Comparative sequence analysis of ribonucleases HII, III, II, PH and D

I. Saira Mian*

Sinsheimer Laboratories, University of California Santa Cruz, Santa Cruz, CA 95064, USA and Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

Received March 19, 1997; Revised and Accepted May 29, 1997

ABSTRACT

Escherichia coli **ribonucleases (RNases) HII, III, II, PH and D have been used to characterise new and known viral, bacterial, archaeal and eucaryotic sequences similar to these endo- (HII and III) and exoribonucleases (II, PH and D). Statistical models, hidden Markov models (HMMs), were created for the RNase HII, III, II, PH and D families as well as a double-stranded RNA binding domain present in RNase III. Results suggest that the RNase D family, which includes Werner syndrome protein and the 100 kDa antigenic component of the human polymyositis scleroderma (PMSCL) autoantigen, is a 3′→5′ exoribonuclease structurally and functionally related to the 3′→5′ exodeoxyribonuclease domain of DNA polymerases. Polynucleotide phosphorylases and the RNase PH family, which includes the 75 kDa PMSCL autoantigen, possess a common domain suggesting similar structures and mechanisms of action for these 3′→5′ phosphorolytic enzymes. Examination of HMM-generated multiple sequence alignments for each family suggest amino acids that may be important for their structure, substrate binding and/or catalysis.**

INTRODUCTION

As the details of RNA metabolism have emerged, there has been a concomitant increase in interest in the enzymes that carry out these events. Although it has been suggested that proteasomes, multiprotein complexes involved in processing and turnover of cellular proteins, could also be involved in cellular RNA breakdown and RNA processing (1), one group of enzymes has long been known to be important in such events. Ribonucleases (RNases) are enzymes involved in many functions such as RNA processing, stability, turnover and degradation (reviewed in 2,3). For example, mRNA stability influences gene expression in virtually all organisms from bacteria to mammals and the abundance of a particular mRNA can fluctuate manyfold following a change in the messenger RNA (mRNA) half-life without any alteration in transcription (reviewed in 4). Another testament to the general importance of these enzymes is evidence that self-incompatibility in flowering plants involves an RNase (reviewed in 5,6).

The focus of this work is identification and characterisation of new and known viral, bacterial, archaeal and eucaryotic sequences similar to *Escherichia coli* RNases HII, III, II, PH and D using the recently developed statistical modelling method of hidden Markov models (HMMs) (7–10). A double-stranded (ds) RNA binding domain present in RNase III is examined also. An HMM of the type created and used here is a sequence of nodes in which each node corresponds to a column in a multiple sequence alignment for a family of related sequences. The HMM technique allows identification, modelling and analysis of the core elements of a family likely to be determinants of the folding, structure and function of that family. For the RNases examined here, the results can provide guidance for further experimental and theoretical work as well as insights into the relationships within and between the different families.

RNases HII, III (also called RNase C), II (also called RNase B), PH and D were selected for study because of their important roles in many organisms (reviewed in 2,3,11,12). In particular, they act on a wide spectrum of substrates and include both endo- (HII and III) and exoribonucleases (II, PH and D). RNase HII degrades the RNA moiety of RNA–DNA hybrids (13,14). Processing of ribosomal RNA precursors (pre-rRNAs) and of some mRNAs requires the ds specific RNase III (15). RNase PH is both a phosphorolytic nuclease that removes nucleotides following the CCA terminus of tRNA and a nucleotidyltransferase which adds nucleotides to the ends of RNA molecules by using nucleoside diphosphates as substrates (16,17). RNase II and polynucleotide phosphorylase (PNPase) are the two principal nucleases involved in processive 3′→5′ degradation of single-stranded (ss) mRNA (see, for example, ref. 18). RNases II, PH and D are three of at least five 3′→5′ nucleases required for 3′ processing of tRNA precursors (pre-tRNAs) (12,19). A number of these RNases also have a role in the efficacy of some therapeutic molecules. Antisense agents such as antisense oligonucleotides and ribozymes bind to DNA or RNA sequences and block the synthesis of cellular or viral proteins by interfering with transcription and translation (reviewed in 20). Antisense oligonucleotides form stable duplexes that are substrates for cleavage by RNase H, which, like RNase HII, acts on RNA–DNA duplexes. In addition, RNase II and PNPase appear to be the major nucleases that degrade hammerhead ribozymes (21) as well as RNA-OUT, a 69 nucleotide antisense RNA that regulates Tn10/IS10 transposition (22). Thus, studies of these RNases may yield insights into intracellular degradation of foreign RNAs and subsequent development of more stable ribozymes and antisense molecules. Furthermore, the five RNase families examined here provide a

*Address correspondence to author at: Life Sciences Division (Mail Stop 29-100), Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA. Tel: +1 510 486 6216; Fax: +1 510 486 6949; Email: smian@lbl.gov glimpse into the myriad of roles that RNases play in how cells grow, differentiate and respond to their environment.

METHODS

Escherichia coli RNases HII, III, II, PH and D were used as query sequences in database searches performed with the BLAST suite of programs (23) run with default parameters and a merged, non-redundant collection of sequences derived from PIR, SwissProt and translated GenBank. Database sequences were considered to exhibit a statistically significant similarity to the query if the smallest sum probability $P(N) \leq 0.05$, $P(N)$ being the lowest probability ascribed to any set of high scoring segment pairs for each database sequence. Partial sequences, fragments and expressed sequence tags (ESTs) identified by these BLAST database searches were retained but not employed for HMM training and database discrimination experiments. HMMs were trained for proteins belonging to the RNase HII, III, II, PH and D families as well as a ds RNA binding domain present in RNase III by the procedure outlined below.

An HMM was created using the SAM (Sequence Alignment and Modeling Software System) suite (7,24) running on a MASPAR MP-2204 with a DEC Alpha 3000/300X frontend at the University of California Santa Cruz (UCSC). In an HMM, use of a match state indicates that a sequence has a residue in that column whereas using a delete state denotes that the sequence does not. Insert states allow sequences to have additional residues between columns and represent regions of the sequence that are not part of the core elements of the family being modelled. To improve the ability of the HMM to generalise, to fit sequences not employed for training, Dirichlet mixture priors (25,26) were employed. Free Insertion Modules (FIMs) were utilised to allow the HMM to model a region or motif within a larger sequence. Multiple models were trained and the best used for further studies.

Any sequence can be compared to a model by calculating the likelihood that the sequence was generated by that model. Taking the negative (natural) logarithm of this likelihood gives the NLL score. For sequences of equal length, the NLL scores measures how ‘far’ they are from the model and can be used to select sequences from the same family. To assess the specificity and sensitivity of an HMM, it can be used in database discrimination experiments to distinguish between sequences that belong to the family used to train it from those that do not. The programme hmmscore was used to evaluate how much better a sequence fits a model than some underlying background distribution or null model (NULL) and to assess the significance of the resultant score. Database searching using the HMM involves computing log-odds (NLL-NULL) (27,28) scores for all sequences in a non-redundant protein database obtained from the NCI (29) and updated weekly at UCSC. Taking into account the number of sequences in this database (~211 000 different proteins in late 1996) and an expected number of false positives of 0.01, a significant log-odds score is 22.6. Scores higher than this value denote fewer expected false positives. A database search was performed and based upon examination of the log-odds scores and an HMM-generated alignment, new family members were identified, added to the training set and the HMM retrained. This cycle of ‘search, align and retrain’ was repeated until no new sequences were identified in databases up to December 1996. This final HMM was utilised to generate a multiple sequence alignment of the final training set and the partial sequences retained from the initial BLAST searches.

RESULTS

An aim of this study was to train and use HMMs that minimised the numbers of false positives and false negatives. Amongst ~211 000 different proteins, sequences that were not part of their respective training set had log-odds scores <15.0 whereas training set sequences had scores >31.0 (RNase HII), >25.0 (RNase III), >27.2 (ds RNA binding domain), >60.4 (RNase II), >26.6 (RNase PH) and >21.1 (RNase D). For all six families, inspection of the HMM-generated alignments and examination of the log-odds scores suggested there were no false positives amongst sequences with log-odds scores >21.1 and that such sequences could be classified as being members of the family being modelled. However, it cannot be assumed that there are no false negatives amongst sequences with log-odds scores <15.0. There may be remote homologues that have diverged to a degree that the current HMMs may be too specific (overfit the data) and thus unable to classify them as belonging to a particular family. Further generalisation of the HMMs is required to detect such distant family members.

