Conserved Regions in Mammalian β-globins: Could They Arise by Cross-species Gene Exchange?

MICHAEL SYVANEN

Department of Microbiology and Molecular Genetics, Harvard Medical School, 25 Shattuck Street, Boston, Massachusetts 02115, *U.S.A.*

Comparison of the nucleotide sequences from the coding regions of the four mammalian β-globin genes shows that different parts of these genes have evolved at two different rates. Those codons designating amino acids 1—20, 41—91 and 109—146 have accumulated substitutions in a random fashion as the molecular clock hypothesis would predict. The codons at positions 21—40 and 91 to 108 behave as if they evolved at a much slower rate. Each of the slowly evolved regions contains an intron. Conservation of the coding sequences flanking the introns are hypothesized to be the result of cross-species gene exchange.

Introduction

There is a growing number of reports of highly similar genes or their products that are found in remotely related species; cases that are most easily explained by the transfer of genetic material across species boundaries. Among the examples are the isolation of a eukaryotic superoxide dismutase from a bioluminescent bacterium (Martin & Fridovich, 1981), the wide-spread distribution of closely related oncogenes between unrelated mammals (Benveniste & Todaro, 1974), the finding of similar oncogenes between arthopods and mammals (Shilo & Weinberg, 1981), a suspected transposable element among vertebrate and insects (Singh, Purdom & Jones, 1980), and the possibility that plant leghemoglobin has derived from animal globin (Hyldig-Nielson *et al.,* 1982). The most impressive example to date is based on a comparison of the total histone gene sequence from remotely related sea urchins which shows preservation of bases at degenerate codon positions (Busslinger, Rusconi & Birnstiel, 1982). In these examples it appears that an entire gene is conserved between the unrelated species.

The major theory of molecular evolution—the molecular clock hypothesis (Wilson, Carlson & White, 1977)—was formulated prior to the evidence of cross-species gene exchange. In fact, given the data upon which the molecular clock is based, there has not been compelling reason to cite

cross-species gene exchange. This is due mainly to the fact that the degree of divergence for homologous proteins appears to depend only on the time at which two lineages have separated. However, divergence of sequence by itself does not rule out frequent exchanges of genetic information between species since the consequence of horizontal gene exchange could be to decrease the amount of divergence, to slow down the molecular clock, not necessarily to alter the qualitative pattern of divergence.

Considering the possibility that cross-species gene exchange could play a noticeable role in evolution, I undertook a reexamination of some assumptions that underlie the molecular clock hypothesis. In particular, one of these assumptions is that the same positions in homologous proteins have an equal probability of undergoing substitutions. King & Jukes (1969) have lent support to this assumption by showing that the number of substitutions in certain neutral positions in cytochrome-cs are predictable in terms of the Poisson distribution. However, this was considered an approximation because somewhat arbitrary judgements must be made in assigning neutral positions. For example, Peacock & Boulter *(1975)* have noted many areas of non-random substitutions for various homologous protein sets. This is probably due to the functional constraint upon amino acid sequence (Dickerson, 1971). With complete gene sequences available, it is now possible to examine both those nucleotide differences that cause amino acid substitutions (replacement site differences) as well as those that do not (silent site differences). Preliminary comparisons of nucleotide sequences seem to be more consistent with the molecular clock hypothesis than are comparative amino acid sequences (Perler *et al.,* 1980; Efstradiatis *et al.,* 1981; Miyata & Hayashida, 1981). This is probably due to the fact that silent site changes are not affected by protein structure. However, still not tested is the assumption that the same positions in homologous genes have an equal probability of undergoing substitutions.

In this paper I have tested the hypothesis that differences between homologous genes are random with respect to position when analyzing silent site codons. The homologous set I have chosen to examine is the four sequenced mammalian ß-globin genes that provide three independent pair-wise comparisons. In order to test the hypothesis of randomness across the gene I have employed a modification of procedures standardly used for comparing homologous genes or proteins.

