Published online 20 October 2010 *Nucleic Acids Research, 2011, Vol. 39, No. 4* ***1187–1196***
doi:10.1093/nar/gkq958

Identification of new homologs of PD-(D/E)XK nucleases by support vector machines trained on data derived from profile–profile alignments

Mindaugas Laganeckas, Mindaugas Margelevičius and Česlovas Venclovas*

Institute of Biotechnology, Graičiūno 8, LT-02241 Vilnius, Lithuania

Received August 19, 2010; Revised September 25, 2010; Accepted September 29, 2010

ABSTRACT

PD-(D/E)XK nucleases, initially represented by only Type II restriction enzymes, now comprise a large and extremely diverse superfamily of proteins. They participate in many different nucleic acids transactions including DNA degradation, recombination, repair and RNA processing. Different PD-(D/E)XK families, although sharing a structurally conserved core, typically display little or no detectable sequence similarity except for the active site motifs. This makes the identification of new superfamily members using standard homology search techniques challenging. To tackle this problem, we developed a method for the detection of PD-(D/E)XK families based on the binary classification of profile–profile alignments using support vector machines (SVMs). Using a number of both superfamily-specific and general features, SVMs were trained to identify true positive alignments of PD-(D/E)XK representatives. With this method we identified several PFAM families of uncharacterized proteins as putative new members of the PD-(D/E)XK superfamily. In addition, we assigned several unclassified restriction enzymes to the PD-(D/E)XK type. Results show that the new method is able to make confident assignments even for alignments that have statistically insignificant scores. We also implemented the method as a freely accessible web server at http://www.ibt.lt/bioinformatics/software/pdexk/.

INTRODUCTION

PD-(D/E)XK nucleases comprise a large and extremely diverse group of proteins that are involved in various processes of nucleic acid metabolism. Typically, different PD-(D/E)XK families share little or no recognizable sequence similarity except for the conserved signature of the active site. As it happens, the name of this large group of nucleases derives from the highly conserved 'PD' (in many cases only 'D') and '(D/E)XK' ('X' denotes the non-conserved position) active site motifs.

Initially, the only known representatives of PD-(D/E)XK nucleases were Type II restriction endonucleases (REases) (1). These enzymes recognize short DNA sequences (usually 4–8-bp long) and cleave both DNA strands either at or in the vicinity of the recognition target. In conjunction with a methylase that recognizes the same DNA sequence, these two proteins make up a restriction–modification (R–M) system that protects the bacterium from foreign DNA. For quite some time it was assumed that, although divergent, all REases are related. However, lately, four other unrelated groups of REases have been discovered making the commonly used term 'restriction endonuclease-like fold' [e.g. in the SCOP database (2)] obsolete. A recent comprehensive survey (3) estimated that out of the five known REase groups, PD-(D/E)XK nucleases comprise the most abundant one. However, still quite a large fraction of REases could not be reliably assigned to any of the known groups (3).

The first instance of the PD-(D/E)XK domain discovered outside of REases was bacteriophage λ exonuclease (4), the protein that functions in recombination and repair of the viral chromosome. Subsequently, PD-(D/E)XK fold domains have been identified in many other protein families in all domains of life. Examples of biological functions represented by these families include DNA damaged repair [MutH (5) and Vsr (6)], Holliday junction resolution [T7 endonuclease I (7), Hjc (8), Hje (9) and XPF/Rad1/Mus81-dependent nuclease (10)] and more recently discovered the RNA processing function [PA subunit of avian influenza RNA polymerase (11,12) and Rai1/Dom3Z (13)].

Although PD-(D/E)XK domains often display little sequence similarity they all share a structurally conserved core. The consensus structural core can be defined as a four-stranded mixed β-sheet flanked by an α-helix on each side (producing αββββαβ topology) (14–15).

*To whom correspondence should be addressed. Tel: +370 5 2691881; Fax: +370 5 2602116; Email: venclovas@ibt.lt

However, the consensus core is often elaborated with very different peripheral elements usually responsible for oligomerization or substrate recognition. Sometimes these non-conserved peripheral elements constitute the majority of the protein effectively concealing the similarity between PD-(D/E)XK domains. This problem has been noted early on (16) starting with the structure determination of the very first representatives of PD-(D/E)XK nucleases, EcoRI and EcoRV. Even as the structural data of the PD-(D/E)XK superfamily has grown considerably new members are still being overlooked (17).

Over the years the number of experimentally determined structures of PD-(D/E)XK domains has been steadily increasing. Currently, the SCOP database (version 1.75) (2) groups 33 'restriction endonuclease-like' families and the 'PD-(D/E)XK clan' in PFAM (version 24.0) (18) includes 44 families. Based on the growth trends, it is reasonable to expect that a significant number of PD-(D/E)XK protein families are yet to be identified. In the absence of experimental evidence, computational methods may provide an effective means in detecting new PD-(D/E)XK families. Such examples include transitive searches with the meta profile comparison method Meta-BASIC (14,19) and profile–profile comparison with HHsearch (20). However, although profile–profile comparison methods at present are state-of-the-art in distant homology detection, non-trivial evolutionary relationships [as is often the case with PD-(D/E)XK domains] may be missed because of the assigned statistically insignificant scores. These relationships 'hidden' among unrelated matches may often be recognized by experts armed with the knowledge that is specific to the protein families of interest, e.g. the nature and the location of the active site residues. However, it is apparent that the expert knowledge is most useful if it can be incorporated in the automated manner. One way to do this is through the use of supervised machine learning methods such as support vector machines (SVMs) (21).

Here, we used a combination of a profile–profile comparison method and SVMs to search for yet undetected PD-(D/E)XK families. For profile comparison we chose HHsearch (version 1.5.0) (22), one of the best performing and fastest methods currently available (23). In addition, HHsearch tends to generate accurate but fairly short alignments (24) avoiding extension into less conserved regions. We reasoned that this would be especially relevant for detecting new putative PD-(D/E)XK families. Their alignments with known PD-(D/E)XK representatives could be expected to preferentially span evolutionary and structurally most conserved region housing the active site motifs (Figure 1). On the other hand, SVMs have been shown to be very efficient in binary classification of data, and our task of distinguishing members from non-members of the PD-(D/E)XK superfamily belongs to this data classification category. We were also encouraged by a recent study, which used HHsearch and SVMs to make a general purpose homology detection tool, HHsvm (25).

