Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 3513–3517, May 1990
Biochemistry

Conservation of the regulatory subunit for the Clp ATP-dependent protease in prokaryotes and eukaryotes

(ATPase/chloroplast)

SUSAN GOTTESMAN*[†], CRAIG SQUIRES[‡], ERAN PICHERSKY[§], MARK CARRINGTON[¶], MATTHEW HOBBS[||], JOHN S. MATTICK[||], BRIAN DALRYMPLE**, HOWARD KURAMITSU[††],[‡‡], TERUAKI SHIROZA[††],[‡‡], TOSHI FOSTER[§], WILLIAM P. CLARK*, BARBARA ROSS[‡], CATHERINE L. SQUIRES[‡], AND MICHAEL R. MAURIZI*

*National Cancer Institute, Building 37, Room 4B03, Laboratory of Molecular Biology, Bethesda, MD 20892; [‡]Department of Biological Sciences, Columbia University, New York, NY 10027; [§]Department of Biology, University of Michigan, Ann Arbor, MI 48109; [¶]Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1QW, United Kingdom; [||]Centre for Molecular Biology and Biotechnology, University of Queensland, St. Lucia Ql. D 4067, Australia; **Commonwealth Scientific and Industrial Research Organization Tropical Animal Production, Indooroopilly, Queensland 4068, Australia; and [††]Northwestern University School of Medicine and Dentistry, Chicago, IL 60611

Communicated by E. R. Stadtman, February 12, 1990

ABSTRACT Bacteria, tomatoes, and trypanosomes all contain genes for a large protein with extensive homology to the regulatory subunit, ClpA, of the ATP-dependent protease of *Escherichia coli*, Clp. All members of the family have between 756 and 926 amino acids and contain two large regions, of 233 and 192 amino acids, each containing consensus sequences for nucleotide binding. Within these regions there is at least 85% similarity between the most distant members of the family. The high degree of similarity among the ClpA-like proteins suggests that Clp-like proteases are likely to be important participants in energy-dependent proteolysis in prokaryotic and eukaryotic cells.

Energy-dependent proteolysis plays a key role in prokaryotic and eukaryotic cells by regulating the availability of certain short-lived regulatory proteins, ensuring the proper stoichiometry for multi-protein complexes, and ridding the cell of abnormal proteins (1–3). A homologous energy-dependent protease shared by evolutionarily diverse organisms has not been previously described. In this paper we provide evidence that the regulatory subunit of the ATP-dependent Clp protease of *Escherichia coli* has been conserved to an unusual degree in numerous prokaryotes and eukaryotes. The Clp protease of *E. coli* is a two-component ATP-dependent protease that contributes to the turnover of abnormal proteins (4–7). The large ClpA subunit (81,000 kDa) has intrinsic ATPase activity. The smaller ClpP subunit (21,000 kDa) has low proteolytic activity that is activated in the presence of the ClpA subunit (ref. 8; M.R.M. and S. H. Kim, unpublished data). The *clpA* and *clpP* genes have been cloned and mutations in the genes have been obtained (ref. 7; M.R.M., W.P.C., and S.G., unpublished data). Neither Clp nor another well-characterized ATP-dependent protease of *E. coli*, Lon, is essential for *E. coli* growth, although cells lacking ClpA and Lon have significantly reduced levels of turnover of abnormal proteins (7, 9).

Sequencing of the *clpA* gene, encoding the regulatory subunit of the Clp protease, allowed us to discover that ClpA is a member of a family of well-conserved proteins with no previously described function. Members of the family include a second *E. coli* gene, called *clpB*, and genes from plants, trypanosomes, and Gram-positive and Gram-negative bacteria. *E. coli* contains two members of the family (see below) as does tomato (E.P., unpublished data) (Fig. 1).[§§] *T. brucei* has one gene with DNA sequence homology to the family (M.C., unpublished data); other organisms listed in Fig. 1 have not been tested for the presence of a second gene. Most of these sequences have not previously been published; the presence in the data banks of a fragment of the C terminus of the *clpB* gene (previously called ORF-BG) allowed others of us encountering homologous sequences to contact the author of the first publication (11) and subsequently contact each other. In the absence of a concerted effort to detect Clp-like genes in eukaryotic or other prokaryotic organisms, the serendipitous discovery of seven such genes and a fragment of an eighth suggests that the family is extremely widespread. The degree of conservation strongly suggests that the members of this family all represent the regulatory subunit for Clp-like energy-dependent proteases, although the possibility exists that the ClpA protein has evolved other ATP-dependent regulatory functions as well.

The proteins predicted from the DNA sequences of members of this family have two regions of particularly high homology, each of which contains a consensus sequence for a nucleotide binding site (Figs. 1 and 2). The regions are defined by extensive homology and by the position of introns in the tomato genes that interrupt the coding sequence at the positions shown in Fig. 2. Within region 1 (amino acids 1–234 in Fig. 2) and region 2 (amino acids 359–550) all family members share at least 85% similar and 50% identical amino acids (Table 1). Eighty-nine amino acids of region 1 and 66 amino acids of region 2 are identical in all members of the family thus far identified. The N-terminal 169–270 amino acids of these proteins show much less conservation than these two highly conserved regions, although specific members of the family have similarities in the N terminus. The proteins also differ in the size and sequence of a spacer domain between the two conserved regions. The *Bacteriodes*, *E. coli* ClpB, and trypanosome proteins, which all have longer spacers, have a number of well-conserved residues within the spacer. The C-terminal regions contain a number of well-conserved patches (see amino acids 588–605 and 646–660), separated by poorly conserved regions.

