BMC Genomics

BioMed Central

Research article

Open Access

# Identification of a new family of putative PD-(D/E)XK nucleases with unusual phylogenomic distribution and a new type of the active site

Marcin Feder and Janusz M Bujnicki*

Address: Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, Trojdena 4, 02-109 Warsaw, Poland

Email: Marcin Feder - marcin@genesilico.pl; Janusz M Bujnicki* - iamb@genesilico.pl

* Corresponding author

Published: 18 February 2005

*BMC Genomics* 2005, **6**:21 doi:10.1186/1471-2164-6-21

Received: 11 November 2004

Accepted: 18 February 2005

This article is available from: http://www.biomedcentral.com/1471-2164/6/21



## Abstract

**Background:** Prediction of structure and function for uncharacterized protein families by identification of evolutionary links to characterized families and known structures is one of the cornerstones of genomics. Theoretical assignment of three-dimensional folds and prediction of protein function even at a very general level can facilitate the experimental determination of the molecular mechanism of action and the role that members of a given protein family fulfill in the cell. Here, we predict the three-dimensional fold and study the phylogenomic distribution of members of a large family of uncharacterized proteins classified in the Clusters of Orthologous Groups database as COG4636.

**Results:** Using protein fold-recognition we found that members of COG4636 are remotely related to Holliday junction resolvases and other nucleases from the PD-(D/E)XK superfamily. Structure modeling and sequence analyses suggest that most members of COG4636 exhibit a new, unusual variant of the putative active site, in which the catalytic Lys residue migrated in the sequence, but retained similar spatial position with respect to other functionally important residues. Sequence analyses revealed that members of COG4636 and their homologs are found mainly in Cyanobacteria, but also in other bacterial phyla. They undergo horizontal transfer and extensive proliferation in the colonized genomes; for instance in *Gloeobacter violaceus* PCC 7421 they comprise over 2% of all protein-encoding genes. Thus, members of COG4636 appear to be a new type of selfish genetic elements, which may fulfill an important role in the genome dynamics of Cyanobacteria and other species they invaded. Our analyses provide a platform for experimental determination of the molecular and cellular function of members of this large protein family.

**Conclusion:** After submission of this manuscript, a crystal structure of one of the COG4636 members was released in the Protein Data Bank (code 1wdj; Idaka, M., Wada, T., Murayama, K., Terada, T., Kuramitsu, S., Shirouzu, M., Yokoyama, S.: Crystal structure of Tt1808 from *Thermus thermophilus* Hb8, *to be published*). Our analysis of the Tt1808 structure reveals that we correctly predicted all functionally important features of the COG4636 family, including the membership in the PD-(D/E)xK superfamily of nucleases, the three-dimensional fold, the putative catalytic residues, and the unusual configuration of the active site.

## Background

The PD-(D/E)XK domain is ubiquitously found in enzymes involved in metabolism of nucleic acids, mostly in nucleases with diverse biological functions. The first structurally characterized members of the PD-(D/E)XK superfamily were restriction enzymes (REases) (reviews: [1,2]). Crystallographic studies revealed that this superfamily groups together many nucleases with different cellular functions, including: phage λ exonuclease [3], bacterial enzymes exerting ssDNA nicking in the context of methyl-directed and very-short-patch DNA repair: MutH [4] and Vsr [5], Tn7 transposase TnsA [6], a family of archaeal Holliday junction resolvases (Hjc and Hje) from different species of Archaea [7-9], a Holliday junction resolvase (endonuclease I) from phage T7 [10], and an archaeal XPF/Rad1/Mus81 family nuclease that cleaves branched structures generated during DNA repair, replication, and recombination [11].

All members of the PD-(D/E)XK superfamily share a common structural core, comprising a mixed β-sheet of 4 or 5 strands flanked on both sides by α-helices [1,2,12]. These secondary structures are often embedded in very different peripheral elements, which sometimes constitute the majority of the protein. The common β-sheet serves as a scaffold for a weakly conserved active site, typically comprising two or three acidic residues (Asp or Glu) and one Lys residue, which together form the hallmark bipartite catalytic motif (P)D...$X_n$...(D/E)XK (where X is any amino acid). The Lys residue serves to position a water molecule for an in-line attack on the scissile phosphodiester bond, while the carboxylate residues coordinate a $Mg^{2+}$ ion, which acts as a cofactor. Despite the wealth of structural and biochemical data, obtained mainly for REases (summarized in a collection of reviews: [13]), there is still controversy over the exact catalytic mechanism and the number of metal ions required (1, 2, or 3) by PD-(D/E)XK nucleases [14,15]. Moreover, it was found that some members of the PD-(D/E)XK superfamily developed different variants of the active site. In Vsr and its homologs, the (D/E)XK half-motif was replaced by "FxH" and an additional, unique catalytic His residue appeared in another part of the common three-dimensional fold [5]. In some REases, the acidic residue from the (D/E)XK half-motif was found to have "migrated" to another region of the polypeptide in a way that the position of the carboxylate group in the active site is generally maintained as in the "orthodox" members of the PD-(D/E)XK superfamily, despite the side chain is attached to another place in the backbone [16-19]. In a few enzymes, the conserved Lys was found to be replaced by a Glu, Gln, or Asn residue [20-22].

Crystallographic analyses have also revealed the PD-(D/E)XK fold in proteins that do not function as deoxyribonucleases at all and exhibit no conservation of the active site with the above-mentioned enzymes. The structure of the C-terminal catalytic domain of tRNA splicing endoribonuclease (RNase) EndA is identical to the minimal core of the PD-(D/E)XK fold [23], yet this protein lacks the $Mg^{2+}$ binding site common to its cousins that cleave phosphodiester bonds in DNA. Remarkably, on the opposite side of the common fold, EndA developed a different active site, whose geometric configuration is very similar to that of a His-Tyr-Lys triad in structurally unrelated RNase A [24]. Finally, the N-terminal domain (NTD) of the RPB5 subunit of RNA polymerase from *Saccharomyces cerevisiae* exhibits perfect conservation of the restriction enzyme-like structure, but lacks any catalytic residues – it is postulated that it functions as a nucleic acid binding domain devoid of any catalytic activity [25].

The divergence exhibited by the members of the PD-(D/E)XK superfamily is remarkable. Even enzymes with very similar biological functions, such as REases that recognize and cleave the same substrate, can exhibit little or no significant sequence similarity. Thus, most of the afore-mentioned enzymes were considered unrelated until the corresponding crystal structures were solved. Only in a few cases the membership in the PD-(D/E)XK superfamily was successfully predicted using bioinformatics (in some cases backed up by mutagenesis of hypothetical catalytic residues) before the actual structures were determined [26-29]. The catalogue of members of the PD-(D/E)XK superfamily is therefore far from being complete and it is expected that new lineages will be discovered as new sequences appear in the databases. Here, we predict that a large uncharacterized protein family with an unusual phylogenetic distribution is likely to represent a new branch of PD-(D/E)XK nucleases.

## Results

### *Sequence analysis of COG4636 reveals remote similarity to PD-(D/E)XK nucleases*

In the course of analyses of proteins with unknown structures, we came across a family of sequences grouped together in the Clusters of Orthologous Groups (COG) database [30] as COG4636 and annotated as "uncharacterized protein conserved in Cyanobacteria". Analyses of cross-references to other databases revealed no functional information about any member of this family. Nonetheless, preliminary analysis of sequence conservation combined with secondary structure prediction revealed a characteristic pattern of α-helices and β-strands associated with conserved carboxylate residues (review: [31]), which suggested that members of COG4636 may belong to the PD-(D/E)XK superfamily (Figure 1). The multiple sequence alignment revealed nearly perfect conservation of a "PD" half-motif, but only partial conservation of the "(D/E)XK" half-motif. Specifically, instead of the Lys

