Structural bioinformatics

I-Ssp6803I: the first homing endonuclease from the PD-(D/E)XK superfamily exhibits an unusual mode of DNA recognition

Jerzy Orlowski, Michal Boniecki and Janusz M. Bujnicki*

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

Received on November 14, 2006; revised on January 7, 2007; accepted on January 11, 2007

ABSTRACT

Motivation: Restriction endonucleases (REases) and homing endonucleases (HEases) are biotechnologically important enzymes. Nearly all structurally characterized REases belong to the PD-(D/E)XK superfamily of nucleases, while most HEases belong to an unrelated LAGLIDADG superfamily. These two protein folds are typically associated with very different modes of protein-DNA recognition, consistent with the different mechanisms of action required to achieve high specificity. REases recognize short DNA sequences using multiple contacts per base pair, while HEases recognize very long sites using a few contacts per base pair, thereby allowing for partial degeneracy of the target sequence. Thus far, neither REases with the LAGLIDADG fold, nor HEases with the PD-(D/E)XK fold, have been found.

Results: Using protein fold recognition, we have identified the first member of the PD-(D/E)XK superfamily among homing endonucleases, a cyanobacterial enzyme I-Ssp6803I. We present a model of the I-Ssp6803I-DNA complex based on the structure of Type II restriction endonuclease R.BglII and predict the active site and residues involved in specific DNA sequence recognition by I-Ssp6803I. Our finding reveals a new unexpected evolutionary link between HEases and REases and suggests how PD-(D/E)XK nucleases may develop a ‘HEase-like’ way of interacting with the extended DNA sequence. This in turn may be exploited to study the evolution of DNA sequence specificity and to engineer nucleases with new substrate specificities.

Contact: iamb@genesilico.pl

1 INTRODUCTION

Homing endonucleases (HEases) and restriction endonucleases (REases) are sequence-specific deoxyribonucleases. In the cells, they function as selfish genetic elements that are transmitted by horizontal gene transfer and promote their own survival and proliferation by cleaving the foreign DNA and either destroying it or inducing DNA repair by recombination, which in turn increases the chance for the duplication of their genes. When isolated and purified, they can be used in vitro as reagents and indispensable tools in the recombinant DNA protocols (comprehensive collections of reviews appeared recently in books dedicated to both types of nucleases; Belfort et al., 2005; Pingoud, 2004).

Structural analyses revealed that HEases originated from at least three unrelated nuclease superfamilies: LAGLIDADG, GIY-YIG, and HNH (Stoddard, 2005). Within these superfamilies, they exhibit easily recognizable sequence conservation. On the other hand, REases typically exhibit little or no significant sequence similarity. Crystallographic studies revealed a number of structures that all contained the catalytic domain from the PD-(D/E)XK superfamily, indicating extreme sequence divergence from a common ancestor (Bujnicki, 2003). However, bioinformatics analyses supported by experiment, and recently validated by X-ray crystallography (at least for one case), revealed that some REases may belong to PLD, GIY-YIG and HNH superfamilies (Aravind et al., 2000; Bujnicki et al., 2001; Grazulis et al., 2005; Saprauanskas et al., 2000). Thus, both REases and HEases are of polyphyletic origin, and moreover some of them share common ancestors from the GIY-YIG and HNH superfamilies. Nonetheless, no functional overlap has been found in the major LAGLIDADG and PD-(D/E)XK superfamilies: neither REases have been identified to exhibit the I-Ssp6803I fold, nor have HEases been found in the PD-(D/E)XK superfamily.

I-Ssp6803I is a HEase encoded by a self-splicing intron in the tRNA\[^{Mtr}\] gene of Synechocystis PCC6803, which has been proposed to have moved laterally within the Cyanobacteria (Biniszkiewicz et al., 1994; Bonocora and Shub, 2001). Its sequence appears to be unrelated to known HEases or, in fact, to any functionally characterized protein in the database (Stoddard, 2005). Thus, we carried out bioinformatics analysis to predict the structure of I-Ssp6803I and to classify it into one of the existing superfamilies.

