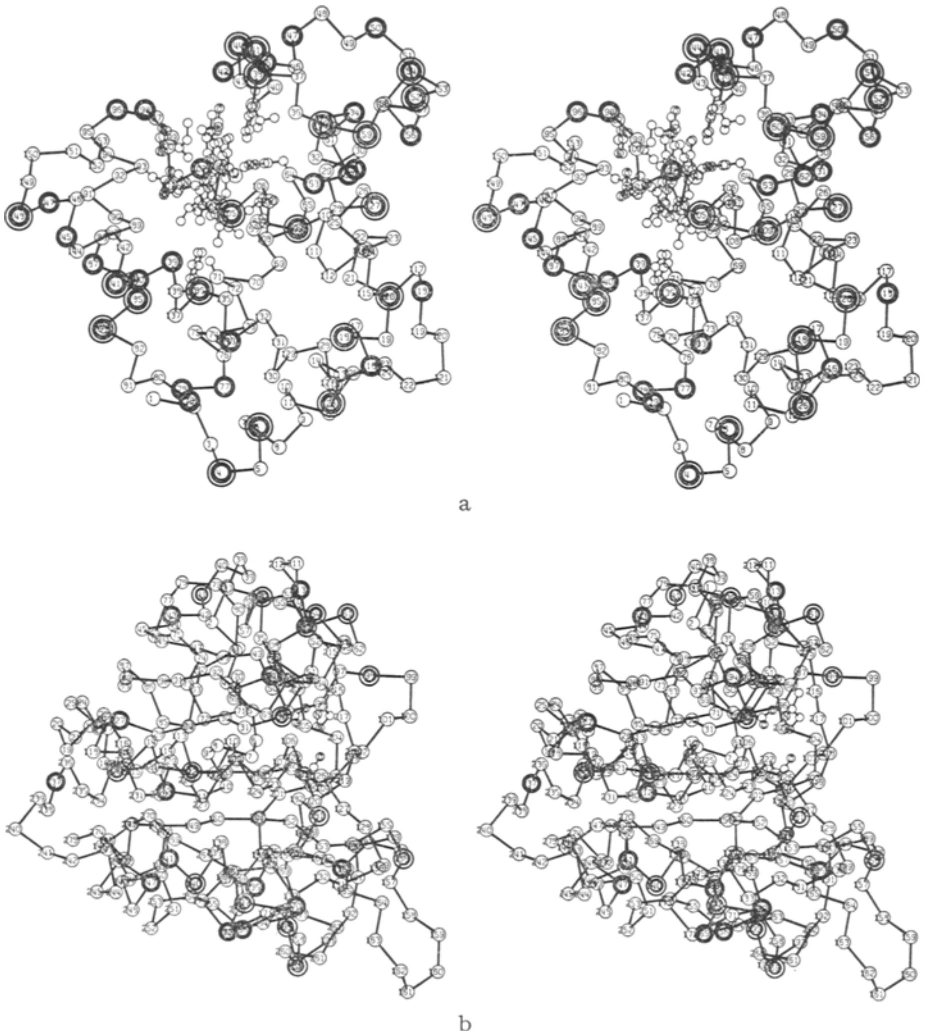

Fig. 5. (a) Stereo diagram of the α-carbon positions in sperm whale myoglobin, with basic side chains indicated by dark rings and acidic residues by double circles around the α-carbons. No localization of positive and negative charge analogous to that of cytochrome *c* is seen in myoglobin. (Coordinates courtesy of J. C. Kendrew, Cambridge.) (b) Stereo diagram of the α-carbon positions in subtilisin BPN', with basic and acidic side chains indicated as before. No cytochrome-like segregation of plus and minus charge on the surface is seen in this protein, either. (Coordinates courtesy of Joseph Kraut, La Jolla)

the mating requirements of one subunit with another place unusual restrictions on surface amino acid changes.

