J Mol Model (2009) 15:173–182
DOI 10.1007/s00894-008-0414-9

ORIGINAL PAPER

Modeling of *Escherichia coli* Endonuclease V structure in complex with DNA

Karolina A. Majorek · Janusz M. Bujnicki

Received: 1 July 2008 / Accepted: 2 October 2008 / Published online: 29 November 2008

Abstract Endonuclease V (EndoV) is a metal-dependent DNA repair enzyme involved in removal of deaminated bases (e.g., deoxyuridine, deoxyinosine, and deoxyxanthosine), with pairing specificities different from the original bases. Homologs of EndoV are present in all major phyla from bacteria to humans and their function is quite well analyzed. EndoV has been combined with DNA ligase to develop an enzymatic method for mutation scanning and has been engineered to obtain variants with different substrate specificities that serve as improved tools in mutation recognition and cancer mutation scanning. However, little is known about the structure and mechanism of substrate DNA binding by EndoV. Here, we present the results of a bioinformatic analysis and a structural model of EndoV from *Escherichia coli* in complex with DNA. The structure was obtained by a combination of fold-recognition, comparative modeling, de novo modeling and docking methods. The modeled structure provides a convenient tool to study protein sequence-structure-function relationships in EndoV and to engineer its further variants.

K. A. Majorek · J. M. Bujnicki (✉)
Institute for Molecular Biology and Biotechnology,
Adam Mickiewicz University,
Umultowska 89,
PL-61-614 Poznan, Poland
e-mail: iamb@genesilico.pl

J. M. Bujnicki
Laboratory of Bioinformatics and Protein Engineering,
International Institute of Molecular and Cell Biology,
Trojdena 4,
02-109 Warsaw, Poland

Keywords Endonuclease V · EndoV · Homology modeling · Protein-DNA docking · Protein fold-recognition

Introduction

Deoxyribonucleic acid (DNA) of all organisms is subjected to a wide range of mutagenic agents. Base deamination is a major type of DNA damage under nitrosative stress, but it can occur spontaneously as well, generating the base analogs, which have pairing specificities different from the original bases. Endonuclease V (EndoV) is a repair enzyme, which initiates removal of deaminated bases from damaged DNA. It is also called deoxyinosine 3′ endonuclease, as it preferentially cleaves DNA containing deoxyinosine, a deamination product of deoxyadenosine. However, EndoV may also recognize deoxyxanthosine, deoxyoxanosine, deoxyuridine, abasic (AP) sites, base mismatches, flap DNA, pseudo-Y structures, and small insertions/deletions in DNA molecules [1, 2]. The cleavage site generated by EndoV occurs at the second phosphodiester bond in the 3′ direction from the lesion, leaving a nick with 5′-phosphate and 3′-hydroxyl groups [1]. EndoV requires Mg^{2+} or Mn^{2+} ions for its activity. Although there is no general agreement on the number of metal ions involved in catalysis, recently a catalytic and regulatory two-metal model has been proposed, similar to the one proposed for restriction endonucleases [3]. According to this model, EndoV possesses two metal binding sites, M1 and M2. Occupation of the M1 site by a catalytic metal (Mg^{2+} or Mn^{2+}) is required for catalysis and the M1 site has relatively high affinity for metal ions. Occupation of the M2 site is not essential for catalysis, but it can regulate the activity catalyzed by the metal ion located in the M1 site. The M2 site can be occupied by Mg^{2+} or Mn^{2+}, as well as by Ca^{2+}. On the other hand, if Ca^{2+} is located in the M1 site, it inhibits the cleavage reaction. Alternatively, EndoV may follow the mechanism proposed for RNase H, in which both metal ions are catalytic [4].

EndoV homologs have been found in Eubacteria, Archaea and Eukaryota. Prokaryotic members of the family are approximately 200 amino acids in length, while the mammalian homologs are about 100 aa longer due to the C-terminal extension. Some EndoV family proteins have additional domains leading to significant enlargement of the entire protein (e.g., *C. elegans* enzyme is 758 aa long). Sequence alignments of EndoV homologs allowed identifying seven conserved regions universal to all EndoV family proteins [5]. Motif I contains an invariant Gln residue that is moderately important for substrate and product binding. Motif II includes the active site Asp residue that is essential for catalysis. Motifs III–VI contain many residues that are directly or indirectly involved in protein–DNA interactions [5]. Site-directed mutagenesis analysis of residues in conserved motifs revealed that D43 in motif II, E89 in motif III, and D110 in motif IV of *T. maritima* EndoV, hereafter referred to as TmEndoV are involved in metal cofactor coordination and catalytic function (in *E. coli* enzyme, hereafter referred to as EcEndoV, these residues correspond to D35, E82, D103, respectively). The fourth highly conserved residue, H214 in TmEndoV (D206 in EcEndoV), has been suggested to play a role in metal binding, nonetheless is the most tolerant to mutagenesis, e.g., it is exchangeable between Asp and His in the EndoV family [3]. Tyrosine at position 80 of TmEndoV (Y73 in EcEndoV) was shown to play a role in substrate and product binding, and to be important in the context of base preferences of mismatch cleavage [5].

Interestingly, substrate preference of EndoV homologs varies among different organisms. EcEndoV has a wide substrate spectrum, while EndoV from *A. fulgidus* and *H. sapiens* recognize only deoxyinosine [6]. It has been suggested that the deoxyinosine cleavage activity is a primordial activity of EndoV enzymes and that the ability of some bacterial members of this family to recognize other DNA lesions was acquired later during the course of evolution [7]. While TmEndoV can rapidly turn over T/U-containing double-stranded DNA [2], *S. typhimurium* EndoV can only turn over deoxyuridine-containing DNA to a limited extent when the substrates are in excess, likely due to tighter binding to these substrates [8]. For EcEndoV the mismatch-specific activity of the enzyme is reduced when the mismatch is flanked by GC pairs, while its deoxyinosine-specific activity is not influenced by the sequence context [9]. Nonetheless, amino acid residues essential for deaminated base recognition and DNA cleavage are highly conserved.