2 METHODS

Sequence searches of the non-redundant (nr) database and putative translations of unfinished genomic and metagenomic DNA sequences were carried out at the NCBI using PSI-BLAST (Altschul et al., 1997). Secondary structure prediction and tertiary fold-recognition (FR) was carried out via the GeneSilico meta-server gateway (Kuworski and Bujnicki, 2003). FR alignments returned by original servers were compared and ranked by the PCONS consensus server (Lundstrom et al., 2001). Because of space constraints in this article, we cannot cite...
the articles describing all methods included in the metaserver: therefore, readers are requested to familiarize themselves with the complete list of original servers at the website http://genesilico.pl/meta/.

Alignments between the I-Ssp6803I sequence and the consensus template structure identified by PCONS (Hjc, 1gef) were used as a starting point for modeling of the I-Ssp6803I monomer using the ‘FRankensteen’s monster’ approach (Kosinski et al., 2003; Kosinski et al., 2005b). For template-based modeling, we used MODELLER (Fiser and Sali, 2003), and for de novo modeling of loops, we used REFINER, a method for folding simulations in real space that uses reduced representation of protein structure and a Monte Carlo sampling scheme. REFINER was found to perform very well in adding missing parts to incomplete models of protein structure (Boniecki et al., 2003). For model evaluation, we used PROQ (Wallner and Elofsson, 2006) and the MetaMQAP method recently developed in our laboratory (Marcin Pawlowski, Ryszard Matlak, J.M.B., manuscript in preparation, server available at https://genesilico.pl/toolkit/).

Model building, evaluation, realignment in poorly scored regions and merging of best-scored fragments was reiterated until all regions in the protein core obtained acceptable MetaMQAP score (predicted deviation <3Å) or their score could not be significantly improved by manipulations of the alignment. The variable loops and the C-terminal part of the protein (aa 105–150), which exhibit no conservation even with the closest homologs of I-Ssp6803I and showed structural variability between the templates, were modeled de novo using REFINER. According to PROQ, the optimized model of the monomer obtained predicted LGscore score: 1.61 and predicted MaxSub score: 0.142, which indicates a ‘fairly good model’, thus supporting our PD-(D/E)XK fold prediction for I-Ssp6803I despite low scores from PCONS.

To construct a model of the I-Ssp6803I dimer in complex with the DNA, we have repeated the modeling, now with the R.BglI dimer–DNA complex (1dmu) as an additional template. This allowed for better modeling of the protein–protein interactions (R.BglI dimersize in a different manner than Hjc) and provided guidance for the positioning of loops with respect to the DNA. The DNA sequence from the R.BglI complex (5’-ATCCGCTTAATAGCGAT-3’) was ‘mutated’ to the I-Ssp6803I target (5’-TCGGGCTCATAACCCGA-3’) using X3DNA (Lu and Olson, 2003). Since REFINER cannot reliably model protein–DNA interactions, amino acid residues at the protein–DNA interface were optimized by NAMD (Phillips et al., 2005).

3 RESULTS

Sequence searches of the non-redundant (nr) database and putative translation products of environmental DNA samples from metagenomic projects revealed no statistically significant similarity between sequence of I-Ssp6803I and any other proteins; the most similar sequence of an uncharacterized protein from Bacillus thuringiensis (GI:49478649) was reported with e-value 0.2. Nonetheless, careful analysis of pairwise alignments between I-Ssp6803I and this and other sequences reported below the threshold of significance revealed conservation of a motif D-X9-11-Q-X-K (where X indicates any amino acid), which suggested that I-Ssp6803I may belong to the PD-(D/E)XK superfamily (Fig. 1). Indeed, FR analysis carried out via the GeneSilico metaserver revealed relationship of I-Ssp6803I to proteins with the PD-(D/E)XK fold (top 5 alignments in the PCONS ranking, with scores 0.2683–0.2047,