HMM-generated multiple alignments of members of the six families examined are shown in Figures 1–6 which were produced using ALSCRIPT (30). § denotes new members of a family identified here and ‡ partial sequences retained from BLAST searches but not employed for HMM training. Existing members of the RNase III (15,31,32), ds RNA binding domain (33,34) and RNase II (35–40) families have been described elsewhere. Subsequent discussions will focus on new family members. Invariant positions are defined as those residues conserved across all the sequences in an alignment and whose locations are marked by filled triangles. Amino acids conserved in the majority of sequences are highlighted and columns that are predominantly hydrophobic boxed. Columns containing ‘.’ correspond to insert states and numbers indicate the lengths of insertions in sequences at that position (if present).

Although all six families have at least one or more yeast (*Saccharomyces cerevisiae*) and human member, only the RNase HII family has an archaeal member. It is possible that the current suite of proteins predicted to occur in the complete genome of the archaeon *Methanococcus jannaschii* (41) contains no homologues for the five other families. However, it may be that members of these families have diverged to an extent in this archaeon that the current HMMs are too specific and thus unable to detect these remote homologues. Another explantion may be that open reading frames that are family members have not been identified yet and thus would not have appeared in the databases that were searched using the HMMs.

DISCUSSION

Figure 1 shows new eucaryotic RNase HII family members (yeast, 12:Sc_N2369, 13:Sp_C4G902; worm, 14:Ce_T13H52; mammals, 15:Mm_ESTs, 16:Hs_EST). Since RNase HII acts on RNA–DNA duplexes, they may be involved in DNA replication as well as being candidates for mediating the effect of antisense oligonucleotides.

Figure 2 shows new RNase III family members from bacteria (10:My_ORF, 12:Si_ORF) and eucarya (yeast, 16:Sp_C8A4.08C; worm, 17:Ce_K12H4.8, 20:Ce_F26E4.b; mammals 21:Mm_EST, 22:Hs_ESTs). A *S.cerevisiae* RNase III (RNase RNT1; 18:Sc_RNT1) cleaves pre-rRNA at a U3 snoRNP- dependent site (15) suggesting that some of the other eucaryotic sequences may

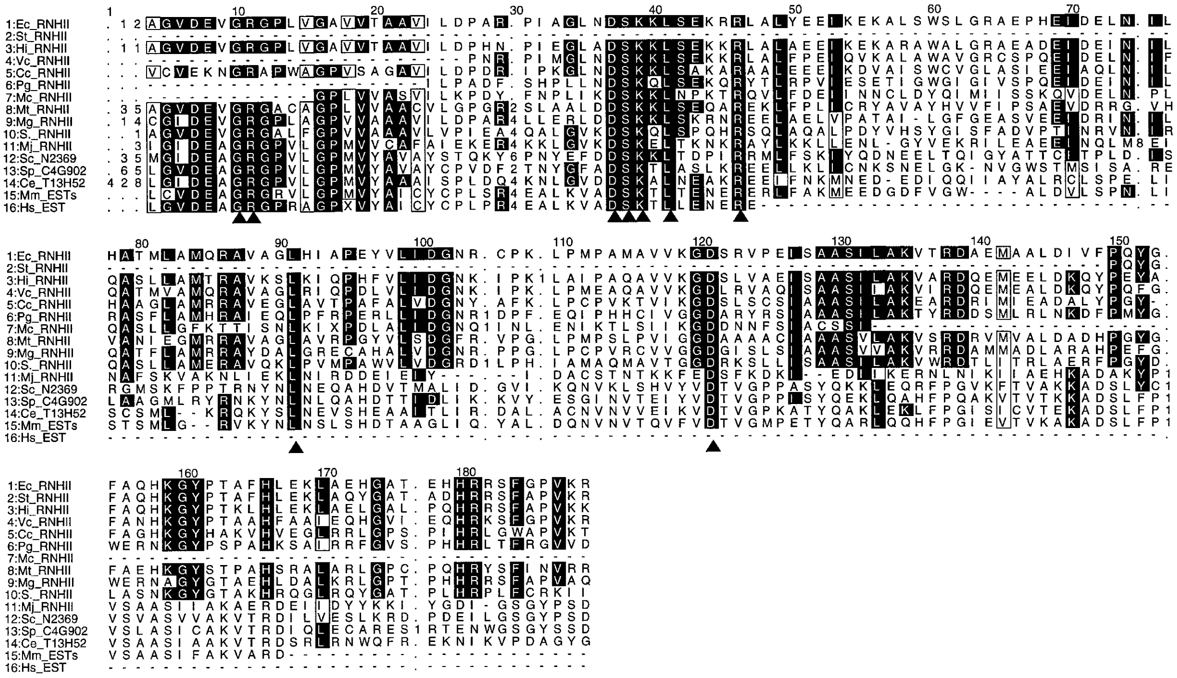

Figure 1. An HMM-generated multiple sequence alignment for RNase HII family members from bacteria [1–10], an archaeon [11] and eucarya [12–16]. 1:Ec_RNHII *E.coli* RNase HII (RNH2_ECOLI); 2:St_RNHII‡ *Salmonella typhimurium* RNase HII (RNH2_SALTY); 3:Hi_RNHII *Haemophilus influenzae* RNase HII (RNH2_HAEIN); 4:Vc_RNHII‡ *Vibrio cholerae* RNase HII (VCU30472); 5:Cc_RNHII *Caulobacter crescentus* RNase HII (S76857); 6:Pg_RNHII‡ *Porphyromonas gingivalis* RNase HII (PGPGAAGEN); 7:Mc_RNHII‡ *Mycoplasma capricolum* RNase HII (S46901); 8:Mt_RNHII *Mycobacterium tuberculosis* RNase HII (MTCY274); 9:Mg_RNHII *Magnetospirillum* sp. RNase HII (MGNMAGA); 10:S._RNHII *Synechocystis* sp. PCC6803 RNase HII (D90899); 11:Mj_RNHII *M.jannaschii* RNase HII (MJU67470); 12:Sc_N2369§ *S.cerevisiae* ORF N2369 (S53908); 13:Sp_C4G902§ *S.pombe* ORF SPAC4G9.02 (SPAC4G9); 14:Ce_T13H52§ *C.elegans* ORF T13H5.2 (CET13H5); 15:Mm_ESTs§‡ *Mus musculus* ESTs (W71720, W76969); 16:Hs_EST§‡ *Homo sapiens* EST (W05602).

be important in pre-rRNA processing. *Schizosaccharomyces pombe* and *Caenorhabditis elegans* each have two RNase III members suggesting involvement in processing different pre-rRNA sites or other RNAs. Three positions in RNase III have been mutated (31,42,43). The first, an invariant Gly (glycine) important for activity in two different sequences, occurs in a highly conserved octapeptide that contains three of the four invariant residues. A second occurs at a variable position. The third is a conserved, functionally important Glu (glutamic acid) present in all RNase III members apart from a bacterium (4:Bs_RNIII) where it is Lys (lysine). In *E.coli* RNase III, a E→K,A mutation uncouples substrate binding from cleavage so that it is unclear whether the *Bacillus subtilis* sequence that has a naturally occuring Lys at this position behaves in a similar manner.

Figure 3 shows new ds RNA binding domain family members from bacteria (9:My_ORF, 10:Cr_ORF, 11:Si_ORF), viruses (68:Va_E3; 33:Rc_NS34; 34:Rs_NS34) and eucarya (yeast 15:Sp_C8A4.08C, 35:Sc_RM03; worm 16:Ce_K12H4.8, 19:Ce_F26E4.b, 46:Ce_F55A44.1, 47:Ce_F55A44.2, 71:Ce_ZK6322, 72:Ce_ORF.1, 73:Ce_ORF.2, 74:Ce_T07D43.1, 75:Ce_T07D43.2, 76:Ce_T22A3.f; mammals 20:Hs_ESTs, 67:Mm_TENR). The underlined sequences are known RNA binding proteins. Although *Paramecium* chlorella virus 1 genome contains a protein that is a member of both this and the RNase III family (12:Pc_A464R; 13:Pc_A464R in Fig. 2), the new viral members lack a catalytic RNase III-like domain suggesting that this activity may reside in a different protein in variola virus and the rotaviruses. A new eucaryotic member may be a link between transcription and RNA metabolism: 71:Ce_ZK6322 also contains a copy of the fork-head-associated (FHA) domain, a putative nuclear signalling domain found in a variety of otherwise unrelated proteins such as transcription factors and kinases (44–46). A subfamily of the ds RNA binding domain is present in an array of tissue types and has a shorter than average α1–β1 loop suggesting a common substrate (52:XI_4F1.1, 53:XI_4F1.2, 54:Hs_NF90.1, 55:Hs_NF90.2, 56:Mm_SPNR.1, 57:Mm_SPNR.2, 58:Rn_RED1.1, 59:Rn_RED1.2). Given the known functions of these members, this substrate could be important in transcription and be a candidate for adenosine to inosine RNA editing.

