A common set of conserved motifs in a vast variety of putative nucleic acid-dependent ATPases including MCM proteins involved in the initiation of eukaryotic DNA replication

Eugene V.Koonin*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bldg. 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received April 1, 1993; Revised and Accepted May 6, 1993

ABSTRACT

A new superfamily of (putative) DNA-dependent ATPases is described that includes the ATPase domains of prokaryotic NtrC-related transcription regulators, MCM proteins involved in the initiation of eukaryotic DNA replication, and a group of uncharacterized bacterial and chloroplast proteins. MCM proteins are shown to contain a modified form of the ATP-binding motif and are predicted to mediate ATP-dependent opening of double-stranded DNA in the replication origins. In a second line of investigation, it is demonstrated that the products of unidentified open reading frames from *Marchantia* mitochondria and from yeast, and a domain of a baculovirus protein involved in viral DNA replication are related to the superfamily III of DNA and RNA helicases that previously has been known to include only proteins of small viruses. Comparison of the multiple alignments showed that the proteins of the NtrC superfamily and the helicases of superfamily III share three related sequence motifs tightly packed in the ATPase domain that consists of 100–150 amino acid residues. A similar array of conserved motifs is found in the family of DnaA-related ATPases. It is hypothesized that the three large groups of nucleic acid-dependent ATPases have similar structure of the core ATPase domain and have evolved from a common ancestor.

INTRODUCTION

Numerous nucleic acid-dependent ATPases play a pivotal role in genome replication, expression, and repair (1). It is common for different enzymes of this type to mediate consecutive steps in a single process (e.g. DNA replication initiation), indicative of fine-tuned functional diversification. A large fraction of DNA-dependent and RNA-dependent ATPases possess helicase activity, i.e. unwind DNA or RNA duplexes concomitantly with ATP hydrolysis (2–4). Other DNA-dependent ATPases, e g. bacterial DnaA protein, do not have bona fide helicase activity but facilitate the so-called opening of DNA duplexes that is a prerequisite for binding of other proteins, including helicases (1,5,6).

Computer-assisted comparative analysis of amino acid sequences has had significant impact on the identification and functional characterization of the nucleic acid-dependent ATPases. The vast majority of these proteins contain the so-called Walker-type purine NTP-binding pattern that consists of two distinct motifs, 'A' and 'B' (7,8). A number of groups of ATPases, particularly DNA and RNA helicases, could be delineated based on unique sets of additional conserved motifs (4, 9–14).

An important, and generally unresolved issue in computer-assisted sequence analysis is the identification of proteins that contain modified but functionally active forms of conserved motifs. Here I report on the identification of a group of eukaryotic proteins involved in DNA replication initiation as putative DNA-dependent ATPases with a modified 'A' motif of the Walker pattern and propose a common arrangement of conserved motifs for three large groups of DNA-dependent and RNA-dependent ATPases.

METHODS

Amino acid sequences were from the GenBank database (Release 74.0) as updated at the National Center for Biotechnology Information (NCBI), NIH. Database searches for sequence similarity were performed using programs based on the BLAST algorithm (15) and the BLOSUM62 matrix for comparison of amino acid residues (16). BLASTP program compares an amino acid sequence with the amino acid sequence databases; TBLASTN program compares an amino acid sequence with the conceptual translation of the nucleotide databases in all six reading frames; and BLAST3 is a modification of BLASTP generating three-way alignments. Multiple alignments of amino acid

* On leave from Institute of Microbiology, Russian Academy of Sciences, 117811 Moscow, Russia

sequences were generated using the programs MACAW (17), CLUSTALV (18), and OPTAL (19). Database search for sequence motifs was performed using programs DBSITE and FPAT (NCBI). Protein secondary structure prediction was performed using the Garnier algorithm (20).

MCM proteins involved in eukaryotic DNA replication initiation contain a modified form of the NTP-binding pattern and belong to a single superfamily with two families of prokaryotic ATPases

Initiation of chromosomal DNA replication in eukaryotes is mediated by the DNA polymerase a holoenzyme, which appears to be an analog of the bacterial primosome. The holoenzyme contains a number of poorly characterized 'auxiliary' proteins, in addition to the catalytic subunits of the DNA polymerase and primase (21,22). Recently a group of related yeast proteins designated MCM (MiniChromosome Maintenance) has been characterized, which appeared to be involved in the initiation of chromosomal DNA replication in an origin-dependent manner (23–26). Experimentally, this was manifest by the differential effect of different *mcm* mutations on the maintenance of minichromosomes containing different ARSs (Autonomously Replicating Sequences). It has been hypothesized that MCM proteins interact directly with ARSs (24,25). Also, it has been shown that the previously characterized CDC46 gene, which is essential for cell division, is identical to MCM5 (26). A mammalian homologue of MCM proteins has been recently identified and it has been shown that this protein (P1) is a loosely bound component of the DNA polymerase α holoenzyme (27).

