Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Structure determination of the minimal complex between Tfb5 and Tfb2, two subunits of the yeast transcription/DNA-repair factor TFIIH: a retrospective study

Denis E. Kainov,‡ Vincent Cura, Marc Vitorino, Helène Nierengarten, Pierre Poussin, Bruno Kieffer, Jean Cavarelli and Arnaud Poterszman*

Institut de Génétique et de Biologie Moléculaire et Cellulaire, CNRS/INSERM/UDS, BP 163, 67404 Illkirch CEDEX, France

‡ Present address: Institute for Molecular Medicine Finland (FIMM), University of Helsinki, FI-00014 Helsinki, Finland.

Correspondence e-mail: arnaud.poterszman@igbmc.fr

Received 19 January 2010
Accepted 16 March 2010

Tfb5 interacts with the Tfb2 subunit of the general transcription factor TFIIH to ensure efficient nucleotide-excision repair in eukaryotes. The crystal structure of the complex between Tfb5 and the C-terminal region of Tfb2 (Tfb2C) from *Saccharomyces cerevisiae* has recently been reported. Here, the structure-determination process is described as a case study. Although crystals were obtained readily, it was not possible to determine experimental phases from a first crystal form ($Tfb2_{412–513}$–$Tfb5_{2–72}$) that diffracted to 2.6 Å resolution. Shortening of the Tfb2C from its N-terminus was decisive and modified the crystal packing, leading to a second crystal form ($Tfb2_{435–513}$–$Tfb5_{2–72}$). These crystals diffracted to 1.7 Å resolution with excellent mosaicity and allowed structure determination by conventional approaches using heavy atoms. The refined structure from the second crystal form was used to solve the structure of the first crystal form by molecular replacement. Comparison of the two structures revealed that the N-terminal region of Tfb2C and (to a lesser extent) the C-terminal region of Tfb5 contributed to the crystal packing. A detailed analysis illustrates how variation in domain boundaries influences crystal packing and quality.

1. Introduction

Nucleotide-excision repair (NER) is a DNA-repair pathway which removes bulky adducts generated by UV sunlight in genomic DNA. A deficiency in NER causes the rare genetic disorders xeroderma pigmentosum (XP), trichothiodystrophy (TTD) and Cockayne syndrome. These diseases are associated with oncogenesis, developmental abnormalities and accelerated ageing.

The widely accepted model of NER includes four steps: identification of damage-induced DNA distortion, verification of the damage, removal of the damaged oligonucleotide and resynthesis, which fills the gap in the DNA. The general transcription factor TFIIH bridges the second and the third steps (Dip *et al.*, 2004; Laine & Egly, 2006; White, 2009).

TFIIH consists of ten subunits, of which seven (Ssl1–2, Rad3, Tfb1–2 and Tfb4–5) form the 'core' and three (Ccl1, Tfb3 and Kin28) associate into the cyclin-activating kinase complex. The molecular architectures of human and yeast TFIIH have been studied by electron microscopy (Chang & Kornberg, 2000; Schultz *et al.*, 2000). High-resolution analysis of individual TFIIH subunits has so far been limited to full-length Kin28, Ccl1, Ssl2, Rad3 and Tfb5, and domains of Tfb3, Ssl1 and Tfb1 from different organisms (Andersen *et al.*, 1996; Fan *et al.*, 2006, 2008; Fribourg *et al.*, 2000; Gervais *et al.*, 2001, 2004; Lolli *et al.*, 2004; Vitorino *et al.*, 2007; White, 2009; Wolski *et al.*, 2008).

Despite the progress that has been made in understanding the functions of individual TFIIH subunits, little is known about the protein–protein interaction network within the complex. Recently, we have reported the crystal structures of minimal complexes between Tfb5 and the C-terminal domain of Tfb2 (Tfb2C). The molecular details of the Tfb5–Tfb2C interaction have shed some light on the understanding of a rare neurodevelopmental repair syndrome called group A trichothiodystrophy (Kainov *et al.*, 2008, 2010). Here, we focus on the structure-determination steps and in particular on the initial experimental phasing. Difficulties were mainly associated with crystal fragility, the relatively poor diffraction quality of the first crystal form, the small number of methionine residues and their location in disordered regions. The removal of disordered residues from the N-terminus of Tfb2C led to a second crystal form which diffracted to higher resolution. This second crystal form was used for structure determination using SIRAS.

2. Material and methods

2.1. Cloning

The *tfb5* gene was PCR-amplified from *Saccharomyces cerevisiae* genomic DNA with GCACCATGGCTAGAGCAAGAAAG and GGTCCTCGAGTTACTGATTTTCTTCTT oligonucleotides. The PCR product was digested with *Nco*I and *Xho*I restriction enzymes and ligated into pACYCDuet or pET21d vectors (Novagen) at the *Nco*I–*Xho*I sites, yielding pTFB5 and pTFB5His expression plasmids.

The *tfb2* gene was PCR-amplified from *S. cerevisiae* genomic DNA with GCACCATGGACTATTCCCTGA and GCCGTCGACTTATTGTTTCTTTTTCA primers. The PCR product was digested with *Nco*I and *Sal*I restriction enzymes and ligated into pACYC vector at the *Nco*I–*Xho*I sites. The resulting pTFB2 plasmid encodes the full-length Tfb2 protein. We recloned the fragment of the *tfb2* gene encoding the C-terminal region of Tfb2 (Tfb2C; residues 412–513) with CACAGCTAGCGCAGAAGAGAAA and CTTCCTCGAGTTATTGTTTCTTTTT oligonucleotides into pSKB2 vector at the *Nhe*I–*Xho*I sites (Deaconescu *et al.*, 2006). The resulting pTFB2-102C plasmid encoded a cleavable N-terminal His tag and 102 residues of Tfb2. We also produced a pTFB2-79C plasmid encoding a shorter fragment (residues 435–513) using 5′-phosphorylated primers CCTACCGTCGTAGATCAAATCA and CCCCTGGAACAGAACTTC, Phusion polymerase (Finnzyme) and pTFB2-102C as a template.

2.2. Expression and purification

For crystallization of $Tfb2_{412–513}$–$Tfb5_{2–71}$ and $Tfb2_{435–513}$–$Tfb5_{2–71}$ complexes, pTFB5 plasmid was co-transformed with pTFB2-102C or pTFB2-79C into *Escherichia coli* B834 (DE3) cells. Bacterial cells were grown at 310 K until the absorbance at 600 nm reached 0.5 either in 2 l LB medium for the production of unlabelled protein or in MOPS minimal medium

