# Small Angle Scattering data archivation

Dissertation submitted to
Faculty of Mathematics, Informatics and Natural Sciences
Department of Chemistry of the Hamburg University

Erica Valentini
European Molecular Biology Laboratory (EMBL)-Hamburg Outstation

2015 in Hamburg

The following evaluators recommend the admission of the dissertation:
Prof. Dr. Ulrich Hahn
Dr. Dmitri I. Svergun

Day of oral defense: 26.06.2015.

# Author publications


[1] C. Bruckmann et al. 2015. In preparation.

[2] K. Berg et al. 2015. In preparation.

[3] M. Kachala, E. Valentini, and I. D. Svergun. Application of SAXS for the structural characterization of IDPs. In *IDPbyNMR*. Springer, 2015. In press.

[4] E. Valentini, A. G. Kikhney, G. Previtali, C. M. Jeffries, and D. I. Svergun. SASBDB, a repository for biological small-angle scattering data. *Nucleic acids research*, 43:D357–63, Jan. 2015.

[5] C. Tallant, E. Valentini, O. Fedorov, L. Overvoorde, F. M. Ferguson, P. Filippakopoulos, I. D. Svergun, S. Knapp, and A. Ciulli. Molecular basis of histone tail recognition by human TIP5 PHD finger and Bromodomain of the chromatin remodelling complex NoRC. *Structure(London, England : 1993)*, pp 1–13, Jan. 2015.

[6] M. Varadi, S. Kosol, P. Lebrun, E. Valentini, M. Blackledge, A. K. Dunker, I. C. Felli, J. D. Forman-Kay, R. W. Kriwacki, R. Pierattelli, et al. pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic acids research*, 42:D326–35, Jan. 2014.

# Contents

**Author publications** **ii**

**List of figures** **vii**

**List of tables** **viii**

**List of acronyms** **xi**

**Abstract** **1**

**Zusammenfassung** **2**

**Introduction** **4**
Scope of this thesis . . . . . 7

**1 SAS theory** **8**
1.1 A general outline of a SAS experiment . . . . . 8
1.1.1 Radiation sources . . . . . 9
1.1.2 Sample requirements . . . . . 12
1.1.3 Detector . . . . . 13
1.2 SAS data processing . . . . . 14
1.2.1 SAS resolution . . . . . 15
1.2.2 SAS plots and structural parameters . . . . . 16
1.3 Modeling . . . . . 22
1.3.1 *Ab initio* modeling . . . . . 24
1.3.2 Rigid body modeling . . . . . 27
1.3.3 Polydisperse samples modeling . . . . . 30
1.4 Conclusions . . . . . 31

**2 Database design** **32**
2.1 Requirements . . . . . 33
2.2 Conceptual schema . . . . . 34
2.3 Normalization . . . . . 34
2.4 SQL . . . . . 38

2.5 Conclusion . . . . . 38

**3 PED: a database of structural ensembles of intrinsically disordered and of unfolded proteins. 39**
3.1 Intrinsically disordered proteins and proteins with intrinsically disordered regions . . . . . 39
3.2 IDPs and SAS . . . . . 41
3.2.1 EOM algorithm. . . . . 41
3.3 The protein ensemble database: PED . . . . . 44
3.4 Conclusions . . . . . 46

**4 SASBDB, a repository for biological small-angle scattering data 48**
4.1 Which data are stored in SASBDB . . . . . 49
4.2 Data display in SASBDB . . . . . 51
4.3 Data validation . . . . . 52
4.4 Implementation: database and tools . . . . . 54
4.5 Availability, searching and browsing . . . . . 55
4.6 Web interface . . . . . 56
4.7 Current status and statistics . . . . . 59
4.8 Conclusions and future perspectives . . . . . 60

**5 Examples of SASBDB to report the results of User projects 63**
5.1 SAXS analysis of the Human TIP5 PHD Finger and Bromodomain of the Chromatin Remodeling Complex NoRC . . . . . 64
5.1.1 Introduction . . . . . 64
5.1.2 SAXS analysis . . . . . 65
5.1.3 Conclusions . . . . . 69
5.2 The complex between the homeodomain proteins PREP1 and PBX1 has an elongated structure formed by hydrophobic interactions and stabilised by DNA . . . . . 70
5.2.1 Introduction . . . . . 70
5.2.2 SAXS results . . . . . 71
5.2.3 Conclusions . . . . . 75
5.3 SAXS studies of aptamer constructs . . . . . 77
5.3.1 Introduction . . . . . 77
5.3.2 SAXS analysis . . . . . 78
5.3.3 Conclusions . . . . . 86

**6 Conclusions 88**

**References 90**

**Appendix 98**

**Acknowledgments** **99**

**Declaration upon oath** **100**

# List of Figures


1 Publications in PubMed referring to SAS . . . . . 5

1.1 A schematics of a SAS experiment. . . . . 9
1.2 BSA SAXS scattering data: sample, buffer and subtracted scattering profile. . . . . 10
1.3 Synchrotrons of the world. . . . . 11
1.4 Concentrations series: scattering curves. . . . . 13
1.5 SAXS-WAXS resolution. . . . . 17
1.6 Scattering profiles depending on particle size and shape. . . . 18
1.7 $P(r)$ distribution depending on the particle shape. . . . . 18
1.8 Guinier plots and related scattering plots. . . . . 20
1.9 Kratky plots and related scattering plots . . . . . 22
1.10 Dimensionless Kratky plot . . . . . 23
1.11 DAMMIN implementation. . . . . 26
1.12 CRYSOL-computed scattering curves. . . . . 28
1.13 Rigid body modeling in comparison with *ab initio* modeling. 29

2.1 Entity-relation diagram illustrating a SAXS database . . . . . 35
2.2 Logical schema illustrating a SAXS database. . . . . 36
2.3 UML representation of SASBDB logical database schema. . . 37
2.4 Example of an SQL query. . . . . 38

3.1 IDP number of publications in PubMed by year. . . . . 40
3.2 Kratky plot and a scattering curve from a model IDP. . . . . 42
3.3 EOM algorithm description. . . . . 43
3.4 PED database schema. . . . . 45
3.5 PED ensamble visualization. . . . . 46

4.1 SASBDB entry detail, figure . . . . . 52
4.1 SASBDB entry detail, caption . . . . . 53
4.2 SASBDB entry organization . . . . . 55
4.3 SASBDB home page. . . . . 57
4.4 Brief representation of a SASBDB entry ("browsing unit"). . 58
4.5 SASBDB usage . . . . . 59
4.6 SASBDB download statistics . . . . . 60

5.1 Conservation of the PHD and BROMO domains . . . . . 65
5.2 Cis and trans conformation for PHD and BRD tandem domains. 66
5.3 Comparison of the entries: SASDA46, SASDA56 and SASDA66. 67
5.4 Molecular Basis of Histone Tail Recognition by Human TIP5 PHD Finger and BRD of the Chromatin Remodeling Complex NoRC: SAXS results. . . . . 68
5.5 Homeodomain bound to DNA. PDB code 1AHD . . . . . 71
5.6 Concentration dependence of PBX1:PREP1 and PBX1:PREP1-DNA complex. . . . . 72
5.7 SASBDB entries related to the complex PBX1:PREP1, PBX1:PREP1-DNA and DNA . . . . . 74
5.8 *Ab initio* model reconstruction for PREP1:PBX1 heterodimer and PREP1:PBX1-DNA complex . . . . . 75
5.9 G-quadruplex motif . . . . . 78
5.10 SASBDB entries of long aptamer constructs . . . . . 80
5.11 Aptamer models . . . . . 82
5.12 SASBDB entries of short aptamer constructs . . . . . 83
5.13 SASBDB entries of IL-6R alone and in complex with AIR-3A. 84
5.14 IL-6R tetramer and dimer . . . . . 85
5.15 IL-6R in complex with AIR-3A . . . . . 86

# List of Tables

1 Databases storing SAS data . . . . . 6

2.1 An example relational database table . . . . . 33

4.1 Structural parameters stored in SASBDB . . . . . 50
4.2 Validation methods to be included in SASBDB . . . . . 62

5.1 Structural parameters resulting from SAXS experiment . . . 66
5.2 Structural parameters resulting from SAXS experiment . . . 73
5.3 Affinity and stability of different aptamers . . . . . 78
5.4 Structural parameters resulting from SAXS experiment . . . 81

# List of acronyms

$R_g$ radius of gyration. 16, 18, 43, 45, 49, 51, 52, 54, 68, 70, 82, 87

$V_p$ Porod volume. 16, 18, 19, 52, 87

***I(0)*** intensity at zero-angle. 16–18, 45, 51, 52

***MW*** molecular weight. 9, 10, 16–19, 49, 52, 54, 70, 75, 84, 87

***nm*** nanometer. 10, 11, 14, 70, 78

**1D** one-dimensional. 5, 23, 24

**2D** two-dimensional. 5, 6, 11, 52, 78

**3D** three-dimensional. 5, 23, 24, 45, 52, 55, 58, 62, 80, 91, 92

**API** application programming interface. 55

**AUL** analytical ultracentrifugation. 9

**BAZ2A** BRD adjacent to zinc finger domain protein 2A. 66

**BAZ2B** BRD adjacent to zinc finger domain protein 2B. 66–68, 70, 72

**BMRB** biological magnetic resonance data bank. 45

**BRD** bromodomain. 66, 67, 70, 72

**BSA** bovine serum albumin. 17, 45, 60

**CD** circular dichroism. 74, 78, 80

**CIF** crystallographic information file. 50, 58, 60, 61, 63, 92

**CS** chemical shifts. 45

**DBMS** database managment system. 31, 32

**DDT** dithiothreitol. 8

**DLS** dynamic light scattering. 9

**EM** electron microscopy. 14

**EOM** ensemble optimization method. 30, 39, 42–44, 55, 68, 70

**ERD** entity-relation diagram. 33–35

**FRET** fluorescence resonance energy transfer. 27

**FTP** file transfer protocol. 46

**G-quadruplex** guanine-quadruplex. 80

**IDEAL** intrinsically disorderd proteins with extensive annotations and literature database. 44

**IDP** intrinsically disordered protein. 1, 29, 39–41, 43, 44, 47, 93

**IDR** proteins with intrinsically disordered regions. 29, 39, 41, 43, 44, 67, 93

**IL-6** interleukin-6. 80, 91

**IL-6R** interleukin-6 receptor. 80, 81, 87, 91

**ITC** isothermal titration calorimetry. 74, 78

**MALS** multiple angle light scattering. 9

**MD** molecular dynamics. 3, 44, 85, 93

**NMR** nuclear magnetic resonance. 2, 3, 14, 41, 42, 44, 45, 93

**NoCR** nucleolar remodeling complex. 66, 72

**nt** nucleotide. 65, 80, 82, 83

**PBX1** pre-b-cell leukemia homeobox. 73–75, 78

**PDB** protein data bank. 2, 3, 50, 55, 57, 63, 87

**PED** protein ensemble database. 3, 4, 35, 36, 39, 43–47, 93

**PHD** plant homeodomain. 66, 67, 70, 72

**PKNOX** PBX/knotted 1 homeobox. 73

**PREP1** PBX regulatory protein. 73–75, 78

**PTM** post translational modification. 72

**RDCs** residual dipolar coupling. 45

**rDNA** ribosomal DNA. 66, 72

**rRNA** ribosomal RNA. 66

**SAS** small angle scattering. 1–5, 9, 10, 13, 16, 17, 23, 26, 29, 30, 32, 33, 41, 47–50, 57, 60–63, 65, 92, 93

**SASBDB** small angle scattering data bank. 3, 36, 47–52, 54–57, 60–63, 65, 67, 68, 72, 74, 78, 82, 83, 87, 91–93

**SAXS** small angle X-ray scattering. 1–7, 9, 11, 14, 29, 33, 39, 41–45, 47, 48, 60, 63, 65–68, 70, 72–75, 78, 80–82, 91–93

**SEC** size-exclusion chromatography. 10

**SQL** structured query language. 32, 36, 55

**SVD** single value decomposition. 29

**TELE** three amino acids loop extension. 73, 78, 79

**TIP5** TTF-I interacting protein 5. 66–68, 70, 72

**UML** unified modeling language. 36

**WAXS** wide angle X-ray scattering. 13, 14, 51, 60

**wwPDB SAStf** world wide protein data bank SAS task force. 2, 3, 48, 50, 61, 63, 92

**Abstract**

Small angle X-ray and neutron scattering (SAXS and SANS) are powerful techniques that allow to structurally characterize proteins, nucleic acids and nanoparticles in solution. The usage of small angle scattering (SAS) has grown dramatically in the last fifteen years thanks to developments of both hardware instrumentation and software applications. The growth is evidenced by a rapid increase in the number of scientific publications referring to biological SAS results. The need for public accessible repositories where to store and easily locate SAS data and derived model has become a necessity as also underlined by the world wide Protein Data Bank SAS task force (wwPDB SAStf).

Since the 1970's the most efficient and widespread method to organize data is through relational databases. Two SAS-related databases have been described in this dissertation: the protein ensemble database (PED) and the small angle scattering biological data bank (SASBDB). The first is a repository of structural ensembles of disordered and denatured proteins where the 3D spatial coordinates are stored together with information and primary data about different experimental techniques (nuclear magnetic resonance –NMR–, molecular dynamics –MD– and SAXS) applied to obtain the ensemble. SASBDB has been developed following the recommendation of the wwPDB SAStf of a comprehensive repository of both SAS experimental data and derived models fully searchable, completely browsable and freely accessible for download. Additional stored information include details about experimental conditions, measured molecule, instrument and detector characteristics as well as derived publications and authors also cross-linked to other biological databases.

The dissertation also contains three examples of projects archived in SASBDB where the author contributed both to SAXS experiment and data analysis. These projects include (i) analysis of the flexibility of the human TIP5 and BAZ2B proteins, both involved in gene transcription; (ii) structural characterization of the heterodimer formed by the transcription factors PBX1 and PREP1 that changes conformation when bound to DNA and it is responsible for embryo development and; (iii) SAXS studies of different RNA constructs named aptamers also in complex with their substrate IL-6R implicated in a number of auto-immune diseases like diabetes and rheumatoidis arthitis. These projects are a small percentage of the total number of entries stored in SASBDB (presently, 114 experimental data and 195 models) which is currently the largest repository of SAS data and models. SASBDB is expected to play a major role in the development of a federated system of SAS repositories foreseen by the wwPDB SAStf and to increase the overall quality assurance of SAS data and models for the growing community of structural biologists applying SAS in their research.

**Zusammenfassung**

Kleinwinkelröntgen- und Neutronenstreuung (zu englisch: small angle X-ray scattering (SAXS); small angle neutron scattering (SANS)) sind leistungsstarke Methoden zur strukturellen Charakterisierung von Proteinen, Nukleinsäuren sowie Nanopartikeln in Lösung. In den letzten fünfzehn Jahren, hat eine weite Ausbreitung der Verwendung von Kleinwinkelstreuung (zu Englisch: small angle scattering (SAS)) stattgefunden. Dies ist auf die Weiterentwicklung der Messgeräte sowie der Softwareanwendungen zurückzuführen und ist durch einen raschen Anstieg in der Anzahl wissenschaftlicher Veröffentlichungen belegt, die auf biologische SAS Ergebnisse beruhen. Diese hat zu der Notwendigkeit einer der Öffentlichkeit zugängige Datenbank geführt, um SAS Daten und die daraus abgeleiteten Modelle zu speichern und wieder leicht abzurufen. Dies wurde auch von der weltweiten Protein Daten Bank SAS Arbeitsgruppe (wwPDB SAStf) erkannt.

Seit den Siebzigern besteht die effizienteste und weitverbreitetste Methode zur Organisation von Daten darin, relationale Datenbanken einzuführen. Zwei SAS bezogene Datenbanken werden in dieser Dissertation beschrieben: das Protein Ensemble-Database (PED) und die Small Angle Scattering Biological Data Bank (SASBDB). Die erste ist eine Sammlung von Strukturensemblen ungeordneter sowie denaturierter Proteinen. Hierfür werden die 3D-Raumkoordinaten zusammen mit Informationen und Primärdaten über die unterschiedlichen Techniken die zur Erhaltung des Struckturensembles verwendet wurden (Kernspinresonanz (zu englisch: nuclear magnetic resonance (NMR), molekulardynamischer Simulationen (MD) und SAXS) gespeichert. Die Datenbank SASBDB ist entsprechend der Empfehlung des wwPDB SAStf entwickelt worden und enthält eine umfassende Sammlung sowohl experimenteller SAS Daten als auch die daraus abgeleiteten Modelle. Diese sind komplett navigier- sowie durchsuchbar und werden freizugänglich zum Download angeboten. Zusätzlich gespeicherte Informationen beinhalten weitergehende Details über die Versuchsbedingungen, den gemessenen Molekülen, den Eigenschaften der Messinstrumente sowie der Detektoren. Zusätzlich können Verknüpfungen sowohl zu Veröffentlichungen, die auf diesen Daten beruhen, gesetzt werden als auch zu anderen biologischen Datenbanken.

In dieser Dissertation sind auch drei Beispiele von Projekten beschrieben, die in der SASBDB archiviert sind. Die Autorin selber, hat zur SAS Datenerfassung sowie zu der Datenanalyse beigetragen. Diese Projekte umfassen: (i) die Analyse der Flexibilität der humanen TIP5 und BAZ2B Proteinen, die beide an der Gentranskription beteiligt sind; (ii) die strukturelle Charakterisierung eines für die embryonale Entwicklung mitverantwortlicher Heterodimers, welches von den Transkriptionsfaktoren PBX1 und PREP1 gebildet wird und eine Konformationsänderung nach der Anbindung von DNS durchläuft; (iii) SAXS Studien verschiedener RNS-Konstrukte, auch Aptamere genannt, die im Komplex mit ihrem Substrat IL-6R an einer Reihe von Autoimmunerkrankungen wie Diabetes und rheumatoide Arthitis verwickelt sind. Diese Projekte machen einen kleinen Prozentsatz der gesamten in der

SASBDB gespeicherten Einträge aus. Mit derzeit 114 Versuchsdaten und 195-Modellen stellt die SASBDB derzeit die größte Sammlung von SAS Daten und Modellen dar. SASBDB wird voraussichtlich eine wichtige Rolle in der weiteren Entwicklung des von der wwPDB SAStf vorgesehenen Systems spielen, das durch das Verknüpfen mehrere Datenbanken das Ziel verfolgt, die allgemeine Qualität der SAS Daten und Modelle für die anwachsende Gemeinschaft von Strukturbiologen, die SAS für ihre Forschung anwenden, sicherzustellen.

# Introduction

Small angle X-ray scattering (SAXS) and small angle neutron scattering (SANS) are powerful techniques to structurally characterize proteins, nucleic acids and nanoparticles in solution. They allow shape reconstruction as well as the detection of conformational changes upon alterations to environmental conditions (e.g. changes in pH or temperature) [1].

André Guinier, a 1930s pioneer in the field of small angle scattering (SAS), first observed that the scattering of X-rays from granular particles in the area close to the direct beam (in the small angle region) depends on the size of the disperse particles [2]. In 1950s SAS was used to acquire information about the overall shape of biological macromolecules and in the 1970s, with the development of large scale radiation sources SAS began to be applied more frequently in structural biology [3, 4, 5]. The expansion was mostly due to the possibility to gain a useful insight into the particle shapes in solution in a short time and with plain sample requirements.

Although the conjunction of SAS and biology is already more than 60 years old, it is only in the last decade that SAS has experienced a renaissance as illustrated by the growth in the number of scientific publications referring to SAS stored in PubMed (`http://www.ncbi.nlm.nih.gov/pubmed`; see Figure 1). The reasons for this renaissance are manyfold: SAS can be applied to investigate the shapes of molecules spanning a wide size range (from few $kDa$ to $GDa$) and with different degrees of flexibility, such as intrinsically disordered proteins(IDPs) [6]. Furthermore, the developments in instrumentation e.g. the high brilliance third generation synchrotrons, allow SAXS experiments to be performed in a very short time using only few micro-liters of sample. When combined with advances in automation and software developments, even high throughput SAXS studies of almost any type of biological macromolecular system are possible without the need to be an expert in SAS, paving the way for structural biologists from all backgrounds to use the technique [7]. Software developments, in particular, allow for the determination of the low resolution structure of the molecules *ab initio* and to reconstruct the hybrid models based on SAS combined with other high resolution structural techniques like X-ray crystallography or nuclear magnetic resonance (NMR).

It is worth mentioning that SANS, although being extremely useful to

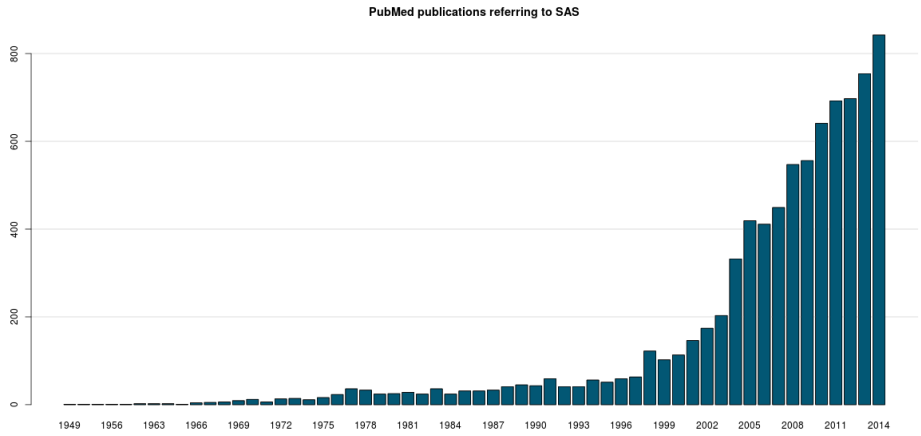


Figure 1: The number of publications stored in PubMed referring to SAS and SAXS. The search has been performed based on the article title or abstract. Using the following search criteria: "((small angle[Title/Abstract] AND scattering[Title/Abstract]) OR SAXS[Title/Abstract])". The word "SANS" returning several false positives was not included in the search criteria.

study macromolecular complexes, may entail more difficult sample preparation (e.g. deuteration of the sample) and longer measurement times due to the inherently low flux of neutron sources compared to X-ray sources. Despite these limitations, computational and instrumental developments for both SAXS and SANS have, overall, resulted in a significant increase in the amount of data produced [8]. The growing amount of SAS data collected and published has lead to the need of a comprehensive and curated repository for SAS data and derived models that is both publicly available and easily accessible and caters to the requirements of the ever-growing SAS community. This need was also underlined by the world wide protein data bank SAS task force (wwPDB SAStf), a selected group of experts in databases, structural biology and SAS [9].

Currently there are four databanks where SAS information are archived (summarized in Table 1). The first and best known is the protein data bank (PDB) [10] that is primarily designed for the storage and dissemination of high-resolution X-ray crystallography and NMR data and models. The PDB contains 47 entries (out of a total of more than 100.000 structures) where SAS was used during the co-refinement process of high-resolution structural models (e.g. PDB code 2A5M [11]). However the PDB, being a repository of high-resolution structures only stores the final models and does not contain primary SAS data or the structural parameters used to build the model. DaRa [12], the second databank, is the first database devoted to

SAXS. It was created to allow users to search for nearest-neighbors similarities between experimental SAS data and scattering patterns calculated from high-resolution models stored in the PDB. In DaRa the models are not stored and users cannot deposit their data. BIOISIS [13] was the first repository developed to store both SAS data and models. The BIOISIS databank is extremely broad in its scope, allowing for the deposition of any SAS project ranging from pure and monodisperse investigations, through to data obtained from completely aggregated samples. At the same time, with a search exclusively based on the BIOISIS ID, without any cross-referencing to external databases, or internal quality assurance mechanisms, BIOISIS –although useful– is limited. Finally, SAXS data and models can also be deposited in the more specialized protein ensemble database (PED) [14], a repository that specifically caters for the storage of structural data about flexible and disordered proteins. Since flexible proteins cannot be described by a single model, PED stores ensembles of structures obtained using different techniques like NMR, molecular dynamics (MD) and SAXS. As PED stores specific types of projects, it cannot be considered –nor does it claim to be– a comprehensive archive for SAXS data.

Table 1: Databases storing SAS data

| Database name | SAS data included | Missing |
|---|---|---|
| wwPDB | 47 models were SAS was used in the refinement stages. | Primary data and structural parameters used to calculate the models. |
| DARA | Scattering curves derived from 20.000 pdb structures. | Derived models and possibility to deposit data. |
| BIOISIS | SAXS data and 97 models. | Complete search, cross references to other databases, quality check on data. |
| PED | 13 scattering curves, ensembles models from IDPs. | SAS data and models from “not IDPs”. |

Despite some attempts to store SAS data and models, a comprehensive repository allowing easy access to both published and unpublished experimental data and models was lacking for the SAS-user community. For this reason we developed the small angle scattering data bank (SASBDB) [15], a resource to store and fetch SAS structural data and models following the standards proposed by the wwPDB SAStf.

## Scope of this thesis

The present dissertation will illustrate the design and implementation of SASBDB including details about database structure, data format, accessibility and user interface (chapter 4). Furthermore, examples of stored data sets will be given (chapter 5). The data sets are the results of three SAXS projects: the first one focusing on different homologous of the same protein showing different degrees of flexibility, the second about DNA-protein complexes and the third dedicated to a specific RNA construct named aptamer. Such examples have been selected because they cover a large spectrum of SAXS applications and the author contributed both in the experimental and in the data analysis phases. Since the author also participated in the development of PED, a description of this database will be detailed in chapter 3. In order to furnish the reader all the means to fully understand the SAS databases, the first chapter of this dissertation will focus on the theory and basics of SAS (chapter 1), while the second will describe the basics of database design (chapter 2).

# Chapter 1

# SAS theory

In a small angle scattering experiment a beam of X-rays or neutrons irradiates a solution of biological molecules, the beam scattered by the solution is then recorded by a detector. The scattering is isotropic and two-dimensional (2D) detector images are reduced to one-dimensional (1D) scattering profiles. These profile (after a number of corrections like background subtraction) are used to acquire information about the molecules such as structural parameters and three-dimensional (3D) shape reconstructions.

In this chapter the general theory of SAS is introduced as applied to experiments performed on biological macromolecules in solution. The chapter is divided in three main sections:

Section 1.1 : experiment design.

Section 1.2 : data analysis and derived structural parameters.

Section 1.3 : modeling processes and software.

## 1.1 A general outline of a SAS experiment

A small angle scattering experiment is conceptually very simple. It requires a radiation source, either X-rays (SAXS) or neutrons (SANS), a sample (consisting of a solution of biological molecules in an appropriate sample holder e.g., a quarts cell or capillary) and a detector. The experiment is illustrated schematically in Figure 1.1.

At the energies used for SAXS (4–30 keV), X-rays scatter from electrons of the sample, while for SANS, scattering primarily occurs as a result of neutron/nuclei interactions. The scattering of the macromolecules in the sample is proportionate to the square of the difference between the total average scattering length density of the macromolecules in the sample relative to the average scattering length density of the solvent (e.g., a buffer; see Figure 1.2). This quantity is called contrast ($\Delta\rho$), which for SAXS relates

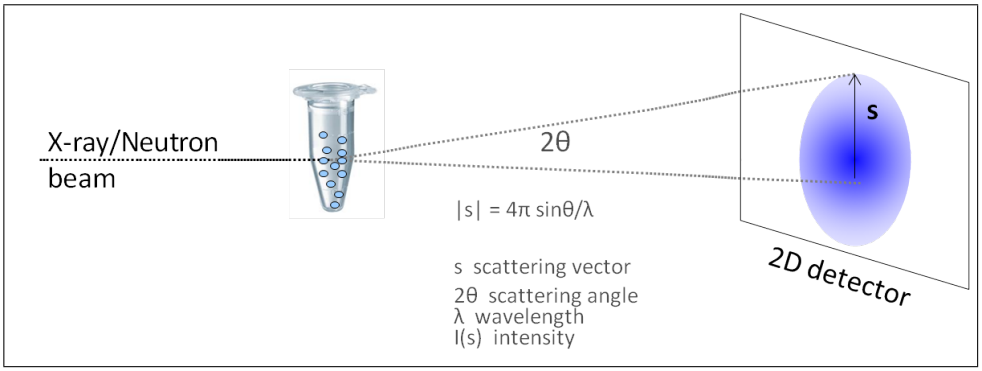


Figure 1.1: A schematics of a SAS experiment. A sample is irradiated by an X-ray or neutron beam. Interactions between the X-rays or neutrons in the sample with atoms cause a portion of the incident beam to scatter at an angle the intensities of which, $I(s)$, are recorded by a 2D detector, where $s = 4\pi \sin\theta/\lambda$ ($\lambda$ is the incident radiation wavelength and $2\theta$ the scattering angle).

to the difference in electron density between a macromolecule and a solvent, while for SANS relates to the nuclear spin-density difference.

SAXS experiments can be performed using synchrotron X-ray radiation or X-ray tubes while SANS experiments are conducted using fission reactor sources or proton accelerator spallation sources. As this dissertation will focus mostly on SAXS measurements using synchrotron sources, this source will be described more in detail. For other source types, the differences from the synchrotron radiation will be outlined. For more information about neutron sources and X-ray tubes please refer to Chapter 3 of Svergun et al. [16].

