5110–5118 *Nucleic Acids Research, 1997, Vol. 25, No. 24*

The proofreading domain of *Escherichia coli* DNA polymerase I and other DNA and/or RNA exonuclease domains

Michael J. Moser, William R. Holley, Aloke Chatterjee and I. Saira Mian*

Life Sciences Division (Mail Stop 29-100), Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA

Received June 5, 1997; Revised and Accepted October 15, 1997

ABSTRACT

Prior sequence analysis studies have suggested that bacterial ribonuclease (RNase) Ds comprise a complete domain that is found also in *Homo sapiens* polymyositis-scleroderma overlap syndrome 100 kDa autoantigen and Werner syndrome protein. This RNase D 3′→5′ exoribonuclease domain was predicted to have a structure and mechanism of action similar to the 3′→5′ exodeoxyibonuclease (proofreading) domain of DNA polymerases. Here, hidden Markov model (HMM) and phylogenetic studies have been used to identify and characterise other sequences that may possess this exonuclease domain. Results indicate that it is also present in the RNase T family; *Borrelia burgdorferi* P93 protein, an immunodominant antigen in Lyme disease; bacteriophage T4 dexA and *Escherichia coli* exonuclease I, processive 3′→5′ exodeoxyribonucleases that degrade single-stranded DNA; *Bacillus subtilis* dinG, a probable helicase involved in DNA repair and possibly replication, and peptide synthase 1; *Saccharomyces cerevisiae* Pab1p-dependent poly(A) nuclease PAN2 subunit, required for shortening mRNA poly(A) tails; *Caenorhabditis elegans* and *Mus musculus* CAF1, transcription factor CCR4-associated factor 1; *Xenopus laevis* XPMC2, prevention of mitotic catastrophe in fission yeast; *Drosophila melanogaster* egalitarian, oocyte specification and axis determination, and exuperantia, establishment of oocyte polarity; *H.sapiens* HEM45, expressed in tumour cell lines and uterus and regulated by oestrogen; and 31 open reading frames including one in *Methanococcus jannaschii*. Examination of a multiple sequence alignment and two three-dimensional structures of proofreading domains has allowed definition of the core sequence, structural and functional elements of this exonuclease domain.

INTRODUCTION

Exonucleases are essential components of many processes such as replication, recombination, repair, turnover, processing and stability. For example, the cellular tumour antigen p53 exhibits 3′→5′ exonuclease activity although this function in p53's role as 'guardian of the genome' is unknown (1). In contrast, the 3′→5′ exodeoxyribonuclease (or proofreading) domain of DNA polymerases has been well studied and shown to possess three characteristic sequence motifs termed Exo I, Exo II and Exo III (2–5). The three-dimensional structures of this domain in *Escherichia coli* Klenow fragment (6–10) and bacteriophage T4 DNA polymerase (11) are similar despite limited sequence identity. The Exo I, II and III motifs are clustered around the active site and contain four negatively charged residues that serve as ligands for the two metal ions required for catalysis as well as a catalytically active tyrosine (Fig. 1). A nucleophilic attack on the phosphorus atom of the terminal nucleotide is postulated to be performed by a hydroxide ion that is activated by one divalent metal ion whilst the expected pentacoordinate transition state and the leaving oxyanion are stabilised by a second divalent metal in close proximity to the first (9). Mutations at the conserved active site positions (red, Fig. 1) affect proofreading activity (reviewed in ref. 12).

Escherichia coli ribonuclease (RNase) D is one of at least five 3′→5′ exoribonucleases required for 3′-end processing of tRNA precursors (13,14). A previous study (15) of the RNase D family of sequences using hidden Markov models (HMMs) indicated that bacterial RNase Ds comprise a complete domain that is present in some eucaryotic proteins including *Saccharomyces cerevisiae* Rrp6p (called Sce_UNC733 in ref. 15), *Homo sapiens* polymyositis-scleroderma overlap syndrome 100 kDa autoantigen (PM-Scl 100) and Werner syndrome protein. Furthermore, this RNase D domain appeared to be similar to the aforementioned proofreading domain suggesting a common structure and mechanism of action for exonucleases acting on DNA and/or RNA. These computational results are consistent with recent experimental studies which suggest that Rrp6p is essential for efficient 5.8S rRNA 3′-end processing (16). Rrp6p, PM-Scl 100 and RNase D have been proposed to function as 3′→5′ exoribonucleases that trim the 3′ end of specific RNA structures to within 3 or 4 nt of a stable base-paired stem (16). Sequence analysis of positionally cloned human disease genes using a different approach (17) has suggested the presence of a nuclease domain homologous to bacterial RNase D and the 3′→5′ exonuclease domain of DNA polymerase I in Werner syndrome protein.

*To whom correspondence should be addressed. Tel: +1 510 486 6216; Fax: +1 510 486 6949; Email: smian@lbl.gov

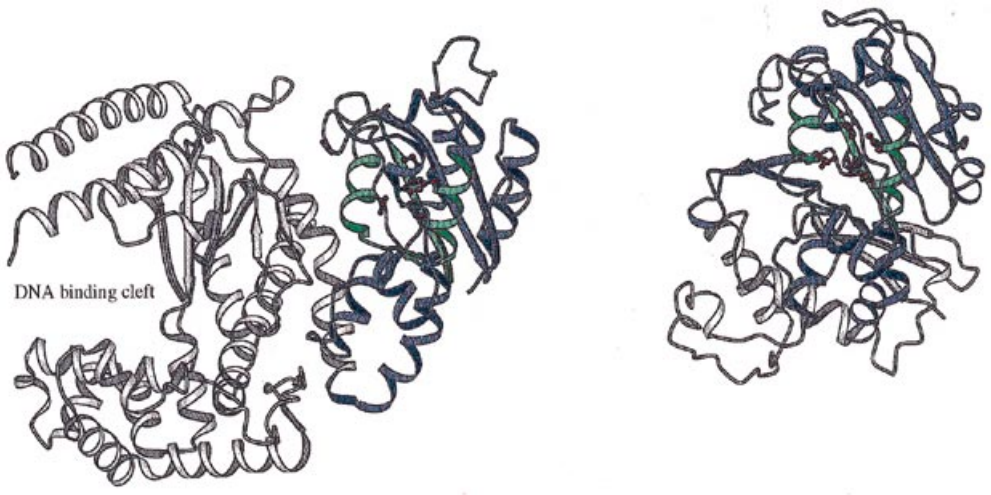

Figure 1. Ribbon diagrams of the Klenow fragment of *E.coli* DNA polymerase I (left, PDB code 1kfd) and T4 DNA polymerase (right, 1noy). The coloured regions represent the exonuclease (proofreading) domain modelled in this work with the Exo I, II and III motifs in green and side chains of the proofreading active site residues in red. The molecules are shown such that the proofreading domains are in approximately the same orientation. Arrows and cylinders denote β-strands and α-helices and are taken from the crystal structures.

