Disentangling structural complexity in proteins by decomposing SAXS data with chemometric approaches

Fatima Herranz-Trillo

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Biochimie et biologie moléculaire

École doctorale Sciences Chimiques et Biologiques pour la Santé CBS2 N°168

Unité de recherche Centre de Biochimie Structurale de Montpellier
CNRS UMR5048 – UM – INSERM U1054

Disentangling structural complexity in proteins by decomposing SAXS data with chemometric approaches

Présentée par Fátima HERRANZ-TRILLO
Le 29 septembre 2017

Sous la direction de Pau BERNADÓ et Bente VESTERGAARD

Devant le jury composé de

Dr. Annette Eva LANGKILDE, University of Copenhagen	Rapporteur
Dr. Haydyn D.T. MERTENS, European Molecular Biology Laboratory (EMBL)	Rapporteur
Dr. Andrey KAJAVA, Centre de Recherche de Biochimie Macromoléculaire (CRBM)	Examinateur
Dr. Javier PÉREZ, Beamline SWING – Synchrotron SOLEIL	Examinateur
Dr. Jean-Michel ROGER, Research group Irstea - SupAgro	Examinateur
Dr. Romà TAULER, Institute of Environmental Assessment and Water Research (IDAEA)	Examinateur

UNIVERSITÉ DE MONTPELLIER

Abstract

Many biological systems are inherently polydisperse, presenting multiple coexisting species differing in size, shape or conformation (i.e. oligomeric mixtures, weakly bound complexes, and species appearing along amyloidogenic processes). The study of such complex systems is challenging due to the instability of the species involved, their low and interdependent relative concentrations, and the difficulties to isolate the pure components. In this thesis, I have developed methodological approaches to apply Small-Angle X-ray Scattering (SAXS), a low-resolution structural biology technique, to the study of polydisperse systems. As an additive technique, the SAXS pattern measured for a polydisperse sample corresponds to the concentration-weighted sum of the contributions from each of the individual components. However, decomposition of SAXS data into species-specific spectra and relative concentrations is laborious and burdened by ambiguity.

In this thesis, I present an approach to decompose SAXS datasets into the individual components. This approach adapts the chemometrics Multivariate Curve Resolution Alternating Least Squares (MCR-ALS) method to the specificities of SAXS data. Our method enables the rigorous and robust decomposition of SAXS data by simultaneously introducing different representations of these data and, consequently, emphasizing molecular changes at different time and structural resolution ranges. We have applied this approach, which we name COSMiCS (Complex Objective Structural analysis of Multi-Component Systems), to study two polydisperse systems: amyloid fibrillation by analysing time-dependent SAXS data, and conformational fluctuations through the analysis of data obtained using on-line size-exclusion chromatography coupled to SAXS (SEC-SAXS).

The importance of studying fibrillation processes lies in their implication in amyloidogenic pathologies such as Parkinson's or Alzheimer's diseases. There exist strong indications that soluble oligomeric species, and not mature fibrils, are the main cause of cytotoxicity and neuronal damage emphasizing the importance of characterizing early stages of fibrillation. The first application of our COSMiCS approach has allowed the study of the amyloidogenic mechanisms of insulin and the familial mutant E46K of α-synuclein, a Parkinson's disease related protein. The analysis enables the structural characterization of all the species present as well as their kinetic transformations.

The second part of the thesis is dedicated to the use of COSMiCS to analyze on-line SEC-SAXS experiments. Using synthetic data, I demonstrate the capacity of chemometric approaches to decompose complex chromatographic profiles. Using this approach, I have studied the conformational fluctuations in prolyl oligopeptidase (POP), a protein related to synaptic functions and neuronal development.

In summary, this thesis presents a novel chemometrics approach that can be generally applied to any macromolecular mixture with a tuneable equilibrium that is amenable to SAXS. Transient biomolecular complexes, folding processes, or ligand-dependent structural rearrangements can be probed structurally using COSMiCS.

Résumé

De nombreux systèmes biologiques sont intrinsèquement polydispersés, présentant de multiples espèces coexistantes, de taille, de forme ou de conformation différentes (c'est-à-dire, mélanges oligomèriques, des complexes faiblement liés se dissociant en composantes individuelles ou des espèces apparaissant lors de processus amyloïdogéniques). L'étude de tels systèmes complexes est une tâche difficile en raison de l'instabilité des espèces concernées, de leurs concentrations relatives faibles et interdépendantes et des difficultés rencontrées pour l'isolation des composantes pures. Dans cette thèse, j'ai développé des approches méthodologiques pour appliquer la diffusion des rayons X aux petits angles (SAXS), une technique de biologie structurale, à l'étude de systèmes polydispersés. SAXS est une technique additive et par conséquent, le diagramme de diffusion mesuré pour un échantillon polydispersé correspond à la somme pondérée en concentration des contributions de chacune des composantes individuelles du mélange. Cependant, la décomposition des données de SAXS en des spectres spécifiques des espèces et de leurs concentrations relatives est extrêmement laborieuse et ambigue.

Dans cette thèse, je présente d'abord une approche objective pour solidement décomposer les jeux de données de SAXS en composantes individuelles. Cette approche adapte la méthode chimiométrique « Multivariable Curve Resolution Alternate Least Squares » (MCR-ALS) aux spécificités des données de SAXS. Notre méthode permet une décomposition rigoureuse et robuste des données de SAXS en introduisant simultanément différentes représentations de ces données et par conséquent, en mettant l'accent sur des changements moléculaires à différentes plages de temps et de résolution structurale. Nous avons appliqué cette approche, que nous appelons COSMiCS (Analyse structurelle objective complexe des systèmes multi-composants) pour étudier deux systèmes polydispersés: la fibrillation des protéines, et les fluctuations conformationnelles de protéines grâce à l'analyse de données obtenues à l'aide d'une technique de couplage de chromatographie d'exclusion de taille (SEC) avec le ligne de SAXS (SEC-SAXS).

L'importance d'étudier les processus de fibrillation réside dans leur implication dans des pathologies amyloïdogéniques telles que les maladies de Parkinson ou d'Alzheimer. Il existe de fortes indications que les espèces oligomériques solubles, et non les fibrilles matures, sont la cause principale de la cytotoxicité et des dommages neuronaux. Cette observation souligne l'importance de caractériser les premiers stades des processus de fibrillation. Notre approche COSMiCS a permis d'étudier les processus amyloïdogéniques de l'insuline et du mutant familial E46K de l'α-synucléine, une protéine associée à la maladie de Parkinson. Cette analyse permet la caractérisation structurale des espèces présentes (y compris les espèces oligomériques) et la caractérisation cinétique de leurs transformations. La deuxième partie de la thèse est consacrée à l'utilisation de COSMiCS pour analyser des données de SEC-SAXS. Le SEC-SAXS est extrêmement populaire et a été implémenté sur plusieurs lignes de SAXS à travers le monde. En utilisant des données synthétiques,

je démontre la capacité des approches chimiométriques à décomposer des profils chromatographiques complexes. À l'aide de cette approche, j'ai décomposé l'ensemble des données SEC-SAXS mesurés pour la Prolyl OligoPeptidase (POP).

En résumé, cette thèse présente une nouvelle approche chimiométrique qui peut être généralement appliquée à tout mélange macromoléculaire pouvant subir une modifacation de son équilibre et pouvant être abordé par SAXS. Les complexes biomoleculaires transitoires, les processus de repliement, les réarrangements structuraux dépendants d'un ligand ou la formation de grands ensembles supramoleculaires peuvent être sondés de façon structurale en utilisant l'approche COSMiCS.

Contents

List of Figures

List of Tables

List of Abbreviations

αSN	α-synuclein
AF	Amyloid Fibrils
AFM	Atomic Force Microscopy
BSA	Bovine Serum Albumin
CD	Circular Dychroism
COSMiCS	Complex Objective Structural analysis of Multi-Component Systems
D_{max}	Maximum intra-particle distance
DLB	Dementia with Lewy Bodies
DTT	Dithiothreitol
EFA	Evolving Factor Analysis
EM	Electron Microscopy
EOM	Ensemble Optimization Method
FSC	Fourier Shell Correlation
FTIR	Fourier Transform infrared microscopy
GA	Genetic Algorithm
IDP	Intrinsically disordered Protein
IDR	Intrinsically disordered Region
kDa	kilo Dalton
LBs	Lewy Bodies
MALLS	Multi angle laser light scattering
MCR-ALS	Multivariate Curve Resolution using Alternating Least Squares
MD	Molecular Dynamics
MSA	Multiple System Atropy
MW	Molecular Weight
NAC	Non-Amyloid-β Component
NMR	Nuclear Magnetic Resonance
PCA	Principal Component Analysis
PD	Parkinson's disease
POP	Prolyl Oligopeptidase
p(r)	Pair-distance distribution
R_g	Radius of gyration
SA	Simulated annealing
SAXS	Small-Angle X-ray Scattering
SEC	Size-Exclusion Chromatography
SLS	Static Light Scattering

SR	Synchrotron radiation
SVD	Singular Value Decomposition
DTT	Tris(2-carboxyethyl)phosphine
TEM	Transmission Electron Microscopy
ThT	Thioflavin T
WAXS	Wide-angle X-ray scattering
WT	Wild Type

Part I

INTRODUCTION

Chapter 1

Small-Angle X-ray Scattering

Small-angle scattering (SAS) of X-rays (SAXS) or neutrons (SANS) is a biophysical method used in many areas of science and technology. In biology, is widely applied for the analysis of macromolecules in solution. SAXS is able to study the overall shape and structural transitions of biological macromolecules in solution. SAXS provides low resolution information on the shape, conformation and assembly state of proteins, nucleic acids and all kinds of macromolecular complexes. In this thesis I will talk about SAXS, which is the method used along this work, but the theory and analysis would be equivalent for SANS.

1.1 Small angle X-ray scattering for proteins

The study of the molecular mechanisms underlying the function of complex biological systems is often the focus of structural biology [1, 2]. The three dimensional structure of a biomolecule determines its functionality in vivo and knowing the 3D structure becomes important when studying the structural bases of biological mechanisms. Small angle X-ray scattering (SAXS) is a powerful method for analyzing the structure and the structural changes of biological macromolecules in solution.

The main advantage of SAXS is that, unlike NMR or X-ray crystallography, does not requires any special sample processing like crystallization, cryo-cooling or isotopic labeling. The sample is measured in solution, providing structural information in nearly native conditions. This characteristic allows its use not only for static structural modelling but also for the analysis of the response to changes in the experimental conditions (pH, temperature, pressure, ionic strength, binding...). It is also possible to follow the time course of processes such as folding/unfolding and assembly/dissociation over several orders of magnitude in time. The capacity of SAXS for studying a protein without need of crystallization allows characterization of proteins that are impossible to crystallize, like intrinsically disordered proteins (IDPs).

Another important advantage of the technique is that it can be applied to particles in a wide range of molecular sizes, from small proteins or peptides to large macromolecular machines [1, 2]. Biophysical parameters such as the radius of gyration (R_g), the maximum intra-particle distance (D_{max}) and the molecular weight (MW) can be estimated in an automated way while the data are collected, which makes SAXS also interesting from a practical point of view. The scattering data are also able to provide structural information that can

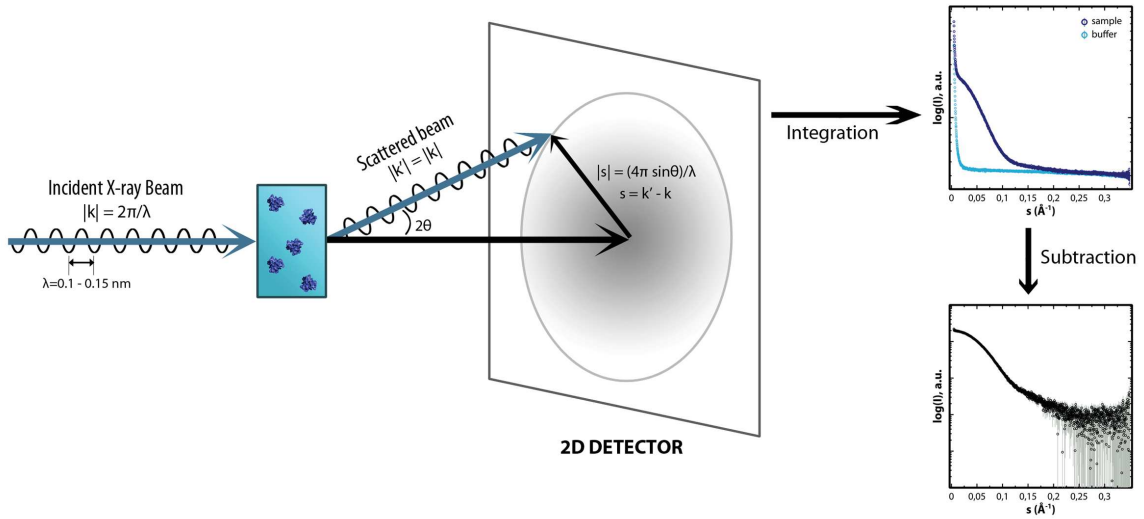


FIGURE 1.1. *Schematic representation of a SAXS experiment*

be exploited to derive low-resolution 3D structures. However, SAXS becomes more informative in combination with other structural, hydrodynamic, computational or biochemical methods. In the following sections I will describe the method and its applications.

1.2 General SAXS theory

In a typical SAXS experiment, a monochromatic (with a well-defined wavelength, λ) and collimated (parallel) X-ray beam is directed orthogonally onto a flow cell or static flat sample holder containing the biological sample in solution and a detector is placed on the opposite site of the sample in line with the beam (Figure 1.1). The 2D detector is placed at a longer distance between the sample and the detector compared with that used for crystallography in order to detect the scattering from the small angle range.

When the sample is illuminated by the monochromatic plane wave, with a modulus $|k| = 2\mu/\lambda$, the electrons within the object interact with the incident radiation becoming a source of spherical waves. For elastic scattering, where the energy and wavelength of the incident and scattered radiation are identical, the modulus of the scattered wave is $|k'| = |k|$. Along this thesis I will consider only the case of elastic scattering, which is the most relevant for structural studies and depends on the momentum transfer, $\mathbf{s = k' - k}$.

The scattering process involves a transformation from the 'real' space coordinates, **r**, where the structure of the scattering is defined, to the 'reciprocal' space of scattering vectors, s, in which the scattered radiation is measured. This process is described by a Fourier transformation, which involves a reciprocity between dimensions in real and reciprocal space implying that the smaller the 'real' size, the larger the corresponding 'reciprocal' size. In solution, the scattering is isotropic and the scattered intensity, $I_{total}(s)$, depends only on the momentum transfer $|s| = (4\mu \sin\theta)/\lambda$, where 2 is the scattering angle between the incident beam and the direction of observation.

The scattered waves from each electron have the same frequency and amplitude and can be summed. The intensity measured represents a summation of all the back scattered waves and is proportional to the square of the amplitude A(*s*), I(*s*) = | A(*s*) |^2 [3]. To describe the scattering of proteins in solution it is convenient to introduce the scattering length density distribution ((r)), that is equal to the total scattering length of the atoms per unit of solution volume [3]. The scattering amplitude is related to *ρ(r)* by a Fourier transform:

$$A(s) = \int_V \rho(r) e^{-isr} dr \qquad \text{(Eq. 1.1)}$$

where the integration is performed over the particle volume (V), **r** is the vector from an arbitrary origin to another point within the sample and s is the scattering length vector. The observed intensity is the product of the amplitude and its complex conjugate:

$$I(s) = |A(s)|^2 = A(s) \cdot A^*(s) = \int \int_V \rho(r)\rho(r^*) e^{-is(r-r^*)} dr dr^* \qquad \text{(Eq. 1.2)}$$

Proteins in the solution and the bulk solvent have different average electron densities ($\rho_s(r)$ 0.33 e^{-1}Å^{-3} for water and $\rho_s(r)$ 0.44 e^{-1}Å^{-3} for proteins). Therefore, the particles are embedded in a homogeneous matrix with a constant scattering density, ρ_s. As a consequence, Eq. 1.1 and Eq. 1.2 should be replaced by the difference between the electron density of the single particle and the solvent, $\Delta\rho(r) = \rho(r) - \rho_s(r)$. The autocorrelation function expresses the correlation between the densities measured at two random points separated by the distance r averaged over the illuminated volume V. The autocorrelation function of the particle $\gamma(r)$ [4] is defined by:

$$\gamma(s) \equiv \Delta\rho^{\,2}(r) = \int_V \Delta\rho(r')\Delta\rho(r'-r) dr' \qquad \text{(Eq. 1.3)}$$

Using Eq. 1.3, Eq. 1.2 can be written as

$$I(s) = |A(s)|^2 = \int_V \Delta\rho^{\sim 2}(r) e^{-isr} dr \qquad \text{(Eq. 1.4)}$$

We assume here that the solution is dilute enough so that inter-particle interference is negligible, and consequently, a spatial average can be made. This means that the intensity depends only on the magnitude and not on the azimuthal dependence of the s vector, **r** = | r | and **s** = | s |. In these conditions, Eq. 1.4 can be spatially averaged and results in:

$$I(s) = \langle |A(s)|^2 \rangle = \langle \int_V \Delta\rho^{\sim 2}(r) e^{-isr} dr \rangle \qquad \text{(Eq. 1.5)}$$

The scattering intensity recorded in the detector, as an isotropic image, can be radially averaged giving the 1D data by applying the Debye formula [5], $\langle e^{-isr} \rangle = \frac{sin(sr)}{sr}$, expressing Eq. 1.5 as:

$$I(s) = 4\pi \int_0^\infty \gamma(r) r^2 \frac{sin(sr)}{sr} dr \qquad \text{(Eq. 1.6)}$$

where r is the distance between two scattering elements within the sample.

In practice the generic scheme of a solution SAXS experiment series is starting with the measurement of the empty cell. Subsequently, collecting the scatter of pure water is useful for assessing the background level of the camera and for using it for absolute calibration and determination of the molecular weight of the solute. Alternatively, a standard protein can be measure at the beginning of the data collection for the molecular weight (MW); one broadly used standard is bovine serum albumin (BSA) or lysozyme. The use of a standard protein for MW determination will be explained in more detail in section 1.2.6. It is very important that the X-ray measurements on solutions of biological macromolecules, both on laboratory instruments and on synchrotron radiation (SR) sources, alternate experiments for the samples and matching solvents, for a correct subtraction of the background. These measurements of the buffer must be done for each sample and in the same cell and close in time to the sample measurements in order to keep identical conditions in both and get a correct background subtraction. The solvent curve is subtracted from the sample data to eliminate the solvent scattering and the instrumental background scattering and obtain the net scattering from the particles (see Figure 1.1) [2].

1.2.1 Structure and form factors

The net SAXS intensity after solvent subtraction may be expressed as a product of two terms, $I_{total}(s) = I(s) \cdot S(s)$. The form factor, I(*s*), arises from the scattering from individual particles in solution and contains the information about their structure. The structure factor, S(*s*), is due to interference of scattered waves emitted by different particles, and contains the information about interparticle interactions (about the structure of the solution), which can be either attractive or repulsive. The ideal sample is a monodisperse sample at low concentration to avoid interparticle interference effects and approaches the limit of infinite dilution. This conditions allow analyzing $I_{total}(s)$, assuming that S(*s*) = 1. SAXS is, however, also useful, and actively used, to study interactions between macromolecules in solution based on the analysis of the structure factor S(*s*) [6].

1.2.2 Goodness-of-fit for SAXS data

The statistical similarity between experimentally obtained intensities, $I_{exp}(s)$ and those computed from a model $I_{calc}(s)$ is evaluated using the reduced χ^2 statistics.

$$\chi^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left[\frac{I_{exp}(s_i) - I_{calc}(s_i)}{\sigma(s_i)} \right]^2 \qquad \text{(Eq. 1.7)}$$

where n is the number of experimental data points. The resulting χ^2 for a perfect model should be in the range $0.9 \leq \chi^2 \geq 1.1$. The experimental error, $\sigma(s_i)$ must be correctly estimated in order to have a statistically valid test. This estimation is used in most of the SAXS-based modeling applications and the method chosen along this thesis.

However, a new promising approach for evaluating differences between one-dimensional spectra, has been developed by Svergun and co-workers, called Correlation Map (CorMap)

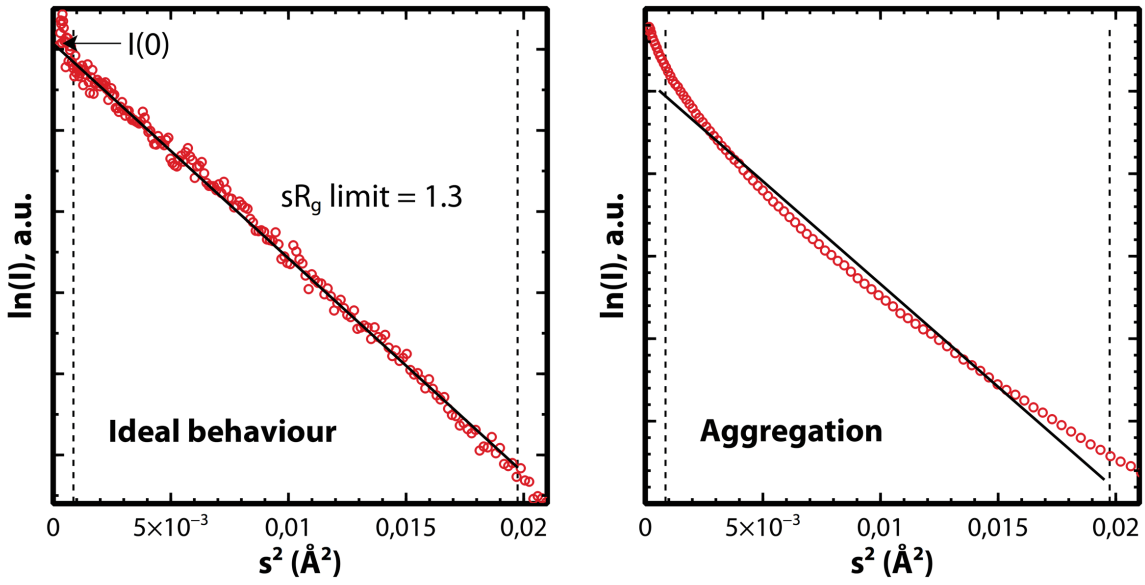


FIGURE 1.2. *Guinier plot for a protein showing good data (left) and aggregation (right)*

[7]. This approach, which uses data point correlation, maintains the power of the reduced χ^2 but has the advantage of being independent of error estimates.

1.2.3 Radius of gyration and forward scattering

The initial slope of the background-corrected scattering curve from a particle can be approximated by a Gaussian function [8] ("Guinier's law"):

$$I(s) \approx I(0)e^{\frac{-s^2R_g^2}{3}} \qquad \text{(Eq. 1.8)}$$

where R_g is the radius of gyration and I(*0*) is the forward scattering. The R_g can be derived by plotting the scattering data in a Guinier plot (log(I(*s*)) against s^2) (Figure 1.2). The R_g is then given as the slope of a straight line going through the data points at the low angles. Equation 8 holds in a range of about $s_{max} < 1.3$, commonly known as the 'Guinier zone' or 'Guinier range'. Depending on the shape of the particles the higher limit can sometimes be larger. Whenever the Guinier plot at low s is not linear, the sample has either aggregation or attractive (upswing) or intermolecular repulsions (downswing), at least in the case of homogeneous particles. Molecules composed of a mixture of different molecules with different scattering length densities also can display an anomalous Guinier behavior.

In the Guinier representation, the intercept of the straight line with the y-axis gives the forward scattering intensity (I(*0*)). I(*0*) is proportional to the number N of particles times the square of the product of the particle volume. Since N is inversely proportional to the molecular mass of the particles for a given particle weight concentration, I(*0*) is proportional to the molecular mass, which can be calculated (see more details in section 1.2.6).

1.2.4 Pair-distance distribution

The scattering intensity of non-interacting particles in dilute solution can be described by an integral that is limited to the maximal dimension (D_{max}) in the particles:

$$I(s)4\pi \int_0^{D_{max}} p(r)\frac{\sin(sr)}{sr}dr \qquad \text{(Eq. 1.9)}$$

where *r* is the distance between two point scatterers within the sample. *p(r)* is the so-called wise-distance distribution function of the particles. The *p(r)* can be obtained from Eq. 1.9 via Fourier transform.

$$p(r) = \frac{r^2}{2\mu^2}\int_0^{\infty} s^2 I(s)\frac{\sin sr}{sr}ds \qquad \text{(Eq. 1.10)}$$

Experimental I(*s*) covers a limited momentum transfer range, and direct Fourier transformation of the scattering curve from this finite number of points is not possible. A solution of this problem is the use of the indirect Fourier transformation [9] that represents *p(r)* as a linear combination of *K* orthogonal functions φ_k in the range [0, D_{max}]:

$$p(r) = \sum_{k=1}^{K} c_k \varphi_k(r) \qquad \text{(Eq. 1.11)}$$

The optimal coefficients c_k are calculated through minimization of

$$\Phi = \chi^2 + \alpha P(p) \qquad \text{(Eq. 1.12)}$$

Where χ^2 is the goodness of fit between the experimental data and that calculated by the direct transform of the *p(r)* function (Eq. 1.7), and the second term, P(p) ensures the smoothness of the *p(r)* function (Eq. 1.13)

$$P(p) = \int_0^{D_{max}} [p']^2 dr \qquad \text{(Eq. 1.13)}$$

The regularizing multiplier α balances between the fit to the data and the smoothness of the *p(r)*.

Since the distance distribution is a function in real space, it is often easier to recognize features of the particles in the *p(r)* function than in the scattering curve (Figure 1.3). Another important parameter that can be derived from the *p(r)* is D_{max}, the maximum intramolecular distance, however polydispersity, flexibility and aggregation may influence this parameter. This often results in a D_{max} estimate different than the actual dimension of the scattering particle [10, 11].

By definition, *p(r)* starts with a value of zero at *p(0)*, and it should terminate smoothly at a maximal dimension D_{max}. A deviation of *p(0)* from zero indicates an incorrect background subtraction, which can be used to estimate the background. A long tail or a shoulder at the high-r end of the *p(r)* should induce caution as it may be a sign of aggregation.

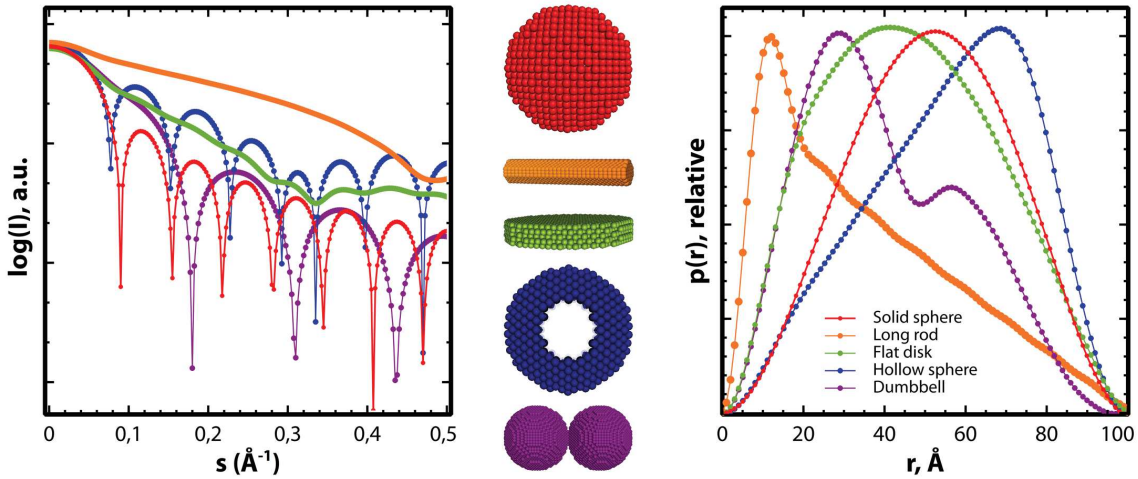


FIGURE 1.3. *Scattering intensities and pair-wise distance distribution functions, p(r), of geometrical bodies. Inspired from [12].*

R_g can also be derived from the *p(r)* through the equation:

$$R_g^2 = \frac{\int_0^{D_{max}} r^2 p(r) dr}{2 \int_0^{D_{max}} p(r) dr} \qquad \text{(Eq. 1.14)}$$

The R_g derived from the *p(r)* is based on the scattering measured across the entire *s* range and it is therefore always a good practice to compare R_g derived from a Guinier estimate with the R_g derived from the *p(r)*.

1.2.5 SAXS data representations and their usage

Besides the usual representation of the scattering curve in a semi-logarithmic scale, different representations of the SAXS data have been developed in order to extract additional information. Each representation enhances different features of the particle and it provides useful information 1.4.

1.2.5.1 Porod representation

The Porod-Debye law describes a fourth power law approximation to the relationship between s and the intensity, I(*s*) [13, 14]. This approximation holds within a limited range of scattering angles and suggests that the scattering of a folded particle decays proportionally to s^{-4},

$$I(s) \approx s^{-d_f} \qquad \text{(Eq. 1.15)}$$

where d_f describes a fractal degree of freedom, which is shape dependent. For spheres d_f = 4. The Porod plot $I(s) \cdot s^4$ plotted against *s* display a curve asymptotically approaching a constant value as s approaches infinity. Because Porod's law assumes uniform density and well-defined borders of contrast for the scattering objects and the solvent, the relationship

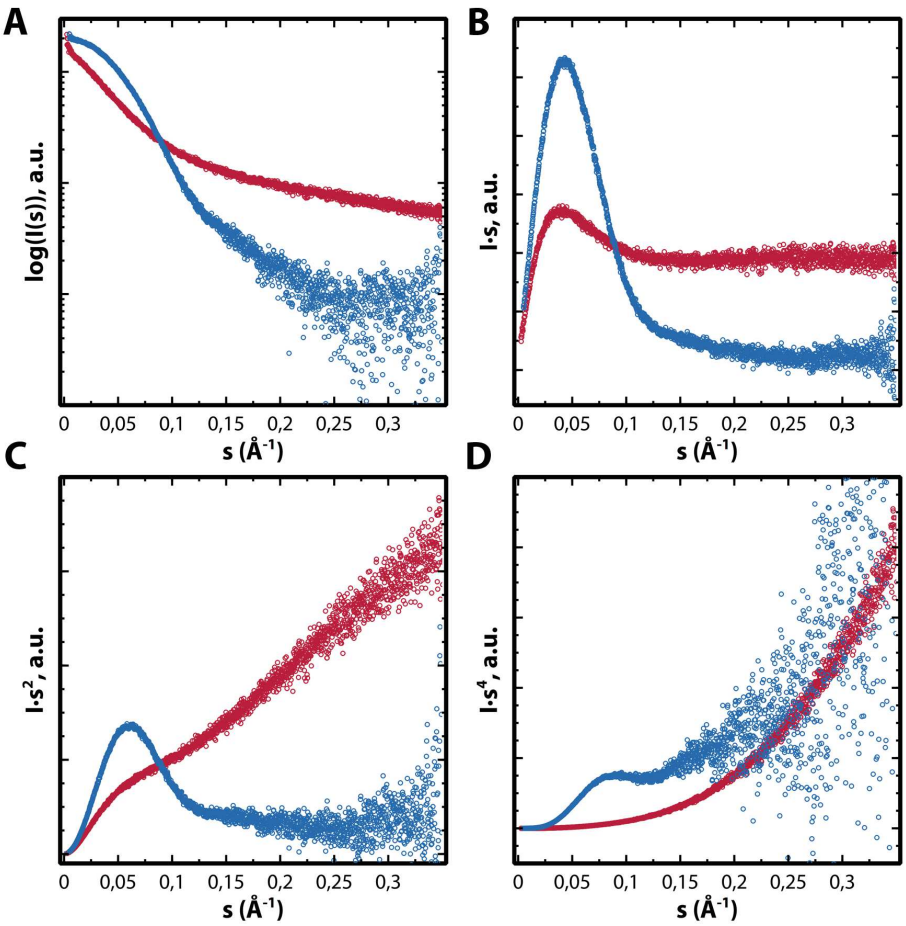


FIGURE 1.4. *Different representations for a folded protein (BSA), curve blue and from an intrinsically disorder protein (α-synuclein), red curve. A) Semi-logarithmic scale. B) Holtzer representation. C) Kratky plot. D) Porod plot.*

is not fulfilled at high angles where the scattering signal is dominated by higher resolution information.

An estimate of the object volume can also be determined from I(*0*) and Porod invariant Q, irrespective of the nature of the scattering sample:

$$Q = \int_0^\infty I(s)s^2 ds = 2\pi^2 \int_V \Delta\rho(r)dr \qquad \text{(Eq. 1.16)}$$

Q is directly related to the excluded particle volume, and, using that I(*0*) = $(\Delta\rho)^2 V_p{}^2$, one obtains

$$Q = 2\pi^2(\Delta\rho)^2 V_p \qquad \text{(Eq. 1.17)}$$

The volume of the particle, Porod volume (V_p) [15], is calculated as

$$V_p = 2\pi^2 \frac{I(0)}{Q} \qquad \text{(Eq. 1.18)}$$

The Porod representation has a practical use in the case of biological systems. In these

situations, it is often difficult to measure the exact contribution from the solvent. Even a small difference can lead to a different incoherent scattering level. In a plot of $I(s)\cdot s^4$ versus s^4 a residual background appears as a slope and $2\pi(\Delta\rho)^2S/V$ as the zero intercept. This allows estimate a flat background.

Additionally, Porod plot gives also information on the flexibility of the protein [16]. Porod-Debye law predicts a plateau within the low resolution region of the SAXS data when transformed by s^4. This plateau can be observed in globular proteins but is not present in unfolded proteins (1.4D).

1.2.5.2 Kratky representation

The Kratky plot [17], where $I(s)\cdot s^2$ is plotted against s, is able to qualitatively distinguish between globular particles and disordered states, and therefore reports on compactness. The Kratky plot is routinely used in SAXS data analysis and provides the first estimate of the folded state of the macromolecule. When plotted for globular proteins, where the intensity decays as s^{-4}, Kratky plot yields bell-shaped curve with a well-defined maximum at the smaller angles that fall in the region of higher angles, see Figure 1.4. Unfolded proteins show a much slower intensity decay: for example, an ideal random chain decays as s^{-2} [18]. The Kratky plot for unfolded proteins therefore presents a plateau over a specific range of s, which is followed by a monotonic increase, instead of the peak that presents folded proteins [19]. Partial unfolding or flexibility of the macromolecule lead to an increase of the scattering at higher angles and display an intermediate behavior between that of the folded protein and of the random chain. This graphical representation enhances the relevant features at higher angles and is a very good tool to qualitatively assess compactness, for example, in folding experiments.

1.2.5.3 Holtzer representation

In the Holtzer representation [20] the total scattered intensity is the integrated area of the SAXS data transformed as $I(s)\cdot s$ versus s. This approximation has been recently visited by Rambo and Tainer [21].

SAXS is capable of providing structural information on all particle types, including flexible systems. However, the analysis of the data using the Porod invariant presents limitations in the study of these flexible systems. As described above, the Porod invariant is an empirical SAXS value defined for compact folded particles. Q is unique to a scattering experiment and requires convergence of the SAS data at high s values in a Kratky plot. Convergence defines an enclosed area where the degree of convergence reflects the compacted (bounded area), flexible or unfolded (unbounded area) solution states. Consequently, Q (and therefore, V_p) is undefined for flexible particles. This observation leaves R_g as the only structural parameter that can be reliably derived from SAS data on flexible systems. Unlike the Kratky plot, the integral of $I(s)\cdot s$ versus s converges for both folded-compact and unfolded-flexible particles. Holtzer plot allows deriving a SAXS invariant,

the V_c. V_c is defined as the ratio of the particle's zero angle scattering intensity, I(*0*), to its total scattered intensity:

$$V_c = \frac{I(0)}{\int sI(s)ds} \qquad \text{(Eq. 1.19)}$$

V_c, like R_g, can be calculated from a single scattering curve and is concentration independent. This value can be used to derive the molecular weight of the particle (see next section).

1.2.6 Calibration to absolute scale and molecular weight

The aim of a scattering experiment is to obtain structural information about the sample. In the case of biological macromolecules this includes the knowledge of the molecular weight (MW) and, therefore, the determination of the oligomeric state. This can be performed in several ways.

1.2.6.1 Forward scattering of standard proteins

Standard proteins of known MWs (such as cytochrome C, lysozyme or bovine serum albumin) are often employed to determine the experimental MW of the given protein from the forward scattering, I(*0*). The standard protein is measured in similar conditions to the protein and the MW can be calculated using the ratio:

$$MW_p = MW_{st} \cdot \frac{I(0)_p/c_p}{I(0)_{st}/c_{st}} \qquad \text{(Eq. 1.20)}$$

where $I(0)_p$, $I(0)_{st}$ are the scattering intensities at zero angle of the studied and the standard protein, respectively, MW_p, MW_{st} are the corresponding molecular weight and c_p, c_{st} are the concentrations.

1.2.6.2 Forward scattering of water

Alternatively, scattering from water can be used to obtain the scattering from the solute in the absolute scale [22, 23] and then to calculate the MW. The water sample is measured in the same cell as the protein samples, and the scattering from the empty holder is subtracted.

1.2.6.3 Porod volume

It is possible to derive the MW from the Porod volume described above. The exact relationship between MW and V_p varies for different proteins depending on a combination of several factors, e.g. particle anisometry, flexibility etc. Using an empirical approach, Svergun's group [15] found that the scattering data range up to about $s_{max} = 8/R_g$ is optimal for a reliable computation of V_p. This upper limit in most cases approximately corresponds to the second minimum in the Porod plot. Using this interval, the average ratio between MW and V_p is 0.625, therefore, the volume in nm^3 is typically 1.6-2.0 times the MW in kDa.

1.2.6.4 *Ab initio* modeling

A similar approach to the Porod volume method is using the particle volume obtained from the volume of the low resolution structure of the molecule, which is 2 times the MW in kDa. The low resolution structure can be derived from several programs and it will be explained in more detail in next section.

1.2.6.5 Apparent volume

Fischer described a method to determine the molecular weight of proteins in dilute solution by using the experimental data of a single small-angle X-ray scattering (SAXS) curve measured on a relative scale [24]. This procedure does not require the measurement of SAXS intensity on an absolute scale and does not involve a comparison with another SAXS curve determined from a known standard protein. However, it is necessary an accurate determination of the protein concentration. The method is able to derive the MW from the value of the apparent volume V′ derived from truncated experimental SAXS data, using an empirical ratio. A program was developed to implement this technique, making it easily available to the scientific community. The web tool 'SAXS MoW' is available at http://www.if.sc.usp.br/ saxs/.

1.2.6.6 Volume of correlation

The above described methods to determine MWs require the knowledge of the protein concentration, the assumption of a compact near-spherical shape, or SAXS measurements on an absolute scale. Recently a new approach was developed to estimate MW of biomolecules without these restrictions. The new method was developed by Rambo & Tainer [21] and is based in the Holtzer representation. The approach determines that a parameter, Q_R, defined as the ratio of the square of V_c to R_g with units of $Å^{-3}$ is linear versus molecular mass in a log–log. The linear relationship is a power-law relationship given by

$$mass = (\frac{Q_R}{e^c})^{1/k} \qquad \text{(Eq. 1.21)}$$

that yields the empirical mass of the scattering biological particle allowing for the direct assessment of oligomeric state and sample quality. Parameters *k* and *c* were empirically determined and are specific for different classes of macromolecular particles. V_c and R_g are both contrast and concentration independent, thus the determination of molecular mass using Q_R can be made from SAXS data collected under diverse buffer conditions and concentrations. V_c can be calculated with the software ScÅtter, developed by Robert Rambo [21].

1.2.7 Radiation damage

One fraction of the X-rays that interact with the sample is absorbed by the particles damaging their structure. This radiation damage can be neglected in laboratory sources but has become a real program in SR sources due to their high brilliance [25]. The major

effect of radiation damage in SAXS is radiation-induced aggregation. The aggregates are readily seen in the low-s region of the scattering patterns and can produce erroneous data that are impossible to analyze. Therefore, it is important reduce the radiation damage.

One way to reduce radiation damage is to use a flow cell where the solute is constantly flowing through the irradiated volume. A disadvantage of this approach is the need for larger amounts of material.

Another option is the use of additives, such as DTT or TCEP, but is not always possible, as they can reduce disulfide bonds in proteins. A universal approach, employed at nearly all modern SR stations, is to slice the data collection into individual successive time frames (for example, ten or twenty). The recorded patterns are then processed separately and compared to each other (essentially to the first frame), and these frames showing systematic changes are not included in the subsequent averaging and further data processing. This analysis is normally done automatically using standard statistical criteria.

1.2.8 Resolution of SAXS data

Most of the intensity scattered by an object of linear size d is concentrated in the range of momentum transfer up to $s = 2\mu/d$. It is therefore assumed that if the scattering pattern is measured in reciprocal space up to s_{max} it provides information about the real space object with a resolution $\Delta = 2\pi/s$. For spherically averaged scattering patterns from solutions, I(s) usually decays rapidly as a function of momentum transfer, and only low-resolution patterns ($d \gg \lambda$) can be obtained. It is thus clear that solution scattering cannot provide information about the atomic position but only about the overall structure of macromolecules in solution.

The s range used for a SAXS experiment determines the molecular dimensions that can be observed. At a typical beamline the s range is 0.006 $Å^{-1}$ – 0.5 $Å^{-1}$ and the real dimensions accessible are then 160 Å – 13 Å. Although in theory would be possible to record scattering from macromolecular solutions up to 0.2 nm, the weak SAXS signal that results from scattering events at the high-s range is not much higher than the level of the noise associated with the technique.

The maximum observable dimension is set by the incoming beam, the beam stop size and the collimation of the beam to avoid divergence of the beam around the beamstop area and the area for small angle detection. When the data are recorded, the R_g estimate dictates the upper limit of the small angle resolution point.

A method very similar to SAXS, Wide-angle X-ray scattering (WAXS), allows access to higher resolution structural information [1, 26]. In this technique, the detector is moved closer to the sample to capture X-rays scattered to higher angles. It is even possible to simultaneously acquire SAXS and WAXS data by placing a detection window near the sample [27]. Although WAXS is easy to implement experimentally, the computational tools required to extract all the information are still under development, and usually it is necessary high resolution models are necessary for interpretation of the data. The strength of WAXS lies in its high sensitivity to small changes, and this can, therefore, be applied to identify structural similarities and characterize structural fluctuations. Currently, WAXS is

employed to study structural fluctuations and ligand binding in proteins [28]. WAXS has been also used to study nucleic acids using molecular dynamics (MD) simulations [29].

1.3 Structural interpretation of SAXS data

Besides providing information on the biophysical parameters such as R_g, D_{max} and MW, SAXS analysis of a protein system can provide information about 3D structure, the oligomeric distribution and the flexibility. SAXS data can also be combined with high-resolution structural information, to obtain structures of multi-subunit systems and is routinely used for the validation of structural models obtained by others methods, such as X-ray crystallography, Nuclear Magnetic Resonance (NMR) or Electronic Microscopy (EM). The measured intensity from an ideal monodisperse system is directly related to the single-particle scattering and provides low-resolution structural information of the molecule. However, some of the biological systems present polydispersity, where particles of different size, shape and/or conformation coexist. To characterize such systems, different data analysis methods are required. These methods will be described in the following sections.

1.3.1 Structural analysis of monodisperse biological systems

In monodisperse systems one considers that an individual species in a single conformation is present in solution. Although biomolecules experience dynamic phenomena at multiple levels, when these motions do not perturb the overall size and shape of the particle, they are not probed by SAXS and the system can be considered as monodisperse. Several approaches to structurally characterize monodisperse systems have been reported depending on the additional information integrated in the analysis.

1.3.1.1 *Ab initio* modeling

The aim of *ab initio* analysis of SAXS data is to recover the three-dimensional structure of the molecules in solution from the one-dimensional scattering pattern. The reconstruction of a 3D model of a molecule from its one-dimensional scattering pattern is a challenging task and different approaches have been developed.

The first *ab initio* shape determination method was proposed in 1970 by Stuhrmann [30]. The particle shape was represented by an angular envelope function describing the particle boundary in spherical coordinates. The method was implemented much later in the program SASHA [31], which was the first available shape determination program for SAXS. The spherical harmonics formalism proved to be very useful for analysis of SAXS data and it has been employed in many other approaches. The use of the angular envelope function is limited to relatively simple shapes without internal cavities. More detailed *ab initio* reconstructions became possible with the development of bead-modeling approaches [32]. A spherical volume with diameter D_{max}, which is obtained from the scattering pattern through *p(r)*, is filled with M densely packed beads (spheres of much smaller radius r0). Each of the beads may belong either to the particle (index = 1) or to the solvent (index

= 0), and the shape is thus described by a binary string X of length M. Starting from a random distribution of 1s and 0s, the model is randomly modified using a Monte Carlo-like search to find a string X fitting the experimental data. The original bead method, program DALAI_GA ([32], uses a genetic algorithm and does not impose explicit constrains. The program DAMMIN (Dummy Atom Model Minimisation) [33] is the most popular *ab initio* bead-modeling program. The program uses a random initial approximation by simulated annealing (SA) procedure [34]. The discrepancy (χ^2) is evaluated between the experimental and calculated scattering intensities. At each step in the SA procedure the assignment of a single bead is randomly changed leading to a new model X′. The solution is constrained to ensure compactness and connectivity of the resulting shape. A new optimized version of the program was implemented in the software DAMMIF [35], which perform this optimization in a faster manner.

One inherent problem of the shape determination methods is the uncertainty. In other words, different starting points yield different structural models with essentially the same fit to the data. To achieve a good solution is a recommendable approach to run shape determination programs several times to produce a diverse set of models corresponding to nearly identical scattering curves and inspect how different these models are as an indicator of the stability of the reconstruction. The uniqueness of the reconstruction is then assessed by *a posteriori* comparison and averaging of the different models. A new *a priori* ambiguity measure has been recently developed, based on the number of distinct shape categories compatible with a given dataset [36]. The models obtained in independent runs can be superimposed and averaged to obtain a most probable model, which is automated in the program package DAMAVER [37]. The ATSAS package employs the program SUPCOMP [38], which aligns two (low or high resolution), structural models represented by ensembles of points and yields a measure of their similarity. All pairs of independent models are aligned with SUPCOMP, and the model giving the smallest average discrepancy with the rest is taken as a reference model. Then, a density map of beads is computed and cut at a threshold corresponding to the excluded particle volume.

The reliability of *ab initio* models can be further improved if additional information about the particle is available. In particular, symmetry restrictions permit to significantly speed up the computations and reduce the discrepancy among models.

It is also important establish the resolution of *ab initio* shape modelling. A resolution based on analysis of the average Fourier shell correlation (FSC) functions within an ensemble of constructions has been developed recently by Svergun and co-workers [39]. This method has been implemented in the ATSAS suite in a program called SASRES [40] and it is able to determine resolution of *ab initio* models obtained using alternative procedures.

1.3.1.2 Computation of scattering patterns from atomic models

SAXS is often used for validation of 3D models obtained by high resolution methods, such as X-ray crystallography, NMR or homology models. Different methods have been developed to compute intensity profiles from a particular macromolecular structure. These methods are based on the Debye equation described above or on spherical harmonics.

Probably the most popular approach to compute SAXS intensity profiles from atomic coordinates is CRYSOL [41], and is the method used in this thesis. In CRYSOL, the scattering body is expanded in terms of an infinite series of spherical harmonics. One of the features of this method is the consideration of a homogeneous hydration shell surrounding the target.

Water modeling is critical to the correct interpretation of SAXS profiles and the more advanced approaches include the explicit modeling of the solvation shell. Programs such as the package PHAISTOS [42], AXES [43], SAXSTER [44], SASTBX [45] or AquaSAXS [46] implement different methods to compute the scatter profile and the solvent shell. The web server FoXs [47] is a tool for several SAXS-based modeling applications, including the computation of a SAXS profile of a given structure. The software also models the first solvation layer based on the atomic solvent accessible areas and provides an optimization of the hydration layer density as well as the excluded volume of the protein, to maximize the fit of the computed profile to the experimental profile. A new web server, called WAXSIS [48], was released recently. This approach computes SAXS (and WAXS) curves based on explicit-solvent all-atom molecular dynamics (MD) simulations. The MD simulations provide a realistic model for both the hydration layer and the excluded solvent. More detailed information about the different methods can be found in [49].

1.3.1.3 Rigid-body modeling

The strength of SAXS is further revealed in hybrid approaches in which this technique is used in combination with other structural information. Rigid body modeling approaches utilize atomic models of individual subunits or domains obtained by high resolution methods to analyze the structure of a complex or multidomain protein in solution. The scattering amplitude of the subunits can be precomputed with the methods described in the previous section and they are moved and rotated with respect to each other to find the configuration that fits better with SAXS data. A number of automated approaches have been developed using SAXS to determine the positions and orientations of subunits within macromolecular complexes.

One of the most used programs is SASREF [40, 50], which, starting from an arbitrary positioning of subunits, conducts random rigid-body movements and rotations, using SA to search for the best fit of the computed complex scattering to the experimental data. SASREF add penalties to avoid solutions without physical sense and allows the simultaneous fitting of multiple scattering curves. Moreover, it is possible improve its results by using information from other techniques, such as symmetry, orientational constrains, inter-residue contacts and inter-subunit distances. SASREF performs an automated global optimization of multi-subunit complexes, but the ATSAS suite includes other programs that complements SASREF, like BUNCH [40], used for multi-domain assembly; and the combination of both, CORAL [50].

There are also other available programs for rigid-body modelling, like the web server

FoXS [47, 51] that, besides being able to calculate the SAXS profile from a molecular structure as described in the previous section, can model the quaternary structures of multidomain proteins with defined rigid domains. The method uses the SAXS profiles, the component structures and information derived from other techniques, like stereochemical restraints. The scoring function is optimized by a biased Monte Carlo protocol and the final prediction corresponds to the best scoring solution in the largest cluster of many independently calculated solutions.

Another software for refining atomic models of multidomain proteins against SAXS is DADIMODO [52]. This program keeps rigid the domain structures and conformational changes are applied cyclically with a genetic algorithm that performs a search in the protein conformation space. The evaluation of the new generated conformation is done through the scoring function S and the goodness-of-fit to the SAXS data is computed by the program CRYSOL. The algorithm guarantees a physically acceptable atomic model of the structure and information from other techniques can be included in the optimization process, such as interdomain distances and orientational restraints.

1.3.2 Structural analysis of polydisperse biological systems

The ideal scenario for a SAXS experiment is a monodisperse sample, where all particles are identical. However, many biological samples display polydispersity. There are multiple sources of polydispersity, (i) where the particles in solution have the same chemical composition but differ in size and/or shape, or (ii) mixtures, where they may differ also in chemical composition.

For polydisperse systems or mixtures consisting of k different components, the measured scattering pattern can be written as a linear combination of the scattering intensity of the k types of particles ($I_k(s)$), where the coefficients $\nu_k > 0$ represent their molar fractions:

$$I(s) = \sum_{k=1}^{K} \nu_k I_k(s) \qquad \text{(Eq. 1.22)}$$

In this case, the overall parameters reflect the average values over the ensemble, but, of course, shapes of individual components cannot be reconstructed given only the experimental scattering from the mixture. To characterize such systems, additional information is required to partially overcome the limitations in the information content and the large number of degrees of freedom. These methods are the aim of this thesis and will be explained with more detail in the following sections.

1.3.2.1 Oligomer distribution

If prior structural knowledge of the components of a heterogeneous system is available, it is possible to derive their molar fractions in the sample. The program OLIGOMER from the ATSAS suite implements a non-negative linear least square algorithm to derive the volume fractions of the included components by minimizing the discrepancy, χ^2, between the experimental scattering profile of the sample and the components considered

[50, 53]. OLIGOMER has been successfully used to characterize oligomeric equilibria and complex formation [54–56].

1.3.2.2 Analysis of flexible systems

The high complexity of biological processes requires flexible systems, which allow to economize genome and protein resources by joining functional domains with flexible linkers, or by facilitating posttranslational modifications [57]. Intrinsically Disordered Proteins (IDPs) have emerged as fundamental molecules in crucial biological tasks. Due to their lack of a permanent secondary and tertiary structure, IDPs can adapt to the structural and chemical features of their partners and perform a broad range of crucial biological functions, complementary to those of ordered proteins [58, 59]. Such relevance makes important the development of methods to characterize their structure and to connect it with their function. Ensemble methods use the experimental data to derive accurate ensemble models to structurally characterize flexible proteins. These ensemble methods have been used with several techniques, including NMR [60] and SAXS. In SAXS, ensemble approaches have become popular in the last decade and several strategies have been developed: Ensemble Optimization Method (EOM) [50, 61], Minimal Ensemble Search (MES) [62]; Basis-Set Supported SAXS (BSS-SAXS) [63]; Maximum Occurrence (MAX-Occ) [64]; Ensemble Refinement of SAXS (EROS) [65]; Broad Ensemble Generator with Re-weighting (BERG) [66]; and Bayesian Ensemble SAXS (BE-SAXS) [67]. These methods share a common strategy but have distinct features. See [68] for details.

The program used in this thesis was the Ensemble Optimization Method (EOM). The program assumes the coexistence of a range of conformations whose average scattering intensity fits the experimental SAXS data. A subensemble of conformations is selected by a genetic algorithm (GA) from the scattering patterns computed from a large pool representing the maximum flexibility allowed by the protein topology. In this sense, the final result is considered a data-driven ensemble optimization strategy. In the EOM algorithm (Figure 1.5), a potential solution is represented by an ensemble containing N different conformers of the same molecule. The appropriate ensemble is selected from a pool containing $M \gg N$ conformers, which should cover the conformational space available to the molecule. Normally, 10,000 possible conformations of the biomolecule expected in the solution are generated. The GA is then used to select the subset of configurations that collectively fit the experimental data. We can assume that the subsets are uniformly populated, so the intensity of a subset I(s) containing N conformers is

$$I(s) = \frac{1}{N} \sum_{n=1}^{N} I_n(s) \qquad \text{(Eq. 1.23)}$$

where $I_n(s)$ is the scattering from the n-th conformer. To speed up the calculations, EOM uses the previously computed scattering curves from all structures in the pool, instead of using the structures directly (see section 1.3.1.2).

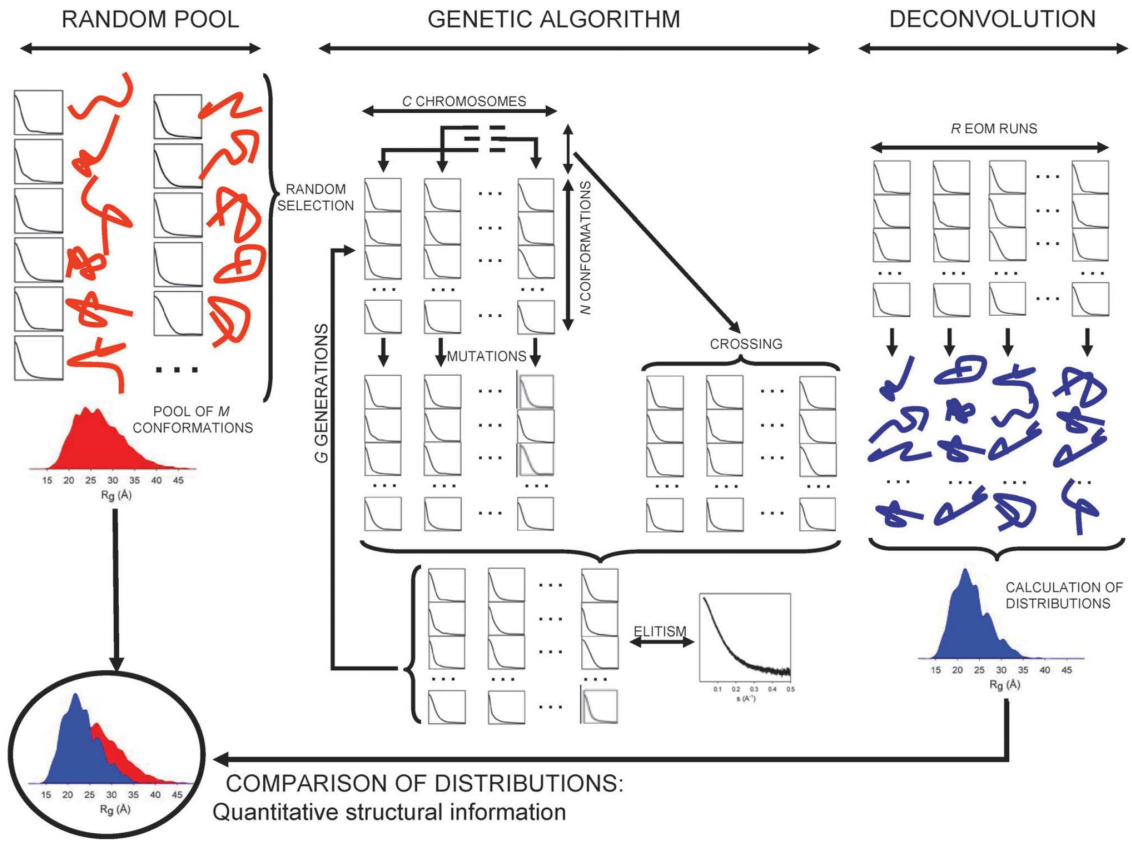


FIGURE 1.5. *Schematic representation of the EOM strategy for the analysis of SAXS data in terms of R_g distributions. The M conformations/curves of the pool (random distribution), left part of the figure, are used to generate the initial C chromosomes and to feed the genetic operators (mutations, crossing and elitism) along the GA process that runs for G generations. The complete process is repeated R independent times, and each run provides N selected structures/curves that fit the experimental profile. The structural analysis of the resulting conformations is displayed on the right part of the scheme, the R_g distribution of the selected (N x R) conformations is compared with that derived from the pool that is considered as a complete conformational freedom scenario. From this comparison it is possible to derive a quantitative structural estimation of the protein conformations coexisting in solution. Figure extracted from [58]*

Following typical GA nomenclature, each subset is called *chromosome* and it contains N scattering profiles, called genes, which correspond to N distinct conformers (typically between 20 and 50). In the first generation, $C = 50$ chromosomes are created by randomly selecting N conformations from the pool. In each generation G, these C chromosomes are submitted to two genetic operators: *mutation* and *crossing* (see details in [61]). In *mutation*, a random number of genes of each chromosome are exchanged for others, either from the pool or from the chromosomes belonging to the same generation. In *crossing*, genes of two randomly selected chromosomes are exchanged, thereby maintaining the size of the chromosome N. After the two genetic operations the populations is composed of $3C$ chromosomes. For each chromosome, the average of its individual SAXS profiles is compared with the experimental scattering to yield the fitness parameter χ^2 (see section 1.2.2). The C chromosomes yielding the lowest χ^2 discrepancy with the experimental profile are selected for further evolution in an elitism fashion. This mutation, crossing, and elitism

process is typically repeated $G = 500 – 1000$ generations. After completion of the optimization, the chromosome that best fits the experimental data is collected for further analysis. The genetic algorithm is often repeated 100 times. The structural examination of the final ensemble composed by the best chromosome of each independent EOM run is performed by analyzing the distribution of low-resolution parameters, R_g and D_{max} of the selected conformations in all GA runs. The selected pool of structures does not describe the actual combination of specific structures that is found in the solution, but rather a collection of structures representing the size and shape of the flexible protein in solution.

EOM provides a new source of structural information for disordered and also for flexible multidomain proteins. For unstructured proteins, the pools of possible chain conformations can be generated by the program Flexible-Meccano [69], which builds consecutively a polypeptide chain assuming that peptide planes are rigid entities connected through C_α atoms. For a system composed by multidomain proteins, it is possible to use a pool of structures generated by Monte-Carlo or molecular dynamic simulations (see an example in section 6.3).

1.3.2.3 Size-exclusion chromatography coupled to SAXS

Size-exclusion chromatography (SEC) is a technique for separating proteins and other biological macromolecules on the basis of molecular size (1.6). SEC has been used for many purposes, including buffer exchange (desalting), removal of non-protein contaminants, protein aggregation separation [70], the study of biological interactions, and protein folding [71]. It also has the important advantage of being compatible with physiological conditions. Protein molecules eluted from the SEC column are most often monitored by absorbance in the ultraviolet range, either at 280 nm or at 260 nm. Other detections, such as refractive index, radiochemical, electrochemical, and light scattering [70, 71], are also available.

SEC on-line with SAXS (SEC-SAXS) was originally implemented at the Advanced Photon Source (APS) BioCAT beamline [72] and is increasingly used in more beamlimes as a standard set-up [70, 73–75]. This combination is only possible at 3rd generation synchrotrons, where data over a large s-range (0.005-0.5 $Å^{-1}$) can be recorded in less than a second for concentrations below 1 mg/ml. The sample elutes from the SEC column and is routed directly and continuously into a BioSAXS flow cell for subsequent acquisition of SAXS data. SAXS sampling of the SEC elution peak can then be performed, therefore reducing the polydispersity of the initial sample.

The use of the SEC-SAXS set-up is preferable in samples with tendency to form aggregates, mixtures and low affinity complexes. Their use has also the advantage of reducing the inter-molecular (inter-particle interactions) at the same concentration and an accurate background subtraction. The disadvantages of the method are the accumulation of radiation damaged sample on the capillary (sample cell) (see section 1.2.7), thus, the need of exhaustive cleaning of the capillary after the data collection and the large sample consumption.

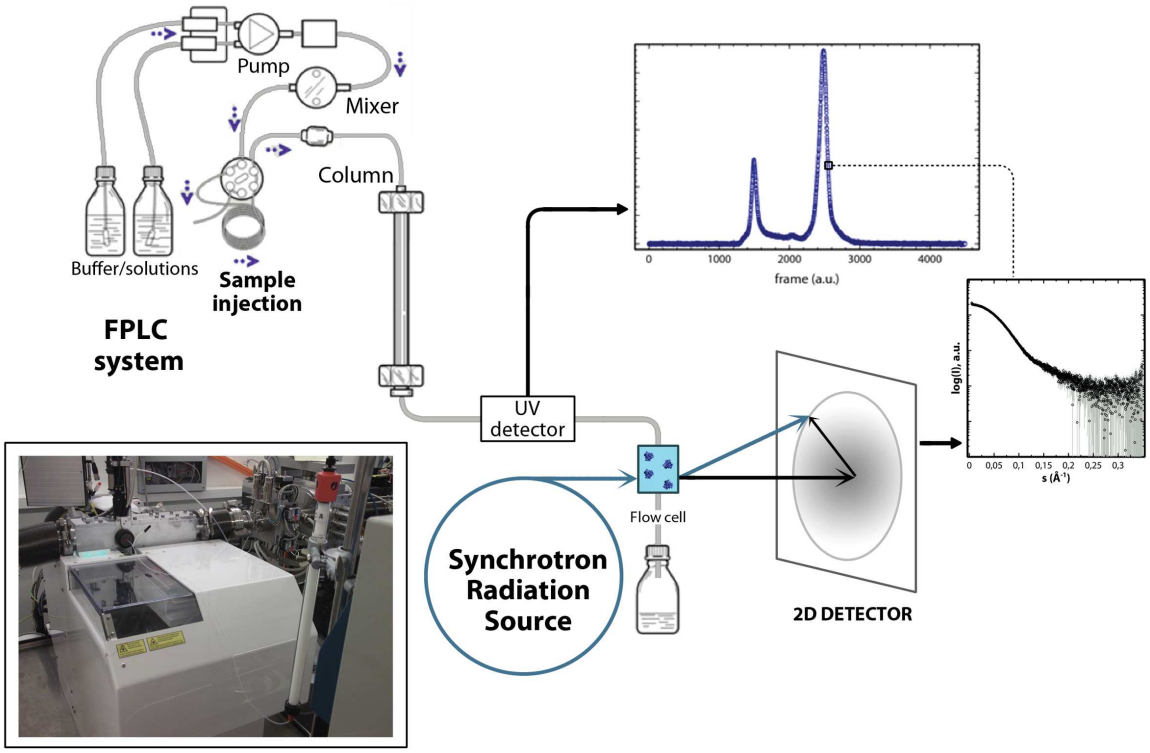


FIGURE 1.6. *Schematic drawing of SEC-SAXS measurement system. In the inset, photo of the SEC-SAXS set-up in P12 Beamline (Petra III, DESY).*

Estimation of the scaling constant for the buffer and/or extrapolation to zero concentration is nontrivial, since determination of the protein concentration requires alignment with a high-quality UV chromatogram. There exist software tools developed to analyze the data from SEC-SAXS experiments, like the program CHROMIXS, released in the version 2.8 of the ATSAS suite [53] or the program ScÅtter (http://www.bioisis.net/). Some beamlines provide their own software to process the data during of the acquisition.

Despite the decrease of polydispersity by using SEC-SAXS, some systems can still present regions with overlapped peaks that are composed by more than one species. For these cases, it is necessary to decompose the data in order to obtain scattering curves of the pure species before proceed to analyze them. One of the tools that has been developed for this task is US-SOMO [76][77] (http://somo.uthscsa.edu), which has a HPLC-SAXS module [78, 79] that is able to analyze such data and it is able to resolve several species co-eluting from a column. The software performs a single value decomposition (SVD) analysis of the dataset to choose the minimal number of components necessary to account for the data. It also includes the correction of baseline drift due to the accumulation of material on the SAXS capillary walls, and the symmetrical and non-symmetrical Gaussian decomposition of non-baseline-resolved HPLC-SAXS elution peaks. We have developed a different approach to analyze and decompose overlapped peaks from SEC-SAXS data using chemometric tools, which will be described in detail in chapter 6.

Chapter 2

Polydispersity in biological systems

Many biological systems are inherently polydisperse, that is, their samples present particles that differ in size, shape or conformation. There exist different kinds, (i) species polydispersity and (ii) conformational polydispersity. In species polydispersity, the sample consists in a mixture of chemically different particles. Typical systems of this type are oligomeric mixtures or weakly bound complexes dissociating into individual components. Conformational polydispersity appears when the system contains particles of identical molecular mass and sequence though these particles may adopt different conformations in solution. The number of conformations may be small such as coexistence of open and closed forms for an enzyme, or enormous such as for multidomains proteins with flexible linkers or intrinsically disordered proteins (IDPs). More complex systems including both kind of polydispersity can also exists [68].

Protein amyloid fibril formation is an example of a very complex biological system with high polydispersity associated with protein misfolding and formation of multiple oligomeric states. This process is going to be explained in detail in chapter 2.3 because its characterization is one of the main goals of this thesis.

2.1 Species polydispersity

In this scenario the solution contains small number of particles with distinctly different shapes and sizes; for example, a monomer-dimer equilibrium. Usually, SAXS data is used in combination with structural methods like Nuclear Magnetic Resonance (NMR) or X-ray to characterize the system in term of the fractions of the components in the mixture (see section 1.3.2).

An example of this kind of analysis is the analysis of the selecase [80] that combines high resolution models from X-ray crystallography with SAXS in order to determine the composition of system in solution. SAXS revealed that the protein does not aggregate at concentrations up to 65 mg/ml, and that the relative population of the oligomeric species in solution was concentration dependent (Figure 2.1). In addition, single value decomposition analysis of the SAXS dataset indicated that four species (monomers, dimers, tetramers, and octamers) were present. More detailed information can be found in the original publication (Paper III).

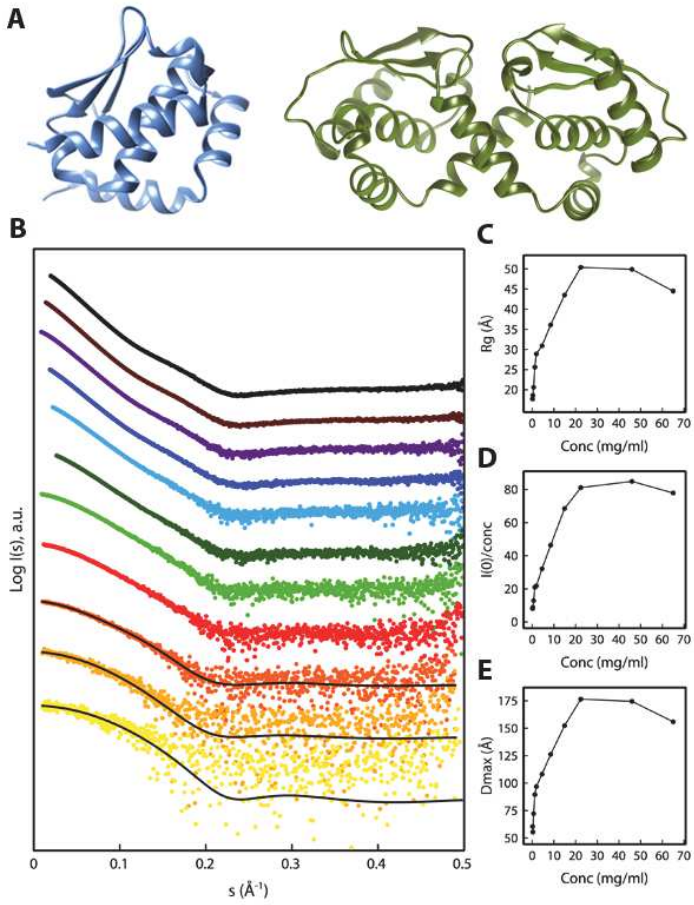


FIGURE 2.1. *A) Structure of the monomeric (blue, PDB entry 4qhf) and dimeric (green, PDB entry 4qhg) selecase. B) SAXS intensity profiles measured for wild-type selecase at 11 concentrations (bottom to top): 0.15 (yellow), 0.30, 0.63, 1.2, 1.9, 4.5, 8.5, 15, 22, 46, and 65mg/ml (black). Profiles have been displaced along the I(s) axis for comparison. The experimental scattering curves at the three lowest concentrations studied indicate a mixed population of monomers and dimers based on the crystallographic structures of slc1 and slc2 (black curves). Variation of the primary SAXS data parameters with concentration: (C) R_g, (D) I(0)/concentration, and (E) D_{max}.*

2.2 Conformational poldydispersity

2.2.1 Large amplitude conformational fluctuations in globular proteins

Proteins under native conditions exhibit a wide range of motions enabling the performance of their biological function. Functional modifications, control of synthesis and degradation are mechanisms requiring the formation of dynamic interfaces and conformational states that regulate and control the function. That control can be either direct, through activation or inactivation of the macromolecule, or indirect through pathways that affect the macromolecule. Each of these levels of control can be manifested through changes in the shapes of the specific macromolecule (Figure 2.2) [81]. Fast protein motions (pico-nanosecond timescale) are local and involve conformational fluctuations of a few angstroms, including side chain rotation and small backbone movements. By contrast,

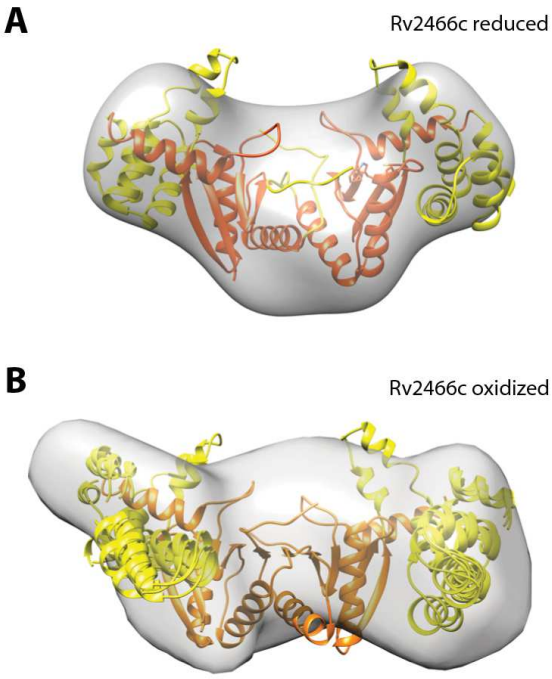


FIGURE 2.2. *Example of the conformational change of oxidized and reduced Rv2466c. Low resolution models, characterized by SAXS, of Rv2466cRED (A) and Rv2466cOX (B) in solution with the high resolution crystal structure of Rv2466cRED (Protein Data Bank code 4NXI) fitted by rigid body docking. Figure extracted from reference [81]*

.

at slow timescales (micro-milliseconds and beyond), large collective motions of protein domains and entire subunits in oligomeric assemblies allow changes in conformation of tenths of angstroms that are typically associated to active site opening/closing in ligand binding and vast structural rearrangements in allosteric events. It is the balanced interplay of this hierarchy of motions that allows proteins to adopt conformations complementary to their binding partners. The larger the amplitude of the motion and the shape change, more sensitive SAXS will be to monitor the perturbation.

2.2.2 Intrinsically disordered proteins

An extreme case of polydispersity can be found in Intrinsically Disordered Proteins (IDPs), which shows an enormous number of conformations. In the last two decades, IDPs have emerged as fundamental molecules for a large variety of biological process [82, 83]. IDPs are proteins that lack for a permanent secondary or tertiary structure and, consequently they are highly plastic and have the capacity of perform specialized functions. It has been predicted that more than 35% of human proteins have significant regions of disorder (intrinsically disordered regions, IDRs) and about 25% are likely to be completely disordered [84]. These IDRs are very well suited for protein-protein interactions. Under physiological conditions these proteins constantly fluctuate between different structural states, resulting in a dynamic mixture of conformations in a polydisperse solution.

The relatively recent discovery that proteins without globular architecture could also have biological activity has revolutionized this classical interpretation of the structure/ function paradigm, one of the bases of biology [85] and has raised intense research efforts that seek to unravel the structural bases of their function. Importantly, functions performed by well-folded globular proteins and IDPs are different and complementary.

There are multiple studies addressing the distinct function that natively unstructured protein can perform facilitated by its inherent plasticity [86–88]. Two are the main biophysical features explaining the biological roles of IDPs: (i) high accessibility, and (ii) conformational adaptability. Moreover, multiple advantages can be envisioned in the context of IDRs connecting two or more folded domains forming the multi-domain or modular proteins. This architecture enables the simultaneous interaction to multiple sites; for example, localizing consecutive catalytic steps in close proximity.

A large amount of disordered regions are functional when they interact with their biological partners to modify their activity. These partners can be proteins, nucleic acids, lipids or small molecules. The disordered nature of IDPs enables them to finely adapt their conformation to the structural and chemical nature of the partner to provide highly specific complexes. In some cases, however, the same disordered region can interact with multiple different partners even adopting different conformations.

Multiple IDPs have been associated with human diseases [89]. The first examples reported of disorder-induced diseases are related to neurodegenerative pathologies such as Alzheimer's, Parkinson's or Huntington's diseases. The hallmark of these pathologies is the presence of large fibrillar aggregates in patients' brains that are formed by the accumulation of non-functional proteins that fully or partially disordered such as amiloyd-β and Tau in Alzheimer's, α-synuclein in Parkinson's (see next chapter), and huntingtin in Huntington's disease. IDPs are also involved in cancer [90] and cardiovascular diseases [89].

2.3 Amyloids

Many human diseases are associated with the formation of extracellular amyloid fibrils (AF) or intracellular inclusions with amyloid-like characteristics. These diseases include the most common neurodegenerative pathologies, such as Alzheimer's, Parkinson's, and Huntington's diseases [91, 92], but also other non-neuropathic localized amyloidosis, like Type II diabetes. These diseases are among the most prevalent, debilitating, and economically and socially impacting disorders in the first-world. Consequently, a big effort has been made from a broad range of disciplines toward understanding the details of the mechanism of aggregation to find pharmaceutical treatments. Amyloids have been one of the focuses of this thesis. In the following sections I will describe the amyloid formation process and the species involved, especially for the proteins used during my work, insulin and α-synuclein (αSN).

2.3.1 Amyloid formation

The formation of amyloids is linked to the failure of a specific peptide or protein to adopt, or remain in, its native functional conformational state. This process is generically known as misfolding [93], and the originated pathologies are referred to as protein misfolding diseases. The largest group of misfolding diseases, however, is associated with the conversion of specific peptides or proteins from their soluble functional states into highly organized fibrillar aggregates. These structures are generally described as amyloid fibrils (AF) or plaques, when they accumulate extracellularly.

There are now approximately 50 disorders, with different symptoms, which are associated with the misfolding of functional peptides and proteins, and their subsequent conversion into aggregates. Interestingly, the ability of polypeptide chains to form amyloid structures is not restricted to that relatively small number of proteins associated with recognized clinical disorders, but it seems to be a generic feature of polypeptide chains [94–97]. A large and increasing number of peptides and proteins have been shown to be able to self-assemble in vitro into insoluble fibrillar aggregates, adopting the distinctive cross-β configuration [91, 93, 98, 99]. Even though the ability to form amyloids fibrils seem to be generic, the propensity to do so under given circumstances can vary markedly between different sequences. The relative aggregation rates for a wide range of peptides and proteins correlates with the physicochemical features of the molecules such as charge, secondary-structure propensities and hydrophobicity [100].

