
Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

ELSEVIER

Biochimica et Biophysica Acta 1434 (1999) 221–247

BIOCHIMICA ET BIOPHYSICA ACTA
BBA

www.elsevier.com/locate/bba

# Similarity relations of DNA and RNA polymerases investigated by the principal component analysis of amino acid sequences

Jinya Otsuka *, Norihiro Kikuchi, Shinji Kojima

*Department of Applied Biological Science, Faculty of Science and Technology, Science University of Tokyo, Noda 278, Japan*

Received 26 May 1999; accepted 12 August 1999

**Abstract**

The principal component analysis based on the physicochemical properties of amino acid residues is applied to DNA and RNA polymerases to assign the sequence motifs for the polymerization activities of these proteins. After the reconfirmation of the sequence motifs of families *A* and *B* of DNA polymerases indicated previously, it elucidates the sequence motifs for the polymerization activity of DNA polymerase III (family *C*) by the similarity to the polymerization center of multimeric DNA dependent RNA polymerases. This identification proceeds to clarify the sequence motifs for polymerization activities of primases; eukaryotic and archaebacterial primases carry motifs similar to those of family *C*, while the motifs of eubacterial primase fall into the category of the motifs in family *B* DNA polymerases such as α, δ, ε and II. This finding means that DNA dependent RNA polymerases are also divided into groups corresponding to three families, *A*, *B* and *C*, because the monomeric DNA dependent RNA polymerases in phages are reconfirmed to carry sequence motifs similar to those of family *A* DNA polymerases. Furthermore, the three families of polymerization motifs are found to fall within the variation range of polymerization motifs displayed by many RNA dependent RNA polymerases, suggesting a close evolutionary relation between them. The sequence motifs for polymerization activities of reverse transcriptase and telomerase seem to be the intermediate between family *A* DNA polymerase and some RNA dependent RNA polymerases, e.g., from Leviviridae. On the contrary, the sequence fragments similar to the nucleotidyltransferase superfamily including DNA polymerase β are not found in any RNA dependent RNA polymerase, suggesting their other lineage of polymerization motifs. 

*Keywords:* DNA polymerase; RNA polymerase; Primase; Reverse transcriptase; Nucleotidyltransferase; Principal component analysis; Sequence motif

## 1. Introduction

Since the three-dimensional structure of the Klenow fragment of *Escherichia coli* DNA polymerase I has been determined by X-ray diffraction analysis [1], much attention has been paid to reveal the sequence characteristics responsible for polymerization activity as well as those for exonuclease activity. The simplest way for this purpose is to find conserved amino acid residues in the structure by the homologous alignment of polypeptide chains exhibiting a similar function. In the early trials along this line, sequence motifs similar to those in the Klenow fragment were found in the DNA dependent DNA polymerases from bacteriophages T7 [2,3] and T5 [4], as well as in DNA polymerases I from *Thermus aquaticus* [5]

* Corresponding author. Fax: +81 (471) 239767;
E-mail: jotsuka@rs.noda.sut.ac.jp

0167-4838/99/$ – see front matter 
PII: S0167-4838(99)00187-9

and *Streptococcus pneumoniae* [6], leading to the proposal of the sequence motifs for polymerization activity of the first family of polymerases. As the second family, human DNA polymerase α has been indicated to carry sequence fragments similar to those in viral DNA dependent DNA polymerases [7]. The sequence similarity between DNA polymerase β and terminal transferase has also been proposed to constitute another family [8,9]. Unification of the sequence fragments responsible for polymerization activities of these families of DNA polymerases and some RNA dependent RNA polymerases was also attempted [10], but this attempt only focused on the presence of aspartic acid residues in the two motifs *A* and *C*. In parallel, the sequence motifs for 3′-5′ exonuclease activity of *E. coli* DNA polymerase I have also been characterized by comparison with other amino acid sequences exhibiting similar exonuclease activity [11], and the conserved regions for 5′-3′ exonuclease activity of *E. coli* DNA polymerase I are suggested from comparison with the amino acid sequences of DNA polymerases I from other species of eubacteria [12].

Besides the above experimental approach, it has been proposed to classify the DNA polymerases into at least four classes or families, *A*, *B*, *C* and *X*, according to the similarities in amino acid sequences [13,14]; families *A*, *B* and *C* are named for their similarities to the products of *polA*, *polB* and *polC* in *E. coli*, respectively, and the remaining family *X* (= *D*, *E*, ...) is also set to accommodate the other DNA polymerases such as eukaryotic DNA polymerase β and new polymerases to be discovered in the future. However, this classification is mainly based on the sequence similarity scores evaluated with the use of the FASTA program developed by Pearson and Lipman [15,16]. Recently, the determination of tertiary structures is extended to T7 bacteriophage DNA dependent RNA polymerase [17] and HIV-1 reverse transcriptase [18], suggesting their structural similarity to the polymerization domain of DNA polymerase I, although a considerable similarity score is not counted between them by the FASTA program. In practice, the FASTA program is not sensitive to detect conserved sequence fragments such as the sequence motifs for polymerization activity, which are distributed with different intervals in the primary structures of different types of polymerases.

In the present paper, the more elaborate method of similarity search on the basis of the principal component analysis [19,20] is applied to detect similar regions between different types of polymerases. The numerical representation of sequence pattern by the principal component makes it easy to assign similar regions between different types of proteins, even if these regions are located with different intervals between the compared proteins. The application of this method not only reconfirms the previously indicated sequence motifs for polymerization activities of families *A* and *B* of DNA polymerases but also assigns the sequence motifs for polymerization activity of DNA polymerase III by the similarity to the polymerization center of multimeric DNA dependent RNA polymerases. This assignment resolves the problem of an apparent difference between eukaryotic and eubacterial primases; the sequence fragments similar to polymerization motifs of family *C* are found in the smallest subunit of eukaryotic primase while the eubacterial primase carries the sequence fragments similar to family *B* polymerization motifs. The similarity of sequence motifs for polymerization activities of DNA and RNA polymerases is also reconfirmed between family *A* DNA polymerase and monomeric DNA dependent RNA polymerase in bacteriophage.

## 2. Method

The present investigation is carried out for the amino acid sequence data of polymerases stored in the databases Swiss Prot release 35 and GenBank release 101.0.

### 2.1. Homologous alignment of the same type of polymerases

The homologous alignment of the amino acid sequences of polymerases of the same name is carried out using the multiple alignment program developed by Thompson et al. [21]. The polymerases from different organisms, but called by the same name in biochemical studies, mostly carry similar amino

acid sequences, and their polypeptide chains are easily aligned homologously in the whole region.

### 2.2. Construction of $z^{(1)}$ diagram: numerical representation of homologous amino acid sequences by the first principal component

The principal component analysis of polypeptide fragments on the basis of the physicochemical properties of constituent amino acid residues and its application to the numerical representation of amino acid sequences aligned homologously are already described in detail [19,20]. Thus, we will denote only the numerical data used in the present principal component analysis and the first principal component obtained from these data.

In the analysis of polypeptide fragments, four variables, (1) polarity, (2) hydrophobicity, (3) volume and (4) p$K_a$, are used to represent the physicochemical properties of each amino acid. The values of these properties are listed for every kind of amino acid in Table 1. The samples, for which the principal component analysis is carried out, are chosen to be polypeptide fragments, each consisting of five amino acid residues. Although $20^5$ kinds of polypeptide fragments are generally considerable in this size of polypeptide fragments, it is sufficient for the present purpose to choose a data set of 36 205 polypeptide fragments that are collected from the amino acid sequences of 45 representative polymerases including four families *A*, *B*, *C* and *X* of DNA polymerases, multimeric and monomeric DNA dependent RNA polymerases, RNA dependent RNA polymerases and reverse transcriptases. Each of the polypeptide fragments thus obtained is numerically characterized by the four kinds of physicochemical properties, each property being taken as an average value over the five constituent amino acid residues. The correlation matrix of the four variables calculated for this data set, $\{x_{ij}\}$ ($i=1, 2, 3, 4$; $j=1, 2, 3, 4$), is shown in Table 2. An eigenvalue problem is then solved for this matrix, and the eigenvalues and percentages of inertia thus obtained are listed in Table 3. Because the percentage of inertia corresponding to the first principal component is 52.7%, a main feature of variation in the samples of used polypeptide fragments may be approximately represented by the first principal component. By this procedure, the coefficients, with which the first principal component is expressed by a linear combination of standardized variables, are calculated to be $a^{(1)}=0.6193$, $a_2^{(1)}=-0.6680$, $a_3^{(1)}=-3.6148$ and $a_4^{(1)}=0.1989$. With the use of these values of coefficients, the first principal component $z_k^{(1)}$ of a polypeptide fragment $k$ is represented as

$$z_k^{(1)} = a_1^{(1)}\left(\frac{x_{k1}-\overline{x_1}}{\sqrt{\sigma_{11}}}\right) + a_2^{(1)}\left(\frac{x_{k2}-\overline{x_2}}{\sqrt{\sigma_{22}}}\right) + a_3^{(1)}\left(\frac{x_{k3}-\overline{x_3}}{\sqrt{\sigma_{33}}}\right) + a_4^{(1)}\left(\frac{x_{k4}-\overline{x_4}}{\sqrt{\sigma_{44}}}\right) \quad (1)$$

Table 1
Numerical values of physicochemical properties for each amino acid residue used in the principal component analysis

| Amino acid | Polarity[a] | Hydrophobicity[b] | Volume[a] | p$K_a^c$ |
|---|---|---|---|---|
| D | 13.0 | 0.0 | 54 | 3.65 |
| N | 11.6 | 0.0 | 56 | 0.00 |
| E | 12.3 | 0.0 | 83 | 4.25 |
| Q | 10.5 | 0.0 | 85 | 0.00 |
| K | 11.3 | 0.0 | 119 | 10.53 |
| R | 10.5 | 0.0 | 124 | 12.48 |
| H | 10.4 | 0.5 | 96 | 6.00 |
| S | 9.2 | 0.0 | 32 | 0.00 |
| T | 8.6 | 0.4 | 61 | 0.00 |
| P | 8.0 | 0.0 | 32.5 | 0.00 |
| A | 8.1 | 0.5 | 31 | 0.00 |
| G | 9.0 | 0.0 | 3 | 0.00 |
| Y | 6.2 | 2.3 | 136 | 10.07 |
| W | 5.4 | 3.4 | 170 | 0.00 |
| C | 5.5 | 0.0 | 55 | 10.28 |
| V | 5.9 | 1.5 | 84 | 0.00 |
| L | 4.9 | 1.8 | 111 | 0.00 |
| I | 5.2 | 1.8 | 111 | 0.00 |
| F | 5.2 | 2.5 | 132 | 0.00 |
| M | 5.7 | 1.3 | 105 | 0.00 |

[a]Cited from Grantham [22].
[b]Cited from Nozaki and Tanford [23].
[c]Cited from Barker [24].

Table 2
Correlation matrix of four variables used in the principal component analysis

| | Polarity | Hydrophobicity | Volume | p$K_a$ |
|---|---|---|---|---|
| Polarity | 1.0000 | | | |
| Hydrophobicity | −0.8001 | 1.0000 | | |
| Volume | −0.2115 | 0.5585 | 1.0000 | |
| p$K_a$ | 0.3833 | −0.2199 | 0.4943 | 1.0000 |

Table 3
Eigenvalues and percentages of inertia corresponding to the four principal components

| | First | Second | Third | Fourth |
|---|---|---|---|---|
| Eigenvalue | 2.109 | 1.531 | 0.274 | 0.087 |
| Inertia (%) | 52.700 | 38.300 | 6.840 | 2.160 |

when the average polarity, hydrophobicity, volume and p$K_a$ values of the five amino acid residues constituting the fragment are inserted into the four variables $x_{k1}$, $x_{k2}$, $x_{k3}$ and $x_{k4}$, respectively, in this equation. Here, the mean value and variance of each variable in the data set are: $\overline{x_1} = 8.41$, $\sigma_{11} = 1.18$, $\overline{x_2} = 0.75$, $\sigma_{22} = 0.14$, $\overline{x_3} = 81.88$, $\sigma_{33} = 283.53$, $\overline{x_4} = 2.66$ and $\sigma_{44} = 3.69$. Approximately, this representation gives a higher positive value of $z^{(1)}$ for the sequence fragment of hydrophilic amino acid residues and a negative value for hydrophobic residues. However, the most important advantage of this representation should be that it reflects different physicochemical properties and thus can predict active centers on the basis of more detailed structural properties than the hydropathy analysis. Such evaluation is made progressively for the successive polypeptide fragments of five sites overlapping by four from the N to the C terminus along a polypeptide chain. For homologously aligned sequences, $z^{(1)}$ values are each evaluated for each of the polypeptide fragments aligned homologously, and the standard deviation width around the average value of them is plotted against the middle sites of the respective fragments along the polypeptide chain. The narrower width of standard deviation means a higher degree of conservation of amino acid residues.

### 2.3. Comparison of $z^{(1)}$ diagrams between different types of polymerases

In most cases, the finding of similar regions between different types of polymerases is possible by the visual comparison of diagrams but the more quantitative method using Kendall's rank correlation coefficient [25] is also applied, especially in the case where two or more candidates for homologous regions are found between different types of polymerases. For the assignment of functionally important regions such as those for polymerization and exonuclease activities, we must find the regions showing a relatively narrow width of standard deviation as well as a similar sequence pattern.

### 2.4. Assignment of sequence motifs in similar regions between different types of polymerases

Each candidate for similar regions is then examined by enumerating identical, and/or the similar properties of, amino acid residues between the compared regions. For this purpose, the amino acids are classified into the following six categories according to their biochemical properties: (1) ambivalent: Gly, Ala, Ser, Thr, Cys and Pro; (2) hydrophobic but not aromatic: Val, Leu, Ile and Met; (3) hydrophobic and carrying an aromatic ring: Phe, Trp and Tyr; (4) hydrophilic and basic: Lys, Arg and His; (5) ambivalent but carrying an amidocarboxyl group as the side chain: Asn and Gln; and (6) hydrophilic and acidic: Asp and Glu. Although such a classification of amino acids is already proposed to identify the homologous sites among cytochromes [26], some modifications are made in the above categories; Gly is incorporated into category 1 together with the other amino acids carrying small hydrophobic or polar side chains, and Asn and Gln are clustered into category 5 separately from other ambivalent amino acids. This is because Asn and Gln carry a side chain stereochemically similar to the carboxyl group of Asp which is known to play an important role in suspending magnesium, zinc and/or manganese ions in the active centers of polymerases and exonucleases.