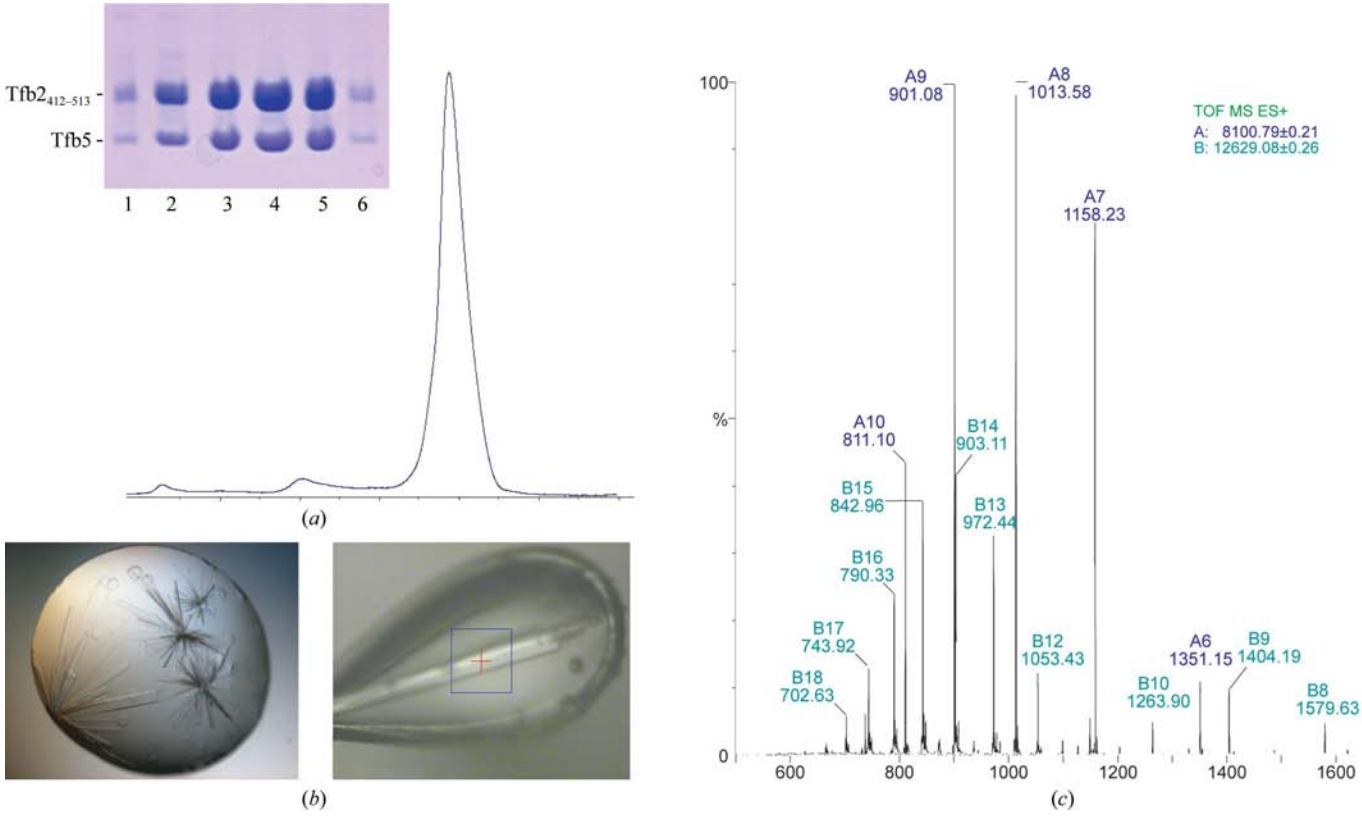

Figure 1
Crystallization of the $Tfb2_{412–513}$–$Tfb5_{2–72}$ complex. (*a*) SDS–PAGE analysis of peak fractions from gel-filtration chromatography. (*b*) Crystals of the complex in the crystallization drop (left) and mounted in a cryoloop (right). (*c*) ESI–TOF MS spectrum of the crystallized complex.

containing 50 mg ml^{-1} selenomethionine (Nanduri *et al.*, 2002) for the production of selenomethionylated complexes. Flasks were transferred to 290 K and IPTG was added to a final concentration of 0.5 m*M*. Cells were grown for 10 h, collected by centrifugation and resuspended in 20 ml TN buffer (20 m*M* Tris–HCl, 50 m*M* NaCl pH 8.9) containing 5 m*M* imidazole. The following steps were performed at 277 K. The cell suspension was sonicated and centrifuged at 120 000*g* for 1 h. The supernatant was loaded onto a HisTrap column (GE Healthcare). Proteins were eluted with a linear gradient of imidazole (10–500 m*M*). The histidine tag was cleaved overnight using HRV 3C protease (Novagen). Complex-containing fractions were diluted with TN buffer and loaded onto connected HisTrap and Q columns (GE Healthcare). The columns were disconnected and proteins were eluted from the anion-exchange column with a linear gradient of NaCl (0.1–1.0 *M*). The fractions containing the complex were pooled and resolved by gel filtration on a Superdex-75 column (GE Healthcare) equilibrated with TN buffer. Both the Tfb2$_{412–513}$–Tfb5$_{2–71}$ and Tfb2$_{435–513}$–Tfb5$_{2–71}$ protein complexes were concentrated to 35 mg ml^{-1} using Amicon Ultra concentration devices (Millipore).

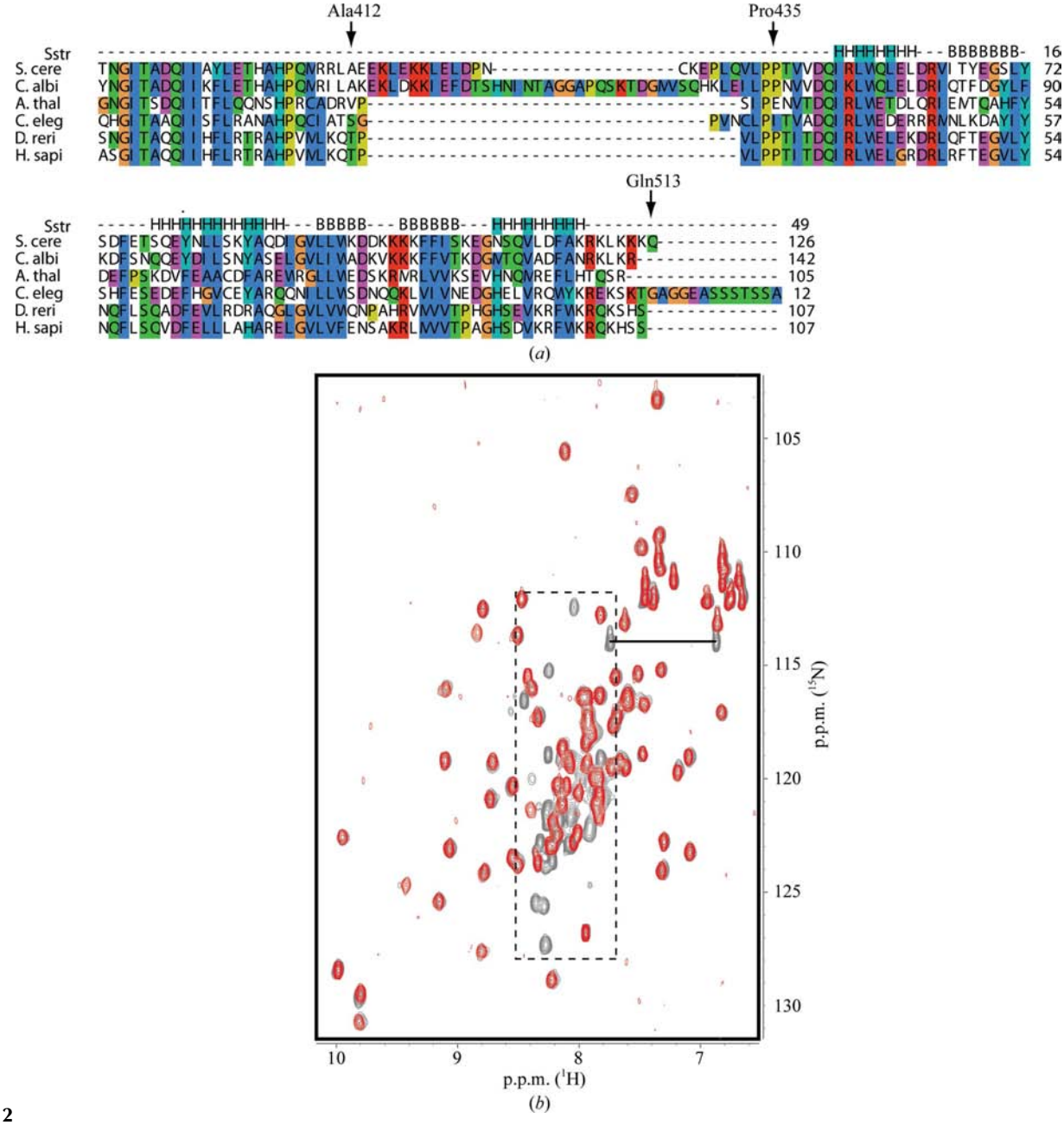

Figure 2
(*a*) Sequence alignment of Tfb2 C-terminal regions from *Saccharomyces cerevisiae* (S. cere), *Candida albicans* (C. albi), *Arabidopsis thaliana* (A. thal), *Caenorhabditis elegans* (C. eleg), *Danio rerio* (D. reri) and *Homo sapiens* (H. sapi). Secondary-structure elements are shown at the top (H, α-helix; B, β-sheet). Arrows indicate Ala412, Pro435 and Gln513. (*b*) Superposition of ^{1}H–^{15}N HSQC spectra from Tfb2$_{412–513}$ (black) and Tfb2$_{435–513}$ (red). The dashed box highlights the spectral region in which most of the resonances are lost on the N-terminal deletion. The NH_2 spin system of Asn425 is indicated by a horizontal black line.

For NMR analysis, $Tfb2_{412–513}$ and $Tfb2_{435–513}$ were expressed from pTFB2-102C and pTFB2-79C in *E. coli* BL21 (DE3) in ^{15}N-labelled M9 minimal medium supplemented with 10%(*v*/*v*) ^{15}N Silantes OD2 medium and purified as described in Kainov *et al.* (2010). Unlabelled His-tagged Tfb5 was purified as described in Kainov *et al.* (2008).

2.3. Crystallization and data collection

Crystallization conditions for the $Tfb2_{412–513}$–$Tfb5_{2–71}$ complex were screened with commercially available kits using the sitting-drop vapour-diffusion method in 96-well CrystalQuick plates (Greiner Bio-One) employing a Tecan robot (0.1 µl protein solution at 10 mg ml^{-1} was mixed with 0.1 µl precipitant solution and equilibrated against 75 µl reservoir solution). Initial hits were obtained using the PEG/Ion Classic screen (Hampton Research). After optimization, thin plate-shaped crystals (120 × 40 × 20 µm) appeared in 24 h in sitting drops (Fig. 1*b*). 1 µl protein solution (10 mg ml^{-1}) was mixed with 1 µl reservoir solution (17.5% PEG 3350, 0.1 *M* HEPES pH 6.8–7.2) and the drops were equilibrated against 600 µl reservoir solution at 294 K. Crystals were flash-frozen in liquid nitrogen after a brief soaking in 5 µl reservoir solution containing 25%(*v*/*v*) glycerol. X-ray diffraction data were collected using cryocooled (100 K) crystals. Single-wavelength data sets were collected on beamlines ID19 and ID23-B at the APS Structural Biology Centre, Chicago, USA using a MAR CCD detector with an attenuated beam and a crystal-to-detector distance of 385 mm; 200 frames were collected with 0.5° oscillation and 0.3 s exposure time (Kainov *et al.*, 2008). A MAD data set was collected on beamline ID23-1 at the ESRF using a Q315 CCD detector from ADSC with a crystal-to-detector distance of 450 mm; 200 frames were collected with 1.0° oscillation and 1.0 s exposure time.

