Chemical Paleogenetics

Molecular "Restoration Studies" of Extinct Forms of Life

LINUS PAULING and EMILE ZUCKERKANDL*

*Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California, USA***

Attention is attracted to the possibility of reconstructing the amino-acid sequence of ancestral polypeptide chains by virtue of a comparison between the amino-acid sequences of related polypeptide chains found in contemporary organisms. A tentative partial structure is proposed for two ancestral hemoglobin polypeptide chains. Some perspectives of paleobiochemistry are outlined.

In different hemoglobin polypeptide chains, derived either from one individual organism (man) or from different vertebrate species, identical amino-acid residues are often found in corresponding positions along the chains (*cf.* Braunitzer[1]). This occurs too frequently to be due to chance, and it appears to be too constant a feature over long periods of evolutionary time to be attributable to convergence by natural selection from primitively heterologous polypeptide starting materials. Consequently, the homology is plausibly interpreted by the assumption of the past existence of common polypeptide-chain ancestors, controlled by common ancestral genes. At least a few times during evolution an evolutionarily effective duplication, *i. e.*, a duplication that has been spread at least temporarily by natural selection, of either a hemoglobin gene or a chromosome carrying such a gene is thought to have occurred. The resulting daughter genes are considered to have differentiated by independent mutation (Itano[2]; Ingram[3]; Zuckerkandl and Pauling[4]).

On the basis of this hypothesis, the degree of difference between two homologous polypeptide chains is a measure of the relative time at which the common ancestor of the structural genes controlling these chains existed, and it is also, within large limits of error, a measure of this time in absolute units (Zuckerkandl and Pauling[4]). We now direct attention to two further types of information that can be derived from the comparison of different homologous polypeptide chains. First, it is possible to determine, with some probability, the amino-acid sequence of their presumed common polypeptide-chain ancestor. Second, when two polypeptide chains do not possess the same amino-acid residue at a certain

* On leave from Centre National de la Recherche Scientifique, Paris.

** Contribution No. 2957.

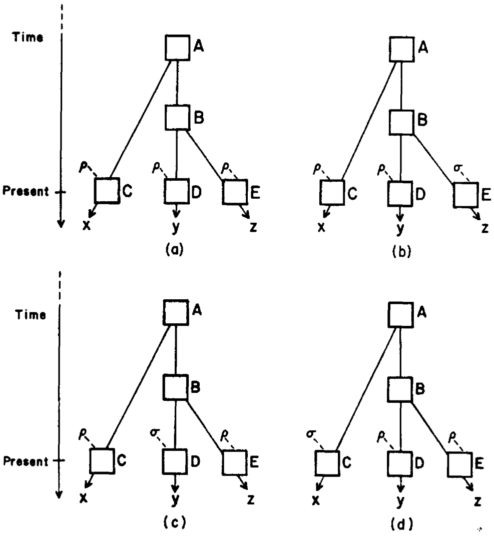

Fig. 1. Different evolutionary relationships of the amino-acid residues ϱ and σ found at corresponding sites in the homologous polypeptide chains C, D, and E. See text.

molecular site, it is possible to determine in which line of descent the mutation responsible for this difference has occurred since the epoch of the common ancestor.

If the amino-acid residue is the same in two chains at a given molecular site, there is a certain likelihood that this residue was also present in the common ancestor of the chains*. If three or more chains are compared and the same residue is found at corresponding sites in two or more chains whereas it differs in one chain, there is a certain likelihood that the mutation responsible for the difference occurred in the line of descent leading to the chain showing this difference, and that it occurred since the time of the most recent molecular ancestor that can be assumed to relate the variant chain to one of the others. The further apart the organisms from which homologous polypeptide chains are derived, the more significant are the identities found, not only because they then suggest that the residue at the molecular site considered has remained the same for a long evolutionary time and therefore must have an important function within the molecule, but also because in widely different forms the occurrence of convergence at the molecular level appears to us to be a negligible possibility.