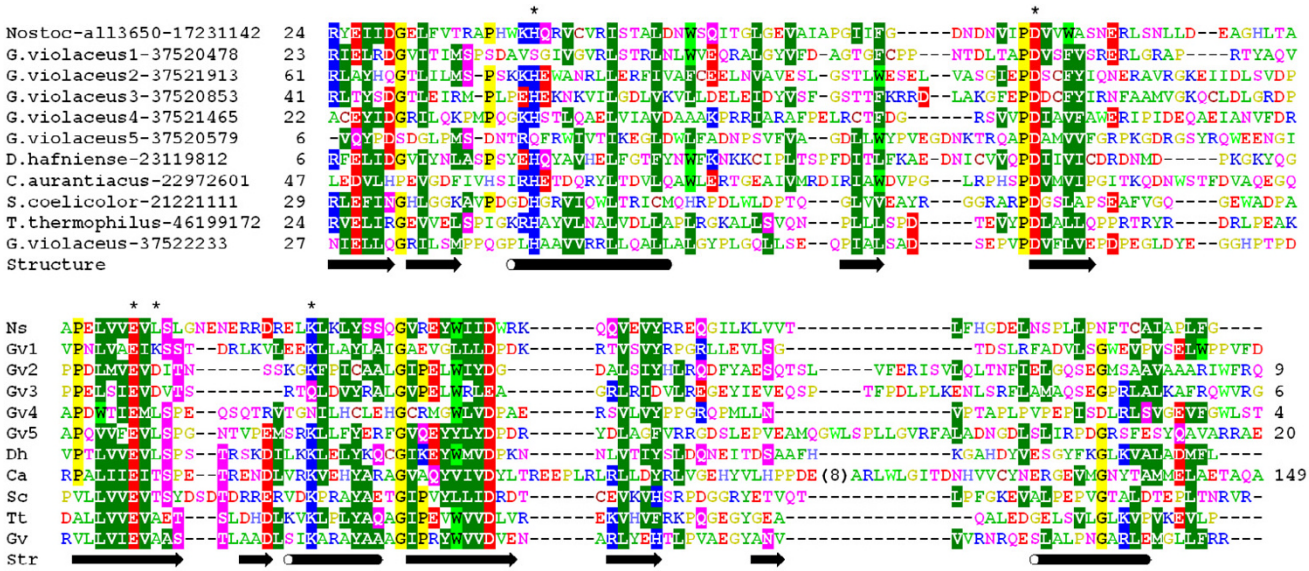


**Figure 1**

**Multiple sequence alignment of selected representatives of the extended COG4636+ family.** The selection of representative sequences includes the modeled protein from Nostoc (motif H-PD-EXX-K, members from *G. violaceus* with different order of putative catalytic residues (Gv1: H-PD-EXK; Gv2: S-PD-EXD-K; Gv3: H-PD-EXD; Gv4: H-PD-EXX-N; Gv5: Q-PD-EXX-K), and members of mono-phyletic clusters from *D. hafniense*, *C. aurantiacus*, *S. coelicolor*, *T. thermophilus*, and *G. violaceus*). The positions of putative catalytic residues are labeled with "*". The variable termini, which could not be confidently aligned, are not shown; the number of omitted residues is indicated. A complete alignment of full-length sequences is available for download from ftp://genesilico.pl/iamb/models/COG4636/. Amino acids are colored according to the physico-chemical properties of their side-chains (negatively charged: red, positively charged: blue, polar: magenta, hydrophobic: green). Conserved residues are highlighted. Elements of predicted secondary structure (helices and strands) are indicated by tubes and arrows, respectively.

residue most members of COG4636 possessed a hydrophobic amino-acid, such as Leu or Val. This suggested that the apparent similarity to the pattern of catalytic residues typical for the PD-(D/E)XK superfamily may be either spurious or indicate a new family of enzymes with an active site devoid of the otherwise conserved residue. We searched for homologs of the analyzed family, beyond sequences from complete genomes grouped together in COG4636, by carrying PSI-BLAST searches of the nr database. Altogether, we collected 435 sequences with significant similarity to COG4636, which will be hereafter referred to as "COG4636+". No statistically significant sequence similarity was detected to any protein with an experimentally determined function.

In order to test the hypothesis of the evolutionary connection between COG4636 and the PD-(D/E)XK superfamily we carried out the fold-recognition analysis, which allows to predict the three-dimensional fold of the target protein by matching its sequence with the available protein structures and assessing the sequence-structure compatibility using a combination of criteria, such as sequence similarity, match of secondary structure elements, compatibility of residue-residue contacts, etc. (review: [32]). Sequences of individual members of COG4636 were therefore submitted to the GeneSilico protein fold-recognition metaserver [33]. Disappointingly, no methods reported statistically significant matches between these sequences and proteins with known structures. Only a few threading methods that explicitly use the structural information from the templates (FUGUE, INBGU, mGenTHREADER, SAM-T02, and 3DPSSM) reported, in some cases, matches to structures of PD-(D/E)XK nucleases, but never at the first position of the ranking. However, in the course of CASP-5 protein structure prediction contest we found that the fold-recognition operation for strongly diverged proteins can be greatly improved by limiting the analysis to the conserved core, i.e. omission of strongly diverged regions and non-conserved insertions, as well as using a refined multiple sequence alignment rather than allowing the servers to build their own sequence profiles from unrefined PSI-BLAST results [34]. Thus, we modified the


**Figure 2**

**Fold-recognition alignment between all3650 and structures of Hjc and Hje.** Amino acids are colored according to the physico-chemical properties of their side-chains. Conserved residues are highlighted. Secondary structure elements experimentally identified in Hjc and Hje and predicted for all3650 are shown between the target and the template sequences. Known and predicted catalytic residues are indicated by "*" (above the alignment for the target, below the alignment for the templates).

multiple sequence alignment of the COG4636+ family by removing strongly diverged termini that could not be reliably aligned, and submitted to the meta-server only the core section, comprising ca. 110 aa. This time, as expected, fold-recognition analysis of a well-defined protein core gave unambiguous results: mGenTHREADER, SPARKS, and FUGUE reported structures of Holliday junction resolvases Hjc and Hje, members of the PD-(D/E)XK fold [7-9], at the first positions of their rankings, with significant scores (0.45, -2.08, and 3.46, respectively). Results obtained from the primary servers have been supported by the consensus server Pcons [35], which reported the Hjc and Hje enzymes at the first four position of its ranking, with scores 1.38-1.20, compared to the insignificant score 0.61 for the subsequent fold in the ranking.

### *Modeling and model-based identification of a putative active site*

In order to identify the putative active site of newly predicted members of the PD-(D/E)XK superfamily, we modeled the structure of one of the COG4636+ members, whose sequence was close to the consensus calculated for the whole family (hypothetical protein all3650 from *Nostoc* sp. PCC 7120, GI: 17231142) and used it as a platform to study the three-dimensional arrangement of conserved residues. A homology model of all3650 was constructed using the "FRankenstein's Monster" approach (see Methods and ref. [34]), starting with the unrefined alignments between the consensus sequence and the structures of Hjc and Hje enzymes (1gef, 1hh1, and 1ob8) reported by threading methods. Initially, the model of the protein core was constructed by iterating the homology modeling procedure, evaluation of the sequence-structure fit by VERIFY3D, merging of fragments with best scores, and local realignment in poorly scored regions. Local realignments were constrained to maintain the overlap between the secondary structure elements found in the template structures, and those predicted for the target. This procedure was stopped when the regions in the protein core (helices and strands) obtained acceptable VERIFY3D score (>0.3) or their score could not be improved by any manipulations, while the average VERIFY3D score for the whole model could not be improved. The final alignment between all3650 and the three structures used as templates is shown in Figure 2. The final model of the core, comprising residues 39–188, obtained a poor average VERIFY3D score of 0.13 due to low scores in the variable loops that could not be modeled with confidence. However, the secondary structure elements (with the exception of the C-terminal helix), obtained an acceptable average score of 0.37. It is important to note that all catalytic residues of the PD-(D/E)XK fold are found in the stable regions of regular secondary structure rather than in loops [36]. The variable N-terminus, which could not be modeled because of the strong divergence and the lack of appropriate template structures, was added "de novo" using the fragment insertion method ROSETTA [37]. The coordinates of the final, full-length model (Figure 3) are available as supplementary material [see Additional file 1] and on-line at ftp://genesilico.pl/iamb/models/COG4636/

The model of all3650 reveals a typical PD-(D/E)XK nuclease-like spatial arrangement of one Lys ε-amino group (from the residue K127) and two carboxylate groups (from residues D83 and E111) (Figure 4). The modeled

