Dissertation

submitted to the

Combined Faculty of Natural Sciences and Mathematics

of the Ruperto Carola University Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

Presented by

M.Sc. (Computer Science) Karen Katrina Manalastas-Cantos

born in: Quezon City, Philippines

Oral examination: 17 December 2019

Development and applications of small-angle scattering-based structure modeling tools for proteins and nucleic acids

Referees: Dr. Thomas Schneider
Prof. Dr. Irmgard Sinning

Abstract

Small-angle scattering (SAS) of X-rays and neutrons allows the study of biological macromolecules in solution, at close to native conditions. The rapidly increasing popularity of the technique is attributed to both improvement in experimental facilities and continuous development of SAS data analysis and structure modeling tools. ATSAS, a software suite developed at the EMBL, is arguably the most comprehensive and utilized computer package for SAS data analysis worldwide. I present here the development of three computational tools, two of which have already been integrated into ATSAS: (1) SAS-guided normal mode analysis in torsion angle space (TNMA); (2) the use of sequence coevolution to reduce the ambiguity of SAS-based modeling; and (3) computation of anomalous scattering (ASAXS) effects in the context of SAXS data. Further, this PhD work contains the results of integrative structural biology projects in collaborations with user groups of the ESRF and EMBL Hamburg SAXS beamlines, where the newly developed methods were utilized.

In normal mode analysis, macromolecular motion is approximated as collective, low frequency harmonic oscillations around an initial, equilibrium structure. NMA in Cartesian space (CNMA) has been demonstrated to reasonably approximate conformational changes for a large set of proteins, and was thus used as the basis for SREFLEX, a method in the ATSAS suite to morph crystallographic structures to fit SAS data. However, it was shown in this work that SAS-guided CNMA results in stereochemically broken structures when applied to RNA. In comparison, SAS-guided TNMA of the same RNA structures resulted in improved models, in terms of both accuracy and stereochemistry. An implementation of SAS-guided TNMA, NMATOR, was thus developed and made available in the latest ATSAS v3.0.0 package. NMATOR was also used to generate SAXS-based solution structure models of Alu RNA, and the condensin HEAT-repeat protein Ycg1, and the ISC proteins HscA and IscU. The solution properties and structure of Ycg1, as determined through SAXS, have been published (Manalastas-Cantos et al, 2019).

SAS modeling ambiguity was also tackled in this work and two ways of ameliorating it through the generation of distance constraints were discussed: (1) experimentally, through anomalous scattering effects; and (2) bioinformatically, by evaluating sequence coevolution. A program to account for energy-dependent anomalous effects when computing SAXS data from macromolecular models was written and is available in ATSAS version 3.0.0, for planning and analyzing ASAXS experiments. In addition, sequence coevolution analysis and the integration of identified coevolving pair distance constraints into SAXS-guided modelling, was shown to improve heterodimer modeling accuracy. Sequence coevolution was utilized to generate distance constraints for HscB-IscU heterodimer modeling.

Zusammenfassung

Die Kleinwinkelstreuung (SAS) von Röntgenstrahlen und Neutronen ermöglicht die Untersuchung biologischer Makromoleküle in Lösung unter nahezu natürlichen Bedingungen. Die schnell zunehmende Popularität der Technik ist sowohl auf die Verbesserung der experimentellen Einrichtungen als auch auf die kontinuierliche Entwicklung von SAS-Datenanalyse- und Strukturmodellierungsmethoden zurückzuführen. ATSAS, eine am EMBL entwickelte Software-Suite, ist das wohl umfassendste und am meisten genutzte Computerpaket für die SAS-Datenanalyse weltweit. Ich präsentiere hier die Entwicklung von drei neuen Analyseprogrammen, von denen zwei bereits in ATSAS integriert wurden: (1) SAS-gesteuerte Normalmodenanalyse im Torsionswinkelraum (TNMA); (2) die Verwendung von Sequenzkoevolution, um die Mehrdeutigkeit der SAS-basierten Modellierung zu verringern; und (3) Berechnung von Anomalous Scattering (ASAXS) -Effekten in SAXS-Daten. Darüber hinaus enthält diese Doktorarbeit die Ergebnisse integrativer strukturbiologischer Projekte in Zusammenarbeit mit Anwendergruppen der SAXS-Beamlines des ESRF und des EMBL Hamburg, bei denen die neu entwickelten Methoden zum Einsatz kamen.

In der Normalmodenanalyse wird die makromolekulare Bewegung als kollektive, niederfrequente harmonische Schwingung um eine anfängliche Gleichgewichtsstruktur angenähert. Es wurde gezeigt, dass NMA im kartesischen Raum (CNMA) Konformationsänderungen für einen großen Satz von Proteinen annähernd annimmt, und es wurde daher als Grundlage für SREFLEX verwendet, eine Methode in der ATSAS-Suite, um kristallographische Strukturen an SAS-Daten anzupassen. In dieser Arbeit wurde jedoch gezeigt, dass SAS-gesteuertes CNMA bei Anwendung auf RNA zu stereochemisch gebrochenen Strukturen führt. Im Vergleich dazu führte SAS-gesteuertes TNMA mit denselben RNA-Strukturen zu verbesserten Modellen sowohl hinsichtlich der Genauigkeit als auch der Stereochemie. Daher wurde eine Implementierung von SAS-gesteuertem TNMA, NMATOR, entwickelt und im neuesten ATSAS v3.0.0-Paket verfügbar gemacht. NMATOR wurde auch verwendet, um SAXS-basierte Lösungsstrukturmodelle von Alu-RNA und dem Kondensin-HEAT-Repeat-Protein Ycg1 sowie den ISC-Proteinen HscA und IscU zu generieren. Die durch SAXS bestimmten Lösungseigenschaften und Strukturen von Ycg1 wurden veröffentlicht (Manalastas-Cantos et al., 2019).

In dieser Arbeit wurde auch die Zweideutigkeit der SAS-Modellierung behandelt, und es wurden zwei Möglichkeiten zur Verbesserung durch die Erzeugung von Abstandsbeschränkungen erörtert: (1) experimentell durch anomale Streueffekte; und (2) bioinformatisch durch Auswertung der Sequenzkoevolution. Ein Programm zur Berücksichtigung energieabhängiger anomaler Effekte bei der Berechnung von SAXS-Daten aus makromolekularen Modellen wurde geschrieben und ist in ATSAS Version 3.0.0 für die Planung und Analyse von ASAXS-Experimenten verfügbar. Darüber hinaus wurde gezeigt, dass eine Sequenzkoevolutionsanalyse und die Integration identifizierter Zwangsbedingungen für Koevolutionspaare in eine SAXS-geführte Modellierung die Genauigkeit der Heterodimer-Modellierung verbessern. Die Sequenzkoevolution wurde verwendet, um Abstandsbeschränkungen für die HscB-IscU-Heterodimer-Modellierung zu erzeugen.

Table of Contents

Abstract 3
Zusammenfassung 4
1 Introduction 7
1.1 Small-angle scattering fundamentals 7
1.2 Modeling biomolecular structure from SAS data 10
1.3 The scope of this work 13
2 Modeling flexible motions with normal mode analysis in torsion angle space 15
2.1 Normal mode analysis and SAS 15
2.2 NMATOR implementation 19
2.2.1 Calculating the torsional normal modes (TNMs) 19
2.2.2 SAS-guided torsional normal mode analysis 21
2.2.3 Generating a pool of conformers with TNMA 22
2.3 NMATOR benchmarking methodology 23
2.4 TNMA benchmarking results 26
2.5 Conclusion and outlook 31
3 Using sequence coevolution to reduce SAXS modeling ambiguity 33
3.1 Sequence coevolution and SAS 33
3.2 Quantifying the accuracy of coevolution-based contact predictions 35
3.3 Using DCA-predicted contacts in SAS-guided modeling 39
3.4 Conclusion and outlook 45
4 Computing anomalous scattering from structure 46
4.1 Anomalous SAXS: concept and biological applications 46
4.2 Adding anomalous scattering to CRYSOL 48
4.3 Using anomalous CRYSOL to compute scattering from biomolecules 51
4.4 Conclusion and outlook 52
5 Modeling the solution structure of *Plasmodium falciparum* Alu RNA and the Alu-SRP9/14 complex 54
5.1 The signal recognition particle (SRP) 54
5.2 Experimental procedures 56
5.2.1 SAXS data collection and processing 56
5.2.2 SAXS data analysis and structure modeling 57
5.3 Solution characteristics and models for *Pf* Alu RNA variants and *Pf* SRP9/14 58
5.4 A consensus model of the AluWT-SRP9/14 interaction 65

5.5 Conclusion and outlook 67

6 Solution structures of condensin HEAT-repeat proteins Ycg1 and Ycs4 from *Chaetomium thermophilum* 68

6.1 Experimental procedures 69

6.1.1 SAXS data collection and processing 69

6.1.2 Molecular weight assessment of Ycg1 oligomers with light scattering 71

6.1.3 SAXS data analysis and structure modeling 71

6.2 The effect of ligand binding on Ycg1 and Ycs4 structure 76

6.3 Characterizing Ycg1 oligomerization 80

6.4 Conclusion and outlook 83

7 Modeling iron-sulfur cluster biosynthesis proteins in *Escherichia coli*: HscA, HscB, IscU and the HscB-IscU complex 84

7.1 Experimental procedures 85

7.1.1 SAXS data collection and processing 85

7.1.2 SAXS data analysis and structure modeling 87

7.2 Solution characteristics and models for HscA, HscB and IscU 90

7.3 Detecting and modeling the HscB-IscU interaction 92

7.4 Conclusion and outlook 95

References 96

Appendix 105

List of Abbreviations 105

Large RNA dataset for TNMA benchmarking 106

Heterodimers used for DCA benchmark 110

Acknowledgements 118

1 Introduction

1.1 Small-angle scattering fundamentals

Small-angle scattering (SAS) is a method that is used to characterize macromolecules in solution, yielding low-resolution information about their structure and interactions. In a typical solution SAS experiment (Figure 1-1A), the macromolecules of interest are suspended in the appropriate buffer, drawn through a capillary, and exposed to a beam of X-rays (SAXS) or neutrons (SANS). The incident beam is scattered due to elastic collisions with electrons in the case of SAXS, or nuclei in the case of SANS, producing interfering waves that are then collected by a detector. Since the macromolecules are found in random orientations in solution, the detector collects isotropic data (Figure 1-1B) that can be radially-averaged into a one-dimensional (1D) profile of scattering intensity $I(s)$ over the range of momentum transfer $s = (4\pi \sin\theta)/\lambda$, where 2θ is the scattering angle, and λ is the wavelength of the incident radiation. The scattering from the buffer without the macromolecules is also measured, radially-averaged, and subtracted as background, yielding the scattering contribution from the macromolecules in the sample (Figure 1-1C) (Svergun *et al.*, 2013).

If the macromolecular solution is ideal and monodisperse (i.e. is pure and sufficiently dilute to prevent interparticle interactions), several parameters can be derived directly from the 1D SAS profile. At very low angles, two parameters provide information about the particles' molecular weight and size: the forward scattering and radius of gyration. The forward scattering is the scattering intensity at zero angle, $I(0)$. $I(0)$ corresponds to in-phase scattering from the whole macromolecule and is thus proportional to molecular weight. The radius of gyration (R_g) is derived from low-angle SAS data, and is related to the forward scattering according to Guinier's law (Guinier, 1939):

$$I(s) = I(0)e^{\frac{-s^2 R_g^2}{3}} \tag{1-1}$$

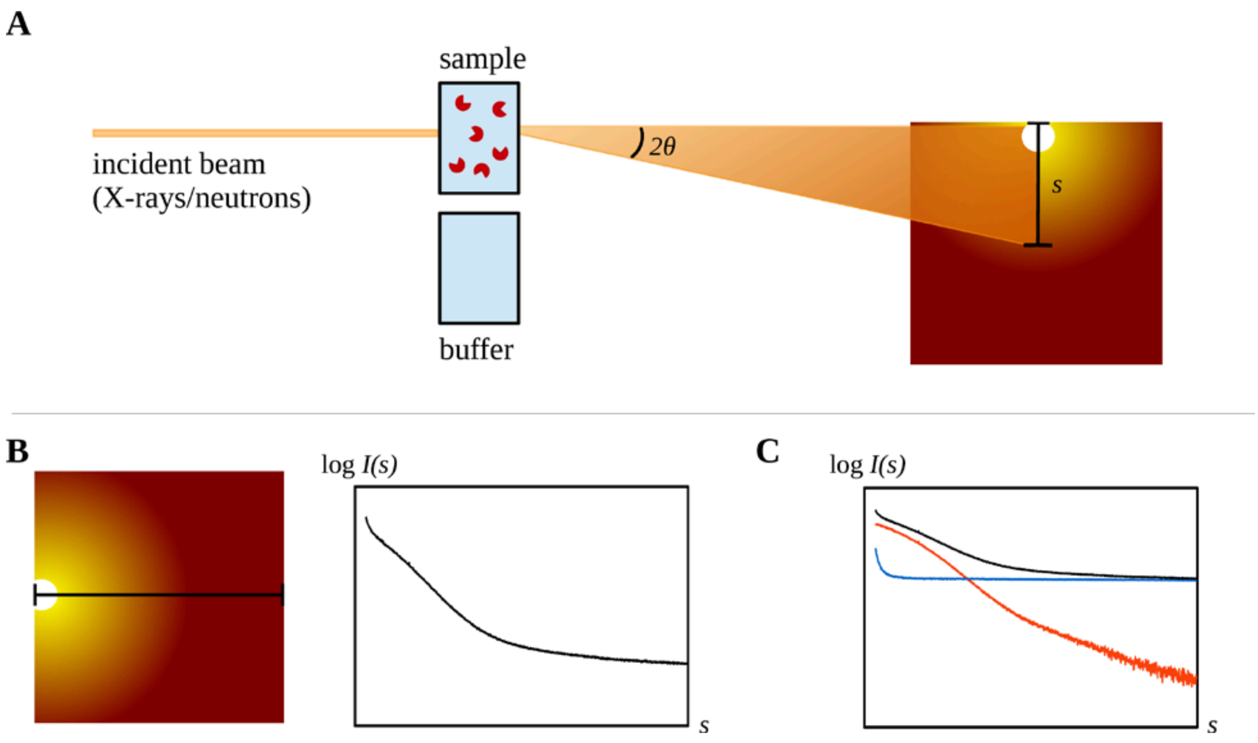

Figure 1-1. (A) Solution SAS experimental setup. (B) Two-dimensional (2D) image data collected by the detector is isotropic; this is radially-averaged to get the one-dimensional SAS profile. (C) Radially-averaged SAS data from the buffer (blue) is subtracted from the data from the macromolecular solution (black), yielding the scattering data from the macromolecule alone (red)

The forward scattering cannot be directly measured since it is in the path of the incident beam, most of which is not scattered during a SAS experiment. However, $I(0)$ can be approximated as the zero-intercept of the Guinier plot, $ln\,I(s)$ vs. s^2 (derived from eq. 1-1), which is linear at small angles ($sR_g < 1.3$), for monodisperse samples free of interparticle effects (aggregation or repulsion) (Figure 1-2A). The radius of gyration can be derived from the slope of the Guinier plot, and is sensitive to both the particle size and shape, i.e. the volume or mass distribution.

In addition to the scattering at low angles, the scattering over the whole angular range can be used to derive the $P(r)$ function (Figure 1-2B), a histogram of distances between all scattering pairs in the macromolecule. The scattering intensity $I(s)$ is the Fourier transform of the $P(r)$ function, as shown:

$$I(s) = 4\pi \int_0^{D_{max}} P(r)\frac{sin(sr)}{sr}dr \qquad (1\text{-}2)$$

And, thus, inversely:

$$P(r) = \frac{1}{2\pi^2}\int_0^\infty I(s)sr\,sin(sr)ds \tag{1-3}$$

Equation 1-3 shows that the calculation of $P(r)$ from $I(s)$ requires integration from zero to infinity, which is not practicable due to the limited angular range that is physically measurable, and the increased noise of experimental SAS data at higher angles. Instead, the $P(r)$ function is derived indirectly, by using the relationship in eq. 1-2 to parametrize $P(r)$, such that the $I(s)$ matches the experimental SAS data (Glatter, 1977; Semenyuk and Svergun, 1991).

An estimate of particle volume in solution can also be directly derived from the SAS data (called the Porod volume, V_p):

$$V_p = \frac{2\pi^2 I(0)}{\int_0^\infty I(s)s^2 ds} \tag{1-4}$$

The denominator of eq. 1-4 is known as the Porod invariant, so-called because it was shown to yield a constant independent of the nature of the scattering particle (Porod, 1951). Although the Porod invariant is another infinite integral, it is approximated by extrapolating to infinity using the Porod asymptotic, which shows that scattering intensity decays proportionally to s^{-4}, assuming a sharp interface between the scattering particle and the solvent.

$$\lim_{s\to\infty} I(s) = \frac{2\pi(\Delta\rho)^2 S_{int}}{s^4 V} \tag{1-5}$$

where $\Delta\rho$ is the excess scattering length density, S_{int} is the sum of internal scattering surfaces, and V is the illuminated volume of the sample (Porod, 1951; Debye, Anderson Jr and Brumberger, 1957). However, this asymptotic behavior is sensitive to the particle's folding state. This sensitivity can be used to qualitatively distinguish between a well-folded, globular particle, and an unfolded, flexible one.

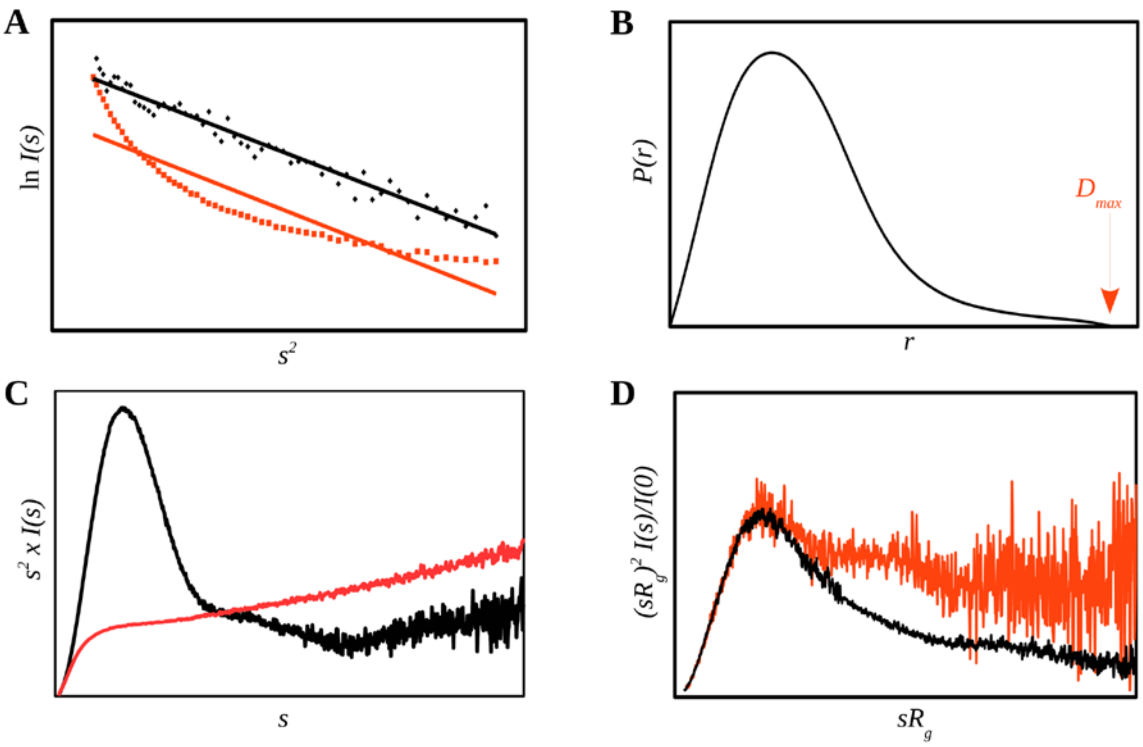

Figure 1-2. (A) Guinier plots for a monodisperse (black) and aggregated (red) sample. (B) The $P(r)$ function is a histogram of all interatomic distances. The largest distance (arrow), corresponds to the maximum dimension, D_{max}, of the particle. (C) The Kratky plot can be used to distinguish between globular (black) and flexible (red) particles. (D) The Kratky plot can be normalized by particle size to facilitate comparison.

Particle flexibility can be qualitatively observed by representing the scattering data as a Kratky plot, $s^2 \times I(s)$ vs. s (Figure 1-2C) (Kratky, 1982), which can be normalized by particle size ($(sR_g)^2 \times I(s)/I(0)$ vs. sR_g) (Figure 1-2D). Globular molecules have the expected intensity decay proportional to s^{-4}, resulting in a bell-shaped Kratky plot. On the other hand, unfolded particles show a slower intensity decay. With random chains, for example, scattering intensity decays proportional to s^{-2} (Debye, 1947). Thus, when viewed as a Kratky plot, flexible molecules show a plateauing signal, instead of a well-defined maximum.

1.2 Modeling biomolecular structure from SAS data

The usefulness of SAS data for structural biology has increased over the past five decades, as structure modeling tools for SAS data has improved with increasing computational resources. These modeling tools range from *ab initio* modeling methods that are conceptually based on either envelope reconstruction (Stuhrmann, 1970; Svergun and

Stuhrmann, 1991), or finite element modeling (Chacon *et al.*, 1998; Svergun, 1999), to those that leverage high-resolution structures as building blocks for modeling (Petoukhov and Svergun, 2005; Schneidman-Duhovny, Hammel and Sali, 2011; Franke *et al.*, 2012; Panjkovich and Svergun, 2016). The specific software tools discussed in this manuscript will primarily come from the ATSAS software suite, which is a comprehensive collection of computer programs for SAS data processing, analysis, and modeling (Franke *et al.*, 2017), though the underlying concepts should hold for other software implementations.

Ab initio modeling tools can be used without any prior information about the structure of a biomolecule. Due to the limited ability of envelope analysis to reconstruct complex shapes, such as those with large concavities, current *ab initio* modeling methods are mostly based on finite element (bead) modeling. In bead modeling, the structure is modeled as an arrangement of beads of similar scattering density to the object being modeled. Methods such as DAMMIN and DAMMIF (as well as multiphase modeling method MONSA) incorporate restrictions such as continuity and compactness to improve modeling quality (Svergun, 1999; Franke and Svergun, 2009). Continuity refers to the interconnectivity of the beads, while compactness means that the beads must be arranged in a way that reflects the compactness of typical biomolecules. These conditions must be specified because *ab initio* modeling from SAS data can often yield many different, yet equally likely models. This is known as the modeling ambiguity problem, and is an intrinsic limitation of the method and a direct consequence of the information lost through the isotropic tumbling of macromolecules that occurs in solution, that results in the time and orientational averaging of scattering amplitudes. Nonetheless, *ab initio* models are often informative, and as long as the ambiguity is adequately characterized, can give structural insights in the absence of high-resolution information. Modeling ambiguity can be quantified *a priori* by examining the SAS data itself for inherent ambiguity (Petoukhov and Svergun, 2015), or by examining the variance of the models from multiple modeling instances (Volkov and Svergun, 2003; Tuukkanen, Kleywegt and Svergun, 2016).

Analogous to the use of heavy atoms for macromolecular phasing in X-ray crystallography, the anomalous scattering of heavy atoms can be used as a molecular ruler in SAXS experiments. In particular, the distances between the heavy atoms can be derived by performing scattering experiments near and at the absorption edge of the atoms, and evaluating the resulting decrease in scattering signal. Although at present, only a few biological studies using anomalous SAXS have been published (Stuhrmann and Notbohm, 1981; Miake-Lye, Doniach and Hodgson, 1983; Pabit *et al.*, 2010), anomalous scattering could conceptually be used as a source of distance constraints to reduce SAXS modeling ambiguity.

Using high-resolution structures as building blocks for modeling (i.e. hybrid modeling) can also somewhat ameliorate, though not completely remove, modeling ambiguity. Cases in which high-resolution structures are used in conjunction with SAS data include (1) validating if the high-resolution structure corresponds to the solution structure, and (2) building the solution structure of the full-length protein, oligomer, or complex, when only partial structures are known.

For the first case, gross structural changes can be seen by SAS. Tools such as CRYSOL can give a measure of the agreement between a high-resolution structure and experimental SAS data (Svergun, Barberato and Koch, 1995). This is achieved by computing the theoretical scattering of the high-resolution model, and comparing the model scattering with the experimental SAS data, through the χ^2 metric:

$$\chi^2 = \frac{1}{N_p}\sum_{i=1}^{N_p}\left[\frac{I_e(s_i)-cI_m(s_i)}{\sigma(s_i)}\right]^2 \tag{1-6}$$

where N_p is the number of experimental points, I_e is the experimental scattering, I_m is the computed scattering from the model, $\sigma(s_i)$ are the experimental errors, and c is the scaling factor. A χ^2 fit of around 1 is considered a good fit, given accurate error estimates. Specifically, if the estimated errors are too large, any differences between two scattering profiles are attributed to error, resulting in a spuriously low χ^2. Conversely, artificially poor fits (high χ^2) could result from underestimated errors.

In the case of poor fit between a high-resolution model and experimental SAS data, conformational changes can be modeled as domain movements simulated by normal mode analysis. SREFLEX is a method in the ATSAS suite that morphs an initial high-resolution structure along its normal modes in Cartesian space, such that it corresponds well with a given scattering profile (Panjkovich and Svergun, 2016).

For the case wherein a full-length protein, oligomer, or complex is built from partial structures and SAS data, rigid-body modeling is often employed. In rigid body modeling, the partial structures are treated as immutable blocks which are arranged in 3D space to optimally fit the experimental SAS data, while also meeting geometric criteria such as structure connectedness and lack of clashes. In the ATSAS suite alone, there are several SAS-guided rigid-body modeling methods that are each suitable to different modeling scenarios: methods that, additionally, reconstruct any missing residues (BUNCH for single proteins, CORAL for complexes), a method that models oligomers and complexes based on the subunit structures (SASREF), and even one that models partially-dissociating oligomers and complexes as a mixture of the oligomer/complex and the constituent subunits (SASREFMX) (Petoukhov and Svergun, 2005; Franke *et al.*, 2012; Petoukhov *et al.*, 2013). As with *ab initio* modeling, SAS-based hybrid modeling could be ambiguous, and as such, benefits from the characterization of this ambiguity, which involves performing multiple modeling runs, and examining the variance of the resulting solutions.

1.3 The scope of this work

The main interests tackled in this work are modeling biomolecular flexibility, and reducing modeling ambiguity. Currently, normal mode analysis in Cartesian space (CNMA) has been shown to reasonably model interdomain motions for a large set of proteins (Krebs *et al.*, 2002), and to combine well with SAS data in order to model protein flexibility in solution (Panjkovich and Svergun, 2016). However, the performance of SAS-guided CNMA for nucleic acid structures had not been extensively tested. In Chapter 2, I discuss the development of an NMA tool in torsion-angle space (TNMA), and show that it is better suited to modeling RNA than CNMA, both in terms of accuracy and stereochemistry.

SAS modeling ambiguity can often be reduced by providing complementary information, such as contacts or solvent exposure data. In the absence of additional experimental data, sequence coevolution was tested here as a way to specify contacts between subunits in a heterodimer. In Chapter 3, I discuss the development, applications, and limitations of a contact-prediction method based on sequence coevolution.

Modeling ambiguity can also be addressed by experimental methods such as ASAXS. In Chapter 4, I discuss the development of a software module to account for anomalous effects when computing scattering from a biomolecule, and the potential of this module to guide ASAXS experiments.

Lastly, in Chapters 5-7, I present several experimental user projects, in which I was involved during the PhD: (1) Alu RNA, (2) condensin HEAT-repeat proteins, and (3) HscA and HscB-IscU bacterial proteins. The experimental SAXS data in these projects were analyzed with ATSAS programs and specifically, with the new methods developed in this work.

2 Modeling flexible motions with normal mode analysis in torsion angle space

2.1 Normal mode analysis and SAS

The simulation of macromolecular dynamics in biologically-relevant timescales is important for understanding macromolecular function. Normal mode analysis approximates macromolecular motion as collective harmonic motions of the component atoms around an initial, equilibrium position (Goldstein, 1950), and as such, is a less computationally-intensive method of simulating protein dynamics than all-atom molecular dynamics (MD). The computational cost of NMA could be further decreased by employing coarse-graining (i.e. representing the structure with a limited set of representative atoms, e.g. the Cα atoms in proteins). Coarse-grained NMA in Cartesian space (CNMA) has been shown to reproduce conformational changes in proteins (Tama and Sanejouand, 2001; Krebs *et al.*, 2002; Alexandrov *et al.*, 2005; Tobi and Bahar, 2005). As a result, one application of coarse-grained CNMA has been to morph high-resolution structures to fit electron density maps from cryo-electron microscopy (Tama, Miyashita and Brooks III, 2004), and more relevant to this work, solution scattering data (Panjkovich and Svergun, 2016). Below, I discuss the mathematical formalism of NMA in a manner agnostic of the coordinate system used, in order to compare between the widely-employed CNMA, and the approach used in this work, NMA in torsion angle space (TNMA).

If the initial structure is taken to be the equilibrium position (represented as a set of N coordinates, $\boldsymbol{q^0}$), the potential energy is assumed to be a quadratic function around this minimum.

$$E_p = \frac{1}{2}\sum_{i,j}^{N} H_{i,j}\left(q_i - q_i^0\right)\left(q_j - q_j^0\right) = \frac{1}{2}(\boldsymbol{q} - \boldsymbol{q^0})\boldsymbol{H}(\boldsymbol{q} - \boldsymbol{q^0}) \tag{2-1}$$

The kinetic energy of the system can be similarly expressed as a quadratic function of the velocities:

$$E_k = \frac{1}{2}\sum_{i,j}^{N} T_{i,j}\,\dot{q}_i\dot{q}_j = \frac{1}{2}\dot{\boldsymbol{q}}^{\boldsymbol{T}}\boldsymbol{T}\dot{\boldsymbol{q}} \tag{2-2}$$

The potential and kinetic energy functions can then be used to solve Lagrange's equations of motion, which are generalizable to any coordinate system:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) = \left(\frac{\partial L}{\partial q_i}\right) \text{ where } L = E_k - E_p \tag{2-3}$$

From eqs. 2-1 and 2-2, we get $\partial L/\partial \dot{q}_i = \sum_j^N T_j\,\dot{q}_j$ and $\partial L/\partial q_i = -\sum_j^N H_j\left(q_j - q_j^0\right)$, which can be substituted into eq. 2-3, as follows:

$$\sum_j^N T_j\,\ddot{q}_j = -\sum_j^N H_j\left(q_j - q_j^0\right) \tag{2-4}$$

Given that each set of coordinates q_j is a function of time, $q_j = q_j^0 + \sum_k^N A_{jk} \propto_k cos(\omega_k t + \delta_k)$, then $\ddot{q}_j = -\sum_k^N A_{jk} \propto_k \omega_k^2 cos(\omega_k t + \delta_k)$. Substituting $q_j - q_j^0$ and $\ddot{q}_j$ into eq. 2-4:

$$-\sum_j^N T_j\left(\sum_k^N A_{jk} \propto_k \omega_k^2 cos(\omega_k t + \delta_k)\right) = -\sum_j^N H_j\left(\sum_k^N A_{jk} \propto_k cos(\omega_k t + \delta_k)\right) \tag{2-5}$$

Which for all values of t simplifies to:

$$\sum_j^N T_j\,A_{jk}\omega_k^2 = \sum_j^N H_j\,A_{jk} \tag{2-6}$$

Which, in matrix notation, can be written as a generalized eigenvalue problem:

$$\boldsymbol{TA\Lambda} = \boldsymbol{HA} \tag{2-7}$$

Where the matrix of eigenvectors $\boldsymbol{A}$ contains the normal modes, and $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues, which are the vibration frequencies associated with each mode. The normal modes are orthogonal, and as such, macromolecular motion is typically approximated by a linear combination of the low frequency modes.

If $\boldsymbol{q}$ is composed of the Cartesian coordinates of identical representative atoms (e.g. coarse-graining using C$_\alpha$ atoms only), the kinetic energy matrix $\boldsymbol{T}$ becomes a diagonal matrix of identical masses, and eq. 2-7 can be reduced to $\boldsymbol{A\Lambda} = \boldsymbol{HA}$, where $\boldsymbol{H}$ is the mass-weighted potential energy Hessian matrix. This simplified form has been used extensively over the years, and has been shown in many instances to reasonably reproduce protein flexibility-related metrics such as crystallographic temperature factors (Atilgan *et al.*, 2001), domain architectures (Hinsen, Thomas and Field, 1999), and conformational changes (Tama and Sanejouand, 2001; Krebs *et al.*, 2002; Alexandrov *et al.*, 2005; Tobi and Bahar, 2005).

The concept of using other non-Cartesian coordinate systems, however, is not new. In fact, several early works using NMA to simulate macromolecular motion used dihedral/torsion angles (Figure 2-1) instead of Cartesian coordinates (Go, Noguti and Nishikawa, 1983; Levitt, Sander and Stern, 1985), because this more accurately represents the degrees of freedom of the macromolecule (i.e. bond rotations occur to a greater extent than bond stretching). In addition, using torsion angles decreases the variables compared to an all-atom representation, consequently reducing the sizes of the potential and kinetic energy matrices. For example, a protein with N amino acids would have at most $2N-2$ backbone torsion angles (the φ and ψ angles, excluding the terminal ones; Figure 2-1A), and at least $4N$ heavy (not hydrogen) atoms, which is at minimum, a two-fold decrease in the number of variables. CNMA-based approaches have gotten around this by employing coarse-graining: for example, by representing each amino acid residue by only the C_α atoms, as in the examples cited above, or by grouping residues into rigid blocks (rotations-translations of blocks, RTB) (Tama *et al.*, 2000).

As mentioned previously, CNMA has been used to refine high-resolution protein structures against SAS data (Gorba and Tama, 2010; Miyashita, Gorba and Tama, 2011; Panjkovich and Svergun, 2016). In particular, it was shown that given a set of proteins with two known conformations, in most of the cases, CNMA was able to morph the structure from one conformation to the other, guided only by the SAS data and the innate domain organization of the protein structure. RTB coarse-graining, specifically the automatic detection of protein domains based on topology, was key to accurately reconstituting the target structure from SAS data, and is currently implemented in the ATSAS program SREFLEX (Panjkovich and Svergun, 2016). However, SAS-guided CNMA was developed and extensively tested only on protein structures; the method’s performance on nucleic acid structures was yet unknown.

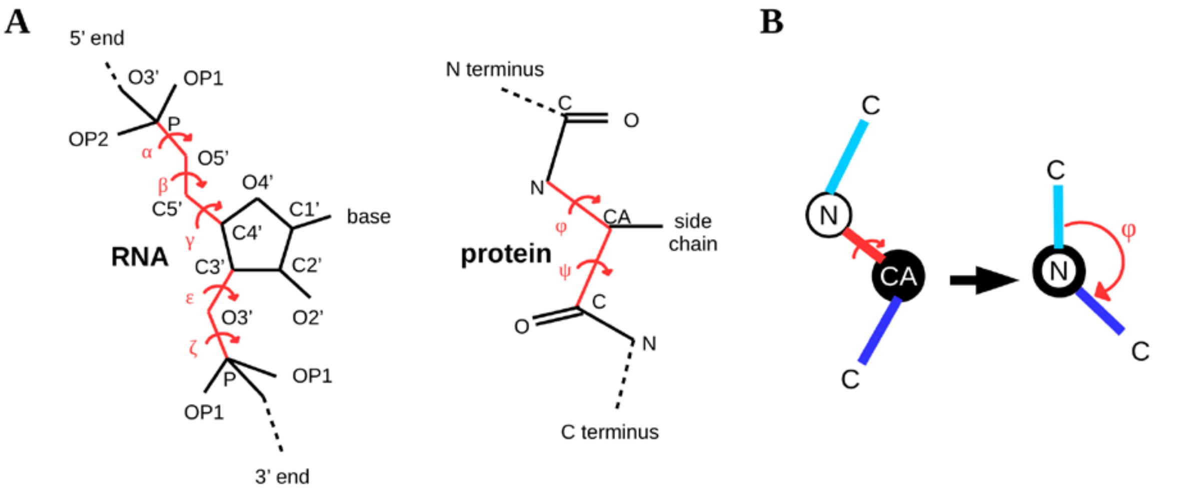

Figure 2-1. (A) Backbone structures of RNA and proteins. Highlighted in red are the backbone bonds considered rotatable by NMATOR, and the torsion angles which they define are specified as Greek letters. The C3'-C4' bond in RNA is assumed rigid, disallowing changes in the ribose moiety. (B) A torsion angle φ as viewed from the upstream N atom.

In this work, we benchmarked and compared two SAS-guided NMA methods on a set of RNA structures, one based on RTB-CNMA (SREFLEX), and the other based on TNMA (NMATOR, discussed below). We show that in most cases, TNMA produced RNA models of greater accuracy and better stereochemistry than RTB-CNMA, when applied to the problem of SAS-guided structure refinement of RNA structures. That the stereochemistry of the resulting models would be better was somewhat an expected result, as CNMA could sometimes result in non-physical motions, such as bonds being stretched beyond what is physically possible. Since bond lengths are kept fixed in TNMA (only bond rotations occur, Figure 2-1), excessive bond stretching does not occur with TNMA. Several works have also shown that that the circular motions from TNMA better approximate structural transitions between two conformations, for both RNA and proteins (Mendez and Bastolla, 2010; Bray, Weiss and Levitt, 2011; Lopéz-Blanco, Garzón and Chacón, 2011).

Currently, there is no published NMA-based tool for modeling RNA structures against SAS data. Instead, the high-resolution structure—either known experimentally, or predicted from sequence with tools such as MC-SYM and FARNA (Das and Baker, 2007; Parisien and Major, 2008)—is used as a starting point for molecular dynamics (MD) simulations, and the resulting pool of conformations are fitted against the scattering data (Chen and Pollack, 2016; Cantara, Olson and Musier-Forsyth, 2017). However, NMA has two main advantages over MD, which are the speed of computation, and ease of use. Both of these factors make NMA accessible to a wider array of users. The minimal computational requirements of NMA means that it can easily be run on a wider range of computers. Also, NMA-based tools like NMATOR usually require less parameter optimization from the user than MD, which requires some basal level of expertise.

2.2 NMATOR implementation

A software implementation of SAS-guided TNMA, called NMATOR (NMA in torsion angle space), was developed and incorporated into the ATSAS software suite. NMATOR can be divided into two main parts: (1) the calculation of normal modes in torsion angle space (TNMs), and (2) morphing an initial high-resolution structure along its TNMs to fit a given SAS profile. A third feature, the generation of a pool of conformations given an initial structure, is also available and under development, and is intended to be an alternative to MD, as a less-computationally expensive way of doing ensemble modeling.