The deoxyinosine or the damaged bases are not removed from DNA by EndoV, and the enzyme forms a stable complex with mutated DNA both before and after cleavage. Therefore, it has been proposed that, besides its endonuclease activity, the enzyme might function to target other repair protein(s), initiating a repair pathway [1]. It was also hypothesized that the cleaved DNA is further repaired through an alternative excision repair (AER) pathway that requires the participation of either a 5′ endonuclease or a 3′–5′ exonuclease to remove the damaged base and DNA polymerase and DNA ligase for repair action [10]. After discovery of EndoV 3′-exonuclease activity, the alternative model has been proposed, in which EndoV plays a dual role in the repair process. According to this model, additional protein(s) may induce a conformational change in EndoV, causing switch from endonuclease to 3′-exonuclease mode, progressive removal of nucleotides from 3′ side to the 5′ side, and gap creation for repair synthesis [5].

Regrettably, no structural information is available for this interesting enzyme to study its sequence-function relationships in a three-dimensional context. The only information available is the discovery that EndoV is a member of the RNase H superfamily [11], but apart from the general three-dimensional fold of RNase H-like enzymes (a very diverged superfamily of proteins), the details of its structure remain unknown. Thus, we have used bioinformatic methods to produce a structural model of EcEndoV in complex with dsDNA and cofactor metal ions. The model has allowed us to provide a structural context for sequence conservation within EndoV proteins family, and to highlight the previously obtained mutations that have been shown to change its specificity.

Materials and methods

Sequence analyses

Searches of the non-redundant (nr) database were carried out at the NCBI using PSI-BLAST [12], using the sequence of EcEndoV protein as a query. Orthologous groups of proteins encoded in complete genomes were obtained from the Clusters of Orthologous Groups database (COGs) [13]. A multiple sequence alignment (MSA) of Endonuclease V family members was generated using PROMALS [14] and manually adjusted to maximize the number of aligned homologous residues and preserve the continuity of predicted secondary structure elements.

Secondary structure prediction, identification of ordered and disordered regions, and fold-recognition (FR) analysis of EcEndoV were carried out via the GeneSilico Meta-Server gateway (for references to original methods see https://genesilico.pl/meta2) [15]. FR alignments to the top-scoring templates from the Protein Data Bank were compared, evaluated and ranked by PCONS [16].

Modeling of protein tertiary structure

Comparative modeling of EcEndoV structure followed the ‘FRankenstein’s monster’ approach [17, 18], which is a method for comparative modeling by optimization of target-template alignments (usually obtained by Fold Recognition, hence capital FR in the method’s name) with the aid of model quality assessment (MQA; see the following section for a more detailed description). It comprises the following steps: First, alternative sequence alignments between the target sequence (here: EcEndoV) and template structures (here: 2nrt for core domain and 1w9h for C-terminal fragment), are obtained from various FR servers queried via the GeneSilico Meta-Server (see above). Only alignments with scores above thresholds of significance are used (the thresholds are based on the Livebench evaluation [19, 20] and are different for each server, as they use different scoring systems). Optionally, these alignments may be refined to move insertions and deletions into biologically realistic positions, e.g., on the protein surface, to avoid disruption of the protein core. Here, alignments were refined manually, taking into account positions of catalytic residues and predicted secondary structure elements. Based on the refined alignments, preliminary models are built using MODELLER [21]. Models are superimposed onto each other and scored by MetaMQAP [22].

A hybrid model (a “FRankenstein monster” itself) is constructed by merging fragments (encompassing one or more elements of secondary structure) based on the following criteria: First, for each regions with consensus alignment between >50% of models, the corresponding structural fragment is taken from the model with the best overall MetaMQAP score. For non-consensus regions, the fragments with locally best MetaMQAP scores are selected. The hybrid model is not optimized directly, but is used as a reference to construct a new hybrid alignment, and only then a new model is built and evaluated again. Subsequently, only regions with poor local MetaMQAP scores in the new model are optimized, other regions are kept unchanged, at least on the level of alignment. For such regions with poor MetaMQAP scores, new alignments are generated by progressively shifting the target sequence within the limits of predicted secondary structures. Locally modified alignments are used to generate new intermediate models, which are again evaluated. The cycles of models building, evaluation, local re-alignment in problematic regions and generation of hybrids, by merging of best scoring fragments, continues until the global MetaMQAP score cannot be improved. We have previously used this approach to successfully predict structures of other nucleases that were subsequently confirmed by crystallographic analyses, e.g., R.SfiI [23] and R.MvaI [24].

The obtained homology model was used as a starting point for de novo folding of the poorly-scoring N-terminal fragment (residues 1–28), while the homology-modeled core (residues 29–223) was kept ‘frozen’. The search of conformational space for the variable regions was carried out with ROSETTA [25], one of the best existing methods for de novo modeling of entire proteins and variable protein fragments. Representatives of the largest clusters (corresponding to the largest free-energy minima), obtained from the analysis of ROSETTA preliminary models (decoys), were selected. Full length models were obtained by merging the homology-modeled core and the best scoring de novo-modeled fragments, and again optimized using MODELLER to improve stereo-chemical parameters at the junctions between fragments and to alleviate minor steric clashes. As the final model we selected the structure that exhibited the best scores according to both MetaMQAP [22] and PROQ [26] (see below).

Model quality assessment

Although the accuracy of structural models remains unknown until the ‘real’ structure is obtained, there is a number of methods that allows for predicting the accuracy from various features of the model. One approach to predict the quality of a model is based on Anfinsen’s hypothesis, which states that the native structure is in the global energy minimum, hence better models should exhibit lower free energy. A relation between the model quality and the energy, as calculated with physical force-fields (e.g., those commonly used in molecular dynamics simulations), holds, however, only for models that are very close to the real structure (e.g., with RMSD less than 3 Å) [27]. Thus, the application of physical force-fields in connection with very extensive sampling can, sometimes, improve the quality of models that are already very good to start with, nonetheless this approach in general is unable to either assess or improve the quality of models that are outside the native energy basin [28]. On the other hand, a number of methods for model quality assessment have been based on statistical evaluation of coarse-grained features that allow for discrimination of models that are outside the global energy minimum (i.e., almost all models produced by bioinformatic approaches such as comparative modeling or de novo folding). A number of model quality assessment programs (MQAPs) have been recently developed and rigorously tested (reviews: [26, 29]). Here, we used our own “meta-server” MetaMQAP that obtains scores from a number of third-party MQAPs and uses a regression model to calculate a predicted deviation between the position of each residue in the model and its (unknown) position in the real structure [22]. MetaMQAP "consensus" scores shows better correlation with the actual model quality than the scores of constituent methods. Independently, global evaluation of model quality was performed with the PROQ method [26].