As a result of this study we identified PD-(D/E)XK domains in several PFAM families of uncharacterized proteins and assigned a number of previously unclassified REases to the PD-(D/E)XK type. We also implemented the method as a freely accessible web server.

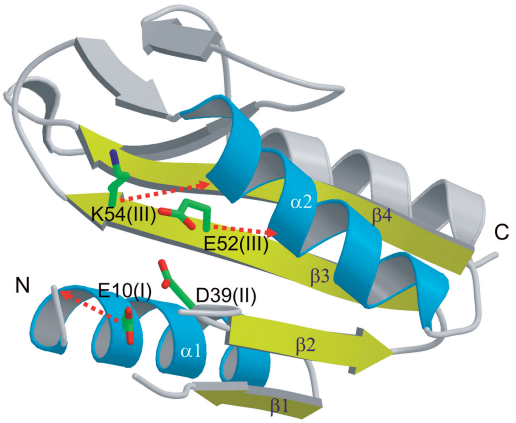

Figure 1. Conserved structural core and typical active site arrangement in PD-(D/E)XK nucleases. Shown is the 3D structure of the archaeal Holliday junction endonuclease (PDB: 1ob8). Secondary structure elements of the conserved core are labeled and colored blue (α-helices) and yellow (β-strands). Side chains are shown for the residues representing the three active site signature motifs (I–III). Red broken arrows indicate observed variants of the active site residue 'migration' into alternative positions.

MATERIALS AND METHODS

SVM training to recognize PD-(D/E)XK families

The SVMs (26) were trained to recognize PD-(D/E)XK families on the basis of HHsearch pairwise alignments that provided both positive and negative data points. To generate alignments, a compiled set of known PD-(D/E)XK domains from SCOP, PDB and PFAM (positive data set, see description below) were searched against the complete SCOP database appended with the positive dataset members from PDB and PFAM (SCOP-1.75-pdb-pfam-pdexk). Before searching, each family was augmented with the secondary structure predicted by PSIPRED (27). Resulting alignments all had a known PD-(D/E)XK family from the positive data set aligned to a family from the SCOP-1.75-pdb-pfam-pdexk database. If both families in an alignment were from the positive PD-(D/E)XK data set, the alignment was considered to represent a positive example. A negative example was represented by the alignment, in which a PD-(D/E)XK family from the positive data set was aligned to a SCOP family that is not part of the 'restriction endonuclease-like' fold (c.52). In other words, the negative examples were extracted from HHsearch results without the necessity of having a predefined negative data set. To make sure that the SVM training would be performed on the structurally and evolutionary conserved core (Figure 1) and not on variable/unrelated regions of PD-(D/E)XK domains, alignments representing both positive and negative examples were considered only if at least one of the two PD-(D/E)XK query active site motifs (II or III) were included into the aligned region.

Positive data set

The positive data set was compiled from PD-(D/E)XK domains classified in SCOP as members of the 'restriction endonuclease-like' superfamily (c.52.1). The set also included PD-(D/E)XK domains of PDB structures not yet classified in SCOP. These were identified either from the assignments made by the authors of published structures or by searching PDB with representative PD-(D/E)XK structures using DaliLite (28) followed by manual validation. Only those SCOP and PDB sequences that were <40% identical to each other and had at least five family members culled at 90% identity were retained. In addition, the positive dataset included consensus sequences of PFAM families that lack experimentally determined 3D structures, yet are assigned to the PD-(D/E)XK clan (CL0236). Active site signature motifs (I–III) (Figure 1) of PD-(D/E)XK domains of the positive data set were extracted manually based on structure analysis and published data. Although in a number of cases active site residues have 'migrated' to alternative positions (Figure 1), we only considered canonical positions. The complete positive data set consisted of 34 SCOP, 24 PDB and 15 PFAM PD-(D/E)XK domains. Corresponding FASTA-formatted sequences are available as Supplementary Data file 1 and the list of active site motifs is provided in Supplementary Figure S1.

Attributes for SVM training

SVMs with the perceptron kernel were trained on a number of attributes derived from the profile–profile alignments (Supplementary Figure S2) and profile regions included in the alignment. Before training, values of all attributes were scaled to fit within the [–1;+1] range. The attributes may be divided into two classes: (i) PD-(D/E)XK-specific attributes derived from the active site positions, and (ii) PD-(D/E)XK-non-specific attributes derived from the alignments between the query and the match.

PD-(D/E)XK-specific attributes include:

(1) The active site position score (how well the PD-(D/E)XK active site positions match those aligned). Initially, we defined the following six active site positions: 'E/Q' (motif-I)—1, 'PD' (motif-II)—2 and 'EXK' (motif-III)—3 positions. However, after experimenting with the motif definitions for the purpose of SVM training we found that 'DX' and 'XE-K' positions for correspondingly motifs II and III made SVMs more specific and therefore were used throughout the study. Motifs II and III were included in all classifiers while the least conserved motif-I was alternatively included/excluded. Each position was scored separately. The position scoring system was taken directly from the HHsearch alignment, but the actual column scores were assigned as follows: very bad match ('='), −10; bad match ('–'), −6; neutral match (' '), 0; good match ('+'), 6; very good match ('|'), 10.
(2) The transition probability of the match to match state (M→M) for each of the active site positions from both the query and the matching profiles.
(3) The entropy E for each of the active site positions computed from the residue frequency distribution at corresponding positions: $E = -\sum_{i=1}^{20} (f_i * \log_2 f_i)$, where f_i is a weighted observed frequency of the i-th amino acid. The entropy values were calculated and used as attributes for both the query and match positions.

PD-(D/E)XK-non-specific attributes:

(1) HHsearch secondary structure score.
(2) The ratio of the query and the match average values for each of the transition probabilites. The average values were obtained from the profile fragments included in the alignment.
(3) The average values of entropy computed from the residue frequency distribution from both the query and the match profile positions included in the alignment.
(4) The ratio of the alignment length with the entire length of the matching sequence.
(5) The ratio of the GRAVY indexes (Grand Average of Hydropathy) calculated for the aligned regions of the query and match profiles using Kyte & Doolittle residue hydropathy values (29). The hydropathy value at each profile position was averaged over the weighted observed residue frequencies at that position.
(6) The ratio of the isoelectric points (pIs) calculated for the aligned fragments of the query and match sequences using residue pK values as in ProMoST (30).