The aim of this work is to extend the previous HMM analysis of the proofreading/RNase D exonuclease domain and to identify other proteins that may possess this exonuclease domain. HMMs are a statistical modelling method (18,19) that have used recently to characterise the common features of a family of related sequences, generate a multiple sequence alignment and recognise related, but divergent family members present in databases (15,20–29). HMMs used to model sequence families can be viewed as 'profiles' recast in a probabilistic framework. A profile is a model for a family consisting of a primary sequence consensus and position-specific residue scores and insertion/deletion penalties (30–34). The results here indicate that the RNase T family possesses not only the Exo I, II and III motifs as suggested elsewhere (35), but the complete exonuclease domain. *Escherichia coli* RNase T is responsible for the 3′ processing and end-turnover of tRNA and the maturation of 5S rRNA (36,37). In addition to DNA proofreading enzymes and the RNase D and T families, a large number of proteins of known and unknown function possess this exonuclease domain (~140 positions in length). The HMM, phylogenetic and structural analyses of the exonuclease domain provide guidance for experimental studies aimed at understanding the diverse functions of proteins possessing this domain.

MATERIALS AND METHODS

Statistical modelling: hidden Markov model

Hidden Markov model creation, training and use was performed with v2.0 of the SAM (Sequence Alignment and Modeling Software System) suite (20,38) running on a MASPAR MP-2204 with a DEC Alpha 3000/300X frontend at the University of California Santa Cruz (UCSC). The HMM trained to model the exonuclease domain of the RNase D family and selected proofreading domains (23 sequences in total) (15) was used as the starting HMM. In Gram+ bacterial DNA polymerase III, the absence of a conventional Exo III motif has led to the identification of a region termed motif III ε (39). The sequence motifs conserved between these proofreading enzymes and RNase T sequences were used as a guide to align RNase T sequences to the starting HMM and thus to create and train an initial exonuclease domain HMM using the aforementioned sequences (the training set). To improve the ability of the HMM to generalise, Dirichlet mixture priors (40,41) were employed. Free Insertion Modules (FIMs) were utilised to allow an arbitrary number of insertions at either end of the HMM to accomodate exonuclease domains that occured within larger sequences.

The SAM programme hmmscore and the initial HMM were used for HMM database searches by calculating log-odds scores (42,43) for all sequences in a non-redundant protein database obtained from the NCI (44) and updated weekly at UCSC. The log-odds score for a sequence is the negative (natural) log-likelihood of the sequence given the model minus a NULL model (a simple FIM loop) (43). The significance of log-odds scores can be ascertained by evaluating *E*, the expected number of false positives above a given log-odds score in a given database search. Since the NULL model, assumed to be a reasonably accurate description of the space the sequences are drawn from, is unlikely to be a good model for the score distribution of all 'random' sequences, the *E* value calculated by SAM is not a true estimate of *E* but an upper bound. SAM log-odds scores provide a conservative estimate of the significance of scores arising from a database search.

Taking into account the number of sequences in the database searched (~242 000 different proteins in early 1997), a significant log-odds score is considered to be 22.7, the value at which $E = 0.01$. Log-odds scores higher than this value denote fewer expected false positives. The approach employed here emphasises training an HMM that discriminates between training and non-training set sequences, i.e., one in which the gap in log-odds scores between the lowest scoring training set sequence and the highest scoring non-training set (database) sequence is relatively large (usually >5.0) and the absolute log-odds score for the lowest training set sequence is >22.7. In addition, efforts were made during training to ensure that, as far as possible, training resulted in an HMM

capable of yielding an alignment such that known enzymatic elements (Exo I, II and III) aligned.

A database search with the initial HMM revealed a number of sequences with log-odds scores higher than or close to that of the lowest scoring training set sequence. The alignment of sequences with log-odds scores >22.7 was examined and those which possessed regions conserved in the initial HMM (primarily Exo I, II and III) were retained and added to the training set. The HMM was then retrained with this expanded training set. Further rounds of 'search, align and retrain' revealed fewer and fewer new sequences with the domain. The gap in log-odds scores between training set and non-training set sequences remained relatively constant. At this point (April 1997) and after ~40 iterations, a final HMM was trained and used for subsequent studies. To avoid overrepresentation of proofreading domains, only sequences from a diverse range of organisms (for example *S.cerevisiae*, *Schizosaccharomyces pombe*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *H.sapiens*) were used for training. Experiments (data not shown) indicated that the excluded sequences were sufficiently similar to one or more training set sequence such that their log-odds scores were in the range of those in the training set.

Phylogenetic analysis

An HMM-generated alignment of the training set containing only match and delete states was utilised for phylogenetic studies. Insert states are not modelled by an HMM because the regions in a sequence they represent are the most divergent parts of the molecules and are likely to be sources of systematic error in phylogenetic analysis. The MOLPHY suite uses a probabilisitic procedure for inferring phylogenetic relationships (45,46). The protml programme in MOLPHY v2.3 was used to generate a maximum likelihood distance matrix using the default JTT model. NJdist and bootstrapping were then used to infer a number of approximate trees from this distance matrix by the neighbour-joining method. Approximate bootstrap probabilities for these trees were computed using the RELL method. Starting from these initial trees, repeated local rearrangements were employed to search for tree better topologies. Amongst these final trees, the one with the highest likelihood was selected.

Figures showing multiple sequence alignments, phylogenetic trees and ribbon diagrams of molecules were produced using ALSCRIPT (47), Treetool (48) and MOLSCRIPT (49), respectively.

RESULTS

Proteins possessing the exonuclease domain

The HMM itself, a complete list of the 148 sequences that comprised the final training set and a multiple sequence alignment of the final training set can be found as supplementary data to this paper via NAR Online (http://www.oup.co.uk/nar). Of ~242 000 sequences searched using the final HMM, the lowest scoring training sequence (sequence name abbreviated to Bsu_PPS1) had a log-odds score of 32.2. Only homologues that had been excluded from HMM training, principally proofreading domains, had log-odds scores >32.2 (data not shown). All other sequences had log-odds scores <26.6. For a log-odds score of 32.2, $E = 7 \times 10^{-7}$ so the sequences in the electronic appendix and the aforementioned homologues are considered to possess an exonuclease domain. The 44 sequences not previously identified as possessing the exonuclease domain are given in the electronic appendix together with additional data where available. The domains in DNA polymerases (2–5) and RNase D (15) have been described in detail elsewhere and so will not be discussed further.

Figure 2 shows an HMM-generated alignment of selected exonuclease domains (subsequent discussions will be based primarily on the alignment of all the training set that can be found in NAR Online supplementary data). Although the exonuclease domain is known to be present in the DNA polymerases of double-stranded (ds) DNA viridae that do not have an RNA stage, the results here indicate this domain may be more widespread because of its occurrence in lactic dehydrogenase virus, a single-stranded (ss) RNA virus without a DNA stage (LDV_ORF1a). Also noteworthy is the presence of the domain in an open reading frame (ORF) from the archaeon *Methanococcus jannaschii* (Mja_MJ0365). The domain forms both complete proteins as well as comprising a domain in larger sequences. In Herpes simplex virus type 1 DNA polymerase, mutation of the active site tyrosine (red, labelled D) reduces both exonuclease and polymerase activities (50). Thus, mutations at the three Exo motifs may affect not only exonuclease activity, but also functions associated with the remainder of the proteins.