Modeling of the protein–DNA complex

Prediction of DNA-binding sites on the protein surface was carried out with PPI-PRED [30], which proposes three sites ranked according to confidence. As there is no protein-DNA complex structure for any homolog of EndoV, and EndoV active site was suggested to be related to Ribonuclease H, we superimposed the EcEndoV model with RNase H-DNA/RNA hybrid complex structure [31], removed the protein moiety of RNase H and considered the resulting EcEndoV-DNA/RNA complex as a very rough approximation of how the EndoV might interact with its nucleic acid substrate. The coordinate file of a B-DNA molecule (PDB Id: 1hq7) was obtained from the RCSB Protein Data Bank (PDB) [32]. The A to I mutation was approximated by replacing the N6 atom of adenosine with an O6 atom, without any further changes of the Watson–Crick base pairing geometry. Cofactor metal ions were copied from the RNase H-DNA/RNA complex structure. To generate a structural model of EcEndoV–DNA interaction we used HADDOCK (High Ambiguity Driven biomolecular DOCKing) version 2.0 [33, 34], a computational docking method that allows to make use of biochemical and/or biophysical information. The active and passive residues defined for HADDOCK to drive the docking process were chosen based on experimental data about EndoV residues involved in protein–DNA interactions [5], residues located in the "de novo" predicted interaction interface (PPI-PRED results), as well as residues buried in the afore-mentioned model with the DNA/RNA hybrid from the superimposed RNase H structure. A 2 Å distance was used to define the ambiguous interaction restraints (AIRs). The lowest-scoring member of the largest cluster of docking solutions was selected as a final low-resolution model of the EcEndoV–DNA complex.

Model analysis

Mapping of sequence conservation, from the EndoV family MSA onto the final model was done via the ConSurf server [35], using the neighbor joining (NJ) algorithm for generating the phylogenetic tree, with the JTT substitution matrix and empirical Bayesian method of calculating the amino acid conservation scores. The distribution of electrostatic potential was calculated for the final model using the APBS software package [36].

Results and discussion

Sequence analysis and protein fold-recognition of *E. coli* EndoV

In order to identify homologs of EcEndoV, we used its full-length sequence as a query to search the non redundant (nr) protein database with PSI-BLAST (see methods). Sequences with significant similarity to the entire query (e-value< 0.005) were found in organisms from all domains of life and included supposed EndoV orthologs as well as Excinuclease ABC subunit C (UvrC) proteins and a group of archaeal hypothetical proteins of unknown function. Unlike EndoV, both these families possess structurally characterized representatives in the Protein Data Bank: 2nrr, 2nrt, 2nrv, 2nrw, 2nrx, 2nrz for the C-terminal part of UvrC from *T. maritima* and 2qh9 (UPF0215 protein AF_1433 — a protein of unknown function from *A. fulgidus*) for the archaeal family of functionally uncharacterized proteins. Archaeal proteins turned out to be members of COG1628 family, functionally uncharacterized and annotated as Endonuclease V homologs [37]. COG1628 contains sequences from Euryarchaeota, Crenarchaeota and one protein from bacterium *Deinococcus radiodurans*.

Expectedly, secondary structure prediction for EcEndoV revealed a β-β-β-α-β-α-β pattern corresponding to common core of the RNase H fold [38, 39], with additional N-terminal α-helix and β-β-β-α-α motif in C-terminal part of the protein. Fold-recognition analysis of EcEndoV suggested that the potentially best templates for modeling of EcEndoV are the C-terminal domain of UvrC or a member of COG1628 from *A. fulgidus*. Structures of these two proteins were found on the top positions of most servers. In particular, HHsearch [40] reported UvrC structure with the score of 200.95 on the first position, and 2qh9 with the score of 186.49 on the second position, while FFAS [41] reported 2qh9 as the best template (score −40.5), followed by UvrC (score −36.1). Another example can be SAM [42], which reported UvrC with the scores in the range of 0.00052–0.0031, and 2qh9 with the score of 0.018. Ultimately, the consensus server PCONS5 [16] validated the C-terminal nuclease domain of UvrC as the best template for the catalytic core of EcEndoV (first six positions in the ranking, with scores from 0.9062 to 0.7438). Thus, we decided to use the UvrC structure as the primary modeling template. From all variants of the UvrC structure (see above) we selected 2nrt as the highest-resolution structure without missing internal fragments [43]. However, because the 2nrt structure did not cover whole EndoV sequence, additional FR searches for the C-terminal part of EndoV (residues: 131–223) were carried out. Among the top templates for this fragment were 2qh9 and *A. fulgidus* Piwi protein (structure 1w9h). Since the 1w9h structure exhibited much better agreement of secondary structure with that predicted for that region of EcEndoV (1w9h contains β-strands instead of α-helices present in 2qh9) it was used as a template for modeling of the EcEndoV C-terminal fragment.

Figure 2 shows a simplified scheme of configuration of known and predicted catalytic residues of EcEndoV, UvrC, Piwi/Argonaute, and RNase HI, with reference to the secondary structure of the catalytic core of these proteins. Position of the first Asp residue (the middle of the first β-strand) is the same in all the enzymes. Position of the second catalytic residue of EndoV (Glu) corresponds to the position of Glu in human RNase HI structure, which is located in the first α-helix of the catalytic core. The following Asp is conserved in all the enzymes, and it is located at the end of the fourth β-strand of the catalytic core (which in 3D structure is located next to the first β-strand). Piwi/Ago proteins posses an additional Glu in the second α-helix of the catalytic core. Following the fifth β-strand there are differences in the structure composition of these four enzymes, nonetheless the position of the C-terminal α-helix, carrying the last catalytic residue, namely Asp or His in EndoV family, His in UvrC and Piwi/Ago, and Asp in the RNase HI, is comparable in all the enzymes, and so is the position of the last catalytic residue.