## 3. Results

### 3.1. DNA polymerase III (family C) and multimeric DNA dependent RNA polymerase

Although DNA polymerase III is considered to play the central role in replicating DNAs in eubacteria [27], the sequence motifs for polymerization activity of this type of DNA polymerase are not well clarified yet, probably because of the difficulty from the approach by determining the crystal structure. In practice, this type of DNA polymerase consists of many subunits. For example, the DNA polymerase

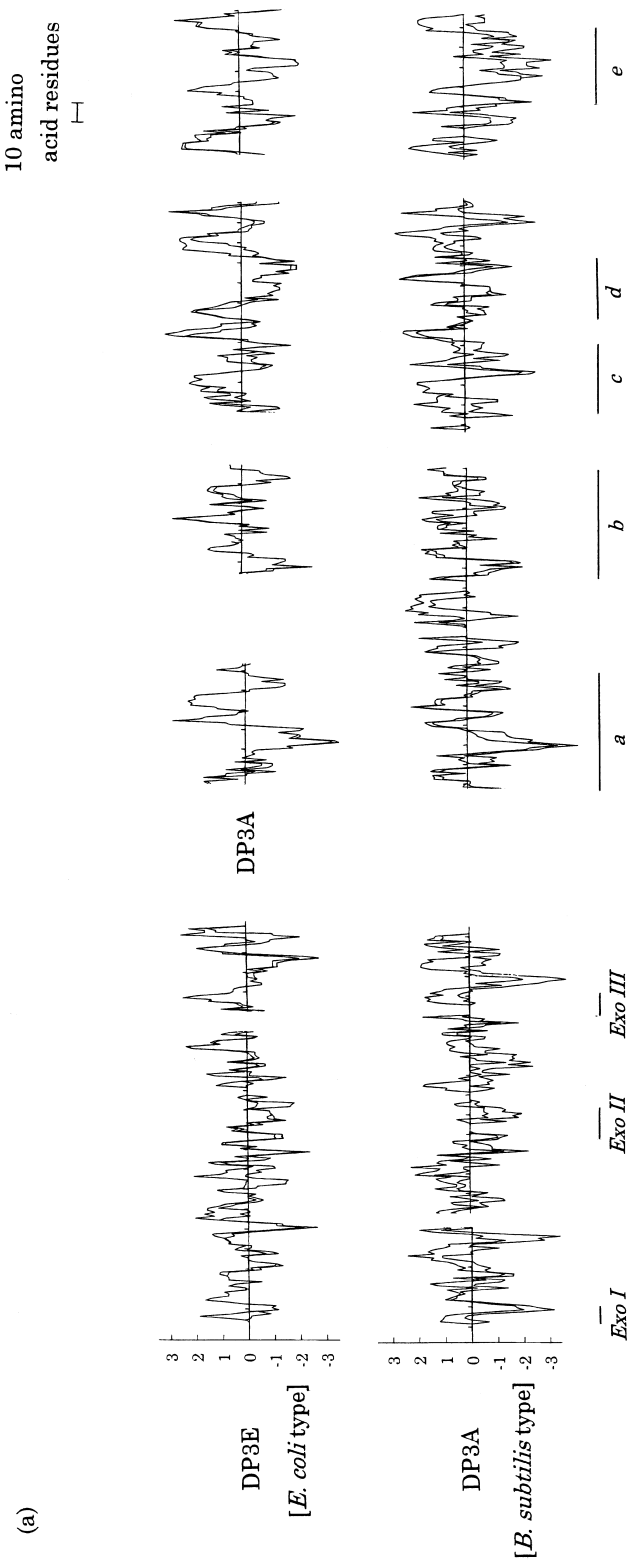

(a)
DP3E
[E. coli type]
DP3A
[B. subtilis type]
10 amino
acid residues
DP3A
Exo I
Exo II
Exo III
a
b
c
d
e

Fig. 1.

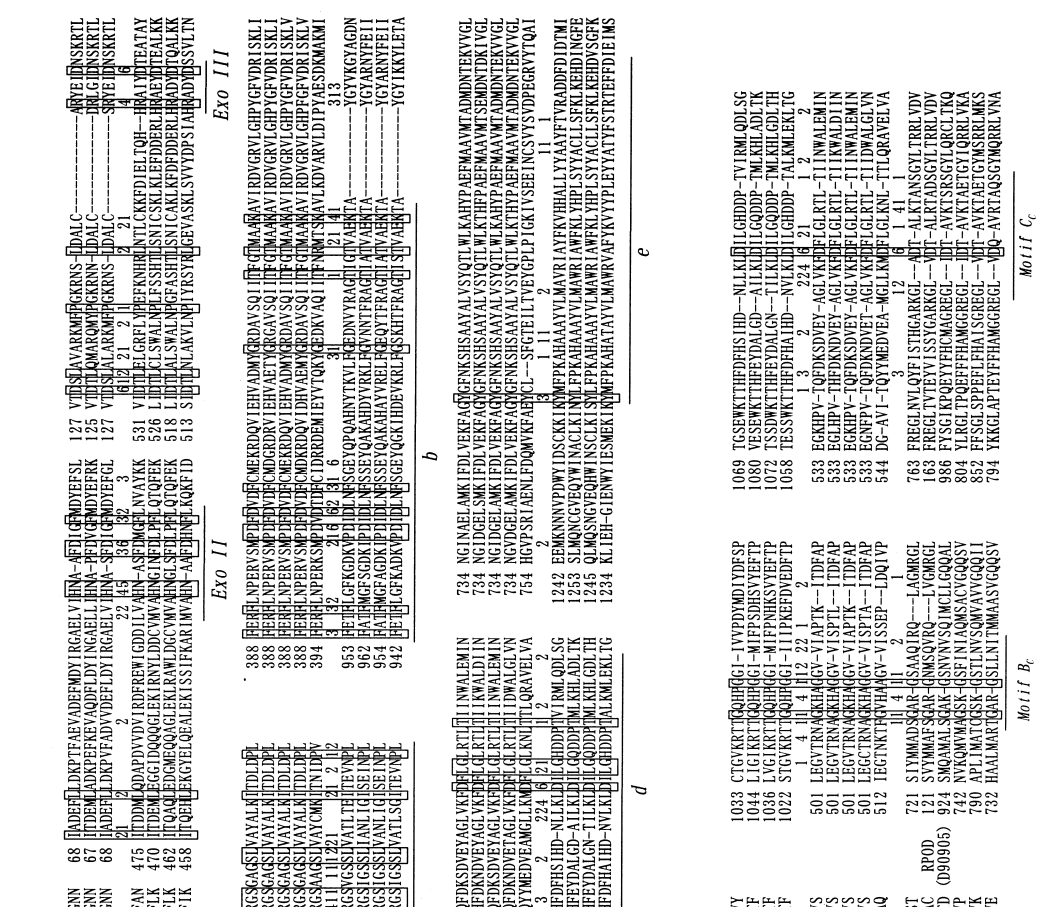


Fig. 1 (*continued*).

III in *E. coli* consists of ten kinds of subunits, α, β, δ, δ′, ε, θ, τ, γ, χ and ψ. In this form of DNA polymerase III, the polymerase activity and 3′-5′ exonuclease activity are attributed to the α and ε subunits, respectively, while both activities are attributed to the α subunit of DNA polymerase III in *Bacillus subtilis* [27]. As far as the available sequence data of DNA polymerases III are concerned, the DNA polymer-

Fig. 1. Homologous alignment of amino acid sequences of ε (DP3E) and α (DP3A) subunits of *E. coli*-type DNA polymerases III to the sequences of α subunits of *B. subtilis*-type DNA polymerases III. (a) Comparison by $z^{(1)}$ diagrams. (b) Comparison by individual amino acid residues. Amino acid residues conserved in both the ε subunit of the *E. coli* type and the α subunit of the *B. subtilis* type correspond to the sequence motifs similar to those for 3′-5′ exonuclease activity identified in families *A* and *B* of DNA polymerases [11,28–30]. The other five regions, each of which contains the amino acid residues highly conserved in the α subunits of both types of DNA polymerases III, are denoted *a*, *b*, *c*, *d* and *e*, respectively. Identical and similar amino acid residues between the two types of DNA polymerases III are boxed and denoted by the category numbers, respectively. (c) Proposal of the sequence motifs for polymerization activity of DNA polymerase III by the similarity to the conserved regions among the largest subunits of multimeric DNA dependent RNA polymerases; β′ subunit (RPOC) of *E. coli* RNA polymerase, γ (RPOG) and δ (RPOD) subunits of *Synechocystis* RNA polymerase, largest subunits (RPA1, RPB1 and RPC1) of eukaryotic RNA polymerases I, II and III, A′ (RPA′) and A″ (RPA″) subunits of archaebacterial RNA polymerase. As for the homologous alignment of these RNA polymerases, see Fig. 2. The proposed motifs for the polymerization activities of these polymerases are denoted as $A_C$, $B_C$ and $C_C$, according to the family name of DNA polymerase III. The organisms from which the compared amino acid sequences are derived are indicated by the following abbreviations: ECO, *E. coli*; HIN, *H. influenzae*; STY, *S. typhimurium*; BSU, *B. subtilis*; MGE, *M. genitalium*; MPN, *M. pneumoniae*; MPU, *M. pulmonis*; VCH, *Vibrio cholerae*; SYN, *Synechocystis* sp.; SCE, *S. cerevisiae*; MVA, *Methanococcus vannielii*. The accession number of each amino acid sequence in the GenBank database is denoted in parentheses after the abbreviation of the organism. The amino acid residue number in the N-terminal of each sequence fragment is denoted on the left side of each fragment.
←

ases III from *Haemophilus influenzae* and *Salmonella typhimurium* belong to the *E. coli* type while the DNA polymerases III from *Mycoplasma genitalium*, *Mycoplasma pneumoniae* and *Mycoplasma pulmonis* belong to the *B. subtilis* type. Recently, a third type of DNA polymerase III is found in cyanobacteria, but its available amino acid sequence of the α subunit from *Synechocystis* is highly similar to that of the α subunit of *E. coli*-type DNA polymerase III in sequence length as well as in individual amino acid residues. Thus, this sequence is also treated as *E. coli* type in the present study. The amino acid sequences of each type of subunit are aligned homologously, and the $z^{(1)}$ diagram is constructed. The comparison of the $z^{(1)}$ diagrams and of amino acid sequences ascertains the amino acid sequence fragments to be similar to those of Exo I, Exo II and Exo III identified in families *A* and *B* of DNA polymerases [11,28–30] in the ε subunit of *E. coli*-type DNA polymerase III and in the region of the α subunit of *B. subtilis*-type DNA polymerase III, which follows the phosphoesterase domain in the N-terminal region of the α subunit indicated previously [31]. These amino acid sequence fragments are D-ETTG in Exo I, HN---FD in Exo II and R---D in Exo III, and they are comparable with the amino acid residues conserved in family *A* DNA polymerases (I and γ) and family *B* DNA polymerases (II, δ, ε and archaebacterial DNA polymerase); D-E in Exo I of both family *A* and *B* DNA polymerases, N---D in Exo II of family *A*, N---FD in Exo II of family *B*, and Y-A-D in Exo III of family *A* and Y---D in Exo III of family *B*. Furthermore, at least five regions are found to carry amino acid residues highly conserved between *E. coli* and *B. subtilis* types of α subunits, and they are tentatively denoted by *a*, *b*, *c*, *d* and *e*. This is shown in Fig. 1a by the comparison of $z^{(1)}$ diagrams and in Fig. 1b by the homologous alignment of amino acid sequences.

Although the amino acid sequence fragments identical to those in motifs *A*, *B* and *C* of families *A* and *B* of DNA polymerases cannot be found in any of the regions *a*, *b*, *c*, *d* and *e*, the amino acid sequence fragments responsible for polymerase activity must be contained in these conserved regions. In practice, our systematic similarity search finds that these regions contain the amino acid sequence fragments conserved in the largest subunits of multimeric DNA dependent RNA polymerases. Thus, we will describe the amino acid sequence fragments conserved in the DNA dependent RNA polymerases before going into the assignment of the sequence fragments responsible for the polymerization activity of DNA polymerase III.

The three types of multimeric DNA dependent RNA polymerases, I, II and III, known in eukaryotes are similar to each other at least with respect to their largest subunit 1 and secondarily largest subunit 2. Moreover, the largest subunit 1 shows considerable similarity to the A′ plus A″ subunits of DNA dependent RNA polymerase in archaebacteria and the β′ subunit of DNA dependent RNA poly-

merase in eubacteria such as *B. subtilis*, *M. genitalium*, *Mycobacterium leprae*, *E. coli* and *Pseudomonas putida*. In the DNA dependent RNA polymerases from cyanobacteria (e.g., *Synechocystis* sp.), the two subunits, γ and δ, correspond to the β′ subunit, and such splitting of the largest subunit into two pieces is also seen in the DNA dependent RNA polymerase encoded in the chloroplast genome. These similarities to the largest subunits 1 are shown in Fig. 2a by comparing the $z^{(1)}$ diagrams of the respective types of polypeptide chains. As seen in this figure, at least five regions are found to show high similarities between subunit 1, A′ plus A′ subunits, β′ subunit, and γ plus δ subunits, and these regions will be tentatively denoted as *a*′, *b*′, *c*′, *d*′ and *e*′. The homologous alignment of amino acid sequences in each of these five regions is shown in Fig. 2b. Among the five regions, region *b*′ contains the amino acid sequence fragment NADFDGDQ/EM that is suggested to be the polymerization active center of *E. coli* DNA dependent RNA polymerase [32]. Such an Asp-rich sequence fragment is also present in region *b* of the α subunit of DNA polymerase III, as seen in Fig. 1b. Moreover, considerable similarities of amino acid sequences are also found between regions *c* and *d* of the α subunit of DNA polymerase III and regions *c*′ and *d*′ of RNA polymerases, although no marked similarity is found either between regions *a* and *a*′ or between regions *e* and *e*′. The direct comparison of the amino acid sequences in these regions is given in Fig. 1c, where the amino acid sequences of regions *b*, *c* and *d* of *E. coli*- and *B. subtilis*-type DNA polymerases III are aligned vertically relative to those of regions *b*′, *c*′ and *d*′ of RNA polymerases, respectively. In this comparison of amino acid sequences, it is found that the amino acid sequence fragments DFD-D are conserved in regions *b* and *b*′, G---G in regions *c* and *c*′, and L---D in regions *d* and *d*′. Although the third aspartic acid residue in region *b* of *B. subtilis*-type DNA polymerase III is replaced by an asparagine residue with a similar size of side chain, the presence of three aspartic acid residues in the first and third regions may be sufficient for forming the polymerization center suspending two metal ions, by analogy with the motifs indicated in families *A* and *B* of DNA polymerases. Thus, the amino acid sequence fragments in regions *b*, *c* and *d* will be proposed as the sequence motifs for the polymerization activity of the family *C* of DNA polymerases, being named as $A_C$, $B_C$ and $C_C$. In fact, the two aspartic acid residues, Asp-401 and Asp-403, in motif $A_C$ are also inspected to be the key residues in the active site for polymerization by the recent approach to the α subunit of DNA polymerase III from *E. coli* [33], and more recently these aspartic acid residues and Asp-555 in motif $C_C$ are indicated to chelate magnesium ions by site-directed mutagenesis [34]. Furthermore, the amino acid residues R/K--G-H-GG conserved in motif $B_C$ of two types of DNA polymerase III seem to be comparable to the residues K------(-)YG in motif *B* of families *A* and *B* of DNA polymerases. The sequence motifs for polymerization activity of DNA dependent RNA polymerases may also reside in regions *b*′, *c*′ and *d*′ which are vertically aligned with motifs $A_C$, $B_C$ and $C_C$ proposed for the polymerization activity of DNA polymerase III, although the residues conserved in the region corresponding to $B_C$ of DNA polymerase III are reduced to G–R/KG with the deletion of one site in DNA dependent RNA polymerases.

→

Fig. 2. Similarity between the largest subunits of DNA dependent RNA polymerases; eubacterial RNA polymerases (RPOC, RPOG+RPOD), eukaryotic RNA polymerases I (RPA1), II (RPB1) and III (RPC1), and archaebacterial RNA polymerase (RPA′+RPA″). (a) Vertical alignment of $z^{(1)}$ diagrams by similar sequence patterns. Because of the high sequence similarity of RPOG and RPOD with RPOC, their amino acid sequences are represented by one $z^{(1)}$ diagram. At least five regions, *a*′, *b*′, *c*′, *d*′ and *e*′, are identified to be similar between the five types of RNA polymerases. (b) Homologous alignment of amino acid sequences in the five regions. These regions contain a considerable number of amino acid residues conserved in all RNA polymerases, as denoted by boxes. Abbreviations for the organisms: MLE, *M. leprae*; PPU, *P. putida*; SYN, *Synechocystis* sp. (strain PCC 6803); SPO, *Schizosaccharomyces pombe*; TBB, *Trypanosoma brucei*; HSP, *Homo sapiens*; DME, *Drosophila melanogaster*; CEL, *Caenorhabditis elegans*; ATH, *Arabidopsis thaliana*; PFA, *Plasmodium falciparum*; GLA, *Giardia lamblia*; MTH, *M. thermoautotrophicum*; HHA, *Halobacterium halobium*; TCE, *Thermococcus celer*; SAC, *Sulfolobus acidocaldarius*; TAC, *Thermoplasma acidophilum*. The abbreviations for the other organisms are described in the legend to Fig. 1.