In Figure 1 three lines of molecular descent, x, y, and z, are represented, and the respective common molecular ancestors at two branching points are A and B. Squares stand for related polypeptide chains. Two different amino-acid residues, ϱ and σ, are considered for one given molecular site in the homologous polypeptide chains, C, D, and E. These chains are located on the time scale at a

* Convergence effects may possibly intervene under some circumstances even at the molecular level, but are not considered to be likely to affect an important proportion of amino-acid residues in polypeptides.

level representing present time. The following probable statements can be made about the four situations pictured: Situation (a): the common chain-ancestors A and B had residue ϱ at the molecular site under consideration. Situation (b): the chain ancestors A and B had residue ϱ; a mutation to residue σ has occurred in evolutionary lineage z after it became distinct from evolutionary lineage y. Situation (c): the chain ancestors A and B had residue ϱ; a mutation to residue σ occurred in evolutionary lineage y after lineage z branched off from lineage y. Situation (d): the chain-ancestor B had residue ϱ, but on the basis of the evidence the nature of the residue in chain-ancestor A is undetermined; further related polypeptide chains have to be drawn into the comparison to permit a conclusion to be reached. The probability of correctness of the other deductions can be increased by use of information about other chains.

Three out of the four types of chains that make up human hemoglobin molecules (there are four chains of two types per molecule), namely the α-, β-, and γ-chains, have been defined by their N-terminal sequence (Schroeder[5]). This sequence may change partially or totally during evolution, and yet the structural genes that control it may still be homologous. We need to be able to refer to homologous structural genes irrespective of the actual amino-acid sequence of their polypeptide products. Let us propose, therefore, to speak of the I^{α}, II^{β}, $\mathrm{III}^{\gamma,}$ and IV^{δ}-hemoglobin-chain genes and the corresponding polypeptide chains, the superscript representing in each case the reference for homology considerations. The common ancestor of two or more of these genes is then designated by juxtaposition of the symbols referring to the genes derived by duplication. Thus, the II^{β}–IV^{δ}-gene is the common ancestor of the II^{β}-gene and the IV^{δ}-gene; the II^{β}–III^{γ}–IV^{δ}-gene is the common ancestor of the III^{δ}-gene and the II^{β}–IV^{δ}-gene; and the I^{α}–II^{β}–III^{δ}–IV^{δ}-gene is the common ancestor of the I^{α}-gene and the II^{β}–III^{γ}–IV^{δ}-gene (Fig. 2).

Tables 1a, 1b, and 1c illustrate the procedure by presenting two partially and tentatively reconstructed ancestral hemoglobin polypeptide chains. The evidence on which these reconstructions are based are the human α-, β-, γ-, and δ-chains, the sperm-whale myoglobin chain (see Schroeder[7] for references and latest information on the structure of these chains), the horse α-chain (Braunitzer and Matsuda, in: Cullis *et al.*[8]), and the human myoglobin chain (Hill[9]). One of us (L. P.[10]) has listed earlier probable residues for 70 loci in the common precursor

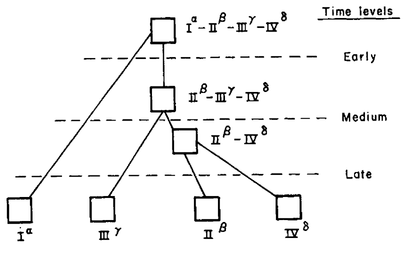
$\mathrm{I}^{\alpha}-\mathrm{II}^{\beta}-\mathrm{III}^{\gamma}-\mathrm{IV}^{\delta}$
Time levels
Early
$\mathrm{II}^{\beta}-\mathrm{III}^{\gamma}-\mathrm{IV}^{\delta}$
Medium
$\mathrm{II}^{\beta}-\mathrm{IV}^{\delta}$
Late
I^{α}
III^{γ}
II^{β}
IV^{δ}

Fig 2. Schematic representation of the successive gene duplications that are presumed to have lead to the hemoglobin genes found in man. *Cf.* Ingram[3]. See text.

Table 1 a. Tentative partial structure of two chain-ancestors of the human hemoglobin polypeptide chains. The numbering of the residues is the one usually applied to the human α-chain. The abbreviations for the amino acids are those commonly employed, except for asparagin (asg) and glutamine (glm). Other abbreviations: E = early epoch; M = medium epoch; L = late epoch; abs = absent. "None", in column (e), means: probably no evolutionrily effective mutation occurred at the site under consideration in the line of descent leading from the ancestral genes to the human genes. The residues and comments are placed in parentheses when the conclusion reached is partly based on the consideration of human and sperm whale myoglobins.

