7016–7045 Nucleic Acids Research, 2012, Vol. 40, No. 15 Published online 25 May 2012
doi:10.1093/nar/gks382

SURVEY AND SUMMARY

Sequence, structure and functional diversity of PD-(D/E)XK phosphodiesterase superfamily

Kamil Steczkiewicz[1], Anna Muszewska[1], Lukasz Knizewski[1], Leszek Rychlewski[2] and Krzysztof Ginalski[1,*]

[1]Laboratory of Bioinformatics and Systems Biology, CENT, University of Warsaw, Zwirki i Wigury 93, 02-089 Warsaw and [2]BioInfoBank Institute, Limanowskiego 24a, 60-744 Poznan, Poland

Received February 13, 2012; Revised April 11, 2012; Accepted April 13, 2012

ABSTRACT

Proteins belonging to PD-(D/E)XK phosphodiesterases constitute a functionally diverse superfamily with representatives involved in replication, restriction, DNA repair and tRNA–intron splicing. Their malfunction in humans triggers severe diseases, such as Fanconi anemia and Xeroderma pigmentosum. To date there have been several attempts to identify and classify new PD-(D/E)KK phosphodiesterases using remote homology detection methods. Such efforts are complicated, because the superfamily exhibits extreme sequence and structural divergence. Using advanced homology detection methods supported with superfamily-wide domain architecture and horizontal gene transfer analyses, we provide a comprehensive reclassification of proteins containing a PD-(D/E)XK domain. The PD-(D/E)XK phosphodiesterases span over 21 900 proteins, which can be classified into 121 groups of various families. Eleven of them, including DUF4420, DUF3883, DUF4263, COG5482, COG1395, Tsp45I, HaeII, Eco47II, ScaI, HpaII and Replic_Relax, are newly assigned to the PD-(D/E)XK superfamily. Some groups of PD-(D/E)XK proteins are present in all domains of life, whereas others occur within small numbers of organisms. We observed multiple horizontal gene transfers even between human pathogenic bacteria or from Prokaryota to Eukaryota. Uncommon domain arrangements greatly elaborate the PD-(D/E)XK world. These include domain architectures suggesting regulatory roles in Eukaryotes, like stress sensing and cell-cycle regulation. Our results may inspire further experimental studies aimed at identification of exact biological functions, specific substrates and molecular mechanisms of reactions performed by these highly diverse proteins.

INTRODUCTION

The large and extremely diverse superfamily of PD-(D/E)XK phosphodiesterases is a remarkable example of adopting a common structural scaffold to various biological activities. These enzymes encompass mainly nucleases (and their inactive homologs) and fill in a variety of functional niches including DNA restriction (1), tRNA splicing (2), transposon excision (3), DNA recombination (4), Holliday junction (HJC) resolving (5), DNA repair (6), Pol II termination (7), or DNA binding (8). The involvement of PD-(D/E)XK enzymes in housekeeping processes suggests that these proteins may be engaged in the development of genetic diseases. It should be noted that PD-(D/E)XK phosphodiesterases exhibit very little sequence similarity, despite retaining a common core fold and a few residues responsible for the cleavage. The extreme sequence diversity, multiple insertions to a relatively small structural core, circular permutations (9) and migration of active site residues (10) render this superfamily a difficult subject to homology inference and hinders a new family identification with traditional sequence- or even structure-based approaches. In the present study our aim was to identify, classify and expand the existing repertoire of proteins belonging to the PD-(D/E)XK fold, in order to obtain a more complete picture of this superfamily.

The common conserved structural core of PD-(D/E)XK phosphodiesterases consists of a central, four-stranded, mixed β-sheet flanked by two α-helices on both sides (with αβββαβ topology), forming a scaffold adopted for

*To whom correspondence should be addressed. Tel: +48 22 5540800; Fax: +48 22 5540801; Email: kginal@cent.uw.edu.pl

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

the active site formation (11) (Figures 1 and 2). This architecture and topology are classified in SCOP (Structural Classification of Proteins) database (12) as a restriction endonuclease-like fold. The active site is located in a characteristic β-sheet Y-shaped bend (the second and third core β-strands) that exposes the catalytic residues (aspartic acid, glutamic acid and lysine, in a canonical active site) from the relatively conserved PD-(D/E)XK motif. In addition to the aforementioned motif, the conserved acidic residues from the core α-helices (usually glutamic acid from the first α-helix) often contribute to active site formation at least in a subset of families (10,13). Altogether, these residues play various catalytic roles which include coordination of up to three divalent metal ion cofactors, depending on the family. In addition, the residues from the second, positively charged α-helix can also contribute to the active site, although their major role is to facilitate the substrate binding and quaternary structure formation (14). The last, fourth core β-strand tends to be strongly hydrophobic as it is buried deeply within the hydrophobic core of the structure. This α/β/α sandwich fold is capable of accommodating a number of modifications (15) that often blur the image of the canonical structure of these enzymes. For a long time, proteins belonging to the PD-(D/E)XK nuclease-like superfamily had been considered as restriction enzymes, exclusively. However, many later experiments showed their contribution to DNA-branched structures resolving (5), double-strand breaks maintenance (16), or RNA maturation (17). In the following years PD-(D/E)XK phosphodiesterases were extensively studied, reclassified (18) and their realm was consequently enlarged. Currently, there are 60 diverse families grouped into the 'PD-(D/E)XK nuclease superfamily' clan in the Pfam 26 database (19). This clan includes restriction enzymes, HJC resolvases, herpes virus exonucleases and various other nucleases from all kingdoms of life, sugar fermentation proteins, and several domains of unknown functions (DUFs). In addition, there are over 100 structures of PD-(D/E)XK nucleases cataloged in SCOP database (12) clustered into four main groups, encompassing restriction endonuclease-like enzymes, tRNA–intron splicing endonucleases, eukaryotic RPB5 N-terminal domain and TBP-interacting protein-like.

The PD-(D/E)XK proteins constitute a functionally diverse superfamily that addresses multiple nucleic acid maintenance issues. For instance, PD-(D/E)XK domain occurs in all classes of restriction enzymes, including those of type I, II, III and IV. Type II restriction endonucleases form the most diverged group of PD-(D/E)XK phosphodiesterases. These enzymes, in concert with methyltransferases, set up the restriction–modification systems which protect bacterial and archaeal genomes against foreign genetic material (21). Host DNA is marked through methylation and therefore it is protected from accidental cleavage by a restriction enzyme which recognizes only unmethylated, foreign nucleic acid. Jeltsch and Pingoud proposed an evolutionary dependence between methyltransferases and restriction endonucleases (22). They managed to show that bacterial cells had acquired both a relevant methyltransferase and a restriction enzyme simultaneously in order to provide sufficient protection of host genetic material. Other restriction endonuclease-like fold proteins include mismatch repairing enzymes MutH and Vsr. These enzymes are a part of the machinery that recognizes and removes nucleotides improperly incorporated during recombination. MutH, which is a part of the MutHLS mismatch repair system, is a methylation- and sequence-specific nuclease (6,23). Vsr nuclease is a part of the Very Short Patch Repair system which aids MutHLS deficiency connected with the methylated cytosine spontaneous deamination. The PD-(D/E)XK proteins can also resolve HJC emerging from homologous recombination. HJC fastens together two homologous DNA molecules which, if unresolved, can lead to mutations (24). There are several PD-(D/E)XK protein families conserved through all kingdoms of life that recognize and cut branched DNA structures. These enzymes include RecU (25) and bacteriophage T7 HJC resolvase (endonuclease I) involved in genetic recombination during viral infection (26). XPF, ERCC4, Mus81 and Dna2 are also PD-(D/E)XK nucleases with structure-based specificity for DNA branched structures (27,28). They may cleave HJC or, as proven for Dna2, cut the remaining long flap RNA primers during the Okazaki fragment maturation (29). XPF was identified to process damaged DNA structures in mammalian nucleotide excision repair (NER) (27). Additionally, together with ERCC1, it cleaves DNA duplexes during homologous recombination. Mus81 participates in recombination and cell-cycle regulation (28). PD-(D/E)XK phosphodiesterases also embrace exoribonucleases involved in homologous recombination and various DNA repair pathways, including RecB and its inactive homolog RecC from the RecBCD complex (16). The assortment of functional niches for PD-(D/E)XK proteins also encompasses mobile genetic element transposition, exemplified by TnsA transposase (3). Viral nucleases constitute another PD-(D/E)XK group. The alkaline exonuclease maintains extensively expressed viral DNA and degrades host mRNA molecules (30). Bacteriophage λ-exonuclease facilitates double strand break repair and single strand annealing (31). An eukaryotic Rai1-like (PF08652, KOG1982) plays an important role in pre-rRNA maturation by removing two phosphates from the 5′-termini leaving a 5′-monophosphate (7). The mitochondrial, membrane-bound Pet127 (PF08634) protein is suggested to process the apocytochrome-b precursor during mRNA maturation (32). RPB5, a universal subunit of all three major eukaryotic RNA polymerase complexes, also retains the PD-(D/E)XK fold. RPB5 interacts with several transcription factors, such as TFIIB or HBx, and the TIP120 pre-initiation complex (8). The tRNA splicing endonucleases that constitute a well distinguishable group of archaeal and eukaryotic proteins within the PD-(D/E)XK phosphodiesterase realm are a very interesting example of alternative function gain through acquisition of a novel active site. They are vital for maturation of tRNA molecules by performing intron excision from an anticodon loop (2). Their activity is crucial for tRNA intron identification and removal, allowing ligases and

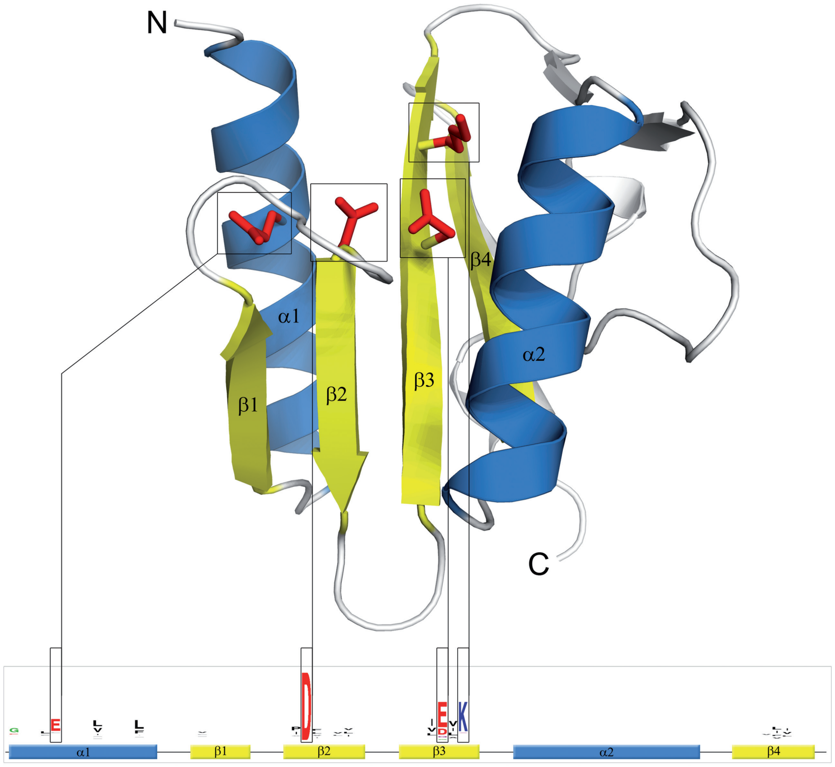


Figure 1. The commonly conserved core of PD-(D/E)XK nuclease fold. Critical active site residues are shown as red sticks and marked in corresponding sequence logo. Sequence logo was derived from multiple sequence alignment for PD-(D/E)XK phosphodiesterase superfamily using WebLogo (20).

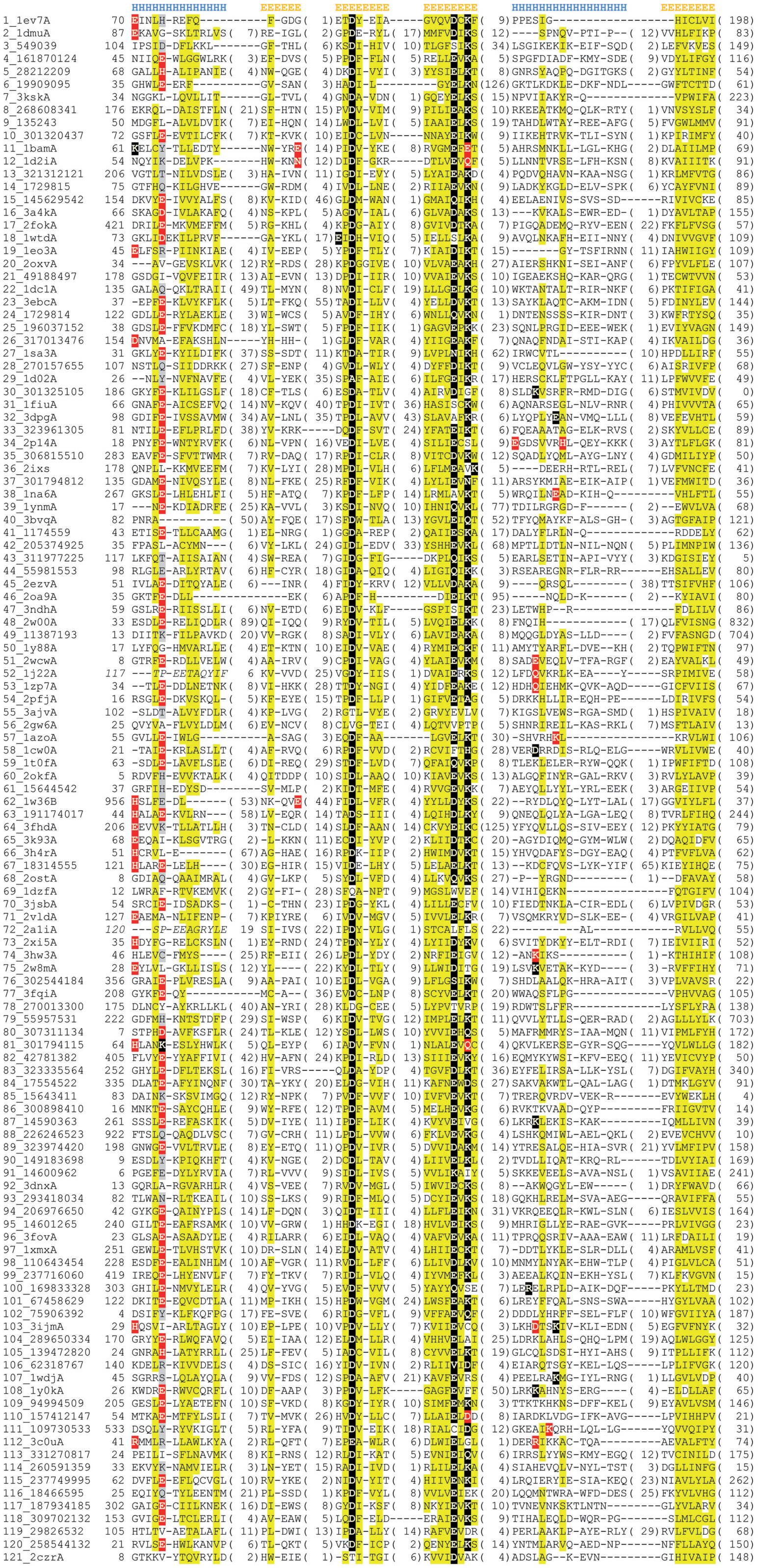
phosphotransferases to complete the tRNA maturation process.

In humans, the malfunction of some PD-(D/E)XK phosphodiesterases is linked to severe, inherited diseases involving neurological abnormalities and susceptibility to develop early onset malignancies. Mutations in tRNA splicing endonuclease lead to pontocerebellar hypoplasia (PCH) (33) which is related to mental and motor impairments. Mutations in XPF–ERCC1, an NER repair pathway structure-dependent endonuclease, are one of the primary causes of xeroderma pigmentosum (XP) (34). XP manifests itself by increased sensitivity to sunlight with the development of carcinomas. Fanconi anemia (FA) is a consequence of mutations in PD-(D/E)XK proteins [e.g. FANCM (35)], participating in DNA repair and involves developmental abnormalities, bone marrow failure, and a predisposition to cancer.

Up to date there have been several attempts to identify and classify new PD-(D/E)XK phosphodiesterases, such as YhgA (36), UL24 (37), NERD (38), CoiA (39), RmuC (39) protein families or various restriction enzymes (1). Those studies were mainly based on remote homology detection methods, as the extreme sequence divergence of the PD-(D/E)XK enzymes remains the main obstacle in detection of new superfamily members. This inspired the development of a dedicated SVM (Support Vector Machines) algorithm for the identification of the PD-(D/E)XK active site signature within protein sequences (11). The discussed analyses covered a large part of the PD-(D/E)XK phosphodiesterase world, however each approach individually relied on a limited set of initial sequences and did not provide a widespread view on the PD-(D/E)XK fold. Therefore, in order to confer our work a broader perspective, first we collected the structures and families annotated as restriction endonuclease-like enzymes. This set was used as a starting point for exhaustive, transitive fold recognition searches aiming to obtain the most complete set of PD-(D/E)XK proteins available in current databases. Here we report a comprehensive reclassification of proteins containing a PD-(D/E)XK domain, including their domain architecture, taxonomic distribution and genomic context.