2.2.1 Calculating the torsional normal modes (TNMs)

In order to compute the TNMs, one must solve the generalized eigenvalue problem in eq. 2-7, which requires the computation of the potential and kinetic energy Hessian matrices. Here we approximate the potential energy as an elastic potential around the initial structure (i.e. atoms within a distance threshold are interacting with a non-atom-specific harmonic potential) (Tirion, 1996). Given N rotatable bonds, each term of the $N \times N$ potential energy Hessian matrix H is obtained as follows:

$$H_{\alpha,\beta} = H_{\beta,\alpha} = \boldsymbol{\chi}_{\boldsymbol{\alpha}}^{\boldsymbol{T}} \boldsymbol{R}^{\boldsymbol{\alpha},\boldsymbol{\beta}} \boldsymbol{\chi}_{\boldsymbol{\beta}} \tag{2-8}$$

Where $\alpha, \beta \leq N$, $\boldsymbol{\chi}_{\boldsymbol{\alpha}} = \begin{bmatrix} \boldsymbol{e}_{\boldsymbol{\alpha}} \\ \boldsymbol{e}_{\boldsymbol{\alpha}} \times \boldsymbol{r}_{\boldsymbol{\kappa}(\propto)} \end{bmatrix}$, $\boldsymbol{e}_{\propto}$ is the unit vector along the rotatable bond α, $\kappa(\propto)$ is the ordinal number of the root atom of α, and $\boldsymbol{r}_{\boldsymbol{\kappa}(\boldsymbol{\alpha})}$ defines the Cartesian coordinates of the root atom (Figure 2-2A). For each pair of rotatable bonds α and β, $\boldsymbol{R}^{\boldsymbol{\alpha},\boldsymbol{\beta}}$ is the Hookean potential between the atoms that are moved by the rotation of each with respect to the other:

$$\boldsymbol{R}^{\boldsymbol{\alpha},\boldsymbol{\beta}} = \sum_{\substack{i \leq \kappa(\alpha) \\ j \geq \kappa(\beta)}} \frac{\gamma_{ij}\delta_{ij}}{r_{ij}^2} \begin{bmatrix} \boldsymbol{r}_{\boldsymbol{i}} \times \boldsymbol{r}_{\boldsymbol{j}} \\ \boldsymbol{r}_{\boldsymbol{i}} - \boldsymbol{r}_{\boldsymbol{j}} \end{bmatrix} \begin{bmatrix} \boldsymbol{r}_{\boldsymbol{i}} \times \boldsymbol{r}_{\boldsymbol{j}} & \boldsymbol{r}_{\boldsymbol{i}} - \boldsymbol{r}_{\boldsymbol{j}} \end{bmatrix}, \; \delta_{ij} = \begin{cases} 0 & if \; r_{ij} > r_{int} \\ 1 & if \; r_{ij} \leq r_{int} \end{cases} \tag{2-9}$$

where atom i is upstream of bond α, and j is downstream of bond β (Figure 2-2B). The Kronecker delta function δ_{ij} restricts the potential to atom pairs with distance $r_{ij} \leq r_{int}$ (here $r_{int} = 10\text{Å}$). The spring constant γ_{ij} used here is the following sigmoid function (Lopéz-Blanco, Garzón and Chacón, 2011):

$$\gamma_{ij} = \frac{1}{1 + \left(\frac{r_{ij}}{3.8}\right)^6}$$

In order to eliminate redundant computations, the computations are done in the following order: $\boldsymbol{R}^{\mathbf{1},\boldsymbol{N}}$, $\boldsymbol{R}^{\mathbf{1},\boldsymbol{N}-\mathbf{1}}$,..., $\boldsymbol{R}^{\mathbf{1},\mathbf{1}}$, $\boldsymbol{R}^{\mathbf{2},\boldsymbol{N}}$, $\boldsymbol{R}^{\mathbf{2},\boldsymbol{N}-\mathbf{1}}$,..., $\boldsymbol{R}^{\mathbf{2},\mathbf{2}}$,..., $\boldsymbol{R}^{\boldsymbol{N},\boldsymbol{N}}$, with consequent $\mathbf{R}^{\boldsymbol{\alpha},\boldsymbol{\beta}}$ using computed values in earlier steps (Abe *et al.*, 1984). A weighting factor, $3min(H_{\propto,\propto})$, is added to the diagonal terms of the resulting Hessian matrix in order to "weigh down" the ends of the structure; otherwise, motions from these floppy ends would dominate the lowest frequency modes (Lu, Poon and Ma, 2006).

Similarly, each term of the $N \times N$ kinetic energy Hessian matrix T is calculated as follows:

$$T_{\alpha,\beta} = T_{\beta,\alpha} = \boldsymbol{\chi}_{\boldsymbol{\alpha}}^{\boldsymbol{T}} \boldsymbol{K}^{\boldsymbol{\alpha},\boldsymbol{\beta}} \boldsymbol{\chi}_{\boldsymbol{\beta}} \qquad (2\text{-}10)$$

$$\boldsymbol{K}^{\boldsymbol{\alpha},\boldsymbol{\beta}} = \frac{1}{M} \begin{pmatrix} \boldsymbol{P}_{\propto}^{\boldsymbol{T}} \\ M_{\propto}\mathbf{1} \end{pmatrix} \begin{pmatrix} \boldsymbol{P}_{\boldsymbol{\beta}} & M_{\beta}\mathbf{1} \end{pmatrix} + \begin{pmatrix} \boldsymbol{I}_{\boldsymbol{\alpha}} \\ \boldsymbol{P}_{\boldsymbol{\alpha}} \end{pmatrix} \boldsymbol{I}^{-\mathbf{1}} \begin{pmatrix} \boldsymbol{I}_{\boldsymbol{\beta}} & \boldsymbol{P}_{\boldsymbol{\beta}}^{\boldsymbol{T}} \end{pmatrix} \qquad (2\text{-}11)$$

Where M = total mass of the molecule, $M_{\propto} = \sum_{i \leq \kappa(\alpha)} m_i$,

$$\mathbf{1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},\ \boldsymbol{P}_{\boldsymbol{i}} = \begin{bmatrix} 0 & -z_i & y_i \\ z_i & 0 & -x_i \\ -y_i & x_i & 0 \end{bmatrix},\ \boldsymbol{P}_{\boldsymbol{\alpha}} = \sum_{i \leq \kappa(\alpha)} m_i \boldsymbol{P}_{\boldsymbol{i}},\ \text{and}\ \boldsymbol{I}_{\boldsymbol{\alpha}} = \sum_{i \leq \kappa(\alpha)} m_i\, \boldsymbol{P}_{\boldsymbol{i}}^{\boldsymbol{T}} \boldsymbol{P}_{\boldsymbol{i}}$$

With the matrices H and T computed, the generalized eigenvalue problem in eq. 2-7 is solved with the LAPACK subroutine DGGEV (Anderson *et al.*, 1999).

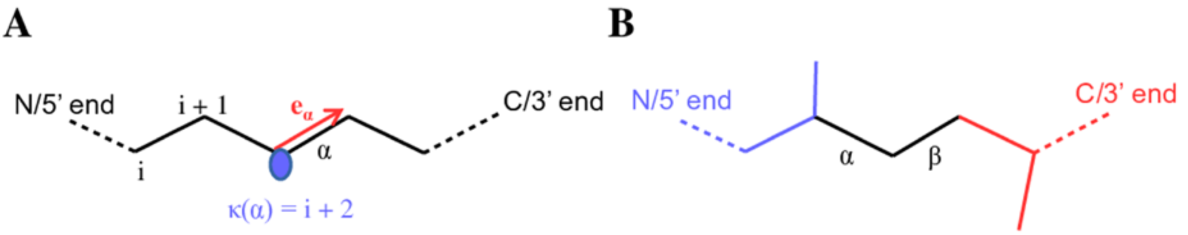

Figure 2-2. (A) The rotatable bond α can be defined by the unit vector e_α pointing in the downstream direction along the bond (to the C terminus for proteins, and the 3' end for nucleic acids). $\kappa(\alpha)$ refers to the ordinal atom number along the chain when going from the N/5' to the C/3' end. (B) Atoms upstream from bond α are shown in blue, while atoms downstream of bond β and shown in red.

2.2.2 SAS-guided torsional normal mode analysis

SAS-guided TNMA was implemented as a greedy algorithm, shown schematically in Figure 2-3. The initial structure is morphed along the first ten normal modes in both positive and negative directions (negative means the sign of the normal mode is flipped), and the χ^2 fit of each of the twenty resulting models against the SAS data is computed. The best-fitting model is selected as the new initial structure. These steps are repeated until one of the three following conditions occurs: (1) a model with fit to SAS data $\chi^2 \leq 1.1$ is obtained; (2) After 30 iterations, a model with $\chi^2 < 2$ was obtained; or (3) after 50 iterations, the χ^2 fits of the resulting models no longer improve.

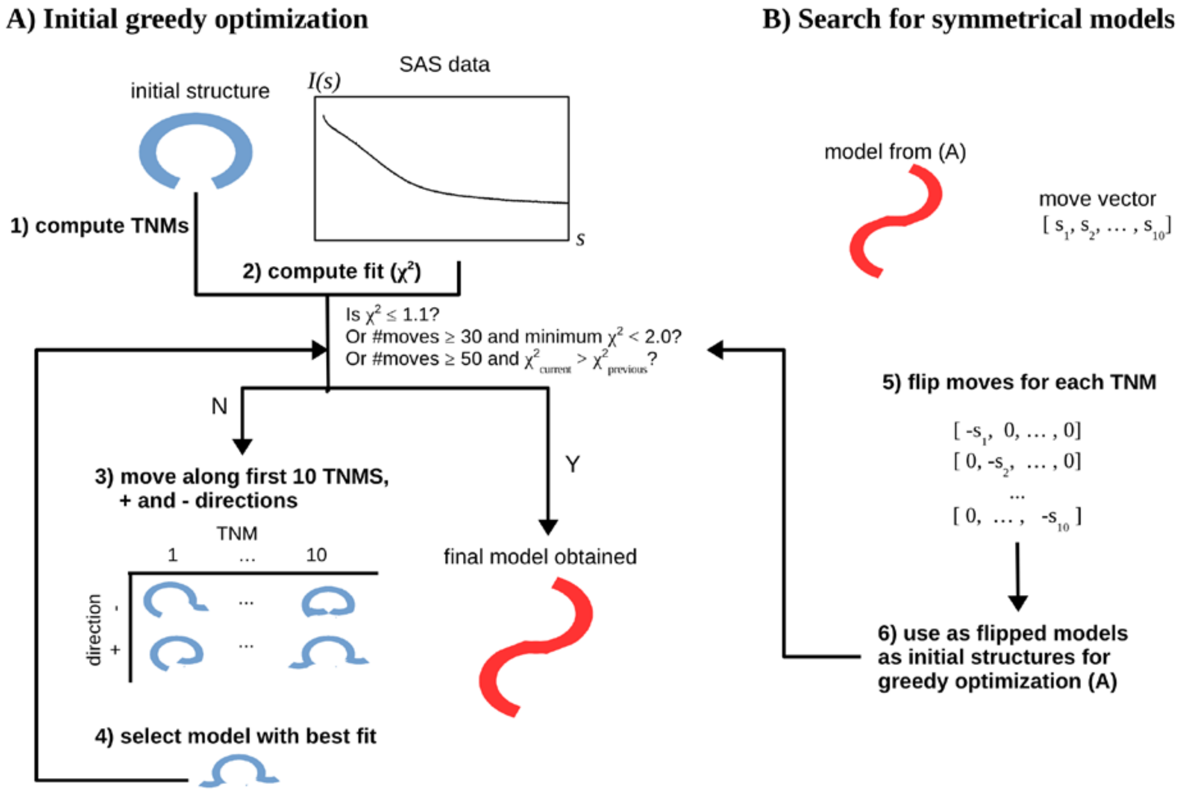

Figure 2-3. SAS-guided structure modeling using torsional normal modes, as implemented in NMATOR

The best model obtained after the initial greedy optimization was used as the seed structure to obtain other possibly symmetric models that fit the SAS data, but with the moves on each of the normal modes flipped. For example, if the moves from the initial structure to the best model can be described by the 10-vector $M_{greedy} = [s_1, \dots, s_{10}]$, where s_i is the number of steps taken along TNM_i, and a negative sign indicates that the TNM was taken in the opposite (negative) direction, then greedy optimization was performed at most more ten times, using as starting structures models of the form $M_{flip,i} = [0, \dots, -M_{greedy}(i), \dots, 0]$, where $i = [1,10]$ and $M_{greedy}(i) \neq 0$.

2.2.3 Generating a pool of conformers with TNMA

The TNMs computed by NMATOR could also be used to generate a pool of conformations, given an initial, high-resolution structure. The pool is generated by taking all possible combinations of the first five normal modes, in both positive and negative directions, for a maximum of five moves. Expressed mathematically, each conformation i in the pool differs from the initial conformation by the move vector $M_i = [s_1, \dots, s_5]$, where $\sum_{j=1}^{5} |s_j| \leq 5$.

2.3 NMATOR benchmarking methodology

As an initial, proof-of-principle run, a small dataset of nonredundant RNA sequences in two different conformations—open and closed, as defined by the radius of gyration (R_g)—was assembled. Five of the RNA structure pairs were taken from a previous study which modeled structural transitions in RNA (Lopéz-Blanco, Garzón and Chacón, 2011). The remaining four RNA structure pairs were obtained by getting a representative sequence from each RNA category as defined in the Nucleic Acid Database (NDB) (Berman *et al.*, 1992), and screening the PDB for at least two structures that share the same sequence, but have an all-atom rmsd of 3 Å or more.

SAXS data were simulated for the structures in the dataset using CRYSOL (Svergun, Barberato and Koch, 1995). Angle-dependent random errors were added to the simulated intensities based on the variance of 1000 independently-measured 1s scattering frames of water, as previously described (Franke, Jeffries and Svergun, 2015). Structures were further screened based on whether or not the open and closed states can be distinguished using the simulated SAXS data: in particular, if $\chi^2 \leq 2$ between the open structure and the simulated data from the corresponding closed structure (or vice versa), the structure pair was excluded. The remaining structure pairs are shown in Table 2-1.

A larger dataset of RNA structure pairs was also collated, that were of redundant sequence but nonredundant structure. The dataset was acquired from the PDB by searching for structures solved by solution NMR consisting of more than one model. Similar to the initial benchmark dataset, SAXS data was simulated for each of the models. Only pairs of models that had at least 10 Å all-atom rmsd between them, and a $\chi^2 > 2$ between one structure in the pair and its partner's simulated SAXS data were kept in the dataset. Additionally, redundant structure pairs were removed by clustering the NMR structures, such that structures with all-atom rmsd < 4 Å were considered the same structure. This left a total of 138 distinct RNA structure pairs: 66 open-to-closed, and 68 closed-to-open cases (Table 2-2; for the full list, see Appendix).

Table 2-1. Small RNA dataset: nonredundant sequences

OPEN		CLOSED		number of bases	rmsd (Å)	Name
PDB ID	R_g	PDB ID	R_g			
3cul_D	24.9	3cun_C	23.8	91	3.1	tRNA aminoacylase (synthetic ribozyme)
1u63_D	19.2	2vpl_B	19.0	48	4.4	Fragment of mRNA for L1
1z2j_13_A	20.1	2l94_9_A	18.2	44	5.4	HIV-1 frameshift inducing element
3knj_W	23.6	1gts_B	22.9	74	5.6	tRNA-Gln of *E. coli*
3fih_Y	23.5	1pns_W	23.0	75	5.7	A/T-site tRNA Phe (synthetic)
1uui_B	15.1	2l8h_5_A	12.2	28	6.6	HIV TAR RNA
2i7z_17_B	20.9	2jyf_1_A	18.7	42	7.2	GAAA tetraloop receptor RNA (synthetic)
3r9w_B	27.3	3r9x_C	21.6	34	16.3	16S rRNA, nt 1506-1542
1u6p_16_B	35.1	1s9s_3_A	32.7	101	21.4	MLV Psi encapsidation site

Table 2-2. Large RNA dataset: redundant sequences, nonredundant structures

PDB ID	number of bases	Name	number of open-to-closed pairs	number of closed-to-open pairs
1anr	29	HIV-1 TAR (cis-acting RNA regulatory element)	8	8
1ikd	22	tRNA-Ala acceptor stem of *E.coli*	1	1
1m5l	38	Modified HIV-1 packaging signal stem-loop 1	4	4
1s9s	101	MLV Psi encapsidation site	34	34
2m58	59	2’-5’ lariat forming ribozyme (synthetic)	9	9
2mtj	47	III-IV-V 3-way junction of the Varkud Satellite ribozyme	0	2
2n3q	62	II-III-VI 3-way junction of the Varkud Satellite ribozyme	4	4
2pcv	35	U65 Box H/ACA snoRNA	5	5
6hag	43	SAM riboswitch	1	1

Torsional NMA was performed with NMATOR on each of the structure pairs, in both open-to-closed, and closed-to-open directions. Open-to-closed here means that the initial structure is the open form which is iteratively morphed along the TNMs to fit the simulated SAXS data from the closed form, which we refer to as the target structure (and vice versa for the closed-to-open case). The Cartesian NMA method SREFLEX was run on the same cases as a comparison.

For both small and large benchmarking runs, the following metrics were computed for each resulting model: χ^2 fit to the simulated SAXS data, rmsd from the target structure, and a stereochemistry score. The all-atom rmsd of models from the target structures ($rmsd_{model}$) were computed using SUPPDB, a program in the ATSAS suite which superimposes two structures optimally with the Kabsch algorithm (Kabsch, 1976). From $rmsd_{model}$, the normalized change in rmsd ($\Delta rmsd_{norm}$) was derived, which is defined as follows:

$$\Delta rmsd_{norm} = \frac{rmsd_{model} - rmsd_{initial}}{rmsd_{initial}} \quad (2\text{-}12)$$

where $rmsd_{initial}$ is the rmsd between the initial and target structures. A negative value for $\Delta rmsd_{norm}$ indicates that the model is closer to the target structure than the initial structure, with $\Delta rmsd_{norm} = -1$ being the ideal case ($rmsd_{model} = 0$). Conversely, a positive $\Delta rmsd_{norm}$ indicates that the model is farther away from the target structure than the initial structure.

Stereochemical integrity was quantified by comparing how much bonds in the model are stretched in comparison to the initial structure. As such, the "breaks score" is defined as follows:

$$\frac{1}{N}\sum_{i=1}^{N}\left(d_i^{model} - d_i^{initial}\right)^2 \quad (2\text{-}13)$$

where d_i is the length of the ith backbone bond, and N is the number of backbone bonds in the structure.

2.4 TNMA benchmarking results

The overall performance of TNMA-based tool NMATOR, and CNMA-based SREFLEX for the small RNA dataset is shown in Table 2-3. For both open-to-closed and closed-to-open categories, TNMA was able to find a model with improved rmsd over the initial structure for more cases than CNMA. CNMA was not able to find models with good fit to the scattering data (we set "good" here as $\chi^2 \leq 2$) for the four cases with the highest rmsd, indicating that these might be topologically inaccessible if moving the structure in Cartesian space. In contrast, models with good fit to the SAXS data were obtained by TNMA for all 18 test cases. However, for both NMA methods, the best rmsd model was not necessarily the one with the best fit to the SAXS data, indicating a need for additional information to resolve this ambiguity.

The results from the individual test cases are shown in Figure 2-4. One notable result is that the magnitude of the initial χ^2 is not necessarily indicative of the rmsd between initial and target structures. An example case where the initial χ^2 suggests a higher initial rmsd than is actually the case is shown in Figure 2-4 (panel B1). In this example, the structure is small, so even sub-10 Å movements could result in a significantly more compact shape (and consequently, a very different scattering profile). In other cases, a high initial χ^2 but small initial rmsd causes cases such as that shown in Figure 2-4 (panel B3), where TNMA resulted in better fit to the scattering data, but worse rmsd than the initial structure. However, this "worsening" puts the model rmsd still in the sub-10 Å range, and shows the overall change in shape (in this case, that the RNA bending increases), which is to be expected, given the resolution limits of SAS.

Table 2-3. NMA benchmarking results for small RNA dataset

	number of cases						
	Initial	model with $\chi^2 \leq 2$ found		model with improved rmsd found		model with best χ^2 has improved rmsd	
		CNMA	TNMA	CNMA	TNMA	CNMA	TNMA
open to closed	9	7 (78%)	9 (100%)	6 (67%)	8 (89%)	5 (56%)	6 (67%)
closed to open	9	7 (78%)	9 (100%)	5 (56%)	6 (67%)	2 (22%)	4 (44%)

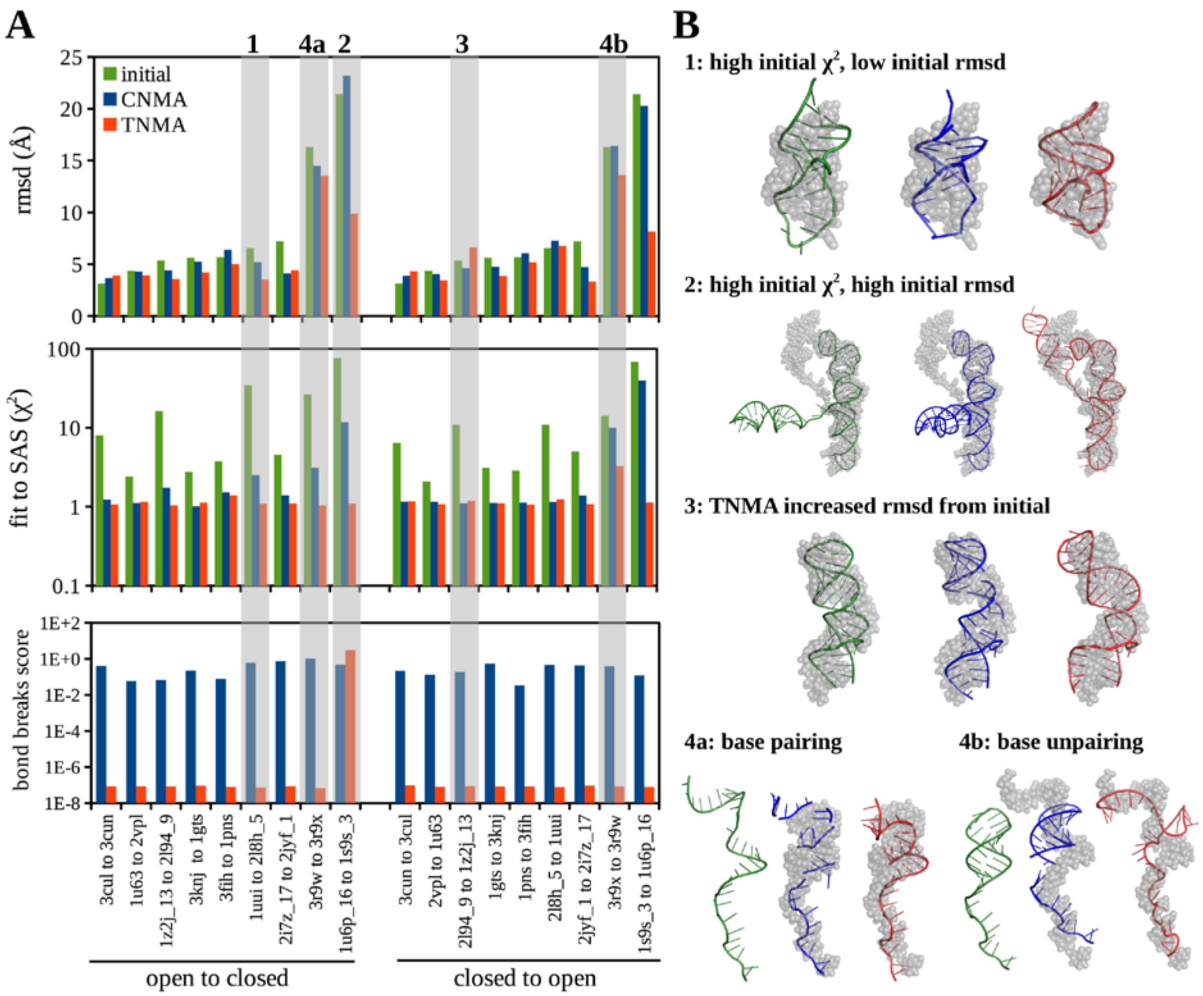

Figure 2-4. The lowest rmsd models from Cartesian and torsional SAS-guided NMA, compared to the initial structure. (A) For the majority of cases, both CNMA and TNMA were able to get models with both a good fit to the SAXS data, and improved rmsd from the initial structure. However, TNMA outperforms CNMA in terms of model stereochemistry in almost all cases. For the highest initial rmsd cases (rightmost bars of open-to-closed and closed-to-open panels), TNMA was able to find models of significantly lower rmsd than CNMA. Interesting cases are highlighted and numbered, and shown in B. (B) Comparison of the initial (green) and target (gray) structures, and the best models found by CNMA (blue), and TNMA (red).

SAS-guided structural modeling clearly provides the best results for cases where the high initial χ^2 is observed due to a pronounced change in the overall structure. Such a case is presented in Figure 2-4 (panel B2). The structural change involves a large pivot of the short helix, which was reached by torsional NMA, but not by CNMA, indicating that this movement was not accessible as linear motions in Cartesian space without significantly breaking stereochemistry.

Both NMA methods, however, were not able to establish hybridized base-pairing from an unpaired initial state (Figure 2-4, panel B4a). Both methods attempted to get the double-helical shape of the self-complementary region by compressing the same end into a tighter single-helix, with the TNMA model having better stereochemistry than the CNMA model. The inverse problem (i.e. undoing self-complementary regions; Figure 2-4, panel B4b) seems to be easier, with torsional NMA able to get a model with almost no base-pairing. This result indicates the need for accurate secondary structure information, which can be obtained from both experiment and prediction tools.

The larger RNA benchmark yielded similar results. Table 2-4 shows the results of TNMA and CNMA benchmarking on the large RNA dataset. Similar to the small benchmark, CNMA was not able to find models with good fit to the SAS data for a significant fraction of the dataset, indicating that these might have necessitated bond-breaking moves in Cartesian space. TNMA, on the other hand, was able to find models that fit the simulated SAXS data in all the cases.

Both NMA methods were able to find models with improved rmsd over the initial structure for the majority of cases, with TNMA doing particularly well in the open-to-closed cases. However, as in the initial benchmark run, these improved models were not always the ones with best χ^2 fit to the SAXS data. This is due to the ambiguity of SAS-based modeling, and highlights the need for orthogonal datasets, to be able to pick the best out of a pool of likely models.

Table 2-4. NMA benchmarking results for large RNA dataset

	number of cases						
	initial	model with $\chi^2 \leq 2$ found		model with improved rmsd found		model with best χ^2 has improved rmsd	
		CNMA	TNMA	CNMA	TNMA	CNMA	TNMA
open to closed	66	44 (67%)	66 (100%)	51 (77%)	64 (97%)	41 (62%)	36 (55%)
closed to open	68	42 (62%)	68 (100%)	57 (84%)	61 (90%)	47 (69%)	52 (76%)

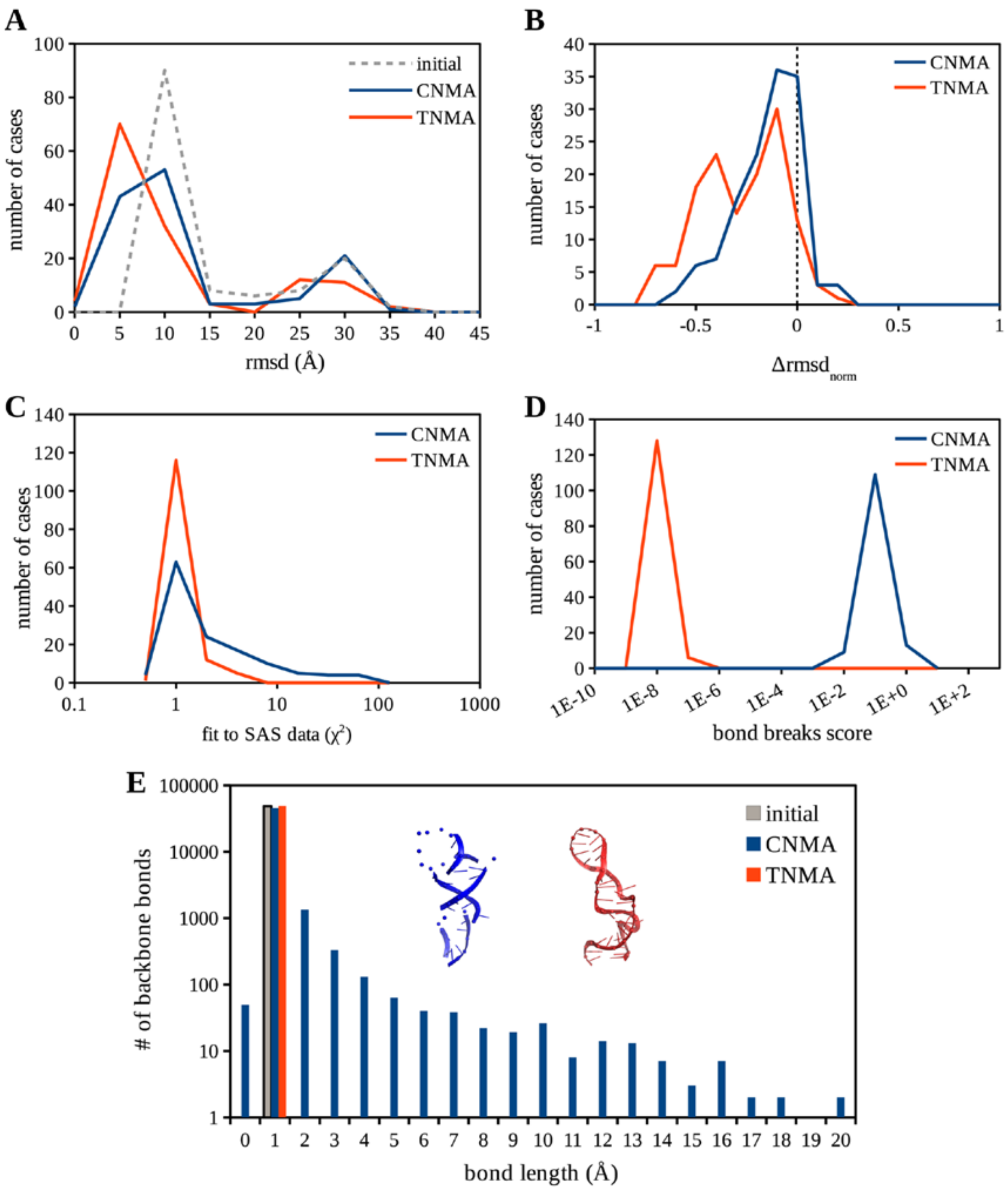

Figure 2-5. Benchmarking results from Cartesian and torsional NMA on a large RNA dataset. (A) shows the rmsd of the initial from the target structures (broken gray line), as compared to the rmsd of the best CNMA (blue), and TNMA (red) models. The rmsd distribution shifted to the left for both NMA methods, indicating an overall improvement compared to the initial structure, with TNMA resulting in a greater improvement, particularly for high initial rmsd cases. (B) The normalized change in rmsd ($\Delta rmsd_{norm}$) shows similar results, with both NMA methods resulting in a net rmsd decrease, but with TNMA resulting in a greater magnitude of improvement for more cases. (C) Goodness-of-fit (χ^2) to the target SAXS data shows that TNMA was able to obtain models with good χ^2 fit (χ^2 ~ 1) for more cases than CNMA. (D) shows the bond breaks score of the pool of best models from CNMA and TNMA. Typical breaks scores from TNMA models are several orders of magnitude lower than from CNMA models, indicating that the bond lengths were largely preserved by torsional NMA. (E) The histogram of backbone bond lengths shows that TNMA largely preserves the bond lengths, while around 5% of backbone bonds in the CNMA models are noticeably stretched or compressed compared to the initial structure.

Metrics from the large RNA benchmark run are shown in Figure 2-5. Both NMA methods improved rmsd from the initial structure for cases where the initial rmsd was around 10 Å (Figure 2-5A). For cases with higher initial rmsd (~30 Å), only torsional NMA resulted in visible improvements, which is similar to the result observed in the smaller initial benchmark.

The normalized change in rmsd ($\Delta rmsd_{norm}$) shows that there is greater net improvement of rmsd with torsional than Cartesian NMA (Figure 2-5B). Also similar to the initial benchmark, most of the TNMA models fit the SAS data well ($\chi^2 \sim 1$), while CNMA had a significant number of models which did not (Figure 2-5C). A comparison of the bond breaks scores shows that the average CNMA model had a score that is several orders of magnitude higher than the average TNMA model (Figure 2-5D), indicating that bonds are stretched to a much greater extent when moving the structure in Cartesian space, as compared to torsion angle space. Figure 2-5E compares the backbone bond lengths between the initial RNA structures and the corresponding models from CNMA and TNMA. It was observed that around 5% of the backbone bonds were stretched to 2 Å or more, or compacted to less than 1 Å, from an initial average length of ~1.5 Å. A pair of representative models from Cartesian and torsional NMA are shown in Figure 2-5E to illustrate the difference.

That torsional NMA would result in less bond stretching or breakage was expected, since the bond lengths are implicitly kept fixed when molecular motions are restricted to bond rotations. However, it is important to note that while the Cartesian approximation has been shown to be more than sufficient in modeling protein flexibility, it has been shown here to be inappropriate for modeling nucleic acid structures undergoing large conformational changes. A possible reason might be that protein domains are generally separated by flexible loop regions, while the double helical RNA domains are separated by stiffer, double-stranded junctions that are not topologically independent of the helical domains (Figure 2-6). Specifically, rotations of the bonds at RNA junctions cannot happen without the torsional strain being released through rotations in the RNA helices themselves.

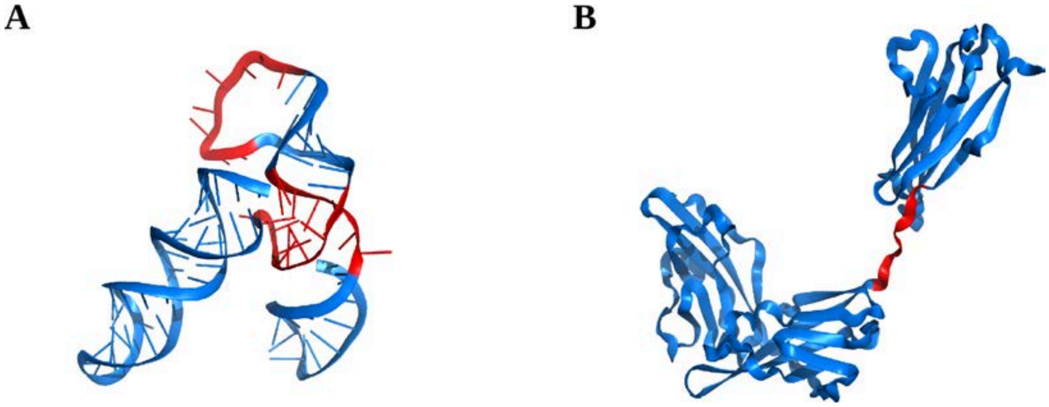

Figure 2-6. Comparison of flexible regions (shown in red) in (A) RNA, and (B) protein structures. RNA flexible regions are the non-base paired regions between double-helical domains, and as such, are often double-stranded. In contrast, protein loop regions are often single-stranded, which results in greater topological freedom compared to the flexible regions in RNA.

The topological constraints on RNA junctions due to base-pairing have been repeatedly observed in high-resolution structures (Lescoute and Westhof, 2006), modeled extensively (Bailor, Sun and Al-Hashimi, 2010; Mustoe *et al.*, 2011), and quantified with RNA-like bodies (Chu *et al.*, 2009). Taken together, these studies seem to indicate that the base-paired regions greatly influence the conformational space that the junctions can occupy, and thus cannot be modeled as independent, rigid bodies, which is the approximation that occurs with RTB-CNMA.

Another interesting thing to note is how well greedy optimization worked for the RNA benchmark. The applicability of the greedy algorithm usually means a corresponding smoothness of the energy function being optimized (in this case, the χ^2 fit) (Cormen *et al.*, 2009). That good models could be reached simply by following the best fitting model per iteration warrants further investigation into the actual smoothness of the χ^2 landscape around the correct model, and whether the method of move generation (moving by bond rotations along normal modes) has a role in this apparent smoothness.

2.5 Conclusion and outlook

A SAS-guided NMA modeling tool in torsion angle space, NMATOR, was developed and tested on RNA structures, and compared to a corresponding SAS-guided Cartesian NMA method, SREFLEX. RNA structure pairs, each with the same sequence but different structures, were used to assess the performance of the two NMA algorithms. Torsional NMA outperformed CNMA in terms of model stereochemistry, as well as in model correctness, predicting a model closer to the target structure in most of the cases. This improvement makes a compelling case for torsional NMA as a method for modeling RNA conformational change, especially given its advantages in terms of ease and accessibility compared to molecular dynamics.

NMATOR is available in ATSAS version 3.0.0, and can be used in the command line in three modes: (1) torsional normal mode computation only, (2) fitting an initial high-resolution structure to a given SAS profile, and (3) generating a pool of conformations from an initial structure. As of this writing, NMATOR only works for single chains, and requires at least a full backbone structure. There is also currently no support for hetero atoms, such as attached metal ions. Multichain and hetero atom support is slated to be added in the next ATSAS release.

Conceptually, NMATOR could be used to perform TNMA on protein and DNA structures, but this has not been extensively tested. This could be a direction of subsequent studies, and based on the results of this work, the approach may provide advantages in modeling conformational changes in proteins with limited portions of defined structure.

3 Using sequence coevolution to reduce SAXS modeling ambiguity

3.1 Sequence coevolution and SAS

The amount of known biological sequence information has increased dramatically in the past decade, propelled by developments in high-throughput nucleic acid sequencing technologies (Goodwin, McPherson and McCombie, 2016). Structural biology, on the other hand, has yet to experience a boom of similar magnitude, despite numerous global initiatives, such as the Structural Genomics Consortium (Chandonia and Brenner, 2006). As of this writing, the number of structures in the PDB (~150 000) differs from the number of gene sequences in UniProt (~500 000 annotated, and ~200 M unannotated) by around a factor of 1000. This very large, accessible set of biological sequences is highly amenable to various statistical and data mining methods.

Among the various approaches that leverage biological sequence information is using coevolution to make predictions of long-range protein contacts. The use of coevolution in protein contact prediction is based on the correlated mutations model: i.e. that there is selective pressure to maintain inter-residue interactions that are essential for function. Thus, if a mutation event occurs in one of the interacting loci, the other locus must also mutate (hence, "coevolve") in order to maintain the interaction. This coevolution signal is detected by analyzing the sequences of the same protein across multiple species, and statistically determining which pairs of positions tend to covary (Juan, Pazos and Valencia, 2013) (Figure 3-1).

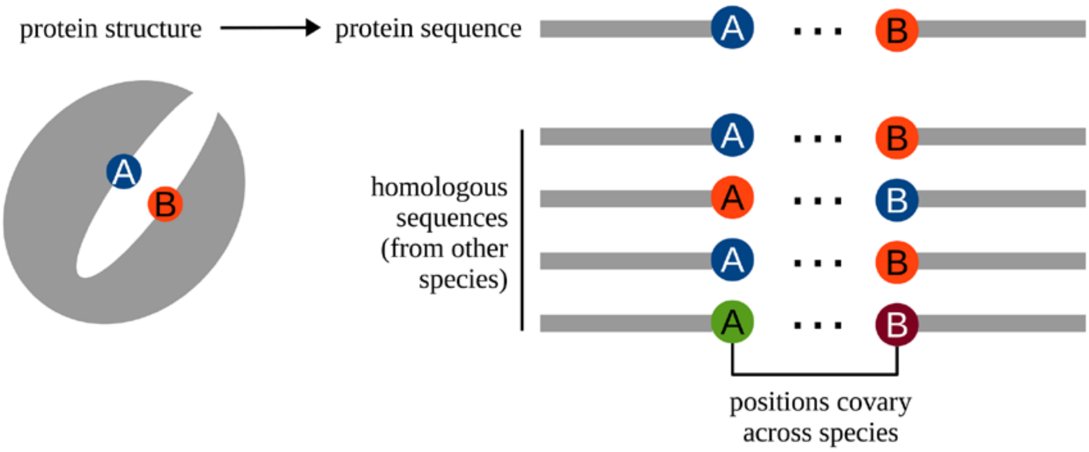
protein structure
protein sequence
A
B
homologous sequences (from other species)
positions covary across species

Figure 3-1.* Using coevolution to derive structural information.** ***The assumption made in coevolution analysis is that if positions A and B in a protein are interacting, and if A mutates, there is selective pressure for B to mutate as well. This coevolution could be tracked by examining positions A and B of the protein across multiple species. Conversely, if two positions in a protein sequence are seen to be coevolving, there is a greater probability that they are interacting in the protein structure.

