Acta Crystallographica Section F
Structural Biology and Crystallization Communications
ISSN 1744-3091

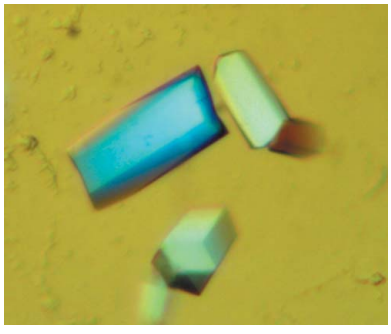
Preliminary X-ray analysis of XC5848, a hypothetical ORFan protein with an Sm-like motif from *Xanthomonas campestris*

Sz-Kai Ruan,[a] Ko-Hsin Chin,[a] Hui-Lin Shr,[b,c] Ping-Chiang Lyu,[d] Andrew H.-J. Wang[b,c] and Shan-Ho Chou[a]*

[a]Institute of Biochemistry, National Chung-Hsing University, Taichung 40227, Taiwan, [b]Institute of Biological Chemistry, Academia Sinica, Nankang, Taipei, Taiwan, [c]Core Facility for Protein Crystallography, Academia Sinica, Nankang, Taipei, Taiwan, and [d]Department of Life Science, National Tsing Hua University, Hsin-Chu, Taiwan

Correspondence e-mail: shchou@nchu.edu.tw

Received 11 October 2006
Accepted 6 December 2006



XC5848, a hypothetical protein from the pathogenic bacterium *Xanthomonas campestris* that causes black rot, has been chosen as a potential target for the discovery of novel folds. It is unique to the *Xanthomonas* genus and has significant sequence identity mainly to corresponding proteins from the *Xanthomonas* genus. In this paper, the cloning, overexpression, purification and crystallization of the XC5848 protein are reported. The XC5848 crystals diffracted to a resolution of at least 1.68 Å. They belong to the orthorhombic space group $P2_12_12_1$, with unit-cell parameters a = 48.13, b = 51.62, c = 82.32 Å. Two molecules were found in each asymmetric unit. Preliminary structural studies nevertheless indicate that XC5848 belongs to the highly conserved Sm-like α-β-β-β-β fold. However, significant differences in sequence and structure were observed. It therefore represents a novel variant of the crucial Sm-like motif that is heavily involved in mRNA splicing and degradation.

1. Introduction

High-throughput genome sequencing has resulted in the determination of an avalanche of genome sequences. To date, more than 300 complete microbial genomes have been deposited at the NCBI website. In the post-genomic era, annotation of these genomic data has become one of the focuses of contemporary life science. Although database mining using programs such as *BLAST* and *PSI-BLAST* (Altschul *et al.*, 1997) has played an important role in annotating unknown protein functions, such methods cannot generally be used to assign the function of proteins bearing less than 30% identity to the query sequences. Of the 4182 ORFs published in the *Xanthomonas campestris* pv. *campestris* (Xcc) ATCC 33913 genome, for example, 1474 ORFs have no assigned function, including 1276 so-called conserved hypothetical proteins that are present across bacterial genera and 198 so-called hypothetical proteins that are only detected in the Xcc genus (da Silva *et al.*, 2002). Elucidation of these protein functions therefore necessitate a different approach.

The structural genomics programme has emerged as a powerful approach to help to solve this problem. It exploits the idea that three-dimensional protein structures are subject to less evolutionary change than protein sequences and can therefore serve to reveal the structures and functions of proteins with less than 30% sequence identity (Zarembinski *et al.*, 1998; Shin *et al.*, 2002; Pal & Eisenberg, 2005). The obtained structural information can disclose the fold, motif, domain or functionally significant residues and finally lead to the revelation of unknown protein function.

