crystallization communications

Ankit Gupta and Clara L. Kielkopf*

Department of Biochemistry and Biophysics, University of Rochester School of Medicine and Dentistry, Rochester, NY 14642, USA

Correspondence e-mail: clara_kielkopf@urmc.rochester.edu

Purification, crystallization and preliminary X-ray crystallographic analysis of a central domain of human splicing factor 1

Pre-mRNA splicing is an essential source of genetic diversity in eukaryotic organisms. In the early stages of splicing, splicing factor 1 (SF1) recognizes the pre-mRNA splice site as a complex with its partner, U2 auxiliary factor 65 kDa subunit (U2AF[65]). A central 'mystery' domain of SF1 (SF1md) lacks detectable homology with known structures, yet is the region of highest phylogenetic sequence conservation among SF1 homologues. Here, steps towards determining the SF1md structure are described. Firstly, SF1md was expressed and purified. The presence of regular secondary structure was verified using circular dichroism spectroscopy and the SF1md protein was then crystallized. A native data set was collected and processed to 2.5 Å resolution. The SF1md crystals belonged to space group $C2$ and have most probable solvent contents of 64, 52 or 39% with three, four or five molecules per asymmetric unit, respectively. Mutually perpendicular peaks on the $\kappa = 180°$ section of the self-rotation function support the presence of four molecules in the asymmetric unit.

1. Introduction

Pre-mRNA splicing serves as an essential source of mRNA diversity in higher eukaryotes by joining discontinuous regions of gene transcripts in alternative patterns (Wang *et al.*, 2008). The fidelity of pre-mRNA splice-site choice is critical, since an error of even a single nucleotide can be lethal or lead to human genetic diseases (reviewed in Wang & Cooper, 2007). A dynamic ribonucleoprotein complex called the spliceosome assembles at consensus pre-mRNA sequences, where it removes the introns and ligates the exons into mature mRNA by two successive transesterification reactions (Weiner, 1993; Michel & Ferat, 1995). A branch-point sequence (BPS) located near the 3' splice site contains an adenosine that ultimately serves as the nucleophile in the splicing reaction (Padgett *et al.*, 1984). The BPS usually accompanies a polypyrimidine tract (Irimia & Roy, 2008) and the proximity of the two sequences is important for splice-site identification (Reed, 1989).

Splicing factor 1 (SF1) recognizes the BPS in the early stages of pre-mRNA splicing. A tight complex between SF1 and U2 auxiliary factor 65 kDa subunit (U2AF[65]) undergoes conformational changes that couple BPS with polypyrimidine-tract recognition (Gupta *et al.*, 2011). In turn, the SF1/U2AF[65]/pre-mRNA complex facilitates stable association of the small nuclear ribonucleoprotein particles to form the active spliceosome (Wahl *et al.*, 2009). The importance of SF1 in early spliceosomal assembly is emphasised by its requirement for embryonic development in mice and viability in human cells and yeast (Shitashige *et al.*, 2007; Tanackovic & Krämer, 2005; Abovich & Rosbash, 1997). Depletion of SF1 from yeast or human extracts decreases the efficiency of spliceosome assembly on traditional splicing substrates (Guth & Valcárcel, 2000; Rutz & Séraphin, 1999). More recently, SF1 has been shown to be required for alternative splicing of transcripts encoding fibroblast growth factor receptor 1 oncogene partner, TNFAIP3-interacting protein 1, procollagen-lysine 1, 2-oxoglutarate 5-dioxygenase 2 and UPF3 regulator of nonsense transcript homologue A (Corioni *et al.*, 2010). SF1 also plays roles in pre-mRNA retention in yeast (Rutz & Séraphin, 2000) and transcription activation and elongation in humans (Zhang *et al.*, 1998; Goldstrohm *et al.*, 2001).
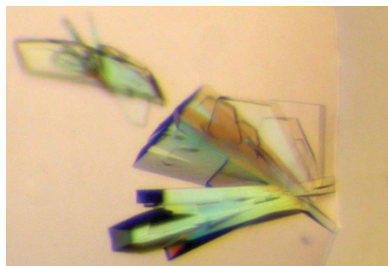
The majority of SF1 domains with known functions in mediating protein–protein or protein–RNA interactions during pre-mRNA splicing belong to well characterized fold families (Fig. 1*a*). A short N-terminal SF1 region termed a 'U2AF ligand motif' (ULM) inserts a conserved tryptophan into a hydrophobic pocket of a C-terminal 'U2AF homology motif' (UHM) of U2AF[65] (Selenko *et al.*, 2003; Kielkopf *et al.*, 2004). A central hnRNP K homology motif and the adjoining quaking 2 (KH-QUA2) region specifically contact the BPS, in particular the key branch-site adenosine (Berglund *et al.*, 1997; Liu *et al.*, 2001). A poorly conserved domain composed of zinc knuckles contributes to the RNA affinity of non-vertebrate SF1 homologues, but is dispensable for RNA binding by vertebrate SF1 homologues (Berglund *et al.*, 1998; Garrey *et al.*, 2006). At the C-terminus, a proline-rich domain of SF1 interacts with WW-domain-containing splicing factors (Abovich & Rosbash, 1997; Goldstrohm *et al.*, 2001; Bedford *et al.*, 1998).