The term AF describes structurally comparable filamentous protein aggregates of only few nanometers of diameter but several micrometers in length. Studied at higher resolution, amyloids consist in multiple protofilaments that are twisted around each other to form mature fibrils. The core of each protofilament has a common 'cross-β' structure (Figure 2.3), which gives rise to repeating distances of 4.8 Å between peptide chains along the fibril axis [101, 102]. Whereas the 4.8 Å distance due to its origin is rather fixed due to the hydrogen bonds constrains [103, 104], the spacing between the β-sheets in the direction perpendicular to the fibril axis (lateral packaging) is highly dependent on the amino acid sequences of the component proteins, and it has hence been reported to span a range from 5-12 Å [98, 103, 105–107]. A more detailed description of the structure of amyloid fibrils and the intermediates of the process can be found in further sections and it will be focused in the proteins used during this work, αSN and insulin.

The transition of a protein from its functional soluble state to the amyloid state is a highly complex process that depends on both the intrinsic characteristics of the protein and the environmental conditions. However, there are similarities in the aggregation behavior of different peptides and proteins [108, 109]. This mechanism is not fully understood although the generic main steps are known (Figure 2.4).

Since the protein is synthetized by the ribosome it can evolve towards different states following different pathways. In the case of globular proteins, the functional state is normally achieved after a complex process of folding that can be assisted by chaperones. However, along this process can exist partially folded states (for example at low pH or as consequence of dynamical fluctuations) that expose the hydrophobic side chains, and the protein

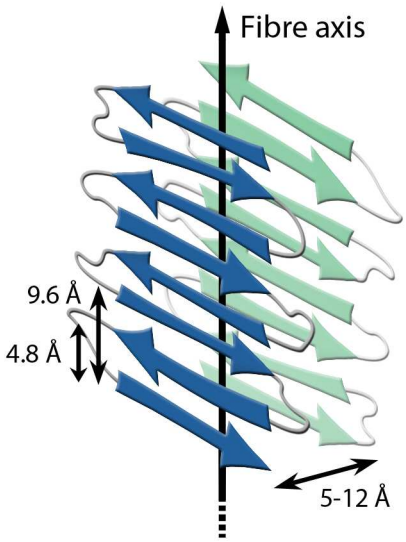


FIGURE 2.3. *Simplified model of the characteristic cross-β spacings from amyloid fibrils. The X-ray fibre diffraction patterns (not shown) reveal a strong 4.8 Å reflection on the meridian that corresponds to the hydrogen bonding distance between β-strands (strong interactions); and a more diffuse reflection on the equator that shows the inter-sheet distance of about 5 - 12 Å, depending on the protein (weaker interactions). A spacing of 9.6 Å would correspond to the repeat distance for an anti-parallel arrangement of β-strands.*

will be particularly vulnerable to misfolding and aggregation. Some of the peptides and proteins that are involved in the most common misfolding diseases are intrinsically disordered in their free soluble forms, such as the Tau protein in Alzheimer's disease [110] or αSN in Parkinson's disease [111, 112].

The overall process follows a nucleation–polymerization model [113] where, in the initial stages of the aggregation process, the soluble species undergo a nucleation step that results in the formation of a heterogeneous array of oligomeric species. These oligomeric species are able to grow through further monomer addition generating early prefibrillar aggregates that transform into species with more distinctive morphologies, often called 'protofilaments' or 'protofibrils'. These structures are commonly short, thin, sometimes curly, fibrillar species that are thought to assemble into mature fibrils, perhaps by lateral association accompanied by some degree of structural reorganization. The growth profile shows a typical sigmoidal reaction time course, which is a feature of nucleated polymerization [114, 115]. The process is usually experimentally monitored by the fluorescent dye Thioflavin T (ThT), that binds to the crossed β-sheet of amyloid fibrils with high selectivity [103, 116, 117]. In this profile we can observe a lag phase, where new oligomers are formed directly from monomers; and a rapid growth phase that reflect the addition of monomers onto existing aggregates. In cases in which the total quantity of protein is limited, the growth phase is followed by a plateau phase in which the reaction rate declines as a result of the depletion of the soluble species that is being monitored as it converts into fibrils. More recently, it has become evident that processes other than primary nucleation and elongation are important, including fibril fragmentation and surface-catalyzed nucleation. These secondary processes can dominate the kinetics of fibril growth under many

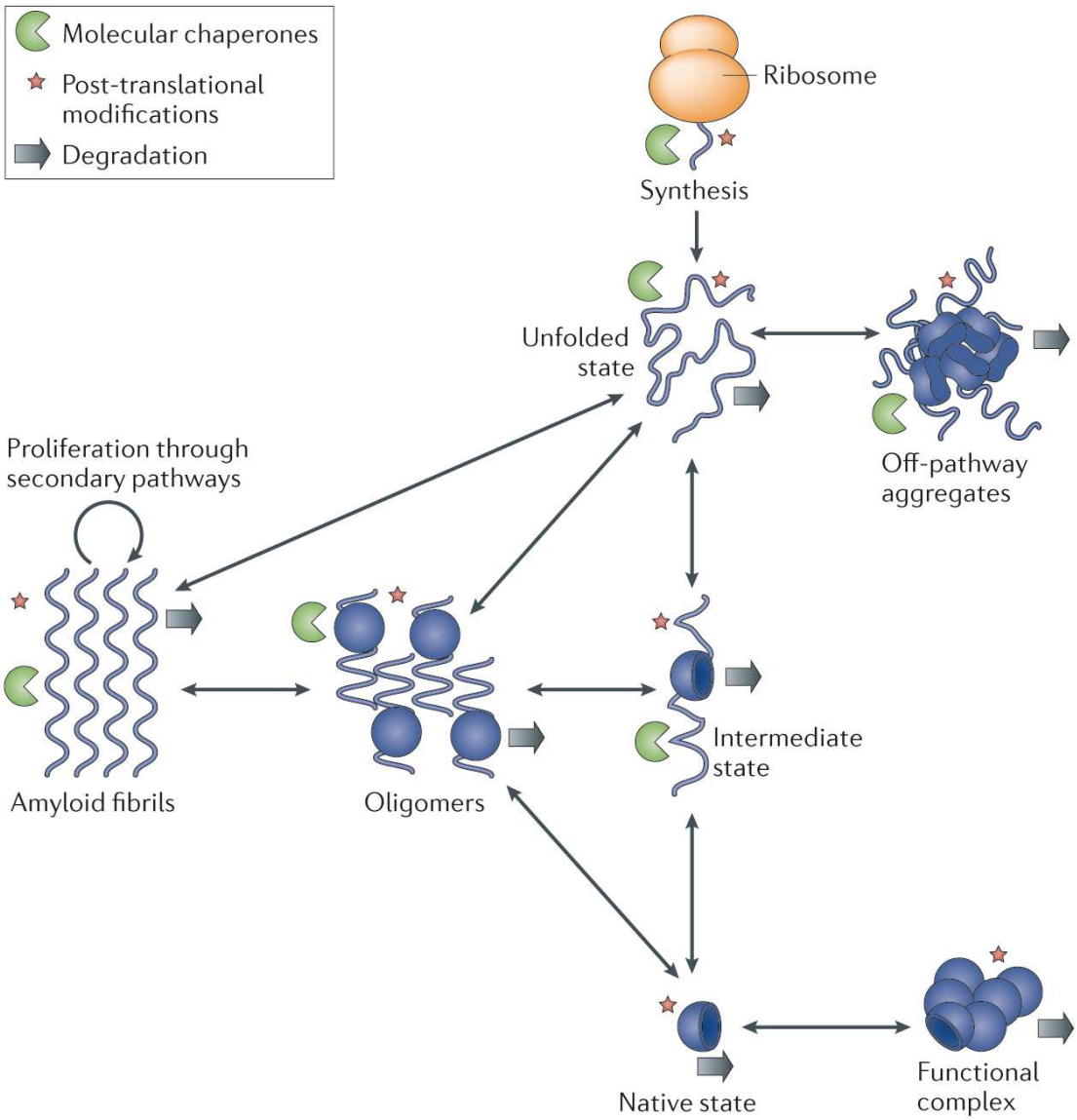


FIGURE 2.4. *Schematic representation of the different states of a protein since it is synthesized by the ribosome. From the unfolded state the protein can evolve towards different states as a function of the thermodynamic stability, the kinetic barriers, and external elements such as chaperones or post-translational modifications. In the pathway towards the amyloids, oligomers of different nature can be formed. Figure extracted from reference [99]*

circumstances, increasing the multiplicity of steps in the formation process [118, 119].

2.3.2 Amyloid oligomers and cytotoxicity

The healthy state of a cell depends on the appropriate population of all the species present in the system. The unbalance of this equilibrium can lead to deleterious effects and disease. Living systems have developed multitude of mechanisms to guarantee this homeostasis including chaperones, proteolysis, autophagy... [99]. One of the causes of the disorder is the loss of function induced by the presence of aggregates that reduce the available

concentration of the functional protein inducing a deficiency in a metabolic pathway. This is the case of systemic amyloidosis, a family of disorders are associated with the presence of large quantities of amyloid deposits in vital organs, including the liver, spleen and kidney [91]. In neurodegenerative disorders, by contrast, there are many cases no detectable correlations between the overall quantity of fibrillar aggregates and the stage of disease advancement [120–123]. It seems likely that the main cause of these type of pathologies is due to misfolding events that induce cellular damage and the gain of toxic function induced by the deleterious interaction of species formed along the fibrillation process with other proteins perturbing the regular signalling or metabolic pathways [124].

A view has emerged over the past 15 years that pre-fibrillar species, rather than mature amyloid fibrils, are likely to represent the primary pathogenic agents in non-systemic conditions, notably neurodegenerative diseases and other organ-specific conditions such as type II diabetes [91, 98, 108, 120, 121, 124–133]. Experimental evidence suggests that the oligomeric assemblies that are almost universally observed as intermediates during the aggregation process are inherently more damaging than the fibrils [108, 121, 126–133]. The origin of the toxicity of the oligomers is still a subject of intense debate [134–137]. One of the possible mechanism of toxicity may arise from their inherently misfolded nature, as they display on their surfaces chemical groups that under normal physiological conditions would not be accessible within the cellular environment [128, 138, 139]. These exposed groups can interact inappropriately with many functional cellular components, ranging from other proteins to nucleic acids and lipid membranes [140, 141].

Considering that the toxicity associated with oligomeric amyloid species appears to be related with their unique features characteristic of these species; the characterization of the structure and the formation process of these early aggregates becomes the key to understand pathologies and to design strategies to fight diseases. However, these studies are inherently challenging due to the often rapid elongation rates of the transient intermediate species and their highly heterogeneous nature.

In order to overcome these difficulties, a number of techniques have been developed, including methods that enable the direct observation and characterization of individual molecular species populated during the aggregation reaction [129, 142–146].

In chapter 4 I will describe a new approach based on SAXS that allows the structural study of the species implicated in the fibrillation process, including the toxic oligomers, without their isolation (therefore, without the potential perturbation of the equilibrium and structure).

2.3.3 α-synuclein

α-synuclein (αSN) is a 140-amino acid protein (14.5 kDa, pI=4.7), which is encoded by a single gene consisting of seven exons located in chromosome 4 and is expressed abundantly in the brain, being 1% of the total protein in soluble cytosolic brain fractions [147]. This protein was first described by Maroteaux in 1988 [148] as a neuron-specific protein localized in the presynaptic nerve termini and nucleus [147], and hence was referred to as synuclein.

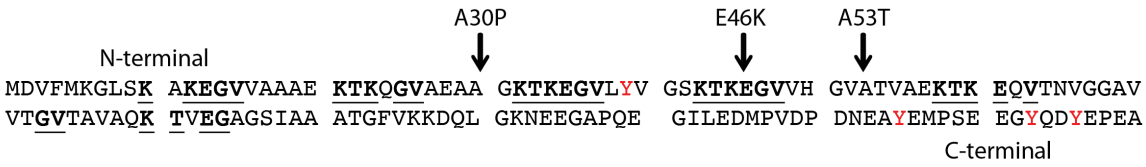


FIGURE 2.5. *The primary structure of WT α-synuclein. The imperfect KTKEGV repeats are highlighted and underlined. The position of the A30P, A53T and E46K early PD onset point mutations are indicated with arrows. The position of the four tyrosine residues in positions 39, 125, 133 and 136 are marked in red.*

The sequence of αSN (Figure 2.5) has neither tryptophan nor cysteine and it is characterized by the presence of 7 imperfect repeats of the amino acid sequence KTKEGV. It is commonly divided into three regions: 1) An amphipathic lysine-rich N-terminal region covering residue 1-60 containing 4 of the imperfect repeats. The N-terminal is playing an important role in modulating membrane interactions of αSN. 2) The hydrophobic and fibrillation-prone Non-Amyloid-β Component (NAC) region from residue 61-95 containing 3 imperfect repeats and 3) the disordered acidic and proline rich C-terminal with 14 negative residues, residue 96-140 [149–153]. The NAC region is named so because a peptide with this sequence was originally isolated from amyloid plaques in Alzheimer's disease [154] and it is very central to the aggregation process (amyloid fibrillation).

There are strong evidences that suggest that the native state of αSN in solution is a natively unfolded monomer, that is, an intrinsically disorder protein (IDP) [151, 155–159]. However, it presents an average hydrodynamic volume significantly smaller than the predicted for an extended random coil conformation [160–164].

Intriguingly, αSN is characterized by a remarkable conformational plasticity, adopting a series of different conformations depending on the environment. The natively unfolded nature of αSN is determined by its relatively low hydrophobicity and high net charge. It is expected that alterations in the protein environment leading to an increase in its hydrophobicity and/or decrease in net charge can induce partial folding [157]. In fact, these two structural parameters can be modulated via changes in the environment. For example, αSN became more ordered at pH 3.0 or at high temperature [157]. Comparable folding/compaction was observed for the protein at high temperatures, and an increase in temperature was sufficient to induce the reversible formation of some ordered secondary structure in αSN [157]. More details about different conformations of the native αSN depending the environmental conditions can be found in [165].

A broad range of studies shows that the function of αSN is somehow related to synaptic vesicle plasticity and neurotransmitter release [157, 159, 166–168] as reviewed by e.g. Lashuel et al. [151] and Stefanis [169]. Although the in vivo function of αSN is not understood in depth, the recent trend has been pointing towards αSN as being involved in different processes, such as membrane binding, modulating the affinity of the protein for the bilayer and acting as an anchor [170], or membrane remodeling [171]. Another possible role of αSN is regulating synaptic trafficking, homeostasis, and neutrotransmitter release

[94, 167, 172–174]. In addition, αSN has been proposed so be involved in other cellular processes including signal transduction, the functioning of mitochondria, and the regulation of oxidative stress [175].

2.3.3.1 Association of α-synuclein with PD and other diseases

αSN has been implicated in Parkinson's disease (PD) as well as in other neurodegenerative disorders including dementia with Lewy bodies (DLB) and multiple system atrophy (MSA), collectively referred to as synucleinopathies [176].

PD is the second most abundant neurodegenerative disease worldwide only surpassed by Alzheimer's disease. PD is expressed clinically by resting tremor, muscular rigidity, postural instability and bradykinesia as well as a number of non-motor deficiencies such as depression, olfactory dysfunction, loss of memory, psychosis and sleep disturbances. Treatment paradigms are all symptomatic as there is no cure available [177, 178].

All synucleinopathies display the accumulation of fibrillar αSN, with the distinctive cross-beta sheet structure of amyloid, in intracellular inclusions called Lewy bodies (LBs); and inside neurons, called Lewy neurites [112, 179, 180]. In MSA, αSN accumulates into the cytoplasm of glial cells in more diffuse inclusions of poorly organized bundles of αSN fibrils [112, 180–182]. Additionally, αSN has also been identified as a component of amyloid inclusions from brain tissue of Alzheimer's disease patients [183].

Several rare single-point mutations in the gene encoding αSN (SNCA) have been identified in familial cases of PD (A53T, A30P, E46K, and more recently H50Q and G51D [111, 152, 184]; with an early age of onset of the disease [185, 186].

The accumulation of αSN in tissues is the hallmark for all patients with PD, along with the genetic evidence, αSN appears to have a central role in the pathogenesis of PD. The contribution of αNS to PD and other synucleinopathies could in principle result from a loss or perturbation of the normal function of αSN, from a toxic gain of function caused by its aggregation, or from a combination of both.

Recently, misfolded forms of αSN, and other proteins associated with neurodegenerative disorders, have been shown to self-propagate and spread, sometimes described as in a "prion-like" manner, between interconnected regions of the central nervous system [187–189]. Indeed, cell-to-cell transmission of αSN aggregates has been experimentally observed [190, 191], pointing to a key role for αSN in the stepwise spreading of LB pathology and the progressive nature of PD and other synucleinopathies. Interestingly, like prions, different types of strains or fibril polymorphs of αSN have been identified. These polymorphs have been proposed to present different degrees of infectivity and induce variable neuronal vulnerability and pathology [135–137], providing some insights into why fibrillar αSN inclusions are associated with distinct types of neurodegenerative disorders. Such "prion-like" behavior is indeed an intrinsic characteristic of the self-assembly process of amyloid fibrils. Indeed, secondary nucleation mechanisms and seeding processes have been shown to be important catalytic processes in the aggregation of both αSN in PD and Aβ42 in Alzheimer's disease [99, 192, 193].

2.3.3.2 Genetic features of the structure of α-synuclein fibrils

αSN fibrils generated in vitro are morphologically and indistinguishable from those extracted from patients [194], showing the typical characteristics of the amyloid cross-β structure [195]. Although the structure of αSN fibrils has not been unambiguously determined at atomic resolution, most experimental data are consistent with the standard model of amyloid fibril, in which αSN monomers adopt antiparallel β-strands with monomeric units stacked in a parallel arrangement forming the fibril protofilament, as described in the model proposed by Riek and collaborators [196]. A recent model shows that αSN fibrils are composed of three filaments, each of them in turn formed by pairs of cross-β protofilaments. The protofilaments are composed of pairs of β-sheets that interact with other protofilaments through specific water-mediated interactions established between the side chains of the residues of the protein that form the core of the fibril [105, 197].

2.3.3.3 Conversion from the monomeric to the fibrillar state

Evidence is accumulating that monomeric αSN is an IDP [161–163]. There is also direct evidence that a similar dynamic structure is present within living cells [151, 198, 199]. By contrast, as explained above, the fibrillar form of αSN adopts mainly a highly stable and compact cross-β structure [196, 200–204]. The transition from the natively unfolded monomeric state to the fibrillar state is therefore a process of acquiring persistent structure within the polypeptide sequence.

Two types of model for the acquisition of the amyloid structure by αSN have been proposed, each of them supported by a range of theoretical and experimental evidence. One is the nucleation–polymerization model, where the structural conversion from random coil to β-sheet structure is assumed to take place at the monomeric level. In this model, the monomeric protein is assumed to adopt fully β-sheet structure, forming small oligomers (dimers or trimers) before the formation of larger ones, which will compose the final fibril. As αSN is an IDP, it is more likely that the monomer forms a β-sheet structure only partially, and this form can then trigger self-assembly such that the aggregated species would adopt the fully formed amyloid structure at a later stage [205–208].

The second model proposed is the nucleation–conversion–polymerization model. In this model, the structural conversion occurs at the oligomeric level through a unimolecular reaction from disordered to β-sheet oligomers. Such β-sheet oligomers, in an extreme case, could have fully formed amyloid-like structure, or partially formed structure that later converts into the fully formed β-sheet structure in a subsequent step. In this model, the two structurally distinct types of oligomers would coexist at the early stages of the self-assembly process. This theory is strongly supported by a number of experimental studies [114, 129, 209–212], and direct experimental evidence [114, 129, 209].

2.3.3.4 Generic features of the structure of α-synuclein oligomers

A detailed understanding of the oligomeric species generated during protein amyloid aggregation is very important for designing new strategies to treat diseases in early

stages. For that, many studies have been focused in its characterization. Highly structurally diverse oligomers of αSN have been found. This structural diversity is translated to variability in cytotoxicity and biological activity. In this section I will review some of the different oligomeric forms that have so far been report. More details can be founded in a recent review [213].

Oligomers identified during in vitro fibril formation

As I described before, one of the main challenges in the study of oligomeric species is that they are found in low concentration in a heterogeneous system, and in a transient state that advance into formation of the fibrils (more stable species) in an often rapid elongation rate. Despite this low stability, there are reports of early species during in vitro fibril formation, and αSN oligomeric species with spherical and annular appearance have been observed by atomic force microscopy (AFM) [214–216] and transmission electron microscopy (TEM) [217, 218]. Incubation of these αSN in the presence of excess monomeric αSN has been found to result in the conversion of oligomers into fibrils [217]. However, the easiest way to study these transient species is by isolating them using different physical or chemical methods.

Stabilization by lyophilization

The main method for isolating the intermediate species is through lyophilisation, that produce remarkably stable oligomers. The isolated oligomers have been found to exist as a heterogeneous population of species with an average size of approximately 30 monomers [219–221], although a wide range of sizes (from 10 mer to 90 mer) have been observed coexisting in samples of this type of oligomers [219]. In general, oligomers formed by lyophilisation appear unable to elongate and form fibrils, or to seed their formation at a significant rate [219, 220, 222].

Apetri et al. found soluble oligomeric intermediates, characterized occasionally as protofibrils, which disappear upon fibril formation; the earliest formed αSN non-fibrillar aggregates appear to be spheroidal with heights between 2 nm and 6 nm. These spheroids have a significant amount of α-helix and transition to β-structure occur during the formation of the oligomers (and further fibrils) are formed [223].

A recent study [219] describes a detailed structural characterization of stable toxic oligomers of αSN; two major size subgroups, designated 10S and 15S, with molecular masses of approximately 260 and 420 kDa on average, with variable β-sheet content and the rest of the protein being largely disordered.

In addition, it was possible to use cryo-EM image reconstruction techniques to obtain three-dimensional structural models. The cryoEM and TEM image analyses reveal essentially two major groups of structural orientations that are independent of the sizes of the oligomers, one with a "doughnut" shape (Figure 2.6) and the other one with a cylindrical appearance, consistent with some previous observations of αSN oligomers [218, 219, 224].

Stabilization by chemical compounds

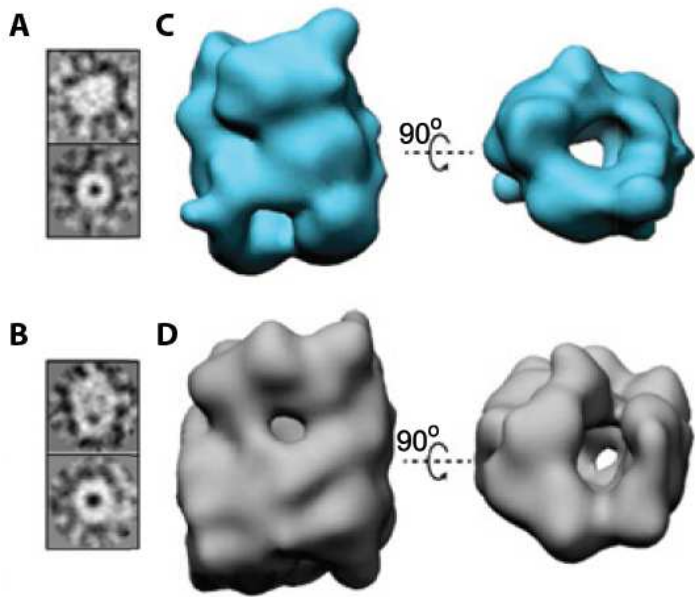


FIGURE 2.6. *Three-dimensional reconstructions of the two main size subgroups of oligomers of purified oligomeric samples of α-synuclein. (A) Example of cryoEM image of an oligomeric sample. Typical side view (Top) and end-on view (Bottom) of the small oligomeric subgroup (corresponding to the 10S oligomer subgroup) according to cryoEM. (B) Two orthogonal views, side (Left) and end-on (Right), of the 3D reconstruction of the average structure for the 10S oligomer subgroup. (C) The same views for the large structural group (corresponding to the 15S oligomer subgroup) from cryoEM and from the 3D reconstruction (D). Figure extracted from reference [219]*

The stabilization of oligomers of αSN is also possible adding chemical compounds. Many polyphenolic molecules, like baicalein [225–227] and Epigallocatechin gallate (EGCG) [228, 229] inhibit the fibril formation by αSN and also destabilize preformed αSN fibrils.

A small-molecule compound denoted FN075 has been reported to trigger the formation of αSN oligomers. The oligomers formed in the presence of FN075 have a radius of 7 nm, with the C-terminal 40 residues remaining highly disordered [230] and they are able to disrupt lipid vesicles [231] and their core structure and overall dimensions are similar to those previously reported for isolated oligomers generated upon agitation and lyophilisation [232], and the fibrils formed with and without FN075 are reported to be similar. An extensive AFM study of the effect of the addition of a wide range of metal ions to αSN indicated that annular and spherical oligomeric αSN species where formed prior to the fibrillar species [233]. Aggregation of αSN in presence of Fe^{3+} and DTT also resulted in the formation of oligomers, ranging from dimers to larger oligomers with annular worm-like structure [234]. While these oligomers possess β-sheet structure, they inhibit the formation of fibrils.

The selective loss of dopamine-containing neurons in the substantia nigra is a key

feature of PD, and as a result, extensive studies have been performed on αSN in the presence of dopamine and its analogues. Several studies show that dopamine and/or analogues of this compound promote the formation of αSN oligomers, inhibiting fibril formation, and disaggregates preformed fibrils to give disordered oligomeric aggregates [235–238].

Oligomers generated upon fibril dissaggregation

The presence of spherical and annular oligomeric species has been observed following cold denaturation of αSN fibrils using supercooling conditions (-15 C) [239, 240]. These oligomers were found to have structures intermediate between that of monomers and fibril, that are able to disrupt lipid membranes. Interestingly, these oligomers were able to elongate and form fibrils efficiently,

2.3.4 Insulin

Insulin is a 51-residue protein hormone, composed of two A and B polypeptide chains which are linked by two disulfide bonds [241], with a largely α-helical structure. It exists in solution as a mixture of different states (hexamer, dimer, monomer), depending on the environmental conditions [242]. In the secretary vesicles of pancreas, the predominant form of insulin is a three-dimers hexamer containing 2–4 zinc ions, but in order to become biologically active, insulin has to take a monomeric form [243, 244].

2.3.4.1 Association of insulin with diseases and pharmacological implications

Insulin fibrillation [242, 245–247] is an interesting process from the pharmacological point of view because amyloid deposits have been observed in patients after subcutaneous insulin infusion [248–252] . These deposits results into poor absorption of subsequently injected insulin, leading to impaired control of hyperglycemia, and a need to increase insulin dosage [253].

In vitro, at low pH and high temperature, insulin is very prone to form amyloid fibrils, causing problems during production, storage, and delivery of insulin-based drugs [242, 254, 255]. Protein aggregation is the most common and troubling manifestation of protein instability, encountered in almost all stages of protein drug development [256]. Protein aggregation, along with other physical and/or chemical instabilities of proteins, remains to be one of the major barriers for the development of protein drugs. For example, insulin can undergo both physical aggregation process, leading to formation of either soluble hexamers or insoluble fibrils and chemical aggregation process, leading to formation of either soluble dimers via cyclic anhydride intermediate or insoluble disulfide-bonded aggregates [257–259]. For that reason, it is important have a clear understanding of the protein aggregation process.

2.3.4.2 Generic features of insulin fibrils

The insulin fibrils share several common structural properties with other amyloid aggregates. Mature fibrils are suggested to be composed of intertwined protofibrils built from two to three protofilaments, each with a typical diameter of 15–50 Å [260, 261] with presence of repeating cross-β-sheets perpendicular to the fibril axis with the typical inter-strand spacing of 4.8 Å found in other amyloid fibrils [102, 182, 262].

At pH 2, mass spectrometry and hydrogen exchange measurements reveal that insulin forms soluble assemblies of up to 12 molecules in equilibrium with monomers and smaller oligomers [263, 264]. At elevated temperatures, these species further assemble into larger, irreversible aggregates and ultimately fibrils. Studies using Fourier transform infrared spectroscopy (FTIR) and circular dichroism (CD) spectroscopy indicate that the initial aggregates retain their predominantly helical structure, but that there is a subsequent conversion to beta-sheet structure [254]. EM and atomic force microscopy studies of other amyloid proteins have indicated similar assembly pathways [265–268].

A cryo-EM reported the 3D reconstructions of insulin fibrils with different morphologies. One of the fibrils was formed by two protofilaments twisting around each other, resulting in a thin type of fibril. A more compact filament was found at higher concentration, formed by a left-handed double helix with four protofilaments. A six protofilament fibril and a twisted ribbon were present also in the sample. The average size of the protofilaments was about 30 x 40 Å [260].

2.3.4.3 Insulin fibrillation process

The kinetics of the fibrillation process is affected by different factors, such as protein concentration, agitation, pH, ionic strength, anions, seeding and addition of chemical compounds [269, 270]. An increase in insulin concentration resulted in shorter lag times and faster growth of fibrils. Shorter lag times and faster growth of fibrils were seen at acidic pH versus neutral pH, whereas an increase in ionic strength resulted in shorter lag times and slower growth of fibrils. There was no clear correlation between the rate of fibril elongation and ionic strength. Agitation during fibril formation attenuated the effects of insulin concentration and ionic strength on both lag times and fibril growth [242, 261].

Insulin fibrillation is proposed to be a nucleation-dependent process [246, 271–274] and exhibits a strong time dependence, with a pronounced lag phase, followed by a very fast growth of fibrils. In a similar model than the fibrillation process for other amyloid aggregation, monomeric insulin shows a conversion of the α-helical monomer to a β-sheet intermediate in the process to form an oligomeric nucleus [245, 263, 275, 276] prior to elongation of protofilaments (by addition of these non-native monomeric intermediate).

Like in the case of αSN, oligomeric precursors of amyloid fibrils seems to be the cytotoxic species, which emphasizes the importance of characterizing such oligomeric species. The structural characterization of the fibrillation nucleus is difficult, because they are present in low concentrations and they have an inherent instability, because is the thermodynamically least favourable species [277].

To study the structure and kinetics of the species involved in the insulin fibrillation, a time-resolved synchrotron SAXS experiment was performed to study the process in solution, starting from monomeric insulin [275]. The scattering pattern from different stages in the fibrillation process was recorded and the spectra and volume fraction of the individual species was decomposed. A low-resolution shape of the oligomer reveals a helical structure with length of 200 Å with a molecular weight of 32 kDa, or 5.6 insulin monomers. The repeating unit has a length of about 700 Å and a cross-section of about 300 Å. The kinetic model proposed was an elongation via structural nucleus, where the oligomer is both the structural nucleus and the elongating building block of insulin amyloid fibrils.

In chapter 4, I will describe a method to perform the decomposition based in the same principle that the proposed in this study but in an automated way.

Chapter 3

Chemometrics

Chemometrics can be briefly described as the interaction of mathematical and statistical methods in chemical measurement analyses [278]. The breakthrough in chemometrics came with the development of new analytical techniques and powerful computers. Chemometric (also called multivariate data analysis) tools involve the analysis of data consisting of numerous variables measured from a number of samples. The aim of multivariate data analysis is to determine all the variations in a data matrix in order to find the relationships between the sample properties and its variables. There exist several different chemometric tools and they can be applied to a great variety of research fields, such as analytical chemistry, pharmaceutical sciences or environmental control. Chemometrics can be used to analyze data from different techniques such as all kind of spectroscopies. This chapter will focus on the chemometric tools used to analyze SAXS data along this thesis: Single Value Decomposition (SVD), Principal Component Analysis (PCA), Multivariate Curve Resolution using Alternating Least Squares (MCR-ALS), and Evolving Factor Analysis (EFA).

3.1 Singular Value Decomposition (SVD)

Singular Value Decomposition (SVD) is the decomposition of a complex matrix **X**, which denote an m x n matrix of real values and rank r[a]. In **X** $m \geq n$, and therefore $r \leq n$. The equation for the singular value decomposition of **X** is the following:

$$X = USV^T \qquad \text{(Eq. 3.1)}$$

where **U** is an m x n matrix with orthonormal columns (**UTU** = **I**), **S** is a n x n diagonal matrix, and **VT** is also an n x n orthonormal matrix (**VTV** = **I**). The elements of **S** are only nonzero on the diagonal, and are called singular values. The columns of **U** are called the left singular vectors, $\{\mathbf{u}_k\}$. The rows of **VT** contain the elements of the right singular vectors, $\{\mathbf{v}_k\}$.This decomposition is graphically described in the Figure 3.1. By convention, the ordering of the singular vectors is determined by the high-to-low sorting of singular values, with the highest singular value in the upper left index of the **S** matrix.

[a]Rank of a matrix is the number of linearly independent rows and columns

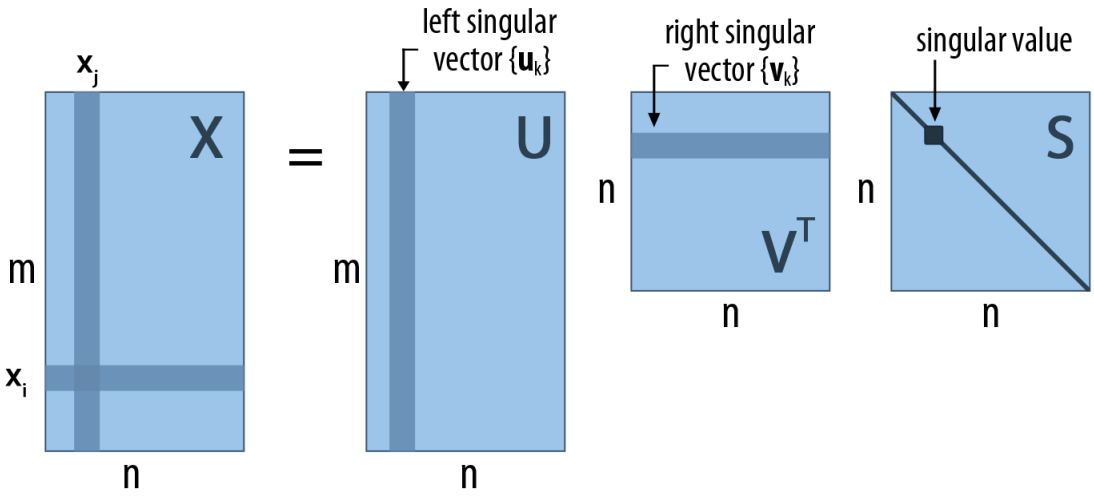


FIGURE 3.1. *Graphical description of SVD of a matrix X*

One way to calculate the SVD is to first calculate $\mathbf{V}^T$ and $\mathbf{S}$ by diagonalizing $\mathbf{X}$. This is done by multiplying both sides by $\mathbf{X^T}$ to construct two positive-definite symmetric matrices, $\mathbf{X}\mathbf{X}^T$ and $\mathbf{X}^T\mathbf{X}$

(1) Multiplying on the left:

$$X^TX = (USV^T)^T\, USV^T = U^TSV\, USV^T, \qquad (U^TU = I), \qquad X^TX = VS^2V^T$$

(2) Multiplying on the right:

$$XX^T = USV^T\, (USV^T)^T = USV^T\, U^TSV, \qquad (V^TV = I), \qquad XX^T = US^2U^T$$

Any real symmetric matrix has a spectral decomposition (also called eigenvector decomposition or eigendecomposition), which is the factorization of a matrix into a canonical form, whereby the matrix is represented in terms of its eigenvalues and eigenvectors. Using that decomposition is possible identify the eigenvectors for $\mathbf{X}^T\mathbf{X}$ as the columns of $\mathbf{V}$ and the eigenvalues as the squared diagonal elements of $\mathbf{S}$:

$$X^TXv = \lambda 2v \qquad \text{(Eq. 3.2)}$$

where $\mathbf{v}$ is the eigenvector and λ the diagonal element.

If we multiply both sides with $\mathbf{X}$:

$$(XX^T)Xv = \lambda 2Xv \qquad \text{(Eq. 3.3)}$$

Which means that there is an eigenvector $\mathbf{u} = \mathbf{Xv}$ (column of matrix $\mathbf{U}$) and eigenvalue $\lambda 2$ for $\mathbf{X}\mathbf{X}^T$. $\mathbf{X}\mathbf{X}^T$ is m x m and $\mathbf{X}^T\mathbf{X}$ which is n x n. These two matrices share n eigenvalues and the remaining $m - n$ eigenvalues of $\mathbf{X}\mathbf{X}^T$ are zero. The singular values in $\mathbf{S}$ are square

roots of eigenvalues from $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ and $\mathbf{X}\mathbf{X}^{\mathrm{T}}$.

3.2 Principal Components Analysis (PCA)

The central idea of Principal Component Analysis (PCA) is to reduce the dimensionality of a dataset consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the dataset. This is achieved by transforming to a new set of variables, which are uncorrelated and ordered so that the first few components retain most of the variation in all the original variables [279].

3.2.1 PCA by SVD

If we denote the matrix of eigenvectors sorted according to eigenvalue by $\tilde{\mathbf{U}}$, we can then do a PCA transformation of the data as $\mathbf{Y} = \tilde{\mathbf{U}}\mathbf{T}\mathbf{X}$. The eigenvectors are called the principal components. By selecting only the first d rows of $\mathbf{Y}$, we have projected the data from n down to d dimensions where $d \ll m$. The way to find the directions in the data with the largest variation is performing an eigenvector decomposition of the variance matrix and to project the data onto these directions. The first step of PCA is to calculate the covariance matrix, $\mathbf{C}$. For a data matrix $\mathbf{X}$ with m rows (samples) and n columns (variables), the covariance matrix of $\mathbf{X}$ is defined as

$$C(X) = \frac{1}{m-1}(X^T X) \qquad \text{(Eq. 3.4)}$$

It is possible decompose the matrix X with SVD to obtain

$$C(X) = \frac{1}{m-1}(US^2U^T) \qquad \text{(Eq. 3.5)}$$

Because SVD routine order the singular values in descending order and $n \leq m$, the first columns in $\mathbf{U}$ correspond to the sorted eigenvalues of $\mathbf{C}$. Finding the eigenvalues and eigenvectors of $\mathbf{C}$ is the same as finding the eigenvalues and eigenvectors of $\mathbf{X}\mathbf{X}^{\mathrm{T}}$.

The transformed data can thus be written as

$$Y = \tilde{U}^T X = \tilde{U}^T USV^T \qquad \text{(Eq. 3.6)}$$

where $\tilde{\mathbf{U}}^{\mathrm{T}}\mathbf{U}$ is a simple n x m matrix with a value one in the diagonal and zero everywhere else.

3.2.2 PCA by SVD in Matlab®

The covariance matrix is very large and difficult to work with; however, using the Eq. 3.2 and Eq. 3.3 it is possible decompose the smaller n x n matrix

$$D \equiv \frac{1}{n}X^T X \qquad \text{(Eq. 3.7)}$$

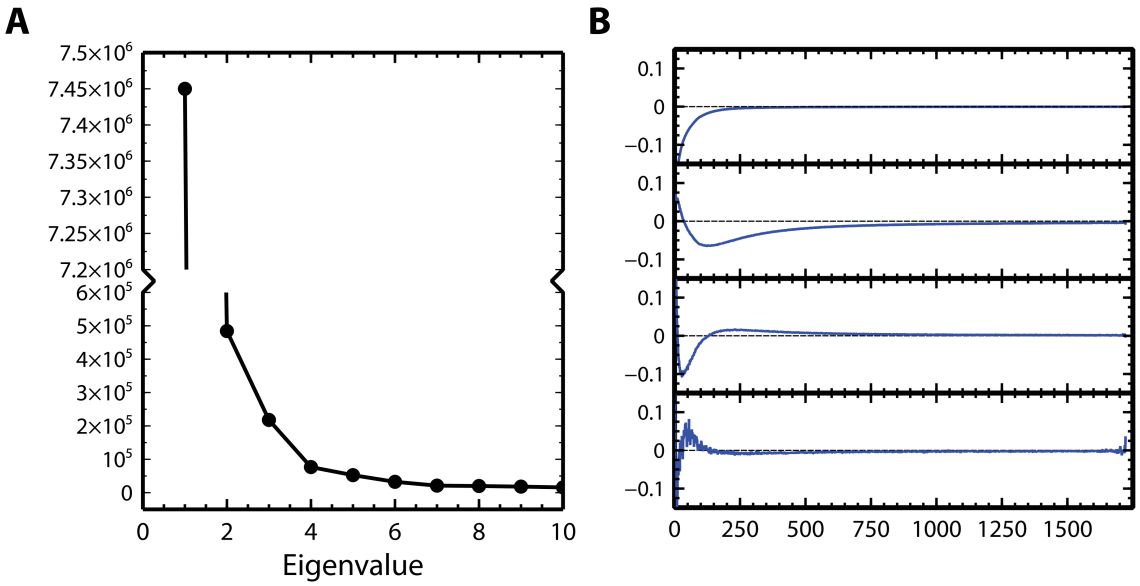


FIGURE 3.2. *A) Scree plot of eigenvalues and B) eigenvectors from a PCA of a system composed by three main components*

Given a decomposition of **D** we can find the interesting non-zero principal directions and components for **C**, **U** = **XVS**$^{-1}$. It is possible to derive the smallest matrix by using the command '`[u s v] = svd(X,0)`'. PCA performed during this thesis has been done using this approach.

3.2.3 Interpretation of PCA results

The results of PCA are shown usually in a Scree plot (Figure 3.2). This graphical approach was proposed in [280]. This approach involves plotting the variance accounted for by each principal component (eigenvalue) sorted from largest to smallest. We then look for an 'elbow' in the curve, that is, a point after which the remaining eigenvalues decline in approximately linear fashion; and retain only those components that are above the elbow. Thus, the *Scree* test calls for a relative judgment of the amount of variance accounted for by the retained components. Another way to determine the number of variables is plotting the eigenvectors, also in order, and counting the number of them that are above the noise level. Given that PCA is based on accounting for variation in the data, it is highly susceptible to the presence of outliers and influential observations.

3.3 Multivariate Curve Resolution using Alternating Least Squares (MCR-ALS)

Multivariate Curve Resolution (MCR) is a powerful chemometric tool for advanced multivariate data analysis. It may be used in the decomposition of any kind of experimental data, organized as a single data matrix or in multiple data matrices when multiple

experiments are analyzed simultaneously. The basic assumption of MCR is that measured data variance may be decomposed as the weighted sum of individual contributions (mixture) coming from the different coexistent components, each one of them defined by a set of profiles corresponding to each technique applied, and weighted according to their composition in the analyzed mixture.

Among multivariate curve resolution methods, the one based on Alternating Least Squares (MCR-ALS,) has become very popular [281–285].

The bilinear model of MCR is described by the matrix Equation:

$$D = CS^T + E \qquad \text{(Eq. 3.8)}$$

where the dimensions of the matrices are: $\mathbf{D}(I,J)$, $\mathbf{C}(I,N)$, $\mathbf{S}^T$ (N,J) and $\mathbf{R}(I,J)$; being I the number of rows in data matrix **D** (e,g. the number of spectra measured for instance at different times); J the number of columns in data matrix **D** (e.g. the number of wavelengths or momentum transfer points); and N the number of components (chemical species contributing to the spectroscopic signal). **C** matrix describes the composition contributions of the N components (concentration profiles of the different components at the different reaction times). $\mathbf{S}^T$ is the matrix describing the instrumental responses (spectra) of these N components (pure spectra profiles). Due to unavoidable experimental (noise, etc.) and modeling uncertainties this matrix decomposition is not perfect, and the differences between the measured data and their decomposition are collected in a matrix **E** of experimental errors and uncertainties. Therefore, the problem to solve in multivariate curve resolution may be mathematically stated in the following way: given the data matrix **D** find: 1) N, the number of chemical components or species causing the observed data variance, **D**; 2) find the concentration profiles of these components in matrix **C**; and 3) find the pure response or spectra profiles of these components in matrix $\mathbf{S}^T$. Stated in this way, and without any further constraint, there is not a unique set of matrices **C** and $\mathbf{S}^T$ that solve Eq. 3.8.

The MCR-ALS strategy consists of the following steps (Figure 3.3):

I. Determination of the number of significant components (**N**) present in the matrix **D** using Singular Value Decomposition (SVD) or Principal Component Analysis (PCA).

II. Generation of initial estimates of concentration or spectra profiles. Once the number of components, **N**, has been determined, an initial estimate of their concentration or spectra profiles can be selected from the experimental data matrix **D**. The initial estimation can be obtained based on methods of finding the purest variables, either using evolving factor analysis (EFA) [286] or SIMPLISMA [287].

Steps I and II can also be done using previous knowledge of the chemical problem under investigation to propose directly the number of components and the initial estimates.

III. Resolution of Equation 3.8 by the alternating least squares (ALS) algorithm.

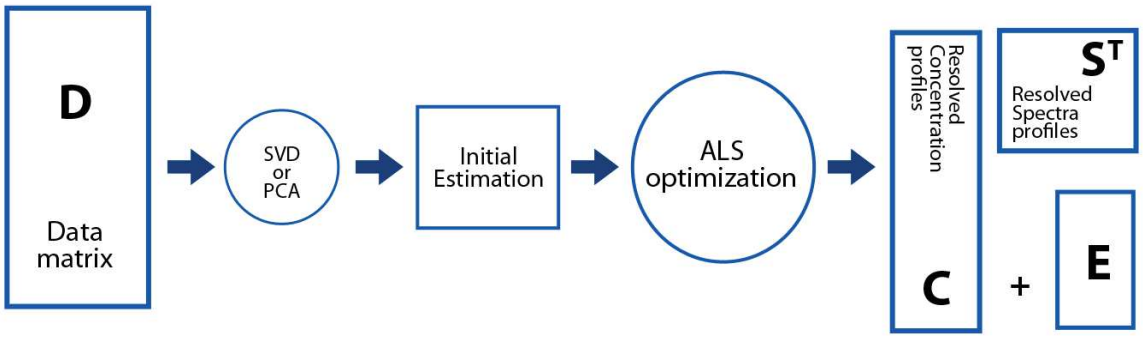


FIGURE 3.3. *Graphical description of the MCR-ALS approach.*

In the unconstrained case, this involves two steps, one that derives the concentration profiles (**C**) and another that resolves the spectra ($\mathbf{S}^T$):

$$C = DS(S^TS)^{-1} \quad \text{(Eq. 3.9)}$$

$$ST = (C^TC)^{-1}CTD \quad \text{(Eq. 3.10)}$$

The percentage of variance explained (Eq. 3.11 and standard deviation of residuals respect experimental data (Eq. 3.12) are calculated according to the following expressions where d_{ij} designs an element of the input data matrix **D** and e_{ij} is the related residual obtained from the difference between the input element and the MCR-ALS reproduction. n_{rows} and $n_{columns}$ are the number of rows and columns in the **D** matrix.

$$R^2 = \frac{\sum_{i,j} d_{ij}^2 - \sum_{i,j} e_{ij}^2}{\sum_{i,j} d_{ij}^2} \quad \text{(Eq. 3.11)}$$

$$\sigma = \sqrt{\frac{\sum_{i,j} e_{ij}^2}{n_{rows}\, n_{columns}}} \quad \text{(Eq. 3.12)}$$

Convergence is achieved when in two consecutive iterative cycles, relative differences in standard deviations of the residuals between experimental and ALS calculated data values are less than the convergence criterion value previously selected.

The lack of fit is defined as the difference among the input data **D** and the data reproduced from the $\mathbf{CS}^T$ product obtained by MCR-ALS. This value is calculated according to the expression:

$$\text{lack of fit (\%)} = 100\sqrt{\frac{\sum_{i,j} e_{ij}^2}{\sum_{i,j} d_{ij}^2}} \quad \text{(Eq. 3.13)}$$

where d_{ij} and e_{ij} are the same as above. Two different values of lack of fit are calculated, differing on the definition of the input data matrix **D** (either the raw experimental data matrix or the PCA reproduced data matrix using the same number of components as in the

MCR-ALS model). These values are useful to understand whether experimental data were well fitted.

3.3.1 Ambiguity

In general, the bilinear decomposition is ambiguous if no additional information is provided. In other words, there is a rotational and a scale freedom in the solutions of Eq. 3.8. This freedom has as a result that it is possible to find an infinite number of mathematically equivalent solutions, since there are an infinite number of matrices **C** and $\mathbf{S}^T$ that, when they are multiplied, will produce the same result, the data matrix $\mathbf{D} = \mathbf{CS}^T$. This phenomenon is called ambiguity [288].

This indeterminacy is described mathematically by the following equation:

$$D^* = C_{old}S^T_{old} = (C_{old}T^{-1})(TS^T_{old}) = C_{new}S^T_{new} \quad \text{(Eq. 3.14)}$$

According with Eq. 3.14, any invertible matrix **T**(*N*,*N*) gives a new set of equivalent solutions of the MCR-ALS model. Or said in other words, any linear combinations of **C** and $\mathbf{S}^T$ solutions will also produce new solutions of the bilinear model.

Fortunately, it is usually possible to reduce considerably the number of possible solutions (and consequently the range of solutions for **C** and $\mathbf{S}^T$ matrices) by introducing constraints (described in detail in the next section) derived from the physical nature of the system and/or from prior knowledge of the problem under study.

3.3.2 Constraints

Constraints can be incorporated into the optimization process in order to render the solution chemically meaningful. A constraint is defined as a particular characteristic of chemical or mathematical nature that the spectra of the pure components or the concentration profiles must obey. The most classical constraints are the following.

3.3.2.1 *Non-negativity*

Non-negativity (Figure 3.4) implies that the decomposed matrix cannot have a negative value. It can always be applied to the concentration profiles and also to multiple types of spectra (not for derivative and difference spectra or for Circular Dichroism). The application of the *non-negativity* constraint can be carried out according to different least squares approaches, the classical non-negative least squares (nnls) [289] and the more recent fast non-negative least squares (fnnls) [290]. An additional option is replacing negative values by zeros. This option is useful when the other algorithms fail for some reason or take too long.

3.3.2.2 *Closure*

Closure constraint (Figure 3.5) is related with mass balance in closed systems, and it can be introduced for the concentration **C** matrix. The total concentration of the system

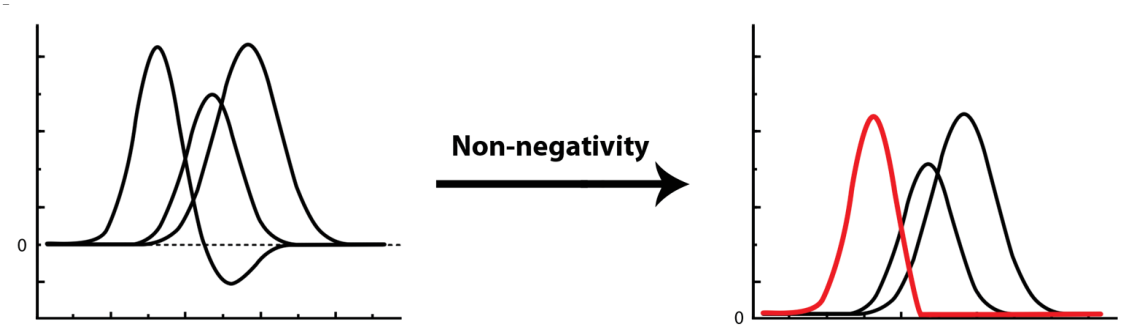


FIGURE 3.4. *Graphical description of non-negativity constraint applied to concentration profiles*

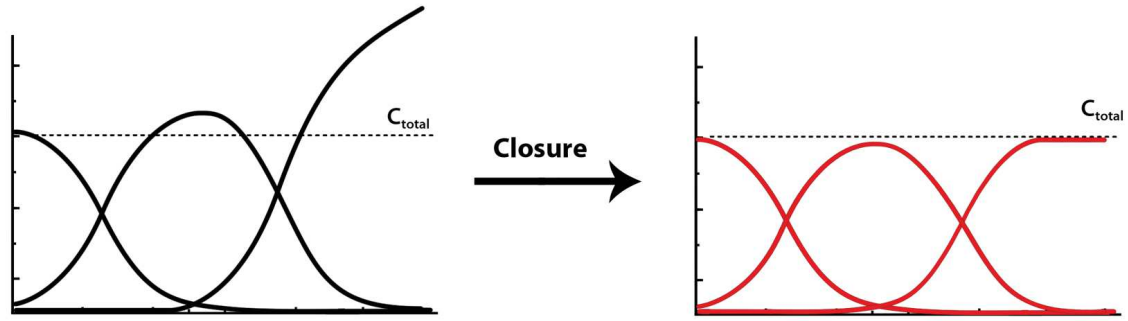


FIGURE 3.5. *Graphical description of closure constraint. A constant closure value, equal to the total concentration of the components, is applied to the concentration profiles.*

can be fixed to a single value (usually equal to 1) or a variable value if the variation of the closure constant along the experiment is known (e.g. titration experiments with known dilutions).

3.3.2.3 *Unimodality*

Unimodality constraint (Figure 3.6) forces the presence of a single maximum per profile, which in some cases can be applied to concentrations profiles (e.g reaction and chromatographic systems). There exist two different options to apply this constraint: 'vertical' and 'horizontal', which mean that the secondary maxima are cut vertically or horizontally respectively [291]. A third implementation, 'average', applies the constraint in a smoother way, taking averages similarly as in unimodal least squares algorithms [292].

3.3.2.4 *Equality*

Equality constraint (Figure 3.7) refers to the possibility to fix known values in the concentration profiles or in the spectra during the optimization, e.g. pure spectra of known compounds or selectivity/local rank information. Selectivity/local rank information can be defined as an *equality* constraint when one or several species are known not to be present

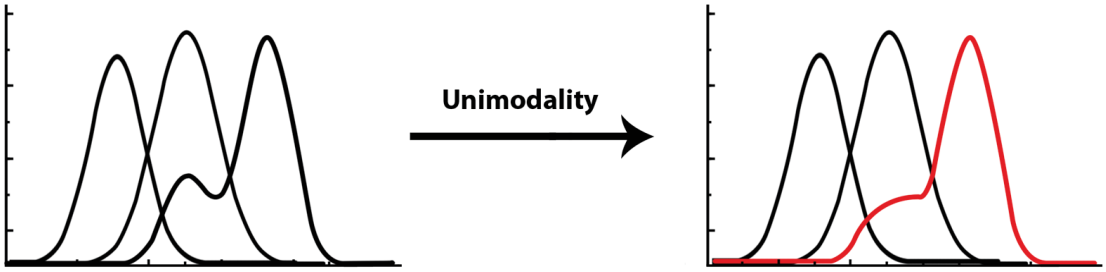


FIGURE 3.6. *Graphical description of unimodality constraint applied to concentration profiles. This example shows the 'average' implementation.*

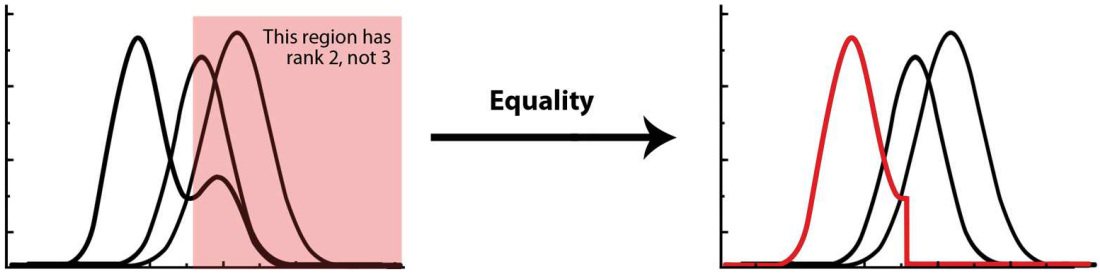


FIGURE 3.7. *Graphical description of the equality constraint. In this case local rank information was used to impose the presence of two species instead of three in a determined region of the concentration profile.*

in a particular region or window of the dataset (in both the concentration or in the spectral directions). The most restrictive use of the selectivity constraint is when only a single species is present in a region of the dataset or for set of experiments, since this constraint significantly reduces (or even suppresses) the ambiguity of results. A method for determination of local rank information regions will be described in the section 3.4.

3.4 Evolving Factor Analysis (EFA)

The use of local rank information (see previous section) is crucial in the resolution of dynamic multicomponent systems. Actually, some decomposition methods, such as MCR-ALS, take advantage of the local rank information to build initial estimates or to set equality constraints, which are important for obtaining a good quality solution during the decomposition of complex systems [284, 285, 293]. Sometimes, the possibility to get unique solutions in resolution depends mainly on the accurate determination of the local rank [294]. Methods based on PCA, such as Evolving Factor Analysis (EFA) [286, 295], are designed to detect zones with a number of compounds smaller than the total rank. In these methods, the interest is not limited to the determination of the total number of components, but to the location and evolution of each of these contributions.

EFA was designed as a chemometric tool to monitor chemical processes [286, 295]. The evolution of a chemical system is gradually known by recording a new response vector

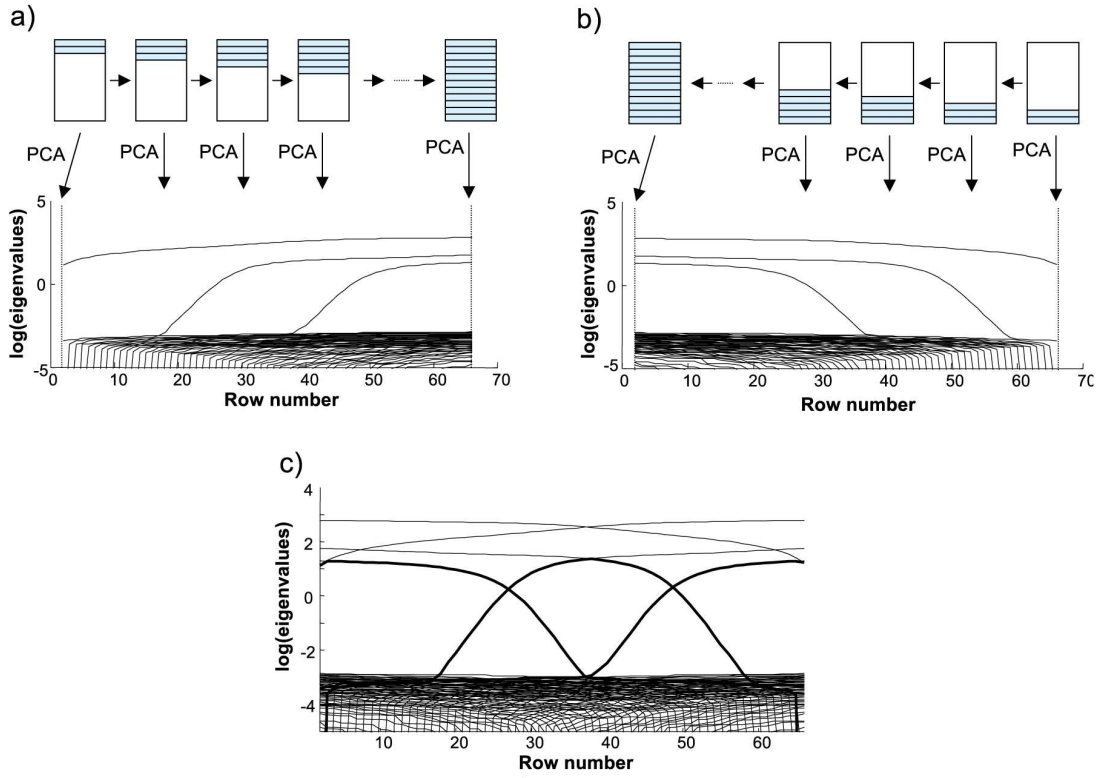


FIGURE 3.8. *Graphical information derived from Evolving Factor Analysis (EFA): (a) forward EFA plot, (b) backward EFA plot, (c) combined forward (lines) and backward (dashed lines) EFA plot and derived initial estimates of concentration profiles (bold lines). The example represents a three-component reaction system. Figure extracted from [296].*

at each stage of the process under study. EFA performs subsequent PCA on gradually increasing submatrices in the process direction, enlarged by adding one new row (response) at a time. This procedure is performed from top to bottom of the dataset (forward EFA) and from bottom to top (backward EFA). Figure 3.8 displays the way to plot and interpret the information provided by EFA [296]. The forward and backward EFA plots are built by representing the singular values (or the log(eigenvalues)) of each PCA analysis vs. the process variable related to the last row included in the window analyzed. The lines connecting all the analogous singular values (s.v.), i.e., all the 1st s.v., the 2nd s.v., the *i*th s.v., ... indicate the evolution of the singular values along the process, as a consequence, the variation of the process contributions. A new singular value line above the noise level defined by the pool of non-significant singular values indicates the emergence (forward EFA, 3.8A) or decay (backward EFA, Figure 3.8B) of a process contribution. Figure 3.8C shows how to build initial concentration profiles from the overlapped forward and backward EFA plots in sequential processes. In general, each element in the derived concentration profile is selected as the smallest value between the forward and backward s.v. lines to be combined.

Part II

RESULTS

Chapter 4

Amyloids, SAXS and chemometrics

As described in the chapter 8, some of the most important challenges in structural biology today concern the study of processes involving highly dynamic large macromolecular complexes, the presence of concerted conformational fluctuations, and unstable, developing systems. The common feature of all these systems is their heterogeneity with multiple species or conformations coexisting in equilibrium, the so-called polydispersity.

Protein amyloid fibril formation is an example of such challenging systems (section 2.3). Protein fibrils are the hallmark of a number of severe diseases, notably the most common neurodegenerative pathologies such as Alzheimer's, Parkinson's, and Huntington's diseases [91, 92]. Protein fibrils are the final stage of the amyloid formation, while several soluble and transient oligomeric states are formed along the process. Increasing evidence places a central role on these transient species in the advancement of the disease, pinpointing the importance of their structural characterization [91, 127, 133, 297]. This is, however, inherently difficult as the oligomers only exist in the context of other amyloidogenic species, and their physical isolation can perturb the equilibrium and potentially modify their structure.

The application of traditional structural biology methods, such as macromolecular crystallography and high-resolution nuclear magnetic resonance (NMR), is not straightforward for the structural analysis of changes associated with such highly complex systems like amyloids. Small Angle X-ray Scattering (SAXS) is an ideal method for such investigations [231, 232, 275, 298, 299] due to its two main advantages (detailed description in section 1.1. One is that SAXS allows the study of a wide range of molecular sizes, which is important because amyloid systems contain a mixture of particles with very different sizes, from small monomers to big fibrils. The other important advantage results from the additive properties of SAXS data, thus measurements performed on a mixture correspond to a population-weighted average of the signal originating from all coexisting species. As a consequence, a dataset consisting of multiple curves obtained from developing mixtures with different relative populations of the same species is inherently very rich in information. In principle, it is possible to decompose such data series into the scattering profiles of the individual constituents (structures) as well as their relative populations (kinetics/thermodynamics) without physically isolating the coexisting species. The isolation of data originating from intermediately occurring species has been previously performed by different laborious and non-automated approaches for several amyloidogenic proteins

[232, 275, 298–300]. Unfortunately, such strategies can only be applied to systems of limited complexity [300], i.e., with a low number of coexisting species, and the imposition of initial and final data curves as species pure. Therefore, it is clear that more robust and objective approaches are needed.

Objective decomposition of large datasets using chemometric approaches is routinely used in many research fields including analytical and organic chemistry and metabolomics [282, 301, 302]. One of the most popular chemometric routines is multivariate curve resolution using alternate least squares (MCR-ALS) [281, 284] (see section 3.3. It has been previously shown that chemometrics in general and MCR-ALS in particular are powerful tools when combined with SAXS, allowing the study of transient biomolecular complexes [303, 304] and folding processes [305–308] by SAXS or wide-angle X-ray scattering [309]. These systems were, however, significantly less complex than amyloidogenesis.

To analyze such complex systems as fibrillation, we have developed a chemometric tool, which we name COSMiCS (Complex Objective Structural analysis of Multi-Component Systems) [310] (Paper I) , to analyze SAXS data measured along the amyloid formation. Using COSMiCS we have investigated two proteins with high biomedical interest: insulin and a familial mutant of α-synuclein (αSN_{E46K}). Details of COSMiCS will be described in section 4.4.2.3.

4.1 Insulin

Time-resolved synchrotron SAXS data were collected along 11 hours from twenty-eight fibrillating insulin samples, under experimental conditions comparable to those applied in a previous study [275] while simultaneously monitoring the fibrillation kinetics with ThT fluorescence. Details of the collection of the data and primary data analysis can be found in next section.

4.1.1 Primary insulin data analysis

The resulting profiles exhibit drastically increasing intensities in the small angle region of the scattering patterns (Figure 4.1) signifying a notable evolution from monomeric to fibril state. Using Guinier's approach, this evolution can be followed from the extrapolated forward scattering (I(*0*)) and the calculated radii of gyration (R_g) (Figure 4.2). It is possible to see how the increasing in the ThT signal intensity is coherent with the increasing of R_g and I(*0*), presenting the same lag-phase, with a growth phase that start at the same point. However, the development of I(*0*) does not relate directly to the molecular weight of a fibrillating species, as it represents an average of all the species in the solution. The quality of the Guinier approximation is strongly declining as the fibrillation takes place, and the amount of data points included in the Guinier approximation is shrinking during fibrillation, as the fibrils grow larger than the resolution limit of the data. However, although the uncertainty of the R_g value is increasing, it still provides a general idea of the progress of the fibrillation.

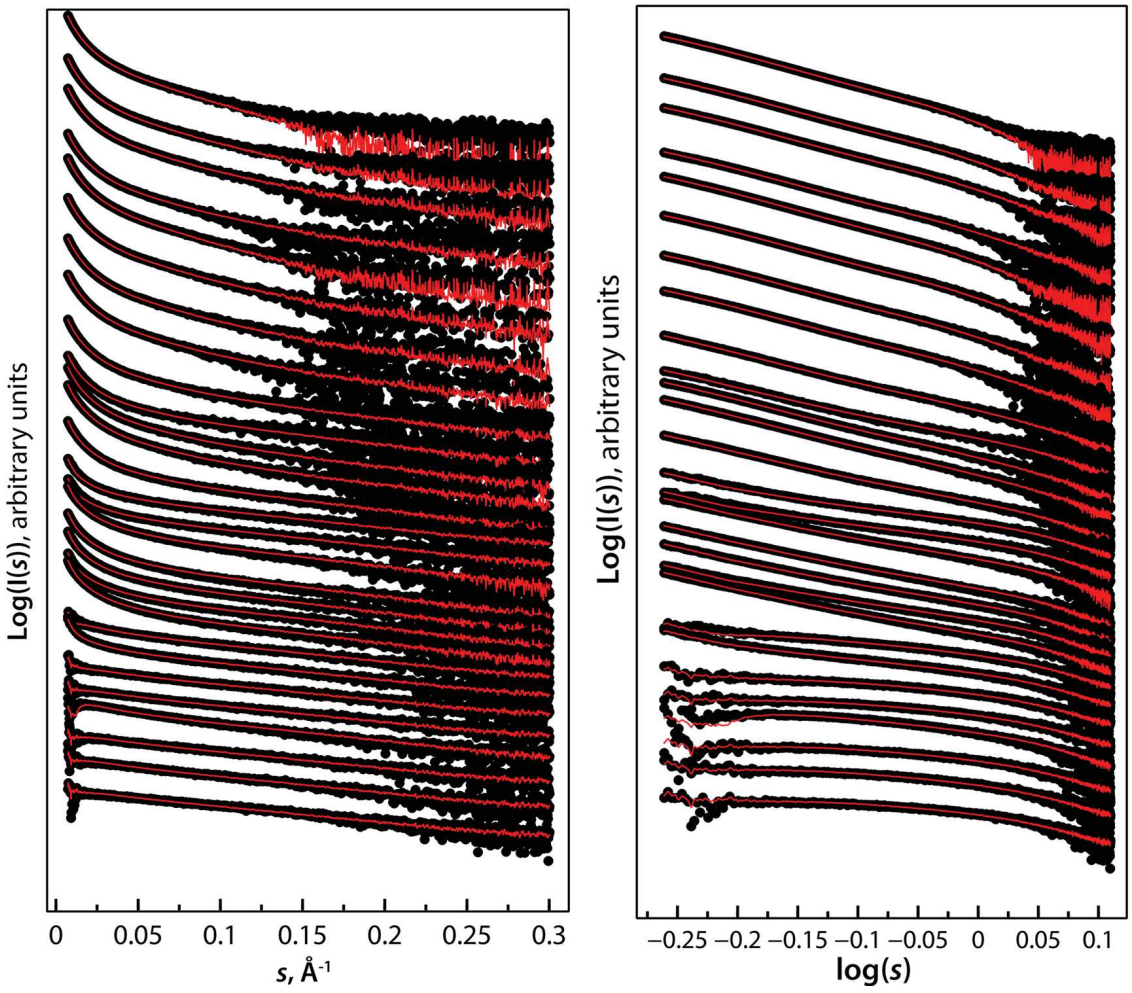


FIGURE 4.1. *SAXS profiles recorded during the evolution of fibrillation of insulin. Scattering intensity profiles (black dots) in logarithmic scale as a function of the momentum transfer ($s = 4\pi\sin(\theta)/\ \lambda$ [2θ, scattering angle; $\lambda = 1.5$ Å, X-ray wavelength]) measured from t = 0 (bottom curve) to t = 11 h (top curve). COSMiCS fitted model combining Absolute and Holtzer (AH) data representations with three species are displayed as solid red lines. Curves are translated arbitrarily along the y-axis for clarity. Semi-logarithmic representation (left) and log-log representation (right).*

The setup of the experiment consists in multiple wells in a platereader and each of them runs an individual fibrillation process (see details in the Material and methods section, 4.4.1). Although all the samples come from the same batch of protein and have the same experimental conditions, the stochasticity of the fibrillation process produces different kinetics among the wells (an example of how different can be these kinetics is showed for αSN_{E46K} in the Figure 4.27). For this reason, the content of the measured sample depends on the status of the fibrillation process in the moment of the collection. This low reproducibility generates a SAXS dataset that follows the trend of a fibrillation process, with increasing I(*0*) and R_g values, but with individual values scattered around the main profile.

A Principal component analysis (PCA) was performed on the complete dataset and

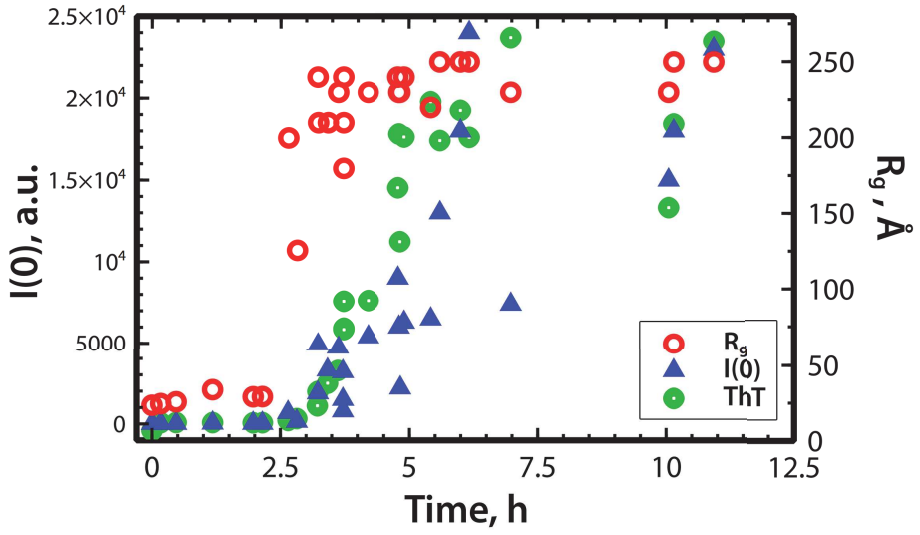


FIGURE 4.2. *Primary SAXS data analysis for insulin. Average radii of gyration, R_g, (hollow red circles), and forward scattering, I(0) (blue triangles) as estimated by Guinier's approximation are displayed as a function of time. ThT fluorescence values are displayed in green on arbitrary scale.*

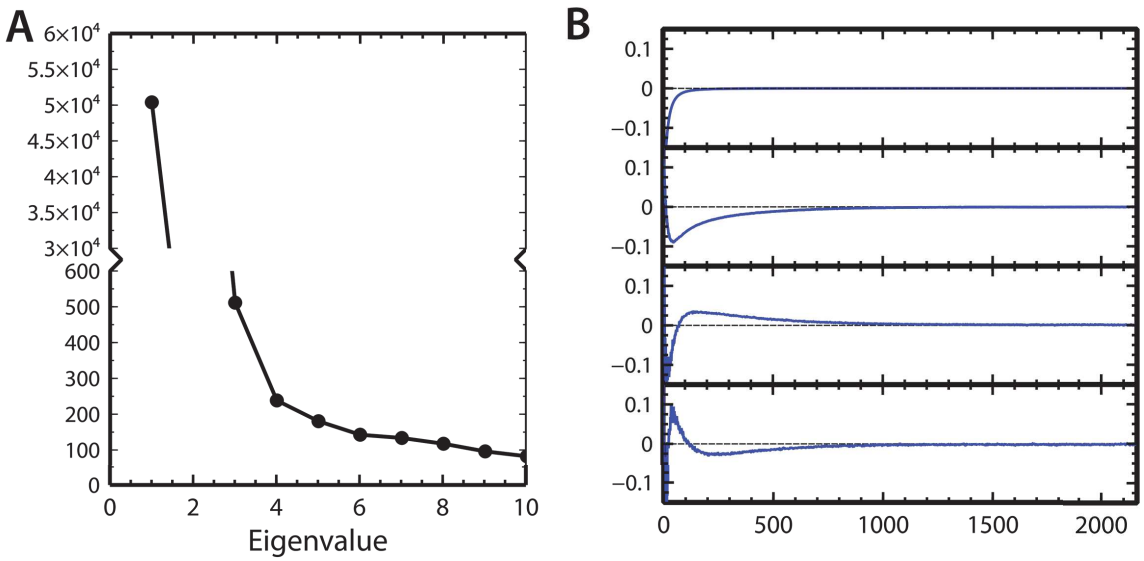


FIGURE 4.3. *Principal Component Analysis (PCA) of the complete insulin datasets. The analysis of the ten first eigenvalues (A) and the five first eigenvectors (B) suggests that three major species are present along the fibrillation process Validation of the presence of three main species contributing to the SAXS datasets is performed using MCR-ALS.*

analyzed according to section 3.2. The results indicated that three individual species significantly contributed to the time-evolution of the SAXS data of insulin (Figure 4.3).

4.1.2 Decomposition of insulin data with MCR-ALS

The decomposition of the insulin SAXS intensities was performed using the MCR-ALS chemometric approach. The principles behind MCR-ALS have been described in the

Representations included					Insulin[a]	
					Complete Dataset	
Code	Absolute I(s)	Holtzer I(s)*s	Kratky $I(s)*s^2$	Porod $I(s)*s^4$	$\langle\chi_i^2\rangle$	χ_i^2 Range
A	+				4.38	1.57-11.85
AH	+	+			3.15[b]	1.37-12.90
AK	+		+		3.75	1.37-10.33
AP	+			+	5.61	1.49-17.02
AHK	+	+	+		3.72	1.37-13.07
AHP	+	+		+	5.45	1.38-16.92
AKP	+		+	+	4.75	1.48-15.80
AHKP	+	+	+	+	4.54	1.39-15.21

TABLE 4.1. *Fitting of the insulin SAXS datasets with COSMiCS using different combinations of data matrices.*

a- Analysis performed using three species.
b- Optimal solutions used for the structural analysis.

section 3.3. However, due to the ambiguity mentioned before, MCR-ALS did not provide a good description of the data, as evidenced by the large average χ_i^2 of the fit to the 28 data curves, $\langle\chi_i^2\rangle$ = 4.38 (Table 4.1). In addition, MCR-ALS-derived SAXS curves had noticeable artifacts and large uncertainties, especially in the low-*s* region; and the decomposed curve for the fibril is truncated from $s = 0.15$ Å^{-1} (Figure 4.4A). These curves yielded non-physical pairwise distance distribution functions, *p(r)*, with large error bars (Figure 4.4). Concentration profiles also displayed non-coherent behavior with a higher initial concentration for the intermediate species compared with that of the native state, and large error bars (Figure 4.4C).

A possible explanation of the poor description of the data using the MCR-ALS method could be the need of another species to be able to explain the data. For that, another MCR-ALS analysis was performed with an additional fourth species. This approach did not improve neither the quality of the fit ($\langle\chi_i^2\rangle$ = 5.08) nor the intelligibility of the results (Figure 4.5). The derived curves using 4 species showed anomalies at the low-*s* range and no-coherent populations. The *p(r)* derived from the curves present big uncertainties and even negative values (non-physical sense), which is a clear evidence of a wrong solution.