Applications of coevolution information include the prediction of long-range contacts within single proteins (Yeang and Haussler, 2007; Morcos *et al.*, 2011; Marks, Hopf and Sander, 2012; Wang and Xu, 2013), and between monomers in a homodimer (Dos Santos *et al.*, 2015). The method can also be extended to predicting contacts between subunits in a complex, simply by concatenating the sequences of the subunits into one long sequence. This approach has been shown to capture inter-protein interactions in highly-conserved systems such as the ribosome, and more generally, bacterial complexes (Halperin, Wolfson and Nussinov, 2006; Hopf *et al.*, 2014; Ovchinnikov, Kamisetty and Baker, 2014; Feinauer *et al.*, 2016). However, there are currently no published studies using coevolution analysis to predict contacts in eukaryotic protein complexes. This is due to the inherent requirement for the protein subunits to be interacting in all of the evaluated species, which is more difficult to establish in eukaryotes than in bacteria, where interacting proteins are often located close together in the genome, often in the same operon. Eukaryotic genomes, with their much less compact organization, are not amenable to the same approach.

In addition, no studies have thus far been published that combined sequence coevolution with SAS data in order to model protein structures. Conceptually, however, the two methods are complementary. SAS provides overall geometric information that defines a protein's gross structural features, while the residue contacts predicted by coevolution could reduce the ambiguity of SAS-guided structure modeling.

Long-range contacts could be obtained through experimental methods, such as cross-linking mass spectrometry (MS) or fluorescence resonance energy transfer (FRET) (Selvin, 1995; Sinz, 2006). However, *in silico* methods could still be a useful part of the structural biology toolkit, in that they can serve as a relatively quick and cheap aid in experimental design. In this case, for example, coevolution-based contact predictions could serve as candidates for contact validation by site-directed mutagenesis.

This work consists of two main parts: (1) developing and evaluating a coevolution-based method for heterodimer contact prediction, and (2) using coevolution-predicted contacts for SAS-guided rigid body modeling.

3.2 Quantifying the accuracy of coevolution-based contact predictions

To evaluate whether sequence coevolution could be used to predict heterodimer contacts, a set of representative heterodimers was acquired from the PDB. The representative heterodimers were queried to each have the following properties:

1. The X-ray resolution of the structure is at worst 3 Å
2. Sequence homology is 30% or less compared to other heterodimers in the set
3. The heterodimer must not have a high homodimeric tendency:
    a. Low structural similarity between subunits: when the subunit structures are aligned, $\frac{rmsd}{\#residues} > 0.03\ \text{Å}/residue$
    b. Low sequence identity between subunits: when the subunit sequences are aligned, the sequence identity is less than 25%

Heterodimers with a high homodimeric tendency were excluded from the representative heterodimer set because for these cases, the coevolution signal is dominated by equivalent positions on the subunits, instead of the interacting pairs of residues.

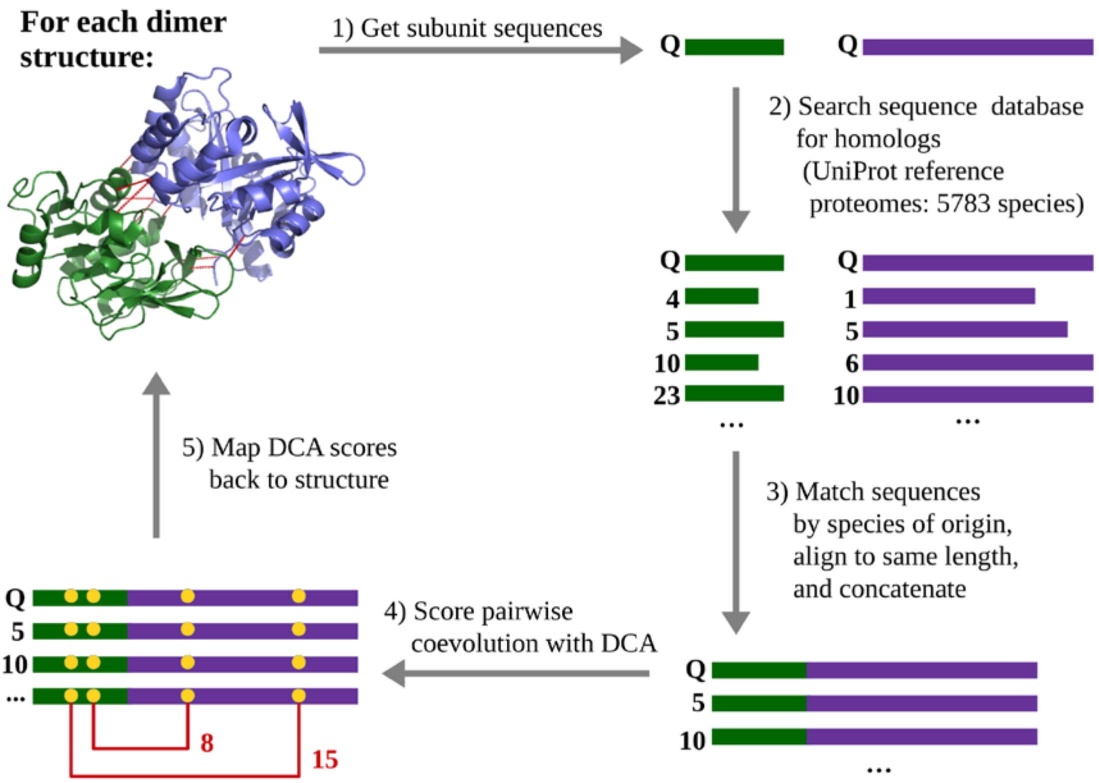
For each dimer structure:
1) Get subunit sequences
Q
Q
2) Search sequence database for homologs (UniProt reference proteomes: 5783 species)
Q
4
5
10
23
...
Q
1
5
6
10
...
3) Match sequences by species of origin, align to same length, and concatenate
Q
5
10
...
4) Score pairwise coevolution with DCA
Q
5
10
...
8
15
5) Map DCA scores back to structure

Figure 3-2. **Heterodimer DCA workflow.** ***Full-length protein sequences were obtained from UniProt. The sequences (labeled Q) were used to query the UniProt Reference Proteomes database (version 2016_08), containing the complete proteomes of 5783 species. The top match from each proteome was taken, with sequences from the same proteome matched (unmatched sequences were discarded). Each remaining sequence was aligned, concatenated with its partner sequence, and the resulting concatenated multiple sequence alignment used for DCA. The DCA scores were then mapped back to the dimer structure.***

Using the above criteria, 177 representative heterodimers were obtained (for full list, see Appendix). The amino acid sequences of the proteins were obtained from UniProt (Wasmuth and Lima, 2016), along with the mapping of each residue in the UniProt sequence to its position in the PDB file (Martin, 2005). The amount of coevolution between each pair of residues in each heterodimer was then computed through the following workflow (summarized graphically in Figure 3-2). Related, homologous sequences to each heterodimer subunit were queried with HMMer (version 3.1b2) (Eddy, 2011), from the UniProt database of reference proteomes (version 2016_08) (UniProt Consortium, 2011), which contained around 6000 complete proteomes at the time of analysis. Only hits with at least 70% the length of the query sequences were kept. Of the remaining hits, only the ones where both sequences could be found in the same species were kept. The underlying assumption behind this species-matching is that if sequence homologs of both subunits of a heterodimer are found in the proteome of another species, they must also be forming a heterodimer in that species. This is a major assumption that is expected to add some noise to the analysis, which will be discussed in the next section.

Each heterodimer subunit sequence was aligned to its remaining homologous sequences with Clustal Omega (version 1.2.3) (Sievers *et al.*, 2011). The resulting multiple sequence alignments were then concatenated into one long MSA for each heterodimer. Each heterodimer MSA was subjected to direct coupling analysis (DCA), a statistical method to detect covarying positions in the alignment (Morcos *et al.*, 2011).

As output, DCA gives a list of all pairs of positions in the concatenated sequence, along with the likelihood that the positions are coevolving. We will call this likelihood the DCA score, and consider it a likelihood that the pair of residues is interacting in the structure. All scores produced by DCA were converted to z-scores (number of standard deviations above mean), so that they can be compared between heterodimers.

$$z = \frac{DCA_{ij} - \mu_{DCA}}{\sigma_{DCA}} \tag{3-1}$$

where DCA_{ij} is the DCA score between the ith and jth residues in the MSA, μ_{DCA} is the mean DCA score for all i, j, and σ_{DCA} the standard deviation.

Since we are looking for long-range contact predictions, residue pairs that are in the same subunit were discarded, leaving only the inter-subunit pairs. This left a total of 5.2 million residue pairs, which were mapped back to the PDB structures. Residue pairs were classified as contacts if they were within 10 Å inter-Cα distance.

From this survey of heterodimers, it was observed that when picking a pair of residues at random, one from each subunit, the likelihood that the pair forms a contact is around 1%. If one takes coevolution into account, this likelihood increases depending on the DCA score and the number of sequences used for coevolution analysis. Figure 3-3A shows how the likelihood that an inter-subunit pair of residues forms a contact increases proportional to the computed DCA score and the number of sequences used in the alignment.

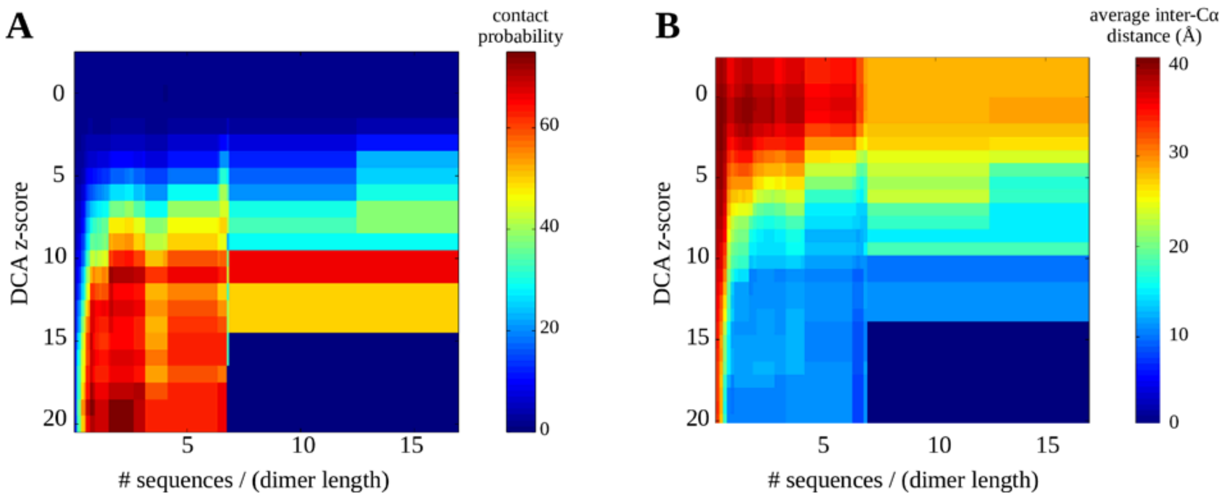

Figure 3-3. (A) A probability map that a pair of residues are within 10 Å, given their coevolution score (y-axis), and the length-normalized number of sequences used for DCA analysis (x-axis). The dark blue area on the bottom right corner represents regions where no information is available (i.e. the coevolution score and/or number of sequences are too high). The contact probability is directly proportional to both the coevolution score and the number of sequences used. In particular, the DCA score required to reach a certain contact probability decreases as the number of sequences used is increased. This indicates that reliability of the coevolution score is highly dependent on having a minimum amount of sequences. (B) The average inter-Cα distance also reflects that a pair of residues are more likely to form a contact (cyan to light blue), the higher the coevolution score (and the greater the number of sequences used to derive the coevolution score).

From Figure 3-3, it can be seen that the reliability of DCA-based contact prediction is highly-dependent on the number of sequences used for analysis. The sequence-dependence of coevolution analysis has been reported in previous work (Ovchinnikov, Kamisetty and Baker, 2014; Dos Santos *et al.*, 2015; Feinauer *et al.*, 2016). The maximum contact probability represented by at least 30 inter-subunit contacts was around 70%. To reach this region of 70% contact probability, one either needs a large number of sequences, a high DCA score or both. There seems to be a drastic reduction in the minimum DCA z-score required to get 70% contact probability at around the 1 sequence/(dimer length) mark (see inflection point on Figure 3-3A). Specifically, a good rule of thumb to get a 70% prediction confidence is to use at least 1 sequence per residue in the heterodimer (i.e. smaller heterodimers require fewer sequences, and larger heterodimers more), and to take the top DCA scoring pairs as contacts if their DCA z-score is at least 15 (i.e. the coevolution score is fifteen standard deviations above the average). For heterodimers where the combined sequence length is 1000 residues, this would mean that at least 1000 species-matched sequences for both subunits is recommended.

This large sequence requirement is around the same as what was previously reported in other work that used DCA. The minimum coevolution score, on the other hand is quite high. This might be caused by the noise introduced by the assumption of interaction, which would not be true for all the sequences included in this heterodimer survey. However, limiting the analysis only to confirmed interactions would have severely limited the dataset. Nonetheless, this method is very suitable for systems where there are many sequences available, such highly-conserved eukaryotic biological pathways. For these cases, DCA can identify contact pairs which can be used as distance constraints for SAS-based structure modeling, when experimental evidence of these contacts are not available.

3.3 Using DCA-predicted contacts in SAS-guided modeling

To test the effect of providing contact predictions from coevolution on SAS-guided modeling, we selected 17 heterodimers from the dataset for which the top contact was predicted at 70% confidence (Table 3-1). SAXS data was simulated for each heterodimer using CRYSOL (Svergun, Barberato and Koch, 1995) . Angle-dependent random errors were added to the simulated intensities based on the variance of 1000 independently-measured 1s scattering frames of water, as previously described (Franke, Jeffries and Svergun, 2015). Each heterodimer structure was then reconstructed from the subunits and the simulated data using the SAS-guided rigid-body modeling method SASREF (Petoukhov and Svergun, 2005). SASREF was performed twenty times for each heterodimer both without additional information, and using the top predicted heterodimer contact as a distance constraint.

Model fitness was quantified using three metrics: (1) the ligand rmsd ($Lrmsd$), (2) the fraction of native contacts (f_{nat}), and (3) the model's χ^2 fit to the simulated SAXS data. Ligand rmsd is the rmsd of the smaller subunit from the target structure, if the larger subunits on the model and the target structures are aligned. The fraction of native contacts refers to how many inter-subunit heavy atom pairs within 5 Å in the target structure can be found in the model. We define a model to be acceptable if the $Lrmsd$ is at most 10 Å, and the $f_{nat} \geq 0.1$, consistent with CAPRI (Critical Assessment of Prediction of Interactions) standards (Lensink, Méndez and Wodak, 2007).

Table 3-1. Heterodimers with DCA-predicted contacts at 70% probability

DIMER PDB ID	SUBUNIT 1	SUBUNIT 2	CONTACT PREDICTED	ACTUAL CONTACT DISTANCE (Å)
1EUD	Mitochondrial succinate-CoA ligase subunit alpha	Mitochondrial succinate-CoA ligase subunit beta	A_106 (GLY) to B_228 (ASN)	4.9
1FM0	Molybdopterin synthase sulfur carrier subunit	Molybdopterin synthase catalytic subunit	D_11 (ARG) to E_53 (GLU)	11.3
1FS0	ATP synthase, epsilon subunit	ATP synthase, gamma subunit	E_81 (ASP) to G_222 (ARG)	8.8
1KA9	Imidazole glycerol phosphate synthase subunit HisF	Imidazole glycerol phosphate synthase subunit HisH	F_220 (GLU) to H_115 (ARG)	10.0
1R6O	ATP-dependent Clp protease adapter protein ClpS	ATP-dependent Clp protease ATP-binding subunit ClpA	C_79 (GLU) to A_83 (SER)	6.2
1RP3	Anti-sigma factor FlgM	RNA polymerase sigma factor FliA	B_76 (ASP) to A_183 (SER)	6.1
3A1P	Ribosome maturation factor, rimM	30S ribosomal protein S19	B_53 (ASN) to A_82 (PRO)	33.4
3EGV	50S ribosomal protein L11	Ribosomal protein L11 methyltransferase	B_43 (ALA) to A_193 (TYR)	32.6
3FPN	UvrB interaction domain	UvrA interaction domain	B_166 (GLU) to A_210 (LYS)	15.8
3ZEU	Putative M22 peptidase yeaZ	tRNA N6-adenosine threonylcarbamoyltransferase	D_43 (GLN) to E_100 (PHE)	8.8
4A9A	Ribosome-interacting GTPase 1	Translation machinery-associated protein 46	A_256 (SER) to C_238 (LEU)	6.2
4LX3	DNA-directed DNA polymerase	Nucleic acid binding, OB-fold, tRNA/helicase-type	A_-1 (HIS) to B_35 (SER)	11.6
4QTT	18S rRNA (guanine(1575)-N(7))-methyltransferase	Multifunctional methyltransferase subunit TRM112	B_105 (PRO) to A_2 (LYS)	15.6
4XD9	Putative ribosome biogenesis protein, Rpf2	Ribosome biogenesis regulatory protein, Rrs1	A_26 (GLY) to B_98 (HIS)	9.9
5DUD	Uncharacterized protein YbgJ	Uncharacterized protein YbgK	B_49 (GLY) to A_227 (HIS)	4.9
5JCA	Sulfide dehydrogenase subunit beta	Sulfide dehydrogenase subunit alpha	S_74 (LYS) to L_287 (ASP)	8.3
5UNI	NAD(P) transhydrogenase subunit alpha 2	NAD(P) transhydrogenase subunit beta	A_32 (THR) to B_81 (MET)	9.0

Table 3-2. SAS-based modeling with and without contact prediction

	# of cases		
	Initial	Found acceptable model*	Model with best χ^2 is acceptable*
Unconstrained	17	9 (53%)	8 (47%)
with DCA contact		13 (76%)	12 (71%)

* acceptable means rmsd ≤ 10 and f_{nat} ≥ 0.1

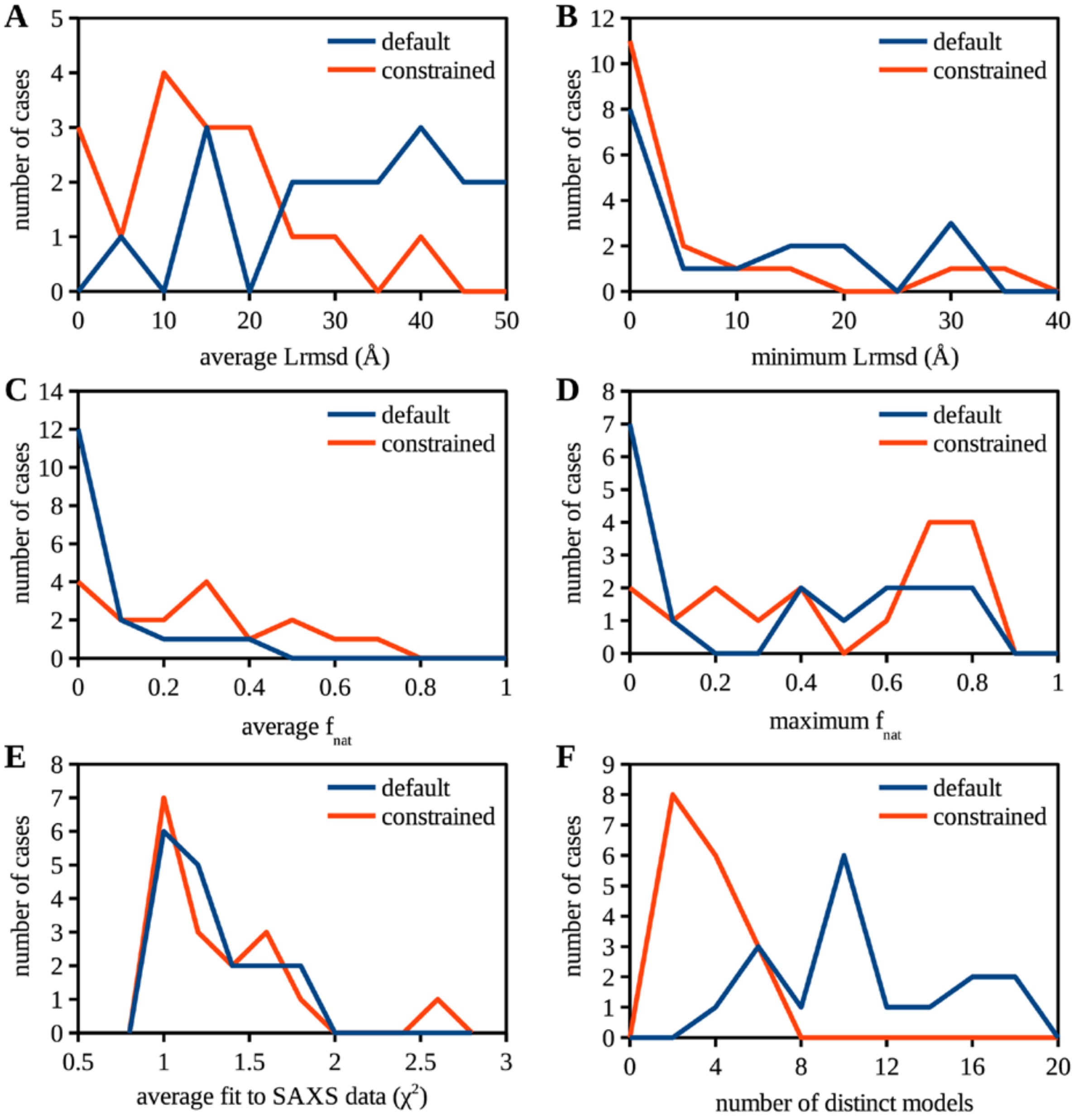

Figure 3-4. Rigid-body modeling with and without a distance constraint from coevolution analysis. (A) shows the average $Lrmsd$ achieved in twenty SASREF runs for each the 19 heterodimers. The average $Lrmsd$ is lower in the constrained case, indicating that the pool of solutions is closer to the target structure in the constrained case, compared to the default. (B) shows the minimum $Lrmsd$ achieved for each heterodimer. The constrained modeling cases were able to reach lower $Lrmsd$ for a larger number of cases, except for when the distance constraint specified was wrong (the two cases with $Lrmsd \sim 30$ Å). (C) shows the average f_{nat} was higher for the pool of constrained models, compared to the default models, indicating an improvement in average model quality with the addition of one distance constraint. (D) The quality of the best model in terms of f_{nat} was also improved with the addition of a distance constraint, indicating that the correct binding interface and ligand orientation was captured. (E) The resulting fits to the SAXS data from both unconstrained and constrained runs were similar, indicating that good fits to the data were found in both cases. However, (F) ambiguity, as measured by the number of distinct models in twenty runs, was markedly decreased in the constrained case.

The overall results of rigid-body modeling on each heterodimer are shown in Table 3-2. Without constraints, a good model was found in around 50% of the cases. Adding one DCA-predicted distance constraint improved this, with an acceptable model being found in 13 out of 17 cases (76%). The likelihood that the model with the best χ^2 fit to the SAXS data also increased in the constrained versus the default case.

Figure 3-4 is a more detailed look at the rigid-body modeling results in terms of $Lrmsd$, f_{nat}, χ^2, and modeling ambiguity. Adding one distance constraint resulted in overall lower $Lrmsd$ models, compared to the unconstrained case. There was also a significant improvement in f_{nat} upon the addition of a distance constraint, indicating that the correct binding interface and ligand orientation was captured more often in the constrained cases. On the other hand, there was no significant overall difference in the model fits between unconstrained and constrained cases, indicating that the fit to the SAXS data alone is insufficient to resolve which set of models are more accurate.

Adding even a single distance constraint noticeably reduced modeling ambiguity, lowering the number of distinct models found in twenty independent SASREF runs. This is an expected result, since the distance constraint effectively reduces the solution search space. One distance constraint was not enough to reduce the number of distinct models to one (the ideal case where there is absolutely no ambiguity), but the addition of complementary information—such as biological insights, or other experimental data—could resolve the remaining ambiguity.

Figure 3-5 is a visual comparison of the results of SAS-guided rigid-body modeling with and without the coevolution-derived distance constraint. In general, it can be seen that adding one distance constrained improved the likelihood of getting a solution close to the PDB model, and also markedly decreased the variability of the set of resulting models.

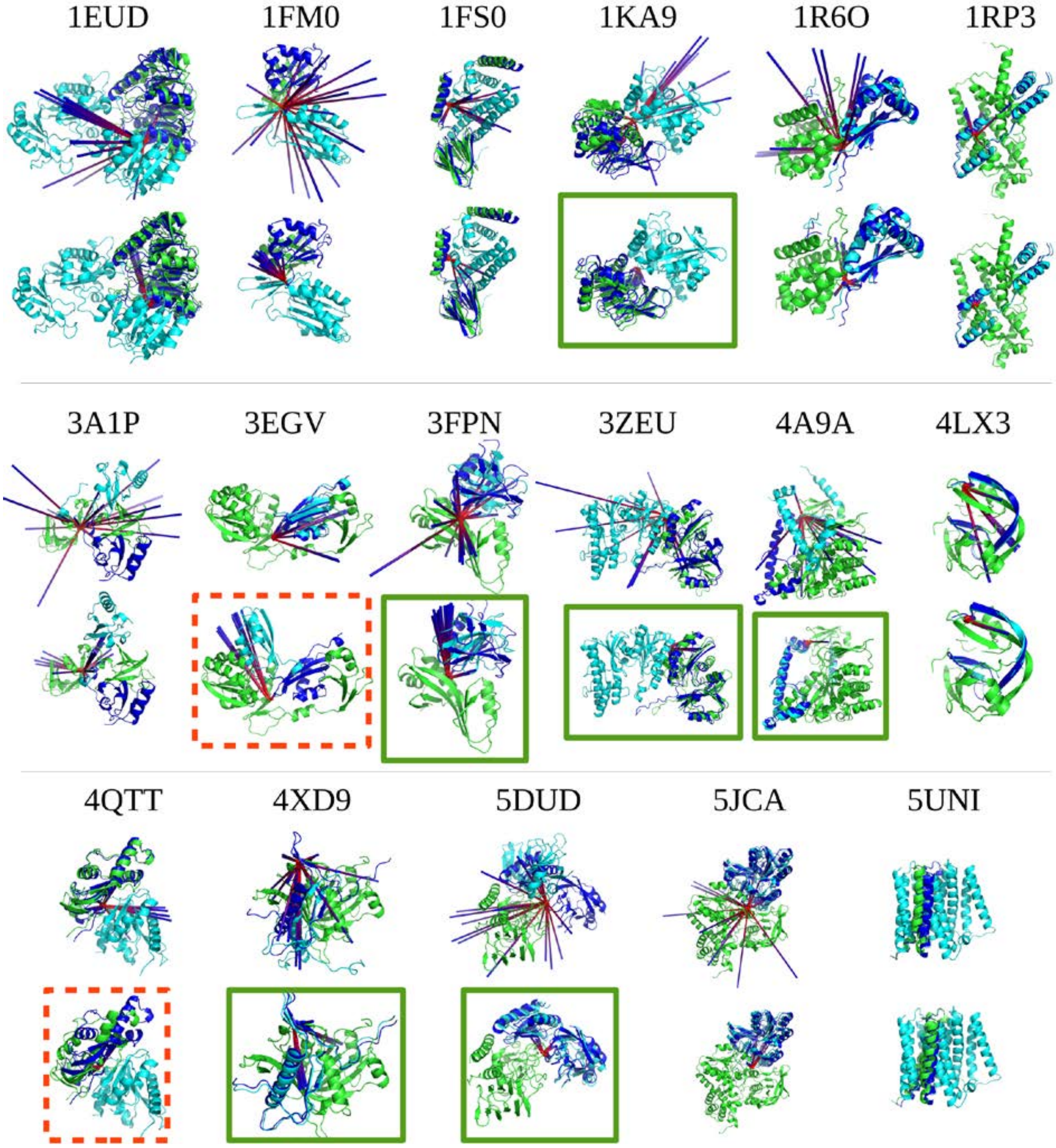

Figure 3-5. Rigid-body modeling results for 17 independent unconstrained (top), and constrained (bottom) SASREF runs. The orientation of the ligand in the PDB structure is shown in dark blue. The purple lines define the shortest inter-subunit distance found in the PDB structure; thus, long lines indicate that the ligand is placed far from the correct position. The lines also give a visual assessment of the variability of the pool of models produced by SASREF (a tighter cluster of lines indicates a less varied set of solutions). Generally, the variability of the models decreased with the addition of a distance constraint. The solid green boxes highlight cases in which an acceptable model was found only in the constrained runs, while the dotted red boxes indicate the opposite case (acceptable model found only in unconstrained runs). Providing one distance constraint to SAS-guided rigid body modeling increased the likelihood of reconstructing a solution near the PDB structure, and also decreases overall model variability.

Among the 17 heterodimers, there were two cases for which the contact predicted by coevolution analysis was known to be very false (inter-Cα distance >> 10 Å): 3A1P, and 3EGV. For these cases, we were effectively biasing the modeling towards the wrong solutions by providing the false information.

Figure 3-6 shows that in cases where contact prediction is wrong, the difference in χ^2 fits between the pool of constrained and unconstrained solutions could be a hint that something is amiss. For 3A1P and 3EGV, it was indeed the case that the SAXS data served as a check for contact prediction accuracy. In particular, in the cases where the distance constraint was wrong, the pool of constrained SASREF models had poorer average fit than the pool of unconstrained models. This same trend does not hold for the cases wherein the constraint was correct. Therefore, just as contact predictions from coevolution can enhance SAS-based modeling by reducing ambiguity, SAS data can in turn validate DCA contact predictions, at least in some cases.

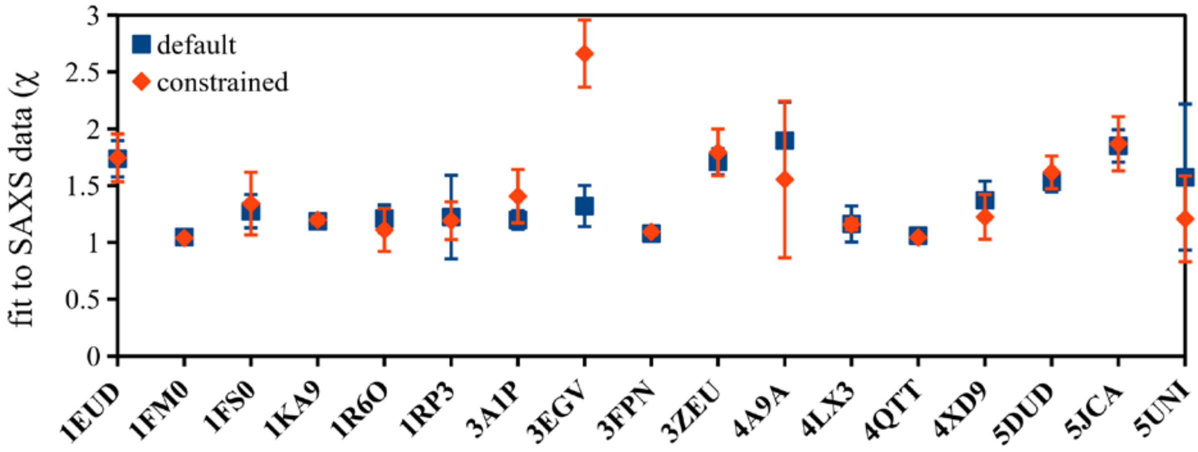

Figure 3-6.* Rigid body modeling with incorrect contact information.** ***For the heterodimers 3A1P and 3EGV, the contacts predicted by coevolution were false (pair distance >> 10 Å). Using these false contacts for a set of rigid body modeling runs resulted in a pool of structures with noticeably worse fit to the SAXS data, compared to the pool of models from unconstrained runs. In contrast, for cases where the contact information was correct, the average fit of the pool of constrained models was not markedly worse than the pool of unconstrained models. This indicates that the SAXS data can verify whether contact predictions are correct, for some cases.

3.4 Conclusion and outlook

The viability of inter-residue coevolution, as quantified with the direct coupling analysis method, in predicting heterodimer contacts was examined. The accuracy of the method has previously been shown to be highly dependent on the quantity of sequence information provided, and the same was observed here. This makes the technique suitable for systems where there much sequence information is available, such as bacterial complexes, or eukaryotic complexes in highly-conserved pathways. While coevolution-based heterodimer contact prediction might not be generally applicable at present, the approach is worth revisiting when the UniProt set of reference proteomes significantly grows in size.

Additionally, we showed that the combination of SAS data and contact information can result in better quality, less ambiguous models than either method by itself. In particular, contact predictions from coevolution can reduce SAS modeling ambiguity, while SAS data can serve as a filter for wrong contact predictions from DCA.

4 Computing anomalous scattering from structure

4.1 Anomalous SAXS: concept and biological applications

So far, our discussion of small-angle scattering has focused on elastic scattering of the incident radiation. Indeed, in the general case, the scattering can be adequately approximated as purely elastic. However, this approximation breaks under certain conditions. In the case of anomalous X-ray scattering, for example, atoms in the sample can alter the total scattering if the incident X-ray wavelength is close to their absorption edge, i.e. to energies that correspond to electronic transitions in a particular element. As a result, if the energy of the incoming photons is close to or at an absorption edge, atoms absorb the incident radiation, and electrons are excited to higher energy states.

At wavelengths far from the absorption edge, photoelectric absorption does not occur to a significant extent. The atomic scattering length is therefore only dependent on the number of electrons, Z. However, at wavelengths close to the absorption edge, the photoelectric effect begins to play a non-negligible role. The X-ray scattering factor, $f(\lambda)$, could then be represented as a function containing a complex term:

$$f(\lambda) = f_0 + f'(\lambda) + if"(\lambda) \tag{4-1}$$

where f_0 is the Z-dependent, λ-independent factor, and the magnitude of the λ-dependent factors f' and $f"$ increase as the wavelength gets closer to the absorption edge of the particular element (James, Bragg and Bragg, 1948).

At present, the most common biological application of anomalous scattering is the *de novo* phasing of X-ray diffraction data (Hendrickson, 2014). Anomalous small-angle X-ray scattering (ASAXS) has had relatively few published biological applications, which include estimating the distances between the four iron atoms bound to hemoglobin (Stuhrmann and

Notbohm, 1981), terbiums in the calcium-binding sites of parvalbumin (Miake-Lye, Doniach and Hodgson, 1983), and metal ion clouds around DNA (Pabit *et al.*, 2010).

A possible reason why ASAXS experiments are not more common, aside from the scarcity of wavelength-tunable X-ray sources, is the low contrast of SAXS experiments due to strong solvent scattering. Whereas anomalous scattering from native sulfur atoms can sometimes be used for phasing in X-ray crystallography experiments (Hendrickson and Teeter, 1981), their effect is very difficult to detect in solution SAXS experiments (D. Svergun, personal communication). However, the use of atoms with higher correction factors f' and f" could result in strong enough absorbance to overcome the low contrast in solution SAXS experiments, due to both strong solvent scattering, and dilute sample concentrations. A method of computationally simulating ASAXS data, therefore, could be a useful tool in guiding experimentalists in which ASAXS experiments have a reasonable chance of overcoming the low contrast of solution SAXS experiments. A number of tools have been developed to simulate SAS data from molecular structure, but the approaches can be roughly divided into two conceptual families: those based on the Debye formula (Pantos and Bordas, 1994; Schneidman-Duhovny, Hammel and Sali, 2010; Stovgaard *et al.*, 2010; Dos Reis, Aparicio and Zhang, 2011), and those based on spherical harmonics (Svergun, Barberato and Koch, 1995; Grishaev *et al.*, 2010).

The scattering intensity can be computed from N discrete spherical bodies using the Debye formula (Debye, 1915), a discrete form of eq. 1-2:

$$I(s) = \sum_{i=1}^{N} f_i^2\,(s) + 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} f_i(s) f_j(s)\,\frac{sin(sr_{ij})}{sr_{ij}} \qquad (4\text{-}2)$$

Where $f_i(s)$ is the form factor of sphere i, and r_{ij} the distance between the centers of spheres i and j. While conceptually simple, the calculation of scattering using the unmodified Debye formula can get slow for even small to medium-sized biological molecules. Put in terms of the big O notation, which gives an estimate of computational run time as a function of input size, computing the scattering from a macromolecule with N atoms necessitates $(N)(N-1)/2$ computations with the Debye formula, making the run time in the order of $O(sN^2)$.

In comparison, spherical harmonics-based methods are particularly suited to cases where N is not small. The scattering is represented as a combination of spherical harmonics

(Y_{lm}), which are a set of angular basis functions in spherical coordinates $(r,\omega)=(r,\theta,\varphi)$, which are composed of trigonometric functions of orders l and m:

$$Y_{lm}(\theta,\varphi)=\sqrt{\frac{(2l+1)(l-m)!}{4\pi(l+m)!}}P_l^{|m|}(cos\,\theta)exp(im\varphi) \tag{4-3}$$

Where l and m are integers, $0\le l<\infty$, $-l\le m\le l$, and $P_l^{|m|}(\cos\theta)$ are associated Legendre polynomials of the first kind. The scattering $I(s)$ of a system of N atoms is then:

$$I(s)=2\pi^2\sum_{l=0}^{\infty}\sum_{m=-l}^{l}|A_{lm}(s)|^2 \tag{4-4}$$

Where $A_{lm}(s)=4\pi i^l\sum_{j=1}^{N}f_j(s)j_l(sr_j)Y_{lm}^*(\omega_j)$, $f_j(s)$ is the form factor of atom j, (r_j,ω_j) its spherical coordinates, and $j_l(sr_j)$ are spherical Bessel functions.

Lower order harmonics define gross structural features, while higher order harmonics describe finer details (Svergun *et al.*, 2013). As it is impracticable to use $L=\infty$, L is set to a maximum value. A cutoff of $L=50$—already a very good approximation of scattering to wide angles—would put run time in the order of $O(sL^2N)=O(2500sN)$, which makes spherical harmonic-based methods faster than the unmodified Debye formula for systems with at least 2500 atoms (equivalent to a ~125 amino acid protein, or nucleic acid molecule with ~80 nucleotides).

In either case, the computation of anomalous scattering simply entails adding wavelength-dependent terms to the form factor, and does not change the underlying mathematics of either method. The method described below adds the wavelength-dependent terms of the form factor to CRYSOL (Svergun, Barberato and Koch, 1995), a spherical harmonics-based method widely used in the biological SAS community.

4.2 Adding anomalous scattering to CRYSOL

The wavelength-dependent anomalous correction terms f' and $f"$ for atoms from calcium to heavier were obtained from the University of Washington Biomolecular Structure Center (http://skuld.bmsc.washington.edu/scatter/AS_periodic.html), spanning each of their L and K absorption edges. The correction factors for bromine, iron and terbium are shown in Figure 4-1 for the recommended energy range for ASAXS experiments at the EMBL P12 SAXS beamline.

