Volume 14 Number 16 1986 Nucleic Acids Research

Sigma factors from *E. coli, B. subtilis*, phage SP01, and phage T4 are homologous proteins

Michael Gribskov*+ and Richard R.Burgess

McArdle Laboratory for Cancer Research, University of Wisconsin, Madison, WI 53706, USA

Received 6 March 1986; Revised and Accepted 23 July 1986

ABSTRACT

We show, using dot matrix comparisons and statistical analysis of sequence alignments, that seven sequenced sigma factors, E. coli sigma-70 and sigma-32, B. subtilis sigma-43 and sigma-29, phage SP01 gene products 28 and 34, and phage T4 gene product 55, comprise a homologous family of proteins. Sigma-70, sigma-32, and sigma-43 each have two copies of a sequence similar to the helix-turn-helix DNA binding motif seen in CRP, and lambda repressor and cro proteins. B. subtilis sigma-29, SP01 gp28, and SP01 gp34 have at least one copy similar to this sequence. We propose that a second sequence, conserved in all seven proteins is the core RNA polymerase binding site. A third region, present only in sigma-70 and sigma-43, may also be involved in interaction with core. Available mutational evidence supports our model for sigma factor structure.

INTRODUCTION

The sigma subunits of bacterial RNA polymerases determine the specificity of the enzyme for promoters. Originally isolated as a factor that stimulated in vitro transcription by Escherichia coli RNA polymerase (1), additional sigma factors important in sporulation (reviewed in 2) and phage development in Bacillus subtilis, and heat shock and phage development in E. coli have been isolated. Six of these sigma factors, the E. coli rpoD and rpoH (htpR) gene products (sigma-70 and sigma-32), T4 gene product 55, the B. subtilis rpoD gene product (sigma-43), and the B. subtilis phage SP01 gene 28 and gene 34 products (gp28 and gp34), are known to bind to RNA polymerase and stimulate transcription with the appropriate specificity in vitro (3,4,5,6). The B. subtilis spoIIG gene product (sigma-29, (7)), while not proven to function as a sigma factor, is included because of its strong sequence similarity to the other sigma factors (8,9).

The nucleotide sequences of seven sigma factors have been determined (8,10,11,12,13,14,15,16). Strong similarities between the sequences of E. coli sigma-70 and B. subtilis sigma-43 (16), E. coli sigma-70 and sigma-32 (12,13), and E. coli sigma-70, B. subtilis sigma-43, and B. subtilis sigma-29

(8,9) have been reported previously. Similarities of sigma-43 and sigma-32 to several DNA binding proteins (12,16), and sigma-32 to CRP (13) have been noted. These similarities, however, were not shown to be statistically significant, and therefore their importance in sigma function was unclear. The aims of this study were twofold. We desired to determine if the apparent homology of sigma factors with each other and with DNA binding proteins could be shown to be real (i.e. statistically more likely than chance). Our second goal was to see whether statistical methods could reveal sequence similarities between functionally similar but (until now) apparently unrelated sigma factors from bacteria and bacteriophage.

Simultaneous analysis of the seven sigma factors reveals additional relationships and strengthens earlier findings. We show that sigma-70, sigma-43, and sigma-32 each have two regions that are significantly similar to proposed DNA binding regions, and that sigma-29, SP01 gp28 and SP01 gp34 have at least one of these regions. This strongly argues for direct interaction with DNA as the basic mechanism of sigma function.

Previous analyses of SP01 gene products 28 and 34 concluded that they were not related to E. coli sigma-70 (11,14), we have found them to be significantly similar to several of the sigma proteins and thus part of the homologous family of sigma factors. T4 gene product 55 (gp55) was also previously thought to be unrelated to other sigma factors (15) but now can be shown to be homologous.

RESULTS

Dot matrix analysis of sigma-70 and sigma-43 (Fig. 1A) reveals two regions of highly similar amino acid sequence as previously noted by Gitt et al. (16). The larger region corresponds to the carboxyl half of sigma-70 and the smaller one to the amino terminal 21%. One of the advantages of dot matrix methods over alignment methods is their ability to detect duplications. Duplications are suggested by a line of dots running parallel to the main diagonal. Every line of dots, however, does not indicate a significant similarity, and each possible duplication must be investigated individually to determine its importance. Arrow 1 in Fig. 1A indicates a small duplication at the amino terminus in which a short sequence present once in sigma-70 is found twice in sigma-43. Taking this duplication into account allows a better alignment of the amino terminal region than was recently reported (16). A second duplication can be seen near the carboxyl end (arrow 2) of sigma-70 and sigma-43 where a sequence element appears to be

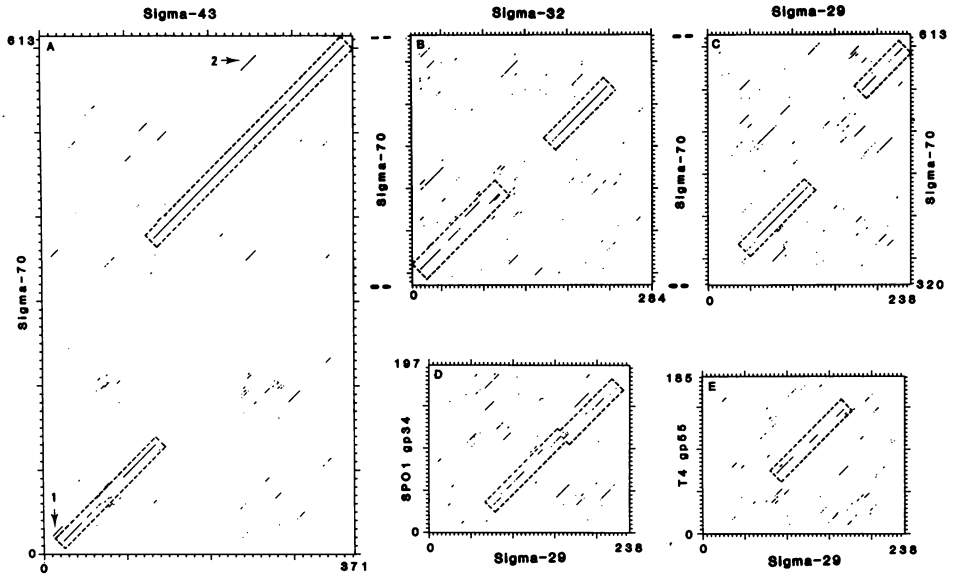

Figure 1. Dot Matrix Comparisons of Sigma Factors.
(A) Vertical axis: sigma-70 (residues 1-613), horizontal axis: sigma-43 (residues 1-371), stringency = 9.5. (B) Vertical axis: sigma-70 (residues 320-613), horizontal axis: sigma-32 (residues 1-284), stringency = 7.5. (C) Vertical axis: sigma-70 (residues 320-613), horizontal axis: sigma-29 (residues 1-238), stringency = 7.5. (D) Vertical axis: SP01 gp34 (residues 1-197), horizontal axis: sigma-29 (residues 1-238), stringency = 7.0. (E) Vertical axis: T4 gp55 (residues 1-185), horizontal axis: sigma-29 (residues 1-238), stringency = 6.8. A window of 20 was used for all comparisons. Boxes enclose the diagonals that can be shown to be homologous by alignment methods. Numbered arrows indicate additional regions of homology mentioned in text.

present in two copies in both proteins. A similar effect is seen in a comparison of sigma-43 and sigma-32 (data not shown). The significance of this duplication is discussed below (region 3 and 4).