Cytochrome *c* evolves at an anomalously slow rate for so small a protein. It is as slow to change as the giant four-subunit enzyme GPDH. It is also anomalous in having a higher ratio of charged side chains—Lys, Arg, Asp and Glu—than other proteins (Table 5). Acidic and basic side chains have two main functions in most globular proteins: to keep those portions of the polypeptide chain where they occur outside, and to regulate the pK of the protein. The percentage of charged groups falls with increasing molecular weight, as the surface to volume ratio falls. Cytochrome *c* has far more charged groups than this trend in the other proteins would suggest.

The slow evolution and large number of charged amino acids in cytochrome *c* are not independent. As we have seen, the x-ray analysis of horse and bonito oxidized cytochrome *c* has shown that acidic and basic residues are not distributed at random over the surface, but are segregated in a negative patch flanked by two positive regions. The charged character of these regions is maintained throughout the entire evolutionary history of cytochrome *c*. No such segregation of charge is seen in any other globular protein whose structure is known. Myoglobin and substilisin illustrate the more common pattern in Fig. 5a and b: positive and negative charges scattered more or less at random over the surface.

In cytochrome *c*, the total picture of the heme in its heme crevice, the hydrophobic patches to right and left surrounded by positive charges, the negative patch at the upper rear of the molecule, and the paired aromatic rings, creates the impression of a molecule carefully structured for complex interactions with other molecules. One of the channels and positive regions is almost certainly the binding site to cytochrome oxidase, and the other channel and the negative patch may be involved with the reductase complex and the mitochondrial membrane. These well-defined surface features of cytochrome *c* are preserved throughout the entire history of evolution of microorganisms, plants, and animals, judging from sequence comparisons from over 30 species. Virtually the entire molecular surface is an "active site" for interaction with other molecules. Viewed another way, cytochrome *c* itself is a carefully tailored substrate for the two enzyme complexes: reductase and oxidase. Selection pressures against variation of the interior are similar to those for any other compact, globular protein. But selection pressures against variation of the *surface* of the cytochrome *c* molecule are more rigorous than for enzymes of comparable size. Only multisubunit proteins such as GPDH and hemoglobin have similar restrictions on surface structure, and it is these molecules that cytochrome *c* most resembles in UEP and rate of evolution.

Conclusion

The sequence comparison data now available for fibrinopeptides, hemoglobin and cytochrome *c* show that to a good first approximation, the rate of evolution of a given protein performing a well established function is constant. Furthermore, the rate of evolution as described by a characteristic *Unit Evolutionary Period* varies with the nature of the specifications for a functioning protein. What is observed is the rate of mutation of the nucleic acids, modified by the elimination of nonfunctional molecules. The more rigid the specifications for the molecule, the slower the accumulation of *acceptable* amino acids, and the longer the *Unit Evolutionary Period* will be.

The UEP in globular proteins generally rises with increasing molecular weight, as the ratio of interior to surface amino acids rises. This is so because the requirements for close-packed interior groups are usually more stringent than for side chains that extend out from the surface of the molecule. The UEP also increases if the protein has unusual requirements for surface structure. The active sites take up much the same fraction of the molecular surface in most enzymes. But multiple-subunit enzymes with strict requirements for structures of contact regions between subunits will have longer UEP's than monomeric enzymes of similar subunit size. Cytochrome *c* has unusual surface specifications for another reason: it must interact with two and possibly three other macromolecular complexes in the mitochondrion. The UEP for cytochrome *c*, therefore, is as long as for the much larger subunit enzyme GPDH. The actual structural requirements on the surface of cytochrome *c* can be seen from the x-ray analysis as a separation in different regions of positive and negative charge, hydrophobic patches, and the disposition of aromatic rings. No such segregation of charged side chains is visible in other globular proteins such as myoglobin or subtilisin. The extensive conservatism of cytochrome *c* and its abnormally long *Unit Evolutionary Period*, in summary, are directly interpretable in terms of the structure and the function of the protein molecule.