The known functions of sequences that possess the domain include DNA replication, repair and recombination, transcription, the initial transport, localisation and long-term maintenence of cytoplasmic RNAs during development, nucleocytoplasmic export and/or processing of RNA and cell cycle progression. Hence, ORFs that possess the domain may have roles in these processes. For example, in many species, early development and/or pattern formation requires targetted movements of molecules and molecular aggregates leading to an asymmetric distribution of proteins within the cell. Some of the new exonuclease domains may play a role in these events. *Drosophila melanogaster* egalitarian (Dme_EGL) and exuperantia (Dme_EXU, Dps_EXU1 in supplementary data) may help to localise mRNA transcripts to opposite poles of the oocyte during development by degradation of the 3′ UTR regions that are known to contain the targetting signal (51). Egalitarian is part of a complex with BicaudalD that has been suggested to link microtubule polarity and RNA transport during oogenesis (52).

Exonuclease domain HMM

The number of expected false positives amongst sequences with log-odds scores >22.7 is 0.01. Sequences with scores in the range 22.7–32.0 are: *Mycoplasma pneumoniae* chromosomal replication initiator protein DNAA (log-odds score 26.6, databank code DNAA_MYCPN); *Drosophila pseudoobscura* exuperantia 2 protein (25.7, EXU2_DROPS); and *Mycobacterium tuberculosis* phosphate transport system permease protein pstA-2 (24.2, MTCY8D9). Other exuperantia proteins are part of the training set (Dme_EXU, Dps_EXU in supplementary data). Inspection of an HMM-generated alignment (data not shown) gives an indication of the residues corresponding to the active site positions A–E of Figure 2 in DNAA ($Asp44^{A}$, $Glu46^{B}$, $Ser102^{C}$, $Ile303^{D}$, $Asp307^{E}$), exuperantia 2 ($Glu36^{A}$, $Asp38^{B}$, $Asp136^{C}$, $Leu247^{D}$, $Glu251^{E}$) and pstA-2 ($Asp71^{A}$, $Gln73^{B}$, $Asp148^{C}$, $Asp268^{D}$, $Asp272^{E}$). Homologues of the latter two had slightly lower scores (*Mycoplasma genitalium* DNAA, 18.3; *M.genitalium* pstA, 18.2) suggesting that DNAA and pstA-2 may possess exonuclease domains. Inclusion of these sequences, DNAA in particular, into the training set for future refinements of the HMM may be

