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

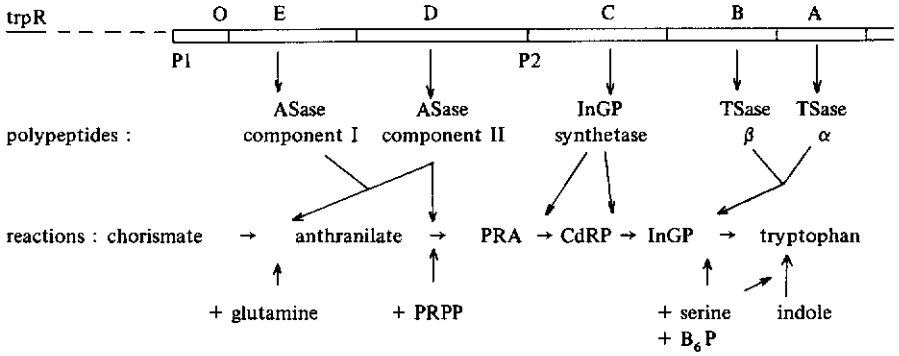


Fig. 1. The tryptophan operon of *E. coli*. The operon consists of 5 structural genes and adjacent controlling elements. TrpR is an unlinked gene specifying a protein repressor of the operon. O is the operator region and P1 and P2 are promoter regions. The various reactions in the pathway and the enzymes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

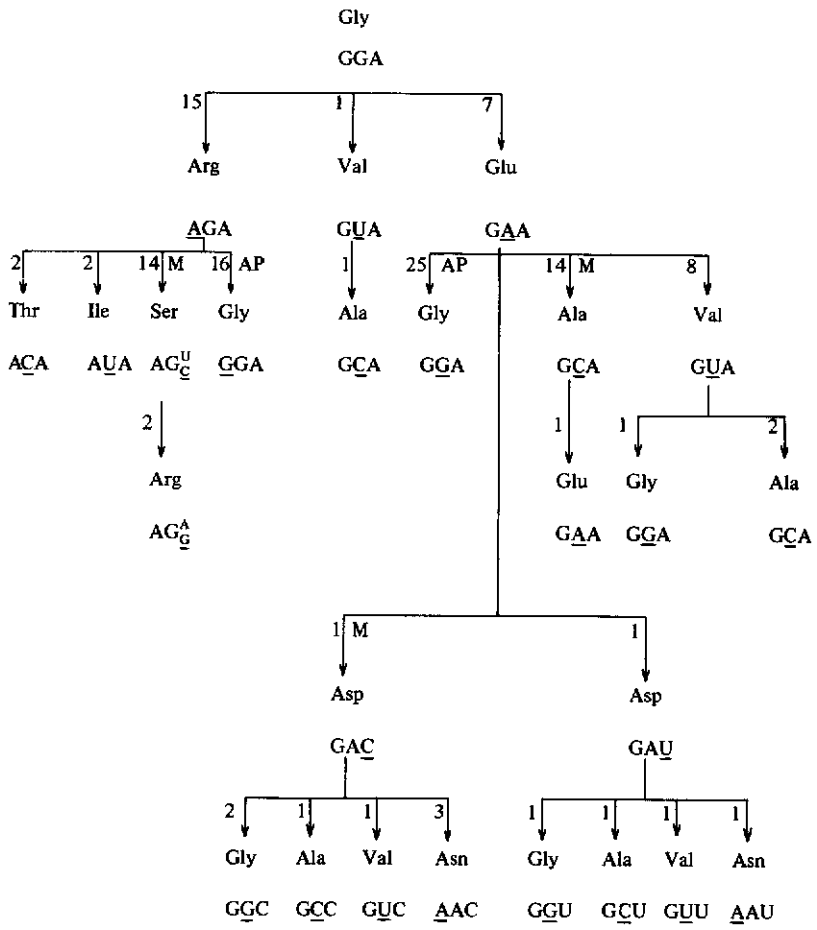


Fig. 2. Amino acid changes that have been observed at position 210 in the A protein and the probable corresponding codons [2, 3]. The number alongside each arrow indicates how many times the change was demonstrated by protein structure analyses. AP and M indicate changes favored by 2-aminopurine and 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

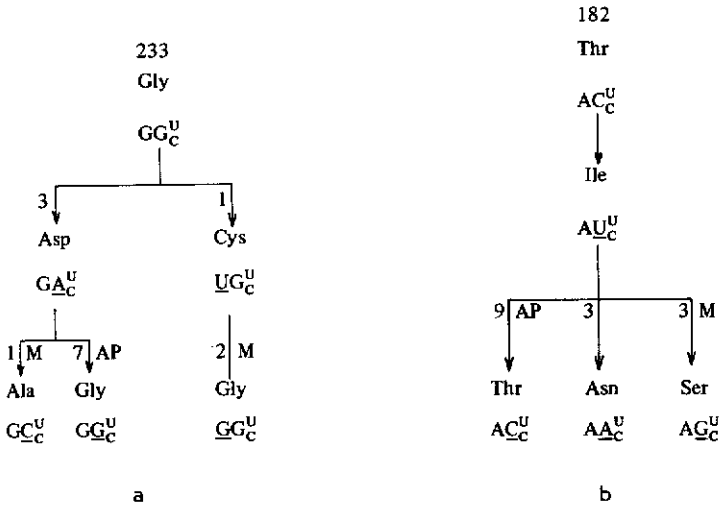


Fig. 3. Amino acid changes at position 182 (Fig. 3 A) and 233 (Fig. 3 B) and the probable corresponding codon changes [3]. See legend to Fig. 2 for other explanatory information.

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

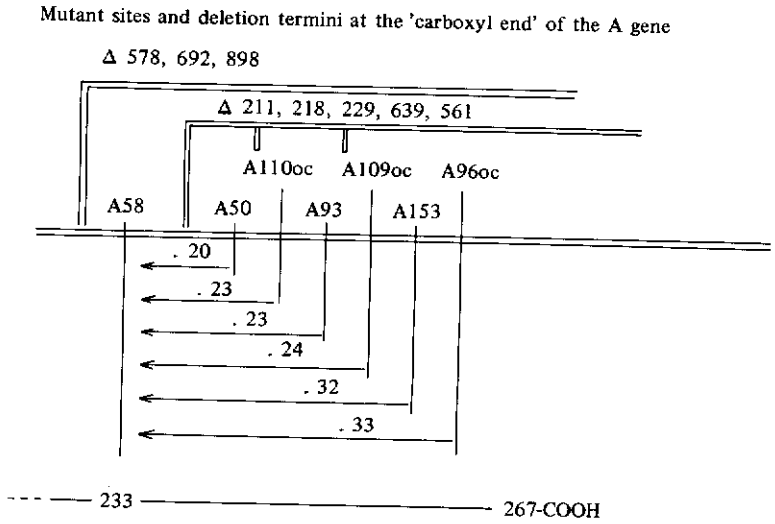


Fig. 6. Mutant sites and deletion termini in the region of the A gene specifying the carboxyl end of the A protein. Map distances are indicated above the arrows. Three of the point mutants (A 110, A 109, A 96) are ochre nonsense mutants. The precise terminus of each deletion is not known but it ends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

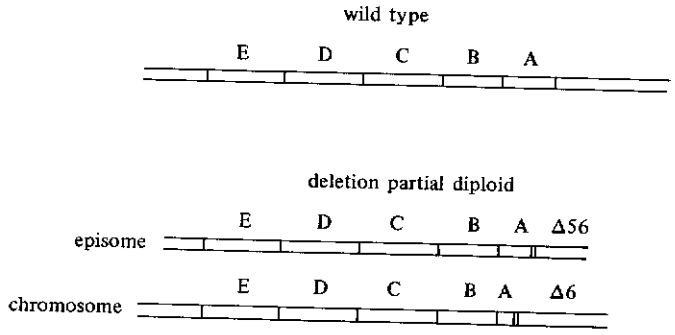


Fig. 7. A diploid strain employed in a prototrophy selection experiment. Deletion 6 removes most of the A gene while deletion 56 removes that region specifying the last thirty or so residues.

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

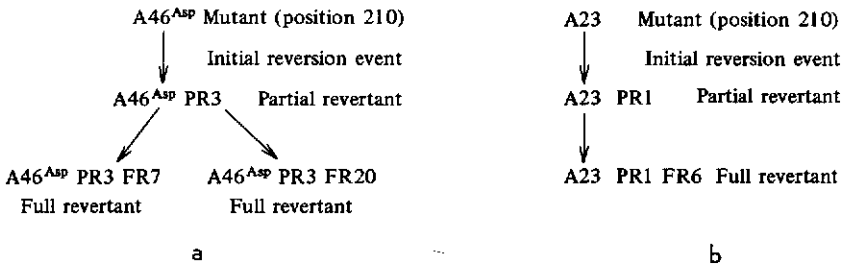


Fig. 8. Sequential reversion events [2] starting with mutant A 46^{Asp} (aspartic acid at position 210), A, and mutant A 23 (arginine at position 210), B.

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

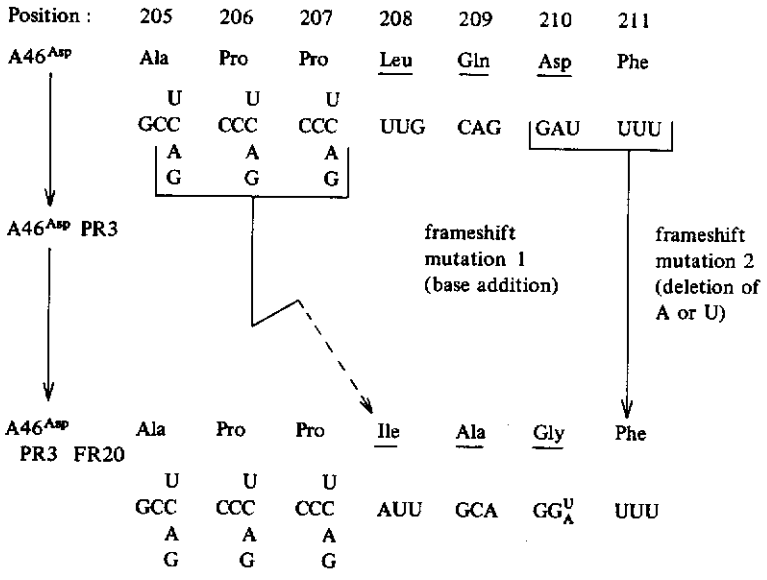


Fig. 9 a

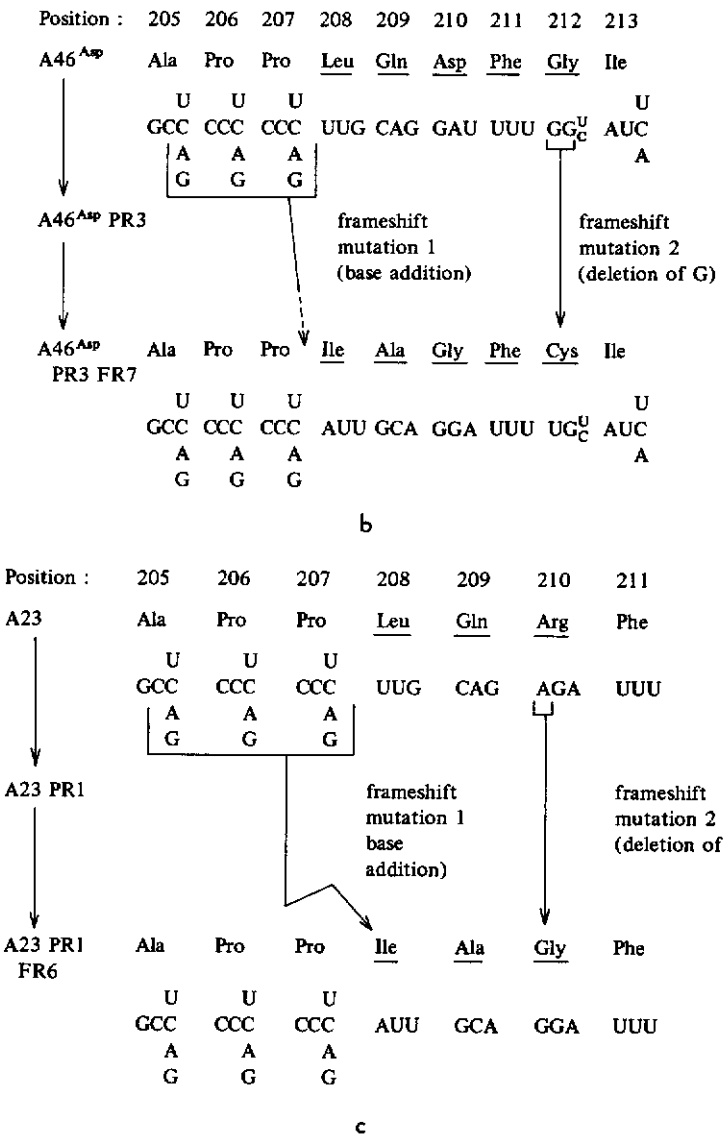


Fig. 9. Amino acid changes and probable corresponding nucleotide sequences in frameshift revertants A 46^{Asp} PR3 FR20 (A), A 46^{Asp} PR3 FR7 (B) and A 23 PR1 FR6 (C) [2].

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.

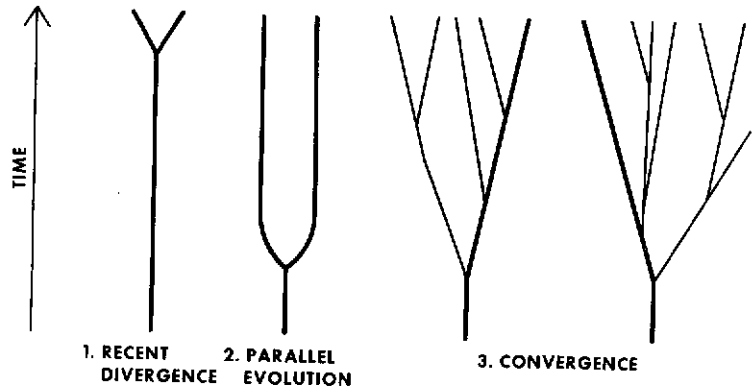


Fig. 1. Possible evolutionary reasons for similarity of polypeptide chains.

