1992–2005 *Nucleic Acids Research, 2006, Vol. 34, No. 7*
doi:10.1093/nar/gkl132

Structural model for the multisubunit Type IC restriction–modification DNA methyltransferase M.EcoR124I in complex with DNA

Agnieszka Obarska, Alex Blundell[1], Marcin Feder, Štěpánka Vejsadová[1,2], Eva Šišáková[2], Marie Weiserová[2], Janusz M. Bujnicki and Keith Firman[1,*]

Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, Trojdena 4, 02-109 Warsaw, Poland, [1]IBBS Biophysics Laboratories, School of Biological Sciences, University of Portsmouth, King Henry Building, King Henry I Street, Portsmouth PO1 2DY, UK and [2]Institute of Microbiology, Czech Academy of Sciences, Videnska 1083, 142 20 Prague 4, Czech Republic

Received November 25, 2005; Revised December 16, 2005; Accepted March 14, 2006

ABSTRACT

Recent publication of crystal structures for the putative DNA-binding subunits (HsdS) of the functionally uncharacterized Type I restriction–modification (R-M) enzymes MjaXIP and MgeORF438 have provided a convenient structural template for analysis of the more extensively characterized members of this interesting family of multisubunit molecular motors. Here, we present a structural model of the Type IC M.EcoR124I DNA methyltransferase (MTase), comprising the HsdS subunit, two HsdM subunits, the cofactor AdoMet and the substrate DNA molecule. The structure was obtained by docking models of individual subunits generated by fold-recognition and comparative modelling, followed by optimization of inter-subunit contacts by energy minimization. The model of M.EcoR124I has allowed identification of a number of functionally important residues that appear to be involved in DNA-binding. In addition, we have mapped onto the model the location of several new mutations of the *hsdS* gene of M.EcoR124I that were produced by misincorporation mutagenesis within the central conserved region of *hsdS*, we have mapped all previously identified DNA-binding mutants of TRD2 and produced a detailed analysis of the location of surface-modifiable lysines. The model structure, together with location of the mutant residues, provides a better background on which to study protein–protein and protein–DNA interactions in Type I R-M systems.

INTRODUCTION

Type I restriction and modification (R-M) systems are encoded by three genes. All three genes are required for production of the restriction endonuclease (REase); *hsdR* is absolutely required for restriction and is transcribed from its own promoter (P_{RES}); while *hsdM* and *hsdS* are transcribed from a separate promoter (P_{MOD}) and together are required for modification [for recent reviews of these enzymes see Sistla and Rao (1) and Loenen (2) or Murray (3)]. The *hsdS* and *hsdM* genes can also produce an independent methyltransferase with a stoichiometry of $HsdM_2$:$HsdS_1$ (4,5), which is the core DNA-binding component of the R-M enzyme.

The Type I restriction and modification systems were originally divided into three families (Type IA e.g. EcoKI, Type IB e.g. EcoAI and Type IC e.g. EcoR124I) based on gene order, amino acid conservation and enzymatic properties (6–8). More recently, additional families [Type ID e.g. StySBLI (9) and Type IE e.g. KpnBI (10)] have been introduced. Within each family there are distinct regions of the HsdS subunit in which amino acid identities are strongly conserved. One such region lies about midway between the C- and N-termini and is known as the central conserved region; while the other region is at the C-terminus (11–13). Outside of these conserved regions the amino acid sequences are highly variable even between members of the same family and these variable regions appear to be responsible for DNA recognition (Figures 1 and 2D). These two variable regions have been named TRD1 and TRD2 (for target recognition domains) and can be 'swapped' between related systems to generate novel DNA specificities (14,15). Accordingly, it was proposed (16) that HsdS comprise two repeats of mutually homologous modules, each comprising one conserved region, and one target-recognition domain (TRD) and this has been confirmed by the isolation of deletion mutants of *hsdS* that produce a

*To whom all correspondence should be addressed. Tel: +44 2392 842059; Fax: +44 2392 842070; Email: keith.firman@port.ac.uk

MTase of stoichiometry $HsdM_1$:$HsdS_{0.5}$ (17,18) in which the one-half, deleted, HsdS subunit can dimerize to produce a MTase with a symmetrical DNA recognition sequence.

This hypothesis was, more recently, also confirmed by the crystallographic analysis of the HsdS subunits of the hypothetical (functionally uncharacterized) Type I R-M systems: MjaXIP (ORF MJ0130m) from *Methanococcus jannaschii* (19) and MgeORF438P (ORF MG3435) from *Mycoplasma genitalium* (20). HsdS(MjaXIP) and HsdS(MgeORF438P) exhibit an overall cyclic topology with an intramolecular 2-fold axis that superimposes two globular TRDs connected by long, conserved α-helices arranged into an antiparallel, coiled-coil structure that comprise most of the central conserved region. Remarkably, the TRDs of Type I HsdS subunits were found to be homologous to the TRD of a Type II MTase—M.TaqI (21) despite the lack of evident sequence similarities. However, neither HsdS(MjaXIP) nor HsdS(MgeORF438P), or their respective putative R-M systems, have been analysed functionally and hence details of sequence–structure–function relationships in these HsdS subunits remain obscure. Second, the orientation of the TRDs and the coiled-coil region are completely different between HsdS(MjaXIP) and HsdS(MgeORF438P). This suggests that significant domain motion occurs in HsdS upon binding of the DNA and the HsdM subunits [cf. Ref. (22)]. However, the putative target DNA sequences of MjaXIP and MgeORF438P that determine the mutual orientation of the TRDs are unknown, thus the respective protein–DNA complexes cannot be modelled reliably. In fact, crude docking models generated for MjaXIP (19) and MgeORF438P (20) differ greatly.

Summarizing, the structures of HsdS(MjaXIP) and HsdS(MgeORF438P) provide useful platforms for the analysis of individual domains, but their quaternary structures should be viewed with caution and models of related sequences should be viewed with an open mind.

In contrast to the aforementioned putative proteins, the EcoR124I R-M system has been studied extensively and a great deal of information, describing the sequence-function relationships in the HsdS and HsdM subunits of this system, exists in the literature.

In this paper, we have used bioinformatic methods to produce a structural model of the M.EcoR124I MTase comprising the HsdM(EcoR124I) and HsdS(EcoR124I) subunits, based on the crystal structures of HsdS(MjaXIP) (19), HsdS(MgeORF438P) (20) and M.TaqI–DNA complex (23), with the docking of domains guided by experimental data on protein–DNA interactions in EcoR124I. At the very last stage of the modelling, we had the opportunity to include, as an additional template, the crystal structure of the EcoKI HsdM subunit, which had been solved in the meantime (2ar0 in the Protein Data Bank, K. R. Rajashankar, R. Kniewel and C. D. Lima, manuscript submitted). The model of the M.EcoR124I complex has allowed us to provide a structural context for sequence conservation between HsdS(EcoR124I) and related HsdS subunits, in particular StySKI (24), discuss the location of a number of DNA-binding mutations within the *hsdS* gene of EcoR124I (25), identify the location of previously described surface-exposed lysines (26,27) and the opportunity to discuss a collection of new point mutations isolated in the central conserved region of HsdS in the context of a structural model, which provides a strong indicator for future analysis of this model structure.

MATERIALS AND METHODS

Sequence alignment

Searches of the non-redundant (nr) database were carried out at the NCBI using PSI-BLAST (28) with the *E*-value threshold of 10^{-30}, using the protein sequences of EcoR124I HsdS and HsdM as queries. The searches converged after the 14th and 9th iteration, respectively, yielding 404 HsdS and 495 HsdM sequences reported above the threshold. Multiple sequence alignments were generated using MUSCLE (29) with default parameters and subsequently adjusted manually, based on the analysis results of secondary structure prediction (see below), to ensure that no unwarranted gaps are introduced within α-helices and β-strands.

Protein structure prediction

Secondary structure prediction and tertiary fold-recognition was carried out via the GeneSilico meta-server gateway at http://genesilico.pl/meta/ (30). Since the meta-server accepts only protein sequences <500 amino acids, the sequence of M.EcoR124I (520 amino acids) was submitted in two variants, each having 20 amino acids deleted from either N- or C-terminus and the predictions were merged. Secondary structure was predicted as a consensus of the following methods: PSIPRED (31), PROFsec (32), PROF (33), SABLE (34), JNET (35), JUFO (36) and SAM-T02 (37). Solvent accessibility for the individual residues was predicted with SABLE (34) and JPRED (35). The fold-recognition analysis (attempt to match the query sequence to known protein structures) was carried out using FFAS03 (38), SAM-T02 (37), 3DPSSM (39), BIOINBGU (40), FUGUE (41), mGENTHREADER (42) and SPARKS (43). Fold-recognition alignments reported by these methods were compared, evaluated and ranked by the Pcons server (44).

Homology modelling of protein monomers

Fold-recognition alignments to the structures of selected templates were used as a starting point for homology modelling using the 'FRankenstein's Monster' approach (45), comprising cycles of model building, evaluation, realignment in poorly scored regions and merging of best scoring fragments. The positions of predicted catalytic residues and secondary structure elements were used as spatial restraints. Briefly, preliminary models were generated based on the alignments to various template structures returned by the FR servers. The sequence–structure fit in these models was assessed using VERIFY3D (46) and visualized using the COLORADO3D server (47). The most common and best-scoring fragments were merged to produce a hybrid model, in which the sequence–structure was re-evaluated. In the poorly scoring fragments shifting the sequences within the limits of predicted secondary structures locally modified the alignment and a next generation of models corresponding to different alignments was generated. The cycles of evaluation of models, generation of hybrids and local re-alignment in problematic regions

