Human Molecular Genetics, 2016, Vol. 25, No. 21 4717–4725

doi: 10.1093/hmg/ddw298
Advance Access Publication Date: 29 August 2016
Original Article

ORIGINAL ARTICLE

A critical examination of the recently reported crystal structures of the human SMN protein

Manfred S. Weiss[1], Kay Diederichs[2], Randy J. Read[3], Santosh Panjikar[4], Gregory D. Van Duyne[5], A. Gregory Matera[6], Utz Fischer[7] and Clemens Grimm[7,*]

[1]Helmholtz-Zentrum Berlin für Materialien und Energie, Macromolecular Crystallography, Berlin, Germany, [2]Department of Biology, University of Konstanz, Germany, [3]Department of Haematology, University of Cambridge, Cambridge Institute for Medical Research, Hills Road, Cambridge, UK, [4]Australian Synchrotron, Clayton, Australia, [5]Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA, [6]University of North Carolina, Chapel Hill, North Carolina, USA and [7]Departement of Biochemistry, Biocenter of the University, University of Wuerzburg, Würzburg, Germany

*To whom correspondence should be addressed at: Clemens Grimm, Department of Biochemistry, Biocenter of the University, University of Wuerzburg, Am Hubland, D-97074 Würzburg, Germany. Tel: +49 931 31 84031; Fax: +49 931 31 84028; Email: clemens.grimm@biozentrum.uni-wuerzburg.de

Abstract

A recent publication by Seng *et al.* in this journal reports the crystallographic structure of refolded, full-length SMN protein and two disease-relevant derivatives thereof. Here, we would like to suggest that at least two of the structures reported in that study are incorrect. We present evidence that one of the associated crystallographic datasets is derived from a crystal of the bacterial Sm-like protein Hfq and that a second dataset is derived from a crystal of the bacterial Gab protein. Both proteins are frequent contaminants of bacterially overexpressed proteins which might have been co-purified during metal affinity chromatography. A third structure presented in the Seng *et al.* paper cannot be examined further because neither the atomic coordinates, nor the diffraction intensities were made publicly available. The Tudor domain protein SMN has been shown to be a component of the SMN complex, which mediates the assembly of RNA-protein complexes of uridine-rich small nuclear ribonucleoproteins (UsnRNPs). Importantly, this activity is reduced in SMA patients, raising the possibility that the aetiology of SMA is linked to RNA metabolism. Structural studies on diverse components of the SMN complex, including fragments of SMN itself have contributed greatly to our understanding of the cellular UsnRNP assembly machinery. Yet full-length SMN has so far evaded structural elucidation. The Seng *et al.* study claimed to have closed this gap, but based on the results presented here, the only conclusion that can be drawn is that the Seng *et al.* study is largely invalid and should be retracted from the literature.

Introduction

The survival motor neuron (SMN) protein has attracted the attention of scientists working in basic and biomedical research alike. This is due to the fact that the encoding gene is mutated in the devastating disorder Spinal muscular atrophy (SMA) and that the protein fulfils a hitherto unknown function in the assembly of RNA-protein complexes.

Received: July 29, 2016. **Revised:** August 25, 2016. **Accepted:** August 26, 2016

Spinal muscular atrophy (SMA) is a common recessive genetic disease with an incidence of approximately 1 in 6000 live births and a carrier frequency of 1 in 35 (1,2). SMA is one of the leading genetic causes of early childhood death. The disease results in the loss of alpha motor neurons in the ventral horn of the spinal cord giving rise to progressive paralysis and premature death. In 1995, the Melki lab uncovered that the vast majority of SMA patients carry deletions or mutations in the survival motor neuron gene 1 (*SMN1*) (1), which is located within an inverted duplication on chromosome 5 along with a nearly identical paralog, *SMN2*. The two genes differ by five nucleotide changes, only one of which is located within the protein coding region, a C to T transition located at position 6 inside exon 7 of *SMN2* (3). This mutation greatly enhances skipping of SMN2 exon 7, resulting in production of a truncated polypeptide, called SMNΔ7 (4). *SMN2* does express low levels of the full-length transcript, whereas the vast majority of transcripts produced from *SMN1* are full-length and represent the predominant contributor of SMN protein levels. Although both transcripts are translated, the SMNΔ7 protein is unstable and degradation-prone, resulting in drastically reduced levels of functional SMN protein produced from the *SMN2* gene. In SMA patients, *SMN2* is the only source of functional SMN protein and thus reduced levels of SMN are highly correlated with the SMA phenotype.

Biochemical investigations of the SMN protein revealed that it is part of a macromolecular complex, together with Gemins 2-8 and unrip (5,6). This so-called SMN complex displays a modular composition in which SMN, Gemin2, and Gemin8 form the backbone onto which the peripheral building blocks Gemin3/4, Gemin6/7 as well as Unrip bind to form the functional unit (7). The SMN complex functions in the biogenesis of pre-mRNA processing UsnRNP particles. In this reaction, a set of 7 Sm proteins is loaded onto the snRNA to form the common core structure of these RNPs. How this reaction is facilitated is largely unknown and can best be understood by a combined approach involving functional biochemistry and structural biology. Towards this end, several laboratories have attempted to crystallize components of the SMN complex.

Of note, the experimental structure determination of the isolated full length SMN protein is exceedingly difficult due the presence of extended (proline-rich) unstructured regions, as well as its tendency to form high molecular weight oligomers of variable stoichiometry. In a recent publication, Seng *et al.* (8) report on the refolding, crystallization and structure determination of full-length SMN, the disease-relevant truncated SMN species SMNΔ7 (lacking exon 7) and an SMN1-4 variant, which is lacking residues encoded by exons 5, 6 and 7. Here, we provide strong evidence that the SMNΔ7 diffraction dataset used in this study is, in fact, derived from a crystal of the bacterial Hfq protein and that the diffraction dataset ascribed to full-length SMN is actually derived from a crystal of the bacterial Gab protein. Therefore, both of the reported structures (8) have to be regarded as incorrect. As neither the coordinates nor the dataset of the SMN1-4 structure have been submitted to the PDB nor have been made available to the public in any other way, no valid information towards the structure of the SMN protein or any of its variants can be deduced from the Seng *et al.* paper.

