A model of restriction endonuclease MvaI in complex with DNA: A template for interpretation of experimental data and a guide for specificity engineering

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · JULY 2007

Impact Factor: 2.63 · DOI: 10.1002/prot.21460 · Source: PubMed

Jan Kosinski
European Molecular Biology Laboratory
27 PUBLICATIONS **629** CITATIONS

Elena Aleksandrovna Kubareva
Lomonosov Moscow State University
135 PUBLICATIONS **884** CITATIONS

A Model of Restriction Endonuclease MvaI in Complex With DNA: A Template for Interpretation of Experimental Data and a Guide for Specificity Engineering

Jan Kosinski,[1*] **Elena Kubareva,**[2] **and Janusz M. Bujnicki**[1]

[1]*Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, Trojdena 4, 02-109 Warsaw, Poland*

[2]*A.N. Belozersky Institute of Physico-Chemical Biology, M.V. Lomonossov Moscow State University, Leninskie gory 1-40, Moscow 119992, Russia*

ABSTRACT **R.MvaI is a Type II restriction enzyme (REase), which specifically recognizes the pentanucleotide DNA sequence 5′-CCWGG-3′ (W indicates A or T). It belongs to a family of enzymes, which recognize related sequences, including 5′-CCSGG-3′ (S indicates G or C) in the case of R.BcnI, or 5′-CCNGG-3′ (where N indicates any nucleoside) in the case of R.ScrFI. REases from this family hydrolyze the phosphodiester bond in the DNA between the 2nd and 3rd base in both strands, thereby generating a double strand break with 5′-protruding single nucleotides. So far, no crystal structures of REases with similar cleavage patterns have been solved. Characterization of sequence-structure-function relationships in this family would facilitate understanding of evolution of sequence specificity among REases and could aid in engineering of enzymes with new specificities. However, sequences of R.MvaI or its homologs show no significant similarity to any proteins with known structures, thus precluding straightforward comparative modeling. We used a fold recognition approach to identify a remote relationship between R.MvaI and the structure of DNA repair enzyme MutH, which belongs to the PD-(D/E)XK superfamily together with many other REases. We constructed a homology model of R.MvaI and used it to predict functionally important amino acid residues and the mode of interaction with the DNA. In particular, we predict that only one active site of R.MvaI interacts with the DNA target at a time, and the cleavage of both strands (5′-CCAGG-3′ and 5′-CCTGG-3′) is achieved by two independent catalytic events. The model is in good agreement with the available experimental data and will serve as a template for further analyses of R.MvaI, R.BcnI, R.ScrFI and other related enzymes. Proteins 2007; 68:324–336.** **© 2007 Wiley-Liss, Inc.**

Key words: **sequence alignment; structure prediction; fold recognition; molecular evolution; extreme divergence; homology modeling; restriction endonuclease MvaI; GeneSilico; model validation**

INTRODUCTION

Classical Type II restriction endonucleases (REases) are homodimeric enzymes that recognize short DNA sequences (typically 4–8 bp long) and in the presence of Mg^{2+} cleave the target in both strands at, or in close proximity to the recognition site. Because of remarkably high specificity in recognizing and cleaving their target sequences, they are of high interest as model systems for analyzing protein-DNA interactions and one of the most frequently used tools for recombinant DNA technology. Type II REases that exhibit structural and functional deviations from the canon (e.g., require more than one target site for efficient cleavage, recognize asymmetrical targets, possess additional enzymatic domains or different subunits, require additional cofactors etc.) have been classified into subtypes (reviewed in Ref. 1). Recently, a new subtype has been characterized, which includes enzymes such as R.MspI[2,3] that possess only one active site, act essentially as monomers and cleave two strands of their target sites independently, in two consecutive reactions.

REases are the largest group of biochemically characterized enzymes. Regrettably, only a limited number of REase crystal structures have been solved and for the vast majority of these enzymes the molecular details of their action are completely unknown. So far it has not been possible to predict protein-DNA interactions and the mode of DNA cleavage from amino acid sequences of REases. In general, sequence comparison reveal no significant sequence similarities between enzymes that recognize different sequences or even between those that act

The Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

Grant sponsor: NIH (Fogarty International Center); Grant number: R03 TW007163-01; Grant sponsor: The DFG program "International Research Training Group; Grant number: GRK 1384.

*Correspondence to: Jan Kosinski, Laboratory of Bioinformatics and Protein Engineering, International Institute of Cell and Molecular Biology, ul. ks. Trojdena 4, 02-109 Warsaw, Poland.
E-mail: kosa@genesilico.pl

Received 15 December 2006; Revised 26 January 2007; Accepted 8 February 2007

Published online 3 April 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21460

on similar sequences but cleave them in a different way (reviewed in Ref. 4). A few exceptions from this rule include mostly remote relationships that were detected with advanced bioinformatics tools and corroborated with experimental analyses.[5–9] In the process of these analyses it has been realized that even enzymes that cleave identical sequences to produce the same patterns of ends may belong to different subfamilies and subtypes, for example R.MboI (class α, subtype IIP), and R.Sau3AI (class β, subtype IIE) both cleave 5′-GATC-3′ to produce 4 bp 5′ overhangs.[10] Moreover, it has been discovered that REases may comprise structurally and evolutionarily unrelated catalytic domains,[11–13] which adds yet another dimension to the difficulty of predicting sequence-function relationships.

To provide useful structural templates for functional analyses of REases, we have initiated a systematic bioinformatics analysis of enzymes with experimentally uncharacterized structures, in particular those, which appear to group together into relatively big subfamilies of sequences that are significantly similar to each other, but dissimilar to other proteins. One of such families of REases, for which so far no structural analyses have been reported and for which nothing is known about the residues important for DNA recognition and catalysis, includes enzymes R.MvaI,[14] R.BcnI,[15] and R.ScrFI.[16] They recognize DNA sequences with an odd number of base pairs: 5′-CC(W = A or T)GG-3′ (R.MvaI), 5′-CC(S = G or C)GG-3′ (R.BcnI), or 5′-CC(N = any base)GG-3′ (R.ScrFI), and cleave recognition site in both strands between the 2nd and 3rd nucleotide in both strands, thereby generating a double strand break with 5′-protruding single nucleotides. Since R.MvaI has been relatively best characterized on the biochemical level, especially with respect to the influence of chemical modifications in the DNA sequence on the cleavage activity,[17–23] we selected this enzyme as the representative enzyme for detailed analyses. Hereafter, we will refer to the group of REases analyzed in this work as 'R.MvaI family.'

MATERIALS AND METHODS

Sequence Analyses

Searches of homologs were carried out using BLAST and its iterative version PSI-BLAST[24] in the following databases: REBASE at New England Biolabs,[25] the nonredundant (nr) database of amino acid sequences, and the DNA sequences from unfinished genomics and metagenomics projects at the National Center for Biotechnology Information.[26] Multiple sequence alignment was calculated using MUSCLE[27] and optimized manually based on the results of structural analyses. Model of evolution (describing the probabilities of substitution of one amino acid by another) was selected using PROTTEST.[28] Phylogenetic inference was done using a modified version of PHYML,[29] which is distributed with PROTTEST. The phylogenetic tree was calculated using the Maximum Likelihood (ML) method. The best model describing the evolution of R.MvaI family according to PROTTEST was the WAG+G+F model (in short, WAG substitution matrix[30] combined with gamma model of rate heterogenity among sites (G) and amino acid frequencies (F) calculated from the R.MvaI family alignment). Bootstrap test with 1000 replications was performed to assess a confidence for each clade of the phylogenetic tree.

Structure Prediction

Protein structure prediction was carried out using a new version of the GeneSilico MetaServer,[31] which is a gateway for a variety of methods for making the predictions and analyzing their results. Sequences of R.MvaI, R.BcnI, R.ScrFI, and their uncharacterized homolog R.TvoORF1416RP were used as prediction targets in two versions: full length and without the apparent insertion in the middle of the catalytic domain. Secondary structure was predicted using a consensus of PSIPRED,[32] PROFsec,[33] PROF,[34] SABLE,[35] JNET,[36] JUFO,[37] PORTER,[38] SSPRO2,[39] and SAM-T02.[40] Solvent accessibility for the individual residues was predicted with SABLE,[35] ACCPRO2,[39] and JNET.[36] The fold-recognition (FR) analysis (attempt to match the query sequence to known protein structures) was carried out using PDB-BLAST (local implementation of a PSI-BLAST[24] search against sequences of proteins from the PDB), HHSEARCH,[41] FFAS03,[42] FORTE,[43] SAM-T02,[40] 3DPSSM,[44] INBGU,[45] FUGUE,[46] mGENTHREADER,[47] and SPARKS.[48] Target-template alignments reported by these methods were compared, evaluated, and ranked by the PCONS server[49] to identify the preferred modeling template and the consensus alignment.

