Functional assignment based on structural analysis: Crystal structure of the yggJ protein (HI0303) ofHaemophilus influenzae reveals an RNA methyltransferase with a deep trefoil kno…

6 AUTHORS, INCLUDING:

Rong Xiao
Rutgers, The State University of New Jersey
181 PUBLICATIONS **2,828** CITATIONS

Thomas B Acton
Rutgers, The State University of New Jersey
128 PUBLICATIONS **2,541** CITATIONS

Giovanni T Monteleone
University of Rome Tor Vergata
255 PUBLICATIONS **7,717** CITATIONS

STRUCTURE NOTE

Functional Assignment Based on Structural Analysis: Crystal Structure of the yggJ Protein (HI0303) of *Haemophilus influenzae* Reveals an RNA Methyltransferase With a Deep Trefoil Knot

Farhad Forouhar,[1] **Jianwei Shen,**[1] **Rong Xiao,**[2] **Thomas B. Acton,**[2] **Gaetano T. Montelione,**[2] **and Liang Tong**[1]*

[1]*Department of Biological Sciences, Northeast Structural Genomics Consortium, Columbia University, New York, New York*
[2]*Center for Advanced Biotechnology and Medicine, Department of Molecular Biology and Biochemistry, Rutgers University, New Brunswick, New Jersey*
[3]*Department of Biochemistry, Robert Wood Johnson Medical School, Northeast Structural Genomics Consortium, Piscataway, New Jersey*

Introduction. Methyltransferases (MTases) constitute a large family of enzymes that transfer the methyl group of S-adenosylmethionine (AdoMet) to carbon, nitrogen, or oxygen atoms of DNA, RNA, proteins, and small molecules.[1,2] In this process, AdoMet is converted to S-adenosylhomocysteine (AdoHcy), which is a potent inhibitor of MTases. Consequently, the enzymatic activities of MTases are regulated by the ratio of AdoMet to AdoHcy concentrations in the cell.

Structures of >20 different MTases are currently available.[3] Most of these structures contain a core domain with a central seven-stranded β-sheet, with the last strand running antiparallel to the first six strands.[3] This core domain is associated with the binding of the AdoMet molecule, and additional domains in these enzymes are involved in binding the other substrates.

Recently, a new class of MTases was characterized on the basis of the structures of the RrmA,[4] MT1,[5] and YibK[6] proteins. These enzymes contain a core domain with a different backbone fold, having a central six-stranded fully parallel β-sheet. More importantly, this core domain has a deep trefoil knot at its C-terminus, and this knot appears to be crucial for the binding of the AdoMet molecule and for the dimerization of these enzymes.

In recent years, several structural genomics initiatives have been established aimed at rapidly elucidating protein structures of functional and biological interest, developing relevant technologies, and providing a more comprehensive picture of protein conformational space. Significantly, the structures of RrmA,[4] MT1,[5] and YibK[6] MTases have all been determined by structural genomics consortia. The Northeast Structural Genomics Consortium (NESG) is particularly focused on clusters of eukaryotic domain families from several model organisms, including humans, and homologous proteins from bacteria and archea (www.nesg.org).

NESG target IR73 is a 245-residue hypothetical protein, yggJ (HI0303), from *Haemophilus influenzae*. HI0303 is a member of a widely conserved protein family, represented in a large number of prokaryotic genomes as well as in *Arabidopsis thaliana* (Fig. 1) and is annotated in Swiss-Prot as a hypothetical protein of unknown function. Here we report the crystal structure of HI0303 at 2.0 Å resolution. On the basis of its structural similarity with the RrmA,[4] MT1,[5] and YibK[6] MTases, and despite the lack of significant sequence similarity to them, we propose that HI0303 is another member of the trefoil-knot class of MTases.

The crystal structure was determined by the selenomethionyl single-wavelength anomalous diffraction (SAD) method[7] and deposited in the Protein Data Bank under the accession code 1NXZ (Table I). Out of the residues in the dimer in the asymmetric unit, 89.1% are in the most favored regions of the Ramachandran plot, and 10.6% are in the additionally allowed regions. One residue (Ser87) is in the generously allowed regions, and none are in disallowed regions.

The structure of the HI0303 monomer is made of two domains. The small domain includes the N-terminal 72 residues of the protein and contains a twisted five-stranded β-sheet (β1–β5) and one helix (α1) [Fig. 2(A)]. The structure of this domain closely resembles that of the RNA-binding domain of the ribosomal protein TL5 (PDB accession code 1FEU),[8] even though the amino acid sequence identity among the structurally equivalent residues of the two domains is only 14%.