In contrast, the structure and function of an approximately 100-amino-acid region of SF1 between the ULM and the KH-QUA2 domains (residues 26–132) are poorly defined at present. Accordingly, we call this region the SF1 'mystery domain' (SF1md). The SF1md is the most highly conserved region of the protein (48% sequence identity between *Saccharomyces cerevisiae* and human SF1md, compared with 35% identity between the full-length

sequences; Fig. 1*b*). In human cells, serines 80 and 82 of an SF1 'SPSP' motif are phosphorylated by U2AF homology motif kinase 1 (UHMK1; Manceau *et al.*, 2006). Although no clear UHMK1 homologues are evident in *S. cerevisiae*, the SPSP motif of yeast SF1 is conserved and an adjoining PP*x*Y motif has been suggested to mediate interactions with the WW domain of the splicing factor Prp40p (Abovich & Rosbash, 1997). Despite these functional implications, the SF1md lacks detectable homology with known structures. As a step towards determining this potentially novel structure, we report the expression, purification, circular-dichroism (CD) spectrum, crystallization and preliminary X-ray analysis of human SF1md.

2. Materials and methods

2.1. Cloning, expression and purification

A region encoding human SF1 residues 26–132 (NCBI RefSeq NP_004621) was PCR-amplified from our corrected SF1 expression plasmid (Thickman *et al.*, 2006) and inserted between the *Bam*HI and *Eco*RI sites of pGEX-6p-2 (GE Healthcare). The SF1md expression plasmid was introduced into *Escherichia coli* strain BL21 Rosetta 2 cells (Merck). Cell cultures were grown in Luria–Bertani (LB) broth at 310 K to an optical density of 0.6 at 600 nm and then induced at
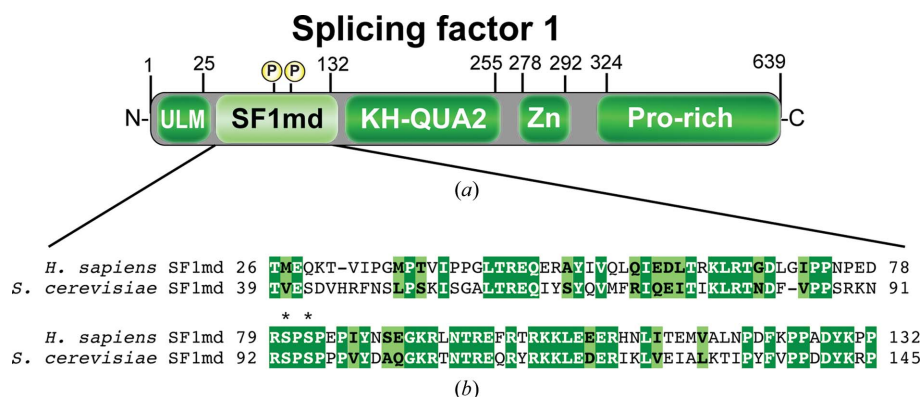

