J.Mol.Evol.6,271-283 (1975)

Detecting Evolutionary Trends from Molecular Data

1. Some Measures of Compositional Nonrandomness

HELMUT VOGEL
Centre de Recherches de Biochimie Macromolêculaire (CNRS)
and Groupe U-67 (INSERM), Montpellier

Received March 7, 1975; September 15, 1975

Summary. The measures of compositional nonrandomness to be discussed as to their physical significance and to their power of detecting evolutionary significant variations are

$Q = \frac{100}{L} \Sigma \left| n_i - p_i L \right|$ (Holmquist, 1974),

$S = \ln L! + \Sigma\, n_i \ln p_i - \Sigma \ln n_i!$ the "compositional entropy", and

$\chi^2 = \Sigma\, (n_i - p_i L)^2 / p_i L$,

(p_i a priori probability for amino acid i, n_i its number of occurrences in a protein of length L). As a concrete example, the p_i are here supposed to represent equal frequencies of all non-stop codons. For each quantity, four levels are defined: The base level, with optimal (i.e. minimal nonrandomness) composition, admitting non-integer values of n_i; the integer level with optimal integer composition; the noise level, represented by a typical random cain; and the real protein level. On all these levels, S, which is the measure with the most direct physical sense, shows the smoothest behavior with the smallest relative fluctuations and thus the highest resolution.

Key words: Compositional Nonrandomness - Entropy of Peptide Chains - Neutrality or Selection - Evolutionary Trends - Selective-Stochastic Balance.

1. INTRODUCTION

What a random chain should look like is a matter of definition. Two main proposals explicitly or implicitly haunt the literature:
a) In a random peptide chain each amino acid occurs with the same a priori probability of $p_i = 0.05$. We call such a chain

a 5% peptide, but will not be further concerned with it except for comparison purposes.
b) In a random chain, each possible non-stop codon triplet is represented with the same probability of 1/61, and consequently the a priori probability for amino acid i is $p_i = m_i/61$, where m_i is its codon multiplicity, i.e. the number of triplets that code for that amino acid. We call such a state the codon equilibrium.

In either case, sequentially speaking, there is no interaction between sites, i.e. each site is occupied according to these probabilities, independently of its neighbors.

Codon equilibrium in a wider sense does of course not require equal probabilities of the four nucleotides. A generalization will be discussed which is characterized by four different probabilities of the nucleotides (Vogel, in prep.). We will treat here the simple case $p_i = m_i/61$.

Several quantities have been proposed to measure the deviation of a real protein from randomness, as far as compositior is concerned. They compare the observed number of amino acids of type i, n_i, with the number predicted by the random hypothesis, p_iL, for a chain with L residues.

Holmquist proposed a simple pseudo-linear measure

$$(1) \qquad Q = \frac{100}{L} \sum_{i=1}^{20} |n_i - p_i L|$$

(without the absolute sign, the sum would of course vanish due to $\Sigma n_i = L$ and $\Sigma p_i = 1$; Holmquist & Laird, 1974).

The χ^2 of the departure between observed and expected amino acid numbers might be used:

$$(2) \qquad \chi^2 = \Sigma \frac{(n_i - p_i L)^2}{p_i L}$$

An entropy-like measure may be introduced, the zero-order compositional entropy

$$(3) \qquad S = \ln L\,! + \Sigma\, n_i \ln p_i - \Sigma \ln n_i\,!$$

We will compare these measures first with respect to their physical meaning, then to their applicability to real data, especially to their freedom of noise and their discriminatory power.

2. PHYSICAL MEANING OF THE NONRANDOMNESS MEASURES

A given chain with letters one by one drown from an inexhaustible bag which contains these letter with abundances p_i, cannot be supposed to display exactly the ideal composition p_iL for two reasons:
- the p_iL will not be all integer, except for $L = \mu\ 61$ (μ integer);
- there will be fluctuations around the equilibrium composition that is as close to p_iL as is possible with integer n_i.