Analytical Procedure for Test of Randomness

If the substitutions along a gene are randomly distributed, the probability that any stretch of codons of length d (gap length) has zero mutations will

be given by the zero term of the poisson distribution, $e^{-d/\breve{d}}$ where $\breve{d}$ is the average gap length (i.e. gene length divided by n +1 where n is the number of substitutions; see below). This will give the probability that the codons are contained within a gap that is greater than or equal to d. Thus, the probability of having a gap between i and j codons long is

$$e^{-i/\breve{d}} - e^{-j/\breve{d}} \qquad (1)$$

This term multiplied by n +1 will give the expected number of gaps of that size within the gene being compared.

The value n is the observed number of substitutions plus a correction for those codons where only zero or one change is detected, although two substitutions have actually occurred—the invisible doubles. To calculate this correction I will assume that doubles are distributed randomly and therefore a Poisson correction can be applied (Holmquist, 1980). Silent sites that have two mutational events will result in no change (49% of the time), a single change (44% of the time) or two changes (7%). This percentage distribution is calculated from the table of the genetic code assuming that the relative abundance of amino acids in proteins is the same as their relative abundance in the table (Doolittle, 1981) (This assumption is only an approximation; the exceptions, however, do not significantly affect the resulting correction). The actual number of substitutions, n, can be calculated by setting the fraction of codons observed with zero substitutions equal to

$$e^{-n/L}(1 +049/2(n/L)^2). \qquad (2)$$

Where L is the total number of codons. If n/L is >0.35 errors will accrue that are not accounted for by this correction.

Distribution of Silent Site Changes in Mammalian ß-globin

I first tested whether base substitutions across the four mammalian ß-globin genes are random with respect to position after the bias introduced by functional constraint upon amino acid sequences was removed. To this end I examined only those silent site changes that are within codons that have not undergone a replacement substitution. This was done by aligning the nucleotide sequences from the coding regions for each pair of ß-globins and deleting those codons that designate different amino acids. The main reason for this decision is that a codon at a non-conserved site has a higher probability of undergoing a "silent site" substitution than a codon at a conserved site. This follows from the ambiguity caused by trying to decide if, in those codons for non-conserved amino acids with two changes, the

codons have undergone two replacement substitutions or one replacement and one silent site change.

Figure 1 shows pair-wise comparisons of silent site differences between the truncated ß-globin genes of the four mammalian species: goat, human, mouse and rabbit. (Only three histograms are given since of the six possible comparisons, only three are independent for the purpose of the probabilities calculated below.)
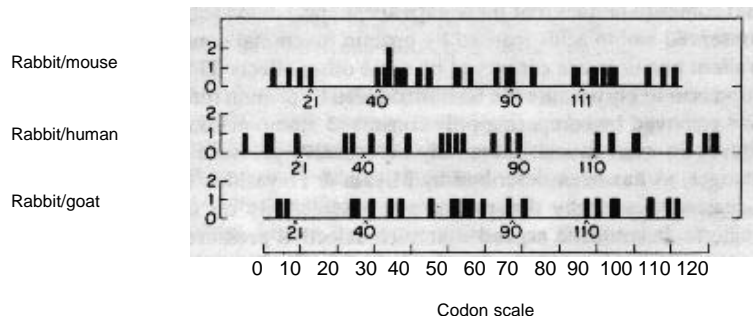

Fig. 1. Distribution of substitutions in the codons of the conserved amino-acids. Codon to codon comparison of the coding sequences for the three pairs of ß-globin genes was made beginning with amino acid 1 on the left to 146 on the right. In the goat gene codons 1 and 2 are deleted. For each comparison a pair of codons were recorded as zero if they were identical, I if a single substitution and 2 for two. Codons with replacement substitutions were deleted from the map, hence the variation in map length. Actual corresponding codon position is indicated with "^" at the various points. Codon scale was used to determine distances between substitutions that are recorded in Table I. The complete nucleotide sequence for the four genes are from Hardison *et al.* (1979) for rabbit; Konkel *et al.* (1979) for mouse; Lawn *et al.* (1980) for human; and Schon *et* al. (1981) for goat.

First, I will show that over each pair-wise comparison the distribution of substitutions across the entire gene is nearly random. In order to test for randomness, the distance between adjacent substitutions is measured from the histogram in Fig. 1; these regions without substitutions are called gaps. If the distribution of substitutions is random across the gene, then the probability that a region of any given size contains no substitution is given by the zero term of the Poisson distribution (see equation (1)). Using the information in Fig. 1, I have compared the observed distribution of gaps against this expectation of random distribution. As shown in Table 1, for each pair-wise comparison the observed number of gaps does not deviate significantly from the random expectation. The result indicates that substitutions in silent sites of the codons of conserved amino acids fulfills the conditions of the molecular clock hypothesis. that substitutions may occur with equal probability at any position across the gene.

