LES CAHIERS DE L'INSTITUT DE LA VIE

1970 No 24

LES
CAHIERS
DE L'INSTITUT
DE LA VIE

Siège : 89, Bd Saint-Michel, Paris-V^e^
Téléphone : 033-94-86
Périodicité : trimestrielle

Prix du numéro :
France 5 F, Etranger 6 F
Abonnement :
France 18 F, Etranger 22 F
Conditions spéciales aux membres de l'Institut de la Vie
Renseignements au Siège

SOMMAIRE

CONFÉRENCE INTERNATIONALE
DE LA PHYSIQUE THÉORIQUE A LA BIOLOGIE
Versailles, 30 juin - 5 juillet 1969

Journée du 3 juillet 1969, 1re séance

Mutations et processus de l'évolution
Président : D. GLASER (Berkeley)

Journée du 3 juillet 1969, 2e séance

Alkali ion carriers : dynamical behavior.
Président : D. GLASER (Berkeley)

Journée du 3 juillet 1969

Première séance

MUTATION ET PROCESSUS DE L'ÉVOLUTION

PRÉSIDENT D. GLASER

De la Physique théorique à la Biologie, C.N.R.S., 1971.

INTRODUCTION

D. GLASER

My name is Glaser and I have been asked to be chairman this morning because Professor Dulbecco has not been able to come to the meeting. Also Professor Bresler of Leningrad will not be here, but we are fortunate in having a contribution from Professor Eigen, who will speak to us after the interval. All the talks before this interval will be devoted to the main topics of this mornings meeting.

For the sake of the physicists it might be useful to make a few general remarks about developments in molecular biology that have made possible a really new confrontation of the classical theory of evolution. You will learn from the 3 talks that we will hear first this morning, that it is beginning to be possible to make a really quantitative examination of the theory of evolution, because it is possible to define evolutionary events at the molecular level in a way that makes one of them strictly comparable with another.

This provides a sound basis for building a quantitative theory in which the rate of evolution would be predicted by the rate of mutation together with specification of selection pressures and such properties of the population as migration and mating patterns.

I could make one remark that the number 10^{11} agrees roughly with the results of measurements of phenotypic mutation rates in bacteria which are in the neighbourhood of 10^{-8}. That's the probability of finding a phenotypic mutant which is an auxotroph, or has acquired drug resistance, per generation per bacterium. But only a small portion of all the base changes will be seen phenotypically. That's based on the fact that a typical cistron has about 1000 nucleotides in it, let's say. Now if you say that the phenotypically detectable mutations constitute only 1 % all the base changes there's a discrepancy of a factor 100 between the estimate you quoted and the final one. Another remark is that I think the mutation rate probably is not dominated by thermal effects, but more likely by errors in the action of polymerases. These inaccuracies are the result of selection for the structure of the polymerase and one can make a qualitative argument that the mutation rate is optimized, and that it may not be the object of evolution to produce polymerases which are the most accurate possible within the limitations of quantum machanics and of kT, but rather to pick one which is a compromise between accurency and a mutation rate which allows evolution. I don't know how to estimate what the accuracy limit of

a polymerase could be. The theoretical chemists have to do that for us some day.

A difficulty arises when one can't define the importance or calculate the probability of a particular step in evolution. Clearly the development of an eye is a much larger event than the change of skin pigment, for example.

The assignment of a quantitative measure to the size of an evolutionary step in gross biology is very difficult. When one can speak of a single base change at the nucleic acid level and can make the chemical statement that a large number of base changes are equally likely, perhaps all base changes under some conditions, then a single base change can be taken as a unit of evolutionary change and the number of such changes per century can be taken as an input to a quantitative theory. The papers that we will hear this morning will contain descriptions of measurements of rates of evolution defined at the molecular level together with explorations of mechanisms which can account for these evolutionary steps at the DNA level. These basic events will be correlated with phenotypic results of evolutionary events at the protein level and to some extent at the organismic level.

With that brief introduction to the physicists describing the significance of these developments in molecular kinetics and their application to evolution, I'd like to call on the first speaker, Professor Yanofsky of Stanford, who will speak on the " Protein structure and evolution ".

De la Physique théorique à la Biologie, C.N.R.S., 1971.

PROTEIN STRUCTURE AND EVOLUTION

CHARLES YANOFSKY

Department of Biological Sciences, Stanford University, Stanford, California 94305

Present day genetic and biochemical techniques provide the means by which we can attempt to answer fundamental questions on the molecular evolution of functional proteins. The considerable knowledge acquired in recent years in studies of gene structure-protein structure relationships serves as the basis for the design of experiments which may reveal why a protein in a particular organism has a unique primary structure, and how that structure changes when the organism is subjected to the forces of evolution. In this article I would like to describe mutational studies we have performed which provide some insight into structure-function relationships in a specific protein. I will also discuss experiments which are directed towards achieving the ' evolution ' of a functional protein.

The tryptophan operon of *E. coli*

The gene cluster we have studied in our analyses of gene structure-protein structure relationships is the tryptophan operon of *Escherichia coli*. This operon

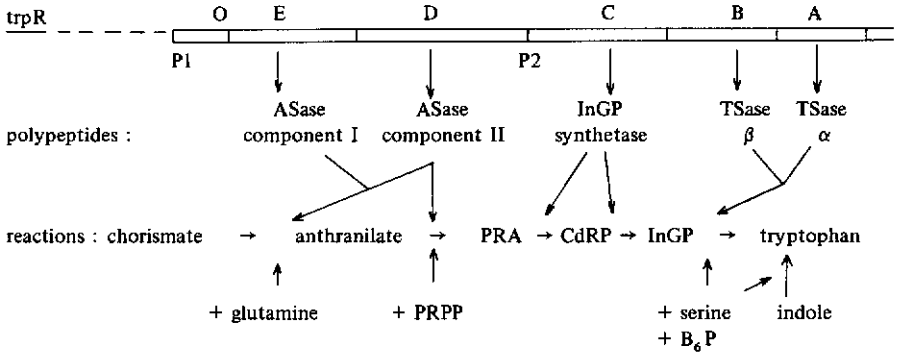

Fig. 1. The tryptophan operon of *E. coli*. The operon consists of 5 structural genes and adjacent controlling elements. TrpR is an unlinked gene specifying a protein repressor of the operon. O is the operator region and P1 and P2 are promoter regions. The various reactions in the pathway and the enzymes and enzyme complexes that serve as catalysts are indicated.

consists of five structural genes, each specifying a polypeptide which by itself or as a component of an enzyme complex catalyzes one or more of the terminal reactions in the biosynthesis of tryptophan (Fig. 1). Extensive mutational studies performed with this operon suggest that no segment of it is concerned with other essential bacterial functions. One implication of the existence of gene clusters of this type is that the component genes were derived from a common ancestral gene. To establish this point is of course one major objective of modern biology.

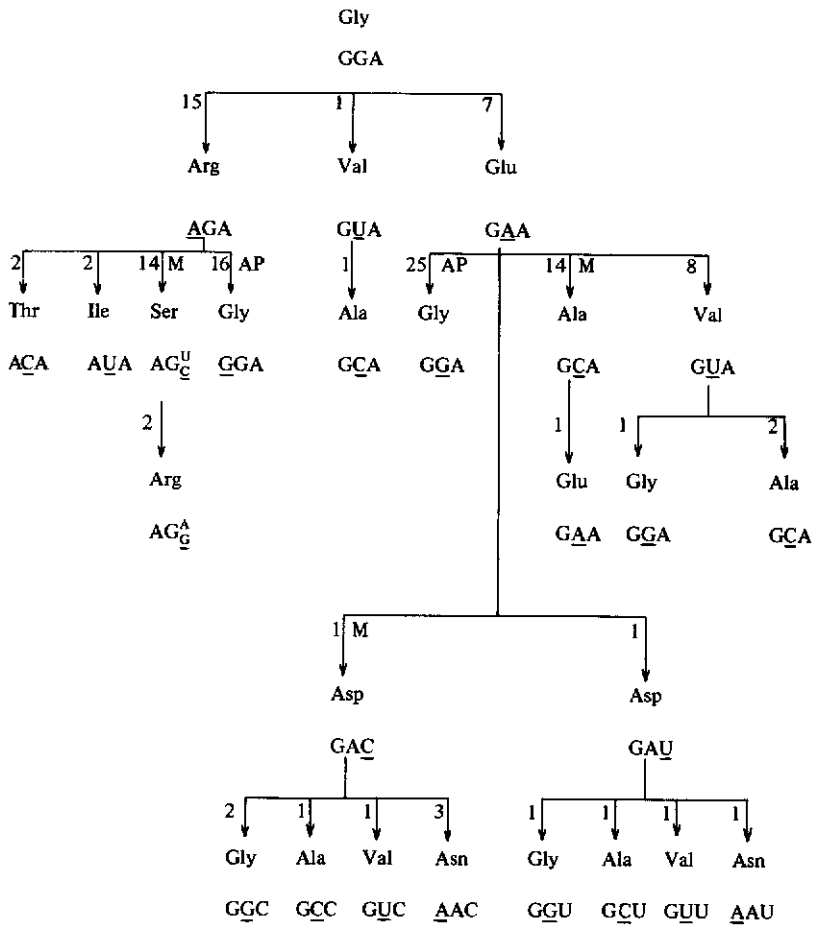

Fig. 2. Amino acid changes that have been observed at position 210 in the A protein and the probable corresponding codons [2, 3]. The number alongside each arrow indicates how many times the change was demonstrated by protein structure analyses. AP and M indicate changes favored by 2-aminopurine and the Treffers mutator gene, respectively. A bar under a codon letter identifies the nucleotide that is presumed to be introduced by the mutational change.

