Structure *16*

Supplemental Data

Structures of the Rare-Cutting Restriction Endonuclease NotI Reveal a Unique Metal Binding Fold Involved in DNA Binding

Abigail R. Lambert, Django Sussman, Betty Shen, Robert Maunus, Jay Nix, James Samuelson, Shuang-Yong Xu, and Barry L. Stoddard

Supplemental Experimental Procedures

Protein Expression and Purification

The IMPACT purification system (NEB) was used to express and purify wt NotI. The *notIR* gene was cloned into the NdeI-SapI sites of pTYB1, and NotI was expressed as a C-terminal fusion to the *Sce* intein and chitin binding domains. The host for over-expression was *E. coli* strain ER2744 carrying pACYC-*eagIM*. Eight liters of cells were grown at 30°C to OD_{600} = 0.65 and then induced with 0.4 mM IPTG for 7 hours at 30°C. The cells were resuspended in chitin column buffer (Tris-HCl (pH 7.8), 500 mM NaCl, 0.1 mM EDTA, 5% glycerol and 0.1% Triton X-100) and cell extract was prepared by sonication. The clarified cell extract was loaded onto 40 mL chitin resin (NEB). After washing, the chitin resin was flushed with column buffer containing 40 mM DTT and incubated overnight at 4°C. Purified NotI was eluted with chitin column buffer containing 200 mM NaCl. The sodium chloride concentration was then reduced to 50 mM for DEAE-Sepharose column loading. Protein was eluted from the DEAE column using a 0.05 – 0.8 M NaCl gradient. Purified fractions were pooled for a final gel filtration step using 10 mM Tris-HCl (pH 7.6) 0.5 M NaCl, 0.1 mM EDTA, 1 mM DTT and 5% glycerol. NotI purity was assessed by SDS-PAGE and the final pool was dialyzed into storage buffer (200 mM NaCl, 20 mM Tris-HCl (pH 7.6), 0.1 mM EDTA, 1 mM DTT and 50% glycerol) and stored at -20°C.

Seleno-methionine labeled NotI was produced by expression in the methionine auxotroph T7 Express Crystal strain (NEB #C3022). The expression strain was transformed with pACYC-*eagIM* and pAII17-*notIR* and

plated on LB agar plus ampicillin and chloramphenicol. Individual colonies were inoculated into 10 mL M9 starter cultures supplemented with 50 ug/mL L-methionine, 0.2% glucose, 1 mM $MgSO_4$, 1 mM $CaCl_2$, 0.0002% ferric ammonium citrate, 100 ug/mL ampicillin and 30 ug/mL chloramphenicol. Four liters of the same M9 media (containing 50 ug/mL L-methionine) were each inoculated with a 10 mL starter culture and grown at 37°C until OD_{600} = 0.8. At this point the cells were pelleted and resuspended in pre-warmed M9 media containing all supplements except methionine. The cells were then grown for 3 hours to deplete methionine. Next, seleno-D/L-methionine was added at 100 ug/mL and 30 minutes later 0.3 mM IPTG was added to induce NotI expression overnight. Seleno-methionine labeled NotI was purified as described for wt NotI except that Heparin HyperD (BioSepra) chromatography substituted for the chitin column step.

Seleno-methionine incorporation was assessed by mass spectrometry (J. Benner, NEB). The mass analysis indicated that greater than 90% of the protein molecules contained seleno-methionine at all 6 positions. Specific activity of the selenium labeled protein is comparable to the specific activity of wt NotI.

Crystallization

Unlabeled protein was dialyzed into crystallization buffer (5% glycerol, 50 mM Tris-HCl pH 7.6, 20 mM NaCl, 1 mM DTT, 1 mM $CaCl_2$) and concentrated to ~10 mg/mL. Crystallization experiments were carried out using vapor diffusion with hanging-drop geometry. Initial sparse matrix crystallization screens were conducted with a series of DNA duplexes varying in length and cohesive overhang. Optimized crystallization conditions based off of initial hits yielded two different crystal forms.

The first crystal form (Form #1), containing a DNA-bound complex, was grown at 4°C with 6-8 mg/mL protein, in 29-39% PEG 4000, 130-240 mM $LiSO_4$, 100mM Tris HCl pH 8.5, and in the presence of a 2:1 molar ratio of a 22bp blunt-ended DNA duplex (DNA:protein dimer). DNA duplexes were prepared by mixing equimolar amounts of two DNA oligos, 5'-CGGAGGCGCGGCCGCGCCGCCG-3' and 5'-CGGCGGCGCGGCCGCGCCTCCG-3', heating at 95°C for 10 minutes, and then gradually cooling under

thermal cycler control to 25°C over six hours. Yellow-colored crystals of the space group $P2_1$ grew after one week. Crystals were briefly dipped in paraffin oil for cryoprotection prior to freezing in liquid nitrogen.