Results

Reported structure determinations of refolded SMN protein variants

Seng *et al.* (8) presented the crystal structures of three variants of bacterially expressed and refolded SMN protein. These include the full-length SMN protein, a truncated variant found in SMA patients expressed from mRNA lacking exon7 (SMNΔ7) and a variant lacking exons 5, 6 and 7 (SMN1-4). Their work is based on a 2.7 Å dataset collected from SMN1-4 crystals (completeness 99.9%, R_{Merge} 15.8%, I/σ(I) 6.4), a 3.0 Å dataset collected from SMNΔ7 crystals (completeness 88.9%, R_{Merge} 13.2%, I/σ(I) 4.1), and a low resolution dataset collected at 5.5 Å resolution from crystals grown from full-length SMN (completeness 61.3%, R_{Merge} 14.5%, I/σ(I) 4.6). Atomic coordinates and diffraction data for the SMNΔ7 and SMN structures were deposited to the PDB (PDB-Ids 4NL7 and 4NL6, respectively), but no data were deposited for the SMN1-4 structure.

The SMN1-4 structure was initially solved by molecular replacement (MR) using a high resolution structure of the SMN Tudor domain (PDB ID code 1MHN, (9)) as a search model. Using the positioned Tudor domain (residues 89–147), the remainder of the structure (residues 1–88 and 148–196) was fitted to the MR-phased electron density without experimental phasing and without the use of NCS symmetry. The final model was refined to R_{work}/R_{free} factors of 22.8%/28.8%. This structure was subsequently used as a search model for MR with the assumed SMNΔ7 dataset (PDB entry 4NL7), where residues 197–282 were fitted into the model-phased density. Refinement of the final SMNΔ7 model resulted in R_{work}/R_{free} factors of 32.7%/34.0% (see Table 1 for the complete statistics). Finally, the SMNΔ7 structure was used as a search model for MR with the low resolution SMN dataset (PDB entry 4NL6). Three copies of the model were placed, the C-termini (residues 279–294) were traced and built into density and the Cα-only model was refined to R_{work}/R_{free} factors of 30.6%/32.9%.

The three SMN structures by Seng *et al.* are remarkable with regard to the presence of two domains which exhibit completely unprecedented folds: the N-domain comprising residues 1–86 and the C-domain comprising residues 208–282. The C-domain is particularly surprising in that it includes a folded proline-rich region with poly-proline helices. In addition, the structure contains a second Tudor domain (Tudor-2; residues 151–207) that was not predicted by bioinformatics approaches from its amino acid sequence. Likewise, to our best knowledge, no secondary structure prediction algorithm is able to predict the Tudor domain secondary structure assigned for Tudor-2.

Re-evaluation of the 4NL7 coordinates and dataset

A closer inspection of the deposited SMNΔ7 model (PDB ID code 4NL7) revealed pronounced discrepancies in side chain orientations when compared to the model of the isolated Tudor domain (9). Strikingly, the majority of those residues that build up the hydrophobic core in the high resolution SMN Tudor domain structure (1MHN) point outwards from the centre of the Tudor-1 domain in 4NL7, such that its hydrophobic core appears exploded. Likewise, the three other domains do not possess any distinct hydrophobic core, as a strikingly large fraction of their hydrophobic residues are highly solvent exposed. The lack of physiological packing of the amino acid side chains is reflected in a packing score calculated by the program WHATIF (10) of -3.6 for the residue range 1–272. This qualifies the 4NL7 crystallographic model as "*certain to be incorrect*". Another striking feature of the model is that despite its good stereochemistry (85% of the residues in the preferred region of the Ramachandran plot, no residues in the disallowed region according to Phenix (11), Table 1), the expected backbone-backbone interactions for helices and β-sheets are either greatly distorted or fully absent.

Given the highly unusual, or even improbable, features of the 4NL7 model, we reasoned that there could be a fundamental

Table 1. Model and refinement statistics for 4NL7/Hfq

	4NL7, Seng *et al.* publication	4NL7, calculated by Phenix	4NL7, Re-refined with Phenix	Hfq model, Refined against reindexed 4NL7 dataset
Space group	C2	C2	C2	$P2_1$
Cell constants (Å, Å, Å, °,°,°)	107.1, 62.3, 57.1, 90.0, 95.1, 90.0	107.1, 62.3, 57.1, 90.0, 95.1, 90.0	107.1, 62.3, 57.1, 90.0, 95.1, 90.0	107.1, 62.3, 57.1, 90.0, 95.1, 90.0
Resolution (Å)	3.0	14 - 3.0 (3.11 - 3.00)	14 - 3.0 (3.11 - 3.00)	14 - 3.0 (3.11 - 3.00)
No. of reflections				
In refinement	7286 {6702}	6702 (657)	6702 (657)	6702 (657)
For R_{free}	–	312 (30)	312 (30)	312 (30)
Dataset completeness (%)	88.9 {96.7}	88.9	88.9	44.5
R_{free} (%)	**34.0 {29.6}**	**36.6 (31.2)**	**42.0 (34.2)**	**30.8 (33.7)**
R_{work} (%)	32.7 {29.6}	36.1 (33.0)	33.5 (31.7)	28.2 (33.0)
Protein residues	–	272	272	792
Water molecules	–	20	20	0
RMS (bonds) (Å)	0.009 {0.013]	0.017	0.011	0.004
RMS (angles) (°)	1.74 {1.93}	3.08	1.56	0.73
Ramachandran favoured (%)	100	85	66	95
Ramachandran allowed (%)	–	15	26	4.7
Ramachandran outliers (%)	0	0.37	8.1	0
Clashscore	–	63.1	38.1	7.7
Average isotropic B-factor (Å^2)	– {26.7}	26.9	17.3	92.8
– Protein	–	26.9	17.4	92.8
– Water	–	20.0	10.4	–

Values in () parentheses for highest resolution shell. Values in {} parentheses given if deviating information available in 4NL7 PDB header. Model statistics calculated with Phenix.

problem with the crystallographic structure solution. We therefore examined the underlying crystallographic datasets used by Seng *et al.* (8). To our great surprise we discovered that neither the SMN1-4 diffraction dataset nor the SMN1-4 model was deposited with the PDB. As SMN1-4 was the initially solved structure that then served as an MR template for the other two crystal structures presented, it is impossible to repeat the MR calculations used to determine the full-length SMN and SMNΔ7 crystal structures.