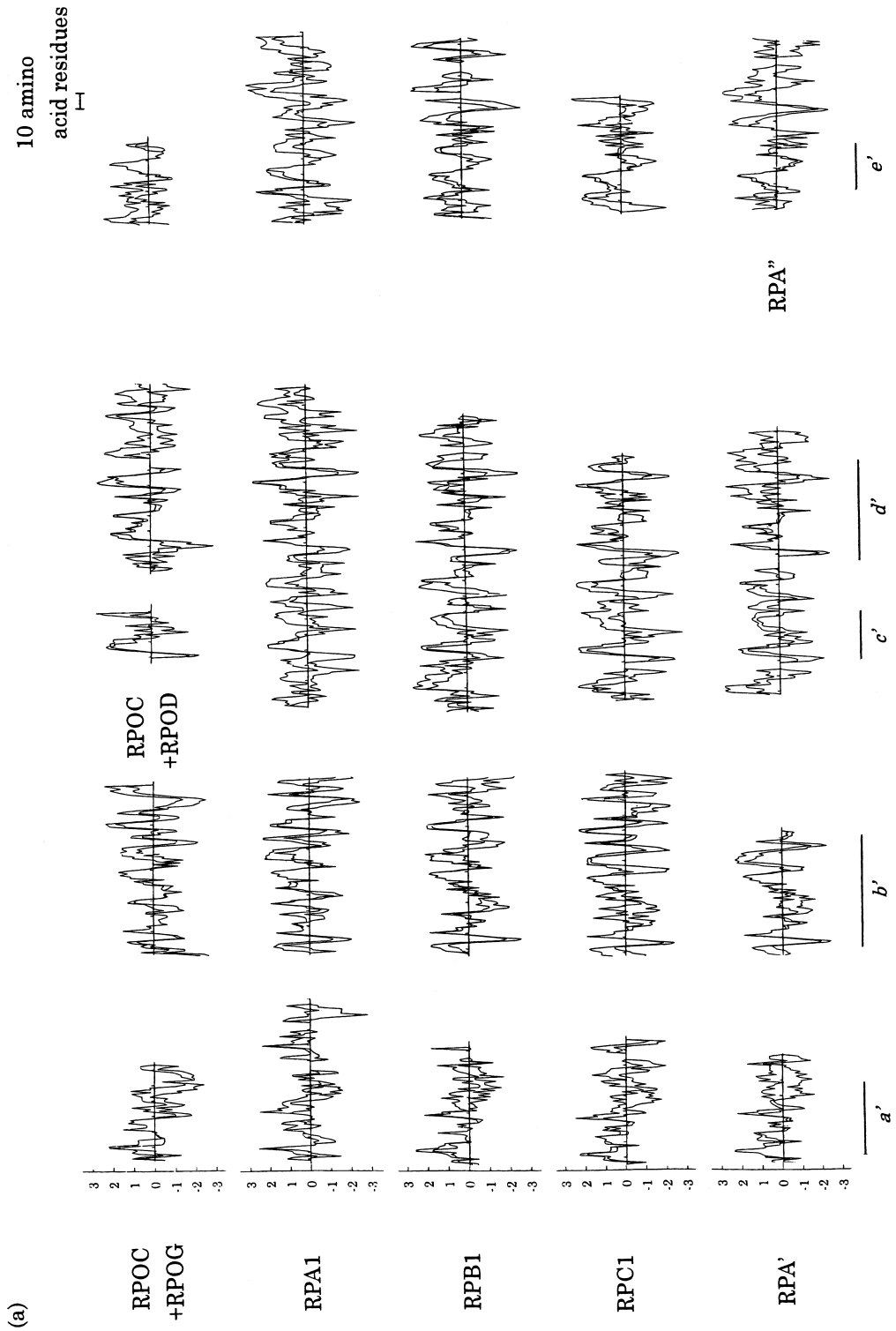

(a)
10 amino
acid residues
RPOC
+RPOG
RPA1
RPB1
RPC1
RPA'
RPOC
+RPOD
RPA"
a'
b'
c'
d'
e'

Fig. 2.

Fig. 2 (*continued*).

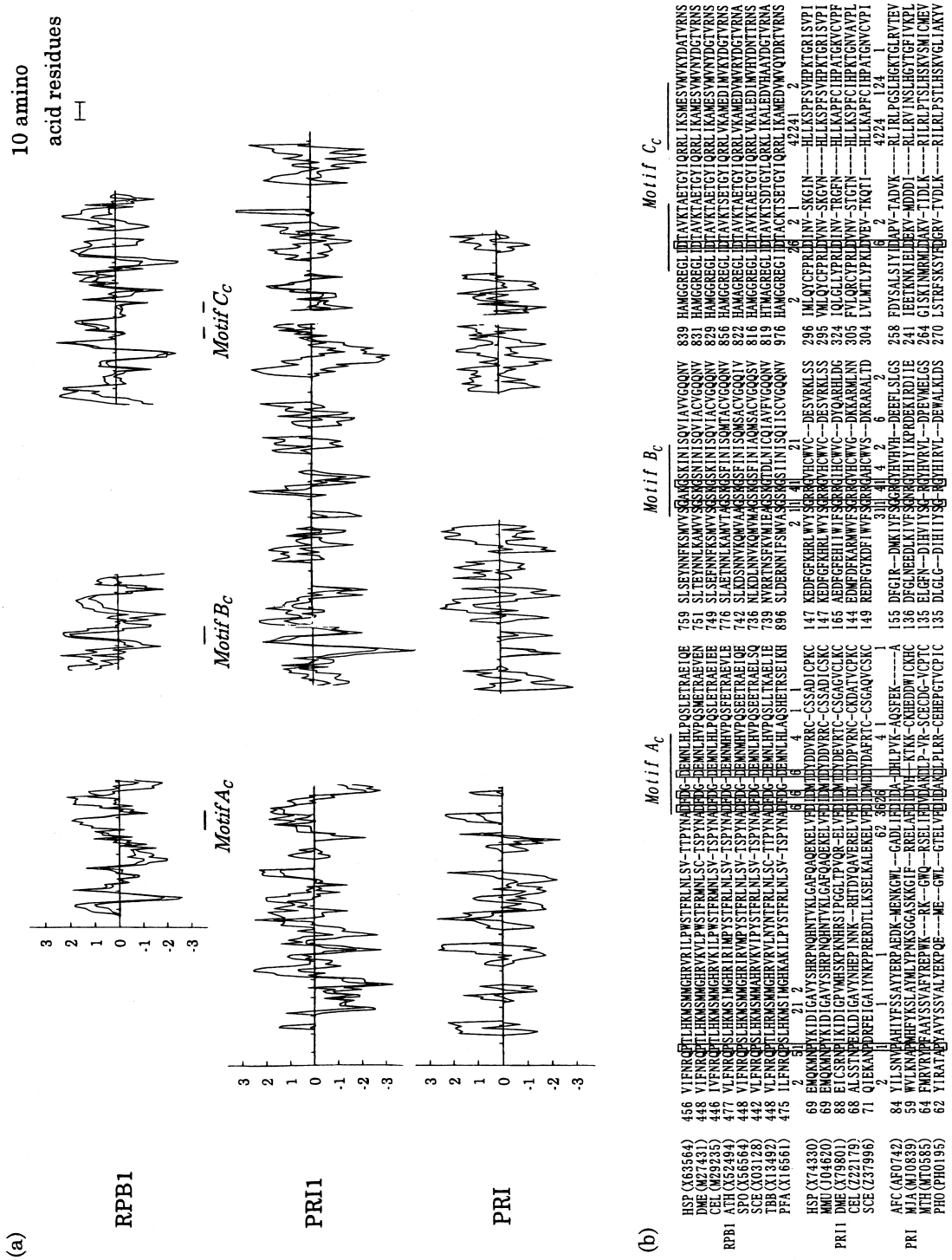



Fig. 3. Similarity of the smallest subunit of eukaryotic primase (PRI1) and an archaebacterial polypeptide chain (PRI) to the polymerization center of the largest subunit of DNA dependent RNA polymerase II (RPB1). (a) Comparison by $z^{(1)}$ diagrams. (b) Homologous alignment of amino acid sequences. The identical and similar amino acid residues between PRI1, PRI and RPB1 are concentrated in the regions corresponding to motifs $A_C$, $B_C$ and $C_C$, respectively, which are proposed for the polymerization activity of RNA polymerase II in Fig. 1. Abbreviations for the organisms: MMU, *Mus musculus*; AFC, *A. fulgidus*; MJA, *M. jannaschii*; PHO, *P. horikoshii* OT3. The abbreviations for the other organisms are already described in the legends to Figs. 1 and 2.

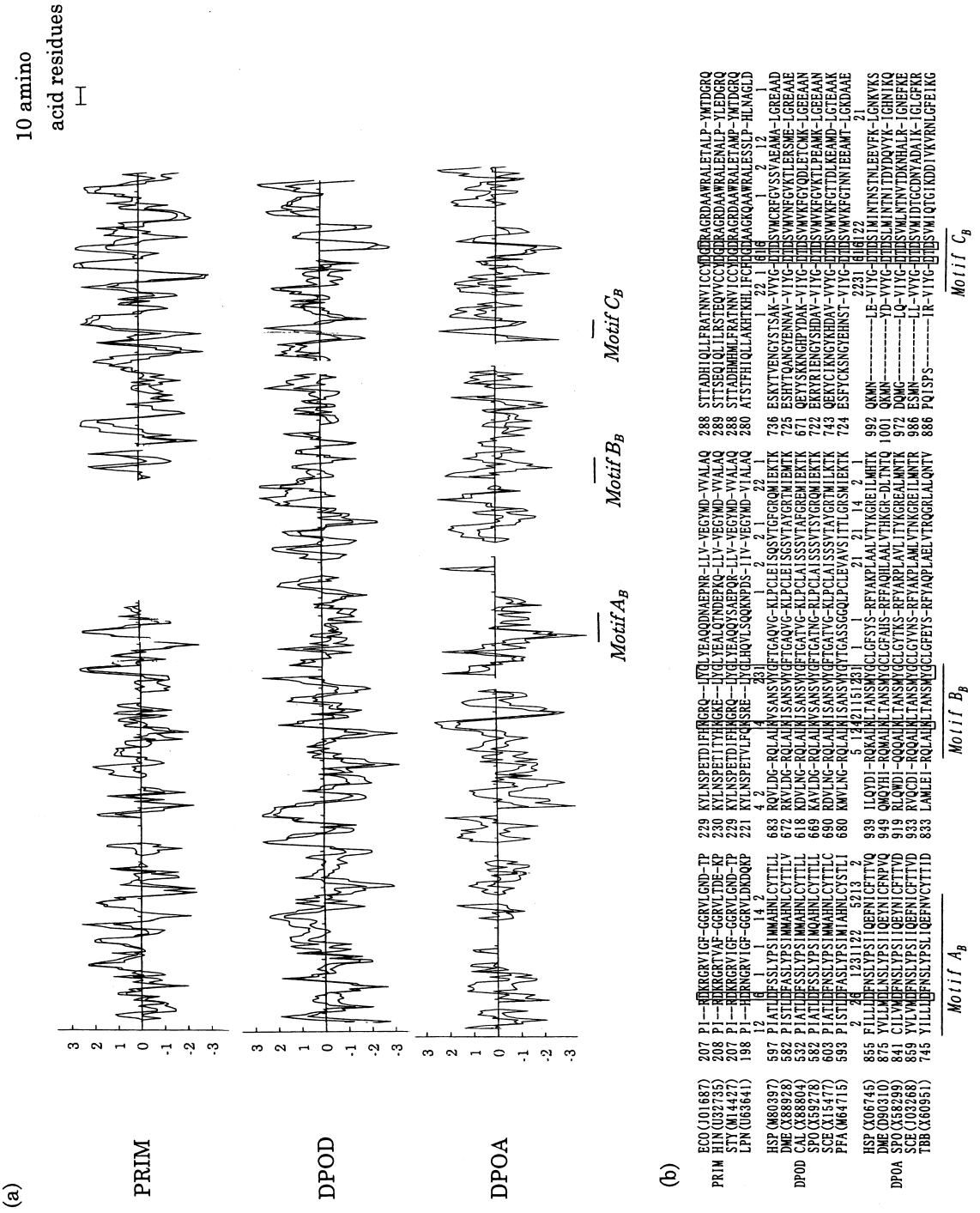



Fig. 4. Similarity of eubacterial primase (PRIM) to the polymerization domains of DNA polymerases δ (DPOD) and α (DPOA). (a) Comparison by $z^{(1)}$ diagrams. (b) Homologous alignment of amino acid sequences. The amino acid residues conserved in PRIM, DPOA and DPOD are boxed. These conserved residues contain D in motif $A_B$, K------YG in motif $B_B$ and D-D in motif $C_B$, which are characteristic of the sequence motifs for the polymerization activity of family *B* DNA polymerases [40–43]. Abbreviations for the sources: LPN, *Legionella pneumophila*; CAL, *C. albicans*. The abbreviations for the other organisms are already described in the legends to Figs. 1–3.

### 3.2. Primases in free living organisms

The assignment of the sequence motifs for the polymerization activity of family *C* DNA polymerase and multimeric DNA dependent RNA polymerase serves to resolve the problem why the eukaryotic primase is not similar to eubacterial primase in its amino acid sequence as well as in the formation of acting complexes.

According to the biochemical studies of primases [27], *E. coli* primase seldom acts alone; most commonly it teams up with the multifunctional dnaB protein in the synthesis of primers to start the DNA chain, and assembling the dnaB protein-primase complex on a template, whether it is ssDNA coated with SSB or duplex DNA, requires additional 'prepriming' proteins, while the primase in eukaryote is extracted as two small subunits (approx. 50 and approx. 60 kDa) of Pol α, together with a large DNA polymerase subunit (approx. 180 kDa, or DNA polymerase α) and an approx. 70 kDa polypeptide chain with no catalytic activity. Recently, the polymerization activity of eukaryotic primase is ascribed to the smallest subunit (approx. 50 kDa) by an experiment on conditional and lethal mutations [35].

The amino acid sequences of the smallest subunits of eukaryotic primases now available from eight species are similar to each other, and the $z^{(1)}$ diagram constructed on the basis of their homologous alignment is compared with the $z^{(1)}$ diagrams of other polymerases. This similarity search found that the smallest subunit of eukaryotic primase contains regions similar to the sequence motifs for the polymerization activities of multimeric DNA dependent RNA polymerases as well as of family *C* DNA polymerases. The sequence patterns represented by $z^{(1)}$ diagrams are compared in Fig. 3a between the smallest subunits of eukaryotic primases and the largest subunits of DNA dependent RNA polymerase II, and the comparison of amino acid sequences in the similar regions is shown in Fig. 3b, indicating the same category of amino acid residues. As seen in Fig. 3b, the smallest subunit of eukaryotic primase not only contains the cluster of three aspartic acid residues characteristic of motif $A_C$ but also shows considerable similarities to motifs $B_C$ and $C_C$ in multimeric DNA dependent RNA polymerase II, and thus to those of RNA polymerases I and III.

Although the primase is not yet identified in archaebacteria, our preliminary similarity search by FASTA program for the genome databases of *Archaeoglobus fulgidus* [36], *Methanococcus jannaschii* [37], *Methanobacterium thermoautotrophicum* [38] and *Pyrococcus horikoshii* OT3 [39] found a polypeptide chain in each of these four archaebacteria that is similar to the smallest subunit of primase in eukaryote with an optimal score of 144–211. The amino acid sequences of the polypeptide chains from these archaebacteria are mutually similar, and are easily aligned homologously. The $z^{(1)}$ diagram constructed on the basis of this homologous alignment and the amino acid sequence of the region, which seems to be responsible for polymerization activity, are also shown in Fig. 3a and b, respectively. As seen in these figures, the polypeptide chains from archaebacteria also carry amino acid sequence fragments similar to those in motifs $A_C$, $B_C$ and $C_C$, although the amino acid residues conserved in motif $B_C$ are further reduced to G-RG.