Tryptophan synthetase A protein alterations

Most of our gene structure-protein structure studies have been performed with the tryptophan synthetase A gene and A protein. The A protein is a single polypeptide chain 267 amino acid residues in length [1]; it has been shown to correspond linearly with the genetic map of the A gene [1]. Mutational changes in the A gene often lead to the production of altered A proteins which have single amino acid differences from the wild-type protein. At several positions in the A protein multiple amino acid changes have been detected. At position 210, for example, ten different amino acids have been inserted [2] (Fig. 2). Each of the observed amino acid substitutions is consistent with the interpretation that single mutational events involve single base-pair changes [2, 3]. Multiple amino acid substitutions have also been observed at positions 182 and 233 [3] (Fig. 3). It is clear from these cases and from comparable ones with other gene-protein systems that different amino acids can occupy a given position in a protein and permit function. We also know from amino acid sequence comparisons that enzymes isolated from related or unrelated species may have many sequence differences and nevertheless exhibit comparable enzymic activity. These observations focus on an important question: To what extent is the amino acid residue at each position in a protein essential for maximum effectiveness of that protein in its respective organism? The same question phrased in terms familiar to the evolutionary biologists is: Are neutral mutational changes preserved during evolution? In order to attack this problem experimentally we sought some means of rigorously assessing the

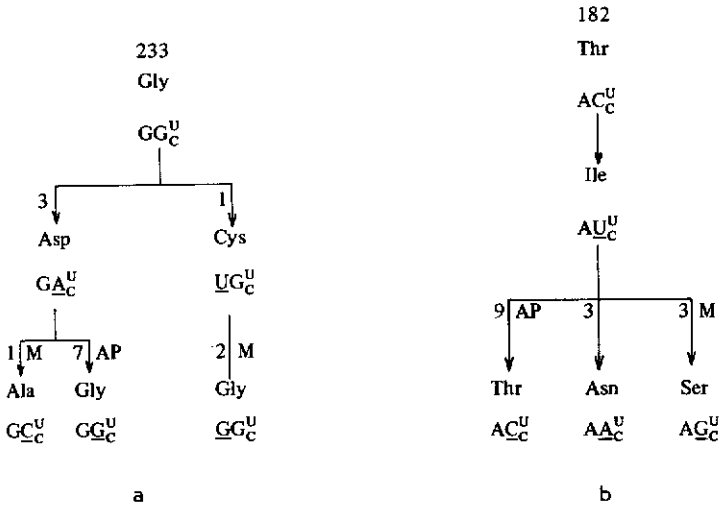

Fig. 3. Amino acid changes at position 182 (Fig. 3 A) and 233 (Fig. 3 B) and the probable corresponding codon changes [3]. See legend to Fig. 2 for other explanatory information.

functional capability of active A proteins with different amino acids at the same position. Many of the properties of the isolated altered A proteins have been examined but, since it is *in vivo* activity which is of concern to the organism, only tests performed in the growing cell could be considered relevant.

Perhaps the most sensitive means of examining the catalytic capability of an altered A protein in an organism such as *E. coli* is to determine the (tryptophan operon) enzyme levels attained when a culture is growing logarithmically in a minimal synthetic medium lacking tryptophan. Since under these conditions the organism must synthesize tryptophan to sustain its growth, a decrease in the catalytic capability of an enzyme should result in the production of elevated levels of all the enzymes specified by the operon; i.e., *E. coli* normally responds to a tryptophan deficiency by derepressing or turning on the synthesis of the enzymes of the pathway. Thus a very sensitive measure of true *in vivo* synthetic capacity is the effect on the levels of the biosynthetic enzymes. As can be seen in Table 1, when experiments of this type are per-

TABLE 1

Tryptophan synthetase B protein specific activities and generation times of strains with different amino acids at position 210 in the A protein.

Amino acid at position 210	Tryptophan synthetase B protein specific activity	Generation time (min)
Gly (wild type)	3.0 3.2 3.3	60, 60
Ala	3.2 3.3 3.5	60, 60
Ser	3.7 3.9 3.9	59, 60
Thr	7.7 8.0 8.2	58, 59
Val	30.1 31.3 31.9	69, 77
Ile	27.9 24.8 27.7	74, 74
Asn	50.9 51.4 55.1	104, 106

To ensure a constant genetic background each A gene was introduced by transduction into the same strain, a mutant with the A gene deleted. Several non-lysogenic colonies from each transduction were isolated and purified. Specific activities were determined with independent cultures harvested during log phase growth on minimal medium. Generation times were determined at 37 °C with cultures growing in minimal medium with glucose as carbon source. Estimates are based on cell population increases from 2×10^8 to 8×10^8/ml.

formed some of the proteins appear to be as active as the wild-type protein while others are probably less efficient and therefore signal the production of increased amounts of the biosynthetic enzymes. The same relationship is evident from the data presented in Table 2 for strains with altered proteins with amino acid changes at other positions in the A protein. Thus it is clear from this test that some amino acids are equally as effective as the wild-type amino acid. However, other amino acids at the same protein positions limit *in vivo* enzyme activity.

TABLE 2

Tryptophan synthetase B protein specific activities and generation times of strains with different amino acids at positions 182 and 233 in the A protein.

Amino acid at	Tryptophan synthetase B protein specific activity	Generation time (min)
position 182		
Thr (wild type)	2.6	58, 59
Ser	2.2	60, 60
Asn	9.1 9.3 10.8	58, 61
position 233		
Gly (wild type)	3.0 3.2 3.3	60, 60
Ala	11.3 12.9 13.1	61, 61

See legend to Table 1 for experimental conditions.

In view of these findings we might ask a related question : When an organism produces elevated enzyme levels does it do so at the expense of its ability to perform other metabolic reactions ? For example, in the wild-type strain growing in minimal medium the tryptophan biosynthetic enzymes constitute 0.4 % of the soluble protein. If the organism were forced to increase this level to ca. 4 % to provide suffiicent tryptophan for maximal growth rates would it do so at the expense of its ability to perform other metabolic reactions, thereby limiting its growth rate ? It is evident from the data in Tables 1 and 2 that significant increases in enzyme levels can be tolerated without any noticeable effect on the generation time. Thus, as can be seen in Table 1, when either serine or threonine occupies position 210 in the A protein the generation time is unaffected (Table 1). However, when the enzyme levels are in-

creased 10-fold (valine and isoleucine proteins) a significant lengthening of the generation time is evident. When still higher levels of enzyme are produced (asparagine protein) even longer generation times are observed. On the basis of the latter finding it seems likely that the enzyme levels and generation times in the valine and isoleucine strains represent the consequence of a *balance* between the rate of tryptophan synthesis and the effect of the formation of large amounts of these proteins on the growth rate of the organism. In Table 2 we also see that significant increases in specific activity are not correlated with appreciable changes in generation time. We might have expected to see such increases when enzyme levels are increased 3- to 4-fold. It should be pointed out, however, that studies performed in the manner described in Tables 1 and 2 are incapable of detecting minor changes in generation time. Despite this, we may tentatively conclude that different amino acids are equally acceptable at certain positions in the A protein and that moderate increases in enzyme levels can be tolerated without exerting a noticeable effect on the growth rate. Thus neutral or near-neutral mutational changes probably can occur-whether they are preserved is a much more difficult question to answer.

Compensating amino acid changes

In many missense mutants reversion events occur at second sites within the A gene as well as in the codon affected by the primary mutation [4, 5]. Several cases of second-site reversion have been analyzed and the findings obtained have revealed structural relationships within the folded protein molecule. For example, the change from glycine to glutamic acid at position 210 in the A protein is reversed by a change from tyrosine to cysteine at position 174 (Fig. 4). Interestingly, only the latter change reverses the effect of the presence of glutamic acid at position 210, i.e., mutational changes in other A gene codons cannot restore functional activity and only the change from tyrosine to cysteine at position 174 is effective. Similarly, mutant A187, an auxotroph with two amino acid differences from the wild-type protein, valines instead of glycines at positions 210 and 212, reverts at three positions,

Strain	Amino acids at corresponding positions		Activity of protein	Locations of genetic changes
	174-175-176	210-211-212		
wild type	-Tyr-Leu-Leu-	33 residues -Gly-Phe-Gly	active	
A46	-Tyr-Leu-Leu-	-Glu-Phe-Gly	inactive	
A46PR8	-Cys-Leu-Leu-	-Glu-Phe-Gly	active	

Fig. 4. Second-site reversion of mutant A46 [4]. As indicated, a Tyr → Cys change at position 174 activates the protein with Glu at position 210.

Strain	Amino acids at corresponding positions		Activity of protein	Locations of genetic changes
	174-175-176	210-211-212		
wild type	Tyr-Leu-Leu- 33 residues	-Gly-Phe-Gly	active	
A46	Tyr-Leu-Leu-	-Glu-Phe-Gly	inactive	
A46PR9	Tyr-Leu-Leu-	-Val-Phe-Val	active	
A187	Tyr-Leu-Leu-	-Val-Phe-Val	inactive	
A187SPR4	Tyr-Leu-Leu-	-Val-Phe-Gly	active	
A187SPR3	Tyr-Leu-Leu-	-Val-Phe-Ala	active	
A187SPR5	Tyr-Leu-Leu-	-Gly-Phe-Val	active	
A187SPR2	Tyr-Leu-Leu-	-Ala-Phe-Val	active	
A187SPR1	Tyr-Leu-Arg-	-Val-Phe-Val	active	

Fig. 5. Second-site reversion of mutant A 187 [5]. The A 187 protein has two changes; the Gly residues at positions 210 and 212 are replaced by Val residues. When either Val is replaced by Gly or Ala, the protein is functional. Both valines are retained in a functional protein in which the Leu residue at position 176 is replaced by Arg.