MATERIALS AND METHODS

A brief overview of our methods is presented below with further details given in Supplementary Materials (see 'Materials and Methods' section). Detection of PD-(D/E)XK families (Pfam, COG, KOG) and structures (PDB90) was performed with a distant homology detection method, Meta-BASIC (40). Non-trivial assignments were additionally confirmed with a consensus of fold recognition, 3D-Jury (41). Sequences of proteins belonging to the identified families were collected with PSI-BLAST (42) searches against NCBI nr database. Multiple sequence alignments were prepared using PCMA (43). In addition, structure-based alignment was derived from a manually curated superimposition of PD-(D/E)XK



```
                   HHHHHHHHHHHHHHH    EEEEEE     EEEEEEEE     EEEEEEEE     HHHHHHHHHHHHHHHHHH    EEEEEEEE
1_1ev7A         70 EINLH-REFQ----------F-GDG(  1)ETDY-EIA-----GVQVDCKF(  9)PPESIG------------------HICLVI( 198)
2_1dmuA         87 EKAVG-SKLTRLVS(  7)RE-IGL(  2)GPDE-RYL( 17)MMFVDIKS( 12)----SPNQV-PTI-P--( 12)VVHLFIKP(  83)
3_549039       104 IPSID-DFLKKL-------GF-THL(  6)KSDI-HIV( 10)TLGFSIKS( 34)LSGIKEKIK-EIF-SQD(  2)LEFVKVES( 149)
4_161870124     45 NIIQE-WLGGWLRK(  3)EF-DVS(  5)PPDF-FLN(  5)RELLEVKA(  5)SPGFDIADF-KMY-SDE(  9)VDYLIFGY( 116)
5_28212209      68 GALLH-ALIPANIE(  4)NW-QGE(  4)DKDI-VYI(  3)YYSIELKT(  8)GNRSYAQPQ-DGITGKS(  2)GYYLTINF(  54)
6_19909095      35 GHWLE-ERF----------GV-SAN(  3)HADI-L---------GYELKN(126)GKTLKDKLE-DKF-NDL-----GWFTCTTD(  61)
7_3kskA         34 NGGKL-LQVLLIT------GL-TVL(  4)GNDA-VDN(  1)GQEYELKS( 15)NPVIIAKYR-Q-------------VPWIFA( 223)
8_268608341    176 EKRQL-DAISTFLN( 21)SF-HTN( 15)PVDV-VIM(  9)PILIEAKS( 10)RKEEATKMQ-QLK-RTY(  1)VNVSYSLF(  34)
9_135243        50 MDGFL-ALVLDVIK(  7)EI-HQN( 14)LWDL-LVI(  5)IAAIELKS( 19)TAHDLWTAY-REE-AFG(  5)FVGWLMMV(  91)
10_301320437    72 GSFLE-EVTILCFK(  7)KT-KVK( 10)EIDC-LVN-----NNAYEIKW(  9)IIKEHTRVK-TLI-SYN(  1)KPIRIMFY(  40)
11_1bamA        61 KELCY-TLLEDTY------NW-YRE( 14)PIDV-YKE(  8)RVGMEFET(  5)AHRSMNKLL-LGL-KHG(  2)DLAIILMP(  69)
12_1d2iA        54 NQYIK-DELVPK-------HW-KNN( 12)DIDF-GKR-----DTLVEVQF(  5)LLNNTVRSE-LFH-KSN(  8)KVAIIITK(  91)
13_321312121   206 VGTLT-NILVDSLE(  3)HA-IVN( 11)IGDI-EVY(  5)LYAIEAKD(  4)PQDVQHAVN-KAA-NSG(  1)KRLMFVTG(  86)
14_1729815      75 GTFHQ-KILGHVE------GW-RDM(  4)IVDL-LNE(  2)RIIAEVKN(  9)LADKYKGLD-ELV-SPK(  6)YCAYFVNI(  89)
15_145629542   154 DKVYE-IVVYALFS(  8)KV-KID( 46)GLDM-WAN(  1)GMAIQIKH(  4)EELAENIVS-SVS-SD------RIVIVCKE(  85)
16_3a4kA        66 SKAGD-IVLAKAFQ(  4)NS-KPL(  5)AGDV-IAL(  6)GLVADAKS( 13)---KVKALS-EWR-ED-(  1)DYAVLTAP( 155)
17_2fokA       421 DRILE-MKVMEFFM(  5)RG-KHL(  4)KPDG-AIY(  8)GVIVDTKA(  7)PIGQADEMQ-RYV-EEN( 22)FKFLFVSG(  57)
18_1wtdA        73 GKLIDEKILPRVF------GA-YKL( 17)EIDH-VIQ(  5)IELLSLKA(  9)AVQLNKAFH-EII-NNY(  4)DNIVVGVF( 109)
19_1eo3A        45 ELFSR-PIINKIAE(  4)IV-EEP(  5)YPDF-TLY(  7)KIAIDIKT( 15)-------GY-TSFIRNN( 11)IAHWIIGY( 109)
20_2oxvA        34 ---AV-GEVSKLVK( 12)RY-RDS( 28)KPDGGIVE( 10)VLVAEAKH( 27)AIERSHKNI-SEI-ANF(  6)FPYVLFLE( 107)
21_49188497    178 GSDGI-VQVFEIIR( 13)AI-EVN( 13)DPDI-IIR( 10)VVAIEVKS( 10)IGEAEKSHQ-KAR-QRG(  1)TECWTVVN(  53)
22_1dc1A       135 GALAQ-QKLTRAII( 49)TL-MYN(  8)NVDI-CLF( 19)LLLGELKG( 10)WKTANTALT-RIR-NKF(  6)PKTIFIGA(  41)
23_3ebcA        37 -EPFE-KLVYKFLK(  5)LT-FKQ( 55)TADI-LLV(  4)YELLDVKT( 13)SAYKLAQTC-AKM-IDN(  5)FDINYLEV( 144)
24_1729814     122 GDLLE-RYLAEKLE(  3)WI-WCS(  5)AVDF-IHY(  6)WNLLQVKN(  1)DNTENSSSS-KIR-DNT(  3)KWFRTYSQ(  35)
25_196037152    38 GDSLE-FFVKDMFC( 18)YL-SWT(  5)FPDF-IIK(  1)GAGVEPKK( 23)SQNLPRGID-EEE-WEK(  1)EVIYVAGN( 149)
26_317013476   154 DNVMA-EFAKSHLN-----YH-HH-(  1)GLDF-VAR(  4)YIIGEAKF(  7)QNAQFNDAI-STI-KAP(  4)IKVAILDG(  36)
27_1sa3A        31 GKLYE-KYILDIFK( 37)SS-SDT( 11)KTDA-TIR(  9)LVPLNIKH( 62)IRWCVTL----------( 10)HPDLLIRF(  55)
28_270157655   107 NSTLQ-SIIDDRKK( 27)SF-ENP(  4)GVDL-WLY(  4)DYFFDIKT(  9)VCLEQVLGW-YSY-YYC(  6)AISRIVFP(  68)
29_1d02A        26 --NLY-NVFNAVFE(  4)VL-YEK( 35)SPAF-AIE(  6)ILFGEIKR( 17)HERSCKLFTPGLL-KAY( 11)LPFWVVFE(  49)
30_301325105   186 GKYFE-KLILGSLF( 18)CF-TLS(  6)ESDA-TLV(  4)AIRVDIGF(  8)SLDKVSRFR-RMD-DIG(  6)STMVIVDV(   0)
31_1fiuA        66 GNAFE-AICSEFVQ( 14)NV-KQV( 40)TPDI-IVT( 36)HASISCKW(  6)AQNARSEGL-NLV-RNR(  4)PHIVVVTA(  65)
32_3dpgA        98 GDIFE-IVSSAVMW( 34)AV-LNL( 35)TPDL-AVV( 43)SLAVAFKR(  6)LYQPLYEAN-VMQ-LLL(  8)VEFEVHTL(  59)
33_323961305    81 NTILE-EFLPRLFD( 38)YV-KRK-----DQDF-SVT( 24)YFATEIKT(  6)FQEAAATAG-ELK-RVS(  2)SKYVLLCE(  89)
34_2p14A        18 PNYFE-WNTYRVFK(  6)NL-VPN( 16)VEDI-LVE(  4)SILIECSL(  9)EGDSVVRHL-QEY-KKK(  3)AYTLFLGK(  81)
35_306815510   283 EAVFE-SFVTTWMR(  7)RV-DAQ( 17)RPDI-CLR(  8)VITCDVKW( 12)SQADLYQML-AYG-LNY(  4)GDMILIYP(  50)
36_2ixs        178 QNPLL-KKMVEEFM(  7)KV-LYI( 28)MPDL-VLH(  6)LFLMEAVK(  5)----DEERH-RTL-REL(  7)LVFVNCFE(  41)
37_301794812   135 GDAME-NIVQSYLE(  8)NL-FKE( 25)RFDF-VIK(  4)LYLIEVNF( 11)ARSYKMIAE-EIK-AIP(  1)VEFMWITD(  36)
38_1na6A       267 GKSLE-LHLEHLFI(  5)HF-ATQ(  7)KPDF-LFP( 14)LRMLAVKT(  5)WRQILNEAD-KIH-Q---------VHLFTL(  55)
39_1ynmA        17 ---NE-KDIADRFE( 25)KA-VVL(  3)KSDI-NVQ( 10)IHNIQVKL( 77)TDILRGRGD-F------(  2)--EWVLVA(  68)
40_3bvqA        82 PNRA----------( 50)AY-FQE( 17)SFDW-TLA( 13)YGVLEIQT( 52)TFYQMAYKF-ALS-GH-(  3)AGTGFAIP( 121)
41_1174559      43 ETISE-TLLCAAMG(  1)EL-NRG(  1)GGDA-REP(  2)QAIIESKA( 17)DALYFLRLN-Q-------------RDDELY(  62)
42_205374925    35 FPASL-ACYMN---(  6)VY-LKL( 24)GIDL-EDV( 33)YSHHEVKL( 68)MPTLIDTLN-NIL-NQN(  5)PLIMNPIW( 136)
43_311977225   117 LKFQT-AIISAIAN(  4)SW-REA(  7)GIDG-FIG-----DKPLQIKS(  5)EARLSETIN-API-VYY(  3)KDGISIEY(   5)
44_55981553     98 RLGLE-ARLYRTAV(  6)HF-CYR( 18)GIDA-QIQ(  4)LIGIQIKK(  3)RSEAREGNR-FLR-RR------EHSALLEV(  80)
45_2ezvA        51 IVLAE-DITQYALE(  6)---INR(  4)FIDY-KRV( 12)VLLVDAKA(  9)----QRSQL--------( 38)TTSIFVHF( 106)
46_2oa9A        35 GKTFE-DLL--------------EK(  6)APDF-H---------DIEIKT( 95)-----NQLD-K------(  2)--KYIAVI(  80)
47_3ndhA        59 GSLRE-RIISSLLI(  6)NV-ETD(  6)EIDV-KLF-----GSPISIKT( 23)LETWHP--R---------------FDLILV(  86)
48_2w00A        33 ESDLE-RELIQDLR( 89)QI-IQQ(  9)RYDV-TIL(  5)LVQIELKK(  8)FNQIH------------( 17)QLFVISNG( 832)
49_11387193     13 DIITK-FILPAVKD( 20)VV-RGK(  8)SADI-VLY(  6)LAVIEAKA(  8)MQQGLDYAS-LLD----(  2)FVFASNGD( 704)
50_1y88A        17 LYFQG-HMVARLLE(  4)ET-KTN( 10)EIDV-VAE(  4)RYMIECKF( 11)AMYTYARFL-DVE--KH(  2)TQPWIFTN(  97)
51_2wcwA         8 GTRFE-RDLLVELW(  4)AA-IRV(  9)CPDI-VAG(  4)YLAIEVKM(  8)SADEVEQLV-TFA-RGF(  2)EAYVALKL(  49)
52_1j22A       117 ---TP-EETAQYIF   6 KV-VVD( 25)VGDY-IIS(  1)DVAIERKS( 12)LFDQVKRLK-EA----Y-----SRPIMIVE(  58)
53_1zp7A        34 --TLE-DDLNETNK(  8)VI-HKK( 28)TTDY-NGI(  4)YIDFEAKE( 12)HDHQIEHMK-QVK-AQD-----GICFVIIS(  67)
54_2pfjA        16 RSGLE-DKVSKQL-(  5)KF-EYE( 15)TPDF-LLP(  1)GIFVETAG(  5)DRKKHLLIR-EQH-PE------LDIRIVFS(  54)
55_3ajvA       102 -SLDT-ALVYFDLR(  4)KP-LVG(  2)RGTL-VYE(  2)GRVYEVLV(  7)KIGSLVEWS-RGA-SMD(  1)HSPIVAIV(  18)
56_2gw6A        25 QVYVA-FLVYLDLM(  6)EV-NCV(  8)CLVG-TEI(  4)LQTVVPTP(  5)SHNRIREIL-KAS-RKL(  7)MSFTLAIV(  13)
57_1azoA        55 GVLLE-IWLG----------A-SAG(  3)EQDF-AA-------LGVELKT( 30)-SHVRHKL----------------KRVLWI( 106)
58_1cw0A        21 -TAIE-KRLASLLT(  4)AF-RVQ(  6)RPDF-VVD(  2)RCVIFTHG( 28)VERDRRDIS-RLQ-ELG-----WRVLIVWE(  40)
59_1t0fA        63 -SDLE-LAVFLSLE(  6)DI-REQ( 29)STDF-LVD(  7)QFAIQVKP(  8)TLEKLELER-RYW-QQK(  1)IPWFIFTD( 108)
60_2okfA         5 RDVFH-EVVKTALK(  4)QITDDP( 10)SIDL-AAQ( 10)KIAVEVKS( 13)ALGQFINYR-GAL-RKV(  3)RVLYLAVP(  39)
61_15644542     37 GRFIH-EDYSD--------SV-MLP(  2)KIDT-MFE(  4)RVVGEVKK(  7)AEYQLLYYL-YRL-EEK(  3)ARGEIIVP(  54)
62_1w36B       956 HSLFE-DL------( 53)NK-QVE( 44)FIDL-VFR(  4)YYLLDYKS( 22)--RYDLQYQ-LYT-LAL( 17)GGVIYLFL(  37)
63_191174017    44 HALAE-KVLRN---( 58)LV-EQR( 14)TADS-III(  4)LQIHDLKY(  9)QNEQLQLYA-LGA-LEQ(  8)TVRLFIHQ( 244)
64_3fhdA       206 EEVVK-TLLATLLH(  3)TN-CLD( 14)SLDF-AAN( 14)CKVYEIKC(125)YFYQVLLQS-SIV-EEY( 12)PKYYIATG(  79)
65_3k93A        68 EEQAI-KLSGVTRG(  2)-L-KKN( 11)ECDI-YVP(  2)KLIIDTKC( 20)-AGYDIQMQ-GYM-WLW(  2)DQAQIDFV(  65)
66_3h4rA        51 HCRVL-E-------( 67)AG-HAE( 16)RPDK-IIP(  2)HWIMDVKT( 15)YHVQDAFYS-DGY-EAQ(  4)PTFVFLVA(  62)
67_18314555    121 HLARE-LELH----( 30)EG-HIR( 15)VIDE-LHY(  5)LELAELKT( 13)--KDCFQVS-LYK-YIF( 65)KIEYIHQE(  75)
68_2ostA         8 GDIAQ-QAAIMRAL(  4)GV-LKP(  5)SYDL-VFD(  4)LLKVQVKS( 27)-P--YRGND--------(  1)--DFAVAY(  58)
69_1dzfA        12 LWRAF-RTVKEMVK(  2)GY-FI-( 28)SFQA-NPT(  9)MGSLWVEF( 14)VIHIQEKN---------------FQTGIFV( 104)
70_3jsbA        54 SRCIE-IDSADKS-(  1)-C-THN( 23)IPDG-YKL(  4)LILLECFV( 10)FIEDTNKLA-CIR-EDL(  6)LVPIVDGR(  53)
71_2vldA       127 EAEMA-NLIFENP-(  7)KPIYRE(  6)IVDV-MGV(  5)IVVLELKR(  7)VSQMKRYVD-SLK-EE-(  4)VRGILVAP(  41)
72_2a1iA       120 ---SP-EEAGRYLE  19 SI-IVS( 22)IPDY-VLG(  1)STCALFLS( 22)----------AL------------RVLLVQ(  55)
73_2xi5A        35 HDYFG-RELCKSLN(  1)EY-RND( 24)TPDN-YLY(  4)LYIIDYKV(  6)SVITYDKYY-ELT-RDI(  7)IEIVIIRI(  52)
74_3hw3A        46 HLEVC-FMYS----( 25)RF-EII( 26)LPDL-YDY(  2)NRFIEIGV( 12)-ANKIKS----------(  1)-KTHIHIF( 108)
75_2w8mA        28 EYLVL-GKLLISLS( 12)YL----( 22)KYDL-TLY(  9)LLWIDITG( 19)LSVKVETAK-KYD----(  3)-RVFFIHY(  71)
76_302544184   356 GRAIE-PLVRESLA(  9)AA-PAI( 10)EIDL-VGA(  9)LFLGSIKW(  7)SHDLAALQK-HRA-AIT(  3)VPLVAVSR(  22)
77_3fqiA       208 GYKFE-QY-----------MC-A--( 36)EVDC-LNP(  8)SCYVELKT( 20)WWAQSFLPG--------------VPHVVAG( 105)
78_270013300   175 DLNCY-AYKRLLKL( 40)AN-YRI( 28)KLDG-CEE(  5)LYPVTVRP( 19)RDWTSLFFR-------------LYSFLYRA( 138)
79_55957531    222 GDFMH-KNTSTDFP( 29)SI-WSP(  6)KIDV-TVG( 12)IMPLELKT( 11)QVVLYTLLS-QER-RAD(  2)AGLLLYLK( 703)
80_307311134     7 STPHD-AVFKSFLR( 24)TL-KLE( 12)YSDL-LWS( 10)YVVIEHQS(  5)MAFRMMRYS-IAA-MQN( 11)VIPMLFYH( 172)
81_301794115    64 HLANK-ESLYHWLK(  5)QL-EYP(  6)IADV-FVN(  1)NLALEVQC(  4)QKVLKERSE-GYR-SQG-----YQVLWLLG( 182)
82_42781382    405 FLVYE-YYAFFIVI( 42)HV-AFN( 24)KPDI-RLD( 13)SIIIEVKY( 16)EQMYKYWSI-KFV-EEQ( 11)YEVICVYP(  50)
83_323335564   252 GHYLE-DFLTEKSL( 16)FI-VRS-----QLDA-YDP(  4)TGVFDLKT( 36)EYFELIRSA-LLK-YSL(  7)DGIFVAYH( 340)
84_17554522    335 DLATE-AFYINQNF( 30)TA-YKY( 20)ELDG-VIH( 11)KAFNEWDS( 27)SAKVAKWTL-QAL-LAG(  1)DTMKLGYV(  91)
85_15643411     83 DAINK-SKSVIMGQ( 12)RY-NPK(  7)PVDF-VVF( 10)IVFVEVKT(  7)TRERQVRDV-VEK-R-------EVYWEKLH(   4)
86_300898410    16 MNKTE-SAYCQHLE(  9)WY-RFE( 12)TPDF-AVM(  5)MELHEVKG(  6)RVKTKVAAD-QYP---------FRIIGVTV(  14)
87_14590363    261 SSSLE-REFASKIK(  5)DV-IYE( 11)IPDF-LIK(  4)KVYVEIVG(  6)LKRKLEKIS-KAK---------IQMLLIVN(  39)
88_226246523   922 FTSLQ-QAQDLVSC(  7)GV-CRH( 11)LPDL-VVW(  6)FKLVEVKG(  4)LSHKQMIWL-AEL-QKL(  2)EVEVCHVV(  10)
89_323974420   198 GNWGE-VVLTRVLE(  8)EY-ETQ( 11)QPDV-IVR(  5)DVVIDAKM( 14)YTRESALQE-HIA-SVR( 21)YVLMFIPV( 158)
90_149183698     9 ESDLY-KPIQKHFT(  4)KV-NGE(  2)DCDL-TAV(  4)LIIVELKL(  2)TVDLLIQAT-KRQ-RL------TDLVYIAI( 169)
91_14600962      6 PGEFE-DYLYRVIA(  7)RL-VVV(  9)SIDL-IVS(  5)VVLIKAIY(  5)SKKEVEELS-AVA-NSL(  1)VSAVIIAE( 241)
92_3dnxA        13 GQRLA-RGVARHLR(  4)VS-VEE(  7)RVDV-MGL(  5)IWVIECKS(  8)--AKWQGYL-EW-----(  1)DRYFWAVD(  66)
93_293418034    82 TLWAN-RLTKEAIL( 10)SS-LKS(  9)RIDF-MLQ(  5)DCYIEVKS( 18)GQKHLRELM-SVA-AEG-----QRAVIFFA(  55)
94_206976650    42 GYKGE-QAINYPLS( 14)LF-DQN(  3)QIDT-III(  4)LLFLEIKN( 31)VKRQEEQLR-KWL-SLN(  6)IESLVVIS( 164)
95_14601265    240 GILTE-EAFRSAMK( 10)VV-GRW(  1)HHDK-EGI( 18)HVLVEIKS(  9)MHRIGLLYE-RAE-GVK-----PRLVIVGG(  23)
96_3fovA        23 GLSAE-ASAADYLE(  4)RILARR(  6)EIDL-VAQ(  4)VAFVEVKA( 11)TPRQQSRIV-AAA-EAW( 11)LRFDAILI(  19)
97_1xmxA       251 GEWLE-TLVHSTVK( 10)DR-SLN( 14)ELDV-ATV(  4)LHIIECKT(  7)-DDTLYKLE-SLR-DLL(  4)ARAMLVSF(  41)
98_110643454   228 ESDFE-EALINHLM( 10)AF-VGR( 11)RVDL-LFF(  6)LLIVDLKV( 10)MNMYLNYAK-EHW-TLP(  4)PIGLVLCA(  51)
99_237716060   419 IREQE-LHYENVLF(  7)FY-TKV(  7)RIDL-VLQ(  4)IYVMEFKL(  3)AEEALKQIN-EKH-YSL(  7)KLFKVGVN(  15)
100_169833328  303 GHILE-NMVYLELR(  4)QV-YVG(  5)EVDF-VVV(  5)YAYYQVSE(  7)LERELRPLD-AIK-DQF-----PKYLLTMD(  23)
101_67458629   122 DKITE-EQVCDTLA( 11)MP-IKH( 15)HPDW-VGM( 24)LWSFEAKT(  6)LREYFFQAL-SNS-SWA-----HYGYLVAA( 102)
102_75906392     4 DSIFY-KLFKQFPG( 17)FE-SVE(  6)RIDG-VFL(  9)VFFAEVQF(  2)DDDLYHRFF-SEL-FLF( 10)WFGVIIYA( 187)
103_3ijmA       29 HQSVI-ARLTAGLY( 10)EP-LPE( 10)VPDV-LLY(  7)KVIIEVCQ(  3)LKHDTSKIV-KLI-EDN(  5)EGFVFNYK(  32)
104_289650334  170 GRYYE-RLWQFAVQ(  5)EI-IAA( 12)ELDM-LLR(  4)VHHVELAI( 25)LDRKLAHLS-QHQ-LPM( 19)AQLWLGGY( 125)
105_139472820   24 GNRAH-LATYRRLL( 25)LF-FEV(  5)IADC-VIL(  8)CYVVELKT( 19)GLCQLSDSI-HYI-AHS(  9)ITPLLIFK( 112)
106_62318767   140 KDELR-KIVVDELS(  4)DS-SIC( 16)YIDV-IVN(  2)RLIIVIDF(  4)EIARQTSGY-KEL-LQS-----LPLIFVGK( 120)
107_1wdjA       45 SGRRS-LQLAYQLA(  9)VV-FDS( 12)SPDA-AFV( 21)KAVFEVRS(  4)PEELRAKMG-IYL-RNG(  1)LLGVLVDP(  40)
108_1y0kA       26 KWDRE-RWVCQRFL( 10)DF-AAP(  3)PPDV-LFK-----GAGFEVFF( 50)LRKKAHNYS-ERG----(  4)-ELDLLAF(  65)
109_94994509   205 GESLE-LYAETEFN( 10)SF-VKD(  8)KGDY-IYR( 10)SIMFEMKN(  3)TTKTKHKNS-DFF-KEL(  9)EYAVLVSM( 146)
110_157412147   54 MTKAE-MTFYLSLI(  7)TV-MVK( 26)HVDY-LLC(  7)LLAIELDD(  8)IARDKLVDG-IFK-AVG-----LPVIHHPV(  21)
111_109730533  533 DSQLY-RYVKIGLT(  5)RL-YFA(  9)TIDV-EIK( 18)RIALCIDG( 12)GKEAIKQRH-LQL-LG------YQVVQIPY(  31)
112_3c0uA       41 RMMLR-LLAWLKYA(  2)RL-QFT(  7)EPEA-WLR(  6)DLWIELGL(  1)DERRIKKAC-TQA---------AEVALFTY(  74)
113_331270817   24 PEILI-SFLNAVMK(  8)KI-RNS( 12)RLDI-KAT(  5)EVNIEIQV(  6)IRRSLYYWS-KMY-EGQ( 12)TVCINILD( 175)
114_260591359   33 EKVYK-NAMVIELR(  3)LN-YET( 15)RADI-IVD(  1)RLILELKA(  4)SIAHEVQLV-NYL-TST(  3)DGLLINFG(  15)
115_237749995   62 DVFLE-EFLQCVGL( 10)RV-YKE(  2)NIDV-YIT(  4)HIIVENKI(  4)LRQIERYIE-SIA-KEQ( 23)NIAVLYLA( 262)
116_18466595    25 EQIYQ-TYLEENTM( 21)PL-SSD(  1)KPDF-VYL(  8)VVLVEIEK( 20)LQQMNTWRA-WFD-DES( 25)FKYVLVHG( 112)
117_187934185  302 GAIGE-CIILKNEK( 16)DI-EWS(  8)GYDI-KSF(  8)NKYIEVKT( 10)TVNEVNKSKTLNTN--------GLYVLARV(  34)
118_309702132  153 GVIGE-LTCLERLI(  4)AV-EAW(  5)LQDF-ELD-----ERAIEVKS(  9)TIHALEQLD-WQR-PG-------SLMLCGL( 112)
119_29826532   105 HTLTV-AETALAFL( 11)PL-DWI( 13)IPDA-LLY( 14)RAFVEVDR(  4)PERLAAKLP-AYE-RLY( 29)RVLFVLDG(  68)
120_258544132   21 RVLSE-HWLKANVP(  6)TL-QKT(  5)VKDF-ICP(  2)QENYELKS( 14)YNTMLARVQ-AED-NP------NFFLLAYQ( 150)
121_2czrA        8 GTKKV-YTQVRYLD(  2)HW-EIE(  1)-STI-TGI(  6)KVVIDVAK(  4)ADSLAG--K-EVN---------GIHIVAIP( 148)
```

