Nature of amino acid substitutions in homologous proteins during evolution

A. S. Kolaskar and V. Ramabrahmam
School of Life Sciences, University of Hyderabad, Hyderabad-500 134, India
(Received 11 May 1981)

Replacement substitutions of mitochondrial cytochrome c and α- and β-chains of haemoglobin have been studied by considering the structural similarity among amino acid residues at the secondary and tertiary structural levels. Secondary structural similarity explains ~70% while tertiary structural similarity explains ~50% of observed replacements for most of the cases. These structural similarities could not account for all the replacement substitutions. The study was extended to consider the composition of codons, and the chemical nature and polarity of the replacing and replaced residues. These also could not individually account for all the affected replacements. In general, no property of amino acid residues is conserved for substitutions occurring at any single position during evolution of proteins.

Keywords: Proteins; amino acids; haemoglobin; cytochrome *c*

Introduction

The study of replacement substitutions between sequences of homologous proteins of different species has been looked to for information about evolutionary pathways[1], and a genetic basis has been given by Fitch and Margoliash[2] for these replacement substitutions. The sequence differences between homologous proteins have also been used to estimate replacement site divergence of corresponding genes by determining the minimum number of base changes needed to generate the observed amino acid replacements and multiple base changes within individual codons[3,4]. However, the limitations in detecting certain replacement substitutions precludes the accurate estimation of rates of fixation of base changes[5].

In spite of these limitations, many groups have successfully made use of these replacement substitutions in establishing molecular phylogenies and taxonomic relationships[6-17]. Thus, the question arises as to why nature has accepted only certain mutations out of the large number possible. It may be that only those mutations which have a minimal effect on the three-dimensional structure of the homologous protein are accepted. This can be seen from the crystal structure data of various cytochrome *c*[18,19] and haemoglobins[20,21]. Structural data indicate that the local three-dimensional structure undergoes little change, which probably helps to increase the efficiency of the protein, or to conserve it, in a new environment. Replacement substitutions in homologous proteins therefore provide valuable structural information.

Two levels of structural similarity among amino acid residues were considered in the present study. First, the secondary structure, which includes preference or indifference to a particular secondary structure by an amino acid residue, and second, the tertiary structure. Amino acid residues that are similar at the level of tertiary structure are termed conformationally similar, and have similar probability distributions in the (φ,ψ)-plane[22]. This structural similarity is found to be an independent property of amino acid residues.

We have studied the replacement substitutions from secondary and tertiary structure points of view in mitochondrial cytochrome *c*, and the α and β-chains of haemoglobin from various species. This study indicated that the observed replacements can not all be explained on a structural basis.

We extended the study by considering the polarity, chemical nature and composition of codons of replaced and replacing residues. Individually, none of these similarities could explain all the replacements observed in the species considered for cytochrome *c* and haemoglobin. We also looked for the conservation of specific properties of residues, such as those mentioned above, at a particular position in the sequence, where a replacement takes place. The results were inconclusive, suggesting that the substitutions accepted by nature during evolution can not be explained by considering any single property that we examined of the amino acid residues.

Method

The amino acid sequences of mitochondrial cytochrome *c*, α and β-chains of haemoglobin for the species considered (see *Tables 1* and *2* below) were taken from 'Handbook of Biochemistry and Molecular Biology'[23]. The replacement substitutions of these sequences were compared as follows. For cytochome *c*, the reference species ranged from Rhesus monkey to Pacific lamprey; for α and β-chains of haemoglobin, the reference species ranged from Rhesus monkey to kangaroo. Replacements in the sequences of other species were studied with respect to the sequence of reference species. In *Table 1* the first row corresponds to replacements observed in the sequence for human cytochrome *c* when the reference species varied

0141-8130/82/030150-05$03.00

Table 1 (a) Percentage of replacements in cytochrome *c* by residues similar in secondary structural affinity