Crystals of $Tfb2_{435–513}$–$Tfb5_{2–71}$ (40 × 40 × 200 µm) grew in a few days in an Eppendorf tube in TN buffer at 277 K. Crystals were flash-frozen in liquid nitrogen after a brief transfer to 5 µl TN buffer containing 25%(*v*/*v*) ethylene glycol as a cryoprotectant and were stored in liquid nitrogen. X-ray diffraction data were collected from cryocooled (100 K) crystals in-house using a Cu rotating-anode generator with Osmic mirrors and a CCD detector at the ESRF. A high-resolution data set was collected between 50 and 1.8 Å resolution as 360 frames of 0.5° oscillation (exposure time of 0.3 s per oscillation with an attenuated beam and a crystal-to-detector distance of 272 mm) on ID29 (Kainov *et al.*, 2008). A MAD data set was collected between 50 and 2.3 Å resolution as 200 frames of 1.0° oscillation (exposure time of 1.0 s per oscillation with an attenuated beam and a crystal-to-detector distance of 227 mm) on ID23.

2.4. Mass spectrometry

The crystal content was analyzed as described by Potier *et al.* (2000). Briefly, crystals were washed in 10 µl TN buffer and dissolved in a 1:1(*v*:*v*) water–acetonitrile mixture containing 1% formic acid to a protein concentration of 5 pmol µl^{-1}. Mass spectrometry was performed in positive-ion mode on an electrospray ionization time-of-flight mass spectrometer with a standard Z-spray source (LCT, Waters, Massachusetts, USA). Samples were continuously infused into the ion source at a flow rate of 5 µl min^{-1} *via* a Harvard Model 11 syringe pump (Havard Apparatus). Denatured horse heart myoglobin (Sigma–Aldrich, St Louis, Missouri, USA) was used as a calibration standard.

2.5. Analytical gel filtration and NMR analysis

The molecular masses of the complexes were estimated using analytical gel filtration on Superdex-75 (GE Healthcare). The column was calibrated with gel-filtration molecular-weight standards (Bio-Rad).

For NMR analysis, purified proteins were dialyzed against 50 m*M* phosphate buffer pH 6.5 containing 50 m*M* NaCl. ^{15}N-labelled $Tfb2_{315–513}$ was mixed with purified His-tagged Tfb5 and a heteronuclear ^{1}H–^{15}N NMR correlation spectrum (HSQC) was acquired on a Bruker DRX600 instrument equipped with a cryoprobe.

2.6. Crystallographic computing

Data were processed with *HKL*-2000 (Otwinowski & Minor, 1997) and refined with *REFMAC*5 (Vagin *et al.*, 2004) and models were analyzed with *Coot* (Emsley & Cowtan, 2004). *SOLVE* (Terwilliger & Berendzen, 1999) and *SHARP* (Vonrhein *et al.*, 2007) were used for MAD phasing attempts, *Phaser* (McCoy *et al.*, 2007) was used for molecular replacement and *ARP*/*wARP* was used for automated model building (Cohen *et al.*, 2008). Rotations were compared using *COMPANG* (Urzhumtseva & Urzhumtsev, 2002). Other crystallographic calculations were carried out with the *CCP*4 program suite (Collaborative Computational Project, Number 4, 1994). Analysis of the interfaces was performed with *PISA* (Krissinel & Henrick, 2007) and *SPIDER* (Porollo & Meller, 2007). Figures were generated using *PyMOL* (DeLano, 2008). The anomalous scattering ratio, $|\Delta F|/F$, was approximated using the formula $(2N_A/N_P)^{1/2}(f''/Z_{eff})$, where N_A is the number of anomalous scatters, N_P is the total number of non-H atoms, $f'' = \sim 6.0$ is the imaginary part of the anomalous scattering for Se and $Z_{eff} = 6.7$ is the effective number of electrons per non-H atom (Hendrickson & Teeter, 1981). In the case of the hexagonal crystal form obtained with the $Tfb2_{435–513}$–$Tfb5_{2–72}$ complex, $N_A = 2$ and $N_P = 1237$, leading to $|\Delta F|/F \simeq 0.05$.

3. Results and discussion

3.1. Crystallization and initial attempts to solve the structure of the Tfb2–Tfb5 complex

The $Tfb2_{412–513}$–$Tfb5_{2–72}$ complex was purified to homogeneity using a combination of affinity, anion-exchange and size-exclusion chromatography. The complex eluted as a single peak from the gel-filtration column with an apparent molecular mass of 20 kDa (the predicted mass of the heterodimer is 21 kDa) and exhibited no signs of aggregation (Fig. 1*a*). The complex was concentrated and crystallized. Thin plate-shaped

Table 1
Data-collection statistics for the $Tfb2_{412–513}$–$Tfb5_{2–72}$ and $Tfb2_{435–513}$–$Tfb5_{2–71}$ complexes.

Values in parentheses are for the outer shell.

	$Tfb2_{412–513}$–$Tfb5_{2–72}$					$Tfb2_{435–513}$–$Tfb_{52–72}$	
			SeMet (Crys21)				SeMet (Crys18)
	Native 1	Native 2†	Peak	Inflection	Remote	Native† (nat02)	Peak
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$			$P6_1$	$P6_1$
Unit-cell parameters (Å)	a = 36.9, b = 101.7, c = 114.8	a = 37.6, b = 103.6, c = 114.3	a = 37.3, b = 103.0, c = 113.9			$a = b$ = 100.0, c = 37.3	$a = b$ = 100.5, c = 37.2
Wavelength (Å)	0.9198	0.9792	0.9803	0.9808	0.9762	0.91980	0.98055
Resolution (Å)	2.7	2.6	2.9	2.9	2.9	1.8	2.4
Unique reflections	11632	13927	10114	10102	10155	19737	15085
Redundancy	2.2 (1.3)	4.0	4.2 (3.6)	4.2 (3.8)	3.6 (4.1)	10.3 (9.2)	4.4 (4.3)
R_{merge}‡ (%)	7.2 (33.2)	5.8 (13.1)	5.7 (10.6)	6.4 (15.1)	7.7 (22.8)	7.2 (24.9)	4.7 (12.0)
$I/\sigma(I)$	25.3 (6.5)	15.9 (8.9)	36.3 (17.5)	31.7 (11.7)	27.3 (9.6)	26.0 (7.3)	49.5 (22.4)
Completeness (%)	92.8 (74.0)	91.9 (96.4)	95.3 (53.9)	97.8 (78.6)	99.2 (93.3)	98.5 (93.7)	96.2 (95.1)

† PDB codes 3dom and 3dgp (Kainov *et al.*. 2008). ‡ $R_{merge} = \sum_{hkl}\sum_i |I_i(hkl) - \langle I(hkl)\rangle|/\sum_{hkl}\sum_i I_i(hkl)$, where $\langle I(hkl)\rangle$ is the mean of i observations of reflection hkl.

crystals (120 × 40 × 20 µm) were obtained using PEG 3350 as a precipitant after two weeks of incubation at 290 K (Fig. 1*b*).

The ESI mass spectrum of the proteins from the crystals showed two series of multiply charged ions corresponding to $Tfb5_{2–72}$ and $Tfb2_{412–513}$. The measured mass of 8100.8 ± 0.2 Da is identical to the expected mass for $Tfb5_{2–72}$ of 8101.3 Da. The measured mass of 12 629.1 ± 0.3 Da corresponds to the expected mass of 12 629.6 Da of $Tfb2_{412–513}$ with the extra N-terminal residues GPHMAS that remained after TEV cleavage of the tag (Fig. 1*c*; Supplementary Table 1[1]).

Although crystals of $Tfb2_{412–513}$–$Tfb5_{2–72}$ could readily be obtained, collection of diffraction data proved to be difficult. The crystals were extremely fragile and difficult to handle. Despite extensive efforts to control cryoprotection, only a minor proportion of the crystals exhibited reasonable diffraction and low mosaicity. Of 250 crystals tested on the ESRF or APS beamlines, only three yielded complete data sets (Table 1). The best crystal diffracted to 2.6 Å resolution. The crystals belonged to the orthorhombic space group $P2_12_12_1$, with unit-cell parameters a = 37.9, b = 103.6, c = 114.3 Å (Native 2 in Table 1). Cell-content analysis suggested the presence of two $Tfb2_{412–513}$–$Tfb5_{2–72}$ complexes per asymmetric unit, with a corresponding Matthews coefficient of 2.68 $Å^3$ Da^{-1} and an estimated solvent content of 54% (Supplementary Table 2[1]).

