J Mol Evol (1990) 31:33–39

Journal of
Molecular Evolution

Extreme Differences in Charge Changes during Protein Evolution

Jack A.M. Leunissen,* Henno W. van den Hooven, and Wilfried W. de Jong

Department of Biochemistry, University of Nijmegen, PO Box 9101, 6500 HB Nijmegen, The Netherlands

Summary. The maintenance of a proper distribution of charged amino acid residues might be expected to be an important factor in protein evolution. We therefore compared the inferred changes in charge during the evolution of 43 protein families with the changes expected on the basis of random base substitutions. It was found that certain proteins, like the eye lens crystallins and most histones, display an extreme avoidance of changes in charge. Other proteins, like phospholipase A2 and ferredoxin, apparently have sustained more charged replacements than expected, suggesting a positive selection for changes in charge. Depending on function and structure of a protein, charged residues apparently can be important targets for selective forces in protein evolution. It appears that actual biased codon usage tends to decrease the proportion of charged amino acid replacements. The influence of nonrandomness of mutations is more equivocal. Genes that use the mitochondrial instead of the universal code lower the probability that charge changes will occur in the encoded proteins.

Key words: Amino acid replacements — Codon usage — Electrostatic interactions — Mitochondrial code — Molecular evolution — Mutations

Introduction

Evolutionary change in proteins is mainly governed by two principles: (1) functionally less important proteins, or parts of proteins, evolve faster than more important ones, and (2) conservative changes, which do not appreciably disrupt the structure or function of a protein, are more readily accepted in evolution than more disruptive ones (Zuckerkandl and Pauling 1965; Ohno 1970; Dickerson 1971; Zuckerkandl 1975; Wilson et al. 1977; Doolittle 1979; Kimura 1983). Similar amino acids are interchanged more easily than chemically dissimilar ones (Clarke 1970; Grantham 1974; Dayhoff et al. 1978).

Many aspects of protein structure and function are dominated by electrostatic interactions (Perutz 1978; Barlow and Thornton 1983; Warshel and Russell 1984; Matthews 1985; Honig et al. 1986). The maintenance of a proper distribution of charged amino acid residues might therefore be an important factor in protein evolution. The actual evolutionary changes in charge properties of a protein will depend on its structural and functional requirements. A comparison of the evolutionary charge changes in a wide variety of proteins may thus reveal important aspects of molecular evolution. Such an endeavor has been made for a small number of proteins by Peetz et al. (1986). They observed that changes in charge in the evolution of cytochrome *c* and fibrinopeptides have occurred to the extent that would be expected on the basis of random base substitutions. These proteins are apparently not subject to noticeable selective forces acting on charged residues. Globins and insulin, on the other hand, apparently sustained significantly fewer changes in charge than expected.

It should be possible to disclose in much more detail the patterns and principles of charge changes in protein evolution by exploiting the extensive data sets of homologous proteins presently accumulated in the data bases. We thus analyzed the evolutionary changes in charge in 43 sets of homologous proteins, with widely different structural and functional properties. For each of the 43 sets of sequences we cal-

* *Present address:* CAOS/CAMM Center, University of Nijmegen, Toernooiveld, 6525 ED Nijmegen, The Netherlands
Offprint requests to: J.A.M. Leunissen at his present address

culated the proportion of amino acid replacements involving a change in charge, expected to have occurred in the absence of selective constraints, and the proportion inferred to have occurred during the actual process of divergent evolution. Certain proteins display strong avoidance of changes in charge during evolution, whereas some others appear to have accepted such changes even more readily than expected. Many proteins in the sample, however, reveal only moderate or no selective forces acting on charge changes.

Methods

Selection and Preparation of Data Sets. Sets of homologous sequences were extracted from the NBRF/PIR protein sequence data base (release 14.0). Data sets, with the exception of sperm histone and αB-crystallin, were selected that contained at least five full-length sequences. All sequences within a set were optimally aligned according to the method of Lipman and Pearson (1985); the number of gaps was then minimized, and their location optimized, by manual editing using a multiple-sequence alignment editor (SALE, written by J.A.M.L.). Highly variable N- and C-terminal extensions that could not satisfactorily be aligned were excluded from the data, as well as sequences containing many poorly determined residues.