MCM proteins and P1 show significant sequence conservation, which is particularly pronounced in the so-called region II, a 200 amino acid residue domain located roughly in the middle of these large proteins (24–27). It has been claimed that the amino acid sequences of MCM proteins and P1 lacked any known nucleic acid or nucleotide-binding motifs (27). However, inspection of the alignments suggested that they did contain a conserved sequence resembling the A motif of the Walker type NTP-binding pattern, with the Gly in the GKS(T) signature substituted by Ala or Ser. Screening of the sequence databases for similarity to MCM proteins revealed moderate similarities with two groups of (putative) prokaryotic ATPases, namely transcription regulators related to NtrC protein, and bacterial (or chloroplast) proteins containing the NTP-binding motif and related to Mg^{2+} chelatase from *Rhodobacter capsulatus*. Although not striking, the similarity in some protein pairs was statistically significant. For instance, the probability of random matching between the sequences of CDC46 and *E.coli* chelatase-related protein YifB was computed to be below 10^{-4}. The regions of the highest similarity between MCM proteins and both groups of bacterial proteins centered at the 'A' and 'B' motifs of the NTP-binding pattern. Multiple alignment of the three proteins groups constructed using the MACAW program revealed four conserved motifs shared by the MCM proteins with each of the two groups of (putative) bacterial ATPases (Fig. 1). In each of these motifs the alignment was statistically highly significant, with the probability of random matching in all sequences below 10^{-13}. In an independent statistical test, the pairwise alignment of the sequences of CDC46 and YifB using the OPTAL program scored 9.5 standard deviations above the random expectation, which is an evidence of significant similarity. The signature from motif II, **$Gx\&x_2[AS]x_2Gx[UC][\&C]UDE[\&C]x_2\&x_7U$** (U—a bulky aliphatic residue, i.e. I, L, V, M; &—a bulky hydrophobic residue, i. e. I, L, V, M, F, Y, W; x— any residue; brackets enclose the residues, all of which are allowed in the given position), which is conserved in the MCM proteins, the proteins of the chelatase family, and the vast majority of the proteins of

```
                                      A
CHEL  Rc     Z11165  42   PGIGGVLVFGDRGTGKSTAVRALAALLPEIEAVEGCPVSS 47
YifB  Ec     M87049 217   AGGHNLLLIGPPGTGKTMLASRINGLLPDLSNEEALESAA 32
C2    At     X51799 109   PKIGGVMIMGDRGTGKSTTVRSLVDLLPEINVVA--GD-- 58
C2    Eg     X65484  32   PKIGGVMIMGDRGTGKSTIVRALVDLLPPIDVIE--ND-- 58
C2    Ol     Z21959  33   PKIGGVMIMGDRGTGKSTTIRAIADLLPEIEVVK--DN-- 45

CDC21 yeast  X58824 535   RGDINILMCGDPSTSKSQILKYVHKIAPRGVYTSGKGSSA 12
CDC46 yeast  A39631 406   RGDINVLLLGDPGTAKSQLLKFVEKVSPIAVYTSGKGSSA 13
MCM2  yeast  X53539 533   RGDINVLLLGDPGTAKSQILKYVEKTAHRAVFATGQGASA 13
MCM3  yeast  X53540 399   RGDINILMVGDPSTAKSQLLRFVLNTASLAIATTGRGSSG 13
P1DP  human  X62154 318   RGDINILLIGDPSVAKSQLLRYVLCTAPRAIPTTGRGSSG 13

NifA  Rl     X16311  30   RSNSPVLLRGESGTGKEFFAKLIHDSSSRHEKPFVKLNC- 21
HoxA  Ae     M64593 183   RYDLPVMVLGESGTGKELLARAIHYASPRAARAFVSENCA 20
xylR  Pp     M10143 253   RGRVSVLLLGETGVGKEVIARSVHLRSERAEQPFVAVNCA 20
NtrC  Kp     M19277 158   RSSISVLINGESGTGKELVAHALHRHSPRAKAPFIALNMA 20
VnfA  Av     M26752 228   RTRTTVLILGESGVGKELVAGAIHYNSPAAKGPFVKFNC- 21
cons                      .....UUU.Gd.GTGKS.&...U.................
                                    e SVA T
                                          S E

                     B                                             C
CHEL   GEKAFEPGLLARANRGYLYIDECNLLEDHIVDLLLDVAQSGENVVERDGLSIRHPARFVLVGSGNP
YifB   GGAIPGPGEISLAHNGVLFLDELPEFERRTLDALREPIESGQIHLSRTRAKITYPARFQLVAAMNP
C2At   GVKAFEPGLLAKANRGILYVDEVNLLDDHLVDVLLDSAASGWNTVEREGISISHPARFILIGSGNP
C2Eg   GKKAFEPGLLAQANRGILYVDEVNLLDDHLVDVLLDSAASGWNTVEREGVSICHPARFILVGSGNP
C2Ol   GVKAFEPGLLAKANRGILYVDEVNLLDDHLVDILLDSAASGLNTVEREGISIRHAARFVLVGSGNP

CDC21  KQLVLESGALVLSDGGICCIDEFDKMSDATRSILHEVMEQQTVTVAKAGIITTLNARTSILASANP
CDC46  REFYLEGGAMVLADGGVVCIDEFDKMRDEDRVAIHEAMEQQTISIAKAGITTVLNSRTSVLAAANP
MCM2   KEWTLEGGALVLADKGVCLIDEFDKMNDQDRTSIHEAMEQQSISISKAGIVTTLQARCSIIAAANP
MCM3   GERRLEAGAMVLADRGVVCIDEFDKMTDVDRVAIHEVMEQQTVTIAKAGIHTTLNARCSVIAAANP
P1DP   GERRLEAGAMVLADRGVVCIDEFDKMSDMDRTAIHEVMEQGRVTIAKAGIHARLNARCSVLAAANP

NifA   GATSQKEGRFELAHGGTLLLDEIGEISAEFQAKLLRVLQEGELERVGGTRTLKVNVR--LVCATNK
HoxA   GAYEDHAGLFQRANGGTIFLDEIGDTSPAFQVKLLRVLQEGEVRPVGSPRWIPVDVRVIAATHCNL
xylR   GAVNARAGRFERANGGTIFLDEVIELTPRAQATLLRVLQEGELERVGGDRTRKVDVR--LITATNE
NtrC   GANTVRQGRFEQADGGTLFLDEIGDMPLDVQTRLLRVLADGQFYRVGGYAPVKVDVR--IIAATHQ
VnfA   GAIGLRKGRFEEAAGGTIFLDEVGEMSLTTQAKLLRVLQERSFERVGGNTTIHVDLRVIAATNRNL
cons   .......G.&..A..G&U&UDE&..&.......U...U..................R..UU...N.
                   S   TCC                  A                     AA   H

                    D
CHEL   0   EEGDLRPQLLDRFGLSVEVLSPRDV 130
YifB  12   YLNRLSGPFLDRFDLSLEIPLPPPG  99
C2At   0   EEGELRPQLLDRFGMHAQVGTVRDA 129
C2Eg   0   EEGELRPQLLDRFGMHAQIKTLKEP 136
C2Ol   0   EEGELRPQLLDRFGMHAVIKTVKDP 129

CDC21 12   KNIDLPPTLLSRFDLVYLILDRVDE 221
CDC46 12   DNIDFQTTILSRFDMIFIVKDDHNE 213
MCM2  12   QNVSLTEPILSRFDILCVVRDLVDE 201
MCM3  12   QNIALPDSLLSRFDLLFVVTDDINE 416
P1DP  12   ENIGLQDSLLSRFDLLFIMLDQMDP 321

NifA   6   AAGEFRADLYYRINVVPITLPPLRQ 168
HoxA  88   SGEEIRAQLFSRKVLHGQPGTVKRG  61
xylR   6   KMGRFRADLFFRLNVFPVHIPPLRE 158
NtrC   6   QEGKFREDLFHRLNVIRVHLPPLRE 156
VnfA  48   MGIEVNRISTPRLNMLQSYQWPGNV  95
cons       ....&...u&.R&.U..........
```

Figure 1. Multiple alignment of the MCM proteins with two groups of prokaryotic ATPases. Only the conserved regions are shown, with their boundaries determined using the MACAW program so as to achieve the maximal statistical significance. The length of the intervening variable regions and the distances from the protein termini are indicated by numbers. C2 are chloroplast proteins. Of the numerous available NifA-related sequences, five relatively dissimilar proteins were chosen. The consensus pattern shows amino acid residues (or groups of two or three related residues) that are conserved in all aligned sequences (upper case) or in all but one sequences (lower case). U indicates a bulky aliphatic residue (I, L, V, M), & indicates a bulky hydrophobic residue, aromatic or aliphatic (I, L, V, M, F, Y, W), and dot indicates any residue. The invariant residues are highlighted by bold typing. Each sequence is accompanied by the GenBank accession number. Three additional highly conserved MCM-related genes from yeast and *Xenopus laevis* have been recently identified and partially sequenced using polymerase chain reaction with primers designed based on the motifs from the conserved region described here as the putative ATPase domain (55). Yet another related sequence has been found in the putative product of an uncharacterized *Caenorhabditis elegans* cDNA clone (Kerlavage, A. R.; GenBank T00192). These incomplete sequences are not shown in the Figure. Organism names are abbreviated as follows: CHEL, Mg chelatase; Rc, *Rhodobacter capsulatus*; Ec, *E.coli*; At, *Arabidopsis thaliana*; Ol, *Olisthodiscus luteus*; Rl, *Rhizobium leguminosporum*; Ae, *Alcaligenes eutrophus*; Pp, *Pseudomonas putida*; Kp, *Klebsiella pneumoniae*; Av, *Azotobacter vinelandii*.

the NtrC family (Fig. 1, and data not shown), was found to be unique upon screening of the available sequence databases. Another descriptor with a similar specificity could be constructed by combining the signatures from motifs I, II, and III in the following regular expression **UUUxGx$_2$[GS][TV][GAS]K[STE]x$_n$Gx$_3$UDE[&C]x$_n$UUx[AS]xN**. Each of these regular expressions is violated only in a few NtrC-related protein, some of which may have lost the ATPase activity (see below). These findings suggest that MCM proteins and the two related groups of bacterial proteins comprise a new superfamily of (putative) DNA-dependent ATPases.