Two conserved sequence motifs have been identified in proteins with adenine nucleotide binding sites (see parts A and B in Fig. 2 and the legend to Fig. 2) (15, 16). Both conserved regions of the members of the ClpA-like family contain well-conserved nucleotide binding sites (marked with dots in Fig. 2). In addition, region 1 but not region 2 contains a grouping

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

[†]To whom reprint requests should be addressed.

[‡‡]Current address: Department of Pediatric Dentistry, University of Texas Health Science Center, San Antonio, TX 78284.

[§§]The following sequences reported in this paper have been deposited in the GenBank data base with the indicated accession numbers: ClpA, M31045; ClpB, M29364; TbB, M32225; T4a, M32603; T4b, M32604; BnB, M32229 and M32230.

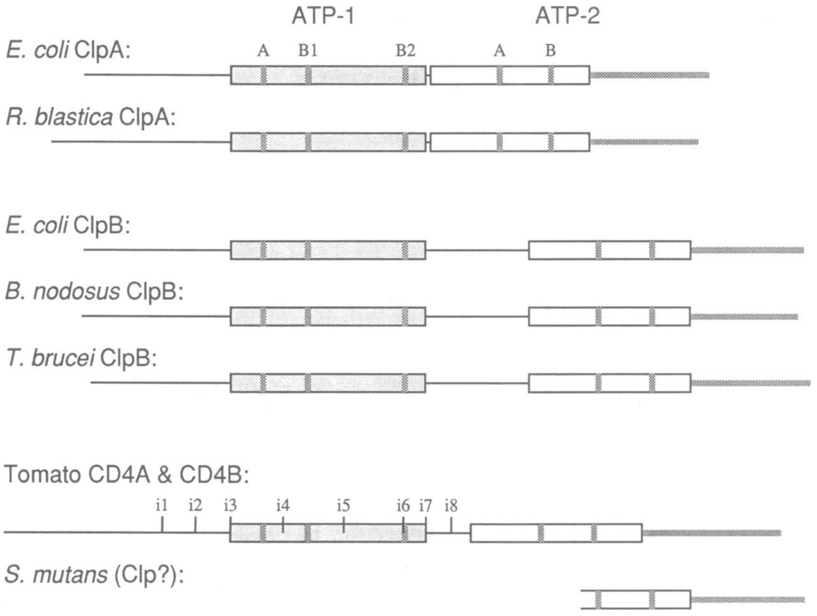

FIG. 1. Organization of ClpA-like proteins. The relative size and organization of the proteins of the ClpA-like family are shown schematically. The boxes represent the well-conserved regions. Within these regions, the sequences homologous to parts A and B of ATP-binding sites are indicated as shaded vertical lines. The thin lines represent the less well-conserved N-terminal and spacer regions. The thick shaded lines represent the partially conserved C-terminal regions. The vertical lines labeled i1, i2, etc., indicate the positions at which the DNA encoding these regions of the protein is interrupted by introns in the tomato genes. The sequence for the *Streptococcus mutans* gene is incomplete (10). *R. blastica, Rhodopseudomonas blastica; B. nodosus, Bacteroides nodosus; T. brucei, Trypanosoma brucei.*

of amino acids (Y-X_8-T-X_{13}-Y, at amino acids 180–207) found in a similar location in the β-subunit of F_1 ATPase and targeted by nucleotide affinity labels (18). Given the observed ATPase activity of purified ClpA, and the dependence of the Clp protease on ATP for proteolysis, the existence of at least a single site is consistent with a similar function for all members of the family. The functional significance of two sites within a polypeptide is not yet clear.

