TWO DIMENSIONAL ANALYSIS OF POLARITY CHANGES IN GLOBIN AND CYTOCHROME *C*

HELMUT VOGEL
CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE, MONTPELLIER

1. Introduction

The classification of amino acid substitutions between protein chains has led to considerable success especially in the construction of phylogenetic trees that correctly, and in complete independence of the paleontological record or morphological facts, retrace many aspects of evolution (for example, M. O. Dayhoff [3]). The changes of physical properties that accompany those substitutions have not been as thoroughly investigated, at least not in close statistical conjunction with the substitutions themselves. The following attempt may open some new perspectives in this direction.

2. Procedure

By "class of proteins" we mean a set of homologous chains, generally functionally defined, that have been completely sequenced and that can include or exclude the reconstructed nodes in a phylogenetic tree (examples: all globins; all cytochromes *c*).

By "group of proteins" we mean a certain subset of a class of proteins which may or may not correspond to functional or taxonomic characteristics (examples: the α globins; all monohemic globins, the cytochromes *c* of all birds).

A "combination of chains" is a nonordered pair of chains.

The "combination of two groups" is the set of all combinations of chains, one member of the pair stemming from one of the groups, the other member from the other group. If the two groups are identical, we speak of an "in-group combination," otherwise of an "out-group combination."

In the present context, every combination is characterized by a point with the coordinates: N the number of sites where the two chains have different amino acids; P the difference of total polarities of the two chains. The polarities p_i of the amino acids are adapted from Woese [5] (Table I). Every combination is, according to the definitions above, listed only once, the sign of P being determined by the arbitrary order in which the chains are numbered. This has been done for 39 globin chains (see Table II) and for 25 cytochrome *c* chains (Table

TABLE I

POLARITIES OF AMINO ACIDS

Woese [5].

Cys	4.8	Pro	6.6	His	8.4
Leu	4.9	Thr	6.6	Gln	8.6
Ile	4.9	Ala	7.0	Arg	9.1
Phe	4.9	Ser	7.5	Asn	10.0
Try	5.2	Gly	7.9	Lys	10.1
Met	5.3			Glu	12.5
Tyr	5.4			Asp	13.0
Val	5.6				

Mean $\bar{p} = 7.43$

Standard deviation $(\overline{p^2} - \bar{p}^2)^{1/2} = 2.45$

II). Figures 1 and 2 show the distribution of the points (N, P). For every combination of groups, the set of points is sliced into an appropriate number of horizontal layers, such that the variation of N within any layer does not seem too great and, on the other hand, the layer still contains a reasonable number of points (at least 6). For every layer, the moments of the distribution of P up to sixth order are computed. The first two moments are fitted by determining the parameters p and ρ of a Bernoulli distribution (see Model 1) that has the same mean and variance (Table III). For this Bernoulli distribution, the higher moments are also computed and compared to the observed ones. If the Bernoulli distribution fits the observed one sufficiently well (as witnessed by a satisfactory correspondence of the higher moments or derived statistics like skewness, kurtosis, and so forth, which holds rather generally), the parameters of the former, p and ρ, are compared to the values of p and ρ that are predicted by totally random polarity changes only subject to the restrictions that the structure of the genetic code provides for one step mutations (see the end of Section 3).

3. Results

Quite roughly speaking, all points (N, P) together fill the inside of a parabola $N = AP^2$. For globins, due to their higher variability, the sector of this parabola traced by the points is much higher than for cytochromes. Nevertheless, for both classes the width is about the same, $A \approx 0.03$. Such a behavior would result from a random polarity change; if an amino acid substitution were connected with an average polarity change p, and if this change could with equal probabilities be positive or negative, within N substitutions the expectancy for the total polarity change would be 0, and its expected standard deviation about $\frac{1}{2}pN^{1/2}$. Thus, there should be a parabola, outside of which at any height some 20 per cent of the points can be found. As a first rough approximation, one thus obtains $p \approx 3.5$.

TABLE II

List of Investigated Chains

The parallel numbering systems for cytochrome *c* correspond to the arrangements used in Figures 2 and 4, respectively.
The printing arrangement within each group is in decreasing order of polarity.
Sources: [1], [2], [4] as stated in table, and [3] if unstated.