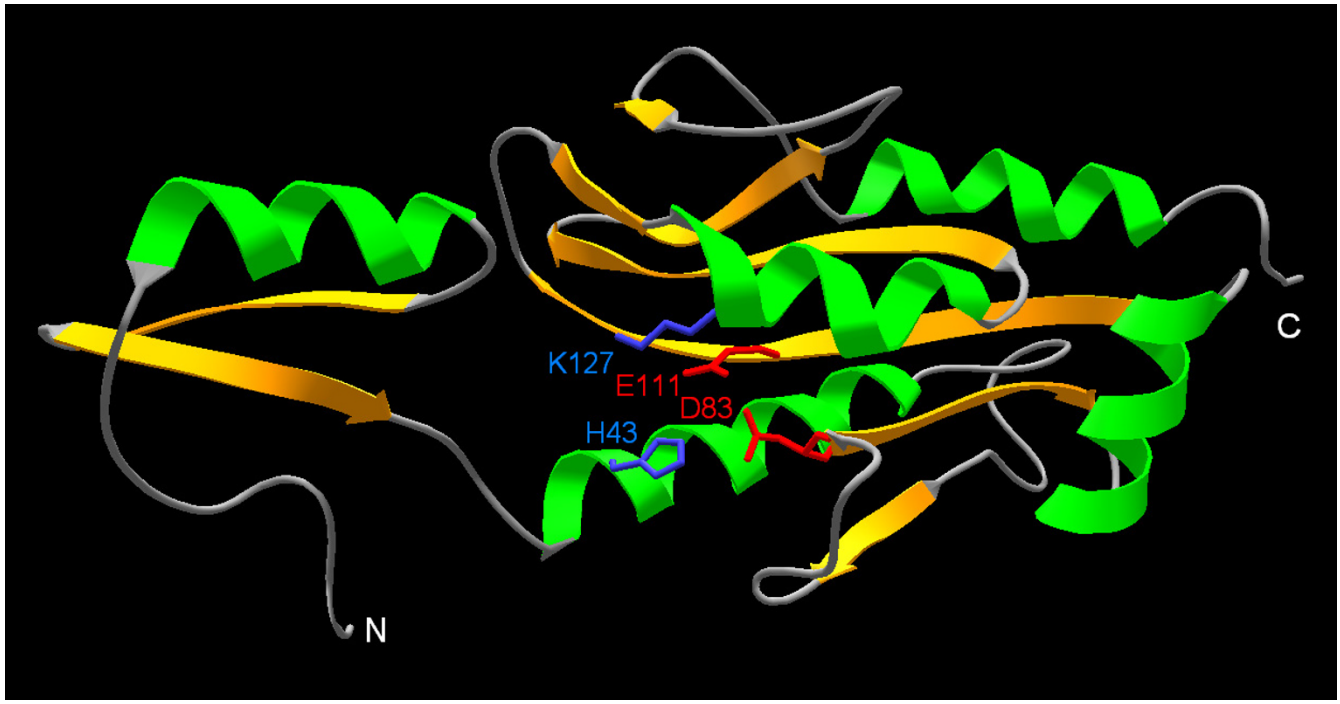


**Figure 3**
**Homology model of all3650.** Helices and strands are shown in green and yellow, respectively. The predicted catalytic residues are shown in the wireframe representation and labeled. The termini are indicated.

structure suggest also an additional highly conserved His residue (H43) that could be a part of the metal ion-binding site or be involved in substrate-binding. Strikingly, in all3650 as well as in the great majority of sequences from the COG4636+ family, the conserved Lys (K127 in all3650) is found not in the common position in the same β-strand as the conserved Glu residue (E111 in all3650), but in a spatially adjacent α-helix. Thus, the predicted active site is formed by a "PD-EXX-K" sequence motif. This "migration" of the presumptive catalytic Lys residue and retention of the original position of the spatially adjacent carboxylate in COG4636+ members resembles the situation reported for a number of restriction enzymes such that as Cfr10I, NgoMIV, Ecl18kI, SsoII, and PspGI [16,18,19,38]. In the latter enzymes, however, it is the carboxylate that is relocated and the original position of the Lys residue is retained, such the active site is formed by a "PD-XXK-E" sequence motif (Figure 4).

Inspection of the multiple sequence alignment reveals that only two carboxylates (corresponding to D83 and E111 in all3650) are practically invariant in the COG4636+ family, while all the others undergo various substitutions (Figure 1). In a small group of sequences (represented by a hypothetical protein gll0909 from *G. violaceus*, GI: 37520478) the Lys residue is present both at the "classical" and alternative position, thereby forming a "PD-EXK-K" variant of the active site. This arrangement resembles a putative evolutionary intermediate between the "classical" active site and the newly discovered rearranged variant. In another lineage of the COG4636+ family, an Asp residue appears in the position normally occupied by Lys in the C-terminal half-motif. Some of the members of this lineage (exemplified by glr2344 from *G. violaceus*, GI: 37521913) exhibit therefore the "PD-EXD-K" motif, but the majority (exemplified by hypothetical protein glr1284 from *G. violaceus*, GI: 37520853) lack the Lys residue and exhibit only the "PD-EXD" variant. In another lineage (represented by gll1896 from *G. violaceus*, GI: 37521465) the Lys residue is replaced by Asn to form the "PD-EXX-N" variant of the predicted active site. The conserved His residue (H43 in all3650) is present in most members of the COG4636+ family, with the exception of a small lineage of closely related proteins (represented by gll1896 from *G. violaceus*, GI: 37520579) in which it is substituted by Gln, and a larger group of more diversified sequences, in which it is substituted by Thr or Ser. Most members of the latter group possess a Lys or Arg residue in the "catalytic" position and hence exhibit "PD-EXK-K" (see above) or "PD-EXR-K" variants of the active site. It

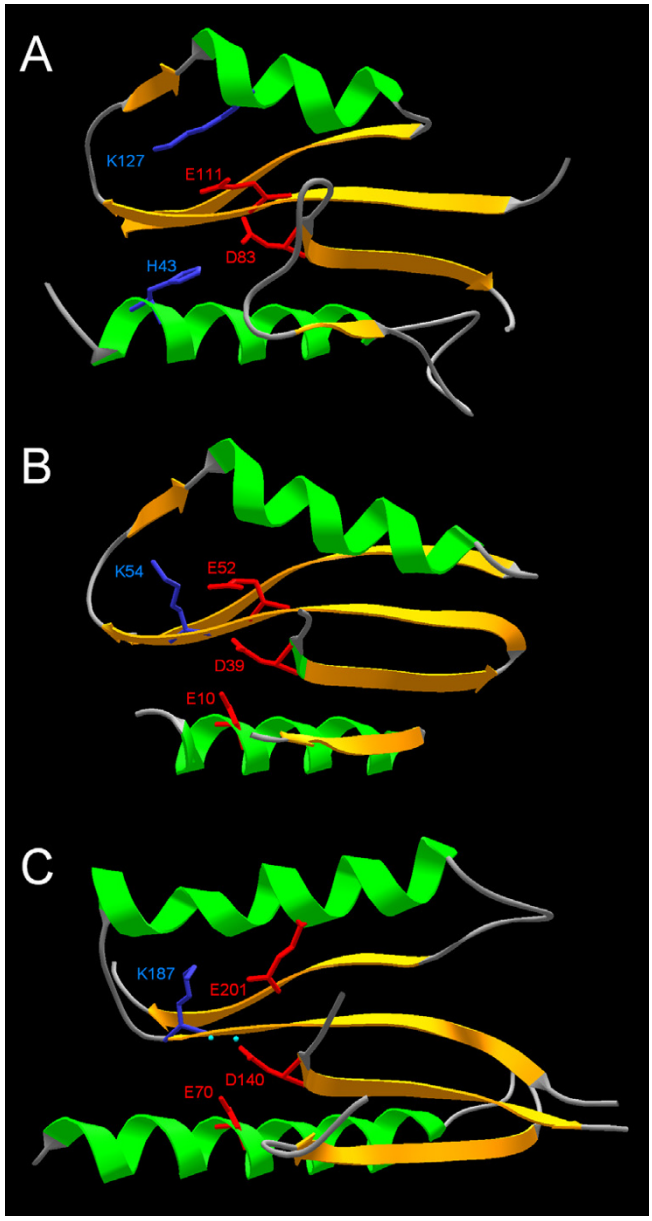


**Figure 4**
**Spatial conservation of the PD-(D/E)XK active site in all3650, Hjc, and NgoMVI.** A) The predicted structure of all360 is shown in the same orientation as the crystal structures of the bona fide PD-(D/E)XK nucleases: B) Holliday junction resolvase Hje (1ob8 in PDB [9]) and C) REase NgoMIV (1fiu in PDB [78] to illustrate the spatial conservation of side-chains in the active site (the carboxylate residues in red and the Lys residue in blue), despite the lack of their conservation in the PD-EXX-K, PD-DXK, and PD-XXK-E variants of the sequence motif. Only the common core is shown, terminal regions and insertions have been omitted for clarity of the presentation.

will be very interesting to determine experimentally, which of those residues in different configurations are involved in catalysis, and which are only auxiliary. In particular, it would be interesting to find if both or either of the Lys residues present in the potential "intermediate" versions of the active site are required for catalysis.

