J Mol Evol (1985) 21:339–345

Journal of
Molecular Evolution

Examination of Protein Sequence Homologies: I. Eleven *Escherichia coli* L7/L12-Type Ribosomal "A" Protein Sequences from Eubacteria and Chloroplast

Eiko Otaka,[1] Tatsuo Ooi,[2] Tsutomu Kumazaki,[1] and Takuzi Itoh[1]

[1] Department of Biochemistry and Biophysics, Research Institute for Nuclear Medicine and Biology, Hiroshima University, 2-3, Kasumi 1 Chome, Minamiku, Hiroshima, 734 Japan
[2] The Institute for Chemical Research, Kyoto University, Uji, Kyoto-Fu, 611 Japan

Summary. Seven complete and four partial sequences of *Escherichia coli* L7/L12-type ribosomal "A" proteins obtained from various bacteria (*E. coli, Bacillus subtilis, Micrococcus lysodeikticus, Rhodopseudomonas spheroides, Desulfovibrio vulgaris, Streptomyces griseus, Bacillus stearothermophilus, Clostridium pasteurianum, Arthrobacter glacialis,* and *Vibrio costicola*) and spinach chloroplast have been reexamined using a computer program that searches for homologous tertiary structures. Comparison matrices for the sequences show that they match the sequence of *E. coli* L7 (EL7) if one assumes the insertion or deletion of certain residues at sites corresponding to residues 1, 38, 49, and 92 of EL7. That two additional insertion points are found only in the spinach chloroplast protein suggests that the chloroplast protein probably diverged from the bacterial forms. Further phylogenetic relationships among these 11 prokaryote-type "A" proteins are discussed with respect to average correlation coefficients computed, taking into account the existence of the gaps.

Key words: Prokaryotes — Ribosomal proteins — Correlation coefficient — Sequence homology — Protein evolution

Introduction

According to Doolittle (1981) the ultimate goal of the study of protein evolution is the reconstruction of the past events that gave rise to the vast inventory of proteins in existence today. If this goal is to be reached, a large number of amino acid sequences are needed. Fortunately, a substantial amount of sequence data on ribosomal proteins has been accumulated in the past three decades (see, e.g., Dayhoff 1972, 1973; Lin et al. 1982; Itoh and Otaka 1984). Ribosomal proteins are very attractive subjects for the study of protein evolution, since these proteins are essential for all organisms. The ribosomes, organelles for protein synthesis, contain 50 (in prokaryotes) to 70 (in eukaryotes) protein species. Eukaryotes also contain prokaryote-type ribosomes, in their mitochondria and chloroplasts.

However, when sequences of *Escherichia coli* L7/L12-type ribosomal proteins from prokaryotes and eukaryotes were examined by the usual method, i.e., by counting the number of identical residues, significant sequence homology was not found (Itoh 1980; Lin et al. 1982). Lin et al. (1982, 1983) therefore inspected their evolutionary relationships by means of a computer program that computed the identical-residue ratio (Barker et al. 1978; Jue et al. 1980). They discovered that one major event in the evolution of such proteins had been a transposition.

We have now used a computer program based on a quite different premise (Kubota et al. 1981, 1982), namely, that selection may work to conserve structural properties rather than to conserve identical residues in sequences. Therefore, the computer searches for homologous tertiary structures. The method has the merit of constructing a comparison frame by introducing the proper gaps. Among comparable sequences, the comparison frames deduced

Offprint requests to: E. Otaka

here show greater than or equal identical-residue ratio for the gap-times, compared to those reported so far. Accordingly, this method can resolve evolutionary relationships even along complex pathways. For such pathways, the average correlation coefficient discussed here should supersede the identical-residue ratio as a criterion of sequence similarity, because it is difficult to detect identical residues in distantly related proteins.

We reexamine in this paper 11 sequences of *E. coli* L7/L12-type ribosomal "A" proteins from prokaryotic organisms and a chloroplast, and discuss the structural and evolutionary divergence of prokaryote-type "A" proteins. In a subsequent paper we will analyze more "A" proteins in order to determine the ancestral relationship between prokaryote- and eukaryote-type as well as among eukaryote-type "A" proteins.