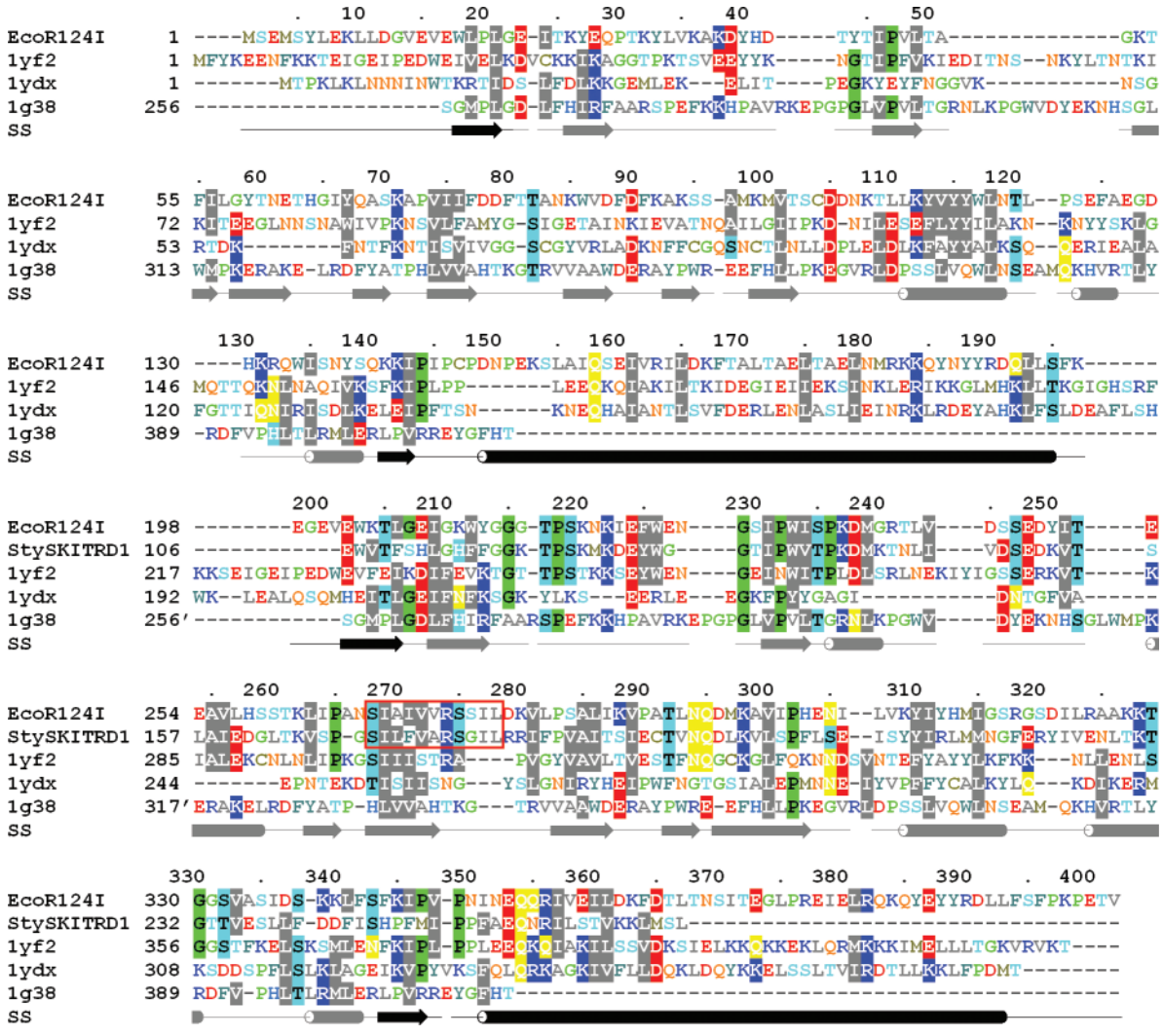

Figure 1. Alignment of the EcoR124I HsdS sequence with the StySKI TRD1 and proteins of known structure identified as closely related by the bioinformatic analysis: Multiple sequence alignment of the EcoR124I HsdS sequence with the StySKI TRD1 (27.9% identity) and proteins of known structure identified as closely related by a bioinformatic analysis: 1yf2 (HsdS(MjaXIP) from *M.jannaschii* DSM26; 19.7% identity to EcoR124I), 1ydx (HsdS(MgeORF438P) from *M.genitalium*; 11.7% identity to EcoR124I) and 1g38 (M.TaqI from *Thermus aquaticus*; 9.0% identity to the TRD1 and 10.4% identity to TRD2 of EcoR124I). The alignment between the crystal structures was derived from their spatial superposition, whereas the alignment of EcoR124I and StySKI was guided by the fold-recognition analysis. The red frame indicates region 268–278, which is conserved between EcoR124I and StySKI (for details see the main text). The single TRD of M.TaqI is present in two copies, aligned to the structures of TRD1 and TRD2 of HsdS. Aliphatic residues are in grey, positively charged in blue, negatively charged in red, and their amides in orange, residues with OH groups are in cyan, residues containing sulphur are in dark yellow, Pro and Gly are in green. Secondary structure prediction of EcoR124I is indicated below the alignment, with grey and black colours indicating variable and conserved regions, respectively.

continued until the global VERIFY3D score could not be improved.

Modelling of the protein–protein–DNA–ligand complex

The preliminary model of the $(HsdM\text{-}AdoMet)_2$–HsdS–DNA complex was constructed by superposition of two copies of the M.TaqI–DNA complex structure (23), with the cofactor analogue replaced by the AdoMet molecule taken from the M.TaqI–AdoMet complex (48), onto the structure of the HsdS(MjaXIP) subunit (49), using the homologous TRD structures from both proteins as a reference. Subsequently, the model of the EcoR124I HsdS subunit was superimposed onto the template HsdS structure, while the models of the EcoR124I HsdM subunits were superimposed onto the M.TaqI structure. Then, the structures of HsdS(MjaXIP) and M.TaqI were removed, leaving only the HsdS subunit of EcoR124I bound to two HsdM subunits of EcoR124I, each with the DNA half-site and the AdoMet molecule. The ends of the DNA half-sites that extruded from the complex were extended in an ideal B-form to facilitate the measurement and adjustment of their angle.

Subsequently, the preliminary model of the $(HsdM\text{-}AdoMet)_2$–HsdS–DNA complex was divided into two rigid parts (each comprising the whole HsdM–AdoMet complex, one TRD and half of the DNA structure). The mutual position of these two parts was adjusted by introducing shifts and rotations to overlay the DNA structures so as to separate the target adenines by exactly 7 bp and yield the angle of 49° between the extruding 'arms'. The break in the HsdS

structure was 'repaired' by superimposing the ends of the isolated coiled-coil domain structure onto the respective amino acids in the shifted halves of the HsdS and merging it with the two TRDs to produce one continuous HsdS polypeptide.

In order to 'ligate' the two halves, the DNA molecules were also merged into one continuous duplex and remodelled using HyperChem 7.1 (Hypercube, Inc.) by 'mutating', deleting and ligating the bases in the original structures to match the recognition site of HsdS(EcoR124I): the two double-stranded DNA molecules 5′-GTTCG**A**TGTC-3′/5′-GACATC-G(m^6A)AC-3′ and 5′-GACATCG(m^6A)AC-3′/5′-GTTCG**A**-TGTC-3′ (where boldface 'A' indicates a flipped-out adenine), from the M.TaqI structure were modified to obtain a single duplex 5′-GTTGA**A**TGT*GACATCGAAC-3′/5′-GT-TCG**A**TGTC*ACATTCAAC-3′ (where the mutated bases are underlined and asterisk indicates the site of deletion of 1 bp and subsequent ligation of two molecules). The geometry of the DNA molecule and the hydrogen-bonding pattern between the newly introduced bases was initially corrected by the energy minimization of modified bases and their immediate neighbours (with the rest of the molecule 'frozen') using the Fleetcher–Greeves steepest descent method (without the protein, *in vacuo*, until convergence) implemented in HyperChem 7.1.

Finally, the structure of the $(HsdM\text{-}AdoMet)_2$–HsdS–DNA complex was energy-minimized to remove steric clashes between the molecules and to allow formation of favourable contacts between all components, in particular between HsdM and HsdS and between the protein and the DNA. This required the determination of electrostatic potential (ESP) charges for AdoMet by the restrained ESP fitting method (50). ESPs were derived from HF/6–31G* PCM quantum mechanical calculations in water performed using Gaussian 03 package (51). The hydrogens were added to the complex structure using the XLEAP module of AMBER 8 (52) and the minimization was carried out using the SANDER module. The step length was set to 0.001 ps. The non-bonded cut-off was set to 18 Å. The Hawkins *et al*. (53,54) pair-wise generalized Born solvation model was used for the non-polarizable force field ff99 (55) with parameters described by Tsui and Case (56). One hundred cycles of steepest descent were followed by 1056 cycles of conjugate gradients. The minimization was stopped when the root mean square deviation of the Cartesian elements of the energy gradient was <0.1 kcal mol^{-1}.

Mutagenesis techniques

(i) *Random mutagenesis*: The central conserved domain of the HsdS subunit with surrounding regions (from 124 to 232 amino acids, between the EcoRI and the NcoI sites of the *hsdS* gene) was subjected to random mutagenesis using Mn^{2+}-induced misincorporation in the PCR amplification reaction as described previously by Weiserova and Firman (57). The plasmid pJS491carrying the wt *hsdS* gene of the EcoR124I system under control of the P_{T7g10} promoter Patel *et al*. (58) serves as a template DNA in the PCR using primers harbouring EcoRI site and NcoI site, respectively, allowing insertion of the PCR product back into the EcoRI–NcoI digest of pJS491.

(ii) *Site-directed mutagenesis*: The Quick-Change XL mutagenesis kit of Stratagene was employed for site-directed mutagenesis of both the wt *hsdS* and the mutant gene *hsdS*K^{184}N present on pJS491, respectively. The top strand of the primers used for the Lys^{384} Asn substitution was 5′-GAAATCGAGTTGCGCCAGAA**C**CAATACGAGT-ACTATCGTG-3′.

DNA manipulations

The *Escherichia coli* XL1-Blue strain, provided with the Quick-Change kit, was used for recovering the plasmids after random and site-directed mutagenesis. Plasmid DNA was isolated using the Perfectprep Plasmid Mini (Eppendorf) or the StrataPrep Plasmid Miniprep kit (Stratagene). All the restriction enzymes, *Taq* polymerase, Klenow enzyme and T4 DNA ligase were supplied by Fermentas. DNA sequences were determined using a Vistra DNA sequencer 725. Manipulations of nucleic acids were performed using the methods described in Sambrook *et al*. (59). Transformation into XL1-Blue was as recommended in the Quick-Change system.