The purified $Tfb2_{412–513}$–$Tfb5_{2–72}$ complex (179 amino acids in total) contained two methionines. Therefore, phasing *via* selenomethionine-derivatized crystals was attempted. SeMet-substituted crystals were grown, their content was analyzed by mass spectrometry (Supplementary Table 1[1]) and data sets were collected at the peak (λ = 0.9803 Å), infection (λ = 0.9808 Å) and remote (λ = 0.9762 Å) wavelengths (Crys21 in Table 1). These data sets extended to 2.9 Å resolution with reasonable statistics but with redundancies of 4.2, 4.2 and 3.9, respectively, that were limited by crystal decay. Unfortunately, neither the Bijvoet nor the dispersive ratios were statistically significant and attempts to locate selenium sites and to phase using *SOLVE* or *SHARP* failed. In the absence of suitable anomalous signal from SeMet-labelled crystals, we tried conventional heavy-atom methods, but were unable to collect a usable data set either after high-pressure treatment with xenon or after soaking with several heavy-atom derivatives.

3.2. Refined domain boundaries led to another crystal form

We then decided to reconsider our constructs. ^{15}N-labelled $Tfb2_{412–513}$ and unlabelled $Tfb5_{2–72}$ were expressed independently in *E. coli* and purified. The ^{15}N-labelled $Tfb2_{412–513}$ protein was mixed with a slight excess of Tfb5 and the heteronuclear ^1H–^{15}N NMR correlation spectrum (HSQC) was recorded (Fig. 2*b*). The spectral dispersion of the correlations indicated the presence of a folded domain. However, the number of observed correlations (85) was lower than expected (95) owing to the specific broadening of a subset of resonances. This observation, which resulted in conformational averaging, prompted us to reconsider the boundaries of the domain in light of the multiple sequence alignment (Fig. 2*a*). Analysis of aligned sequences of Tfb2 orthologues showed that the N-terminal part of $Tfb2_{412–513}$ contains several proline residues together with nonconserved charged residues, features that are typical of disordered regions. Residues 412–434 were deleted from Tfb2C, resulting in a construct starting with the sequence G**PTVVDQIR** (after tag removal). The deletion did not affect the ability of the Tfb2 C-terminal domain to associate with Tfb5. A heteronuclear ^1H–^{15}N NMR correlation spectrum of the $Tfb2_{435–513}$–$Tfb5_{2–72}$ complex was measured. Superpositioning of the spectra of $Tfb2_{412–513}$–$Tfb5_{2–72}$ and $Tfb2_{435–513}$–$Tfb5_{2–72}$ (Fig. 2*b*) showed that most of the correlation peaks were common to both complexes, except for a subset of 16 correlations with proton frequencies clustered around 8.2 p.p.m. These correlations that were specifically observed in the $Tfb2_{412–513}$–Tfb5 complex are indicative of a disordered state of the N-terminal extension of Tfb2 encompassing residues 412–434. It is noteworthy that the removal of this region had only marginal effects on the

[1] Supplementary material has been deposited in the IUCr electronic archive (Reference: DZ5196). Services for accessing this material are described at the back of the journal.

frequencies of the remaining amide protons, indicating a lack of interaction between the N-terminal tail and the folded part of the domain. Moreover, the number of correlation peaks obtained for the short Tfb2C construct was very close to the expected number (74 *versus* 76), indicating efficient suppression of conformational averaging. Thus, the short version of Tfb2C was used for co-expression with Tfb5 and complex crystallization.

Deletion of the N-terminal tail had no visible effect on the chromatographic behaviour of the complex (not shown). Crystallization trials with commercial kits and protein solution at 10 mg ml^{-1} resulted in crystalline precipitates and needle-shaped crystals in a number of conditions. The largest crystals, however, grew in batch from the concentrated protein stock (35 mg ml^{-1} protein in a buffer containing 20 m*M* Tris–HCl pH 8.9 and 50 m*M* NaCl) that was stored at 277 K. ESI–MS analysis of Tfb5–Tfb2$_{435–513}$ revealed two populations of multiply charged ions (data not shown). The Tfb5$_{2–72}$ measured mass of 8101.4 ± 0.2 Da coincided with the theoretical mass of 8101.3 Da. The Tfb2$_{435–513}$ measured mass of 9459.9 ± 0.9 Da corresponded to the expected mass of 9459.9 Da with an extra N-terminal glycine residue remaining after tag cleavage.

Cryoprotected crystals diffracted to 2.9 Å resolution on our home source with a mosaicity close to 0.3°. A complete 1.8 Å resolution data set was collected on beamline ID29 at ESRF. The crystals belonged to the hexagonal space group $P6_1$, with unit-cell parameters $a = b = 100.0$, $c = 37.3$ Å (Table 1). The asymmetric unit contained one complex, with a corresponding Matthews coefficient of 3.24 Å^3 Da^{-1} and a solvent content of 62%.

3.3. Structure determination and packing analysis

To obtain phases for the Tfb2$_{435–513}$–Tfb5$_{2–52}$ crystal form, MAD experiments with SeMet-substituted crystals and conventional heavy-atom searches were attempted in parallel. Heavy atoms were screened on our home source and gold cyanide was rapidly identified as a potential derivative. Soaking with 5 m*M* KAu(CN)$_2$ for 10 min did not lead to a visible loss of diffraction, but resulted in isomorphous differences of 19.5% with the reference set collected under identical conditions. A single heavy-atom site was identified using Patterson methods. SIRAS phases (up to 3.2 Å) with an overall figure of merit of 0.64 were computed using *SOLVE* and improved by density modification with *RESOLVE* for space groups $P6_1$ and $P6_5$. Space group $P6_1$ led to an interpretable map with an average figure of merit of 0.85. This map allowed chain tracing after a few cycles of rebuilding/refinement first against the 2.9 Å data set and then against the 1.8 Å synchrotron data set (Table 1). Fig. 3(*a*) shows a portion of the electron-density map of the interface