The alignments between the sequence of R.MvaI and the structure of the best template identified by FR were used to carry out comparative modeling using the "FRankenstein's Monster" approach,[50,51] which comprises cycles of local realignments in uncertain regions, building of alternative models and their evaluation, realignment in poorly scored regions and merging of the best scoring fragments. This method was found as one of the most accurate approach for comparative modeling and FR in the rankings of CASP5 and CASP6.[52,53] We have also used this approach for successful prediction of the structure and accurate identification of protein-DNA contacts of R.SfiI,[54] later confirmed by the crystallographic study.[55]

For the evaluation of models we used PROQ,[56,57] and a MetaMQAP method recently developed in our group (URL: https://genesilico.pl/toolkit/unimod?method=MetaMQAPII, Marcin Pawlowski, Ryszard Matlak, J.M.B. manuscript in preparation), which allow predicting the deviation of individual residues in the model from their counterparts in the native structure.

The insertion comprising residues 76–153 was modeled '*de novo*,' because no reliable modeling template was found for this region. For this purpose we used ROSETTA, which attempts to generate native-like global conformations from 3 and 9 AA backbone fragments of experimentally solved protein structures.[58] Fragment selection is based on profile–profile, sequence, and secondary structure comparison between the target sequence (here: residues 76–153 of R.MvaI and the corresponding sequences of its homologs) and a database of fragments derived from the PDB database. ROSETTA is one of the best currently existing methods for '*de novo*' modeling[59] and is capable of adding structurally variable regions (insertions) to the core built with comparative modeling methods.[60] We carried out the fragment assembly with default parameters and medium level of sidechains rotamers optimization. Modeling was performed for R.MvaI and its homologs (R.BcnI, R.ScrFI, and R.TvoORF1416P) yielding 10,000–20,000 preliminary decoys for each homolog. 4000 decoys with the lowest energy were selected from each set and clustered based on structural similarity. The conformation most common to low-energy decoys from four representative members of the R.MvaI family was regarded as the one most likely resembling the native structure.

Modeling of the Protein–DNA Complex

Modeling of R.MvaI in complex with cognate DNA was performed with the assumption that R.MvaI binds DNA in a similar way as its preferred modeling template, MutH endonuclease. First, the model of R.MvaI was superimposed onto the structure of MutH bound to unmethylated DNA (2aoq[61]) by minimizing the root mean square deviation (RMSD) of Cα atoms in regions of protein-DNA contacts in the MutH-DNA complex. The DNA from the MutH structure was merged with the R.MvaI model and rebuilt to match the recognition site of R.MvaI (CCAGG) using the method implemented in 3DNA,[62] by 'mutating' the bases in the original sequence. A few steric clashes between amino acid sidechains and DNA were removed by choosing alternative rotamers from the Dunbrack library.[63] PyMol[64] and Swiss-PDBViewer[65] were used as graphical interfaces for inspection and manipulation of the structures, and to generate rendered figures.

RESULTS AND DISCUSSION

Sequence Analysis of R.MvaI and its Homologs

We carried out extensive searches of publicly available sequence databases (see Materials and Methods section) to identify as many members of the R.MvaI family as possible. The search revealed 19 sequences with a common domain of about 240 AA, which exhibits a pattern of conserved residues E-X_{18}-(P/A/G)D-X_4-E-X-K (Fig. 1). The motif strikingly resembles the (E)-X_N-(P)D-X_N-(D/E)-X-K motif (commonly referred to as the PD-(D/E)XK motif) present in many REases with known structure and in related nucleases with other cellular functions, for example DNA repair.[66,67] However, the distance between the conserved 'PD' and 'E-X-K' peptides is unusually short in the R.MvaI family.

We searched for potential homologs of the R.MvaI family using the FR approach, which allows identification of homologs among proteins with known structure even in the absence of evident sequence similarity (reviewed in Ref. 4; see Materials and Methods for details). Thus, full length sequences of R.MvaI, R.BcnI, R.ScrFI, and R.TvoORF1416P were submitted to the GeneSilico MetaServer[31] to identify structurally similar proteins that could serve as modeling templates. The FR analysis revealed that the closest relative of R.MvaI family among proteins with known structures is the MutH endonuclease involved in bacterial DNA mismatch repair, which exhibits the PD-(D/E)XK fold common to the majority of REases with known structure and to many other nucleases.[61,68] In the cell, MutH requires stimulation by MutS and MutL that detect mismatched nucleotides introduced by DNA polymerase in the newly replicated strand. MutH identifies the hemimethylated GATC sequence close to the mismatch site and cleaves the newly synthesized, and hence still unmethylated strand (i.e. introduces a single-stranded 'nick'), to prompt removal and resynthesis of the sequence including the site of the mismatch. Interestingly, MutH has been already found to be related to another family of REases, including R.Sau3AI,[5,69] which comprises a tandem duplication of two MutH-like domains.

The analysis of FR results in the context of multiple sequence alignment and predicted secondary structure (Figure 1) indicated that R.MvaI-related sequences comprise a MutH-like PD-(D/E)XK domain and a long insertion (AA 76–153 in MvaI) with predicted high content of β-strands. To optimize the alignment, we carried out the FR separately for the PD-(D/E)XK domain without the insertion. Among the FR servers used via the GeneSilico metaserver, 5 methods (FFAS03, HHSEARCH, FORTE, SAM-T02, and SPARKS) reported matches to the PD-(D/E)XK fold. All these servers selected structures of MutH endonucleases from *Escherichia coli* (1azo) and *Haemophilus influenzae* (2aoq, 2aor) as the best templates, albeit with medium scores (FFAS03: 1azo, score: −10.4; 2aoq, score: −9.61; HHSEARCH: 2aor, score: 60.66; 1azo, score: 45.31; FORTE: 1azo, score: 6.12; SAM-T02: 2aor, score: 1.2; 1azo, score: 2.2; SPARKS: 1azo, score: −1.68). 4 servers (PDBBLAST, mGENTHREADER, FROST, and 3DPSSM) did not report any PD-(D/E)XK nucleases among the 10 top ranked matches. Nonetheless, MutH was selected as the best template by the PCONS consensus server that does not make its own alignments, but selects the best alignment among all results proposed by the 'primary' FR servers. Similar results were obtained in the case of R.BcnI and R.TvoORF1416P. For R.ScrFI, only HHSEARCH reported a match to MutH, albeit with a very low score of 23.62. No other nuclease structures were reported by the FR servers with scores indicative of any significance. On the basis of our experience with detection of remote homology using FR methods, these results confidently identify the R.MvaI family as a member of the PD-(D/E)XK superfamily.

Eight uncharacterized members of the R.MvaI family contain N-terminal extensions of ~150–200 AA. On the basis of sequence similarity they can be divided into two groups. Fold recognition analysis revealed that N-termi-