XC5848 (gi|21114923) is a hypothetical protein from a local strain of Xcc (http://xcc.life.nthu.edu.tw/). It contains 102 amino acids and has a molecular weight of 11 945 Da. A *BLAST* search against the PDB revealed no similar tertiary structures for XC5848; hence, it serves as a potential target for the discovery of a novel fold. Preliminary structural studies nevertheless indicate that it belongs to the well conserved Sm-like α-β_1-β_2-β_3-β_4 fold (Kambach *et al.*, 1999). However, substantial differences from the Sm-like motif were observed; XC5848 contains a long structured N-terminal region, with a much-extended β_2-β_3 hairpin motif. The N-terminal α-helix is longer and the loop between the β_3 and β_4 strands is shorter. Since Sm-like proteins have been found to associate with small nuclear RNAs to form the core of the small nuclear ribonucleoproteins

Table 1
Data-collection statistics for native XC5848.

Values in parentheses are for the highest resolution shell.

Space group	$P2_12_12_1$
Unit-cell parameters (Å, °)	a = 48.12, b = 51.62, c = 82.31, $\alpha = \beta = \gamma = 90$
Temperature (K)	100
Wavelength (Å)	0.96408
Resolution range (Å)	50–1.68 (1.75–1.68)
Unique reflections	264403 (14214)
Redundancy	10.2 (10.0)
Mosaicity (°)	0.3
Completeness (%)	96.6 (98.7)
R_{merge} (%)	4.5 (25.0)
Mean $I/\sigma(I)$	39 (9.3)
Solvent content† (%)	50.1
Cryoprotectant	Mother liquor

† The solvent content was obtained using the method of Matthews (1968).

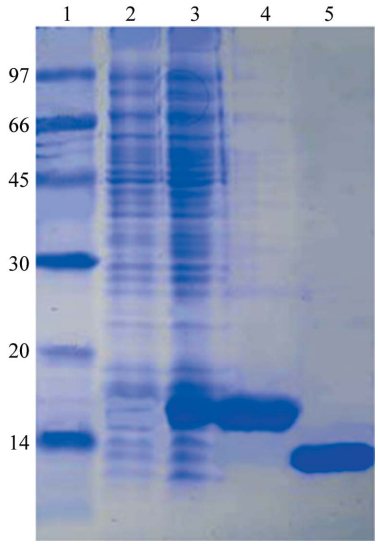


Figure 1
SDS–PAGE monitoring of the overexpression and purification of XC5848. Lane 1, molecular-weight markers (kDa). Lane 2, whole cell lysate before IPTG induction. Lane 3, whole cell lysate after IPTG induction. Lane 4, purified XC5848 before TEV cleavage. Lane 5, purified XC5848 after TEV cleavage. The location is close to the expected molecular weight of 11 945 Da.

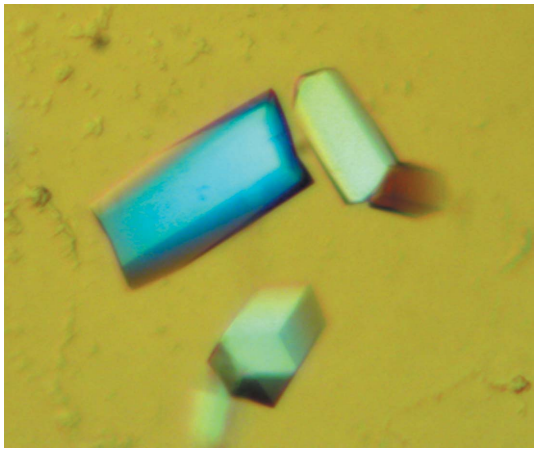


Figure 2
Crystallization of Se-XC5848 from Xcc. Crystals of Se-XC5848 were grown by the hanging-drop vapour-diffusion method under the final optimized crystallization conditions of 0.1 *M* sodium citrate pH 5.6, 0.2 *M* ammonium acetate and 30%(*w*/*v*) PEG 4K. After one week, the approximate dimensions of these crystals were 0.05 × 0.05 × 0.15 mm.

