FOR THE RECORD

P100, a transcriptional coactivator, is a human homologue of staphylococcal nuclease

CHRISTOPHER P. PONTING

Fibrinolysis Research Unit, University of Oxford, The Old Observatory, South Parks Road, Oxford OX1 3RH, United Kingdom

Abstract: *Staphylococcus aureus* nuclease (SNase) homologues, previously thought to be restricted to bacteria and archaea, are demonstrated by sequence analysis to be present also in eukaryotes. The human cellular coactivator p100 is shown to contain four repeats, each of which is a SNase homologue. Surprisingly, these repeats are unlikely to possess SNase-like activities as each lacks equivalent SNase catalytic residues, yet they may mediate p100's single-stranded DNA-binding function. Products of *Corydalis sempervirens* and *Saccharomyces cerevisiae* open reading frames are predicted to adopt the same fold and possess similar functions as SNase. Five additional hypothetical proteins of bacterial origin are also predicted to be active SNase-like nucleases, including one that appears to be C-terminally truncated in a manner analogous to an engineered active SNase variant. Conservation of Asp-19 and Asp-83 among these homologues suggests a re-evaluation of the roles of these residues in Ca^{2+}-binding and/or catalysis.

Keywords: Ca^{2+}-binding; DNA-binding; homology; tandem repeats

Staphylococcus aureus nuclease (SNase; or "nuclease A") is a small (≈ 150 amino acids) globular enzyme that catalyzes the Ca^{2+}-dependent hydrolyses of single- or double-stranded DNA and RNA at the 5' position of phosphodiester bonds (Cuatrecasas et al., 1967a; Tucker et al., 1978). X-ray crystallographic (Cotton et al., 1978; Loll & Lattman, 1989; Hynes & Fox, 1991; Judice et al., 1993; Libson et al., 1994) and NMR (Torchia et al., 1989; Wang et al., 1990) structures of wild-type and variant SNases have implicated several residues in this enzymatic process. The side-chain carboxylates of Asp-21 and Asp-40, and the backbone carbonyl oxygen of Thr-41 ligate a single Ca^{2+}, as do two water molecules that are hydrogen-bonded to Glu-43. In crystal structures, the 5'-phosphate of a presumed substrate analogue (deoxythymidine 3',5'-bisphosphate) is almost completely buried and binds Arg-35, Arg-87, and Ca^{2+} (Cotton et al., 1978; Loll & Lattman, 1989). Cotton et al. (1978) orig-

inally proposed Glu-43 as acting as a general base in the reaction mechanism, yet more recent evidence (Wang et al., 1990; Judice et al., 1993; Libson et al., 1994) indicates otherwise. Clearly the catalytic mechanism for this enzyme is more complex than previously proposed (Cotton et al., 1978; Loll & Lattman, 1989), particularly given the requirement of not one but two equivalents of Ca^{2+} for maximal activity (Cuatrecasas et al., 1967b; Tucker et al., 1979a, Tucker et al., 1979b).

To date, homologues of *S. aureus* SNase have been found to be encoded only in bacteria (Gerlitz et al., 1990; Yoshioka et al., 1990; Chesneau & El Solh, 1992; Close & Kado, 1992; Chesneau & El Solh, 1994) and archaea (Bult et al., 1996). Importantly, those residues in *S. aureus* SNase that are thought to be involved in substrate-binding and catalysis (Asp-21, Arg-35, Asp-40, Glu-43, and Arg-87) are conserved in these homologues (Chesneau & El Solh, 1994).

Repeats in P100 as candidate SNase homologues: Interest was aroused in human transcriptional coactivator p100 (Tong et al., 1995) as a result of its identification as a member of a novel family containing "tudor" domains (Ponting, 1997). P100 is a nuclear protein that binds single-stranded DNA, the Epstein-Barr virus nuclear antigen 2 (EBNA 2), and both subunits of transcription factor TFIIE (Tong et al., 1995). It appears to be essential for normal cell growth and may act as a bridge between EBNA 2-type acidic domains and the basal transcription machinery during Epstein-Barr virus infection of B lymphocytes (Tong et al., 1995). Self-comparisons of the p100 sequence using dot plots (Thompson et al., 1994a) and REPRO, an algorithm that detects distant repeat sequences (Heringa & Argos, 1993), indicated the presence of four repeats each containing approximately 150 residues. Analysis using MACAW (Schuler et al., 1991) indicated that the four repeats in human p100 are homologues, as multiple alignments generated by Gibbs sampling (Lawrence et al., 1993) revealed the probabilities (*p*-values) of four separate blocks aligning by chance of 3.4×10^{-17}, 2.3×10^{-4}, 4.2×10^{-3}, and 9.7×10^{-3} (here a maximal searchspace N was chosen as $N = 885^4$; human p100 contains 885 residues). Similar results (not shown) were obtained for a second p100 sequence (F10g7.2), known as a result of the *Caenorhabditis elegans* genome project (Wilson et al., 1994).