TABLE **1**

Distribution of gaps are random for each pairwise comparison. Gap size is the number of codons separating adjacent substitutions. The observed number of gaps for each interval is determined from the data in Fig. 1 *with corrections from equation* (2). *Expected values are calculated from equation* (1) *where the range in gap sizes, shown in the Table, determines the values of i and j. p is the probability that the* over-all distribution *of gaps does not deviate from random based on a* χ^2 *test of significance applied to the entire data set; 5° of freedom in each test. Long gaps, those* >11 *are excluded from the* χ^2 *test; the probability of their occurrence by chance is given by expected value*

Gap size	Observed	Expected	Gap size	Observed	Expected
Rabbit/mouse ($n=37$; $\overline{d}=2{\bullet}95$; $p=0{\bullet}82$)			Rabbit/human	($n=28$; $\overline{d}=4{\bullet}4$;	$p=0{\bullet}82$)
0—1	9•75	10•9	0—1	4•75	5•8
1—2	6•5	7•8	1—2	4	4•6
2—3	5•5	5•5	2—3	5•75	3•7
3—4	6•5	3•9	3—6	6•5	8•7
4—11	7•75	8•8	6—12	5	4•1
>11	2	0•9	>12	3	2
>18	1	0•085	>18	2	0•5
11 and 18		0•076	>20	1	0•33
			12, 18 and 20		0•16

Rabbit/goat ($n = 27$; $\overline{d}=4{\bullet}03$; $p = 0{\bullet}95$)					
0—1	5•25	6•1			
1—2	4•75	4•8			
2—4	6•75	6•6			
4—6	5•25	4•1			
6—16	4	5•8			
>16	2	0•53			
>17	1	0•41			
16 and 17		0•22			

Unexpected Long Gaps

Even though the *overall* distribution of substitutions for each pair-wise cross is random—as judged by the χ^2 analysis—there is one feature that was unexpected. This is the occurrence of long gaps in each comparison. These can be seen as gaps of 10 to 20 codons which have, in the three comparisons, a 7% to 50% chance of occurring if the distribution of substitutions is truly random. For any given pair-wise cross this is not too interesting. What is unexpected is that these long gaps are found between codons 21—40 and 91—108 in each comparison, as can be seen by inspection of Fig. 1. (These gaps correspond to regions of highly conserved amino acids seen in the ß-globin sequence; they are not the result of deleting nonconserved
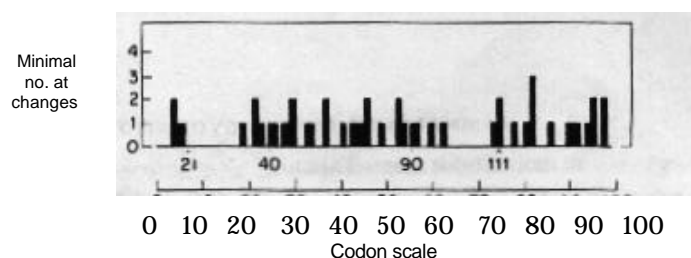
FIG. 2. Distribution of substitutions in the four β-globins in codons of conserved amino acids. Of the 146 amino acids, 98 of them are the same in all four β-globins: substitutions at these sites are recorded. The number of substitutions required to produce the observed differences are calculated assuming that the four lineages leading to these four species diverged simultaneously and that the ancestral sequence can be reconstructed from the least number of substitutions. For example, if silent sites at the four codons are identical it is assumed that is the ancestral sequence, if two are of one base and two of another, one of the two is ancestral and two substitutions occurred to produce the differences and so on.

amino acids during the construction of the histograms.) Clearly the probability of finding such long gaps as these preserved over the same regions in three independent comparisons will be much lower than the probability of finding the long conserved gaps in a single comparison. This probability can be determined by directly comparing the silent site changes in the sequences of all four genes. Figure 2 presents these data. The number of substitutions in Fig. 2 represents the minimal number of changes which the four mammalian sequences must have undergone since their divergence from a