PD-(D/E)XK-specific attributes were chosen to focus on the recognition of the active site. Some of the general attributes were selected on the basis of their use by human experts in the assessment of tentative homologous relationships. For example, a good match between secondary structures of two protein families is suggestive of the homology even if the corresponding sequence similarity is low. Also, a longer region of a PD-(D/E)XK domain aligned to unknown family is often more indicative of the true relationship than the shorter one. Calculations of the GRAVY index and pI for the aligned regions were included to better reflect structural and functional (e.g. nucleic acids binding) similarities. The meaning of the remaining attributes may not be entirely intuitive, but together they produce a sufficiently rich object description for machine learning. Every new feature was first tested and included in the set of attributes only if its addition increased the training accuracy. The training accuracy was obtained using 5-fold cross validation. The training data set was split into five parts. SVMs were trained on each of the four parts, and the newly created classifier was tested on the fifth part. The training accuracy was calculated as ('number-of-samples' – 'number-of-errors')*100%/'number-of-samples'.

Following the previously reported observations (25) the HHsearch alignment probability was not included as a feature into SVM classifiers. However, it was used in the

SVM training indirectly. Three different SVM classifiers were trained using alignments representing positive examples that had HHsearch probabilities above 50, 70 and 80% respectively (no filtering was applied to negative examples). The rationale for using these cutoffs was to perform training on positive alignments of higher quality and also to make their numbers comparable to the smaller set of alignments representing negative examples.

Analysis of candidate PD-(D/E)XK families

PFAM families classified by the SVM procedure as new candidate members of the PD-(D/E)XK superfamily were analyzed further in order to validate the results. First, we classified the same HHsearch alignments with HHsvm (25), a general purpose homology detection method, and compared with our results. Second, we analyzed the obtained HHsearch probability (statistical significance estimation) values. Since our classification method does not explicitly include HHsearch probabilities, this was considered to provide important non-redundant evidence. In adition, we used another recently developed sensitive profile–profile comparison method, COMA (version 1.10) (24) to search SCOP, PDB and PFAM databases with selected representatives of the newly detected PD-(D/E)XK families. Results were inspected for statistically significant matches with known PD-(D/E)XK representatives. Query alignments to SCOP and PDB PD-(D/E)XK structures obtained with either of the two profile–profile comparison methods were used to generate crude models for inspection of the active site residue positions and compatibility of the alignments with the structurally conserved core. Analysis of multiple models superimposed with representative PD-(D/E)XK structures were used to generate consensus alignments for the conserved core regions similarly as described previously (31). We also analyzed whether additional domains, if present, support the predicted relationship.

RESULTS

Based on the analysis of training results, we selected five best performing SVM classifier variants. They differ in the HHsearch probability threshold used to select positive alignments for SVM training and whether or not motif-I ('E/Q') is included together with the modified motifs II ('DX') and III ('X(D/E)-K') to calculate active site positional scores and entropy values (Table 1). However, we noticed that even these best classifiers with similar training accuracy in some cases were producing different results. Therefore, we introduced a consensus classifier, which gives the overall estimated probability, calculated as a simple average of the five above mentioned classifier probabilities. The consensus classifier probability of 0.8 or higher is considered to indicate a reliable assignment to the PD-(D/E)XK superfamily. If the probability is <0.5 the query is unlikely to be a member of the superfamily, and the probability in the 0.5–0.8 interval indicates an uncertain assignment.

We applied the five classifiers in two settings. In the first setting, we used to query all families in the PFAM database aiming primarily at the identification of PD-(D/E)XK nucleases among uncharacterized protein families. In the second setting, we asked whether any of unclassified REases can be confidently assigned to the PD-(D/E)XK superfamily.

Table 1. Description of the SVM classifiers

SVM classifier	HHsearch probability threshold (%)	Number of positive/negative training examples	Use of motif-I (E/Q) in the definition of the active site	Training accuracy (%)
SVM-1	50	381/257	No	95.9
SVM-2	50	381/257	Yes	95.9
SVM-3	70	285/257	No	97.4
SVM-4	70	285/257	Yes	98.0
SVM-5	80	233/257	Yes	98.6

Identification of new PD-(D/E)XK families in PFAM

Among PFAM families that were classified as PD-(D/E)XK with very high consensus probabilities (0.9 and above) we detected three new families of unknown function and refined the active site motif assignment for the fourth (Figure 2). A brief summary of the analysis for each family is presented below.

DUF511 (PF04373). The DUF511 domain is found in bacteria (predominantly in Helicobacter and Campylobacter species). A typical length of proteins in this family is 310–320 residues. DUF511 is typified by *Helicobacter pylori* protein of unknown function encoded by *hrgA*, a *H. pylori* restriction endonuclease-replacing gene A (32–33). The *hrgA* gene was discovered by analyzing the *H. pylori* Type II *Hpy*III R–M system, which is homologous to *MboI*, specific for the DNA sequence GATC. It was found that in some strains *Hpy*IIIR, the REase component of the *Hpy*III R–M system, is replaced by *hrgA*. Strikingly, 208 strains examined had either *hrgA* or *hpy*IIIR, but not both. This suggested that *hrgA* and *hpy*IIIR can be exchanged by homologous recombination and this hypothesis was confirmed experimentally (32). Neither *hrgA*, nor *hpy*IIIR are essential. It is interesting that initially HrgA was hypothesized to be associated with gastric cancer (33). However, a subsequent comprehensive study examined ~500 *H. pylori* isolates and concluded that there is no specific association of HrgA with the disease (34).