The amino acid sequences of eubacterial primases are available from four species including *E. coli*, and they are so similar that they can be easily aligned homologously. The $z^{(1)}$ diagram of the first principal component constructed on the basis of the homologous alignment of eubacterial primases is then compared with the $z^{(1)}$ diagrams of other polymerases. This comparison found that the eubacterial primases carry sequence fragments similar to the sequence motifs for polymerization activity of family *B* DNA polymerases. To illustrate this, the sequence pattern and amino acid sequences of eubacterial primases are compared with those of the polymerization domains of DNA polymerases δ and α in Fig. 4a and b, respectively, where the three sequence motifs for polymerization activity identified already in family *B* DNA polymerases are denoted by $A_B$, $B_B$ and $C_B$. Among family *B* DNA polymerases, DNA polymerase α has been investigated experimentally with respect to the amino acid residues responsible for polymerization and its associated functions [40–43], and these residues are contained in the following set of amino acid residues conserved in DNA polymerase II and archaebacterial DNA polymerases including DP1 and DP2 identified in *Pyrodictium occultum* as

well as in DNA polymerases α and δ: D--SLYPS in motif $A_B$, K------YG in motif $B_B$ and DTD in motif $C_B$, although SLYPS in motif $A_B$ is replaced by S/AMYPN in DNA polymerase ε and YG in motif $B_B$ is replaced by GY in DNA polymerase ζ. The three regions of eubacterial primase, which are assigned to motifs $A_B$, $B_B$ and $C_B$ in the present study, also contain D, K----YG and D-D, respectively, although two sites between K and Y are vacant in the middle region of the primase in comparison with family *B* DNA polymerases.

### 3.3. Monomeric DNA dependent RNA polymerases

The DNA dependent RNA polymerases in viruses are of multiple subunits, and each of them contains the two subunits that are considerably similar to the largest and secondarily largest subunits of eukaryotic DNA dependent RNA polymerase. On the other hand, the DNA dependent RNA polymerases in bacteriophages are monomeric, and the tertiary structure of T7 bacteriophage RNA polymerase is indicated to be similar to that of the Klenow fragment by X-ray diffraction analysis [17], although no considerable similarity score is counted between T7 bacteriophage RNA polymerase and DNA polymerase I by the FASTA program. This type of bacteriophage RNA polymerase also shows high similarity to the monomeric DNA dependent RNA polymerase for transcribing the genes encoded in the mitochondrial DNA (e.g., an optimal score of 708 between bacteriophage T3 RNA polymerase and yeast mitochondrial RNA polymerase). In practice, the homologous alignment of phage RNA polymerase and mitochondrial RNA polymerase is already carried out with the indication of 11 conserved domains, I–XI [44]. For the reconfirmation of these previous indications, $z^{(1)}$ diagrams are constructed for four species of bacteriophages RNA polymerases, two species of mitochondrial RNA polymerases and seven species of DNA polymerase I. These diagrams are compared in Fig. 5a, and the homologous alignment of their amino acid sequences in the region which seems to be responsible for the polymerization activity is shown in Fig. 5b. In these figures, the sequence motifs proposed for the polymerization activity of DNA polymerase I [2–6,10] are denoted by $A_A$, $B_A$ and $C_A$. Among the amino acid sequence fragments in the three motifs, D----E in motif $A_A$, K-------YG in motif $B_A$ and HDE in motif $C_A$ are also conserved in DNA polymerase γ. As seen in Fig. 5b, the monomeric DNA dependent RNA polymerases share the residue characteristic of the motifs for polymerization activity, i.e., the D in motif $A_A$, K-------YG in motif $B_A$ and HD in motif $C_A$, with family *A* DNA polymerases, although residue E in motif $A_A$ is not conserved and residues HDE in motif $C_A$ are replaced by HDS in monomeric RNA polymerase.

### 3.4. RNA dependent RNA polymerases

Most of the RNA dependent RNA polymerases in viruses and bacteriophages function in the form of multisubunits, but the polymerase activity is attributed to a single polypeptide chain in most cases. Such polypeptide chains are collected from databases (Swiss Prot release 35 and GenBank release 101.0) and then classified into several groups by the criterion of the optimal similarity scores of more than 200 evaluated with the FASTA program. The polypeptide chains clustered into the same group are mostly those from the sources in the same taxonomic category, corresponding to a 'family' defined by the international committee on the taxonomy of viruses [45], and their amino acid sequences are relatively easily aligned homologously except for those in var-

→

Fig. 5. Homology of bacteriophage RNA polymerase (RPOL) and mitochondrial RNA polymerase (RPOM), and their similarity to the polymerization domain of family *A* of DNA polymerase I (DPOI). (a) Comparison by $z^{(1)}$ diagrams. (b) Homologous alignment of amino acid sequences. The amino acid residues conserved in RPOL and RPOM are boxed, and some of them are also conserved in DPOI, consistent with the sequence fragments D, K-------YG and HD in the motifs $A_A$, $B_A$ and $C_A$, respectively, identified for the polymerization activity of family *A* of DNA dependent DNA polymerases [2–6,10]. Abbreviations for the organisms: B11, bacteriophage K11; SP6, bacteriophage SP6; BT3, bacteriophage T3; BT7, bacteriophage T7; NCR, *Neurospora crassa*; BCA, *Bacillus caldotenax*; DRA, *Deinococcus radiodurans*; MTU, *Mycobacterium tuberculosis*; SPN, *S. pneumoniae*; TAQ, *T. aquaticus*. Abbreviations for the other organisms are already described in the legend to Fig. 1.

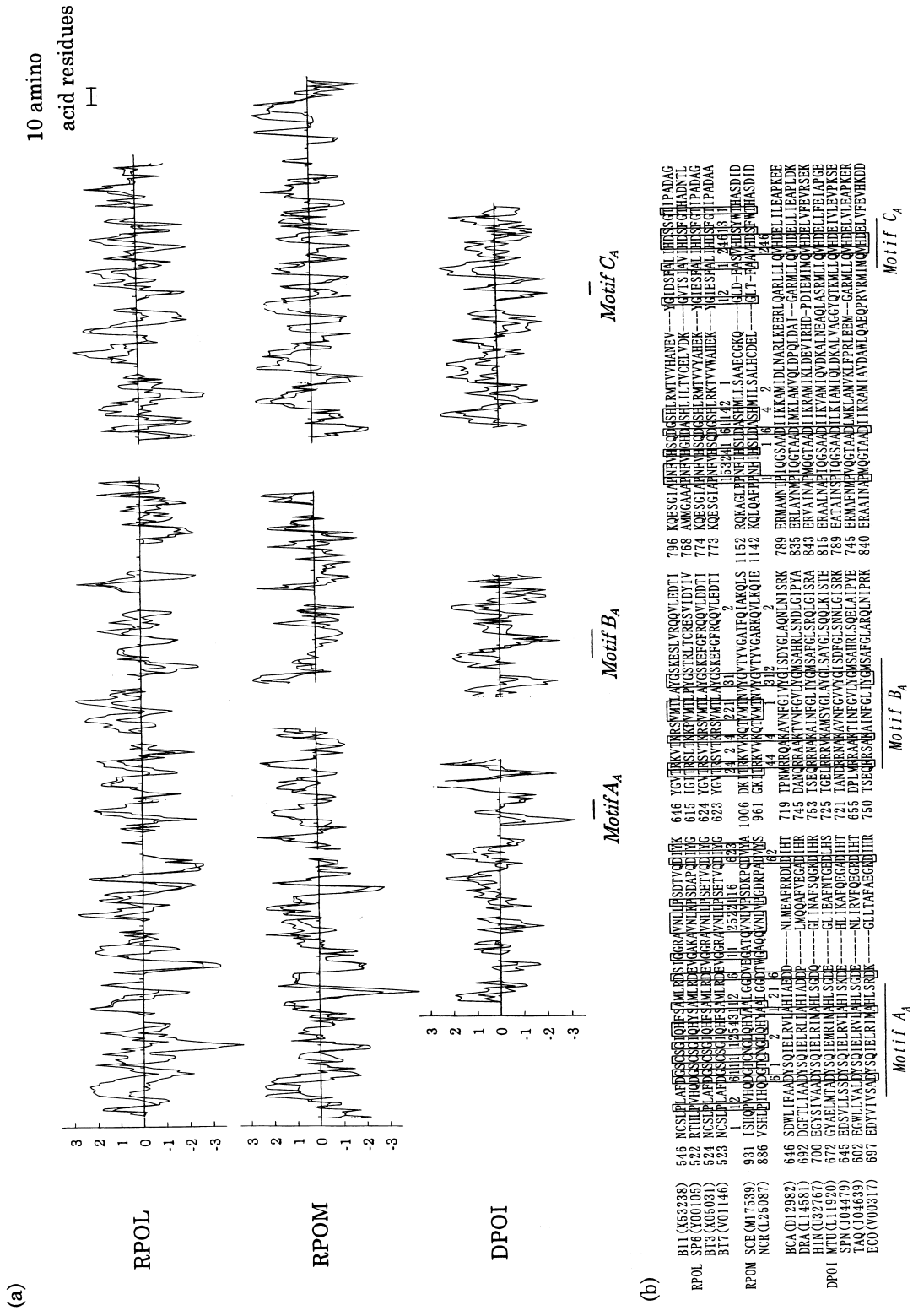

(a)
10 amino acid residues
RPOL
RPOM
DPOI
Motif $A_A$
Motif $B_A$
Motif $C_A$
(b)
RPOL
RPOM
DPOI

Fig. 5.

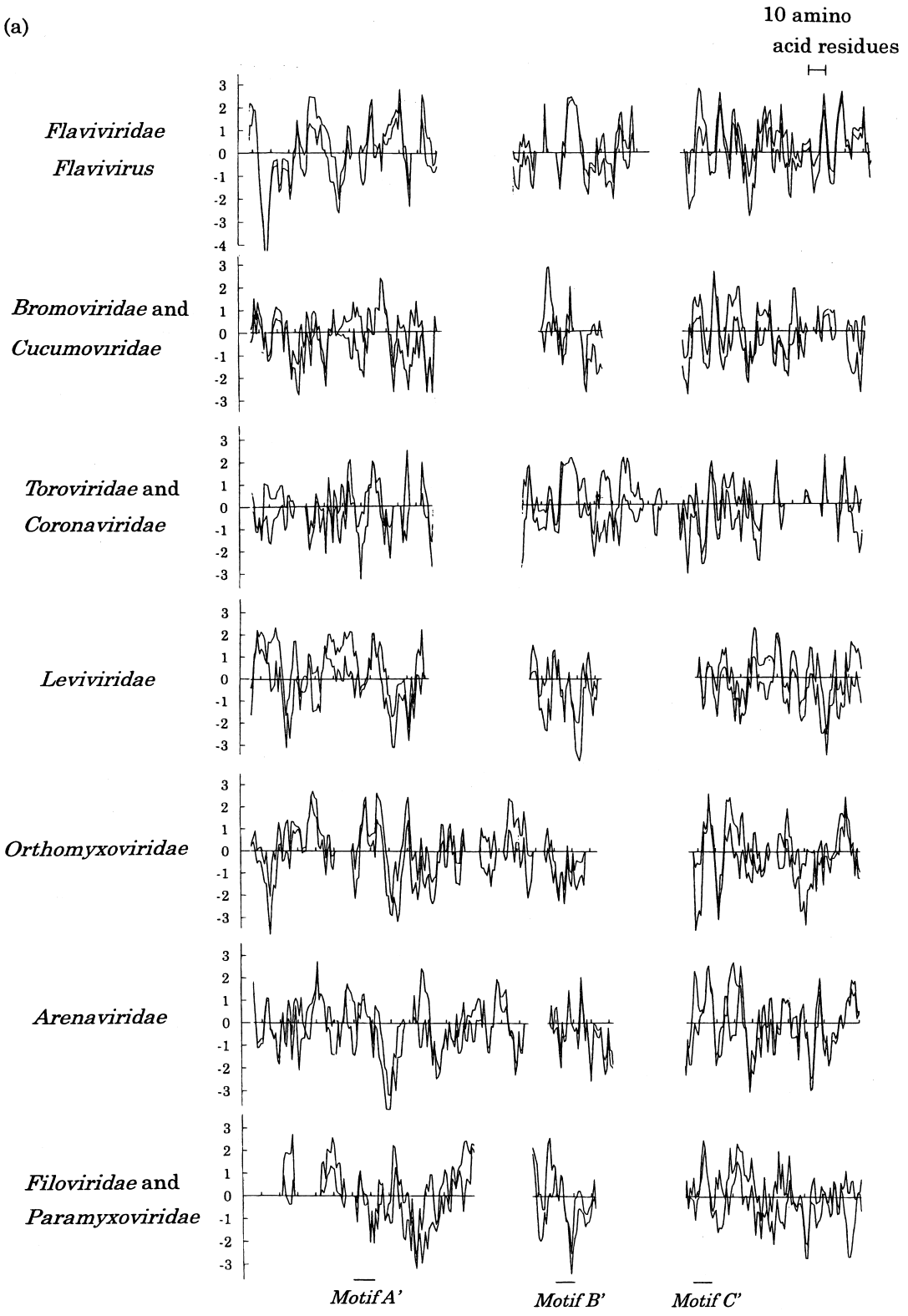

(a)
10 amino acid residues
Flaviviridae
Flavivirus
Bromoviridae and
Cucumoviridae
Toroviridae and
Coronaviridae
Leviviridae
Orthomyxoviridae
Arenaviridae
Filoviridae and
Paramyxoviridae
Motif A'
Motif B'
Motif C'

Fig. 6.

