Multiple internal repeats within protein S1 from the *Escherichia coli* ribosome

Brigitte Wittmann-Liebold, Keith Ashman and Michael Dzionara

Max-Planck-Institut für Molekulare Genetik, Abteilung Wittmann, D-1000 Berlin 33 (Dahlem), Germany

Received 31 January 1983

The complete sequence determination of protein S1, the largest protein from the *Escherichia coli* ribosome, revealed that it is composed of repeated internal duplications, mainly at the central region of the molecule which contains the mRNA-binding domain [Eur. J. Biochem. (1982) 123, 37–53]. With the aid of computer programs the statistical significance of the internal repeats in S1 was proven. Auto-comparison of the S1-sequence showed that it is composed of 87-residue strings with 44-residue subunits: 3 strings (residues 189–447) are highly related; 3 strings (residues 13–188 and 448–533) are less but significantly related. Statistical analysis revealed a more distant relatedness for the 44-residue subunits than for the 87-residue strings. Protein S1 was compared to all other *E. coli* ribosomal proteins and to the 1100 primary structures listed in the last Atlas of Protein Sequence and Structure (1978) showing parts of S1 distantly related with parts of several ribosomal proteins. However, distinct homologies between protein S1 and the other ribosomal proteins can be ruled out. The strongest repeats within the S1 sequence were mainly found corresponding to the mRNA-binding domain. Distantly related partial sequences were also found with ribosome-associated and nucleotide-binding proteins, with some enzymes, with several peptide hormones and with contractile proteins.

Ribosomal protein S1 *Internal repeats, significance of* *Protein comparison* *Computer analysis* *Nucleotide-binding and contractile protein similarities* *S1 gene evolution*

1. INTRODUCTION

Protein S1 plays an important role in the translation of mRNA. It is required for the binding of the messenger to the ribosome during the initiation step (details in [1]). As the α-subunit of the $Q\beta$ phage replicase complex it is involved in the transcription of the phage RNA plus strand. Gentle proteolysis [1], small angle X-ray scattering [2,3] and in vivo mRNA translation experiments with an S1 mutant [4] have shown that S1 has several structural domains. The N-terminal region has ribosome binding ability, whilst the central region is concerned with the mRNA binding activity [1].

The complete primary structure of S1 was determined in [5,6]. These studies showed that the functions of this protein are reflected in its primary, secondary and, probably, tertiary structure. Protein S1 is very large in comparison to all the other *Escherichia coli* ribosomal proteins (M_r 61 159). It is highly elongated (~25 nm) and has homologous sequence repeats. Three 24-residue stretches have 67% identical amino acids and additional conservative replacements. Further, these repetitious sequence areas extend throughout the S1 molecule and can be grouped into regions of 87 residues, which may be subdivided into 40–45-residue units [6].

Here, we present the details of the statistical analysis performed with the computer programs ALIGN and RELATE [7,8] and the results obtained with our search programs developed to detect homologies between distantly related proteins. The significance of the relationship between the various 87-residue stretches and between the various

Published by Elsevier Science Publishers
00145793/83/0000–0000/$3.00

40–45-residue strings is given, and the evolutionary aspects of the occurrence of these strings and their correlation to the functional domains are discussed. The entire S1 sequence and the repeats were compared with all the other *E. coli* ribosomal proteins. The S1-sequence was also compared with all the protein sequences contained in [9]. The results of these comparisons are given and discussed.

2. METHODS

2.1. *Search program*

Searches for internal homologies in protein S1, and for sequence stretches homologous to S1 in the other *E. coli* ribosomal proteins, as well as some proteins from the literature, were made with the FORTRAN program SEEK. This was developed to search for short regions of homologous sequence in ribosomal proteins. All possible segments of a defined length, parameter *A*, (e.g., 10, 20 or 30 residues) of one sequence are compared with all possible segments of the same length of another sequence. The program requires two further parameters. Parameter *B*, which is the minimum number of amino acid residues of identical type and position which must be present within a segment, and parameter *C* which limits the non-identical amino acids by demanding that $\geqslant 30\%$ of them differ only by one base in their codons. All segments containing more than the specified number of matches and allowed mismatches are listed. The values of *B* and *C* were chosen such that they were $\geqslant 30\%$ of *A* (e.g., $\geqslant 30\%$ of the residues must be identical). The following parameters for *A*, *B* and *C* were employed: 20/7/7 for the search of internal repeats within protein S1; 15/5/7, 10/5/3 or 10/4/2 for the detection of related proteins or of distantly related sequence stretches, respectively.

2.2. *Statistical analysis*

RELATE and ALIGN [7,8] were used to calculate the statistical significance of the homologies found, and to assess possible evolutionary relationships. The alignment scores (from ALIGN) and the segment comparison scores (from RELATE) give, in standard deviation (SD) units, estimates of the relatedness of two sequences. As scoring matrix the mutation data matrix [9] was predominantly employed: > 10 SD units indicated strong homology; 3–10 SD units indicated a distant but still significant evolutionary relationship; 2–3 SD units indicated a potential homology [10]. To arrive at the scores, both ALIGN and RELATE compare the results obtained with the actual sequences to those obtained with random permutations of the sequences being examined.

