BMC Structural Biology

BioMed Central

Research article

Open Access

Realm of PD-(D/E)XK nuclease superfamily revisited: detection of novel families with modified transitive meta profile searches

Lukasz Knizewski[1], Lisa N Kinch[2], Nick V Grishin[2], Leszek Rychlewski[3] and Krzysztof Ginalski*[1]

Address: [1]Interdisciplinary Centre for Mathematical and Computational Modelling, Warsaw University, Pawinskiego 5a, 02-106 Warsaw, Poland, [2]Howard Hughes Medical Institute and Department of Biochemistry, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390-9050, USA and [3]BioInfoBank Institute, Limanowskiego 24a, 60-744 Poznan, Poland

Email: Lukasz Knizewski - lukas@icm.edu.pl; Lisa N Kinch - lkinch@chop.swmed.edu; Nick V Grishin - grishin@chop.swmed.edu; Leszek Rychlewski - leszek@bioinfo.pl; Krzysztof Ginalski* - kginal@icm.edu.pl

* Corresponding author

Published: 20 June 2007

BMC Structural Biology 2007, **7**:40 doi:10.1186/1472-6807-7-40

Received: 13 March 2007
Accepted: 20 June 2007

This article is available from: http://www.biomedcentral.com/1472-6807/7/40

Abstract

Background: PD-(D/E)XK nucleases constitute a large and highly diverse superfamily of enzymes that display little sequence similarity despite retaining a common core fold and a few critical active site residues. This makes identification of new PD-(D/E)XK nuclease families a challenging task as they usually escape detection with standard sequence-based methods. We developed a modified transitive meta profile search approach and to consider the structural diversity of PD-(D/E)XK nuclease fold more thoroughly we analyzed also lower than threshold Meta-BASIC hits to select potentially correct predictions placed among unreliable or incorrect ones.

Results: Application of a modified transitive Meta-BASIC searches on updated PFAM families and PDB structures resulted in detection of five new PD-(D/E)XK nuclease families encompassing hundreds of so far uncharacterized and poorly annotated proteins. These include four families catalogued in PFAM database as domains of unknown function (DUF506, DUF524, DUF1626 and DUF1703) and YhgA-like family of putative transposases. Three of these families represent extremely distant homologs (DUF506, DUF524, and YhgA-like), while two are newly defined in updated database (DUF1626 and DUF1703). In addition, we also confidently identified an extended AAA-ATPase domain in the N-terminal region of DUF1703 family proteins.

Conclusion: Obtained results suggest that detailed analysis of below threshold Meta-BASIC hits may push limits further for distant homology detection in the 'midnight zone' of homology. All identified families conserve the core evolutionary fold, secondary structure and hydrophobic patterns common to existing PD-(D/E)XK nucleases and maintain critical active site motifs that contribute to nucleic acid cleavage. Further experimental investigations should address the predicted activity and clarify potential substrates providing further insight into detailed biological role of these newly detected nucleases.

Background

Restriction endonuclease-like proteins, also called a PD-(D/E)XK nucleases, constitute a large and diverse superfamily of enzymes that are involved in numerous nucleic acid cleavage events important for various cellular processes. The SCOP [1] database currently groups 23 families of known structure in the restriction endonuclease-like superfamily, including among others 15 different restriction endonucleases [2], holiday junction resolvases (endonuclease I, Hjc) [3,4], lambda exonuclease [5] and very short patch repair (Vsr) endonuclease [6]. Their function varies from repairing damaged DNA (Vsr), resolving holliday junctions (endonuclease I, Hjc), performing additional cleavage events in DNA recombination (lambda exonuclease), to protection of host organisms against foreign DNA invasion (restriction endonucleases).

Despite displaying very little sequence similarity, the restriction endonuclease-like enzymes retain a common core fold that consists of a central four-stranded mixed β-sheet flanked by an α-helix on either side (with αβββαβ topology). The general architecture of the restriction endonuclease-like fold allows recognition of diverse nucleic acid substrates, which may vary from specific palindromic DNA sequences (type II restriction endonuclease and Vsr) to unique DNA backbone structures (Hjc). The substrate specificity in many cases arises from various insertions to the core fold that may even encompass entire domains [7,8].

The restriction endonuclease-like superfamily possesses a relatively conserved active site PD-(D/E)XK signature (motif II and motif III), critical for cleaving the nucleic acid phosphodiester bond [9-11]. The signature lysine residue is responsible for positioning water for in-line attack on the substrate phosphodiester bond, while the carboxylates coordinate up to three metal ions that serve as cofactors in the reaction. Several variations to this named motif exist, and additional family conserved charged or polar residues located in core helices of the fold (motif I and motif IV) contribute to active site architecture in many cases [11].

Standard homology detection methods usually fail to recognize novel PD-(D/E)XK nucleases due to lack of significant sequence similarity between families and numerous insertions to the core fold. In previous work we applied a transitive search approach using the meta profile comparison method Meta-BASIC and identified nine new restriction endonuclease-like fold families among hypothetical proteins [12]. Some of these families were also detected by others using HHsearch method [13]. In this study we employed a modified transitive search procedure by including additional below threshold score Meta-BASIC hits in updated PFAM and PDB databases and identified five more novel PD-(D/E)XK nuclease families.

Results and discussion

This study is a continuation of our previous work identifying novel restriction endonuclease-like fold families among catalogued PFAM families (mainly domains of unknown function, DUFs) using transitive searches with the Meta-BASIC method. Although existing restriction endonuclease-like structures retain similar active site residues within the same core fold (with αβββαβ topology) (Figure 1), they exhibit extreme structural diversity (structure comparison scores can be below threshold [12]). Since Meta-Basic scores are benchmarked using rigorous structural criteria to define confidence thresholds (predictions with Z-score above 12 have <5% probability of being incorrect [14]), the structural diversity of PD-(D/E)XK nuclease fold can be reflected as lower than threshold Meta-BASIC scores. Accordingly, we modified our previous transitive search approach to consider potentially correct Meta-BASIC predictions placed (according to Z-score) among unreliable or incorrect ones. While this method extension provides an effective strategy for detecting extremely distant homologs, the resulting non-trivial predictions require additional criteria for justification. Therefore, we demand that family profiles of low-scoring hits retain all core secondary structure elements comprising the nuclease fold and conserve critical functional residues essential for function. Confirmation of these predictions by a consensus of fold recognition methods, such as those produced by 3D-Jury must also support the non-trivial links. Consequently, an extensive search of the updated GRDB database, which stores Meta-BASIC connections between PFAM families and PDB structures, led to identification of five new putative PD-(D/E)XK nuclease families (not detectable with standard sequence search methods) encompassing hundreds of so far uncharacterized proteins of unknown function. Three of these families represent extremely distant homologs (DUF506, DUF524, and YhgA-like), while two are newly defined in the updated GRDB system (DUF1626 and DUF1703). In addition to conserving critical features of the restriction endonuclease-like fold, all families display similar hydrophobicity patterns to known PD-(D/E)XK nucleases (Figure 2). All predictions are discussed in details below.