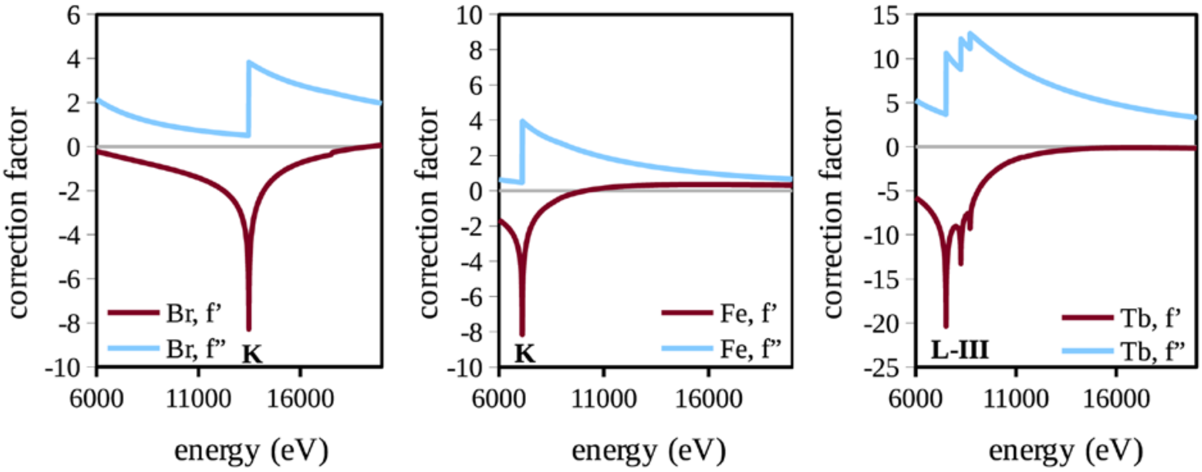

Figure 4-1. Anomalous correction factors f' and f" for bromine, iron, and terbium, shown for the range of energies recommended for ASAXS experiments in the EMBL P12 SAXS beamline. The relevant absorption edges are indicated (K-edge for bromine and iron, L-III edge for terbium).

The f' and $f"$ tables were stored as constants in ATSAS (in libatsas/constants/asaxs_fprimes.for). CRYSOL was then amended to add the wavelength-dependent terms to the scattering factor of affected atoms if a wavelength is specified by the user. In the command line, the ASAXS mode of CRYSOL can be used by using the following command:

```
$ crysol <pdbfile> -en <energy in eV>
```

Optionally, the user can also specify a comma-separated file containing the wavelength-dependent terms of the form factor of a particular element (each line in the format `energy in eV,f’,f”`). CRYSOL can be used in this mode as follows:

```
$ crysol <pdbfile> -en <energy, eV> -ff <.csv file with f’ and f”>
    -el <two-letter element name>
```

The ASAXS module of CRYSOL was compared to an implementation of the Debye formula for a 30 Å-diameter bead composed of a 10 Å-diameter core of bromine atoms surrounded by carbon atoms. Optimal atomic packing was generated using Packmol (Martinez *et al.*, 2009). Scattering data was simulated for this test object at a range of energies around the K-absorption edge of bromine (13474 eV). Figure 4-2 shows that there is good correspondence between the *in vacuo* scattering profiles computed by both methods. The scattering intensities dip at the same values of s, and the absorbance is greatest at the absorption edge, resulting in the lowest scattering intensities at and near 13474 eV.

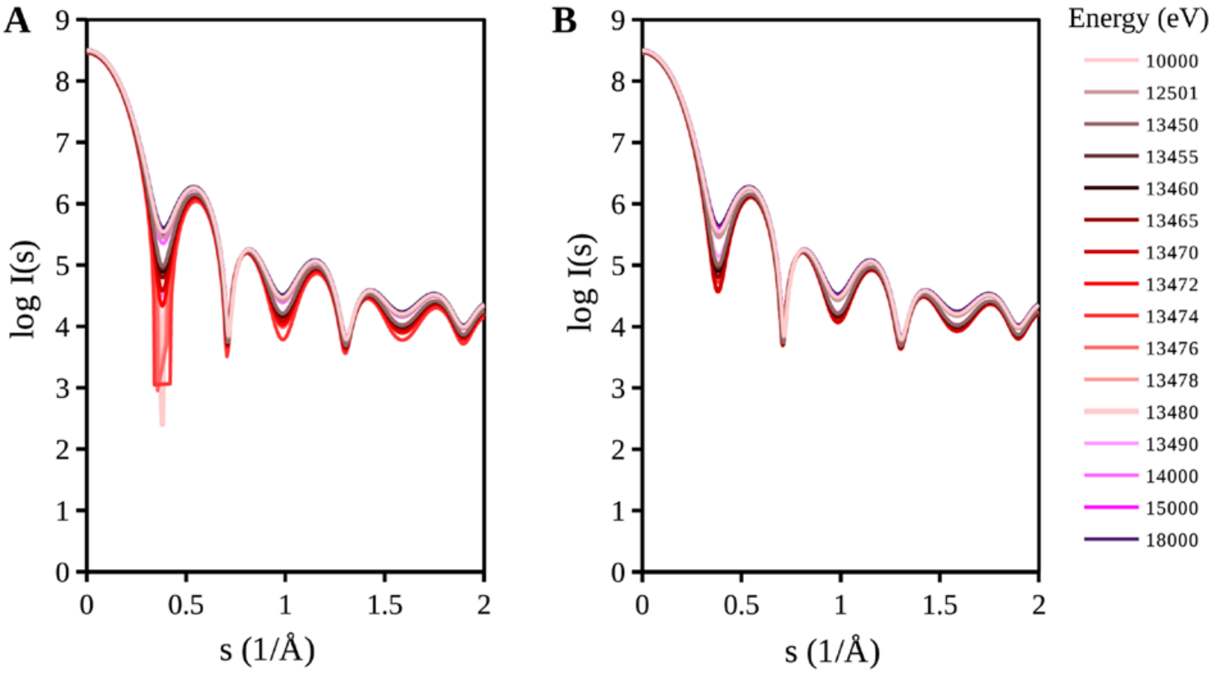

Figure 4-2. Simulated scattering data from a Br-C bead, computed with (A) the Debye formula, and (B) anomalous CRYSOL, at a range of energies at and around the Br K-absorption edge (13474 eV).

However, anomalous CRYSOL has some important advantages over the Debye formula. First, this particular test object consisted of exactly 2500 atoms, which is the inflection point at which a spherical harmonics implementation should be faster than the Debye formula. This was certainly observed during scattering data simulation, with anomalous CRYSOL with a run time in the order of minutes, and the Debye formula implementation in the order of hours. Granted, this dramatic difference might be due to a lack of optimization in the Debye formula implementation, but a speedup was nonetheless expected based on mathematical principles. The difference is even more stark considering that the scattering profile produced with anomalous CRYSOL was of tenfold higher resolution (5000 points) than the one produced with the Debye formula implementation (~500 points).

Another useful feature of anomalous CRYSOL is that the solvated scattering is also computed along with the *in vacuo* scattering. This is makes it particularly suited for simulating anomalous scattering from biomolecules, which is the intended use case for anomalous CRYSOL. In fact, there is at least one documented case of CRYSOL being utilized to compute anomalous scattering of biomolecules. In this case, the anomalous scattering due to iron in myoglobin was approximated by replacing the iron atom with an atom of the same effective form factor (i.e. reduced number of electrons), as iron at its K-absorption edge (Makowski *et al.*, 2012). This case of CRYSOL utilization suggests that there is an interest from the community, which along with in-house projects in EMBL Hamburg, would profit from the developed ASAXS module.

4.3 Using anomalous CRYSOL to compute scattering from biomolecules

The anomalous mode of CRYSOL was used to simulate anomalous scattering at and away from the absorption edge, for two historical biological ASAXS examples: (1) human hemoglobin with four bound iron atoms (Stuhrmann and Notbohm, 1981), and (2) rabbit parvalbumin with terbium atoms at the two calcium binding sites (Miake-Lye, Doniach and Hodgson, 1983).

In 1981, Stuhrmann and Notbohm derived the distances and tetrahedral geometry of the four bound iron atoms in human hemoglobin from ASAXS data. They predicted that the total contribution of the iron atoms would be very small, causing relative changes in scattering intensity in the order of 0.001-0.01. The anomalous scattering of solvated human hemoglobin was simulated from the PDB structure (PDB ID: 1a3n) using anomalous CRYSOL. The difference in scattering intensity measured away and at the K-absorption edge of iron is indeed quite subtle (Figure 4-3A), due to both the relative scarcity of iron atoms compared to the size of the protein (four iron atoms in a ~600 residue protein), and the low f' and f" anomalous correction factors at the K-absorption edge of iron. A couple of years later, Miake-Lye and colleagues did a similar experiment with a smaller protein parvalbumin, which binds calcium at two sites. They replaced the calcium atoms with terbium, which has the advantage of having an absorption edge with relatively high magnitude f' and f" correction factors, at an energy that is reachable by synchotron radiation (Figure 4-1). Expectedly, the difference between scattering intensity at and away from the absorption edge of terbium was quite pronounced. We observed a similar phenomenon upon simulating anomalous scattering of parvalbumin at similar energies (Figure 4-3B).

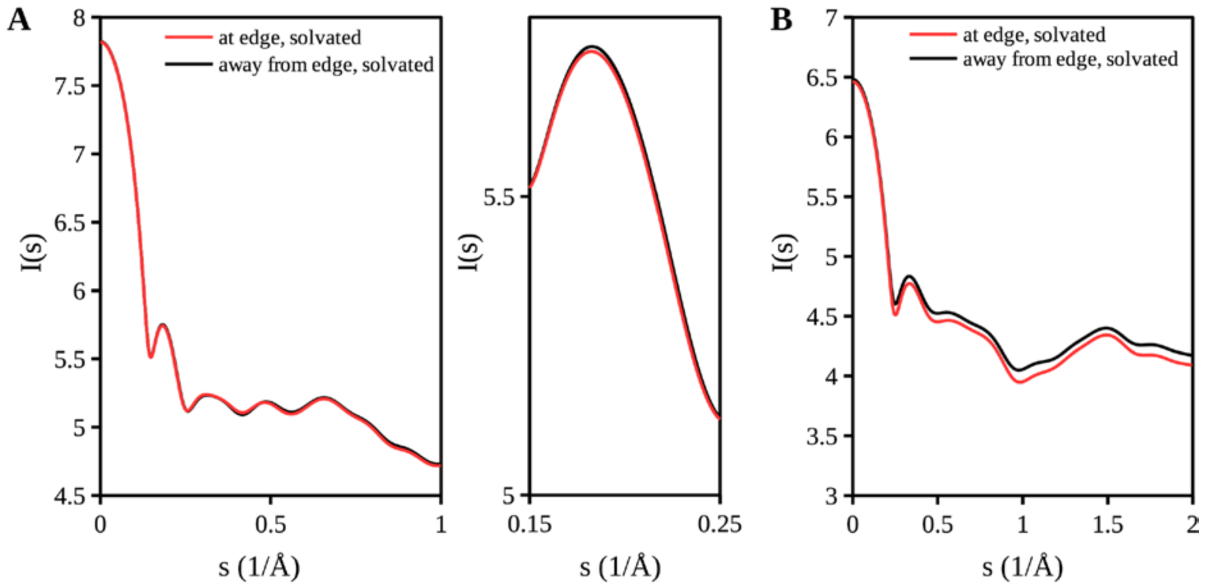

Figure 4-3. (A) The scattering of solvated hemoglobin (PDB ID: 1a3n), as computed with anomalous CRYSOL, both at and away from the Fe K-absorption edge (7115 eV). The overall contribution of the four Fe atoms to the scattering is small. A zoomed in version emphasizing the change is shown at the center panel. (B) The scattering of solvated rabbit parvalbumin (PDB ID: 1pal) with two bound terbium atoms in the calcium-binding sites, as computed with anomalous CRYSOL, both at and away from the Tb L-III absorption edge (7517 eV). The two Tb atoms have a large contribution to the resulting scattering, due to the relatively large f' and f" values at the L-III absorption edge.

These two cases show that it is possible to extract useful information from ASAXS data from biological systems, although one must be judicious about whether ASAXS is applicable to one's system of interest. ASAXS experiments are significantly longer than regular SAXS, since the beam has to be tuned to multiple wavelengths, but it is convenient to do if one already has samples and beamtime (one essentially gets additional information "for free", notwithstanding the time investment). As such, anomalous CRYSOL could serve as a useful tool for users to determine *a priori* whether ASAXS can be applied to solve their particular biological problem.

4.4 Conclusion and outlook

We have developed a module in CRYSOL allowing for the computation of anomalous scattering effects. Further developments could include making the simulated data more realistic by adding varying levels of noise. A systematic test could also be undertaken using the anomalous CRYSOL module to determine the required ratio of anomalous scatterers to size of the biomolecule to get a detectable signal, analogous to the Bijvoet ratio in X-ray crystallographic phasing. Experimental validation of the simulated ASAXS data would also be extremely useful.

As currently implemented, however, anomalous CRYSOL can be a useful sandbox for experimentalists to try out their anomalous scattering experiments *in silico*, before doing actual ASAXS experiments. The anomalous mode of CRYSOL is available in ATSAS v3.0.0.

5 Modeling the solution structure of *Plasmodium falciparum* Alu RNA and the Alu-SRP9/14 complex

In this and the following two sections, SAXS was used for the structural characterization of different macromolecular systems, including proteins, nucleic acids, and their complexes. The work described here stems from collaborative projects with user groups at the ESRF and the EMBL P12 SAXS beamline, where I was the responsible contact. The data analysis methods described in previous sections, in particular the NMA-based approaches for model refinement and coevolution analysis, were used wherever appropriate.

5.1 The signal recognition particle (SRP)

The signal recognition particle (SRP) is a universally-conserved RNA-protein complex that is involved in the transport of nascent proteins from the ribosome. The ~300nt RNA component of the eukaryotic SRP folds into two domains of about equal length: an Alu domain and an S domain. The Alu domain forms a complex with protein heterodimer SRP9/14. The Alu-SRP9/14 complex causes the retardation of protein translation, possibly through binding competitively to the elongation factor binding site on the ribosome (Wild and Sinning, 2014).

Previously, it was demonstrated that drug-induced disruption of the transport of SRP9 and SRP14 across the nuclear membrane prevented the formation of the SRP complex in *Plasmodium falciparum*, drastically reducing cell growth *in vitro* (Panchal *et al.*, 2014). As a result, the drug (ivermectin) is a candidate for an anti-malarial prophylactic (Metzger *et al.*, 2019). Thus, elucidating the molecular details of the Alu-SRP9/14 interaction in *P. falciparum* is of interest, both as a basic scientific question, and as a source of drug targets.

We analyzed solution SAXS data from the SRP9/14 heterodimer and the Alu domain of the SRP RNA from *Plasmodium falciparum* (*Pf*), for both the wildtype sequence (AluWT, 118 nucleotides) and several synthetic variants. Size-exclusion chromatography SAXS (SEC-SAXS) was also performed for each Alu variant in combination with SRP9/14 to check for complex formation in each case (Graewert *et al.*, 2015). Models of the solution structures of the *Pf* Alu RNA variants, SRP9/14 heterodimer, and the AluWT-SRP9/14 complex are proposed.

Table 5-1. Sample properties

	AluWT	AluRigid	AluH1	Alu106	Alu76	SRP9/14
Organism	*P. falciparum*					
Source	See Figure 1					Panchal et al., 2014
Molecule name	Alu, wild-type	Alu, rigid variant (synthetic)	Alu, with extended H1 helix (synthetic)	Alu, 106-nt truncation variant (synthetic)	Alu, 76-nt truncation variant (synthetic)	Signal recognition particle heterodimer, 9/14 kDa

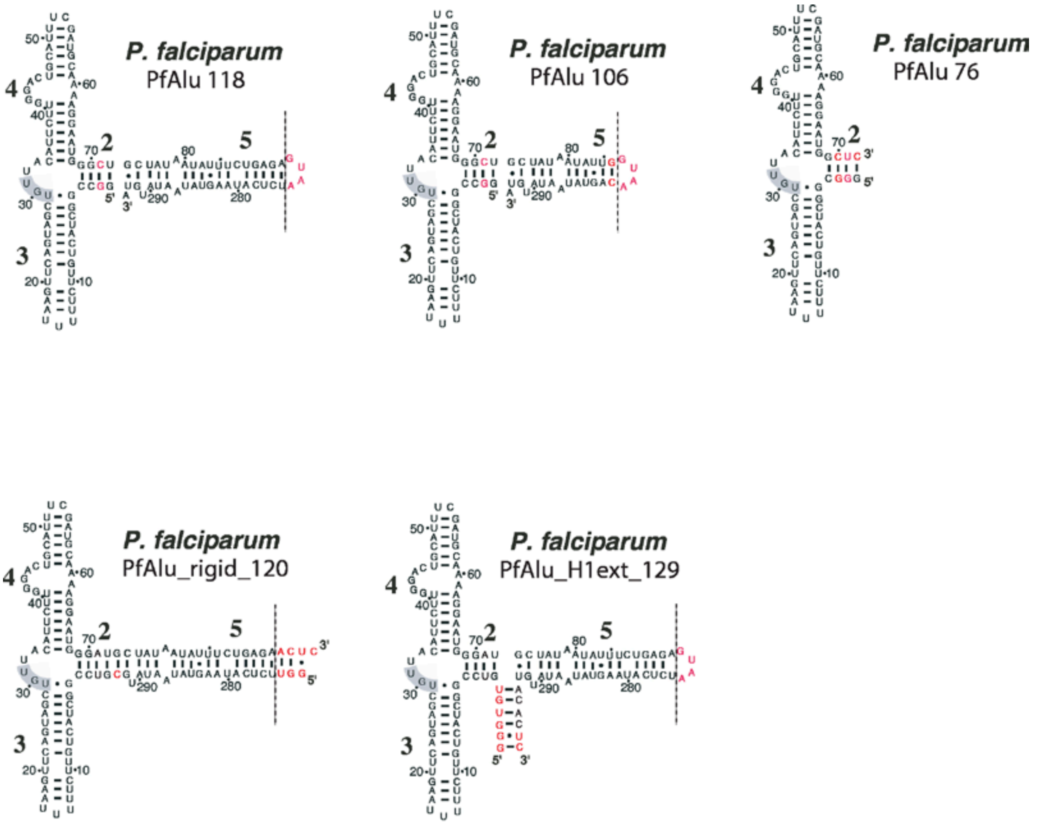

Figure 5-1. Pf Alu wild-type RNA (top left) and its length and flexibility variants.

5.2 Experimental procedures

5.2.1 SAXS data collection and processing

Pf SRP9/14 protein and *Pf* Alu RNA (wild-type sequence, and several length and flexibility variants) were produced and subjected to small-angle X-ray scattering at the SAXS beamline BM29 of the European Synchotron Radiation Facility (Grenoble) (Pernot *et al.*, 2013), by Komal Soni (Sinning group, Heidelberg University). Sample details are summarized in Table 5-1, while the sequence and secondary structure of each Alu RNA variant is shown in Figure 5-1. SAXS data were measured for each sample individually, and for each Alu variant in combination with SRP9/14. The Alu-SRP9/14 mixtures were subjected to size exclusion chromatography (SEC) upstream of the SAXS measurement, in order to separate the monomeric fractions from any complexes formed (Mathew, Mirza and Menhart, 2004). The radially-averaged time-course SEC-SAXS data were viewed with CHROMIXS (Panjkovich and Svergun, 2017), from where buffer and sample frames were manually selected. Data averaging and buffer subtraction were then done using Primus (Konarev *et al.*, 2003), to produce the scattering profile from the putative complex.

The analysis of the SAXS data was performed using the ATSAS 2.8 suite (Franke *et al.*, 2017). The concentration series SAXS data for each Alu RNA variant (Figure 5-2), and SRP9/14 (Figure 5-5, panels A and B), as well as the SEC-SAXS profiles of each Alu-SRP9/14 mixture (Figure 5-2), were assessed for the absence of aggregation and interparticle effects, by checking for a linear Guinier region. Based on these criteria, the scattering profiles from the following samples were selected: 2.8 mg/ml AluWT, 2.9 mg/ml AluRigid, 1.6 mg/ml AluH1, 2.6 mg/ml Alu106, 0.45 mg/ml Alu76, and 5 mg/ml SRP9/14.

For each scattering profile, the forward scattering $I(0)$ and the radius of gyration R_g were obtained from the Guinier approximation (Guinier, 1939), following the standard procedures (Konarev *et al.*, 2006). The distribution of pair distances $P(r)$ was computed using the indirect Fourier transformation method implemented in GNOM (Semenyuk and Svergun, 1991). From the $P(r)$ function, an alternative estimate of R_g and the maximum particle dimension D_{max} were obtained. Molecular weights (MW) in solution of the Alu variants were assessed from the SAXS data with two methods: (a) from $I(0)$, with the data adjusted to absolute scale (Jeffries *et al.*, 2016), (b) from the volume of correlation, V_c, which has a correction factor for RNA samples (Rambo and Tainer, 2013). For SRP9/14, MW estimates were derived from $I(0)$ using a bovine serum albumin standard, and through a consensus Bayesian MW assessment method (Hajizadeh *et al.*, 2018). Since no concentration data was available for the Alu-SRP SEC-SAXS data, only V_c-based MW estimation was performed on these scattering profiles. To adapt V_c-based MW estimation to the RNA-protein complexes, MW_{Vc} was computed assuming that the data represented pure protein or pure RNA. The resulting MW estimates were averaged and reported as the MW of the complex.

5.2.2 SAXS data analysis and structure modeling

The *ab initio* modeling program DAMMIF (Franke and Svergun, 2009) was used to produce low-resolution bead models of the Alu variants and SRP9/14 from their respective scattering profiles. Ten independent DAMMIF models were generated, superimposed with SUPCOMB (Kozin and Svergun, 2001), compared and averaged using DAMAVER (Volkov and Svergun, 2003), and the resolution computed with SASRES (Tuukkanen, Kleywegt and Svergun, 2016).

The theoretical scattering curve from the high-resolution structure of human SRP9/14 (PDB ID: 4uyk) was computed, and its χ^2 fit against the experimental SAXS data evaluated using CRYSOL (Svergun, Barberato and Koch, 1995). To obtain a model that better fits the experimental scattering data, missing loop regions were added to the human SRP9/14 structure using CORAL (Franke *et al.*, 2012). The resulting model was further refined against the SAXS data with Cartesian NMA using SREFLEX (Panjkovich and Svergun, 2016). Since there were no available high-resolution structures for any of the *Pf* Alu RNA variants, models of each were built from sequence using the MC-SYM pipeline (Parisien and Major, 2008). The resulting models were refined to fit the SAXS data using NMA in torsion angle space (approach discussed in Chapter 2). Complex formation between each Alu variant and SRP9/14 was probed from the SAXS data using OLIGOMER (Konarev *et al.*, 2003). OLIGOMER was used to approximate the SAXS data from the Alu-SRP9/14 mixtures as a sum of the monomer scattering profiles. Inability to fit the mixture data as a combination of monomers indicates the presence of another scattering species, which in this case is the Alu-SRP9/14 complex. Using this procedure, complex formation was detected for all Alu-SRP9/14 mixtures.

Ab initio models for each Alu-SRP complex were generated by multiphase modeling using MONSA (Svergun, 1999), wherein the RNA and protein are modeled as separate phases, and the model is built to fit the scattering data from the RNA, the protein, and the RNA-protein mixture simultaneously. The MONSA models of the different Alu-SRP complexes were used to identify a consensus binding site for SRP9/14 on the Alu RNA. Using the information from the Alu-SRP9/14 MONSA models, as well as contact information derived from the crystal structure of the complex between *Pyrococcus horikoshii* Alu RNA and human SRP9/14 (PDB ID: 4uyk) (Bousset *et al.*, 2014), hybrid models for *Pf* AluWT-SRP were constructed through constrained rigid-body modeling with SASREF (Petoukhov and Svergun, 2005).

5.3 Solution characteristics and models for *Pf* Alu RNA variants and *Pf* SRP9/14

The scattering profiles from the Alu RNA variants, both in unbound form and in combination with SRP9/14, are shown in Figure 5-3 (panels A and B), along with the OLIGOMER profile, which represents a combination of the component monomers. In each Alu-SRP9/14 mixture, the OLIGOMER profile does not adequately approximate the experimental data, suggesting complex formation between components. Guinier analysis of each Alu variant indicates that no concentration-dependent oligomerization is occurring (Figure 5-2, left panel insets), while the MW and Porod volume estimates suggest a monomeric state for each (Table 5-2). For each Alu-SRP scattering profile, the Porod volume and MW estimates (Table 5-3) suggest the formation of a 1:1 complex, which in the case of Alu76-SRP9/14, might be partially dissociated.

The $P(r)$ function for each Alu variant indicates a two-domain structure (seen from the shoulder in each plot) that disappears upon binding of the SRP9/14 protein (Figure 5-3, panels C and D). The Kratky plots (Figure 5-3, E and F) confirm this two-domain structure, and show that these two domains have a small amount of flexibility between them that is decreased by SRP9/14 binding.

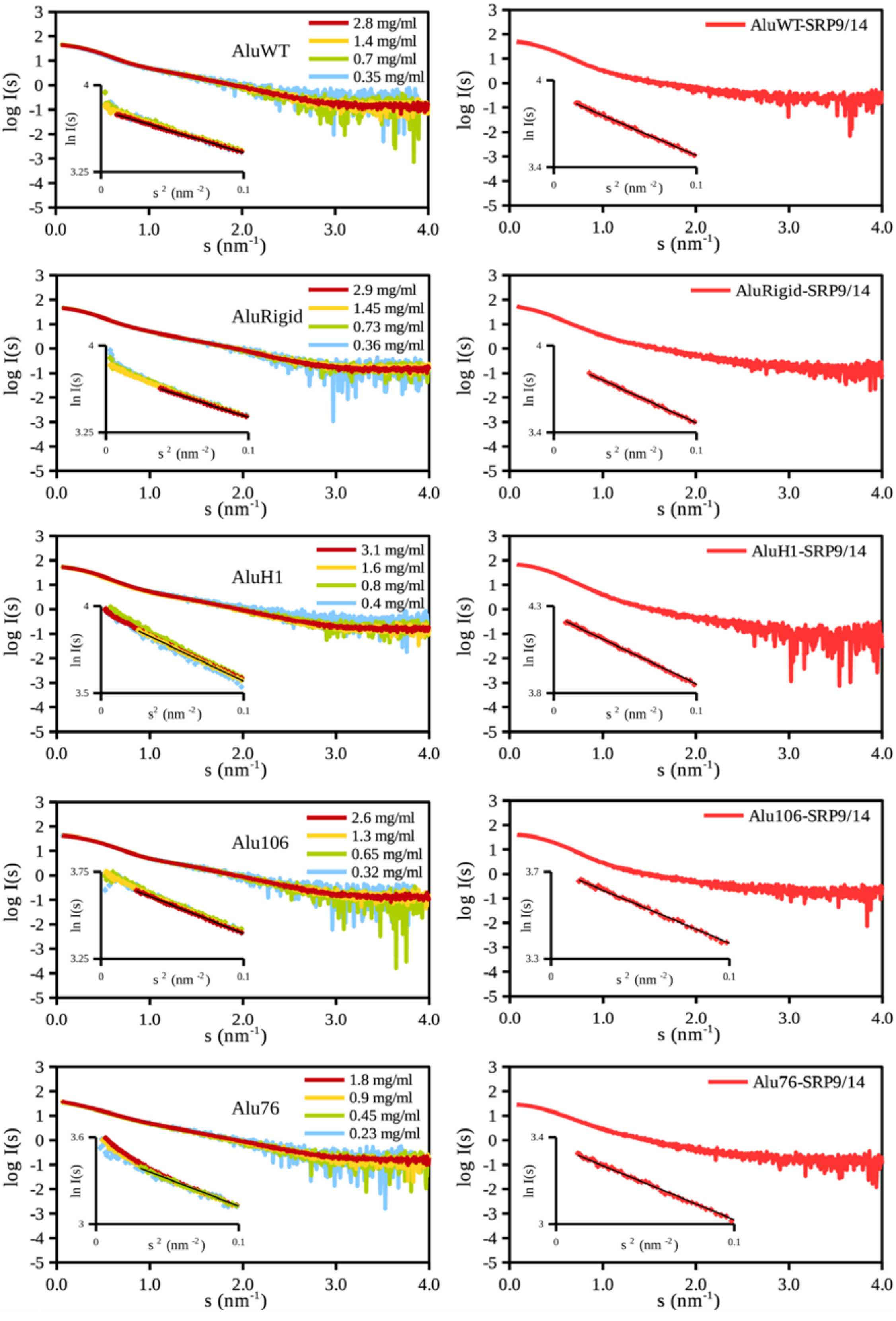

Figure 5-2. Concentration series data from Alu RNA variants (left) and SEC-SAXS data from each Alu-SRP complex (right), with Guinier plots for each scattering profile (inset).

Table 5-2. Data collection and structure statistics for small angle X-ray scattering analysis (monomers)

Data collection parameters	**AluWT**	**AluRigid**	**AluH1**	**Alu106**	**Alu76**	**SRP9/14**
Instrument	BM29 (ESRF, Grenoble)					
Beam geometry (mm^2)	0.7×0.7					
Wavelength (Å)	0.99					
s range ($Å^{-1}$)	0.003-0.5					
Exposure time (s)	5 (10×0.5s)					
Temperature (K)	293					
Concentration range measured ($mg\ ml^{-1}$)	0.4 – 2.8	0.4 – 2.9	0.4 – 3.1	0.3 – 2.6	0.2 – 1.8	0.3 – 5
Concentration used ($mg\ ml^{-1}$)	2.8	2.9	1.6	2.6	0.45	5
Structural parameters						
R_g (Å) (from $P(r)$)	33±1	35±1	34±1	31±1	33±1	21±1
R_g (Å) (from Guinier plot)	33±1	35±1	34±1	31±1	34±2	21±1
D_{max} (Å)	110±10	120±10	120±10	110±10	110±10	74±7
Porod volume estimate, V_p ($10^3Å^3$)	75	72	77	54	60	44
Excluded volume estimate[§] ($10^3Å^3$)	79	81	87	70	52	48
Molecular weight determination (kDa)						
From volume of correlation (V_c)	40±2	40±2	45±3	37±2	31±2	22±4
From Bayesian assessment	n.a.	n.a.	n.a.	n.a.	n.a.	19±2
From $I(0)$	28±6	49±9	44±8	38±7	40±8	23±5
Calculated monomeric *MW* from sequence	38	38	42	34	25	24
Software employed						
Primary data reduction	SASFLOW					
Data processing	PRIMUS					
Ab initio analysis	DAMMIF					
Validation and averaging	DAMAVER					
3D structure prediction	MC-SYM					
Flexibility modeling	NMATOR	NMATOR	NMATOR	NMATOR	NMATOR	SREFLEX
Computation of model intensities	CRYSOL					
3D graphics representations	PyMOL[+]					

[§] excluded volume calculations made with human SRP9/14, and the MC-SYM RNA structures, using Mol_volume, Version 1.0, Theoretical Biophysics Group, University of Illinois (retrieved from "http://www.ks.uiuc.edu/Development/MDTools/molvolume/")

[+] The PyMOL Molecular Graphics System, Version 1.7.2.1, Schrödinger, LLC

Table 5-3. Data collection and structure statistics for small angle X-ray scattering analysis (complexes)

Data collection parameters	**AluWT-SRP9/14**	**AluRigid-SRP9/14**	**AluH1-SRP9/14**	**Alu106-SRP9/14**	**Alu76-SRP9/14**
Instrument	BM29 (ESRF, Grenoble)				
Beam geometry (mm^2)	0.7×0.7				
Wavelength (Å)	0.99				
s range (Å^{-1})	0.003-0.5				
Exposure time (s)	1s ×2500 frames				
Temperature (K)	293				
Concentration range (mg ml^{-1})	unknown (SEC)				
Structural parameters					
R_g (Å) (from $P(r)$)	35±1	37±1	35±1	32±1	33±1
R_g (Å) (from Guinier plot)	36±1	37±1	34±1	32±1	32±1
D_{max} (Å)	120±10	125±10	120±10	110±10	110±10
Porod volume estimate, V_p (10^3Å^3)	121	118	131	112	77
Excluded volume estimate[§] (10^3Å^3)	127	129	135	118	100
Molecular weight determination (kDa)					
From volume of correlation (V_c)	60±6	62±6	77±9	59±6	28±4
Calculated monomeric *MW* from Sequence	62	63	66	58	49
Software employed					
Primary data reduction	SASFLOW	SASFLOW	SASFLOW	SASFLOW	SASFLOW
Data processing	CHROMIXS, PRIMUS	CHROMIXS, PRIMUS	CHROMIXS, PRIMUS	CHROMIXS, PRIMUS	CHROMIXS, PRIMUS
Ab initio analysis	MONSA	MONSA	MONSA	MONSA	MONSA
Detection of complex formation	OLIGOMER	OLIGOMER	OLIGOMER	OLIGOMER	OLIGOMER
Rigid body modelling	SASREF	n.a.	n.a.	n.a.	n.a.
Computation of model intensities	CRYSOL	CRYSOL	CRYSOL	CRYSOL	CRYSOL
3D graphics representations	PyMOL[+]	PyMOL[+]	PyMOL[+]	PyMOL[+]	PyMOL[+]

[§] excluded volume calculations made with human SRP9/14, and the MC-SYM RNA structures, using Mol_volume, Version 1.0, Theoretical Biophysics Group, University of Illinois (retrieved from "http://www.ks.uiuc.edu/Development/MDTools/molvolume/")

[+] The PyMOL Molecular Graphics System, Version 1.7.2.1, Schrödinger, LLC

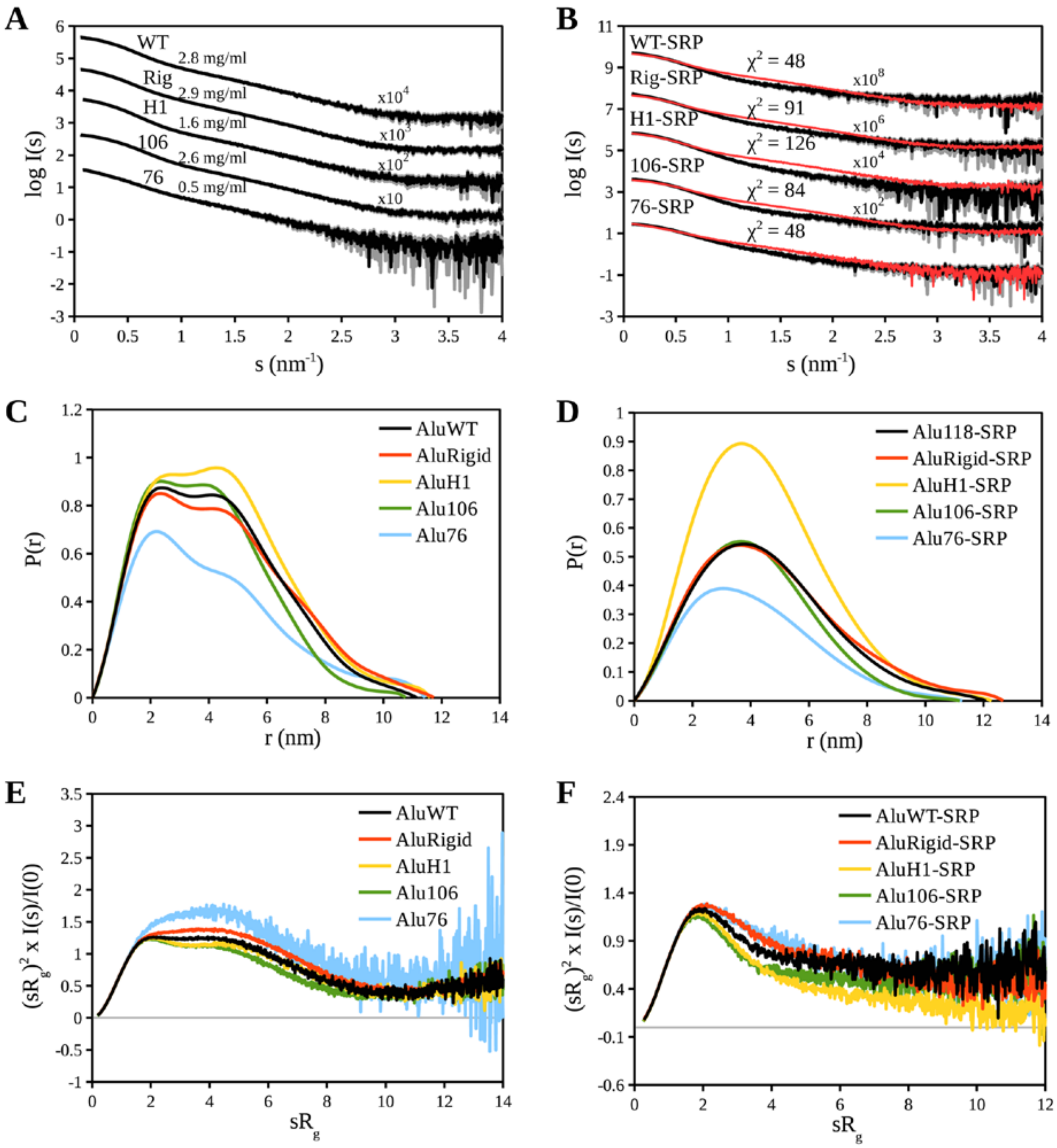

Figure 5-3. SAXS profiles, and SAXS-derived geometric and flexibility information for Alu RNA variants, with and without bound SRP9/14. (A and B) are the scattering profiles for each Alu variant without and with SRP9/14, respectively. The red overlay in B is the scattering profile from a combination of monomers that best fit the scattering data. The χ^2 fit is poor in each case, indicating complex formation. (C and D) show the pair distance distributions, $P(r)$, of each Alu variant without and with SRP9/14, respectively. There is a shoulder in the $P(r)$ function of each Alu variant, indicating a two-domain structure that disappears when SRP9/14 is added. (E and F) are the normalized Kratky plots, which show a similar two-domain structure with a small amount of flexibility between them, in the unbound Alu variants. Upon the addition of SRP9/14, this flexibility decreases, resulting in a relatively rigid Alu-SRP9/14 complex in each case.

The solution structure of each Alu variant was further characterized through *ab initio* and hybrid modeling, and the typical reconstructed models for each Alu variant are presented in Figure 5-4. Both bead and hybrid models were consistent with the two-domain architecture seen from the P(r) and Kratky plots of each Alu RNA variant. The bead models varied in whether this two-domain structure was represented as a hook shape or a branched shape. On the other hand, due to the constraints of secondary structure, the hybrid models from TNMA consistently modeled the Alu RNA variants as branched structures, with a long double helix present in all the variants, connected by a flexible junction to a shorter helix that varied in length for the truncation variants Alu76 and Alu106.