Figure 1
(*a*) Domain boundaries of human SF1. ULM, U2AF ligand motif; KH-QUA2, hnRNP K homology and quaking 2 motif; Zn, zinc knuckle; Pro-rich, proline-rich domain. (*b*) Sequence alignment of human (NCBI RefSeq NP_004621) and *S. cerevisiae* (NCBI RefSeq NP_013217) SF1md. Identical residues are shown with a white font on a dark green background and similar residues are shown with a black font on a light green background. The Ser80 and Ser82 phosphorylation sites are marked with a P in (*a*) and with asterisks in (*b*).
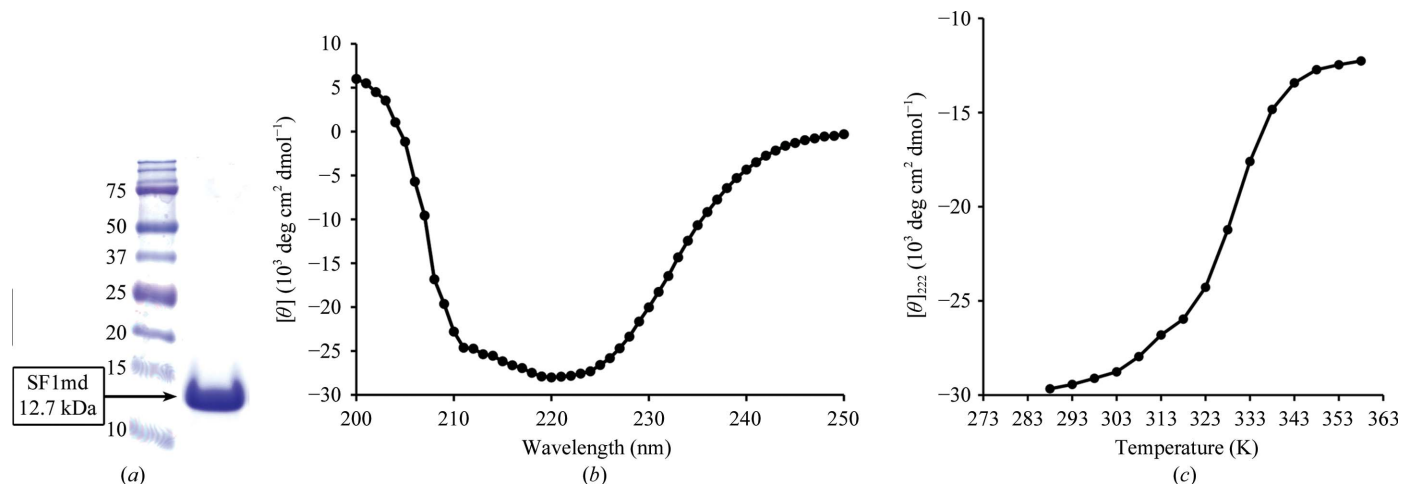

Figure 2
(*a*) SDS–PAGE of purified SF1md following Coomassie Blue staining. (*b*) CD spectrum of SF1md. (*c*) Thermal unfolding transition of SF1md from the mean residue ellipticity at 220 nm plotted as a function of temperature.