Comparison of sigma-70 and sigma-32 shows a pattern of regions of high similarity separated by regions of lower similarity (Fig. 1B). A region near the amino terminus of sigma-32 and another near the carboxyl terminus are the most closely related to sequences present in sigma-70. As previously noted (12,13), sigma-32 is related only to the carboxyl terminal half of sigma-70, and shows no similarity to the amino terminal end of sigma-70 (but see also Other Proposed Homologous Regions).

Sigma-29, SP01 gp34 and T4 gp55 are smaller than sigma-70, sigma-43, and sigma-32. As can be seen in Fig. 1C, most of this size difference is

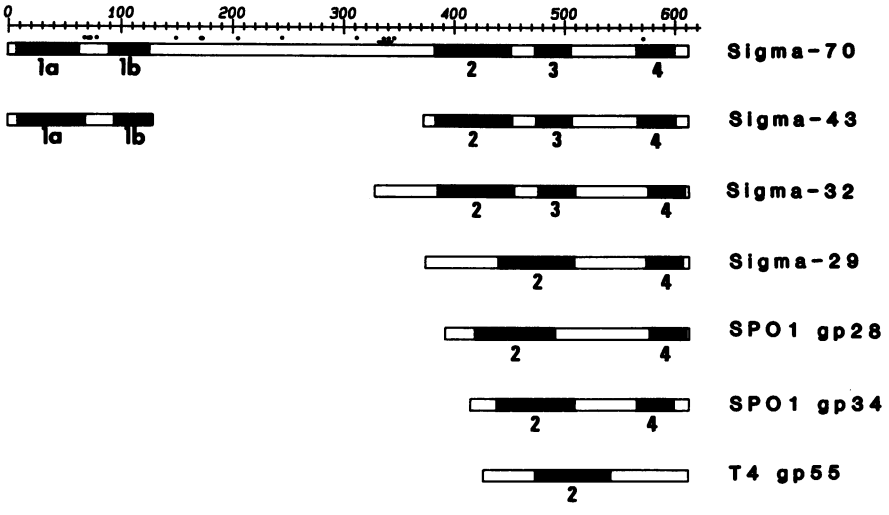

Figure 2. Location of Highly Conserved Regions.
The shaded section of the boxes indicate the locations of the highly conserved regions discussed in the text. The locations of amino acid differences between E. coli and S. typhimurium sigma-70 are indicated by dots above the sigma-70 bar. The location of the rpoD800 mutation is shown as a short line above the sigma-70 bar.

accounted for by a large block of sequence present in sigma-70, but not in sigma-29. SP01 gp34 and T4 gp55 are weakly related to the other sigma factors, their strongest similarity being to sigma-29. Their essential colinearity with sigma-29 can be seen in Fig. 1D and 1E. This indicates that they also lack the same block of sequence missing in sigma-29 as compared to sigma-70. Note that the boxes indicating the main regions of homology have been determined after further analyses and can not be derived from the dot matrix comparison alone.

From these dot matrix comparisons, and others not shown, a general idea of the regions conserved between the various sigma factors can be formed. Fig. 2 shows a schematic diagram of the most conserved sequence elements which, for convenience, we refer to as regions 1 to 4. Region 1 is present only in sigma-70 and sigma-43. Region 2 is found in all the proteins, region 3 in sigma-70, sigma-43 and sigma-32, and region 4 in all of the proteins except T4 gp55.

Regions 2 to 4 represent highly similar sequence elements found in three or more of the proteins. In the case of sigma-70, sigma-43, and sigma-32, the intervening regions are also clearly related (12,16). In general,

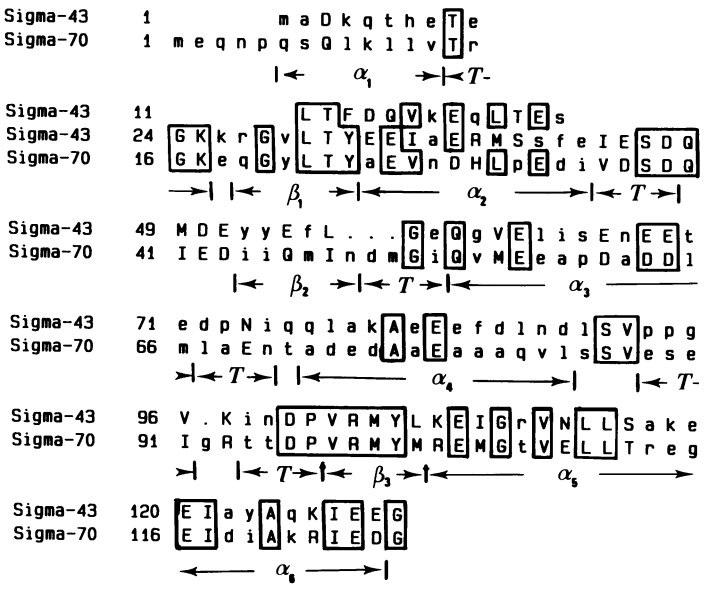

Figure 3. Alignment of Sigma-43 and Sigma-70 in Region 1. Chemically similar residues (ILMV, KHR, DENQ, ST, AG, and FWY) are shown in capitals when aligned. Identical residues are boxed when aligned. Periods indicate gaps inserted to give the best alignment. Note that sigma-43 has two sequence elements corresponding to residues 22-34 of sigma-70. Secondary structures predicted from the aligned sequences are shown below the aligned sequences (α = alpha helix, β = beta structure, T = turn).

though, the similarity of the intervening regions is less than within regions 2 to 4. Sigma-29 and SP01 gp34, and sigma-29 and T4 gp55, can also be aligned over the entire length of the sequences, although the intervening regions in these alignments are much more weakly conserved than regions 2 to 4.