Our analysis showed that DUF511 sequences have all three motifs associated with the canonical PD-(D/E)XK active site signature (Figure 2). In addition to the PD-(D/E)XK domain, DUF511 family members are predicted to have the winged-helix motif (a variant of the DNA-binding helix-turn-helix motif) at the very N-terminus. Both HHsearch and COMA detect the N-terminal domain of *Bacillus subtilis* δ subunit of RNA polymerase [PDB id: 2krc; (35)] as the closest match for this region in PDB. In addition, a number of other winged-helix motifs such as fork head DNA-binding

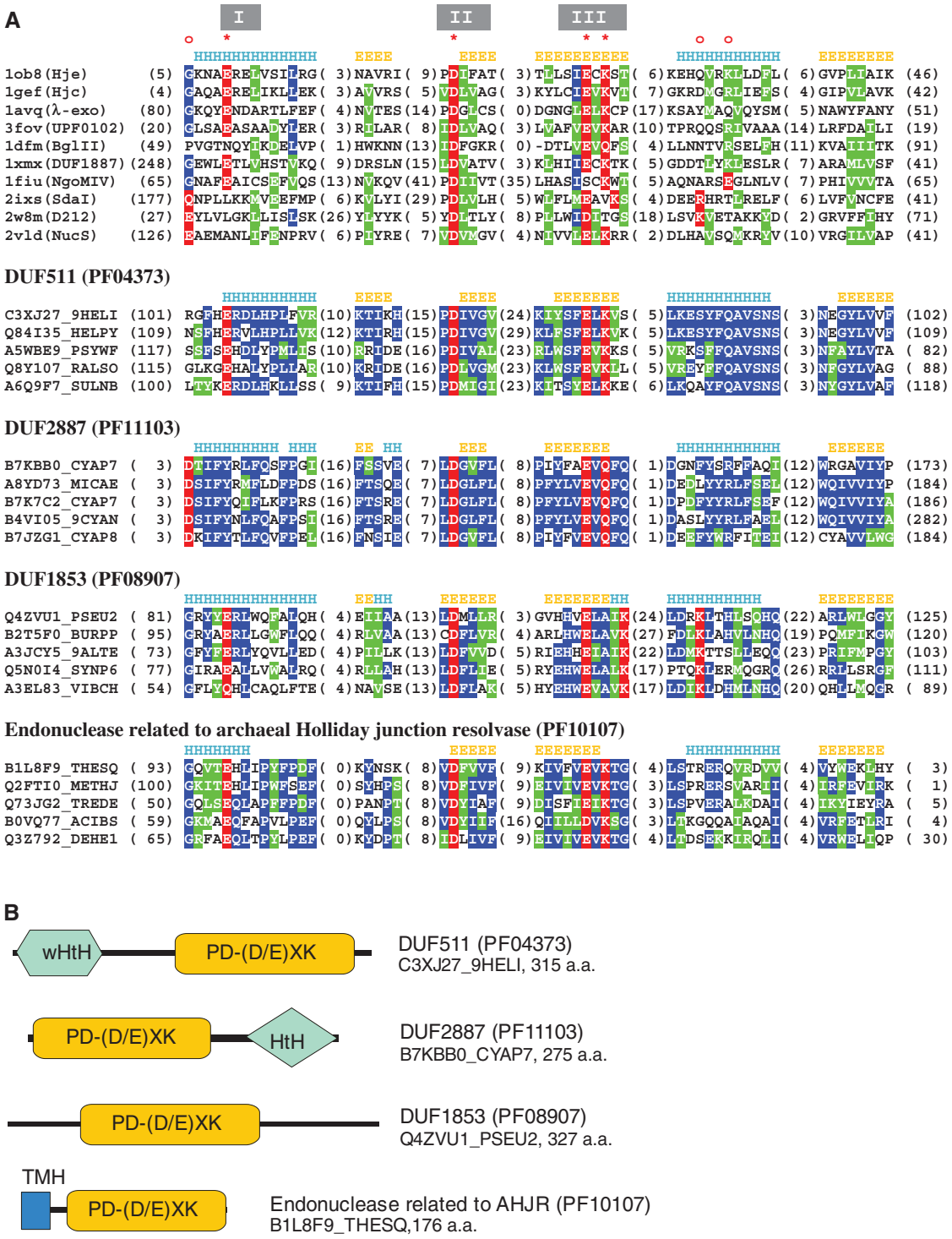

Figure 2. Putative PD-(D/E)XK families. (**A**) Sequence alignments with PD-(D/E)XK representatives having known experimental 3D structures. Each family is denoted by the PFAM name and the accession number. Sequences within families are labeled with Uniprot (www.uniprot.org) accession codes. PD-(D/E)XK structural representatives are labeled with corresponding PDB codes and common protein/family names in parentheses. PD-(D/E)XK signature motifs (I–III) are indicated at the top. Red asterisks and open circles denote respectively canonical and alternative positions of the active site residues. Six aligned blocks correspond to secondary structure elements of the conserved structural core characteristic of the superfamily as shown in Figure 1. Numbers in parentheses indicate excluded residues. Alignments are colored according to sequence conservation: identical residues have blue background, similar ones—green. Known or putative active site residues are highlighted in red. Observed (PDB: 1ob8) or predicted consensus secondary structure for each family is displayed above the sequences (H, α-helix; E, β-strand). (**B**) Domain composition of representative sequences from each family. Additional domains/motifs are denoted as follows: wHtH, winged-helix; HtH, helix-turn-helix; TMH, transmembrane helix.

domain and ferric uptake regulator-like (FUR) transcription factors are also detected. The presence of the DNA-binding domain provides an additional support for the assignment of DUF511 to the PD-(D/E)XK nuclease superfamily. Surprisingly, the closest PD-(D/E)XK domain is detected by HHsearch with only poor statistical significance value (23% probability). On the other hand, COMA detects a PD-(D/E)XK structure (PDB: 3dnx) as the best match in PDB with *E*-value of 0.002. Reciprocal search with COMA using the 3dnx sequence to query PFAM in turn detects DUF511 (discounting the query itself) as the best match albeit with worse *E*-value (0.01). Thus, COMA results provide an independent support for the SVM-based assignment. Noteworthy, none of the four HHsvm classifiers predicts the presence of the PD-(D/E)XK domain in DUF511.

The exact cellular function of DUF511 family proteins is not obvious. Based on findings that HrgA in *H. pylori* replaces HpyIIIR it might be hypothesized that HrgA serves as a replacement restriction endonuclease. However, the fact that the HpyIIIM methylase is inactive in a number of *H. pylori* strains carrying *hrgA* (32) argues against it. In addition, there does not seem to be a positional conservation of a methylase in the neighborhood of DUF511 members in other bacterial genomes. This suggests that the regulation of the DUF511 function is not dependent on the methylase activity.