### *Phylogenomic analysis of the COG4636+ family*

Sequence searches of the nr database at the NCBI revealed that the great majority of members of the COG4636+ family (382 of total 435) originate from Cyanobacteria; of these, 84% were found in just 6 genomes (*G. violaceus* PCC 7421, *Nostoc punctiforme* PCC 73102, *Crocosphaera watsonii* WH 8501, *Nostoc* sp. PCC 7120, *Anabaena variabilis* ATCC 29413, *Synechocystis* sp. PCC 6803). It is astonishing that members of COG4636+ represent over 2% of all protein-encoding genes of *G. violaceus* PCC 7421 (95 of 4430 total [39]), other completely sequenced genomes of Cyanobacteria are completely devoid of them or encode only 1 or 2 sequences from this family. We were not able to identify any members of the COG4636+ family in the sequences derived from seawater samples collected from the Sargasso Sea [40] and deposited in the "environmental samples" database at the NCBI. Since the prevalent Cyanobacteria found in the Sargasso Sea are *Synechococcus* and *Prochlorococcus*, the lack of COG4636+ members in the environmental samples is in good agreement with the paucity of these genes in the fully sequenced genomes of these species.

In order to reconstruct the evolutionary history of the COG4636+ family, we calculated the phylogenetic tree, based on the same reliable section of the multiple sequence alignment that was used for protein structure prediction (see Methods). Unfortunately, in all trees obtained with different methods and parameters, the majority of deep branches received very low bootstrap support (data not shown), hence the relationships within the whole family must be regarded as unresolved. We were able, however, to identify a number of branches with bootstrap support >90%. Many of such branches comprise members from one species only. This situation is characteristic for sequences found in a few non-Cyanobacterial species; for instance 8 sequences from *D. hafniense* DCB-2 (Firmicutes), 7 sequences from *C. aurantiacus* (Chloroflexales), and 6 from *S. coelicolor* (Actinobacteria) each form a separate species branch on the phylogenetic tree, while 14 sequences from *T. thermophilus* HB27 (Deinococcus-Thermus lineage) form three separate branches. Several monophyletic groups of closely related sequences are also observed in *G. violaceus* (e.g. a sub-family comprising 7 sequences with GI numbers: 37522824, 37520777, 37522646, 37521452, 37522233, 37520151, 37522558). There is also one branch comprising 6 closely related sequences in *C. watsonii*, GI numbers: 45527153,

45527776, 45524526, 45527777, 45527775, 45527774). Other statistically significant branches, however, comprise members from different species, suggesting that they were either formed prior to speciation or that their members were transmitted horizontally between different genomes of already existing species.

To identify if members of COG4636+ are encoded by any known mobile genetics elements or if they are preferentially associated with any other proteins, we analyzed the genomic neighborhood of all members of the family. Although we carefully examined annotations of predicted open reading frames (ORFs) in the range of 3000 bp upstream and downstream, we weren't able to identify any recurrent type of proteins, either with respect to the molecular or cellular function or the predicted three-dimensional fold (data not shown). Also no preference for occurrence of COG4636+ family members within or near any apparent mobile genetic elements (putative prophages etc.) was observed. Thus, insertion of the genes encoding putative COG4636+ nucleases seems virtually random. The only notable exception is a neighborhood of another member of COG4636+, suggesting tandem duplication. We identified one instance of 4 consecutively arranged genes in the genome of *C. watsonii* WH8501, all from the above-mentioned branch of 6 closely related sequences (the other two relatives are located elsewhere on the chromosome). We also found a few tandem duplications: 9 in *C. watsonii* WH8501 and 5 in *G. violaceus* PCC7421, 5 in Nostoc sp. PCC6803, 2 in *N. punctiforme* PCC73102, 2 in *A. variabilis* ATCC 29413, 2 in Synechocystis sp. PCC6803, 2 in *T. thermophilus* HB27, 1 in *T. erythraeum* IMS101 and 1 in *M. magnetotatcticum* MS-1. In general, however, tandem duplications are rare and the distribution of COG4636+ family members along the chromosomes of Cyanobacteria with completed genomes seems completely erratic (Figure 5).

## Discussion

Our results suggest that functionally uncharacterized proteins grouped together in COG4636 are a branch of the PD-(D/E)XK superfamily, which has not been identified to date due to a presence of an unusual variant of the active site, which lacks the conserved Lys residue at the typical position in the primary sequence. That the catalytic Lys can migrate in the framework of the active site of PD-(D/E)XK nucleases has been suggested earlier, based on the sequence analysis of another nuclease domain found in site-specific, non-long terminal repeat retrotransposable elements [2], but to date no molecular model was offered to suggest the alternative point for the attachment of the side chain to the protein backbone. Our sequence analysis of the COG4636+ family and the structural model of one of its members explain the problems with identification of the PD-(D/E)XK motif on the sequence level and provide a platform for further studies. Specifically, our analysis points at the most interesting members of the family, which display previously not observed variants of the PD-(D/E)XK active site. Experimental analyses of these proteins and determination of the role of individual amino acids in the evolutionary context may help to better understand the plasticity of the PD-(D/E)XK active site and may settle down the controversy in the field of nucleases regarding the mechanism(s) of the reaction.

Phylogenomic analyses show that putative nucleases grouped in the COG4636+ family are exceptionally abundant in genomes of certain Cyanobacteria, but absent in others. They are typically abundant in the sequenced genomes of freshwater species, but scarce in the genomes of marine species, with the exception of *C. watsonii* WH 8501, which was isolated from tropical waters of the Western Atlantic and Pacific oceans. It is remarkable that members of COG4636+ are almost absent from the genomes of *Synechococcus* and *Prochlorococcus* species thriving in the Sargasso sea, as well as in the environmental samples isolated from that region. On the other hand, in *G. violaceus* PCC 7421 they comprise over 2% of all protein-encoding genes. This phylogenetic distribution resembles that of mobile genetic elements such as introns or insertion sequences (reviews: [41,42]) and suggests that the contemporary COG4636+ family originated from a few predecessors that underwent extensive horizontal gene transfer and massive proliferation in certain genomes. Monophyly of COG4636+ sequences in non-Cyanobacterial species strongly suggests that proliferation occurred in each of these species independently, following a single event of colonization by horizontal transfer from a Cyanobacterium (or in the case of *T. thermophilus* – three independent successful colonizations).