DUF506

DUF506 family (PF04720) embraces a number of plant proteins including 18 copies from *Oryza sativa*, 23 copies from *Arabidopsis thaliana* and single sequences from *Pinus pinaster* and *Medicago truncatula.* Meta-BASIC mapped the consensus sequence of DUF506 onto holliday junction resolvase (Hjc) structure (pdb:1hh1[15], Meta-BASIC score 8.2) and PFAM family UPF0102 (PF02021, Meta-BASIC score 7.6), which has been predicted as distantly

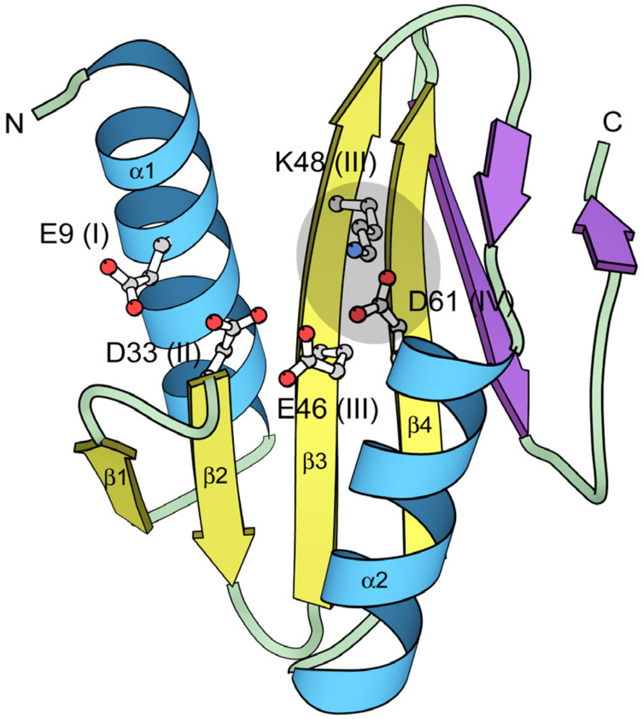


Figure 1
Typical PD-(D/E)XK nuclease structure. Ribbon representation for archaeal holliday junction resolvase (Hjc, pdb:1gef) with the critical motif I (E9), motif II (D33), motif III (E46, K48) and motif IV (D61) active site residues shown as balls-and-sticks. Secondary structure elements of the conserved core fold are labeled and colored blue (α-helix) and yellow (β-strand), while additional insertions to the core are colored violet. In several detected families, conserved motif III lysine migrates to motif IV where (frequently mutated to arginine) it may occupy a similar spatial position (highlighted in grey) and thus play an equivalent functional role.

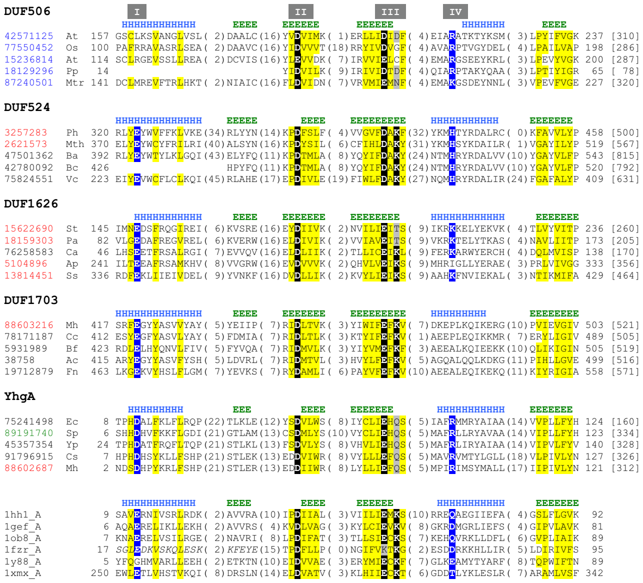


Figure 2
Novel PD-(D/E)XK nuclease families. Multiple sequence alignment of representative sequences from newly detected families and selected PD-(D/E)XK nuclease structures for the conserved structural core. Sequences are labeled according to gi number or PDB code and an abbreviation of the species name: Ac *Azotobacter chroococcum*, Ap *Aeropyrum pernix*, At *Arabidopsis thaliana*, Ba *Bacillus anthracis*, Bc *Bacillus cereus*, Bf *Bacteroides fragilis*, Ca *Chloroflexus aurantiacus*, Cs *Chromohalobacter salexigens*, Ct *Clostridium tetani*, Ec *Escherichia coli*, Fn *Fusobacterium nucleatum*, Mh *Methanospirillum hungatei*, Mtr *Medicago truncatula*, Mth *Methanothermobacter thermautotrophicus*, Os *Oryza sativa*, Pa *Pyrobaculum aerophilum*, Ph *Pyrococcus horikoshii*, Pp *Pinus pinaster*, Sp *Sodalis phage*, Ss *Sulfolobus solfataricus*, St *Sulfolobus tokodaii*, Vc *Vibrio cholerae*, Yp *Yersinia pesti*. Sequence gi numbers are colored according to taxonomy: with bacterial in black, archeal in red, eukaryotic in blue and viral in green. The first and last residues numbers are indicated before and after each sequence with total sequence length following in square bracket. The numbers of excluded residues are specified in parentheses. Residue conservation is denoted with following scheme: uncharged, highlighted in yellow; charged or polar highlighted in grey; small, letters in red. Active site PD-(D/E)XK signature residues are highlighted in black while other conserved polar/charged residues in alternative active site positions are highlighted in blue. Restriction endonuclease-like motifs (I-IV) are labeled on the top. Predicted (gi:42571125, gi:3257283, gi:15622690, gi:88603216 and gi:75241498) and observed (pdb:1hh1) secondary structure elements (E, β-strand; H, α-helix) are indicated above the sequences. Italicized sequence corresponds to domain-swapped region of 1fzr.

related to Hjc. According to our benchmarks, predictions with Z-score >8 have <10% probability of being incorrect. In addition, several fold recognition servers also selected restriction endonuclease-like fold among first 5 models. DUF506 proteins possess a predicted conserved core αα$\underline{\alpha\beta\beta\beta\alpha\beta}$αα secondary structure pattern (with putative restriction endonuclease-like core elements underlined). DUF506 family members retain the characteristic motif II (xDxxx motif located in the second core β-strand, where x is in general any hydrophobic residue) of the PD-(D/E)XK signature, with the exception of three conservative replacements of aspartate for glutamate in gi:15236814, gi:18399441 and gi:20147393. Motif III is represented by a (D/E)X(D/N/S/C/G) pattern (Figure 2). The missing positively charged residue in motif III is possibly replaced by a conserved arginine in motif IV (R218 gi:42571125) located in the proceeding α-helix (Figure 1). A similar migration of this critical active site residue can be observed for the DUF790 family [12]. Although one of the family members, *Pinus pinaster* protein (gi:18129296), has been reported to be expressed in native cells [16], this prediction represents the only functional information available for this family of sequences. DUF506 proteins lack any identified fused domains that might hint at biological function, and detailed analysis of a genomic context did not help identify potential physiological roles for the family.