required for diverse processes such as pre-mRNA splicing (Kambach *et al.*, 1999) and mRNA degradation (Pannone & Wolin, 2000) in the archaea (Toro *et al.*, 2001, 2002; Mura *et al.*, 2001; Collins *et al.*, 2001) and eukarya (Kambach *et al.*, 1999; Collins *et al.*, 2003) and also bind to uridine-rich tracks on regulatory sRNAs such as OxyS (Zhang *et al.*, 2002), Spot42 (Moller *et al.*, 2002) or DsrA (Brescia *et al.*, 2003) to act as the Hfq protein of the RNA chaperone (Moll *et al.*, 2003) in the bacteria (Sauter *et al.*, 2003; Schumacher *et al.*, 2002), their structural studies are thus crucial to understanding their functional roles in RNA metabolism. Interestingly, while Sm-like proteins in eukarya or archaea host several types of heteroheptamers to accomplish various functions (Toro *et al.*, 2001; Kambach *et al.*, 1999; Mura *et al.*, 2001; Collins *et al.*, 2003; Urlaub *et al.*, 2001; Achsel *et al.*, 1999), bacterial Sm-like protein complexes are simpler and mainly adopt a homohexamer architecture (Sauter *et al.*, 2003; Schumacher *et al.*, 2002). XC5848 thus turns out to be an interesting target for structural and functional studies.

2. Materials and methods

2.1. Cloning, expression and purification

The XC5848 gene fragment was PCR-amplified directly from a local Xcc genome (*X. campestris* pv. *campestris* strain 17) with forward primer 5′-TACTTCCAATCCAATGCTATGCCCAAA-TACGCCCCCCA and reverse primer 5′-TTATCCACTTCC-AATGTCAGGGCATGACTTGCGGGTC. A ligation-independent cloning (LIC) approach according to a previously published protocol (Wu *et al.*, 2005) was carried out in order to obtain the desired construct. The final construct codes for an N-terminal His_6 tag, a 17-amino-acid linker and the XC5848 target protein (102 amino acids) under the control of a T7 promoter. Overexpression of the Hig_6-tagged target protein was induced by the addition of 1 m*M* IPTG at 293 K for 20 h. The target protein was purified by immobilized metal-affinity chromatography (IMAC) on a nickel column (Sigma). The His_6 tag and linker were cleaved from XC5848 by TEV (tobacco etch virus) protease at 288 K for 24 h. For crystallization, XC5848 was further purified using an AKTA anion-exchange column (Pharmacia Inc.). The final target protein (102 amino acids) has greater than 99% purity (Fig. 1) and contains only an extra tripeptide (SNA) at the N-terminal end. The overexpression and purification of XC5848 was monitored by SDS–PAGE as shown in Fig. 1.

2.2. Crystallization

For crystallization, the protein was concentrated to 28 mg ml^{-1} in 20 m*M* Tris pH 8.0 and 80 m*M* NaCl using an Amicon Ultra-10 (Millipore). Screening for crystallization conditions was performed using sitting-drop vapour diffusion in 96-well plates (Hampton Research) at 295 K by mixing 0.5 µl protein solution with 0.5 µl reagent solution. Initial screens, including the Hampton sparse-matrix Crystal Screens 1 and 2, a systematic PEG–pH screen and the PEG/Ion Screen, were performed using a Gilson C240 crystallization workstation. Hexagonal pillar-shaped crystals appeared in 3 d from a reservoir solution comprising 0.1 *M* sodium citrate pH 5.6, 0.2 *M* ammonium acetate and 30%(*w*/*v*) PEG 4K. Crystals suitable for diffraction experiments were grown by mixing 1.5 µl protein solution with 1.5 µl reagent solution at 295 K and reached maximum dimensions of 0.05 × 0.05 × 0.15 mm after one week (Fig. 2).

