COMPARISON OF POLYPEPTIDE SEQUENCES

THOMAS H. JUKES
UNIVERSITY OF CALIFORNIA, BERKELEY

1. Introduction

Evolution is able to take place because the inherited information in living organisms is subject to change. Most of this information is constant from generation to generation, but changes that take place in a very small part of it are necessary for evolution. Hereditary information is stored in the base sequences of DNA. Molecular evolution is, therefore, concerned with alterations in DNA. Changes in DNA are of three general types: *point mutations*, in which one base pair is substituted for another; *recombination*, in which two double strands of DNA cross over, so that there is a lengthening of one double strand and a shortening of the other; and *duplication*, proliferation, or multiplication, in which the amount of DNA per cell becomes increased. Any change of these three types is evolutionary if it is predominantly adopted into the genome of a species.

Part of the DNA, probably only a small fraction, consists of base sequences that provide information for the synthesis of proteins, and it is through proteins that most of the phenotypic expression of hereditary information takes place. This paper is concerned with evolutionary changes in proteins, as detected by analyzing proteins and by comparing the proteins of different species of living organisms with each other. Differences between two similar proteins, such as the hemoglobins obtained from man and monkey, will represent evolutionary changes in DNA that have entered into the makeup of the species. The last-mentioned two types of change, recombination and duplication, are infrequently adopted; for example, there seem to have been only six evolutionary events of recombination in the hemoglobins in about 500 million years. In contrast, base replacements take place incessantly. These produce the "evolutionary clock" that ticks slowly in proteins, independent of speciation, generation time or gene duplication, but *not* independently of the type of protein as defined by its essential function. We shall illustrate this by comparing the amino acid sequences of polypeptide chains, with the use of the amino acid code. A family of proteins, the globins, that occurs in all vertebrates, will be used as illustrative examples.

Supported by a grant from the National Aeronautics and Space Administration NgR 05-003-020 "The Chemistry of Living Systems" to the University of California.

Most of the studies on polypeptide chains have been carried out by protein chemists, and hence, biochemical viewpoints are quite prominent in this field. In considering amino acid substitutions during evolution, emphasis has been placed on the amino acids themselves, rather than on the nucleic acid code which is responsible for the changes. One result of this is a great underestimate of the number of evolutionary events during the divergence of two homologous proteins. Another result is that amino acids have been placed by several authors into so-called similar groups.

I shall include all and any base substitutions that are incorporated in the DNA of a species in the definition of *evolutionary events*, regardless of where the substitution is, or whether it produces a change in an amino acid residue. For example, there is an evolutionary substitution of CUA for UUA, both of which code for leucine, in the coat protein cistrons of the RNA viruses f2 and R17. Such neutral changes can affect the results produced by subsequent mutations.

Perhaps the first demonstration that amino acid replacements were an expression of evolutionary differentiation was the discovery by Sanger and co-workers [3], [13], [29] of differences between insulin A chains as shown in Table I. The remainders of the chains are identical.

TABLE I

Species	Residue No. 8 9 10
Cow	-Ala-Ser-Val-
Horse	-Thr-Gly-Ile-
Sheep	-Ala-Gly-Val-
Pig	-Thr-Ser-Ile-

By using the genetic code (Table II), the preceding series can be rewritten as alterations of a single strand of DNA (see Table III), where A, C, G, T are the four bases present in DNA, N is any one of them, and Y is C or T.

The molecular steps in this evolutionary process would seem to be simplicity itself, but this is not so. Consider the comparisons of the β chains of hemoglobins portrayed in Table IV.

The difference between the second and third lines is the result of a point mutation, as originally identified by Ingram [16]. There are more than 100 different identified point mutations in human hemoglobin, and each corresponds to a single base change in the genetic code (Table II). For this and other reasons, we conclude that the change from Glu to Val in hemoglobin S was caused by a point mutation from A to T in the codon for glutamic acid at site 6. But we do not know the steps in the evolutionary change from Gly to Pro at site 5 in lines

TABLE II

THE AMINO ACID CODE

Y = U or C; R = A or G; N = U, C, A or G.

Phenylalanine (Phe)	UUY	Serine (Ser)	UCN	Tyrosine (Tyr)	UAY	Cysteine (Cys)	UGY
Leucine (Leu)	UUR			Chain Termination (CT)	UAR	Chain Termination (CT)	UGA
						Tryptophan (Trp)	UGG
Leucine (Leu)	CUN	Proline (Pro)	CCN	Histidine (His)	CAY	Arginine (Arg)	CGN
				Glutamine (Gln)	CAR		
Isoleucine (Ile)	AUY, AUA	Threonine (Thr)	ACN	Asparagine (Asn)	AAY	Serine (Ser)	AGY
Methionine (Met)	AUG			Lysine (Lys)	AAR	Arginine (Arg)	AGR
Valine (Val)	GUN	Alanine (Ala)	GCN	Aspartic acid (Asp)	GAY	Glycine (Gly)	GGN
				Glutamic acid (Glu)	GAR		

TABLE III

Species	Base Sequence
Cow	-G-C-N-A-G-Y-G-T-Y-
Horse	-A-C-N-G-G-Y-A-T-Y-
Sheep	-G-C-N-G-G-Y-G-T-Y-
Pig	-A-C-N-A-G-Y-A-T-Y-

1 and 2. The interpretation, as Gatlin [12] says of matters related to the information in DNA, depends on the context. Indeed, the question of the extent to which two amino acids are interchangeable in a sequence depends on the context more than on their chemical structures. There may have been two changes, such as Gly to Ala and Ala to Pro, or three changes, such as Gly to Ala, Ala to Ser, and Ser to Pro, in the millions of years during which human beings and horses differentiated by separate pathways from a common ancestor.

TABLE IV

Horse hemoglobin		-Ser-Gly-Glu-Glu-Lys-
Human hemoglobin	A	-Thr-Pro-Glu-Glu-Lys-
Human hemoglobin	S	-Thr-Pro-Val-Glu-Lys-

If we next compare the same sequence in the β hemoglobin chain with the corresponding region in myoglobin, the difference is so great that no homology is perceptible (see Table V). Only by comparing the entire sequences and the tertiary structures of these two proteins can their homology be established. Their evolutionary separation took place probably more than 500 million years ago. We wonder if the glutamic acid at site 7 remained constant during this period, or if it perhaps changed, let us say, from Glu to Asp and back to Glu again, because Asp is found at this site in two other hemoglobin chains.

TABLE V

	4	5	6	7	8
β chain	-Thr-	Pro-	Glu-	Glu-	Lys-
Myoglobin	-Ser-	Asp-	Gly-	Glu-	Trp-

There is a way of measuring the rate of change in DNA molecules directly (Kohne [23]). It is only a rough measurement, and does not identify the changes. I will describe the principle of the method very simply. When two strands of DNA are separated by heating, they reassociate on cooling—the bases seek out their original partners. Most of the DNA in higher organisms consists of what is termed "unique sequence DNA," meaning that its sequence of bases is not repeated elsewhere in the chromosomes. If this fraction of DNA from human cells is heated, and mixed with the corresponding fraction from heated monkey DNA, the single strands of human DNA will form double strands with some of their corresponding partners from monkey DNA, during slow cooling. However, the partnership is imperfect because the human and monkey DNA, which have both descended from a common ancestor, have accumulated base changes during evolution so that the sequences are somewhat different from each other. The hybrid DNA has a lower melting point than the nonhybrid DNA, corresponding to a lowering of about 1°C for each 1.5 per cent of unmatched base sites. This quantitation may be used as an evolutionary clock. If we deduce from the fossil record that the lines of descent leading to human beings and green monkeys separated about 30 million years ago (Wilson and Sarich [31]), then there have been about 18×10^{-10} substitutions per nucleotide pair per year. Most proteins diverge during evolution at a rate lower than this. We deduce from comparing homologous proteins that Darwinian natural selection acts to reject some of the changes in them caused by differentiation of DNA. If we could separate and examine the region of DNA that codes for proteins, we should expect to find that it had been held more constant than other regions in DNA (except for the regions that are transcribed into specialized molecules of RNA, such as transfer RNA and ribosomal RNA). The majority of the unique sequence DNA apparently differentiates sufficiently fast [23], that, evidently, most of its base changes are neutral and are incorporated into the genomes of the species by genetic drift. These changes originate mainly as point mutations. When these occur in structural genes, many of them are translated into amino acid changes in proteins. The genetic drift of such changes occurs at different rates in different proteins, depending on how much the protein can change without losing its function. Cytochromes change slowly, hemoglobins fairly fast, and antibodies somewhat faster.