Restriction–modification phenotype analysis

Phenotypes of resulting plasmids were analysed in complementation assay as described by Abadjieva *et al*. (18). Briefly this assay is based upon competition between the HsdS subunit of specificity EcoR124I, introduced on the plasmid pJS491 (wt or mutant), and the HsdS subunit of specificity EcoR124II, expressed from plasmid pKF650, to produce an active endonuclease. When the wt HsdS(R124) is present in the cell, the restriction and modification activities of both specificities are expressed. The virulent mutant of phage λ (60) was used for testing of restriction and modification. All assays were carried out in JM109(DE3) (61) in the absence of isopropyl-β-D-thiogalactopyranoside [the background level of T7 RNA polymerase has been found to be sufficient for restriction and modification activity (18)]. For screening of large collection of potential random mutants the spot tests were used as described in Colson *et al*. (62). Clones expressing even slight variance from the wt restriction phenotype were further quantitatively analysed for precise estimations of restriction and modification levels as described in Hubacek and Glover (63) and mutations were identified by sequencing of *hsdS* genes present on the appropriate plasmids used for transformation of JM109(DE3)[pKF650]. The cultivation media and antibiotics were used as previously described (64). The *E.coli* strain C122 (prototroph, Δ*hsd* British Culture Collection strain No 122) itself, or with either R124 (64) of R124/3 (65) plasmids, served for *in vivo* modification assays.

RESULTS

Structure Prediction of the HsdS and HsdM subunits of EcoR124I

The protein fold-recognition (FR) analysis [for review see (66)] was used to identify the best modelling templates for sequences of the HsdS(EcoR124I) and HsdM(EcoR124I) subunits. For HsdS(EcoR124I), all the FR servers suggested that the best template was the structure of HsdS(MjaXIP)

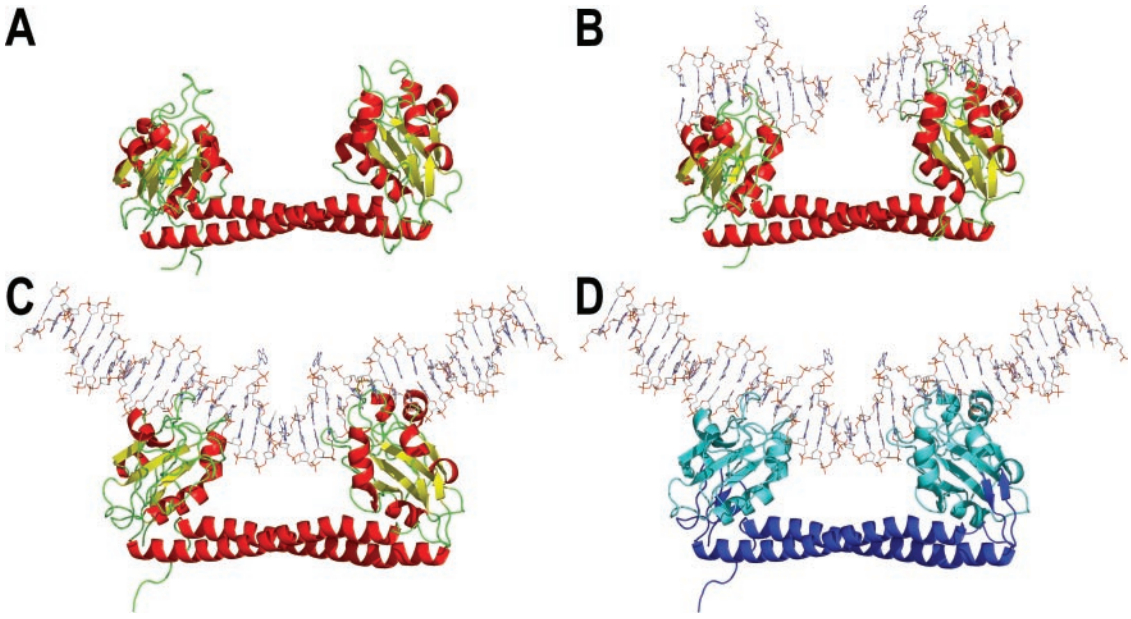

Figure 2. Predicted 3D structure of the DNA-binding subunit (HsdS) of EcoR124I. (**A**) The HsdS(MjaXI) crystal structure (1yf2). (**B**) The preliminary model of HsdS(EcoR124I) produced using the 'FRankenstein monster' approach as detailed in Materials and Methods, with two DNA half-sites. (**C**) The final energy-minimized model of HsdS(EcoR124I), with the DNA fitted to the experimentally measured 49° bend. (**D**) Mapping of the conserved (blue) and variable (cyan) regions (16) onto the HsdS(EcoR124I) structure.

(19); some servers suggested, as the second-best template, the C-terminal TRD of M.TaqI, which exhibits the same fold as the TRDs of the Type I HsdS subunits. HsdS(MgeORF438P) (20) was suggested as the second best template by only a few servers. Apparently, even a few months after the publication of this structure, it has not yet been included in the template libraries of other servers. Ultimately, the consensus server Pcons assigned a high reliability score of 6.5 to the HsdS(MjaXIP) structure as the best template. The set of FR alignments, produced by different methods, differed slightly (data not shown) and were used as starting points for the modelling of HsdS(EcoR124I), according to the 'FRankenstein's monster' approach (67) for simultaneous optimization of the target-template alignment in 1D and evaluation of the corresponding protein structure in 3D (for details see Materials and Methods). It is noteworthy that this approach has been evaluated as one of the best modelling methods in the recent CASP-6 competition (http://predictioncenter.org/casp6/). Figure 1 shows the final alignment between HsdS(EcoR124I) and its selected close homologs, the template secondary structures (ss) and the repeated, conserved regions within the domain structure. The resulting model of HsdS(EcoR124I) (Figure 2) exhibited a good VERIFY3D score (0.297), with the core regions scoring >0.3, which suggests that all major errors in the initial alignments were corrected and that the model may contain significant inaccuracies only in the extended loops. It must be emphasized, however, that the mutual orientation of domains in this initial model is arbitrarily identical to that in HsdS(MjaXIP) and is likely to be modified in the protein–DNA complex (see below).

A similar approach was used to model the HsdM subunit of EcoR124I. All the FR algorithms, run via the GeneSilico metaserver, indicated, with a very high confidence (Pcons consensus score >5.4), that the best template for modelling of the central region of HsdM(EcoR124I) (amino acids 201–420) was the catalytic, N-terminal domain, of the Type II DNA:m^6A MTase M.TaqI (23). All other protein structures (nearly all of them being various MTases) received significant, but much lower scores (<3.0). The model of the catalytic domain of HsdM was built using the FRankenstein monster approach, as described above for HsdS and was evaluated as acceptable according to VERIFY3D (average score 0.259). The final target-template alignment (data not shown) was similar to that reported earlier, in a related work on modelling of the HsdM subunit of EcoKI (68). At the very last stage of the modelling, we had the opportunity to include, as an additional template, the crystal structure of the EcoKI HsdM subunit, which had been solved (2ar0 in the PDB, K. R. Rajashankar, R. Kniewel and C. D. Lima, manuscript submitted). The catalytic domain of HsdM(EcoKI) was very similar to the catalytic domain of our model of HsdM(EcoR124I), which strongly supported the accuracy of the initial prediction, based solely on M.TaqI. The HsdM(EcoKI) structure was used as the template to model the additional domain composed of the N- and C-terminal regions of HsdM(EcoR124I) (amino acids 1–200 and 421–520). This domain was placed in an arbitrary orientation with respect to the catalytic domain, because the mutual orientation of the domains is unknown. In the crystal structure of HsdM(EcoKI) the orientation of catalytic domains and the additional domains seems to be dictated mostly by crystal packing (data not shown) and is probably irrelevant to function, as the HsdS subunit of EcoKI is missing. In the final model of the full-length DNA methyltransferase M.EcoR124I (Figure 3), we have also included the target DNA structure (with flipped-out target adenines) copied from the template structure 1g38 of M.TaqI (23) and the methyl group donor AdoMet copied from another M.TaqI structure 2adm (48). The M.TaqI DNA sequence included

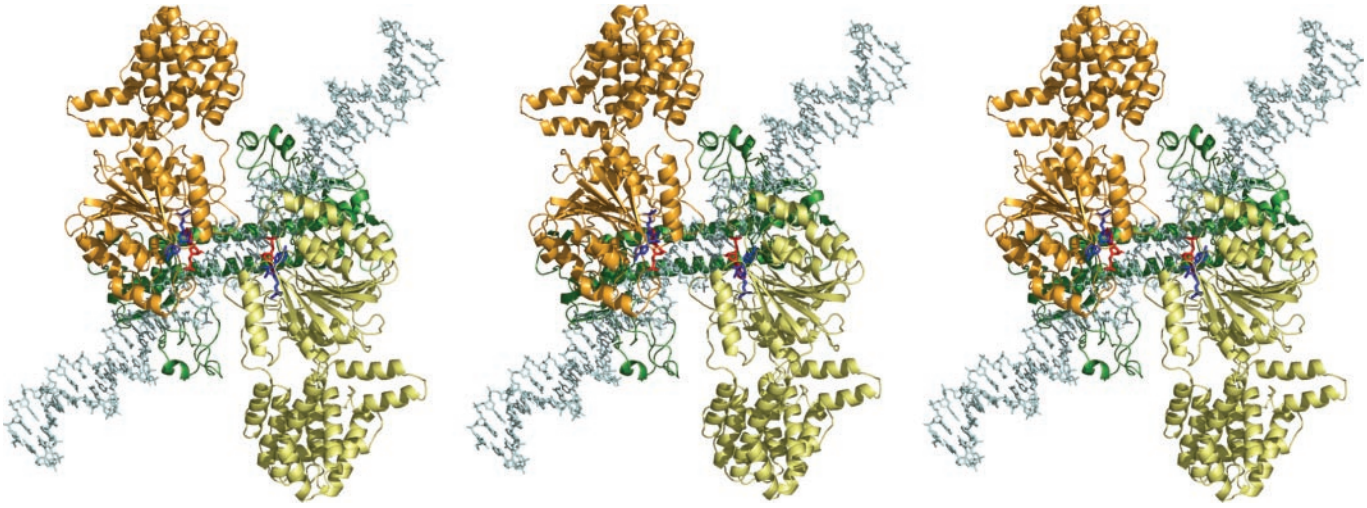

Figure 3. Predicted structure of the M.EcoR124I DNA methyltransferase and substrate. The full model of the M.EcoR124I DNA methyltransferase, as a series of three images at different angles, allowing stereoscopic vision, 'cross-eye' using the left and middle image, and 'wall-eye' using the middle and right image. The HsdS subunit is shown in green and can be seen at the bottom of the figure, with the helical coiled-coil region below the DNA. The two HsdM subunits (in yellow and orange) can be seen located on either side of the DNA in a symmetrical arrangement. The adenines of the recognition sequence, which are to be methylated are shown 'flipped-out' (70,81,82), coloured red, and the methyl donor is present above these bases (in blue).

the TCG**A**T segment (target adenine in bold/underline), which agreed with one half-site (CG**A**Y) recognized by HsdS(EcoR124I). A second variant of the model was generated in which the DNA was 'mutated' to yield the second HsdS(EcoR124I) half-site GA**A** (for details see Materials and Methods). However, the DNA substrate present in this model is far from the observed conformation of the DNA in the methyltransferase (69), where the DNA is bent through 49°, an observation that was also made for the related EcoKI MTase (70,71).