### 1.1.1 Radiation sources

Synchrotrons are large facilities where charged particles, electrons or positrons, circulate inside storage rings at relativistic velocity generating X-rays as a by-product. Bending magnets, wigglers and undulators are used to optimize the radiation while constant or injection-based refilling is performed to compensate for the particle loss. Currently there are fifty operational synchtrotrons located worldwide (`www.lightsources.org`, illustrated in Figure 1.3) twelve of which have SAXS beam lines suitable or dedicated for biological investigations [17]. The modern third generation facilities like PETRA III in Hamburg (Germany) or ESRF in Grenoble (France), present the great advantage of having high brilliance beams which allow short measurement times (in the range of milliseconds) combined with low sample volume requirements (in the microliters range). Together with the advances in automation and software (described in section 1.3), these facilities have led to

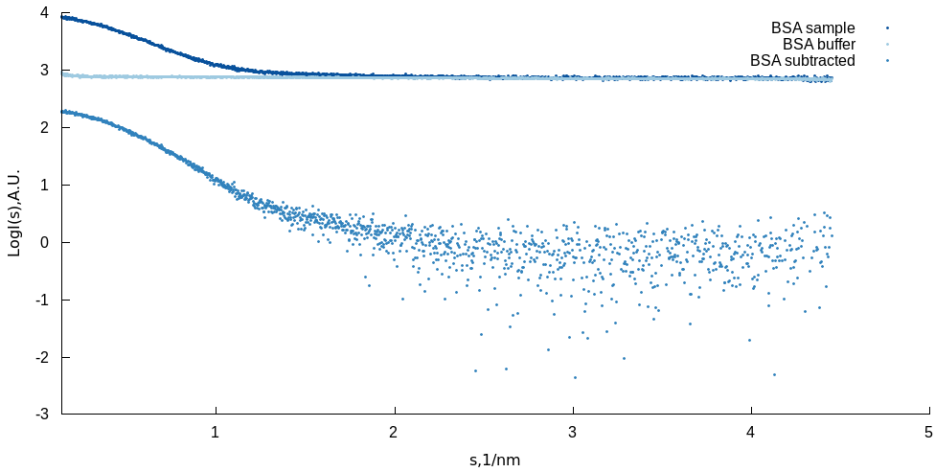


Figure 1.2: BSA SAXS scattering data: sample, buffer and subtracted scattering profile. In this Figure the scattering curves from BSA sample (in dark blue), buffer (in light blue) and subtracted curve (in medium blue) are shown in the same plot ($\log_{10} I(s)$ vs. $s$ where $I(s)$ is in arbitrary units $A.U.$ and $s$ is in inverse nanometers $nm^{-1}$). The subtracted curve has been scaled for visualization purposes. In a SAS experiment the scattering $I(s)$ of the buffer is subtracted from the $I(s)$ of the sample in order to obtain $I(s)$ from the solute alone.

developments in high-throughput operations [13] and time-resolved SAXS experiments. The latter can access sub-millisecond time scales and have been applied to monitor assembly processes and proteins folding/unfolding mechanisms [18].

The high brilliance of the synchrotron radiation, although being very effective, presents one mayor disadvantage for solution-based biological studies. On irradiating a sample at the X-ray energies and fluxes encountered at synchrotron sources, water undergoes photolysis into free radicals and solvated electrons which chemically activate biological macromolecules. As a consequence, unspecific bonds are formed between macromolecules resulting in the formation of large aggregates. These aggregates severely affect the scattering intensities in the small angle region and prevent a reliable analysis of the data. For proteins in particular, possible solutions to circumvent radiation damage is to add glycerol or dithiothreitol (DDT) to the sample. Another solution is to continuously flow the sample during measurement. Although effective, these solutions come with drawbacks. Small-molecule addition to the solvent can chemically change the sample environment, that may adversely affect the structure of a target molecule of interest, while continuous sample flow increases the amount of sample required for an experiment. Additional alternatives to reduce the influence of radiation damage


Figure 1.3: Synchrotrons of the world. The total number of synchrotrons presently in the world is fifty, of which the locations of thirty are displayed above (extracted from the url: `http://www.veqter.co.uk/assets/drgalleries/93/big_map-of-world---synchrotron.png`).

are to attenuate the incident beam or to reduce exposure time. By slicing the data collection process into several short successive frames, radiation damage effects can be monitored via a comparison of the different frames. As a practical example at the EMBL P12 SAXS beam at PETRA III [19] – where most of the data presented in the last chapter of this dissertation were collected– a common practice is to measure data in 20 successive frames, using 50 $ms$ exposures with continuous flow enabled. Using this approach 30 microliters of sample are required, which for most user applications is sufficient to avoid radiation damage.

While synchrotron facilities offer fast SAXS measurements, in-house machines afford the opportunity to perform experiments "in house". Laboratory instruments usually consist of X-ray tubes produced by several firms (e.g. Anton Paar or Bruker) which accelerate charged particles between a catode and an anode. The low incident flux requires measurement times of minutes to hours and the effects of radiation damage are generally reduced. Finally, neutron radiation sources for SANS can only be accessed at large facilities, for example the ILL in France or spallation sources like SINQ in Switzerland. When compared to X-ray sources, neutron fluxes are much lower requiring larger beam sizes and thus larger sample volumes and longer measurement times. However, one of the advantages of SANS is that neu-

trons interact primarily with the nuclei of atoms and do not induce chemical changes, and radiation damage is thus unlikely to occur.

### 1.1.2 Sample requirements

In order to obtain meaningful results from a SAS experiment where the aim is to extract structural parameters and restore the shape of a macromolecule in solution the sample has to fulfill three main requirements: (i) it has to be pure and monodisperse, (ii) interparticle interactions must be kept to a minimum and, (iii) the solvent scattering contributions have to be accurately subtracted from the sample scattering, requiring the measurement of exactly-matched solvent blanks.

The monodispersity is intended as sample purity of at least 95% and it is essential because aggregates or the presence of high-molecular weight contaminants complicate data analysis. The reason is that, as already explained for the radiation damage effects, large particles scatter more intensely (proportionate to the square of the particle volume) compared to small particles to affect the signal intensities in the small angle range where most of the information regarding the shape are located. Monodisperse proteins should migrate as a single species on a native gel (a single bar on a native gel corresponds to a single molecular weight species) and the purity can be assured using multiple angle light scattering (MALS), dynamic light scattering (DLS), size exclusion chromatography and analytical ultracentrifugation (AUL). SAS can be applied in specific cases of polidispersity like mixtures of proteins at different oligomerization states or flexible proteins, and both cases will be explained later in this chapter (see section 1.3).

Different concentrations are needed because, ideally, the macromolecules in the sample should not interact with each other that otherwise causes "across particle", or interparticle interference contributions to the scattering. In order to achieve this conditions the solution should be infinitely diluted, such level of dilution is impossible to obtain in the real world. This problem can be overcome by measuring dilution series of different concentrations and then extrapolating them to infinite dilution. As an example, for a protein of about 50 $kDa$ a concentration series of 10, 5, 2 and 1 $mg/ml$ is recommended (see Figure 1.4). The concentrations should be exactly measured in order to allow a correct determination of the molecular weight ($MW$) (the procedure to obtain the $MW$ will be explained in the next section). The concentration can be determined using UV or visible light extinction coefficients (based on the Beer-Lambert law), which for many proteins can be assessed at absorbance at 280 nanometer ($nm$) ($A_{280}$). Importantly, different concentrations are also needed because the net signal depends on the number of particle in the illuminated volume. Therefore on one hand, the higher is the solute concentration the stronger is the signal but also higher is the risk of interparticle interaction effects. On the other hand, lower

solute concentrations will lead to weak and noisy signal but will exclude interparticle interactions effects.

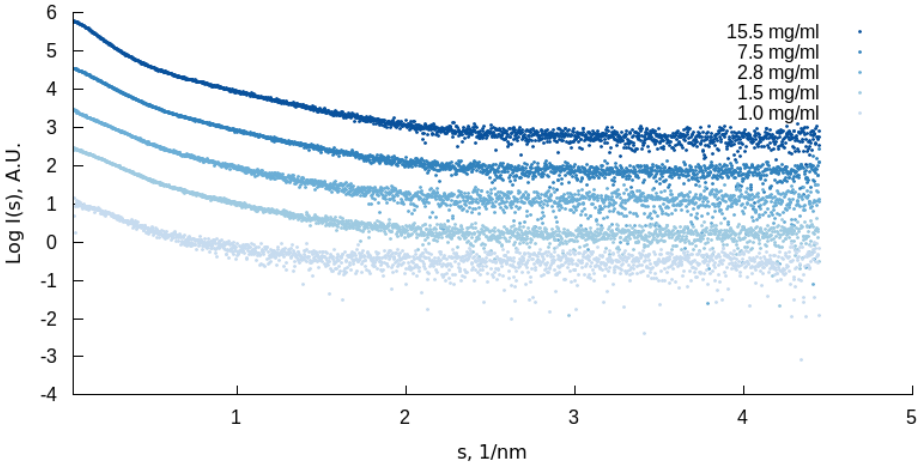


Figure 1.4: Concentrations series: scattering curves. The scattering curves from the same protein complex of 73 kDa are compared. The different concentrations have been scaled for visualization purposes. The plot is $\log_{10} I(s)$ vs. $s$ where $I(s)$ is in arbitrary units $A.U.$ and $s$ is in inverse $nm$ $nm^{-1}$). The system will be described in the chapter 5 of this dissertation. The data were measured at the EMBL-P12 SAXS beam line at the PETRA III synchrotron, DESY.

Finally, as explained before, all the atoms of the illuminated volume scatter when they encounter the incident beam. For this reason, the scattering of the solvent (buffer) must be subtracted from that of the solution (sample) to obtain the pure scattering by the molecules in solution (see Figure 1.2). Hence, each SAS sample measurement has to be followed by a measurement of the corresponding solvent-blank that is exactly-matched to the sample solvent. The best ways to obtain the exact solvent match include dialysis and size-exclusion chromatography (SEC). For the same reason, it is also important that the electron density of the buffer differs from the one of the macromolecules dissolved in it and this is one of the reason why aqueous solutions are commonly used as buffer for SAS experiments.

### 1.1.3 Detector

The type of beam needed for SAS experiments is the monochromatic beam (except for neutrons spallation sources) where the wavelength ($\lambda$) is fixed and the scattering intensities are derived from elastic scattering events. In elastic scattering there is no transfer of energy between the radiation and the scattering centers (electrons and nuclei) and therefore the wavelength of

the incident and scattered beam are the same.

The detector collects the X-ray photons or neutrons scattered by the sample and records their intensities. The SAXS detector is usually a 2D plate made of pixels where the scattered radiation counts are ultimately converted into an electrical signal. Importantly, a beam stop has to be positioned corresponding to the coordinates of the direct beam to avoid both damages of the detector damage and to mask incident beam intensities at low angle that could affect the measurement of the very-weak sample scattering intensities masking the signal coming from the sample which is much weaker than the signal from the direct beam.

## 1.2 SAS data processing

If the sample has been prepared correctly it will contain macromolecules that are (at the resolutions measured for a SAS experiment) structurally identical and randomly oriented with respect to each other without significant interparticle interactions. In such a case the background-corrected scattering intensities will be proportional to the intensity of a single particle in solution. The intensity $I(s)$ is recorded as a function of the momentum transfer (or scattering vector) $s$ (sometimes also denoted as $q$) ($\mathbf{s}$ was already defined in the caption of Figure 1.1). $I(s)$ is expressed as:

$$I(s) = \langle I(s) \rangle_\Omega = \langle A(s) A*(s) \rangle_\Omega \tag{1.1}$$

The angle brackets $\langle \cdots \rangle_\Omega$ mean that the net $I(s)$ derived from the scattering centers within a particle is rotationally averaged over all orientations. The scattering amplitude $A(s)$ (which cannot be directly measured in a scattering experiment but only the intensity, proportional to the number of scattering photons) is defined as:

$$A(\mathbf{s}) = \int_V \Delta\rho(\mathbf{r}) \exp(i\mathbf{s} \cdot \mathbf{r}) \, \mathrm{d}\mathbf{r} \tag{1.2}$$

Where the integration is performed over the particle volume $V$ and each atom pair, connected by the vector $\vec{r}$, emits a spherical wave whose form is expressed as $\exp(i\mathbf{s} \cdot \mathbf{r})$. $A(s)$ is related through a Fourier transform to the excess of electron density $\Delta\rho$(r):

$$\Delta\rho(\mathbf{r}) = \rho(\mathbf{r}) - \rho_s \tag{1.3}$$

Where $\rho(\mathbf{r})$ is the electron density of the macromolecule and $\rho_s$ is the electron density of the solvent.

For solutions of identical particles, $I(s)$ can be expressed as:

$$I(s) = \sum_{i}^{n} \left[ (\Delta\rho_i V_i)^2 \, P_i(s) \right] S(s) \tag{1.4}$$

Where $n$ is the total number of scattering particles in the illuminated sample volume and $P_i(s)$ and $S(s)$ are respectively the form factor and the structure factor terms. The form factor is the scattering depending on the single particle size and shape. The structure factor, instead, is the scattering arising from correlated distances between particles in the irradiated volume, i.e., interparticle interference. Equation 1.4 means that the $I(s)$ is proportional to (i) the summed contribution of all the scattering particles, i.e. to the concentration; (ii) the square of the difference in excess of average scattering length between a macromolecule and a solvent, i.e to the squared contrast $\Delta\rho^2$; (iii) the particle volume squared and; (iv) the product of the form and structure factors.

If the sample is sufficiently dilute, monodisperse and randomly oriented, the contribution of the structure factor $S(s)$ is unitary. In such cases the $S(s)$ amplitudes are negligible and the scattering intensity is proportional only to the scattering by a single particle averaged all over the orientations, weighted by the contrast and volume squared:

$$I(s) = \langle P(s) \rangle \tag{1.5}$$

This explains the sample requirements of monodispersity, concentration series and matching buffer illustrated in the previous section.

### 1.2.1 SAS resolution

Although the scattering profile contains information about the single particle, the resolution of SAS is limited due to the random orientation of the particles in solution. Vector length information is preserved in the data ($P(s)$), but the relative spatial orientations of the vectors are lost. For this reason SAS is defined as low resolution technique.

Such a resolution is due to the fact that the intensity of an object of size $d$ is contained in the momentum transfer $s = 2\pi/d$, therefore giving a maximum resolution:

$$\delta = 2\pi/s_{max} \tag{1.6}$$

While this equation is applicable for X-ray crystallography where the particles are stacked in a fixed lattice and the resolution is close to the wavelength $\delta \sim \lambda$, it does not apply to solution techniques where the particles are randomly oriented in the space and the intensity rapidly decays in the higher angles. Therefore the maximum resolution that SAS can achieve is $\delta \gg \lambda$.

Long distances occurring within the particle will produce scattering intensities at low angles and vice versa. This means that information about the overall particle size is represented in the low-angle regime while distances representing secondary structure and other higher-resolution features are represented in higher angles. For the latter analysis wide angle X-ray scattering (WAXS) was developed (for $s > 6nm^{-1}$). However, being a solution technique, the information from WAXS is also limited by isotropic tumbling of the sample. At present, WAXS is used mainly to detect small differences (e.g. changes in the secondary structures) but is not suitable for generating atomic-scale models like other high resolution techniques (the different resolutions are illustrated in Figure 1.5). Still, WAXS has progressed in the last years especially thanks to its applications in conjunction with high resolution techniques [17].

### 1.2.2 SAS plots and structural parameters

The distribution of inter-atomic distances in a particle weighted by their contrast is given by the pair distance distribution function ($p(r)$ vs. $r$). Such a plot illustrates the frequency of distances in real space and can furnish information about the particle shape. Both the $p(r)$ distribution and the scattering curve ($I(s)$ vs $s$) plots change significantly depending on the particle shape and size. The calculated scattering for a set of archetypal shapes like sphere, dumbbell or elongated particle are shown in Figure 1.6 and and the corresponding real-space probable pair-distance distributions in Figure 1.7.

The $p(r)$ distribution is related to the scattering curve $I(s)$ through an inverse Fourier transform:

$$I(s) = 4\pi \int_0^{D_{max}} p(r) \frac{\sin(sr)}{sr} \, \mathrm{d}r \tag{1.7}$$

$$p(r) = \frac{r^2}{2\pi} \int_0^{\infty} s^2 I(s) \frac{\sin(sr)}{sr} \, \mathrm{d}s \tag{1.8}$$

Where $D_{max}$ is the maximal distance inside the particle and $p(r) = 0$ at long vector lengths ($r \gg D_{max}$). This parameter is one of the model-independent structural values that can be directly obtained from a SAS experiment. The other parameters include: radius of gyration ($R_g$), intensity at zero-angle ($I(0)$) –which is related to the molecular weight ($MW$) of the particle– and its excluded so-called Porod volume ($V_p$) [3]. Some of this parameters are provided by different transformations of the scattering curves: e.g. $I(0)$ and $R_g$ can be calculated from the $p(r)$ distribution but also from the Guinier plot [2], while the $V_p$ can be calculated from the Kratky

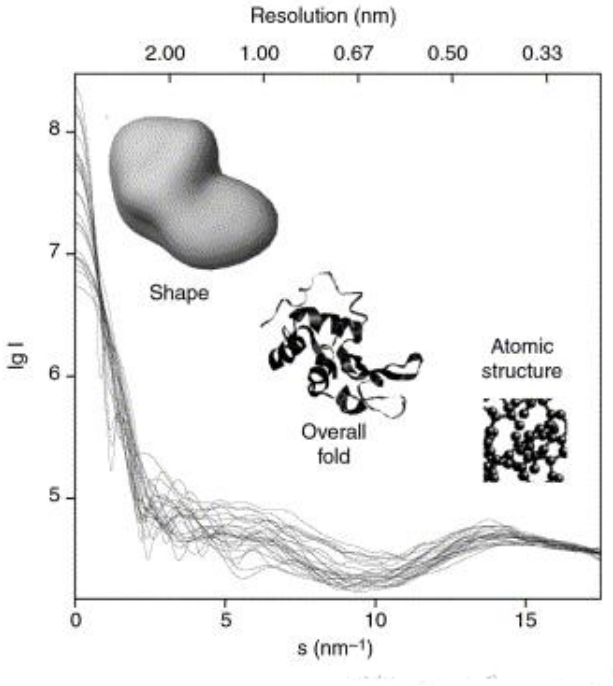


Figure 1.5: SAXS-WAXS resolution. The SAXS measurements (up to 6 $nm^{-1}$) reach a resolution up to 1 nm from which information about the overall particle shape can be obtained. The WAXS can access distances of 1-0.5 $nm$ because of momentum transfer up to 18 $nm^{-1}$. Such a resolution allows one to gain information about secondary structures and overall fold. The atomic resolution can be achieved only with X-ray crystallography, NMR and sometimes EM where the orientations are not averaged. Picture extracted from [20] where the theoretical SAXS curves were computed from the atomic models of 25 different proteins ranging in size between 10 and 300 kDa.

plot [4]. This set of plots and derived parameters will be described in this section.

As already stated, *I(0)* and $R_g$ can be both calculated from the $p(r)$ distribution. The intensity at zero-angle (*I(0)*) is defined as the contrast-weighted number of scattering centers in the squared volume and it can be derived from Equation 1.7 as:

$$I(0) = 4\pi \int_0^{D_{max}} p(r)\,\mathrm{d}r = (\Delta\rho)^2 V^2 \tag{1.9}$$

Where $\Delta\rho$ is the excess scattering density and $V$ is the particle volume. Therefore *I(0)* is proportional to the number of excess electrons in

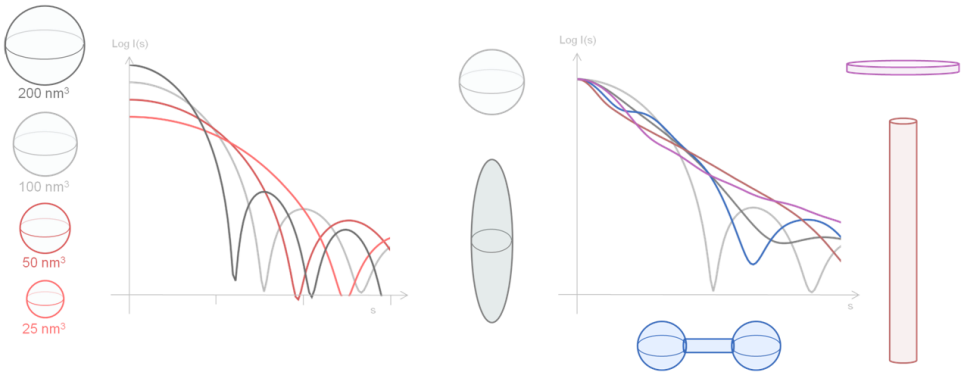


Figure 1.6: The calculated scattering profile ($\log I(s)$ vs $s$) of different particle size and shape. The plot on the left panel shows six particles of different size: the larger particle has an higher intensity at zero angle $I(0)$ and the curve decays steeper than smaller particles. While the plot on the right panel shows that scattering curves of elongated shapes appears to decay gradually, globular and dumbbell shapes present several distinct minima. Figure courtesy of Alexey Kikhney.

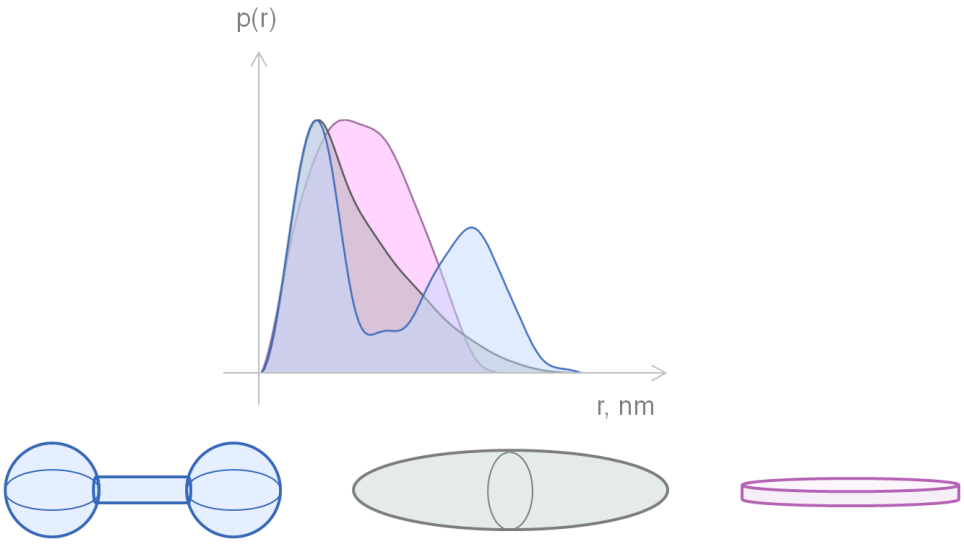


Figure 1.7: The probable real-space pair-distance distribution function $p(r)$ present a different trend according to the particle shapes. Namely, a dumbbell particle will generate a $p(r)$ distribution with multiple peaks while spherical shapes will generate a single peak $p(r)$ that slowly decays to zero. Figure courtesy of Alexey Kikhney.

the particle and can be used to calculate its $MW$. Having the experimental $MW$ allows the identification of oligomerization as well as non-specific ag-

gregation, which can be used to guide the modeling process and assess the quality of the data. In practice, *MW* is obtained from *I(0)* by comparison of the experimental forward scattering value with that of a known standard with known concentration. Such a standard can be a protein e.g. bovine serum albumin (BSA), lysozyme or a macromolecule with the same scattering length density as the sample in question. In order to have a reliable estimate the standard should be measured before any set of SAS experiments at exactly the same experimental conditions (e.g. wavelength, temperature or sample-detector-distance).

When a macromolecule standard is used to calibrate the sample scattering, the $MW_{particle}$ can be obtained with a simple proportion:

$$\frac{MW_{particle}}{I(0)_{particle}} = \frac{MW_{standard}}{I(0)_{standard}} \tag{1.10}$$

$$MW_{particle} = \frac{MW_{standard}}{I(0)_{standard}} I(0)_{particle} \tag{1.11}$$

Another option is to use water as standard; this procedure allows one to measure the $MW_{particle}$ placing the scattering curve on an absolute scale (in untis of $cm^{-1}$) instead of arbitrary scale and unit:

$$MW_{particle} = \frac{I(0)_{particle} N_A}{c_{particle}(\Delta\rho\nu)^2} \tag{1.12}$$

Where $N_A$ is the Avogadro number, $c$ (in $g/cm^3$) is the particle concentration, $\Delta\rho$ is the contrast (in $e/cm^3$)and $\nu$ is the partial specific volume (in $cm^3/g$).

The *I(0)* for *MW* determination, can be estimated also from the Guinier plot ($\ln I(s)$ vs. $s^2$). Such a plot shows a linear trend in the lower angular range ($s_{min}R_g < 1.3$) but only for samples that do not present inter-particle interactions [2]. This property it is extremely important because it allows one to estimate the sample quality: in case of aggregated samples the Guinier plot has an upward trend in the lower angle while, in case of repulsive interaction between particles the trend is downward. Since the *I(0)* is obtained from the intercept of the y-axis of the Guinier plot, in aggregated samples the *I(0)* (and therefore the *MW*) is overestimated while the opposite occurs for repulsive samples where the *I(0)* (and the *MW*) are underestimated (see Figure 1.8).

Both the Guinier plot and the $p(r)$ distribution allow one to obtain another important structural parameter, the radius of gyration ($R_g$). The $R_g$ is defined as the average of square center-of-mass distances in the particle weighted by the scattering length density.

The Guinier approximation:

$$I(0) = I(0)\mathrm{e}^{\frac{-s^2R_g^2}{3}} \tag{1.13}$$

demonstrates the relation between the scattering intensity *I(0)* and the $R_g$ and it is valid only in the low angular range. Indeed, the slope of the linear Gunier plot corresponds to the $R_g$ value. And, as explained for the *I(0)*, it can be overestimated or underestimated according to the sample quality (see Figure 1.8).

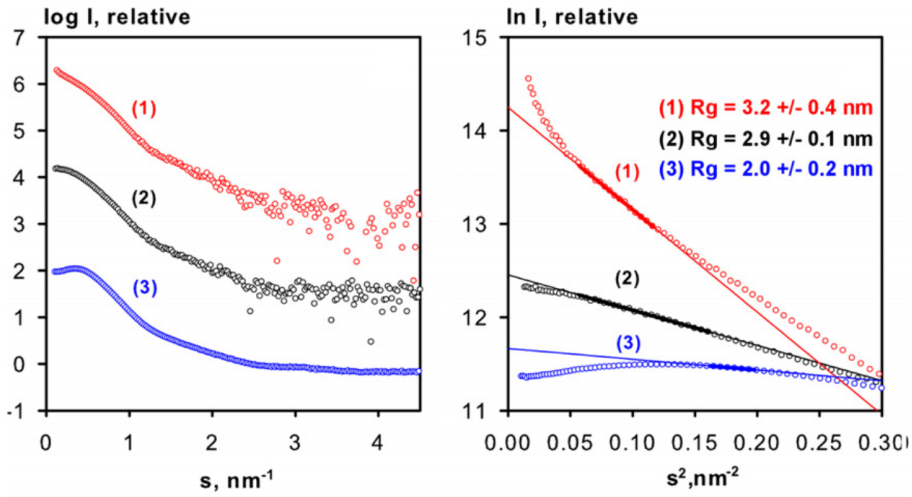


Figure 1.8: Guinier plots and related scattering plots. The Guinier plot ($\ln I(s)$ vs. $s^2$) on the right panel is compared to the corresponding scattering plot ($I(s)$ vs. $s$) on the left. The red and the blue curves present interparticle interactions, namely aggregation (the red) and repulsion (the blue). The interparticle interactions are also clear by the overestimation and underestimation of the $R_g$ and $I(0)$ calculated by the curve slopes, respectively. The Guinier plot is displayed in the $s_{min} - s_{max}$ range. Figure extracted from [21].

The $R_g$ can also be estimated from the $p(r)$ distribution through the following relationship:

$$R_g^2 = \frac{\int_0^{D_{max}} p(r) r^2 \mathrm{d}r}{2 \int_0^{D_{max}} p(r) \mathrm{d}r} \tag{1.14}$$

This second method to estimate the $R_g$ is often more reliable because it considers the entire angular range and not only the data points in the small angle region.

Yet another structural parameter is the Porod volume ($V_p$) [3], which is the excluded geometrical volume of the hydrated particle. Assuming a

constant scattering density inside a macromolecule and sharp boundaries with the solvent, the Porod volume is defined as:

$$V_p = \frac{2\pi^2 I(0)}{Q} \tag{1.15}$$

Where $Q$ is the so-called Porod's invariant that can be calculated from:

$$Q = \int_0^\infty I(s)s^2 \mathrm{d}s \tag{1.16}$$

From here one can estimate that the Porod's invariant corresponds to the area under the curve $s^2I(s)$ vs. $s$ calculated from zero to infinity. Different approaches were developed to assess $Q$ from the finite experimental range and thus to calculate $V_p$ [22].

The Porod volume allows one to estimate the $MW$ of a particle without knowing the exact sample concentration and that can prove crucial whenever the concentration is difficult to determine (e.g. because of the absence of aromatic resudes). Indeed it has been estimated that for globular proteins $MW = V_p/(1.6 or 1.8)$ [7].

The Kratky plot $s^2I(s)$ vs. $s$ can further be useful for a qualitative assessment of the flexibility of a macromolecule in solution. Indeed the appearance of a Kratky plot depends on the compactness of a particle: globular particles produce the plots with a single well-defined maximum while flexible particles (e.g. a random coiled coil) present a plateau with a slower intensity decay (see Figure 1.9). This different behavior was analyzed by both Porod [5] and Debye [23] who stated that the scattering intensity of globular particles should decay proportional to $s^{-4}$ while coiled coils decay as $s^{-2}$.

A different version of the Kratky plot is the normalized (or dimensionless) Kratky plot where the intensities $I(s)$ on the y-axis are divided by $I(0)$ and multiplied by $(sR_g)^2$ and the $s$ on the x-axis is multiplied by $R_g$. Such a plot enables the identification of different scaling relationships in the intensity decay at increasing angles, e.g., the effects of particle flexibility but with the advantage of being independent from the particle size. For example, the peak of globular particle should approximate that of a sphere corresponding to $sR_g = \sqrt{3}$ and $(sR_g)^2 I(s)/I(0) = 1.104$, also random coils, at values above $sR_g = \sqrt{3}$, generate a plateau in the dimensionless Kratky plot at values of $(sR_g)^2 I(s)/I(0) = 1.5 - 2$ [24] as showed in Figure 1.10.

The described set of curves and derived parameters is used as input for the modeling software that will be illustrated in the following section.

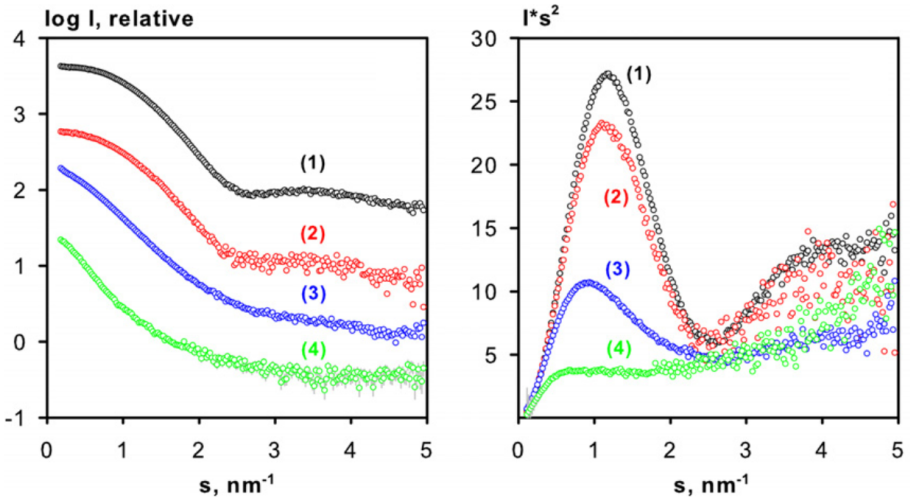


Figure 1.9: Kratky plots and related scattering plots. The Kratky plot ($I(s)$ vs. $s^2$) on the right panel is compared to the corresponding scattering plot ($I(s)$ vs. $s$) on the left. In this experiment the folded lysozyme (**1**) is displayed together with lysozyme in 8 M urea (**2**), lysozyme at 90° C (**3**) and with lysozyme in 8 M urea and at 90° C (**4**). Adding urea and increasing temperature make the protein unfold and this is reflected in the different behavior of the Kratky plot. The data have been measured at the beam line X33 in the DORIS synchrotron (Hamburg, Germany). Figure extracted from [21].