210, 212 and 176. At positions 210 and 212 the replacement of valine by either glycine or alanine restores activity, demonstrating that the A 187 protein is inactive only because both valines are present. One further point of interest is that the position of the distal reversion change, at 176, is two residues from the position of the second-site reversion in mutant A 46. These observations suggest that the two regions of the polypeptide chain indicated in Figs. 4 and 5 interact in the native molecule. We may conclude from these studies that because of the spatial relationships in the folded molecule the effects of an amino acid change in one region of the molecule can only be overcome by distal changes by specific alterations in an interacting region.

These observations raise the possibility that a neutral mutational change at one site may permit a subsequent change to confer a selective advantage. This gain in functional acceptability would then preserve what originally was a neutral event.

Attemps to "evolve" a functional A protein in strains lacking a segment of the A gene

I would like to known whether it is possible to produce a functional A protein by mutationally altering a protein fragment lacking the 20 or so amino acid residues at the carboxyl end of the molecule. To determine this, deletion mutants lacking the end of the A gene were subjected to mutagenic treatments and the treated populations were added to a medium which would only sustain the growth of cells with a functional A protein. The deletion

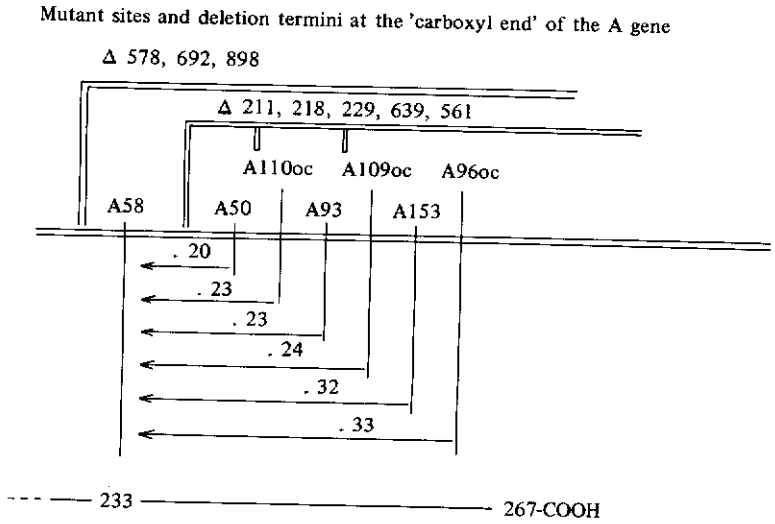

Fig. 6. Mutant sites and deletion termini in the region of the A gene specifying the carboxyl end of the A protein. Map distances are indicated above the arrows. Three of the point mutants (A 110, A 109, A 96) are ochre nonsense mutants. The precise terminus of each deletion is not known but it ends in the region indicated.

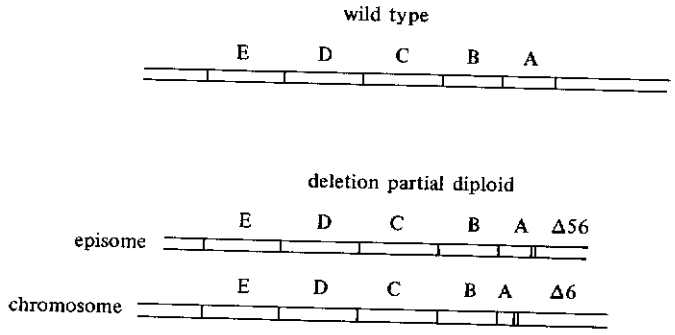

Fig. 7. A diploid strain employed in a prototrophy selection experiment. Deletion 6 removes most of the A gene while deletion 56 removes that region specifying the last thirty or so residues.

mutants examined are listed in Figs. 6 and 7. They were heavily mutagenized with nitrosoguanidine and ICR 191, powerful mutagens which produce base substitutions and base deletions and additions, respectively. Selective conditions were employed which would have permitted the outgrowth of bacteria with a functional A protein. In none of the deletions listed was a functional A protein detected — this despite the fact that large populations of bacteria were employed and the bacteria were permitted to divide several times before the

tryptophan supplement in the growth medium was depleted. These negative results suggest that the carboxyl-terminal end of the A protein is essential for enzyme activity and that alterations in the remaining portion of the molecule cannot compensate for the loss. This conclusion is supported by the fact that nonsense codons in the terminal region of the A gene result in enzyme inactivity (see Fig. 6). Diploid strains with terminal deletions were also examined in these studies to eliminate the possibility that inactivating mutations occurred in other genes of the operon concomitantly with mutations in the A gene. Furthermore, episomes with A gene terminal deletions were transferred out of a heavily mutagenized population into haploid cells lacking only the A gene. In every case except one, to be described below, we did not detect a functional A protein. Despite these negative results other findings to be described subsequently suggest that the amino acid sequence at the carboxyl terminus of the wild-type A protein is not the only sequence that will permit this protein to be catalytically active. In the diploid strain described in Fig. 7, in which A gene deletions were present on both chromosome and episome, an active A protein was formed as a consequence of mutational changes in the A gene segment. The prototrophic strain obtained grows very poorly without tryptophan, however, suggesting that the functional A protein that is produced is at best inefficient. To eliminate the possibility that in the diploid strain one of the other genes of the tryptophan operon was assuming the function of the A gene, the *trp* operon of the episome was introduced by transduction into a haploid strain, replacing the operon of the recipient; i.e., the transductants had only one copy of each of the genes of the operon. These transductants were slow-growing prototrophs, suggesting that the mutation or mutations responsible for A protein activity were in or near the A gene.

In related studies we attempted to modify mutationally the E, D, C or B gene in a diploid strain so that the altered protein it produced could function as an A protein. The partial diploid prepared for the experiment had 90 % of the A gene deleted on both chromosome and episome. To date, these experiments have also given negative results, suggesting that each of the operon

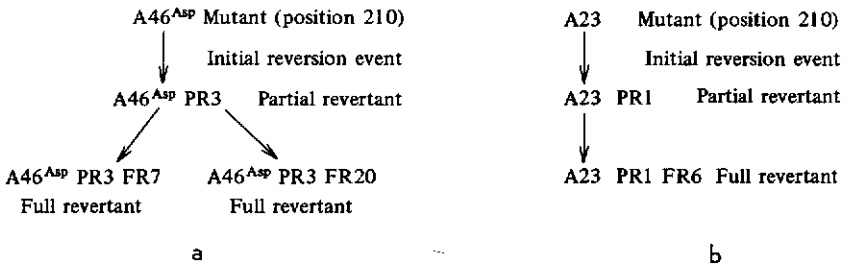

Fig. 8. Sequential reversion events [2] starting with mutant A 46^{Asp} (aspartic acid at position 210), A, and mutant A 23 (arginine at position 210), B.

proteins is considerably different from its ancestral protein and thus many amino acid changes would be required for it to acquire A protein activity.

Frameshift mutations and protein function

Mutant A46^{Asp} and A23 both yield slow-growing partial revertant strains in which prototrophy is due to second-site mutations [2]. In order to analyze the effect of these mutations on the structure of the A protein, faster-growing full revertants were selected from the partial revertant strains (Fig. 8). This extra step was necessary because the partial revertant A proteins were extremely labile and could not be isolated. When the full revertant A proteins were analyzed [2] we were surprised to find that several contiguous amino acids had changed in each (Fig. 9). The amino acid differences in each revertant could be explained by assuming that the primary mutational event resulted in a single base addition and the second mutation involved a single base deletion. The greater activity of the full revertant A proteins compared to the partial revertant proteins is readily understandable since, with the exception of the residues in the vicinity of position 210, the amino acid sequences would be unaltered. The activity of the partial revertant A proteins is surprising, particularly in view of the conclusions reached in the previous section. We would expect that in these strains (A46^{Asp} PR3 and A23 PR1) the entire sequence of the terminal portion of the A protein would be altered as a result of the single