DUF524

The DUF524 family (PF04411, COG1700) includes a number of hypothetical proteins of bacterial origin (in addition to two sequences from *Bacillus cereus* [17] with a fused McrB restriction GTPase domain that are annotated as 5-methylcytosine-specific restriction related enzymes) and two sequences from archeal species (*Pyrococcus horikoshii* and *Methanothermobacter thermautotrophicus*). The C-terminal region of DUF524 consensus sequence was mapped by Meta-BASIC onto the PD-(D/E)XK nuclease families DUF1064 (PF06356, Meta-BASIC Z-score 9.3) [12] and Type I restriction enzyme R protein N terminus (HSDR_N, PF04313, Meta-BASIC Z-score 8.8) [18]. Additional support for this prediction was obtained with 3D-Jury although with below threshold scores due to inconsistent alignments generated by servers.

DUF524 family proteins consist of at least two domains: a C-terminal PD-(D/E)XK nuclease domain and an N-terminal region of yet unknown function with a predicted all β secondary structure pattern followed by mainly α-helical structure. The DUF524 restriction endonuclease-like domain has two additional β-strands inserted to the core fold after the first core α-helix (αβββββαβ topology, conserved core elements are underlined). Similar insertion in this region can be found in BsoBI restriction endonuclease (pdb:1dc1) [19]. The PD-(D/E)XK signature is clearly conserved among DUF524 family members and corresponds to invariant PD (motif II) and DAK (motif III) motifs (Figure 2). Additionally, DUF524 proteins conserve a glutamic acid in motif I (E323 in gi:3257283), most likely involved in metal ion binding. Lastly, the second core α-helix contains an invariant MHXYRD motif (motif IV).

The COG corresponding to DUF524 (COG1700) has a confidently detected (STRING score 0.73) genomic neighbourhood association to a unique family of restriction endonuclease GTPase subunits (COG1401). These GTPases have been assigned to AAA+ class chaperonin-like ATPases [20] and include McrB of the *E. coli* methylation-dependent restriction system (McrBC). In this system, DNA cleavage by the McrC subunit is strictly coupled to GTP hydrolysis by the McrB subunit [21] instead of the typical ATP cofactor requirement of most restriction modification systems (for example Type I and Type III restriction endonucleases [22]. The McrC subunit responsible for cleavage is a PD-(D/E)XK endonuclease [21], which supports assignment of DUF524 to the restriction endonuclease-like superfamily and suggests a function of methylation-dependent restriction for this group of unknown proteins.

DUF1626

DUF1626 family (PF07788, COG5493) includes 19 proteins from certain archeal (*Sulfolobus tokodaii*, *Sulfolobus solfataricus*, *Sulfolobus acidocaldarius*, *Pyrobaculum aerophilum*, *Thermofilum pendens* and *Aeropyrum pernix*) and bacterial (*Chloroflexus aurantiacus*, *Roseiflexus sp.*, *Candidatus Kuenenia stuttgartiensis* and delta proteobacterium MLMS-1) organisms. The consensus sequence of this family was mapped with an above threshold scores to both PD-(D/E)XK PFAM families: DUF91 (PF01939, Meta-BASIC score 17.2) [12], restriction endonuclease (PF04471, Meta-BASIC score 14.9) and PDB structures: holliday junction resolvases Hje (pdb:1ob8[23], Meta-BASIC score 13.7) and Hjc (pdb:1hh1, Meta-BASIC score 12.1).

Majority of DUF1626 proteins possess an additional N-terminal α-helical region, mainly coiled-coil (as predicted with Coils2 [24]) and are frequently annotated as tropomyosin, coiled-coil or microtubule binding proteins. Specifically, RPS-BLAST searches of the Conserved Domain Database (CDD) [25] detect sequence similarity to several coiled-coil containing families, although the repeated sequence patterns found in coiled-coils render this similarity unreliable for any type of functional or evolutionary assumptions.

The DUF1626 restriction endonuclease-like domain has predicted αβββαβαβ topology (with conserved endonuclease elements underlined), where the C-terminal elements possibly extend the domain core. Motif III lysine of the PD-(D/E)XK nuclease fingerprint is often substituted by threonine (Figure 2) and it is likely that, similarly to DUF506 or DUF790 [12], the lysine migrated to an α-helix (motif IV) following the third core β-strand (Figure 1). Specifically, DUF1626 proteins with threonine in motif III possess a conserved patch of positively charged lysine and arginine residues in the motif IV α-helix that might be involved in substrate binding or contribute to active site formation.

DUF1703

The DUF1703 family (NCOG44579) groups together a set of uncharacterized proteins from one archeal (*Methanospirillum hungatei*) and various bacterial organisms. The C-terminal region of DUF1703 consensus sequence has detectable similarity to PD-(D/E)XK nucleases (DUF91, above threshold Meta-BASIC score 16.7). Specifically, the DUF1703 C-terminal domain has the predicted secondary structure pattern of the restriction endonuclease-like fold core with an additional β-strand at C-terminus (αβββαββ, nuclease core underlined) and conserves all restriction endonuclease-like superfamily motifs (Figure 2). These include both PD-(D/E)XK signature motifs II and III as well as two other family conserve positions: a charged aspartate residue in the first core α-helix (E420 in gi:88603216, motif I), presumably responsible for metal ion binding, and a glutamine residue in the second core α-helix (Q480 in gi:88603216, motif IV), that may take part in binding of nucleic acid substrate.