To analyse the internal homologies of S1, the entire sequence was examined using RELATE. The segment lengths employed were 10, 20, 25, 40, 80, 87 and 174, and for each run 100 randomisations of the sequence were performed. The 87-residue strings were all compared with each other. The segment lengths were 10, 15, 20, 25, 40 and 43. Similarly, the 44-residue strings were investigated, but with segment lengths of 8, 12, 20 and 30. In all cases 100 randomisations of the sequence were carried out.

The 87-residue strings were also subjected to analysis with ALIGN. In this case, both the mutation data and genetic code scoring matrices were used [9], and break penalties of 2, 10, 20, 40 and 80 were employed (N.B., the bias was set at 0).

RELATE was used for the comparison of S1 with all the other *E. coli* ribosomal proteins (fragment lengths 12 and 25, with 50 randomisations) and for comparison with the proteins in [9] (fragment length 25; 50 randomisations). When segment comparison scores over 2.0 SD units were obtained those regions of S1 and of the various proteins that RELATE suggested to be related were further analysed with ALIGN.

The computer analyses with SEEK were performed on a DEC-10 computer and the RELATE and ALIGN runs were made on a VAX/VMS system.

2.3. *Proteins used for the comparison*

The following proteins were compared with protein S1: all 53 ribosomal proteins derived from the *E. coli* ribosome (for references see [11], except for S2 [12], L2 [13], L9 [14], L17 [15] and S14 [16]); proteins associated with the ribosome, namely NS1 and NS2 [17] as well as the factors, EC-IF-1 [18], EC-IF-3 [19] and EC-EF-Tu [20]; all protein sequences contained in [9].

3. RESULTS AND DISCUSSION

3.1. *Multiple-repeated segments within the sequence of protein S1*

Visual inspection of the S1 sequence showed it to have several similar segments of considerable homology. Interestingly, it was found that the part of the sequence shown to contain the mRNA-binding ability [1] consists of repeated sequences [6]. Employing statistical methods we have determined the lengths of repeats that have the highest significance and whether these repeats can be further subdivided and still maintain a statistically significant relationship. Further, the statistical approach offers an objective means for testing whether or not the structural relationship of the predominant repeats can be followed throughout the entire sequence.

With RELATE [8] the complete sequence of S1 was examined. Fragments of a defined length were compared with all other possible fragments of the same length, and various lengths were tried. The highest segment comparison scores (in SD units), which are a measure for the internal relatedness of the sequence, were obtained for an 87-residue segment (table 1). These scores strongly support the idea of internal repeats within S1. The 87-residue displacement was found 43–102 times (depending on the segment length investigated) and is the most significant. Furthermore, displacements of 172–174 residues were less frequently obtained, suggesting that the 87-residue strings are possibly internal repeats of stretches, double the length.

Considering the distribution of seldomly occurring amino acids, such as tryptophan, tyrosine, phenylalanine and histidine, visual inspection allows the 87-residue strings to be aligned (fig.1). By permitting a few deletions, almost the entire sequence of S1 can be divided into six 87-residue strings. Independently of this, the subdivision of the sequence was made by SEEK, to compare and plot homologous sequence stretches. The result is presented in fig.2 and shows a similar alignment for the 87-residue strings as in fig.1.

Table 1

Segment comparison scores for protein S1

Fragment length	Compared fragments	Number of top scores used	Segment comparison scores (SD units)	Displacement
10	149878	269	18.9	43 × − 87 13 × −172 12 × −174
20	144453	259	29.9	57 × − 87 15 × −174 10 × −172
25	141778	254	30.5	69 × − 87 14 × −174 8 × −172
40	133903	239	32.2	96 × − 87 6 × −174
80	114003	199	33.9	101 × − 87
87	110685	192	34.3	101 × − 87
174	73536	105	31.2	102 × − 87

The complete sequence was compared with RELATE (Scoring matrix: mutation data matrix; fragment lengths: 10, 20, 25, 40, 80, 87 and 174 residues; 100 random runs)