DUF2887 (PF11103)

The DUF2887 family is represented by predominantly bacterial and few archaeal proteins that typically are ~300 residues long. The predicted PD-(D/E)XK domain is located in the N-terminal region, while the C-terminus contains the DNA-binding helix-turn-helix (HtH) domain (Figure 2). Within the PD-(D/E)XK domain, motif-I is in an alternative position, and motif-III is represented by the EXQ motif instead of the canonical EXK motif. Among PD-(D/E)XK proteins of known structure, the identical active site signature motif-III is shared by Type II restriction enzymes BglII (36) and BstY (37). However, the most extensive similarity is detected with the PFAM family of YhgA-like putative transposases, also featuring EXQ as motif-III (19). Both HHsearch and COMA identify the YhgA-like family as the best match (62% probability and *E*-value = 8e–13, respectively). Two of the HHsvm variants also classified this family as PD-(D/E)XK. The similarity with YhgA-like proteins extends beyond the PD-(D/E)XK domain and into the all α-helical C-terminal region, harboring the HtH domain. The HtH domain is clearly homologous to the DNA-binding domain of the resolvase/invertase family, exemplified by γδ-resolvase. PD-(D/E)XK and HtH domains in DUF2887 proteins appear to be connected via a long α-helix, much like the arrangement of catalytic and HtH domains in γδ-resolvase [(38); PDB id: 1gdt]. It is intriguing that in contrast to the homologous C-terminal region, the catalytic domain of γδ-resolvase is not at all related to PD-(D/E)XK nucleases. Both the structure and the catalytic mechanism are different. The cleavage of DNA by γδ-resolvase proceeds through a covalent intermediate formed by the side chain of the Ser residue (39), while PD-(D/E)XK nucleases hydrolyze phosphodiester bond by direct inline nucleophile attack resulting in the inversion of configuration at the phosphorous atom (40). Thus, DUF2887/YhgA-like proteins and γδ-resolvase respectively represent two different domain fusion events linked by the deployment of a common C-terminal domain. The detected homology between the C-terminal regions of DUF2887/YhgA-like proteins and γδ-resolvase suggests similar protein–DNA interactions, in which both the long helix (E-helix) and the HtH domain are involved (38). Furthermore, it might be expected that DUF2887/YhgA-like putative transposases and γδ-resolvase share a similar higher order arrangement as the E-helix connecting catalytic and HtH domains contributes significantly to the interface of both γδ-resolvase dimer and its synaptic tetramer complex (41).

DUF1853 (PF08907). The DUF1853 domain (~320 a.a.) is predominantly found in bacteria with few instances identified in eukaryotes (plants, algae, phytophthora). In eukaryotes DUF1853 is present in the context of significantly longer sequences compared to those in bacteria. Both HHsearch and COMA detect UPF0102, a widely distributed domain of unknown function (PDB: 3fov), as the closest PD-(D/E)XK match with 96% probability and *E*-value = 0.005 respectively). Furthermore, all four versions of HHsvm also consistently (probabilities >0.9) classify DUF1853 as the PD-(D/E)XK domain.

Both motif-I (E/Q) and motif-II (D) are arranged as in canonical PD-(D/E)XK domains. In contrast, the conserved Lys residue in motif-III (EXK) appears to have migrated by two positions along the protein chain to produce the EXXXK-motif instead. To our knowledge, this particular variant of the EXK-motif has not yet been identified in any of the experimentally characterized PD-(D/E)XK proteins. Yet, it is reminiscent of the active site residue arrangement in type II REase SdaI (42) (Figure 2), in which Lys251 is part of the EXXK motif, and an additional arginine residue (Arg260) resides in the second conserved α-helix. DUF1853 sequences similarly have an additional Lys present in the corresponding position of the α-helix (Figure 2A). In SdaI both residues have been shown to contribute to the catalytic activity (42), with Arg260 in the α-helix being more important. A recent structure of the PD-(D/E)XK nuclease D212 from an archaeal virus revealed another example of the putative active site lysin (Lys123) occupying equivalent position in the second conserved α-helix (43) (Figure 2A). Thus, either one or both conserved Lys residues might also be expected to contribute to the formation of the active site in DUF1853.

Endonuclease related to archaeal Holliday junction resolvase (PF10107). This protein family consists of 150–200 residue-long proteins, found in bacteria and archaea. To our knowledge, no protein in this family has been experimentally characterized yet, but a previous computational study assigned a subset of this family (COG4741) to the PD-(D/E)XK superfamily (20).

Our results corroborate the presence of a compact PD-(D/E)XK domain in proteins of this family and refine the assignment of the structural core by identifying the conserved motif-I of the putative active site (Figure 2A). In a previous study (20), the predicted N-terminal transmembrane α-helix raised a question of whether these proteins are intracellular or secreted. To further address this question, we explored a number of family members present in CoBaltDB, a database that compiles predictions by various methods concerning protein localization of complete prokaryotic proteomes (44). The CoBaltDB data in agreement with the previous finding (20) predict that these proteins are not secreted, but located in cytoplasm anchored to the membrane through the N-terminus. However, it remains to be investigated whether they are involved in protecting cells from translocating foreign DNA or in some kind of DNA repair/recombination activity.

DUF1173 (PF06666): a family lacking the conserved PD-(D/E)XK active site signature. In addition to the highly reliable assignments described above, we identified a putative PD-(D/E)XK relative among families classified with lower probabilities. The DUF1173 family grouping proteins of unknown function was classified as PD-(D/E)XK with 0.6 probability. DUF1173 includes mainly bacterial proteins of ~400 residues in length. The putative PD-(D/E)XK-related domain is located within the C-terminal region, while the very N-terminus features a putative zinc-binding motif, $CysXCysX_nHisX_{3\text{-}5}Cys$. The inspection revealed that unlike the above four families, DUF1173 does not conserve the PD-(D/E)XK active site signature, and that could be the reason for a lower probability assignment by SVM classifiers. For example, similar low probability (0.55) is assigned to the 'NMDA receptor-regulated gene protein 2′ family (PF10505), recently identified as a PD-(D/E)XK-related family lacking active site signature motifs (17). At the same time, the relationship between DUF1173 and PD-(D/E)XK domains is supported by additional evidence. Using DUF1173 as a query HHsearch detects DUF790 [a member of the PFAM PD-(D/E)XK clan] with 63% probability, and COMA finds another PD-(D/E)XK clan family (DUF1064) with *E*-value = 2e−04. In a reciprocal search using DUF1064 as a query, both HHsearch and COMA detect DUF1173 respectively with 88% probability and *E*-value = 0.006, with many other known PD-(D/E)XK families scoring worse. This makes a strong case for the evolutionary link between the C-terminus of DUF1173 and PD-(D/E)XK nucleases although manifested only at a structural rather than at a functional level. There are no obvious clues as to the function of this family, especially that the closest PD-(D/E)XK relatives are themselves proteins of unknown function.

