Protein Science (1996), 5:178–180. Cambridge University Press. Printed in the USA.

FOR THE RECORD

Two domains of superfamily I helicases may exist as separate proteins

EUGENE V. KOONIN AND KENNETH E. RUDD

National Center for Biotechnology Information, National Library of Medicine, NIH, Bethesda, Maryland 20894

(RECEIVED September 27, 1995; ACCEPTED October 19, 1995)

Abstract: DNA and RNA helicases of superfamily I are characterized by seven conserved motifs. The five N-terminal motifs are separated from the two C-terminal ones by a spacer that is highly variable in both sequence and length, suggesting the existence of two distinct domains. Using computer methods for protein sequence analysis, we show that PhoH, an ATP-binding protein that is conserved in *Escherichia coli* and *Mycobacterium leprae*, is homologous to the putative N-terminal domain of the helicases, whereas the putative *E. coli* protein YjhR is homologous to the C-terminal domain. These findings suggest that the N- and C-terminal domains of superfamily I helicases have distinct activities, with only the N-terminal domain having the ATPase activity. It is speculated that PhoH and YjhR have evolved from helicases through deletion of the portions of the helicase genes coding for the C- and N-terminal domain, respectively.

Keywords: ATPases; conserved amino acid motifs; domain evolution; helicases

DNA and RNA helicases are ubiquitous enzymes that mediate ATP-dependent unwinding of DNA and RNA duplexes and play a key role in gene replication and expression (reviewed in Matson & Kaiser-Rogers, 1990; Matson, 1991; Lohman, 1993). The numerous experimentally characterized and putative helicases demonstrate complex patterns of amino acid sequence conservation. Two large superfamilies, as well as several smaller groups of helicases, have been characterized as the result of detailed sequence comparisons (Hodgman, 1988; Gorbalenya et al., 1988a, 1988b, 1989b; Gorbalenya & Koonin, 1993). Helicase superfamilies I and II each have an array of seven conserved motifs; these sets of motifs are easily distinguished from each other, even though they are thought to be distantly related (Gorbalenya et al., 1989b). In the helicases of superfamily I, the five proximal motifs are separated from the two distal motifs by a spacer that varies widely in length, from about 50 to almost 500 amino acid residues (Gorbalenya & Koonin, 1993), and it has been suggested that the distal motifs belong to a separate domain (Gorbalenya et al., 1989a; Koonin, 1992).

Here we show that distinct proteins homologous to the N-terminal and C-terminal portions of superfamily I helicases actually exist, even though they are much less common than complete helicases containing both regions.

The observations that led to the above conclusion were made in the course of our systematic analysis of the protein sequences encoded by the *Escherichia coli* genome (Koonin et al., 1995, 1996). First, inspection of the results of nonredundant (NR) protein sequence database search using the BLASTP program (Altschul et al., 1990) showed that the ATP-binding protein PhoH belonging to the phosphate regulon (Kim et al., 1993) and its homologue from *Mycobacterium leprae* (Smith & Robison, 1994) share a pattern of conserved amino acid residues that strongly resembles the superfamily I helicase motifs I–V. Specifically, the tripartite amino acid signature $UX_2[GA]X_2GX$ $GK[TS]X_nUX_2DEXQX_nUX_2GDX_2Q$, comprising motifs I, II, and III, is unique for superfamily I helicases and PhoH. (A number of helicases, however, show some deviations from this pattern.) In BLASTP searches, PhoH did not show statistically significant similarity to helicases or to any other proteins, except for the *Mycobacterium* homologue and a partial C-terminal sequence of the *Bacillus subtilis* homologue (Kim et al., 1995), even though a marginal similarity to putative RNA helicases from some RNA viruses was observed (not shown). Nevertheless, when the PhoH sequences from *E. coli* and *M. leprae* were compared with the sequences of five superfamily I helicases (or helicase subunits) from *E. coli* using the MACAW program (Schuler et al., 1991), the five helicase motifs readily aligned (Fig. 1); the alignment blocks containing motifs I, Ia, II, and III were highly statistically significant (probability of aligning by chance, $p < 10^{-5}$). An alignment of the PhoH sequences with those of helicases throughout the putative ATPase domain was difficult to construct owing to a large difference in the distance between motifs Ia and II. However, in the RecD protein, this region is even shorter than in PhoH (Fig. 1). We generated an alignment of two PhoH sequences and RecD using the OPTAL program (Gorbalenya et al., 1989a). The alignment scored 5.5 SD above the random expectation, which is typical for distantly related proteins belonging to the same superfamily. The *E. coli* PhoH and RecD contain 24.1% identical amino acid residues

Reprint requests to: Eugene V. Koonin, National Center for Biotechnology Information, National Library of Medicine, NIH, Bldg. 38A, Bethesda, Maryland 20894; e-mail: koonin@ncbi.nlm.nih.gov.