303 K for 4 h by the addition of 1 mM isopropyl β-D-1-thiogalacto-pyranoside (IPTG). Cell pellets were collected by centrifugation at 2700g for 20 min, resuspended in lysis buffer [1 M NaCl, 15%(v/v) glycerol, 25 mM HEPES pH 7.4, 0.5 mM EDTA, 1 mM PMSF and a protease-inhibitor cocktail tablet (Roche)], flash-cooled in liquid nitrogen and stored at 200 K until use. Thawed cell suspensions were lysed by sonication on ice following addition of MgCl$_2$, lysozyme and DNAse to final concentrations of 2 mM, 1 mg ml^{-1} and 0.1 mg ml^{-1}, respectively. The proteins were maintained at approximately 277 K throughout the purification steps. The cell lysate was clarified by centrifugation at 34 000g for 30 min and then loaded onto a GSTrap FF column (GE Healthcare) pre-equilibrated with wash buffer (1 M NaCl, 25 mM HEPES pH 7.4). After approximately ten column volumes of wash buffer had returned the absorbance at 280 nm to baseline, the GST-SF1md fusion protein was eluted by exposure to two column volumes of elution buffer (10 mM glutathione, 150 mM NaCl, 100 mM Tris pH 8.0).

The GST tag was cleaved from SF1md by addition of PreScission Protease (GE Healthcare) followed by overnight dialysis against 100 mM NaCl, 25 mM Tris pH 8.0, 5%(v/v) glycerol, 0.5 mM EDTA, 0.1 mM PMSF. Five residues (GPLGS) from the cloning and protease sites remained present at the N-terminus of the recombinant SF1md protein following cleavage. The GST tag, any uncleaved GST-fusion protein and the GST-fused protease were separated from the SF1md by a second pass through a GSTrap column previously equilibrated with dialysis buffer. The majority of contaminant *E. coli* proteins were removed by a subtractive pass through a HiTrap Q FF column (GE Healthcare). The remaining contaminants and any aggregated proteins were removed by a final step of size-exclusion chromato-graphy through a 120 ml Superdex-75 HiLoad column (GE Health-care) equilibrated with gel-filtration buffer (100 mM NaCl, 15 mM HEPES pH 7.4). SDS–PAGE with Coomassie Blue staining was used to monitor the purity of the protein at each step and to verify that the final protein was >98% pure (Fig. 2a).

2.2. Circular dichroism spectroscopy

For CD measurements, the SF1md was dialyzed into a buffer with low absorbance in the UV range [50 mM Na$_2$HPO$_4$ pH 7.4, 0.2 mM tris(2-carboxyethyl)phosphine]. The sample was placed in a quartz cell with 0.1 cm path length and maintained at 298 K by a jacketed cell holder connected to a circulating water bath. The CD spectra were recorded using an Aviv Instruments CD model 202 spectro-polarimeter with 1.0 nm bandwidth and 10 s averaging time. Protein denaturation was followed by the change of the CD signal at 220 nm. All spectra were plotted in units of mean molar residue ellipticity following subtraction of buffer scans (Figs. 2b and 2c).

2.3. Crystallization

The purified SF1md protein was concentrated to 15 mg ml^{-1} in gel-filtration buffer using a 5 kDa molecular-weight cutoff Vivaspin 15 centrifugal concentrator (Sartorius Stedim Biotech) prior to crystal screening. The protein concentration was estimated from the absor-bance at 280 nm and an extinction coefficient of 4470 M^{-1} cm^{-1}, which was calculated using the *ProtParam* tool available on the Expert Protein Analysis System (ExPASy) proteomic server (http://expasy.org/tools/protparam.html). The monodispersity of the final protein solution was checked by dynamic light scattering using a DynaPro MS800 instrument (Wyatt).