Globins			Cytochrome *c*									
Alpha	Beta	Others										
1 Human	14 Human	30 Human γ	2	1	Horse	10	8	Pigeon	13	17	Turtle	
2 *Rhesus*	15 *Rhesus*	31 Human δ	3	2	Sheep	8	9	Duck	16	16	Tuna	
3 Dog [4]	16 Lemur	32 Sheep fetal	6	3	Rabbit	9	10	Penguin	17	15	Lamprey	
4 Mouse	17 Dog [4]	33 Bovine fetal	7	4	Kangaroo	11	11	Chicken	14	14	Bullfrog	
5 Rabbit	18 Rabbit	34 Lamprey	5	5	Grey Whale				15	13	Dogfish	
6 Horse	19 Horse	35 Myogl. Whale	4	6	Dog				12	12	Rattlesnake	
7 Pig	20 Pig	36 Myogl. Kangaroo	1	7	Human							
8 Bovine	21 Llama	37 Myogl. Horse		21	21	Horn Worm Moth			22	22	*Neurospora*	
9 Sheep A	22 Bovine	38 *Chironomus* [1]		20	20	Silk Worm Moth			23	23	Yeast	
10 Goat A	23 Sheep B	39 Myogl. Bovine		19	19	Screw Worm Fly			24	24	*Candida*	
11 Llama	24 Sheep C			18	18	*Drosophila*			25	25	Wheat	
12 Kangaroo	25 Sheep A											
13 Carp	26 Goat A											
	27 Barbary Sheep											
	28 Kangaroo											
	29 Frog [2]											

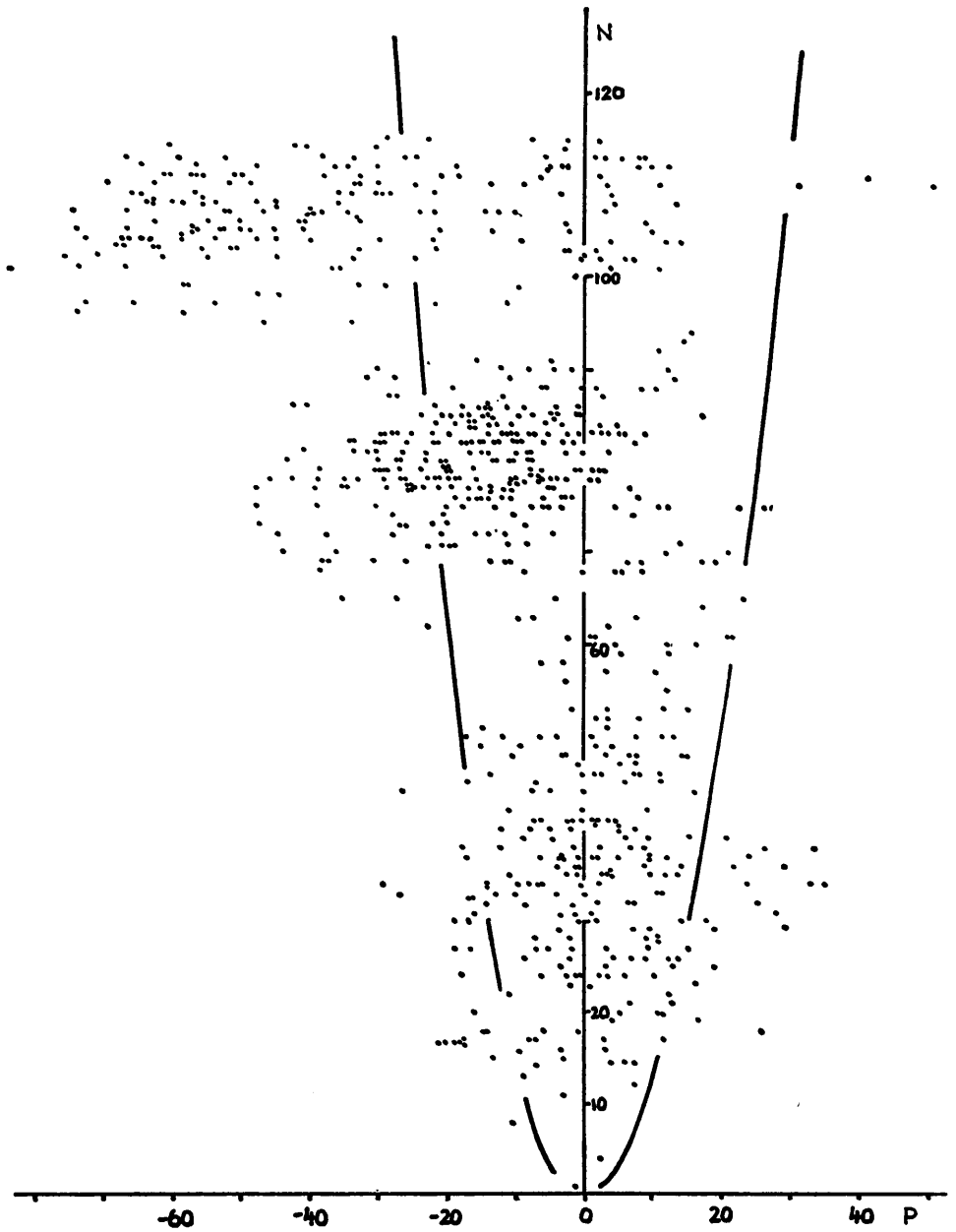

FIGURE 1

Number of substitutions N *versus* polarity difference P for globins.