(a) Residue number
(b) Partial sequence of the II$^{\beta}$–III$^{\gamma}$–IV$^{\delta}$-chain (late form)
(c) Partial sequence of the I$^{\alpha}$–II$^{\beta}$–III$^{\gamma}$–IV$^{\delta}$-chain (late form)
(d) Chain(s) in whose direct ancestry the mutation(s) seem to have occurred
(e) Nature of the substitution
(f) Qualitative evaluation of the time of mutation
(g) From evidence relating to a number of different animals (cf. Gratzer and Allison[6])
(h) Not after the time of ancestor common to horse and man
(i) After the time of ancestor common to horse and man
(j) Amino-acid residue x present at a homologous site in the two myoglobins
(k) Polypeptide chain I$^{\alpha}$ or II$^{\beta}$–III$^{\gamma}$–IV$^{\delta}$

(a)	1	1a	2			5					10		12
(b)	val	his	leu	thr	pro	glu	asp	lys	?	?	val	thr	ala
(c)	val	?	leu	?	pro	?	asp	lys	?	?	val	?	ala
(d)	III$^{\gamma}$	(k)	III$^{\gamma}$	(k)	III$^{\gamma}$ and horse I$^{\alpha}$	(k)	II$^{\beta}$—IV$^{\delta}$				III$^{\gamma}$	(k)	III$^{\gamma}$
(e)	val→gly		leu→phe		III$^{\gamma}$: pro→glu; horse I$^{\alpha}$: pro→ala		asp→glu	none			val→ileu		ala→ser
(f)	M or L	E(g)	M or L	E or M(h)	Man: M or L Horse: (i)	E or M(h)	M					E or M(h)	M or L

(a)	13		15					20					25			28
(b)	leu	try	gly	lys	val	asg	val	abs	abs	?	glu	?	gly	gly	glu	ala
(c)	?	try	gly	lys	val	?	?	?	?	?	glu	?	gly	?	glu	ala
(d)	(k)		I$^{\alpha}$ of horse			(k)	(k)				III$^{\gamma}$			(k)		III$^{\gamma}$
(e)		none	gly→ser	none	none						glu→asp		none		none	ala→thr
(f)	E or M(h)		(i)			E or M(h)	E or M(h)				M or L			E or M(h)		M or L

Table 1 b. Continuation of Table 1 a. See Table 1 a for legend.

(a)	29	30					35					40	41
(b)	leu	gly	arg	leu	leu	val	val	tyr	pro	try	thr	glm	arg
(c)	leu	?	arg	(leu)	(phe)	?	?	?	pro	?	thr	?	?
(d)		(k)		(I^{α}) (leu→met)	$(II^{\beta}\text{-}III^{\gamma}\text{-}IV^{\delta})$ (phe→leu)	(k)	(k)	(k)		(k)		(k)	(k)
(e)	none		none	x=leu (j)	x=phe(j)				none		none		
(f)		E or M(h)		(E or M)	(E)	E or M(h)	E or M(h)	E or M(h)		E or M(h)		E or M(h)	E or M(h)

(a)	42			45		46a	47			50					54a	54b
(b)	phe	phe	(asp)	ser	phe	gly	asp	leu	ser	ser	?	ser	ala	?	met	gly
(c)	?	phe	(asp)	?	phe	?	asp	leu	ser	?	?	ser	ala	?	?	?
(d)	(k)		(I) and $(II^{\beta}\text{-}IV^{\delta})$	(k)			III^{γ}			(k)	II^{β}	$II^{\beta}\text{-}IV^{\delta}$				
(e)		none	(?→pro) (asp→glu) x=asp(j)		none		asp→asg	none	none		ser→thr	ser→asp	none			
(f)	E or M(h)		(E or M(h)) (M)	E or M(h)			M or L			E or M(h)	L	M				

(a)	54c	54d	54e	55					60					65			68
(b)	asg	pro	lys	val	lys	ala	his	gly	lys	lys	val	leu	?	ala	leu	(gly)	asp
(c)	?	?	?	val	lys	ala	his	gly	lys	lys	val	?	?	ala	leu	(gly)	?
(d)						I^{α}						(k)		III^{γ}	$II^{\beta}\text{-}IV^{\delta}$	(I^{α}) and $(II^{\beta}\text{-}IV^{\delta})$	(k)
(e)				none	none	ala→gly	none	none	none	none	none			ala→ser	leu→phe	(gly→thr) (gly→ser) x=gly(j)	
(f)						M or L(i)						E or M(h)		M or L	M	E or M(h) M	E or M(h)

(a)	69	70					75					80			83
(b)	ala	?	ala	his	leu	asp	asp	leu	lys	gly	thr	phe	ala	?	leu
(c)	ala	?	ala	his	leu	asp	asp	?	?	?	?	?	?	?	leu
(d)	$II^{\beta}\text{-}IV^{\delta}$		III^{γ}		I^{α}		$II^{\beta}\text{-}IV^{\delta}$ or II^{β}	(k)	(k)	(k)	(k)	(k)	(k)		
(e)	ala→gly		ala→lys	none	leu→val	none	asp→asg								none
(f)	M		M or L		L		M or L	E or M(h)	E or M(h)	E or M(h)	E or M(h)	E or M(h)	E or M(h)		

Table 1 C. Continuation of Table 1 a. See Table 1 a for legend.