## 1.3 Modeling

As explained in the previous section, the $p(r)$ distribution represents the inverse Fourier transform of $I(s)$ (see equations 1.7 and 1.8). In the real world such a computation is impossible because of two main reasons: the scattering intensity is not measured in an infinite range but in a discrete number of points (from $s_{min}$ to $s_{max}$) and the real data are not perfect containing statistical and sometimes also systematic errors.

In order to overcome this problem, in 1977 Glatter proposed an approximated distance distribution function $p_a(r)$:

$$p_a(r) = \sum_{i=1}^{n_s} c_i \phi_i(r) \quad \text{for } 0 \leq r \leq D_{max} \tag{1.17}$$

where $n_s$ is the number of coefficient-weighted orthogonal functions $\phi_i$ only valid in the range from 0 to $D_{max}$ where the $D_{max}$ can be defined *a priori* by the user. $P_a(r)$ is therefore represented as a linear combination of functions (B-splines are used in the original publication) whereby the coeffi-

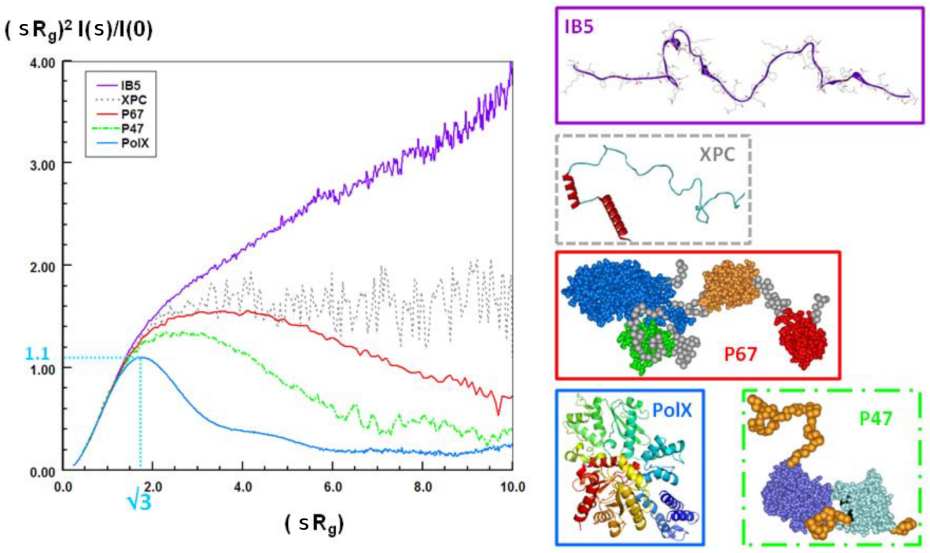


Figure 1.10: Dimensionless Kratky plot. The dimensionless Kratky plot is independent on the particle size and exhibits a distinct behavior depending on the particle flexibility. Globular particles like the folded protein PolX (blue line) have a peak in the coordinates $[\sqrt{3},\ 1.104]$, globular proteins with flexible linkers like P47 (green line) have a higher and less defined peak, flexible extended proteins with globular domains like P67 (red line) decays very slowly, disordered protein with few elements of secondary structures like XPC (gray line) reach a plateau between 1.5 and 2 on the y-axis and finally, completely disordered proteins like IB5 (purple line) grows constantly up to 4 or more on the y-axis. Picture adapted from [24] and available at the address `http://openi.nlm.nih.gov/detailedresult.php?img=3394175_CPPS-13-55_F2&req=4`.

cients $c_i$ are obtained through the minimization of the following functional:

$$\Phi = \chi^2 + \alpha P(p) \tag{1.18}$$

where the first term $\chi^2$ is the discrepancy or goodness of fit defined as:

$$\chi^2 = \frac{1}{N-1} \sum_{j=1}^{N} \left[ \frac{I_{exp}(s_j) - cI_{calc}(s_j)}{\sigma(s_j)} \right]^2 \tag{1.19}$$

The calculated scattering function $I_{calc}(s)$ can also be obtained from 1.17 as a linear combination of partial scattering functions $\Phi_i$ with the same

coefficients:

$$I_{calc}(s) = 4\pi \sum_{i=1}^{n_s} \int_0^{D_{max}} \varphi_i(r) \frac{\sin(sr)}{sr} \mathrm{d}r = 4\pi \sum_{i=1}^{n_s} c_i \Phi_i(s) \tag{1.20}$$

The discrepancy $\chi^2$ is a measure of the difference between the calculated scattering curve $I_{calc}$ and the experimental one $I_{exp}$ in every point of the measured range $s_j$ within the experimental error $\sigma$. This criterion is extensively used to qualitatively assess fits to experimental data in different physical experiments.

The second term $\alpha P(p)$ is a penalty term:

$$\alpha P(p) = \int_0^{D_{max}} \left[ \frac{\mathrm{d}p}{\mathrm{d}r} \right]^2 \mathrm{d}r \tag{1.21}$$

which ensures the smoothness of the $p_a(r)$ function. The regularization of the parameter $\alpha > 0$ allows one to avoid unstable and oscillating $p_a(r)$ distributions when the goodness of fit $\chi^2$ is minimized. When the value of $\alpha$ is too small the goodness of fit is prioritized but the $p(r)$ vs. $r$ function will present too many oscillations. When $\alpha$ is too high the function $p(r)$ vs. $r$ will be smooth but it will compromise the goodness of fit [25]. An optimum value of $\alpha$ exists ensuring good fits by a smooth distance distribution.

Once the $D_{max}$ is estimated, it is possible to produce the reconstruct $p(r)$ using a software that performs the indirect inverse Fourier transformation fit to the data. Currently the most widespread software used for this purpose is GNOM [26]. A more modern version of the program AutoGNOM [27] calculates the $p(r)$ distribution without giving the $D_{max}$ as input. The program will automatically try a range of $D_{max}$ values until the best fitting is achieved.

Minimization of the discrepancy is performed in varius modeling programs based on SAS data. The software can be divided in three main types: the *ab initio* reconstructions, rigid body modeling (or hybrid modeling) and the modeling of polydisperse systems.

### 1.3.1 *Ab initio* modeling

*Ab initio* three-dimensional modeling of macromolecules based on a solution scattering profile is an ill-posed problem. Here, the term *ab initio* means that modelling utilizes only the 1D scattering data without any *a priori* knowledge about the particle shape. The first attempts to solve this problem were based on simple geometrical shapes like spheres, ellipsoids or prisms, but such shapes are not able to describe accurately the complexity of the real macromolecules. In 1970, Stuhrmann [28] proposed that the scattering

curve can be conveniently described using a sum of spherical harmonics and based on this approach Svergun and co workers developed efficient methods applied to reconstruct 3D shapes from 1D data.

Currently the most widespread method for *ab initio* reconstruction is based on "dummy atoms" or beads. In the program DAMMIN [29], based on this approach, a sphere with a diameter corresponding to the maximum dimension of a particle (derived from $p(r)$, i.e. $D_{sphere} = D_{max}$) is filled with $N$ densely packed beads whose size is much smaller than that of the initial sphere. The beads are initially randomly assigned to a "phase" that can be particle (phase = 1) or solvent (phase = 0), and a Monte Carlo-like algorithm is applied to randomly change the assigned phases. Simulated annealing algorithm is applied to minimize the discrepancy 1.19 between the calculated (based on the model) and the experimental data (see Figure 1.11). In addition, penalties for disconnectivity and non-compactness of the current bead model are computed and added to the discrepancy similar to 1.18. This is an extremely important feature of the reconstruction allowing one to create physically and biologically sound models.

An optimized re-implementation of DAMMIN named DAMMIF [30] has been developed later and runs 20–45 times faster than DAMMIN. DAMMIF does not start with a defined restricted search volume but with a variable volume that can be automatically extended during the run in order to adapt to the model.

Obviously, reconstructions of a 3D model based on a 1D curve do not produce unique solutions and different models that fit the data equally well can be obtained. In order to overcome this problem it is recommended to run DAMMIN/F several times (at least 10) and then overlap the models to compare them. If the models obtained are similar one can create an average (most probable) model, otherwise the different shapes can be divided in clusters and then averaged. Finally, the averaged models can be filtered according to the occupancy of the beads. Such a process is done automatically using the programs SUPCOMB [31] to overlap and compare, DAMAVER [32] to make an average model, DAMCLUST to cluster the results and DAMFILT to filter the average models.

An extension of the *ab initio* bead modeling technique can be applied to study complexes like protein or nuclear acid-protein complexes associating each component of the complex to a different phase e.g. protein phase, DNA phase, solvent phase. This multi-phase modeling approach is particularly appropriate for SANS experiments where complexes consisting of different regions of different average scattering length density are often the object of study. In MONSA [29, 33] where this approach is implemented, multiple scattering curves can be fitted (e.g. the scattering curves of the single components and the one of the complex) simultaneously. For complexes, the approach is valid if the single components do not change conformation and oligomeric state upon binding (for contrast valriation in SANS, these

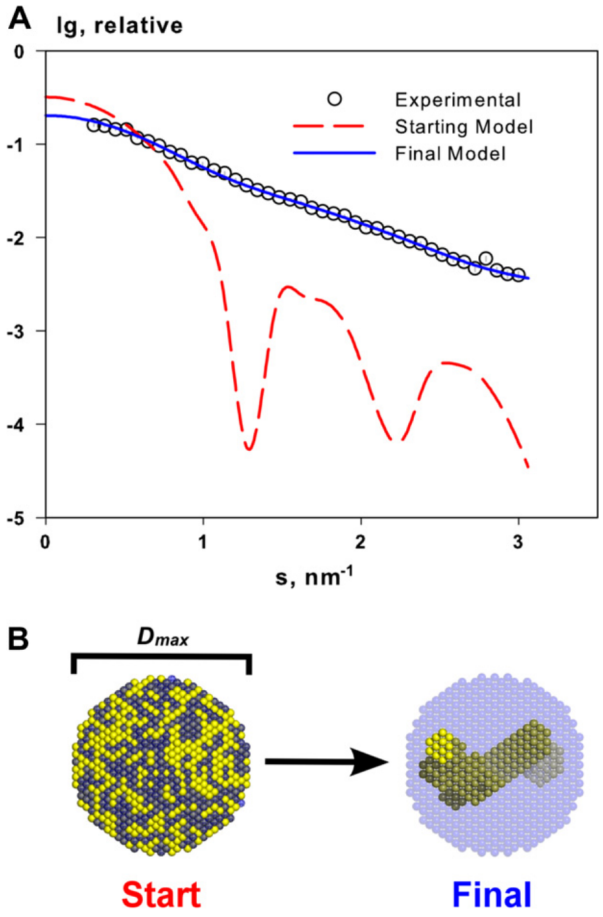


Figure 1.11: DAMMIN implementation. A DAMMIN run starts from a sphere (as large as the $D_{max}$) filled with densely packed dummy beads. **B** Initially the beads are randomly assigned to solvent (blue beads) or particle (yellow beads). **A** The corresponding scattering curve calculated for the initial dummy-atom model is shown in red. During the algorithm run the phase of the beads is randomly reassigned and the corresponding scattering curve is calculated. The procedure is repeated until the discrepancy between the experimental curve (circles in the plot) and the computed curve is minimized. The curve computed from the final model is shown blue in the plot. Figure extracted from [21].

conditions are fulfilled automatically).

One of the limitations of the phase-based reconstruction is that the macromolecules are assumed to have a constant average scattering length density inside the envelope shape, but this is not the case in the real world. A way to overcome this problem is to model proteins using beads that do

not fill the space but are instead placed in a chain (as in a pearl necklace) to simulate a polypeptide chain. Such a modeling approach is implemented in the program GASBOR [20] where the beads are as large as an average amino acid, their distance in the space approximates the distance between C-alpha atoms and the number of beads corresponds to the number of amino acids. Such information is given as input from the user and it is usually available from the protein sequence. GASBOR can be used to model proteins also at higher angular scattering range (up to about 1 $Å^{-1}$). Taking into account higher resolution data is not feasible with DAMMIN/F, which can represent the shape but not the internal structure of proteins. On the other hand, GASBOR is not recommended for large proteins ($> 700\ kDa$) due to the longer computational times compared to DAMMIN/F.

Whenever the symmetry of the particle is known in advance, it can be given as a hard constraint during the *ab initio* modeling process to build a symmetric model. Still, it is always useful to also run the programs in P1 symmetry models and compare the unbiased models with the symmetric ones.

### 1.3.2 Rigid body modeling

Whenever a high resolution structure of a macromolecule is available, it is possible to compute the scattering curve derived from its atomic coordinates [34]. In such a calculation it is necessary to take into account the scattering of the particle *in vacuo*, the scattering of the excluded volume (the volume of the solution occupied by the molecule) and the scattering from the hydration layer defined as the layer of solvent that surrounds the particle. It has been shown that the hydration layer has a higher scattering density compared to the rest of the bulk solvent. This higher density, of approximately 10%, is due to a higher number of solvent molecules per unit volume at the particle-surface boundary surface compared to the free solvent [35] (see Figure 1.12).

CRYSOL [34] and CRYSON [35] are two programs developed for X-rays and neutron scattering, respectively, which use the spherical harmonics to calculate the solution scattering from the atomic structures. Such programs are commonly applied to validate crystallography structures i.e. to check if the macromolecular conformation in the densely packed crystallographic lattice is maintained also in solution.

This approach can be used as starting point for hybrid modeling using SAS in conjunction with high resolution techniques. The hybrid modeling approach can be applied to determine the quaternary structures of multi-subunits complexes, in this case by rigid body modeling. SASREF [36] has been developed for this purpose and it can be applied whenever the high resolution structures of the single subunits are available.

A typical SASREF run begins from the subunits randomly positioned in space, and they are then randomly rotated and translated. Each time

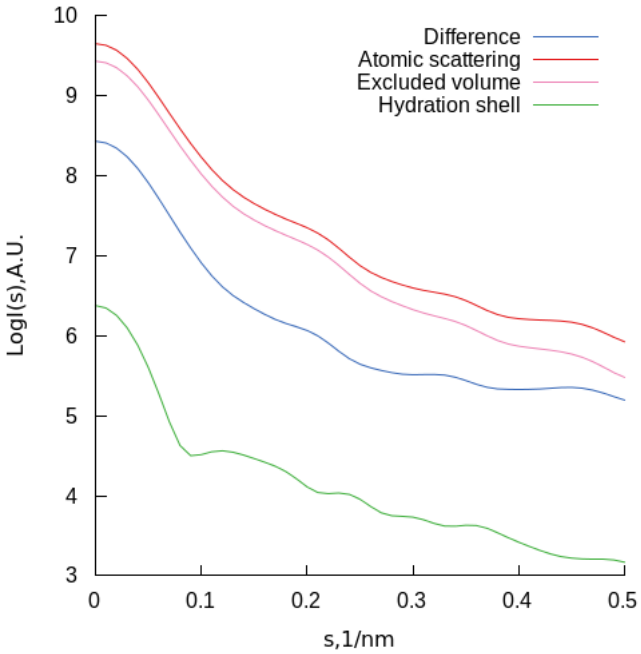


Figure 1.12: CRYSOL-computed scattering curves. In order to calculate the scattering curve from a particle in solution CRYSOL takes in consideration: the scattering of the atoms in vacuum (red curve), the scattering of the excluded volume (pink curve) and the hydration shell (green curve). The difference curve (blue curve) takes into account all the contributions. The curves are calculated from the coordinates of the bovine serum albumin crystal structure (PDB entry 3E0V).

the subunits are positioned in different arrangements, the scattering curve is recomputed and the discrepancy from the experimental data is evaluated. The aim of the simulated annealing algorithm is to minimize the discrepancy of the model scattering against the scattering data taking into account penalty values associated with model interconnectivity and steric clashes. Multiple experimental scattering curves can be fitted at the same time e.g. the scattering curve obtained from the complex and from the single subunits alone. Importantly, SASREF allows for incorporation of data obtained with other biophysical techniques like: contacting amino acids evaluated with mutagenesis or distance between amino acids evaluated with fluorescence resonance energy transfer (FRET). Finally, whenever available, information about symmetry constraints can be applied to the models.

Although being very useful for quaternary structure modeling, SASREF provides only approximate results for high resolution structures with missing amino acids, like structures obtained with X-ray crystallography without N-, C-terminus or without flexible linkers. In such cases two other programs

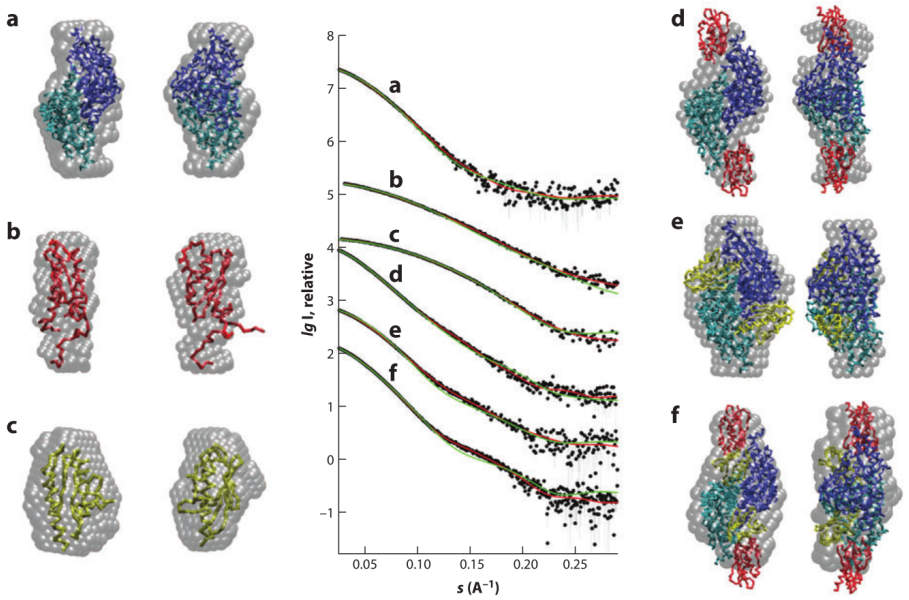


Figure 1.13: Rigid body modeling in comparison with *Ab initio* modeling. The scattering curves and the *ab inito* and rigid body models from IscS (**a**), IscU (**b**), CyaY (**c**), IscU/IscS (**d**), CyaY/IcsS (**e**) and CyaY/IscS/IscU (**f**). The *ab inito* models are displayed as grey spheres and green corresponding scattering curves. The rigid body models theoretical scattering curves are red. The experimental scattering curves are represented as black dots. The data corresponding to this list of proteins are stored in SASBDB with the following identification codes: IscS (SASDAV6), IscU (SASDAW6), CyaY (SASDAX6), IscU/IscS (SASDAY6), CyaY/IcsS (SASDAZ6), and CyaY/IscS/IscU (SASDA27). Figure extracted from [1].

can be used BUNCH [36] and CORAL [7]. The two programs have similar features and can be applied using rigid body modeling for known structures to build multi-meric models while the missing amino acids are represented as dummy beads. Each bead in the missing part corresponds to a residue, similar to GASBOR and these beads are interconnected to maintain the $c\alpha - c\alpha$ distance between the subsequent dummy residues. A simulated annealing algorithm is used for the minimization, multiple curves can be fitted at the same time, and symmetry as well as information derived from other biochemical techniques can be used as constrains in the modeling process. The main difference between BUNCH and CORAL is that the former is limited to single chain modeling while the latter allows for multiple chains. Both methods can be applied for SAXS (not for SANS) data.

### 1.3.3 Polydisperse samples modeling

All the modeling approaches described until now can only be applied to monodisperse systems where all the particles have the same conformation and oligomerization state. Still, SAS can also be applied to polydisperse systems. The scattering from a solution containing $K$ distinct components (types of particles) with different shapes and structures can be written as

$$I(s) = \sum_{k=1}^{K} \nu_k I_k(s) \tag{1.22}$$

Where $\nu_k$ is the volume fraction of each component and $I_k(s)$ is the corresponding scattering intensity. When the number of components and the relative scattering intensity is known in advance, it is possible to calculate the volume fraction of each component. This approach can be applied to characterize chemical reactions or equilibria among different components. Such a procedure it is currently used by the program OLIGOMER [37]. OLIGOMER uses the information from experimental or calculated (obtained with CRYSOL) scattering curves to determine the volume fractions of each component. When the number of components is unknown, it is still possible to apply OLIGOMER together with the single value decomposition (SVD) approach [38]. The latter allows for the determination of the number of components.

SAS is one of the few techniques that can be applied to model multi-components and heteogeneous systems. For this reason, some of the previously described programs (GASBOR and SASREF) have been extended to accommodate mixtures [39]. GASBORMX can be used to obtain *ab initio* shape models of multimeric states of proteins in equilibrium with their monomers while SASREFMX can build rigid body models from the scattering by mixtures of different oligomerization states.

A distinct type of polydisperse system are flexible proteins and their subgroups of intrinsically disordered protein (IDP) or proteins with intrinsically disordered regions (IDR). Such proteins are considered to be polydisperse because they do not assume a stable tertiary conformation in solution therefore are defined by a mixture of -an astronomic number of- different conformations.

Such a set of different conformations may be defined as an ensemble and it is possible to select a refined ensemble composition that describes the SAS data. The ensemble optimization method (EOM) [40] begins from a pool comprising a large conformational space (by default 50.000 structures) and, using the genetic algorithm, selects an ensemble of conformations that fits the experimental scattering curve with the minimum discrepancy 1.19. The EOM algorithm will be explained in detail in chapter 3.

## 1.4 Conclusions

In this chapter the basics of SAS have been considered going from the experimental setup to the extraction of key structural parameters and different modeling approaches. The presented information provides the basis for the development of the repositories illustrated in the following chapters of this dissertation. The next chapter will focus on the technical basis of database design and development.

# Chapter 2

# Database design

A database is a large collection of data grouped and organized such that it can be managed by a software system named database managment system (DBMS). Dealing with large amounts of data has always represented a challenge for computer science. Till 1970s, when the first database were devised, data storage was primarily implemented using long concatenated files. While these files are easy to store, they are limited in size and scalability and it is difficult to edit and exchange information concurrently between multiple users. Databases, together with the related DBMS have been developed to overcome this problem, in particular to:

1. Share and exchange data among different users and applications at the same time.
2. Be consistent across time.
3. Offer a reliable platform for data storage that is not dependent on hardware or software performance.
4. Maintain the accuracy of records and privacy of the data (as each user has a specific set of permissions).
5. Optimize efficiency and be effective using limited resources (in terms of time and space) while maintaining the requirements of a user.

The data in a database are organized and described according to a specific database model. In 1983 Edgar F. Codd [41] published an article describing the relational model that remains the most widespread set of principles for data organization. In the relational model data are organized in tables connected by specific relations, each column of the table is defined as an *attribute* and each row as *tuple*, i.e. an ordered list of elements. Importantly, each tuple is identified by a unique code named *key*.

An example of a relational database table is shown in Table 2.1. This table describes the elements of SAS software following relational database

Table 2.1: An example relational database table

| Key | Software | Type | Author |
|---|---|---|---|
| 1 | DAMMIF | *Ab initio* | D. Franke |
| 2 | SASREF | Rigid body modeling | M. Pethoukov |
| 3 | OLIGOMER | Polydisperse system | P. Konarev |

rules. The attributes of the table are "software", "type" and "author"; each tuple describes a different software and it is identified by a unique key.

From the creation of the relational model, different DBMS have been developed to deal specifically with that model. An example of DBMS specific for the relational model is the open source MySQL `https://www.mysql.com/` which uses the structured query language (SQL) database language [42]. The SQL database language allows the definition, modification and query of a database. Importantly, creation of the physical database using, for example, MySQL as DBMS is only the last phase of the database design process. Most of the design process is independent of the underlying DBMS and can be divided in four main phases described in the following sections:

Section 2.1 : Collection and analysis of database requirements.

Section 2.2 : Organization of the information in a conceptual schema.

Section 2.3 : Application of normalization rules to create a logical schema.

Section 2.4 : Development of the physical model using the designed DBMS.

## 2.1 Requirements

The first phase of database design consists of defining the scope of the database in terms of requirements, i.e., defining what types of information are to be stored and what do the end users of the database need with respect to database management. During the requirements phase an outline of the functionalities of the system is mapped out in order to produce an accurate description of both the data to be stored and of the database operations. The operations will be different according to the type of the database user. As an example, for the development of a SAS database one has to take in consideration different types of users (or use cases): one user could be a biologist interested in the scattering data collected about the protein he/she is working on, another could be a scientist who performed a SAXS experiment and wants to deposit the derived data onto the database, a third user could be a journal referee who wants access to the primary data and models related to the publication submitted to a journal. This list of three possible

users will have different requirements and will perform different operations on the database. Once all the requirements information is collected, the database designer has to organize them in groups and identify the relations among those groups. The information in each group should be conceptually related (e.g. data of the same type should be grouped) because eventually they will become the tables of the relational database.

## 2.2 Conceptual schema

Once the database requirements, tables and relations have been defined, the second phase consists of organizing these elements into a conceptual schema. The most common conceptual schema is the entity-relation diagram (ERD) that illustrates the entire database through entity and relation objects. An entity is a class of objects with similar properties while a relation object is the logical link among two or more entities. Both entities and relations are identified by a unique name and can have a set of attributes related to it. One or more of the attributes of an entity must be unique and define the entity key.

Whenever a relation among entities is established, one has to define the minimum and maximum number of entities involved in each relation. These numbers are defined as *cardinality* and the most common are: one-to-one, one-to-many or many-to-many. A cardinality can also be set to zero, meaning that the relation is optional. In the example in figure 2.1 the entity "experiment" is connected to the entity "publication" trough the relation named "result". The small digits close to the connecting lines indicate that one experiment can optionally result in many publications (0,N) while each publication is the result of at least one experiment (1,N).

A correct ERD has to fulfill a set of requirements: (i) An ERD has to be complete i.e. it has to include all the requirements obtained during the first phase of database design; (ii) The ERD has to be understandable for non-experts and (iii) The ERD should avoid data redundancy i.e. each information has to be stored in one location only.

## 2.3 Normalization

After the conceptual schema has been completed, the next phase of the database design consists of applying a set of optimization rules (i.e., normalization) to transform the conceptual ERD schema in an equivalent logical schema. Once assured the correctness and completeness of the conceptual schema, the next phase consists in applying a set of optimization rules to transform the conceptual schema into an equivalent logical schema. Multiple rules govern the different steps of normalization, the details of which are

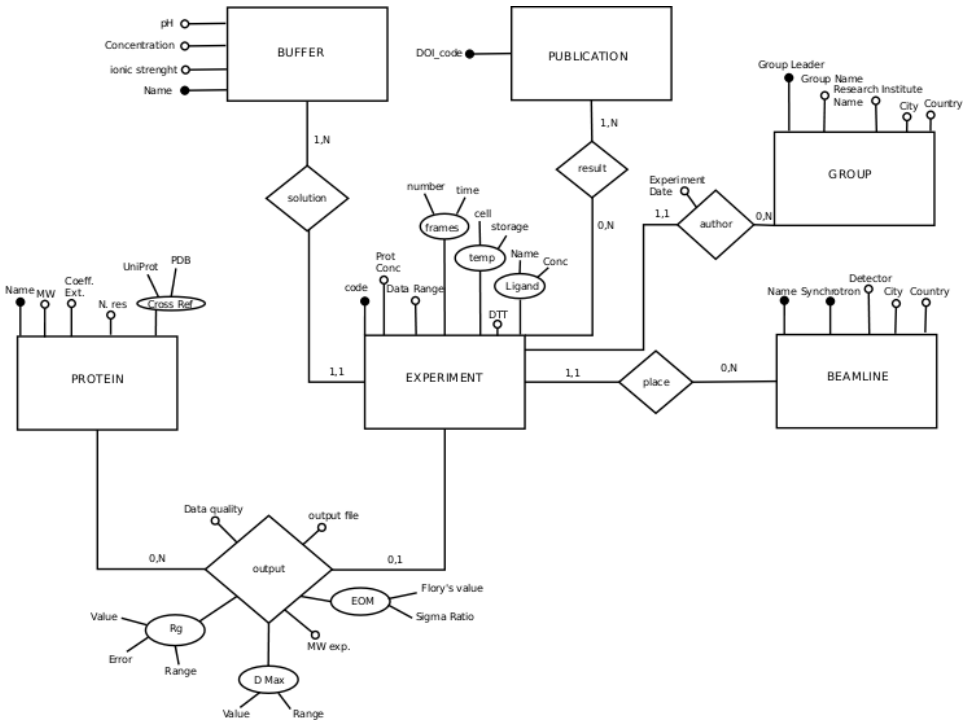


Figure 2.1: Entity-relation diagram illustrating a SAXS database. In an ERD the entities are represented by rectangular shapes while relations by diamond shapes. The attributes of both entities and relations are small dots linked by a line and the black dots represent the entity keys. This ERD is a conceptual schema for a database storing SAXS data about IDPs (the first version of the PED database that will be described in the next chapter 3).

outside of the scope of this dissertation. Briefly, the optimization rules to be applied may be grouped into three main categories:

1. Elimination of duplicates and redundancy.

2. Check for atomicity i.e. the data related to each entity have to be logically related. This means that the entities of the ERD can be merged or divided according to the needs.

3. Analysis of the relationships: in the logical schema some of the relations of the ERD are merged into the related entities according to the cardinality.