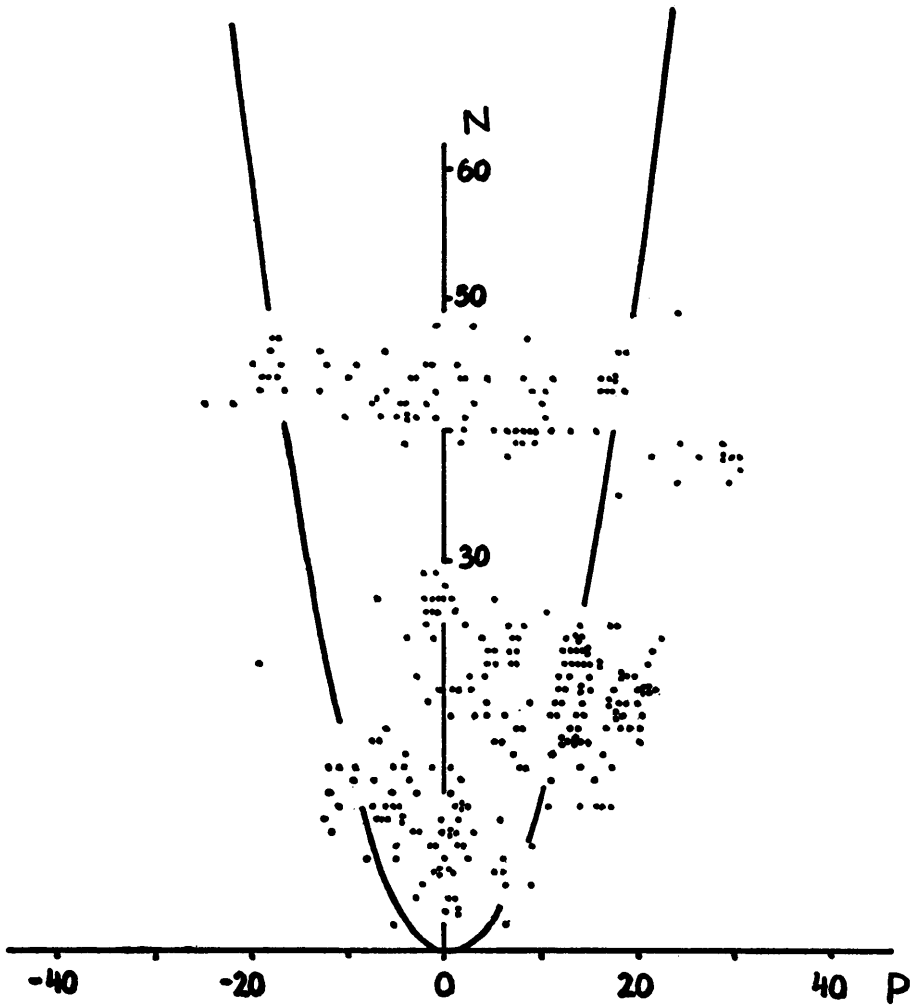

FIGURE 2

Number of substitutions N *versus* polarity change P for cytochrome *c*.

For globins and cytochromes alike, the distribution of the points is uneven: discrete islands are sometimes separated by almost empty areas. These islands are not always centered around $P = 0$. Each of them, generally speaking, corresponds to a certain combination of (taxonomical or functional) groups. Their eccentricity is often particularly patent if the two groups combined are taxonomically or functionally different (for example, birds/reptilia, amphibia, fish, displaced to the right, centered around $P \approx 16$; human/other mammals, birds displaced to the left, centered around $P \approx -10$). This evidently expresses

a corresponding tendency in the polarity differences; reptilia (with the exception of the turtle), amphibia and fishes have a lower polarity than birds, man has a lower one than the rest of the warm blooded animals.

But also within the same group (that is, for an in-group combination), displacements of the center of the point cloud from $P = 0$ may occur. Such is the case, for example, for the insect cytochromes or, in a lesser degree, for the α globins. This obviously means that the order of numbering the proteins within that group, being generally at least somewhat systematic, just corresponds to the order of rising or falling polarities (see below).

Each of these islands representing in-group or out-group combinations displays a much smaller dispersion of the P values than all the points taken together. On the other hand, the parabolic shape of those individual islands is less pronounced (better still for β globins), mostly due to the generally very limited extension in N direction for every group. If one nevertheless, on the testimony of the total distribution and of the fortunate cases like β globins, accepts the parabolic shape and the corresponding Bernoulli model, one finds for every individual combination of groups a much smaller mean polarity change p than for the total distribution. In some instances, generally for in-group and out-group combinations of relatively uniform taxonomy, p goes down to about 0.3.