To study such changes adequately, it is necessary to use the genetic code, for changes in amino acid sequences are responses to changes in DNA. The genetic code is the alphabet of molecular evolution. Students of this subject should be so familiar with the code that when they see ACG on an automobile license plate they should immediately think of threonine.

2. Methods for comparing polypeptides

I shall now describe two of the models used in comparing homologous polypeptide sequences. The first of these ignores the code, and uses a first order decay

curve to predict the number of evolutionary amino acid replacements that have occurred at each site. The model has been used by Margoliash and Smith [25], Zuckerkandl and Pauling [32], Dayhoff [5], Kimura [21], and King and Jukes [22]. It says that one or more hits (that is, base substitutions) at any one codon site produce an end result of substitution of the amino acid at that site. The changes are enumerated without reference to the genetic code. Kimura [21] used this procedure to estimate the number of evolutionary events (that is, amino acid replacements) per unit of time as follows. Let x be the average number of unchanged amino acid sites per site in the comparison of two polypeptide chains. For the human: carp α Hb comparison, $x = 0.514$. Then, from the Poisson distribution, $m = 0.665$. This is the number of evolutionary events per site. The evolutionary interval between human and carp is 2×375 million years $= 0.075 \times 10^{10}$ years, that is, twice the length of time since each has separated from a common ancestor. Therefore, evolutionary events in α Hb have occurred at the rate of $0.665 \div (0.075 \times 10^{10}) = 8.9 \times 10^{-10}$ per codon per year since the separation of the carp and human lines. This model makes the following assumptions:

(i) evolutionary amino acid replacements are distributed at random along the polypeptide chains;

(ii) invariable sites, such as the histidines at positions 63 and 121, are included as being subject to evolutionary change;

(iii) no replacements are considered to be revertants to an original amino acid residue.

All three assumptions are erroneous.

I have used a second model, based on minimum base differences per codon (MBDC) (Jukes [18]). The procedure makes no assumptions. It consists of counting all the visible base changes in the codons when two polypeptide chains are compared. Visible base changes are defined as the minimum number of base changes needed for an interchange between two amino acids which are compared. Using this procedure, the minimum number of evolutionary events since the separation of the human and carp lines is as given in Table VI. This corresponds

TABLE VI

Sites compared	Minimum base differences				Average per codon
	0	1	2	3	
140	72	41	27	0	0.68

to 9.1×10^{-10} per codon per year, which is quite close to the value calculated by Kimura's procedure. However, it is an obvious underestimate, for it is calculated in terms of *minimum* base replacements per codon. Assuming randomness, the value of 0.68 implies that there has been the Poisson distribution of events in

TABLE VII

0	1	2	3	4	5
71	48	16	4	0.6	0.1

140 loci shown in Table VII; that is, only 21 codons should have undergone more than one base replacement. The actual number is 27, which indicates that the number of unhit sites *plus* sites hit only once is not less than 113. Referring again to the Poisson distribution, a minimum of 0.805 evolutionary events per codon separate the carp and human α Hb lines, or 10.7×10^{-10} per codon per year. Yet this is still a gross underestimate. Evolutionary events must be measured in terms of base replacements in DNA. As shown by Holmquist [14], the number of primary mutagenic events may be as much as 3.5 times as great as the expected number of amino acid differences between two homologous polypeptide chains. To illustrate this matter, let us examine the effect of a series of base replacements on an individual codon. Single base substitutions in the threonine codon ACU, lead to the example in Table VIII.

TABLE VIII

UCU	CCU	GCU	AUU	AAU	AGU	ACC	ACA	ACG
Ser	Pro	Ala	Ile	Asn	Ser	Thr	Thr	Thr

The three changes to other Thr codons are not expressed, and are, therefore, not included in the next step. This examines the next series of changes, each attributable to two substitutions in the original codon (see Table IX). By

TABLE IX

UCU	to	CCU	ACU	GCU	UUU	UAU	UGU	UCC	UCA	UCG
CCU	to	UCU	ACU	GCU	CUU	CAU	CGU	CCC	CCA	CCG
GCU	to	UCU	CCU	ACU	GUU	GAU	GGU	GCC	GCA	GCG
AUU	to	UUU	CUU	GUU	ACU	AAU	AGU	AUC	AUA	AUG
AAU	to	UAU	CAU	GAU	AUU	ACU	AGU	AAC	AAA	AAG
AGU	to	UGU	CGU	GGU	AUU	ACU	AAU	AGC	AGA	AGG

following this procedure for each of the other three Thr codons the totals for changes of two bases per Thr codon are as shown in Table X, together with examples from other amino acids.

TABLE X

RESULTS OF TWO-BASE SUBSTITUTIONS IN CODONS FOR THREONINE GLYCINE, ASPARAGINE, AND METHIONINE

The results are derived from single base changes, that is, from the codons for amino acids that are mutations resulting from one base substitutions in the original sets of codons for threonine, glycine, asparagine, and methionine, respectively.

(a) Revertants to original amino acid.
(b) Two-base changes that simulate single base changes.
(c) Recognizable two-base changes.

From threonine (ACN)			From glycine (GGN)		
(a)	(b)	(c)	(a)	(b)	(c)
Thr-24	Ala- 20	Asp- 4	Gly-23	Arg-28	His- 4
	Pro- 20	Cys- 4		Ser- 17	Gln- 4
	Ser- 30	Gln- 4		Trp- 4	Pro- 8
	Lys- 10	Glu- 4		Cys- 8	Leu- 11
	Asn- 10	Gly- 8		Glu- 10	Met- 2
	Arg- 18	His- 4		Asp- 10	Ile- 6
	Ile- 15	Leu- 12		Ala- 20	Thr- 8
	Met- 5	Phe- 4		Val- 20	Lys- 4
		Trp- 2			Asn- 4
		Tyr- 4			Phe- 4
		Val- 8			Tyr- 4
Totals 24	128	58	23	117	59
From asparagine (AAY)			From methionine (AUG)		
(a)	(b)	(c)	(a)	(b)	(c)
Asn-20	Asp- 6	Ala- 4	Met-9	Leu- 12	Ala- 2
	Tyr- 6	Arg- 12		Val- 8	Asn- 4
	His- 6	Cys- 4		Lys- 4	Gln- 2
	Ser- 10	Gln- 8		Thr- 8	Glu- 2
	Thr- 14	Glu- 8		Arg- 6	Gly- 2
	Ile- 12	Gly- 4		Ile- 6	Phe- 4
	Lys- 4	Leu- 2			Pro- 2
		Met- 2			Ser- 6
		Phe- 2			Trp- 2
		Pro- 4			
		Val- 4			
Totals 20	58	54	9	44	26

The percentage distribution of these changes is seen in Table XI. Extended calculations along these lines show that, roughly speaking, 19 recognizable two-base changes correspond to 54 codons that have been hit more than once. The remaining 35 codons either show no change in amino acid assignment, or show changes that could correspond to a single base replacement.

I shall not pursue the intricacies of the comparison as regards two-base changes any further. Holmquist [14] has presented some mathematical treatment of this question in an accompanying paper. However, the existence of the above relationships must be considered when homologous proteins are compared. For example, the β hemoglobin chains of human and horse differ at 25 loci, and six

TABLE XI

	Thr	Gly	Asn	Met
(a) Revertants	11	11	13	11
(b) Two-base changes simulating single base changes	61	59	43	56
(c) Recognizable two-base changes	28	30	44	33

of the differences correspond to changes that are recognizable as resulting from a minimum of two base replacements in a codon. We shall, for convenience, term these "two-base changes." On the basis of randomization, the number of such changes should be about 2.3. Therefore, there must be some channeling of evolutionary changes in a nonrandom manner, as noted in the cytochromes *c* by Fitch and Margoliash [8] and developed further by Fitch and Markowitz [9]. This channeling seems to take place without regard to much of the often emphasized chemical similarity between certain pairs of amino acids, such as the presence of hydrophobic side chains. The recognizable two-base interchanges in the comparison of human and horse β chains are between Pro and Gly; Thr and Leu; Gly and His; Ala and His; Cys and Val; Pro and Glu. The members of these pairs of amino acids do not show a close chemical resemblance to each other. Evidently protein molecules pay very little attention to the rules laid down by some biochemists as to which amino acids are suitable for evolutionary replacements.