In order to investigate whether these repeats are homologues of any other proteins, an alignment of human and *C. elegans* p100 repeats was generated using ClustalW (Thompson et al., 1994b) and used to generate profiles that were compared with current protein sequence databases (Gribskov et al., 1987; Birney et al., 1996). Surprisingly, in two types of searches seven of the nine highest scoring sequences were known SNase homologues; the remaining two sequences were later shown to be previously-unknown SNase homologues (below). Using PROFILESEARCH (Gribskov et al., 1987) these sequences yielded Z-scores of greater than 6.5, suggesting that the p100 repeats and nuclease family are diverged from a common ancestor. This is consistent with the results of Blastp (Altschul et al., 1994) searches. A search of databases with the human p100 sequence yielded *p*-values of 8.8×10^{-3} and 1.3×10^{-2} when aligned with *S. aureus* and *Methanococcus jannaschii* SNases, whereas a *C. elegans* p100 search yielded a *p*-value of 9.4×10^{-6}

with *Staphylococcus hyicus* SNase. Reciprocal searches using SNase sequences provided similar results (not shown).

These results immediately suggest that the p100 repeats are homologues of the nuclease family. However, a multiple alignment of all these sequences reveals that absolutely conserved amino acids in known SNases thought to be involved either in binding Ca^{2+} (Asp-21 and Asp-40; *S. aureus* SNase numbering) or in catalysis or in substrate-binding (Arg-35, Arg-87, and Glu-43) are not conserved in any of the human and *C. elegans* p100 repeats (Fig. 1). Two possible interpretations of this were considered: either the p100 repeats are not SNase homologues and their sequence similarities to them arise by chance, or the repeats are SNase homologues that have dispensed with their catalytic activities.