The probability of generating a chain with the composition n_i is given by the multinomial expression

$$(4) \qquad P\ (n_i,\ L) = L\ !\ \frac{\Pi p_i^{\,n_i}}{\Pi n_i\ !}\ .$$

If noninteger values of n_i were allowed, the maximum of P would be given by the equilibrium composition $n_i = p_iL$. For real chains with integer n_i, the highest possible P is slightly lower than (4) (see section 4). The typical chain generated by the random source has a yet lower P (see section 5). For a real protein, the distance between its P and the maximum one permitted for the given length L will have a bearing on the selective constraints that governed the evolution of that protein. Without any such constraints, and provided the assumptions underlying the randomness concept are valid, the protein should resemble the chains generated by the random drawing process. If there is a significant departure from such a random composition, this must have a functional or evolutionary reason.

The entropy S as defined above is, as usual, the natural logarithm of P and thus shares its physical sense. We will show that χ^2 can under certain assumptions be taken as a first approximation to S.

In contrast, a direct justification for Q in probabilistic or other terms seems not to exist. Its analytical form (the absolute value is a "kinked" function) already seems to preclude that. Q is thus a purely phenomenological quantity, possibly sometimes more convenient than the others.

Our "compositional entropy" S should not be confused with Shannon's informational entropy $H = -\ \Sigma p_i\ \ln\ p_i$, nor with the "sequential entropy" $\Sigma n_i\ \ln\ p_i$, the log of the probability of a chain with given sequence being produced by random drawing out of the infinite bag.

3. SOME PROPERTIES OF S

We first consider $p_i L$ as integers, which is literally valid only if $L = \mu\ 61$ (μ an integer), and derive an approximation for S:

$$S = \ln L! + \Sigma n_i \ln p_i - \Sigma \ln n_i!$$

$$S_o = \ln L! + \Sigma L\, p_i \ln p_i - \Sigma \ln (Lp_i)!$$

$$S_o - S = \Sigma s_i - \Sigma \nu_i \ln p_i \quad ,$$

where

$$\nu_i = n_i - L\, p_i$$

and

$$s_i = \ln \frac{n_i!}{(Lp_i)!} \approx \nu_i \ln Lp_i + \frac{\nu_i(\nu_i + 1)}{2\, Lp_i} -$$

$$- \frac{\nu_i(\nu_i + 1)\ (2\nu_i + 1)}{12\ L^2\ p_i^{\,2}}$$

whichever is the sign of ν_i (obtained by developing the log).

Thus

$$(5) \qquad S_o - S = \Sigma \frac{\nu_i(\nu_i + 1)}{2\ L\ p_i} - \Sigma \frac{\nu_i(\nu_i + 1)\ (2\nu_i + 1)}{12\ L^2 p_i^2}$$

Real proteins generally have n_i that deviate from $p_i L$ by some units. Then the χ^2 approximation is adequate:

$$s_i \approx \frac{1}{2} \quad \frac{\nu_i}{p_i L} \quad .$$

Only in very exceptional cases like histones, where one amino acid is several times more abundant than expected, the second sum in (5) with its ν_i^3 may take over for some i:

$$s_i \approx \frac{1}{12} \quad \frac{\nu_i^3}{L^2 p_i^2} \quad .$$

For *noninteger* $p_i L$, the factorials in the $\Sigma \ln (p_i L)!$ have to be smoothed out, e.g. by the gamma function:

$$(6) \qquad S_o = \ln L! + L \Sigma p_i \ln p_i - \Sigma \ln \Gamma (L\, p_i + 1) \quad .$$

This is still the maximum of all possible S for a given L. As a function of L, the maximal entropy per amino acid, S_o/L, has been plotted in Fig.2.