Species studied	Reference species: Man	Rhesus monkey	Horse	Donkey	Cow	Camel	Elephant seal	Dog	Bat	Rabbit	Kangaroo	Chicken	Emu	King penguin	Pekin duck	Pigeon	Snapping turtle	Rattlesnake	Bull frog	Tuna	Bonito	Carp	Dog fish	Pacific lamprey
Man	—	100	67	55	73	70	67	73	82	67	70	85	77	85	82	83	80	93	78	85	85	92	67	75
Rhesus monkey	—	—	64	64	70	67	64	70	80	63	73	83	75	83	80	82	79	93	76	85	85	92	66	75
Horse	—	—	—	100	75	40	72	83	86	83	86	73	75	83	80	73	73	82	86	94	94	100	70	69
Donkey	—	—	—	—	50	60	57	67	71	67	75	64	67	75	70	64	64	76	79	88	88	89	65	60
Cow	—	—	—	—	—	67	60	75	83	80	86	70	73	82	78	70	70	81	83	94	93	100	65	67
Camel	—	—	—	—	—	—	25	33	60	100	67	89	78	89	86	75	75	89	91	100	100	88	69	60
Elephant seal	—	—	—	—	—	—	—	100	75	67	75	60	50	60	50	56	44	81	67	82	81	75	61	60
Dog	—	—	—	—	—	—	—	—	67	60	61	70	60	70	63	56	56	81	75	88	88	88	59	57
Bat	—	—	—	—	—	—	—	—	—	20	50	50	40	50	38	50	67	81	61	78	76	78	55	56
Rabbit	—	—	—	—	—	—	—	—	—	—	67	88	75	88	83	86	89	93	92	100	100	89	71	65
Kangaroo	—	—	—	—	—	—	—	—	—	—	—	83	80	80	80	73	82	86	77	88	88	89	75	76
Chicken	—	—	—	—	—	—	—	—	—	—	—	—	50	100	67	50	63	83	82	93	93	74	77	76
Emu	—	—	—	—	—	—	—	—	—	—	—	—	—	100	67	50	67	85	73	88	88	80	68	83
King penguin	—	—	—	—	—	—	—	—	—	—	—	—	—	—	67	50	63	85	84	94	94	91	65	79
Pekin duck	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	67	43	88	73	88	88	80	65	72
Pigeon	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	63	94	67	82	82	73	63	68
Snapping turtle	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	86	70	88	88	78	68	74
Rattlesnake	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	71	76	76	70	54	67
Bull frog	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	86	86	67	70	62
Tuna	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100	100	68	67
Bonito	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100	68	67
Carp	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	46	40
Dog fish	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	74

Table 1 (b) Percentage of replacements by conformationally similar residues in cytochrome *c*

Species studied	Reference species: Man	Rhesus monkey	Horse	Donkey	Cow	Camel	Elephant seal	Dog	Bat	Rabbit	Kangaroo	Chicken	Emu	King penguin	Pekin duck	Pigeon	Snapping turtle	Rattlesnake	Bull frog	Tuna	Bonito	Carp	Dog fish	Pacific lamprey
Man	—	100	42	27	36	30	25	27	45	33	30	31	31	31	27	33	40	50	33	35	40	38	21	20
Rhesus monkey	—	—	36	27	30	22	18	20	40	25	27	25	25	25	20	27	36	47	29	30	35	31	17	15
Horse	—	—	—	50	25	40	29	33	57	33	14	45	50	50	50	36	27	41	29	28	29	40	24	25
Donkey	—	—	—	—	0	20	14	17	43	17	25	37	42	42	40	27	18	38	21	24	25	33	18	20
Cow	—	—	—	—	—	33	20	25	50	20	29	30	36	36	33	30	20	38	25	24	25	33	12	20
Camel	—	—	—	—	—	—	0	0	40	0	17	33	33	33	29	25	13	42	27	25	27	25	13	20
Elephant seal	—	—	—	—	—	—	—	0	50	0	13	30	30	30	25	33	11	43	25	24	25	25	17	27
Dog	—	—	—	—	—	—	—	—	67	0	14	40	40	40	38	33	22	43	33	29	31	38	18	29
Bat	—	—	—	—	—	—	—	—	—	0	13	50	50	50	50	38	33	43	31	28	29	33	28	31
Rabbit	—	—	—	—	—	—	—	—	—	—	17	38	38	38	33	29	11	50	27	25	31	22	12	18
Kangaroo	—	—	—	—	—	—	—	—	—	—	—	50	60	50	50	45	36	48	46	41	47	56	25	35
Chicken	—	—	—	—	—	—	—	—	—	—	—	—	0	0	67	50	38	53	45	31	38	27	37	22
Emu	—	—	—	—	—	—	—	—	—	—	—	—	—	0	67	50	50	45	45	31	38	30	37	22
King penguin	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100	75	50	50	50	35	41	36	40	26
Pekin duck	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	67	43	41	45	31	38	30	29	22
Pigeon	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	38	44	50	35	41	36	32	21
Snapping turtle	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	50	30	24	25	22	37	26
Rattlesnake	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	33	40	38	30	27	22
Bull frog	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	36	43	22	30	24
Tuna	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	50	33	32	28
Bonito	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	33	32	28
Carp	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	31	30
Dog fish	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	25