Figure 4 shows new RNase II family members from bacteria (3:Tf_ORF, 17:S._ZAM) and eucarya (yeast 11:Sc_ORF, 16:Sc_MSU1; malaria parasite 18:Pf_EST; mammal 19:Hs_ESTs). These are candidates for 3′→5′ nucleases involved in processive RNA degradation and since some of the other yeast members are essential for cell division, they may be important for control of mitosis.

Figure 5 shows that RNase PH sequences comprise a complete domain that is also present in PNPase from bacteria (11:Ec_PNPase, 12:PI_PNPase, 13:Hi_PNPase, 14:Bs_PNPase, 15:S._PNPase, 17:Sa_GPS) and eucarya (16:So_PNPase). A number of eucaryotic sequences also possess this domain and are thus members of the RNase PH family (yeast 9:Sc_YG87, 18:Sc_YGR095C, 20:Sc_ORF2315, 21:Sp_YAXE, 23:Sc_D99541; worm 19:Ce_C14A45, 24:Ce_F37C1213; plant 26:Zm_EST; mammals 22:Hs_ORF, 25:Hs_PMSC75, 27:Mm_ESTs, 28:Hs_ESTs). Since both RNase PH and PNPase are 3′→5′ phosphorolytic nucleases, the presence of a common domain suggests that they may have a similar structure and/or mechanism of action. PNPase has been shown to be involved in 3′ adenylation-mediated degradation of mRNA (47). A multiprotein complex mediating

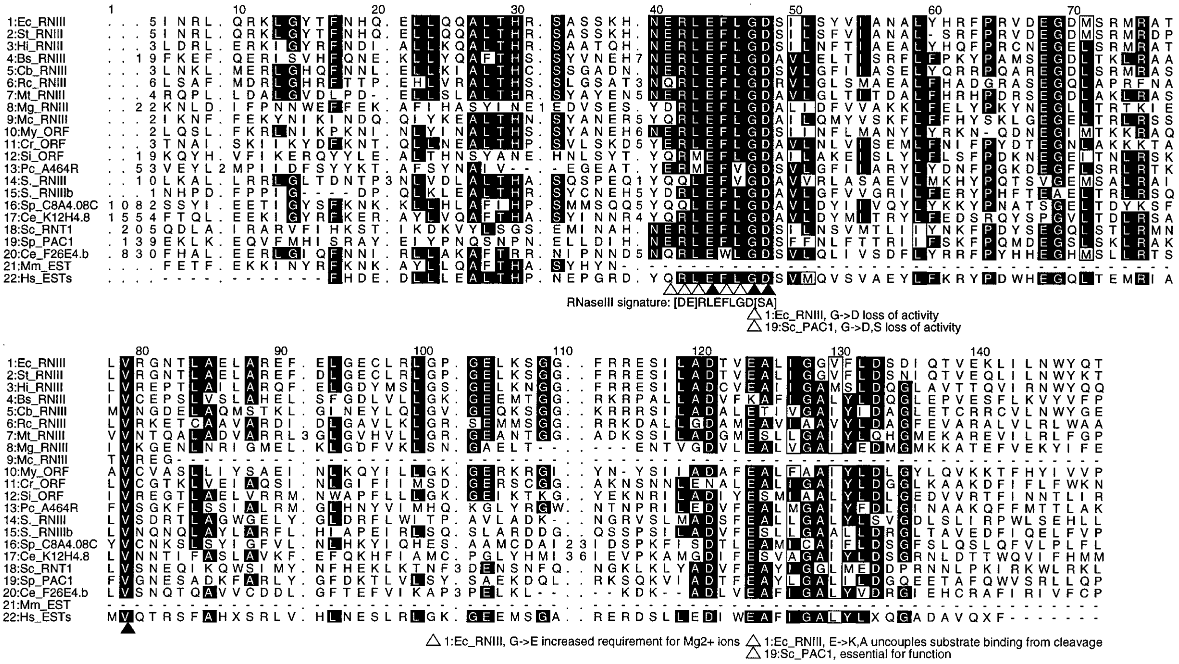
RNaseIII signature: [DE]RLEFLGD[SA]
1:Ec_RNIII, G->D loss of activity
19:Sc_PAC1, G->D,S loss of activity
1:Ec_RNIII, G->E increased requirement for Mg2+ ions
1:Ec_RNIII, E->K,A uncouples substrate binding from cleavage
19:Sc_PAC1, essential for function

Figure 2. An HMM-generated multiple sequence alignment for RNase III family members from bacteria [1–12, 14 and 15], a virus [13] and eucarya [17–22]. The location of the PROSITE RNase III signature (RIBONUCLEASE_III) and the pattern itself are indicated. Mutations in 1:Ec_RNIII (42,43) and 19:Sp_PAC1 (31) are shown. 1:Ec_RNIII *E.coli* RNase III (RNC_ECOLI); 2:St_RNIII *S.typhimurium* RNase III (STU48415); 3:Hi_RNIII *H.influenzae* RNase III (RNC_HAEIN); 4:Bs_RNIII *B.subtilis* RNase III (BACORF1G); 5:Cb_RNIII *Coxiella burnetii* RNase III (COXRER); 6:Rc_RNIII *Rhodobacter capsulatus* RNase III (RCLEPRNCG); 7:Mt_RNIII *M.tuberculosis* RNase III (MTCY338); 8:Mg_RNIII *Mycoplasma genitalium* RNase III (RNC_MYCGE); 9:Mc_RNIII *Mycoplasma capricolum* RNase III (MC235); 10:My_ORF§ *Mycoplasma*-like organism ORF (obtained by translation of the nucleic acid sequence in a different reading frame to that in MOU15224); 11:Cr_ORF§ *Cowdria ruminantium* ORF (obtained by translation of the nucleic acid sequence in a different reading frame to that in CRPCS20); 12:Si_ORF§ *Spiroplasma citri* ORF orfb (SCU28972); 13:Pc_A464R *Paramecium bursaria* chlorella virus 1 ORF A464R similar to RNase III (PBU42580); 14:S._RNIII *Synechocystis* sp. RNase III (SYCSLRB); 15:S._RNIIIb *Synechocystis* sp. PCC6803 RNase III (D90914); 16:Sp_C8A4.08C§ *S.pombe* hypothetical helicase C8A4.08C (YAH8_SCHPO); 17:Ce_K12H4.8§ *C.elegans* hypothetical helicase K12H4.8 (YM68_CAEEL); 18:Sc_RNT1 *S.cerevisiae* RNase RNT1 (SCU27016); 19:Sp_PAC1 *S.pombe* RNase PAC1 (PAC1_SCHPO); 20:Ce_F26E4.b§ *C.elegans* ORF F26E4.b (CEF26E4); 21:Mm_EST§‡ *M.musculus* EST (W54380); 22:Hs_ESTs§‡ *H.sapiens* ESTs (HSA64D051, HSA52B011).

mRNA degradation in *E.coli* and including at least PNPase has been proposed (48). Since RNase PH can catalyse the phosphorolytic cleavage of poly(A) (16), it may be a part of this processing complex and thus involved in coordinated control of mRNA degradation. Given that this RNase PH domain is present in a variety of both bacterial and eucaryotic proteins, control of mRNA degradation in these two kingdoms may be similar. Of particular interest is the human protein 25:Hs_PMSC75, one of two antigenically unrelated proteins recognised by sera from patients suffering from the polymyositis/scleroderma overlap syndrome (PMSCL) and which is defined by idiopathic chronic inflammation in skeletal muscle (reviewed in 49). These 75 and 100 kDa autoantigens, part of a particle localised in the granular component of the nucleolus, belong to the RNase PH and D familes, respectively, suggesting that aberrant RNA processing may be a factor in PMSCL. The role(s) of the other human RNase PH members in this and other disorders remains to be seen.