Calculation of Expected Values. The expected values (EV) indicate the proportion of amino acid replacements that would be expected to cause a charge change, under the assumption of unbiased codon usage and random nucleotide substitutions, and in the absence of selective constraint. The chance fa_i for any amino acid a_i to undergo a change in charge, assuming that all nucleotide substitutions are equally probable, is given by

$$fa_i = \frac{\sum_{j=1}^{n_i} (L_{ij} \times g_{ij})}{\sum_{j=1}^{n_i} (R_{ij} \times g_{ij})} \quad (1)$$

where L_{ij} is the number of possible point mutations in codon c_{ij}, resulting in a charge change, R_{ij} the number of point mutations in codon c_{ij} that result in an amino acid replacement, n_{ij} the number of codons for amino acid a_i, and g_{ij} the relative frequency of codon c_{ij}. From Eq. (1) the expected value for any amino acid sequence follows according to

$$P(L|R) = \frac{\sum_{i=1}^{20} p_i \times \sum_{j=1}^{n_i} (L_{ij} \times g_{ij})}{\sum_{i=1}^{20} p_i \times \sum_{j=1}^{n_i} (R_{ij} \times g_{ij})} \quad (2)$$

where p_i is the frequency of occurrence of amino acid a_i. Assuming that all synonymous codons are equally frequent, g_{ij} reduces to $1/n_i$.

From this general formula, the EV for a set of homologous protein sequences can be calculated in several ways: (1) As the EV of the ancestral sequence: the last common ancestor of a protein family is reconstructed using a modification of the method of Dayhoff (Dayhoff et al. 1972; Leunissen, unpublished). Because it is not always possible to construct a reliable phylogeny for a set of sequences, this method could only be applied in a limited number of cases. (2) As the mean value of the EVs of all individual sequences in a data set: this method has the advantage that the standard deviation can be calculated. (3) All sequences in the data set were combined into one large sequence, of which the EV was calculated. This gives the EV of the average protein for a set of sequences.

Comparable EV values were obtained by all three methods; the mean difference between these methods was <1%. The values displayed in Tables 1 and 2 were calculated using the second method (i.e., mean value). It may be noted that our model necessarily only takes into account codon changes resulting from single nucleotide substitutions, whereas during actual evolution an amino acid replacement often corresponds with more than one nucleotide substitution. However, the validity of our approach is supported by the use of actual nucleotide sequences, as shown in Table 2.

Calculation of Observed Values. The observed value (OV) for any pair of sequences was calculated as the quotient of all observed charge changes and the total number of replacements. The OV for a protein family is then calculated as the mean of the OVs resulting from all pairwise comparisons of the sequences in a data set. In a limited number of cases the last common ancestor of the sequences in the set was first reconstructed; the OV was then calculated as the mean value of the observed charge change fractions for all individual contemporary sequences as compared with their ancestor. This value correlated well with the value obtained by the former method.

Correction for Biased Codon Usage and Nonrandom Base Substitutions. The expected value can be corrected for biased codon usage by using the actual codon frequencies for variable g_{ij} in Eq. (2). These codon frequencies were obtained from corresponding nucleic acid sequences in the EMBL nucleotide data bank.

Correction for nonrandom base substitutions was obtained by recalculation of the values L_{ij} and R_{ij}, using the nucleotide substitution probabilities as given by Li et al. (1984).

Results

Expected Changes in Charge

The purpose of the present work was to determine to what extent the actual changes in charge in the evolution of different proteins deviate from expectation. The difference between the expected and actually observed values is an indication of the selective constraints working at the level of protein charge. For a given protein the expected proportion of amino acid replacements leading to changes in charge can easily be estimated if we assume absence of selective constraints, random base substitutions, and unbiased codon usage. This value then simply depends on the amino acid composition, as each amino acid has a given chance to be involved in a charge change when a random mutation hits a gene.