A predicted model of the active form of EcoR124I MTase, (HsdM-AdoMet)$_2$–HsdS–DNA complex

The close structural similarity and evident homology between both the TRDs and the catalytic domains of Type I MTases and the Type II MTase M.TaqI, suggests that all these domains may bind DNA in a similar manner. Based on this premise, very preliminary models of protein–DNA complexes were already reported for HsdS(MjaXIP) (19) and HsdS (MgeORF438P) (20) by superposing two copies of the M.TaqI–DNA complex (23) onto the HsdS structures, such as to overlay the homologous TRDs. This modelling suggested a separation of 8 bases between the two adenines that would be methylated. The HsdS(MjaXIP)–DNA model suggested some kinking (about 25°) and unwinding of the DNA between the two half-sites contacted by the TRDs. While the HsdS(MgeORF438P)–DNA model suggested a straight B-form of the DNA. However, in neither model were the ends of the two DNA duplexes from M.TaqI molecules perfectly aligned, suggesting that these models should be regarded, at best, as very rough approximations of protein–DNA interactions and that a much better model could be built by taking the known structure of a real DNA target into account.

The HsdS(EcoR124I) subunit imparts specificity for the sequence 5′-GA**A**NNNNNNR**T**CG-3′/5′-CG**A**YNNNNNN**T**TC-3′ (with the target adenines and complementary thymines in bold/underline), i.e. with a separation of 7 bases between the methylation sites. It has also been shown that the DNA, in the complex with the EcoR124I MTase, is bent by 49° (69). Thus, modelling of the functional form of the M.EcoR124I–DNA complex requires considerable modification of the DNA structure and rearranging of the domains compared with that described for HsdS(MjaXIP) or HsdS(MgeORF438P).

The initial model of the EcoR124I MTase was constructed by using the superimposed structures of HsdS(MjaXIP) and M.TaqI as templates. Then, breaks were introduced within the coiled-coil region that separates the two TRDs, to subdivide the whole structure into two parts, each comprising the HsdM–AdoMet complex, a half of the HsdS subunit, and the DNA molecule including either CG**A**Y or GA**A** half-site (Figure 2B). These two parts were mutually rotated and shifted so as to produce a continuous DNA duplex with the target adenines separated by exactly 7 bp and an angle of 49° between the extruding 'arms' (69) (Figure 2C). After repairing the 'break' in the coiled-coil structure, to make the HsdS structure contiguous, the final model (Figure 3) was energy minimized to remove local steric clashes and to introduce favourable interactions between the protein, DNA and the cofactor molecules (for details see Materials and Methods). By this way, we generated the first model of a Type I MTase (a total of 7555 non-hydrogen atoms in the protein components) in which all domains are parts of the same molecule [previous models comprised mixtures of subunits from R–M systems that do not form complexes in the nature—i.e. M.TaqI with HsdS, or the model of M.EcoKI with the TRD of M.HhaI—(19,20,68)]. In our model, the DNA sequence represents a real biological target, and its structure conforms to experimental data (all previous models were arbitrary in this respect, or used non-cognate DNA from other R-M systems). The coordinates of the model are available online from the URL ftp://genesilico.pl/iamb/models/M.EcoR124I/ as well as from the *Nucleic Acids Research* website (Supplementary Data).

We are very aware that our structural model may contain errors, both in details of conformations of individual residues and in the mutual orientation of domains. In particular, the exact structure of the coiled-coil linker region is uncertain, because with computational methods alone it is impossible to determine where the bend is located. The same uncertainty applies to the DNA structure, although the angle between the extruding ends was fitted to the experimental data, in the region of protein–DNA contacts and in the non-specific linker between the half-sites there may be local bends and regions of unwound DNA that are unaccounted for in the present model. This model, however, represents the best fit to the existing experimental data, and will be refined as more data become available. It provides a useful model that can be tested by site-directed mutagenesis and allows us to make suggestions of target sites for such mutagenesis.

Mutations within the central conserved region of HsdS

Weiserova *et al*. (57) described the isolation of a mutation within the central conserved region of HsdS(EcoR124I) produced by misincorporation mutagenesis. This mutation was found to produce an unexpected phenotype (r^-m^+), which was described as 'non-classical'; this was because, 'classically' (63,72,73), mutations in *hsdS* produce an r^-m^- phenotype. Therefore, it was suggested that this mutation might alter interactions between the MTase and the HsdR subunit, but further experimental studies *in vitro* showed that the effect of the mutation was more complex and appeared to alter the ability of the MTase to undergo conformational changes required for DNA-binding (74).

Therefore, further mutations were produced within or close to the central conserved domain of *hsdS* using misincorporation mutagenesis and analysed for their R-M phenotype. The six new substitutions identified were $Ser^{154}Pro(r^+m^+)$, $Arg^{163}Gln(r^+m^+)$, $Glu^{200}Gly(r^+m^+)$, $Leu^{175}Pro(r^{\pm}m^+)$, $Lys^{184}Asn(r^-m^+)$ and $Pro^{218}Ser(r^-m^-)$. Of these mutations, three stand out (Ser^{154}, Arg^{163} and Glu^{200}, all of which produce a wt R–M phenotype), because they are located at either end of the long coiled-coil structure, which links both the TRDs (Figure 4). It seems likely that the region connecting the coiled-coil structure and the TRDs will serve as flexible hinges allowing movement of the two TRDs during the conformational changes associated with DNA binding and bending in the MTase upon DNA binding (19,20,22,70). The $Ser^{154}Pro$ substitution within this hinge might be expected to produce a loss of function because of the introduction of a proline, but in fact this is a phenotypic-silent mutation and careful analysis, using the model structure, suggests that the presence of the proline, in the mutant protein, may in fact stabilize the bend at the end of the helix, but still allow overall flexibility of the coiled-coil structure. A similar argument can be made about the mutation $Glu^{200}Gly$, which is in a loop at the hinge region, but is unlikely to decrease flexibility of the structure.

The change, $Leu^{175}Pro$ (Figure 4) is in the centre of the coiled-coil domain of the central conserved region. Such a change would introduce a distortion in the helix and produce a bend, which would suggest a major structural change. However, the mutation only slightly lowers the ability of the restriction enzyme to cut DNA (it produces an intermediate level of restriction activity), most probably this is because the

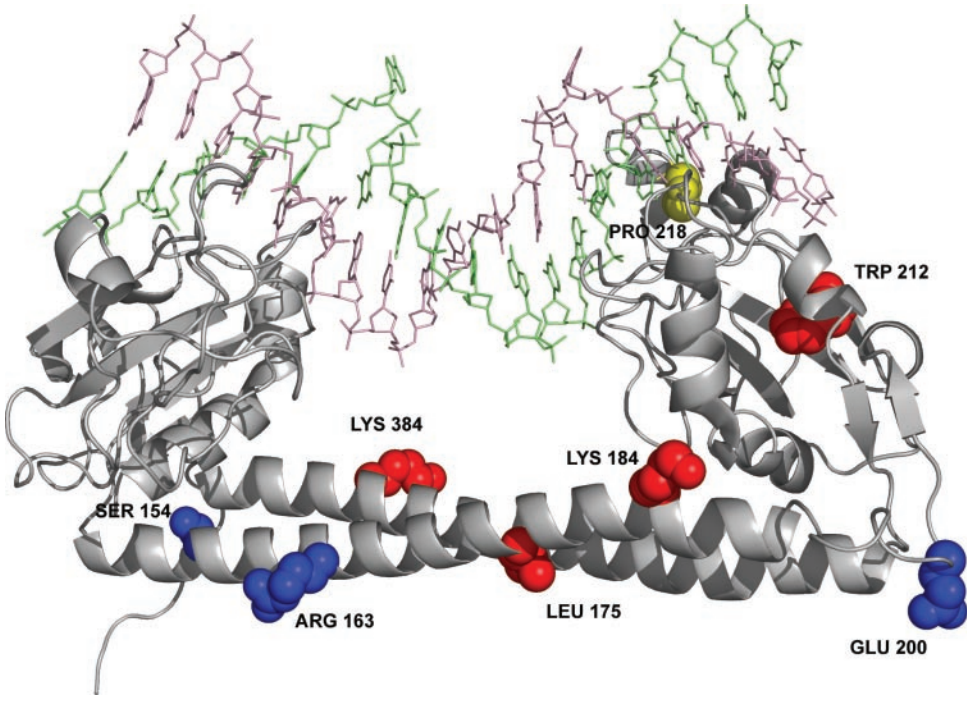

Figure 4. Location of 'non-classical' mutations within HsdS. Classically, mutations within the *hsdS* gene of a Type I R–M enzyme produce a r^-m^- phenotype due to the loss of DNA-binding properties (72,73). The mutations illustrated as red spheres were identified as 'non-classical' because they were identified by a r^-m^+ phenotype, which is thought to be due to alterations to protein–protein interactions (57,74). Blue spheres indicate silent mutations. The yellow amino acid is Pro^{218}, in which a Serine substitution produced an r^-m^- phenotype.

mutation simply reinforces the changes to the orientation of the two TRDs that are required during DNA-binding (22).

The mutation $Lys^{184}Asn$ is another 'non-classical' mutation resulting in an r^-m^+ phenotype. However, the position and 3D localization of this mutation is somewhat surprising. Figure 4 shows that this positively charged amino acid extends from the side of the coiled-coil region. It seems likely that the restriction-deficiency of the mutant protein may reflect loss of protein–protein interactions between HsdS and HsdR and that this mutant may map part of an important region involved in such protein–protein interactions. It became apparent, from studies with the structural model, that an equivalent residue is also present on the parallel helix (Lys^{384}) and we immediately realized that a mutation at this residue should produce a similar phenotype. Therefore, to further test this hypothesis, we have prepared the equivalent $Lys^{384}Asn$ mutation and analysed the phenotype *in vivo*. The result confirmed the symmetry of this situation, as shown in the 3D model, because the phenotype of both the $Lys^{384}Asn$ and the $Lys^{184}Asn$/$Lys^{384}Asn$ double mutant was the predicted 'non-classical' r^-m^+ phenotype.

