METHOD FOR THE DETERMINATION OF PROTEIN EVOLUTION RATES BY AMINO ACID COMPOSITION. EVOLUTION RATE OF ACTINS

M. M. OGIEVETSKAYA*
Institute of Biophysics, USSR Academy of Sciences, Puschino, U.S.S.R.

(Received 24 January, 1977)

Abstract. A method has been developed to determine the actin evolution rate. The method is based on amino acid composition. The actin evolution rate has been established to be extremely low. Only three or less amino acid changes per hundred amino acid residues have accumulated for a 100 million years. One can explain the conservative nature of actin evolution as a sequence of its unique tightly fitted structure rich in biologically active centres at short distances from each other. The peculiar invariability of polar amino acids leads to a conclusion that some given distribution of charges is necessary for the unique functioning of actin molecules.

1. Introduction

The amino acid sequence has now been deciphered for a rather large number of proteins and it now becomes possible to study the evolution of living things on a molecular level [8]. To study protein evolution is important not only for the reconstruction of the history of living things but also for other reasons, namely, for the understanding of the genetic mechanism which fixes changes in protein molecules during the evolution process and for the deeper understanding of some molecular biology problems [9], including those of molecular immunobiology [16,7,3].

The modern methods of protein evolution investigation are based on the analyses of the primary structure of homological proteins. The comparative analysis of amino acid sequence of two or more proteins of a given family or a given homology allows one to count the number of amino acid changes and thus one can express the similarity or dissimilarity of proteins under investigation. The number of amino acid changes for a given geological period reflects the protein evolution rate.

The method of estimation of the evolution rate of proteins by their sequences is, in principle, the most sensitive one. However, it has some shortcomings. The main of them are the following: when performing the comparative analysis of homological proteins their amino acid sequences are aligned with respect to each other so as to achieve the maximal number of coincidences [8]. If the difference is small this procedure may not result in incorrect interpretation. But if the difference is essential, one has to introduce artificially a large number of deletions and insertions on the protein level. In this

* Present address: Laboratory of Neutron Physics, Joint Institute for Nuclear Research, H.P.O. Box 79, Moscow, U.S.S.R.

Origins of Life 8 (1977) 145–154.

connection a real possibility appears for making errors. The determination of protein primary structure is labour-consuming and includes a lot of complications caused by the purification of peptide-fragments and the establishment of their sequences. Although there are great achievements in this field [10, 18, 21] the method of the determination of the amino acid sequence is still complicated and expensive. The accumulation of knowledge on the primary structure of proteins proceeds rather slowly, especially for high molecular proteins.

At the same time there is a long list of proteins for which only the amino acid composition is established. This list is still being added, since the analysis of amino acid composition is simple and automated. Although the common amino acid composition of proteins is less informative than their amino acid sequences, it characterizes a protein too.

Some attempts have been made to compare proteins by their amino acid composition [25, 26] to estimate the degree of relatedness among proteins with similar functions [22] or to find the common ancestor among proteins of different homologies [24].

The present paper is aimed at estimating the evolution rate of proteins on the basis of their amino acid composition. The proper method is proposed and its correctness is established. This method is called ERPAAC (the Evolution Rate of Proteins by Amino Acid Composition).

2. Description of ERPAAC

Each protein is made up of 20 amino acids but in this paper only 18 amino acids are analysed since asparagine and aspartic acid and also glutamine and glutamic acid are taken as their sums. The quantity of each amino acid may be figuratively called a 'coordinate'. Then the protein can be presented as a point in an 18-dimensional space. The resemblance of evolutionary related proteins is reflected in the similarity of their amino acid compositions or, in other words, in the short distances of the corresponding points in the 18-dimensional space under consideration. The degree of relatedness among proteins can be established by the dispersion of their amino acid composition. The proposed quantitative measure of dispersion is the mean value of standard deviations over all amino acids. Since only the relation of amino acids is essential, the quantitative contribution of each amino acid can be expressed in molar percents. Note that no information about the protein molecular weight is needed.