Molecular modeling of EcEndoV structure

Comparative modeling technique requires a homologous template with known structure to be identified and the sequence of the modeled protein (a target) to be correctly aligned to the template. Although the FR analysis revealed clear homology of EndoV to proteins with the RNase H fold, the alignments returned by different methods were similar only in the regions corresponding to first (residues: 28–39), fourth (residues: 95–104) and fifth (residues: 124–131) β-strands of the EcEndoV catalytic core (altogether less than 30 residues of a 223 residue-long protein), and differed significantly in the remaining regions. Thus, modeling of EndoV should be considered as very challenging. In order to overcome the uncertainty of target-template alignments we used the 'FRankenstein's monster' approach to identify (and combine) local alignment variants that yield best-scoring models [17, 18]. Alignments of EndoV sequence to UvrC and Piwi structures reported by FR methods were manually refined to shift insertions from the protein core to loops and used as starting points for automated comparative modeling. Each model was then evaluated by MetaMQAP and all models were superimposed to generate a multiple structure alignment, where regions with alternative alignments and different structures could be compared to identify variants with locally best scores. Based on this superposition, new hybrid alignments were generated and used to generate further models. Thus, a comparative model was constructed by iterating model building, evaluation, local modification of alignments and merging of fragments with best scores (see Methods for details), until no significant improvement in MetaMQAP score was observed. The final optimized alignment is shown in Fig. 1. This template-based modeling failed to produce any model with well-scored N-terminal helix, therefore we decided to model this part de novo, using ROSETTA [25]. As the final model we selected the one with the best scores according to PROQ and MetaMQAP. In particular, PROQ predicted that the difference between our model of EcEndoV and the true (currently unknown) can be expressed as an LGscore of 3.221, indicating a potentially "very good" model. MetaMQAP also predicted that global accuracy of our model is likely to be high, with RMSD to the native structure predicted to be ~3.7 Å and the GDT_TS score predicted to be 46.749 (Fig. 2).

Figure 3 shows the predicted structure of EcEndoV, indicating conserved residues as well as the results of local accuracy prediction according to MetaMQAP. In agreement with common sense, the central β-sheet is predicted to be modeled most accurately, while peripheral helices and loops are predicted to exhibit relatively larger deviation from the true structure. Mapping of the evolutionary information from the multiple sequence alignment of EcEndoV homologs onto the surface of the model reveals that conserved residues form an elongated patch. Analysis of the electrostatic potential reveals accumulation of negative charge in the center of the conserved patch, which is generated by a cluster of carboxylate residues involved in binding of divalent metal ions. On the other hand, the more peripheral conserved residues tend to exhibit positive charge, suggesting that they may be involved in binding of the negatively-charged DNA backbone.

Modeling of the protein–DNA complex

Modeling of the protein–DNA complex was carried out in three steps. First, we predicted the DNA-binding site at the surface of the EcEndoV model, using both de novo and comparative approaches. Second, we modeled the DNA substrate. Third, we docked the DNA substrate to the protein model with restraints on the active site and predicted binding site. In the first step, we used PPI Pred [30] to identify potential sites of protein–DNA interactions on the EcEndoV models structure. The highest-scoring predicted binding patch comprised the following amino acid residues: L3, R7, V36, G37, F38, I67, A68, T69, T70,

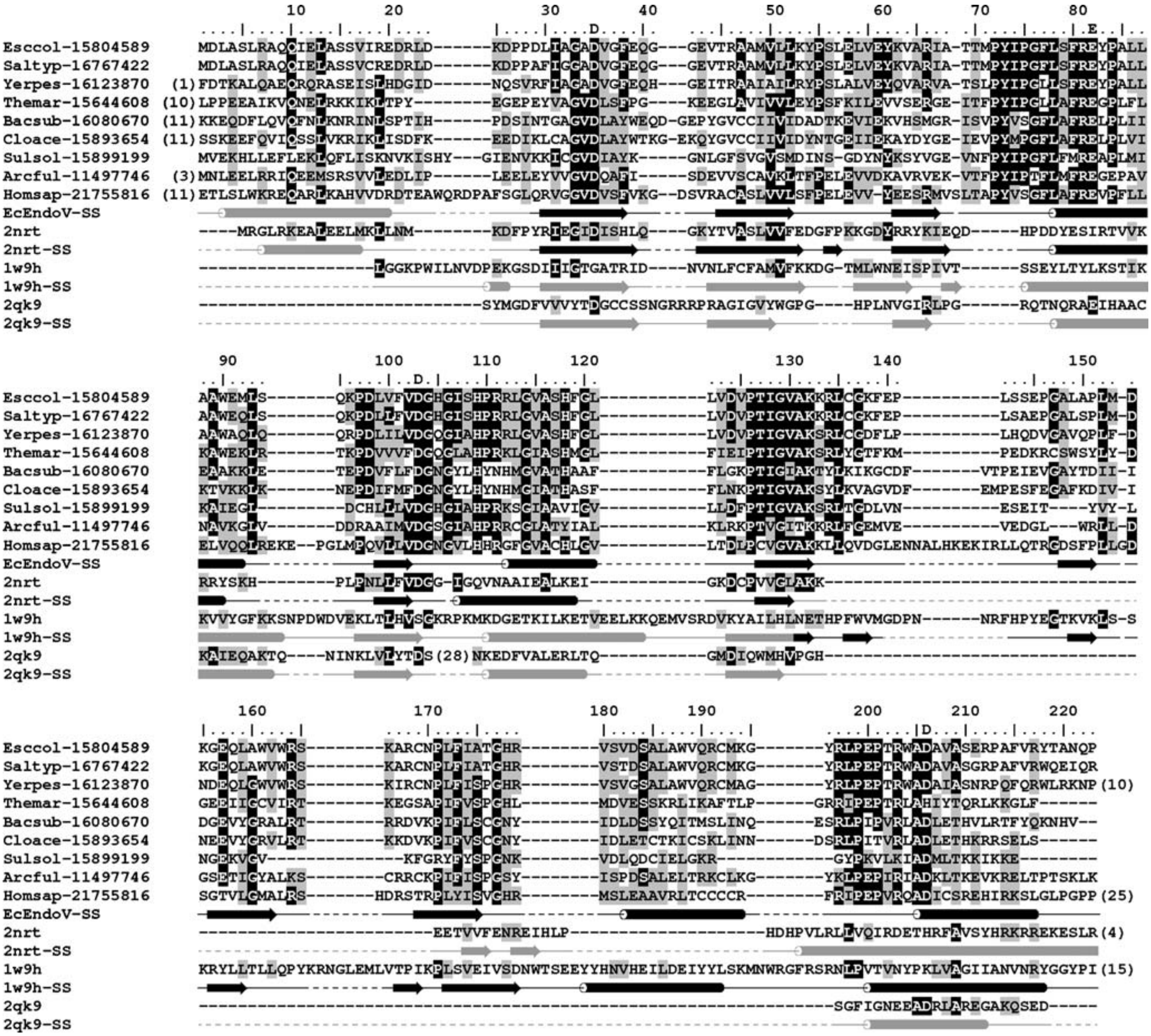
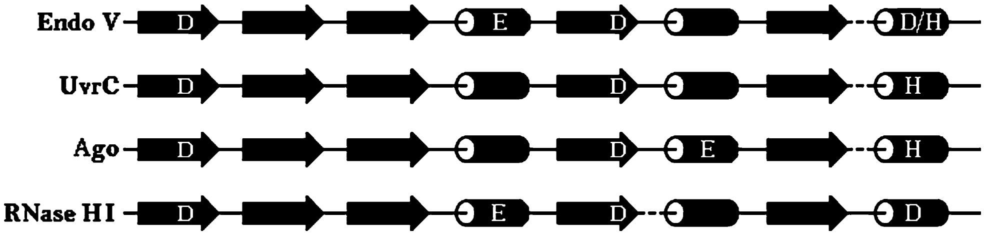


