Volume 16 Number 16 1988 Nucleic Acids Research

Relatedness of archaebacterial RNA polymerase core subunits to their eubacterial and eukaryotic equivalents

Bruno Berghöfer, Lothar Kröckel, Christiane Körtner[+], Mathias Truss[§], Jürgen Schallenberg and Albrecht Klein*

Molecular Genetics, Department of Biology, Philipps-University, Karl-von-Frisch-Strasse, D-3550 Marburg, FRG

Received June 17, 1988; Accepted July 26, 1988 Accession no. X08038

ABSTRACT

The sequence of the genes encoding the four largest subunits of the RNA polymerase of the archaebacterium Methanobacterium thermoautotrophicum was determined and putative translation signals were identified. The genes are more strongly homologous to eukaryotic than to eubacterial RNA polymerase genes. Analysis of the polypeptide sequences revealed colinearity of two pairs of adjacent archaebacterial genes encoding the B" and B' or A and C genes, respectively, with two eubacterial and two eukaryotic genes each encoding the two largest RNA polymerase subunits. This difference in sequence organization is discussed in terms of gene fusion in the course of evolution. The degree of conservation is much higher between the archaebacterial and the eukaryotic polypeptides than between the archaebacterial and the eubacterial enzyme. Putative functional domains were identified in two of the subunits of the archaebacterial enzyme.

INTRODUCTION

DNA dependent RNA polymerases are enzymes common to all organisms. They must have evolved during the establishment of the gene expression apparatus of an ancient cell. Since gene expression is similar in all kingdoms of organisms it is reasonable to assume a monophyletic descent of the presently known cellular RNA polymerases. However, the enzymes have considerably changed in the course of evolution, which is obvious from the different numbers of RNA polymerases in prokaryotic and eukaryotic cells in which there exist three different nuclear enzymes with specialized functions, while prokaryotes possess only one RNA polymerase (1). The architecture of the enzymes also differs. Eukaryotic RNA polymerases are made up of a relatively large number of subunits, while the eubacterial enzymes are less complex.

© IRL Press Limited, Oxford, England.

The discovery of a separate kingdom of organisms, the prokaryotic archaebacteria (2), has made a third line of evolution of the organisms visible. Archaebacteria combine features of eubacteria and eukaryotes and unique traits, with respect to their cytology, physiology and gene expression apparatus.

The analysis of their RNA polymerases has therefore been of major interest. It has been found that their subunit composition resembles the eukaryotic types rather than the typical eubacterial enzymes (3). Immunological cross-reactions have indicated relationships among the enzymes of members of all three kingdoms (4). They have also suggested that the four largest subunits A, B', B" and C of the RNA polymerases of halophilic and methanogenic archaebacteria are homologous to the two largest subunits of their eubacterial and eukaryotic counterparts.

We have previously located the genes encoding these four subunits on a contiguous DNA fragment in the methanogenic archaebacterium *Methanobacterium thermoautotrophicum* where they are arranged in the order B", B', A and C (5). In this communication we present the complete sequences of the four genes and the deduced amino acid sequences of the subunits, which allows a comparison with the homologous subunits of eubacterial and eukaryotic RNA polymerases. In agreement with the previous immunological studies we show that two pairs of the archaebacterial subunits are homologous to the two largest subunits of both eukaryotic and eubacterial enzymes. The sequence conservation is more pronounced with the eukaryotic enzymes. Putative nucleotide and DNA binding sites which are also strongly conserved can be assigned to two subunits of the *M. thermoautotrophicum* RNA polymerase.

MATERIALS AND METHODS

Bacteria and Plasmids.

Methanobacterium thermoautotrophicum (Winter) was obtained from J. Winter, Regensburg. *E. coli* strains JM83 (6), DH5 (7), and SG4044 (8) were used for cloning and plasmid preparation according to standard procedures (9). Plasmids pEx31 (10), pUC8 (6), pIC19 and pIC20 (11) were from our collection.

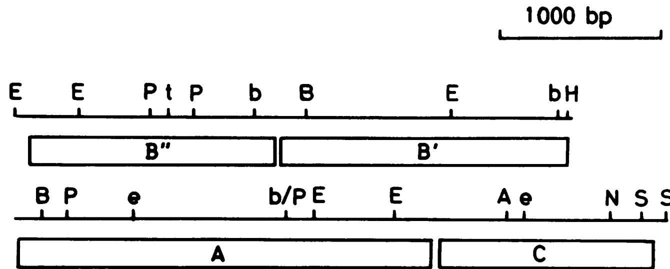

Figure 1. Restriction sites used for subcloning and sequencing RNA polymerase genes of *M. thermoautotrophicum* (Winter). A, *Asp*718; B, *Bam*HI; b, *Bgl*II; E, *Eco*RI; e, *Eco*RV; H, *Hin*dIII; N, *Nco*I; P, *Pst*I; S, *Sau*3A. Fragments B/E in gene B', P/P in gene A, and E/e spanning the intergenic region of genes A and C were shortened by combined exonuclease III and exonuclease S1 treatment (13) in order to obtain overlapping sequences.

Enzymes.

Restriction enzymes were purchased from Boehringer Mannheim (Mannheim, FRG), Pharmacia (Freiburg, FRG), or Biolabs (Bad Schwalbach, FRG). Exonuclease III and T4 DNA ligase were products of Boehringer Mannheim. S1 nuclease was purchased from SIGMA (Munich, FRG). *E. coli* DNA polymerase I was obtained from Biolabs.

Sequencing strategy.

The genomic DNA fragment of *M. thermoautotrophicum* carrying the structural genes encoding the four RNA polymerase genes considered in this study was previously cloned (5). Fig. 1 shows the restriction sites used for terminal 5'- or 3'-labeling and recutting in the course of the sequence determination according to the chemical cleavage method (12). As indicated in the legend, three fragments were shortened according to the method of Guo and Wu (13). Suitable pIC19 or pIC20 vectors were used for this purpose. They contain restriction sites leading to 3'-protruding ends in their polylinkers, which allows directional exonucleolytic digestion of the inserts. All sequences were obtained independently for both strands.

Computer analysis.

Sequence analysis was performed with the help of the SEQNCE

program (Delaney Software) and programs written by Ralf Tolle in our laboratory.

RESULTS

Sequence of the genes encoding the four largest subunits of the RNA polymerase of M. thermoautotrophicum and of the derived polypeptides.

The sequence of the contiguous DNA fragment encoding the four largest subunits of the RNA polymerase of M. thermoautotrophicum (Winter) is shown in fig. 2. The derived polypeptides B", B', A, and C have the relative molecular masses of 62021, 75570, 97408 and 50738, respectively. These values are in agreement with those previously determined by SDS polyacrylamide gel electrophoresis (14). All four genes are preceded by Shine-Dalgarno sequences, which have been found in front of all known genes in methanogenic archaebacteria (15). GTG is used as the start codon of the genes coding for the B' and C subunits and TTG in the A gene. GTG is also found at the start of the B' gene in strain Marburg of M. thermoautotrophicum (B. Berghöfer, unpublished result). TTG is a rare start codon, but its occurrence has been described for a nif gene in Methanococcus voltae. The ATG encoded methionine following the TTG codon in the A gene is certainly an internal amino acid of the polypeptide for two reasons: 1) it is not preceded by a Shine-Dalgarno sequence (16) and 2) the derived amino acid sequence preceding this methionine is found within a conserved region as compared to the homologous genes of Drosophila (17) and Saccharomyces (18).

Gene organization.

