JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 5, Number 1, 1998
Mary Ann Liebert, Inc.
Pp. 57–72

Statistical Modelling and Phylogenetic Analysis of a Deaminase Domain

I. SAIRA MIAN, MICHAEL J. MOSER, WILLIAM R. HOLLEY, and ALOKE CHATTERJEE

ABSTRACT

Deamination reactions are catalyzed by a variety of enzymes including those involved in nucleoside/nucleotide metabolism and cytosine to uracil (C→U) and adenosine to inosine (A→I) mRNA editing. The active site of the deaminase (DM) domain in these enzymes contains a conserved histidine (or rarely cysteine), two cysteines and a glutamate proposed to act as a proton shuttle during deamination. Here, a statistical model, a hidden Markov model (HMM), of the DM domain has been created which identifies currently known DM domains and suggests new DM domains in viral, bacterial and eucaryotic proteins. However, no DM domains were identified in the currently predicted proteins from the archaeon *Methanococcus jannaschii* and possible causes for, and a potential means to ameliorate this situation are discussed. In some of the newly identified DM domains, the glutamate is changed to a residue that could not function as a proton shuttle and in one instance (*Mus musculus* spermatid protein TENR) the cysteines are also changed to lysine and serine. These may be non-competent DM domains able to bind but not act upon their substrate. Phylogenetic analysis using an HMM-generated alignment of DM domains reveals three branches with clear substructure in each branch. The results suggest DM domains that are candidates for yeast, platyhelminth, plant and mammalian C→U and A→I mRNA editing enzymes. Some bacterial and eucaryotic DM domains form distinct branches in the phylogenetic tree suggesting the existence of common, novel substrates.

Key words: hidden Markov model, RNA editing, deamination, TENR protein, DM domain.

INTRODUCTION

DEAMINATION REACTIONS OCCUR IN A VARIETY OF PROCESSES including nucleoside/nucleotide metabolism and base substitution RNA editing, one of a set of co- or posttranscriptional events in which nucleotide insertion, deletion, or base substitution results in the production of an RNA whose sequence differs from that of its template (reviewed in Benne, 1996; Scott, 1995; Smith and Sowden, 1996; Herbert, 1996). A number of enzymes known to catalyze such deamination reactions share an active site containing a conserved histidine (His, rarely cysteine), two cysteine (Cys) and a glutamic acid (Glu) residue believed to act as a proton donor or shuttle during the hydrolytic deamination reaction. The part of the enzyme encompassing this active site will be referred to as the DM (deaminase) domain.

The best characterized DM domain containing enzymes bind monomeric and polymeric nucleoside/nucleotide substrates and are cytidine deaminase (CDD) which converts cytidine to uridine (Yang *et al.*, 1992); deoxycytidylate deaminase (DCTD) which hydrolyses dCMP into dUMP (Moore *et al.*, 1993); the catalytic

Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California.

subunit of the mammalian apolipoprotein B (apoB) mRNA editing enzyme (APOBEC) which is responsible for the cytosine to uracil (C→U) conversion that alters a specific glutamine (CAA) codon into a stop codon (UAA) in the apoB mRNA (Navaratnam *et al.*, 1995); double-stranded (ds) adenosine deaminase (DRADA) which is required for the conversion of specific adenosines to inosines (A→I) in brain-expressed pre-mRNAs for glutamate receptor (GluR) subunits (Lai *et al.*, 1995; Herb *et al.*, 1996; Rueter *et al.*, 1995) and in the antigenome of hepatitis delta virus (HDV) (Polson *et al.*, 1996). GluR editing requires a second ds adenosine deaminase, RED1, which has a distinct but overlapping substrate specificity with DRADA (Maas *et al.*, 1996; Melcher *et al.*, 1996). Inosines are read as guanosines (G) by the translational machinery (Dabiri *et al.*, 1996) resulting in change of the encoded amino acid from glutamine (CAG) to arginine (CGG) and arginine (AGA) to glycine (GGA) in GluR pre-mRNAs, and stop (UAG) codon to tryptophan (UGG) in the HDV antigenome. In addition to the site-selective editing of mammalian mRNAs of neural origin, DRADA has been implicated in the generation of biased hypermutations (clusters of certain transitions) in some RNA viruses (Cattaneo, 1994; Polson and Bass, 1994).

Three of the aforementioned enzymes, CDD, DCTD and APOBEC, bind zinc ions via the active site His and Cys residues (Betts *et al.*, 1994; Moore *et al.*, 1993; Navaratnam *et al.*, 1995). A motif comprized of these zinc ion-binding residues and the active site Glu has been observed in the sequences of a number of other proteins (Reizer *et al.*, 1994; Bhattacharya *et al.*, 1994): riboflavin biosynthesis protein ribG, which converts 2,5-diamino-6-(ribosylamino)-4(3H)-pyrimidinone 5′-phosphate (ribG) into 5-amino-6-(ribosylamino)-2,4(1H,3H)-pyrimidinedione 5′-phosphate (Sorokin *et al.*, 1993); *Bacillus cereus* blasticidin S deaminase (BSD), which catalyzes deamination of the cytosine moiety of the antibiotic blasticidin S and its derivatives but not cytosine nucleosides (Kobayashi *et al.*, 1991; Nawa *et al.*, 1995); *Bacillus subtilis* open reading frame (ORF) CME2, which is part of the comE operon required for the binding and uptake of transforming DNA (Hahn *et al.*, 1993); *Bacillus subtilis* ORF YAAJ (Ogasawara *et al.*, 1994; Struck *et al.*, 1990); *Escherichia coli* ORF YFHC (Poulsen *et al.*, 1992); *Vibrio fischeri* ORF YLXG (Lee *et al.*, 1993); and *Saccharomyces cerevisiae* ORF YJD5 (Pohl and Aljinovic, 1996). Reizer and colleagues (Reizer *et al.*, 1994) generalized the zinc ion-binding motif and derived a PROSITE (Bairoch *et al.*, 1996) regular expression, [CH][AV]Ex(2)[LIVMFGA][LIVM]x(17,33)PCx(2,8)Cx(3)[LIVM] (accession number PS00903, CYT_DCMP_DEAMINASES). Although DRADA is likely to bind zinc ions, this PROSITE descriptor fails to capture the DM domain in DRADA and thus presumably in other proteins possessing a DM domain. For example, most, if not all mitochondrial mRNAs from vascular plants and bryophytes undergo mRNA editing involving the conversion of some Cs to Us thereby correcting multiple genomically encoded missense codons and permitting functional protein synthesis (reviewed in Gray (1996)). Recent evidence indicates that this mRNA editing occurs via a deamination mechanism (Blanc *et al.*, 1995; Yu and Schuster, 1995) but the proteins responsible have yet to be characterized.

Identification of new DM domains can suggest candidates responsible for catalyzing deamination of known and novel substrates as well as yield insights into normal physiology and disease processes. For example, mRNA editing of GluR subunits leads to cation specific channels that recover much faster from desensitization and may integrate signals better (Lomeli *et al.*, 1994). The neurofibromatosis type I (NF1) tumor suppressor mRNA undergoes a C→U change causing an arginine (CGA) codon to be changed to an inframe stop (UGA) codon (Skuse *et al.*, 1996). NF1 editing occurs in normal tissues but is higher in tumors and appears not to be mediated by APOBEC (Skuse *et al.*, 1996).

Here, a hidden Markov model (HMM) of the DM domain has been trained and an HMM-generated alignment of eighty-three DM domains (thirty-seven of which were identified in this work) employed for subsequent phylogenetic analysis. HMMs are a statistical modelling method (Rabiner and Juang, 1986; Rabiner, 1989; Krogh *et al.*, 1994; Baldi *et al.*, 1994; Eddy, 1996; Fujiwara *et al.*, 1994) that have been used recently to characterize the common features of a family of related sequences and to recognize related, but divergent family members present in databases (Mian, 1997; Dalgaard *et al.*, 1997; Bateman and Chothia, 1996; Bateman *et al.*, 1996; Hazes, 1996; Shub *et al.*, 1994; Baldi *et al.*, 1994; Grundy *et al.*, 1997). HMMs 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 (Waterman and Perlwitz, 1986; Barton and Sternberg, 1990; Gribskov *et al.*, 1987; Taylor, 1986; Bowie *et al.*, 1991). Although both HMMs and profiles can overcome the shortcomings of PROSITE regular expressions, the HMM formalism provides two advantages over profiles. First, HMMs can be trained from unaligned as well as aligned sequences (profiles require an existing alignment) and second, HMMs treat scores and penalties and evaluate alignments in a justifiable statistical manner. The DM domain HMM is sensitive in that the numbers of false positives (sequences incorrectly identified by the HMM as possessing a DM domain) and false negatives (sequences

not identified by the HMM as possessing a DM domain) are likely to be small. However, although new and known viral, bacterial and eucaryotic DM domains were identified by the HMM, the absence of any archaeal DM domains suggests methodological improvements that might allow detection of such remote homologs.