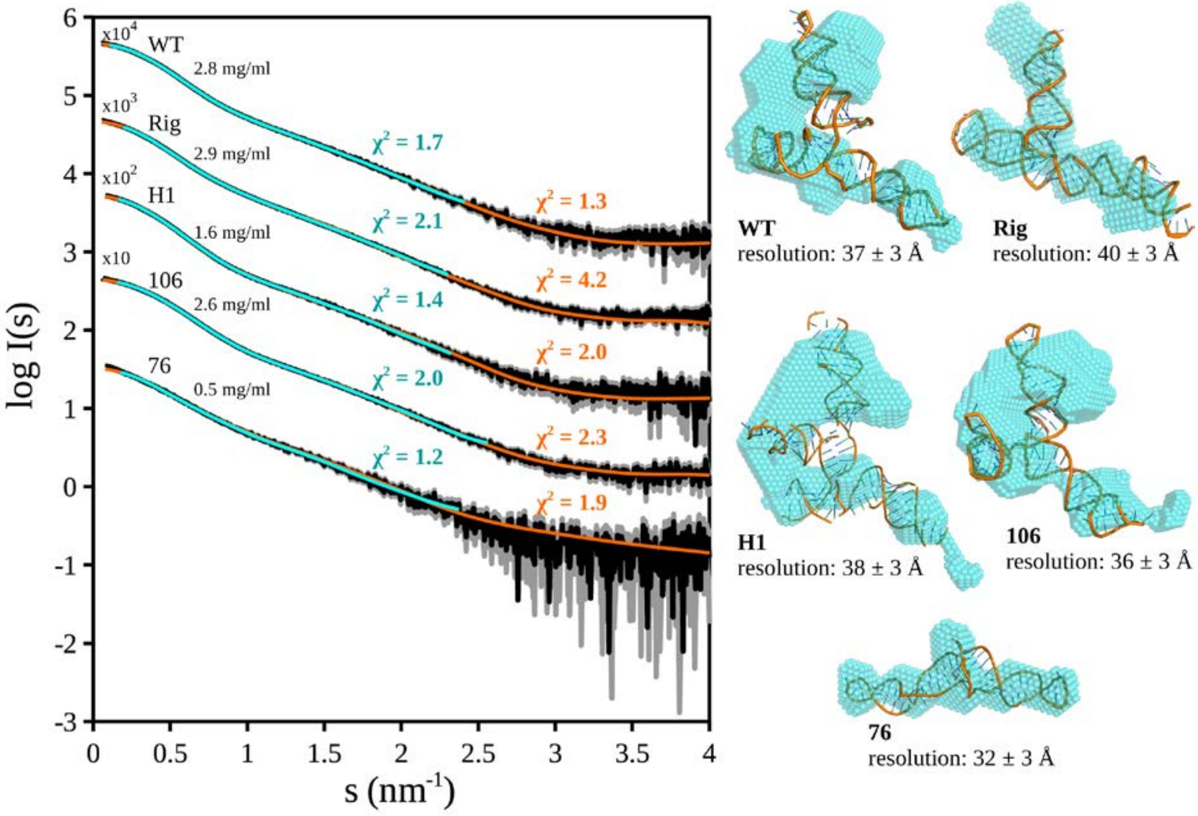

Figure 5-4. Bead and hybrid models of Alu RNA variants. For each, a typical DAMMIF model (cyan), and a hybrid model from MC-SYM-TNMA (orange) is shown, along with the respective model fits to the SAXS data. The resolution from ten independent ab initio reconstructions are shown below each bead model.

The properties and structure of SRP9/14 in solution was characterized from the SAXS data. No concentration-dependent effects were observed (Figure 5-5A), so the SAXS profile from the highest measured concentration was selected (Figure 5-5B). MW estimates indicated that SRP9/14 was not forming higher order structures in solution (Table 5-2). Based on the $P(r)$ function (Figure 5-5C) and Kratky plot (Figure 5-5D), the SRP9/14 heterodimer was observed to be globular and compact. Comparison of typical models from *ab initio* (Figure 5-5E) and hybrid modeling (Figure 5-5F) show that the bead model is roughly the same configuration as the high-resolution structure, save for some extra volume which could be due to the missing loop residues in the PDB structure. The addition of these missing residues improved the fit to the SAXS data, with subsequent CNMA only changing the loop conformations. The final *Pf* SRP9/14 hybrid model is mostly consistent with the human version, suggesting that the structure of the heterodimer is conserved between the two species.

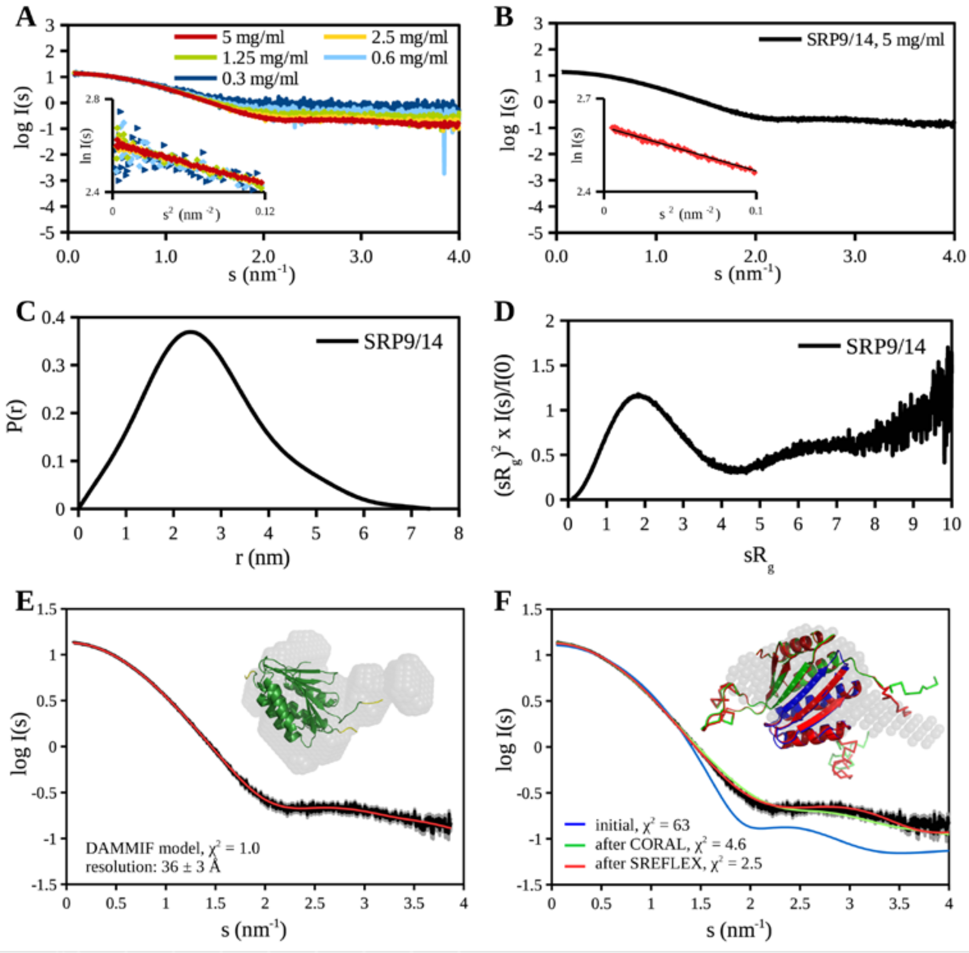

Figure 5-5. (A) Concentration series data from SRP9/14 shows no concentration effects, so (B) the SAXS profile from 5 mg/ml SRP9/14 was selected. (C) The $P(r)$ function and (D) normalized Kratky plot show that SRP9/14 is compact and globular. (E) A typical ab initio ***model (gray) has some volume unaccounted for in the high-resolution model, which is missing some loop regions. (F) Missing loops were filled in (green), after which the resulting structure was refined by CNMA (red). Model resolution and fits to the SAXS data and are indicated where appropriate.***

5.4 A consensus model of the AluWT-SRP9/14 interaction

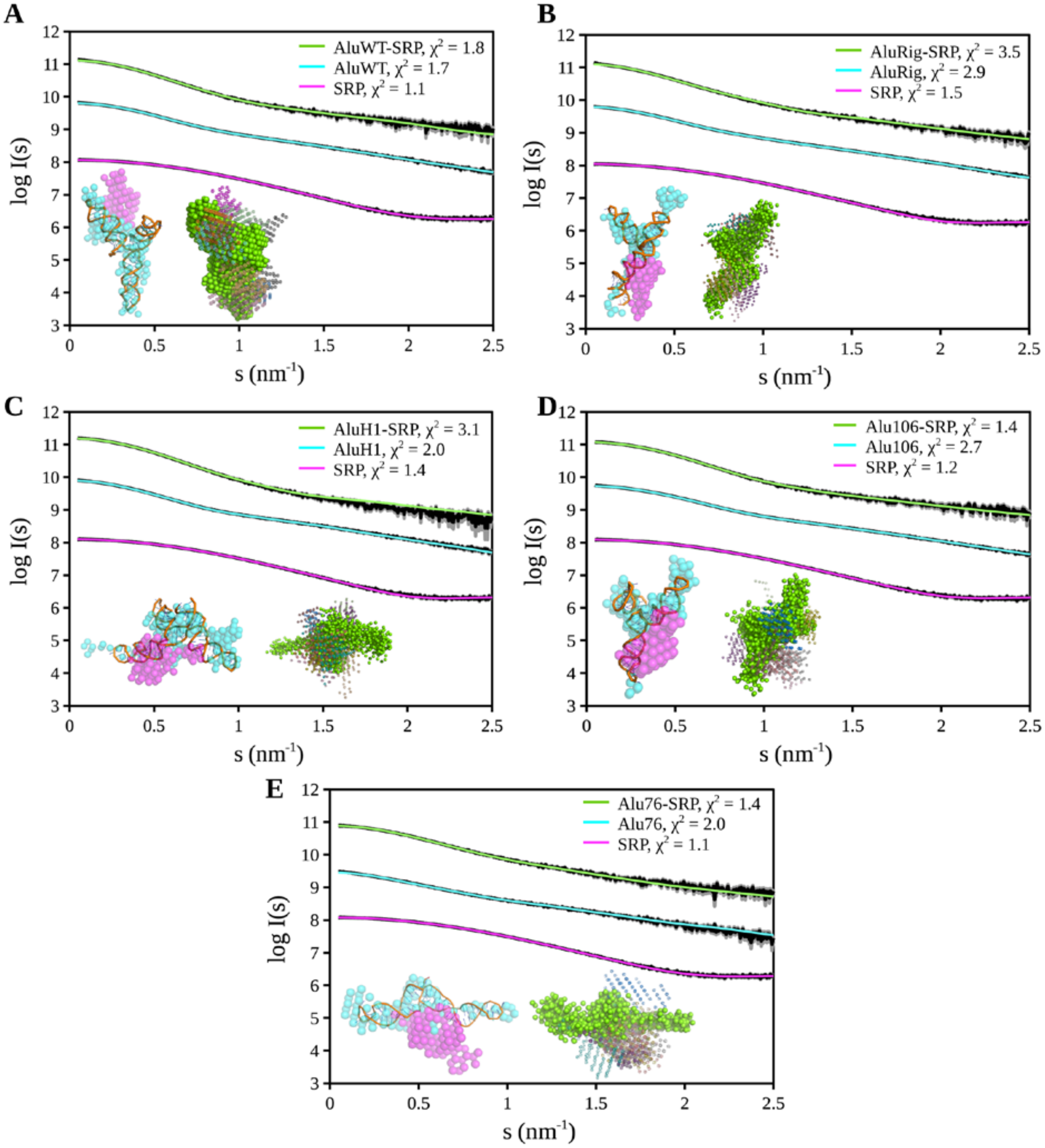

Figure 5-6. Ab initio MONSA models of the Alu-SRP9/14 complexes (A) AluWT-SRP9/14, (B) AluRigid-SRP9/14, (c) AluH1-SRP9/14, (d) Alu106-SRP9/14, and (e) Alu76-SRP9/14. In each case, the structure on the left shows the model with the best fit out of ten MONSA runs, with the RNA phase in cyan, and the protein phase in magenta. Each hybrid RNA model is superimposed on the RNA phase as comparison. The structure on the right shows all of the ten MONSA models, aligned with respect to the RNA phase, and shows the variability of the MONSA solutions obtained.

Ab initio models of each Alu-SRP9/14 complex were constructed by multiphase modeling using MONSA, in order to narrow down the docking site of the SRP protein. Figure 5-6 shows typical *ab initio* models of each Alu-SRP complex, with the hybrid model of the Alu RNA molecule optimally superimposed on the RNA phase. Although the models are still quite variable with respect to the SRP docking site, it can be seen that most of the MONSA models either have the SRP9/14 bound to the side of the long helix (AluRigid-SRP9/14, Figure 5-6B), or near the junction between the short and long helices (AluWT-SRP9/14, Figure 5-6A). Of these two possible SRP9/14 binding sites, only SRP9/14-binding near the junction would seem to have the observed effect of decreasing the flexibility of the Alu RNA upon binding, suggesting that there is a greater likelihood that the SRP9/14 heterodimer binds at the junction.

With the SRP9/14-binding site narrowed down, the next thing to consider was the orientation of the SRP9/14 protein. The crystal structure of the complex between *Pyrococcus horikoshii* Alu RNA and human SRP9/14 was used as a reference for possible contacts between *Pf* AluWT and *Pf* SRP (Figure 5-7).

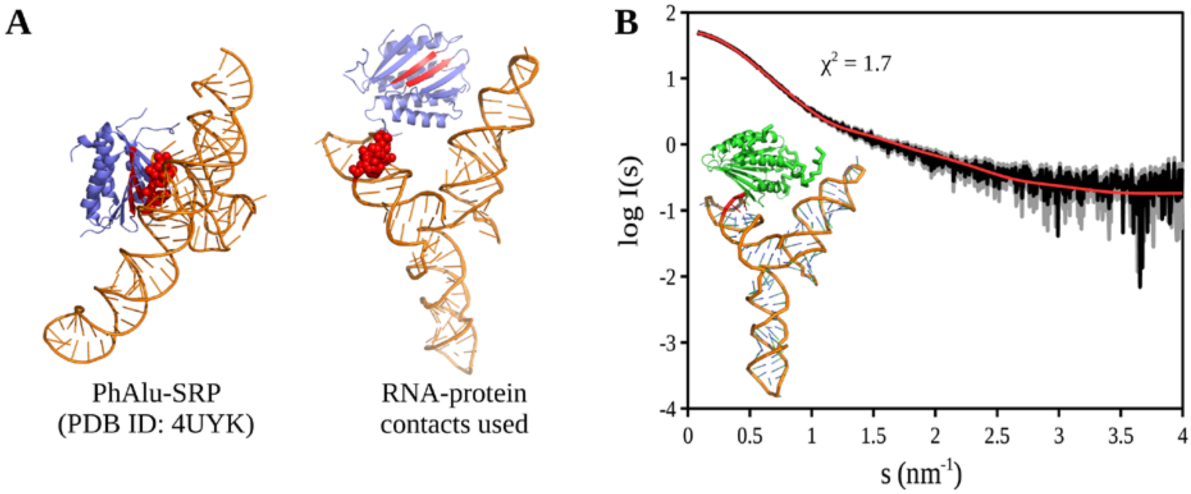

Figure 5-7. Rigid body modeling of the AluWT-SRP9/14 complex. (A) The crystal structure of* P. horikoshii *Alu RNA in complex with human SRP (PDB ID: 4uyk) was used to specify distance constraints for modeling. Highlighted in red are the most proximal regions between the RNA and protein. The same RNA motif was found for the P. falciparum AluWT model, and the regions used as distance constraints for SASREF are similarly shown in red. (B) shows the best-fitting AluWT-SRP model which meets the distance constraint, and in which the SRP also interacts with the short helix, which could account for the reduction in flexibility observed upon SRP binding.

At the binding interface of the PhAlu-SRP9/14 complex, it was observed that a cluster of basic residues on the beta-sheet region of the SRP9/14 heterodimer was interacting with a short UGU motif on the Alu RNA (Figure 5-6A) (Bousset *et al.*, 2014). This same UGU motif was seen near the hinge region of the AluWT model. This information was used to specify distance constraints for rigid body modeling with SASREF. Figure 5-7B shows a model with reasonable fit to the SAXS data, that shares a similar binding interface as the related crystallographic structure.

5.5 Conclusion and outlook

Solution SAXS studies on *P. falciparum* SRP9/14 protein, and Alu RNA (along with Alu RNA synthetic variants) have shown the Alu RNA likely forms a branched structure with some flexibility in solution. Upon binding of the SRP9/14 heterodimer, this branched structure disappears, along with its concomitant flexibility. Based on these observations, plausible branched models for the solution structures of the Alu RNA variants, were constructed using a combination of 3D structure prediction and TNMA. The AluWT-SRP9/14 complex was then built using distance constraints from a related crystallographic structure to guide SAXS-based rigid body modeling.

Additional validation experiments are recommended, to verify both the assumptions made to construct these models, as well as to check the accuracy of the models themselves.

6 Solution structures of condensin HEAT-repeat proteins Ycg1 and Ycs4 from *Chaetomium thermophilum*

Condensins are protein complexes that are involved in the segregation of eukaryotic chromosomes during mitotic and meiotic cell divisions (Houlard *et al.*, 2015; Uhlmann, 2016). Condensins consist of a dimer of Structural Maintenance of Chromosomes (SMC) subunits and a kleisin subunit Brn1 (Ivanov and Nasmyth, 2005; Cuylen, Metz and Haering, 2011; Wilhelm *et al.*, 2015), which recruits two additional subunits, Ycg1 and Ycs4, that are composed of tandem repeats of short, amphiphilic α-helices known as HEAT repeats (named after four proteins that contain this motif: **H**untingtin, **E**longation factor 3, the **A** subunit of protein phosphatase 2A (PP2A) and the signaling kinase **T**OR1) (Andrade and Bork, 1995; Neuwald and Hirano, 2000).

Ycg1 has previously been shown to have a horseshoe-shaped structure when bound to Brn1 in several co-crystal structures. The Ycg1-Brn1 complex was demonstrated to bind to double-stranded DNA via the formation of a positively-charged groove that contacts the negative charges in the DNA backbone, as well as the entrapment of the DNA helix within a flexible Brn1 loop, which thereby acts analogous to a safety-belt that pins the DNA double helix in place (Kschonsak *et al.*, 2017). Ycs4 has also had its high-resolution structure in complex with Brn1 elucidated, although its exact function in the condensin complex is less clear (Hassler *et al.*, 2019).

HEAT-repeat proteins have been shown to exhibit significant flexibility (Grinthal *et al.*, 2010; Kappel *et al.*, 2010), and solution scattering experiments were expected to validate whether this flexibility existed for both Ycg1 and Ycs4. Our hypothesis was that the binding of the Brn1 ligand was responsible for both recruiting the two HEAT-repeat proteins to the condensin complex, and also restricting the conformations of both to their functional form.

Here, we examined the structure and flexibility of the *Chaetomium thermophilum* (*Ct*) condensin subunits Ycg1 and Ycs4 in solution using SAXS, each in their unbound form as well as in complex with the kleisin Brn1. Both Ycg1 and Ycs4 were shown to be flexible in solution, with this flexibility restricted upon binding to Brn1. The average solution structures of unbound Ycg1 and Ycs4 were modeled, and seen to be significantly different from their conformations in the condensin complex, suggesting that the binding to Brn1 induces structural transitions to the functional conformation for both HEAT-repeat proteins. In addition, unbound Ycg1 was found to oligomerize in solution in the absence of Brn1, which might also have a role in regulating condensin function (Manalastas-Cantos *et al.*, 2019).

6.1 Experimental procedures

6.1.1 SAXS data collection and processing

Ct Ycg1 and Ycs4, alone or in complex with various ligands in the condensin complex were produced by Marc Kschonsak (Haering group, EMBL Heidelberg), as previously described (Haering *et al.*, 2017; Hassler *et al.*, 2019). Ycg1 and Ycs4 and their respective complexes are described in Table 6-1. SAXS data for each sample were collected at the SAXS beamline P12 of the PETRA III storage ring (Deutsches Elektronen-Synchrotron, Hamburg) (Blanchet *et al.*, 2015). The details of the data collection conditions are summarized in Table 6-2 and Table 6-3. The scattering data were collected with a PILATUS 2M pixel detector at a distance of 4.0 m from the sample. For each sample, solute concentrations ranging from 0.25 to 10 mg/ml were measured. The samples were loaded using an automatic sample changer, constantly flowed through the capillary during the X-ray exposure in order to minimize radiation damage. The two-dimensional pixel data from the detector were converted to one-dimensional scattering profiles using the automated pipeline SASFLOW, which performed radial averaging, outlier removal, data averaging, and buffer subtraction (Franke, Kikhney and Svergun, 2012).

Table 6-1. Sample properties

	Ycg1	**Ycg1–Brn1**	**Ycs4**	**Ycs4-Brn1$_{\text{short}}$**	**Ycs4-Brn1**	**Ycs4-Brn1-Smc4**
Organism	*C. thermophilum*					
Source (reference)	Kschonsak et al., 2017			Hassler et al., 2019		
Protein name (residues in construct)	Ycg1$_{24\text{-}1006}$	Ycg1$_{24\text{-}1006}$ in complex with Brn1$_{515\text{-}634}$	Ycg1$_{3\text{-}1222}$	Ycs4$_{3\text{-}1222}$ in complex with Brn1$_{336\text{-}418}$	Ycs4$_{3\text{-}1222}$ in complex with Brn1$_{225\text{-}418}$	Ycs4$_{3\text{-}1222}$ in complex with Brn1$_{225\text{-}418}$ and Smc4$_{263\text{-}1542}$
Solvent	300 mM NaCl, 25 mM tris-HCl pH 7.5, 1 mM DTT					

The analysis of the SAXS data was performed using the ATSAS 2.8 suite (Franke *et al.*, 2017). The concentration series SAXS data for each sample were assessed for the absence of aggregation and concentration effects, by checking for linearity in the Guinier region (Figure 6-1). For Ycg1, noticeable concentration effects were observed and the scattering data from 0.25 and 0.5 mg/ml were extrapolated to zero concentration using Primus (Konarev *et al.*, 2003). Only minor concentration effects were observed for the other samples, which could be ameliorated by data merging. A composite scattering profile for Ycg1–Brn1 was generated by merging the low-angle scattering at 0.5 mg/ml and high-angle scattering at 5 mg/ml. For Ycs4, a composite scattering curve was produced by merging low-angle scattering from 0.5 mg/ml and high-angle scattering from 10 mg/ml. For the Ycs4 complexes, monodispersity was validated by performing singular value decomposition (SVD) (Konarev *et al.*, 2003) on buffer-subtracted concentration series SAXS data from Ycs4-Brn1$_{\text{short}}$, Ycs4-Brn1, and Ycs4-Brn1-Smc4. Based on SVD, composite scattering curves for both the Ycs4-Brn1$_{\text{short}}$ and the Ycs4-Brn1 complexes were generated by merging low angle data from 0.5 mg/ml with high angle data from 5 mg/ml. For the Ycs4-Brn1-Smc4 complex, the data at 0.5mg/ml was computed to consist of only one component, and was selected for further analysis.

All relevant, either derived or selected scattering profiles, were used for further analysis and modeling. The $I(0)$ and R_g were obtained from the Guinier approximation (Guinier, 1939), following the standard procedures (Konarev *et al.*, 2006). The $P(r)$ function was computed using the indirect Fourier transformation method implemented in GNOM (Semenyuk and Svergun, 1991). From the $P(r)$ function, alternative estimates of R_g and D_{max} were obtained. Molecular weights in solution were assessed from the SAXS data with three methods: (a) using $I(0)$ and comparing against a reference solution of bovine serum albumin (Jeffries *et al.*, 2016), (b) from the excluded (Porod) volume V_p (given that V_p in nm^3 is about 1.6 times the MW in kDa) (Franke *et al.*, 2012), and (c) a consensus Bayesian MW assessment method (Hajizadeh *et al.*, 2018).

6.1.2 Molecular weight assessment of Ycg1 oligomers with light scattering

The oligomeric states of Ycg1 were analyzed by analytical size-exclusion chromatography (SEC), coupled to a multiangle static light scattering (MALS) detector. SEC was performed using an Agilent 1260 Infinity II Bio-inert LC system, and an analytical Superdex200 10/300 GL column (GE Healthcare) equilibrated with the sample buffer (25 mM tris, 300 mM NaCl, 1 mM DTT, pH 7.5) at 20°C. Seven microliters of Ycg1 at 10 mg/ml was injected, with the experiment performed at a flow rate of 0.8 ml/min. Protein elution was detected by absorbance at 280 nm, and protein concentration quantified with differential refractometry using an Optilab T-rEX detector (Wyatt). Light scattering data was measured with a miniDAWN TREOS multiangle light scattering detector (Wyatt). Molecular weights were computed from the refractometry and light scattering data using the software ASTRA version 7.1.3.15 (Wyatt)

Batch dynamic light scattering (DLS) measurements were also performed for Ycg1 at concentrations 0.6, 1.4, 2.5, 5.5, 10.8, and 22.1 mg/ml with a DynaPro NanoStar DLS detector (Wyatt). Light scattering data collection and analysis was performed with the software DYNAMICS version 7.6.0 (Wyatt). For each concentration, ten 5-second acquisitions were performed.

6.1.3 SAXS data analysis and structure modeling

For Ycg1 and Ycg1-Brn1, the *ab initio* modeling program DAMMIF (Franke and Svergun, 2009) was used to produce low-resolution bead models from the SAXS data. Ten independent DAMMIF models were generated, superimposed with SUPCOMB (Kozin and Svergun, 2001), compared and averaged using DAMAVER (Volkov and Svergun, 2003) , and the resolution computed with SASRES (Tuukkanen, Kleywegt and Svergun, 2016).

For all samples, the theoretical scattering curves from available high-resolution models were computed, and their χ^2 fits against the experimental SAXS data evaluated using

CRYSOL (Svergun, Barberato and Koch, 1995). For Ycg1 and Ycg1-Brn1, the crystal structure of *Saccharomyces cerevisiae* Ycg1-Brn1 (PDB ID: 5oqq) was used. For Ycs4 and its complexes, models were derived from the *C. thermophilum* Ycs4-Brn1-Smc4 crystal structure (PDB ID: 6qj4).

To obtain models that better fit the experimental scattering data, hybrid modeling was performed for Ycg1, Ycs4, and their respective complexes. Normal mode analysis in Cartesian space (CNMA) was performed on Ycg1 and Ycg1-Brn1 using the program SREFLEX (Panjkovich and Svergun, 2016). Torsional NMA (TNMA) was also performed for Ycg1 as comparison, and yielded similar results to CNMA. For the Ycs4 complexes, CORAL was first used to build the missing loop regions in the crystal structure, keeping the high-resolution structure fragments fixed in space (Franke *et al.*, 2012). To build the Ycs4-Brn1$_{short}$ model, the loop regions were built such that the resulting structure fit the SAXS data from both Ycs4 and Ycs4-Brn1$_{short}$ simultaneously. Similarly, for the Ycs4-Brn1-Smc4 complex, three SAXS datasets (Ycs4, Ycs4-Brn1, and Ycs4-Brn1-Smc4) were used simultaneously. The Ycs4-Brn1$_{short}$ model was further refined with NMA to fit the scattering data.

As the unbound Ycg1 exhibited signs of oligomerization, a dimer structure and its proportion at elevated concentrations in solution was further modeled using SASREFMX (Petoukhov and Svergun, 2005; Franke *et al.*, 2012; Petoukhov *et al.*, 2013). The SAXS data for Ycg1 at 5 and 10 mg/ml were further modeled with SASREFMX as a mixture of monomers, dimers and tetramers, with the tetrameric structure built as a dimer of dimers.

The results for Ycg1 and Ycg1-Brn1 have been published (Manalastas-Cantos *et al.*, 2019). The experimental SAXS data and the models for Ycg1 and Ycg1-Brn1 were also deposited on SASBDB (Small Angle Scattering Biological Data Bank; accession numbers SASDFC4 [Ycg1 monomer], SASDFD4 [Ycg1-Brn1 monomer], SASDFE4 [Ycg1 tetramer], SASDFG4 [Ycg1 dimer], and SASDFF4, SASDFH4, SASDFJ4, SASDFK4 [Ycg1 concentration series]) (Valentini *et al.*, 2014).

Table 6-2. Data collection and structure statistics for small angle X-ray scattering analysis

Data collection parameters	**Ycg1**	**Ycg1- Brn1$_{515\text{-}634}$**
Instrument	EMBL P12 (PETRA III, DESY, Hamburg)	
Beam geometry (mm^2)	0.2×0.12	
Wavelength (Å)	1.24	
s range (Å^{-1})	0.002-0.38	
Exposure time (s)	1 (20×0.05s)	
Temperature (K)	283	
Concentration range measured (mg ml^{-1})	0.25–10	
Concentration used (mg ml^{-1})	0 (extrapolated)	0.5 and 5 (merged)
Structural parameters		
R_g (Å) (from $P(r)$)	46±1	42±1
R_g (Å) (from Guinier plot)	46±1	43±1
D_{max} (Å)	160±10	140±10
Porod volume estimate, V_p (10^3Å^3)	190	230
Excluded volume estimate[§] (10^3Å^3)	186	205
Molecular weight determination (kDa)		
From Porod volume (V_p/~1.6)	119±24	144±29
From consensus Bayesian assessment	109 ±11	138±15
From $I(0)$	102±9	122±10
Calculated monomeric *MW* from Sequence	109	125
Software employed		
Primary data reduction	SASFLOW	SASFLOW
Data processing	PRIMUS	PRIMUS
Ab initio analysis	DAMMIF	DAMMIF
Validation and averaging	DAMAVER	DAMAVER
Rigid body modelling	SASREFMX	n.a.
Flexibility modeling	SREFLEX, NMATOR	SREFLEX
Computation of model intensities	CRYSOL	CRYSOL
3D graphics representations	PyMOL[+]	PyMOL[+]

[§] excluded volume calculations made with the crystal structures, using Mol_volume, Version 1.0, Theoretical Biophysics Group, University of Illinois (retrieved from "http://www.ks.uiuc.edu/Development/MDTools/molvolume/")

[+] The PyMOL Molecular Graphics System, Version 1.7.2.1, Schrödinger, LLC

Table 6-3. Data collection and structure statistics for small angle X-ray scattering analysis (continued)

Data collection parameters	**Ycs4**	**Ycs4-Brn1$_{short}$**	**Ycs4- Brn1**	**Ycs4- Brn1-Smc4**
Instrument	EMBL P12 (PETRA III, DESY, Hamburg)			
Beam geometry (mm^2)	0.2×0.12			
Wavelength (Å)	1.24			
s range (Å^{-1})	0.003-0.51			
Exposure time (s)	1 (20×0.05s)			
Temperature (K)	293			
Concentration range measured (mg ml^{-1})	0.25–10			
Concentration used (mg ml^{-1})	0.5 and 10 (merged)	0.5 and 5 (merged)	0.5 and 5 (merged)	0.5
Structural parameters				
R_g (Å) (from $P(r)$)	54±1	50±1	50±1	52±1
R_g (Å) (from Guinier plot)	52±2	50±1	49±2	51±2
D_{max} (Å)	180±10	170±10	160±10	180±10
Porod volume estimate, V_p (10^3Å^3)	285	286	308	341
Excluded volume estimate§ (10^3Å^3)	206	210	219	301
Molecular weight determination (kDa)				
From Porod volume (V_p/~1.6)	178±36	179±36	193±39	213±43
From consensus Bayesian assessment	159±17	173±22	192±29	199±22
From *I*(0)	142±14	150±15	167±17	197±20
Calculated monomeric *MW* from Sequence	138	147	159	219
Software employed				
Primary data reduction	SASFLOW	SASFLOW	SASFLOW	SASFLOW
Data processing	PRIMUS	PRIMUS	PRIMUS	PRIMUS
Rigid body modelling	CORAL	CORAL	CORAL	CORAL
Flexibility modeling	SREFLEX	SREFLEX	n.a.	n.a.
Computation of model intensities	CRYSOL	CRYSOL	CRYSOL	CRYSOL
3D graphics representations	PyMOL+	PyMOL+	PyMOL+	PyMOL+

§excluded volume calculations made with the crystal structures, using Mol_volume, Version 1.0, Theoretical Biophysics Group, University of Illinois (retrieved from "http://www.ks.uiuc.edu/Development/MDTools/molvolume/")

+The PyMOL Molecular Graphics System, Version 1.7.2.1, Schrödinger, LLC

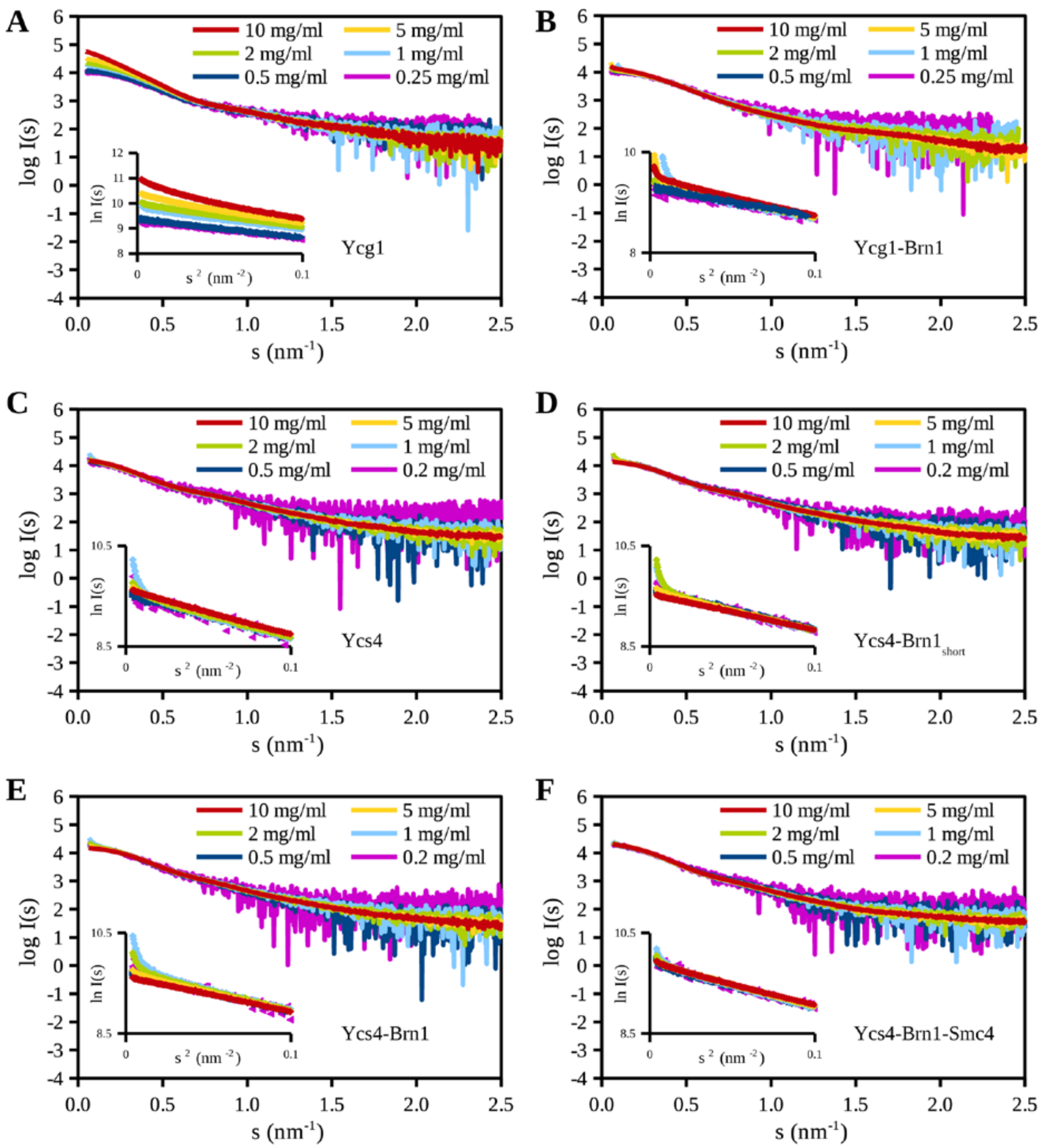

Figure 6-1. Concentration series data and corresponding Guinier plots (lower left inset) of (A) Ycg1, (B) Ycg1-Brn1, (C) Ycs4, (D) Ycs4-Brn1$_{short}$, (E) Ycs4-Brn1, and (F) Ycs4-Brn1-Smc4. The slope of the Guinier plots (and hence the R_g) for Ycg1 increased systematically with increasing concentration, indicating concentration-dependent oligomerization. For the rest of the samples, only minimal concentration effects were observed.

6.2 The effect of ligand binding on Ycg1 and Ycs4 structure

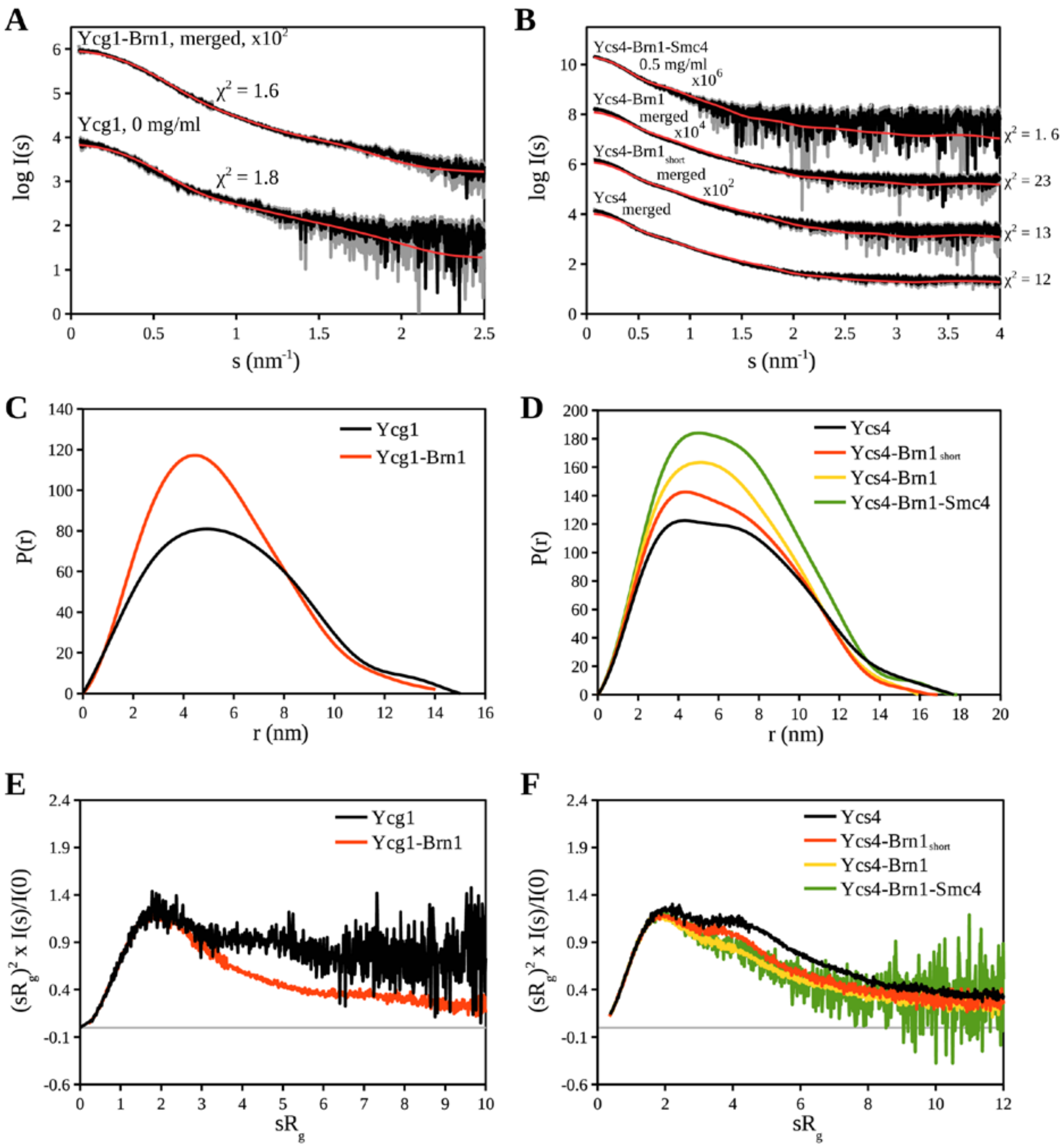

Figure 6-2. SAXS profiles, and SAXS-derived geometric and flexibility information for HEAT-repeat proteins Ycg1 and Ycs4, with and without additional ligands. (A and B) are the scattering profiles for the Ycg1 and Ycs4 samples, respectively. The red overlay in each case indicates the computed scattering profiles of the existing crystallographic structures. (C and D) show the pair distance distributions, $P(r)$, of the Ycg1 and Ycs4 samples, respectively. The unbound forms of both Ycg1 and Ycs4 appear to have the greatest maximum dimension (D_{max}) compared to their ligand-bound forms, possibly indicating that the unbound forms have a more open conformation. In addition, Ycs4 and Ycs4-Brn1$_{short}$ have a shoulder in their $P(r)$ functions, indicating a two-domain structure. (E and F) are the dimensionless Kratky plots, which show that unbound Ycg1, Ycs4, and Ycs4-Brn1$_{short}$ have some flexibility which is reduced with the binding of additional ligands.