Evolutionary changes in proteins, in addition to replacements of one amino acid by another, include duplication, lengthening and shortening. All these are reflections of corresponding changes in DNA. Lengthening and shortening are produced by genetic recombination. Shortening results in gaps that are found by comparing two polypeptide sequences; for example, there is a gap of five amino acids in the Gun Hill mutant of the β chain of human hemoglobin (Bradley, Wohl, and Rieder [2]). Note (Table XII) that the gap may be written in any of three ways, because of the repetition of Leu-His. When a deletion becomes incorporated in one member of a homologous pair of proteins, the subsequent differentiation of the polypeptide chains can make it very difficult to locate the gap, or even to justify the assumption that it has occurred. The matter was discussed and analyzed by Cantor [4]. The problem is particularly acute if the

TABLE XII

Normal		-Ser-Glu-Leu-His-Cys-Asp-Lys-Leu-His-Val-Asp-Pro-
Mutant		-Ser-Glu-Leu-His -Val-Asp-Pro-
	or	-Ser-Glu-Leu -His-Val-Asp-Pro-
	or	-Ser-Glu- -Leu-His-Val-Asp-Pro-

gap is near the end of a polypeptide sequence, where the homology on one side of the gap can almost never be good enough to justify the gap. As Cantor points out, "When two relatively dissimilar sequences are compared, it is almost always possible to improve the homology by the insertion of one or more gaps $\cdots$. In general, adding gaps increases the number of comparisons by about L^{N_g} where N_g is the number of gaps and L is the number of amino acid pairs so that the extra homology introduced by a gap must more than compensate for the increased number of comparisons."

The measurement of such homology is facilitated by using the genetic code. Cantor's method for searching for gaps is to introduce gaps into computerized comparisons of sequences of amino acids. These comparisons depend on minimum base differences per codon (Table XIII), and include the use of a computer program as described by Fitch [10]. Cantor's procedure, which is illustrated in Figure 1, enables the gap to be placed at a point where the difference between the two polypeptide chains is minimized.

The locating of gaps is helped by the availability of a large number of sequences from a family of homologous proteins. The hemoglobins are unusually useful for this procedure, because gene duplications are superimposed on speciation. This makes it possible to locate a gap of five consecutive residues at positions 53 to 57 in all the α chains by comparing them with the myoglobin and β chains, none of which contain the gap. A large gap is in lamprey hemoglobin between the G and H helical regions (Fujiki and Braunitzer [11], Li and Riggs [24]). Carp and lamprey hemoglobins contain extra Met residues, inserted between positions 70 and 71, and 85 and 86, respectively. The gaps and insertions in the globins are summarized in Table XIV.

Figure 2, modified from the proposal by Ingram [17], shows how the different globin chains have evolved by duplication and differentiation from a common ancestral gene. Below the diagram are arrows indicating the order at which various lines of descent have branched off as compared with the order in which the events of duplication occurred. The time relationship shows that the lamprey should have only α type chains, because its line of descent separated before the origin of β chains, that is, before the appearance of tetrameric hemoglobins, which are not present in lampreys. The carp should have both α and β chains, but not the γ chain, because on the line of descent leading to human γ this duplication apparently originated at about the time of the marsupial-eutherian divergence. This was estimated as about 130 million years ago (Air, Thompson, Richardson, and Sharman [1]). A comparison of the human β and γ chains shows about the same divergence as that between the human and kangaroo β chains (Table XV). This would place the γ chain as having appeared at about the same time as the placental circulatory system, and after the origin of amphibians and birds. The embryonic frog and chicken hemoglobins (which have not been "sequenced") should, therefore, have diverged prior to the separation of the mammalian β and γ chains.

TABLE XIII

Minimum Base Differences per Codon for the 61 Codons for the Amino Acids

The assignments of the codon are in Table I.

		GCN Ala	CGN Arg	AAY Asn	GAY Asp	UGY Cys	CAR Gln	GAR Glu	GGN Gly	CAY His	AUY Ile	AUA Ile	CUN Leu	AAR Lys	AUG Met	UUY Phe	CCN Pro	UCN Ser	ACN Thr	UGG Trp	UAY Tyr	GUN Val	AGY Ser	AGR Arg	UUR Leu
GCN	Ala	0	2	2	1	2	2	1	1	2	2	2	2	2	2	2	1	1	1	2	2	1	2	2	2
CGN	Arg	2	0	2	2	1	1	2	1	1	2	2	1	2	2	2	1	2	2	1	2	2	1	1	2
AAY	Asn	2	2	0	1	2	2	2	2	1	1	2	2	1	2	2	2	2	1	3	1	2	1	2	3
GAY	Asp	1	2	1	0	2	2	1	1	1	2	3	2	2	3	2	2	2	2	3	1	1	2	3	3
UGY	Cys	2	1	2	2	0	3	3	1	2	2	3	2	3	3	1	2	1	2	1	1	2	1	2	2
CAR	Gln	2	1	2	2	3	0	1	2	1	3	2	1	1	2	3	1	2	2	2	2	2	3	2	2
GAR	Glu	1	2	2	1	3	1	0	1	2	3	2	2	1	2	3	2	2	2	2	2	1	3	2	2
GGN	Gly	1	1	2	1	1	2	1	0	2	2	2	2	2	2	2	2	2	2	1	2	1	1	1	2
CAY	His	2	1	1	1	2	1	2	2	0	2	3	1	2	3	2	1	2	2	3	1	2	2	3	3
AUY	Ile	2	2	1	2	2	3	3	2	2	0	1	1	2	1	1	2	2	1	3	2	1	1	2	2
AUA	Ile	2	2	2	3	3	2	2	2	3	1	0	1	1	1	2	2	2	1	3	3	1	2	1	1
CUN	Leu	2	1	2	2	2	1	2	2	1	1	1	0	2	1	1	1	2	2	2	2	1	2	2	1
AAR	Lys	2	2	1	2	3	1	1	2	2	2	1	2	0	1	3	2	2	1	2	2	2	2	1	2
AUG	Met	2	2	2	3	3	2	2	2	3	1	1	1	1	0	2	2	2	1	2	3	1	2	1	1
UUY	Phe	2	2	2	2	1	3	3	2	2	1	2	1	3	2	0	2	1	2	2	1	1	2	3	1
CCN	Pro	1	1	2	2	2	1	2	2	1	2	2	1	2	2	2	0	1	1	2	2	2	2	2	2
UCN	Ser	1	2	2	2	1	2	2	2	2	2	2	2	2	2	1	1	0	1	1	1	2	2	2	1
ACN	Thr	1	2	1	2	2	2	2	2	2	1	1	2	1	1	2	1	1	0	2	2	2	1	1	2
UGG	Trp	2	1	3	3	1	2	2	1	3	3	3	2	2	2	2	2	1	2	0	2	2	2	1	1
UAY	Tyr	2	2	1	1	1	2	2	2	1	2	3	2	2	3	1	2	1	2	2	0	2	2	3	2
GUN	Val	1	2	2	1	2	2	1	1	2	1	1	1	2	1	1	2	2	2	2	2	0	2	2	1
AGY	Ser	2	1	1	2	1	3	3	1	2	1	2	2	2	2	2	2	2	1	2	2	2	0	1	3
AGR	Arg	2	1	2	3	2	2	2	1	3	2	1	2	1	1	3	2	2	1	1	3	2	1	0	2
UUR	Leu	2	2	3	3	2	2	2	2	3	2	1	1	2	1	1	2	1	2	1	2	1	3	2	0