Materials and Methods

The amino acid sequences used were as follows: *E. coli* EL7 (Terhost et al. 1973); *Bacillus subtilis* BL9 (Itoh and Wittmann-Liebold 1978); *Micrococcus lysodeikticus* MA1 (Itoh 1981); *Rhodopseudomonas spheroides* RsA (Itoh and Higo 1983); *Desulfovibrio vulgaris* DvA (Itoh and Otaka 1984); *Streptomyces griseus* SA1 (Itoh et al. 1982); *Clostridium pasteurianum* CpA, *Bacillus stearothermophilus* BsA, *Arthrobacter glacialis* AgA, and *Vibrio costicola* VcA (Visentin et al. 1979); and spinach chloroplast SpCA (Bartsch et al. 1982).

Briefly, the method used to search for sequence homology was as follows: Two protein sequences X and Y were compared by computing correlation coefficients between the i-th residues of X (X(i)) and j-th residues of Y (Y(j)); the results are expressed as a comparison matrix (Kubota et al. 1981). To compute the correlation coefficient, sequences expressed as arrangements of amino acids were replaced by a series of values for physicochemical properties (e.g., hydrophobicity) corresponding to each amino acid. With the sequences now represented as numbers, we can compute a correlation coefficient for a property k between a pair at X(i) and Y(j), $C_k(i, j)$, according to the equation

$$C_k(i, j) = \frac{\sum_{m=-a}^{a} (X_k(i + m) - \langle X_k \rangle)(Y_k(j + m) - \langle Y_k \rangle)}{\left\{\left[\sum_{m=-a}^{a} (X_k(i + m) - \langle X_k \rangle)^2\right]\left[\sum_{m=-a}^{a} (Y_k(j + m) - \langle Y_k \rangle)^2\right]\right\}^{1/2}} \quad (1)$$

where $X_k(i)$ and $Y_k(j)$ are the values of property k for the amino acids at positions X(i) and Y(j), and X_k and Y_k are average values for property k. The length of regions centered around the i-th and j-th residues, $2a + 1$, was taken as 11 unless otherwise noted.

The sequence homology shows good correspondence to the homology of tertiary structures when the following six properties are used: propensity to form turns, polarity, partial specific volume, pK value of alpha-amino group, pK value of alpha-carboxyl group, and mutability. Thus, the correlation coefficient over all these properties, C(i, j), is obtained as an arithmetic average:

$$C(i, j) = \frac{1}{6} \sum_{k=1}^{6} C_k(i, j) \quad (2)$$

All computations were performed on a Facom 180AD computer at the Computer Center of the Institute for Chemical Research of Kyoto University.

Results and Discussion

Sequence Comparisons with EL7

Bacillus subtilis (BL9). Figure 1a shows a comparison matrix for the sequences of BL9 and EL7 in which are plotted C(i, j) values greater than 0.4. The heavier dots, which represent values greater than 0.7, cluster near the diagonal, indicating the existence of three sequential well-matching segments covering almost the entirety of the sequences. Figure 2 shows the alignment derived from this matrix, along with the individual correlation coefficients (multiplied by 10) computed from Eqs. (1) and (2) for each pair. The first segment with values greater than 0.4 consists of residues 1–44 of BL9, corresponding to the same residues of EL7, and the second such segment consists of residues 50–92 of BL9 matched with the EL7 residues of the same number. We thus obtain a continuous alignment of the segments (i.e., one without gaps) for residues 1–92. It is more natural than other alignments, despite correlations less than 0.4 for residues 45–49. The third segment, a match between residues 94–122 of BL9 and residues 92–120 of EL7, is connected with the second as marked with circles in Fig. 2 by following the fall in the value. It is clear that there was either an insertion of residues 93 and 94 in BL9 or a deletion of two residues between residues 92 and 93 in EL7. This contrasts with the comparison frame of Itoh and Wittmann-Liebold (1978), and matched residues 1–92 and 95–122 of BL9 with residues 1–120 of EL7.

Micrococcus lysodeikticus (MA1). A similar comparison matrix for MA1 and EL7 shows two long matching segments lying nearly on the diagonal. At the connecting point there is a deletion of at least one residue corresponding to residue 38 of EL7; in total, the MA1 sequence is two residues shorter than that of EL7, since there is also a deletion of one residue corresponding to residue 1 of EL7 (Fig. 2). The resultant comparison frame is not much different from that of Itoh (1981).