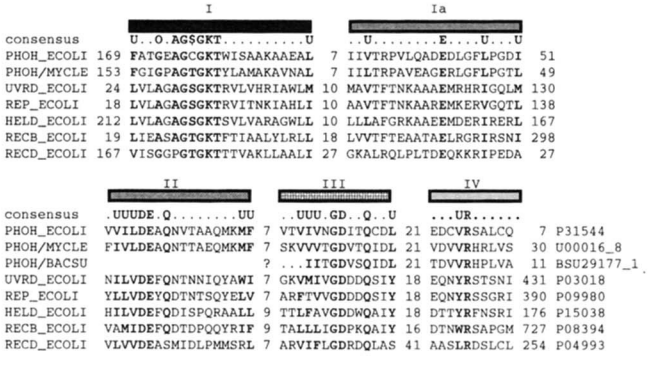

Fig. 1. Conserved motifs in PhoH and the N-terminal domain of superfamily I helicases from *E. coli*. The alignment was constructed using the MACAW program (Schuler et al., 1991). MACAW detects conserved blocks and allows subsequent change of their boundaries accompanied by recalculation of the statistical significance. Thus, the conserved motifs are delimited so as to ensure maximum significance. Designation of the motifs (I–IV) is after Gorbalenya et al. (1989a). Motifs I and II correspond to the A and B motifs that are thought to be essential parts of many NTP-binding sites (Walker et al., 1982; Gorbalenya & Koonin, 1989). The consensus line shows the conserved amino acid residues with one exception allowed (residues conforming to the consensus are highlighted by bold type). Distances from the protein termini and distances between the conserved blocks are indicated for each protein. Accession numbers from the SWISS-PROT database or the GenBank database (*M. leprae* and *B. subtilis* PhoH) are indicated in the rightmost column. *B. subtilis* PhoH is a partial sequence.

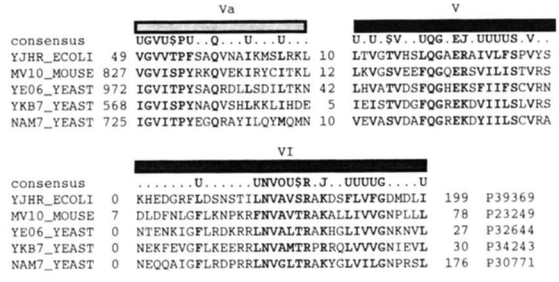

Fig. 2. Conserved motifs in YjhR and the C-terminal domain of a distinct group of putative superfamily I helicases. Motifs V and VI are designated after Gorbalenya et al. (1989a); motif Va has not been described previously. A distinct group of putative superfamily I helicases, which contains the yeast NAM7 protein involved in mitochondrial function (Altamura et al., 1992), MV10, and yeast protein SEN1, has been described previously (Koonin, 1992); SEN1 showed a lower similarity to YjhR and was not included in the alignment. Sequences of the related uncharacterized yeast proteins YE06 and YKB7 have been reported since. The method of alignment construction and designations are as in Figure 1.

and 54.1% similar residues in a 180-residue overlap, which is close to the threshold of sequence similarity sufficient to infer analogous structures (Sander & Schneider, 1991). Thus, we conclude that PhoH is homologous to the N-terminal part of superfamily I helicases.

When a stand-alone version of a domain that normally belongs to a multidomain protein is discovered, there is a concern that the rest of the protein might have been lost because of cloning or sequencing errors. However, in the case of PhoH, this concern is alleviated by the fact that the three independently cloned and sequenced homologues from distantly related bacteria have similar C termini (Fig. 1), and they all lack the putative C-terminal helicase domain containing motifs V and VI. This indicates that the entire PhoH protein indeed corresponds to the N-terminal helicase domain.