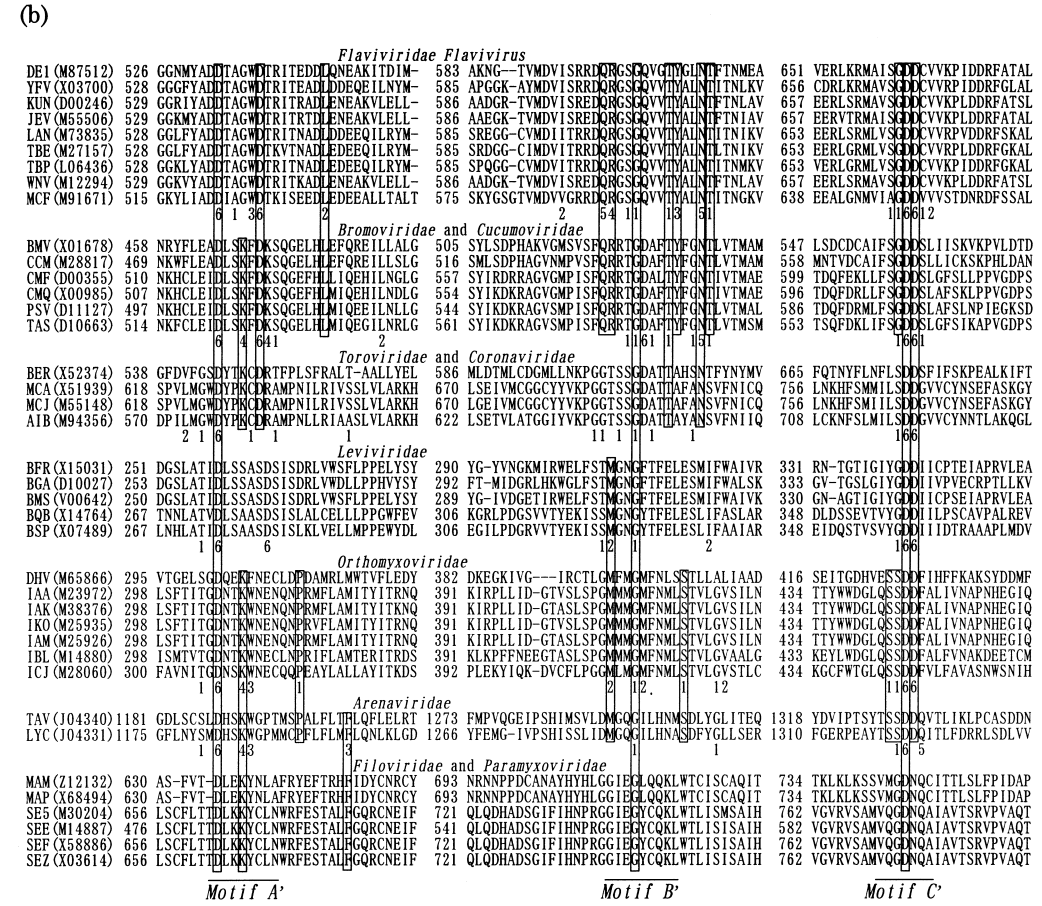



Fig. 6. Seven groups of RNA dependent RNA polymerases from viruses and bacteriophages, and the sequence motifs $A'$, $B'$ and $C'$ proposed for their polymerization activities. (a) Comparison of $z^{(1)}$ diagrams among seven groups of RNA dependent RNA polymerases. (b) Homologous alignment of amino acid sequences in the region showing similar sequence patterns. Identical and similar amino acid residues between different groups are boxed and denoted by the category numbers, respectively. Although all the groups contain the aspartic acid residues in motifs $A'$ and $C'$, and one glycine residue in motif $B'$, these groups represent the variation encompassing the polymerization motifs seen in families $A$, $B$ and $C$ of DNA dependent polymerases; the number of aspartic acid residues varies from three to one in motif $A'$, two to one in motif $C'$, and lysine or the same category of amino acid residues as well as the conserved glycine appear in motif $B'$ of some groups of RNA dependent RNA polymerases. The sources from which these RNA dependent RNA polymerases are derived are denoted by the following abbreviations: DE1, dengue virus type 1; YFV, yellow fever virus; KUN, Kunjin virus; JEV, Japanese encephalitis virus; LAN, Langat virus; TBE, tick-borne encephalitis virus; TBP, tick-borne powassan virus; WNV; West Nile virus; MCF, mosquito cell fusing agent (CFA flavivirus); BMV, Brome mosaic virus; CCM, cowpea chlorotic mottle virus, CMF; cucumber mosaic virus (strain FNY); CMQ, cucumber mosaic virus (strain Q); PSV, peanut stunt virus; TAS, tomato aspermy virus; BER, Berne virus; MCA, murine coronavirus MHV (strain A59); MCJ, murine coronavirus MHV (strain JHM); AIB, avian infectious bronchitis virus; BFR, bacteriophage FR; BGA, bacteriophage GA; BMS, bacteriophage MS2; BQB, bacteriophage Qβ; BSP, bacteriophage SP; DHV, Dhori virus; IAA, influenza A virus (strain A/Ann Arbor/6/60); IAK, influenza A virus (strain A/Kiev/59/79); IKO, influenza A virus (strain A/Korea/426/68); IAM, influenza A virus (strain A/Mallard/New York/6750/78); IBL, influenza B virus (strain B/LEE/40); ICJ, influenza C virus (strain C/JJ/50); TAV, Tacaribe virus; LYC, lymphocytic choriomeningitis virus; MAM, Marburg virus (strain Musoke); MAP, Marburg virus (strain Popp); SE5, Sendai virus (strain Z/host mutants); SEE, Sendai virus (strain Enders); SEF, Sendai virus (strain Fushimi); SEZ, Sendai virus (strain Z).

iable regions. Thus, the $z^{(1)}$ diagram is constructed on the basis of the homologous alignment of each group of amino acid sequences, and different groups of RNA dependent RNA polymerases are compared by the respective diagrams in Fig. 6a. This comparison found at least three regions, each of which shows a similar sequence pattern among the groups, and the amino acid sequence fragments underlying

the similar sequence patterns are compared in Fig. 6b.

Although the amino acid residues conserved in all these groups of RNA dependent RNA polymerases are only D in the first region, G in the second region and D in the third region, much more residues are conserved in each group. Among them, the amino acid residues similar to those in the polymerization motifs of families *A*, *B* and *C* are found in some groups of RNA dependent RNA polymerases. As seen in Fig. 6b, the three types of sequence fragments, DD----D, R---RGSG and GDD, are conserved in the group of polypeptide chains from *Flavivirus* of Flaviviridae. These three sequence fragments are also observed in the polypeptide chains exhibiting the polymerization activities from Potyviridae, Luteoviridae, Picornaviridae and Tobamoviridae, although they are omitted from the figure because of their resemblance to those from Flaviviridae *Flavivirus*. Similar but somewhat changed sequence motifs can be found in the RNA dependent RNA polymerases from other groups of viruses and bacteriophages. In the polypeptide chains from Bromoviridae and Cucumoviridae, the first and third motifs remain D----D and GDD, respectively, but the second motif is replaced by TG. Toroviridae and Coronaviridae carry the first and second motifs of D----D and K------SG, but SDD in the third motif. In the polypeptide chains from bacteriophage Leviviridae, the third motif GDD is found, but the first motif is D-----D and the second motif is changed into M--G instead of SG. In the polypeptide chains from Orthomyxoviridae, the first motif is replaced by D-----E, although the second and third motifs are M--G and SDD, respectively. In the polypeptide chains from Arenaviridae, the first motif is D-KW, the second motif M--G and the third motif is SDD. In the polypeptide chains from Filoviridae and Paramyxoviridae, the number of aspartic acid residues is reduced to one in both the first and third motifs, and only one glycine residue in the second motif is common to other groups of RNA dependent RNA polymerases, although several other amino acid residues such as H---GG-G seem to be specifically conserved in the second region.

Thus, the center of polymerization activity in most RNA dependent RNA polymerases seems to be characterized by the three sequence motifs in the same way as the sequence motifs *A*, *B* and *C* in DNA dependent DNA polymerases. These sequence motifs will be denoted as *A′*, *B′* and *C′*, respectively, hereafter. Although only two aspartic acid residues, one in motif *A′* and another in motif *C′*, were noted in the previous alignment of RNA dependent RNA polymerases from a few species to family *A* DNA polymerases [10], our systematic comparison of much more sequence data reveals that RNA dependent RNA polymerases are full of variety encompassing the sequence motifs of three families *A*, *B* and *C* of polymerases; the number of aspartic acid residues varies from one to three in motif *A′*, one to two in motif *C′*, and categories 1 and 4 of amino acid residues, in addition to the conserved glycine, tend to appear in motif *B′*. However, it should be noted that these sequence motifs, especially motifs *A′* and *C′*, are located at more hydrophilic regions than the polymerization motifs of DNA dependent DNA polymerases.

### 3.5. RNA dependent DNA polymerases (reverse transcriptases) and telomerases

The sequence motifs for polymerization activities of reverse transcriptases as well as RNA dependent RNA polymerases have been investigated on the analogy of the polymerization motifs of family *A* DNA polymerases [18,44,46–48]. However, the variety of RNA dependent RNA polymerases indicated in Section 3.4 requires more careful examination of reverse transcriptase sequences. The amino acid sequences of now available RNA dependent DNA polymerases from viruses are classified into three groups by the criterion of similarity scores of more than 200 evaluated with the FASTA program. These groups correspond to those from Caulimoviridae, Lentivirinae and Oncovirinae, respectively. The amino acid sequences of the polymerases in the same group are easily aligned homologously, and the $z^{(1)}$ diagram is constructed on the basis of their homologous alignment in each group. Although the $z^{(1)}$ diagrams of the three groups are considerably different from each other, a careful comparison of these diagrams detects three regions, each of which shows a similar sequence pattern among the three groups. These regions are also found in telomerase, which is considered to be a specialized form of a reverse transcriptase that syn-

thesizes a DNA sequence using its own RNA template to seal the ends of a linear DNA.

The sequence patterns of these three regions are compared between telomerase and the three groups of reverse transcriptases in Fig. 7a, and their amino acid sequence fragments are compared in Fig. 7b. As seen in Fig. 7b, the first, second and third regions contain conserved sequence fragments D, P-G and DD, respectively, in all groups of reverse transcriptases and telomerases. From the tertiary structure of reverse transcriptase from human immunodeficiency virus HIV-1 (Lentivirinae), the N-terminal aspartic acid residue in the first region and two aspartic acid residues in the third region are suggested to be stereochemically arranged to highlight the relative position of the 'catalytic triad' accepting a divalent cation in a similar way to the polymerase active center of the Klenow fragment [48]. Moreover, Asp-530 in the first region of telomerase from *Saccharomyces cerevisiae* is suggested to play an essential role in telomerase activity by mutational study [49]. However, it should be noted that the amino acid residues conserved in the third region are DD just like those in motif $C'$ of RNA dependent RNA polymerases from most viruses, in contrast to HDE conserved in motif $C_A$ of family $A$ DNA polymerases. The amino acid residues conserved in the second region are also different from those conserved in motif $B_A$ of family $A$ DNA polymerases. Thus, the amino acid sequence fragments in the three regions of these reverse transcriptases will be denoted as motifs $A''$, $B''$ and $C''$, respectively, in Fig. 7a and b, being distinguished from those of family $A$ DNA polymerases. Motifs $A''$, $B''$ and $C''$ all locate on the border from hydrophilic to hydrophobic parts, and, in this sense, the environmental structure of these motifs may be similar to that of DNA dependent DNA polymerases rather than that of RNA dependent RNA polymerases. Thus, reverse transcriptase may be an intermediate between RNA dependent RNA polymerase and family $A$ of DNA dependent DNA polymerase.

### 3.6. Poly(A) polymerase, tRNA nucleotidyltransferase, DNA polymerase β and deoxynucleotidyltransferase

The similarity of sequence motifs for polymerization of nucleotides and deoxynucleotides has also been suggested for poly(A) polymerase, tRNA nucleotidyltransferase, DNA polymerase β and deoxynucleotidyltransferase [50]. In order to ascertain this suggestion, we constructed $z^{(1)}$ diagrams separately for eukaryotic poly(A) polymerases, eubacterial poly(A) polymerases, eubacterial and eukaryotic tRNA nucleotidyltransferases, archaebacterial tRNA nucleotidyltransferases, DNA polymerase β and deoxynucleotidyltransferase. These $z^{(1)}$ diagrams are compared in Fig. 8a, and the six types of amino acid sequences in the regions showing a similar sequence pattern are vertically aligned in Fig. 8b. Curiously, the largest number of identical and/or similar properties of amino acid residues are shared between eubacterial poly(A) polymerase and eubacterial tRNA nucleotidyltransferase. In fact, a high similarity score of 280 is counted between *E. coli* tRNA nucleotidyltransferase and *B. subtilis* poly(A) polymerase by the FASTA program, while the similarity of eubacterial poly(A) polymerase to eukaryotic poly(A) polymerase is not as high as the similarity between eubacterial and eukaryotic tRNA nucleotidyltransferases, e.g., the optimal score calculated between *E. coli* poly(A) polymerase and *Candida albicans* poly(A) polymerase is only 42. At any rate, several amino acid residues are found to be conserved in all available amino acid sequences of poly(A) polymerases and tRNA nucleotidyltransferases. Among them, we can find the three aspartic acid residues that are also conserved in DNA polymerase β and terminal deoxynucleotidyltransferase. These three aspartic acid residues correspond to those indicated to be essential for the catalysis of mammalian poly(A) polymerase by mutational analysis [50]. In rat DNA polymerase β, Asp-190, Asp-192 and Asp-256 are indicated to suspend magnesium ions in site $A$ and site $B$ [51]. Thus, the two regions containing the two and one aspartic acid residues are denoted as motifs $A_X$ and $C_X$, respectively, in Fig. 8a and b, according to the family name of DNA polymerase β.

In the description of the nucleotidyl transfer reaction inferred from the structures of ternary complexes of rat DNA polymerase β, DNA template primer and ddCTP [51], Gly-274 and Ser-275 are considered to facilitate the release of the enzyme from the template primer by a conformational change of *cis*- to *trans*-peptide at these residues and

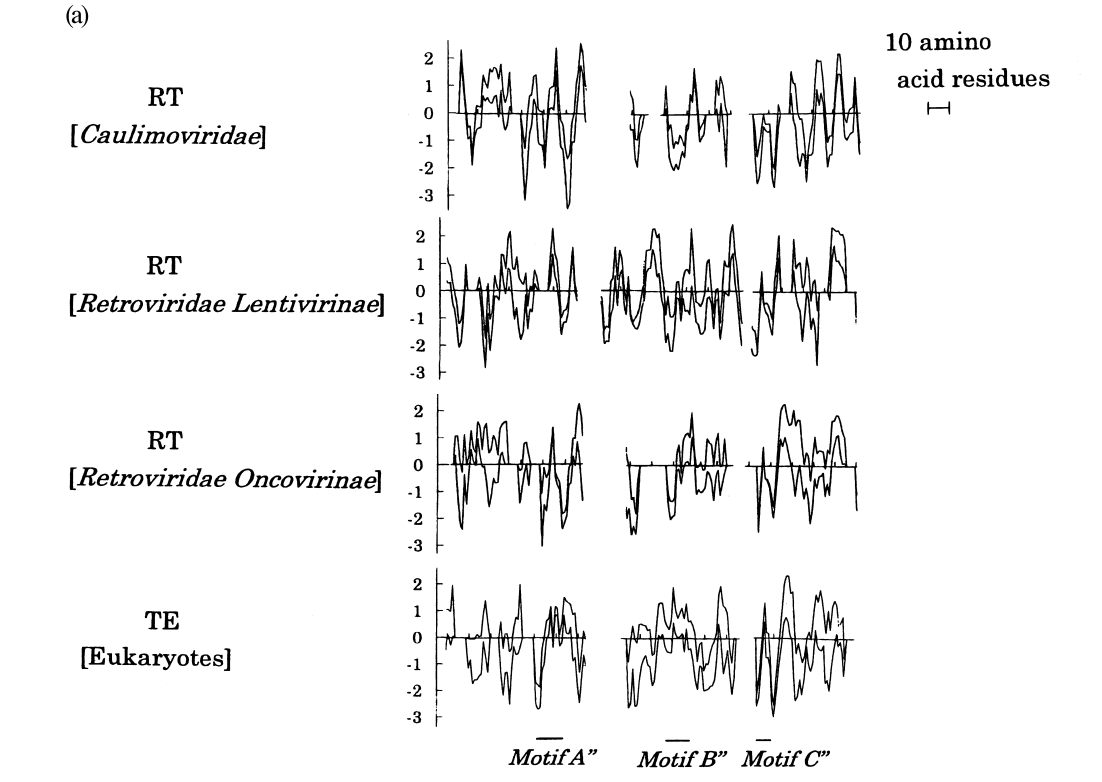
some of the amino acid residues (Asn-279, Asp-276 and Tyr-271) in the C-terminal side of the metal suspending residues are considered to play a role in positioning an incoming nucleotide into the active site of the two metal ions. Certainly, these glycine and serine residues are conserved in terminal deoxynucleotidyltransferases as well as DNA polymerases β from different species. However, the residues corresponding to Tyr-271 and Asp-276 are not found in DNA polymerases β from human and *S. cerevisiae* while Arg as well as Asn are conserved. In particular, it should be noted that Arg belongs to the same category as Lys. Thus, the sequence fragment TGS---N---R conserved in DNA polymerase β and deoxynucleotidyltransferase is denoted as motif $B_X$ in Fig. 8b. Although the arrangements of these three motifs $A_X$, $B_X$ and $C_X$ as well as of the amino acid residues in motif $B_X$ are different from those in other families of DNA polymerases, the tertiary structure of the polymerization domain of DNA polymerase β is known to be quite different from the structures of families *A* and *B* of DNA polymerases, as will be discussed in Section 4.