4. THE "PERIODIC TABLE OF PROTEINS"

Which is the chain with *integer* n_i that best approaches the ideal values S_o according to (6) and $Q_o = 0$ respectively? Since chains with best S and with best Q will be slightly different, we first maximize S and call the value thus obtained S_1. To do this, we start from a chain with $L = \mu\ 61$ (μ an integer), for which $S = S_o$ is attainable by $n_i = \mu m_i$ (m_i number of codons for amino acid i). We might call such a chain a full shell chain. Next we gradually add always that amino acid that yields the highest increase in S or in P. The P increase by adding an amino acid i is given by the factor $(L+1)p_i/(n_i+1)$ which, as long as $n_i = Lp_i$, i.e. as long as there are no duplications in the construction of the new shell, is the larger, the higher p_i. Thus one first adds sextet, then quartet, then triplet coded amino acids. Before adding the first doublet acid, however, addition of still another sextet acid becomes more advantageous. One thus arrives at the following pattern of shell construction:

Multiplicity:

6 4 3 6 (2 4 6) 3 6 4 6 (1 2 3 4 6) (7)

Total no. of aa added:

3 8 9 12 21 26 29 30 33 38 41 43 52 53 58 61 .

The order in the last 5 columns does not matter for S, nor does it in the other 3 columns put into parentheses.

Each period of 61 amino acids can be split into 4 subperiods, of 12, 17, 12, 20 elements. Within the second and the fourth subperiod, the filling order actually does not matter, in the first and the third it does. The atomic physicist would speak of degenerate S values in the second and the fourth subperiod. Exceptions of the rare earth type do not occur in this periodic system.

The most apparent kinks in the function $S_1(L)$ occur near the full shell state $L = \mu 61$. For instance, $S_1(\mu 61 - 1) = S_1(\mu 61)$, whereas $S_o(\mu 61 - 1) = S_o(\mu 61) - \ln(4.5/61.\mu)$. The difference $S_o - S_1$ is, aside from this "jumpiness", fairly proportional to L, such that generally $(S_o - S_1)/L \lesssim 0.02$.

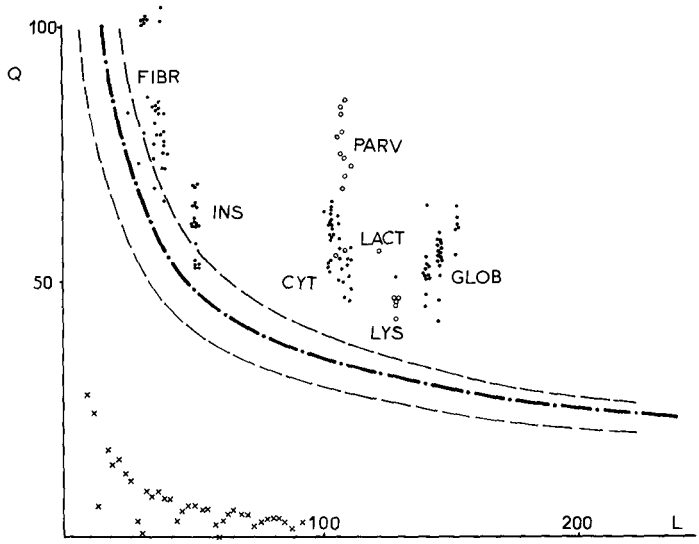

Fig.1
Holmquist's measure of compositional nonrandomness $Q = \frac{100}{L} \Sigma \left| n_i - p_i L \right|$ and its 4 levels: Base line: Q of an equilibrium chain; $Q_O = 0$; crosses: Optimal chain with integer n_i (Q_1); heavy line: Typical randomly generated chain (Q_2 with its standard deviation, indicated by broken lines); points and open circles: Real proteins

The compositions with lowest Q look somewhat different, at least for some lengths L, since the amino acids to be added in the course of their gradual construction are chosen by a different criterion. If one notes the Q-optimal compositions in the same notation as for the S-optimal ones above, together with their Q·L values within a period of 61, say between L = 62 and L = 122, they form a palindrome:

$$6 \quad 4 \quad 3 \quad \begin{matrix} 6 \\ 2 \end{matrix} \quad 4 \quad 6 \quad \begin{matrix} 3 \\ 1 \end{matrix} \quad 6 \quad 4 \quad \begin{matrix} 6 \\ 2 \end{matrix} \quad 3 \quad 4 \quad 6 \quad .$$