The general architecture of the RNA polymerase genes of M. thermoautotrophicum was compared to homologous genes of other organisms (fig.3). As previously concluded from immunological cross-reactions, the A and C subunits of M. thermoautotrophicum form the joint equivalent of the homologous largest subunits of the eukaryotic and eubacterial enzymes, respectively. The combined B" and B' subunits have their homologous equivalent in the second largest subunits of the eubacterial or eukaryotic RNA polymerases. Fig. 3 also shows that conserved homologous regions

```
 -79 CGTTCAGGACTGATTTTCCAGATTAAGATAGGGCCGGCATGATAATTATTCATCATTTAATTTGTTTTGGAGGAA
           M  K  K  S  A  W  G  L  V  D  A  F  F  D  K  Y  D  L  V  D  H  H  I  H
  -4 GTCCATGAAAAAAAGTGCATGGGGATTGGTAGACGCGTTTTTTGATAAGTACGACCTGGTGGACCACCACATCCA
       S  Y  N  D  F  V  S  N  R  I  Q  E  I  I  D  T  S  E  P  I  E  L  E  Q  G
  71 TTCATACAACGACTTTGTGAGTAACAGGATACAGGAAATAATAGACACCAGCGAACCTATAGAACTTGAACAGGG
       Q  Y  T  V  E  T  G  K  V  T  I  E  K  P  F  I  K  E  A  D  G  S  K  S  K
 146 CCAATACACCGTTGAGACCGGGAAGGTTACCATCGAAAAGCCCTTCATAAAGGAGGCTGATGGTTCCAAGAGCAA
       I  Y  P  T  E  A  R  L  R  N  L  T  Y  S  A  H  M  S  L  E  M  R  L  L  K
 221 GATATACCCTACAGAGGCAAGGCTAAGGAACCTGACCTACTCCGCCCATATGAGCCTTGAAATGAGGCTCCTCAA
       E  G  G  S  E  T  E  F  E  K  V  H  I  G  E  L  P  V  M  L  K  S  E  I  C
 296 GGAGGGCGGCTCCGAGACAGAATTCGAGAAGGTCCACATCGGTGAACTGCCTGTTATGCTCAAATCAGAGATATG
       H  L  H  G  L  G  R  D  E  L  I  E  K  G  E  D  P  A  D  L  G  G  Y  F  I
 371 CCACCTGCATGGTCTCGGAAGAGATGAACTCATTGAAAAGGGCGAGGACCCGGCAGACCTTGGAGGCTACTTCAT
       V  N  A  S  E  R  S  I  V  T  M  E  E  I  A  P  N  K  I  I  L  E  R  I  G
 446 CGTCAATGCATCAGAGAGGTCCATTGTAACAATGGAGGAGATAGCCCCCAACAAGATAATCCTTGAAAGGATAGG
       E  E  D  E  K  R  A  R  A  I  V  T  S  I  R  S  G  F  R  A  R  I  S  L  E
 521 TGAAGAGGATGAGAAGAGGGCAAGGGCCATCGTCACATCCATAAGGAGCGGTTTCAGGGCAAGAATATCCCTCGA
       Y  R  K  P  R  K  T  G  V  F  L  R  I  S  F  P  Y  V  P  G  E  L  P  L  V
 596 GTACAGGAAGCCCCGGAAGACAGGCGTCTTCCTAAGGATATCGTTCCCCTACGTCCCGGGAGAACTTCCACTTGT
       I  L  L  R  A  L  G  L  A  T  D  Q  E  I  I  T  S  I  S  D  D  F  N  Y  Q
 671 AATCCTCCTGAGGGCACTGGGCCTTGCAACCGACCAAGAGATCATCACAAGCATATCAGATGACTTCAACTATCA
       M  I  A  A  D  D  I  Q  V  S  L  D  K  L  K  L  D  S  D  K  M  E  E  E  M
 746 GATGATAGCTGCAGACGACATCCAGGTATCCCTTGACAAGCTGAAACTTGACAGCGATAAAATGGAAGAAGAAAT
       D  E  E  E  R  R  E  Y  L  I  R  S  A  I  K  Y  I  G  N  R  V  A  K  G  M
 821 GGATGAAGAAGAGAGAAGGGAATACCTCATAAGAAGCGCCATAAAATACATCGGTAACCGCGTTGCCAAGGGCAT
       T  E  D  Y  R  I  K  R  A  E  D  V  I  D  R  Y  L  L  P  H  I  G  T  E  P
 896 GACCGAGGATTACCGTATAAAAAGGGCCGAGGATGTTATTGACCGTTATCTTTTACCGCACATAGGAACAGAACC
       D  K  R  L  E  K  A  V  Y  L  A  E  M  T  E  M  L  L  Q  V  I  S  G  E  R
 971 CGATAAGAGACTTGAAAAGGCAGTGTACCTTGCTGAGATGACAGAGATGCTCCTGCAGGTCATCTCAGGGGAGCG
       K  P  H  D  K  D  H  Y  T  N  K  R  L  R  V  S  G  D  L  M  E  D  L  F  R
1046 CAAACCCCACGACAAGGACCACTACACAAACAAGCGCCTCAGGGTATCAGGAGACCTCATGGAGGACCTCTTCAG
       V  A  F  T  S  L  T  R  D  M  S  Y  R  L  E  R  S  L  A  R  G  K  E  P  S
1121 GGTTGCCTTCACAAGCCTAACCAGGGATATGAGCTACCGGCTCGAGAGGAGCCTTGCAAGGGGAAAGGAACCCTC
       V  K  Q  A  V  R  S  D  V  L  S  E  N  L  K  H  A  I  A  T  G  N  W  V  G
1196 TGTAAAGCAGGCCGTGAGATCCGACGTCCTGAGCGAAAACCTGAAGCACGCCATCGCAACAGGTAACTGGGTCGG
       G  R  A  G  V  S  Q  L  L  D  R  T  S  Y  M  G  T  L  S  H  M  R  R  V  V
1271 TGGAAGGGCAGGTGTGAGCCAGCTACTGGACAGGACAAGTTACATGGGTACACTCTCCCATATGAGGAGGGTTGT
       S  P  L  S  R  S  Q  P  H  F  E  A  R  D  L  H  P  T  Q  F  G  K  I  C  P
1346 ATCACCACTGAGCAGGAGCCAGCCCCACTTCGAGGCAAGGGACCTTCACCCGACACAGTTCGGTAAGATCTGCCC
       N  E  T  P  E  S  P  N  C  G  L  V  K  N  L  A  L  M  A  K  I  S  E  G  S
1421 CAACGAGACCCCTGAAAGTCCGAACTGTGGTCTTGTCAAGAACCTCGCCCTCATGGCCAAGATATCCGAGGGCTC
       D  P  D  E  I  E  E  V  I  K  K  M  G  V  I  N  #
1496 AGACCCTGATGAAATTGAAGAAGTCATCAAAAAAATGGGGGTTATTAACTAATTTTTCCACCGGGAGGCTGTTA
       M  N  K  T  K  I  Y  I  N  G  K  L  I  G  T  C  D  N  P  E  E  F  V  E  E
1571 CGTGAATAAAACCAAGATCTACATTAATGGAAAACTTATAGGAACCTGCGACAATCCTGAAGAATTTGTTGAGGA
       I  R  A  K  R  R  S  G  E  V  S  H  E  M  N  I  T  H  Y  P  E  N  H  E  I
1646 GATACGGGCTAAGAGAAGGTCCGGAGAAGTATCCCATGAAATGAACATAACCCACTACCCTGAAAACCATGAGAT
       Y  I  F  T  D  P  G  R  A  R  R  P  L  I  I  V  E  D  G  E  P  L  L  K  E
1721 ATACATCTTCACGGATCCCGGGAGGGCCAGAAGACCCCTGATCATCGTGGAGGACGGCGAACCCCTCCTCAAGGA
       E  H  L  E  K  L  S  S  G  E  M  E  W  D  D  L  I  S  Q  G  I  I  E  Y  L
1796 GGAACACCTTGAAAAACTCAGCTCCGGTGAAATGGAATGGGACGACCTCATATCACAGGGGATTATAGAGTACCT
       D  A  E  E  E  E  N  T  Y  I  A  M  S  P  E  E  V  T  E  E  H  T  H  L  E
1871 TGACGCCGAGGAGGAGGAGAACACCTACATAGCAATGAGCCCCGAGGAGGTCACAGAGGAACATACCCACCTTGA
       I  D  P  S  T  M  L  G  I  C  A  G  I  I  P  F  A  N  H  N  S  S  P  R  N
1946 AATTGACCCGTCAACCATGCTGGGTATCTGTGCAGGGATCATACCGTTCGCAAACCACAACTCATCCCCAAGGAA
       T  M  E  A  G  M  T  K  Q  A  L  G  L  Y  A  S  N  Y  N  L  R  T  D  T  R
2021 CACCATGGAGGCGGGGATGACAAAACAGGCCCTTGGTCTCTACGCATCAAACTATAACCTCCGTACGGATACACG
       A  H  L  L  H  H  P  Q  V  P  I  V  K  T  R  I  I  D  V  T  G  Y  D  E  R
2096 CGCCCACCTCCTCCACCACCCCCAGGTCCCCATAGTCAAGACAAGGATAATCGACGTGACAGGCTACGATGAAAG
       P  S  G  Q  N  F  V  V  A  V  M  S  Y  E  G  Y  N  M  E  D  A  L  I  L  N
2171 GCCATCCGGTCAGAACTTCGTGGTGGCAGTCATGTCCTACGAGGGCTACAACATGGAGGACGCCCTCATCCTCAA
       K  A  S  L  E  R  G  L  A  R  S  S  F  F  R  S  Y  E  A  T  E  R  R  Y  P