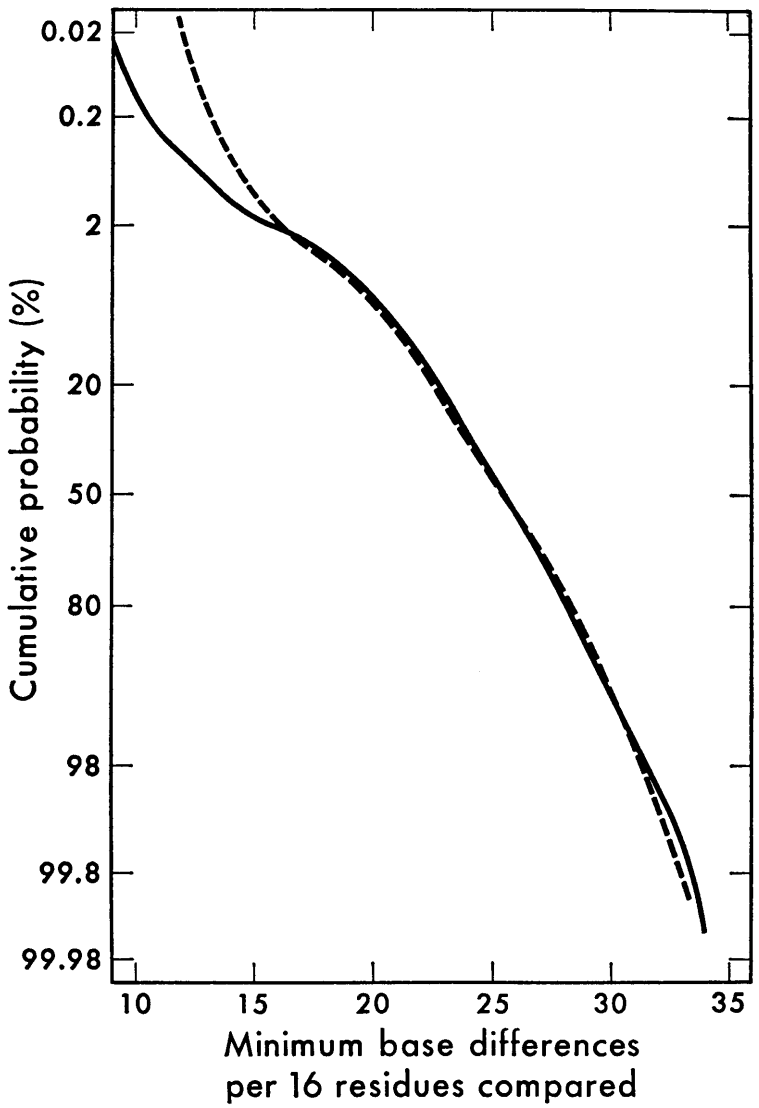


FIGURE 1

Computer comparisons of the first 40 amino acids of α and β human hemoglobin chains including a gap of two residues on the β chain (solid line); and on the α chain (broken line). The shift in the β distribution is evidence for a gap of two on the β chain. (From Jukes and Cantor [19].)

TABLE XIV

EVOLUTIONARY CHANGES IN LENGTH IN GLOBIN CHAINS

Globin	Nature of change (Residues numbered to include gaps)	Time of occurrence
β, γ and δ hemoglobins	Addition of one residue at NH_2-terminus Deletion of 19 and 20	Before origin of birds
α hemoglobin of terrestrial species	Deletion of 48	Subsequent to divergence of bony fishes
α hemoglobin of bony fishes	Deletion of 64. Addition of 70.5	Subsequent to divergence of bony fishes
α hemoglobin	Deletion of 53–57	Subsequent to divergence of lamprey
Lamprey hemoglobin	Addition of 9 residues to N-terminus. Addition of 78.5. Deletion of residues 110–119 and 148	Subsequent to divergence of lamprey
Frog β hemoglobin	Deletion of 1–6	After divergence of amphibians

The duplication of the α chain to give rise to the α and β chains enables comparisons to be made between the α and β chains of any two groups of animals that appeared subsequently. Such comparisons are not possible with the cytochromes *c* because these have evolved throughout the vertebrate line without

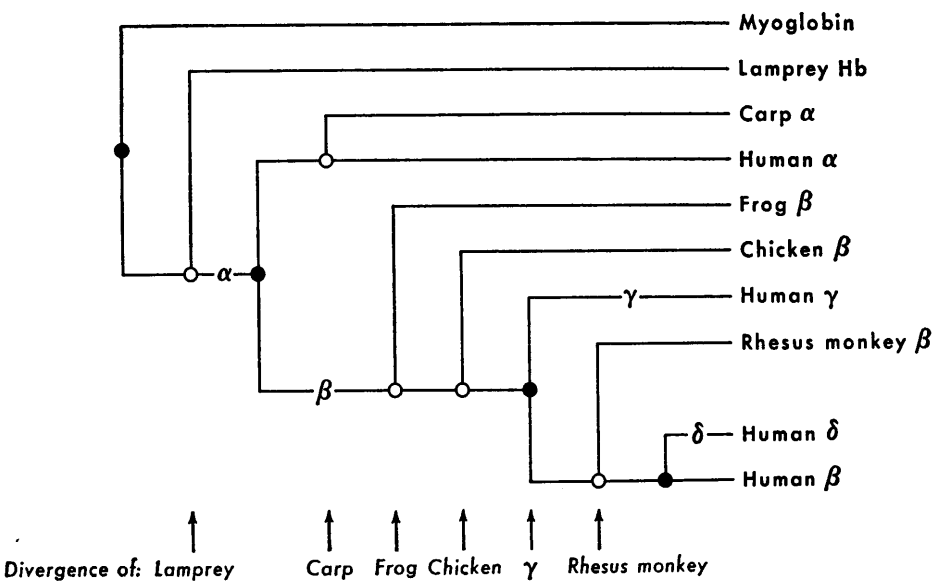


FIGURE 2

Divergence of globin chains by gene duplication (dots) and speciation (circles).

TABLE XV

COMPARISONS OF VARIOUS HEMOGLOBIN (Hb) AND MYOGLOBIN (Mb) AMINO ACID SEQUENCES IN TERMS OF AMINO ACID DIFFERENCES (AAD) AND MINIMUM BASE DIFFERENCES PER CODON (MBDC)

	Sites compared	Minimum base differences per codon 0	1	2	3	Per site AAD	MBDC
(i) α chains							
Human with Horse	141	123	14	4	0	0.13	0.16
Bovine	141	124	14	3	0	0.12	0.14
Mouse	141	125	13	3	0	0.11	0.13
Rabbit	141	116	22	3	0	0.18	0.20
Horse with Bovine	141	125	11	5	0	0.16	0.15
Mouse	141	119	17	5	0	0.16	0.19
Rabbit	141	116	19	6	0	0.18	0.22
Bovine with Mouse	141	122	17	2	0	0.13	0.15
Rabbit	141	116	22	3	0	0.18	0.20
Mouse with Rabbit	141	118	19	4	0	0.16	0.19
Carp with Human	140	72	41	27	0	0.49	0.68
Rhesus	140	72	39	29	0	0.49	0.69
Horse	140	72	39	29	0	0.49	0.69
Bovine	140	74	38	28	0	0.47	0.67
Rabbit	140	70	46	24	0	0.50	0.67
Mouse	140	72	46	22	0	0.49	0.64
Chicken	140	68	44	28	0	0.51	0.71
(ii) Mammalian α and β chains							
Human α with Human β	139	64	53	22	0	0.54	0.70
Rhesus β	139	64	50	25	0	0.54	0.72
Horse β	139	64	52	23	0	0.54	0.70
Rabbit β	139	61	55	23	0	0.56	0.73
Rabbit α with Human β	139	75	53	21	0	0.53	0.68
(iii) Lamprey with some Hb chains							
Lamprey with Human α	130	45	57	28	0	0.65	0.87
Rabbit α	130	42	54	34	0	0.68	0.94
Carp α	130	39	53	38	0	0.70	0.99
Chicken α	130	36	59	34	1	0.72	1.00
(iv) Human β with Horse β	146	122	18	6	0	0.16	0.20
Human γ	146	107	29	10	0	0.27	0.34
Kangaroo β	146	108	24	14	0	0.26	0.36
Frog β	140	80	39	20	1	0.43	0.59
(v) Myoglobins (Mb) with Hemoglobins							
Horse Mb with Horse α	141	36	55	50	0	0.74	1.10
Human α	141	39	54	47	1	0.72	1.04
Mouse α	141	36	62	43	0	0.75	1.05
Carp α	141	35	66	39	1	0.75	1.05
Horse β	145	35	63	45	2	0.76	1.10
Human β	145	38	56	49	2	0.74	1.10
Frog β	140	36	52	51	1	0.74	1.13
Lamprey	135	33	60	41	1	0.75	1.09

indications of duplication. Comparisons of groups of α hemoglobin chains with groups of β chains are shown in Table XV. There are fewer two base changes in the α:β comparison than in the carp α: mammalian α comparison, although the total difference is greater in the α:β comparison. This points to different evolutionary patterns of amino acid replacements in the two cases. A possible explanation is that the α:β divergence is modified by the fact that the two chains have to fit each other to form an α:β dimer (Perutz [27]), while the comparisons of α chains with each other may show a pattern depending on the fact that mammalian α chains must form dimers with both β and γ chains.

A comparison of horse myoglobin with various hemoglobin chains shows a surprising constancy of difference, clearly indicating divergent rather than convergent evolution. Lamprey hemoglobin has apparently differentiated from the common ancestor of the globins (Figure 2) at the same rate as the other hemoglobins. The "primitive" anatomy and physiology of the lamprey have led to its being regarded as a "living fossil." Nevertheless, lamprey hemoglobin seems to have changed during evolution to about the same extent as the other hemoglobins.