Nearly all combinations of groups (out-group and in-group) show a p that is significantly smaller than $p_{\text{rand}} = 2.6$. Exceptions are all the α globins combined with each other ($p = 2.45 \pm 0.66$) and all the monohemic globins combined with each other ($p = 3.68 \pm 1.06$). If one picks out of the latter group the only really related subset of chains, namely, the myoglobins, one finds again $p = 2.80 \pm 0.10$, that is, practically the random value. Some combinations have p values down to $\frac{1}{6}$ or $\frac{1}{9}$ of the random value (birds/birds, insects/insects, human/insects). For the globins, the combination β/β, the one most numerous in points, presents the most perfect example for an absolutely symmetric, non-skew distribution with a kurtosis close to the value three of a normal distribution. Here, p (1.33 ± 0.31) is about half of the random value. This is true although the fetal sheep and cattle chains as well as human γ and δ are included. On the other hand, the α/α distribution is almost as wide as randomness prescribes.

"Right" and "left" with respect to the ordinate axis are evidently matters of the arrangement of the chains: since polarity establishes an ordering relation, there exists an arrangement such that all points lie to the right, for example. In Figure 4, this principle has been followed within each taxonomical group. Permuting the groups themselves would not quite succeed in putting all points to the right; the best solution would be birds—mammals—reptilia—amphibia—fish—insects, but there is an overlap between each pair of this series, also expressed by the fact that out-group combination point sets of the respective two groups cross the ordinate axis.

The most extended in-group point set is understandably that of the com-

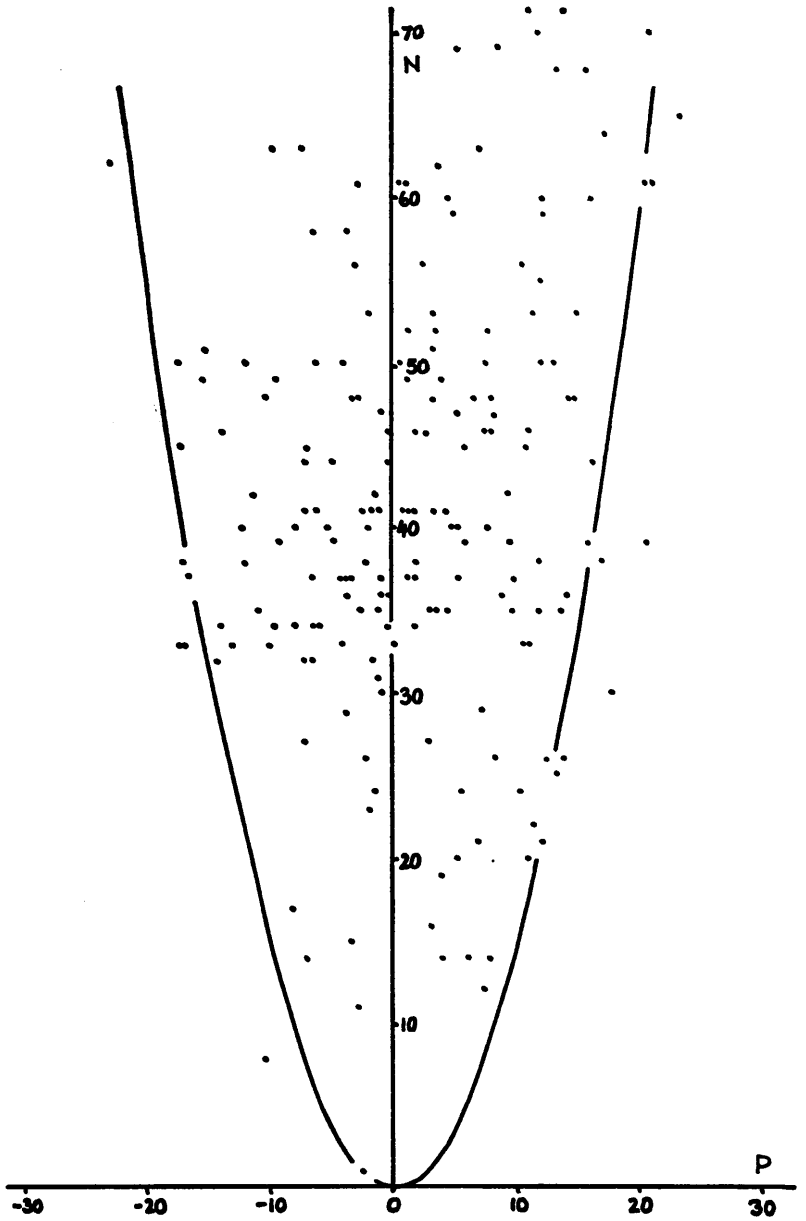

FIGURE 3

Number of substitutions N *versus* polarity change P for β globins (including human γ, δ, and sheep, bovine fetal).