2246 CAAGGCATCCCTTGAAAGGGGCCTTGCAAGGTCATCATTTTTCAGGTCATATGAGGCAACGGAACGCAGATACCC
       G  G  Q  E  D  R  F  E  I  P  E  K  G  V  R  G  Y  R  S  E  R  D  Y  R  H
2321 CGGGGGGCAGGAGGACCGCTTCGAGATACCTGAGAAGGGTGTCCGGGGTTACCGTTCAGAGAGGGACTACAGGCA

       L  D  E  D  G  I  I  N  P  E  T  E  V  S  S  G  D  V  L  I  G  K  T  S  P
2396 CCTTGACGAGGACGGTATAATCAACCCAGAAACAGAGGTGTCATCAGGCGATGTGCTCATCGGTAAAACATCACC
       P  R  F  L  E  E  I  D  E  F  G  T  V  A  E  R  R  R  E  T  S  V  T  V  R
2471 GCCCCGTTTCCTTGAAGAGATCGATGAATTTGGAACGGTTGCAGAGAGAAGACGTGAAACATCGGTAACCGTGAG
       H  G  E  E  G  I  V  D  A  V  L  L  T  E  T  V  E  G  S  R  L  A  K  I  R
```

```
2542 GCACGGGGAAGAGGGAATCGTTGACGCCGTTCTCCTCACAGAAACCGTTGAGGGAAGCAGACTTGCCAAGATAAG
       V  R  E  Q  R  Q  P  E  F  I  G  D  K  F  A  S  R  H  G  Q  K  G  V  V  G
2621 GGTGCGTGAACAGAGACAGCCAGAATTCATCGGTGACAAGTTCGCATCAAGGCACGGTCAGAAGGGGGTTGTGGG
       L  I  V  S  Q  E  D  M  P  F  T  E  D  G  V  V  P  D  L  I  V  N  P  H  A
2696 CCTCATAGTATCACAGGAGGACATGCCCTTCACAGAGGACGGTGTTGTACCCGACCTCATAGTGAACCCCCACGC
       I  P  S  R  M  S  V  G  Q  V  L  E  M  L  A  G  K  A  A  C  M  E  G  R  R
2771 CATACCCTCAAGGATGTCGGTGGGACAGGTCCTTGAGATGCTTGCAGGTAAGGCCGCCTGCATGGAGGGGCGGCG
       V  D  G  T  P  F  T  G  E  E  E  K  D  L  K  E  A  L  K  A  N  G  F  E  S
2846 CGTGGACGGAACACCCTTCACCGGTGAGGAAGAGAAGGACCTGAAGGAAGCCCTGAAGGCCAACGGATTTGAAAG
       A  G  V  E  T  L  Y  N  G  I  T  G  E  R  I  E  A  E  I  F  I  G  V  A  Y
2921 TGCAGGTGTTGAAACACTCTACAACGGCATAACCGGGGAAAGGATAGAGGCCGAGATATTCATAGGCGTTGCATA
       Y  Q  K  L  H  H  M  T  T  D  R  I  Y  A  R  S  R  G  P  V  Q  V  L  T  R
2996 CTACCAGAAACTTCACCACATGACAACAGACAGGATATACGCGCGTTCAAGGGGTCCGGTGCAGGTCCTTACAAG
       Q  P  T  E  G  R  A  R  E  G  G  L  R  F  G  E  M  E  R  D  C  L  I  A  H
3071 ACAACCCACAGAGGGAAGGGCGAGGGAAGGAGGTCTCAGATTCGGTGAAATGGAAAGGGACTGTCTGATCGCCCA
       G  A  A  L  A  L  K  E  R  L  L  D  E  S  D  K  Y  E  A  L  V  C  A  E  C
3146 TGGAGCAGCTCTGGCCCTGAAAGAGAGGCTTCTTGATGAATCAGACAAGTACGAGGCCCTCGTATGTGCCGAATG
       G  M  I  A  I  Y  D  K  I  R  D  K  K  Y  C  P  I  C  E  D  S  E  S  F  P
3221 TGGTATGATAGCCATCTACGACAAGATCAGGGACAAGAAGTACTGTCCAATATGTGAGGACTCTGAATCATTCCC
       V  E  I  S  Y  A  F  K  L  L  L  D  E  L  K  S  L  C  I  F  P  K  L  V  L
3296 TGTGGAGATCTCATACGCATTCAAATTACTCCTAGACGAGCTTAAAAGTCTATGCATCTTCCCAAAGCTTGTACT
       E  D  K  A  -                                      M  R  G  I  L  K  K
3371 TGAAGATAAGGCATGATAATGGATTTAAGGGAATAAACCAAAGGAGAGAATACCTTGAGAGGAATTTTAAAGAAA
       I  S  Q  I  N  F  G  L  M  S  P  E  D  I  R  K  M  S  V  A  Q  I  V  T  P
3446 ATTTCCCAGATAAACTTTGGGCTCATGTCCCCTGAGGATATAAGGAAGATGTCCGTGGCCCAGATAGTGACACCG
       D  T  Y  D  E  D  G  Y  P  I  E  N  E  L  M  D  P  R  L  G  V  I  G  S  S
3521 GACACCTATGACGAGGACGGGTACCCCATCGAGAACGAGCTCATGGATCCAAGGTTAGGGGTTATAGGTTCAAGT
       L  R  C  R  S  C  G  A  K  G  G  E  C  P  G  H  F  G  S  I  N  L  A  R  P
3596 CTAAGGTGCAGAAGCTGCGGTGCCAAGGGTGGTGAGTGCCCCGGGCACTTTGGAAGCATAAACCTTGCAAGGCCC
       V  I  H  V  G  F  A  D  T  I  H  K  I  L  S  S  I  C  R  K  C  S  R  I  L
3671 GTGATCCATGTTGGTTTTGCAGACACCATACACAAGATACTCAGTTCCATCTGCAGGAAATGCAGCAGGATACTC
       L  T  E  T  E  I  Q  D  Y  R  Q  R  I  L  E  A  M  E  R  E  E  S  L  T  P
3746 CTGACGGAAACAGAAATCCAGGACTACCGGCAGAGGATACTTGAAGCCATGGAGAGGGAGGAAAGCCTCACACCC
       I  I  K  E  I  Y  A  E  A  R  R  D  R  C  P  H  C  E  E  E  Q  E  E  I  K
3821 ATAATAAAGGAGATATACGCAGAGGCAAGACGTGACAGGTGCCCCCACTGTGAGGAGGAACAGGAAGAAATAAAA
       L  D  K  P  V  S  I  V  E  G  D  Y  K  L  T  P  S  E  V  R  E  R  L  E  R
3896 CTTGACAAACCGGTTTCAATCGTTGAGGGCGACTACAAGCTAACACCAAGTGAAGTCAGGGAGCGCCTTGAGAGG
       I  T  D  N  D  S  L  L  L  G  V  N  P  E  V  A  R  P  E  W  M  V  L  T  V
3971 ATAACCGACAATGATTCCCTTCTCCTCGGAGTCAACCCGGAGGTTGCAAGACCCGAATGGATGGTCCTCACCGTA
       L  P  V  P  P  V  T  V  R  P  S  I  T  L  E  T  G  E  R  S  E  D  D  L  T
4046 CTCCCGGTCCCCCCGGTTACCGTGAGACCATCAATCACCCTGGAGACAGGTGAAAGGTCAGAGGACGACCTCACC
       H  K  L  V  D  I  L  R  I  N  Q  R  L  K  E  N  M  E  A  G  A  P  Q  L  I
4121 CATAAACTCGTTGATATCCTCAGGATAAACCAGCGCCTCAAGGAGAACATGGAGGCCGGAGCGCCCCAGCTCATC
       V  E  D  L  W  E  L  L  Q  Y  H  V  T  T  Y  F  D  N  E  A  S  G  V  P  P
4196 GTTGAGGACCTCTGGGAGCTCCTCCAGTACCATGTGACAACCTACTTCGACAACGAGGCATCAGGAGTGCCGCCT
       A  R  H  R  S  G  R  P  L  K  T  L  A  Q  R  L  K  G  K  E  G  R  F  R  S
4271 GCAAGGCACAGGTCAGGAAGGCCCCTCAAGACCCTGGCCCAGAGACTGAAGGGTAAGGAGGGTCGTTTCAGGAGC
       N  L  S  G  K  R  V  N  F  S  A  R  T  V  V  S  P  D  P  N  V  S  V  N  E
4346 AACCTTTCAGGTAAAAGGGTTAACTTCTCAGCCCGTACAGTCGTGTCCCCTGACCCCAACGTAAGCGTGAACGAG
       V  G  V  P  E  L  I  A  K  E  V  T  V  P  V  Y  V  T  E  W  N  I  D  R  M
4421 GTGGGCGTCCCTGAACTGATAGCAAAGGAGGTCACGGTACCGGTCTATGTCACCGAGTGGAACATAGACCGCATG
       K  E  H  I  E  N  G  P  D  V  H  P  G  A  N  Y  V  I  R  P  D  G  R  K  I
4496 AAGGAGCACATAGAAAACGGACCGGACGTCCACCCCGGCGCAAACTACGTTATAAGACCCGACGGCCGGAAGATA
       R  A  Y  N  E  T  K  D  V  V  L  E  N  L  K  P  G  Y  I  V  E  R  H  L  K
4571 AGGGCCTACAATGAGACAAAGGATGTTGTGCTGGAAAACCTCAAACCCGGCTACATAGTTGAAAGGCACCTCAAG
       D  G  D  I  V  L  F  N  R  Q  P  S  L  H  R  M  S  M  M  A  H  E  V  R  V
4646 GACGGGGACATAGTGCTCTTCAACCGTCAGCCATCACTCCACAGGATGTCAATGATGGCCCATGAGGTCAGGGTC
       L  P  Y  K  T  F  R  L  N  L  C  V  C  P  P  Y  N  A  D  F  D  G  D  E  M
4721 CTCCCCTACAAGACCTTCCGTCTAAACCTCTGCGTCTGCCCCCCATACAACGCTGACTTTGACGGTGACGAGATG

       N  M  H  V  F  Q  T  E  E  S  R  A  E  A  K  T  L  M  R  V  Q  D  H  I  L