The scattering profiles from the HEAT-repeat proteins Ycg1 and Ycs4, both in unbound form and with various ligands from the condensin complex, are shown in Figure 6-2 (panels A and B), along with the computed scattering of their corresponding crystallographic structures. Except for the full complexes, there was poor agreement between the experimental and computed scattering data for both Ycg1 and Ycs4, indicating that the less ligand-bound forms of these HEAT-repeat proteins might have a significantly different conformation than what is found in the crystal. The pair distance distributions of Ycg1 and Ycs4 (Figure 6-2, panels C and D) show that the maximum dimension of the unbound HEAT-repeat proteins are either comparable or slightly larger than their ligand-bound counterparts, suggesting that unbound Ycg1 and Ycs4 might have a more extended conformation in solution, compared to the conformation in the crystal structure. The normalized Kratky plots (Figure 6-2, panels E and F) show significant flexibility for Ycg1, and a moderate amount of hinge-like flexibility for Ycs4 and Ycs4-Brn1$_{short}$, which suggests that this conformational difference might be due to the inherent flexibility of the two HEAT-repeat proteins in the absence of a ligand.

The MW estimates from $I(0)$ confirm a monomeric form for both Ycg1 and Ycs4 (Table 6-2 and Table 6-3), indicating that the difference in scattering is due to a difference in conformation, and not due to polydispersity. MW estimates from $I(0)$ for Ycg1-Brn1, Ycs4-Brn1$_{short}$, and Ycs4-Brn1 also indicated a 1:1 stoichiometry, and the Ycs4-Brn1-Smc4 complex a 1:1:1 stoichiometry, as was expected based on the crystal structures. In order to characterize the conformational differences between free and ligand-bound HEAT-repeat proteins, solution structure models were constructed from the SAXS data. Typical bead and hybrid models of Ycg1 and Ycg1-Brn1 are shown in Figure 6-3.

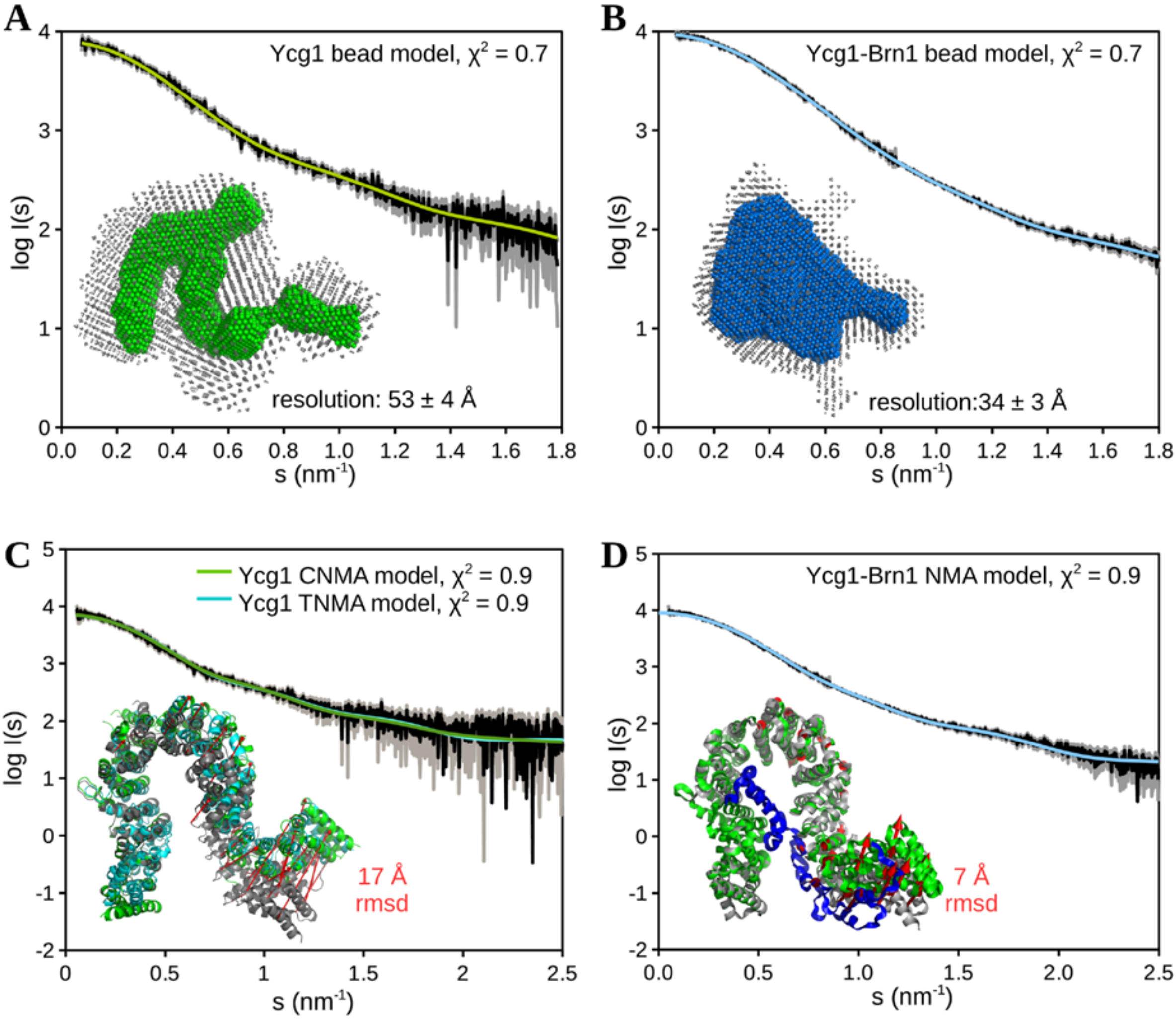

Figure 6-3. Ab initio models of (A) Ycg1, and (B) Ycg1–Brn1. A typical DAMMIF model is shown on the lower left corner, with the combined envelopes from 10 DAMMIF runs overlaid in gray. Ycg1 bead models appear elongated compared with bead models of Ycg1–Brn1. NMA models of (C) Ycg1, and (D) Ycg1–Brn1 show similar features to the bead models. Initial Ycg1 structures are shown in gray, with the Brn1 peptide shown in blue. Ycg1 structures after CNMA are shown in green, TNMA in cyan. Ycg1 CNMA and TNMA models are very similar. Red arrows on the initial structures depict the movement of the domains after CNMA. Ycg1 has a much larger rmsd (17 Å) than Ycg1–Brn1$_{515\text{-}634}$ (7 Å), which could be attributed to an increased flexibility in the absence of the ligand. (Figure edited from Manalastas-Cantos et al., 2019)

Both the bead and hybrid models show that Ycg1 solution structure is more elongated and open than Ycg1-Brn1. SAXS-guided NMA of the Ycg1-Brn1 crystal structure resulted in a net movement of only 7 Å, which is not a significant change at low resolution. On the other hand, both Cartesian and torsional NMA refinement of the Ycg1 crystal conformation resulted in a larger swivel movement of around 17 Å. This suggests that Brn1 binding corrals the protein into its functional, horseshoe shaped conformation that is capable of binding double-stranded DNA. Interestingly, Ycg1 by itself has previously been shown to have much lower DNA binding activity than the Ycg1-Brn1 complex (Kschonsak *et al.*, 2017). Conformational selection and restriction through Brn1 binding might be a mechanism behind this observed phenomenon.

SAXS-based hybrid models were also built for Ycs4-Brn1$_{short}$ and Ycs4-Brn1-Smc4. The Ycs4-Brn1$_{short}$ and Ycs4-Brn1-Smc4 models are shown in Figure 6-4. The Ycs4-Brn1$_{short}$ model (Figure 6-4A) the fit the scattering data from both Ycs4 and Ycs4-Brn1$_{short}$, indicating that the binding of the truncated Brn1 ligand did not significantly affect Ycs4 structure and flexibility. On the other hand, the Ycs4-Brn1-Smc4 model (Figure 6-3B) only fit the SAXS data from Ycs4-Brn1 and Ycs4-Brn1-Smc4, but fit the SAXS data from unbound Ycs4 poorly. This indicates a significant conformational difference between free and Brn1-bound Ycs4, which was quantified as a 9.5 Å swivel motion of one side of the structure in the unbound Ycs4 structure (Figure 6-3C). Thus, similar to Ycg1, Ycs4 undergoes conformational restriction to a more closed structure upon Brn1 binding. The functional significance of this closed Ycs4 structure is still unclear, but it has been proposed that the Ycs4-Brn1-Smc4 complex forms a compartment for DNA-binding, similar to Ycg1-Brn1. The timing of DNA association and dissociation may allow the condensin molecule to "walk" along the DNA in one direction (Hassler *et al.*, 2019).

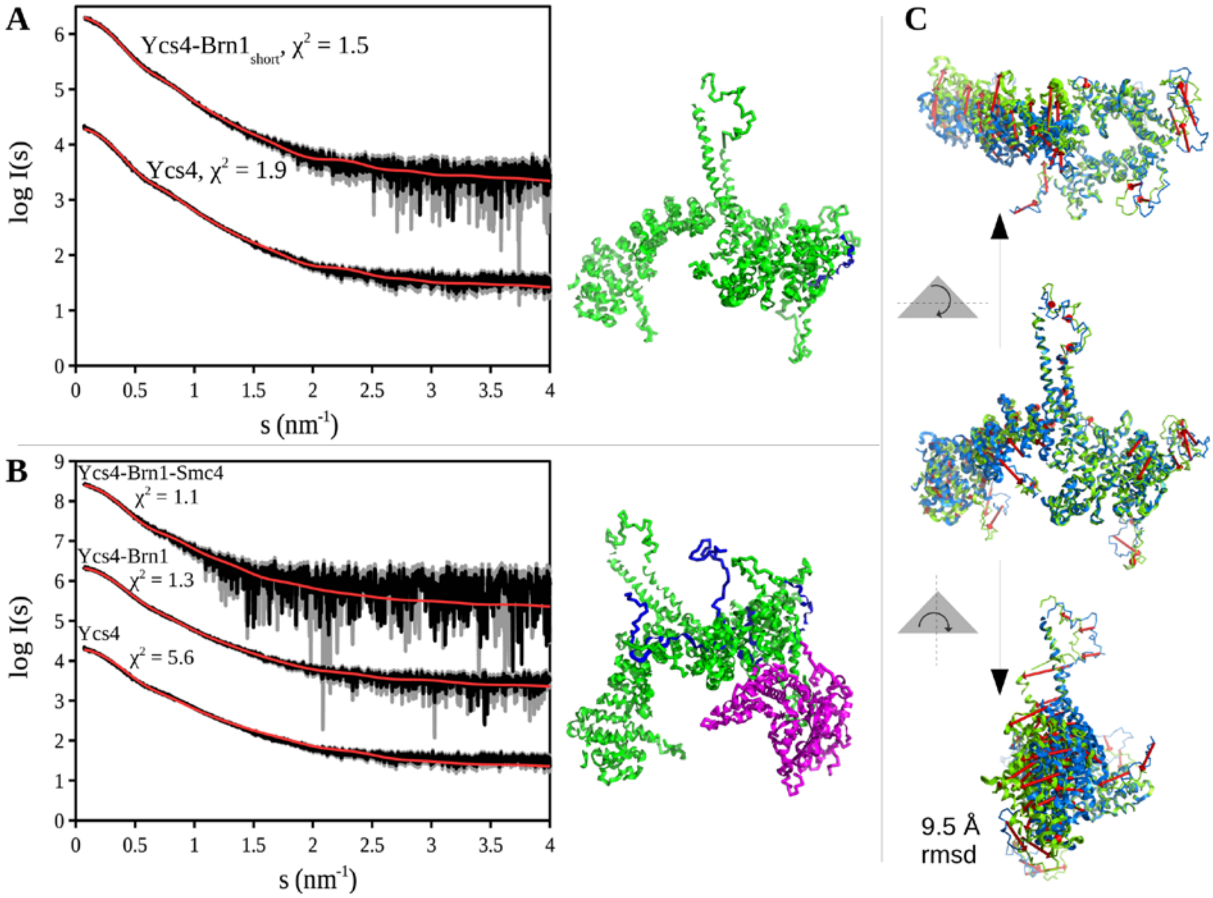

Figure 6-4. Hybrid models of Ycs4-Brn1$_{short}$, and Ycs4-Brn1-Smc4. (A) The Ycs4- Brn1$_{short}$ model was further refined by NMA to fit the scattering data. (B) The Ycs4-Brn1-Smc4 model is consistent with the SAXS data from both Ycs4-Brn1, and Ycs4-Brn1-Smc4, but fits the data from free Ycs4 poorly, indicating a conformational difference between free and Brn1-bound Ycs4. This conformational difference is shown in (C), and involves a 9.5 Å swivel motion of one side of the structure.

6.3 Characterizing Ycg1 oligomerization

Ycg1 oligomerization was observed from the concentration series SAXS data, which showed the formation of increasingly large, non-aggregated particles with increasing concentration (Figure 6-1). Dimer and tetramer models and their proportions at different concentrations were obtained with SASREFMX and shown in Figure 6-4. In this model, the Ycg1 dimerization interface partially overlaps with the Brn1 binding site, which might explain the lack of oligomerization in the Brn1-bound Ycg1 sample.

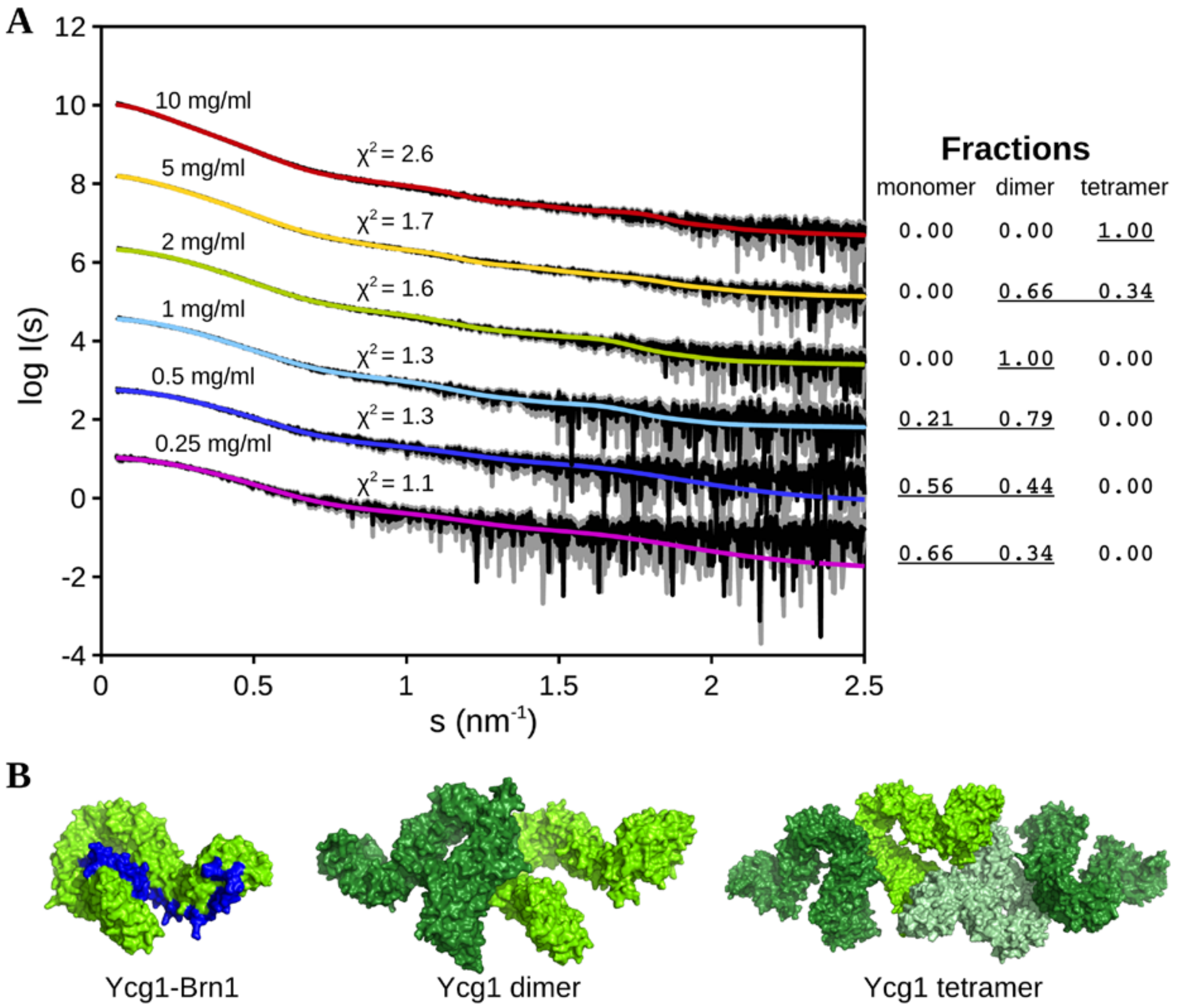

Figure 6-5. Ycg1 concentration-dependent oligomerization. (A) Scattering data from a Ycg1 concentration series were modeled as mixtures of monomer, dimer, and tetramer molecules. As Ycg1 concentration increases, the amount of dimeric and tetrameric species in solution increases. (B) The Ycg1-Brn1 crystal structure, compared with dimer and tetramer Ycg1 models. In the dimer model, the dimerization interface partly overlaps with the Brn1 (blue) binding site, which might explain why oligomerization was not observed to occur for Ycg1–Brn1 (Figure is from Manalastas-Cantos et al., 2019)

Ycg1 oligomerization was confirmed with SEC-MALS and DLS experiments. The DLS measurements revealed a systematic increase in the average hydrodynamic radius R_h (from about 6.5 to about 9 nm) and apparent MW (from about 250 to about 500 kDa) with increasing solute concentration (Figure 6-4A). Size exclusion chromatography coupled to multiangle static light scattering (SEC-MALS) revealed three components in the elution profile (Figure 6-4B). These components correspond to MW values of monomeric, dimeric and tetrameric Ycg1 (~100, ~200 and ~400 kDa), which is in excellent agreement with SAXS modeling results.

Although the condensin HEAT-repeat subunits have previously been speculated to self-assemble (‘phase separate’) via multivalent, weak interactions (Yoshimura and Hirano, 2016), there had been no direct experimental evidence to support this notion. The function of such self-assembling behavior is unclear in the context of Ycg1 and the condensin complex. Nonetheless, the implications of this oligomerization behavior, combined with the oligomer-dissociating effect of Brn1-binding, is an interesting avenue to explore in future studies.

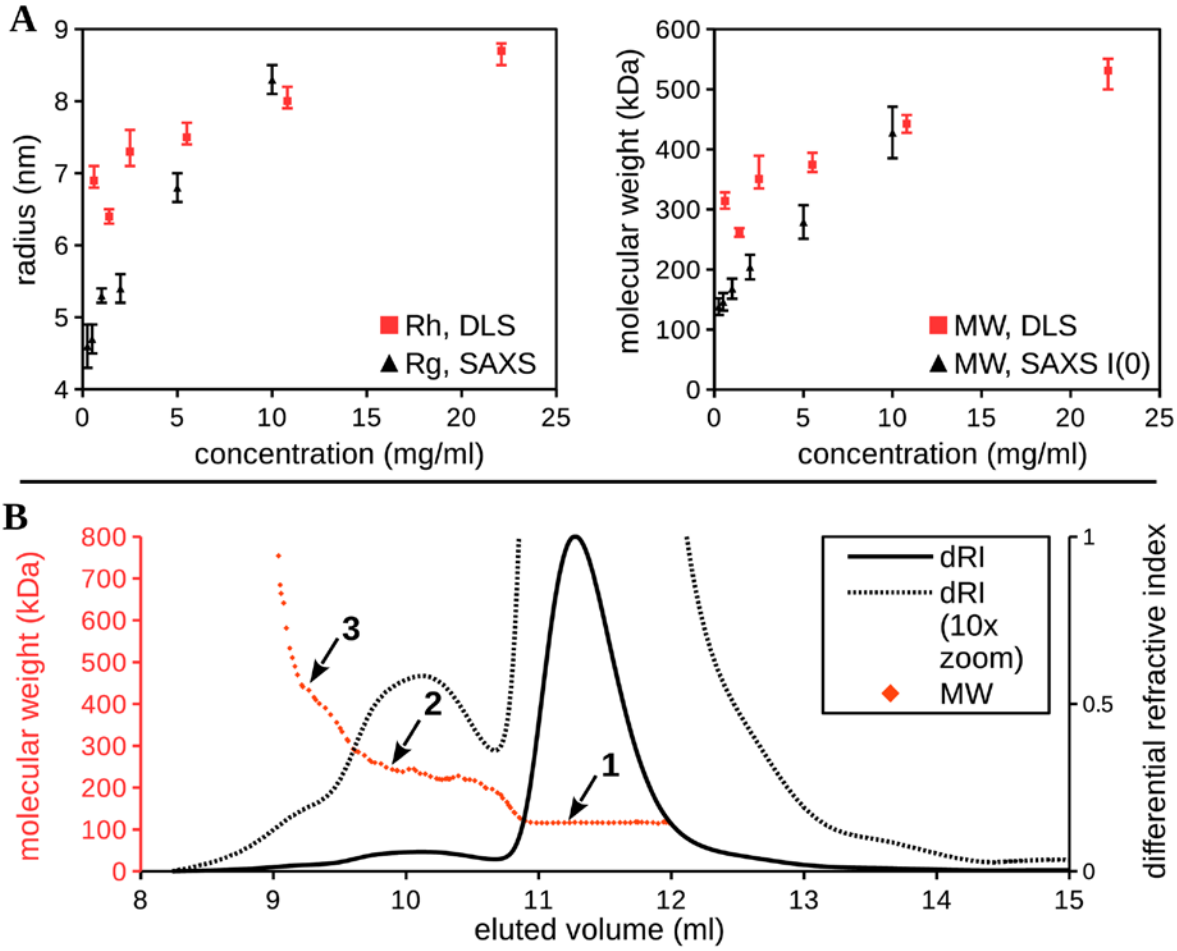

Figure 6-6. Concentration-dependent oligomerization of Ycg1 assessed by light scattering. (A) Batch DLS measurements show R_h and MW increasing with concentration, similar to SAXS. (B) SEC-MALS of Ct Ycg1 confirms the presence of oligomeric species with monomeric (1: ~100 kDa), dimeric (2: ~200 kDa), and tetrameric (3: ~400 kDa) MWs. Note that the dimers and tetramers may be dissociating during chromatography due to dilution, causing them to be present in much smaller amounts compared to the SAXS experiment. (Figure is from Manalastas-Cantos et al., 2019)

6.4 Conclusion and outlook

The condensin HEAT-repeat proteins Ycg1 and Ycs4 were shown to be flexible in solution, with this flexibility constrained by the binding of Brn1. For both Ycg1 and Ycs4, Brn1-binding appears to bring the flexible proteins into their functional conformation in the condensin complex. Additionally, Ycg1 was observed to form dimers and tetramers in solution, that are dissociated upon Brn1 binding. This was the first time condensin HEAT-repeat protein oligomerization was observed experimentally. The functional role of this oligomerization, as well as how frequently it occurs for HEAT-repeat proteins in general, would be interesting to explore in future work.

For detailed information regarding the SAXS data analysis and modeling of Ycg1, readers are advised to consult the paper (Manalastas-Cantos *et al.*, 2019).

7 Modeling iron-sulfur cluster biosynthesis proteins in *Escherichia coli*: HscA, HscB, IscU and the HscB-IscU complex

Iron–sulfur (Fe-S) clusters are used by more than 200 different types of proteins and as such, represent one of the most ubiquitous biological prosthetic groups across multiple branches of life. Hence, the Fe-S cluster biosynthesis pathway is remarkably conserved in both prokaryotic and eukaryotic organisms. For bacteria such as *Escherichia coli*, the ISC operon contains the primary set of genes involved in Fe–S cluster biosynthesis Figure 7-1. Moreover, aside from a few additional components, the eukaryotic mitochondrial machinery also uses the ISC system for Fe–S cluster biogenesis. This ubiquity, functional importance, and high-degree on conservation makes the ISC system of considerable interest (Bandyopadhyay, Chandramouli and Johnson, 2008).

The ISC system consists of a scaffold protein (IscU or IscA), upon which a cysteine desulfurase IscS attaches transient [2Fe–2S]$^{2+}$ and [4Fe–4S]$^{2+}$ clusters, to be transferred to acceptor proteins. The operon contains additional genes that encode a regulatory protein (IscR), a ferredoxin (Fdx), and two heat-shock like proteins (HscA and HscB).

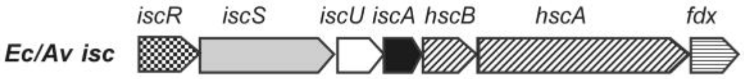

Figure 7-1. The E. coli ISC operon (Figure modified from Bandyopadhyay, Chandramouli and Johnson, 2008)

The high-resolution structures of almost all the individual *E. coli* ISC proteins have been solved (aside from HscA, which has only its substrate binding domain elucidated), as well as some of the interactions (Cupp-Vickery and Vickery, 2000; Kakuta *et al.*, 2001; Cupp-Vickery, Peterson, *et al.*, 2004; Cupp-Vickery, Silberg, *et al.*, 2004; Shi *et al.*, 2010; Kim, Tonelli and Markley, 2012; Rajagopalan *et al.*, 2013). Solution SAXS provides a quick way of screening for novel interactions: the scattering from a complex is markedly different from the sum of scattering from the components alone, and this difference can be detected by software tools such as OLIGOMER (Konarev *et al.*, 2003). SAXS data can also be used in combination with the available high-resolution structures of the individual proteins to construct solution structure models of the complexes in the Fe-S biosynthesis apparatus.

In this work, we measured SAXS data from ISC proteins HscA, HscB, and IscU individually and as pairs, screening for pairwise interactions. Solution structures for the monomers IscU, HscB, and HscA were obtained and compared to existing high-resolution structures. In addition, HscB and IscU heterodimer formation was detected with solution SAXS, and models of the HscB-IscU heterodimer were produced from the scattering data. Due to the ubiquity and conservation of the ISCU system, coevolution was used to predict heterodimer contacts between HscB and IscU, thus reducing SAXS modeling ambiguity.

7.1 Experimental procedures

7.1.1 SAXS data collection and processing

The ISC system proteins HscA, HscB, and IscU were produced by Rita Puglisi (Pastore group, King’s College London), and described in Table 7-1. SAXS data for each sample were collected at the SAXS beamline P12 of the PETRA III storage ring (Deutsches Elektronen-Synchrotron, Hamburg) (Blanchet *et al.*, 2015). The details of the data collection conditions are summarized in Table 7-2. The scattering data in the momentum transfer range $0.003 < s < 0.73$ Å^{-1} were collected with a PILATUS 6M pixel detector at a distance of 3.0 m from the sample. For HscB and IscU, solute concentrations ranging from 1.25 to 10 mg/ml were measured, while for HscA, the concentration range 3.75 - 30 mg/ml was examined. The samples were also mixed in a 1:1 molar ratio of all the possible two-protein combinations (HscA-HscB, HscA-IscU, and HscB-IscU), then subjected to size-exclusion chromatography

(SEC) directly upstream of the SAXS capillary (Graewert *et al.*, 2015). SEC was performed using an Agilent 1260 Infinity II Bio-inert LC system, and an analytical Superdex200 increase 5/150 GL column (GE Healthcare) equilibrated with the sample buffer (20 mM tris, 150 mM NaCl, 2 mM DTT, pH 8) at 20°C. Forty microliters of 10 mg/ml sample was injected for HscA-HscB and HscA-IscU (6 mg/ml for HscB-IscU), with the experiment performed at a flow rate of 0.45 ml/min.

The samples were constantly flowed through the capillary during X-ray exposure in order to minimize radiation damage. The two-dimensional pixel data from the detector were converted to one-dimensional scattering profiles using the automated pipeline SASFLOW, which performed radial averaging, outlier removal, data averaging, and buffer subtraction (Franke, Kikhney and Svergun, 2012). For the SEC-SAXS data, radially-averaged time course data from SASFLOW were viewed with CHROMIXS (Panjkovich and Svergun, 2017), from where buffer and sample frames were manually selected. Data averaging and buffer subtraction were then done using Primus (Konarev *et al.*, 2003), to produce the scattering profile from the putative complex. The analysis of the SAXS data was performed using the ATSAS 2.8 suite (Franke *et al.*, 2017). The concentration series SAXS data for each monomer sample were assessed for monodispersity and the absence of repulsive or attractive interactions, by checking for linearity in the Guinier region (Figure 7-2). No major concentration effects were observed for the monomer samples. The selected scattering profiles were HscA at 15 mg/ml, HscB at 10 mg/ml, and IscU at 10 mg/ml. These selected scattering profiles were used for further analysis and modeling.

Table 7-1. Sample properties

	HscA	**HscB**	**IscU**
Organism	*P. falciparum*		
Source (UniProt ID)	P0A6Z1	P0A6L9	P0ACD4
Protein name (residues in construct)	$\text{HscA}_{1\text{-}616}$	$\text{HscB}_{1\text{-}171}$	$\text{IscU}_{1\text{-}128}$
Solvent	300 mM NaCl, 25 mM tris-HCl pH 7.5, 1 mM DTT		

For each scattering profile, the $I(0)$ and R_g were obtained from the Guinier approximation (Guinier, 1939), following the standard procedures (Konarev *et al.*, 2006). The $P(r)$ function was computed using the indirect Fourier transformation method implemented in GNOM (Semenyuk and Svergun, 1991). The $P(r)$ function was used to derive alternative estimates of R_g and D_{max}. Molecular weights in solution were assessed from the SAXS data with three methods: (a) using the forward scattering, and comparing against a reference solution of bovine serum albumin (Jeffries *et al.*, 2016), (b) from the excluded (Porod) volume V_p (given that V_p in nm^3 is about 1.6 times the MW in kDa) (Franke *et al.*, 2012), and (c) a consensus Bayesian MW assessment method (Hajizadeh *et al.*, 2018).

7.1.2 SAXS data analysis and structure modeling

The scattering curves from available high-resolution models of HscA, HscB, and IscU were computed, and their χ^2 fits against the experimental SAXS data evaluated using CRYSOL (Svergun, Barberato and Koch, 1995). The structures used as reference were the *Escherichia coli* HscB crystal structure (PDB ID: 1fpo), the *E. coli* IscU structure from solution NMR (PDB ID: 2l4x), and a chimeric model of HscA. The HscA model was constructed using the crystal structure of the *E.coli* HscA substrate binding domain (PDB ID: 1u00), with its missing ATPase domain taken from cognate protein Hsp70 (PDB ID: 2kho). The Hsp70 structure was also used as a reference for the relative orientations of the substrate binding and ATPase domains of the HscA model. To obtain models that better fit the experimental scattering data, CNMA and TNMA were performed on HscA and HscB using the programs SREFLEX (Panjkovich and Svergun, 2016) and NMATOR (discussed in Chapter 2), respectively.

Complex formation was probed from the SAXS data of each two-protein combination (HscA-HscB, HscA-IscU, and HscB-IscU) using OLIGOMER (Konarev *et al.*, 2003). OLIGOMER was used to approximate the SAXS data from the two-protein mixtures as a sum of the monomer scattering profiles. Inability to fit the mixture data as a combination of monomers indicates the presence of another scattering species, which in this case is the protein complex. Using this procedure, complex formation was detected for HscB-IscU.

The structure and proportion in solution of the HscB-IscU dimer was modeled ten times using SASREFMX (Petoukhov and Svergun, 2005; Franke *et al.*, 2012; Petoukhov *et al.*, 2013). To decrease modeling ambiguity, heterodimer contacts between HscB and IscU were predicted using coevolution analysis (discussed in Chapter 3). Homologous sequences to HscB and IscU were queried using HMMer (version 3.1b2) (Eddy, 2011). The homologous sequences were matched by species-of-origin, with any unmatched sequences discarded. The remaining sequences were aligned to the HscB and IscU sequences with Clustal Omega (version 1.2.3) (Sievers *et al.*, 2011), and concatenated into a single long multiple sequence alignment, which was analyzed for pairwise positional coevolution using direct coupling analysis (Morcos *et al.*, 2011). The top scoring heterodimer contact from DCA was used as a distance constraint for another ten SASREF modeling runs.

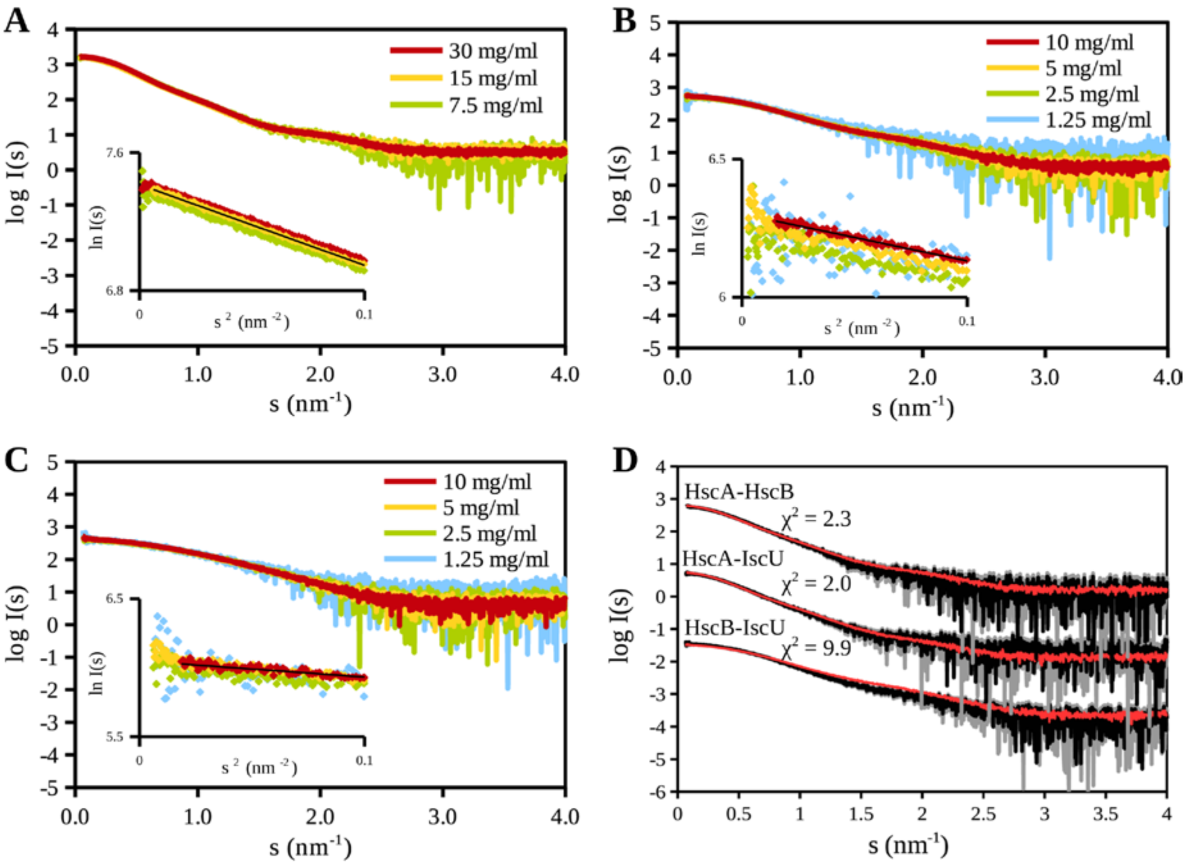

Figure 7-2. Concentration series SAXS data and Guinier plots (left inset) for (A) HscA, (B) HscB, and (C) IscU. For all three proteins, linear Guinier plots indicate a negligible amount of aggregation or interparticle repulsion, while the constant slope indicates that R_g remains constant in the concentration ranges at which SAXS data were measured, demonstrating absence of concentration-dependent oligomerization. (D) SEC-SAXS data the two-protein combinations HscB-IscU, HscA-IscU, and HscA-HscB. The red overlay is the scattering profile of the combination of monomers that best fits the scattering data. For HscA-HscB and HscA-IscU, the difference between the OLIGOMER profiles and the SAXS data were insufficient to unambiguously indicate complex formation. For HscB-IscU, the combination of monomers noticeably does not account for the observed scattering.

Table 7-2. Data collection and structure statistics for small angle X-ray scattering analysis

Data collection parameters	**HscA**	**HscB**	**IscU**	**HscB-IscU**
Instrument	EMBL P12 (PETRA III, DESY, Hamburg)			
Beam geometry (mm^2)	0.2×0.12			
Wavelength (Å)	1.24			
s range ($Å^{-1}$)	0.003-0.73			
Exposure time (s)	1 (20×0.05s)			
Temperature (K)	293			
Concentration range measured ($mg\ ml^{-1}$)	3.75 - 30	1.25 - 10	1.25 – 10	unknown (SEC)
Concentration used ($mg\ ml^{-1}$)	15	10	10	unknown (SEC)
Structural parameters				
R_g (Å) (from $P(r)$)	38±1	23±1	19±1	24±1
R_g (Å) (from Guinier plot)	38±1	23±1	18±1	24±1
D_{max} (Å)	130±10	75±5	70±5	75±5
Porod volume estimate ($10^3 Å^3$)	135	29	17	31
Excluded volume estimate[§] ($10^3 Å^3$)	135	25	17	42
Molecular weight determination (kDa)				
From Porod volume, V_p (V_p/~1.6)	84±17	18±4	10±2	19±4
From consensus Bayesian assessment	89±7	22±1	14±1	29±1
From $I(0)$	68±7	20±2	12±1	n.a.
Calculated monomeric *MW* from sequence	66	20	14	34
Software employed				
Primary data reduction	SASFLOW	SASFLOW	SASFLOW	SASFLOW
Data processing	PRIMUS	PRIMUS	PRIMUS	CHROMIXS, PRIMUS
Detection of complex formation	n.a.	n.a.	n.a.	OLIGOMER
Coevolution analysis	n.a.	n.a.	n.a.	DCA
Rigid body modelling	n.a.	n.a.	n.a.	SASREFMX
Flexibility modeling	SREFLEX, NMATOR	SREFLEX, NMATOR	n.a.	n.a.
Computation of model intensities	CRYSOL	CRYSOL	CRYSOL	CRYSOL
3D graphics representations	PyMOL[+]	PyMOL[+]	PyMOL[+]	PyMOL[+]

[§] excluded volume calculations made with the reference structures, using Mol_volume, Version 1.0, Theoretical Biophysics Group, University of Illinois (retrieved from "http://www.ks.uiuc.edu/Development/MDTools/molvolume/")

[+] The PyMOL Molecular Graphics System, Version 1.7.2.1, Schrödinger, LLC

7.2 Solution characteristics and models for HscA, HscB and IscU

The scattering profiles of the monomers HscA, HscB, and IscU are shown on Figure 7-3A, along with the simulated scattering data from their respective high-resolution structures (Figure 7-3B). Unsurprisingly, the solution NMR structure for IscU corresponded well to the solution SAXS data. On the other hand, both HscB crystal structure and the HscA chimeric model did not have good agreement with the experimental scattering data, indicating the need for further modeling.