### 3.7. The sequence motifs for 5′-3′ exonuclease activity

Lastly, we will briefly describe the sequence motifs for 5′-3′ exonuclease activity. The 5′-3′ exonuclease activity is known in the N-terminal fragments of eubacterial polymerases I [12], 5′-3′ exonucleases from T-phages and flap endonuclease [52]. Furthermore, the N-terminal fragment of DNA polymerase I shows considerable similarity to exonuclease, DNA repair proteins, RAD2, RAD13, RAD27 and XP-G, and DNA damage inducible protein DIN7 as well as flap endonuclease, all of which are called the XP-G/RAD2 family [53]. The comparison of $z^{(1)}$ diagrams and the homologous alignment of these polypeptide chains show that the amino acid residues conserved in the six regions *A–F* of DNA polymerase I indicated previously [12] are also found in the XP-G/RAD2 family. However, the conservation of amino acid residues is hardly observed in regions *B* and *C*, when the sequence comparison is advanced to the 5′-3′ exonucleases of T-phages. Although regions *C* and *F* are suggested to be required for 5′-3′ exonuclease activity from the mutations of DNA polymerase I defective in this activity [52,54], the present comparison of three groups of proteins strongly suggests that regions *A*, *D*, *E* and *F* are essential for 5′-3′ exonuclease activity. In practice, the recent results of X-ray diffraction analyses on DNA polymerase I from *T. aquaticus* [55] and on 5′-3′ exodeoxyribonuclease from bacteriophage T5 [56] indicate that the aspartic acid and glutamic acid residues in these four regions are associated with the suspension of manganese and zinc ions, although the number of suspended metal ions is reported to be three in DNA polymerase I in contrast to two in exodeoxyribonuclease.

→

Fig. 7. Sequence motifs proposed for the polymerization activities of reverse transcriptases (RT) and telomerases (TE). For convenience of homologous alignment, available amino acid sequences of reverse transcriptases are divided into three groups corresponding to those from Caulimoviridae, Lentivirinae and Oncovirinae, respectively. (a) Comparison of sequence patterns, each represented by $z^{(1)}$ diagram. (b) Vertical alignment of amino acid sequences in the region showing similar sequence patterns. The amino acid residues conserved in the three groups of reverse transcriptases and telomerases are boxed, and the similar amino acid residues between these polypeptide chains are denoted by the category numbers. The sequence motifs proposed for the polymerization activities of reverse transcriptases including telomerases are denoted as *A″*, *B″*, and *C″*. The sources from which the compared sequences are derived are indicated by the following abbreviations: CMV, cauliflower mosaic virus; CER, carnation etched ring virus; CYM, Commelina yellow mottle virus; FMV, figwort mosaic virus; RTB, rice tungro bacilliform virus; SCM, soybean chlorotic mottle virus; CAE, caprine arthritis encephalitis virus; EIA, equine infectious anemia virus; FIV, feline immunodeficiency virus; HIV, human immunodeficiency virus; JSR, Jaagsiekte sheep retrovirus; SIV, simian immunodeficiency virus; VIL, Visna lentivirus; BIV, bovine immunodeficiency virus; BLV, bovine leukemia virus; GAL, gibbon ape leukemia virus; HTL, human T-cell leukemia virus type 1; AML, AKV murine leukemia virus; MMT, mouse mammary tumor virus; SMP, simian mason-pfizer virus; RSV, Rous sarcoma virus; SFV, simian foamy virus; SMR, squirrel monkey retrovirus; SRV, simian retrovirus SRV-1; BEN, baboon endogenous virus; TTH, *Tetrahymena thermophila*; EAE, *Euplotes aediculatus*; OTR, *Oxytricha trifallax*. The abbreviations for the names of the other eukaryotes are already described in the legends to Figs. 1–3.




(b)

*Caulimoviridae*

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | CMV(V00140) | 327 | LTLIRGKKIFSSH | DCKSGFWQVLLDQESRPL | 358 | TAFTCP-------QGHYEWNVVPF | GLKQAPSIFQR----- | 388 | DEAFRVFRKFCCVYV | DDILVFSNN- |
| | CER(X04658) | 309 | LTLVRGKKIYSSH | DCKSGLWQVLLDKESQLL | 340 | TAFTCP-------QGHYQWNVVPF | GLKQAPSIFPKT---- | 371 | NSHSNQYSKYCCVYV | DDILVFSNTG |
| RT | CYM(X52938) | 1490 | ISKVGRSKIYSKH | DLKSGFWQVAMEEEESVPW | 1521 | TAFLAG-------NKLYEWLVMPF | GLKNAPAIFQR----- | 1551 | DNVFKGTEKFIAVYI | DDILVFSET- |
| | FMV(X06166) | 320 | LTLLRGKSIFSSH | DCKSGFWQVVLDEESQKL | 351 | TAFTCP-------QGHFQWKVVPF | GLKQAPSIFQR----- | 381 | QTALNGADKFCMVYV | DDIIVFSNS- |
| | RTB(X57924) | 1266 | INLIQKANIFSKH | DLKAGFHHMKLKDDFKDW | 1297 | TTFTCS-------EGLYTWNVCPF | GIANAPCAFQR----- | 1327 | QESFGDL-KFALLYI | DDILIASNN- |
| | SCM(X15828) | 283 | SEKIKGSLWFSSL | DAKSGYYQLRLHENTKPL | 314 | TAFSCP------PQKHYEWNVLSF | GLKQAPCIYQR----- | 345 | DQSLKGLDHIYLAYI | DDILIFTKGS |

*Retroviridae Lentivirinae*

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | CAE(M33677) | 243 | PGGLQKKKHVTIL | DIGDAYFTIPLYEPYREY | 274 | TCFTLLSPNNLGPCKRYYWKVLPQ | GWKLSPSVYQFTMQEI | 316 | DWIQQHPEIQFGIYM | DDIYIGSDLE |
| | EIA(M16575) | 282 | PGGLIKCKHMTVL | DIGDAYFTIPLDPEFRPY | 313 | TAFTIPSINHQEPDKRYVWNCLPQ | GFVLSPYIYQKTLQEI | 355 | PFRERYPEVQLYQYM | DDLFVGSNGS |
| | FIV(M25381) | 252 | PAGLQIKKQVTVL | DIGDAYFTIPLDPDYAPY | 283 | TAFTLPRKNNAGPGRRFVWCSLPQ | GWILSPLIYQSTLDNI | 325 | PFIRQNPQLDIYQYM | DDIYIGSNLS |
| RT | HIV(M26727) | 252 | PAGLKKKKSVTVL | DVGDAYFSVPLDKDFRKY | 283 | TAFTIPSINNETPGIRYQYNVLPQ | GWKGSPAIFQSSMTKI | 325 | PFRKQNPDIVIYQYM | DDLYVGSDLE |
| | JSR(M80216) | 107 | PSAIPDKSYIIVI | DLKDCFYTIPLAPQDCKR | 138 | FAFSLPSVNFKEPMQRYQWRVLPQ | QGMTNSPTLCQKFVATA | 180 | PVRQRFPQLYLVHYM | DDILLAHTDE |
| | SIV(M29975) | 288 | PAGLKKMKQTITI | DVGDAYYSIPLDPEFRKY | 319 | TAFTIPTVNNEGPGIRYQFNCLPQ | GWKGSPTIFQNTASKI | 361 | EIKKELKQLTIVQYM | DDLWVGSQEE |
| | VIL(M10608) | 238 | PGGLQRKKHVTIL | DIGDAYFTIPLYEPYRQY | 269 | TCFTMLSPNNLGPCVRYYWKVLPQ | QGWKLSPAVYQFTMQKI | 311 | GWIEEHPMIQFGIYM | DDIIYIGSDLG |
| | BIV(M32691) | 253 | PPGIKECEHLTAI | DIKDAYFTIPLHEDFRPF | 284 | TAFSVVPVNREGPIERFQWNVLPQ | QGWVCSPAIYQTTTQKI | 326 | NIKKSHPDVMLYQYM | DDLLIGSNRD |

*Retroviridae Oncovirinae*

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | BLV(K02120) | 75 | TAIPTHPPHIICL | DLKDAFFQIPVEDRFRFY | 106 | LSFTLPSPGGLQPHRRFAWRVLPQ | QFINSPALFERALQEP | 148 | QVSAAFSQSLLVSYM | DDILYASPTE |
| | GAL(M26927) | 254 | SSLPPSYTWYSVL | DLKDAFFCLRLHPNSQPL | 285 | FAFEWKDPE-KGNTGQLTWTRLPQ | GFKNSPTLFDEALHRD | 326 | PFRALNPQVVLLQYV | DDLLVAAPTY |
| | HTL(J02029) | 101 | SSLPTTLAHLQTI | DLRDAFFQIPLPKQFQPY | 101 | FAFTVPQQCNYGPGTRYAWKVLPQ | QGFKNSPTLFEMQLAHI | 134 | PIRQAFPQCTILQYM | DDILLASPSH |
| | AML(J01998) | 257 | SGLPPSHRWYTVL | DLKDAFFCLRLHPTSQPL | 288 | FAFEWRDPG-MGISGQLTWTRLPQ | GFKNSPTLFDEALHRD | 329 | DFRIQHPDLILLQYV | DDILLAATSE |
| | MMT(K01707) | 102 | VAVPKGW-EIIII | DLQDCFFNIKLHPEDCKR | 132 | FAFSVPSPNFKRPYQRFQWKVLPQ | QGMKNSPTLCQKFVDKA | 174 | TVRDKYQDSYIVHYM | DDILLAHPSR |
| RT | SMP(M12349) | 108 | VAIPQGY-LKIII | DLKDCFFSIPLHPSDQKR | 138 | FAFSLPSTNFKEPMQRFQWKVLPQ | QGMANSPTLCQKYVATA | 180 | KVRHAWKQMYIIHYM | DDILIAGKDG |
| | RSV(V01197) | 95 | SALPRGW-PLMVL | DLKDCFFSIPLAEQDREA | 125 | FAFTLPSVNNQAPARRFQWKVLPQ | QGMTCSPTICQLVVGQV | 167 | PLRLKHPSLCMLHYM | DDLLLAASSH |
| | SFV(X58484) | 252 | SSIYRGK-YKTTL | DLTNGFWAHPITPESYWL | 282 | TAFTWQ-----GK--QYCWTRLPQ | GFLNSPALFTADVVDL | 317 | EIPN------VQAYV | DDIYISHDDP |
| | SMR(K01706) | 105 | VAIPLNY-HKIVI | DLKDCFFTIPLHPEDRPY | 135 | FAFSVPQINFQSPMPRYQWKVLPQ | QGMANSPTLCQKFVAAA | 177 | PVRSQWPEAYILHYM | DDILLACDSA |
| | SRV(M11841) | 108 | VAIPQGY-LKIII | DLKDCFFSIPLHPSDQKR | 138 | FAFSLPSTNFKEPMQRFQWKVLPQ | QRMANSPTLCQKYVATA | 180 | KVRHAWKQMYIIHYM | DDILIAGKDG |
| | BEN(M16550) | 255 | STLKPDYSWYTVL | DLKDAFFCLPLAPQSQEL | 286 | FAFEWKDPE-RGISGQLTWTRLPQ | GFKNSPTLFDEALHRD | 327 | DFRTQHPEVTLLQYV | DDLLLAAPTK |

*Eukaryotes*

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | HSP(AF015950) | 699 | AQDPPPELYFVKV | DVTGAYDTIPQDRLTEVI | 810 | LRFMCHHAVRIRGKSYVQCQGIPQ | GSILSTLLCSLCYGDM | 859 | RD-----G-LLLRLV | DDFLLVTPHL |
| | MMU(AF051911) | 689 | ALDQTPRMYFVKA | DVTGAYDAIPQGKLVEVV | 803 | LHFLRHSVVKIGDRCYTQCQGIPQ | GSSLSTLLCSLCFGDM | 852 | RD-----G-LLLRFV | DDFLLVTPHL |
| | TTH(AF061284) | 606 | NKGR-PQLYYVTL | DIKKCYDSIDQMKLLNFF | 750 | KHISQYNVISFNKVKFRQKRGIPQ | GLNISGVLCSFYFGKL | 800 | AEQVNGSINLLMRLT | DDYLFISDSQ |
| TE | EAE(U95964) | 591 | QVGQ-PKLFFATM | DIEKCYDSVNREKLSTFL | 714 | INICQYNYINFNGKFYKQTKGIPQ | GLCVSSILSSFYYATL | 766 | MNPENPNVNLLMRLT | DDYLLITTQE |
| | OTR(AF060230) | 696 | QINS-PKLYFVAM | DIEKCYDNVDCERVVNFL | 819 | LKICQNNYVTFNKKQYKQMKGIPQ | GLCVSYILSSFYYANL | 871 | MDPEKPEINLLMRLT | DDYLLMTTEK |
| | SPO(AF015783) | 580 | MFGR--KKYFVRI | DIKSCYDRIKQDLMFRIV | 684 | KEHLSGHIVKIGNSQYLQKVGIPQ | GSILSSFLCHFYMEDL | 733 | KK-----GSVLLRV | DDFLFITVNK |
| | SCE(U20618) | 518 | NNVL-PELYFMKF | DVKSCYDSIPRMECMRIL | 609 | EMEIFKTALWVEDKCYIREDGLPQ | GSSLSAPIVDLVYDDL | 661 | QD-----T-LILKLA | DDFLIISTDQ |

*Motif A"* *Motif B"* *Motif C"*

Fig. 7.

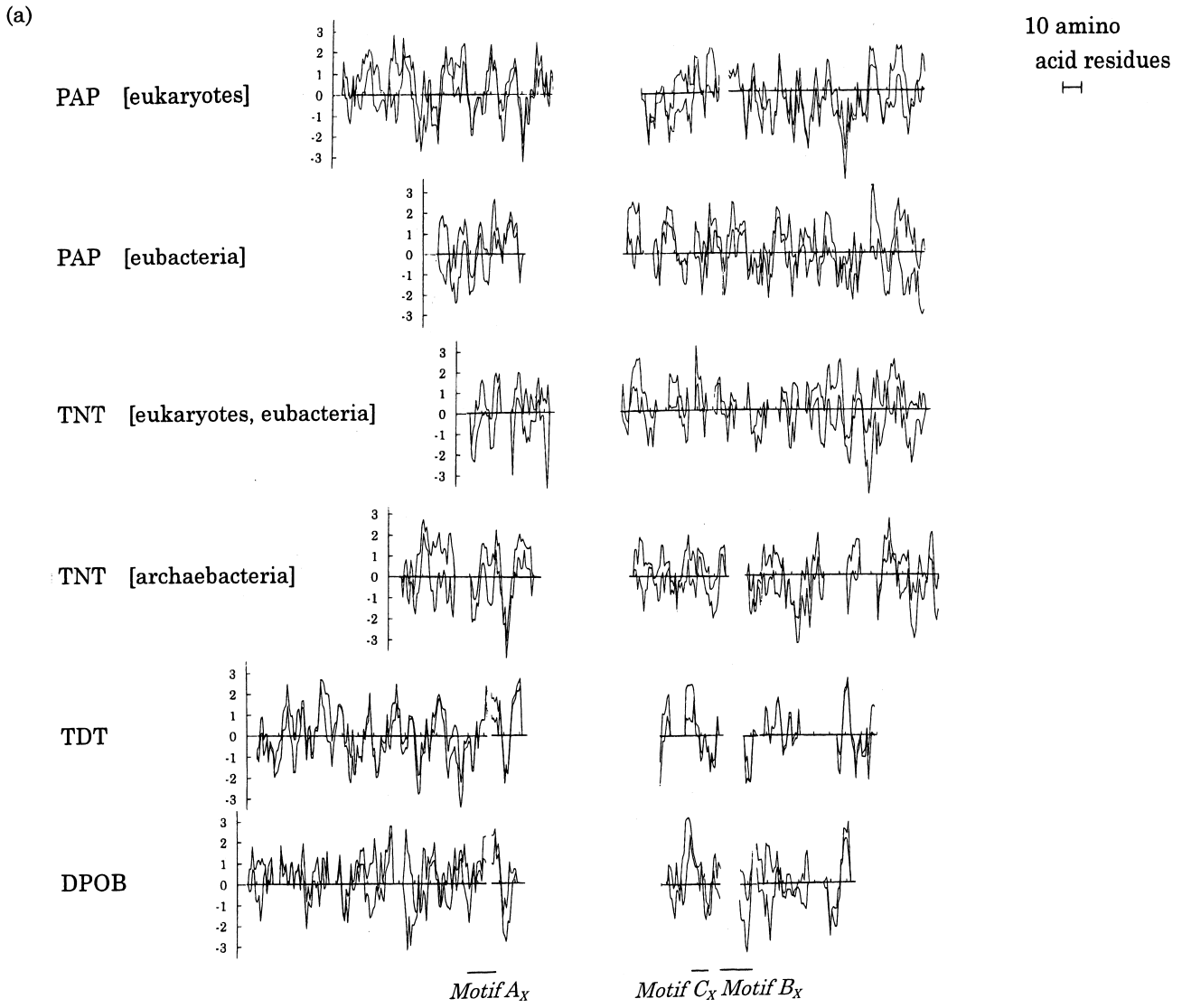



Fig. 8. Homology between poly(A) polymerase (PAP) and tRNA nucleotidyltransferase (TNT), and their similarity to terminal deoxynucleotidyltransferase (TDT) and DNA polymerase β (DPOB). (a) Comparison by $z^{(1)}$ diagrams. (b) Homologous alignment of amino acid sequences. The identical and same categories of amino acid residues between eukaryotic PAP, eubacterial PAP, eukaryotic and eubacterial TNT and archaebacterial TNT are boxed and denoted by the category numbers, respectively. These conservation patterns of amino acid residues are compared to those of TDT and DPOB shown in the lower rows, where the three aspartic acid residues identified to suspend two magnesium ions in rat DPOB [51] are denoted. The sequence motifs proposed for polymerization activities of TDT and DPOB are denoted as $A_X$, $B_X$ and $C_X$, according to the family name of these polymerases. Abbreviations for the organisms: BTA, *Bos taurus*; MDO, *Monodelphis domestica*; GGA, *Gallus gallus*; XLA, *Xenopus laevis*; RAT, *Rattus norvegicus*. The abbreviations for the other organisms are already described in the legends to Figs. 1–3.