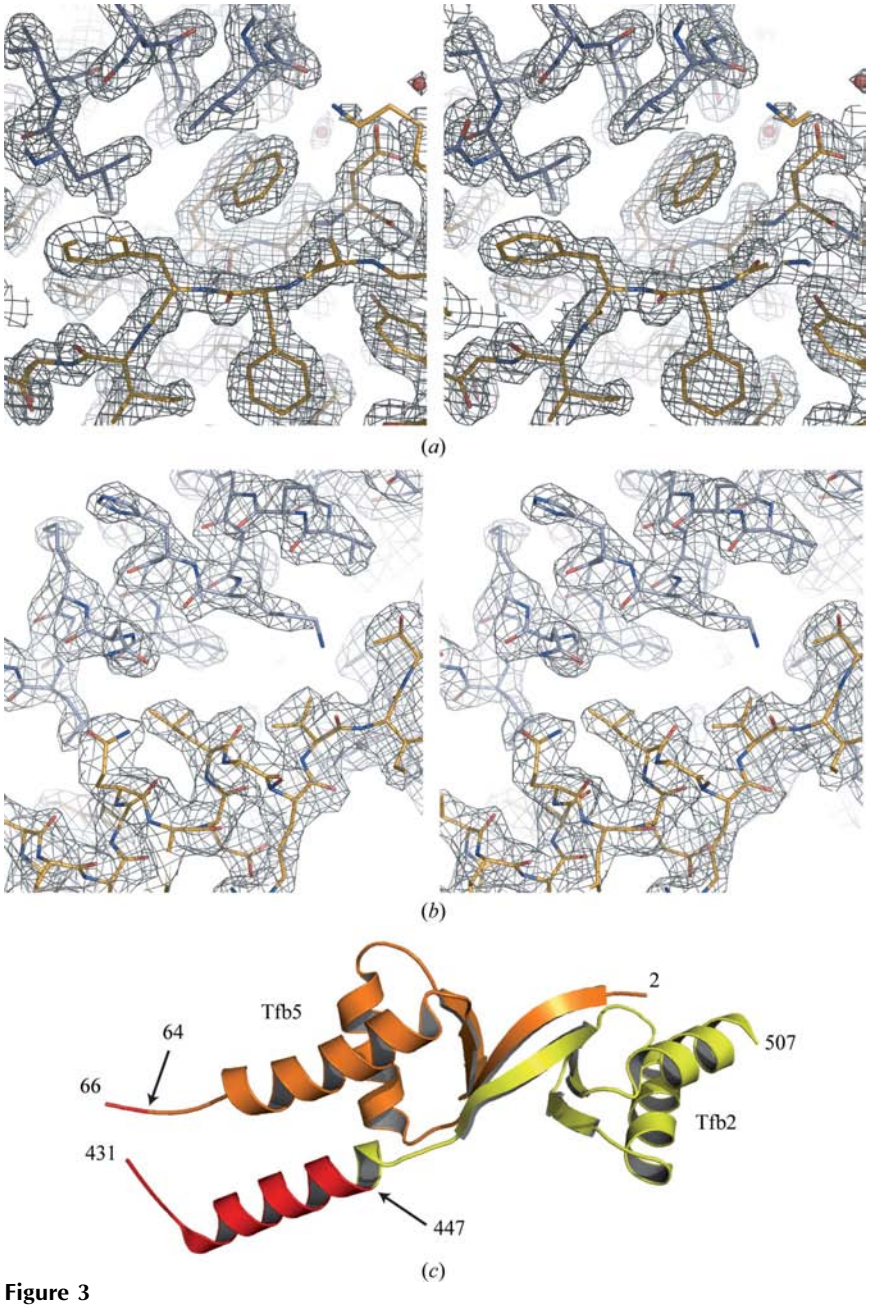

Figure 3
Stereoviews of the Tfb2$_{435–513}$–Tfb5$_{2–72}$ and Tfb2$_{412–513}$–Tfb5$_{2–72}$ complexes. (*a*) Electron-density map from the hexagonal crystal form ($P6_1$; Tfb2$_{435–513}$–Tfb5$_{2–72}$) contoured at 1σ, showing a portion of the interface between Tfb2 and Tfb5. (*b*) Electron-density map from the Tfb2$_{412–513}$–Tfb5$_{2–72}$ orthorhombic crystal form ($P2_12_12_1$; Tfb2$_{412–513}$–Tfb5$_{2–72}$), showing the N-terminal helix of Tfb2C (αN), which is not visible in the hexagonal form, and the Tfb5 (α2) helix. (*c*) Overall view of the Tfb2C–Tfb5 complex (chains *C* and *D* from PDB entry 3dom; Kainov *et al.*, 2008). Tfb2C is in yellow and Tfb5 is in orange. Residues 431–446 of Tfb2C and 65–66 of Tfb5, which could not be traced from the hexagonal form, are shown in red.

between Tfb2 and Tfb5 (Fig. 3*a*). The final model includes residues 447–508 of Tfb2 and residues 2–64 of Tfb5 (Kainov *et al.*, 2008). Residues 435–446 of Tfb2 and 65–72 of Tfb5 are not visible in the electron-density map.

The refined structure of the $Tfb2_{435–513}$–$Tfb5_{2–72}$ heterodimer was used as a search model to solve the $Tfb2_{412–513}$–$Tfb5_{2–72}$ structure by molecular replacement. A unique and contrasted solution with two heterodimers in the asymmetric unit was obtained by *Phaser*. The log-likelihood gain was 1572. *Z* scores for the rotation and translation functions were 12.2 and 19.9 for the first copy and 11.6 and 39.8 for the second copy, respectively. In the $2F_o - F_c$ electron-density map, several residues which could not be traced in the high-resolution $P6_1$ crystal form were clearly visible. They correspond to the N-terminus of Tfb2C and to the C-terminus of Tfb5, which form antiparallel helices (Fig. 3*b*). After refinement, the *R* and R_{free} factors were 20.2% and 26.1%, respectively. The first heterodimer encompasses residues 437–509 of Tfb2 and residues 2–59 of Tfb5 (chains *A* and *B*) and the second encompasses residues 431–507 of Tfb2 and residues 2–66 of Tfb5 (chains *C* and *D*; see Fig. 3*c*). The two copies can be superimposed with an average r.m.s.d. of 0.57 Å based on 123 C^{α} atoms and are related by a 170° rotation around an axis almost parallel to the *x* axis (ω = 88.97°, φ = 0.062°, κ = 169.27°). Owing to crystal packing, minor differences are observed for the N-terminal extremity of Tfb2C and the C-terminus of Tfb5.

When the two molecules from the orthorhombic crystal form are superimposed with the complex in the hexagonal

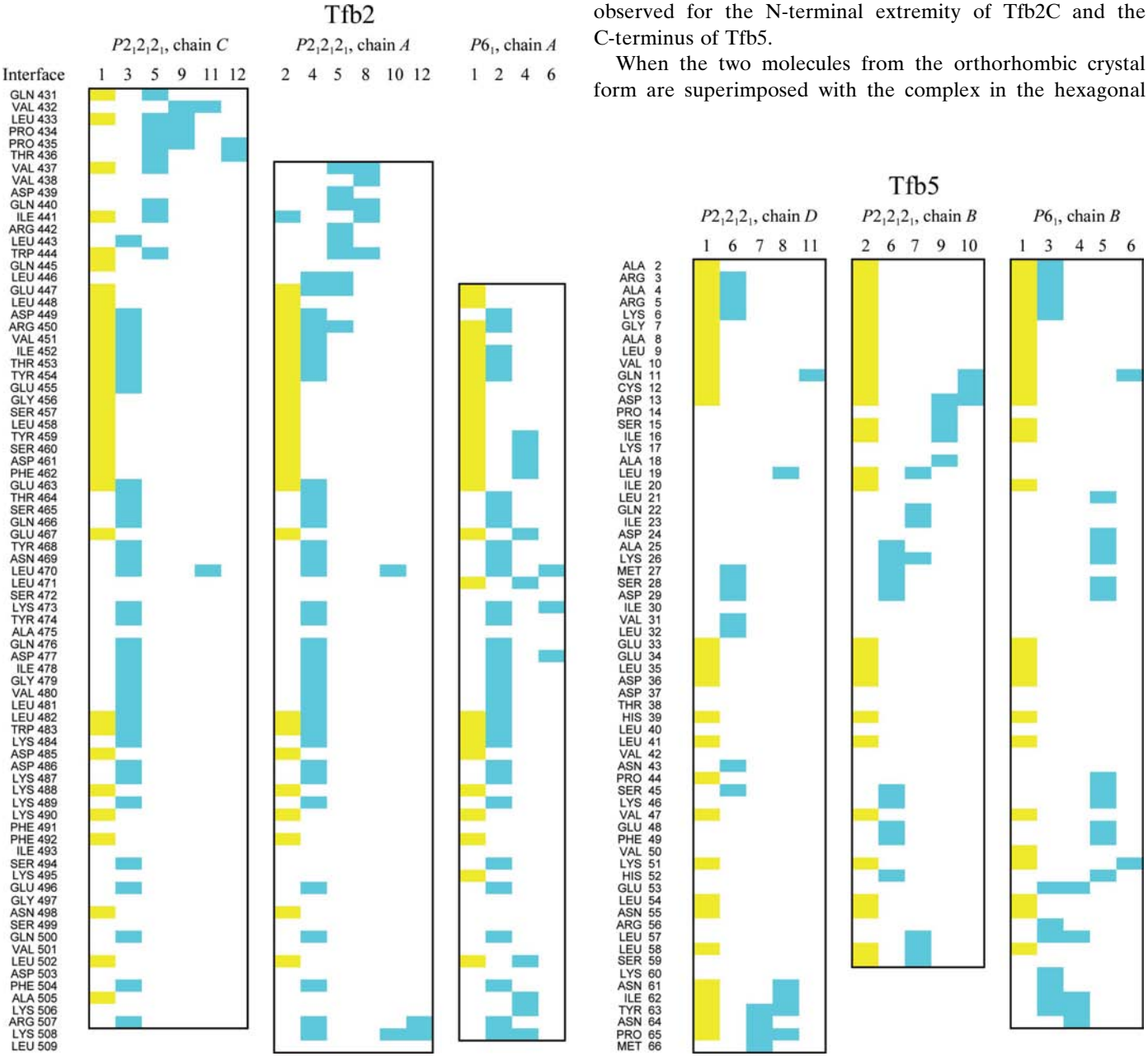

Figure 4
Interfaces in the orthorhombic ($P2_12_12_1$; $Tfb2_{412–513}$–$Tfb5_{2–72}$) and hexagonal ($P6_1$; $Tfb2_{435–513}$–$Tfb5_{2–72}$) crystal forms. Residues implicated in biologically relevant interfaces are shown in yellow. Crystal-packing contacts are shown in cyan. Data were obtained using the *PISA* software with PDB entries 3dgp and 3dom as input.

form, the average r.m.s.d.s are 0.57 and 0.87 Å based on 120 and 122 C^{α} atoms, indicating that the molecular structures are almost identical. Both forms have similar crystal volumes per molecule and solvent contents (V_M of 2.68 and 3.05 Å^3 Da^{-1} and 542 and 59% solvent content for the orthorhombic and hexagonal forms, respectively). This is reflected in the equal number of contacts: five contacts per heterodimer in the hexagonal form and ten for the two copies of the complex in the orthorhombic form (Fig. 4 and Supplementary Table 3). Identical Tfb2C–Tfb2C packing contacts are observed in both crystal forms (interface 2 in the hexagonal form; interfaces 3 and 4 in the orthorhombic form). This interface involves ∼650 Å^2, which accounts for 50% of the total packing interface. For comparison, the Tfb2C–Tfb5 interface occupies 1100 Å^2 (Supplementary Table 3).