Figure 2. Multiple sequence alignment for the conserved core regions of the PD-(D/E)XK superfamily. Each group of closely related Pfam, COG, KOG families and PDB90 structures (detectable with PSI-BLAST) is represented by available PDB90 sequence or selected representative if the cluster does not contain solved structure. Sequences are labeled according to the group number followed by NCBI gene identification number or PDB code. The first residue numbers are indicated before each sequence, while the numbers of excluded residues are specified in parentheses. Sequence given in italic corresponds to circularly permuted α-helix. Residue conservation is denoted with the following scheme: uncharged, highlighted in yellow; polar, highlighted in grey; active site PD-(D/E)XK signature residues, highlighted in black; other conserved polar/charged residues augmenting the active site, highlighted in red. Locations of secondary structure elements are shown above the corresponding alignment blocks.

structures. The final alignment for PD-(D/E)XK superfamily was assembled from sequence-to-structure mappings using a consensus alignment and 3D assessment approach (44). The collected PD-(D/E)XK fold proteins were clustered into groups of closely related families and structures based on detectable sequence similarity with both PSI-BLAST and RPS-BLAST. Structure similarity based searches were performed with ProSMoS program (45). Domain architecture was analyzed with RPS-BLAST against COG, KOG and Pfam, and with HMMER3 against Pfam. Transmembrane regions were detected with a TMHMM server (46). Cellular localization for prokaryotic sequences was predicted with PSORTb (47) and for eukaryotic with Cello (48), WoLF PSORT (49) and Multiloc (50). Taxonomic assignment was based on NCBI taxonomic identifiers. HGT events were identified using a phylogenetic approach. Phylogenetic trees for each cluster were calculated using PhyML. The genomic context was analyzed with The SEED (51), GeContII (52), MicrobesOnline (53) and NCBI genomic resources. Clustering of all 21 911 sequences was performed with CLANS (54), with high resolution figures drawn with an in-house script based on CLANS scores.

RESULTS

In order to broaden the repertoire of PD-(D/E)XK proteins we performed sensitive distant homology searches using as the initial dataset 44 Pfam 25 families and 60 representative restriction endonuclease-like proteins of known structure cataloged in SCOP database. The exhaustive, transitive fold recognition searches against Pfam, COG, KOG and PDB90 databases resulted in a collection of various PD-(D/E)XK families that altogether span 21 911 sequences from the NCBI nr protein database (a list of all identified proteins is provided as Supplementary Dataset S1). For instance, we found that 99 PDB90 structures, 49 COG, 11 KOG and 118 Pfam families retain the PD-(D/E)XK fold. This is significantly more than the currently reported in Pfam 26 database in PD-(D/E)XK nuclease superfamily clan which defines only 60 families. In addition, we found six PD-(D/E)XK fold families to be classified also in two other Pfam clans: (i) Restriction endonuclease-like (Endonuc-FokI_C, PF09254; MutH, PF02976; RE_AlwI, PF09491) and (ii) tRNA–intron endonuclease catalytic domain-like (Sen15, PF09631; tRNA_iecd, PF12858; tRNA_int_endo, PF01974).

All PD-(D/E)XK proteins were identified with a single procedure as described in our previous work (36). This exemplifies a major progress in comparison with previous studies on the diversity of PD-(D/E)XK phosphodiesterase superfamily. All collected families and structures were clustered into 121 groups of closely related proteins. The average sequence similarity between different PD-(D/E)XK groups is very low, which is reflected by low Meta-BASIC scores (Supplementary Table S1) and is below the confident recognition both with standard and even more advanced sequence comparison methods. This high sequence divergence implies the need for complex sequence and structure search strategies. Many of the identified protein groups contain uncharacterized and poorly annotated proteins or functionally studied proteins without structural annotations. Eventually, upon further manual literature inspection, the majority of these families were linked to the PD-(D/E)XK superfamily. However, such an assignment was feasible with a list of proteins in question. The remaining 11 identified groups embrace the newly found PD-(D/E)XK fold families.

We detected PD-(D/E)XK sequences in multiple genomes from all forms of life. The versatility of this superfamily convinced us to perform a variety of structure- and sequence-based analyses. We thoroughly examined every family in our dataset in order to determine its characteristic sequence and structure features. Here, we describe in detail the results of sequence and literature searches, domain architecture analysis, structural comparisons and phylogenetic inference, that eventually shed new light on functional diversity of PD-(D/E)XK proteins. Table 1 summarizes the details of all identified PD-(D/E)XK phosphodiesterase groups. Human genes encoding PD-(D/E)XK proteins are shown in Supplementary Table S2. One should note that most of the human PD-(D/E)XK genes are involved in diseases.

Newly identified PD-(D/E)XK families

According to extensive database and literature searches 11 groups (3, 5, 14, 15, 26, 90, 91, 116, 117, 118, 119; Table 1) include proteins not annotated previously to PD-(D/E)XK fold superfamily. Five of them embrace completely uncharacterized proteins from DUF4420 (PF14390), DUF3883 (PF13020), DUF4263 (PF14082), COG5482 and COG1395 families. The remaining six newly detected groups cover functionally studied protein families which, however, lacked fold assignment. These include restriction endonucleases Tsp45I (PF06300), HaeII (PF09554), Eco47II (PF09553), ScaI (PF09569) and HpaII (PF09561) and Replic_Relax (PF13814)—a predicted transcriptional regulator. We studied in detail all newly detected families to hint at additional functional information. COG1395, COG5482 and Replic_Relax (PF13814) usually occur in a fusion with HTH DNA-binding domains, which suggests their role in transcription regulation. DUF4263 (PF14082) and DUF3883 (PF13020) are often present in proteins encoding an ATPase domain. Additionally, DUF3883 appears in a variety of domain architectures, including fusions with helicases, TF domains, protein kinases and MTases. Details of identification of new families are summarized in Supplementary Table S3. One should note that only two of them were assigned to the PD-(D/E)XK superfamily with Meta-BASIC scores above confidence threshold of 40.

Structure analysis

A comprehensive analysis of the identified structures allows us to better understand how the PD-(D/E)XK fold adapt to particular functions. The structural analyses are critical to further detection and classification

Table 1. One hundred and twenty-one groups of proteins retaining PD-(D/E)XK nuclease fold

No.	Name	Pfam, COG/KOG, PDB90 structure	Reference to fold assignment	Biological function	Taxonomy					HGTs
					Viruses	Bacteria	Archaea	Eukaryota	Detailed distribution	
1	NaeI	PF09126 1ev7	(58)	Type II Restriction Endonuclease (58)		+			Bacteria *(α-proteobacteria, Actinobacteria)*	*Deinococcus maricopensis* sequence is found in a clade with Roseobacteriales (α-proteobacteria) & Actinomycetales. The Roseobacteriales clade locates within a Actinomycetales tree.
2	BglI	1dmu	(59)	Type II Restriction Endonuclease (59)		+			Bacteria	Only four sequences from distant taxa: *Bacillus atrophaeus* (Bacilli), *Microcoleus* (Oscillatoriales), *Deinococcus deserti* (Deinococci) suggest a HGT.
3	HpaII	PF09561	New	Type II Restriction Endonuclease (60)		+			Bacteria (*Bacillus/ Clostridium, Bacteroidetes*)	*Streotibacillus moniliformis* (Fusobacteriales) forms a clade with *Sulfurimonas denitrificans* (Campylobacteriales). *Bacillus thuringiensis* (Bacillales) groups with *Flexibacter tractuosus* (Cytophagales). Single sequences of Fusobacteria, ε-proteobacteria, β-proteobacteria and γ-proteobacteria.
4	NgoBV, NlaIV	PF09564	(1)	Type II Restriction Endonuclease (61)		+			Bacteria *(mostly Neisseria)*	Multiple transfers, animal related bacteria. Single representatives of: Spirochaetes, Fusobacteria, Tenericutes, ε-proteobacteria, Clostridia, Bacilli.
5	ScaI	PF09569	New	Type II Restriction Endonuclease (62)		+			Bacteria	Multiple transfers. Ecologically and taxonomically unrelated bacteria from Bacilli, Proteobacteria, Cyanobacteria, Bacterioidetes.
6	LlaMI, ScrFI	PF09562	(63)	Type II Restriction Endonuclease (63)		+			Bacteria (*Cyanobacteria, Bacillus/ Clostridium, γ-proteobacteria)*	One clade grouping: Lachnospiraceae bacterium (Clostridiales), *Lactococcus lactis* subsp. *cremoris* (Lactobacillales), *Prochlorococcus marinus* (Cyanobacteria), *Vibrio parahaemolyticus* (γ-proteobacteria).
7	PvuII	PF09225 3ksk	(64)	Type II Restriction Endonuclease (64)		+			Bacteria	*Meiothermus ruber* (Thermales), *Bacteroides cellulosilyticus* (Bacteroidales) and *Arthrospira maxima* (Burkholderiales) are single representatives of corresponding taxa suggesting a transfer event from Enterobacteriales.
8	XamI	PF09572	(11)	Type II Restriction Endonuclease (65)		+	{1}		Bacteria	Patchy distribution including a Haloarcheon—*Halogeometricum borinquense* grouping with good support within a bacterial clade.
9	XhoI	PF04555	(1)	Type II Restriction Endonuclease (66)		+	{1}		Bacteria (*mostly Proteobacteria and Actinobacteria*)	*Leptospirillum rubarum* and 3 Actinobacteria within a Proteobacteria clade.
10	ApaLI	PF09499	(67)	Type II Restriction Endonuclease (62)		+			Bacteria	Multiple transfers, *Helicobacter felis* (ε-proteobacteria) with *Microscilla marina* (Bacterioidetes). Patchy distribution including single sequences from Bacillales, Chloroflexales, Xantomonadales, Fusobacteriales, Beggiatoales, Borrelomycetales, Campylobacteriales.

(continued)

Table 1. Continued

No.	Name	Pfam, COG/KOG, PDB90 structure	Reference to fold assignment	Biological function	Taxonomy					HGTs
					Viruses	Bacteria	Archaea	Eukaryota	Detailed distribution	
11	BamHI	PF02923 1bam, 3odh	(68)	Type II Restriction Endonuclease (68)		+			Bacteria	Multiple transfers, extremophilic and/or aquatic bacteria.
12	BstYI, BglII	PF09195 1sdo, 1d2i	(69)	Type II Restriction Endonuclease (69)	{1}	+			Bacteria	Multiple transfers for example *B. subtilis* sequence grouped with Cyanobacteria. *Ethanoligenens harbinense (Clostridiales)* is located in a Proteobacteria clade.
13	SacI	PF09566	(1)	Type II Restriction Endonuclease (70)		+			Bacteria (*Bacilli*)	Multiple transfers. Patchy distribution: single sequences Bacteroides, Actinobacteria, γ-proteobacteria, ε-proteobacteria.
14	Eco47II	PF09553	New	Type II Restriction Endonuclease (71)	{1}	+			Bacteria	*Helicobacter pylori* sequence groups within a Mycoplasma clade, multiple transfers.
15	HaeII	PF09554	New	Type II Restriction Endonuclease (72)		+			Bacteria *(mostly γ- and β-proteobacteria)*	Cyanobacteria sequences not grouped. Single sequences from Cyanobacteria, Bacterioidetes.
16	HindIII	PF09518 3a4k	(73)	Type II Restriction Endonuclease (73)		+			Bacteria *(mostly γ-proteobacteria)*	Multiple transfers: *Citrobacter* (*γ-proteobacteria)* within a Bacilli clade, oral bacterium *Streptococcus downei* grouped together with *Haemophilus influenzae*.
17	FokI	PF09254 2fok	(14)	Type II Restriction Endonuclease (14)		+			Bacteria *(Bacillus/ Clostridium)*	*Haemophilus influenzae* within a *Streptococcus sanguinis* clade.
18	EcoO109I	1wtd	(74)	Type II Restriction Endonuclease (74)		+			Bacteria *(Escherichia coli)*	No HGT observed
19	EcoRV	PF09233 1eo3	(75)	Type II Restriction Endonuclease (75)		+	{2}		Bacteria	*Escherichia coli* in a clade with *Streptococcus mitis* (Lactobacillales), *Listeria innocua* (Bacillales), *Vibrio orientalis* (Vibrionales) and Thiomonas (Burkholderiales)[a]
20	EcoRI	PF02963 2oxv	(76)	Type II Restriction Endonuclease (76)		+	{1}		Bacteria (*BCF group,* Proteobacteria*, Bacillus/ Clostridium)*	*Methanobrevibacter smithii*, *Staphylococcus aureus*, *Fusobacterium ulcerans* and *Brucella melitensis* group together with 5 *E. coli* Migula 1895 sequences. Multiple transfers
21	XcyI	PF09571	(77)	Type II Restriction Endonuclease (77)		+			Bacteria *(γ-proteobacteria, Clostridium)*	*Pseudomonas alcaligenes* (soil bacterium) in a plant pathogenic Xanthomonas clade, Proteobacteria in a extremophilic Clostridium clade. Multiple transfers
22	BsoBI	PF09194 1dc1	(78)	Type II Restriction Endonuclease (78)		+			Bacteria *(mostly Cyanobacteria)*	*Roseiflexus castenholzii* phototrophic bacterium and intestinal *Alistipes* sp. within a mostly Cyanobacteria clade
23	HincII	PF09226 3ebc	(79)	Type II Restriction Endonuclease (79)		+			Bacteria (*mostly γ-proteobacteria)*	Oral bacterium *Capnocytophaga ochracea* within a Haemophilus & Actinobacillus clade. Additionally, *Prevotella bivia* pathogen, joins this clade
24	SinI, AvaII	PF09570	(1)	Type II Restriction Endonuclease (22)		+			Bacteria	Patchy distribution

(continued)

Table 1. Continued

No.	Name	Pfam, COG/KOG, PDB90 structure	Reference to fold assignment	Biological function	Taxonomy					HGTs
					Viruses	Bacteria	Archaea	Eukaryota	Detailed distribution	
25	NgoPII	PF09521	(1)	Type II Restriction Endonuclease (80)		+	+		Prokaryota	Patchy distribution, possible transfer between *Desulfurobacterium thermolithotrophicum* (Aquificiae) and *Methanothermobacter thermautotrophicus* and Candidatus *Parvarchaeum acidiphilum* (Euryarchaeota)
26	Tsp45I	PF06300	New	Type II Restriction Endonuclease (81)		+			Bacteria	Possible transfer between *Simonsiella muelleri* (β-proteobacteria) and *Fusobacterium periodonticum* (Fusobacteria). Patchy distribution including: *Prevotella*, *Treponema* and *Chlorobium*
27	MspI	PF09208 1sa3	(82)	Type II Restriction Endonuclease (82)		+			Bacteria *(mostly Bacilli/Clostridia)*	Two γ-proteobacteria (*Idiomarina loihiensis*, *Moraxella*) within a Firmicutes clade. *Moraxella* opportunistic pathogen groups with *Clostridium botulinum*. Deep sea *I. loihiensis* groups with *Anoxybacillus flavithermus* thermophile. Patchy distribution
28	MjaII	PF09520	(11)	Type II Restriction Endonuclease (83)		+	+		Prokaryota	Possible transfer between Archaea and Bacteria. Patchy distribution
29	MunI	PF11407 1d02	(83)	Type II Restriction Endonuclease (83)		+	{1}		Bacteria	*Desulfurivibrio alkaliphilus* and *Prevotella copri* prossible transfer. *Cenarchaeum symbiosum* groups together with Tenericutes and Clostridia. *Cenarchaeum symbiosum* is a partner of a marine sponge (84)
30	CfrBI	PF09516	(1)	Type II Restriction Endonuclease (85)		+			Bacteria (*mostly proteobacteria*)	Anaerobic ammonium-oxidizing candidatus *Kuenenia stuttgartiensis*, thermophilic *Geobacillus stearothermophilus* and *Thermodesulfovibrio yellowstonii* group within a Proteobacteria tree
31	NgoMIV	PF09015 1fiu	(85)	Type II Restriction Endonuclease (85)		+			Bacteria	*Bacteroides finegoldii* groups within *Heliobacterium modesticaldum* and *Faecalibacterium prausnitzii* (Clostridiales) clade. *Thermomonospora curvata* (Actinomycetaceae), *Opitutaceae bacterium* TAV2 (Opitutaceae) and *Idiomarina baltica* (Alteromionadaceae) group together
32	Cfr10I, Bse634I, SgrAI	PF07832 1cfr, 1knv 3dpg	(86)	Type II Restriction Endonuclease (86)		+			Bacteria	*Pseudomonas stutzeri* (Pseudomonadales), *Nodularia spumigena* (Nostocales) and *Streptomyces griseus* (Actinomycetales) sequences group together
33	Bpu10I	PF09549	(87)	Type II Restriction Endonuclease (87)		+			Bacteria	Multiple transfer events. One clade encompasses representatives of Cyanobacteria (Cyanothece and Nodularia), Proteobacteria (*E. coli*, *Allochromatium vinosum*, *Plesiocystis pacifica*), Chloroflexi (*Chloroflexus aurantiacus*) and Actinobacteria (*Gardnerella vaginalis*)