Figure 6 shows new RNase D family members from bacteria (3:MI_U1764U, 4:Dn_ORFQ, 5:S._sII0320). These RNase D sequences comprise a complete domain that is also present in eucaryotic proteins (yeast 6:Sp_SPAC1F301, 7:Sc_UNC733; worm 8:Ce_C14A44, 10:Ce_ZK10988, 11:Ce_ZK10983; mammals 9:Hs_PMSC100, 12:Hs_WRN). This RNase D domain is itself similar to the proofreading 3′→5′ nuclease domain found in many DNA polymerases (14:Ec_DPOL1-23:C31_ORF are selected sequences from this family). Following this putative exoribonuclease domain is a region not present in the DNA proofreading enzymes. Given the nucleolar location of the 100 kDa PMSCL autoantigen (9:Hs_PMSC100), the particle of which it is a component may be involved in pre-rRNA processing. Werner syndrome is a rare autosomal recessive disorder that mimics some of the characteristics of many age-related features (reviewed in 50). The Werner syndrome protein (WRN) contains a recQ class helicase domain suggesting possible involvement in nucleotide excision repair and transcription (51). The RNase D domain in WRN suggests that it could act upon RNA also and possibly have a role in processing heterogeneous nuclear RNA. How and whether the helicase and 3′→5′ exoribonuclease domains interact and their role(s) in aging remain to be determined.

The RNase D family not only possesses the three motifs characteristic of the DNA proofreading enzymes (52) (Exo I, Exo II and Exo III in Fig. 6), but also shares a number of conserved residues outside these motifs. Since the three-dimensional structure of the 3′→5′ exodeoxyribonuclease domain of *E.coli* Klenow fragment is known, the RNase D family is predicted to possess a similar structure. The structure of the Klenow fragment (14:Ec_DPOL1 in Fig. 6) resembles closely the corresponding region in the T4 DNA polymerase despite limited sequence identify (53). In both cases, the side chains of four negatively charged residues that serve as ligands for the two metal ions required for catalysis are located in geometrically equivalent positions. These charged residues correspond to four of the five invariant residues in