A specific feature of the orthorhombic form involves the N-terminal residues of Tfb2C and (to a lesser extent) the C-terminal residues of Tfb5 in crystal packing (Fig. 4). These residues, which are disordered in the hexagonal crystal form, bridge the two copies of the Tfb2C–Tfb5 complex in the orthorhombic crystal form, forming a heterotetrameric assembly (Fig. 5*a*). The residues at the interface are mainly engaged in hydrophobic interactions that stabilize the extremities of Tfb2C and Tfb5 (interfaces 5, 7, 8 and 9; Supplementary Table 4; Kabsch & Sander, 1983). These interactions are responsible for the continuity of the packing along the *c* axis in the orthorhombic crystal form (Fig. 5*b*). They stabilize the N-terminus of Tfb2C. Residues Val432, Pro434 and Pro435 from Tfb2 are packed against Asp13, Pro14, Ser15 and Ala18 from the Tfb5 moiety of another heterodimer (Fig. 5*c*). This

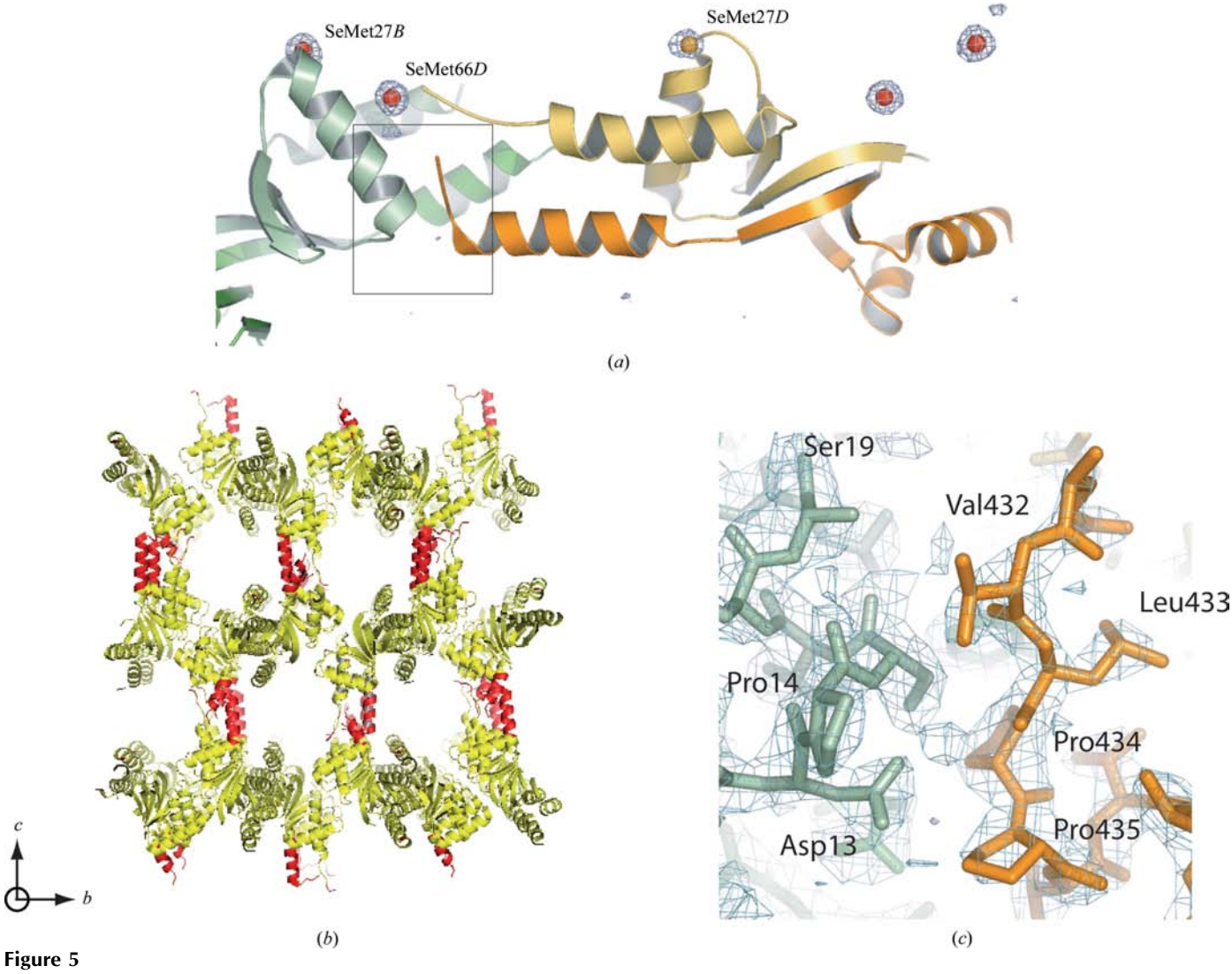

Figure 5
Packing analysis of the orthorhombic crystal form. (*a*) Interface between two copies of the Tfb2$_{412–513}$–Tfb5$_{2–72}$ complex. Chains *A* (*x*, *y*, *z*) and *B* (*x*, *y*, *z*) from 3dom for the first complex are shown in dark and pale green, respectively. Chains *C* (−*x* − 1/2, −*y*, *z* + 1/2) and *D* (−*x* − 1/2, −*y*, *z* + 1/2) of the second complex are shown in dark and pale orange, respectively. An anomalous difference Fourier (contoured at +4σ) calculated with the measured structure factors collected at the peak for selenium and phases from the refined model (3dom) is superimposed. The three highest peaks in the map correspond to the Se atoms of SeMet27 of chain *B* (18σ), SeMet66 of chain *D* (15σ) and SeMet27 of chain *D* (7σ). (*b*) Crystal packing in the *bc* plane in the orthorhombic crystal form showing the chains of molecules connected by the N- and C-terminal helices of Tfb2 and Tfb5, respectively. Residues 431–446 of Tfb2 and 65–66 of Tfb5, which could not be traced in the $P6_1$ crystal form, are shown in red. (*c*) Electron-density map ($3F_o - 2F_c$) contoured at 1σ showing the N-terminus of Tfb2 (chain *C*) and part of the symmetry-related Tfb5 (chain *B*) molecule (shown in a box in *a*).

interface engages only ∼120 Å^2, which accounts for 30% of the total packing interface but is critical for crystal packing. Indeed, the short Tfb2C construct lacking Val432 and Pro434 crystallized in a different space group ($P2_12_12_1$ instead of $P6_1$).

3.4. Phasing with SeMet-substituted crystals

Having determined the structure of the Tfb2$_{412–513}$–Tfb5$_{2–72}$ complex in the orthorhombic space group, we re-analyzed the original data sets in an attempt to understand why MAD phasing failed. Analysis of the anomalous difference Fourier map calculated with data collected at the peak wavelength and phases from the refined model located three of the four expected anomalous sites (Fig. 5*a*). Two of them are clearly above the background and correspond to the SeMet27 residue of chain *B* (18σ) and to SeMet66 of chain *D* (15σ) which is stabilized by packing contacts. The third corresponds to SeMet27 of chain *D* (7σ) and is slightly above the level of background peaks (5σ). The anomalous signal is nevertheless insufficient for location of the selenium substructure and phasing. The anomalous difference Patterson computed with peak data is not interpretable: two maxima that might correspond are found in the x = 0.5 section but have no counterparts in the y = 0.5 and z = 0.5 sections (Supplementary Fig. 1). With the known Se position, we attempted to phase *a posteriori* both with SAD (using the peak wavelength) and with MAD (using two or three wavelengths) but this was also unsuccessful, suggesting that data sets of better quality would be required to improve the signal-to-noise ratio.