β1　　β2　　α1　　β3　　β4　　β5

```
HI0303    MRIPRIYHPISLENQTQCYLSEDAAN-HVARVLRMTEGEQLELFDGSNHIYPAKIIESNKKSVKVEILGRE--LADK  74
yggJ      MRIPRIYHPEPLTSHSHIALCEDAAN-HIGRVLRMGPGQALQLFDGSNQVFDAEITSASKKSVEVKVLEGQ--IDDR
yqeU      MQRYFIELTKQQIEEAPTFSITGEEVH-HIVNVRMNEGDQIICCSQDGFEAKCELQSVSKDKVSCLVIEWT--NENR
TM1380      MPHLFYGT---AQNGEVIFDEREA-HHMRVVRLKEGDVIEATDGNGFSYTCILKSLKKTAAAKIVKVE--EKEK
NP_175422 34-TVSSSPPSDLNLRNQSRVVRVQGSEFWHMAKVLRLKQEDRVELFNGKGGLVEGCIQSIDKTGVDFIAQEDQKVILPQ
```

β6　　α2　　β7　　α3　　β8

```
HI0303    ESHLKIHLGQVISRGERMEFTIQKSVELGVNVITPLWSERCGVKLDAERMDKKIQQWQKIAIAACEQCGRNIVPEIRPL-  153
yggJ      ESPLHIHLGQVMSRGEKMEFTIQKSIELGVSLITPLFSERCGVKLDSERLNKKLQQWQKIAIAACEQCGRNRVPEIRPA-
yqeU      ELPIKVYIASGLPKGDKLEWIIQKGTELGAHAFIPFQAARSVVKLDDKKAKKKRERWTKIAKEAAEQSYRNEVPRVMDVH
TM1380    EPTEKLSVVVPIGRWERTRFLIEKCVELGVDEIFFHKFERSQHEIS-------LDKAKIVVREAAKQCKRYLFPKVSFL-
NP_175422 GMQWVFAAFGTLKGGRADWLIEKCTELGASSVTPLLTERSPI-----ISENRVDRLERVSFAAAKQCQRLHQMVLNPPI
```

α4　　β9　　β10　　α5　　β11

```
HI0303    ------MKLQDW---CAENDGALKLNLHPRAHYSIKTLPTIPAGGVRLLIGSEGG-LSAQEIAQTEQQGPTEILLGKRVL  223
yggJ      ------MDLEAW---CAEQDEGLKLNLHPRASNSINTLP-LPVERVRLLIGPEGG-LSADEIAMTARYQFTDILLGPRVL
yqeU      SFQQLLQRMQDFDKCVVAYEESSKQGEISAFSAIVSSLP--KGSSLLIVPGPEGG-LTEAEVERLTEQDGVTCGLGPRIL
TM1380    ------EKLE-------FSGN--VITLDLDASQ--NLLDANLEGSITVVVGPEGG-FSEKEREL-LRSSTTIVSLGKKIL
NP_175422 KFGTLL----------DHSKLCLVATAEAKPLLNAVNSSAKESSGL-LIVGPEGADFTKKEVEMMLEAGGTAVGLGPHRL
```