This palindrome property can be generally proven as follows: Take a chain with $L = \mu \cdot 61 \pm \lambda$, which has a composition $n_i = \mu \cdot 61 \cdot p_i \pm \nu_i$. Then $Q \cdot L = 100 \Sigma \left| \mu \cdot 61 \cdot p_i \pm \nu_i - \mu \cdot 61 \cdot p_i \pm \lambda p_i \right| = 100 \Sigma \left| \nu_i - \lambda p_i \right|$, which only depends on the excesses or deficiencies λ and ν_i, but not on their sign.

For the quantity Q·L, the palindrome is moreover almost symmetrical around $L = (\mu + 1/4) \cdot 61$.

Except for the cases $\begin{matrix} 6 \\ 2 \end{matrix}$ and $\begin{matrix} 3 \\ 1 \end{matrix}$ in the above filling order, there is no indeterminacy like there was in large regions the S-optimal period.

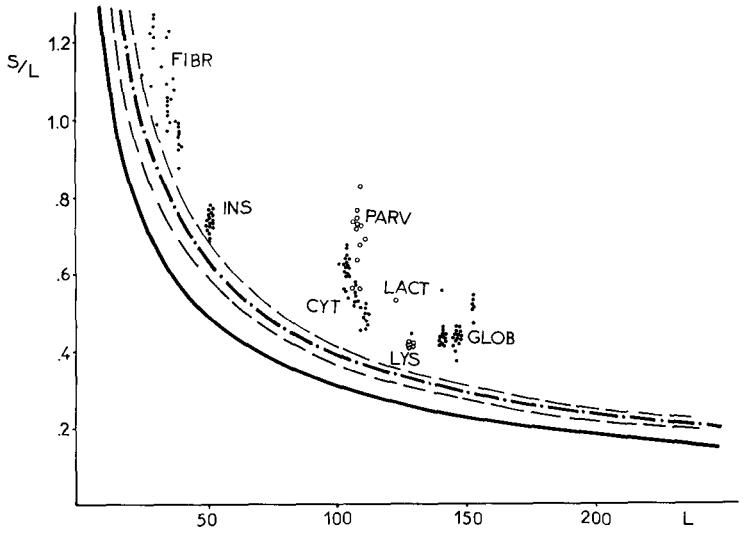

Fig.2
Entropy S and its 4 levels: Base line: S = O; Solid line: Equilibrium chain (S_O); the S_1 values of the optimal chain with integer n_i merge with the S_O line within its width; Broken heavy line: Typical randomly generated chain (S_2 with its standard deviation, indicated by broken lines); Points and open circles: Real proteins. Note that the S axis has been inverted in order to facilitate comparison with Fig.1

The Q values for the Q-optimal compositions, which we will call $Q_1(L)$, are relatively much farther from their base line $Q_O = O$ than the $S_1(L)$ are from their base $S_O(L)$. This becomes evident if one compares these distances with those obtained for the noise peptides ($S_2(L)$ and $Q_2(L)$ respectively) and with real proteins (see below and Figs.1 and 2). Most Q_1 (but the few around $L = \mu \cdot 61$ and $L = (\mu+1/2) \cdot 61$) attain about 20% of Q_2 or more, whereas only the highest S_1-S_O attain 10% of S_2-S_O for $L \approx 50$. At higher L (above 100) this relation becomes more favorable for S: while Q_1 stays around 15-20%, S_1-S_O decreases to 1% or less.

5. THE NOISE PEPTIDE

The third level of S or Q is represented by the chains that a random source emits (see section 2). We generated such chains by a computer, ten for each length L, at L = 10, 20, ... 200, and determined the means of S and Q, called $S_2(L)$ & $Q_2(L)$, with their standard deviation, for each L. Figs.1 and 2 show the results.