The second crystal form (Form #2), containing unbound protein, was grown at room temperature with 6 mg/mL protein in 2.0 M $(NH_4)_2SO_4$, 100 mM MES pH 6.5, 10 mM $CaCl_2$, and 20 mM NaCl. A 1.5:1 molar ratio of a 12 bp blunt-ended DNA duplex was also present in the initial crystallization experiments, but the DNA was not visible in the resulting electron density maps. Subsequent experiments demonstrated that the DNA was not essential for growth of this crystal form, but greatly facilitated growth via an indirect (presumably buffering) effect. Deep yellow crystals of the space group $P4_32_12$ grew sporadically amidst heavy protein precipitation after 1-2 weeks. Improved efficiency in crystal growth was achieved using both micro- and macroseeding techniques. Artificial mother liquor (2.25 M $(NH_4)_2SO_4$, 100 mM MES pH 6.5, 10 mM $CaCl_2$, 20 mM NaCl) with 25% glycerol was used for cryoprotection, and the crystals were frozen in liquid nitrogen.

Derivitization for Phasing

Both crystal forms were also grown with SeMet-labeled protein for the purpose of phase determination, and a high quality mercury derivative of unbound protein (crystal form #2) was also prepared. SeMet NotI was dialyzed into crystallization buffer (see above) and concentrated to ~10 mg/mL. Form #1 SeMet crystals (DNA-bound complex) were successfully grown in the same range of conditions as the native crystals. Form #2 SeMet crystals (unbound protein) were grown in pre-equilibrated drops (with SeMet protein) seeded with small (50 micron) crystals of unlabeled protein. A large, native crystal of the apo enzyme (form #2) was soaked in a 10 uL drop for 2.5 hours in the presence of 10 mM $HgCN_2$. The crystal was then back-soaked for 30 seconds in artificial mother liquor (AML) with no mercury, cryo-protected with 25% glycerol in AML, and frozen in liquid nitrogen.

Data Collection and Processing

Datasets were collected at the Advanced Light Source (ALS) beamlines 5.0.1, 5.0.2, and 8.2.1 (Table 1). Form #1 SeMet datasets (DNA-bound complex) were collected at selenium peak and high-energy remote wavelengths; Form #2 SeMet data were collected at the selenium peak, inflection, and high-energy remote wavelengths. Additionally, a mercury derivative dataset for Form #2 was collected at the mercury peak wavelength. Images were processed using the Denzo, Scalepack, and HKL2000 software packages (Otwinowski and Minor, 1997).

Structure Determination and Refinement

The structure of unbound NotI enzyme (Form #2) was solved first to 2.8Å resolution using phases derived from a combination of selenium multiple anomalous dispersion (MAD) and mercury single isomorphous replacement (SIR). Reflection intensities were converted to structure factors (TRUNCATE) (CCP4, 1994) and datasets were merged and scaled using the CCP4 software suite (CCP4, 1994). An initial mercury site was manually identified from an isomorphous difference Patterson map (**Figure S1**), and a second mercury site was determined by the program MLPHARE (CCP4). Experimental SIR phases were then generated by MLPHARE. No additional mercury sites were found when difference Fourier maps were inspected.

Four selenium sites were then identified from SeMet remote-wavelength structure factors cross-phased with Hg phases in CCP4 (FFT, CCP4). These four Se sites were input into an automated MAD phasing run by SHARP (Fortelle and Bricogne, 1997) which produced electron density sufficient for identification of two Fe atoms and 10 Se sites per asymmetric unit. A final phasing run in SHARP combined phasing power from 10 Se sites at three different wavelengths and 2 mercury sites from a single isomorphous dataset to produce electron density maps into which the model was built.

455 of 766 (dimer) residues were placed (173 in sequence, the rest as Ala/Gly placeholders) by the autobuild portion of the PHENIX program (Adams, 2002). The remainder of the model was built by hand using the

molecular modeling program COOT (Emsley and Cowtan, 2004). Restrained refinement was performed in REFMAC (Murshudov et al., 1997), including restraints from two-fold non-crystallographic symmetry (NCS). Sulfate ions were added by hand where obvious in difference maps and followed by final rounds of TLS refinement in REFMAC (Winn, 2001). Ramachandran statistics for the free enzyme structure, as calculated by PROCHECK (Laskowski, 1993) are as follows: Core 84.2%, Allowed 13.3%, Generously allowed 1.5%, Disallowed 1.0 %.

The structure of the DNA-bound complex was then solved. The DNA duplex was first located and modeled, using a 3.0 Å resolution Se-Met MAD-phased electron density map calculated from peak and high-energy remote SeMet datasets (**Figure S2**). The bound protein was then located by molecular replacement using a 2.5 Å resolution dataset collected from a crystal containing unlabeled protein.

Twenty-two basepairs of DNA were modeled into the MAD-phased density using COOT, using a template of B-form DNA structure and sequence generated by the 'make-na' server (http://structure.usc.edu/make-na/). This hand-built DNA and a model of the endonuclease catalytic core from the apo enzyme structure were used individually as search models for molecular replacement by PHASER (McCoy et al., 2005). Two copies of the REase core and one copy of the 22bp DNA were located using the 2.5Å native data to generate an initial model of the core enzyme dimer bound to its DNA. This catalytic core plus DNA model was fixed in a second round of molecular replacement, and the metal-binding domain (residues 1-85) was added as an additional search model. The resulting placement of the metal-binding domain permitted contiguous modeling of the peptide chain between residues 85 and 86. The placement of the protein domains was closely superimposable with the location of selenium sites calculated from lower resolution data sets collected using derivatized crystals (**Figure S3**).