For families of related proteins, as we have used, there are different approaches to calculate the expected values for charge changes in that family. Ideally, one would like to know the sequence of the last common ancestor of the proteins in a family. From that ancestral sequence the expected changes in charge could then prospectively be estimated. However, reconstructing ancestral sequences from pres-

Table 1. Charge changes in the evolution of proteins

Protein family	EV	OV	OV/EV	L	ChR/100	Ch/100	Hy	PAM	*n*
αA-crystallin	0.39	0.03	0.08	172.0	26.1	−2.9	−0.49	5.0	69
Histone H4	0.43	0.07	0.16	96.8	31.5	17.7	−0.55	0.1	6
Histone H2B	0.38	0.07	0.18	98.4	28.2	11.9	−0.56	0.9	14
αB-crystallin	0.36	0.08	0.21	175.5	27.5	−0.7	−0.49	1.5	4
Histone H3	0.41	0.09	0.23	132.7	31.4	15.1	−0.63	0.1	6
γ-crystallin	0.37	0.11	0.29	169.3	25.8	0.0	−0.81		17
Actin	0.34	0.12	0.34	365.5	23.1	−3.2	−0.21		13
Cytochrome *c* oxidase I[c]	0.24	0.09	0.39	511.3	7.9	−1.2	0.73		7
Cytochrome *c* oxidase III[c]	0.23	0.11	0.46	260.6	7.4	−1.8	0.55		8
Cytochrome *b*[c]	0.24	0.11	0.48	380.2	8.9	0.3	0.73		9
Rib. bisph. carb. L[a]	0.36	0.22	0.63	474.5	23.7	−1.1	−0.26		6
Hemoglobin α	0.32	0.20	0.64	138.8	19.8	1.0	−0.04	12.0	96
Histone H2A	0.39	0.26	0.67	121.9	25.0	12.2	−0.32	0.5	12
Myoglobin	0.38	0.26	0.68	152.6	28.3	0.5	−0.43	8.5	64
Calmodulin	0.41	0.28	0.68	148.3	35.4	−16.3	−0.64		6
β-crystallin	0.37	0.25	0.69	139.6	23.9	−1.0	−0.78		11
Metallothionein	0.32	0.22	0.71	60.9	18.3	7.2	0.08		8
Cytochrome *c* oxidase II[c]	0.29	0.21	0.74	229.7	16.1	−5.9	0.26		9
Hemoglobin β	0.34	0.26	0.78	145.6	20.8	0.1	−0.03	12.0	97
Sperm histone	0.56	0.45	0.81	49.3	54.8	54.8	−1.96		4
Troponin *c*	0.43	0.36	0.85	159.3	38.7	−18.4	−0.61	1.5	6
Leghemoglobin	0.34	0.29	0.87	146.0	22.4	−1.5	0.03		9
Fibrinopeptide A	0.43	0.38	0.88	16.3	32.6	−17.2	−0.61	29.4	20
Cytochrome *b*5	0.38	0.36	0.94	119.7	28.3	−8.0	−0.67	4.5	7
Fibrinopeptide B	0.48	0.47	0.97	17.0	47.5	−20.4	−1.32	45.5	13
Dihydrofolate reductase	0.38	0.37	0.99	186.6	28.0	0.3	−0.47		5
Parvalbumin α	0.41	0.41	1.00	109.2	35.3	−5.0	−0.35	7.0	6
Carbonic anhydrase	0.35	0.35	1.01	259.4	21.8	−0.6	−0.57	12.5	8
Cytochrome *c*	0.39	0.39	1.01	107.0	25.8	6.1	−0.68	2.2	77
Lysozyme *c*	0.36	0.37	1.01	129.1	21.3	6.6	−0.52	9.8	14
RNAse pancreatic	0.34	0.34	1.02	124.7	19.5	2.4	−0.73	21.7	20
ATPase *b*	0.35	0.37	1.07	463.1	22.5	−2.8	−0.07		7
G3PDH[b]	0.35	0.37	1.08	330.7	22.3	0.2	−0.06	2.2	6
Superoxide dismutase	0.36	0.39	1.08	152.2	21.3	−3.5	−0.33		12
Parvalbumin β	0.38	0.43	1.12	108.3	30.9	−5.8	−0.12	7.0	10
Cytochrome *c*6	0.35	0.39	1.12	84.6	20.3	−1.9	−0.23		13
Plastocyanin	0.33	0.37	1.14	99.3	19.4	−8.3	−0.09	3.5	15
α-lactalbumin	0.38	0.45	1.20	122.9	27.8	−5.8	−0.47	21.7	8
Rib. bisph. carb. S[a]	0.35	0.42	1.21	122.4	24.2	−1.3	−0.34		7
Phospholipase A2	0.36	0.45	1.27	120.7	22.7	0.7	−0.56	19.0	43
Long neurotoxin	0.34	0.44	1.28	71.5	22.2	5.0	−0.43	55.6	21
Ferredoxin	0.33	0.43	1.30	56.6	23.0	−16.4	0.20	1.9	14
Rubredoxin	0.39	0.52	1.35	52.6	30.4	−16.0	−0.48		5