MATERIALS AND METHODS

Definition of the DM domain

In the three-dimensional structure of *Escherichia coli* CDD, the active site residues (the His and Cys zinc ligands and Glu proton shuttle) are part of an α-β-α structure (Betts *et al.*, 1994). The region between the first α-helix and the β-strand will be termed L1 and the region between the β-strand and the second α-helix L2. Inspection of the CDD structure indicated that a contiguous region of the protein chain that included this α-β-α structure constituted the substrate binding region. This region is shown in Figure 1 and defines the boundaries of the DM domain for which an HMM was created. In *Escherichia coli* CDD, this region is approximately 60 residues in length. However, when the active site residues of an A→I editing enzyme such as DRADA are aligned to those of CDD, the corresponding region is well over twice this length. It was reasoned that if the active site of DRADA has the same geometry (α-β-α structure) as CDD, then the "extra" residues in DRADA could be accommodated by treating them as insertions between the secondary structure elements present in CDD, i.e., in regions L1 and L2.

Starting set of DM domains

The initial set of DM domains consisted of previously known examples and comprized those obtained from the literature (3 A→I editing enzymes) and those possessing the PROSITE CYT_DCMP_DEAMINASES signature (22 sequences). Each of the sequences was used as the query for a database search. The BLAST suite of programs (Altschul *et al.*, 1990) were run with default parameters and a merged, nonredundant collection of sequences derived from PIR, SwissProt and translated Genbank. Database sequences were considered to exhibit a statistically significant similarity to the query if smallest sum probability $P(N) \leq 0.01$, $P(N)$ being the lowest probability ascribed to any set of high scoring segment pairs for each database sequence. Partial

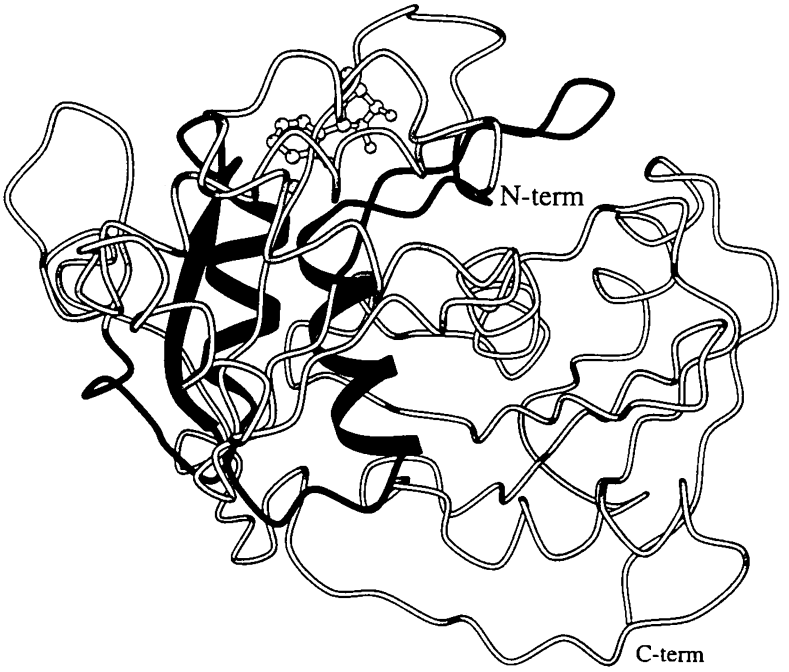


FIG. 1. Three-dimensional structure of *Escherichia coli* cytidine deaminase (CDD) complexed with 3,4-dihydrozebularine (PDB: 1CTT). The substrate is shown in ball-and-stick form, the zinc ion is a sphere and the N- and C-terminii are labelled. The DM domain of CDD modelled by the HMM in this work is shown in ribbon form and shaded more darkly.

sequences (fragments and expressed sequence tags (ESTs)) were retained but not employed for training an HMM and subsequent phylogenetic analysis. Complete DM domains from these two sources were used as the starting training set for a DM domain HMM trained using the SAM (Sequence Alignment and Modeling Software System) suite (Krogh *et al.*, 1994; Hughey and Krogh, 1996) running on a MASPAR MP-2204 with a DEC Alpha 3000/3000X front end at the University of California Santa Cruz (UCSC).

Hidden Markov model

Since a more comprehensive description of the HMMs employed here can be obtained elsewhere (Krogh *et al.*, 1994; Hughey and Krogh, 1996), only a summary is provided. The HMM consists of a series of nodes corresponding to columns in a multiple sequence alignment. Its architecture captures most of the features of a family of related sequences as follows. The match, delete and insert states in a node and transitions between these states model (i) a sequence of positions each with its own distribution over the 20 characters in an amino acid alphabet (the degree of residue conservation at each position); (ii) the possibility of skipping a position (equivalent to a deletion); (iii) the possibility of inserting amino acids between consecutive positions (a region in a family member that is not part of core structure of the family); and (iv) allowing for the possibility that continuing an insertion or deletion is more likely than starting one (position-dependent scores). The parameters of an HMM are the position-dependent character distributions, position-dependent transition frequencies between states and the number of nodes. Training an HMM involves estimating these parameters by the following procedure. An initial stochastic model representing the family is created by describing transitions into a match, delete or insert state and the occurrence of a given residue in a particular match or delete state. Using this initial model and the training sequences (some or all of the family members), possible paths for each sequence through the model are evaluated to obtain new estimates of the parameters that will increase the likelihood of the model. This procedure is repeated until the model converges. Estimating the parameters using only the training set leads to the problem of overfitting in which a model fits the training sequences well but gives a poor fit to related data not included in the training set. Thus, to improve the ability of the HMM to generalize, to fit sequences not employed for training, SAM employs Dirichlet mixture priors (Brown *et al.*, 1993; Sjölander *et al.*, 1996). Dirichlet mixture priors are an effective means to estimate the distribution of characters in a specific context given a small sample of characters from that distribution (Karplus, 1995; Tatusov *et al.*, 1994). These priors, estimated from a Blocks database of multiple sequence alignments (Henikoff *et al.*, 1997; Henikoff and Henikoff, 1991), are designed to be combined with the observed position-dependent character distributions in the training set to form estimates of expected character probabilities.

An HMM is a model that defines a probability distribution over possible sequences. 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 that are from the same family. An HMM trained to model a family assumed to have a common underlying structure assigns high likelihood to family members and low likelihood to non-members. A multiple sequence alignment for a set of sequences is generated by computing, for each sequence in turn, the most likely path through the model given possible paths generated by the training sequences.

A rough multiple alignment of the starting training set of DM domains was created manually in which the active site residues were aligned. This alignment was used to calculate the parameters for an initial HMM. Two features of the DM domain necessitated the use of special types of nodes termed FIMs (free-insertion modules). Because DM domains occur as domains within longer proteins, external FIMs were used at the beginning and end of the HMM to allow an arbitary number of insertions at either end. Since regions L1 and L2 have length distributions much greater than can be modelled adequately by insert states, two internal FIMs were used to model the variable length L1 and L2 regions by permitting insertions of any length between specific nodes in the HMM. Figure 2 shows a detailed view of the DM domain for which an HMM was created and highlights the locations of the active site residues and the L1 and L2 FIMs. To avoid changes in the nodes representing the active site residues and thus deviations from the initial manually-generated alignment, the parameters for these particular nodes were fixed. Multiple models were trained to reduce the problem of local minima and the best were used for further studies.

The specificity and sensitivity of the DM domain HMM was examined by using it to discriminate between sequences that possess a DM domain from those that do. This was achieved by evaluating how much better sequences in a database fit the model than some underlying background distribution or null model (NULL) and assessing the significance of the resultant score. Such database searching using the HMM involves computing

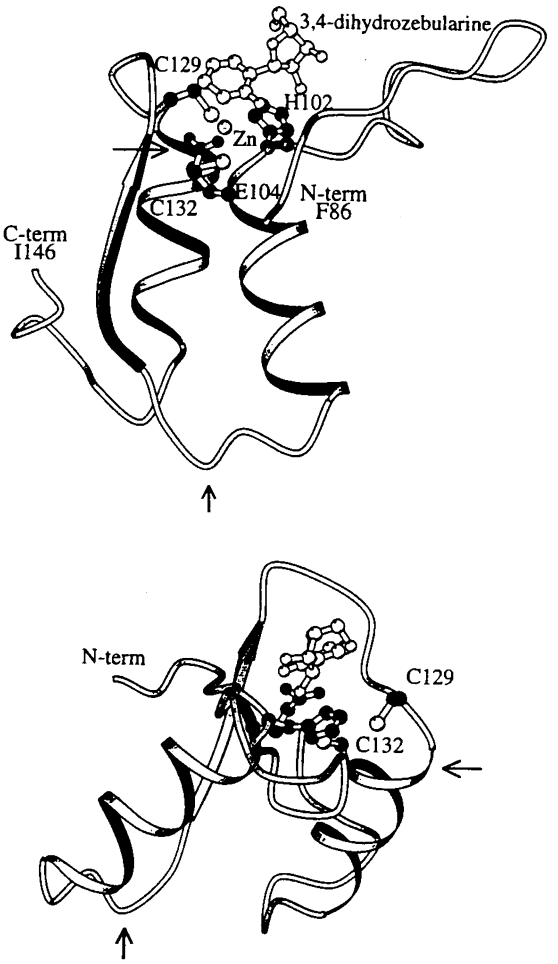


FIG. 2. Orthogonal views of the CDD DM domain shown in Figure 1 for which an HMM was created. The active site residues and bound substrate (white) are depicted in ball-and-stick form and the zinc ion is a sphere. Vertical and horizontal arrows mark the positions where the two internal FIMs employed to model regions L1 and L2 respectively occur in the structure of CDD.