The sequences around the highly conserved nucleotide binding consensus sequences of regions 1 and 2 are distinctly different from each other, suggesting that they serve somewhat different functions in the intact protein. Region 1 extends far beyond the conserved hydrophobic core (part B1 in Fig. 2, amino acids 99–105) and includes a second highly conserved stretch of amino acids with similarities to the hydrophobic core (B2 in Fig. 2, amino acids 218–222). Region

```
Variable leader region:
-270          '         '         '         '         '         '         '         '         '         -161
T4a:MMARALVQSTNILPSVAGERAGQFNGSRKDQRTVRMLCNVKCCSSRLNNFAGLRGCNALDTLLVKSGETLHSKVAAATFVRRPRGCRFVPKAMFERFTEKAIKVIMLAQE
T4b: MARALVQSTSIPSSVAGERTTKFNGSGKTKRAVTMLCNAQSSSLTLRDFTGLRGCNAIDTL VRSGETLQSKVAAATYVRRPRGCRFVPKAMFERFTEKAIKVIMLAQE
BnB:                                                                                               MEFMQMSKKFTARFQEAL
EcB:                                                                                                   MRLDRLTNKFQLAL
TbB:                                                                                                  MAHSDRQCTNAAQTAL
EcA:                                                                                         MLNQELELSLNMAFAREHRHEFMTVEH
RbA:                                                                 MRPRSNAGSSPPDPEEQEHAQVPSFSSTLEQAIHAALALANARRHELATLEH

-160          '         '         '         '         '    |i1       '         '         '         ' |i2 -51
T4a:EARRLGHNFVGTEQILLGLIGEGTGIAAKVLKSMGINLKDARVEVEKIIGRGSGFIAVEIPFTPRAKRVLELSLEEARQLGHNYIGSEHLLLGLLREGEGVAARVLENLG
T4b:EARRLGHNFVGTEQILLGLIGEGTGIAAKVLKSMGINLKDARVEVEKIIGRGSGFVAVEIPFTPRAKRVLELSLEEARQLGHNYIGSEHLLLGLLREGEGVAARVLENLG
BnB:SAAQSLAVGKDHAYIEPLHIFSALLDQEGSSIIAIAQRAGGQIGAVRQAVMQALERLPRVKNPTGDVNISPESMRLLNLCDKYAQQNGDEYISSELFLRAVYDAKGDLEQ
EcB:ADAQSLALGHDNQFIEPLHLMSALLNQEGGSVSPLLTSAGINAGQLRTDINQALNRLPQVEGTGGDVQPSQDLVRVLNLCDNVAQKRGDNFISSELFVLAALESRGTVAD
TbB:SDAVESARKHNNGFVDPAHLALVLFKNEDGLASRVLRKLNV LEPLAARVG ALPEQRPRPRSITFSSDGGCAQHR    RAEANRVGDSLIAVDHLLIGLFE    VEA
EcA:LLLALLSNPSAREGLEACSVDLVALRQELEAFIEQTTPVLPASEEERDTQPTLSFQRVLQRAVFHVQSSGRNEVTGANVLVAIFSEQESQAAYLLRKHEVSRLDVVNFIS
RbA:LLLALIDEPDAARVMKACSVDLEDLKKTLTDFIDDDLSTLVTSVEGSEAVPTAAFQRVIQRAAIHVQSSGRTEVTGANVLVAIFAERESNAAYFLQEQDMTRYDAVNFIA

                                                      ATP-1 conserved region:          Part A
 -50          '         '         '         '     1(i3)     '         '         ' •    •••'         '         60
T4a:ADPTNIRTQVIRMVGESSEAVGASVGGGTSGLKMPTLEEYGTNLTKLAEEGKLDPVVGRQAQIERVTQILGRRTKNNPCLIGEPGVGKTAIAEGLAQRIANGDVPETIEG
T4b:ADPSNIRTQVIRMVGESNEAVGASVGGGTSGQKMPTLEEYGTNLTKLAEEGKLDPVVGRQPQIERVTQILGRRTKNNPCLIGEPGVGKTAIAEGLAQRIANGDVPETIEG
BnB:LLRANGLEKAAVVAAIDAVRGGEAVTDEYAEDKRGALKKYTLDVTQRPLDGKIDPVIGRDEEIRRAMQILQRRSKNNPVLIGEPGVGKTAIVEGLAQRIADRAVPESLKG
EcB:ILKAAGATTANITQAIEQMRGGESVNDQGAEDQRQALKKYTIDLTERAEQGKLDPVIGRDEEIRRTIQVLQRRTKNNPVLIGEPGVGKTAIVEGLAQRIINGEVPEGLKG
TbB:IMKAAHASKKAVEGALLELRKGKKVTSEFQEENYQALEKYATDLCKLAEEGKLDPVIGRTDEVLRTIRVLSRRTKNNPILIGEPGVGKTAIAEGIAQRIVRGDVPDTLLN
EcA:HGTRKDEPTQSS          DPGSQPNSEEQAGGEERMENFTTNLNQLARVGGIDPLIGREKELERAIQVLCRRRKNNPLLVGESGVGKTAIAEGLAWRIVQGDVPEVMAD
RbA:HGVAKDPSYGESRPVQGADEPQETPKAEAGEAKESALSKYCVDLNIKARKGDVDPLIGREAEVERAIQVLCRRRKNNPLLVGDPGVGKTAIAEGLAWKIIKKEVPEVLSG

                                              Part B1
 61 |i4         '         '         '    •     • •••    '         '         '   |i5'         '         '         170
T4a:KKVITLDMGLLVAGTKYRGEFEERLKKLMEEIK QSDEIIILFIDEVHTLIGAGAAE GAIDAANILKPALARGELQCIGATTLDEYRKHIEKDPALERRFQPVKVPEPSV
T4b:KKVITLDMGLLVAGTKYRGEFEERLKKLMEEIK QSDEIIILFIDEVHTLIGAGAAE GAIDAANILKPALARGELQCIGATTLDEYRKHIEKDPALERRFQPVKVPEPTV
BnB:KRLLSLDLAALLAGTKYRGEFEERLKAVLDEITKADGQIILFIDEIHTMVGAG KSEGSLDAGNMLKPALARGDLHCIGATTLDEYRQYMEKDAALERRFQKVIVDEPSV
EcB:RRVLALDMGALVAGAKYRGEFEERLKGVLNDLAKQEGNVILFIDELHTMVGAG KADGAMDAGNMLKPALARGELHCVGATTLDEYRQYIEKDAALERRFQKVFVAEPSV
TbB:TRLFSLDLGALIAGSSLRGEFEERLKSVLNEVKESSNGVILFIDEIHLVLGAG KSGGSMDAANLLKPMLARGELRTIGATTLEEYRTYVEKDAAFERRFMPVYVTEPSV
EcA:CTIYSLDIGSLLAGTKYRGDFEKRFKALLKQLE QDTNSILFIDEIHTIIGAGAASGGQVDAANLIKPLLSSGKIRVIGSTTYQEFSNIFEKDRALARRFQKIDITEPSI
RbA:ATIFSLDMGALLAGTRYRGDFEERLKAVVKELE DHPDAILFIDEIHTVIGAGATSGGAMDASNLLKPALQGGKLRCMGSTTYKEFRQHFEKDRALSRRFQKIDVNEPSV

                                                      Part B2             Variable spacer region:
 171          '         '         '      i6| •     • •••     ' 235(i7)'        '         '     |i8  '         280
T4a:DETIQILKGLRERYEIHHKLHYTDEAIEAAAKLSHQYISDRFLPDKAIDLIDEAGSRVRLRHAQLPEEARELEKELRQITKEKNEAVRGQDFEKAGELRDREMDLKAQIS
T4b:DETIQILKGLRERYEIHHKLRYTDEDLVAAAQLSYQYISDRFLPDKAIDLIDEAGSRVRLRHAQLPEEAKELEKELRQITKEKNEAVRGQDFEKAGELRDREMDLKAQIT
BnB:EDTIAILRNLQERYEVHHGINITDPALVAAAQLSHRYISGRKLPDKAIDLMDEAAAQIRMELDSKPEVMDKIDRRLIQLQIERMALEKETDAASKRRLSDLEAEIAAQEK
EcB:EDTIAILRGLKERYELHHHVQITDPAIVAAATLSHRYIADRQLPDKAIDLIDEAASSIRMQIDSKPEELDRLDRRIIQLKLEQQALMKESDEASKKRLDMLNEELSDKER
TbB:EECISILRGLKDRYEAHHGVQITDNAVVVAAQLANRYITNRFMPDKAIDLIDEACANVRVQLSSRPEAIDILERKKRQLEIEAKALERDKEAASRERLKLVKADIQRVEE
EcA:EETVQIINGLKPKYEGHHDVRYTAKAVRAAVELAVKYINDRHLPDKAIDVIDEAGARARVMPVS          KRKKT
RbA:PDAIKILMGLKPHFEEHHELRYTADAIKSAVELASRYINDRKLPDSAIDVIDEAGAAQHLVSDS          KRKKV
```