(continued)

Table 1. Continued

No.	Name	Pfam, COG/KOG, PDB90 structure	Reference to fold assignment	Biological function	Taxonomy					HGTs
					Viruses	Bacteria	Archaea	Eukaryota	Detailed distribution	
34	BspD6I, AlwI, MlyI	PF09491 2ewf, 2p14	(88)	Type II Restriction Endonuclease Restriction Endonuclease (88)		+	{1}		Bacteria	*Micrococcus lylae* (Actinomycetales) and *Methanohalobium evestigatum* (Euryarchaeota) forming a common clade or *Mannheimia haemolytica* (γ-proteobacteria) within a Firmicutes clade are examples of possible HGT. *M. haemolytica* causes intramammary infection in sheep. *Micrococcus lylae* is a denitrifying soil bacterium whereas *M. evestigatum* is an extreme halophilic methanogen
35	LlaJI, McrBC	PF09563 PF10117 COG4268	(89)	Type II Restriction Endonuclease (89)		+	+	{1}	Prokaryota	*Mobiluncus curtisii* subsp. *curtisii* (Actinomycetales) within a Clostridium clade. *Gardnerella vaginalis* (Actinomycetales) forms a clade with *L. lactis* (Lactobacillales) and *Anaerostipes caccae* (Clostridiales). *Batrachochytrium dendrobatidis* JAM81 (Chytrydiomycota, Fungi) forms a clade with *Desulfotomaculum nigrificans* (Clostridiales). *Methanobrevibacter ruminantium* DSM 1093 (Euryarchaeota) locates in a mostly Firmicutes clade
36	SdaI, BsuBI	PF06616 2ixs	(90)	Type II Restriction Endonuclease (90)		+	{1}		Bacteria	*Treponema vincentii* (Spirochaetales), *B. subtilis* and *Paenibacillus larvae* subsp. *larvae* (Bacillales) within a Proteobacteria clade. *Shewanella sediminis* (Enterobacteriales) sequence groups with *Clostridium sticklandii* (Clostridiales). *Methanobrevibacter ruminantium* (Euryarchaeota) forms a clade with 2 *Prevotella* (Bacteroidales) sequences. *Methanobrevibacter ruminantium* is a rumen bacterium of cattle and *Prevotella* is involved in periodontal infections
37	DpnII, MboI	PF04556	(91)	Type II Restriction Endonuclease (91)		+	+		Prokaryota	*Carboxydothermus hydrogeniformans* in a Mycoplasma clade. Extremophilic *Dictyoglomus thermophilum* (Dictyoglomi) with *M. smithii* & *Methanosphaera stadtmanae* (Euryarchaeota)
38	Ecl18kI, EcoRII, PspGI	PF09019 2fqz, 1na6 3bm3	(92)	Type II Restriction Endonuclease (92)	{2}	+	{1}		Bacteria	*Photobacterium damselae* subsp. *piscicida* (Vibrionales) sequence locates within an Enterobacteriaceae clade (*Klebsiella*, *Shigella*, *Escherichia* and *Yersinia*)
39	HinP1I	PF11463 1ynm	(93)	Type II Restriction Endonuclease (93)		+			Bacteria *(Proteobacteria)*	*Leptotrichia goodfellowii* (Fusobacteriales) in a Proteobacteria clade. *Moraxella catarrhalis* (Pseudomonadaceae) in a Haemophilus clade (Pasteruellaceae). *Haemophilus somnous* is a bovine pathogen, *L. goodfellowii* is found in dental plaque. *Moraxella catarrhalis* was recently described as a respiratory pathogen

(continued)

Table 1. Continued

No.	Name	Pfam, COG/KOG, PDB90 structure	Reference to fold assignment	Biological function	Taxonomy					HGTs
					Viruses	Bacteria	Archaea	Eukaryota	Detailed distribution	
40	NotI	PF12183 3bvq	(94)	Type II Restriction Endonuclease (94)		+			Bacteria	*Desulfobacterium* sp. (Deltaproteobacteria) and *Syntrophomonas wolfei* (Clostridiales) in a green sulfur bacteria *Chlorobium phaeobacteroides* clade
41	Bsp6I	PF09504	(95)	Type II Restriction Endonuclease (95)	{1}	+			Bacteria	*Fusobacterium nucleatum* (Fusobacteria) sequence localizes in a Ureaplasma/Mycoplasma (Borrellomycetales) clade
42	HindVP, HgiDI, BsaHI	PF09519	(96)	Type II Restriction Endonuclease (96)		+			Bacteria	Patchy taxonomic distribution[a]
43	MjaI	PF09568	(67)	Type II Restriction Endonuclease	{1}	+	+		Prokaryota	*Methanothermobacter thermautotrophicus* within a BCF group clade
44	TaqI	PF09573	(97)	Type II Restriction Endonuclease (97)		+			Bacteria *(Thermus, Aquficae, Nitrospirae)*	*Thermodesulfovibrio yellowstonii* (Nitrospirae) in a *Hydrogenivirga* sp. (Aquificae) clade
45	SfiI	PF11487 2ezv	(98)	Type II Restriction Endonuclease (98)		+			Bacteria	No HGT observed, the phylogeny could not be resolved with reliable confidence
46	MvaI, BcnI	2odh, 2oa9	(99)	Type II Restriction Endonuclease (99)		+	{2}		Bacteria	*Thermoplasma volcanium* (Euryarchaeota) within mixed bacterial clades
47	ThaI	3ndh	(100)	Type II Restriction Endonuclease (100)			+		Archaea *(Thermoplasmata)*	No HGT observed
48	HSDR_N, HSDR_N_2, EcoR124I	PF04313 PF13588 COG4748 COG2810 COG0610 2w00, 3h1t	(101)	Type I Restriction Endonuclease (101); EcoR124I cleaves DNA at a location distant from specific recognition site (102). Type IV Restriction Endonuclease (predicted, found mostly in Archaea)		+	+	{1}	Prokaryota	*Simonsiella muelleri* (β-proteobacteria) in a *H. influenzae* (γ-proteobacteria) clade. A single sequence from *Vibrio splendidus* (Vibrionales) locates in an *Actinobacillus pleuropneumoniae* & *Haemophilus parasuis* (Pastereullaceae) clade
49	HindVIP, EcoPI	COG4889 COG4096 COG3421 COG3587 3s1s	(103)	Type I Restriction Endonuclease Type II Restriction Endonuclease (104) Type III Restriction Endonuclease (103) *Broad sequence and function profile due to wide, multidomain definitions of COG entities*	+	+	+		Prokaryota & phages	*Lactobacillus helveticus* (Lactobacillales) and *Pseudomonas stutzeri* (Pseudomonadales) form a perfectly supported group. Phylogeny is not well resolved[a]
50	Mrr_cat, DUF2034	PF04471 PF10356 COG4127 COG1715 COG1787 1y88	(105)	Mrr restriction endonuclease (Methylated adenine recognition and restriction) restricts both adenine- and cytosine-methylated DNA (106). DUF2034 function is unknown.	{2}	+	+	+	Eukaryota (*without Plantae*) Bacteria Archaea Phages	No HGT observed, the phylogeny could not be resolved with reliable confidence

(continued)

Table 1. Continued

No.	Name	Pfam, COG/KOG, PDB90 structure	Reference to fold assignment	Biological function	Taxonomy					HGTs
					Viruses	Bacteria	Archaea	Eukaryota	Detailed distribution	
51	Archaeal HJC	PF01870 COG1591 1hh1, 1gef 1ob8, 2wcw 2eo0	(24)	HJC resolvase (107)	+	+	+		Prokaryota *(mostly Archaea)* & Archaeal phages	A handful of unrelated bacteria: *Fusobacterium nucleatum* subsp. *polymorphum*, *Fusobacterium* sp., *Hydrogenobaculum* sp., *Rhizobium leguminosarum* bv. *viciae*, *Ralstonia solanacearum*, *E. coli* TA206, *Nitratiruptor* sp. and *Synechococcus* sp. form a clade within the Archeal tree
52	ERCC4, XPF, Mus81	PF02732 KOG0442 KOG2379 COG1948 1j22, 2bgw 2ziu, 2zix 2ziv	(9)	HJC resolvase (108) DNA repair, structure specific endonuclease			+	+	Archaea & Eukaryota	No HGT observed
53	RecU, HJC Resolvase, Penicillin-binding protein-related factor A	PF03838 COG3331 1zp7, 1y1o	(24)	HJC resolvase (109). The genomic context is well conserved and includes a penicilin-binding protein, a methylase and HhH domain containing proteins. Penicillin-binding proteins are involved in cell-wall biosynthesis.		+			Bacteria *(Bacillus/ Clostridium)*	*Catonella morbi* (Clostridiales) in a Lactobacillales clade. *Acholeplasma laidlawii* (Tenericutes) in a Bacillus clade
54	Bacteriophage T7 endonuclease I, Phage_endo_I	PF05367 2pfj	(110)	HJC resolvase (110)	+	+	+		Prokaryota & phages	*Halanaerobium hydrogeniformans* (Firmicutes) locates with *Dehalococcoides* sp. and *Thermomicrobium roseum* (Chloroflexi). Patchy distribution suggesting multiple transfers. Phages group with their hosts
55	tRNA intron endonuclease	PF01974 KOG4133 KOG4685 COG1676 1a79, 2cv8, 2gjw 2zyz, 2ohe, 3iey 3if0, 3ajv, 3p1y	(17)	tRNA intron endonuclease, in the proximity of various tRNA synthases in archaeal genomes.			+	+	Archaea & Eukaryota	No HGT observed
56	Sen15	PF09631 PF12858 2gw6	(111)	A structural subunit of eukaryotic tRNA intron endonuclease (111)				+	Eukaryota *(Ophisthokonta, Amoebozoa)*	No HGT observed
57	MutH	PF02976 COG3066 1azo, 2aoq	(6)	Mismatch repairing enzyme (6). MutH cleaves a newly synthesized and unmethylated daughter strand 5′ to the sequence d(GATC) in a hemi-methylated duplex.		+			Bacteria *(γ-proteobacteria)*	*Plautia stali* symbiont (unclassified bacterium) in a γ-proteobacteria clade

(continued)

Table 1. Continued

No.	Name	Pfam, COG/KOG, PDB90 structure	Reference to fold assignment	Biological function	Taxonomy					HGTs
					Viruses	Bacteria	Archaea	Eukaryota	Detailed distribution	
58	VSR, DUF559, DUF2726	PF04480 PF03852 COG3727 COG2852 1cw0, 3hrl, 3r3p	(112)	Very short patch repair (Vsr) endonuclease that specifically removes T/G mismatches in DNA sequences targeted to cytosine methyltransferase (112). Group I intron homing endonuclease (113)	{1}	+	+		Prokaryota	No HGT observed
59	TnsA	PF08722 1t0f	(114)	Transposase (114)		+		{1}	Bacteria	*Ricinus communis* and *Vibrio harvei* form a clade, might be a long branch attraction phenomenon. *Deinococcus proteolyticus* in a Proteobacteria clade. Mixed clades containing: Bacilli, Chloroflexi, Cyanobacteria and Proteobacteria
60	XisH	PF08814 2inb, 2okf	Pfam	fdxN element excision controlling factor (115)		+			Bacteria *(mostly Cyanobacteria)*	*Herpetosiphon aurantiacus* in a Cyanobacteria clade. *Beggiatoa* sp. (γ-proteobacteria) in a Cyanobacteria clade[a]
61	DUF83, Cas_Cas4	PF01930 COG1468 COG2251	(5)	Cas1 protein (YgbT) has nuclease activity against single-stranded and branched DNAs including HJC, replication forks and 5′-flaps (116).		+	+	{1}	Prokaryota	Not resolved phylogeny. *Aureococcus anophagefferens* (Stramenopile, Eukaryota) sequence is localized in a mixed Bacteria clade. *Aureococcus anophagefferens* causes algal blooms. Planctomycetes are isolated from marine water
62	RecBCD, Exonuclease V	PF04257 COG1330 COG3857 COG1074 1w36	(16)	Exonuclease/helicase, a component of the RecBCD complex that handles double-strand breaks (DSB) (16). RecB alone has a weak helicase activity (117) and its nuclease domain generates single-strand regions at the ends of DSBs (5).		+		{1}	Bacteria *(Clostridium/ Bacillus, Chlorobiales, γ-proteobacteria)*	*Oryza sativa* protein groups in an Enterobacteriaceae clade within a Serratia proteins
63	DUF2800, PDDEXK_1	PF10926 PF12705 COG2887	(118)	RecB-like, probable prophage proteins	+	+			Bacteria phages	*Dehalococcoides ethernogenes* (Chloroflexi) sequence resides in a Clostridiales clade
64	Viral alkaline exonuclease	PF01771 2w45, 3fhd	(30)	Exonuclease processing viral genome during recombination (4). The enzyme displays RNase activity used in mRNA degradation pathways (4).	+				Herpesvirales	No HGT observed

(continued)

Table 1. Continued

No.	Name	Pfam, COG/KOG, PDB90 structure	Reference to fold assignment	Biological function	Taxonomy					HGTs
					Viruses	Bacteria	Archaea	Eukaryota	Detailed distribution	
65	YqaJ, lambda-exonuclease	PF09588 COG5377 1avq, 3k93 3slp	(31)	Exonuclease facilitating phage DNA recombination (31). The λ exonuclease is an ATP-independent enzyme that binds to dsDNA ends and processively digests the 5′-ended strand to form 5′-mononucleotides and a long 3′-overhang (119).	+	+		+	Bacteria Eukaryota phages	No HGT observed
66	RecE, DUF3799	PF12684 3h4r, 3l0a	(120)	Exonuclease from RecET recombination system (120)	+	+			Bacteria phage	No HGT observed, the phylogeny could not be resolved with reliable confidence
67	DEM1, EXO5	PF09810 KOG4760	Pfam	Mitochondrial, single-strand-specific 5′-exonuclease releasing dinucleotides as the main products of catalysis. EXO5 binds to 5′-RNA termini of chimeric DNA–RNA molecules and, after sliding across the RNA substrate, cuts the DNA 2 nt from the RNA–DNA junction (121).	{1}	+	+	+	Archaea *(Euryarchaeota)* Eukaryota	*Methanocella paludicola* in a Actinobacteria clade. *Methanocella paludicola* is a methanogen isolated form rice paddy soil. *Eubacterium eligens* (Clostridiales) in an Ascomycota clade (very long branch)
68	ssp6803i	PF11645 2ost	(122)	Homing endonuclease with a specificity profile extending over a long (17-bp) target site (122)		+	+		Prokaryota	Patchy distribution including 5 Haloarcheales and 2 Ktedonobacter sequences as well as Bacillus forming a sister clade to 5 sequences Cyanobacteria suggest a HGT history
69	Rpb5 N-terminal domain	PF03871 KOG3218 1dzf, 3h0g	(8)	RNA Polymerase (8). It may hold together the Rpb1-β24/25 and Rpb1-α44/47-fold of RNA polymerase II, or their counterparts in the archaeal, viral and RNA polymerase I and III enzymes (123).				+	Eukaryota	No HGT observed
70	Arenavirus RNA polymerase N-terminal domain, virus L-Protein	PF06317 3jsb	(124)	RNA Polymerase N-terminal domain that utilizes ‘cap snatching’ mechanism for viral mRNA transcription (125). *Similar to groups 73 and 74*	+				Arenavirus	No HGT observed
71	RecB, DUF91	PF01939 COG1637 2vld	(126)	DNA endonuclease specialized in cleavage at double-stranded DNA (dsDNA)/ssDNA junctions on branched DNA substrates (126)		+	+		Prokaryota *(Actinobacteria, β-proteobacteria)*	All 3 sequences from Deinococcus-Thermus are located within the Archaea clade. The Proteobacteria sequences are close to the root, this topology is not well resolved

(continued)

Table 1. Continued

No.	Name	Pfam, COG/KOG, PDB90 structure	Reference to fold assignment	Biological function	Taxonomy					HGTs
					Viruses	Bacteria	Archaea	Eukaryota	Detailed distribution	
72	ERCC1-XPF, Swi10, Rad10	PF03834 KOG2841 COG5241 2a1i	(127)	Nuclease of NER system incising oligonucleotide from damaged DNA strand (128)				+	Eukaryota	No HGT observed
73	La crosse virus L-protein	2xi5	(129)	Cap-snatching Endonuclease; cleaves short and capped host primers that are subsequently used by viral RNA-dependent RNA polymerase to transcribe viral mRNAs (129)	+				Bunyaniviridae	No HGT observed
74	Viral L-protein	PF00603 3hw3	(130)	Cap-snatching Endonuclease, mechanism identical to that described above (131)	+				Influenza A virus	Phylogeny not resolved
75	D212	PF12187 2w8m	(132)	Uncharacterized nuclease suggested to take part in DNA replication, repair, or recombination (132)	+		+		Archaea (*Sulfolobus*) archaeal phages	Phages and prophages of Sulfolobus, together form one coherent clade
76	Archaea bacterial proteins of unknown function, DUF234	PF03008 COG1672	(5)	DEXX-box ATPase belonging to AAA+ superfamily; DEXX-box ATPases act to transduce the energy of ATP-hydrolysis into a conformational stress required for the remodeling of nucleic acid or protein–nucleic acid structure (133).		+	+		Prokaryota	Two *Treponema vincentii* (Spirochaetales) sequences are in a *Butyrivibrio proteoclasticus*/*Ruminococcus bromii*/*Roseburia inulinivorans* rumen bacteria (Clostridiales) clade
77	RAI1-like, Dom-3z	PF08652 KOG1982 3fqg, 3fqi	(7)	Exoribonuclease. Has a pyrophosphohydrolase activity towards 5′-triphosphorylated RNA (7).				+	Eukaryota	No HGT observed[a]
78	NARG2	PF10505	(134)	Nuclear protein involved in thickness of the brain's cortical gray matter regulation (57)				+	Eukaryota *(without Plantae & Chromoalveolata)*	No HGT observed
79	DUF911, Dna2	PF06023 PF08696 KOG1805 COG4343	(39)	Dna2 processes common structural intermediates that occur during diverse DNA processing (e.g. lagging strand synthesis and telomere maintenance) (135). Dna2 is a dual polarity exo/endonuclease, and 5′ to 3′ DNA helicase involved in Okazaki Fragment Processing (OFP) (136) and DSB Repair (137). DUF911 function is unknown.		+	+	+	Prokaryota & Eukaryota	Very long branches, dubious positioning of various taxons

(continued)

Table 1. Continued

No.	Name	Pfam, COG/KOG, PDB90 structure	Reference to fold assignment	Biological function	Taxonomy					HGTs
					Viruses	Bacteria	Archaea	Eukaryota	Detailed distribution	
80	YhgA-like	PF04754 COG5464	(36)	Putative transposase (138). The genomic context is not conserved even among strains of one species suggesting recent mobility.		+			Bacteria (*γ*-proteobacteria)	Three *Burkholderia rhizoxinica* (β-proteobacteria) sequences are present on a Enterobacteriales clade forming a sister clade to a Yersinia clade
81	CoiA-like	PF06054 COG4469	(39)	Negative regulator of competence. CoiA is probably involved after DNA uptake, either in DNA processing or recombination (139).		+			Bacteria *(Bacillus, Lactobacillus)*	No HGT observed
82	DUF524	PF04411 COG1700	(36)	Predicted restriction endonuclease (36). Co-occurs with a restriction GTPase or ATPase.		+	+		Bacteria & Euryarchaeota	Mixed clades like: *Geobacter uraniireducens* (Deltaproteobacteria) together with *Gallionella capsiferriformans* (β-proteobacteria) and *Chlorobium luteolum* (Chlorobia)
83	Mitochondrial protein Pet127	PF08634	(134)	5′-exonuclease responsible for processing the precursor to the mature form (140) involved in modulation of mtRNAP activity				+	Alveolata Fungi Myxomycota Excavata	Distribution limited to different unicellular eukaryote, not enough sequencing data for a HGT hypothesis
84	Eukaryotic translation initiation factor 3 subunit 7, eIF-3-zeta, eIF3 p66, moe1	PF05091 KOG2479	(134)	eIF3 p66 is the major RNA-binding subunit of the eIF3 complex; Cdc48, Yin6 and Moe1 act in the same protein complex to concertedly control ERAD and chromosome segregation (141).				+	Eukaryota	No HGT observed
85	Secreted endonuclease distantly related to HJC resolvase	PF10107 COG4741	(11)	Predicted secreted endonuclease distantly related to archaeal HJC resolvase		+	+	{1}	Prokaryota	A sequence of a bacteria feeding nematode *Caenorhabditis remanei* in an Acintobacter clade. Archaea sequences in Bacteria clades
86	DUF1064	PF06356	(39)	Unknown, In firmicutes co-occurs with: RecT, DnaC, DnaB, SSB what suggest a role in recombination. In Proteobacteria phage proteins are also present.	+	+			Bacteria phages	*Beggiatoa* sp. (γ-proteobacteria) within a Clostridiales clade
87	DUF790	PF05626 COG3372	(39)	Unknown. Co-occurs with ResIII and helicase domains.		+	+		Prokaryota	A single sequence of *Rubrobacter xylanophilus* (Actinobacteria) locates with Cyanobacteria and Deinococci

