THE ORIGIN OF THE GENETIC MATERIAL IN THE ABNORMALLY LONG HUMAN HEMOGLOBIN α AND β CHAINS

Lois T. Hunt and Margaret O. Dayhoff
National Biomedical Research Foundation
Georgetown University Medical Center
3900 Reservoir Road, N.W.
Washington, D.C. 20007

Received April 6, 1972

SUMMARY

The additional carboxy-terminal segment of the abnormal human Hemoglobin Constant Spring (Hb CS) α chain and positions 68-98 of the normal human α chain have nine identical residues; this was the maximum number revealed by computer search of 518 known protein sequences. Searches were also made with 31-residue segments having randomized sequences of an average composition. Because the probability is less than 1% that the similarity is due to chance alone, the results favor a common genetic origin for the Hb CS α chain additional segment and the region 68-98 of the normal α chain; it may also be true that the additional segment of Hemoglobin Tak β chain arose in a similar manner.

Recently two variant human hemoglobins having elongated chains have been described; Clegg, Weatherall and Milner[1] have found that the α chain of Hemoglobin Constant Spring (Hb CS) has 31 additional carboxy-terminal residues, and Flatz, Kinderlerer, Kilmartin and Lehmann[2] have found that the β chain of Hemoglobin Tak (Hb Tak) has 10 additional carboxy-terminal residues. The normal messenger RNAs for the hemoglobin chains are believed to have a length greater than that needed to code for the actual proteins. Therefore, an abnormally elongated protein form may result from a mutation in the codon which terminates the translation of mRNA into protein; the additional residues in the abnormal chain would then reflect chromosome structure which is transcribed into mRNA but not usually translated into protein. Alternatively, the genetic material which codes for the extended region may have been

TABLE 1

Segments Similar to the Carboxy-terminal 31 Residues of the α Chain of Human Hemoglobin Constant Spring (CS)

	PROTEIN	POSITION IN SEQUENCE		SEQUENCE
	Hemoglobin α chain — Human CS	142	173	Q A G A S V A V P P A R W A S Q R A L L P H S L R P F L V F E
NINE IDENTICAL RESIDUES	Hemoglobin α chain — Human	68	98	N A V A H V D D M P N A L S A L S D L H A H K L R V D P V N F
	Hemoglobin α chain — Mouse	68	98	N A G(A.H.L)D D L P G A L S A L S D L H A H K L R V D P V N F
	Trypsin-like Enzyme — Myxobacter 495	161	191	G N V Q S N G N N C G I P A S Q R S S L F E R L Q P I L S Q Y
EIGHT IDENTICAL RESIDUES	Globin — Gastropod Mollusc	32	62	G F V A S V A A P P A G A D A W T K L F G L I I D A L K A A G
	Hemoglobin α chain — Kangaroo	68	98	Q A V E H I D D L P G T L S K L S D L H A H K L R V D P V N F
	Hemoglobin α chain — Llama	68	98	K(A.A,A.H.L.B.B.L.P.S,A.L.S.B.L.S.B.L.H.A.H)K.L R(V.B.P.V.B.F.
	Hemoglobin α chain — Irus Macaque	68	98	L.A.V.G.H.V.D.D.M.P.Q.A.L.S.A.L.S.D.L.H.A.H)K.L R(V.D.P.V.N.F.
	Hemoglobin α chain — Rhesus Macaque	68	98	L A V G H V D D M P N A L S A L S D L H A H K L R V D P V N F
	Histone IIb2 — Bovine	83	113	Y N K R S T I T S R E I Q T A V R L L L P G E L A K H A V S E
	Immunoglobulin κ chain — Human BAT	6	36	Q S P L S L P V T P G E P A S I S C R S S Q S L L H S B G B B
	Immunoglobulin κ chain — Human CUM	7	37	Q T P L S L P V T P G E P A S I S C R S S Q S L L D S G D G N
	Immunoglobulin κ chain — Human TEW	6	36	Q S P L S L P V T P G E P A S I S C R S S Q H(G,B)S F L N W Y
	Lipotropin β — Pig	7	37	P E P A R D P E A P A E G A A A R A E L E Y G L V A E A Q A A

introduced through recent chromosomal aberrations. In either case, the origin of the extra segments would be of interest. For each abnormal protein, the authors[1,2] noted that the sequence of the extra segment did not resemble any portion of any known hemoglobin sequence. In an effort to discover similarities, we have compared the sequences of the additional segments with all of the sequences in the "Protein Sequence Data Deck" from the *Atlas of Protein Sequence and Structure 1972*[3]. The data include 518 sequences, with a total of 48,292 residues, derived from 154 different protein families. The results of these computer searches are shown in Tables 1 and 2.

Let us consider first the additional segment from the α chain of HB CS. The highest number of identities found for the 31-residue segment was 9. It is surprising that one of the three high-scoring segments occurred in the α chain of human hemoglobin (positions 68-98), unless the genetic material represented in the extra segment of the HB CS α chain was derived from the structural genes of this family. Some estimate of the probability of this similarity occurring by chance can be derived in two ways. The first is theoretical and the second involves a computer simulation. About 32,500 segments with a length of 31 residues occur in the collected data, and 111 such segments are found in the α chain of human hemoglobin. Thus, if one hypothesizes no relationship, there is one chance in 293 that a single high score which we might find would have come from this chain, and about one chance in 100 that one of the three segments which we actually found would be derived from it.