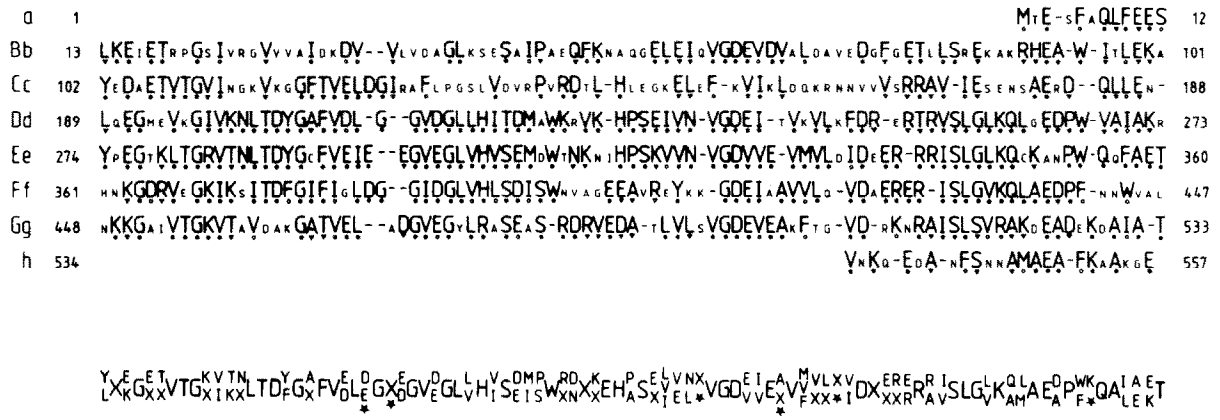

Fig.1. Alignment of the 87-residue strings within protein S1. The alignment was made with the aid of ALIGN and by visual inspection of the sequence: capital letters, most conserved residues; symbols beneath, identical amino acids (black symbols) or related amino acids (open symbols); bottom line, conserved residues within the 87-residue strings; *means, assumed deletion of a residue in some of the strings.

Fig.2. Auto-comparison of string Ee (position 274–360) of S1 with the other 87-residue segments. The alignment was made by SEEK under the conditions: 20/7/7 (see section 2 for details). Only the strongest repeats, namely Ee:Dd, Ee:Ff and Ee:Gg were matched under these conditions. (*) Identical residues at identical positions within the compared segments; (+) codons of the matched amino acids differ by one nucleotide; (−) codons differ by two nucleotides. None of the listed homologous stretches had residues that were unrelated in their codons (> 2 nucleotide replacements).

3.2. *Relatedness of the 87-residue strings within protein S1*

Besides the distribution of seldomly occurring amino acids the arrangement of the hydrophobic amino acids and the strong conservation of the glycine residues is remarkable within the sequence of S1. Most of Trp, His, Phe and Tyr and of the charged residues show the same distribution (fig.1). Strongly homologous are the strings Dd, Ee, Ef and Gg (pos. 189–273, 274–360, 361–447 and 448–533, respectively). Similarities in alternate strings can be observed by inspecting certain Glu/Asp and Ile/Val distributions, and this serves as a guide for locating the 172/174-residue stretches. When counting the number of identical residues occurring in identical positions within the

87-residue strings the greatest similarity is between strings Dd and Ee (if two deletions are allowed) with 49% identity. Strings Dd and Ff are almost as related (41% identity, allowing two deletions). The comparison of the other possible combinations of the 87-residue strings from the sequence between positions 189–533 results in 28–31% identity. The relatedness for the other 87-residue combinations are less pronounced.

The significance of the relatedness of the 87-residue strings was investigated with RELATE (table 2). The highest segment comparison scores obtained are listed. According to these data, the strings Ee and Ff are most closely related, contrary to the above conclusion (where deletions were permitted). The strings are listed in table 3 in decreasing order of their segment comparison scores. Both RELATE and ALIGN were employed to examine similarities and both programs gave the best fit for strings Ee:Ff, followed by Ee:Dd and Dd:Ff. The homology of these 3 strings to string Gg (pos. 448–533) is weaker, but still significant (fig.3, table 3).

Results in tables 2 and 3 show the distant relationship of string Bb (pos. 13–101) to strings Ee, Dd, and the C-terminal sequence of S1, strings Gg

Table 2

Segment comparison scores of the 87-residue strings of protein S1

String positions	Bb (13–101)	Cc (102–188)	Dd (189–273)	Ee (274–360)	Ff (361–447)
Cc (102–188)	−0.85				
	−0.38				
	−0.13				
	−0.42				
	−0.65				
Dd (189–273)	0.93	2.06			
	1.20	3.69			
	1.46	3.62			
	0.61	3.58			
	0.15	3.47			
Ee (274–360)	2.14	2.11	16.02		
	2.27	2.09	15.94		
	2.99	2.19	14.73		
	1.20	1.69	11.63		
	0.79	1.46	11.05		
Ff (361–447)	0.19	1.38	13.41	12.82	
	0.59	2.23	14.54	14.73	
	1.28	2.25	12.45	14.75	
	1.64	1.22	7.80	17.50	
	1.37	1.04	7.23	18.41	
Gg (448–533)	2.13	1.30	3.74	6.07	4.38
	2.23	2.40	5.84	7.73	5.70
	1.88	2.18	5.84	8.42	5.11
	2.75	1.59	8.82	6.06	4.72
	2.78	1.53	9.29	5.73	4.36

Comparison was made with RELATE (scoring matrix: mutation data matrix; fragment lengths, 15, 20, 25, 40 and 43 residues; 100 random runs). Listed are the fragment comparison scores (in SD units) in order of increasing fragment length (from top to bottom)