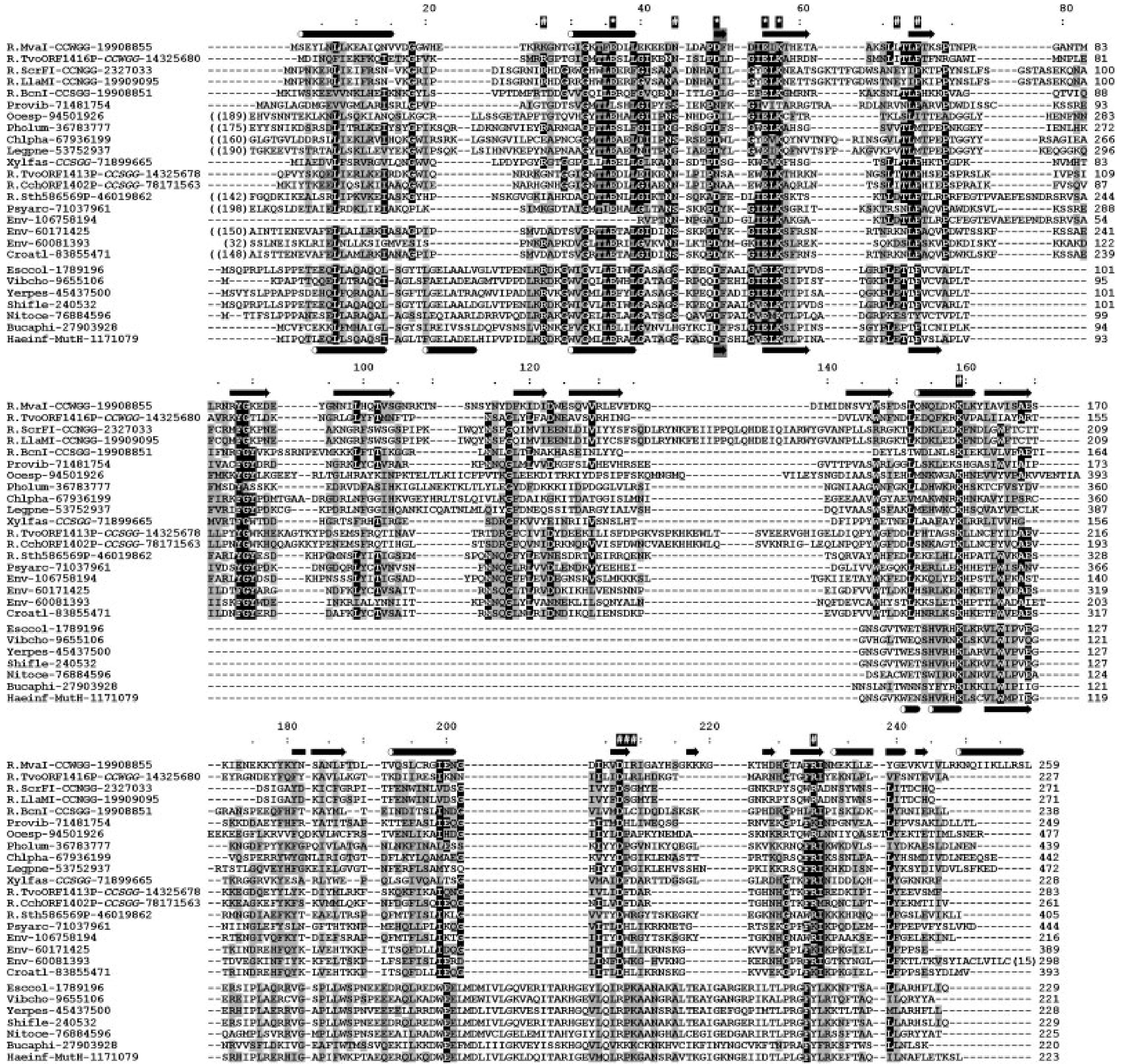

Fig. 1. Sequence alignment of R.MvaI and MutH and their homologs. Similar or identical amino acid residues are highlighted. Sequences present in REBASE are named according to nomenclature from REBASE (an organism name is omitted). Other sequences are named according to organism names (shortened to contain three letters of genus name and three letters of species name). Environmental sequences are named with "Env." For experimentally characterized enzymes known DNA sequence specificities are indicated in sequence names. Predicted DNA sequence specificities are shown in italic. All names contain NCBI Entrez GI numbers. Numbering in the top ruler corresponds to residue numbering of R.MvaI. Each fifth residue is indicated by a dot and each 10th one is indicated by a colon. Numbers in brackets in the N- and C-terminal part of the alignment indicate how many aa have been omitted for the sake of the clarity. Amino acid residues which are predicted to form catalytic site are indicated above the alignment by "*". Residues predicted to be important for DNA-binding are indicated by "#". Secondary structure (derived from crystal structure of MutH (2aoq) and the model of R.MvaI using DSSP program) is shown below the homologs of the target and the template. β-strands are shows as arrows and helices are shown as cylinders. Structurally degenerated (and not detected by DSSP program) β-strand harboring the 'PD' half-motif is shown on a grey background.

nal extensions of the group represented by an uncharacterized protein CA2559_00250 from *Croceibacter atlanticus* exhibit weak similarity to the ribosomal protein S8. Despite quite good match of predicted secondary structure elements (data not shown), the sequence similarity between those sequences and protein S8 sequence is rather low (10–20% of sequence identity). The scores of the matches reported by FR servers are also low (data not shown), which precludes straightforward answer about their relationship. However, the N-terminal extensions of

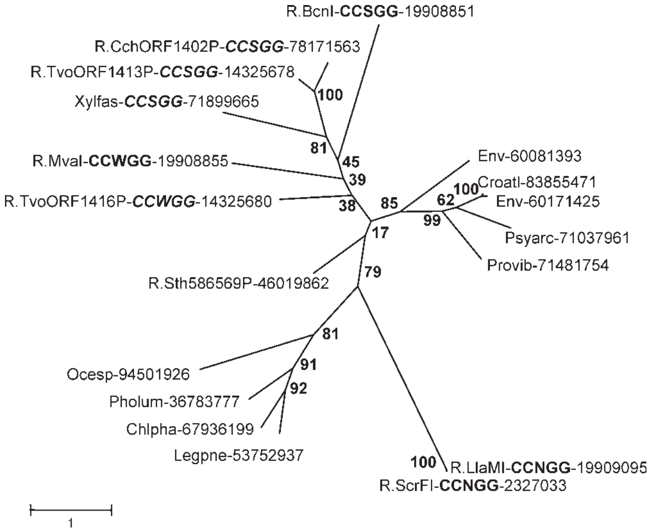

Fig. 2. Unrooted Maximum Likelihood phylogenetic tree of R.MvaI family, based on the multiple sequence alignment of conserved PD-(D/E)XK domains. Values at the nodes indicate the statistical support for the particular clades calculated according to bootstrap test. DNA sequence specificities of experimentally characterized enzymes are indicated in sequence names and shown in regular bold font. Predicted DNA sequence specificities are shown in italic bold.

proteins from the second group represented by an uncharacterized protein RED65_02328 from Oceanobacter sp. *RED65*, exhibit a different pattern of secondary structures and we were unable to detect their similarity to any known domains.

Phylogenetic analysis of the R.MvaI family based on the multiple sequence alignment of PD-(D/E)XK domains indicates that sequence specificity of these enzymes correlates with their evolutionary relationship (Fig. 2). However, the cleavage specificities of only four members (R.MvaI, R.BcnI, R.ScrFI, and R.LlaMI) have been confirmed experimentally. According to REBASE, recognition sequences of three other members (R.TvoORF1413P, R.CchORF1402P, and R.TvoORF1416P) are only putative. Nevertheless, the predictions available in the database are substantiated by the fact that they group with members of known specificities on the phylogenetic tree. Ten sequences (in particular two aforementioned groups of sequences with large N-terminal extensions) form too distant branches on the tree to confidently predict their specificities. Thus, experimental characterization of more R.MvaI family members is required to trace the evolution of DNA sequence specificity in this family, and for example infer the specificity of the ancestor.

Modeling of R.MvaI

The identification of relationship between the R.MvaI family and MutH provides the opportunity to build its molecular model by template-based methods. However, sequence alignments between R.MvaI and MutH proposed by FR servers differed substantially. The only region of consensus between different methods was in the vicinity of the catalytic core (AA 25–57 and 121–174 of R.MvaI). Moreover, most of alignments were incomplete in the C-terminus and some of them in N-terminus. In a search for other modeling templates we looked for other nucleases with known structures. Among known structures of PD-(D/E)XK enzymes, the closest homolog of MutH is R.PvuII REase,[70] which contains an additional β-strand at the very C-terminus. According to secondary structure predictions, R.MvaI also contains a C-terminal β-strand with no counterpart in the MutH template, which suggests that R.PvuII structure may be a useful template for modeling the C-terminal part of R.MvaI. Unfortunately, we could not identify such additional template for modeling the long insertion in R.MvaI (AA. 76–153), which was nevertheless predicted to be structured and composed of 5 or 6 β-strands.

We used the “FRankenstein’s monster” approach[50,51] to build a homology model of R.MvaI using the coordinates of MutH and PvuII as the templates, and at the same time refine the target-template alignment by validation of the sequence-structure fit at the three-dimensional level (see Materials and Methods for details). The structure of MutH from *H. influenzae* bound to unmethylated DNA substrate (PDB code: 2aoq) and the structure of PvuII D34G mutant bound to its cognate DNA (PDB code: 3pvi) substrate were used as modeling templates. The aforementioned insertion comprising residues 76–153 of R.MvaI could not be modeled using standard tools and represented significant obstacle for correct modeling of the remaining ‘well-behaved’ part of the protein. Therefore, it was initially omitted from the analysis and inserted only at the later stage. The preliminary model without the insertion was constructed by iterating the homology modeling procedure (initially based on the raw FR alignments), evaluation of the sequence-structure fit by MetaMQAP, 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 MutH structure used as the modeling template, and predicted for R.MvaI. This procedure was stopped when all regions in the protein core obtained acceptable MetaMQAP score (predicted deviation from the native structure <4 Å) or their score could not be improved by any manipulations. Almost all residues of the resulting model were modeled based on the MutH structure and only the last C-terminal β-strand was modeled based on the R.PvuII structure. The C-terminal helix of R.MvaI (AA 246–259) with no counterpart in MutH or R.PvuII structures, was added using ROSETTA. Because the target-template alignment in the N-terminal helix (AA 1–17) was uncertain according to FR analysis, this helix was also remodeled with ROSETTA. The resulting orientation of the helix is similar to that found in MutH and PvuII. Nonetheless, remodeling of the helix with ROSETTA improved the score of the model, according to MetaMQAP and PROQ. The model of R.MvaI without the insertion obtained a PROQ predicted LGscore 1.914 (which indicates ‘fairly good model’).