We hypothesize that the mechanism by which these putative nucleases induce their proliferation in a genome is similar to that displayed by homing nucleases and restriction enzymes [43], namely to incise the DNA by introducing nicks or double-strand breaks, which stimulates recombination and may lead to tandem duplications and a variety of genomic rearrangements [44-47]. Frequent cleavage of the genomic DNA would be lethal for the cell, therefore if members of COG4636+ are indeed active as nucleases, then they should target rare sequences (in a manner similar to homing endonucleases; review: [48]) or unusual structures in the DNA (similarly to the structure-specific Holliday junction resolvases), or their activity would have to be somehow regulated (inhibited) by interactions with other proteins or cellular processes (for instance by DNA modification). There are known examples of Holliday junction resolvases carried on defective lambdoid prophages [49]. Unfortunately, analysis of the genomic neighborhood shows no preferred association of

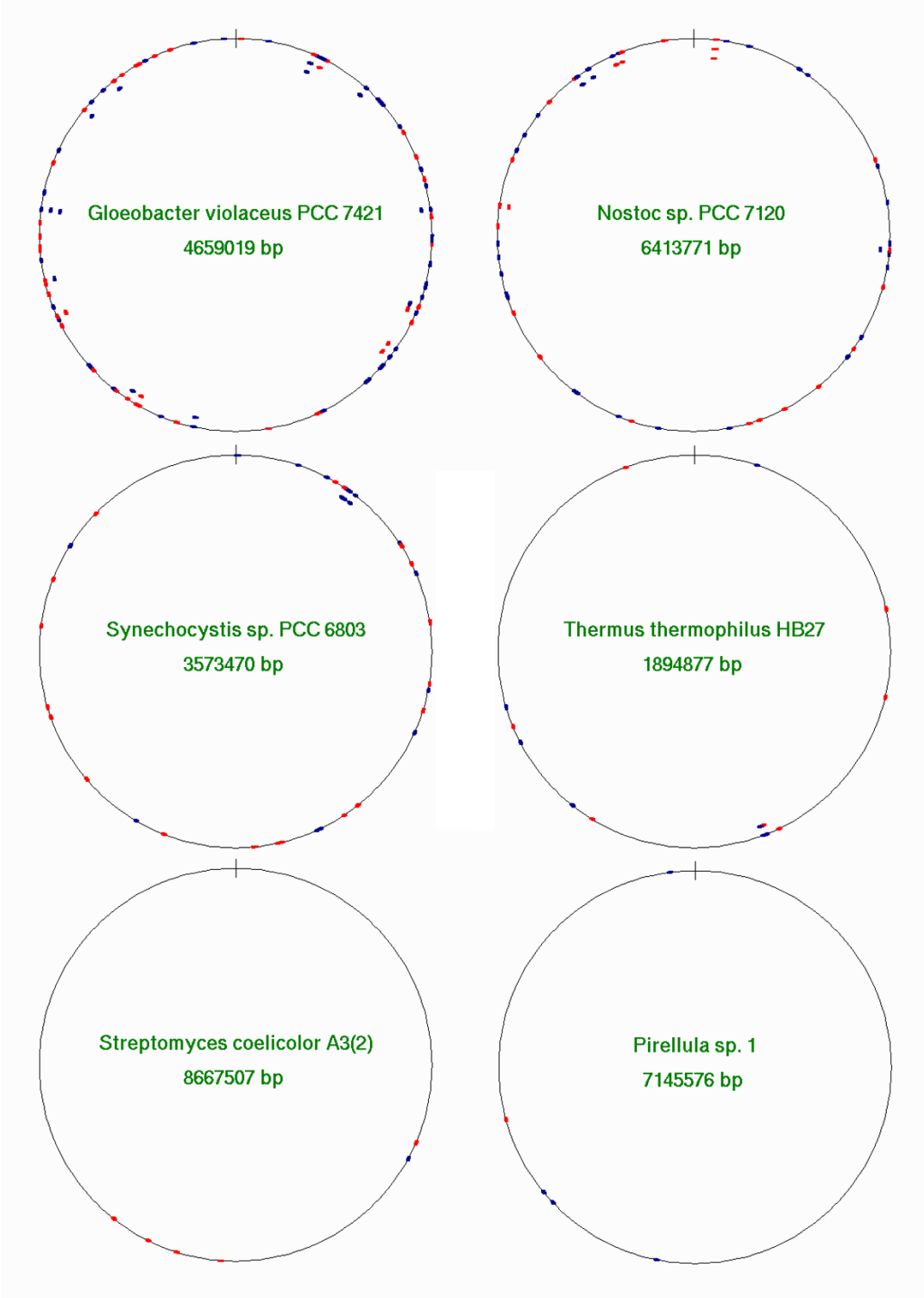


**Figure 5**
**Localization of COG4636+ family members in the chromosomes of Cyanobacteria with completed genomes.** Circular chromosome maps of genomes with at least three genes encoding COG4636+ members (indicated by dots). Genes shown in dark blue are transcribed clockwise (positive reading frame) and those in red are transcribed anticlockwise (negative reading frame). Dots plotted inside the circle indicate that more than one gene is localized in the same region of the map (1/360 of the genome length).

COG4636+ members with any mobile genetic elements or particular gene families that could give us hints about the cellular processes they could be part of or suggest how their predicted nuclease activity could be inhibited or regulated. Especially, we found no correlation with the presence of known or putative methyltransferases. This suggests that despite sharing the common PD-(D/E)XK fold with REases, COG4636+ members are unlikely to serve as parts of restriction-modification systems, which are known to be abundant in Cyanobacteria [50,51]. It must be noted, however, that multiple solitary DNA methyltransferases were reported in *Anabaena* PCC 7120 [51], and these enzymes could potentially provide protection against the cleavage of the chromosomal DNA by at least some of the COG4636+ members found in this organism.

One possibility is that COG4636+ members serve as a part of the restriction barrier, similarly to the unrelated NucA family of extracellular nucleases found in Cyanobacteria, e.g. *Anabaena* sp. PCC 7120 [52] and *Microcystis* sp. [53]. They could also fulfill a role in maintenance of the identity of the species by controlling the flow of incoming DNA, as recently suggested for restriction-modification systems [54]. From the genomic analyses it appears, however, that the primary function of COG4636+ members is to spread and multiply, and their cellular roles may be merely side-effects of this selfish expansion. It is very likely that their nuclease activity is recombinogenic and may increase the frequency of genomic rearrangements. Moreover, the multiplication of closely related COG4636+ members in certain genomes leads to an abundance of dispersed related DNA sequences, which by themselves may increase the frequency of genome rearrangements by homologous recombination. It was suggested that in the marine Cyanobacteria the factors that increase the genome plasticity might not be promoted by natural selection due to the homeostatic environment of the open ocean [55]. Conversely, the unstable environment of fresh waters might promote the spreading of factors that destabilize the genome by increasing the frequency of recombination and thereby increase the diversity of the population. This is in good agreement with our finding of prevalence of COG4636+ members in Cyanobacteria that thrive in fresh waters and their paucity in marine species (with the exception of *C. watsonii* WH 8501). Summarizing, it is plausible that members of COG4636+ fulfill an important role in the genome dynamics of Cyanobacteria and other species they colonize. We hope that our predictive study will facilitate experimental determination of the molecular and cellular function of members of this intriguing protein family.

## Methods

### *Sequence analysis*

Searches of the non-redundant (nr) database were carried out at the NCBI using PSI-BLAST [56] with default parameters, using different sequences from COG4636 as queries. Significantly similar sequences were retrieved from all searches and pooled together. Identical sequences from the same organism were removed. A multiple sequence alignment was generated using MUSCLE [57] with default parameters and subsequently adjusted manually, based on the analysis results of secondary structure prediction (see below), to ensure that no unwarranted gaps are introduced within α-helices and β-strands. Phylogenetic inference was carried out using the reliable central section of the multiple sequence alignment. The matrix of pairwise distances was calculated from sequences according to the JTT model [58] with gaps treated as missing data. The neighbor-joining (NJ) tree was inferred according to the method of Saitou and Nei [59].

### *Phylogenomic analysis*

The Eutils module from the Biopython package was used as an interface to access remotely the NCBI databases [60]. The Gene Identification numbers of proteins included in the final multiple alignment sequences were used to identify the corresponding GenPept entries, which were downloaded into a local Barkeley database using an in-house developed parser based on the SAX package http://sourceforge.net/projects/pyxml. The "coded_by" field from each GenPept file was used to identify the corresponding DNA sequence, which were also downloaded into the database. The sequence in the range of 3000 bp upstream or downstream from the region encoding a COG4636+ member were scanned for the presence of annotated Open Reading Frames (ORFs). Initially, the functional categorization of these ORFs was carried out based on the automatic assignment into the PFAM and COG families. In the absence of any recurrent function, the annotations of all ORFs were carefully re-analyzed visually and in uncertain cases, additional searches against the CDD database were carried out [61]. The distribution of COG4636+ members on the chromosome maps was visualized using a program developed in-house specifically for that purpose.