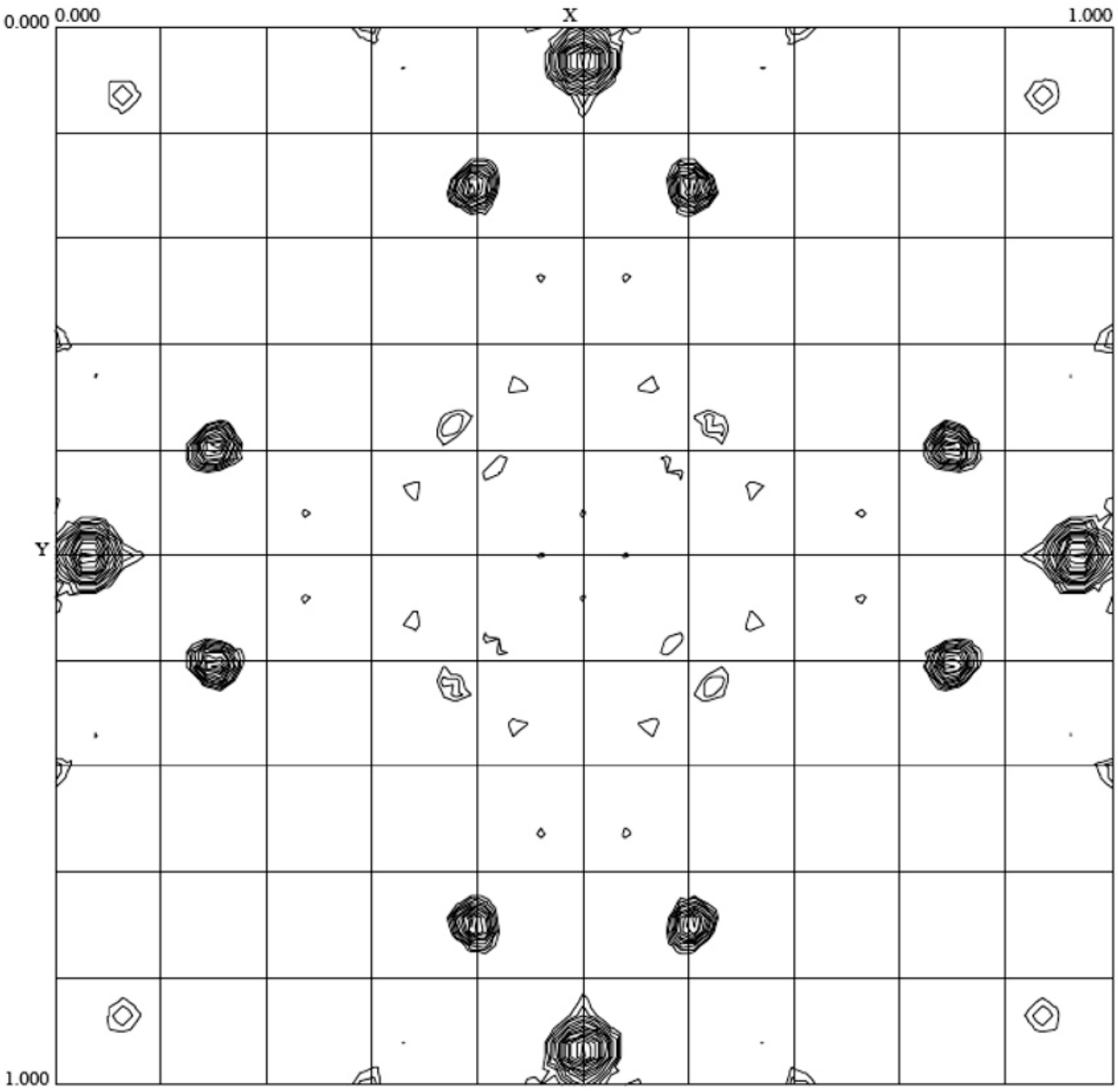
The model was subjected to a round of rigid body refinement in Refmac. The electron density after rigid body refinement allowed for manual adjustment of chain position and building of missing loop regions. Further restrained refinement was performed in Refmac including restraints from the two-fold NCS between chains of

the protein dimer. Water ions were added initially using automated methods and then further supplemented manually. Final rounds of refinement in Refmac included TLS refinement. Ramachandran statistics for the DNA-bound structure (PROCHECK) are as follows: Core 88.4%, Allowed 11.1%, Generously allowed 0.5%, Disallowed 0.0%.