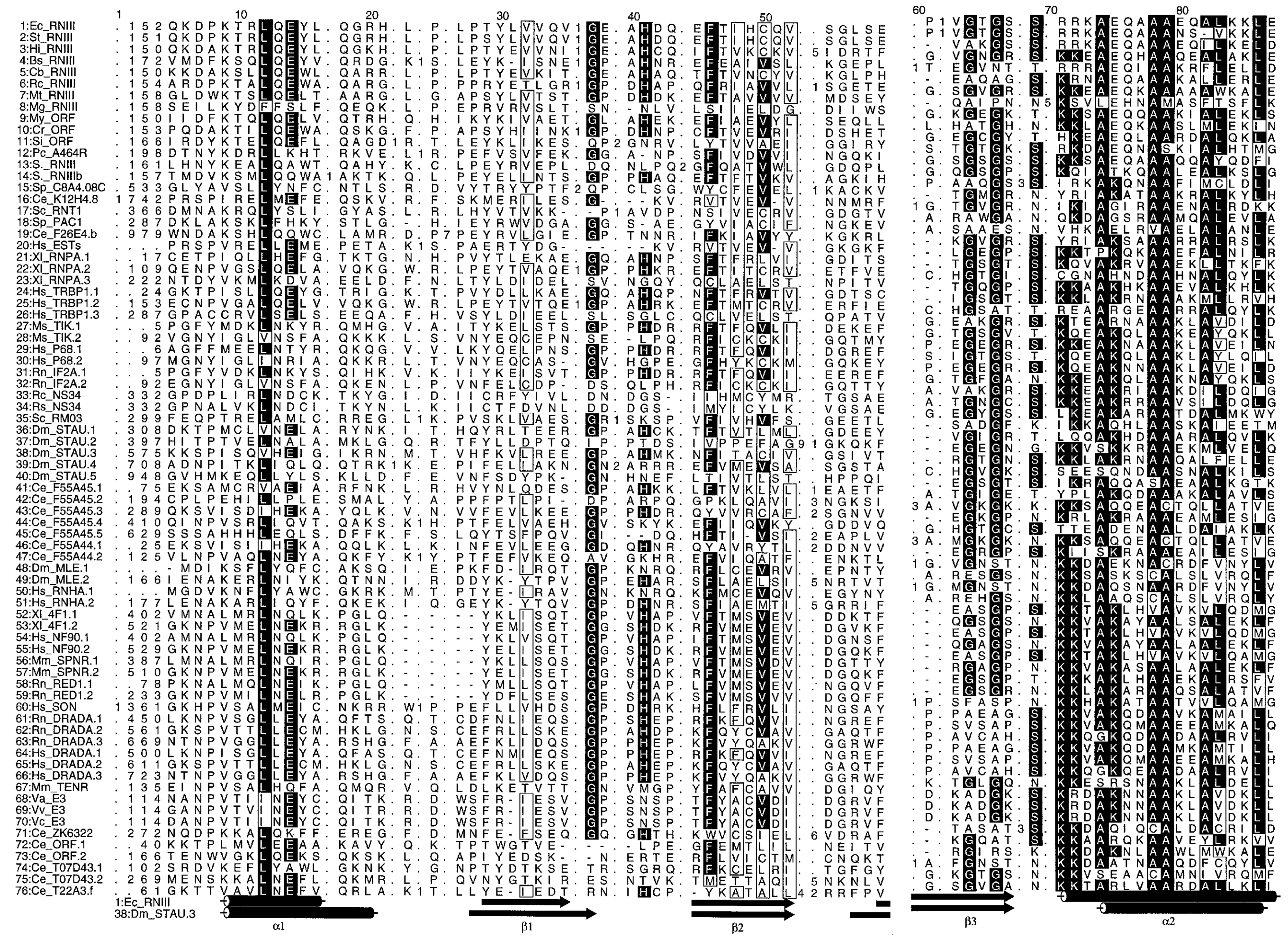

Figure 3. An HMM-generated multiple sequence alignment of ds RNA-binding domain family members from bacteria [1–11, 13 and 14], viruses [12 and 68–70] and eucarya [all other sequences]. Cylinders and arrows denote the α-helices and β-strands in the structures of 1:Ec_RNIII (60) and 38:Dm_STAU.3 (61). 1:Ec_RNIII *E.coli* RNase III (RNC_ECOLI); 2:St_RNIII *S.typhimurium* RNase III (STU48415); 3:Hi_RNIII *H.influenzae* RNase III (RNC_HAEIN); 4:Bs_RNIII *B.subtilis* RNase III (BACORF1G); 5:Cb_RNIII *C.burnetii* RNase III (COXRER); 6:Rc_RNIII *R.capsulatus* RNase III (RCLEPRNCG); 7:Mt_RNIII *M.tuberculosis* RNase III (MTCY338); 8:Mg_RNIII *M.genitalium* RNase III (RNC_MYCGE); 9:My_ORF§ *Mycoplasma*-like organism ORF (translation of MOU15224); 10:Cr_ORF§ *C.ruminantium* ORF (translation of CRPCS20); 11:Si_ORF§ *S.citri* ORFB (SCU28972); 12:Pc_A464R *P.bursaria* chlorella virus 1 ORF A464R similar to RNase III (PBU42580); 13:S._RNIII *Synechocystis* sp. RNase III (SYCSLRB); 14:S._RNIIIb *Synechocystis* sp. PCC6803 RNase III (D90914); 15:Sp_C8A4.08C§ *S.pombe* hypothetical helicase C8A4.08C (YAH8_SCHPO); 16:Ce_K12H4.8§ *C.elegans* hypothetical helicase K12H4.8 (YM68_CAEEL); 17:Sc_RNT1 *S.cerevisiae* RNase RNT1 (SCU27016); 18:Sp_PAC1 *S.pombe* RNase PAC1 (PAC1_SCHPO); 19:Ce_F26E4.b§ *C.elegans* ORF F26E4.b (CEF26E4); 20:Hs_ESTs§ *H.sapiens* ESTs (W60364, R82247); 21:Xl_RNPA.1–23:Xl_RNPA.3 *Xenopus laevis* ds RNA-binding protein A (S27945); 24:Hs_TRBP1.1–26:Hs_TRBP1.3 *H.sapiens trans*-activation-responsive (TAR) RNA-binding protein (A38430); 27:Ms_TIK.1–28:Ms_TIK.2 *M.musculus* serine/threonine protein kinase TIK (A40813); 29:Hs_P68.1–30:Hs_P68.2 *H.sapiens* ds RNA-activated protein kinase p68 (KP68_HUMAN); 31:Rn_IF2A.1–32:Rn_IF2A.2 *Rattus norvegicus* initiation factor 2α (S50216); 33:Rc_NS34§ porcine rotavirus non-structural RNA-binding protein 34 (NS34) (VN34_ROTPC); 34:Rs_NS34§ bovine rotavirus NS34 (VN34_ROTBS); 35:Sc_RM03§ *S.cerevisiae* mitochondrial ribosomal protein L3 (RM03_YEAST); 36:Dm_STAU.1–40:Dm_STAU.5 *Drosophila melanogaster* maternal effect protein staufen (STAU_DROME); 41:Ce_F55A45.1–45:Ce_F55A45.5 *C.elegans* ORF F55A4.5 similar to staufen (CELF55A4); 46:Ce_F55A44.1§–47:Ce_F55A44.2§ *C.elegans* ORF F55A4.4 (CELF55A4); 48:Dm_MLE.1–49:Dm_MLE.2 *D.melanogaster* maleless (MLE) required for increased transcription of X-linked genes in males (A40025); 50:Hs_RNHA.1–51:Hs_RNHA.2 *H.sapiens* RNA helicase A homologous to MLE (RNHA_HUMAN); 52:Xl_4F1.1–53:Xl_4F1.2 *X.laevis* ds RNA-binding protein 4F.1 (XLU07155); 54:Hs_NF90.1–55:Hs_NF90.2 *H.sapiens* NF90, the 90 kDa subunit of cyclosporin A- and FK506-sensitive nuclear transcription factor of activated T-cells (B54857); 56:Mm_SPNR.1–57:Mm_SPNR.2 *M.musculus* spermatid Spnr localised to cytoplasmic microtubles (MMSPNR); 58:Rn_RED1.1–59:Rn_RED1.2 *R.norvegicus* brain ds RNA-specific editase (RED1) (RNU43534); 60:Hs_SON *H.sapiens* son protein (SON_HUMAN); 61:Rn_DRADA.1–63:Rn_DRADA.3 *R.norvegicus* ds RNA adenosine deaminase (DRADA) (RNU18942); 64:Hs_DRADA.1–66:Hs_DRADA.3 *H.sapiens* DRADA (HSU10439); 67:Mm_TENR§ *M.musculus* spermatid RNA binding protein Tenr (MMTENR); 68:Va_E3§ variola virus protein E3L (VE03_VARV); 69:Vv_E3 vaccinia virus strain WR E3L (VE03_VACCV); 70:Vc_E3 vaccinia virus strain Copenhagen E3L (VE03_VACCC); 71:Ce_ZK6322§ *C.elegans* ORF ZK632.2 (YOT2_CAEEL); 72:Ce_ORF.1§–73:Ce_ORF.2§ *C.elegans* ORF (S42378); 74:Ce_T07D43.1§–75:Ce_T07D43.2§ *C.elegans* ORF T07D4.3 (CET07D4); 76:Ce_T22A3.f§ *C.elegans* ORF T22A3.f (CET22A3).

the RNase D family, whilst the fifth corresponds to catalytically active tyrosine (Tyr). This suggests that a 3′→5′ exoribonuclease has a mechanism of action and active site structure similar to a 3′→5′ exodeoxyribonuclease. Thus, mutations at the invariant positions that affect 3′→5′ nuclease activity in DNA proofreading enzymes (reviewed in 54) may have a similar effect on the RNase D family.

Comparative examination of all the families indicates that each possesses at least one invariant Asp and/or Glu, amino acids

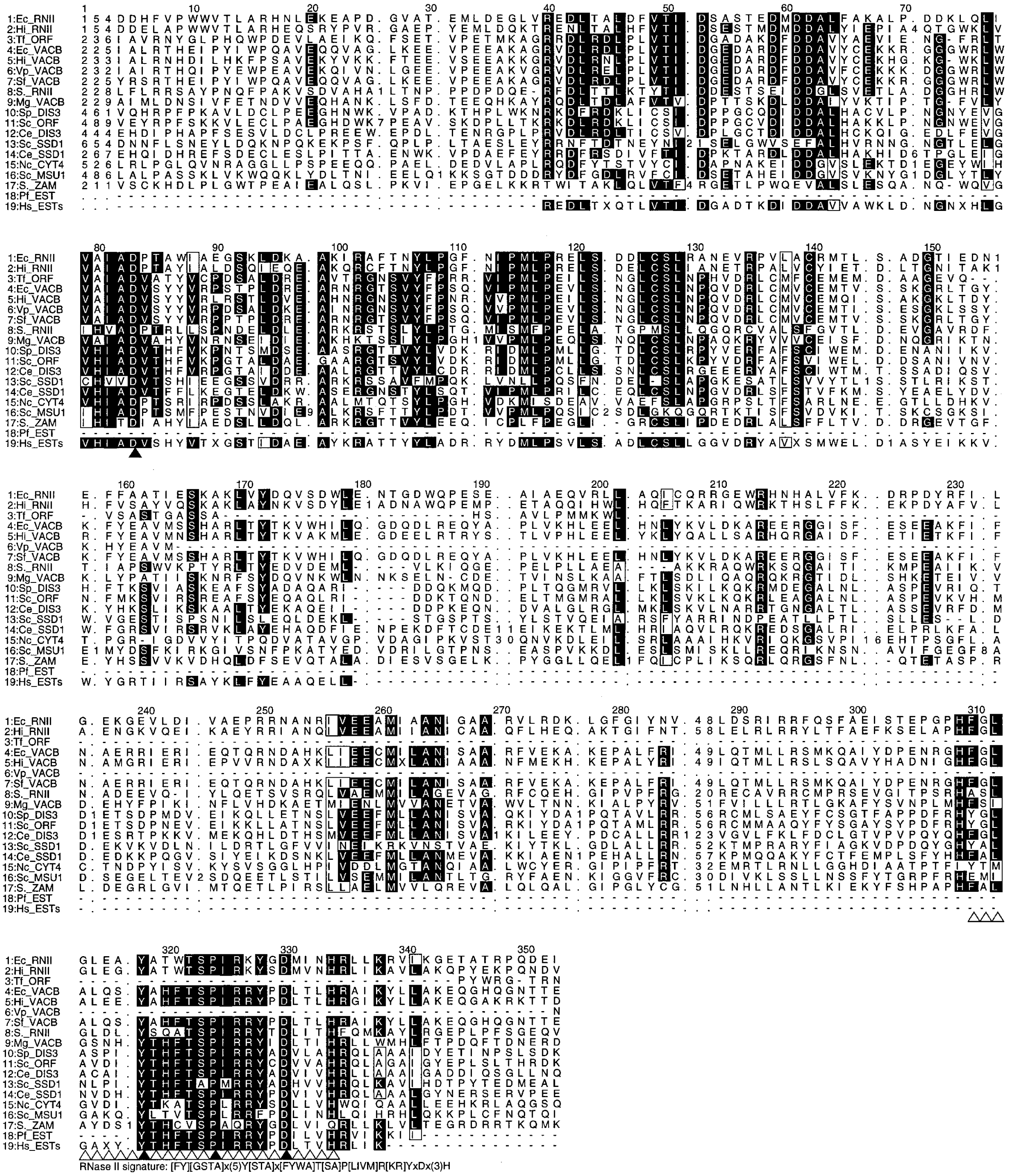

Figure 4. An HMM-generated multiple sequence alignment for RNase II family members from bacteria [1–9 and 17] and eucarya [10–16, 18 and 19]. The location of the PROSITE RNase II signature (RIBONUCLEASE_II) and the pattern itself are indicated. 1:Ec_RN11 *E.coli* RNase II (RNB_ECOLI); 2:Hi_RNII *H.influenzae* RNase II (RNB_HAEIN); 3:Tf_ORF§‡ *Thiobacillus ferrooxidans* ORF (S23260); 4:Ec_VACB *E.coli* vacB (VACB_ECOLI); 5:Hi_VACB *H.influenzae* vacB (HIU32767); 6:Vp_VACB‡ *Vibrio parahaemolyticus* vacB (VACB_VIBPA); 7:Sf_VACB *Shigella flexneri* vacB required for posttranscriptional expression of virulence genes on the large plasmid (VACB_SHIFL); 8:S._RNII *Synechocystis* sp. PCC6803 RNase II (D90904); 9:Mg_VACB *M.genitalium* vacB (MGU39690); 10:Sp_DIS3 *S.pombe* mitotic control protein dis3 (DIS3_SCHPO); 11:Sc_ORF§ *S.cerevisiae* ORF YOL021c (SCYOL021C); 12:Ce_DIS3 *C.elegans* ORF C04G2.6 similar to dis3 (CEC04G2); 13:Sc_SSD1 *S.cerevisiae* cell cycle control protein SSD1/SRK1 (SSD1_YEAST); 14:Ce_SSD1 *C.elegans* ORF F48E8.6 similar to SSD1 (CELF48E8); 15:Nc_CYT4 *Neurospora crassa* mitochondrial RNA-splicing regulatory protein phosphatase CYT-4 (A38227); 16:Sc_MSU1§ *S.cerevisiae* MSU1 essential for mitochondrial biogenesis (MSU1_YEAST); 17:S._ZAM§ *Synechocystis* sp. PCC6803 protein zam which controls resistance to the carbonic anhydrase inhibitor acetazolamide (S46946); 18:Pf_EST§‡ *Plasmodium falciparum* EST (N97483); 19:Hs_ESTs§‡ *H.sapiens* ESTs (HSA63B121, W38990, HSA54A071).