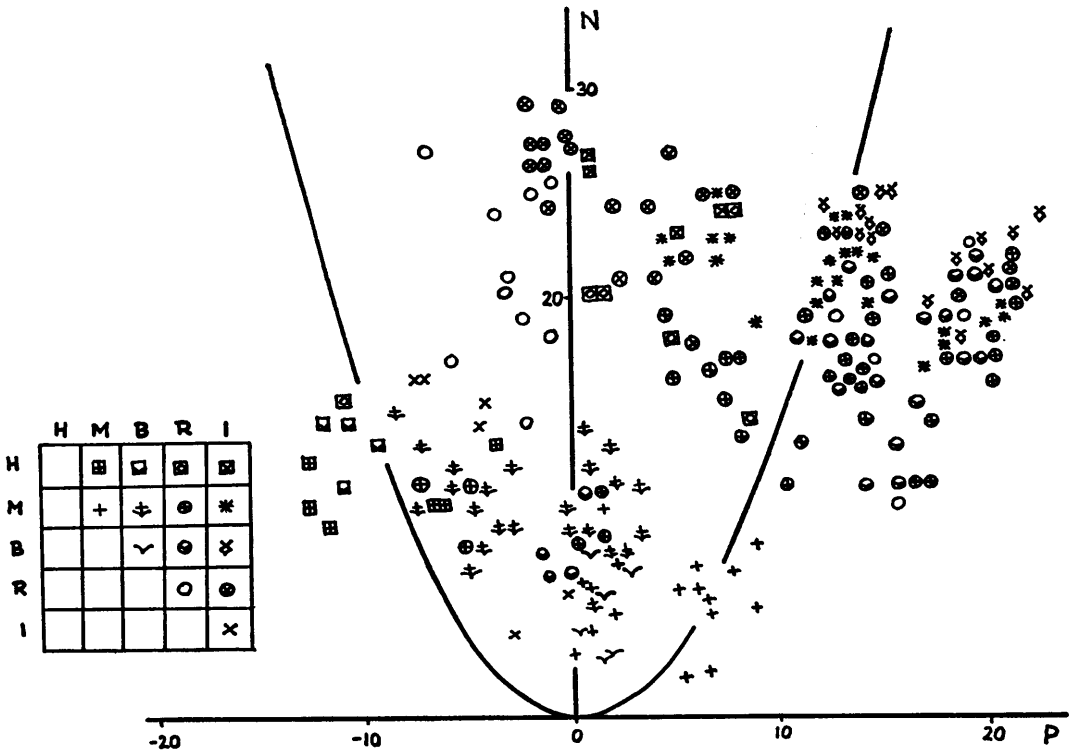

FIGURE 4

Number of substitutions N *versus* polarity change P for animal cytochrome *c*. H: human, M: mammals, B: birds, R: reptilia, amphibia, fish, cyclostoma, I: insects. Notice the different arrangement of the chains in each group as compared to Figure 2.

pound reptilia, amphibia, fish, cyclostoma. Mammals vary more in polarity than in amino acid composition, being a relatively young but extremely diversified group. The contrary holds true for the insects. Birds are very homogeneous in both respects.

Man and turtle each have to be set apart from their respective groups, but for different reasons: man is not too extravagant in polarity (if put at the end after dog, having the smallest mammalian polarity, it would not change the p and ρ that hold for the rest of the mammals too much); but man has a much higher number of amino acid differences N compared to the other mammals than these have among each other. Conversely, turtle has a high polarity totally outside of the rest of its group, rattlesnake included, but is very conservative in its N. In this sense, man is something between bird and reptile; turtle is a sort of bird.

4. Random polarity changes as dictated by the genetic code

Consider a protein in which amino acid i has an abundance a_i and a polarity p_i. If all possible base replacements occur with the same probability, the frequency of the amino acid substitution $i \rightarrow k$ will be given by a weight factor ν_{ik} that expresses the number of possible base replacements leading from i to k. This is done under the assumption that the different codons for a certain amino acid occur with equal frequency.

A substitution between amino acids represented by codon quartets, if possible in one step, will get a weight 1 (for example, Ala-Val). For the substitution of a quartet coded acid by a duet coded one, like Thr-Asn, the weight will be ½, which may be interpreted by cutting the Thr abundance in two, since only the two codons that permit a transition to Asn are in question. Likewise, for Leu-Ile, one gets ⅔, conversely for Ile-leu, ⅓.

Quite generally, one obtains the weight ν_{ik} for the substitution $i \rightarrow k$ by counting all possible transitions leading from any codon of i to any codon of k and dividing by the number of codons for i. The weights lie between ⅙ (for example, Ser-Try) and 3 (for example, Phe-Leu).

With a polarity change $\delta_{ik} = p_k - p_i$ for the substitution $i \rightarrow k$, the expectation for the *mean* polarity change per substitution and its standard deviation in a randomly selected substitution are plainly

$$(1) \qquad \bar{\delta} = \frac{\sum_{i,k} \delta_{ik}\nu_{ik}a_i}{\sum_{i,k} \nu_{ik}a_i}, \qquad \sigma = (\overline{\delta^2} - \bar{\delta}^2)^{1/2}.$$

Unless the above summation is arbitrarily restricted, $\bar{\delta}$ is very close to 0, some possible departure from 0 being only due to a preponderance in abundance either of the polar or the unpolar amino acids. Anyhow, the "equilibrium protein" has $\bar{\delta} = 0$, whether it is defined as a protein with five per cent abundance for every amino acid, or as one in which the abundances are proportional to the number of codons available for the different amino acids. The actual abundances in globin and cytochrome *c* yield very small $\bar{\delta}$. The appropriate measure for average polarity change per substitution, as far as only its absolute size is concerned, is evidently the standard deviation σ. This will be used under the name of p_{rand} for comparison with the observed polarity changes according to Model 1.