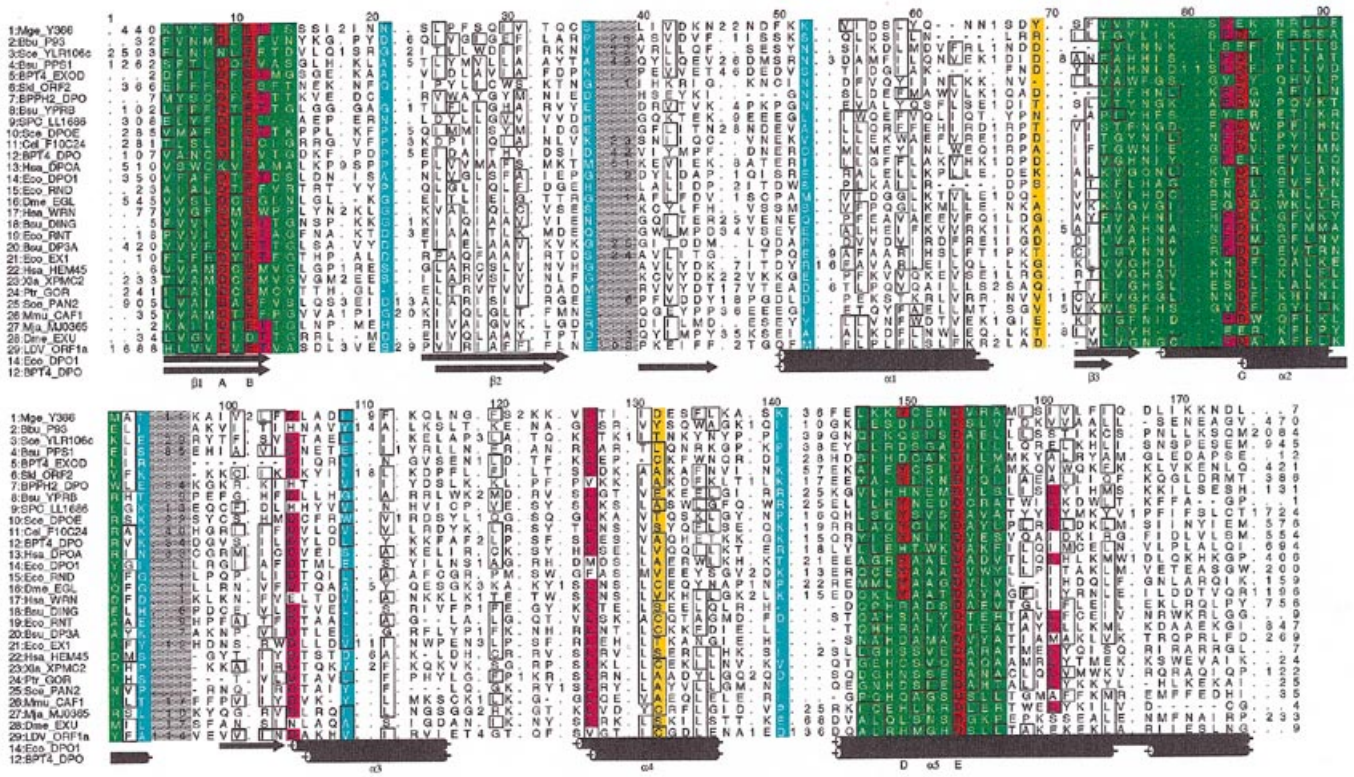

Figure 2. An HMM-generated alignment of selected exonuclease domains. Amino acids conserved in the majority of the sequences are highlighted and columns that are predominantly hydrophobic are boxed. Columns containing a full point correspond to insert states and numbers indicate the length of insertions at that position (if present). The secondary structures of two proofreading domains, Eco_DPO1 and BPT4_DPO, are shown and correspond to the regions coloured blue, green and red in Figure 1. α-Helices and β-strands conserved between the domains are labelled sequentially and not according to the designations in the individual Protein Databank entries. The Exo I, II and III motifs are green, active site residues are red (labelled A–E), residues preceeding insert states where two or more sequences have insertions longer than 10 residues are cyan, conserved residues are magenta and residues which correspond to Cys 112 and 168 in *E.coli* RNase T (Eco_RNT) are yellow. The sequences shown are as follows (exonuclease domains identified here are in bold text). **Mge_Y366**, *M.genitalium* ORF (databank code Y366_MYCGE); **Bbu_P93**, *Borrelia burgdorferi* P93 protein, an immunodominant antigen present in patients with Lyme disease (70) (BBTROP93); **Sce_YLR106c**, *S.cerevisiae* probable membrane protein YLR106c (S64942); **Bsu_PPS1**, *B.subtilis* peptide synthetase 1 (71) (PPS1_BACSU); **BPT4_EXOD**, bacteriophage T4 dexA protein, a processive 3′→5′ exodeoxyribonuclease that degrades single-stranded (ss) DNA (72,73) (EXOD_BPT4); **Skl_ORF2**, *Saccharomyces kluyveri* ORF 2 (S15961); BPPH2_DPO, bacteriophage Phi-29 DNA polymerase (DPOL_BPPH2); **Bsu_YPRB**, *B.subtilis* ORF yprB (BACPONAYPP); **SPC_LL1686**, *Synechocystis* ORF sll1686 (D90900); Sce_DPOE, *S.cerevisiae* DNA polymerase ε (DPOE_YEAST); **Cel_F10C24**, *C.elegans* ORF F10C2.4 (CEF10C2); BPT4_DPO, bacteriophage T4 DNA polymerase (DPOL_BPT4); Hsa_DPOA, *H.sapiens* DNA polymerase α (DPOA_HUMAN); Eco_DPO1, *E.coli* DNA polymerase I (DPO1_ECOLI); Eco_RND, *E.coli* RNase D (RND_ECOLI); **Dme_EGL**, *D.melanogaster* egalitarian, co-localises with BicaudalD and is involved in oocyte specification and axis determination. Its complex with BicD may link microtubule polarity and RNA transport (52) (DMU86404); Hsa_WRN, *H.sapiens* Werner syndrome protein (HUMDR); **Bsu_DING**, *B.subtilis* dinG, a probable ATP-dependent helicase involved in DNA repair and perhaps replication (74) (DING_BACSU); Eco_RNT, *E.coli* RNase T (RNT_ECOLI); Bsu_DP3A, *B.subtilis* DNA polymerase III α chain (DP3A_BACSU); **Eco_EX1**, *E.coli* exodeoxyribonuclease I (also called SbcB or xona), a processive 3′→5′ exodeoxyribonucleases that degrades ss DNA (75) and has been suggested to play a role in the the RecBCD-dependent recombination pathway (76) (EX1_ECOLI); **Hsa_HEM45**, *H.sapiens* HEM45, expressed in tumour cell lines and rat uterus and regulated by oestrogen (HSU88964); **Xla_XPMC2**, *X.laevis* XPMC2, a nuclear protein which prevents premature entry into mitosis (termed mitotic catastrophe) in fission yeast (77) (S53818); **Ptr_GOR**, *Pan troglodytes* GOR antigen, antibodies against GOR are present in individuals with hepatitis C (GOR_PANTR); **Sce_PAN2**, *S.cerevisiae* Pab1p-dependent poly(A) nuclease PAN2 subunit, required for shortening mRNA poly(A) tails (60,61) (SCU39204); **Mmu_CAF1**, *M.musculus* CCR4-associated factor 1 (CAF1), CAF1 from *S.cerevisiae* (also called POP2) and *C.elegans* interact genetically with the CCR4 component of the transcriptional regulatory complex *in vivo* and POP2 associates physically with CCR4 (63) (MMU21855); **Mja_MJ0365**, *M.jannaschii* ORF MJ0365 (MJU67489); **Dme_EXU**, *D.melanogaster* maternal exuperantia, ensures the proper localisation of bicoid mRNA to the anterior region of the oocyte and thus establishment of oocyte polarity (78–80) (EXU_DROME); **LDV_ORF1a**, lactic dehydrogenase virus polyprotein ORF1a (LDVGLYPOL).

warranted but awaits assessment of the false positive rate by experimental validation of exonuclease activity in the new domains identified here.

All other database sequences had scores <20.7 and include some potential false negatives such as proteins implicated in nuclear-cytoplasmic transport. One noteworthy example of a possible false negative is *S.cerevisiae* transcription initiation protein SPT6 (15.6, SPT6_YEAST) in which residues 417–624 match the HMM and which has active site residues identical to those of most of the other exonuclease domains (Asp421^{A}, Glu423^{B}, Asp559^{C}, Tyr600^{D}, Asp603^{E}). Two other transcription-associated proteins (Mmu_CAF1 in Fig. 2, Cel_CAF1 in supplementary data) are part of the training set. Both the log-odds score and alignment to the HMM suggest that p53 does not appear to possess the exonuclease domain modelled here. This may be because either p53 possesses a divergent form of this domain which cannot be identified by the current HMM because it is too specific, or because p53 has an exonuclease domain unrelated to that examined here.

Although the high log-odds scores of *M.musculus* and *C.elegans* CAF1 indicate that they possess an exonuclease domain, the very

low log-odds score of *S.cerevisiae* CAF1 (2.0) suggests that it does not. In contrast, BLAST database searches run with default parameters (53) and the metazoan HMM-defined domain indicate that *S.cerevisiae* CAF1 possesses significantly related regions ($P < 10^{-42}$), some of which correspond to insertions in the HMM. Inspection of an HMM-generated alignment suggests that in spite of considerable overall sequence similarity, the apparent discrepancy lies in changes to the key catalytic residues in Exo I and II as well as the (probable) absence of Exo III in *S.cerevisiae* CAF1. It is unclear whether *S.cerevisiae* CAF1 possesses an extremely divergent exonuclease domain that has a tertiary structure and/or activity similar to that in metazoan CAF1. Two similar discrepencies between BLAST and the HMM results occur (Bsu_PPS1 and Xla_XPMC2, data not shown). These results suggest that one origin for the difference may lie in how SAM scores HMMs and evaluates the significance of the score.

Further analysis is required to assess whether incorporation of significant BLAST-derived sequences into the training set results in a less specific and sensitive HMM and thus entry into the HMM twilight zone in which false postives and false negatives are likely to have log-odds scores similar to genuine domains. Overall, however, previous studies (15,24,54–56) have shown that utilising both BLAST and HMMs is an effective approach to modelling protein domains. BLAST can be considered an ungapped HMM that identifies segments (motifs) present in a pair of related sequences and thus can help to define an initial training set. Given this training set, HMMs can be employed to model the ordered series of motifs that define the family as a whole and to detect remote homologues in an iterative approach.

Sequence and structural features of the exonuclease domain

Amongst the 148 training set sequences, 53 do not have tyrosine at active site position D in Exo III: three possess histidine at D, 43 histidine preceeding D and seven serine or threonine preceeding or following D. In one of these, Bsu_DP3A in Figure 2, mutation of the histidine preceeding D (histidine 565) to alanine leads to polymerase and exodeoxyribonuclease activities 50% and 0.001%, respectively, of wild type (39). The corresponding values for mutation of the aspartate at E (aspartate 570) to alanine or glycine are 75% and 0.001%. Thus, for the 53 domains that lack tyrosine at D, these mutational studies provide support for the identification of Exo III as shown. Therefore, the aforementioned histidine, serine and threonine residues are most likely to be the functional equivalents of the catalytically important tyrosine in the conventional two-metal-ion mechanism and all 53 domains could behave as exonucleases.