The presence of the putative NTPase domain conserved from yeast to mammals indicates that MCM proteins may be involved in an ATP-consuming step in the initiation of DNA replication in all types of eukaryotic cells. The ATP-dependence of the assembly of the yeast replication initiation complex has been demonstrated (28,29). The similarity between MCM proteins and NtrC-like proteins may be indicative of a direct functional analogy between these proteins, which are involved in the initiation of replication and transcription, respectively. Recently it has been shown that NtrC is a DNA-stimulated ATPase and that ATP hydrolysis is required for DNA duplex opening, which is a prerequisite of transcription activation (30,31). MCM proteins may act similarly in the initiation of eukaryotic DNA replication.

The functional implications of the even more notable similarity with the chelatase-related family remains unclear as these prokaryotic proteins have not been studied in any detail. It remains to be determined experimentally whether these proteins also are DNA-dependent ATPases or the observed sequence conservation is due solely to structural and functional analogies in the ATPase domain as such. The latter possibility cannot be ruled out as illustrated by the striking sequence similarity between the ATPase domains of transcription termination factor Rho, which is a DNA-RNA helicase, and proton-translocating ATPases (4, 32).

```
HELICASE M.p. MIT (28-302) VKL-GPQK--LKKCIIVFKDAVLDTMTGRVEEFSPDRF
                           :**  **     : :* *:::****  *     **  :
PRIMASE  P4      (399-679) LKLIIPQQEAPSRRLIGFRNGVLDTQNGTFHPHSPSHW

HELICASE M.p. MIT CNAKLPYNIGMSQLEDIPCPDIPGDLCPTFTEFLDSFTGGKDDLKKF
                     :   ::  :   *    :      * *  :**   **: : :
PRIMASE  P4       --MRTLCDVDFTPPVD---GETLETHAPAFWRWLDRAAGGRAEKRDV

                                          A
HELICASE M.p. MIT IRAYFNHLLRSDNLSQR FLYMMGPTGTGKS VFSLVSEVLVGSINT
                  : *    :*      *  ** : ** *:*** :   :  :* *  *
PRIMASE  P4       ILAALFMVLANRYDWQL FLEVTGPGGSGKS IMAEIATLLAGEDNA

                                           B
HELICASE M.p. MIT CHTTLARMNQPWGLISSNLKKM LIVVND SPFYKGDTAILRQLVGG
                     *:  :: *            ** : *    : ** * *: : **
PRIMASE  P4       TSATIETLESPRE-RAALTGFS LIRLPD QEKWSGDGAGLKAITGG

                                          C
HELICASE M.p. MIT DRISCKLKHANVRHEFSYSG WVLIVGN EYLGMSETSGALARRMIV
                  * :*   *  :           :* * * : :  :: **:: ** ::
PRIMASE  P4       DAVSVDPKYRDA-YSTHIPA VILAVNN NPMRFTDRSGGVSRRRVI

HELICASE M.p. MIT --FPARNAVHLKKFLIKE----EHGLFMGPLAEEISE-IAKWALSMP
                    **   *   :   :*:    * :: :  * :  *: :    *
PRIMASE  P4       IHFPEQIAPQERDPQLKDKITRELAVIVRHLMQKFSDPMLARSLLQS

HELICASE M.p. MIT DQDATE---IMRDPEIHCPSL
                  :::  *   * ** :     :
PRIMASE  P4       QQNSDEALNIKRDADPTFDFI
```

Figure 2. Amino acid sequence alignment of the putative helicase domain of bacteriophage primase and concatenated products of ORF180 and ORF167 from *Marchantia polymorpha* mitochondrion.The positions of the aligned regions in the protein sequences are shown in parentheses. Asterisks show identical amino acid residues and colons show similar residues. 'A', 'B', and 'C' motifs are boxed. The overlap between ORF180 and ORF167 where the frameshift is supposed to be located and the actual encoded sequence is uncertain is underlined. Amino acid sequence of the putative helicase domain of bacteriophage phiR173 primase (56) was identical to the P4 sequence.

Cellular homologues of viral helicases

Comparative analysis of the amino acid sequences of DNA and RNA helicases has revealed well defined conserved motifs and allowed the delineation of several distinct (super)families (4; 8–14). The two largest groups of helicases, provisionally designated superfamilies I and II, each include DNA and RNA helicases encoded by prokaryotic, eukaryotic, DNA viral, and RNA viral genes (9–12).

In contrast, superfamily III has been shown to consist entirely of (putative) helicases of small viruses, namely positive-strand RNA picornaviruses, caliciviruses, comoviruses, and nepoviruses, single-stranded DNA parvoviruses and geminiviruses, and double-stranded DNA papovaviruses and bacteriophage P4 (13). Recently a protein related to the NS1 protein of parvoviruses has been discovered in human herpesvirus 6 (HSV6), a large DNA virus (33). The relatively close similarity between this protein and NS1 of the adeno-associated virus (AAV) and the absence of a homologue in the other herpesviruses suggested recombinational origin for the putative helicase of HSV6 (33).

The DNA and RNA helicase activity has been explored in detail for the T antigen of SV40 papovavirus (34) and has been demonstrated also for parvovirus NS1 protein (35,36), and for the P4 bacteriophage primase (37). For 2C proteins of picornaviruses, the importance of the conserved helicase motifs for virus reproduction has been proved by site-directed mutagenesis (38,39), and very recently the RNA-dependent ATPase activity has been demonstrated (Mirzayan, C. and Wimmer, E. Abstracts of the Third International Symposium on Positive Strand RNA viruses. Clearwater, FL, 1992, Abstract P4–28).

The helicases of superfamily III have only three conserved sequence motifs tightly packed in an approximately 100 amino acid domain (13). Two of these are specific versions of the 'A' and 'B' motifs of the purine NTP-binding pattern, and only the third C-proximal motif ('C') has been found to be specific for this superfamily. The scarcity of conserved motifs made it difficult to derive a unique sequence signature for this superfamily although the multiple alignments between its members were statistically significant (13).