For concrete cases, one obtains the values shown in Table III for $\sigma = p_{\text{rand}}$. By its very construction, polar and apolar amino acids each group residing together in their special "quarters," the code keeps the mean polarity change smaller than it would be for really random changes. If any amino acid could be freely substituted for any other one, the average absolute size of the polarity change would be

$$(2) \qquad p_{\text{free}} = 1.45\ (\overline{p_i^2} - \bar{p}_i^2)^{1/2},$$

TABLE III

AVERAGE POLARITY CHANGE PER SUBSTITUTION

Abundance	p_{rand}	p_{free}
5% protein	2.68	3.55
Globin	2.98	4.10
Cytochrome *c*	2.60	3.58
Codon equilibrium protein	2.58	3.53

p_i meaning the polarity of amino acid i (Table I). The completely random p values thus obtained and listed in the second column of Table III are generally about 40 per cent bigger than those taking account of the restrictions dictated by the code.

5. Model 1

Every amino acid substitution is connected with a polarity change of equal absolute value p; this change can be positive or negative, with probabilities ρ and $1 - \rho$, respectively.

If there is a total of N substitutions between two chains, the probability that among these exactly μ correspond to a polarity increase (and $N - \mu$ to a decrease) is

$$W(\mu) = \binom{N}{\mu} \rho^{\mu}(1 - \rho)^{N-\mu}. \tag{3}$$

The resulting Bernoulli distribution of the total polarity change P, which is expressed in terms of μ by

$$P = p(2\mu - N), \tag{4}$$

has the following moments (in a form convenient for recursive computation):

$$\overline{P} = pN(2\rho - 1), \tag{5}$$

$$\overline{P^2} = 4p^2N\rho(1 - \rho) + \overline{P}^2, \qquad \sigma_P = 4p^2N\rho(1 - \rho), \tag{6}$$

$$\overline{P^3} = \overline{P}\sigma_P^2(3 - 2/N) + \overline{P}^3, \tag{7}$$

$$\overline{P^4} = 3\sigma_P^4(1 - 2/N) + 4p^2\sigma_P^2 + 2\overline{P}^2\sigma_P^2(3 - 4/N) + \overline{P}^4, \tag{8}$$

$$\overline{P^5} = 10\overline{P}^3\sigma_P^2(1 - 2/N) + \overline{P}\sigma_P^4(15 - 50/N + 24/N^2) + 4\overline{P}p^2\sigma_P^2(5 - 2/N) + \overline{P}^5. \tag{9}$$

From the observed values for P and ρ, according to (5) and (6), the parameters of the model can be computed:

$$p = \frac{1}{N}(N\sigma_P^2 + \overline{P}^2)^{1/2}, \tag{10}$$

$$\rho = \tfrac{1}{2}[1 + \overline{P}/(N\sigma_P^2 + \overline{P}^2)^{1/2}]. \tag{11}$$

TABLE IV

PARAMETERS OF MODEL 1

M: mammals; B: birds; RAF: reptilia, amphibia, fish, *Cyclostoma;* I: insects; P: plants; Hum: human; Tur: turtle; α: α globins; β: non-α globins (β, γ, δ, fetal); Myo: myoglobins; Mono: monohemic chains (myoglobin, lamprey *Chironomus);* Chir: *Chironomus.*

Combination	No. of points evaluated	Mean no. of substitutions	Parameters of model 1	
			Step length	Probability for right step
Cytochrome *c*				
Hum/MB	11	11.9	1.17	0.15
Hum/RAF	6	18.5	0.74	0.66
Hum/I	4	25.2	0.63	0.62
Hum/P	4	40.3	1.97	0.50
M/M	15	5.4	1.54	0.78
M/B	24	10.6	1.14	0.42
M/RAF	30	16.6	1.49	0.76
M/I	24	21.0	1.20	0.74
M/P	24	41.3	2.04	0.55
B/B	6	5.2	0.40	0.81
B/RAF	20	17.7	1.11	0.90
B/I	16	22.3	1.02	0.87
B/P	16	41.7	1.80	0.56
RAF/RAF	10	21.1	0.47	0.35
RAF/I	20	24.2	1.34	0.56
RAF/P	20	42.8	1.74	0.48
I/I	6	11.8	0.80	0.27
I/P	16	43.9	1.49	0.43
P/P	6	40.8	1.04	0.66
M/Tur	6	11.2	1.15	0.40
B/Tur	4	9.0	0.30	0.46
Tur/RAF	5	17.4	1.01	0.96
Tur/I	4	22.0	1.01	0.88
Tur/P	4	42.8	1.82	0.58
Globin				
α/α	63	26.1	2.45 ± 0.66	0.52 ± 0.09
α/α carp	12	68.2	1.32	0.47
α/β	74	74.1	1.50 ± 0.65	0.39 ± 0.08
α/Mono	78	104.0	1.31	0.49
α/Myo	52	106.6	2.06	0.42
α/Chir	13	101.2	1.09	0.19
β/β	218	45.0	1.33 ± 0.31	0.50 ± 0.05
β/Mono	68	111.6	2.04 ± 0.24	0.45 ± 0.01
β/Chir	19	105.3	0.78	0.13
Mono/Mono	15	71.3	3.68 ± 1.06	0.48 ± 0.04
Myo/Myo	6	22.5	2.80	0.49