Initial crystallization screening was carried out manually at 277 and 293 K using the hanging-drop vapour-diffusion method in 24-well VDX plates (Hampton Research). Crystal Screen (Hampton

Research) and the High Probability Screen (CP-CUSTOM-I; Axygen Inc.) were used for the initial crystallization experiments. A 1.2 µl aliquot of protein solution layered with 1.0 µl well solution was equilibrated against a 600 µl well volume. The initial SF1md crystals were obtained from condition E6 of CP-CUSTOM-I (1.34 M sodium malonate pH 7.0, 0.1 M imidazole maleate pH 6.5) after three weeks at 277 K. After optimizing the precipitant concentration and the pH of the solution, crystals with sharper edges were obtained using 1.34 M sodium malonate pH 6.0, 0.1 M imidazole maleate pH 5.5. A 96-condition additive screen (Hampton Research) identified 3%(w/v) D-(+)-trehalose dihydrate as an additive that improved crystal quality. Nevertheless, crystals continued to grow as stacked plates (Fig. 3a), so single crystals for data collection were isolated by gently dissecting the clusters.

2.4. X-ray data collection and processing

The crystals were cryoprotected by sequential soaks of 1–2 min each in crystallization solution supplemented with 11, 19 or 27%(w/v)
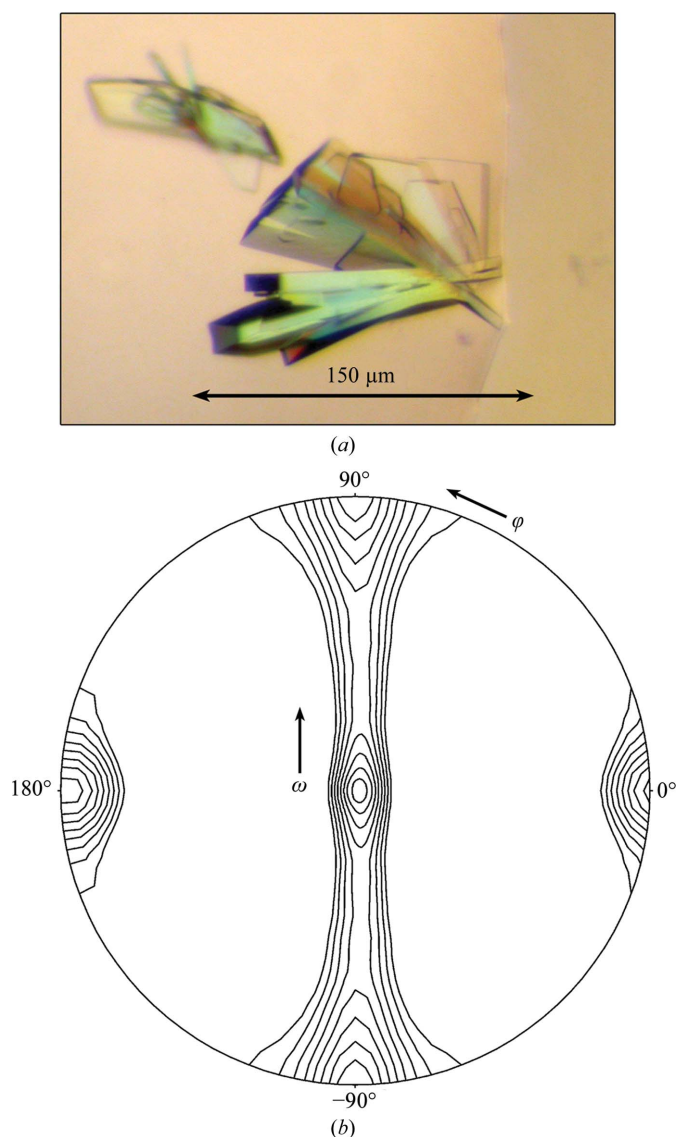

(a)

(b)

Figure 3
(a) Crystals of SF1md grown as described. (b) The κ = 180° section of a self-rotation function contoured from 2 r.m.s. at 0.5 r.m.s. intervals and calculated with an integration radius of 18 Å using data in the 15.0–3.5 Å resolution range.

Table 1
Data-collection and processing statistics.