Rhodopseudomonas spheroides (RsA). Three long and one short (11 residues) segment are found nearly on the diagonal of the comparison matrix for RsA and EL7. The alignment of the first long segment reveals a deletion corresponding to residue 1 of EL7 (Fig. 2). At the three points connecting the four segments are found an insertion of one residue, an in-

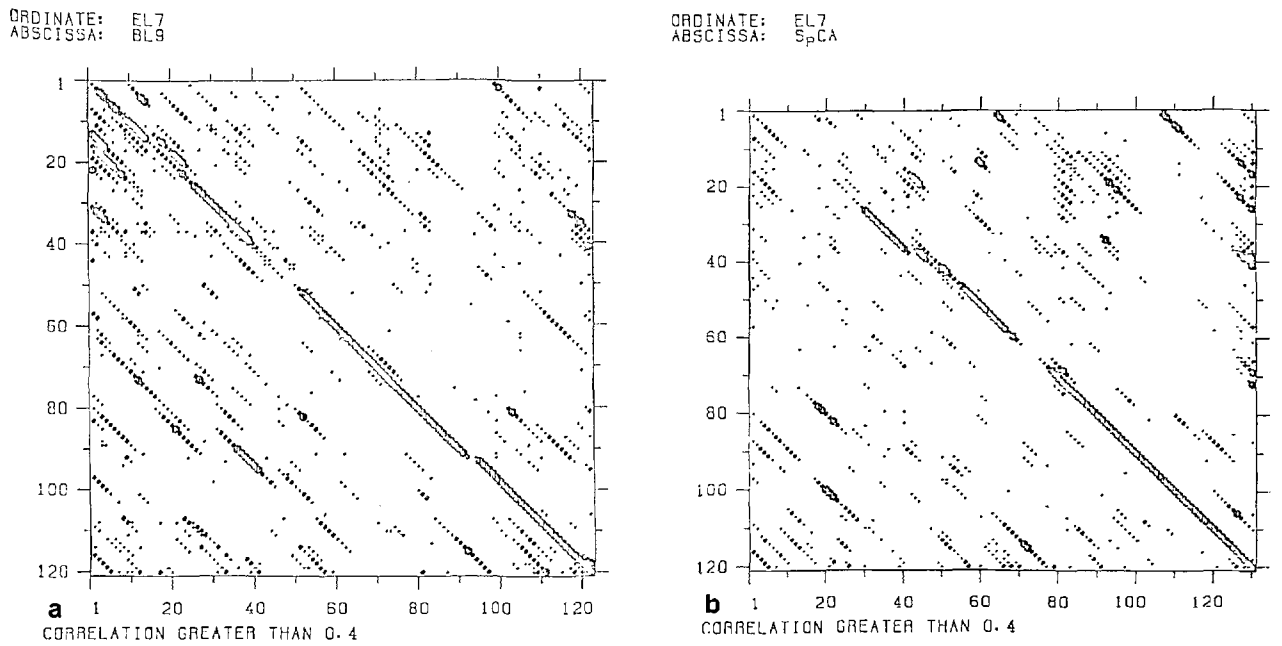


Fig. 1a, b. Comparison matrices obtained using a computer program that searches for homology between tertiary structures (Kubota et al. 1981, 1982). **a** Comparison between proteins EL7 (*E. coli* L7/L12) and BL9 (*B. subtilis* A protein). **b** Comparison between proteins EL7 and SpCA (spinach chloroplast A protein)

sertion of five residues, and a deletion of one residue, respectively, in RsA as compared with EL7. In total, the RsA sequence is four residues longer than that of EL7.

Desulfovibrio vulgaris (DvA). Two long matching segments are located nearly on the diagonal of the comparison matrix for DvA and EL7, with a short matching segment overlapping the two long ones as shown in Fig. 2. At the two connecting points with the short segment, two residues and three residues, respectively, are inserted in DvA as compared with EL7. These insertions, together with that of the first residue of DvA, yield a sequence longer by six residues than that of EL7.

Streptomyces griseus (SA1). As shown in Fig. 2, the comparison frame deduced for SA1 and EL7 has a short matching segment similar to that for DvA, and a segment of relatively low correlation in the latter part of the sequence. SA1, as compared with EL7, has an insertion of the first residue, and a deletion of one residue and an insertion of five residues at the two connecting points. This gives a total sequence five residues longer than that of EL7.

Clostridium pasteurianum (CpA), Bacillus stearothermophilus (BsA), Arthrobacter glacialis (AgA), and Vibrio costicola (VcA). Correlations of partial sequences of CpA, BsA, AgA, and VcA with that of EL7 are also shown in Fig. 2. CpA and BsA each have a deletion of one residue, and AgA has an insertion of one residue, at their respective N-termini.