Region 1

The alignment of sigma-70 and sigma-43 recently reported by Gitt et al. (16) juxtaposed 62 (24.4%) identical residues and 36 (14.2%) chemically similar residues. The alignment shown here (Fig. 3) differs from theirs in the first three quarters of region 1, and aligns 31.3% of the residues in this region with identical amino acids and 21.5% with chemically similar amino acids. There are two strongly related segments in region 1, corresponding to residues 14-64 (1A) and 86-116 (1B) of sigma-70 (Fig. 3). More than 60% of the amino acids in these segments are identical or chemically similar in sigma-70 and sigma-43, a level that approaches the 84%

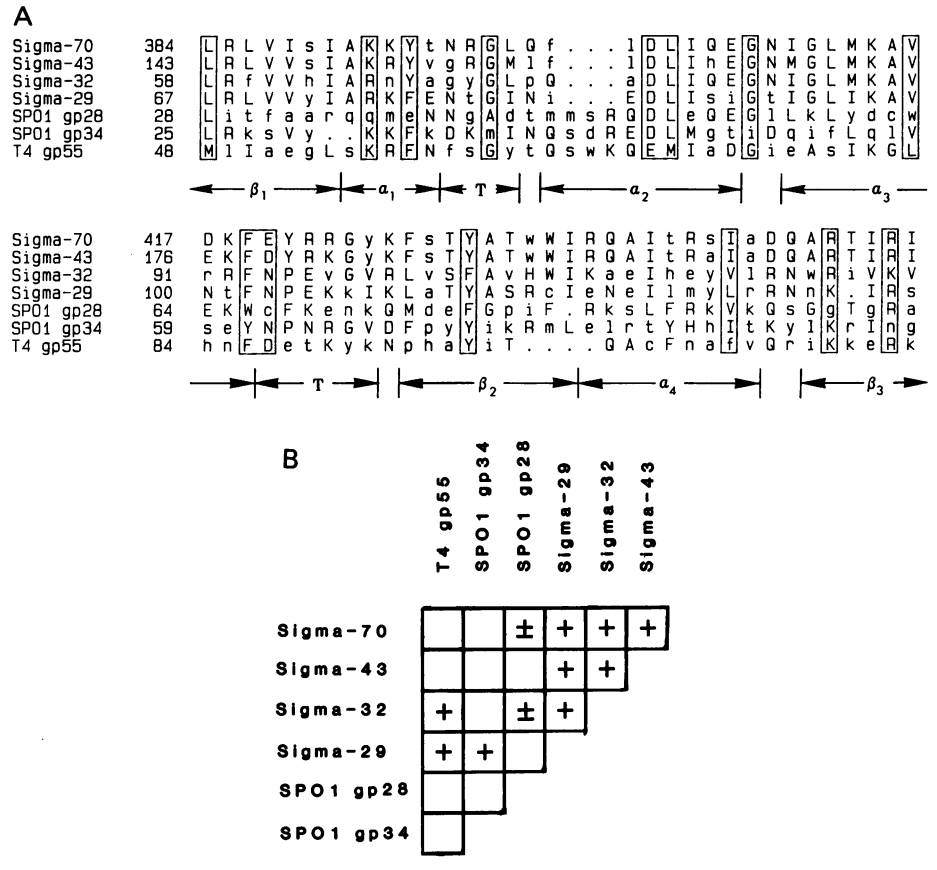

Figure 4. Alignment of Sigma Factor Sequences in Region 2.
(A) Chemically similar residues present in three or more of the sequences are shown in capitals. Boxes indicate positions where at least six of the sequences have chemically similar residues. Secondary structure predicted from the aligned sequences are shown below the aligned sequences (symbols as in Fig. 3).
(B) Significance of alignments shown in Fig. 4A, (+) denotes the indicated alignment has an adjusted score of 3.0 or greater, (+/-) indicates an adjusted score of 2.0 to 3.0.

seen in the alignment of the carboxyl terminal ends of sigma-70 and sigma-43. Structural predictions of the aligned sequences in region 1A show several potential alpha and beta structures, while we predict that region 1B forms several helices separated by reverse turns.

Region 2

Similarities in region 2 have previously been noted between sigma-70 and

Table 1
Alignment Statistics for region 2

Sequence 1	From	To	Sequence 2	From	To	Score	Mean	S.D.	Adj. Score
Sigma-70	364	456	Sigma-43	123	215	105.8	25.2	2.1	38.2
Sigma-70	364	456	Sigma-32	38	130	52.6	21.4	2.1	15.2
Sigma-70	364	456	Sigma-29	46	139	52.0	20.9	2.3	13.4
Sigma-70	364	456	SP01 gp28	6	103	25.0	19.1	2.1	2.7
Sigma-43	123	215	Sigma-32	38	130	53.7	18.8	2.1	16.6
Sigma-43	123	215	Sigma-29	46	139	51.3	19.8	2.1	15.0
Sigma-32	38	130	Sigma-29	46	139	58.1	20.6	2.4	15.8
Sigma-32	38	130	SP01 gp28	6	103	21.6	17.1	2.2	2.1
Sigma-32	38	130	T4 gp55	28	123	24.9	18.8	2.0	3.0
Sigma-29	46	139	SP01 gp34	6	98	33.0	18.9	2.2	6.4
Sigma-29	46	139	T4 gp55	28	123	27.9	19.8	2.3	3.5

Statistics for the alignments shown in Fig. 4A. Score is the alignment score for the alignments shown in Fig. 4A. Mean and S.D. are the average score and standard deviation for one hundred alignments of random sequences of the same length and composition. The adjusted score is calculated from the relation: Adj. Score. = (Score - Mean)/S.D.

sigma-32 (12,13), sigma-70 and sigma-29 (8,9), and sigma-70 and sigma-43 (16). Fig. 4 shows the alignment of the seven sigma factors in this region. The sequence similarity in region 2 is striking, especially in the region corresponding to residues 403 to 421 of sigma-70. Not all of these sequences can be shown to be significantly similar in pairwise comparisons, but each protein can be clearly shown to be related to at least one of the others (Table 1). Sigma-29 is particularly important as it is significantly related to both bacterial and phage encoded sigma factors, and thus provides the bridge between these two groups. Residues corresponding to positions 384, 392, 398, 403 - 405, 408, 415, 419, and 430 of sigma-70 appear to be particularly strongly conserved. These positions include several where aromatic residues are seen in all seven proteins.

Secondary structure predictions for the aligned sequences suggest there are two beta structures (residues 384 - 391 and 427 - 435 of sigma-70) separated by a long region of alpha helix. A third region of beta structure is predicted at the C-terminal end of region 2 (residues 449 to 456 of sigma-70). These three beta strands are of nearly the same length suggesting that they form a beta sheet structure. If they are arranged in the order beta-1, beta-3, beta-2, the long alpha helices at residues 397 to 418 of sigma-70 are the correct length to connect beta-1 and beta-2. In agreement with this model, these helices would have a hydrophobic side suitable for packing against a beta sheet. This model suggests that the most remarkably

