J. Mol. Biol. (1984) **173**, 265–269

LETTER TO THE EDITOR

Nucleotide Sequence of the Gene for Bacteriophage T7 RNA Polymerase

Differences between two previously published nucleotide sequences for bacteriophage T7 gene *1* have been resolved. The revised sequence has eight changes from the sequence that was used to compile the complete nucleotide sequence of T7 DNA. The revisions do not change the total number of nucleotides in T7 DNA or the predicted number of amino acids in T7 RNA polymerase. Only one of the changes introduces any change in predicted cleavage sites for known restriction endonucleases, and the correctness of the revised sequence at this position has been confirmed by cutting T7 DNA with the appropriate enzyme. However, the revisions do make a substantial difference in the amino acid sequence predicted for T7 RNA polymerase: 37 of the 883 amino acids are changed, 35 because of a shift in reading frame for one stretch of 37 amino acids. The predicted reading frame through this region now agrees with that predicted for the same region of the homologous T3 RNA polymerase. The calculated molecular weight for T7 RNA polymerase is now 98,856.

We have recently published a nucleotide sequence for the entire bacteriophage T7 DNA, 39,936 base-pairs (Dunn & Studier, 1983). The only part of this sequence not determined by us was nucleotides 3283 to 5901, the region that contains most of the coding sequence of gene *1* (T7 RNA polymerase, 3171 to 5822), a promoter for T7 RNA polymerase (5831 to 5853) and an RNAase III cleavage site (5847 to 5899). The sequence we used for nucleotides 3283 to 5821 was determined by Stahl & Zinn (1981), who sequenced a cloned copy of gene *1* (3127 to 5821), and the sequence we used for nucleotides 5822 to 5901 was determined by Oakley & Coleman (1977), who sequenced an *Hpa*II fragment directly from T7 DNA (5764 to 5901).

Grachev & Pletnev (1981) have also reported a sequence for nucleotides 2858 to 5855, but their sequence is different from ours in six places in the region where the two sequences overlap, and it disagrees with the Stahl & Zinn (1981) sequence in an additional 37 places. We used the Stahl & Zinn sequence in compiling the complete sequence of T7 DNA because it was in complete agreement with our sequence in the region of overlap, and with all of the restriction mapping we had done. The Grachev & Pletnev sequence, on the other hand, was discrepant in several respects with our restriction mapping data, most notably in restriction sites for *Ava*II, *Bst*NI, *Eco*B and *Xmn*I.

In terms of the biology of T7, the most serious discrepancies between the Stahl & Zinn and the Grachev & Pletnev sequences for gene *1* are the four places where single nucleotides are inserted or deleted relative to the other sequence. These insertions and deletions shift the reading frame for translation, giving sequences that predict substantial differences in the amino acid sequence in two regions

0022-2836/84/060265-05 $03.00/0

TABLE 1

Revisions to the Stahl & Zinn (1981) nucleotide sequence for T7 gene 1, and to the Dunn & Studier (1983) nucleotide sequence for T7 DNA

Position	Stahl & Zinn	Revised	DNA strand sequenced†
4331	AG_A	AGGA	*l, r*
4442	GTTC	GT C	*l, r*
4498–9	CGTC	CTGC	*l, l*
4590	TTTC	TTCC	*l*
4595	TTTC	TTCC	*l*
4916	CTGC	CTAC	*r*
5222	TCCG	TCTG	*l*

Hyphens in sequences have been omitted for clarity.

† The *l* strand has its 5′ end at the genetic left end of T7DNA; the *r* strand is its complement.

totalling about 20 amino acids. The problem was compounded by the recent determination of the nucleotide sequence of the RNA polymerase gene of a related bacteriophage, T3, by McAllister *et al.* (1983). Their sequence of this homologous gene agrees with the Stahl & Zinn sequence for two of the four insertions or deletions, and agrees with the Grachev & Pletnev sequence for the other two.

Because of the central importance of T7 RNA polymerase in the biology of T7, the interest in the enzyme itself, and the question of evolutionary divergence between the T3 and T7 RNA polymerases, we decided to resolve the discrepancies between the previous nucleotide sequences by making a third independent determination of the nucleotide sequence of gene *1*. In so doing, we have sequenced the entire region of T7 DNA not previously determined by us, that is, nucleotides 3283 to 5901. This was done by the techniques of Maxam & Gilbert (1979) on fragments of DNA isolated from wild-type T7 phage particles. Because sequences were already available, the majority of the region was sequenced on one strand only, and not every restriction cleavage site was overlapped. The sequence we obtained seems unambiguous, and agrees with at least one of the previously published sequences at every position. The changes from the Stahl & Zinn (1981) sequence (and therefore the changes from the sequence given by Dunn & Studier, 1983) are listed in Table 1. The revised nucleotide sequence of gene *1* and the amino acid sequence it predicts for T7 RNA polymerase are given in Figure 1.