In an attempt to identify possible cellular members of the helicase superfamily III, I performed systematic database searches for sequence similarity to these viral proteins. Highly significant similarity (probability of matching by chance 6.2×10^{-5} as computed using the BLASTP program) was revealed between the helicase domain of bacteriophage P4 primase and the product of an unidentified open reading frame consisting of 180 codons (ORF180) in *Marchantia polymorpha* mitochondrial genome (40). ORF180 aligned only with the N-terminal portion of the helicase domain of the phage primase, terminating closely after the 'A' motif. Inspection of the sequence of the downstream overlapping ORF167 revealed counterparts to motifs 'B' and 'C', suggesting that these two ORFs probably encode a single protein and have been separated by an artifactual frameshift. Alignment of the P4 helicase domain with the concatenated products of ORF180 and ORF167 (Fig. 2) generated by OPTAL program scored 11.9 standard deviations (SD) above the random expectation, which is indicative of a genuine evolutionary and

```
                             A
NS1  MVM   J02275  465 VLFHGPASTGKSIIAQAI--AQAVGNV-GCYNA-ANVNFPF---NDCT
NS1  ADV   M20036  455 IWFYGPGGTGKTLLASLI--CKATVNY-GMVTT-SNPNFPW---TDCG
NS1  AAV   J01901  328 IWLFGPATTGKTNIAEAI--AHTVPFY-GCVNW-TNENFPF---NDCV
REP  HSV6  D11134  333 VSFIGPPGCGKSMLTGAI--LENIPLH-GILHG-SLNTKNL---RAYG
HEL  M.p.  M68929  127 LYMMGPTGTGKSVFSLVS--EVLVGSINTCHTTLARMNQPWGLISSNL
PRIM P4    X05623  496 LEVTGPGGSGKSIMAEIA--TLLAGEDNATSATIETLESPRE--RAAL
p143 ACNPV M57687  912 IYMPGEPGSGKSSFFELLDYLVLMHKFDDDNHSGESNKETSDKEVSKL
cons                   U.&.Gp...GKS.&..U...............................
                                  T    A

                   B                   B'                      C
NS1  MVM    NKN-LIWVEEAGNFGQQVNQFKAICSGQTIRIDQKGKG-SKQIEPTP--VIMTTNE 185
NS1  ADV    NRN-IIWAEECGNFGNWVEDFKAITGGDVKVDTKNKQ-PQSIK-GC--VIVTSNT  71
NS1  AAV    DKM-VIWWEEGKMTAKVVESAKAILGGSKVRVDQKCKS-SAQIDPTP--VIVTSNT 114
REP  HSV6   QVL-VLWWKDISINFDNFNIIKSLLGGQKIIFPINEND-HVQIGPCP--IIATSCV  64
HEL  M.p.   KKM-LIVVNDSPFYKGDTAILRQLVGGDRISCKLKHAN-VRHEFSYSGWVLIVGNE 115
PRIM P4     TGFSLIRLPDQEKWSGDGAGLKAITGGDAVSVDPKYRD-A-YSTHIPAVILAVNNN 183
p143 ACNPV  NSQ-LYTINELKQ-CSESYFKKHADSSKSDSKSRKYQGLLKYEANYK--MLIVNNK 209
cons        ....U&.&.E...........K.U.GG..u....k .............UUUV.n.
                     D           R A SS                        AT
```

Figure 3. Alignment of the helicase domains from phage P4, *Marchantia* mitochondrion, AcNPV and parvoviruses. The designations are as in Fig. 1. Abbreviations: MVM, minute virus of mice; ADV, Aleutian disease (of minks) virus; AAV, adeno-associated virus, type 2; HSV6, herpes simplex virus type 6; Mp, *Marchantia polymorpha*; ACNPV, *Autographa californica* nuclear polyhedrosis virus; HEL, helicase; PRIM, primase.

probably functional relationship. A common unusual feature shared by these two putative helicases was the substitution of the first of the doublet of negatively charged amino acid residues typical of the NTP-binding motif 'B' by either asparagine or proline (Fig. 2). This feature was shared also by the putative helicase of *Autographa californica* nuclear polyhedrosis virus (AcNPV, a baculovirus; ref. 41) that showed statistically significant similarity (7.0 SD) with the phage and mitochondrial helicases. The sequences of these three proteins were compared to other groups of helicases in superfamily III. The closest similarity was observed with the parvovirus helicases, with the alignment with six parvovirus sequences scoring 10.9 SD. A unique pattern, **Ux&xGPx$_3$GK[ST]x$_n$[KR]xUx[GS]Gx$_2$Ux[&C]x$_2$[KN]** could be defined that selectively retrieved from the NRDB the P4 and *Marchantia* sequences together with those of the parvovirus helicases (Fig. 3). The second part of this pattern was derived from an additional conserved motif (B′) that is located between motifs 'B' and 'C' (Fig. 3).

Several proteins showed marginal similarities to the helicases of superfamily III in the region around motif A. Further analysis suggested that one of these, the product of an unidentified ORF located 3′ of the yeast ADE3 gene (hereafter ADE3x), was indeed related to superfamily III helicases. ADE3x sequence scored 9.3 SD when aligned with 23 helicase sequences. Moreover, this sequence shared additional similarity, including an equivalent to the B′ motif, with the putative helicases of geminiviruses; the pattern **U&&xG[PD]x$_2$[ST]GK[ST]x$_2$[AS]x$_2$Ux$_2$Hx$_n$Kx&xGx$_7$[DN]xK&** was shown to be unique for ADE3x and the geminivirus proteins (Fig. 4).

Interestingly, I observed that the putative NTPases encoded by coconut foliar decay virus (CFDV) and banana bunchy top virus (BBTV), unclassified plant viruses with a very small ssDNA genome (42,43), are also related to this group of putative helicases, some deviations in the conserved motifs notwithstanding (Fig. 4).

These observations showed that analogously to the two other superfamilies of helicases, superfamily III includes both viral and cellular proteins.

```
                               A
ORF1 CFDV  M29963 176  LWICG-RDGGDGKSVFAKYLGLKPD-WFYT-CGGTRKDV--LYQYIEDP
ORF1 BBTV  NA     176  IWVYGPNGGGEGKTTYAKHLMKTRN-AFYS-PGGKSLDICRLYNY-EDI
AL1 ABMV   X15983 216  LIVEG--DSRTGKTMWARALGPH---NYLS-GHLDFNGR----VYSNEV
AL1 BGMV   M10070 217  IIVEG--DSRTGKTMWARALGPH---NYLS-GHLDFNSL----VYSNSV
AL1 TGMV   K02029 220  IIIEG--DSRTGKTMWARSLGPH---NYLS-GHLDLNSR----VYSNKV
AL1 CLV    J02057 217  IVIEG--DSRTGKTIWARSLGPH---NYLC-GHLDLSPK----VFNNAA
AL1 BCTV   M24597 216  IIVEG--DSRTGKTMWARSLGAH---NYIT-GHLDFSPR----TYYDEV
AL1 CSMV   M20021 245  LYICG--PTRTGKTSWARSLGTH---HYWQ-HSVNFLE-----EWNCQA
AL1 MSV    X01089 230  LYIVG--PTRTGKSTWARSLGVH---NYWQ-NNVDWS------SYNEDA
AL1 WDV    X02869 225  IYICG--PTRTGKTSWARSLGTH---NYYN-SLVDFT------TYDVNA
ADE3X Sc   M12878  33  IFFSG--PQGSGKSFTSIQIYNHLMEKYGGEKSIGY-------ASIDDF
cons                   U&&.G  D..tGKT.@Ar.U..h....@&....u.&........&....
                              P  s  S  Sk

                B                  B'                     C
ORF1 CFDV  KRNLILDVPRCNLEYLNYALLECVKNRAFSSDKYEPLSYLGFDHVHVLVFANV  17
ORF1 BBTV  ---VIFDIPRCKEDYLNYGLLEEFKNGIIQSGKYEPVLKI-VEYVEVIVMANF  17
AL1 ABMV   EYNVIDDVAPHYLKLKHWKELLGAQKDWQSNCKLAKPVQIK-GGIRAIVLCNP  48
AL1 BGMV   EYNVIDDITPNYLKLKDWKELIGEQKDWQSNCKYGKPVQIK-GGIPSIVLCNP  45
AL1 TGMV   EYNVIDDVTPQYLKLKHWKELIGAQRDWQTNCKYGKPVQIK-GGIPSIVLCNP  43
AL1 CLV    WYNVIDDVDPHY--LKHFKEFMGSQRDWQSNTKYGKPVQIK-GGIPTIFLCNP  54
AL1 BCTV   EYNVIDDVDPTYLKMKHWKHLIGAQKEWQTNLKYGKPRVIK-GGIPCIILCNP  50
AL1 CSMV   QFNIIDDIPFKF--VPCWKGLVGSQYDLTVNPKYGKKKRIP-NGIPCIILVNE  55
AL1 MSV    IYNIVDDIPFKF--CPCWKQLVGCQRDFIVNPKYGKKKKVQKKSKPTIILANS  43
AL1 WDV    KYNIIDDIPFKF--TPNWKCFVGAQRDFIVNPKYGKRKVIR-GGIPCIILVNP  46
ADE3X Sc   -YLTHEDQLKLNEQFKNNKLLQGRGLPGTHDMKLL--QEV----LNTIFNNME 171
consensus2 .@nuudDu.........@k.&.g......n.K&.....U....u..U&&.N.
                e
```