(a)	84	85					90					95				99
(b)	ser	glu	leu	his	cys	asp	lys	leu	his	val	asp	pro	glu	asg	phe	lys
(c)	ser	?	leu	his	(ala)	?	lys	leu	?	val	asp	pro	?	asg	phe	lys
(d)		(k)			(II$^{\beta}$-III$^{\gamma}$-IV$^{\delta}$)	(k)			(k)				(k)			II$^{\beta}$-IV$^{\delta}$ or II$^{\beta}$
(e)	none		none	none	(ala→cys) (x=ala (j))		none	none		none	none	none		none	none	lys→arg
(f)		E or M(h)			(E)	E or M(h)			E or M(h)				E or M(h)			M or L

(a)	100					105					110					115
(b)	leu	leu	gly	asg	val	leu	val	?	val	leu	ala	?	his	phe	gly	lys
(c)	leu	leu	(ser)	?	?	leu	?	?	?	leu	ala	?	his	?	?	?
(d)			(II$^{\beta}$-III$^{\gamma}$-IV$^{\delta}$)	(k)	(k)		(k)		(k)					(k)	(k)	(k)
(e)	none	none	(ser→gly) (x = gly (j))			none				none	none					
(f)			(E)	E or M(h)	E or M(h)		E or M(h)		E or M(h)					E or M(h)	E or M(h)	E or M(h)

(a)	116				120					125					130	
(b)	glu	phe	thr	pro	?	val	glm	ala	ser	?	glm	lys	?	val	ala	gly
(c)	?	phe	thr	pro	?	val	?	ala	ser	?	(asp)	lys	?	?	ala	?
(d)	(k)						(k)		II$^{\beta}$-IV$^{\delta}$ or II$^{\beta}$		(II$^{\beta}$-III$^{\gamma}$-IV$^{\delta}$)			(k)	III$^{\gamma}$	(k)
(e)		none	none	none				none	ser→ala		asp→glm (x = asp (j))	none			ala→thr	
(f)	E or M(h)						E or M(h)		M or L		(E)			E or M(h)	M or L	E or M(h)

(a)	132	133		135					140	
(b)	val	ala	?	ala	leu	?	ser	lys	tyr	his
(c)	val	?	?	?	leu	?	ser	lys	tyr	?
(d)		(k)		(k)			II$^{\beta}$-IV$^{\delta}$	III$^{\gamma}$		(k)
(e)	none				none		ser→his	lys→arg	none	
(f)		E or M(h)		E or M(h)			M	M or L		E or M(h)

of α-chains and non-α-chains. A qualitative estimate is made of the time at which a given mutation supposedly occurred. (In a majority of cases the changes probably involve more than one substitution, according to the genetic codes proposed by Yukes[11] and by Smith[12]). Three periods are distinguished: early, medium, and late, according to the time off the mutation or mutations in relation to the successive gene duplications (Fig. 2).

Fossil remains no doubt express the activity of only a fraction of the genes of a given organism (although perhaps a significant fraction) and this fraction cannot be analyzed into its components. Paleobiochemistry, through molecular restoration studies on the basis of existing related polypeptide chains, provides the means of investigating the structure of such components for any part of the genome of extinct organisms. This holds, however, only in relation to structural genes, as long as the object of such studies is confined to the polypeptide products rather than extended to the genic material itself. Yet, once the structures of ancestral polypeptide chains are known, it will in the future be possible to synthesize these presumed components of extinct organisms. Thus one will be able to study the physico-chemical properties of these molecules and to make inferences about their functions. For instance, the oxygen affinity and its dependence on pH of ancestral hemoglobins might be studied as well as the affinity of ancestral enzymes for various substrates and the probable nature of these substrates in past evolutionary history. As information about various paleogenes belonging to a given group of extinct organisms will accumulate, some deductions concerning these organisms will be possible in relation to levels of biological integration higher than the level of individual macromolecules. When a fossil record is available, knowledge about the organisms concerned will go far beyond what has so far been believed possible. Important information will also be provided about forms that have left no fossil record whatsoever, such as many soft-bodied animals.