The S_2 points, reduced to the base line S_o and to one amino acid, i.e. the points $(S_2-S_o)/L$ can be represented by a fit function which for small L runs like $L^{-1/2}$, for high L like L^{-1}. The best fit (linear regression of $(S_2-S_o)^{-1} \cdot L^{1/2}$ versus $L^{1/2}$) is

$$(8) \qquad S_o - S_2 = \frac{11.89}{1 + 4.22\ L^{-1/2}}$$

with a correlation coefficient r = 0.906.

In contrast, the points $Q_2(L)$ are rather "jumpy". The best fit by a similar function as above is

$$(9) \qquad Q_2 \approx \frac{1}{0.00278\ \sqrt{L} + 0.000011\ L}$$

i.e. practically $Q_2 \approx L^{-1/2}$, but with a correlation of r = 0.255 only.

Relative to the total values of S_2, the standard deviation of this quantity is much smaller than the corresponding fraction for Q_2. However, with respect to the distance from the base line, S_2 has the higher coefficient of variation (st. dev./mean), namely about 20% as compared to about 12% for Q_2.

If one asks for the *S and Q of the most probable composition*, the answer is $S = S_o$, $Q = 0$. But if one asks for the *most probable values of S and Q*, the answer is S_2, Q_2 with $S_2 \neq S_o$, $Q_2 \neq 0$. This apparent paradox, so common in statistical physics, is resolved by a distinction of micro- and macrostates. Different values of S or Q are realized by very different numbers of compositions. The maximal S, namely S_o, and the minimal Q, namely 0, are by definition realized by only one composition, if any. The number of possible compositions rises steeply with the distance from equilibrium.

The most probable value of S can be estimated from first principles rather than from simulation as $S_2 = S_o - 8.5$ (see appendix), the width of the distribution around this value is $\Delta S_2 \approx 2.9$. This predicted $S_o - S_2 = 8.5$ looks slightly different from the empirical relation (8), but coincides with it for L = 112, which is close to the mean length of the chains used in (8). The mean of the 200 values of $S_o - S_2$ for 20 different chain lengths used in Fig.2 is 8.59. A hundred more chains generated with L = 61 furnished $(S_o - S_2)/L = 0.128 \pm 0.041$ compared to a prediction of 0.139 ± 0.048. Fifty more chains with L = 140 had an observed 0.065 ± 0.021 against a predicted 0.061 ± 0.021.

6. REAL PROTEINS

As a fourth level, we plot the entropies S and the Q values for some proteins from the following classes:

34 globins (α, β, myoglobins and other monomeric chains),
36 cytochromes (c),
32 fibrinopeptides (A and B combined),
18 insulins,
6 lysozymes and lactalbumins,
12 parvalbumins and related chains.

Of rabbit TN-C and ALC 1,2, only the sections have been used that align with the parvalbumins. (Thatcher et al., 1974; Pechère et al., 1973; Joassin, 1974; Capony et al., 1973; Gerday, 1974; Frankenne et al., 1973; Coffee et al., 1973, 1974; Capony et al., 1974, 1975a, 1975b; Collins, 1974; Frank et al., 1974).

All other sequences are from Dayhoff (1972, 1973). Chains were only used if their sequence is known without any ambiguity. Even chains containing B and Z (for Asx and Glx) were discarded.

It is obvious from Figs.1 and 2 that S discriminates better than Q. The clouds of low molecular weight proteins (fibrinopeptides and insulins) penetrate into the standard deviation margins of Q_2 (some points almost touch the Q_2 curve itself), whereas they are all outside of the S_2 margins (but fibrinopeptide of Rhesus which touches this margin). The distances of the means for the 6 or 7 protein classes from the Q_2 or S_2 curve, expressed in standard deviations of Q_2 or S_2, are always higher for S:

Protein:	Fibrin.	Insul.	Cytochr.	Parvalb.	Lactalb.	Lysoz.	Globin
$(S_2-S)/\Delta S_2$	3.6	2.5	7.4	10.3	7.6	4.2	9
$(Q_2-Q)/\Delta Q_2$	2.1	1.6	4.9	8.4	6.0	3.8	8