It is generally accepted that one Se per 75–100 amino acids is sufficient for MAD or SAD phasing. Analysis of the selenomethionine density for proteins deposited in the PDB (Fig. 6*a*) shows that 18% have less than one ordered SeMet per 75 amino acids and that only 3% have less than one per 150 residues. In the case of the Tfb2C–Tfb5 complex, with two well ordered Se atoms and a third that is detected just above background in an anomalous difference Patterson, we were almost in this situation. The presence of only a single well ordered Se atom per 150 amino acids

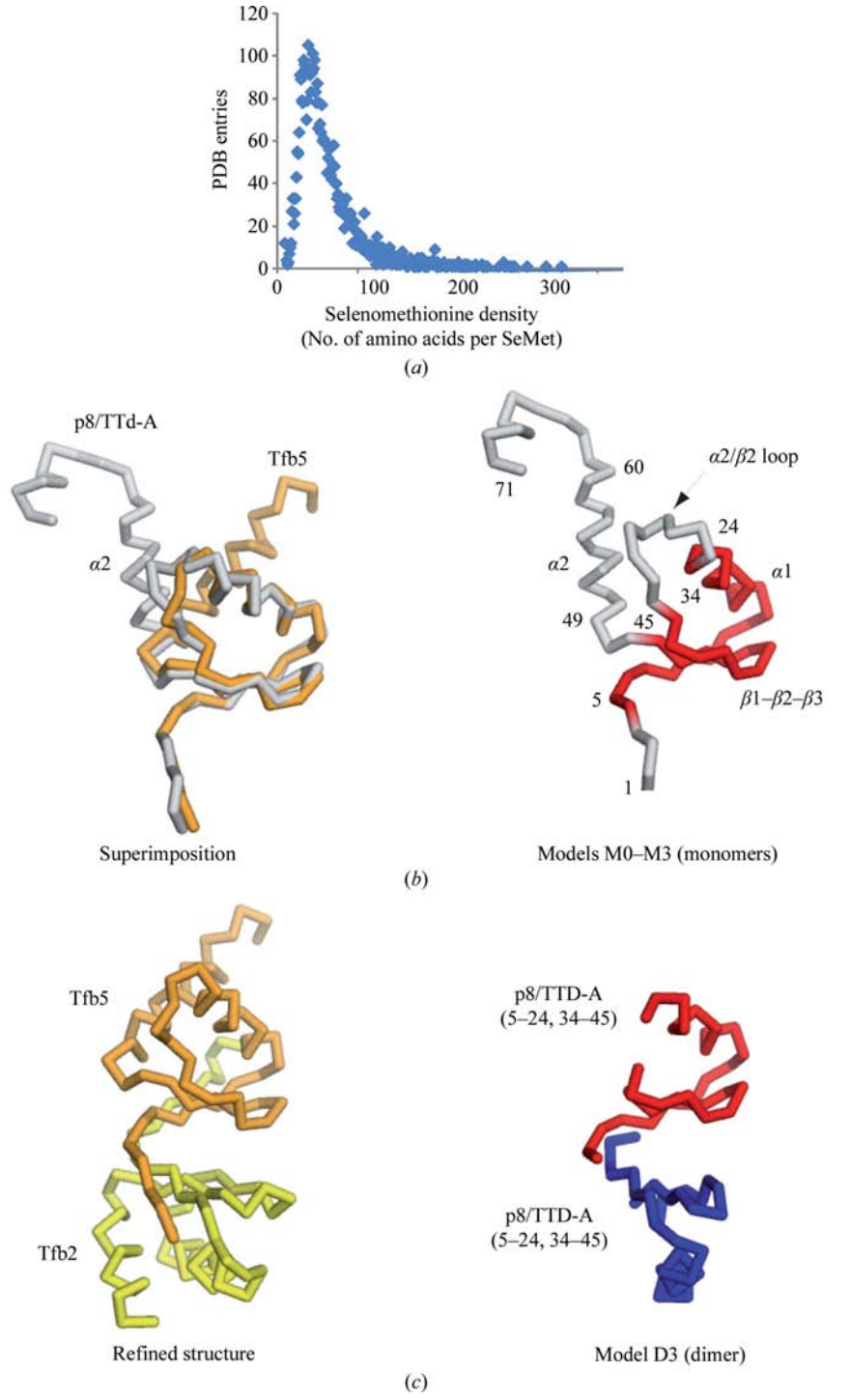

Figure 6
Retrospective analysis. (*a*) Statistical survey of SeMet in the PDB. A search for proteins containing selenomethionines was performed in the PDB and for each entry the SeMet content, expressed as a density, *i.e.* as the number of amino acids per SeMet, was computed. The graph represents the number of PDB entries (y axis) as a function of their selenomethionine density (x axis). (*b*) Monomeric model. The left panel shows the C$^\alpha$ trace of human p8/TTD-A (light grey) compared with that of the refined Tfb5 structure (orange). The right panel shows the C$^\alpha$ trace of residues 5–24 and 34–45 of Tfb5 (search model M3; red). (*c*) The C$^\alpha$ trace of the refined Tfb2C–Tfb5 structure (left panel) is shown in the same orientation as that of the truncated p8/TTD-A dimer (residues 5–24 and 34–45; search model M4; right panel).

Table 2
Molecular replacement.

Search probes positioned in the hexagonal form ($P6_1$; Tfb2$_{435–513}$–Tfb5$_{2–72}$) using *Phaser*. Trials were performed with models derived from the X-ray structure (1ydl). The refined structure of the Tfb5 subunit (3dgp chain *B*) was used as a control. For clarity, only data obtained with the mixed model are shown. The RFZ and TFZ (*Z* scores for the rotation and the translation function, respectively), as well as the log-likelihood gain (LLG), are from the *Phaser* solution file. Dm is the angular distance between the rotation peak and the solution (evaluated with *COMPANG*; Urzhumtseva & Urzhumtsev, 2002).

Model	Resolution (Å)	Code	Solution	Dm (°)	RFZ	TFZ	LLG	LLG	Rotation (Euler)	Translation (fractional)
3dgp chain *B*, 2–59	15–4.0	71	1/1		2.1	3.5	57	120	349.6 134.8 137.6	0.1489 0.0155 −0.1840
	15–2.5	72	1/1		6.4	13.7	110	120	349.8 134.8 137.7	0.1481 0.0149 −0.1823
M0 (1ydl), 1–71	15–4.0	70	—		—	—	—	—	—	—
	15–2.5	69	—		—	—	—	—	—	—
M1 (1ydl), 1–28, 32–60	15–4.0	151	—		—	—	—	—	—	—
	15–2.5	150	—		—	—	—	—	—	—
M2 (1ydl), 1–28, 32–50	15–4.0	94	2/11	20.8	3.4	3.6	15	18	76.0 140.9 174.1	—
	15–2.5	93	10/17	11.3	3.0	3.7	17	17	335.0 141.4 123.9	—
M3 (1ydl), 5–24, 34–45	15–4.0	66	9/40	15.3	2.5	4.5	13	13	306.7 133.7 159.9	—
	15–2.5	92	1/17	5.3	2.9	3.7	17	18	351.3 136.7 143.6	0.1666 0.0356 −0.1850
D3 (1ydl), 5–24, 34–45	15–4.0	96	1/23	2.9	3.1	4.0	19	29	350.5 137.5 139.7	0.1524 0.0201 −0.2017
	15–2.5	95	1/17	2.3	2.9	4.9	29	30	350.7 137.4 139.8	0.1521 0.0198 −0.1602

combined with the limited quality of our data set (Table 1) provides a reasonable explanation for the failure of our attempts to obtain experimental phases.

In the case of the hexagonal crystal form (Tfb2$_{435–513}$–Tfb5$_{2–72}$ complex) which diffracts to high resolution, the situation differs. A SAD data set was collected from an SeMet-modified complex at the peak wavelength for selenium (Table 1). In the resolution range 34–2.7 Å the average anomalous difference ($\Delta F/F$) is 5.2% and the signal-to-noise ratio is 0.7 (see Supplementary Table 5). Although weak, anomalous differences higher than the estimated errors of up to 3.5 Å are present.

After the deletion of outliers [$|F(+) - F(-)| > 3\sigma$], the anomalous Patterson calculated at the peak is interpretable. Harker sections and cross-peak analysis can be interpretated as a correct single ordered selenium site. SAD phasing leads to a quite low average FOM of 0.29 (resolution range 35–2.7 Å, output value from *SOLVE*), but interpretable electron density can nevertheless be obtained after density modification (using *RESOLVE*). *A posteriori* analysis with the anomalous difference Fourier map calculated with phases from the refined model shows that the ordered selenium site corresponds to SeMet27, which is located in a solvent-accessible loop of Tfb5 (α_1–β_2), and that the second site that is not detected corresponds to SeMet66, a residue that is disordered in the hexagonal crystal form.

3.5. Could the structure be solved by molecular replacement?

Tfb5 shares 28% sequence identity with p8/TTD-A, the three-dimensional structure of which has been determined both by X-ray crystallography and NMR. These structures constitute a suitable search model for molecular-replacement trials. At this stage, the fold of Tfb2 was unknown and could not be predicted. Therefore, molecular replacement was performed with p8/TTD-A monomers, which accounted for 50% of the structure of the complex.

In the case of the Tfb2$_{435–513}$–Tfb5$_{2–71}$ complex (hexagonal space group $P6_1/P6_5$, complete data set at 1.7 Å resolution), cell-content analysis indicated that the asymmetric unit is most likely to contain one heterodimer (V_M = 3.2 Å^3 Da^{-1}). Searches were conducted in a single-model mode using experimentally determined structures of human p8/TTD-A [PDB codes 1ydl (F. Forouhar, W. Edstrom, R. Xiao, T. B. Acton, G. T. Montelione, L. Tong & J. F. Hunt, unpublished work) and 2jnj (Vitorino *et al.*, 2007)]. Based on sequences analysis, several models have been constructed and used as a probe [initial unmodified 1ydl or 2jnj models, homology models (full, poly-Ala/poly-Ser) and mixed models]. Trials were performed with *Phaser* between 15 and 4 Å or between 15 and 2.5 Å using the full-length protein (model M0, residues 1–71) as well as several truncated versions (models M1, residues 1–28 and 32–60, and M2, residues 1–28 and 32–50) (Table 2). However, a clear solution did not emerge (*Z* scores for translation function of <6, poor contrast).