To address this question, profile and motif searching algorithms were employed to complement previous results. Results indicate that p100 repeats are, indeed, homologues of SNase that lack cat-
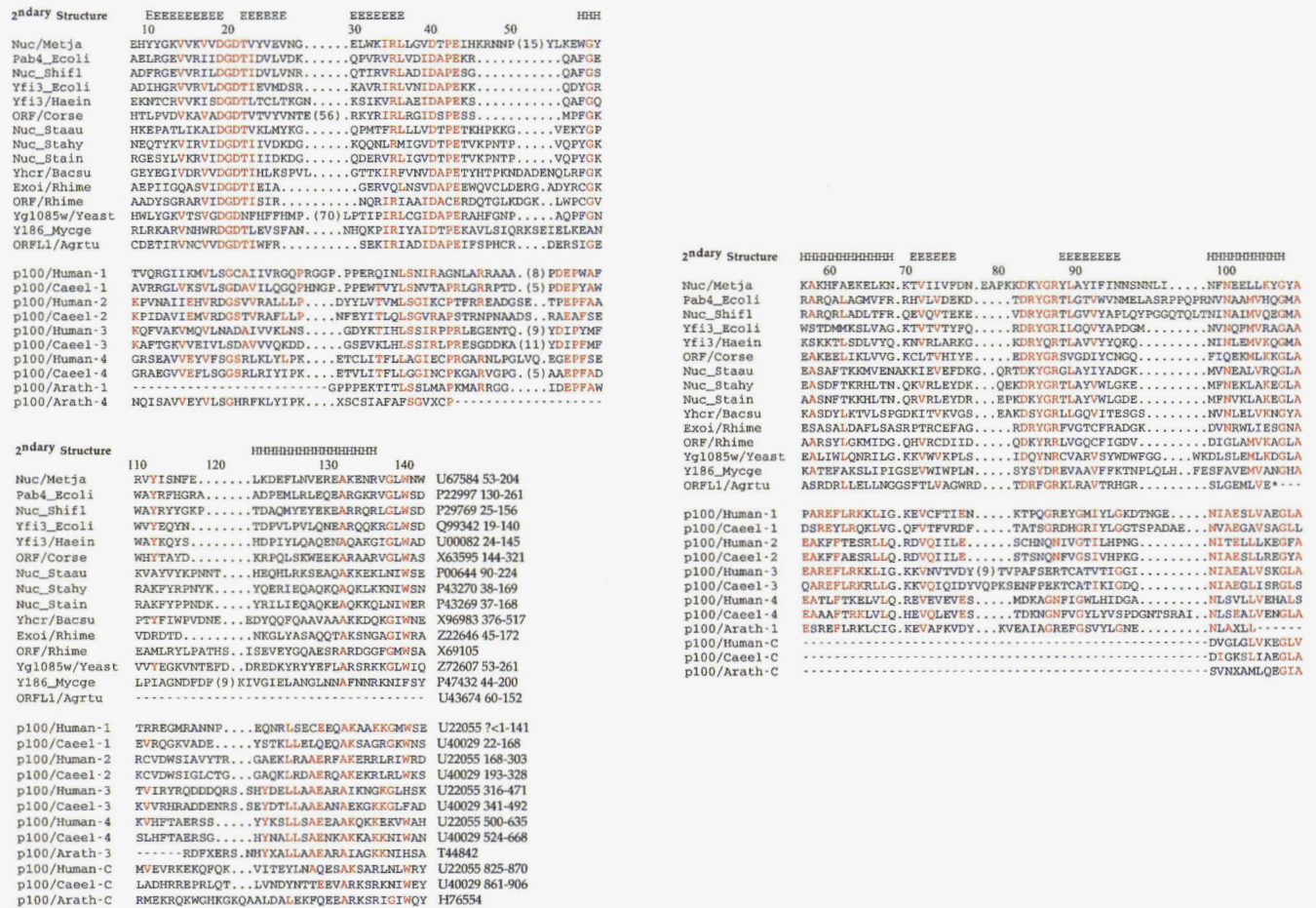
```
2ndary Structure   EEEEEEEEEE EEEEEE        EEEEEE                   HHH
                   10         20        30        40        50
Nuc/Metja      EHYYGKVVKVVDGDTVYVEVNG......ELWKIRLLGVDTPEIHKRNNP(15)YLKEWGY
Pab4_Ecoli     AELRGEVVRIIDGDTIDVLVDK......QPVRVRLVDIDAPEKR..........QAFGE
Nuc_Shifl      ADFRGEVVRILDGDTIDVLVNR......QTIRVRLADIDAPESG..........QAFGS
Yfi3_Ecoli     ADIHGRVVRVLDGDTIEVMDSR......KAVRIRLVNIDAPEKS..........QDYGR
Yfi3_Haein     EKNTCRVVKISDGDTLTCLTKGN.....KSIKVRLAEIDAPEKS..........QAFGQ
ORF/Corse      HTLPVDVKAVADGDTVTVYVNTE(56).RKYRIRLRGIDSPESS...........MPFGK
Nuc_Staau      HKEPATLIKAIDGDTVKLMYKG......QPMTFRLLLVDTPETKHPKKG......VEKYGP
Nuc_Stahy      NEQTYKVIRVIDGDTIIVDKDG......KQQNLRMIGVDTPETVKPNTP.....VQPYGA
Nuc_Stain      RGESYLVKRVIDGDTIIIDKDG......QDERVRLIGVDTPETVKPNTP.....VQPYGK
Yhcr/Bacsu     GEYEGIVDRVVDGDTIHLKSPVL.....GTTKIRFVNVDAPETYHTPKNDADENQLRFGK
Exoi/Rhime     APIIGQASVIDGDTIEIA..........GERVQLNSVDAPEEWQVCLDERG.ADYRCGK
ORF/Rhime      AADYSGRARVIDGDTISIR..........NQRIRIAAIDACERDQTGLKDGK..LWPCGV
Ygl085w/Yeast  HWLYGKVTSVGDGDNFHFFHMP.(70)LPTIPIRLCGIDAPERAHFGNP.....AQPFGN
Y186_Mycge     RLRKARVNHWRDGDTLEVSFAN.....NHQKPIRIYAIDTPEKAVLSIQRKSEIELKEAN
ORFL1/Agrtu    CDETIRVNCVVDGDTIWFR.........SEKIRIADIDAPEIFSPHCR....DERSIGE

p100/Human-1   TVQRGIIKMVLSGCAIIVRGQPRGGP.PPERQINLSNIRAGNLARRAAA.(8)PDEPWAF
p100/Caeel-1   AVRRGLVKSVLSGDAVILQGQPHNGP.PPEWTVYLSNVTAPRLGRRPTD.(5)PDEPYAW
p100/Human-2   KPVNAIIEHVRDGSVRALLLP....DYYLVTVMLSGIKCPTFRREADGSE..TPEPFAA
p100/Caeel-2   KPIDAVIEMVRDGSTVRAFLLP....NFEYITLQLSGVRAPSTRNPNAADS..RAEAFSE
p100/Human-3   KQFVAKVMQVLNADAIVVKLNS.....GDYKTIHLSSIRPPRLEGENTQ.(9)YDIPYMF
p100/Caeel-3   KAFTGKVVEIVLSDAVVVQKDD.....GSEVKLHLSSIRLPRESGDDKA(11)YDIPFMF
p100/Human-4   GRSEAVVEYVFSGGSKLRLKLYPK....ETCLITFLLAGIECPRGARNLPGLVQ.EGEPFSE
p100/Caeel-4   GRAEGVVEFLSGGSRLRIYIPK....ETVLITFLLGGINCPKGARVGPG.(5)AAEPFAD
p100/Arath-1   ------------------------GPPPEKTITLSSLMAPKMARRGG.....IDEPFAW
p100/Arath-4   NQISAVVEYVLSGHRFKLYIPK....XSCSIAFAFSGVXCP-----------------
```

```
2ndary Structure               HHHHHHHHHHHHHHHH
                   110       120       130       140
Nuc/Metja      RVYISNFE.......LKDEFLNVEREAKENRVGLWNW  U67584 53-204
Pab4_Ecoli     WAYRFHGRA......ADPEMLRLEQEARGKRVGLWSD  P22997 130-261
Nuc_Shifl      WAYRYYGKP......TDAQMYEYEKEARRQRLGLWSD  P29769 25-156
Yfi3_Ecoli     WVYEQYN........TDFVLPVLQNEARQQKRGLWSD  Q99342 19-140
Yfi3_Haein     WAYKQYS........HDPIYLQAQENAQAKGIGLWAD  U00082 24-145
ORF/Corse      WHYTAYD........KRPQLSKWEEKARAARVGLWAS  X63595 144-321
Nuc_Staau      KVAYVYKPNNT....HEQHLRKSEAQAKKEKLNIWSE  P00644 90-224
Nuc_Stahy      RAKFYRPNYK.....YQERIEAQKQAQKLKKNIWSN  P43270 38-169
Nuc_Stain      RAKFYPPNDK.....YRILIEAQKEAQKDKKNIWSN  P43269 37-168
Yhcr/Bacsu     PTYFIWPVDNE...EDYQQFQAAVAAAKKDQKGIWNE  X96983 376-517
Exoi/Rhime     VDRDTD........NKGLYASAQQTAKSNGAGIWRA  Z22646 45-172
ORF/Rhime      EAMLRYLPATHS..ISEVEYGQAESRARDGGFGMWSA  X69105
Ygl085w/Yeast  VVYEGKVNTEFD..DREDKYRYYEPLARSRKKGLWIQ  Z72607 53-261
Y186_Mycge     LPIAGNDPDF(9)KIVGIELANGLNNAFNNRKNIFSY  P47432 44-200
ORFL1/Agrtu    ------------------------------------  U43674 60-152

p100/Human-1   TRREGMRANNP....EQNRLSECEEQAKAAKKGMWSE  U22055 ?<1-141
p100/Caeel-1   EVRQGKVADE.....YSTKLLELQEQAKSAGRGKWNS  U40029 22-168
p100/Human-2   RCVDWSIAVYTR...GAEKLRAAERFAKERRLRIWRD  U22055 168-303
p100/Caeel-2   KCVDWSIGLCTG...GAQKLRDAERQAKEKRLRLWKS  U40029 193-328
p100/Human-3   TVIRYRQDDDQRS.SHYDELLAAEARAIKNGKGLHSK  U22055 316-471
p100/Caeel-3   KVVRHRADDENRS.SEYDTLLAAEANAEKGKKGLFAD  U40029 341-492
p100/Human-4   KVHFTAERSS.....YYKSLLSAEEAAKQKKEKVWAH  U22055 500-635
p100/Caeel-4   SLHFTAERSG.....HYNALLSAENKAKKAKKNIWAN  U40029 524-668
p100/Arath-3   ------RDFXERS.NHYXALLAAEARAIAGKKNIHSA  T44842
p100/Human-C   MVEVRKEKQFQK...VITEYLNAQESAKSARLNLWRY  U22055 825-870
p100/Caeel-C   LADHRREPRLQT...LVNDYNTTEEVARKSRKNIWEY  U40029 861-906
p100/Arath-C   RMEKRQKWGHKGKQAALDALEKFQEEARKSRIGIWQY  H76554
```

```
2ndary Structure   HHHHHHHHHHH EEEEEE          EEEEEEEE           HHHHHHHHHH
                   60         70        80        90        100
Nuc/Metja      KAKHFAEKELKN.KTVIIVFDN.EAPKKDKYGRYLAYIFINNSNNLI.....NFNEELLKYGYA
Pab4_Ecoli     RARQALAGMVFR.RHVLVDEKD....TDRYGRTLGTVWVNMELASRPPQPRNVNAAMVHQGMA
Nuc_Shifl      RARQRLADLTFR.QEVQVTEKE.....VDRYGRTLGVVYAPLQYPGGQTQLTNINAIMVQEGMA
Yfi3_Ecoli     WSTDMMKSLVAG.KTVTVTYFQ.....RDRYGRILGQVYAPDGM........NVNQFMVRAGAA
Yfi3_Haein     KSKKTLSDLVYQ.KNVRLARKG.....KDRYQRTLAVVYYQKQ........NINLEMVKQGMA
ORF/Corse      EAKEELIKLVVG.KCLTVHIYE.....EDRYGRSVGDIYCNGQ.........FIQEKMLKKGLA
Nuc_Staau      EASAFTKKMVENAKKIEVEFDKG..QRTDKYGRGLAYIYADGK........MVNEALVROGLA
Nuc_Stahy      EASDFTKRHLTN.QKVRLEYDK..QEKDRYGRTLAYVWLGKE........MFNEKLAKEGLA
Nuc_Stain      AASNFTKKHLTN.QRVRLEYDR..EPKDKYGRTLAYVWLGDE........MFNVKLAKEGLA
Yhcr/Bacsu     KASDYLKTVLSPGDKITVKVGS..EAKDSYGRLLGQVITESGS.......NVNLELVKNGYA
Exoi/Rhime     ESASASALDAFLSASRPTRCEFAG....RDRYGRFVGTCFRADGK......DVNRWLIESGNA
ORF/Rhime      AARSYLGKMIDG.QHVRCDIID....QDKYRRLVGQCFIGDV........DIGLAMVKAGLA
Ygl085w/Yeast  EALIWLQNRILG.KKVWVKPLS.....IDQYNRCVARVSYWDWFGG...WKDLSLEMLKDGLA
Y186_Mycge     KATEFAKSLIPIGSEVWIWPLN.....SYSYDREVAAVFFKTNPLQLH..FESFAVEMVANGHA
ORFL1/Agrtu    ASRDRLLELLNGGSFTLVAGWRD....TDRFGRKLRAVTRHGR.......SLGEMLVE*----

p100/Human-1   PAREFLRKLIG.KEVCFTIEN.....KTPQGREYGMIYLGKDTNGE.....NIAESLVAEGLA
p100/Caeel-1   DSREYLRQKLVG.QFVTFVRDF.....TATSGRDHGRIYLGGTSPADAE..NVAEGAVSAGLL
p100/Human-2   EAKFFTESRLLQ.RDVQIILE......SCHNQNIVGTILHPNG.......NITELLLKEGYA
p100/Caeel-2   EAKFFAESRLLQ.RDVQIILE......STSNQNFVGSIVHPKG.......NIAESLLREGYA
p100/Human-3   EAREFLRKKLIG.KKVNVTVDY(9)TVPAFSERTCATVTIGGI........NIAEALVSKGLA
p100/Caeel-3   QAREFLRKRLLG.KKVQIQIDYVQPKSENPFEKTCATIKIGDQ........NIAEGLISRGLS
p100/Human-4   EATLFTKELVLQ.REVEVEVES....MDKAGNFIGWLHIDGA.......NLSVLLVEHALS
p100/Caeel-4   EAAAFTRKLVLQ.HEVQLEVES.....TDKNGNFVGYLYVSPDGNTSRAI.NLSEALVENGLA
p100/Arath-1   ESREFLRKLCIG.KEVAPKVDY..KVEAIAGREFGSVYLGNE........NLAXLL------
p100/Human-C   ------------------------------------------------DVGLGLVKEGLV
p100/Caeel-C   ------------------------------------------------DIGKSLIAEGLA
p100/Arath-C   ------------------------------------------------SVNXAMLQEGIA
```