Spinach chloroplast (SpCA). Figure 1b shows the comparison matrix for SpCA and EL7, containing some clusters of thick dots located near the diagonal. The most reasonable alignment is that shown in Fig. 2, according to which SpCA has insertions of four residues from SpCA positions 10 to 13, five residues from positions 42 to 46, and one residue around position 74 compared with EL7. In total, SpCA is 10 residues longer than EL7. This comparison frame differs considerably from that of Bartsch et al. (1982).

Drastic Events During Protein Evolution

Overall in Fig. 2, it is found that such drastic events as deletions and insertions (gaps) have occurred in very limited regions around the sites corresponding to residues 1, 9, 38, 49, 64 (or so), and 92 of EL7. These points are marked with arrows in Fig. 2, and are designated A, B, C, D, E, and F, respectively, in Fig. 3, which summarizes schematically the correlations of the 10 "A" proteins with EL7 shown in Fig. 2. It is suggestive that the region of weak correlation in BL9 is located at site D. Presumably, the fact that the insertions at sites B and E are seen only in SpCA reflects the phylogenetic branching of the chloroplast from the bacteria treated here.

Since we now see that the gaps were introduced systematically, the gap must be recognized anew as a drastic substitution during the evolutionary his-

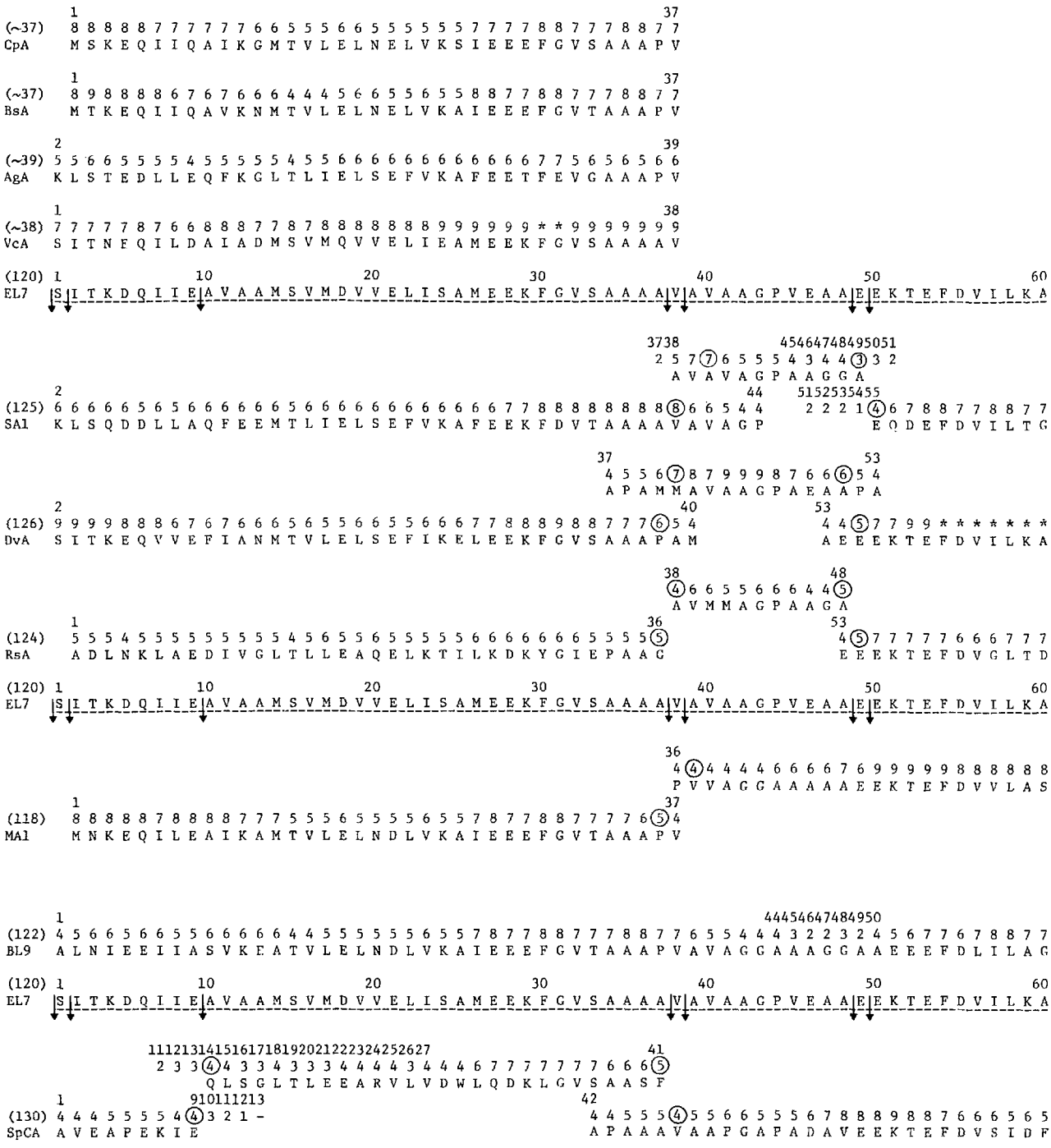
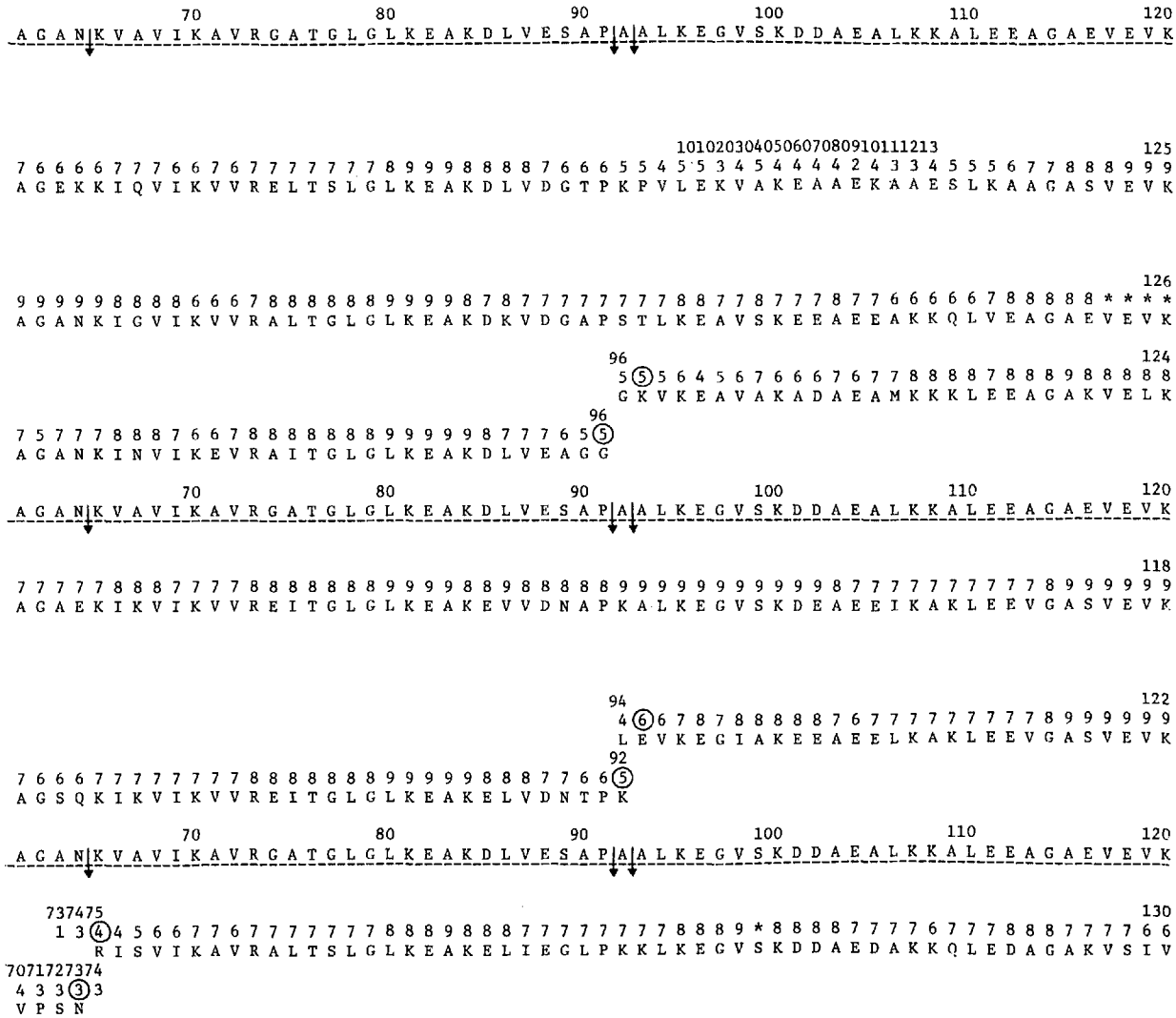