The two well-studied proofreading domains serve as guides for inferring sequence, structural and functional features of the exonuclease domain examined here. Figure 3 shows some of the information in the alignment (Fig. 2) mapped onto the structures of the proofreading domains depicted in Figure 1. Overall, there is considerable sequence divergence because only 7.2% (10/138) of the positions in the alignment are conserved. In addition to the known conserved metal ion-ligand and nucleophilic residues in Exo I, II and III (red, A–E), there are additional conserved residues around the active site (magenta) that may be necessary for the structure, function and/or folding of the domain. Conservation and variation at the primary sequence level is also evident at the tertiary structure level. Examination of the proofreading domains and the alignment suggests that the core of the exonuclease domain consists of three β-strands and five α-helices (labelled β1–β3 and α1–α5). This core is subject to insertions at the periphery (cyan) whose main functions may be to modulate the stereochemistry of the active site and hence recognition of specific substrates and/or be part of regions involved in intra- and/or intermolecular contacts. Two notable examples of such insertions present in many domains occur between α2–α3 and α4–α5. An asparagine, often present as a histidine–asparagine dipeptide, occurs immediately after β3 and may have a functional role because of its proximity to the active site.

The core structure permits rationalisation of experimental data. *Escherichia coli* RNase T (Eco_RNT) functions as a homodimer dependent on cysteine 168. Cysteine 112 and 168 to serine mutations reduce activity both *in vivo* and *in vitro* (57). However, cysteine 168 is believed not to be involved directly in substrate binding but to contribute to a hydrophobic core that influences the structure of the enzyme and thus its activity. Figure 3 indicates that these cysteine residues (yellow) correspond to positions distal from the active site and are thus unlikely to be major factors in substrate binding. Cysteine 168 would lie on the surface of the domain suggesting that the region it is part of might represent the dimer interface in RNase T and the possible site of intra- and/or intermolecular interactions in other exonuclease domains. Residues 650–715 of *H.sapiens* DNA polymerase α (Hsa_DPOA), believed to be a putative DNA binding region (58), correspond to the segment from the end of α2 to the end of α4 that is likely to interact with substrate in proofreading domains (compare Figs 1 and 3). The temperature sensitive alanine 176 to arginine mutation in a phage DNA polymerase (BPPH2_DPO) (59) maps to a conserved leucine (magenta, α5) in proximity to a conserved aspartate (magenta, α3) and leucine (magenta, α4). In Bsu_DP3A, mutation of this aspartate (aspartate 533) to alanine or glycine leads to polymerase and exodeoxyribonuclease activities 40% and 0.001% of wild type (39). The conserved leucine in α5 would be on the same face of the helix that contains active site positions D and E in Exo III.

Phylogenetic analysis

Given the large number of domains and their considerable divergence, the relative arrangement of branches in a phylogenetic tree is likely to exhibit variability so the focus is primarily on sequences within branches. Figure 4 shows the phylogenetic tree based upon the HMM-generated alignment of all the training set. Examination of the tree suggests the existence of five major exonuclease domain subfamilies which are defined as A, DNA polymerase (2–5); B, RNase D (15,16); C, RNase T (see also ref. 35); D, PAN2; and E, CAF1. Most of the newly identified domains (red) belong to subfamilies D and E although new members of A–C are present. A common behaviour of domains in subfamilies A–D is removal of one or more bases at the 3′-end of specific ssRNA or DNA structures within a few nucleotides of a stable base-paired stem. Whether this function is common to all the domains remains to be determined. All members of subfamily B possess tyrosine at active site position D whereas C and D all have histidine preceeding E. The functional significance of the tyrosine/histidine difference is unknown. Possiblities include discrimination between RNA and DNA substrates and functionality other than, or in addition to, exonuclease activity.

Sequences with similar functions tend to group together suggesting related functions for previously uncharacterised