Restriction endonucleases assigned to the PD-(D/E)XK superfamily

Many REases remain unassigned to any of the five structurally and evolutionary unrelated superfamilies (3), raising the question of whether they form new superfamilies or belong to the known ones. We analyzed PFAM REase families and the set of REBASE REase families previously classified as 'unknown' (3) in an attempt to identify REases of the PD-(D/E)XK type. Since the method is based on comparison of multiple sequence alignments (represented as profiles) we only considered REase families that had at least four members sharing no >90% sequence identity. Therefore, a number of REAses, represented by 'orphan' sequences (without detectable homologs) or having too few homologs, were not analyzed.

We identified a number of PFAM REase families, for which no experimental 3D structures are available, as the PD-(D/E)XK type. To our knowledge, two of them (Table 2) are novel assignments. Two other detected PD-(D/E)XK REases, Eco47II (PF09553) and ScaI (PF09569), corroborate a previous theoretical assignment of their respective homologs (33 and 39% sequence identity), Sau96I and LladI (3). We also assigned five 'unknown' REBASE enzyme families (Table 2). The confidence for all these new assignments is indicated by high SVM probabilities [0.9–1.0]. Noteworthy, the high confidence is achieved even for alignments with very low

Table 2. REases, newly assigned to PD-(D/E)XK superfamily

REase family	SVM probability	HHsearch probability, (%)	Putative active site motifs	Recognition site	Subtype
PFAM families					
XamI (PF09572)	1.0	85	171-E(42)AD(13)ECK	GTCGAC	P
MjaII (PF09520)	0.99	67	149-AD(12)ELK	GGNCC	P
REBASE enzymes					
BlopNAC1P	1.0	100	111-E(25)PD(23)ASK(11)E	CCWGG	P
NcoI	0.99	82	63-LD(19)EAA(1)R	C^CATGG	P
BseMII	1.0	99	48-PD(16)ELK	CTCAG (10/8)	G,S
BseRI	1.0	98	74-VD(8)EYE	GAGGAG (10/8)	G,S
AluI	0.88	7	248-YD(15)DLK	AG^CT	P

For each REase family the SVM probability of assignment to the PD-(D/E)XK superfamily, HHsearch probability, putative active site motifs, DNA recognition sequence and a subtype according to REBASE are indicated. Putative active site motifs are annotated with starting residue numbers and number of residues in between the motifs. Predicted active site residues are in red color, motif-III residues that 'migrated' to non-canonical positions are underlined. Where known, cleavage sites within recognition sequences are indicated with '^' and those outside of recognition sequences—with two numbers in parentheses—for top and bottom strands respectively. REBASE subtypes are as follows: P—symmetric target and cleavage sites; G—symmetric or asymmetric target, affected by AdoMet; S—asymmetric target and cleavage sites.

HHsearch probability (e.g. 7% for AluI). HHsvm turned out to be much less effective in recognizing REases. Although REBASE enzymes (Table 2) except for AluI were recognized as PD-(D/E)XK nucleases, Eco47II and ScaI were not. XamI and MjaII each was assigned high probabilities (>0.9) only by one of the four HHsvm classifiers.

Newly classified REases (Table 2) feature both canonical and alternative active site signature motifs. NcoI and BlopNAC1P both have a non-canonical motif-III, from which an active site residue has 'migrated' into a different position. BlopNAC1P is closely related to the catalytic domain of EcoRII, a well characterized PD-(D/E)XK REase (45–46), and the similarity can be recognized even by BLAST. BseRI and BseMII both resemble the bifunctional REase-methylase Eco57I, in which endonuclease, methylase and recognition domains are arranged sequentially within a single polypeptide chain (47–49). However, only the BseMII putative active site as suggested earlier (50) and substantiated here closely resembles that of Eco57I, while BseRI instead has the BamHI-like motif-III (ExE).

PD-(D/E)XK recognition web server

To make the method accessible for broader biological community we implemented it as a user-friendly web server accessible at http://www.ibt.lt/bioinformatics/software/pdexk/. The input for the server is either a single sequence or a multiple sequence alignment. If the input is a single sequence, a PSI-BLAST search is initiated against locally maintained sequence database to collect homologous sequences. If the input is a multiple sequence alignment, the user can choose whether to use it directly for the query HMM construction by HHsearch or to use it to jump-start a PSI-BLAST search. The user may bookmark the web page, in which results will be displayed, or choose to receive an HTML link by e-mail instead of waiting for a job to finish. Results page gives the consensus estimated probability of the query sequence or sequence family (alignment) to be related to the PD-(D/E)XK nucleases. In addition, the 'Job details' section provides links to intermediate data files generated along the path of obtaining the final result. These include the multiple sequence alignment used as an input for construction of the query-based HMM, the raw HHsearch results and individual probabilities of different SVM classifiers. If HHsearch aligns query with active site motif(s) of at least one of the members of the positive data set, the annotated multiple sequence alignment is provided through the embedded Jalview (51) applet. The annotated alignment includes predicted secondary structure and the positions of the active site motifs (I–III), enabling the user to get an immediate overview of some of the evidence leading to the assignment.

Underlying sequence databases [derived from the NCBI non-redundant (nr) database] and the HHsearch profile (SCOP-pdb-pfam-pdexk) database are updated on a regular basis. The PD-(D/E)XK positive data set is updated manually as the new superfamily members identified by experimental or computational studies are reported.