Listed are the protein data sets, the expected value (EV), observed value (OV), the ratio OV/EV, the average chain length (L), percentage charged residues (ChR/100), net charge per 100 residues (Ch/100), hydrophobicity (Hy) (Kyte and Doolittle 1982), the evolutionary rate of change (PAM), and the number of sequences in a data set (*n*). The evolutionary rate, where available, is expressed as the number of replacements per 100 residues per 100 million years, corrected for multiple replacements at the same site; values were taken from Wilson et al. (1977), Dayhoff et al. (1978), Stapel et al. (1985), and Peterson and Piatigorsky (1986)

[a] Ribulose bisphosphate carboxylase large and small chain
[b] Glyceraldehyde 3-phosphate dehydrogenase
[c] Mitochondrially encoded protein

ent-day proteins introduces additional assumptions and uncertainties and requires knowledge of the topological relationships between the involved sequences. The other approaches directly use the actual sequences in the data set, in slightly different manners.

In Table 1 the EVs are given that are obtained as the mean of the EVs of all individual sequences in the data set. This method has the advantage that a standard deviation can be calculated. The alternative methods all give EVs that are very similar to the values shown in Table 1, the mean difference between the methods being less than 1%. The EVs in Table 1 range from a low value of 0.23 for cytochrome *c* oxidase III to a high value of 0.56 for sperm histone. The values correlate well, as expected, with the fraction of charged residues in the proteins.

Table 1 includes four proteins that are encoded by the mitochondrial genome. For these proteins the EVs have been calculated using the appropriate mitochondrial codes. Interestingly, when a gene uses the mitochondrial instead of the universal code, it lowers the probability that charge changes will occur by approximately 10%. For cytochrome *b* the EVs based on the mitochondrial and the universal code are 0.216 and 0.239, respectively, and for cytochrome *c* oxidase I, II, and III these values are 0.208 and 0.237, 0.268 and 0.288, and 0.208 and 0.233, respectively.

Corrections for Codon Preference and Nonrandom Base Substitutions

The calculation of the expected changes in charge assumes necessarily an oversimplified model of unbiased codon usage and random base substitutions. Actual usage of synonymous codons shows a considerable bias, which varies between organisms (Grantham et al. 1986; Maruyama et al. 1986), but also among genes within a species (e.g., Sharp et al. 1988). Biased codon usage obviously influences the chance that mutations result in changes in charge. This influence is considerable in the extreme cases that a gene would exclusively use only those synonymous codons that give a maximum or a minimum fraction of charge changes upon mutation. Although the expected value for charge changes in a protein with average amino acid composition (Dayhoff et al. 1978), making equal use of synonymous codons, is 0.355, this value becomes 0.403 or 0.314 when codons are used that give maximum or minimum charge changes, respectively. The actual codon usage is not known for most of the proteins in the data bases used for Table 1. To assess the possible influence of biased codon usage, we compared for a small number of proteins the EV values under the assumption of random codon usage with the values obtained using their actually known codon usage. Table 2 reveals a slight decrease for most expected values under biased codon usage. This is confirmed when the average biased codon usage in human, mouse, chicken, and *Drosophila* genes (Grantham et al. 1986) is applied to EV calculations for the average protein composition (Dayhoff et al. 1978, p. 363). This yields values of 0.351, 0.352, 0.351, and 0.347, respectively, as compared with EV = 0.355 for unbiased codon usage.