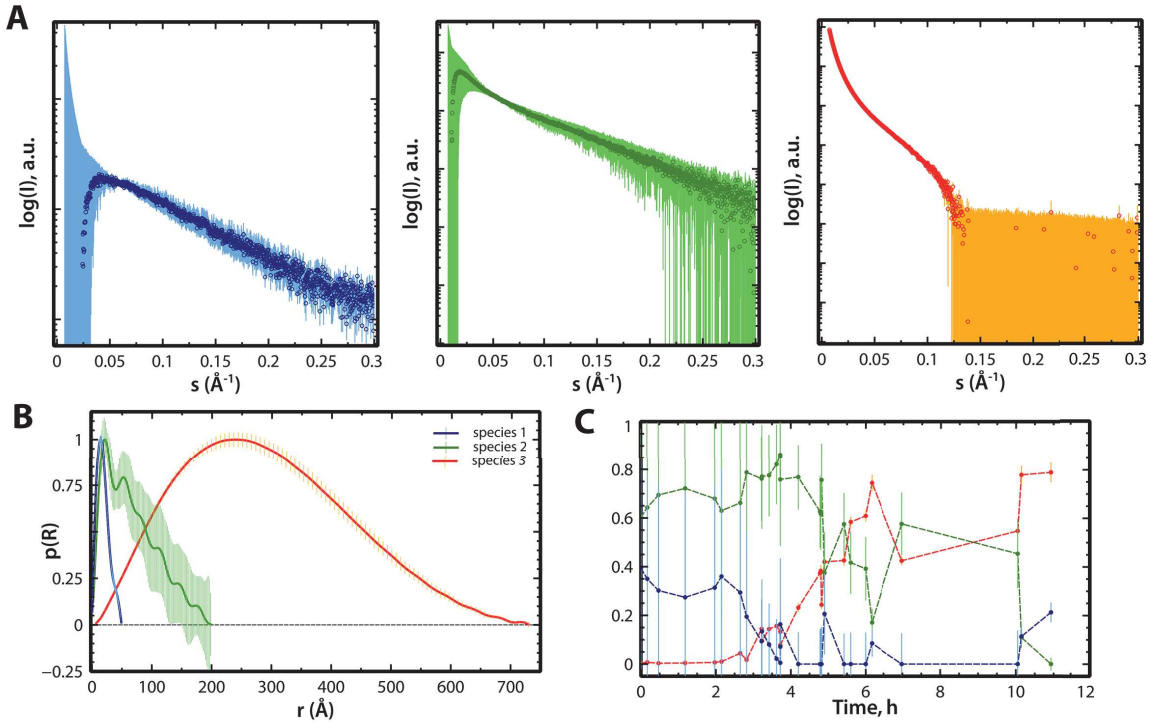


FIGURE 4.4. *Optimized results from the decomposition of the insulin data with MCRALS using only the Absolute scale data representation, and imposing the presence of three species in the mixture. (A) Spectra profiles for each species: native-like species (blue), intermediate oligomer (green) and fibril (red). (B) Pair-wise distance distribution functions, p(r), for the three estimated scattering species computed from the derived curves using GNOM [311]. (C) Concentration profiles (same color code). Error bars computed using our Monte Carlo approach are displayed in both SAXS spectra and concentration profiles.*

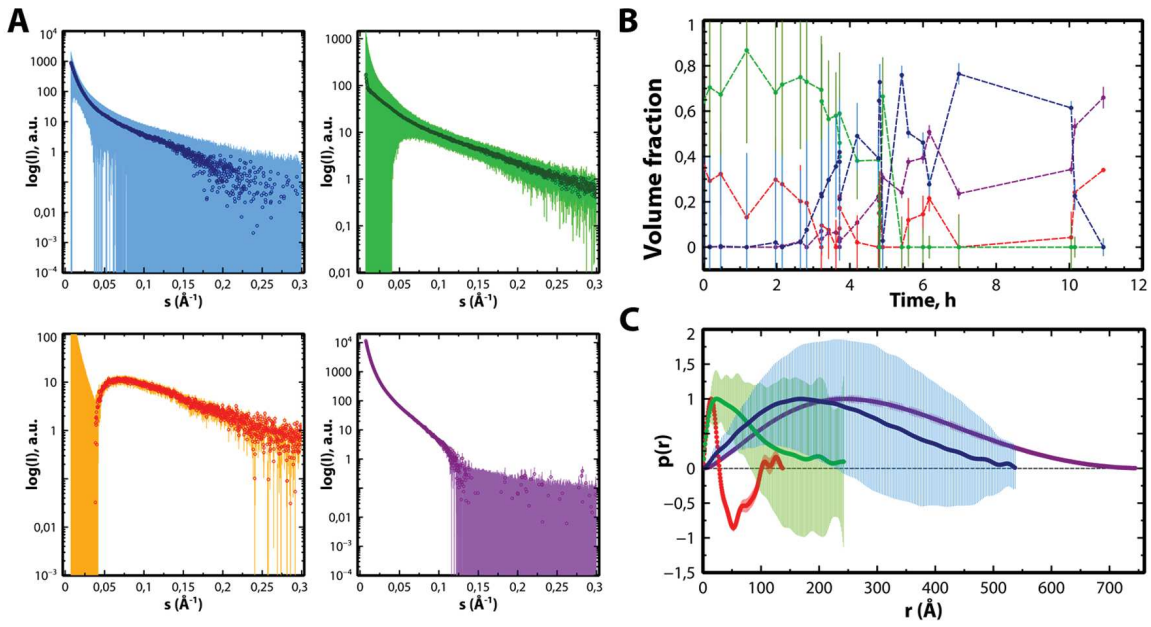


FIGURE 4.5. *(A) SAXS profiles in semi-logarithmic scale from the decomposition of the insulin dataset using MCR-ALS using 4 species. (B) Time-dependent concentration profiles derived from MCR-ALS for each species with the same color code. (C) Pairwise distribution functions derived from the individual curves*

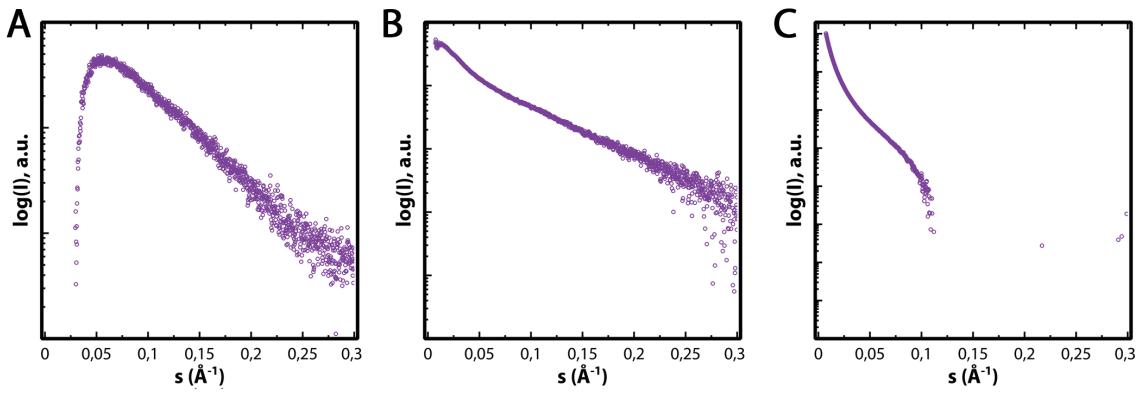


FIGURE 4.6. *Optimized SAXS curves for insulin dataset obtained using MCR-ALS 2.0. MCR-ALS 2.0, which uses the experimental errors along the optimization process, has been applied to the absolute value representation of the SAXS dataset. Severe artifacts are observed for the curves which demonstrates that weighting the experimental intensity values by their associated errors along the optimization is not enough to allow the correct decomposition of the pure species.*

4.1.3 Decomposition with MCR-ALS using weighted data

SAXS data have different level of experimental error depending on the momentum transfer. The classic MCR-ALS does not include errors in the decomposition, but just the intensity values. A new version of MCR-ALS has been developed recently, where a weighted least-squares analysis is implemented. However, the use of experimental errors to weight the agreement to the scattering intensities along the optimization process did not improve the quality of the resulting species-pure curves, which presented strong artifacts at low angles (Figure 4.6).

4.1.4 COSMiCS analysis of insulin data

Degeneracy of mathematical solutions poses an intrinsic limitation to chemometric methods [285, 312, 313]. This ambiguity problem is the origin of a non-optimal solution when a large SAXS dataset is analyzed. Besides the use of constraints, the most efficient way to reduce ambiguity in MCR-ALS is the simultaneous analysis of multiple datasets measured under different experimental conditions and/or including additional data simultaneously measured using complementary techniques. However, in present SAXS beamlines the simultaneous measurement of complementary spectroscopic data is not straightforward. We overcome this limitation by introducing different SAXS representations in the analysis. We call this new approach COSMiCS, which simultaneously fits multiple representations of the same SAXS dataset, including, in addition to the absolute values (I(s)), the commonly used data representations introduced by Kratky (I(s)·s^2) [17], Porod (I(s)·s^4) [14], and Holtzer (I(s)·s) [20]; the latter has been recently visited by Rambo and Tainer [314]. Although this does not enrich the information content in our

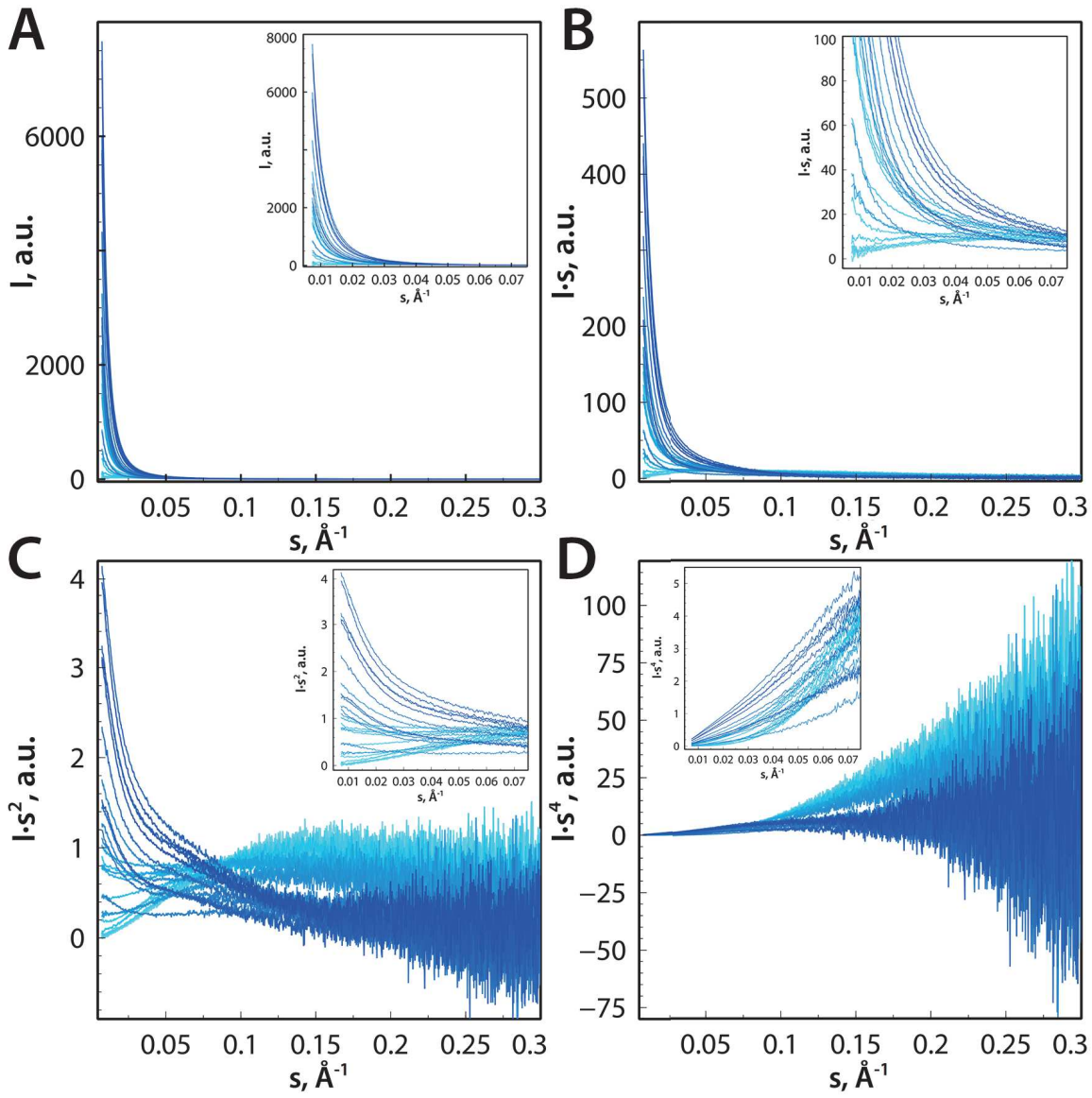


FIGURE 4.7. *Representations of the SAXS data measured along the fibrillation of insulin. (A) Absolute values, I(s). (B) Holtzer, I(s)·s. (C) Kratky, I(s)·s2. (D) Porod, I(s)·s4. Different features are observed in the momentum transfer range displayed along the fibrillation process represented with a blue color scale, from light blue (t=0) to dark blue (t=11 h).*

input dataset, it emphasizes the structural changes at different time points along the fibrillation pathway. Species appearing along the fibrillation process present distinct structural features that emerge at specific momentum transfer ranges (resolution) and are captured differently by SAXS data representations (Figure 4.7). The consequent enhancement of data variability along the fibrillation process increases the discrimination power of the MCR-ALS optimization by reducing the ambiguity of the mathematical solutions. The simultaneous use of multiple SAXS data representations was tested first on simulated data (see section 5.7 for details).

We used all combinations of dataset representations for COSMiCS analyses (Table 4.1). With the exception of Porod's representation, the inclusion of a second SAXS data

representation yielded a systematic improvement in the quality of the fit to the 28 experimental profiles, while no improvement in the overall fit to the dataset was observed when increasing the number of data matrix representations to three. The best agreement to the dataset was obtained by combining the absolute value representation with the Holtzer representation (AH; $\langle {\chi_i}^2 \rangle$ = 3.15). The fitting to the 28 curves (Figure 4.1) demonstrates that the linear combination of three species and the population profile properly describes the complete experimental dataset with no systematic deviations along the time evolution.

Choosing the right number of species to describe the system is very important. COSMiCS is a more powerful method to discriminate between different solutions than PCA, which is better to be used to obtain an estimation of the number of species. For this, is a good practice to repeat the COSMiCS analysis with one species less and one species more than the estimated by PCA to confirm the minimum number of species necessary to explain the data. We repeated the complete COSMiCS analysis using two and four species (Figure 4.8A and 4.8B). When fitting with two species, $\langle {\chi_i}^2 \rangle$ significantly increased to $\langle {\chi_i}^2 \rangle$ = 7.66, and the optimized spectra from the analysis with four yield a not physically meaningful solution (Figure 4.8C and 4.8D).

4.1.5 Structural analysis of the components of insulin

The pure SAXS curves of the insulin species obtained from the COSMiCS analysis are shown in Figure 4.8. The capacity to derive species-specific information enables their detailed structural investigation. An important piece of information that can be derived is the oligomeric state through the molecular weight (MW) estimation. We have applied several available strategies to derive these two parameters (Table 4.2) [24, 50, 314, 315] providing coherent MW estimations despite their distinct approaches and inherent limitations (see section 1.2.6). The use of an external standard depends on correct estimation of protein concentration and is very sensitive to the presence of small quantities of non-properly decomposed species, such as unspecific aggregates, while SAXS invariant volume of correlation, V_c, has not been calibrated for very large species [314]. Consequently, we have chosen V_c for small species, and the external standard for large aggregates.

The *p(r)* functions derived from the profiles indicate that the three species are a small particle, an intermediate oligomer, and a large aggregate (Figure 4.9B). This result is in very good agreement with previous manual decomposition applied to a similar fibrillation series of insulin [275]. The smallest species corresponds to a slightly elongated particle (Figure 4.10) with a R_g of 23.2 $\pm$ 0.2 Å, in full agreement with the need for a monomeric partially unfolded species to trigger insulin fibrillation [271]. The partially unfolded nature of this species is also evident when plotting the isolated curve in the Kratky presentation and the low agreement with the crystallographic structure of the monomeric insulin (Figure 4.11). The oligomer, which has an estimated mass of 4–8 protomers (Table 4.2), features an elongated shape with a R_g of 48.5 $\pm$ 0.1 Å and a D_{max} =196 $\pm$ 10 Å according to the derived *p(r)* (Figure 4.9B). The *p(r)* function was used to derive a low-resolution structure of this elusive intermediate species. The resulting structure, which perfectly describes the experimental curve, shows that the oligomer is an elongated particle, with a bent/helical

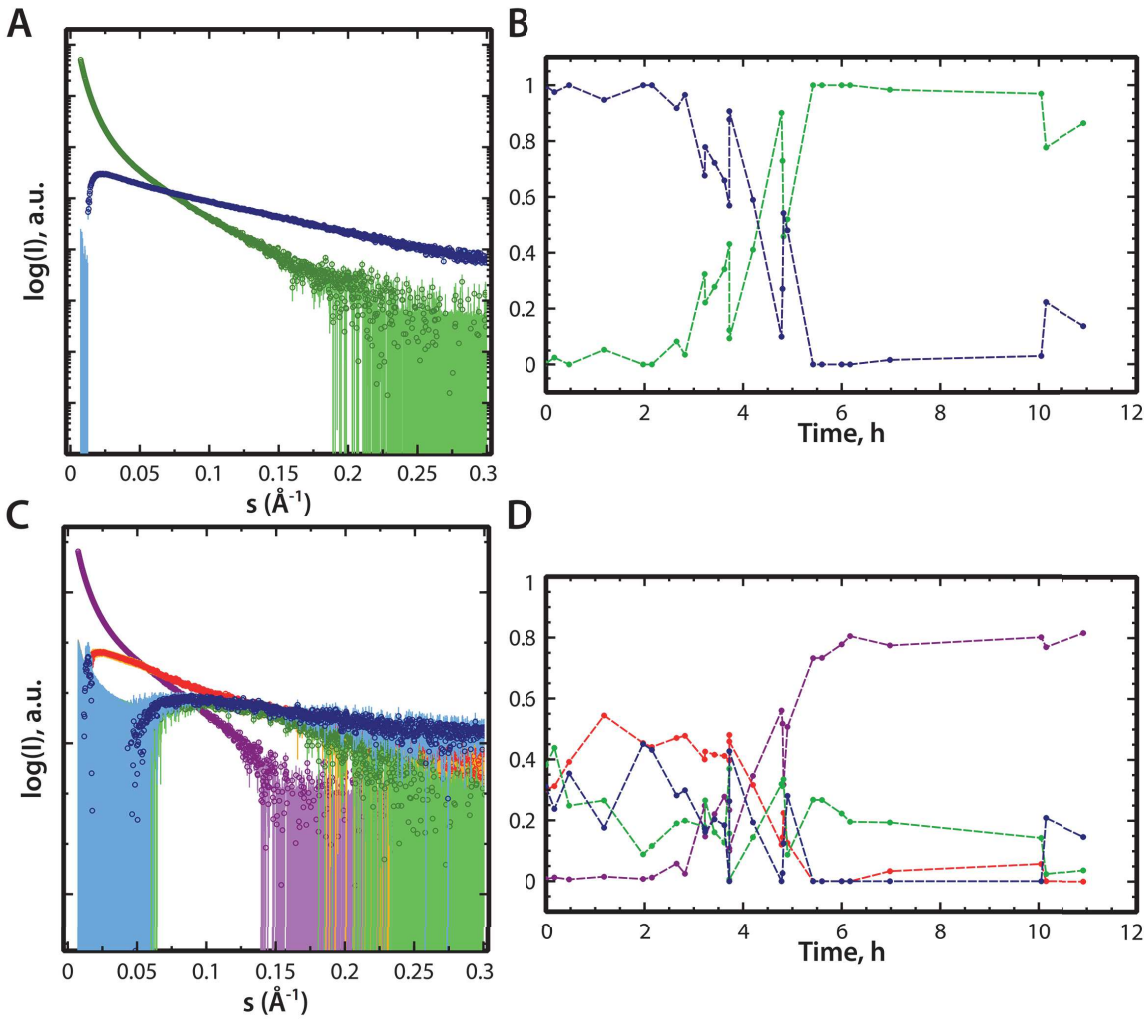


FIGURE 4.8. *Use of different number of components for the insulin data optimization, resulting in non-physical or poorly fitted results. Optimized results from the decomposition of the insulin data with 2 species using the combination AKP, with $\langle \chi_i^2 \rangle$ = 7.66, (A) Spectra profiles and (B) concentration profiles for each species. Optimized results from the decomposition of the data with 4 species using the combination of matrices AHK, with $\langle \chi_i^2 \rangle$ = 2.87. (C) Spectra profiles and (D) concentration profiles for each species. Two of the resulting spectra are not SAXS-like curves (blue and green).*

				Molecular Weight, kDa (oligomeric state)				
	R_g, Å	**D_{max}, Å**	**I(0), a.u.**	**BSA**	**Scatter**	**SAXSMoW**	**Porod (Vp/1.7)**	**Dammin**
Species 1	25.8	103.4	0.11	21.8 (3.8)	8.0 (1.4)	20.0 (3.4)	6.3 (1.1)	36.8 (6.3)
Species 2	48.5	196.0	0.22	45.7 (7.9)	23.0 (4.0)	62.7 (10.8)	35.1 (6.1)	186.3 (32.1)
Species 3	225.1	760.0	62.24	10036.0 (1730.3)	30600 (5275.9)	9423 (1624.7)	17799 (3068.9)	51401 (5807.7)

TABLE 4.2. *Structural information from the pure species of insulin derived with COSMiCS.*

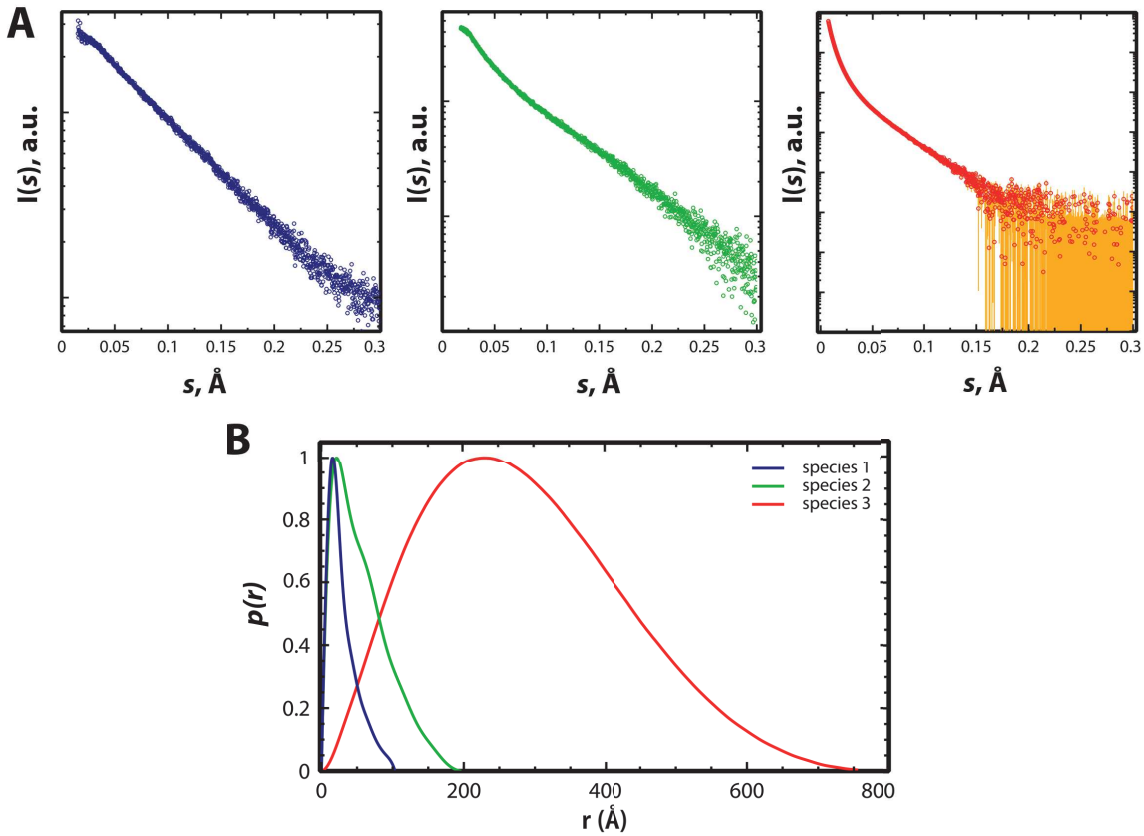


FIGURE 4.9. *(A) SAXS profiles in logarithmic scale from the decomposition of the insulin dataset using COSMiCS using the combination AH, displaying the monomer (blue), oligomer (green), and fibril (red) curves. Fits of the Ab initio reconstructions are displayed as solid lines. (B) Pairwise distribution functions derived from the individual curves for each species computed with the program GNOM [311]. The same color code as in (A) is used.*

form (Figure 4.10) in excellent agreement with the structure previously derived from a more concentrated fibrillation series [275]. The third species, which represents the repeating unit of insulin fibrils, is composed by 1730 insulin monomers with a $R_g = 225.1 \pm 0.7$ Å and a D_{max} of 760 ± 10 Å. The low-resolution structure indicates that this fibril unit consists of several intertwining protofibrils, resulting in the relatively globular and flat appearance, in accordance with previous studies [275].

4.1.6 Kinetics of the insulin fibrillation process

Figure 4.12 displays the time-dependent concentration profiles of the three species. The derived concentration profiles clearly identify the population behavior of the three species, but display spikes along the fibrillation due to the stochasticity of the fibrillation process in the individual sample wells, as discussed before. This deleterious effect could be overcome using SAXS laboratory sources whereby a single sample could be measured along the whole fibrillation process. Monomeric insulin, which is the most populated species at the beginning of the reaction, was not present in significant amounts after 3 hours of incubation. The intermediate oligomer is present during almost the complete

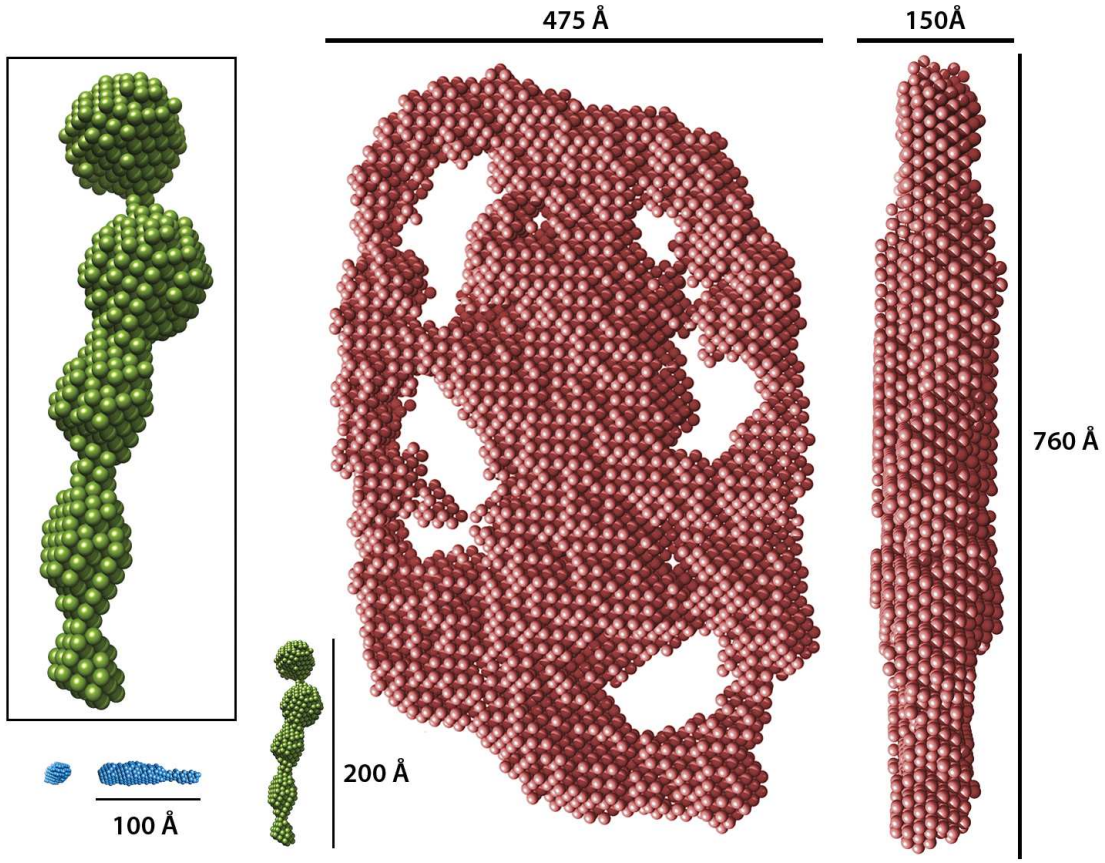


FIGURE 4.10. *Ab initio reconstructions of the three components obtained from the COSMiCS analysis of the SAXS data measured along the insulin fibrillation. Structures of the monomer (blue), oligomer (green), and the repeating unit of the fibril (red) are displayed in their relative sizes. The monomer is displayed in two orientations, one rotated 90°. The oligomer is displayed in more detail in the inset.*

observation time and becomes the major species around 4 hours, indicating that it is a relatively stable species. This species can hence not be considered as a thermodynamic nucleus, which per definition is the least stable species along the reaction coordinate. Rather, the species must be described as a structural nucleus [316], and most likely is a building block of protofibirils [275]. We have superimposed the ThT fluorescence profile from the same samples, which is sensitive to the fibrillar forms of the protein, and serves as an independent measure of the presence of the amyloidogenic fibrils. The SAXS-derived volume fractions of the large fibrillar species follows a sigmoidal growth after the lag phase that is in excellent agreement with the ThT fluorescence profile (Figures 4.12 and 4.13) and substantiates the results derived from the chemometric SAXS decomposition. Our analysis indicates that all samples measured, with the exception of the last point, contain at least two coexisting species. This observation highlights the importance of chemometric approaches whereby no a priori assumptions on the composition of the individual SAXS curves are made.

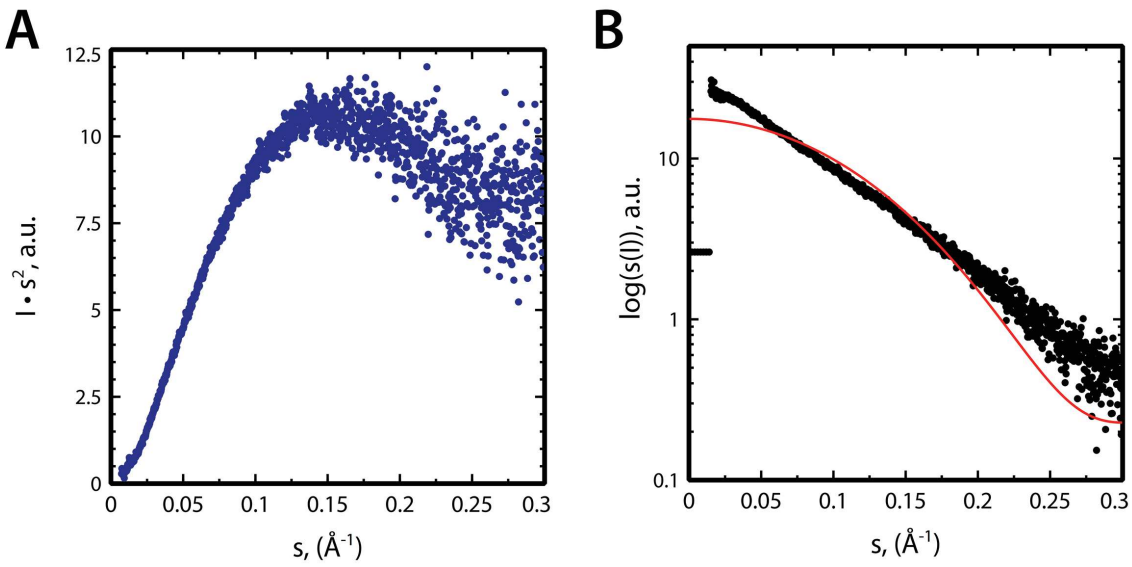


FIGURE 4.11. *Structural analysis of some of the monomeric species of insulin derived from the COSMiCS analysis (A) Kratky representation of the monomeric insulin derived from COSMiCS indicating that the isolated monomeric species is partially unfolded. (B) Fitting with CRYSOL of the folded monomeric insulin extracted from pdb 1EV6 to the curve isolated using COSMiCS for the monomeric species*

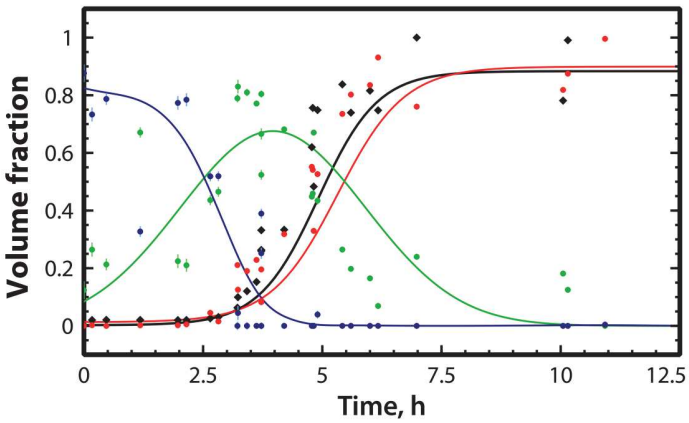


FIGURE 4.12. *Time-dependent concentration profiles derived from COSMiCS for each species: monomer (blue), oligomer (green) and fibril (red). Smoothing of the data in solid lines. The ThT fluorescence signal (black) is included to highlight its excellent correlation with the population of fibrils derived from the COSMiCS analysis.*

4.2 α-synuclein

A similar strategy than the performed for insulin was subsequently applied to data from the fibrillation of the αSN_{E46K}, associated with early-onset Parkinson's [317]. Although this mutant has been widely investigated, its fibrillation process has never been studied at structural level by SAXS. Time-resolved SAXS data were obtained following the protocols previously described [300] and detailed in section 4.4.2.

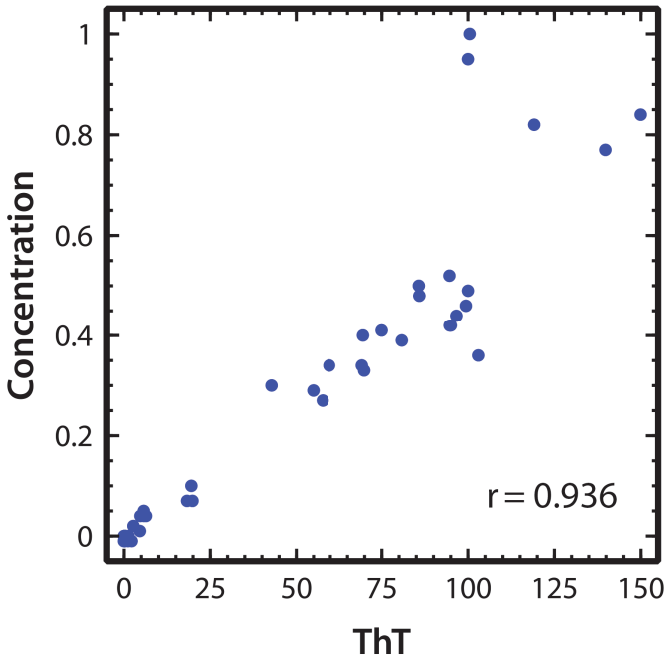


FIGURE 4.13. *Correlation between ThT signal and concentration of fibril species derived with COSMiCS for insulin.*

4.2.1 Primary α-synuclein E46K SAXS data analysis

A total of 51 SAXS curves were measured during 25 h, starting from the monomeric protein at 12 mg/ml. Two of these curves presented severe radiation damage and were discarded. The remaining curves displayed a distinct evolution in scattering intensity along the fibrillation process (Figure 4.14). The initial analysis of the raw data revealed, as expected, the formation of very large species during the fibrillation process. The average R_g and I(*0*) of the individual curves were estimated using Guinier's approximation (Figure 4.15). Interestingly, the steep increase in molecular mass occurs significantly later than the increase in average radii and the initiation of the elongation phase as indicated by ThT fluorescence. A PCA of the complete αSN_{E46K} SAXS dataset indicates that it can be safely described with three coexisting species (Figure 4.16).

4.2.2 Decomposition of α-synuclein E46K SAXS data with MCR-ALS

A MCR-ALS analysis was performed to the data in absolute scale, imposing the presence of three species in the system. The fitting with MCR-ALS yield a $\langle {\chi_i}^2 \rangle$ of 2.38, which is not a very bad adjustment, especially considering the number of curves used. However, after a closer inspection of the results it is clear that the solution is not correct, both in the spectra (Figure 4.17A) and its calculated *p(r)* (Figure 4.17B), as well as in the populations (Figure 4.17C). The pure spectrum of the smaller species (blue curve) is a non SAXS-like curve that leads to a non-physical solution in the *p(r)* with negative values. The second

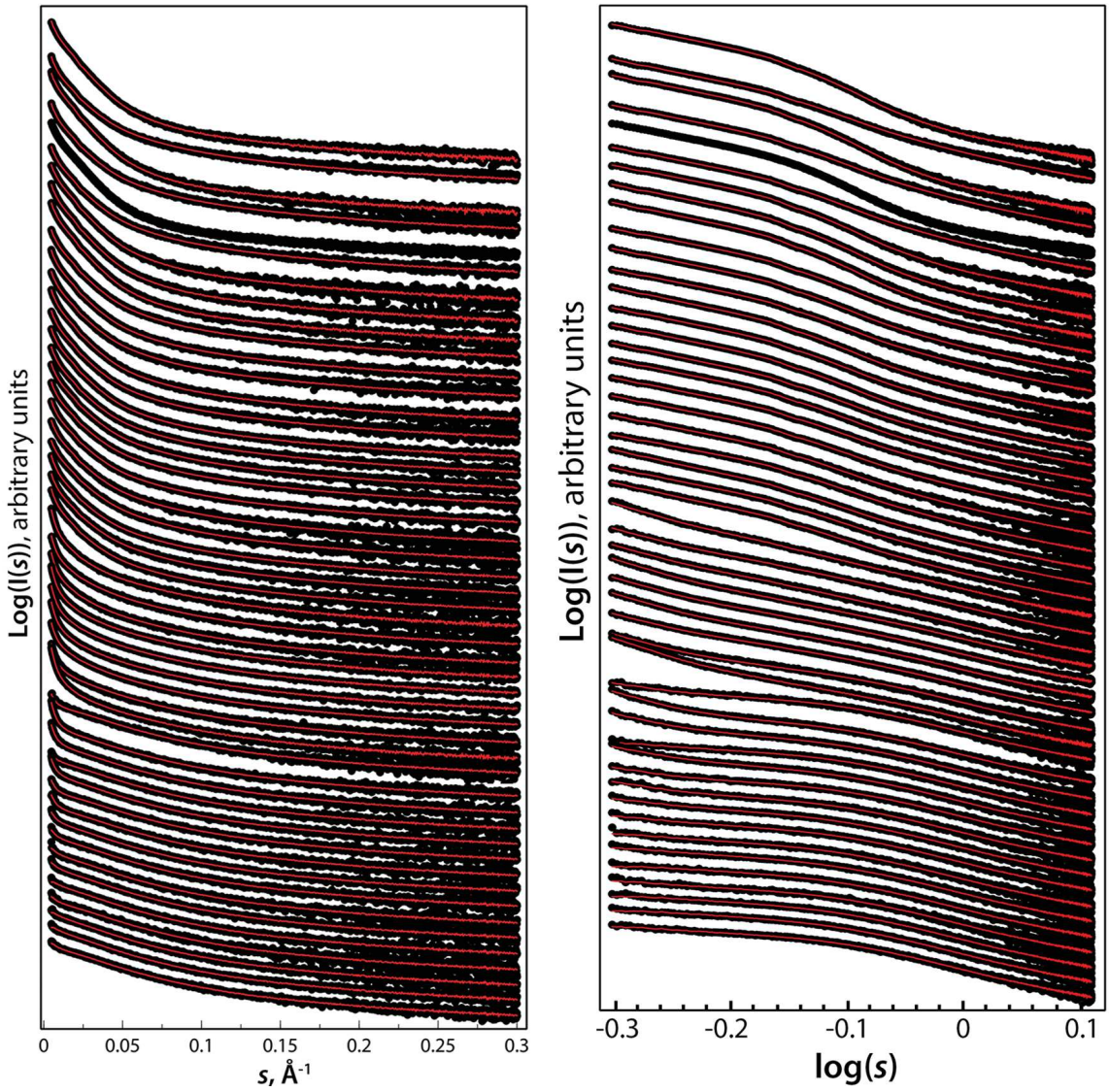


FIGURE 4.14. *SAXS profiles showing the evolution of fibrillation of αSN_{E46K} (black dots) in logarithmic scale as a function of the momentum transfer ($s = 4\pi\sin(\theta)/\lambda$ [2θ, scattering angle; λ = 1.5 Å, X-ray wavelength]). Bottom curve corresponds to the first curve (time = 0 h) and top curve corresponds to t = 25.6 h. Red lines are the COSMiCS fits obtained from the three-component mixture using Absolute and Holtzer and Kratky representations (AHK) of the SAXS data. The curve corresponding to t = 25.33 h has been identified as an outlier and is not used for the global fitting (and hence no COSMiCS fit is superposed). Curves are translated arbitrarily along the y-axis for visualization purposes. Semi-logarithmic scale (left) and log-log (right) representations of the data.*

species (green) is a big species with large uncertainties and generates a *p(r)* with an anomalous shape. Moreover, the populations are not coherent, with a very low contribution of the fibril species (red) during the whole fibrillation. This low contribution of one of the species can be an indication that a larger number of species than needed were used to explain the data during the optimization, leading to redundant solutions that could cause artifacts in the shape of the pure spectra. For this reason, we performed the analysis of the data imposing two species to describe the system. The high $\langle {\chi_i}^2 \rangle$, 7.69, indicates that two species is not enough to describe the data. PCA is just an estimation of the number

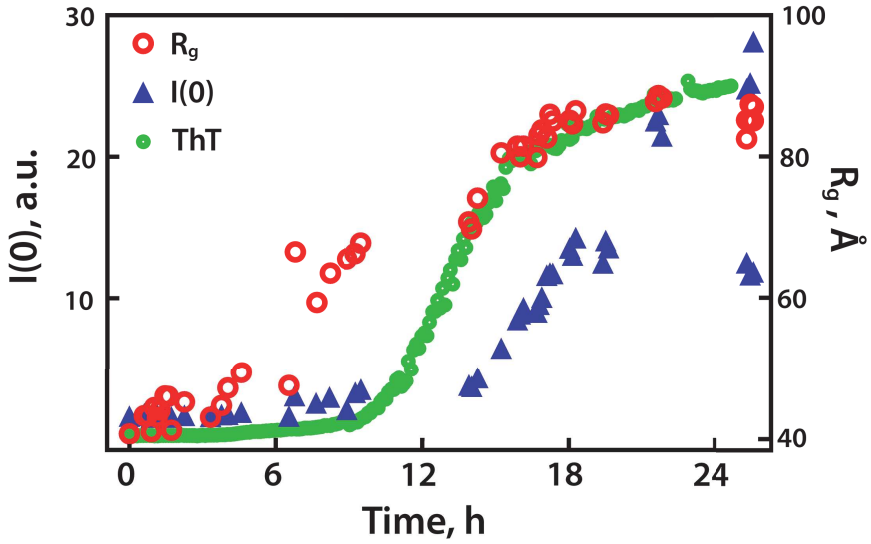


FIGURE 4.15. *Primary SAXS data analysis for αSN_{E46K}. Average radii of gyration, R_g, (hollow red circles), and forward scattering, I(0) (blue triangles) as estimated by Guinier's approximation are displayed as a function of time. ThT fluorescence values are displayed in green on arbitrary scale.*

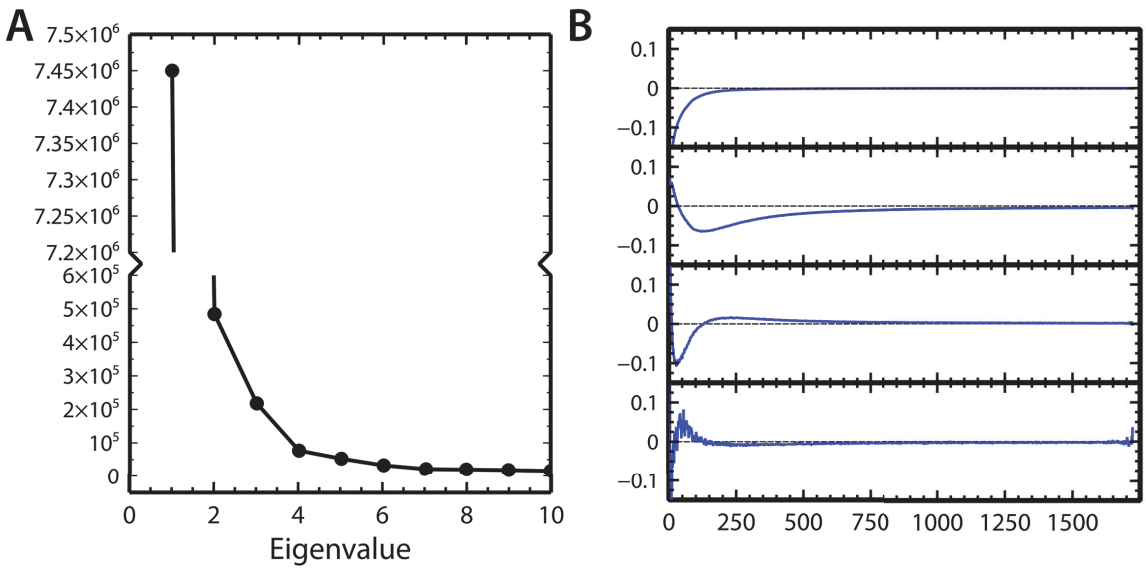


FIGURE 4.16. *Principal Component Analysis (PCA) of the complete αSN_{E46K} datasets. The analysis of the ten first eigenvalues (A) and the five first eigenvectors (B) suggests that three major species are present along the fibrillation process. Validation of the presence of three main species contributing to the SAXS datasets is performed using MCR-ALS.*

of species and is possible that the system is better explained with one species more. A MCR-ALS analysis with four species decreases the $\langle \chi_i^2 \rangle$ to 1.90, but the SAXS curves of the pure components were not SAXS-like, and yielded a *p(r)* without physical meaning (Figure 4.18A and 4.18C). Moreover three out of the four components showed a very similar volume fraction, and even negative values.

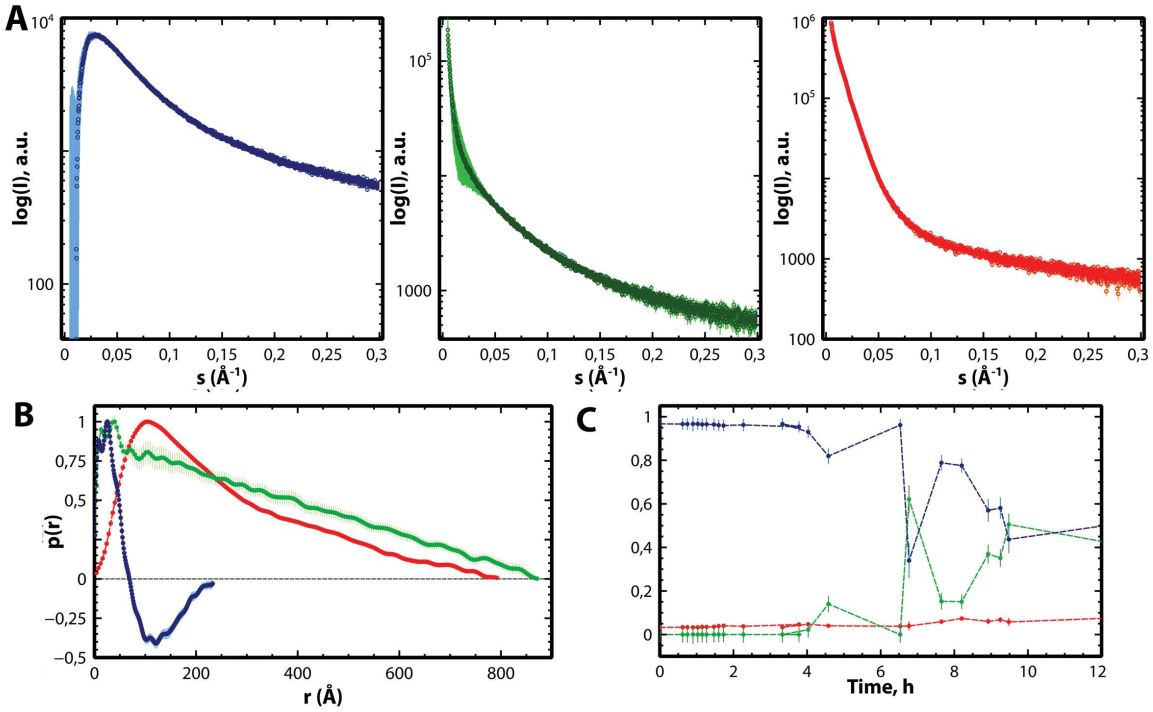


FIGURE 4.17. *Optimized results from the decomposition of the αSN_{E46K} data with MCR-ALS using only the absolute scale data representation. (A) Pure spectra. (B) p(r) from these pure spectra. (C) Populations derived from MCR-ALS.*

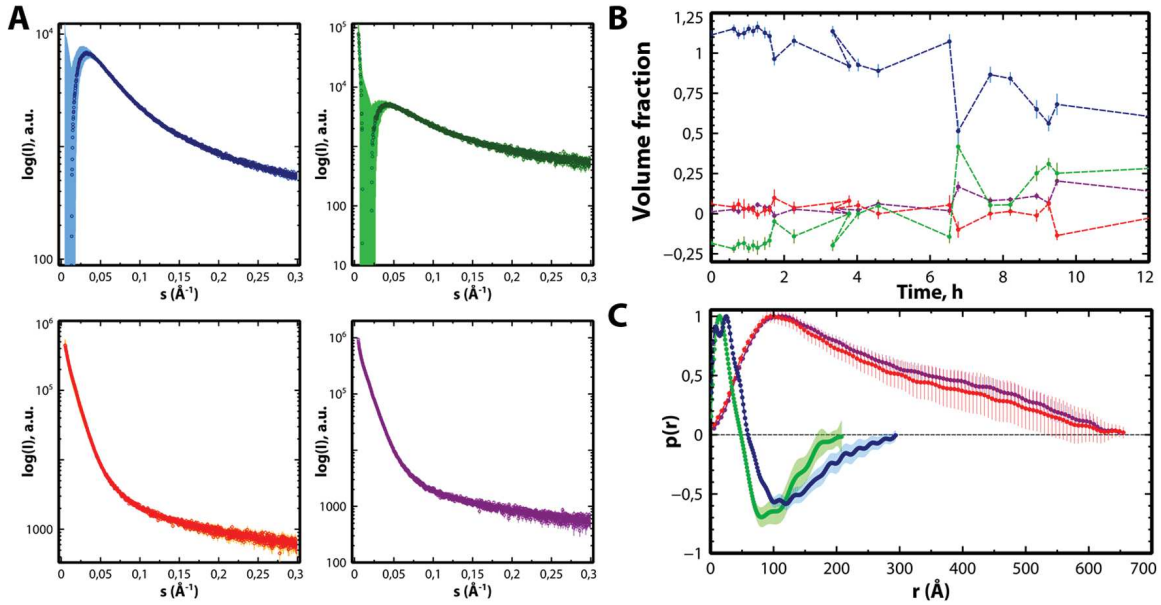


FIGURE 4.18. *(A) SAXS profiles in semi-logarithmic scale from the decomposition of the αSN_{E46K} dataset using MCR-ALS using 4 species. (B) Time-dependent concentration profiles derived from MCR-ALS for each species with the same color code. (C) Pairwise distribution functions derived from the individual curves*

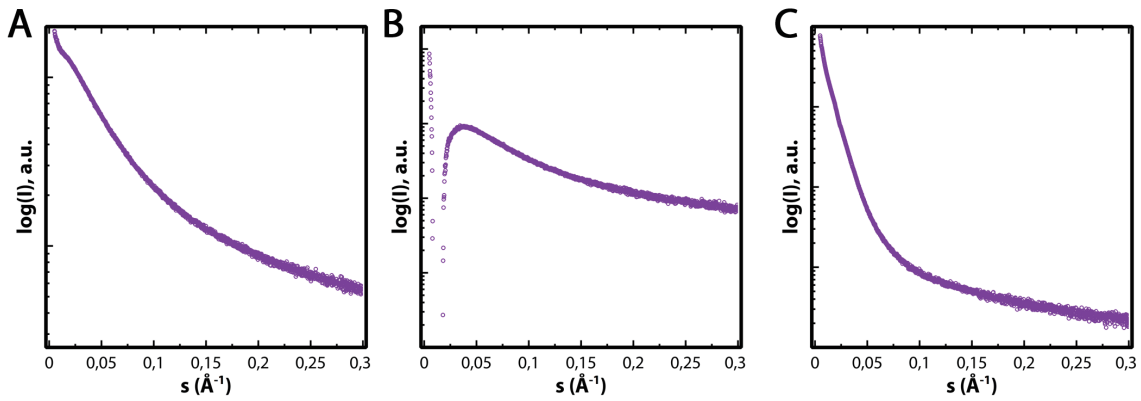


FIGURE 4.19. *Results of the optimized SAXS curves for aSNE46K obtained using MCR-ALS 2.0.*

4.2.3 Decomposition with MCR-ALS using weighted data

In a similar way we did with the insulin, we have performed a decomposition using weighted data using the new version of MCR-ALS. The results are showed in the Figure 4.19. Like in the case of insulin, severe artifacts were observed for the curves, which demonstrate that weighting the experimental intensity values by their associated errors along the optimization is not enough to allow the correct decomposition of the data.

4.2.4 COSMiCS analysis of α-synuclein E46K SAXS data

The αSN_{E46K} dataset was subsequently successfully decomposed with COSMiCS using the combinations of SAXS data representations (Table 4.3 and Figure 4.20). As occurred for the case of insulin, a systematic improvement in the quality of the fit was observed when multiple data representations were introduced, with $\langle {\chi_i}^2 \rangle$ decreasing from 2.62 to a range of 1.16-1.33. This systematic amelioration of the data description was also observed in the range of individual ${\chi_i}^2$ obtained among the 49 curves (Table 4.3). However, not all data combinations guided the decomposition equivalently. Again, the inclusion of Porod's representation (P) only modestly decreased the $\langle {\chi_i}^2 \rangle$ compared with inclusion of Holtzer's or Kratky's representations (AH or AK). Four of the solutions presented very similar $\langle {\chi_i}^2 \rangle$, between 1.16 and 1.18. Although quantitatively equivalent, a closer inspection of the decomposed curves showed that representations AHK and AHKP provideed solutions yielding non-physical *p(r)* functions with negative values. Therefore, the AKP solution, with a $\langle {\chi_i}^2 \rangle$ = 1.16, was used for the subsequent analyses.

To confirm the PCA analysis we repeated the complete COSMiCS analysis using two and four species (Figure 4.21A and 4.21B). When fitting with two species the $\langle {\chi_i}^2 \rangle$ significantly increased to $\langle {\chi_i}^2 \rangle$ = 7.66, and the optimized spectra from the analysis with four species were not physically meaningful (4.21C and 4.21D).

The inspection of the level of agreement of the individual curves indicated that the vast majority of $\langle {\chi_i}^2 \rangle$ values were around 1.0 (Figure 4.21A), indicating that our optimization protocol was not overfitting the data. However, curve 45 (measurement at 25.33 h)

Representations included					α-SN$_{E46K}$[a]			
					Complete Dataset		w/o outlier[b]	
Code	Absolute I(s)	Holtzer I(s)*s	Kratky I(s)*s^2	Porod I(s)*s^4	$\langle\chi_i^2\rangle$	χ_i^2 Range	$\langle\chi_i^2\rangle$	χ_i^2 Range
A	+				2.62	9.46-0.54	2.38	8.96-0.73
AH	+	+			1.22	3.42-0.54	1.08	2.32-0.55
AK	+		+		1.18	3.69-0.53	1.05	2.24-0.53
AP	+			+	1.33	4.82-0.57	1.18	3.05-0.58
AHK	+	+	+		1.16	3.55-0.56	1.04[c]	2.18-0.53
AHP	+	+		+	1.24	4.31-0.54	1.09	2.71-0.54
AKP	+		+	+	1.16[c]	3.98-0.53	1.05	2.08-0.53
AHKP	+	+	+	+	1.16	3.85-0.53	1.03	2.11-0.52

TABLE 4.3. *Fitting of the αSN$_{E46K}$ SAXS datasets with COSMiCS using different combinations of data matrices.*
a- Analysis performed using three species.
b- COSMiCS analysis after extraction of the curve corresponding to t = 25.22 h.
b- Optimal solutions used for the structural analysis.

was not properly described by the model and presented a relatively large χ_i^2 (3.98). In fact, this curve presented the largest $\langle\chi_i^2\rangle$ in all data combinations. After removing this curve the complete COSMiCS decomposition was repeated, yielding a smaller $\langle\chi_i^2\rangle$ (1.03-1.08) for various data combinations (Table 4.3). Importantly, the improvement observed after discarding a single curve did not simply originate from the elimination of an outlier but corresponded to a systematic improvement of the individual $\langle\chi_i^2\rangle$ for the vast majority of the curves of the dataset (Figure 4.21B). A further refinement by removing an additional potential outlier curve was tested but no systematic improvement in the fit was observed (Figure 4.21C). These observations underline the robustness of the final solution and the ability of COSMiCS to detect outlier SAXS curves. The final fits from the COSMiCS analysis using the AHK combination ($\langle\chi_i^2\rangle$ =1.04) are displayed in Figure 4.14, and the solution was subsequently analyzed in terms of structures and kinetics.

4.2.5 Structural analysis of the components of α-synuclein

As it can be seen from the COSMiCS-derived scattering curves and their *p(r)* functions (Figure 4.22), two of the species are very large while the first species is a low MW particle. We estimated a MW of the first species of 13 kDa, in agreement with a monomeric state of the protein (14.6 kDa). Additionally the Kratky representation (not shown) of this species, along with the skewed *p(r)* function and the relatively large values for R_g (47.1 ± 0.72 Å) and D_{max} (209 ± 5 Å), indicates the disordered nature of αSN$_{E46K}$ [58]. The

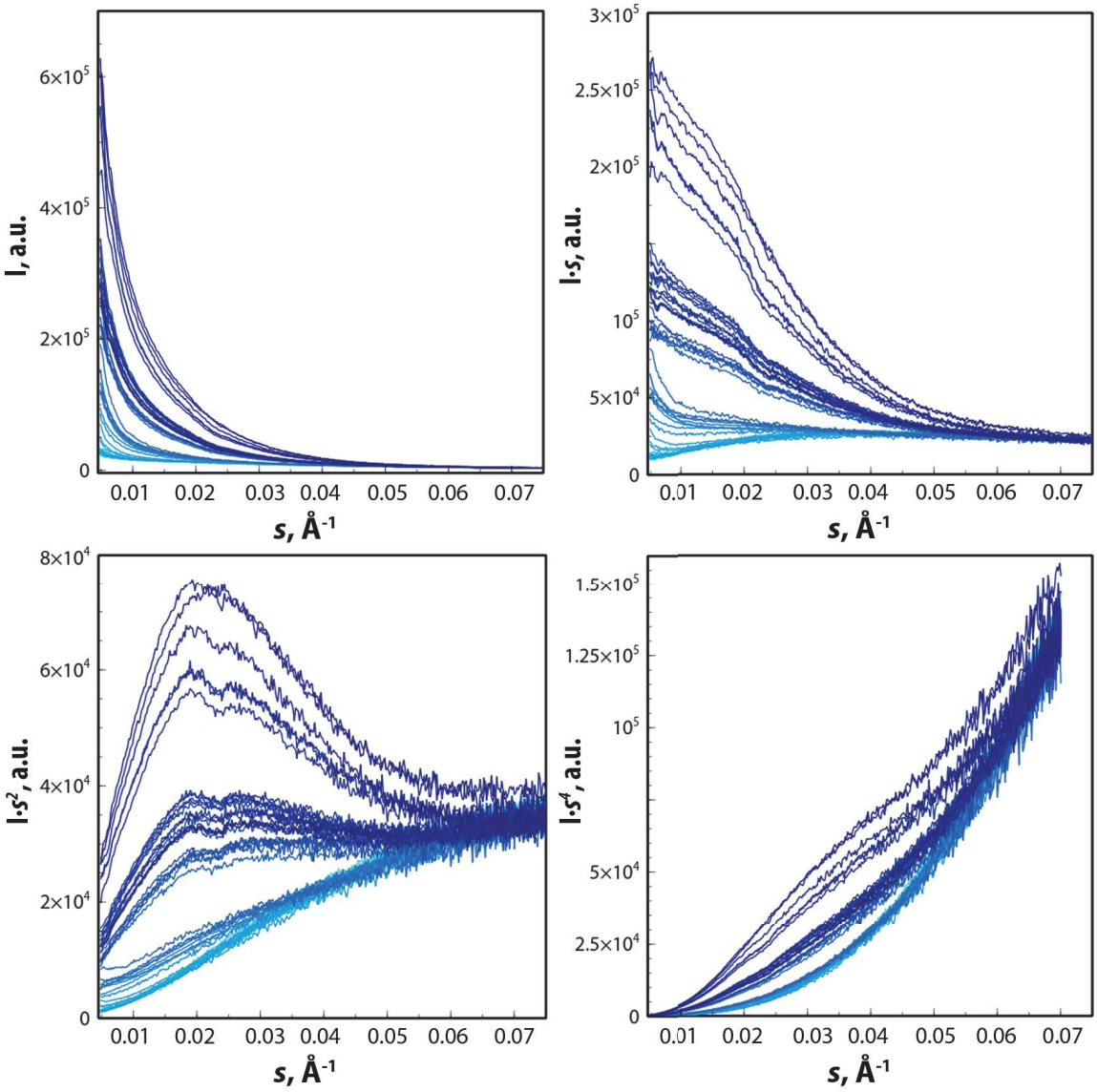


FIGURE 4.20. *Representations of the αSN_{E46K} SAXS data measured along the fibrillation. (A) Absolute values, I(s). (B) Holtzer, I(s)·s. (C) Kratky, I(s)·s^2. (D) Porod, I(s)·s^4. Different features are observed in the momentum transfer range displayed along the fibrillation process represented with a blue color scale, from light blue (t=0) to dark blue (t=25.6 h). A significant change can be observed in the Kratky plot, where it is possible to see how the curve change from a typical shape of an IDP curve to a curve of a more compacted protein (the fibril), with the characteristic bell-shape form.*

disordered nature of the monomeric species was substantiated by the ensemble optimization method (EOM) analysis of the SAXS profile [61] (see section 1.3.2.2 for details). The subensemble of conformations that collectively describe the SAXS curve displayed a broad range of R_g values indicating the large degree of flexibility of the protein in solution (Figure 4.23).

The second oligomeric species is large (≈40 protomers) with an R_g of 282.4 ± 4.3 Å and a D_{max} of 960±10 Å (i.e. at the resolution limit of our measurements), approaching values obtained for the final fibril species (D_{max} = 920 ± 10 Å and R_g = 256.8 ± 1.9 Å). However, when comparing the mass and overall dimensions of these two species (Table

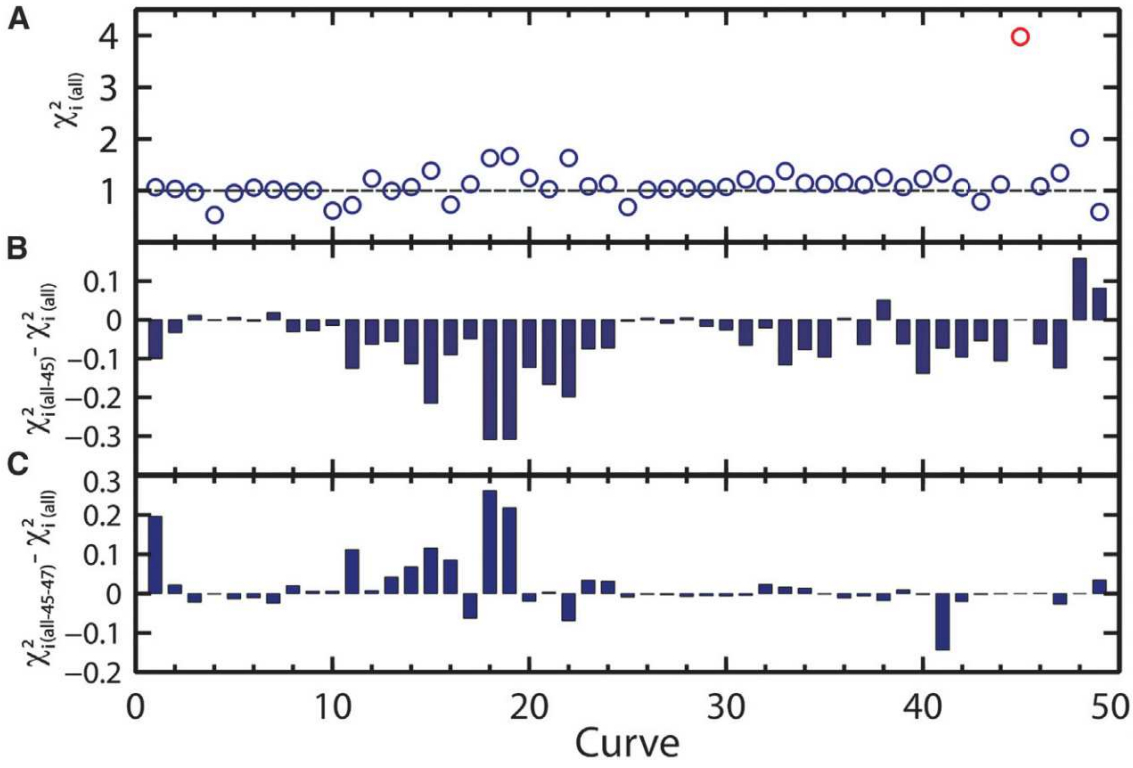


FIGURE 4.21. *Assessment of Outlier Curves in the αSN_{E46K} SAXS Dataset. (A) Individual χ^2 values obtained from the COSMiCS analysis of the αSN_{E46K} complete dataset using the AKP combination of matrices. One single value appears as a potential outlier (curve 45, marked in red). (B) The variation in the χ^2 values of the individual curves when extracting curve 45 (corresponding to a time of 25.33 h) from the analysis using the AHK combination. A systematic improvement for the vast majority of curves of the dataset is observed upon extraction of the curve from the global fitting, and it is hence concluded that curve 45 is an outlier. (C) Difference in individual χ^2 values derived from the COSMiCS analysis with the AHK combination using 48 curves of the αSN_{E46K} dataset and the AKP combination using 47 curves in the dataset after removing the SAXS curve corresponding to t = 25.6 h from the analysis. Either no effect or an increase in the χ^2 values is observed. It can be concluded that the extraction of this second curve does not improve the derived model and it is therefore not justifiable.*

4.4) the mass density of the intermediate species is much lower than that of the fibrils. This is in agreement with the observation that the average R_g of the mixture increases significantly earlier than the average mass (Figure 4.15), and suggests that the intermediate species is a large and disordered oligomer. It is evident also from the *p(r)* functions (Figure 4.22C) that the two species present distinct overall shapes. While the intermediate species is represented as an overall globular shape, the final species presents the typical elongated fibrillar shape [232, 298]. The COSMiCS curves for the large oligomer and the fibril were used to derive low-resolution structures. Whereas the fibril repetitive unit is a large and elongated particle (Figure 4.24), attempts to derive a structure for the oligomer were unsuccessful. Indeed, this observation is in line with the disordered nature of this species that precludes the determination of its *ab initio* structure.

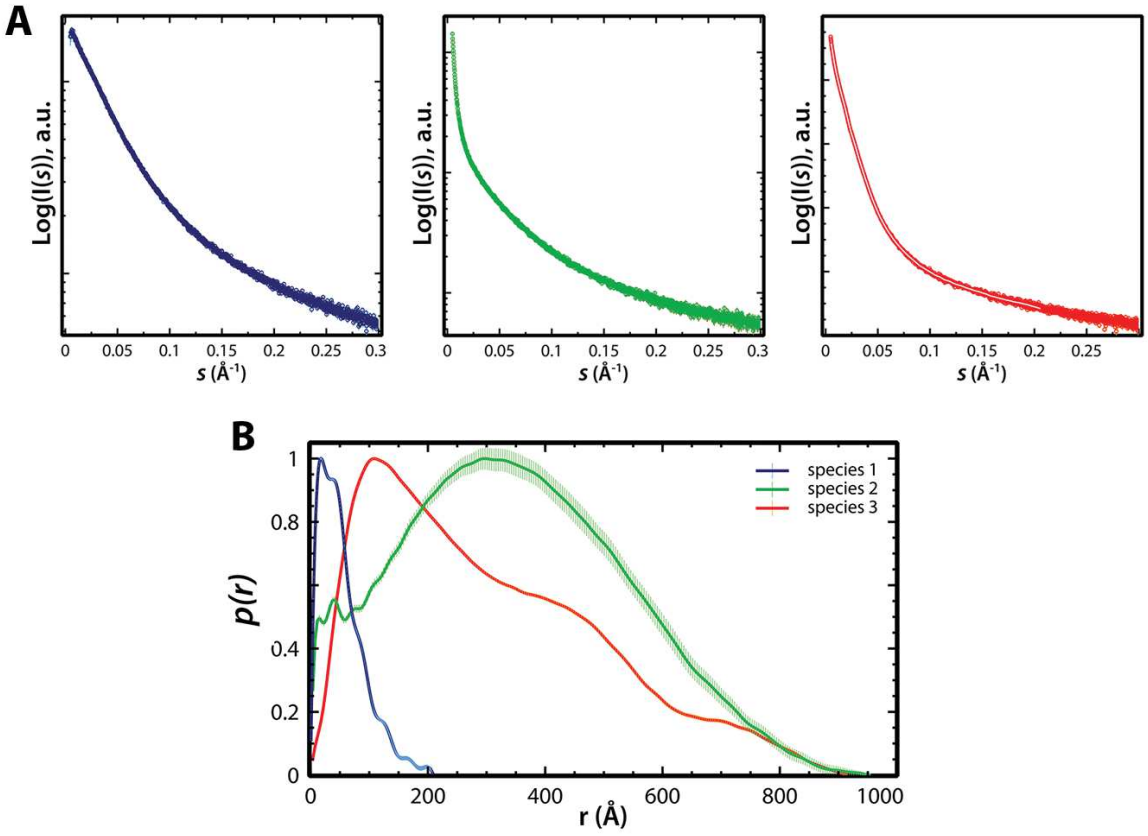


FIGURE 4.22. *COSMiCS Analysis of αSN_{E46K} Fibrillation with COSMiCS using AHK combination of matrices. (A) Decomposed SAXS profiles for the monomer (blue), oligomer (green), and fibril (red) species. (B) p(r) functions of the three species computed from the SAXS profiles of (A) using GNOM [311]*

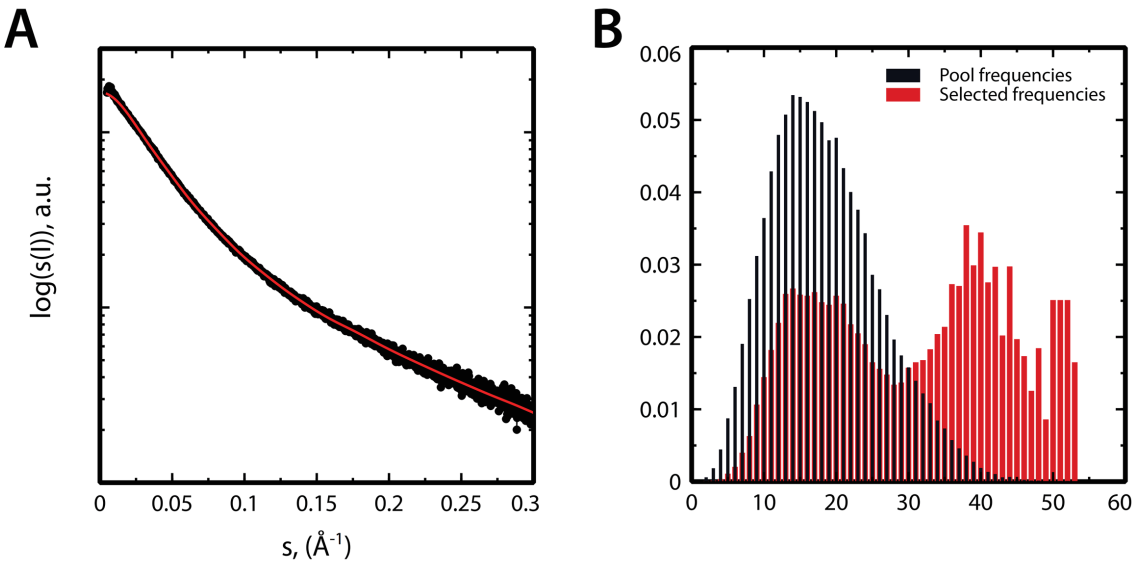


FIGURE 4.23. *A) EOM fitting (red curve) of the αSN_{E46K} curve isolated with COSMiCS (black dots), with $\chi^2 = 1.12$. (B) The distributions of radii of gyrations for the pool of αSN_{E46K} conformations (black)) and the EOM selected ones (red).*

				Molecular Weight, kDa (oligomeric state)				
	R_g, Å	D_{max}, Å	I(0), a.u.	BSA	Scatter	SAXSMoW	Porod (Vp/1.7)	Dammin
Species 1	47.1	209.5	2.56	47.6 (41.0)	13.0 (0.9)	52.3 (3.6)	29.3 (2.0)	N.D.
Species 2	281.8	960.0	52.72	594.3 (41.0)	630.0 (43.4)	493 (34.0)	11146 (768.7)	N.D.
Species 3	51.7	920.0	111.22	2565.5 (176.9)	1470 (101.4)	1781.7 (122.9)	6342 (437.4)	20837 (1437.0)

TABLE 4.4. *Structural information from the pure species of αSN_{E46K} derived from COSMiCS.*

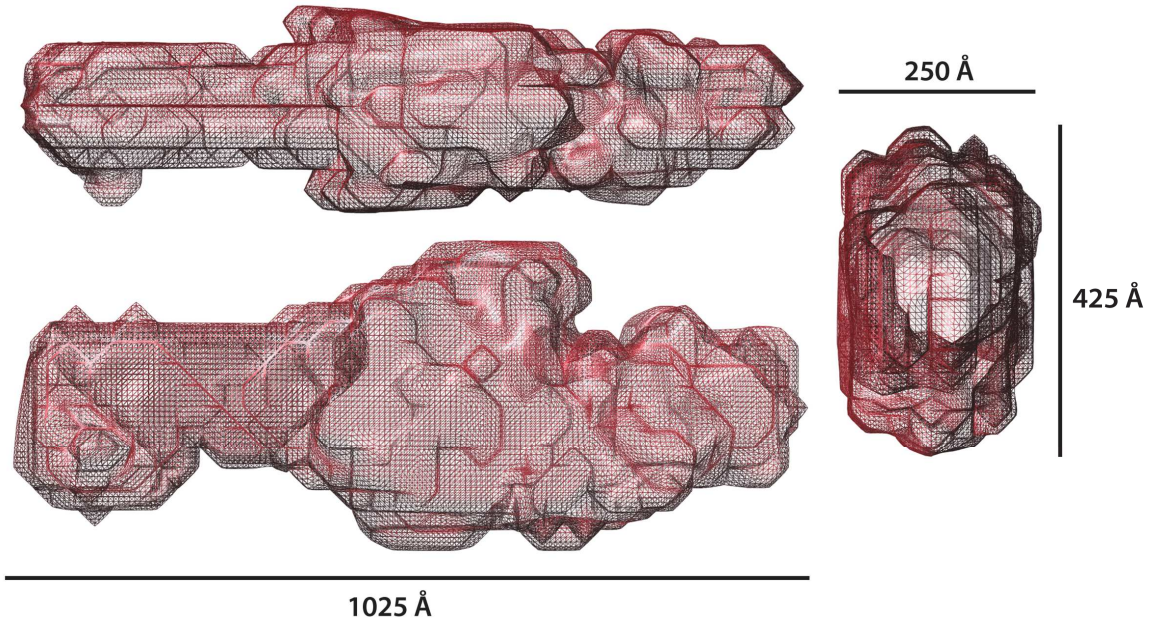


FIGURE 4.24. *Three orientations of the ab initio structure of the fibril repeating unit of αSN_{E46K} determined from the decomposed curve with COSMiCS. Average from 20 refined models computed with the program DAMMIN [33].*

4.2.6 Kinetics of the α-synuclein fibrillation process

The distinct structural characteristics of both forms are corroborated when observing their time-dependent evolution, which is plotted together with the ThT fluorescence curves in Figures 4.25 and 4.26. Clearly, the evolution of the ThT signal coincides with the occurrence of the fibril-shaped species, whereas the second large species is not ThT-active. Both large species coexist after the lag phase, but the decrease of the second species at the final steps of the aggregation suggests a transformation, through an unknown mechanism (see discussion, section 4.3), from the large disordered aggregates to amyloidogenic fibrils. Interestingly, the monomeric form is present throughout the whole experiment, suggesting

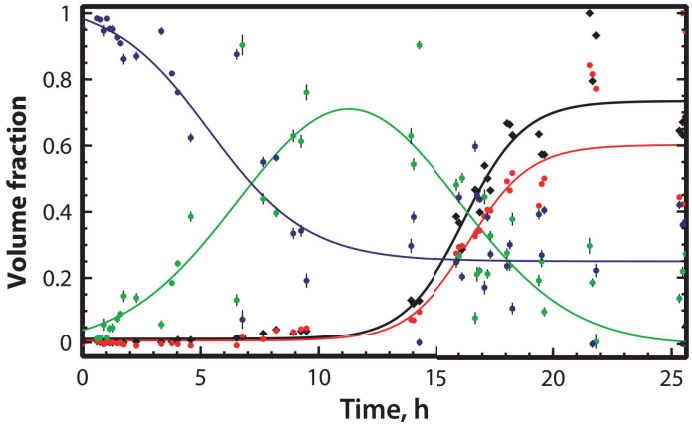


FIGURE 4.25. *Concentration profiles for the monomer (blue), oligomer (green), and fibril (red) species of αSN_{E46K} derived from COSMiCS using AHK combination. ThT fluorescence signal superimposed (solid black line). The ThT profile is in excellent agreement with the population of fibrils.*

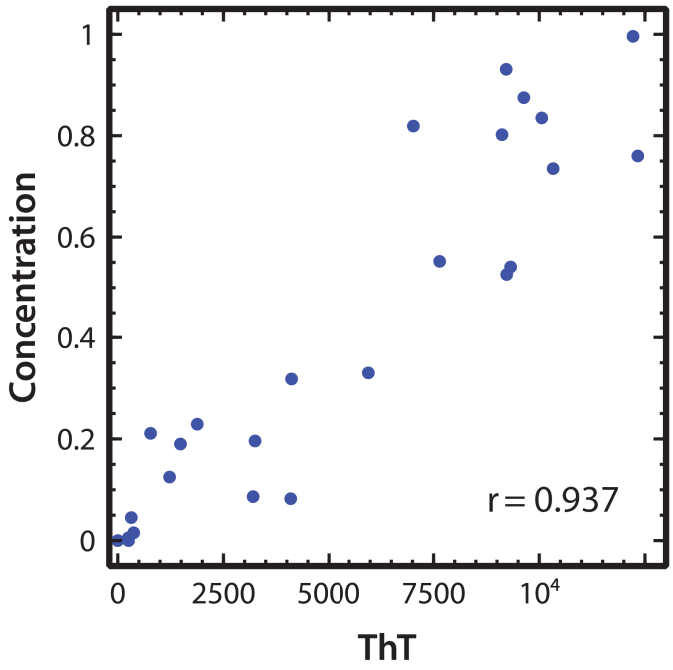


FIGURE 4.26. *Correlation between ThT signal and concentration of fibril species derived with COSMiCS for αSN_{E46K}. The excellent correlation observed substantiates the COSMiCS decompositions.*

that the disordered aggregate has to disassemble into monomers before forming amyloidogenic fibrils.