α6

```
HI0303    RTETASLAAISALQICFGDLGE    245    Haemophilus influenzae
yggJ      RTETTALTAITALQVRFGDLG           Escherichia coli
yqeU      RTETAPLYALSAISYQTELLRGDQ        Bacillus subtilis
TM1380    RFETAAILTVGYIALKKQKI           Thermotoga maritima
NP_175422 RVETATIALLATLVMW-SDSQEMI       Arabidopsis thaliana
```
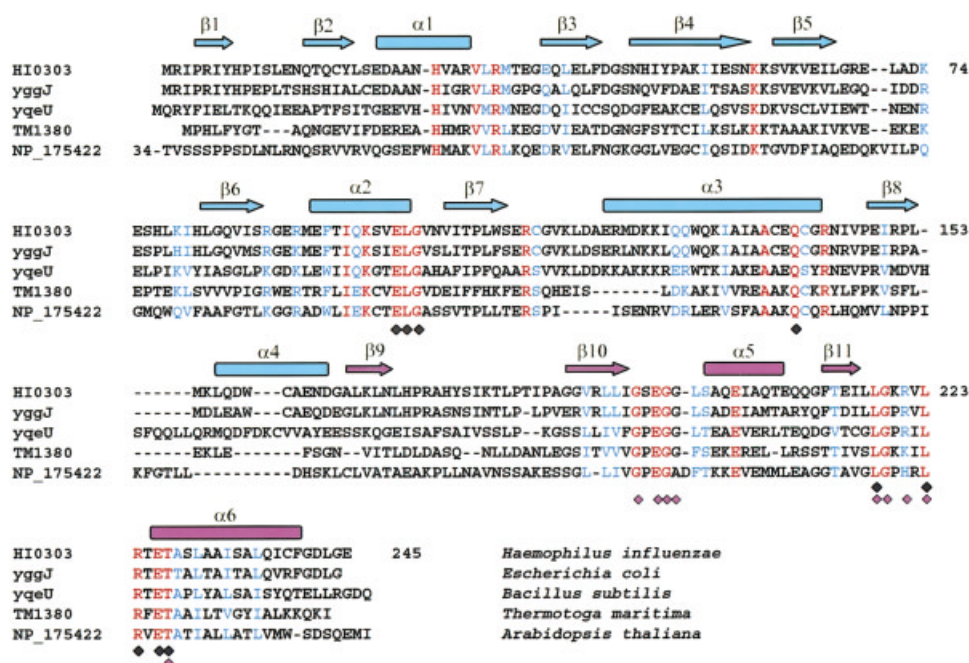
Fig. 1. Sequence aligment of HI0303 and its closely related homologs. The secondary structure elements in the crystal structure of HI0303 are shown above the alignment; the knot and the remaining part of HI0303 are colored magenta and cyan, respectively. Strictly conserved and conservatively substituted residues are colored red and blue, and residues involved in the HI0303 dimerization and cofactor-binding are marked below the alignment as black and magenta diamonds. The numbering on the right of each row corresponds to HI0303. The first 34 residues of the *Arabidopsis* protein are not shown.

TABLE I. Summary of Crystallographic Information

Maximum resolution (Å)	2.0
No. of observations	134,504
R_{merge} (%)[a]	6.2 (41.0)
No. of reflections	59,377
Figure-of-merit from SAD phasing	0.24
Resolution range for refinement	29.5–2.0
Completeness (%)	85 (61)
R factor (%)[b]	21.4 (22.7)
Free R factor (%)	26.4 (26.1)
RMS d in bond lengths (Å)	0.008
RMS d in bond angles (°)	1.1

[a]$R_{\mathrm{merge}} = \sum_h \sum_i |I_{hi} - \langle I_h \rangle| / \sum_h \sum_i I_{hi}$. The numbers in parenthesis are for the high-resolution shell.
[b]$R = \sum_h |F_h^o - F_h^c| / \sum_h F_h^o$

The core domain contains a central six-stranded parallel β-sheet (β6–β11) that is flanked by five helices (α2–α6) [Fig. 2(A)]. Near the C-terminus of this domain, the structure contains a deep trefoil knot,[9] so that the β11–α6 segment (~25 residues) is threaded through the β9–β10 loop. The structure of this domain is remarkably similar to the core domain of RrmA, MT1, and YibK [Fig. 2(B)]. However, the degree of amino acid sequence conservation among these proteins is very low (<15% identity for structurally equivalent residues).

A dimer of the HI0303 protein is observed in the crystal [Fig. 2(C)], consistent with solution light-scattering results (data not shown). The dimer is formed through contacts of the core domains of the two monomers, with little contribution from the small domains. The knot in the core domain (the β11–α6 segment) mediates a substantial portion of the dimer interface. A total of 1650 Å^2 of the surface area of each monomer is buried in the dimer interface, which is twice the value of 700 Å,^{2} that is found in many biologically relevant protein–protein interactions.[10] Residues in this interface are generally well conserved among the homologs of this protein (Fig. 1), suggesting that HI0303 and its homologs may function as dimers.