log-odds (NLL-NULL) (Altschul, 1991; Barrett *et al.*, 1997) scores for all sequences in a non-redundant protein database obtained from the NCI (NCI, 1996) and updated weekly at UCSC. Taking into account the number of sequences in this database (approximately 211,000 different proteins in mid-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 false positives. A database search was performed with the HMM and based upon examination of the log-odds scores and an HMM-generated alignment, new DM domains were identified. These new DM domains were added to the training set and the HMM was retrained. At each round of this iterative procedure, the newly identified sequences improve the ability of the HMM to generalize. The cycle of "search, align and retrain" was repeated until no new sequences were identified in databases up to November, 1996. This final HMM was utilized for all subsequent work.

Phylogenetic analysis

Having modelled existing and new DM domains by means of the HMM, the relationships between individual sequences were examined by phylogenetic analysis. An HMM-generated multiple sequence alignment of the training set was utilized as the starting point. The alignment contained only match and delete states. Insertions, including the L1 and L2 FIMs, were excluded. Insert states are not modelled by an HMM and because the regions in a sequence they represent are the most divergent parts of the molecules, they are likely to be sources of systematic error in phylogenetic analysis. The MOLPHY (MOLecular PHYlogenetics) suite uses

a statistical procedure for inferring phylogenetic relationships (Adachi, 1995; Adachi and Hasegawa, 1992). PROTML, the main program in MOLPHY, infers evolutionary trees from amino acid sequences by means of a maximum likelihood method. The star decomposition algorithm of PROTML 2.2 was used to determine automatically the best tree for the sequences in the aforementioned HMM-generated multiple alignment.

Figures showing multiple sequence alignments, phylogenetic trees and ribbon diagrams of molecules were produced using ALSCRIPT (Barton, 1993), Treetool (Maciukenas, 1992) and MOLSCRIPT (Kraulis, 1991) respectively.

RESULTS

Hidden Markov model

Table 1 lists known and new DM domains identified in this work and includes both complete sequences used for HMM training and partial sequences that were not used. Figure 3 shows an HMM-generated alignment of

TABLE 1. ABBREVIATIONS EMPLOYED FOR DM DOMAINS

Abbreviation	Species	Protein description and Databank code in []
Apl_RIBG	*Actinobacillus pleuropneumoniae*	riboflavin biosynthesis protein ribG [APU27202]
Ate_BSD	*Aspergillus terreus*	blasticidin S deaminase (BSD) [S41571]
Ath_EST1	*Arabidopsis thaliana*	EST [H36401]
Ath_EST2	*Arabidopsis thaliana*	EST [H76529]
Ath_HPP	*Arabidopsis thaliana*	vacuolar H+-pyrophosphatase (HPPase) [AVP3_ARATH]
Bam_RIBG	*Bacillus amyloliquefaciens*	riboflavin biosynthesis protein ribG [BARIBGENS]
Bce_BSD‡	*Bacillus cereus*	blasticidin S deaminase (BSD) [BSR_BACCE]
Bpa_CDD	*Brugia pahangi*	cytidine deaminase (CDD) [BPRNAS5]
Bna_EST	*Brassica campestris*	flower bud EST [BNAF1868]
Bja_NFP	*Bradyrhizobium japonicum*	nitrogen fixation protein [RHBNFXP]
Bsu_CDD‡	*Bacillus subtilis*	cytidine deaminase (CDD) [CDD_BACSU]
Bsu_CME2‡	*Bacillus subtilis*	comE operon protein 2 [CME2_BACSU]
Bsu_RIBG‡	*Bacillus subtilis*	riboflavin biosynthesis protein ribG [RIBG_BACSU]
Bsu_YAAJ‡	*Bacillus subtilis*	ORF YAAJ [YAAJ_BACSU]
BT2_DCTD‡	Bacteriophage T2	deoxycytidylate deaminase (DCTD) [DCTD_BPT2]
BT4_DCTD‡	Bacteriophage T4	deoxycytidylate deaminase (DCTD) [DCTD_BPT4]
Bvu_HPP1	*Beta vulgaris*	vacuolar H+-pyrophosphatase (HPPase) [BEUPYROA]
Bvu_HPP2	*Beta vulgaris*	vacuolar H+-pyrophosphatase (HPPase) [BEUPYRO]
Ccr_ORF	*Caulobacter crescentus*	ORF obtained by translation of nucleic acid sequence [CCU27301]
Cel_CDD1	*Caenorhabditis elegans*	ORF F49E8.4 similar to cytidine deaminase (CDD) [CELF49E8]
Cel_CDD2	*Caenorhabditis elegans*	ORF C47D2.2 similar to cytidine deaminase (CDD) [CELC47D2]
Cel_D2005	*Caenorhabditis elegans*	ORF D2005.1 similar to ds RNA adenosine deaminase (DRADA) [CED2005]
Cel_DCTD‡	*Caenorhabditis elegans*	probable deoxycytidylate deaminase (DCTD) [DCTD_CAEEL]
Cel_EST	*Caenorhabditis elegans*	embryo EST [C08551]
Cel_T20H4	*Caenorhabditis elegans*	ORF T20H4.4 similar to ds RNA adenosine deaminase (DRADA) [CELT20H4]
Dme_ORF	*Drosophila melanogaster*	ORF obtained by translation of nucleic acid sequence [DRO28DC4Z]
Eco_BSD	*Escherichia coli*	blasticidin S deaminase (BSD) [S81409]
Eco_CDD‡	*Escherichia coli*	cytidine deaminase (CDD) [CDD_ECOLI]
Eco_RIBG‡	*Escherichia coli*	riboflavin biosynthesis protein ribG [RIBG_ECOLI]
Eco_YFHC‡	*Escherichia coli*	ORF YFHC [YFHC_ECOLI]

(continued)

TABLE 1. *(Continued)*

Abbreviation	Species	Protein description and Databank code in []
Hin_CDD‡	*Haemophilus influenzae*	cytidine deaminase (CDD) [CDD_HAEIN]
Hin_RIBG‡	*Haemophilus influenzae*	riboflavin biosynthesis protein ribG [RIBG_HAEIN]
Hin_YFHC‡	*Haemophilus influenzae*	ORF HI0906 [YFHC_HAEIN]
Hsa_APOBEC‡	*Homo sapiens*	apoB mRNA editing enzyme APOBEC [ABME_HUMAN]
Hsa_CDD‡	*Homo sapiens*	cytidine deaminase (CDD) [CDD_HUMAN]
Hsa_DCTD‡	*Homo sapiens*	deoxycytidylate deaminase (DCTD) [DCTD_HUMAN]
Hsa_DRADA	*Homo sapiens*	ds RNA adenosine deaminase (DRADA) [HSU10439]
Hsa_EST1	*Homo sapiens*	foetal heart EST [R58006]
Hsa_EST2	*Homo sapiens*	EST similar to apoB mRNA editing enzyme APOBEC [R10201]
Hsa_EST3	*Homo sapiens*	sequence constructed from an EST from foetal heart [W78087, 5′ similar to an Alu repetitive element] and one from senescent fibroblasts [W63795]
Hsa_EST4	*Homo sapiens*	EST [T97494]
Hsa_EST5	*Homo sapiens*	pregnant uterus EST [AA040828]
Hsa_phorI	*Homo sapiens*	phorbolin I from psoriatic keratinocytes [HSU03891]
Hvu_HPP	*Hordeum vulgare*	vacuolar H+-pyrophosphatase (HPPase) [AVP3_HORVU]
Lpo_PPLZ20	*Lupinus polyphyllus*	pPLZ20 protein [LPPLZ20]
Mmu_EST1	*Mus musculus*	placenta EST [AA013683]
Mmu_EST2	*Mus musculus*	foetus EST similar to apoB mRNA editing enzyme APOBEC [W30319]
Mmu_EST3	*Mus musculus*	embryo EST [W71850]
Mmu_EST4	*Mus musculus*	embryo EST similar to cytidine deaminase (CDD) [W98165]
Mmu_EST5	*Mus musculus*	embryo EST similar to deoxycytidylate deaminase (DCTD) [W44079]
Mmu_TENR	*Mus musculus*	testis nuclear RNA binding protein (TENR) [MMTENR]
Mca_DCTD	*Mycoplasma capricolum*	ORF similar to deoxycytidylate deaminase (DCTD) [MC037]
Mcv_E2L	Molluscum contagiosum virus subtype 1	homologue of vaccinia and variola virus E2L protein [MCU60315]
Mge_CDD	*Mycoplasma genitalium*	cytidine deaminase (CDD) [CDD_MYCGE]
Mle_ORF	*Mycobacterium leprae*	ORF obtained by translation of nucleic acid sequence [MSGL611CS]
Mpi_CDD	*Mycoplasma pirum*	cytidine deaminase (CDD) [CDD_MYCPI]
Mtu_RIBG‡	*Mycobacterium tuberculosis*	ORF MTCY21B4.26 similar to riboflavin biosynthesis protein ribG [MTCY21B4]
Nta_HPP1	*Nicotiana tabacum*	vacuolar H+-pyrophosphatase (HPPase) [S54173]