### *Protein structure prediction*

Secondary structure prediction and tertiary fold-recognition was carried out via the GeneSilico meta-server gateway at http://genesilico.pl/meta/[33]. Secondary structure was predicted using PSIPRED [62], PROFsec [63], PROF [64], SABLE [65], JNET [66], JUFO [67], and SAM-T02 [68]. Solvent accessibility for the individual residues was predicted with SABLE [65] and JPRED [66]. The fold-recognition analysis (attempt to match the query sequence to known protein structures) was carried out using FFAS03 [69], SAM-T02 [68], 3DPSSM [70], BIOINBGU [71],

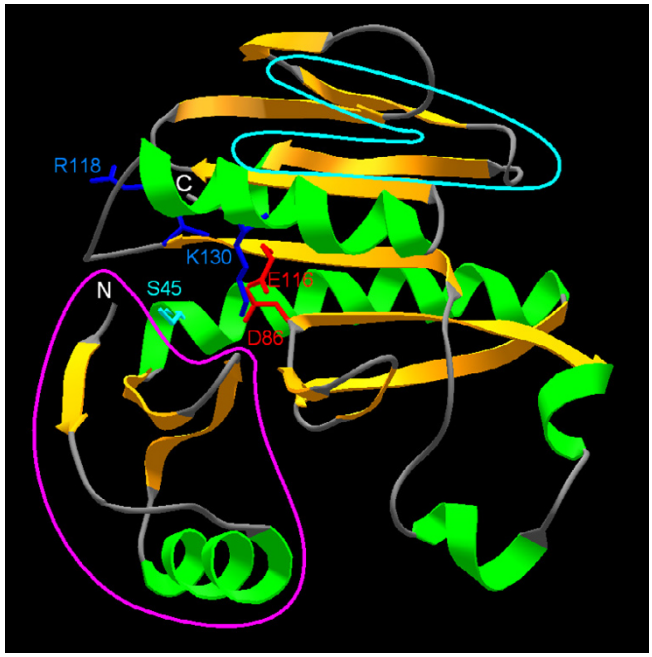


**Figure 6**
**The crystal structure of Tt1808 (1wdj in PDB).** Tt1808 is shown in the same orientation and is colored and labeled in the same way as the homology model of all3650 on Figure 3. Two regions of differences between Tt1808 and the model of all3650 are indicated: the N-terminal subdomain has a similar fold, but different orientation (magenta line) and the C-terminal region folds as a β-harpin (cyan line) rather than as an α-helix.

FUGUE [72], mGENTHREADER [73], and SPARKS [74]. Fold-recognition alignments reported by these methods were compared, evaluated, and ranked by the Pcons server [35].

### *Homology modeling*

Fold-recognition alignments to the structures of selected templates were used as a starting point for homology modeling using the "FRankenstein's Monster" approach [34], comprising cycles of model building, evaluation, realignment in poorly scored regions and merging of best scoring fragments. The positions of predicted catalytic residues and secondary structure elements were used as spatial restraints. Briefly, preliminary models were generated based on the alignments to various template structures returned by the FR servers. The sequence-structure fit in these models was assessed using VERIFY3D [75] and visualized using the COLORADO3D server [76]. The most common and best-scoring fragments were merged to produce a hybrid model, in which the sequence-structure was re-evaluated. In the poorly scoring fragments the alignment was locally modified by shifting the sequences within the limits of predicted secondary structures and a next generation of models corresponding to different alignments was generated. The cycles of evaluation of models, generation of hybrids and local re-alignment in problematic regions continued until the global VERIFY3D score could not be improved. Regions, which could not be modeled because of the lack of the appropriate template structure, were added "de novo" using the fragment insertion method ROSETTA [37].

## Note added in Proof

After submission of this manuscript, a crystal structure of one of the COG4636+ members was released in the Protein Data Bank (code 1wdj; Idaka, M., Wada, T., Murayama, K., Terada, T., Kuramitsu, S., Shirouzu, M., Yokoyama, S.: Crystal Structure of Tt1808 from *Thermus thermophilus* Hb8 To be Published). Our analysis of the Tt1808 structure and its comparison with the model of all3650 confirms our predictions. Tt1808 does indeed exhibit the PD-(D/E)xK fold: the DALI [77] search of the the Protein Data Bank (PDB) database with 1wdj revealed that its 8 closest structural matches with Z-scores in a range of 5.3-3.7 are members of the PD-(D/E)xK superfamily, including the Holliday junction resolvases we used as templates to model the all365 protein. Analysis of the Tt1808 structure (Figure 6) reveals that we correctly predicted the topology of the catalytic domain in all365. We only mispredicted an α-helix in the C-terminus of all365; in Tt1808 this element is replaced by a β-hairpin. We have also successfully modeled the structure of the N-terminal subdomain but failed to predict the interaction between this part and two loops of the catalytic domain (compare Figure 3 and Figure 6). It is important to note that these errors concern regions that do not influence any of our functional interpretations based on the all3650 model. Most importantly, the identity of presumed catalytic residues of all365 was predicted correctly, including the postulated unusual position of the Lys residue (in our model of all365 the side chain of K127 has a different orientation than K130 in Tt1808, but such details are irrelevant to our functional interpretations). It is interesting to note that Tt1808 has the S-PD-EXR-K variant of the active site, and that the side chain of the R118 residue, which replaced the "classical" catalytic Lys, points away from other catalytic residues, on the opposite side of the loop between the "EXR" and "K" elements. Summarizing, we correctly predicted all functionally important features of the COG4636+ family, including the membership in the PD-(D/E)xK superfamily of nucleases, the three-dimensional fold, the putative catalytic residues, and the unusual configuration of the active site.

**Table 1: Distribution of COG4636+ family members among different bacteria.**

| organism / genome | phylum | habitat | data source | COG4636+ members | |
|---|---|---|---|---|---|
| | | | | total | disrupted |
| Gloeobacter violaceus PCC 7421 | Cyanobacteria | calcareous rock | C | 95 | 1 |
| *Nostoc punctiforme* PCC 73102 | Cyanobacteria | cycad (endosymbiont) | WGS | 71 | 7 |
| *Crocosphaera watsonii* WH 8501 | Cyanobacteria | marine water | WGS | 62 | 1 |
| *Nostoc* sp. PCC 7120 | Cyanobacteria | fresh water | C | 58 | 1 |
| *Anabaena variabilis* ATCC 29413 | Cyanobacteria | fresh water | WGS | 45 | 5 |
| *Synechocystis* sp. PCC 73102 | Cyanobacteria | fresh water | C | 36 | 1 |
| *Thermus thermophilus* HB27 | Deinococcus-Thermus | thermal environment | C | 14 | - |
| *Trichodesmium erythraeum* IMS101 | Cyanobacteria | marine water | WGS | 10 | 3 |
| *Desulfitobacterium hafniense* DCB-2 | Firmicutes | sewage sludge | WGS | 8 | - |
| *Chloroflexus aurantiacus* | Chloroflexi | fresh water (hot springs) | WGS | 7 | - |
| *Streptomyces coelicolor* A3(2) | Actinobacteria | soil | C | 6 | - |
| *Rhodopirellula baltica* SH 1 | Planctomycetes | marine water | C | 5 | 1 |
| *Moorella thermoacetica* ATCC 29413 | Firmicutes | fresh water (ponds) | WGS | 3 | - |
| *Deinococcus radiodurans* R1 | Deinococcus-Thermus | unknown | C | 3 | - |
| *Magnetospirillum magnetotacticum* MS-1 | Proteobacteria | fresh water (ponds) | WGS | 2 | 1 |
| *Synechococcus elongatus* PCC 73102 | Cyanobacteria | fresh water | WGS | 2 | - |
| *Aquifex aeolicus* VF5 | Aquificae | fresh water (hot springs) | C | 2 | - |
| *Kineococcus radiotolerans* SRS30216 | Actinobacteria | unknown (isolated from radioactive work area) | WGS | 2 | - |
| *Caulobacter crescentus* CB15 | Proteobacteria | fresh water | C | 1 | - |
| *Thermosynechococcus elongatus* BP-1 | Cyanobacteria | fresh water (hot springs) | C | 1 | - |
| *Synechococcus* sp. PCC 73102 | Cyanobacteria | brackish (euryhaline) and/or marine water | UGS | 1 | - |
| *Microcystis aeruginosa* | Cyanobacteria | fresh water (lakes, ponds and rivers) | NR | 1 | - |
| Prochlorococcus marinus str. MIT9313 | Cyanobacteria | marine water | C | - | - |
| *Prochlorococcus marinus* subsp. marinus CCMP1375 | Cyanobacteria | marine water | C | - | - |
| *Prochlorococcus marinus* subsp. pastoris CCMP1986 | Cyanobacteria | marine water | C | - | - |
| *Synechococcus* sp. WH 8102 | Cyanobacteria | marine water | C | - | - |

C – Completed genomic sequence, WGS – Whole Genome Shotgun, UGS – Unfinished Genomic Sequence, NR – non-redundant database (NCBI). ORFs were regarded as "disrupted" if they bear frameshift mutations or stop codons.