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

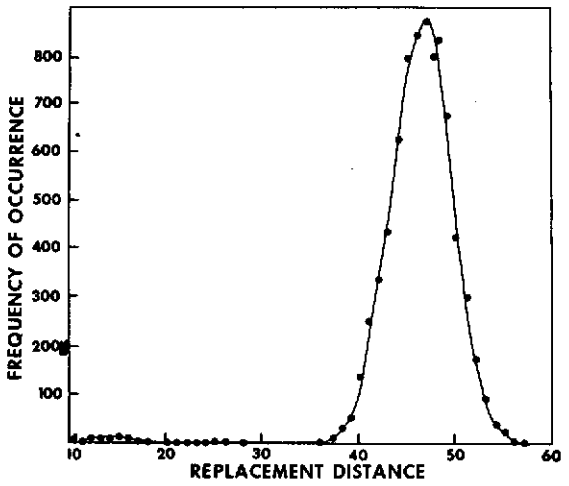


Fig. 2. Comparison of replacement distances for all possible 30 residue segments of human cytochrome *c* (6) and bakers' yeast iso-1 cytochrome *c* (7) by the procedure of Fitch [5]. The number of times various replacement distances occur in the comparisons are given on the ordinate (Frequency of occurrence).

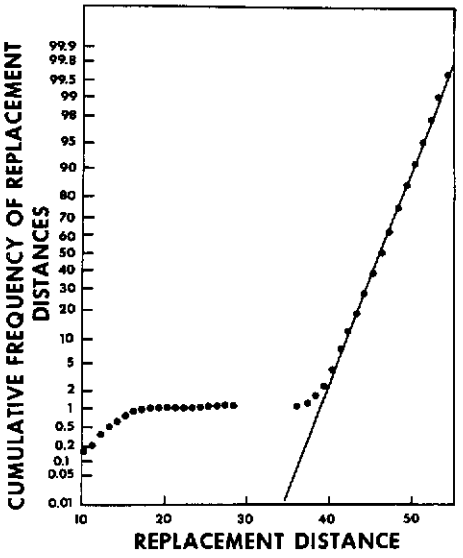


Fig. 3. Probability plot of the data from Figure 2. The random part of the distribution is given by the linear portion of the curve. The comparisons for which the replacement distances are smaller than expected for a random distribution are represented by the points which deviate from the straight line 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).

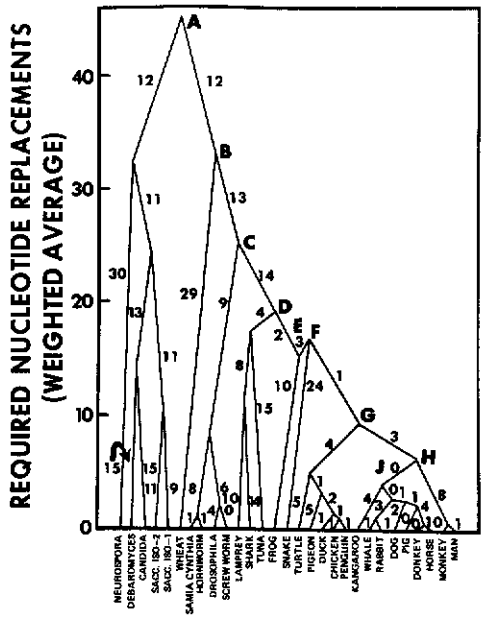


Fig. 4. Statistical phylogenetic tree based on the replacement distances between the cytochromes *c* of the species listed, as obtained by the procedure of Fitch and Margoliash [30]. Each number on the figure is the replacement distance along the line of descent as counted by the procedure of Fitch [36]. 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.

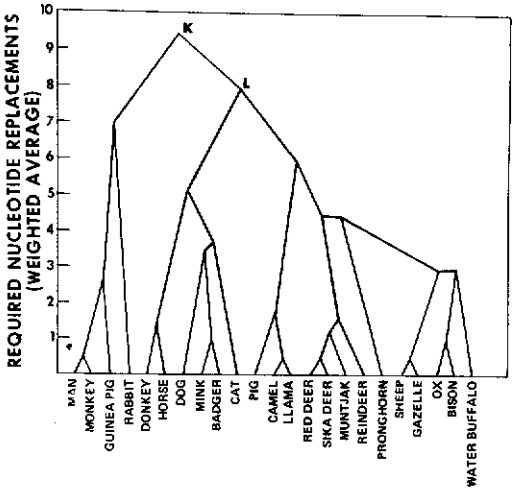


Fig. 5. Statistical phylogenetic tree based on the replacement distances between the fibrinopeptides A of the species listed. The topology of the tree was obtained by the procedure of Fitch and Margoliash [30]. The nucleotide replacements were counted by the procedure of Fitch [36]. Other markings as 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.

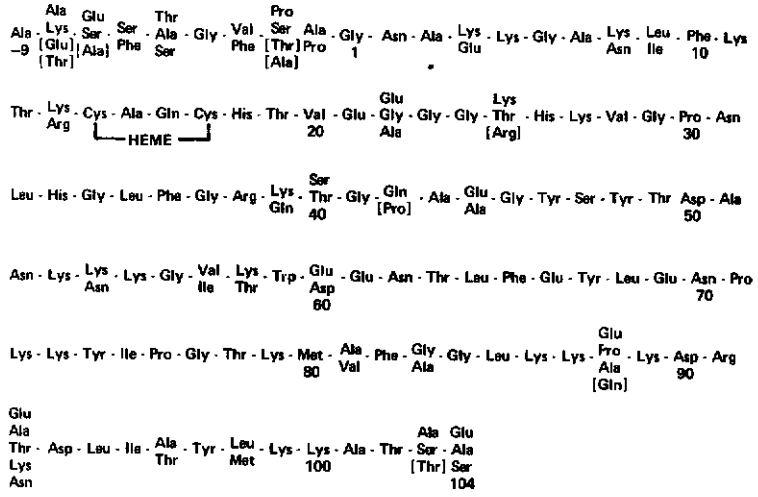


Fig. 6. Amino acid sequence of the ancestral form of cytochrome *c* at the topmost apex of the phylogenetic tree. Any of the amino acids shown would permit the evolution of the 29 descendent cytochromes *c* in the minimum number of 366 nucleotide replacements, assuming the topology shown in Figure 4. 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-

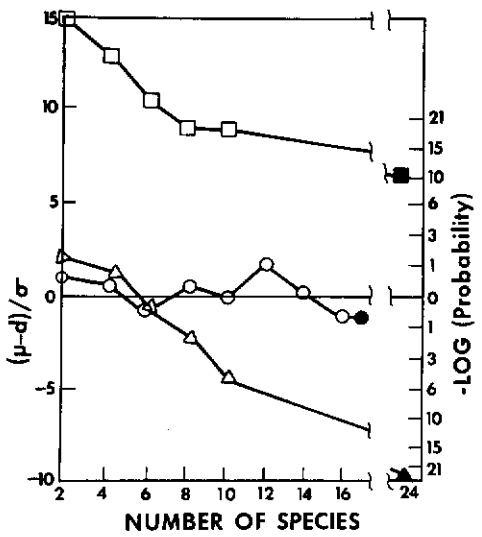


Fig. 7. Convergence and divergence as a function of the number of species. The abcissa gives the total number of sequences examined. Open symbols indicate that equal numbers of sequences were present in the two groups compared, closed symbols that they were divided unequally between the two groups. The 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

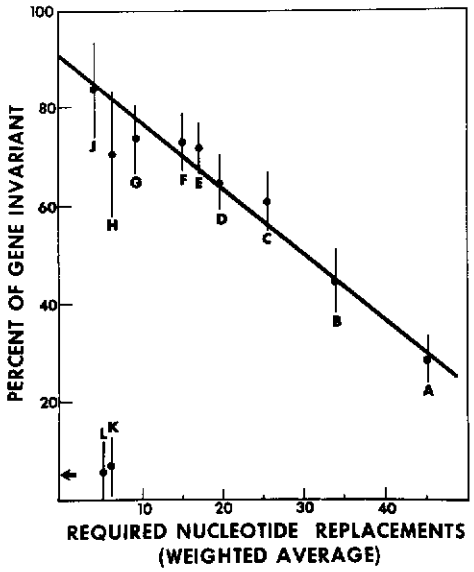


Fig. 8. Concomitantly variable codons. The percent of the gene found to be invariant is plotted as a function of the weighted average of required nucleotide replacements (height of peaks in Figures 4 and 5 for all the species in each comparison). Letters A to J represent the groups of cytochromes *c* 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

of such estimates, these two values cannot be significantly different and should be thought of as in remarkably good concordance.

If similar values are obtained for other proteins when a sufficiency of primary structures become available it will be possible to conclude that the overall rate of evolutionary change of a protein is determined by the number of its covarions. Such a result would also be compatible with the contention that for those positions that are amenable to change the occurrence of variations is governed by random processes, namely, that the mutations fixed are selectively neutral. The pros and the cons of such an interpretation have been discussed elsewhere [26, 38, 43-49] and it would not be practicable to reproduce the entire, rather inconclusive, argument here. However, regardless of whether the mutations are or are not neutral, wide applicability of the relation between covarions and evolutionary rate of change would provide, given a suitable paleontological reference point in time, a method for dating evolutionary events, such as speciations in the case of species phylogenies and gene duplications in the case of gene phylogenies. This procedure would be a statistically elaborated counterpart of earlier crude attempts to date evolutionary events directly from the average numbers of residue or nucleotide differences for the eukaryotic cytochromes *c* of different taxonomic groups of species and a paleontological reference point [41-43, 50].

Early Evolutionary History of Replicating Macromolecules

The most fascinating possibility opened up by the statistical approach to the evolutionary information contained in the structures of present-day macromolecules, is an approach to the distant past. Our ability to extrapolate backward in biological history is indeed not limited to the latest common ancestor of present species. By examining numerous orthologous sets of proteins and nucleic acids, it will be possible to reconstruct the structure of the ancestral form for each set. Having in this way eliminated the mutations that were accumulated between the time when these early species existed and the present, one may well observe similarities between the ancestral sequences which are not obvious in their extant descendants. This procedure would thus identify paralogous relationships when none could be observed in present day macromolecules and make it possible to establish an extensive gene phylogeny. One could work out a tree based on the ancestral sequences and thus estimate the structure of the ancestral sequence of these ancestral sequences. This could in principle be done both with proteins and nucleic acids, independently. The second order ancestral sequences, if these can be attained, may well take us back to very early stages at the borders of chemical and biological evolutions at which the genetic code and machinery actually originated. The earliest " protein " and " nucleic acid " should indeed correspond one to the other if our present

speculative concepts of the evolution of the genetic code [51-54] are factually correct.

References

[1] E. Zuckerkandl and L. Pauling, *J. Theoret. Biol.*, **8**, 357 (1965).

[2] H. Neurath, K.A. Walsh and W.P. Winter, *Science*, **158**, 1638 (1967).

[3] B.S. Hartley, This Volume.

[4] K. Brew, T.C. Vanaman and R.L. Hill, *J. Biol. Chem.*, **242**, 3747 (1967).

[5] W.M. Fitch, *J. Mol. Biol.*, **16**, 9 (1966).

[6] H. Matsubara and E.L. Smith, *J. Biol. Chem.*, **238**, 2732 (1963).

[7] Y. Yaoi, K. Titani and K. Narita, *J. Biochem.*, **59**, 247 (1966).

[8] W.M. Fitch, *J. Mol. Biol.*, **49**, 1 (1970).

[9] W.M. Fitch, *Biochem. Gen.*, **3**, 99 (1969).

[10] M. Tanaka, T. Nakashima, A.M. Benson, M.F. Mower and K.T. Yasunobu, *Biochemistry*, **5**, 1666 (1966).

[11] A.M. Benson, H.F. Mower and K.T. Yasunobu, *Arch. Biochem. Biophys.*, **121**, 563 (1967).

[12] W.M. Fitch, *J. Mol. Biol.*, **16**, 17 (1966).

[13] R.V. Eck and M.O. Dayhoff, *Science*, **152**, 363 (1966).

[14] T.H. Jukes, *Molecules and Evolution*, Columbia Univ. Press, New York, 1966.

[15] S.J. Singer and R.F. Doolittle, *Science*, **153**, 13 (1966).

[16] R.L. Hill, R. Delaney, R.E. Fellows and H.E. Lebowitz, *Proc. Natl. Acad. Sci. U.S.*, **56**, 1762 (1966).

[17] R.L. Hill, H.E. Lebowitz, R.E. Fellows and R. Delaney, in J. Killander, Editor, *Gamma Globulins*, Interscience Publishers, New York, 1967, p. 109.

[18] U. Rutishauser, B.A. Cunningham, C. Bennett, W.H. Konigsberg and G.M. Edelman, *Proc. Natl. Acad. Sci. U.S.*, **61**, 1414 (1968).

[19] G.M. Edelman, B.A. Cunningham, W.E. Gall, P.D. Gottlieb, U. Rutishauser and M.J. Waxdal, *Proc. Natl. Acad. Sci. U.S.*, **63**, 78 (1969).

[20] C. Baglioni, *Proc. Natl. Acad. Sci. U.S.*, **48**, 1880 (1962).

[21] C. Baglioni, *Biochim. Biophys. Acta*, **97**, 37 (1965).

[22] C.C. Curtain, *Australian J. Exptl. Biol. Med. Sci.*, **42**, 89 (1964).

[23] J. Barnabas and C.J. Muller, *Nature*, **194**, 931 (1962).

[24] D. Labie, W.A. Schroeder and T.H.J. Huisman, *Biochim. Biophys. Acta*, **127**, 428 (1966).

[25] J.A. Black and G.H. Dixon, *Nature*, **218**, 736 (1968).