The application of the optimization rules results in a logical schema like the one in Figure 2.2 where the previously illustrated conceptual schema of Figure 2.1 has been revised according to the optimization rules. The

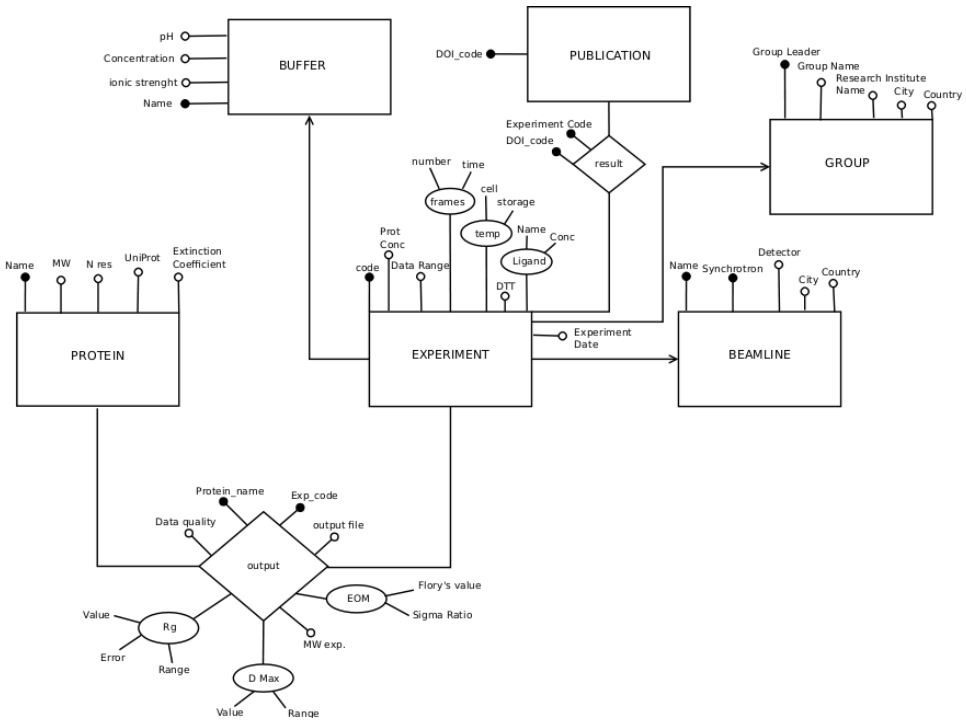


Figure 2.2: Logical schema illustrating a SAXS database. In the logical schema the entities are represented as rectangles and the relations as diamond shapes. Entities are linked by pointers according to the type of relation between them. This logical schema was created for the database PED that will be detailed in the next chapter 3.

resulting schema contains a diminished number of relations and the linkers between entities transformed in pointers. Importantly, during the normalization phase one must also define the restraints of each attribute in terms of data type (e.g. character or integer) and also space occupied. These restraints will assure stability and correct usage of the database. A widespread way to illustrate the logical schema displaying also the restraints is the unified modeling language (UML) `http://www.uml.org/`, which is a graphic language commonly applied not only to databases but also to software applications. Figure 2.3 illustrates the UML-based logical schema of another database, namely SASBDB that will be described in Chapter 4.

The logical schema is the last step before creating the relational database. Indeed, all the objects of the logical schema will ultimately be transformed into tables connected by pointers. The entity identifier will become the primary key and the attributes will be defined according to the given restraints.

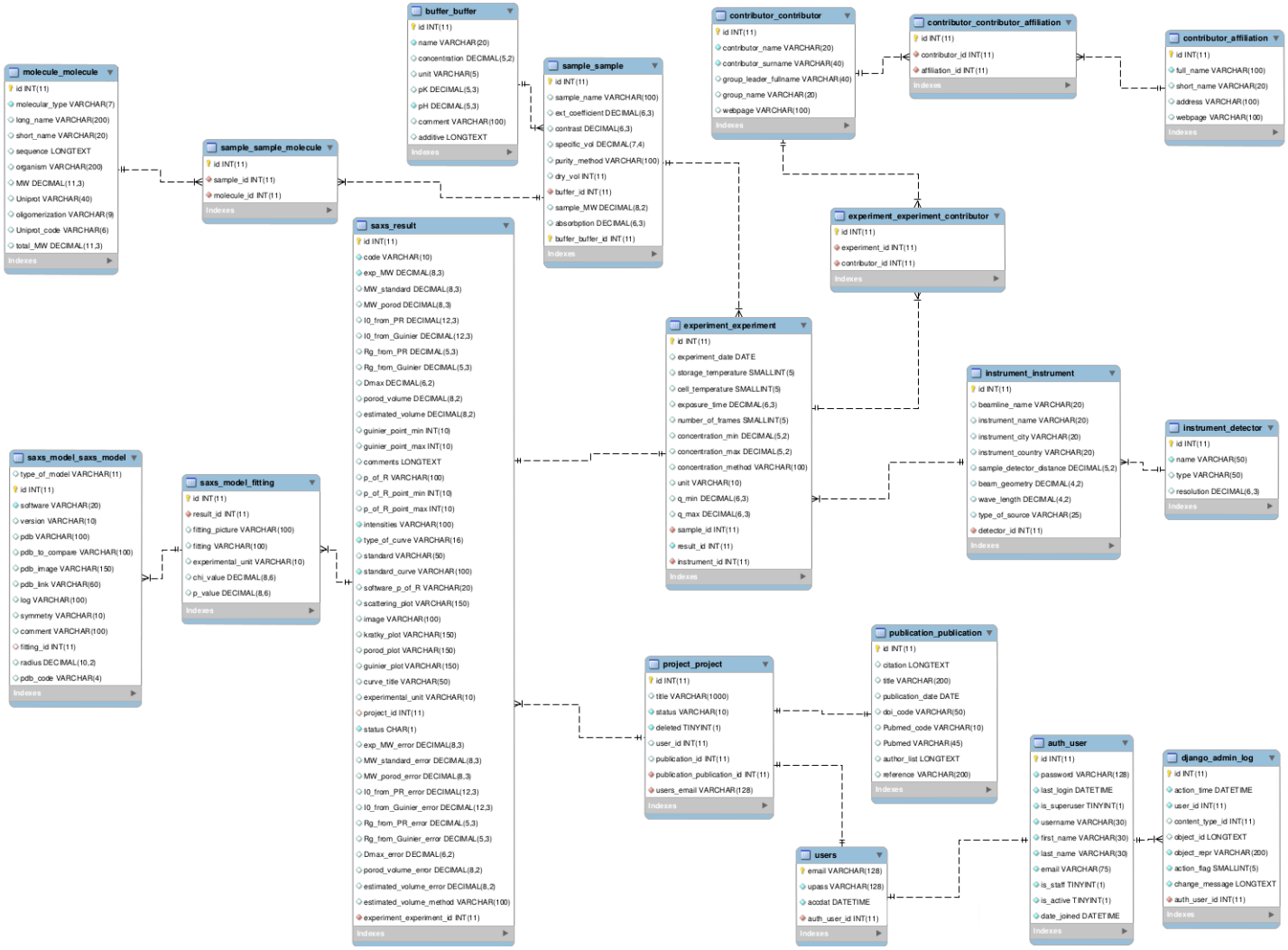


Figure 2.3: UML representation of SASBDB logical database schema. In this type of representation each object is a diagram containing the list of attributes with related restraints, primary keys are represented as small key symbols and the relations among objects are identified with dotted lines with different termini according to the type of relation. For more information refer to Chapter 4.

## 2.4 SQL

The last phase of developing a database takes advantage of the SQL language that actually creates the database. The SQL contains a set of standardized commands which allow for database creation and database query. Indeed with SQL it is possible to insert, update, delete and select data. An example of a simple database query to a database in SQL is illustrated in figure 2.4.

```
mysql> select * from instrument_detector;
+----+-----------------------------------------+--------------------+------------+
| id | name                                    | type               | resolution |
+----+-----------------------------------------+--------------------+------------+
|  1 | Pilatus 1M-W pixel                      | 2D Photon counting |      0.172 |
|  7 | Roper Scientific KAF 2084 x 2084 SCX CCD | CCD                |     24.000 |
|  3 | 1D Gas detector                         |                    |       NULL |
|  4 | Pilatus 2M                              |                    |       NULL |
|  5 | Pilatus-1M                              |                    |       NULL |
|  6 | MAR Image Plate                         |                    |       NULL |
|  8 | VANTEC                                  |                    |       NULL |
|  9 | AVIEX                                   | CCD                |       NULL |
+----+-----------------------------------------+--------------------+------------+
8 rows in set (0.00 sec)
```

Figure 2.4: Example of an SQL query. The command used in this case is aimed at showing all the tuples in the table named instrument_detector. In this case the result of the request is the instrument_detector table as stored in the SAS database SASBDB that will be described later in this dissertation (Chapter 4).

## 2.5 Conclusion

The application of the relational database design procedures allows for the development of systems that are reliable, persistent, consistent and efficient. In addition, the spread of the relational model lead to the production of multiple tools for software developers and to the growth of a large open source community. For these reasons the relational model was chosen to develop the databases detailed in the next chapters of this dissertation namely PED (Chapter 3) and SASBDB (Chapter 4).

# Chapter 3

# PED: a database of structural ensembles of intrinsically disordered and of unfolded proteins.

*M. Varadi, S. Kosol, P. Lebrun, E. Valentini, M. Blackledge, A. K. Dunker, I. C. Felli, J. D. Forman-Kay, R. W. Kriwacki, R. Pierattelli, et al. Nucleic acids research, 42:D326–35, 2014.*

Both IDPs and IDRs are becoming more and more common as targets for structural biologists. The popularity of the topic -proven by an increase in the number of scientific publication (as illustrated in figure 3.1)-, together with the need to better understand the structural behavior of such proteins, brought to the development of a database for the deposition of protein ensembles named the protein ensemble database (PED). SAXS is one of the few techniques that are able to characterize IDPs and IDRs structurally. In this chapter PED is described with a particular emphasis on theSAXS-based aspects of the database.

In order to fully understand the content of the repository, some details about flexible proteins and the EOM technique will be reviwed prior to the PED database.

## 3.1 Intrinsically disordered proteins and proteins with intrinsically disordered regions

IDPs are proteins that lack of a defined tertiary structure under physiological conditions. As a consequence of their high flexibility, a single IDP can be involved in multiple interactions usually involving transient complexes with both proteins and nucleic acids. In some cases IDPs assume stable

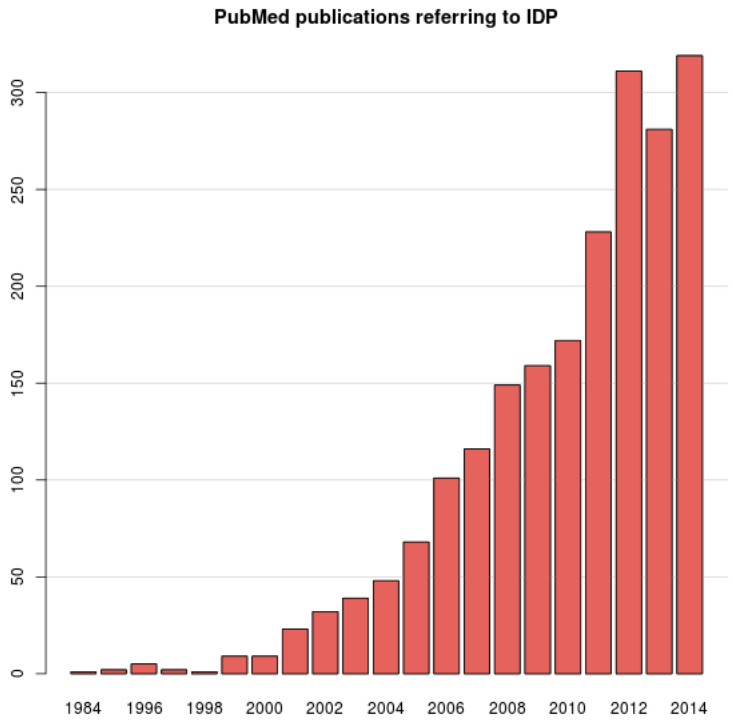


Figure 3.1: IDP number of publications in PubMed by year. The search has been performed based on the article title or abstract. Using the following search criteria: "(inherently OR natively OR intrinsically) AND (disordered OR unfolded OR unstructured) AND protein". Figure adapted from [43].

tertiary structure upon complex formation, undergoing a disorder-to-order transition. In other cases they remain in a disordered state forming so-called "fuzzy complexes" [44].

The traditional structure-function paradigm states that the function of a protein is determined by its structure. In contrast to this paradigm IDPs perform key functions for the cell despite their lack of a defined tertiary structure [45]. Such functions include, but are not limited to, differentiation, signaling, transcription regulation, DNA condensation and cell division [43]. In addition, bioinformatic analysis suggests that IDPs and IDRs are particularly abundant in eukaryotes (from 33% to 50% depending on the predictor) and in disease-related proteins. Indeed IDPs have shown to play crucial roles in diseases like cancer, cardiovascular disease, amyloidoses, neurodegenerative diseases and diabetes [46].

Many prediction servers to identify IDPs and IDRs have been developed over the last few years. These servers use predictions based on amino acid

composition (like GlobPlot [47] and also on estimated interaction energy (like IUPred [48]) or machine learning methods (like DISOPRED2 [49]).

In addition to the bioinformatic approach, a number of experimental techniques can be applied to identify IDPs and IDRs. Those structural and biochemical techniques include protease digestion, NMR and SAXS. Protease digestion aids with the identification of IDPs because IDRs show a much higher propensity toward proteolytic hydrolysis compared to globular domains.

NMR and SAXS are efficient techniques to structurally investigate the IDPs mainly because these methods are applied in solution and do not require crystallization. NMR allows for the acquisition of spectra that can yield atomic-scale distance information and dynamics but is limited in protein size to maximum 100 kDa, while SAXS gives indications about the overall shape and does not have limitations in size. For these reasons a combination of NMR and SAXS is considered very effective in structural biology and for the the studies of IDPs in particular [50, 51].

## 3.2 IDPs and SAS

SAS is one of the few techniques that can be give structural insight into polydisperse samples. As explained in the previous chapter, equation 1.22 can be applied to determine the volume fraction of different components in a polydisperse mixture. In case of IDPs, the number of components corresponds to the different conformations that the protein can assume in solution. This is an astronomical number and the determination of the volume fractions is by far not trivial. Being the volume-fraction weighted sum of so many conformations, the scattering curves of IDPs usually display featureless slowly decaying patterns, while the Kratky plots do not show any defined maximum and demonstrate a continuous growth. This is evident from Figure 3.2 where a Kratky plot and the scattering curve of an average over 10.000 structures are presented [6].

Since equation 1.22 cannot be directly applied, new methods have been developed to investigate IDPs. The most widespread method is based on selecting an ensemble from a pool covering the whole space of possible conformations. A concept of ensemble modelling was initially developed for NMR and this principle was successively adapted to SAXS with the program EOM [53]. Figure 3.3 illustrates the general implementation of the EOM algorithm.

### 3.2.1 EOM algorithm.

The EOM algorithm begins with creation of a pool of 10.000 different protein conformations of $n$ amino acids represented by beads, where $n$ is provided by the user. The scattering profile of each conformer is calculated (using

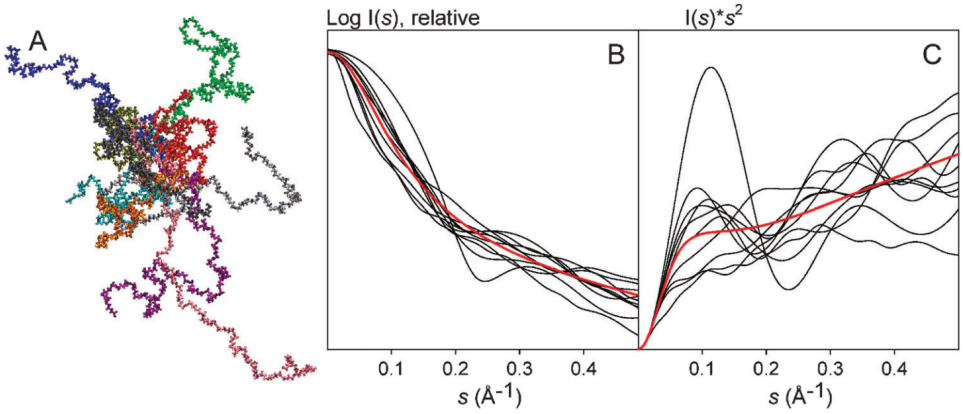


Figure 3.2: IDP Kratky plot and a scattering curve from a model IDP. 10.000 poly-alanin chains have been created with Flexible Meccano [52], 10 structures have been selected (showed in panel **A**) and their relative scattering curves and Kratky plots are displayed in black in **B** and **C** respectively. The average curve from the 10.000 conformers are displayed in red, assuming the same volume fraction for each conformation. Picture extracted from [6]

the program CRYSOL) following steric restrictions and clash-avoidance (i.e. the chains do not self-cross) through the incorporation of Ramachandran $\phi$ and $\psi$ angle constraints. The genetic algorithm is then applied to select a subset of $N$ conformations composing the ensemble. In the genetic algorithm each ensemble is defined as a *chromosome* and each conformation is a *gene*. For each chromosome the average of the individual scattering patterns is calculated using equation 1.22 and the discrepancy from the experimental curve is evaluated using equation 1.19. At each generation of the genetic algorithm some of the genes are changed by selecting other random chromosomes or exchanged between two selected chromosomes (*mutation* and *crossing*). The chromosomes with the minimum discrepancy are selected for the next generation (*elitism*). This process allows for optimization of a single chromosome (ensemble) fit to the experimental data. Importantly, the number of conformations composing the ensemble is not fixed *a priori* and is determined during the optimization process itself. In addition, in the last version of the program [54] symmetry restrictions, folded subunits, nucleic acids and multiple pools can also be included as part of the modeling process.

Different approaches have been developed to quantitatively assess the flexibility of the protein based on the EOM results. The first approach is the size of the ensemble selected by EOM which can span fifty (in case of very flexible particles) down to two (in case of rigidity) protein conformations. The second qualifier is the comparison of the $R_g$ and $D_{max}$ distribution of

the selected ensemble with the distributions in the initial pool. The protein is considered moderately flexible when the distribution of the optimized ensemble is narrower than the initial pool of structures. On the other hand the protein is considered extremely flexible when the ensemble distribution is as wide (or even wider) then the pool.

EOM improved the previous applications of SAXS to flexible systems because it does not only return average structural parameters but also the distribution of these parameters, furnishing precious indications about the particles. Such distributions resulting from EOM together with the Kratky plots belong to the SAXS-related information stored in PED that will be described in the next section.

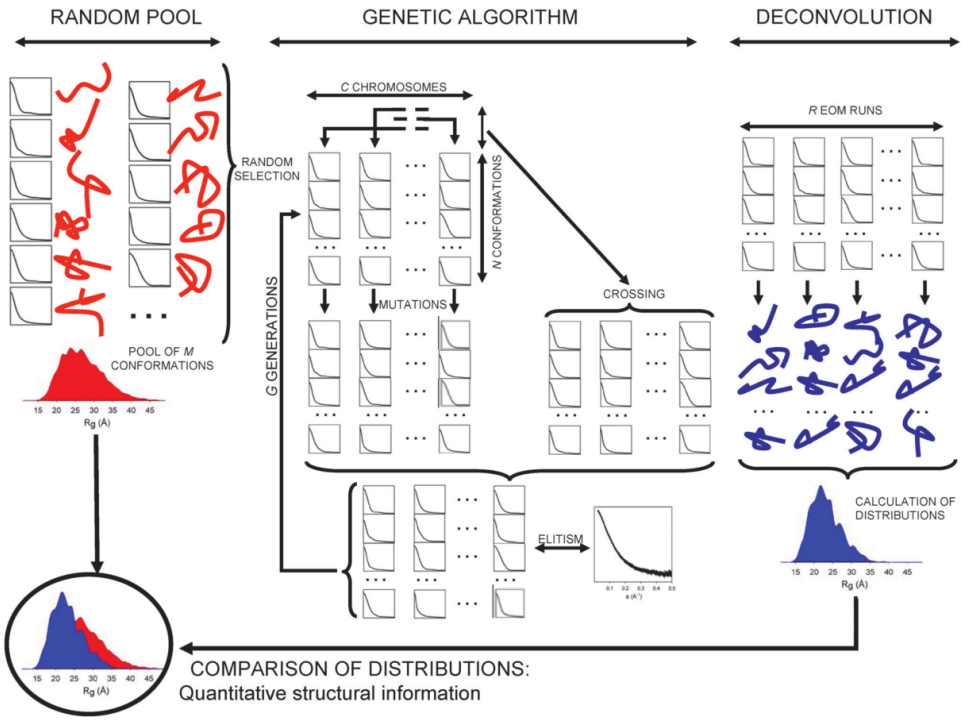


Figure 3.3: EOM algorithm description. A pool of $N$ protein conformations is initially generated, for each conformation the scattering curve is calculated and the $R_g$ and $D_{max}$ distributions of the initial pool are evaluated. At the end of the EOM optimization, a pool of structures is selected by the genetic algorithm that represents a volume fraction-weighted sub-set of structures from the initial pool. The average calculated scattering profile of the refined ensemble fits the experimental data. Picture extracted from [6].

## 3.3 The protein ensemble database: PED

As explained in the previous section, IDPs can assume a huge number of conformations. This behavior is due to the low free-energy barriers between the different conformations that allow for kinetic exchange between various structural states. PED [14] is the first database aimed at storing structural information obtained from ensembles of IDPs, IDRs and denatured proteins. The IDP sequences and other biophysical data are stored in DisProt [55] or in the data bank of intrinsically disorderd proteins with extensive annotations and literature database (IDEAL) [56]. PED aims at centralizing the data and structures obtained from various structural biological techniques (SAXS, NMR and MD) applied to IDPs and IDRs. In addition, special attention is given to storing the information about the software applied to determine the ensemble. In general the software tools are based on two main approaches: *1*) selection of a subset of conformations from a pool covering the whole conformational space; *2*) the use of MD to sample the conformational space of a specific protein [57]. The first approach was already described in the previous section: in EOM the experimental scattering patterns are used as constraints to select the ensembles. Similarly, NMR-derived values like chemical shifts (CS), residual dipolar coupling (RDCs) or J-coupling can be applied to select the ensemble. The second approach explores the free energy landscape specific of the given protein but, due to computational limitations, application of experimental techniques is recommended to obtain reliable results.

PED stores two main types of data: the primary experimental data (e.g. scattering patterns in case of SAXS data) and the 3D-atomic coordinates of the conformations composing the ensemble (typically as pdb files). Metadata describing the experimental details are also provided. The data are organized in a schema where each technique has a dedicated section: SAXS-related information includes beamline details, experimental parameters, sample conditions and structural parameters like $R_g$, $D_{max}$ and *I(0)*. Importantly, several cross-links to other biological databases are stored, namely UniProt [58], Ensembl [59], biological magnetic resonance data bank (BMRB) [60] and Disprot [55]. A simplified schema of the database is illustrated in figure 3.4.

Each entry can be related to multiple ensembles and each ensemble may contain up to thousands of conformations. In order to describe each conformation correctly, the web-interface (available at `pedb.vib.be`) has been developed with specific characteristics. These characteristics include, for SAXS, a normalized Kratky plot and images of three selected structural models, namely the one corresponding to the most compact, the average and the most extended $R_g$. To help the flexibility assessment based on the Kratky plot two reference curves are added to the plot: one based on the globular protein bovine serum albumin (BSA) and one based to the com-

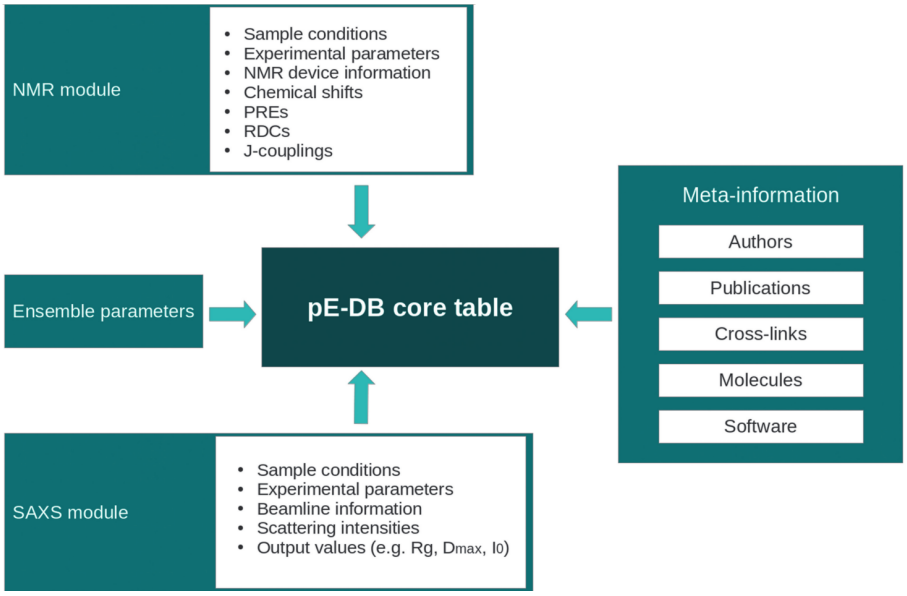


Figure 3.4: PED database schema. PED contains a core of tables describing the ensemble to which method-specific and meta-information tables are connected. The method-related tables contain general information like sample conditions and experimental parameters and method-specific data. The SAXS-related tables include the scattering intensity data and structural parameters. The NMR-related tables include NMR instrument information, chemical shifts, RDCs and J-couplings. Metadata include information about the authors of the experiment, related publications, cross-links to other biological parameters, details about the measured molecules and about the software used to determine the ensemble. Picture extracted from [14].

pletely disordered Tau protein [61]. On clicking on the structural model images, an interactive JSmol window is loaded. From this interface the selected model is visualized in 3D and is flanked by the $R_g$ distribution of the ensemble. Below the $R_g$ distribution the complete list of conformations is made available and, on clicking on each of those, the corresponding model is loaded in the JSmol interactive window. This interface is showed in figure 3.5.

The web interface includes multiple ways to browse the content of PED according to the technique, protein name, author and publication. The same criteria can be used to restrict the search that which by default, i.e. when no option is selected, is performed on every field. The primary data related to each entry are available for download and the whole database can also be downloaded as a logical schema.

In order to submit data, users have to fill in a pre-submission request form. The approval by the PED team will allow the users to submit data

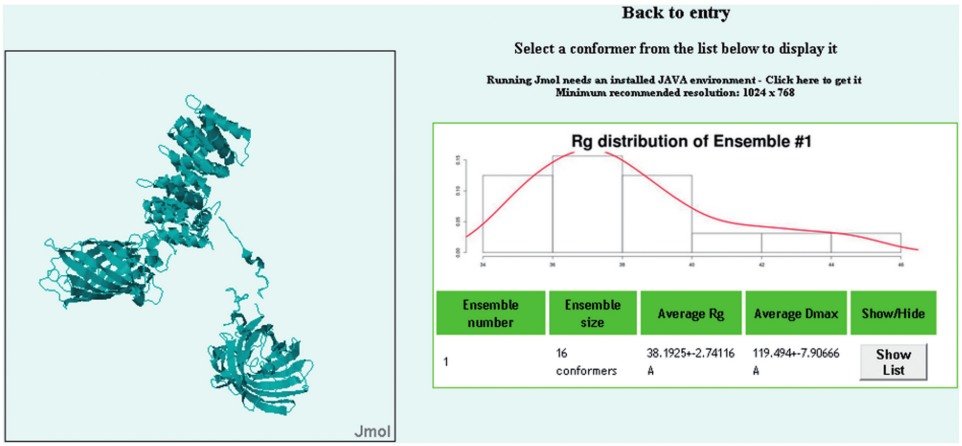


Figure 3.5: PED ensemble visualization. The JSmol window allows the interactive visualization of the structure. The $R_g$ distribution related to the experiment is displayed on the right, followed by a table representing the ensemble. In this case the ensemble contains 16 conformers, the list of conformers can be extended and, for each conformer, the $R_g$ and $D_{max}$ values are displayed. On clicking one conformer, the corresponding structure is loaded in the JSmol window. Figure extracted from [14].

through an online form and send experimental data and the ensemble structures by a standard file transfer protocol (FTP). PED currently contains 21 entries, each containing a different number of ensembles. Each ensemble is composed by a number of conformers which range from a minimum of three up a maximum of 4939 conformations. The users are encouraged to download the experimental data, re-calculate the ensembles using the same or different software and eventually submit any newly-calculated ensemble derivations.

## 3.4 Conclusions

PED is the first database aimed at the structural characterization of IDPs. The main purpose of PED is to act as a repository of structural data, parameters and models pertaining to protein ensembles. The data is accessible for the entire structural biology community thus further promoting the information exchange and facilitating the structural investigations on flexible proteins.

The author contributed to this project in designing the structure and the content of the SAXS-related part of the relational database schema. This collaboration formed the basis for the development of SASBDB, a comprehensive database of SAS data and models spanning various macromolecule types and archiving both primary data and derived models. SAS-

BDB database will be described in the next chapter.

# Chapter 4

# SASBDB, a repository for biological small-angle scattering data

*E. Valentini, A. G. Kikhney, G. Previtali, C. M. Jeffries, and D. I. Svergun. Nucleic acids research, 43 pp. D357–63, Jan. 2015.*

The small angle scattering data bank (SASBDB) is a repository of SAS primary data and models deposited together with the relevant information about the experiment including macromolecule characteristics, instrument details, cross links to other biological databases and information about the author and publication. The establishment of a SAS database has become a necessity with the increasing community of structural biologists applying SAXS and SANS in their research. In response to the requests of the SAS community, in 2012 two articles were published stressing the need to standardize and make publicly available the primary data underlying scientific publications. The first work, written by SAS experts, outlines the guidelines to be followed for presenting SAS results in publications [62]. In this article the authors define a list of quality checks and structural parameters that should be mandatory in any publication including SAS data. In addition, they clearly state that "... making scattering data publicly available is necessary, or at least desirable". The second publication is yet more specific about the importance of a SAS data repository for the structural biology community. Here, the world wide protein data bank SAS task force (wwPDB SAStf) outlines a list of recommendations for a SAS repository [9]. These recommendations encompass:

1. The need of a global repository that is searchable and freely accessible for download.

2. The development of a standard dictionary to import and export SAS data.

3. The possibility to store SAS-derived and atomistic models with details about the software used to obtain them.

4. Quality assessment of the data and accuracy of the models.

5. The possibility to archive diverse data and models.

SASBDB has been developed in agreement with the requirements outlined by these publications by SAS experts. The data stored in SASBDB describe important details of a SAS experiment with the purpose of improving reproducibility and transparency of data reporting. Indeed, the lack of reproducibility and transparency is an urgent matter in modern research as also indicated by the NIH directors [63].