Fig. 1. Multiple alignment of SNase homologues, including ORFL1 that appears to be a C-terminally truncated SNase homologue, and p100 SNase-like repeats. The latter include partial sequences encoded by *A. thaliana* ESTs. P100 repeats are denoted by numbers, except the putative C-terminal truncated repeat, denoted by "C." The known secondary structure of *S. aureus* SNase (Hynes & Fox, 1991) is shown above the alignment (E = β-strand, H = α-helix). Numbers in parentheses represent numbers of amino acids excised from the alignment, and dots and dashes represent insertions/deletions and incomplete sequences, respectively. EMBL and/or SwissProt accession codes and domain limits are given following the alignment. The proposed initiating Met in human p100 (Tong et al., 1995) may be in error as there is substantial similarity to the *C. elegans* p100 sequence in a preceding region. The alignment was produced using BOXSHADE (K. Hofmann & M.D. Baron, unpublished) using default parameters. Numbering follows the *S. aureus* sequence. Species: Metja, *Methanococcus jannaschii*; Shifl, *Shigella flexneri*; Ecoli, *Escherichia coli*; Haein, *Haemophilus influenzae*; Corse, *Corydalis sempervirens*; Staau, *Staphylococcus aureus*; Stahy, *Staphylococcus hyicus*; Stain, *Staphylococcus intermedius*; Bacsu, *Bacillus subtilis*; Rhime, *Rhizobium meliloti*; Yeast, *Saccharomyces cerevisiae*; Mycge, *Mycoplasma genitalium*; Agrtu, *Agrobacterium tumefaciens*; Caeel, *Caenorhabditis elegans*; and Arath, *Arabidopsis thaliana*.
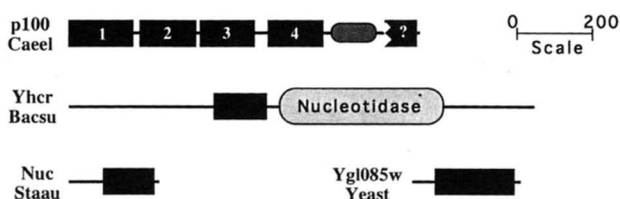
Fig. 2. Schematic representation of the domain organizations of selected SNase homologues (approximately to scale). Rectangles represent SNase homologous domains; the shaded domain in p100 is a tudor domain. The possibility of a p100 repeat-like domain fragment in the C-terminal region of p100 is indicated by a question mark. Abbreviations as in Figure 1 legend.