Having a refined template-based model of the R.MvaI PD-(D/E)XK domain, we inserted residues 76–153 as an unstructured loop and attempted to model it de novo with ROSETTA, while keeping the remaining part of R.MvaI intact. However, no well-folded model could be obtained in this way. Therefore, we modeled the structure of the insertion separately assuming that it forms independently folded domain. The modeling was performed for the insertion from R.MvaI as well as from several homologs: R.BcnI, R.ScrFI, and R.TvoORF1416P (see Materials and Methods section). The most representative model of the insertion, with the arrangement of β-strands observed commonly for all four sequences, was combined with the rest of the R.MvaI model and refined with Modeller. However, we must emphasize that the relative orientation of the domains in our model of R.MvaI is completely arbitrary. Unfortunately, the insertion scores very poorly according to PROQ (predicted LGscore −0.139) and the model of full-length MvaI exhibits lower score than the PD-(D/E)XK core alone (predicted LGscore 1.189). Thus, the 'insertion' part of the model must be regarded as unreliable. The problems with modeling of this region may however suggest a biologically important feature: we predict that the insertion is not well folded on its own, it may be involved in protein-protein or protein-DNA interactions (e.g., binding of the second subunit of the enzyme) that are not taken into account by the present model or that it undergoes conformational changes in the course of R.MvaI action.

Figure 1 shows the final alignment between R.MvaI and MutH obtained after several rounds of optimization, whereas Figure 3 shows the comparison of the template and the model structures. The sequence identity between R.MvaI and MutH over all aligned residues is 12.7%. As expected from comparative modeling, which assumes similarity to the template, the model R.MvaI shares many features of MutH and other PD-(D/E)XK nucleases, in particular the overall fold, comprising an antiparallel β-sheet in which the common active site is located (residues D50, E55, and K57 in R.MvaI), flanked on both sides by α-helices. R.MvaI, MutH, and R.PvuII belong to the β ('EcoRV-like') lineage[71,72] of PD-(D/E)XK nucleases that typically possess two features: an antiparallel β-hairpin (rather than a parallel βαβ motif) positioned C-terminally to the strand that harbors the 'E-X-K' half-motif in the common β-sheet, and an additional β-sheet used as a scaffold to mount sidechains involved in protein-DNA recognition.[72] Nonetheless, the model of R.MvaI reveals a number of important differences from typical 'PD-(D/E)XK' nucleases.

Although the residues of the putative active site are conserved between R.MvaI and MutH as well as other PD-(D/E)XK enzymes, their structural environment is different. The deletion of two amino acid residues between the 'PD' and 'E-X-K' peptides in R.MvaI results in a truncation of the region that forms a β-meander structure in PD-(D/E)XK enzymes. As a result, the sequence fragment including the 'PD' half-motif and the following residues is no longer able to form an extensive network of hydrogen bonds with the strand comprising the 'E-X-K' half-motif, leading to the structural degeneration of the β-hairpin that in other REases usually form a considerable extension of the common β-sheet.[73] This structural modification exposes the R.MvaI active site at the edge of the common β-sheet. In contrast to the degenerated common 'catalytic' β-sheet, the 'DNA-recognizing' β-sheet is expanded to form a wall of a very deep, positively charged cleft [Fig. 4(B)]. The expansion of this structural element has been previously observed also in MspI and HinP1I restriction enzymes, which form extensive contacts with the dsDNA target using only one domain (thus blocking the access of a second domain from the opposite side) and cleave each strand of the DNA separately.[3,74] It is interesting to note that in the R.MvaI family the number of residues on the exposed side of the additional sheet is conserved, even though several residues are variable [Figs. 1 and 4(A)]. This is in agreement with that REases from R.MvaI family recognize a partially conserved DNA sequence (5′-CCXGG-3′). Overall, these features suggests that R.MvaI binds DNA quite deeply, making contacts both to the minor and the major groove by a single subunit.

Modeling of the Protein-DNA Complex

REases from the R.MvaI family recognize similar sequences: 5′-CCWGG-3′ (R.MvaI), 5′-CCSGG-3′ (R.BcnI) or 5′-CCNGG-3′ (R.ScrFI). It would be interesting to determine residues responsible in these enzymes for the different specificity against the middle base pair. In the simplest case, residues contacting the outer CC and GG base pairs should be conserved in all R.MvaI homologs, while residues in contact with the central variable base pair should be unique to different classes of specificities in R.MvaI family. Contacts of R.MvaI with its recognition DNA sequence were studied quite extensively by analysis of cleavage rates of synthetic DNA substrates containing base analogs and altered phosphate backbone.[18,20,21,75–78] Based on these studies, a number of functional groups of DNA bases within the R.MvaI target sequence were proposed to be important for specific recognition. However, in the absence of a structure of R.MvaI in complex with DNA, residues responsible for specific recognition could not be predicted in the course of the previous analyses. Besides, experimental evidence suggests that R.MvaI may dimerize upon DNA binding,[79] but it is not known if one or two subunits are directly involved in sequence recognition and cleavage. To address these uncertainties we constructed a model of R.MvaI-DNA complex and confronted it with the available experimental data.

MutH is apparently the closest homolog of R.MvaI among proteins with known structure. Taking into account that MutH is monomeric in the functional complex with DNA, it seems plausible that R.MvaI may also binds to the DNA as a monomer. Indeed, a model of a dimeric R.MvaI-DNA complex, generated by duplicating the R.MvaI monomer and positioning the second copy symmetrically onto the DNA, contains severe clashes between the protein subunits (data not shown). These

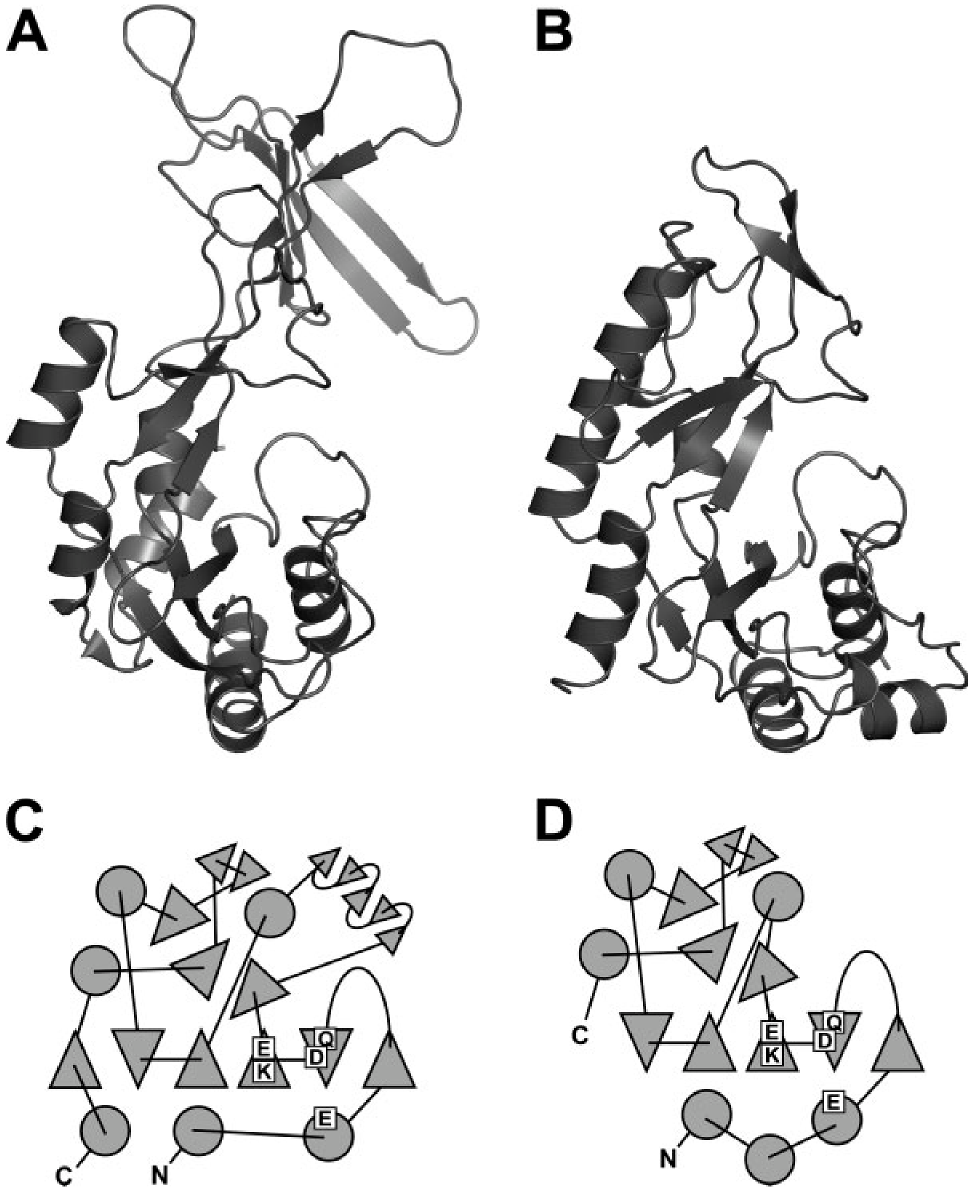

Fig. 3. Comparison of structures of the MutH template and the homology model of R.MvaI and their corresponding topology diagrams: (**A**) and (**C**) the final model of the R.MvaI based on coordinates of MutH and R.PvuII, (**B**) and (**D**) MutH (2aoq).