## List of abbreviations

aa, amino acid(s); bp, base pair(s); nt, nucleotide; e, expectation; REase, restriction endonuclease; ORF, product of an open reading frame,

## Authors' contributions

MF carried out all sequence analyses and structure predictions using fold-recognition methods and ROSETTA. JMB built the homology model, analyzed spatial vs. sequential conservation of the putative active site, and wrote the manuscript. Both authors have read and accepted the final version of the manuscript.

## Additional material

### Additional File 1

*The additional data file all3650.pdb contains the coordinates of the original all3650 model (obtained before the Tt1808 structure was published) in the PDB format.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-6-21-S1.pdb]

## Acknowledgements

This analysis was funded by KBN (grant 3P04A01124 to JMB). JMB was also supported by the EMBO/HHMI Young Investigator Award and by a Fellowship from the Foundation for Polish Science. The work of MF was supported by the NIH (Fogarty International Center grant R03 TW007163-01).

## References


1. Aggarwal AK: **Structure and function of restriction endonucleases.** *CurrOpinStructBiol* 1995, **5:**11-19.
2. Bujnicki JM: **Molecular phylogenetics of restriction endonucleases.** In *Restriction Endonucleases Volume 14*. Edited by: Pingoud A. Berlin, Springer-Verlag; 2004:63-87.
3. Kovall RA, Matthews BW: **Structural, functional, and evolutionary relationships between lambda-exonuclease and the type II restriction endonucleases.** *Proc Natl Acad Sci U S A* 1998, **95:**7893-7897.
4. Ban C, Yang W: **Structural basis for MutH activation in E.coli mismatch repair and relationship of MutH to restriction endonucleases.** *Embo J* 1998, **17:**1526-1534.
5. Tsutakawa SE, Muto T, Kawate T, Jingami H, Kunishima N, Ariyoshi M, Kohda D, Nakagawa M, Morikawa K: **Crystallographic and functional studies of very short patch repair endonuclease.** *Mol Cell* 1999, **3:**621-628.
6. Hickman AB, Li Y, Mathew SV, May EW, Craig NL, Dyda F: **Unexpected structural diversity in DNA recombination: the restriction endonuclease connection.** *Mol Cell* 2000, **5:**1025-1034.
7. 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 (Camb)* 2001, **9:**197-204.
8. 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:**5509-5514.
9. 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:**5442-5451.
10. 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:**62-67.
11. Nishino T, Komori K, Ishino Y, Morikawa K: **X-ray and biochemical anatomy of an archaeal XPF/Rad1/Mus81 family nuclease: similarity between its endonuclease domain and restriction enzymes.** *Structure (Camb)* 2003, **11:**445-457.
12. Venclovas C, Timinskas A, Siksnys V: **Five-stranded beta-sheet sandwiched with two alpha-helices: a structural link between restriction endonucleases EcoRI and EcoRV.** *Proteins* 1994, **20:**279-282.
13. Pingoud A: **Restriction endonucleases.** In *Nucleic Acids and Molecular Biology Volume 14*. Edited by: Gross HJ. Berlin, Heidelberg, Springer-Verlag; 2004:442.
14. Kovall RA, Matthews BW: **Type II restriction endonucleases: structural, functional and evolutionary relationships.** *Curr Opin Chem Biol* 1999, **3:**578-583.
15. Horton JR, Blumenthal RM, Cheng X: **Restriction endonucleases: Structure of the conserved catalytic core and the role of metal ions in the DNA cleavage.** In *Restriction endonucleases Volume 14*. Edited by: Pingoud AM. Berlin, Springer-Verlag; 2004:361-392.
16. Skirgaila R, Grazulis S, Bozic D, Huber R, Siksnys V: **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* 1998, **279:**473-481.
17. Bujnicki JM, Rychlewski L: **Identification of a PD-(D/E)XK-like domain with a novel configuration of the endonuclease active site in the methyl-directed restriction enzyme Mrr and its homologs.** *Gene* 2001, **267:**183-191.
18. Pingoud V, Kubareva E, Stengel G, Friedhoff P, Bujnicki JM, Urbanke C, Sudina A, Pingoud A: **Evolutionary relationship between different subgroups of restriction endonucleases.** *J Biol Chem* 2002, **277:**14306-14314.
19. Tamulaitis G, Solonin AS, Siksnys V: **Alternative arrangements of catalytic residues at the active sites of restriction enzymes.** *FEBS Lett* 2002, **518:**17-22.
20. Newman M, Strzelecka T, Dorner LF, Schildkraut I, Aggarwal AK: **Structure of restriction endonuclease BamHI and its relationship to EcoRI.** *Nature* 1994, **368:**660-664.
21. Lukacs CM, Kucera R, Schildkraut I, Aggarwal AK: **Understanding the immutability of restriction enzymes: crystal structure of BglII and its DNA substrate at 1.5 A resolution.** *Nat Struct Biol* 2000, **7:**134-140.
22. Pingoud V, Sudina A, Geyer H, Bujnicki JM, Lurz R, Luder G, Morgan R, Kubareva E, Pingoud A: **Specificity changes in the evolution of Type II restriction endonucleases: a biochemical and bioinformatic analysis of restriction enzymes that recognize unrelated sequences.** *J Biol Chem* 2004.
23. Bujnicki JM, Rychlewski L: **Unusual evolutionary history of the tRNA splicing endonuclease EndA: relationship to the LAGLIDADG and PD-(D/E)XK deoxyribonucleases.** *Protein Sci* 2001, **10:**656-660.
24. Li H, Trotta CR, Abelson J: **Crystal structure and evolution of a transfer RNA splicing enzyme.** *Science* 1998, **280:**279-284.
25. Todone F, Weinzierl RO, Brick P, Onesti S: **Crystal structure of RPB5, a universal eukaryotic RNA polymerase subunit and transcription factor interaction target.** *Proc Natl Acad Sci U S A* 2000, **97:**6306-6310.
26. Daiyasu H, Komori K, Sakae S, Ishino Y, Toh H: **Hjc resolvase is a distantly related member of the type II restriction endonuclease family.** *Nucleic Acids Res* 2000, **28:**4540-4543.
27. Kvaratskhelia M, Wardleworth BN, Norman DG, White MF: **A conserved nuclease domain in the archaeal Holliday junction resolving enzyme Hjc.** *J Biol Chem* 2000, **275:**25540-25546.
28. 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:**3417-3432.
29. Bujnicki JM, Rychlewski L: **Grouping together highly diverged PD-(D/E)XK nucleases and identification of novel superfamily members using structure-guided alignment of sequence profiles.** *J Mol Microbiol Biotechnol* 2001, **3:**69-72.
30. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4:**41.
31. 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:**327-337.
32. Godzik A: **Fold recognition methods.** *Methods Biochem Anal* 2003, **44:**525-546.
33. Kurowski MA, Bujnicki JM: **GeneSilico protein structure prediction meta-server.** *Nucleic Acids Res* 2003, **31:**3305-3307.
34. Kosinski J, Cymerman IA, Feder M, Kurowski MA, Sasin JM, Bujnicki JM: **A "FRankenstein's monster" approach to comparative modeling: merging the finest fragments of Fold-Recognition models and iterative model refinement aided by 3D structure evaluation.** *Proteins* 2003, **53 Suppl 6:**369-379.
35. Lundstrom J, Rychlewski L, Bujnicki JM, Elofsson A: **Pcons: a neural-network-based consensus predictor that improves fold recognition.** *Protein Sci* 2001, **10:**2354-2362.
36. Fuxreiter M, Simon I: **Protein stability indicates divergent evolution of PD-(D/E)XK type II restriction endonucleases.** *Protein Sci* 2002, **11:**1978-1983.
37. Simons KT, Kooperberg C, Huang E, Baker D: **Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions.** *J Mol Biol* 1997, **268:**209-225.
38. Pingoud V, Conzelmann C, Kinzebach S, Sudina A, Metelev V, Kubareva E, Bujnicki JM, Lurz R, Luder G, Xu SY, Pingoud A: **PspGI, a type II restriction endonuclease from the extreme thermophile Pyrococcus sp.: structural and functional studies to investigate an evolutionary relationship with several mesophilic restriction enzymes.** *J Mol Biol* 2003, **329:**913-929.
39. Nakamura Y, Kaneko T, Sato S, Mimuro M, Miyashita H, Tsuchiya T, Sasamoto S, Watanabe A, Kawashima K, Kishida Y, Kiyokawa C, Kohara M, Matsumoto M, Matsuno A, Nakazaki N, Shimpo S, Takeuchi C, Yamada M, Tabata S: **Complete genome structure of Gloeobacter violaceus PCC 7421, a cyanobacterium that lacks thylakoids.** *DNA Res* 2003, **10:**137-145.
40. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO: **Environmental genome shotgun sequencing of the Sargasso Sea.** *Science* 2004, **304:**66-74.