Meta-BASIC linked the DUF1703 N-terminal region to an archeal ATPase family (PF01637, above threshold Meta-BASIC score 20.6). The initial assignment was further supported with 3D-Jury that mapped this region with above threshold scores (>50) to several structures from the SCOP extended AAA-ATPases (AAA+) family of P-loop containing nucleoside triphosphate hydrolases fold (for instance pdb:1fnn[26], pdb:1sxj[27], pdb:1iqp[28], pdb:1jr3[29]). The DUF1703 AAA+ module consists of two domains: an N-terminal α/β domain (referred as domain 1) [20] with characteristic Walker A and Walker B motifs, and a small anti-parallel four-helix bundle domain (referred as domain 2). In typical AAA+ structures, residues from the Walker A motif (GXXGXGK(T/S)) and the Walker B motif (referred as DEAD or DEXX motif) together bind nucleoside and Mg^{2+} in the deep cleft between two domains. While DUF1703 family proteins possess classical Walker B motif (DEYD), residues in the Walker A motif differ from the canonical definition: arginines replace the first and second glycines i.e. 44RPRRFGKS51 in gi:88603216 (described residues underlined) (Figure 3). Detailed analysis of AAA+ family sequences and structures revealed that similar changes to Walker A motif are possible (for instance GXX**R**XGKT in pdb:1v5w[30] and **L**XXSXGRS in pdb:1rif[31]), but they usually correlate with mutations in other positions in order to maintain active site pocket function, shape and accessibility. Other defined AAA+ family elements, such as the Sensor 1 and Sensor 2 regions [20] that help catalyze hydrolysis lack corresponding functional residues in DUF1703 sequences (Figure 3) have diverged in this family. Specifically, the Sensor 1 region (194FLTGKVS199 in gi:88603216), situated in a helical turn after strand 4 (Figure 3), conserves a glycine residue (G197 in gi:88603216) instead of a polar amino acid, while the Sensor 2 region (260GQQVYNP266 in gi:88603216), located at the N-termini of helix 7, lacks the typical arginine finger. We hypothesize that second and third conserved arginines in the DUF1703 family Walker A motif (44RPRRFGKS51 in gi:88603216) substitute for the Sensor I polar residue and the Sensor II arginine, respectively (Figure 3).

In summary, DUF1703 family proteins share a common four-domain structure, with an N-terminal AAA+ module (domains 1 and 2), a C-terminal restriction endonuclease-like domain and a small α-helical region (according to secondary structure predictions) in-between. Similar domain architecture can be found for other PD-(D/E)XK nucleases (for example, Mrr restriction endonuclease fused to NACHT ATPase domain in gi:68548712), providing further support for functional predictions.

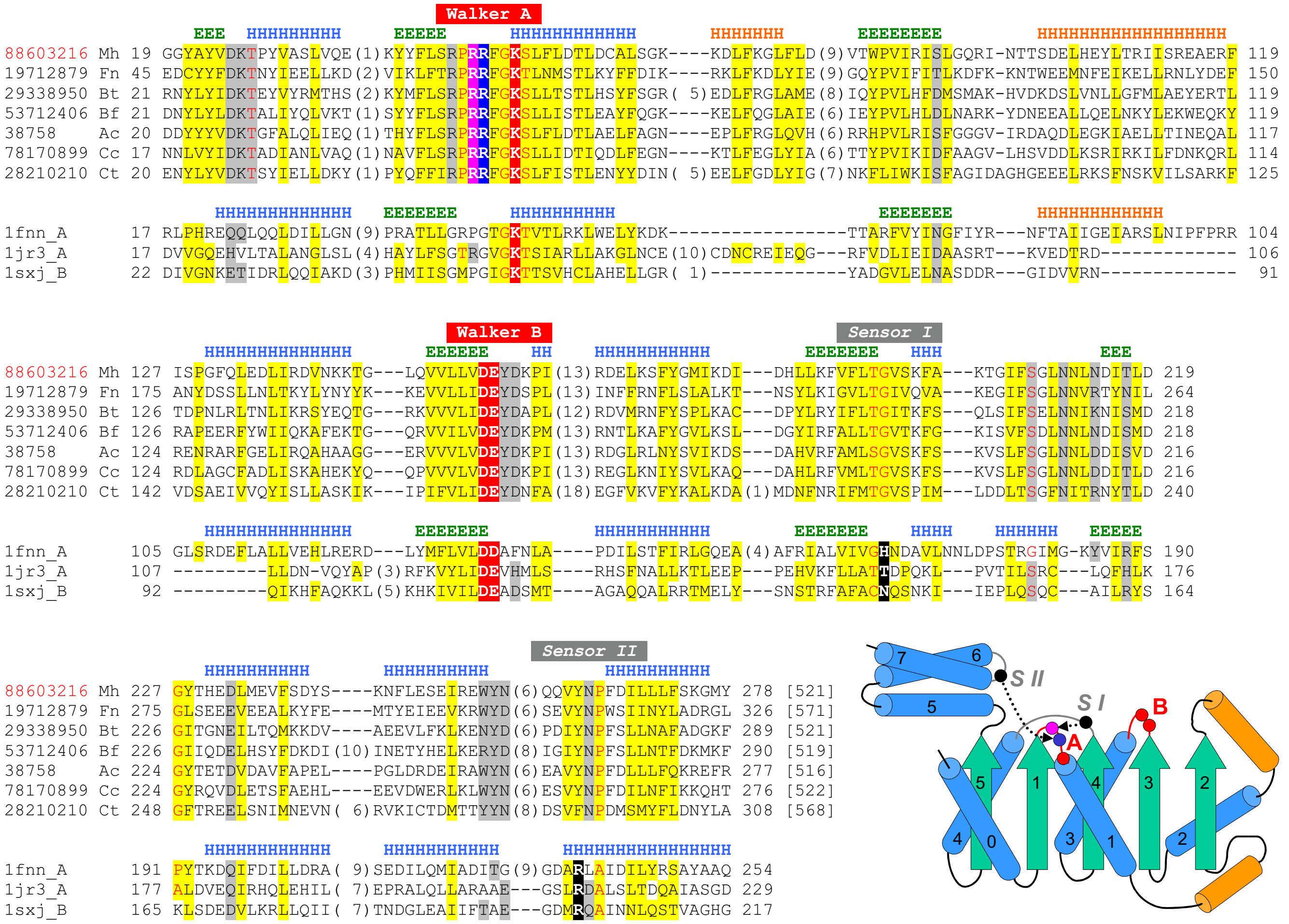
YhgA-like

The YhgA-like family (PF04754, COG5464) of putative transposases encompasses hundreds of bacterial proteins in addition to a few archeal (*Methanospirillum hungatei*) and one viral sequence from *Sodalis phage*. These proteins are assigned to the PD-(D/E)XK nuclease superfamily based on a detected Meta-BASIC connection to herpes virus protein UL24 (PF01646, Meta-BASIC Z-scores 11.6 for whole length consensus YhgA-like family sequence and above threshold 15.3 for selected N-terminal region encompassing putative restriction endonuclease-like domain), which has been recently identified as an additional member of this superfamily [32].