(continued)

the sequences in Table 1. Sequences will be referred to by their number in the alignment shown in Figure 3 and the abbreviation given in Table 1, for example, 1:Sce_L8543. Amongst approximately 211,000 sequences in the most recent non-redundant protein database that was searched using the final DM domain HMM, all sequences with log-odds scores higher than 35.3 were part of the training set. Only three training set sequences had scores below 35.3: 19:Mmu_TENR (33.0), 20:Cel_D2005 (29.8) and 27:Hsa_EST3 (27.0). The only other sequences in the range 27.0–35.3 were *Sorangium cellulosum* soraphen A polyketide synthase (33.6; databank code SCU24241) and hepatitis C virus E1 and E2/NS1 envelope glycoprotein (28.0; HPVHCVN, HCU01214 and other, almost identical sequences). Alignment of these sequences to the model (data not shown) indicated the L1 and L2 FIMs were 1759/1244 residues and 64/103 residues respectively. These FIMs are considerably longer than those in the training set which range in length from 1–35 (L1) and 0–72 (L2) residues (the A→I mRNA editing enzymes have the longest L1 and L2 regions). Furthermore, for sequences in the range 27.0–35.3, there were differences in active site residues: in polyketide synthase, a Cys is changed to leucine and in the hepatitis

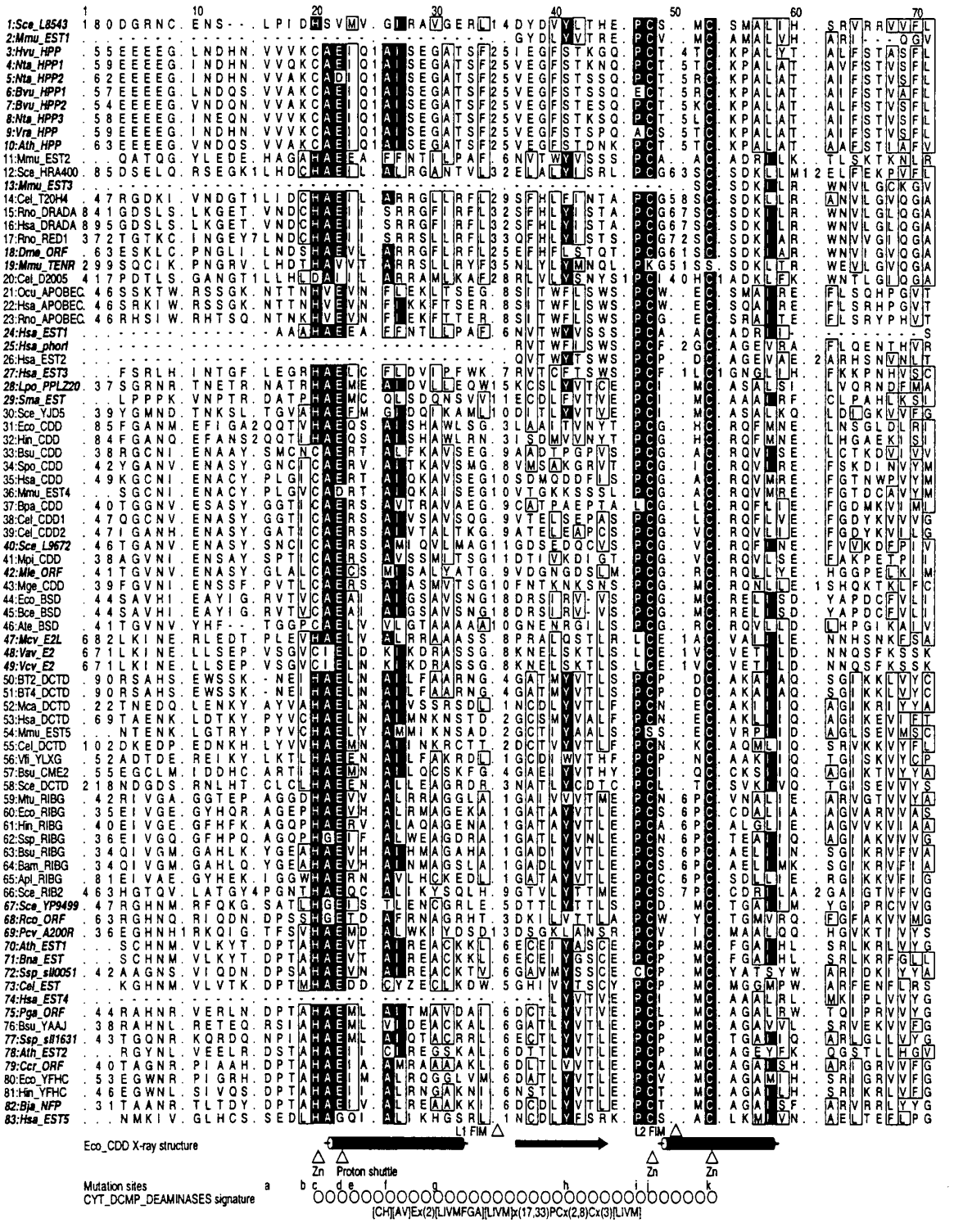
TABLE 1. (*Continued*)

Abbreviation	Species	Protein description and Databank code in []
Nta_HPP2	*Nicotiana tabacum*	vacuolar H+-pyrophosphatase (HPPase) [S42893]
Nta_HPP3	*Nicotiana tabacum*	vacuolar H+-pyrophosphatase (HPPase) [S54172]
Ocu_APOBEC	*Oryctolagus cuniculus*	apoB mRNA editing enzyme APOBEC [ABME_RABIT]
Pcv_A200R	Paramecium bursaria Chlorella virus 1	ORF A200R [PBU42580]
Pga_ORF	*Porphyromonas gingivalis*	ORF ORF150 [PGPGAAGEN]
Rco_ORF	*Rhodococcus corallinus*	ORF ORF1 [RERTRZA]
Rno_APOBEC‡	*Rattus norvegicus*	apoB mRNA editing enzyme APOBEC [ABME_RAT]
Rno_DRADA	*Rattus norvegicus*	adult brain ds RNA adenosine deaminase (DRADA) [RNU18942]
Rno_RED1	*Rattus norvegicus*	ds RNA-specific editase (RED1) [RNU43534]
Sce_DCTD‡	*Saccharomyces cerevisiae*	deoxycytidylate deaminase (DCTD) [DCTD_YEAST]
Sce_HRA400	*Saccharomyces cerevisiae*	HRA400 protein [S53937]
Sce_L9672	*Saccharomyces cerevisiae*	ORF L9672.13 [S59391]
Sce_L8543	*Saccharomyces cerevisiae*	ORF L8543.6 [S53395]
Sce_RIB2	*Saccharomyces cerevisiae*	RIB2 protein [S50972]
Sce_YJD5‡	*Saccharomyces cerevisiae*	ORF YJL035C [YJD5_YEAST]
Sce_YP9499	*Saccharomyces cerevisiae*	ORF YP9499.17 [S54083]
Spo_CDD	*Schizosaccharomyces pombe*	putative cytidine deaminase (CDD) homologue [SPARZPCD]
Sma_EST	*Schistosoma mansoni*	EST [N21734]
Ssp_RIBG	*Synechocystis* sp. PCC6803	ORF slr0066 homologous to riboflavin biosynthesis protein ribG [SYCCPNC]
Ssp_sll0051	*Synechocystis* sp. PCC6803	ORF sll0051 [SYCCPNC]
Ssp_sll1631	*Synechocystis* sp. PCC6803	ORF sll1631 [D90909]
Vav_E2	Variola virus	protein E2 [VE02_VARV]
Vcv_E2	Vaccinia virus	protein E2 [VE02_VACCV]
Vfi_YLXG‡	*Vibrio fischeri*	ORF in luxg 3′ region [YLXG_VIBFI]
Vra_HPP	*Vigna radiata*	vacuolar H+-pyrophosphatase (HPPase) [VRU31467]

New DM domains identified in this work are shown in a different font. The symbol ‡ denotes that the DM domain possesses the PROSITE CYT_DCMP_DEAMINASES signature.