The mutation $Pro^{218}Ser$ (a restriction and modification deficient phenotype), alters a structurally important residue which lies close to the phosphate backbone of the DNA. The proline appears to serve two functions at this position. Firstly, it allows the protein to wrap around the DNA and secondly, interactions with His^{258}, Phe^{225}, Trp^{226} and Trp^{233} give shape and structural support to this entire region of the protein. By changing this proline to a serine, essential protein–DNA interactions are made impossible from amino acids such as the Lys^{220}, also, the keystone of the interactions between the aromatic ring structures in this region has been removed resulting in a significant re-ordering of this DNA-binding domain. The phenotype of r^-m^- is most likely to be as a result of the loss of ability of the protein to bind DNA and this can be confirmed by analysis of this mutant *in vitro*.

The DNA sequence analysis, associated with identification of the sequence changes produced by the misincorporation mutagenesis, revealed some unexpected information regarding the published DNA sequence of *hsdS* from EcoR124I. Comparison of the published sequence of *ecoR124IhsdS* and the sequence of the wild-type gene determined prior to mutagenesis revealed, in the region 1012–1022, two extra adenines and in the region 1035–1039, one extra adenine, which results in correction of the amino acid sequence from N^{341} Y^{342} F^{343} H^{344} L^{345} to L^{341} F^{342} S^{343} F^{344} (the accession number, X13145, has been updated with this information). These changes have been incorporated into the predicted model, and are used in all descriptions of the protein below.

DISCUSSION

The recent publication of the crystal structure of the DNA-binding subunit of two putative Type I R-M systems (19,20) and the imminent publication of the crystal structure of the HsdM subunit of EcoKI has opened the possibility of modelling the structure of the well-described Type IC R-M DNA-methyltransferase M.EcoR124I from this background information. The availability of information about specific point mutations within the *hsdS* gene of this enzyme [particularly DNA-binding mutants (25,75,76)], the details of modifiable lysines on the surface of the MTase (27) and the details of a related R-M enzyme, with an overlapping DNA-specificity (24), provides a strong background of information on which a model of the structure can be discussed (see below).

In this paper, we describe the first structural model of a complete Type I DNA methyltransferase at the level of atomic detail, which is based on known crystal structures (used as templates), but is constructed entirely from subunits that are known to assemble and cooperate in nature. While we readily acknowledge that this model structure must now be thoroughly tested through further mutagenesis of key amino acids, we have already (within this paper) commenced this process. This work will now be extended in an attempt to identify specific residues that are involved in HsdS–HsdM and HsdS–DNA interactions.

The availability of a model of the M.EcoR124I–DNA complex and the structure-based alignment between TRDs of different MTases now allows us to gain further insight regarding amino acids of HsdS(EcoR124I) potentially involved in sequence-specific DNA recognition. In particular, it is interesting to examine the conservation of amino acids involved in sequence recognition by TRD2 of EcoR124I (specific for the 'RTCG' half-site, i.e. CGAY if read from the opposite strand), TRD1 of StySKI, which recognizes CGAT (77), and M.TaqI, which recognizes TCGA, and whose target in the crystal structure includes the 'TCGAT' sequence, i.e. includes the targets of EcoR124I TRD2 and StySKI TRD1.

Comparison of the protein sequence of HsdS(EcoR124I) and HsdS(StySKI)

Figure 5 shows the numbering system we used to facilitate identification of specific bases in both strands of the EcoR124I target DNA. Figure 1 and Supplementary Figure 3.1 show the identical and similar amino acids between the TRD2 of EcoR124I and the TRD1 of StySKI mapped onto the model structure of HsdS(EcoR124I). A large concave surface is predicted to be involved in DNA recognition and appears to be nearly identical between these two proteins. Interestingly, significant conservation is also observed between the presumed DNA-binding surfaces of TRD2 of HsdS(EcoR124I) and the TRD2 of HsdS(MjaXIP), suggesting a similar mechanism for protein–DNA interactions and perhaps similar DNA specificity. On the other hand, there is little conservation between M.TaqI and the aforementioned Type I enzymes, suggesting that similar sequence specificity for the common 'CGA' trinucleotide may be achieved by different protein–DNA contacts. This is not entirely an unknown

```
5′--G1--A2--A3-N-N-N-N-N-N--R10--T11--C12--G13-3′
3′-*C1-*2T-*3T-N-N-N-N-N-N-*Y10-*A11-*G12-*C13-5′
```

Figure 5. Numbering scheme of nucleosides in the EcoR124I target DNA used in this work. Base pairs in the 13 bp target of EcoR124I are numbered from 1 to 13, with nucleosides in the bottom strand indicated by asterisks. The half-sites recognized by TRD1 and TRD2 comprise bp 1–3 and 10–13, respectively. Unspecified bases are indicated by N. The bold and underlined adenines within the DNA sequence are those that are methylated by the EcoR124I MTase.

phenomenon, and an example of recognition of a similar DNA sequence by different amino acids, attached to a similar structural scaffold (homologous fold), is provided by the remotely related Type II restriction enzymes BglII and BamHI (78).

Closer inspection of the predicted protein–DNA interface reveals a number of candidates for specificity determinants in TRD2 of HsdS(EcoR124I). The conserved residues Arg^{274} and Gln^{295} [homologous to Arg^{176} and Gln^{197} in HsdS(StySKI) and Arg^{305} and Gln^{322} in HsdS(MjaXIP), respectively] make close contacts with the specific bases *G12 and G13 (complementary to *C13) in the recognition sequence *C13–*G12–*A11–*T10 and appear to be the key specificity determinants for this region. There seems to be no specific direct contacts made to the target adenine (*A11), which should be flipped out into the catalytic pocket of the HsdM subunit for methylation, nor to the 'orphaned' T11 base in the other DNA strand. This suggests that the recognition of the 11th base pair may be controlled primarily by the HsdM subunit, or that indirect contact, or recognition of sequence-dependent DNA conformation, plays a role. The current model is, however, of too low an accuracy to provide a conclusive answer. The model, however, suggests an explanation for the difference in specificity between StySKI and EcoR124I with respect to the 10^{th} base pair of the target. Both proteins possess a substitution compared to the HsdS(MjaXIP) template (Figure 1): Ala^{334} in HsdS(EcoR124I) corresponds to (Glu^{236}) in HsdS(StySKI). The backbone of Ala^{334} is positioned in such a way that the longer side chain of a Glu could reach from this position to bases of the R10–*Y10 basepair (A–T or G–C), thereby restricting the specificity to CGAT [as in HsdS(StySKI)], as compared with the more relaxed CGAY in HsdS(EcoR124I). Therefore, the model allows us to predict that a mutation within HsdS(EcoR124I) producing $Ala^{334}Glu$ may well increase the specificity of the EcoR124I enzyme and this prediction can now be readily tested. One of the residues that probably contributes to the recognition of the *Y10 residue is Ser^{275} of HsdS(EcoR124I) [the serine at this position is conserved also in HsdS(StySKI), but not in HsdS(MjaXIP)]. However, this region of the model corresponds to an insertion, whose conformation is uncertain and thereby contacts cannot be predicted in detail.

In contrast to TRD2, TRD1 of HsdS(EcoR124I) is considerably more divergent and exhibits similarities to the corresponding TRD1 of the HsdS(MjaXIP) template only in the protein core, but not on the surface (Figure 1). Nonetheless, the availability of the structural model allows us to predict that Lys^{32} is probably involved in the recognition of the G1–*C1 base pair and Asp^{79} and Arg^{132} could be involved in recognition of either of the A–T pairs in the G1–A2–A3 sequence. Finally, we predict that Lys^{131} makes a contact with the phosphate backbone of the DNA target.

Summarizing, despite the limited accuracy of the model structure, which permits prediction only at the level of amino acid residues, but not individual atoms, we can infer a number of protein–DNA interactions in both TRD1 and TRD2 of HsdS(EcoR124I). This includes specific recognition of all 3 bp of the GAA half-site and 2 bp of the CGAY half-site, as well as being able to infer the molecular basis of the different specificity of HsdS(EcoR124I) and HsdS(StySKI). We also predict that TRD2 of HsdS(MjaXIP) (if this enzyme is found to be active) would recognize CGA, or a related sequence.

Analysis of DNA-binding mutants of TRD2 of HsdS(EcoR124I)

The model structure has allowed us to identify those amino acids that make contact with the DNA substrate; although, somewhat surprisingly, these residues were not amongst those residues identified previously as DNA-binding mutants of TRD2 (25,75,76). Several such mutants were identified as having the appropriate phenotype to indicate a mutation affecting DNA-binding and the presence of a DNA change was confirmed by C-track, one-lane, DNA sequencing (25). It is now interesting to discuss the location of these mutants with reference to the predicted 3D structure of HsdS (although the exact residue changes obtained were not confirmed by the single-track sequencing approach used for this work). Several of these mutations alter an amino acid that is identical between HsdS(EcoR124I) and HsdS(StySKI) (Figure 1), which suggests that these residues are important for DNA sequence reading: Pro^{236}, Asp^{238}, Glu^{248}, Asp^{249}, Ser^{268} and Val^{272}. Pro^{236} is a highly-conserved residue that assumes a critical position at the N-terminus of an α-helix (amino acids 236–240, see Figures 1 and 6). This residue is also located close to the DNA backbone and is preceded by a semi-conserved Ser or Thr residue [Ser^{235} in HsdS(EcoR124I) and Thr^{294} in M.TaqI], which stabilizes the phosphodiester backbone of the T11 nucleoside in the M.EcoR124I model, and the corresponding nucleoside in the M.TaqI structure (23). It is likely that substitution of Pro^{236} changes the local conformation of the polypeptide and perhaps destabilizes interactions with the backbone, leading to propagation of conformational changes and disruption of specific interactions between the protein and the DNA. Asp^{238} is another residue from this region, which fulfils a structural role. Its homologues in the experimentally determined structures [Asp^{265} in HsdS(MjaXIP) and Asn^{297} in M.TaqI] are not involved in any contacts with the DNA, but hydrogen-bond to two regions of the polypeptide backbone, thereby stabilizing the tertiary fold of the TRD. It is very difficult to predict the precise effect of mutation of Asp^{238}, but almost certainly it leads to conformational changes in the TRD that would indirectly affect its ability to interact with the DNA and hence produce the observed r^- phenotype. Glu^{248} and Asp^{249} are confidently predicted not to interact with the DNA directly. Glu^{248} is partially conserved, while Asp^{249} is not [except for HsdS(StySKI); Figure 1]. Comparison of TRD2 in the crystal structures of both putative HsdS subunits and M.TaqI reveals significant conformational variability in the corresponding region. The only common feature is the involvement of Glu in a salt-bridge with an Arg residue, e.g. Glu^{248}-Lys^{237} in HsdS(EcoR124I), Glu^{306}–Arg^{348} in M.TaqI, Glu^{279}–Arg^{268} in HsdS(MjaXIP). A homologue of Arg^{268} of HsdS(MjaXIP) is also conserved in HsdS (EcoR124I) (Arg^{241}) and, therefore, we predict that the primary function of these residues is, again, stabilization of the protein (which may lead to the appropriate DNA contacts by other residues), rather than direct interaction with any other molecule. The role of Asp^{249} is more difficult to assign. It may form a salt-bridge with Lys^{220}, but the latter residue could also rotate away in the opposite direction and bind to the phosphate