Figure 4.27 shows the single ThT curves of all the wells where the experiment was performed. As we mentioned before, the setup of the experiment produces a high variability between the samples due to the low reproducibility of the fibrillation process. This high stochasticity is very prominent in the case of αSN.

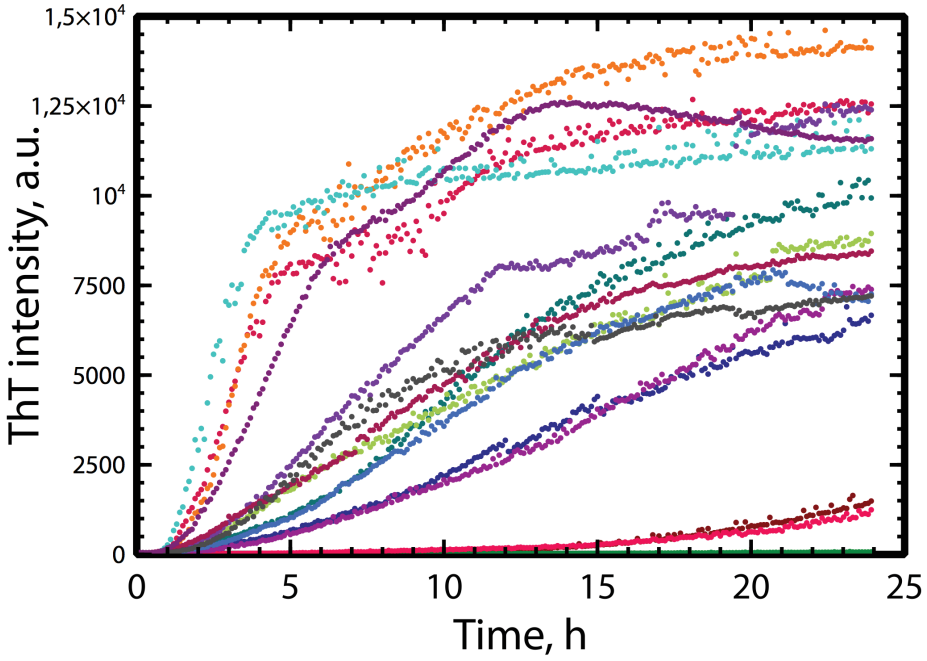


FIGURE 4.27. *ThT curves of the individual wells from which the samples were withdrawn. Each fibrillation shows different kinetics, with a different long phase, elongation phase and maximum intensity at the plateau. All wells start in the same conditions (12 mg/ml).*

4.3 Discussion

The interest in structural studies of molecular conversions, functional and structural heterogeneity, and time-evolving processes is significant but highly challenging. SAXS is an extremely well-suited technique to address the characterization of such complex mixed systems, but robust approaches are necessary to address the data decomposition process. Here, we present a method for the chemometric decomposition of multiple SAXS curves measured along a structurally developing process. This strategy, implemented as an extension of the popular MCR-ALS method [281, 284], has been applied to fibrillating proteins but could be extended to other macromolecular systems with tuneable equilibrium phenomena such as the study of protein folding, transient interactions, intermediate structural states, viral capsid formation, and supramolecular assemblies.

Initial attempts to analyze SAXS data measured along insulin and αSN_{E46K} fibrillation with MCR-ALS were unsuccessful, likely because the algorithm is trapped in a local minimum, yielding acceptable fits but non-physical solutions. The classical approach to resolve this ambiguity-related problem is to include orthogonal datasets [283, 318]. However, the simultaneous measurement of complementary data is not available in present SAXS beamlines. To overcome the ambiguity problem, we simultaneously analyze multiple representations of the same SAXS data. The empowerment of the decomposition when using multiple data representations reflects one of the central aspects of solution scattering data. The scattering curve arises from pairwise distances at both short and longer scales,

thereby probing the shape of solutes, covering several orders of magnitude, from nanometer to micrometer sizes. In the case of fibril development this phenomenon is fully exploited as starting particles are of nanometer size, while fibrils are several micrometers in length. During the fibrillation process the structural features, which are coded in the scattering curves, dramatically change in a time-dependent manner. However, at different time periods of the fibrillation reaction these changes occur in different parts of the momentum transfer range measured, and can be highlighted depending on the representation of the data used.

The different sensitivity of SAXS data representations to structural features can be understood when considering their mathematical nature. Absolute scale curves have a large dynamic range and the most important intensity variations occur at the smallest angles, linked to the lowest-resolution structural information. In the other representations the intensity is multiplied by momentum transfer with increasing power functions: 1, 2, and 4 for Holtzer, Kratky, and Porod, respectively. This successively decreases the emphasis on intensities at low scattering angles (low-resolution) and increases the emphasis at high scattering angles (high-resolution). Therefore, the combination of the absolute scale, which enhances the low-resolution part of the SAXS curve, with the other representations, which enhance the high-resolution part of the SAXS profile, facilitates decomposition. Surprisingly, Porod's representation, in the examples tested, does not increase (or even decrease) the decomposition power of COSMiCS. Porod's law is valid for smooth interfaces between solutes and solvent. In the case of fibrillation, it is known that the fibril interface is disordered and poorly defined [201], and presents high-entropy solvent around the interface [319, 320]. This observation does not exclude, however, that Porod's representation could be an important one for the decomposition of other types of data.

A few examples from the present analysis are included here for clarity (Figure 4.28). The first example is the conversion of the intrinsically disordered protein, αSN_{E46K}, into a large and more ordered intermediate that finally evolves toward elongated fibrils. The Kratky fingerprint of an intrinsically disordered protein is very characteristic with a lack of low-angle features, and a steadily rising profile at high angles [58]. These features contrast with those found in ordered systems where a pronounced peak is found at intermediate momentum transfer ranges. Therefore, the Kratky representation is extremely sensitive to the initial conversion from a disordered to a more ordered species. In contrast, the Holtzer representation is highly sensitive to overall changes in mass, which is very significant at later time points in the fibrillation process. Insulin, in contrast to αSN_{E46K}, fibrillates from a (partially) folded-like species. Here, however, the subsequently formed species is elongated with a very distinct scattering curve that is easily discriminated from the starting species. The oligomeric intermediate is subsequently transformed into a much larger mature fibril. In this second transition, the difference in size, which induces a strong differentiation in the initial part of the SAXS curve and in the peak position in Kratky representation, facilitates their discrimination. Hence, simultaneous fitting of multiple data representations used in COSMiCS exploits complementary features that appear at different time points, enhancing the capacity of the chemometric approach to discriminate within

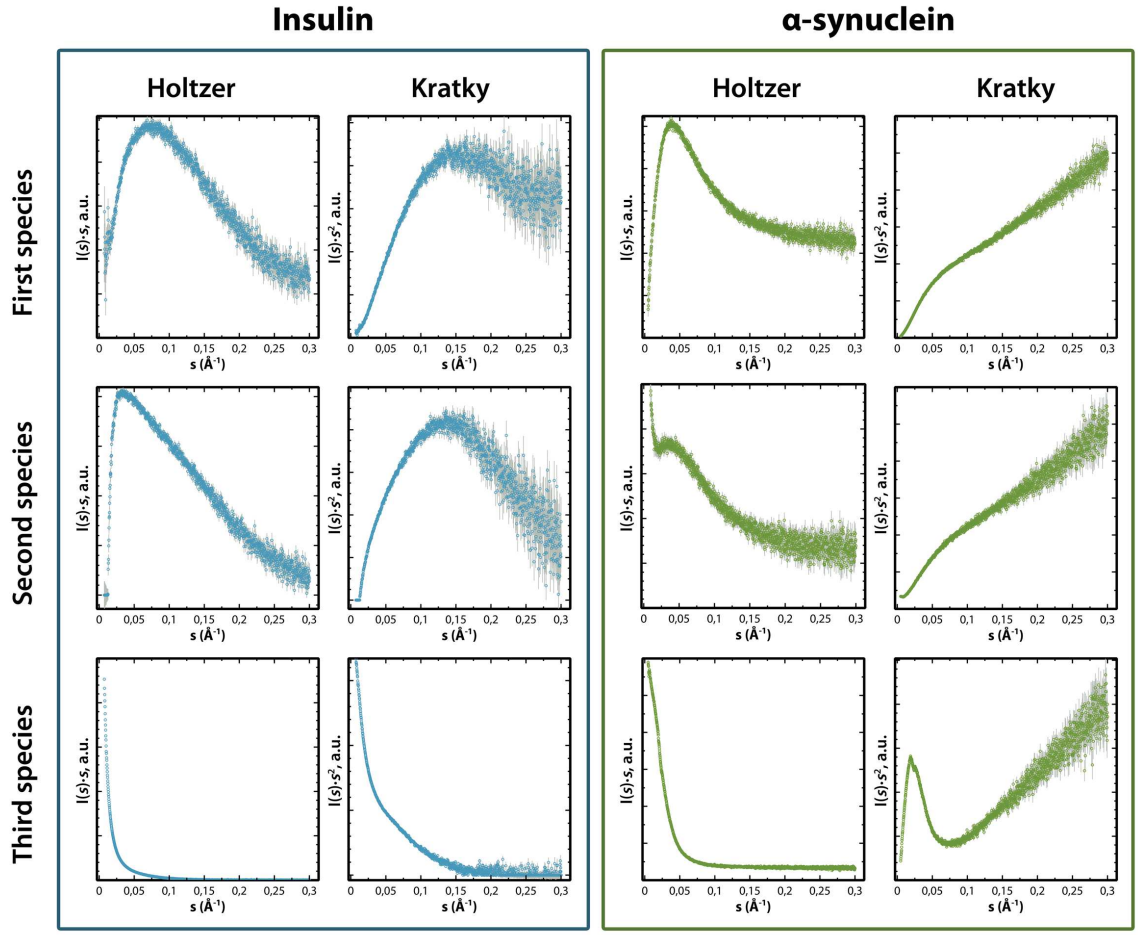


FIGURE 4.28. *Holtzer and Kratky representations of the decomposed species from COSMiCS for insulin (blue) and αSN_{E46K} (green)*

the vast space of potential solutions.

It is worth noting that no specific shapes for the resolved scattering profiles were imposed as a constraint throughout the optimization procedure. We have observed that when multiple SAXS data representations were considered in the same analysis, the overall fit to the data is better and the possibility to obtain physically meaningful solutions increases. The presence of these local minima solutions with unreasonable SAXS profiles highlights the importance of identifying and discarding them. Here, all wrong solutions present strong artifacts and large uncertainties that identify them as unphysical. This observation points toward the need to find a proper mathematical description of SAXS curves to allow for the introduction of a new constraint to further decrease solution ambiguity. Some attempts to elucidate these problems are explained in the discussion section.

Our approach represents a great advantage compared with other techniques that probe a single species. This is exemplified for the case of αSN_{E46K}. In the decomposition of the SAXS data of this familial mutant, we reveal that two very large aggregated species coexist for a long period of time. One of these species is ThT-inactive and would remain invisible when using traditional fluorescence experiments, thereby dramatically biasing the

interpretation of a protein fibrillation study solely based on ThT fluorescence [321]. Importantly, by applying our approach we uncover the existence of an intermediate species of a structural nature that has hitherto not been described in the context of fibrillation kinetics. This oligomer has a low density of mass, is predominantly disordered, and is of a reversible nature.

4.4 Materials and Methods

4.4.1 Insulin

4.4.1.1 Insulin sample preparation and fluorescence measurements

Human zinc insulin was obtained from Novo Nordisk. Zinc content was 0.37% (w/w) corresponding to approximately two Zn^{2+} ions per insulin hexamer. ThT was purchased as the chloride salt from Sigma-Aldrich. ThT was recrystallized three times in demineralized water before use. All other chemicals were of analytical grade. For ThT fluorescence assays, a Polarstar Optima platereader from BMG Labtechnologies was used with 96-well, black, polystyrene, nonsterile plates with optical bottom from Nalge Nunc International. The wells were covered with Polyolefin non-sterile sealing tape from Nalge Nunc International, and bottom/bottom measurements were performed. ThT fluorescence measurements were performed with λ_{ex} = 440 nm (10 nm bandpass) and λ_{em} = 480 nm (12 nm bandpass). A pellet of insulin was dissolved at 2.5 mg/ml insulin in 20% acetic acid (pH 2.0) with 0.1 M NaCl and 20 mM ThT. Afterwards 100 ml of the solution was transferred to each well. The plate was placed in the platereader and ThT fluorescence measurements were conducted at 45 C without shaking. The measurement of the fluorescence intensity was performed every 300 s. The platereader was paused at appropriate time intervals and a sample of 80 ml was withdrawn from a well.

4.4.1.2 COSMiCS analysis of insulin dataset

All SAXS representations used for the MCR-ALS optimization were used in a momentum transfer range of $0.0074 < s < 0.3$ Å^{-1}. Scattering curves numbers 3 (28 min), 4 (1 h 11 min), and 28 (10 h 56 min) were selected as starting points for the optimization. The number of maximum iterations was set to 50 and the convergence criteria were set to 0.1. ALS optimization is performed under the standard constraints for fibrillating system, *non-negativity* (both for spectra and concentration profiles) and *closure* [282] for concentration profiles.

4.4.1.3 *Ab initio* modeling of insulin components

Ab Initio structures of the oligomer and fibril of insulin were obtained using the program DAMMIN [33]. The program employs a simulated annealing protocol to search for a complex bead model minimizing the discrepancy between the experimental and calculated curves at low resolution (up to s of about 0.15 Å^{-1}). The search volume, evaluated

with the program BODIES [50], for the fibril repeat was an ellipsoid with half-axes of 800, 500, and 150 Å using 31,985 spheres. Individual jobs were loaded, and 20 independent models were averaged using the program DAMAVER [15] and filtered with DAMFILT [50]. The oligomer was calculated inside a sphere with a diameter of 98 Å, obtaining a final averaged and filtered structure from 18 individual models. The structure of the monomer was obtained using DAMMIF [35], starting from 20 arbitrary initial models to obtain the final model.

4.4.2 α-synuclein E46K

4.4.2.1 α-synuclein E46K sample preparation

αSN_{E46K} was produced in *Escherichia coli* BL21 using a pET-11a vector, and expressed and purified as described previously. Lyophilized αSN_{E46K} was dissolved in 20 mM PBS buffer with 150 mM NaCl (pH 7.4). After filtration through 0.22-μm spin filters (Millipore) the concentration was determined by A_{280} nm using a Nanodrop UV-vis spectrophotometer (Thermo Scientific) with an extinction coefficient of 5,120 $M^{-1}cm^{-1}$. Solutions were prepared containing 12 mg/ml αSN_{E46K} and 20 mM ThT. Fibrillation of 150-ml aliquots was induced in 96-well optical bottom plates (Thermo Scientific) using a Fluostar Optima Platereader (BMG Labtech) under heating (37 C) and orbital shaking with 3-mm glass beads [232].

4.4.2.2 SAXS data collection and primary data evaluation

Three aliquots of protein were extracted from each well and SAXS data were recorded immediately after extracting the sample. Scattering data of αSN_{E46K} were recorded at the P12 beamline at European Molecular Biology Laboratory (EMBL) on the Petra III storage ring (Deutsches Electronen Synchrotron [DESY]) using the automated loading system [322]. The data were collected on a PILATUS detector with a momentum transfer range of $0.0049 < s < 0.35$ $Å^{-1}$ by 20 individual exposures of 45 ms each. The data reduction and buffer subtraction was performed by the beamline automated procedure [323] followed by a subsequent manual control. After close inspection of the multiple frames recorded at each time point, two curves were discarded, which exhibited unsystematic features. The insulin samples were collected on the X33 beamline at the EMBL on DORIS III (DESY) at a wavelength of 1.5 Å, using a MAR345 Image Plate Detector, in the momentum transfer range $0.006 < s < 0.51$ $Å^{-1}$ with 2-min exposure time. No radiation damage was detected when performing repeated exposures. Zinc acetate was added to the background buffers corresponding to two Zn^{2+} ions per insulin hexamer in the protein sample, and buffer measurements were performed immediately before and after each protein sample measurement. An averaged buffer measurement used for background subtraction. When the previously reported buffer effect, typical for fibril scattering data, was observed, background correction was applied as previously reported [320].

Data analysis was performed using the software suite ATSAS [50], and molecular masses were estimated relative to that of a standard reference solution of BSA. Guinier's

approximation was applied to provide rough estimates of the extrapolated forward scattering (I(*0*)) and radii of gyration (R_g) for the evolving samples. The *p(r)* functions were evaluated by the program GNOM [311], providing the maximal dimension (D_{max}) within the particle, and a second estimate of I(*0*) and R_g values.

The EOM was applied to structurally describe the COSMiCS-derived curve of monomeric αSN_{E46K} [61]. A pool of 4,000 conformations of αSN_{E46K} was built with Flexible-Meccano [69]. After addition of side chains with SCWRL4.0 [324], the individual theoretical SAXS profiles were computed with CRYSOL [41], and were used to select a subensemble of conformations that collectively described the experimental curve.

4.4.2.3 COSMiCS analysis of α-synuclein E46K SAXS dataset

The momentum transfer ranges used for the MCR-ALS analysis were $0.0074 < s < 0.3$ Å^{-1} for the absolute values, $0.0074 < s < 0.16$ Å^{-1} for Holtzer and Kratky representations, and $0.0074 < s < 0.07$ Å^{-1} for Porod's representation. The scattering curves selected as initial estimations were the curves 2 (37 min), 17 (6 h 46 min), and 47 (16 h 1 min). The maximum number of iterations, the convergence criteria, and the constraints were equivalent to those used in the insulin case. A Monte Carlo approach similar to the previously used by [325] was applied to estimate the standard deviations of the scattering intensities and the populations of the final solutions of the COSMiCS analyses of insulin and αSN_{E46K} datasets.

4.4.2.4 *Ab initio* modeling of α-synuclein

The structure of the repeating unit of αSN_{E46K} fibril was calculated with DAMMIN 4.24 using as starting point an ellipsoid of 200, 500, and 100 Å in 20 individual runs that were averaged and filtered. The structures were rendered with the program CHIMERA [326].

Chapter 5

COSMiCS

In this section I am going to describe in more detail the software used in the previous chapter, COSMiCS (Complex Objective Structural analysis of Multi-Component Systems). In addition of studying amyloids, COSMiCS can be used for decomposition of SAXS data from other polydisperse systems, such as transient interactions, oligomerization or conformational changes. In order to decompose the signal of the pure species coexisting in a polydisperse system, their population must change along the dataset. Therefore, a rational modification of some factor that produces this change has to be applied during the data collection; e.g. changes in pH or addition of a denaturant agent in folding studies, fibrillating conditions and time for amyloid studies, or increasing concentration in concentration-dependent oligomerizations.

I will explain the different parts composing the software, and how to perform the analysis, using the α-synuclein E46K (αSN_{E46K}) dataset described in the section 4.4.2 as example. In that case, the total protein concentration remains constant along the experiment. Finally, I will describe the results for the COSMiCS decomposition of two synthetic systems for which the concentration changes.

5.1 Implementation of COSMiCS

COSMiCS software is divided in different subroutines that are executed in a linear way using a command line procedure. All the steps will be described in this chapter in the same sequential order than in the program. The display of the program will be shown, with examples in answers and results in blue; the default choice is between [ ]. She selection of the options is not case-sensitive. An overview of the process is shown in Figure 5.1. The software has been implemented under MATLAB® 8.5 (Release 2015a) and does not need any toolbox apart from the MATLAB® standard core program. The software has been tested in computers under Linux Ubuntu 12 and Mac OS 10.10 "Yosemite" and 10.12.3 "Sierra" with no need of any particular additional resources.

5.2 Importing the data

The program loads the data from the files to the MATLAB® workspace and is organized in a data matrix of SAXS intensities.

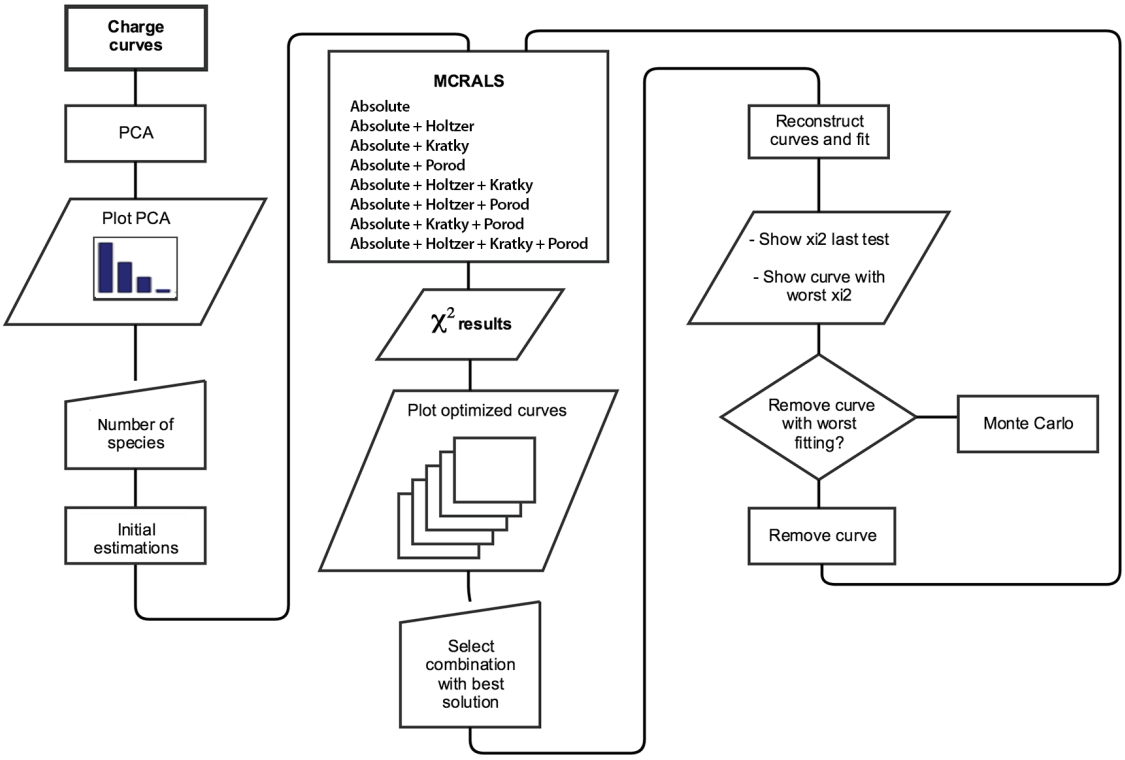


FIGURE 5.1. *COSMiCS flowchart*

5.2.1 Selection of folders

Selection of the folder containing the experimental data

A directory dialogue box will appear containing the data

```
» Select the folder containing the data
```

Selection of the folder to save the output

The user has to select the main output folder. The program will create additional folders inside with the different solutions of the optimization (see section 5.4.1)

```
» Where do you want to save the output?
```

Important: It is recommendable to name all the folders in the pathway without spaces because depending the operative system the program could crash.

5.2.2 Format of the experimental files

Name of the SAXS files

All the experimental SAXS files have to be in the same directory and they must have the same root and be sequentially named. COSMiCS will import only the files that fit

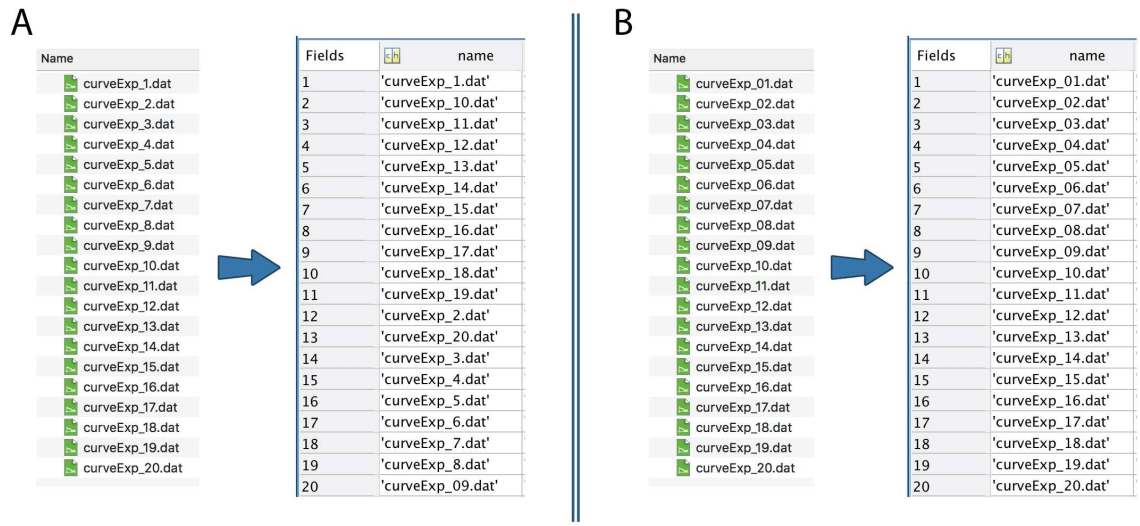


FIGURE 5.2. *Example of the order in which COSMiCS loads the files. A) Different digits for the numbering will produce an array of disordered files. B) Same number of digits will keep the same order in the loaded files than in the folder.*

the pattern, in alphabetical order, and it will give a correlative index to each curve. For this reason, it is important that the files have the same number of digits in the name. An example of the order how COSMiCS charge the files in the workspace is illustrated in the Figure 5.2.

For choosing the pattern of the files the user has to write the name of the files and an asterisk in the place of the numbers. COSMiCS will show the first and last files inside the input folder to help the user to choose the pattern of the data files.

```
» What is the pattern of the files (e.g.  curveExp_.dat ) ??
```

Number of header lines

Usually, SAXS files have certain number of header lines with text. All curves from the dataset must have the same number of header lines in order to only import the numeric values of each curve. COSMiCS do not remove the header lines from the original data. The first 10 lines of the first file will be shown to help the user select the number of header lines. Press Intro to keep the default value (3 lines).

```
» Number of header lines [3]:
```

Number of columns

Most of SAXS data formats have three columns: momentum transfer (s), intensity ($I(s)$), and experimental error ($\sigma(s)$). If the data have more than three columns you must to specify this information. COSMiCS will ignore the extra columns.

Important: columns must be in that order [s I(s) $\sigma(s)$]. If your data has only two columns [s I(s)], COSMiCS will work but it will not be able to calculate the χ^2. Press intro for keep the default value (3)

```
» Number of columns [3]:
```

5.2.3 Displaying curves

COSMiCS plots all the loaded dataset in semi-logarithmic scale and displays the number of curves in the command window, so the user can check that all the curves from the dataset have been included correctly.

```
» You added 25 files
» Curves (semi-logarithmic scale)...  Press any key to continue
```

5.2.4 Units of the experimental data

Select data units

The user has to specify the units of the original dataset: Angstroms or nanometers

```
» Units of your curves:  1/Angstrom (A) or 1/nanometers [n]:
```

Change data units

The user has the option to change the data units of the data. All the data that COSMiCS will generate will be in these chosen units.

```
» Do you want to change the units??  Y/[N]:
```

5.2.5 Removing initial points of the curves

It is usual that SAXS curves present artifacts at low momentum transfer values due to interparticle interactions or the proximity to the incident beam. It is possible for the user to remove these points from the analysis. The user selects the number of data points to be deleted, not the momentum transfer range. The program will show a plot with the dataset without these points for helping the user to decide the number of points.

Important: if applied, these points will be removed from all curves of the dataset as COSMiCS requires that all the curves have the same size.

```
» Do you want eliminate some data points at the beginning??  Y/[N]:
» Number of data points [0]:
```

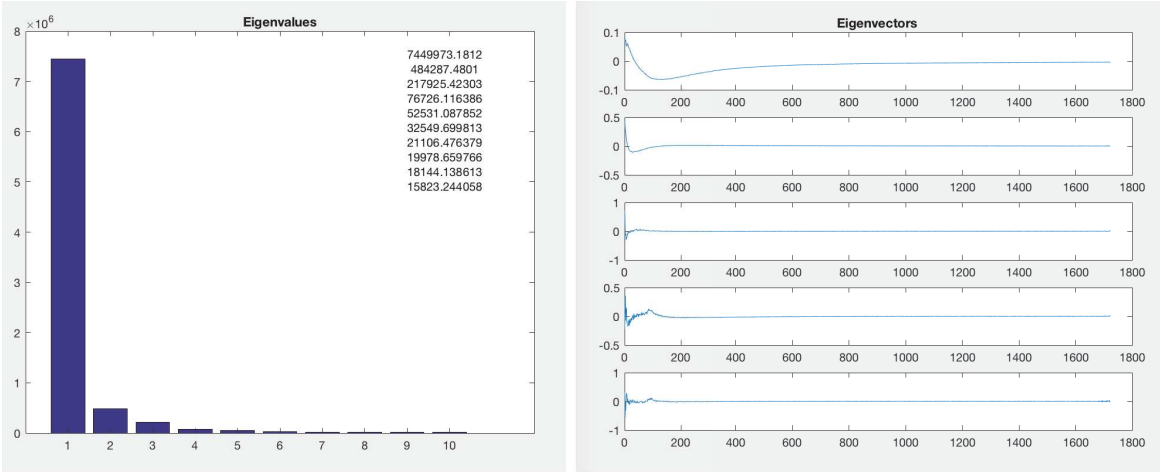


FIGURE 5.3. *PCA results for the αSN_{E46K} dataset. Ten first eigenvalues (left) and five first eigenvectors (right). In this case the PCA analysis suggests the presence of 3 species, although performing the COSMiCS analysis of 2 and 4 species would be useful for confirmation.*

5.3 Optimization parameters

Principal Component Analysis (PCA)

PCA has been described in the section 3.2. The main goal of this step in COSMiCS is to determine the minimum number of species needed to describe the experimental data.

PCA analyzes the experimental data without using any external information, in order to do an objective estimation of the number of species present. The program shows the results of the PCA plotting both the eigenvalues, in a Scree plot, and the eigenvectors (Figure 5.3). It is important to check both plots before deciding the number of species present. That information helps the user to decide the number of species of the system. However, prior information about the system can be used to determine this number. The COSMiCS optimization itself can be used also to confirm the optimal number of species by performing analysis using one species more or one species less, and deciding the best solution using statistical tools.

```
» Do you want see results of PCA??  Y/N [Y]:
```

Selection of the number of species

Independently of whether the user inspects the PCA results or not, it is necessary to introduce the number of species that the program will use to perform the optimization.

```
» Number of species [3]:
```

5.3.1 Number of species

Independently of whether the user inspects the PCA results or not, it is necessary to introduce the number of species that the program will use to perform the optimization.

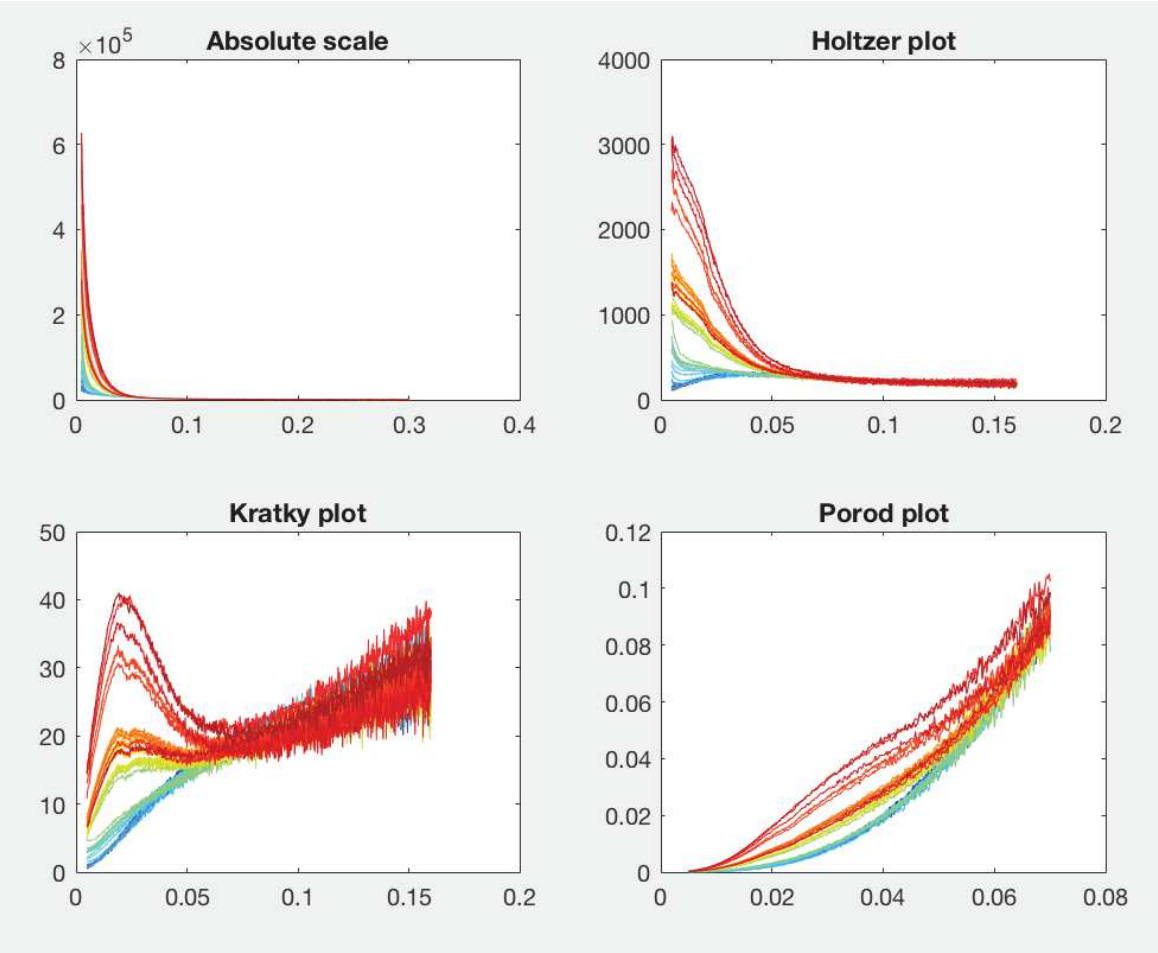


FIGURE 5.4. *SAXS dataset of αSN plotted in the four representations that will be used by COSMiCS.*

```
» Number of species [3]:
```

5.3.2 Selection of the momentum transfer range

During the optimization, COSMiCS will combine the Absolute values of the SAXS profiles with different representations of the dataset: Holtzer, Kratky and Porod. These representations present an increasing amount of noise at the wide-angle momentum transfer range, especially in Porod's representation. The user can select an individual momentum transfer range for each representation to be analyzed with COSMiCS. The use of a specific momentum transfer range for each data representation is recommended. COSMiCS will plot the different representations of the data (Figure 5.4) to help the user choose the best one depending the type of data.

```
» Do you want to cut the matrices??  Y/N [Y]:
```

If the user wants to select different momentum transfer for the different representations the program will plot the different representations with suggested ranges for each one. The user can use the suggested values or insert manually the momentum transfer for

each of the representations. COSMiCS will show a plot with the selection and will ask for confirmation before proceeding with the next step.

5.3.3 Initial estimations

Search for the initial estimations of the SAXS curves

Initial estimations of the pure scattering curves best describing the complete set of SAXS curves are obtained by searching the 'purest' curves among all the experimental ones these are subsequently used as the starting point for the MCR-ALS optimization (see section 3.3). This process is performed with the program PURE [327].

```
»Calculating initial estimations
»
» Curve 1 - curveExp_01.dat
» Curve 2 - curveExp_02.dat
...
» Curve 20 - curveExp_20.dat

first purest variable:  17
next purest variable:  4
next purest variable:  3
```

Selection of initial estimations

The user can use the initial estimations selected by PURE or to introduce manually other experimental curves of the dataset as starting points for the optimization process. The user must introduce one by one the estimations using the index of loaded curve, not the number in the file name. COSMiCS shows a list with the number of curve and the name of the original file to help the user to know the index of the curve. COSMiCS will plot the selected initial estimations (Figure 5.5).

```
» Do you want use the initial estimations selected by pure??  [17
4 3] [Y]/N: n

If the user decides introduce the estimations manually:
» Curve 1 - curveExp_01.dat
» Curve 2 - curveExp_02.dat
...
» Curve 20 - curveExp_20.dat
» Introduce the curves that you want as initial estimations:  »

Initial estimation:  1
» Initial estimation:  10
» Initial estimation:  20
```

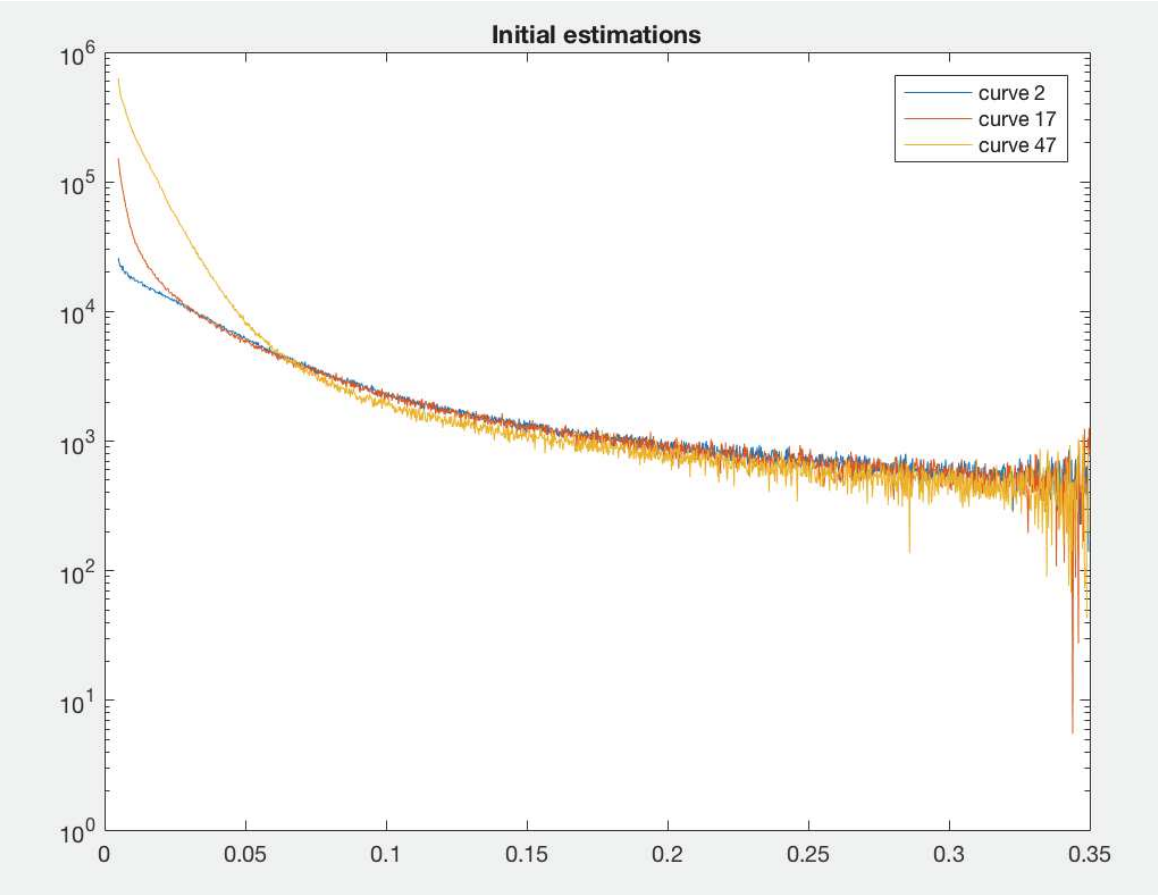


FIGURE 5.5. *Selected initial estimations for the MCR-ALS optimization after the sorting.*

Sorting the initial estimations

When COSMiCS does the search for the initial estimations, their index can appear disorganized (As in the previous example [17 4 3]). In some cases, especially fibrillation experiments, it can be interesting to sort the initial estimations, so the first species corresponds to the species that appears before in time during the fibrillation. In the example the curves will be [3 4 17] after the sorting. This sorting helps the user to visualize the results, but has no effect in the optimization.

```
» Do you want sort the curves??  [Y]/N:
```

5.3.4 Selection of constraints

Type of experiment

The use of constraints is key for the successful application of COSMiCS (see section 3.3.2 for more details). In the present version of COSMiCS, two types of experiments are proposed. Importantly, the type of experiment determines the constraints that will be applied.

```
» Type of experiment
» (1) Fibrillation or folding
» (2) Titration
» Select type of experiment:
```

(1) Fibrillation or folding (default). Constrains used in this type of experiments are *non-negativity* (ffnls) for both concentrations and spectra, and *closure* for concentrations. This constraint scheme is applicable to datasets measured for systems in which the sample concentration is maintained and the species evolve in a time-dependent manner or by changing the experimental conditions (temperature, pH, ionic strength, denaturants...).

```
» Standard options for fibrillation or folding:
» - Non-negativity (fnnls) for concentrations and spectra
» - Closure for concentrations (equal to 1.0)
» Press any key to continue
```

(2) Titration. Constraints used in this type of experiments are *Non-negativity* (fnnls) for both concentrations and spectra, and *Unimodality* for concentrations. No mass conservation (*closure)* is applied. This constraint scheme is applicable to systems in which the total concentration is changed such as titration experiments to study protein oligomerization, or low-affinity protein-protein interactions.

```
» Standard options for titration:
» - Non-negativity (fnnls) for concentrations and spectra
» - Unimodality for concentrations
» - No closure.
```

Equality constraint

In case of a system where the curve of a pure species is already available, it is possible to fix it during the optimization. This is called equality constraint. It is possible to use this constrain in both types of experiments.

Important: This fixed curve must be one of the initial estimations (see above). In case that the algorithm does not detect it as one of the purest curves, it is necessary to do it manually. COSMiCS will ask to the user if fixing one or more initial estimation curves is desired.

```
» Do you want to fix any species?  Y/[N]
» Initial estimation #1 - curve 1 (curveExp_01.dat)
» Do you want to fix this species?  Y/[N] y
»
```

```
» Initial estimation #2 - curve 10 (curveExp_10.dat)
» Do you want to fix this species?  Y/[N] n
»
» Initial estimation #3 - curve 20 (curveExp_20.dat)
» Do you want to fix this species?  Y/[N] n
```

5.3.5 Convergence criterion

In two consecutive iteration cycles the relative standard deviation difference of the residuals between experimental and ALS calculated data values are compared. If this difference is less than the convergence criterion value, convergence is achieved and the optimization finishes. This value is 0.1% by default, but it can be modified by the user depending on the stage of the optimization. Usually at the beginning of the study a higher value is used (i.e. 1%) for exploratory purposes. In contrast, once a good model has been found, lower values are attempted to see whether there is an improvement in the solution.

```
» Convergence criterion (0.1% by default)
```

5.3.6 Maximum number of iterations

The user can decide a maximum number of iterations allowed in the optimization process. If the process does not achieve convergence before this number, the optimization will stop in a divergence situation. In case of divergence, this information will appear in the results (see below).

```
» Maximum number of iterations (50 by default)
```

5.3.7 Graphical output

A graphical display of the optimization process is possible (Figure 5.6). This possibility enables the visualization of the optimization process but it will make the total running time longer.

Note: Only spectra corresponding to the Absolute scale optimization are displayed in a semi-logarithmic scale.

```
» Do you want a graphical output during the ALS optimization [Y]/N)?
```

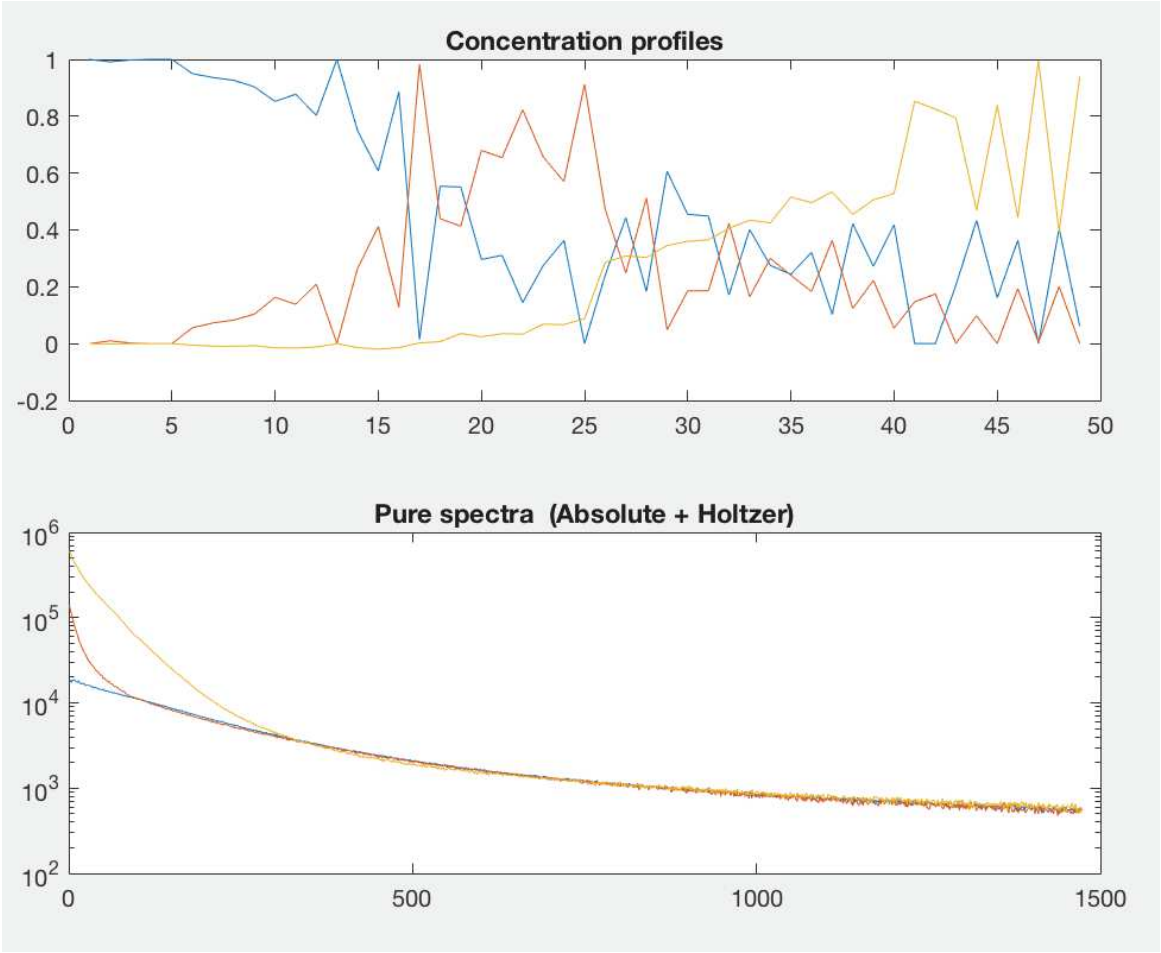


FIGURE 5.6. *If the graphical output is selected, this window will be shown and updated with the current combination during the optimization. The combination showed (Absolute + Holtzer) is plotted in semi-logaritmic scale for better inspection.*

5.4 Optimization process

5.4.1 MCR-ALS with different representations

The program performs the optimization for all possible combinations of SAXS data representations. In an initial step, COSMiCS generates for each combination the data matrix, which is scaled by dividing all the intensity values of each representation by the first eigenvalue of its individual PCA. This scaling step is necessary to not overweight a certain representation. The final dataset is a row-wise augmented matrix of spectra that is used as input. Then COSMiCS performs the MCR-ALS optimization in the following order:

```
1.  Absolute [A]
2.  Absolute + Holtzer [AH]
3.  Absolute + Kratky [AK]
4.  Absolute + Porod [AP]
5.  Absolute + Holtzer + Kratky [AHK]
6.  Absolute + Holtzer + Porod [AHP]
7.  Absolute + Kratky + Porod [AKP]
8.  Absolute + Holtzer + Kratky + Porod [AHKP]
```

Optimization results

After each optimization, the COSMiCS procedure back-calculates the solution curves and compares them with the experimental dataset to derive individual χ^2 values for each curve and the average χ^2 for each combination. The graphical interface of COSMiCS displays all the optimized curves for each combination (Figure 5.7).

The program will show a message when the optimization for each combination is finished, showing if convergence has been achieved and the different statistical parameters used to measure the quality of the fitting (see section 3.3). This message will appear for each combination, but the next combination will start immediately. This information will be saved in a file for each combination (see section 5.5).

```
Example of the information displayed at the end of the optimization:
» CONVERGENCE IS ACHIEVED !!!!
» Fitting error (lack of fit, lof) in % at the optimum = 3.984(PCA)
5.0693(exp)
» Percent of variance explained (r2) at the optimum is 99.743
» Relative species conc.  areas respect matrix (sample) at the
optimum
» Plots are at optimum in the iteration 6
```

When the program performs the optimization with all the combinations it shows a list with the averaged χ^2 for each of them:

```
» Test 1:  xi2 (A) = 2.62
» Test 2:  xi2 (A+H) = 1.22
» Test 3:  xi2 (A+K) = 1.18 DIVERGENCE!
» Test 4:  xi2 (A+P) = 1.33
» Test 5:  xi2 (A+H+K) = 1.16
» Test 6:  xi2 (A+H+P) = 1.24
» Test 7:  xi2 (A+K+P) = 1.16
» Test 8:  xi2 (A+H+K+P) = 1.16» Press any key to continue
```

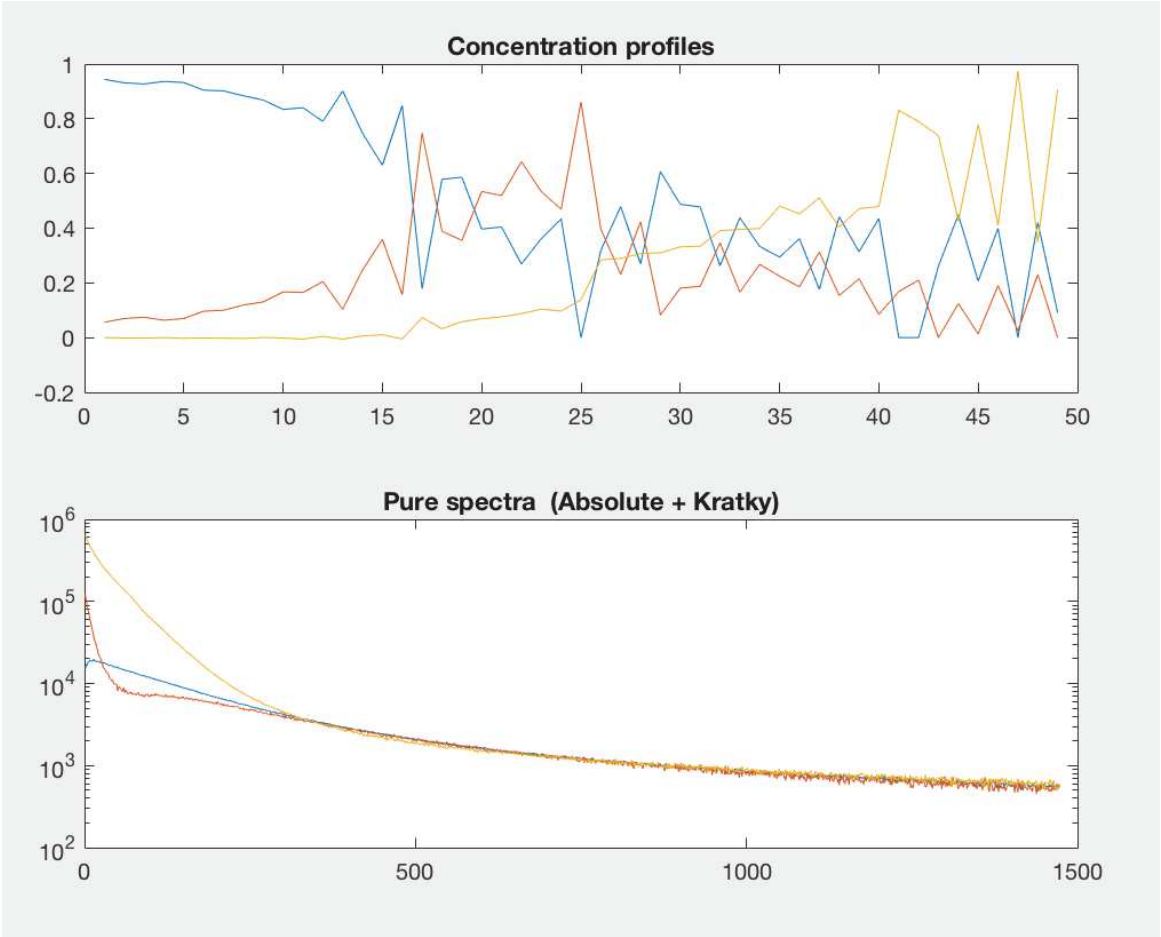


FIGURE 5.7. *Example of the final solution corresponding to the AK combination. The other 7 solutions will be displayed. In addition of the average $\langle\chi^2\rangle$, the inspection of the solutions helps to decide the best combination. For example, in this case we obtain a non SAXS-like curve for one of the species (bottom plot), The population profile can be used also as a criteria to select the best solution (e.g. presence of negative values).*

5.4.2 Remove outliers

It is possible that the dataset has one or several curves that are outliers and cannot be well described with the model. In this case, it is possible to identify and remove these curves. First, the user can choose the best solution. The program suggests the best combination according the χ^2 among the combinations that converged; the user can accept this solution as the best or choose a different one. It is recommendable to inspect the graphical output of the solutions for choosing the best solution instead using only the χ^2 as the only criteria. After the selection of the best solution, the program will provide a sorted list with the curves with the largest individual χ^2. The user has the option of removing the curve with worst χ^2 from the analysis.

Note: COSMiCS will not erase the previous solution with the complete dataset.

```
» The best combination is Absolute + Holtzer + Kratky + Porod (xi2
= 1.1602)
» Do you want use this combination?  Y/N [Y]:
» Now we going to detect outliers...
» Average xi2:  1.16
» The 10 worst xi2 are:
» 3.8519
» 2.0559
...
» 1.4034

» The curve with worst xi2 (45) - xi2 = 3.8519
» Do you want remove this curve?  Y/N [Y]: y
```

In case that the curve is removed and the analysis is performed again. COSMiCS will start again the optimization for the eight combinations of matrices. COSMiCS will provide a $\Delta\chi^2$ between both optimizations (the best solution previously selected and the new solution for each combination) for each individual curve to confirm the improvement of the results and the overall benefit of removing the experimental curve (bottom plot Figure 5.8).

```
» Now we going to repeat the different combinations without this
curve
» Press any key to continue
```

5.5 Output

COSMiCS will create different files and folders inside the output folder selected by the user (Figure 5.9).

5.5.1 Output files

- Eigenvalues.txt. All the PCA eigenvalues in one column
- Eigenvectors.txt. Ten first PCA eigenvectors in different columns.
- Info.txt. It contains general information about the optimization:
  - Folder containing the experimental data
  - List of files: index and corresponding file names
  - Number of species used
  - Used constrains
  - Species fixed (if any)
  - Convergence criterion

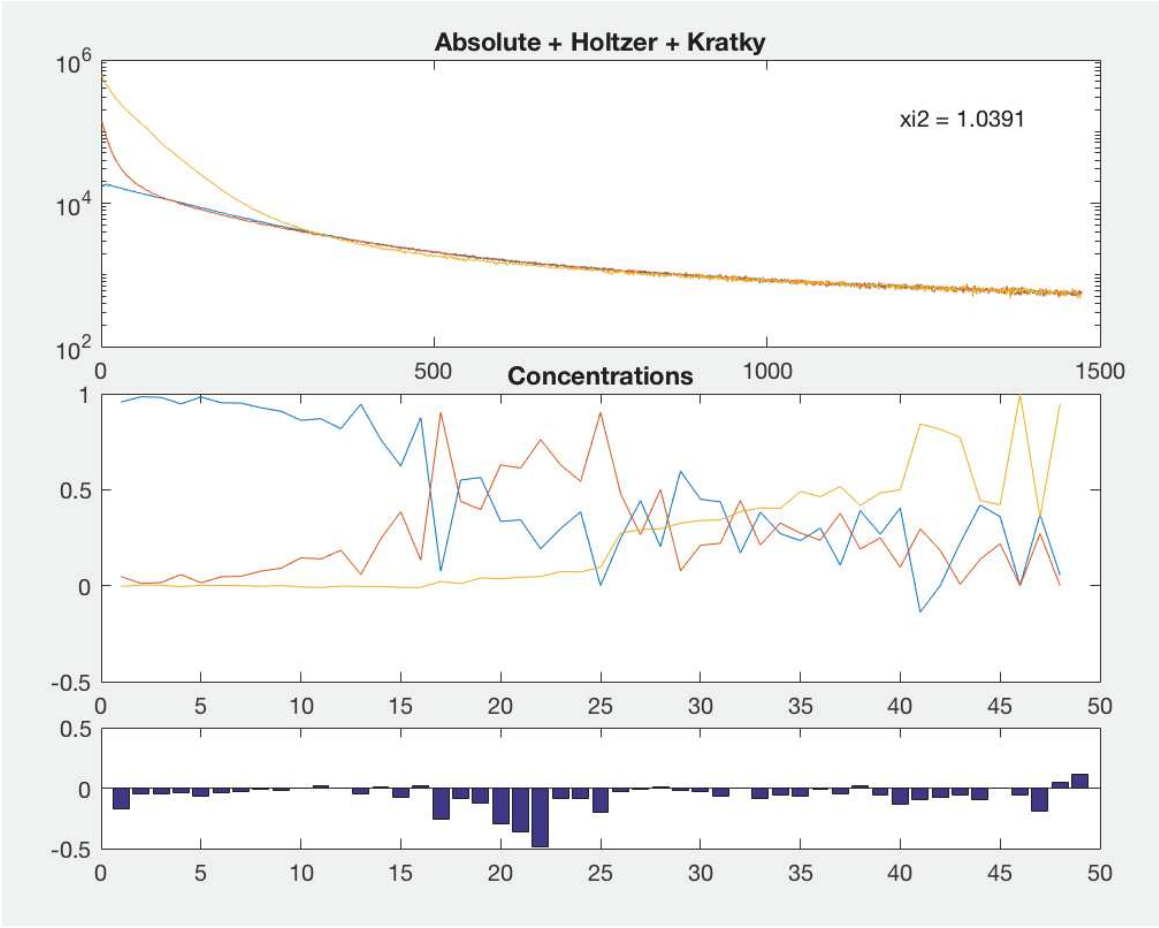


FIGURE 5.8. *Example of the resulting optimization after removing an outlier. The bottom plot shows the difference of χ^2 ($\Delta\chi^2$) for the individual curves between the previous best solution and the actual solution without the outlier. Negative values of $\Delta\chi^2$ indicate a better description of the individual curve after removing the outlier.*

  - Maximum number of iterations

- A different folder is created for each combination of representations. The folder name consists in a correlative number and the combination of representations used in capital letters (e.g. Test05_AHK). This folder contains:
  - Spectra in individual species (species*.dat).
  - Concentrations of the individual species along the dataset (concentrations.txt).
  - χ^2 of each curve against the experimental curve. One column format.
  - Other information about the optimization results (infoTest#.txt).
    * Combination of representations used
    * Initial estimations
    * *s* range selected for each representation
    * Results of the optimization
      · Lack of fit (exp)

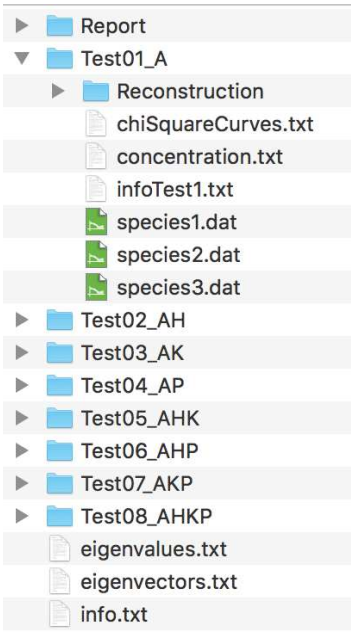


FIGURE 5.9. *Screenshot of the output folders and files created by COSMiCS*

- Lack of fit (PCA)
- Percent of variance explained

* Average χ^2

5.5.2 Reconstruction of the curves

The user has the option of reconstruct the dataset using the solutions obtained with COSMiCS. Reconstructed curves are written as individual files (curveMCRALS_*.dat) in the folder named 'Reconstruction' and placed in the output folder. In addition, the experimental dataset (curveExp_*.dat) is copied in the folder to facilitate the visualization of the results with external programs. It is possible to do the reconstruction only for some combinations to save space in the disk.

```
» Do you want reconstruct the curves??  [Y]/N:
» Which combination do you want the reconstruction??  (i.e.  [1
5 7]) - 0 for all:  [3 4]
```

5.5.3 Report

A report with all the selected options and solutions will be created with the name report.html in a folder named 'Report' inside the output folder (Figure 5.10 to 5.14). In addition of the report, the folder contains the images used in the report in a png format.

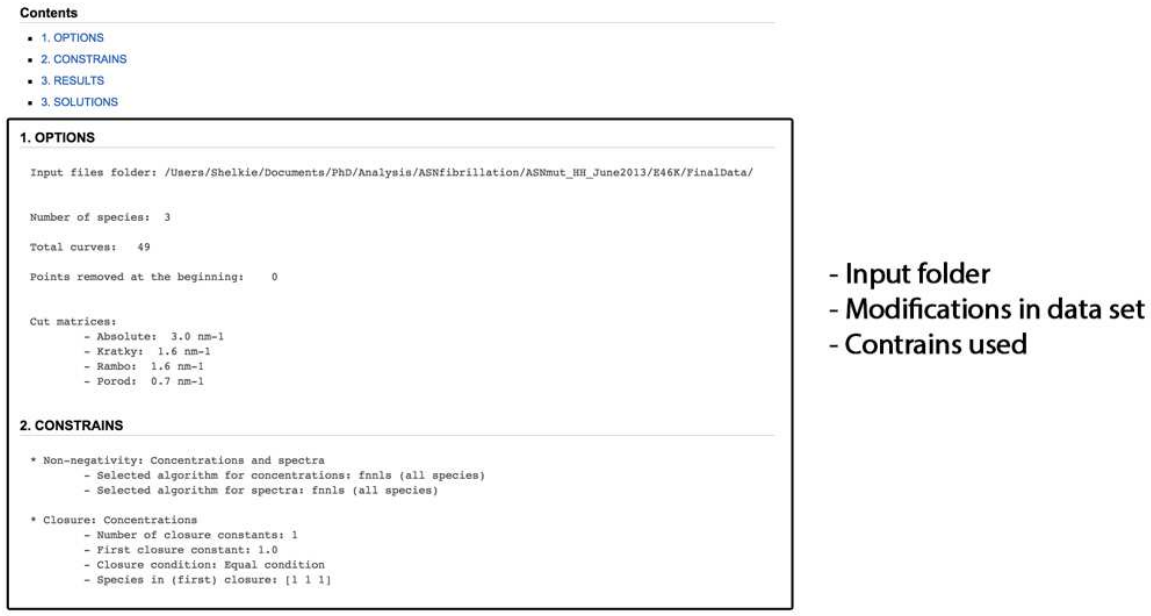


FIGURE 5.10. *Screenshot of the report html file (1 of 5)*

5.6 Monte Carlo error analysis (optional)

5.6.1 Monte Carlo approach

SAXS-based structural modeling procedures require proper estimations of the uncertainty for each of the SAXS intensity values. The uncertainties on both SAXS curves and concentration profiles are derived using a Monte Carlo error analysis similar to what was previously used by Svergun and Pedersen [325]. To perform this analysis, random noise based on the experimental error bars is added to each of the reconstructed (back-calculated) curves. This new synthetic dataset is then submitted to the MCR-ALS optimization. This process is repeated a minimum of one hundred times using different values of random noise. The optimized SAXS curves and the concentration profiles in each Monte Carlo cycle that converge are stored and the standard deviations for the intensities, I(s), and for the concentration profiles are calculated.

The number of Monte-Carlo simulations is selected by the user (100 by default). This process is time consuming, so it exists the option of performing the analysis just for one combination. New files with the pure species and the errors from the Monte Carlo error analysis will be created in the corresponding folders of the solutions. The old files are kept and the new ones will be named species*MC.dat.

```
    » Do you want to compute the uncertainty of the solution using
the Monte Carlo method??  [Y]/N: y
    » How many runs do you want for the Monte Carlo??  (default 100)
100
```

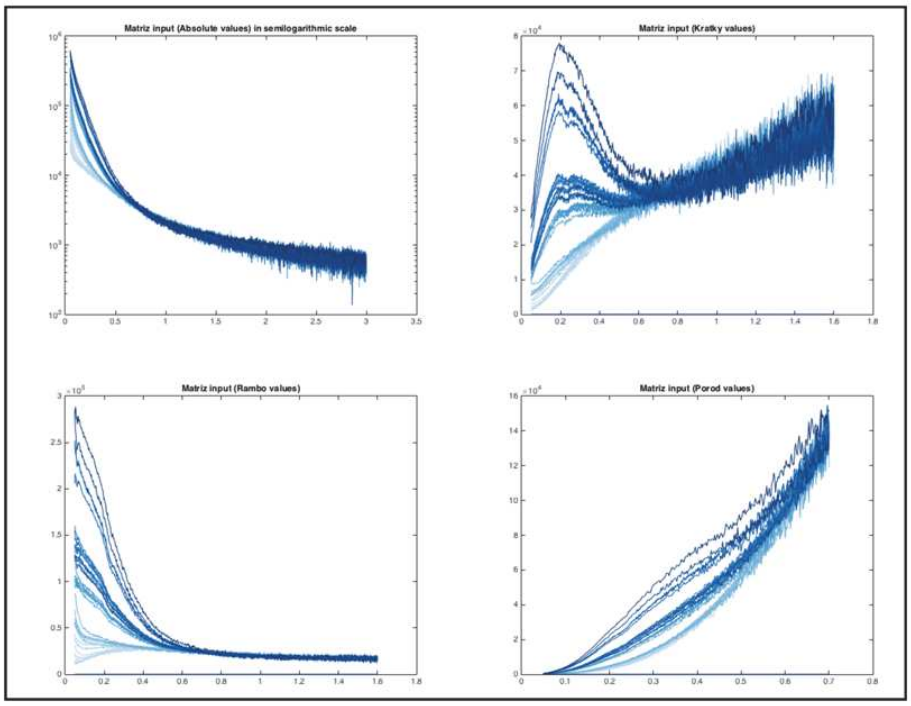
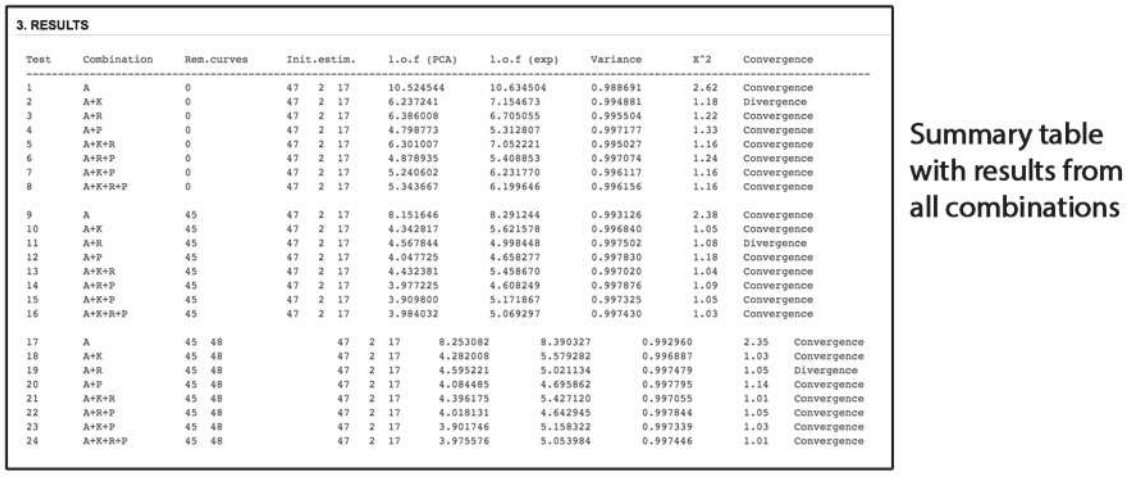


FIGURE 5.11. *Screenshot of the report html file (2 of 5)*



3. RESULTS

Test	Combination	Rem.curves	Init.estim.	l.o.f (PCA)	l.o.f (exp)	Variance	X^2	Convergence
1	A	0	47 2 17	10.524544	10.634504	0.988691	2.62	Convergence
2	A+K	0	47 2 17	6.237241	7.154673	0.994881	1.18	Divergence
3	A+R	0	47 2 17	6.386008	6.705055	0.995504	1.22	Convergence
4	A+P	0	47 2 17	4.798773	5.312807	0.997177	1.33	Convergence
5	A+K+R	0	47 2 17	6.301007	7.052221	0.995027	1.16	Convergence
6	A+R+P	0	47 2 17	4.878935	5.408853	0.997074	1.24	Convergence
7	A+K+P	0	47 2 17	5.240602	6.231770	0.996117	1.16	Convergence
8	A+K+R+P	0	47 2 17	5.343667	6.199646	0.996156	1.16	Convergence
9	A	45	47 2 17	8.151646	8.291244	0.993126	2.38	Convergence
10	A+K	45	47 2 17	4.342817	5.621578	0.996840	1.05	Convergence
11	A+R	45	47 2 17	4.567844	4.998448	0.997502	1.08	Divergence
12	A+P	45	47 2 17	4.047725	4.658277	0.997830	1.18	Convergence
13	A+K+R	45	47 2 17	4.432381	5.458670	0.997020	1.04	Convergence
14	A+R+P	45	47 2 17	3.977225	4.608249	0.997876	1.09	Convergence
15	A+K+P	45	47 2 17	3.909800	5.171867	0.997325	1.05	Convergence
16	A+K+R+P	45	47 2 17	3.984032	5.069297	0.997430	1.03	Convergence
17	A	45 48	47 2 17	8.253082	8.390327	0.992960	2.35	Convergence
18	A+K	45 48	47 2 17	4.282008	5.579282	0.996887	1.03	Convergence
19	A+R	45 48	47 2 17	4.595221	5.021134	0.997479	1.05	Divergence
20	A+P	45 48	47 2 17	4.084485	4.695862	0.997795	1.14	Convergence
21	A+K+R	45 48	47 2 17	4.396175	5.427120	0.997055	1.01	Convergence
22	A+R+P	45 48	47 2 17	4.018131	4.642945	0.997844	1.05	Convergence
23	A+K+P	45 48	47 2 17	3.901746	5.158322	0.997339	1.03	Convergence
24	A+K+R+P	45 48	47 2 17	3.975576	5.053984	0.997446	1.01	Convergence

Summary table with results from all combinations

FIGURE 5.12. *Screenshot of the report html file (3 of 5)*

```
    » For which tests do you want to perform the error analysis??  (i.e.
[1 5 7]) - 0 for all:  3
```

5.7 Examples of the use of COSMiCS using synthetic data

5.7.1 Example of a system in equilibrium with mass conservation

This first example with synthetic data is a system with constant concentration along the dataset, conceptually equivalent to the one presented in the chapter 4 for amyloids.