(continued)

Table 1. Continued

No.	Name	Pfam, COG/KOG, PDB90 structure	Reference to fold assignment	Biological function	Taxonomy					HGTs
					Viruses	Bacteria	Archaea	Eukaryota	Detailed distribution	
88	VRR-NUC	PF08774 KOG2143	(39)	A DNA repair nuclease recruited to DNA damage by monoubiquitinated FANCD2 (142) exhibits endonuclease activity toward 5′ flaps and has 5′ exonuclease activity. In γ-proteobacteria co-occurs with DEAD_2 helicase and bacterial extracellular solute-binding protein family POTD/POTF.	+	+		+	Bacteria & Eukaryota & phages	No HGT observed
89	RmuC	PF02646 COG1322	(39)	Molecular function unknown. Involved in DNA recombination (143), neighborhood of metallopeptidases and MFS1 transporters		+			Bacteria *(mostly γ-proteobacteria)*	*Lentisphaera araneosa* (Lentisphaere) in a Oceanospirillales (Proteobacterial) clade, forms a clade together with *Neptuniibacter caesariensis*. Both bacteria were isolated from a surface water sample (144,145)
90	Uncharacterized conserved protein	COG5482	New	Unknown	{2}	+		{1}	Bacteria *(mostly α-proteobacteria)* & phages	*Ricinus communis* (Plantae) forms a clade with a tumorogenic *Agrobacterium radiobacter* (Rhizobiales) within a Rhizobiales clade
91	Predicted transcriptional regulator	COG1395	New	The function is unknown but it likely binds nucleic acids. Harbors a HTH motif, co-occurs with a two-domain protein consisting of DUF1743 and tRNA_anti (PF01336) nucleic acid-binding OB-fold domain.			+		Archaea	No HGT observed
92	DUF1052	PF06319 COG5321 3dnx	Pfam	Co-occurs with HisKA and Lactamase_B or YkuD (PF03734) which also gives β-lactam resistance.	{1}	+			Bacteria *(mostly α-proteobacteria)*	An uncultured Acidobacterium within a Rhizobiales clade with *Nitrobacter*, *Bradyrhizobium* and *Rhodopseudomonas palustris*. *Acidobacteria*, *Nitrobacter*, *Bradyrhizobium* are soil related bacteria, but *R. palustris* is found in sea sediments
93	Sugar fermentation stimulation protein SfsA	PF03749 COG1489	(146)	Unknown, SfsA protein binds to DNA non-specifically (147). Connected with maltose metabolism (147). In γ-proteobacteria in the proximity of LigT and Pol A or with a C4-type zinc finger and nucleotidyl-transferase domain. In Cyanobacteria co-occurs		+			Bacteria *(mostly Proteobacteria)*	*Plautia sali* symbiont (unclassified bacterium) groups with a *Pantoea* sp. clade (γ-proteobacteria)[a]

(continued)

Table 1. Continued

No.	Name	Pfam, COG/KOG, PDB90 structure	Reference to fold assignment	Biological function	Taxonomy					HGTs
					Viruses	Bacteria	Archaea	Eukaryota	Detailed distribution	
				with transport proteins related to virulence. In Archaea with a MSF_1 transporter or Lactamase_B.						
94	NERD	PF08378	(38)	Unknown, described as nuclease-related (38)		+	{2}		Bacteria	*Planctomyces limnophilus* (Planctomycetales) groups with *Puniceispirillum marinum* (α-proteobacteria). *Mannheimia succiniciproducens* (γ-proteobacteria) locates in a Neisseria (β-proteobacteria) clade. Clades with mixed taxonomic groups
95	DUF1626	PF07788 COG5493	(36)	Unknown		+	+		Prokaryota	*Thermodesulfovibrio yellowstonii* (Nitrospirales) within a Cyanobacterial clade mostly *C. raciborskii*. *Cylindrospermopsis raciborskii* is bloom-forming and potentially toxic river cyanobacteria. *T. yellowstonii* was isolated form thermal vent water. Patchy distribution in Bacteria suggesting multiple HGT events
96	UPF0102, RPA0323	PF02021 COG0792 COG4998 3fov	Pfam	Is often found with a TP_methylase (PF00590) domain. Tetrapyrrole (Corrin/Porphyrin) Methylases use S-AdoMet in the methylation of diverse substrates. The genomic context is well conserved for each bacterial class.		+	+		Prokaryota	*Cryptobacterium curtum* (Actinobacteria) in a Clostridium clade[a]
97	DUF1887	PF09002 1xmx	Pfam	Occasionally co-occurs with phosphorylase superfamily PNP_UDP_1 (PF01048) (uridine phosphorylase) and zinc/cadmium/mercury/lead-transporting ATPase.		+	+		Prokaryota	Three *M. smithii* (Euryarchaeota) sequences form a clade with 2 sequences from *Synechococcus* sp. from Yellowstone (Cyanobacteria) and *M. ruber (Thermales)*. *Methanobrevibacter smithii* is a methanogenic archeon highly resistant to antibiotics
98	DUF1016	PF06250 COG4804	(39)	Co-occurs with restriction MTase, ResIII and ResI S domains, and mobile element domains (phage integrase, DDE). Might act as nucleic acid-binding element in restriction enzymes.	{1}	+	{3}	{2}	Bacteria	*Trichoplax adhaerens* (Plecozoa) groups with a Bacterioidales clade with two additional HGT transfered sequences: *Rickettsia felis* (α-proteobacteria) and *Legionella longbeachae* (γ-proteobacteria). *Ricinus communis* (Plantae) locates with a Burkholderiales clade harboring other unrelated taxa from γ-proteobacteria: *Thioalkalivibrio* sp., *Pseudomonas aeruginosa* and *Dickeya dadantii*

(continued)

Table 1. Continued

No.	Name	Pfam, COG/KOG, PDB90 structure	Reference to fold assignment	Biological function	Taxonomy					HGTs
					Viruses	Bacteria	Archaea	Eukaryota	Detailed distribution	
99	DUF1703	PF08011	(36)	There are 9 DUF1703 proteins in *Methanospirillum hungatei* DSM 864. Some of them reside in the proximity of multiple PAS fold domains and CheY sensor related genes. In Bacterioidetes the genomic context is not conserved due to a duplication.		+	{1}		Bacteria *(mostly Bacterioidetes)*	Nine sequences from *M. hungatei* form a sister clade to a Proteobacteria clade. This clade is grouped together with a Treponema clade. The rest of the tree belongs to Bacterioidetes
100	DUF4143	COG1373, PF13635	Pfam	Unknown		+	+		Prokaryota	*Ilyobacter polytropus* (Fusobacteriales) forms a clade with *C. sticklandii* (Clostridiales). *Ilyobacter polytropus* was isolated from marine anoxic mud
101	DUF511	PF04373 COG2958	(11)	Unknown		+			Bacteria	Unrelated sequences from Fibrobacterales, Chlorobiales, Clostridiales, Flavobacteriales and Bacteroidales on a Proteobacteria tree
102	DUF2887	PF11103	(11)	Unknown. Co-occurs with transport related proteins.		+			Bacteria *(Cyanobacteria)*	*Methylococcus capsulatus* and Beggiatoa sequences are found within a Cyanobacteria clade
103	Restriction endonuclease-like fold superfamily protein	3ijm	PDB	Unknown		+			*Spirosoma linguale (Cytophagales)*	No HGT observed
104	DUF1853	PF08907 COG3782	(11)	Unknown. The genomic context is conserved within bacterial families.		+			Bacteria (*mostly Proteobacteria*)	*Anacystis nidulans* (Cyanobacteria), Planctomycetes and Flavobacteria within a Proteobacteria clade
105	UL24	PF01646	(36)	The molecular mechanism is unknown however the UL24 protein is able to induce G2 cell-cycle arrest (148), disperse nucleolin (149) and alter the nuclei. The PD-(D/E)XK motif preservation is crucial for these functions (150).	+			+	Herpesvirales	No HGT observed
106	DUF506	PF04720	(36)	Unknown					Plantae Green algae	No HGT observed
107	TT1808, DUF820, Uma2	PF05685 COG4636 1wdj, 3ot2	(39)	Predicted endonuclease. In Cyanobacteria the genomic context is well conserved. In γ-proteobacteria the context is not conserved and involves mobile elements suggesting recent mobility and/or acquisition.		+			Bacteria	Proteobacteria sequences within Firmicutes or Cyanobacteria clades. Very long branches. Multiple transfer
108	DUF1780	PF08682 1y0k	SCOP	Unknown. Well conserved context		+			Bacteria *(Pseudomonadales)*	No HGT observed

(continued)

Table 1. Continued

No.	Name	Pfam, COG/KOG, PDB90 structure	Reference to fold assignment	Biological function	Taxonomy					HGTs
					Viruses	Bacteria	Archaea	Eukaryota	Detailed distribution	
109	DUF2130	PF09903 COG4487	Pfam	Unknown		+	{1}		Bacteria	*Parascardovia denticolens* and *Scardovia inopinata* (Bifidobacteriales) in a Lactobacillaes clade. One archeon *M. paludicola*
110	DUF2726	PF10881	Pfam	Unknown. In Fusobacteria DUF2726 proteins are surrounded by mobile elements. This feature is less pronounced in other bacteria.		+		+	Bacteria	Multiple transfers. *Pirellula staleyi* (Plantomyces) forms a clade with *Anaerolinea thermophila* (Chloroflexi)
111	RAP domain	PF08373	Pfam	Unknown. Initially claimed to bind RNA and abundant in Apicomplexans, present in proteins involved in mitochondrial stress sensing (151) and plant immunity (152).		{1}		+	Eukaryota	*Parachlamydia acanthamoebae* is located with a lycophyte, *Selaginella moellendorffii*, long branches
112	YaeQ	PF07152 COG4681 2ot9, 2g3w 3c0u	(153)	Located with bleomycin resistance (Glyoxalase) and Aceltyltransf_1 (GNAT). In *P. aeruginosa* biofilms a YaeQ mutant has decreased expression of genes encoding NADH dehydrogenase activity and cobalamin biosynthetic process and increased expression of secretion and pathogenesis genes (e.g. exoY, pscU and exsC). This mutant has biofilm-exclusive tobramycin fitness advantages. Tobramycin is an aminoglycoside antibiotic. YaeQ compensates (154) or does not (155) the hemolysin transcription elongation protein RfaH function.		+			Bacteria *(Proteobacteria)*	*Nitrospira defluvii* on a Proteobacteria tree forms a clade with *Leptothrix cholodnii*. *Ricinus communis* (Plantae) groups with *Methylotenera mobilis*
113	PDDEXK_2	PF12784	Pfam	Putative transposase		+	{1}		Bacteria	Phylogeny not resolved
114	PDDEXK_3	PF13366	Pfam	Unknown	+	+	+		Prokaryota & Viruses	Multiple transfers, mixed clades for Bacteria and Archaea or different Bacterial divisions

(continued)

Table 1. Continued

No.	Name	Pfam, COG/KOG, PDB90 structure	Reference to fold assignment	Biological function	Taxonomy					HGTs
					Viruses	Bacteria	Archaea	Eukaryota	Detailed distribution	
115	PDDEXK_4	PF14281	Pfam	Unknown		+	+	{1}	Prokaryota	*Ricinus communis* (Plantae) is present in a Proteobacteria clade. *Parabacteroides merdae* a human gut bacterium found also in wounds forms a clade with a bacteria from termite hindguts *Treponema primitia*
116	DUF4263	PF14082	New	Unknown	{1}	+	{2}	{1}	Bacteria	*Populus balsamifera* subsp. *trichocarpa* (Plantae) sequence forms a clade with a non-pathogenic metal resistant bacterium *Ralstonia metallidurans*
117	DUF3883	PF13020	New	Unknown		+	+	+	Eukaryota & Prokaryota	Phylogeny not well resolved[a]
118	DUF4420	PF14390	New	Putative transposase		+	{2}		Bacteria	*Methanoplanus petrolearius* (Euryarchaeota) and an uncultured archaeon locate within a Bacteria (Bacterioidetes/Actinobacteria) clade. Multiple transfers
119	Replic_Relax	PF13814	New	Plasmid replication (156) and plasmid DNA relaxation (157)	{1}	+			Bacteria (*Bacillus*/*Clostridium* & *Actinobacteria*)	*Streptococcus* (Lactobacillales) locates within an Actinobacteria clade. *Paenibacillus* (Bacillales) sequence is found in an Actinobacteria clade
120	Dam-replacing protein	PF06044	(158)	DNA adenine methyltransferase replacing protein (DRP), a restriction endonuclease (158)	{2}	+	{3}		Bacteria	Patchy distribution possibly due to multiple transfers
121	TBP-interacting protein	2czr	(159)	A family of proteins, that interact with TATA-binding protein (TBP) (159).			+		Archaea *(Thermococcales)*	No HGT observed

Groups include closely related families and structures that share relatively high sequence similarity detectable with PSI-BLAST and RPS-BLAST.

[a]The tree was not rooted due to dubious position of the rooting sequence.

The curly brackets in the taxonomy columns indicate the number of sequences if kingdom is represented only by a few sequences.

of PD-(D/E)XK proteins and provide a solid background for rational hypotheses about structurally unstudied families. In the next section we describe multiple aspects of structural changes that blur a commonly recognized image of the restriction endonuclease-like proteins.

Core variability

The structural core of PD-(D/E)XK phosphodiesterase fold includes only six major elements: four β-strands and two α-helices (Figure 3A). We believe that this minimalism contributes to structural diversity of the superfamily. The first and the second core β-strands can embrace only a few residues (pdb|1y0k, Figure 3B), hardly forming a well-defined part of the central β-sheet. On the other hand, they can also be very long, forming a hairpin, which barely interacts with the rest of the β-sheet and keeps the remaining region bent away from the core structure (RecBCD nuclease, pdb|1w36 chain C, Figure 3C). Even if all core secondary structures are present, their spatial arrangement may still vary significantly. In a canonical PD-(D/E)XK enzyme α-helices remain in a roughly parallel orientation, whereas in the Pa4535 protein (pdb|1y0k, Figure 3B) they are almost perpendicular. In addition, we also observed circular permutations, e.g. in HJC resolving enzyme (pdb|1j22), where the first core α-helix is formed by the C-terminal sequence region, while N-termini encodes the first core β-strand (Figure 3D). Finally, the repertoire of structural variation within restriction endonuclease-like proteins is additionally enriched by domain swapping. For instance, bacteriophage T7 endonuclease I (pdb|2pfj) exchanges the first core α-helix and the first core β-strand between separate chains, both forming catalytically active, dimerized domains (Figure 3E).

Insertions to core

In order to investigate the capabilities of the fold to handle additional structural elements we studied the structures of known PD-(D/E)XK proteins. The PD-(D/E)XK structural core is often decorated with plenty of insertions that tune the substrate-binding capabilities or enable protein-protein interactions (Supplementary Figure S1). The structure of *Bacillus subtilis* RecU resolvase (pdb|1zp7) is a remarkable example of tweaking canonical restriction endonuclease core for a specific function. It has a characteristic stalk formed by the first and the second β-strands extensions that fits into a four-way junction central region and provides a scaffold for substrate destabilizing interactions.

Interestingly, using topology based-searches we identified PD-(D/E)XK core fold in many unrelated structures (Supplementary Figure S2). The so called 'Russian-doll' effect is discussed in more detail in Supplementary Materials [PD-(D/E)XK fold in other unrelated structures].

Active site variation

A PD-(D/E)XK active site residues fingerprint varies between the families (Figure 4). For instance, the signature motif proline can be replaced by any residue (mainly hydrophobic). Having a vast collection of PD-(D/E)XK proteins we analyzed possible alterations to the archetypical active site architecture. Such information is fundamental for further effective searches for new, putative PD-(D/E)XK enzymes within uncharacterized protein families. The canonical active site is formed by aspartic acid placed in the N-termini of the second core β-strand and glutamic acid, followed by lysine from the third β-strand, placing the carboxyl and amino groups in a suitable spatial arrangement. Interestingly, the glutamic acid and lysine may be shifted into nearby structural elements, tending however to position their chemical groups towards the active site and preserving its catalytic functionality (10). We observed such migration in several structures: (i) Cfr10I restriction endonuclease (pdb|1cfr), where glutamic acid migrates from the third β-strand to the adjacent, second core α-helix resulting in the PD-XXK-E motif; (ii) EcoO109I restriction enzyme (pdb|1wtd), where glutamic acid E moves from the expected position 124 into position 108 and now precedes aspartic acid from the PD motif (motif EPD-XXK); (iii) Pa4535 structural genomics hypothetical protein (pdb|1y0k), where lysine migrates from the expected position 70 into position 125 in the adjacent second core α-helix (motif PD-EXX-K). Interestingly, tRNA splicing endonucleases acquired a different active site within restriction endonuclease-like fold. These enzymes conserve three catalytic residues: tyrosine, histidine and lysine (Y115, H125, K156 in a *Methanococcus jannaschii* endonuclease) that form an active site located on the opposite edge of the central β-sheet. Even though tRNA-splicing endonucleases share a common PD-(D/E)XK fold, they eventually recognize a different substrate and possess a distinct catalytic mechanism.

Sequence analyses

Although most of the PD-(D/E)XK proteins have a nuclease activity, they may also perform other diverse functions. Adaptation to a particular functional niche may involve the presence of additional protein domains encoded separately or together with the PD-(D/E)XK domain. Some functions are restricted to a certain taxonomic unit while others are widely distributed across the tree of life. In order to gain a general overview of sequence similarities, all 21 911 protein sequences were clustered with CLANS. The obtained clustering was colored based on both sequence taxonomic distribution and protein function (Figure 5). One should note that restriction endonucleases exhibit high sequence divergence, whereas house-keeping genes form tight clusters. Bacterial sequences are present all over the sequence space in contrast to viral sequences which appear only in a handful of sequence groups. Our analysis of taxonomic distribution, genomic context and domain architecture of PD-(D/E)XK proteins should help understand their biological relevance.