Determination of metal content by X-ray fluorescence measurements

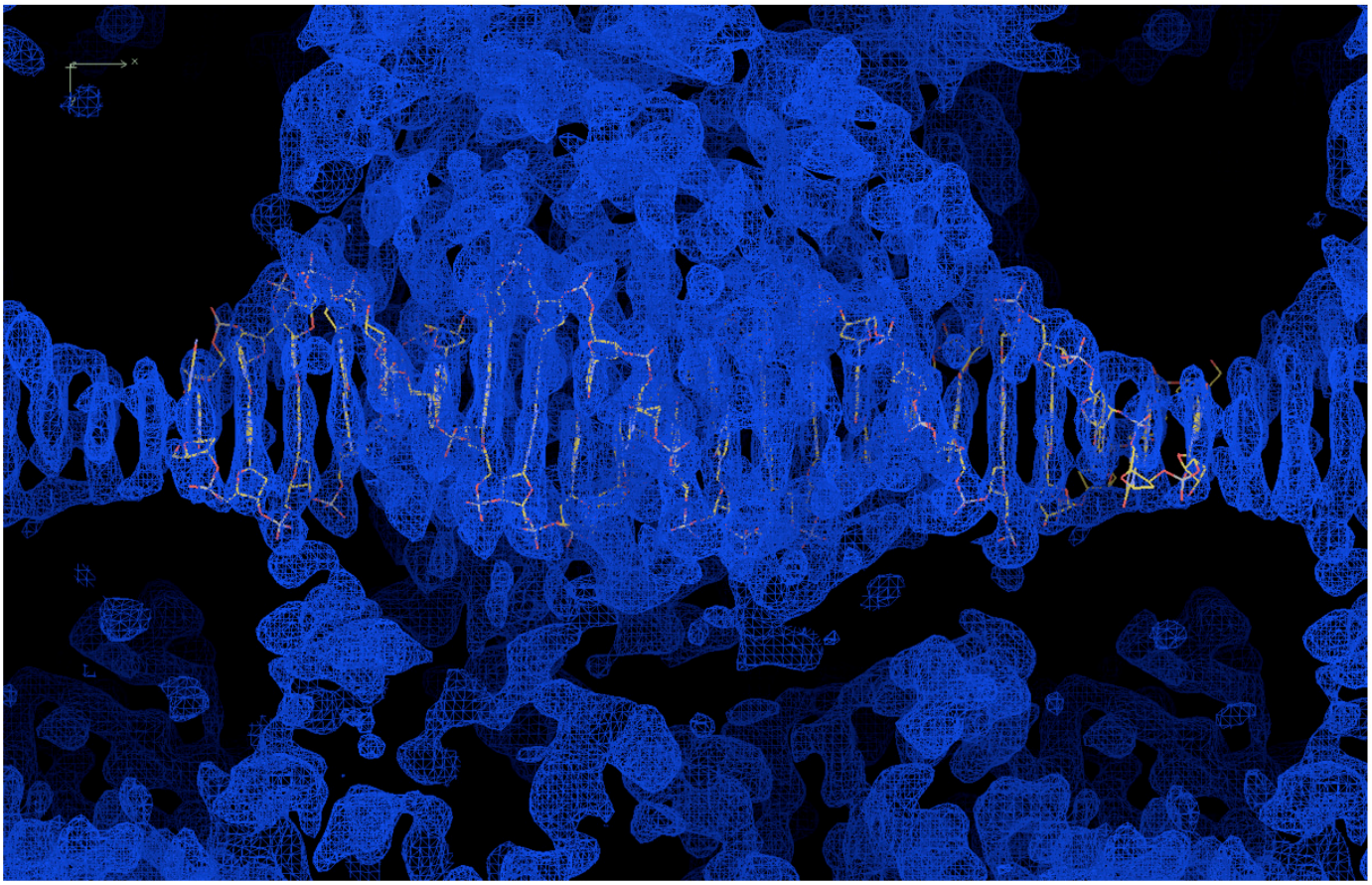
X-ray fluorescence data was collected at the Advanced Light Source beamline 4.2.2 on both native and selenium derivative crystals. Fluorescence emission spectra collected at an energy of 12,700eV (0.9762 Å) were used to determine the metals present in each sample. A clear emission line at 6400eV indicating the presence of iron was found in both samples; no emission lines at the zinc emission energy of 8640eV were seen in either sample. The presence of iron (and absence of zinc) in the NotI crystals was verified by fluorescence absorption scans at the K absorption edge of iron, 7.1120 keV and at the K absorption edge of zinc, 9.6586 keV (**Figure S4**).