Denote the amino acid by the index 'a' and the analysed proteins by the index 'i'. The molar percent of the amino acid 'a' in the protein 'i' is expressed by x_{ia}. Then the mean contents of amino acid 'a' in the given family of 'n' proteins is defined by the relation

$$\bar{x}_a = \frac{1}{n} \sum_{i=1}^{n} x_{ia} \tag{1}$$

For each protein of an individual family x_{ia} differs from x_a. The more similar proteins are, the smaller the differences in x_{ia}. As a natural dispersion measure we take the value of the standard or mean square deviation, widely used in mathematical statistics, [4].

$$\sigma_a = \frac{1}{n} \sum_{i=1}^{n} (x_{ia} - \bar{x}_a)^2 \ . \tag{2}$$

The measure of σ_a characterizes the mean-square dispersion of the amino acid 'a' of the family of 'n' proteins. The mean value of standard deviations over each of 18 amino acids is proposed as a general measure of dispersion

$$\sigma = \frac{1}{18} \sum_{a=1}^{18} \sigma_a \ . \tag{3}$$

The value of σ describes the degree of relatedness of proteins under consideration. The smaller σ, the more similar the proteins. In the case of proteins of an individual homology the value of σ is thought to be characteristic. It can serve as a quantitative measure for the comparative estimation of the protein evolution rate. The value of σ will below be called the σ-measure.

The statistical confidence of the σ-measure rises with increasing the number of proteins in a sample. However, a possibility of increasing the number of proteins is limited by the present knowledge about the amino acid composition of proteins. It should be noted that the σ-measure depends not only on the number of proteins in a sample but also on their set. For example, the σ-measure for a sample consisting of five homological proteins, namely, cytochromes c of rabbits, horses, pigs, bulls, sheep is equal to 0.28, while for a sample of five cytochromes c of rabbits, dogs, pigs, bulls, sheep is smaller and equal to 0.21. The reason for this difference is due to the fact that the cytochrome c of sheep, bulls and pigs are very similar to each other but the degree of their relatedness is different from the cytochromes c of dogs and horses. Thus it is possible to get some knowledge about the evolution rate of proteins on the basis of the σ-measure but only by comparing the values of the σ-measure computed for completely identical samples of homological proteins of two or more families.

To present the evolution rate of protein under investigation as the rate of amino acid change accumulation for a chain of 100 amino acid residues per 100 mln years this σ-measure for the protein family under consideration must be compared with that calculated for an analogous sample of another protein family whose evolution rate has been expressed by the amino acid sequence technique.

3. The Correctness of ERPAAC

In mathematics a method is called correct if for a problem with the known answer it gives the same one.

Now estimate by the ERPAAC method the evolution rate for the proteins of the known amino acid sequence and then compare the results obtained with those of the evolution rates obtained by their amino acid sequence. Such a comparison permits also interpret in terms of our method the PAM which is the unit of the evolution rate introduced by Dayhoff [8]. The PAM unit (the first letters of 'Accepted Point Mutation')

TABLE I

Standard deviations of 18 amino acids for equivalent (see the text) protein samples: fibrinopeptides, hemoglobins α, hemoglobins β, insulins and cytochromes c

Amino acid	Family of proteins				
	Fibrinopeptides	Hemoglobins α	Hemoglobins β	Insulins	Cytochromes c
1. Alanine	3.63	2.06	0.50	1.60	0.36
2. Phenylalanine	2.56	0.26	0.54	0.00	0.36
3. Lysine	1.91	0.26	0.38	0.00	0.36
4. Tyrosine	1.52	0.26	0.34	0.00	0.36
5. Serine	4.35	1.04	1.10	1.35	0.56
6. Cysteine	–	0.26	0.26	0.00	0.00
7. Glycine	4.20	0.86	0.95	0.92	0.72
8. Leucine	2.54	0.95	0.53	0.00	0.00
9. Proline	3.63	0.49	0.86	0.00	0.36
10. Threonine	3.69	0.95	1.18	1.35	0.36
11. Asparagine + Aspartic acid	4.10	1.26	1.31	0.00	0.72
12. Histidine	2.28	0.26	0.52	0.00	0.00
13. Methionine	–	0.33	0.53	–	0.36
14. Glutamine + glutamic acid	5.35	0.89	0.92	0.00	0.91
15. Valine	1.92	0.41	0.49	0.92	0.36
16. Isoleucine	1.48	0.79	0.32	0.92	0.72
17. Arginine	0.83	0.00	0.56	0.00	0.00
18. Tryptophan	–	0.00	0.004	–	0.00
Total	43.99	11.33	11.29	7.06	7.01

corresponds to one amino acid change in the protein chain consisting of 100 amino acid residues.