Only one of the changes from the Stahl & Zinn sequence creates or eliminates predicted cleavage sites for restriction endonucleases whose specificities are currently known (Roberts, 1983). The G-T to T-G change at position 4498 to 4499 (Table 1) eliminates the *Aat*II (G-A-C-G-T-C) and *Hgi*DI (G-Pu-C-G-Py-C)

FIG. 1. Nucleotide sequence of T7 gene *1* mRNA and predicted amino acid sequence of T7 RNA polymerase. The gene *1* mRNA is the product of RNAaseIII cleavages (Dunn & Studier, 1983), and corresponds to nucleotides 3139 to 5887 of the *l* strand of T7 DNA (see Table 1 for definition). The DNA just past the coding sequence contains a promoter for T7 DNA polymerase that initiates RNA chains at nucleotide 5848 (Oakley & Coleman, 1977).

5'-AGGTACGATTTACTAACTGGAAGAGGCACTAAATGAACACGATTAACATCGCTAAGAACGAC 3200
MET ASN THR ILE ASN ILE ALA LYS ASN ASP
10

TTCTCTGACATCGAACTGGCTGCTATCCCGTTCAACACTCTGGCTGACCATTACGGTGAGCGTTTAGCTCGCGAACAGTTGGCCCTTGAGCATGAGTCTT 3300
PHE SER ASP ILE GLU LEU ALA ALA ILE PRO PHE ASN THR LEU ALA ASP HIS TYR GLY GLU ARG LEU ALA ARG GLU GLN LEU ALA LEU GLU HIS GLU SER
20 30 40

ACGAGATGGGTGAAGCACGCTTCCGCAAGATGTTTGAGCGTCAACTTAAAGCTGGTGAGGTTGCGGATAACGCTGCCGCCAAGCCTCTCATCACTACCCT 3400
TYR GLU MET GLY GLU ALA ARG PHE ARG LYS MET PHE GLU ARG GLN LEU LYS ALA GLY GLU VAL ALA ASP ASN ALA ALA ALA LYS PRO LEU ILE THR THR LEU
50 60 70

ACTCCCTAAGATGATTGCACGCATCAACGACTGGTTTGAGGAAGTGAAAGCTAAGCGCGGCAAGCGCCCGACAGCCTTCCAGTTCCTGCAAGAAATCAAG 3500
LEU PRO LYS MET ILE ALA ARG ILE ASN ASP TRP PHE GLU GLU VAL LYS ALA LYS ARG GLY LYS ARG PRO THR ALA PHE GLN PHE LEU GLN GLU ILE LYS
80 90 100 110

CCGGAAGCCGTAGCGTACATCACCATTAAGACCACTCTGGCTTGCCTAACCAGTGCTGACAATACAACCGTTCAGGCTGTAGCAAGCGCAATCGGTCGGG 3600
PRO GLU ALA VAL ALA TYR ILE THR ILE LYS THR THR LEU ALA CYS LEU THR SER ALA ASP ASN THR THR VAL GLN ALA VAL ALA SER ALA ILE GLY ARG
120 130 140

CCATTGAGGACGAGGCTCGCTTCGGTCGTATCCGTGACCTTGAAGCTAAGCACTTCAAGAAAAACGTTGAGGAACAACTCAACAAGCGCGTAGGGCACGT 3700
ALA ILE GLU ASP GLU ALA ARG PHE GLY ARG ILE ARG ASP LEU GLU ALA LYS HIS PHE LYS LYS ASN VAL GLU GLU GLN LEU ASN LYS ARG VAL GLY HIS VAL
150 160 170

CTACAAGAAAGCATTTATGCAAGTTGTCGAGGCTGACATGCTCTCTAAGGGTCTACTCGGTGGCGAGGCGTGGTCTTCGTGGCATAAGGAAGACTCTATT 3800
TYR LYS LYS ALA PHE MET GLN VAL VAL GLU ALA ASP MET LEU SER LYS GLY LEU LEU GLY GLY GLU ALA TRP SER SER TRP HIS LYS GLU ASP SER ILE
180 190 200 210