Fig. 1 Optimized fold-recognition alignment of selected members of the Endonuclease V family (indicated by a six-letter abbreviation for genus and species, followed by the NCBI gene identification (GI) number). The template structures detected by fold-recognition and used for modeling are shown at the bottom, indicated by their PDB accession numbers: 2nrt for UvrC, 1w9h for Piwi and 2qk9 for human RNase HI. Conserved residues are highlighted. The amino acid residues of EcEndoV are numbered. Secondary structure of EcEndoV (taken from the final model) and of the templates used for modeling is shown as tubes (helices) and arrows (strands). Secondary structure of the EcEndoV is colored black in regions modeled by homology modeling, while region folded de novo is shown in gray. Secondary structures of the template structures are colored black in regions used as a template for the homology modeling of corresponding regions of the EcEndoV, while regions from particular templates not used for the homology modeling are shown in gray. Residues important for the catalytic activity of EndoV proteins are indicated above the alignment by a single-letter code of the amino acid

Fig. 2 Simplified scheme of configuration of known and predicted catalytic residues with respect to the secondary structure in different members of the EndoV family that represent families used in the analysis



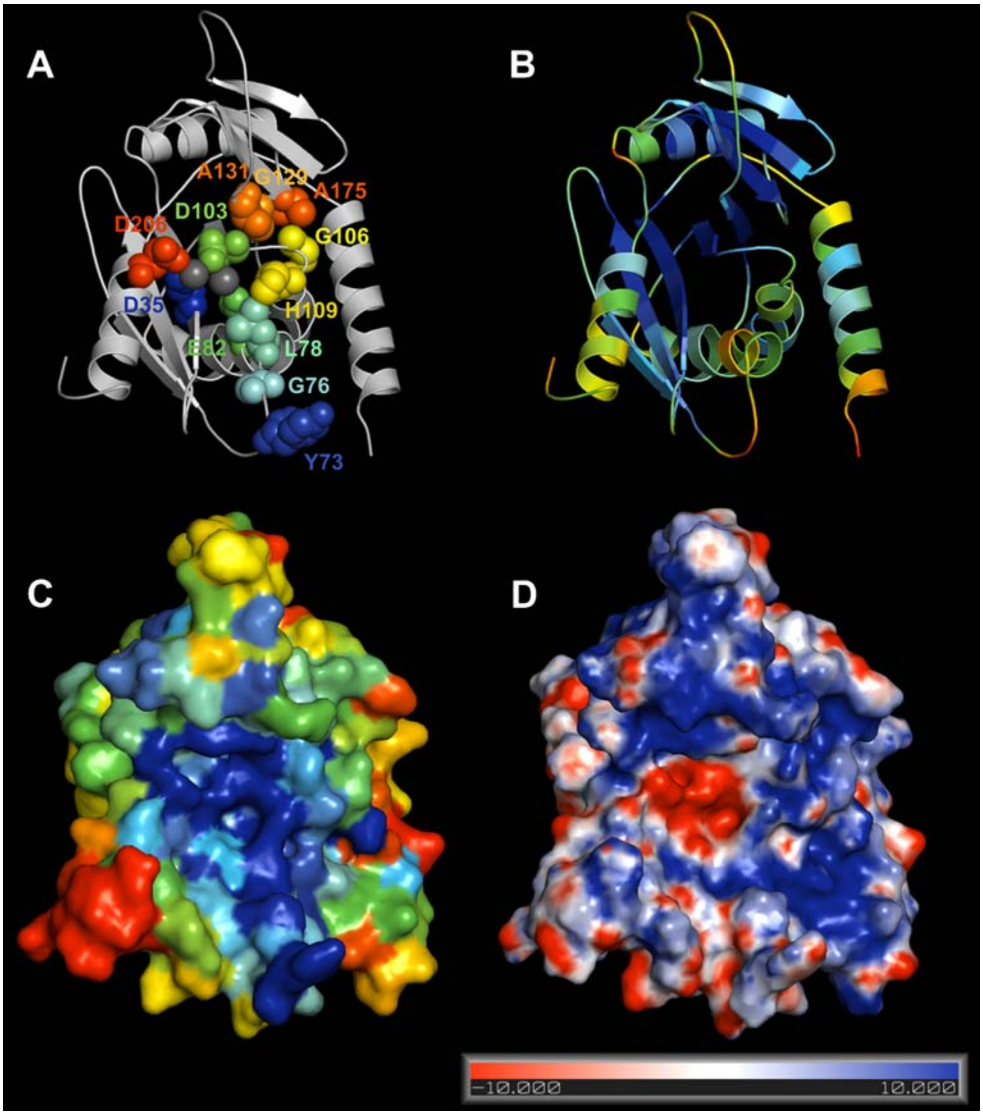


Fig. 3 A structural model of EcEndoV. Coordinates are available for download from ftp://genesilico.pl/iamb/models/EndoV/ (**a**) Functionally important residues of EcEndoV. The protein backbone is shown in the "cartoon" representation (light gray), residues important for catalysis and protein–DNA interactions are shown in the space-filled representation, labeled and colored according to the sequence index, from blue (N-terminus) to red (C-terminus). The magnesium ions are indicated by dark grey spheres. (**b**) The model colored according to the predicted accuracy (agreement with the native structure estimated for individual residues using MetaMQAP), from blue (highly confident, low predicted deviation of Cα atoms, predicted error ~1 Å), through intermediate values indicated by green to orange, to red (predicted low accuracy, error difficult to estimate (>5 Å). (**c**) Model in the surface representation, colored according to sequence conservation in the EndoV family, from deep blue (invariant) to light blue (conserved), to yellow/red (highly variable). (**d**) Model colored according to the distribution of electrostatic potential, from red (−10 kT) to blue (+10 kT)