Also, the assumption of random base substitution is not in agreement with the actual evolutionary processes. Directional trends, still poorly understood, are clearly present in molecular evolution (Perrin and Bernardi 1987; Preparata and Saccone 1987; Bernardi et al. 1988; Sueoka 1988). From the analysis of pseudogene sequences, which are supposedly not subject to functional constraints, it appears that the pattern of spontaneous point mutation deviates considerably from randomness (Li et al. 1984). Most notably, transitional mutations occur almost twice as frequently as expected under random mutation. This nonrandomness also influences the proportion of charged amino acid replacements expected to occur. Li et al. (1984) estimate that nonrandom mutation tends to reduce the proportion of charge changes by 9% (where histidine is considered to be positively charged). The nonrandom mutation pattern as obtained by Li et al. has been applied to some of the protein families in our data sets (Table 2). It appears that nonrandom base substitutions in some cases decrease the expected number of charge changes and in other cases increase this number. Because the estimate of the spontaneous mutation pattern is based on a limited number of mammalian pseudogenes, the observed trends may not be generally applicable. We therefore preferred to use the model of random base substitutions to calculate the EV values in Table 1, although we recognize that nonrandomness may considerably influence these values, in either direction.

Table 2. Influence of biased codon usage and nonrandom base substitutions on the expected charge changes in protein evolution

Protein family	EV_u	EV_c	Diff%	*n*
A) Influence of biased codon usage on EV				
αA-crystallin	0.352	0.337	−4.3	4
β-crystallin	0.353	0.348	−1.4	4
γ-crystallin	0.370	0.369	−0.2	15
Histone H4	0.426	0.406	−4.7	7
Cytochrome *c*	0.403	0.395	−2.0	6
B) Influence of nonrandom base substitutions on EV				
αA-crystallin	0.348	0.344	−1.2	69
αB-crystallin	0.355	0.341	−4.1	4
β-crystallin	0.366	0.386	4.9	11
γ-crystallin	0.369	0.387	4.7	17
Histone H4	0.432	0.435	0.7	6
Hemoglobin α	0.321	0.291	−10.3	96
Cytochrome *c*	0.385	0.370	−4.1	77

Listed are the protein data sets, the uncorrected expected value (EV_u), the corrected expected value (EV_c), the percentage difference between both values (Diff%), and the number of sequences in the set (*n*)

Observed Changes in Charge

The actual amino acid replacements that have occurred in the evolution of a protein family can only be inferred from comparisons of present-day sequences and are therefore deemed to remain uncertain. The only practically feasible way to infer the charge changes that have occurred since the divergence of two homologous proteins is to compare their aligned sequences position by position and

count the number of charge changes that are observed relative to the total number of amino acid replacements. The numbers and types of replacements can reliably be determined, under the assumption of parsimony, when few differences exist between the sequences. With increasing evolutionary distance, multiple superimposed replacements will diminish the accuracy of the inferred number and nature of amino acid replacements. This does not, however, appreciably influence the observed fraction of charged replacements. It appears, indeed, that in the largest data sets (cytochrome *c*, hemoglobin α and β) the fractions of charged replacements determined by comparison of sequences with more versus those with less sequence divergence are only marginally different.

The proportion of amino acid replacements involving a change in charge in the evolution of a protein family (OV in Table 1) was calculated as the average of the OVs of all sequences in the data set. The OV for each individual sequence was determined by pairwise comparison with all other sequences in the data set. The observed values in Table 1 show that in αA-crystallin only 0.03 of the amino acid replacements involve a change in charge, whereas this is 0.52 in rubredoxin. The observed values thus display a much greater variation between proteins than the expected values. The ratio OV/EV is a convenient measure to assess the deviation of the actual frequency of charged amino acid replacements from expectation for each protein family. The OV/EV values in Table 1 then are a measure of the constraints acting on changes in charge in different protein families. It is immediately obvious that extreme differences exist in the extent to which proteins are allowed to accept charge changes in evolution.

Discussion

The positively charged lysine and arginine residues and the negative aspartic and glutamic acid residues form a special category of amino acid side chains in proteins. These residues, being very polar, are mostly located at the surface of proteins (Rose et al. 1985; Miller et al. 1987). In monomeric globular proteins these residues constitute on average 27% of the protein surface and only 4% of the interior (Miller et al. 1987). Charged groups are not distributed randomly on the surface of proteins; they are usually surrounded by charges of opposite sign (Wada and Nakamura 1981). One-third of charged residues in a protein are on average involved in intramolecular ion pairs (or salt bridges) (Barlow and Thornton 1983). Charged groups on the surface of proteins may be of general importance for recognition and interactions between molecules, whereas ion pairs stabilize the tertiary structure and contribute to the thermostability of proteins (Perutz 1978; Barlow and Thornton 1983; Warshel and Russell 1984). Insight into electrostatic effects is an essential requirement for successful protein engineering (e.g., Blundell et al. 1987; Sternberg et al. 1987), and a knowledge of the evolutionary behavior of charged residues may be of considerable help in this respect.