This work was performed with the support of National Science Foundation Grant GB-6617 and National Institutes of Health Grant GM-12121, whose help is gratefu ly acknowledged.

Note Added in Proof. In order to facilitate comparisons between this paper and the preceeding ones by Kimura and Ohta, the following conversions may be useful:

This paper *Kimura and Ohta*

$$m/100 = K_{aa}$$

$$\mathrm{UEP} = \mathrm{T}/m \quad \text{vs.} \quad k_{aa} = K_{aa}/2\mathrm{T}$$

$$\mathrm{UEP} = 1/200\, k_{aa}$$

or, in more convenient form:

$$\mathrm{UEP}\ (\text{in MY}) = 5/k_{aa}\ (\text{in Paulings}).$$

The Unit Evolutionary Period and Kimura's k_{aa} convey precisely the same information. We prefer the UEP, however, because of the immediate impression that it gives of *rates* of molecular evolution in comparison with geological time periods.

References

1. Bradshaw, R. A., Neurath, H., Tye, R. W., Walsh, K. A., Winter, W. P.: Nature (Lond.) **226**, 237 (1970).
2. Chance, R. E., Ellis, R. M., Bromer, W. W.: Science **161**, 165 (1968).
3. Dayhoff, M. O.: Atlas of protein sequence 1969. Silver Spring, Maryland: Nat. Biomed. Res. Found. 1969.
4. DeLange, R. J., Fambrough, D., Smith, E. L., Bonner, J.: J. biol. Chem. **244**, 5669 (1969).
5. Dickerson, R. E., Geis. I.: The structure and action of proteins. New York: Harper & Row 1969.
6. — Takano, T., Eisenberg, D., Kallai, O. B., Samson, L., Cooper, A., Margoliash, E.: J. biol. Chem. **246**, 1511 (1971).
7. — — Kallai, O. B., Samson, L.: Proceedings of the Wenner-Gren Symposium on Oxidation-Reduction Enzymes. Stockholm, August 1970. In press.
8. Fitch, W. M., Margoliash, E.: Science **155**, 279 (1967).
9. Glaessner, M. F.: Earth Sci. Rev. **1**, 29 (1966).
10. Ingram, V. M.: The hemoglobins in genetics and evolution. New York: Columbia U.P. 1963.
11. King, J. L., Jukes, T. H.: Science **164**, 788 (1969).
12. Kulp, J. L.: Science **133**, 1105 (1961).
13. Margoliash, E.: Proc. nat. Acad. Sci. (Wash.) **50**, 672 (1963).
14. — Fitch, W. M.: Ann. N.Y. Acad. Sci. **151**, 359 (1968).
15. — — Dickerson, R. E.: Brook. Symp. Biol. **21**, 259 (1968).
16. — Smith, E. L.: In: Evolving genes and proteins (V. Bryson and H. J. Vogel, eds.). New York: Academic Press 1965.
17. Mross, G. A., Dolittle, R. F.: Arch. Biochem. **122**, 674 (1967).
18. Nolan, C., Margoliash, E.: Ann. Rev. Biochem. **37**, 727 (1968).
19. Romer, A. S.: The vertebrate body, 3rd ed. Philadelphia: Saunders 1962.
20. — The vertebrate story. Chicago: Univ. Chicago Press 1962.
21. Szalay, F. S.: Evolution **22**, 19 (1968).
22. Young, J. Z.: The life of vertebrates, 2nd ed. Oxford: Univ. Press 1962.
23. Zuckerkandl, E.: Sci. Amer. **212**, 110 (1965).
24. — Pauling, L.: In: Evolving genes and proteins. (V. Bryson and H. J. Vogel, eds.) New York: Academic Press 1965.

R. E. Dickerson
Division of Chemistry and Chemical Engineering
California Institute of Technology
Pasadena, Calif. 91109
U.S.A.