4796 AACATGCACGTGTTCCAGACAGAAGAGTCCCGTGCAGAGGCCAAGACCCTCATGCGTGTCCAGGATCATATCCTT
       S  P  R  F  R  D  L  S  S  G  V  Y  T  T  I  S  Q  E  H  T  S  S  Q  G  K
4871 TCACCAAGGTTCCGCGACCTATCATCGGGGGTATACACGACCATATCTCAGGAGCATACCTCCTCACAAGGAAAA
       R  L  F  S  V  R  S  R  P  P  D  P  Q  E  G  R  A  P  P  P  D  R  E  G  R
4946 AGGCTGTTTTCAGTGAGGAGCAGGCCTCCAGATCCTCAAGAAGGCAGGGCTCCCCCTCCCGACAGGGAGGGCCGC
       E  W  T  V  K  E  I  F  S  M  V  L  P  D  D  L  N  M  V  Y  L  A  E  I  C
5021 GAGTGGACAGTTAAGGAAATATTCAGCATGGTCCTCCCGGACGACCTCAACATGGTTTACCTGGCGGAGATCTGC
       R  K  C  D  E  C  L  E  M  D  W  E  N  D  A  Y  V  V  I  E  N  G  Q  L  I
5096 AGGAAATGTGACGAATGCCTTGAGATGGACTGGGAAAACGACGCCTACGTTGTCATAGAGAACGGTCAGCTGATA
       T  G  V  I  D  E  K  A  Y  G  A  F  A  G  K  I  L  D  Q  I  V  K  E  Y  G
```

```
5171 ACAGGTGTCATTGACGAGAAGGCCTACGGAGCATTTGCAGGTAAGATACTTGACCAGATAGTGAAGGAGTATGGA
      S  D  A  A  K  E  F  L  D  S  A  T  K  L  A  I  A  G  I  M  H  A  G  F  T
5246 AGCGACGCCGCCAAGGAATTCCTCGACTCCGCCACAAAACTGGCTATTGCCGGTATAATGCATGCAGGTTTCACC
      T  S  T  N  D  E  E  I  P  E  E  A  K  E  R  I  E  A  H  L  R  N  A  E  A
5321 ACCAGTACCAACGATGAGGAGATCCCTGAGGAGGCCAAGGAACGCATAGAGGCCCATCTCAGGAATGCTGAGGCC
      R  V  D  Q  L  I  E  A  Y  E  N  G  E  L  E  P  L  P  G  R  S  L  E  E  T
5396 CGTGTAGACCAGCTCATCGAGGCATATGAGAACGGTGAACTCGAACCCCTCCCCGGAAGGAGCCTTGAGGAAACC
      L  E  M  K  I  M  Q  V  L  G  E  A  K  D  K  S  G  E  I  A  E  S  Y  F  D
5471 CTCGAGATGAAGATAATGCAGGTCCTCGGTGAGGCCAAGGACAAGTCAGGGGAGATTGCAGAGAGCTACTTTGAC
      M  D  E  N  H  A  V  I  M  A  L  T  G  A  R  G  A  M  L  N  L  T  Q  I  T
5546 ATGGACGAGAACCACGCGGTTATAATGGCACTTACAGGTGCAAGGGGTGCCATGCTTAACCTGACGCAGATAACC
      A  C  V  G  Q  Q  S  V  H  G  G  R  I  T  R  G  Y  D  N  R  T  L  P  H  F
5621 GCATGTGTCGGTCAGCAGTCTGTCCATGGTGGTCGTATAACAAGGGGGTACGACAACAGGACGCTTCCCCACTTC
      K  K  G  E  L  G  A  K  S  R  G  F  V  H  S  S  Y  K  E  D  S  I  L  L  E
5696 AAGAAGGGAGAACTCGGTGCTAAGTCACGTGGATTCGTGCACTCAAGTTACAAGGAGGACTCGATTCTATTGGAA
      F  M  G  G  R  E  G  L  V  D  T  A  I  R  T  A  Q  S  G  Y  M  Q  R  R  L
5771 TTCATGGGTGGAAGGGAAGGTCTTGTGGACACCGCCATCCGTACAGCCCAGAGCGGTTACATGCAGAGGCGTCTT
      V  N  A  L  Q  D  L  T  V  D  E  N  G  R  V  V  D  N  R  G  V  I  I  Q  T
5846 GTGAACGCCCTCCAGGACCTCACGGTGGATGAGAATGGAAGGGTCGTTGACAACAGGGGAGTTATCATACAGACC
      R  F  G  E  D  G  V  D  P  A  K  S  D  Y  G  K  I  V  D  L  D  K  L  V  Q
5921 CGCTTCGGTGAGGATGGCGTGGACCCGGCGAAGAGTGATTACGGTAAGATAGTTGACCTTGACAAACTCGTGCAG
      E  I  R  L  K  S  G  K  -                    M  Q  D  V  I  K  K  I  E  D
5996 GAGATAAGGCTTAAATCAGGGAAGTGAAGGGGTAAGGTGATTCTGTGCAGGACGTTATTAAAAAGATAGAGGATT
      Y  S  S  R  N  G  I  M  L  P  N  P  V  T  E  Y  V  A  G  I  A  E  E  E  K
6071 ACTCCTCCAGGAACGGGATTATGCTCCCGAACCCTGTCACCGAATACGTGGCAGGGATTGCTGAGGAGGAAAAAC
      L  K  D  P  E  L  H  D  L  V  R  L  F  S  R  I  C  E  R  N  R  G  L  E  G
6146 TGAAGGACCCCGAGCTCCATGACCTTGTAAGGCTCTTCAGCAGGATATGTGAGAGAAACAGGGGCCTTGAAGGAG
      E  E  L  L  K  A  V  E  D  E  Y  L  R  I  L  K  V  R  E  L  V  K  R  K  R
6221 AAGAACTGTTAAAGGCCGTGGAGGATGAATACCTCAGGATACTGAAGGTCAGGGAACTCGTCAAAAGGAAGAGGG
      A  K  F  P  P  R  L  I  E  E  I  A  E  A  I  K  R  H  G  L  S  D  D  E  L
6296 CCAAGTTCCCCCCAAGGCTCATTGAAGAGATAGCTGAAGCAATAAAAAGGCACGGCCTCAGTGACGATGAACTCG
      D  E  L  V  R  G  V  R  R  A  Y  E  R  A  M  V  E  A  G  E  A  V  G  T  V
6371 ATGAACTTGTAAGGGGTGTCAGAAGGGCCTATGAACGTGCCATGGTTGAGGCCGGTGAAGCCGTTGGTACAGTGG
      A  A  Q  S  V  G  E  P  G  T  Q  M  T  M  R  T  F  H  Y  A  G  V  A  E  L
6446 CCGCCCAGTCAGTCGGTGAACCAGGTACCCAGATGACGATGCGTACCTTCCACTATGCAGGGGTGGCTGAGCTCA
      N  V  T  L  G  L  P  R  L  I  E  I  V  D  A  R  K  K  I  S  T  P  T  M  S
6521 ACGTTACACTCGGTCTCCCCAGGCTCATCGAGATAGTGGATGCCAGGAAGAAGATATCCACCCCAACAATGAGCA
      I  Y  F  E  G  D  R  K  Y  D  E  E  F  V  R  K  K  A  N  K  I  C  K  S  T
6596 TATACTTTGAAGGCGACAGGAAATACGATGAGGAGTTCGTGAGGAAAAAGGCCAATAAGATATGTAAGAGCACCC
      L  N  D  V  L  K  N  F  S  I  Q  Y  A  D  M  S  V  E  A  E  L  D  E  E  K
6671 TCAACGATGTCCTTAAAAATTTCAGCATCCAGTATGCCGACATGTCAGTTGAAGCAGAACTGGATGAGGAGAAAA
      I  R  E  K  H  L  E  Y  D  D  I  I  A  K  V  E  K  T  F  K  K  V  E  I  D
6746 TCCGTGAAAAACACCTTGAATATGATGATATAATAGCCAAGGTGGAAAAAACTTTTAAGAAAGTAGAAATAGATA
      N  N  I  L  R  F  E  P  P  K  P  T  I  R  E  L  R  L  L  A  D  K  V  R  K
6821 ACAACATACTGAGATTTGAACCCCCAAAACCAACCATAAGGGAGCTCAGACTGCTTGCAGATAAAGTAAGGAAAC
      L  Q  I  S  G  V  K  N  I  G  K  V  V  I  R  K  E  D  D  E  W  V  I  H  T
6896 TCCAGATAAGCGGGGTTAAAAATATAGGCAAGGTGGTTATCCGTAAAGAGGATGATGAATGGGTAATCCATACAG
      E  G  S  N  L  G  D  R  F  K  E  E  G  V  D  K  V  R  T  T  T  N  D  I  H
6971 AGGGTTCAAACCTTGGAGACCGTTTTAAAGAAGAGGGTGTAGATAAGGTTAGGACGACCACCAACGACATTCATG
      E  I  E  T  V  L  G  I  E  A  A  R  N  A  I  I  H  E  A  K  R  T  M  E  E
7046 AGATAGAAACAGTCCTTGGTATAGAGGCAGCCAGAAACGCGATAATTCACGAAGCAAAGAGAACCATGGAGGAAC
      Q  G  L  T  V  D  I  R  H  I  M  L  V  A  D  M  M  T  A  D  G  S  V  K  S
7121 AGGGTCTCACCGTGGACATAAGACATATAATGCTCGTTGCAGATATGATGACTGCCGATGGTTCTGTCAAGTCCA

      I  G  R  H  G  I  S  G  E  K  A  S  V  L  A  R  A  S  F  E  E  T  G  K  H
7196 TTGGAAGACATGGTATTAGCGGTGAGAAAGCGAGCGTCCTTGCAAGGGCTTCTTTTGAAGAAACCGGGAAACACC
      L  L  R  A  S  I  R  G  E  V  D  H  L  T  G  I  I  E  N  I  I  I  G  Q  P
7271 TTCTAAGAGCAAGTATCAGGGGAGAGGTCGATCATCTCACCGGTATTATAGAGAACATTATTATAGGGCAACCCA
      I  P  L  G  T  G  S  V  S  V  I  M  K  K  *
7346 TACCCCTGGGTACAGGTTCCGTTAGCGTTATAATGAAAAAATAGGAGGCAGTAGATGGACATAGATAGAGGAATA

7421 CGAGTCGCTGTAGATACTGGTAATGTTATTCTTGGATC
```