C virus protein, the proton shuttle Glu is valine. Although a few training set sequences have changes in their active site residues, the unusually long L1 FIM and inspection of HMM-generated alignment for non-active site residues suggest that in the absence of other data, these should be considered false positives. (It should be noted, however, that hepatitis delta virus does undergo RNA editing (reviewed in Casey (1996))). Overall, there are (at most) two false positives amongst sequences with scores higher than 27.0 and none with scores higher

FIG. 3. An HMM-generated multiple sequence alignment of DM domains listed in Table 1. New DM domains identified in this work are shown in a different font. Sequences 2, 13, 24, 25, 26 and 74 are partial sequences not used for training the HMM. Amino acids conserved in the majority of the sequences are shown in bold and columns that are predominantly hydrophobic are boxed. Columns containing "." correspond to insert states and numbers indicate the lengths of insertions in sequences at that position (if present). Lines below the alignment highlight some of the features depicted in Figures 1 and 2 as well as other data pertinent to the DM domain. L1 FIM and L2 FIM: open triangles identify the locations of the internal FIMs employed to model regions L1 and L2; Eco_CDD X-ray structure: cylinders and arrow mark the α-helices and β-strand of the DM domain in *Escherichia coli* CDD; Zn, Proton shuttle: open triangles mark the zinc ligands and proton transferring Glu residue in *Escherichia coli* CDD (residues H102, C129, C132 and E104); Mutation sites: positions labelled a-l have been mutated in some sequences. Eco_YFHC: mutation at position a renders the mutant strain resistant to the cell-killing functions encoded by a specific gene family (Poulsen *et al.*, 1989). Rno_APOBEC: mutations at positions c, d, f, h and j abolish mRNA editing but one at position g has no effect (Navaratnam *et al.*, 1995). Hsa_DRADA: mutations at c, d, j and k abolish mRNA editing but mutations at b and l have no effect (Lai *et al.*, 1995). CYT_DCMP_DEAMINASES signature: open circles mark the region covered by the PROSITE signature for the cytidine and deoxycytidine deaminase zinc ion-binding region; the pattern itself is shown.


L1 FIM
L2 FIM
Eco_CDD X-ray structure
Zn
Proton shuttle
Zn
Zn
Mutation sites
CYT_DCMP_DEAMINASES signature
[CH][AV]Ex(2)[LIVMFGA][LIVM]x(17,33)PCx(2,8)Cx(3)[LIVM]

than 35.3. Thus, all complete sequences shown in Figure 3 are classified as possessing a DM domain by the HMM.

Although no sequences used for training have scores below 27.0, it cannot be assumed that there are no false negatives. There may be some DM domains that have diverged to a degree that the current HMM may be too specific and thus unable to classify them as possessing the domain. In addition, deaminase domains that may differ in their primary sequence characteristics from that modelled here would not be detected by the current HMM.

New DM domains

Thirty-seven of the eighty-three sequences in the alignment are new DM domains (sequences depicted in a different font in Table 1 and Figure 3). These new DM domains fall into the following categories: bacterial nitrogen fixation protein (82:Bja_NFP); mouse spermatid protein (19:Mmu_TENR); plant HPPases (10:Ath_HPP, 6:Bvu_HPP1, 7:Bvu_HPP2, 3:Hvu_HPP, 4:Nta_HPP1, 5:Nta_HPP2, 8:Nta_HPP3, 9:Vra_HPP); viral proteins (47:Mcv_E2L, 69:Pcv_A200R, 48:Vav_E2, 49:Vcv_E2); eucaryotic ESTs (70:Ath_EST1, 78:Ath_EST2, 71:Bna_EST, 73:Cel_EST, 24:Hsa_EST1, 27:Hsa_EST3, 74:Hsa_EST4, 83:Hsa_EST5, 2:Mmu_EST1, 29:Sma_EST); bacterial ORFs (79:Ccr_ORF, 42:Mle_ORF, 75:Pga_ORF, 68:Rco_ORF, 72:Ssp_sll0051, 77:Ssp_sll1631); and eucaryotic ORFs (18:Dme_ORF, 25:Hsa_phorI, 28:Lpo_PPLZ20, 40:Sce_L9672, 1:Sce_L8543, 67:Sce_YP9499).

Phylogenetic relationships of DM domains

Figure 4 shows the DM domain phylogenetic tree and reveals three subtrees each with clear substructure.

Subtree A contains only eucaryotic sequences and includes DM domains present in A→I and C→U mRNA editing enzymes (subtrees A1b and A2 respectively). Analysis of the HMM-generated alignment and phylogenetic tree suggests possible substrates for some ORFs possessing the PROSITE CYT_DCMP_DEAMINASES signature. Thus, 30:Sce_YJD5 may be a yeast DM domain involved in C→U mRNA editing; and 56:Vfi_YLXG and 57:Bsu_CME2 may be bacterial DCTDs. Similarly for the remaining DM domains, 11:Mmu_EST2, 13:Mmu_EST3, 14:Cel_T20H4, 18:Dme_ORF, 20:Cel_D2005 and 24:Hsa_EST1 may be fly, human and mouse DM domains involved in A→I mRNA editing; 25:Hsa_phorI, 26:Hsa_EST2, 27:Hsa_EST3, 28:Lpo_PPLZ20 and 29:Sma_EST may be human, plant and platyhelminth DM domains involved in C→U mRNA editing; 36:Mmu_EST4, 40:Sce_L9672 and 42:Mle_ORF may be mouse, yeast and bacterial CDDs; 47:Mcv_E2L, 48:Vav_E2 and 49:Vcv_E2 may be viral proteins with a CDD/BSD-like substrate; 54:Mmu_EST5 may be a mouse DCTD; and 67:Sce_YP9499, 68:Rco_ORF and 69:Pcv_A200R may be yeast, bacterial and viral proteins with a DCTG/ribG-like substrate. Phorbolin I (25:Hsa_phorI) is a protein that is highly expressed in noncultured psoriatic keratinocytes (Rasmussen and Celis, 1993) suggesting the occurrence of C→U mRNA editing in a novel cell type (keratinocytes).

The plant HPPases (subtree A1a) are vacuolar H(+)-translocating inorganic pyrophosphatases, one of two electrogenic proton pumps present in the membrane surrounding the central vacuole of plant cells. HPPases, ubiquitous in plants but otherwise known in only a few phototrophic bacteria, lack sequence identity to any other characterized ion pump implying a different evolutionary origin for this translocase (Rea *et al.*, 1992). The HPPases are most similar to the A→I mRNA editing enzymes raising the possibility that in addition to C→U editing, A→I editing may occur in plants. Alternatively, HPPases may be C→U deaminases acting on ds RNA.

Subtree B has three branches, each containing bacterial, eucaryotic and viral sequences except for B3 which has no viral members. Subtree B1a encompasses the BSDs and CDDs, B2a the DCTDs and B2b the ribG proteins. Three of the four new viral DM domains, 47:Mcv_E2L, 48:Vav_E2 and 49:Vcv_E2 belong to the CDD/BSD family (subtree B1b) whilst the fourth, 69:Pcv_A200R, belongs to the DCTD/ribG family (subtree B2). The two known bacteriophage DM domains (50:BT2_DCTD and 51:BT4_DCTD) are in subtree B2a. The roles of these virally-encoded DM domains in pathogenesis remains to be seen. Subtree B3 contains no sequences which have been characterized so the substrate for this family remains to be defined (70:Ath_EST1, 71:Bna_EST1, 72:Ssp_sll0051, 73:Cel_EST, 74:Hsa_EST4, 75:Pga_ORF, 76:Bsu_YAAJ, 77:Ssp_sll1631 and 78:Ath_EST2).

Subtree C contains bacterial and eucaryotic sequences of unknown function.