41. Lambowitz AM, Belfort M: **Introns as mobile genetic elements.** *Annu Rev Biochem* 1993, **62:**587-622.
42. Mahillon J, Leonard C, Chandler M: **IS elements as constituents of bacterial genomes.** *Res Microbiol* 1999, **150:**675-687.
43. Sadykov M, Asami Y, Niki H, Handa N, Itaya M, Tanokura M, Kobayashi I: **Multiplication of a restriction-modification gene complex.** *Mol Microbiol* 2003, **48:**417-427.
44. Gimble FS: **Invasion of a multitude of genetic niches by mobile endonuclease genes.** *FEMS Microbiol Lett* 2000, **185:**99-107.
45. Chinen A, Uchiyama I, Kobayashi I: **Comparison between Pyrococcus horikoshii and Pyrococcus abyssi genome sequences reveals linkage of restriction-modification genes with large genome polymorphisms.** *Gene* 2000, **259:**109-121.
46. Kobayashi I: **Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution.** *Nucleic Acids Res* 2001, **29:**3742-3756.
47. Handa N, Nakayama Y, Sadykov M, Kobayashi I: **Experimental genome evolution: large-scale genome rearrangements associated with resistance to replacement of a chromosomal restriction-modification gene complex.** *Mol Microbiol* 2001, **40:**932-940.
48. Jurica MS, Stoddard BL: **Homing endonucleases: structure, function and evolution.** *Cell Mol Life Sci* 1999, **55:**1304-1326.
49. Mahdi AA, Sharples GJ, Mandal TN, Lloyd RG: **Holliday junction resolvases encoded by homologous rusA genes in Escherichia coli K-12 and phage 82.** *J Mol Biol* 1996, **257:**561-573.
50. Lyra C, Halme T, Torsti AM, Tenkanen T, Sivonen K: **Site-specific restriction endonucleases in cyanobacteria.** *J Appl Microbiol* 2000, **89:**979-991.
51. Matveyev AV, Young KT, Meng A, Elhai J: **DNA methyltransferases of the cyanobacterium Anabaena PCC 7120.** *Nucleic Acids Res* 2001, **29:**1491-1506.
52. Muro-Pastor AM, Flores E, Herrero A, Wolk CP: **Identification, genetic analysis and characterization of a sugar-non-specific nuclease from the cyanobacterium Anabaena sp. PCC 7120.** *Mol Microbiol* 1992, **6:**3021-3030.
53. Takahashi I, Hayano D, Asayama M, Masahiro F, Watahiki M, Shirai M: **Restriction barrier composed of an extracellular nuclease and restriction endonuclease in the unicellular cyanobacterium Microcystis sp.** *FEMS Microbiol Lett* 1996, **145:**107-111.
54. Jeltsch A: **Maintenance of species identity and controlling speciation of bacteria: a new function for restriction/modification systems?** *Gene* 2003, **317:**13-16.
55. Carr NG, Mann NH: **The oceanic cyanobacterial picoplankton.** In *The molecular biology of Cyanobacteria* Edited by: Bryant DA. Dordrecht, Kluwer Academic Publishers; 1994.
56. 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.** *NucleicAcidsRes* 1997, **25:**3389-3402.
57. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32:**1792-1797.
58. Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices from protein sequences.** *Comput Appl Biosci* 1992, **8:**275-282.
59. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4:**406-425.
60. Chapman B, Chang J: **Biopython: python tools for computational biology.** *ACM SIGBIO Newslett* 2000, **20:**15-19.
61. Marchler-Bauer A, Anderson JB, DeWeese-Scott C, Fedorova ND, Geer LY, He S, Hurwitz DI, Jackson JD, Jacobs AR, Lanczycki CJ, Liebert CA, Liu C, Madej T, Marchler GH, Mazumder R, Nikolskaya AN, Panchenko AR, Rao BS, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Vasudevan S, Wang Y, Yamashita RA, Yin JJ, Bryant SH: **CDD: a curated Entrez database of conserved domain alignments.** *Nucleic Acids Res* 2003, **31:**383-387.
62. McGuffin LJ, Bryson K, Jones DT: **The PSIPRED protein structure prediction server.** *Bioinformatics* 2000, **16:**404-405.
63. Rost B, Yachdav G, Liu J: **The PredictProtein server.** *Nucleic Acids Res* 2004, **32:**W321-6.
64. Ouali M, King RD: **Cascaded multiple classifiers for secondary structure prediction.** *Protein Sci* 2000, **9:**1162-1176.
65. Adamczak R, Porollo A, Meller J: **Accurate prediction of solvent accessibility using neural networks-based regression.** *Proteins* 2004, **56:**753-767.
66. Cuff JA, Barton GJ: **Application of multiple sequence alignment profiles to improve protein secondary structure prediction.** *Proteins* 2000, **40:**502-511.
67. Meiler J, Baker D: **Coupled prediction of protein secondary and tertiary structure.** *Proc Natl Acad Sci U S A* 2003, **100:**12105-12110.
68. Karplus K, Karchin R, Draper J, Casper J, Mandel-Gutfreund Y, Diekhans M, Hughey R: **Combining local-structure, fold-recognition, and new fold methods for protein structure prediction.** *Proteins* 2003, **53 Suppl 6:**491-496.
69. Rychlewski L, Jaroszewski L, Li W, Godzik A: **Comparison of sequence profiles. Strategies for structural predictions using sequence information.** *Protein Sci* 2000, **9:**232-241.
70. Kelley LA, MacCallum RM, Sternberg MJ: **Enhanced genome annotation using structural profiles in the program 3D-PSSM.** *J Mol Biol* 2000, **299:**499-520.
71. Fischer D: **Hybrid fold recognition: combining sequence derived properties with evolutionary information.** *Pacific Symp Biocomp* 2000:119-130.
72. Shi J, Blundell TL, Mizuguchi K: **FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties.** *J Mol Biol* 2001, **310:**243-257.
73. Jones DT: **GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences.** *J Mol Biol* 1999, **287:**797-815.
74. Zhou H, Zhou Y: **Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition.** *Proteins* 2004, **55:**1005-1013.
75. Luthy R, Bowie JU, Eisenberg D: **Assessment of protein models with three-dimensional profiles.** *Nature* 1992, **356:**83-85.
76. Sasin JM, Bujnicki JM: **COLORADO3D, a web server for the visual analysis of protein structures.** *Nucleic Acids Res* 2004, **32:**W586-9.
77. Holm L, Sander C: **Protein structure comparison by alignment of distance matrices.** *J Mol Biol* 1993, **233:**123-138.
78. Deibert M, Grazulis S, Sasnauskas G, Siksnys V, Huber R: **Structure of the tetrameric restriction endonuclease NgoMIV in complex with cleaved DNA.** *Nat Struct Biol* 2000, **7:**792-799.