## 4. Conclusions and discussion

Although some other motifs for polymerization probably correlated with template or substrate specificity have been proposed, e.g., T/R--GR only found in the N-terminal side of motif *A* of DNA dependent polymerase, G--h---K in the C-terminal side of motif *C′* or *C″* in RNA dependent polymerase and LG in the further C-terminal side of RNA dependent RNA polymerase [57], we thoroughly compare the amino acid sequences of many types of polymerases in the present paper, mainly focusing on the minimal set of motifs that are most widely distributed in polymerases. This investigation identifies

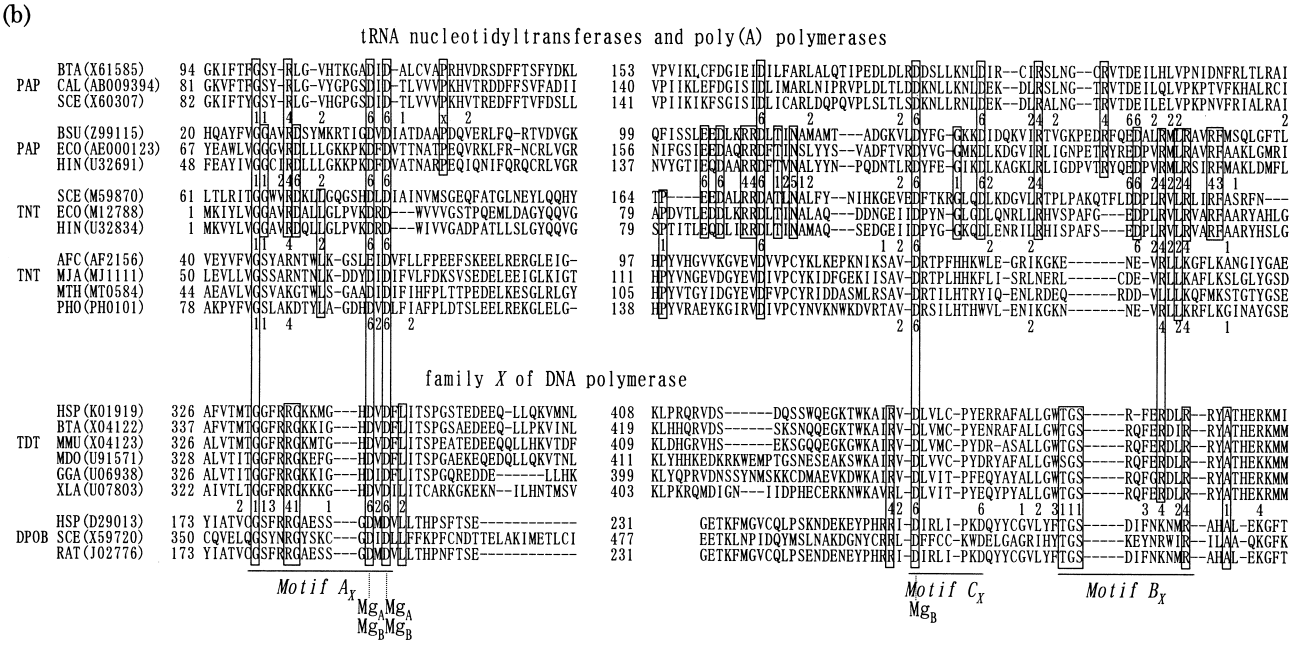



Fig. 8 (*continued*).

the sequence motifs for polymerization activity of family *C* DNA polymerase, giving an open view for many RNA polymerases as well as DNA polymerase III itself. The similarity relations of DNA and RNA polymerases, mainly focused on the sequence motifs for polymerization, are summarized in Fig. 9. One of the most noticeable points is that the sequence motifs for the polymerization activities of DNA dependent RNA polymerases are also divided into three types in accordance with those of families *A*, *B* and *C* of DNA dependent DNA polymerases. This result not only clarifies the lineages of sequence motifs for polymerization activities of DNA dependent RNA polymerases including primases but also serves to fill the gap between DNA dependent DNA polymerases and RNA dependent RNA polymerases.

In the early comparison of families *A* and *B* of DNA polymerases and some RNA dependent RNA polymerases [10], motif *B* was speculated to associate with DNA template strand, because the sequence fragment K------(-)YG conserved in family *A* and *B* DNA polymerases were not found in compared RNA dependent RNA polymerases. However, the amino acid residues R/K--G-H-GG conserved in motif $B_C$ of DNA polymerase III are somewhat different from K------(-)YG and they are changed into G-K/RG in multimeric DNA dependent RNA polymerases and in eukaryotic and archaebacterial primases. With such variation in mind, we can also find a region fairly well conserving G and some other residues such as Y, S and M between motifs *A′* and *C′* in RNA dependent RNA polymerases from many sources, although category 4 amino acid residues such as K, R and H are not found in this region of all RNA dependent RNA polymerases. Thus, the residues in this region may also play a role in releasing the enzyme from the RNA template, although they are not so strongly conserved as in DNA dependent DNA polymerases.

The structural similarity of polymerization domains between DNA dependent DNA polymerase, DNA dependent RNA polymerase and reverse transcriptase is already indicated for DNA polymerase I, T7 bacteriophage monomeric DNA dependent RNA polymerase and HIV-1 reverse transcriptase; motifs *A* ($A_A$ and *A″* in our notation) and *C* ($C_A$ and *C″*) form three strands of a β sheet and a short segment of α-helix within the core of the palm subdomain and motif *B* ($B_A$ and *B″*) is located in the fingers domain [18,48,57]. As ascertained in the present investigation, the polymerases showing a similar tertiary structure also carry similar sequence motifs for polymerization, although a considerable similarity

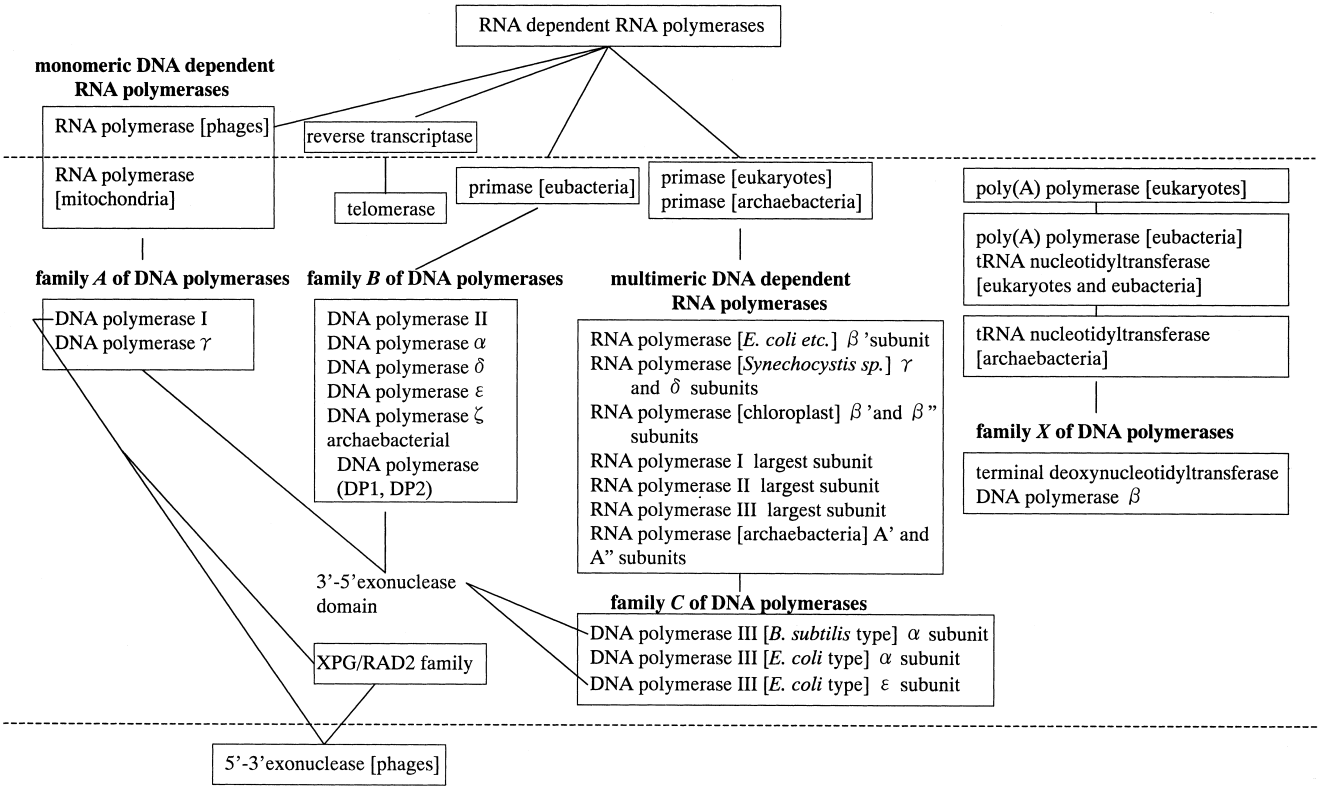



Fig. 9. Similarity relationships of DNA and RNA polymerases, mainly based on the similar amino acid sequences in the polymerase domains. The polymerases within the same box show the optimal score of more than 200 by the FASTA program, and carry almost the same sequence motifs for polymerization. Furthermore, DNA dependent RNA polymerases are also divided into four types according to their similarities to sequence motifs for polymerization activities of families *A*, *B*, *C* and *X* of DNA dependent DNA polymerases, i.e., monomeric DNA dependent RNA polymerases to family *A*, eubacterial primases to family *B*, eukaryotic and archaebacterial primases and multimeric RNA polymerases to family *C*, and poly(A) polymerases and tRNA nucleotidyltransferases to family *X*, although the FASTA program does not count a considerable similarity score between these DNA and RNA polymerases. The variety of sequence motifs for polymerization activities of RNA dependent RNA polymerases encompasses the polymerization motifs characteristics of family *A*, *B* and *C* polymerases as well as the sequence motifs for polymerization activities of reverse transcriptase including telomerase, while the sequence fragment similar to the family *X* polymerization motif is not found in any RNA dependent RNA polymerase. In contrast to the variation in sequence motifs for polymerization activities, the sequence motifs for 3′-5′ exonuclease activity are mostly common to families *A*, *B* and *C* of DNA dependent DNA polymerases, and the sequence motifs for 5′-3′ exonuclease activity present in DNA polymerase I are highly similar to those in the XPG/RAD2 family of proteins and those in 5′-3′ exonucleases from bacteriophages. These similarity relations of sequence motifs for polymerization and/or exonuclease activities are represented by the connection with a straight line. Proteins surrounded by two dotted lines are those encoded in the genomes of free living organisms.

score is not necessarily counted between them by the current method such as the FASTA program. Although T7 phage DNA dependent RNA polymerase carries extra 1 and 2 regions in comparison with DNA polymerase I, this may be reasonable because the DNA dependent RNA polymerase should be capable of sequence specific (promoter) DNA binding and template unwinding. The recently determined structure of bacteriophage RB69 DNA polymerase, which is considered to be homologous to DNA polymerase α, also shows a similar palm subdomain containing motifs *A* ($A_B$) and *C* ($C_B$), but its thumb subdomain is topologically different from that of family *A* polymerases and its fingers subdomain is simpler [58]. In practice, the number of sites between motifs $B_B$ and $C_B$ is smaller in DNA polymerase α than in DNA polymerase I, and this is also the case in eubacterial primase. With this difference of tertiary structures, the similar sequence motifs for polymerization activities of families *A* and *B* of DNA

polymerases have been proposed as an example of convergence [58]. However, this proposal seems too narrow to consider the evolution of polymerases. Although the study of crystal structure does not reach DNA polymerase III and multimeric DNA dependent RNA polymerase yet, it is true that the relative position of the three motifs is the same in the three families *A*, *B* and *C* of polymerases, reverse transcriptases including HVI-1 transcriptase show a similar sequence pattern to RNA dependent RNA polymerases and some of the RNA dependent RNA polymerases carry sequence motifs similar to those of family *C* polymerases. These facts seem to support the possibility that the polymerization domains of these three families of polymerases are homologous, because they could have diverged by insertion and/or deletion of some amino acid sequence fragments and some degree of substitutions. In this sense, the RNA dependent RNA polymerases as well as families *A*, *B* and *C* of DNA dependent DNA and RNA polymerases may be clustered as a superfamily of nucleic acid polymerase.

On the other hand, the polymerization motifs similar to those of DNA polymerase β and terminal deoxynucleotidyltransferase are not found in any RNA dependent RNA polymerase. In fact, the tertiary structure of DNA polymerase β solved recently [51,59,60] shows the catalytic domain containing a single helix packing against a five-stranded β sheet with an unusual topology of the mixture of antiparallel and parallel sheets, which is quite different from the families *A* and *B* of DNA polymerases. A similar structure of a single helix packing against a four-stranded β sheet is also seen in kanamycin nucleotidyltransferase solved at a low resolution [61]. On the basis of the structural similarity between these two proteins and a scan of the SWISS-PROT database using the sequence pattern in and around motif $A_X$, the nucleotidyltransferase superfamily is proposed to encompass polymerase family *X*, poly(A) polymerase, kanamycin nucleotidyltransferase, protein-$P_{II}$ uridyltransferase, streptomycin 3′-adenyltransferase, (2′-5′) oligoadenylate synthetase, and glutamine synthase adenyltransferase [62]. Thus, the lineage of these polymerases may be different from RNA dependent RNA polymerases and families *A*, *B* and *C* of polymerases.