Table 2
Alignment Statistics for Sigma Factors in Regions 3 and 4

Sequence 1	From	To	Sequence 2	From	To	Score	Mean	S.D.	Score
Sigma-70 (3)	465	516	Sigma-43 (3)	224	275	57.2	13.1	1.9	23.6
Sigma-70 (3)	465	516	Sigma-43 (4)	316	369	16.0	12.7	1.5	2.2
Sigma-70 (3)	465	516	Sigma-32 (3)	139	191	18.1	11.5	1.5	4.4
Sigma-70 (3)	465	516	Sigma-32 (4)	237	284	18.8	11.6	1.9	3.8
Sigma-70 (3)	465	516	SP01 gp34	141	193	17.0	11.5	1.5	3.6
Sigma-70 (4)	557	610	Sigma-43 (3)	224	275	20.1	13.3	1.8	3.9
Sigma-70 (4)	557	610	Sigma-43 (4)	316	369	61.4	13.6	1.9	24.6
Sigma-70 (4)	557	610	Sigma-32 (3)	139	191	15.8	11.9	1.8	2.2
Sigma-70 (4)	557	610	Sigma-32 (4)	237	284	29.8	11.4	1.8	10.4
Sigma-70 (4)	557	610	Sigma-29	190	238	20.6	11.0	1.4	7.0
Sigma-43 (3)	224	275	Sigma-32 (3)	139	191	21.7	12.0	2.0	4.9
Sigma-43 (3)	224	275	Sigma-32 (4)	237	284	19.4	11.2	1.6	5.2
Sigma-43 (3)	224	275	Sigma-29	190	238	14.2	10.9	1.6	2.1
Sigma-43 (3)	224	275	SP01 gp34	141	193	16.5	12.6	1.8	2.2
Sigma-43 (4)	316	369	Sigma-32 (3)	139	191	16.3	11.9	1.8	2.3
Sigma-43 (4)	316	369	Sigma-32 (4)	237	284	32.5	11.2	1.6	13.6
Sigma-43 (4)	316	369	Sigma-29	190	238	24.4	10.8	1.6	8.5
Sigma-32 (4)	237	284	Sigma-29	190	238	22.1	11.5	1.9	5.6
Sigma-32 (4)	237	284	SP01 gp34	141	193	16.6	10.1	1.6	4.1
Sigma-29	190	238	SP01 gp34	141	193	17.3	12.4	2.0	2.5
Sigma-29	190	238	SP01 gp28	119	170	16.8	10.6	1.6	3.9

Statistics for the alignments shown in Fig. 5A. Score is the alignment score for the alignments shown in Fig. 5A. Mean and S.D. are the average score and standard deviation for one hundred alignments of random sequences of the same length and composition. The adjusted score is calculated as in Table 1. Numbers in parentheses indicate region 3 or region 4.

conserved part of this region (residues 403 - 431 of sigma-70) would lie on the protein surface. We speculate that it would be available for interaction with core RNA polymerase.

Regions 3 and 4

Earlier analyses have noted conserved sequence elements in region 3 (12) and region 4 (12,13,16). Region 3 is clearly conserved between sigma-70, sigma-43, and sigma-32, as is region 4. Furthermore, many of the region 3 sequence elements can be shown to be related to region 4 elements (Table 2, Fig. 5). This indicates that region 3 and 4 are two repeats of a single homologous sequence. Sigma-29, SP01 gp28, and SP01 gp34 may have only one sequence element corresponding to this sequence, since only one element from each can be shown to be significantly related to either region 3 or region 4. It is, however, possible that these proteins have two related elements that are structurally similar, but not detectably similar by sequence analysis.

The most striking aspect of the structure of regions 3 and 4 is their strong similarity to the helix-turn-helix (HTH) DNA binding motif of lambda repressor (17), lambda cro (18), and E. coli CRP (19). The resemblance of

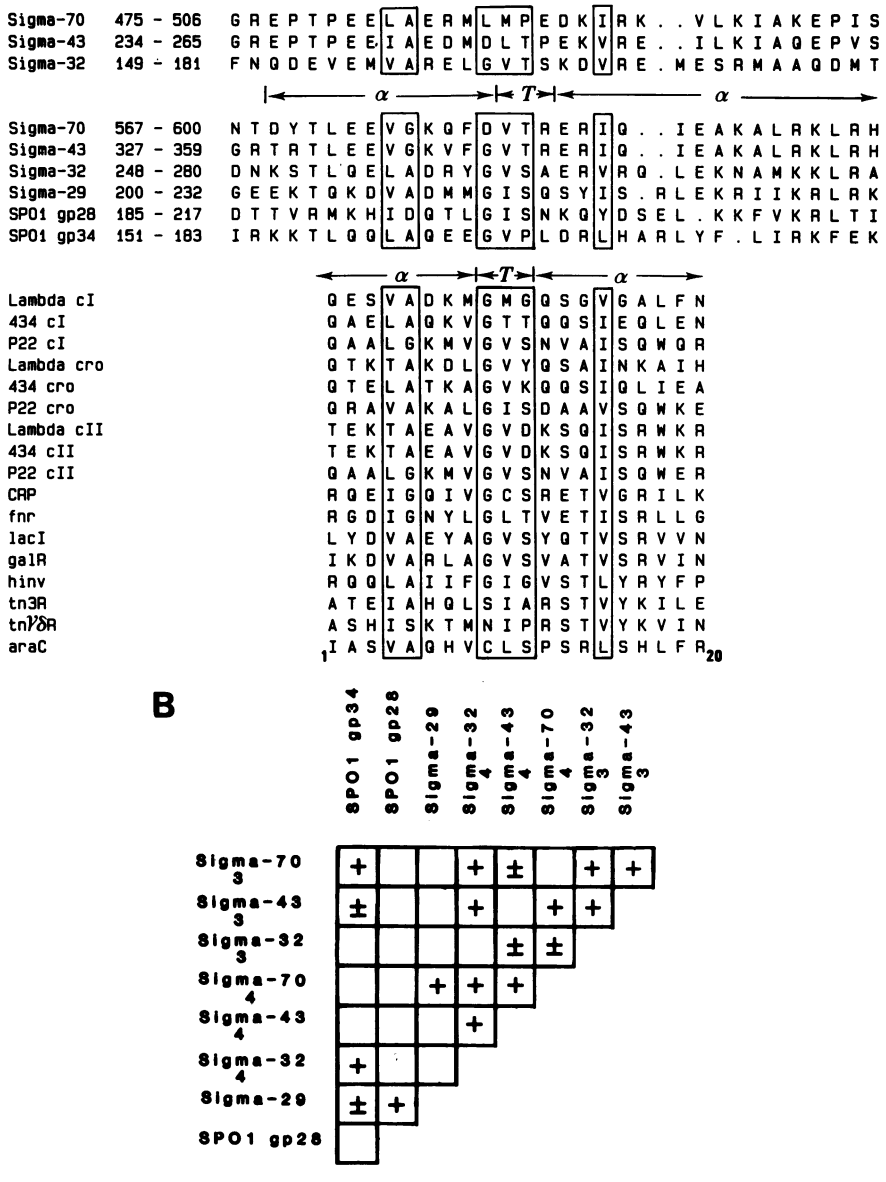

Figure 5. Alignment of Sigma Factor Sequences in Regions 3 and 4. (A) Regions 3 and 4 of the sigma factors are shown aligned above sequences thought to contain the helix-turn-helix DNA binding motif. Secondary structures predicted from the aligned sigma factor sequences are shown between the aligned region 3 and region 4 sequences. The structures predicted for the aligned HTH sequences are shown above the HTH sequences (symbols as in Fig. 3A). (B) Significance of sigma factor alignments shown in Fig. 5A (+) denotes the indicated alignment has an adjusted score of 3.0 or greater, (+/-) indicates an adjusted score of 2.0 to 3.0.