On the basis of this structural similarity, we propose that this trefoil knot in the core domain of HI0303 mediates the binding of the AdoMet/AdoHcy substrate, as observed in the structure of the YibK-AdoHcy complex.[6] Considering the structural similarity between YibK and HI0303, we generated a model for the complex of AdoHcy and HI0303 [Fig. 2(A), (C), (D)]. The AdoHcy molecule assumes the more commonly observed extended conformation in our model, instead of the strained conformation seen in the YibK-AdoHcy complex. In this model, AdoHcy is located in a cavity near the knot, surrounded by two highly conserved segments 195-GSEGG-199 and 218-LGKRVLRTET-225 in HI0303 (Fig. 1). The AdoHcy molecule shows generally favorable interactions with these conserved residues of the protein, with no bad steric clashes. Therefore, it is likely that HI0303 and its homologs can also bind AdoMet/AdoHcy and are thus also MTases. Attempts at cocrystallizing HI0303 with AdoMet were not successful because the protein precipitated shortly after addition of the compound.

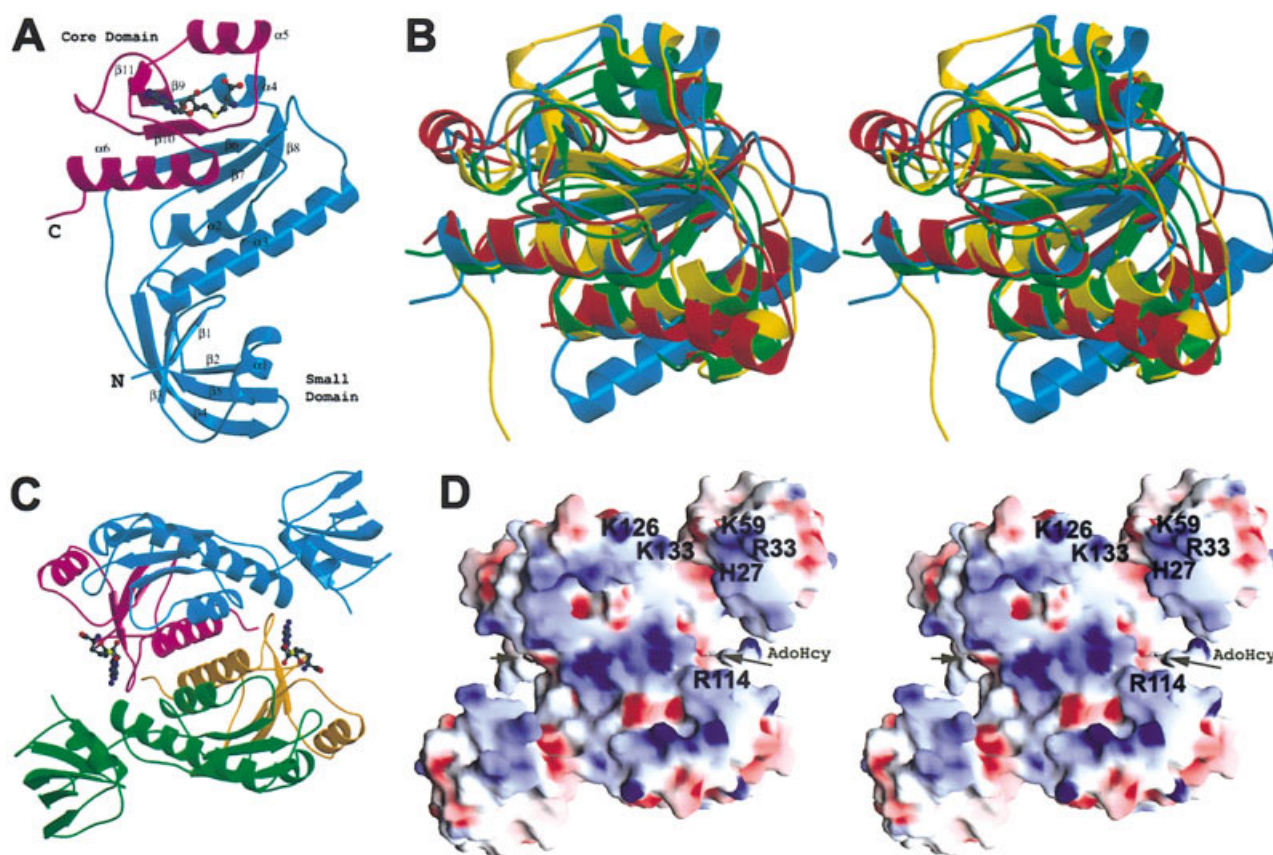Among the structural homologs of HI0303, Rrma and