[26] W.M. Fitch and E. Margoliash, *Evolutionary Biology*, **4**, 67 (1970).

[27] H.A. Itano, *Advan. Protein Chem.*, **12**, 215 (1957).

[28] V.M. Ingram, *Nature*, **189**, 704 (1961).

[29] V.M. Ingram, *The Hemoglobins in Genetics and Evolution,* Columbia Univ. Press, New York, 1963.

[30] W.M. Fitch and E. Margoliash, *Science,* **155,** 279 (1966).

[31] W.M. Fitch and E. Margoliash, *Brookhaven Symp. Biol.,* **21,** 217 (1968).

[32] R.R. Sokal and P.H.A. Sneath, *Principles of Numerical Taxonomy,* Freeman and Co., San Francisco, 1963.

[33] M.O. Dayhoff, *Atlas of Protein Sequence and Structure,* Nacional Biomedical Research Foundation, Silver Spring, Md., 1969.

[34] G. A. Mross and R.F. Doolittle, *Arch. Biochem. Biophys.,* **122,** 674 (1967).

[35] R.F. Doolittle, personal communication.

[36] W.M. Fitch, *Systematic Zoology,* **19,** 99 (1970).

[37] W.M. Fitch and E. Margoliash, *Biochem. Gen.,* **1,** 65 (1967).

[38] W.M. Fitch and E. Markowitz, *Biochem. Gen.,* **4,** 579 (1970).

[39] E. Markowitz, *Biochem. Gen.,* **4,** 595 (1970).

[40] E. Margoliash and A. Schejter, *Advan. Protein Chem.,* **21,** 113 (1966).

[41] E. Margoliash and W.M. Fitch, *Trans. N. Y. Acad. Sci.,* **151,** 359 (1968).

[43] E. Margoliash, W.M. Fitch and R.E. Dickerson, *Brookhaven Symp. Biol.,* **21,** 259 (1968).

[44] M. Kimura, *Nature,* **217,** 624 (1968).

[45] J.M. Smith, *Nature,* **219,** 1114 (1968).

[46] J.L. King and T.H. Jukes, *Science,* **164,** 788 (1969).

[47] K.W. Corbin and T. Uzzell, *American Naturalist,* **140,** 37 (1970).

[48] P. O'Donald, *Nature,* **221,** 815 (1969).

[49] N. Arnheim and C.E. Taylor, *Nature,* **223,** 900 (1969).

[50] E. Margoliash and E. L. Smith in V. Bryson and H.J. Vogel, Editors, *Evolving Genes and Proteins,* Academic Press, New York, 1965, p. 221.

[51] T.H. Jukes, *Biochem. Biophys. Res. Comm.,* **19,** 391 (1965).

[52] C.R. Woese, *The Genetic Code,* Harper and Row, New York, 1967.

[53] F.H.C. Crick, *J. Mol. Biol.,* **38,** 367 (1968).

[54] L.E. Orgel, *J. Mol. Biol.,* **38,** 381 (1968).

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

DISCUSSIONS

B. B. LLOYD : I wonder whether there is any nutritional significance in the change from the horse to the donkey as between serine and threonine. Serine, as we know, can easily be made by mammalian metabolism, but threonine can't. Now between man and the monkey, the change is from threonine to isoleucine, both of which are essential amino-acids, but threonine is a rather difficult amino-acid, in that you have got to have it in the L-form, whereas with isoleucine you can use the D- or L-isomer.

There is one other question and that is that the computer-produced tree is terribly like the sort of tree you get when you do textual criticism, according to a procedure that Dom Quentin, a French Priest, devised for the comparison of texts. I wonder if the statistics are the same. Thank you.

E. MARGOLIASH : In answer to your first question, I do not believe there is any relation between nutrition and amino acid substitutions in a series of homologous proteins. To take the example you cite, one cannot see how a serine for threonine substitution at position 48 in the cytochromes *c* from the donkey and the horse could possibly have a nutritional or amino acid biosynthetic concommittant, since both species carry numerous other proteins which contain both serines and threonines.

I cannot comment on your second question being unaware of the procedure you cite.

Question as reconstituted : Are the changes observed in the amino acid sequences of the cytochromes *c* of different species related to the genetic code ?

E. MARGOLIASH : Clearly there must be a relation to the genetic code since cytochrome *c* like other proteins is made by the usual mechanism. In fact, the majority of the mutations fixed in the course of evolution as given by the statistical phylogenetic tree correspond to single nucleotide changes, and the more or less expected number correspond to two nucleotide changes. The only quite remarkable phenomenon is that, when the numbers of mutations in the various classes are compared to the numbers one would expect on the assumption that all mutations are equiprobable, there is a major preponderance of $G \rightarrow A$ changes over all others, without a corresponding increase of the complementary $C \rightarrow U$ transitions. This is, by the way, also true of comparisons of normal and abnormal human hemoglobins, and of fibrinopeptides, and appears to be an interesting characteristic of the mutations accumulated in the course of the evolutionary divergence of proteins from a common ancestral form.

J. DUCHESNE : I would like to mention that it has been found in studies in my laboratory that the CO_2 exhaled by higher living systems varies in constitu-

tion when going from one class to another. The best examples which we have found up to now concern human beings and birds. In these classes, the isotopic ratio $C^{12}O_2/C^{13}O_2$ is not the same and indeed varies by about 0.5 per cent, in the sense of an enrichment in C^{13}, in the particular case of the domestic cock. This result, if it should be proved to be general may become, in its turn, a basis for observing molecular evolution, at the nuclear level. Certainly this isotopic effect expresses some complicated situation connected with enzymatic activity within the Krebs cycle. The advantage of the method is the simplicity and the accuracy of the measurements, performed by mass spectrometry, which reach 0.01 %. It should be mentioned that the isotopic changes are unsignificant between individuals belonging to a given class. This might therefore become a new method which could be added to the one you describe.

J. MONOD : I just wish to point out — and I think Dr. Margoliash should agree — that kinetic measurements *in vitro* of an enzyme or an electron carrier like this one, give us only an incomplete view of the complex interactions that protein has to realize *in vivo* in order to do his job. For instance, one very simple thought is that a protein which functions in relation with a complex system like mitochondria has to find the right place in the morphogenesis of the cell. It cannot be allowed to lie around in other places where it would have no business being. Therefore, this problem, which is very difficult to approach, of protein, — protein interactions *in vivo,* of functional proteins making the right connexions — loosely or strongly depending on which protein it is— must have had a very important role in evolution. Evolution of a given protein cannot be independant of the evolution of other neighbouring proteins, I mean, neighbouring in the cell. If something happens to another protein in the system such that the interactions are not strong enough, then there might be a compensating change, just like Dr. Yanofsky has shown that there are compensations from one site to another in tryptophane synthetase. It is most likely that compensations of this kind, between different proteins, which are not related genetically, must be rule.

E. MARGOLIASH : I cannot agree with you more. Certainly *in vitro* measurements of enzyme kinetics are far removed from the normal biological milieu in which a protein functions and with which it must interact appropriately. In the case of cytochrome *c*, however, we are in a rather fortunate position. The experiments I showed with the cytochrome *c*-cytochrome *c* oxidase system represent only one facet of the available evidence. Another, and one possibly even more pertinent to the question of the functional attributes that are evolutionary effective, is given by cytochrome *c* repletion experiments of mitochondria lacking cytochrome *c*. One can osmotically shock mitochondria in a mild way, cause up to about 97 % of the cytochrome *c* in them to flow out at the appropriate ionic strength, and then reintroduce under the proper conditions any cytochrome *c* you will.

Doing this with rat liver mitochondria, for example, one can show that human, horse, various bird, reptile, fish, insect and fungal cytochromes *c* can all indiscriminately be used to titrate back both substrate oxidations and oxidative phosphorylation in precisely identical fashions, even though the proteins employed vary from a few to about half the amino acid residues in the polypeptide chain. Substrate oxidations are recovered completely and phosphorylation to about 75 % of the original rate at amounts of added cytochrome *c* essentially equal to the amounts originally removed from the mitochondria. The one objection to this sort of experiment is that the depleted mitochondria are not precisely the same as the original organelles since the outer membrane has been irreversibly stretched, so that to remove the reintroduced cytochrome *c* no second osmotic shock is required, only a change of ionic conditions. However, functions of the inner membrane are maintained so one can assume that no drastic changes have occurred. Moreover, this objection has been elegantly countered by Mattoon and Sherman who used a strain of yeast which carried a chain terminating mutation at the codon for Residue 12 of the cytochrome *c* gene. Mitochondria from such a yeast appear to be quite normal except that they lack cytochrome *c*. Such mitochondria will take up cytochrome *c* from solution without any pretreatments and recover full function, and whether yeast or horse cytochrome *c* is used makes no difference.

In summary, I would say that both types of experiments are useful. Those that deal in the kinetics of purified systems such as with cytochrome *c* oxidase or reductase complement the experiments with mitochondria which represent a situation very near indeed to the *in vivo* situation in which cytochrome *c* normally functions. In evolutionary terms, since the structure of any natural cytochrome *c* does not seem to affect the two known functions of the protein, electron transport and a possible involvement in oxidative phosphorylation, even though the proteins come from many taxonomic groups and vary by as much as 50 % of their amino acid sequences, one of two conditions must prevail. Either, there are functions of cytochrome *c* which are presently unknown and provide the basis for evolutionary selection by virtue of functional differences related to the known structural differences, or the structural differences are evolutionarily neutral, that is have neither positive nor negative selective influences. There is at this time no experimental basis for a clear cut decision between the two possibilities for cytochrome *c*, though studies related to the distribution of evolutionary variations on the protein structure, as determined by X-ray cristallography, tend to favor the first.

J. BRONOWSKI : I want to comment on what Monod has just said. The point of the analyses that we have seen is that there are some 20 sites or so of this total of just over 100 which are absolutely invariant. In the context of this conference, the crucial question then is the following : are they invariant (and is their configuration invariant) because of some initial accident which has sur-

vived throughout evolution, or because without them the molecule cannot work at all. Two hypotheses are therefore possible : but I imagine that most of us here, who have seen this and similar evidence, would lean (as I do) to the view that the invariant sites owe their stability not to chance but to their function in the molecule. If we are right in this interpretation, then Monod's question is answered by the internal logic of the system, as follows. *In vivo* or *in vitro,* the molecule will only work if there are these 20 invariant sites, and they are therefore absolved from taking part in interactions with other proteins that are specific to one species. But the remaining sites are occupied by different amino-acids in different species, and at these sites the interaction with other species-specific proteins can take place. This is an important concept in explaining how the horse and the donkey, once they are separated by an evolutionary accident, can each function as animals in which the same kinds of proteins form, as it were, a basic scaffolding of life which nevertheless is differently interconnected in the different species.

E. MARGOLIASH : I am afraid I must disagree. There appear to be about 32 residues which are invariant in all cytochromes *c,* as estimated by a statistical calculation from the distribution of codons which have undergone 1, 2, 3, etc. mutations in the complete phylogenetic tree. These are positions which require the particular residues in them and no others. Any mutation which causes a change in residue is malefic, namely the line of descent bearing it will invariably disappear in evolution, so that our chances of observing such a substitution in the cytochrome *c* of an extant species is essentially nil. However, this does not at all mean that all other positions in the protein can carry any residue. There are very definite limitations on what many positions can contain. For example, at Residue 13 only lysine or arginine will do. There are many similar examples of so-called conservative substitutions. Other positions can accommodate a wider range, and still others appear to be able to carry as many as 8 or 9 different amino acids. The non-invariant positions, even though conservative, are very difficult to deal with in a quantitative manner. How can one decide, for example, if a lysine is more or less similar to an arginine than an isoleucine is to a leucine ? So we have put them all in the non-invariant class without differenciating between more or less non-invariant. It is only ignorance of how to manipulate such data that forces us to do so.

C. LEVINTHAL : Going back to the question of function due to related things, is there any evidence one way or the other as to whether or not there is an enzyme required to attach the heme to the peptide ?

E. MARGOLIASH : I wish we knew. There is some genetic evidence that can be interpreted to mean that the heme is put on rather early in the biosynthesis of cytochrome *c*. The process probably requires an enzyme as it seems unlikely

to procede spontaneously except with reduced intermediates of heme biosynthesis, as demonstrated by Sano. Of course, all this is only speculative.

C. LEVINTHAL : This clearly is a kind of constraint of function which would not have been tested in the experiment you mentioned.

E. MARGOLIASH : Yes, certainly. From the X-ray crystallographic structure it is certain that the thioether bond holding Residue 14 to pyrrole ring I is deep inside the protein, and also that a large proportion of the stabilization of the tertiary structure comes from side chain interactions with the bulky heme. An enzyme would be unlikely to be able to perform the making of a thioether link at the bottom of the heme crevice and conversely the protein would probably not take its native conformation until the heme is locked into place. This is all very well, but it is nevertheless also unlikely that variations of amino acid sequence of some 50 % of the structure could be accounted for by the requirements of the presumably somewhat different enzymes which may be involved in making the two thioether bonds in different species, whether the chain is unfolded or folded at the time this takes place.

M. EIGEN : You said you measured overall velocity. Was it maximum velocity ? Were the P.H. conditions the same for all the species you have tested ?

E. MARGOLIASH : What was discussed were first order rate constants for the reaction of the various cytochromes *c* with beef cytochrome *c* oxidase at a particular concentration of the cytochromes *c*. The pH and other conditions were optimal, and these do not differ for the cytochromes *c* of different species. What was varied was the concentration of the cytochromes *c*, since because of the peculiar kinetics of the reaction it is necessary to show identity of rates at different cytochrome *c* concentrations. This was done and identity was observed. What was not done was to vary conditions away from optimal and see whether it was then possible to detect differences between the cytochromes *c* of different species.

Question : Have you tested the thermal stability of the different cytochromes ?