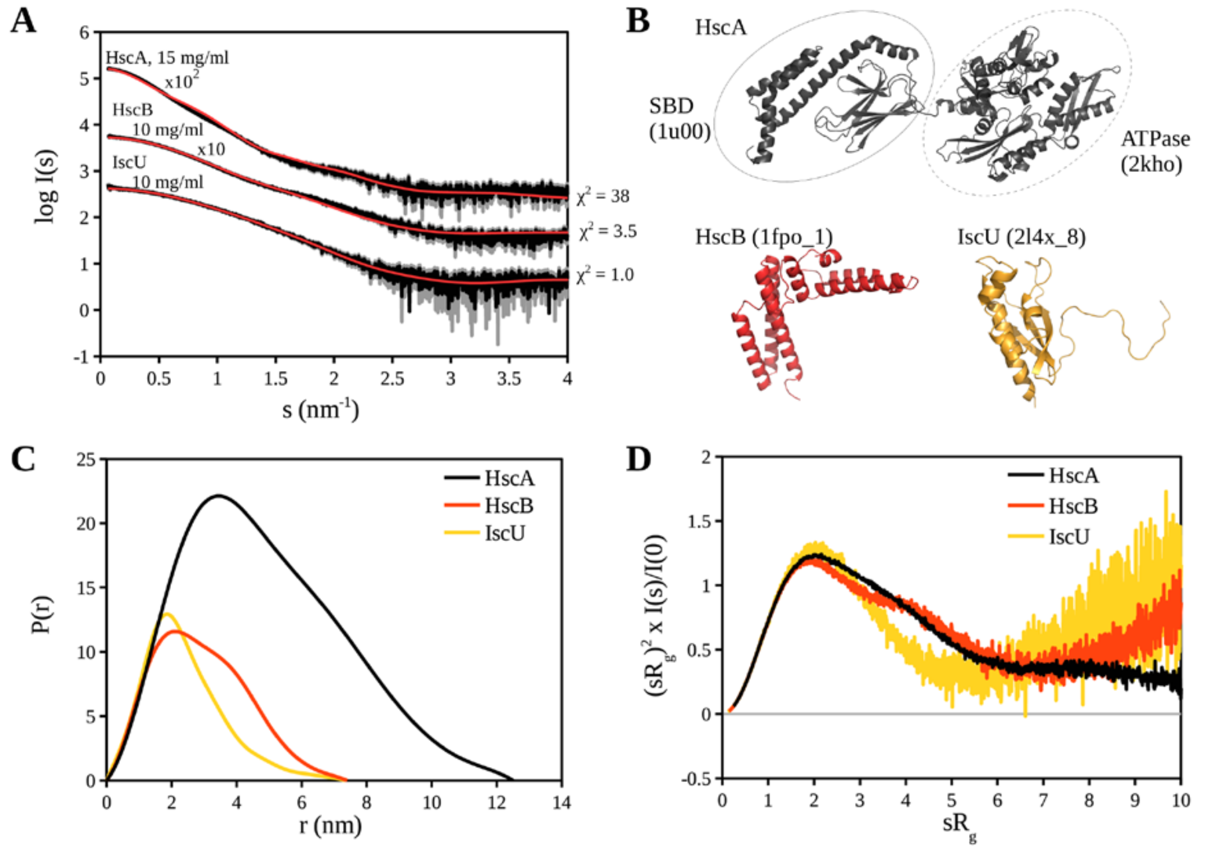

Figure 7-3. (A) Scattering data from the HscA, HscB, and IscU monomers. The red overlay in each case represents the computed scattering from the corresponding high-resolution models. (B) Full-length, high-resolution structures of HscB, and IscU, and a chimeric HscA model, built by aligning the HscA SBD with the SBD of the homologous protein Hsp70 (PDB ID: 2kho), and using both the ATPase domain and the relative domain orientations from the Hsp70 structure. This resulting model does not coincide well with the SAXS data from (A), shown by the poor fit between the scattering profiles. (C) The pair distance distribution, and (D) normalized Kratky plots show that HscA, HscB, and IscU are largely globular, although HscB shows a subtle two-domain structure with a small amount of flexibility between the domains.

The pair distance distribution functions and Kratky plots (Figure 7-3, panels C and D) for HscA and IscU indicate mostly compact, rigid structures, but with the tail of the $P(r)$ distribution of IscU corresponding well to the "tail" in the solution NMR structure. For HscB, both the $P(r)$ function and Kratky plot show a two-domain structure with some hinge-like flexibility between them, which could possibly map back to flexible motions in the junction of HscB's L-shaped structure (Figure 7-2B).

In order to further characterize the solution structures of HscA and HscB, hybrid models were built from the initial, reference structures using Cartesian and torsional NMA. Figure 7-4 shows solution structure models for HscA and HscB that fit the experimental scattering data. For HscB, both Cartesian and torsional NMA resulted in a swiveling motion of one of its helical bundles, originating near the junction of the L-shaped structure. For HscA, CNMA resulted in a large domain repositioning, indicating that the domain configuration in the Hsp70 crystal structure is significantly different from the domain orientations of HscA in solution, even though their sequences are so similar. Interestingly, this might be linked to the observation that while HscA is functionally a molecular chaperone like Hsp70, its substrates are very specific to the Fe-S biosynthesis pathway (Vickery and Cupp-Vickery, 2007), unlike the more promiscuous Hsp70 (Mayer and Bukau, 2005).

On the other hand, torsional NMA did not find an HscA model that fit the scattering data well. Noticeable conformational change within one of the domains could be seen in the TNMA HscA model (Figure 7-2A), since domains are not kept rigid in the currently implemented NMATOR. In order to be applicable to large, multidomain proteins, some possible features to add to NMATOR could be (1) automatic domain detection, as in SREFLEX, and (2) the capacity to keep domains rigid during refinement. Overall, the results obtained in Chapters 6-7 indicate that the performance of CNMA on proteins is generally somewhat better, and at least not worse than that of TNMA. Therefore, as indicated above, TNMA should be largely considered for nucleic acids (see the application in Chapter 5), where CNMA shows its limitations.

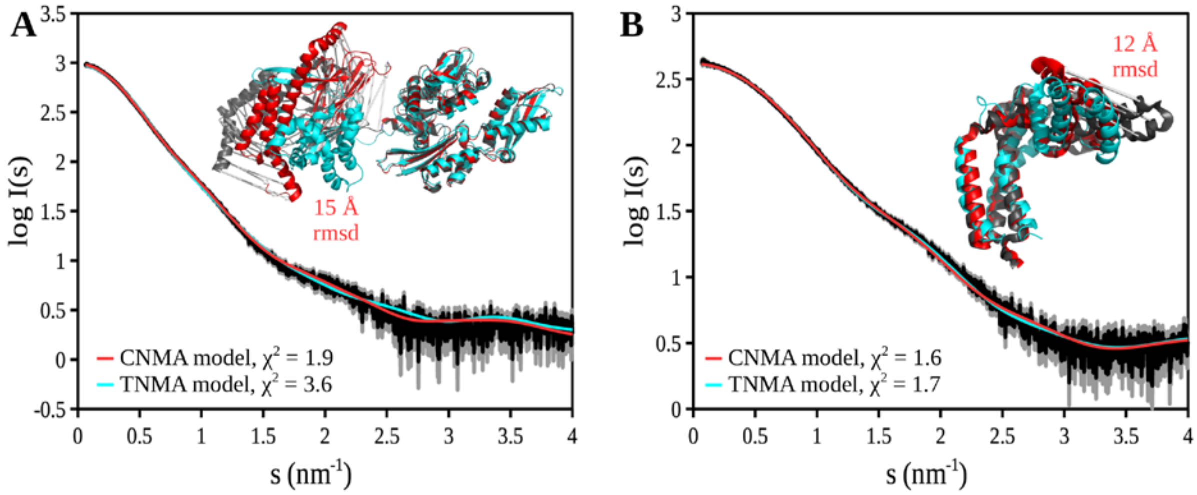

Figure 7-4. Hybrid models of (A) HscA, and (B) HscB from CNMA with SREFLEX (red), and TNMA with NMATOR (cyan), along with the fits of the resulting models to the SAXS data. Initial structures are shown in gray, and the net movement shown by gray arrows, with the rmsd shown alongside. TNMA found a similar model as CNMA for HscB, but did not find a well-fitting model for HscA, suggesting that the RTB approach might better describe domain movements for larger proteins.

7.3 Detecting and modeling the HscB-IscU interaction

The results from OLIGOMER showed that while the scattering data from the HscA-IscU and HscA-HscB mixtures could be fully accounted for by the monomers alone, the SAXS data from the HscB-IscU mixture could not (Figure 7-5). This unambiguously indicated that another scattering species was being formed. From the Porod volume and MW estimates (Table 7-1), the complex seems to be a partially-disssociated 1:1 complex. The $P(r)$ function and Kratky plot indicates a globular, relatively rigid complex.

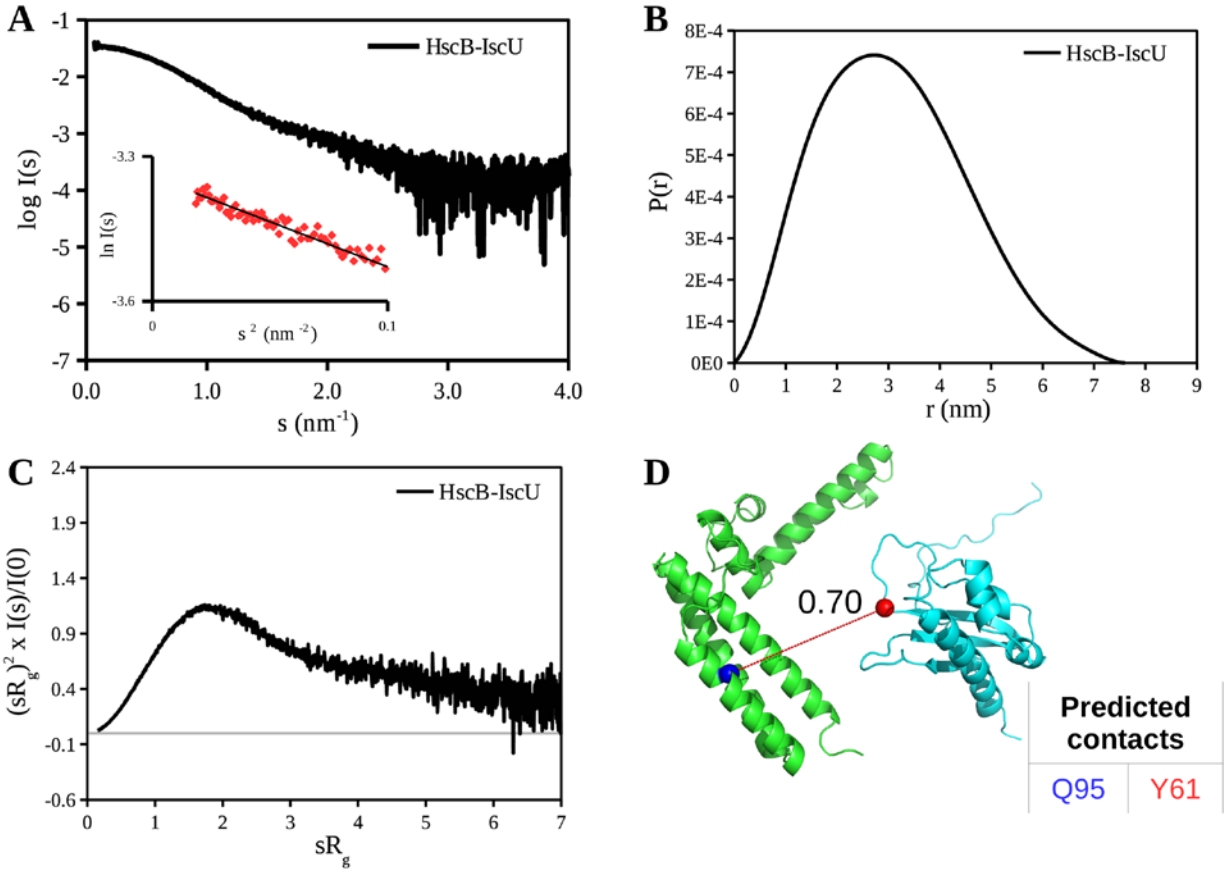

Figure 7-5. (A) SEC-SAXS data from the 1:1 HscB-IscU, with the associated Guinier plot (inset) (B) The pair distance distribution P(r), and (C) the dimensionless Kratky plot from the HscB-IscU SAXS data both indicate a compact, globular complex. (D) shows the top contact predicted by coevolution, along with the prediction confidence.

Heterodimer models for HscB-IscU were constructed through SAXS-guided rigid body modeling. Since the ISC system of proteins is highly conserved, it was deemed a good candidate for coevolution-based contact prediction. The top scoring heterodimer contact (Figure 7-5D) was used as a distance constraint for another round of rigid body modeling. The resulting models that were built with and without coevolution information are shown in Figure 7-6.

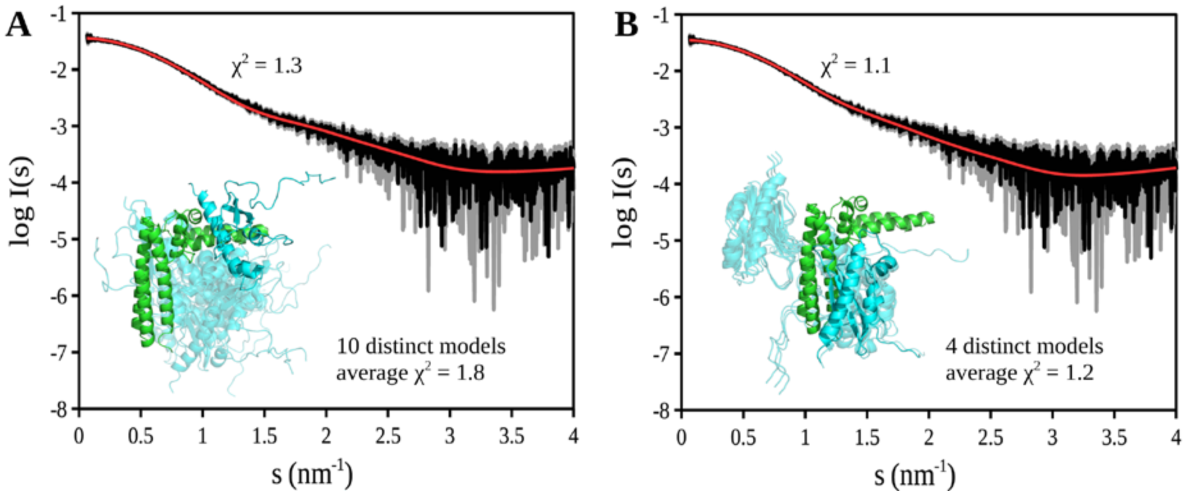

Figure 7-6. Rigid-body models of the HscB-IscU heterodimer from SASREFMX. (A) shows the models from ten unconstrained runs, with the best fitting model shown in opaque colors, along with the fit of the model to the scattering data. (B) similarly shows the models from ten runs with distance constraints from coevolution analysis. Adding a constraint significantly reduces model variability.

The HscB-IscU models consistently placed IscU such that it filled the cavity in the corner of the L-shaped HscB. The unconstrained models were highly variable in the orientation of IscU, however. The addition of a distance constraint greatly reduced the variability of the heterodimer models, since it restricted the possible orientations of IscU to those that would preserve the predicted contact. Whether the predicted contact has a biological or chemical significance, or is just a statistical artifact picked up by coevolution analysis could be validated by site-directed mutagenesis.

A somewhat surprising result is the lack of interaction detected between HscA and IscU. Because of its role as a scaffold protein, IscU was expected to interact with all of the proteins. In addition, the HscA has been cocrystallized with a bound IscU peptide (Cupp-Vickery, Peterson, *et al.*, 2004). However, the size difference between HscA and IscU might have made it difficult to distinguish between HscA and HscA-IscU complex at low resolution. The OLIGOMER-derived scattering profile consisted around 90% HscA and 10% IscU, indicating that just the HscA monomer is mostly sufficient to account for scattering from the mixture. Another possibility is that the HscA-IscU dissociated in solution, due to dilution during chromatography.

7.4 Conclusion and outlook

SAXS data were measured at the EMBL P12 beamline for the ISC proteins HscA, HscB, and IscU, from the individual proteins and their binary complexes. The IscU solution structure was found to be consistent with the known high-resolution structure. Models for HscA and HscB solution structure were derived through from high-resolution models through normal mode analysis.

HscB and IscU were demonstrated to form a heterodimer in solution. Possible structures of the HscB-IscU complex were derived with SAXS-guided rigid-body modeling. The ambiguity of the computed models was decreased by applying distance constraints derived from coevolution analysis. The derived HscB-IscU models would greatly benefit from validation with further lab experiments.

References

Abe, H. *et al.* (1984) ‘Rapid calculation of first and second derivatives of conformational energy with respect to dihedral angles for proteins general recurrent equations’, *Computers & Chemistry*. Elsevier, 8(4), pp. 239–247.

Alexandrov, V. *et al.* (2005) ‘Normal modes for predicting protein motions: a comprehensive database assessment and associated Web tool’, *Protein science*. Wiley Online Library, 14(3), pp. 633–643.

Anderson, E. *et al.* (1999) *{LAPACK} Users’ Guide*. Third. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Andrade, M. A. and Bork, P. (1995) ‘HEAT repeats in the Huntington’s disease protein’, *Nature Genetics*, 11, pp. 115–116.

Atilgan, A. R. *et al.* (2001) ‘Anisotropy of fluctuation dynamics of proteins with an elastic network model’, *Biophysical journal*. Elsevier, 80(1), pp. 505–515.

Bailor, M. H., Sun, X. and Al-Hashimi, H. M. (2010) ‘Topology links RNA secondary structure with global conformation, dynamics, and adaptation’, *Science*. American Association for the Advancement of Science, 327(5962), pp. 202–206.

Bandyopadhyay, S., Chandramouli, K. and Johnson, M. K. (2008) ‘Iron--sulfur cluster biosynthesis’. Portland Press Limited.

Berman, H. M. *et al.* (1992) ‘The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids.’, *Biophysical journal*. The Biophysical Society, 63(3), p. 751.

Blanchet, C. E. *et al.* (2015) ‘Versatile sample environments and automation for biological solution X-ray scattering experiments at the P12 beamline (PETRA III, DESY)’, *Journal of Applied Crystallography*. International Union of Crystallography, 48(2), pp. 431–443. doi: 10.1107/S160057671500254X.

Bousset, L. *et al.* (2014) ‘Crystal structure of a signal recognition particle Alu domain in the elongation arrest conformation’, *RNA*. Cold Spring Harbor Lab, 20(12), pp. 1955–1962.

Bray, J. K., Weiss, D. R. and Levitt, M. (2011) ‘Optimized torsion-angle normal modes reproduce conformational changes more accurately than cartesian modes’, *Biophysical journal*. Elsevier, 101(12), pp. 2966–2969.

Cantara, W. A., Olson, E. D. and Musier-Forsyth, K. (2017) ‘Analysis of RNA structure using small-angle X-ray scattering’, *Methods*. Elsevier, 113, pp. 46–55.

Chacon, P. *et al.* (1998) ‘Low-resolution structures of proteins in solution retrieved from X-ray scattering with a genetic algorithm’, *Biophysical Journal*. Elsevier, 74(6), pp. 2760–2775.

Chandonia, J.-M. and Brenner, S. E. (2006) ‘The impact of structural genomics: expectations and outcomes’, *Science*. American Association for the Advancement of Science, 311(5759), pp. 347–351.

Chen, Y. and Pollack, L. (2016) ‘SAXS studies of RNA: structures, dynamics, and interactions with partners', *Wiley Interdisciplinary Reviews: RNA*. Wiley Online Library, 7(4), pp. 512–526.

Chu, V. B. *et al.* (2009) 'Do conformational biases of simple helical junctions influence RNA folding stability and specificity?', *RNA*. Cold Spring Harbor Lab, 15(12), pp. 2195–2205.

Consortium, U. (2011) 'Reorganizing the protein space at the Universal Protein Resource (UniProt)', *Nucleic acids research*. Oxford University Press, 40(D1), pp. D71--D75.

Cormen, T. H. *et al.* (2009) *Introduction to algorithms*. MIT Press.

Cupp-Vickery, J. R., Silberg, J. J., *et al.* (2004) 'Crystal structure of IscA, an iron-sulfur cluster assembly protein from Escherichia coli', *Journal of molecular biology*. Elsevier, 338(1), pp. 127–137.

Cupp-Vickery, J. R., Peterson, J. C., *et al.* (2004) 'Crystal structure of the molecular chaperone HscA substrate binding domain complexed with the IscU recognition peptide ELPPVKIHC', *Journal of molecular biology*. Elsevier, 342(4), pp. 1265–1278.

Cupp-Vickery, J. R. and Vickery, L. E. (2000) 'Crystal structure of Hsc20, a J-type Co-chaperone from Escherichia coli', *Journal of molecular biology*. Elsevier, 304(5), pp. 835–845.

Cuylen, S., Metz, J. and Haering, C. H. (2011) 'Condensin structures chromosomal DNA through topological links', *Nature Structural and Molecular Biology*. Nature Publishing Group, 18(8), pp. 894–901. doi: 10.1038/nsmb.2087.

Das, R. and Baker, D. (2007) 'Automated de novo prediction of native-like RNA tertiary structures', *Proceedings of the National Academy of Sciences*. National Acad Sciences, 104(37), pp. 14664–14669.

Debye, P. (1915) 'Zerstreuung von röntgenstrahlen', *Annalen der Physik*. Wiley Online Library, 351(6), pp. 809–823.

Debye, P. (1947) 'Molecular-weight determination by light scattering.', *The Journal of Physical Chemistry*. ACS Publications, 51(1), pp. 18–32.

Debye, P., Anderson Jr, H. R. and Brumberger, H. (1957) 'Scattering by an inhomogeneous solid. II. The correlation function and its application', *Journal of applied Physics*. AIP, 28(6), pp. 679–683.

Eddy, S. R. (2011) 'Accelerated profile HMM searches', *PLoS Computational Biology*, 7(10). doi: 10.1371/journal.pcbi.1002195.

Feinauer, C. *et al.* (2016) 'Inter-Protein Sequence Co-Evolution Predicts Known Physical Interactions in Bacterial Ribosomes and the Trp Operon', pp. 1–18. doi: 10.1371/journal.pone.0149166.

Franke, D. *et al.* (2012) ' New developments in the ATSAS program package for small-angle scattering data analysis ', *Journal of Applied Crystallography*. International Union of Crystallography, 45(2), pp. 342–350. doi: 10.1107/s0021889812007662.

Franke, D. *et al.* (2017) 'ATSAS 2.8: A comprehensive data analysis suite for small-angle scattering from macromolecular solutions', *Journal of Applied Crystallography*. International Union of Crystallography, 50, pp. 1212–1225. doi: 10.1107/S1600576717007786.

Franke, D., Jeffries, C. M. and Svergun, D. I. (2015) ‘Correlation Map, a goodness-of-fit test for one-dimensional X-ray scattering spectra’, *Nature Methods*. Nature Publishing Group, 12(5), pp. 419–422. doi: 10.1038/nmeth.3358.

Franke, D., Kikhney, A. G. and Svergun, D. I. (2012) ‘Automated acquisition and analysis of small angle X-ray scattering data’, *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*. Elsevier, 689, pp. 52–59. doi: 10.1016/j.nima.2012.06.008.

Franke, D. and Svergun, D. I. (2009) ‘DAMMIF , a program for rapid ab-initio shape determination in small-angle scattering ’, *Journal of Applied Crystallography*. International Union of Crystallography, 42(2), pp. 342–346. doi: 10.1107/s0021889809000338.

Glatter, O. (1977) ‘A new method for the evaluation of small-angle scattering data’, *Journal of Applied Crystallography*. International Union of Crystallography, 10(5), pp. 415–421.

Go, N., Noguti, T. and Nishikawa, T. (1983) ‘Dynamics of a small globular protein in terms of low-frequency vibrational modes’, *Proceedings of the National Academy of Sciences*. National Acad Sciences, 80(12), pp. 3696–3700.

Goldstein, H. (1950) *Classical mechanics*. Reading, Massachusetts: Addison-Wesley-Longman.

Goodwin, S., McPherson, J. D. and McCombie, W. R. (2016) ‘Coming of age: ten years of next-generation sequencing technologies’, *Nature Reviews Genetics*. Nature Publishing Group, 17(6), p. 333.

Gorba, C. and Tama, F. (2010) ‘Normal mode flexible fitting of high-resolution structures of biological molecules toward SAXS data’, *Bioinformatics and biology insights*. SAGE Publications Sage UK: London, England, 4, p. BBI--S4551.

Graewert, M. A. *et al.* (2015) ‘Automated Pipeline for Purification, Biophysical and X-Ray Analysis of Biomacromolecular Solutions’, *Scientific Reports*. Nature Publishing Group, 5(1), p. 10734. doi: 10.1038/srep10734.

Grinthal, A. *et al.* (2010) ‘PR65, the HEAT-repeat scaffold of phosphatase PP2A, is an elastic connector that links force and catalysis’, *Proceedings of the National Academy of Sciences*. National Acad Sciences, 107(6), pp. 2467–2472.

Grishaev, A. *et al.* (2010) ‘Improved fitting of solution X-ray scattering data to macromolecular structures and structural ensembles by explicit water modeling’, *Journal of the American Chemical Society*. ACS Publications, 132(44), pp. 15484–15486.

Guinier, A. (1939) ‘La diffraction des rayons X aux très petits angles: application à l’étude de phénomènes ultramicroscopiques’, in *Annales de physique*, pp. 161–237.

Haering, C. H. *et al.* (2017) ‘Structural Basis for a Safety-Belt Mechanism That Anchors Condensin to Chromosomes’, *Cell*. Elsevier, 171(3), pp. 588-600.e24. doi: 10.1016/j.cell.2017.09.008.

Hajizadeh, N. R. *et al.* (2018) ‘Consensus Bayesian assessment of protein molecular mass from solution X-ray scattering data’, *Scientific reports*. Nature Publishing Group, 8(1), p. 7204.

Halperin, I., Wolfson, H. and Nussinov, R. (2006) ‘Correlated Mutations: Advances and Limitations. A Study on Fusion Proteins and on the Cohesin-Dockerin Families’, 845(February), pp. 832–845. doi: 10.1002/prot.

Hassler, M. *et al.* (2019) ‘Structural Basis of an Asymmetric Condensin ATPase Cycle’, *Molecular cell*. Elsevier, 74(6), pp. 1175–1188.

Hendrickson, W. A. (2014) ‘Anomalous diffraction in crystallographic phase evaluation’, *Quarterly reviews of biophysics*. Cambridge University Press, 47(1), pp. 49–93.

Hendrickson, W. A. and Teeter, M. M. (1981) ‘Structure of the hydrophobic protein crambin determined directly from the anomalous scattering of sulphur’, *Nature*. Nature Publishing Group, 290(5802), p. 107.

Hinsen, K., Thomas, A. and Field, M. J. (1999) ‘Analysis of domain motions in large proteins’, *Proteins: Structure, Function, and Bioinformatics*. Wiley Online Library, 34(3), pp. 369–382.

Hopf, T. A. *et al.* (2014) ‘Sequence co-evolution gives 3D contacts and structures of protein complexes’, pp. 1–45. doi: 10.7554/eLife.03430.

Houlard, M. *et al.* (2015) ‘Condensin confers the longitudinal rigidity of chromosomes’, *Nature cell biology*. Nature Publishing Group, 17(6), p. 771.

Ivanov, D. and Nasmyth, K. (2005) ‘A topological interaction between cohesin rings and a circular minichromosome’, *Cell*, 122(6), pp. 849–860. doi: 10.1016/j.cell.2005.07.018.

James, R. W., Bragg, S. L. and Bragg, W. L. (1948) ‘The Optical Principles of the Diffraction of X-rays’. Bell London.

Jeffries, C. M. *et al.* (2016) ‘Preparing monodisperse macromolecular samples for successful biological small-angle X-ray and neutron-scattering experiments’, *Nature Protocols*. Nature Research, 11(11), pp. 2122–2153. doi: 10.1038/nprot.2016.113.

Juan, D. De, Pazos, F. and Valencia, A. (2013) ‘Emerging methods in protein co-evolution’, *Nature Publishing Group*. Nature Publishing Group, 14(4), pp. 249–261. doi: 10.1038/nrg3414.

Kabsch, W. (1976) ‘A solution for the best rotation to relate two sets of vectors’, *Acta Crystallographica Section A*, 32(5), pp. 922–923. doi: 10.1107/S0567739476001873.

Kakuta, Y. *et al.* (2001) ‘Crystal Structure of Escherichia coli Fdx, an Adrenodoxin-Type Ferredoxin Involved in the Assembly of Iron- Sulfur Clusters’, *Biochemistry*. ACS Publications, 40(37), pp. 11007–11012.

Kappel, C. *et al.* (2010) ‘An unusual hydrophobic core confers extreme flexibility to HEAT repeat proteins’, *Biophysical Journal*, 99(5), pp. 1596–1603. doi: 10.1016/j.bpj.2010.06.032.

Kim, J. H., Tonelli, M. and Markley, J. L. (2012) ‘Disordered form of the scaffold protein IscU is the substrate for iron-sulfur cluster assembly on cysteine desulfurase’, *Proceedings of the National Academy of Sciences*. National Acad Sciences, 109(2), pp. 454–459.

Konarev, P. V. *et al.* (2003) ‘PRIMUS: A Windows PC-based system for small-angle scattering data analysis’, *Journal of Applied Crystallography*. International Union of Crystallography, 36(5), pp. 1277–1282. doi: 10.1107/S0021889803012779.

Konarev, P. V *et al.* (2006) ‘ATSAS 2.1, a program package for small-angle scattering data analysis’, *Journal of applied crystallography*. International Union of Crystallography, 39(2), pp. 277–286.

Kozin, M. B. and Svergun, D. I. (2001) ‘Automated matching of high- and low-resolution structural models’, *J. Appl. Cryst.*, 34, pp. 33–41.

Kratky, O. (1982) ‘Natural high polymers’, in Glatter, O. and Kratky, O. (eds) *Small angle X-ray scattering*. Academic press.

Krebs, W. G. *et al.* (2002) ‘Normal mode analysis of macromolecular motions in a database framework: developing mode concentration as a useful classifying statistic’, *Proteins: Structure, Function, and Bioinformatics*. Wiley Online Library, 48(4), pp. 682–695.

Kschonsak, M. *et al.* (2017) ‘Structural basis for a safety-belt mechanism that anchors condensin to chromosomes’, *Cell*. Elsevier, 171(3), pp. 588–600.

Lensink, M. F., Méndez, R. and Wodak, S. J. (2007) ‘Docking and scoring protein complexes: CAPRI 3rd Edition’, *Proteins: Structure, Function, and Bioinformatics*, 69(4), pp. 704–718. doi: 10.1002/prot.

Lescoute, A. and Westhof, E. (2006) ‘Topology of three-way junctions in folded RNAs’, *Rna*. Cold Spring Harbor Lab, 12(1), pp. 83–93.

Levitt, M., Sander, C. and Stern, P. S. (1985) ‘Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme’, *Journal of molecular biology*. Elsevier, 181(3), pp. 423–447.

Lopéz-Blanco, J. R., Garzón, J. I. and Chacón, P. (2011) ‘iMod: Multipurpose normal mode analysis in internal coordinates’, *Bioinformatics*, 27(20), pp. 2843–2850. doi: 10.1093/bioinformatics/btr497.

Lu, M., Poon, B. and Ma, J. (2006) ‘A new method for coarse-grained elastic normal-mode analysis’, *Journal of chemical theory and computation*. ACS Publications, 2(3), pp. 464–471.

Makowski, L. *et al.* (2012) ‘Multi-wavelength anomalous diffraction using medium-angle X-ray solution scattering (MADMAX)’, *Biophysical journal*. Elsevier, 102(4), pp. 927–933.

Manalastas-Cantos, K. *et al.* (2019) ‘Solution structure and flexibility of the condensin HEAT-repeat subunit Ycg1’, *Journal of Biological Chemistry*. ASBMB, 294(37), pp. 13822–13829.

Marks, D. S., Hopf, T. A. and Sander, C. (2012) ‘Protein structure prediction from sequence variation’, *Nature biotechnology*. Nature Publishing Group, 30(11), p. 1072.

Martinez, L. *et al.* (2009) ‘PACKMOL: a package for building initial configurations for molecular dynamics simulations’, *Journal of computational chemistry*. Wiley Online Library, 30(13), pp. 2157–2164.

Martin, A. C. R. (2005) ‘Mapping PDB chains to UniProtKB entries’, *Bioinformatics*. Oxford University Press, 21(23), pp. 4297–4301.

Mathew, E., Mirza, A. and Menhart, N. (2004) ‘Liquid-chromatography-coupled SAXS for accurate sizing of aggregating proteins’, *Journal of synchrotron radiation*. International Union of Crystallography, 11(4), pp. 314–318.

Mayer, M. P. and Bukau, B. (2005) ‘Hsp70 chaperones: cellular functions and molecular mechanism’, *Cellular and molecular life sciences*. Springer, 62(6), p. 670.

Mendez, R. and Bastolla, U. (2010) ‘Torsional Network Model: Normal Modes in Torsion Angle Space Better Correlate with Conformation Changes in Proteins’, 228103(June), pp. 1–4. doi: 10.1103/PhysRevLett.104.228103.

Metzger, W. *et al.* (2019) ‘Ivermectin for Causal Malaria Prophylaxis: A Randomized Controlled Human Infection Trial’.

Miake-Lye, R. C., Doniach, S. and Hodgson, K. O. (1983) ‘Anomalous x-ray scattering from terbium-labeled parvalbumin in solution’, *Biophysical journal*. Elsevier, 41(3), pp. 287–292.

Miyashita, O., Gorba, C. and Tama, F. (2011) ‘Structure modeling from small angle X-ray scattering data with elastic network normal mode analysis’, *Journal of structural biology*. Elsevier, 173(3), pp. 451–460.

Morcos, F. *et al.* (2011) ‘Direct-coupling analysis of residue coevolution captures native contacts across many protein families.’, *Proceedings of the National Academy of Sciences of the United States of America*, 108(49), pp. E1293-301. doi: 10.1073/pnas.1111471108.

Mustoe, A. M. *et al.* (2011) ‘New insights into the fundamental role of topological constraints as a determinant of two-way junction conformation’, *Nucleic acids research*. Oxford University Press, 40(2), pp. 892–904.

Neuwald, A. F. and Hirano, T. (2000) ‘HEAT repeats associated with condensins, cohesins, and other complexes involved in chromosome-related functions.’, *Genome research*, 10(10), pp. 1445–52. doi: 10.1101/gr.147400.subunits.

Ovchinnikov, S., Kamisetty, H. and Baker, D. (2014) ‘Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information’, *eLife*, 2014(3), pp. 1–21. doi: 10.7554/eLife.02030.

Pabit, S. A. *et al.* (2010) ‘Counting ions around DNA with anomalous small-angle X-ray scattering’, *Journal of the American Chemical Society*. ACS Publications, 132(46), pp. 16334–16336.

Panchal, M. *et al.* (2014) ‘Plasmodium falciparum signal recognition particle components and anti-parasitic effect of ivermectin in blocking nucleo-cytoplasmic shuttling of SRP’, *Cell death & disease*. Nature Publishing Group, 5(1), p. e994.

Panjkovich, A. and Svergun, D. I. (2016) ‘Deciphering conformational transitions of proteins by small angle X-ray scattering and normal mode analysis’, *Physical Chemistry Chemical Physics*. Royal Society of Chemistry, 18(8), pp. 5707–5719. doi: 10.1039/c5cp04540a.

Panjkovich, A. and Svergun, D. I. (2017) ‘CHROMIXS: automatic and interactive analysis of chromatography-coupled small-angle X-ray scattering data’, *Bioinformatics*. doi: 10.1093/bioinformatics/btx846.

Pantos, E. and Bordas, J. (1994) ‘Supercomputer simulation of small angle X-ray scattering, electron micrographs and X-ray diffraction patterns of macromolecular structures’, *Pure and applied chemistry*. De Gruyter, 66(1), pp. 77–82.

Parisien, M. and Major, F. (2008) ‘The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data', *Nature*. Nature Publishing Group, 452(7183), p. 51.

Pernot, P. *et al.* (2013) 'Upgraded ESRF BM29 beamline for SAXS on macromolecules in solution', *Journal of synchrotron radiation*. International Union of Crystallography, 20(4), pp. 660–664.

Petoukhov, M. V. *et al.* (2013) 'Reconstruction of Quaternary Structure from X-ray Scattering by Equilibrium Mixtures of Biological Macromolecules', *Biochemistry*. ACS Publications, 52(39), pp. 6844–6855. doi: 10.1021/bi400731u.

Petoukhov, M. V. and Svergun, D. I. (2015) 'Ambiguity assessment of small-angle scattering curves from monodisperse systems', *Acta Crystallographica Section D Biological Crystallography*. International Union of Crystallography, 71(5), pp. 1051–1058. doi: 10.1107/S1399004715002576.

Petoukhov, M. V and Svergun, D. I. (2005) 'Global rigid body modeling of macromolecular complexes against small-angle scattering data.', *Biophysical journal*. Elsevier, 89(2), pp. 1237–50. doi: 10.1529/biophysj.105.064154.

Porod, G. (1951) 'Die Röntgenkleinwinkelstreuung von dichtgepackten kolloiden Systemen', *Colloid & Polymer Science*. Springer, 124(2), pp. 83–114.

Rajagopalan, S. *et al.* (2013) 'Studies of IscR reveal a unique mechanism for metal-dependent regulation of DNA binding specificity', *Nature structural & molecular biology*. Nature Publishing Group, 20(6), p. 740.

Rambo, R. P. and Tainer, J. A. (2013) 'Accurate assessment of mass, models and resolution by small-angle scattering', *Nature*. Nature Publishing Group, 496(7446), p. 477.

Dos Reis, M. A., Aparicio, R. and Zhang, Y. (2011) 'Improving protein template recognition by using small-angle X-ray scattering profiles', *Biophysical Journal*. Biophysical Society, 101(11), pp. 2770–2781. doi: 10.1016/j.bpj.2011.10.046.

Dos Santos, R. N. *et al.* (2015) 'Dimeric interactions and complex formation using direct coevolutionary couplings.', *Scientific reports*. Nature Publishing Group, 5, p. 13652. doi: 10.1038/srep13652.

Schneidman-Duhovny, D., Hammel, M. and Sali, A. (2010) 'FoXS: a web server for rapid computation and fitting of SAXS profiles', *Nucleic acids research*. Oxford University Press, 38(suppl_2), pp. W540--W544.

Schneidman-Duhovny, D., Hammel, M. and Sali, A. (2011) 'Macromolecular docking restrained by a small angle X-ray scattering profile', *Journal of Structural Biology*, 173(3), pp. 461–471. doi: 10.1016/j.jsb.2010.09.023.

Selvin, P. R. (1995) '[13] Fluorescence resonance energy transfer', in *Methods in enzymology*. Elsevier, pp. 300–334.

Semenyuk, A. V and Svergun, D. I. (1991) 'GNOM--a program package for small-angle scattering data processing', *Journal of Applied Crystallography*. International Union of Crystallography, 24(5), pp. 537–540.

Shi, R. *et al.* (2010) 'Structural basis for Fe--S cluster assembly and tRNA thiolation mediated by IscS protein--protein interactions', *PLoS biology*. Public Library of Science, 8(4), p.

e1000354.