TABLE 2

Four-fold comparison shows unexpected long gaps. See legend to Table 1. The value of n is the sum of the total number of changes from Fig. 2 plus invisible double substitutions of 9. Equation (2) cannot be used to estimate n since there are more codons with zero substitutions than expected from the Poisson distribution. The value of 9 comes from

Table 1

Gap size	Observed	Expected
0—1	25	26•7
1—2	13•75	14•6
2—3	7•25	8•0
3—4	7•25	4•4
4—11	4•75	5•2
11	2	0•07
13	1	0•023
11 and 13		(5×10^{-4})

common ancestral ß-globin. The gap information, given in Table 2, shows the probabilities. Again, the over-all distribution of substitutions appears to be random (p=0•82). However, there are two long gaps of 13 and 11 codons whose expectation of occurring is extremely low (5×10^{-4}).

Taken together, the data in Figs 1 and 2 show that the ß-globin gene can be divided into two regions. The major part of the structural gene (roughly codons 1—20, 41—90, 109-146) has accumulated substitutions in a random fashion, as predicted by the molecular clock hypothesis. The second part of this gene (codons 21—40, and 91—108) is conserved when compared to rest of the sequences. The probability that this difference occurred by chance is 5×10^{-4}.

There are a number of possible explanations for the two unexpected long gaps. One that comes to mind is that these gaps are simply coding sequences for highly conserved amino acids (caused by protein functional constraint) and that the silent positions are conserved by some other effects. However, this explanation is unlikely because the bias introduced by protein functional constraint was removed by comparing only conserved amino acids, as was described above. So even though there may be selective pressure against silent site changes, as has been described by Miyata & Hayashida (1981), there is no reason to see why this pressure acts selectively on the two conserved regions. It could be argued that such selective pressure arises because, just by chance, the amino acids through these regions happen to be populated with those amino acids that display a large bias in codon usage (Hardison, *et al,* 1979). If so, we would predict that the highly preferred codons would be used through these two regions. This is clearly not the case. The codons GUC for Val and CUC for Leu are rarely used in mammalian globins (about 10% of the time for each) but both are conserved in these long gaps (at positions 34 and 105 respectively). In addition many of the amino acids in this region show no codon bias. In general, the amino acids and respective codons for these two regions are not unusual.

There remains the possibility of functional constraint acting upon either mRNA structure or genome structure that conserves these two regions. In this explanation nucleotide sequence in the two regions are important for expression of active hemoglobin. Undoubtedly important to any explanation for the long gaps is the fact that the two introns of ß-globin are located within each of the two conserved regions; they are inserted between codons 30 and 31 and 104 and 105 respectively. Thus, one functional constraint explanation would be that intron processing conserves these sequences, implying that 50 to 60 nucleotides in exons are involved in processing of each intron. This number is much larger than the number of nucleotides which constitute the known consensus sequence (Breathnach *et al.,* 1978).

Evidence for functional constraint acting upon either RNA or DNA structure could possibly be found from an examination of variant ß-globins that have been seen in human populations.

Distribution of Amino Acid Substitutions Among Variant Human β-globins

If silent site substitutions through the long gap regions are not neutral because of functional constraint at the level of nucleic acid function, then it seems quite likely that replacement site mutations would not be tolerated in these regions for this same reason. For example, such nucleotide substitutions may lead to aberrant ß-chain synthesis, or perhaps are important for chromosome structure or replication. If such mutations cause aberrant synthesis of ß-chains or cause cis-dominant lethal mutations then support for either notion should be obtained from an examination of the phenotypes associated with the variant ß-globins found in the human population. There are many known variant ß-globins with phenotypes that range from asymptomatic to severe blood disorders. The variants fall into two classes:
the majority synthesize normal amounts of the mutated ß-chains while the minority, known as ß-thalassemias, are defective in ß-chain synthesis. The position of the amino acid substitutions that constitute this major class of variant ß-chains is shown in the histogram of Fig. 3. As can be seen, these substitutions are found more or less evenly across the molecule. This shows that replacement site substitutions are tolerated between codons 2 1—40 and codons 9 1—108 indicating that alteration of these sequences does not cause cis-dominant lethal mutations and, furthermore, that ß-chain expression is not affected. It should be pointed out that many of the variants in Fig. 3 could never be fixed in the human population since the resulting hemoglobins are defective; but these are defective at the level of hemoglobin function,
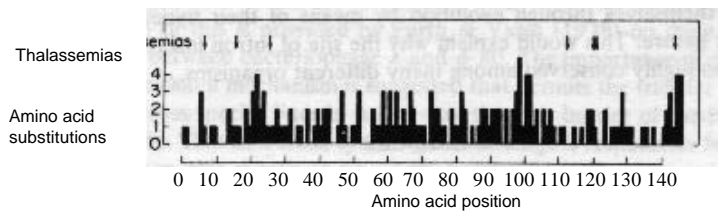