---

Fig. 1. Sequence alignment of I-Ssp6803I and its predicted homologs (borderline BLAST hits that also matched the PD-(D/E)XK fold according to the metaserver), together with the conserved elements of PD-(D/E)XK nucleases used as a scaffold for modeling. The predicted active site residues are indicated by asterisks. Secondary structures of the final model and of the template structures (1gef and 1dmu) are shown as arrows and cylinders. Sequences are indicated by species names or PDB accession numbers. Symbols: Env—environmental sample, T7g—T7 phage, Ana—Anaeba variabilis, Bac—Bacillus thuringiensis, Bur—Burkholderia vietnamiensis, Aci—Acidiphilium cryptum, She—Shewanella sp., Rho—Rhodopirellula baltica. Regions present in templates but structurally divergent and hence omitted from template-based modeling are indicated by the number of corresponding residues. Thus, regions of I-Ssp6803I sequence that were modeled de novo are indicated by the lack of the alignment to any template structure/sequence.
The first homing endonuclease from the PD-(D/E)XK superfamily exhibits an unusual mode of DNA recognition

while the next unrelated fold was reported at position 6 with much lower score of 0.1478. Interestingly, also the sequences reported in database searches with low e-values, but containing the characteristic motif, were found to match the PD-(D/E)XK fold, often with higher PCONS scores (e.g. 0.4762 for the environmental sequence, GI:59985358). The highest-ranked template was always the Holliday junction resolvase Hjc (1gef) (Nishino et al., 2001). Thus, our results suggest that I-Ssp6803I is the first HEase with the PD-(D/E)XK fold characteristic for REases, and that its residues D36, Q49 and K51 form the nuclease active site. However, it must be emphasized that the PCONS scores reported above are below the recommended threshold of 1.5 and, therefore, the fold prediction has to be supported by modeling and analysis of the model at the three-dimensional level.

Most REases recognize very short sites (4–8 base pairs) with very high specificity (sites that differ by one base pair are cleaved very poorly or not at all (Pingoud et al., 2005), while most HEases recognize very long sequences but do cleave variants that differ even by several base pairs (Stoddard, 2005). These two different modes of DNA recognition require different positions of polypeptide loops that form specific interactions with the substrate. REases usually employ multiple loops that encircle the DNA and form a very complicated network of interactions, both at the side of the major and the minor groove, providing redundant contacts to a few bases, while REases form very extended binding sites, often with only one loop extending along the major groove of the DNA. Thus, it is not surprising that REases and HEases usually utilize different folds as preferred scaffolds: the PD-(D/E)XK fold with its β-strands positioned perpendicularly to the DNA, which facilitates the ‘REase’ mode of recognition (Pingoud et al., 2005), and the LAGLIDADG fold with β-strands positioned along the axis of the DNA (Stoddard, 2005), respectively.

I-Ssp6803I is very unusual as a HEase that recognizes a long site, yet exhibits the PD-(D/E)XK fold. To provide insight into the possible mode of its action, we constructed a model of I-Ssp6803I dimer in complex with the DNA (see Methods for details). The quaternary structure prediction was based on the empirical finding that enzymes from the PD-(D/E)XK superfamily that produce similar cleavage patterns, usually dimerize in a very similar way (Pingoud et al., 2005). Thus, we decided to model the interactions between I-Ssp6803I monomers and conformation of the target DNA based on the structure of REase R.BglII (1dmu) (Newman et al., 1998), which exhibits the same pattern of cleavage (3 nt 3’ overhangs) as I-Ssp6803I. Our prediction of protein–protein interactions in the I-Ssp6803I dimer is supported by the fact that the PROQ score of the dimer increased from 1.61 to 2.0 (predicted LGscore, still within the range of a ‘fairly good model’), owing mostly to the formation of contacts between hydrophobic residues that were exposed on the surface of the monomer (data not shown). It must be emphasized that PROQ and other methods for model evaluation cannot take into account the contacts between amino acid residues and DNA or metal ions and often identify such sites as potentially misfolded, which indicates that the our model may be more accurate than indicated by the score.