E. MARGOLIASH : We have tested thermal stabilities and they do vary, but in a way which does not seem to make any evolutionary sense. We even believe we may have identified one reason for the relative stabilities. This sort of data has, however, best been worked out in terms of digestibility with proteolytic enzymes, particularly chymotrypsin, so let me describe these results. When Residue 3 of the polypeptide chain is an isoleucine, as in many bird cytochromes *c*, digestion is very poor, when it is valine, digestion is better, and when it is alanine or serine complete digestions are achieved in minutes rather than hours. Again, there is no obvious evolutionary meaning, as the proteins of phylogenetically relatively close species may show large variations in these parameters, while in some cases the proteins of widely divergent species may be very similar.

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

POPULATION GENETICS AND MOLECULAR EVOLUTION

J. MAYNARD SMITH

School of Biological Sciences, University of Sussex

1. Introduction

Population genetics and evolution theory, like other branches of biology, must now be reconsidered in the light of advances in molecular biology. In particular, two conclusions seem to be emerging :

(i) A much larger proportion (perhaps as high as one third) of gene loci are polymorphic in natural populations than had previously been suspected (Lewontin & Hubby, 1966; Harris, 1966).

(ii) The rate of gene substitution in evolution may have been much greater than previously suspected (Kimura, 1968).

These discrepancies between observation and expectation have reopened an old controversy about the relative importance of " genetic drift " and natural selection. The proponents of drift, or as it has more often been called recently, " non-Darwinian evolution ", hold that a large proportion of protein variants in natural populations, and of gene substitutions in evolution, occur because they are selectively neutral, conferring neither advantage nor disadvantage. Such neutral variants may spread through a population, and may ultimately be established as the sole common variant, by the chances of sampling in a finite population.

Of course no one denies that some polymorphisms are maintained by natural selection (e.g. by heterozygous advantage, or by frequency-dependant selection), or that some gene substitutions in evolution occur by natural selection. The argument is between those who hold that natural selection is the only important cause of protein polymorphism and evolution, and those who hold that most polymorphisms and gene substitutions are of selectively neutral variants.

A direct biochemical approach to the problem can show that the drift hypothesis is at least plausible. Many amino-acid substitutions are known which do not destroy the enzymic activity of a protein, or have any obvious selective effect. King & Jukes (1969) have recently reviewed some of this evidence. The data presented by Yanowsky and by Margoliash at this Conference also supports the drift hypothesis. I am not competent to discuss the problem from the point of view of a protein chemist; instead I shall consider it from the viewpoint of evolution theory.

2. The arguments for non-Darwinian evolution

The two strongest arguments in favour of the drift interpretation derive from the frequency distribution of amino acids in proteins, and from the uniformity of the rate of evolution for a given class of proteins. I will consider them in turn.

Figure 1 compares the observed frequencies of amino acids in 53 mammalian polypeptides which have been sequenced with the frequencies to be expected from the genetic code if the bases were arranged in a random sequence. The expected frequencies of amino acids coded for by 6 codons (e.g. leucine) are much higher than of amino acids coded for by one codon (e.g. methionine). The correspondence between theory and observation is remarkably close. It is precisely what would be expected on the drift theory. The correspondence does not of course imply that all amino acid substitutions are selectively neutral, only that a majority of them are.

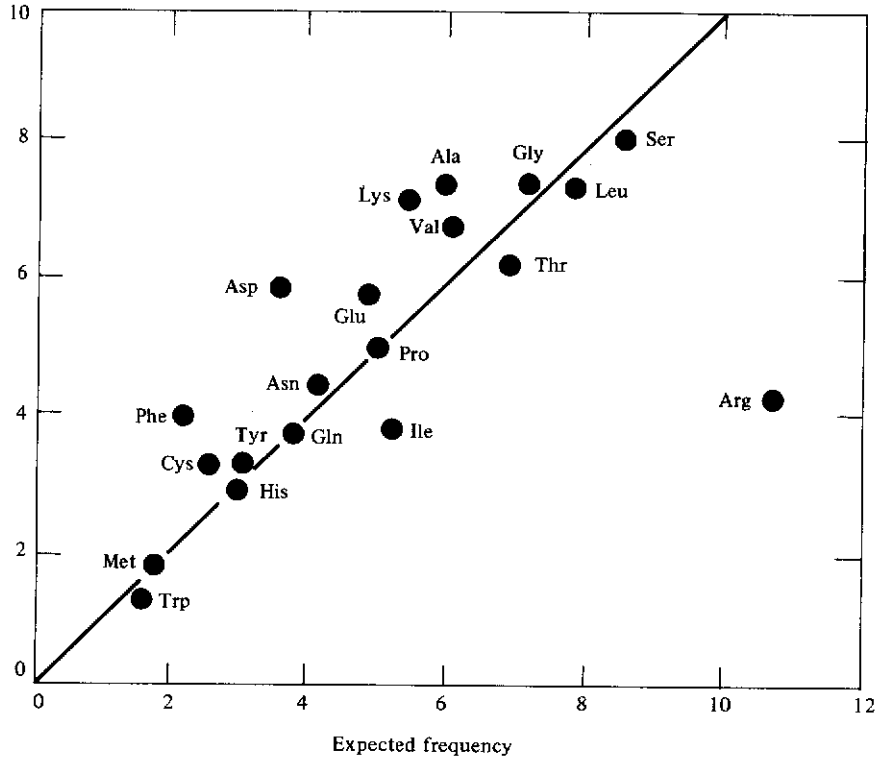


Fig. 1. A comparison of the observed number of amino acids in 53 completely sequenced mammalian proteins and the frequencies to be expected from the genetic code if most substitutions in evolution are selectively neutral (from King & Jukes, 1969).

A Darwinist might discount this correspondence by arguing that the code evolved under the influence of natural selection to provide a larger number of codons for amino acids which were more often needed. Other selectionist interpretations may be possible, but it has to be admitted that this correspondence can be predicted from the drift hypothesis, whereas a strict selectionist can only explain it away.

It is a prediction of the drift theory that the rate of evolution of a protein should be proportional only to the rate of neutral mutation. Thus let the rate of selectively neutral mutation per cistron per generation be u. Then if the population size is N, the number of new mutations is $2Nu$ per generation. Since in any generation there are $2N$ selectively equal genes, the probability that a new mutation will ultimately be established in the population is $1/2N$. Hence the rate of substitution of new mutants in the population is u per generation.

If we then suppose that for a given class of protein (e.g. haemoglobin) the total rate of mutation $\bar{u}$, and the fraction f of all mutations which are selectively neutral, are both constant (for different species and at different times), then $u = \bar{u}f$ is constant, and hence the rate of gene substitution per generation is constant. An estimate of the rate of gene substitution can be obtained by comparing protein sequences of existing species.

Figure 2 shows the relationships of some existing haemoglobins, and the approximate time scale involved. Table 1 shows rates of gene substitution for

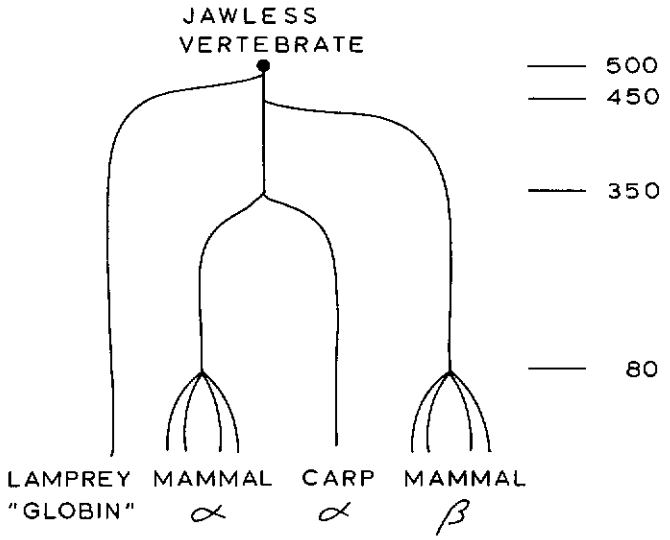


Fig. 2. Relationships between some existing haemoglobin chains. Lines represent continuity of descent by DNA replication. The figures on the right represent time in millions of years. It is assumed that a single gene specifying a globin in ancestral jawless vertebrates duplicated, and that the resulting genes diverged to specify the α and β chains.

TABLE 1

The rate of amino acid substitution in the evolution of haemoglobin, after Kimura, 1969.

Comparison	substitutions per amino acide site per year X 10^{10}
various mammalian α chains	9.1 – 10.9
various mammalian β chains	9.3 – 14.9
carp α cf. human α	8.9
human α cf. human β	8.9
human β cf. lamprey globin	12.8

some of the available comparisons. Two points should be made about these estimates. First, they may be underestimates, since if the same amino acid substitution has occurred in two lineages (parallel evolution), it will not be included. Second, the estimates are not all fully independant, since the same lineage is sometimes included in two estimates (e.g. comparing man-horse, and man-carp, the lineage from primitive mammal to man is included in both). There is, however, sufficient independance for the agreement to be remarkable. This agreement again is one which would be predicted by the drift theory, but which might be explained away by a strict selectionist.

Two further points need to be made. First, Table 1 shows uniformity of rate per year, not per generation. This would fit the drift theory if it turns out that for vertebrates mutation rates are likewise uniform with time rather than per generation.

The second point is that in so far as data is available (see King & Jukes, 1969) different types of protein evolve at different rates. Some (e.g. cytochrome c) have evolved more slowly than haemoglobin, others (e.g. fibrino peptide) more rapidly. This is to be expected if f, the fraction of all amino acid substitutions which are selectively neutral, is different for different protein species. There is some evidence for sudden changes in evolutionary rate. For example, guinea-pig insulin differs by 17 substitutions from most other mammalian insulins. This suggests either that these substitutions were the result of natural selection, or that in the ancestors of the guinea-pig insulin was removed from some selective requirements, so that in this lineage f was greater.

3. The cost of selection argument

A third argument which has been used in support of the drift hypothesis derives from the concept of the " cost of selection " (or, what amounts to the same thing, of the " genetic load "). Briefly, it is argued that to maintain the

observed polymorphisms and to produce the observed gene substitutions would require a greater number of selective deaths than is compatible with the continued survival of the species. In the case of genetic polymorphism, this argument was proposed by Lewontin and Hubby (1966), and countered by Sved, Reed & Bodmer (1967), by King (1967) and by Milkman (1967). In the case of gene substitution it was proposed by Kimura (1968), and countered by Sved (1968) and by Maynard Smith (1968). The argument is similar in the two contexts, and will be illustrated briefly for the rate of gene substitution.

The concept of the " cost of selection " originated with Haldane (1957). It can be illustrated without algebraic complications for a haploid organism. Suppose that allele a is being replaced by allele A which confers greater fitness. Initially the frequency of A is 0.01; how many generations will it take to establish A as the predominant allele in the population ? Clearly this depends on the fraction of the population which is killed selectively; i.e. the fraction of the population which die because they do not have the optimal genotype. Thus suppose this fraction to be 0.5. Then in successive generations the frequency of A would be 0.01, 0.02, 0.04, 0.08, 0.16, 0.32, 0.64 and 1.00. In other words seven generations of selection, with 50 % selective deaths in each (except the last), are required to establish A as the predominant allele. In Haldane's terms, the " cost " of establishing A is a number of selective deaths equal to $7 \times 0.5 = 3.5$ times the population size.

Using essentially this argument, Haldane estimated that in a diploid population with small selective advantages, the typical " cost " of establishing a new allele would be approximately 30 times the population size. If as he suggested the total number of selective deaths is approximately 10 % per generation, we may expect one gene substitution per 300 generations. More recently, Kimura (1968) has estimated that the actual rate may be 1 gene substitution per two years. His estimate rests on the assumption that the total DNA of the human haploid genome codes non-redundantly for proteins evolving at approximately the same rate as haemoglobin. If there turns out to be a lot of redundancy, or of nonsense, in this genome, then his estimate will be seriously too high. But accepting it for the moment, it is clearly incompatible with Haldane's.

Kimura's conclusion is that most gene substitutions have occurred by drift. An alternative explanation is that the cost of two gene substitutions occurring simultaneously need not be twice the cost of a single substitution, as is tacitly assumed by Haldane. Suppose that two substitutions, $a \rightarrow A$ and $b \rightarrow B$, are occurring simultaneously. If selection acts independently on the two loci, then the cost of increasing the frequency of both alleles from 0.01 to 0.02 is indeed twice the cost of a single allele. If on the other hand selection acts on the two alleles together, eliminating ab genotypes at the expense of aB, Ab and AB, then the cost for two alleles is no greater than the cost for a single allele. It can

in fact be shown that if selection acts on the genotype as a whole, eliminating the least favourable genotypes, then a rate of evolution of 1 gene substitution per generation could easily be achieved without an excessive number of selective deaths.

It is not immediately obvious which assumption is more appropriate. In some contexts selection will act independently at different loci. This will be the case if selection acts on two loci at different times : if for example, one allele influences the resistance of a larval insect to a virus disease and the other the chance that the adult will be eaten by a predator. In such cases Haldane's argument will hold. If on the other hand both loci influence success in competition for a limiting resource (e.g. food), then selection will act on the genotype as a whole, and the " cost of selection " argument fails.

It follows that the " cost of selection " argument cannot by itself establish the drift hypothesis, since it is logically possible for selection to produce the observed rates of evolution, and degree of polymorphism.

4. The rate of evolution and the Extent of Polymorphism

It has already been shown that the rate of substitution of neutral alleles in evolution will be u per generation, where u is the rate of neutral mutation. Thus the rate of non-Darwinian evolution is independant of population size.

It is also possible to derive an expression for the extent of polymorphism expected on the neutral mutation theory.

If I = probability that an individual is homozygous at a locus,
N_e = effective population size and
u = rate of selectively neutral mutation at that locus per generation,
then at equilibrium between mutation and random elimination of alleles (Kimura & Crow, 1964),

$$I = 1/(1 + 4 N_e u). \tag{1}$$

This relation can be tested against observation. For example Kimura assumes that for *Drosophila pseudoobscura* $I \simeq 0.65$, and $u \simeq 1.5 \times 10^{-5}$. This gives $N_e \simeq 9000$, which Kimura suggests is approximately correct for the size of local populations in Drosophila.