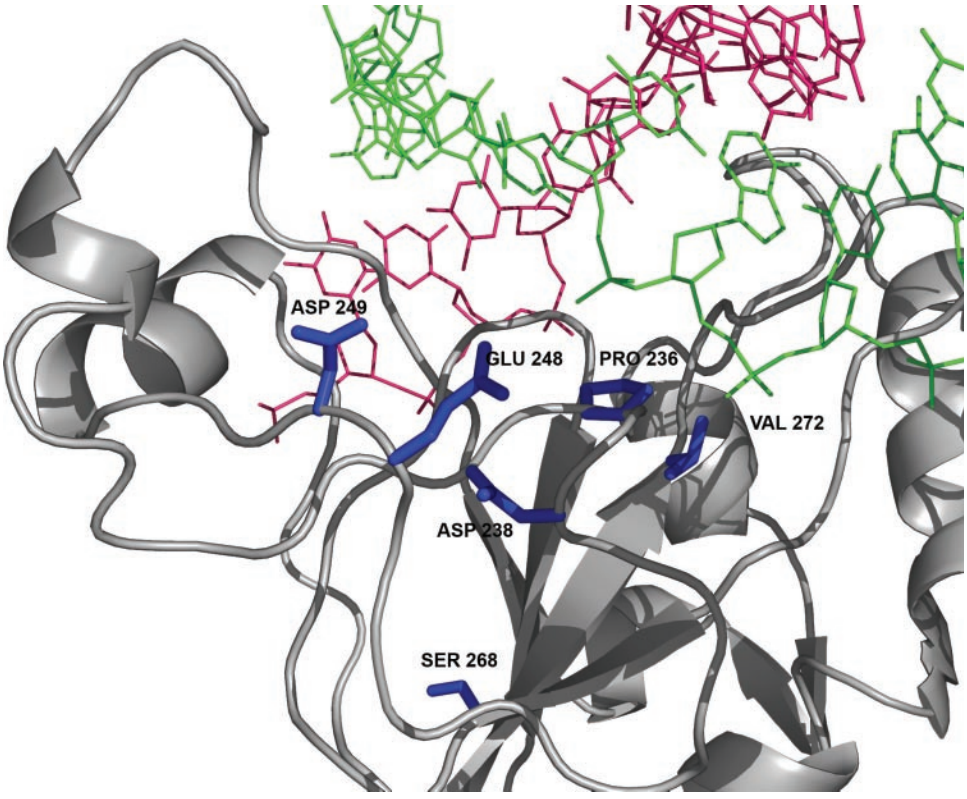

Figure 6. Location of DNA-binding mutants within TRD2 of the HsdS subunit of EcoR124I. The mutations, originally identified as DNA-binding mutants using an *in vivo* competitive complementation assay, which are identical between HsdS(EcoR124I) and HsdS(StySKI), are shown as blue sticks.

backbone of the DNA (the available methodology does not allow confident prediction of conformations of the side chains).

A further series of DNA-binding mutants, produced using the same technique, are not strongly conserved between EcoR124I and MjaXPI: Ser^{230}, Asp^{245}, Ala^{270}, Asp^{279}, Ala^{291}, Val^{256} and Val^{273}. Within the region from residue 268 to 278 lie the final two mutants, Ser^{268} and Val^{272}, that are conserved between HsdS(EcoR124I) and HsdS(StySKI). Val^{272} is not directly involved in DNA recognition, but provides structural support for a loop that interacts with the DNA phosphate backbone (i.e. $^{235}SPK^{237}$ in EcoR124I). Thus the substitution of this residue may significantly perturb protein–DNA contacts. On the other hand, Ser^{268} is located on the opposite side of the TRD and almost certainly participates in stabilization of the protein structure rather than in recognition, as its counterpart in MjaXIP (S^{299}) is involved in hydrogen-bonding to the protein backbone in a neighbouring tight turn. Asp^{279} is located in an insertion that we predict to participate in recognition of the 10th base pair (R–Y) in the HsdS(EcoR124I) target. Since this residue is not conserved even in HsdS(StySKI), where it is substituted by Arg (Figure 1), we speculate that it may be involved either in stabilization of the loop or in making water-mediated contacts with this semi-specific base pair. In addition, this region lies within the sixth beta strand of TRD2, which is also present in the crystal structure of the HsdS(MjaXPI) (19) subunit and this beta strand forms the core of the TRD. Mutations within this region are likely to destabilize the DNA-binding domain. The role of Ser^{230} is difficult to reconcile. It is substituted with Thr in HsdS(StySKI) and its counterpart Glu^{257} in HsdS(MjaXIP) is fully exposed and not involved in any interactions with other amino acids. It is also located on the opposite side from the DNA-binding site. Further mutagenesis at this codon may clarify the role of this amino acid. Asp^{245} may form a salt bridge with Lys^{288}, while Val^{256}, Ala^{270}, Val^{273} and Ala^{291} are all buried in the protein core. All these residues are likely to be important for protein stability, but these point mutations must not totally 'unfold' the protein in a way that prevents protein–protein interactions, because of the nature of the complementation assay used for the screening process. Therefore, the role of these residues must be to stabilize the position of the residues that make DNA contacts rather than stabilize the whole protein structure.

This model and the predictions regarding protein-DNA contacts, from this model, can be used to guide future mutagenesis work that might identify new DNA-binding mutants and either confirm or modify the model as appropriate. Perhaps the most exciting prediction from the model is the mechanism by which the two closely related enzymes EcoR124I and StySKI differ in their recognition of the sequence R(/G)TCG and the possibility of increasing the degree of specificity for EcoR124I by changing the amino acid involved in discriminating the purine at the start of this sequence.

Of some significance, is the observation of two 'mirrored' lysines at either end of the coiled-coil spacer region, which only became apparent through analysis of the 3D model structure and was not predicted from studies with the 2D structure. The fact that we were able to predict the resulting phenotype of this mutation is to our knowledge the first example of a

single point mutation leading to a predicted phenotype for a multisubunit enzyme.

The $Trp^{212}Arg$ mutant of HsdS(EcoR124I)

Mutagenesis carried out previously by Weiserova *et al*. (74) identified an unusual 'non-classical' mutation in *hsdS* (that is a mutation that does not result in a r^- m^- phenotype). They proposed that this mutation ($Trp^{212}Arg$) altered the precise alignment of the HsdM subunits, onto the HsdS subunit, so as to prevent DNA-binding of the MTase through the required conformational change observed upon DNA-binding (22). Figure 4 shows that this mutant is located on the 'elbow' between the central conserved region and TRD2. This suggests that this location is extremely important for HsdS–HsdM interactions and the required flexibility of the MTase subunit. This tryptophan makes edge-to-edge contacts with two other aromatic residues (Phe^{342} and Phe^{344}), which make a stable stacking structure within this 'elbow' region. This introduction of a positive charge in the protein core most likely leads to a local structural rearrangement, within this elbow region, that reduces the ability of HsdS to bind to the HsdM subunit. Therefore, the structural model supports the previous explanation for the effect of this mutation.

Mapping surface-modifiable lysines of the EcoR124I MTase

Further, previously available information, which can be discussed in more detail, using the structural model as a background, is the availability of surface lysines within the MTase. Taylor *et al*. (27) identified surface lysines in the M.EcoR124I methyltransferase and suggested some of the lysines might be involved in DNA binding. Based on the translation of the DNA sequence information that was available at the time, they indicated that the M.EcoR124I MTase contained 109 lysines. However, the changes to the DNA sequence identified in this paper show that the MTase actually contains 111 lysines. Of these, Taylor *et al*. (27) showed that ~18 of all the possible lysines in HsdS were available for chemical modification and ~11 of the lysines in HsdM were also available for surface modification. Taylor *et al*. (27) were also able to show that the presence of DNA significantly reduced the rate of lysine modification, indicating protection of certain surface lysines by the DNA substrate. With the available predicted structure of M.EcoR124I it is now possible to examine the location of the accessible lysines more closely.

Taylor *et al*. (27) describe at least six highly modifiable lysines in HsdS [Lys^{197}, Lys^{204}, Lys^{211} Lys^{262}, Lys^{298} and Lys^{328}; the numbering system used here, and in the pdb file, is that used for the model structure and includes the *N*-formyl methionine at the *N*-terminus of HsdS, which Taylor *et al*. (27) did not include]. They suggested that at least four of these lysines lead to a loss of DNA binding when they were modified. The location of Lys^{197}, Lys^{204} and Lys^{211} is shown in Supplementary Figure 3.2a and, interestingly, they appear in Supplementary Data to be surface accessible and unlikely to be protected by DNA binding or by HsdS–HsdM protein–protein interactions. The structural model shows that there are four lysines (in red in Supplementary Figure 3.2b) that appear to be covered by the DNA—Lys^{94}, Lys^{99}, Lys^{131}, which could not have been mapped by Taylor *et al*. (27) (because their limited proteolysis did not separate any peptides covering TRD1) and Lys^{298}, which was identified by Taylor *et al*. (27) as one of the highly modifiable lysines in the absence of DNA. In addition, as discussed earlier Lys^{32}, which also could not have been mapped by Taylor *et al*. (27), is probably involved in contacting and recognition of the DNA sequence.