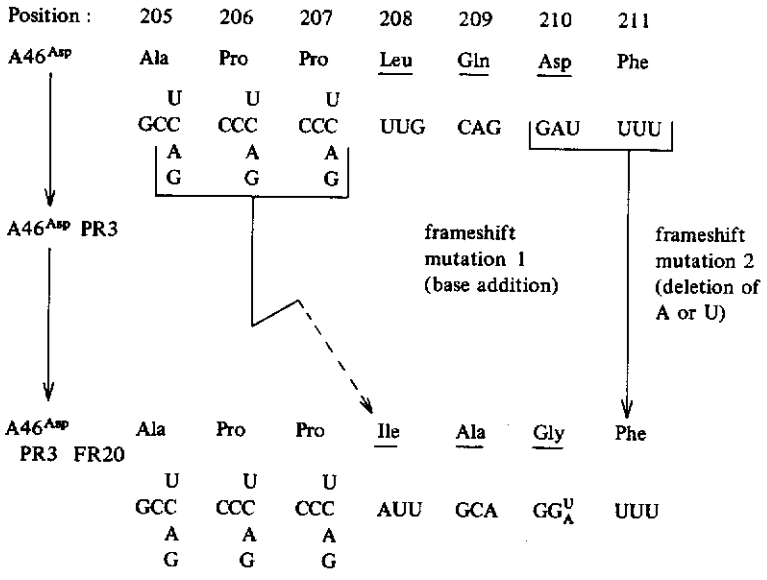

Fig. 9 a

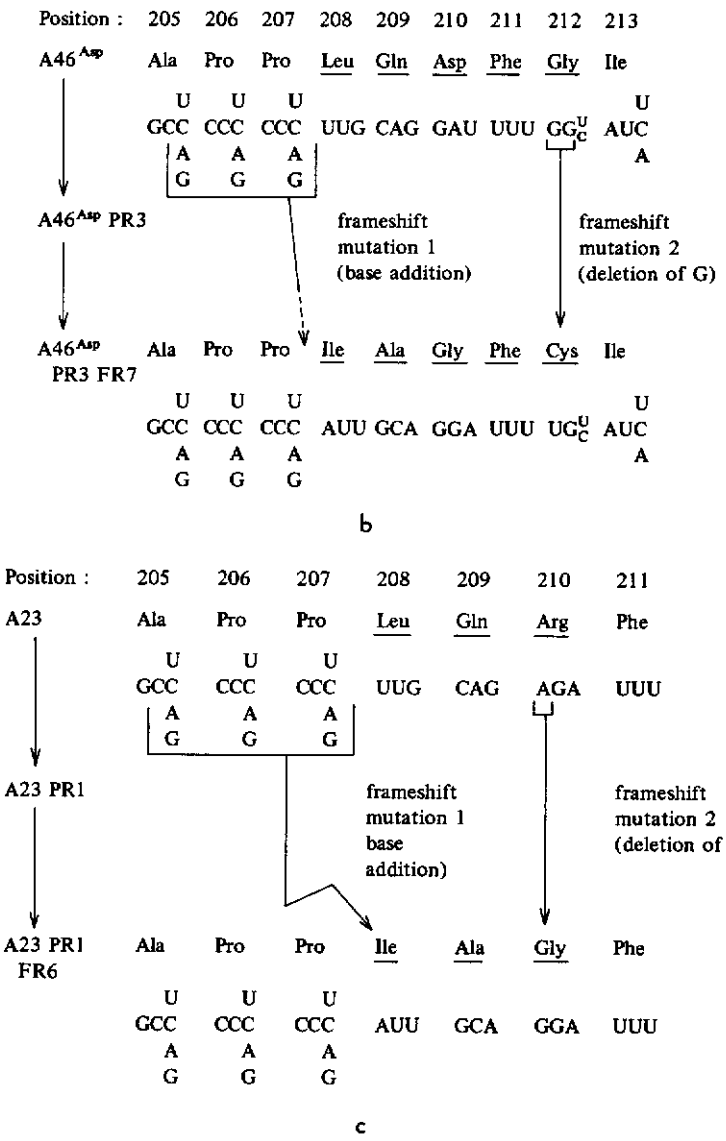

Fig. 9. Amino acid changes and probable corresponding nucleotide sequences in frameshift revertants A 46^{Asp} PR3 FR20 (A), A 46^{Asp} PR3 FR7 (B) and A 23 PR1 FR6 (C) [2].

base addition. The new sequences that would be generated in the vicinity of position 210 are shown in Fig. 10. In strain A46^{Asp} PR3 a glycine residue would presumably replace the aspartic acid residue which is present at position 210 in A46^{Asp} and is responsible for enzyme *inactivity*. In A23 PR1, however, the arginine residue at position 210 would be replaced by a charged

	207 –	208 –	209 –	210 –	211 –	212
wild type sequence	Pro –	Leu –	Gln –	Gly –	Phe –	Gly
A46^{Asp}	Pro –	Leu –	Gln –	Asp –	Phe –	Gly
	CCA	UUG	CAG	GAU	UUU	GGU
	+ 1					
A46^{Asp} PR3	CC·	AUU	GCA	GGA	UUU	UGG
	Pro –	Ile –	Ala –	Gly –	Phe –	Trp
A23	Pro –	Leu –	Gln –	Arg –	Phe –	Gly
	CCA	UUG	CAG	AGA	UUU	GGU
	+ 1					
A23 PR1	CC·	AUU	GCA	GAG	AUU	UGC
	Pro –	Ile –	Ala –	Glu –	Ile –	Trp

Fig. 10. Hypothetical amino acid and nucleotide sequences in partial revertants A46^{Asp} PR3 and A23 PR1. It is assumed that the reading frame would remain shifted and the polypeptide would terminate at the first " new phase " terminator.

amino acid, glutamic acid. Thus it is not obvious from these hypothetical sequences why the partial revertant proteins are active, especially since we would expect that the entire carboxyl sequence starting at position 208 would be different from that of the wild-type protein. We do not know the length of the partial revertant proteins; as many as 14 terminator codons could be introduced as a result of the frameshift in these strains [2]. Since this number is large it seems likely that the protein is less than 267 residues in length in the partial revertant strains. The most reasonable explanation for these findings is that the new sequence that is generated as a consequence of the frameshift can perform the function peculiar to the carboxyl end of the normal protein.

If we compare the hypothetical wild-type amino acid sequence which would result from the initial base addition with the presumed sequences in PR3 and PR1 (Fig. 11), we see that the sequences are quite similar — in fact, the only difference between the PR3 and wild-type sequences is at posi-

	207 –	208 –	209 –	210 –	211 –	212
wild type	Pro –	Leu –	Gln –	Gly –	Phe –	Gly
	CCA	UUG	CAG	GGA	UUU	GGU
	+ 1					
	CC·	AUU	GCA	GGG	AUU	UGG
hypothetical sequence	Pro –	Ile –	Ala –	Gly –	Ile –	Trp
A46^{Asp} PR3 sequence	Pro –	Ile –	Ala –	Gly –	Phe –	Trp
A23 PR1	Pro –	Ile –	Ala –	Glu –	Ile –	Trp

Fig. 11. Hypothetical amino acid and nucleotide sequences if the frameshift occurred in the wild-type strain.

tion 211. Here, different hydrophobic amino acids are present — isoleucine and phenylalanine. In view of these comparisons it is perhaps surprising that functional A proteins were not detected in the mutagenesis studies with the deletion mutants. This may indicate that in $A46^{Asp}$ PR3 and A23 PR1 the polypeptide chain has a near-normal length.

Comparative studies with the A proteins of *Escherichia coli*, *Salmonella typhimurium* and *Aerobacter aerogenes*

We are presently determining the amino acid sequences of the tryptophan synthetase A proteins from *Salmonella typhimurium* and *Aerobacter aerogenes* so that they may be compared with the sequence from *E. coli*. Our principal reason for performing these studies is based on the different GC contents of the DNA's of these organisms. *E. coli* DNA has approximately 50 % GC base pairs, *S. typhimurium* DNA about 51 % GC base pairs, while *A. aerogenes* DNA contains approximately 56-57 % GC base pairs [6]. At the present time about two-thirds of the *S. typhimurium* sequence and one-half of the *A. aerogenes* sequence are known. We estimate that there are ca. 10-15 % amino acid differences when we compare either sequence with that of *E. coli* [7]. The differences seem to be randomly distributed throughout the proteins and the *Salmonella* differences and *Aerobacter* differences are not at identical positions [8]. Thus it is not possible on the basis of these data to establish the evolutionary order of the three bacterial species. Most of the amino acid differences can be explained by a single base change per codon; thus the present structures probably differ from an ancestral molecule by no more than a single base change per codon. When we deduce the probable base change responsible for each amino acid change we find that the evolution of an *E. coli*-type protein to a *S. typhimurium*-type protein involved seven A or T → G or C changes and eight G or C → A or T changes. Of the seven identified AT changes, three were AT → GC and four were AT → CG. If we consider the differences proceeding from an *E. coli* protein to an *A. aerogenes* protein, there are nine A or T → G or C changes and only five G or C → A or T changes. This distribution is, of course, consistent with the higher GC content of *A. aerogenes* DNA. Furthermore, of the nine AT changes, two involved AT → GC and seven AT → CG. It appears, therefore, that the higher GC content of *A. aerogenes* DNA may be due to a preferential increase in the proportion of mutations from AT → CG. This conclusion, if substantiated by further studies, would be particularly interesting in view of our findings with the mutator gene of *E. coli* discovered by Treffers [9]. This mutator gene preferentially increases the base-pair change AT → CG [10, 11]. The presence of such a mutator gene at some period in the evolution of *A. aerogenes* would explain its high GC content and the apparent increase in the proportion of

AT → CG changes. In this connection, some recent work of Drake [12] should be cited. Drake has shown that certain mutations in the gene of phage T4 which specifies its DNA polymerase result in a reduction of the spontaneous mutation frequency. This finding suggests that at least in this organism DNA polymerase mistakes are responsible for a significant portion of spontaneous mutations. A bacterium with a highly active mutator gene may be being subjected to an exaggeration of a specific mistake-making mechanism.

The randomness of the positions of amino acid differences in the various A proteins examined and the fact that most of the amino acid differences can be explained by a single base change per codon, in my opinion support the conclusion that neutral mutations do occur and are preserved during evolution. More extensive and convincing data on this point have been discussed by King and Jukes [13] and by Margoliash (see article in this volume).

Acknowledgment

The author is indebted to Miriam Bonner, Virginia Horn and Susan Stasiowski for their excellent assistance with the studies described in this paper. These investigations were supported by grants from the National Science Foundation and the United States Public Health Service.

References

[1] C. Yanofsky, G.R. Drapeau, J.R. Guest and B.C. Carlton, The complete amino acid sequence of the tryptophan synthetase A protein (α subunit) and its colinear relationship with the genetic map of the A gene. *Proc. Natl. Acad. Sci. U.S.*, **57**, 296-298 (1967).

[2] H. Berger, W.J. Brammar and C. Yanofsky, Analysis of amino acid repacements resulting from frameshift and missense mutations in the tryptophan synthetase A gene of *Escherichia coli*. *J. Mol. iBol.*, **34**, 219-238 (1968).