Table 3

Comparison of the 87-residue strings of protein S1

String compared	Alignment scores (SD units)[a]	Segment comparison scores (SD units)[b]
Ee:Ff	29.6 (40)	17.5 (40)
Ed:Dd	25.3 (40,20)	15.9 (20)
Dd:Ff	21.1 (20)	14.5 (20)
Ee:Gg	9.7 (20)	7.7 (20)
Dd:Gg	9.4 (20)	8.8 (40)
Ff:Gg	8.0 (20)	5.7 (20)
Cc:Dd	5.3 (10)	3.7 (20)
Bb:Ee	5.0 (40)	2.3 (20)
Cc:Gg	4.8 (10)	2.4 (20)
Bb:h	3.5 (10)	2.9 (40)
Bb:Gg	–	2.7 (40)
Cc:Ff	–	2.2 (20)
Cc:Ee	–	2.1 (20)
Bb:Dd	3.4 (2)	1.4 (10)
a:Dd	2.2 (10)	0.7 (20)
a:Bb	1.3 (10)	0.7 (20)
a:Cc	0.9 (2)	−0.4 (20)
a:Ee	0.8 (10)	
a:Ff	0.4 (2)	

[a] Break penalty values in parentheses
[b] Fragment lengths in parentheses

The comparison was made with ALIGN (scoring matrix: mutation data matrix; break penalties, 40, 20, 10 and 2; 100 random runs) and with RELATE (same matrix; segment lengths, 20, 40; 100 random runs). The highest alignment and segment comparison scores (both in SD units) obtained are listed in decreasing order

and h, but no significant similarity was found between Bb, Cc and Ff. Further, string Cc (pos. 102–188) is less related to most of the other strings and only distantly related to strings Dd and Gg.

Most of the amino acid alterations observed when superimposing the 87-residue strings on each other can be regarded as conservative replacements, such as the change from one hydrophobic amino acid to another; e.g., Val to Ile or Leu, Ile to Trp, or Phe to Tyr. Charged residues (e.g., Arg, Lys, Glu or Asp) or small residues (e.g., Ala, Gly, Thr or Ser) frequently replace one another. The most frequent amino acids in a particular position in the 87-residue strings are listed on the bottom line of fig.1. This sequence shows the special conservation of glycine, the hydrophobic and charged residues (mainly that of glutamic acid).

3.3. *Occurrence of 44-residue repeats within protein S1*

The distribution of the tryptophans in the 87-residue strings, 41–44 positions apart, the locations of the histidines, of some prolines and the repeated Gly–Leu sequences made it likely that the 87-residue strings might be repeats of smaller sized units. A manual alignment of the S1 sequence into 44-residue repeats (fig.4) was made by subdividing each of the 87-residue strings into two portions: the N-terminal half, labelled with a capital letter; and the C-terminal half, with a small letter. The origin of each 44-residue unit can be easily correlated with the 87-residue strings, such as Cc, Dd,

```
Ee : Ff   >   Ee : Dd   >   Dd : Ff   >   Ee : Gg   >   Dd : Gg   >   Ff : Gg
 29.6          25.3          21.1          9.7           9.4           8.0

>   Cc : Dd   >   Bb : Ee   >   Cc : Gg   >   Bb : H   >   Bb : Dd
      5.3           5.0           4.8           3.5         3.4
```

Fig.3. Relationship of the 87-residue strings of protein S1. With ALIGN (conditions of table 3) alignment scores > 3.0 SD units were found for 11 pairs of the 87-residue strings (listed below). According to these values 6 out of these pairs show a strong homology (upper line), and 5 pairs are weaker but still significantly related (lower line). The related 87-residue string pairs are listed in decreasing order of their relatedness.