This raises the question whether N_e in equation (1) should be taken as the local population size, or the population size of the species as a whole. Thus suppose that a species of total effective number N is divided into r partially isolated populations each of effective number N_e (so that $N = rN_e$), and that there is some migration between populations, so that the probability that an individual born in one population will breed in another is m. Then it turns out (Maynard Smith, 1970) that if $m > ur$, the population is effectively panmictic, and the appropriate population size in equation (1) is N, and not N_e as

assumed by Kimura. Alternatively, if $m < ur$, N_e is the appropriate population size, but in this case hybrids between populations would be heterozygous at almost all loci, and would be polymorphic for different alleles at any given locus. There is evidence (Stone *et al.*, 1968) that for Drosophila on pacific islands, this is not the case. In fact, if $m < ur$ for a period long enough to affect the equilibrium described by equation (1), we should probably describe the different populations as separate species.

It seems therefore that the agreement with equation (1) obtained by Kimura for Drosophila is spurious, and that if neutral mutation is as common as he supposes then a far larger proportion of loci should be heterozygous in wild populations than is in fact the case. Unfortunately, this discrepancy is insufficient to disprove the neutral mutation hypothesis, since it may be that Kimura has seriously overestimated the neutral mutation rate. The difficulty arises because in Drosophila we cannot estimate u from the rate of gene substitution, as might be possible in the case of mammals, since no proteins have been sequenced.

Equation (1) should however provide a way of testing the neutral mutation hypothesis. For example, comparing different species of mammals, one would expect a far higher degree of genetic polymorphism in, for example, a rodent with a continental distribution than in a large ungulate confined to a small geographical range.

There is however a serious difficulty in using equation (1) to test the drift hypothesis. This is that it describes an equilibrium which is reached very slowly. The number of generations required to approach the equilibrium. (e.g. after a disturbance due to a change in N_e) is of the same order of magnitude as N_e; hence no large population will be anywhere near its equilibrium. This is particularly serious when considering a species such as our own, which has increased by a factor of perhaps 10^4 in the last 400 generations. Nevertheless, there may be a way of testing the neutral mutation theory from data on the frequencies of variant proteins in man.

5. Gene frequency distributions in man

The approach to an equilibrium between mutation and elimination is exceedingly slow in a large population. Hence if we want to test the neutral mutation theory by comparing observed and theoretical gene frequency distributions, we need to work out what the gene frequency will be in a population changing in size. It turns out that this may be possible for a population such as our own which has increased by several orders of magnitude in the recent past.

Every gene at a particular locus in the present population is a copy, without mutation, of a gene which arose by mutation some specific number n generations in the past. Let F_n be the probability that a gene, chosen at

random from the present population, arose exactly n generations ago. Let N_0 and N_n be the population size now, and n generations in the past respectively.

The number of new selectively neutral mutations arising exactly n generations ago was $2N_nu$. If, as we must on the drift hypothesis, we ignore selectively advantageous mutation, the expected number of copies now of every gene (including new mutations) present n generations ago is $N_0(1 - P_{n-1})/N_n$, where P_{n-1} is the fraction of genes now which are copies of genes arising by mutation during the last $(n - 1)$ generations. Hence if $P_{n-1} \ll 1$, we have $F_n = 2N_nu \times N_0/N_n \div 2N_0 = u$ *.

Thus the neutral mutation theory leads to the very simple conclusion that, provided $nu \ll 1$, the fraction of genes now present which arose by mutation during the period from r to n generations ago is $(n - r)u$. This can form the basis for a frequency distribution of variant alleles in the present population.

The derivation of frequency distributions is now being undertaken. However, it already seems likely that the frequency of rare haemoglobins in man will not agree well with the frequencies predicted from the neutral mutation theory. Considering the observational data first, Lehmann (quoted in Harris, 1970) in a sample of 10971 individuals from Western Europe found 10 rare electrophoretically separable haemoglobin variants, 3 in the α and 7 in the β chain. Of the β chain variants, 3 are common in other parts of the world and thought to be maintained by heterozygous advantage. Of the remaining 7 variants, 4 occurred once and 3 twice. Hence all had frequencies less than 10^{-4}. Together, these very rare variants contribute a fraction of approximately 2.5×10^{-4} of the genes at each of the α and β loci.

An idea of the distribution to be expected on the neutral mutation theory can be obtained as follows. Let the probability that a new mutation arising n generations ago is still represented in the population be P_n. For all mutations occurring n generations ago the expected number of copies now is N_0/N_n, and hence for each surviving mutation the expected number of copies is N_0/N_nP_n, and hence its expected frequency now is $1/N_nP_n$. For $n > 400$ the human population was probably small - perhaps of the order of 10^5. P_n for $n > 400$ will be of the order of 10^{-2} or less. Hence alleles arising by mutation 400 generations or more in the past will, if still present, have frequencies of 10^{-3} or greater. No such alleles were found in Lehmann's samples (if we discount the alleles thought to be maintained by heterozygous advantage), although if the period from 0 to 400 generations ago has provided a fraction of 2.5×10^{-4} of existing genes, we would expect that the period from 400 to 4000 generations should provide 2.5×10^{-3} of existing genes, or perhaps one or two alleles with frequencies of 10^{-3} at each of the α and β loci. We would also expect some alleles with still higher frequencies originating still further in the past.

* The precise expression for F_n, when P_{n-1} is not very small, is $F_n = u\,e^{-un}$.

The objection might be raised that it is inappropriate to consider electrophoretically separable variants as typical of selectively neutral variants. But in the evolutionary divergence of the α and β chains, there have been 60 substitutions which do not involve a charge change and 24 which do, corresponding roughly to the ratio to be expected if the two classes are equally likely to be selectively neutral.

The simplest explanation of the discrepancy is that selectively neutral mutations do not occur. The very rare variants are slightly disadvantageous, but have not yet had time to be eliminated. Variants arising more than 400 generations ago have been eliminated.

There is however one other possible explanation, consistent with retaining the view that most gene substitutions in evolution are selectively neutral. Suppose that our species has in the comparatively recent past (say 5000 generations, or 100,000 years) gone through a " bottleneck" of very small numbers, leading to genetic homozygosity at many loci, including haemoglobin. On this view, the absence of variant haemoglobins with frequencies of 10^{-3} or greater is explained by inbreeding *. This view would be consistent with the existence of other loci in man in which two or a small number of selectively neutral alleles were common, since the degree of inbreeding might have been sufficient to produce homozygosity at some but not all loci. But it would be inconsistent with the presence of loci at which an appreciable number of alleles have frequencies of 10^{-3} or more.

If there has been such a bottleneck, it should be possible to estimate approximately how small the population must have been. The bottleneck must have occurred in the last 100,000 years, since if it had occurred earlier we would expect to detect more variants which have arisen since that time.

More work, both in collecting data on protein variants and in deriving the theoretical distributions, is needed before drawing any firm conclusions. Perhaps the most interesting possibility is this. If we are obliged by other evidence to accept the neutral mutation theory, then we may be able to obtain an idea of the size of populations ancestral to our own from a study of protein variation in existing populations.

* The idea that close inbreeding would lead to an absence of variant proteins is not inconsistent with the equation $F_n = u\,e^{-un}$. In this equation, F_n is the *mean* fraction of genes in the present population which originated n generations ago. If the human population was closely inbred in the past, the variance of F_n would be very large. Thus imagine a large number of populations each having the same past history of numbers as our own. Then if this past history involved recent close inbreeding, F_n gives (approximately) the probability that a particular population is homozygous for a gene originating n generations ago. But if there has been no recent inbreeding, F_n will approximate to the fraction of each population consisting of such genes.

References

J.B.S. Haldane, *J. Genet.*, **55,** 511 (1957).

H. Harris, *Proc. Roy. Soc. B,* **164,** 298 (1966).

H. Harris, *The principles of human biochemical genetics.* North-Holland Publishing Co. (1970).

M. Kimura, *Nature, Lond.,* **217,** 624 (1968).

M. Kimura, *Proc. Natl. Acad. Sci. U. S.,* **63,** 1181.

M. Kimura & J.F. Crow, *Genetics,* **49,** 725 (1964).

J.L. King, *Genetics,* **55,** 483 (1967).

J.L. King & T.H. Jukes, *Science,* **164,** 788 (1969).

R.C. Lewontin & J.L. Hubby, *Genetics,* **54,** 595 (1966).

J. Maynard Smith, *Nature, Lond.,* **219,** 1114 (1968).

J. Maynard Smith, *Amer. Nat.,* **104,** 231 (1970).

R.D. Milkman, *Genetics,* **55,** 493 (1967).

W.S. Stone, M.F. Wheeler, F.M. Johnson & K. Kojima, *Proc. Natl. Acad. Sci. U.S.,* **59,** 102 (1968).

J.A. Sved, *Amer. Nat.,* **102,** 283 (1968).

J.A. Sved, T.E. Read & W.F. Bodmer, *Genetics,* **55,** 469 (1967).

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

DISCUSSIONS

C. LEVINTHAL : What is the relationship between Lewontin's experiments and the lack of variation you discuss ?

J. MAYNARD SMITH : Lewontin was one of the first population geneticists to exploit protein variation. He is one of the people responsible for my dogmatic statement that populations are a great deal more variable than we had thought. When he first found this, he thought it must be explained by neutral mutation, because the extent of it is too great to be accounted for by natural selection. He said this because he, as we all did at that time, accepted Haldane's cost of selection argument. However I think Lewontin would now accept that this protein variation could be explained either way. Since that time, Lewontin and others have studied protein variation in flies from different local populations of the same species — for example on different islands. If you have a population divided up into small groups and it turns out that each group has the same variant, then this is very difficult to explain on a random basis. It suggests that each group must be stable, otherwise they wouldn't be the same, they would drift apart. A lot of work of this kind is going on at that moment. It's difficult to interpret precisely because of the difficulty that the equilibrium is approached very slowly. Unless you know quite a lot about the past numbers of the animal you are looking at, it's difficult to interpret the data. I have a feeling that our own species, for which we have more idea about past population numbers than for the fruit fly, may be a better bet than fruit flies because of this particular problem.

H. FRÖHLICH : Is there anything known about the physical processes that lead to a mutation ?

D. GLASER : I have been asked to answer this, which I will do by expressing, the prevailing prejudice that, except for occasional hot spots, the probability of a base change is the same everywhere in the chromosome. Whether a base change is reflected in an amino-acid change which can have a phenotypic effect, depends on the considerations that Pr. Margoliash outlined for us. Since leucine is represented by four possible base triplets and methionine by only one, a particular protein rich in leucine, can tolerate more base changes without suffering a change in the amino-acid leucine, then if it is rich in

methionine. In general, I think, the prevailing prejudice is that the probability of a base change is uniform along the DNA, but doesn't produce uniform consequences in the amino-acids. And then it's a very open question which Pr. Margoliash discussed at length, whether amino acid changes produce phenotypic changes. But maybe we can get a professional answer.

D. GLASER : I have a question : is it really necessary that we decide between the two competing theories. Isn't it more likely that there is a range of proteins, some of which can sustain large numbers of neutral mutations, and others of which, because of the sensitivity of their structure to their sequence or because of their critical importance to the organism, cannot tolerate the slightest change in their structure.

E. MARGOLIASH : It is important to point out something different. Lewontin's experiments are also based on electrophoresis of *Drosophila* extracts. Therefore, the estimate of heterogeneity obtained is a minimal one, since not all changes of structure necessarily produce an electrophoretic variation. We may in fact be dealing with much more than he actually observed.

J. MAYNARD SMITH : I think you are right that we don't have to take a black or white choice here. However I should perhaps explain that in the history of population genetics almost religious feelings were aroused by the problem of whether anything could be selectively neutral. There are men, you know, who will go to the length of saying that nothing can possibly be selectively neutral. So for such people, once you admit that any mutation is selectively neutral, you have really sold out to the devil. Now I don't take this view. I think it perfectly possible to say that some proteins, say fibrinopeptide, can vary a lot without mattering, others are very tightly constrained. But in Dr. Margoliash's study of cytochrome c., it will probably turn out that one of two things is true. Either the great majority of the variations are selectively neutral, or the great majority have a selective interpretation which for the moment we cannot see. I'd be surprised if it turned out that 50 % was of one class, and 50 % was of the other. But that's just hunch.

J. BRONOWSKI : I want to ask a question about this last argument which is particularly interesting because it is quite new. Are not the calculations going to be seriously affected by whatever assumptions you make as to when perfect mixing began to take place ? In other words, how many highly homogenous human populations which were separated 500 generations ago have entered the stream that you are measuring now ? And, of course, the analysis also makes assumptions about selective mating which run all through the argument.

J. MAYNARD SMITH : This is the real difficulty. However it may not be as severe a difficulty as at first one might think. I have been looking at the problem of how much interchange of genes there must be between populations in order, for the purpose of these calculations, for them all to be effectively one population. A degree of migration of the order of the mutation rate is all you need. So, unless one supposes total isolation, then I think probably the distortion won't be too big.

D. GLASER : I would like to make a comment about this calculation you say is in progress, and express pessimism that it will be useful. That process is not unlike the problem in bacterial genetics of counting the number of mutants in a population and trying to determine the mutation rate, as was done in the interpretation of the Luria-Delbrück experiment. The mathematics for that problem after many years of approximations has finally been worked out exactly by Benoit Mandelbrot and leads to a statistical formula in which enormous fluctuations are found in the final number of mutations. The fluctuations are so large that it is extremely difficult to give any kind of estimate about the mutation rate which one wants to determinate.

J. MAYNARD SMITH : I was worried by this as well. Of course you can calculate the means; it is calculating their variances which is difficult. But there is this difference. If we wanted to know how many mutations we expected to find in the existing human population which originated exactly 500 generations ago, it is fairly easy to estimate the mean. But its variance might be so big that it would mean nothing. This is essentially the problem in the Luria-Delbrück situation. We do not have this problem because I'm not interested in how many mutations originated exactly 500 generations ago. What I'm interested in is how many originated between 500 and say, 5 000 generations ago. Then I think the variances are going to shrink fairly fast. However the variance might remain large if the human population has been through a bottleneck of small numbers of the order of 100.