alytic residues. Comparison of a profile, prepared from an alignment of the eight SNase homologues known from the literature, with databases (Gribskov et al., 1987; Birney et al., 1996) revealed that both human and *C. elegans* p100 sequences scored above perceived "false-positives" at levels comparable to four additional candidate SNases. These candidate homologues are: the exoI gene product from *Rhizobium meliloti* (Becker et al., 1993), a *Bacillus subtilis* gene (*yhcR*) encoding an hypothetical protein that is homologous to 5'-nucleotidases and UDP-sugar hydrolases in its C-terminal portion (M.A. Noback, P. Terpstra, S. Holsappel, G. Venema, S. Bron, unpublished, EMBL code X96983), and putative ORFs from *S. cerevisiae* (Ygl085w) (M. Rieger, S. Mueller-Auer, M. Brueckner, M. Schaefer, unpublished, EMBL code Z72607) and pink corydalis (*Corydalis sempervirens*). The latter ORF had been previously noted to be similar to the plasmid RP4 *parB* gene but not Staphylococcal nucleases (Schaller et al., 1992). *S. cerevisiae* Ygl085w and *C. sempervirens* ORF represent the first known eukaryotic homologues of SNase. One subsequent iteration of this procedure and Tblastn (Altschul et al., 1994) searches revealed two further SNase homologues: a previously unidentified ORF in plasmid pRmeGR4a of *R. meliloti* (Mercado-Blanco & Olivares, 1994) and an hypothetical protein MG186 from *Mycoplasma genitalium*. Each of these six additional SNase homologues conserve putative calcium-binding and/or active site residues Asp-21, Arg-35, Asp-40, Glu-43, and Arg-87, with the exception of an Arg-35 → Gln substitution in *C. sempervirens* ORF (Fig. 1). This suggests that these six sequences represent active Ca^{2+}-dependent nucleases.