[3] C. Yanofsky, H. Berger and W.J. Brammar, *In vivo* studies on the genetic code. *Proceedings of the XII International Congress of Genetics*, **3**, 155-165 (1969).

[4] D.R. Helinski and C. Yanofsky, A genetic and biochemical analysis of second site reversion. *J. Biol. Chem.*, **238**, 1043-1048 (1963).

[5] C. Yanofsky, V. Horn and D. Thorpe, Protein structure relationships revealed by mutational analysis. *Science*, **146**, 1593-1594 (1964).

[6] N. Sueoka. Compositional correlation between deoxyribonucleic acid and protein. *Cold Spring Harbor Symposium on Quantitative Biology*, **26**, 35-43 (1961).

[7] T.E. Creighton, G. Johnson and C. Yanofsky, unpublished observations.

[8] T.E. Creighton, D.R. Helinski, R.L. Somerville and C. Yanofsky, Comparison of the tryptophan synthetase α subunits of several species of Enterobacteriaceae *J. Bact.*, **91**, 1819-1826 (1966).

[9] H.P. Treffers, V. Spinelli and N.O. Belser, A factor (or mutator gene) influencing mutation rates in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.*, **40,** 1064-1071 (1954).

[10] C. Yanofsky, E.C. Cox and V. Horn, The unusual mutagenic specificity of an *E. coli* mutator gene. *Proc. Natl. Acad. Sci. U.S.*, **55,** 274-281 (1966).

[11] E.C. Cox and C. Yanofsky, Altered base ratios in the DNA of an *Escherichia coli* mutator strain. *Proc. Natl. Acad. Sci. U.U.*, **58,** 1895-1902 (1967).

[12] J.W. Drake, E.F. Allen, S.A. Forsberg, R. Preparata and E.O. Greening, Genetic control of mutation rates in Bacteriophage T4. *Nature*, **221,** 1128-1132 (1969).

[13] J.L. King and T.H. Jukes, Non-Darwinian evolution. *Science*, **164,** 788-798 (1969).

De la Physique théorique à la Biologie, C.N.R.S., 1971.

DISCUSSIONS

Ch. YANOFSKY : If I had the time, the next example I was going to show was one in which the initial amino acid change was two residues from the change in the A protein that I did describe. In that strain, the compensating change also occured two residues from the compensating change in the first strain. These findings plus the additional fact that such secondary changes do not occur throughout the protein indicate that there are restricted sites at which one amino acid change can compensate for an initial inactivating change. I have no doubt that we are looking at properties of the folded protein molecule in these cases of compensating changes.

Dr. BENNETT : My question to Dr. Yanofsky relates to the interpretation of the data in which you show that the activity of a protein coded for in the operon which makes the precursor of tryptophan would show activity perhaps three times that of wild types, whereas the rates of growth and the rates of replication are normal. Under these circumstances, one can presume that the pool of precursor would be three times larger than in wild types. Do I understand the circumstance correctly ?

Ch. YANOFSKY : In fact, if you test for the accumulation of biosynthetic intermediates in the strains I have described you find that accumulation increases as the activity of the tryptophan synthetase A protein decreases.

S. BENNETT : If I understood your test situation correctly, you look for the generation time of your mutant in pure culture. If one makes a competitive situation, when your mutant is competing with a wild type or some other type which does not accumulate, what is the performance of your mutant under such circumstances ?

Ch. YANOFSKY : We have performed competition experiments with mixtures of bacterial populations with different amino acids at the same position in the A protein — e.g., bacteria with glycine at position 210 mixed with bacteria with alanine at this position. The two classes of bacteria in each mixture were distinguished by alternative forms of a non-selective genetic marker (milibiose utilizing or non-utilizing; cultures grown in glucose). The different A genes were introduced by transduction into the same genetic background (but *mel*$^+$ or *mel*$^-$) in order to ensure uniformity. Reciprocal mixtures were prepared (Gly, *mel*$^+$ and Ala, *mel*$^-$; Gly, *mel*$^-$ and Ala, *mel*$^+$) and the

proportion of the two types of bacteria determined after growth in a minimal-glucose medium. All bacteria were F^-, thereby eliminating gene transfer and recombination as a possible complication. Mixtures were also prepared in which one of the bacterial types employed had an amino acid at position 210 which was growth-limiting--e.g., Gly, *mel*$^+$ and Val *mel*$^-$. The results of these competition experiments were somewhat disappointing although in retrospect the findings could have been anticipated. With Gly-Ala mixtures neither type appeared to be selectively favored in short-term experiments, while after many generations either type began to outgrow the other. With Gly-Val mixtures, in short-term experiments, the proportion of Gly bacteria increased but after many generations, again, either type predominated in the mixed population. I interpret these findings as indicating that mutations in any one of a fairly large number of genes in *E. coli* can confer a greater selective advantage than is possible on the basis of the differences in tryptophan synthetase activity of the strains that were mixed. Thus in attempting to assess relative selective values we cannot consider one trait alone, if the differences between the competing strains are not great. On the other hand, these studies illustrate how a neutral mutational change could be fixed in a population. If a mutation conferring a selective advantage occurred in the individual with the neutral mutational change then, of course, the neutral change could be preserved.

J. POLONSKY : I would like to put a question. Is this mutation in the active site or not in the active site of the protein ?

Ch. YANOFSKY : I don't know.

J. MONOD : I suppose you have studied the kinetic parameters of some of these proteins. Do you know whether it is the specific activity which is changed or more often the K_m I would presume that depending on whether it is one of the two parameters which is modified, the selective value or disadvantage would not be the same.

Ch. YANOFSKY : In one of the revertants the K_m is altered; the affinity of the A protein for its substrate is reduced.

D. GLASER : I think we should go on to the next paper which is by E. Margoliash, who will talk on the evolutionary information content of protein amino acid sequences.

De la Physique théorique à la Biologie, C.N.R.S., 1971.

THE EVOLUTIONARY INFORMATION CONTENT OF PROTEIN AMINO ACID SEQUENCES *

E. MARGOLIASH and W.M. FITCH

Department of Molecular Biology, Abbott Laboratories, North Chicago, Illinois 60064 and Department of Physiological Chemistry, University of Wisconsin, Madison, Wisconsin 53706

It has long been obvious that because of the processes by which life perpetuates itself, living organisms are an excellent repository of the evidence of their own evolutionary history. Every material of which an organism is composed and every phase of its activities are results of that history. However, as pointed out by Zuckerhandl and Pauling [1], some biological substances retain the traces of the past in a relatively easily identifiable form, while for others the relation to evolution is much more difficult to discern. There is in this regard a very fundamental difference between so-called " informational macromolecules ", DNA, RNA and proteins, and the other substances found in living organisms. The former are simple images of each other in which the linear sequence of chemical building blocks carries the biological information, so that whether one determines the amino acid sequence of a protein chain or the structure of a transfer RNA molecule, one is merely examining the fine structure of a very small segment of the genome at its simpliest molecular level. This simplicity is the crucial advantage. All other biological substances which are elaborated by the organism represent a far more complex interplay of sources of biological information. For example, chemically simple micromolecular substances, such as flavinoids or any of the intermediates of metabolic energy cycles, are the products of whole assembly lines of enzymes, each derived from one or more structural genes, each of which is in turn likely to be controlled by one or more so-called regulatory genetic influences. Thus, though the number of genes controlling the synthesis of a micromolecular biological entity may be in the order of 100, far less than the possibly 10^6 genes which may affect a complex morphological character, such as the shape of the human nose, the genetic complexity of the micromolecule is more than enough to give it the same status as that of the ordinary morphological characters employed for classical evolu-

* Reproduced by permission from Miami Winter Symposia 1, pp. 33-51, North-Holland Publishing Co., Amsterdam, 1970.

tionary appraisals. The advantage of chemistry, represented by the understanding of the structure of substances at the molecular level, has been entirely lost.

This is not the case with the amino acid sequences of proteins, since the chemical structure is itself an expression of the structure of a gene. Thus, with information on the primary structure of a sufficient number of different proteins from a sufficient number of different and properly chosen species, it may eventually be possible, independently of any other knowledge, to read directly the record of the evolutionary history of these species encoded in the proteins they synthesize. An obvious attraction of the molecular taxonomy of proteins is the possibility of reconstituting today the temporal order of long past evolutionary changes in terms of unit mutational events. This could possibly lead to an estimate of the structures of informational macromolecules, proteins and nucleic acids, as they occurred further and further back to that shadowy point in biological history when chemical evolution ended and replicating biological systems took over. In this process one can expect to obtain a wealth of information concerning evolutionary mechanisms as they relate to protein structure and function. This short review attempts to summarize the present status of the endeavour.

The Significance of Amino Acid Sequence Similarities

Similarities between different proteins in the same species or between ostensibly similar proteins of different species are apparent by any of the large variety of techniques which define their structural and functional parameters. These extend from similarities in tissue and cellular localizations, similarities in physiological and physico-chemical modes of function, all the way to precise details of amino acid sequence and of three-dimensional spatial structure. Since the primary structures are the direct expression of the organism's store of biological information, this paper will concern itself solely with amino acid sequences. However, proteins were classified before their primary structures were known and the search for similarities is still to a large extent limited to groups defined by criteria other than primary structures. This will necessarily exclude descendants of the ancestral form which have varied to the extent of acquiring new functions and the physico-chemical attributes which fit the new functions. It is only when primary structures will have become available for a large proportion of all proteins that it will be possible to discuss relations of proteins which are no longer apparent in their functions. In the meantime, expected similarities in function of proteins that have undergone relatively small divergences, as in the case of the digestive proteolytic enzymes [2, 3], or quite unexpected similarities, as in the case of lysozyme and α-lactalbumin [4], have already provided vivid illustrations of the evolutionary shaping at the molecular level of new functions from old structures.