Figure 2. Sequence of the genes encoding the four largest subunits of the RNA polymerase of *M. thermoautotrophicum* (Winter) and derived polypeptide sequences. The numbering starts at the A of the B" gene initiation codon ATG. Putative ribosome binding sites are underlined.

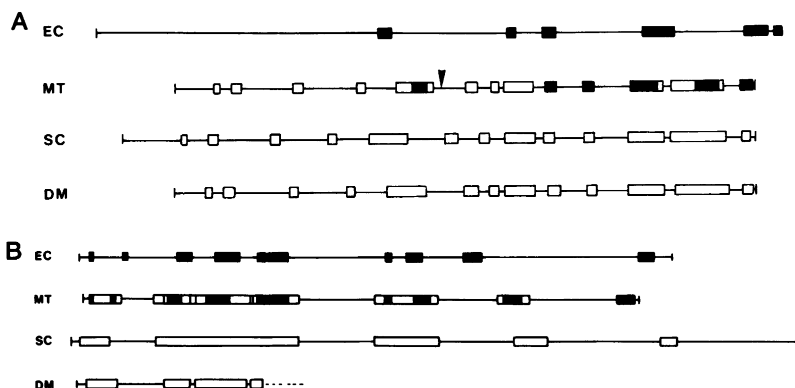

Figure 3. Comparison of the organization of the RNA polymerase encoding sequences of M. thermoautotrophicum and homologous sequences of other eubacterial and eukaryotic organisms. (A) B genes of M. thermoautotrophicum (MT) compared to rpoB (22) of Escherichia coli (EC), RPB2 (23) of Saccharomyces cerevisiae (SC), and DmRP140 (24) of Drosophila melanogaster (DM); (B) A+C genes of M. thermoautotrophicum compared to rpoC (25) of E. coli, rpo21 (18) of S. cerevisiae and the 5'-terminal part of the gene encoding the 215 kD subunit of RNA polymerase II of D. melanogaster (17). The Methanobacterium genes B" and B', as well as A and C, respectively, are shown as contiguous structures disregarding the intergenic regions. The gene borders are indicated by arrow heads. Boxes show homologies of at least 50% on the polypeptide level between the archaebacterial and eubacterial (filled bars) or eukaryotic coding sequences (open bars), respectively.

are found in the same orders in all the homologous polypeptides. The molecular architecture is strikingly similar among the archaebacterial and eukaryotic equivalents. It is most conserved among the B"+B' polypeptide of the methanogen and its equivalent in Drosophila, both with respect to the total polypeptide size and the arrangement of the strongly conserved sequence blocks. In contrast, the N-terminus of the ß subunit of E. coli has very little homology to both the archaebacterial and eukaryotic equivalents.