Fig. 2. Correlation of sequences of ten prokaryote-type A proteins with that of EL7 (underlined). See Materials and Methods for sources of A proteins sequences. The correlation coefficient (multiplied by 10) for each residue pair is shown above the letter symbol for the residue from the aligned sequence. Residue numbers are placed above coefficients of the first and the last residues of well-matching sequences (i.e., stretches with correlation values greater than 0.4). Continuous strings of residue numbers indicate stretches with correlation values below 0.4. See text for further details

tory of "A" proteins. The resultant fixed gaps represent the so-called neutral mutation sites that have been contrastingly described as frequent occurrences by Ohno (1970) and as very rare mutations by Simpson (1964). After due consideration of cases that were not neutral and not fixed, we conclude that such events have occurred very frequently. Although we emphasize the evolutionary importance of such gaps, we unfortunately do not know how to interpret gaps numerically from a phylogenetic standpoint.

Since the parameters for computation of correlation coefficients are selected so as to reveal homology among tertiary structures, the common re-



gions separated by sites A–F in Fig. 3 must represent important units for the architecture of "A" proteins. The report following this one will show that these units exist in eukaryote-type "A" protein sequences as well. It is worth mentioning that the consensus residues conserved in all the species treated here are located in three places: residues 54–56 (54 and 55 only for BL9) of EL7, in the D–E region; and residues 68–70 and 78–84, in the E–F region. Two modified (methylated) lysines in DvA are included in the consensus residues. These facts suggest the important role of these regions for structure.

Average Correlation Coefficient

The comparison matrices were constructed according to correlation coefficients computed using Eq. (1), in which segments of 11 residues (including the five residues before and after the i-th residue of one sequence and the j-th residue of the other) are considered. This expression is convenient for locating homologous regions in sequences. Once a comparison frame is constructed, however, an average correlation coefficient over all the entire sequence, or a particular region, yields a quantitative estimate of