If all codons occurred with equal frequency in genes, 8/61 or 13.1% of amino acids in an average protein would be basic and only 4/61 or 6.6% acidic. In this case 32.6% of the amino acid replacements caused by random base substitution would be expected to cause a change in charge (Nei 1975). In a pool of 314 sequences from different protein families 11.5% of residues was earlier found to be actually acidic and precisely the same percentage was basic (Dayhoff et al. 1978). In the most recent release of the NBRF/PIR data base (release 19.0), including 10,527 sequences, these percentages are 11.4% and 11.1%, respectively. Not considering the unpredictable charge contribution of histidine residues, it thus appears that the average protein tends to have a neutral net charge at physiological pH. From the average amino acid composition of proteins one expects that 35.5% of the amino acid replacements would change the charge, assuming unbiased codon usage and random base substitutions. On the other hand, based on an accumulation of 1572 accepted point mutations deduced from ancestral sequence reconstructions of many closely related sequences, one can see that 27.9% of the replacements involve a change in charge (Dayhoff et al. 1978, p. 346). This is an indication that selective constraints acting during the evolution of proteins result in a certain avoidance of charge changes. Considering the widely varying functional role and importance of charged residues in different proteins and in different parts of proteins, one expects considerable differences in the acceptance of charge changes in different protein families.

Earlier studies indicated that the evolutionary variability of polar and nonpolar amino acids does not differ significantly (Vogel and Zuckerkandl 1971). Also, from a comparison of tertiary structures of different globins, of lysozymes, and of serine proteases, it was concluded that in these families charged interactions between ion pairs are poorly conserved, unless the residues involved have more specific functions to perform (Barlow and Thornton 1983). Yet, Peetz et al. (1986) calculated that globins accumulate charge changes at rates slower than those predicted by a model of random substitutions. Also, insulin was found to accept fewer charge changes than expected, but cytochrome *c* and fibrinopeptides

accumulated charge changes as predicted by a random model. The present analysis of 43 protein families shows extreme differences in the extent to which proteins accept charge changes in evolution. No simple and general parameter, like molecular size, content of charged residues, overall charge, variation in overall charge, or hydropathy, was found to correlate with the variation in OV/EV values (Table 1). Also, a comparison with estimated rates of evolution (Wilson et al. 1977; Dayhoff et al. 1978; Stapel et al. 1985; Peterson and Piatigorsky 1986) does not reveal that faster-evolving proteins, which are supposed to be less constrained, are more free to accumulate charge changes.

The actual evolutionary fate of charged residues must depend on the specific properties of each individual protein. Most striking is the extreme conservation of charge in most crystallins and histones. Charged interactions between the eye lens crystallins are supposed to be of extreme importance to maintain the transparency of the lens fiber cells. A close and even packing of these soluble proteins is required to prevent light scattering (Delaye and Tardieu 1983; Slingsby 1985). The tertiary and quaternary structure of α-crystallin is not known, but the αA subunits are more exposed in the large aggregates than are the αB subunits (Tardieu et al. 1986). The lower OV/EV value of αA thus suggests that the surface charges, especially, are under selective constraint in these proteins. The monomeric γ-crystallins are indeed characterized by conserved networks of ion pairs covering large surface areas of the molecule (Summers et al. 1986). In the related β-crystallins a redistribution of charge has occurred, probably in connection with the formation of oligomers of various sizes by these proteins, and this may explain the lesser constraint on charge changes (Slingsby et al. 1988). The conservation of charge in most histones may logically be explained by the required contacts between arginine side chains and the DNA phosphate backbone (Kornberg 1977).

Equally interesting is the apparent tendency of some proteins to accumulate more changes in charge than expected by random substitutions. In the case of phospholipase A2 it can be envisaged that positive selection of charge changes may occur in relation with the charge and packing properties of the various substrate phospholipids (Waite 1988). In the electron transfer protein ferredoxin it has been observed (Perutz 1978) that, as in other proteins (e.g., Tomazic and Klibanov 1988), salt bridges are responsible for thermal stability, and the high proportion of charged replacements may reflect adaptations to different stability requirements.