Table 3
Alignment Statistics for Sigma Factors Regions 3 and 4 and HTH elements

Sequence 1	From	To	Sequence 2	From	To	Score	Mean	S.D.	Adj. Score
Sigma-29	190	238	galR	1	34	16.4	7.8	1.2	7.1
Sigma-29	190	238	lambda cro	6	45	16.0	9.1	1.5	4.6
Sigma-43 (4)	316	369	galR	1	34	13.8	8.4	1.2	4.5
Sigma-43 (3)	224	275	lexA	18	57	16.7	10.1	1.6	4.2
Sigma-29	190	238	lambda cII	14	53	17.6	10.8	1.7	4.1
Sigma-29	190	238	lacI	1	35	13.0	7.6	1.3	4.1
Sigma-32 (4)	237	284	434 cro	9	48	16.8	11.0	1.5	4.0
Sigma-32 (4)	237	284	lambda cI	23	62	14.2	8.8	1.5	3.8
Sigma-43 (4)	316	369	CRP	159	198	14.5	9.4	1.6	3.3
Sigma-43 (3)	224	275	CRP	159	198	14.4	9.4	1.6	3.3
Sigma-43 (3)	224	275	tn3R	151	185	14.2	9.6	1.4	3.3
Sigma-32 (3)	139	191	lambda cI	23	62	13.6	8.8	1.4	3.3
Sigma-29	190	238	hinv	157	190	12.3	8.0	1.4	3.2
Sigma-32 (4)	237	284	galR	1	34	14.7	10.2	1.5	2.9
Sigma-70 (3)	465	516	lexA	18	57	14.6	10.3	1.5	2.9
Sigma-43 (4)	316	369	lambda cro	6	45	14.3	9.9	1.6	2.8
Sigma-43 (4)	316	369	434 cro	9	48	15.2	10.9	1.5	2.8
Sigma-32 (4)	237	284	CRP	159	198	13.0	9.3	1.3	2.8
Sigma-70 (4)	557	610	CRP	159	198	12.9	9.2	1.4	2.7
Sigma-29	190	238	lambda cI	23	62	16.4	11.6	1.8	2.7
Sigma-70 (3)	465	516	tn3R	151	185	13.5	10.0	1.4	2.6
Sigma-70 (3)	465	516	CRP	159	198	12.8	9.2	1.4	2.6
Sigma-70 (4)	557	610	lambda cro	6	45	13.7	10.0	1.4	2.6
Sigma-70 (4)	557	610	galR	1	34	11.4	8.7	1.1	2.5
Sigma-32 (4)	237	284	lacI	1	35	12.4	8.8	1.5	2.5
Sigma-43 (3)	224	275	lambda cI	23	62	15.7	11.9	1.7	2.2
SP01 gp28	174	220	lexA	18	57	12.2	9.1	1.4	2.2
SP01 gp34	141	193	CRP	159	198	11.7	8.9	1.3	2.2
Sigma-43 (3)	224	275	lambda cII	14	53	13.4	10.5	1.4	2.0
Sigma-32 (3)	139	191	fnr	197	216	11.2	8.5	1.3	2.0

Numbers in parentheses indicate region 3 or region 4. Sequences are referred to by their genetic loci except for tn3r, the transposon 3 repressor, and tn R, the transposon gamma-delta repressor. The HTH elements are those suggested in (20). The protein sequences derived from the following genes were compared to each region 3 and region 4 element; only alignments with adjusted scores of 2.0 of higher are reported. Lambda cI (44), lambda cII (45), lambda cro (46), 434 cI (47), P22 cII (48), araC (49), crp (50, 51), fnr (52), galR (53), lacI (protein sequence, (54)), lexA 55,56), tn3 repressor (57), tn gamma-delta repressor (58), and S. typhimurium hinv (59). Sequences for the HTH region of 434 cI and cII, and P22 cI and cro, were taken from (20).

region 4 of sigma-32 to the HTH region of CRP was previously reported (13). Landick et al. (12) also noted the resemblance of region 3 and 4 of sigma-32 to the A-N-N-N-G-N-N-N-N-N-V sequence conserved among DNA binding proteins. The significance of these similarities, however, was not tested. We find that several of the comparisons of region 3 and 4 with proteins thought to

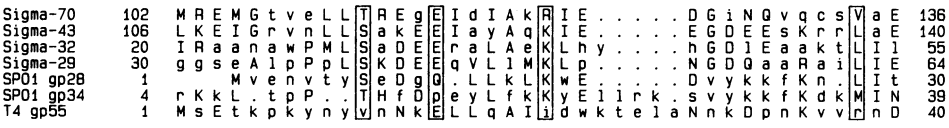

Figure 6. Alignment of Region 2 Proximal Sigma Factor Sequences. Chemically similar residues present in three or more of the sequences are shown in capitals. Boxes indicate positions where at least six of the sequences have chemically similar residues.

employ the HTH motif (20) show a significant relationship (Table 3). Although the sequences in Fig. 5 have been aligned to highlight the similarities of the sigma factors, it is obvious that both region 3 and region 4 strongly resemble the HTH consensus.

Residues thought to be important in the HTH motif are strongly conserved in regions 3 and 4. Hydrophobic residues at positions 4 and 15 of the HTH consensus are thought to be important in stabilizing the interaction of the two helices, and are present in all the sigma sequences. Positions 9 - 11 of the HTH consensus form the characteristic turn of the motif. A glycine residue at position 9 is favored due to its conformational flexibility. A glycine residue is seen in this position in most of the sigma sequences. Two of the exceptions, region 3 of sigma-43 and region 4 of sigma-70, have an aspartate residue in place of the glycine normally found. This may not be as unfavorable as it might seem since it has been shown (21) that a gly to glu mutation at this position in lambda repressor does not destroy its DNA binding activity. An alanine or glycine residue at position 5 of the HTH consensus is favored, probably because the close approach of the two alpha helices at this point leaves little space for a large side chain. Alanine or glycine residues are seen in the corresponding position for all but one of the sigma sequences shown. As might be expected, the consensus secondary structure predicted for both the aligned region 3 and 4 sequences, and the aligned DNA binding proteins, predict the same helix-turn-helix structure observed by x-ray crystallography.