A posteriori, knowing the correct solution for this crystal form, the possibility of solving the structure using molecular replacement was reinvestigated and additional models were tested. The correct solution was found with a truncated version of the probe (1ydl) that corresponded to the core of the molecule (strands β_1-β_2-β_3 and helix α_1) when the high-resolution data set (at least 2.5 Å) was used for the translation function (Table 2). This partial model truncated to 32 residues (model M3, residues 5–24 and 32–45) corresponds to 21% of the structure. The solution ranked first was poorly contrasted, with a *Z* score for the translation search of 3.7. The solution ranked second had a *Z* score of 3.8 and was incorrect. Significant differences between the two molecules were observed in the α1–β2 loop as well as in the orientation of the C-terminal helix α2, which differed by 40° (Fig. 6*b*). This explains why searches with the full-length protein (model M0) and with probes in which these regions were still present (models M1 and M2) were unsuccessful. Having positioned the Tfb5 subunit of the complex, we used the different partial models in density modification as well as in automated model-building programs such as *DM* or *ARP*/*wARP*. Despite intensive efforts, we were not able to improve the quality of the phases and to complete the models.

The structure of the Tfb2C–Tfb5 complex revealed that Tfb5 and the C-terminus of Tfb2 adopt the same fold (r.m.s.d. of 1.9 Å for 42 equivalent C^{α} atoms) and that the Tfb2C–Tfb5 heterodimer mimics the p8/TDD-A homodimer (r.m.s.d. of 1.8 Å for 87 equivalent C^{α} atoms; Kainov *et al.*, 2008). Molecular-replacement trials using two independent copies of p8/TTD-A as a probe failed to locate the second copy corresponding to Tfb2C. However, searches using a truncated version of the p8/TTD-A homodimer (model D3, residues 5–24 and 32–45; Fig. 6*c*) led to the correct solution. This model, corresponding to 42% of the whole structure, was sufficient for automated model building. *ARP*/*wARP* run with default parameters built 60 additional amino acids, leading to a model composed of 114 residues with an *R* value of 19.6% at 1.8 Å resolution. This model is almost identical to our refined structure (PDB code 3dgp; r.m.s.d. of 0.097 Å for 114 equivalent C^{α} atoms, true sequence; Kainov *et al.*, 2008) and differs only by 11 residues that were not automatically traced. As mentioned above, the structural homology between Tfb2C and Tfb5 and p8/TTD-A could not be detected by sequence analysis or fold prediction and therefore the homodimer of p8/TTD-A was unfortunately not used as a probe for molecular replacement.

4. Conclusions

We have reported the process of the structure determination of the minimal complex between Tfb2 and Tfb5, which are two subunits of the transcription/DNA-repair factor TFIIH. The limited quality of the initial crystals was a bottleneck in the structure determination. These crystals diffracted to 2.6 Å resolution but were difficult to handle. Despite extensive efforts to control cryoprotection, only a minor proportion of the crystals exhibited reasonable diffraction and mosaicity, which hampered the possibility of solving the structure using heavy-atom derivatives. Of 250 crystals that were tested on synchrotron beamlines, only three yielded usable data sets. A key step was the shortening of the Tfb2 construct, which affected part of the Tfb2–Tfb5 interface as well as crystal packing along the *c* axis, leading to a new crystal form. These crystals diffracted to 1.7 Å resolution on a synchrotron beamline and to 2.9 Å in-house, which facilitated heavy-atom screening and structure determination.

This work was funded by CNRS, the INSERM and the Université de Strasbourg, and benefited from grants from the Agence Nationale de la Recherche (ANR-maladies rares ANR-05-MRAR-005-02 and ANR-08-042-02), the Association de la Recherche sur le Cancer and the European Commission (SPINE2-complexes contract No. LSHG-CT-2006-031220). DK was supported by EMBO-LTF. We thank Jean-Claude Thierry for constructive discussions, André Mitschler and Ruslan Sanishvili for help with data collection and the staff at the ESRF, Grenoble and APS, Chicago (GM/CA CAT, ANL). We also thank the members of the Structural Biology and Genomics platform and the members of the IGBMC's common services.

References

Andersen, G., Poterszman, A., Egly, J. M., Moras, D. & Thierry, J. C. (1996). *FEBS Lett.* **397**, 65–69.
Chang, W. H. & Kornberg, R. D. (2000). *Cell*, **102**, 609–613.
Cohen, S. X., Ben Jelloul, M., Long, F., Vagin, A., Knipscheer, P., Lebbink, J., Sixma, T. K., Lamzin, V. S., Murshudov, G. N. & Perrakis, A. (2008). *Acta Cryst.* D**64**, 49–60.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Deaconescu, A. M., Chambers, A. L., Smith, A. J., Nickels, B. E., Hochschild, A., Savery, N. J. & Darst, S. A. (2006). *Cell*, **124**, 507–520.
DeLano, W. L. (2008). *PyMOL Molecular Viewer.* DeLano Scientific LLC, Palo Alto, California, USA.
Dip, R., Camenisch, U. & Naegeli, H. (2004). *DNA Repair*, **3**, 1409–1423.
Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.
Fan, L., Arvai, A. S., Cooper, P. K., Iwai, S., Hanaoka, F. & Tainer, J. A. (2006). *Mol. Cell*, **22**, 27–37.
Fan, L., Fuss, J. O., Cheng, Q. J., Arvai, A. S., Hammel, M., Roberts, V. A., Cooper, P. K. & Tainer, J. A. (2008). *Cell*, **133**, 789–800.
Fribourg, S., Kellenberger, E., Rogniaux, H., Poterszman, A., Van Dorsselaer, A., Thierry, J. C., Egly, J. M., Moras, D. & Kieffer, B. (2000). *J. Biol. Chem.* **275**, 31963–31971.
Gervais, V., Busso, D., Wasielewski, E., Poterszman, A., Egly, J. M., Thierry, J. C. & Kieffer, B. (2001). *J. Biol. Chem.* **276**, 7457–7464.
Gervais, V., Lamour, V., Jawhari, A., Frindel, F., Wasielewski, E., Dubaele, S., Egly, J. M., Thierry, J. C., Kieffer, B. & Poterszman, A. (2004). *Nature Struct. Mol. Biol.* **11**, 616–622.
Hendrickson, W. A. & Teeter, M. M. (1981). *Nature (London)*, **290**, 107–113.
Kabsch, W. & Sander, C. (1983). *Biopolymers*, **22**, 2577–2637.
Kainov, D. E., Selth, L. A., Svejstrup, J. Q., Egly, J. M. & Poterszman, A. (2010). *DNA Repair*, **9**, 33–39.
Kainov, D. E., Vitorino, M., Cavarelli, J., Poterszman, A. & Egly, J. M. (2008). *Nature Struct. Mol. Biol.* **15**, 980–984.
Krissinel, E. & Henrick, K. (2007). *J. Mol. Biol.* **372**, 774–797.
Laine, J. P. & Egly, J. M. (2006). *Trends Genet.* **22**, 430–436.
Lolli, G., Lowe, E. D., Brown, N. R. & Johnson, L. N. (2004). *Structure*, **12**, 2067–2079.
McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.
Nanduri, B., Byrd, A. K., Eoff, R. L., Tackett, A. J. & Raney, K. D. (2002). *Proc. Natl Acad. Sci. USA*, **99**, 14722–14727.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Porollo, A. & Meller, J. (2007). *Proteins*, **66**, 630–645.
Potier, N., Lamour, V., Poterszman, A., Thierry, J. C., Moras, D. & Van Dorsselaer, A. (2000). *Acta Cryst.* D**56**, 1583–1590.
Schultz, P., Fribourg, S., Poterszman, A., Mallouh, V., Moras, D. & Egly, J. M. (2000). *Cell*, **102**, 599–607.
Terwilliger, T. C. & Berendzen, J. (1999). *Acta Cryst.* D**55**, 849–861.
Urzhumtseva, L. & Urzhumtsev, A. (2002). *J. Appl. Cryst.* **35**, 644–647.
Vagin, A. A., Steiner, R. A., Lebedev, A. A., Potterton, L., McNicholas, S., Long, F. & Murshudov, G. N. (2004). *Acta Cryst.* D**60**, 2184–2195.
Vitorino, M., Coin, F., Zlobinskaya, O., Atkinson, R. A., Moras, D., Egly, J. M., Poterszman, A. & Kieffer, B. (2007). *J. Mol. Biol.* **368**, 473–480.
Vonrhein, C., Blanc, E., Roversi, P. & Bricogne, G. (2007). *Methods Mol. Biol.* **364**, 215–230.
White, M. F. (2009). *Biochem. Soc. Trans.* **37**, 547–551.
Wolski, S. C., Kuper, J., Hanzelmann, P., Truglio, J. J., Croteau, D. L., Van Houten, B. & Kisker, C. (2008). *PLoS Biol.* **6**, e149.