implicated in metal complex-promoted phosphodiester bond hydrolysis in a number of RNases. For example, RNase PH (17) and *S.cerevisiae* pac1 RNase (RNase III family) (55) require divalent metal ions for activity. A single magnesium (Mg^{2+}) ion-binding site containing a Glu that ligates the metal ion is essential for RNase HI catalytic activity (56). *Escherichia coli* RNase H, reverse transcriptase and reverse transcriptase-like entities in eucaryotic genomes (57), possess four invariant acidic

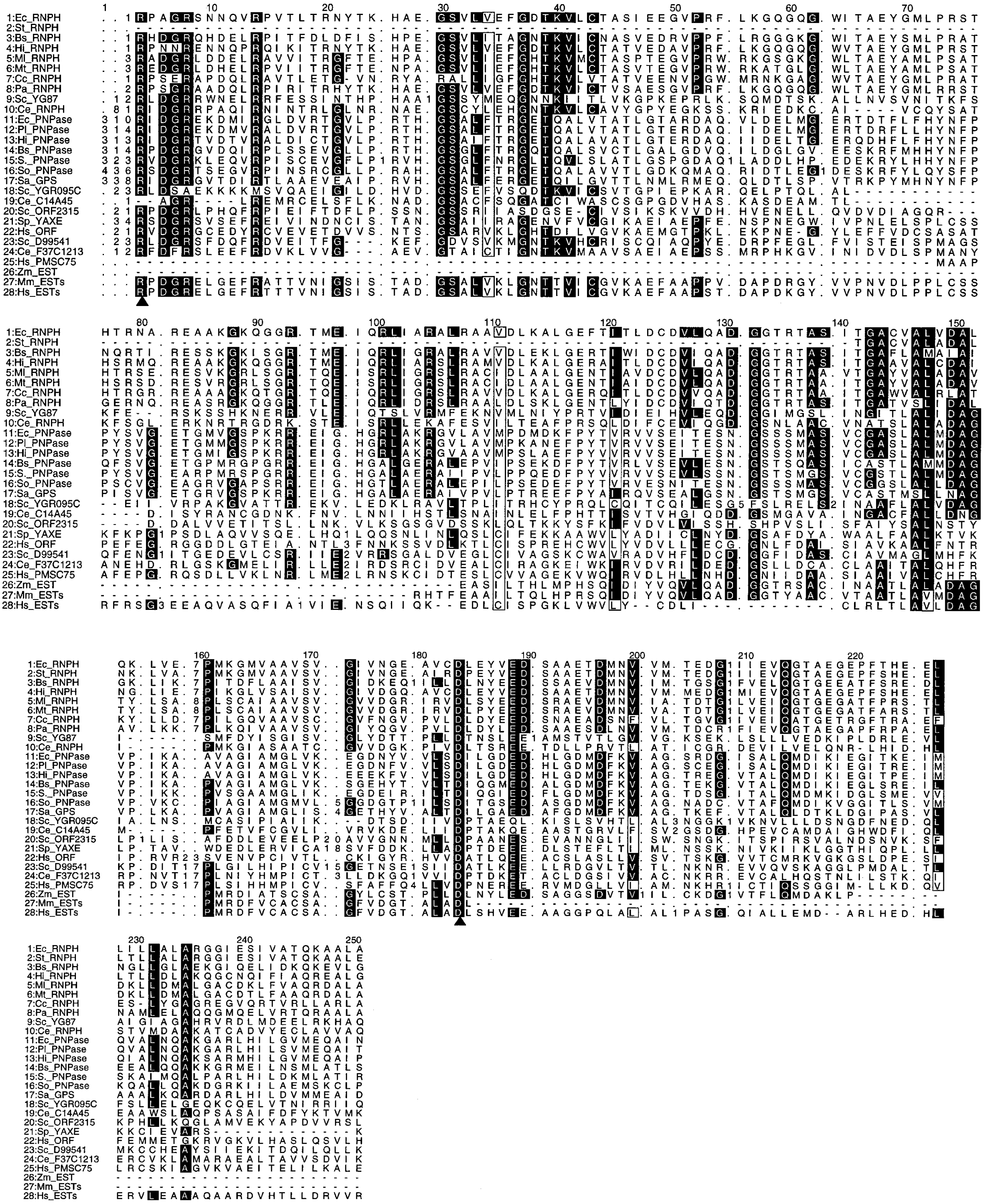

Figure 5. An HMM-generated multiple sequence alignment of RNase PH family members from bacteria [1–8 and 11–15] and eucarya [9, 10 and 16–28]. 1:Ec_RNPH *E.coli* RNase PH (RNPH_ECOLI); 2:St_RNPH‡ *S.typhimurium* RNase PH (RNPH_SALTY); 3:Bs_RNPH *B.subtilis* RNase PH (RNPH_BACSU); 4:Hi_RNPH *H.influenzae* RNase PH (RNPH_HAEIN); 5:Ml_RNPH *Mycobacterium leprae* RNase PH (RNPH_MYCLE); 6:Mt_RNPH *M.tuberculosis* RNase PH (MTCY130); 7:Cc_RNPH *C.crescentus* (CCU33324); 8:Pa_RNPH *Pseudomonas aeruginosa* RNase PH (PAU38241); 9:Sc_YG87§ *S.cerevisiae* ORF YG87 (YG87_YEAST); 10:Ce_RNPH *C.elegans* RNase PH (CEB0564); 11:Ec_PNPase§ *E.coli* PNPase (PNP_ECOLI); 12:Pl_PNPase§ *Photorhabdus luminescens* PNPase (PNP_PHOLU); 13:Hi_PNPase§ *H.influenzae* PNPase, (PNP_HAEIN); 14:Bs_PNPase§ *B.subtilis* PNPase (BSU29668); 15:S._PNPase§ *Synechocystis* sp. PCC6803 PNPase (D90899); 16:So_PNPase§ *Spinacia oleracea* PNPase (SOU52048); 17:Sa_GPS§ *Streptomyces antibioticus* guanosine pentaphosphate synthetase which has shown to be a PNPase (62) (SAU19858); 18:Sc_YGR095C§ *S.cerevisiae* ORF YGR095c (SCYGR095C); l9:Ce_C14A45§ *C.elegans* ORF C14A4.5 (CEC14A4); 20:Sc_ORF2315§ *S.cerevisiae* ORF 2315 (SCCHRIVLA); 21:Sp_YAXE§ *S.pombe* ORF YAXE (YAXE_SCHPO); 22:Hs_ORF¤ *H.sapiens* ORF (HUMORFA10); 23:Sc_D99541§ *S.cerevisiae* ORF D9954.1 (YSCD9954); 24:Ce_F37C1213§ *C.elegans* ORF F37C12.13 (CELF37C12); 25:Hs_PMSC75§ *H.sapiens* 75 kDa polymyositis/scleroderma overlap syndrome (PMSCL) autoantigen (JH0446); 26:Zm_EST‡ *Zea mays* RNase PH EST (T18324); 27:Mm_ESTs§‡ *M.muscu1us* ESTs (AA000401, W14321); 28:Hs_ESTs§‡ *H.sapiens* ESTs (W58718, HSPD03858).