Table 2 (a) Percentage of replacements by residues similar in secondary structural affinity

	Haemoglobin α-chain									Haemoglobin β-chain								
Reference species	Man	Rhesus monkey	Mouse NB	Rabbit	Dog	Sheep	Pig	Horse, slow	Reference species	Man	Rhesus monkey	Mouse C57BL	Rabbit	Dog	Sheep	Pig	Horse	Elephant
Kangaroo	75	74	73	78	76	73	79	80	Kangaroo	74	78	70	78	74	70	73	70	82
									Elephant	69	63	73	73	70	82	67	68	—
Horse, slow	50	56	58	68	44	50	62	—	Horse	64	71	67	68	70	68	73	—	—
Pig	42	42	57	55	39	40	—	—	Pig	20	41	67	21	33	67	—	—	—
Sheep	59	63	56	70	47	—	—	—	Sheep	48	54	69	52	61	—	—	—	—
Dog	70	75	72	75	—	—	—	—	Dog	53	73	73	57	—	—	—	—	—
Rabbit	68	76	62	—	—	—	—	—	Rabbit	50	63	76	—	—	—	—	—	—
Mouse NB	79	72			—	—	—	—	Mouse C57BL	71	82	—	—	—	—	—	—	—
Rhesus monkey	50	—	—	—	—	—	—	—	Rhesus monkey	63	—	—	—	—	—	—	—	—

Table 2 (b) Percentage of replacements by conformationally similar residues

	Haemoglobin α-chain									Haemoglobin β-chain								
Reference species	Man	Rhesus monkey	Mouse NB	Rabbit	Dog	Sheep	Pig	Horse, slow	Reference species	Man	Rhesus monkey	Mouse C57BL	Rabbit	Dog	Sheep	Pig	Horse	Elephant
Kangaroo	32	30	37	35	36	20	27	27	Kangaroo	34	33	32	43	40	42	50	40	36
									Elephant	27	27	30	36	33	30	24	24	—
Horse, slow	17	25	38	52	33	25	23	—	Horse	32	29	42	44	30	18	20	—	—
Pig	17	25	33	50	30	27	—	—	Pig	47	41	48	64	53	33	—	—	—
Sheep	9	5	24	37	20	—	—	—	Sheep	37	32	44	48	46	—	—	—	—
Dog	30	38	36	32	—	—	—	—	Dog	20	20	30	33	—	—	—	—	—
Rabbit	44	44	38	—	—	—	—	—	Rabbit	36	25	31	—	—	—	—	—	—
Mouse NB	42	44	—	—	—	—	—	—	Mouse C57BL	29	21	—	—	—	—	—	—	—
Rhesus monkey	25	—	—	—	—	—	—	—	Rhesus monkey	50	—	—	—	—	—	—	—	—

from Rhesus monkey to Pacific lamprey. Similarly, when examining replacements in the cytochrome *c* sequence of cow, the reference species varied from camel to Pacific lamprey. In the haemoglobin comparisons, the reference species are given in separate columns in *Table 2* for α and β-chains. The amino acid residue in the sequence of the reference species which can be replaced is termed the replaced residue and the residue which occupies that position in the sequence of other species is the replacing residue.

Each replacement was studied by taking into consideration the following characteristics of amino acid residue preferences and similarities.

Secondary structure preference

Levitt's classification of secondary structural preference of amino acid residues was used[24], since this classification has no overlap and clearly states whether a particular amino acid residue favours, is indifferent to or breaks a particular secondary structure. A replacing residue that prefers or is indifferent to the secondary structure preferred by the replaced residue is considered to have similar secondary structural preference.

Conformational similarity

Conformationally similar residues were taken from Table 2 of our previous paper[22]. This table gives a set of amino acid residues having a high probability to adopt similar main chain conformations. Thus, if the replacing residue is conformationally similar to the replaced residue, a minimal change in the tertiary structure is expected.

Composition of codons

Residues that have two bases in common, irrespective of the position in respective codons, were considered to be similar in this study.

Chemical nature

Classification of the chemical nature of residues was adapted from Stryer[25].

Polarity

The residues are divided as hydrophobic, neutral polar or polar, as described by Lehninger[26].