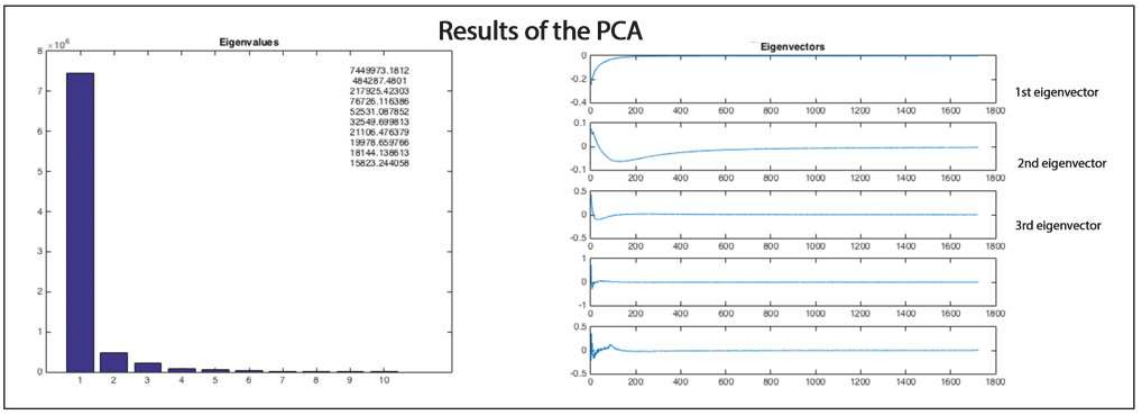


FIGURE 5.13. *Screenshot of the report html file (4 of 5)*

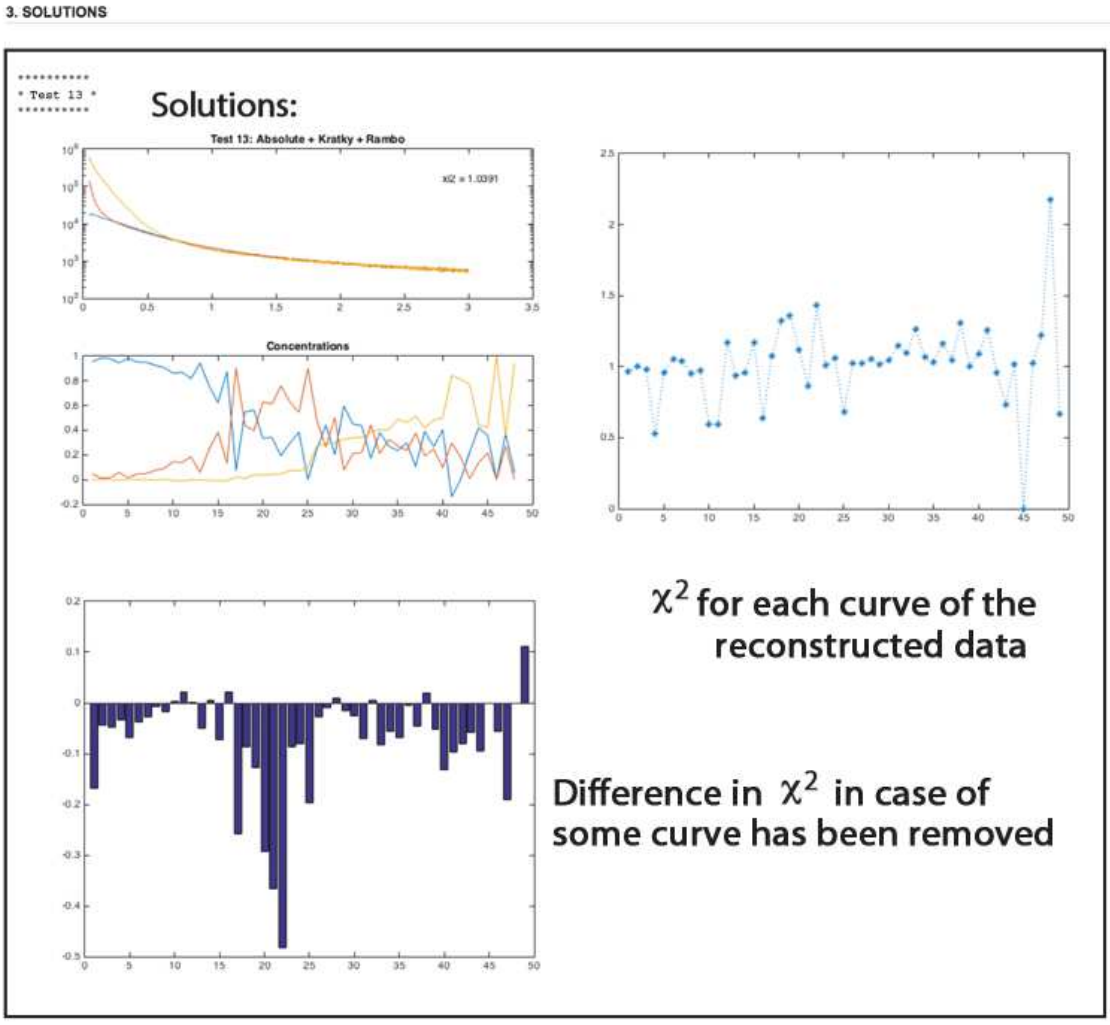


FIGURE 5.14. *Screenshot of the report html file (5 of 5)*

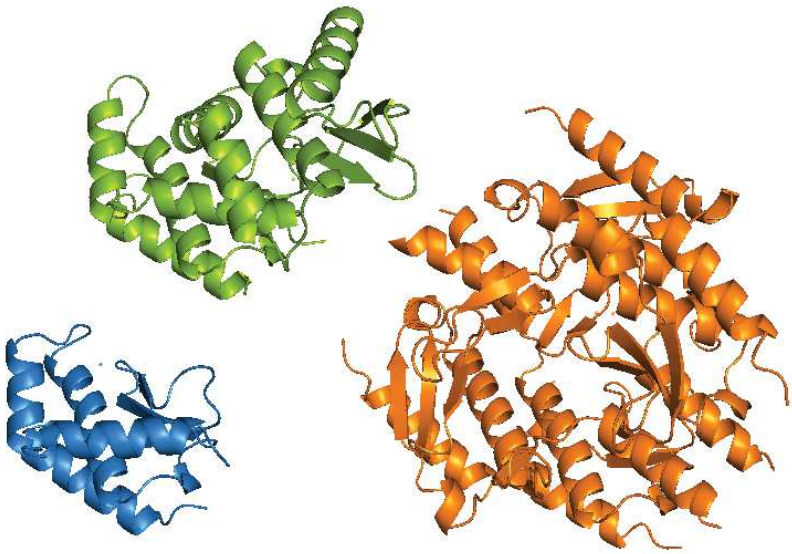


FIGURE 5.15. *Crystallographic structures of the monomer (4QHF, blue), dimer (4QHG, green) and tetramer (4QHH, orange) of the selecase, determined in [80]*

5.7.1.1 Generation of synthetic data

We have generated a synthetic SAXS dataset based on the selecase oligomerization [80]. Concretely, we designed a monomer-dimer-tetramer model corresponding to pdb codes 4QHF, 4QHG and 4QHH, respectively (Figure 5.15). The theoretical scattering profiles for the three structures were computed with CRYSOL [41] using the maximum number of spherical harmonics and maximum order for the Fibonacci's grid representations. All other parameters were in default setting. These three curves were scaled in order to have a monomeric species curve with a forward scattering, I(*0*), equivalent to the curve at 8.5 mg/ml selecase concentration from the experimental dataset [80]. A concentration model that included a Gaussian population for the dimeric species was generated making sure that the sum of molar fractions at each time-point was 1.0. The final dataset of curves was calculated for 50 time-points using the synthetic profiles for the three species and the molar fractions of the kinetic model.

Synthetic noise was added to these curves based on the experimental error estimations of the experimental SAXS curve of selecase at 8.5 mg/ml. The relative noise from the experimental dataset was calculated by dividing the experimental noise [$\sigma(s_{\text{exp}}$] by the intensity [I(s)$_{\text{exp}}$] observed:

$$k(s)_{\text{exp}} = \frac{\sigma(s)_{\text{exp}}}{I(s)_{\text{exp}}} \qquad \text{(Eq. 5.1)}$$

The resulting factor, $k(s)_{\text{exp}}$, was related to the I(0)$_{\text{sim}}$ through its ratio with I(0)$_{\text{exp}}$ and multiplied by the intensity of the simulated scattering curve [I(0)$_{\text{sim}}$] giving the noise $\sigma(s)_{\text{sim}}$:

$$\sigma(s)_{\text{sim}} = k(s)_{\text{exp}} \sqrt{I(0)_{\text{sim}}/I(0)_{\text{sim}}} I(s)_{\text{sim}} \qquad \text{(Eq. 5.2)}$$

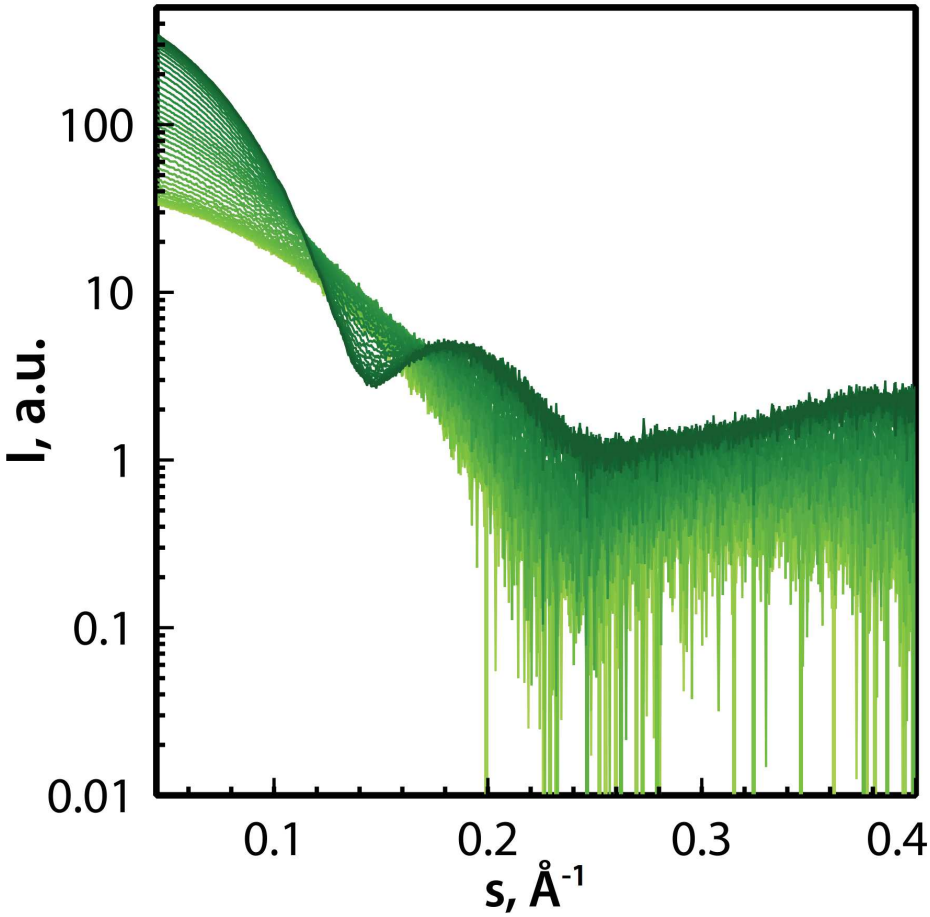


FIGURE 5.16. *Complete synthetic dataset (50 curves) of the selecase example in semi-logarithmic scale. A noise level of 2·$\sigma(s)_{sim}$ was used in this example.*

For the inclusion of noise in the simulated scattering curve, values from a Gaussian distribution with a standard deviation, $\sigma(s)_{\text{sim}}$, centred around zero were added to the curve. Two independent datasets were generated: one using the standard deviation $\sigma(s)_{\text{sim}}$, and a second one multiplying $\sigma(s)_{\text{sim}}$ by 2.0 in order to increase the level of noise of the curves (Figure 5.16). Both datasets yielded very similar results with the COSMiCS analysis, only results of second dataset are presented here.

5.7.1.2 COSMiCS analysis

COSMiCS was applied to the synthetic dataset, using the standard options for a fibrillation experiment (see section 5.3.4).

The resulting χ^2 resulting from the different data representations are displayed in the Table 5.1. A very good agreement to the complete dataset is observed for the majority of combinations. Some combinations including Porod's representation present larger χ^2. Importantly, all combinations with χ^2 smaller than 1.0 present almost equivalent results. The results of the AH combination are presented in Figure 5.17. The resulting decomposed

Representations included					
Code	Absolute I(s)	Holtzer I(s)*s	Kratky $I(s)*s^2$	Porod $I(s)*s^4$	$\langle\chi_i^2\rangle$
A	+				0.95
AH	+	+			0.93
AK	+		+		0.93
AP	+			+	1.24
AHK	+	+	+		0.93
AHP	+	+		+	1.22
AKP	+		+	+	0.95
AHKP	+	+	+	+	0.95

TABLE 5.1. *Results of the COSMiCS analysis of the synthetic dataset for the selecase test that has mass conservation*

curve profiles for the three species and their relative populations are in excellent agreement with the theoretical values used to simulate the data.

5.7.1.3 Results

The three optimized curves are in excellent agreement with the curves used to calculate the dataset. The one displaying the lowest level of uncertainty is the tetramer due to the fact that it is the largest one, and its curve presents more features. Despite the fact that the dimer and monomer curve are more similar between them, the monomer has a better agreement (χ^2=1.46) than the dimer because it is virtually alone in the first part of the dataset (Figure 5.17B). The dimer displays the worst agreement (χ^2 = 1.67) because it is always present simultaneously with the other species. The inspection of the populations (Figure 5.17B) shows that the parts of the dataset where the three species are together presents larger error bars from the Monte Carlo analysis, especially for monomer and dimer.

5.7.2 Example of a synthetic SAXS dataset along a titration experiment

In this example I present a system where one of the species concentration is constant and a second species is added stepwise. Both species bind to form a complex with a known dissociation constant, k_d. The total concentration is not constant along the system, so the constraints applied are different than in the previous example.

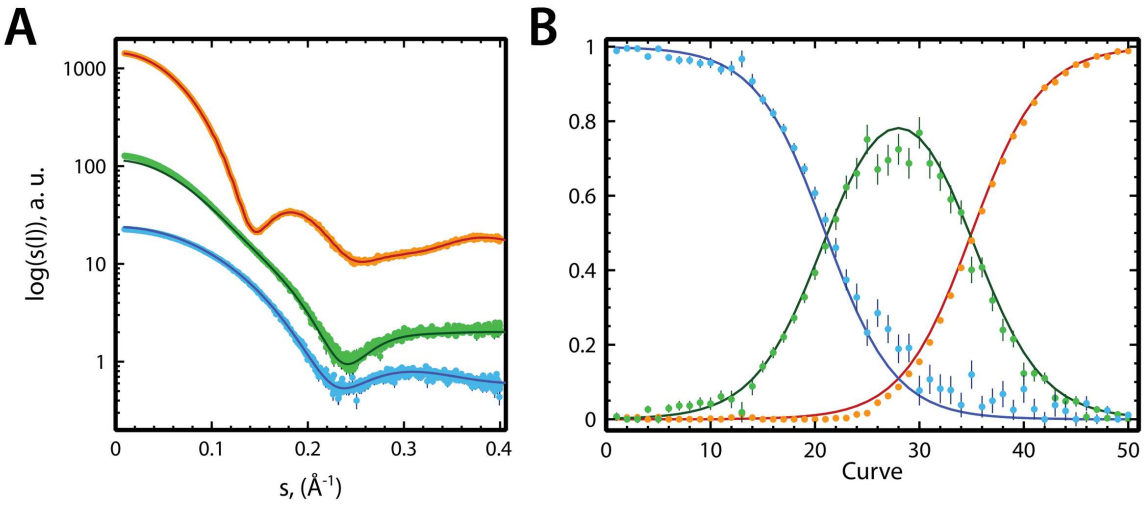


FIGURE 5.17. *COSMiCS analysis of the selecase dataset. (A) SAXS profiles for the isolated species using the AH combination: monomer (blue), dimer (green) and tetramer (orange) and the original CRYSOL curves superimposed in darker colors. (B) Kinetic model used to create the dataset (solid lines) and the concentration profiles derived from COSMiCS (dots), with the same color code as the spectra profiles.*

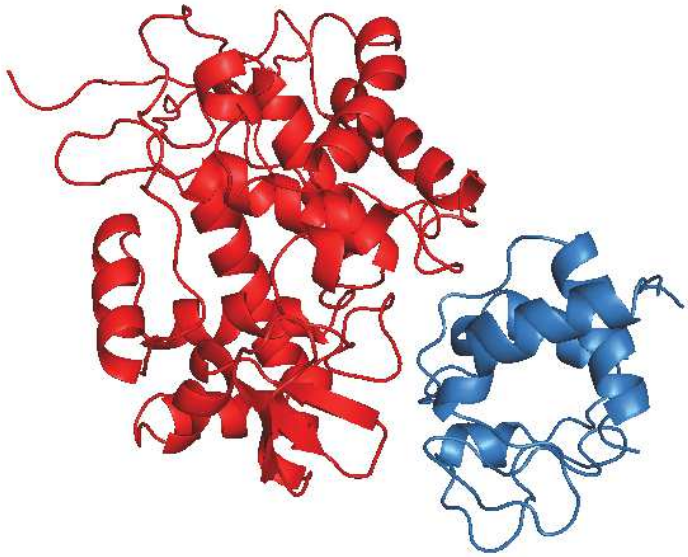


FIGURE 5.18. *Crystallographic structure of the complex formed by the cytochrome c peroxidase (2PCC, chain A, in red) and the iso-1-cytochrome c (2PCC, chain B, in blue) [328] used to generate the synthetic titration SAXS dataset.*

5.7.2.1 Generation of synthetic data

We have generated a synthetic SAXS dataset following a typical biomolecular equilibrium reaction, A + B $\rightleftharpoons$ AB. A population (thermodynamic) model was generated increasing the amount of the reactant [A] with a fixed amount of the reactant [B]. The equilibrium concentration of the product [AB] was determined using an equilibrium dissociation constant (K_d) equal to 0.1 μM.

The partners used for the model are extracted from the pdb 2PCC, which corresponds

to the complex between yeast cytochrome c peroxidase and yeast iso-1 cytochrome c [328]. In the original pdb, subunits A and C correspond to two units of cytochrome c peroxidase and the subunits B and D correspond to two units of iso-1-cytochrome c. To simplify the model, I have used just one unit of each partner; that is reactant [A] corresponds to cytochrome c peroxidase and reactant [B] corresponds to cytochrome c (Figure 5.18).

In the same way than the previous synthetic data, the theoretical scattering profiles of the individual partners and the complex were computed with CRYSOL [41] using the maximum number of spherical harmonics and maximum order for the Fibonacci's grid representation. All other parameters were in default setting.

The final dataset of curves was calculated for 50 concentration points using the synthetic profiles for the three species and the molar fractions of the thermodinamic model.

The noise was added to the final curves using the same experimental curves than in the previous case. The three synthetic curves were scaled in order to have the minor species curve (compound B, cythocrome c) with a forward scattering, I(*0*), equivalent to the curve of 8.5 mg/ml selecase concentration from the experimental dataset [80]. The noise, $\sigma(s)_{\text{sim}}$, was calculated using Eq. 5.2, like in the previous synthetic dataset. For the inclusion of noise in the simulated scattering curve, values from a Gaussian distribution with a standard deviation, $\sigma(s)_{\text{sim}}$, centred around zero were added to the curve. In order to generate a dataset with a more realistic noise level, the $\sigma(s)_{\text{sim}}$ value was increased multiplying by 10 (Figure 5.19).

5.7.2.2 COSMiCS analysis

COSMiCS was applied to the synthetic dataset, using the standard options for a titration experiment (see section 5.3.4). In the case of a titration experiment the pure curves of the compounds A and B can be measured individually, therefore they are available to use as additional information, fixing the curves during the optimization process. Obviously, the optimization must be done using as much information as is available to obtain the best solution. In order to illustrate how the use of supplementary information improves the solution, two analyses were performed, one fixing the pure curve of both compounds A and B (equality constraint, see section 3.3.2.4) and another without fixing any curve.

5.7.2.2.1 COSMiCS analysis fixing curves from known species

The resulting χ^2 from the different data representations are displayed in the Table 5.2. The agreement of the reconstructed and experimental curves is good, although the combinations including Porod's representation present larger χ^2. When we compare the original curve of the AB complex generated by CRYSOL and the second species obtained by COSMiCS, we can see that the fitting is perfect, χ^2 = 0.53 (Figure 5.20A, green curve). To achieve this fitting, it was necessary to scale the intensity axis. The same scaling factor was applied to the population profile of this species computed by COSMiCS, showing a perfect match (5.20B, green hollow spheres). The need for scaling comes from the inherent rotational ambiguity of the system that appears where closure cannot be applied, as it is

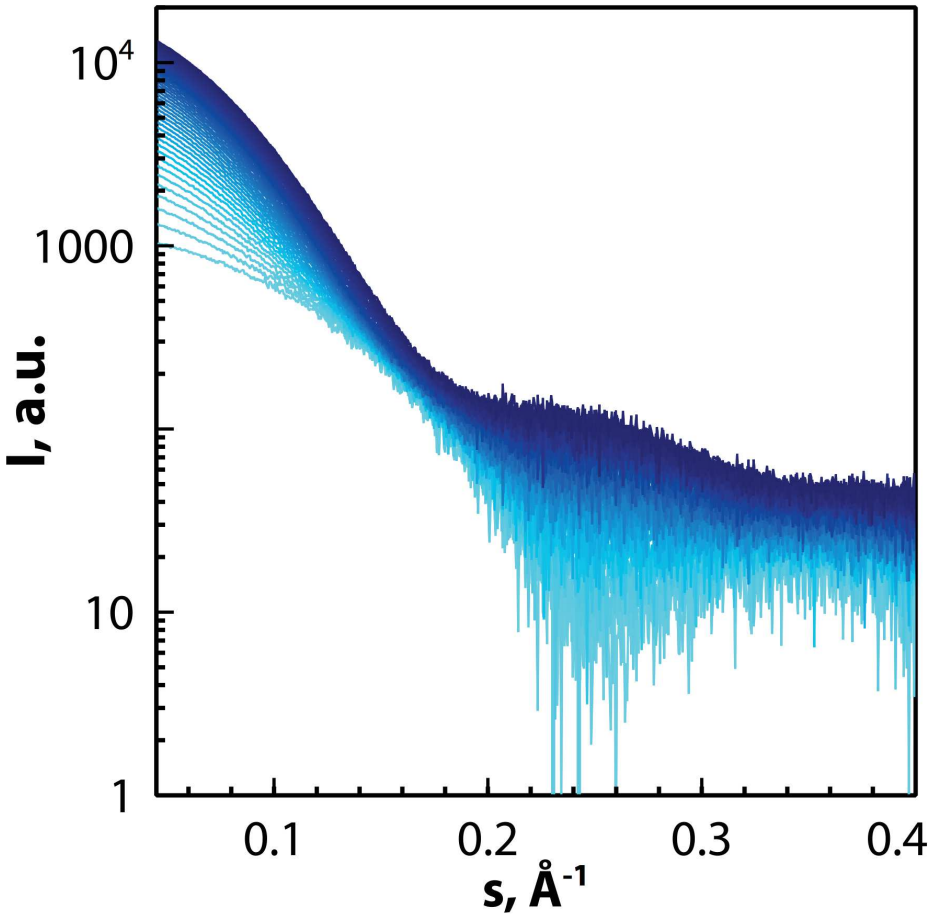


FIGURE 5.19. *Complete synthetic dataset (50 curves) in semi-logarithmic scale for the transient interaction (K_d = 0.1 μM) between yeast cytochrome c peroxidase and yeast iso-1 cytochrome c.*

the case for all titration systems. The first and third species do not require scaling because their curves were fixed during the optimization.

5.7.2.2.2 COSMiCS analysis without fixing curves

In order to evaluate the restriction power of fixing individual curves, another COSMiCS analysis was performed to the same dataset, with the same constraints than the previous one but without fixing any pure spectra. The resulting χ^2 from the different data representations are displayed in the Table 5.2. In contrast with the previous case, a better agreement to the complete dataset is observed when no additional information is added to the optimization. The reason for this is the larger number of degrees of freedom, so the program is able to find a solution that explains better the experimental data. All combinations with χ^2 smaller than 1.0 present almost equivalent results. The results of the Absolute data are presented in Figure 5.21A, fitted with their corresponding original curve used to generate the dataset. The pure curve of the AB complex has perfect agreement with the original, χ^2 = 0.43 (Figure 5.21A, green curve). However, the fitting of the decomposed curve for the subunits A and B shows large discrepancies. The Monte Carlo analysis shows a big error bars for the subunit B curve and its associated population profile. For that reason, the χ^2

Representations included						
Code	**Absolute I(s)**	**Holtzer I(s)*s**	**Kratky $I(s)*s^2$**	**Porod $I(s)*s^4$**	**$\langle\chi_i^2\rangle$ Fixing curves**	**$\langle\chi_i^2\rangle$ Without fixing**
A	+				**1.29**	1.06
AH	+	+			1.29	1.06
AK	+		+		1.25	**1.24**
AP	+			+	1.41	1.12
AHK	+	+	+		1.25	1.20
AHP	+	+		+	1.39	1.07
AKP	+		+	+	1.39	1.06
AHKP	+	+	+	+	1.39	1.03

TABLE 5.2. *Results of the COSMiCS analysis of the synthetic dataset for titration case representing a transient biomolecular interaction.*

of the fitting with the original curve is very low, $\chi^2 = 0.13$, even when is possible to see from the inspection of the fitting that the agreement is not good (Figure 5.21A, blue curve). The subunit A also presents a bad agreement with the theoretical curve, and it is reflected in the high $\chi^2 = 7.26$, mainly due to the smaller error bar associated (Figure 5.21A, orange curve). The bad agreement of the AB complex and the smaller species is also reflected in the population (Figure 5.21B), which shows also a bad agreement and large uncertainties.

These results highlight the importance of introducing additional information, such as kinetic/thermodynamic model or pure spectra curves. Importantly, this latest information is available in titration experiments and it is implemented in COSMiCS.

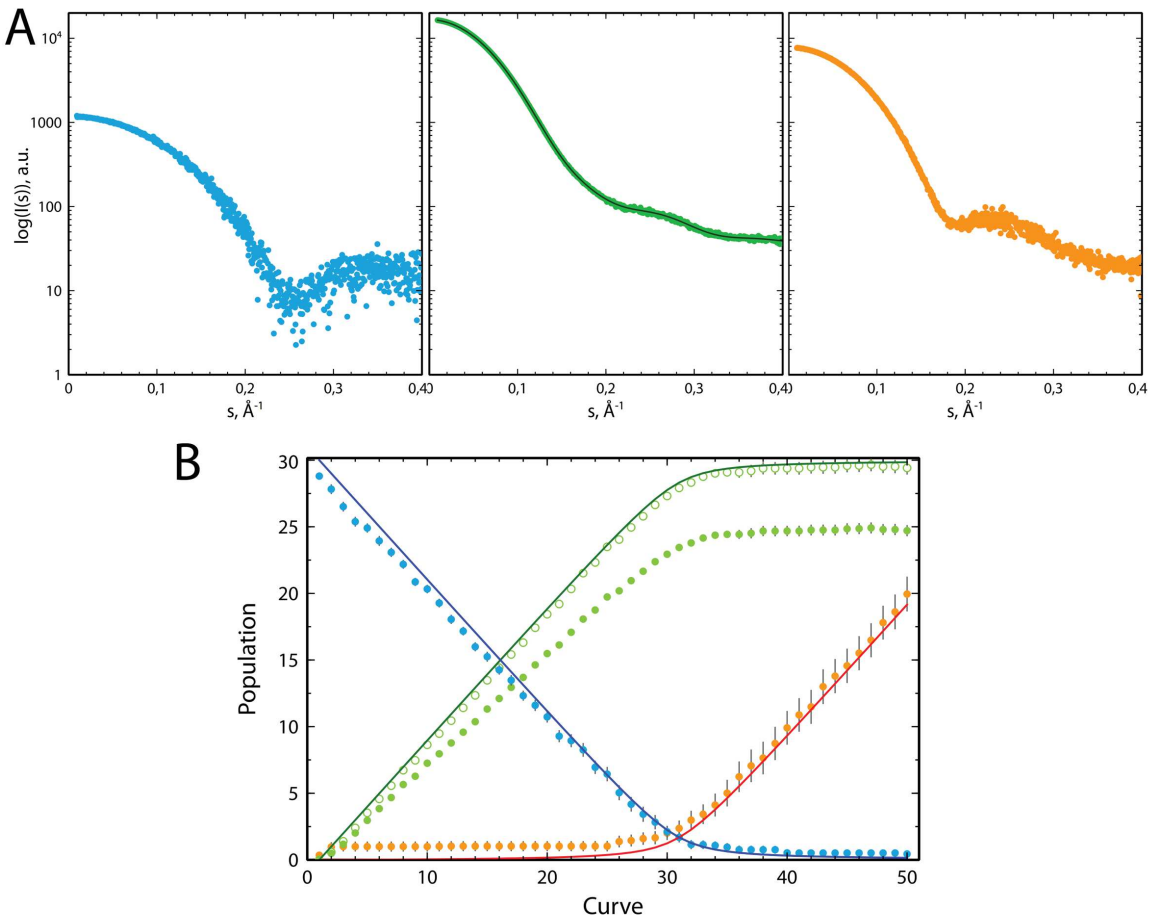


FIGURE 5.20. *Results of the COSMiCS analysis of the system A + B ⇌ AB for the AK combination, fixing the curves from known species (subunit A and B). A) Pure spectra of the subunit A (orange), subunit B (blue) and complex AB (light green), the latter one is fitted with the original curve computed by CRYSOL used to generate the dataset (dark green). The curves are in different plots for a better inspection. B) In the same color code, the original populations used to create the dataset in solid lines. The concentration profile decomposed by COSMiCS is shown in solid circles with the error bars from the Monte Carlo analysis. The green hollow circles represent the population of the complex AB using the same scale factor applied to fit the spectra. This uncertainty in the scale arises from the rotational ambiguity.*

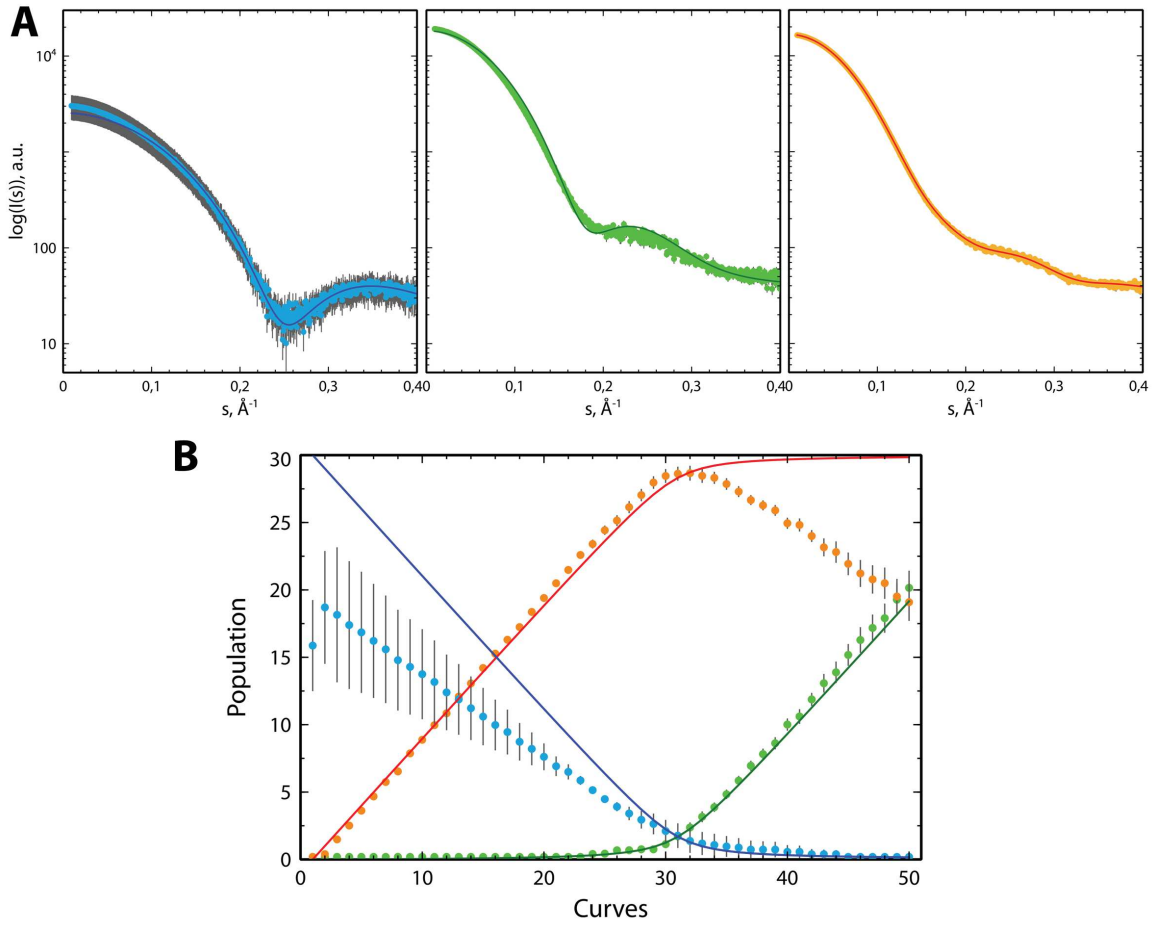


FIGURE 5.21. *Results of the COSMiCS analysis of the system $A + B \rightleftharpoons AB$ using just the Absolute values, without fixing curves. A) Pure spectra of the subunit A (orange), subunit B (blue) and complex AB (light green), fitted with their corresponding original curves computed by CRYSOL used to generate the dataset (solid lines). The curves are in different plots for a better inspection. B) In the same color code, the original populations used to create the dataset in solid lines. The concentration profile decomposed by COSMiCS is shown in solid circles with the error bars from the Monte Carlo analysis.*

Chapter 6

Chemometrics analysis of SEC-SAXS data from mixtures

6.1 Introduction

Structural analysis of SAXS data requires scattering profiles corresponding to a single species or to a well-defined mixture of species. In the previous sections I have explained how to decompose data from polydisperse samples using COSMiCS. However, this approach is not common and, in some circumstances, the experiments can be difficult to perform from a practical point of view. As the scattering from a particle is related to the square of its excess scattering length density, aggregates contribute disproportionately to the signal and in most cases render the data non-interpretable. In practice, samples should be at least 95% monodisperse in order to obtain useful information. A great effort has been done in the recent years to implement online size exclusion chromatography coupled to SAXS (SEC-SAXS) in BioSAXS beamlines to achieve this required monodispersity.

Recently, several beamlines have combined SAXS with online SEC purification as a standard set-up [70, 73, 74] (see section 1.3.2.3). This combination is only possible at 3rd generation synchrotrons, where statistically significant data over a broad s-range (0.005 – 0.5 Å^{-1}) can be recorded in less than a second for concentrations below 1 mg/ml. Extensive SAXS sampling of the HPLC elution peaks can then be performed, similarly to the popular SEC-MALLS combination. Therefore, these experiments provide information for individual species. SEC is the most common chromatography coupled to SAXS due to the homogeneity of the buffer. Recently, efforts to use affinity or ionic columns, where components are eluted using buffers with variable composition, have been performed [329].

In some cases, the chromatographic peaks can be still overlapped after elution from the column, yielding polydisperse SAXS data. In these cases, decomposition of the data is necessary in order to structurally analyze the species present in the system. One strategy to decompose this data has been implemented in the software US-SOMO [78, 79], which has been briefly described in the section 1.3.2.3.

In this chapter we explore the use of COSMiCS (chapter 5) to decompose data from SEC-SAXS experiments composed by a mixture of species. In order to be able to analyze the power of COSMiCS in decomposing these data, as well as to test the limits of the method, we created synthetic datasets that simulate SEC-SAXS experiments with overlapped peaks.

Finally, real experimental data from a SEC-SAXS experiment will be presented as an example of the application of this strategy.

6.2 Results

6.2.1 Generation of synthetic data

We have generated a synthetic SAXS dataset that emulates a SEC-SAXS experiment of a protein that forms dimers that are in equilibrium with the monomeric species. A population model was created using as original concentration profiles two overlapped Gaussians corresponding to the monomer and the dimer (Figure 6.1B). The Gaussian corresponding to the dimer starts at frame 100 and finishes at frame 700, with a maximum in frame 400, and the monomer population starts at frame 400 and finish at frame 1000, with a maximum at frame 700. The structures used to create the synthetic data are the same than these used in section 5.6, corresponding to the pdb codes 4QHF for the monomer and 4QHG for the dimer [80]. Their theoretical scattering profiles were computed with CRYSOL [41] using the maximum number of spherical harmonics and maximum order of the Fibonacci's grid representations. All other parameters were in default setting (Figure 6.1A). The primary analysis of the two species was performed using the program PRIMUS [15], obtaining a R_g(dim) = 21.10 Å and R_g(mon) = 15.47 Å, and a ratio $I(0)_{dim}$ / $I(0)_{mon}$ = 3.6.

The dataset of scattering curves was generated as the concentration weighted sum of the species present in each frame of the experiment. A total of 1200 curves were created (Figure 6.1C). Noise was added to the dataset using the experimental error values, $\sigma(s)_{exp}$, from a SEC-SAXS study of protein POP [331] (see section 6.3) in the following way. Synthetic curves were scaled such that the maximum I(*0*) among all the theoretical curves corresponds to the maximum I(*0*) among all the experimental curves. Importantly, the concentration along the dataset is changing, therefore the curves have a different level of noise depending on their concentration. In order to add a concentration-dependent level of noise, the I(*0*) of each synthetic curve was calculated (Figure 6.1D) and the error of the experimental curve with the closest I(*0*) was assigned. For the inclusion of noise in the simulated scattering curve, values from a Gaussian distribution with a standard deviation, $\sigma(s)_{exp}$, centred around I(s) were added to the curve (Figure 6.1C). The resulting level of noise is high due to the low concentration of the experimental data used. In order to test the effect of the level of noise in the decomposition using COSMiCS, two additional datasets were generated with reduced levels of noise. These datasets were calculated multiplying $\sigma(s)_{exp}$ by 0.2 ($\sigma_{0.2}$) to obtain a dataset with high signal-to-noise level, and by 0.5 ($\sigma_{0.5}$) to obtain a medium signal-to-noise level.

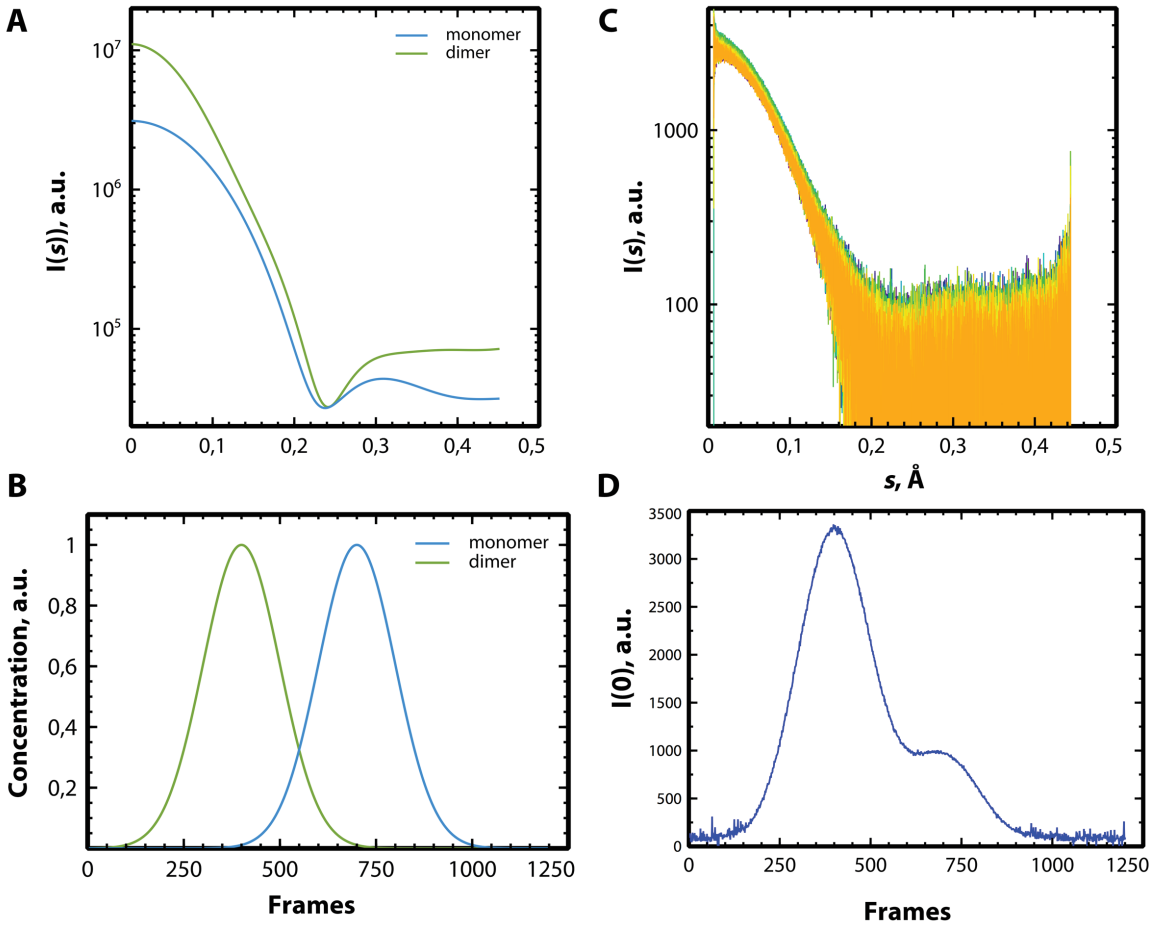


FIGURE 6.1. *Synthetic datasets. A) Synthetic curves computed with CRYSOL for the monomer (blue) and dimer (green). B) Gaussians corresponding to the individual populations of each species simulating a SEC-SAXS experiment with the same colour code than panel A. C) Final dataset with noise ($\sigma_{0.2}$) in semi-logarithmic scale (only frames from 350 to 450 are showed for clarity). D) I(0) of the complete synthetic dataset displayed in panel C and computed with autorg from ATSAS package [330].*

6.2.2 COSMiCS analysis of synthetic data

COSMiCS was used to analyze the three synthetic datasets, using only *Non-negativity* as constraint. The first step, like in every decomposition approach, was performing a Principal Component Analysis (PCA) to the complete dataset to determine the total number of species present. PCA analysis identified two species for the three datasets. Although the inspection of the eigenvalues and eigenvectors indicates the presence of two species in the datasets with the lowest noise level (Figure 6.2A), the dataset with lowest signal-to-noise level (Figure 6.2B and 6.2C) presents a large value for the third eigenvalue, which originates from the more important amount of noise.

The second step, previous to the ALS optimization, is the selection of initial estimations. We have used the approach based in SIMPLISMA [287], like in the COSMiCS analysis (section 3.3). In the case of the datasets with lower noise ($\sigma_{0.2}$), the software selects as initial estimation for the decomposition frames 396 and 724. These frames are very close

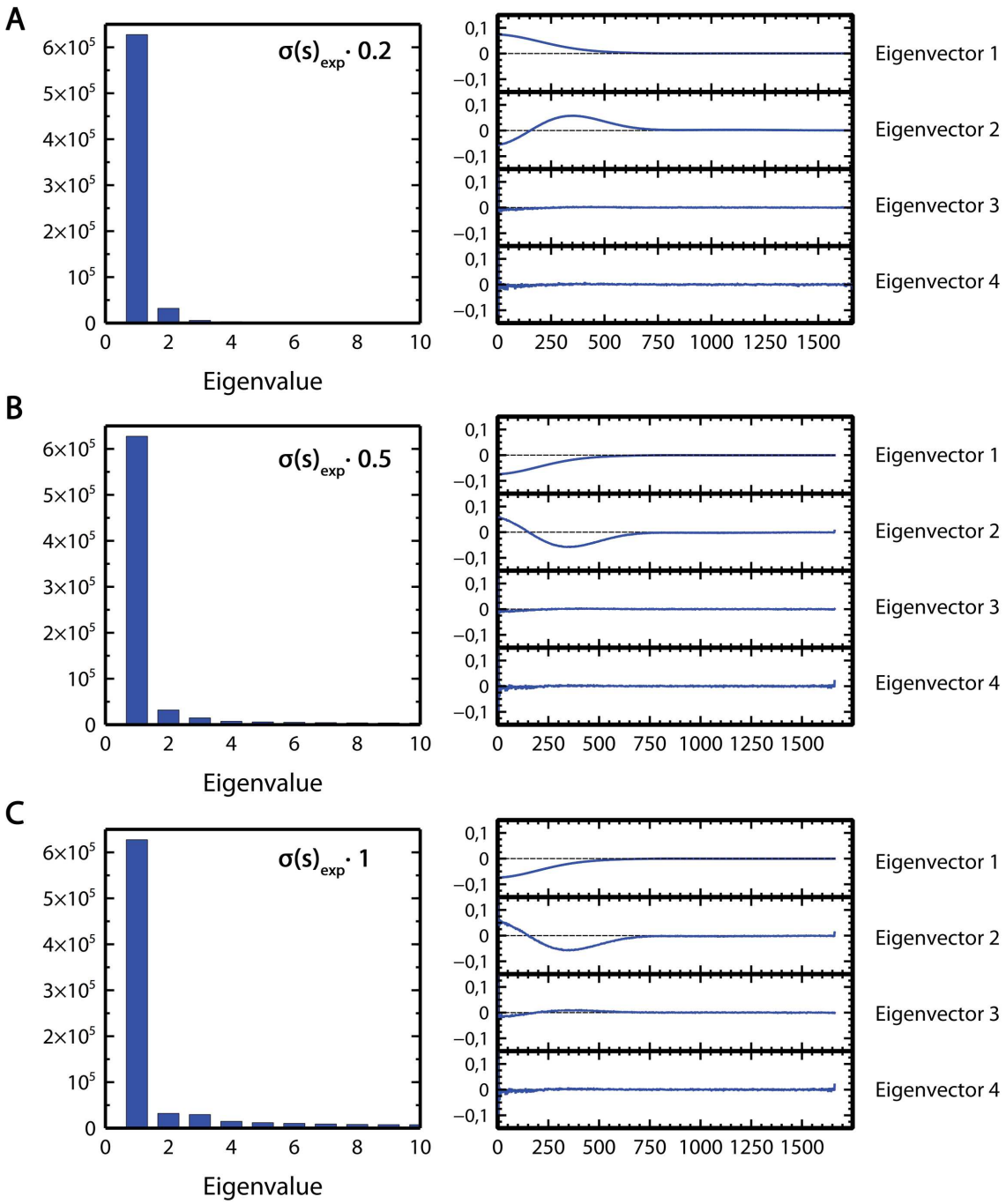


FIGURE 6.2. *Results from PCA for the three synthetic datasets with increasing amount of noise. First ten eigenvalues and four first eigenvectors for A) low level of noise ($\sigma_{0.2}$), B) medium level of noise ($\sigma_{0.5}$) and C) high level of noise ($\sigma_{1.0}$). The inspection of eigenvalues and eigenvectors allows the estimation of the number of coexisting species.*

to the center of each of the Gaussians (400 and 700). However, when the noise is higher ($\sigma_{0.5}$), the program selects frames 385 and 1057 as initial estimations for the decomposition. Frame 1057 corresponds to a zone of buffer and it is not a good initial estimation. For unifying all the analyses, we have selected the same frames as initial estimations for every dataset: frames 400 and 700, which correspond to the frames with the highest protein

Analysis	R_g(dim)	R_g(mon)	$I(0)_{dim}/I(0)_{mon}$	C(dim)	C(mon)	$P^{400}_{dim}/P^{700}_{mon}$
Theoretical	**21.10**	**15.47**	**3.6**	**-**	**-**	**1**
Noise $\sigma_{0.2}$	21.04	15.31	3.5	62	94	1.03
Noise $\sigma_{0.5}$	21.05	14.97	3.5	103	96	1.11
Noise $\sigma_{1.0}$	21.30	13.47	3.6	103	101	1.52

TABLE 6.1. *Analysis of the COSMiCS decomposed scattering curves and concentration profiles from datasets with increasing levels of noise compared with the theoretical values used to generate the data.*

concentration for the monomer and the dimer, respectively (Figure 6.1D).

After the COSMiCS decomposition, a Monte Carlo analysis (see section 5.6.1) was performed to obtain the error bars of the curves. Notice that these error bars cannot be used to compare with the theoretical ones using the traditional goodness of fit analysis, χ^2, because this statistical value depends on the size of the error bars, which will be different for the different analyses. However, the error bars are useful to monitor the uncertainties of the solution. The parameter that we used to check the validity of the solution is the C-value, obtained with the software CorMap [7] (See section 1.2.2), which is the highest number of consecutive points where the fit is systematically higher (or lower) than the experimental data. However, the best approach to check the accuracy of the solution is the visualization of the ratio between the theoretical curve and the decomposed curve by COSMiCS ($I(s)_{COSMiCS}/I(s)_{theoretical}$). To monitor the accuracy of the populations derived from the analyses, besides the overall inspection, the ratio between the maximum concentration of both species, situated in frames 400 and 700 for the dimer and monomer respectively, was performed and compared with the theoretical one. Finally, an analysis of the decomposed SAXS curves was performed in order to compare R_gs and the ratio $I(0)_{dim}/I(0)_{mon}$ with the theoretical values. A comparison table with these parameters from the decomposition of the three datasets, and the theoretical one, are shown in Table 6.1.

COSMiCS decomposition for the $\sigma_{1.0}$ dataset was performed for all the combinations of SAXS data representations. The decomposed solutions for the spectra and the populations were very similar for all the combinations. The ratios between the original and decomposed curves ($I(s)_{COSMiCS}/I(s)_{theoretical}$) for each combination are shown in Figure 6.3, and they reveal that the decomposed curves do not agree with the theoretical ones. The decomposed curves show a bad agreement with the original curves (Figure 6.4A), with R_g(dim) = 21.30 Å and R_g(mon) = 13.47 Å, which is notably lower than the R_g of the original monomer, 15.30 Å. However, the relative ratio $I(0)_{dim}/I(0)_{mon}$ = 3.6 is the same than in the original curves. The decomposed peaks are shown in Figure 6.5. All the combinations have similar concentration profiles and they present an additional contribution of the second species (monomer) in the first 400 frames of the dataset that was not present in the

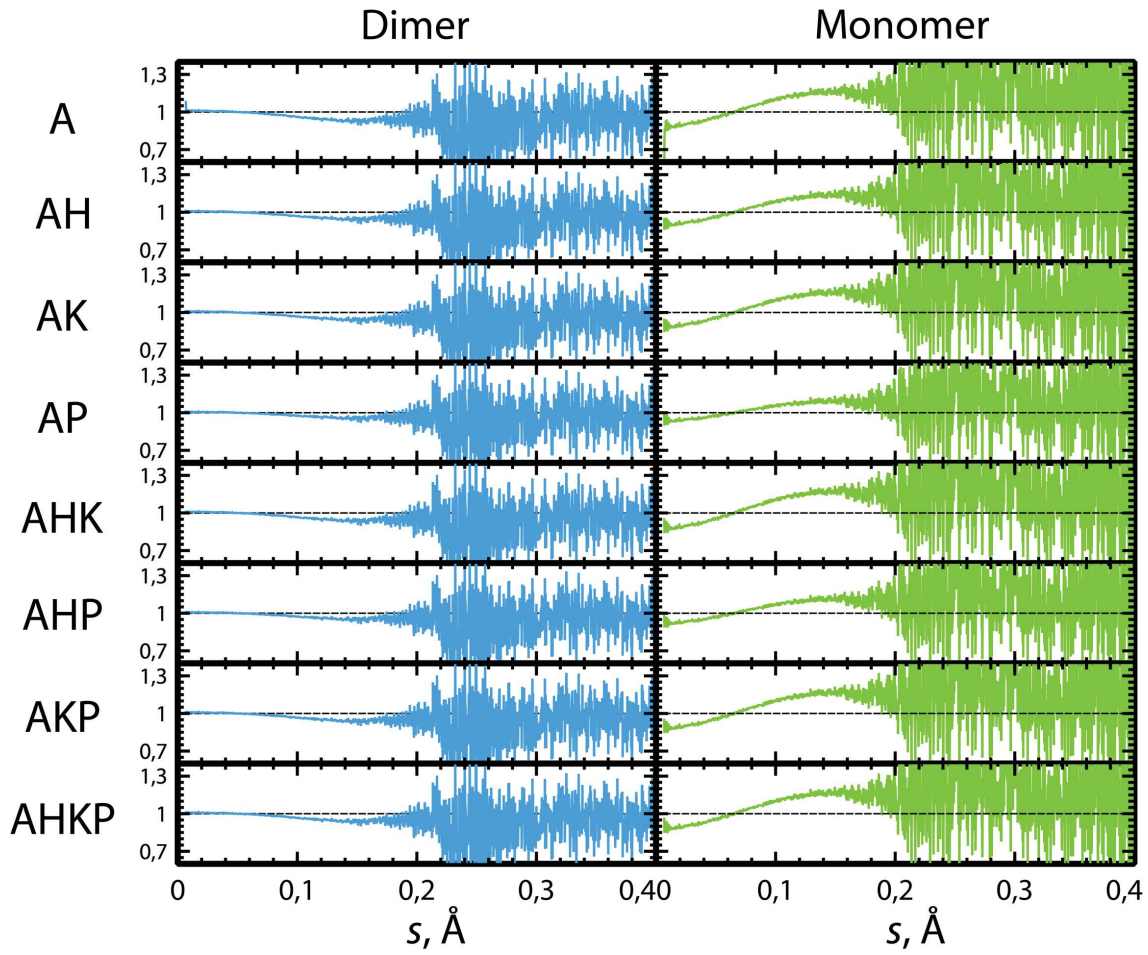


FIGURE 6.3. *COSMiCS analysis for the monomer-dimer synthetic dataset with low signal-to-noise ($\sigma_{1.0}$). Ratios $I(s)_{COSMiCS}/I(s)_{theoretical}$ for the dimer (blue) and monomer (green) for each of the combinations of representations. All the solutions are equivalent and do not agree with the theoretical curves.*

theoretical population. The decomposed concentration (Figure 6.4B) is also wrong, especially in the monomer concentration, which is smaller than the original one. It is noticeable that the decomposed concentration using only the absolute scale shows more noise than the concentrations from the decomposition using one or more representations simultaneously.

In order to compare the effect of the level of noise in data decomposition, we performed a COSMiCS analysis of the other two datasets: $\sigma_{0.2}$ and $\sigma_{0.5}$. Figure 6.6 shows the results of the analysis for the $\sigma_{0.2}$ dataset, using only Absolute values to simplify the examination of the results. The agreement with the original SAXS curves (Figure 6.6A) is perfect, which is obvious after the visual inspection of the ratio $I(s)_{COSMiCS}/I(s)_{theo}$. The analysis of the decomposed spectra gives R_g(dim) = 21.04 Å, R_g(mon) = 15.31 Å and the ratio $I(0)_{dim}/I(0)_{mon}$ = 3.5, all values very close to the original ones. The concentration profiles of the COSMiCS solution (Figure 6.6B) show also a good agreement with the original populations. A larger uncertainty (error bars) was observed in the overlapped part of the populations, especially for the monomer due to its smaller size.

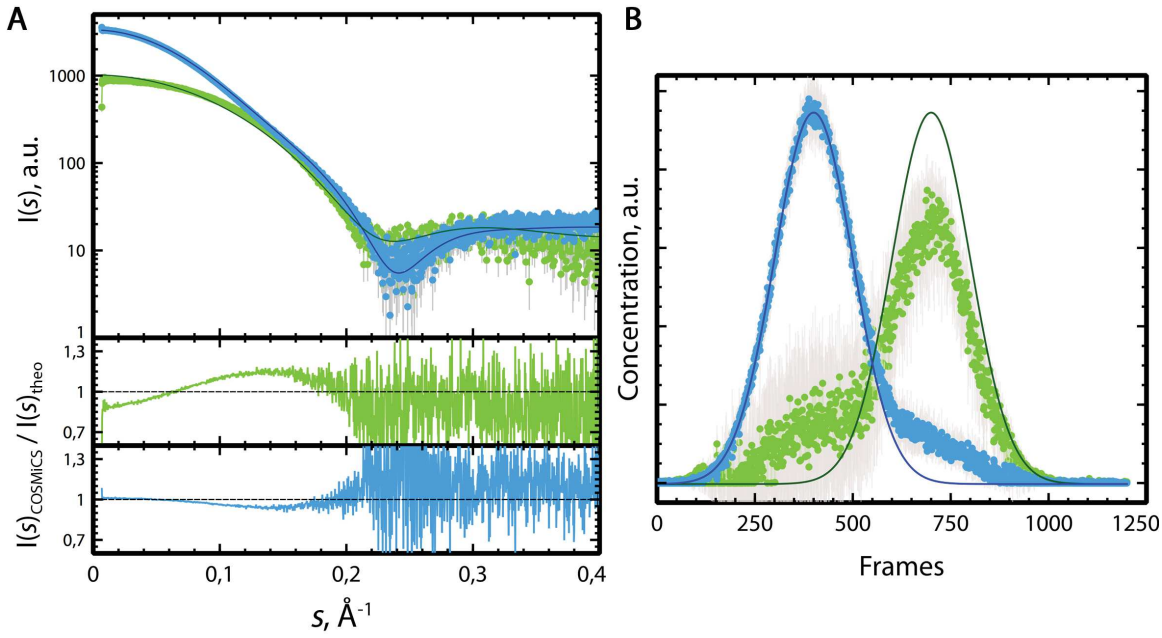


FIGURE 6.4. *Results from the COSMiCS analysis of the synthetic SEC-SAXS dataset using the final noise ($\sigma_{1.0}$). A) Pure spectra of the monomer (green) and dimer (blue) and their fitting with the original curves computed with CRYSOL. At the bottom, the ratios $I(s)_{COSMiCS}/I(s)_{theoretical}$ for both species are shown. B) Concentration profiles of the two species (same color code) and the original populations used for generate the synthetic data*

When we performed the same analysis with data with medium signal-to-noise ($\sigma_{0.5}$) using Absolute values, the solution obtained with COSMiCS was still in good agreement with the original curves (Figure 6.7A), although slightly worse compared with the $\sigma_{0.2}$ case. R_g(dim) = 21.05 Å, R_g(mon) = 14.97 Å, and the ratio $I(0)_{dim}/I(0)_{mon}$ = 3.5, all values still close to the original curves. The concentrations (Figure 6.7B) also worsen, with a smaller monomer concentration compared with the theoretical one used to generate the data. In the same way than the previous analysis, the decomposed data have an increase of the ambiguity in the overlapped part, which also presents an enhanced contribution of the monomer that was not present in the theoretical population.

6.2.3 Adding more information to the system: Equality constraint

We have seen in the previous section that, if the noise is high, we cannot decompose correctly neither the pure species nor the concentration profiles. This is due to the inherent ambiguity of the method (see section 3.3.1). In order to decrease the ambiguity, it is necessary to add more information to the decomposition process. We performed the COSMiCS analysis adding information about the part of the dataset composed by a single species (monodisperse zone). This information can be introduced in COSMiCS as *Equality* constraint for concentrations (section 3.3.2.4).

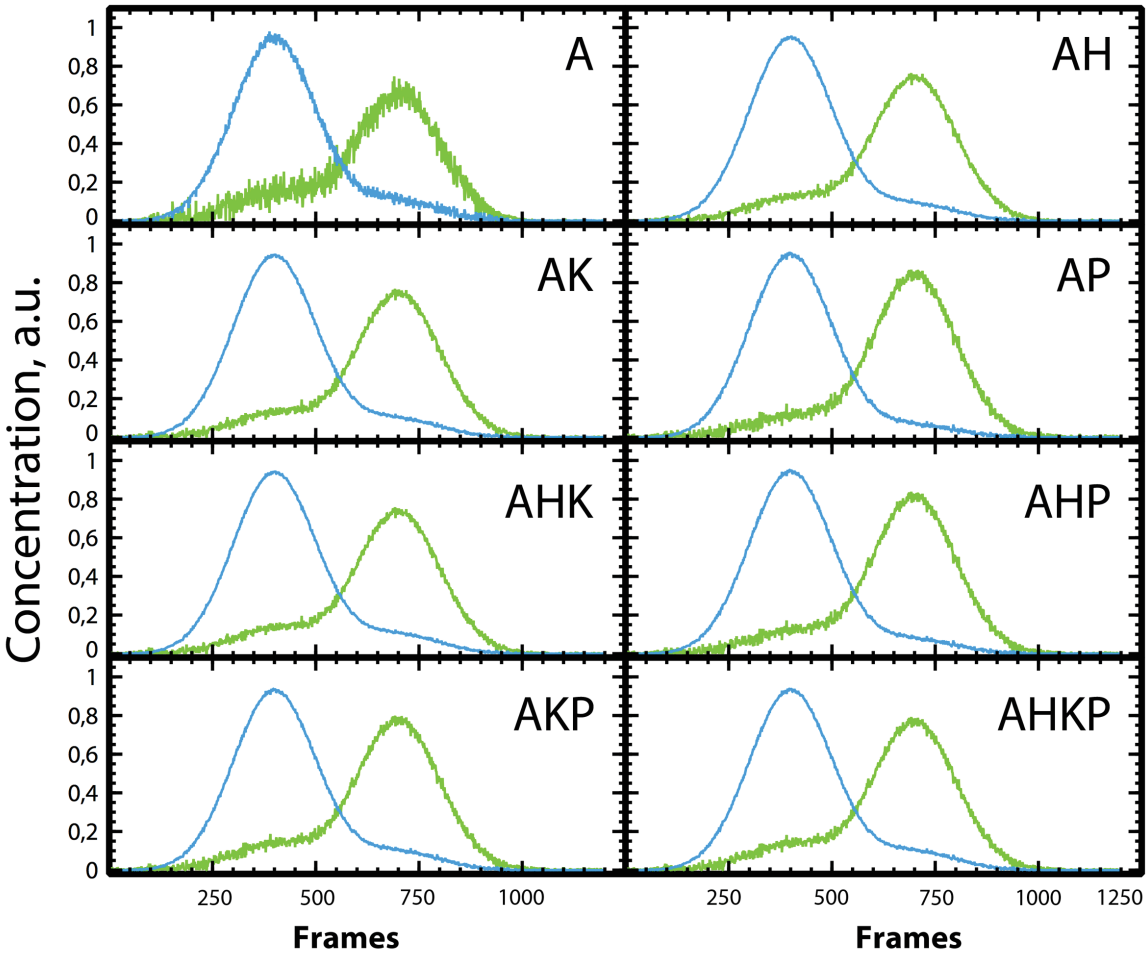


FIGURE 6.5. *COSMiCS population profiles for the monomer-dimer synthetic dataset with low signal-to-noise ($\sigma_{1.0}$). Decomposed population of dimer in blue and monomer in green for each of the combinations of representations. All the solutions are equivalent, except for the Absolute scale, which presents a more important level of noise. None of the combinations agree with the theoretical population profiles.*

6.2.3.1 COSMiCS analysis adding theoretical *Equality* constraint

To assess the ambiguity reduction capacity, the *Equality* constraint was applied into the COSMiCS analysis for the decomposition of the dataset with low signal-to-noise ($\sigma_{1.0}$). For that, we introduced in the program the information about which species are present in each of the frames. According to the original populations, we defined that the first species start in frame 1 and ends in frame 700, and the second species start in the frame 400 until the last frame (1200).

The results of this analysis showed an important improvement of the decomposition (Figure 6.8A), with a better agreement with the original curves than in the previous analysis without constraints (Figure 6.7A), with a ratio of intensities centred in 1. R_g(dim) = 20.98 Å and R_g(mon) = 15.58 Å, which are also better than the previous unconstrained analysis. The description of the population (Figure 6.8B) also improves respect the previous analysis as it does not present artifacts in the concentration profiles. However, the population profiles do

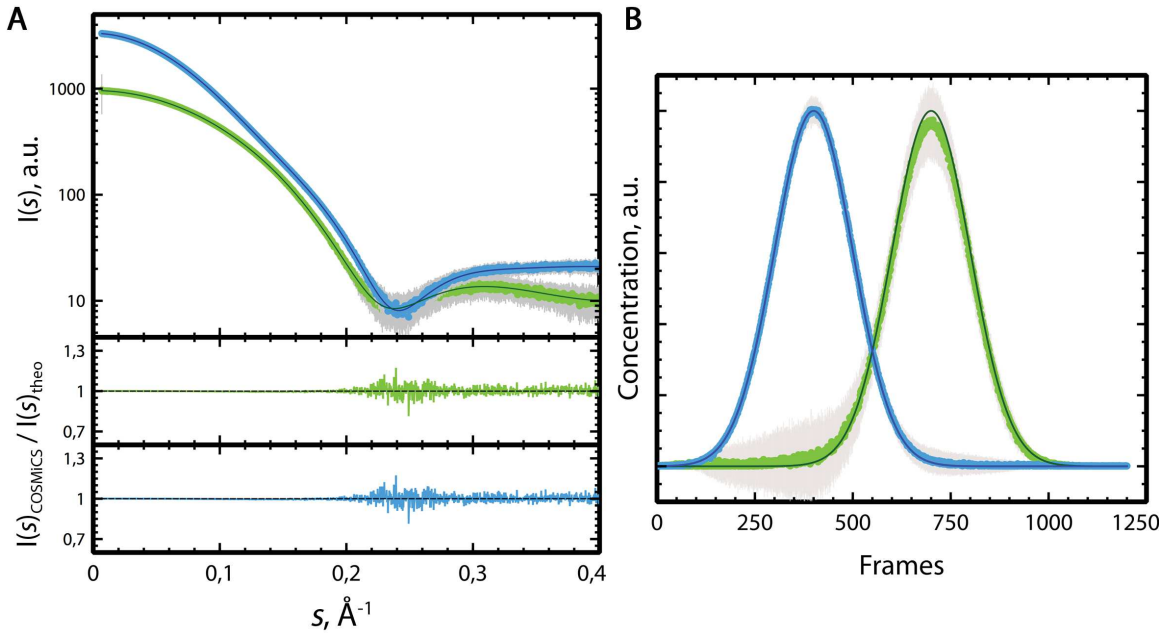


FIGURE 6.6. *Results from the COSMiCS analysis of the synthetic SEC-SAXS data using a scale of 0.2 of the final noise ($\sigma_{0.2}$). A) Pure spectra of the monomer (green) and dimer (blue) and their fitting with the original curves computed with CRYSOL. At the bottom, the ratios $I(s)_{COSMiCS}/I(s)_{theoretical}$ of both species are shown. B) Concentration profiles of the two species (same color code) and the original populations used for generate the synthetic data*

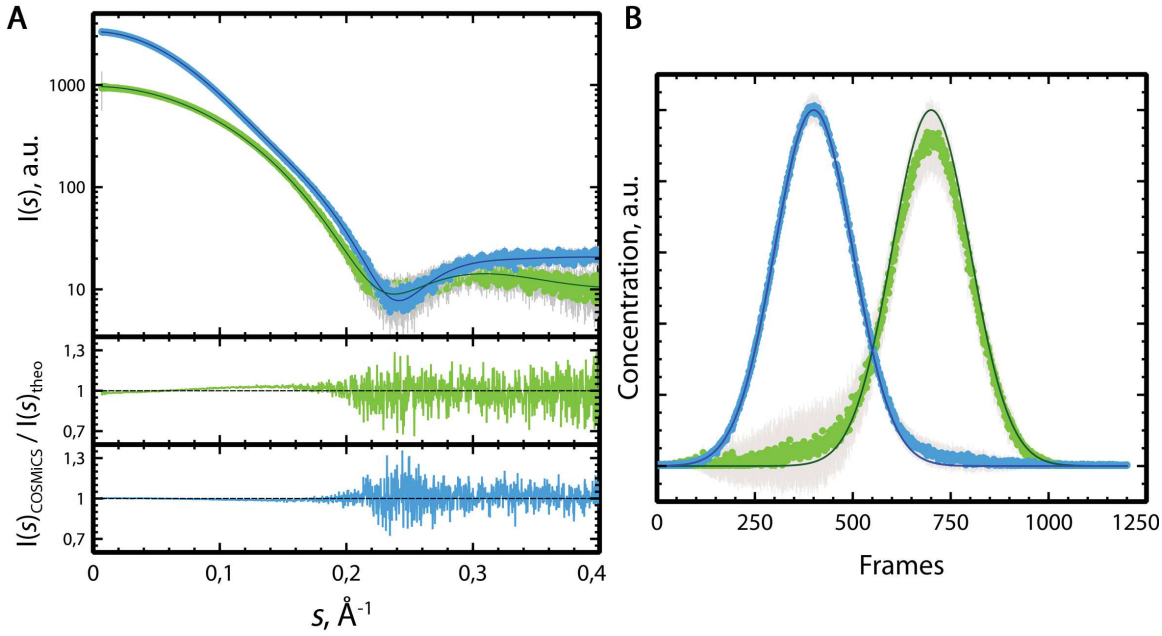


FIGURE 6.7. *Results from the COSMiCS analysis of the synthetic SEC-SAXS data using a scale of 0.5 of the final noise ($\sigma_{0.5}$). A) Pure spectra of the monomer (green) and dimer (blue) and their fitting with the original curves computed with CRYSOL. At the bottom, the ratios $I(s)_{COSMiCS}/I(s)_{theoretical}$ of both species are shown for a better inspection. B) Concentration profiles of the two species (same color code) and the original populations used for generate the synthetic data*

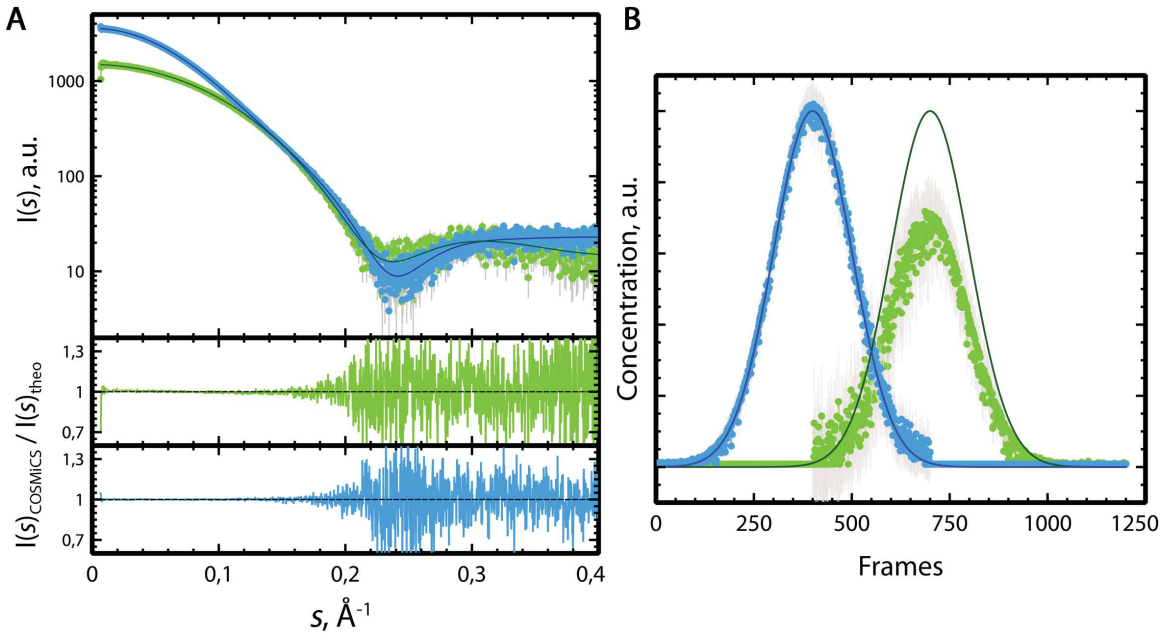


FIGURE 6.8. *Results from the COSMiCS analysis of the synthetic SEC-SAXS dataset using the final noise ($\sigma_{1.0}$), adding an equality constraint that indicates that the dimer ends at frame 700 and monomer starts at frame 400. A) Pure decomposed curves for the dimer (blue) and monomer (green). In solid line is shown the original SAXS curves of the two species (same color code). At the bottom, the ratios $I(s)_{COSMiCS}/I(s)_{theoretical}$ of both species are shown B) Concentration decomposed profile of the two species.*

not match with the original populations, especially for the monomer profile. In this case, the rotational ambiguity gives still a bad determination of the I(*0*), giving a ratio between the I(*0*) of two species of 2.4 instead 3.6 of the original curves, although better than in the analysis without *Equality* constraint.

6.2.3.2 Determining monodisperse zones

Considering the results obtained, it seems important to identify the monodisperse zones along a chromatography elution in SEC-SAXS experiments to be able to precisely add this information to the system, in order to obtain better solutions. Different approaches have been proposed to obtain the zones of monodispersity. The methods that we tested with our synthetic dataset were R_g, which is the most commonly used, and the Evolving Factor Analysis (EFA) (section 3.4), that has been recently proposed to analyze monodispersity in SEC-SAXS data.

6.2.3.2.1 Determining monodisperse zones using R_g

The most common strategy to determine the monodisperse regions in SAC-SAXS is calculating the R_g for each frame of the dataset and monitoring its changes. These zones with a constant R_g are considered to be monodisperse. Figure 6.9 shows the R_g from the three synthetic datasets used in previous sections. The R_g calculation presents higher uncertainty when the signal-to-noise of the data decreases. However, the general shape remains constant independently of the level of noise of the dataset.

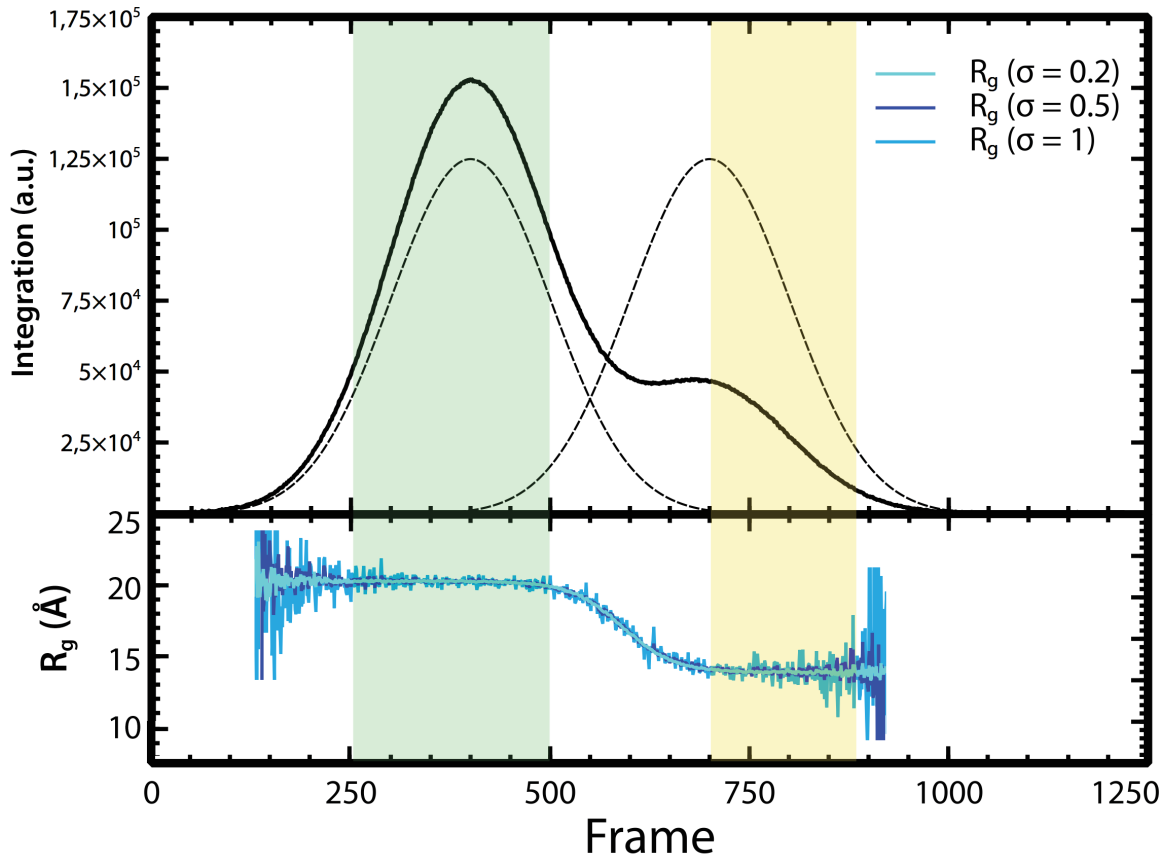


FIGURE 6.9. *Integration of first 50 I(s) points of the SAXS data (solid line) and original Gaussian populations used to generate the synthetic data (dashed lines). R_g (bottom) along the frames of the dataset computed with autorg from the ATSAS package [ref] for the three datasets with different noise levels. The green and yellow areas display the monodisperse regions for the dimer and monomer, respectively, according to the R_g analysis.*

According to the R_g profile, it is possible to determine a monodisperse zone, where the R_g is stable, in the range of frames from 250 to 500 corresponding to the dimer. Importantly, the contribution from the second peak (monomer) in the frames 400 to 500 does not cause a noticeable reduction of the R_g. This is because the concentration of the monomer is relatively low with respect to the concentration of dimer in this region and the contribution of monomer (smaller species) is not detectable. The R_g profile has a second stable zone from frames 700 to 900, which corresponds to the presence of a single monomeric species. Interestingly, this second stable R_g zone appears just when dimer population disappears. This behavior is expected due to the larger size of the dimer, which provokes a more important contribution to the mixed curves at smaller concentrations.

6.2.3.2.2 Determining monodisperse zones using EFA

Another method used to determine the monodisperse regions is EFA (see section 3.4), a chemometric method that applies PCA to estimate the number of species present in different regions of the dataset. To this aim, the eigenvalues of the PCA are monitored upon

sequentially increasing the size of the dataset. The inspection of the eigenvalue profiles enables to detect the appearance of new species along the column elution.

First, we need to understand how EFA works for SEC-SAXS data, and how to interpret the information that we obtain from it. For that, we created a simplified synthetic dataset without noise, in order to see the results without any noise-related artifact. Figure 6.10 shows the results of this EFA analysis, the first (blue) and second (green) eigenvalues in the forward and backward directions. The size of the eigenvalues increases when signal appears, indicating the presence of the first species in the case of the first eigenvalue, and the presence of the second species for the second eigenvalue. In forward direction the first species is the dimer and the second is the monomer, and vice versa for the backward direction. It is very clear in Figure 6.10, where the first eigenvector in the forward direction starts growing exactly in the frame than the dimer peak starts and in backward direction the eigenvalue increases when the monomer appears. Interestingly, the first eigenvalue in backward direction (blue dotted line) is affected by the presence of the dimer, which has larger size. Conversely, the first eigenvalue in the forward direction is not affected by the presence of the smaller monomer. Importantly, the second eigenvalue follows perfectly the increase of the theoretical populations in both directions when the second species appears, and it is therefore the profile that must be followed to identify the monodisperse zones in a SEC-SAXS experiment.