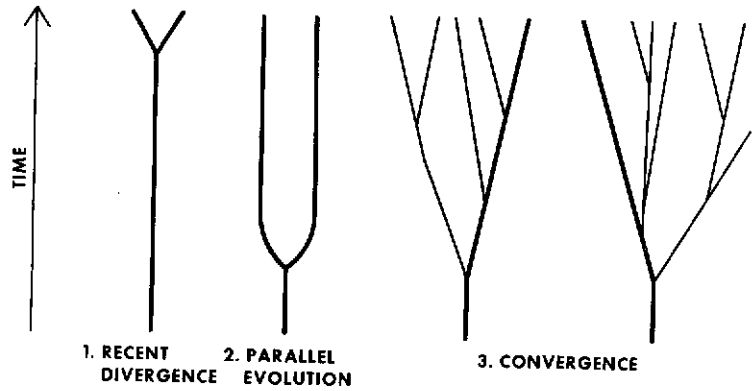

Fig. 1. Possible evolutionary reasons for similarity of polypeptide chains.

However, common ancestry is not the only possible basis for similarity in amino acid sequence. Indeed, two proteins may be similar at the time they are examined not only because they diverged from a common origin relatively recently in their evolutionary history or because having diverged a long time ago they have followed largely parallel pathways, but also because having arisen from different ancestral origins they have tended to evolve to similar or identical functions in different lines of evolutionary descent, and have therefore acquired the degree of similarity of structure required by this similarity of function. These possibilities are diagramed in Fig. 1. Thus, before one can conclude that a set of proteins of apparently similar amino acid sequence has a common evolutionary origin, i.e. are *homologous* in the ordinary biological usage of the term, one must answer two questions, as follows :

1) *Are the similarities of primary structure greater than could occur by chance ?*

A systematic approach to this question [5] requires in essence the ability to calculate the probability of random similarity. This can be done by comparing all possible pairs of segments of a fixed length (such as 20 or 30 residues long) between the two protein chains under consideraction. For example, in a comparison of two sequences 100 residues long, for segments of 20 residues there are (100-20 + 1) (100-20 + 1) or 6561 possible pairs. One can calculate the minimal number of single nucleotide changes required to transform the gene segment coding for one member of each pair into that coding for the other (" mutation " or " replacement distance "), and plot the total number of times each particular replacement distance occurs in all the comparisons as a function of the replacement distance. Such a plot is given in Fig. 2 for human and the iso-1-cytochrome *c* of bakers' yeast. The average replacement distance for any randomly chosen pair of amino acids is 1.5, so that for a pair of random 30-residue segments it would be 45. In Fig. 2, the random comparisons are given

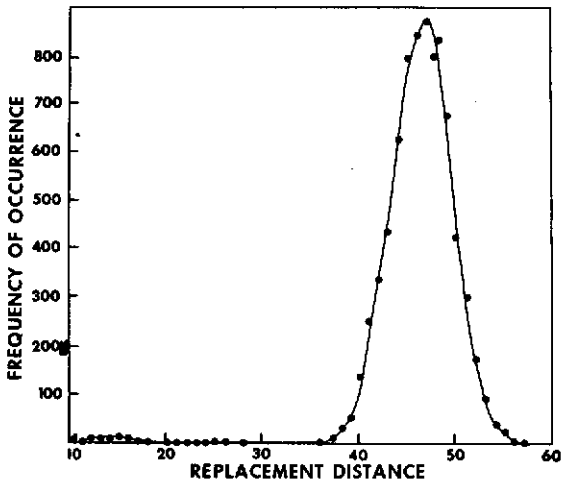

Fig. 2. Comparison of replacement distances for all possible 30 residue segments of human cytochrome *c* (6) and bakers' yeast iso-1 cytochrome *c* (7) by the procedure of Fitch [5]. The number of times various replacement distances occur in the comparisons are given on the ordinate (Frequency of occurrence).

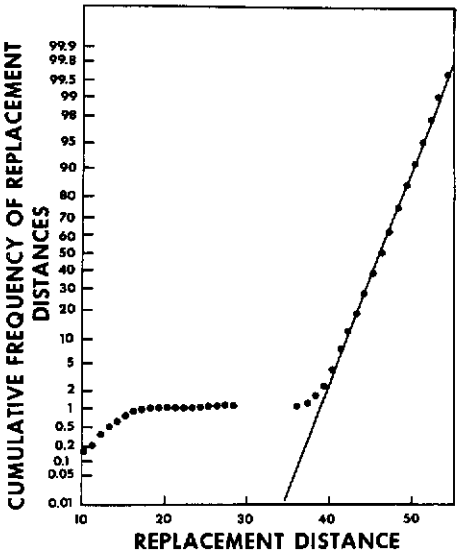

Fig. 3. Probability plot of the data from Figure 2. The random part of the distribution is given by the linear portion of the curve. The comparisons for which the replacement distances are smaller than expected for a random distribution are represented by the points which deviate from the straight line at the lower left. The probability that such a distribution occurred by chance is less than 10^{-80} [8].

by the Gaussian portion of the curve, very nearly centered, as expected, on a replacement distance of 45. The long tail of the curve to the left represents those comparisons for which the replacement distances are smaller than would be expected on a random basis, and which therefore indicate that the degree of similarity between human and yeast cytochromes *c* is greater than could be accounted for by chance.

Data of the type given in Fig. 2 can be recalculated to give cumulative distributions. When these are plotted on probability paper, the Gaussian portion of the curve becomes a straight line, and non-random comparisons are detected as deflections from the linear curve towards the lower left [5]. Such a probit plot is shown for the human-yeast cytochrome *c* comparison in Fig. 3. To supplement the graphic comparisons, particularly in cases in which the degree of non-randomness is not as obvious as in the example given in Figs. 2 and 3, an arbitrary statistic having the characteristic distribution of χ^2 when unrelated amino acid sequences are compared, can be employed. This permits one to determine the probability that a given departure from linearity would occur by chance [8]. Thus, for example, the data in Fig. 3 indicate that the probability that such a distribution would occur by chance is less than 10^{-80}.

The proper alignment of two amino acid sequences, for which a degree of similarity greater than random has been established, is to a large extent obtained merely by considering those pairs of segments which yielded the non-random portion of the distribution curve. Furthermore, a method has been devised to locate the gaps required to align two primary structures so as to minimize the total number of nucleotide replacements, deletions and insertions necessary to account for the differences between the sequences [9].

The search for significant similarities need not be limited to different proteins, but may also be usefully conducted with portions of a single polypeptide chain. If such are found, one may reasonably infer that partial internal duplications have occurred during the evolution of the corresponding structural genes. Such phenomena could result from unequal crossing over within one gene, as is considered to account for the remarkable similarity between the first and last 26 amino acids of bacterial ferredoxins [10-14], the two segments of the light chains and the four segments of the heavy chains of γ immunoglobulins [15-19]. Moreover, equal crossing over can take place between adjoining genes, a phenomenon which presumably accounts for the non-α chains of the abnormal human Lepore hemoglobins, hybrids of δ and β chains [20-24]. It also may occur between two alleles in a heterozygote, as must have been the case for the 2-α chain of human haptoglobin, the 142 residues of which are derived from the amino-terminal and carboxylterminal segments of the 83-residue 1Fα and 1Sα common haptoglobin allelic chains [25].

2) *Are significant similarities of primary structure due to common ancestry or to functional convergence ?*

Statistical answers to this question require the techniques employed in estimating evolutionary relations from amino acid sequence information (phylogenetic trees), and the assessment of the structures of ancestral forms of the protein under consideration (reconstructed ancestral sequences). Since both these topics are considered below, any discussion of the distinction between divergence from a common ancestral form and convergence from different phylogenetic origins is best postponed till after these procedures have been considered.

Statistical Phylogenetic Trees

If the amino acid sequences for a set of proteins have been shown to possess similarities greater than random, and one further assumes that this is due to evolutionary homology, one can then set out to attempt to determine the phylogenetic relations of the species carrying these proteins purely on the basis of their structures. However, it must not be overlooked that not all homologous relationships justify such a procedure. If, in the common ancestor of all the species considered, the protein was represented by a single gene, then the descendent genes can be called *orthologous* (from ortho, meaning exact) [26], and precisely reflect, in a one-to-one fashion, the lineage of the species. As long as the evolutionary variations of this protein represent a statistically valid sample of the overall evolutionary variations of species carrying it, then one may expect to extract proper phylogenetic information from the corresponding amino acid sequences. However, homologous genes may have undergone duplication and remained side by side in all or many of the species descending from the earliest ancestor in which the duplication occurred. These may be termed *paralogous* (from para, meaning in parallel) [26], and clearly cannot be utilized indiscriminately to ascertain phylogenetic relations. For example, in most vertebrates, homoglobins are tetrameric and have at least two types of chains, α and β. Moreover, there often are other types of non-α chains, such as the γ and δ human chains. Vertebrates also carry another protein of the same homologous series, the monomeric myoglobin. There is general agreement that the genes for all these proteins are homologous [27-29], but if one were to utilize for the construction of a vertebrate phylogeny the amino acid sequences of the α chains of some species, those of the β chains of others and those of the myoglobins of still others, the result would be an absurdity. Indeed, the species would be mainly segregated into 3 groups, one each for those species for which the α, β or myoglobin chains were used for the analysis. This is because the gene duplications which gave rise to the three varieties of chains had occurred before the evolutionary appearance of the common ancestor of the species examined, and

these genes had since evolved more or less independently. Each gene separately would be orthologous and the species variations of α, or β, or myoglobin chain structures could in principle provide data for three independent assessments of vertebrate evolutionary relations. (In the phylogenetic tree for eukaryotic cytochromes *c* shown in Fig. 4, all the proteins are orthologous, except for the iso-1 and iso-2 cytochromes *c* of bakers' yeast which are paralogous. Since this is the only such relationship, it does not introduce any errors in the rest of the tree).