FIG. 2. (*Figure continues on the opposite page.*)

```
                                                                              ATP-2 conserved region:
 281        '         '         '         '         '         '         '      359'        '         '   390
T4a:ALIDKNKEKSKAESEAGDAAGPI                                                       VTEADIQHIVSSWTGIPVEKVSTDESDRLLKM
T4b:ALIDKNKEVSKAESEAAD TGPL                                                       VTEADIQHIVSSWTGIPVEKVSTDESDRLLKM
BnB:EYADLEEIWLAEKAGNAGAAEIKEQLDKLRVELEAAKRRGDFARASEIQYGLIPAKEKQLLENEQQTEQRPHRLMRNKVTAEEIAEIVSRWTGIPVAKMMEGEKERLLHL
EcB:QYSELEEEWKAEKASLSGTQTIKAELEQAKIAIEQARRVGDLARMSELQYGKIPELEKQL EAATQLEGKTMRLLRNKVTDAEIAEVLARWTGIPVSRMMESEREKLLRM
TbB:ELQPLVSKYNDERQRIDELQEMQSRLDEKK KLERAVRDGKMDLAADLQYNVIPLIQDRIRSLKEDIERQKATLVQEKVTEGDVAAVVARWTGIPVVKLSQTDRERLLNL
EcA:                                                                              VNVADIESVVARIARIPEKSVSQSDRDTLKNL
RbA:                                                                              LGTKEIEAVVAKIARIPPRNVSKDDAETLRDL

                                                      Part A
 391        '         '         '         '  •    •••   '         '         '         '         '  •  500
T4a:EETLHTRVIGQDEAVKAISRAIRRARVGLKNPNRPIASFIFSGPTGVGKSELAKSLATYYFGSEEAMIRLDMSEFMERHTVSKLIGSPPGYVGYTEGGQLTEAVRRRPYT
T4b:EETLHTRIIGQDEAVKAISRAIRRARVGLKNPNRPIASFIFSGPTGVGKSELAKALAAYYFGSEEAMIRLDMSEFMERHTVSKLIGSPPGYVGYTEGGQLTEAVRRRPYT
Sm1:                                    nsgwqtshwlFLFFGadsVGKTELAKALAEVLFDDESALIRFDMSEYMEKFAASRLNGAPPGYVGYEEGGELTEKVRNKPYS
BnB:ETVLNERVVGQKTAVEAVANAIRRNRAGLSDPKRPIGSFLFLGPTGVGKTELCRTLAQFLFDSEENMVRIDMSEFMEKHSVARLIGAPPGYVGYDQGGYLTEAVRRKPYS
EcB:EQELHHRVIGQNEAVDAVSNAIRRSRAGLADPNRPIGSFLFLGPTGVGKTELCKALANFMFDSDEAMVRIDMSEFMEKHSVSRLVGAPPGYVGYEEGGYLTEAVRRRPYS
TbB:SMHLHRRVKGQDEAVERVADAIIRARAGLSRPNSPTASFLFLGPTGVGKTELVKAVAAELFDDEKHMVRIDMSEYMEQHSVSRLIGAPPGYIGHDEGGQLTEPVRRRPHA
EcA:GDRLKMLVFGQDKAIEALTEAIKMARAGLGHEHKPVGSFLFAGPTGVGKTEVTVQLSKAL GIE  LLRFDMSEYMERHTVSRLIGAPPGYVGFDQGGLLTDAVIKHPHA
RbA:ERTLKRLVFGQDKAIEALSASIKLARAGLREPEKPIGNYLFTGPTGVGKTEVAKQLAATL GVE  LLRFDMSEYMEKHAVSRLIGAPPGYVGFDQGGMLTDGVDQHPHC

  Part B                                                 Trailer region:
  • •••      '         '         '         '         551        '         '         '         '         '      610
T4a:VVLFDEIEKAHPDVFNMMLQILEDGRLTDSKGRTVDFKNTLLIMTSNVGSSVIEKGGRRIGFDLDFDEKDSSYNRIKSLVTEELKQYFRPEFLNRLSEMIVFRQLTKLEV
T4b:VVLFDEIEKAHPDVFNMMLQILEDGRLTDSKGRTVDFKNTLLIMTSNVGSSVIEKGGRRIGFDLDLDEKDSSYNRIKSLVTEELKQYFRPEFLNRLDEMIVFRQLTKLEV
Sm1:VLLFDEVEKAHPDIFNILLQVLDDGVLTDSRGRKVDFSNTIIIMTSNLGa    tafvmikplvlelkafpriirpwkvvf*
Sm2:                                                                SFSQDYKAMESRILEELKKVYRPEFINRIDEKVVFHNLGQEDI
BnB:VVLFDEVEKAHSDVFNTLLQVLDEGRLTDGQGRTVDFRHTVIIMTSNLGSDMIQ         LLAEK SYEEMKSAVMEIVMAHFRPEFINRIDEAIVFHGLAKTHM
EcB:VILLDEVEKAHPDVFNILLQVLDDGRLTDGQGRTVDFRNTVVIMTSNLGSDLIQ   ERFGEL       DYAHMKELVLGVVSHNFRPEFINRIDEVVVFHPLGEQHI
TbB:VVLFDEVEKAHPNVYNVLLQVLDDGRLTDSRGRTVDFSNTIIVMTSNLGSEHLL        NPEETNE SYEVLRENVLAAVRSYFRPELINRLDDIVVFRRLRTEDL
EcA:VLLLDEIEKAHPDVFNILLQVMDNGTLTDNNGRKADFRNVVLVMTTNAGVRETER KSIGLIHQDNSTD         AMEEIKKIFTPEFRNRLDNIIWFDHLSTDVI
RbA:VLLLDEIEKAHPDVYNILLQVMDHGKLTDHNGRAVDFRNVILIMTSNVGAADMAK EAIGFGRERRTGE         DTAAVERTFTPEFRNRLDAVISFAPLGREII

 611        '         '         '         '         '         '         '         '         '         '    717
T4a:KEIADIMLKEVFVRLKNKEIELQVTERFRDRVVDEGYNPSYGARPLRRAIMRLLEDSMAEKMLAGEIKEGDSVIVDVDSD GNVTVLNGTSGAPSDSAPEPILV*
T4b:KEIADIMLKEVFERLKVKEIELQVTERFRDRVVDEGYNPSYGARPLRRAIMRLLEDSMAEKMLANEIKEGDSVIVDVDSD GNVTVLNGSSGTPSDPAPEPIPV*
Sm2:RHVVKIMVAPLIAHLADQGITLKFQPSALKHLALAGYDAEMGARPLRRTLQTEVEDKLAELILSGKLASGQTLKIGIQK  EKLKFDIV*
BnB:YRIAQIQLERLRQRLQTRELLLSVEEDAINQLVELGYDPLFGARPLKRAIQNYIENPLAQALLAGQYLPQSTITIGFDG  TNFTFH*
EcB:ASIAQIQLKRLYKRLEERGYEIHISDEALKLLSENGYDPVYGARPLKRAIQQQIENPLAQQILSGELVPGKVIRLEVNE  DRIVAVQ*
TbB:RGVVDNLIAGVNERLKSSGFSVLLDDGVKDFILEHGHDANMGARPLRRWIEKNIVTEIGRMLIAKELPPNSTLRVSLPEGGNKLTFGVKRGLTSDEWE*
EcA:HQVVDKFIVELQVQLDQKGVSLEVSQEARNWLAEKGYDRAMGARPMARVIQDNLKKPLANELLFGSLVDGGQVTVDALKEKNELTYGFQSAQKHKAEAAH*
RbA:LQVVEKFVLQLEAQLIDRNVHIELTPEAAAWLGEKGYDDKMGARPLGRVIQEHIKKPLAEELLFGKLTKGGLVKVGV KD DAIVLEVQEPQKPRLTGQKPPLLTAE*
```