M71, P72, Y73, I74, P75, G76, F77, L78, S79, F80, R81, S108, H109, and R112. Encouragingly, this list includes most of residues defined experimentally by Cao and coworkers as involved in protein–DNA interactions [5], and is in agreement with the position of catalytic center. Thus, it is reasonable to assume that EcEndoV is likely to bind its nucleic acid substrate in a similar manner to its homologs. In the absence of a known structure of a close homolog of EndoV in complex with DNA, we used a structure of RNaseH-RNA/DNA complex [31] as a proxy for illustrating potential protein–nucleic acid interactions. Superposition of EcEndoV onto the RNase H structure revealed that the RNA/DNA hybrid from the RNase H complex structure covered the predicted DNA-binding site of EcEndoV, suggesting that the orientation of the substrate in both enzymes may be similar. We expect that the substrate for EndoV is more likely a distorted B-DNA structure rather than A-like structure observed in the RNA/DNA duplex of RNase H complex. Therefore, in the second step of analysis, we constructed a model of the EndoV substrate as a nearly ideal B-DNA duplex, containing one crudely modeled inosine (I) instead of A. The third step—docking of the DNA substrate to the EcEndoV model was carried out with HADDOCK2.0 software [33, 34], using experimental data about amino acid residues involved in protein–DNA interactions, predictions of interaction interface and solvent accessibility as a restraints (see Methods for details).

Analysis of the EcEndoV model

The model of EcEndoV-DNA complex (Fig. 4) is in accordance with our expectations that protein and DNA structures should exhibit steric and electrostatic compatibility. It satisfies essentially all restraints used for its generation, including proximity between the DNA and the

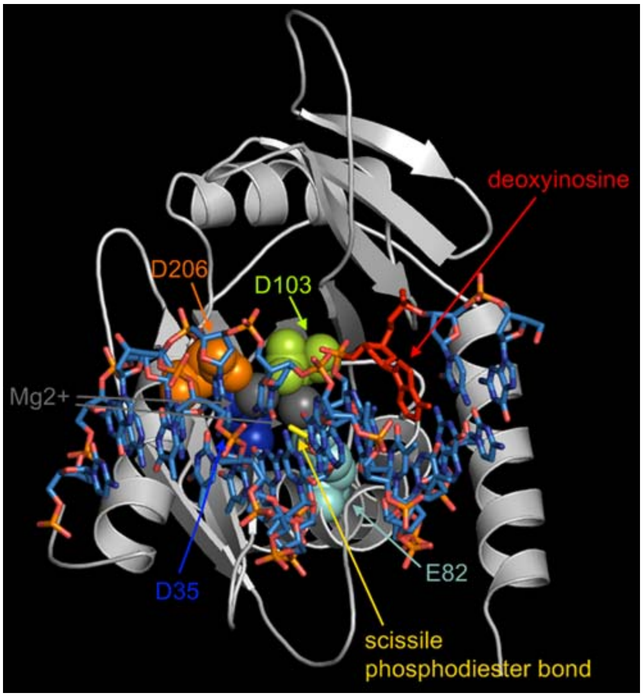


Fig. 4 A docking model of EcEndoV-DNA complex. The interaction between the molecules has been predicted by HADDOCK based on restraints from experimental data and putative protein–DNA interaction site predicted by PPI-PRED. The protein backbone is shown as a gray ribbon, the DNA is shown as sticks, with the scissile phosphodiester bond and deoxyinosine residue colored are labeled. The catalytic residues of EcEndoV and the magnesium ions are indicated by spheres and labeled

afore-mentioned experimentally determined DNA-binding residues [5]. Among residues determined to have a profound effect on both substrate and product binding, the only exception is G114, which seems to be buried and not involved in the protein–DNA interaction, although it is possible that mutations of this residue (G121 in TmEndoV) may affect protein–DNA interaction indirectly, by perturbing protein structure. The scissile phosphodiester bond (the second phosphodiester bond at the 3′ side of the deoxyinosine residue) is correctly located next to the highly conserved catalytic residues (D35, D103).