This divergence is also found within the homologous regions of the subunit polypeptides (fig. 4). The total homology between Saccharomyces and Methanobacterium is only marginally higher than Drosophila and the archaebacterium. This comparison neglects conservative exchanges of amino acids, which are

difficult to evaluate in the absence of information about secondary and tertiary structures of the proteins.

Putative nucleotide and DNA binding sites.

Two major binding functions must be assigned to the subunits constituting the core of DNA dependent RNA polymerases: 1) DNA binding, and 2) binding of the nucleoside triphosphates in the reactive center. Oligopeptide loops stabilized by chelation of a metal atom involving pairs of cysteine and histidine residues and exhibiting conserved hydrophobic amino acids are believed to function in binding proteins to DNA (19). Sequences which could conceivably form such zinc finger structures were found in the largest subunits of the Drosophila and yeast RNA polymerases and are also present in the equivalent subunit A of M. thermoautotrophicum (fig. 5A). In the archaebacterium one of the putatively involved histidine residues is replaced by serine.

Nucleotide analogs can be used for labeling nucleotide binding sites of RNA polymerases (20). Labeled analogs of the starting nucleoside triphosphate have been covalently attached to their binding sites and subsequently linked to a second nucleotide thus forming a ternary starting complex at the reactive center of the enzyme. Using this method, it has been shown that the RNA polymerase B' subunit of the M. thermoautotrophicum strain used in our study contains this reactive center (21). Previously, similar experiments had been carried out on RNA polymerases of E. coli and wheat germ resulting in the labeling of the ß or B subunits respectively, by reaction with the amino group of the side chain of positively charged amino acid residues. Two such potential reaction partner were identified as lysines 1048 and 1051 of the ß subunit of E. coli RNA polymerase. Other reactive analogs have revealed different reactive sites. One such site is characterized by a histidine residue at position 1237. As discussed above, the ß subunit in E. coli and the eukaryotic B subunit are the homologs of the B"+B' subunits in Methanobacterium. The amino acid residues in the ß subunit of E. coli to which two different nucleotide analogs used as affinity labels bind, are surrounded by peptide sequences which show at least moderate homology among

E. coli, Saccharomyces, Drosophila and Methanobacterium (fig. 5B). They are located in or close to neighboring conserved sequence blocks of the polypeptides. These regions in the B' subunit of the M. thermoautotrophicum RNA polymerase can therefore be assumed to form (a) reactive center(s) of the enzyme.

DISCUSSION

Gene order and translation signals.