D. GLASER : In Mandelbrot's analysis, the statistical distributions are of the Pareto type, $P(x) \propto x^{-\alpha}$, $0 < \alpha \leqslant 2$, and do not possess mean values, but only percentiles.

G.M. EDELMAN : In your last example, wouldn't recombination play a big role if there were selection for heterozygosis in the population. I seem to remember that Wills and Lewontin have talked about this possibility. Wouldn't it tend to increase the amount of variation and possibly account for these results ?

J. MAYNARD SMITH : Yes indeed, but only if selection is involved, not if the mutations are neutral. The beauty of the neutral mutation theory is the mathe-

matics are much easier, so that we can work out the consequences of the neutral theory with adequate precision. We can then compare the predictions of the neutral theory with observation, and if that doesn't fit, then we know there's some selection, and that is the basic point we want to decide. If there is selection we shall have to start worrying about the point you raised about recombination, but on the whole population genetic calculations become dreadfully difficult as soon as you remember that genes are on chromosomes and are linked.

P.O. LÖWDIN : I will try to comment on Pr. Frölich's question concerning the physical reasons for mutations. The problem is not to give a series of possible mutation mechanisms, but to explain why mutations are so exceedingly rare in each one of them. If you look at the Watson-Crick model of DNA and the genetic code, it may be a good idea to start out with doubts and try to prove that the model is wrong or physically impossible — which is the starting point we chose in Uppsala in 1962 in our quantum-mechanical studies. The genetic code consists of a pattern of protons and electron-pairs forming hydrogen bonds, and, in each bond, the proton has two classical equilibrium positions corresponding to different information storage :

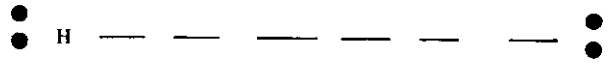


Fig. A

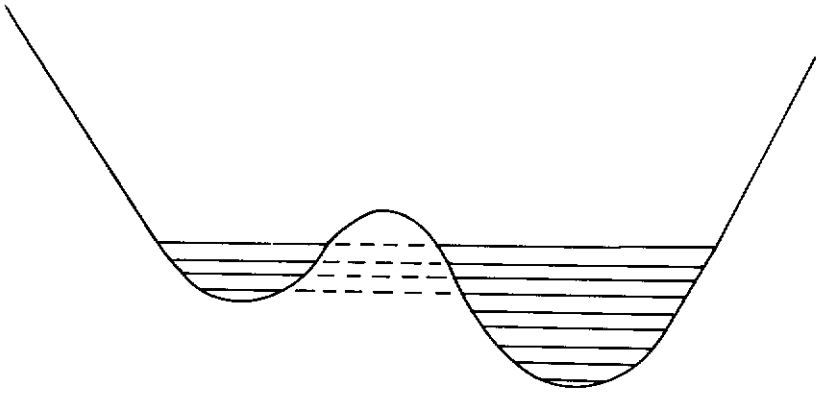


Fig. B

In quantum mechanics, the wave packet associated with the proton may " tunnel " between the two classical equilibrium positions, and the occupation of the tunneling levels depends on the temperature. At body temperature (310 °K), the tunnel effect corresponds to an " error " of 10^{-10} and 10^{-11} for the GC- and AT-base pairs, respectively, and these errors are associated with spontaneous mutation probabilities per base pair and generation. The

fact that these figures are so small indicate an essential stability of the genetic code in the Watson-Crick model; the corresponding experimental figures are estimated to lie between 10^{-8} and 10^{-12}. It is interesting that the AT-pair seems to be more stable than the GC-pair. If the nucleic acid is exposed to weak ultraviolet radiation, there are electron transfers in the base pair which may enhance the probability for a proton transfer by a factor 10^6.

References :

Per-Olov Löwdin, " Quantum Genetics " (*International Science and Technology*, May 1963).

Per-Olov Löwdin, " Quantum Genetics and the Aperiodic Solid. Some aspects on the Biological Problems of Heredity, Mutations, Aging, and Tumors in View of the Quantum Theory of the DNA Molecule " (*Adv. Quant. Chem.*, **2**, 213 (1965)).

Per-Olov Löwdin, " Some Properties of the Hydrogen Bonds in Biochemistry with Particular Reference to the Stability of the Genetic Code " (*Pontificiae Academiae Scientiarum Scripta Varia* 31, « Semaine d'Etude sur les forces moléculaires » (1966)).

S. BENNETT : With reference to the question by Professor Bronowski, I wish to express my lack of understanding at the choice of 500 generations as fitting some significant point in human evolution. If I judge this correctly, this would be about the end of the Pleistocene and somewhere in the neighborhood of 15000 years ago. At that time there were many isolated human populations in many parts of North America and the Old World. To postulate some single genetic pool, or even a small bottleneck of some kind, strikes me as unreasonable. For this reason, I don't understand the merits of choosing this kind of a basis for the argument you present. It seems to me a much older, much longer series of generations would be necessary, if one is to draw on our present knowledge of human evolution.

J. MAYNARD SMITH : The choice is not based on our knowledge of human evolution in this case. The point is that comparatively recent mutations, let us say those occuring in the last 200 or 300 generations, would have very low frequencies in the present population. So that the known observation that there are a number of variants of haemoglobin with a very low frequency of 10^{-4} or less is consistant with the idea that these have originated in the last 300 generations. But if a mutation originated more than about 500 generations ago, and if such a mutation is still present in the population now, then it would probably have a frequency of more than 10^{-4}, and such variants have not been found.

K. MENDELSSOHN : May I just ask the chairman one thing : is there really a difference between the thermal mutation rate and mistakes in arrangement.

Are mistakes in arrangement necessarily part of the thermal or statistical variation under kT ?

D. GLASER : I'm sure that if you did a calculation you would probably identify several terms that contribute to the total rate. One term would be thermal dislocations in complete double helix structures; another would be thermal damage to a temporarily open singlestranded structure; a third one would be thermal disturbances in the fitting of the polymerase to the precursor and single-strand structure that it was working on; and probably another one would be a simply a quantum mechanical calculation of barrier penetration probabilites independent of kT. What I had in mind was the last one, when I spoke of quantum mechanical effects which were not thermal. I'm sure all these things would come together in determining the temperature dependence of mutation rates, for instance, but I imagine you have to estimate them separatly.

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

PHYSICAL AND CHEMICAL PROCESSES LEADING TO A MUTATION

S. BRESLER
Leningrad B 164
Physico-technical Institute Academy of Sciences, U.S.S.R.

Mutation and selection are generally regarded as the main driving force of evolution. We have now a good classification of mutations : nonsense, missense, frame shift, deletions and chromosome aberrations are the main kinds, covering most of the events. A more complicated problem is how mutations are generated, what is the mechanism of the premutational act, how they become fixed and replicated in the progeny ? I shall concentrate on two different types of mutations — the spontaneous and induced ones. Both were studied in our laboratory using bacteria as a genetic model. What are spontaneous mutations ? Their probability is on the order of 10^{-7} - 10^{-9} for a specific gene per generation of cells. If we take for the number of acting genes 10^2 (the overall number of genes per bacterium is on the order of 10^3, but only a part of them is derepressed), so the number of mutations in any active gene would be 10^{-5} - 10^{-7} per generation. If we take 10^3 nucleotide pairs as an average dimension of a cistron it gives us the probability of replication mistake on the order of 10^{-8} - 10^{-10} per generation. We will take 10^{-9} as a mean. If this is an intrinsic probability of erroneous replication by DNA-polymerase during DNA synthesis we must consider a rate constant of the reaction of noncomplementary DNA replication. For bacteria the rate constant of regular DNA replication is on the order of 10^4 nucleotide pairs/second. Hence, for the rate constant of erroneous noncomplementary replication we obtain $10^{-9} \times 10^4$ nucleotide per/sec. = 1 nucleotide per/day. We see that if we could stop the reaction of regular DNA replication but keep the cell alive and the DNA-polymerase active, we would obtain one wrong nucletide addition per day. This would be a premutational event. If the cell would grow afterwards and replicate its DNA in a normal way a mutant would result for every cell in an entire population. This experiment was really performed recently. We took thymine deficient cells of Bac. subtilis and E. coli and cultivated them in a medium with a very low concentration of thymine (0.2 - 0.3 μg/ml). It is a usual practice that bacterial cells die out in such conditions during 2 - 3 hours. The phenomenon of thymineless death was much studied last years and was regarded as a very general one. But we found that on a solid medium i.e. agar plates with a minimal medium supplemented with a very small amount

of thymine or devoid of thymine, thymineless death is practically absent. Cells can be washed off after 50 - 80 hours of thymine starvation and the cell count is constant. When a small thymine concentration is given, the cells even grow and divide. The surface of agar, where at first 10^4 - 10^5 cells were plated (on a Petri dish) becomes covered by microclones visible in the microscope. The cells are washed off afterwards and cultivated in a medium with 20 μg/ml of thymine and then tried for mutants. In conditions of thymine starvation an overall mutagenesis in the whole cell population develops. It starts after 20 hours, when the growth of the cell population comes to a stop. During 2 - 3 additional days the per cent of auxotrophic mutants increases and comes to 70 - 80 % of the whole cell population. Practically most of the cells are mutants and many of them multiple mutants. It is easy to observe the auxotrophic cells with deficiencies for aminoacids, nucleotides, vitamines. They do not grow on minimal agar or grow at a small rate (if the mutants are leaky). We identified some 10 - 12 of the mutants in different loci. We studied also revertants to prototrophy in tryp locus and drug resistant mutants (especially streptomycine resistant).

The frequency of any particular mutation was some 10^3 times higher in our conditions than in the control. (For instance the number of streptomycine resistant cells increased without selection till 10^{-2} % of the cell population, with less than 10^{-5} % in the control, growing without thymine starvation). Of course some of the mutations are lethal and this gives a small decrease of the cell population (some 30 - 40 %). It is obvious that we deal here with a phenomenon of complete mutagenesis in an entire cell population, mainly without death of the cells involved. We think that the only possible explanation is erroneous DNA replication. The scale of time for the occurence of mutations by replication is just what we estimated earlier. The frequency of erroneous replications is a function not only of DNA structure and the properties of nucleotides. It depends also on the enzyme specificity. This factor was studied by Yanofsky and others and special strains with mutated DNA-polymerase revealed increased probabilities of replication mistakes. This is obviously the explanation of special mutator genes- i.e. genetic loci, which imply to the organism and increased mutability and genetic instability.

Now we turn our attention to induced mutations, generated by some chemical substances or by UV-irradiation. We studied them by means of in vitro reaction with isolated and purified DNA with subsequent biological control of the DNA changes by means of transformation of bacteria (Bac. subtilis particularly). Of course we do not take here into consideration the mutations caused by aberrant DNA replication like those considered above. We can mention as example of the latter the increased mutability of bacteria with-5-bromouracil incorporated into their genome instead of thymine, probably because of tautomeric changes in the molecule, and also some specific muta-

genic poisons acting on DNA during replication (for instance acridine dyes). We shall concentrate on factors inducing stable chemical modifications in DNA and will consider the complicated sequence of events leading to a mutation by this mechanism. The chemical modification of DNA effected in vitro is the first event. Then we observe the uptake of DNA by competent cells, its integration into the genome and as a final result - new hereditory properties. All the agents studied by us (UV-light, nitrous acid, dimethylsulfate, nitrosonitromethylguanidine, hydroxylamine, hydrazine) cause mutations when acting on DNA in vitro. But with a probability 300 - 10,000times higher they causes nonspecific damages, which prevent the integration of damaged DNA into the chromosome. It was shown by us that the uptake of DNA by cells is decreased insignificantly by DNA modification. The crucial stage is apparently the synapsis of DNA with the recipient cell chromosome. Here the nonspecific hits are acting and their lack of specificity is obvious because the action of different inactivating agents is strictly additive. This allowed us to formulate a simple phenomenologic theory of DNA inactivation which was confirmed both in our laboratory and in many others. The inactivating hits are not integrated into the genome, but there occur, as mentioned earlier also some chemical events which are integrated and lead to mutations. Their probability is much inferior to that of inactivating hits. What are the premutational events in this case ? We could study them in case of UV-irradiation by means of the photoreactivating enzyme. The latter is well known to repair only pyrimidine dimers. We isolated this enzyme from yeast and acted upon UV-inactivated DNA before using it in transformation. We found that both inactivating damages and premutational damages are repaired at the same rate, i.e. are chemically equivalent. What is the difference then between the specific mutational damages and the nonspecific inactivating hits ? Probably this is a question of secondary DNA structure. The integrity of a straight DNA strand is important for its recombination ability. In particular crosslinks between both DNA strands and also dimerization with the formation of loops in one strand may be responsible for inactivating hits. If so, then only in case of dimerization of neighbouring pyrimidines we obtain a structure which is able to be inserted into the recipient genome. We shall not go into details of chemical interactions. The technique used enables us in principle to study the mechanism of premutational events and of nonspecific hits. The former and the latter are partly repaired in the recipient cell by dark repair enzymes. The use of special strains deficient in repair enzymes is a powerfull tool for the investigation of premutational DNA-modifications and their fate in the cell.

On the whole a study of the mechanism of mutagenesis both in the course of replication and chemically induced, gives us an insight into the details of evolutionary changes which took place on our planet at early times and proceed to act nowadays.

Journée du 3 juillet 1969

Deuxième séance

ALKALI ION CARRIERS : DYNAMICAL BEHAVIOUR

PRESIDENT D. GLASER

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

ALKALI ION CARRIERS: SPECIFICITY, ARCHITECTURE AND MECHANISMS

M. EIGEN and R. WINKLER

Göttingen, Germany

The " carrier " is an entity which facilitates a " dynamical " process, namely the transport of some molecule or ion through a membrane. An understanding of its functional mechanism requires studies of its *dynamical* behaviour.