Using these in (7), (8), (9), one can determine the higher moments of the Bernoulli distribution and compare them to observed ones, either directly or after transforming them into skewness and kurtosis:

$$\begin{aligned} \beta_1 &= \overline{P^3}/\sigma_P^3 \\ \beta_2 &= \overline{P^4}/\sigma_P^4. \end{aligned} \tag{12}$$

6. Model 2

The polarity change connected with a substitution is itself a random variable p with a probability distribution $f(p)$, that does not depend on the site at which the substitution occurs nor on time. If

$$\varphi(\zeta) = \int_{-\infty}^{\infty} e^{\zeta p} f(p)\, dp \tag{13}$$

is the generating function of $f(p)$, that is, essentially its Laplace transform, and if N subsequent independent substitutions are considered, the total polarity change P by those N substitutions has a distribution with the generating function

$$\Phi(\zeta) = \varphi(\zeta)^N. \tag{14}$$

The moments of that distribution are, consequently,

$$\begin{aligned} \overline{P} &= N\bar{p}, \\ \overline{P^2} &= N\overline{p^2} + N(N-1)\bar{p}^2, \qquad \sigma_P^2 = N\sigma_p^2, \\ \overline{P^3} &= N\overline{p^3} + 3N(N-1)\bar{p}\overline{p^2} + N(N-1)(N-2)\bar{p}^3, \\ \overline{P^4} &= N\overline{p^4} + 3N(N-1)(\bar{p}\overline{p^3} + \bar{p}^2) \\ &\quad + 6N(N-1)(N-2)\bar{p}^2\overline{p^2} + N(N-1)(N-2)(N-3)\bar{p}^4, \end{aligned} \tag{15}$$

and so forth.

Thus, given the observed distribution of the P, one could get an idea of the underlying distribution $f(p)$, for example, from its moments, the coefficients of the MacLaurin expansion of its Laplace transform:

$$\begin{aligned} \bar{p} &= \overline{P}/N \\ \overline{p^2} &= \overline{P^2}/N - (N-1)\overline{P}^2/N^2 \\ \overline{p^3} &= \overline{P^3}/N - 3(N-1)\bar{p}\overline{p^2} - (N-1)(N-2)\bar{p}^3, \end{aligned} \tag{16}$$

and so forth.

Heuristically, this procedure suffers from hyperparametritis, because *any* distribution of the P could be exactly fitted this way. However, one could decide whether the resulting $f(p)$ looks anyway reasonable, for example, like the random distribution predicted by the genetic code.

At the moment, the available set of data hardly permits doing this in a meaningful manner. Therefore, throughout this paper only Model 1 is referred to.

7. Conclusions

7.1. Model 1 satisfactorily describes all the combinations of groups investigated, as is established by the fact that once the model parameters p and ρ have been determined from the first two observed moments, the higher moments are predicted by the model with the accuracy to be expected considering the size of the sample of points. There is no need at present to recur to the more complete distribution of Model 2, which in Model 1 is simplified to two δ shaped peaks at $\pm p$.

7.2. In the case of cytochrome *c*, every combination of taxonomical groups, even for a group as wide as "plants," yields a point set that is much narrower than the genetic code predicts. During the evolution of these groups as well as for larger portions of the phylogenetic tree that include some of the corresponding branchings, a mechanism has been acting that has kept polarity at a desired level, evidently by suppressing substitutions connected with too high a polarity change. In cases of extreme constancy of polarity, as within the birds or the insects, practically only the "central tower" of Figure 5 can have been used, that corresponds to substitutions as harmless as, for example, Phe-Val or Ala-Ser. Even Ser-Thr, for example, must already have been too radical.

7.3. Just within this framework, the adaptive character of the relatively large changes of polarity from group to group is particularly accentuated. Since the distribution of *all* the cytochrome *c* points is even somewhat wider than the parabola predicted by the code, one has to admit that during the whole not only of animal, but even of vertebrate, evolution not only the reins that have curbed polarity changes have been let loose, but that even to a certain degree substitutions with higher than average polarity effect have been encouraged.

7.4. For the globins, the taxonomical grouping has not yet been done. Within the functional groups (α, β, monohemic), the above conclusions also hold true, although in a lesser degree.