The second dataset presented in the Seng *et al.* study is derived from a crystal of full-length SMN and extends to only about 5.5 Å resolution. It also exhibits a rather low completeness of 61.3%, whereas usually a completeness in the 90% region or higher is desired. We therefore decided to focus first on the presumptive SMNΔ7 dataset extending to 3.0 Å resolution and exhibiting a reasonable completeness of 88.9% in space group C2. The overall R_{Merge} of 13.2% for this dataset is high, but still in a reasonable range while the R_{Merge} for the outer shell is comparably low with 25.0%. Finally, its low overall I/σ(I) value of 4.1 is unusual and would be characteristic for very weak data or could point to a processing or a space group problem. Importantly, the values deposited in the PDB for the 4NL7 dataset (completeness 96.7%, R_{Merge} 35.2%, no I/σ(I) value deposited) do not correspond to those presented in the paper. It might therefore be possible that the dataset used in the published work is different from that deposited with PDB entry 4NL7, however the values for the cell constants are identical and the actual completeness corresponds to the value published in the paper. If the extremely high overall R_{Merge} reported in the PDB were true, it would point to a severe problem with data processing, symmetry determination or data collection. Despite several requests made to the authors of the original publication by email, no unmerged or raw data were made available. This left us with the deposited merged datasets to solve the issues presented above.

Molecular replacement of the purported 4NL7 dataset with Hfq as search model

Many crystal structures display symmetric properties that cannot be described by crystallographic symmetry operators. If such a non-crystallographic symmetry (NCS) has a translational character, it can often be detected by calculating a native Patterson map, if it has a rotational characteristic, it can often be detected by the calculation of the self-rotation function of the crystal structure or the dataset. The native Patterson map of the deposited dataset shows a peak at 21% of the origin peak, indicating significant translational NCS. The self-rotation function of the data shows a pronounced peak of 71% of the crystallographic peak height for the $\kappa = 60°$ and of 72% of the crystallographic peak height for the $\kappa = 180°$ section (Fig. 1 and Table 2). These symmetry features are indicative of a 622 point group symmetry. Since the basic architecture of the 4NL7 model is not symmetric in any way, these features could not be reproduced by calculated structure factors derived from the model. This then led to the consideration of the possibility that the dataset might have been collected from a crystal of a different, unrelated protein.

As the SMN protein was overexpressed in *Escherichia coli* and purified and refolded on a nickel-NTA affinity chromatography column, we inferred that the protein in question might be a hexameric bacterial protein with affinity for Ni-NTA. As the *E. coli* Hfq protein is known to possess both of these properties and one or two Hfq hexamers (Fig. 2A) would fit the asymmetric unit, we considered this to be a likely candidate for a crystallized contaminant. Likewise, we noted that several Hfq entries in the PDB, e.g. 4RCB (12) showed a remarkable similarity of unit cell parameters to those of entry 4NL7. In addition, the primitive cell for 4NL7 features striking resemblance to several *E. coli* Hfq datasets deposited to the PDB (Table 3). A MR run in Phaser (13) with the biological, hexameric unit of 4RCB and the 4NL7