Profile-independent searches for SNase homologues employed the MoST algorithm (Tatusov et al., 1994), which iteratively generates conserved blocks (ungapped alignments) of sequences that are significantly similar to user-supplied blocks. Separate alignments of the 8 previously-known SNase homologues encompassing six secondary structure regions $\beta1$-$\beta2$, $\beta3$, $\alpha1$, $\beta5$, $\alpha2$, and $\alpha3$ were constructed as initial alignment blocks and scanned against protein databases. Repeats 1 and 2 in human p100, and repeats 1, 2, and 4 in *C. elegans* p100 were identified in one or more of these scans as significantly similar ($p < 0.02$) to the 8 SNase homologues, as were the *C. sempervirens* ORF, *B. subtilis* ychR gene product, *S. cerevisiae* Ygl085w and *M. genitalium* hypothetical protein MG186. ORFL1 from *Agrobacterium tumefaciens* was also identified with $p < 0.02$; this sequence conserves each of the putative active site residues, yet interestingly appears to be C-terminally truncated (Fig. 1).

SNase-homologous domains in p100: The SNase-homologous domains in p100 appear to lack residues essential for nuclease ac-

tivity. However, the relatively high level of conservation among p100 sequences from diverse eukaryotes (Fig. 1) suggests that these domains mediate one or more conserved functions. These may include the known single-stranded DNA-binding function of p100 (Tong et al., 1995) via a site(s) analogous to the DNA-binding site of SNase homologues. Alternatively these repeats may harbor binding sites for TFIIE subunits or EBNA 2.

Surprisingly, a MoST search for sequences similar to the SNase C-terminal ($\alpha3$) helix identified such a sequence in human p100, C-terminal to its tudor domain, with high significance ($p = 5 \times 10^{-3}$; this value is considered to underestimate the significance given that four other SNase-homologous domains occur in the same sequence and no "false-positive" sequences were identified). Moreover, similarity to SNase $\alpha2$- and $\alpha3$-helices is apparent for human, *C. elegans*, and *Arabidopsis thaliana* p100 C-terminal sequences, including conservation of Gly-107, Ala-132, and Trp-140, residues that are characteristic of SNase homologues (Fig. 1). However, sequences similar to the N-terminal β-barrel of SNase homologues are not apparent, suggesting that the C-terminal region of p100 contains an N-terminally truncated fifth SNase homologous domain (Fig. 2). If so, this bi-helical region might rely on other structural elements of the p100 sequence as contributors to its stability.

Conservation of catalytic and Ca^{2+}-binding residues in active SNase homologues: Identification of SNase homologues in eukaryotes, as well as archaea (Bult et al., 1996), demonstrates that this nuclease is represented in each of the three domains from which all extant life forms have evolved (Woese et al., 1990). A multiple alignment of SNase homologues (Fig. 1) now allows a re-evaluation of the participation of particular residues in nuclease function. Although Ca^{2+}-dependent hydrolysis of single- or double-stranded DNA and RNA has been demonstrated only for *S. aureus*, *Staphylococcus intermedius*, and *S. hyicus* SNases (Tucker et al., 1978; Chesneau & El Solh, 1992; Chesneau & El Solh, 1994), conservation of Ca^{2+}-binding residues (Asp-21 and Asp-40) and a putative catalytic residue (Glu-43) in the newly-identified sequences (not including those in p100) suggests that each of these possesses a Ca^{2+}-dependent nuclease activity. SNase residues Arg-35 and Arg-87 that bind the 5'-phosphate of a presumed substrate analogue (Cotton et al., 1978; Loll & Lattman, 1989) are also conserved in these sequences, except for an Arg-35 → Gln substitution in the *C. sempervirens* ORF (Fig. 1). The proposed interaction of SNase lysine-49 with the 3'-phosphate of substrates (Weber et al., 1993) appears not to be conserved for other homologues.

Several other residues are conserved in non-p100 SNase homologues, implying that these too are essential for nuclease activity and/or fold. Asp-19 is absolutely conserved and may, as suggested elsewhere (Hynes & Fox, 1991), contribute to the binding of a second Ca^{2+} that is known to be required for maximal activity (Cuatrecasas et al., 1967b; Tucker et al., 1979a, Tucker et al., 1979b). Asp-83 also is absolutely conserved, except for Asp → Tyr in MG186, and may contribute to a trinucleotide-binding site (Weber et al., 1992). Conserved hydrophobic and/or small residues (Fig. 1) are known to contribute to *S. aureus* SNase protein stability (Shortle et al., 1990; Green et al., 1992). Variants containing substitutions of Gly-107 or Ala-132 are among the most unstable of SNase mutants (Green et al., 1992), and it is notable that these are among the most conserved residues of SNase homologues, including those in p100 orthologues (Fig. 1). The SNase fold appears to readily accommodate relatively large insertions between $\beta3$ and $\alpha1$, but more particularly between $\beta2$ and $\beta3$ (Fig. 1).