The presented sequence of the continuous fragment of M. thermoautotrophicum DNA comprising the four genes encoding the largest subunits of RNA polymerase has confirmed the gene order

```
A MT  * PTEARLRNLTYS * KVHIGELPVMLKSEICHLHGL * RALGLATDQEIITSI *
  SC  * PQEARLRNLTYS * KVFIGRLPIMLRSKNCYLSEA * RALGIIPDGEILEHI *
  DM  * PNEARLRNLTYS * KTFIGKIPIMLRSTYCLLSQL * RALGFVADRDILEHI *

  EC                                                          * FEVRDVHP-
  MT  * DKDHYTNKRLRVSGDL * RAGVSQLLDRTSYMGTLSHMRRVVSPLSRSQPHFEARDLHP-
  SC  * DRDHFGKKRLDLAGPL * RAGVSQVLNRYTYSSTLSHLRRTNTPIGRDGKLAKPRQLHN-
  DM  * DRDHYGNKRLDLAGPL * RAGVSQVLNRLTFASTLSHLRRVNSPIGRDGKLAKPRQLHN-

  EC   -THYGRVCPIATPEGPNIGL*
  MT   -TQFGKICPNETPESPNCGLVKNLALMAKISEGSDP *
  SC   -THWGLVCPAETPEGQACGLVKNLSLMSCISVGTDP *
  DM   -TLWGMLCPAETPEGAAVGLVKNLALMAYISVGSQP *

  MT  * IFTDAGRVYRPLFIVEDDESLGHKEL * EYLDAEEEENTYIAMSPE *
  SC  * IFTDPGRARRPLIIVEDGEPLLKEEH * EYIDAEEEESILIAMQPE *
  DM  * IYTDAGRICRPLLIVENGSLLLKKTH * EYMYTLEEETVMIAMSPY *

  MT  * THLEIDPSTMLGICAGIIPFANHNSSPRNTMEAGMTKQALGLYASNYNLRTDTRAHLL *
  SC  * THCEIHPSMILGVAASIIPFPDHNQSPRNTYQSAMGKQAMGVFLTNYNVRMDTMANIL *
  DM  * THCEIHPAMILGVCASIIPFPDHNQSPRNTYQSAMGKQAMGVYITNFHVRMDTLAHVL *

  EC  * GQNMRVAFMPWNGYNFED *    * LDESGIVYIGAEVTGGDILVGK *
  MT  * GQNFVVAVMSYEGYNMEDALILN * LDEDGIINPETEVSSGDVLIGKT *
  SC  * GQNAIVAIACYSGYNQEDSMIMN * LDDDGLIAPGVRVSGEDVIIGKT *
  DM  * GINSIVAILCYTGYNQEDSVILN * LDDDGIIAPGIRVSGDDVVIGKT *

  EC   * GDKMAGRHGNKGVISKINPIEDMPYDENGTPVDIVLNPLGVPSRMNIGQILETHLGMAA *
  MT  * IGDKFASRHGQKGVVGLIVSQEDMPFTEDGVVPDLIVNPHAIPSRMSVGQVLEMLAGKAA-
  SC  * IGDKFASRHGQKGTIGITYRREDMPFTAEGIVPDLIINPHAIPSRMTVAHLIECLLSKVA-
  DM  * IGDKFASRHGQKGTCGIQYRQEDMAFTCEGLAPDIIINPHAIPSRMTIGHLIECLQGKLG-

  MT   -CMEGRRVDGTPF * GFESAGVETLYNGITGERIEAEIFIGVAYYQKLHHMTTDRIYARS-
  SC   -ALSGNEGDASPF * GYQSRGFEVMYNGHTGKKLMAQIFFGPTYYQRLRHMVDDKIHARA-
  DM   -SNKGEIGDATPF * GYHLRGNEVMYNGHTGRKINAQVFLGPTYYQRLKHMVDDKIHSRA-

  EC         * TQQPLGGKAQFGGQRFGEMEVWALEAYGAAYTLQEMLTVKSD *
  MT   -RGPVQVLTRQPTEGRAREGGLRFGEMERDCLIAHGAALALKERLLDESDKYEALVCAECGM *
  SC   -RGPMQVLTRQPVEGRSRDGGLRFGEMERDCMIAHGAASFLKERLMEASDAFRVHICGICGL *
  DM   -RGPVQILVRQPMEGRARDGGLRFGEMERDCQISHGAAQFLRERLFEVSDPYRVHICNFCGL *

  EC    * FNVLLKEIRSLGINIEL *
  MT  * YAFKLLLDELKSLCIFPKL *
  SC  * YAAKLLFQELMAMNITPRL *
  DM  * YAAKLLFQELMSMNIAPRL *
```

```
B EC    * LASPDMIRSWS *
  MT  * FGLMSPEDIRKMSVAQIVTPDTYDEDGYPIEN ELMDPRLGVIGSSLRCRSCGAKGGECPG-
  SC  * FGLFSPEEVRAISVAKIRFPETMDETQTRAKIGGLNDPRLGSIDRNLKCQTCQEGMNECPG-
  DM  * FGILSPDEIRRMSV *         * GGRPKLGGLMNPRQGVIDRTSRCQTCAGNMTECPG-

  EC    * GHIELASPTAHIWF *
  MT   -HFGSINLARPVIHVGFADTIHK *
  SC   -HFGHIDLAKPVFHVGFIAKIKK *
  DM   -HFGHIDLAKPVFHIGFITKTIK *

  EC                            * PEWMILTVLPVLPPDLRPLVPLDGGRFATSDL *
  MT  * LTPSEVRERLERITDNDSLLLGVNPEVARPEWMVLTVLPVPPVTVRPSITLETGERSEDDLT-
  SC  * LSTEEILNIFKHISVKDFTSLGFNEVFSRPEWMILTCLPVPPPPVRPSISFNESQRGEDDLT-
  DM                     * LGMDPKYARPDWMIVTVLPVPPLAVRPAVVMFGAAKNQDDLT-

  EC                                                        * RPLKSL-
  MT   -HKLVDILRINQRLKENMEAGAPQLIVEDLWELLQYHVTTYFDNEASGVPPARHRSGRPLKTL-
  SC   -FKLADILKANISLETLEHNGAPHHAIEEAESLLQFHVATYMDNDIAGQPQALQKSGRPVKSI-
  DM   -HKLSDIIKANNELRKNEASGA *      * MLQFHVATLVDNDMPGMPGYA KVGKPLKAI-

  EC   -ADMIKGKQGRFRQNLLGKRVDYSGRSVITVGPYLRLHQCGLPKKMALEL *
  MT   -AQRLKGKEGRFRSNLSGKRVNFSARTVVSPDPNVSVNEVGVPELIAKEVTVPVYVTEWNIDR-
  SC   -RARLKGKEGRIRGNLMGKRVDFSARTVISGDPNLELDQVGVPKSIAKTLTYPEVVTPYNIDK-
  DM   -KARLKGKEGRIRGNLMGKRVDFSARTVITPDPNLRIDQVGVPRSIAQNLTFPELVTPFNIDR-

  EC                                                       * PVLLNRAP-
  MT   -MKEHIENGPDVHPGANYVIRPDGRKIRAYNETKDVVLENLKPGYIVERHLKDGDIVLFNRQP-
  SC   -LTQLVRNGPNEHPGAKYVIRDSGDKIGLRYSKRAGDIQLQYGWK VERHIMDNDPVLFNRQP-
  DM   -MQELVRRGNSQYPGAKYIVRDNGERIDLRFHPKSSDL *     * VERHLRDDDLVIGTRQP-

  EC   -TLHRLGIQAFE *      * LHPLVCAAYNADFDGDQMAVHVPLTLEAQLEARALMMSTNN-
  MT   -SLHRMSMMAHEVRVLPYKTFRLNLCVCPPYNADFDGDEMNMHVFQTEESRAEAKTLMRVQDH-
  SC   -SLHKMSMMAHRVKVIPYSTFRLNLSVTSPYNADFDGDEMNLHVPQSEETRAELSQLCAVPLQ-
  DM   -RCTTMSMMVTGESV * (END OF PUBLISHED SEQUENCE)

  EC   -ILSP *                                                   * MADS-
  MT   -ILSPRFRDLSSGVYTTISQEHTSSQGKRLF * LGEAKDKSGEIAESYFDMDENHAVIMALT-
  SC   -IVSPQSNKPCMGIVQDTLCGIRKLTLRDTF * LNEARDKAGRLAE VNLKDLNNVKQMVMA-

  EC   -GARGSAAQIRQ *
  MT   -GARGAMLNLTQITACVGQQSVHGGRITRGYDNRTLPHFKKGELGAKSRGFVHSSY *
  SC   -GSKGSFINIAQMSACVGQQSVEGKRIAFGFVDRTLPHFSKDDYSPESKGFVENSY *

  EC   * GARKGLADTALKTANSGYLTRRLVDVAQDLVVTED *
  MT  * MGGREGLVDTAIRTAQSGYMQRRLVNALQDLTVDENGRVVDNRGVIIQTRFGEDGVDPA *
  SC  * MGGREGLIDTAVKTAETGYIQRRLVKALEDIMVHYDNTTRNSLGNVIQFIYGEDGMDAA *

  EC        * GEAIGVIAAQSIGEPGTQLTMRTFHIGGAASRAAAESSIQVK *
  MT  * RAMVEAGEAVGTVAAQSVGEPGTQMTMRTFHYAGVAELNVTLGLPRLIEIVDARKKISTP-
  SC  * RSVVHPGEMVGVLAAQSIGEPATQMTLNTFHFAGVASKKVTSGVPRLKEILNVAKNMKTP-

  EC                      * ASFQETTRVLTEAAVAGKRDELRGLKENVIVGRLIPAGTG *
  MT   -TMSIYFEGDRKYDEE * LARASFEETGKHLLRASIRGEVDHLTGIIENIIIGQPIPLGTG *
  SC   -SLTVYLEPGHAADQE * LMRCSFEETVEILFEAIASAELDDCRGVSENVILGQMAPIGTG *
```

Figure 4. Highly conserved areas within the polypeptide sequences of the homologous RNA polymerase subunits compared in fig. 3. (A) B genes, (B) A and C genes of *M. thermoautotrophicum* and their homologs. The sequence blocks are shown in the same N-terminal to C-terminal order as in fig. 3. MT, *M. thermoautotrophicum*; EC, *E. coli*; SC, *S. cerevisiae*; DM, *D. melanogaster*. Asterisks signify borders of more weakly (<50%) or non-homologous adjacent sequences. Hyphens indicate continuous sequences.

A

MT 60 CrsCgakggecpgHfgSinlarpvihvgfadtihkilssiCrkC 103

SC 67 CqtCqegmnecpgHfgHidlakpvfhvgfiakikkvcecvCmhC 110

DM 67 CqtCagnmtecpgHfgHidlakpvfhigfitktikilrcvCfyC 110

B

EC 1047 lkivKvylavkr-x_{174}-lklnHlvddk 1242

MT (859) srlaKirvreqr-x_{121}-qklhHmttdr (1001)

SC 961 lkfvKvrvrttk-x_{120}-qrlrHmvddk 1102

DM 856 ykfcKirvrsvr-x_{121}-qrlkHmvddk 1007

Figure 5. Comparison of putative functional sites of the homologous RNA polymerase subunits. (A) Potential zinc fingers; (B) sequences involved in nucleotide binding and polymerization. The capital letters show identified functional amino acids (see discussion in the text). The amino acid positions given for M. thermoautotrophicum are the positions in an artificial head to tail "fusion" of the B' to the B" polypeptide, which would be the equivalent of the subunit polypeptides of the other organisms. The actual positions are 344 and 486 in the B' polypeptide.

previously determined (5). The identical orientation of all four genes suggests that they could form a transcription unit like their homologs in E. coli.

Each of the genes is preceded by a Shine-Dalgarno (16) sequence in proper distance to the start codon which confirms the identified gene starts. The use of start codons other than AUG in three out of the four cases is remarkable. These codons, rarely used in other prokaryotes, have previously been found in other genes of methanogenic bacteria (26,27) including the very highly expressed gene coding for the ß subunit of methyl CoM reductase in Methanosarcina barkeri. Their relatively frequent occurrence suggests a special structure of the anticodon loop in the corresponding initiator tRNA, which would facilitate wobble in the 5'-position of the start codon.

Gene organization and homology with other organisms.