The predicted PD-(D/E)XK nuclease domain resides in the N-terminal region of YhgA-like proteins and displays a common predicted secondary structure αααβαββαβ pattern (putative restriction endonuclease-like core elements underlined), with two α-helical insertions to the core fold. Similar insertions of two α-helices before and a single α-helix after the first core β-strand can be found in RecB structure (pdb:1w36) [33]. In the YhgA-like family, the predicted restriction endonuclease-like domain is followed by a C-terminal α-helical region (~150 aa).

The YhgA-like family members clearly conserve active site motifs II and III (Figure 2), where motif III is identical to that of the Coi-A-like family (EXQ in the third core β-strand). Additional charged residues include an invariant motif I putative metal ion binding glutamate (D11 in gi:75241498) [8] and a motif IV arginine (R94 in gi:75241498). Importantly, a few bacterial sequences have glutamine substituted by lysine in motif III, nevertheless, glutamine is found in motif IV instead of conserved arginine. This evident sequence correlation in critical active site positions further stresses the correctness of the prediction for the YhgA-like family.

Using COG5464 (YhgA-like) as an input, the STRING database assigns a high confidence combined neighborhood and domain fusion score (0.755) to a group of uncharacterized cyanobacterial proteins (COG4636) that correspond to DUF820. DUF820 was previously identified as a PD-(D/E)XK nuclease (with a representative structure pdb:1wdj) [12], suggesting that the two families arose from a genetic duplication and may perform similar functions.



```
                                          Walker A
                  EEE  HHHHHHHHH     EEEEE      HHHHHHHHHHHH       HHHHHHH       EEEEEEEE          HHHHHHHHHHHHHHHHHH
88603216 Mh 19 GGYAYVDKTPYVASLVQE(1)KYYFLSRPRRFGKSLFLDTLDCALSGK----KDLFKGLFLD(9)VTWPVIRISLGQRI-NTTSDELHEYLTRIISREAERF 119
19712879 Fn 45 EDCYYFDKTNYIEELLKD(2)VIKLFTRPRRFGKTLNMSTLKYFFDIK----RKLFKDLYIE(9)GQYPVIFITLKDFK-KNTWEEMNFEIKELLRNLYDEF 150
29338950 Bt 21 RNYLYIDKTEYVYRMTHS(2)KYMFLSRPRRFGKSLLTSTLHSYFSGR( 5)EDLFRGLAME(8)IQYPVLHFDMSMAK-HVDKDSLVNLLGFMLAEYERTL 119
53712406 Bf 21 DNYLYLDKTALIYQLVKT(1)SYYFLSRPRRFGKSLLISTLEAYFQGK----KELFQGLAIE(6)IEYPVLHLDLNARK-YDNEEALLQELNKYLEKWEQKY 119
38758    Ac 20 DDYYYVDKTGFALQLIEQ(1)THYFLSRPRRFGKSLFLDTLAELFAGN----EPLFRGLQVH(6)RRHPVLRISFGGGV-IRDAQDLEGKIAELLTINEQAL 117
78170899 Cc 17 NNLVYIDKTADIANLVAQ(1)NAVFLSRPRRFGKSLLIDTIQDLFEGN----KTLFEGLYIA(6)TTYPVIKIDFAAGV-LHSVDDLKSRIRKILFDNKQRL 114
28210210 Ct 20 ENYLYVDKTSYIELLDKY(1)PYQFFIRPRRFGKSLFISTLENYYDIN( 5)EELFGDLYIG(7)NKFLIWKISFAGIDAGHGEEELRKSFNSKVILSARKF 125

                    HHHHHHHHHHHHHH  EEEEEEE     HHHHHHHHHH                         EEEEEEE         HHHHHHHHHHHH
1fnn_A      17 RLPHREQQLQQLDILLGN(9)PRATLLGRPGTGKTVTLRKLWELYKDK-----------------TTARFVYINGFIYR---NFTAIIGEIARSLNIPFPRR 104
1jr3_A      17 DVVGQEHVLTALANGLSL(4)HAYLFSGTRGVGKTSIARLLAKGLNCE(10)CDNCREIEQG---RFVDLIEIDAASRT---KVEDTRD------------- 106
1sxj_B      22 DIVGNKETIDRLQQIAKD(3)PHMIISGMPGIGKTTSVHCLAHELLGR( 1)-------------YADGVLELNASDDR---GIDVVRN-------------  91

                                          Walker B                     Sensor I
                   HHHHHHHHHHHHHHH      EEEEEE    HH    HHHHHHHHHHH         EEEEEEE   HHH               EEE
88603216 Mh 127 ISPGFQLEDLIRDVNKKTG---LQVVLLVDEYDKPI(13)RDELKSFYGMIKDI---DHLLKFVFLTGVSKFA---KTGIFSGLNNLNDITLD 219
19712879 Fn 175 ANYDSSLLNLTKYLYNYYK---KEVVLLIDEYDSPL(13)INFFRNFLSLALKT---NSYLKIGVLTGIVQVA---KEGIFSGLNNVRTYNIL 264
29338950 Bt 126 TDPNLRLTNLIKRSYEQTG---RKVVVLIDEYDAPL(12)RDVMRNFYSPLKAC---DPYLRYIFLTGITKFS---QLSIFSELNNIKNISMD 218
53712406 Bf 126 RAPEERFYWIIQKAFEKTG---QRVVILVDEYDKPM(13)RNTLKAFYGVLKSL---DGYIRFALLTGVTKFG---KISVFSDLNNLNDISMD 218
38758    Ac 124 RENRARFGELIRQAHAAGG---ERVVVLVDEYDKPI(13)RDGLRLNYSVIKDS---DAHVRFAMLSGVSKFS---KVSLFSGLNNLDDISVD 216
78170899 Cc 124 RDLAGCFADLISKAHEKYQ---QPVVVLVDEYDKPI(13)REGLKNIYSVLKAQ---DAHLRFVMLTGVSKFS---KVSLFSGLNNLDDITLD 216
28210210 Ct 142 VDSAEIVVQYISLLASKIK---IPIFVLIDEYDNFA(18)EGFVKVFYKALKDA(1)MDNFNRIFMTGVSPIM---LDDLTSGFNITRNYTLD 240

                   HHHHHHHHHHHHHHH     EEEEEEE          HHHHHHHHHHH        EEEEEEE    HHHH    HHHHHH   EEEEE
1fnn_A      105 GLSRDEFLALLVEHLRERD---LYMFLVLDDAFNLA----PDILSTFIRLGQEA(4)AFRIALVIVGHNDAVLNNLDPSTRGIMG-KYVIRFS 190
1jr3_A      107 ---------LLDN-VQYAP(3)RFKVYLIDEVHMLS----RHSFNALLKTLEEP---PEHVKFLLATTDPQKL---PVTILSRC---LQFHLK 176
1sxj_B       92 ---------QIKHFAQKKL(5)KHKIVILDEADSMT----AGAQQALRRTMELY---SNSTRFAFACNQSNKI---IEPLQSQC---AILRYS 164

                                                   Sensor II
                   HHHHHHHHHH       HHHHHHHHHH           HHHHHHHHHH
88603216 Mh 227 GYTHEDLMEVFSDYS----KNFLESEIREWYN(6)QQVYNPFDILLLFSKGMY 278 [521]
19712879 Fn 275 GLSEEEVEEALKYFE----MTYEIEEVKRWYD(6)SEVYNPWSIINYLADRGL 326 [571]
29338950 Bt 226 GITGNEILTQMKKDV----AEEVLFKLKENYD(6)PDIYNPFSLLNAFADGKF 289 [521]
53712406 Bf 226 GIIQDELHSYFDKDI(10)INETYHELKERYD(8)IGIYNPFSLLNTFDKMKF 290 [519]
38758    Ac 224 GYTETDVDAVFAPEL----PGLDRDEIRAWYN(6)EAVYNPFDLLLFQKREFR 277 [516]
78170899 Cc 224 GYRQVDLETSFAEHL----EEVDWERLKLWYN(6)ESVYNPFDILNFIKKQHT 276 [522]
28210210 Ct 248 GFTREELSNIMNEVN( 6)RVKICTDMTTYYN(8)DSVFNPDMSMYFLDNYLA 308 [568]

                   HHHHHHHHHHHH     HHHHHHHHHHH      HHHHHHHHHHHHHHHH
1fnn_A      191 PYTKDQIFDILLDRA( 9)SEDILQMIADITG(9)GDARLAIDILYRSAYAAQ 254
1jr3_A      177 ALDVEQIRHQLEHIL( 7)EPRALQLLARAAE---GSLRDALSLTDQAIASGD 229
1sxj_B      165 KLSDEDVLKRLLQII( 7)TNDGLEAIIFTAE---GDMRQAINNLQSTVAGHG 217
```