CATGTAGGAGTACGCTGCATCGAGATGCTCATTGAGTCAACCGGAATGGTTAGCTTACACCGCCAAAATGCTGGCGTAGTAGGTCAAGACTCTGAGACTA 3900
HIS VAL GLY VAL ARG CYS ILE GLU MET LEU ILE GLU SER THR GLY MET VAL SER LEU HIS ARG GLN ASN ALA GLY VAL VAL GLY GLN ASP SER GLU THR
220 230 240

TCGAACTCGCACCTGAATACGCTGAGGCTATCGCAACCCGTGCAGGTGCGCTGGCTGGCATCTCTCCGATGTTCCAACCTTGCGTAGTTCCTCCTAAGCC 4000
ILE GLU LEU ALA PRO GLU TYR ALA GLU ALA ILE ALA THR ARG ALA GLY ALA LEU ALA GLY ILE SER PRO MET PHE GLN PRO CYS VAL VAL PRO PRO LYS PRO
250 260 270

GTGGACTGGCATTACTGGTGGTGGCTATTGGGCTAACGGTCGTCGTCCTCTGGCGCTGGTGCGTACTCACAGTAAGAAAGCACTGATGCGCTACGAAGAC 4100
TRP THR GLY ILE THR GLY GLY GLY TYR TRP ALA ASN GLY ARG ARG PRO LEU ALA LEU VAL ARG THR HIS SER LYS LYS ALA LEU MET ARG TYR GLU ASP
280 290 300 310

GTTTACATGCCTGAGGTGTACAAAGCGATTAACATTGCGCAAAACACCGCATGGAAAATCAACAAGAAAGTCCTAGCGGTCGCCAACGTAATCACCAAGT 4200
VAL TYR MET PRO GLU VAL TYR LYS ALA ILE ASN ILE ALA GLN ASN THR ALA TRP LYS ILE ASN LYS LYS VAL LEU ALA VAL ALA ASN VAL ILE THR LYS
320 330 340

GGAAGCATTGTCCGGTCGAGGACATCCCTGCGATTGAGCGTGAAGAACTCCCGATGAAACCGGAAGACATCGACATGAATCCTGAGGCTCTCACCGCGTG 4300
TRP LYS HIS CYS PRO VAL GLU ASP ILE PRO ALA ILE GLU ARG GLU GLU LEU PRO MET LYS PRO GLU ASP ILE ASP MET ASN PRO GLU ALA LEU THR ALA TRP
350 360 370

GAAACGTGCTGCCGCTGCTGTGTACCGCAAGGACAAGGCTCGCAAGTCTCGCCGTATCAGCCTTGAGTTCATGCTTGAGCAAGCCAATAAGTTTGCTAAC 4400
LYS ARG ALA ALA ALA ALA VAL TYR ARG LYS ASP LYS ALA ARG LYS SER ARG ARG ILE SER LEU GLU PHE MET LEU GLU GLN ALA ASN LYS PHE ALA ASN
380 390 400 410

CATAAGGCCATCTGGTTCCCTTACAACATGGACTGGCGCGGTCGTGTTTACGCTGTGTCAATGTTCAACCCGCAAGGTAACGATATGACCAAAGGACTGC 4500
HIS LYS ALA ILE TRP PHE PRO TYR ASN MET ASP TRP ARG GLY ARG VAL TYR ALA VAL SER MET PHE ASN PRO GLN GLY ASN ASP MET THR LYS GLY LEU
420 430 440

TTACGCTGGCGAAAGGTAAACCAATCGGTAAGGAAGGTTACTACTGGCTGAAAATCCACGGTGCAAACTGTGCGGGTGTCGATAAGGTTCCGTTCCCTGA 4600
LEU THR LEU ALA LYS GLY LYS PRO ILE GLY LYS GLU GLY TYR TYR TRP LEU LYS ILE HIS GLY ALA ASN CYS ALA GLY VAL ASP LYS VAL PRO PHE PRO GLU
450 460 470

GCGCATCAAGTTCATTGAGGAAAACCACGAGAACATCATGGCTTGCGCTAAGTCTCCACTGGAGAACACTTGGTGGGCTGAGCAAGATTCTCCGTTCTGC 4700
ARG ILE LYS PHE ILE GLU GLU ASN HIS GLU ASN ILE MET ALA CYS ALA LYS SER PRO LEU GLU ASN THR TRP TRP ALA GLU GLN ASP SER PRO PHE CYS
480 490 500 510