Other proposed homologous regions

Stragier et al (9) have proposed that another short region of homology exists just before the element we call region 2. This short region is contiguous with region 2 in all the proteins except sigma-70, where it is interupted by a 245 residue insertion relative to the other sigma factors. These segments represent a continuation of the alignment of region 2 towards the amino terminus of the proteins. The alignment of the elements reported

Table 4
Alignment statistics for Region 2 Proximal Elements

Sequence 1	From	To	Sequence 2	From	To	Score	Mean	S.D.	Adj. Score
Sigma-70	102	141	Sigma-43	106	145	28.0	14.1	1.4	10.3
Sigma-70	102	141	Sigma-32	19	58	14.2	12.8	1.5	0.7
Sigma-70	102	141	Sigma-29	40	79	15.0	11.7	1.6	2.2
Sigma-70	102	141	SP01 gp28	5	44	13.6	13.2	1.7	0.3
Sigma-43	106	145	Sigma-32	19	58	22.6	13.1	1.4	6.5
Sigma-43	106	145	Sigma-29	40	79	21.1	13.2	2.3	3.5
Sigma-43	106	145	SP01 gp28	5	44	12.8	13.3	1.6	-
Sigma-32	19	58	Sigma-29	40	79	17.5	12.4	1.7	3.0
Sigma-32	19	58	SP01 gp28	5	44	13.1	12.6	1.5	0.3
Sigma-29	40	79	SP01 gp28	5	44	15.7	12.1	1.7	2.1

Statistics for the alignments shown in Fig. 6. Score is the alignment score for the alignments shown in Fig. 6. Mean and S.D. are the average score and standard deviation for one hundred alignments of random sequences of the same length and composition. The adjusted score is calculated as in Table 1. No adjusted score for SP01 gp34 or T4 gp55 exceeded 0.5.

by Stragier et al is shown in Fig. 6 and Table 4, along with additional segments from SP01 gp28, SP01 gp34, and T4 gp55. Although the alignments in Fig 6. generally show positive adjusted scores, most of these elements can not be shown to be significantly related because of their short length. However, if considered as amino terminal extensions of region 2, these alignments increase the significance of the alignments shown in Fig. 4 and Table 2. The strong similarity seen in this region 2 proximal region begins at a position corresponding to residue 111 of sigma-70 and therefore overlaps region 1B by 16 residues. This indicates that the large insertion in sigma-70 must have arisen after the divergence of the multiple sigma factors from a common ancestor. There does not appear to be a clear correlation between the adjusted score for the alignment and whether the sigma proteins are the major vegetative sigma factor or "minor" sigma factors as proposed by Stragier et al (9).

A possible relationship of sigma factors with the E. coli NusA protein has been postulated (13). This similarity seems to result primarily from the alignment of an HTH - like region of NusA with the HTH - like regions of the sigma factors (data not shown). The best alignment score for this HTH - like segment of NusA with any region 3 or 4 element is 3.64 (residues 372 to 406 of NusA aligned with residues 245 to 279 of sigma-32). Only one other alignment with region 3 and 4 elements gives an adjusted score greater than 3.0. It is not at all clear whether this sequence similarity represents a case of homology or convergent evolution. We therefore do not, at this time,

include NusA protein in the family of homologous sigma factors.

DISCUSSION

The demonstration that similar sequence elements are more closely related than random sequences is crucial to the establishment of homology (22). It is particularly important that the statistics of random alignments be considered when gaps are inserted in aligned sequences since the admission of gaps greatly increases the probability that unrelated sequences will yield an alignment with a high score. As explained by Doolittle (22), simple consideration of the scores of aligned sequences can easily lead to claims of homology for proteins which are no more closely related than random sequences.

Our results show that the seven proteins studied here, *E. coli* sigma-70 and sigma-32, *B. subtilis* sigma-43 and sigma-29, SP01 gp34 and gp28, and T4 gp55 are homologous sigma factors. Although each and every pair of these sequences can not be demonstrated to be related at a statistically significant level, enough of the pairwise comparisons show significant relationships to conclude that all seven protein comprise a related family. This concusion is further supported by the presence of multiple sequence elements with significant similarity in most of the sequences. Since unrelated proteins would not be expected to maintain the order of these similar sequence elements, as do the sigma factors, the overall significance of the alignments must be greater than the simple segment by segment analyses indicate. This reasoning is strongly supported by *in vitro* experiments that show that each of these proteins acts as a sigma factor (4,5,6,23).

Region 1 is apparently not essential for the stimulatory activity of sigma factors since it is present only in sigma-70 and sigma-43. The highly conserved nature of the sequence region, however, implies that it is important in the structure and/or function of the vegetative sigma factors. The most obvious chemical difference between these proteins and the rest of the sigma factors lies in their affinity for core RNA polymerase. Sigma-70 and sigma-43 are released from core during chromatography on BioRex-70 or phosphocellulose (1,3,24), the other sigma proteins are not (25). This suggests that sigma-70 and sigma-43 bind less tightly to core, or at least that the binding can be weakened by the presence of acidic groups such as those on the column matrices. The minor sigmas of *B. subtilis*, SP01 gp28 , and sigma-32 are able to bind to core and function *in vivo*, in spite of the presence of large amounts of sigma-43 or sigma-70, again suggesting a difference in their affinity for core. In the case of SP01 gp28, this is

supported by *in vitro* experiments suggesting that gp28 binds more tightly to core RNA polymerase than does sigma-43 (26). We postulate that region 1 is involved in interaction with core, and may be specifically involved in weakening the sigma - core interaction under physiological conditions where a shift in sigma factors is necessary.

Region 2 is found in all of the sigma factors. It does not have any similarity to known DNA binding sites, and we therefore believe that it is also involved in interaction with core RNA polymerase. This is strongly indicated by the remarkable resemblance between sigma-70, sigma-43, and sigma-32 in region 2, since all of them have been shown to bind and function with *E. coli* core (3,4). We speculate that this region may form a beta sheet constituting the structural center of the sigma molecule. This structure also seems reasonable for an intersubunit binding region.

Promoter sequences have two important regions, the -10 and -35 regions, important for RNA polymerase binding. Each sigma factor isolated to date seems to specify a unique sequence in both of these regions (27). The discovery that several of the sigma factors have two sequence elements related to the helix-turn-helix DNA binding motif observed in lambda repressor, lambda cro protein, and CRP is interesting as it suggests that sigma factors recognize both the -10 and -35 regions of the promoter directly, and that the recognition is mediated by distinct parts of the polypeptide. SP01 gp34 and gp28, sigma-29, and T4 gp55 do not have two *obvious* copies of a HTH - like sequence. This may be because the sequences have diverged so far as to be no longer significantly similar, because other factors are involved in promoter recognition by these proteins, or because of a real difference in the way these proteins interact with core RNA polymerase and DNA. Another possibility is, since core RNA polymerase probably supplies a substantial part ot the free energy for binding to DNA (28), that core polymerase acts to recognize part of the promoter sequence when these proteins are present. In the case of SP01 gp34, another protein, gp33, is thought to be required for promoter recognition (5). T4 gp55 also represents a special case since it recognizes a promoter with only one conserved region, roughly corresponding to the -10 region (29,30,31). RNA polymerase isolated from T4 infected cells contains several other T4 encoded proteins including gp33 and gp45 (32), although the presence of gp55 appears to be sufficient to direct late gene specific transcription (6).