clashes are formed by a region that is modeled with relatively low confidence (AA 20–30 and 208–227 in R.MvaI) and which may adopt a different conformation than in the present model. Since dimerization of R.MvaI onto the DNA cannot be excluded based solely on the model structure, we confronted the model with the biochemical data concerning the interference of chemical modifications on protein-DNA interactions and DNA cleavage by R.MvaI.

Gromova and coworkers carried out a series of experiments on R.MvaI cleavage rates of synthetic DNA substrates containing base analogs and altered phosphate backbone.[20] They observed that a substrate with the A nucleoside substituted by a trimethylene bridge is still cleaved in the intact strand (82% of the cleavage rate of the unmodified DNA duplex) while cleavage is hindered in modified strand (cleavage rate close to 0). Indeed, in the monomeric model of R.MvaI-DNA complex, this modification would not interfere with R.MvaI binding and cleavage of the unmodified strand. Moreover, it would not allow for symmetrical binding of two R.MvaI subunits at the same time on both sides of the pentanucleotide recognition sequence. This suggests that the target sequence is bound by only one R.MvaI subunit, and one side of the DNA remains partially exposed to the solvent (or if R.MvaI is a dimer, then the second subunit would not be directly involved in recognition of the target sequence). Interestingly, a cleavage of both strands was blocked when the central T nucleoside was substituted by a tri-

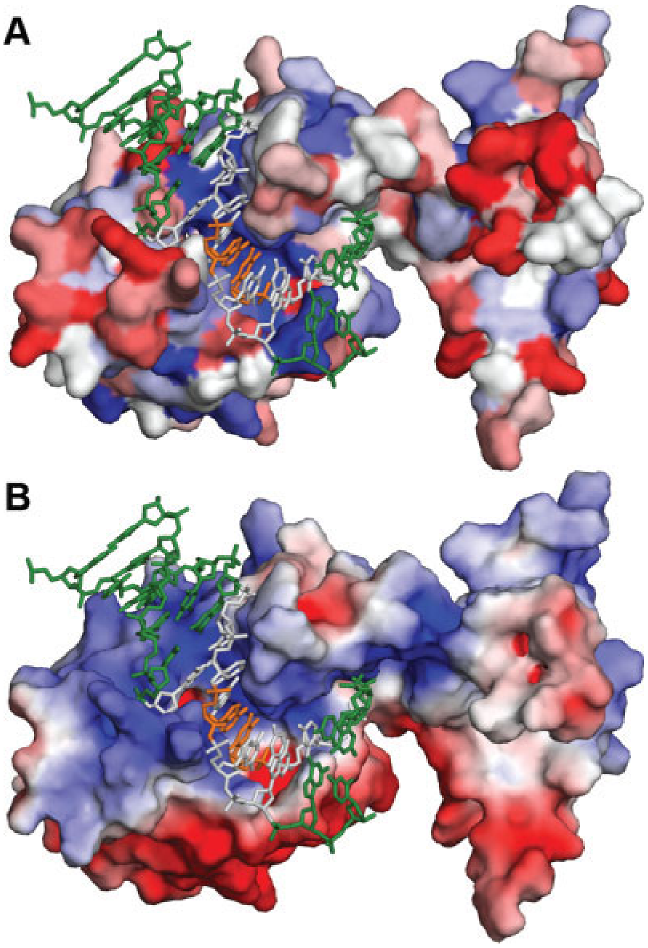

Fig. 4. DNA-binding mode of R.MvaI. The protein is shown in a molecular surface representation. DNA is shown as green with specifically recognized bases in light grey. The central base pair is shown in orange. (**A**) Sequence conservation mapped onto the molecular surface of the R.MvaI (conserved residues are colored blue, variable residues are colored red) (**B**) Electrostatic potential mapped onto the molecular surface of R.MvaI (positively and negatively charged regions are colored in blue and red, respectively).

methylene bridge. On the other hand, when a similar modification was introduced to replace the phosphodiester bond between C and T residues, the unmodified strand was still cleaved (65% of the original cleavage rate). This suggests that R.MvaI requires interactions with the central T base for efficient cleavage. Moreover, it has been shown that removing one of scissile phosphodiester bonds or substituting one of scissile phosphodiester bonds by a non-hydrolysable pyrophosphate or a bulky modification does not hinder the cleavage of the unmodified strand.[20] This is in agreement with our monomeric model of R.MvaI, in which the enzyme does not form direct contacts with the scissile phosphodiester bond in the complementary strand.

Experiments on cleavage rates of different strands of quasipalindromic substrate DNA[21] demonstrated that a strand containing "A" in the recognition site ("A-strand") is cleaved more efficiently by R.MvaI than a strand containing "T" ("T-strand"). After the A-strand of the dsDNA is hydrolyzed, the T-strand in the nicked duplex DNA is cleaved faster than in the continuous dsDNA. This suggests that R.MvaI is in fact a nicking nuclease, which cleaves dsDNA in two independent events of binding and cleavage.

Certain regions of our model that may take part in protein-DNA interactions (e.g. AA 20–30, 42–46, and 208–227) are predicted with moderate confidence and may have different conformations in the native structure of R.MvaI. Nevertheless, a number of probable protein-DNA contacts that are in agreement with the available experimental data can be predicted from the model (Fig. 5) and the analysis of conservation in the sequence alignment (Figure 1).

Contacts with DNA bases

According to our model, most of the contacts with DNA are formed by R.MvaI and the specifically recognized bases in the target sequence (5′-CCWGG-3′) [Fig. 5(A)]. The two DNA strands are recognized in different ways, thus we will refer to the strand containing the phosphodiester bond which is to be cleaved as the proximal strand, and to the opposite strand as the distal strand. For the purpose of this work we define that two groups are in contact if the shortest distance between atoms of the groups is lower than 4 Å. The most prominent contact with the external C—G pairs is formed between the invariant D207 and the external cytosine in the distal strand [Fig. 5(B)]. The carboxyl group of D207 could form a hydrogen bond with the exocyclic NH_2 of cytosine. However, Gromova and coworkers found that methylation of exocyclic NH_2 groups in either of the external cytosines in the recognition site blocks cleavage of the strand with that methylated base, whereas the unmethylated strand is cleaved as efficiently as in the completely unmethylated DNA.[20] This suggests that the contact of D207 with external cytosine from the distal strand is not important for a cleavage rate of the proximal strand. However, this contact can still be important for the specificity of DNA recognition by R.MvaI and related enzymes. According to our model of R.MvaI-DNA complex, the external cytosine in the proximal strand may be recognized by residues from the sequence region 208–227 [Fig. 5(A)]. Thus, methylation of NH_2 group of the external cytosine in the proximal strand could interfere with protein-DNA contacts and thereby prevent the cleavage. However, this protein region is modeled with low confidence and its exact conformation is too uncertain to predict direct contacts with the external cytosine from a proximal strand. Our model provides several explanations of the fact that N4-methylation of internal cytosine in both strands of the recognition site by the M.MvaI methyltransferase abolishes cleavage.[14] Probably, N4-methylation of internal cytosines generates steric clashes with R209 (which most likely recognizes the central base pair) and disturbs contacts with protein backbone of residues AA 208–211. N4-methylation of