Values in parentheses are for the highest resolution shell.

No. of crystals	1
Beamline	SSRL 9-2
Wavelength (Å)	0.9794
Detector	MAR 325 CCD
Crystal-to-detector distance (mm)	315
Rotation range per image (°)	0.5
Total rotation range (°)	330
Exposure time per image (s)	10
Resolution range (Å)	48.00–2.48 (2.61–2.48)
Space group	$C2$
Unit-cell parameters (Å, °)	$a = 96.1$, $b = 37.8$, $c = 147.6$, $\beta = 110.8$
Mosaicity (°)	1.1
Total reflections	82410
Unique reflections	16211
Multiplicity	5.1 (5.3)
Mean $I/\sigma(I)$	4.53 (1.90)
Completeness (%)	89.2 (75.4)
R_{merge}† (%)	10.8 (26.0)
R_{meas} (%)	12.0 (41.0)
Overall B factor from Wilson plot (Å^2)	37.2

† $R_{\mathrm{merge}} = \sum_{hkl} \sum_i |I_i(hkl) - \langle I(hkl) \rangle| / \sum_{hkl} \sum_i I_i(hkl)$, where $I_i(hkl)$ is the ith measurement of intensity for a reflection with indices hkl and $\langle I(hkl) \rangle$ is the weighted mean of all intensity measurements for reflection hkl. R_{meas} is R_{merge} corrected for the multiplicity of the observed reflection.

sucrose. A native data set was collected from flash-cooled SF1md crystals with a MAR 325 CCD detector on beamline 9-2 of the Stanford Synchrotron Radiation Lightsource (SSRL; Menlo Park, California, USA) using a wavelength of 0.9794 Å, 0.5° oscillation per 10 s exposure and 315 mm crystal-to-detector distance. Given that the long unit-cell axis naturally fell parallel to the X-ray beam, we adjusted the angle of the cryoloop to resolve the overlapping reflections. The data set was indexed and integrated using *iMOSFLM* (Leslie, 2006) and scaled using *SCALA* (Evans, 2006). All subsequent manipulations were carried out using programs from the *CCP4* software suite (Collaborative Computational Project, Number 4, 1994). We investigated the possibility of noncrystallographic symmetry (NCS) in the asymmetric unit by calculating self-rotation functions with the program *POLARRFN* (Kabsch *et al.*, 1976; Fig. 3*b*). To reduce cross-vectors and retain self-vectors, a 15–20 Å radius of integration was tested based on the theoretical radius of gyration for a 13 kDa globular protein (Putnam *et al.*, 2007).

3. Results and discussion

The highly conserved SF1md was successfully expressed in soluble form as a GST-fusion protein from the pGEX-6p vector. The final protein used for crystallization included residues 26–132 of human SF1 along with five N-terminal residues (GPLGS) from the cloning and protease sites. The *E. coli* BL21 Rosetta 2 host strain facilitated expression by providing tRNAs for rare codons, including those for four arginines, three prolines, two isoleucines and two glycines in the human SF1md coding sequence. The final yield of purified protein was approximately 4 mg per litre of cell culture. The protein was >98% pure by SDS–PAGE (Fig. 2*a*) and was monodisperse by dynamic light scattering (>98% of the scattering was contributed by appropriately sized molecules; data not shown).

A *BLAST* search (Altschul *et al.*, 1990) of human SF1md against the Protein Data Bank failed to identify reliable matches. The closest match for the N-terminal region of SF1md (human SF1 residues 36–69) was a β-strand, loop and α-helix (residues 89–122) from the ribosomal protein L11 methyltransferase (PDB entry 3cjt; E value

2.8; 39% identity for 31% coverage). A short peptide from the C-terminal region of SF1md (human SF1 residues 86–99) matched a turn and β-strand (residues 167–180) from the trigger-factor chaperone (PDB entry 3gty; E value 7.4; 58% identity for 13% coverage). Neither of these matches included the key SPSP motif of SF1 (residues 80–83) that is known to be phosphorylated by UHMK1 (Manceau *et al.*, 2006).