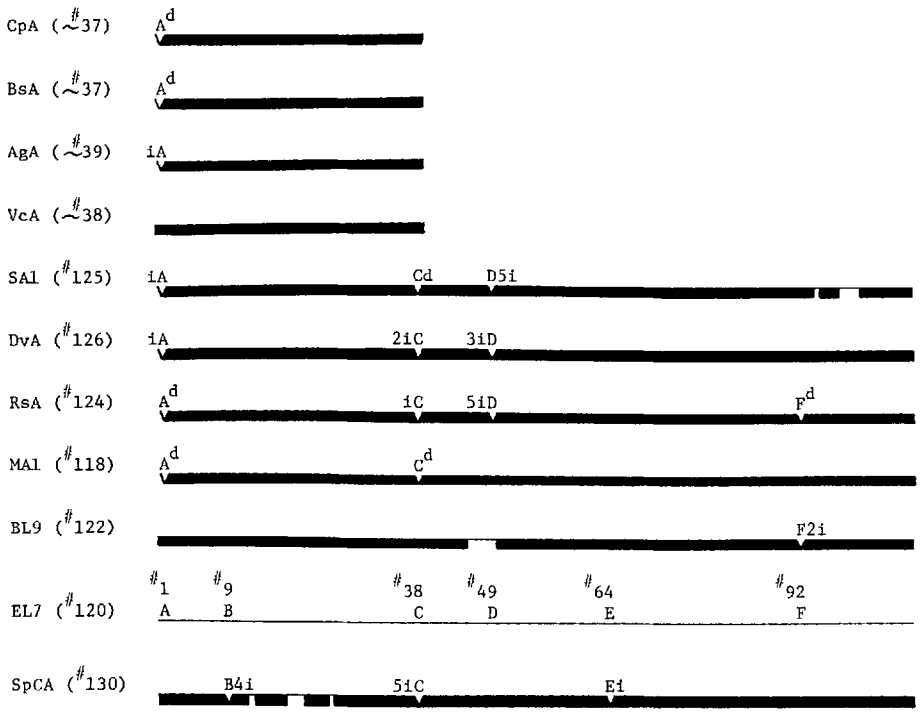


Fig. 3. Schematized correlation of ten prokaryote-type A proteins with EL7. Thick lines represent well-matching stretches indicated in Fig. 2. Letters A–F on the EL7 line correspond to the residue numbers shown. On other sequence lines the letters are accompanied by the number of residues deleted (d) or inserted (i) before and/or after these sites. For example, "B4i" on the SpCA line indicates an insertion of four residues after site B of SpCA, corresponding to residue 9 of EL7. Similarly, "5iC" on the SpCA line indicates an insertion of five residues before site C in SpCA, corresponding to residue 38 of EL7. A superscript "d" indicates deletion of the corresponding residue only; thus "A^d" indicates deletion of site A (residue 1 of EL7). Numbers in parentheses next to sequence names indicate total lengths of sequences shown

Table 1. Average correlation coefficients between A–C regions (see Fig. 3) of prokaryote-type "A" proteins

	"A" protein	EL7	VcA	DvA	MA1	CpA	BsA	SA1	BL9	AgA	RsA
VcA	(*Vibrio costicola*)	0.82									
DvA	(*Desulfovibrio vulgaris*)	0.67	0.66								
MA1	(*Micrococcus lysodeikticus*)	0.69	0.58	0.74							
CpA	(*Clostridium pasteurianum*)	0.61	0.54	0.74	0.91						
BsA	(*Bacillus stearothermophilus*)	0.60	0.59	0.82	0.91	0.91					
SA1	(*Streptomyces griseus*)	0.63	0.60	0.63	0.65	0.61	0.64				
BL9	(*Bacillus subtilis*)	0.55	0.48	0.67	0.87	0.83	0.86	0.63			
AgA	(*Arthrobacter glacialis*)	0.52	0.52	0.66	0.70	0.71	0.66	0.78	0.69		
RsA	(*Rhodopseudomonas spheroides*)	0.52	0.40	0.48	0.53	0.53	0.49	0.43	0.54	0.48	
SpCA	(spinach chloroplast)	0.47	0.42	0.41	0.41	0.46	0.42	0.45	0.46	0.46	0.43

sequence homology. This provides an alternative to the usual method of comparison, i.e., the identical-residue ratio. Tables 1 and 2 give examples of such average correlation coefficients. Since the A–C region is known for all 11 proteins, the values of all sequence pairs were computed for this region (Table 1). The relatively low values for comparisons with SpCA imply that the evolutionary distances from the eubacteria to spinach are almost equal, supporting a finding in the preceding section (see also Fig. 3). EL7 and VcA are both from Gram-negative bacteria, and naturally show good homology. Although *Desulfovibrio vulgaris*, a Gram-positive bacterium, is considered a member of the oldest group of eubacteria, its "A" protein is shown here to be very similar even to EL7 and VcA. However, insertion/deletion events have occurred at least three times as it and EL7 have diverged.

Higher values for the D–F region than for the A–C region are found for the proteins whose whole sequences are known (Table 2), suggesting that this region is the most characteristic for "A" proteins, as mentioned above. Even SpCA shows a high correlation with EL7 in this region. The evolutionary distances suggested by the average correlation coefficients for the complete sequences (Table 2) are not quite compatible with those estimated from the A–C region alone. While DvA and MA1 are close to EL7, others indicate similar distances to EL7. Since the extent of measured sequence conservation varies with the sequence region considered, complete sequencing of proteins is a prerequisite for examinations of protein evolution. Further comparison should be done after elucidation of the gaps. Presumably, it signifies a shortage of the gap-factor in the values that the correlation coefficient values in Tables 1 and 2 do not always suggest phylogenetic relations identical to those deduced from RNA se-

Table 2. Average correlation coefficients between prokaryote-type A proteins and *E. coli* protein L7

		Sequence region[a]			
	"A" protein	A–C	D–F	D–end	Whole sequence
DvA	(*Desulfovibrio vulgaris*)	0.67	0.86	0.83	0.79
MA1	(*Micrococcus lysodeikticus*)	0.69	0.85	0.84	0.75
SA1	(*Streptomyces griseus*)	0.63	0.72	0.65	0.60
BL9	(*Bacillus subtilis*)	0.55	0.78	0.78	0.66
RsA	(*Rhodopseudomonas spheroides*)	0.52	0.74	0.73	0.66
SpCA	(spinach chloroplast)	0.47	0.78	0.68	0.62

[a] See Fig. 3

quence studies (Hori and Osawa 1979; Hori et al. 1982).