2.3. Data collection

Crystals were flash-cooled without cryoprotectant at 100 K in a stream of cold nitrogen to prevent crystal cracking. X-ray diffraction

data were collected using the National Synchrotron Radiation Research Center (NSRRC, Taiwan) beamline 13B1 and a Q315 area detector. A data set was obtained to 1.68 Å resolution for native XC5848. The data were indexed and integrated using the *HKL*-2000 processing software (Otwinowski & Minor, 1997), giving a data set that was 96.6% complete with an overall $R_{\rm merge}$ of 4.5%. The crystals belong to the orthorhombic space group $P2_12_12_1$. The data-collection statistics are summarized in Table 1.

3. Results and discussion

The gene sequence of XC5848 was confirmed after cloning and consists of 309 bp coding for 102 amino-acid residues. The purified XC5848 contains only an extra SNA tripeptide at the N-terminal end and is greater than 99% pure, with a single band of approximately 12 kDa on SDS–PAGE (Fig. 1), which is in good agreement with the calculated molecular weight of 11 673.21 Da. It has a theoretical pI of 4.87 calculated using the *Compute pI/MW* tool at the *ExPASy* website (http://us.expasy.org/tools/pi_tool.html).

Sm and Sm-like (Lsm) proteins, which share two sequence motifs Sm1 and Sm2, were first discovered in the eukarya (Kambach *et al.*, 1999) and archaea (Collins *et al.*, 2001; Toro *et al.*, 2001; Mura *et al.*, 2001). The motifs were found to contain hydrophobic residues that maintain the core of the Sm-fold (Kambach *et al.*, 1999) and highly conserved residues involved in specific RNA binding (Toro *et al.*, 2001). Recently, a comprehensive *BLAST* search against 140