FIG. 3. Distribution of amino acid substitutions in variant human beta-globins. Nearly 200 different human variant beta globins are known carrying a single amino acid substitution. The histogram shows which amino acids are mutated and the number of different mutations seen at each position. Only those variants that synthesize the mutated chain are given in the histogram. The position of those variants that cause beta-thalassemia are indicated by the arrows. The data is taken from compilations by Lehmann & Kynoch (1976) and Bunn & Forget (1984).

not hemoglobin synthesis or chromosome integrity. As I have shown above, the conserved regions from Fig. 2 cannot be due to functional constraint at the level of protein action. A second way that we can examine whether changes between amino acids 2 1—40 and 9 1—108 cause reduced levels of ß-chain synthesis may be seen in the distribution of substitutions that lead to ß-thalassemia — the disease characterized by a depression in ß-chain synthesis. These are indicated by the arrows above the histogram in Fig. 3. Of the seven ß-chain amino acid substitutions that lead to thalassemia only one (position 26) is found in the long gaps. To conclude, the information in Fig. 3 shows that nucleotide changes in the long gaps (that result in amino acid substitutions) have little affect upon ß-chain expression. We may infer that silent site changes in these codons would not likewise affect ß-chain expression. This raises the possibility that the silent site changes are indeed neutral.

A caveat to the above conclusion. There is a possibility that a subtle defect in ß-chain expression results from mutations in the long gaps—too subtle to be detected in the affected patients but sufficiently severe to prevent such mutations from ever being fixed in the respective populations.

Implications for a Theory of Cross-species Gene Exchange

The explanation that I would now like to consider for these highly conserved sequences is the possibility that they are caused by polynucleotide sequence exchange among these mammalian lineages at a time significantly after their evolutionary divergence. This is the only possibility that accounts for the conserved gaps in the absence of any clear functional constraint. Here I postulate that the different mammalian germ lines are occasionally exposed to ß-globin sequences from other mammalian species (perhaps by viral infection) and that the information on these sequences becomes incorporated by a gene conversion event.

Let it be granted that the silent site substitutions in the ß-globin coding sequences are neutral (at least through most of the positions that define the long gaps) and that the lack of substitutions is brought about by cross-species gene exchange. If so, then we should reevaluate the meaning of the molecular clock. The most important consequence is that the molecular clock will be slower than predicted by Kimura (Kimura, 1968; King & Jukes, 1969). This conclusion can be directly inferred if we compare the distributions in Figs 2 and 3. Within the human population, all positions give rise to substitutions (Fig. 3) but when we examine which substitutions are being fixed (Fig. 2), those in one region are preferred over another. We can then recast Kimura's quantitative statement in the following manner:

"the rate of random fixation of neutral mutations in evolution (per species per generation) is equal to the rate of occurrence of neutral mutations (per gamete per generation)" (King & Jukes, 1969) *minus* those that undergo a gene conversion event. The gene conversion event occurs between the two species being compared, thus erasing any differences between them. The qualitative pattern of the molecular clock will not be disrupted if gene conversion events between closely related species are more probable than those between distantly related species. This situation is consistent with the finding that broad host range viruses (the probable vectors of cross-species gene transfer) usually infect a group of species that are closely related.