Most of the proteins in Table 1, however, have OV/EV values around or somewhat below unity (0.7–1.1). This would be in keeping with a limited functional significance of most of the charged residues in these proteins.

The validity of our inferences greatly depends on the reliability of the calculated expected and observed values. The predicted fractions of charged replacements are based on the assumption of random base substitution and unbiased codon usage. We found indications that the actual codon preference of genes may tend to slightly decrease the EV. In this light biased codon usage could be partially the result of an evolutionary pressure to diminish the chances for radical amino acid replacements. More difficult to assess are the effects of nonrandom base substitutions resulting from directional mutation pressure or base drift. Such directionality can go toward both higher or lower G+C content of DNA and can correspondingly have considerable directional effects on amino acid replacements (Perrin and Bernardi 1987; Preparata and Saccone 1987; Bernardi et al. 1988; Sueoka 1988). The actual nonrandom pattern of spontaneous point mutations results, at least in mammalian genes and pseudogenes, in particular in a much higher proportion of transitional mutations than expected under random mutation (Li et al. 1984). This reduces, in a set of six mammalian and one *Drosophila* genes, the proportion of charged changes by about 9%. However, applying the observed nonrandom mutation pattern to some of our data sets revealed a variable influence on expected charge changes (Table 2).

A more trivial cause for uncertainty is that both OV and EV values will vary when different sets are used for the protein families. We have included in our study many more globin and cytochrome *c* sequences than Peetz et al. (1986), and omitted some fibrinopeptide B sequences. This indeed changes the calculated OV and EV values, but the same general conclusions about the acceptance of charge changes in these proteins are reached.

In conclusion, more accurate EV values can in principle be expected when actual codon preference and substitution patterns are known for all investigated proteins; and OV values may change, becoming more representative, when the data sets become larger. This will, however, not invalidate our main observation that functional constraint at the protein level indeed is a major determinant of charge changes during the evolution of many proteins.

The present work shows that conservation of physicochemical properties of amino acids is dependent upon the protein that is considered. This implies that transition probabilities for amino acids are protein specific. It should be realized that this aspect is not taken into account in the commonly used methods for the comparison of protein sequences (viz. homology searching, alignment, dot

matrix), which use a common similarity measure (e.g., Dayhoff's PAM matrix).

Acknowledgments. We thank Dr. Ton de Haan for statistical advice and Dr. Christian Gautier for valuable comments. The use of the services and facilities of the Dutch CAOS/CAMM Center is gratefully acknowledged.

References

Barlow DJ, Thornton JM (1983) Ion-pairs in proteins. J Mol Biol 168:867–885

Bernardi G, Mouchiroud D, Gautier C, Bernardi G (1988) Compositional patterns in vertebrate genomes: conservation and change in evolution. J Mol Evol 28:7–18

Blundell TL, Sibanda MJ, Sternberg MJE, Thornton JM (1987) Knowledge-based prediction of protein structures and the design of novel molecules. Nature 326:347–352

Clarke B (1970) Selective constraints on amino-acid substitutions during evolution of proteins. Nature 228:159–160

Dayhoff MO, Park CM, McLaughlin PJ (1972) Building a phylogenetic tree: cytochrome *c*. In: Dayhoff MO (ed) Atlas of protein sequence and structure, vol 5. National Biomedical Research Foundation, Washington DC, pp 7–16

Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In: Dayhoff MO (ed) Atlas of protein sequence and structure, vol 5, Suppl 3. National Biomedical Research Foundation, Washington DC

Delaye M, Tardieu A (1983) Short-range order of crystallin-proteins accounts for eye lens transparency. Nature 302:415–417

Dickerson RE (1971) The structure of cytochrome *c* and the rates of molecular evolution. J Mol Evol 1:26–45

Doolittle RF (1979) Protein evolution. In: Neurath H, Hill RL (eds) The proteins, ed 2, vol 4. Academic Press, New York, pp 1–118

Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 185:862–864

Grantham R, Perrin P, Mouchiroud D (1986) In: Dawkins R, Ridley M (eds) Oxford surveys in evolutionary biology, vol 3, Oxford University Press, Oxford