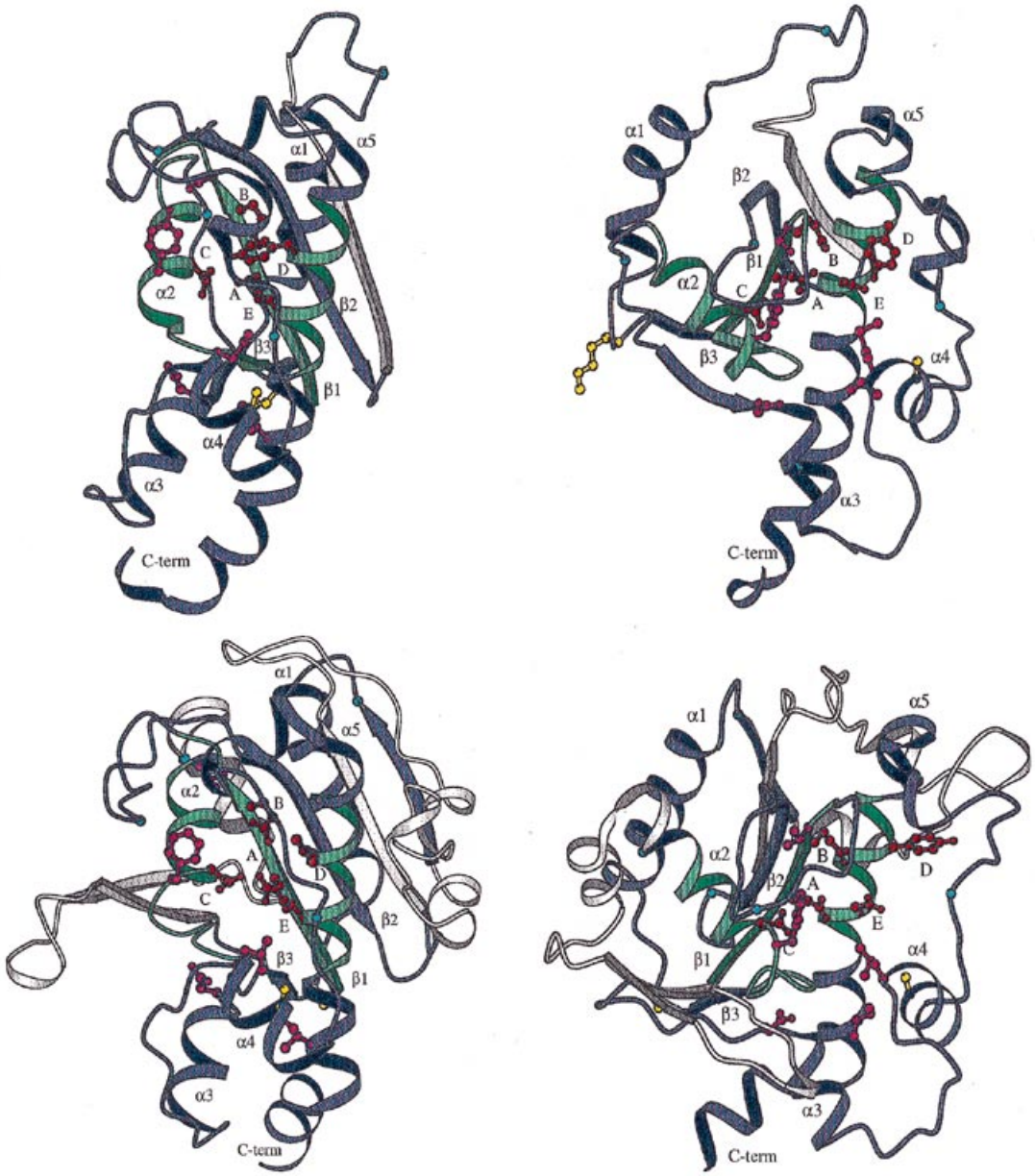

Figure 3. Ribbon diagrams of the two proofreading domains Eco_DPO1 (top) and BPT4_DPO (bottom) shown in Figure 1 (regions coloured green, red and blue). The orthogonal pairs are in approximately the same orientations and the left hand view is the same as that in Figure 1. The Exo I, II and III motifs are green, active site residues are red (A–E), residues preceeding insert states where two or more sequences have insertions longer than 10 residues are cyan, conserved residues are magenta and residues which correspond to Cys 112 and 168 in Eco_RNT are yellow. Helices and strands that occur as insertions in the HMM-generated alignment are in grey. Secondary structure designations are those given in Figure 2. C-term designates the C-terminus.

domains in the same subfamily. For example, new exonuclease domains (red) in subfamily A could have a role in proofreading since they are most like the exonuclease domain in DNA polymerases. *Saccharomyces cerevisiae* PAN2 (Sce_PAN2) is a 3′→5′ exoribonuclease that shortens mRNA poly(A) tails (60,61) and requires Mg^{2+} for activity (62). Thus, the other bacterial and eucaryotic sequences in subfamily D may have a role in RNA and/or DNA degradation. Subfamily E possesses both RNA-binding proteins (Dme_EXU, Dps_EXU1) and those associated with transcription regulation [Mmu_CAF1, Cel_CAF1 (63)]. Whether the archaeal, viral and eucaryotic sequences in subfamily E play a role in RNA and/or DNA processing remains to be seen. The five subfamilies do not appear to reflect whether the substrate is RNA (B: Eco_RND, Sce_RRP6P; C: Eco_RNT; D: Sce_PAN2) or DNA (A, B, C: DNA polymerases; A: BPT4_EXOD; C: Eco_EX1) because B–D each have one domain with an RNA substrate.

Further studies are necessary to resolve the nature of the substrate: RNA, DNA, both DNA and RNA or a non-nucleic acid. This is of particular interest for proteins as Werner syndrome protein (Hsa_WRN) in subfamily B because it may assist in elucidating some of the mechanisms involved in aging. One group of proteins in this subfamily are key enzymes that catalyse

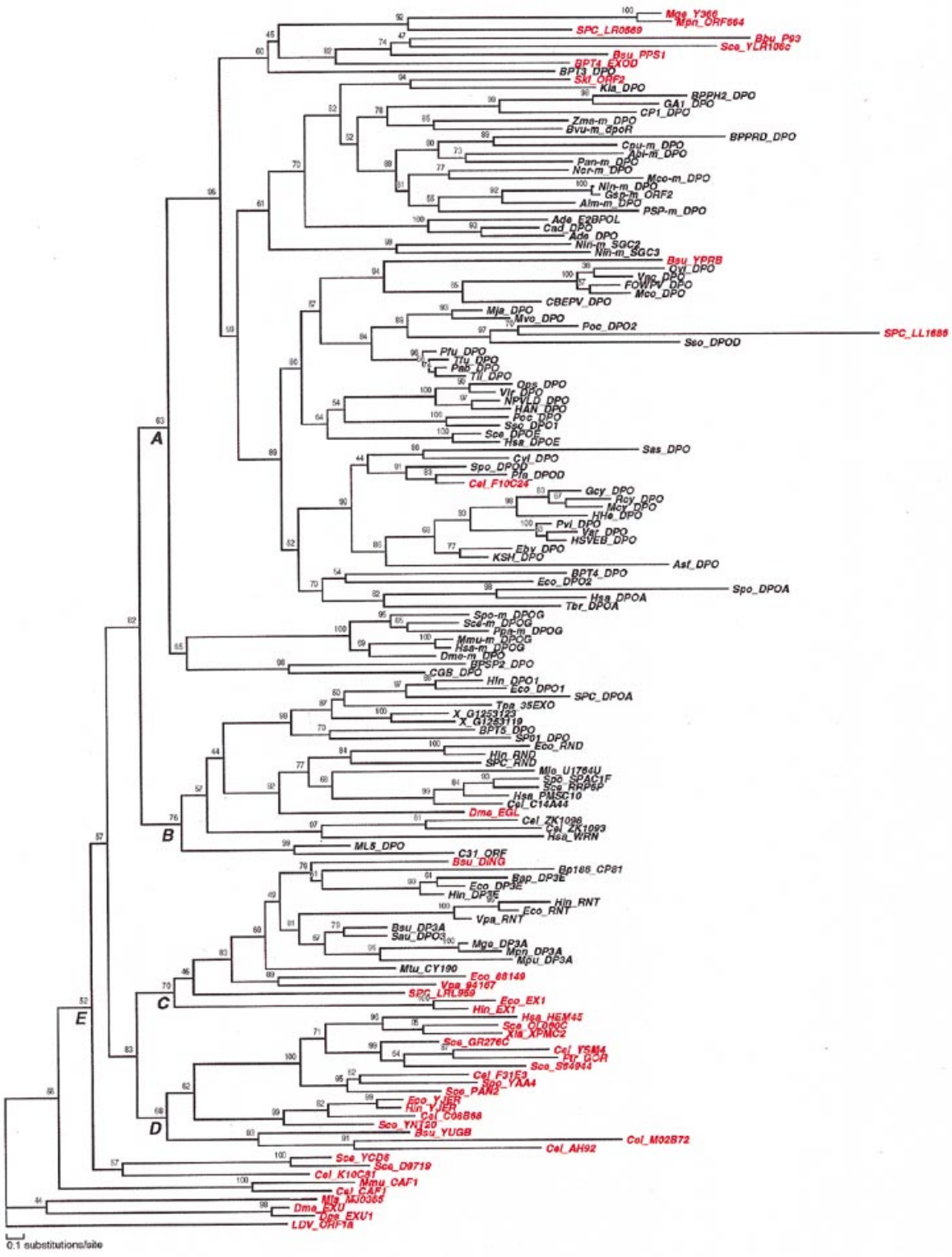

Figure 4. Phylogenetic tree for the exonuclease domain computed using a maximum likelihood and neighbour-joining approach and based upon an alignment of all the training set. Sequence identifiers are given in the supplementary data to this manuscript. New domains identified in this work are in red and subfamilies discussed in the text are labelled. Local bootstrap probabilities are given for each branch and indicate the bootstrap probability of that branch when the other parts of the tree are correct.

the accurate replication of DNA. The contribution of proof-reading to frameshift fidelity during replication of repetitive DNA sequences diminishes as the number of repeats increases indicating that only errors close to the growing DNA point can be repaired by this mechanism (64,65). Thus, one role for Werner syndrome protein may be accurate replication of the repetitive DNA sequences whose instability is associated with several human diseases. An alternative and/or additional role may be RNA processing. Clearly, experimental characterisation of the preferred substrate(s) and precise activities for one or more sequence in each subfamily is required.