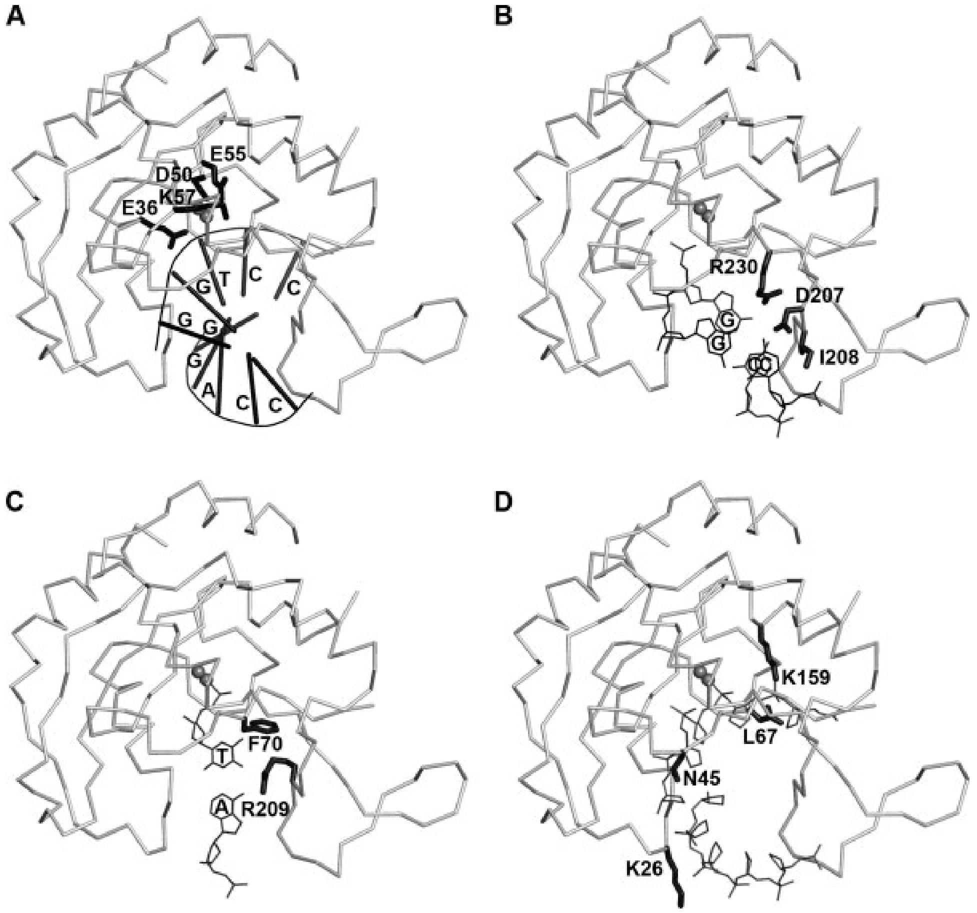

Fig. 5. Predicted functionally important residues of R.MvaI. For the sake of clarity, the insertion corresponding to AA 77–152 is not shown. The protein is shown as a grey Cα trace. The sidechains of important residues are shown as black thick sticks. Metal ions are shown as grey spheres. (**A**) Predicted catalytic residues. DNA is shown as grey sticks where one stick correspond to single nucleoside (only specifically recognized DNA residues are shown) (**B**) Protein-DNA contacts with guanosines of proximal strand and cytosines of distal strand (**C**) Protein-DNA contacts with the central AT base pair (**D**) Protein-DNA contacts with DNA backbone.

internal cytosine can also affect local conformation of DNA required for efficient binding and cleavage.

The other possible contact between R.MvaI and cytosines in target sequence is a hydrophobic interaction of I208 with positions 4 and 5 of either internal or external cytosine [Fig. 5(B)] in the distal strand. Indeed, according to Kubareva et al.,[18] methylation at the C5 position in either of the internal cytosines of R.MvaI recognition site slightly affects the cleavage of unmodified strand only (relative initial rate about 55%[18]). In our monomeric model C5-methylation of either cytosine in the proximal strand would not generate any steric clashes with the enzyme, while the C5-methylation of either cytosine in the distal strand would generate small clashes with I208.

Our model suggests that R230 in R.MvaI is involved in recognition of one of the G residues. In the model it makes a contact with the internal G in the proximal strand [Fig. 5(B)]; however a slight conformational rearrangement may allow it to reach to the external G in the same strand, thus we cannot predict any of these interactions with confidence. R230 is almost invariant in the R.MvaI family (exchanged to K in one homolog—a hypothetical protein CvibDRAFT_0836 from *Prosthecochloris vibrioformis*), which is in agreement with the conservation of the "GG" module in the recognition site of all experimentally characterized members. The contacts between R.MvaI and G bases are important in the light of the experiments demonstrating that substitution of one of G residue by an abasic site reduced the cleavage of both DNA strands (the cleavage of the unmodified strand is blocked and the cleavage rate of the modified strand is reduced to 9–19% compared to the unmodified DNA).[75]

Our model also suggests that the central base pair is recognized by R209 [Fig. 5(C)]. N6-methylation of the central adenine blocks the cleavage of the modified strand by R.MvaI.[20] On the other hand, the complementary thymine also seems to be important for the recognition, as R.MvaI does not cleave an unmodified strand when the T nucleoside is substituted by a trimethylene bridge, but it cleaves that strand when the trimethylene bridge is substituted for the phosphodiester bond between the C and T residues (albeit with a reduced rate, 65% of the rate for the unmodified DNA). Interestingly, R209 is replaced by G in R.ScrFI and R.LlaMI, which recognize the CCNGG sequence, and by C in R.BcnI, which recognizes the CCSGG sequence. This suggests that R209 could be responsible for restricting the specificity of R.MvaI to W in the CCWGG sequence. Site-directed mutagenesis of R209 and homologous residues in other enzymes will provide a direct test of our prediction, and if successful, could help to engineer enzymes with new specificities (e.g., restricted to nicking of asymmetric substrates with a single defined base in the central position).

Another contact is formed between the sidechain of F70 and the central base in the proximal strand [Fig. 5(C)]. Substituting dU for the central residue T in the R.MvaI recognition site (resulting in the removal of the exocyclic CH_3 group of thymine) does not affect the cleavage rate of neither strand, while substituting bromine for CH_3 group (thereby introducing a group with a covalent radius greater than that of CH_3) lowers the cleavage activity up to twofold.[20] Thus, Gromova et al. proposed that the methyl group of the central T may not form hydrophobic contacts with the protein and the introduction of Br may generate moderate steric clashes. F70 is almost invariant residue in the R.MvaI family (rarely substituted by M or I), as well as in the MutH family where it is postulated to be important for proper positioning of K186 in MutH, which recognizes G/C pair in GATC sequence.[61] In the R.MvaI family, F70 could be important for positioning of a sidechain that recognizes the central base pair (R209 in the case of R.MvaI). Bulky bromine substitution of the methyl group in the central T could 'push away' F70 and thereby disturb this interaction, while the removal of methyl group would not affect the position of F70.

Contacts with the DNA backbone

According to our model, R.MvaI makes multiple contacts with the backbone of the target DNA, mostly with the proximal strand. This is in agreement with observations that altering the conformation of scissile phospodiester bonds (and thus protein-DNA backbone contacts) completely prevents the cleavage of the modified strand, and only decreases the cleavage of the unmodified strand, but does not abolish it completely. It was reported that alterations of a conformation of the TpG phosphodiester bond (by introducing derivatives of the central T residue) reduce the cleavage rate of the unmodified strand.[18,20,78] Thus, it was suggested that to efficiently bind and cleave one DNA strand, R.MvaI has to make contacts with the phosphate opposite to the scissile phosphate. In our model such a contact is formed by K26, which is moderately conserved in the R.MvaI family (in other members it may be exchanged to other polar amino acids). All phosphates in the proximal strand form extensive contacts with the sidechains and protein backbone of R.MvaI [Fig. 5(D)]. Residues N45, L67, and K159 of R.MvaI, which take part in protein-DNA backbone contacts, are strongly conserved in the R.MvaI family-N45 and L67 are typically exchanged to amino acids with similar physicochemical properties, while K159 is almost invariant. The scissile phosphate forms contacts with the catalytic D50 and, through metal ion(s), with E36 and E55. Only a few contacts are found between R.MvaI and phosphates outside the R.MvaI recognition sequence.

CONCLUSIONS

Our FR analysis reveals that R.MvaI and its homologs are members of the PD-(D/E)XK superfamily of nucleases and that they are related to the DNA repair enzyme MutH. This discovery is very interesting from the evolutionary point of view, as it provides another example of a group of structurally similar REases with completely different sequence specificities: 5′-CC↓XGG-3′ in the R.MvaI family versus 5′-↓GATC-3′ in MutH ("↓" indicates the site of cleavage). Our results suggest that R.MvaI cleaves the DNA as a monomeric enzyme, that is only one active site of R.MvaI interacts with the DNA target at a time, and the cleavage of both strands (5′-CCAGG-3′ and 5′-CCTGG-3′) is achieved by two independent catalytic events. This mode of action is intermediate between that of MutH, a monomeric nicking enzyme that acts only on one strand of its target, and that of typical Type II REases, which are homodimers that simultaneously cleave both strands. R.Sau3AI is a representative of another group of MutH-like nucleases, and it exhibits yet another mode of action, namely dimerization upon binding of two copies of the DNA sequence, of which one is cleaved in both strands, and the other is not cleaved at all. At this point it is difficult to infer the mechanism of action of the ancestor of these nucleases, although it is tempting to speculate that it acted as a monomer.