## 4.1 Which data are stored in SASBDB

As explained in the first chapter of this dissertation, the SAS data are usually displayed as a plot where the scattering vector $s$ is a function of the scattered intensities $I(s)$ on a logarithmic scale. The indirect inverse Fourier transform of $I(s)$ vs. $s$ gives the distance distribution function $p(r)$. Other plots include the Guinier plot ($\ln I(s)$ vs. $s^2$) and the Kratky plot ($s^2 I(s)$ vs. $s$). From these plots it is possible to obtain key structural parameters like the radius of gyration ($R_g$), the maximum particle dimension ($D_{max}$), the particle volume estimates (Vol) and the molecular weight ($MW$) of the macromolecule. The whole set of structural parameters of SASBDB is detailed in table 4.1.

All these portions of information are stored in specific and standardized formats. The SAS curve is a three column ASCII format (.dat file) where the first column is the momentum transfer $s$ (in inverse nanometers $nm^{-1}$), the second column is the intensity $I(s)$ (arbitrary unit or absolute scale) and the third is the associated error. The distance distribution function $p(r)$, instead is stored in the GNOM format [26] (.out file). These are the formats currently used by the ATSAS package [7] which is the most widespread suite of SAS software and already counts a community of more than 10.000 SAS users. Other formats and units will be accepted in agreement with the further expansion of SASBDB.

Several approaches are available to reconstruct the shape of the macromolecules by fitting the experimental SAS data. The two main types of models are low-resolution *ab initio* models and the so-called hybrid models based on high-resolution structures. These two types of models can also be combined in the same particle reconstruction whenever the high-resolution structures have missing portions. In case of mixtures it is possible to calculate the volume fractions of the single components and finally, for flexible proteins, an ensemble of conformations can be built (further details on modeling approaches can be found in chapter 1).

The spatial coordinates of both low and high resolution models are stored in SASBDB in pdb (or pdb-like) formats. In case of *ab initio* models, the pdb format is adapted in a way that each bead is defined as a dummy c-alpha atom having a defined size. The scattering computed from the models is compared to the experimental data and stored as .fit (or .fir) files. The fit against the SAS data has the same format as the three column ASCII file of the scattering curve, but includes an extra column with the calcuated intensities $I(s)$ derived from the model.

As indicated by the wwPDB SAStf it is necessary to adopt an uniform data format to guarantee information exchange between different databases. The data format chosen as such a standard is sasCIF. The format was designed in 2000 [64] as an extension of the crystallographic information file (CIF) dictionary, analogously of mmCIF [65] which is the current standard employed by the PDB for the deposition of high resolution data and models. The sasCIF format has recently been updated and extended (Kachala M. et al., in preparation) such that now it allows the description of a complete SAS experiment in a single file. Indeed a single sasCIF file includes information about: measured sample, experimental settings, authors and related publications, experimental scattering data, models and model fits as well as structural parameters (all the information related to a SASBDB entry). This set of information is automatically extracted from SASBDB through a specialized library and tools developed for this purpose. At the same time users can submit sasCIF files to SASBDB, in such cases all the information contained in the file is automatically extracted to be included into SASBDB. Despite being hardly human-readable, the CIF format presents an advantage of being conveniently machine-readable and therefore easy to parse with computational approaches.

Table 4.1: Structural parameters stored in SASBDB

| Parameter | Based on | Abbreviation | Units |
|---|---|---|---|
| Radius of gyration | Guinier approximation | $R_g^{Guinier}$ | $nm$ |
| Forward scattering | Guinier approximation | $I(0)^{Guinier}$ | $A.U.$* |
| Radius of gyration | $p(r)$ function | $R_g^{p(r)}$ | $nm$ |
| Forward scattering | $p(r)$ function | $I(0)^{p(r)}$ | $A.U.$* |
| Maximum intra-particle distance | $p(r)$ function | $D_{max}$ | $nm$ |
| Molecular weight | Guinier approximation | $MW^{I(0)}$ | $kDa$ |
| Excluded volume | Porod invariant | $V^{Porod}$ | $nm^3$ |
| Molecular weight | Sequence | $MW^{expected}$ | $kDa$ |

* A.U. = Arbitrary unit. Data in absolute scale ($cm^{-1}$) can also be stored

## 4.2 Data display in SASBDB

The data, the models and the structural parameters previously described are presented in SASBDB in a way that maximizes the information content and gives indication about the data quality. In the scattering plot $s$ vs. $I(s)$ (figure 4.1,**E**) the information content is color coded. Most of the scattering intensities are dark-blue except for the initial very-low angle data points affected by the beam-stop (color-coded in light-blue) and the negative intensity points that may occur at high-angles (color coded in red). In order not to distort the plots and facilitate the comparison of different scattering patterns, the relative height of the $s$ vs. $I(s)$ plot is kept fixed and the width is directly dependent on the angular range (one unit-one pixel). A scroll bar allows to display plots larger than 6 $nm^{-1}$ (e.g. WAXS data). Aside from the primary reduced and solvent-subtracted scattering data, four other plots are also displayed on the main SASBDB entry. (i) The Guinier plot (Figure 4.1,**K**) that comes with a red line representing the linear fit to the Guinier region, the slope of which relates to the $R_g$ of the scattering particle. (ii) The calculated pair-distance distribution (Figure 4.1,**M**) is filled to facilitate the visualization of its shape. (iii) Whenever the$R_g$ and the *I(0)* are available, the normalized (or dimensionless) Kratky plot ($(sR_g)^2 I(s)/I(0)$ vs. $sR_g$) replaces the standard Kratky plot and is displayed (figure 4.1,**L**). As explained in the first chapter, such a plot is particularly convenient since it is independent of the particle size. Standard values associated with the plot, for example the values on the x and y-axis associated with a sphere (respectively $\sqrt{3}$ and 1.104) allows one to estimate the general compactness of a macromolecule. (iv) The plots that report the model-fits against the scattering data display the experimental points in blue and the calculated scattering intensities of the model in red (Figure 4.1, **F**, **G**).

Spatial models are displayed in the SASBDB interface as highly rendered 2D images that, upon clicking, become pseudo 3D interactive JSmol windows. In the interactive mode it is possible to rotate and zoom into and out of the models. *Ab initio* models (figure 4.1,**N**) can be distinguished from the hybrid models (figure 4.1,**O**) through the rendering mode. While the former are represented as densely packed gray beads, the latter, based on atomistic structures, are displayed as rainbow colored ribbons. In the *ab initio* models the bead size can be defined by the user whenever it is not available from the pdb file (in the latter case it is automatically extracted). Whenever present, glycans and non-bonded atoms are also displayed in the high resolution hybrid models. The models obtained through a combination of the two approaches i.e. containing high and low resolution portions (e.g. CORAL and BUNCH models) are displayed accordingly: the high resolution parts as ribbons and the *ab initio* pieces as beads.

## 4.3 Data validation

The validation and the quality assessment of deposited data in SASBDB is performed manually by database curators using a defined checklist list outlined by Jacques and Trewhella [66] and Jacques *et al.* [62]. These quality checks include:

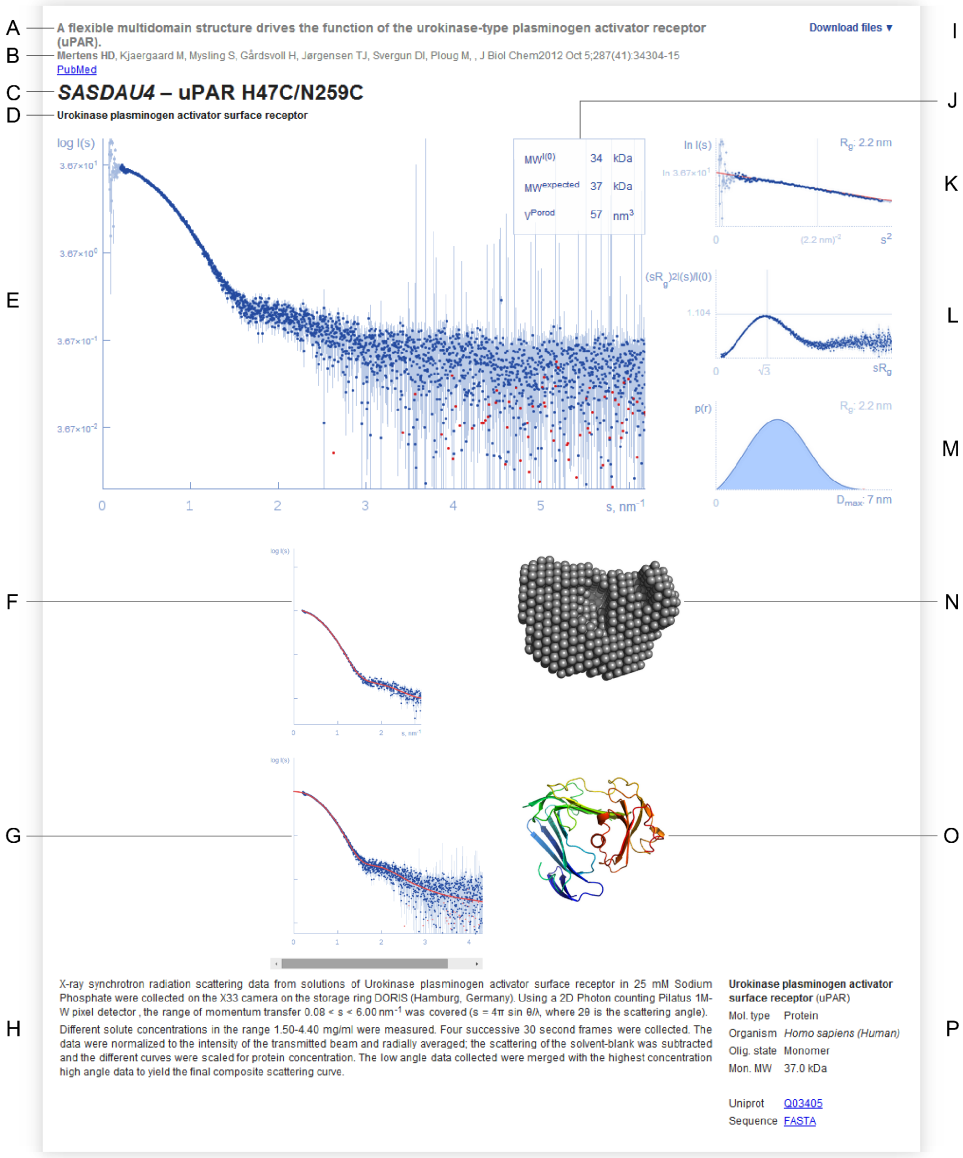


Figure 4.1: (Caption next page.)

Figure 4.1: (Figure on the previous page). Detailed representation of a SASBDB entry (`http://www.sasbdb.org/data/SASDAU4/`). **(A)** Title of the publication (or project in case of unpublished data). **(B)** List of authors, main contributor(s) to the SAS results highlighted in bold; journal reference and a link to PubMed if available. **(C)** SASBDB ID and the title of the entry. **(D)** Name of the macromolecule(s) that was(were) measured. **(E)** Experimental scattering data (Log-linear plot). **(F, G)** Fit to the experimental data from the respective model(s) displayed on the right. **(H)** Experimental details, buffer and a brief description of the data reduction steps. **(I)** Drop-down list of files available for download. **(J)** Summary of the overall parameters: $MW$ estimated from the forward scattering intensity at zero angle, $MW$ expected based on the sequence and the oligomeric state, Porod volume. **(K)** Guinier plot with the linear fit and estimated values of the forward scattering $I(0)$ and the radius of gyration $R_g$. **(L)** Dimensionless Kratky plot. **(M)** Distance distribution function and the maximum particle dimension $D_{max}$. **(N)** *Ab initio* model. **(O)** Hybrid model. By clicking the models, the 2D rendition is converted into an interactive 3D mode. **(P)** Biological details of each measured macromolecule, sequence and a link to UniProt if available.

1. The agreement between structural parameters calculated through different approaches, e.g. $R_g$ derived from Guinier plot and from the $p(r)$ distribution.

2. The agreement between the experimental and the expected $MW$. The former can be derived by $V_p$ or *I(0)*, the latter is based on the macromolecular sequence.

3. The quality of the Guinier region evaluated by AutoRg. The quality estimation is based on the number of points, the chosen range and the linearity of the plot.

4. The quality of the $p(r)$ distribution evaluated by GNOM [26]. The quality estimation is based on the fitting, the point range and the value chosen for the $\alpha$ value (see Equation 1.21).

5. The quality of the scattering curve including the signal-to-noise ratio and the number of negative points.

6. The goodness-of-fit of any model scattering curve and the experimental one. The goodness-of-fit is currently quantified by two values: reduced $\chi^2$ and the p-value computed from the Correlation Map [67].

Quality checks are performed whenever a new entry is submitted and selected parameters are made available on the web interface e.g. the agreement

among the different $MW$ (see frame in Figure 4.1,**J**) and $R_g$ estimations. Those data, that have been positively assessed, are made publicly available for the community.

## 4.4 Implementation: database and tools

Each submission to SASBDB is defined as a project or, whenever the data are published, a publication. A project can contain one or several scattering curves. Each scattering curve corresponds to an entry in the database (Figure 4.2 shows the entry structure). Each entry is characterized by a unique ID in the format SASXXXN where X are alphanumeric characters and N is a digit. Multiple models of different types can be associated into a single scattering curve through the corresponding fits to the data. In addition, multiple models can be associated to a single fit to the data, in order to accommodate mixtures or polydisperse systems e.g. an ensemble described by EOM. In turn, the sample is defined as a combination of a buffer (and related additives) with one or multiple macromolecules in order to accommodate the description complexes (or mixtures) where each macromolecule is a component of the complex (or mixture). A set of metadata is linked to the scattering curve that include publication details, experimental conditions, instrument used, detector characteristics and information about the user who submitted the data (defined as contributor).

The complete relational database developed in MySQL is shown in Figure 2.3, chapter 2. The SASBDB database is directly linked to the ATSAS database of users in order to allow the users with an ATSAS account to use their credentials to register for SASBDB. In addition, the data from UniProt [58] and PubMed are automatically fetched using the WSDbfetch application programming interface (API) from PDBe [68] for the former and a BioPython [69] tool for the latter source.

SASBDB has been implemented in python 2.7 (`http://www.python.org`) using the web application framework Django 1.6 (`https://djangoproject.com`). Django was preferred to other frameworks because is scalable, stable and enables the programming of database-driven websites. The framework is based on three main layers: the model, the view and the template. In the model the database is defined as a set of python objects, the view is a medium between the database and the interface and is used to select which data to display in which URL and, finally, the template defines how the data are presented.

Plots and model images are automatically generated using, respectively, gnuplot (`http://gnuplot.info`) and PyMOL (The PyMOL Molecular Graphics System, Version 1.7.0.1 Schrödinger, LLC). The interactive 3D model visualization is based on JSmol (`http://wiki.jmol.org/index.php/JSmol#JSmol`). JSmol was preferred to the more widespread Jmol (`http://www.`

jmol.org) because, being a JavaScript applet, it doesn't require any Java authorization. Database searching is implemented using the Django extension Haystack 2.0 (http://haystacksearch.org/) and Elasticsearch 0.90.2 (http://www.elasticsearch.org/) as a search backend engine which continously indexes all the data submitted to SASBDB.

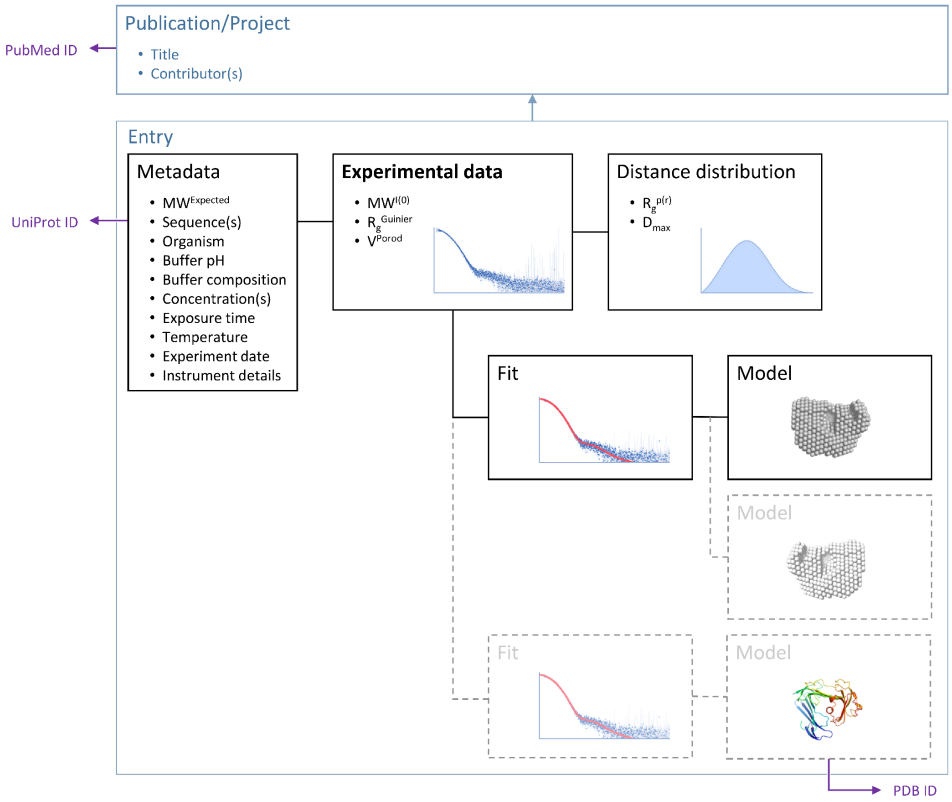


Figure 4.2: SASBDB entry organization. Multiple entries in SASBDB can be linked to a single project (or an article when published). Each entry is linked to a set of metadata ranging from macromolecular sequence information to details about the instrumentation used for the experiment. Each entry is also related to one or multiple fit(s) from one or multiple models. Whenever available the model can be linked to a PDB ID, the molecule can be linked to an UniProt ID and publications can be linked to their PubMed IDs.

## 4.5 Availability, searching and browsing

SASBDB is available at www.sasbdb.org. Registration with email address and password is required only to submit new data. ATSAS-online users can

use their existing credentials. Once registered, the users have access to an online submission form to be filled with the relevant information about the SAS experiment they want to submit. The submission form is divided into three main categories:

1. Information about the sample and publication references (if applicable).

2. Experimental details (hardware, software, sample, solvents, etc).

3. Experimental results consisting in derived structural parameters, fitting files and models.

Users have the possibility of either making the deposited entry immediately available for the community ("public entry") or receiving a private link ("on hold entry") to be shared with article referees or collaborators. Once the submission form is completed, the user receives a reply and the entry is passed to a SASBDB curator for evaluation. The curator communicates the assigned SASBDB accession code as well as any additional requirements necessary to complete the submission process, e.g., missing information or data/model clarification.

The search in SASBDB is extensive and flexible. SASBDB entries can be located in the database using a number of search criteria that include the SASBDB entry code, macromolecule name, publication title or journal, author name or affiliation, instrument details and corresponding UniProt, PubMed or PDB [10] identifiers. Two types of search interfaces are made available, "standard" and "advanced" search at the top of each page of the website.

The standard seach option consists on a search bar where seach items can be entered with few suggestions just below the bar (see Figure 4.3,**D**). The advanced search brings the user to a dedicated page `http://www.sasbdb.org/search/advanced/` with a set of fields that can be applied to restrict the search. The advanced search is expected to become more relevant as the database content increases and depositions are made by multiple users investigating the same macromolecule. As a result of the search one gets a list of corresponding entries in a compact format (see Figure 4.4) where the searched item is highlighted in yellow.

## 4.6 Web interface

The home page, the browsing page and the details page are the main user interfaces of the SASBDB website. The home page (in Figure 4.3) is designed to summarize the information about the database in brief. The home page contains statistics about the database content displayed in four pie charts (figure 4.3, **B**). By clicking on the different sections of the pie chart one is

redirected to a browsing page where the entries are filtered according to the selected section e.g. by clicking on the number of human macromolecules stored one can browse all the entries coming from human samples included in the database.

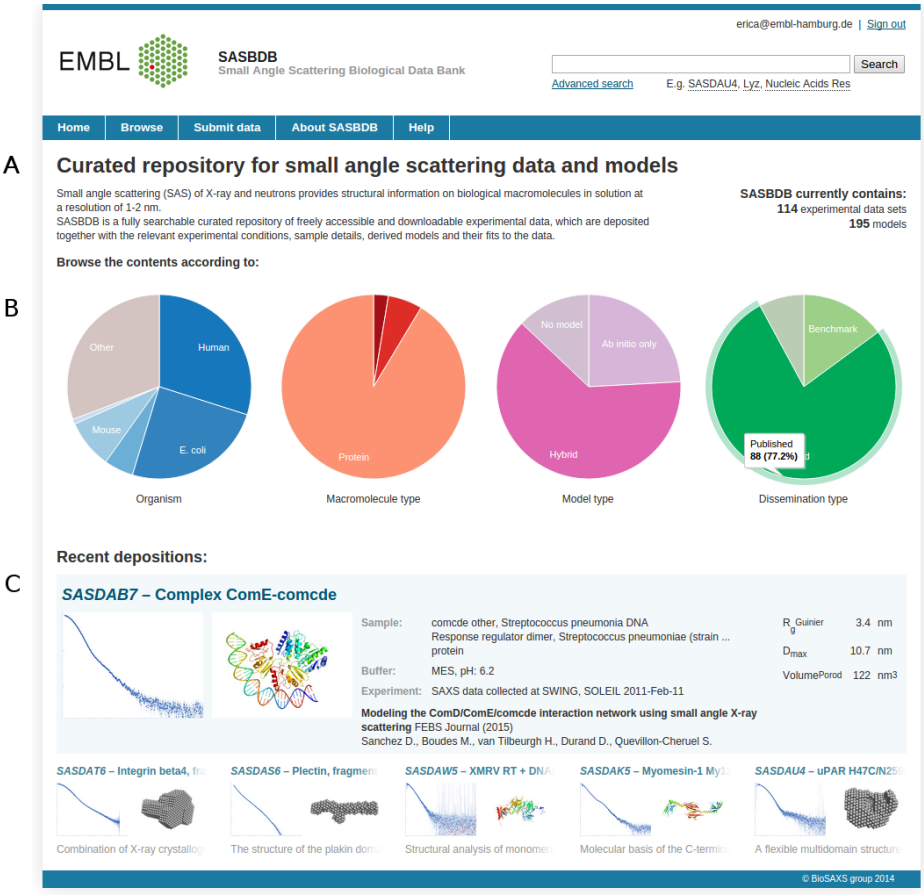


Figure 4.3: SASBDB home page can be divided in five main sections: **(A)** General description of the data bank. **(B)** Pie charts illustrating the database content (macromolecule organism and type, type of model and dissemination). **(C)** Six recently deposited entries. **(D)** Search bar with suggested items and link to advanced search page. **(E)** Number of experimental data sets and models currently stored (and publicly available).

In the browsing page each entry is displayed as a "browsing unit" which contains the overall information about the entry (see Figure 4.4). The browsing unit contains details about the measured macromolecule, buffer, instrument and publication(s) (or project for unpublished data). In addition, the browsing unit displays thumbnails of the scattering curve and of one of the models related to the entry. Whenever there is no model related to the

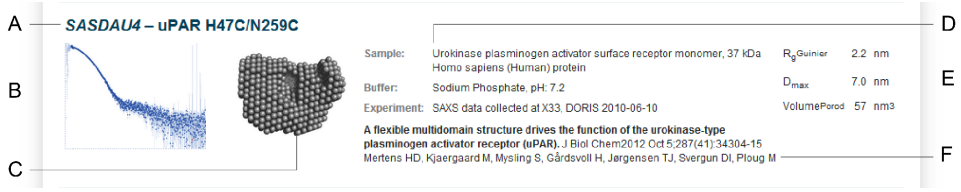


Figure 4.4: Brief representation of a SASBDB entry ("browsing unit"). **(A)** SASBDB ID and the title of the entry. **(B)** Experimental scattering data (Log-linear plot thumbnail). **(C)** One of the models. **(D)** Name of the macromolecule(s) that were measured, expected molecular weight, organism, polymer type, buffer details, experiment type, instrument and data collection date. **(E)** Structural parameters: radius of gyration, maximum particle dimension, Porod volume. **(F)** Title of the publication (or project in case of unpublished data), journal reference and authors.

entry, a Kratky plot is displayed instead. The Kratky plot has been chosen because it can be used to qualitatively assess macromolecular compactness without the need for a 3D-spatial model.

By clicking on the browsing unit one is redirected to the detail page of the entry, which displays all the information related to the selected data set (see Figure 4.4). From the detail page it is possible to download the experimental data, the distance distribution function, the fit to the experimental data and the spatial coordinates of the 3D-models as pdb files. All the listed items can also be downloaded at once as a compressed directory named after the entry code. In addition a sasCIF file of the entry can also be downloaded from this interface. Importantly, the detail page presents a set of convenient shortcuts to the search results. By clicking on the project title, the publication title or the macromolecule name one is redirected to the list of entries related to the selected item e.g. the entries related to that project. In addition, for published data, the list of authors contains one name automatically highlighted in bold, and this name corresponds to the database contributor. By clicking on the contributor name one is redirected to the list of entries submitted by the contributor. At the very bottom of the detail page there is a text summary containing information about the experimental settings. The text is automatically created out of the data submitted by the user, with possibilities to write it manually whenever preferred and/or to add comment fields uploaded by the submitter.

One of the aims of SASBDB is the dissemination of SAS as technique, and to achieve this objective two help pages have been included that inform a first time visitor about the database interface, the database content and the technique. In addition, a set of 17 SAXS and WAXS benchmark data sets is archived in SASBDB. These are experimental data from well known,

commercially available proteins, which have been further purified in order to serve as benchmark for teaching and algorithm testing purposes. Such a set of entries is accessible from the homepage by clicking on the "Benchmark" section of the "Dissemination type" pie chart (see Figure 4.3,**B**). And, as any other entry from SASBDB, the benchmark set is available for download.

## 4.7 Current status and statistics

Since the first of August 2014 when SASBDB was publicly announced, more than 2500 users have visited the website. The users were connecting to SASBDB from all over the world as illustrated in Figure 4.5

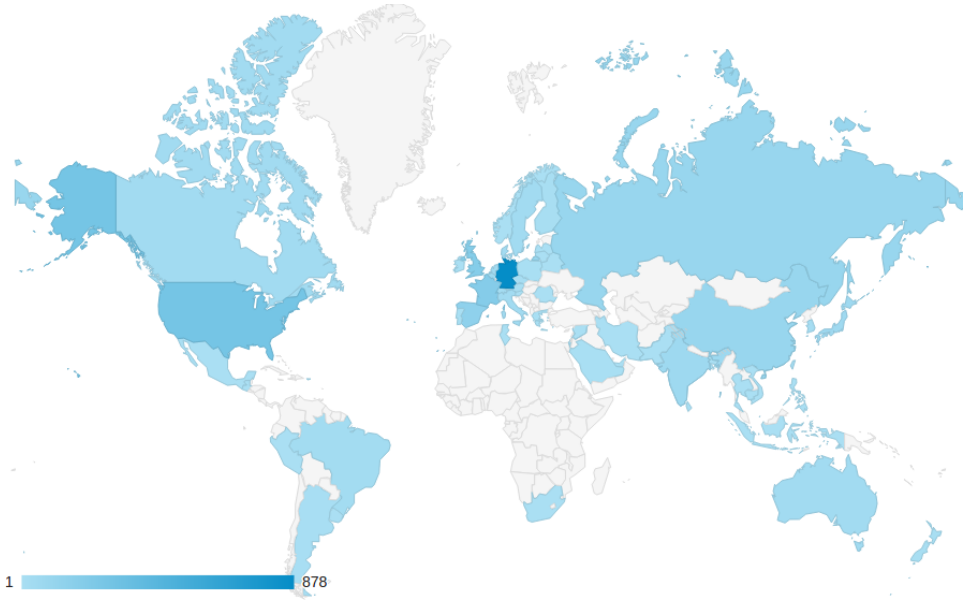


Figure 4.5: SASBDB usage around the world as of February 2015. The intensity of the blue in each country is proportional to the number of sessions started from that country.

The number of downloads as well as the most searched items are currently monitored in order to gain insight into the website usage. From a total of 334 files downloaded, the most downloaded are the .zip, the .dat and the .pdb files. This suggests that the website is visited by SAS experts and novices users –assuming that SAS experts would download .dat files while the structural biologists would look for .pdb files–. The least downloaded is the sasCIF file, which is not surprising given that this format has only been recently included at the website. It is expected that the sasCIF format downloads will increase in the future as soon as CIF will become

the standard reporting mechanism for both SAS and X-ray crystallographic data (mmCIF), (statistics about downloads is illustrated in Figure 4.6).

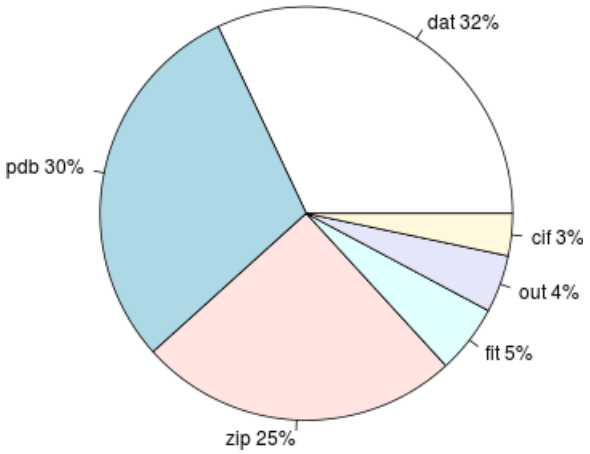


Figure 4.6: SASBDB pie chart download percentages as of March 2015. Out of a total of 334 downloads the .dat file has been downloaded 107, the .pdb 99, the .zip 84, the .fit 18, the .out 15 and the .cif 11 times.

Approximately 15% of the visitors took advantage of the search bar, the three most searched terms are: "standard proteins", "lysozyme" and "BSA". Interestingly both lysozyme and BSA belong to the set of initial entries composing the benchmark set of "standard proteins". This proves the usefulness for users of having access to such a data set of well known proteins. The large majority of the searches, however, have been unique (83%). Finally, the most visited pages of the website are the homepage, the browsing and the help page "about SASBDB".

## 4.8 Conclusions and future perspectives

SASBDB is a repository of SAS data and SAS derived models developed in agreement with the recommendations of the wwPDB SAStf and includes:

1. Publicly available and freely accessible data and models for download, including *ab initio* and hybrid models and associated fits of the experimental data.