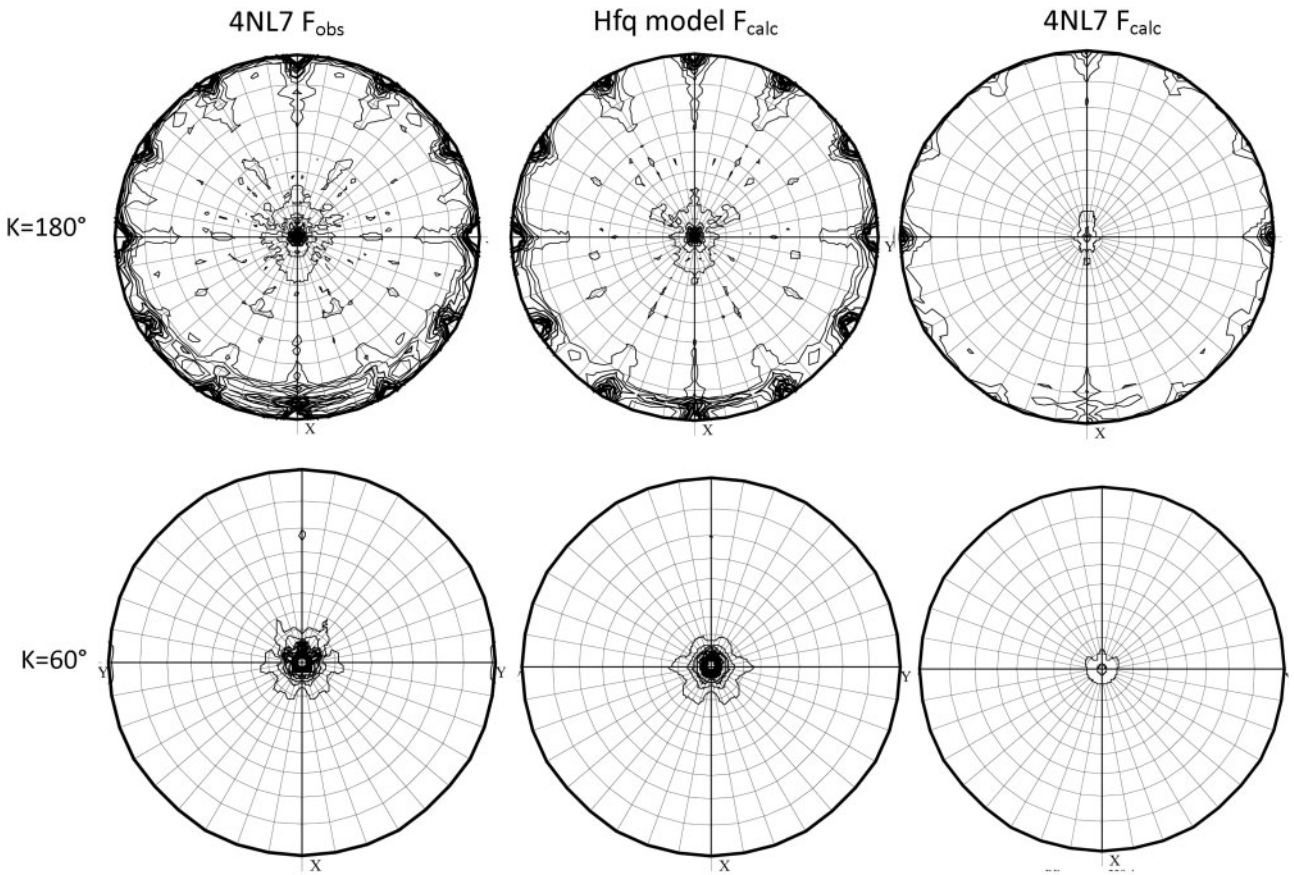

Figure 1. Self rotation function of the 4NL7 dataset, our Hfq model and the 4NL7 model. The three self-rotation functions were scaled relative to one another so that the contour lines are at approximately the same absolute self rotation function value. The $\kappa = 180°$ sections for each of the three self rotation functions are displayed in the upper row, the $\kappa = 60°$ sections in the lower row.

Table 2. Self-rotation function peaks at $\kappa = 60°$ and $\kappa = 180°$ indicative of a 622 point group symmetry as percentage of crystallographic peak height for the 4NL7 dataset, our Hfq model, and the 4NL7 model

Resolution	3 Å		5 Å		6 Å	
κ (°)	60	180	60	180	60	180
4NL7 dataset	71%	72%	64%	57%	61%	66%
Hfq model	78%	77%	71%	68%	79%	79%
4NL7 model	31%	32%	26%	27%	n. d.	30%

dataset in the original C2 space group instantly yielded Z-scores of roughly 15 for the rotation function and Z-scores of more than 8 for the translation function, indicative for a high probability of a correct MR solution. In total, two independent solutions for the hexamer were found of which one was rejected by the packing test. The significance of the second, rejected solution will be discussed later.

Visual inspection of the crystal lattice showed large gaps, indicating an incomplete solution. We therefore assumed that the space group might have been wrongly assigned and expanded the dataset to space group P1. We also considered that the lattice type might have been determined wrongly because of overlooked weak reflections. Assuming a primitive instead of a C-centred lattice with the same cell results in a P2 or $P2_1$ dataset with half the original completeness. Re-running Phaser with these re-indexed datasets, a convincing solution could be found in space group $P2_1$ with two tightly packed hexamers in the asymmetric unit (Fig. 2B and C) that without further manual adjustments of the model or the refinement protocol refined to an R_{free} of 34.9% using Phenix.refine with NCS restraints imposed on the 12 Hfq protein chains within the asymmetric unit. Further refinement with the high resolution *Salmonella typhimurium* Hfq structure (14) from PDB entry 2YLB as a reference model and comprising TLS refinement resulted in a model with R_{work}/R_{free} factors of 28.2%/30.8% and excellent geometry. The observed R-factors are still in the acceptable range for a structure of the observed resolution, but might be elevated due to symmetry or other problems of the dataset, which is discussed below. Finally, we note that the amino acid sequences of bacterial Hfq and SMN do not display any obvious sequence homology and hence cannot be reasonably aligned with the usual bioinformatics tools.

Comparison between our hfq model and the 4NL7 SMN model

The R_{work}/R_{free} factors registered in the PDB for entry 4NL7 are 29.5%/29.5%, the values reported in the publication are 32.7%/34.0%. Phenix calculates values of 36.0/36.6 for the deposited model and dataset. During re-refinement with Phenix.refine the R_{free} quickly rises to values above 40% and the model stereochemistry deteriorates with Ramachandran outliers in the 8%