DISCUSSION

PD-(D/E)XK domains are found to perform ever expanding set of functions within a wide variety of proteins. However, the reliable identification of protein domains belonging to the PD-(D/E)XK superfamily remains a serious challenge. In part this can be explained by the fact that PD-(D/E)XK domains have a relatively small evolutionary and structurally conserved core (four-stranded mixed β-sheet flanked by an α-helix on each side), which is often outweighted by extensive variable structural elements. The observed plasticity of the PD-(D/E)XK active site, which can be assembled by 'moving' active site residues (except for 'D' from motif-II) around [e.g. (52)], makes the assignment from sequence data alone even more challenging.

Here, we developed an approach for PD-(D/E)XK recognition by combining a state-of-the-art homology detection method (HHsearch) with a machine learning method (SVMs). We considered that the SVM framework will be able to provide means to 'dig up' true assignments even if they are 'buried' among those below the statistical significance level. First, SVMs are known to be well-suited for binary data classification problems. Second, SVMs can be trained on both explicit family-specific (e.g. active site properties) and also on more general profile features.

The results show that the method is able to recognize PD-(D/E)XK domains independently of the statistical significance level assigned to the alignments by HHsearch. In a couple of cases (DUF511, AluI) HHsearch probability was <25%, yet the consensus SVM classifier assigned probabilities >0.8, suggesting the importance of the family-specific training. HHsvm classifiers directed at the non-family-specific homology identification, although performed reasonably well for the PD-(D/E)XK detection task, were less efficient. The difference in the performance was most obvious when either HHsearch probability values for PD-(D/E)XK matches were low or the family had a small number of members, which was true for some REase families. Apparently, in those cases features, specific to PD-(D/E)XK domains such as active site properties played a very important role in distinguishing true from false relationships. Thus, recognition of REases of the PD-(D/E)XK type and identification of their putative active sites may be one of the most prospective applications of our method. On the other hand, the DUF1173 (PF06666) case shows that the absence of the active site signature does not necessarily preclude our method from finding the relationship with PD-(D/E)XK nucleases.

One must bear in mind that the basis for the reported classification of protein families into PD-(D/E)XK and non-PD-(D/E)XK is an alignment and the properties of the aligned regions of two profiles. If none of the query alignments include active site regions of the positive PD-(D/E)XK domain set they are not even considered by SVMs. Therefore, improvements in the description of a

family (profile) and the alignment quality are important for finding new PD-(D/E)XK families. In general, the addition of new sequences to the existing families is expected to automatically improve the performance of the method. However, sometimes the increase in sequence data may aggravate profile corruption by a PSI-BLAST error known as homologous over-extension (53), when alignments begin in a homologous region, but are extended into neighboring non-homologous regions. This may be especially relevant for sequences that have non-PD-(D/E)XK domains highly abundant in nature such as methylase or helicase domains.

We considered that the PD-(D/E)XK recognition method presented here would be most useful if it were accessible to researchers without special training in computational biology. Therefore, we implemented the method as a web server, which could be used in the discovery of new PD-(D/E)XK families and assignment of 'unknown' REases. This method may also be adopted for other 'problematic' protein superfamilies, in which evolutionary relationships are not easily detectable by standard homology search methods.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Johannes Söding for the help regarding HHsearch data formats and for suggestions on the use of secondary structure score weights. The authors also wish to thank Albertas Timinskas for stimulating discussions and useful comments, Virgis Siksnys and Ana Venclovienė for critically reading the manuscript.

FUNDING

Howard Hughes Medical Institute and Ministry of Education and Science of Lithuania. Funding for open access charge: Howard Hughes Medical Institute (grant no. 55005627 to Č.V.).

Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. 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.
3. 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.
4. Kovall,R. and Matthews,B.W. (1997) Toroidal structure of lambda-exonuclease. *Science*, **277**, 1824–1827.
5. 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.
6. 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.
7. 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.
8. Nishino,T., Komori,K., Tsuchiya,D., Ishino,Y. and Morikawa,K. (2001) Crystal structure of the archaeal holliday junction resolvase Hjc and implications for DNA recognition. *Structure*, **9**, 197–204.
9. Middleton,C.L., Parker,J.L., Richard,D.J., White,M.F. and Bond,C.S. (2004) Substrate recognition and catalysis by the Holliday junction resolving enzyme Hje. *Nucleic Acids Res*., **32**, 5442–5451.
10. 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.
11. Dias,A., Bouvier,D., Crepin,T., McCarthy,A.A., Hart,D.J., Baudin,F., Cusack,S. and Ruigrok,R.W. (2009) The cap-snatching endonuclease of influenza virus polymerase resides in the PA subunit. *Nature*, **458**, 914–918.
12. Yuan,P., Bartlam,M., Lou,Z., Chen,S., Zhou,J., He,X., Lv,Z., Ge,R., Li,X., Deng,T. *et al*. (2009) Crystal structure of an avian influenza polymerase PA(N) reveals an endonuclease active site. *Nature*, **458**, 909–913.
13. 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.
14. 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.
15. 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.
16. Venclovas,Č., Timinskas,A. and Siksnys,V. (1994) Five-stranded β-sheet sandwiched with two α-helices: a structural link between restriction endonucleases *Eco*RI and *Eco*RV. *Proteins*, **20**, 279–282.
17. Margelevičius,M., Laganeckas,M. and Venclovas,Č. (2010) COMA server for protein distant homology search. *Bioinformatics*, **26**, 1905–1906.
18. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al*. (2010) The Pfam protein families database. *Nucleic Acids Res*., **38**, D211–D222.
19. 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.
20. 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.
21. Vapnik,V.N. (1995) *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.
22. Söding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.
23. Hildebrand,A., Remmert,M., Biegert,A. and Söding,J. (2009) Fast and accurate automatic structure prediction with HHpred. *Proteins*, **77(Suppl. 9)**, 128–132.
24. Margelevičius,M. and Venclovas,Č. (2010) Detection of distant evolutionary relationships between protein families using theory of sequence profile-profile comparison. *BMC Bioinformatics*, **11**, 89.