Given the apparent absence of structural homologues for the domain, we used CD spectroscopy to probe whether SF1md adopts a regular secondary structure (Figs. 2*b* and 2*c*). The CD spectrum exhibits features of regular secondary structure, including minima characteristic of α-helices near 208 and 222 nm (Fig. 2*b*). The midpoint of thermal unfolding from changes in the spectra with increasing temperature is approximately 328 K (Fig. 2*c*), a value that is typical of other mesophilic folded proteins (Razvi & Scholtz, 2006). The older programs *PEPCOIL* (Lupas *et al.*, 1991) and *Predict-Protein* (Rost & Sander, 1994) were used in previous studies to predict an α-helical coiled-coil structure for SF1md (Rain *et al.*, 1998). In our hands, the more recent structure-prediction program *I-TASSER* (Roy *et al.*, 2010) also produced α-helical structures for the top-scoring models (C scores of -3.9 to -4.8). Nevertheless, given the presence of significant negative molar ellipticity at 215 nm in the CD spectrum of SF1md, the presence of β-structures remains possible (Greenfield & Fasman, 1969).

With the knowledge in hand that SF1md adopts a stable folded structure, we proceeded to screen crystallization conditions. Following initial screens, a hit was obtained from condition E6 of the High Probability Precipitating Screen CP-CUSTOM I (Axygen Biosciences). The crystals remained small ($0.1 \times 0.05 \times 0.025$ mm) with a stacked-plate morphology (Fig. 3*a*) despite the optimized conditions [1.34 *M* sodium malonate pH 6.0, 0.1 *M* imidazole maleate pH 5.5, 3%(*w/v*) D-(+)-trehalose dihydrate in hanging-drop format at 277 K]. Regardless, we were able to tease single-plate crystals from the clustered growth and collect a 2.5 Å resolution native data set (Table 1).

The SF1md crystals belonged to space group $C2$, with unit-cell parameters $a = 96.1$, $b = 37.8$, $c = 147.6$ Å, $\beta = 110.8°$. The Matthews coefficients (Matthews, 1968) indicated that the presence of three, four or five molecules in the asymmetric unit would result in solvent contents of 64, 52 or 39%, respectively, for the crystal. The $\kappa = 180°$ section of the self-rotation function supported the presence of four molecules per asymmetric unit, based on three mutually perpendicular twofold rotation axes. These included two noncrystallographic axes (96% of the origin peak) and the crystallographic twofold along the b axis (Fig. 3*b*). Given the absence of SF1md-related structures to use as search models for molecular replacement, we are currently preparing crystals of selenomethionine-labelled SF1md and plan to determine this structure using multiwavelength anomalous dispersion methods.