The analysis described above may be generally applied to any property of a protein chain other than polarity that depends additively on the corresponding property of the component amino acids. By its two dimensional character, the method disposes of the objection that can sometimes be raised against similar discussions, namely, that a lack or smallness of differences in a given property be just due to the fact that there are so few substitutions between the considered chains. More detailed investigations are under way, taking into account also polarity variations between sections of chains.

REFERENCES

[1] G. Buse, S. Braig, and G. Braunitzer, "The constitution of the hemoglobin (erythrocruorin) of an insect (*Chironomus thummi thummi, Diptera*)," *Hoppe-Seyler's Z. Physiol. Chem.*, Vol. 350 (1969), pp. 1686–1691.

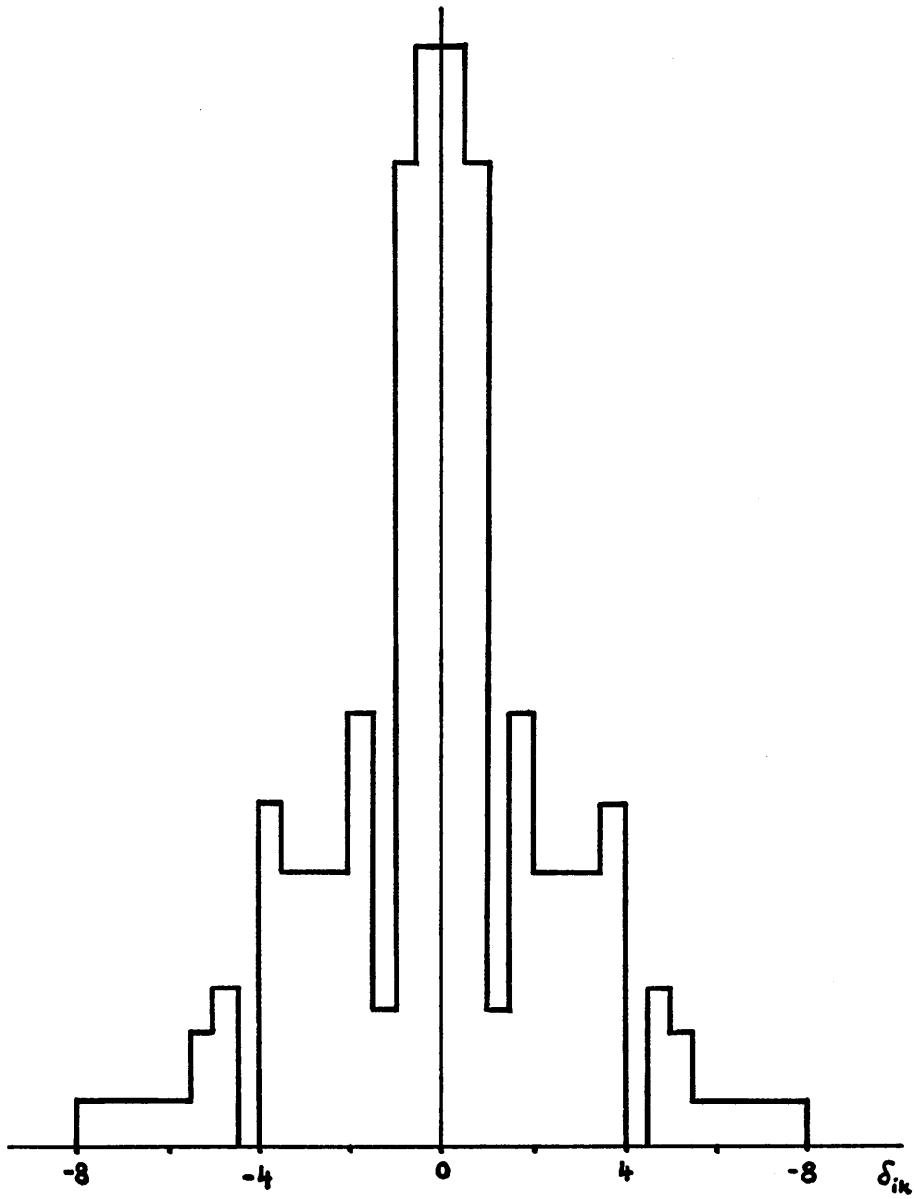

FIGURE 5

Frequency of polarity changes δ_{ik} in one step mutations on random movement within the genetic code; abundances of amino acids for codon equilibrium protein.

[2] J. CHAUVET and R. ACHER, "Sequence of frog hemoglobin β," *FEBS-Letters*, Vol. 10 (1970), pp. 136–140.
[3] M. O. DAYHOFF, *Atlas of Protein Sequence and Structure*, Vol. 4, Silver Spring, Md., National Biomedical Research Foundation, 1969.
[4] R. T. JONES, Personal communication.
[5] C. W. WOESE, in *The genetic code*, New York, Harper and Row, 1967, p. 172.