But even with the less powerful measure, Q, one sees at once significant differences between the protein classes and within each class, both in their mean distance from S_2 or Q_2 and in their spread around this mean. The best example for the first point (different class means) is cytochrome c compared to the typical parvalbumins (except rabbit ALC): Both have approximately the same length, and yet the parvalbumins are much more distant. This effect evidently survives any base shift, but it is much emphasized by not using S but $S-S_2$ as a measure. For the second point (different class spreads) one can quote the extremely scattered fibrinopeptides as compared to the very compact lysozymes. However, this spread has

to be judged against the background of substitutional change within the molecule. A variable chain like fibrinopeptide might well scatter more than a relatively slowly substituting one like lysozyme, even if the latter covers a wider taxonomic range. But much "slower" proteins like cytochrome and insulin also scatter more than lysozyme. The question to be asked is not "How much scatter?", but "Given the observed number of substitutions, did these substitutions cause an entropy (or Q) change as expected if they were random, or more, or less?" In this perspective, the problem is treated in another paper of this series.

As to the differences between means of subgroups within protein classes, we also think them to be evolutionary and functionally significant. The possible meaning of myoglobin being more distant than most other globins, and of α and β globins being alike, will be discussed elsewhere (H. Vogel, in prep.). Cytochromes of plants, fungi, and insects are less distant than those of vertrebrates as has been pointed out before (Vogel, 1972). Quite generally, the comparison with the behavior of other properties than S and Q helps distinguish between significant and spurious differences and helps explain the functional reasons for the significant ones.

7. CONCLUSIONS

From the preceding discussion it appears that of the three measures for compositional nonrandomness that have been considered - Q, χ^2 and S - the compositional entropy S has the clearest sense with respect to the model chosen, since it describes the distance from equilibrium or from a noise peptide in direct probabilistic terms without any approximation or additional assumption.

Phenomenologically, the behavior on the three levels (integer base level, noise level, real level) is much smoother and less subject to fluctuations in terms of S than in terms of Q. Most importantly, S offers a much better resolution than Q, i.e. variations between classes of proteins and between individual chains within one class can be much more clearly discriminated.

Acknowledgements. I thank Miss D. Hillaire for skillful technical assistance, Dr. J.F. Pechère for communicating to me the aligned sequences of some parvalbumins and related chains, some of which he has elucidated himself, and Drs. L. Gatlin, J. King and E. Zuckerkandl for interesting discussions.

APPENDIX:
COMPUTATION OF S_2, THE MOST PROBABLE VALUE OF S FOR THE NOISE PEPTIDE

We first simplify the situation by assuming all p_i to be equal, $p_i = 0.05$. Then, for deviations ν_i from equilibrium that are not much greater than $p_i L$, we need only to maintain the χ^2-like term in (5):

$$S' = S_o - S = \frac{1}{2} \Sigma \frac{\nu_i (\nu_i + 1)}{p_i L} = \frac{10}{L} \Sigma (\nu_i^2 + \nu_i) = \frac{10}{L} \Sigma \nu_i^2$$

since $\Sigma \nu_i = 0$. Abstracting for the moment from the discreteness of the ν_i, one sees that in the 20-dimensional space of the ν_i, the surface with a fixed value of S' is a sphere around the origin with a radius $R = \sqrt{S' L/10}$. The condition $\Sigma \nu_i = 0$ reduces this sphere to its intersection with a plane passing through the center, i.e. to a 19-dimensional sphere. The volume enclosed by this sphere, corresponding to S' values higher than the fixed S'_1, is

$$V = A R^{19} = A (S'_1 L/10)^{9.5} .$$

(A is a constant whose value does not interest here). Increasing S'_1 by dS' leads to an increase of V by

$$dV = \frac{\partial V}{\partial S'} dS' = 9.5 A (0.1 L)^{9.5} S'^{8.5}_1 dS' .$$