2. Options to use sasCIF as a standard dictionary format for data exchange.

3. Storing and dissemination of SAS models with details about the software employed, symmetry and other information.

4. Quality assessments.

In addition, SASBDB contains a number of valuable features to facilitate data exchange. These features include possibility to browse the archive according to different criteria (e.g. macromolecule or dissemination type), an extensive and flexible search, highly informative plots, interactive visualization of 3D models, cross links to other biological databases (PubMed, UniProt and PDBe) and a set of highly purified standard proteins that can be used as benchmark for algorithm testing purposes. Importantly, these features are publicly available from the very beginning providing overall transparency for the scientific publication and reporting process.

This total number of entries, currently at 114 SAS data sets and 195 models, does not include the entries kept as "secret" upon the request of the contributor. Those are about 30 entries belonging to unpublished data which have been purposefully designed for confidential review purposes. In such cases, it is possible for the submitter, to have access to the SASBDB accession code and a corresponding url that may be shared with journal referees and editors. Only when permission is granted by the submitter the entry is released to the public.

As explained before in this chapter, SASBDB data undergo a set of validation steps and are, at present, manually curated and reviewed by the SASBDB maintainers. Such a curation system is currently sustainable because the number of submission is limited. Automated validation tools will eventually be required to keep pace with an expected increase in the number of depositions. A set of automatic validation procedures is currently under development (summarized in table 4.2). An automated pipeline based on the currently running at some SAXS beam lines [8] will be applied to compute structural parameters and evaluate the quality of the Guinier region and of the distance distribution function. These values will be compared to those submitted by the user in order to assess the quality and reliability of the information. New methods to evaluate the quality of both rigid body and *ab initio* models are also under development. For the hybrid models, a set of physical parameters will be assessed including comparison with interfaces of existing complexes stored in the PDB. For *ab initio* models, the beam occupancy upon averaging of multiple shape reconstructions will be analyzed. Finally, two new methods have recently been developed to evaluate the scattering curves and fits to the data. The first technique applies a Shannon sampling approach to estimate the useful range of the experimental data (Konarev et al., in press). The second method is a statistical estimation of the differences between two scattering curves in the absence of experimental errors [67]. These automatic quality assessment methods will be integrated with SASBDB.

Importantly, the wwPDB SAStf also proposed the creation of a confederation of SAS databases using a similar principle as adopted by the world

Table 4.2: Validation methods to be included in SASBDB

| Assessment of | Method | Status |
|---|---|---|
| Structural parameters | Automatic pipeline including ATSAS software. | Currently in usage at SAXS beam lines [8]. |
| *Ab initio* models | Evaluation of the beam occupancy based on the comparison of different reconstructions superimposed. | Tuukkanen et al., in preparation |
| Rigid body models | Computation of physical parameters also including the comparison with known complex interfaces. | Tuukkanen et al., in preparation |
| Useful data range | Shannon sampling based approach. | Konarev et al., in press |
| Discrepancy between data sets | Statistical method which doesn't account for errors in the data points to compute a p-value. | [67] |

wide protein data bank (PDB). The wwPDB is an organization with a single storage and deposition system, but has four confederated websites that report the data: PDBe, RCSB PDB, PDBj and BMRB. In order to achieve this objective for a future SAS database confederation, the sasCIF file format has been updated and included in SASBDB. This format will facilitate the exchange of information between current and future SAS database hubs (e.g. BIOISIS). Integration of automated quality checks is also expected to contribute to a highly advanced submission interface, which could at ultimately serve as a single deposition system for the confederation.

# Chapter 5

# Examples of SASBDB to report the results of User projects

Some of the entries stored in SASBDB come from projects where the author of this dissertation participated in the different stages of the SAXS experiments and data analysis. The first project was published early in 2015 [70] and focuses on two related proteins, TIP5 and BAZ2B that interact with histone tails. Histones are proteins with long disordered tails responsible for DNA packaging and gene regulation. The second project describes the interaction and global structural transitions that occur when a heterodimeric complex of transcription factors (PREP1 and PBX1) interacts with a 22 nucleotide (nt) DNA fragment. These transcription factors are involved in embryo development and tumorigenesis. The third project describes scattering experiments performed on a RNA construct named aptamer, developed to specifically realease drugs and other molecules to the binding partners.

In this chapter the three experiments will be described in detail with emphasis on the SAS components of the projects and reporting in SASBDB.

## 5.1 SAXS analysis of the Human TIP5 PHD Finger and Bromodomain of the Chromatin Remodeling Complex NoRC

*C. Tallant, E. Valentini, O. Fedorov, L. Overvoorde, F. M. Ferguson, P. Filippakopoulos, D. I. Svergun, S. Knapp, and A. Ciulli. Structure , pp. 1-13, Jan. 2015.*

In this study SAXS was applied to determine the differences in solution conformations of the plant homeodomain (PHD) finger and bromodomain (BRD) of the chromatin nucleolar remodeling complex (NoCR) in two human proteins: TTF-I interacting protein 5 (TIP5), also known as BRD adjacent to zinc finger domain protein 2A (BAZ2A) and the second BRD adjacent to zinc finger domain protein 2B (BAZ2B).

### 5.1.1 Introduction

The chromatin remodeling complex NoCR is known to interact with promoters of ribosomal DNA (rDNA) in order to silence the transcription of ribosomal RNA (rRNA). This action is accomplished by promoting the acetylation of histone tail H4, which triggers heterocromatin formation. The largest component of the NoCR in human, TIP5, is a 211 kDa protein consisting of a number of domains, including protein-protein and protein-DNA interacting domains. The human protein BAZ2B (240 kDa) shares 29% sequence identity with TIP5 and is known to be also involved in transcriptional regulation [71]. Both TIP5 and BAZ2B contain a PHD zinc finger adjacent to BRD on the C terminus. These domains have a yet higher sequence identity, increases up to 50%–65% suggesting similar functions. The C terminus domains PHD-BRD are evolutionary conserved and have been found in several proteins associated with chromatin remodeling functions (see Figure 5.1).

Both domains of the PHD-BRD tandem are known to interact with histone tails [72], but the structural details about those interactions are still unclear. In particular a question arises whether the domains engage a single histone tail utilizing a *cis*-type interaction or if each domain interacts with a different histone tail using a *trans*-type interaction (see Figure **??**).

In the present study different structural techniques have been applied. Biophysical methods were used to screen different histone epigenetic modifications and identify, which ones promote binding to the PHD and to the BRD domains in both TIP5 and BAZ2B. X-ray crystallography was utilized to determine the structure of the complexes of the single subunits with selected histone tails. SAXS was employed to determine low resolution structures of the two domains (including the linker regions) in solution both alone and in complex with the histone tail in order to understand the type of interaction are they engaged (i.e. to distinguish between cis and trans

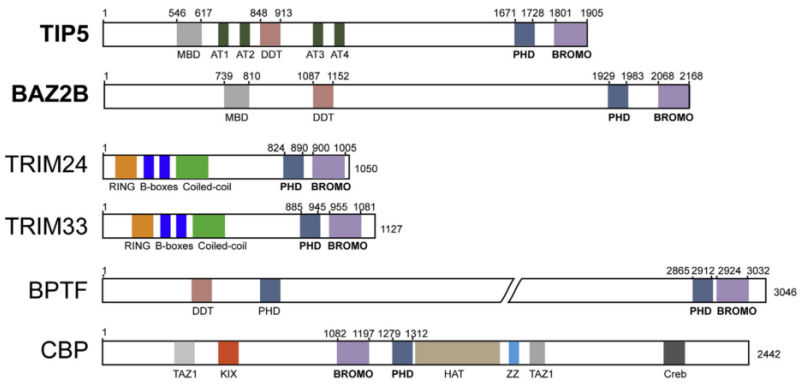


Figure 5.1: Conservation of the PHD and BROMO domain in different proteins involved in chromatine remodeling functions. Figure extracted from [70].

interactions).

### 5.1.2 SAXS analysis

In both TIP5 and BAZ2B the tandem domains PHD and BRD are connected by a linker, 65 and 70 amino acids long, respectively. Different bioinformatics approaches (IUPred and DISprot) identify two IDRs in the two linkers and estimate a higher disorder predisposition for BAZ2B. The flexible linkers may have prevented the crystallization of the two full constructs, and in such cases SAXS can be applied to determine not only the structural conformation but also the flexibility of the full constructs. The two tandem constructs (PHD and BRD domains with intra-domain linkers in TIP5 and BAZ2B) were measured both alone and in complex with peptides (histone tails). While BAZ2B in complex with the histone tail H4 acetilated on the K14 yelded an interpretable scattering profile, TIP5 complex formation was apparently not sufficiently specific, the scattering data were not interpretable and therefore were excluded from this study. The results of the experiments have been stored in SASBDB with the accession codes: SASDA46 (TIP5), SASDA56 (BAZ2B) and SASDA66 (BAZ2B in complex with histone tail).

The structural parameters derived from the experiments are summarized in table 5.1. Molecular weight estimates determined from the SAXS data indicate that both TIP5 and BAZ2B are monomeric in solution. From the $D_{max}$ and $R_g$ estimations, BAZ2B is more extended than TIP5. From the appearance of the normalized Kratky plots in SASBDB (in Figure 5.3 plots framed in red) there is a clear difference between the two proteins. While TIP5 displays a slow decay in $s^2I(s)$ after a peak, the BAZ2B Kratky plot

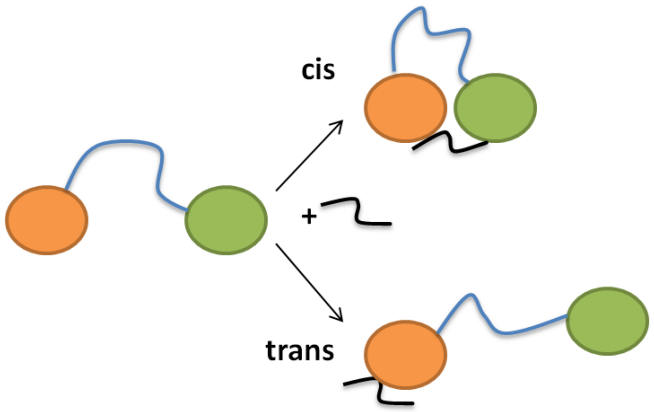


Figure 5.2: Cis and trans conformation for PHD and BRD tandem domains. Since both domains are known to interact with histone tails, they could engage a cis conformation where both domains interact with the same histone or a trans conformation having only one domain interacting.

does not show a show a maximum, suggesting that BAZ2B is either more extended than TIP5 or displays significant conformational flexibility corroborating the bioinformatics predictions previously described. Still none of the two protein reveals a Kratky plot typical of a globular folded protein. Indeed the Kratky plots of both proteins and the asymmetric $p(r)$ distributions are typical of elongated macromolecules (see Figure 5.3, plots framed in green).

Table 5.1: Structural parameters resulting from SAXS experiment

| Protein(s) | $R_g$ | $D_{max}$ | $MW^{I(0)}$ | $MW^{expected}$ |
|---|---|---|---|---|
| TIP5 | $3 \pm 0.1$ nm | $10 \pm 1$ nm | $28 \pm 5$ kDa | 27 kDa |
| BAZ2B | $4.2 \pm 0.1$ nm | $14.5 \pm 1$ nm | $33 \pm 5$ kDa | 28 kDa |
| BAZ2B, H3K14ac | $4.5 \pm 0.1$ nm | $16 \pm 1$ nm | $45 \pm 5$ kDa | 32 kDa |

EOM was applied to further investigate the flexibility of the linkers. As explained in chapter 3 EOM selects an ensemble of conformations fitting the experimental data from a pool covering the whole conformational space. In case of the TIP5 construct, EOM selected an ensemble with $R_g$ and $D_{max}$ distributions narrower and with lower average values compared to the pool (as shown in Figure 5.4, **C**, **D**, **E**). These results suggest that the TIP5 ensemble is more compact and the linker less flexible then a random pool. Different results were obtained for the BAZ2B tandem construct. In this case the ensemble selected by EOM has a distribution of $R_g$ and $D_{max}$ as

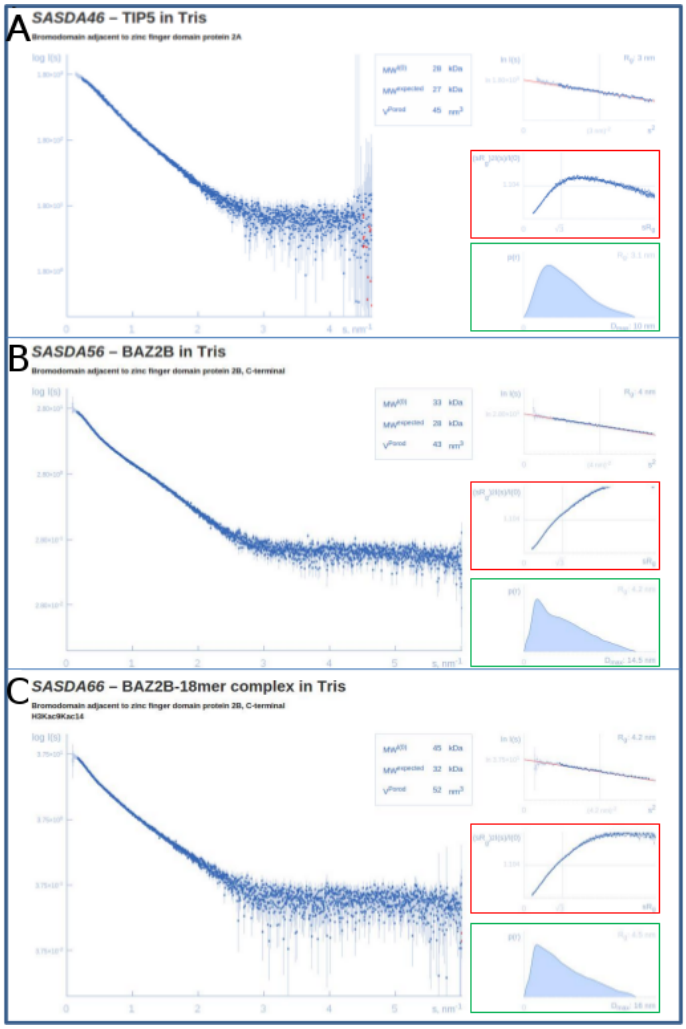


Figure 5.3: Comparison of the entries: SASDA46 (**A**), SASDA56 (**B**) and SASDA66 (**C**). The entries have been extracted from the SASBDB pages: `http://www.sasbdb.org/data/SASDA46/`, `http://www.sasbdb.org/data/SASDA56/` and `http://www.sasbdb.org/data/SASDA66/` respectively. For each entry, from left to bottom right, are shown: the scattering plot, the Guinier region, the normalized Kratky plot and the distance distribution function $p(r)$. For visualization purposes the normalized Kratky plot has been framed in red, while the $p(r)$ in green.

wide and extended as the pool (as shown in Figure 5.4, **F**, **G**, **H**). This suggests that, in contrast to TIP5, BAZ2B is likely very flexible with the two PHD and BRD domains connected by a disordered linker.

Once established that TIP5 presents a limited flexibility, an *ab initio*

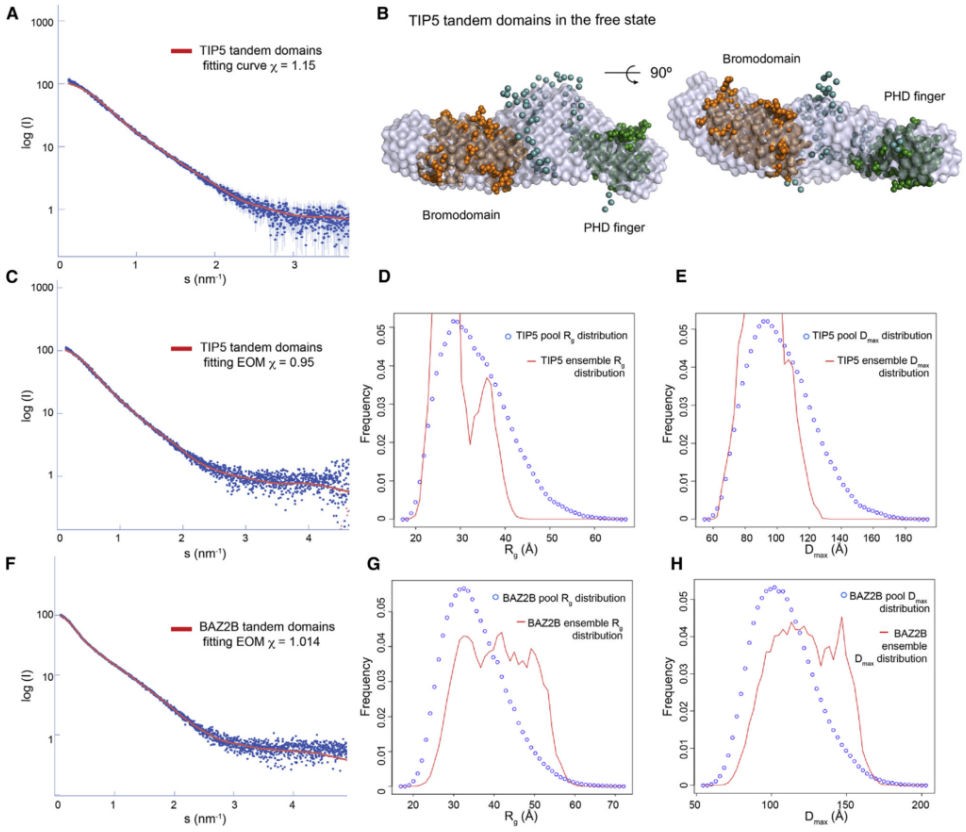


Figure 5.4: Molecular Basis of Histone Tail Recognition by Human TIP5 PHD Finger and BRD of the Chromatin Remodeling Complex NoRC: SAXS results. **A** Fitting of the BUNCH model of the TIP5 tandem domains (red curve) on the experimental data in blue. **B** Superimposition of the BUNCH model on the DAMMIF model of the TIP5 tandem domains. **C** Fitting of the EOM selected ensemble for the TIP5 tandem domains. **D** $R_g$ distribution of the EOM selected ensemble in red line compared to the pool $R_g$ distribution in blue dots for the TIP5 tandem domains. **E** $D_{max}$ distribution of the EOM selected ensemble in red line compared to the pool $D_{max}$ distribution in blue dots for the TIP5 tandem domain. **F** Same as **C** but for BAZ2B tandem domain. **G** Same as **D** but for BAZ2B tandem domain. **H** Same as **E** but for BAZ2B tandem domain. Picture extracted from [70].

(DAMMIF) reconstruction was performed (such a reconstruction would have been meaningless for a disordered protein). The DAMMIF models that fit the data well ($\chi = 0.948$) are all characterized by a slightly bent ellipsoid shape (see Figure 5.4, **B**). Since the high resolution structures of the two PHD and BRD domains were available, BUNCH was also used to recon-

struct an atomistic hybrid TIP5 model using single domain high resolution structures as rigid bodies and reconstructing *ab initio* the conjunction linker with a string of 65 dummy beads. The model obtained using BUNCH fits well the experimental scattering curve ($\chi^2 = 1.15$ ) and spatially aligns with the corresponding DAMMIF reconstruction (as illustrated in Figure 5.4, **A**, **B**).

The SAXS data were also acquired from TIP5 and BAZ2B constructs in complex with histone tails, interpretable data were obtained only for BAZ2B. The BAZ2B-H3K14ac histone tail complex has an increased $MW$ of 45 kDa as expected for a complex formation, and both $R_g$ and $D_{max}$ also increase compared to the BAZ2B free construct ($R_g$ from 4.2 to 4.5 $nm$ and $D_{max}$ from 14.5 to 16 $nm$) (see Figure 5.3, **C**). Such an increase is not compatible with the engagement of both domains in an interaction with the histone tail (cis-type) which would have led at least to a decrease in the $D_{max}$. The SAXS results therefore clearly point to the trans-type of the interaction.

### 5.1.3 Conclusions

The SAXS experiments indicate that both TIP5 and BAZ2B have extended conformations with the two globular domains PHD and BRD separated by a linker. The latter was found to have a high degree of flexibility in case of BAZ2B, while being rather rigid in the TIP5 construct. The TIP5 tandem construct was also modeled using both an *ab initio* and a combination of rigid body modeling and *ab initio* reconstruction.

The SAXS analysis of BAZ2B in complex with H3 histone tail suggest an extended conformation with the two globular domains distinctly separated. Such a conformation corroborates the hypothesis of the two globular domains being involved in different interactions with different histone tails, rather than both domains interacting with a single peptide. The SAXS experimental data and derived models describing these results are archived in SASBDB.

Further analysis in this study included a biophysical screening to identify the histone tails with (and without) specific post translational modifications(PTMs), more prone to form complexes with each of the globular domains PHD and BRD. The complexes of the identified peptides with the corresponding domains have been crystallized and structurally analyzed using X-ray crystallography.

The NoCR complex engages in the interactions with the histone tails in order to silence specific rDNA genes, and the described study gives a valuable structural insight in this mechanism.

## 5.2 The complex between the homeodomain proteins PREP1 and PBX1 has an elongated structure formed by hydrophobic interactions and stabilised by DNA

This work describes conformational changes of the heterodimer pre-b-cell leukemia homeobox (PBX1) and PBX regulatory protein (PREP1) –also known as PBX/knotted 1 homeobox (PKNOX)– on forming a complex with DNA. A number of biophysical techniques have been applied to analyze changes in overall structure, secondary structures and to determine, which amino acids are involved in the protein-protein heterodimeric interface. SAXS analysis revealed changes in flexibility and in the overall shape upon complex formation with DNA. For this work an article is in preparation in collaboration with Chiara Bruckmann and Francesco Blasi from IFOM-IEO Campus (Milan, Italy).

### 5.2.1 Introduction

The homeodomain proteins are a group of transcription factors characterized by a conserved 60 amino acids sequence and by a typical H-T-H (Helix-Turn-Helix) fold where three alpha-helices are connected by a loop 5.5 [73]. The TELE proteins share the typical homeodomain fold but present a three-aminoacids extension (P-Y-P) between the first and the second helix (hence the acronym three amino acids loop extension (TELE) [74]). The TELE class of transcription factors includes PREP1 and PBX1 that are involved in embryo development. The knock out of *Pbx1* or *Prep1* in mice prevents the embryogenesis generating lethal phenotypes [73]. The formation of the PREP1-PBX1 heterodimer has been shown to regulate the transcription of genes involved in embryonic patterning in both vertebrates and insects. In addition, both transcription factors play key roles in tumorigenesis. Indeed, the suppression of PREP1 has been related to lung, breast and colon cancers and the binding of PBX1 to PREP1 has tumor-inhibiting effects (Bruckmann et al., in preparation).

The importance of the complex PBX1-PREP1 for embryo development and tumorigenesis makes it an ideal drug target candidate. Yet to date, the heterodimer protein-protein interface, as well as overall complex with DNA, have been investigated only via indirect methods [75]. High resolution structures are currently available but they cover only the single homodomain with flanking amino acids (e.g. pdb 1ADH, see figure 5.5). The full-length proteins have never been crystallized, possibly because the amino acids outside the structured domains are disordered (a significant level of disorder is predicted). For these reasons SAXS was applied, together with other biophysical techniques like thermofluor, circular dichroism (CD) spectra and isother-

Figure 5.5: Homeodomain "Antennapedia" from *Drosophila melanogaster* bound to a fragment of DNA. The three helices, trademark of the homodeodomain, are shown in cyan while the unstructured flanking amino acids interacting with the DNA are in magenta. The DNA is displayed as colorful sticks. PDB code 1AHD.

mal titration calorimetry (ITC), to investigate the overall structure of the heterodimer and to detect changes in the structure upon DNA binding. The results of the SAXS experiments are archived in SASBDB with the accession codes: SASDAP7 (PBX1:PREP1 complex), SASDAQ7 (PBX1:PREP1 and DNA complex) and SASDAR7 (DNA).

### 5.2.2 SAXS results

The SAXS experiments were performed at the beam line P12 of the synchrotron PETRA III in Hamburg (Germany). Both samples, PBX1:PREP1 and PBX1:PREP1 + DNA, were measured at different concentrations, ranging from 1 to 15 mg/ml. The scattering data (compared in Figure 5.6) point to the development of concentration-dependent aggregates, and the latter are more evident for the heterodimeric complex than for the tertiary complex formed on binding the DNA. The *MW* estimates extracted from the

SAXS data indicate that the protein-protein complex and the corresponding complex with DNA are monodisperse at the lowest solute concentrations as shown in Table 5.2. The 22mer DNA sample was also measured for modeling purposes and resulted in a calculated Porod volume value in agreement with a double stranded DNA (Table 5.2). The scattering data collected from the protein-DNA complex and from the DNA molecule alone allowed a multiple-phase *ab initio* modeling of the protein-DNA complex described in the next section.

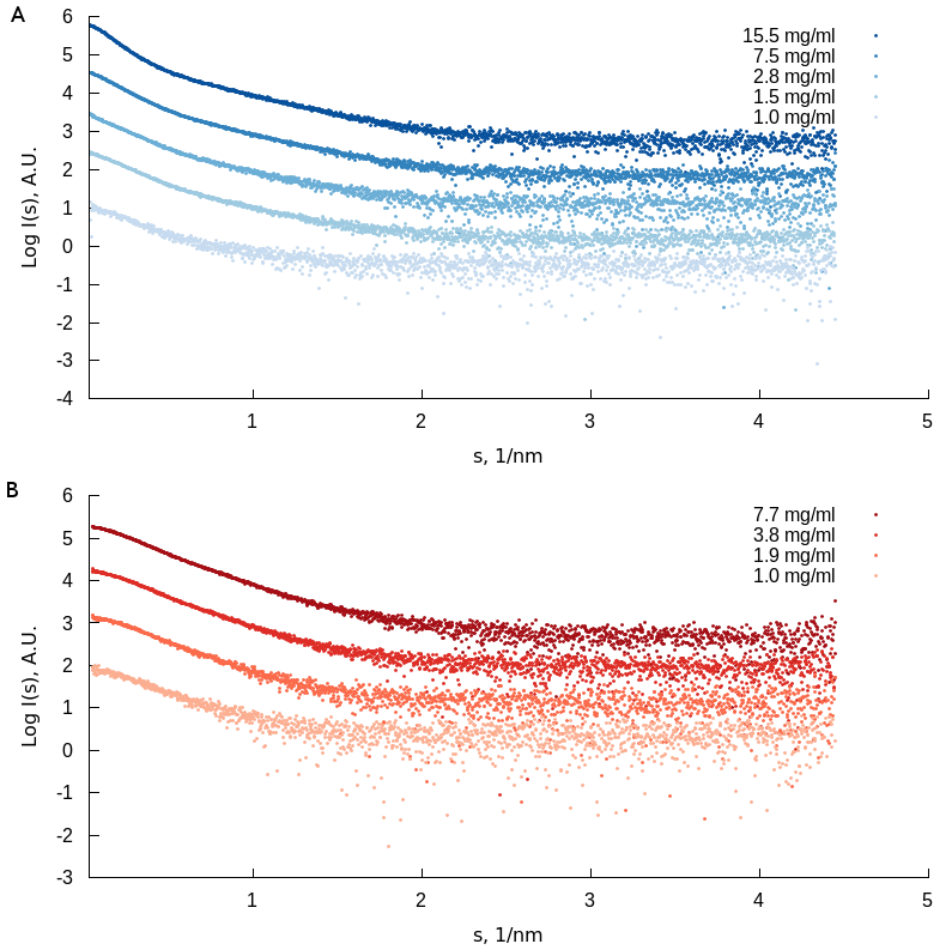


Figure 5.6: Concentration dependence of the PBX1:PREP1 (**A**) and PBX1:PREP1-DNA complex (**B**). The curves have been scaled for visualization purposes.

Interestingly, despite being lower in $MW$, the heterodimeric complex presents higher values for both $R_g$ and $D_{max}$ compared to the complex with DNA, as summarized in Table 5.2. The different $R_g$ and $D_{max}$ values are in

Table 5.2: Structural parameters resulting from SAXS experiment

| Molecule | $R_g$ | $D_{max}$ | $MW^{experimental*}$ | $MW^{expected}$ |
|---|---|---|---|---|
| PREP1, PBX1, DNA | $4.8 \pm 0.1$ nm | $16 \pm 1$ nm | $80 \pm 5$ kDa | 85 kDa |
| PREP1, PBX1 | $5.8 \pm 0.1$ nm | $19 \pm 1$ nm | $77 \pm 5$ kDa | 73 kDa |
| DNA | $2.2 \pm 0.1$ nm | $7 \pm 1$ nm | $19 \pm 5$ | 12 kDa |

*The $MW$ for DNA is based on the Porod volume given that $MW \approx PorodVolume$ for nucleic acids. The other experimental $MW$ are based, instead, on the $I(0)$.

agreement with a conformational transition adopted by the PBX1:PREP1 complex upon binding to DNA. The effect is evident from the comparison of the two $p(r)$ distributions where the complex with DNA seems to acquire a more compact conformation compared to the more extended conformation of the complex in absence of DNA (in the green frame of Figure 5.7**A**,**B**).

The normalized Kratky plots also provide an interesting insight into the conformational changes between the two complexes. While the normalized Kratky plot of the heterodimer shows an increase that reaches a plateau, the heterodimer in complex with DNA presents a distinct maximum that decreases in the higher $sR_g$ range indicating less flexible structure (in the red frame of Figure 5.7**A**,**B**).

The complexes were modeled using DAMMIF. Ten different models were reconstructed, subsequently averaged and filtered to produce a consensus overall shape for the PREP1:PBX1 heterodimer and its complex with DNA. The final models obtained are a rod-like elongated shape for the heterodimeric complex (Figure 5.8**A**) and a more globular ellipsoidal shape for the heterodimer in complex with DNA (Figure 5.8**B**).