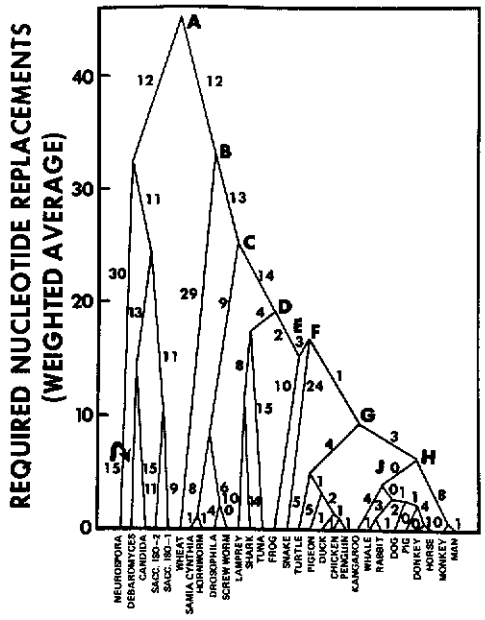

Fig. 4. Statistical phylogenetic tree based on the replacement distances between the cytochromes *c* of the species listed, as obtained by the procedure of Fitch and Margoliash [30]. Each number on the figure is the replacement distance along the line of descent as counted by the procedure of Fitch [36]. Each apex is placed at an ordinate value which is the weighted average of the sums of all nucleotide replacements in the lines of descent from that apex. References to the amino acid sequences of the cytochrome *c* are given in References 28 and 45.

Just as for the polypeptide segments utilized to establish the random or non-random nature of the similarities between two amino acid sequences (see above), it is possible to calculate the minimal replacement distances between any two orthologous amino acid sequences. For a set of n sequences, there are $n(n\text{-}1)/2$ such distances, which can be used to construct a phylogenetic tree, such as that shown in Fig. 4 for the eukaryotic cytochromes *c* of 29 species [26, 30, 31]. Initially, each protein is assigned to a separate subset. Those two

subsets which show the lowest replacement distance are joined and are henceforth treated as a single subset. The procedure is repeated until all proteins have been joined to provide the initial phylogenetic tree, which is merely a graphical representation of the order in which the subsets were joined. The replacement distances between various branch points of the tree can be calculated, and the distance between the proteins of any two species can be reconstructed by summing the appropriate branch lengths to give an " output " replacement distance. Such output distances will differ from so-called " input " distances, namely, those calculated directly from the amino acid sequences. This is because, after the first two subsets are joined, the distances from the other proteins to the new joined subset can only be calculated in terms of the average of the distances from every other protein to those in the first subset, and the utilization of average distances necessarily continues throughout the computation. Therefore, the initial tree constructed need not necessarily represent the best utilization of the data. One procedure for seeking an optimal tree is to calculate a percent standard deviation between the distances reconstructed from the tree and the original input replacement distances. Alternative trees are examined, and that which shows the smallest percent standard deviation is considered to be the best.

However, this is not the only criterion that can be used in seeking an optimal tree. One could, for example, choose the three for which the total number of mutations is the least. Moreover, it is not possible to examine all possible trees since there are a very large number of such trees, and there are no known alogorithins which can choose the one best tree, by either of the above criteria, without examining too many trees to be practicable. Thus, for n species there are

$$(2n-3)\frac{(2n-5)!}{(n-3)!\,2^{n-3}}$$

trees [31]. For 29 species this corresponds to more than 10^{36} different trees. Several variations of common numerical taxonomic methods are therefore used by different authors [30, 32, 33] to pick " reasonable " trees for examination.

Whatever criteria are utilized, it is remarkable that the resulting phylogenies are generally in good accord with ordinary biological classifications, even though the amino acid sequences of the set of orthologous proteins, the genetic code and a simple set of statistical calculations were strictly the only information employed. The phylogenetic tree derived from the structures of eukaryotic cytochromes *c* (Fig. 4) is not by any means perfect. Some of the relations depicted are certainly erroneous. Thus, primates branch off the ancestral mammalian line before marsupials, the turtle is nearer the birds than the other reptile (the rattlesnake) in the set, and the shark appears to relate more closely to the lamprey than to the tuna. Nevertheless, before this type of procedure was avail-

able, a phylogeny as accurate as this could not be derived from a single trait, let alone a single gene. Clearly, this must be because an examination of the number of mutations fixed in the course of the evolution of a single gene yields a considerably more precise estimate of the extent of evolutionary divergence than that from a single morphological trait. Indeed, one can expect that when sufficient amino acid sequence data for various sets of proteins become available, precise phylogenies will be readily obtainable by such procedures.

Of the other proteins for which structural information has accumulated, fibrinopeptide A, cleaved by thrombin from the amino-terminal segment of fibrinogen in blood clotting, has led to a satisfactory phylogenetic tree (Fig. 5) for a set of 23 species much more closely related than those represented in the cytochrome *c* tree (Fig. 4). This segment of fibrinogen varies rather rapidly during evolution, which together with its small size (19 residues) makes it most useful in examining a narrow taxonomic span of species. As shown by Mross and Doolittle [34, 35], the structures of the fibrinopeptides from 19 artiodactyls fit very well the classical phyletic relations of these species. Other such relatively small groups can surely be studied as effectively on this basis.

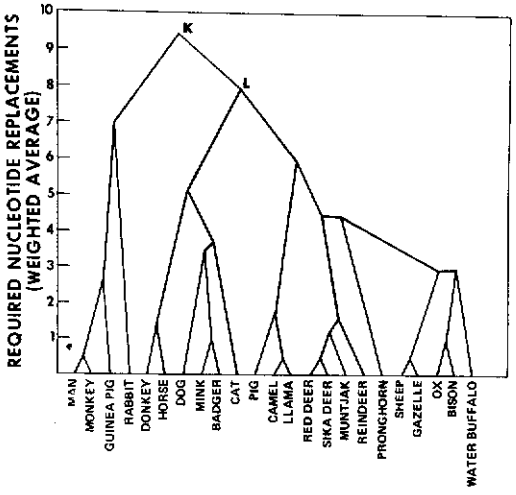

Fig. 5. Statistical phylogenetic tree based on the replacement distances between the fibrinopeptides A of the species listed. The topology of the tree was obtained by the procedure of Fitch and Margoliash [30]. The nucleotide replacements were counted by the procedure of Fitch [36]. Other markings as for Figure 4. References to the amino acid sequences of the fibrinopeptides are given in References 34 and 35.

An insufficient number of structures of orthologous proteins in the hemoglobin-myoglobin and in the ferredoxin series are as yet available to lead to useful phylogenetic trees.

Reconstruction of Ancestral Amino Acid Sequences and the Distinction between Divergent and Convergent Evolutionary Processes

The reconstruction of the amino acid sequence of the ancestral form of the protein at each of the branching points of the phylogenetic tree can be carried out, following certain rules, from a phylogenetic tree and the amino acid sequences of the present day proteins [30, 31, 36]. An example of the result of such a procedure is the ancestral cytochrome *c* sequence (Fig. 6) corresponding to the structure derived for the cytochrome *c* of the ancestral species at the topmost apex of the phylogenetic tree. The ambiguities result from the lack of sufficient data to decide unequivocally what is the codon for every residue position.

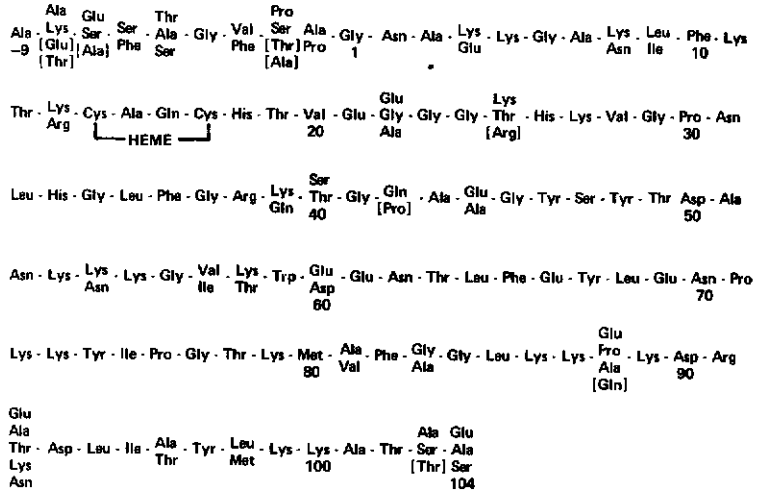

Fig. 6. Amino acid sequence of the ancestral form of cytochrome *c* at the topmost apex of the phylogenetic tree. Any of the amino acids shown would permit the evolution of the 29 descendent cytochromes *c* in the minimum number of 366 nucleotide replacements, assuming the topology shown in Figure 4. Amino acids in brackets have not yet been observed in any present day cytochrome *c*.