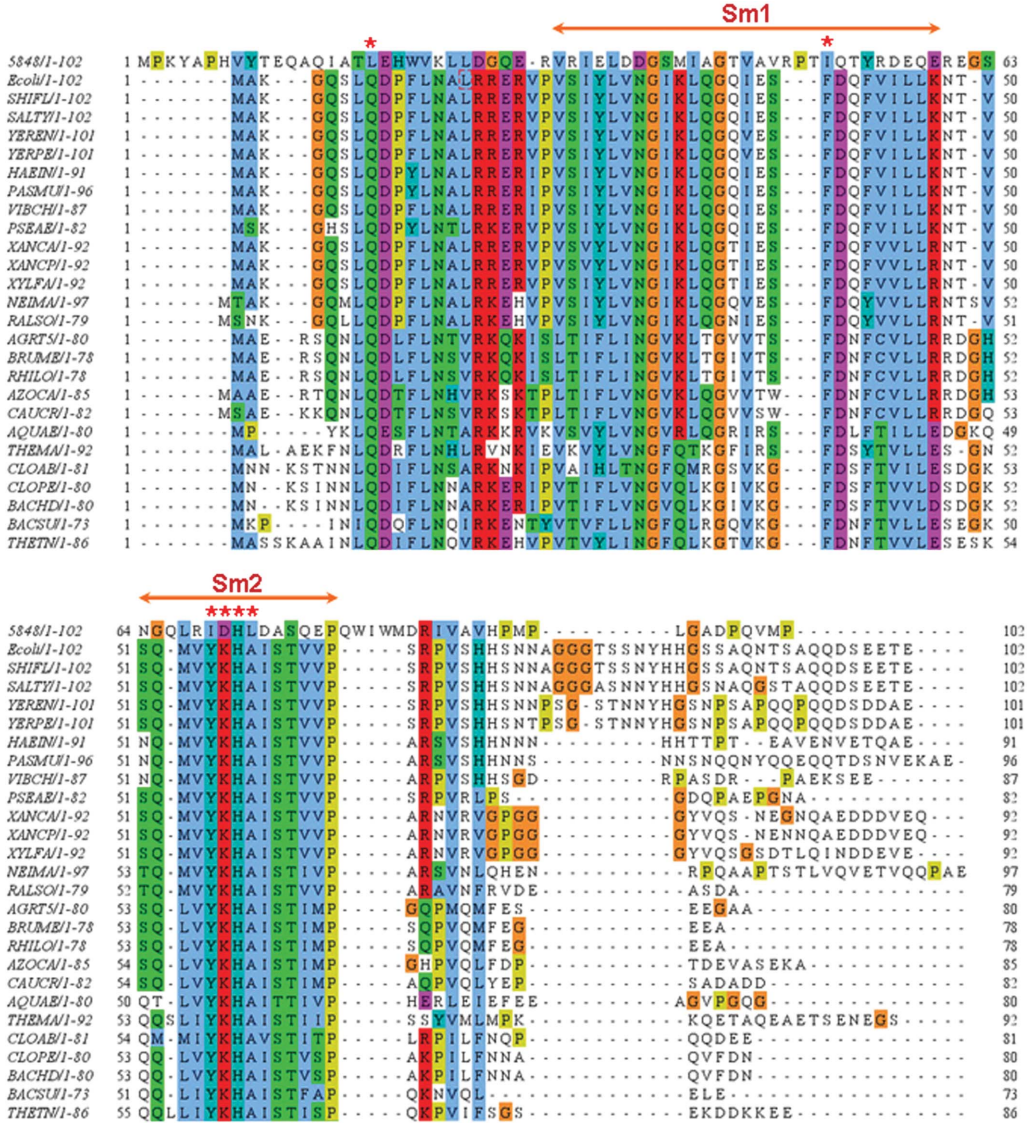


Figure 3
Multiple sequence alignment of Hfq proteins using *ClustalW* (Thompson *et al.*, 1994). The organisms corresponding to the sequences are indicated on the left along with their lengths, except for the top sequence 5848 from Xcc. The proteins are from *E. coli* (ECOLI), *Shigella flexneri* (SHIFL), *Salmonella typhimurium* (SALTY), *Yersinia enterocolitica* (YEREN), *Yersinia pestis* (YERPE), *Erwinia carotovora* (ERWCA), *Haemophilus infuenzae* (HAEIN), *Pasteurella multocida* (PASMU), *Vibrio cholerae* (VIBCH), *Pseudomonas aeruginosa* (PSEAE), *Xanthomonas axonopodis* (XANAC), *Xanthomonas campestris* (XANCP), *Xylella fastidiosa* (XYLFA), *Neisseria meningitidis* (NEIMA), *Ralstonia solanacearum* (RALSO), *Agrobacterium tumefaciens* (AGRT5), *Brucella melitensis* (BRUME), *Rhizobium loti* (RHILO), *Azorhizobium caulinodans* (AZOCA), *Caulobacter crescentus* (CAUCR), *Aquifex aeolicus* (AQUAE), *Thermotoga maritima* (THEMA), *Clostridium acetobutylicum* (CLOAB), *Clostridium perfringens* (CLOPE), *Bacillus halodurans* (BACHD), *Bacillus subtilis* (BACSU) and *Thermoanaerobacter tengcongensis* (THETN). The conserved Sm1 and Sm2 motifs in the Hfq proteins are annotated with double-headed arrows. Conserved polar, basic and acidic residues are marked in green, red and pink, respectively, and Gly in orange and Pro in yellow; a star indicates those residues involved in RNA binding (Sauter *et al.*, 2003; Schumacher *et al.*, 2002). Blue boxes are conserved patches of hydrophobic residues.