Sievers, F. *et al.* (2011) 'Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.', *Molecular systems biology*, 7(1), p. 539. doi: 10.1038/msb.2011.75.

Sinz, A. (2006) 'Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein--protein interactions', *Mass spectrometry reviews*. Wiley Online Library, 25(4), pp. 663–682.

Stovgaard, K. *et al.* (2010) 'Calculation of accurate small angle X-ray scattering curves from coarse-grained protein models', *BMC bioinformatics*. BioMed Central, 11(1), p. 429.

Stuhrmann, H. B. (1970) 'Ein neues Verfahren zur Bestimmung der Oberflächenform und der inneren Struktur von gelösten globulären Proteinen aus Röntgenkleinwinkelmessungen', *Zeitschrift für Physikalische Chemie*. De Gruyter Oldenbourg, 72(4_6), pp. 177–184.

Stuhrmann, H. B. t and Notbohm, H. (1981) 'Configuration of the four iron atoms in dissolved human hemoglobin as studied by anomalous dispersion', *Proceedings of the National Academy of Sciences*. National Acad Sciences, 78(10), pp. 6216–6220.

Svergun, D. I. (1999) 'Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing', *Biophysical journal*. Elsevier, 76(6), pp. 2879–2886.

Svergun, D. I. *et al.* (2013) *Small angle X-ray and neutron scattering from solutions of biological macromolecules*. Oxford University Press.

Svergun, D. I., Barberato, C. and Koch, M. H. J. (1995) 'CRYSOL--a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates', *Journal of applied crystallography*. International Union of Crystallography, 28(6), pp. 768–773.

Svergun, D. I. and Stuhrmann, H. B. (1991) 'New developments in direct shape determination from small-angle scattering. 1. Theory and model calculations', *Acta Crystallographica Section A: Foundations of Crystallography*. International Union of Crystallography, 47(6), pp. 736–744.

Tama, F. *et al.* (2000) 'Building-block approach for determining low-frequency normal modes of macromolecules', *Proteins: Structure, Function, and Bioinformatics*. Wiley Online Library, 41(1), pp. 1–7.

Tama, F., Miyashita, O. and Brooks III, C. L. (2004) 'Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM', *Journal of structural biology*. Elsevier, 147(3), pp. 315–326.

Tama, F. and Sanejouand, Y.-H. (2001) 'Conformational change of proteins arising from normal mode calculations', *Protein engineering*. Oxford University Press, 14(1), pp. 1–6.

Tirion, M. M. (1996) 'Large amplitude elastic motions in proteins from a single-parameter, atomic analysis', *Physical review letters*. APS, 77(9), p. 1905.

Tobi, D. and Bahar, I. (2005) 'Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state', *Proceedings of the National Academy of Sciences*. National Acad Sciences, 102(52), pp. 18908–18913.

Tuukkanen, A. T., Kleywegt, G. J. and Svergun, D. I. (2016) ‘Resolution of ab initio shapes determined from small-angle scattering’, *IUCrJ*. International Union of Crystallography, 3(6), pp. 440–447.

Uhlmann, F. (2016) ‘SMC complexes: From DNA to chromosomes’, *Nature Reviews Molecular Cell Biology*, 17(7), pp. 399–412. doi: 10.1038/nrm.2016.30.

Valentini, E. *et al.* (2014) ‘SASBDB, a repository for biological small-angle scattering data’, *Nucleic acids research*. Oxford University Press, 43(D1), pp. D357--D363.

Vickery, L. E. and Cupp-Vickery, J. R. (2007) ‘Molecular chaperones HscA/Ssq1 and HscB/Jac1 and their roles in iron-sulfur protein maturation’, *Critical reviews in biochemistry and molecular biology*. Taylor & Francis, 42(2), pp. 95–111.

Volkov, V. V and Svergun, D. I. (2003) ‘Uniqueness of ab initio shape determination in small-angle scattering’, *Journal of applied crystallography*. International Union of Crystallography, 36(3), pp. 860–864.

Wang, Z. and Xu, J. (2013) ‘Predicting protein contact map using evolutionary and physical constraints by integer programming’, *Bioinformatics*, 29(13), pp. 266–273. doi: 10.1093/bioinformatics/btt211.

Wasmuth, E. V and Lima, C. D. (2016) ‘UniProt: the universal protein knowledgebase’, *Nucleic Acids Research*, 45(November 2016), pp. 1–12. doi: 10.1093/nar/gkw1152.

Wild, K. and Sinning, I. (2014) ‘RNA gymnastics in mammalian signal recognition particle assembly’, *RNA biology*. Taylor & Francis, 11(11), pp. 1330–1334.

Wilhelm, L. *et al.* (2015) ‘SMC condensin entraps chromosomal DNA by an ATP hydrolysis dependent loading mechanism in Bacillus subtilis’, *eLife*, 4(MAY), pp. 1–18. doi: 10.7554/eLife.06659.

Yeang, C. and Haussler, D. (2007) ‘Detecting Coevolution in and among Protein Domains’, 3(11). doi: 10.1371/journal.pcbi.0030211.

Yoshimura, S. H. and Hirano, T. (2016) ‘HEAT repeats – versatile arrays of amphiphilic helices working in crowded environments?’, *Journal of Cell Science*, p. jcs.185710. doi: 10.1242/jcs.185710.

Appendix

List of Abbreviations

ASAXS	anomalous small-angle X-ray scattering
CAPRI	Critical Assessment of Prediction of Interactions
CNMA	normal mode analysis in Cartesian space
Ct	*Chaetomium thermophilum*
DCA	direct coupling analysis
DLS	dynamic light scattering
DNA	deoxyribonucleic acid
DTT	dithiothreitol
EMBL	European Molecular Biology Laboratory
ESRF	European Synchotron Radiation Facility
FRET	fluorescence resonance energy transfer
Lrmsd	ligand root-mean-square deviation
MALS	multiangle static light scattering
MD	molecular dynamics
MS	mass spectrometry
MW	molecular weight
NDB	Nucleic Acid Database
NMA	normal mode analysis
NMR	nuclear magnetic resonance
PDB	Protein Data Bank
Pf	*Plasmodium falciparum*
rmsd	root-mean-square deviation
RNA	ribonucleic acid
RTB	rotations-translations of blocks
SANS	small-angle neutron scattering
SAS	small-angle scattering
SAXS	small-angle X-ray scattering
SEC	size exclusion chromatography
SRP	signal recognition particle
SVD	singular value decomposition
TNMA	normal mode analysis in torsion angle space
UniProt	Universal Protein Resource

Large RNA dataset for TNMA benchmarking

initial structure	target structure	Rg, initial	Rg, target	rmsd	transition
1anr_12	1anr_18	17.2	13.9	11.6	open-to-closed
1anr_12	1anr_2	17.2	13.5	11.2	open-to-closed
1anr_16	1anr_18	17.5	13.9	11.9	open-to-closed
1anr_16	1anr_9	17.5	13.8	10.9	open-to-closed
1anr_19	1anr_6	15.8	15.2	10.3	open-to-closed
1anr_6	1anr_18	15.2	13.9	11.9	open-to-closed
1anr_6	1anr_2	15.2	13.5	11.3	open-to-closed
1anr_7	1anr_18	17.0	13.9	11.1	open-to-closed
1ikd_14	1ikd_2	15.5	11.3	10.7	open-to-closed
1m5l_3	1m5l_10	20.3	19.2	10.8	open-to-closed
1m5l_3	1m5l_15	20.3	17.1	10.4	open-to-closed
1m5l_5	1m5l_10	23.3	19.2	11.9	open-to-closed
1m5l_5	1m5l_15	23.3	17.1	12.5	open-to-closed
1s9s_1	1s9s_8	34.6	29.9	14.4	open-to-closed
1s9s_11	1s9s_20	31.4	31.2	12.3	open-to-closed
1s9s_11	1s9s_8	31.4	29.9	11.7	open-to-closed
1s9s_12	1s9s_14	32.0	30.5	26.2	open-to-closed
1s9s_12	1s9s_19	32.0	31.5	22.7	open-to-closed
1s9s_12	1s9s_20	32.0	31.2	26.5	open-to-closed
1s9s_12	1s9s_4	32.0	30.2	24.7	open-to-closed
1s9s_13	1s9s_12	35.6	32.0	23.7	open-to-closed
1s9s_13	1s9s_15	35.6	32.8	12.0	open-to-closed
1s9s_13	1s9s_18	35.6	33.1	30.9	open-to-closed
1s9s_13	1s9s_20	35.6	31.2	11.6	open-to-closed
1s9s_13	1s9s_3	35.6	32.7	32.5	open-to-closed
1s9s_13	1s9s_4	35.6	30.2	13.3	open-to-closed
1s9s_13	1s9s_8	35.6	29.9	14.1	open-to-closed
1s9s_16	1s9s_20	33.0	31.2	12.3	open-to-closed
1s9s_18	1s9s_10	33.1	31.5	30.7	open-to-closed
1s9s_18	1s9s_14	33.1	30.5	32.4	open-to-closed
1s9s_18	1s9s_17	33.1	33.0	11.9	open-to-closed
1s9s_18	1s9s_19	33.1	31.5	29.5	open-to-closed
1s9s_18	1s9s_20	33.1	31.2	33.2	open-to-closed
1s9s_19	1s9s_11	31.5	31.4	13.7	open-to-closed
1s9s_2	1s9s_18	34.9	33.1	13.1	open-to-closed
1s9s_2	1s9s_3	34.9	32.7	16.3	open-to-closed
1s9s_20	1s9s_8	31.2	29.9	12.6	open-to-closed
1s9s_3	1s9s_14	32.7	30.5	34.4	open-to-closed
1s9s_3	1s9s_19	32.7	31.5	30.8	open-to-closed
1s9s_3	1s9s_20	32.7	31.2	34.5	open-to-closed
1s9s_3	1s9s_4	32.7	30.2	33.3	open-to-closed

initial structure	**target structure**	**Rg, initial**	**Rg, target**	**rmsd**	**transition**
1s9s_9	1s9s_12	34.6	32.0	27.3	open-to-closed
1s9s_9	1s9s_18	34.6	33.1	34.0	open-to-closed
1s9s_9	1s9s_20	34.6	31.2	12.6	open-to-closed
1s9s_9	1s9s_3	34.6	32.7	36.2	open-to-closed
1s9s_9	1s9s_5	34.6	33.4	14.5	open-to-closed
1s9s_9	1s9s_8	34.6	29.9	11.2	open-to-closed
2m58_1	2m58_4	26.7	19.4	18.1	open-to-closed
2m58_1	2m58_5	26.7	20.0	18.5	open-to-closed
2m58_1	2m58_7	26.7	24.6	11.5	open-to-closed
2m58_1	2m58_8	26.7	23.7	12.1	open-to-closed
2m58_6	2m58_4	22.7	19.4	11.8	open-to-closed
2m58_6	2m58_5	22.7	20.0	14.8	open-to-closed
2m58_7	2m58_10	24.6	19.5	10.5	open-to-closed
2m58_7	2m58_6	24.6	22.7	13.3	open-to-closed
2m58_8	2m58_5	23.7	20.0	12.5	open-to-closed
2n3q_4	2n3q_6	21.7	21.4	11.2	open-to-closed
2n3q_7	2n3q_6	23.9	21.4	10.9	open-to-closed
2n3q_9	2n3q_18	22.7	21.0	10.3	open-to-closed
2n3q_9	2n3q_4	22.7	21.7	10.1	open-to-closed
2pcv_3	2pcv_1	27.5	23.5	10.7	open-to-closed
2pcv_3	2pcv_2	27.5	22.0	13.8	open-to-closed
2pcv_3	2pcv_4	27.5	19.6	17.2	open-to-closed
2pcv_8	2pcv_4	24.4	19.6	10.4	open-to-closed
2pcv_9	2pcv_4	26.2	19.6	13.5	open-to-closed
6hag_1	6hag_8	18.8	16.4	10.5	open-to-closed
1anr_18	1anr_12	13.9	17.2	11.6	closed-to-open
1anr_18	1anr_16	13.9	17.5	11.9	closed-to-open
1anr_18	1anr_6	13.9	15.2	11.9	closed-to-open
1anr_18	1anr_7	13.9	17.0	11.1	closed-to-open
1anr_2	1anr_12	13.5	17.2	11.2	closed-to-open
1anr_2	1anr_6	13.5	15.2	11.3	closed-to-open
1anr_6	1anr_19	15.2	15.8	10.3	closed-to-open
1anr_9	1anr_16	13.8	17.5	10.9	closed-to-open
1ikd_2	1ikd_14	11.3	15.5	10.7	closed-to-open
1m5l_10	1m5l_3	19.2	20.3	10.8	closed-to-open
1m5l_10	1m5l_5	19.2	23.3	11.9	closed-to-open
1m5l_15	1m5l_3	17.1	20.3	10.4	closed-to-open
1m5l_15	1m5l_5	17.1	23.3	12.5	closed-to-open
1s9s_10	1s9s_18	31.5	33.1	30.7	closed-to-open
1s9s_11	1s9s_19	31.4	31.5	13.7	closed-to-open
1s9s_12	1s9s_13	32.0	35.6	23.7	closed-to-open
1s9s_12	1s9s_9	32.0	34.6	27.3	closed-to-open

initial structure	target structure	Rg, initial	Rg, target	rmsd	transition
1s9s_14	1s9s_12	30.5	32.0	26.2	closed-to-open
1s9s_14	1s9s_18	30.5	33.1	32.4	closed-to-open
1s9s_14	1s9s_3	30.5	32.7	34.4	closed-to-open
1s9s_15	1s9s_13	32.8	35.6	12.0	closed-to-open
1s9s_17	1s9s_18	33.0	33.1	11.9	closed-to-open
1s9s_18	1s9s_13	33.1	35.6	30.9	closed-to-open
1s9s_18	1s9s_2	33.1	34.9	13.1	closed-to-open
1s9s_18	1s9s_9	33.1	34.6	34.0	closed-to-open
1s9s_19	1s9s_12	31.5	32.0	22.7	closed-to-open
1s9s_19	1s9s_18	31.5	33.1	29.5	closed-to-open
1s9s_19	1s9s_3	31.5	32.7	30.8	closed-to-open
1s9s_20	1s9s_11	31.2	31.4	12.3	closed-to-open
1s9s_20	1s9s_12	31.2	32.0	26.5	closed-to-open
1s9s_20	1s9s_13	31.2	35.6	11.6	closed-to-open
1s9s_20	1s9s_16	31.2	33.0	12.3	closed-to-open
1s9s_20	1s9s_18	31.2	33.1	33.2	closed-to-open
1s9s_20	1s9s_3	31.2	32.7	34.5	closed-to-open
1s9s_20	1s9s_9	31.2	34.6	12.6	closed-to-open
1s9s_3	1s9s_13	32.7	35.6	32.5	closed-to-open
1s9s_3	1s9s_2	32.7	34.9	16.3	closed-to-open
1s9s_3	1s9s_9	32.7	34.6	36.2	closed-to-open
1s9s_4	1s9s_12	30.2	32.0	24.7	closed-to-open
1s9s_4	1s9s_13	30.2	35.6	13.3	closed-to-open
1s9s_4	1s9s_3	30.2	32.7	33.3	closed-to-open
1s9s_5	1s9s_9	33.4	34.6	14.5	closed-to-open
1s9s_8	1s9s_1	29.9	34.6	14.4	closed-to-open
1s9s_8	1s9s_11	29.9	31.4	11.7	closed-to-open
1s9s_8	1s9s_13	29.9	35.6	14.1	closed-to-open
1s9s_8	1s9s_20	29.9	31.2	12.6	closed-to-open
1s9s_8	1s9s_9	29.9	34.6	11.2	closed-to-open
2m58_10	2m58_7	19.5	24.6	10.5	closed-to-open
2m58_4	2m58_1	19.4	26.7	18.1	closed-to-open
2m58_4	2m58_6	19.4	22.7	11.8	closed-to-open
2m58_5	2m58_1	20.0	26.7	18.5	closed-to-open
2m58_5	2m58_6	20.0	22.7	14.8	closed-to-open
2m58_5	2m58_8	20.0	23.7	12.5	closed-to-open
2m58_6	2m58_7	22.7	24.6	13.3	closed-to-open
2m58_7	2m58_1	24.6	26.7	11.5	closed-to-open
2m58_8	2m58_1	23.7	26.7	12.1	closed-to-open
2mtj_13	2mtj_6	17.5	19.5	12.8	closed-to-open

initial structure	target structure	Rg, initial	Rg, target	rmsd	transition
2mtj_6	2mtj_13	19.5	24.6	12.8	closed-to-open
2n3q_18	2n3q_9	21.0	22.7	10.3	closed-to-open
2n3q_19	2n3q_8	21.5	23.9	10.1	closed-to-open
2n3q_6	2n3q_4	21.4	21.7	11.2	closed-to-open
2n3q_6	2n3q_7	21.4	23.9	10.9	closed-to-open
2pcv_1	2pcv_3	23.5	27.5	10.7	closed-to-open
2pcv_2	2pcv_3	22.0	27.5	13.8	closed-to-open
2pcv_4	2pcv_3	19.6	27.5	17.2	closed-to-open
2pcv_4	2pcv_8	19.6	24.4	10.4	closed-to-open
2pcv_4	2pcv_9	19.6	26.2	13.5	closed-to-open
6hag_8	6hag_1	16.4	18.8	10.5	closed-to-open

Heterodimers used for DCA benchmark

DIMER PDB ID	SUBUNIT 1	SUBUNIT 2	COMBINED SEQUENCE LENGTH	# RESIDUES IN STRUCTURE	#SEQUENCES FOR DCA
1bh9	TAFII18	TAFII28	335	134	107
1eud	SUCCINYL-COA SYNTHETASE, ALPHA CHAIN	SUCCINYL-COA SYNTHETASE, BETA CHAIN	779	698	3210
1f3u	TRANSCRIPTION INITIATION FACTOR IIF, BETA SUBUNIT	TRANSCRIPTION INITIATION FACTOR IIF, ALPHA SUBUNIT	766	257	99
1fm0	MOLYBDOPTERIN CONVERTING FACTOR, SUBUNIT 1	MOLYBDOPTERIN CONVERTING FACTOR, SUBUNIT 2	231	223	1598
1fs0	ATP SYNTHASE EPSILON SUBUNIT	ATP SYNTHASE GAMMA SUBUNIT	426	347	2769
1h32	DIHEME CYTOCHROME C	CYTOCHROME C	447	394	139
1h6k	CBP80	20 KDA NUCLEAR CAP BINDING PROTEIN	946	805	186
1hcn	HUMAN CHORIONIC GONADOTROPIN	HUMAN CHORIONIC GONADOTROPIN	281	195	34
1jmt	SPLICING FACTOR U2AF 35 KDA SUBUNIT	SPLICING FACTOR U2AF 65 KDA SUBUNIT	715	121	455
1ka9	imidazole glycerol phosphate synthase	imidazole glycerol phosphate synthase	452	446	3058
1mqs	Sly1 Protein	Integral Membrane Protein SED5	1006	604	407
1oo0	Mago nashi protein	CG8781-PA	312	236	284
1ory	flagellar protein FliS	Flagellin	642	159	136
1r6o	ATP-dependent Clp protease ATP-binding subunit clpA	ATP-dependent Clp protease adaptor protein clpS	864	236	1278
1rp3	RNA polymerase sigma factor SIGMA-28 (FliA)	anti sigma factor FlgM	324	312	369
1usu	HEAT SHOCK PROTEIN HSP82	AHA1	1059	378	548
1wpx	Carboxypeptidase Y	Carboxypeptidase Y inhibitor	751	625	128
1wui	Periplasmic [NiFe] hydrogenase small subunit	Periplasmic [NiFe] hydrogenase large subunit	884	799	398
1x3z	peptide: N-glycanase	UV excision repair protein RAD23	761	380	239
1xqs	HSPBP1 protein	Heat shock 70 kDa protein 1	1003	429	144
1y96	Gem-associated protein 6	Gem-associated protein 7	298	171	46
1ykh	RNA polymerase II mediator complex protein MED7	RNA polymerase II holoenzyme component SRB7	362	209	151
1z3e	Regulatory protein spx	DNA-directed RNA polymerase alpha chain	445	186	2352

DIMER PDB ID	SUBUNIT 1	SUBUNIT 2	COMBINED SEQUENCE LENGTH	# RESIDUES IN STRUCTURE	#SEQUENCES FOR DCA
1z5y	Thiol:disulfide interchange protein dsbD	Thiol:disulfide interchange protein dsbE	750	255	340
2apo	Probable tRNA pseudouridine synthase B	Ribosome biogenesis protein Nop10	396	356	113
2bh1	GENERAL SECRETION PATHWAY PROTEIN L	GENERAL SECRETION PATHWAY PROTEIN E,	910	306	226
2blf	Sulfite:cytochrome c oxidoreductase subunit A	Sulfite:cytochrome c oxidoreductase subunit B	513	454	85
2byk	CHRAC-16	CHRAC-14	268	161	242
2ejf	235aa long hypothetical biotin--[acetyl-CoA-carboxylase] ligase	149aa long hypothetical methylmalonyl-CoA decarboxylase gamma chain	384	304	958
2f6m	Suppressor protein STP22 of temperature-sensitive alpha-factor receptor and arginine permease	Vacuolar protein sorting-associated protein VPS28	627	172	111
2f9z	chemotaxis protein CheC	PROTEIN (chemotaxis methylation protein)	362	347	518
2fh5	Signal recognition particle receptor alpha subunit	Signal recognition particle receptor beta subunit	907	312	267
2gsk	Vitamin B12 transporter btuB	protein TONB	853	671	218
2h6f	Protein farnesyltransferase/geranylgeranyltransferase type I alpha subunit	Protein farnesyltransferase beta subunit	816	729	264
2hqs	Protein tolB	Peptidoglycan-associated lipoprotein	603	520	934
2hrk	Glutamyl-tRNA synthetase, cytoplasmic	GU4 nucleic-binding protein 1	1084	298	282
2ido	DNA polymerase III epsilon subunit	Hot protein	330	247	84
2o3b	Nuclease	Sugar-non-specific nuclease inhibitor	409	376	42
2ode	Guanine nucleotide-binding protein G(k) subunit alpha	Regulator of G-protein signaling 8	534	447	283
2pi2	Replication protein A 32 kDa subunit	Replication protein A 14 kDa subunit	391	246	167
2pqr	Mitochondria fission 1 protein	WD repeat protein YKR036C	798	138	95
2q1z	RpoE, ECF SigE	Anti-Sigma factor ChrR, transcriptional activator ChrR	394	361	333
2rd7	Complement component C8 alpha chain	Complement component C8 gamma chain	786	487	79

DIMER PDB ID	SUBUNIT 1	SUBUNIT 2	COMBINED SEQUENCE LENGTH	# RESIDUES IN STRUCTURE	#SEQUENCES FOR DCA
2v3b	RUBREDOXIN REDUCTASE	RUBREDOXIN 2	439	435	1144
2v8s	CLATHRIN INTERACTOR 1	VESICLE TRANSPORT THROUGH INTERACTION WITH T-SNARES HOMOLOG 1B	857	230	97
2wus	ROD SHAPE-DETERMINING PROTEIN MREB	PUTATIVE UNCHARACTERIZED PROTEIN	527	422	165
2xjy	RHOMBOTIN-2	LIM DOMAIN-BINDING PROTEIN 1	569	166	138
2xze	STAM-BINDING PROTEIN	CHARGED MULTIVESICULAR BODY PROTEIN 3	646	158	127
2za4	Ribonuclease	Barstar	247	197	55
2zae	Ribonuclease P protein component 1	Ribonuclease P protein component 4	247	203	48
3a1p	Ribosome maturation factor rimM	30S ribosomal protein S19	255	246	1007
3a8g	Nitrile hydratase subunit alpha	Nitrile hydratase subunit beta	419	407	191
3a8k	Aminomethyltransferase	Glycine cleavage system H protein	493	488	3190
3ajb	Peroxisomal biogenesis factor 3	Peroxisomal biogenesis factor 19	672	319	92
3aji	26S proteasome non-ATPase regulatory subunit 10	Proteasome (Prosome, macropain) 26S subunit, ATPase, 4	649	301	188
3aon	V-type sodium ATPase subunit D	V-type sodium ATPase subunit G	313	279	556
3awu	Tyrosinase	MelC	399	356	72
3ayh	DNA-directed RNA polymerase III subunit rpc9	DNA-directed RNA polymerase III subunit rpc8	332	317	293
3cjs	Ribosomal protein L11 methyltransferase	50S ribosomal protein L11	401	130	2533
3d3b	N utilization substance protein B	30S ribosomal protein S10	242	226	3015
3dbo	Uncharacterized protein	Uncharacterized protein	221	160	64
3dgp	RNA polymerase II transcription factor B subunit 2	RNA polymerase II transcription factor B subunit 5	585	125	100
3dpl	Cullin-5	RING-box protein 1	888	465	532
3e0j	DNA polymerase subunit delta-2	DNA polymerase subunit delta-3	935	551	94
3egv	Ribosomal protein L11 methyltransferase	50S ribosomal protein L11	401	331	2533

DIMER PDB ID	SUBUNIT 1	SUBUNIT 2	COMBINED SEQUENCE LENGTH	# RESIDUES IN STRUCTURE	#SEQUENCES FOR DCA
3fav	ESAT-6-like protein esxB	6 kDa early secretory antigenic target	195	146	305
3fpn	Geobacillus stearothermophilus UvrA interaction domain	Geobacillus stearothermophilus UvrB interaction domain	225	212	3827
3gni	Protein Mo25	STRAD alpha	772	639	100
3h7h	Transcription elongation factor SPT4	Transcription elongation factor SPT5	1204	214	445
3hi2	HTH-type transcriptional regulator mqsA(ygiT)	Motility quorum-sensing regulator mqsR	229	167	64
3htu	Vacuolar protein-sorting-associated protein 25	Vacuolar protein-sorting-associated protein 20	377	110	501
3hzh	Chemotaxis response regulator (CheY-3)	Chemotaxis operon protein (CheX)	307	282	367
3ixs	E3 ubiquitin-protein ligase RING2	RING1 and YY1-binding protein	564	142	75
3kmu	Integrin-linked kinase	Alpha-parvin	824	375	136
3kse	Cathepsin L1	Cystatin-A	431	317	143
3lcb	Isocitrate dehydrogenase kinase/phosphatase	Isocitrate dehydrogenase [NADP]	994	969	317
3lpe	Putative transcription antitermination protein nusG	DNA-directed RNA polymerase subunit E	206	143	278
3mca	Elongation factor 1 alpha-like protein	Protein dom34	982	718	477
3mcb	Nascent polypeptide-associated complex subunit alpha	Transcription factor BTF3	421	112	396
3ml1	Periplasmic nitrate reductase	Diheme cytochrome c napB	1000	908	212
3mp7	Preprotein translocase subunit secY	Preprotein translocase subunit secE	529	441	559
3n7s	Calcitonin gene-related peptide type 1 receptor	Receptor activity-modifying protein 1	609	178	63
3nv0	Nuclear RNA export factor 2	NTF2-related export protein	558	332	160
3ny7	Sulfate transporter	Acyl carrier protein	637	195	2042
3o2p	Defective in cullin neddylation protein 1	Cell division control protein 53	1084	285	370
3oss	TYPE 2 SECRETION SYSTEM, GSPC	TYPE 2 SECRETION SYSTEM, SECRETIN GSPD	962	222	175
3p8b	DNA-directed RNA polymerase, subunit e	Transcription antitermination protein nusG	213	207	180

DIMER PDB ID	SUBUNIT 1	SUBUNIT 2	COMBINED SEQUENCE LENGTH	# RESIDUES IN STRUCTURE	#SEQUENCES FOR DCA
3pge	SUMO-modified proliferating cell nuclear antigen	Proliferating cell nuclear antigen	359	335	642
3r07	Lipoate-protein ligase A subunit 1	Putative lipoate-protein ligase A subunit 2	356	343	158
3t5x	PCI domain-containing protein 2	26S proteasome complex subunit DSS1	469	225	379
3tgo	UPF0169 lipoprotein yfiO	Lipoprotein 34	589	399	210
3uz0	Stage III sporulation protein AH	Stage II sporulation protein Q	501	252	53
3v96	Metalloproteinase inhibitor 1	Stromelysin-2	683	323	82
3vep	Uncharacterized protein Rv3413c/MT3522	Probable RNA polymerase sigma-D factor	511	112	80
3vrd	Flavocytochrome c heme subunit	Flavocytochrome c flavin subunit	630	572	134
3vz9	Uncharacterized protein	Spc24 protein	307	163	40
3zeu	PUTATIVE M22 PEPTIDASE YEAZ	PROBABLE TRNA THREONYLCARBAMOYL ADENOSINE BIOSYNTHESIS PROTEIN GCP	568	569	1534
4a9a	RIBOSOME-INTERACTING GTPASE 1	TRANSLATION MACHINERY-ASSOCIATED PROTEIN 46	714	462	582
4c0o	TRANSPORTIN-3	SERINE/ARGININE-RICH SPLICING FACTOR 1	1171	994	361
4c9b	EUKARYOTIC INITIATION FACTOR 4A-III	PRE-MRNA-SPLICING FACTOR CWC22 HOMOLOG	1319	666	304
4cbu	ACTIN-1	GELSOLIN	1156	482	227
4clq	RIBOSOME BIOGENESIS PROTEIN BMS1	RIBOSOME BIOGENESIS PROTEIN BMS1	1550	402	212
4cvn	PUTATIVE ADENYLATE KINASE	30S RIBOSOMAL PROTEIN S11	317	283	965
4e6n	Metallophosphoesterase	Methyltransferase type 12	1335	625	216
4fou	FimX	Type IV fimbriae assembly protein	806	357	187
4geh	Programmed cell death protein 10	Serine/threonine-protein kinase MST4	628	263	139
4gzr	ESAT-6-like protein 6	ESAT-6-like protein 7	192	135	28
4hi8	Integrin-linked protein kinase	LIM and senescent cell antigen-like-containing domain protein 1	777	242	141
4i0x	ESAT-6-like protein MAB_3112	ESAT-6-like protein MAB_3113	208	158	65

DIMER PDB ID	SUBUNIT 1	SUBUNIT 2	COMBINED SEQUENCE LENGTH	# RESIDUES IN STRUCTURE	#SEQUENCES FOR DCA
4iyp	Immunoglobulin-binding protein 1	Serine/threonine-protein phosphatase 2A catalytic subunit alpha isoform	648	317	188
4jeh	Syntaxin-binding protein 1	Syntaxin-1A	882	777	507
4joi	CST complex subunit STN1	CST complex subunit TEN1	491	259	69
4kbm	DNA-directed RNA polymerase subunit beta	RNA polymerase-binding transcription factor CarD	1340	534	590
4kbq	E3 ubiquitin-protein ligase CHIP	Heat shock cognate 71 kDa protein	949	221	400
4kdi	Transitional endoplasmic reticulum ATPase	Ubiquitin thioesterase OTU1	1107	222	260
4l9p	CaaX farnesyltransferase alpha subunit Ram2	CaaX farnesyltransferase beta subunit Ram1	872	792	149
4lx3	DNA polymerase III, alpha subunit	Nucleic acid binding, OB-fold, tRNA/helicase-type	1326	133	2036
4n6o	legumain	cystatin-M	582	377	80
4nqw	ECF RNA polymerase sigma factor SigK	Anti-sigma-K factor RskA	419	221	208
4onm	Ubiquitin-conjugating enzyme E2 variant 2	Ubiquitin-conjugating enzyme E2 N	297	289	658
4pw9	Putative sulfite oxidase	Putative cytochrome C	512	446	73
4q35	LPS-assembly protein LptD	LPS-assembly lipoprotein LptE	977	908	548
4qjv	DNA-directed RNA polymerase subunit D	DNA-directed RNA polymerase subunit L	353	354	234
4qtt	Multifunctional methyltransferase subunit TRM112	Putative methyltransferase BUD23	410	306	516
4rr2	DNA primase small subunit	DNA primase large subunit	929	602	502
4tps	Sporulation inhibitor of replication protein SirA	Chromosomal replication initiator protein DnaA	594	222	98
4txv	Thiol:disulfide interchange protein TlpA	Cytochrome c oxidase subunit 2	500	317	354
4ue8	EUKARYOTIC TRANSLATION INITIATION FACTOR 4E	4E-BINDING PROTEIN THOR	376	199	52
4un1	PUTATIVE TRANSCRIPTIONAL REGULATOR, ASNC FAMILY	PUTATIVE TRANSCRIPTIONAL REGULATOR, ASNC FAMILY	332	304	255
4uqz	HSIE1	HSIB1	453	282	142
4uzy	FLAGELLAR ASSOCIATED PROTEIN	INTRAFLAGELLAR TRANSPORT PROTEIN IFT52	1101	674	178
4ww7	EKC/KEOPS complex subunit BUD32	EKC/KEOPS complex subunit CGI121	442	409	290

DIMER PDB ID	SUBUNIT 1	SUBUNIT 2	COMBINED SEQUENCE LENGTH	# RESIDUES IN STRUCTURE	#SEQUENCES FOR DCA
4wxa	EKC/KEOPS complex subunit PCC1	EKC/KEOPS complex subunit GON7	211	152	31
4x33	Diphthamide biosynthesis protein 3	Protein ATS1	415	384	89
4x8k	RNA polymerase sigma factor SigA	RNA polymerase-binding protein RbpA	639	154	256
4xax	DNA-directed RNA polymerase subunit beta domain 1	CarD	1283	342	581
4xd9	Ribosome biogenesis protein, putative (AFU_orthologue AFUA_8G04790)	Ribosome biogenesis protein (Rrs1), putative (AFU_orthologue AFUA_7G04430)	549	324	458
4xga	Outer membrane protein assembly factor BamB	Outer membrane protein assembly factor BamA	1202	521	814
4xwj	Regulator of sigma D	Phosphocarrier protein HPr	243	236	264
4xxb	60S ribosomal protein L11	E3 ubiquitin-protein ligase Mdm2	669	193	79
4ygb	Protein ERGIC-53	Multiple coagulation factor deficiency protein 2	656	255	144
4yh8	Splicing factor U2AF 23 kDa subunit	Splicing factor U2AF 59 kDa subunit	733	235	399
4zgn	Cell division cycle protein 123	Eukaryotic translation initiation factor 2 subunit gamma	846	382	356
4zhy	YfiR	YfiB	358	262	112
5bw0	Type II secretion system protein J	Type II secretion system protein I	366	241	279
5by8	Rpf2	Rrs1	549	316	458
5czd	Malonyl-CoA-[acyl-carrier-protein] transacylase	Acyl-carrier-protein	409	378	141
5d6h	CsuC	CsuA/B	457	317	69
5dmb	Flagellar assembly factor FliW	Carbon storage regulator homolog	226	212	604
5dud	YbgK	YbgJ	528	500	942
5eb1	YfiR	YfiB	358	271	112
5f5t	Putative uncharacterized protein	Putative uncharacterized protein	708	279	134
5fvk	VACUOLAR PROTEIN SORTING-ASSOCIATED PROTEIN 4	VPS4-ASSOCIATED PROTEIN 1	640	103	173
5gna	Flagellar protein FliT	Flagellar hook-associated protein 2	589	152	156
5gpy	General transcription factor IIE subunit 1	Transcription initiation factor IIE subunit beta	730	278	159

DIMER PDB ID	SUBUNIT 1	SUBUNIT 2	COMBINED SEQUENCE LENGTH	# RESIDUES IN STRUCTURE	#SEQUENCES FOR DCA
5gxg	Camphor 5-monooxygenase	Putidaredoxin	522	507	629
5hxg	Uncharacterized protein STM1697	Flagellar transcriptional regulator FlhD	351	276	51
5hy7	Putative pre-mRNA splicing protein	YSF3	1308	1228	461
5ja1	Enterobactin synthase component F	Enterobactin biosynthesis protein YbdZ	1365	1297	307
5jca	NADH-dependent Ferredoxin:NADP Oxidoreductase (NfnI) subunit alpha	NADH-dependent Ferredoxin:NADP Oxidoreductase (NfnI) subunit beta	752	750	1274
5jff	Probable adenosine monophosphate-protein transferase fic	Uncharacterized protein YhfG	255	239	26
5jwo	Circadian clock protein kinase KaiC	Circadian clock protein KaiB	626	305	195
5lda	JAMM1	SAMP2	209	182	44
5m72	Signal recognition particle subunit SRP72	Signal recognition particle subunit SRP68	1298	171	377
5o8w	Elongation factor 1-alpha	Elongation factor 1-beta	664	537	375
5o9e	Putative U3 small nucleolar ribonucleoprotein	Putative U3 small nucleolar ribonucleoprotein protein	1082	281	150
5tdy	Flagellar M-ring protein	Flagellar motor switch protein FliG	867	127	861
5tqb	60S ribosomal protein L4-like protein	Assembly chaperone of ribosomal protein L4 (Acl4)	763	572	245
5u9m	Superoxide dismutase [Cu-Zn]	Superoxide dismutase 1 copper chaperone	403	379	369
5uni	NAD(P) transhydrogenase subunit alpha 2	NAD(P) transhydrogenase subunit beta	550	355	1749
5v8w	Integrator complex subunit 9	Integrator complex subunit 11	1258	181	130
5v8z	Endoplasmic reticulum resident protein 29	Calmegin	881	124	101
5wwo	Essential nuclear protein 1	Protein LTV1	946	307	171
5wxl	Ribosome biogenesis protein RPF2	Regulator of ribosome biosynthesis	547	304	497
5wy5	Melanoma-associated antigen G1	Non-structural maintenance of chromosomes element 1 homolog	570	425	60
5xly	Chemotaxis protein methyltransferase 1	Cyclic diguanosine monophosphate-binding protein PA4608	399	388	229

Acknowledgements

Writing this manuscript—and of course, the sheer volume of work that preceded it—has been an experience that has tried my capabilities to their limits (i.e. probably the universal PhD student experience). During this process, I depended on the help and kindness of many people, whom I would like to thank from the bottom of my heart:

- My supervisor, Dmitri Svergun, whose generosity of spirit extends not only to the sharing of his scientific expertise, but also to giving his time and empathy. It has been an honor to work with you.
- The BioSAXS group, past and present: Maxim and Chris, who answered my earliest questions; Sasha, who originally conceptualized the coevolution project, and from whom I learned a lot; Andrey, for help with anomalous scattering concepts; Daniel, for building much of the software infrastructure that made my own work possible; Cy and Melissa for all of the laboratory guidance, manuscript checking and emotional support, I can't overstate how much this has helped me; Martin, for the career advice and for checking the German abstract on this thesis; Clément, for keeping P12 in tiptop shape; Haydyn, for being one of the best people to learn user contacting from; Al and Clemente, for SASBDB development and interesting political news/insights; Stefano and Tobias for helping me out at the beamline several times; Nelly, for being a paragon of good time management, whom I unfortunately failed to emulate; Taia and Dima for the tea, meals, and conversations.
- My collaborators (Komal and Irmi, Marc and Christian, Rita and Annalisa), who trusted me with their samples/data, without whom everything would be merely theoretical.
- The EMBL Hamburg community, particularly the fabulous administration team, who really took care of me when the timings got tight. My fellow predocs, with whom I shared laughter, stress, and gossip.
- The EMBL Heidelberg Graduate Office for patiently answering all my questions, arranging the logistics on the Heidelberg end, and submitting this manuscript in my stead. The Photolab, for quickly printing this (I hope).
- The panel of examiners, who responded so quickly and made room in their schedules for the thesis defense
- My parents and parents-in-law, who provided the extra motivation for me to defend in December, so that they would be around to see it.
- Jeric, for being steadfast and supporting me through the most difficult bits.