In understanding the behaviour of alkali ions it may be of help to give a short survey of what is known about the dynamical properties of other ions in solution.

Figure 1 [1, 2] contains a summary of characteristic rates of substitution of water molecules from the inner coordination shell of various metal ions. This step turns out to be rate limiting for metal complex formation reactions. The presentation of such a " periodic table " of rate constants is meaningful

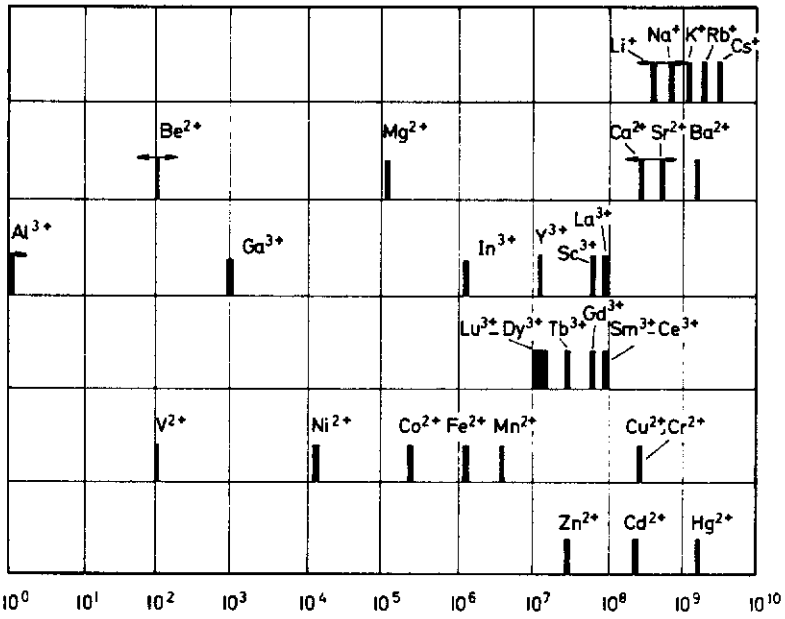


Fig. 1. Characteristic rate constants (sec^{-1}) for substitution of water molecules of the inner coordination sphere of various metal ions.

only, if the rates can be correlated to the properties (i.e. the electronic structure) of the metal ion only, independent of the nature of the incoming ligand. This indeed turned out to be so for most metal ions.

There are two facts to be taken from figure 1 which are of importance for the further discussion of alkali ion carriers: As will be seen this even holds for the formation of complexes where all the water molecules of the hydration sphere have to be substituted by ligands.

1. Any specificity in rates not following a simple metal ion radius dependence, can be found only with the transition metal ions [3], where typically chemical effects govern the substitution behaviour (as an example cf. the non monotonic radius dependence of substitution rates of transition metal ions as depicted in figure 2). In particular the well studied substitution processes

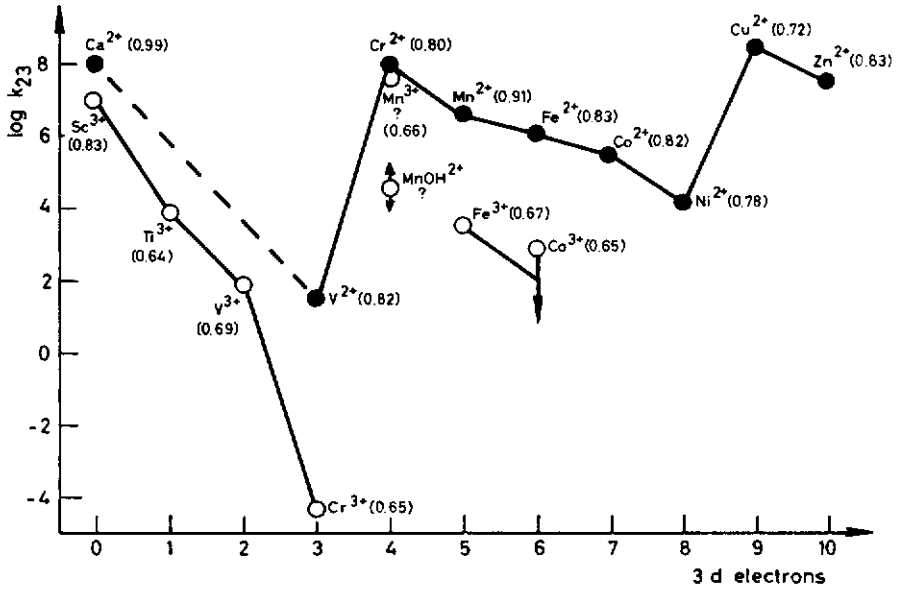


Fig. 2. Radius dependence of substitution rate constant (sec^{-1}) of transition metal ions.

for alkaline earth ions — irrespective of the ligand — show always high rate constants for Ca^{2+} and more than three orders of magnitude lower values for Mg^{2+} [4].

2. The alkali ions all are very fast in substituting single solvent molecules. The time constants are in the neighbourhood of 10^{-9} sec, with a slight radius dependence, i.e. increasing from Cs^{+} to Li^{+} [4].

Especially the second finding may be surprising if it is correlated with the well known solvation behaviour of alkali ions. Figure 3 shows the free energy of solvation for the different alkali ions as a function of the ionic

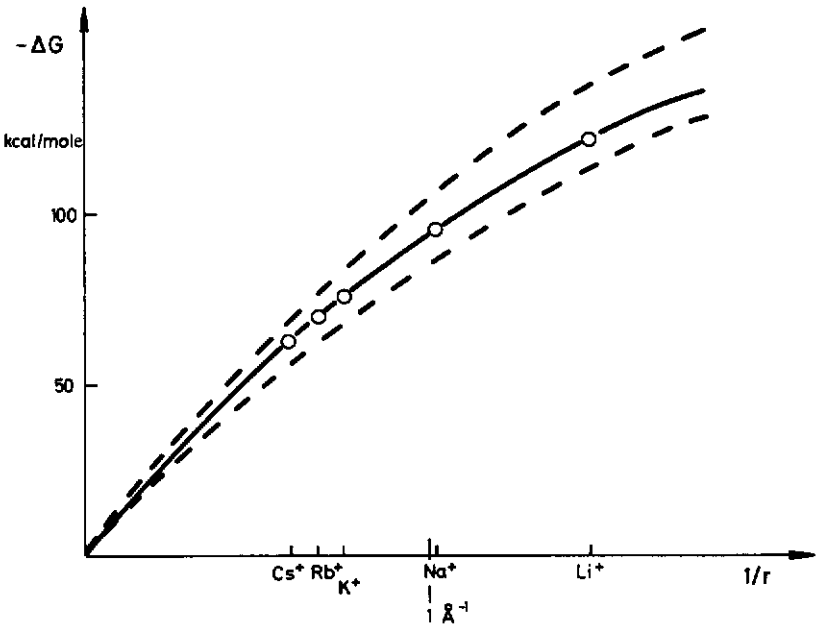


Fig. 3. Free energy of hypothetical ligand binding (dotted lines) and solvation (solid line, experimental values) as function of the reciprocal radius of alkali ions. (The ligand binding curve is related to a fixed ligand concentration.)

radius. The increase of the free energy of solvation with decreasing radius is monotonic as to be expected from any simple electrostatic picture. What is surprising is the relatively small correlation between solvation energy and rate. Despite the high solvation energy values, substitution is an extremely rapid process involving only a few kcal/mole of free energy of activation.

The two dotted curves in figure 3 indicate standardized free energies to be expected for complexes with ligands which are either more or less tightly bound than solvent molecules. No specific carrier behaviour can be deduced from such curves. The free energy of complex formation in the given solvent (here water) would be proportional to the differences between the broken and solid line. For simple ligands only monotonic behaviour, i.e. no maximum at any intermediate metal ion radius could be expected. Thus, the metal ion does not provide any specific property which would explain the specificity of a carrier. The metal ion specificity must be the consequence of a peculiar property of the carrier molecule utilizing the difference of solvation energy for the various alkali ions.

Figure 4 shows how such a behaviour can be envisaged. The two upper curves represent the free energy of binding for two different chelating agents. They consist of multidentate ligands which for complexation have to substitute

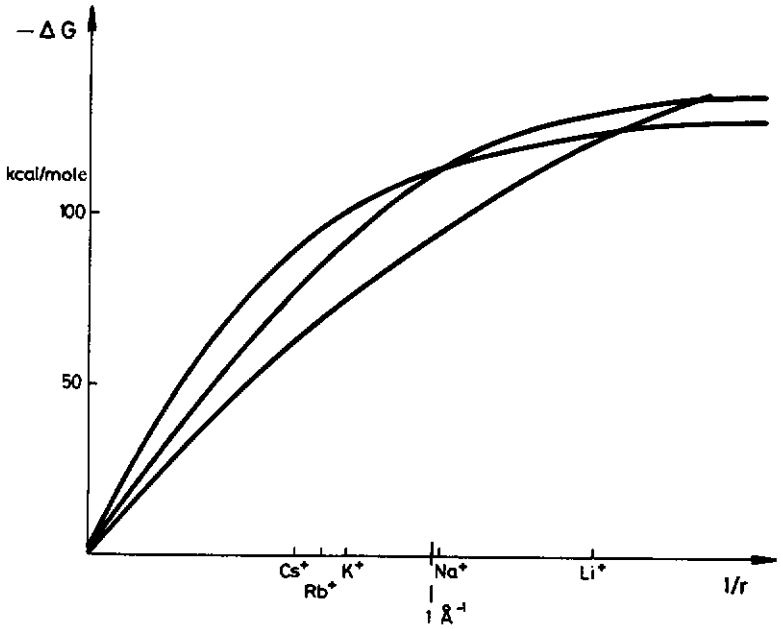
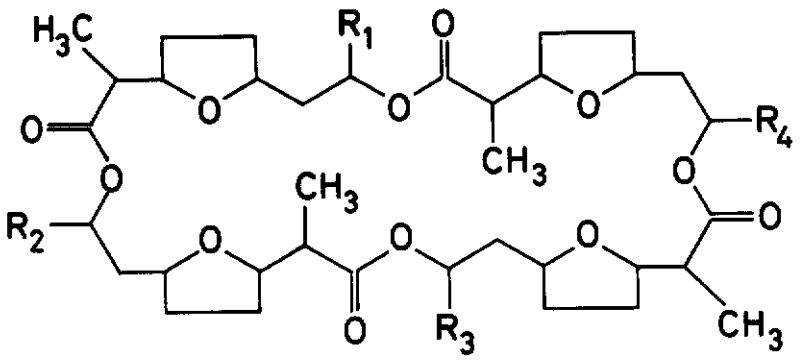


Fig. 4. Free energy of solvation (lower curve) and chelation (upper curve for two hypothetical cases) as function of the reciprocal radius of alkali ions.

the entire solvation sphere of the metal ion thereby enclosing the metal ion into a cavity. At large metal ion radius the free energy of interaction will increase monotonically with decreasing metal ion radius (even relative to the solvent if the ligand is favoured with respect to the solvent molecule as expressed by the higher interaction energy). A decrease of the metal ion radius is accompanied by a shrinkage of the size of the cavity. Due to steric hindrance and repulsion between the binding groups of the carrier the cavity soon will approach a minimum size, which will optimally fit a given metal ion radius. A further decrease of the metal ion size will then not result in any appreciable increase of binding energy (since the binding groups are " frozen " into fixed positions). Thus, the difference between ligand binding and solvation energy will pass through a maximum at a given size of the metal ion.

We have to conclude that binding specificity is a consequence of a specific architecture of the carrier utilizing the difference between free energy of solvation and ligand binding. This difference involves appreciable entropy increments in favour of the chelating ligand. A question remains : How can the process be facilitated quickly, since *total* desolvation is required which — even if one single solvent molecule can be substituted rapidly — may finally lead to quite high free energies of activation. On the other hand — as will be seen below — successful carrier action *depends* on a quick uptake and release of the metal ion.



$R_1 = R_2 = R_3 = R_4 = CH_3$		Nonactin
$R_1 = R_2 = R_3 = CH_3$	$R_4 = C_2H_5$	Monactin
$R_1 = R_3 = CH_3$	$R_2 = R_4 = C_2H_5$	Dinactin
$R_1 = CH_3$	$R_2 = R_3 = R_4 = C_2H_5$	Trinactin

Fig. 5. Chemical structure of macrotetrolides.

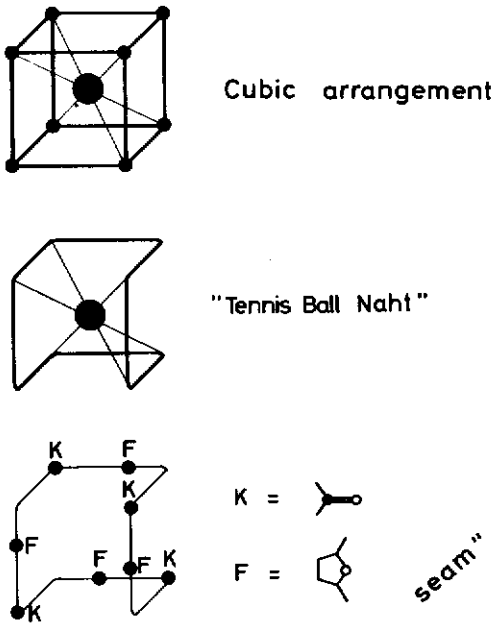


Fig. 6. Simplified spatial structure of metal complexes of the macrotetrolides.

Equilibrium and rate studies have been carried out with monactin and Na^+ as a model system. Monactin belongs to a group of antibiotics which can be isolated from microorganisms such as actinomycetes. The chemical structure as shown in figure 5 was determined by Gerlach and Prelog [5]. X-ray analysis carried out by Dunitz [11] indicates that the metal ion is surrounded by eight oxygens in a quasi-cubic arrangement, whereby the cyclic molecule has a conformation similar to that of a tennis-ball seam (cf. Fig. 6). All polar groups are " inside " the cyclic molecule, whereas the nonpolar groups are situated on the periphery of the complex. In the closed form the macrotetrolide molecule appears to be a hydrophobic sphere. Several problems had to be solved to make such measurements possible.