Honig BB, Hubbell WL, Flewelling RF (1986) Electrostatic interactions in membranes and proteins. Annu Rev Biophys Biophys Chem 15:163–193

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Kornberg RD (1977) Structure of chromatin. Annu Rev Biochem 46:931–954

Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105–132

Li W-H, Wu C-I, Luo C-C (1984) Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J Mol Evol 21:58–71

Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searches. Science 227:1435–1441

Maruyama T, Gojobori T, Aota SI, Ikemura T (1986) Codon usage tabulated from the GenBank genetic sequence data. Nucleic Acids Res 14:r151–r153

Matthews JB (1985) Electrostatic effects in proteins. Annu Rev Biophys Biophys Chem 14:387–417

Miller S, Janin J, Lesk AM, Chothia C (1987) Interior and surface of monomeric proteins. J Mol Biol 196:641–656

Nei M (1975) Molecular population genetics and evolution. North-Holland Publishing Company, Amsterdam, p 25

Ohno S (1970) Evolution by gene duplication. Springer-Verlag, New York

Peetz EW, Thomson G, Hedrick PW (1986) Charge changes in protein evolution. Mol Biol Evol 3:84–94

Perrin P, Bernardi G (1987) Directional fixation of mutations in vertebrate evolution. J Mol Evol 26:301–310

Perutz MF (1978) Electrostatic effects in proteins. Science 201: 1187–1191

Peterson CA, Piatigorsky J (1986) Preferential conservation of the globular domains of the βA3/A1-crystallin polypeptide of the chicken eye lens. Differentiation 19:134–153

Preparata G, Saccone C (1987) A simple quantitative model of the molecular clock. J Mol Evol 26:7–15

Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH (1985) Hydrophobicity of amino acid residues in globular proteins. Science 229:834–838

Sharp PM, Cowe E, Higgins DG, Shields DC, Wolfe KH, Wright F (1988) Codon usage patterns in *Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. Nucleic Acids Res 16:8207–8211

Slingsby C (1985) Structural variation in lens crystallins. Trends Biochem Sci 10:281–284

Slingsby C, Driessen HPC, Mahadevan D, Bax B, Blundell TL (1988) Evolutionary and functional relationships between the basic and acidic β-crystallins. Exp Eye Res 46:375–403

Stapel SO, Zweers A, Dodemont HJ, Kan JH, de Jong WW (1985) ε-crystallin, a novel avian and reptilian eye lens protein. Eur J Biochem 147:129–136

Sternberg MJE, Hayes FRF, Russell AJ, Thomas PG, Fersht AR (1987) Prediction of electrostatic effects of engineering of protein charges. Nature 330:86–88

Sueoka N (1988) Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci USA 85:2653–2657

Summers LJ, Slingsby C, Blundell TL, den Dunnen JT, Moormann RJM, Schoenmakers JGG (1986) Structural variation in mammalian γ-crystallins based on computer graphics analyses of human, rat and calf sequences. I. Core packing and surface properties. Exp Eye Res 43:77–92

Tardieu A, Laporte D, Licinio P, Krop B, Delaye M (1986) Calf lens α-crystallin quaternary structure. A three-layer tetrahedral model. J Mol Biol 192:711–724

Tomazic SJ, Klibanov AM (1988) Why is one *Bacillus* α-amylase more resistant against irreversible thermoinactivation than another? J Biol Chem 263:3092–3096

Vogel H, Zuckerkandl E (1971) The evolution of polarity relations in globins. In: Neyman J (ed) Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, vol 5, Darwinian, neo-Darwinian, and non-Darwinian evolution. University of California Press, Berkeley, pp 155–176

Wada A, Nakamura H (1981) Nature of the charge distribution in proteins. Nature 293:757–758

Waite M (1988) The phospholipases. In: Hanahan DJ (ed) Handbook of lipid research, vol 5. Plenum, New York

Warshel A, Russell ST (1984) Calculations of electrostatic interactions in biological systems and in solutions. Q Rev Biophys 17:283–422

Wilson AC, Carlson SS, White TJ (1977) Biochemical evolution. Annu Rev Biochem 46:573–639

Zuckerkandl E (1975) The appearance of new structures and functions in proteins during evolution. J Mol Evol 7:1–57

Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ (eds) Evolving genes and proteins. Academic Press, New York, pp 97–116

Received July 10, 1989/Revised January 3, 1990