Interestingly, the model reveals no direct contacts between EndoV and the inosine base. Instead, extensive contacts with the DNA backbone are observed, suggesting that non-standard bases are recognized by EndoV using an indirect readout mechanism, e.g., based on distortion of the backbone resulting from atypical base pairing or mismatches in the substrate. The exact nature of this recognition must however await solution of a high-resolution crystal structure of EndoV-DNA complex, as such subtleties of interactions are way beyond the resolution of the current model. As mentioned earlier, the predicted RMSD of our model to the native structure is ~3.7 Å, which suggests that the structure is almost certainly outside the global energy minimum, therefore techniques of energy optimization are not expected to improve its quality. Given the predicted relative (in)accuracy of our model and its relatively large size, a fine-grained search for the global energy minimum would be prohibitively expensive (months of supercomputer time) and actually none of the available methods would guarantee that the lowest-energy structure would correspond to the correct solution. Thus, we have not attempted to optimize the geometry of the protein, DNA or the complex. At the present stage, however, the model can be used to make predictions at the level of individual residues.

The conserved Y80 of TmEndoV (Y73 in EcEndoV) was shown to be important for substrate recognition. A single alanine substitution at this position switched the base preference from purine mismatches to C-specific mismatches, while histidine substitution caused T-containing mismatch preference [44]. In our model of EcEndoV this Tyr residue makes close contact with the T29* base (complementary to the inosine). However, this region of the model corresponds to a fragment where alignments returned by different methods differed significantly, thus conformation of this region is uncertain. Substitutions of H116 in TmEndoV (H109 in EcEndoV) have led to preference for A-containing mismatches [44]. According to our model, H109 makes a contact with A4 (first base in the 3′ direction from the deoxyinosine residue). Another residue, whose mutations caused A-containing mismatch preference in TmEndoV is A86. The corresponding residue in EcEndoV is S79, which makes contact with the phosphate backbone, further supporting our suggestion that EndoV enzymes employ indirect readout to discriminate between different substrates.

Summarizing, on the level of individual residues our model represents a good fit to the existing experimental data and therefore can be used to make new predictions. In particular, we suggest that residues Y73, G76, L78, G106, H109, G129, and A131 whose counterparts have been studied by mutagenesis in TmEndoV, and additionally H105, R134 and K133 predicted by the model to be involved in EcEndoV-DNA interactions, may be interesting targets for mutagenesis aiming at altering EndoV substrate specificity. Although computational models cannot fully substitute for high-resolution crystal structures, we hope that until such a structure is obtained for EndoV-DNA complex, the models presented in this work will serve as a helpful guide for experimental analyses of this interesting family of enzymes.

Another practical application might be to use the model as a starting structure for molecular replacement (MR), should the crystallographic data become available. This might help to solve EndoV structure from the native data without the necessity of, e.g., obtaining selenomethionine derivatives. Several groups have recently demonstrated that theoretical models can be used to obtain successful MR solutions, also in cases where the original template structures used to build these models do not lead to good solutions [45, 46]. The availability of the MQAP assessment (i.e., prediction of how much each modeled residue is likely to deviate from its real position) might be very helpful in such case, as MR is very sensitive to errors in models. Hence, we would recommend using only such "core" of the model, whose predicted deviation from the true structure would be significantly lower than the nominal resolution of the crystallographic data used as an input.

Acknowledgements We thank Marcin Nowotny for critical reading of the manuscript. This work was supported by the EU 6FP (grant MRTNCT-2005-019566).

References

1. Yao M, Kow YW (1997) Further characterization of Escherichia coli endonuclease V. Mechanism of recognition for deoxyinosine, deoxyuridine, and base mismatches in DNA. J Biol Chem 272:30774–30779. doi:10.1074/jbc.272.49.30774
2. Huang J, Lu J, Barany F et al. (2001) Multiple cleavage activities of endonuclease V from Thermotoga maritima: recognition and strand nicking mechanism. Biochemistry 40:8738–8748. doi:10.1021/bi010183h
3. Feng H, Dong L, Cao W (2006) Catalytic mechanism of endonuclease v: a catalytic and regulatory two-metal model. Biochemistry 45:10251–10259. doi:10.1021/bi060512b
4. Nowotny M, Gaidamakov SA, Crouch RJ et al. (2005) Crystal structures of RNase H bound to an RNA/DNA hybrid: substrate specificity and metal-dependent catalysis. Cell 121:1005–1016. doi:10.1016/j.cell.2005.04.024
5. Feng H, Dong L, Klutz AM et al. (2005) Defining amino acid residues involved in DNA-protein interactions and revelation of 3′-exonuclease activity in endonuclease V. Biochemistry 44:11486–11495. doi:10.1021/bi050837c
6. Moe A, Ringvoll J, Nordstrand LM et al. (2003) Incision at hypoxanthine residues in DNA by a mammalian homologue of the *Escherichia coli* antimutator enzyme endonuclease V. Nucleic Acids Res 31:3893–3900. doi:10.1093/nar/gkg472
7. Liu J, He B, Qing H et al. (2000) A deoxyinosine specific endonuclease from hyperthermophile, *Archaeoglobus fulgidus*: a homolog of Escherichia coli endonuclease V. Mutat Res 461:169–177
8. Feng H, Klutz AM, Cao W (2005) Active site plasticity of endonuclease V from Salmonella typhimurium. Biochemistry 44:675–683. doi:10.1021/bi048752j
9. Yao M, Kow W (1994) Strand-specific cleavage of mismatch-containing DNA by deoxyinosine 3′- endonuclease from Escherichia coli. J Biol Chem 269:31390–31396
10. Kow YW (2002) Repair of deaminated bases in DNA. Free Radic Biol Med 33:886–893. doi:10.1016/S0891-5849(02)00902-4
11. Rand TA, Ginalski K, Grishin NV et al. (2004) Biochemical identification of Argonaute 2 as the sole protein required for RNA-induced silencing complex activity. Proc Natl Acad Sci USA 101:14385–14389. doi:10.1073/pnas.0405913101
12. Altschul SF, Madden TL, Schaffer AA et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402. doi:10.1093/nar/25.17.3389
13. Tatusov RL, Natale DA, Garkavtsev IV et al. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 29:22–28. doi:10.1093/nar/29.1.22
14. Pei J, Grishin NV (2007) PROMALS: towards accurate multiple sequence alignments of distantly related proteins. Bioinformatics 23:802–808. doi:10.1093/bioinformatics/btm017
15. Kurowski MA, Bujnicki JM (2003) GeneSilico protein structure prediction meta-server. Nucleic Acids Res 31:3305–3307. doi:10.1093/nar/gkg557
16. Lundstrom J, Rychlewski L, Bujnicki J et al. (2001) Pcons: a neural-network-based consensus predictor that improves fold recognition. Protein Sci 10:2354–2362. doi:10.1110/ps.08501
17. Kosinski J, Cymerman IA, Feder M et al. (2003) 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 53(Suppl 6):369–379. doi:10.1002/prot.10545
18. Kosinski J, Gajda MJ, Cymerman IA et al. (2005) FRankenstein becomes a cyborg: the automatic recombination and realignment of fold recognition models in CASP6. Proteins 61(Suppl 7):106–113. doi:10.1002/prot.20726
19. Bujnicki JM, Elofsson A, Fischer D et al. (2001) LiveBench-1: continuous benchmarking of protein structure prediction servers. Protein Sci 10:352–361. doi:10.1110/ps.40501
20. Bujnicki JM, Elofsson A, Fischer D et al. (2001) LiveBench-2: large-scale automated evaluation of protein structure prediction servers. Proteins Suppl 5:184–191. doi:10.1002/prot.10039
21. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234:779–815. doi:10.1006/jmbi.1993.1626
22. Pawlowski M, Gajda MJ, Matlak R et al. (2008) MetaMQAP: a meta server for the quality assessment of protein models. BMC Bioinformatics 9:403. doi:10.1186/1471-2105-9-403
23. Chmiel AA, Bujnicki JM, Skowronek KJ (2005) A homology model of restriction endonuclease SfiI in complex with DNA. BMC Struct Biol 5:2. doi:10.1186/1472-6807-5-2
24. Kosinski J, Kubareva E, Bujnicki JM (2007) A model of restriction endonuclease MvaI in complex with DNA: a template for interpretation of experimental data and a guide for specificity engineering. Proteins 68:324–336. doi:10.1002/prot.21460
25. Simons KT, Kooperberg C, Huang E et al. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 268:209–225. doi:10.1006/jmbi.1997.0959
26. Wallner B, Elofsson A (2003) Can correct protein models be identified? Protein Sci 12:1073–1086. doi:10.1110/ps.0236803
27. Bradley P, Misura KM, Baker D (2005) Toward high-resolution de novo structure prediction for small proteins. Science 309:1868–1871. doi:10.1126/science.1113801
28. Chen J, Brooks CL 3rd (2007) Can molecular dynamics simulations provide high-resolution refinement of protein structure? Proteins 67:922–930. doi:10.1002/prot.21345
29. Cozzetto D, Kryshtafovych A, Ceriani M et al. (2007) Assessment of predictions in the model quality assessment category. Proteins 69(Suppl 8):175–183. doi:10.1002/prot.21669
30. Bradford JR, Westhead DR (2005) Improved prediction of protein-protein binding sites using a support vector machines