Figure 4. Alignment of the amino acid sequences of yeast Ade3x protein and putative helicase domains of geminiviruses. The designations are as in Fig. 1; @ indicates an aromatic residue (F, Y, W). Abbreviations: CFDV, coconut foliar decay virus; BBTV, banana bunchy top virus; ABMV, abutilon mosaic virus; BGMV, bean golden mosaic virus; TGMV, tomato golden mosaic virus; CLV, cassava latent virus; BCTV, beet curly top virus; CSMV, *Chloris* striate mosaic virus; MSV, maize streak virus; WDV, wheat dwarf virus; Sc, *Saccharomyces cerevisiae*. NA, accession number not available; the BBTV sequence was from ref. 43.

The NtrC superfamily, the helicase superfamily III and the DnaA family contain similar sets of conserved motifs

Database searches using the BLAST programs showed marginal similarities between some of the proteins of the NtrC superfamily, the helicase superfamily III, and the family of DnaA-related DNA-dependent ATPases (44,45). Comparison of the multiple alignments of these three groups of DNA(RNA)-dependent ATPases revealed surprisingly similar arrangement of the conserved motifs. In each case the 'A', 'B', and 'C' motifs were packed in a compact domain consisting of 100–150 amino acid residues (the 'ABC' domain). Apparently, large inserts between the conserved motifs were not allowed (with the possible exception of the spacer between motifs 'A' and 'B' in the chloroplast chelatase-related proteins; Fig. 1), in contrast to what is observed in the helicase superfamilies I and II (9–12), and in UvrA-related DNA-dependent ATPases (14). In addition to the similar spacing of the conserved motifs, the 'C' motif had a conserved structure, with a hydrophilic residue, usually asparagine, preceded by a hydrophobic stretch probably forming a β-strand (Fig. 5). In the helicase superfamily III, the three principal motifs almost precisely delineated the conserved domain while both the NtrC superfamily and the DnaA family contained a fourth, distal motif (Fig. 1 and refs. 44, 45).

The presence of only three conserved motifs, two of which are common to a wide variety of NTPases, precluded identification of the three groups of the DNA-dependent ATPases as a single class in a database search. Nevertheless, comparison of the profiles of amino acid residue conservation in the three motifs and the distance intervals between them revealed remarkable similarity (Fig. 5b).