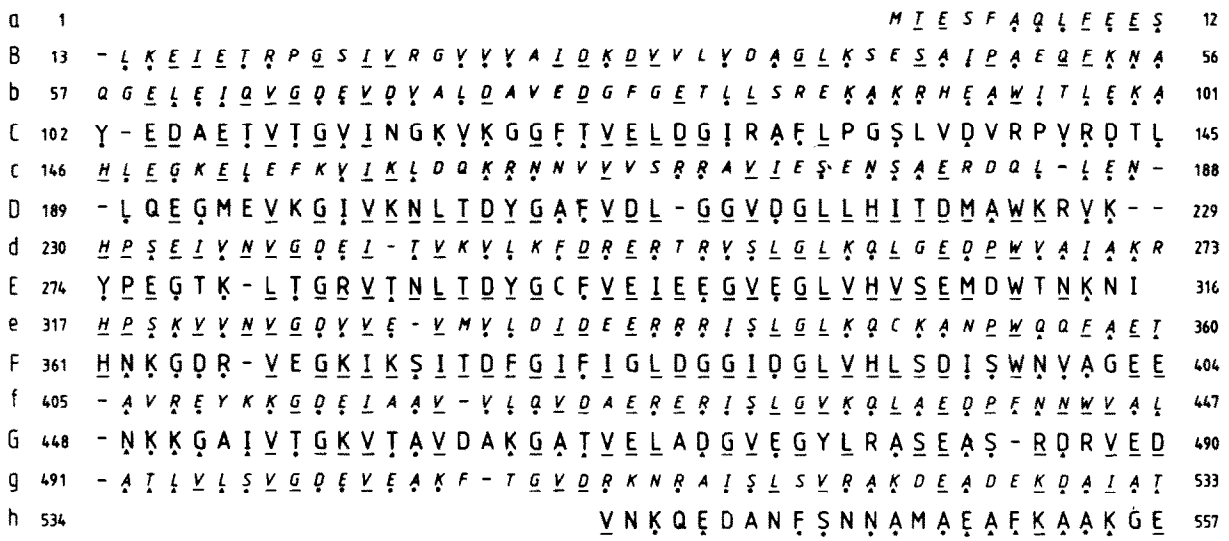

Fig.4. Alignment of the 44-residue strings within protein S1. Origin of the 87-residue strings: capital letters, N-terminal half of the 87-residue segment; lower case letter, C-terminal half; related sequences are given in italics and upright letters, respectively.

Ff. Visual inspection of the 44-residue strings gives best relatedness for those strings printed in italics, whereas the strings in upright letters are more related to each other than to the ones given in italics. String Bb seems not to follow the otherwise alternating similarity. The significance of these relationships has been examined by the above programs. The maximum scores obtained for different fragment lengths are listed in table 4. Not surprisingly a significant relatedness was found for those strings deriving from homologous areas of the 87-residue strings: e.g., E is significantly related with F; D with F; and G with E; all deriving from the N-terminal portions of the 87-residue strings. Similarly, the strings d:f, e:f and e:g from the C-terminal halves are strongly related. But also, 44-residue strings, that derive from *different* parts of the 87-residue strings were found to be related, with significant segment comparison scores; i.e., > 3.0 SD units: B:e, B:d and B:f. Segment comparison scores of 2.0–3.0 SD units, indicating distantly related sequences were obtained for the pairs D:e, b:C, G:g and G:d. However, string D and d were not found to be related at all, according to the scores values obtained. Summarizing the results with RELATE: at least some of the 44-residue strings were found to be homologous or distantly related with each other although they derived from N- and C-terminal halves of the 87-residue segments.

With SEEK similar results were obtained if less rigorous search conditions were employed than applied for the comparison of the 87-residue strings, again indicating the more distant relationship of these repeats (not shown).

3.4. *Subdivision of the 44-residue strings into smaller sized segments*

The observed 172–174/87/44-residue repeats suggested that they might be constructed of still smaller sized segments. A similar construction from small-sized units which were the result of multiple condensations of the corresponding genes during evolution has been observed for other protein classes, such as the Ca^{2+}-binding proteins of muscle, made from repeats of 7-residue segments, or the proteins of the collagen-types, built up from units of 3 amino acids of the type $(Gly–X–Y)_m$ [10]. More examples of multiple repetitions are given in [9,10].

We tested the possibility that S1 is constructed from small-sized segments by trying different short segment lengths for the intra-comparison of all 44- and 87-residue strings with RELATE. However, the maximum values of the segment comparison scores obtained were all < 3.0 (i.e., the threshold

Table 4

Comparison of the 44-residue strings of protein S1 with RELATE

	cond.	B	b	C	c	D	d	E	e	F	f	G	g	h
a	8/100	-1.2	0.9	-0.2	-0.6	0.6	0.8	0.6	-0.7	0.8	0.4	-0.1	-0.9	1.2
B	20/100		0.6	0.8	0.8	0.3	2.1	0.2	4.9	0.4	2.0	1.1	-0.2	
	30/100		0.3	1.6	0.8	0.6	3.7	1.3	6.8	1.0	3.2	2.2	0.4	0.8
b	8/100			0.7	-0.3	-0.5	4.0	0.8	1.5	0.2	1.3	0.1	2.0	1.1
	20/100			2.2	-0.5	-0.4	2.0	0.5	2.4	-0.4	0.5	1.7	1.2	
	30/100			2.2	0.7	0.1	1.3	-0.8	2.1	-0.4	0.9	1.1	1.6	
C	8/100				-0.6	2.3	-1.3	1.6	0.7	2.3	0.4	3.3	-0.8	-0.6
	20/100				-1.0	2.7	-0.9	0.8	1.2	1.1	0.8	3.4	-0.4	
	30/100				-0.8	4.2	-0.1	1.6	1.7	1.4	3.8	-0.2		
c	8/100					-1.1	1.9	-0.0	1.1	-1.2	0.5	0.8	-0.1	0.0
	20/100					-1.6	2.9	-0.9	2.1	-2.3	2.7	0.9	2.9	
	30/100					-0.5	3.0	-0.7	1.5	-1.4	2.5	-0.1	3.9	
D	8/100						0.5	10.3	1.9	8.3	0.2	2.6	-1.0	-1.5
	12/100						0.3	11.9	2.7	9.6	0.1	3.3	-1.0	
	20/100						-0.3	11.0	0.8	9.7	0.2	2.8	-1.4	
d	8/100							1.2	9.1	1.2	7.3	-0.7	1.5	-1.5
	20/100							1.9	13.3	0.2	11.3	0.5	3.8	
	30/100							0.2	13.5	0.8	11.2	-0.9	5.2	
E	12/100								1.9	12.4	0.5	5.2	0.9	-1.4
	20/100								1.5	15.3	0.7	7.1	-0.5	
	30/100								1.2	15.9	0.1	10.0	-0.6	
	40/ 50								2.9	12.7	0.6	11.7		
	G15/ 50								3.2	8.4	2.1	6.7		
e	20/100									0.7	10.7	2.8	4.7	-0.5
	30/100									0.2	10.6	-0.8	5.7	
	G20/ 50									1.9	10.3	2.5		
F	20/100										-0.3	6.3	0.2	-2.0
	30/100										-0.4	8.7	0.4	
	40/ 50											8.9		
f	8/100											1.2	2.0	0.3
	20/100											1.4	2.6	
	30/100											-0.3	4.6	
	G20/ 50												2.5	
G	20/100												2.0	-1.9
g	8/100													0.9