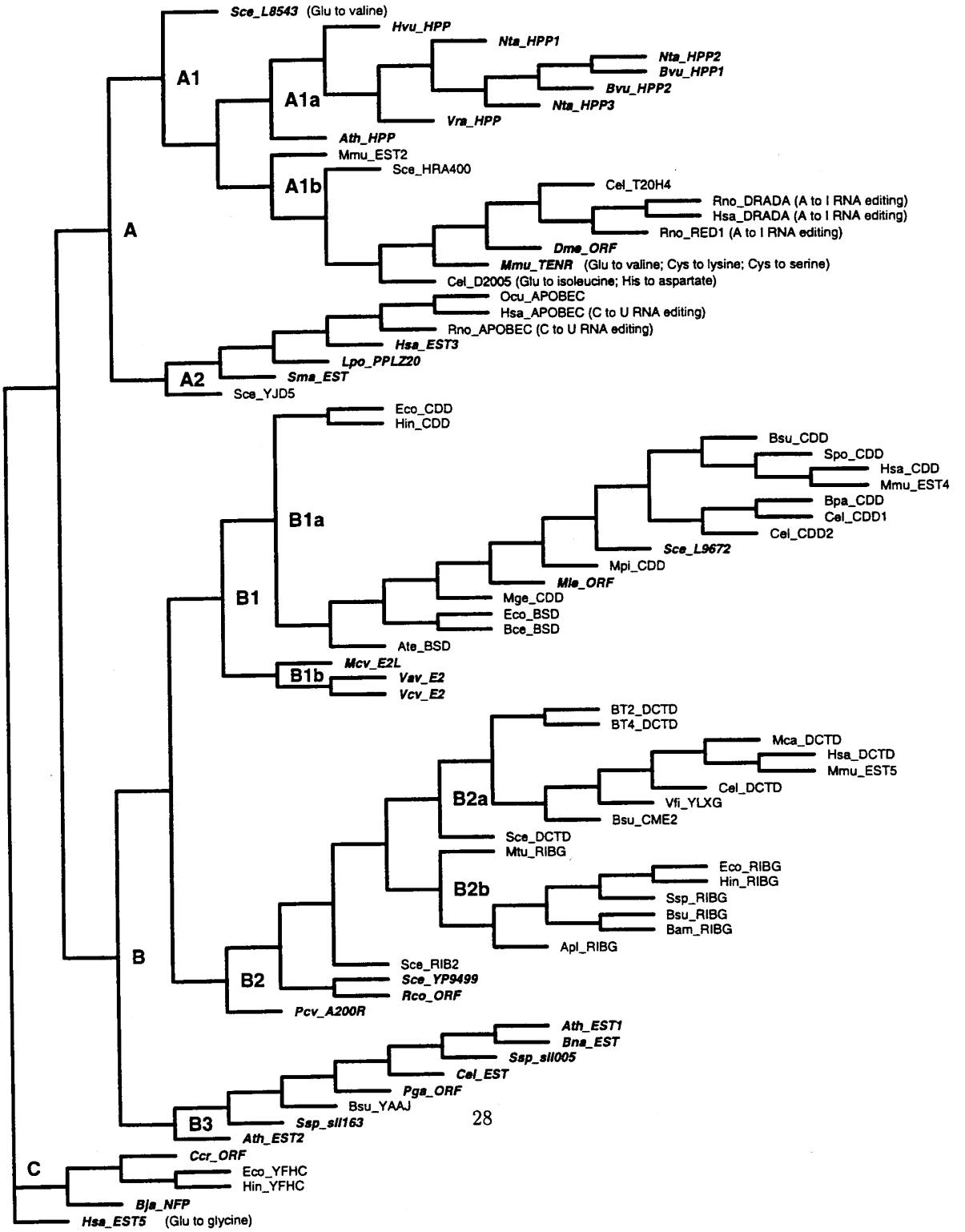


FIG. 4. Phylogenetic tree of DM domains computed using a maximum likelihood approach (based upon the alignment shown in Figure 3). New DM domains in this work are shown in a different font. Sequences in which the active site His, Cys, Cys and proton shuttle Glu are not conserved and those known to be involved in mRNA editing are identified. Subtrees discussed in the text are labelled at their root.

DISCUSSION

The work here has focussed on creating and using a statistical model of a DM domain involved in performing deamination reactions. The results demonstrate that the DM domain HMM overcomes the shortcomings of the PROSITE pattern and highlights the ability of HMMs to model a family of related sequences and to detect remote homologs. The HMM is sensitive in that the numbers of false positives and false negatives are likely to be small. However, the assessment of two false positives is based on primary sequence characteristics but an assessment based on actual function could differ and may be higher. Given that the regions flanking a DM domain are not considered here (for example, the region in white in Figure 1), possession of a DM domain may be a necessary but not sufficient condition in determining whether a protein can perform a deamination reaction. Thus, although sequences in Table 1 are assumed to have a common underlying structure, a DM domain, only additional data can indicate how many behave as deaminases.

Figure 5 shows the most highly conserved residues in the DM domain (residues in bold in Figure 3) and indicates they are generally confined to active site residues and to those in the core of the domain. It is unclear how many DM domains in which the active site residues are not conserved catalyze deamination reactions. In 54:Mmu_EST5, a Cys is changed to serine but this is not expected to have any consequences on structure or catalytic activity. A similar situation arises in 5:Nta_HPP2 and 36:Mmu_EST4 where the proton shuttle is changed from Glu to a functionally equivalent aspartic acid. However, in four other sequences, 1:Sce_L8543, 19:Mmu_TENR, 20:Cel_D2005 and 83:Hsa_EST5, the active site Glu becomes valine, isoleucine or glycine and in one case (19:Mmu_TENR) the Cys are changed to lysine and serine. Inspection of Figure 5 indicates that all these changes to active site residues could be accommodated structurally. However, in the absence of a proton shuttle, these DM domains would be unable to perform a deamination reaction (unless an editing event restored the active site Glu). Based upon their positions in the tree, 19:Mmu_TENR and 20:Cel_D2005 would be unable to participate in A→I mRNA editing. These results reinforce the notion that substrate binding and catalysis can be segregated, as was demonstrated by previous studies using APOBEC in which mutations that inactivated the enzyme did not affect RNA binding (Navaratnam *et al.*, 1995). Furthermore, they suggest that some DM domains may have a regulatory function in that they could bind but not act upon their substrate. In the case of mRNA editing, this would suggest sequestering of some mRNAs to prevent them from being edited. These DM domain containing proteins would be candidates for trans-acting factors able to protect a site from being edited.

The existence of two new branches in the DM domain phylogenetic tree (subtrees C and B3), each containing diverse bacterial and eucaryotic sequences, suggests the involvment of the DM domain in fundamental cellular processes yet to be characterized. Subtree C contains a bacterial protein, 82:Bja_NFP, required for free-living growth and bacteroid development (Weidenhaupt *et al.*, 1995). In the case of another sequence in this branch, 80:Eco_YFHC, a D→E mutation at position a in Figure 3 renders the mutant strain resistant to the cell-killing functions encoded by a specific gene family which kills cells from within by damaging the cell membrane

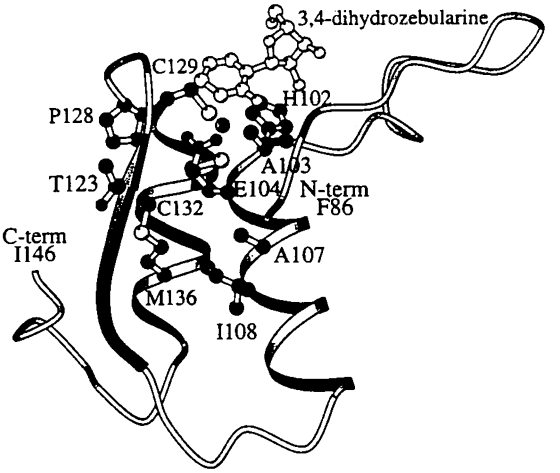


FIG. 5. The DM domain shown in Figures 1 and 2 highlighting all the conserved residues present in the alignment (Figure 3). Residues in bold in the alignment and the bound substrate (white) are depicted in ball-and-stick form; the zinc ion is a sphere.

(Poulsen *et al.*, 1989). Whether this DM domain has a direct or indirect role in cell death and whether the eucaryotic DM domains in this subtree play a comparable role in apoptosis remain to be determined.

With regards to possible false negatives, the simplest explanations for no DM domains having been identified amongst the currently predicted proteins from the archaeon *Methanococcus jannaschii* (Bult *et al.*, 1996) are (i) there are no DM domain containing proteins in this organism, (ii) DM domain containing ORFs have not yet been identified and thus would not have appeared in the databases that were searched using the HMM and (iii) deaminases in this organism may contain a domain distinct from the DM domain modelled here. If, however, such archaeal remote homologs do exist, then the most probable reason for the inability to detect them lies in overfitting and the Dirichlet mixture priors employed currently. Although the training set contained viral, bacterial and eucaryotic sequences, the priors were able to generalize the HMM only to the extent that new DM domains from these phylogenetic groups but not archaea were identified. The Blocks alignments employed to compute the priors used during HMM training (Brown *et al.*, 1993; Sjölander *et al.*, 1996) varied in their ratios of sequences from the three phylogenetic kingdoms. Therefore, their skewed compositions and the relative underrepresentation of archaeal sequences could be the primary factor underlying the inability to detect *M. jannaschii* DM domains. One means to reduce the redundancy and emphasize the diversity of Blocks would be to devise "weighting schemes" for each Block before its use in calculation of the priors. However, given the unequal and unknown evolutionary rates at different sites in biological sequences, this may be problematic. We are exploring a different approach to improving the data sets used to compute Dirichlet mixture priors and thus improve their capacity to generalize HMMs. The approach employs results from genome sequencing projects, avoids weighting schemes, and employs alignments for proteins proposed to be part of the minimal set likely to be necessary for cellular life (Mushegian and Koonin, 1996). Hence, given the current availability of the complete genomes of *E. coli, M. jannaschii* and *S. cerevisiae*, each alignment should contain a minimam of one protein from these organisms and thus from each phylogenetic kingdom. Inclusion of viral sequences, when available, should improve the ability of HMMs to generalize even further. Calculation of intra- and interphylogenetic domain specific priors will be explored.