Figure 3

N-terminal extended AAA-ATPase (AAA+) domain in DUF1703 family. Multiple sequence alignment for the N-terminal region of representative sequences from DUF1703 family and selected AAA+ structures together with the schematic diagram of the fold. Sequences are labeled according to gi number or PDB code and an abbreviation of the species name: Ac *Azotobacter chroococcum*, Bf *Bacteroides fragilis*, Bt *Bacteroides thetaiotaomicron*, Cc *Chlorobium chlorochromatii*, Ct *Clostridium tetani*, Fn *Fusobacterium nucleatum*, Mh *Methanospirillum hungatei*. Sequence gi numbers are colored according to taxonomy: with bacterial in black and archeal in red. The first and last residues numbers are indicated before and after each sequence with total sequence length following in square bracket. The numbers of excluded residues are specified in parentheses. Residue conservation is denoted in following scheme: uncharged, highlighted in yellow; charged or polar highlighted in grey; small, letters in red. Invariant active site residues are highlighted in red, while additional active site residues that have migrated in DUF1703 family sequences from typical Sensor I and Sensor II positions (highlighted in black) are highlighted in pink and blue. AAA+ motifs (Walker A, Walker B, Sensor I and Sensor II) are labeled above corresponding residue columns. Locations of predicted (gi:88603216) and observed (pdb:1fnn) secondary structure elements (E, β-strand; H, α-helix) are marked above the sequences and are colored according to the schematic representation shown in the bottom right corner. Core AAA+ helices and strands are colored blue and green, respectively, while inserted helices are colored orange. Walker A and B motif loops are labeled and colored red, with invariant DUF1703 family residues that correspond to AAA+ active site residues depicted as red circles. Typical Sensor I and Sensor II sites are labeled and colored gray, with positions of functional sites depicted as black circles. Migrated DUF1703 residues (indicated by broken black arrows) that could substitute for Sensor I and Sensor II are denoted by pink and blue circles, respectively.

Conclusion

The PFAM database currently defines a PD-(D/E)XK nuclease superfamily clan that groups 15 different families: including restriction endonuclease, archaeal holliday junction resolvase (Hjc), RmuC, herpes virus protein UL24, competence protein CoiA-like, sugar fermentation stimulation protein, VRR-NUC, herpesvirus alkaline exonuclease, DUF91, DUF790, DUF911, DUF1016, DUF1052, DUF1064 and UPF0102. Many of these families were identified in our previous studies [12,32]. In this work we performed systematic searches in updated PFAM database with transitive Meta-BASIC approach to further expand the realm of PD-(D/E)XK nuclease superfamily. We analyzed below threshold Meta-BASIC predictions to identify correct hits that due to their large evolutionary distance were assigned below cut-off confidence scores. Selection of these highly non trivial but reliable assignments was based on consistency of a predicted secondary structure pattern with that of the restriction endonuclease-like fold core, general conservation of hydrophobicity patterns, and presence of the signature PD-(D/E)XK nuclease motifs critical for function. This strategy resulted in identification of five new PD-(D/E)XK nuclease families in the PFAM database (DUF506, DUF524, DUF1626, DUF1703 and YhgA-like) encompassing hundreds of uncharacterized or hypothetical proteins. Additionally, analysis of genomic context for examined families strengthened several of our predictions. Altogether, obtained results suggest that combination of transitive Meta-BASIC searches with various other analyses (including sequence conservation, secondary structure prediction, domain architecture and genomic context) of below threshold hits may push limits further for distant homology detection in the 'midnight zone' of homology. Finally, using top-of-the line fold recognition methods we also identified AAA+ domain in the N-terminal region of DUF1703 proteins that is not detectable by standard sequence comparison methods.

Methods

Detection of putative PD-(D/E)XK nuclease families

Identification of novel PD-(D/E)XK nuclease families was carried out using GRDB system [14], which includes precalculated Meta-BASIC connections between PFAM (version 20.0) [34] families and proteins of known structure (PDB) [35]. Meta-BASIC is a consensus meta profile alignment method capable of finding very distant similarity between proteins through a comparison of sequence profiles enriched by predicted secondary structures (meta profiles).