One of the first criticisms levelled against the notion that cross-species gene exchange may have a significant influence on evolution involved the problem of integration of phenotypes or coadaptation (see West (1971) in response to Anderson (1970)). The problem of coadaptation may be illustrated by a consideration of protein-protein or protein-nucleic acid interactions. Most proteins have multiple functional domains, many of which are involved in highly specific interactions with other proteins or nucleic acid sequences. For such proteins, possible amino acid substitutions will be limited by its partner's conformation and, further, as neutral changes do occur, these will influence the types of changes possible in this partner. Over time, this interacting pair will drift, slowly changing the interactive surface between them. If a lineage split in two, and the resulting homologous pair drifted sufficiently far apart, then eventually we can see that hybrids between these pairs might lose their functional interactions. At this point transfer of a functional homologous gene from one lineage to another would no longer be possible. That is, these two homologous pairs are no longer coadapted. There are numerous concrete examples of this seen with hybrids between distantly related bacteriophages (Susskind & Botstein, 1978), and many phenomena from population genetics can be interpreted in this way (Mayr, 1954). Thus, for cross-species gene exchange to be able to contribute traits to complex interacting systems, it would be important to be able to introduce the parts of the gene that carry a specific trait while leaving intact those parts that determine interactions with other elements. A beautiful example of this is provided by Furth & Yates (1978) on their studies of hybrids between bacteriophage *A* and ø 80. The importance of my current results is that a mechanism is suggested that permits the transfer of genetic traits across species lines in such a way that this barrier of coadaptation is overcome. That is, if cross-species gene exchange is responsible for the long gaps, then this means there exists a relatively efficient mechanism, in eukaryotes, for transferring *parts* of genes. This provides a general means for overcoming the problem of coadaptation during transfer of a trait.

Involvement of Introns

Given that the conserved gaps flank ß-globin's two introns, we can extend the hypothesis to say that introns serve to facilitate the incorporation of nucleotide sequences from the homologous genes of foreign species. Introns may have specifically arisen to overcome the barrier of coadaptation. This idea would be an extension of some current ideas about the functional significance of introns. Gilbert (1978) suggested that introns may serve as recombinators that facilitate genetic rearrangements that are of importance in the evolution of new proteins. The finding that introns separate protein functional domains rather precisely in globulins and approximately in betaglobins has been offered in support of this hypothesis (Blake, 1979; Craik, Buchman & Beychok, 1980; Eaton, 1980; Go, 1981). Possibly relevant to this point is that all nine amino acids in ß-globin that are involved in the $\alpha1\ \beta2$ contact are contained within the long gap region described here (they are amino acids 36, 37, 39, 40, 97, 98, 99, 101, 102, Perutz, 1965). In the hemoglobin tetramer, these are the only amino acids contacts between subunits that change during the R-T conformational transition upon 02 binding (Perutz *et al.,* 1968; Edelstein, 1975). They are important for the allosteric interactions of hemoglobin and have been postulated to be the most recent functional addition to the evolution of hemoglobin (Eaton, 1980).

I do not wish to imply that the amino acid sequence contained in the long gaps arose since the mammalian dispersion. This is probably not the case since these same or very similar amino acid sequences are found in both the embryonic beta like globins (Slighton, Bechl & Smithis, 1980) and the chicken and duck ß-globins (Robinson & Ingram, 1982); however the silent sites in these other ß-globins show divergence from the adult mammalian version. More likely, the cross-species gene conversion events among the mammals are neutral.

According to my interpretation introns at these two ß-globin loci evolved originally because they helped establish the $\alpha1\ \beta2$ interactions among many different species before birds and mammals diverged, but these introns now preserve themselves through evolution by means of their hyper-recombinogenic nature. This would explain why the site of intron insertion is, in general, so highly conserved among many different organisms.

REFERENCES

ANDERSON, A. (1970). *Nature* **209**, 637**.**
BENVENISTE, R. E. & TODARO, G. J. (1974). *Nature* **252,** 456.
BLAKE, C. C. F. (1979). *Nature* **277,** 598.

BREATHNACH, R., BENOIST, C., O'HARE, K., GANNON, F. & CHAM RON, P. (1978). *Proc. natn. Acad. Sci. USA.* **75**, 4853.

BONN, H. F. & FORGET, B. G. (1984). *Hemoglobin: Structure, Function Molecular Genetics, Clinical Aspects.* New York: Sanders.