In accordance with the above requirements five identical samples were composed for the homological proteins of the following families: fibrinopeptides, α and β hemoglobins, insulins and cytochromes c from six mammals (human beings, horses, rabbits, pigs, bulls and sheep). The values of standard deviations from each amino acid are given in Table I. It should be noted that fibrinopeptides and insulins are deficient in three and two amino acids, respectively. From the data on each amino acid standard deviation it is possible to compute the σ-measure for an individual protein family, taking into account the number of deficit amino acids. These σ-measures are given in Table II. It presents also the evolution rates for these proteins calculated by the sequence technique and expressed in PAM's [8]. It should be noted that the computation of the σ-measure was carried out by using proteins of mammals. The time of divergence for mammals is 75 million years by an approximate estimation according to the geological data [8]. Therefore, the values of the σ-measures obtained must be ascribed to the period of 75 million years or shorter, since the divergence of the mammals under consideration could have occurred considerably later.

TABLE II

Protein evolution rate established by alternative techniques: (a) By amino acid sequences [8] expressed in PAM (see the text); (b) by the general amino acid composition (ERPAAC) expressed in the σ-measure

Protein family	Alternative technique	
	By amino acid sequences (PAM)	By the general amino acid composition (σ-measure)
1. Fibrinopeptides	90	2.93
2. Hemoglobins α	14	0.63
3. Hemoglobins β	14	0.63
4. Insulins	4	0.44
5. Cytochromes *c*	3	0.39

As is seen from Table II, fibrinopeptides have the largest σ-measure (2.93), while cytochromes *c* have the smallest one (0.39). The method of the determination of the evolution rate by the sequence also characterizes fibrinopeptides as the proteins with the extremely high evolution rate (90 PAM), while it gives for cynchromes *c* the extremely low evolution rate (3 PAM). One can see also that in the order of decreasing both the σ-measure and the quantity of PAM units these proteins are arranged in the same sequence.

So, qualitative conclusions on the evolution rate of proteins analysed by alternative methods coincide in spite of the fact that the given values take into account different information and show the rates of amino acid change accumulation for different geological periods.

The above comparison allows one to consider the ERPAAC method to be correct. So one can apply the σ-measure to the relative estimation on the evolution rate for proteins of unknown amino acid sequence. The ERPAAC method has been applied to actins.

4. Estimation of the Evolution Rate for Actins

At present the amino acid composition is known for a large number of actins, while the sequence has been established only for the rabbit skeletal muscle actin [11]. For this purpose two identical samples of different protein families were set, see Table III. One sample consisted of the proteins investigated, namely, actins. The other one consisted of proteins which could serve as a standard for the proteins under investigation. This standard was made up of cytochromes *c*. The values of standard deviations over each of 18 amino acids and the σ-measure for each sample are presented in Table IV. As can be seen from Table IV, the σ-measure for actins turns out to be equal to 0.14, while for cytochromes *c* it is 0.21. This shows logically that the evolution conservatism of actins is observed in still larger degree than that of the cytochromes *c*.