a

```
                                        A                      B             C
                           bbbbl lltttttaaaaaaaaa      bbbbttaaaa      lbbbbbtt
T    SV40     (422-530) WLFKG-PIDSGKTTLAAALLEL 26 LV/FEDVKGT 44 PGIVTMNE
T    PyV      (569-677) ILFRG-PVNSGKTGLAAALISL 26 VVCFEDVKGQ 44 PCVCTMNE
E1   HPV18    (429-524) LVFCG-PANTGKSYFGMSFIHF 24 VAMLDDATTT 33 PILLTTNI
E1   HPV35    (393-487) ILIYG-APNTGKSLFGMSLMHF 24 IAMLDDATSP 32 TFTYYINI
E1   BPV1     (474-569) LAFIG-PPNTGKSMLCNSLIHF 24 AALVDDATHA 33 PLLVTSNI
E1   CRPV     (427-522) MVFYG-PPNSGKSYFCMSLIRL 24 LALVDDATSA 33 PLLITTNV
NS1  DNVb     ( 70-164) FQIVS-PPSAGKNFFIETVLAF 25 VNYWDEPNFE 31 PVIITANY
NS1  DNVj     (405-495) FLIIS-PPSAGKNFFFDMIFGL 25 VLLWNEPNYE 31 PVIILTNN
NS1  DNVa     (879-985) MVLEG-ITNAGKSLILDNLLAM 24 SILFEEPMIT 31 PTWITTAT
NS1  MVM      (466-557) VLFHG-PASTGKSIIAQAIAQA 23 LIWVEEAGNF 31 PVIMTTNE
NS1  ADV      (456-547) IWFYG-PGGTGKTLLASLICKA 23 IIWAEECGNF 30 CVIVTSNT
NS1  AAV      (329-422) IWLFG-PATTGKTNIAEAIAHT 23 VIWWEEGKMT 31 PVIVTSNT
REP  HSV6     (334-426) VSFIG-PPGCGKSMLTGAILEN 23 VLWWKDISIN 31 PIIATSCV
HEL  Mp       (128-227) LYMMG-PTGTGKSVFSLVSEVL 28 LIVVNDSPFY 33 WVLIVGNE
PRIM P4       (497-593) LEVTG-PGGSGKSIMAEIATLL 27 LIRLPDQEKW 32 VILAVNNN
HEL  ANPV    (913-1012) IYMPG-EPGSGKSSFFELLDYL 30 LYTINELKQC 30 KMLIVNNK
ORF1 CFDV     (177-273) LWICGRDGGDGKSVFAKYLGLK 23 RNLILDVPRC 34 HVLVFANV
AL1  ABMV     (217-307) LIVEG-DSRTGKTMWARALGPH 19 YNVIDDVAPH 33 RAIVLCNP
AL1  BGMV     (218-308) IIVEG-DSRTGKTMWARALGPH 19 YNVIDDITPN 33 PSIVLCNP
AL1  TGMV     (221-311) IIIEG-DSRTGKTMWARSLGPH 19 YNVIDDVTPQ 33 PSIVLCNP
AL1  CLV      (218-306) IVIEG-DSRTGKTIWARSLGPH 19 YNVIDDVDPH 31 PTIFLCNP
AL1  BCTV     (217-307) IIVEG-DSRTGKTMWARSLGAH 19 YNVIDDVDPT 33 PCIILCNP
AL1  CSMV     (246-333) LYICG-PTRTGKTSWARSLGTH 18 FNIIDDIPFK 31 PCIILVNE
AL1  MSV      (231-328) LYIVG-PTRTGKSTWARSLGVH 17 YNIVDDIPFK 32 PTIILANS
AL1  WDV      (226-312) IYICG-PTRTGKTSWARSLGTH 17 YNIIDDIPFK 31 PCIILVNP
ADE3X Sc      ( 34-119) IFFSG-PQGSGKSFTSIQIYNH 19 YLTHEDQLKL 28 NTIFNNNE
2C   PV       (125-224) LLVHG-SPGTGKSVATNLIARA 26 VVIMDDLNQN 35 YVLASTNS
2C   HAV      (150-253) CYLYG-KRGGGKSLTSIALATK 30 VCIIDDIGQN 35 FIIATSNW
2C   FCV      (480-576) YILTG-PPGCGKTTAAQALAKK 27 VCIIDEF-DS 36 YIIMTSNS
2C   RTSV     (363-469) VCMLG-APGVGKSTIAHVVINH 30 VILYDDLGAI 38 YVFSCTNV
p58  CPMV     (164-268) IFFQG-KSRTGKSLLMSQVTKD 29 FVLMDDFAAV 36 FVFVSTNF
p72  TBRV     (211-315) IYLFG-QRHCGKSNFMATLDNA 29 FFHVDDLSSV 35 FIISSSNF

DnaA Ec       (168-270) LFLYG-GTGLGKTHLLHAVGNG 38 ALLIDDIQFF 22 QIILTSDR
DnaA Pp       (207-309) LFLYG-GVGLGKTHLMHAVGNH 38 ALLIDDIQFF 22 QVILTSDR
DnaA Mc       (144-248) LFIYG-ESGMGKTHLLKAAKNY 40 VLIIDDVQFL 22 QLFFSSDK
DnaA Bs       (147-249) LFIYG-GVGLGKTHLMHAIGHY 38 VLLIDDIQFL 22 QIVISSDR
DnaA Ml       (210-312) LFIYG-ESGLGKTHLLHAIGHY 38 ILLIDDIQFL 22 QVVITSDL
IstB Bt       (104-203) IVFLG-PSGVGKTHLATSIGIA 36 LLIIDEIGYL 21 STILTTNI
IstB Ec       (100-199) VILLG-PPGVGKTHLAIALGVK 36 VLILDEIGYL 21 SIILTSNK
OrfB Bst      (101-200) ILFLG-PPGIGKTHLAISIGME 36 VLIIDEMGYL 21 PIILTSNK
DnaC Ec       (102-204) FIFSG-KPGTGKNHLAAAICNE 38 LLVIDEIGVQ 22 PTGMLTNS
ORF311 Bs     (164-265) LYLYG-KFGVGKTFMLAAIANE 36 VLMLDDIGAE 23 PTFFSSNF
B251 SSV1     ( 32-137) AIIFG-KQGTGKTTYALKVAKE 38 IIIFDDAGIW 26 GVIFTT-P

 CHEL   Rc ( 48-196)  VLVFG-DRGTGKSTAVRALAAL 78 YLYIDECNLL 32 VLVGSGNP
 YifB   Ec (223-356)  LLLIG-PPGTGKTMLASRINGL 63 VLFLDELPEF 32 QLVAAMNp
 C2     At (115-270)  VMIMG-DRGTGKSTTVRSLVDL 85 ILYVDEVNLL 32 ILIGSGNP
 C2     Eg ( 38-193)  VMIMG-DRGTGKSTIVRALVDL 85 ILYVDEVNLL 32 ILVGSGNP
 C2     Ol ( 39-180)  VMIMG-DRGTGKSTTIRAIADL 71 ILYVDEVNLL 32 VLVGSGNP
 CDC21  Sp (541-654)  ILMCG-DPSTSKSQILKYVHKI 43 ICCIDEFDKM 32 SILASANP
 CDC46  Sc (412-525)  VLLLG-DPGTAKSQLLKFVEKV 43 VVCIDEFDKM 32 SVLAAANP
 MCM2   Sc (539-652)  VLLLG-DPGTAKSQILKYVEKT 43 VCLIDEFDKM 32 SIIAAANP
 MCM3   Sc (405-518)  ILMVG-DPSTAKSQLLRFVLNT 43 VVCIDEFDKM 32 SVIAAANP
 P1DP   HU (324-437)  ILLIG-DPSVAKSQLLRYVLCT 43 VVCIDEFDKM 32 SVLAAANP
 NifA   Rl ( 36-154)  VLLRG-ESGTGKEFFAKLIHDS 49 TLLLDEIGEI 30 -LVCATNK
 HoxA   Ae (189-309)  VMVLG-ESGTGKELLARAIHYA 50 TIFLDEIGDT 32 IAATHCNL
 HupR1  Rc (189-308)  VLLRG-EPGSGRAQLARAMHYV 49 TLFVAGVEAA 32 ITGAAADL
 HydG   Ec (164-280)  VLIHG-DSAR-KELVARGLHAS 50 T-CLDEIGDI 30 RLIAATHR
 PgtA   St (166-282)  VWFYG-EHGTGRMTGARYLHQL 49 ARLQSLEHRP 29 MTQIACQS
 XylR   Pp (259-377)  VLLLG-ETGVGKEVIARSVHLR 50 TIFLDEVIEL 30 -LITATNE
 NtrC   Kp (164-282)  VLING-ESGTGKELVAHALHRH 50 TLFLDEIGDM 30 -IIAATHQ
 VnfA   Av (234-354)  VLILG-ESGVGKELVAGAIHYN 50 TIFLDEVGEM 32 IAATNRNL
0.5                   &&&&G ESGTGKT&&A..&...    &&&&DE&..&    .&&.TTN.
0.6                   &&& G   GTGKT&&A  &       &&&&DE&  &     &&  TN
0.7                   &&& G   GTGK  &   &        &&&DE&        &&  TN
0.8                   &&& G   GTGK  &   &        &&&DE&        &&
0.9                   &&& G     GK  &   &         &&DE          &
1.0                   & & G      K      &          & E          &
```