BUSSLINGER, M., RUSCONI, S. & BIRNSTIEL, M. L. (1982). *EMBO J.* **1,** 27.

CRAIK, C. S., BUCHMAN, S. R. & BEYCHOK, S. (1980). *Proc. natn. Acad. Sci. U.S.A.* **77,** 1384.

DICKERSON, R. E. (1971). *J. mol. Evol.* **1,** 26.

DOOLITTLE, R. F. (1981). *Science* **214,** 149.

EATON, W. A. (1980). *Nature* **284,** 183.

EDELSTEIN, S. (1975). *Ann. Rev. Biochem.* **44,** 209.

EFSTRATIADIS, A., POSAKONY, J. W., MUNIATIS, T., LAWN, R., O'CONNEL, C., SPRITZ, R. A., DERIEL, J. K., FORGET, H. G., WEISMAN, S. M., SLIGHTON, J. T., BLECHL, A. E., SMITHIES, 0., BARALLE, F. E., SHOULDERS, C. C. & PROUDFOOT, N. J. (1980). *Cell* **21,** 653.

FURTH, M. E. *&* YATES, J. L. (1978). *J. mol. Biol.* **126,** 227.

GILBERT, W. (1978). *Nature* **271,** 501.

GO, M. (1981). *Nature* **291,** 90.

HARDISON, R. C., BUTLER, E. T., LACY, E., MANITIS, T., ROSENTHAL, N. & EFSTRATIADIS, A. (1979). *Cell* **18,** 1285.

HOLMOUIST, H. (1980). *J. mol. Evol* **15,** 149.

HYLDIG-NIELSON, J. J., JENSEN, E. 0., PALUDAN, K., WIBORG, 0., GARRETT, R., JORGENSEN, P. & MARCKER, K. A. (1982). *Nucleic Acids Res.* **10,** 689.

KIMURA, M. (1968). *Nature* **217,** 624.

KING, J. L. & JUKES, T. H. (1969). *Science Journal* **164,** 788.

KONKEL, D. A., MAIZEL, J. V. & LEDER, P. (1979). *Cell* **18,** 865.

LAWN, R. M., EFSTRATIADIS, A., O'CONNELL, C. & MANIATIS, T. (1980). *Cell* **21,** 647.

LEHMANN, H. & KYNOCH, P. A. M. (1976). *Human Haemoglobin Variants and their Characteristics.* Amsterdam: North-Holland Co.

MARTIN, J. P. & FRIDOVICH, I. (1981). *J. biol. Chem.* **256,** 6080.

MAYR, B. (1954). In: *Evolution as a Process* (Huxley, J. ed.) pp. 157—180. London: Collier.

MIYATA, T. & HAYASHIDA, H. (1981). *Proc. natn. Acad. U.S.A.* **78,** 5739.

PEACOCK, D. & BOULTER, D. (1975). *J. mol. Biol.* **95,** 513.

PERLER, F., EFSTRATIADIS, A., LOMEDICO, P., GILBERT, W., KOLODNER, R. & DODGSON, J. (1980). *Cell* **20,** 555.

PERUTZ, M. F. (1965). *J. mol. Biol.* **13,** 646.

PERUTZ, M. F., MUIRHEAD, H., COX, J. M. & GOAMAN, L. C. G. (1968). *Nature* **219,** 131.

ROBINSON, I. H. & INGRAM, V. M. (1982). *Cell* **28,** 515.

SCHON, E. A., CLERY, M. L., HAYNES, J. R. & LINGREL, J.H. (1981). *Cell* **27,** 359.

SHILO, H. & WEINBERG, R. (1981). *Proc. natn. Acad. Sci. U.S.A.* **78,** 6789.

SINGH, L., PURDOM, I. & JONES, K. (1980). *Cold Spring Harbor Symp. Quant. Biol.* **78,** 6789.

SLIGHTON, J. L., BECHL, A. E. & SMITHIES, 0. (1980). *Cell* 21, 627.

SUSSKIND, M. M. *&* BOTSTEIN, D. (1978). *Microb. Rev.* **42,** 385.

WEST, C. (1971). *Nature* **229,** 637.

WILSON, A. C., CARLSON, S. S. & WHITE, T. J. (1977). *Ann. Rev. Biochem.* **46,** 573.