Of the other highly modifiable lysines perhaps the most unexpected observation of Taylor *et al*. (27) involved Lys^{328}, because the adjacent residue, Lys^{327} was not significantly (10-fold lower) modified, while Lys^{328} was highly modifiable. These two lysines were found to be arranged, in the model structure, in such a way that one is relatively exposed on the surface (Lys^{328}); although partially covered by the DNA, while Lys^{327} is buried into the HsdS subunit and is inaccessible.

The solvent accessibility surface area (SASA) of the surface lysines within the structural model of both HsdS(R124I) and M.EcoR124I+DNA was analysed using *in silico* techniques (79). The predicted availability was compared to the experimental data available from Taylor *et al*. (27) and was found to be in general agreement (Supplementary Data and Supplementary Table 1). However, caution has to be exercised not to over-interpret the values of solvent-exposure calculated based on the model, as we observed that slight variations of the modelling procedure can result in dramatic changes of exposure/burial of side-chains close to the protein–solvent interface. For instance, Lys^{197} was found to be buried in the model, which does not fit the experimental data. This residue and its neighbours were all flexible during the minimization procedure used to produce the model structure. Thus, we refined the model based on the SASA information by rotating the ϵ group of Lys^{197} into the solvent and adjusting the conformation of its neighbours by energy minimization. This refinement of the model did not alter the conformation of the main chain, only allowed a better fit of the side-chain conformation to experimental data.

Differences between EcoR124I and EcoR124II

When the DNA sequence of the *hsdS* genes of EcoR124I and EcoR124II were compared, it was found that EcoR124II possessed an extra 12 bp repeat within the central conserved region [three repeats compared with two repeats in EcoR124I (80)]. According to the structural model of HsdS (EcoR124I), the insertion of one such repeat creates an additional tetrapeptide within the coiled-coil region. This tetrapeptide is predicted to form an additional turn of a helix, which could extend the length of the coiled coil by ~5.6 Å. This in turn would lead to increased spacing between the two TRDs, and would require 'stretching' of the DNA between the two target adenines. This 'stretching' could be compensated by insertion of an additional base pair within the non-specific region of the recognition target DNA site, or increased bending, or a combination of the two. In agreement with this model, the recognition sequences of EcoR124I and EcoR124II differ by one extra nucleotide in the non-specific spacer.

In conclusion, the predicted protein structure of the M.EcoR124I enzyme (comprising three subunits) plus DNA substrate and cofactor AdoMet provides a model that explains a large body of experimental data and suggests the mechanism of interactions between the components of the complex. In the

absence of a crystal structure for any functionally competent form of a Type I enzyme, the model of M.EcoR124I MTase presents the most comprehensive and biologically relevant structural model to date and will guide an *in silico* 'assembly' of the entire endonuclease and future mutagenesis of the MTase.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank David Dryden for discussions and comments on the manuscript. We are grateful to the authors of the EcoKI HsdM structure (K. R. Rajashankar, R. Kniewel and C. D. Lima) for making the coordinates available in the Protein Data Bank prior to publication. A.B. was supported by an IBBS Studentship from the University of Portsmouth, and A.O. and J.M.B. were supported by the Polish Ministry of Scientific Research and Information Technology (grant PBZ-KBN-088/P04/2003). M.F. was supported by the NIH (Fogarty International Center grant R03 TW007163-01). J.M.B. was also supported by the Young Investigator award from EMBO and HHMI. E.S., S.V. and M.W. were supported by fund provided by The Grant Agency of the Czech Republic (204/03/1011) and by Institutional Research Concept No.AV0Z5020903. K.F. and J.M.B. would like to thank Darek Gorecki for initiating their contact and therefore prompting this collaborative work. Funding to pay the Open Access publication charges for this article was provided by K.F.

Conflict of interest statement. None declared.

REFERENCES

1. Sistla,S. and Rao,D.N. (2004) S-Adenosyl-L-methionine-dependent restriction enzymes. *Crit. Rev. Biochem. Mol. Biol.*, **39**, 1–19.
2. Loenen,W.A. (2003) Tracking EcoKI and DNA fifty years on: a golden story full of surprises. *Nucleic Acids Res.*, **31**, 7059–7069.
3. Murray,N.E. (2000) Type I restriction systems: sophisticated molecular machines (a legacy of Bertani and Weigle). *Microbiol. Mol. Biol. Rev.*, **64**, 412–434.
4. Taylor,I., Patel,J., Firman,K. and Kneale,G.G. (1992) Purification and biochemical characterisation of the EcoR124 modification methylase. *Nucleic Acids Res.*, **20**, 179–186.
5. Dryden,D.T.F., Cooper,L.P. and Murray,N.E. (1993) Purification and characterisation of the methyltransferase from the type I restriction and modification system of *Escherichia coli* K12. *J. Biol. Chem.*, **268**, 13228–13236.
6. Bickle,T.A. and Krüger,D.H. (1993) Biology of DNA restriction. *Microbiol. Rev.*, **57**, 434–450.
7. Wilson,G.G. (1991) Organisation of restriction–modification systems. *Nucleic Acids Res.*, **19**, 2539–2566.
8. Redaschi,N. and Bickle,T.A. (1996) DNA restriction and modification systems. In Neidhardt,F.C., Curtiss,R.III, Ingraham,J.L., Lin,E.C.C., Low,K.B., Magasanik,B. and Reznikoff,W.S. (eds), *Escherichia coli and Salmonella: Cellular and Molecular Biology*. 2nd edn. American Society for Microbiology, Washington, DC, pp. 773–781.
9. Titheradge,A.J.B., King,J., Ryu,J. and Murray,N.E. (2001) Families of restriction enzymes: an analysis prompted by molecular and genetic data for type ID restriction and modification systems. *Nucleic Acids Res.*, **29**, 4195–4205.
10. Chin,V., Valinluck,V., Magaki,S. and Ryu,J. (2004) KpnBI is the prototype of a new family (IE) of bacterial type I restriction-modification system. *Nucleic Acids Res.*, **32**, e138.
11. Cowan,G.M., Gann,A.A.F. and Murray,N.E. (1989) Conservation of complex DNA recognition domains between families of restriction enzymes. *Cell*, **56**, 103–109.
12. Kannan,P., Cowan,G.M., Daniel,A.S., Gann,A.A.F. and Murray,N.E. (1989) Conservation of organisation in the specificity polypeptides of two families of type I restriction enzymes. *J. Mol. Biol.*, **209**, 335–344.
13. Murray,N.E., Gough,J.A., Suri,B. and Bickle,T.A. (1982) Structural homologies among type I restriction and modification systems. *EMBO J.*, **1**, 535–539.
14. Gann,A.F.F., Campbell,A.J.B., Collins,J.F., Coulson,A.F.W. and Murray,N.E. (1987) Reassortment of DNA recognition domains and the evolution of new specificities. *Mol. Microbiol.*, **1**, 13–22.
15. Fuller-Pace,F.V., Bullas,L.R., Delius,H. and Murray,N.E. (1984) Genetic recombination can generate altered restriction specificity. *Proc. Natl Acad. Sci. USA*, **81**, 6095–6099.
16. Kneale,G.G. (1994) A symmetrical model for the domain structure of type I DNA methyltransferases. *J. Mol. Biol.*, **243**, 1–5.
17. MacWilliams,M.P. and Bickle,T.A. (1996) Generation of new DNA binding specificity by truncation of the type IC EcoDXXI *hsdS* gene. *EMBO J.*, **15**, 4775–4783.
18. Abadjieva,A., Patel,J., Webb,M., Zinkevich,V. and Firman,K. (1993) A deletion mutant of the type IC restriction endonuclease EcoR124I expressing a novel DNA specificity. *Nucleic Acids Res.*, **21**, 4435–4443.
19. Kim,J.-S., DeGiovanni,A., Jancarik,J., Adams,P.D., Yokota,H., Kim,R. and Kim,S.-H. (2005) Crystal structure of DNA sequence specificity subunit of a type I restriction–modification enzyme and its functional implications. *Proc. Natl Acad. Sci. USA*, **102**, 3248–3253.
20. Calisto,B.M., Pich,O.Q., Pinol,J., Fita,I., Querol,E. and Carpena,X. (2005) Crystal structure of a putative Type I restriction–modification S subunit from *Mycoplasma genitalium*. *J. Mol. Biol.*, **351**, 749–762.
21. Labahn,J., Granzin,J., Schluckebier,G., Robinson,D.P., Jack,W.E., Schildkraut,I. and Saenger,W. (1994) Three-dimensional structure of the adenine-specific DNA methyltransferase M·*Taq*I in complex with the cofactor *S*-adenosyl methionine. *Proc. Natl Acad. Sci. USA*,, **91**, 10957–10961.
22. Taylor,I.A., Davis,K.G., Watts,D. and Kneale,G.G. (1994) DNA binding induces a major structural transition in a type I methyltransferase. *EMBO J.*, **13**, 5772–5778.
23. Goedecke,K., Pignot,M., Goody,R.S., Scheidig,A.J. and Weinhold,E. (2001) Structure of the N6-adenine DNA methyltransferase M.TaqI in complex with DNA and a cofactor analog. *Nature Struct. Biol.*, **8**, 121–125.
24. Thorpe,P.H., Ternent,D. and Murray,N.E. (1997) The specificity of *Sty*KI, a type I restriction enzyme, implies a structure with rotational symmetry. *Nucleic Acids Res.*, **25**, 1694–1700.
25. Patel,J. (1992) The cloning, expression and mutagenesis of the recognition subunit, HsdS, from the EcoR124 R-M system. PhD Thesis, University of Portsmouth, Portsmouth, UK.
26. Taylor,I.A. and Webb,M. (2001) Chemical modification of lysine by reductive methylation. A probe for residues involved in DNA binding. *Methods Mol. Biol.*, **148**, 301–314.
27. Taylor,I.A., Webb,M. and Kneale,G.G. (1996) Surface labelling of the type I methyltransferase M.EcoR124I reveals lysine residues critical for DNA binding. *J. Mol. Biol.*, **258**, 62–73.
28. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
29. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
30. Kurowski,M.A. and Bujnicki,J.M. (2003) GeneSilico protein structure prediction meta-server. *Nucleic Acids Res.*, **31**, 3305–3307.
31. McGuffin,L.J., Bryson,K. and Jones,D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–405.
32. Rost,B., Yachdav,G. and Liu,J. (2004) The PredictProtein server. *Nucleic Acids Res.*, **32**, W321–W326.
33. Ouali,M. and King,R.D. (2000) Cascaded multiple classifiers for secondary structure prediction. *Protein Sci.*, **9**, 1162–1176.
34. Adamczak,R., Porollo,A. and Meller,J. (2004) Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins*, **56**, 753–767.