The analysis of datasets without noise is useful to illustrate the behavior of the EFA profiles along a SEC-SAXS dataset; however, that is not the real situation that we can find in experimental data. We have performed the EFA analysis in the three datasets with different levels of noise to see how this method behaves in the presence of increasing amounts of noise. Figure 6.11 shows the EFA for the first eigenvalue (forward and backward) for the different datasets, which corresponds to the appearance of signal. The method detects the signal when it rises above the noise level, and for that it requires the inclusion of more frames in the analysis when the signal-to-noise is lower. In other words, the change in the slope in the eigenvalue appears systematically later as we increase the noise in the dataset for both the forward and the backward directions. According to this analysis, it is clear that the level of noise affects the capacity of the method to detect the presence of signal.

An equivalent analysis was done with the second eigenvector profile, which is more informative because it indicates the appearance of a second species in the dataset, and allows the identification of monodisperse zones. Figure 6.12 shows the EFA for the second eigenvalue in the forward and the backward directions. The change in the slope of the second eigenvalue profile indicates the appearance of a second species, and it is indicated by a circle in the forward direction and by a square in the backward direction. The second species corresponds to the monomer or the dimer for the forward and the backward directions, respectively. As in the case of the first eigenvalue, these changes in the EFA profiles appear later on in the datasets as the level of noise increases. When the signal-to-noise is low, the method is less sensitive to the presence of a new species and it is necessary add more frames of the second peak to be able to detect the appearance of the new species. In the case of high signal-to-noise dataset ($\sigma_{0.2}$) the monomer peak appears in the frame 440,

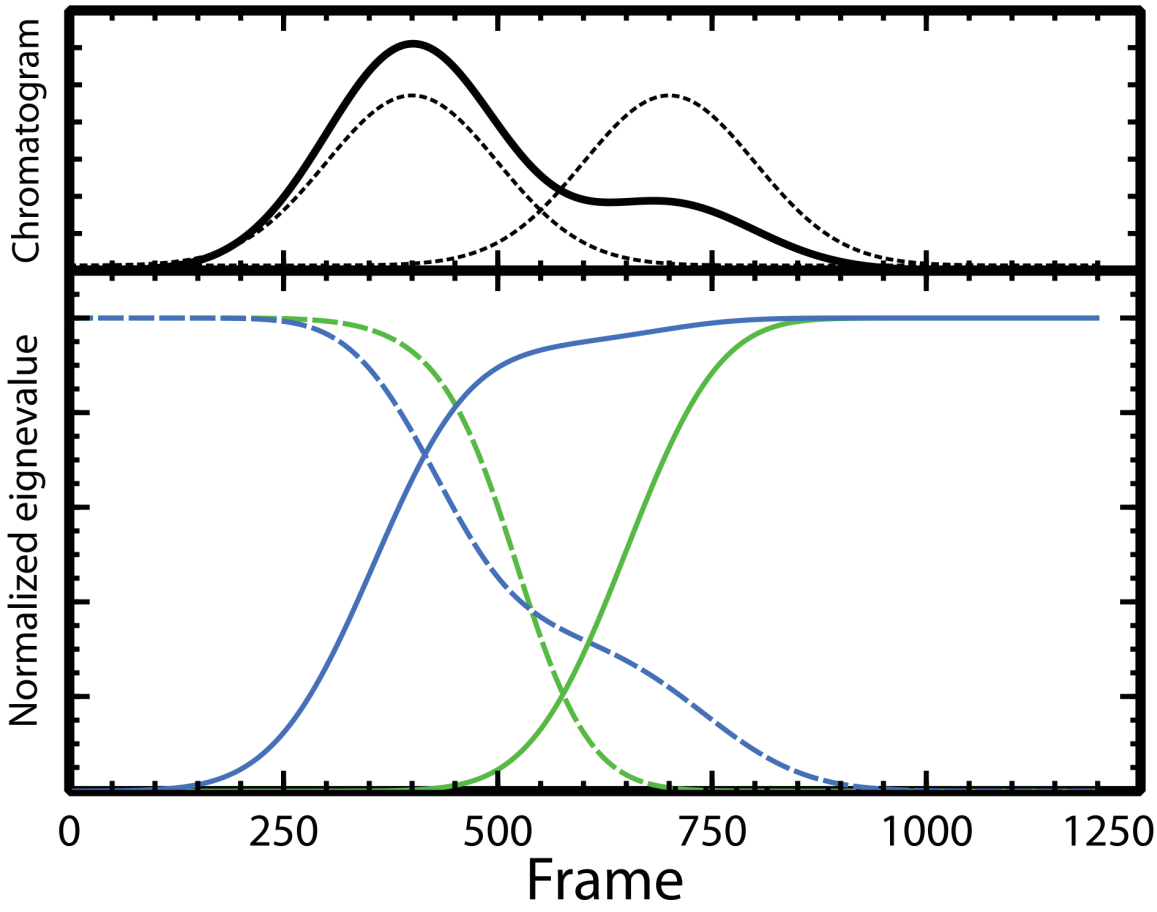


FIGURE 6.10. *EFA of synthetic data (monomer-dimer system) without noise, in the forward (solid line) and backward (dashed line) directions. First eigenvalue (blue) shows when the first peak appears, dimer peak in the forward analysis and monomer peak in the backward analysis. The second eigenvalue (green) increases when the second species appears, monomer in the forward analysis and dimer in the backward analysis. The chromatogram (black solid line) and the original populations (dotted black line) are represented on the top panel.*

very close to the theoretical value of 400. When the noise increases ($\sigma_{0.5}$), it is necessary add more frames to the analysis to detect the appearance of the monomer, which occur at the frame 470. Finally, for the low signal-to-noise dataset ($\sigma_{1.0}$) the change appears even further, at the frame 500. In the backward direction, the change of the slope for the $\sigma_{0.2}$, the $\sigma_{0.5}$, and the $\sigma_{1.0}$ datasets happens at the frame 690, 660, and 630, respectively. Thus, the EFA indicates that for low levels of noise the detection of the populations is very close to the theoretical value, 700. Therefore, the behavior of the second eigenvalue EFA profile with respect to noise level is the same than that of the first eigenvalue, as more frames are needed to detect a new species when increasing the noise level. Importantly, in the backward direction, less frames are necessary to detect the appearance of the dimer (70 frames for the $\sigma_{1.0}$ dataset) than the detection of the monomer in the forward direction (100 frames). The explanation of this difference arises from the difference in size between the two species.

In summary, the EFA is a good method to determine monodisperse zones, obtaining

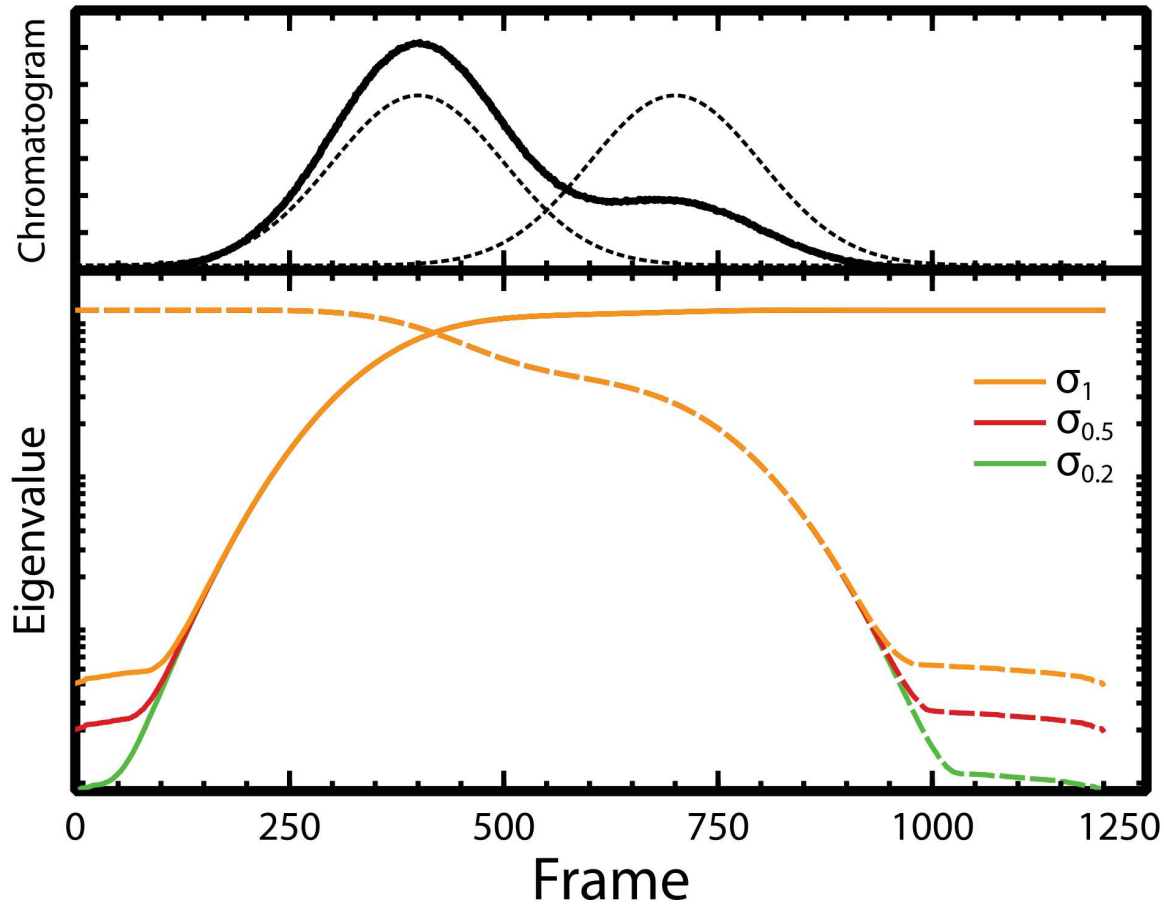


FIGURE 6.11. *EFA from synthetic data (monomer-dimer system). Only the first eigenvalue is plotted along the dataset in the forward (solid line) and the backward direction (dashed line). Different levels of noise are plotted (0.2, 0.5 and 1). First eigenvalue profiles, which indicates the detection of signal over the noise level, for the three datasets tested are overlapped for the vast majority of the chromatogram. The capacity to identify the appereance of the first species is virtually the same for all levels of noise.*

the same results than R_g for the $\sigma_{1.0}$ dataset (see previous section). However, the EFA is more sensitive to the level of noise, and its application is advantageous in high signal-to-noise datasets, and in scenarios where the R_g value of the mixed species is similar.

6.2.3.2.3 COSMiCS analysis using Equality constraint from R_g or EFA

In section 6.2.3.1, we used the information from the theoretical populations to create an Equality constraint in order to use it in the COSMiCS decomposition. That constraint decreased the ambiguity of the analysis and the results from the decomposition improved notably. Obviously, this information is not available in a real experiment, so it is necessary to use the information obtained by R_g and/or EFA to create the Equality constraint and use it in a COSMiCS analysis. The information from R_g and EFA is slightly different than the theoretical one, previously used to test the *Equality* constraint (see section 6.2.3.1). We want to test if the addition of monodisperse zones derived from R_g/EFA analyses is able to improve the decomposition in the same way. Concretely, the Equality constraint based on the R_g was introduced limiting the first species from the frames 1 to 700, and the second

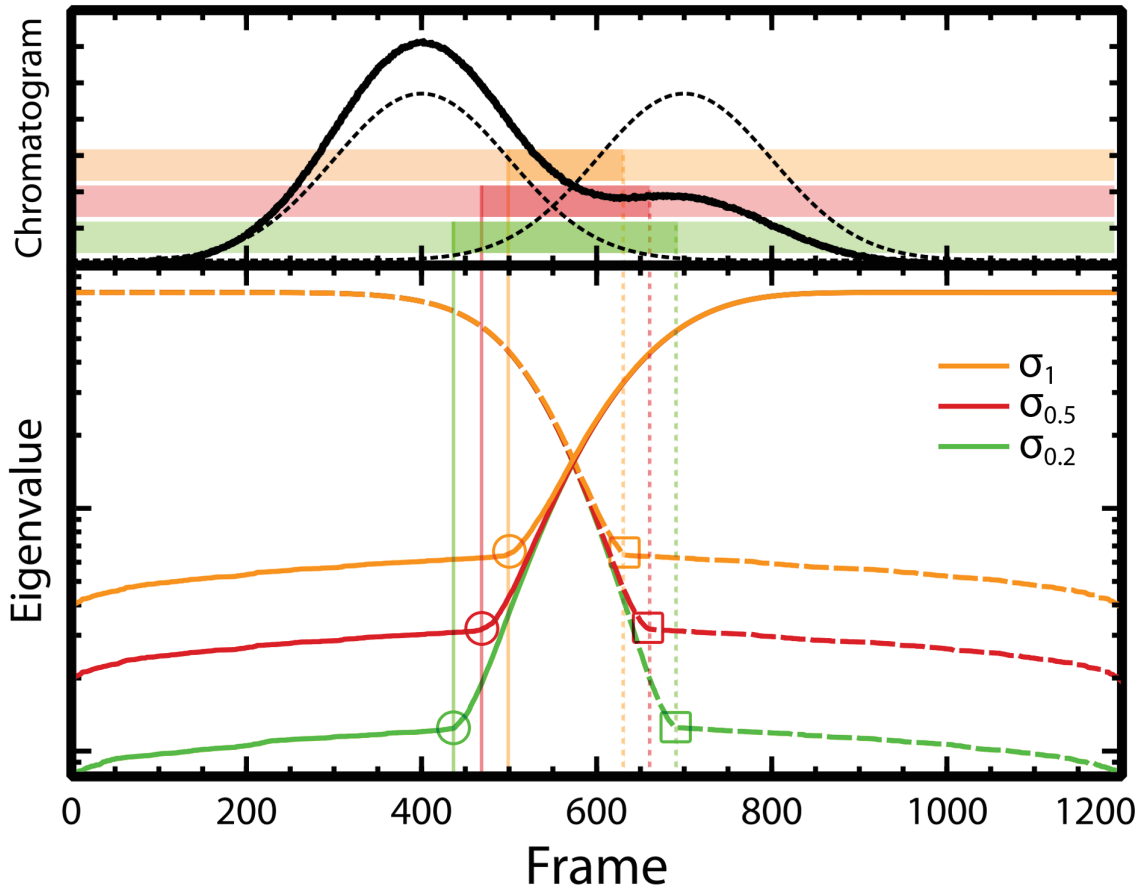


FIGURE 6.12. *EFA from synthetic data (monomer-dimer system) for the different noise levels tested ($\sigma_{0.2}$, $\sigma_{0.5}$ and $\sigma_{1.0}$). The second eigenvalue is plotted along the dataset in the forward (solid line) and the backward directions (dashed lines). The second eigenvalue indicates the appereance of a second species. Abrupt slope changes are highlighted with circles (forward) and squares (backward). Increasing noise levels leads to a progressive delay of the slope change point (abrupt change in the singular value). The ranges corresponding to monodisperse regions (light color) and overlapped regions (darker color) are shown on top of the chromatogram for each of the datasets, in the same color code.*

species from the frame 500 until the last frame (1200). Note that these limits are very similar to these obtained for the EFA analysis.

When we compare the results from the analysis using theoretical limits (Figure 6.8) with these derived using the R_g limits (Figure 6.13) we can see that both solutions are very similar. This means that the small contribution of the monomer peak that R_g or EFA does not detect, does not affect to the decomposition. The description of the populations is not optimal in either case, and the concentration profiles do not match with the original ones, especially for the monomer. In Table 6.2 we can see that the R_gs, ratio between I(*0*) of the species, and the difference between the maximum of the populations are similar in both analyses. However, with the R_g/EFA information the solution presents artifacts in the concentration profiles, with truncated populations. These truncated population profiles observed come from the need of additional frames and depends on the noise of the dataset. That error can be corrected by selecting an earlier frame corresponding to the appearance of the monomer in the *Equality* constraint in a new COSMiCS analysis, which will give

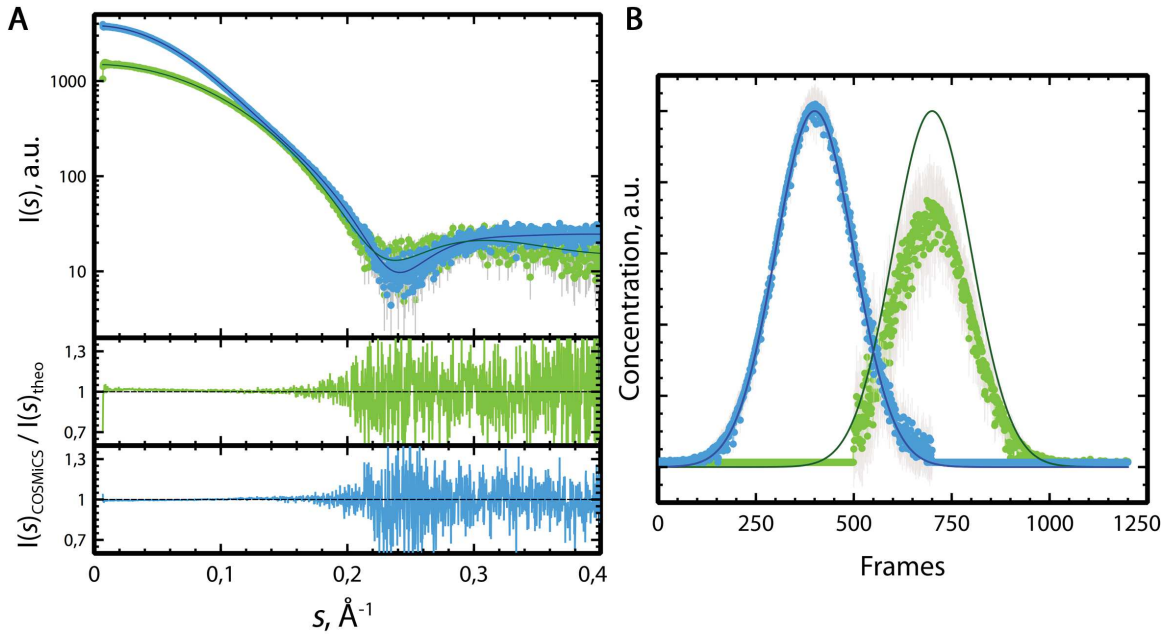


FIGURE 6.13. *Results from the COSMiCS analysis of the synthetic SEC-SAX data using the final noise without scaling, adding an equality constraint obtained using the R_g analysis. A) Pure decomposed curves for the dimer (blue) and monomer (green). In solid line is shown the original SAXS curves of the two species (same color code). At the bottom, the ratios $I(s)_{COSMiCS}/I(s)_{theoretical}$ of both species are shown. B) Decomposed Concentration profiles of the two species (same color code) and the original populations used to generate the synthetic data*

same results that would be closer to these obtained using theoretical information (Figure 6.8).

In summary, the use of monodisperse zones derived, which can be determined from the R_g or EFA analyses, is an excellent strategy to improve the decomposition with COSMiCS to derive precise SAXS profiles from overlapped SEC-SAXS chromatograms. However, the rotational ambiguity, which is inherent to chemometric analyses, produces wrong populations. Therefore, it is necessary to add more information to decompose correctly the populations too.

6.2.4 Adding more information to the system: UV-Vis data

In the previous analysis using Equality constraint, COSMiCS was able to decompose SEC-SAXS dataset in SAXS curves with very good agreement with the original ones. However, the concentration profiles were not correct due to the rotational ambiguity. To improve the solutions, more information has to be introduced in the analysis. UV-Vis absorbance is extra information that is normally measured in chromatography and reports on the populations of the species. UV-Vis absorbance can be measured in parallel to the collection of SAXS data or measured off-line, depending the set-up of the beamline.

We have tested whether the UV absorbance can be used to further restraint the solutions. In order to test the UV absorbance data we have calculated a theoretical Absorbance profile of the chromatogram. Extinction coefficients (ε) of the monomer and

	R_g(dim)	R_g(mon)	$I(0)_{dim}/I(0)_{mon}$	C(dim)	C(mon)	$P^{400}_{dim}/P^{700}_{mon}$
Theoretical	**21.10**	**15.47**	**3.6**	**-**	**-**	**1**
Noise $\sigma_{1.0}$	21.30	13.47	3.6	103	101	1.52
Noise $\sigma_{1.0}$ & Equality (theoretical)	20.98	15.58	2.4	14	41	1.58
Noise $\sigma_{1.0}$ + Equality (R_g/EFA)	20.92	15.58	2.5	44	41	1.52
Noise $\sigma_{1.0}$ + Equality (R_g/EFA) + Absorbance Closure	20.91	15.65	3.5	51	47	1.05
Noise $\sigma_{1.0}$ + Absorbace Closure	21.24	12.77	4.6	103	101	1.3

TABLE 6.2. *Analysis of the COSMiCS decomposed species from datasets with high level of noise ($\sigma_{1.0}$) using different constraints.*

dimer were calculated using ProtParam http://web.expasy.org/protparam/, determining a ε_{mon} = 5960 $M^{-1}cm^{-1}$ and ε_{dim} = 11920 $M^{-1}cm^{-1}$. Using these extinction coefficients and the original populations, a theoretical UV absorbance profile for the SEC synthetic experiment was computed (Figure 6.14), using the expression:

$$A = c_{mon} \cdot \varepsilon_{mon} + c_{dim} \cdot \varepsilon_{dim} \qquad \text{(Eq. 6.1)}$$

The implementation as a constraint of the Absorbance was done using the same strategy used in the *Closure*. In each iteration of the classic version of *Closure*, MCR-ALS does an estimation of the concentrations using least squares, and then it forces their concentrations to sum a given value provided as input. In our modified version of *Closure*, we forced the estimated concentrations for each iteration to match the synthetic UV absorbance values using Eq. 6.1.

6.2.4.1 COSMiCS analysis using UV-vis absorbance data in Closure and Equality constraint

COSMiCS analysis was performed to the dataset with low signal-to-noise, applying Equality constraint and Closure of Absorbance data during the analysis. Using the information from R_g (Figure 6.9) and EFA (Figure 6.12), the monodisperse zones were defined as the first species from the frames 1 to 700, and the second species from the frame 500 until the last frame (1200).

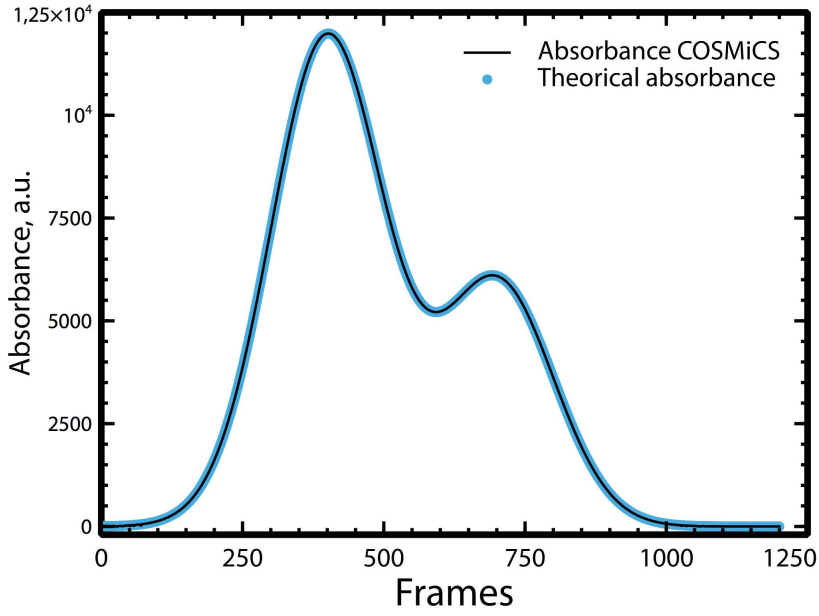


FIGURE 6.14. *Theoretical (blue) and COSMiCS back-calculated (black) Uv-vis absorbance when using the Absorbance data as a closure constraint in the COSMiCS analysis. Both profiles are perfectly overlapped.*

This analysis showed a good agreement of the curves (Figure 6.15A), similar to the analysis with only *Equality* constraint; R_g(dim) = 20.91 Å and R_g(mon) = 15.65 Å. Importantly, in this case, the concentration profiles (Figure 6.15B) dramatically improved, and they agree perfectly with the original populations. The relative ratio of I(*0*) is also similar to the original ratio ($I(0)_{dim}$ / $I(0)_{mon}$ = 3.5). Therefore, under these application of the monodispersity and the UV-vis COSMiCS provides an excellent decomposition of the SEC-SAXS dataset.

6.2.4.2 COSMiCS analysis using UV-vis absorbance data in *Closure*

We performed also a COSMiCS analysis using the data with low signal-to-noise and just the Absorbance data to test if this constraint can compensate the absence of *Equality*. The resulting SAXS profiles for both species were bad (Figure 6.16A). They were not in agreement with the theoretical ones, and similar to these obtained from the analysis without any constraint, with R_g(dim) = 21.24 Å and R_g(mon) = 12.77 Å and the ratio $I(0)_{dim}$ / $I(0)_{mon}$ = 4.6. Moreover, the application of the absorbance constraint only slightly improved populations (Figure 6.16B), especially for the monomer species, with a ratio between the two maxima was 1.3. This observation shows that *Equality* constraint is a better strategy to decrease ambiguity and decompose correctly the pure spectra.

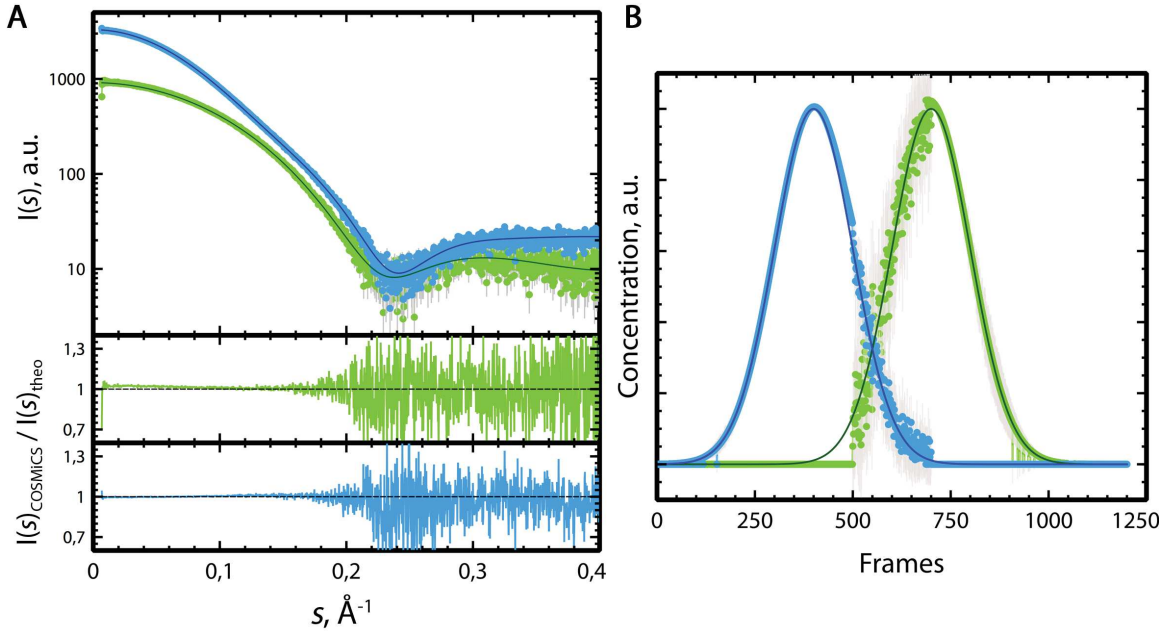


FIGURE 6.15. *Results from the COSMiCS analysis of the synthetic low signal-to-noise SEC-SAXS dataset ($\sigma_{1.0}$) including an equality constraint that indicates that the dimer ends at frame 700 and monomer starts at frame 500, and a closure using the UV-vis Absorbance information of Figure 6.14. A) Pure decomposed curves for the dimer (blue) and monomer (green). In solid line is shown the original SAXS curves of the two species (same color code). In the bottom, the ratios $I(s)_{COSMiCS}/I(s)_{theoretical}$ of both species are shown. B) Concentration profiles for the two species (same color code) and the original populations used to generate the synthetic data.*

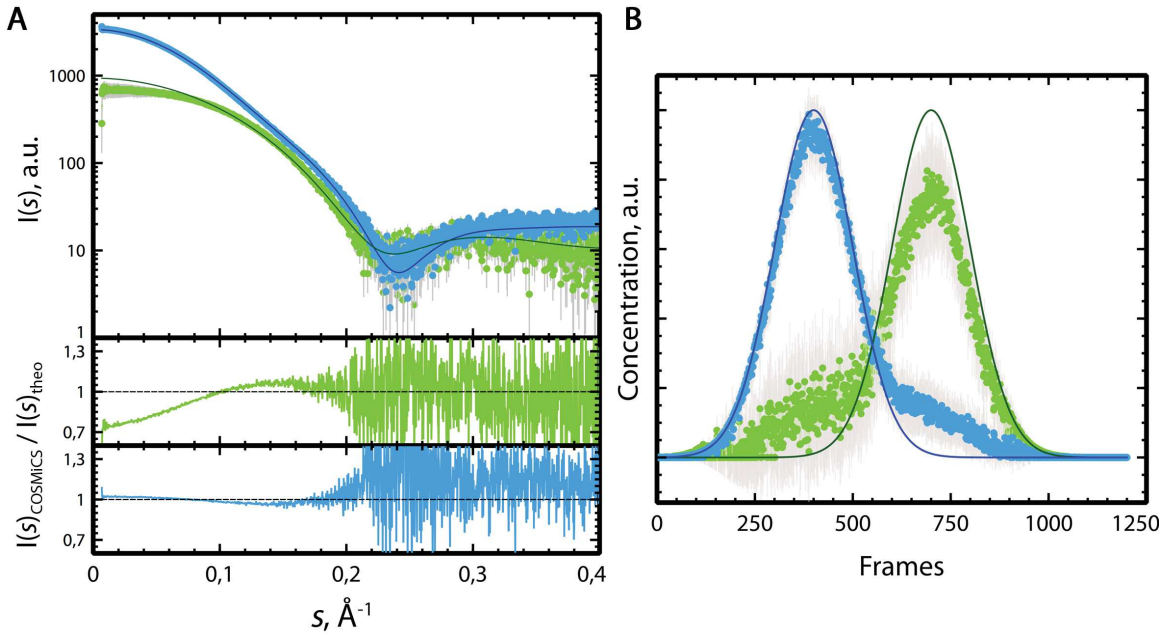


FIGURE 6.16. *Results from the COSMiCS analysis of the synthetic low signal-to-noise SEC-SAXS dataset ($\sigma_{1.0}$) including a closure using the UV-vis Absorbance information. A) Pure decomposed curves for the dimer (blue) and monomer (green). In solid lines are shown the original SAXS curves of the two species (same color code). At the bottom, the ratios $I(s)_{COSMiCS}/I(s)_{theoretical}$ for both species are shown. B) Concentration profiles for the two species (same color code) and the original populations used to generate the synthetic data*

6.3 Real-case study: POP

6.3.1 Prolyl Oligopeptidase (POP) system

Prolyl Oligopeptidase (POP; EC 3.4.21.26) is a monomeric 81-KDa cytosolic serine endopeptidase that hydrolyses peptides under 30 residues at the carboxyl side of proline [332]. POP is ubiquitous in mammals, but a relatively high concentration of this protein is found in the central nervous system (CNS). The first X-ray structure was obtained by Fulop et al. (PDB entries 1QFS and 1QFM) from porcine muscle POP [333]. The crystallographic structure showed that this enzyme has an overall cylindrical shape, constituted by two domains: the α/β-hydrolase and the β-propeller that are linked by a pair of hinge polypeptide chains (Figure 6.17). The X-ray structure of the mammalian enzyme shows the α/β-hydrolase and β-propeller domains packed together in a closed conformation (Figure 6.17A) [333]. However, the crystal structures of two bacterial POPs show a large hinge separation between domains [334, 335] (Figure 6.17B), thus suggesting that the enzyme undergoes interdomain motions [336]. Two studies, one based on ^{15}N line broadening NMR experiments [337] and the other on X-ray crystallography combined with MD simulations [338], strongly support that POP is a highly flexible enzyme, but several fundamental aspects concerning the conformational landscape of POP in solution and the effects on this flexibility exerted by inhibitors are largely unknown.

The in vivo role of POP is related to synaptic functions and neuronal development. It has been discovered that POP interacts with the intrinsically disordered proteins such as α-synuclein and GAP-43 [339]. Recent studies have demonstrated that the direct interaction between POP and α-synuclein accelerates the aggregation of α-synuclein in vitro and in cells [340].

Nevertheless, the lack of knowledge on the conformational dynamics of POP and the effects of inhibitors represents a major drawback for exploring the mechanisms underlying POP-mediated aggregation of α-synuclein. Here, we analyzed in detail the conformational equilibrium of POP in solution and how this is affected by the binding of active-site-directed inhibitors. We analyzed SEC-SAXS experiments complemented with MD simulations to probe large-scale structural fluctuations in solution [341, 342]. Moreover, the effects of binding of covalent active-site-directed inhibitor benzyloxycarbonyl-prolyl-prolinal (ZPP, Figure 6.17C) [343] on POP conformational dynamics were examined using the same combined approach (See chapter 6.3.2).

During the production of POP some aggregates are formed. Although, gel filtration successfully removes POP aggregates, spontaneous aggregation slowly takes place in solution, especially in concentrated samples (SEC chromatogram of POP is shown in Figure 6.18). Note that a significant amount of irreversible aggregates would strongly interfere in direct SAXS measurements of concentrated POP samples (i.e. batch measurement). Hence, an online SEC-SAXS measurement was the best solution to overcome this problem. The separation between POP aggregates and monomers was excellent by conventional Superdex 200 gel filtration columns, and moreover, the facility of P12 beamline allowed the

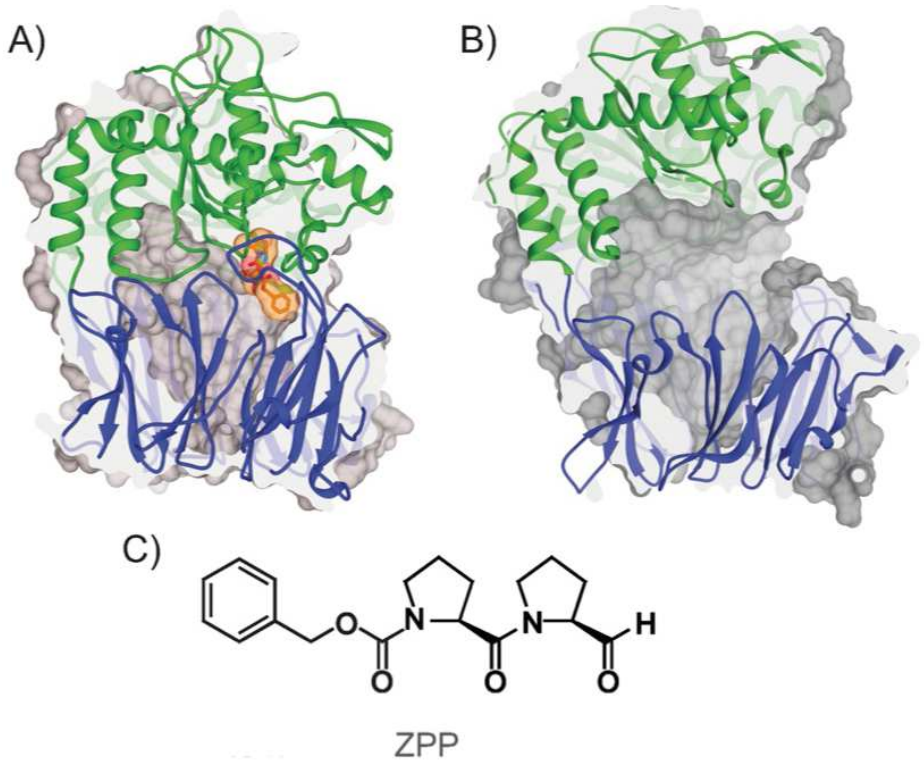


FIGURE 6.17. *A) Porcine POP (PDB ID: 1QFS) [333] in the closed conformation covalently bound to the active-site-directed inhibitor ZPP (orange). The a/b-hydrolase domain is shown in green and the b-propeller is in blue. B) Aeromonas punctata POP in the open conformation (PDB ID: 3IUJ) [334]. C) Inhibitor ZPP. Figure extracted from [331]*

option of connecting a Malvern FPLC instrument to the SAXS capillary. Free and ZPP-bound POP samples were analyzed by this technique. Our objective here was to extract free and inhibitor-bound POP SAXS scattering profiles, and compare them with the theoretical profiles obtained from X-ray structures, homology models, and MD simulations.

This study was carried out in collaboration with Ernest Giralt and Abraham López (Chemistry and Molecular Pharmacology Program, IRB - Barcelona), and Víctor Guallar (BSC-Barcelona).

6.3.2 ZPP-bound POP (closed form)

6.3.2.1 Primary analysis

The SEC-SAXS experiment of ZPP-bound POP generated a dataset of 4500 scattering curves (Figure 6.19). The first step consisted in the visual inspection of all SAXS frames. This procedure introduces a time-consuming intervention of the user into the data analysis as SAXS frames randomly present artifacts derived from large particles in suspension (e.g. dust). These artifacts were easily identified by visual inspection due to their uncommon scattering profiles, and were eliminated by removing the corresponding frame.

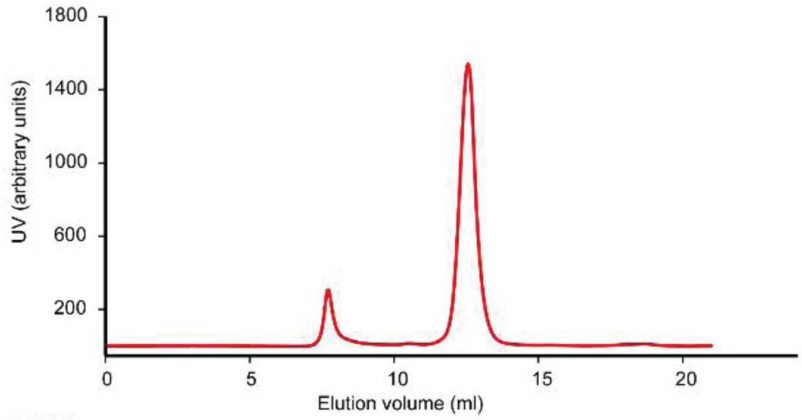


FIGURE 6.18. *Size exclusion chromatography coupled to SAXS. Representative SEC chromatogram of free POP at room temperature, using a Superdex 200 column and standard POP buffer (50 mM Tris-HCl pH=8, 20 mM NaCl). The first eluting peak corresponds to an aggregate whereas the second corresponds to the monomeric species.*

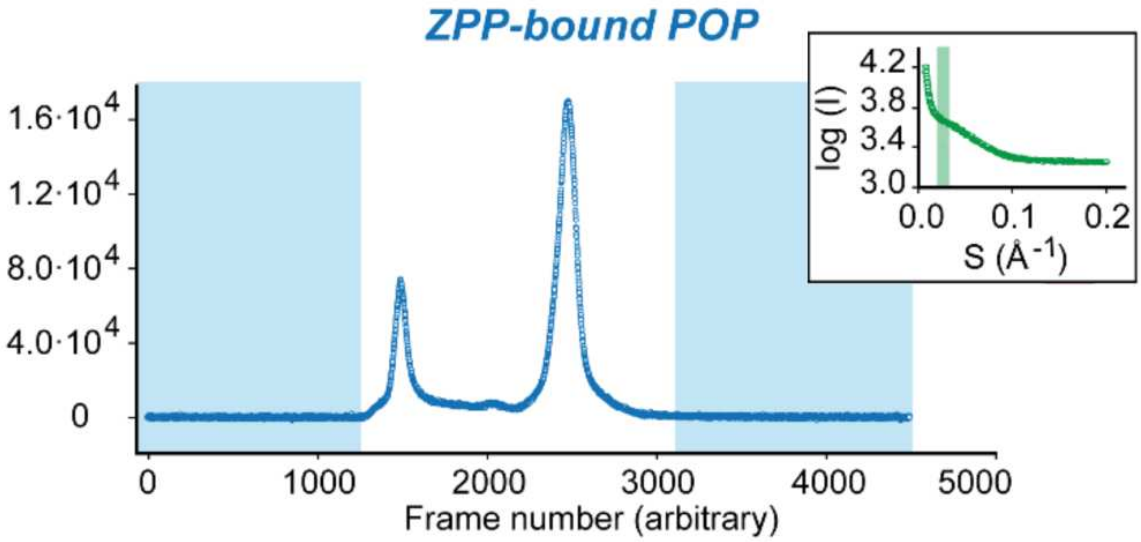


FIGURE 6.19. *SEC-SAXS integration chromatograms of ZPP-bound POP. Blue regions indicate the intervals used to derive the averaged buffer curve. Inset box shows the region of the scattering profiles which was integrated to obtain the integration chromatograms (green area)*

For the detection of buffer frames, the whole SEC-SAXS chromatogram was represented by plotting the integration of each SAXS curve from 50 to 100 data points (Figure 6.18) as a function of frame number (an absolute time scale was not possible due to the manual injection of the sample and other experimental limitations).

According to integration chromatograms, the following buffer regions were selected: from frame 0 to 1250 and from 3100 to 4500 and from 3100 to 4500 (Figure 6.19) and averaged with datsub program from the ATSAS data analysis software [330]. The buffer profile was subsequently subtracted from all frames of chromatogram. This operation was carried

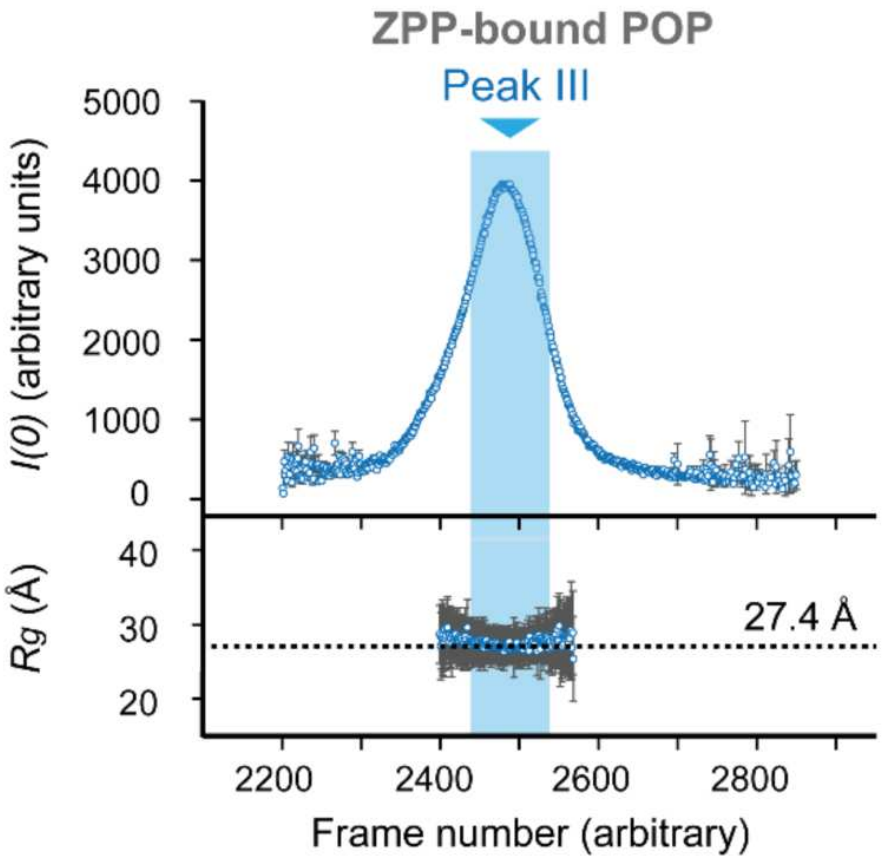


FIGURE 6.20. *SEC-SAXS I(0) chromatogram for the monomer frames of ZPP-bound POP. The blue area corresponds to the monomer area, as localized by PCA (see below). R_g plot is shown at the bottom; R_g derived from the averaged SAXS curve is marked in black dotted lines, 27.4 Å*

out with MATLAB®. The monomer region of ZPP-bound POP was found between frames 2200 and 2850.

Afterwards, I(*0*) and R_g for the subtracted frames were calculated from the Guinier approximation (see chapter 1.2.3). Guinier plots were calculated with autorg program [ref atsas]. Finally, I(*0*) and R_g were plotted against SAXS frame number (Figure 6.20), which can be assigned to an arbitrary elution time; these representations are known as I(*0*) and R_g chromatograms, respectively. The I(*0*) and the R_g chromatograms reflect the concentration and the size and of the eluting species, respectively. For this reason, the inspection of these plots is highly useful to monitor SEC-SAXS experiments.

6.3.2.2 PCA of monomer region of SEC-SAXS I(*0*) chromatograms

In the case of ZPP-bound POP, the chromatogram shows two isolated peaks corresponding to the aggregate and the monomeric species a single peak. In this case, it is not necessary to perform a decomposition of the data, and it is enough to perform an averaging of the frames to obtain a high-quality curve. However, it is important ensure that the frames correspond a monodisperse sample before averaging them. For that purpose,

the chromatogram was subjected to PCA in increasing intervals centred at the monomeric peak (frame 2488) (Figure 6.20). The peak presents only one significant eigenvector and eigenvalue when 101 frames were subjected to PCA (Figure 6.21). Increasing the number of frames to 401 in the PCA analysis shows that the second eigenvalue presents shape, although noisy. When the PCA was performed with the whole dataset, third eigenvector presents also shape, which probably is caused by the high level of noise of the data. Consequently, we used the 101 central frames to be certain that the data correspond to a single species. This central region of 101 frames also presents a stable R_g (Figure 6.20), which is another indication of monodispersity. Therefore, the 101 central frames of the peak were averaged to produce a high-quality SAXS curve corresponding to the pure species (Figure 6.22), which has a R_g = 27.40 Å. A complete structural analysis of this curve will be described in next section.

6.3.2.3 Ensemble optimization fitting and theoretical SAXS-scattering profiles

Once we have obtained the high-quality SAXS curve of the ZPP-bounded POP, structural information can be extracted directly. This information is limited only to I(*0*), Rg, and qualitative information from Porod's and Kratky's representation. In order to obtain more information, we performed the fitting of the theoretical SAXS profile with structures from MD simulations using the Ensemble Optimization Method (EOM [61], section 1.3.2.2).

The calculation of theoretical SAXS profiles from structures was carried out with CRYSOL [41], implemented in ATSAS data analysis software. The scattering profiles where calculated up to s = 0.5 $Å^{-1}$, with a total of 201 data points. The parameters of CRYSOL program were a maximum order of 30 harmonics and of 20 Fibonacci grids.

The pool of ZPP-bound POP profiles consisted in those calculated from the 1019 structures from the MD5 simulation described in section 6.3.4.2 (Figure 6.23). The curves obtained from the simulation were used to perform the EOM analysis of the experimental curve of ZPP-bound POP. The size of conformers in a chromosome, N, depends on the flexibility of the system and it is normally N = 50 for unfolded systems. However, the conformational fluctuations expected for POP are restricted to interdomain separation and local loop flexibility. For this reason, N was set at 20. Indeed, the use of N = 50 was discarded as it did not improve the quality of the result. The number of chromosomes was the typical value C = 50. Each EOM run consisted in 1500 generations, and 100 independent EOM runs were performed.

Given the high quality of the experimental scattering profiles, this analysis was performed for data points with s < 0.3 $Å^{-1}$. An excellent fitting was obtained between the experimental and the averaged theoretical curve selected by the EOM (χ^2 = 0.054, Figure 6.22). Note that the low χ^2 value arises from the large error bars derived from the average of the SEC-SAXS frames. Importantly, all the 20 theoretical curves selected by the EOM closely resembled the initial 1QFS X-ray structure of inhibited POP (maximum RMSD value of 1.226 Å, Figure 6.23). Therefore, this result shows that ZPP-bound POP exists in solution in a highly stabilized closed conformation.

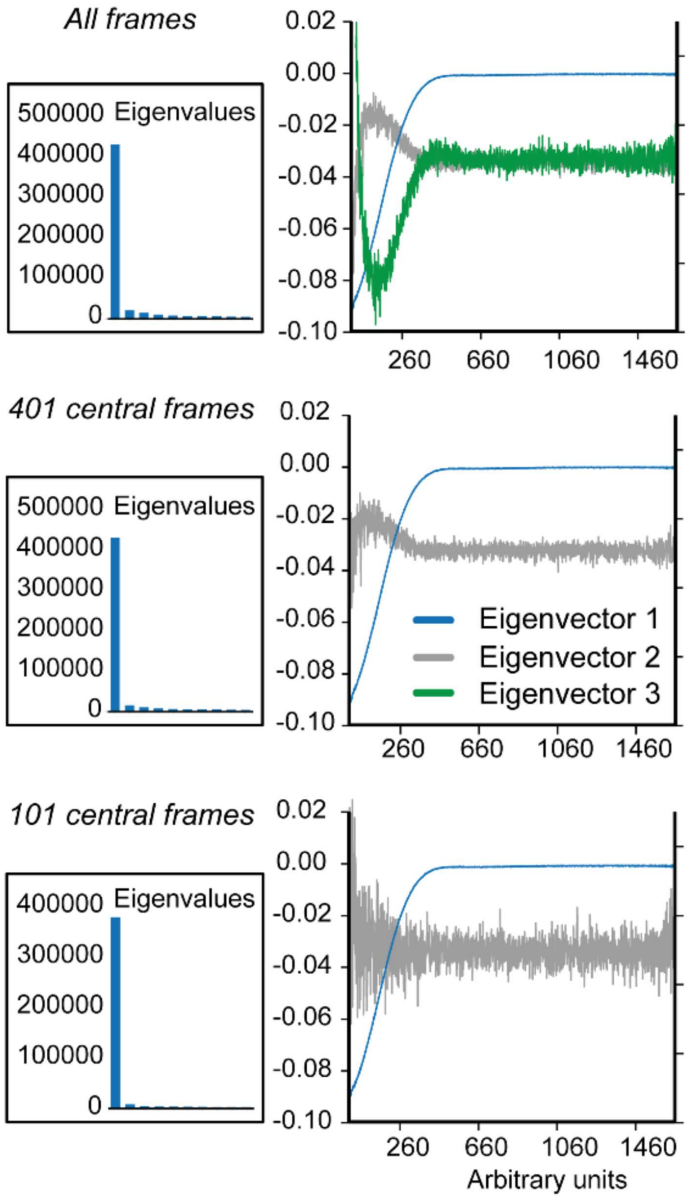


FIGURE 6.21. *Singular Value Decomposition (SVD) of the monomer region of SEC-SAXS I(0) chromatograms at different intervals; eigenvalues are displayed in the insets. SVD calculations of peak III of ZPP-bound POP showed only one predominant eigenvector and eigenvalue, independently of the interval used in the algorithm. In the case of the SVD analysis for the complete peak, a third eigenvector is shown for clarity.*

6.3.3 Free POP

6.3.3.1 Primary analysis

An equivalent SEC-SAXS experiment was performed for free POP obtaining a complete dataset of 4500 SAXS curves (Figure 6.24). A visual inspection of all SAXS frames was performed in order to remove artifacts, in the same way than in the previous dataset. According to integration chromatogram, the following buffer regions were selected: from frame 810 to 1260 and from 4220 to 4500 (Figure 6.23). All buffer frames comprised in these

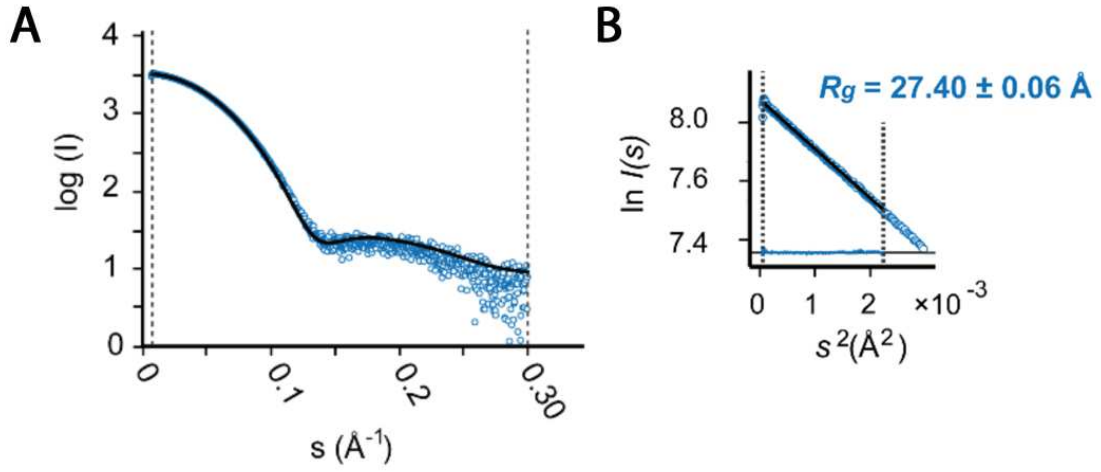


FIGURE 6.22. *A) EOM fitting of the scattering profile of ZPP-bound POP (blue dots). The theoretical curve is shown as black solid line. B) Guinier plot and R_g value.*

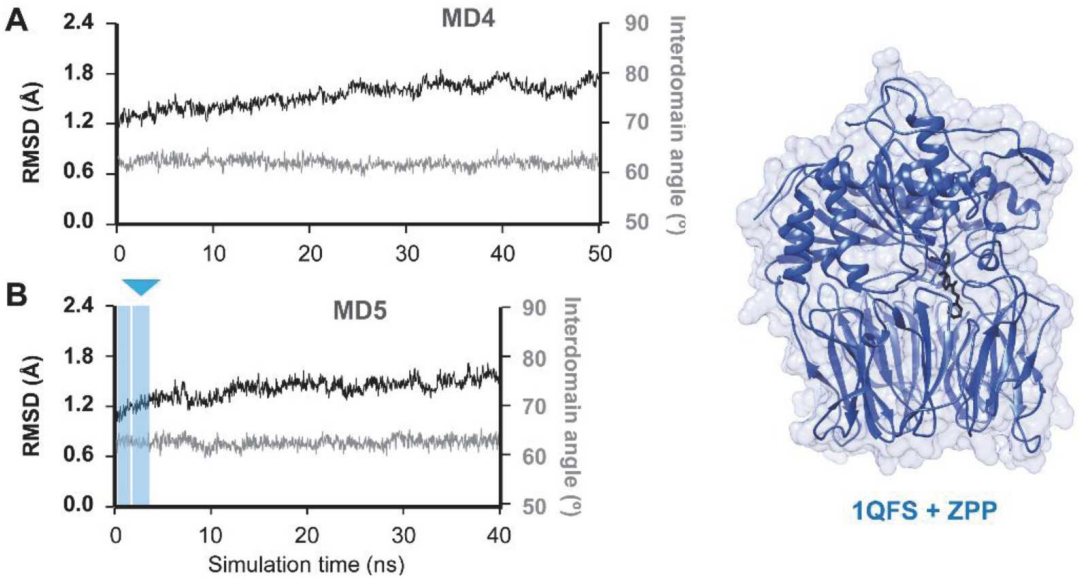


FIGURE 6.23. *MD simulations of inhibited POP. A) MD4 simulation of X-ray structure 1QFS with the hemiacetal bond between Ser 554 and ZPP removed. B) MD5 simulation of 1QFS structure with ZPP inhibitor covalently bound. RMSD (in Å) is shown in black, and interdomain angle is shown in gray. Blue sections in MD5 simulation correspond to the intervals of conformations selected by EOM when fitting the SAXS curve. 1QFS structure is shown in blue; ZPP inhibitor is depicted in black. Figure extracted from [344]*

regions were averaged in order to obtain the final buffer profile. The buffer profile was subsequently subtracted to all frames of the chromatogram. This operation was carried out with MATLAB® program.

The monomer region of free POP was found approximately between frames 2350 and 3000. Interestingly, the SEC-SAXS I(*0*) chromatogram of free POP disclosed two coexisting peaks at room temperature, referred as peaks I and II (Figure 6.25).

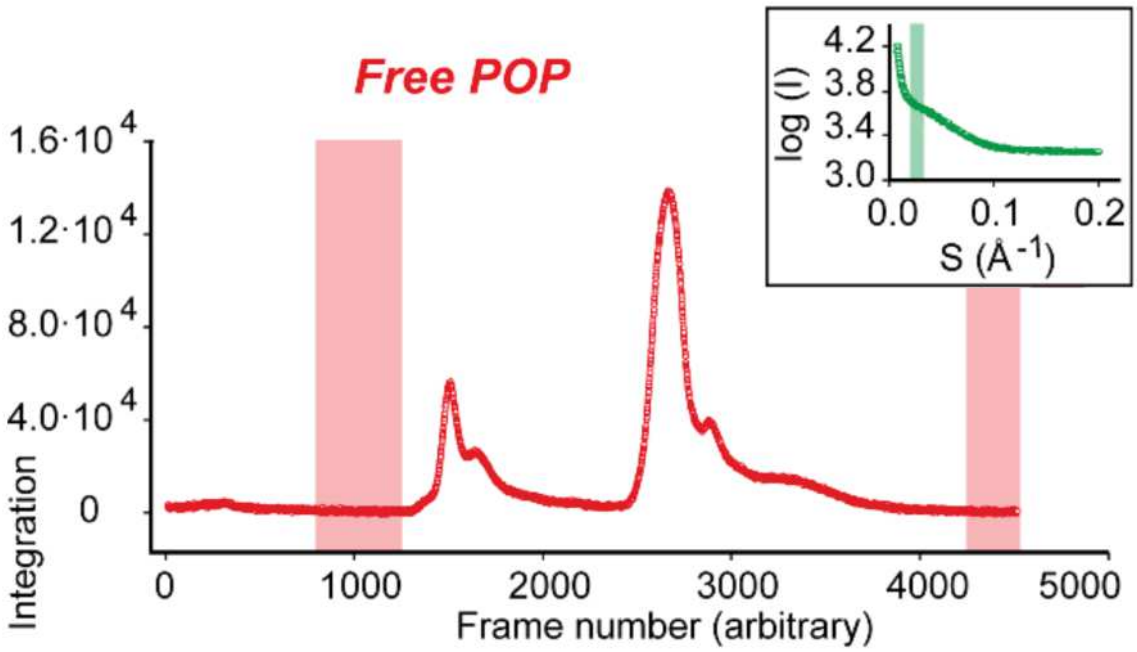


FIGURE 6.24. *SEC-SAXS integration chromatogram of free POP. Red regions define the intervals used for the averaging of the buffer curve. Inset box shows the region of the scattering profiles which was integrated to obtain the integration chromatogram (green area)*

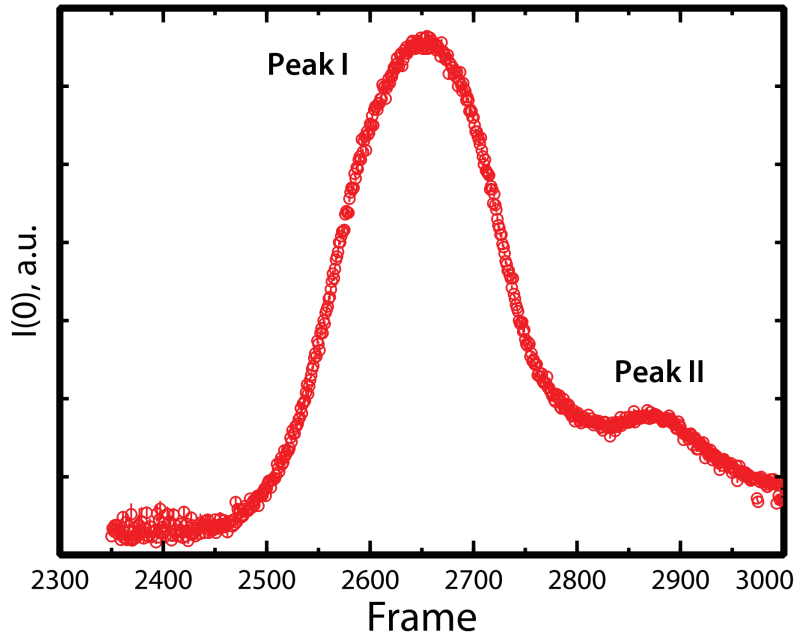


FIGURE 6.25. *SEC-SAXS I(0) chromatogram displaying peaks I and II of free POP.*

6.3.3.2 Peak composition in the free-POP SEC-SAXS dataset: R_g and EFA analysis

Due to the presence of two overlapped peaks, it was not possible to perform the analysis of the free POP simply averaging the central part of the peaks, like in the case of ZPP-bound POP. In this case, it is necessary to decompose the data to obtain scattering curves for both species. We have used COSMiCS for this purpose. As we established in the previous section, it is very important to determine the monodisperse zones on the dataset in order to obtain an accurate solution. For that, the first step is to inspect the R_g along the dataset and performing an EFA.

Figure 6.26 shows the R_g profile (top, blue) and the second eigenvalue calculated from EFA (forward and backward direction), from frame 2450 to frame 3000. It is possible see how the R_g is stable along the initial part of the I(*0*) chromatogram to subsequently start decreasing. The change in the R_g profile corresponds to the frame where the second EFA eigenvalue starts increasing (in forward direction), and it indicates the appearance of the second species, which corresponds to the peak II. The eigenvalue in backward direction has the contribution of the species from peak I but does not shows a change of slope, which suggests that the species from peak I is present along the complete dataset.

6.3.3.3 COSMiCS analysis

In order to extract the pure SAXS curve corresponding to peak I and II, all the frames were decomposed with COSMiCS, in the same way than the previous analysis with synthetic data. In this case, our data consisted in a collection of scattering profiles (i.e. the SEC-SAXS chromatogram) of a two-component mixture, in which each SAXS frame is a mixture of the scattering profiles of species corresponding to peaks I and II at variable concentrations (Figure 6.25). We removed frames from 2450 to 2500 and from 2900 to 3000 because according to R_g (Figure 6.26) these scattering curves corresponded to buffer.

COSMiCS analysis was performed using Non-negativity and Equality as constraints. Concretely, the Equality constraint was based on the EFA analysis, which indicates that the first species is present in the whole dataset, but limiting the second species from the frame 2640 to the end of the dataset (frame 2900). COSMiCS results are shown in Figure 6.27. The decomposed scattering curve from the first peak has a R_g = 28.86 Å and the one for the second peak is smaller, with R_g = 23.35 Å (Figure 6.27A). The Guinier plot of the interval with $s \cdot Rg < 1.3$ of the curve corresponding to peak I was linear, indicating high quality of the data (Figure 6.22B). Concentration profiles of the species (Figure 6.27B) present a truncated curve for the second species. This can be due to inherent uncertainty of EFA in detecting the appearance of the second species due to the high level of noise in peak tails. Using the I(*0*) and a standard BSA we estimated to be 0.18 mg/ml the concentration in the maximum of peak II.

6.3.3.4 Structural analysis of peak I

We have performed a COSMiCS analysis of the free POP SEC-SAXS dataset and we obtained a decomposed scattering curve from the first peak (see section above). Previously

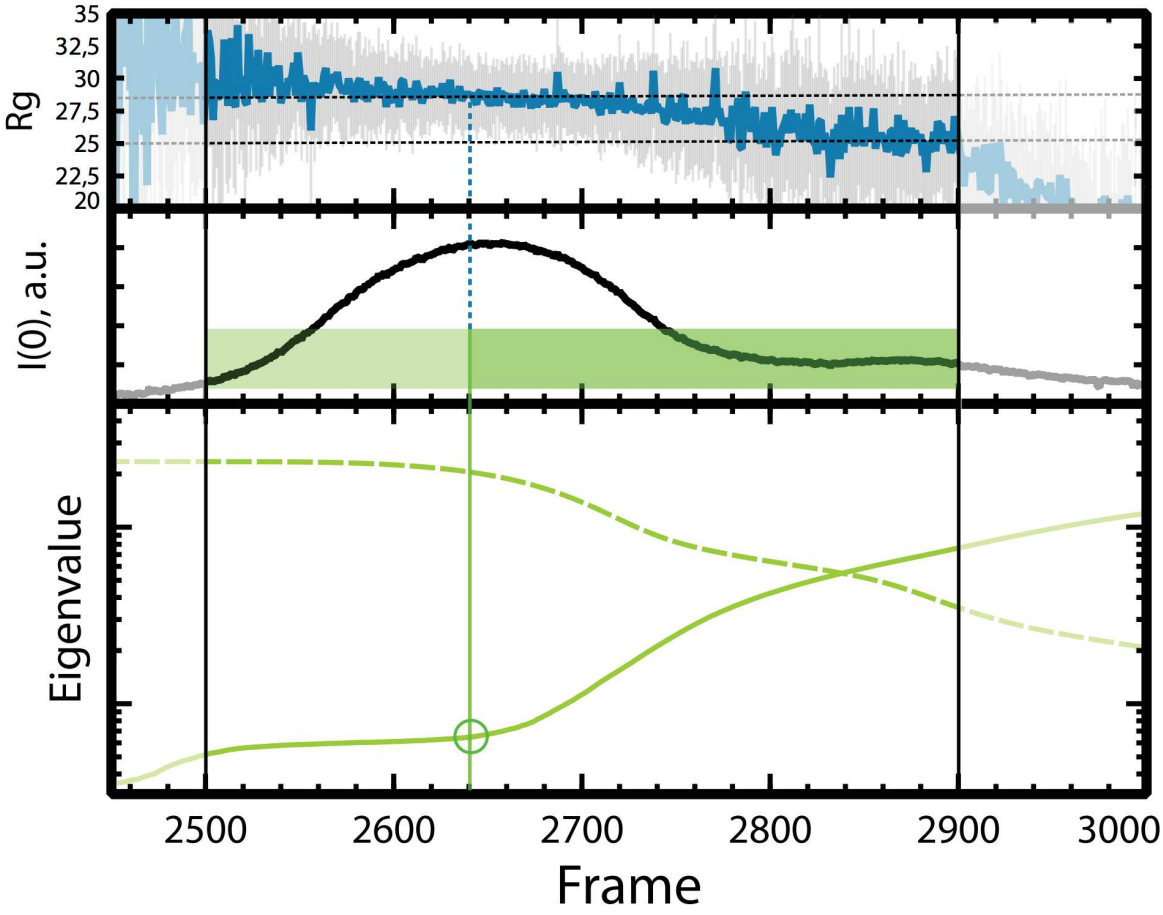


FIGURE 6.26. *$_g$ (blue) and EFA (green) along the SEC-SAXS chromatogram of free POP. The second eigenvalue is plotted along the dataset in the forward (solid line) and the backward directions (dashed lines). The second eigenvalue indicates the appereance of a second species. The slope change of the eigenvalue profile in forward direction is highlighted with a circle and it indicates the appearance of a second species. R_g shows a slow decrease from that point. The ranges corresponding to the monodisperse region (light green) and overlapped region (darker green) are shown on top of the chromatogram. The final range of the chromatogram selected for the COSMiCS analysis is limited by the black lines (frames 2500 to 2900)*

[see analysis described in the published article – (Paper II)], a curve of peak I was derived by simply averaging the part of the dataset corresponding to a single species (see below). The curves obtained using these two methods are very similar. The structural analysis that I describe in this section corresponds to that performed for the averaged curve. Figure 6.28 shows the PCA analysis of the free POP dataset using the 101, 401 central frames of the peak and the whole dataset. The PCA analysis of 401 frames and the whole dataset present a second eigenvector with shape, which could be produced by the increasing level of noise due to the inclusion of low protein concentration curves, or the presence of a second species, which was confirmed by EFA a posteriori. The curve used for the structural analysis was obtained by averaging 101 frames, which we are sure that corresponds to a single species.

To perform the structural analysis of the free POP SAXS curve we compared the experimental curve with theoretical ones from different structural models of the protein. The

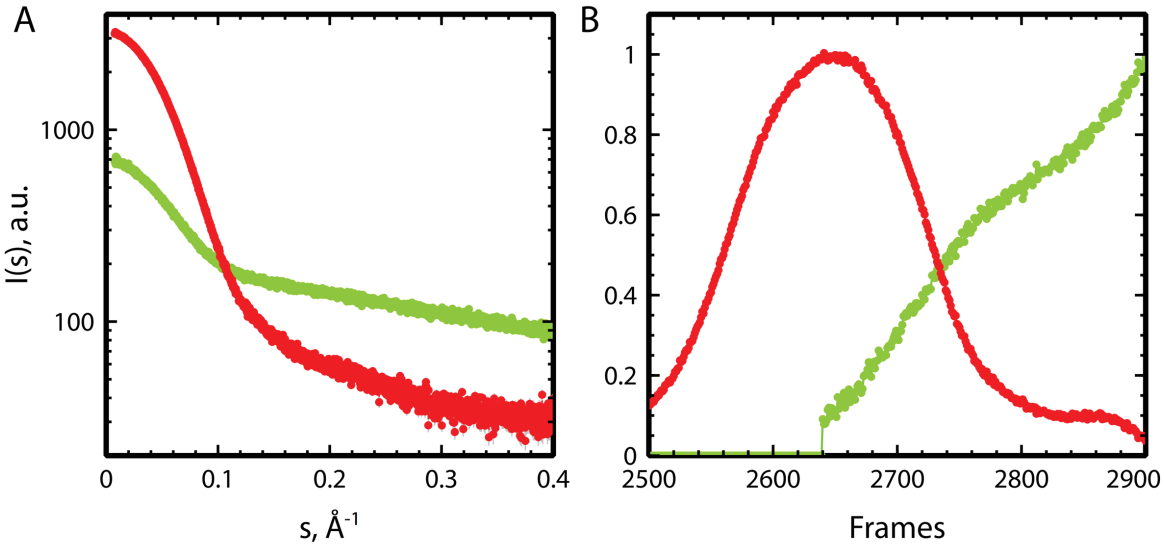


FIGURE 6.27. *Results from the COSMiCS analysis of the SEC-SAXS dataset for free POP. A) Pure spectra of the peak I (red) and peak II (green). B) Concentration profiles of the two species (same color code).*

χ^2 was monitored to quantify the degree of similarity between both curves. This fitting was performed separately with the averaged SAXS curves corresponding to each MD trajectory, in order to estimate the predominant conformation present at peak I. Fittings of the averaged SAXS curves suffered from poor quality, and the derived χ^2 values were not acceptable (0.282, 0.396 and 0.131 for MD1, MD2 and MD3 simulations, respectively). Note that the low χ^2 values are related to the large error bars, and not to the good agreement to the experimental curve. These results indicated that the structure or the ensemble of coexisting conformations of POP in solution do not correspond to the trajectories derived from MD simulations.

For this reason, a more flexible methodology was used in order to extract the ensemble of conformations that collectively described the SAXS experimental data. Specifically, we performed an EOM analysis that has the capacity to bias the population of the coexisting conformers based on SAXS data [61] (See section 1.3.2.2). In the same way that we analyzed the data from the ZPP-bounded POP, we described the free POP experimental curve with the theoretical SAXS profiles obtained from the conformations derived from multiple MD simulations.

The calculation of theoretical SAXS profiles from computational structures was carried out with CRYSOL program [41], implemented in ATSAS data analysis software. The scattering profiles where calculated up to $s = 0.5$ Å^{-1}, with a total of 201 data points. The parameters of CRYSOL program were a maximum order of 30 harmonics and of 20 Fibonacci grids. Given the large number of structures generated along MD1, MD2, MD3 simulations (see methods section 6.3.4.2 for the description of these trajectories), only certain structures were periodically collected from the trajectories to compute the theoretical profile with CRYSOL (Table 6.3 and Figure 6.29). The pool of free POP conformations was generated by joining snapshots from MD1, MD2 and MD3 simulations as described in Table 6.3. The

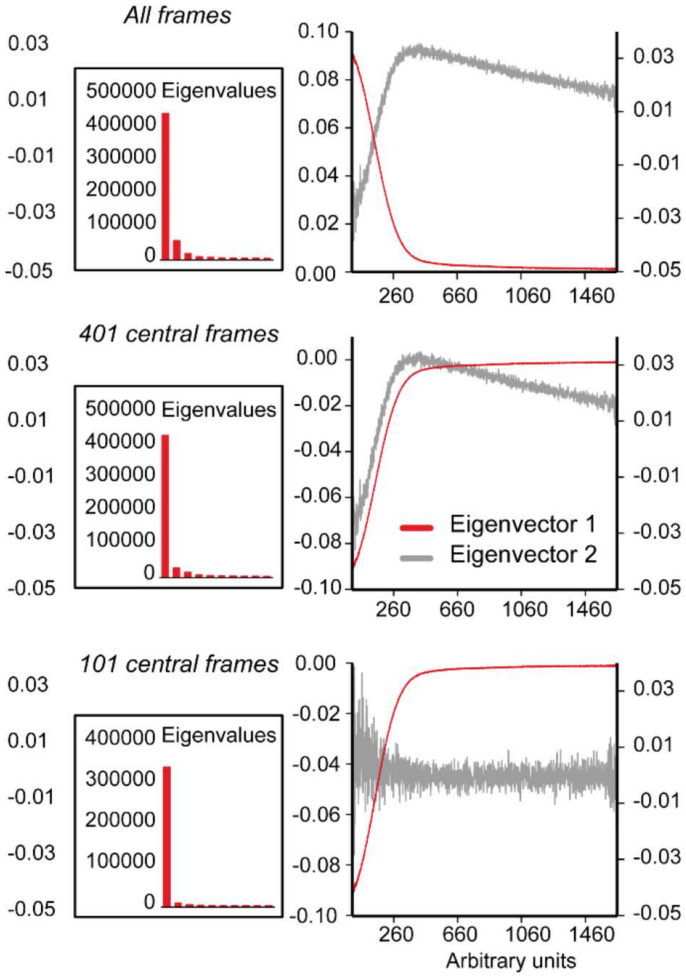


FIGURE 6.28. *Singular Value Decomposition (SVD) of the monomer region of SEC-SAXS I(0) chromatogram of free POP at different intervals; eigenvalues are displayed in the insets. SVD calculations both from the whole region and the case of 401 central frames, show a significant contribution of a second eigenvector. Only in the case of 101 frames an unique eigenvector is present.*

pool was composed by 13766 conformations.

The fitting of the curve was carried out for data points with $s < 0.15$ Å^{-1}. An excellent fitting was obtained with the EOM ($\chi^2 = 0.021$, Figure 6.30A), indicating the suitability of this method to describe the flexibility of POP. The EOM selected 45% of the frames from MD1 and 55% from MD2 (Figure 6.29A and 6.29B). The selected frames corresponded either to completely open or completely closed forms, with averaged separation angles of 86° and 62° and theoretical R_g values of 29.17 Å and 26.03 Å for the open and closed forms, respectively. These results are coherent with the R_g extracted directly from the experimental curve ($R_g = 28.50 \pm 0.06$ Å), and shows that free POP in solution exists in a dynamic equilibrium between a fully open and a closed conformation.

The low intensity of the scattering profile obtained using COSMiCS for peak II of free POP was a serious limitation for the success of the EOM. For that reason I do not present these results. Nevertheless, given the low R_g value associated to this peak ≤ 25 Å, it is unlikely that open structures are present.