The hypothetical *E. coli* protein YjhR (Burland et al., 1995) shows statistically significant sequence similarity to a distinct group of putative superfamily I helicases (Koonin, 1992). In particular, the alignments of the YjhR sequence with those of the mouse protein MV10 and yeast protein YE06 produced by searching the NR database with BLASTP had a very low probability of occurring by chance ($p = 1.8 \times 10^{-5}$ and $p = 3 \times 10^{-4}$, respectively). YjhR aligned only with the C-terminal region of the putative helicases containing motifs V and VI and did not contain a counterpart to the N-terminal domain with motifs I–IV (Fig. 2). Multiple-alignment analysis revealed a striking conservation between YjhR and this particular group of (putative) helicases in motifs V and VI, as well as in an additional, upstream motif; the multiple alignment was highly statistically significant ($p < 10^{-19}$) for each of these blocks. In the *E. coli* chromosome, the *yjhR* open reading frame is preceded by a "gray hole"—a region more than 1 kb long containing no apparent genes (Burland et al., 1995). Comparison of the nucleotide sequence of the "gray hole" with protein sequence databases using the BLASTX program (Altschul et al., 1994) did not show similarity to helicases or any other proteins. Coding-potential analysis using the GeneMark program (Borodovsky & McIninch, 1993; Borodovsky et al., 1994) indicated that *yjhR* is likely to be an expressed gene that is preceded by a noncoding region. Furthermore, application of oligonucleotide frequency matrices that differentially identify two classes of *E. coli* genes (Borodovsky et al., 1995) suggested that *yjhR* may be a horizontally transferred gene. Thus, the available data appear to rule out the possibility that *yjhR* is a part of a helicase gene that has been artifactually disrupted by a frameshift error, and indicate that it is a complete gene encoding a stand-alone version of the C-terminal domain of the helicases.

We found that PhoH and YjhR are distinct proteins homologous to the N-terminal and C-terminal portions of superfamily I helicases, respectively (Fig. 3). This is compatible with the hypothesis that these two parts of the helicases constitute structurally distinct domains (Gorbalenya et al., 1989a). PhoH and the homologous N-terminal helicase domain contain the A and B motifs typical of a broad variety of NTP-using enzymes (Fig. 1; Walker et al., 1982; Gorbalenya & Koonin, 1989) and are likely to possess ATPase activity, in agreement with the experimental demonstration of ATP binding by the *E. coli* PhoH protein (Kim et al., 1993). The properties of the putative PhoH ATPase, and, in particular, the question of whether or not its activity is nucleic-acid-dependent, remain to be studied.

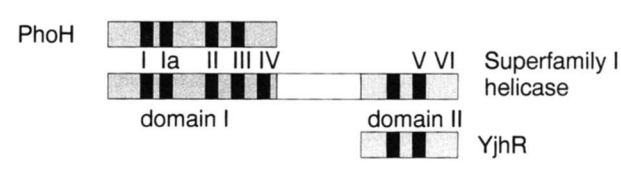

Fig. 3. Superfamily I helicases and stand-alone versions of the N- and C-terminal domains. The scheme shows the general domain organization of a hypothetical superfamily I helicase, PhoH, and YjhR. The positions of the seven conserved motifs are indicated approximately.

There is no experimental data on a possible activity of YjhR. One may speculate that this is a new type of nucleic acid-binding protein; the homologous domain in the helicases may be directly involved in duplex unwinding. This hypothesis seems to be compatible with the finding that, in superfamily II helicases, the distal conserved motifs are essential for the helicase but not for the ATPase activity (Pause & Sonenberg, 1992).

Our database screening performed using BLASTP and motif searches with the DBSITE (Claverie, 1994) and MoST (Tatusov et al., 1994) programs failed to identify any stand-alone homologues of the two helicase domains other than PhoH and YjhR. Although genes coding for such proteins may be discovered upon further accumulation of nucleotide sequences, it is obvious that they are much less widespread than bona fide helicases (and "helicase-like" proteins) containing both domains. In accord with this conclusion, the genome of *Haemophilus influenzae*, the first bacterial genome for which the complete sequence has been reported (Fleischmann et al., 1995), does not encode counterparts to PhoH or YjhR (E.V.K., unpubl. obs.), indicating that these genes are not essential to bacterial cell survival. It seems most likely that the *phoH* and *yjhR* genes have evolved from helicase genes by deletion of the portions coding for the C-terminal domain and the N-terminal domain, respectively. In the case of YjhR, this deletion might have occurred concomitantly with horizontal recombinational transfer of the gene from another organism.

Like many groups of proteins with various functions that have a complex and flexible domain organization (reviewed by Doolittle, 1995), numerous helicases are multidomain proteins; moreover, in some cases, large domains may be inserted between the conserved helicase motifs (Gorbalenya & Koonin, 1993). However, the observations reported here indicate for the first time that the helicase moiety itself may be split into distinct, apparently functional proteins.

References

Altamura N, Groudinsky O, Dujardin G, Slonimski PP. 1992. NAM7 nuclear gene encodes a novel member of a family of helicases with a Zn-ligand motif and is involved in mitochondrial functions in *Saccharomyces cerevisiae*. *J Mol Biol 224*:575-587.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol 215*:403-410.

Borodovsky M, McIninch J. 1993. GenMark [sic]: Parallel gene recognition for both DNA strands. *Comput Chem 17*:123-133.

Borodovsky M, Rudd, KE, Koonin, EV. 1994. Intrinsic and extrinsic approaches for detecting genes in a bacterial genome. *Nucleic Acids Res 22*:4756-4767.