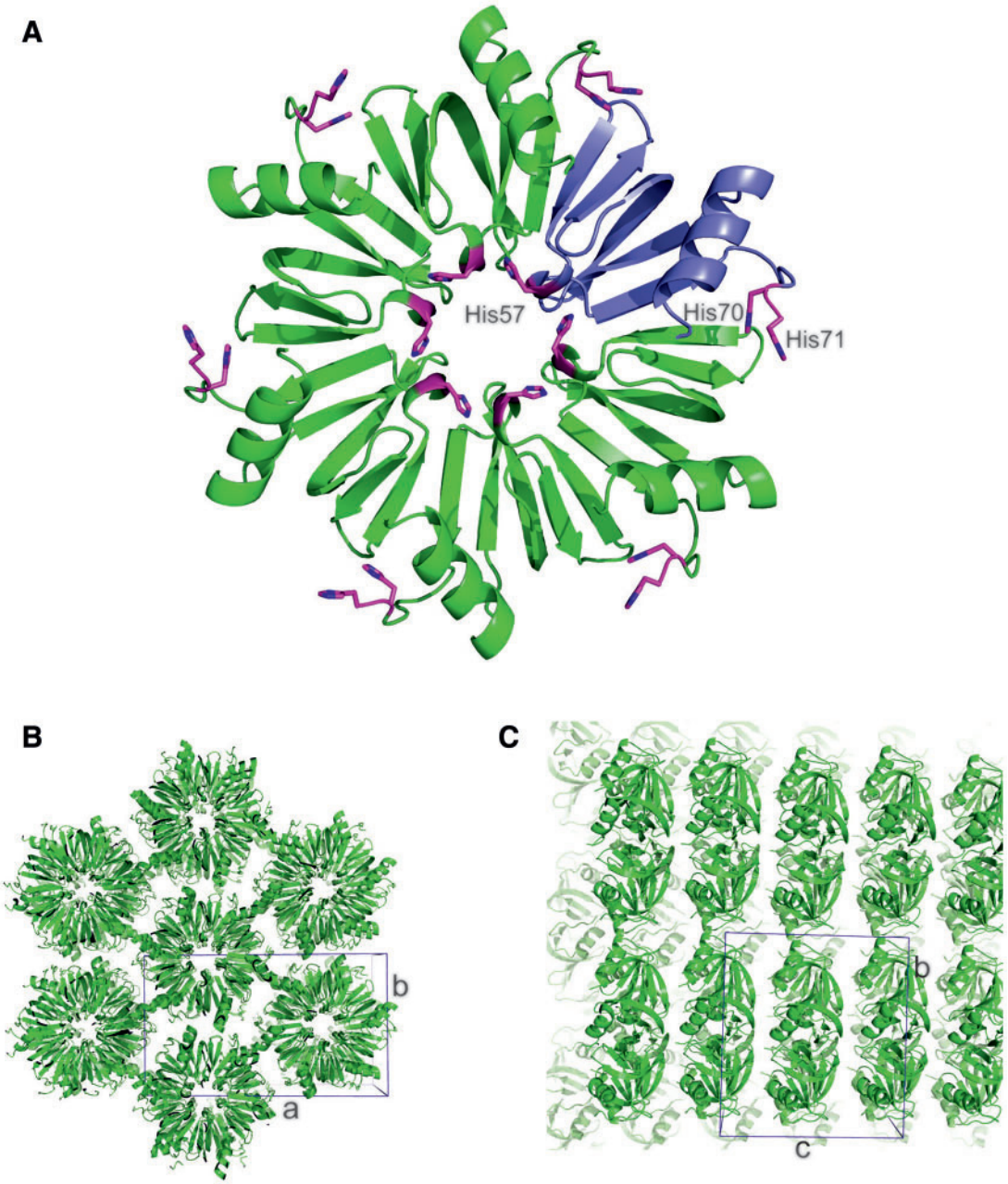

Figure 2. Crystal structure of the *E. coli* Hfq protein solved from the 4NL7 dataset. (**A**) - The Hfq hexamer in ribbon depiction. One protomer is coloured in purple, the His residues are shown as pink sticks. (**B**) - View of the crystal packing along the c axis. (**C**) - View of the crystal packing along the a axis.

Table 3. *E. coli* Hfq entries in the PDB with cell constants (axis lengths in A, cell angles in °) comparable to the 4NL7 cell reduced to P1

PDB ID code	Space group	a	b	c	α	β	γ
4NL7	Reduced to P1	62.0	62.0	94.4	57.1	94.4	60.4
2YHT	P1	61.2	61.2	53.1	82.6	87.3	60.0
3QHS	P1	61.9	62.2	81.3	78.6	86.2	59.9

range (Table 1). It therefore seems likely that the refinement of the deposited model was carried out with strong Ramachandran restraints and that the R_{free} set might have changed after the final refinement. In fact, the internal evidence in the coordinates deposited with the PDB reveals that the structure was actually refined without cross-validation, using a least-squares target instead of the currently preferred maximum likelihood target (which would have required the use of cross-validation data). In the deposited PDB entry, REMARK records state that all of the data belonged to the Rfree set, and Rwork and Rfree have identical values. Given these inconsistencies we refer to the refinement statistics generated from the 4NL7 model after re-refinement with five macrocycles with the standard protocol in Phenix.refine (Table 1) in the discussion that follows.

Based on refinements using Phenix, our refined Hfq model fits with a roughly 10% better R_{free} to the 4NL7 dataset than does the deposited 4NL7 model. To further corroborate the