The hypothesis that the conserved sequence elements identified here are the most important for sigma function is supported by available mutational

evidence. Sigma-70 of Salmonella typhimurium is very similar to E. coli sigma-70, having about 97% identical residues (33). The residues that differ between E. coli and S. typhimurium sigma-70 can be considered to be mutations not affecting sigma function. Only six of the 16 amino acid differences fall in the conserved regions shown in Fig. 2. Five of the differences are in the less conserved internal part of region 1 between regions 1A and 1B, and the sixth occurs at residue 573, just before the HTH - like element of region 4. The rpoD800 allele encodes a ts mutant of sigma-70 in which residues 330 to 343 of the wild type polypeptide are deleted. Amino acid differences between E. coli and S. typhimurium are also found in this region at position 334, 338, 340 and 345. The complete dispensability of this region suggests it lies in a linker region between the carboxyl terminal domain and a domain formed by the extra 245 residues found in sigma-70 but not other sigma factors.

The alt mutation (rpoD2) is a sigma mutation that allows transcription of the lac operon in the absence of adenyl cyclase or CRP (35). Precise mapping of this mutation (36) shows that it is an arg to his mutation at residue 596 of sigma-70. Hu and Gross (36) have also isolated additional mutations at the same position which allow growth on arabinose in cya^- strains. Some of these mutants are identical to the original alt mutant, and others are arg to ser or cys changes. The presence of these mutants near an HTH - like site suggests that they stimulate arabinose expression by improving the binding of RNA polymerase to the arabinose promoter, although it is also consistent with more efficient interaction of RNA polymerase with the araC stimulatory protein in these strains.

The pattern of conserved sequence elements among the sigma proteins suggests that they evolved from a proto - sigma factor consisting of a core binding domain and a single HTH DNA binding site. At some point the DNA binding region was duplicated giving rise to the structure seen in sigma-70, sigma-43, and sigma-32. The clear presence of two HTH - like elements in sigma-32 suggests the precursors of the minor sigma factors arose after this internal duplication in the proto - sigma. We can not say with certainty whether sigma-29, SP01 gp28, and gp34 have only one element corresponding to this region (or in the case of T4 gp55, no copies), since we have only negative evidence that this is the case; i.e. we cannot show a significant relationship between any region (other than those shown in Fig. 5) of these proteins and region 3 or region 4. If, in fact, they have only one HTH - like region, it may have arisen either by the loss of one DNA binding

site from a sigma-32 like proto - sigma, or by the divergence of sigma proteins before the internal HTH region duplication that led to sigma-70, sigma-43, and sigma-32.

Several predictions can be made based on the model we present based on sequence analysis. Mutations in region 2 might be expected to perturb the interaction of sigma with core RNA polymerase, and to be revertable by mutations in one or more of the core subunits. Mutations in regions 3 and 4 should affect primarily DNA binding, and presumably would cause a change in the promoter sequence best recognized by the holoenzyme carrying the mutant sigma factor. If, as suggested above, there are two independent DNA binding sites, it should be possible to assign the specificity for the -10 and -35 regions of the promoter to particular parts of the polypeptide.

EXPERIMENTAL PROCEDURES

The sequences of the sigma factors have been derived from the reported nucleotide sequences of their respective genes. In the case of sigma-32, two versions of the nucleotide sequence have been reported; we have adopted the version of Yura et al. (13), which differs from that of Landick et al. (12) in having thr at position 2, ala at 185, his at 193, and ala at 194, instead of ala, ser, glu, and pro, respectively. Note that the B. subtilis major vegetative sigma, previously termed sigma-55 due to its migration on SDS polyacrylamide gels, has been referred to as sigma-43 because of its molecular weight calculated from the amino acid sequence.

We have used the programs of the University of Wisconsin Genetics Computer Group (UWGCG) in our analyses (37). We have relied primarily on dot-matrix analysis (38) and the local homology alignment algorithm of Smith and Waterman (39) as implemented by UWGCG. In both cases, we have used a scoring table in which comparisons of identical residues score 1.5, and non-identical residues a score derived from the mutational difference matrix (MDM78) of Schwartz and Dayhoff (40). The scores for non-identical comparisons have been adjusted to have a mean of -0.17 and a standard deviation of 0.364. Scores for non-identical comparisons range from 1.491 for phe-tyr to -0.677 for ala-trp. A stringency of 7 with a window of 20 thus requires two sequences to contain the equivalent of 23% identical residues to produce a dot.

The dot matrix method compares two short "windows" of each sequence and puts a dot at the position of the center of each window whose score exceeds a certain score or stringency. The method is particularly useful for qualitative assessment of the relationship of two sequences because it is

less sensitive to insertions or deletions than are alignment methods.

Statistical analysis of the alignments was performed by calculating the average and standard deviation of the scores for 100 alignments of random sequences with the same composition and length as the sequences of interest. Because the proteins have regions of strong sequence similarity separated by regions with lower, or in some cases no, sequence similarity, each region of strong similarity has been evaluated independently. In Tables 1 to 4, score refers to the alignment score of the original sequences using the scoring table described above. Mean and S.D. are the mean and standard deviation for the 100 alignments of randomized sequences. The mean and score are dependent on the length of the sequence elements aligned as can be seen from a comparison of Tables 1 and 2. The adjusted score is calculated from the relation

$$\text{adj. score} = (\text{score}-\text{mean})/\text{S.D.}$$

The adjusted score is less length dependent than the original score and mean values, and can, with caution, be used to compare alignments of different lengths. An adjusted score of 3.0 or greater is required for reasonable confidence that two sequences are significantly related (21,41).

We have predicted the secondary structures of the proteins by two methods (42,43) with computer programs developed by M.G. (60) using a procedure library provided by UWGCG. Consensus structures have been predicted for the regions of strongest similarity using the Chou - Fasman method by averaging the alpha, beta, and turn potentials of the residues of the aligned sequences at each position. Complete predictions for each protein are available on request.

Aknowledgements

We thank John Devereux of the University of Wisconsin Genetics Computer Group for assistance with the sequence alignments. This work was supported by NIH grants GM 28575, CA 07175, and CA 09135.

+Present address: Molecular Biology Institute, University of California, Los Angeles, CA 90024, USA

*To whom correspondence should be addressed

REFERENCES

1. Burgess, R.R., Travers, A.A., Dunn, J.J. and Bautz, E.K.F. (1969). Nature 221, 43-46.
2. Losick, R. and Youngman, P. (1984) in Microbial Development, Losick, R. and Shapiro, L., Eds., pp. 63-68, Cold Spring Harbor Laboratory, N.Y.