crystallization communications

References

Abovich, N. & Rosbash, M. (1997). *Cell*, **89**, 403–412.
Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). *J. Mol. Biol.* **215**, 403–410.
Bedford, M. T., Reed, R. & Leder, P. (1998). *Proc. Natl Acad. Sci. USA*, **95**, 10602–10607.
Berglund, J. A., Chua, K., Abovich, N., Reed, R. & Rosbash, M. (1997). *Cell*, **89**, 781–787.
Berglund, J. A., Fleming, M. L. & Rosbash, M. (1998). *RNA*, **4**, 998–1006.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Corioni, M., Antih, N., Tanackovic, G., Zavolan, M. & Kramer, A. (2010). *Nucleic Acids Res.*, doi:10.1093/nar/gkq1042.
Evans, P. (2006). *Acta Cryst.* D**62**, 72–82.
Garrey, S. M., Voelker, R. & Berglund, J. A. (2006). *J. Biol. Chem.* **281**, 27443–27453.
Goldstrohm, A. C., Albrecht, T. R., Suñé, C., Bedford, M. T. & Garcia-Blanco, M. A. (2001). *Mol. Cell. Biol.* **21**, 7617–7628.
Greenfield, N. & Fasman, G. D. (1969). *Biochemistry*, **8**, 4108–4116.
Gupta, A., Jenkins, J. L. & Kielkopf, C. L. (2011). *J. Mol. Biol.* **405**, 1128–1138.
Guth, S. & Valcárcel, J. (2000). *J. Biol. Chem.* **275**, 38059–38066.
Irimia, M. & Roy, S. W. (2008). *PLoS Genet.* **4**, e1000148.
Kabsch, W., Kabsch, H. & Eisenberg, D. (1976). *J. Mol. Biol.* **100**, 283–291.
Kielkopf, C. L., Lücke, S. & Green, M. R. (2004). *Genes Dev.* **18**, 1513–1526.
Leslie, A. G. W. (2006). *Acta Cryst.* D**62**, 48–57.
Liu, Z., Luyten, I., Bottomley, M. J., Messias, A. C., Houngninou-Molango, S., Sprangers, R., Zanier, K., Krämer, A. & Sattler, M. (2001). *Science*, **294**, 1098–1102.
Lupas, A., Van Dyke, M. & Stock, J. (1991). *Science*, **252**, 1162–1164.
Manceau, V., Swenson, M., Le Caer, J. P., Sobel, A., Kielkopf, C. L. & Maucuer, A. (2006). *FEBS J.* **273**, 577–587.
Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
Michel, F. & Ferat, J. L. (1995). *Annu. Rev. Biochem.* **64**, 435–461.
Padgett, R. A., Konarska, M. M., Grabowski, P. J., Hardy, S. F. & Sharp, P. A. (1984). *Science*, **225**, 898–903.
Putnam, C. D., Hammel, M., Hura, G. L. & Tainer, J. A. (2007). *Q. Rev. Biophys.* **40**, 191–285.
Rain, J. C., Rafi, Z., Rhani, Z., Legrain, P. & Krämer, A. (1998). *RNA*, **4**, 551–565.
Razvi, A. & Scholtz, J. M. (2006). *Protein Sci.* **15**, 1569–1578.
Reed, R. (1989). *Genes Dev.* **3**, 2113–2123.
Rost, B. & Sander, C. (1994). *Proteins*, **19**, 55–72.
Roy, A., Kucukural, A. & Zhang, Y. (2010). *Nature Protoc.* **5**, 725–738.
Rutz, B. & Séraphin, B. (1999). *RNA*, **5**, 819–831.
Rutz, B. & Séraphin, B. (2000). *EMBO J.* **19**, 1873–1886.
Selenko, P., Gregorovic, G., Sprangers, R., Stier, G., Rhani, Z., Krämer, A. & Sattler, M. (2003). *Mol. Cell*, **11**, 965–976.
Shitashige, M., Satow, R., Honda, K., Ono, M., Hirohashi, S. & Yamada, T. (2007). *Cancer Sci.* **98**, 1862–1867.
Tanackovic, G. & Krämer, A. (2005). *Mol. Biol. Cell*, **16**, 1366–1377.
Thickman, K. R., Swenson, M. C., Kabogo, J. M., Gryczynski, Z. & Kielkopf, C. L. (2006). *J. Mol. Biol.* **356**, 664–683.
Wahl, M. C., Will, C. L. & Lührmann, R. (2009). *Cell*, **136**, 701–718.
Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P. & Burge, C. B. (2008). *Nature (London)*, **456**, 470–476.
Wang, G.-S. & Cooper, T. A. (2007). *Nature Rev. Genet.* **8**, 749–761.
Weiner, A. M. (1993). *Cell*, **72**, 161–164.
Zhang, D., Paley, A. J. & Childs, G. (1998). *J. Biol. Chem.* **273**, 18086–18091.