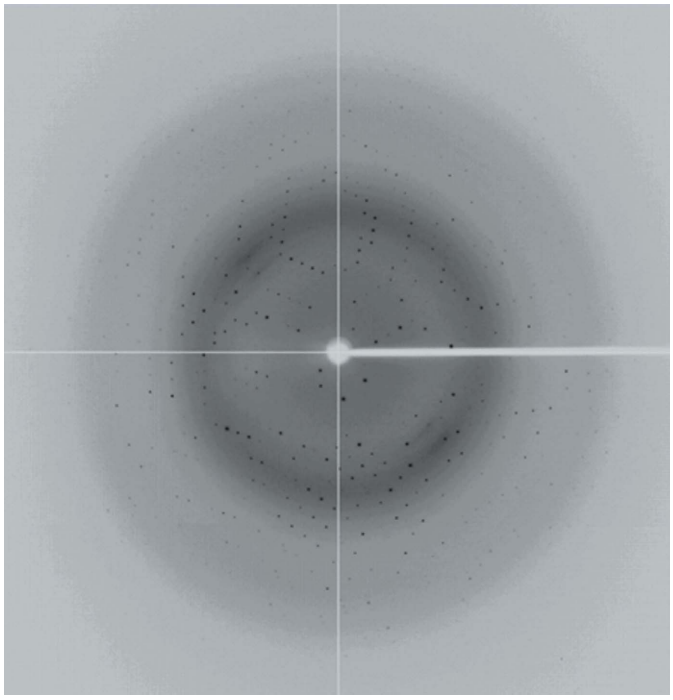


Figure 4
The X-ray pattern of a Se-substituted XC5848 crystal diffracting to a resolution of 2.0 A. The phases were successfully determined from this data set.

complete or nearly complete microbial genomes using the *Escherichia coli* sequence as a query suggested that the Hfq protein may be the Sm analogue in bacteria (Sun *et al.*, 2002). Because our preliminary structural analysis of XC5848 indicates that it resembles the Hfq fold, we have performed a multiple sequence alignment of XC5848 with other bacterial Hfq proteins using the *ClustalW* program (Thompson *et al.*, 1994) as shown in Fig. 3. Surprisingly, the sequence identities are rather low, ranging from 12% to 25%, and the specific RNA-binding residues are not even present in XC5848. Significant sequence changes were detected throughout the entire sequence; the highly conserved *E. coli* Hfq sequence has been changed from ^{16}RRER^{19} to ^{26}DGQE^{29}, the hydrophobic patch from ^{42}FVILL^{46} to ^{54}YRDEQ^{58} and the characteristic Sm2 signature sequence (Sauter *et al.*, 2003; Schumacher *et al.*, 2002; Sun *et al.*, 2002) from ^{55}YKHA^{58} to ^{69}IDHL^{72}. This information indicates that XC5848 is unlikely to be a Hfq protein, but that through divergent evolution its gene has been duplicated (Xcc also has a separate Hfq gene as shown in Fig. 3) and appended with two sequences at the N-terminus and between the β2 and β3 regions to adopt an Lsm-motif variant that perhaps performs other biological functions. The structure determination of XC5848 is thus crucial to understanding the function of this novel Hfq-protein analogue.

A three-wavelength MAD data set to 2.0 Å resolution has been collected at the remote, peak and reflection points of Se absorption using beamline 13B1 at the NSRRC, Taiwan (Fig. 4). A good preliminary structure of XC5848 has been obtained by the MAD approach using the *SOLVE* and *RESOLVE* programs (Hendrickson & Ogata, 1997; Terwilliger & Berendzen, 1999). The starting structure was refined against the native data set to 1.68 Å resolution (Table 1). Detailed structural refinement of this interesting protein is currently ongoing.

This work was supported by an Academic Excellence Pursuit grant from the Ministry of Education and the National Science Council, Taiwan to S-HC. We also thank the Core Facilities for Protein X-ray Crystallography in the Academia Sinica, Taiwan and the National Synchrotron Radiation Research Center, Taiwan for assistance during X-ray data collection. The National Synchrotron Radiation Research Center is a user facility supported by the National Science Council, Taiwan and the Protein Crystallography Facility is supported by the National Research Program for Genomic Medicine, Taiwan.