Scoring matrices, mutation data matrix and genetic code matrix; fragment lengths, 8, 20, 30 and 40; 50–100 random runs as indicated. Listed are the segment comparison scores (in SD units) obtained with the mutation data matrix. With the genetic code matrix the values were lower and are only listed in cases where they are higher or similar to the other ones (labelled with G)

value for significant homology) and < 2.0 SD units, the limit for distantly related sequences. Therefore, no significance for 5-, 6- or 7-residue repeats was found within S1. However, some similar dipeptide-, tripeptide- or tetrapeptide sequences within the same 44-residue string can be found, such as the homologous sequence Val–Asp–Gly–Leu and Val–Lys–Gly–Ile within string D, which might be an indication of distantly related stretches within the 44-residue strings, condensed from 22 units or from 5–7-residue segments. With SEEK, a further attempt was made to detect such small-sized repeats (e.g., fig.5).

```
EC S1    274-293    YPEGTKLTGRVTNLTDYGCF
         285-289         TNLTD
                         *+***+
         278-282               TKLTG
                               *+***+
```

Fig.5. Observation of short homologous peptides within a 20-residue segment of string E. This and other short similar sequences within one string indicate that the 44-residue segments in turn are composed of 20–22 residue segments which may be repeats of 5–7 residue units.

These examples suggest that the existence of small repeats cannot be ruled out completely. But significant proof of their occurrence cannot be given due to the fact that they are not related enough to allow a direct statistical approach.

3.5. *Evolution of the S1 gene*

A statistical analysis in partial agreement with our repetitious sequence results within protein S1 appeared in [21]. However, the conclusions drawn from that analysis for the segmental evolution of the S1 gene as given in a schematic depiction of unequal crossover events simplifies the results. The differences in the relatedness of the 44- and 87-residue strings as obtained in the statistical approach show a complicated divergence. A successive figure of the pairwise similarities of the segments is difficult to reach as slight changes made to the parameters of the statistical analysis (e.g., to the selected fragment length in RELATE or to the number of allowed deletions in ALIGN) cause changes in the sequence of relative similarities for rather related pairs of strings, such as Dd, Ee and Ff. Therefore, any direct conclusion of the evolutionary progression of such a gene and whether it has expanded in a continuous or discontinuous mode might over-interpret the results obtainable from the statistical approach.

3.6. *Comparison of protein S1 with all other ribosomal proteins of* E. coli

The computer programs were used to compare S1 with all other proteins derived from the *E. coli* ribosome and with ribosome-associated proteins (sequence references in [11]). The comparison showed no or very distant relationships between S1 and the other proteins. The segment comparison scores (RELATE: fragment length 25; 50 random runs) were below 1.0 SD units for most of the proteins, 1.5–2.0 for the comparison of S1 with proteins S2, S19, L21, and L23, and 2.0–2.8 for proteins L6, L7, L19 and L25. Under the same conditions EC-IF-1, EC-IF-3 and EC-Tu gave scores of 4.07, 1.77 and 1.92 SD units.