Some of the proteins that possess the exonuclease domain may be microtubule–nucleic acid interface proteins. Inhibition of DNA synthesis induces transcription of DNA damage-inducible genes and prevention of mitotic entry through the action of the S phase checkpoint. A known exonuclease, REC1, couples DNA repair and completion of DNA synthesis to a mitotic checkpoint (66). Thus, some of the proteins examined here may act as sensors of DNA replication and coordinate the transcriptional and cell cycle responses to replication blocks. Oncoprotein 18 (Op18)/stathmin, which is highly expressed in leukaemia cells, interacts with tubulin dimers and increases the catastrophe rate of microtubules in mitosis (67). This suggests that *X.laevis* XMPC2 (Xla_XMPC2) which prevents mitotic catastrophe in fission yeast may be important in regulating microtubule dynamics in response to external signals and thus cell cycle progression. Similar roles might be possible for other domains in subfamily D that are involved in tumour formation (Hsa_HEM45), hepatitis C infection (Ptr_GOR) and mRNA poly(A) processing (Sce_PAN2).

DISCUSSION

An HMM has been trained that captures the core elements of an exonuclease domain present in proteins from viruses, bacteria, archaea and eucarya. In general, the results suggest that HMM-based analysis is a valuable tool in defining domains and in identifying remote homologues. Whilst the major features of the exonuclease domain such as the locations of conserved regions Exo I, II and III are unlikely to change, further refinement of the HMM and inclusion of additional sequences should assist in revising and improving the detailed aspects of the model and thus identifying new exonuclease domains. The current HMM represents a good estimate of the features that characterise this domain but how many exhibit exonuclease and/or other activities remains to be determined.

Saccharomyces cerevisiae Sep1 (also called Kem1, Xrn1, Rar5, Dst2) is a multifunctional nuclear protein with an array of proposed roles including nucleic acid binding (e.g., G4 tetraplex DNA), RNA turnover, 5′→3′ exonuclease activity on ss, ds DNA as well as ss RNA, association with cytoplasmic microtubules through β-tubulin, necessity for transition through meiotic prophase and catalysis of DNA strand transfer reactions *in vitro* (see 68,69 and references therein). Although Sep1 is a 5′→3′ exonuclease, many of its suspected roles are present in proteins containing the exonuclease domain modelled here. Thus, exonucleases acting at the 5′- and/or 3′-ends of DNA and RNA molecules play key roles in how cells grow, divide and respond to their environment. Experimental characterisation of the exonuclease domain modelled here may provide insights into these processes.

ACKNOWLEDGEMENTS

We thank our colleagues at UCSC for use of computer time and equipment. This work was supported by the Director, Office of Energy Research, Office of Health and Environmental Research, Division of the US Department of Energy under Contract No. DE-AC03-76SF00098. The data and multiple alignments are available in electronic form upon request.

REFERENCES

1 Mummenbrauer,T., Janus,F., Muller,B., Wiesmuller,L., Deppert,W. and Grosse,F. (1996) *Cell*, **85**, 1089–1099.
2 Bernad,A., Blanco,L., Lazaro,J., Martin,G. and Salas,M. (1989) *Cell*, **59**, 219–228.
3 Delarue,M., Poch,O., Tordo,N., Moras,D. and Argos,P. (1990) *Protein Engng*, **3**, 461–467.
4 Ito,J. and Braithwaite,D. (1991) *Nucleic Acids Res.*, **19**, 4045–4057.
5 Braithwaite,D. and Ito,J. (1993) *Nucleic Acids Res.*, **21**, 787–802.
6 Ollis,D., Brick,P., Hamlin,R., Xuong,N. and Steitz,T. (1985) *Nature*, **313**, 762–766.
7 Freemont,P., Friedman,J., Beese,L., Sanderson,M. and Steitz,T. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 8924–8928.
8 Derbyshire,V., Freemont,P., Sanderson,M., Beese,L., Friedman,J., Joyce,C. and Steitz,T. (1988) *Science*, **240**, 199–201.
9 Beese,L. and Steitz,T. (1991) *EMBO J.*, **10**, 25–33.
10 Beese,L., Derbyshire,V. and Steitz,T. (1993) *Science*, **260**, 352–355.
11 Wang,J., Yu,P., Lin,T., Konigsberg,W. and Steitz,T. (1996) *Biochemistry*, **35**, 8110–8119.
12 Joyce,C. and Steitz,T. (1994) *Annu. Rev. Biochem.*, **63**, 777–822.
13 Reuven,N. and Deutscher,M. (1993) *FASEB J.*, **7**, 143–148.
14 Zhang,J. and Deutscher,M. (1988) *J. Biol. Chem.*, **263**, 17909–17912.
15 Mian,I. (1997) *Nucleic Acids Res.*, **25**, 3187–3195.
16 Briggs,M., Dacanay,K. and Butler,J. (1997) *Mol. Cell. Biol.*, in press.
17 Mushegian,A., Bassett,D.,Jr, Boguski,M., Bork,P. and Koonin,E. (1997) *Proc. Natl. Acad. Sci. USA*, **94**, 5831–5836.
18 Rabiner,L. and Juang,B. (1986) *IEEE ASSP Magazine*, **3**, 4–16.
19 Rabiner,L. (1989) *Proc. IEEE*, **77**, 257–286.
20 Krogh,A., Brown,M., Mian,I., Sjölander,K. and Haussler,D. (1994) *J. Mol. Biol.*, **235**, 1501–1531.
21 Baldi,P., Chauvin,Y., Hunkapiller,T. and McClure,M. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 1059–1063.
22 Eddy,S. (1996) *Curr. Opin. Struct. Biol.*, **6**, 361–365.
23 Fujiwara,Y., Asogawa,M. and Konagaya,A. (1994) *ISMB*, **2**, 121–129.
24 Dalgaard,J., Moser,M., Hughey,R. and Mian,I. (1997) *J. Comp. Biol.*, **4**, 193–214.
25 Bateman,A. and Chothia,C. (1996) *Curr. Biol.*, **6**, 1544–1547.
26 Bateman,A., Eddy,S. and Chothia,C. (1996) *Protein Sci.*, **5**, 1939–1941.
27 Hazes,B. (1996) *Protein Sci.*, **5**, 1490–1501.
28 Shub,D., Goodrich-Blair,H. and Eddy,S. (1994) *Trends Biochem. Sci.*, **19**, 402–404.
29 Grundy,W., Bailey,T., Elkan,C. and Baker,M. (1997) *Biochem. Biophys. Res. Comm.*, **231**, 760–766.
30 Waterman,M. and Perlwitz,M. (1986) *Bull. Math. Biol.*, **46**, 567–577.
31 Barton,G. and Sternberg,M. (1990) *J. Mol. Biol.*, **212**, 389–402.
32 Gribskov,M., McLachlan,A. and Eisenberg,D. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 4355–4358.
33 Taylor,W. (1986) *J. Mol. Biol.*, **188**, 233–258.
34 Bowie,J., Lüthy,R. and Eisenberg,D. (1991) *Science*, **253**, 164–170.
35 Koonin,E. and Deutscher,M. (1993) *Nucleic Acids Res.*, **21**, 2521–2522.
36 Li,Z. and Deutscher,M. (1994) *J. Biol. Chem.*, **269**, 6064–6071.
37 Li,Z. and Deutscher,M. (1995) *Proc. Natl. Acad. Sci. USA*, **92**, 6883–6886.
38 Hughey,R. and Krogh,A. (1996) *Comput. Applic. Biosci.*, **12**, 95–107. The hidden Markov model software can be accessed at URL http://www.cse.ucsc.edu/research/compbio/sam.html
39 Barnes,M., Spacciapoli,P., Li,D. and Brown,N. (1995) *Gene*, **165**, 45–50.
40 Brown,M., Hughey,R., Krogh,A., Mian,I., Sjölander,K. and Haussler,D. (1993) *ISMB*, **1**, 47–55.
41 Sjölander,K., Karplus,K., Brown,M., Hughey,R., Krogh,A., Mian,I. and Haussler,D. (1996) *Comput. Appl. Biosci.*, **12**, 327–345.
42 Altschul,S. (1991) *J. Mol. Biol.*, **219**, 555–565.
43 Barrett,C., Hughey,R. and Karplus,K. (1997) *Comput. Appl. Biosci.*, **13**, 191–199.