In order to compare experimentally, by computer simulation, the results of the search shown in Table 1 with a null case, we constructed segments whose composition was average for the protein families in the data collection: they had 3 each of Ser and Ala; 2 each of Asp, Glu, Pro, Thr, Gly, Val, Leu, and Lys; and 1 each of Ile, Asn, Gln, Arg, Phe, Tyr, Cys, His, and Met. We generated 12 sequences with this composition by lottery and searched the data collection for similar segments. There were five segments or an average of

TABLE 2

Segments Similar to the Carboxy-terminal 10 Residues of the β Chain of Human Hemoglobin TAK

	PROTEIN	POSITION IN SEQUENCE		SEQUENCE
	Hemoglobin β chain — Human TAK	147	156	T K L L A(N,S,L)F Y
FIVE IDENTICAL RESIDUES	Elastase — Pig	119	128	T I L A N N S P C Y
	Hemoglobin α chain — Bovine	98	107	F K L L.S H S L L.V
	Hemoglobin αA chain — Goat	98	107	F)K L L.S H S L L.V
	Hemoglobin βB chain — Bovine	102	111	F K(L.L.G.N.V.L.V.V.
	Hemoglobin β chain — Dog	103	112	F K L L G N V L V C
	Hemoglobin β chain — Brown Lemur	103	112	F)K(L.L.G.B.S,L.B,S,
	Hemoglobin β chain — Rhesus Macaque	103	112	F K L L G N V L V C
	Hemoglobin γ chain — Human	103	112	F K L L G N V L V T
	Lysozyme — Bacteriophage T4	59	68	T K D E A E K L F N
	Growth Hormone — Human	93	102	R S V F A N S L V Y
	Coat Protein — Bacteriophage FR	120	129	T A I A A N S G I Y

0.4 sequences per search with a score of 9 identities. None of those found was from a hemoglobin chain. There were 7.7 sequences per search with a score of 8. Among the 73 protein families with scores of 8 or 9 represented in the 12 searches, the α hemoglobin family appeared three times. The chance that a hemoglobin α chain segment with a score of 9 would appear can then be estimated by the product of the chance that the sequence of any segment will have a score of 9 and the chance that a segment from the hemoglobin α family will have a high score: $0.4 \times \frac{3}{73} = 0.016$.

The chance that a segment from the human hemoglobin α chain itself would score high is somewhat less; only five occur among the 47 hemoglobin α chain segments having 7, 8, and 9 identities (high scores) with the random segments. The chance that a segment from human hemoglobin α chain would have a score of 9 would then be $0.016 \times \frac{5}{47} \sim 0.002$.

For this value 0.002 we derived an estimate of the standard deviation,

which is ±0.002.* If the distribution is normal, then the probability of a segment from human hemoglobin α chain having a score of 9 is less than 1%, with a confidence level greater than 99%.

From either point of view, the theoretical one or the computer experiment, the probability is low, only about 1%, that the similarity of the extra segment of HB CS α chain to the human hemoglobin α chain is due to chance alone.

Although a probability of such magnitude by itself does not establish a conclusion, nevertheless, the known mechanisms of genetic aberration and the proximity on the chromosome of the genetic material for the extra segment to that for the similar region of the α chain of human hemoglobin make it very probable that the segment including residues 68 to 98 of the α chain and the additional segment of HB CS α chain had a common evolutionary origin. Because several of the other mammalian α chains are almost as similar, we would surmise that the appearance of the extra segment occurred as far back as the divergence of these species from each other, possibly 75 million years ago. If this is so, the rate of acceptance of point mutations is considerably faster in the additional segment than in the functional part of the α chain.

The results of a search for segments similar to the additional carboxy-terminal segment of the β chain of Hb Tak are shown in Table 2. Hemoglobin chains are prominent in the results. For this short segment, it is not possible to make a strong case for its origin from a portion of the functional hemoglobin chain. However, the data are not inconsistent with such an origin.

---

* Theoretical estimates of the standard deviations of the random variables involved here are given by $\sqrt{N}$, where N counts are observed. The standard deviation of the product of three functions was estimated as follows:

Let δ_n = the standard deviation of the nth variable divided by its mean; then the fractional standard deviation of the product of three independent distributions is given by

$$\delta_{ijk} = \sqrt{\delta_i^2 + \delta_j^2 + \delta_k^2 + \delta_i^2 \delta_j^2 + \delta_i^2 \delta_k^2 + \delta_j^2 \delta_k^2}$$

The carboxyl ends of the two abnormal chains are not very similar to each other. Their origins may well have been two different series of genetic events. This is to be expected if the additional α chain segment arose from a duplication at the time of the mammalian radiation, long after the duplication leading to α and β chains of hemoglobin.

ACKNOWLEDGMENTS

This investigation was supported by Grants GM-08710 and RR-05681 from the National Institutes of Health.

REFERENCES

1. Clegg, J.B., Weatherall, D.J. and Milner, P.F., *Nature*, 234: 337 (1971)
2. Flatz, G., Kinderlerer, J.L., Kilmartin, J.V. and Lehmann, H., *Lancet*, 1: 732 (1971)
3. *Atlas of Protein Sequence and Structure 1972*, ed. Dayhoff, M.O., 5: in press, National Biomedical Research Foundation, Washington, D.C. (1972)