Similar procedures can be utilized to distinguish between divergent and convergent evolutionary processes [36]. Consider two sets of orthologous proteins which are to be tested for homology; it is possible to reconstruct the probable nucleotide (or nucleotides where some ambiguity may exist) for every position of the two ancestral genes for the two sets. If the same nucleotide is present in a certain position in both ancestral genes, then any differences in present day sequences are of a divergent character. If, on the other hand, a different nucleotide occurs in a given position in the two ancestral genes, then any similarity between the present day proteins of the two sets is of a conver-

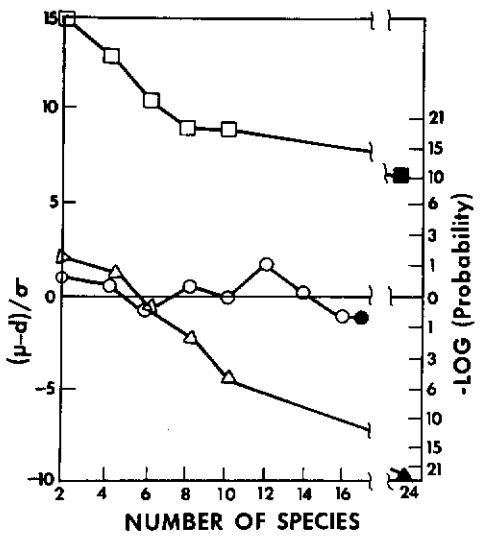

Fig. 7. Convergence and divergence as a function of the number of species. The abcissa gives the total number of sequences examined. Open symbols indicate that equal numbers of sequences were present in the two groups compared, closed symbols that they were divided unequally between the two groups. The ordinate gives the deviation (δ) from expectation ($\mu - d$), in standard deviation units on the left, and the equivalent probability of a result being due to chance is given as negative powers of 10 on the right. Points above the zero line represent an excess of divergent comparisons, below the line, an excess of convergent comparisons. Random sequences of amino acids are shown by circles (o——o), convergent sequences by triangles (△——△), and divergent sequence of squares (□——□). The convergent sequences were obtained by computer simulation. The divergent sequences compare fungal to non-fungal cytochromes *c*. According to Fitch [36].

gent character for that particular position. To consider the complete structural genes, one can, assuming that the descendent nucleotide sequences are completely unrelated, estimate how many of the some 300 ancestral nucleotide comparisons (for proteins of 100 residues) would be expected to be of a divergent and how many of a convergent type. A significant excess of one or the other type would make it possible to decide whether the two sets were divergently related or only similar because of convergence. Typical results are shown in Fig. 7. The abcissa plots the number of different species in the two trees being compared. The horizontal line is the line of mean expectation. The curve fluctuating about it was obtained using random sequences of 100 amino acids, showing that when the proteins are entirely unrelated there is no excess of either divergent or convergent relationships. The ordinate gives the standard deviation from expectation, so that points above the line of mean expectation indicate excess of divergent over convergent comparisons, and points below the line the opposite situation. The lower curve was obtained from two sets of amino acid

sequences that were made to simulate a convergent evolutionary process by a computer. The curve above the line is for two sets of eukaryotic cytochromes *c* composed of fungal and non-fungal proteins. The result clearly shows that fungal and non-fungal eukaryotic cytochromes *c* had a common evolutionary origin. It should be noted that orthology within each of the two groups is the only required assumption.

Invariant Codons and Covarions

Possibly the most useful of all present applications of statistical phylogenetic trees is the estimation of the number of invariant codons in the structural gene for the protein considered [37]. These represent positions in the polypeptide chain for which only one particular amino acid can fulfill the required function satisfactorily, so that the probability of a line of evolutionary descent surviving the fixation of a mutation in these codons is essentially nil. All mutations in such codons are termed *malefic* [37].

The phylogenetic tree based on cytochrome *c* structures (Fig. 4) prescribes the distribution of codons in the structural gene which have undergone 0, 1, 2, 3 or more replacements in their descent from the common ancestral form. That distribution can be accounted for if one assumes that there are three classes of codons. One class is invariant. The other two vary in a random fashion according to two different rates, one, the " hypervariable " set of codons, changing much more rapidly than the other [37-39]. There are probably more than two rates of variation, but two rates are sufficient to fit the presently available data [38]. All codons belonging to the same variable set are equally likely to fix the next nucleotide replacement, and for each, therefore, the number of codons that have undergone 1, 2, 3 ... replacements will follow a Poisson distribution. Fitting such distributions to the data obtained from the cytochrome *c* phylogenetic tree in Figure 4 makes it possible to estimate the size of the three sets. The best fit is for an invariant set of 32 residues, a normally variable set of 65 residues and a " hypervariable " set of 16 residues [38]. This last appears to fix mutations in the course of evolution some 3.2 times faster than the normally variable codons [38].

The above calculation employed 29 different cytochromes *c* of species ranging from fungi to vertebrates (see Fig. 4), and yielded an estimate of the percent of the cytochrome *c* gene that was invariant of about 25 % [38]. A similar estimate was made earlier using the cytochromes *c* of only 20 species, but covering the same taxonomic range [37]. However, if one selectively and gradually excludes the proteins of the more remote groups of species from the calculation, the resulting percent of the gene found to be invariant increases. If these values are plotted as a function of the average replacement distance for all the species taken into account for each recalculation (Fig. 8), a roughly

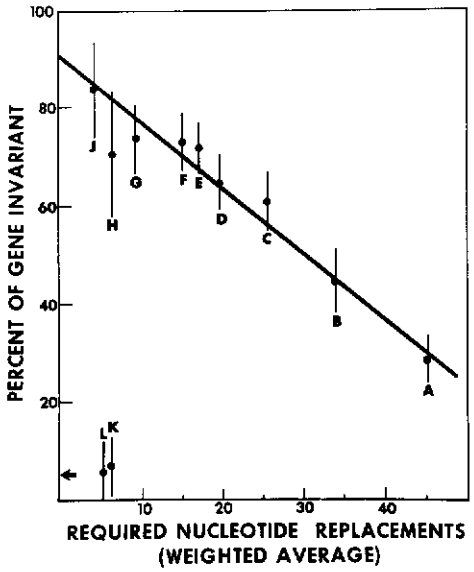

Fig. 8. Concomitantly variable codons. The percent of the gene found to be invariant is plotted as a function of the weighted average of required nucleotide replacements (height of peaks in Figures 4 and 5 for all the species in each comparison). Letters A to J represent the groups of cytochromes *c* indicated in Figure 4, letters L and K the groups of fibinopeptides indicated in Figure 5. The arrow on the ordinate is the position equivalent to one invariant residue out of the 19 residues of fibrinopeptide. The line at each point is an estimate of the standard deviation of the ordinate value of the point. A weighted least squares fit to the results for cytochrome *c* is extrapolated to the abcissa to estimate the fraction of the gene for which all mutations are lethal or malefic. According to Fitch and Markowitz[38].

linear regression is obtained. On extrapolation, a value showing over 90 % of the gene to be invariant is obtained when the replacement distance is zero [38]. This demonstrates that in any one mammalian cytochrome *c* at the present time, only about 10 residues can undergo changes without leading to a lethal or malefic change [38]. Moreover, it seems reasonable that if enough data were available to make similar extrapolations towards fungal cytochromes *c* or insect cytochromes *c*, for example, an essentially similar result would be obtained. The codons corresponding to those amino acid positions which, in any one species and at any one time in the course of evolution, are free to fix mutations may be termed *concomitantly variable codons* or *covarions* [38].

This conclusions is particularly important as it demonstrates that in the cytochrome *c* of any one species only a very small proportion of all residue positions that have varied, as among the cytochromes *c* of the more than 30 species investigated, are in fact variable. This very stringent limitation on evo-

lutionary change in protein structure must be due to the complex interplay of structural-functional requirements. In cytochrome *c*, in addition to the types of residue interaction common to other proteins, the relatively short peptide chain must essentially wholly enclose the evolutionarily invariant heme, in a way that requires a relatively large number of internal residues to be in contact with the prosthetic group [40-42]. Moreover, provision must be made to adapt the outer surface to specific interactions with three different macromolecular surfaces, those of cytochrome oxidase, cytochrome reductase and the mitochondrial membrane binding site for cytochrome *c*. These contacts could well involve a major proportion of the surface of the protein.

In order to account for the observed variation of over two thirds of the residues of cytochrome *c* in the proteins of a wide taxonomic range of species, one must assume that when a mutation is fixed in a particular covarion, it may also change some of the members of the set of covarions. Thus, over extended periods of evolutionary history, more than 70 residue positions have shown substitutions.

The number of covarions, obviously an expression of the extend and tightness of structural-functional requirements, appears to represent a fundamental parameter which is nothing else than a quantitative expression of the effect of function on the evolutionary behavior of proteins. Though the number of covarions may well vary somewhat for the same protein in different species, it nevertheless appears to impose the average rate of evolutionary change so characteristic of every protein.

An excellent example is provided by the comparison of cytochromes *c* and fibrinopeptides A [38]. As depicted in Fig. 8, 18 of the 19 residues of fibrinopeptides A are variable if one considers the fibrinopeptides of all the species listed in Fig. 5, and remarkably, this number does not appear to change as the range of species is decreased. The number of covarions for fibrinopeptide A thus appears to be 18. [Because of the relatively small range of species for which the data are available, this estimate is probably not as accurate as that for cytochrome *c*, and the correct value could be 17]. Since the time of the common ancestor of the horse and the pig, the phylogenetic tree for cytochrome *c* indicates that 5 nucleotide replacements were fixed in the 104 codons in both lines of descent to the present day genes, while the tree for fibrinopeptide A shows 13 nucleotide replacements for 19 codons. This corresponds to 0.048 and 0.684 fixations/codon, as expected from the known slow conservative nature of evolutionary changes in cytochromes *c* [40-43] and the very rapid changes of fibrinopeptides [34]. However such calculations include not only the codons which can undergo changes, namely covarions, but also all the codons for which variations are either lethal or malefic. If one excludes the latter, the values become $5/10 = 0.50$ for cytochrome *c*, and $13/18 = 0.72$ for fibrinopeptide A in fixations/covarion [38]. Considering the probable error