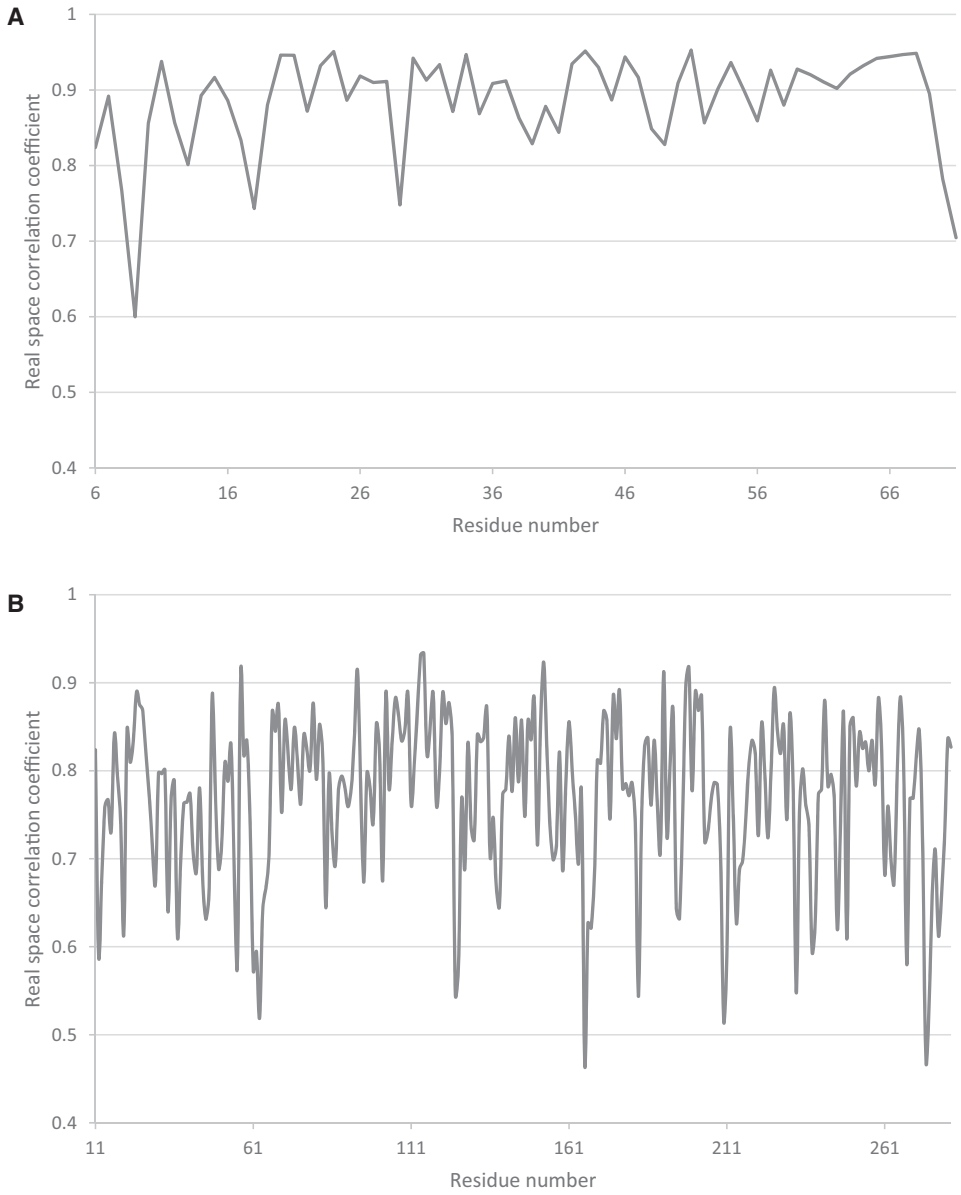

Figure 3. Real space correlation coefficient plot as a function of residue number. (**A**) Plot for peptide chain A of our Hfq model refined against the 4NL7 data in space group $P2_1$. (**B**) Plot for the SMN model from Seng *et al.* from PDB entry 4NL7 without modifications.

assumption that the 4NL7 model might be wrong, we compared the self-rotation function of the data to the self-rotation function of each of the two models at different resolutions (Table 2). The significant peaks for $\kappa = 60^\circ$ and $\kappa = 180^\circ$ at around 70% of the crystallographic symmetry peak height for the data indicative of a 622 point group symmetry is reproduced well in our Hfq model at all tested resolutions. Conversely, the 4NL7 model features only lower peaks at $\kappa = 60^\circ$ and $\kappa = 180^\circ$ that quickly diminish for $\kappa = 60^\circ$ when the self-rotation function is calculated at lower resolution. This observed resolution dependency for the self-rotation function of the 4NL7 model is most likely due to adaptation of finer model details to the dataset during the automated refinement.

Despite the low (44.5%) completeness of the dataset in space group $P2_1$, our Hfq model shows reasonable real space correlation coefficients; only the N- and C-termini are less well defined (Fig. 3A). In contrast, the 4NL7 SMN model has significantly worse correlation coefficients with no discernible trend (Fig. 3B). We thus conclude that our model fits the 4NL7 data better than the SMN model reported by Seng *et al.* in all investigated aspects and hence the SMNΔ7 model presented by Seng *et al.* cannot be correct.

Examination of the 4NL6 dataset attributed to full-length SMN

The findings reported above at the same time challenge the model of the full-length SMN protein presented by Seng *et al.* as this low resolution Cα structure was solved by molecular replacement with SMNΔ7 as a search model. We therefore moved on to examine also the dataset attributed to a crystal of the full-length SMN protein by Seng *et al.* (PDB entry 4NL6). These data were

Table 4. Model and refinement statistics for 4NL6/Gab

	4NL6, Seng *et al.* publication	4NL6, calculated by Phenix	4NL6 dataset, reindexed and refined against bacterial Gab
Space group	C2	C2	I422
Cell constants (Å, Å, Å, °,°,°)	137.0, 169.8, 108.8 90.0, 128.5, 90.0	137.0, 169.8, 108.8 90.0, 128.5, 90.0	120.2, 120.2, 137.0 90.0, 90.0, 90.0
Resolution (Å)	5.5	5.5 (5.649 - 5.5)	5.5 (5.649 - 5.5)
No. of reflections			
In refinement	7946	4511 (428)	4511 (429)
For R_{free}	–	198 (22)	198 (22)
Dataset completeness (%)	61.3	71.0	71.0
R_{free} (%)	**32.9 {34.2}**	**48.3 (46.6)**	**25.0 (30.8)**
R_{work} (%)	30.6 {29.9}	45.4 (49.2)	15.3 (21.0)
Protein residues	–	880	1228
Water molecules	–	–	0
RMS (bonds) (Å)	–	–	0.010
RMS (angles) (°)	–	–	1.25
Ramachandran favoured (%)	–	–	97
Ramachandran allowed (%)	–	–	2.7
Ramachandran outliers (%)	–	–	0.66
Clashscore	–	30.7	10.92
Average isotropic B-factor (Å^2)	–	30.7	199

Values in () parentheses for highest resolution shell. Values in {} parentheses given if deviating information available in 4NL6 PDB header. Model statistics calculated with Phenix.

processed in space group C2, with cell dimensions a = 137.0Å, b = 169.8Å, c = 108.8Å and beta = 128.5°. However, the program phenix.xtriage (11) suggested that the true symmetry of the data set could be I422 or $I4_122$, with cell dimensions a = b = 120.2 Å, and c = 137.0 Å. A search of the PDB for similar cell dimensions and one of these space groups came up with two possibilities, either D-amino acid oxidase from *Rhodosporidium toruloides* (PDB entry 1C0I, (15)) or the Gab protein from *E. coli* (PDB entry 1JR7, (16)). Of the two, the Gab protein seemed a much more likely candidate, both because it comes from the expression organism and because it has been identified as a contaminant that can yield crystals (17). Molecular replacement trials were carried out in Phaser, searching for four copies of either model in the reported C2 space group. The search for four copies of Gab yielded a very clear unique solution, with a final translation function Z-score of 21.6 and a log-likelihood-gain of 632, indicating a most likely correct solution. The solution obeys I422 symmetry, like entry 1JR7. The resulting model could then be refined to R_{work}/R_{free} factors of 15.3%/24.9%. We note that most of the statistics included with the deposited 4nl6 model do not match those given in the original publication. In particular, the R_{free} calculated by phenix amounts to 48.3% (a value which is unacceptable for a final refined model) rather than 32.9%, given in the publication (see Table 4 for comprehensive model and refinement statistics). We finally conclude that our model fits the 4NL6 data significantly better than the SMN model provided by Seng *et al.* Hence, the data for what was reported as full-length SMN is unambiguously from the *E. coli* Gab protein, which could have crystallized as a contaminant.