TTCCTTGCGTTCTGCTTTGAGTACGCTGGGGTACAGCACCACGGCCTGAGCTATAACTGCTCCCTTCCGCTGGCGTTTGACGGGTCTTGCTCTGGCATCC 4800
PHE LEU ALA PHE CYS PHE GLU TYR ALA GLY VAL GLN HIS HIS GLY LEU SER TYR ASN CYS SER LEU PRO LEU ALA PHE ASP GLY SER CYS SER GLY ILE
520 530 540

AGCACTTCTCCGCGATGCTCCGAGATGAGGTAGGTGGTCGCGCGGTTAACTTGCTTCCTAGTGAAACCGTTCAGGACATCTACGGGATTGTTGCTAAGAA 4900
GLN HIS PHE SER ALA MET LEU ARG ASP GLU VAL GLY GLY ARG ALA VAL ASN LEU LEU PRO SER GLU THR VAL GLN ASP ILE TYR GLY ILE VAL ALA LYS LYS
550 560 570

AGTCAACGAGATTCTACAAGCAGACGCAATCAATGGGACCGATAACGAAGTAGTTACCGTGACCGATGAGAACACTGGTGAAATCTCTGAGAAAGTCAAG 5000
VAL ASN GLU ILE LEU GLN ALA ASP ALA ILE ASN GLY THR ASP ASN GLU VAL VAL THR VAL THR ASP GLU ASN THR GLY GLU ILE SER GLU LYS VAL LYS
580 590 600 610

CTGGGCACTAAGGCACTGGCTGGTCAATGGCTGGCTTACGGTGTTACTCGCAGTGTGACTAAGCGTTCAGTCATGACGCTGGCTTACGGGTCCAAAGAGT 5100
LEU GLY THR LYS ALA LEU ALA GLY GLN TRP LEU ALA TYR GLY VAL THR ARG SER VAL THR LYS ARG SER VAL MET THR LEU ALA TYR GLY SER LYS GLU
620 630 640

TCGGCTTCCGTCAACAAGTGCTGGAAGATACCATTCAGCCAGCTATTGATTCCGGCAAGGGTCTGATGTTCACTCAGCCGAATCAGGCTGCTGGATACAT 5200
PHE GLY PHE ARG GLN GLN VAL LEU GLU ASP THR ILE GLN PRO ALA ILE ASP SER GLY LYS GLY LEU MET PHE THR GLN PRO ASN GLN ALA ALA GLY TYR MET
650 660 670

GGCTAAGCTGATTTGGGAATCTGTGAGCGTGACGGTGGTAGCTGCGGTTGAAGCAATGAACTGGCTTAAGTCTGCTGCTAAGCTGCTGGCTGCTGAGGTC 5300
ALA LYS LEU ILE TRP GLU SER VAL SER VAL THR VAL VAL ALA ALA VAL GLU ALA MET ASN TRP LEU LYS SER ALA ALA LYS LEU LEU ALA ALA GLU VAL
680 690 700 710

AAAGATAAGAAGACTGGAGAGATTCTTCGCAAGCGTTGCGCTGTGCATTGGGTAACTCCTGATGGTTTCCCTGTGTGGCAGGAATACAAGAAGCCTATTC 5400
LYS ASP LYS LYS THR GLY GLU ILE LEU ARG LYS ARG CYS ALA VAL HIS TRP VAL THR PRO ASP GLY PHE PRO VAL TRP GLN GLU TYR LYS LYS PRO ILE
720 730 740

AGACGCGCTTGAACCTGATGTTCCTCGGTCAGTTCCGCTTACAGCCTACCATTAACACCAACAAAGATAGCGAGATTGATGCACACAAACAGGAGTCTGG 5500
GLN THR ARG LEU ASN LEU MET PHE LEU GLY GLN PHE ARG LEU GLN PRO THR ILE ASN THR ASN LYS ASP SER GLU ILE ASP ALA HIS LYS GLN GLU SER GLY
750 760 770

TATCGCTCCTAACTTTGTACACAGCCAAGACGGTAGCCACCTTCGTAAGACTGTAGTGTGGGCACACGAGAAGTACGGAATCGAATCTTTTGCACTGATT 5600
ILE ALA PRO ASN PHE VAL HIS SER GLN ASP GLY SER HIS LEU ARG LYS THR VAL VAL TRP ALA HIS GLU LYS TYR GLY ILE GLU SER PHE ALA LEU ILE
780 790 800 810