Borodovsky M, McIninch J, Koonin EV, Rudd KE, Medigue C, Danchin A. 1995. Detection of new genes in a bacterial genome using Markov models for three gene classes. *Nucleic Acids Res 23*:3554-3562.

Burland V, Plunkett G III, Sofia HJ, Daniels DL, Blattner FR. 1995. Analysis of the *Escherichia coli* genome VI: DNA sequence of the region from 92.8 through 100 minutes. *Nucleic Acids Res 23*:2105-2119.

Claverie JM. 1994. Some useful statistical properties of position-weight matrices. *Comput Chem 18*:287-294.

Doolittle RF. 1995. The multiplicity of domains in proteins. *Annu Rev Biochem 64*:287-314.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BQ, Merrick JM. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science 269*:496-512.

Gorbalenya AE, Blinov VM, Donchenko AP, Koonin EV. 1989a. An NTP-binding motif is the most conserved sequence in a highly diverged group of proteins involved in positive strand RNA viral replication. *J Mol Evol 28*:256-268.

Gorbalenya AE, Koonin EV. 1989. Virus proteins containing the purine NTP-binding pattern. *Nucleic Acids Res 17*:8413-8440.

Gorbalenya AE, Koonin EV. 1993. Helicases: Amino acid sequence comparisons and structure-function relationship. *Curr Opin Struct Biol 3*: 419-429.

Gorbalenya AE, Koonin EV, Donchenko AP, Blinov VM. 1988a. A conserved NTP-motif in putative helicases. *Nature 333*:22-23.

Gorbalenya AE, Koonin EV, Donchenko AP, Blinov VM. 1988b. A novel superfamily of nucleoside triphosphate-binding motif-containing proteins which are probably involved in duplex unwinding in DNA and RNA replication and recombination. *FEBS Lett 239*:16-24.

Gorbalenya AE, Koonin EV, Donchenko AP, Blinov VM. 1989b. Two related superfamilies of putative helicases involved in replication, recombination, repair and expression of DNA and RNA genomes. *Nucleic Acids Res 17*:4713-4730.

Hodgman TC. 1988. A new superfamily of replicative proteins. *Nature 333*: 22-23;579 [Erratum].

Kim K, Hwang S, Suh J, Song BH, Hong S, Kim J. 1995. Nucleotide sequence upstream of the *cdd* locus in *Bacillus subtilis*. *J Microbiol Biotechnol*. Forthcoming.

Kim SK, Makion K, Amemura M, Shinagawa H, Nakata A. 1993. Molecular analysis of the phoH gene, belonging to the phosphate regulon in *Escherichia coli*. *J Bacteriol 175*:1316-1324.

Koonin EV. 1992. A new group of putative helicases. *Trends Biochem Sci 17*:496-497.

Koonin EV, Tatusov RL, Rudd KE. 1995. Sequence similarity analysis of *Escherichia coli* proteins – Functional and evolutionary implications. *Proc Natl Acad Sci USA*. Forthcoming.

Koonin EV, Tatusov RL, Rudd KE. 1996. *Escherichia coli* protein sequences – Functional and evolutionary implications. In: Ingraham JL, Low KB, Magasanik B, Schaechter M, Umbarger HE, eds. *Escherichia coli and Salmonella typhimurium. Cellular and molecular biology*, 2nd ed. Washington, DC: American Society for Microbiology. Forthcoming.

Lohman TM. 1993. Helicase-catalyzed DNA unwinding. *J Biol Chem 268*: 2269-2272.

Matson SW. 1991. DNA helicases of *Escherichia coli*. *Prog Nucl Acids Res Mol Biol 1991*:289-326.

Matson SW, Kaiser-Rogers KA. 1990. DNA helicases. *Annu Rev Biochem 59*:289-329.

Pause A, Sonenberg N. 1992. Mutational analysis of a DEAD box RNA helicase: The mammalian translation initiation factor eIF-4A. *EMBO J 11*:2643-2654.

Sander C, Schneider R. 1991. Database of homology derived protein structures and the structural meaning of sequence alignment. *Proteins Struct Funct Genet 9*:56-68.

Schuler GD, Altschul SF, Lipman DJ. 1991. A workbench for multiple alignment construction and analysis. *Protein Struct Funct Genet 9*:180-190

Smith DR, Robison K. 1994. GenBank U00016.

Tatusov RL, Altschul SF, Koonin EV. 1994. Detection of conserved segments in proteins: Iterative scanning of sequence databases with alignment blocks. *Proc Natl Acad Sci USA 91*:12091-12095.

Walker JE, Saraste M, Runswick MJ, Gay NJ. 1982. Distantly related sequences in the a- and b-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J 1*:945-951.