The model of I-Ssp6803I-DNA complex (Fig. 2) reveals that this HEase may recognizes its long target mainly via the major groove, in particular using a long loop (aa 53–74) with a cluster of Arg residues that may make specific contacts with bases in the DNA. Additional contacts may be formed with the middle part of the recognition sequence via the minor groove, using the positively charged N-terminus (aa 2-6); however, the conformation of this region is highly uncertain, and no precise predictions about its function can be made at this stage.

The active/metal-binding sites in the I-Ssp6803I model resembles the classical structure observed in most of other PD-(D/E)XK nucleases. The presence of Q instead of D or E in the (D/E)XK part of the catalytic motif is a little bit unusual, but not unprecedented—Q at this position has been already reported in PD-(D/E)XK nucleases, particularly in the TnsA transposase whose crystal structure has been determined (1f1z; Hickman et al., 2000), and in other enzymes, including a methyl-directed REase Mrr and nuclease ExiS of MAV1 phage (Kosinski et al., 2005a). This residue is not involved in recognition of the DNA, but in coordination of the metal ion, and we believe that it does not influence sequence specificity or the reaction mechanism in any dramatic way.

The I-Ssp6803I dimer is symmetrical on the level of the protein, while its native DNA target site is not, with the exception of the external pentanucleotide TCGGG. This suggests that this HEase may tolerate base substitutions in the middle of the target site, while it requires conservation of the external part of the DNA sequence. The extended loop of I-Ssp6803I enters the major groove of the DNA and appears to make contact exactly with the symmetrically conserved part of the DNA target. In particular, arginine residues R67, R68 and R70 make contact with the three conserved guanosines
(this might be, however, an artifact, as the quality of the model in this region is too low to predict interactions with atomic detail, despite our efforts to optimize the local geometry and the energy of interactions).

Although PD-(D/E)XK enzymes in general possess loops in the corresponding position, thus far they have not been found to introduce them into the major groove of the DNA in the ‘LAGLIDAGD-like’ way. Such a mode of sequence recognition may be a specific ‘invention’ of I-Ssp6803I and its close homologs. Interestingly, homologous sequences identified by database searches exhibit strong variability in the potential DNA-binding loop, despite the relative conservation of the surrounding elements of predicted secondary structure (Fig. 1). This variability suggests that these proteins may exhibit different sequence specificity than I-Ssp6803I, which makes them attractive targets for experimental characterization. Moreover, site-directed or random mutagenesis of the unusual loop associated with the PD-(D/E)XK scaffold may generate enzymes with altered specificities that could have commercial value.

It must be remembered that the current model of I-Ssp6803I is of low resolution and may exhibit local errors (which are inevitable at such a low level of target-template similarity). In particular, conformations of loops and details of protein–DNA interactions should be considered with a grain of salt. For instance, we cannot exclude the possibility that the C-terminal region (modeled de novo) has a different conformation in the real structure and is involved in direct interactions with the DNA. Nonetheless, we believe that our preliminary model accurately predicts the protein fold, the overall quaternary structure, the catalytic residues and the candidates for DNA-binding residues of I-Ssp6803I, which may be targeted, e.g. by site-directed mutagenesis experiments. Thus, the model may serve as a convenient platform for further experiments, whose results will be used to refine it.

Summarizing, our discovery that I-Ssp6803I is the first HEase with the PD-(D/E)XK fold provides a new, unexpected and exciting evolutionary link between HEases and REases. Our molecular model suggests that REase-like structures may be exploited to develop HEase-like specificities, which opens up new doors to the study of the evolution of DNA sequence specificity in these biotechnologically important nucleases. It would be extremely useful to engineer HEases and REases with new specificities, and I-Ssp6803I may be an excellent target for systematic modification of the substrate preference towards the ‘REase-like’ site with few bases recognized with high stringency or perhaps extension of the target site by insertion of residues into the presumed DNA-binding loop.

ACKNOWLEDGEMENTS
This analysis was funded by the NIH (Fogarty International Center grant R03 TW007163-01).

Conflict of Interest: none declared.

REFERENCES