3. Silent changes in codons

Changes in the third base of codons for amino acids will in most cases not produce a change in the amino acid assignment, so that in such cases they are, therefore, selectively neutral. Evidence has recently come to light that such "silent" changes take place during evolution. Three small RNA viruses that grow in *E. coli* are f2, MS2, and R17. The coat proteins of these contain a sequence of 129 amino acids. R17 and MS2 coat proteins are identical and f2 differs only by having leucine instead of methionine at residue 88. Recently the base sequences of large portions of these viruses have been determined, including those portions that code for the coat proteins. There are six "silent" base differences in cistrons for the identical coat proteins of MS2 and R17. Other "silent" changes include UCC(f2) *versus* UCU(R17); CUA(f2) *versus* UUA(R17); and GGC(f2) *versus* GGU(R17).

These are evidently neutral changes that have occurred by mutation and have been incorporated by genetic drift. Subsequent changes in such codons may cause marked differences in two lines, as shown by the hypothetical example in Figure 3.

Doubtless many other silent changes in DNA will soon come to light. There could well be between 10 and 15 such silent differences in the genes for the identical molecules of human and chimpanzee hemoglobins, as calculated from the rate of evolutionary separation of the hemoglobins and the five million years estimated for the chimpanzee-human divergence by Wilson and Sarich [31].

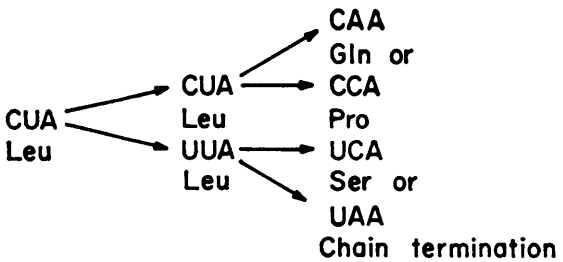


FIGURE 3

4. The disappearance of homology during evolution

The evolutionary differentiation of two polypeptide chains can eventually proceed to a point where their homology can no longer be perceived. This is a gradual process, and, as a result, some sequences are at the borderline of disappearing homology. The process of divergence may be followed phylogenetically by comparing the complete sequences of various globins, but it is perhaps more interesting to observe it in internal duplications in proteins, because the homology can then be followed to its point of disappearance.

The best example of internal repetition is human haptoglobin α-2 (Table XVI). In this protein segment 2, of 59 residues, 13 to 71, is repeated exactly at residues 72 to 130 except for one (single base) interchange, Glu to Lys. This may be diagrammed as shown in Figure 4. Moreover, the same segment occurs once, that is, without repetition, in haptoglobin α-1, residues 13 to 71 (see Figure 5). Evidently this repetition in haptoglobin α-2 is a recent evolutionary event, and, obviously, partial duplication of the gene has taken place. This example establishes the existence of the phenomenon of internal duplication in a protein molecule.

The next example is clostridial ferredoxin, for example, that of *Clostridium butyricum* (Table XVI). Using a similar diagram, the duplication is represented in Figure 6. Segments 1 and 1′ are, respectively, 25 and 26 amino acid residues in length; 1′ has an additional amino acid at the fifth residue. After adjusting for this, segments 1 and 1′ are identical in 14 out of 25 residues at corresponding sites. The existence of the internal duplication is clear, but the event took place so long ago in evolutionary time that considerable differentiation has taken place, and four of the eleven differing residues are separated by two-base codon changes.

The third example (see Figure 7) is the A and B chains of insulin. These occur in the single molecule of proinsulin. The B chain comes first. Residues 1 to 20 repeat at 65 to 84; if three gaps are inserted in each of the two regions (Table XVI) nine out of the 17 pairs of homologous residues are identical and four of the eight differing residues are separated by two or more base changes per codon.

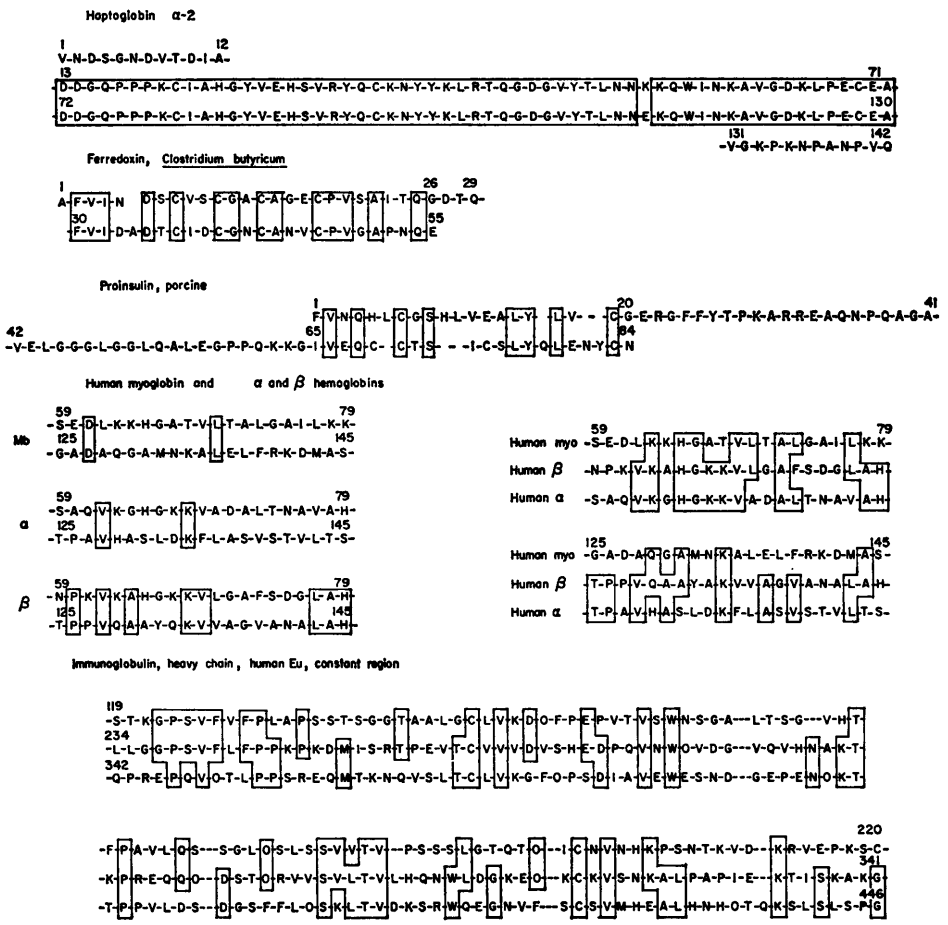
The next two examples illustrate the eventual disappearance of homology.

TABLE XVI

EXAMPLES OF INTERNAL REPETITION

Abbreviations:

A	Ala	G	Gly	M	Met	S	Ser
C	Cys	H	His	N	Asn	T	Thr
D	Asp	I	Ile	P	Pro	V	Val
E	Glu	K	Lys	Q	Gln	W	Trp
F	Phe	L	Leu	R	Arg	Y	Tyr
				—	Gap	Z = Glx	



The first of these considers two cytochromes *c*: those of *Neurospora* and tuna. Residues 20 to 34 in *Neurospora* are a repetition of 5 to 19. There has been much evolutionary differentiation but the repetition is still distinct (Table XVII). When the same sequences are compared in tuna, there is no sign of homology

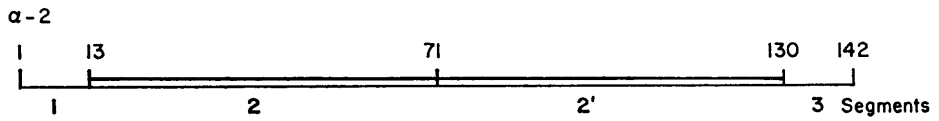


FIGURE 4

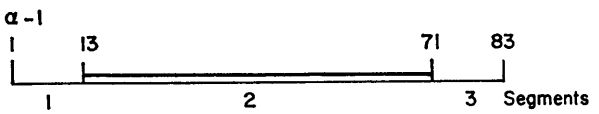


FIGURE 5

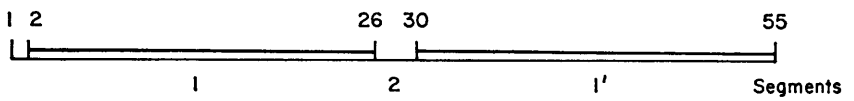


FIGURE 6

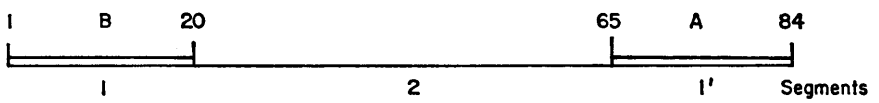


FIGURE 7

(Table XVIII). However, when the segments of *Neurospora* and tuna are compared directly with each other, the homology of 5 to 19 *Neurospora* with 5 to 19 tuna, is obvious, and so are the homologies of the two 20 to 34 regions, as seen in Table XIX. The tuna internal comparison (5 to 19 *versus* 20 to 34) is a clear case of complete evolutionary erosion of homology.