Discussion

Immobilized metal affinity chromatography (IMAC) is one of the most common protein purification techniques used in crystallography and other fields of molecular biology. Typically, Ni ions bound to chromatographic media interact with a 6-10 residue histidine tag at the N- or C-terminus of the overexpressed protein. It is obvious that proteins from the overexpression host possessing clusters of histidine residues or an elevated content of this amino acid in flexible regions are possible contaminants to this purification method. In fact, Hfq (18) and Gab are, among other bacterial contaminants (19), well-known impurities in IMAC-purified samples. Both crystallize readily, Hfq in a remarkable number of different crystal forms (20,21). It is therefore expedient for the protein crystallographer to be alert for false-positive crystallization hits from non-target proteins. Notably, for the 4NL7 dataset the ContaMiner webserver (22) identifies Hfq as the highest scoring molecular replacement template from its database of likely crystallization contaminants (23). Consequently, during the process of structure solution, well-established quality indicators should be observed critically. In the present case, the unfortunate combination of limited resolution and limited quality of the dataset, structural similarities (24) between the intended target protein and the actually crystallized contaminant and the choice of the phasing method might have contributed to refinement of an incorrect structure.

The evidence presented in this study unambiguously proves that the dataset used by Seng *et al.* to solve the SMNΔ7 structure is indeed derived from a crystal of bacterial Hfq. Unless raw or unmerged data are made available, the actual crystal symmetry of PDB entry 4NL7 cannot be verified, and the considerable fraction of missing reflections of the reindexed $P2_1$ dataset cannot be recovered. The low completeness of 44.5% would, under normal circumstances, disqualify the dataset as useful for a structure determination. However, the application of 12-fold NCS in conjunction with the fact that a near-atomic resolution crystal structure could be used as a reference model allowed a robust refinement of the Hfq structure against the 4NL7 data. We also note that our MR runs identified a second solution with equal likelihood scores that was discarded because of an elevated number of clashes. Interestingly, the relationship between the two structures is that they differ by a fractional translation of the second hexamer of 1/2, 1/2, 0, which corresponds to the crystallographic

C-centering operator that was discarded by expanding the original dataset from C2 to $P2_1$. The structure factors calculated for the two models are exactly the same for the h + k even reflections that are present in the expanded data, whereas the h + k odd reflections that would be able to distinguish between them are missing. Because the h + k odd reflections are missing in our reindexed dataset, the electron density for the second hexamer is actually expected to be an average of the densities for the two possible MR solutions. This agrees well with our actual observation of a significantly weaker density and higher overall B factor for the second hexamer. It would require access to the original diffraction data to sort out exactly what is going on, as there may also be complications from statistical disorder, twinning or possibly other crystallographic pathologies.

Overall, the good refinement statistics (Table 1), notably the low R_{free} factor of 30.8% and the good agreement of the model self-rotation function and the self-rotation function of the data (Table 2) support the fundamental correctness of our model. We therefore conclude that the SMNΔ7 model presented by Seng *et al.* is incorrect. We furthermore have provided clear evidence that the dataset used to solve the structure of full-length SMN is derived from a crystal of bacterial Gab as, after reindexing to I422, our molecular replacement solution of this bacterial protein refines to a R_{free} value that qualifies this solution as almost certain to be correct. Finally, neither the coordinates nor the dataset of the SMN1-4 model have been submitted to the PDB or made available in any other way. Therefore, the results presented in Seng *et al.* are either wrong or not verifiable in their entirety and all conclusions based on the three models are hence invalid. We would also like to take this opportunity and call upon the authors of the Seng *et al.* paper to make their diffraction images available to the community to that the data can be reprocessed and that this case can definitely be closed.

Materials and Methods

Self-rotation functions were calculated with Molrep (25) using a 30 Å integration radius. In order to make the three self-rotation functions depicted in Figure 1 comparable, they were scaled relative to one another so that the contour lines are at approximately the same absolute self-rotation function value. Molecular replacement was carried out with Phaser (13) with data up to 3 Å, 4 Å and 4.5 Å resolution, respectively, for 4NL7 and with all data to 5.5 Å for 4NL6. At all resolutions for 4NL7 similar top solutions for Hfq were observed, and a unique, unambiguous solution was obtained for 4NL6 using the Gab protein as a model. After minor manual corrections at the C-termini in Coot (26), the Hfq model containing two hexamers in the asymmetric unit was subjected to 8 macrocycles of automated refinement with Phenix.refine (11) under application of 12-fold NCS restraints including overall B-factor refinement. After eight more macrocycles cycles of refinement in Phenix performed with additional torsion angle restraints generated from PDB entry 2YLB (14) as a reference model and TLS refinement treating each hexamer as a single group the refinement converged. The Gab model was refined against the reindexed 4NL6 dataset with phenix.refine, restraining non-crystallographic symmetry and using 1JR7 as a reference model because of the low resolution of 5.5 Å for this data set. The model quality of the resulting Hfq model, the 4NL7 and 4NL6 (8) models and the Gab model was assessed and documented (Tables 1 and 4) with Phenix.

Acknowledgements

U. F. and C. G. were supported by a grant of the Deutsche Forschungsgemeinschaft. R. J. R. was supported by a grant of the Wellcome Trust.

Conflict of Interest statement. None declared.

Funding

U. F. and C. G. were supported by grant Fi573 /7-2 of the Deutsche Forschungsgemeinschaft. R. J. R. was supported by Principal Research Fellowship No. 082961/Z/07/Z of the Wellcome Trust. Funding to pay the Open Access publication charges for this article was provided by the University of Würzburg and the Helmholtz-Zentrum Berlin für Materialien und Energie GmbH.