44 {NCI} (1997) 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.
45 Adachi,J. (1995) Modelling of molecular evolution and maximum likelihood inference of molecular phylogeny (PhD dissertation). Institute of Statistical Mathematics, Tokyo.
46 Adachi,J. and Hasegawa,M. (1992) MOLPHY: Programs for Molecular Phylogenetics, I. PROTML: Maximum Likelihood Inference of Protein Phylogeney Computer Science Monographs 27. Institute of Statistical Mathematics, Tokyo. MOLPHY is available from ftp://sunmh.ism.ac.jp/pub/molphy
47 Barton,G. (1993) *Protein Engng*, **6**, 37–40.
48 Maciukenas,M. (1992) Treetool: an interactive tool for displaying, editing and printing phylogenetic trees. Currently, Treetool is modified and maintained by Mike McCaughey, Ribosomal Database Project, University of Illinois. It is available from ftp://rdp.life.uiuc.edu/rdp/programs/TreeTool
49 Kraulis,P. (1991) *J. Appl. Crystallog.*, **24**, 946–950.
50 Kuhn,F. and Knopf,C. (1996) *J. Biol. Chem.*, **271**, 29245–29254.
51 Micklem,D. (1995) *Dev. Biol.*, **172**, 377–395.
52 Mach,J. and Lehmann,R. (1997) *Genes Dev.*, **11**, 423–435.
53 Altschul,S., Gish,W., Miller,W., Myers,E. and Lipman,D. (1990) *J. Mol. Biol.*, **215**, 403–410.
54 Herbert,A., Alfken,J., Kim,Y.-G., Mian,I., Nishijura,K. and Rich,A. (1997) *Proc. Natl. Acad. Sci. USA*, **94**, 8421–8426.
55 Mian,I., Moser,M., Holley,W. and Chatterjee,A. (1998) *J. Comp. Biol.*, In press.
56 Dalgaard,J., Klar,A., Moser,M., Holley,W., Chatterjee,A. and Mian,I. (1997) *Nucleic Acids Res.*, **25**, 4626–4638.
57 Li,Z., Zhan,L. and Deutscher,M. (1996) *J. Biol. Chem.*, **271**, 1127–1132.
58 Wong,S., Wahl,A., Yuan,P., Arai,N., Pearson,B., Arai,K., Korn,D., Hunkapiller,M. and Wang,T. (1988) *EMBO J.*, **7**, 37–47.
59 Blasco,M., Blanco,L., Pares,E., Salas,M. and Bernad,A. (1990) *Nucleic Acids Res.*, **18**, 4763–4770.
60 Brown,C. Tarun,S.,Jr, Boeck,R. and Sachs,A. (1996) *Mol. Cell. Biol.*, **16**, 5744–5753.
61 Boeck,R., Tarun,S.,Jr, Rieger,M., Deardorff,J., Muller-Auer,S. and Sachs,A. (1996) *J. Biol. Chem.*, **271**, 432–438.
62 Lowell,J., Rudner,D. and Sachs,A. (1992) *Genes Dev.*, **6**, 2088–2099.
63 Draper,M., Salvadore,C. and Denis,C. (1995) *Mol. Cell. Biol.*, **15**, 3487–3495.
64 Strauss,B., Sagher,D. and Acharya,S. (1997) *Nucleic Acids Res.*, **25**, 806–813.
65 Kroutil,L., Register,K., Bebenek,K. and Kunkel,T. (1996) *Biochemistry*, **35**, 1046–1053.
66 Onel,K., Koff,A., Bennett,R., Unrau,P. and Holloman,W. (1996) *Genetics*, **143**, 165–174.
67 Belmont,L. and Mitchison,T. (1996) *Cell*, **84**, 623–631.
68 Szankasi,P. and Smith,G. (1996) *Curr. Genet.*, **30**, 284–293.
69 Bashkirov,V., Solinger,J. and Heyer,W. (1995) *Chromosoma*, **104**, 215–222.
70 Dykhuizen,D., Polin,D., Dunn,J., Wilske,B., Preac-Mursic,V., Dattwyler,R. and Luft,B. (1993) *Proc. Natl. Acad. Sci. USA*, **90**, 10163–10167.
71 Tognoni,A., Franchi,E., Magistrelli,C., Colombo,E., Cosmina,P. and Grandi,G. (1995) *Microbiology*, **141**, 645–648.
72 Gruber,H., Kern,G., Gauss,P. and Gold,L. (1988) *J. Bacteriol.*, **170**, 5830–5836.
73 Gauss,P., Gayle,M., Winter,R. and Gold,L. (1987) *Mol. Gen. Genet.*, **206**, 24–34.
74 Koonin,E. (1993) *Nucleic Acids Res.*, **21**, 1497.
75 Phillips,G. and Kushner,S. (1987) *J. Biol. Chem.*, **262**, 455–459.
76 Miesel,L. and Roth,J. (1996) *J. Bacteriol.*, **178**, 3146–3155.
77 Su,J. and Maller,J. (1995) *Mol. Gen. Genet.*, **246**, 387–396.
78 MacDonald,P., Luk,S. and Kilpatrick,M. (1991) *Genes Dev.*, **5**, 2455–2466. Published erratum appears in (1992) *Genes Dev*, **6**, 690.
79 Luk,S., Kilpatrick,M., Kerr,K. and MacDonald,P. (1994) *Genetics*, **137**, 521–530.
80 Wang,S. and Hazelrigg,T. (1994) *Nature*, **369**, 400–403.