Supplementary Figure S1



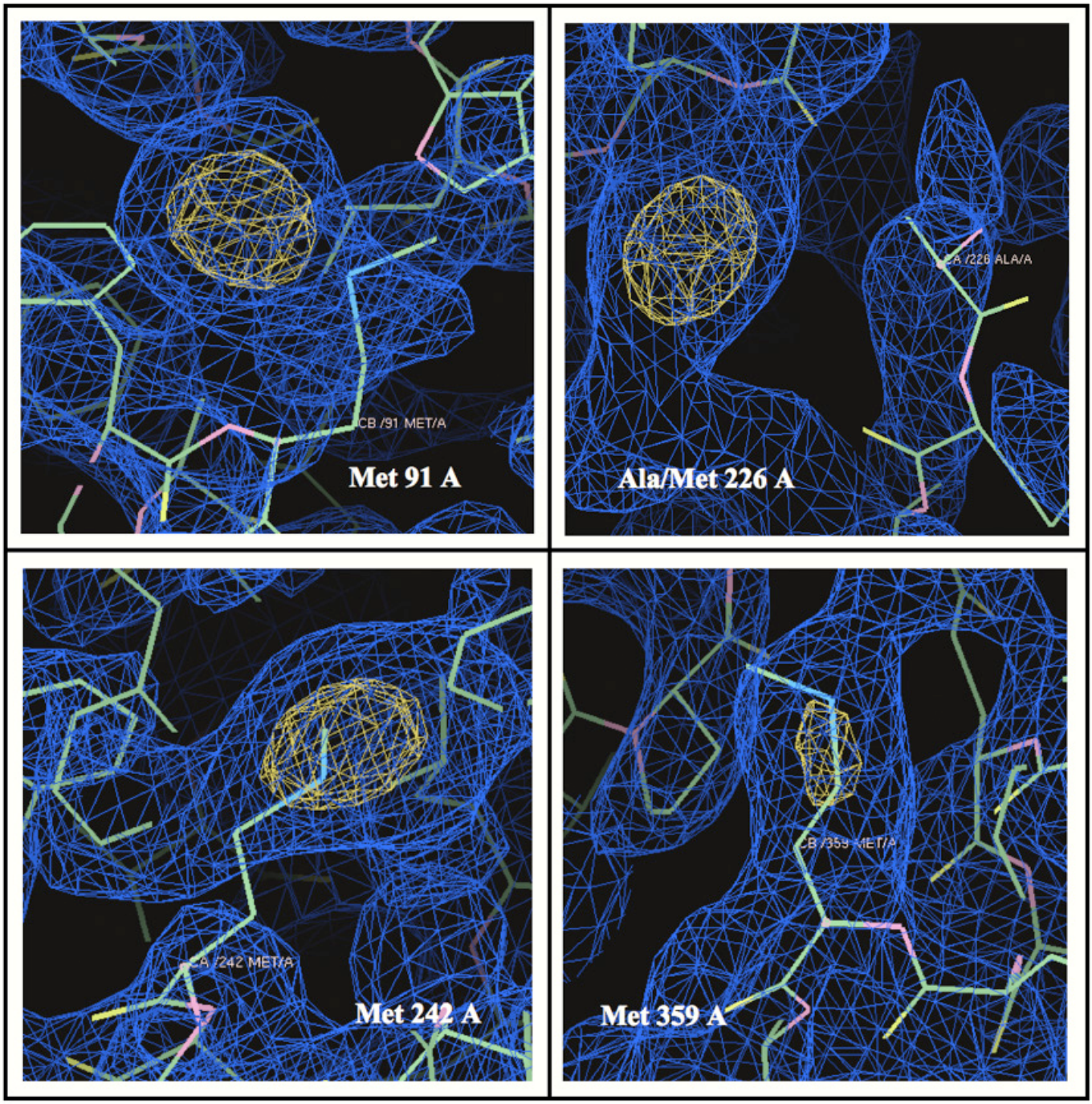
Supplementary Figure S1. Isomorphous difference Patterson map of NotI $Hg(CN)_2$ derivative which was solved for 2 mercury atoms per asymmetric unit (1 per protein chain, Cys254). Sample Harker section at w = 0.5 (space group $P4_32_12$).

Supplementary Figure S2



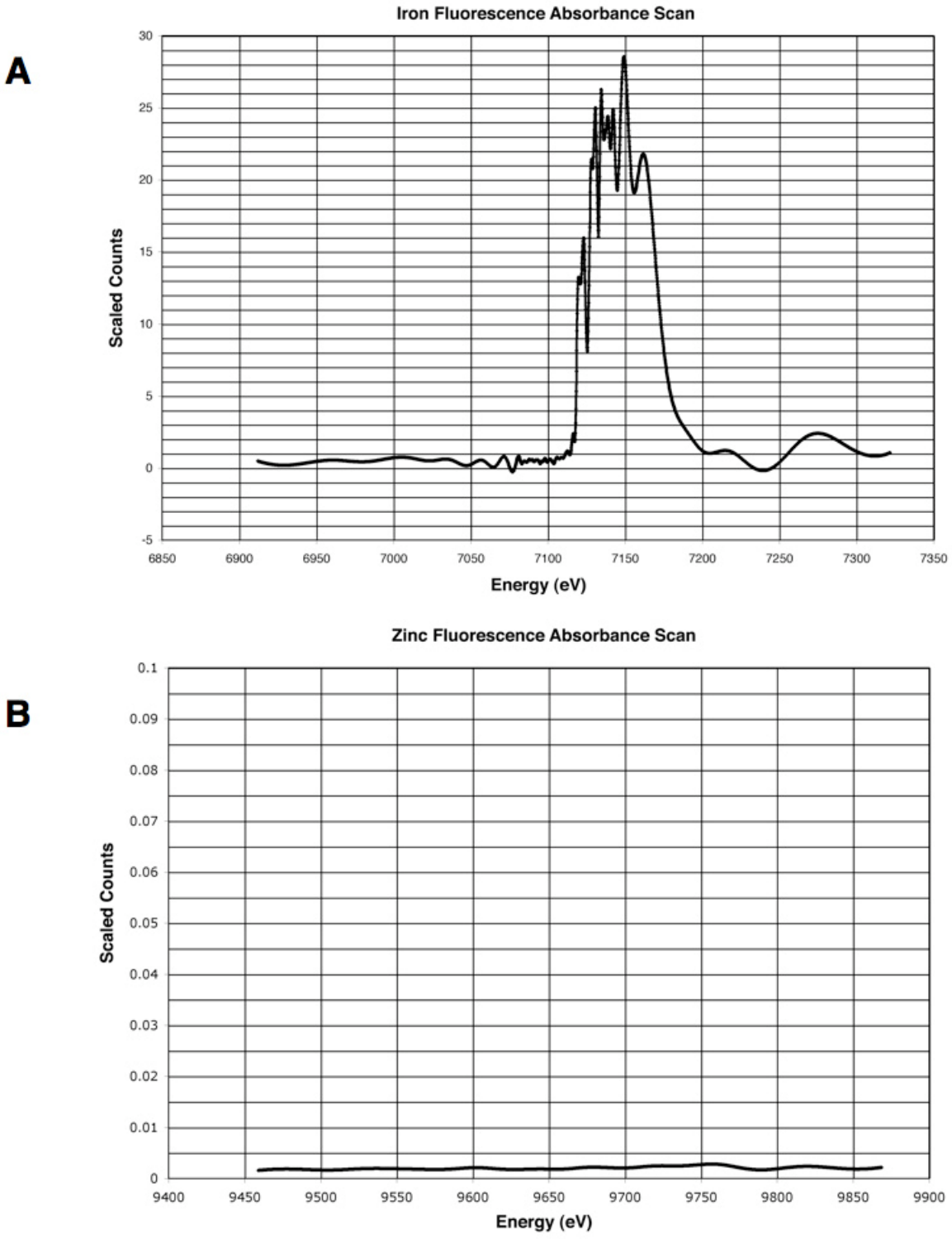
Supplementary Figure S2. Coordinates depict 22bp of DNA built into a 2-wavelength (peak and high energy remote) 3.0Å SeMet MAD-phased electron density map. NotI protein density was not of sufficient quality for backbone tracing or modeling of sidechains. The hand-built DNA was subsequently used as a search model for molecular replacement into a native dataset. (Screenshot taken from inside the program COOT.)