The molecular model described in this article provides a structural template for interpretation of a large body of experimental data on R.MvaI-DNA interactions that so far could not be analyzed on the molecular level at the side of the protein. Thus, combination of protein structure prediction and experimental analyses allows for characterization of sequence-structure-function relationships in the R.MvaI family. Given the remote homology of R.MvaI and MutH the atomic details of protein-DNA contacts present in the model should be taken with caution. Nevertheless, our model strongly suggests that the protein-DNA interface is asymmetric and the majority of contacts are with the proximal strand. Moreover, based on the model we can predict that D207 and R230 are responsible for the recognition of the external C—G pairs in the CCWGG, CCSGG, or CCNGG sequences, which is consistent with the observation that these residues are almost invariant in the R.MvaI family. On the other hand, R209 is predicted to be responsible for the recognition of the central base pair (A—T or T—A in R.MvaI).

In addition to explaining the data from previous analyses, the model and the comparative sequence analysis presented in this work suggest new experiments to be carried out for R.MvaI, R.BcnI, R.ScrFI, and other related enzymes. In particular, substitution of R209 (and perhaps also its structural neighbors) may result in the alteration of sequence specificity towards the central base pair. It will be also interesting to study the role of the big insertion comprising residues 76–153, which we were not able to model confidently, that is deletion mutagenesis could help to validate the prediction that this region may be required for DNA binding. Besides, our model suggests that the kinetic mechanism of dsDNA cleavage by R.MvaI should be reexamined, as it may reveal a distinct model of action than in the case of orthodox, homodimeric REases. Finally, our phylogenetic analysis identifies such functionally uncharacterized members of the R.MvaI family, which are evolutionarily most distant to the characterized members. Experimental analysis of these 'putative' REases will help to correlate the pattern of amino acid residues at the protein-DNA interface with the substrate specificity, at the level of the whole R.MvaI family. Some of these enzymes may exhibit new specificities, and could immediately become commercially valuable reagents.

ACKNOWLEDGMENTS

J.K. is a recipient of a scholarship from the Postgraduate School of Molecular Medicine at the Medical University of Warsaw.

NOTE ADDED IN PROOF

After the acceptance of the final version of this manuscript, the crystal structure of R.MvaI was published (ref. 80). Although the overall similarity between the model and the structure is moderate (RMSD 3.05 Å over 99 superimposable Cα atoms), our model correctly predicts all the key features of MvaI: its monomeric structure, the PD-(D/E)XK fold, relationship to MutH, all catalytic residues, several DNA contacting residues including D207, R230 and R209 and that R209 residue is involved in the recognition of the central base pair. Some details of protein-DNA recognitions have not been predicted correctly. In the crystal structure the inserted subdomain forms an extension of the DNA-recognition lobe, in agreement with our prediction that it does not fold on its own and may be involved in protein-DNA and protein-protein interactions.

REFERENCES

1. Roberts RJ, Belfort M, Bestor T, Bhagwat AS, Bickle TA, Bitinaite J, Blumenthal RM, Degtyarev S, Dryden DT, Dybvig K, Firman K, Gromova ES, Gumport RI, Halford SE, Hattman S, Heitman J, Hornby DP, Janulaitis A, Jeltsch A, Josephsen J, Kiss A, Klaenhammer TR, Kobayashi I, Kong H, Kruger DH, Lacks S, Marinus MG, Miyahara M, Morgan RD, Murray NE, Nagaraja V, Piekarowicz A, Pingoud A, Raleigh E, Rao DN, Reich N, Repin VE, Selker EU, Shaw PC, Stein DC, Stoddard BL, Szybalski W, Trautner TA, Van Etten JL, Vitor JM, Wilson GG, Xu SY. A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. Nucleic Acids Res 2003;31:1805–1812.
2. Xu QS, Kucera RB, Roberts RJ, Guo HC. An asymmetric complex of restriction endonuclease MspI on its palindromic DNA recognition site. Structure 2004;12:1741–1747.
3. Xu QS, Roberts RJ, Guo HC. Two crystal forms of the restriction enzyme MspI-DNA complex show the same novel structure. Protein Sci 2005;14:2590–2600.
4. 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.
5. Friedhoff P, Lurz R, Luder G, Pingoud A. Sau3AI, a monomeric type II restriction endonuclease that dimerizes on the DNA and thereby induces DNA loops. J Biol Chem 2001;276:23581–23588.
6. 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.
7. Pawlak SD, Radlinska M, Chmiel AA, Bujnicki JM, Skowronek KJ. Inference of relationships in the 'twilight zone' of homology using a combination of bioinformatics and site-directed mutagenesis: a case study of restriction endonucleases Bsp6I and PvuII. Nucleic Acids Res 2005;33:661–671.
8. Chmiel AA, Radlinska M, Pawlak SD, Krowarsch D, Bujnicki JM, Skowronek KJ. A theoretical model of restriction endonuclease NlaIV in complex with DNA, predicted by fold recognition and validated by site-directed mutagenesis and circular dichroism spectroscopy. Protein Eng Des Sel 2005;18:181–189.
9. Skowronek KJ, Kosinski J, Bujnicki JM. Theoretical model of restriction endonuclease HpaI in complex with DNA, predicted by fold recognition and validated by site-directed mutagenesis. Proteins 2006;63:1059–1068.
10. 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 2005;280:4289–4298.
11. Sapranauskas R, Sasnauskas G, Lagunavicius A, Vilkaitis G, Lubys A, Siksnys V. Novel subtype of type IIs restriction enzymes. BfiI endonuclease exhibits similarities to the EDTA-resistant nuclease Nuc of *Salmonella typhimurium*. J Biol Chem 2000;275:30878–30885.
12. Bujnicki JM, Radlinska M, Rychlewski L. Polyphyletic evolution of type II restriction enzymes revisited: two independent sources of second-hand folds revealed. Trends Biochem Sci 2001;26:9–11.
13. Grazulis S, Manakova E, Roessle M, Bochtler M, Tamulaitiene G, Huber R, Siksnys V. Structure of the metal-independent restriction enzyme BfiI reveals fusion of a specific DNA-binding domain with a nonspecific nuclease. Proc Natl Acad Sci USA 2005;102:15797–15802.
14. Butkus V, Klimasauskas S, Kersulyte D, Vaitkevicius D, Lebionka A, Janulaitis A. Investigation of restriction-modification enzymes from M. varians RFL19 with a new type of specificity toward modification of substrate. Nucleic Acids Res 1985;13: 5727–5746.
15. Janulaitis AA, Petrusite MA, Jaskelavicene BP, Krayev AS, Skryabin KG, Bayev AA. A new restriction endonuclease BcnI from Bacillus centrosporus RFL 1. FEBS Lett 1982;137:178–180.
16. Fitzgerald GF, Daly C, Brown LR, Gingeras TR. ScrFI: a new sequence-specific endonuclease from *Streptococcus cremoris*. Nucleic Acids Res 1982;10:8171–8179.
17. Pein CD, Cech D, Gromova ES, Orezkaya TS, Shabarova ZA, Kubareva EA. Interaction of the MvaI restriction enzyme with synthetic DNA fragments. Nucleic Acids Symp Ser 1987:225–228.
18. Kubareva EA, Pein CD, Gromova ES, Kuznezova SA, Tashlitzki VN, Cech D, Shabarova ZA. The role of modifications in oligonucleotides in sequence recognition by MvaI restriction endonuclease. Eur J Biochem 1988;175:615–618.
19. Kubareva EA, Gromova ES, Romanova EA, Oretskaia TS, Shabarova ZA. Cleavage by restriction endonucleases MvaI and EcoRII of substrates modified in amino groups of heterocyclic bases. Bioorg Khim 1990;16:501–506.
20. Gromova ES, Kubareva EA, Vinogradova MN, Oretskaya TS, Shabarova ZA. Peculiarities of recognition of CCA/TGG sequences in DNA by restriction endonucleases MvaI and EcoRII. J Mol Recognit 1991;4:133–141.
21. Kubareva EA, Gromova ES, Pein CD, Krug A, Oretskaya TS, Cech D, Shabarova ZA. Oligonucleotide cleavage by restriction endonucleases MvaI and EcoRII: a comprehensive study on the influence of structural parameters on the enzyme-substrate interaction. Biochim Biophys Acta 1991;1088:395–400.