The comparison of the four RNA polymerase genes with those of other organisms yields obvious homologies. It can therefore

be concluded that the archaebacterial enzyme is of the same monophyletic descent. The colinear arrangement of conserved areas among the homologous coding sequences was noticed previously (18, 23) when eubacterial and eukaryotic genes were compared. The fact that two pairs of adjacent genes in the archaebacterium are homologs to two genes in eubacteria and eukaryotes indicates gene fusions or splitting of genes with subsequent introduction of expression signals in the course of evolution, which, however, would not appear very likely. It is remarkable that the RNA polymerase of the sulfur dependent archaebacterium Sulfolobus acidocaldarius shows an intermediate situation. Here, the homologs to the B genes of the methanogenic and halophilic archaebacteria - whose RNA polymerase gene organization corresponds to that in the methanogens - are fused as in eubacteria and eukaryotes (4).

The homology of the Methanobacterium RNA polymerase genes with the coding sequences of their eukaryotic counterparts is much more pronounced than with the homologs in E. coli. This reflects an early evolutionary separation of the prokaryotes of the two different prokaryotic urkingdoms (2). The overall degree of homology on the polypeptide level appears to be slightly higher among Methanobacterium and yeast than between the methanogen and Drosophila. However, the similarity of the architecture of the Methanobacterium genes encoding the B subunits and the DmRP140 coding sequences of Drosophila is striking. It is remarkable in this respect, that the introns found in the Drosophila gene (24) have no positional relationship to the intergenic region between the two archaebacterial genes.

Functional sites.

All DNA dependent RNA polymerases must similarly be able to bind DNA and the nucleotide substrates while they may differ in the recognition of special signals leading to correct initiation and termination of transcription. In complex proteins the formation of functional domains often involves several subunits. Therefore it is difficult to assign functions to individual subunits. In addition to the biochemical data mentioned above, mutational analysis has however led to some clues concerning the involvement of E. coli RNA polymerase subunits in both DNA and

nucleotide binding. Mutants resistant to drugs which specifically interfere with the initiation of transcription have been found in the ß-coding rpoB gene, while certain temperature sensitive mutants impaired in DNA binding have been localized in the rpoC gene encoding the ß' subunit (see 28 for review).
We have found that the amino acids identified as reaction partners for the covalent attachment of the nucleotide analogs in eukaryotic enzymes are located in regions with homology to sequences of the B' subunit of the Methanobacterium polymerase. Two such homologous areas are seen at similar distances from each other as in the eukaryotic examples and next to or within highly conserved amino acid sequence blocks. This arrangement reminds of the results of photoaffinity labeling experiments (29) which were consistent with the assumption of two different reactive sites on the ß subunit of the E. coli RNA polymerase.
In many DNA binding proteins sequences have been found which are capable of metal binding and are supposed to form "finger" structures, which supposedly interact with the DNA. The conserved structure found in the A subunit of the Methanobacterium polymerase (fig. 5A) shows the common features of such putative DNA binding regions except for the replacement of one of the histidine residues by serine. This amino acid, however, is also capable of binding metal due to the potential action of its OH group as a nucleophile in proper surrounding of neighboring proton attracting groups (30). Therefore, it is likely that the identified peptide sequence is capable of forming a finger structure. It might thus be involved in the constitution of the DNA binding domain of the RNA polymerase.

ACKNOWLEDGEMENTS

We thank H. Bestgen for excellent technical assistance, H. Steinebach for help with the preparation of the manuscript, B. Bowien and R. Tolle for computer analyses, G. Hartmann and W. Zillig for communication of unpublished results and helpful discussions. This work was supported by the Deutsche Forschungsgemeinschaft and Fonds der Chemischen Industrie.

Present addresses: +Institut für Mikrobiologie, Theodor-Stern-Kai, D-6000 Frankfurt-M and §Institut für Molekularbiologie und Tumorforschung, Philipps-University, Emil-Mannkopff-Str.1, D-3550 Marburg, FRG

*To whom correspondence should be addressed

REFERENCES

1. Chamberlin, M.L. (1976) in: Losick, R. , Chamberlin, M.L. (eds.) RNA polymerase. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., pp. 17-67.
2. Woese, C.R. (1982) Zbl. Bakt. Hyg. I. Abt. Orig. C 3, 1-17.
3. Zillig, W., Schnabel, R., Thomm, M., Reiter, W.D. (1985) in: Schleifer, K.H., Stackebrandt, E. (eds.) The Evolution of the Prokaryotes, Academic Press, New York, pp. 45-82.
4. Schnabel, R., Thomm, M., Gerargy-Schahn, R., Zillig, W., Stetter, K.O., Huet, J. (1983) EMBO J. **2**, 751-755.
5. Schallenberg, J., Moes, M., Truss, M., Reiser, W., Thomm, M., Stetter, K.O., Klein, A. (1988) J. Bacteriol. **170**, 2247-2253.
6. Vieira, J., Messing, J. (1982) Gene **19**, 259-268.
7. Hanahan, D. (1985) in: Glover, D. M. (ed.) DNA cloning vol.I, IRL Press, Oxford, pp. 109-135.
8. Konheiser, U., Pasti, G., Bollschweiler, C., Klein, A. (1984) Mol. Gen. Genet. **198**, 146-152.
9. Maniatis, T., Fritsch, E.F., Sambrook, J. (1982) Molecular cloning: A Laboratory Manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
10. Strebel, K., Beck, E., Strohmaier, K., Schaller, H. (1986) J. Virol. 57, 983-991.
11. Marsh, J.L., Erfle, M., Wykes, E.J. (1984) Gene **32**, 481-485.
12. Maxam, A.M., Gilbert T.W. (1980) Methods Enzymol. **65**, 499-559.
13. Guo, L.-H., Wu, R. (1983) Methods Enzymol. **100**, 60-96.
14. Stetter, K.O., Winter, J., Hartlieb, R. (1980) Zbl. Bakt. Hyg. I. Abt. Orig. C **1**, 201-214.
15. Klein, A. (1988) Forum Mikrobiologie **11**, 82-86.
16. Shine, J., Dalgarno, L. (1974) Proc. Natl. Acad. Sci. USA **71**, 1342-1346.
17. Biggs, J., Searles, L.L., Greenleaf, A.L. (1985) Cell **42**, 611-621.
18. Allison, L.A., Moyle, M., Shales, M., Ingles, C.J. (1985) Cell **42**, 599-610.
19. Berg, J.M. (1986) Science **232**, 485-487.
20. Grachev, M.A., Lukhtanov, E.A., Mustaev, A.A., Richter, V.A., Rabinov, I.V., Skoblov, Y.S., Abdukayumov, M.N. (1987) Bioorg. Khim. **13**, 552-555.
21. Thomm, M., Lindner, A.J., Hartmann, G.R., Stetter, K.O. (1988) Syst. Appl. Microbiol. **10**, 101-105.
22. Ovchinnikov, Y.A., Monastyrskaya, G.S., Gubanov, V.V., Guryev, S.O., Chertov, O.Y., Modjanov, N.N., Grinkevich, V.A., Makarkova, I.A., Marchenko, T.V., Polovnikova, I.N.,

Lipkin, V.M., Sverdlov, E.D. (1981) Eur. J. Biochem. **116**, 621-629.
23. Sweetser, D., Nonet, M., Young, R.A. (1987) Proc. Natl. Acad. Sci. USA **84**, 1192-1196.
24. Falkenburg, D., Dworniczak, B., Faust, D.M., Bautz, E.K.F. (1987) J. Mol. Biol. **195**, 929-937.
25. Ovchinnikov, Y.A., Monastyrskaya, G.S., Buganov, V.V., Guryev, S.O., Salomatina, I.S., Shuvaeva, T.M., Lipkin, V.M., Sverdlov, E.D. (1982) Nucleic Acids Res. **10**, 4035-4044.
26. Souillard, N., Sibold, L. (1986) Mol. Gen. Genet. **203**, 21-28.
27. Bokranz, M., Klein, A. (1987). Nucleic Acids Res. **15**, 4350-4351.
28. Matzura, H. (1980) Curr. Top. Cell. Regulation 17, 89-136.
29. Panka, D., Dennis, D. (1985) J. Biol. Chem. **260**, 1427-1431.
30. Creighton, T.E. (1984) in: Proteins, W.H. Freeman and Company, New York, p. 431.