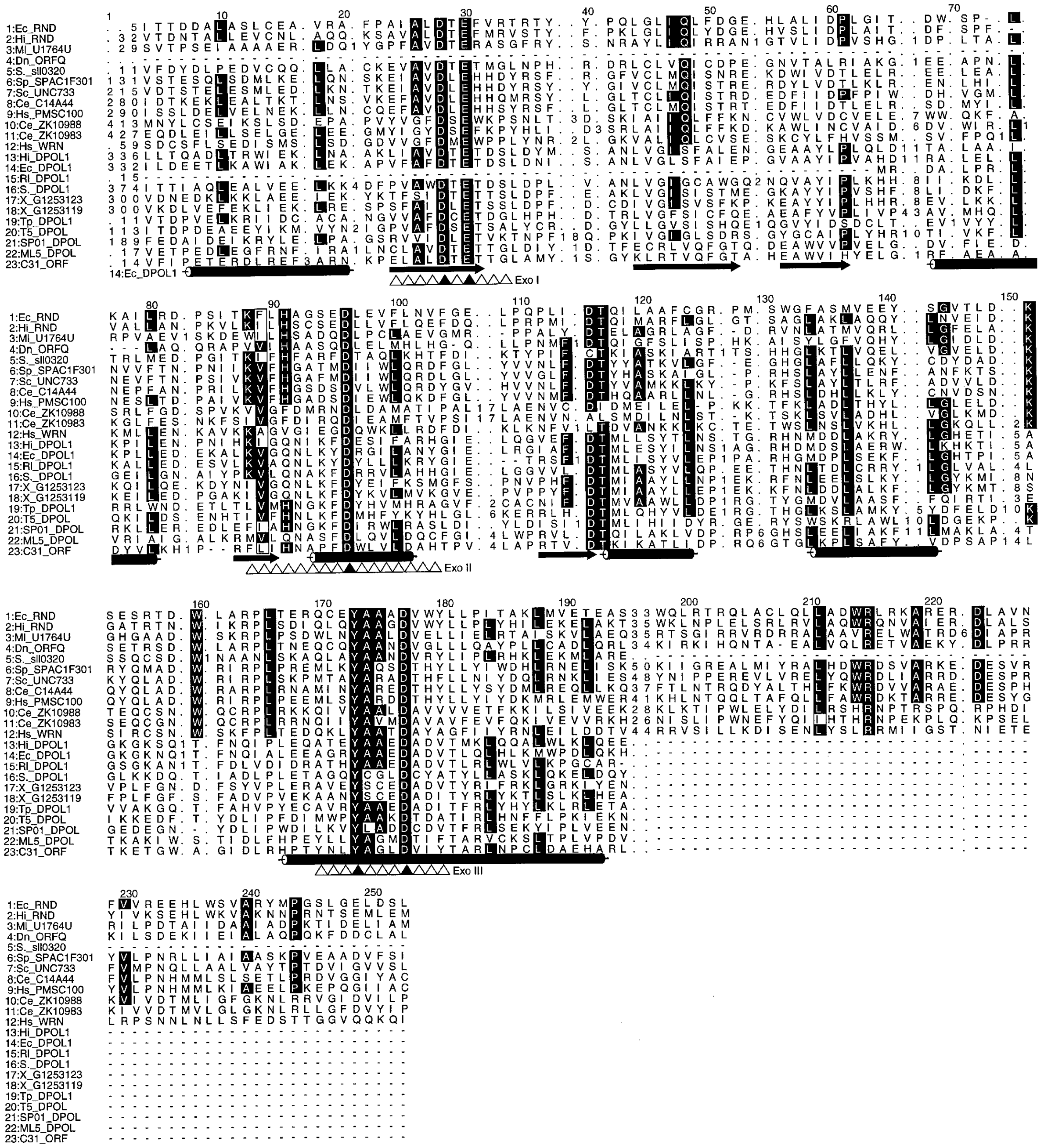

Figure 6. An HMM-generated multiple sequence alignment for the RNase D family. Cylinders and arrows denote the α-helices and β-strands in the structure of 14:Ec_DPOL1 (63–65). 1:Ec_RND *E.coli* RNase D (RND_ECOLI); 2:Hi_RND *H.influenzae* RNase D (RND_HAEIN); 3:Ml_U1764U§ *M.leprae* ORF U1764U (MLU15181); 4:Dn_ORFQ§‡ *Dichelobacter nodosus* ORFQ (DNU17138); 5:S._sII0320§‡ *Synechocystis* sp. PCC6803 ORF sII0320 (SYCSLRB); 6:Sp_SPAC1F301§ *S.pombe* ORF SPAC1F3.01 (SPAC1F3); 7:Sc_UNC733§ *S.cerevisiae* ORF UNC733 (SCU43491); 8:Ce_C14A44§ *C.elegans* ORF C14A4.4 similar to PMSCL autoantigen (CEC14A4); 9:Hs_PMSC100§ *H.sapiens* 100 kDa PMSCL autoantigen (PMSC_HUMAN); 10:Ce_ZK10988§ *C.elegans* ORF ZK1098.8 (YO68_CAEEL); 11:Ce_ZK10983§ *C.elegans* ORF ZK1098.3 (YO63_CAEEL); 12:Hs_WRN§ *H.sapiens* Werner syndrome protein (WRN) (HUMDR); 13:Hi_DPOL1§ *H.influenzae* DNA polymerase I (DPO1_HAEIN); 14:Ec_DPOL1§ *E.coli* DNA polymerase I (DPO1_ECOLI); 15:Rl_DPOL1§ *Rhizobium leguminosarum* DNA polymerase I (S43892); 16:S._DPOL1§ *Synechocystis* sp. PCC6803 DNA polymerase I (SYCSLRG); 17:X_G1253123§ sequence 12 from patent US 5466591 (1253123); 18:X_G1253119§ sequence 4 from patent US 5466591 (1253119); 19:Tp_DPOL1§ *Treponema pallidum* DNA polymerase I (TPU57757); 20:T5_DPOL§ bacteriophage T5 DNA polymerase (DPOL_BPT5); 21:SP01_DPOL§ bacteriophage SP01 DNA polymerase (DPOL_BPSP1); 22:ML5_DPOL§ Mycobacteriophage L5 DNA polymerase (DPOL_BPML5); 23:C31_ORF§ Phage ϕ-C31 ORF11 (S38923).

residues of which at least three are involved in Mg^{2+} binding in RNase H (58). Thus, invariant negatively charged residues could be metal ion ligands important for catalytic activity and/or necessary for stabilising local RNA conformation. In the RNase HII and PH families, invariant Lys, Arg and serine (Ser) may interact with key phosphate groups and/or bases in the substrate. Invariant Gly and Pro in the RNase HII, III and II families could be structurally important whereas invariant Leu and Val residues in the first two of the aforementioned RNases may participate in substrate recognition. By analogy with the RNase D family, the invariant Tyr in RNase II may be an active site residue that interacts with the scissile phosphate of the mRNA substrate. Although invariant and other conserved residues are scattered throughout the primary sequences of the families, folding of the RNases may juxtapose them to form a metal ion-containing active site. Finally, reversible phosphorylation of residues situated in the substrate binding site may be a means to regulate these RNases and the pathways in which they act. Thus, phosphorylation of such Ser, Thr and Tyr residues could lower affinity for RNA by increasing electrostatic repulsion between the phosphate backbone and the phosphorylated amino acid.

ACKNOWLEDGEMENTS

This work was supported by NIH grant number GM17129 and by the Director, Office of Energy Research, Office of Health and Environmental Research, Division of the US Department of Energy under Contract No. DE-AC03-76F00098. The data and multiple alignments are available in electronic form upon request.