approach. Bioinformatics 21:1487–1494. doi:10.1093/bioinformatics/bti242

31. Nowotny M, Gaidamakov SA, Ghirlando R et al. (2007) Structure of human RNase HI complexed with an RNA/DNA hybrid: insight into HIV reverse transcription. Mol Cell 28:264–276. doi:10.1016/j.molcel.2007.08.015
32. Berman HM, Westbrook J, Feng Z et al. (2000) The protein data bank. Nucleic Acids Res 28:235–242. doi:10.1093/nar/28.1.235
33. Dominguez C, Boelens R, Bonvin AM (2003) HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc 125:1731–1737. doi:10.1021/ja026939x
34. van Dijk M, van Dijk AD, Hsu V et al. (2006) Information-driven protein-DNA docking using HADDOCK: it is a matter of flexibility. Nucleic Acids Res 34:3317–3325. doi:10.1093/nar/gkl412
35. Landau M, Mayrose I, Rosenberg Y et al. (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res 33:W299–W302. doi:10.1093/nar/gki370
36. Baker NA, Sept D, Joseph S et al. (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. Proc Natl Acad Sci USA 98:10037–10041. doi:10.1073/pnas.181342398
37. Makarova KS, Koonin EV (2003) Comparative genomics of Archaea: how much have we learned in six years, and what's next? Genome Biol 4:115. doi:10.1186/gb-2003-4-8-115
38. Katayanagi K, Miyagawa M, Matsushima M et al. (1990) Three-dimensional structure of ribonuclease H from E. coli. Nature 347:306–309. doi:10.1038/347306a0
39. Yang W, Hendrickson WA, Crouch RJ et al. (1990) Structure of ribonuclease H phased at 2 A resolution by MAD analysis of the selenomethionyl protein. Science 249:1398–1405. doi:10.1126/science.2169648
40. Soding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21:951–960. doi:10.1093/bioinformatics/bti125
41. Jaroszewski L, Rychlewski L, Li Z et al. (2005) FFAS03: a server for profile–profile sequence alignments. Nucleic Acids Res 33: W284–W288. doi:10.1093/nar/gki418
42. Karplus K, Katzman S, Shackleford G et al. (2005) SAM-T04: what is new in protein-structure prediction for CASP6. Proteins 61 (Suppl 7):135–142. doi:10.1002/prot.20730
43. Karakas E, Truglio JJ, Croteau D et al. (2007) Structure of the C-terminal half of UvrC reveals an RNase H endonuclease domain with an Argonaute-like catalytic triad. EMBO J 26:613–622. doi:10.1038/sj.emboj.7601497
44. Gao H, Huang J, Barany F et al. (2007) Switching base preferences of mismatch cleavage in endonuclease V: an improved method for scanning point mutations. Nucleic Acids Res 35:e2. doi:10.1093/nar/gkl916
45. Raimondo D, Giorgetti A, Giorgetti A et al. (2007) Automatic procedure for using models of proteins in molecular replacement. Proteins 66:689–696. doi:10.1002/prot.21225
46. Schwarzenbacher R, Godzik A, Jaroszewski L (2008) The JCSG MR pipeline: optimized alignments, multiple models and parallel searches. Acta Crystallogr D Biol Crystallogr 64:133–140. doi:10.1107/S0907444907050111