Domain architecture

We extensively studied a domain organization for all collected PD-(D/E)XK proteins that might provide a broader view on the diversity of functional associations in this

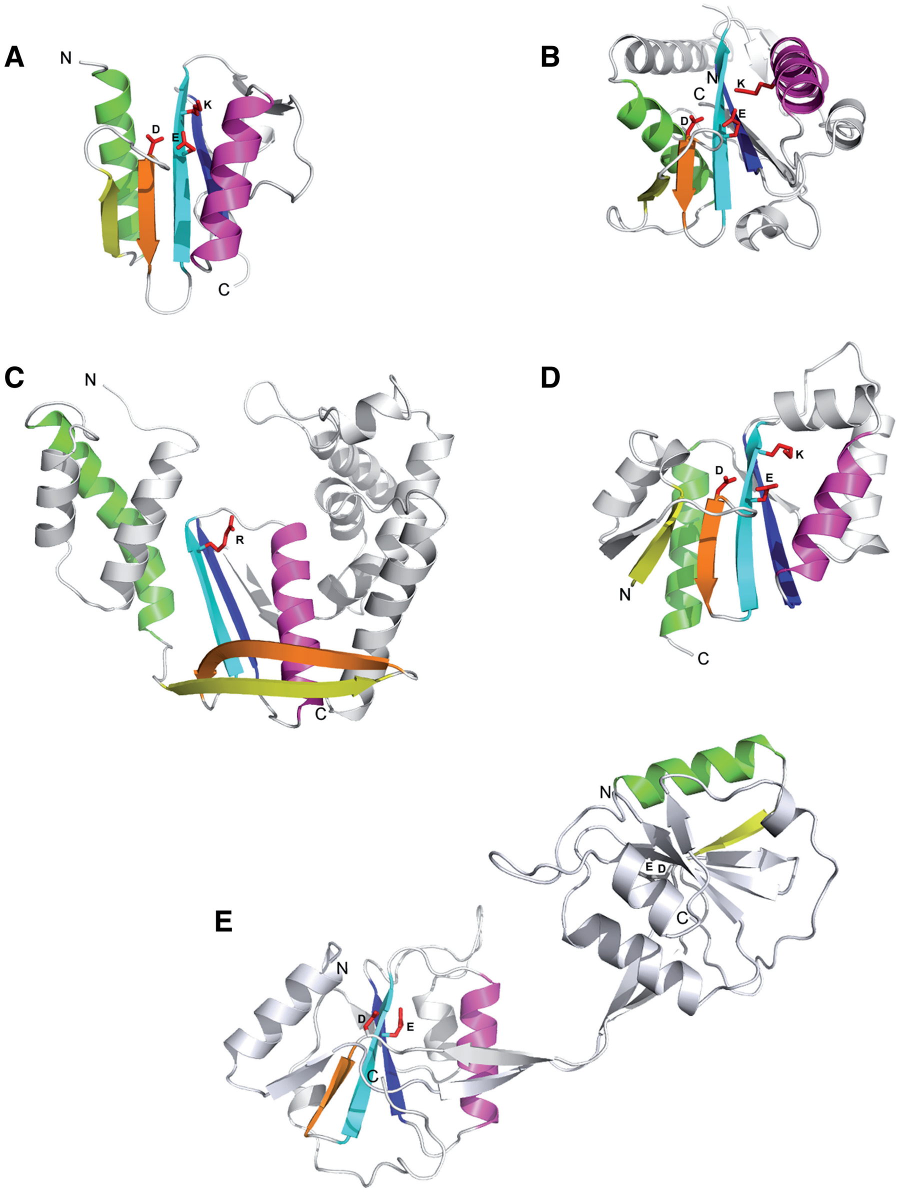


Figure 3. Examples of structural diversity in the PD-(D/E)XK phosphodiesterase superfamily. (**A**) typical PD-(D/E)XK enzyme (Holiday junction resolvase, *Pyrococcus furiosus*, pdb|1gef); (**B**) highly diverged structure with short first β-strand and perpendicular orientation of core α-helices (Pa4535 protein, *P. aeruginosa*, pdb|1y0k); (**C**) structure deterioration and the loss of active site (RecC, *E. coli*, pdb|1w36C); (**D**) circular permutation of the first core α-helix (Hef endonuclease, *Pyrococcus furiosus*, pdb|1j22); (**E**) domain swapping (endonuclease I, Enterobacteria phage T7, pdb|2pfj). Active site PD-(D/E)XK signature residues are shown as red sticks.

superfamily and also hint at specific functions for uncharacterized and poorly annotated proteins. In particular, we identified fused protein domains, internal repeat regions, coiled-coils and transmembrane elements. We observed various interesting domain arrangements that adjust the PD-(D/E)XK protein function to a specific role (Supplementary Figure S3), although most of the analyzed proteins harbor a single PD-(D/E)XK domain. Altogether, we identified 535 fused protein domains of distinct functions in 79 PD-(D/E)XK groups (Supplementary Table S4). Some of the most interesting and newly observed domain architectures are described in Supplementary Materials [Domain architecture], whereas a complete list of domain arrangements is included as Supplementary Figure S3.

Taxonomic distribution and horizontal gene transfers

The abundance of possible functions within PD-(D/E)XK phosphodiesterase proteins raises a question of the origin of these enzymes. In order to gain some insight into evolutionary history of these proteins we looked at the taxonomic distribution of the 121 PD-(D/E)XK groups (Table 1 and Supplementary Dataset S2). The housekeeping genes such as: HJC resolvase, RecBCD or tRNA intron endonuclease exhibit a broad taxonomic distribution. On the other hand, restriction endonucleases are

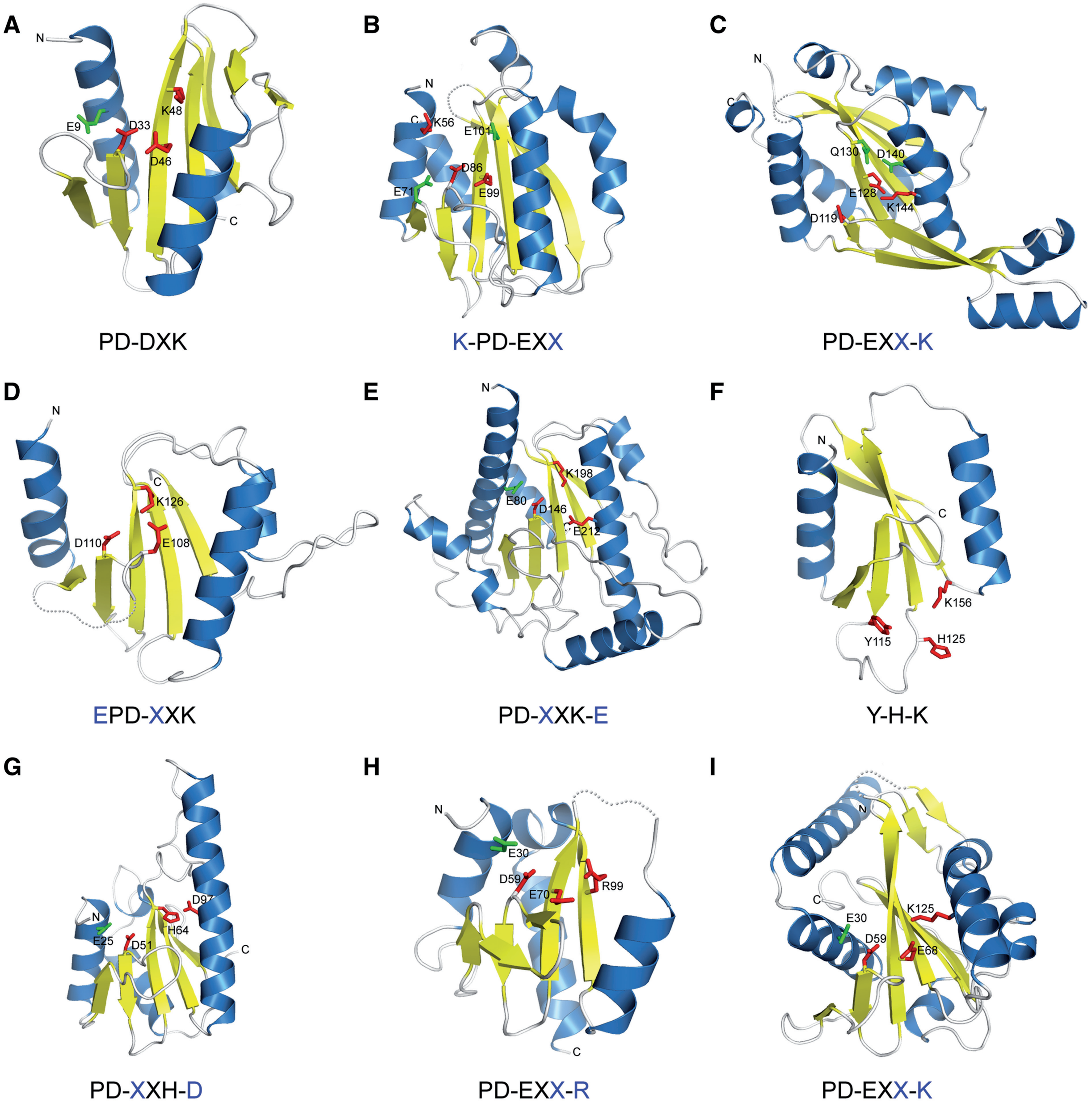


Figure 4. Active site variations observed in the PD-(D/E)XK phosphodiesterase superfamily structures. Observed variant of 'PD-(D/E)XK' signature motif is given below each structure with residue migration denoted in blue. (**A**) archaeal HJC resolvase (*P. furiosus*, pdb|1gef); (**B**) BamHI restriction endonuclease (*Oceanobacter kriegii*, pdb|3odh); (**C**) BstYI restriction endonuclease (*Geobacillus stearothermophilus*, pdb|1sdo); (**D**) EcoO109I restriction endonuclease (*E. coli*, pdb|1wtd); (**E**) Bse634I restriction endonuclease (*Geobacillus stearothermophilus*, pdb|1knv); (**F**) tRNA splicing endonuclease (*Methanocaldococcus jannaschii*, pdb|1a79); (**G**) Vsr repair endonuclease (*E. coli*, pdb|1cw0); (**H**) a putative endonuclease-like protein (*Neisseria gonorrhoeae*, pdb|3hrl); (**I**) Pa4535 protein (*P. aeruginosa*, pdb|1y0k).

usually unevenly distributed among a handful of specific orders of Prokaryota. Some PD-(D/E)XK proteins display a special taxonomic distribution. For example, the occurrence of Sen15 tRNA, a subunit of a splicing endonuclease is limited to Amebozoa and Ophistokonta. Noteworthy, in plants only two pre-tRNA molecules undergo splicing (tRNA^{Tyr} and tRNA^{Met}) (55) and the observed introns are significantly related in structure. The remaining Eukaryotic lineages could display alternative modes of tRNA intron endonuclease action. NARG2 and Pet127 proteins, also absent in plants, are known to participate in vital processes (32,56), but their molecular function is unknown. Pet127 is a mitochondrial protein involved in mtRNA polymerase regulation and mitochondrial mRNA maturation (32). The absence of Pet127 in plants raises a question of the differences in mtRNA polymerase performance and mtRNA maturation in these organisms. Initial studies on NARG2 claimed it is restricted to higher vertebrates and is involved in development (56). For example, human NARG2 protein is involved in the regulation of brain cortical gray matter thickness (57). Importantly, we

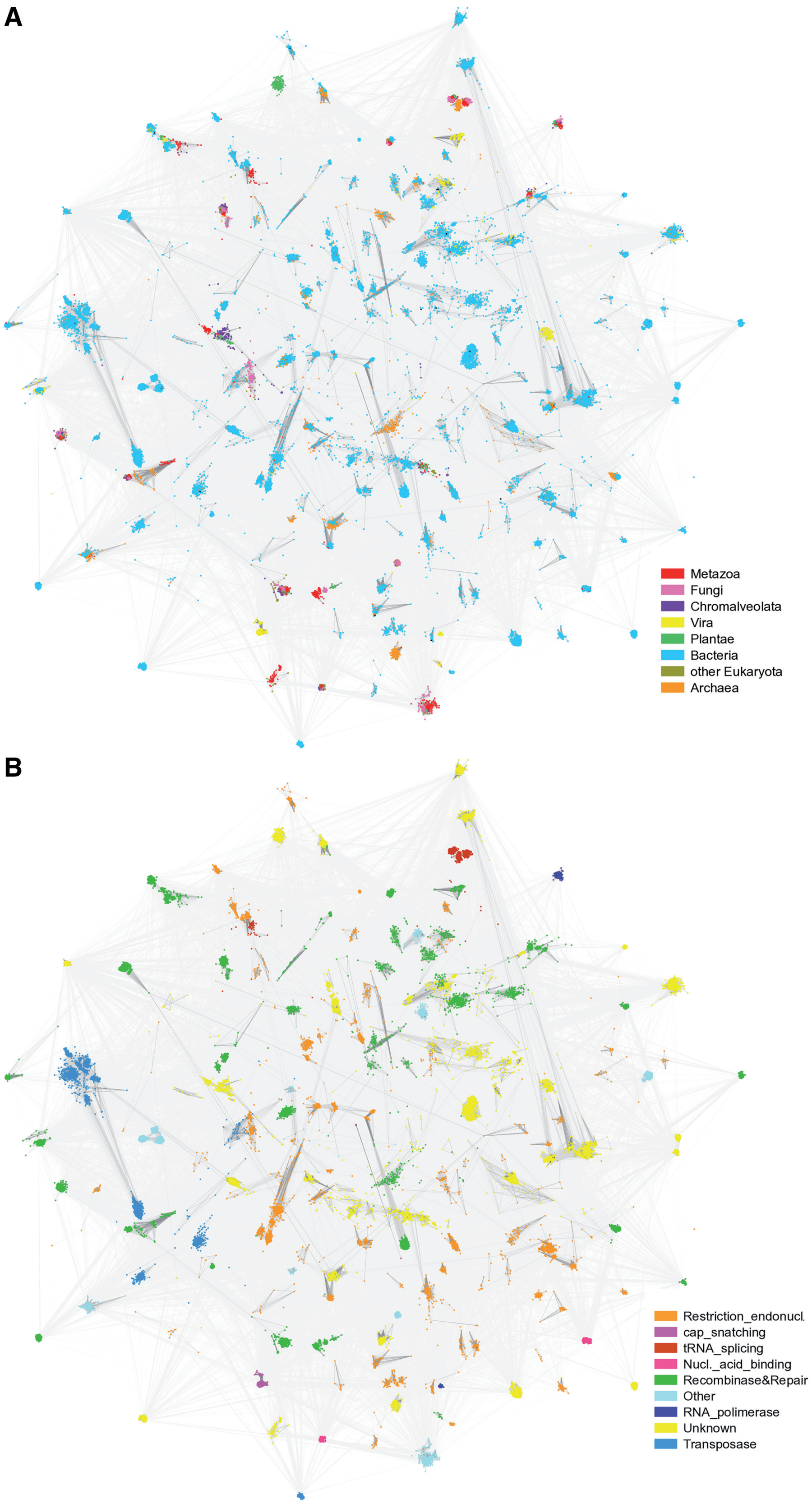


Figure 5. CLANS clustering of 21 911 sequences belonging to 121 clades of the PD-(D/E)XK superfamily. The image was drawn with an in-house script based on CLANS run files. (**A**) illustrates the taxonomic distribution of analyzed sequences and (**B**) summarizes their functional annotation.

found NARG2-like proteins to be also present in Amebozoa and Metazoa.

We observed Horizontal Gene Transfers (HGTs) in the majority of the PD-(D/E)XK groups. In the families that span multiple proteins originated from one taxon, together with a protein from evolutionary distant species, the HGT is the most parsimonious hypothesis which explains such uneven distribution. Derived tree topologies are often obscured by long and deep, unresolved branches. The distorted clades occasionally encompass sequences of mixed taxonomic origin and may intriguingly group together Archaea and Bacteria sequences. In 1996 Jeltsch and Pingoud (22) hypothesized that HGT affected the distribution and evolution of type II restriction enzymes. Our results corroborate their hypothesis. Patchy taxonomic distribution of restriction enzymes usually covers many unrelated taxonomic ranges, but is limited to a handful of representatives of each taxon. House-keeping genes such as HJC, Vsr do not transfer laterally. The event possibilities for each of the 121 PD-(D/E)XK clades are summarized in Table 1. In Supplementary Materials (Taxonomic distribution and HGTs), we describe some of the most interesting HGT events, with special attention paid to human pathogenic bacteria, and Prokaryota to Eukaryota transfers.

Summarizing, the patchy, narrow, or wide taxonomy distribution along with multiple HGT events greatly contribute to the complexity of the world of PD-(D/E)XK proteins that significantly vary in their structural features and display a wide range of domain architectures.

DISCUSSION

The PD-(D/E)XK proteins play important roles in many vital processes including the nucleic acid maintenance. Probably for this reason they are found in all living organisms. Across the superfamily, these proteins display a broad collection of general scaffold alterations which tweak their basic function to perform more specialized actions. The abundance of functions and distant evolutionary distances between particular PD-(D/E)XK families encouraged us to split the whole set of identified proteins into groups of sequences displaying obvious homology in terms of sequence comparison (Table 1). We expected such grouping to reflect the differences between functions and taxonomic distributions. Indeed, most of the defined groups show very coherent functions. The restriction enzymes and tRNA splicing endonucleases may be one of the most prominent examples here. However, some of the groups are blurred in terms of sequence similarity and cover many, yet connected functions including helicases, repair endonucleases, exonucleases and others. The difficulty of reproducing functional partition in our grouping procedure is 2-fold. The consensus sequence definitions that were used in our search included COG and KOG sequences which tend to cover multiple domains. This might lead to extended sequence alignment and boost of sequence similarity measure between distinct protein families. The other reason for grouping deficiency is the complex biological context of the analyzed proteins, especially that observed for housekeeping enzymes, like structure-specific repair nucleases. The alternative functions may emerge relatively fast, because homologous proteins may easily gain a new activity by fusing or interacting with unconventional protein domains. In our opinion, the precision of the grouping also strongly depends on the protein family concept which remains unclear.

PD-(D/E)XK phosphodiesterases exhibit great variability in sequence and structure. There are potentially two major reasons for that. These enzymes are involved in a variety of biological processes which require a very diverse range of substrates to be recognized in both the sequence- and structure-specific manner. High sequence dissimilarity, especially between restriction endonucleases is the result of evolutionary arms race between phages and bacteria (160).

A detailed analysis of insertions to the common conserved core observed in the existing structures across multiple PD-(D/E)XK families inspire a reflection that the majority of structural diversities are focused on the substrate-binding side (Supplementary Figure S1). The opposite side to the active site remains relatively unchanged.

The PD-(D/E)XK fold can be described as gregarious (161) referring to its presence in several evolutionary unrelated protein structures. N-acetyltransferases, lipases, dehydrogenases containing the PD-(D/E)XK domain as a substructure represent different folds (even fold classes) according to SCOP database. This finding provides novel challenges to protein structure classification that should probably describe structural space for the α/β sandwich architecture as the continuum rather than distinct folds. This also sheds new light on the possible mechanisms of fold change in the evolution of protein structure through the structural drift (162), and may also provide some hints about the evolutionary history of these proteins suggesting that some of them might have evolved from a common ancestor.

We observed many rare multiple domain architectures what is a general feature of sequence space (163). We identified PD-(D/E)XK domains that co-occur with the domains acting on nucleic acids, including methylases, helicases, resolvases, RNAse H, excision repair endonucleases, chromatin remodelers, or DNA ligases. These domain architectures follow the main functional niche occupied by nucleases. However, proteins with the PD-(D/E)XK domain can also be involved in protein structure maintenance. An interesting example is provided here by a hypothetical protein from *Vitis vinifera* discussed above (gi|147821195) which may be involved in both nucleic acid and histone protein structure upkeep, or Rai 1 from *Polysphondylium pallidum* (gi|281203778) followed by COBRA domain, a BRCA1 related protein that contributes to chromatin remodeling. Also intriguing domain association includes nucleases co-occurring with kinases. This might suggest that such proteins are somehow involved in triggering the response to nucleic acid aberrancies.

We observed the PD-(D/E)XK groups limited to one Archaea group (ThaI REase in Thermoplasmata), present in a few unrelated taxa (BglI REase) or conserved and essential in all domains of life (Dem1/EXO5). This phenomenon might be explained by the different roles played by conserved and patchy distributed proteins. The former are rarely transferred and inherited vertically, their mutations are strongly deleterious. In consequence, they appear in broad taxonomic groups in a fixed number of copies per genome and in all representatives of a taxon. The latter offer additional adaptive advantages, useful in a defined ecological niche and are frequently transferred laterally rather than inherited. The reported cases of HGT between human pathogenic bacteria or from bacteria to Eukaryotes additionally exemplify the complex evolution of the PD-(D/E)XK proteins.

CONCLUDING REMARKS

The aim of this project was to identify the most complete set of proteins retaining the PD-(D/E)XK fold. Such a collection is indispensable for a comprehensive view on this fold and enables further insight into detailed biological functions, exact substrates and the molecular mechanisms undergoing in the processes connected with nucleic acid cleavage.

The large and extremely diverse PD-(D/E)XK superfamily covers both specialized and multifunctional enzymes, as well as proteins that lost their enzymatic activity and now serve as structural or nucleic acid-binding units. Some of the PD-(D/E)XK fold families are restricted to a single bacterial family while others are present in all living organisms. The PD-(D/E)XK domains may co-occur solely, with one additional protein domain or in elaborated domain contexts. Moreover, some of the PD-(D/E)XK families harbor proteins appearing once per genome and others can display an increased number of copies. In humans the PD-(D/E)XK proteins can be linked to severe neurological diseases and may increase the probability of cancer.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–4, Supplementary Figures 1–4, Supplementary Materials, Supplementary Datasets 1–2 and Supplementary References [164–199].