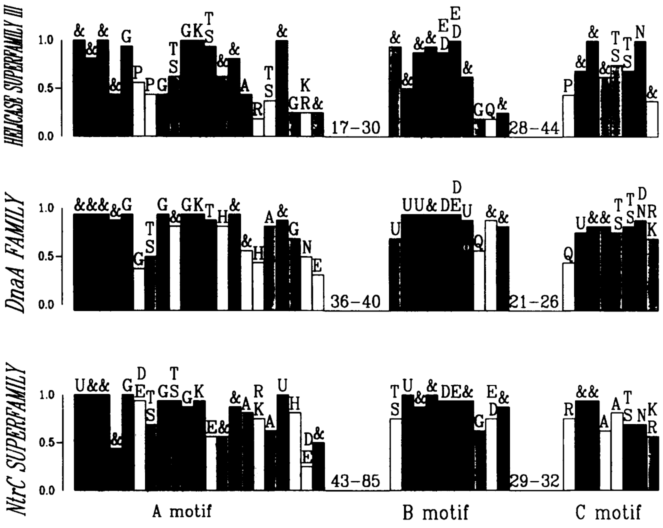

Figure 5. Three large groups of DNA-dependent ATPases contain a similar set of conserved motifs. **(a)** Alignment of the three conserved motifs in NtrC superfamily, helicase superfamily III, and DnaA family. The consensus is shown as the frequency profile calculated over the complete set of 101 sequences (available upon request). In the consensus, pairs of related residues are represented by one with the higher frequency (e.g. E stands for E *or* D; and T stands for T *or* S). Residues found with the frequency of >0.95 are highlighted by bold typing. Secondary structure was also predicted for the complete sequence set and the majority rule consensus is shown; a indicates a-helix, b indicates b-strar 1, l indicates loop, and t indicates b-turn. Sequences not shown in Figs. 1–4: T antigen of SV40 (GenBank J02400) and human polyoma virus (PyV; J02228); E1 protein of bovine papilloma virus type 1, (BPV1; X02346), human papilloma virus type 18 (HPV18; X05015), and human papilloma virus type 35 (HPV35; M74117); NS1 protein of densonucleosis viruses of *Bombyx mori* (DNVb; M15123), *Junonia coenia* (DNVj; S47266), and *Aedes albopictus* (DNVa; M37899); putative 2C helicases of poliovirus type 1 (PV; J02281) and hepatitis A virus (HAV; K02990); putative 2C-related helicases of feline calicivirus (FCV; M86379), rice tungro spherical virus (RTSV; M95497), cowpea mosaic virus (CPMV; X00206), and tomato black ring virus (TBRV; D00322); DnaA proteins from *E.coli* (Ec; J01602), *Pseudomonas putida* (Pp; M58352), *Mycoplasma capricolum* (Mc; D90426), *Bacillus subtilis* (Bs; X02369), and *Micrococcus luteus* (Ml; M34006); IstB proteins of Ec insertion sequence IS21 (X14793), *Bacillus thuringiensis* (Bt) insertion sequence IS232 (M38370), and of *Bacillus stearothermophilus* (Bs) insertion sequence IS5376 (X67861); HupR1 protein of *Rhodobacter capsulatus* (Rc; M63670), Ec HydG protein (M28369), and PgtA protein of *Salmonella typhimurium* (St; M13923). The sequence of the putative helicase of BBTV that is shown in Fig. 4 is not included. **(b)** Comparison of profiles of amino acid residue conservation in the three groups of nucleic acid-dependent ATPases. The diagram depicts the three conserved motifs and the variance of distances between them. Each rectangle denotes the amino acid residue(s) which is most frequent in the given position. The scale in the left shows the frequency of the amino acid residues in the motifs. The designations of amino acid residue classes are as in Fig. 1. The profiles were calculated for the set of 101 sequences. Identical or similar residues in all three consensus patterns are shown in black and identical or similar residues in any two of the patterns are shown in gray.

I suggest that the three tightly packed motifs define an ancient ancestral domain of DNA-dependent ATPases. Conceivably, this domain appeared very early in evolution. The helicase superfamily III and the NtrC superfamily combine prokaryotic and eukaryotic proteins. Thus the distinct ancestors of each of these large groups of ATPases containing the 'ABC' domain appeared to have evolved before the divergence of prokaryotes and eukaryotes. Accordingly, the initial ancestral domain could already exist in very primitive organisms.

Inspection of the available alignments of other groups of NTPases for this arrangement of conserved motifs revealed similarities in a subset of the large superfamily that includes regulatory subunits of ATP-dependent proteases and other ATPases with diverse functions (46,47). However, conservation of the 'C' motif could not be traced throughout the superfamily. Hence, these similarities were not pursued further although it cannot be ruled out that they reflect a genuine ancestral relationship.

Different types of deviations in the ATP-binding motifs

Motif A (the so called P loop) has been shown to directly interact with the phosphate moiety of the NTP substrate in several ATPases and GTPases (48,49). Generally, this motif has the formula hydrophobic sequence-(G)x_2(G)xGK[ST] (Gly residues shown in parentheses may occur alternatively) forming a Gly-rich flexible loop preceded by a β-strand and succeeded by an α-helix. MCM proteins contain alanine or serine instead of the third conserved glycine, whereas the putative helicases of densoviruses (insect parvoviruses) and CFDV have other deviations in this motif (Fig. 5a). As discussed elsewhere (50,51), various substitutions in the 'A' motif may be compatible with ATPase activity provided that the typical β-strand-loop-α-helix conformation is maintained. Secondary structure predictions indicated that this is likely to be the case for the MCM proteins and the viral helicases (Fig. 5a). In addition, the perfect conservation of the other two motifs also suggests that these proteins have ATPase activity.

The 'B' motif typically has the structure <hydrophobic stretch>D[ED]. Considerable variability in this motif is observed in the helicases of superfamily III (Fig. 5a). Apparently, the only common denominator for all variants of the 'B' motif is at least one negatively charged residue preceded by a stretch of bulky hydrophobic residues conceivably forming a β-strand. As variants of the 'B' motif with striking deviations from the consensus are found in proteins with demonstrated helicase activity, e.g. T antigen and primase of phage T4, these features should be sufficient for mediating the interaction with ATP via Mg^{2+} cation, in which this motif is implicated (49).

In several proteins of the NtrC superfamily (HydG from *E.coli*, PgtA from *Salmonella typhimurium*, HupR1 from *Rhodobacter capsulatus*), both 'A' and 'B' motifs contain drastic modifications including small deletions and elimination of the hydrophobic region in the 'B' motif (Fig. 5a). It can be predicted that these proteins have lost the ATPase activity.

Domain structure of nucleic acid-dependent ATPases

In nucleic acid dependent ATPases, the small conserved 'ABC' ATPase domain is frequently combined with different other domains that mostly perform related functions. These include the DNA-binding domains of the helix-turn-helix type (NtrC superfamily) and the Zn-finger type (T antigen, primase of P4, putative helicase of AcNPV); the primase domain (51); and the endonuclease domain involved in rolling circle DNA replication (the putative helicases of parvoviruses and geminiviruses; 52,53). In only a few proteins, i.e. the putative RNA helicases of picornaviruses and the putative DNA helicase from *Marchantia* mitochondria, the ATPase domain is accompanied by relatively short additional sequences that may be involved in protein-protein and/or membrane interaction. Thus the ancient 'ABC' ATPase domain may be viewed as a movable module, perhaps performing a mechanistically similar ATPase reaction in different contexts.

CONCLUSIONS

The analysis summarized here indicates that a simple basic arrangements of three structural motifs is conserved in a small ancestral ATPase domain found in a vast variety of nucleic acid-dependent ATPases, which may or may not have helicase activity. Despite the wide range of specific activities, all these proteins may have a degree of functional similarity in that some of them mediate unwinding and the others 'opening' of DNA or RNA duplexes.

A remarkable flexibility of the sequence constraints in the ATPase motifs should be emphasized. The relaxed definition of the 'A' motif of the NTP-binding pattern led to the prediction of DNA-dependent ATPase activity in a new group of proteins involved in eukaryotic DNA replication, the MCM proteins. Extensive search for motifs that despite sequence variation, maintain their structural and functional identity has significant potential in identification of functional domains in proteins.