Fig. 2. Structure of HI0303. **a:** Ribbon representation of the HI0303 monomer. The trefoil knot is depicted in magenta. An AdoHcy inhibitor, modeled into its putative binding site, is shown as a ball-and-stick model. **b:** Stereoview showing structural overlay of the core domain of HI0303 (cyan) and those of other knot-containing RNA MTases, color coded as follows: RrmA (green, accession code 1ipa), MT1 (red, 1k3r), and YibK (gold, 1MXI). **c:** Schematic drawing of the HI0303 dimer. One monomer is colored as Figure 1(a) and the other as green and gold (knot). Two modeled AdoHcy molecules are shown as ball-and-stick models. **d:** Stereoview showing surface electrostatic potential of the HI0303 dimer. The view is identical to Figure 2(c), and the modeled AdoHcy molecules are labeled. Blue and red colors represent positively charged and negatively charged residues. (a)–(c) were prepared by Raster3D,[18] and GRASP[19] was used for generating (d).

MT1 have been classified as RNA 2′-O-ribose MTases. The gene for MT1 is located in an operon for ribosomal proteins, whereas Rrma contains the three conserved motifs[1,2] that have been identified for some members of this family. The first motif corresponds to the β6–α2 segment in the structure of HI0303, but the conformation of residues in this region is different among the enzymes [Fig. 2(B)]. The second and third motifs correspond to the β10–α5 and β11–α6 segments, respectively, which comprises the knot of the core domain. These are the same segments that are highly conserved among the HI0303 family members and may interact with the AdoMet molecule. However, the sequence homology between HI0303 and Rrma for the residues in these two motifs is very low.

Our observation that the small domain of HI0303 shares structural similarity with a RNA-binding domain leads to the suggestion that HI0303 may also be an RNA MTase. This hypothesis is supported by an examination of the electrostatic surface features of the HI0303 dimer [Fig. 2(D)]. There is a long groove on the surface of the dimer, which is surrounded from two sides by many highly conserved basic residues [Fig. 2(D)]. Notably, the AdoHcy-binding site is in close proximity to a cluster of conserved basic residues from the small domain, His27, Arg33, Lys59, and one residue from the interacting monomer (Arg214) [Fig. 1(D)], which may mediate the positioning of the RNA substrate near the active site. Therefore, our structural analyses suggest the hypothesis that HI0303 and its homologs are RNA 2′-O-ribose methyltransferases.

Materials and Methods. The full length IR73 (*yggJ*) gene from *Haemophilus influenzae* was cloned into a pET21d (Novagen) derivative, generating plasmid pIR73-21-1. The resulting IR73 open reading frame contains 11 non-native residues (AAALEHHHHHH) at the C-terminus of the protein. *Escherichia coli* BL21 (DE3) pMGK cells, a rare codon enhanced strain, were transformed with pIR73-21-1. A single isolate was cultured in either LB for native protein or MJ9 minimal media[11] containing selenomethionine, lysine, phenylalanine, threonine, isoleucine, leucine, and valine, for selenomethionyl IR73.[12] Initial growth was carried out at 37°C until the OD_{600} of the culture reached 1.0 units. The incubation temperature was then decreased to 17°C, and protein expression was induced by the addition of IPTG (isopropyl-β-D-thiogalactopyranoside) at a final concentration of 1 mM. After

overnight incubation at 17°C, the cells were harvested by centrifugation.

Native and selenomethionyl IR73 were purified by standard methods. Cell pellets were resuspended in lysis buffer [50 mM NaH_2PO_4 300 mM NaCl, 10 mM imidazole, and 5 mM β-mercaptoethanol (pH 8.0)] and disrupted by sonication. The resulting lysate was clarified by centrifugation at 26,000 × g for 45 min at 4°C. The supernatant was loaded onto an Ni-NTA column (Qiagen) and eluted in lysis buffer containing 250 mM imidazole. Fractions containing the partially purified IR73 were pooled and loaded onto a gel filtration column (Superdex 75, Amersham Biosciences) and eluted in buffer 2 [10 mM Tris, 5 mM DTT, 100 mM NaCl (pH 8.0)]. The resulting purified IR73 protein was concentrated to 10 mg/mL, and sample purity (>97%) and molecular weight (28.6 kDa) were verified by SDS-PAGE and MALDI-TOF mass spectrometry, respectively. The yield of purified protein was ~50 mg/L.