FUNDING

EMBO Installation, Foundation for Polish Science (Team, Focus), Ministry of Science and Higher Education [N N301 159435, 0376/IP1/2011/71]; European Regional Development Fund under Innovative Economy Programme [POIG.01.01.02-14-054/09-00]; European Social Fund [UDA-POKL.04.01.01-00-072/09-00] grants. Funding for open access charge: Waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Orlowski,J. and Bujnicki,J.M. (2008) Structural and evolutionary classification of Type II restriction enzymes based on theoretical and experimental analyses. *Nucleic Acids Res.*, **36**, 3552–3569.
2. Belfort,M. and Weiner,A. (1997) Another bridge between kingdoms: tRNA splicing in archaea and eukaryotes. *Cell*, **89**, 1003–1006.
3. Hickman,A.B., Li,Y., Mathew,S.V., May,E.W., Craig,N.L. and Dyda,F. (2000) Unexpected structural diversity in DNA recombination: the restriction endonuclease connection. *Mol. Cell*, **5**, 1025–1034.
4. Dahlroth,S.L., Gurmu,D., Schmitzberger,F., Engman,H., Haas,J., Erlandsen,H. and Nordlund,P. (2009) Crystal structure of the shutoff and exonuclease protein from the oncogenic Kaposi's sarcoma-associated herpesvirus. *FEBS J.*, **276**, 6636–6645.
5. Aravind,L., Makarova,K.S. and Koonin,E.V. (2000) SURVEY AND SUMMARY: holliday junction resolvases and related nucleases: identification of new families, phyletic distribution and evolutionary trajectories. *Nucleic Acids Res.*, **28**, 3417–3432.
6. Ban,C. and Yang,W. (1998) Structural basis for MutH activation in E.coli mismatch repair and relationship of MutH to restriction endonucleases. *EMBO J.*, **17**, 1526–1534.
7. Xiang,S., Cooper-Morgan,A., Jiao,X., Kiledjian,M., Manley,J.L. and Tong,L. (2009) Structure and function of the 5′->3′ exoribonuclease Rat1 and its activating partner Rai1. *Nature*, **458**, 784–788.
8. Todone,F., Weinzierl,R.O., Brick,P. and Onesti,S. (2000) Crystal structure of RPB5, a universal eukaryotic RNA polymerase subunit and transcription factor interaction target. *Proc. Natl Acad. Sci. USA*, **97**, 6306–6310.
9. Nishino,T., Komori,K., Ishino,Y. and Morikawa,K. (2003) X-ray and biochemical anatomy of an archaeal XPF/Rad1/Mus81 family nuclease: similarity between its endonuclease domain and restriction enzymes. *Structure*, **11**, 445–457.
10. Feder,M. and Bujnicki,J.M. (2005) Identification of a new family of putative PD-(D/E)XK nucleases with unusual phylogenomic distribution and a new type of the active site. *BMC Genomics*, **6**, 21.
11. Laganeckas,M., Margelevicius,M. and Venclovas,C. (2011) Identification of new homologs of PD-(D/E)XK nucleases by support vector machines trained on data derived from profile-profile alignments. *Nucleic Acids Res.*, **39**, 1187–1196.
12. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
13. Pingoud,A., Fuxreiter,M., Pingoud,V. and Wende,W. (2005) Type II restriction endonucleases: structure and mechanism. *Cell. Mol. Life Sci.*, **62**, 685–707.
14. Wah,D.A., Bitinaite,J., Schildkraut,I. and Aggarwal,A.K. (1998) Structure of FokI has implications for DNA cleavage. *Proc. Natl Acad. Sci. USA*, **95**, 10564–10569.
15. Reeves,G.A., Dallman,T.J., Redfern,O.C., Akpor,A. and Orengo,C.A. (2006) Structural diversity of domain superfamilies in the CATH database. *J. Mol. Biol.*, **360**, 725–741.
16. Singleton,M.R., Dillingham,M.S., Gaudier,M., Kowalczykowski,S.C. and Wigley,D.B. (2004) Crystal structure of RecBCD enzyme reveals a machine for processing DNA breaks. *Nature*, **432**, 187–193.
17. Li,H., Trotta,C.R. and Abelson,J. (1998) Crystal structure and evolution of a transfer RNA splicing enzyme. *Science*, **280**, 279–284.
18. Roberts,R.J., Belfort,M., Bestor,T., Bhagwat,A.S., Bickle,T.A., Bitinaite,J., Blumenthal,R.M., Degtyarev,S., Dryden,D.T., Dybvig,K. *et al.* (2003) A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res.*, **31**, 1805–1812.
19. Punta,M., Coggill,P.C., Eberhardt,R.Y., Mistry,J., Tate,J., Boursnell,C., Pang,N., Forslund,K., Ceric,G., Clements,J. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.
20. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.

21. Bickle,T.A. and Kruger,D.H. (1993) Biology of DNA restriction. *Microbiol. Rev.*, **57**, 434–450.
22. Jeltsch,A. and Pingoud,A. (1996) Horizontal gene transfer contributes to the wide distribution and evolution of type II restriction-modification systems. *J. Mol. Evol.*, **42**, 91–96.
23. Lee,J.Y., Chang,J., Joseph,N., Ghirlando,R., Rao,D.N. and Yang,W. (2005) MutH complexed with hemi- and unmethylated DNAs: coupling base recognition and DNA cleavage. *Mol. Cell*, **20**, 155–166.
24. Bond,C.S., Kvaratskhelia,M., Richard,D., White,M.F. and Hunter,W.N. (2001) Structure of Hjc, a Holliday junction resolvase, from Sulfolobus solfataricus. *Proc. Natl Acad. Sci. USA*, **98**, 5509–5514.
25. McGregor,N., Ayora,S., Sedelnikova,S., Carrasco,B., Alonso,J.C., Thaw,P. and Rafferty,J. (2005) The structure of Bacillus subtilis RecU Holliday junction resolvase and its role in substrate selection and sequence-specific cleavage. *Structure*, **13**, 1341–1351.
26. Hadden,J.M., Convery,M.A., Declais,A.C., Lilley,D.M. and Phillips,S.E. (2001) Crystal structure of the Holliday junction resolving enzyme T7 endonuclease I. *Nat. Struct. Biol.*, **8**, 62–67.
27. Newman,M., Murray-Rust,J., Lally,J., Rudolf,J., Fadden,A., Knowles,P.P., White,M.F. and McDonald,N.Q. (2005) Structure of an XPF endonuclease with and without DNA suggests a model for substrate recognition. *EMBO J.*, **24**, 895–905.
28. Chang,J.H., Kim,J.J., Choi,J.M., Lee,J.H. and Cho,Y. (2008) Crystal structure of the Mus81-Eme1 complex. *Genes Dev.*, **22**, 1093–1106.
29. Kang,M.J., Lee,C.H., Kang,Y.H., Cho,I.T., Nguyen,T.A. and Seo,Y.S. (2010) Genetic and functional interactions between Mus81-Mms4 and Rad27. *Nucleic Acids Res.*, **38**, 7611–7625.
30. Buisson,M., Geoui,T., Flot,D., Tarbouriech,N., Ressing,M.E., Wiertz,E.J. and Burmeister,W.P. (2009) A bridge crosses the active-site canyon of the Epstein-Barr virus nuclease with DNase and RNase activities. *J. Mol. Biol.*, **391**, 717–728.
31. Kovall,R. and Matthews,B.W. (1997) Toroidal structure of lambda-exonuclease. *Science*, **277**, 1824–1827.
32. Fekete,Z., Ellis,T.P., Schonauer,M.S. and Dieckmann,C.L. (2008) Pet127 governs a 5′ -> 3′-exonuclease important in maturation of apocytochrome b mRNA in Saccharomyces cerevisiae. *J. Biol. Chem.*, **283**, 3767–3772.
33. Budde,B.S., Namavar,Y., Barth,P.G., Poll-The,B.T., Nurnberg,G., Becker,C., van Ruissen,F., Weterman,M.A., Fluiter,K., te Beek,E.T. *et al.* (2008) tRNA splicing endonuclease mutations cause pontocerebellar hypoplasia. *Nat. Genet.*, **40**, 1113–1118.
34. Naegeli,H. and Sugasawa,K. (2011) The xeroderma pigmentosum pathway: decision tree analysis of DNA quality. *DNA Repair*, **10**, 673–683.
35. Deans,A.J. and West,S.C. (2009) FANCM connects the genome instability disorders Bloom's Syndrome and Fanconi Anemia. *Mol. Cell*, **36**, 943–953.
36. Knizewski,L., Kinch,L.N., Grishin,N.V., Rychlewski,L. and Ginalski,K. (2007) Realm of PD-(D/E)XK nuclease superfamily revisited: detection of novel families with modified transitive meta profile searches. *BMC Struct. Biol.*, **7**, 40.
37. Knizewski,L., Kinch,L., Grishin,N.V., Rychlewski,L. and Ginalski,K. (2006) Human herpesvirus 1 UL24 gene encodes a potential PD-(D/E)XK endonuclease. *J. Virol.*, **80**, 2575–2577.
38. Grynberg,M. and Godzik,A. (2004) NERD: a DNA processing-related domain present in the anthrax virulence plasmid, pXO1. *Trends Biochem. Sci.*, **29**, 106–110.
39. Kinch,L.N., Ginalski,K., Rychlewski,L. and Grishin,N.V. (2005) Identification of novel restriction endonuclease-like fold families among hypothetical proteins. *Nucleic Acids Res.*, **33**, 3598–3605.
40. Ginalski,K., von Grotthuss,M., Grishin,N.V. and Rychlewski,L. (2004) Detecting distant homology with Meta-BASIC. *Nucleic Acids Res.*, **32**, W576–W581.
41. Ginalski,K., Elofsson,A., Fischer,D. and Rychlewski,L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics*, **19**, 1015–1018.
42. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
43. Pei,J., Sadreyev,R. and Grishin,N.V. (2003) PCMA: fast and accurate multiple sequence alignment based on profile consistency. *Bioinformatics*, **19**, 427–428.
44. Ginalski,K. and Rychlewski,L. (2003) Protein structure prediction of CASP5 comparative modeling and fold recognition targets using consensus alignment approach and 3D assessment. *Proteins*, **53 (Suppl 6)**, 410–417.
45. Shi,S., Zhong,Y., Majumdar,I., Sri Krishna,S. and Grishin,N.V. (2007) Searching for three-dimensional secondary structural patterns in proteins with ProSMoS. *Bioinformatics*, **23**, 1331–1338.
46. Sonnhammer,E.L., von Heijne,G. and Krogh,A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **6**, 175–182.
47. Yu,N.Y., Wagner,J.R., Laird,M.R., Melli,G., Rey,S., Lo,R., Dao,P., Sahinalp,S.C., Ester,M., Foster,L.J. *et al.* (2010) PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, **26**, 1608–1615.
48. Yu,C.S., Lin,C.J. and Hwang,J.K. (2004) Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. *Protein Sci.*, **13**, 1402–1406.
49. Horton,P., Park,K.J., Obayashi,T., Fujita,N., Harada,H., Adams-Collier,C.J. and Nakai,K. (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res.*, **35**, W585–W587.
50. Hoglund,A., Donnes,P., Blum,T., Adolph,H.W. and Kohlbacher,O. (2006) MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics*, **22**, 1158–1165.
51. Overbeek,R., Begley,T., Butler,R.M., Choudhuri,J.V., Chuang,H.Y., Cohoon,M., de Crecy-Lagard,V., Diaz,N., Disz,T., Edwards,R. *et al.* (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.*, **33**, 5691–5702.
52. Martinez-Guerrero,C.E., Ciria,R., Abreu-Goodger,C., Moreno-Hagelsieb,G. and Merino,E. (2008) GeConT 2: gene context analysis for orthologous proteins, conserved domains and metabolic pathways. *Nucleic Acids Res.*, **36**, W176–W180.
53. Dehal,P.S., Joachimiak,M.P., Price,M.N., Bates,J.T., Baumohl,J.K., Chivian,D., Friedland,G.D., Huang,K.H., Keller,K., Novichkov,P.S. *et al.* (2010) MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res.*, **38**, D396–D400.
54. Frickey,T. and Lupas,A. (2004) CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics*, **20**, 3702–3704.
55. Michaud,M., Cognat,V., Duchene,A.M. and Marechal-Drouard,L. (2011) A global picture of tRNA genes in plant genomes. *Plant J.*, **66**, 80–93.
56. Sugiura,N., Dadashev,V. and Corriveau,R.A. (2004) NARG2 encodes a novel nuclear protein with (S/T)PXX motifs that is expressed during development. *Eur. J. Biochem.*, **271**, 4629–4637.
57. Kochunov,P., Glahn,D.C., Nichols,T.E., Winkler,A.M., Hong,E.L., Holcomb,H.H., Stein,J.L., Thompson,P.M., Curran,J.E., Carless,M.A. *et al.* (2011) Genetic analysis of cortical thickness and fractional anisotropy of water diffusion in the brain. *Front Neurosci.*, **5**, 120.
58. Huai,Q., Colandene,J.D., Chen,Y., Luo,F., Zhao,Y., Topal,M.D. and Ke,H. (2000) Crystal structure of NaeI-an evolutionary bridge between DNA endonuclease and topoisomerase. *EMBO J.*, **19**, 3110–3118.
59. Newman,M., Lunnen,K., Wilson,G., Greci,J., Schildkraut,I. and Phillips,S.E. (1998) Crystal structure of restriction endonuclease BglI bound to its interrupted DNA recognition sequence. *EMBO J.*, **17**, 5466–5476.
60. Gemmen,G.J., Millin,R. and Smith,D.E. (2006) DNA looping by two-site restriction endonucleases: heterogeneous probability distributions for loop size and unbinding force. *Nucleic Acids Res.*, **34**, 2864–2877.
61. Stein,D.C., Gunn,J.S., Radlinska,M. and Piekarowicz,A. (1995) Restriction and modification systems of Neisseria gonorrhoeae. *Gene*, **157**, 19–22.
62. Xu,S.Y., Xiao,J.P., Ettwiller,L., Holden,M., Aliotta,J., Poh,C.L., Dalton,M., Robinson,D.P., Petronzio,T.R., Moran,L. *et al.* (1998)

Cloning and expression of the ApaLI, NspI, NspHI, SacI, ScaI, and SapI restriction-modification systems in Escherichia coli. *Mol. Gen. Genet.*, **260**, 226–231.

63. Kosinski,J., Kubareva,E. and Bujnicki,J.M. (2007) A model of restriction endonuclease MvaI in complex with DNA: a template for interpretation of experimental data and a guide for specificity engineering. *Proteins*, **68**, 324–336.
64. Athanasiadis,A., Vlassi,M., Kotsifaki,D., Tucker,P.A., Wilson,K.S. and Kokkinidis,M. (1994) Crystal structure of PvuII endonuclease reveals extensive structural homologies to EcoRV. *Nat. Struct. Biol.*, **1**, 469–475.
65. Arrand,J.R., Myers,P.A. and Roberts,R.J. (1978) A new restriction endonuclease from Streptomyces albus G. *J. Mol. Biol.*, **118**, 127–135.
66. Theriault,G., Roy,P.H., Howard,K.A., Benner,J.S., Brooks,J.E., Waters,A.F. and Gingeras,T.R. (1985) Nucleotide sequence of the PaeR7 restriction/modification system and partial characterization of its protein products. *Nucleic Acids Res.*, **13**, 8441–8461.
67. Bujnicki,J.M. and Rychlewski,L. (2001) Grouping together highly diverged PD-(D/E)XK nucleases and identification of novel superfamily members using structure-guided alignment of sequence profiles. *J. Mol. Microbiol. Biotechnol.*, **3**, 69–72.
68. Newman,M., Strzelecka,T., Dorner,L.F., Schildkraut,I. and Aggarwal,A.K. (1995) Structure of Bam HI endonuclease bound to DNA: partial folding and unfolding on DNA binding. *Science*, **269**, 656–663.
69. Lukacs,C.M., Kucera,R., Schildkraut,I. and Aggarwal,A.K. (2000) Understanding the immutability of restriction enzymes: crystal structure of BglII and its DNA substrate at 1.5 A resolution. *Nat. Struct. Biol.*, **7**, 134–140.
70. Bilcock,D.T., Daniels,L.E., Bath,A.J. and Halford,S.E. (1999) Reactions of type II restriction endonucleases with 8-base pair recognition sites. *J. Biol. Chem.*, **274**, 36379–36386.
71. Stankevicius,K., Povilionis,P., Lubys,A., Menkevicius,S. and Janulaitis,A. (1995) Cloning and characterization of the unusual restriction-modification system comprising two restriction endonucleases and one methyltransferase. *Gene*, **157**, 49–53.
72. Puchkova,L.I., Ushakova,T.A., Mikhailova,V.K., Serov,G.D., Krivopalova,G.N. and Repin,V.E. (2002) [Testing and isolation of high-purity restriction endonucleases]. *Prikl. Biokhim. Mikrobiol.*, **38**, 20–24.
73. Dahai,T., Ando,S., Takasaki,Y. and Tadano,J. (1999) Site-directed mutagenesis of restriction endonuclease HindIII. *Biosci. Biotechnol. Biochem.*, **63**, 1703–1707.
74. Hashimoto,H., Shimizu,T., Imasaki,T., Kato,M., Shichijo,N., Kita,K. and Sato,M. (2005) Crystal structures of type II restriction endonuclease EcoO109I and its complex with cognate DNA. *J. Biol. Chem.*, **280**, 5605–5610.
75. Winkler,F.K., Banner,D.W., Oefner,C., Tsernoglou,D., Brown,R.S., Heathman,S.P., Bryan,R.K., Martin,P.D., Petratos,K. and Wilson,K.S. (1993) The crystal structure of EcoRV endonuclease and of its complexes with cognate and non-cognate DNA fragments. *EMBO J.*, **12**, 1781–1795.
76. Sapienza,P.J., Rosenberg,J.M. and Jen-Jacobson,L. (2007) Structural and thermodynamic basis for enhanced DNA binding by a promiscuous mutant EcoRI endonuclease. *Structure*, **15**, 1368–1382.
77. Withers,B.E., Ambroso,L.A. and Dunbar,J.C. (1992) Structure and evolution of the XcyI restriction-modification system. *Nucleic Acids Res.*, **20**, 6267–6273.
78. van der Woerd,M.J., Pelletier,J.J., Xu,S. and Friedman,A.M. (2001) Restriction enzyme BsoBI-DNA complex: a tunnel for recognition of degenerate DNA sequences and potential histidine catalysis. *Structure*, **9**, 133–144.
79. Little,E.J., Babic,A.C. and Horton,N.C. (2008) Early interrogation and recognition of DNA sequence by indirect readout. *Structure*, **16**, 1828–1837.
80. Sullivan,K.M. and Saunders,J.R. (1989) Nucleotide sequence and genetic organization of the NgoPII restriction-modification system of Neisseria gonorrhoeae. *Mol. Gen. Genet.*, **216**, 380–387.
81. Wayne,J., Holden,M. and Xu,S.Y. (1997) The Tsp45I restriction-modification system is plasmid-borne within its thermophilic host. *Gene*, **202**, 83–88.
82. Xu,Q.S., Kucera,R.B., Roberts,R.J. and Guo,H.C. (2004) An asymmetric complex of restriction endonuclease MspI on its palindromic DNA recognition site. *Structure*, **12**, 1741–1747.
83. Deibert,M., Grazulis,S., Janulaitis,A., Siksnys,V. and Huber,R. (1999) Crystal structure of MunI restriction endonuclease in complex with cognate DNA at 1.7 A resolution. *EMBO J.*, **18**, 5805–5816.
84. Moissl-Eichinger,C. and Huber,H. (2011) Archaeal symbionts and parasites. *Curr. Opin. Microbiol.*, **14**, 364–370.
85. Deibert,M., Grazulis,S., Sasnauskas,G., Siksnys,V. and Huber,R. (2000) Structure of the tetrameric restriction endonuclease NgoMIV in complex with cleaved DNA. *Nat. Struct. Biol.*, **7**, 792–799.
86. Bozic,D., Grazulis,S., Siksnys,V. and Huber,R. (1996) Crystal structure of Citrobacter freundii restriction endonuclease Cfr10I at 2.15 A resolution. *J. Mol. Biol.*, **255**, 176–186.
87. Stankevicius,K., Lubys,A., Timinskas,A., Vaitkevicius,D. and Janulaitis,A. (1998) Cloning and analysis of the four genes coding for Bpu10I restriction-modification enzymes. *Nucleic Acids Res.*, **26**, 1084–1091.
88. Kachalova,G.S., Rogulin,E.A., Yunusova,A.K., Artyukh,R.I., Perevyazova,T.A., Matvienko,N.I., Zheleznaya,L.A. and Bartunik,H.D. (2008) Structural analysis of the heterodimeric type IIS restriction endonuclease R.BspD6I acting as a complex between a monomeric site-specific nickase and a catalytic subunit. *J. Mol. Biol.*, **384**, 489–502.
89. O'Driscoll,J., Heiter,D.F., Wilson,G.G., Fitzgerald,G.F., Roberts,R. and van Sinderen,D. (2006) A genetic dissection of the LlaJI restriction cassette reveals insights on a novel bacteriophage resistance system. *BMC Microbiol.*, **6**, 40.
90. Tamulaitiene,G., Jakubauskas,A., Urbanke,C., Huber,R., Grazulis,S. and Siksnys,V. (2006) The crystal structure of the rare-cutting restriction enzyme SdaI reveals unexpected domain architecture. *Structure*, **14**, 1389–1400.
91. Pingoud,V., Sudina,A., Geyer,H., Bujnicki,J.M., Lurz,R., Luder,G., Morgan,R., Kubareva,E. and Pingoud,A. (2005) Specificity changes in the evolution of type II restriction endonucleases: a biochemical and bioinformatic analysis of restriction enzymes that recognize unrelated sequences. *J. Biol. Chem.*, **280**, 4289–4298.
92. Bochtler,M., Szczepanowski,R.H., Tamulaitis,G., Grazulis,S., Czapinska,H., Manakova,E. and Siksnys,V. (2006) Nucleotide flips determine the specificity of the Ecl18kI restriction endonuclease. *EMBO J.*, **25**, 2219–2229.
93. Yang,Z., Horton,J.R., Maunus,R., Wilson,G.G., Roberts,R.J. and Cheng,X. (2005) Structure of HinP1I endonuclease reveals a striking similarity to the monomeric restriction enzyme MspI. *Nucleic Acids Res.*, **33**, 1892–1901.
94. Lambert,A.R., Sussman,D., Shen,B., Maunus,R., Nix,J., Samuelson,J., Xu,S.Y. and Stoddard,B.L. (2008) Structures of the rare-cutting restriction endonuclease NotI reveal a unique metal binding fold involved in DNA binding. *Structure*, **16**, 558–569.
95. Pawlak,S.D., Radlinska,M., Chmiel,A.A., Bujnicki,J.M. and Skowronek,K.J. (2005) Inference of relationships in the 'twilight zone' of homology using a combination of bioinformatics and site-directed mutagenesis: a case study of restriction endonucleases Bsp6I and PvuII. *Nucleic Acids Res.*, **33**, 661–671.
96. Neely,R.K. and Roberts,R.J. (2008) The BsaHI restriction-modification system: cloning, sequencing and analysis of conserved motifs. *BMC Mol. Biol.*, **9**, 48.
97. Cao,W. and Barany,F. (1998) Identification of TaqI endonuclease active site residues by Fe2+-mediated oxidative cleavage. *J. Biol. Chem.*, **273**, 33002–33010.
98. Vanamee,E.S., Viadiu,H., Kucera,R., Dorner,L., Picone,S., Schildkraut,I. and Aggarwal,A.K. (2005) A view of consecutive binding events from structures of tetrameric endonuclease SfiI bound to DNA. *EMBO J.*, **24**, 4198–4208.
99. Sokolowska,M., Kaus-Drobek,M., Czapinska,H., Tamulaitis,G., Szczepanowski,R.H., Urbanke,C., Siksnys,V. and Bochtler,M. (2007) Monomeric restriction endonuclease BcnI in the apo form and in an asymmetric complex with target DNA. *J. Mol. Biol.*, **369**, 722–734.