FIG. 2. Sequence alignments for ClpA-like proteins. The protein sequences are those predicted from DNA sequencing done in the following laboratories: EcA, *E. coli* ClpA, National Institutes of Health group; EcB, *E. coli* ClpB, Columbia group, and includes sequences published in ref. 11; TbB, *T. brucei* ClpB, Cambridge group; T4a and T4b, tomato (*Lycopersicon esculentum*) Cd4A and Cd4B from chromosomes 3 and 12, University of Michigan group; BnB, *B. nodosus* ClpB, Australian group; Sml and Sm2, amino acid sequences of ORF1 (where ORF indicates open reading frame) (Northwestern University group) and ORF2 (10) from *S. mutans*. RbA is the *R. blastica* ClpA that is described as URF-2 by Tybulewicz *et al.* (12). Sequences were aligned and compared with the aid of the alignment programs IFASTA (13) and MSA (14). Boldface characters indicate amino acids that are in the majority at that position in the alignment. Characters that are boldface and italic indicate amino acids conserved in only a particular Clp subfamily (i.e., ClpA, ClpB, or tomato Clp) in the variable leader and spacer regions. This convention was not used for sequences between amino acids −29 and −1 in the leader and between amino acids 235 and 246 (corresponding to exon 8 in the tomato sequences), where all sequences show weak convergence. The bold lines in the tomato leader sequences mark two 32-amino acid regions that have very similar sequences. The ATP consensus sites are indicated by dots above the aligned sequences. Part A of such sites (15, 16) includes the sequence K/R-X$_4$-G-X$_4$-G-K-T (X indicates unspecified amino acid). Part B consists of R/K-X$_{5-8}$-ϕ_4-D (ϕ indicates hydrophobic amino acids). Numbering starts with 1 at the start of the first region of conserved sequences and proceeds to the C terminus, using spacing for the longest of the aligned sequences. Negative numbers extend from the beginning of the first conserved region back toward the N terminus of the aligned proteins. The limits of region 1 are defined by a fairly sharp line between the poorly conserved leader and spacer sequences and the well-conserved region of the proteins; these defined limits are supported by the location of introns 3 and 7 at the borders of the first region in the tomato genes. Examination of the Sml sequence from *S. mutans* suggests that several reading frame alterations would improve the conformation of this sequence to the region 2 consensus. A frame shift is also required to bring the end of ORF1 (Sml) into the same reading frame as ORF2 (Sm2). Amino acid sequences that deviate from a better fit to the consensus and a single ORF are indicated by faint dotted letters (C.S. and H.K., unpublished observations). Pairwise sequence comparisons of Sm2 with the trailer sequences of the other proteins show that it has greater similarity to *E. coli* and *B. nodosus* ClpB sequences than to *E. coli* or *R. blastica* ClpA (data not shown).