References

1. Lefebvre, S., Burglen, L., Reboullet, S., Clermont, O., Burlet, P., Viollet, L., Benichou, B., Cruaud, C., Millasseau, P., Zeviani, M., *et al.* (1995) Identification and characterization of a spinal muscular atrophy-determining gene. *Cell*, **80**, 155–165.
2. Wirth, B. (2000) An update of the mutation spectrum of the survival motor neuron gene (SMN1) in autosomal recessive spinal muscular atrophy (SMA). *Hum. Mutat.*, **15**, 228–237.
3. Monani, U.R., Lorson, C.L., Parsons, D.W., Prior, T.W., Androphy, E.J., Burghes, A.H. and McPherson, J.D. (1999) A single nucleotide difference that alters splicing patterns distinguishes the SMA gene SMN1 from the copy gene SMN2. *Hum. Mol. Genet.*, **8**, 1177–1183.
4. Lorson, C.L., Hahnen, E., Androphy, E.J. and Wirth, B. (1999) A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. *Proc. Natl. Acad. Sci. U S A*, **96**, 6307–6311.
5. Cauchi, R.J. (2010) SMN and Gemins: 'we are family' ... or are we?: insights into the partnership between Gemins and the spinal muscular atrophy disease protein SMN. *Bioessays*, **32**, 1077–1089.
6. Matera, A.G. and Wang, Z. (2014) A day in the life of the spliceosome. *Nat. Rev. Mol. Cell. Biol.*, **15**, 108–121.
7. Otter, S., Grimmler, M., Neuenkirchen, N., Chari, A., Sickmann, A. and Fischer, U. (2007) A comprehensive interaction map of the human survival of motor neuron (SMN) complex. *J. Biol. Chem.*, **282**, 5825–5833.
8. Seng, C.O., Magee, C., Young, P.J., Lorson, C.L. and Allen, J.P. The SMN structure reveals its crucial role in snRNP assembly. *Hum. Mol. Genet.*, **24**, 2138–2146.
9. Sprangers, R., Groves, M.R., Sinning, I. and Sattler, M. (2003) High-resolution X-ray and NMR structures of the SMN Tudor domain: conformational variation in the binding site for symmetrically dimethylated arginine residues. *J. Mol. Biol.*, **327**, 507–520.
10. Hooft, R.W., Vriend, G., Sander, C. and Abola, E.E. (1996) Errors in protein structures. *Nature*, **381**, 272.
11. Adams, P.D., Afonine, P.V., Bunkoczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.W., Kapral, G.J., Grosse-Kunstleve, R.W., *et al.* (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.*, **66**, 213–221.
12. Feng, S.Q., Si, Y.L., Song, C.Y., Wang, P.Q. and Ji-Yong, S. (2015) Limited proteolysis improves *E. coli* Hfq crystal structure resolution. *Chinese J. Biochem. Mol. Biol.*, **31**, 1102–1108.

13. McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C. and Read, R.J. (2007) Phaser crystallographic software. *J. Appl. Crystallogr.*, **40**, 658–674.
14. Sauer, E. and Weichenrieder, O. Structural basis for RNA 3'-end recognition by Hfq. *Proc. Natl. Acad. Sci. U S A*, **108**, 13065–13070.
15. Pollegioni, L., Diederichs, K., Molla, G., Umhau, S., Welte, W., Ghisla, S. and Pilone, M.S. (2002) Yeast D-amino acid oxidase: structural basis of its catalytic properties. *J. Mol. Biol.*, **324**, 535–546.
16. Chance, M.R., Bresnick, A.R., Burley, S.K., Jiang, J.S., Lima, C.D., Sali, A., Almo, S.C., Bonanno, J.B., Buglino, J.A., Boulton, S., *et al.* (2002) Structural genomics: a pipeline for providing structures for the biologist. *Protein Sci.*, **11**, 723–738.
17. Lohkamp, B. and Dobritzsch, D. (2008) A mixture of fortunes: the curious determination of the structure of Escherichia coli BL21 Gab protein. *Acta Crystallogr. D Biol. Crystallogr.*, **64**, 407–415.
18. Milojevic, T., Sonnleitner, E., Romeo, A., Djinovic-Carugo, K. and Blasi, U. False positive RNA binding activities after Ni-affinity purification from Escherichia coli. *RNA Biol.*, **10**, 1066–1069.
19. Robichon, C., Luo, J., Causey, T.B., Benner, J.S. and Samuelson, J.C. Engineering Escherichia coli BL21(DE3) derivative strains to minimize E. coli protein contamination after purification by immobilized metal affinity chromatography. *Appl. Environ. Microbiol.*, **77**, 4634–4646.
20. Hammerle, H., Beich-Frandsen, M., Vecerek, B., Rajkowitsch, L., Carugo, O., Djinovic-Carugo, K. and Blasi, U. Structural and biochemical studies on ATP binding and hydrolysis by the Escherichia coli RNA chaperone Hfq. *PLoS One*, **7**, e50892.
21. Wang, W., Wang, L., Zou, Y., Zhang, J., Gong, Q., Wu, J. and Shi, Y. Cooperation of Escherichia coli Hfq hexamers in DsrA binding. *Genes Dev.*, **25**, 2106–2117.
22. Hungler, A., Momin, A., Diederichs, K. and Arold, S.T. (2016) ContaMiner: a webserver for early identification of unwantedly crystallised protein contaminants. *J. Appl. Cryst.*, submitted.
23. Bolanos-Garcia, V.M. and Davies, O.R. (2006) Structural analysis and classification of native proteins from E. coli commonly co-purified by immobilised metal affinity chromatography. *Biochim. Biophys. Acta*, **1760**, 1304–1313.
24. Selenko, P., Sprangers, R., Stier, G., Buhler, D., Fischer, U. and Sattler, M. (2001) SMN tudor domain structure and its interaction with the Sm proteins. *Nat. Struct. Biol.*, **8**, 27–31.
25. Vagin, A. and Teplyakov, A. Molecular replacement with MOLREP. *Acta Crystallogr. D Biol. Crystallogr.*, **66**, 22–25.
26. Emsley, P., Lohkamp, B., Scott, W.G. and Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.*, **66**, 486–501.