To test whether or not these latter proteins have homologous stretches in common with S1, we compared different areas of the S1 sequence [e.g., the N-terminal portion (pos. 1–193) and the central region (pos. 224–354)] with these proteins or parts of them. Most of the segment comparison

```
S1      d     HPSE-IVNVGDEIT-VKVLKFDRERTRVSLGLKQLGEDPWVAIAKR
S1      e     HPSK-VVNVGDVVE-VMVLDIDEERRRISLGLKQCKANPWQQFAET
S1      f     -AVR-EYKKGDEIAAV-VLQVDAERERISLGVKQLAEDPFNNWVAL

L1    102     QIKKGEM-NFDVVIASPDAMRVVGQLGQVLGPRGLMPNPKVGTVTP
L2     77      VERLEYDPNRSANIALVLYKDGER-RYILAPKGLKAGDQIQSGVD
L18    74     VGK-AVAERALEKGIKDV-SFDRSGFQYHGRVQALADAAR
L21    59     IKAE-VVAHGRGEKVKIV-KFRRRK---HY-RKQQGHRQW
EC-IF1 41     KNYIRIL-TGDKVT-VELTPYDLSKGRIVFRSR


S1      D     -LQEGMEVKGIVKNLTDYGAFVDL-GGVD-GLLHITDMAWKRVK
S1      E     YP-EGTKLTGRVTNLTDYGCFVEIEEGVE-GLVHVSEMDWTNKNI
S1      F     H-NKGDRVEGKIKSITDFGIFIGLDGGID-GLVHLSDISWNVAGEE

S4    117      SHKAIMVNGRVVNIASYQVDPNSVVIRE-KAKKESRVKAALELAE
S13    65     EGDLRREISMSIKRLMDLGCYRGLRHRR--GLPVRGQRTKTNARTR
NS1    31     SVTE-SLKEGDDVALVGFGTFAVKERAARTGRNPQTGKEITIAAA
G5P    14     TTRSGVSRQGKPYSLNEQLCYVDLGNEYPV-LVKITLDEGQPAYAP
```

Fig.6. Homologous stretches of the strongest 44-residue repeats, d, e, f and D, E, F of protein S1 with some parts of other ribosomal proteins and related sequences as indicated by ALIGN.

scores obtained (RELATE: fragment length 12–20; 50 random runs) were < 1.5 SD units. Some proteins gave scores of 2.0–3.0, suggesting distant relationships. Proteins S3, S4, S5, S7, S12, S20, L6, L21, L23, L28, L30 and L34 showed similarities with the N-terminal part of S1 (segment comparison scores 1.0–3.0; underlined proteins 2.0–3.0). Similarly, proteins S4, S9, S13, S14, S21, L2, L6, L18, L21, L27, L30 and L34 had similar sequence stretches with the central region of S1 (underlined proteins as above). In fig.6 examples of the homologous stretches found are listed. A fairly marked degree of conservation was noticed at the site of protein S1 where the two cysteine residues are located.

With SEEK some short segments of almost all proteins were found similar to the N-terminal area of S1 (pos. 1–193) having $\geqslant 30\%$ identities at identical positions and most of the other residues being related. But since these similar segments are rather short their significance is low. Some proteins gave repeated homologous stretches in common with the middle part of S1 (e.g., L3, L9 and L12) or had a sequence similar to 2–3 of the strings Dd, Ee, Ff and Gg, such as L9 (2–18), L14 (48–68), L19 (19–39), L25 (60–81), S13 (53–78) and S14 (7–23).

Most of the observed homologous segments of these proteins match parts of one of the strings Dd, Ee, and Ff with the following sequences:

Dd (238–258): G–D–E–I–T–V–K–V–L–K–F–D–R –E–R–T–R–V–S–L–G;

Ee (316–344): I–H–P–S–K–V–V–N–V–G–D–V–V –E–V–M–V–L–D–I–D–E–E–R–R– –R–I–S–L;

Ff (405–432): A–V–R–E–Y–K–K–G–D–E–I–A–A –V–V–L–Q–V–D–A–E–R–E–R–I– –S–L–G;

which are rather monotonous sequence regions containing mainly hydrophobic and charged residues, such as valine (V), lysine (K), arginine (R), glutamic acid (E) and aspartic acid (D). The highest number of identities (with parts of string Ff) had proteins L3 and L14 with the sequences:

L3 (172–192): V–Q–S–L–D–V–V–R–V–D–A–E–R –N–L–L–L–V–K–G–A;

L14 (48–68): P–R–G–K–V–K–K–G–D–V–L–K–A –V–V–V–R–T–K–K–G.

The results obtained with the statistical approach are in accordance with a direct sequence comparison of S1 with the other proteins [6] or with immuno-chemical studies [22]; i.e., the entire S1 sequence is not significantly homologous with the total sequences of the other ribosomal proteins. The longest identical peptide was the pentapeptide in common with L3 above. No identical hexapeptide has been observed for the comparison of S1 with these sequences. Yet, faint similarities exist with some sequence segments of the other proteins, suggesting a common evolutionary source for some sequence regions.

3.7. *Comparison of protein S1 with proteins from the literature*

Protein S1 was compared with RELATE to sequences of 1100 proteins listed in [9]. No comparison gave segment comparison scores > 4.5 SD units (MDM matrix: fragment length 25; 50 random runs); 26 proteins gave 3.0–4.0 SD units; about 50 sequences gave 2.0–3.0 SD units.