ACKNOWLEDGMENTS

We thank the two referees for their comments and criticisms and 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-AC03-76F00098. The data are available in electronic form upon request.

REFERENCES

Adachi, J. 1995. *Modelling of Molecular Evolution and Maximum Likelihood Inference of Molecular Phylogency*. Ph.D. dissertation, Institute of Statistical Mathematics, Tokyo.

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.

Altschul, S. 1991. Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* 219, 555–565.

Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

Bairoch, A., Bucher, P., and Hofmann, K. 1996. The PROSITE database, its status in 1995. *Nucleic Acids Research* 24, 189–196.

Baldi, P., Chauvin, Y., Hunkapiller, T., and McClure, M. 1994. Hidden Markov models of biological primary sequence information. *Proc. Natl. Acad. Sci.* 91, 1059–1063.

Barrett, C., Hughey, R., and Karplus, K. 1997. Scoring hidden Markov models. *CABIOS* 13, 191–199.

Barton, G. 1993. ALSCRIPT—A tool to format multiple sequence alignments. *Protein Engineering* 6, 37–40.

Barton, G., and Sternberg, M. 1990. Flexible protein sequence patterns: A sensitive method to detect weak structural similarities. *J. Mol. Biol.* 212 (2), 389–402.

Bateman, A., and Chothia, C. 1996. Fibronectin type III domains in yeast detected by a hidden Markov model. *Current Biology* 6, 1544–1547.

Bateman, A., Eddy, S., and Chothia, C. 1996. Members of the immunoglobulin superfamily in bacteria. *Protein Science* 5, 1939–1941.

Benne, R. 1996. RNA editing: How a message is changed. *Current Opinion Gen. Devel.* 6, 221–231.

Betts, L., Xiang, S., Short, S., Wolfenden, R., and Carter, C., Jr. 1994. Cytidine deaminase. The 2.3Å} crystal structure of an enzyme: transition-state analog complex. *J. Mol. Biol.* 235, 635–656.

Bhattacharya, S., Navaratnam, N., Morrison, J., Scott, J., and Taylor, W. 1994. Cytosine nucleoside/nucleotide deaminases and apolipoprotein B mRNA editing. *TIBS* 19, 105–106.

Blanc, V., Litvak, S., and Araya, A. 1995. RNA editing in wheat mitochondria proceeds by a deamination mechanism. *FEBS Letts.* 373, 56–60.

Bowie, J., Lüthy, R., and Eisenberg, D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253, 164–170.

Brown, M., Hughey, R., Krogh, A., Mian, I., Sjölander, K., and Haussler, D. 1993. Using Dirichlet mixture priors to derive hidden Markov models for protein families. *ISMB* 1, 47–55.

Bult, C., White, O., Olsen, G., Zhou, L., Fleischmann, R., Sutton, G., Blake, J., FitzGerald, L., Clayton, R., Gocayne, J., Kerlavage, A., Dougherty, B., Tomb, J.-F., Aams, M., Reich, C., Overbeek, R., Kirkness, E., Weinstock, K., Merrick, J., Glodek, A., Scott, J., Geoghagen, N., Weidman, J., Fuhrmann, J., Presley, E., Nguyen, D., Utterback, T., Kelley, J., Peterson, J., Sadow, P., Hanna, M., Cotton, M., Hurst, M., Roberts, K., Kaine, B., Borodovsky, M., Klenk, H.-P., Fraser, C., Smith, H., Woese, C., and Venter, J. 1996. Complete genome sequence of the methanogenic archeon, *Methanococcus jannaschii. Science* 273, 1058–1073.

Casey, J. 1996. Hepatitis delta virus. Genetics and pathogenesis. *Clinics in Laboratory Medicine* 16, 451–464.

Cattaneo, R. 1994. Biased (A→I) hypermutation of animal RNA virus genomes. *Current Opinion Gen. Devel.* 4, 895–900.

Dabiri, G., Lai, F., Drakas, R., and Nishikura, K. 1996. Editing of the GLuR-B ion channel RNA in vitro by recombinant double-stranded RNA adenosine deaminase. *EMBO J.* 15, 34–45.

Dalgaard, J., Moser, M., Hughey, R., and Mian, I. 1997. Statistical modeling, phylogenetic analysis and structure prediction of a protein splicing domain common to inteins and hedgehog proteins. *J. Comp. Bio.* 4, 193–214.

Eddy, S. 1996. Hidden Markov models. *Current Opinions in Structural Biology* 6, 361–365.

Fujiwara, Y., Asogawa, M., and Konagaya, A. 1994. Stochastic motif extraction using hidden Markov model. *ISMB* 2, 121–129.

Gray, M. 1996. RNA editing in plant organelles: a fertile field. *Proc. Natl. Acad. Sci.* 93, 8157–8159.

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

Grundy, W., Bailey, T., Elkan, C., and Baker, M. 1997. Hidden Markov model analysis of motifs in steroid dehydrogenases and their homologs. *Biochemical and Biophysical Research Communications* 231, 760–766.

Hahn, J., Inamine, G., Kozlov, Y., and Dubnau, D. 1993. Characterization of *comE*, a late competence operon of *Bacillus subtilis*, required for the binding and uptake of transforming DNA. *Mol. Microbiol.* 10, 99–111.

Hazes, B. 1996. The (QxW)3 domain: A flexible lectin scaffold. *Protein Science* 5, 1490–1501.

Henikoff, J., Pietrokovski, S., and Henikoff, S. 1997. Recent enhancements to the Blocks Database servers. *Nucleic Acids Research* 25, 222–225.

Henikoff, S., and Henikoff, J. 1991. Automated assembly of protein blocks for database searching. *Nucleic Acids Research* 19, 6565–6572.

Herb, A., Higuchi, M., Sprengel, R., and Seeburg, P. 1996. Q/R site editing in kainate receptor GluR5 and GluR6 pre-mRNAs requires distant intronic sequences. *Proc. Natl. Acad. Sci.* 93, 1875–1880.

Herbert, A. 1996. RNA editing, introns and evolution. *TIGS* 12, 6–9.

Hughey, R., and Krogh, A. 1996. Hidden Markov models for sequence analysis: Extension and analysis of the basic method. *CABIOS* 12, 95–107. The hidden Markov model software can be accessed at URL http://www.cse.ucsc.edu/research/compbio/sam.html.

Karplus, K. 1995. Evaluating regularizers for estimating distributions of amino acids. *ISMB* 3, 188–196.

Kobayashi, K., Kamakura, T., Tanaka, T., Yamaguchi, I., and Endo, T. 1991. Nucleotide sequence of the *bsr* gene and N-terminal amino acid sequence of blasticidin S deaminase from blasticidin S resistant *Escherichia coli* TK121. *Agric. Biol. Chem.* 55, 3155–3157.

Kraulis, P. 1991. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *Journal of Applied Crystallography* 24, 946–950.

Krogh, A., Brown, M., Mian, I. S., Sjölander, K., and Haussler, D. 1994. Hidden Markov models in computational biology: Applications to protein modelling. *J. Mol. Biol.* 235, 1501–1531.

Lai, F., Drakas, R., and Nishikura, K. 1995. Mutagenic analysis of double-stranded RNA adenosine deaminase, a candidate enzyme for RNA editing of glutamate-gated ion channel transcripts. *J. Biol. Chem.* 270, 17098–17105.

Lee, C., Szittner, R., Miyamoto, C., and Meighen, E. 1993. The gene convergent to *luxG* in *Vibrio fischeri* codes for a protein related in sequence to RibG and deoxycytidylate deaminase. *Biochim. Biophys. Acta* 1143, 337–339.

Lomeli, H., Mosbacher, J., Melcher, T., Hoger T., Geiger, J., Kuner, T., Monyer, H., Higuchi, M., Bach, A., and Seeburg, P. 1994. Control of kinetic properties of AMPA receptor channels by nuclear RNA editing. *Science* 266, 1709–1713.

Maas, S., Melcher, T., Herb, A., Seeburg, P., Keller, W., Krause, S., Higuchi, M., and O'Connell, M. 1996. Structural

requirements for RNA editing in glutamate receptor pre-mRNAs by recombinant double-stranded RNA adenosine deaminase. *J. Biol. Chem.* 271, 12221–12226.

Maciukenas, M. 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.

Melcher, T., Maas, S., Herb, A., Sprengel, R., Seeburg, P., and Higuchi, M. 1996. A mammalian RNA editing enzyme. *Nature* 379, 460–464.

Mian, I. 1997. Sequence analysis of ribonucleases HII, III, II, PH and D. *Nucleic Acids Research* 25, 3187–3195.