The same phenomenon has occurred in hemoglobin (Table XVI) except that the homologous segments do not occur consecutively, but are separated by

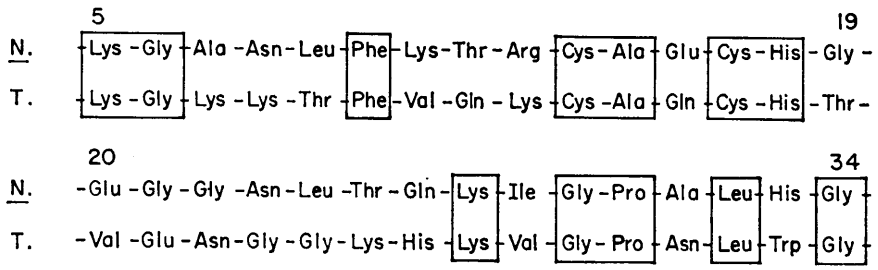
TABLE XVII

5																		19
-Lys	Gly	Ala	Asn	Leu	Phe	Lys	Thr	Arg	Cys	Ala	Glu	Cys	His	Gly				
20																		
-Glu	Gly	Gly	Asn	Leu	Thr	Gln	Lys	Ile	Gly	Pro	Ala	Leu	His	Gly				

TABLE XVIII

5																		19
-Lys	Gly	Lys	Lys	Thr	Phe	Val	Gln	Lys	Cys	Ala	Gln	Cys	His	Thr				
20																		34
-Val	Glu	Asn	Gly	Gly	Lys	His	Lys	Val	Gly	Pro	Asn	Leu	Trp	Gly				

TABLE XIX



about 45 residues. Residues 59 to 79 and 125 to 145 in the human β chain are clearly homologous (Fitch [10], Jukes and Cantor [19]) because eight of the 21 pairs of corresponding amino acids are identical. The corresponding regions in horse myoglobin show no significant homology with each other (only two identities in 21 pairs, with an expectation of one identity in 20 random sequences of amino acids).

The same is true of the human α chain. Direct comparisons of corresponding sequences with each other, however (Table XVI), show that regions 59 to 79 are homologous and 125 to 145 are homologous in myoglobin, α Hb, and β Hb. The similarity of the myoglobin sequences to the hemoglobins is, as to be expected, less than the similarity of the α and β hemoglobin regions to each other. For some reason, the internal homology in the β chain, but not in the other two, has been retained. This is unexpected in view of the fact that the β chains of various species of vertebrates are diverging from each other more rapidly than is the case with either the myoglobins or the α chains that have been examined.

To facilitate further examination and because of the interest in this particular comparison, the internal homology in the hemoglobins is shown in more detail in Table XX. The prototype is derived by comparing all the codons for the amino acid residues in a vertical row, and assigning the predominant nucleotides to the codon for the amino acid in the prototype. The polypeptide sequences for residues 59 to 79 and 125 to 145 are both from helical regions (E and H) in which the hydrophobic side chains point inwards. This would tend to favor homology, whether or not the two regions had a common origin. However, other helical regions such as A, F, and G, which also have hydrophobic, inner directed, side chains do not show homology, in the β and γ chains, with residues 59 to 79 and 125 to 145. Further discussion of this homology in the globins is in Jukes and Cantor [19].

There are three possible reasons for homology in two polypeptide sequences: it may be by chance, by convergent evolution, or by divergent evolution. Chance homology is examined in terms of probability. The decision between convergent and divergent evolution is aided by probability, molecular structure, and phylogeny. As an example, let us examine the sequences 125–145 in various

TABLE XX

Comparisons of Sequences in Polypeptide Chains of Certain Globins

Mb = human myoglobin; α = α chain of human hemoglobin; β = β chain; γ = γ chain; Lamprey = *Petromyzon* hemoglobin. The numbering system is identical with that in Table XVI. The underlined residues are identical with residues in the chain above them.

Mb	50–58	-Leu-Lys-Ser - -Glu-Asp-Glu-Met-Lys-Ala- (53 = Glu)
	59–79	-Ser -Glu-Asp-Leu-Lys-Lys-His-Gly-Ala-Thr-Val -Leu-Thr-Ala-Leu-Gly-Gly- Ile -Leu-Lys-Lys-
	125–145	-Gly-Ala-Asp-Ala-Gln-Gly-Ala-Met-Asn-Lys-Ala -Leu-Glu-Leu-Phe-Arg-Lys-Asp-Met-Ala -Ser
α	50–58	-Leu-Ser -His — — — — — — Gly-
	59–79	-Ser -Ala-Gln-Val-Lys-Gly-His-Gly-Lys-Lys-Val -Ala-Asp-Ala-Leu-Thr-Asn-Ala -Val -Ala -His-
	125–145	-Thr-Pro-Ala-Val-His-Ala-Ser-Leu-Asp-Lys-Phe-Leu-Ala -Ser -Val-Ser -Thr-Val -Leu-Thr-Ser
β	50–58	-Leu-Ser -Thr- -Pro-Asp-Ala -Val -Met-Gly-
	59–79	-Asn-Pro-Lys-Val-Lys-Ala-His-Gly-Lys-Lys-Val -Leu-Gly-Ala-Phe-Ser -Asp-Gly-Leu-Ala -His-
	125–145	-Thr-Pro-Pro-Val-Gln-Ala-Ala-Tyr-Gln-Lys-Val -Val-Ala -Gly-Val-Ala-Asn-Ala -Leu-Ala -His
γ	50–58	-Leu-Ser -Ser - -Ala-Ser -Ala - Ile -Met-Gly-
	59–79	-Asn-Pro-Gly-Val-Lys-Ala-His-Gly-Lys-Lys-Val -Leu-Thr-Ser -Leu-Gly-Asp-Ala - Ile -Lys-His-
	125–145	-Thr-Pro-Glu-Val-Gln-Ala-Ser-Trp-Gln-Lys-Met-Val-Thr-Gly-Val-Ala-Ser -Ala -Leu-Ser -Ser
Lamprey	50–58	-Leu-Thr-Thr- -Ala-Asp-Gln-Leu-Lys-Lys-
	59–79	-Ser -Ala-Asp-Val-Arg-Trp-His-Ala -Glu-Arg- Ile -Ile -Asn-Ala-Val-Asn-Asp-Ala -Val -Ala -Ser-
	145	— — — — — — Gly-Phe-Glu-Lys-Leu-Ser -Met-Cys-Ile -Ile -Leu-Met-Leu-Arg-Ser
Prototype		-Ser -Pro-Asp-Val-Lys-Ala-His-Gly-Lys-Lys-Val -Leu-Thr-Ala-Val-Ala-Asp-Ala -Leu-Ala -Ser-

human globins (Table XX). There are only two differences between the β and δ chains, at positions 127 and 128 and the homology seems apparent by probability alone. However, there are 8 differences between β and γ, 13 between β and α, and 17 between β and Mb (Table XX). Therefore, in terms of probability, there is no visible homology between these regions of β and Mb. Despite this, the globins are regarded as a family of proteins that have divergently evolved by gene duplication followed by differentiation, and, by this argument, the sequences 125–145 are all homologous. This conclusion is supported by an examination of the tertiary structures of the globins; these structures are homologous. Furthermore, a progressive divergence can be seen when the sequences of the α Hb chains of vertebrates are compared. The amino acid differences between human and *Rhesus* average 0.03 per site; human and horse, 0.13; human and chicken, 0.25; human and carp, 0.49; human and lamprey, 0.65. The human: lamprey difference reaches the edge of probability, yet the homology is accepted because the progressive phylogenetic divergence mirrors the zoological and fossil comparisons. This pattern of thought favors divergent, rather than convergent, evolution as an explanation for the homology in this series. This example is perhaps the most elegant illustration that has come to light of the disappearance of homology as a result of evolutionary erosion.