We applied a similar transitive Meta-BASIC search approach as in our recent work [12], which identified a number of PD-(D/E)XK nuclease families exhibiting below threshold (<12) scores (according to rigorous structural criteria, scores above 12 have <5% probability of being incorrect [14]). Considering the structural divergence of restriction endonuclease-like fold [12] that is likely reflected as lower than threshold Meta-BASIC scores, we extended our previous transitive search method to consider the top 200 ranked Meta-BASIC hits for each query (including those hits that rank lower than the first false positive). Additionally, the steady increase in protein database sizes (PFAM, PDB and NCBI nr) may lead to straightforward detection of novel restriction endonuclease-like enzymes that were not detectable before or not included in previous versions of the GRDB system.

Using the collective set of previously identified PD-(D/E)XK nuclease families as queries, we applied transitive Meta-BASIC searches to an updated GRDB system. Selection of potentially correct, yet highly diverged families among the numerous low scoring Meta-BASIC predictions was based on two essential defining criteria for PD-(D/E)XK nucleases. First, family profiles must include correctly aligned conserved acidic residues from motifs II and III that contribute to cleavage. Second, family profiles must include all secondary structure elements (predicted by PSI-PRED [36]) that correspond to the evolutionary core of PD-(D/E)XK nuclease fold (αβββαβ). Consensus sequences and a few representative members of families that met the above criteria were submitted to the Meta Server [37] that assembles various secondary structure prediction and top-of-the-line fold recognition methods. Results produced by these diverse structure prediction methods were screened with 3D-Jury [38,39], a consensus approach that uses structural comparisons to select the best models (the most abundant structures) from a group of assembled models.

Generation of multiple sequence-to-structure alignment

To collect protein sequences that belong to newly identified restriction endonuclease-like families, PSI-BLAST [40] searches (E-value threshold 0.001) were performed against the NCBI non-redundant (nr) protein database with the consensus sequences of analyzed families. Multiple sequence alignments of the families were generated using PCMA program [41] followed by manual adjustments. All PD-(D/E)XK nuclease structures identified in fold recognition searches were used to derive structure-based alignments in the conserved core regions of the fold encompassing four β-strands and two α-helices. Sequence-to-structure mapping between putative PD-(D/E)XK nuclease families and selected structures in the core region was built manually using consensus alignment approach and 3D assessment [42] based on the results of Meta-BASIC, 3D-Jury and secondary structure predictions (mainly with PSI-PRED) as well as conservation of critical active site residues and hydrophobic patterns.

Domain architecture and genomic analysis

To detect other conserved protein domains in identified putative restriction endonuclease-like families, their sequences were analyzed with CDD [25] and SMART [43] This analysis also included searches for transmembrane segments (with TMHMM2 [44]), signal peptides (SignalP [45]), low compositional complexity (CEG [46]) and coiled coil (Coils2 [24]) regions as well as internal repeats (Prospero [47]). Regions with no significant sequence similarity to known protein domains were submitted to Meta-BASIC and then to the Meta Server coupled with 3D-Jury system. All identified domains were checked for the conservation of essential elements, including the presence of domain-specific residues. Genomic analysis (neighborhood and/or gene fusion) was performed with STRING [48] to detect possible functional associations.

Authors' contributions

LK and LNK performed the analyses and drafted the manuscript. LR developed GRDB system. NVG provided thoughtful insights and participated in drafting the manuscript. KG designed and coordinated the study as well as critically edited the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by NIH grant (GM67165) to NVG, MNI and European Commission grants (LSHG-CT-2003-503265 and LSHG-CT-2004-503567) to LR, and by EMBO Installation grant, Foundation for Polish Science (FOCUS) and MNiSW (PBZ-MNiI-2/1/2005) grant to KG.