100. Firczuk,M., Wojciechowski,M., Czapinska,H. and Bochtler,M. (2011) DNA intercalation without flipping in the specific ThaI-DNA complex. *Nucleic Acids Res.*, **39**, 744–754.
101. Lapkouski,M., Panjikar,S., Janscak,P., Smatanova,I.K., Carey,J., Ettrich,R. and Csefalvay,E. (2009) Structure of the motor subunit of type I restriction-modification complex EcoR124I. *Nat Struct. Mol. Biol.*, **16**, 94–95.
102. Sisakova,E., Stanley,L.K., Weiserova,M. and Szczelkun,M.D. (2008) A RecB-family nuclease motif in the Type I restriction endonuclease EcoR124I. *Nucleic Acids Res.*, **36**, 3939–3949.
103. Sears,A., Peakman,L.J., Wilson,G.G. and Szczelkun,M.D. (2005) Characterization of the Type III restriction endonuclease PstII from Providencia stuartii. *Nucleic Acids Res.*, **33**, 4775–4787.
104. Shen,B.W., Xu,D., Chan,S.H., Zheng,Y., Zhu,Z., Xu,S.Y. and Stoddard,B.L. (2011) Characterization and crystal structure of the type IIG restriction endonuclease RM.BpuSI. *Nucleic Acids Res.*, **39**, 8223–8236.
105. Bujnicki,J.M. and Rychlewski,L. (2001) Identification of a PD-(D/E)XK-like domain with a novel configuration of the endonuclease active site in the methyl-directed restriction enzyme Mrr and its homologs. *Gene*, **267**, 183–191.
106. Waite-Rees,P.A., Keating,C.J., Moran,L.S., Slatko,B.E., Hornstra,L.J. and Benner,J.S. (1991) Characterization and expression of the Escherichia coli Mrr restriction system. *J. Bacteriol.*, **173**, 5207–5219.
107. Komori,K., Sakae,S., Shinagawa,H., Morikawa,K. and Ishino,Y. (1999) A Holliday junction resolvase from Pyrococcus furiosus: functional similarity to Escherichia coli RuvC provides evidence for conserved mechanism of homologous recombination in Bacteria, Eukarya, and Archaea. *Proc. Natl Acad. Sci. USA*, **96**, 8873–8878.
108. Gaillard,P.H., Noguchi,E., Shanahan,P. and Russell,P. (2003) The endogenous Mus81-Eme1 complex resolves Holliday junctions by a nick and counternick mechanism. *Mol. Cell*, **12**, 747–759.
109. Canas,C., Carrasco,B., Ayora,S. and Alonso,J.C. (2008) The RecU Holliday junction resolvase acts at early stages of homologous recombination. *Nucleic Acids Res.*, **36**, 5242–5249.
110. Hadden,J.M., Declais,A.C., Carr,S.B., Lilley,D.M. and Phillips,S.E. (2007) The structural basis of Holliday junction resolution by T7 endonuclease I. *Nature*, **449**, 621–624.
111. Song,J. and Markley,J.L. (2007) Three-dimensional structure determined for a subunit of human tRNA splicing endonuclease (Sen15) reveals a novel dimeric fold. *J. Mol. Biol.*, **366**, 155–164.
112. Tsutakawa,S.E., Jingami,H. and Morikawa,K. (1999) Recognition of a TG mismatch: the crystal structure of very short patch repair endonuclease in complex with a DNA duplex. *Cell*, **99**, 615–623.
113. Taylor,G.K., Heiter,D.F., Pietrokovski,S. and Stoddard,B.L. (2011) Activity, specificity and structure of I-Bth0305I: a representative of a new homing endonuclease family. *Nucleic Acids Res.*, **39**, 9705–9719.
114. Ronning,D.R., Li,Y., Perez,Z.N., Ross,P.D., Hickman,A.B., Craig,N.L. and Dyda,F. (2004) The carboxy-terminal portion of TnsC activates the Tn7 transposase through a specific interaction with TnsA. *EMBO J.*, **23**, 2972–2981.
115. Ramaswamy,K.S., Carrasco,C.D., Fatma,T. and Golden,J.W. (1997) Cell-type specificity of the Anabaena fdxN-element rearrangement requires xisH and xisI. *Mol. Microbiol.*, **23**, 1241–1249.
116. Babu,M., Beloglazova,N., Flick,R., Graham,C., Skarina,T., Nocek,B., Gagarinova,A., Pogoutse,O., Brown,G., Binkowski,A. *et al.* (2011) A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. *Mol. Microbiol.*, **79**, 484–502.
117. Kowalczykowski,S.C., Dixon,D.A., Eggleston,A.K., Lauder,S.D. and Rehrauer,W.M. (1994) Biochemistry of homologous recombination in Escherichia coli. *Microbiol. Rev.*, **58**, 401–465.
118. Yu,M., Souaya,J. and Julin,D.A. (1998) Identification of the nuclease active site in the multifunctional RecBCD enzyme by creation of a chimeric enzyme. *J. Mol. Biol.*, **283**, 797–808.
119. Zhang,J., McCabe,K.A. and Bell,C.E. (2011) Crystal structures of lambda exonuclease in complex with DNA suggest an electrostatic ratchet mechanism for processivity. *Proc. Natl Acad. Sci. USA*, **108**, 11872–11877.
120. Zhang,J., Xing,X., Herr,A.B. and Bell,C.E. (2009) Crystal structure of E. coli RecE protein reveals a toroidal tetramer for processing double-stranded DNA breaks. *Structure*, **17**, 690–702.
121. Burgers,P.M., Stith,C.M., Yoder,B.L. and Sparks,J.L. (2010) Yeast exonuclease 5 is essential for mitochondrial genome maintenance. *Mol. Cell Biol.*, **30**, 1457–1466.
122. Zhao,L., Bonocora,R.P., Shub,D.A. and Stoddard,B.L. (2007) The restriction fold turns to the dark side: a bacterial homing endonuclease with a PD-(D/E)-XK motif. *EMBO J.*, **26**, 2432–2442.
123. Zaros,C., Briand,J.F., Boulard,Y., Labarre-Mariotte,S., Garcia-Lopez,M.C., Thuriaux,P. and Navarro,F. (2007) Functional organization of the Rpb5 subunit shared by the three yeast RNA polymerases. *Nucleic Acids Res.*, **35**, 634–647.
124. Morin,B., Coutard,B., Lelke,M., Ferron,F., Kerber,R., Jamal,S., Frangeul,A., Baronti,C., Charrel,R., de Lamballerie,X. *et al.* (2010) The N-terminal domain of the arenavirus L protein is an RNA endonuclease essential in mRNA transcription. *PLoS Pathog.*, **6**, e1001038.
125. Salvato,M., Shimomaye,E. and Oldstone,M.B. (1989) The primary structure of the lymphocytic choriomeningitis virus L gene encodes a putative RNA polymerase. *Virology*, **169**, 377–384.
126. Ren,B., Kuhn,J., Meslet-Cladiere,L., Briffotaux,J., Norais,C., Lavigne,R., Flament,D., Ladenstein,R. and Myllykallio,H. (2009) Structure and function of a novel endonuclease acting on branched DNA substrates. *EMBO J.*, **28**, 2479–2489.
127. Tsodikov,O.V., Enzlin,J.H., Scharer,O.D. and Ellenberger,T. (2005) Crystal structure and DNA binding functions of ERCC1, a subunit of the DNA structure-specific endonuclease XPF-ERCC1. *Proc. Natl Acad. Sci. USA*, **102**, 11236–11241.
128. Xu,H., Swoboda,I., Bhalla,P.L., Sijbers,A.M., Zhao,C., Ong,E.K., Hoeijmakers,J.H. and Singh,M.B. (1998) Plant homologue of human excision repair gene ERCC1 points to conservation of DNA repair mechanisms. *Plant J.*, **13**, 823–829.
129. Reguera,J., Weber,F. and Cusack,S. (2010) Bunyaviridae RNA polymerases (L-protein) have an N-terminal, influenza-like endonuclease domain, essential for viral cap-dependent transcription. *PLoS Pathog.*, **6**, e1001101.
130. Zhao,C., Lou,Z., Guo,Y., Ma,M., Chen,Y., Liang,S., Zhang,L., Chen,S., Li,X., Liu,Y. *et al.* (2009) Nucleoside monophosphate complex structures of the endonuclease domain from the influenza virus polymerase PA subunit reveal the substrate binding site inside the catalytic center. *J. Virol.*, **83**, 9024–9030.
131. Nakagawa,Y., Oda,K. and Nakada,S. (1996) The PB1 subunit alone can catalyze cRNA synthesis, and the PA subunit in addition to the PB1 subunit is required for viral RNA synthesis in replication of the influenza virus genome. *J. Virol.*, **70**, 6390–6394.
132. Menon,S.K., Eilers,B.J., Young,M.J. and Lawrence,C.M. (2010) The crystal structure of D212 from sulfolobus spindle-shaped virus ragged hills reveals a new member of the PD-(D/E)XK nuclease superfamily. *J. Virol.*, **84**, 5890–5897.
133. Durr,H., Flaus,A., Owen-Hughes,T. and Hopfner,K.P. (2006) Snf2 family ATPases and DExx box helicases: differences and unifying concepts from high-resolution crystal structures. *Nucleic Acids Res.*, **34**, 4160–4167.
134. Margelevicius,M., Laganeckas,M. and Venclovas,C. (2010) COMA server for protein distant homology search. *Bioinformatics*, **26**, 1905–1906.
135. Kang,Y.H., Lee,C.H. and Seo,Y.S. (2010) Dna2 on the road to Okazaki fragment processing and genome stability in eukaryotes. *Crit. Rev. Biochem. Mol. Biol.*, **45**, 71–96.
136. Kang,H.Y., Choi,E., Bae,S.H., Lee,K.H., Gim,B.S., Kim,H.D., Park,C., MacNeill,S.A. and Seo,Y.S. (2000) Genetic analyses of Schizosaccharomyces pombe dna2(+) reveal that dna2 plays an essential role in Okazaki fragment metabolism. *Genetics*, **155**, 1055–1067.
137. Budd,M.E., Antoshechkin,I.A., Reis,C., Wold,B.J. and Campbell,J.L. (2011) Inviability of a DNA2 deletion mutant is due to the DNA damage checkpoint. *Cell Cycle*, **10**, 1690–1698.

138. Merkel,W.K. and Nichols,B.P. (1996) Characterization and sequence of the Escherichia coli panBCD gene cluster. *FEMS Microbiol. Lett.*, **143**, 247–252.
139. Desai,B.V. and Morrison,D.A. (2006) An unstable competence-induced protein, CoiA, promotes processing of donor DNA after uptake during genetic transformation in Streptococcus pneumoniae. *J. Bacteriol.*, **188**, 5177–5186.
140. Schonauer,M.S., Kastaniotis,A.J., Hiltunen,J.K. and Dieckmann,C.L. (2008) Intersection of RNA processing and the type II fatty acid synthesis pathway in yeast mitochondria. *Mol. Cell Biol.*, **28**, 6646–6657.
141. Otero,J.H., Suo,J., Gordon,C. and Chang,E.C. (2010) Int6 and Moe1 interact with Cdc48 to regulate ERAD and proper chromosome segregation. *Cell Cycle*, **9**, 147–161.
142. MacKay,C., Declais,A.C., Lundin,C., Agostinho,A., Deans,A.J., MacArtney,T.J., Hofmann,K., Gartner,A., West,S.C., Helleday,T. *et al.* (2010) Identification of KIAA1018/FAN1, a DNA repair nuclease recruited to DNA damage by monoubiquitinated FANCD2. *Cell*, **142**, 65–76.
143. Slupska,M.M., Chiang,J.H., Luther,W.M., Stewart,J.L., Amii,L., Conrad,A. and Miller,J.H. (2000) Genes involved in the determination of the rate of inversions at short inverted repeats. *Genes Cells*, **5**, 425–437.
144. Arahal,D.R., Lekunberri,I., Gonzalez,J.M., Pascual,J., Pujalte,M.J., Pedros-Alio,C. and Pinhassi,J. (2007) Neptuniibacter caesariensis gen. nov., sp. nov., a novel marine genome-sequenced gammaproteobacterium. *Int. J. Syst. Evol. Microbiol.*, **57**, 1000–1006.
145. Thrash,J.C., Cho,J.C., Vergin,K.L., Morris,R.M. and Giovannoni,S.J. (2010) Genome sequence of Lentisphaera araneosa HTCC2155T, the type species of the order Lentisphaerales in the phylum Lentisphaerae. *J. Bacteriol.*, **192**, 2938–2939.
146. Kosinski,J., Feder,M. and Bujnicki,J.M. (2005) The PD-(D/E)XK superfamily revisited: identification of new members among proteins involved in DNA metabolism and functional predictions for domains of (hitherto) unknown function. *BMC Bioinformatics*, **6**, 172.
147. Takeda,K., Akimoto,C. and Kawamukai,M. (2001) Effects of the Escherichia coli sfsA gene on mal genes expression and a DNA binding activity of SfsA. *Biosci. Biotechnol. Biochem.*, **65**, 213–217.
148. Nascimento,R., Dias,J.D. and Parkhouse,R.M. (2009) The conserved UL24 family of human alpha, beta and gamma herpesviruses induces cell cycle arrest and inactivation of the cyclinB/cdc2 complex. *Arch. Virol.*, **154**, 1143–1149.
149. Lymberopoulos,M.H. and Pearson,A. (2007) Involvement of UL24 in herpes-simplex-virus-1-induced dispersal of nucleolin. *Virology*, **363**, 397–409.
150. Bertrand,L., Leiva-Torres,G.A., Hyjazie,H. and Pearson,A. (2010) Conserved residues in the UL24 protein of herpes simplex virus 1 are important for dispersal of the nucleolar protein nucleolin. *J. Virol.*, **84**, 109–118.
151. Simarro,M., Gimenez-Cassina,A., Kedersha,N., Lazaro,J.B., Adelmant,G.O., Marto,J.A., Rhee,K., Tisdale,S., Danial,N., Benarafa,C. *et al.* (2010) Fast kinase domain-containing protein 3 is a mitochondrial protein essential for cellular respiration. *Biochem. Biophy, Res. Commun.*, **401**, 440–446.
152. Katiyar-Agarwal,S., Gao,S., Vivian-Smith,A. and Jin,H. (2007) A novel class of bacteria-induced small RNAs in Arabidopsis. *Genes Dev.*, **21**, 3123–3134.
153. Guzzo,C.R., Nagem,R.A., Barbosa,J.A. and Farah,C.S. (2007) Structure of Xanthomonas axonopodis pv. citri YaeQ reveals a new compact protein fold built around a variation of the PD-(D/E)XK nuclease motif. *Proteins*, **69**, 644–651.
154. Wong,K.R., Hughes,C. and Koronakis,V. (1998) A gene, yaeQ, that suppresses reduced operon expression caused by mutations in the transcription elongation gene rfaH in Escherichia coli and Salmonella typhimurium. *Mol. Gen. Genet.*, **257**, 693–696.
155. Vicari,D. and Artsimovitch,I. (2004) Virulence regulators RfaH and YaeQ do not operate in the same pathway. *Mol. Genet. Genomics*, **272**, 489–496.
156. Zou,X., Caufield,P.W., Li,Y. and Qi,F. (2001) Complete nucleotide sequence and characterization of pUA140, a cryptic plasmid from Streptococcus mutans. *Plasmid*, **46**, 77–85.
157. Nunez,B. and De La Cruz,F. (2001) Two atypical mobilization proteins are involved in plasmid CloDF13 relaxation. *Mol. Microbiol.*, **39**, 1088–1099.
158. Cantalupo,G., Bucci,C., Salvatore,P., Pagliarulo,C., Roberti,V., Lavitola,A., Bruni,C.B. and Alifano,P. (2001) Evolution and function of the neisserial dam-replacing gene. *FEBS Lett.*, **495**, 178–183.
159. Yamamoto,T., Matsuda,T., Inoue,T., Matsumura,H., Morikawa,M., Kanaya,S. and Kai,Y. (2006) Crystal structure of TBP-interacting protein (Tk-TIP26) and implications for its inhibition mechanism of the interaction between TBP and TATA-DNA. *Protein Sci.*, **15**, 152–161.
160. Stern,A. and Sorek,R. (2011) The phage-host arms race: shaping the evolution of microbes. *Bioessays*, **33**, 43–51.
161. Harrison,A., Pearl,F., Mott,R., Thornton,J. and Orengo,C. (2002) Quantifying the similarities within fold space. *J. Mol. Biol.*, **323**, 909–926.
162. Krishna,S.S. and Grishin,N.V. (2005) Structural drift: a possible path to protein fold change. *Bioinformatics*, **21**, 1308–1310.
163. Levitt,M. (2009) Nature of the protein universe. *Proc. Natl Acad. Sci. USA*, **106**, 11079–11084.