The SeMet HI0303 was crystallized in 10 mM Tris (pH 7.5), 20% PEG3350, 200 mM potassium thiocyanate, 50 mM NaCl, and 10 mM DTT at 22°C using the hanging drop vapor diffusion method. A single-wavelength anomalous diffraction (SAD) data set was collected at the peak absorption wavelength of selenium to 2 Å resolution at the X4A beamline of NSLS. The crystal belongs to $P2_12_12_1$ space group with cell parameters of $a = 61.3$ Å, $b = 73.5$ Å, and $c = 109.5$ Å. There are two molecules in the asymmetric unit. After data processing with the HKL package (Table I),[13] 8 of 10 possible selenium (Se) sites were found by SnB.[14] The Se sites were then used in SOLVE[15] for phasing and automated model building, which correctly placed 50% of the residues. It is of interest that most of the residues in the knot were built incorrectly by this procedure, but it was possible to readjust the model and trace the remaining residues of the dimer by XtalView[16] as a consequence of a clear map. A complete model was built during the refinement with twofold noncrystallographic restraint using CNS.[17]

REFERENCES

1. Kagan RM, Clarke S. Widespread occurrence of three sequence motifs in diverse S-adenosylmethionine-dependent methyltransferases suggests a common structure for these enzymes. Arch Biochem Biophys 1994;310:417–427.

2. Cavaille J, Chetouani F, Bachellerie JP. The yeast *Saccharomyces cerevisiae* YDL112w ORF encodes the putative 2′-O-ribose methyltransferase catalyzing the formation of Gm18 in tRNAs. RNA 1999;5:66–81.

3. Cheng X, Roberts RJ. AdoMet-dependent methylation, DNA methyltransferases and base flipping. Nucleic Acids Res 2001;29:3784–3795.

4. Nureki O, Shirouzu M, Hashimoto K, Ishitani R, Terada T, Tamakoshi M, Oshima T, Chijimatsu M, Takio K, Vassylyev DG, Shibata T, Inoue Y, Kuramitsu S, Yokoyama S. An enzyme with a deep trefoil knot for the active-site architecture. Acta Crystallogr D Biol Crystallogr 2002;58:1129–1137.

5. Zarembinski TI, Kim Y, Peterson K, Christendat D, Dharamsi A, Arrowsmith CH, Edwards AM, Joachimiak A. Deep trefoil knot implicated in RNA binding found in an archaebacterial protein. Proteins 2003;50:177–183.

6. Lim K, Zhang H, Tempczyk A, Krajewski W, Bonander N, Toedt J, Howard A, Eisenstein E, Herzberg O. Structure of the YibK methyltransferase from Haemophilus influenzae (HI0766): a cofactor bound at a site formed by a knot. Proteins 2003;51:56–67.

7. Hendrickson WA. Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. Science 1991;254:51–58.

8. Fedorov RV, Meshcheryakov VA, Gongadze GM, Fomenkova, NP, Nevskaya NA, Selmer M, Laurberg M, Kristensen O, Al-Karadaghi S, Liljas A, Garber MB, Nikonov SV. Structure of ribosomal protein TL5 complexed with RNA provides new insights into the CTC family of stress proteins. Acta Crystallogr 2001;D57:968–

9. Taylor WR. A deeply knotted protein structure and how it might fold. Nature 2000;406:916–919.

10. Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. J Mol Biol 1999;285:2177–2198.

11. Jansson M, Li Y-C, Jendeberg L, Anderson S, Montelione GT, Nilsson BJ. High level production of uniformly ^{15}N- and ^{13}C-enriched fusion proteins in *Escherichia coli*. Biomol NMR 1996;7:131–141.

12. Doublie S, Kapp U, Aberg A, Brown K, Strub K, Cusack S. Crystallization and preliminary X-ray analysis of the 9 kDa protein of the mouse signal recognition particle and the selenomethionyl-SRP9. *FEBS Lett* 1996;384:219–221.

13. Otwinowski Z, Minor W. Processing of X-ray diffraction data collected in oscillation mode. Methods Enzymol 1997;276:307–326.

14. Weeks CM, Miller R. The design and implementation of SnB v2.0. J Appl Crystallogr 1999;32:120–124.

15. Terwilliger TC, Berendzen J. Automated structure solution for MIR and MAD. Acta Crystallogr 1999;D55:849–861.

16. McRee DE. XtalView/Xfit—a versatile program for manipulating atomic coordinates and electron density. J Struct Biol 1999;125:156–165.

17. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL. Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr 1998;D54:905–921.

18. Merritt EA, Bacon DJ. Raster3D: photorealistic molecular graphics. Methods Enzymol 1997;277:505–524.

19. Nicolls A, Sharp KA, Honig B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. Proteins 1991;11:281–296.