A simple algorithm was developed to study these

features of replacement substitutions in mitochondrial cytochrome *c* and the α and β-chains of haemoglobin in different species.

Results and discussion

Tables 1 and *2* give the percentage of observed replacement substitutions for respective comparisons when a particular property or similarity of amino acid residues is considered. For example, the value 67 in the third column of the first row of *Table 1a* signifies that if the sequence of human cytochrome *c* is compared with that of horse as the reference species, 8 out of 12 replacements (67%), are such that the replacing residues have similar secondary structural affinity to the replaced residues. Similar results are given in *Table 2a* for the α and β-chains of haemoglobin. These Tables show that similarity in secondary structural affinity of the replacing and replaced residues explains between 100 and 20% of replacement substitutions, although for most cases the percentage is around 65–75. In this study, we considered substitutions throughout the polypeptide chain, so that the results should not be interpreted as meaning that the secondary structural regions are conserved. Indeed, study of the crystal structure of Tuna and Bonito cytochrome *c* indicates that the secondary structural region is not the same in both species. Tuna cytochrome *c* consists of five α-helices while Bonito cytochrome *c* consists of three α-helices and two 3_{10} helices. At position 61, which lies in α-helical region of both proteins, the residue was Asn in Tuna and Gln in Bonito; Asn is an α-helix breaker and Gln is an α-helix preferrer. Thus, the secondary structural regions differ, and the replacing and replaced residues do not prefer the same secondary structure.

We then examined the replacement substitutions by conformationally similar residues. The results are given in *Tables 1b* and *2b* for cytochrome *c*, and α and β-chains of haemoglobin, respectively. Similarity of the tertiary structure explains a considerable number of replacements; the percentage of replacements explained varies between 20 and 60. Analysis of crystal structure data indicates that almost all the replacing residues can adopt a similar main chain conformation to that of the replaced residue. This conclusion is derived from our observation that when the (φ,ψ)-maps are compared the replacing and replaced residues have very low discrepancy values between the grids which represent these observed conformations. Thus, the replacing amino acid residues are those which have a minimal effect on the three-dimensional structure of the protein.

Since none of the two structural similarities could fully explain the replacements, we studied the replacement substitutions from the other three aspects mentioned in the Method to see whether consideration of any one of these properties could explain the replacements.

When the composition of codons of replacing and replaced residues was considered, i.e. genetic similarity, it could not explain all the replacements, only ~70%. Similarity in chemical nature or polarity explained ≤50% of replacements*.

Since no single feature could explain all the replacements, we examined the data to see whether any of these properties is conserved in a position affected by replacements during evolution. For example, residue 89 of the cytochrome *c* sequence for different species, going from Pacific lamprey to man is Gly, Ser, Gly, Gly, Gly, Gly, Lys, Ala, Ala, Ser, Ser, Ser, Ser, Gly, Asp, Ala, Gly, Gly, Gly, Gly, Thr, Thr, Glu, Glu. Several observations can be made for this example.

(1) The residues Ser, Asp and Gly prefer chain reversals, Ala, Lys and Glu prefer α-helix and Thr prefers β-sheet. Gly is an α-helix breaker and Ala is a breaker of chain reversals. Thus, similarity at the secondary structural level is not maintained.

(2) Conformational similarity is not preserved. Lys, Glu, Ala, Asp, Ser, Thr are all conformationally different from Gly.

(3) Codons for Lys and Thr differ from those for Gly by not having two bases in common, irrespective of position in the codon. Thus, the codons for the replacing and replaced residues are not similar.

(4) Amino acids having different chemical natures occur at the same position: Ala and Gly are aliphatic, Thr and Ser are hydroxy amino acid residues, Asp and Glu are acidic and Lys is basic.

(5) The polarity of the residue varies: Gly, Ser and Thr are neutral polar, Asp, Glu and Lys are polar, and Ala is hydrophobic.

Thus, none of these properties is conserved at residue 89 of cytochrome *c*. Similar observations are made at other positions in the cytochrome *c* sequence as well as the α and β-chains of haemoglobin. During evolution, therefore, the emphasis seems to be on conservation of the three-dimensional structure of homologous proteins rather than conservation of local structure, chemical nature or polarity. This may reflect species-characteristic functions of these homologous proteins in addition to their common function.