Although the correlation coefficients determined here are compatible with the identical-residue ratios, we expect that use of the former will supersede use of the latter in comparisons among more distantly related proteins (e.g., comparisons between prokaryotic and eukaryotic proteins).

We believe that these computed data will provide a new perspective on protein evolution.

Acknowledgments. We are grateful to Professor J.R. Warner (Albert Einstein College of Medicine of Yeshiva University) for improving the English and commenting helpfully on the manuscript. This work was supported by Grant-in-Aid for Scientific Research no. 58212014 to T.I. from the Ministry of Education of Japan.

References

Barker WC, Ketcham LK, Dayhoff MO (1978) A comprehensive examination of protein sequences for evidence of internal gene duplication. J Mol Evol 10:265–281

Bartsch M, Kimura M, Subramanian A (1982) Purification, primary structure, and homology relationships of a chloroplast ribosomal protein. Proc Natl Acad Sci USA 79:6871–6875

Dayhoff MO (ed) (1972) Atlas of protein sequence and structure, vol 5. National Biomedical Research Foundation, Washington DC

Dayhoff MO (ed) (1973) Atlas of protein sequence and structure, vol 5, suppl 1. National Biomedical Research Foundation, Washington DC

Doolittle RF (1981) Similar amino acid sequences: chance or common ancestry? Science 214:149–159

Hori H, Osawa S (1979) Evolutionary change in 5S RNA secondary structure and a phylogenic tree of 54 5S RNA species. Proc Natl Acad Sci USA 76:381–385

Hori H, Itoh T, Osawa S (1982) The phylogenic structure of the metabacteria. Zentralbl Bakteriol Mikrobiol Hyg [C] 3: 18–31

Itoh T (1980) Primary structure of yeast acidic ribosomal protein YPA1. FEBS Lett 114:119–123

Itoh T (1981) Primary structure of an acidic ribosomal protein from *Micrococcus lysodeikticus*. FEBS Lett 127:67–70

Itoh T, Higo K (1983) Complete amino acid sequence of an L7/L12-type ribosomal protein from *Rhodopseudomonas spheroides*. Biochim Biophys Acta 744:105–109

Itoh T, Otaka E (1984) Complete amino acid sequence of an L7/L12-type ribosomal protein from *Desulfovibrio vulgaris*, Miyazaki. Biochim Biophys Acta 789:229–233

Itoh T, Wittmann-Liebold B (1978) The primary structure of *Bacillus subtilis* acidic ribosomal protein B-L9 and its comparison with *Escherichia coli* proteins L7/L12. FEBS Lett 96: 392–394

Itoh T, Sugiyama M, Higo K (1982) The primary structure of an acidic ribosomal protein from *Streptomyces griseus*. Biochim Biophys Acta 701:164–172

Jue RA, Woodbury NW, Doolittle RF (1980) Sequence homologies among *E. coli* ribosomal proteins: evidence for evolutionarily related groupings and internal duplications. J Mol Evol 15:129–148

Kubota Y, Takahashi S, Nishikawa K, Ooi T (1981) Homology in protein expressed by correlation coefficients. J Theor Biol 91:347–361

Kubota Y, Nishikawa K, Takahashi S, Ooi T (1982) Correspondence of homologies in amino acid sequence and tertiary structure of protein molecules. Biochim Biophys Acta 701: 242–252

Lin A, Wittmann-Liebold B, McNally J, Wool IG (1982) The primary structure of the acidic phosphoprotein P2 from rat liver 60S ribosomal subunits. J Biol Chem 257:9189–9197

Lin A, McNally J, Wool IG (1983) The primary structure of rat liver ribosomal protein L37. Homology with yeast and bacterial ribosomal proteins. J Biol Chem 258:10664–10671

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

Simpson GG (1964) Organisms and molecules in evolution. Science 146:1535–1538

Terhorst C, Möller W, Laursen R, Wittmann-Liebold B (1973) The primary structure of an acidic protein from 50-S ribosomes of *Escherichia coli* which is involved in GTP hydrolysis dependent on elongation factors G and T. Eur J Biochem 34: 138–152

Visentin LP, Yaguchi M, Matheson AT (1979) Structural homologies in alanine-rich acidic ribosomal proteins from procaryotes and eucaryotes. Can J Biochem 57:719–726

Received May 7, 1984/Revised December 20, 1984