2 does not contain this second part B-like sequence, and the significance of this second core-like sequence in region 1 is not known. Introns in and around region 1 in the tomato genes may suggest how this part of the gene was put together. The fact that region 2 does not contain introns suggests that region 1 and region 2 may have existed as independent entities for some period of time. Alternatively, the two regions may have always existed as two parts of the same gene, but preferential loss (or gain) of introns has occurred in these two regions.

Clp-like proteins can be assigned to two general groups, exemplified by the *E. coli* genes *clpA* and *clpB*, based on the size of the central spacer segment, the need for gaps in aligning the overall sequences, and sequence similarities in the well-conserved regions and the N- and C-terminal segments. Such considerations suggest that the *clpA* gene of *E. coli* and the gene of *R. blastica* (called URF-2 in ref. 12) are quite similar. These two proteins have 74% similar amino acids over the full length of the protein and >90% amino acid similarity within the conserved regions (Table 1). By contrast, the two *E. coli* proteins, ClpA and ClpB, share an overall similarity of 63% and of 85% within the conserved regions. The ClpB spacer region is 124 amino acids, whereas that of ClpA is 5 amino acids. The *T. brucei* and *B. nodosus* Clp proteins are more similar to the *E. coli* ClpB than to ClpA, both in their leader and spacer sequences and within the conserved regions (Table 1). The tomato genes [which are the result of a recent gene duplication (E.S., unpublished data) and are almost identical to each other except for the first 50 amino acids] may represent a third subfamily, based on their divergent leader and spacer region sequences. Comparisons of the sequences within regions 1 and 2 suggest that the tomato proteins are more like ClpB than ClpA (Table 1).

The degree of conservation seen among members of the ClpA-like family of proteins is comparable to that for other well-conserved families, such as the hsp70 class of heat shock proteins, membrane ATPases, and ribosomal proteins. *T.*

Table 1. Comparison of ClpA-like protein sequences: Percent amino acid similarity within conserved regions

Sequence	EcB	EcA	RbA	TbB	T4A	BnB
EcB	100	54.3	58.5	66.5	69.6	77.7
	0	33.5	30.5	24.9	24.1	20.2
		87.8	89.0	91.4	93.7	97.9
EcA	52.9	100	67.5	54.8	55.9	53.0
	36.0	0	27.5	32.6	34.1	34.8
	88.9		95.0	87.4	90.0	87.8
RbA	53.4	70.7	100	55.2	57.2	56.8
	35.5	22.2	0	31.3	31.9	32.2
	88.9	92.9		86.5	89.1	89.0
TbB	66.7	53.9	54.8	100	61.3	64.6
	26.6	32.3	30.2	0	29.7	27.9
	93.3	86.2	85.0		91.0	92.3
T4A	67.2	50.8	51.6	62.5	100	66.4
	29.2	37.0	37.0	29.2	0	25.9
	90.4	87.8	88.6	91.7		92.3
BnB	78.1	51.3	52.4	66.7	64.1	100
	20.6	36.0	34.9	26.6	32.3	0
	98.9	87.3	87.3	93.3	69.4	