REFERENCES

1 Schmid,H., Pouch,M., Petit,F., Dadet,M., Badaoui,S., Boissonnet,G., Buri,J., Norris,V. and Briand,Y. (1995) *Mol. Biol. Rep.*, **21**, 43–47.
2 Deutscher,M. (1993) *J. Biol. Chem.*, **268**, 13011–13014.
3 Apirion,D. and Miczak,A. (1993) *Bioessays*, **15**, 113–120.
4 Rost,B. (1995) *Intelligent Systems Mol. Biol*. **3**, 314–321.
5 Golz,J., Clarke,A. and Newbigin,E. (1995) *Curr. Opin. Genet. Dev.*, **5**, 640–645.
6 Kao,T. and McCubbin,A. (1996) *Proc. Natl. Acad. Sci. USA*, **93**, 12059–12065.
7 Krogh,A., Brown,M., Mian,I.S., Sjölander,K. and Haussler,D. (1994) *J. Mol. Biol.*, **235**, 1501–1531.
8 Baldi,P., Chauvin,Y., Hunkapiller,T. and McClure,M. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 1059–1063.
9 Eddy,S. (1996) *Curr. Opin. Struct. Biol.*, **6**, 361–365.
10 Fujiwara,Y., Asogawa,M. and Konagaya,A. (1994) *Intelligent Systems Mol. Biol*. **2**, 121–129.
11 Li,Z. and Deutscher,M. (1994) *J. Biol. Chem.*, **269**, 6064–6071.
12 Reuven,N. and Deutscher,M. (1993) *FASEB J.*, **7**, 143–148.
13 Tomasiewicz,H. and McHenry,C. (1987) *J. Bacteriol.*, **169**, 5735–5744.
14 Itaya,M. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 8587–8591.
15 Elela,S., Igel,H. and Ares,M.,Jr (1996) *Cell*, **85**, 115–124.
16 Jensen,K.F., Andersen,J.T. and Poulsen,P. (1992) *J. Biol. Chem.*, **267**, 17147–17152.
17 Kelly,K.O. and Deutscher,M.P. (1992) *J. Biol. Chem.*, **267**, 17153–17158.
18 Coburn,G. and Mackie,G. (1996) *J. Biol. Chem.*, **271**, 15776–15781.
19 Zhang,J. and Deutscher,M. (1988) *J. Biol. Chem.*, **263**, 17909–17912.
20 Putnam,D. (1996) *Am. J. Health-System Pharmacy*, **53**, 151–160.
21 Wang,J., Qiu,L., Wu,E. and Drlica,K. (1996) *J. Bacteriol.*, **178**, 1640–1645.
22 Pepe,C., Maslesa-Galic,S. and Simons,R. (1994) *Mol. Microbiol.*, **13**, 1133–1142.
23 Altschul,S., Gish,W., Miller,W., Myers,E. and Lipman,D. (1990) *J. Mol. Biol.*, **215**, 403–410.
24 Hughey,R. and Krogh,A. (1996) *Comp. Appl. Biosci.*, **12**, 95–107. The hidden Markov model software can be accessed from http://www.cse.ucsc.edu/research/compbio/sam.html.
25 Brown,M., Hughey,R., Krogh,A., Mian,I., Sjölander,K. and Haussler,D. (1993) *Intelligent Systems Mol. Biol*. **1**, 47–55.
26 Sjölander,K., Karplus,K., Brown,M., Hughey,R., Krogh,A., Mian,I. and Haussler,D. (1996) *Comp. Appl. Biosci.*, **12**, 327–345.
27 Altschul,S. (1991) *J. Mol. Biol.*, **219**, 555–565.
28 Barrett,C., Hughey,R. and Karplus,K. (1997) *Comp. Appl. Biosci.*, **13**, 191–199.
29 NCI (1996) NRP (Non-Redundant Protein) and NRN (Non-Redundant Nucleic Acid) Database. Distributed on the Internet via anonymous FTP from ftp.ncifcrf.gov, under the auspices of the National Cancer Institute’s Frederick Biomedical Supercomputing Center.
30 Barton,G. (1993) *Protein Engng*, **6**, 37–40.
31 Rotondo,G., Gillespie,M. and Frendewey,D. (1995) *Mol. Gen. Genet.*, **247**, 698–708.
32 Kutish,G., Li,Y., Lu,Z., Furuta,M., Rock,D.L. and Van Etten,J. (1996) *Virology*, **223**, 303–317.
33 St. Johnston,D., Brown,N., Gall,J. and Jantsch,M. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 10979–10983.
34 Gatignol,A., Buckler,C. and Jeang,K. (1993) *Mol. Cell. Biol.*, **13**, 2193–2202.
35 Zilhao,R., Camelo,L. and Arraiano,C. (1993) *Mol. Microbiol.*, **8**, 43–51.
36 Tobe,T., Sasakawa,C., Okada,N., Honma,Y. and Yoshikawa,M. (1992) *J. Bacteriol.*, **174**, 6359–6367.
37 Kinoshita,N., Goebl,M. and Yanagida,M. (1991) *Mol. Cell. Biol.*, **11**, 5839–5847.
38 Sutton,A., Immanuel,D. and Arndt,K. (1991) *Mol. Cell. Biol.*, **11**, 2133–2148.
39 Wilson,R., Brenner,A., White,T., Engler,M., Gaughran,J. and Tatchell,K. (1991) *Mol. Cell. Biol.*, **11**, 3369–3373.
40 Turcq,B., Dobinson,K., Serizawa,N. and Lambowitz,A. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 1676–1680.
41 Bult,C., White,O., Olsen,G., Zhou,L., Fleischmann,R., Sutton,G., Blake,J., FitzGerald,L., Clayton,R., Gocayne,J. *et al.* (1996) *Science*, **273**, 1058–1073.
42 Davidov,Y., Rahat,A., Flechner,I. and Pines,O. (1993) *J. Gen. Microbiol.*, **139**, 717–724.
43 Li,H. and Nicholson,A. (1996) *EMBO J.*, **15**, 1421–1433.
44 Hofmann,K. and Bucher,P. (1995) *Trends Biochem. Sci.*, **20**, 347–349.
45 Stone,J., Collinge,M., Smith,R., Horn,M. and Walker,J. (1994) *Science*, **266**, 793–795.
46 Navas,T., Zhou,Z. and Elledge,S. (1995) *Cell*, **80**, 29–39.
47 Xu,F. and Cohen,S. (1995) *Nature*, **374**, 180–183.
48 Py,B., Causton,H., Mudd,E. and Higgins,C. (1994) *Mol. Microbiol.*, **14**, 717–729.
49 Plotz,P., Rider,L., Targoff,I., Raben,N., O’Hanlon,T. and Miller,F. (1995) *Ann. Internal Med.*, **122**, 715–724.
50 Epstein,C. and Motulsky,A. (1996) *Bioessays*, **18**, 1025–1027.
51 Yu,C., Oshima,J., Fu,Y., Wijsman,E., Hisama,F., Alisch,R., Matthews,S., Nakura,J., Miki,T., Ouais,S., Martin,G., Mulligan,J. and Schellenberg,G. (1996) *Science*, **272**, 258–262.
52 Bernad,A., Blanco,L., Lazaro,J., Martin,G. and Salas,M. (1989) *Cell*, **59**, 219–228.
53 Wang,J., Yu,P., Lin,T., Konigsberg,W. and Steitz,T. (1996) *Biochemistry*, **35**, 8110–8119.
54 Joyce,C. and Steitz,T. (1994) *Annu. Rev. Biochem.*, **63**, 777–822.
55 Rotondo,G. and Frendewey,D. (1996) *Nucleic Acids Res.*, **24**, 2377–2386.
56 Katayanagi,K., Okumura,M. and Morikawa,K. (1993) *Proteins*, **17**, 337–346.
57 Doolittle,R., Feng,D., Johnson,M. and McClure,M. (1989) *Q. Rev. Biol.*, **64**, 1–30.
58 Katayanagi,K., Miyagawa,M., Matsushima,M., Ishikawa,M., Kanaya,S., Nakamura,H., Ikehara,M., Matsuzaki,T. and Morikawa,K. (1992) *J. Mol. Biol.*, **223**, 1029–1052.
59 Briggs,M., Dacanay,K. and Butler,J. (1997) *Mol. Cell. Biol.*, in press.
60 Kharrat,A., Macias,M., Gibson,T., Nilges,M. and Pastore,A. (1995) *EMBO J.*, **14**, 3572–3584.
61 Bycroft,M., Grünert,S., Murzin,A., Proctor,M. and St Johnston,D. (1995) *EMBO J.*, **14**, 3563–3571.
62 Jones,G. and Bibb,M. (1996) *J. Bacteriol.*, **178**, 4281–4288.
63 Derbyshire,V., Freemont,P., Sanderson,M., Beese,L., Friedman,J., Joyce,C. and Steitz,T. (1988) *Science*, **240**, 199–201.
64 Beese,L. and Steitz,T. (1991) *EMBO J.*, **10**, 25–33.
65 Beese,L., Derbyshire,V. and Steitz,T. (1993) *Science*, **260**, 352–355.

Note added in proof

Sequence 7:Sc_UNC733 in Figure 6 is identical to Rrp6p which is essential for efficient 5.8S rRNA 3′ end processing (59). Rrp6p, 25:Hs_PMSC75 and bacterial RNase Ds have been suggested to function as 3′→5′ exoribonucleases that trim the 3′ end of specific RNA structures to within 3 or 4 nucleotides of a stable base-paired stem (59).