3. Shorenstein, R.G. and Losick, R. (1973) J. Biol. Chem. 248, 6170-6173.
4. Grossman, A.D., Erickson, J.W. and Gross, C.A. (1984) Cell 38, 383-390.
5. Tjian, R. and Pero, J. (1976) Nature 262, 753-757. Proc. Natl. Acad. Sci. USA 82, 4189-4192.
6. Kassavetis, G.A. and Geiduschek, E.P. (1984) Proc. Natl. Acad. Sci. USA 81, 5101-5105.
7. Trempy, J.E., Bonamy, C., Szulmajster, J. and Haldenwang, W.G. (1985) Proc. Natl. Acad. Sci. USA 82, 4189-4192.
8. Stragier, P., Bouvier, J., Bonamy, C. and Szulmajster, J. (1984) Nature 312, 376-378.
9. Stragier, P., Parsot, C. and Bouvier, J. (1985) FEBS Lett. 187, 11-15.
10. Burton, Z.F., Burgess, R.R., Lin, J., Moore, D., Holder, S. and Gross, C.A. (1981). Nucl. Acids. Res. 9, 2889-2903.
11. Costanzo, M. and Pero, J. (1983) Proc. Natl. Acad. Sci. USA 80,1236-1240.
12. Landick, R., Vaughn, V., Lau, E.T., Van Bogelen, R.A., Erickson, J.W. and Neidhardt, F.C. (1984) Cell 38, 175-182.
13. Yura, T., Takashi, Y., Tobe, T., Ito, K. and Osawa, T. (1984) Proc. Natl. Acad. Sci. 81, 6803-6807.
14. Costanzo, M., Brzustowicz, L., Hannett, N. and Pero, J. (1984) J. Mol. Biol. 180, 533-547.
15. Gram, H. and Ruger, W. (1985) EMBO J. 4, 257-264.
16. Gitt, M.A., Wang, L-F. and Doi, R.H. (1985) J. Biol. Chem. 260, 7178-7185.
17. Pabo, C.O. and Lewis, M. (1982) Nature 298, 443-447.
18. Anderson, W.F., Ohlendorf, D.H., Takeda, Y. and Matthews, B.W. (1981). Nature 290, 754-758.
19. McKay, D.B. and Steitz, T.A. (1981) Nature, 744-749.
20. Pabo C. and Sauer, R.T. (1984) Ann. Rev. Biochem. 53, 293-323.
21. Hochschild, A., Irwin, N. and Ptashne, M. (1983) Cell 32, 319-325.
22. Doolittle, R.F. (1981) Science 214, 149-159.
23. Costanzo, M. and Pero, J. (1984) J. Biol. Chem. 259, 6681-6685.
24. Lowe, P.A., Hager, D.A. and Burgess, R.R. (1979) Biochemistry 18, 1344-1352.
25. Doi, R.H. (1982) Arch. Biochem. Biophys. 214, 777-781.
26. Chelm, B., Beard, C. and Geiduschek, E.P. (1981). Biochemistry 20, 6564-6569.
27. Losick, R. and Pero, J. (1981) Cell 25, 582-584.
28. DeHaseth, P.L., Lohman, T.M., Burgess, R.R. and Record M.T. (1978) Biochemistry 17, 1612-1622.
29. Christensen, A.C. and Young, E.T. (1982). Nature 299, 369-371.
30. Christensen, A.C. and Young, E.T. (1983). In Bacteriophage T4, Matthews, C.K., Kutter, E.M., Mosig, G., Berget, P.B., eds. Am. Soc. Microbiol. Washington D.C., pp. 184-188.
31. Elliot, T. and Geiduschek, E.P. (1984) Cell 36, 211-219.
32. Rabussay, D. (1983) in Bacteriophage T4, Matthews, C.K., Kutter, E.M., Mosig, G. and Berget, P.B. Eds, pp. 167-173, Am. Soc. Microbiol., Washington D.C.
33. Erickson, B.D., Burton, Z.F., Watanabe, K.K., and Burgess, R.R. (1985) Gene 40,67-78.
34. Hu, J.C. and Gross, C.A. (1983) Mol. Gen. Genet. 191, 492-498.
35. Silverstone, A.E., Goman, M. and Scaife, J.G. (1972) Mol. Gen. Genet. 118, 223-234.
36. Hu, J.C. and Gross, C.A. (1985) Mol. Gen. Genet. 199, 7-13.

37. Devereux, J., Haeberli, P. and Smithies, O. (1984) Nucl. Acids Res. 12, 387-395.
38. Maizel, J.V. Jr. and Lenk, R.P. (1981) Proc. Natl. Acad. Sci. USA 78, 7665-7669.
39. Smith, T.F. and Waterman, M.S. (1981) Adv. Appl. Math. 2, 482-489.
40. Schwartz, R.M. and Dayhoff, M.O. (1979) in Atlas of Protein Sequence and Structure, Dayhoff, M.O. Ed, pp. 353-358, National Biomedical Research Foundation, Washington D.C.
41. Barker, W.C. and Dayhoff, M.O. (1972) in Atlas of Protein Sequence and Structure, Dayhoff, M.O. Ed, vol. 5, pp. 101-110, National Biomedical Research Foundation, Washington D.C.
42. Chou, P.Y. and Fasman, G. (1978). Adv. Enz. 47, 45-147.
43. Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) J. Mol. Biol. 120, 97-120.
44. Sauer, R.T. (1978) Nature 276, 301-302.
45. Hsiang, M.W., Cole, R.D., Takeda, Y. and Echols, H. (1977) Nature 270, 274-277.
46. Schwarz, E., Scherer, G., Hobom, G. and Kossel, H. (1978) Nature 272, 410-414.
47. Ovchinnikov, Yu.A., Guryev, S.O., Krayev, A.S., Monastryskaya, G.S., Skryabin, K.G., Sverdlov, E.D., Zakharyev, V.M. and Bayev, A.A. (1979) Gene 6, 235-249.
48. Sauer, R.T., Pan, J., Hopper, P., Hehir, K., Brown, J. and Poteete, A.N. (1981) Biochemistry 20, 3591-3598.
49. Stoner, C.M. and Schleif, R. (1982) J. Mol. Biol. 154, 649-652.
50. Aiba, H., Fujimoto, S. and Ozaki, N. (1982). Nucl. Acids Res.
51. Cossart, P. and Gicquel-Sanzey, B. (1982). Nucl. Acids Res. 10, 1363-1378. 10, 1345-1361.
52. Shaw, D.J. and Guest, J.R. (1982) Nucl. Acids Res. 10, 6119-6130.
53. Von Wilcken-Bergmann, B. and Muller-Hill, B. (1982) Proc. Natl. Acad. Sci. USA 79, 2427-2431.
54. Bayreuther, K., Adler, K., Giesler, N. and Klemmn, A. (1973). Proc. Natl. Acad. Sci. USA 70, 3576-3580.
55. Markham, B.E., Little, J.W. and Mount, D.W. (1981) Nucl. Acids Res. 9, 4149-4161.
56. Horii, T., Ogawa, T. and Ogawa, H. (1981) Cell 23, 689-697.
57. Heffron, F., McCarthy, B.J., Ohtsubo, H. and Ohtsubo, E. (1979) Cell 18, 1153-1163.
58. Reed, R.R., Shibuya, G.I. and Steitz, J.A. (1982) Nature 300, 381-383.
59. Silverman, M., Zieg, J., Mandel, G. and Simon, M. (1980) Cold Spring Harbor Symp. Quant. Biol. 45, 17-26.
60. Gribskov, M., Burgess, R.R., and Devereux, J. (1986) Nucl. Acids Res. 14, 327-334.