22. Sheflyan G, Kubareva EA, Volkov EM, Oretskaya TS, Gromova ES, Shabarova ZA. Chemical cross-linking of MvaI and EcoRII enzymes to DNA duplexes containing monosubstituted pyrophosphate internucleotide bond. Gene 1995;157:187–190.
23. Sheflyan GY, Kubareva EA, Gromova ES, Shabarova ZA. Conformational transition of restriction endonuclease MvaI-substrate complex under the influence of Mg2+ probed by DNA-protein cross-linking studies. Bioconjug Chem 1998;9:703–707.
24. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
25. Roberts RJ, Vincze T, Posfai J, Macelis D. REBASE–restriction enzymes and DNA methyltransferases. Nucleic Acids Res 2005;33: D230–D232 (Database Issue).
26. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Church DM, DiCuccio M, Edgar R, Federhen S, Helmberg W, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pontius JU, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Sirotkin K, Starchenko G, Suzek TO, Tatusov R, Tatusova TA, Wagner L, Yaschenko E. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2005;33:D39–D45 (Database issue).
27. Edgar RC. MUSCLE. Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;321792–1797.
28. Abascal F, Zardoya R, Posada D. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 2005;21:2104–2105.
29. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 2003;52:696–704.
30. Whelan S, Goldman N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 2001;18:691–699.
31. Kurowski MA, Bujnicki JM. GeneSilico protein structure prediction meta-server. Nucleic Acids Res 2003;31:3305–3307.
32. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
33. Rost B, Yachdav G, Liu J. The predict protein server. Nucleic Acids Res 2004;32:W321–W326 (Web Server Issue).
34. Ouali M, King RD. Cascaded multiple classifiers for secondary structure prediction. Protein Sci 2000;9:1162–1176.
35. Adamczak R, Porollo A, Meller J. Combining prediction of secondary structure and solvent accessibility in proteins. Proteins 2005;59:465–475.
36. Cuff JA, Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 2000;40:502–511.
37. Meiler J, Baker D. Coupled prediction of protein secondary and tertiary structure. Proc Natl Acad Sci USA 2003;100:12105–12110.
38. Pollastri G, McLysaght A. Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics 2005;21: 1719–1720.
39. Cheng J, Randall AZ, Sweredoski MJ, Baldi P. SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res 2005;33:W72–W76 (Web Server Issue).
40. 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.
41. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics 2005;21:951–960.
42. Jaroszewski L, Rychlewski L, Li Z, Li W, Godzik A. FFAS03: a server for profile–profile sequence alignments. Nucleic Acids Res 2005;33:W284–W288 (Web Server Issue).
43. Tomii K, Akiyama Y. FORTE: a profile-profile comparison tool for protein fold recognition. Bioinformatics 2004;20:594–595.
44. Kelley LA, MacCallum RM, Sternberg MJ. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 2000;299:499–520.
45. Fischer D. Hybrid fold recognition: combining sequence derived properties with evolutionary information. Pac Symp Biocomput 2000:119–130.
46. 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.
47. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287: 797–815.
48. 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.
49. 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.
50. 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.
51. Kosinski J, Gajda MJ, Cymerman IA, Kurowski MA, Pawlowski M, Boniecki M, Obarska A, Papaj G, Sroczynska-Obuchowicz P, Tkaczuk KL, Sniezynska P, Sasin JM, Augustyn A, Bujnicki JM, Feder M. FRankenstein becomes a cyborg: the automatic recombination and realignment of Fold-Recognition models in CASP6. Proteins 2005;61 (Suppl 7):106–113.
52. Tramontano A, Morea V. Assessment of homology-based predictions in CASP5. Proteins 2003;53 (Suppl 6):352–368.
53. Wang G, Jin Y, Dunbrack RL, Jr. Assessment of fold recognition predictions in CASP6. Proteins 2005;61 (Suppl 7):46–66.
54. Chmiel AA, Bujnicki JM, Skowronek KJ. A homology model of restriction endonuclease SfiI in complex with DNA. BMC Struct Biol 2005;5:2.
55. Vanamee ES, Viadiu H, Kucera R, Dorner L, Picone S, Schildkraut I, Aggarwal AK. A view of consecutive binding events from structures of tetrameric endonuclease SfiI bound to DNA. EMBO J 2005;24:4198–4208.
56. Wallner B, Elofsson A. Can correct protein models be identified? Protein Sci 2003;12:1073–1086.
57. Wallner B, Elofsson A. Identification of correct regions in protein models using structural, alignment, and consensus information. Protein Sci 2006;15:900–913.
58. 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.
59. Vincent JJ, Tai CH, Sathyanarayana BK, Lee B. Assessment of CASP6 predictions for new and nearly new fold targets. Proteins 2005;61 (Suppl 7):67–83.
60. Rohl CA, Strauss CE, Chivian D, Baker D. Modeling structurally variable regions in homologous proteins with ROSETTA. Proteins 2004;55:656–677.
61. Lee JY, Chang J, Joseph N, Ghirlando R, Rao DN, Yang W. MutH complexed with hemi- and unmethylated DNAs: coupling base recognition and DNA cleavage. Mol Cell 2005;20:155–166.
62. Lu XJ, Olson WK. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. Nucleic Acids Res 2003;31:5108–5121.
63. Bower MJ, Cohen FE, Dunbrack RL, Jr. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. J Mol Biol 1997;267:1268–1282.
64. DeLano WL. The PyMOL molecular graphics system. San Carlos, CA: DeLano Scientific; 2002.
65. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 1997;18:2714–2723.
66. 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.
67. Kosinski J, Feder M, Bujnicki JM. The PD-(D/E)XK superfamily revisited: identification of new members among proteins involved in DNA metabolism and functional predictions for domains of (hitherto) unknown function. BMC Bioinformatics 2005; 6:172.
68. 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.

69. Bujnicki JM. A model of structure and action of Sau3AI restriction endonuclease that comprises two MutH-like endonuclease domains within a single polypeptide. Acta Microbiol Pol 2001; 50:219–231.
70. Cheng X, Balendiran K, Schildkraut I, Anderson JE. Structure of PvuII endonuclease with cognate DNA. EMBO J 1994;13: 3927–3935.
71. Huai Q, Colandene JD, Chen Y, Luo F, Zhao Y, Topal MD, Ke H. Crystal structure of NaeI-an evolutionary bridge between DNA endonuclease and topoisomerase. EMBO J 2000;19:3110–3118.
72. Bujnicki JM. Understanding the evolution of restriction-modification systems: clues from sequence and structure comparisons. Acta Biochim Pol 2001;48:935–967.
73. Bujnicki JM. Molecular phylogenetics of restriction endonucleases. In: Pingoud A, editor. Restriction Endonucleases. Nucleic Acids and Molecular Biology, Vol. 14. Berlin: Springer-Verlag; 2004. pp 63–87.
74. Yang Z, Horton JR, Maunus R, Wilson GG, Roberts RJ, Cheng X. Structure of HinP1I endonuclease reveals a striking similarity to the monomeric restriction enzyme MspI. Nucleic Acids Res 2005;33:1892–1901.
75. Kubareva EA, Petrauskene OV, Karyagina AS, Tashlitsky VN, Nikolskaya II, Gromova ES. Cleavage of synthetic substrates containing non-nucleotide inserts by restriction endonucleases. Change in the cleavage specificity of endonuclease SsoII. Nucleic Acids Res 1992;20:4533–4538.
76. Sheflian G, Kubareva EA, Gromova ES, Shabarova ZA. Hydrolysis of DNA-duplexes, containing 5-fluorodeoxycytidine by restriction endonucleases. Biokhimiia 1993;58:1806–1811.
77. Brevnov MG, Kubareva EA, Volkov EM, Romanova EA, Oretskaia TS, Gromova ES, Shabarova ZA. The effect of point modifications of sugar-phosphate backbone of DNA on function of EcoRII, MvaI, and BstN1 restriction endonucleases. Mol Biol (Mosk) 1995;29:1294–1300.
78. Petrauskene OV, Yakovleva JN, Alekseev YI, Subach FV, Babkina OV, Gromova ES. DNA duplexes containing altered sugar residues as probes of EcoRII and MvaI endonuclease interactions with sugar-phosphate backbone. J Biomol Struct Dyn 2000;17:857–870.
79. Ovechkina LG, Popova SP, Zinoviev VV, Vaitkevicius DP, Janulaitis AA, Gorbunov YA, Malyigin EG. Induced dimerization of MvaI endodeoxyribonuclease by oligonucleotide substrate. Biopolimery i kletka 1988;4(N5):269–272 (in Russian).
80. Kaus-Drobek M, Czapinska H, Sokolowska M, Tamulaitis G, Szczepanowski RH, Urbanke C, Siksnys V, Bochtler M. Restriction endonuclease MvaI is a monomer that recognizes its target sequence asymmetrically. Nucleic Acids Res. 2007 Mar 7; [Epub ahead of print] doi:10.1093/nar/gkm064