Repetitive sequences in polypeptide chains must be examined by the same criteria for homology: probability, convergence, and divergence. But in this case, phylogeny is not available for support, and tertiary structure can actually work against us. We are thrown back on probability. The homology between 59–79 and 125–145 in the β Hb chains or between the two halves of clostridial ferredoxin (Table XVI) must be evaluated by probability alone. If two such sequences are in unrelated regions of the tertiary structure, the credibility of the proposal that they are a repetition may be questioned. But it should be remembered that, after such an internal repetition, a protein may evolve towards a more "useful" tertiary structure and, during this evolution, the homology of the repeated regions may change rapidly. Therefore, an existing homology between two separated portions of a single polypeptide chain may indeed be indicative of an evolutionary duplication.

The establishment of cases of convergent evolution, on the other hand, would seem to be entirely dependent upon similarities in tertiary structure, such as recently noted for subtilisin and trypsinogen. In such cases, evidence of strong homology in comparisons of the amino acid sequences would probably override convergent evolution in favor of divergent evolution. A case in point is chymotrypsinogen and trypsinogen, in which the amino acid difference is only 0.44 per site when the appropriate gaps are inserted. The similarity of these two sequences is, on this basis, too great to be attributed to convergent evolution.

Homology may still be detectable in proteins when it is not possible to demonstrate it by a comparison of the amino acid sequence. An example is the similarity of the tertiary structures of *Chironomus* and mammalian hemoglobins, which have similar configurations (Huber, Formanek, and Epp [15]) when compared by

X-ray methods. In such a case it may be difficult to decide whether the homology is the result of convergent as opposed to divergent evolution.

4.1. *Immunoglobulins.* These proteins, the antibodies, contain repetitious sequences of about 110 amino acid residues. The IgG immunoglobulins, for example, consist of two "light" chains, each with two such sequences and two "heavy" chains, each with four such sequences. The N-terminal sequences of both the light and heavy chains are highly variable with respect both to length and sequences. The constant regions are easier to compare with each other. The internal repetition of a constant region in the heavy chain of the Eu (γ type) immunoglobulins is shown in Table XVI. The homology of the three repetitions is clearly visible and the alignment is made easy by the presence of 13 identical amino acids at corresponding sites in all three sequences. These sequences have diverged from each other at a constant rate as estimated by amino acid differences and minimum base differences per codon.

The tertiary structure of immunoglobulins is unknown, but the repeating sequences evidently have similar secondary structures, characterized by disulfide bridges (Putnam [28]). These connect pairs of cysteine residues shown at positions 144 and 200; 259 and 321; 367 and 425 in Table XVI. Most of the residues are needed to preserve the structure and function of the molecule. Some of these residues are common also to the constant regions of the light chain and to the variable regions of light and heavy chains (Jukes and Holmquist [20]).

5. Fibrinopeptides

During blood clotting, which typically takes place outside the circulatory system of animals, a complex series of reactions leads to the conversion of prothrombin to thrombin. This then reacts with fibrinogen, splitting off two fibrinopeptides (A and B) and producing fibrin. The two fibrinopeptides markedly differ from each other in the same animal. Each of them shows great evolutionary diversity, so that closely related species have different fibrinopeptides A and the same is true for fibrinopeptide B (Table XXI).

The enzyme thrombin, however, is quite nonspecific, and the thrombins of all mammalian species so far examined will cross-react with any mammalian fibrinogen (Doolittle [6]). The point of attack is the first Arg-Gly linkage in each of the A and B chains. Evidently the adjoining amino acids have no measurable effect on the affinity of thrombin for the Arg-Gly peptide bond, for none of the A chains in Table XXI have the same amino acid preceding this bond as is found in the corresponding B chain. Thrombin, which is a protease, is so nonspecific that it is standardized by its reaction with a methyl ester of tosyl-blocked arginine. Nonspecificity is typical of proteases.

The marked differences between the A and B chains in the same animal may be a clue to the reason for nonspecificity of thrombin, because the same thrombin molecule releases two different fibrinopeptides A and B. Obviously, if the enzyme were more specific, for example, tailored to fit a special sequence of amino acids,

TABLE XXI

SEQUENCES OF FIBRINOPEPTIDES

Abbreviations as in Table XVI. Z = Glx, a derivative of glutamic acid. The asterisks indicate invariant residues.

(1) 39 Fibrinopeptides A

																*				*
Human	–	–	–	A	D	S	G	–	E	G	D	F	L	A	E	G	G	G	V	R
Others	A	D	G	S	N	T	K	D	S	S	E	S	I	T	A		A	T	G	
	E	T	D	T	K	P	A	T	G	–	T		E	D	D			A	I	
				P	E	V	E		K		S			S	G			V		
				G	–	A	D		T					H				H		
				V		G	V		D					G						
				–		–	Q		–					P						
							P							E						
							–													

(2) 30 Fibrinopeptides B

																						*
Human	–	–	–	–	–	–	–	–	Z	G	V	N	D	N	E	E	G	F	F	S	A	R
Others	Z	H	A	L	D	D	D	D	E	E	E	E	E	R	A	K	V	H	L	D	G	
		F	S	D	Y	Y	Y	H	T	V	Q	G	G	G	P	V	L	T	V	G		
		G	P	I		S			–	D	D	D	N	–	V		S	R	P			
		P	L	A						–	–	–	–		D		D	G				
		A	Y	E											I			L				
		S	I	H											T			P				
			T	T											K			I				
			F															V				
																		S				

the evolutionary rate of differentiation of the fibrinopeptides would be much slower than it is, because each change in a fibrinopeptide would have to be accompanied by a corresponding change in the thrombin molecule. An analogy is the slow rate of evolutionary differentiation of cytochrome *c*. The biological function of cytochrome *c* necessitates its interaction with two other proteins, cytochrome c_1 and cytochrome oxidase. It is postulated that this circumstance helps to slow down the rate of evolution of cytochrome *c*.

The amino acid residues at the various fibrinopeptide sites are shown in Table XXI. Comparisons of the sequences have been made by various authors in order to relate differences in the fibrinopeptides to phylogeny and evolutionary rate of change. There are, however, certain peculiarities in the fibrinopeptides that make such comparisons difficult to evaluate. First, the sequences are very short so that not many amino acid residues can be compared, and, second, the chains vary in length, and it is difficult to align them without using arbitrary gaps.

It is obvious, however, that the fibrinopeptides change more rapidly in evolution than any of the polypeptide sequences that have so far been compared, with the possible exception of the calcitonins (Milhaud [26]). It seems likely that most of the changes in the fibrinopeptides are neutral for reasons described above.

6. Discussion

The examples in Table XVI show that internal repetitions are common rather than exceptional in protein evolution. Proteins are large molecules, usually containing more than one hundred amino acid residues. The probability of an existing sequence having emerged by chance from the superastronomical number of possibilities for permuting twenty variables (the different amino acids) into a sequence of more than one hundred units is almost infinitely small. It seems more likely that a short sequence of amino acids having some catalytic activity could be formed during the early stages of the emergence of life. Repeated end to end duplication of such a sequence could perhaps preserve and enhance its activity. Meanwhile, numerous point mutations would occur, and some of these, in nonfunctional regions of the molecule, might favorably influence the development of an advantageous tertiary structure which would be preserved by natural selection.

It is axiomatic that any mutation that changes an amino acid residue simultaneously alters the chemical properties of a protein. We have proposed (King and Jukes [22]) that many of such alterations are selectively neutral. The opposite conclusion is that any alteration of properties has an adaptive effect which enters into the process of natural selection, so that all changes which enter into the makeup of an organism during evolution are incorporated on a Darwinian basis. These two contrary viewpoints, in various guises, have been debated at this Symposium.

In support of the neutral concept, I wish to advance the suggestion that alterations in the properties of an enzyme may often be selectively neutral. It is well known in biochemistry that wide "margins of safety" exist in many physiological processes. The capacity of an enzymatic process in a living organism may far exceed the demands that are ever placed upon it. Söderqvist and Blombäck [30] state that, "A species-specific thrombin has varying clotting-times with different fibrinogens while a species-specific fibrinogen gives fairly constant clotting-times with different thrombins. The results suggest that mammalian thrombins have changed slowly while corresponding fibrinogens have undergone a more pronounced change during mammalian evolution. The differences in clotting-times are as great as ten times and cause no obvious negative adaptive effects."