A. tumefaciens ORFL1: A C-terminally truncated SNase?: The SNase structure contains an N-terminal OB-fold (Murzin, 1993) and C-terminal helices. Database searches demonstrated sequence similarity between ORFL1 from *A. tumefaciens* plasmid Ti (Alt-Mörbe et al., 1996) and SNase homologues, including conservation of catalytic and Ca^{2+}-binding site residues (Fig. 1). However, ORFL1 appears to contain only the OB-fold region and lacks residues contributing to two C-terminal α-helixes, α2 and α3. This appears to be a result of neither frame-shift nor stop codon sequencing errors, although a more substantial sequencing error can not be discounted. A more exotic explanation is that ORFL1 may represent a truncated SNase homologue lacking most of these C-terminal helices. As supporting evidence for this unusual proposition, it is noted that an OB-fold SNase variant similarly lacking residues from α2 and all of α3, and containing two amino acid substitutions (Val-66 → Leu and Gly-88 → Val), has been found to be structurally stable and active, albeit at a level 10^3-fold lower than that of the wild-type enzyme (Alexandrescu et al., 1995). In addition, substitution of Val-66 with Leu, which has little or no effect on the stability of intact SNase yet stabilizes the truncated variant (Alexandrescu et al., 1995), occurs naturally in the ORFL1 sequence (Fig. 1).

Note added in proof

Callebut and Mornon (Callebut I, Mornon JP, 1997. The human EBNA-2 coactivator p100: Multidomain organization and relationship to the staphylococcal nuclease fold and to the tudor protein involved in *Drosophila* melanogaster development. *Biochem J* *321*:125–132) have independently reported the SNase- and tudor-homologous domains of p100.

References

Alexandrescu AT, Gittis AG, Abeygunawardana C, Shortle D. 1995. NMR structure of a stable "OB-fold" sub-domain isolated from Staphylococcal nuclease. *J Mol Biol 250*:134–143.

Alt-Mörbe J, Stryker JL, Fuqua C, Li P-L, Farrand SK, Winans SC. 1996. The conjugal transfer system of *Agrobacterium tumefaciens* octopine-type Ti plasmids is closely related to the transfer system of an IncP plasmid and distantly related to Ti plasmid *vir* genes. *J Bacteriol 178*:4248–4257.

Altschul SF, Boguski MS, Gish W, Wootton JC. 1994. Issues in searching molecular sequence databases. *Nature Genet 6*:119–129.

Becker A, Kleickmann A, Kuster H, Keller M, Arnold W, Puhler A. 1993. Analysis of the *Rhizobium meliloti* genes *exoU, exoV, exoW, exoT,* and *exoI* involved in exopolysaccharide biosynthesis and nodule invasion: *exoU* and *exoW* probably encode glucosyltransferases. *Mol Plant Microbe Interact 6*:735–744.

Birney E, Thompson JD, Gibson TJ. 1996. PairWise and SearchWise: Finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. *Nucl Acids Res 24*:2730–2739.

Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD, Kerlavage AR, Dougherty BA, Tomb J-F, Adams MD, Reich CI, Overbeek R, Kirkness EF, Weinstock KG, Merrick JM, Glodek A, Scott JL, Geoghagen NSM, Weidman JF, Fuhrmann JL, Presley EA, Nguyen D, Utterback TR, Kelley JM, Peterson JD, Sadow PW, Hanna MC, Cotton MD, Hurst MA, Roberts KM, Kaine BP, Borodovsky M, Klenk H-P, Fraser CM, Smith HO, Woese CR, Venter JC. 1996.

Complete genome sequence of the methanogenic archeon, *Methanococcus jannaschii. Science 273*:1058–1073.

Chesneau O, El Solh N. 1992. Nucleotide sequence of a nuc gene encoding the thermonuclease of *Staphylococcus intermedius. Nucl Acids Res 20*:5232–5232.

Chesneau O, El Solh N. 1994. Primary structure and biological features of a thermostable nuclease isolated from *Staphylococcus hyicus. Gene 145*:41–47.

Close SM, Kado CI. 1992. A gene near the plasmid pSa origin of replication encodes a nuclease. *Mol Microbiol 6*:521–527.

Cotton FA, Hazen EE Jr, Legg MJ. 1978. Staphylococcal nuclease: Proposed mechanism of action based on structure of enzyme-thymidine 3',5'-bisphosphate-calcium ion complex at 1.5-Å resolution. *Proc Natl Acad Sci USA 76*:2551–2555.

Cuatrecasas P, Fuchs S, Anfinsen, CB. 1967a. Catalytic properties and specificity of the extracellular nuclease of *Staphylococcus aureus. J Biol Chem 242*:1541–1547.

Cuatrecasas P, Fuchs S, Anfinsen CB. 1967b. The binding of nucleotides and calcium to the extracellular nuclease of *Staphylococcus aureus.* Studies by gel filtration. *J Biol Chem 242*:3063–3067.

Gerlitz M, Hrabak O, Schwab H. 1990. Partitioning of broad-host-range plasmid RP4 is a complex system involving site-specific recombination. *J Bacteriol 172*:6194–6203.

Green SM, Meeker AK, Shortle D. 1992. Contributions of the polar, uncharged amino acids to the stability of Staphylococcal nuclease: Evidence for mutational effects on the free energy of the denatured state. *Biochemistry 31*:5717–5728.

Gribskov M, McLachlan AD, Eisenberg D. 1987. Profile analysis: Detection of distantly related proteins. *Proc Natl Acad Sci USA 84*: 4355–4358.

Heringa J, Argos P. 1993. A method to recognize distant repeats in protein sequences. *Proteins: Struct Funct Genet 17*:391–411.

Hynes TR, Fox RO. 1991. The crystal structure of Staphylococcal nuclease refined at 1.7 Å resolution. *Proteins: Struct Funct Genet 10*:92–105.