Supplementary Figure S3



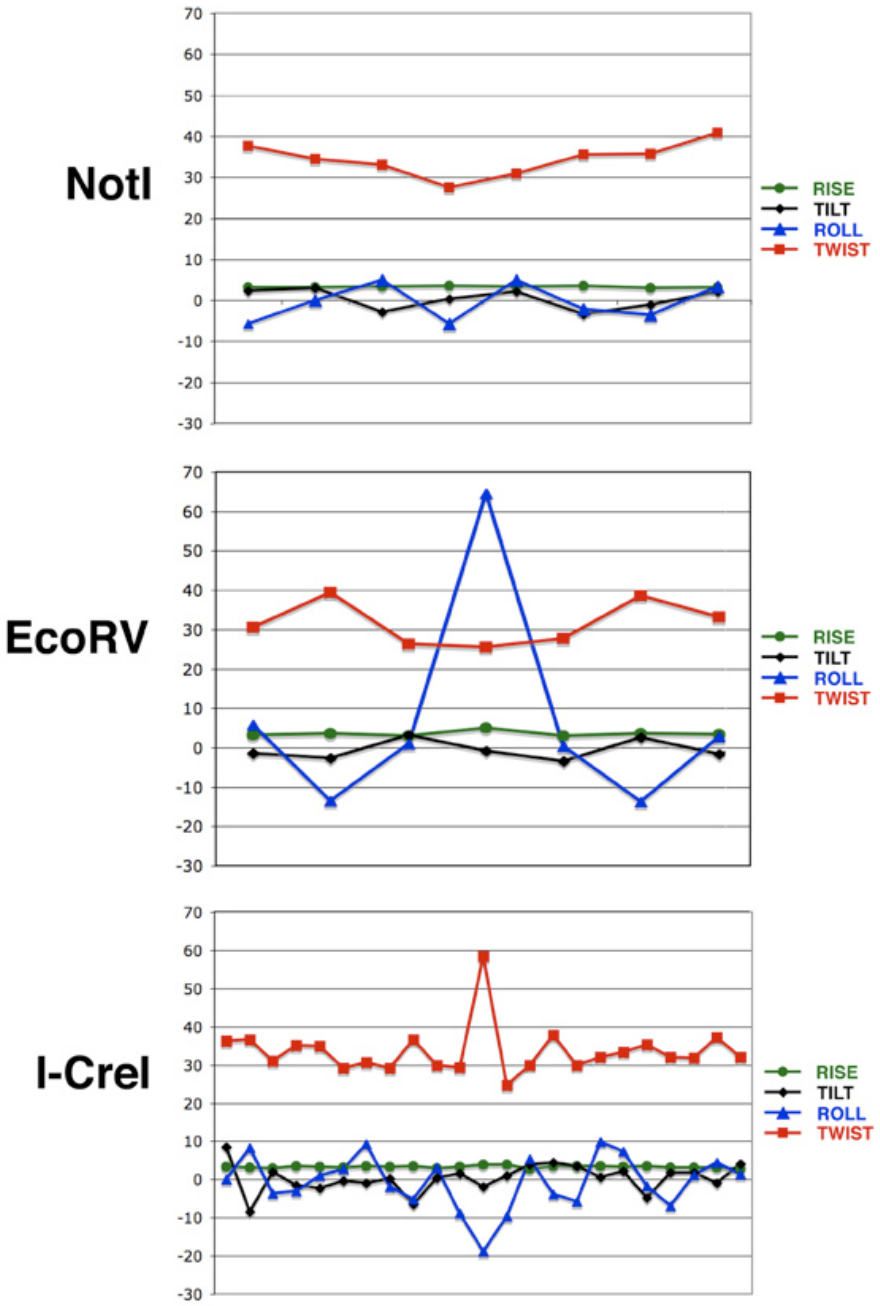
Supplementary Figure S3. Coordinates of molecular replacement (MR) solution shown inside 3.0Å SeMet MAD experimental electron density map. The positions of methionine residues in the MR solution coincide nicely with respect to yellow, high-contoured peaks (7.0σ for Met 91, 226, 242, and 5.0σ for residue 359) representing selenium density. Met186 was not included in the molecular replacement search model and is therefore not shown.