Moore, J., Silversmith, R., Maley, G., and Maley, F. 1993. T4-phage deoxycytidylate deaminase is a metalloprotein containing two zinc atoms per subunit. *J. Biol. Chem.* 268, 2288–2291.

Mushegian, A., and Koonin, E. 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl. Acad. Sci.* 93, 10268–10273.

Navaratnam, N., Bhattacharya, S., Fujino, T., Patel, D., Jarmuz, A., and Scott, J. 1995. Evolutionary origins of apoB mRNA editing: catalysis by a cytidine deaminase that has acquired a novel RNA-binding motif at its active site. *Cell* 81, 187–195.

Nawa, K., Tamura, Y., Sato, K., Hattori, J., Shimotohno, K., and Endo, T. 1995. Inactivation of blasticidin S by *Bacillus cereus*. V. Purification and characterization of blasticidin S-deaminase mediated by a plasmid from blasticidin S resistant *Bacillus cereus* K55-S1. *Biological and Pharmaceutical Bulletin* 18, 350–354.

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.

Ogasawara, N., Nakai, S., and Yoshikawa, H. 1994. Systematic sequencing of the 180 kilobase region of the *Bacillus subtilis* chromosome containing the replication origin. *DNA Research* 1, 1–14.

Pohl, T., and Aljinovic, G. Unpublished.

Polson, A., and Bass, B. 1994. Preferential selection of adenosines for modification by double-stranded RNA adenosine deaminase. *EMBO J.* 13, 5701–5711.

Polson, A., Bass, B., and Casey, J. 1996. RNA editing of hepatitis delta virus antigenome by dsRNA-adenosine deaminase. *Nature* 380, 454–456.

Poulsen, L., Larsen, N., Molin, S., and Andersson, P. 1989. A family of genes encoding a cell-killing function may be conserved in all gram-negative bacteria. *Mol. Microbiol.* 3, 1463–1472.

Poulsen, L., Larsen, N., Molin, S., and Andersson, P. 1992. Analysis of an *Escherichia coli* mutant strain resistant to the cell-killing function encoded by the *gef* gene family. *Mol. Microbiol.* 6.

Rabiner, L. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77, 257–286.

Rabiner, L., and Juang, B. 1986. An introduction to hidden Markov models. *IEEE ASSP Magazine* 3, 4–16.

Rasmussen, H., and Celis, J. 1993. Evidence for an altered protein kinase C (PKC) signaling pathway in psoriasis. *Journal of Investigative Dermatology* 101, 560–566.

Rea, P., Kim, Y., Sarafian, V., Poole, R., Davies, J., and Sanders, D. 1992. Vacuolar H (+)-translocating pyrophosphatases: a new category of ion translocase. *TIBS* 17, 348–353.

Reizer, J., Buskirk, S., Bairoch, A., Reizer, A., and Saier, M., Jr. 1994. A novel zinc-binding motif found in two ubiquitous deaminase families. *Protein Science* 3, 853–856.

Rueter, S., Burns, C., Coode, S., Mookherjee, P., and Emeson, R. 1995. Glutamate receptor RNA editing in vitro by enzymatic conversion of adenosine to inosine. *Science* 267, 1491–1494.

Scott, J. 1995. A place in the world for RNA editing. *Cell* 81, 833–836.

Shub, D., Goodrich-Blair, H., and Eddy, S. 1994. Amino acid sequence motif of group I intron endonucleases is conserved in open reading frames of group II introns. *TIBS* 19, 402–404.

Sjölander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I., and Haussler, D. 1996. Dirichlet mixtures: a method for improving detection of weak but significant protein sequence homology. *CABIOS* 12, 327–345.

Skuse, G., Cappione, A., Snowden, M., Metheny, L., and Smith, H. 1996. The neurofibromatosis type I messenger RNA undergoes base-modification RNA editing. *Nucleic Acids Research* 24, 478–486.

Smith, H., and Sowden, M. 1996. Base-modification mRNA editing through deamination—the good, the bad and the unregulated. *TIGS* 12, 418–424.

Sorokin, A., Zumstein, E., Azevedo, V., Ehrlich, S., and Serror, P. 1993. The organization of the *Bacillus subtilis* 168 chromosome region between the *spoVA* and *serA* genetic loci, based on sequence data. *Mol. Microbiol.* 10, 385–395.

Struck, J., Far, R., Schröder, W., Hucho, F., Toschka, H., and Erdmann, V. 1990. Characterization of a 17 kDa protein gene upstream from the small cytoplasmic RNA gene of *Bacillus subtilis*. *Biochim. Biophys. Acta* 1050, 80–83.

Tatusov, R., Altschul, S., and Koonin, E. 1994. Detection of conserved segments in proteins: Iterative scanning of sequence databases with alignment blocks. *Proc. Natl. Acad. Sci.* 91, 12091–12085.

Taylor, W. 1986. Identification of protein sequence homology by consensus template alignment. *J. Mol. Biol.* 188, 233–258.

Waterman, M., and Perlwitz, M. 1986. Line geometries for sequence comparisons. *Bull. Math. Biol.* 46, 567–577.

Weidenhaupt, M., Schmid-Appert, M., Thony, B., Hennecke, H., and Fischer, H. 1995. A new *Bradyrhizobium japonicum* gene required for free-living growth and bacteroid development is conserved in other bacteria and in plants. *Molecular Plant-Microbe Interactions* 8, 454–464.

Yang, C., Carlow, D., Wolfenden, R., and Short, S. 1992. Cloning and nucleotide sequence of the *Escherichia coli* cytidine deaminase (ccd) gene. *Biochemistry* 31, 4168–4174.

Yu, W., and Schuster, W. 1995. Evidence for a site-specific cytidine deamination reaction involved in C to U RNA editing of plant mitochondria. *J. Biol. Chem.* 270, 18227–18233.

Address Correspondence to:
I. Saira Mian
Life Sciences Division (Mail Stop 29-100)
Lawrence Berkeley National Laboratory
1 Cyclotron Road
Berkeley, CA 94720

SMian@lbl.gov

Received for publication March 26, 1997; accepted as revised August 8, 1997.

This article has been cited by:

1. Arka Mallela, Kazuko Nishikura. 2012. A-to-I editing of protein coding and noncoding RNAs. *Critical Reviews in Biochemistry and Molecular Biology* **47**:6, 493-501. [CrossRef]
2. Eva Maria Novoa, Mariana Pavon-Eternod, Tao Pan, Lluís Ribas de Pouplana. 2012. A Role for tRNA Modifications in Genome Structure and Codon Usage. *Cell* **149**:1, 202-213. [CrossRef]
3. C. P. Godfried Sie, M. Kuchka. 2011. RNA Editing adds flavor to complexity. *Biochemistry (Moscow)* **76**:8, 869-881. [CrossRef]
4. Bjorn-Erik Wulff, Kazuko Nishikura. 2010. Substitutional A-to-I RNA editing. *Wiley Interdisciplinary Reviews - RNA* n/a-n/a. [CrossRef]
5. Pufeng Du, Yang Chen, Yanda Li. 2009. Computational evidence of A-to-I RNA editing in nucleus transcriptome of Arabidopsis thaliana. *Frontiers of Electrical and Electronic Engineering in China* **4**:4, 349-361. [CrossRef]
6. Pufeng Du, Yanda Li. 2009. Computational analysis of RNA editing: seeking tiny discrepancies between transcriptome and genome. *Frontiers of Electrical and Electronic Engineering in China* **4**:3, 251-258. [CrossRef]
7. Harold C. SmithMeasuring Editing Activity and Identifying Cytidine-to-Uridine mRNA Editing Factors in Cells and Biochemical Isolates **424**, 387-416. [CrossRef]
8. Charles M. Connolly, Andrea T. Dearth, Robert E. Braun. 2005. Disruption of murine Tenr results in teratospermia and male infertility. *Developmental Biology* **278**:1, 13-21. [CrossRef]
9. Y ZHAO. 2005. Identification of the activation-induced cytidine deaminase gene from zebrafish: an evolutionary analysis*1. *Developmental & Comparative Immunology* **29**:1, 61-71. [CrossRef]
10. K. Xie, M. P. Sowden, G. S. C. Dance, A. T. Torelli, H. C. Smith, J. E. Wedekind. 2004. The structure of a yeast RNA-editing deaminase provides insight into the fold and function of activation-induced deaminase and APOBEC-1. *Proceedings of the National Academy of Sciences* **101**:21, 8114-8119. [CrossRef]
11. Sara L. Sawyer, Michael Emerman, Harmit S. Malik. 2004. Ancient Adaptive Evolution of the Primate Antiviral DNA-Editing Enzyme APOBEC3G. *PLoS Biology* **2**:9, e275. [CrossRef]
12. Joseph E. Wedekind, Geoffrey S.C. Dance, Mark.P. Sowden, Harold C. Smith. 2003. Messenger RNA editing in mammals: new members of the APOBEC family seeking roles in the family business. *Trends in Genetics* **19**:4, 207-216. [CrossRef]