The only partial explanation of replacements between sequences of any two species by individual properties of amino acid residues, and the non-conservation of any of these properties in replacement substitutions during evolution, shows that this type of study alone cannot provide enough information to establish an evolutionary pathway. Phylogenetic trees constructed using data from *Tables 1* and *2* and methods such as those developed either by Dayhoff[27], or the modified method[28], may provide a better basis. Generation of such phylogenetic trees is in progress, and will include not only the amino acid substitutions but will also indirectly account for the additional functions characteristic of a given species, incorporated during evolution.

This study also points to the need for such studies, to aid in understanding whether micro-evolution and macro-evolution are in phase, i.e. to establish whether the changes observed in proteins, nucleic acids and hormones correspond one to one with morphological changes of species.

* Data available from the publisher.

References

1 Schwartz, R. M. and Dayhoff, M. O. *Science* 1978, **199**, 395
2 Fitch, W. M. and Margoliash, E. in 'Evolutionary Biology', (Eds T. Dobzhansky, M. K. Hecht and W. C. Steere), Appleton-Century-Crofts, Educational Division/Meredith Corporation, New York, 1970, Vol. 4, pp. 67–109
3 Wilson, A. C., Carloson, S. S. and White, T. J. *Annu. Rev. Biochem.* 1977, **46**, 573

4 Dickerson, R. E. and Geis, I. in 'Proteins: Structure, function and evolution', Menlo Park, Benjamin/Cummings Publishing, California, 1980
5 Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Connell, C., Spritz, R. A., De Riel, J. K., Forget, B. G., Weissmann, S. M., Slightom, J. L., Blechl, A. E., Smithies, O., Baralle, F. E., Shoulders, C. C. and Proudfoot, N. J. *Cell* 1980, **21**, 653
6 Margoliash, E. *Proc. Natl. Acad. Sci. USA* 1963, **50**, 672
7 Smith, E. L. and Margoliash, E. *Fed. Proc.* 1964, **23**, 1243
8 Smith, E. L. and Margoliash, E. *Fed. Prc.* 1964, **23**, 1276
9 Margoliash, E. and Smith, E. L. in 'Evolving Genes and Proteins' (Eds V. Bryson and H. J. Vogel) Academic Press, New York, 1965, p. 221
10 Fitch, W. M. and Margoliash, E. *Science* 1967, **155**, 279
11 Nolan, C. and Margoliash, E. *Annu. Rev. Biochem.* 1968, **37**, 727
12 Dayhoff, M. O. *Sci. Am.* July 1969, 86
13 Smith, E. L. in 'The Enzymes' (Ed. P. D. Boyer) Academic Press, New York, 1970, Vol. 1, pp. 267–339
14 Dickerson, R. E. *J. Mol. Evol.* 1971, **1**, 26
15 Williams, J. in 'Chemistry of Macromolecules' MTP International Review of Science, Biochemistry Series One, (Ed. H. Gutfreund) Butterworths, London, 1974, Vol. 1, pp. 1–56
16 Dayhoff, M. O., McLaughlin, P. J., Barker, W. C. and Hunt, L. T. *Naturwissenschaften* 1975, **62**, 154
17 Barnabas, J., Mathew, P. A., Ratna Parikhi, M. V. and Barnabas, S. *Ind. J. Biochem. Biophys.* 1978, **15**, 388
18 Dickerson, R. E. and Timkovich, R. in 'The Enzymes' (Ed. P. D. Boyer) Academic Press, New York, 1975, Vol. 11, pp. 397–547
19 Srinivasan, A. R. *Int. J. Pept. Protein Res.* 1980, **16**, 111
20 Perutz, M. F. *J. Mol. Biol.* 1965, **13**, 646
21 Perutz, M. F., Kendrew, J. C. and Watson, H. C. *J. Mol. Biol.* 1965, **13**, 669
22 Kolaskar, A. S. and Ramabrahmam, V. *Int. J. Biol. Macromol.* 1981, **3**, 171
23 'Handbook of Biochemistry and Molecular Biology, Proteins III', (Ed. G. D. Fasman), CRC Press, 1976
24 Levitt, M. *Biochemistry* 1978, **17**, 4277
25 Stryer, L. in 'Biochemistry', W. H. Freeman, San Francisco, 1975
26 Lehninger, A. L. in 'Biochemistry', 2nd Ed., Worth, New York
27 'Atlas of Protein Sequence and Structure', (Ed. M. O. Dayhoff) National Biomedical Research Foundation, Washington, D. C., 1972
28 Barnabas, J., Goodman, M. and Moore, G. W. *J. Mol. Biol.* 1972, **69**, 249