Judice JK, Gamble TR, Murphy EC, de Vos AM, Shultz PG. 1993. Probing the mechanism of Staphylococcal nuclease with unnatural amino acids: Kinetic and structural studies. *Science 261*:1578–1581.

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. 1993. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science 262*:208–214.

Libson AM, Gittis AG, Lattman EE. 1994. Crystal structures of the binary Ca^{2+} and pdTp complexes and the ternary complex of the Asp^{21} → Glu mutant of Staphylococcal nuclease. Implications for catalysis and ligand binding. *Biochemistry 33*:8007–8016.

Loll PJ, Lattman EE. 1989. The crystal structure of the ternary complex of Staphylococcal nuclease, Ca^{2+}, and the inhibitor pdTp, refined at 1.65 Å. *Proteins: Struct Funct Genet 5*:183–201.

Mercado-Blanco J, Olivares J. 1994. The large nonsymbiotic plasmid pRmeGR4a of *Rhizobium meliloti* GR4 encodes a protein involved in replication that has homology with the REPC protein of Agrobacterium plasmids. *Plasmid 32*: 75–79.

Murzin AG. 1993. OB (oligonucleotide/oligosaccharide binding)-fold: Common structural and functional solution for non-homologous sequences. *EMBO J 12*:861–867.

Ponting CP. 1997. Tudor domains in proteins that interact with RNA. *Trends Biochem Sci.* In press.

Schaller A, Schmid J, Amrhein N. 1992. Plant cDNA similar to a bacterial plasmid partition locus. *Plant Physiol 99*:777–778.

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

Shortle D, Stites WE, Meeker AK. 1990. Contributions of the large hydrophobic amino acids to the stability of Staphylococcal nuclease. *Biochemistry 29*:8033–8041.

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.

Thompson JD, Higgins DG, Gibson TJ. 1994a. Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Comput Applic Biosci 10*:19–30.

Thompson JD, Higgins DG, Gibson TJ. 1994b. CLUSTAL-W—Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties, and weight matrix choice. *Nucl Acids Res 22*:4673–4680.

Tong X, Drapkin R, Yalamanchill R, Mosialos G, Kieff E. 1995. The Epstein-Barr virus nuclear protein 2 acidic domain forms a complex with a novel cellular coactivator that can interact with TFIIE. *Mol Cell Biol 15*:4735–4744.

Torchia DA, Sparks SW, Bax A. 1989. Staphylococcal nuclease—sequential assignments and solution structure. *Biochemistry 28*: 5509–5524

Tucker PW, Hazen EE Jr, Cotton FA. 1978. *Staphylococcal nuclease* reviewed:

A prototypic study in contemporary enzymology. I. Isolation, physical, and enzymatic properties. *Mol Cell Biochem 22*: 67–77.

Tucker PW, Hazen EE Jr, Cotton FA. 1979a. *Staphylococcal nuclease* reviewed: A prototypic study in contemporary enzymology. II. Solution studies of the nucleotide binding site and the effects of nucleotide binding. *Mol Cell Biochem 23*:3–17.

Tucker PW, Hazen EE Jr, Cotton FA. 1979b. *Staphylococcal nuclease* reviewed: A prototypic study in contemporary enzymology. III. Correlation of the three-dimensional structure with the mechanisms of enzymatic action. *Mol Cell Biochem 23*:67–86.

Wang J, LeMaster DM, Markley JL. 1990. 2-dimensional NMR-studies of staphylococcal nuclease. 1. Sequence-specific assignments of ^{1}H signals and solution structure of the nuclease H124L-thymidine 3′,5′-bisphosphate-Ca^{2+} ternary complex. *Biochemistry 29*:88–101

Weber DJ, Gittis AG, Mullen GP, Abeygunawardana C, Lattman EE, Mildvan AS. 1992. NMR docking of a substrate into the X-ray structure of Staphylococcal nuclease. *Proteins: Struct Funct Genet 13*:275–287.

Weber DJ, Serpersu EH, Gittis AG, Lattman EE, Mildvan AS. 1993. NMR docking of the competitive inhibitor thymidine 3′,5′-diphosphate into the X-ray structure of Staphylococcal nuclease. *Proteins: Struct Funct Genet 17*:20–35.

Wilson R, Ainscough R, Anderson K, Baynes C, Berks M, Bonfield J, Burton J, Connell M, Copsey T, Cooper J, Coulson A, Craxton M, Dear S, Du Z, Durbin R, Favello A, Fulton L, Gardner A, Green P, Hawkins T, Hillier L, Jier M, Johnston L, Jones M, Kershaw J, Kirsten J, Laister N, Latreille P, Lightning J, Lloyd C, McMurray A, Mortimore B, O'Callaghan M, Parsons J, Percy C, Rifken L, Roopra A, Saunders D, Shownkeen R, Smaldon N, Smith A, Sonnhammer E, Staden R, Sulston J, Thierry-Mieg J, Thomas K, Vaudin M, Vaughan K, Waterston R, Watson A, Weinstock L, Wilkinson-Sproat J, Wohldman P. 1994. 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans*. *Nature 368*:32–38.

Woese CR, Kandler, O, Wheelis ML. 1990. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci USA 87*:4576–4579.

Yoshioka Y, Fujita Y, Ohtsuba E. 1990. Nucleotide sequence of the promoter-distal region of the *tra* operon of plasmid R100, including *traI* (DNA helicase I) and *traD* genes. *J Mol Biol 214*:39–53.