REFERENCES

1. Kornberg, A. and Baker, T. S. (1992) DNA Replication, 2nd Ed. Freeman, San Francisco.
2. Matson, S. W. and Kaiser-Rogers, K. (1990) *Annu. Rev. Biochem.*, **59**, 289–329.
3. Lohman, T. M. (1993), *J. Biol. Chem.*, **268**, 2269–2272.
4. Gorbalenya, A. E. and Koonin, E. V. (1993) Curr. Opin. Struct. Biol., in press.
5. Bramhill, D. and Kornberg, A. (1988) *Cell*, 52,: 743–755.
6. Georgoupoulos, C. (1989) *Trends Genet.*, **5**, 319–321.
7. Walker J. E., Saraste M., Runswick M. J. and Gay N. J. (1982) *EMBO J.*, 1, 945–951.
8. Gorbalenya A. E. and Koonin E. V. (1989) *Nucleic Acids Res.*, 17, 8413–8440.
9. Gorbalenya, A. E., Koonin, E. V., Donchenko, A. P. and Blinov, V. M. (1988) *FEBS Lett.*, **239**, 16–24.
10. Hodgman, T. C. (1988) *Nature*, **333**, 22–23; 578 (Erratum).
11. Gorbalenya, A. E., Koonin, E. V., Donchenko, A. P. and Blinov, V. M. (1989) *Nucleic Acids Res.*, **17**, 4713–4730.
12. Lain, S., Riechman, J. L., Martin, M. T. and Garcia, J. A. (1989) *Gene*, 82, 357–362.
13. Gorbalenya, A. E., Koonin, E. V. and Wolf, Yu. I. (1990) *FEBS Lett.* **252**: 145–148.
14. Gorbalenya, A.E. and Koonin, E. V. (1990) *J. Mol. Biol.*, **213**, 583–591.
15. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) *J. Mol. Biol.*, **215**, 403–410.
16. Henikoff, S. and Henikoff, J. (1992) *Proc. Natl. Acad. Sci. USA*, 89, 10915–10919.
17. Schuler, G. D., Altschul, S. F. and Lipman, D. J. (1991) *Proteins Struct. Funct. Genet.*, **9**, 180–190.
18. Higgins, D. G. and Sharp, P. M. (1988) *Gene*, **73**, 237–244.
19. Gorbalenya, A. E., Blinov, V. M., Donchenko, A. P. and Koonin, E. V. (1989) *J. Mol. Evol.*, **28**, 256–268.
20. Garnier, J. P., Ostguthorpe, D. J. and Robson, B. (1978) *J. Mol. Biol.*, **120**, 97–120.
21. Borowiec, J. A., Dean, F. B., Bullock, P. A. and Hurwitz, J. (1990) *Cell*, **60**, 181–184.
22. Wang, T. S. F. (1991) *Annu. Rev. Biochem.* **60**, 513–552.

23. Gibson, S. I., Surosky, R. T. and Tye, B.-K. (1990) *Mol. Cell. Biol.*, **10**, 5707–5720.
24. Hennessy, K. M., Lee, A., Chen, E. and Botstein, D. (1991) *Genes Dev.*, **5**, 958–969.
25. Yan, H., Gibson, S. I. and Tye, B.-K. (1991) *Genes Dev.*, **5**, 944–957.
26. Chen, Y., Hennessy, K. M., Botstein, D. and Tye, B.-K. (1992) *Proc. Natl. Acad. Sci. USA*, 89, 10459–10463.
27. Thommes, P., Fett, R., Schray, B., Burkhart, R., Barnes, M., Kennedy, C., Brown, N. C. and Knippers, R. (1992) *Nucleic Acids Res.* **20**, 1069–1074.
28. Bell, S. P. and Stillman, B. (1992) *Nature*, 357, 128–134.
29. Estes, H. G., Robinson, B. S. and Eisenberg, S. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 11156–11160.
30. Weiss, D. S., Batut, J., Klose, K. E., Keener, J. and Kustu, S. (1991) *Cell*, 67, 155–167.
31. Austin, S. and Dixon, R. (1992) *EMBO J.*, **11**, 2219–2228.
32. Dombroski, A. J. and Platt, T (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 2538--2542
33. Thomson, B. J., Efstathiou, S. and Honess, R. W. (1991) *Nature*, **351**: 78–80.
34. Fanning, E. and Knippers, R. (1992) *Annu. Rev. Biochem.*, **61**: 55–85.
35. Im, D. S. and Muzyczka, N. (1990) *Cell*, **61**, 447–457.
36. Im, D. S. and Muzyczka, N. (1992) *J. Virol.*, **66**, 1119–1128.
37. Lanka, E. Scherzinger, E., Lurz, R. and Strack, B. (1992) J. Cell. Biochem., Suppl. **16B**, 49.
38. Mirzayan, C. and Wimmer, E. (1992) *Virology*, **189**, 547–555.
39. Teterina, N. L., Kean, K. M., Gorbalenya, A. E., Agol, V. I. and Girard, M. (1992) *J. Gen. Virol.*, **73**, 1977–1986.
40. Oda, K., Yamato, K., Ohta, E., Nakamura, Y., Takemura, M., Nozato, N., Akashi, K., Kanegae, T., Ogura, Y., Kohchi, T. and Ohyama, K. (1992) *J. Mol. Biol.*, **223**, 1–7.
41. Lu, A. and Carstens, E. B. (1991) *Virology* , **181**, 336–347.
42. Rohde, W., Randles, J. W., Langridge, P. and Hanold, B. (1990) *Virology*, **176**, 648–651.
43. Harding, R. M., Burns, T. M., Hafner, G., Dietzgen, R. G. and Dale, J. L. (1993) *J. Gen. Virol.* **74**, 323–328.
44. Koonin, E. V. (1992) *Nucleic Acids Res.*, **20**, 1143.
45. Koonin, E. V. (1992) *Nucleic Acids Res.*, **20**, 1997.
46. Gottesman, S., Squires, C., Pichersky, E., Carrington, M., Hobbs, M., Mattick, J. S., Dalrymple, B., Kuramitsu, H., Shiroza, T., Foster, T. et al. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 3513–3517.
47. Erdmann, R., Wiebel, F. F., Flessau, A., Rytka, J., Beyer, A., Frohlich, K. U. and Kunau, W. H. (1991) *Cell*, **64**, 499–510.
48. Saraste, M., Sibbald, P. R. and Wittinghofer, A. (1990) *Trends Biochem. Sci.*, **15**, 430–434.
49. Schulz, G. E. (1992) *Curr. Opin. Struct. Biol.*, **2**: 61–67.
50. Koonin, E. V. (1992) *FEBS Lett.*, **312**, 3–6.
51. Koonin, E. V. (1993) *J. Mol. Biol.*, **229**, 1165–1174.
52. Ilyina, T. V., Gorbalenya, A. E. and Koonin, E. V. (1992) *J. Mol. Evol.*, **34**, 351–357.
53. lyina, T. V. and Koonin, E. V. (1992) *Nucleic Acids Res.*, **20**, 3279–3285.
54. Koonin, E. V. and Ilyina, T. V. (1992) J. Gen. Virol. 73, 2763–2766.
55. Coxon, A., Maundrell, K. and Kearsey, S. E. (1992) *Nucleic Acids Res.*, 20, 5571–5577.
56. Sun, J., Inouye, M. and Inouye, S. (1991) *J. Bacteriol.*, **173**, 4171–4181.