Comparisons of the 13 amino acids within the conserved regions indicated on Fig. 2 were done using IFASTA (13) and the PAM250 matrix to define similar amino acids (17). Column and row headings are abbreviations for genes as in Fig. 2. For each comparison, line 1 shows the % identical amino acids within the region, line 2 shows the % similar amino acids, and line 3 shows the sum of these two numbers. Comparisons for region 1 are to the right of the diagonal; comparisons for region 2 are to the left of the diagonal. The limits of regions 1 and 2 were chosen by inspection as amino acids 1–234 for region 1 and amino acids 359–550 for region 2; the limits of region 1 are also the locations of introns 3 and 7 in the tomato genes.

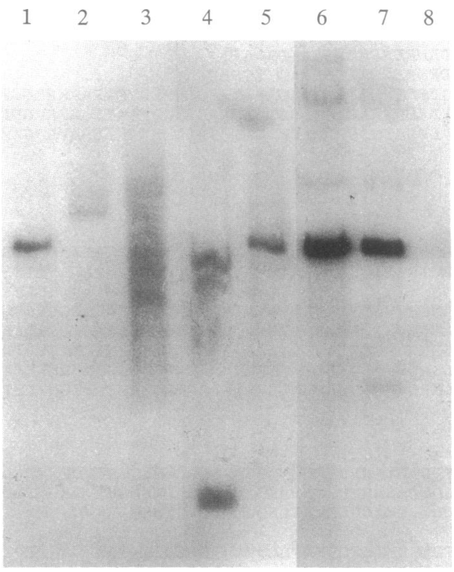

FIG. 3. Southern blots of DNA from different organisms using an *E. coli clpB* probe. Lanes 1–5, DNA from *E. coli* strain MC1000 (lane 1); *S. cerevisiae* (lane 2); *Drosophila melanogaster* (lane 3); archaebacterium *Methanosarcana acetivorans* (lane 4); *E. coli* strain OP50-1 (used as food for the *C. elegans*) (lane 5). Lanes 6–8 [from a different (partial) digestion and gel], DNA from *E. coli* strain OP50-1 (lane 6); *C. elegans* (lane 7); coliphage λ (lane 8). The stronger, upper band in lane 7 apparently is derived from hybridization to DNA from the *E. coli* strain fed to the *C. elegans*; the faster-moving faint band is specific to *Caenorhabditis* and is not seen with the *E. coli* strain. DNAs from different sources were digested with *Pst* I and *Bam*HI endonucleases and the digests were separated on 0.8% agarose gels. The probe was made from an M13 single-stranded DNA template (isolate BP1) that contains the proximal 69% of the *clpB* gene (to position 421 in Fig. 2). Label was incorporated into the probe by extension from five primers that are distributed throughout the gene fragment.

brucei hsp70 (19) and *E. coli dnaK* (20) share 65% similar amino acids over 670 amino acids, whereas the *T. brucei* Clp-like protein and *E. coli* ClpB share 73% similar amino acids over 843 amino acids. Therefore, it seems likely that the functional conservation between the Clp-like proteins extends beyond the ATPase function and reflects a common role as the regulatory subunits of ATP-dependent proteases. The prediction from such a conclusion is that the small subunit of the Clp protease, ClpP, will also be widely conserved. In fact, the sequence of the *clpP* gene shares a high degree of homology with a gene for an open reading frame found in tobacco and liverwort chloroplast DNA (M.R.M., W.P.C., and S.G., unpublished data; B. Wallner, Y. Zhu, R. Tizzard, K. Xia, C. H. Chung, and A. L. Goldberg, personal communication).

How widespread is the Clp family of proteases? In this report, we have detected closely related genes in Gram-negative (*E. coli, R. blastica*, and *B. nodosus*) and Gram-positive bacteria (*S. mutans*), the eukaryotic parasite *T. brucei*, and a plant (tomato). Southern blots to a variety of other organisms (Fig. 3) are able to detect cross-hybridizing species in *Saccharomyces cerevisiae, Caenorhabditis elegans*, and the archaebacterium *Methanosarcana acetivorans*. Therefore, we believe that this protein will in fact be rather ubiquitous and represents a previously unrecognized class of highly conserved functions found in prokaryotic and eukaryotic cells.

In *E. coli*, energy-dependent proteolysis is responsible for the degradation of abnormal proteins, including protein fragments and missense proteins (reviewed in ref. 1). *clpA* mutants have been shown to decrease the residual turnover of abnormal, canavanine-containing proteins in cells lacking the Lon ATP-dependent protease (7). In addition, *clpA* mutants are defective for the turnover of specific protein fusions to β-galactosidase that are not subject to degradation by Lon (S.G., W.P.C., and M.R.M., unpublished data). Therefore, it seems likely that the Clp protease has a role in turnover of abnormal proteins but may also serve to degrade some specific, not yet identified, naturally unstable proteins.