Blood clotting is, of course, an extremely complex phenomenon involving a multitude of chemical reactions. Blood clotting within the circulatory system can be fatal to an animal. Speed of clotting may, therefore, be disadvantageous; the important feature is that the clot arrest the flow of blood from a ruptured vessel within a reasonably prompt time. Söderqvist and Blombäck conclude that some variation in this time can occur without selective advantage.

I conclude that it is possible that amino acid differences, between two homologous proteins in similar organisms may well be the result of a selectively neutral

divergence. An example might be horse and bovine myoglobins, which differ in 17 of 153 amino acid residues.

It is to be expected that supporters of the neo-Darwinian viewpoint will continue to fit all molecular changes in evolution to the Procrustean bed of panselectionism. Time will not permit the resolution of the argument, for we vanish from the scene much more rapidly than evolutionary changes take place. We can, however, take advantage of modern scientific methods to study the result of evolutionary change at the molecular level. Much new information is available as a result of determinations of the sequences of amino acids in proteins, thus permitting new theoretical insights into molecular evolutionary processes.

7. Summary

7.1. Polypeptide sequences in families of homologous proteins may be used to study molecular evolution. When two such sequences are compared, amino acid differences are usually found at many corresponding sites. These differences are considered to be the result of replacements of nucleotides in DNA during divergent evolution.

7.2. It is possible to interpret the differences in terms of the amino acid code, and such interpretations show that the rate of nucleotide replacement is more rapid than is usually estimated on the basis of amino acid differences.

7.3. Evolutionary changes in amino acid sequences produce changes in the properties of proteins. Such changes may be deleterious, adaptive, or neutral in their phenotypic effects.

7.4. The detectable homology between two polypeptides with a common origin may disappear during evolution. This disappearance is the result of progressive evolutionary differentiation following gene duplication. Examples of the phenomenon are given.

◇ ◇ ◇ ◇ ◇

Note added in proof. Since this manuscript was written, McLachlan [33], [34] has discussed at length methods for searching for repetitive homology in globin and other polypeptide chains.

◇ ◇ ◇ ◇ ◇

The author thanks Doctors Richard Holmquist, Lila Gatlin, and Jack King for suggestions, and Mrs. Regina Cepelak for preparing the manuscript.

REFERENCES

[1] G. M. Air, E. O. P. Thompson, B. J. Richardson, and G. B. Sharman, "Amino-acid sequences of kangaroo, myoglobin and haemoglobin and the date of marsupial-eutherian divergence," *Nature*, Vol. 229 (1971), pp. 391–394.

[2] T. B. BRADLEY, JR., R. C. WOHL, and R. F. RIEDER, "Hemoglobin Gun Hill: deletion of five amino acid residues and impaired hemoglobin binding," *Science*, Vol. 157 (1967), pp. 1581–1583.
[3] H. BROWN, F. SANGER, and R. KITAI, "The structure of pig and sheep insulins," *Biochem. J.*, Vol. 60 (1955), pp. 556–565.
[4] C. R. CANTOR, "The occurrence of gaps in protein sequences," *Biochem. Biophys. Res. Commun.*, Vol. 31 (1968), pp. 410–416.
[5] M. O. DAYHOFF, *Atlas of Protein Sequence and Structure*, Vol. 4, Silver Spring, Md., National Biomedical Research Foundation, 1969.
[6] R. DOOLITTLE, Personal communication, 1971.
[7] G. M. EDELMAN, B. A. CUNNINGHAM, W. E. GALL, P. D. GOTTLIEB, U. RUTISHAUSER, and M. J. WAXDAL, "The covalent structure of an entire γG immunoglobulin molecule," *Proc. Nat. Acad. Sci. U.S.A.*, Vol. 63 (1969), pp. 78–85.
[8] W. M. FITCH and E. MARGOLIASH, "A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome *c* as a model case," *Biochem. Genet.*, Vol. 1 (1967), pp. 65–71.
[9] W. M. FITCH and E. MARKOWITZ, "An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution," *Biochem. Genet.*, Vol. 4 (1970), pp. 579–593.
[10] W. M. FITCH, "The relation between frequencies of amino acids and ordered trinucleotides," *J. Mol. Biol.*, Vol. 16 (1966), pp. 1–27.
[11] H. FUJIKI and G. BRAUNITZER, "The primary structure of lamprey haemoglobin and its contribution to the evolutionary aspects of haemoglobin molecules," *Folia Bioch. et Biol. Graeca*, Vol. 7 (1970), pp. 68–73.
[12] L. L. GATLIN, "Evolutionary Indices," *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1972, Vol. 5, pp. 277–296.
[13] J. I. HARRIS, F. SANGER, and M. A. NAUGHTON, "Species differences in insulin," *Arch. Biochem. Biophys.*, Vol. 65 (1956), pp. 427–438.
[14] R. HOLMQUIST, "Theoretical foundations of paleogenetics," *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1972, Vol. 5, pp. 315–350.
[15] R. HUBER, H. FORMANEK, and O. EPP, "Kristallstrukturanalyse des Met-Erythrocruorins bei 5, 5Å Auflösung," *Naturwissenschaften*, Vol. 55 (1968), pp. 75–77.
[16] V. M. INGRAM, "Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin," *Nature*, Vol. 180 (1957), pp. 326–328.
[17] ———, "Gene evolution and the haemoglobins," *Nature*, Vol. 189 (1961), pp. 704–708.
[18] T. H. JUKES, "Some recent advances in studies of the transcription of the genetic message," *Advan. Biol. Med. Phys.*, Vol. 9 (1963), pp. 1–41.
[19] T. H. JUKES and C. R. CANTOR, "Evolution of protein molecules," *Mamm. Prot. Metab.*, Vol. 3 (1969), pp. 21–132.
[20] T. H. JUKES and R. HOLMQUIST, "Estimation of evolutionary changes in certain homologous polypeptide chains," *J. Mol. Biol.*, Vol. 64 (1972), pp. 163–179.
[21] M. KIMURA, "The rate of molecular evolution considered from the standpoint of population genetics," *Proc. Nat. Acad. Sci. U.S.A.*, Vol. 63 (1969), pp. 1181–1188.
[22] J. L. KING and T. H. JUKES, "Non-Darwinian evolution," *Science*, Vol. 164 (1969), pp. 788–798.
[23] D. E. KOHNE, "Evolution of higher organism DNA," *Quart. Rev. Biophys.*, Vol. 3 (1970), pp. 327–375.
[24] S. L. LI and A. RIGGS, "The amino acid sequence of hemoglobin V from the lamprey *Petromyzon marinus*," *J. Biol. Chem.*, Vol. 245 (1970), pp. 6149–6169.
[25] E. MARGOLIASH and E. L. SMITH, "Structural and functional aspects of cytochrome *c* in

relation to evolution," *Evolving Genes and Proteins* (edited by V. Bryson and H. J. Vogel), New York, Academic Press, 1965, pp. 221–242.

[26] G. MILHAUD, Personal communication, 1971.

[27] M. F. PERUTZ, "A model of oxyhaemoglobin," *J. Mol. Biol.*, Vol. 13 (1965), pp. 646–668.

[28] F. W. PUTNAM, "Immunoglobulin structure: variability and homology," *Science*, Vol. 163 (1969), pp. 633–644.

[29] F. SANGER and E. O. P. THOMPSON, "The amino-acid sequence in the glycyl chains of insulin. 1. The identification of lower peptides from partial hydrolysates," *Biochem. J.*, Vol. 53 (1953), pp. 353–374.

[30] T. SÖDERQVIST and B. BLOMBÄCK, "Fibrinogen structure and evolution," *Naturwissenschaften*, Vol. 58 (1971), pp. 16–23.

[31] A. C. WILSON and V. M. SARICH, "A molecular time scale for human evolution," *Proc. Nat. Acad. Sci. U.S.A.*, Vol. 63 (1969), pp. 1088–1093.

[32] E. ZUCKERKANDL and L. PAULING, "Evolutionary divergence and convergence in proteins," *Evolving Genes and Proteins* (edited by V. Bryson and H. J. Vogel), New York, Academic Press, 1965, pp. 97–166.

Added in proof.

[33] A. D. MCLACHLAN, "Tests for comparing related amino-acid sequences. Cytochrome *c* and cytochrome c_{551}," *J. Mol. Biol.*, Vol. 61 (1971), pp. 409–424.

[34] ———, "Repeating sequences and gene duplication in proteins," *J. Mol. Biol.*, Vol. 64 (1972), pp. 417–438.