The number of integer-coordinate points inside this shell, equal to αV, gives the number of compositions that realize a S' out of $(S'_1, S'_1 + dS')$. Each of these compositions has a probability $e^{-S'}$ relative to the equilibrium composition. The probability distribution of S' reads thus

$$f(S') = A' L^{9.5} S'^{8.5} e^{-S'} ,$$

A' comprising all constants. f(S') is a curve slightly skewed to the right. Its maximum is given by

$$f'(S') = 0, \text{ i.e. } S'_m = S_o - S_2 = 8.5.$$

This mode has a height of

$$f(S'_m) = 16164 A' L^{9.5} .$$

We estimate the width of the distribution from the modal curvature

$$f''(S'_m) = -f(S'_m)/8.5 \ .$$

The distance $\Delta S' = S' - S'_m$ that reduces the height to half its modal value is in good approximation

$$\Delta S' = 2.9 \ .$$

In the actual case of different p_i, the 20-dimensional sphere is deformed to an ellipsoid whose center is generally not exactly at the origin:

$$S' = \frac{1}{2} \Sigma \frac{x_i^2}{L p_i} + \frac{1}{8L} \Sigma p_i^{-1} \quad , \quad x_i = \nu_i + \frac{1}{2} \ .$$

The half-axes are $a_i = \sqrt{2 p_i LS' + \frac{1}{4} \Sigma p_k^{-1}}$, the center is at $\nu_i = -1/2$. The volume becomes $V = A \Pi a_i$. These complications make a complete treatment harder. It is however evident that the S' independent terms in the a_i flatten the original $S'^{8.5} \cdot e^{-S'}$ distribution to a $L \to 0$ limit of simply $e^{-S'}$ with $S'_m = 0$. The transition is tentatively described by the $\sqrt{L}$ term in the denominator of (8).

The prediction of Q_2 and its standard deviation would probably be possible along similar lines, but is more cumbersome due to the analytical brittleness of the absolute sign.

REFERENCES

Capony, J.-P., Rydèn, L., Demaille, J., Pechère, J.-F. (1973). Eur.J. Biochem. 32, 97-108

Capony, J.-P., Rochat, H., Pina, C., Pechère, J.-F. (1974). C.R.Acad.Sci. 279 D, 1789-1791

Capony, J.-P., Demaille, J., Pina, C., Pechère, J.-F. (1975a). Eur.J. Biochem., in press

Capony, J.-P., Pina, C., Pechère, J.-F. (1975b). C.R.Acad.Sc., in press

Coffee, C.J., Bradshaw, R.A. (1973). J.Biol.Chem. 248, 3305-3312

Coffee, C.J., Bradshaw, R.A., Kretsinger, R.H. (1974). Adv.Exptl.Med. Biol. 48, 211-222

Collins, J.H. (1974). Biochem.Biophys.Res.Comm. 58, 301-308

Dayhoff, M.O. (1972, Suppl.1973). Atlas of Protein Sequence and Structure

Frank, G., Weeds, A.G. (1974). Eur.J.Biochem. 44, 317-325

Frankenne, F., Joassin, L., Gerday, C. (1973). FEBS-Lett. 35, 145-148

Gerday, C. (1974). Abstr. 9th FEBS-Meeting, Budapest, p.21
Holmquist, R. (in press). J.Mol.Evol.
Joassin, L. (1974). Abstr. 9th FEBS-Meeting, Budapest, p.20
Laird, M., Holmquist, R. (in press). J.Mol.Evol.
Pechère, J.-F., Capony, J.-P., Demaille, J. (1973). Syst.Zool. 22, 533-548
Thatcher, D.R., Pechère, J.-F. (unpublished)
Vogel, H., Zuckerkandl, E. (1971). Biochemical Evolution and Origin of Life, E. Schoffeniels, ed., p.352
Vogel, H. (in press). J.Mol.Evol.

Dr. Helmut Vogel
CNRS, BP 5051
F-34033 Montpellier, France