Energy-dependent proteases have been detected in isolated mitochondria (21), in chloroplasts (22), and in a variety of eukaryotic cells (reviewed in ref. 2). A major energy-dependent pathway for protein turnover in eukaryotic cells is thought to depend on ubiquitination of target proteins (reviewed in refs. 2 and 23). Protease complexes capable of degrading ubiquitin-tagged proteins in an ATP-dependent fashion have recently been identified (24–26); the relationship of this complex to the Clp-like proteases is not yet clear. A possible role for Clp-like proteins in organelles is suggested by two lines of evidence: (*i*) the finding of genes homologous to ClpP in the genome of chloroplasts and (*ii*) the observation that the composition and order of the N-terminal 50 or so residues in both of the tomato sequences resemble that of transit peptides, the sequences that are responsible for the import into chloroplasts of cytosolically synthesized proteins (27). An active energy-dependent protease such as Clp in chloroplasts could help to explain, for example, the observed degradation of excess small subunits of ribulose 1,5-bisphosphate carboxylase (28) in chloroplasts. Whatever the eventual function, the striking conservation of this gene throughout eukaryotes and prokaryotes strongly suggests ClpA-like proteins play an important role as the regulatory subunits for ATP-dependent proteases.

C.L.S. gratefully acknowledges the contribution of DNAs by M. Chalfie, A. Tzagaloff, S. Mount, D. Robins, and R. Gunsalus. B.D., M.H., and J.S.M. acknowledge support from the Australian Meat and Livestock Research and Development Corporation and National Health and Medical Research Council; E.P. acknowledges support from the National Science Foundation.

1. Gottesman, S. (1989) *Annu. Rev. Genet.* **23**, 163–198.
2. Bond, J. S. & Butler, P. E. (1987) *Annu. Rev. Biochem.* **56**, 333–364.
3. Goldberg, A. L., Voellmy, R., Chung, C. H., Menon, A. S., Desautels, M., Meixsell, T. & Waxman, L. (1985) in *Intracellular Protein Catabolism*, eds. Khairallah, E. A., Bond, J. S., & Bird, J. W. C., (Liss, New York), pp. 33–45.
4. Katayama-Fujimura, Y., Gottesman, S. & Maurizi, M. R. (1987) *J. Biol. Chem.* **262**, 4477–4485.
5. Hwang, B. J., Park, W. J., Chung, C. H. & Goldberg, A. L. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 5550–5554.
6. Hwang, B. J., Woo, K. M., Goldberg, A. L. & Chung, C. H. (1988) *J. Biol. Chem.* **263**, 8727–8734.
7. Katayama, Y., Gottesman, S., Pumphrey, J., Rudikoff, S., Clark, W. P. & Maurizi, M. R. (1988) *J. Biol. Chem.* **263**, 15226–15236.
8. Woo, K. M., Chung, W. J., Ha, D. B., Goldberg, A. L. & Chung, C. H. (1989) *J. Biol. Chem.* **264**, 2088–2091.
9. Maurizi, M. R., Trisler, P. & Gottesman, S. (1985) *J. Bacteriol.* **164**, 1124–1135.
10. Shiroza, T. & Kuramitsu, H. (1988) *J. Bacteriol.* **170**, 810–816.
11. Shen, W. F., Squires, C. & Squires, C. L. (1982) *Nucleic Acids Res.* **10**, 3303–3313.
12. Tybulewicz, V. L. J., Falk, G. & Walker, J. E. (1984) *J. Mol. Biol.* **179**, 185–214.
13. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448.
14. Lipman, D. J., Altschul, S. F. & Kececioglu, J. D. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 4412–4415.
15. Walker, J. E., Saraste, M., Runswick, M. J. & Gay, N. J. (1982) *EMBO J.* **1**, 945–951.
16. Chin, D. T., Goff, S. A., Webster, T., Smith, T. & Goldberg, A. L. (1988) *J. Biol. Chem.* **263**, 11718–11728.
17. Dayhoff, M., Schwartz, R. M. & Orcutt, B. C. (1978) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. (Natl. Biomed. Res. Found., Silver Spring, MD), Vol. 5, Suppl. 3, pp. 345–352.
18. Senior, A. E. (1988) *Physiol. Rev.* **68**, 177–231.
19. Glass, D. J., Polver, R. I. & Van der Ploeg, L. H. T. (1986) *Mol. Cell. Biol.* **6**, 4657–4666.
20. Bardwell, J. C. A. & Craig, E. A. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 848–852.
21. Desautels, M. & Goldberg, A. L. (1982) *J. Biol. Chem.* **257**, 11673–11679.
22. Malek, L., Bogorad, L., Ayers, A. R. & Goldberg, A. L. (1984) *Fed. Eur. Biochem. Soc.* **166**, 253–257.
23. Rechsteiner, M. (1987) *Annu. Rev. Cell. Biol.* **3**, 1–30.
24. Hough, R., Pratt, G. & Rechsteiner, M. (1987) *J. Biol. Chem.* **262**, 8303–8313.
25. Fagan, J. M., Waxman, L. & Goldberg, A. L. (1987) *Biochem. J.* **243**, 335–343.
26. Ganoth, D., Leshinsky, E., Eytan, E. & Hersko, A. (1988) *J. Biol. Chem.* **263**, 12412–12419.
27. Keegstra, K., Olsen, L. J. & Theg, S. M. (1989) *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **40**, 471–501.
28. Schmidt, G. W. & Mishkind, M. L. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 2632–2636.