Supplementary Figure S4



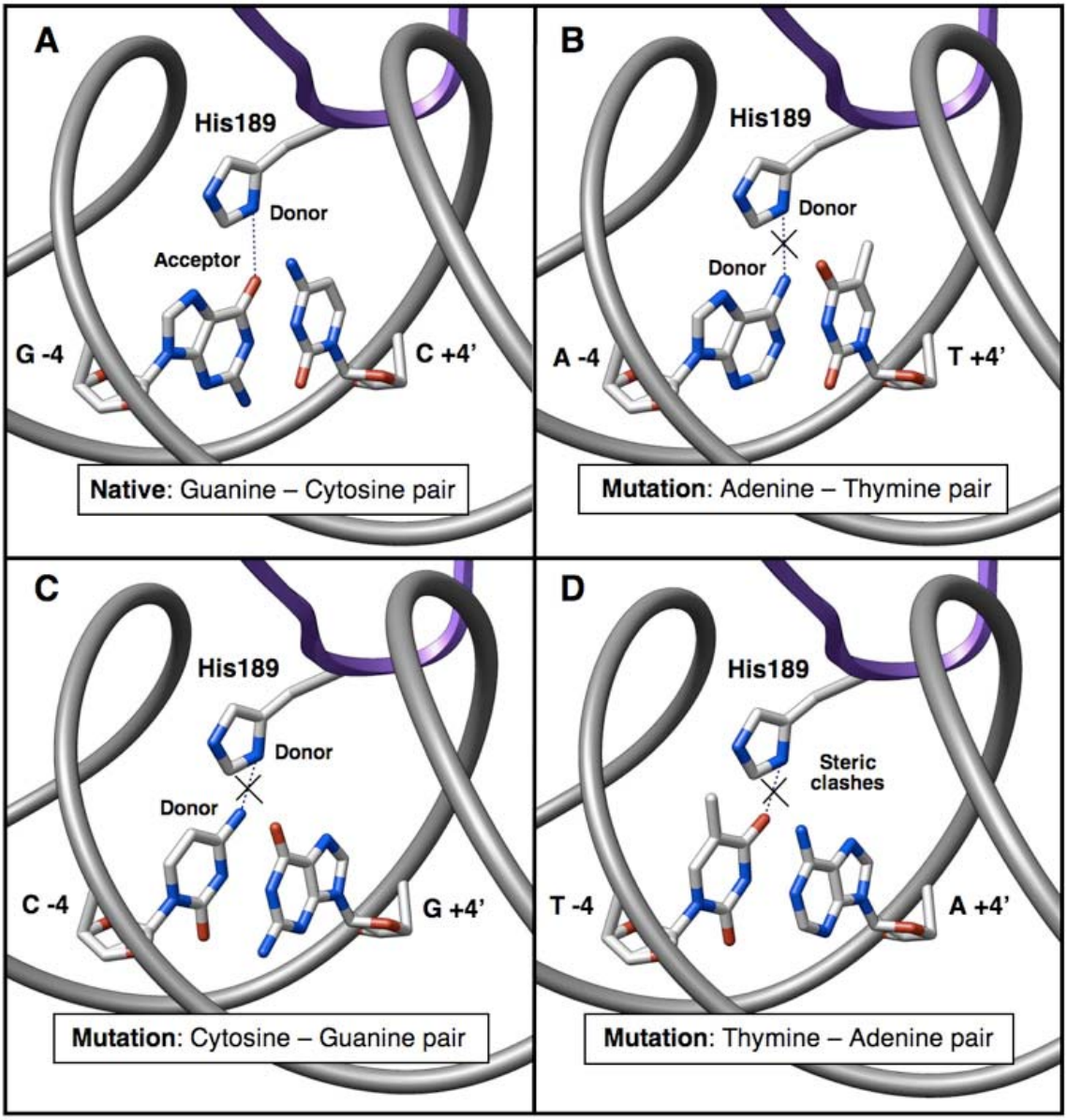
Supplementary Figure S4. X-ray fluorescence absorption spectra were collected across the K absorption edge of a.) iron, 7.1120 keV and b.) zinc, 9.6586 keV. A peak at the expected iron absorption edge identifies the presence of iron in the NotI crystals, while the absence of signal at the zinc edge signifies no zinc metal present. This method was also used to verify the incorporation of mercury and selenium in the derivatized crystals (data not shown).