When a gene-duplication has occurred and one of the duplicate genes is not needed for carrying out the function for which its partner suffices, the polypeptide chain synthesized under its control may undergo a change in structure, including a change in spatial conformation, and the latter change especially may sometimes lead to the appearance of a new function. If the new function is retained by natural selection, it appears likely that during the subsequent evolutionary period many more mutational changes will be preserved by the corresponding gene than by genes whose functions have long been established. The latter genes are submitted to forces of selection that are conservative most of the time. Two reasons may be advanced for a more rapid alteration of the amino-acid sequence in a conformationally and functionally altered polypeptide: First, a newly evolved function is likely to be perfectible through further structural changes; second, while over very long periods of evolution the amino-acid sequence of homologous polypeptide chains can be deeply transformed even when the spatial conformation is kept nearly constant, as the comparison of myoglobin and hemoglobin suggests, a faster transformation is to be expected when the spatial conformations differ, since in that case the residues required to remain unchanged in order to preserve a given spatial conformation are not any more the same in the precursor polypeptide and in the polypeptide derived from it. Can paleobiochemical studies be anticipated to trace the filiation between genes that control polypeptides endowed with different functions? In a polypeptide whose function has been replaced by another one the amino-acid sequence may eventually be modified to such an extent that it becomes unrecognizable in terms of the molecular ancestor. Thus the comparison of the amino-acid sequences in contemporary polypeptide chains endowed with different functions probably will lead only in a minority of cases to the detection of the evolutionary relationships that actually obtain, *viz.*, mainly when the acquisition of a different molecular function by a polypeptide has been relatively recent. In such cases convincing homologies in amino-acid sequence will still be found, although the spatial conformations of the proteins may differ. Even then the discovery of polypeptides related by evolution but functionally different might appear to involve either an unlikely coincidence or a formidable task. This task could, however, be substantially reduced, if a systematic investigation were made of the amino-acid sequences in polypeptides that have been shown by genetic studies to be controlled by closely linked genes, notably by genes controlled by a common operator gene (Jacob and Monod[13]). The functions carried out by the genes included in a given operon (the unit controlled by an operator gene) can be quite diversified, and it is to be expected that the spatial conformations of the corresponding polypeptides are diverse also. Yet, since at least some genes derived through the duplication of a mother gene will probably remain closely linked during a considerable evolutionary time and since some of them may change in their function, the chances of discovering homologies between apparently unrelated proteins are likely to be greatest in a survey of the amino-acid sequences of polypeptides controlled by closely linked genes. An effort in that direction may result in a worthy contribution to the theory of evolution, in that it might show that even apparently unrelated proteins can indeed have a common molecular ancestor. Thus suggestive evidence would be furnished in support of the view that most or all apparently heterologous genes derive ultimately from a common gene-ancestor.

REFERENCES

1. Braunitzer, G., Hilschmann, N., Rudloff, V., Hilse, K., Liebold, B. and Müller, R. *Nature* **190** (1961) 480.
2. Itano, H. A. *Adv. Protein Chem.* **12** (1957) 216.
3. Ingram, V. M. *Nature* **189** (1961) 704.
4. Zuckerkandl, E. and Pauling, L. In Kasha, M. and Pullman, B. *Horizons in Biochemistry*, Academic Press, New York and London, 1962, p. 189.
5. Schroeder, W. A. *Fortschr. der Chem. Org. Naturstoffe* **17** (1959) 322.
6. Gratzer, W. B. and Allison, A. C. *Biol. Rev. Cambridge Phil. Soc.* **35** (1960) 459.
7. Schroeder, W. A. *Ann. Rev. Biochem.* **32** (1963). *In press.*
8. Braunitzer, G. and Matsuda, G. *Cf.* Cullis, A. F., Muirhead, H., Perutz, M. F. and Rossmann, M. G. *Proc. Roy. Soc. London* **A 265** (1962) 161.
9. Hill, R. *Personal communication* (1962).
10. Pauling, L., communicated at Hemoglobin Workshop, Arden House, Columbia University, Nov. 1962. *Cf.* Ingram, V. M., Richards, D. W. and Fishman, A. P. *Science* **138** (1962) 996.
11. Jukes, T. H. *Proc. Natl. Acad. Sci. U. S.* **48** (1962) 1809.
12. Smith, E. *Proc. Natl. Acad. Sci. U. S.* **48** (1962) 859.
13. Jacob, F. and Monod, J. *J. Mol. Biol.* **3** (1961) 318.

Received March 29, 1963.