CACGACTCCTTCGGTACCATTCCGGCTGACGCTGCGAACCTGTTCAAAGCAGTGCGCGAAACTATGGTTGACACATATGAGTCTTGTGATGTACTGGCTG 5700
HIS ASP SER PHE GLY THR ILE PRO ALA ASP ALA ALA ASN LEU PHE LYS ALA VAL ARG GLU THR MET VAL ASP THR TYR GLU SER CYS ASP VAL LEU ALA
820 830 840

ATTTCTACGACCAGTTCGCTGACCAGTTGCACGAGTCTCAATTGGACAAAATGCCAGCACTTCCGGCTAAAGGTAACTTGAACCTCCGTGACATCTTAGA 5800
ASP PHE TYR ASP GLN PHE ALA ASP GLN LEU HIS GLU SER GLN LEU ASP LYS MET PRO ALA LEU PRO ALA LYS GLY ASN LEU ASN LEU ARG ASP ILE LEU GLU
850 860 870

GTCGGACTTCGCGTTCGCGTAACGCCAAATCAATACGACTCACTATAGAGGGACAAACTCAAGGTCATTCGCAAGAGTGGCCTTTAT-3'
SER ASP PHE ALA PHE ALA
880

FIG. 1.

TABLE 2

Predicted amino acid composition of T7 RNA polymerase

Amino acid	No. per molecule
Ala	100
Arg	41
Asn	40
Asp	43
Cys	12
Gln	33
Glu	67
Gly	54
His	22
Ile	52
Leu	67
Lys	66
Met	26
Phe	37
Pro	37
Ser	41
Thr	44
Trp	19
Tyr	24
Val	58
Total	883
M_r	98,856

cleavage sites that are predicted by the Stahl & Zinn sequence. We have confirmed the absence of an *Hgi*DI site at this position by analyzing the pattern of fragments produced from T7 DNA by this enzyme (obtained from New England Biolabs).

The changes in the predicted amino acid sequence are much more extensive, 37 of the 883 positions being affected. Two of these changes are due to base substitutions at nucleotides 4498 and 4590 (amino acids 443 and 474); the others are due to a shift in reading frame between nucleotides 4331 and 4442, affecting amino acids 388 to 424. This reading frame for T7 RNA polymerase is now the same as that for the equivalent region of T3 RNA polymerase, as predicted by the McAllister *et al.* (1983) nucleotide sequence. The amino acid composition predicted by the revised nucleotide sequence is given in Table 2; the calculated molecular weight for T7 RNA polymerase is now 98,856 instead of 98,092.

We thank W. Crockett for able technical assistance. This research was carried out at Brookhaven National Laboratory under the auspices of the United States Department of Energy. Barbara Moffatt, a doctoral student in the Department of Medical Genetics, University of Toronto, was supported by a fellowship from the Ontario Government and the Department of Medical Genetics, University of Toronto.

Biology Department
Brookhaven National Laboratory
Upton, N.Y. 11973, U.S.A.

BARBARA A. MOFFATT
JOHN J. DUNN
F. WILLIAM STUDIER

Received 17 October 1983

REFERENCES

Dunn, J. J. & Studier, F. W. (1983). *J. Mol. Biol.* **166**, 477–536.
Grachev, M. A. & Pletnev, A. G. (1981). *FEBS Letters*, **127**, 53–56.
Maxam, A. M. & Gilbert, W. (1979). In *Methods in Enzymology* (Grossman, L. & Moldave, K., eds), vol. 65, pp. 499–559, Academic Press, New York.
McAllister, W. T., Horn, N. J., Bailey, J. N., MacWright, R. S., Jolliffe, L., Gocke, C., Klement, J. F., Dembinski, D. R. & Cleaves, G. R. (1983). In *Gene Expression, UCLA Symposia on Molecular and Cellular Biology, New Series* (Hamer, D. & Rosenberg, M., eds), vol. 8, Alan R. Liss, Inc., New York (in the press).
Oakley, J. L. & Coleman, J. E. (1977). *Proc. Nat. Acad. Sci., U.S.A.* **74**, 4266–4270.
Roberts, R. J. (1983). *Nucl. Acids Res.* **11**, r135–r167.
Stahl, S. J. & Zinn, K. (1981). *J. Mol. Biol.* **148**, 481–485.

Edited by S. Brenner