TABLE III

Amino acid composition of cytochromes *c* and actins in molar per cent

Amino acid	Protein family									
	Cytochromes *c* [8]					Actins [13]				
	Bull	Pig	Sheep	Dog	Rabbit	Pig	Bull	Sheep	Dog	Rabbit
1. Alanine	5.77	5.77	5.77	6.73	6.73	7.88	8.36	8.14	8.03	8.17
2. Phenylalanine	3.85	3.85	3.85	3.85	3.85	3.25	3.28	3.19	3.17	3.17
3. Lysine	17.31	17.31	17.31	17.31	17.31	5.11	5.36	5.44	5.11	5.05
4. Tyrosine	3.85	3.85	3.85	3.85	3.85	4.01	3.79	4.10	4.15	3.94
5. Serine	0.96	0.96	0.96	0.96	0.96	6.71	6.92	6.80	6.56	6.66
6. Cysteine	1.92	1.92	1.92	1.92	1.92	1.30	0.99	1.05	1.06	1.28
7. Glycine	13.46	13.46	13.46	13.46	12.50	7.53	7.81	7.52	7.70	7.67
8. Leucine	5.77	5.77	5.77	5.77	5.77	7.14	7.29	7.30	7.07	6.95
9. Proline	3.85	3.85	3.85	3.85	2.88	4.98	4.74	4.86	4.94	5.11
10. Threonine	7.69	7.69	7.69	8.65	7.69	7.84	8.01	7.83	7.55	7.88
11. Asparagine + aspartic acid	7.69	7.69	7.69	6.73	9.62	9.45	9.56	9.36	9.36	9.38
12. Histidine	2.88	2.88	2.88	2.88	2.88	2.00	2.08	2.12	2.04	2.00
13. Methionine	1.92	1.92	1.92	1.92	1.92	3.83	2.68	3.15	4.29	3.38
14. Glutamine + glutamic acid	11.54	11.54	11.54	10.58	9.62	11.10	10.81	11.71	10.73	10.89
15. Valine	2.88	2.88	2.88	2.88	3.85	4.71	5.09	4.99	5.16	5.09
16. Isoleucine	5.77	5.77	5.77	5.77	5.77	7.31	7.37	7.32	7.22	7.51
17. Arginine	1.92	1.92	1.92	1.92	1.92	4.82	4.88	5.01	4.90	4.87
18. Tryptophan	0.96	0.96	0.96	0.96	0.96	1.03	0.97	0.68	0.98	0.99

TABLE IV

Standard deviations of 18 amino acids for equivalent (see the text) samples of cytochromes *c* and actins

Amino acid	Protein family	
	Cytochromes *c*	Actins
1. Alanine	0.47	0.16
2. Phenylalanine	0	0.05
3. Lysine	0	0.16
4. Tyrosine	0	0.13
5. Serine	0	0.12
6. Cysteine	0	0.13
7. Glycine	0.39	0.11
8. Leucine	0	0.13
9. Proline	0.39	0.12
10. Threonine	0.39	0.15
11. Asparagine + aspartic acid	0.94	0.08
12. Histidine	0	0.05
13. Methionine	0	0.55
14. Glutamine + glutamic acid	0.77	0.17
15. Valine	0.39	0.16
16. Isoleucine	0	0.09
17. Arginine	0	0.06
18. Tryptophan	0	0.13
Total	3.74	2.55
σ-measure	0.21	0.14

In fact the σ-measure (0.14) for actins can be even smaller since in the determination of the amino acid composition some errors can be introduced which are difficult to take into account. These errors can artificially increase the σ-measure. One can obtain the approximate value of these errors summed up for all the 18 amino acids. This can be calculated by comparing the amino acid compostion data for the same protein, namely, the rabbit skeletal muscle actin which has been determined in several laboratories.

The degree of discrepancy between amino acid data of two laboratories can be expressed by means of the coefficient of difference. This coefficient is derived by Formula (3), where 'n' equals 2. The coefficients obtained are presented in Table V. It is seen from Table V that the data from the Laboratory denoted 9 are in sharp disagreement with the remainder of the data and they are, therefore, not discussed here. However, the remaining coefficients also reveal dispersion. The average difference coefficient is 0.047 and its standard deviation is 0.015. The data of the second and the third Laboratories are outside the statistical limit of 0.047 ± 0.015. We believe the data of Elzinga *et al.* [11] to be reliable, as they have been obtained from the primary structure. These data are given under number 10 in Table V.