References

1. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *J Mol Biol* 1995, **247(4):**536-540.
2. Bujnicki JM: **Crystallographic and bioinformatic studies on restriction endonucleases: inference of evolutionary relationships in the "midnight zone" of homology.** *Curr Protein Pept Sci* 2003, **4(5):**327-337.
3. Hadden JM, Convery MA, Declais AC, Lilley DM, Phillips SE: **Crystal structure of the Holliday junction resolving enzyme T7 endonuclease I.** *Nat Struct Biol* 2001, **8(1):**62-67.
4. Nishino T, Komori K, Tsuchiya D, Ishino Y, Morikawa K: **Crystal structure of the archaeal holliday junction resolvase Hjc and implications for DNA recognition.** *Structure* 2001, **9(3):**197-204.
5. Kovall R, Matthews BW: **Toroidal structure of lambda-exonuclease.** *Science* 1997, **277(5333):**1824-1827.
6. Tsutakawa SE, Jingami H, Morikawa K: **Recognition of a TG mismatch: the crystal structure of very short patch repair endonuclease in complex with a DNA duplex.** *Cell* 1999, **99(6):**615-623.
7. Tatusov RL, Koonin EV: **A simple tool to search for sequence motifs that are conserved in BLAST outputs.** *Comput Appl Biosci* 1994, **10(4):**457-459.
8. Zhou XE, Wang Y, Reuter M, Mucke M, Kruger DH, Meehan EJ, Chen L: **Crystal structure of type IIE restriction endonuclease EcoRII reveals an autoinhibition mechanism by a novel effector-binding fold.** *J Mol Biol* 2004, **335(1):**307-319.
9. Kovall RA, Matthews BW: **Type II restriction endonucleases: structural, functional and evolutionary relationships.** *Curr Opin Chem Biol* 1999, **3(5):**578-583.
10. Galburt EA, Stoddard BL: **Catalytic mechanisms of restriction and homing endonucleases.** *Biochemistry* 2002, **41(47):**13851-13860.
11. Aravind L, Makarova KS, Koonin EV: **SURVEY AND SUMMARY: holliday junction resolvases and related nucleases: identification of new families, phyletic distribution and evolutionary trajectories.** *Nucleic Acids Res* 2000, **28(18):**3417-3432.
12. Kinch LN, Ginalski K, Rychlewski L, Grishin NV: **Identification of novel restriction endonuclease-like fold families among hypothetical proteins.** *Nucleic Acids Res* 2005, **33(11):**3598-3605.
13. Kosinski J, Feder M, Bujnicki JM: **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* 2005, **6:**172.
14. Ginalski K, von Grotthuss M, Grishin NV, Rychlewski L: **Detecting distant homology with Meta-BASIC.** *Nucleic Acids Res* 2004, **32(Web Server issue):**W576-81.
15. Bond CS, Kvaratskhelia M, Richard D, White MF, Hunter WN: **Structure of Hjc, a Holliday junction resolvase, from Sulfolobus solfataricus.** *Proc Natl Acad Sci U S A* 2001, **98(10):**5509-5514.
16. Dubos C, Plomion C: **Identification of water-deficit responsive genes in maritime pine (Pinus pinaster Ait.) roots.** *Plant Mol Biol* 2003, **51(2):**249-262.
17. Okstad OA, Hegna I, Lindback T, Rishovd AL, Kolsto AB: **Genome organization is not conserved between Bacillus cereus and Bacillus subtilis.** *Microbiology* 1999, **145 (Pt 3):**621-631.
18. Piekarowicz A, Klyz A, Kwiatek A, Stein DC: **Analysis of type I restriction modification systems in the Neisseriaceae: genetic organization and properties of the gene products.** *Mol Microbiol* 2001, **41(5):**1199-1210.
19. van der Woerd MJ, Pelletier JJ, Xu S, Friedman AM: **Restriction enzyme BsoBI-DNA complex: a tunnel for recognition of degenerate DNA sequences and potential histidine catalysis.** *Structure* 2001, **9(2):**133-144.
20. Neuwald AF, Aravind L, Spouge JL, Koonin EV: **AAA+: A class of chaperone-like ATPases associated with the assembly, operation, and disassembly of protein complexes.** *Genome Res* 1999, **9(1):**27-43.
21. Pieper U, Brinkmann T, Kruger T, Noyer-Weidner M, Pingoud A: **Characterization of the interaction between the restriction endonuclease McrBC from E. coli and its cofactor GTP.** *J Mol Biol* 1997, **272(2):**190-199.
22. Bourniquel AA, Bickle TA: **Complex restriction enzymes: NTP-driven molecular motors.** *Biochimie* 2002, **84(11):**1047-1059.
23. Middleton CL, Parker JL, Richard DJ, White MF, Bond CS: **Substrate recognition and catalysis by the Holliday junction resolving enzyme Hje.** *Nucleic Acids Res* 2004, **32(18):**5442-5451.
24. Lupas A: **Predicting coiled-coil regions in proteins.** *Curr Opin Struct Biol* 1997, **7(3):**388-393.
25. Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Liebert CA, Liu C, Lu F, Marchler GH, Mullokandov M, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Yamashita RA, Yin JJ, Zhang D, Bryant SH: **CDD: a Conserved Domain Database for protein classification.** *Nucleic Acids Res* 2005, **33(Database issue):**D192-6.
26. Liu J, Smith CL, DeRyckere D, DeAngelis K, Martin GS, Berger JM: **Structure and function of Cdc6/Cdc18: implications for origin recognition and checkpoint control.** *Mol Cell* 2000, **6(3):**637-648.
27. Bowman GD, O'Donnell M, Kuriyan J: **Structural analysis of a eukaryotic sliding DNA clamp-clamp loader complex.** *Nature* 2004, **429(6993):**724-730.
28. Cann IK, Ishino S, Yuasa M, Daiyasu H, Toh H, Ishino Y: **Biochemical analysis of replication factor C from the hyperthermophilic archaeon Pyrococcus furiosus.** *J Bacteriol* 2001, **183(8):**2614-2623.
29. Jeruzalmi D, O'Donnell M, Kuriyan J: **Crystal structure of the processivity clamp loader gamma (gamma) complex of E. coli DNA polymerase III.** *Cell* 2001, **106(4):**429-441.
30. Kinebuchi T, Kagawa W, Enomoto R, Tanaka K, Miyagawa K, Shibata T, Kurumizaka H, Yokoyama S: **Structural basis for octameric ring formation and DNA interaction of the human homologous-pairing protein Dmc1.** *Mol Cell* 2004, **14(3):**363-374.
31. Sickmier EA, Kreuzer KN, White SW: **The crystal structure of the UvsW helicase from bacteriophage T4.** *Structure* 2004, **12(4):**583-592.

32. Knizewski L, Kinch L, Grishin NV, Rychlewski L, Ginalski K: **Human herpesvirus 1 UL24 gene encodes a potential PD-(D/E)XK endonuclease.** *J Virol* 2006, **80(5):**2575-2577.
33. Singleton MR, Dillingham MS, Gaudier M, Kowalczykowski SC, Wigley DB: **Crystal structure of RecBCD enzyme reveals a machine for processing DNA breaks.** *Nature* 2004, **432(7014):**187-193.
34. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR: **The Pfam protein families database.** *Nucleic Acids Res* 2004, **32(Database issue):**D138-41.
35. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28(1):**235-242.
36. Jones DT: **Protein secondary structure prediction based on position-specific scoring matrices.** *J Mol Biol* 1999, **292(2):**195-202.
37. Bujnicki JM, Elofsson A, Fischer D, Rychlewski L: **Structure prediction meta server.** *Bioinformatics* 2001, **17(8):**750-751.
38. Ginalski K, Elofsson A, Fischer D, Rychlewski L: **3D-Jury: a simple approach to improve protein structure predictions.** *Bioinformatics* 2003, **19(8):**1015-1018.
39. Ginalski K, Rychlewski L: **Detection of reliable and unexpected protein fold predictions using 3D-Jury.** *Nucleic Acids Res* 2003, **31(13):**3291-3292.
40. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25(17):**3389-3402.
41. Pei J, Sadreyev R, Grishin NV: **PCMA: fast and accurate multiple sequence alignment based on profile consistency.** *Bioinformatics* 2003, **19(3):**427-428.
42. Ginalski K, Rychlewski L: **Protein structure prediction of CASP5 comparative modeling and fold recognition targets using consensus alignment approach and 3D assessment.** *Proteins* 2003, **53 Suppl 6:**410-417.
43. Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P: **SMART 5: domains in the context of genomes and networks.** *Nucleic Acids Res* 2006, **34(Database issue):**D257-60.
44. Sonnhammer EL, von Heijne G, Krogh A: **A hidden Markov model for predicting transmembrane helices in protein sequences.** *Proc Int Conf Intell Syst Mol Biol* 1998, **6:**175-182.
45. Nielsen H, Engelbrecht J, Brunak S, von Heijne G: **Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites.** *Protein Eng* 1997, **10(1):**1-6.
46. Wootton JC: **Non-globular domains in protein sequences: automated segmentation using complexity measures.** *Comput Chem* 1994, **18(3):**269-285.
47. Mott R: **Accurate formula for P-values of gapped local sequence and profile alignments.** *J Mol Biol* 2000, **300(3):**649-659.
48. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P: **STRING: known and predicted protein-protein associations, integrated and transferred across organisms.** *Nucleic Acids Res* 2005, **33(Database issue):**D433-7.