Finally, a multi-phase *ab initio* program MONSA was used to model the PREP1:PBX1-DNA complex by fitting multiple curves (using the approach described in chapter 1). In this case the phases were: the heterodimeric protein-protein complex, the DNA and the solvent. Since we were aware of the conformational changes upon binding of the protein-protein heterodimer, the scattering curve of the heterodimer was not used for the MONSA reconstruction. The multi-phase *ab initio* model (Figure 5.8, **C**) shows an ellipsoidal shape with the DNA close to one of the termini (because of the radial averaging it is not clear whether it is the C- or N- termini). A portion of the DNA molecule is buried inside the complex, explaining the $D_{max}$ of 7 *nm* (about half of the $D_{max}$ of the full complex) which could appear

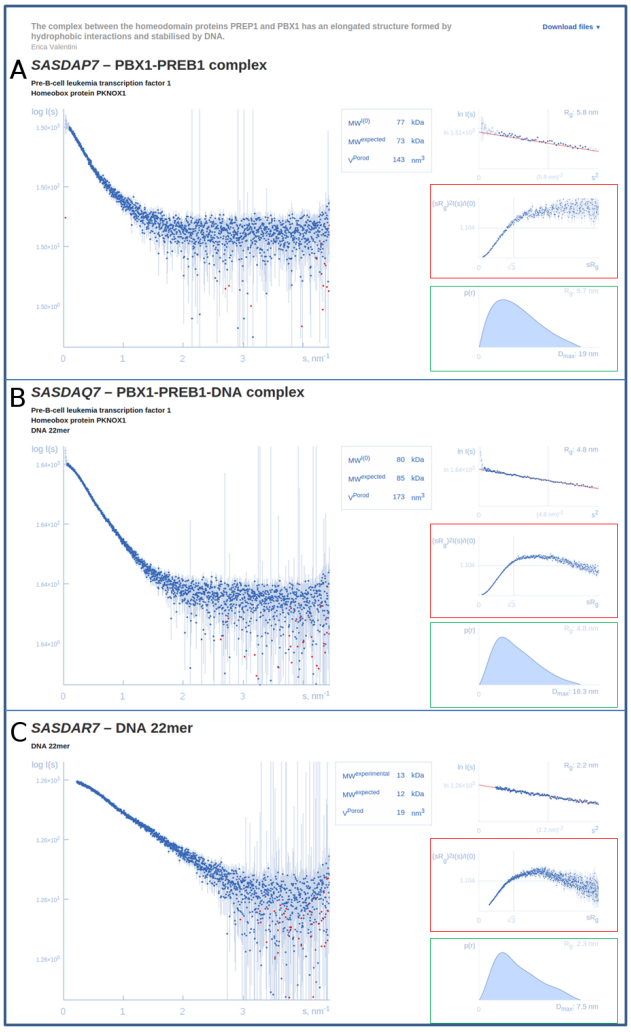


Figure 5.7: SASBDB entries related to the complex PBX1:PREP1, PBX1:PREP1-DNA and DNA. (**A**) The entry SASDAP7 illustrates the complex PBX1:PREP1 url: `http://www.sasbdb.org/data/SASDAP7/`. (**B**) SASDAQ7 illustrates the complex PBX1:PREP1-DNA url: `http://www.sasbdb.org/data/SASDAQ7/`. (**C**) SASDAR7 shows the DNA url: `http://www.sasbdb.org/data/SASDAR7/`. For visualization purposes the normalized Kratky plot is highlighted by a red frame, while the $p(r)$ distribution is highlighted by a green frame.

overestimated by looking at the 2D image.

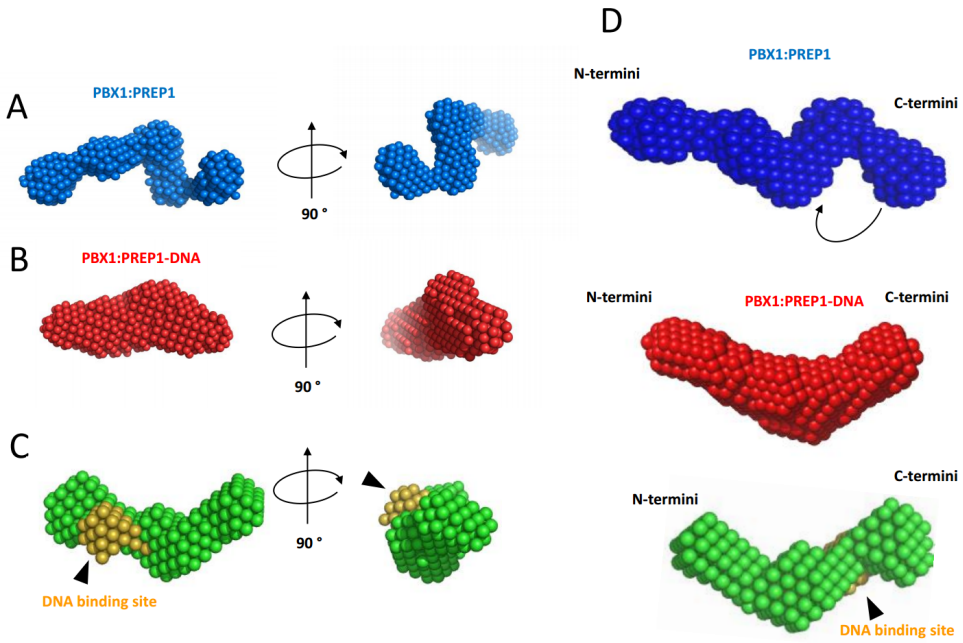


Figure 5.8: *Ab initio* model reconstruction for PREP1:PBX1 heterodimer and PREP1:PBX1-DNA complex. **A** Front and size view of the DAMMIF model of the PBX1:PREP1 complex. **B** Front and size view of the PBX1:PREP1-DNA complex. **C** Multiphase MONSA model of the PBX1:PREP1-DNA complex where the heterodimer is green and the DNA is yellow. **D** hypothetical conformational change that the C-termini of the PBX1:PREP1 complex (in blue) undergoes upon complex formation with DNA (in red). In the multiphase model instead, the heterodimeric complex is green and the DNA is yellow.

### 5.2.3 Conclusions

In this study multiple biophysical techniques have been applied to gain insight into the interface, secondary structure and overall shape of the complex formed by the TELE transcription factors PBX1 and PREP1 belonging to the homodomain family. A systematic mutational analysis revealed that the complex formation is due to hydrophobic interactions. In addition, the heterodimer undergoes a conformational change when bound to DNA, and this has been proved by a number of techniques. In particular, CD spectra analysis suggest a decrease in the alpha-helices content, and thermofluor analysis reveal an increase in stability and ITC point to a conformational transition.

The SAXS-derived structural parameters and *ab initio* models show a decrease in flexibility and a more compact structure of the heterodimer

when bound to DNA. Multiphase *ab initio* modeling suggests that the DNA crosses the complex close to one terminus of the heterodimer. One could speculate that, upon complex formation, the C-termini of the PREP1:PBX1 complex tend to close around the DNA binding site giving rise to a rather compact shape (Figure 5.8, **D**). The SAXS data and the resulting models related to this project are archived in SASBDB but they are currently not listed in the public accessible repository because the results are not published yet.

The described structural and biophysical analysis gave a precious insight into the behavior of an important class of transcription factors. Indeed the TELE proteins play key roles in the embryo development and in tumor suppression mechanisms. The development of crucial drugs could be aided by knowing which amino acids are involved in the protein complex formation and how the overall shape changes when the protein complex binds to the DNA.

## 5.3 SAXS studies of aptamer constructs

This study was focused on the structural characterization of RNA aptamer constructs. Aptamers are oligonucleotides that bind to specific substrates with high affinity and specificity. The aptamers analyzed in this study are two different RNA constructs named AIR-3 and AIR-3A. Both constructs bind to interleukin-6 receptor (IL-6R); the former is a large construct 106 nts long, while the latter is a truncated version only 19 nt long.

Both constructs, and two mutants of each of them, have been measured with SAXS together with the target protein IL-6R in the apo and holo form (in complex with AIR-3A). The results of the SAXS analysis will be illustrated in this section.

### 5.3.1 Introduction

Aptamers are DNA or RNA oligonucleotides with specific 3D shapes, which allow them to bind their targets with high affinity and specificity. The affinity is the tendency of two molecules to form a complex while the specificity is the capability to distinguish among different targets. The targets can be ions, antibiotics, proteins but also larger objects like viruses or entire cells. Aptamers have multiple applications spanning from diagnostics to delivery of specific molecules (like drug molecules) into the target cell or tissue [76]. To this last category belong the objects of this study, namely aptamers AIR-3 and its truncated form AIR-3A.

The truncated form AIR-3A has been investigated because, despite being only 19 nt long (instead of 106 nt of the full construct) it shows the same affinity and specificity for the substrate. In addition, the AIR-3A construct is characterized by a motif named guanine-quadruplex (G-quadruplex) which consists of layers and can be detected via melting point experiments and CD spectroscopy analysis [77, 78] (see Figure 5.9).

The substrate of both AIR-3 and AIR-3A is IL-6R, a natural receptor of interleukin-6 (IL-6), a multifunctional protein involved in anti-inflammatory and immune response [79]. IL-6 and its receptor are involved in many autoimmune diseases like diabetes, systemic lupus erythematosus and rheumatoid arthritis. In addition, they are also involved in cancer progression and infectious diseases. A promising therapeutic treatment to tackle these diseases is the delivery of drugs directly in the infected cells through IL-6R. Importantly, the selected aptamers have been observed to bind to the substrate even in presence of the natural ligand IL-6.

For both aptamers different modifications with artificial nt have been tested to achieve more stability in media retaining the substrate affinity. Indeed, the different modified aptamers shows different stability and affinity as illustrated in Table 5.3. SAXS has been applied to determine if the differences in stability and affinity are coupled with differences in the confor-

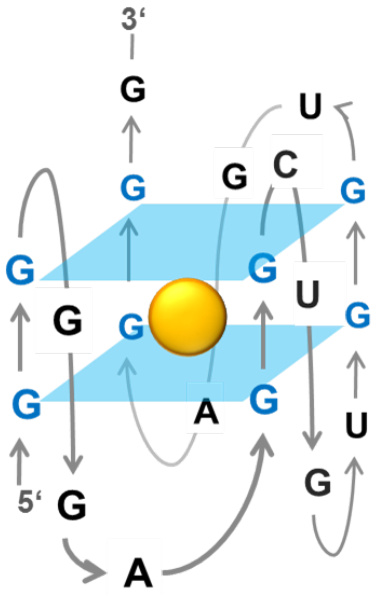


Figure 5.9: G-quadruplex motif. This motif of oligonucleotides presents two layers, each formed by four guanine nucleotides. Figure courtesy of Sven Kruspe.

mations in solution. In addition, SAXS was also applied to investigate the structure of the target IL-6R, both alone and in complex with the aptamer AIR-3A.

Table 5.3: Affinity and stability of different aptamers

| Aptamer | Number of nt | Stable in serum | Retained affinity |
|---|---|---|---|
| AIR-3 | 106 | − | + |
| AIR-3 5FU* | 106 | − | + |
| AIR-3 5FdU* | 106 | − | + |
| AIR-3A | 19 | − | + |
| AIR-3A G18U* | 19 | − | − |
| AIR-3A 2'FU* | 19 | + | − |

*Modified aptamers

### 5.3.2 SAXS analysis

The SAXS measurements can be divided in two groups: the first performed at the beam line P12 (PETRA III, Hamburg, Germany) focused on the different RNA constructs, the second performed at the beam line B21 (DI-

AMOND, Oxfordshire, UK) focused on the IL-6R protein and its complex with RNA. The usage of a different beam line was made necessary by the one-year shout down of the PETRA III synchrotron. All the results of this project have been stored in the database SASBDB.

### RNA aptamers

The aptamer measurements were thwarted by the fact that most of the RNA samples revealed a tendency to form unspecific aggregates down to the low solute concentrations, which ranged from 0.2 to 0.4 mg/ml. Still, the use of the lowest concentrations, approximating the infinite dilution conditions, allowed us to proceed with the SAXS data analysis for many not-aggregating samples despite the absence of multiple measurements of the concentration series. Initially the long (106 nt) constructs were measured. The data obtained for the unmodified construct were pointing to aggregation and impossible to analyze while good experimental data were obtained for the modified constructs AIR-3 5FU and AIR-3 5FdU. In both latter cases the position 5' of the uridines contained a modified atom (the uridines in the second construct were deoxyuridines). The structural parameters obtained for AIR-3 (summarized in the first part of Table 5.4) show $D_{max}$ and $R_g$ in agreement with a dimer formation for both constructs. In addition, both the scattering curves and the distance distribution functions suggest a bent elongated shape (see green framed plots stored in SASBDB in Figure5.10).

DAMMIF was applied to reconstruct the *ab initio* shapes of the aptamers (see Figure5.10, **A** and **B**) confirming, for both constructs, their bent elongated shapes. These structural parameters and derived models suggest a similar shape for the two modified constructs in agreement with their similar affinity and stability (shown in Table 5.3).

Successively, the shorter constructs (19 nt) were analyzed. In this case three experimental data sets presented a limited aggregation rate and could be further interpreted. The three constructs were the unmodified AIR-3A and the modified AIR-3A 2'FU with the 2' position of every uridin replaced by a fluorine atom and AIR-3A G18U with the 18th nt mutated. As expected, in this second set of measurements the aptamers appeared smaller and with a globular shape (see plots from SASBDB in Figure 5.12). The obtained structural parameters (summarized in the second half of Table 5.4) are similar to each other for the first two aptamers, and indicate a larger macromolecule for the third aptamer (AIR-3A G18U). Indeed, the $MW$ estimation suggests a dimer for the first two aptamers and a tetramer for the third. At the same time, the third construct (AIR-3A G18U) presents an higher aggregation rate, therefore it is not clear whether the estimated values derive from an higher oligomerization state or from the formation of unspecific aggregations. For the second construct (AIR-3A 2'FU) some problems were encountered in the higher angular range where the buffer

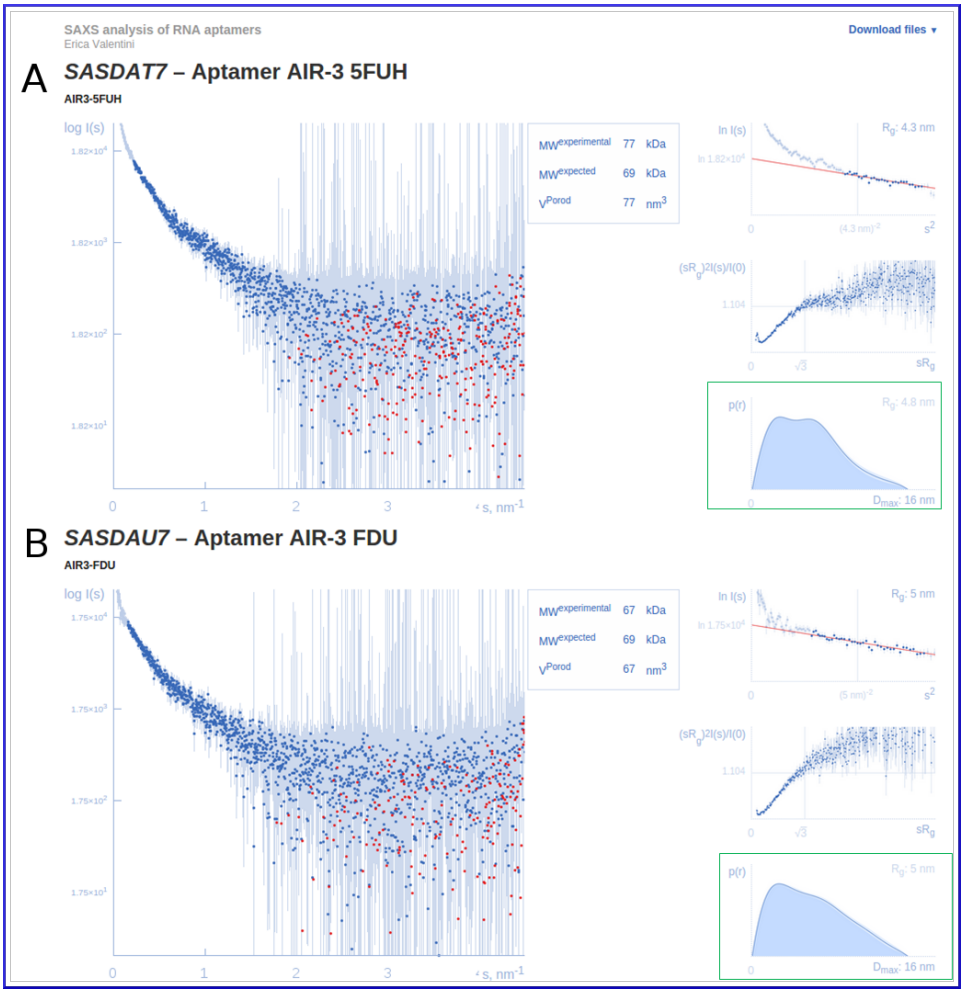


Figure 5.10: SASBDB entries of long aptamer constructs: **A** `http://www.sasbdb.org/data/SASDAT7/`, **B** `http://www.sasbdb.org/data/SASDAU7/`. The green frame highlights the distance distribution functions. The scattering curves appear noisy and slightly aggregated, but still it was possible to obtain the structural parameters.

$I(s)$ appears higher than the sample $I(s)$. This problem is known as buffer mismatch and it prevents a correct modeling based on the experimental scattering curve. For this reason further modeling was only applied to the first construct AIR-3A.

AIR-3A was firstly modeled *ab initio* using the software DAMMIN applying a P2 symmetry constrain to account for the presumed dimeric state of the construct. After it, rigid body modeling approach was applied using the software SASREF and imposing, also in this case, a P2 symmetry.

Table 5.4: Structural parameters resulting from SAXS experiment

| Molecule | $R_g$(nm) | $D_{max}$(nm) | $MW^{exp}$(kDa)* | $MW^{seq}$(kDa) | O.S.§ |
|---|---|---|---|---|---|
| AIR-3 5FU | $4.8 \pm 0.5$ | $16 \pm 0.5$ | $70 \pm 10$ | 34.5 | 2 |
| AIR-3 5FdU | $5.0 \pm 0.1$ | $16 \pm 0.5$ | $70 \pm 10$ | 34.5 | 2 |
| AIR-3A | $1.9 \pm 0.1$ | $6.5 \pm 0.5$ | $14 \pm 3$ | 6.4 | 2 |
| AIR-3A 2'FU | $1.9 \pm 0.1$ | $6 \pm 0.5$ | $12 \pm 2$ | 6.4 | 2 |
| AIR-3A G18U | $2.5 \pm 0.2$ | $8.5 \pm 1$ | $24 \pm 2$ | 6.3 | 4 |
| IL-6R | $5 \pm 0.1$ | $20 \pm 2$ | $140 \pm 2$ | 41 | 3? |
| IL-6R, AIR-3A | $6.6 \pm 0.1$ | $25 \pm 1$ | $150 \pm 10$ | 47.4 | 2:4 |

*The expected $MW$ ($MW^{exp}$) is based on the Porod volume given that $MW \approx Vol^{Porod}$ for nucleic acids and $MW \approx Vol^{Porod}/1.7$ for proteins.
§O.S. = Oligomerization state (based on $MW^{exp}/MW^{seq}$). In the table 2 stands for dimer, 3 for trimer and 4 for tetramer. 2:4 means a complex with dimeric protein and four aptamer molecules.

For the rigid body analysis, a predicted model of AIR-3A was generated using MD by our collaborator (Martin Zacharias). The two reconstructions superposed well as illustrated in Figure 5.11,**C**.

### IL-6R and AIR-3A aptamer complex analysis

The measurements of the IL-6R protein alone and in complex with the short AIR-3A aptamer were performed at the beamline B21 at the synchrotron DIAMOND (Oxfordshire, UK). For the apo form the scattering curve and the distance distribution function suggest a rather globular protein (see Figure 5.13,**A**). In addition, the structural parameters (summarized in the last part of Table 5.4) suggest the oligomerization state close to a trimer. It is important to mention that the trimeric oligomeric state can be misleading; unless indicated by the symmetry of the crystallographic packing, the sample may simply be a mixture of different oligomerization states.

The high resolution crystallographic structure of IL-6R is available in the PDB (code: 1N26) with some amino acids missing at both termini i.d. 28 amino acids in the N-termini and 29 in the C-termini. Therefore, a combina-

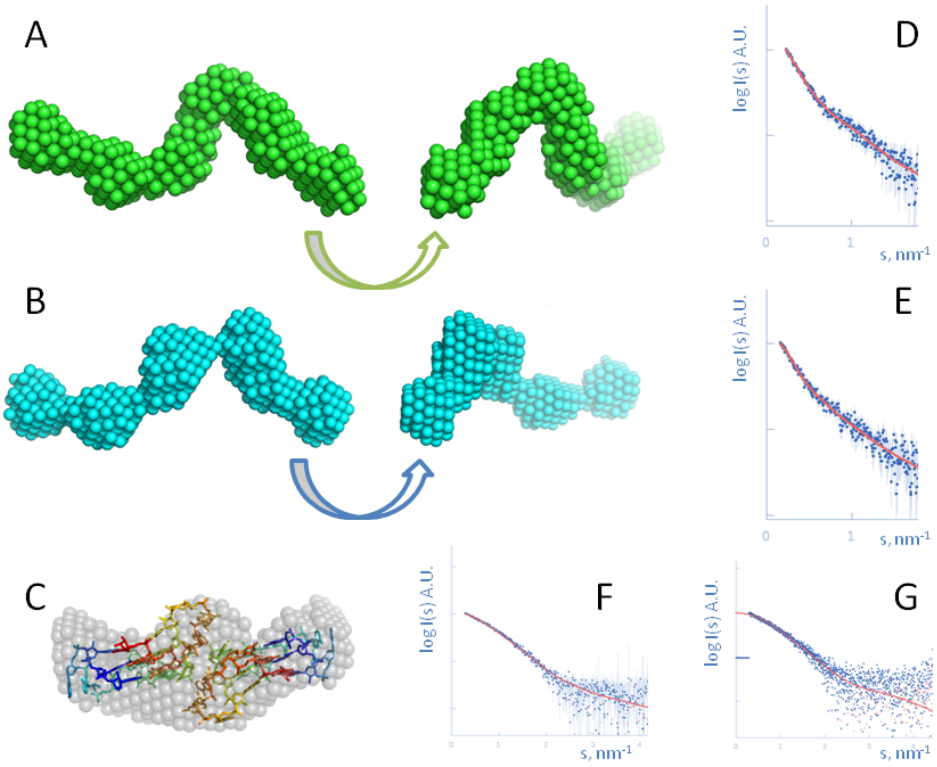


Figure 5.11: Aptamer models. **A** DAMMIF model of AIR-3 5FU front and size view, **D** corresponding goodness-of-fit ($\chi^2 = 0.887$). **B** DAMMIF model of AIR-3 5FdU front and size view, **E** corresponding goodness-of-fit ($\chi^2 = 0.850$). **C** DAMMIN model of AIR-3A superimposed to SASREF model both in P2 symmetry, **F** corresponding DAMMIN goodness-of-fit ($\chi^2 = 1.118$) and **G** corresponding SASREF goodness-of-fit ($\chi = 1.043$).

tion of the *ab initio* and rigid body approach was applied with the program CORAL in order to model the termini as chains of dummy beads. The crystallographic unit cell of the PDB structure contains a tetrameric assembly which was kept fixed during the modeling process as a rigid body. The resulting model has a sub-optimal fitting to the experimental data ($\chi^2 = 2.49$) with a discrepancy in the lower angle given by the larger size compared to the measured sample (Figure 5.14,**D**).

In order to improve the model we used OLIGOMER to understand if the measured sample displayed an equilibrium of different oligomerization states. The tetramer reconstructed with CORAL could be divided in two possible dimers with different conformations: two compact dimers and two extended dimers (see Figure 5.14). The first equilibrium tetramer-compact

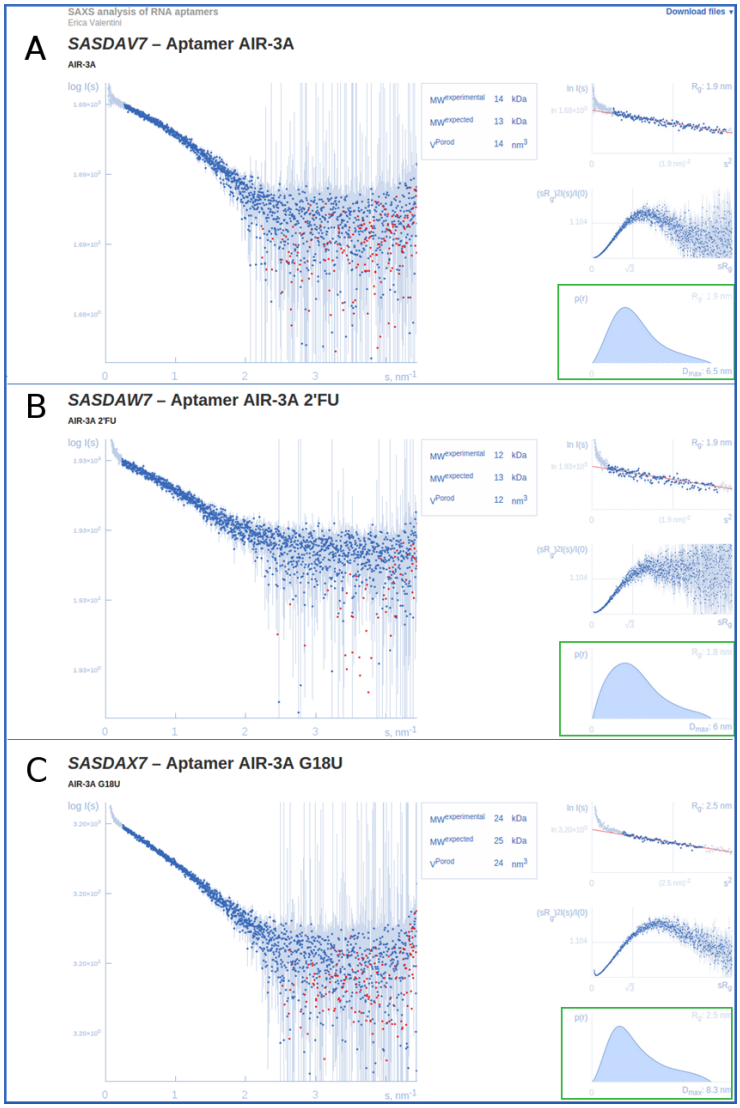


Figure 5.12: SASBDB entries of short aptamer constructs:**A** `http://www.sasbdb.org/data/SASDAV7/`, **B** `http://www.sasbdb.org/data/SASDAW7/` and **C** `http://www.sasbdb.org/data/SASDAX7/`. The green frame highlights the distance distribution functions. The scattering curves appear oversubtracted for SASDAW7 and slightly aggregated for SASDAX7, but still it was possible to obtain structural parameters.

dimer resulted in a volume fraction of 65% for the tetramer and 35% for the dimer fitting the data with the goodness-of-fit of 1.92 (Figure 5.14,**B**). The second equilibrium tetramer-extended dimer resulted in a more balanced volume fraction of 47% for the tetramer and 53% for the dimer with a sig-

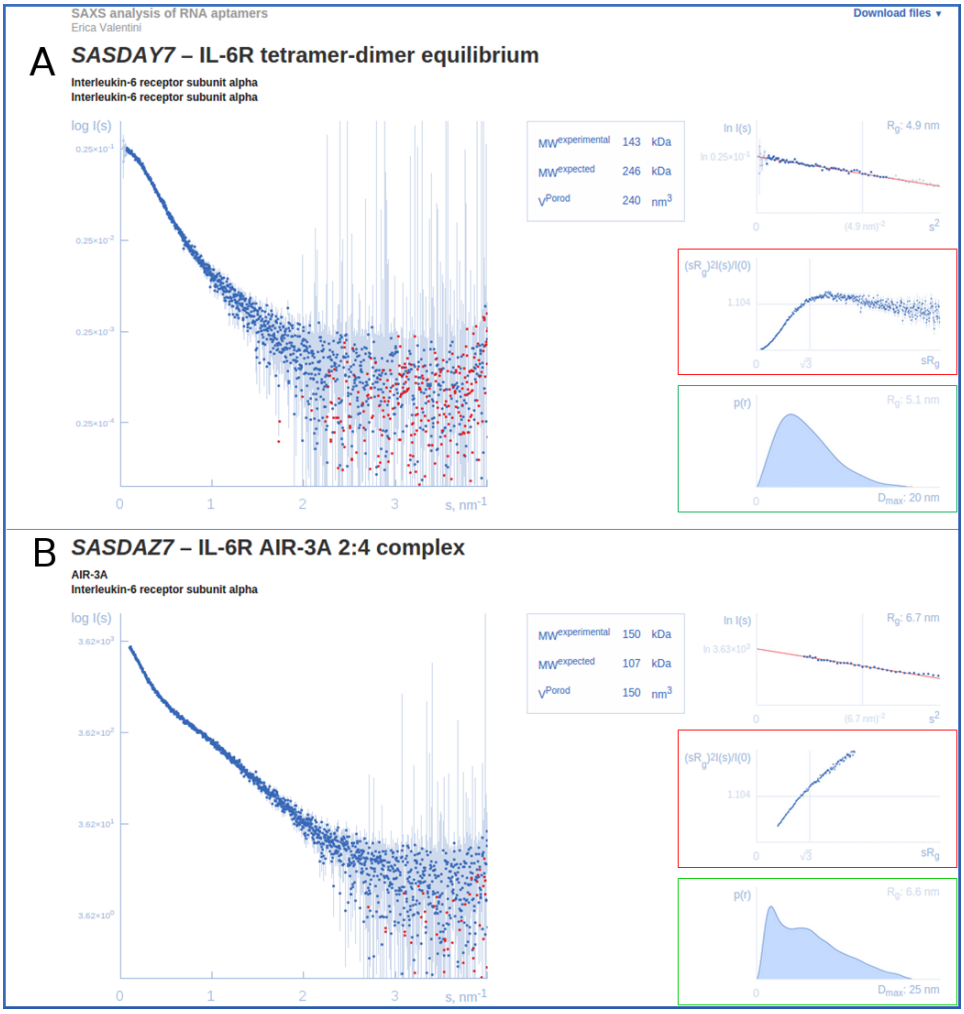


Figure 5.13: SASBDB entries of IL-6R alone and in complex with AIR-3A: **A** `http://www.sasbdb.org/data/SASDAY7/` and **B** `http://www.sasbdb.org/data/SASDAZ7/`. The green frame highlights the distance distribution functions while, the red frame highlights the normalized Kratky plot. In SASDAY7 the plots show that the protein has a rather globular and not flexible while, in SASDAZ7 the plots show a rather flexible protein-RNA complex very extended with a turn in the shape.

nificantly improved goodness-of-fit of 1.33 (Figure 5.14,**C**). The improved fitting suggests that the dimeric species in solution have an extended conformation. In addition, the volume fraction of the two forms is in agreement with the estimated $MW$ based on the $V_p$.