35. Cuff,J.A. and Barton,G.J. (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*, **40**, 502–511.
36. Meiler,J. and Baker,D. (2003) Coupled prediction of protein secondary and tertiary structure. *Proc. Natl Acad. Sci. USA*, **100**, 12105–12110.
37. Karplus,K., Karchin,R., Draper,J., Casper,J., Mandel-Gutfreund,Y., Diekhans,M. and Hughey,R. (2003) Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins*, **53**, 491–496.
38. Rychlewski,L., Jaroszewski,L., Li,W. and Godzik,A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.*, **9**, 232–241.
39. Kelley,L.A., MacCallum,R.M. and Sternberg,M.J. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.*, **299**, 499–520.
40. Fischer,D. (2000) Hybrid fold recognition: combining sequence derived properties with evolutionary information. In Altman,R.B., Dunker,A.K., Hunter,L., Lauderdale,K. and Klein,T.E. (eds), *Pacific Symposium on Biocomputing*. World Scientific Publishing, Hawaii, pp. 119–130.
41. Shi,J., Blundell,T.L. and Mizuguchi,K. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.*, **310**, 243–257.
42. Jones,D.T. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.*, **287**, 797–815.
43. Zhou,H. and Zhou,Y. (2004) Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. *Proteins*, **55**, 1005–1013.
44. Lundstrom,J., Rychlewski,L., Bujnicki,J.M. and Elofsson,A. (2001) Pcons: a neural-network-based consensus predictor that improves fold recognition. *Protein Sci.*, **10**, 2354–2362.
45. Kosinski,J., Cymerman,I.A., Feder,M., Kurowski,M.A., Sasin,J.M. and Bujnicki,J.M. (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**, 369–379.
46. Luthy,R., Bowie,J.U. and Eisenberg,D. (1992) Assessment of protein models with three-dimensional profiles. *Nature*, **356**, 83–85.
47. Sasin,J.M. and Bujnicki,J.M. (2004) COLORADO3D, a web server for the visual analysis of protein structures. *Nucleic Acids Res.*, **32**, W586–W589.
48. Labahn,J., Granzin,J., Schluckebier,G., Robinson,D.P., Jack,W.E., Schildkraut,I. and Saenger,W. (1994) Three-dimensional structure of the adenine-specific DNA methyltransferase M.TaqI in complex with the cofactor *S*-adenosylmethionine. *Proc. Natl Acad. Sci. USA*, **91**, 10957–10961.
49. Kim,J.S., DeGiovanni,A., Jancarik,J., Adams,P.D., Yokota,H., Kim,R. and Kim,S.H. (2005) Crystal structure of DNA sequence specificity subunit of a type I restriction-modification enzyme and its functional implications. *Proc. Natl Acad. Sci. USA*, **102**, 3248–3253.
50. Bayly,C.I., Cieplak,P., Cornell,W.D. and Kollman,P.A. (1993) A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges—the RESP model. *J. Phys. Chem.*, **97**, 10269–10280.
51. Frisch,M.J., Trucks,G.W., Schlegel,H.B., Scuseria,G.E., Robb,M.A., Cheeseman,J.R., Montgomery,J.J.A., Vreven,T., Kudin,K.N., Burant,J.C. *et al.* (2004) Gaussian 03. Revision C.02 edn. Gaussian, Inc, Wallingford.
52. Case,D.A., Darden,T.A., Cheatham,T.E.III, Simmerling,C.L., Wang,J., Duke,R.E., Luo,R., Merz,K.M., Wang,B., Pearlman,D.A. *et al.* (2004) AMBER 8. University of California, San Francisco.
53. Hawkins,G.D., Cramer,C.J. and Truhlar,D.G. (1995) Pairwise solute descreening of solute charges from a dielectric medium. *Chem. Phys. Lett.*, **246**, 122–129.
54. Hawkins,G.D., Cramer,C.J. and Truhlar,D.G. (1996) Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *J. Phys. Chem.*, **100**, 19824–19839.
55. Wang,J., Cieplak,P. and Kollman,P.A. (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.*, **21**, 1049–1074.
56. Tsui,V. and Case,D.A. (2000) Theory and applications of the generalized Born solvation model in macromolecular simulations. *Biopolymers*, **56**, 275–291.
57. Weiserova,M. and Firman,K. (1998) Isolation of a non-classical mutant of the DNA recognition subunit of the type I restriction endonuclease R.EcoR124I. *Biol. Chem.*, **379**, 585–589.
58. Patel,J., Taylor,I., Dutta,C.F., Kneale,G.G. and Firman,K. (1992) High-level expression of the cloned genes encoding the subunits of and the intact DNA methyltransferase,M.EcoR124. *Gene*, **112**, 21–27.
59. Sambrook,J.C., Fritsch,E.F. and Maniatis,T. (1989) *Molecular Cloning: A Laboratory Manual*. 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
60. Jacob,F. and Wollman,E.L. (1954) Etude genetique d'un bacteriophage tempere d'*Escherichia coli*: le systeme genetique du bacteriophage lambda. *Ann. Inst. Pasteur.*, **87**, 653–673.
61. Yanisch-Perron,C., Vieira,J. and Messing,J. (1985) Improved M13 phage cloning vectors and host strains: nucleotide sequence of the M13mp18 and pUC19 vectors. *Gene*, **33**, 103–119.
62. Colson,C., Glover,S.W., Symonds,N. and Stacey,K.A. (1965) The location of the genes for host-controlled modification and restriction in *Escherichia coli*. *Genetics*, **52**, 1043–1050.
63. Hubácek,J. and Glover,S.W. (1970) Complementation analysis of temperature-sensitive host specificity mutations in *Escherichia coli*. *J. Mol. Biol.*, **50**, 111–127.
64. Hedges,R.W. and Datta,N. (1972) R124 an fi^+ *R*-factor of a new compatibility class. *J. Gen. Microbiol.*, **71**, 403–405.
65. Hughes,S.G. (1977) Studies of plasmid encoded restriction and modification systems. PhD Thesis, University of Edinburgh, UK.
66. Cymerman,I.A., Feder,M., Pawlowski,M., Kurowski,M.A. and Bujnicki,J.M. (2004) Computational methods for protein structure prediction and fold recognition. In Bujnicki,J.M. (ed.), *Practical Bioinformatics*. Springer-Verlag, Berlin, Vol. 15, pp. 1–21.
67. Kosinski,J., Gajda,M.J., Cymerman,I.A., Kurowski,M.A., Pawlowski,M., Boniecki,M., Obarska,A. and Papaj,G., Sroczynska-Obuchowicz,P., Tkaczuk,K.L. *et al.* (2005) FRankenstein becomes a cyborg: the automatic recombination and realignment of fold recognition models in CASP6. *Proteins*, **61**, 106–113.
68. Dryden,D.T.F., Sturrock,S.S. and Winter,M. (1995) Structural modelling of a type I DNA methyltransferase. *Nature Struct. Biol.*, **2**, 632–635.
69. van Noort,J., van der Heijden,T., Dutta,C.F., Firman,K. and Dekker,C. (2004) Initiation of translocation by Type I restriction–modification enzymes is associated with a short DNA extrusion. *Nucleic Acids Res.*, **32**, 6540–6547.
70. Su,T.J., Tock,M.R., Egelhaaf,S.U., Poon,W.C. and Dryden,D.T. (2005) DNA bending by M.EcoKI methyltransferase is coupled to nucleotide flipping. *Nucleic Acids Res.*, **33**, 3235–3244.
71. Walkinshaw,M.D., Taylor,P., Sturrock,S.S., Atanasiu,C., Berge,T., Henderson,R.M., Edwardson,J.M. and Dryden,D.T.F. (2002) Structure of Ocr from bacteriophage T7, a protein that mimics B-form DNA. *Mol. Cell*, **9**, 187–194.
72. Glover,S.W. and Colson,C. (1969) Genetics of host-controlled restriction and modification in *Escherichia coli*. *Genet. Res. (Camb.)*, **13**, 227–240.
73. Glover,S.W. (1970) Functional analysis of host-specificity mutants in *Escherichia coli*. *Genet. Res. (Camb.)*, **15**, 237–250.
74. Weiserova,M., Dutta,C.F. and Firman,K. (2000) A novel mutant of the type I restriction-modification enzyme EcoR124I is altered at a key stage in the subunit assembly pathway. *J. Mol. Biol.*, **304**, 301–310.
75. Patel,J. and Firman,K. (1992) The domain structure of the DNA specificity subunit of Type I restriction endonucleases. I. Cloning, mutagenesis and over-production of the EcoR124 DNA methyltransferase. In Balla,E., Berensci,G. and Szentirmai,A. (eds), *Proceedings of the 11th European Meeting on Genetic Transformation*. Intercept Ltd, Budapest, pp. 179–187.
76. Abadjieva,A., Zinkevich,V., Weiserová,M., Patel,J. and Firman,K. (1992) The domain structure of the DNA specificity subunit of type I restriction endonucleases. II. Mutations affecting subunit assembly. In Balla,E., Berensci,G. and Szentirmai,A. (eds), *Proceedings of the 11th European Conference on Genetic Transformation*. Intercept Ltd, Budapest, pp. 189–196.
77. Thorpe,P.H. (1995) The DNA specificity of type I restriction and modification enzymes. PhD Thesis, University of Edinburgh, Edinburgh.
78. Lukacs,C.M., Kucera,R., Schildkraut,I. and Aggarwal,A.K. (2000) Understanding the immutability of restriction enzymes: crystal structure

of BglII and its DNA substrate at 1.5 A resolution. *Nature Struct. Biol.*, **7**, 134–140.
79. Fraczkiewicz,R. and Braun,W. (1998) Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *J. Comput. Chem.*, **19**, 319–333.
80. Price,C., Lingner,J., Bickle,T.A., Firman,K. and Glover,S.W. (1989) Basis for changes in DNA recognition by the EcoR124 and EcoR124/3 type I DNA restriction and modification enzymes. *J. Mol. Biol.*, **205**, 115–125.
81. Xiaodong,Cheng and Blumenthal,R.M. (1996) Finding a basis for flipping bases. *Structure*, **4**, 639–645.
82. Su,T.-J., Connolly,B.A., Darlington,C., Mallin,R. and Dryden,D.T.F. (2004) Unusual 2-aminopurine fluorescence from a complex of DNA and the EcoKI methyltransferase. *Nucleic Acids Res.*, **32**, 2223–2230.