TABLE V

The difference coefficients of actins from rabbit skeletal muscles according to the data of the amino acid analysis of 10 laboratories

No.	1	2	3	4	5	6	7	8	9	10	Ref.
1.	0	0.056	0.066	0.043	0.036	0.029	0.024	0.029	0.220	0.031	[2]
2.		0	0.058	0.057	0.054	0.059	0.051	0.053	0.178	0.076	[14]
3.			0	0.069	0.059	0.057	0.070	0.052	0.212	0.082	[17]
4.				0	0.039	0.032	0.046	0.049	0.217	0.049	[6]
5.					0	0.028	0.031	0.036	0.215	0.041	[15]
6.						0	0.036	0.036	0.228	0.038	[23]
7.							0	0.031	0.210	0.035	[12]
8.								0	0.211	0.039	[1]
9.									0	0.234	[20]
10.										0	[11]

However, in the above analysis the data of Carsten and Katz [6] have been used as just this laboratory gives data on actins for other mammals [13]. Note that the data [6] agree with those of Elzinga *et al.* [11] and that they are within the limits of permissible errors, since the corresponding difference coefficient is 0.049.

Thus, the σ-measure of actin evolution rate which we obtained and which is equal to 0.14 includes, along with the differences among mammal actins accumulated as a result of the evolution process, also the errors introduced in amino acid composition determination. Data of some laboratories for the same actin have the mean difference coefficient equal to 0.047 instead of zero. *So, the true σ-measure can turn out still smaller.*

The evolution rate of cytochromes *c* by the amino acid sequence has been established to be three PAMs for 100 mln years. This technique allows one to state that only three amino acid changes took place in the chain of 100 amino acid residues for 100 mln years. The σ-measure for actins is much smaller than for the cytochromes *c*. Thus, one can consider that less than three amino acid changes took place in the pieces of 100 amino acid residues of the actin molecule for 100 mln years. The conclusion about the low evolution rate of actins has been made on the comparative analysis of mammals actins. Our preliminary report [19] has been devoted to the analysis of actins of other living things. The conclusion on the striking actin evolution conservatism has been confirmed for a wide variety of living things up to mono-cell organisms.

5. Discussion of the Evolution Conservatism of Actins

The low evolution rate of actins as has been established by REPAAC evidences that the amino acid composition of actins from different living things is invariable. This proves in turn that in actin molecules there is a very small number of neutral regions, i.e. the regions which accumulate amino acid changes without breaking its biological function.

We have no reason to think that the actin coding piece of DNA is less subjected to mutagenic effects than other protein coding pieces of DNA. So, weak actin evolution can be explained as a result of extremely severe selection. Actins are proteins of great importance to life as they are responsible for the mobility of living things. The organisms perish if their actins change even slightly as the result of mutation. Actin-mutated organisms lose their ability to move, they cannot search for food and struggle for life.

One should also take into account the following well-known properties of actins. Actin molecules have a rather small molecular weight of 41.719 as has been established by the amino acid sequence [11]. At the same time this molecule is poly-functional and able to interact with many structures. It is known that actin can interact with other actin molecules and this results in their fibrillar form, with myosin, or its fragments, increasing the myosin ATP-ase activity, with tropomyosin, troponins, alpha-actinins and also with nucleotides and cations [5]. For realizing each of these reactions the actin molecule should have a corresponding active centre capable of specific interaction with a proper substrate.

Thus, the severe selection of actin-mutant organisms could be understood properly if actin is imagined as a protein having a unique, tightly fitted, structure. The absence of neutral regions and polyfunctionality of actin, the protein of comparatively low molecular weight, accounts for extremely tight fit of multiple biologically active areas on the molecule.

Therefore, actin seems to have a structure whose parts are nearly all biologically active and are engaged in the sequential chain of biological events. Blocking of any biologically active centre, may lead to the discommunication of the whole mechanism responsible for the mobility of the living things.

From Table III it can be seen that actins are loaded with polar amino acids: histidine, lysine, arginine, asparagine and glutamine together with the proper acid. Each third amino acid may be carrying a charge. As has been shown in [11], polar amino acids are distributed nonuniformly. They appear to form clusters or charged regions. The analysis of Table IV, where standard deviations of all amino acids are given, shows that the dispersion of the majority of polar amino acids – histidine, arginine and asparagine together with the proper acid – in the actins of mammals is rather small. Hence, one can draw the conclusion that some given distribution of charges is necessary for the unique functioning of the actin molecule. The striking conservatism of actins allows one to think that there are also evolutionally conservative regions (invariable to the composition of amino acids and to their sequence) in the proteins interacting with the actin, just in myosin, its S_1-fragment, tropomyosin, troponins and alpha-actinin. One may consider that the study of these regions will make it possible to understand many details of the mechanism of muscular and nonmuscular contraction.