MD simulation	Production run interval (ns)	Sampled structures
MD1	0-1800	1 of every 20
MD2	0-50	All
MD3	90-300	1 of every 2
MD5	0-40	1 of every 2
TOTAL GENERATED FRAMES		**14785**

TABLE 6.3. *Intervals and structures from MD simulations taken for the EOM analysis*

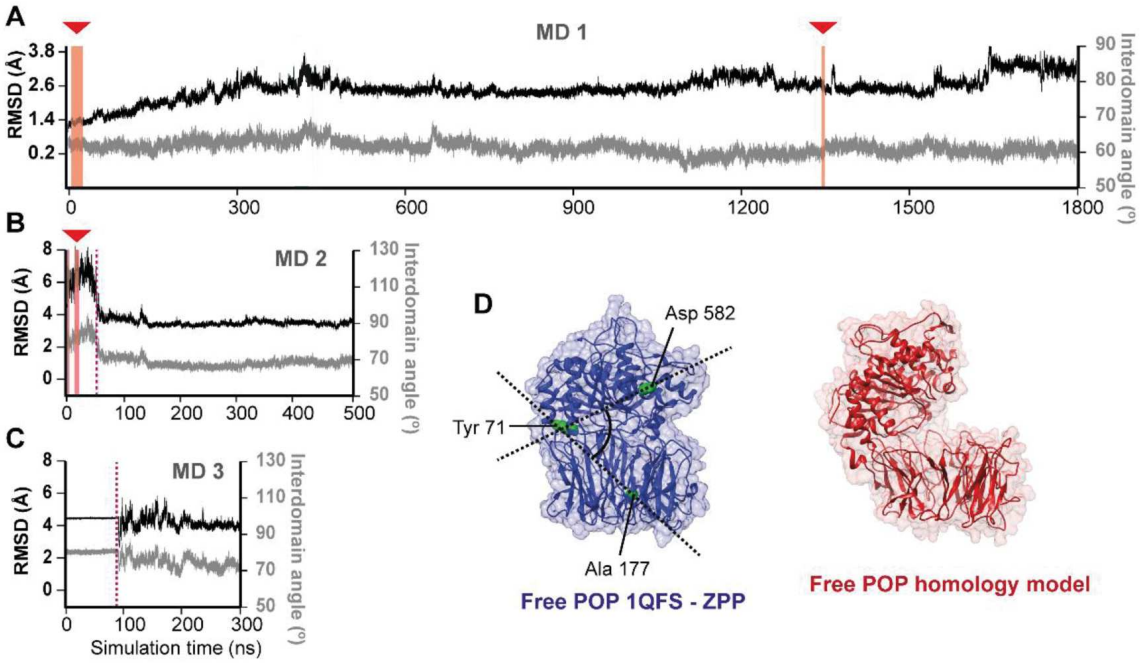


FIGURE 6.29. *MD simulations of free POP. A) 1.8-μs trajectory of MD1 simulation of the closed form of free POP. B) MD2 simulation of the homology model of free POP without constraints. Dashed line show the interval which has been selected for further data analyses (0-50 ns). C) MD3 simulation of the homology model of free POP with α-carbon constraints during the first 90 ns of the trajectory (dashed line, excluded from data analyses). RMSD (in Å) with respect to the initial structure is shown in black, and interdomain angle between residues 582, 71 and 177 is shown in gray. The sections in red and green correspond to the intervals selected by EOM of curves corresponding to peaks I and II, respectively (see chapter 2). D) The starting X-ray structure of POP in a closed conformation with the inhibitor removed is shown in blue (PDB entry 1QFS); the porcine POP homology model of Aeromonus punctata POP in an open conformation (PDB entry 3IUJ) is shown in red. The residues used for the determination of interdomain angle are displayed as green spheres. Figure extracted from [344]*

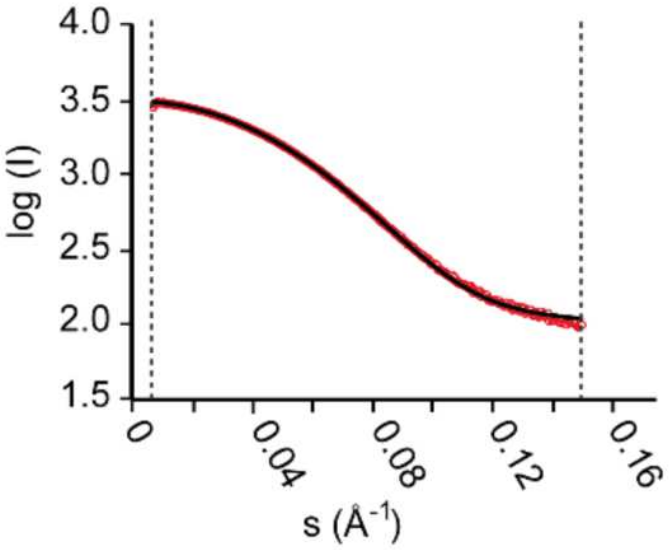


FIGURE 6.30. *EOM fitting (black) of peak I.*

6.3.4 Material and Methods

6.3.4.1 Online gel filtration coupled to SAXS

HisTag-cleaved samples of free and ZPP-bound POP were subjected to an online Superdex 200 10/300 SEC column coupled to EMBL beamline P12 of PETRA III (DESY, Hamburg, Germany) with a PILATUS2M pixel detector (DECTRIS, Baden-Daettwil, Switzerland). The column was run at 0.35 ml·min^{-1} to record 1 frame per second (X-ray wavelength 1.24; momentum transfer covered 0.007–0.444 Å^{-1}). The scattering profiles of all frames were inspected, and strangely behaving profiles were discarded. The scattering profiles corresponding to the pure buffer frames of free and ZPP-bound POP datasets were averaged and subtracted from all profiles with MATLAB®. The same program was used to average the subtracted scattering profiles of monomer species of free and inhibited POP, and to derive forward scattering (I(*0*)) and R_g from the Guinier approximation. *P(r)* distribution functions were obtained with the GNOM program [311].

6.3.4.2 Molecular Dynamic simulations

MD simulations were carried out by Martin Kotev (Joint BSV-CRG-IRB Research Program in Computational Biology, Barcelona), under the supervision of Dr. Victor Guallar.

All free-POP MD simulations were performed with AMBER12 software [345]. The ff99SB force field [346] for proteins was used, and explicit water molecules were incorporated as the TIP3P water model [347]. Protein structures were neutralized, and additional sodium and chloride ions were added to simulate physiological saline solution. Protein

plus ions were then solvated in pre-equilibrated water molecules in a truncated octahedron box with a 15 layer. After energy minimization, the temperature was progressively raised to 300 K with constant pressure dynamics. All production runs were performed with 2.0 fs steps in NPT ensemble (1 bar, 298 K). The shorter MD simulation for ZPP-bound POP was computed with the Desmond molecular dynamics program [348]. The OPLS-AA force field and TIP3P water model were used [347]. The default relaxation protocol in Desmond was used, followed by the production run in the NPT ensemble.

MD1 The long trajectory of MD1 simulation showed the evolution of the closed conformer of POP (Figure 6.29A). A global view of the structures displaying maximal interdomain angles showed that they corresponded to a general, yet slight, interdomain separation. Moreover, the highly flexible loops A and B were displaced from their original position; this favoured the existence of a significant cavity in the interdomain loop region that exposed the catalytic centre. At the same time, a second small cavity appeared near His 680 loop and the first polypeptide hinge. The exposure of buried areas has a significant impact in the global SASA of POP: the solvent exposure also reached maximal values simultaneously to interdomain separation. The maximum SASA were of 30100 $Å^2$ and 29100 $Å^2$ for the peaks at 422 and 651 ns, respectively. Taken together, these results highlight the high solvent exposition of the semi-open POP conformers. The reiteration of two consecutive interdomain angle maximums during the same MD simulation trajectory indicates that this breathing of POP structure might occur periodically.

MD2 The first MD simulation starting from the homology model of free POP is shown in Figure 6.29B. The initial part of the trajectory (50 ns) displayed a significant RMSD and interdomain angle fluctuations, but the overall value of the interdomain angle indicated that the open structure of free POP was maintained during this interval (Figure 6.29B). However, the RMSD and intedomain angle values drop after this first interval of 50 ns, probably as a result of a spontaneous closing of the structure. From 160 ns to the end of the trajectory (510 ns) only the closed conformation was exclusively present. The interdomain angle of the closed conformer generated in the second part of the trajectory was found to be higher compared to that of MD1 simulation ($67 \pm 6°$ vs $61 \pm 2°$, respectively); nevertheless, this difference was attributed to the differences in the starting structures.

MD3 The second MD simulation starting from POP homology model was carried out including α-carbon restraints during the first 90 ns of the trajectory in order to allow side chain relaxation (Figure 6.29C). Obviously, RMSD and interdomain angle during this relaxation interval remained fixed and were not considered in further analyses. After the elimination of the α-carbon restraints, RMSD and interdomain angle underwent marked fluctuations, which reflected the high degree of flexibility of the open structure. Of interest, the interdomain angle and the global SASA (32200 $Å^2$, compared to 28900 $Å^2$ averaged for MD1 simulation) were coherent with a long-range interdomain separation. Together, these results indicated that the homology model of the relaxed open conformation was relatively

stable. The initial side chain relaxation of the homology model structure was found to be a critical factor for the stability of the open conformer of free POP in MD simulations.

MD4 and MD5 Two MD simulations of inhibited POP, with shared closely similar features, were generated. By comparing the RMSD and interdomain plots, both trajectories reflected high structural stability and poor flexibility (Figure 6.23). During the initial equilibration period, the RMSD slightly of MD4 and MD5 simulations increased to a plateau approximately at 1.6 Å and 1.5 Å, respectively. At this point, only small amplitude thermal motions occurred during the rest of the production run. Interdomain angles remained constant during the whole trajectory, indicating the absence of interdomain separations. In spite of the high degree of similarity between the two MD simulations, only MD5 simulation was chosen for further studies due to the presence of the covalent bond between the inhibitor and Ser 554 of the active site. Although this characteristic did not affect the general evolution of the structure during the MD simulation, it might influence the local configuration of the active site environment. This would lead to wrong conclusions when analysing the active site configuration at atomic detail.

6.3.5 Discussion

COSMiCS analysis of SEC-SAXS data

SEC-SAXS data are becoming popular during the last years due to the necessity of obtain SAXS data from monodisperse samples for subsequent structural analysis. (see section 1.3.2.3). It is therefore not surprising that all modern BioSAXS beamlines have incorporated on-line SEC-SAXS as sample environment. Moreover, this method presents the big advantage of the possibility of measuring data from other techniques simultaneously such as UV-visible absorbance, light scattering or refractive index, which provide information about the concentration and the molecular weight of the protein for each frame. However, there are still cases where the separation of the species is not complete, and the chromatogram presents overlapped areas that are populated by more than one species. For these situations, it is necessary to decompose the data to the individual scattering and concentration profiles. Only after this decomposition, it is possible to analyze the SAXS curves from the pure components and to derive relevant structural information.

As we have seen along this chapter, COSMiCS has the capacity to decompose these SEC-SAXS datasets from overlapped peaks into the pure scattering and concentration profiles. However, according to our results using synthetic data, there are several factors that affect the capacity of decomposition. One of these factors is the level of noise. High signal-to-noise datasets contain enough information to perform an accurate decomposition of the pure components. Nevertheless, low signal-to-noise datasets present difficulties in the decomposition with COSMiCS, and the accuracy of both the spectra and concentrations are not correct. In the presence of noise, the information content of the data does not permit resolve the high rotational ambiguity of the data. This can be in theory improved by

using chromatographic columns with a reduced dilution effect, and/or high-brilliance synchrotron radiation sources as both improvements allow the acquisition of curves of higher quality.

Our results demonstrate that there are ways to decrease the degrees of freedom of the system by adding more information in order to be able to analyze low signal-to-noise datasets. Nowadays, the most common approach to decompose overlapped peaks uses Gaussians as concentration constraint, which decreases the ambiguity and, therefore, allows an accurate decomposition. This approach is the one used in the software US-SOMO [78, 79] (see section 1.3.2.3), and it has the advantage of obtaining good decomposed curves even in extremely complex systems with very overlapped populations. This approach, however, can lead to overfitting in some cases. The approach that we have used, with the aim of decreasing the ambiguity, is the use of local-rank information as a constraint for COSMiCS (see section 6.2.3). The local-rank information gives us the number of species that are present in different regions of the chromatogram and, used as a COSMiCS constraint, simplifies the decomposition enough to obtain correct decomposed spectra profiles (Figure 6.13A). This local-rank information can be obtained by analysing the chromatogram with the R_g profile or the EFA. Our results with synthetic data indicate that the sensibility of one or the other is defined by the nature of the system. The R_g analysis is more sensitive than EFA for systems composed by species with very different Rg, even if the molecular weight is similar. Moreover, R_g seems less affected by noise than the EFA. Conversely, EFA, which accounts by changes in the complete momentum transfer range, seems more sensitive to identify different species of a similar size [349]. For this, it is recommendable the use of the combined information from the both methods to obtain the maximum information about the peak composition. Our approach has limitations of in very complex systems, i.e. a system with an important degree of overlapped region and minimal monodispersity. In this case, the detection of the species along the chromatogram becomes difficult, and the ambiguous regions would be too large for the program being able to decompose.

Both decomposition approaches, US-SOMO and COSMiCS are able to derive the pure scattering profiles of the species involved in the system. However, they present limitations in the correct decomposition of the concentration profiles, even after using constraints. The wrong determination of the concentrations leads to a wrong determination of the original relative I(*0*) between the species and, consequently the relative molecular weight. The reason of this is that the shape of the chromatogram obtained by SAXS depends on the molecular weight of the species involved, generating chromatograms that do not reflect the real concentration of the species. To overcome this limitation, it is necessary introduce additional information to the software. For that, we implement in COSMiCS the use of UV-visible absorbance data (see section 6.2.4), which is commonly obtained during the acquisition of the SEC-SAXS data. The addition of UV-visible information allows COSMiCS to decompose the correct spectra profile, as well as the pure concentrations (Figure 6.15). However, we demonstrate that the use of absorbance data is not useful unless is used together with local-rank information (see Figure 6.16).

SEC-SAXS study of the conformational fluctuations in POP

The tandem SEC-SAXS experiment allowed the isolation of the monomeric species of free POP and ZPP-bound POP, and the concomitant measurement of the monomer scattering profiles without any oligomer contamination. In order to monitor the elution of POP conformers through the gel filtration column, SEC-SAXS I(*0*) chromatograms were generated: in the case of free POP, two different enzyme forms were partially resolved by the tandem SEC-SAXS. This was confirmed by mass spectrometry experiments that showed the purity of the preparation and the absence of proteolytic products. However, the slow exchange regime between the two free POP species observed here might not reflect the real situation. It should be stressed that species eluting in the gel filtration matrix are in an altered environment, with high local crowding and strong interactions with the stationary matrix [350].

The PCA and EFA delimited the monodispersity region for peak I of free POP, and indicated the purity of ZPP-bound peak. By COSMiCS analysis or by averaging the SAXS curves of these pure regions, high quality scattering profiles for both peaks were obtained. R_gs extracted corresponding to peaks I free POP and ZPP-bound POP using Guinier's approximation were 28.50 $\pm$ 0.06 Å and 27.40 $\pm$ 0.06 Å, respectively. The difference in R_g values between these two species indicates that they exist in different conformational states. In this regard, the R_g of free POP is much higher compared to that calculated for the X-ray closed structure 1QFS (25.82 Å), suggesting that large conformational rearrangements are occurring in this species. In order to have a more detailed estimation of the spatial arrangement of atoms in both species, the pair-distance distribution function *p(r)* was calculated for the pure SAXS curves of the two species with GNOM program [311] (ATSAS data analysis software, Figure 6.31. This distribution is sensitive to small changes in the protein structure, which alter the distribution of atoms, and lead to changes in the *P(r)* distribution (see Paper II). In the case of free POP, the pure curve corresponding to free POP yielded a multimodal distribution, pointing that the tertiary structure of this species features an irregular global shape. In contrast, for the pure curve of ZPP-bound a bell-shaped distribution was obtained. Hence, ZPP-bound POP seems to adopt a more globular shape.

These SAXS curves in combination with conformations derived from multiple MD simulations were used to study the structural features of POP in these two conditions. NMR (see section 6.3.2.3, 6.3.3.4 and Paper II) and SAXS experimental data complemented by MD simulations of different POP structures showed that free POP exists in a μs-ms equilibrium between completely open and closed conformers. According to our experimental data, this long-range opening and closing transition consists in a composition of several motions of different amplitudes rather than a single hinge motion. In contrast, inhibitor-bound POP appears exclusively in a closed conformation.

Overall, the analysis of MD simulations correlated with experimental data stresses the highly dynamic nature of free POP at different time scales. MD simulations also showed that interdomain opening have important effects in POP structure, exposing the active site and other buried areas to the solvent. MD simulations indicated faster events involving

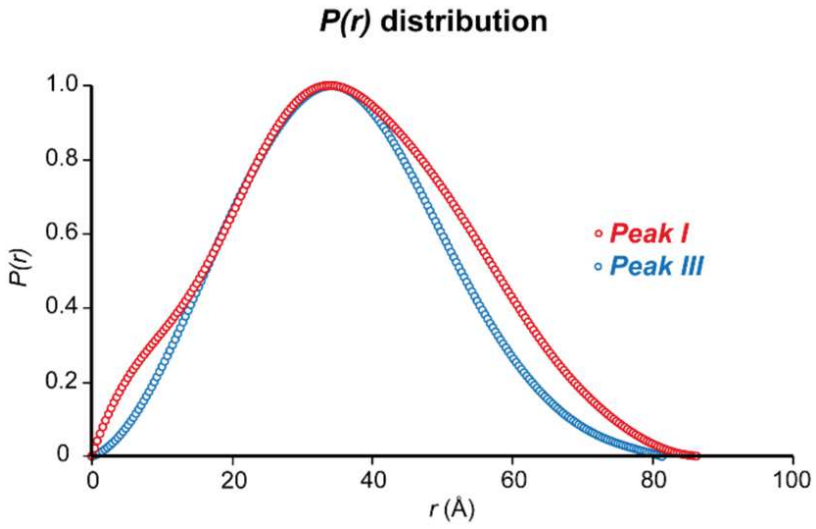


FIGURE 6.31. *P(r) function of species corresponding to peaks I (red) and III (blue).*

loops surrounding the active site. The detailed analysis of these unstructured regions discloses the mechanistic role of residues Asp 149 – Lys 172 and Asp 642 – Arg 643 in the active configuration of the catalytic center of POP. Of interest, the active configuration of the active site only occurred in the closed conformation, pointing that interdomain fluctuations switch between active/inactive conformations.

In the case of the scattering profile of peak II extracted by COSMiCS, the corresponding radius of gyration was between 24 and 25 Å depending on the method used for deriving the curve. This result is consistent with the order of elution observed in the online SEC-SAXS, and also discloses significant structural changes between the two species. Unfortunately, the low intensity of the scattering profile of peak II precluded the fitting of theoretical data. The R_g value of this peak is slightly smaller than the POP crystallographic structure 1QFS. We speculate that this secondary POP conformer arises from a tight arrangement of β-propeller blades [333]; the collapse of the internal tunnel of this domain would result in smaller Rg. However, further work is required to elucidate the structure this minor form.

Taken together, our results yielded unprecedented evidences of the conformational equilibrium of POP in solution, and the effects of inhibitors in this process. The preliminary analysis of SEC-SAXS I(*0*) chromatograms, R_g values and *p(r)* functions confirmed that POP undergo important conformational changes in the presence of inhibitors. Our SEC-SAXS analysis could provide an efficient tool to systematically characterize the conformational perturbations exerted by pharmaceutical molecules to POP. These studies, in combination with activity assays could represent an invaluable tool to study the structural bases of the enzymatic function of POP and the development of efficient inhibitors.

Chapter 7

Simultaneous use of COSMiCS with multiple techniques

7.1 Introduction

In previous chapters, we have use COSMiCS to decompose data from polydisperse samples in order to perform a structural analysis of individual coexisting species (i.e. amyloids or conformational fluctuations). Chemometrics is a powerful method and its application for the analysis of SAXS data, as we have seen along this thesis, allows the study of these complex polydisperse samples that are difficult to study otherwise. However, due to the inherent ambiguity of the decomposition method, it is necessary provide additional information in order to achieve accurate solutions. We have seen previously that it is possible to decrease the ambiguity by adding more information using different constraints, such as *Non-negativity*, *Closure* or *Equality*. These constraints change depending on the system of interest and the information available. For the study of amyloids, for instance, we have used the combination of different representations of the SAXS data in order to add more information, and we have demonstrated to be an excellent strategy to improve the solutions (see chapter 4). Another strategy broadly used for reducing ambiguities in decomposition methods is to enrich the dataset with additional experimental information, simultaneously measured, from complementary techniques [351–355]. This strategy is very powerful because not only allows to reduce ambiguity to achieve more accurate solutions, but it also yields additional information that is not available with the use of a single technique. For instance, in the systems studied we have seen that SAXS is a powerful technique that allows study biomolecular systems in terms of structure, dynamics and transformation kinetics. However, other techniques can provide complementary information that cannot be captured by SAXS, such that secondary structure or hydrogen bond formation that, in the context of complex polydisperse systems, can be extremely valuable.

The most challenging system that we have addressed during this thesis is the amyloid fibrillation and therefore it is the system that can benefit most from the combined use of multiple techniques. The selected technique to simultaneously measure with SAXS was fluorescence in the presence of probes sensitive to different oligomeric or fibrillar species. We have performed a series of experiments measuring fluorescence in parallel with SAXS along several fibrillation processes. Time did not permit the analysis of these datasets or the

implementation in COSMiCS but it will be the subject of future work. In this chapter I will describe the measured datasets, as well as the potential applications, and the experimental limitations.

7.2 Flourescence as additional source of information

One of the amyloid formation process that we have studied along this this thesis is the fibrillation of α-synuclein (see section 2.3.3), which has been monitored by thioflavin T (ThT), which detects the presence of amyloid fibrils. However, the understanding of the aggregation process requires the identification of all the conformational states and oligomeric structures adopted by the polypeptide chain during this process. We were able to decompose the pure signal of the main species present in the system using COSMiCS. However, one of the limitations of ThT is that this dye is only sensitive to the presence of fibrils, but not any of the other coexisting species. The identification and characterization of prefibrillar states preceding the formation of well-defined fibrils are of particular interest both because of their likely role in the mechanism of fibril formation, and their critical role in the pathogenesis of amyloidogenic diseases [121, 131, 356–360]. Recently, novel chemically defined pentameric thiophene derivatives, denoted luminescent conjugated oligothiophenes (LCOs), have been described as useful probes to monitor fibrillation processes [361, 362]. In contrast with the traditional small hydrophobic fluorescent amyloid ligands, LCOs contain highly flexible conjugated thiophene backbones. When these dyes bind to protein aggregates, the rotational freedom of this flexible thiophene backbone is restricted and the emission properties of the probes are affected in a conformation-dependent manner. Hence, an optical fingerprint is obtained, unique to the structure of the protein. This phenomenon has been used to distinguish prion strains and for discrimination of heterogeneous Aβ plaques [361, 363]. These oligothiophenes have been also tested in insulin, lysozyme and prion protein amyloid fibrils [364–367]. These evidences indicate that LCOs might be superior to conventional amyloid ligands such as ThT and ThS in identifying prefibrillar states during the fibrillation process. In addition, LCOs have high multiphoton excitation capabilities, allowing real-time imaging of protein aggregates in vivo in animal disease models [361].

We have used for our experiments the p-FTAA, q-FTAA and h-FTAA LCO probes. A preliminary test with each probe was performed in Nilsson's group (University of Linköping, Sweden) to test their behavior with α-synuclein (αSN). The fibrillation process using p-FTAA was performed at 8.0 mg/ml of αSN in 20 mM Na_3PO_4 buffer with 150 mM NaCl with a final concentration of p-FTAA of 300 nM. The excitation wavelength used was 430 nm. Figure 7.1A shows the chemical structure of the probe p-FTAA and the complete fluorescence spectra for the free probe and with αSN, measured in different times during the fibrillation process. The spectra of the probe change completely along the fibrillation, both in intensity and in shape, displaying distinct and specific spectroscopic changes for early and late formed species. Two peaks appear at 510 and 540 nm for early species, being the peak at 510 nm smaller than the peak at 540. As the fibrillation advance, the intensity of the signal increases and the shape changes again, with a small shift of the peak at 505 nm

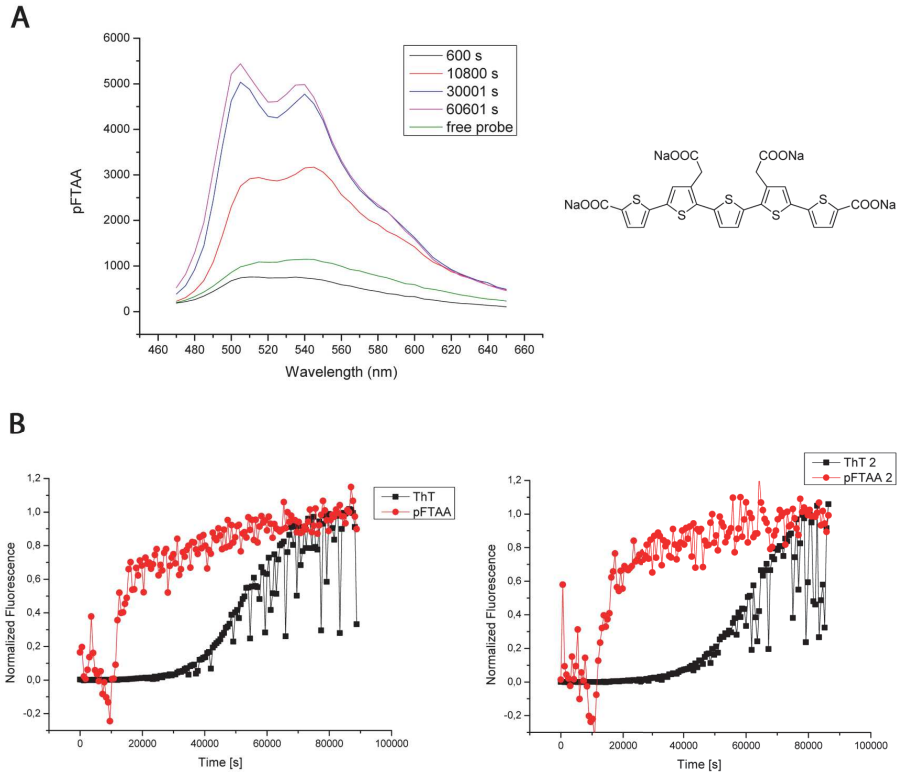


FIGURE 7.1. *A) Chemical structure (right) and emission spectra (left) of p-FTAA with α-synuclein in 20 mM Na_3PO_4 buffer with 150 mM NaCl at different times along the fibrillation process. B) Normalized fluorescence along the fibrillation process for p-FTAA at 510-545 nm (averaged) and ThT signal at 480 nm. The two graphs correspond to two different experiments to test reproducibility. Analysis performed by Per Hammarström.*

to 510 nm, which become more intense than the peak at 540 nm. The intensity of p-FTAA at 510-545 nm (averaged intensity) was monitored during the αSN fibrillation in parallel with the ThT fluorescence (Figure 7.1B). The Lag phase, monitored by ThT, lasted 8h and, at that point, the p-FTAA fluorescence reached the maximum intensity and the shape of the profile changed completely. Two samples were measured with very similar results, showing an acceptable reproducibility (Figure 7.1B). These results indicate that p-FTAA is sensitive to the appereance of prefibrillar oligomeric species.

The same test was performed with q-FTAA in the same conditions than the previous experiment. The concentration of the probe was in this case 600 nM and the excitation wavelength was kept at 430 nm. Figure 7.2A shows the fluorescence spectra measured at the same time points than the previous experiment. The spectra of this probe also changes along the fibrillation, with a major peak appearing at 500 nm in the last part of the lag phase of ThT. The increase in intensity and the change of shape occurs early in the lag phase (Figure 7.2B). Both samples tested gave very similar kinetics.

Finally, the h-FTAA probe was tested under the same conditions and a final concentration of 300 nM. The probe is sensitive also to species that appears at very early stages of fibrillation, with an increase of intensity and a change of shape, with a maximum at 545

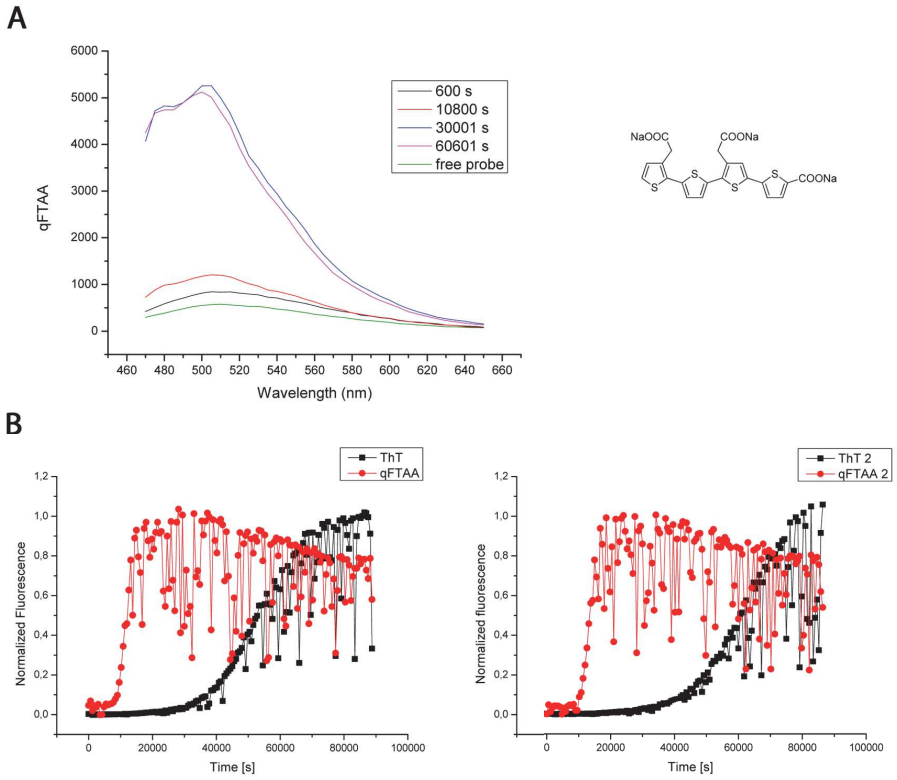


FIGURE 7.2. *A) Chemical structure (right) and emission spectra (left) of q-FTAA with α-synuclein in 20 mM Na_3PO_4 buffer with 150 mM NaCl at different times along the fibrillation process. B) Normalized fluorescence along the fibrillation process for q-FTAA at 500 nm and ThT signal at 480 nm. The two graphs correspond to two different samples to test reproducibility. Analysis performed by Per Hammarström.*

nm. As the fibrillation advances but still during lag phase, a change in shape occurs, appearing two peaks at 550 and 590 nm in the spectra (Figure 7.3A). Unlike the other probes, h-FTAA shows fluorescence intensity response from the beginning of the fibrillation process, with a steep decrease at early stages that slowly grows along the lag and exponential phases (Figure 7.3B).

This study was carried out in collaboration with Profs. Per Hammarström and Peter R. Nilsson (Department of Chemistry, IFM, Linköping University, Sweden), who provided the LCO probes and performed the fibrillation tests with αSN.

7.3 Experimental set-up

BioSAXS beamlines are sensitive to the need of additional information from other techniques and, therefore, this possibility is being implemented as an option for the user, for example the measurement of UV-vis absorbance, refractive index or DLS during the SEC-SAXS experiments [70, 73, 368]. Monitoring a complex phenomenon such as amyloid fibrillation by SAXS is difficult (see section 4.4), and performing the experiment in parallel with another technique is an extremely challenging task. The main issue that must be

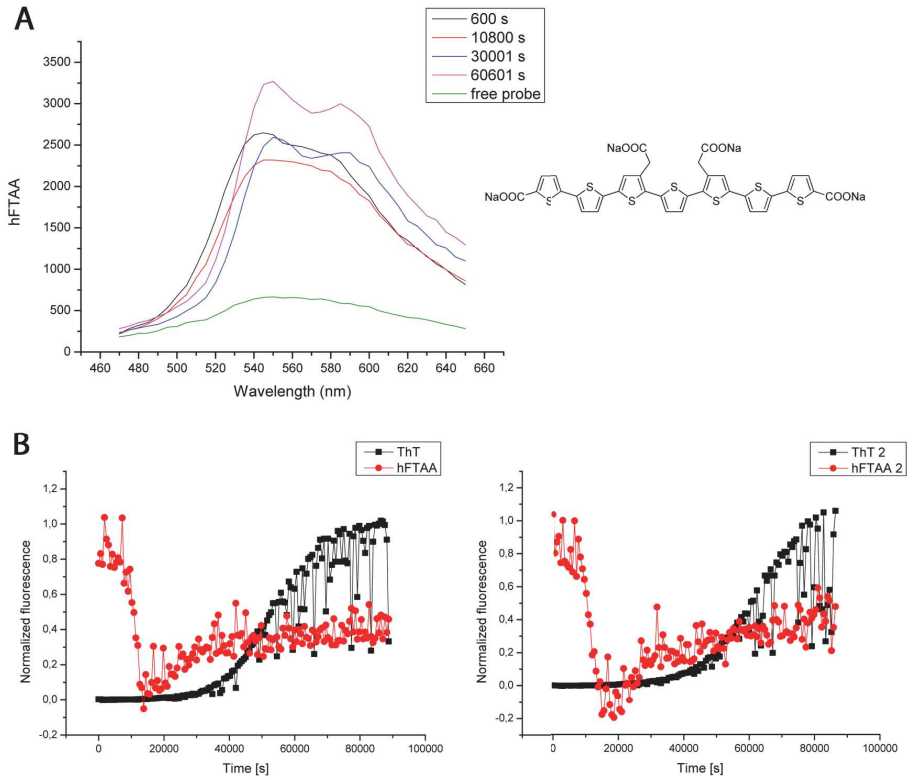


FIGURE 7.3. *A) Chemical structure (right) and emission spectra (left) of h-FTAA with α-synuclein in 20 mM Na_3PO_4 buffer with 150 mM NaCl at different times along the fibrillation process. B) Normalized fluorescence along the fibrillation process for h-FTAA at 550-590 nm (averaged) and ThT signal at 480 nm. The two graphs correspond to two different samples to test reproducibility. Analysis performed by Per Hammarström.*

solved is the need to record fluorescence data corresponding to the same time-point measured by SAXS. This is mandatory in order to subsequently use the previously described decomposition method (see section 4.3). In the previous studies (chapter 4), fibrillation processes were performed in a plate-reader for which every individual well was considered as an individual sample (see section 4.4). The distinct samples/wells evolved at slightly different rates and the collected data do not follow a perfect sequential fibrillation model. This lack of reproducibility caused the scattered profile of the decomposed concentrations (see sections 4.1.6 and 4.2.6). However, this problem could be treated due to the general trend observed in the data that followed realistic fibrillation kinetics.

To allow the simultaneous fluorescence measurements, we have used a different methodology. The complete fluorescence spectrum for each probe was measured with *ProbeDrum* (Figure 7.4, http://probedrum.se), a titrating spectrometer created by Dr. Thom Leiding and Dr. Sindra Petersson. *ProbeDrum* is able to measure UV-visible absorbance over the whole detection range (from 220 to 790 nm), static light scattering (SLS) (650 or 635 nm laser), and fluorescence from 260 to 650 nm. The optical lines monitoring the sample comprise a CCD-based detector and a total of 16 different light sources in two spatial orientations. The sample cell is a standard optical cuvette, base 12.5x12.5 mm, with an internal volume of 1 ml or 3 ml with a Z Dimension (Z) of 8 mm, which is large enough

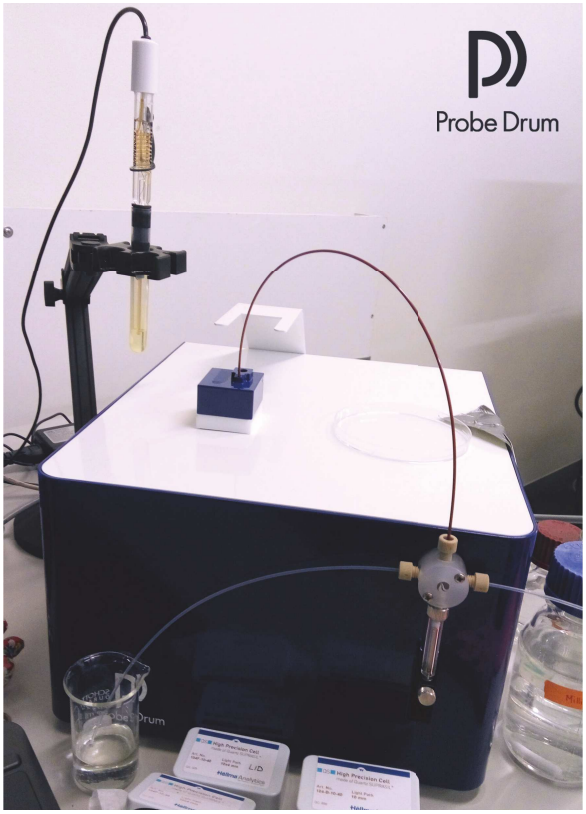


FIGURE 7.4. *Photo of the ProbeDrum set-up.*

to allow multiple extractions for SAXS measurements. The use of a single cuvette is an advantage of this new fibrillation protocol because it avoids different stochastic onsets of the kinetics, and it makes the handling of both the samples and the data easier, in addition of reducing the amount of protein required. *ProbeDrum* allows the agitation of the sample through a small spin magnet that is introduced in the cuvette that is located in one of the internal sides during the experiment.

The first step is the optimization of the set-up to find the best conditions of the experiment before performing the measurements in parallel with SAXS in a synchrotron. For that, we need to transfer the fibrillation conditions of αSN from the classical protocol used in a plate-reader, with beads and agitation, to the stirring set-up of the *ProbeDrum*.

7.3.1 Optimization of the experiment

In order to find the best conditions for the fibrillation of αSN in the *ProbeDrum*, we performed a series of experiments changing different parameters of the protocol. Table 7.1 shows the different conditions tested for the fluorescence experiments. All experiments were performed at 37 C, and αSN was dissolved and filtered in a PBS buffer at pH 7.4. The optimized parameters were:

Concentration of α-synuclein

	Concentration αSN	Probe	Concentration probe	Excitation wavelength (nm)	Integration Time (ms)	Number of scans	Wavelength Range (nm)
1	5.23	ThT	20 µM	392	500	9	240-720
2	8.13	ThT	20 µM	392	400	9	410-720
3	**8.13**	**ThT**	**20 µM**	**392**	**500**	**9**	**410-720**
4	7.80	ThT	20 µM	363	5	20	240-720
5	7.80	ThT	20 µM	363	15	20	240-720
6	7.80	ThT	20 µM	375	300	20	240-720
7	7.80	ThT	20 µM	375	500	20	240-720
8	8.10	ThT	20 µM	363	5	20	240-720
9	8.10	ThT	20 µM	363	15	20	240-720
10	8.10	ThT	20 µM	363	300	20	240-720
11	8.10	ThT	20 µM	363	500	20	240-720
12	8.86	p-FTAA	300 nM	392	400	9	410-720
13	8.86	p-FTAA	300 nM	392	500	9	410-720
14	8.20	p-FTAA	300 nM	449	500	15	460-700
15	8.20	p-FTAA	300 nM	449	800	15	460-700
16	8.20	p-FTAA	300 nM	449	1000	15	460-700
17	8.80	p-FTAA	1.2 µM	375	500	50	240-720
18	8.80	p-FTAA	1.2 µM	375	700	50	240-720
19	8.80	p-FTAA	1.2 µM	375	800	50	240-720
20	8.80	p-FTAA	1.2 µM	375	1000	50	240-720
21	8.80	p-FTAA	1.2 µM	351	300	50	240-720
22	8.80	p-FTAA	1.2 µM	351	1500	50	240-720
23	8.80	p-FTAA	1.2 µM	375	300	50	240-720
24	**8.80**	**p-FTAA**	**1.2 µM**	**375**	**1500**	**50**	**240-720**
25	8.88	q-FTAA	600 nM	392	500	9	450-720
26	8.88	q-FTAA	600 nM	392	600	9	450-720
27	8.80	q-FTAA	600 nM	449	1000	50	460-700
28	8.80	q-FTAA	600 nM	449	1500	50	460-700
29	8.80	q-FTAA	600 nM	449	2000	50	460-700

TABLE 7.1. *Optimization experiments performed with the ProbeDrum. Bold lines correspond to the optimized parameters for ThT and p-FTAA. The optimal parameters for q-FTAA and h-FTAA can be used for p-FTAA, just changing the concentration of the probe (1.2 µM for p-FTAA and h-FTAA and 2.4 µM for q-FTAA)*

The concentration used for fibrillation rangeed from 3 mg/ml to 12 mg/ml. The required time to form the final fibrils is inversely proportional to the concentration of the protein. Due to the limited beam-time available at the beamline, the concentration must be optimized in order to keep the total fibrillation experiment within 12 hours.

Concentration of the LCO probes

Previous experiments made with the LCO probes used a concentration of 300 nM (for p-FTAA and h-FTAA) and 600 nM (q-FTAA). With our set up that quantity was not enough to get a clear signal above the noise level. For this reason, we had to optimize these

concentrations in order to measure the spectroscopic details for each probe.

Excitation wavelengths

Each probe has an optimal excitation wavelength. However, this signal also causes direct scattering from the sample, which is recorded by the detector together with the emission signal of the probes. When the excitation and emission wavelengths are relatively close, perturbations in the emission signal are observed. To obtain the best results I had to choose a non-optimal excitation wavelength to be different enough of the emission signal.

Integration time

The integration time is the time of exposure for each experiment. Increasing it yields higher signal relative to the dark noise. However, it requires more time to perform the measurement, therefore it is necessary reach a compromise between acquisition time and quality of the data.

Number of scans

The increase in the Number of scans improves the quality of the spectra. However, in the same way than the integration time, a larger number of scans also increases the time required to record the data.

From the results of the experiments performed using the *ProbeDrum* we have defined the optimum parameters to be used during the experiments in the P12 Bio-SAXS beamline at PETRAIII in Hamburg (Table 7.2). The optimal results were obtaining using 8 mg/ml of αSN, that allows to perform the complete fibrillation process within 12 hours. The usual concentration of ThT, 20 μM, was enough to follow the fibrillation. However, larger concentrations of the LCOs probes were necessary to be able to visualize the details of the emission signal. We selected a final concentration of 1.2 μM for p-FTAA and h-FTAA, and 2.4 μM for q-FTAA. The spectra profile of ThT does not present changes in the shape, just a systematic increase in intensity (Figure 7.5) with a regular spectra shape that starts at 450 nm. However, LCO probes, for instance p-FTAA, show extended emission fluorescence spectra, starting at 400 nm, which is affected by the scatter of the signal at the excitation wavelength (Figure 7.6). For this reason, the optimal excitation wavelength was set to 375 nm for the LCO probes (far from the emission signal) and 392 nm for ThT. The total measurement time depends on both the integration time and averaged values, so it was necessary adjust the values to find an optimal compromise. ThT has a strong signal, so 20 scans were enough; however, LCO probes required 70 scans. Integration time of 15 ms gives enough signal-to-noise for ThT and the LCOs.

Finally, it is very important to test if the final fibrils obtained using the *ProbeDrum* were equivalent to the fibrils obtained with the classic protocol in the plate-reader. For that, we performed Transmission Electron Microscopy (TEM) experiments of the final fibrils. From the different experimental conditions, six samples were imaged with TEM. In

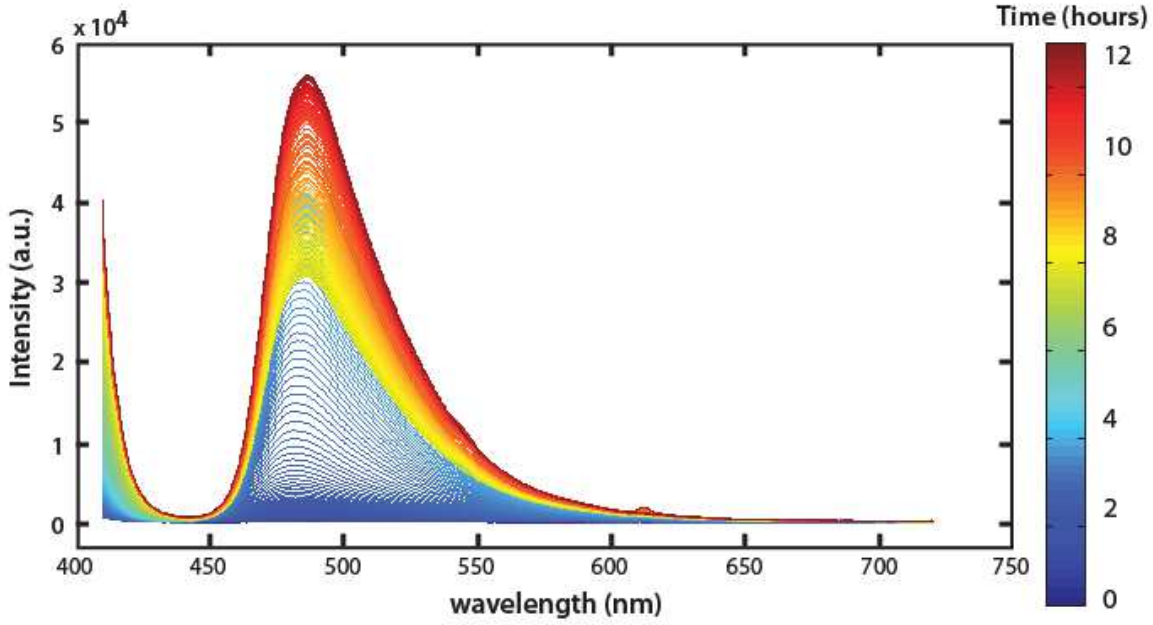


FIGURE 7.5. *ThT fluorescence spectra of the fibrillation experiment for 8.13 mg/ml αSN and 20 μM ThT, corresponding to the experiment 3 in Table 7.1. Total time of 12 hours.*

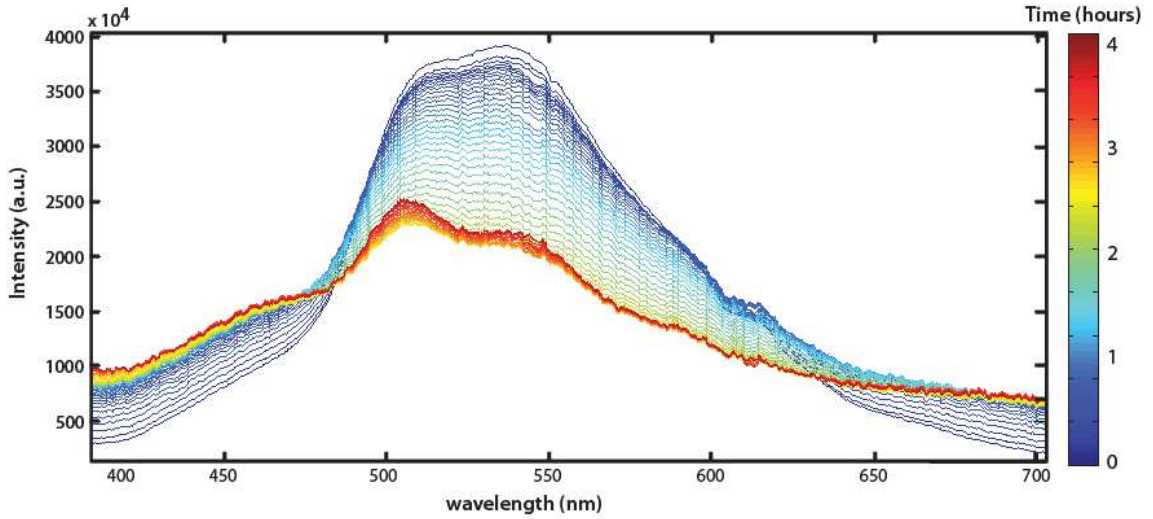


FIGURE 7.6. *p-FTAA fluorescence spectra of the fibrillation experiment for 8.80 mg/ml αSN and 1.2 μM p-FTAA for the first 4 hours of the fibrillation process. Changes in the shape are observed, such as the appearance of a peak at 510 nm. After this period, both the shape and the intensity remain stable. These spectra correspond to the experiment 3 in Table 7.1.*

all cases the resulting fibrils were similar to these obtained with the traditional protocol. Figure 7.7 shows a TEM image for one of the samples, corresponding to αSN with ThT.

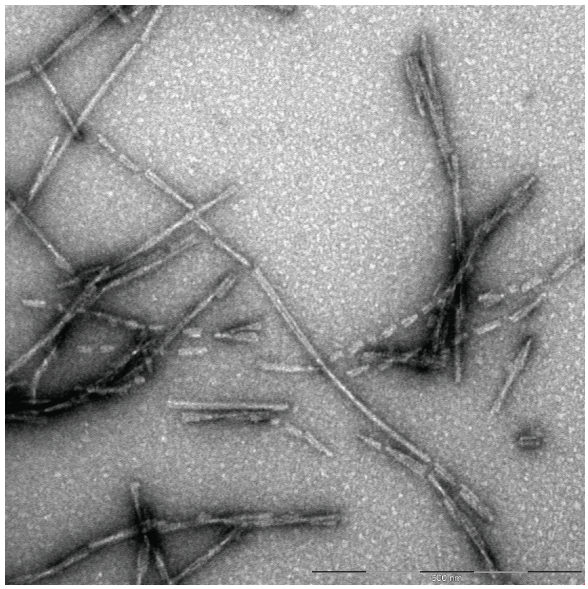


FIGURE 7.7. *TEM image from the ProbeDrum fibrillation from a sample of 8.1 mg/ml of αSN and 20 μM of ThT after 24 hours of fibrillation.*

7.4 Recorded datasets in parallel with SAXS

Fibrillation of α-synuclein and p-FTAA

A fibrillation experiment was performed in the *ProbeDrum* with αSN at 8.3 mg/ml and 1.2 μM p-FTAA along 10.3 hours, with the parameters selected previously (Table 7.2). The complete fluorescence and static light scattering (SLS) spectra were recorded each 6 minutes. During this time, samples were extracted from the spectrometer and brought to the beamline to measure the SAXS profile. A total of 46 scattering curves were obtained (Figure 7.8A) that, in general, displayed the expected behavior of a fibrillation process with a systematic increase in intensity and profile changes. However, it was possible to observe that some curves appear before or after were expected, which can be due to the polydispersity present in the system. Even if the whole process is performed in the same cuvette, fibrillation increases the viscosity of the sample, and some fibrils remain attached to the inner walls of the cuvette. In these conditions, it becomes difficult to extract a homogeneous sample. Figure 7.8B shows the intensity of SLS at 636 nm, which presents a decrease after 6 hours probably due to the decrease of fibrillated αSN in the cuvette. Figure 7.8C shows all the fluorescence spectra measured the first 8.2 hours. The spectra behave in the same way than the test performed previously (Figure 7.6), with two peaks appearing at 505 nm and 540 nm, with a decrease in intensity due to the scattering of the excitation signal. However, during this experiment there were some artifacts due to the sample volume of the cuvette after systematically withdrawing volume for SAXS measurements. For this reason, the spectra have been plotted separately to facilitate the visualization. The spectra corresponding to the final part of the fibrillation (from 8.2 to 10.3 hours) present artifacts, so they were removed from the figure. SLS data from that point were also removed due to

	ThT	p-FTAA/h-FTAA	q-FTAA
Concentration probe	20 µM	1.2 µM	2.4 µM
Excitation wavelength	392 nm	375 nm	375 nm
Integration time, ms (fluorescence)	500	3000	3000
Integration time, ms (static light scattering)	15	15	15
Number of scans	20	70	70

TABLE 7.2. *Optimal selected conditions of the fibrillation experiment for αSN in the ProbeDrum*

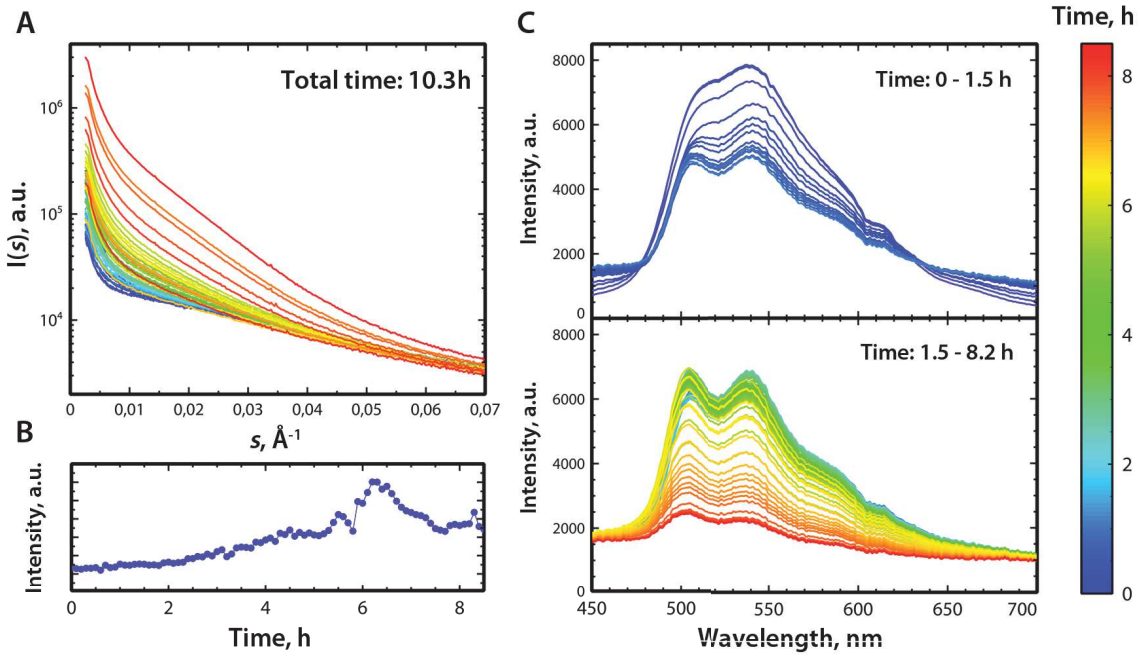


FIGURE 7.8. *p-FTAA series. A) SAXS curves of the fibrillation process of αSN at 8.3 mg/ml and 1.2 µM p-FTAA during 10.3 hours (46 curves). B) SLS intensity at 636 nm. C) Fluorescence spectra from 450 – 720 nm for the first 1.5 hours (14 spectra) and from 1.5 to 8.2 hours (67 curves).*

their lack of reliability.

After the fibrillation, a final sample was extracted to measure TEM. Figure 7.9 shows how the fibrils present the same features than the fibrils formed in the classic protocol and the fibrils obtained during the optimization without withdrawing samples and using ThT (Figure 7.7). Importantly, these results show that p-FTAA does not change the fibrillation process, generating the same fibrils than with ThT.

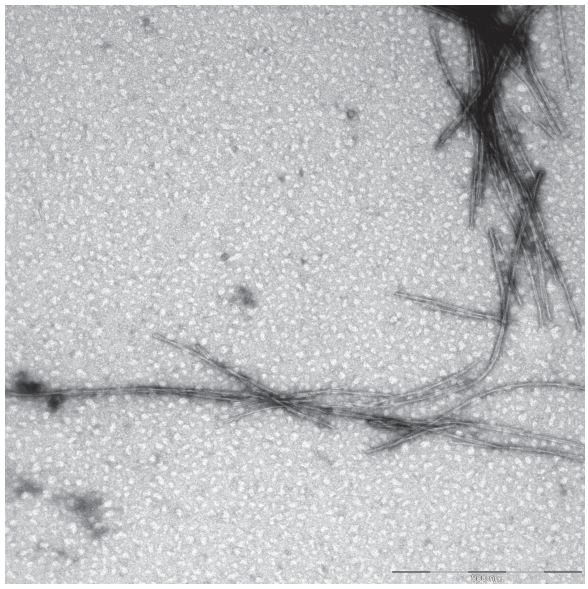


FIGURE 7.9. *TEM image of α-synuclein fibrils in the presence of pFTAA after 10 hours of fibrillation. The fibrils display the same features than these obtained using standard conditions with the plate-reader.*

Fibrillation of α-synuclein and q-FTAA

In the same way than the p-FTAA probe, a fibrillation experiment was performed with αSN at 8.6 mg/ml and 2.4 μM p-FTAA along 8.5 hours, with the previously selected parameters (Table 7.2). The complete fluorescence and SLS spectra were measured every 6 minutes. During this time, samples were extracted to record a SAXS curve. A total of 38 scattering curves were obtained (Figure 7.8A). Unlike in the previous experiment with p-FTAA, the fibrillation evolved very fast, reaching a stable point where the size of the species was not changing, although is possible to see that the shape of the SAXS curves slightly change. That fast fibrillation was probably caused by the fact than the starting point of the fibrillation already contained some aggregated αSN. The preparation of the experiment was done in the same way than the previous one, therefore, we speculate that the there was some problem during one of the purification steps and the starting point was not purely monomeric. The fast fibrillation was also evident in the SLS data (Figure 7.10B), which presented a fast increase in intensity at the beginning. Figure 7.10C shows all the fluorescence spectra measured along the fibrillation (8.5 hours). The spectra behave in the same way than the test performed previously (Figure 7.2), with a peak at 500 nm. The intensity decreases due to the scattering of the excitation signal, which should be subtracted for a correct analysis of the spectra.

After the fibrillation, a final sample was extracted to measure TEM. Figure 7.11 shows how the fibrils present the same features than the fibrils formed in the classic protocol and the fibrils obtained during the optimization without withdrawing samples and using ThT (Figure 7.7). Therefore, q-FTAA does not change the fibrillation process and generates the same fibrils than these in the presence of ThT.

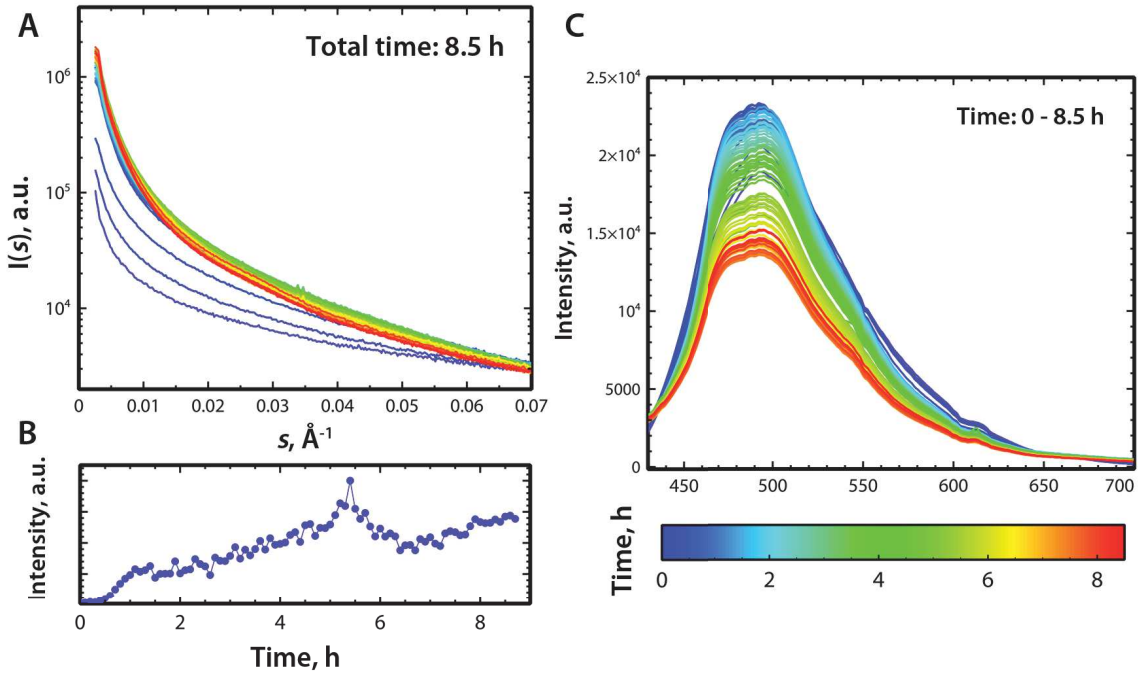


FIGURE 7.10. *q-FTAA series. A) SAXS curves of the fibrillation process of αSN at 8.6 mg/ml and 2.4 μM q-FTAA during 8.5 hours (38 curves). B) SLS intensity at 636 nm. C) Fluorescence spectra from 450 – 720 nm for the 8.5 hours (88 spectra).*

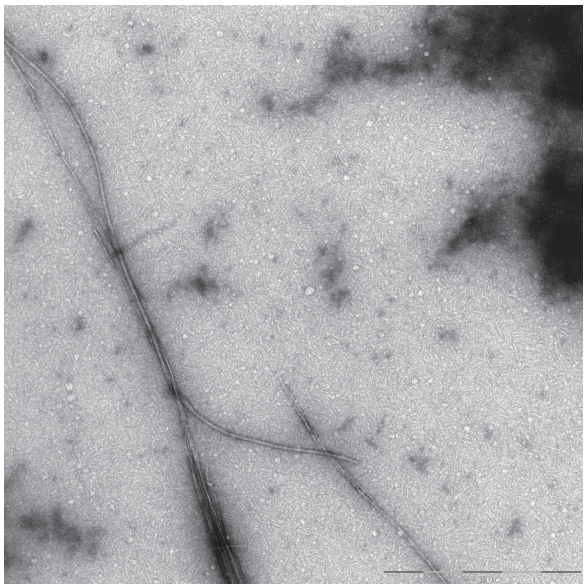


FIGURE 7.11. *TEM image that shows the fibrils of α-synuclein in the presence of qFTAA probe after 8.5 hours of fibrillation. Fibrils display the same features than the fibrils formed using standard conditions. Length of the scale bar, 1 μm*

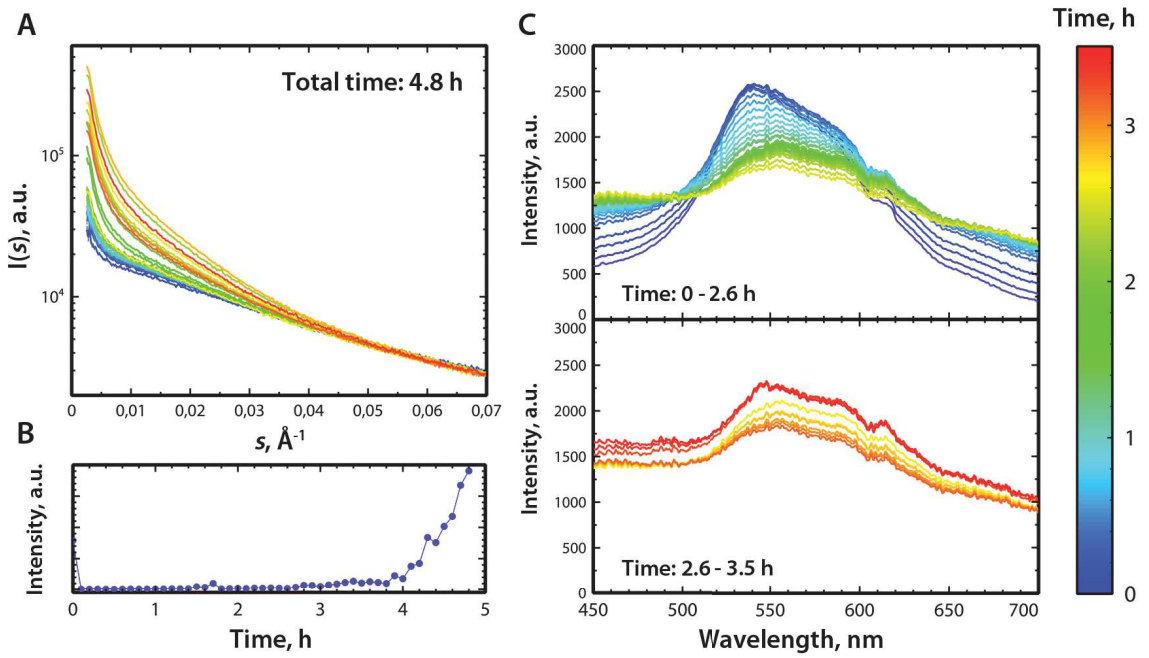


FIGURE 7.12. *h-FTAA series. A) SAXS curves of the fibrillation process of alphaSN at 7.3 mg/ml and 1.2 μM h-FTAA during 4.7 hours (25 curves). B) SLS intensity at 636 nm. C) Fluorescence spectra from 450 – 720 nm for the first 2.6 hours (26 curves) and from 2.6 to 3.5 hours (9 spectra).*

Fibrillation of α-synuclein and h-FTAA

Finally, a third fibrillation experiment was performed with αSN at 7.3 mg/ml and 1.2 μM h-FTAA, with the parameters selected previously (Table 7.2). Due to lack of time, the experiment was shorter (4.8 hours). A total of 25 scattering curves were measured (Figure 7.12A). The time was not enough to achieve full formation of fibrils but it was possible observe the evolution from the native species to more aggregated ones. This fibrillation started with a smaller species than the previous one, and consequently the development to fibrils was slower presenting a long-lag phase according to the SLS data (Figure 7.12B). Figure 7.12C shows all the spectra measured during 3.5 hours, divided in two plots for an easier inspection. The fluorescence spectra behaved in the same way than the test performed previously (Figure 7.3), appearing a peak at at 540 nm at the beginning of the fibrillation. It is not possible to see the two peaks that appeared in the test experiments because, as observed in the tests, these peaks appear at 8.3 hours (Figure 7.3).

We confirmed by TEM (not shown) that the final species were not fibrils.

7.5 Discussion

We were able to measure fluorescence spectra simultaneously with SAXS data and SLS during a fibrillation process, using probes that bind oligomeric species different than the fibrils. These probes can provide more information than the classical ThT due to the

specific shape of the fluorescence spectra that changes along the fibrillation process. In addition of reporting on the specificities of these oligomeric species, the simultaneous use of two techniques can improve the decomposition with COSMiCS by decreasing the ambiguity and, therefore, achieving more accurate results for complex systems. However, the measurement of the two datasets simultaneously still has experimental limitations. The use of a single cuvette to perform the fibrillation reduces, in theory, the problem that we had with the traditional protocol, i.e. the stochasticity of the fibrillation process. Nevertheless, the new set-up present other problems that are necessary to solve in order to combine the two techniques to obtain useful data to improve the decomposition. One of the problems that we face is the decrease in homogeneity within the cuvette as a consequence of the formation of fibrils. This problem becomes evident when we withdraw sample to perform SAXS measurements. As the volume decreases, some fibrils remain attached to the walls of the cuvette. The main consequence of this phenomenon is that the total concentration of fibrils in the remaining volume decreases, creating artifacts in the data and, more important, is no longer possible to apply the *Closure* constraint in the subsequent decomposition because the total concentration of protein is not constant along the dataset. Another problem that the new set-up presents is the scattering of the emission signal that produces disturbances in the emission spectra of the LCO probes. This signal should be subtracted from the emission signal before the decomposition analysis to avoid its contribution.

Another conclusion from these experiments is that the initial state of the αSN is very important for be able to record a good fibrillation dataset allowing the decomposition and detection of the main species present in the system. To achieve this aim a purely monomeric form of the protein must be used. Efforts are performed in the laboratory in Copenhagen to derive a robust protocol to obtain large quantities of highly pure monomeric αSN.

Due to the limit of time and the lack of precise data, I was not able to address the issues previously described, but they will be the goal of future work. As a perspective of these experiments, we have to define a clear and robust protocol to simultaneously measure fluorescence spectra and SAXS data without the problems of sample homogeneity while reducing the kinetic stochasticity. In that context, it seems that a plate-reader with the capacity to measure the full emission spectra can be the optimal set-up to derive the desired data. Then, the application of our COSMiCS program to the combined dataset of spectroscopic and SAXS spectra will allow us to define the spectroscopic fingerprints of both the soluble oligomeric species and the final amyloid fibrils. As a long-term application we envision the possibility of searching for these fingerprints in a cellular context using tissue samples. In this way, we can indirectly pinpoint the presence of particular non-fibrillar species in cellular assays.

Part III

DISCUSSION AND PERSPECTIVES

Chapter 8

Discussion and perspectives

SAXS and chemometrics for the study of complex systems

Along this thesis I have studied complex biological systems that are polydisperse, presenting multiple co-existing species differing in size, shape or conformation. This is a challenging task due to the instability of the species involved, their low and interdependent relative concentrations, and the difficulties to isolate the pure components. Usually, a single isolated component of these processes is monitored, due to the complexity of studying all the components simultaneously. Our strategy, however, is based in the study of such processes as a whole, decomposing the data of the mixture to obtain information of the pure components, instead isolating them. For that, we have chosen SAXS as technique due to its ability to monitor molecules of very different sizes in biologically relevant conditions, without the need for crystallization or freezing. Furthermore, another interesting property of SAXS, which I have exploited to study all the components of a mixture simultaneously, is that SAXS is an additive technique. This means that the scattered curve corresponds to the weighted average of the curves of the individual components. Therefore, SAXS data from a complex process is also complex, and we need tools adapted to this complexity in order to be able to disentangle and to interpret the data in terms of 3D structures and the kinetics/thermodynamics of these processes. The method that we have used during this thesis to analyze these complex systems is MCR-ALS (section 3.3.1), a chemometric tool that has been broadly used for analysis of mixtures monitored by diverse techniques such as NMR [369, 370] or fluorescence [371, 372]; and that has also been used to decompose SAXS data [303, 306, 373]. Multivariate resolution models are able to recover the pure response profiles of the species of an unresolved mixture that can be explained by a bilinear model using a limited number of components. This is the case for SAXS data recorded for systems in which variations of an experimental condition change the resulting scattering. However, one of the inherent limitations of decomposition methods is the ambiguity, which produces solutions that fit the experimental data but that are not correct. To address this limitation and improving the solutions obtained, it is necessary the use of constraints by the addition of more information about the system along the optimization. The use of constraints helps to improve the results, adapting the general method to the specificities of the system while avoiding non-physical solutions.

We have adapted MCR-ALS to the specificities of SAXS and implemented in a software package called COSMiCS (Complex Objective Structural analysis of Multi-Component Systems). The main development of COSMiCS is the simultaneous use of multiple representations of SAXS data in order to increase the discrimination capacity of the method. This feature is based on the different sensitivity of these representations to structural phenomena happening at different resolutions. COSMiCS performs the optimization process in a semi-automatic way as an attempt to facilitate the interpretation of datasets from complex systems while allowing the user to decide in several critical points. COSMiCS also includes the use of χ^2, which is a statistical metrics typically used in SAXS analysis for the comparison of the final results of the optimization with the experimental data. A Monte Carlo analysis was implemented to obtain error bars of the decomposed spectra for their further use with other software to derive structural models from experimental data. Despite these changes, COSMiCS conserves the flexibility of the original MCR-ALS to include several constraints along the optimization. In COSMiCS I have adapted these constraints to the specific features of SAXS experiments. During this thesis I have applied COSMiCS to two types of biological systems monitored by SAXS in different manners: amyloid fibrillation using time-dependent data, and SEC-SAXS data.

Amyloid fibrillation processes

We have performed time-dependent measurements of two proteins that forms fibrils: insulin and the familial mutant E46K of αSN. Insulin has a great pharmacological interest (section 2.3.4), and αSN is directly related with Parkinson's disease (section 2.3.3).

From the COSMiCS analysis, I was able to detect three main species in both systems: native state, intermediate oligomer and mature fibrils. It seems reasonable that these complex systems are formed for more than three species, but these are the species that we are able to identify under these experimental conditions. We have to take in account the limitations of the method, which is not able to detect species that are present at very low concentration. Moreover, species that present parallel kinetics are detected by COSMiCS as a single component. Despite these limitations, the description of the system with a reduced number of components provides an overall picture of such complex processes. In the case of insulin fibrillation is possible that three represents a realistic number of species along fibrillation, although we cannot discard small amounts of other oligomers. However, in the case of αSN there are indications of the presence of more species that we are not able to detect with our method. These indications are (i) the impossibility of performing a ab-initio modelling of the oligomer, (ii) a D_{max} similar to that of the final fibril, and (iii) a *p(r)* that suggests that the detected oligomer is polydisperse including species of very distinct size (Figure 4.22).

I was able to derive an overview of the kinetics for both fibrillation processes. Insulin seems to evolve as an on-pathway process (see section 4.1.6) because the population of the native state decreases due the formation of the oligomers, whose concentration also

decreases when that of the fibrils starts to grow. Interestingly, αSN displays a different kinetics (see section 4.2.6) that probably evolves as an off-pathway process [219]. In αSN, our results suggest that the large oligomer disassembles again in smaller species, which can restructure in the nuclei that forms the fibrils. This is indicated by the presence of the native form along the whole fibrillation process, and also the delay observed between the increase of the fibril population and the decay of the oligomer population. The lack of ThT signal during the lag phase while the oligomeric species is growing indicates that protein form does not have an organized β-sheet structure. Interestingly, the same observation was done for insulin. A very important evidence for the accuracy of our COSMiCS decomposition is that ThT signal measured along the process and not included in the COSMiCS analysis, is in very good agreement with the decomposed population of the final fibril (Figures 4.12 and 4.25).

Unfortunately, our structural analysis does not provide insights into the cytotoxicity of the derived species. It is possible that these mixtures formed along the fibrillation contain one or some species that are able to break vesicles or interfere with other biological mechanisms leading to a cytotoxic effect. However, it would be necessary to perform additional experiments [232, 374, 375].

Importantly, the detailed structural and kinetic knowledge derived from COSMICS provides novel insights into these fibrillation processes. Moreover, it opens the possibility to perform and interpret new experiments, such as evaluating inhibitors of the formation of cytotoxic species during early stages of fibrillation processes.

Data collection from fibrillation processes

It is important to note the difficulty of monitoring fibrillation experiments by SAXS. In the experiments presented in chapter 4, they have been performed in different wells in a platereader [275]. The low reproducibility of the fibrillation process yields scattered solutions of the populations. Despite this problem, we have been able to obtain information about the general behaviour of the species along the process. However, the precision of the populations derived is low making impossible to fit the results with a kinetic model. In the future, these problems in the set-up have to be improved, but with certain considerations imposed by the use of SAXS: (i) the fibrillation protocol has to be performed in a time frame compatible with the normally available beamtime in synchrotrons ($\approx$24h); (ii) it has to allow withdrawing samples to be brought to the beam; and (ii) the concentration has to be suitable for SAXS measurements (few mg/mls). Our results demonstrate that it is very important to have a total control of the starting point of the fibrillation as it decreases the stochasticity of the process. One of the attempts that we did for improving the set-up was the use of the spectrometer *ProbeDrum* (see chapter 7), which allows performing the whole fibrillation in a single experiment and, consequently reducing the reproducibility problems observed in the plate-reader. However, this set-up brings other difficulties, like the systematic loss of protein, probably in the form of fibrils, which attaches to the glass of the cell along the experiment. Importantly, this continuous concentration decrease prevents the

use of *closure* constraint in COSMiCS, and has precluded its application for the analysis of data described in section 7.4. The improvement of these limiting factors could provide a more precise description of the population of the different species present along the fibrillation. Under these circumstances, our data could be used to test or validate complex kinetic models that have been proposed to describe the formation of amyloids [99, 118, 193]. It is worth noting that COSMiCS provides a holistic picture of the fibrillation, and therefore its results could be used to have a better understanding of the kinetics of amyloidogenesis. From the chemoetrics point of view, the use of kinetic models along the optimization process would severely constrain the potential solution and increase the accuracy of the results.

COSMiCS as a tool to analyse SEC-SAXS experiments

As we described in chapter 6, COSMiCS allows analyzing data from SEC-SAXS experiments in cases where there is peak overlap. Our results demonstrate that in the vast majority of cases additional information is needed to decompose the data. Again, the definition of appropriate constraints is key to disentangle complex data. Our results demonstrate the need to define local-rank regions to accurately decompose overlapped chromatographic peaks. When defining these local ranks, regardless of the noise level, the derived curves are correct. The definition of the local rank has been attempted by the analysis of the R_g profile and the EFA. According to our results, the combination of both approaches seems the most appropriate strategy to define the peak composition. Our results also show that additional information is required to derive precise populations (chromatographic peaks). The inclusion of absorbance data, if the extinction coefficient of the species involved are known or can be computed, can be used to obtain these populations accurately. This piece of information is very important as it is related with the rotational ambiguity, and with the possibility to estimate the molecular weight of the species involved.

Using SEC-SAXS data, we have studied the conformational fluctuations in prolyl oligopeptidase (POP), a protein related to synaptic functions and neuronal development. POP presents a conformational equilibrium between an open and closed form that is displaced towards the closed, inactive conformation in the presence of an inhibitor.

There are still several tests that we need to perform with data from SEC-SAXS experiments to check the limits of the usage of COSMiCS for their analysis. Some of the tests to be performed would be the design and analyse more complex chromatographic profiles with more peak overlap, more species, and different concentrations. It is worth to test the two methods to derive local ranks in scenarios with species that have similar molecular weight but different R_g and vice versa. All these test will enable to evaluate the power and the limits of COSMiCS in the analysis of SEC-SAXS data, and compare it with US-SOMO, which is the reference program for this kind of analyses.

Further improvements in COSMiCS

From the conclusions of the analysis done for amyloids and SEC-SAXS data, it is clear the necessity of the method for additional information to improve the accuracy of the solutions. We have seen that the simultaneous use of multiple representations and the inclusion of different constraints (e.g. closure, chromatogram composition) improve the derived solutions and decrease the ambiguity of the system. Moreover, we have seen how powerful can be the addition of complementary data to the system in the analysis of SEC-SAXS data using UV-visible absorbance data (see section 6.2.4). We have also explored the use of other techniques introducing fluorescence signal to the dataset using LCO probes to monitor species that are not detected by a classical dyes such as ThT (chapter 7). This later implementation is promising, but it will depend on the optimization of the experimental set-up, as described above. These efforts in combining SAXS with complementary techniques are in accordance with the general tendency in the field to complement SAXS data with other information for the system of interest. This is evident in the new set-ups of the beamlines all over the world that make possible, simultaneously to SAXS, other data from UV-visible or fluorescence spectroscopy, refractive index or light scattering.