References

Achsel, T., Brahms, H., Kastner, B., Bachi, A., Wilm, M. & Luhrmann, R. (1999). *EMBO J.* **18**, 5789–5802.
Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). *Nucleic Acids Res.* **25**, 3389–3402.
Brescia, C. C., Mikulecky, P. J., Feig, A. L. & Sledjeski, D. D. (2003). *RNA*, **9**, 23–30.
Collins, B. M., Cubeddu, L., Naidoo, N., Harrop, S. J., Kornfeld, G. D., Dawes, I. W., Curmi, P. M. G. & Mabbutt, B. C. (2003). *J. Biol. Chem.* **278**, 17291–17298.
Collins, B. M., Harrop, S. J., Kornfeld, G. D., Dawes, I. W., Curmi, P. M. G. & Mabbutt, B. C. (2001). *J. Mol. Biol.* **309**, 915–923.
da Silva, A. C. R. *et al.* (2002). *Nature (London)*, **417**, 459–463.
Hendrickson, W. A. & Ogata, C. M. (1997). *Methods Enzymol.* **276**, 494–523.
Kambach, C., Walke, S., Young, R., Avis, J. M., de La Fortelle, E., Raker, V. A., Luhrmann, R., Li, J. & Nagai, K. (1999). *Cell*, **96**, 375–387.
Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
Moll, I., Leitsch, D., Steinhauser, T. & Blasi, U. (2003). *EMBO Rep.* **4**, 284–289.
Moller, T., Franch, T., Hojrup, P., Keene, D. R., Bachinger, H. P., Brennan, R. G. & Valentin-Hansen, P. (2002). *Mol. Cell*, **9**, 23–30.
Mura, C., Cascio, D., Sawaya, M. R. & Eisenberg, D. S. (2001). *Proc. Natl Acad. Sci. USA*, **98**, 5532–5537.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Pal, D. & Eisenberg, D. (2005). *Structure*, **13**, 121–130.
Pannone, B. K. & Wolin, S. L. (2000). *Curr. Biol.* **10**, R478–R481.
Sauter, C., Basquin, J. & Suck, D. (2003). *Nucleic Acids Res.* **31**, 4091–4098.
Schumacher, M. A., Perarson, R. F., Moller, T., Valentin-Hansen, P. & Brennan, R. G. (2002). *EMBO J.* **21**, 3546–3556.
Shin, D. H., Yokota, H., Kim, R. & Kim, S.-H. (2002). *Proc. Natl Acad. Sci. USA*, **99**, 7980–7985.
Sun, X., Zhulin, I. & Wartell, R. M. (2002). *Nucleic Acids Res.* **30**, 3662–3671.
Terwilliger, T. C. & Berendzen, J. (1999). *Acta Cryst.* D**55**, 849–861.
Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). *Nucleic Acids Res.* **22**, 4673–4680.
Toro, I., Basquin, J., Teo-Dreher, H. & Suck, D. (2002). *J. Mol. Biol.* **320**, 129–142.
Toro, I., Thore, S., Mayer, C., Basquin, J., Seraphin, B. & Suck, D. (2001). *EMBO J.* **20**, 2293–2303.
Urlaub, H., Raker, V. A., Kostka, S. & Luhrmann, R. (2001). *EMBO J.* **20**, 187–196.
Wu, Y.-Y., Chin, K.-H., Chou, C.-C., Lee, C.-C., Shr, H.-L., Lyu, P.-C., Wang, A. H.-J. & Chou, S.-H. (2005). *Acta Cryst.* F**61**, 902–905.
Zarembinski, T. I., Hung, L.-W., Mueller-Dieckmann, H.-J., Kim, K.-K., Yokota, H., Kim, R. & Kim, S.-H. (1998). *Proc. Natl Acad. Sci. USA*, **95**, 15189–15193.
Zhang, A., Wassarman, K. M., Ortega, J., Steven, A. C. & Storz, G. (2002). *Mol. Cell*, **9**, 11–22.