The results have been calculated by using the BESM-6 computer.

Acknowledgements

Thanks are due to Acad. G. M. Frank for support and contant interest in the investigation, to Prof. V. I. Ogievetsky for helpful discussions and O. V. Ogievetsky for his help in programming.

References

[1] Adelstein, R. S. and Kuhel, W. M.: 1970, *Biochemistry* **9**, 1355–1364.
[2] Asatoor, A. M. and Armstrong, M. D.: 1947, *Biochem. Biophys. Commun.* **26**, 168–175.
[3] Atassi, M. Z.: 1975, *Immunochemistry* **12**, 1–6.
[4] Bailey, P.: 1970, *Mathematics in Biology and Medicine*, MIR, Moscow.
[5] Bendall, J. R.: 1970, *Muscles, Molecules and Movement*, MIR, Moscow.
[6] Carsten, M. E. and Katz, A. M.: 1964, *Biochem. Biophys. Acta*, **90**, 534–546.
[7] Crumpton, M. J.: 1975, *The Antigens* (ed. by M. Sela), Acad. Press, N.Y., Vol. **2** 1–75.
[8] Dayhoff, M.: 1972, *Atlas of Protein Sequence and Structure*, Nat. Biomed. Foundation, Silver Spring.
[9] Dickerson, R. E. and Timkovich, R.: 1975, *The Enzymes* (ed. by P. Boyer), third ed., V. XI, pp. 397–547, Acad. Press, N.Y.
[10] Edman, P. and Begg, G.: 1967, *Eur. J. Biochem.* **1**, 80–93.
[11] Elzinga, M. E., Collins, J. H., Kuhel, W. M., and Adelstein, R. S.: 1973, *Proc. Nat. Acad. Sci. U.S.A.* **70**, 2687–2691.
[12] Johnson, P. and Perry, S. V.: 1968, *Biochem. J.* **110**, 207.
[13] Katz, A. M. and Carsten, M. E.: 1963, *Circulation Res.* **13**, 474–480.
[14] Kominz, D. R., Hough, A., Symonds, R., and Laki, K.: 1954, *Arch. Biochem. Byophys.* **50**, 148–170.
[15] Krans, H. M. J., Van Eijk, H. G., and Westenbrink, H. G. K.: 1965, *Biochem. Biophys. Acta* **100**, 193–200.
[16] Kubo, R. T., Zimmerman, B., and Grey, H. M.: 1973, *The Antigens* (ed. by M. Sela), Acad. Press, N.Y., Vol. 1, pp. 417–474.
[17] Laki, K., Maruyna, K., and Komins, D. R.: 1962, *Arch. Biochem. Biophys.* **98**, 323–330.
[18] Laursen, R. A.: 1971, *Eur. J., Radiochem.* **20**, 89–93.
[19] Ogievetskaya, M. M.: 1976, in *Proc. of the Symp. on Biophysica and Biochemistry of Muscles*, Tbilisi 1974, Nauka, p. 129.
[20] Pinaev, G. P. and Khaitlina, S. Yu.: 1972, *J. Evolution of Biochem. Physiol.* **VIII**, 369–373.
[21] Prager, E. M., Arnheim, N., Mross, G., and Wilson, A. C.: 1972, *J. Biol. Chem.* **247**, 2905–2910.
[22] Shapiro, H. M.: 1971, *Biochem. Biophys. Acta* **236**, 725–738.
[23] Trayer, I. P. and Perry, S. V.: 1966, *Biochem. Z.* **345**, 87–90.
[24] Weltman, J. K. and Dowben, R. M.: 1973, *Proc. Nat. Ac. Sci., U.S.A.* **70**, 3230–3234.
[25] Zagalsky, P. F.: 1972, *Comp. Bioch.* **41B**, 385–395.
[26] Zagalsky, P. F. and Herring, P. J.: 1972, *Comp. Biochem. Physiol.* **41B**, 397–415.