In contrast to the variation in the amino acid sequence fragments responsible for polymerization activities, the sequence motifs for 3′-5′ exonuclease activity are almost the same in all DNA dependent DNA polymerases of families *A*, *B* and *C*, although this activity is lost in DNA polymerase α and DNA polymerases I from many species of eubacteria except for *E. coli* and *H. influenzae*. The sequence motifs for 5′-3′ exonuclease activity carried by DNA polymerase I are essentially the same as those in the 5′-3′ exonucleases of bacteriophages and in the eukaryotic proteins clustered in the XP-G/RAD2 family. Thus, most processive DNA dependent DNA polymerases would have been generated by the fusion of 3′-5′ and/or 5′-3′ exonuclease domains to the polymerization domains after their differentiation into the three types. In practice, the editing 3′-5′ exonuclease domain of T7 bacteriophage RB69 is homologous to that of *E. coli* DNA polymerase I but lies on the opposite side of the polymerization active site [58]. The presence of phosphoesterase in both the α subunit of DNA polymerase III and DNA polymerase β [31] may also be due to the result of domain fusion.

## References

[1] D.L. Ollis, P. Brick, R. Hamlin, N.G. Xuong, T.A. Steitz, Structure of large fragment of *Escherichia coli* DNA polymerase I complexed with dTMP, Nature 313 (1985) 762–766.

[2] D.L. Ollis, C. Kline, T.A. Steitz, Domain of *E. coli* DNA polymerase I showing sequence homology to T7 DNA polymerase, Nature 313 (1985) 818–819.

[3] P. Argos, A.D. Tucker, L. Philipson, Primary structural relationships may reflect similar DNA replication strategies, Virology 149 (1986) 208–216.

[4] M.C. Leavitt, J. Ito, T5 DNA polymerase: structural-functional relationships to other DNA polymerases, Proc. Natl. Acad. Sci. USA 86 (1989) 4465–4469.

[5] F.C. Lawyer, S. Stoffel, R.K. Saiki, K. Myambo, R. Drummond, D.H. Gelfand, Isolation, characterization, and expression in *Escherichia coli* of the DNA polymerase gene from *Thermus aquaticus*, J. Biol. Chem. 264 (1989) 6427–6437.

[6] P. Lopez, S. Martinez, A. Diaz, M. Espinosa, S.A. Lacks, Characterization of the pol A gene of *Streptococcus pneumoniae* and comparison of the DNA polymerase I; it encodes homologous enzymes from *Escherichia coli* and phage T7, J. Biol. Chem. 264 (1989) 4255–4263.

[7] T.S.F. Wang, S.W. Wong, D. Korn, Human DNA polymer-

ase α: predicted functional domains and relationships with viral DNA polymerases, FASEB J. 3 (1989) 14–21.
[8] B.Z. Zmudzka, D. SenGupta, A. Matsukage, F. Cobianchi, P. Kumar, S.H. Wilson, Structure of rat DNA polymerase beta revealed by partial amino acid sequencing and cDNA cloning, Proc. Natl. Acad. Sci. USA 83 (1986) 5106–5110.
[9] A. Matsukage, K. Nishikawa, T. Ooi, Y. Seto, M. Yamaguchi, Homology between mammalian DNA polymerase beta and terminal deoxynucleotidyltransferase, J. Biol. Chem. 262 (1987) 8960–8962.
[10] M. Delarue, O. Pock, N. Tordo, D. Moras, P. Argos, An attempt to unify the structure of polymerases, Protein Eng. 3 (1990) 461–467.
[11] A. Bernad, L. Blanco, J.M. Lazaro, G. Martin, M. Salas, A conserved 3′-5′ exonuclease active site in prokaryotic and eukaryotic DNA polymerase, Cell 59 (1989) 219–228.
[12] P.D. Gutman, K.W. Minton, Conserved sites in the 5′-3′ exonuclease domain of *Escherichia coli* DNA polymerase, Nucleic Acids Res. 21 (1993) 4406–4407.
[13] J. Ito, D.K. Braithwaite, Compilation and alignment of DNA polymerase sequences, Nucleic Acids Res. 19 (1991) 4045–4057.
[14] A.W. Braithwhite, J. Ito, Compilation, alignment and phylogenetic relationship of DNA polymerases, Nucleic Acids Res. 21 (1993) 787–802.
[15] W.R. Pearson, D.J. Lipman, Improved tools for biological sequence comparison, Proc. Natl. Acad. Sci. USA 85 (1988) 2444–2448.
[16] W.R. Pearson, Rapid and sensitive sequence comparison with FASTP and FASTA, Methods Enzymol. 183 (1990) 63–98.
[17] R. Sousa, Y.J. Chung, J.P. Rose, B. Wang, Crystal structure of bacteriophage T7 RNA polymerase at 3.3 Å resolution, Nature 364 (1993) 593–599.
[18] L.A. Kohlstaedt, J. Wang, J.M. Friedman, P.A. Rice, T.A. Steitz, Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase coupled with an inhibitor, Science 256 (1992) 1783–1790.
[19] J. Otsuka, H. Miyachi, K. Horimoto, Structure model of core proteins in photosystem I inferred from the comparison with those in photosystem II and bacteria; an application of principal component analysis to detect the similar regions between distantly related families of proteins, Biochim. Biophys. Acta 1118 (1992) 194–210.
[20] K. Horimoto, H. Yamamoto, K. Yanagi, K. Ohshima, J. Otsuka, A simple procedure for assigning a sequence motif with an obscure pattern: an application to the basic/helix-loop-helix motif, Protein Eng. 7 (1994) 1433–1440.
[21] J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res. 22 (1994) 4673–4680.
[22] R. Grantham, Amino acid difference formula to help explain protein evolution, Science 185 (154) (1974) 862–864.
[23] Y. Nozaki, C. Tanford, The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale, J. Biol. Chem. 246 (7) (1971) 2211–2217.
[24] R. Barker, Organic Chemistry of Biological Compounds, Prentice-Hall, Englewood Cliff, NJ, 1971.
[25] M.G. Kendall, J.D. Gibson, Rank Correlation Methods, 5th edn., Edward Arnold, London, 1990.
[26] R.E. Dickerson, Cytochrome c and the evolution of energy metabolism, Sci. Am. 242 (1980) 137–153.
[27] A. Kornberg, T.A. Barker, DNA Replication, 2nd edn., W.H. Freeman and Co., New York, 1992.
[28] L.S. Beese, T.A. Steitz, Structural basis for the 3′-5′ exonuclease activity of *Escherichia coli* DNA polymerase I: a two metal ion mechanism, EMBO J. 10 (1991) 25–33.
[29] A. Morrison, J.B. Bell, T.A. Kunkel, A. Sigino, Eukaryotic DNA polymerase amino acid sequence required for 3′-5′ exonuclease activity, Proc. Natl. Acad. Sci. USA 88 (1991) 9473–9477.
[30] Y. Ishino, H. Iwasaki, I. Kato, H. Shinagawa, Amino acid sequence motifs essential to 3′-5′ exonuclease activity of *Escherichia coli* DNA polymerase II, J. Biol. Chem. 269 (1994) 14655–14660.
[31] L. Aravind, E.V. Koonin, Phosphoesterase domains associated with DNA polymerases of diverse origins, Nucleic Acids Res. 26 (1998) 3746–3752.
[32] E. Zaychikov, E. Martin, L. Denissova, M. Kozlov, V. Markovtsov, M. Kashlev, H. Heumann, V. Nikiforov, A. Goldfarb, A. Mustaev, Mapping of catalytic residues in the RNA polymerase active center, Science 273 (1996) 107–109.
[33] D.R. Kim, A.E. Prichard, C.S. McHenry, Localization of the active site of the alpha subunit of the *Escherichia coli* DNA polymerase III holoenzyme, J. Bacteriol. 179 (1997) 6721–6728.
[34] A.E. Prichard, C.S. McHenry, Identification of the acidic residues in the active site of DNA polymerase III, J. Mol. Biol. 285 (1999) 1067–1080.
[35] S. Francesconi, M.P. Longhese, A. Piseri, C. Santocanale, G. Lucchini, P. Plevani, Mutations in conserved yeast DNA primase domains impair DNA replication in vivo, Proc. Natl. Acad. Sci. USA 88 (1991) 3877–3881.
[36] H.P. Klenk, R.A. Clayton, J.F. Tomb, O. White, K.E. Nelson, K.A. Ketchum, R.J. Dodson, M. Gwinn, E.K. Hickey, J.D. Peterson, D.L. Richardson, A.R. Kerlavage, D.E. Graham, N.C. Kyrpides, R.D. Fleischman, J. Quackenbush, N.H. Lee, G.G. Sutton, S. Gill, E.F. Kirkness, B.A. Dougherty, K. Mckenney, M.D. Adams, B. Loftus, J.C. Venter, The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*, Nature 390 (1997) 364–370.
[37] C.J. Bult, O. White, G.J. Olsen, L. Zhou, R.D. Fleischmann, G.G. Sutton, J.A. Blake, L.M. FitzGerald, R.A. Clayton, J.D. Gocayne, A.R. Kerlavage, B.A. Dougherty, J.F. Tomb, M.D. Adams, C.I. Reich, E.F. Overbee, K.G. Weinstock, J.M. Merrick, A. Glodek, J.L. Scott, N.S.M. Geoghagen, J.C. Venter, Complete genome sequence of the

methanogenic archaeon, *Methanococcus jannaschii*, Science 273 (1996) 1058–1073.
[38] D.R. Smith, L.A. Doucette-Stamm, C. Deloughery, H. Lee, J. Dubois, T. Aldredge, R. Bashirzadeh, D. Blakely, R. Cook, K. Gilbert, D. Harrison, L. Hoang, P. Keagle, W. Lumm, B. Pothier, D. Qiu, R. Spadafora, R. Vicaire, Y. Wang, J. Wierzbowski, R. Gobson, N. Jiwani, A. Caruso, D. Bush, J.N. Reeve, Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics, J. Bacteriol. 179 (22) (1997) 7135–7155.
[39] Y. Kawarabayasi, M. Sawada, H. Horikawa, Y. Haikawa, Y. Hino, S. Yamamoto, M. Sekine, S. Baba, H. Kosugi, A. Hosoyama, Y. Nagai, M. Sakai, K. Ogura, R. Otsuka, H. Nakazawa, M. Takamiya, Y. Ohfuku, T. Funahashi, T. Tanaka, Y. Kudoh, J. Yamazaki, N. Kushida, A. Oguchi, K. Aoki, H. Kikuchi, Complete sequence and gene organization of the genome of hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3, DNA Res. 5 (2) (1998) 55–76.
[40] W.C. Copeland, T.S.F. Wang, Mutational analysis of the human DNA polymerase α; the most conserved region in α-like DNA polymerases is involved in metal-specific catalysis, J. Biol. Chem. 268 (1993) 11028–11040.
[41] Q. Dong, W.C. Copeland, T.S.F. Wang, Mutational studies of human DNA polymerase α: identification of residues critical for deoxynucleotide binding and misinsertion fidelity of DNA synthesis, J. Biol. Chem. 268 (1993) 24163–24174.
[42] Q. Dong, W.C. Copeland, T.S.F. Wang, Mutational studies of human DNA polymerase α: serine 867 in the second most conserved region among α-like DNA polymerases is involved in primer binding and mispair primer extension, J. Biol. Chem. 268 (1993) 24175–24182.
[43] Q. Dong, T.S.F. Wang, Mutational studies of human DNA polymerase α: lysine 950 in the third most conserved region of α-like DNA polymerases is involved in binding the deoxynucleoside triphosphate, J. Biol. Chem. 270 (1995) 21563–21570.
[44] W.T. McAllister, C.A. Raskin, Micro review: the phage RNA polymerases are related to DNA polymerases and reverse transcriptases, Mol. Microbiol. 10 (1993) 1–6.
[45] R.E.F. Matthews, Classification and nomenclature of viruses. Fourth report of the international committee on taxonomy of viruses, Intervirology 17 (1-3) (1982) 1–199.
[46] R. Sousa, D. Patra, E.M. Lafer, Model for the mechanism of bacteriophage T7 RNAP transcription initiation and termination, J. Mol. Biol. 224 (1992) 319–334.
[47] L. Gross, W.J. Chen, W.T. McAllister, Characterization of bacteriophage T7 RNA polymerase by linker insertion mutagenesis, J. Mol. Biol. 228 (1992) 488–505.
[48] A. Jacobo-Molina, J. Ding, R.G. Nanni, A.D. Clark Jr., X. Lu, C. Tantillo, R.L. Williams, G. Kamer, A.L. Ferris, P. Clark, A. Hizi, S.H. Hughes, E. Arnold, Crystal structure of human immunodeficiency virus type 1 reverse transcriptase complexed with double-stranded DNA at 3.0 Å resolution shows bent DNA, Proc. Natl. Acad. Sci. USA 90 (1993) 6320–6324.
[49] C.M. Counter, M. Meyerson, E.N. Eaton, R.A. Weinberg, The catalytic subunit of yeast telomerase, Proc. Natl. Acad. Sci. USA 94 (1997) 9202–9207.
[50] G. Martin, W. Keller, Mutational analysis of mammalian poly (A) polymerase identifies a region for primer binding and a catalytic domain, homologous to the family *X* polymerase and to other nucleotidyltransferases, EMBO J. 15 (1996) 2593–2603.
[51] H. Pelletier, M.R. Sawaya, A. Kumar, S.H. Wilson, J. Kraut, Structures of ternary complexes of rat DNA polymerase β, a DNA template-primer, and ddCTP, Science 264 (1994) 1891–1903.
[52] J.J. Harrington, M.R. Lieber, The characterization of a mammalian DNA structure-specific endonuclease, EMBO J. 13 (1994) 1235–1246.
[53] J.M. Murray, M. Tavassoli, R. al-Harithy, K.S. Sheldrick, A.R. Lehman, A.M. Carr, F.Z. Watts, Structural and functional conservation of the human homology of the *Schizosaccharomyces pombe* rad2 gene, which is required for chromosome segregation and recovery from DNA damage, Mol. Cell. Biol. (1994) 4878–4888.
[54] Y. Ishino, A. Takahashi-Fujii, T. Uemori, M. Imamura, I. Kato, H. Doi, The amino acid sequence required for 5′→3′ exonuclease activity of *Bacillus caldotenax* DNA polymerase, Protein Eng. 8 (1995) 1171–1175.
[55] Y. Kim, S.H. Eom, J. Wang, D.-S. Lee, S.W. Suh, T.A. Steitz, Crystal structure of *Thermus aquaticus* DNA polymerase, Nature 376 (1995) 612–616.
[56] T.A. Ceska, J.R. Sayers, G. Stier, D. Suck, A helical arch allowing single-stranded DNA to thread through T5 5′-exonuclease, Nature 382 (1996) 90–93.
[57] R. Sousa, Structural and mechanistic relationships between nucleic acid polymerases, Trends Biochem. Sci. 21 (1996) 186–190.
[58] J. Wang, A.K.M.A. Scatter, C.C. Wang, J.D. Karam, W.H. Konigsberg, T.A. Steitz, Crystal structure of a pol α family replication DNA polymerase from bacteriophage RB69, Cell 89 (1997) 1087–1099.
[59] M.R. Sawaya, H. Pelletier, A. Kumar, S.H. Wilson, J. Kraut, Crystal structure of rat DNA polymerase beta: evidence for a common polymerase mechanism, Science 264 (1994) 1930–1935.
[60] J.F. Davies, R.J. Almassy, Z. Hostomska, R.A. Ferre, Z. Hostomsky, 2.3 Å crystal structure of the catalytic domain of DNA polymerase beta, Cell 76 (1994) 1123–1133.
[61] J. Sakon, H.H. Liao, A.M. Kanikula, M.M. Benning, I. Rayment, H.M. Holden, Molecular structure of kanamycin nucleotidyltransferase determined to 3.0-Å resolution, Biochemistry 32 (1993) 11977–11984.
[62] L. Holm, C. Sander, DNA polymerase β belongs to an ancient nucleotidyltransferase superfamily, Trends Biochem. Sci. 20 (1995) 345–347.