Finally, the scattering curve and the distance distribution function of

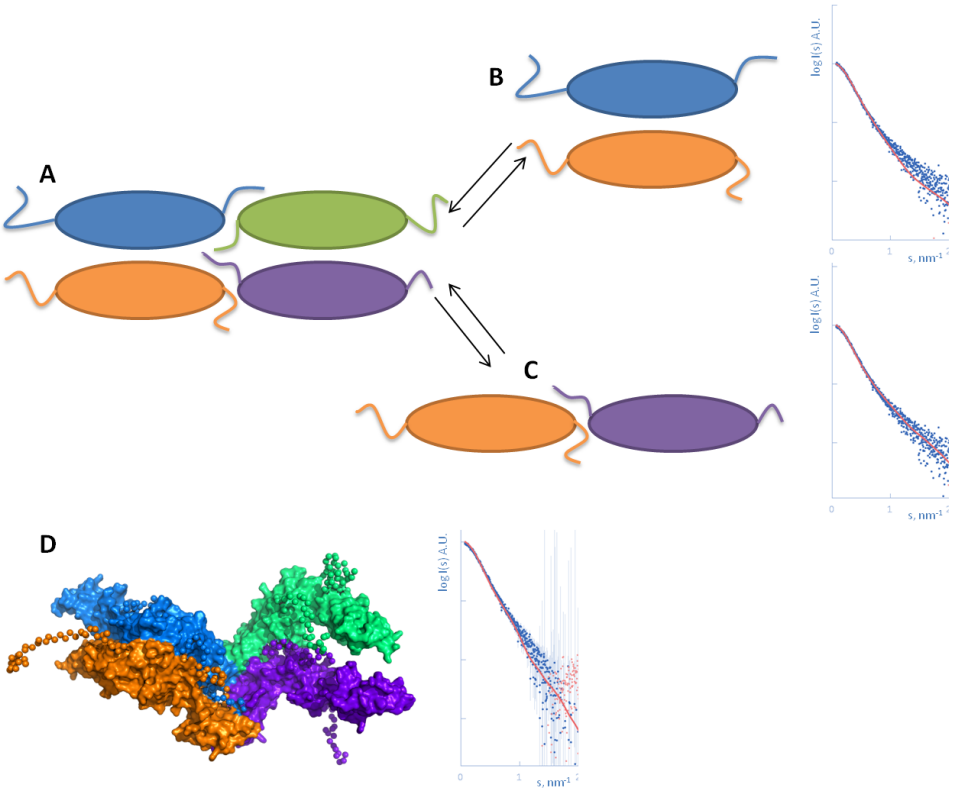


Figure 5.14: IL-6R tetramer and dimer. **A** Schematic representation of the IL-6R tetramer in equilibrium with **B** a compact dimer with corresponding goodness-of-fit ($\chi^2 = 1.92$) and with **C** an extended dimer with corresponding goodness-of-fit ($\chi^2 = 1.33$). The goodness-of-fit was estimated with OLIGOMER with the volume fractions for **A-B** equilibrium of 65% and 35% and of 47% and 53% for **A-C**. **D** CORAL model of the tetramer with the goodness-of-fit of $\chi = 2.49$.

the complex IL-6R-AIR-3A are typical of a very extended bent shape (as showed by the plots stored in SASBDB 5.13,**B**). The structural parameters in Table 5.3 suggest an increase in both the $R_g$ and $D_{max}$ as compared to the apo protein form. The estimation of the oligomerization state is complicated by the different scattering density of RNA compared to proteins and by the fact that the solution was prepared with an excess of RNA (0.2 mg/ml of protein and 0.4 mg/ml of RNA). Nevertheless, the values were in agreement with an oligomerization state ranging from 2:2 to 2:4 (protein:RNA). In order to model the protein:RNA complex the *ab initio* approach with DAMMIF was applied. The resulting model has an "L" shape

as shown in Figure 5.15,**B**. As for the apo protein, CORAL was applied because the missing N- and C- termini. Since the shape was estimated to be very elongated, the extended dimer was used as starting point for the rigid body modeling. Since the oligomerization state was difficult to estimate *a priori* different protein-RNA combinations were tried. The 2:4 (protein-RNA) oligomerization state settings resulted in a model comparable with the *ab initio* model (see superposition in Figure 5.15,**A**) and a better goodness-of-fit ($\chi^2 = 1.43$) (showed in Figure 5.15,**C**).

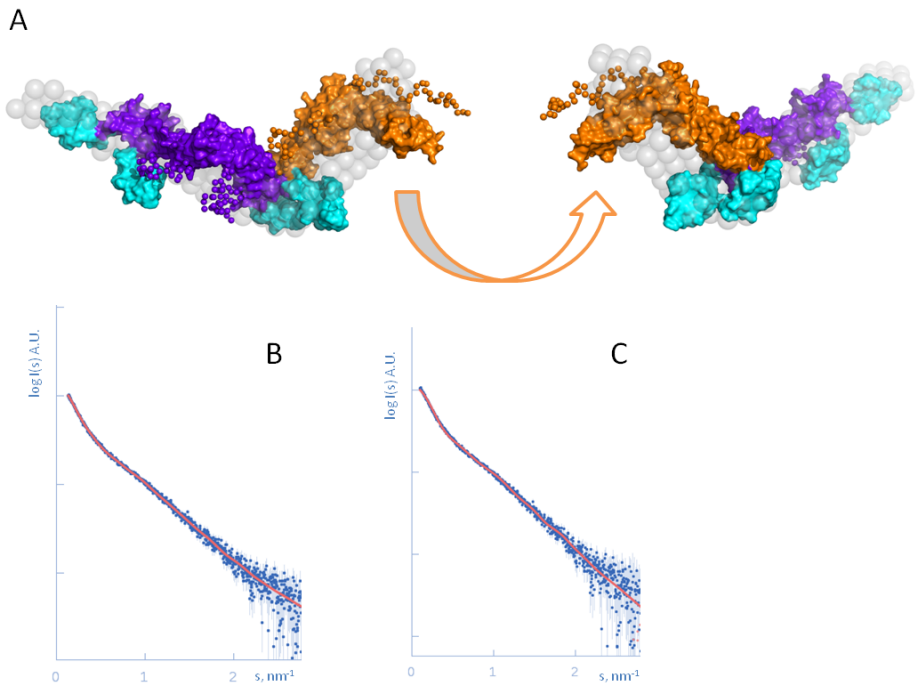


Figure 5.15: IL-6R in complex with AIR-3A. **A** CORAL model superimposed to DAMMIF model front and size view. **B** Goodness-of-fit of the DAMMIF model ($\chi^2 = 1.036$). **C** Goodness-of-fit of the coral model ($\chi^2 = 1.43$).

### 5.3.3 Conclusions

The SAXS analysis performed allowed the structural characterization of different aptamers. The results indicated that the long AIR-3 aptamers are dimeric structures with an extended bent shape. The small G-quadruplex aptamers are also dimeric constructs but have a rather globular shape. Finally the protein IL-6R was measured in its apo and holo form. The apo form displays an equilibrium of crystallographic tetramers with dimers, the

latter being in an extended conformation. The holo form reveals dimeric protein in complex with four AIR-3A molecules. Also in this case all the data and models derived from this set of SAXS experiments have been stored in SASBDB and still on hold pending the publication.

Aptamers allow one to target specific substrates to, for example, deliver drugs in a precise way. A better understanding of the aptamer 3D structure and complex formation could be important for the development of new treatments for diseases like diabetes, cancer or multiple sclerosis and other syndromes involving the IL-6 receptor.

# Chapter 6

# Conclusions

The three projects described above focus on three different scenarios: two homologous proteins with different degrees of flexibility [70], a protein-protein heterodimer that changes conformation upon binding DNA (Bruckmann et al., in preparation) and finally different RNA aptamer constructs in complex with the target protein (Berg et al., in preparation). In all these cases applications of SAXS was crucial to gain an insight into the 3D shapes and oligomeric compositions allowing one to make educated functional guesses.

The three projects are just a small selection from the manifold of projects published where SAS is applied. The number of SAXS and SANS-based publications is increasing every year also thanks to high brilliance synchrotron facilities and improvements in the software that allow very short measurement times and high quality results. The steadily growing amount of data published has to be make available for the increasing community of structural biologists applying SAS in their experiments.This was the main reason to develop SASBDB [15].

The small angle scattering data bank (SASBDB) follows the requirements of the wwPDB SAStf to create a repository where the SAS data and derived models are freely accessible for download, stored in a standard format (sasCIF), and are easy to browse and to locate. In addition, SASBDB presents interactive visualization of models, highly informative plots automatically generated, cross-links to other biological databases and a number of highly purified proteins that can be used as benchmark for different purposes. SASBDB reached in a short time a large number of publicly accessible entries –114 experimental data sets and 195 models– with an increasing number of users accessing the website to search and download the different types of data stored. SASBDB is currently the largest repository of SAS data and models available. Other repositories include the protein ensemble database (PED) [14] where SAXS is one of the techniques applied –other techniques include NMR and MD– to determine the ensembles of conforma-

tions describing IDPs, proteins with IDRs and denatured proteins.

The author contributed to different extents towards all of the described projects. The SASBDB was fully developed by the author starting from the design of the relational database structure to the development of the web application, while for PED the contribution consisted in designing the SAXS-related part of the database schema and suggesting the essential SAXS information that needed to be stored. Finally, for what concerns the users projects, the author participated at the different phases of the experiments spanning from the experiment design to data analysis and writing the publications (one of the project has been published while for the other two papers are in preparation).

# References


[1] C. E. Blanchet and D. I. Svergun, "Small-angle X-ray scattering on biological macromolecules and nanocomposites in solution.," *Annual review of physical chemistry*, vol. 64, pp. 37–54, Jan. 2013.

[2] A. Guinier, "La diffraction des rayons x aux très petits angles; application a l'étude de phénomènes ultramicroscopiques.," *Annales de Physique*, vol. 12, pp. 161–237, 1939.

[3] G. Porod, "General theory," in *Small-Angle X-Ray Scattering* (O. Glatter and O. Kratky, eds.), pp. 17–51, New York: Academic, 1982.

[4] O. Kratky and G. Porod, "Diffuse small-angle scattering of x-rays in colloid systems.," *Journal of Colloid Science.*, vol. 4, pp. 35–70, 1949.

[5] G. Porod, "Die ŕontgenkleinwinkelstreuung von dichtgepackten kolloiden systemen," *Kolloidnyi Zhurnal*, vol. 124, pp. 83–114, 1951.

[6] P. Bernadó and D. I. Svergun, "Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering.," *Molecular bioSystems*, Sept. 2011.

[7] M. V. Petoukhov, D. Franke, A. V. Shkumatov, G. Tria, A. G. Kikhney, M. Gajda, C. Gorba, H. D. T. Mertens, P. V. Konarev, and D. I. Svergun, "New developments in the ATSAS program package for small-angle scattering data analysis," *Journal of Applied Crystallography*, vol. 45, pp. 342–350, 2012.

[8] D. Franke, A. G. Kikhney, and D. I. Svergun, "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*, vol. 689, pp. 52–59, 2012.

[9] J. Trewhella, W. A. Hendrickson, G. J. Kleywegt, A. Sali, M. Sato, T. Schwede, D. I. Svergun, J. A. Tainer, J. Westbrook, and H. M. Berman, "Report of the wwPDB Small-Angle Scattering Task Force: Data Requirements for Biomolecular Modeling and the PDB," *Structure*, vol. 21, pp. 875–881, 2013.

[10] H. Berman, K. Henrick, H. Nakamura, and J. L. Markley, "The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data.," *Nucleic acids research*, vol. 35, pp. D301–3, 2007.

[11] A. Grishaev, J. Wu, J. Trewhella, and A. Bax, "Refinement of multidomain protein structures by combination of solution small-angle X-ray scattering and NMR data," *Journal of the American Chemical Society*, vol. 127, no. 47, pp. 16621–16628, 2005.

[12] A. V. Sokolova, V. V. Volkov, and D. I. Svergun, "Prototype of a database for rapid protein classification based on solution scattering data," *Journal of Applied Crystallography*, vol. 36, pp. 865–868, 2003.

[13] G. L. Hura, A. L. Menon, M. Hammel, R. P. Rambo, F. L. Poole, S. E. Tsutakawa, F. E. Jenney, S. Classen, K. A. Frankel, R. C. Hopkins, *et al.*, "Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS).," *Nature methods*, vol. 6, pp. 606–12, 2009.

[14] M. Varadi, S. Kosol, P. Lebrun, E. Valentini, M. Blackledge, A. K. Dunker, I. C. Felli, J. D. Forman-Kay, R. W. Kriwacki, R. Pierattelli, *et al.*, "pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins.," *Nucleic acids research*, vol. 42, pp. D326–35, 2014.

[15] E. Valentini, A. G. Kikhney, G. Previtali, C. M. Jeffries, and D. I. Svergun, "SASBDB, a repository for biological small-angle scattering data.," *Nucleic acids research*, vol. 43, pp. D357–63, Jan. 2015.

[16] I. D. Svergun, H. J. M. Koch, A. P. Timmins, and P. M. Roland, *Small Angle X-Ray and Neutron Scattering from Solutions of Biological Macromolecules*, vol. 19 of *IUCr Texts on Crystallography*. Oxford science publications, May 2013.

[17] M. A. Graewert and D. I. Svergun, "Impact and progress in small and wide angle X-ray scattering (SAXS and WAXS).," *Current opinion in structural biology*, vol. 23, pp. 748–54, 2013.

[18] A. Spilotros and I. D. Svergun, "Advances in small- and wide-angle x-ray scattering saxs and waxs of proteins," in *Encyclopedia of Analytical Chemistry*, Chichester, UK: John Wiley & Sons, Ltd, Sept. 2014.

[19] C. E. Blanchet, A. Spilotros, F. Schwemmer, M. A. Graewert, A. Kikhney, C. M. Jeffries, D. Franke, D. Mark, R. Zengerle, F. Cipriani, S. Fiedler, M. Roessle, and D. I. Svergun, "Versatile sample environments and automation for biological solution X-ray scattering experiments at the P12 beamline (PETRA III, DESY)," *Journal of Applied Crystallography*, vol. 48, pp. 431–443, Apr 2015.

[20] D. I. Svergun, M. V. Petoukhov, and M. H. Koch, "Determination of domain structure of proteins from X-ray solution scattering.," *Biophysical journal*, vol. 80, pp. 2946–53, 2001.

[21] H. D. T. Mertens and D. I. Svergun, "Structural characterization of proteins and complexes using small-angle X-ray solution scattering.," *Journal of structural biology*, vol. 172, pp. 128–41, 2010.

[22] O. Glatter and O. Kratky, *Small Angle X-ray Scattering*. London: Academic Press INC., 1982.

[23] P. Debye, H. R. Anderson, and H. Brumberger, "Scattering by an Inhomogeneous Solid. II. The Correlation Function and Its Application," *Journal of Applied Physics*, vol. 28, no. 6, p. 679, 1957.

[24] D. Durand, C. Vivès, D. Cannella, J. Pérez, E. Pebay-Peyroula, P. Vachette, and F. Fieschi, "NADPH oxidase activator p67(phox) behaves in solution as a multidomain protein with semi-flexible linkers.," *Journal of structural biology*, vol. 169, pp. 45–53, Jan. 2010.

[25] C. M. Jeffries and D. I. Svergun, "High-throughput studies of protein shapes and interactions by synchrotron small-angle x-ray scattering.," in *Structural Proteomics* (R. J. Owens, ed.), Methods in Molecular Biology, New York, NY: Springer New York, 2015.

[26] D. I. Svergun, "Determination of the regularization parameter in indirect-transform methods using perceptual criteria," *Journal of Applied Crystallography*, vol. 25, pp. 495–503, 1992.

[27] M. V. Petoukhov, P. V. Konarev, A. G. Kikhney, and D. I. Svergun, "ATSAS 2.1 – towards automated and web-supported small-angle scattering data analysis," *Journal of Applied Crystallography*, vol. 40, pp. s223–s228, 2007.

[28] H. B. Stuhrmann, "Ein neues Verfahren zur Bestimmung der Oberflaechenform und der inneren Struktur von geloesten globularen Proteinen aus Roentgenkleinwinkelmessungen," *Zeitschr. Physik. Chem. Neue Folge*, no. 72, pp. 177–198, 1970.

[29] D. I. Svergun, "Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing.," *Biophysical journal*, vol. 76, pp. 2879–86, 1999.

[30] D. Franke and D. I. Svergun, "DAMMIF , a program for rapid abinitio shape determination in small-angle scattering," *Journal of Applied Crystallography*, vol. 42, pp. 342–346, 2009.

[31] M. B. Kozin and D. I. Svergun, "Automated matching of high- and low-resolution structural models," *Journal of Applied Crystallography*, vol. 34, pp. 33–41, Feb. 2001.

[32] V. V. Volkov and D. I. Svergun, "Uniqueness of ab initio shape determination in small-angle scattering," *Journal of Applied Crystallography*, vol. 36, pp. 860–864, 2003.

[33] D. I. Svergun and K. H. Nierhaus, "A map of protein-rRNA distribution in the 70 S Escherichia coli ribosome.," *The Journal of biological chemistry*, vol. 275, pp. 14432–9, 2000.

[34] D. Svergun, C. Barberato, and M. H. J. Koch, "CRYSOL a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates," *Journal of Applied Crystallography*, vol. 28, pp. 768–773, Dec. 1995.

[35] I. D. Svergun, S. Richard, M. H. J. Koch, Z. Sayers, S. Kuprin, and G. Zaccai, "Protein hydration in solution : Experimental observation by x-ray," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. March, pp. 2267–2272, 1998.

[36] M. V. Petoukhov and D. I. Svergun, "Global rigid body modeling of macromolecular complexes against small-angle scattering data.," *Biophysical journal*, vol. 89, pp. 1237–50, 2005.

[37] P. V. Konarev, V. V. Volkov, A. V. Sokolova, M. H. J. Koch, and D. I. Svergun, "PRIMUS: a Windows PC-based system for small-angle scattering data analysis," *Journal of Applied Crystallography*, vol. 36, pp. 1277–1282, Oct. 2003.

[38] T. E. Williamson, B. a. Craig, E. Kondrashkina, C. Bailey-Kellogg, and A. M. Friedman, "Analysis of self-associating proteins by singular value decomposition of solution scattering data.," *Biophysical journal*, vol. 94, no. 12, pp. 4906–4923, 2008.

[39] M. V. Petoukhov, I. M. Billas, M. Takacs, M. A. Graewert, D. Moras, and D. I. Svergun, "Reconstruction of quaternary structure from x-ray scattering by equilibrium mixtures of biological macromolecules," *Biochemistry*, vol. 52, no. 39, pp. 6844–6855, 2013. PMID: 24000896.

[40] P. Bernadó, E. Mylonas, M. V. Petoukhov, M. Blackledge, and D. I. Svergun, "Structural characterization of flexible proteins using SAXS," *Journal of the American Chemical Society*, vol. 129, no. 17, pp. 5656–5664, 2007.

[41] E. F. Codd, "A relational model of data for large shared data banks," *Commun. ACM*, vol. 26, pp. 64–69, Jan. 1983.

[42] IBM, DB2, and Universal Database, *SQL Reference Version 7*. 1993.

[43] C. J. Oldfield and a. K. Dunker, "Intrinsically disordered proteins and intrinsically disordered protein regions.," *Annual review of biochemistry*, vol. 83, pp. 553–84, 2014.

[44] P. Tompa and M. Fuxreiter, "Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions.," *Trends in biochemical sciences*, vol. 33, pp. 2–8, Jan. 2008.

[45] P. E. Wright and H. J. Dyson, "Intrinsically unstructured proteins: reassessing the protein structure-function paradigm.," *Journal of molecular biology*, vol. 293, pp. 321–331, 1999.

[46] V. N. Uversky, C. J. Oldfield, and a. K. Dunker, "Intrinsically disordered proteins in human diseases: introducing the D2 concept.," *Annual review of biophysics*, vol. 37, pp. 215–46, Jan. 2008.

[47] R. Linding, R. B. Russel, V. Neduva, and T. J. Gibson, "GlobPlot: exploring protein sequences for globularity and disorder," *Nucleic Acids Research*, vol. 31, pp. 3701–3708, July 2003.

[48] Z. Dosztányi, V. Csizmok, P. Tompa, and I. Simon, "IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content.," *Bioinformatics (Oxford, England)*, vol. 21, pp. 3433–4, Aug. 2005.

[49] D. W. a. Buchan, F. Minneci, T. C. O. Nugent, K. Bryson, and D. T. Jones, "Scalable web services for the PSIPRED Protein Analysis Workbench.," *Nucleic acids research*, vol. 41, pp. W349–57, July 2013.

[50] S. Feuerstein, Z. Solyom, A. Aladag, A. Favier, M. Schwarten, S. Hoffmann, D. Willbold, and B. Brutscher, "Transient structure and SH3 interaction sites in an intrinsically disordered fragment of the hepatitis C virus protein NS5A.," *Journal of molecular biology*, vol. 420, pp. 310–23, July 2012.

[51] P. Bernadó, K. Modig, P. Grela, D. I. Svergun, M. Tchorzewski, M. Pons, and M. Akke, "Structure and Dynamics of Ribosomal Protein L12: An Ensemble Model Based on SAXS and NMR Relaxation.," *Biophysical journal*, vol. 98, pp. 2374–82, May 2010.

[52] V. Ozenne, F. Bauer, L. Salmon, J.-R. Huang, M. R. b. Jensen, S. Segard, P. Bernadó, C. Charavay, and M. Blackledge, "Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables.," *Bioinformatics (Oxford, England)*, vol. 28, pp. 1463–70, June 2012.

[53] P. Bernadó, E. Mylonas, M. V. Petoukhov, M. Blackledge, and D. I. Svergun, "Structural characterization of flexible proteins using SAXS," *Journal of the American Chemical Society*, vol. 129, no. 2, pp. 5656–5664, 2007.

[54] G. Tria, H. D. T. Mertens, M. Kachala, and D. I. Svergun, "Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering," *IUCrJ*, vol. 2, no. 2, pp. 207–217, 2015.

[55] M. Sickmeier, J. a. Hamilton, T. LeGall, V. Vacic, M. S. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V. N. Uversky, Z. Obradovic, and a. K. Dunker, "DisProt: the Database of Disordered Proteins.," *Nucleic acids research*, vol. 35, pp. D786–93, Jan. 2007.

[56] S. Fukuchi, T. Amemiya, S. Sakamoto, Y. Nobe, K. Hosoda, Y. Kado, S. D. Murakami, R. Koike, H. Hiroaki, and M. Ota, "IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners.," *Nucleic acids research*, vol. 42, pp. D320–5, Jan. 2014.

[57] J. R. Allison, P. Varnai, C. M. Dobson, and M. Vendruscolo, "Determination of the Free Energy Landscape of r -Synuclein Using Spin Label Nuclear Magnetic Resonance Measurements," *Journal of the American Chemical Society*, vol. 131, no. 51, pp. 18314–18326, 2009.

[58] T. U. Consortium, "Activities at the Universal Protein Resource (UniProt).," *Nucleic acids research*, vol. 42, pp. D191–8, 2014.

[59] F. Cunningham, M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, L. Gil, C. G. Giron, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek, N. Johnson, T. Juettemann, a. K. Kahari, S. Keenan, F. J. Martin, T. Maurel, W. McLaren, D. N. Murphy, R. Nag, B. Overduin, a. Parker, M. Patricio, E. Perry, M. Pignatelli, H. S. Riat, D. Sheppard, K. Taylor, a. Thormann, a. Vullo, S. P. Wilder, a. Zadissa, B. L. Aken, E. Birney, J. Harrow, R. Kinsella, M. Muffato, M. Ruffier, S. M. J. Searle, G. Spudich, S. J. Trevanion, a. Yates, D. R. Zerbino, and P. Flicek, "Ensembl 2015," *Nucleic Acids Research*, vol. 43, no. October 2014, pp. D662–D669, 2014.

[60] E. L. Ulrich, H. Akutsu, J. F. Doreleijers, Y. Harano, Y. E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C. F. Schulte, D. E. Tolmie, R. Kent Wenger, H. Yao, and J. L. Markley, "BioMagResBank," *Nucleic Acids Research*, vol. 36, no. November 2007, pp. 402–408, 2008.

[61] E. Mylonas, A. Hascher, P. Bernadó, M. Blackledge, E. Mandelkow, and D. I. Svergun, "Domain conformation of tau protein studied by solution small-angle X-ray scattering," *Biochemistry*, vol. 47, no. 39, pp. 10345–10353, 2008.

[62] D. a. Jacques, J. M. Guss, D. I. Svergun, and J. Trewhella, "Publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution.," *Acta crystallographica. Section D, Biological crystallography*, vol. 68, pp. 620–6, June 2012.

[63] F. S. Collins and L. A. Tabak, "Policy: NIH plans to enhance reproducibility.," *Nature*, vol. 505, pp. 612–3, 2014.

[64] M. Malfois and D. Svergun, "sasCIF: an extension of core Crystallographic Information File for SAS," *Journal of Applied Crystallography*, vol. 33, pp. 812–816, 2000.

[65] N. Deshpande, K. J. Addess, W. F. Bluhm, J. C. Merino-Ott, W. Townsend-Merino, Q. Zhang, C. Knezevich, L. Xie, L. Chen, Z. Feng, *et al.*, "The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema.," *Nucleic acids research*, vol. 33, pp. D233–7, 2005.

[66] D. a. Jacques and J. Trewhella, "Small-angle scattering for structural biology–expanding the frontier while avoiding the pitfalls.," *Protein science : a publication of the Protein Society*, vol. 19, pp. 642–57, Apr. 2010.

[67] D. Franke, C. M. Jeffries, and D. I. Svergun, "Correlation map, a goodness-of-fit test for one-dimensional x-ray scattering spectra," *Nature Methods*, vol. 12, no. 4, 2015.

[68] S. Velankar, Y. Alhroub, A. Alili, C. Best, H. C. Boutselakis, S. Caboche, M. J. Conroy, J. M. Dana, G. van Ginkel, A. Golovin, *et al.*, "PDBe: Protein Data Bank in Europe.," *Nucleic acids research*, vol. 39, pp. D402–10, 2011.

[69] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, *et al.*, "Biopython: freely available Python tools for computational molecular biology and bioinformatics.," *Bioinformatics (Oxford, England)*, vol. 25, pp. 1422–3, 2009.

[70] C. Tallant, E. Valentini, O. Fedorov, L. Overvoorde, F. M. Ferguson, P. Filippakopoulos, D. I. Svergun, S. Knapp, and A. Ciulli, "Molecular Basis of Histone Tail Recognition by Human TIP5 PHD Finger and Bromodomain of the Chromatin Remodeling Complex NoRC.," *Structure (London, England : 1993)*, pp. 1–13, Jan. 2015.

[71] A. Eberharter, I. Vetter, R. Ferreira, and P. B. Becker, "Acf1 improves the effectiveness of nucleosome mobilization by iswi through phd–histone contacts," *The EMBO Journal*, vol. 23, no. 20, pp. 4029–4039, 2004.

[72] Y. Zhou and I. Grummt, "The PHD finger/bromodomain of NoRC interacts with acetylated histone H4K16 and is sufficient for rDNA silencing.," *Current biology : CB*, vol. 15, pp. 1434–8, Aug. 2005.

[73] E. Longobardi, D. Penkov, D. Mateos, G. De Florian, M. Torres, and F. Blasi, "Biochemistry of the tale transcription factors PREP, MEIS, and PBX in vertebrates," *Developmental Dynamics*, vol. 243, no. July 2013, pp. 59–75, 2014.

[74] T. R. Bürglin, "Analysis of TALE superclass homeobox genes (MEIS, PBC, KNOX, Iroquois, TGIF) reveals a novel domain conserved between plants and animals," *Nucleic Acids Research*, vol. 25, no. 21, pp. 4173–4180, 1997.

[75] E. Ferretti, H. Marshall, H. Pöpperl, M. Maconochie, R. Krumlauf, and F. Blasi, "Segmental expression of Hoxb2 in r4 requires two separate sites that integrate cooperative interactions between Prep1, Pbx and Hox proteins.," *Development (Cambridge, England)*, vol. 127, no. 1, pp. 155–166, 2000.

[76] C. Meyer, U. Hahn, and A. Rentmeister, "Cell-specific aptamers as emerging therapeutics.," *Journal of nucleic acids*, vol. 2011, p. 904750, 2011.

[77] E. Magbanua, T. Zivkovic, B. Hansen, N. Beschorner, C. Meyer, I. Lorenzen, J. Grötzinger, J. Hauber, A. E. Torda, G. Mayer, S. Rose-John, and U. Hahn, "d(GGGT) 4 and r(GGGU) 4 are both HIV-1 inhibitors and interleukin-6 receptor aptamers.," *RNA biology*, vol. 10, pp. 216–27, Feb. 2013.

[78] S. Zhang, Y. Wu, and W. Zhang, "G-quadruplex structures and their interaction diversity with ligands," *ChemMedChem*, vol. 9, no. 5, pp. 899–911, 2014.

[79] C. Meyer, K. Eydeler, E. Magbanua, T. Zivkovic, N. Piganeau, I. Lorenzen, J. Grötzinger, G. Mayer, S. Rose-john, and U. Hahn, "Interleukin-6 receptor specific RNA aptamers for cargo delivery into target cells," *RNA Biology*, no. January, pp. 67–80, 2012.

# Appendix

No hazardous substances according to GHS were used in the study.

# Acknowledgments

First of all I would like to acknowledge Dmitri Svergun for his excellent work as supervisor and for following my inclinations. I also would like to thank, for the continuous help and support, the present and former members of the BioSAXS group: Clément Blanchet, Alejandro De Maria Antolinos, Daniel Franke, Gustavo Fuertes Vives, Matt Franklin, Nelly Hajizadeh, Andrew Huang, Mikhail Kachala, Chris Kerr, Al Kikhney, Tomáš Klumpler, Peter Konarev, Na Li, Haydyn Mertens, Sasha Panjkovich, Maxim Petoukhov, Zuzanna Pietras, Manfred Roessle, Darja Ruskule, Gundolf Schenk, Alessandro Spilotros, Giancarlo Tria and Anne Tuukkanen. A special thanks goes to Cy Jeffries and Melissa Gräwert for proofreading the thesis and giving me precious suggestions.

I also acknowledge both the Marie Curie Initial Training Network IDPbyNMR and EMBL for the financial support.

Finally, I would like to thank my university supervisor Ulrich Hahn and the other members of my Thesis Advisory Committee.

# Declaration upon oath

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

I hereby declare on oath, that I have written the present dissertation by my own and have not used other than the acknowledged resources and aids.

Date: _____

Signature: _____