1. A specific indicator for alkali ions in methanol had to be found. (Since lipophilic carrier molecules are not sufficiently soluble in water, measurements had to be carried out in methanol as solvent.)

2. A relaxation method had to be adjusted to the particular reaction system in methanol. Relaxational amplitudes provide the information about equilibrium parameters such as stability constant, reaction enthalpy etc., whereas relaxation times yield the rate constants for uptake and release of the metal ion.

Ad 1). Murexide [6], the ammonium salt of purpuric acid (Fig. 5) turned out to be an ideal indicator for alkali ions in methanol [7]. A titration curve is shown in figure 6, demonstrating the characteristic colour change upon addition of sodium ions. Figure 7 shows that addition of monactin results in a detectable absorption change. Thus, murexide can be used as an indicator for the reactions of alkali ions with carriers. It is ideal for the present system, because

a) The stability constant of the Na^+-murexide is in a range most suitable for competition with the carrier. (" Half-binding " at concentrations between 10^{-3} and 10^{-4} M.)

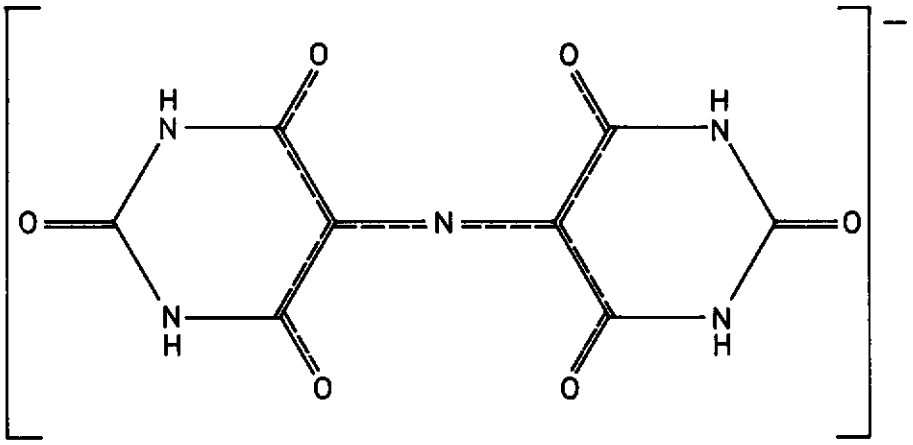


Fig. 7. Chemical structure of the murexide anion.

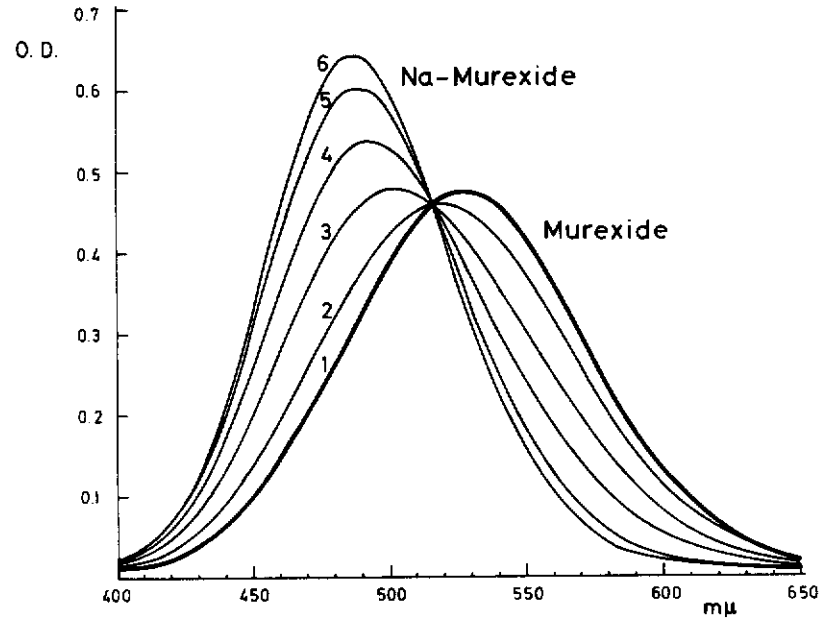


Fig. 8. Spectrophotometric titration of murexide with Na^+ (25 °C; $c_{Mu} = 4.10^{-5}$ M)

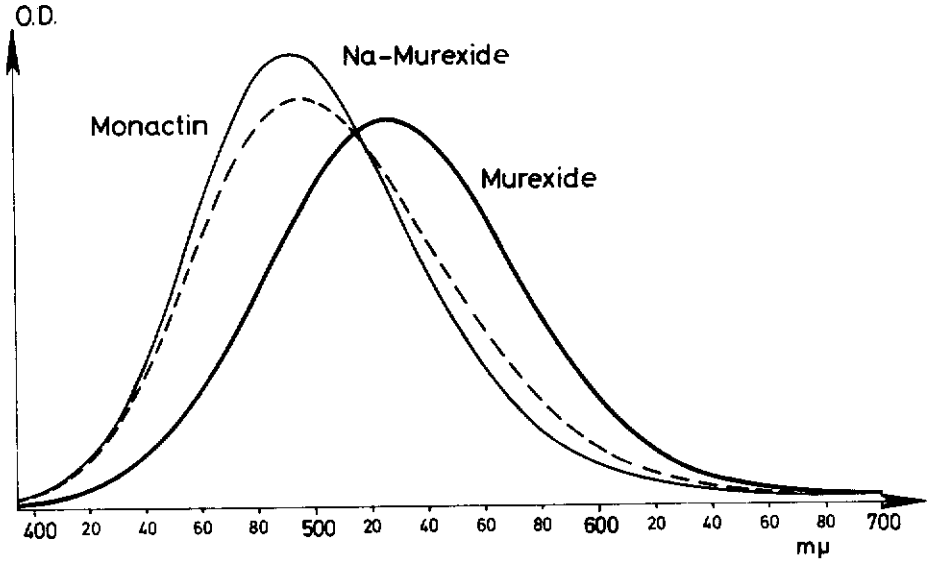


Fig. 9. Decrease of Na^+-murexide absorption upon addition of monactin (dotted line).

b) The indication is very fast. The reaction of murexide with Na^+ is almost diffusion controlled [7] as found from relaxation studies using an electrical travelling wave technique developed by G. Ilgenfritz [8]. The relaxation time is around 100 nanoseconds.

25° C	$k_R\left[M^{-1}\cdot sec^{-1}\right]$	$k_D\left[sec^{-1}\right]$	$K\left[M^{-1}\right]$
Li^+	$5.5\cdot10^9$	$7.7\cdot10^6$	$7.1\cdot10^2$
Na^+	$1.5\cdot10^{10}$	$5.9\cdot10^6$	$2.55\cdot10^3$
K^+	$\sim2\cdot10^{10}$ *)	$\geq10^7$	$1.1\cdot10^3$

*) diffusioncontrolled

Fig. 10. Stability and rate constants of alkali-murexide-complexes in MeOH.

Figure 8 summarizes the properties of murexide. A detailed description of the measurements is found in ref. [7] (cf. also ref. [4 and 9]).

Dissociation constant of the Na^+ monactin complex :	$K = 2.10^{-3}$ [M]
Reaction enthalpy :	$\Delta H \approx -6.0$ [kcal/mole]
Complex formation rate constant :	$k_R = 3.10^8$ [M^{-1} . sec^{-1}]
Life-time of complex :	$1/k_D = 1.5$ [μ sec]

Ad 2). A special technique utilizing differences of temperature jump amplitudes was worked out [7], which allowed a simultaneous determination of stability constants and reaction enthalpies for the metal ion carrier complex. The technique is quite precise and of special advantage in the study of biochemical reactions involving substances which tend to denature. It is described in more detail elsewhere [9].

Using this technique in combination with other relaxation techniques (E-field pulse [8], sound-absorption [10]) for the determination of relaxation times the following results for the system sodium-monactin were obtained.

The most surprising result is the high value for the rate constant of complex formation. A detailed analysis [7] shows that each solvent molecule of the Na^+ solvation shell is substituted within less than 10^{-9} sec. in agreement with the data shown in fig. 1 for single ligand substitution of alkali ions. The mechanism must be a stepwise substitution process, in which each solvent molecule in the coordination shell of the alkali ion is stepwise replaced by the polar groups of the carrier, keeping the total coordination number essentially constant. Such a mechanism requires a quite open form of the carrier molecule in which all the polar groups are easily accessible to the solvated alkali ion. Thus, the uptake of the alkali ion is accompanied by a conformation change of the carrier from an open ring to a closed sphere with the cavity containing the metal ion inside, well shielded by hydrophobic groups. The twisted tennis ball seam structure allows the existence of the two alternate conformations, depending on the compensation of the negative charge of the polar groups by the metal ion.

Four rules for the " design of the carrier " follow from these mechanistic studies :

1. The carrier molecule should possess electrophilic groups which are able to compete with the solvent molecules for metal ion binding. These groups should be located inside an otherwise lipophilic structure, which easily dissolves in membranes.

2. As many solvent molecules of the inner coordination sphere as possible should be replaced by the coordinating sites of the carrier molecule. For two ions of different size the reference state may then involve as much as the total difference of free energy of solvation.

3. The ligand should form a cavity adapted to the size of the metal ion. " Optimal fit " is related to an arrangement where the difference of the free energies of ligand binding and solvation is maximal. It often coincides almost with " fittest " geometrical arrangement. Cavity formation involves ligand - ligand repulsion as well as steric fixation of the chelate.

4. The carrier molecule should possess sufficient flexibility in order to allow for a stepwise substitution of the solvent molecules. Otherwise — i.e. if complete or substantial desolvation would be required for the ion to slip into the cavity — the activation barrier would be rather high and the reaction quite low.

Rule 1 fulfills the biological requirement, i.e. to gate the ion through a (lipid) membrane. Rule 2 and 3 take care of a high selectivity whereas rule 4 allows fast loading and unloading of the carrier. There are not too many low molecular weight structures known which would allow a simultaneous observation of all four rules. Almost all of the classical complexing agents are poor in one or the other respect.

The high rate of complex formation is of great significance for the use of such carriers in biological membranes. The carrier can only be selective if the overall transport rate is not limited by the rate of metal ion release. If this were the case the high selectivity reflected by a high binding constant would be compensated by the slower rate of release. (The ratio of rate constants for uptake and release yield the binding constant). For Na^+-monactin the time constant for release lies in the microsecond range, for the more selectively bound K^+ ion it would reach almost the millisecond range. The release times would be appreciably longer if the recombination process would not have a rate constant of $\sim 10^8$ [$M^{-1}.sec^{-1}$]. It turns out that the carrier can play its selective role in alkali ion transport across molecular bilayer membranes only due to its high — almost diffusion controlled — recombination rate.

Literature

[1] M. EIGEN, *Pure Appl. Chem.*, **6**, 97 (1963).

[2] M. Eigen and R.G. Wilkins, in Mechanism of Inorganic Reactions. *Advan. Chem. Soc.*, **49**, 55 (1965).

[3] M. Eigen, *Ber. Bunsenges. phys. Chem.*, **67,** 753-762 (1963).
[4] H. Diebler, M. Eigen, G. Ilgenfritz, G. Mass and R. Winkler, *Pure Appl. Chem.*, **20,** 93 (1969).
[5] H. Gerlach and V. Prelog, *Liebigs. Ann. Chem.*, **669,** 121 (1963).
[6] G. Schwarzenbach and H. Gysling, *Helv. Chim. Acta*, **32,** 1314 (1949).
[7] R. Winkler, *Dissertation*, Göttingen-Wien (1969).
[8] G. Ilgenfritz, *Dissertation*, Göttingen (1966).
[9] R. Winkler and M. Eigen, to be published.
[10] F. Eggers, *Acustica*, **19,** 323 1967/68).
[11] B.T. Kilbourn, J.D. Dunitz, L.A.R. Pioda and W. Simon, *J. Md. Biol.*, **30,** 553 (1967)

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

DISCUSSIONS

E.P. KENNEDY : I think the beautiful exposition of Pr. Eigen reveals the high degree of specificity in the interaction of cations with molecules like valinomycin. The comments Pr. Eigen made about the possibility that proteins rather than small molecules might be involved in systems function *in vivo*, should be seriously considered. As we discussed on Tuesday the systems that have been investigated, even though information about them is fragmentary, do require proteins. Whether these proteins have also prosthetic groups which are similar to ionophoric antibiotics still is not clear. However, studies of ion transport in bacteria, especially in *E. coli* reveal that the cell membranes are not freely permeable under all conditions to ions, such as magnesium or potassium ions. If one takes cells of *E. coli* and loads them with radioactive magnesium, then the equilibration of the internal and external magnesium pools is a function of the metabolic state of the cell; that is if the cell has sources of metabolic energy available, it rapidly equilibrates the external and internal magnesium pool. But if one gives energy poisons to these cells, then the equilibration process itself is prevented. This appeared from our study on magnesium transport and studies carried out by Silver in St. Louis on potassium transport. This would argue that you do not have low molecular weight, freely diffusible lipid molecules in the membrane, since these would equilibrate internal and external pools, without regard to the energy state of the cell. At the very least there has to be a system that responds to the energy state of the cell. There has to be regulation and, of course, this very strongly suggests that the apparatus which links the transport system to metabolism is enzymatic in nature.

D. HODGKIN : When I earlier introduced the discussion of the properties of nonactin as an ion carrier, it was not because I thought this was the usual kind of carrier of ions in membranes, but because nonactin had geometrical properties which were known and could be simulated by other systems, perhaps by two or more molecules coming together. I should be very much interested to know whether, in your other model system, more than one molecule is involved in the ion complex.

Directeur de la publication : M. Maurice Marois
Imprimé en France
Imprimerie Louis-Jean — 05 - Gap
Dépôt légal n° 399 - 1971