25. Dlakic,M. (2009) HHsvm: fast and accurate classification of profile-profile matches identified by HHsearch. *Bioinformatics*, **25**, 3071–3076.
26. Lin,H.T. and Li,L. (2008) Support vector machinery for infinite ensemble learning. *Journal of Machine Learning Res.*, **9**, 285–312.
27. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
28. Holm,L., Kaariainen,S., Rosenstrom,P. and Schenkel,A. (2008) Searching protein structure databases with DaliLite v.3. *Bioinformatics*, **24**, 2780–2781.
29. Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.
30. Halligan,B.D., Ruotti,V., Jin,W., Laffoon,S., Twigger,S.N. and Dratz,E.A. (2004) ProMoST (Protein Modification Screening Tool): a web-based tool for mapping protein modifications on two-dimensional gels. *Nucleic Acids Res.*, **32**, W638–W644.
31. Venclovas,Č. and Margelevičius,M. (2009) The use of automatic tools and human expertise in template-based modeling of CASP8 target proteins. *Proteins*, **77(Suppl. 9)**, 81–88.
32. Ando,T., Aras,R.A., Kusugami,K., Blaser,M.J. and Wassenaar,T.M. (2003) Evolutionary history of hrgA, which replaces the restriction gene hpyIIIR in the hpyIII locus of Helicobacter pylori. *J. Bacteriol.*, **185**, 295–301.
33. Ando,T., Wassenaar,T.M., Peek,R.M. Jr, Aras,R.A., Tschumi,A.I., van Doorn,L.J., Kusugami,K. and Blaser,M.J. (2002) A Helicobacter pylori restriction endonuclease-replacing gene, hrgA, is associated with gastric cancer in Asian strains. *Cancer Res.*, **62**, 2385–2389.
34. Lu,H., Graham,D.Y. and Yamaoka,Y. (2004) The Helicobacter pylori restriction endonuclease-replacing gene, hrgA, and clinical outcome: comparison of East Asia and Western countries. *Dig. Dis. Sci.*, **49**, 1551–1555.
35. Motackova,V., Sanderova,H., Zidek,L., Novacek,J., Padrta,P., Svenkova,A., Korelusova,J., Jonak,J., Krasny,L. and Sklenar,V. (2010) Solution structure of the N-terminal domain of Bacillus subtilis delta subunit of RNA polymerase and its classification based on structural homologs. *Proteins*, **78**, 1807–1810.
36. 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.
37. Townson,S.A., Samuelson,J.C., Vanamee,E.S., Edwards,T.A., Escalante,C.R., Xu,S.Y. and Aggarwal,A.K. (2004) Crystal structure of BstYI at 1.85A resolution: a thermophilic restriction endonuclease with overlapping specificities to BamHI and BglII. *J. Mol. Biol.*, **338**, 725–733.
38. Yang,W. and Steitz,T.A. (1995) Crystal structure of the site-specific recombinase gamma delta resolvase complexed with a 34 bp cleavage site. *Cell*, **82**, 193–207.
39. Reed,R.R. and Moser,C.D. (1984) Resolvase-mediated recombination intermediates contain a serine residue covalently linked to DNA. *Cold Spring Harb. Symp. Quant. Biol.*, **49**, 245–249.
40. Pingoud,A., Fuxreiter,M., Pingoud,V. and Wende,W. (2005) Type II restriction endonucleases: structure and mechanism. *Cell. Mol. Life Sci.*, **62**, 685–707.
41. Li,W., Kamtekar,S., Xiong,Y., Sarkis,G.J., Grindley,N.D. and Steitz,T.A. (2005) Structure of a synaptic gammadelta resolvase tetramer covalently linked to two cleaved DNAs. *Science*, **309**, 1210–1215.
42. 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.
43. 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.
44. Goudenege,D., Avner,S., Lucchetti-Miganeh,C. and Barloy-Hubler,F. (2010) CoBaltDB: Complete bacterial and archaeal orfeomes subcellular localization database and associated resources. *BMC Microbiol.*, **10**, 88.
45. Zhou,X.E., Wang,Y., Reuter,M., Mucke,M., Kruger,D.H., Meehan,E.J. and Chen,L. (2004) Crystal structure of type IIE restriction endonuclease EcoRII reveals an autoinhibition mechanism by a novel effector-binding fold. *J. Mol. Biol.*, **335**, 307–319.
46. Tamulaitis,G., Mucke,M. and Siksnys,V. (2006) Biochemical and mutational analysis of EcoRII functional domains reveals evolutionary links between restriction enzymes. *FEBS Lett.*, **580**, 1665–1671.
47. Janulaitis,A., Vaisvila,R., Timinskas,A., Klimasauskas,S. and Butkus,V. (1992) Cloning and sequence analysis of the genes coding for Eco57I type IV restriction-modification enzymes. *Nucleic Acids Res.*, **20**, 6051–6056.
48. Rimseliene,R. and Janulaitis,A. (2001) Mutational analysis of two putative catalytic motifs of the type IV restriction endonuclease Eco57I. *J. Biol. Chem.*, **276**, 10492–10497.
49. Rimseliene,R., Maneliene,Z., Lubys,A. and Janulaitis,A. (2003) Engineering of restriction endonucleases: using methylation activity of the bifunctional endonuclease Eco57I to select the mutant with a novel sequence specificity. *J. Mol. Biol.*, **327**, 383–391.
50. Jurenaite-Urbanaviciene,S., Kazlauskiene,R., Urbelyte,V., Maneliene,Z., Petrusyte,M., Lubys,A. and Janulaitis,A. (2001) Characterization of BseMII, a new type IV restriction-modification system, which recognizes the pentanucleotide sequence 5′-CTCAG(N)(10/8). *Nucleic Acids Res.*, **29**, 895–903.
51. Waterhouse,A.M., Procter,J.B., Martin,D.M., Clamp,M. and Barton,G.J. (2009) Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
52. Skirgaila,R., Grazulis,S., Bozic,D., Huber,R. and Siksnys,V. (1998) Structure-based redesign of the catalytic/metal binding site of Cfr10I restriction endonuclease reveals importance of spatial rather than sequence conservation of active centre residues. *J. Mol. Biol.*, **279**, 473–481.
53. Gonzalez,M.W. and Pearson,W.R. (2010) Homologous over-extension: a challenge for iterative similarity searches. *Nucleic Acids Res.*, **38**, 2177–2189.