Among the proteins which gave (weak) indications of relatedness were: some enzymes (subtilisin, asparaginase, carboxypeptidase and pepsinogen); several peptide hormones; myoglobins; virus and phage coat proteins. Keratins, fibrinogen A-type peptides and lactalbumin gave ~2.5 SD units. The comparison of these proteins with parts of S1 showed, in most cases, relatively more similarity to strings Dd and/or Ee than to others. However, the similarities again are restricted to small areas (except for pepsinogen) with rare amino acids at similar positions. Flagellin, gastric inhibitory peptide and testis-specific basic protein had similarities to the C-terminal part of S1. No immunoglobulins showed relatedness nor in most cases did the histones.

Interestingly, the proteins known to have internal repeats, such as the Ca^{2+}-binding muscle proteins (details in [9]), showed some relationship to protein S1, mainly with its region of strongest repeats. This suggests that protein S1 is similar to proteins that cause conformational changes. S1 might have a common ancestor with the contractile proteins. Among the ribosomal proteins it is exceptional in binding the messenger. Therefore, it may well be that S1 acts within the ribosome as a contractile protein; e.g., by attaching the mRNA to the ribosome [23] and/or by facilitating or directly inducing the movement of the messenger. It would be interesting to investigate some of the strongest repeats of the middle region of S1 which have been shown to contain the mRNA-binding domain [1], in a contractile system.

REFERENCES

[1] Suryanarayana, T. and Subramanian, A.R. (1979) J. Mol. Biol. 127, 41–54.
[2] Laughrea, M. and Moore, P.B. (1977) J. Mol. Biol. 112, 399–421.
[3] Labischinski, H. and Subramanian, A.R. (1979) Eur. J. Biochem. 95, 359–366.
[4] Subramanian, A.R. and Mizushima, S. (1979) J. Biol. Chem. 254, 4309–4312.
[5] Schnier, J., Kimura, M., Foulaki, K., Subramanian, A.R., Isono, K. and Wittmann-Liebold, B. (1982) Proc. Natl. Acad. Sci. USA 79, 1008–1011.
[6] Kimura, M., Foulaki, K., Subramanian, A.R. and Wittmann-Liebold, B. (1982) Eur. J. Biochem. 123, 37–53.
[7] George, D.G., Orcutt, B.C., Dayhoff, M.O. and Barker, W.C. (1982) National Biomedical Research Foundation, Georgetown University Medical Center, Washington DC.
[8] Orcutt, B.C., Dayhoff, M.O. and Barker, W.C. (1982) National Biomedical Research Foundation, Georgetown University Medical Center, Washington DC.

[9] Dayhoff, M.O. (1978) in: Atlas of Protein Sequence and Structure, vol.5, suppl.3, National Biomedical Research Foundation, Washington DC.
[10] Barker, W.C., Ketcham, L.K. and Dayhoff, M.O. (1978) J. Mol. Evol. 10, 265–281.
[11] Wittmann-Liebold, B. (1981) in: Chemical Synthesis and Sequencing of Peptides and Proteins (Liu, T.H. et al. eds) pp.75–110, Elsevier Biomedical, Amsterdam, New York.
[12] Wittmann-Liebold, B. and Bosserhoff, A. (1981) FEBS Lett. 129, 10–16.
[13] Kimura, M., Mende, L. and Wittmann-Liebold, B. (1982) FEBS Lett. 149, 304–312.
[14] Kamp, R.M. and Wittmann-Liebold, B. (1982) FEBS Lett. 149, 313–319.
[15] Rombauts, W., Feytons, V. and Wittmann-Liebold, B. (1982) FEBS Lett. 149, 320–327.
[16] Yaguchi, M., Roy, C., Reithmeier, R.A.F., Wittmann-Liebold, B. and Wittmann, H.G. (1983) FEBS Lett. 154, 21–30.
[17] Mende, L., Timm, B. and Subramanian, A.R. (1978) FEBS Lett. 96, 395–398.
[18] Pon, C.L., Wittmann-Liebold, B. and Gualerzi, C. (1979) FEBS Lett. 101, 157–160.
[19] Brauer, D. and Wittmann-Liebold, B. (1977) FEBS Lett. 79, 269–275.
[20] Arai, K., Clark, B.F.C., Duffy, L., Jones, M.D., Kaziro, Y., Laursen, R.A., L'Italien, J.L., Miller, D.L., Nagarkatti, S., Nakamura, S., Nielsen, K.M., Petersen, T.E., Takahashi, K. and Wade, M. (1980) Proc. Natl. Acad. Sci. USA 77, 1326–1330.
[21] Doolittle, R.F., Woodbury, N.W. and Jue, R.A. (1982) Biosc. Rep. 2, 405–412.
[22] Stöffler, G. (1974) in: Ribosomes (Nomura, M. et al. eds) pp.615–667, Cold Spring Harbor Lab. Press, Long Island NY.
[23] Subramanian, A.R. (1983) Progr. Biophys. Mol. Biol., in press.