Supplementary Figure S5



Supplementary Figure S5. DNA bend parameters were quantitated using program 'Readout' (34), via a web-based server located at http://gibk26.bse.kyutech.ac.jp/jouhou/readout/). The values of 'tilt' and 'roll' indicate rigid body rotation of basepairs, along torsion angles that are orthogonal to the DNA axis, resulting in base unstacking. "Twist" and "rise" indicate relative under- or over-winding of DNA that results in widening or narrowing of the DNA major or minor grooves.

Supplementary Figure S6



Supplementary Figure S6. Fidelity at the outermost basepair of the NotI target site. Panel a: NotI uses His189 to make a single base-specific contact in the major groove to the outer-most basepairs (+/- 4) of its 8 basepair recognition sequence. (A single water-mediated contact is also made to G -4 in the minor groove, which is *not* shown here.) This single contact appears to be sufficient for maintaining specificity for G-C at this position. Mutation to **(Panel b)** an A-T basepair or **(Panel c)** a C-G basepair disrupts hydrogen bonding possibilities between protein and bound DNA. **Panel d**: Mutation to a T-A basepair retains the correct chemistry for hydrogen bonding, but conformational changes necessary to adopt the correct geometry to form this bond would introduce unfavorable steric clashes.

REase	1st acidic residue	2nd acidic residue	General Base	P**D...(D/E)**-X-**K**	Outliers/Exceptions
BamHI	Asp94	Glu111	Glu113	I**D.**$_{16}$**.E**-X-**E**	Glu as general base
BcnI	Asp55	Glu60	Lys62	G**D.**$_{4}$**.E**-X-**K**	
BglI	Asp116	Asp142	Lys144	P**D.**$_{25}$**.D**-X-**K**	
Bgl2	Asp84	Glu93	Gln95	I**D.**$_{8}$**.E**-X-**Q**	Gln as general base
Bse634I	Asp146	Glu212	Lys198	P**D.**$_{51}$**.K**-(X$_{13}$)-**E**	Lys preceeds 2nd acidic residue
BsoBI	Asp212	Glu240	Lys242	V**D.**$_{27}$**.E**-X-**K**	
BstYI	Asp119	Glu128	Gln130	T**D.**$_{8}$**.E**-X-**Q**	Gln as general base
Cfr10I	Asp134	Glu204	Lys190	P**D.**$_{55}$**.K**-(X$_{13}$)-**E**	Lys preceeds 2nd acidic residue
Ecl18kI	Asp160	Glu195	Lys182	V**D.**$_{21}$**.K**-(X$_{12}$)-**E**	Lys preceeds 2nd acidic residue
EcoO1091	Asp77	Asp110	Lys126	I**D.**$_{32}$**.D**-(X$_{15}$)-**K**	
EcoRI	Asp91	Glu111	Lys113	P**D.**$_{19}$**.E**-X-**K**	
EcoRII	Asp299	Glu337	Lys324	P**D.**$_{24}$**.K**-(X$_{12}$)-**E**	Lys preceeds 2nd acidic residue
EcoRV	Asp74	Asp90	Lys92	P**D.**$_{15}$**.D**-X-**K**	
FokI	Asp450	Asp467	Lys469	P**D.**$_{16}$**.D**-X-**K**	
HincII	Asp114	Asp127	Lys129	A**D.**$_{12}$**.D**-X-**K**	
HinP1I	Asp62	Gln81	Lys83	S**D.**$_{18}$**.Q**-X-**K**	Gln as second acidic residue
MspI	Asp99	Asn117	Lys119	T**D.**$_{17}$**.N**-X-**K**	Asn as second acidic residue
MunI	Asp83	Glu98	Lys100	P**D.**$_{14}$**.E**-X-**K**	
MvaI	Asp50	Glu55	Lys57	P**D.**$_{4}$**.E**-X-**K**	
NaeI	Asp86	Asp95	Lys97	T**D.**$_{8}$**.D**-X-**K**	
NgoMIV	Asp140	Glu201	Lys187	P**D.**$_{46}$**.K**-(X$_{13}$)-**E**	Lys preceeds 2nd acidic residue
NotI	Asp160	Glu182	Gln184	F**D.**$_{21}$**.E**-X-**Q**	Gln as general base
PvuII	Asp58	Glu68	Lys70	N**D.**$_{9}$**.E**-X-**K**	
SdaI	Asp233	Glu248	Lys251	P**D.**$_{14}$**.E**-(X$_{2}$)-**K**	
SfiI	Asp79	Asp100	Lys102	I**D.**$_{20}$**.D**-X-**K**	

Figure S7. Table of the 25 PD...(D/E)xK Restriction Enzyme Structures Solved to Date
The three active site residues of the core REase nuclease motif are listed, along with the customized/modified PD...(D/E)xK motif for each enzyme. Structures that deviate in any way from the canonical PD...(D/E)xK REase are noted in the far right column.