Other improvement that is possible to do in COSMiCS is the implementation of non-experimental information, such as kinetic models that can be used for amyloids and folding experiments, as well as thermodynamic models that can be used in titration experiments. The inclusion of kinetic or thermodynamic models will assure the physical validity of the solution obtained by COSMiCS. Moreover, as these models will severely constrain the population profiles, the number of degrees of freedom to be optimized will be reduced and overall the amount of data required to decompose a mixture will decrease.

As I have explained above, we have used constraints such as closure or non-negativity to provide physical meaning to the solutions from COSMiCS. However, there are still situations where COSMiCS finds solutions that are unphysical. The most notorious example found along this thesis is the obtention of non SAXS-like curves. This phenomenon arises because the method analyses the data from a mathematical perspective, and all the experimental points are treated independently. This is different to the usual way to manage SAXS data, where individual intensities are highly correlated, and the density of points describing a profile depends on the detector used. It would be desirable that COSMiCS could identify SAXS-like solutions along the optimization, and add a penalty for solutions that do not fulfill this requirement as an additional constraint. We have done a preliminary analysis (not shown) for different SAXS-like datasets of curves (from fibrillation processes and globular proteins), and non SAXS-like solutions obtained with COSMiCS. We have tried to describe both families of datasets using Fourier series, and this analysis suggests that the number of Fourier series necessary to fit the non SAXS-like curves is larger than these needed for a real curve. We will perform further analyses in that direction to be able to validate this preliminary observation. I believe that this differential behavior between physical and non-physical SAXS curves can be coded into a numerical restraint to drive COSMiCS optimization. This term accounting for the number of Fourier series needed to describe a SAXS curve can be added as a penalty to a general pseudoenergy function that will be minimized along the optimization. This approach would be similar to the

traditional NMR approaches to determine protein structure, which integrate multiple distinct experimental observables. In addition to penalties, these pseudoenergy functions are based on the capacity of the solution to describe the experimental dataset. In that context, the present version of COSMiCS could be improved by using the traditional χ^2 to evaluate the accuracy of intermediate solutions in addition or substitution of the lack of fit. In this manner, our optimization process would account for the experimentally determined errors for each of the intensity values.

We believe that COSMiCS has a broad range of applications in complex biological systems amenable to SAXS and that can be rationally perturbed by modifying the experimental conditions: Transient bimolecular interactions, protein and RNA folding, and the formation of large supramolecular assemblies could be now studied at structural level by overcoming their inherent polydispersity. Coupling this analysis strategy to microfluidics or stop-flow devices could ensure rapid and economic access to hundreds of curves reporting on complex equilibria and time-evolving systems. Moreover, the implementation of additional constrains either experimental (complementary data) or theoretical (kinetic/thermodynamic models and definition of SAXS-like profiles) could make COSMiCS a general tool to decompose complex biological systems in a very accurate manner.

Bibliography

[1] Maxim V. Petoukhov, Isabelle M. L. Billas, Maria Takacs, Melissa A. Graewert, Dino Moras, and Dmitri I. Svergun. Reconstruction of quaternary structure from X-ray scattering by equilibrium mixtures of biological macromolecules. *Biochemistry*, 52(39):6844–55, oct 2013.

[2] Christopher D. Putnam, Michal Hammel, Greg L. Hura, and John A. Tainer. X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Q. Rev. Biophys.*, 40(3):191–285, 2007.

[3] Michel H. J. Koch, Patrice Vachette, and Dmitri I. Svergun. *Small-angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution*, volume 36. may 2003.

[4] P. Debye and A. M. Bueche. Scattering by an inhomogeneous solid. *J. Appl. Phys.*, 20(6):5128–25, 1949.

[5] P. Debye. Zerstreuung von R{ö}ntgenstrahlen. *Ann. Phys.*, 351(6):809–823, jan 1915.

[6] Dmitri I. Svergun, Michel H. J. Koch, Peter A. Timmins, and Roland P. May. *Small Angle X-Ray and Neutron Scattering from Solutions of Biological Macromolecules*. 2013.

[7] Daniel Franke, Cy M. Jeffries, and Dmitri I. Svergun. Correlation Map, a goodness-of-fit test for one-dimensional X-ray scattering spectra. *Nat. Methods*, 12(5):419–422, 2015.

[8] A. Guinier. La Diffraction des Rayons X aux Tres Faibles Angles: Applications a l'Etude des Phenomenes Ultramicroscopiques. *Ann. Phys*, 12:161–236, 1939.

[9] Otto Glatter. Data evaluation in small-angle scattering: calcularion of radial electron-density distribution by means of indirect Fourier transformation. *Acta Phys. Austriaca*, 47:83–102, 1977.

[10] Pau Bernadó. Effect of interdomain dynamics on the structure determination of modular proteins by small-angle scattering. *Eur. Biophys. J.*, 39(5):769–780, apr 2010.

[11] William T. Heller. Influence of multiple well defined conformations on small-angle scattering of proteins in solution. *Acta Crystallogr. D. Biol. Crystallogr.*, 61(Pt 1):33–44, jan 2005.

[12] Dmitri I. Svergun and Michel H. J. Koch. Small-angle scattering studies of biological macromolecules in solution. *Reports Prog. Phys.*, 66(10):1735–1782, oct 2003.

[13] P. Debye, H. R. Anderson, and H. Brumberger. Scattering by an inhomogeneous solid. 2. The correlation function ant its application. *J. Appl. Phys.*, 28:679–83, 1957.

[14] G Porod. Die R{ö}ntgenkleinwinkelstreuung von dichtgepackten kolloiden Systemen - I. Teil. *Kolloid-Zeitschrift*, 124:83–114, 1951.

[15] Petr V. Konarev, Vladimir V. Volkov, Anna V. Sokolova, Michel H. J. Koch, and Dmitri I. Svergun. PRIMUS : a Windows PC-based system for small-angle scattering data analysis. *J. Appl. Crystallogr.*, 36(5):1277–1282, sep 2003.

[16] Robert P. Rambo and John A. Tainer. Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law. *Biopolymers*, 95(8):559–71, aug 2011.

[17] Otto Glatter and O. Kratky. *Small angle X-ray Scattering*. Academic Press Inc., London, 1982.

[18] P. Debye. Molecular-weight determination by light scattering. *J. Phys. Colloid. Chem.*, 51:18–32, 1947.

[19] Sebastian Doniach, J. Bascle, T. Garel, and H. Orland. Partially folded states of proteins: characterization by X-ray scattering. *J. Mol. Biol.*, 254(5):960–7, dec 1995.

[20] Alfred Holtzer. Interpretation of the angular distribution of the light scattered by a polydisperse system of rods. *J. Polym. Sci.*, 17(85):432–434, 1955.

[21] Robert P. Rambo and John A. Tainer. Super-resolution in solution X-ray scattering and its applications to structural systems biology. *Annu. Rev. Biophys.*, 42:415–41, jan 2013.

[22] L. A. Feigin, Dmitri I. Svergun, and G. W. Taylor. *Structure analysis by small-angle X-ray and neutron scattering*. 1987.

[23] Doris Orthaber, Alexander Bergmann, and Otto Glatter. SAXS experiments on absolute scale with Kratky systems using water as a secondary standard. *J. Appl. Crystallogr.*, 33(2):218–225, apr 2000.

[24] H. Fischer, M. de Oliveira Neto, H. B. Napolitano, I. Polikarpov, and A. F. Craievich. Determination of the molecular weight of proteins in solution from a single small-angle X-ray scattering measurement on a relative scale. *J. Appl. Crystallogr.*, 43(1):101–109, feb 2010.

[25] Cy M. Jeffries, Melissa A. Graewert, Dmitri I. Svergun, and Clément E. Blanchet. Limiting radiation damage for high-brilliance biological solution scattering: Practical experience at the EMBL P12 beamline PETRAIII. *J. Synchrotron Radiat.*, 22(2):273–279, 2015.

[26] Lee Makowski. Characterization of proteins with Wide-angle X-ray Solution Scattering (WAXS). *J. Struct. Funct. Genomics*, 11(1):9–19, 2010.

[27] Marc Allaire and Lin Yang. Biomolecular solution X-ray scattering at the national synchrotron light source. *J. Synchrotron Radiat.*, 18(1):41–44, 2011.

[28] Lee Makowski, David Gore, Suneeta Mandava, David D. L. Minh, Sanghyun Park, Diane J. Rodi, and Robert F. Fischetti. X-ray solution scattering studies of the structural diversity intrinsic to protein ensembles. *Biopolymers*, 95(8):531–542, aug 2011.

[29] Suzette A. Pabit, Andrea M. Katz, Igor S. Tolokh, Aleksander Drozdetski, Nathan Baker, Alexey V. Onufriev, and Lois Pollack. Understanding nucleic acid structural changes by comparing wide-angle x-ray scattering (WAXS) experiments to molecular dynamics simulations. *J. Chem. Phys.*, 144(20), 2016.

[30] H. B. Stuhrmann. Ein neues Verfahren zur Bestimmung der Oberflaechenform und der inneren Struktur von geloesten globularen Proteinen aus Roentgenkleinwinkelmessungen. *Zeitschr. Phys. Chem. Neue Folge*, 72:177–198, 1970.

[31] Dmitri I. Svergun, Vladimir V. Volkov, M. B. Kozin, and H. B. Stuhrmann. New Developments in Direct Shape Determination from Small-Angle Scattering. 2. Uniqueness. *Acta Crystallogr. Sect. A Found. Crystallogr.*, 52(3):419–426, may 1996.

[32] P. Chacón, F. Morán, J.F. Díaz, E. Pantos, and J.M. Andreu. Low-Resolution Structures of Proteins in Solution Retrieved from X-Ray Scattering with a Genetic Algorithm. *Biophys. J.*, 74(6):2760–2775, jun 1998.

[33] Dmitri I. Svergun. Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biophys. J.*, 76(6):2879–86, jun 1999.

[34] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by Simulated Annealing. *Science (80-.).*, 220(4598):671–680, may 1983.

[35] Daniel Franke and Dmitri I. Svergun. DAMMIF , a program for rapid ab-initio shape determination in small-angle scattering. *J. Appl. Crystallogr.*, 42(2):342–346, jan 2009.

[36] Maxim V. Petoukhov and Dmitri I. Svergun. Ambiguity assessment of small-angle scattering curves from monodisperse systems. *Acta Crystallogr. Sect. D Biol. Crystallogr.*, 71:1051–1058, 2015.

[37] Vladimir V. Volkov and Dmitri I. Svergun. Uniqueness of ab initio shape determination in small-angle scattering. *J. Appl. Crystallogr.*, 36(3):860–864, apr 2003.

[38] M. B. Kozin and Dmitri I. Svergun. Automated matching of high- and low-resolution structural models. *J. Appl. Crystallogr.*, 34(1):33–41, feb 2001.

[39] Anne T. Tuukkanen, Gerard J. Kleywegt, and Dmitri I. Svergun. Resolution of ab initio shapes determined from small-angle scattering. *IUCrJ*, 3:440–447, 2016.

[40] Maxim V. Petoukhov and Dmitri I. Svergun. Global Rigid Body Modeling of Macromolecular Complexes against Small-Angle Scattering Data. *Biophys. J.*, 89(2):1237–1250, 2005.

[41] Dmitri I. Svergun, C Barberato, and Michel H. J. Koch. CRYSOL-a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. . . .*, pages 768–773, 1995.

[42] Kasper Stovgaard, Christian Andreetta, Jesper Ferkinghoff-Borg, and Thomas Hamelryck. Calculation of accurate small angle X-ray scattering curves from coarse-grained protein models. *BMC Bioinformatics*, 11:429, aug 2010.

[43] Alexander Grishaev, Liang Guo, Thomas C. Irving, and Ad Bax. Improved fitting of solution X-ray scattering data to macromolecular structures and structural ensembles by explicit water modeling. *J. Am. Chem. Soc.*, 132(44):15484–6, nov 2010.

[44] Marcelo Augusto dos Reis, Ricardo Aparicio, and Yang Zhang. Improving Protein Template Recognition by Using Small-Angle X-Ray Scattering Profiles. *Biophys. J.*, 101(11):2770–2781, dec 2011.

[45] Haiguang Liu, Richard J. Morris, Alexander Hexemer, Scott Grandison, and Peter H. Zwart. Computation of small-angle scattering profiles with three-dimensional Zernike polynomials. *Acta Crystallogr. Sect. A Found. Crystallogr.*, 68(2):278–285, mar 2012.

[46] F. Poitevin, H. Orland, Sebastian Doniach, P. Koehl, and M. Delarue. AquaSAXS: a web server for computation and fitting of SAXS profiles with non-uniformally hydrated atomic models. *Nucleic Acids Res.*, 39(suppl):W184–W189, jul 2011.

[47] Dina Schneidman-Duhovny, Michal Hammel, and Andrej Sali. FoXS: a web server for rapid computation and fitting of SAXS profiles. *Nucleic Acids Res.*, 38(Web Server issue):W540–4, jul 2010.

[48] Christopher J. Knight and Jochen S. Hub. WAXSiS: a web server for the calculation of SAXS/WAXS curves based on explicit-solvent molecular dynamics. *Nucleic Acids Res.*, 43(W1):W225–W230, jul 2015.

[49] Daniel K. Putnam, Edward W. Lowe, and Jens Meiler. Reconstruction of Saxs Profiles From Protein Structures. *Comput. Struct. Biotechnol. J.*, 8(11):1–12, 2013.

[50] Maxim V. Petoukhov, Daniel Franke, Alexander V. Shkumatov, Giancarlo Tria, Alexey G. Kikhney, Michal Gajda, Christian Gorba, Haydyn D. T. Mertens, Petr V. Konarev, and Dmitri I. Svergun. New developments in the ATSAS program package for small-angle scattering data analysis. *J. Appl. Crystallogr.*, 45(2):342–350, mar 2012.

[51] Friedrich Förster, Benjamin Webb, Kristin A. Krukenberg, Hiro Tsuruta, David A. Agard, and Andrej Sali. Integration of Small-Angle X-Ray Scattering aata into Structural modeling of proteins and their assemblies. *J. Mol. Biol.*, 382(4):1089–1106, oct 2008.

[52] Guillaume Evrard, Fabien Mareuil, Francois Bontems, Christina Sizun, and Javier Perez. DADIMODO: A program for refining the structure of multidomain proteins and complexes against small-angle scattering data and NMR-derived restraints. *J. Appl. Crystallogr.*, 44(6):1264–1271, dec 2011.

[53] Petr V. Konarev, Maxim V. Petoukhov, Vladimir V. Volkov, and Dmitri I. Svergun. ATSAS 2.1, a program package for small-angle scattering data analysis. *J. Appl. Crystallogr.*, 39(2):277–286, mar 2006.

[54] Pau Bernadó, Yolanda Pérez, Jascha Blobel, Juan Fernández-Recio, Dmitri I. Svergun, and Miquel Pons. Structural characterization of unphosphorylated STAT5a oligomerization equilibrium in solution by small-angle X-ray scattering. *Protein Sci.*, 18(4):716–726, apr 2009.

[55] Hartmut H. Niemann, Maxim V. Petoukhov, Michael Härtlein, Martine Moulin, Ermanno Gherardi, Peter A. Timmins, Dirk W Heinz, and Dmitri I. Svergun. X-ray and neutron small-angle scattering analysis of the complex formed by the Met receptor and the Listeria monocytogenes invasion protein InlB. *J. Mol. Biol.*, 377(2):489–500, mar 2008.

[56] Stefano Paravisi, Gianluca Fumagalli, Milena Riva, Paola Morandi, Rachele Morosi, Petr V. Konarev, Maxim V. Petoukhov, Stéphane Bernier, Robert Chênevert, Dmitri I. Svergun, Bruno Curti, and Maria A. Vanoni. Kinetic and mechanistic characterization of Mycobacterium tuberculosis glutamyl-tRNA synthetase and determination of its oligomeric structure in solution. *FEBS J.*, 276(5):1398–417, mar 2009.

[57] Zhirong Liu and Yongqi Huang. Advantages of proteins being disordered. *Protein Sci.*, 23(5):539–550, may 2014.

[58] Pau Bernadó and Dmitri I. Svergun. Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. *Mol. Biosyst.*, 8(1):151–67, jan 2012.

[59] Veronique Receveur-Brechot and Dominique Durand. How random are intrinsically disordered proteins? A small angle scattering perspective. *Curr. Protein Pept. Sci.*, 13(1):55–75, feb 2012.

[60] Pau Bernadó and Martin Blackledge. Proteins in dynamic equilibrium. *Nature*, 468(7327):1046–1047, dec 2010.

[61] Pau Bernadó, Efstratios Mylonas, Maxim V. Petoukhov, Martin Blackledge, and Dmitri I. Svergun. Structural Characterization of Flexible Proteins Using Small-Angle X-ray Scattering. *J. Am. Chem. Soc.*, 129(17):5656–5664, may 2007.

[62] Martin Pelikan, Greg L. Hura, and Michal Hammel. Structure and flexibility within proteins as identified through small angle X-ray scattering. *Gen. Physiol. Biophys*, 28:174–189, 2009.

[63] Sichun Yang, L. Blachowicz, Lee Makowski, and Benoît Roux. Multidomain assembled states of Hck tyrosine kinase in solution. *Proc. Natl. Acad. Sci.*, 107(36):15757–15762, sep 2010.

[64] Ivano Bertini, Lucio Ferella, Claudio Luchinat, Giacomo Parigi, Maxim V. Petoukhov, Enrico Ravera, Antonio Rosato, and Dmitri I. Svergun. MaxOcc: a web portal for maximum occurrence analysis. *J. Biomol. NMR*, 53(4):271–280, aug 2012.

[65] Bartosz Różycki, Young C Kim, and Gerhard Hummer. SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions. *Structure*, 19(1):109–116, jan 2011.

[66] Stepan Kashtanov and F. Marty Ytreberg. Characterizing Structural Ensembles of Intrinsically Disordered Proteins Determined via a Broad Ensemble Generator with Reweighting. *Biophys. J.*, 102(3):631a, jan 2012.

[67] L. D. Antonov, S. Olsson, W. Boomsma, and T. Hamelryck. Bayesian inference of protein ensembles from SAXS data. *Phys. Chem. Chem. Phys.*, 18(8):5832–8, feb 2016.

[68] Tiago N. Cordeiro, Fátima Herranz-Trillo, Annika Urbanek, Alejandro Estaña, Juan Cortés, Nathalie Sibille, and Pau Bernadó. Small-angle scattering studies of intrinsically disordered proteins and their complexes. *Curr. Opin. Struct. Biol.*, 42:15–23, 2017.

[69] Pau Bernadó, Laurence Blanchard, Peter A. Timmins, Dominique Marion, Rob W .H. Ruigrok, and Martin Blackledge. A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. *Proc. Natl. Acad. Sci. U. S. A.*, 102(47):17002–7, nov 2005.

[70] Yasushi Watanabe and Yoji Inoko. Size-exclusion chromatography combined with small-angle X-ray scattering optics. *J. Chromatogr. A*, 1216(44):7461–7465, 2009.

[71] Kensal E. Van Holde, W. Curtis. Johnson, and Pui Shing Ho. *Principles of physical biochemistry*. Prentice-Hall Inc., Upper Saddle River, NJ, 1998.

[72] Elizabeth Mathew, Ahmed Mirza, and Nick Menhart. Liquid-chromatography-coupled SAXS for accurate sizing of aggregating proteins. *J. Synchrotron Radiat.*, 11(Pt 4):314–8, jul 2004.

[73] G. David and J. Pérez. Combined sampler robot and high-performance liquid chromatography: a fully automated system for biological small-angle X-ray scattering experiments at the Synchrotron SOLEIL SWING beamline. *J. Appl. Crystallogr.*, 42(5):892–900, oct 2009.

[74] Natalie J. Gunn, Michael A. Gorman, Renwick C. J. Dobson, Michael W. Parker, and Terrence D. Mulhern. Purification, crystallization, small-angle X-ray scattering and preliminary X-ray diffraction analysis of the SH2 domain of the Csk-homologous kinase. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.*, 67(3):336–339, mar 2011.

[75] Javier Pérez and Yoshinori Nishino. Advances in X-ray scattering: From solution SAXS to achievements with coherent beams. *Curr. Opin. Struct. Biol.*, 22(5):670–678, 2012.

[76] Emre Brookes, Borries Demeler, and Mattia Rocco. Developments in the US-SOMO bead modeling suite: New features in the direct residue-to-bead method, improved grid routines, and influence of accessible surface area screening. *Macromol. Biosci.*, 10(7):746–753, 2010.

[77] Emre Brookes, Borries Demeler, Camillo Rosano, and Mattia Rocco. The implementation of SOMO (SOlution MOdeller) in the UltraScan analytical ultracentrifugation data analysis suite: Enhanced capabilities allow the reliable hydrodynamic modeling of virtually any kind of biomacromolecule. *Eur. Biophys. J.*, 39(3):423–435, 2010.

[78] Emre Brookes, Javier Pérez, Barbara Cardinali, Aldo Profumo, Patrice Vachette, and Mattia Rocco. Fibrinogen species as resolved by HPLC-SAXS data processing within the UltraScan Solution Modeler (US-SOMO) enhanced SAS module. *J. Appl. Crystallogr.*, 46(Pt 6):1823–1833, dec 2013.

[79] Emre Brookes, Patrice Vachette, Mattia Rocco, and Javier Pérez. US-SOMO HPLC-SAXS module: Dealing with capillary fouling and extraction of pure component patterns from poorly resolved SEC-SAXS data. *J. Appl. Crystallogr.*, 49(5):1827–1841, 2016.

[80] Mar López-Pelegrín, Núria Cerdà-Costa, Anna Cintas-Pedrola, Fátima Herranz-Trillo, Pau Bernadó, Juan R. Peinado, Joan L. Arolas, and F. Xavier Gomis-Rüth. Multiple Stable Conformations Account for Reversible Concentration-Dependent Oligomerization and Autoinhibition of a Metamorphic Metallopeptidase. *Angew. Chemie Int. Ed.*, 53(40):10624–10630, 2014.

[81] David Albesa-Jové, Natalia Comino, Montse Tersa, Elisabeth Mohorko, Saioa Urresti, Elisa Dainese, Laurent R. Chiarelli, Maria Rosalia Pasca, Riccardo Manganelli, Vadim Makarov, Giovanna Riccardi, Dmitri I. Svergun, Rudi Glockshuber, and Marcelo E Guerin. The redox state regulates the conformation of Rv2466c to activate the antitubercular prodrug TP053. *J. Biol. Chem.*, 290(52):31077–31089, dec 2015.

[82] Gary W. Daughdrill, M. S. Chadsey, J. E. Karlinsey, K. T. Hughes, and F. W. Dahlquist. The C-terminal half of the anti-sigma factor, FlgM, becomes structured when bound to its target, sigma 28. *Nat. Struct. Biol.*, 4(4):285–91, apr 1997.

[83] R. W. Kriwacki, L. Hengst, L. Tennant, S. I. Reed, and P. E. Wright. Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity. *Proc. Natl. Acad. Sci. U. S. A.*, 93(21):11504–9, oct 1996.

[84] Satoshi Fukuchi, Kazuo Hosoda, Keiichi Homma, Takashi Gojobori, and Ken Nishikawa. Binary classification of protein molecules into intrinsically disordered and ordered segments. *BMC Struct. Biol.*, 11(1):29, jun 2011.

[85] Jing-Shan Zhao, Kai Zhou, and Zhi-Jing Feng. A theory of degrees of freedom for mechanisms. *Mech. Mach. Theory*, 39(6):621–643, jun 2004.

[86] A. Keith Dunker, Celeste J Brown, J David Lawson, Lilia M Iakoucheva, and Zoran Obradović. Intrinsic disorder and protein function. *Biochemistry*, 41(21):6573–82, may 2002.

[87] H. Jane Dyson and Peter E. Wright. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.*, 6(3):197–208, mar 2005.

[88] Robin van der Lee, Marija Buljan, Benjamin Lang, Robert J. Weatheritt, Gary W. Daughdrill, A. Keith Dunker, Monika Fuxreiter, Julian Gough, Joerg Gsponer, David T. Jones, Philip M. Kim, Richard W. Kriwacki, Christopher J. Oldfield, Rohit V. Pappu, Peter Tompa, Vladimir N. Uversky, Peter E. Wright, and M. Madan Babu. Classification of Intrinsically Disordered Regions and Proteins. *Chem. Rev.*, 114(13):6589–6631, jul 2014.

[89] Vladimir N. Uversky, Christopher J. Oldfield, and A. Keith Dunker. Intrinsically Disordered Proteins in Human Diseases: Introducing the D2 Concept. *Annu. Rev. Biophys.*, 37(1):215–246, jun 2008.

[90] Lilia M Iakoucheva, Celeste J Brown, J David Lawson, Zoran Obradović, and A. Keith Dunker. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.*, 323(3):573–84, oct 2002.

[91] Fabrizio Chiti and Christopher M. Dobson. Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.*, 75:333–66, jan 2006.

[92] Ines Moreno-Gonzalez and Claudio Soto. Misfolded protein aggregates: mechanisms, structures and potential for disease transmission. *Semin. Cell Dev. Biol.*, 22(5):482–7, jul 2011.

[93] Christopher M. Dobson. Protein folding and misfolding. *Nature*, 426(December):884–890, dec 2003.

[94] Jacqueline Burré, Manu Sharma, and Thomas C. Südhof. Alpha-synuclein assembles into higher-order multimers upon membrane binding to promote SNARE complex formation. *Proc. Natl. Acad. Sci. U. S. A.*, 111(40):E4274–83, oct 2014.

[95] Mark R Cookson. Alpha-Synuclein and neuronal cell death. *Mol. Neurodegener.*, 4(1):9, 2009.

[96] Christopher M. Dobson. Protein misfolding, evolution and disease. *Trends Biochem. Sci.*, 24(9):329–32, sep 1999.

[97] J. I. Guijarro, M. Sunde, J. A. Jones, I. D. Campbell, and Christopher M. Dobson. Amyloid fibril formation by an SH3 domain. *Proc. Natl. Acad. Sci. U. S. A.*, 95(8):4224–8, apr 1998.

[98] David S. Eisenberg and Mathias Jucker. The Amyloid State of Proteins in Human Diseases. *Cell*, 148(6):1188–1203, mar 2012.

[99] Tuomas P. J. Knowles, Michele Vendruscolo, and Christopher M. Dobson. The amyloid state and its association with protein misfolding diseases. *Nat. Rev. Mol. Cell Biol.*, 15(6):384–96, 2014.

[100] Christopher A. Ross and Michelle A. Poirier. Protein aggregation and neurodegenerative disease. *Nat. Med.*, 10(7):S10–S17, jul 2004.

[101] Kyle L. Morris and Louise C. Serpell. X-Ray Fibre Diffraction Studies of Amyloid Fibrils. pages 121–135. 2012.

[102] M. Sunde and C. Blake. The structure of amyloid fibrils by electron microscopy and X-ray diffraction. *Adv. Protein Chem.*, 50:123–59, 1997.

[103] M. Fändrich. On the structural definition of amyloid fibrils and other polypeptide aggregates. *Cell. Mol. Life Sci.*, 64(16):2066–2078, aug 2007.

[104] O. Sumner Makin, E. Atkins, P. Sikorski, J. Johansson, and Louise C. Serpell. Molecular basis for amyloid fibril formation and stability. *Proc. Natl. Acad. Sci.*, 102(2):315–320, jan 2005.

[105] A. W. P. Fitzpatrick, G. T. Debelouchina, M. J. Bayro, D. K. Clare, M. A. Caporini, V. S. Bajaj, C. P. Jaroniec, L. Wang, V. Ladizhansky, S. A. Muller, C. E. MacPhee, C. A. Waudby, H. R. Mott, Alfonso De Simone, Tuomas P. J. Knowles, H. R. Saibil, M. Vendruscolo, E. V. Orlova, R. G. Griffin, and Christopher M. Dobson. Atomic structure and hierarchical assembly of a cross-beta amyloid fibril. *Proc. Natl. Acad. Sci.*, 110(14):5468–5473, apr 2013.

[106] Michael R. Sawaya, Shilpa Sambashivan, Rebecca Nelson, Magdalena I. Ivanova, Stuart A. Sievers, Marcin I. Apostol, Michael J. Thompson, Melinda Balbirnie, Jed J. W. Wiltzius, Heather T. McFarlane, Anders Ø. Madsen, Christian Riekel, and David S. Eisenberg. Atomic structures of amyloid cross-beta spines reveal varied steric zippers. *Nature*, 447(7143):453–7, may 2007.

[107] Margaret Sunde, Louise C. Serpell, Mark Bartlam, Paul E. Fraser, Mark B. Pepys, and Colin C.F. Blake. Common core structure of amyloid fibrils by synchrotron X-ray diffraction. *J. Mol. Biol.*, 273(3):729–739, oct 1997.

[108] Byron Caughey and Peter T. Lansbury. Protofibrils, pores, fibrils and neurodegeneration: Separating the Responsible Protein Aggregates from The Innocent Bystanders. *Annu. Rev. Neurosci.*, 26(1):267–298, mar 2003.

[109] G. Bitan, M. D. Kirkitadze, A. Lomakin, S. S. Vollers, G. B. Benedek, and D. B. Teplow. Amyloid -protein (Abeta) assembly: Abeta 40 and Abeta 42 oligomerize through distinct pathways. *Proc. Natl. Acad. Sci.*, 100(1):330–335, jan 2003.

[110] Tong Guo, Wendy Noble, and Diane P. Hanger. Roles of tau protein in health and disease. *Acta Neuropathol.*, 133(5):665–704, may 2017.

[111] M. H. Polymeropoulos, C. Lavedan, E. Leroy, S. E. Ide, A. Dehejia, A. Dutra, B. Pike, H. Root, J. Rubenstein, R. Boyer, E. S. Stenroos, S. Chandrasekharappa, A. Athanassiadou, T. Papapetropoulos, W. G. Johnson, A. M. Lazzarini, R. C. Duvoisin, G. Di Iorio, L. I. Golbe, and R. L. Nussbaum. Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. *Science*, 276(5321):2045–7, jun 1997.

[112] Maria Grazia Spillantini, Marie Luise Schmidt, Virginia M.-Y. Lee, John Q. Trojanowski, Ross Jakes, and Michel Goedert. Alpha-synuclein in Lewy bodies. *Nature*, 388(6645):839–840, aug 1997.

[113] J. T. Jarrett and Peter T. Lansbury. Amyloid fibril formation requires a chemically discriminating nucleation event: studies of an amyloidogenic sequence from the bacterial protein OsmB. *Biochemistry*, 31(49):12345–52, dec 1992.

[114] T. R. Serio, A. G. Cashikar, A. S. Kowal, G. J. Sawicki, J. J. Moslehi, L. Serpell, M. F. Arnsdorf, and S. L. Lindquist. Nucleated conformational conversion and the replication of conformational information by a prion determinant. *Science*, 289(5483):1317–21, aug 2000.

[115] P. R. ten Wolde and D. Frenkel. Enhancement of protein crystal nucleation by critical density fluctuations. *Science*, 277(5334):1975–8, sep 1997.

[116] Minna Groenning. Binding mode of Thioflavin T and other molecular probes in the context of amyloid fibrils—current status. *J. Chem. Biol.*, 3(1):1–18, mar 2010.

[117] Hongxia Zhao, Esa K. J. Tuominen, and Paavo K. J. Kinnunen. Formation of Amyloid Fibers Triggered by Phosphatidylserine-Containing Membranes. *Biochemistry*, 43(32):10302–10307, aug 2004.

[118] Samuel I. A. Cohen, Michele Vendruscolo, Mark E. Welland, Christopher M. Dobson, Eugene M. Terentjev, and Tuomas P. J. Knowles. Nucleated polymerization with secondary pathways. I. Time evolution of the principal moments. *J. Chem. Phys.*, 135(6):065105, aug 2011.

[119] Tuomas P. J. Knowles, C. A. Waudby, G. L. Devlin, Samuel I. A. Cohen, A. Aguzzi, M. Vendruscolo, E. M. Terentjev, Mark E. Welland, and Christopher M. Dobson. An

Analytical Solution to the Kinetics of Breakable Filament Assembly. *Science (80-.).*, 326(5959):1533–1537, dec 2009.

[120] Samuel I. A. Cohen, Sara Linse, Leila M. Luheshi, Erik Hellstrand, Duncan A. White, Luke Rajah, Daniel Erik Otzen, Michele Vendruscolo, Christopher M. Dobson, and Tuomas P. J. Knowles. Proliferation of amyloid-beta42 aggregates occurs through a secondary nucleation mechanism. *Proc. Natl. Acad. Sci. U. S. A.*, 110(24):9758–9763, jun 2013.

[121] Christian Haass and Dennis J. Selkoe. Soluble protein oligomers in neurodegeneration: lessons from the Alzheimer's amyloid β-peptide. *Nat. Rev. Mol. Cell Biol.*, 8(2):101–112, feb 2007.

[122] Eric Karran, Marc Mercken, and Bart De Strooper. The amyloid cascade hypothesis for Alzheimer's disease: an appraisal for the development of therapeutics. *Nat. Rev. Drug Discov.*, 10(9):698–712, aug 2011.

[123] L. F. Lue, Y. M. Kuo, A. E. Roher, L. Brachova, Y. Shen, L. Sue, T. Beach, J. H. Kurth, R. E. Rydel, and J. Rogers. Soluble amyloid beta peptide concentration as a predictor of synaptic change in Alzheimer's disease. *Am. J. Pathol.*, 155(3):853–62, sep 1999.

[124] Massimo Stefani and Christopher M. Dobson. Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution. *J. Mol. Med.*, 81(11):678–699, nov 2003.

[125] S. Baglioni, F. Casamenti, Monica Bucciantini, Leila M. Luheshi, N. Taddei, Fabrizio Chiti, Christopher M. Dobson, and M. Stefani. Prefibrillar Amyloid Aggregates Could Be Generic Toxins in Higher Organisms. *J. Neurosci.*, 26(31):8160–8167, aug 2006.

[126] Lauren M. Billings, Salvatore Oddo, Kim N. Green, James L. McGaugh, and Frank M. LaFerla. Intraneuronal Aβ Causes the Onset of Early Alzheimer's Disease-Related Cognitive Deficits in Transgenic Mice. *Neuron*, 45(5):675–688, mar 2005.

[127] Monica Bucciantini, Elisa Giannoni, Fabrizio Chiti, Fabiana Baroni, Lucia Formigli, Jesús Zurdo, Niccolò Taddei, Giampietro Ramponi, Christopher M. Dobson, and Massimo Stefani. Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature*, 416(6880):507–511, apr 2002.

[128] Silvia Campioni, Benedetta Mannini, Mariagioia Zampagni, Anna Pensalfini, Claudia Parrini, Elisa Evangelisti, Annalisa Relini, Massimo Stefani, Christopher M. Dobson, Cristina Cecchi, and Fabrizio Chiti. A causative link between the structure of aberrant protein oligomers and their toxicity. *Nat. Chem. Biol.*, 6(2):140–147, feb 2010.

[129] Nunilo Cremades, Samuel I. A. Cohen, Emma Deas, Andrey Y. Abramov, Allen Y. Chen, Angel Orte, Massimo Sandal, Richard W. Clarke, Paul Dunne, Francesco A.

Aprile, Carlos W. Bertoncini, Nicholas W. Wood, Tuomas P. J. Knowles, Christopher M. Dobson, and David Klenerman. Direct observation of the interconversion of normal and toxic forms of alpha-synuclein. *Cell*, 149(5):1048–59, may 2012.

[130] R. M. Koffie, M. Meyer-Luehmann, T. Hashimoto, K. W. Adams, M. L. Mielke, M. Garcia-Alloza, K. D. Micheva, S. J. Smith, M. L. Kim, Virginia M.-Y. Lee, B. T. Hyman, and T. L. Spires-Jones. Oligomeric amyloid beta associates with postsynaptic densities and correlates with excitatory synapse loss near senile plaques. *Proc. Natl. Acad. Sci.*, 106(10):4012–4017, mar 2009.

[131] Sylvain Lesné, Ming Teng Koh, Linda Kotilinek, Rakez Kayed, Charles G. Glabe, Austin Yang, Michela Gallagher, and Karen H. Ashe. A specific amyloid-β protein assembly in the brain impairs memory. *Nature*, 440(7082):352–357, mar 2006.

[132] Dominic M. Walsh, Igor Klyubin, Julia V. Fadeeva, William K. Cullen, Roger Anwyl, Michael S. Wolfe, Michael J. Rowan, and Dennis J. Selkoe. Naturally secreted oligomers of amyloid β protein potently inhibit hippocampal long-term potentiation in vivo. *Nature*, 416(6880):535–539, apr 2002.

[133] Beate Winner, Roberto Jappelli, Samir K. Maji, Paula A. Desplats, Leah Boyer, Stefan Aigner, Claudia Hetzer, Thomas Loher, Marçal Vilar, Silvia Campioni, Christos Tzitzilonis, Alice Soragni, Sebastian Jessberger, Helena Mira, Antonella Consiglio, Emiley Pham, Eliezer Masliah, Fred H. Gage, and Roland Riek. In vivo demonstration that alpha-synuclein oligomers are toxic. *Proc. Natl. Acad. Sci. U. S. A.*, 108(10):4194–9, mar 2011.

[134] Kelvin C. Luk, Victoria Kehm, Jenna Carroll, Bin Zhang, Patrick O'Brien, John Q. Trojanowski, and Virginia M. Lee. Pathological alpha-synuclein transmission initiates Parkinson-like neurodegeneration in nontransgenic mice. *Science*, 338(6109):949–53, nov 2012.

[135] W. Peelaerts, L. Bousset, A. Van der Perren, A. Moskalyuk, R. Pulizzi, M. Giugliano, C. Van den Haute, R. Melki, and V. Baekelandt. alpha-Synuclein strains cause distinct synucleinopathies after local and systemic administration. *Nature*, 522(7556):340–4, jun 2015.

[136] Stanley B. Prusiner, Amanda L. Woerman, Daniel A. Mordes, Joel C. Watts, Ryan Rampersaud, David B. Berry, Smita Patel, Abby Oehler, Jennifer K. Lowe, Stephanie N. Kravitz, Daniel H. Geschwind, David V. Glidden, Glenda M. Halliday, Lefkos T. Middleton, Steve M. Gentleman, Lea T. Grinberg, and Kurt Giles. Evidence for alpha-synuclein prions causing multiple system atrophy in humans with parkinsonism. *Proc. Natl. Acad. Sci.*, 112(38):E5308–E5317, sep 2015.

[137] Amanda L. Woerman, Jan Stöhr, Atsushi Aoyagi, Ryan Rampersaud, Zuzana Krejciova, Joel C. Watts, Takao Ohyama, Smita Patel, Kartika Widjaja, Abby Oehler, David W. Sanders, Marc I. Diamond, William W. Seeley, Lefkos T. Middleton,

Steve M. Gentleman, Daniel A. Mordes, Thomas C. Südhof, Kurt Giles, and Stanley B. Prusiner. Propagation of prions causing synucleinopathies in cultured cells. *Proc. Natl. Acad. Sci.*, 112(35):E4949–E4958, sep 2015.

[138] Benedetta Bolognesi, Janet R. Kumita, Teresa P. Barros, Elin K. Esbjorner, Leila M. Luheshi, Damian C. Crowther, Mark R. Wilson, Christopher M. Dobson, Giorgio Favrin, and Justin J. Yerbury. ANS Binding Reveals Common Features of Cytotoxic Amyloid Species. *ACS Chem. Biol.*, 5(8):735–740, aug 2010.

[139] Mookyung Cheon, Iksoo Chang, Sandipan Mohanty, Leila M. Luheshi, Christopher M. Dobson, Michele Vendruscolo, and Giorgio Favrin. Structural reorganisation and potential toxicity of oligomeric species formed during the assembly of amyloid fibrils. *PLoS Comput. Biol.*, 3(9):1727–38, sep 2007.

[140] Priyanka Narayan, Kristina A. Ganzinger, James McColl, Laura Weimann, Sarah Meehan, Seema Qamar, John A. Carver, Mark R. Wilson, Peter St. George-Hyslop, Christopher M. Dobson, and David Klenerman. Single Molecule Characterization of the Interactions between Amyloid-β Peptides and the Membranes of Hippocampal Cells. *J. Am. Chem. Soc.*, 135(4):1491–1498, jan 2013.

[141] Heidi Olzscha, Sonya M. Schermann, Andreas C. Woerner, Stefan Pinkert, Michael H. Hecht, Gian G. Tartaglia, Michele Vendruscolo, Manajit Hayer-Hartl, F. Ulrich Hartl, and R. Martin Vabulas. Amyloid-like aggregates sequester numerous metastable proteins with essential cellular functions. *Cell*, 144(1):67–78, jan 2011.

[142] Hao Ding, Pamela T. Wong, Edgar L. Lee, Ari Gafni, and Duncan G. Steel. Determination of the Oligomer Size of Amyloidogenic Protein β-Amyloid(1–40) by Single-Molecule Spectroscopy. *Biophys. J.*, 97(3):912–921, aug 2009.

[143] Mathew H. Horrocks, Steven F. Lee, Sonia Gandhi, Nadia K. Magdalinou, Serene W. Chen, Michael J. Devine, Laura Tosatto, Magnus Kjaergaard, Joseph S. Beckwith, Henrik Zetterberg, Marija Iljina, Nunilo Cremades, Christopher M. Dobson, Nicholas W. Wood, and David Klenerman. Single-Molecule Imaging of Individual Amyloid Protein Aggregates in Human Biofluids. *ACS Chem. Neurosci.*, 7(3):399–406, mar 2016.

[144] Marcus Kostka, Tobias Högen, Karin M. Danzer, Johannes Levin, Matthias Habeck, Andreas Wirth, Richard Wagner, Charles G. Glabe, Sabine Finger, Udo Heinzelmann, Patrick Garidel, Wenzhen Duan, Christopher A. Ross, Hans Kretzschmar, and Armin Giese. Single particle characterization of iron-induced pore-forming alpha-synuclein oligomers. *J. Biol. Chem.*, 283(16):10992–1003, apr 2008.

[145] Angel Orte, Neil R. Birkett, Richard W. Clarke, Glyn L. Devlin, Christopher M. Dobson, and David Klenerman. Direct characterization of amyloidogenic oligomers by single-molecule fluorescence. *Proc. Natl. Acad. Sci. U. S. A.*, 105(38):14424–9, sep 2008.

[146] M. Pitschke, R. Prior, M. Haupt, and D. Riesner. Detection of single amyloid beta-protein aggregates in the cerebrospinal fluid of Alzheimer's patients by fluorescence correlation spectroscopy. *Nat. Med.*, 4(7):832–4, jul 1998.

[147] A. Iwai, E. Masliah, M. Yoshimoto, N. Ge, L. Flanagan, H. A. de Silva, A. Kittel, and T. Saitoh. The precursor protein of non-A beta component of Alzheimer's disease amyloid is a presynaptic protein of the central nervous system. *Neuron*, 14(2):467–75, feb 1995.

[148] L. Maroteaux, J. T. Campanelli, and R. H. Scheller. Synuclein: a neuron-specific protein localized to the nucleus and presynaptic nerve terminal. *J. Neurosci.*, 8(8):2804–15, aug 1988.

[149] M. Bisaglia, S. Mammi, and L. Bubacco. Structural insights on physiological functions and pathological effects of alpha-synuclein. *FASEB J.*, 23(2):329–340, sep 2008.

[150] Leonid Breydo, Jessica W. Wu, and Vladimir N. Uversky. Alpha-Synuclein misfolding and Parkinson's disease. *Biochim. Biophys. Acta - Mol. Basis Dis.*, 1822(2):261–285, 2012.

[151] Hilal A. Lashuel, Cassia R. Overk, Abid Oueslati, and Eliezer Masliah. The many faces of α-synuclein: from structure and toxicity to therapeutic target. *Nat. Rev. Neurosci.*, 14(1):38–48, dec 2013.

[152] Suzanne Lesage, Mathieu Anheim, Franck Letournel, Luc Bousset, Aurélie Honoré, Nelly Rozas, Laura Pieri, Karine Madiona, Alexandra Dürr, Ronald Melki, Christophe Verny, Alexis Brice, and French Parkinson's Disease Genetics Study Group. G51D α-synuclein mutation causes a novel Parkinsonian-pyramidal syndrome. *Ann. Neurol.*, 73(4):459–471, apr 2013.

[153] Yonaton Zarbiv, Dganit Simhi-Haham, Eitan Israeli, Suaad Abed Elhadi, Jessica Grigoletto, and Ronit Sharon. Lysine residues at the first and second KTKEGV repeats mediate α-Synuclein binding to membrane phospholipids. *Neurobiol. Dis.*, 70:90–98, 2014.

[154] M. Hashimoto, T. Takenouchi, M. Mallory, Eliezer Masliah, and A. Takeda. The role of NAC in amyloidogenesis in Alzheimer's disease. *Am. J. Pathol.*, 156(2):734–6, feb 2000.

[155] Andres Binolfi, Francois-Xavier Theillet, and Philipp Selenko. Bacterial in-cell NMR of human α-synuclein: a disordered monomer by nature? *Biochem. Soc. Trans.*, 40(5):950–954, oct 2012.

[156] Jacqueline Burré, Sandro Vivona, Jiajie Diao, Manu Sharma, Axel T. Brunger, and Thomas C. Südhof. Properties of native brain alpha-synuclein. *Nature*, 498(7453):E4–E6, jun 2013.

[157] Deborah E. Cabin, Kazuhiro Shimazu, Diane Murphy, Nelson B. Cole, Wolfram Gottschalk, Kellie L. McIlwain, Bonnie Orrison, Amy Chen, Christopher E. Ellis, Richard Paylor, Bai Lu, and R. L. Nussbaum. Synaptic vesicle depletion correlates with attenuated synaptic responses to prolonged repetitive stimulation in mice lacking alpha-synuclein. *J. Neurosci.*, 22(20):8797–807, oct 2002.

[158] Bruno Fauvet, M. K. Mbefo, M.-B. Fares, C. Desobry, S. Michael, M. T. Ardah, E. Tsika, P. Coune, M. Prudent, N. Lion, David Eliezer, D. J. Moore, B. Schneider, P. Aebischer, Omar M. A. El-Agnaf, Eliezer Masliah, and Hilal A. Lashuel. Alpha-Synuclein in Central Nervous System and from Erythrocytes, Mammalian Cells, and Escherichia coli Exists Predominantly as Disordered Monomer. *J. Biol. Chem.*, 287(19):15345–15364, may 2012.

[159] David A. Scott, Iustin Tabarean, Yong Tang, Anna Cartier, Eliezer Masliah, and Subhojit Roy. A pathologic cascade leading to synaptic dysfunction in alpha-synuclein-induced neurodegeneration. *J. Neurosci.*, 30(24):8083–95, jun 2010.

[160] Pau Bernadó, Carlos W. Bertoncini, Christian Griesinger, Markus Zweckstetter, and Martin Blackledge. Defining long-range order and local disorder in native alpha-synuclein using residual dipolar couplings. *J. Am. Chem. Soc.*, 127(51):17968–9, dec 2005.

[161] Carlos W. Bertoncini, Young-Sang Jung, Claudio O. Fernandez, Wolfgang Hoyer, Christian Griesinger, Thomas M. Jovin, and Markus Zweckstetter. Release of long-range tertiary interactions potentiates aggregation of natively unstructured alpha-synuclein. *Proc. Natl. Acad. Sci. U. S. A.*, 102(5):1430–5, feb 2005.

[162] Matthew M. Dedmon, Kresten Lindorff-Larsen, John Christodoulou, Michele Vendruscolo, and Christopher M. Dobson. Mapping Long-Range Interactions in α-Synuclein using Spin-Label NMR and Ensemble Molecular Dynamics Simulations. *J. Am. Chem. Soc.*, 127(2):476–477, jan 2005.

[163] J. C. Lee, R. Langen, P. A. Hummel, H. B. Gray, and J. R. Winkler. Alpha-Synuclein structures from fluorescence energy-transfer kinetics: Implications for the role of the protein in Parkinson's disease. *Proc. Natl. Acad. Sci.*, 101(47):16466–16471, nov 2004.

[164] Christofer Lendel, Carlos W. Bertoncini, Nunilo Cremades, Christopher A. Waudby, Michele Vendruscolo, Christopher M. Dobson, Dale Schenk, John Christodoulou, and Gergely Toth. On the mechanism of nonspecific inhibitors of protein aggregation: dissecting the interactions of alpha-synuclein with Congo red and lacmoid. *Biochemistry*, 48(35):8322–34, sep 2009.

[165] Vladimir N. Uversky. A protein-chameleon: conformational plasticity of alpha-synuclein, a disordered protein involved in neurodegenerative disorders. *J. Biomol. Struct. Dyn.*, 21(2):211–34, oct 2003.

[166] Asa Abeliovich, Yvonne Schmitz, Isabel Fariñas, Derek Choi-Lundberg, Wei-Hsien Ho, Pablo E. Castillo, Natasha Shinsky, Jose Manuel Garcia Verdugo, Mark Armanini, Anne Ryan, Mary Hynes, Heidi Phillips, David Sulzer, and Arnon Rosenthal. Mice Lacking α-Synuclein Display Functional Deficits in the Nigrostriatal Dopamine System. *Neuron*, 25(1):239–252, 2000.

[167] Venu M. Nemani, Wei Lu, Victoria Berge, Ken Nakamura, Bibiana Onoa, Michael K. Lee, Farrukh A. Chaudhry, Roger A. Nicoll, and Robert H. Edwards. Increased Expression of α-Synuclein Reduces Neurotransmitter Release by Inhibiting Synaptic Vesicle Reclustering after Endocytosis. *Neuron*, 65(1):66–79, jan 2010.

[168] L. Wang, Utpal Das, David A. Scott, Yong Tang, Pamela J. McLean, and Subhojit Roy. Alpha-Synuclein Multimers Cluster Synaptic Vesicles and Attenuate Recycling. *Curr. Biol.*, 24(19):2319–2326, oct 2014.

[169] Leonidas Stefanis. Alpha-synuclein in Parkinson's disease. *Cold Spring Harb. Perspect. Med.*, 2(2):a009399, feb 2012.

[170] Giuliana Fusco, Alfonso De Simone, Tata Gopinath, Vitaly Vostrikov, Michele Vendruscolo, Christopher M. Dobson, and Gianluigi Veglia. Direct observation of the three regions in α-synuclein that determine its membrane-bound behaviour. *Nat. Commun.*, 5:3827, may 2014.

[171] Myriam M. Ouberai, Juan Wang, Marcus J. Swann, Celine Galvagnion, Tim Guilliams, Christopher M. Dobson, and Mark E. Welland. α-Synuclein senses lipid packing defects and induces lateral expansion of lipids leading to membrane remodeling. *J. Biol. Chem.*, 288(29):20883–95, jul 2013.

[172] Jacqueline Burré, Manu Sharma, T. Tsetsenis, V. Buchman, M. R. Etherton, and T. C. Sudhof. Alpha-Synuclein Promotes SNARE-Complex Assembly in Vivo and in Vitro. *Science (80-.).*, 329(5999):1663–1667, sep 2010.

[173] Jiajie Diao, Jacqueline Burré, Sandro Vivona, Daniel J. Cipriano, Manu Sharma, Minjoung Kyoung, Thomas C. Südhof, and Axel T. Brunger. Native alpha-synuclein induces clustering of synaptic-vesicle mimics via binding to phospholipids and synaptobrevin-2/VAMP2. *Elife*, 2:e00592, apr 2013.

[174] Ying Lai, Sunae Kim, Jobin Varkey, Xiaochu Lou, Jae-Kyun Song, Jiajie Diao, Ralf Langen, and Yeon-Kyun Shin. Nonaggregated alpha-Synuclein Influences SNARE-Dependent Vesicle Docking via Membrane Binding. *Proc. Natl. Acad. Sci. U. S. A*, 53:3889–3896, 2014.

[175] Nancy M. Bonini and Benoit I. Giasson. Snaring the Function of α-Synuclein. *Cell*, 123(3):359–361, nov 2005.

[176] Heather McCann, Claire H. Stevens, Heidi Cartwright, and Glenda M. Halliday. α-Synucleinopathy phenotypes. *Parkinsonism Relat. Disord.*, 20:S62–S67, jan 2014.

[177] Christian Hansen and Jia-Yi Li. Beyond α-synuclein transfer: pathology propagation in Parkinson's disease. *Trends Mol. Med.*, 18(5):248–255, may 2012.

[178] Mehmet Ozansoy and A. Nazli Başak. The Central Theme of Parkinson's Disease: α-Synuclein. *Mol. Neurobiol.*, 47(2):460–465, apr 2013.

[179] M. Baba, S. Nakajo, P. H. Tu, T. Tomita, K. Nakaya, V. M. Lee, John Q. Trojanowski, and T. Iwatsubo. Aggregation of alpha-synuclein in Lewy bodies of sporadic Parkinson's disease and dementia with Lewy bodies. *Am. J. Pathol.*, 152(4):879–84, apr 1998.

[180] M. G. Spillantini, R. A. Crowther, R. Jakes, M. Hasegawa, and M. Goedert. Alpha-Synuclein in filamentous inclusions of Lewy bodies from Parkinson's disease and dementia with lewy bodies. *Proc. Natl. Acad. Sci. U. S. A.*, 95(11):6469–73, may 1998.

[181] M. I. Papp, J. E. Kahn, and P. L. Lantos. Glial cytoplasmic inclusions in the CNS of patients with multiple system atrophy (striatonigral degeneration, olivopontocerebellar atrophy and Shy-Drager syndrome). *J. Neurol. Sci.*, 94(1-3):79–100, dec 1989.

[182] Louise C. Serpell, M. Sunde, Paul E. Fraser, P. K. Luther, E. P. Morris, O. Sangren, E. Lundgren, and C. Blake. Examination of the structure of the transthyretin amyloid fibril by image reconstruction from electron micrographs. *J. Mol. Biol.*, 254(2):113–118, nov 1995.

[183] K. Uéda, H. Fukushima, E. Masliah, Y. Xia, A. Iwai, M. Yoshimoto, D. A. Otero, J. Kondo, Y. Ihara, and T. Saitoh. Molecular cloning of cDNA encoding an unrecognized component of amyloid in Alzheimer disease. *Proc. Natl. Acad. Sci. U. S. A.*, 90(23):11282–6, dec 1993.

[184] Silke Appel-Cresswell, Carles Vilarino-Guell, Mary Encarnacion, Holly Sherman, Irene Yu, Brinda Shah, David Weir, Christina Thompson, Chelsea Szu-Tu, Joanne Trinh, Jan O. Aasly, Alex Rajput, Ali H. Rajput, A. Jon Stoessl, and Matthew J. Farrer. Alpha-synuclein p.H50Q, a novel pathogenic mutation for Parkinson's disease. *Mov. Disord.*, 28(6):811–813, jun 2013.

[185] Marie-Christine Chartier-Harlin, Jennifer Kachergus, Christophe Roumier, Vincent Mouroux, Xavier Douay, Sarah Lincoln, Clotilde Levecque, Lydie Larvor, Joris Andrieux, Mary Hulihan, Nawal Waucquier, Luc Defebvre, Philippe Amouyel, Matthew J. Farrer, and Alain Destée. Alpha-synuclein locus duplication as a cause of familial Parkinson's disease. *Lancet*, 364(9440):1167–1169, sep 2004.

[186] A. B. Singleton, M. Farrer, J. Johnson, A. B. Singleton, S. Hague, J. Kachergus, M. Hulihan, T. Peuralinna, A. Dutra, R. L. Nussbaum, S. Lincoln, A. Crawley, M. Hanson, D. Maraganore, C. Adler, M. R. Cookson, M. Muenter, M. Baptista, D. Miller, J. Blancato, J. Hardy, and K. Gwinn-Hardy. Alpha-Synuclein Locus Triplication Causes Parkinson's Disease. *Science (80-.).*, 302(5646):841–841, oct 2003.

[187] Patrik Brundin, Ronald Melki, and Ron Kopito. Prion-like transmission of protein aggregates in neurodegenerative diseases. *Nat. Rev. Mol. Cell Biol.*, 11(4):301–307, apr 2010.

[188] Karin M. Danzer, Simon K. Krebs, Michael Wolff, Gerald Birk, and Bastian Hengerer. Seeding induced by α-synuclein oligomers provides evidence for spreading of α-synuclein pathology. *J. Neurochem.*, 111(1):192–203, oct 2009.

[189] Mathias Jucker and Lary C. Walker. Self-propagation of pathogenic protein aggregates in neurodegenerative diseases. *Nature*, 501(7465):45–51, sep 2013.

[190] K. M. Danzer, W. P. Ruf, P. Putcha, D. Joyner, T. Hashimoto, C. Glabe, B. T. Hyman, and P. J. McLean. Heat-shock protein 70 modulates toxic extracellular -synuclein oligomers and rescues trans-synaptic toxicity. *FASEB J.*, 25(1):326–336, jan 2011.

[191] P. Desplats, H. J. Lee, E.-J. Bae, C. Patrick, E. Rockenstein, L. Crews, B. Spencer, E. Masliah, and S. J. Lee. Inclusion formation and neuronal cell death through neuron-to-neuron transmission of alpha-synuclein. *Proc. Natl. Acad. Sci.*, 106(31):13010–13015, aug 2009.

[192] A. K. Buell, C. Galvagnion, R. Gaspar, E. Sparr, M. Vendruscolo, Tuomas P. J. Knowles, Sara Linse, and Christopher M. Dobson. Solution conditions determine the relative importance of nucleation and growth processes in alpha-synuclein aggregation. *Proc. Natl. Acad. Sci.*, 111(21):7671–7676, may 2014.

[193] Samuel I. A. Cohen, Michele Vendruscolo, Christopher M. Dobson, and Tuomas P. J. Knowles. Nucleated polymerization with secondary pathways. II. Determination of self-consistent solutions to growth processes described by non-linear master equations. *J. Chem. Phys.*, 135(6):065106, aug 2011.

[194] R. A. Crowther, S. E. Daniel, and M. Goedert. Characterisation of isolated alpha-synuclein filaments from substantia nigra of Parkinson's disease brain. *Neurosci. Lett.*, 292(2):128–30, oct 2000.

[195] Louise C. Serpell, J. Berriman, R. Jakes, M. Goedert, and R. A. Crowther. Fiber diffraction of synthetic alpha-synuclein filaments shows amyloid-like cross-beta conformation. *Proc. Natl. Acad. Sci. U. S. A.*, 97(9):4897–902, apr 2000.

[196] M. Vilar, H. T. Chou, T. Luhrs, S. K. Maji, D. Riek-Loher, R. Verel, G. Manning, H. Stahlberg, and R. Riek. The fold of alpha-synuclein fibrils. *Proc. Natl. Acad. Sci.*, 105(25):8637–8642, jun 2008.

[197] Jose A. Rodriguez, Magdalena I. Ivanova, Michael R. Sawaya, Duilio Cascio, Francis E. Reyes, Dan Shi, Smriti Sangwan, Elizabeth L. Guenther, Lisa M. Johnson, Meng Zhang, Lin Jiang, Mark A. Arbing, Brent L. Nannenga, Johan Hattne, Julian Whitelegge, Aaron S. Brewster, Marc Messerschmidt, Sébastien Boutet, Nicholas K. Sauter, Tamir Gonen, and David S. Eisenberg. Structure of the toxic core of α-synuclein from invisible crystals. *Nature*, 525(7570):486–490, sep 2015.

[198] Francois-Xavier Theillet, Andres Binolfi, Beata Bekei, Andrea Martorana, Honor May Rose, Marchel Stuiver, Silvia Verzini, Dorothea Lorenz, Marleen van Rossum, Daniella Goldfarb, and Philipp Selenko. Structural disorder of monomeric α-synuclein persists in mammalian cells. *Nature*, 530(7588):45–50, jan 2016.

[199] Christopher A. Waudby, Carlo Camilloni, Anthony W. P. Fitzpatrick, Lisa D. Cabrita, Christopher M. Dobson, Michele Vendruscolo, and John Christodoulou. In-Cell NMR Characterization of the Secondary Structure Populations of a Disordered Conformation of α-Synuclein within E. coli Cells. *PLoS One*, 8(8):e72286, aug 2013.

[200] M. Chen, M. Margittai, J. Chen, and R. Langen. Investigation of alpha-synuclein fibril structure by site-directed spin labeling. *J. Biol. Chem.*, 282(34):24970–24979, aug 2007.

[201] Gemma Comellas, Luisel R. Lemkau, Andrew J. Nieuwkoop, Kathryn D. Kloepper, Daniel T. Ladror, Reika Ebisu, Wendy S. Woods, Andrew S. Lipton, Julia M. George, and Chad M. Rienstra. Structured Regions of alpha-synuclein Fibrils Include the Early-Onset Parkinson's Disease Mutation Sites. *J. Mol. Biol.*, 411(4):881–895, aug 2011.

[202] Ani Der-Sarkissian, Christine C. Jao, Jeannie Chen, and Ralf Langen. Structural organization of alpha-synuclein fibrils studied by site-directed spin labeling. *J. Biol. Chem.*, 278(39):37530–5, sep 2003.

[203] Julia Gath, Birgit Habenstein, Luc Bousset, Ronald Melki, Beat H. Meier, and Anja Böckmann. Solid-state NMR sequential assignments of α-synuclein. *Biomol. NMR Assign.*, 6(1):51–55, apr 2012.

[204] Hirotomo Miake, Hidehiro Mizusawa, Takeshi Iwatsubo, and Masato Hasegawa. Biochemical characterization of the core structure of alpha-synuclein filaments. *J. Biol. Chem.*, 277(21):19213–9, may 2002.

[205] Santiago Esteban-Martín, Jordi Silvestre-Ryan, Carlos W. Bertoncini, and Xavier Salvatella. Identification of Fibril-Like Tertiary Contacts in Soluble Monomeric α-Synuclein. *Biophys. J.*, 105(5):1192–1198, sep 2013.

[206] Sigurður Aegir Jónsson, Sandipan Mohanty, and Anders Irbäck. Distinct phases of free α-synuclein–a Monte Carlo study. *Proteins*, 80(9):2169–77, aug 2012.

[207] Massimo Sandal, Francesco Valle, Isabella Tessari, Stefano Mammi, Elisabetta Bergantino, Francesco Musiani, Marco Brucale, Luigi Bubacco, and Bruno Samorì. Conformational Equilibria in Monomeric α-Synuclein at the Single-Molecule Level. *PLoS Biol.*, 6(1):e6, jan 2008.

[208] Shahin Zibaee, O. Sumner Makin, Michel Goedert, and Louise C. Serpell. A simple algorithm locates β-strands in the amyloid fibril core of α-synuclein, Aβ, and tau using the amino acid sequence alone. *Protein Sci.*, 16(5):906–918, may 2007.

[209] Marija Iljina, Gonzalo A. Garcia, Mathew H. Horrocks, Laura Tosatto, Minee L. Choi, Kristina A. Ganzinger, Andrey Y. Abramov, Sonia Gandhi, Nicholas W. Wood, Nunilo Cremades, Christopher M. Dobson, Tuomas P. J. Knowles, and David Klenerman. Kinetic model of the aggregation of alpha-synuclein provides insights into prion-like spreading. *Proc. Natl. Acad. Sci.*, 113(9):E1206–E1215, mar 2016.

[210] Jiyong Lee, Elizabeth K. Culyba, Evan T. Powers, and Jeffery W. Kelly. Amyloid-beta forms fibrils by nucleated conformational conversion of oligomers. *Nat. Chem. Biol.*, 7(9):602–609, jul 2011.

[211] Ashwani K. Thakur, Murali Jayaraman, Rakesh Mishra, Monika Thakur, Veronique M. Chellgren, I. J. Byeon, Dalaver H. Anjum, Ravindra Kodali, Trevor P. Creamer, James F. Conway, Angela M. Gronenborn, and Ronald Wetzel. Polyglutamine disruption of the huntingtin exon 1 N terminus triggers a complex aggregation mechanism. *Nat. Struct. Mol. Biol.*, 16(4):380–9, apr 2009.

[212] L. Wei, P. Jiang, W. Xu, H. Li, H. Zhang, L. Yan, M. B. Chan-Park, X.-W. Liu, K. Tang, Y. Mu, and K. Pervushin. The Molecular Basis of Distinct Aggregation Pathways of Islet Amyloid Polypeptide. *J. Biol. Chem.*, 286(8):6291–6300, feb 2011.

[213] Nunilo Cremades, Serene W. Chen, and Christopher M. Dobson. *Structural Characteristics of alpha-Synuclein Oligomers*, volume 329. Elsevier Inc., 1 edition, 2017.

[214] K. A. Conway, J. D. Harper, and Peter T. Lansbury. Accelerated in vitro fibril formation by a mutant alpha-synuclein linked to early-onset Parkinson disease. *Nat. Med.*, 4(11):1318–20, nov 1998.

[215] K. A. Conway, S. J. Lee, J. C. Rochet, T. T. Ding, J. D. Harper, R. E. Williamson, and Peter T. Lansbury. Accelerated oligomerization by Parkinson's disease linked alpha-synuclein mutants. *Ann. N. Y. Acad. Sci.*, 920:42–5, 2000.

[216] K. A. Conway, J. D. Harper, and Peter T. Lansbury. Fibrils formed in vitro from alpha-synuclein and two mutant forms linked to Parkinson's disease are typical amyloid. *Biochemistry*, 39(10):2552–63, mar 2000.

[217] Hilal A. Lashuel, Benjamin M. Petre, Joseph Wall, Martha Simon, Richard J. Nowak, Thomas Walz, and Peter T. Lansbury. Alpha-synuclein, especially the Parkinson's disease-associated mutants, forms pore-like annular and tubular protofibrils. *J. Mol. Biol.*, 322(5):1089–102, oct 2002.

[218] Hilal A. Lashuel, Dean Hartley, Benjamin M. Petre, Thomas Walz, and Peter T. Lansbury. Neurodegenerative disease: Amyloid pores from pathogenic mutations. *Nature*, 418(6895):291–291, jul 2002.

[219] Serene W. Chen, S. Drakulic, Emma Deas, Myriam M. Ouberai, Francesco A. Aprile, Rocío Arranz, Samuel Ness, Cintia Roodveldt, Tim Guilliams, Erwin J. De-Genst,

David Klenerman, Nicholas W. Wood, Tuomas P. J. Knowles, Carlos Alfonso, Germán Rivas, Andrey Y. Abramov, José María Valpuesta, Christopher M. Dobson, and Nunilo Cremades. Structural characterization of toxic oligomers that are kinetically trapped during alpha-synuclein fibril formation. *Proc. Natl. Acad. Sci. U. S. A.*, 112(16):E1994–2003, apr 2015.

[220] Nikolai Lorenzen, Søren Bang Nielsen, Alexander K. Buell, Jørn Døvling Kaspersen, Paolo Arosio, Brian Stougaard Vad, Wojciech Paslawski, Gunna Christiansen, Zuzana Valnickova-Hansen, Maria Andreasen, Jan J. Enghild, Jan Skov Pedersen, Christopher M. Dobson, Tuomas P. J. Knowles, and Daniel Erik Otzen. The role of stable alpha-synuclein oligomers in the molecular events underlying amyloid formation. *J. Am. Chem. Soc.*, 136(10):3859–68, mar 2014.

[221] Niels Zijlstra, Christian Blum, Ine M. J. Segers-Nolten, Mireille M. A. E. Claessens, and Vinod Subramaniam. Molecular composition of sub-stoichiometrically labeled α-synuclein oligomers determined by single-molecule photobleaching. *Angew. Chemie Int. Ed.*, 51(35):8821–8824, aug 2012.

[222] María Soledad Celej, Rabia Sarroukh, Erik Goormaghtigh, Gerardo D. Fidelio, Jean-Marie Ruysschaert, and Vincent Raussens. Toxic prefibrillar α-synuclein amyloid oligomers adopt a distinctive antiparallel β-sheet structure. *Biochem. J.*, 443(3):719–726, may 2012.

[223] Mihaela M. Apetri, Nakul C. Maiti, Michael G. Zagorski, Paul R. Carey, and Vernon E. Anderson. Secondary Structure of α-Synuclein Oligomers: Characterization by Raman and Atomic Force Microscopy. *J. Mol. Biol.*, 355(1):63–71, jan 2006.

[224] A. Quist, I. Doudevski, H. Lin, R. Azimova, D. Ng, B. Frangione, B. Kagan, J. Ghiso, and R. Lal. Amyloid ion channels: A common structural link for protein-misfolding disease. *Proc. Natl. Acad. Sci.*, 102(30):10427–10432, jul 2005.

[225] Dong-Pyo Hong, Anthony L. Fink, and Vladimir N. Uversky. Structural characteristics of alpha-synuclein oligomers stabilized by the flavonoid baicalein. *J. Mol. Biol.*, 383(1):214–223, oct 2008.

[226] Jia-Hong Lu, Mustafa Taleb Ardah, Siva Sundara Kumar Durairajan, Liang-Feng Liu, Li-Xia Xie, Wang-Fun David Fong, Mohamed Y. Hasan, Jian-Dong Huang, Omar M. A. El-Agnaf, and Min Li. Baicalein inhibits formation of α-synuclein oligomers within living cells and prevents Aβ peptide fibrillation and oligomerisation. *Chembiochem*, 12(4):615–24, mar 2011.

[227] M. Zhu, S. Rajamani, J. Kaylor, S. Han, F. Zhou, and A. L. Fink. The flavonoid baicalein inhibits fibrillation of alpha-synuclein and disaggregates existing fibrils. *J. Biol. Chem.*, 279(26):26846–26857, jun 2004.

[228] Jan Bieschke, Jenny Russ, Ralf P. Friedrich, Dagmar E. Ehrnhoefer, Heike Wobst, Katja Neugebauer, and Erich E. Wanker. EGCG remodels mature α-synuclein and

amyloid-β fibrils and reduces cellular toxicity. *Proc. Natl. Acad. Sci.*, 107(17):7710–7715, apr 2010.

[229] Dagmar E. Ehrnhoefer, Jan Bieschke, Annett Boeddrich, Martin Herbst, Laura Masino, Rudi Lurz, Sabine Engemann, Annalisa Pastore, and Erich E. Wanker. EGCG redirects amyloidogenic polypeptides into unstructured, off-pathway oligomers. *Nat. Struct. Mol. Biol.*, 15(6):558–66, jun 2008.

[230] Istvan Horvath, Christoph F. Weise, Emma K. Andersson, Erik Chorell, Magnus Sellstedt, Christoffer Bengtsson, Anders Olofsson, Scott J. Hultgren, Matthew Chapman, Magnus Wolf-Watz, Fredrik Almqvist, and Pernilla Wittung-Stafshede. Mechanisms of protein oligomerization: inhibitor of functional amyloids templates alpha-synuclein fibrillation. *J. Am. Chem. Soc.*, 134(7):3439–3444, feb 2012.

[231] Martin Nors Pedersen, Vito Foderà, Istvan Horvath, Andreas van Maarschalkerweerd, Katrine Nørgaard Toft, Christoph F. Weise, Fredrik Almqvist, Magnus Wolf-Watz, Pernilla Wittung-Stafshede, and Bente Vestergaard. Direct Correlation Between Ligand-Induced alpha-Synuclein Oligomers and Amyloid-like Fibril Growth. *Sci. Rep.*, 5:10422, jan 2015.

[232] Lise Giehm, Dmitri I. Svergun, Daniel Erik Otzen, and Bente Vestergaard. Low-resolution structure of a vesicle disrupting alpha-synuclein oligomer that accumulates during fibrillation. *Proc. Natl. Acad. Sci. U. S. A.*, 108(8):3246–51, mar 2011.

[233] Rachel Lowe, Dean L. Pountney, Poul Henning Jensen, Wei Ping Gai, and Nicolas H. Voelcker. Calcium(II) selectively induces α-synuclein annular oligomers via interaction with the C-terminal domain. *Protein Sci.*, 13(12):3245–3252, jan 2004.

[234] Nelson B. Cole, D. D. Murphy, J. Lebowitz, L. Di Noto, R. L. Levine, and R. L. Nussbaum. Metal-catalyzed oxidation of alpha-synuclein: helping to define the relationship between oligomers, protofibrils, and filaments. *J. Biol. Chem.*, 280(10):9678–9690, mar 2005.

[235] K. A. Conway, J. C. Rochet, R. M. Bieganski, and Peter T. Lansbury. Kinetic stabilization of the alpha-synuclein protofibril by a dopamine-alpha-synuclein adduct. *Science*, 294(5545):1346–9, nov 2001.

[236] Jie Li, Min Zhu, Sudha Rajamani, Vladimir N. Uversky, and Anthony L. Fink. Rifampicin inhibits alpha-synuclein fibrillation and disaggregates fibrils. *Chem. Biol.*, 11(11):1513–21, nov 2004.

[237] E. H. Norris, B. I. Giasson, R. Hodara, S. Xu, John Q. Trojanowski, H. Ischiropoulos, and V. M. Lee. Reversible Inhibition of alpha-Synuclein Fibrillization by Dopaminochrome-mediated Conformational Alterations. *J. Biol. Chem.*, 280(22):21212–21219, jun 2005.

[238] Chi L.L. Pham, Su Ling Leong, Feda E. Ali, Vijaya B. Kenche, Andrew F. Hill, Sally L. Gras, Kevin J. Barnham, and Roberto Cappai. Dopamine and the Dopamine Oxidation Product 5,6-Dihydroxylindole Promote Distinct On-Pathway and Off-Pathway Aggregation of α-Synuclein in a pH-Dependent Manner. *J. Mol. Biol.*, 387(3):771–785, apr 2009.

[239] Hai-Young Kim, Min-Kyu Cho, Dietmar Riedel, Claudio O. Fernandez, and Markus Zweckstetter. Dissociation of Amyloid Fibrils of alpha-Synuclein in Supercooled Water. *Angew. Chemie Int. Ed.*, 47(27):5046–5048, jun 2008.

[240] Hai-Young Kim, Min-Kyu Cho, Ashutosh Kumar, Elke Maier, Carsten Siebenhaar, Stefan Becker, Claudio O. Fernandez, Hilal A. Lashuel, Roland Benz, Adam Lange, and Markus Zweckstetter. Structural Properties of Pore-Forming Oligomers of α-Synuclein. *J. Am. Chem. Soc.*, 131(47):17482–17489, dec 2009.

[241] Edward N. Baker, Thomas L. Blundell, John F. Cutfield, Susan M. Cutfield, Eleanor J. Dodson, Guy G. Dodson, Dorothy M. Crowfoot Hodgkin, Roderick E. Hubbard, Neil W. Isaacs, Colin D. Reynolds, Kiwako Sakabe, Norioshi Sakabe, and Numminate M. Vijayan. The Structure of 2Zn Pig Insulin Crystals at 1.5 Å Resolution. *Philos. Trans. R. Soc. London B Biol. Sci.*, 319(1195), 1988.

[242] Liza Nielsen, R. Khurana, A. Coats, S. Frokjaer, J. Brange, S. Vyas, Vladimir N. Uversky, and A. L. Fink. Effect of environmental factors on the kinetics of insulin fibril formation: elucidation of the molecular mechanism. *Biochemistry*, 40(20):6036–46, may 2001.

[243] U. Derewendat, Z. Derewendaf, Eleanor J. Dodson, G. G. Dodsonl, Xiao Bing, and Jan Markussen. X-ray Analysis of the Single Chain B29-A1 Peptide-linked Insulin Molecule. A Completely Inactive Analogue. *J. Mol. Rid*, 220:425–433, 1991.

[244] Vincent Zoete, Markus Meuwly, and Martin Karplus. A Comparison of the Dynamic Behavior of Monomeric and Dimeric Insulin Shows Structural Rearrangements in the Active Monomer. *J. Mol. Biol.*, pages 913–929, 2004.

[245] A. Ahmad, Vladimir N. Uversky, D. Hong, and A. L. Fink. Early Events in the Fibrillation of Monomeric Insulin. *J. Biol. Chem.*, 280(52):42669–42675, dec 2005.

[246] Liza Nielsen, S. Frokjaer, J. Brange, Vladimir N. Uversky, and A. L. Fink. Probing the mechanism of insulin fibril formation with insulin mutants. *Biochemistry*, 40(28):8397–8409, 2001.

[247] Alessandro Podestà, Guido Tiana, Paolo Milani, and Mauro Manno. Early events in insulin fibrillization studied by time-lapse atomic force microscopy. *Biophys. J.*, 90(2):589–97, jan 2006.

[248] S.G. Albert, J. Obadiah, S.A. Parseghian, Yadira M. Hurley, and A.D. Mooradian. Severe insulin resistance associated with subcutaneous amyloid deposition. *Diabetes Res. Clin. Pr.*, 75:374–376, 2007.

[249] F. E. Dische, C. Wernstedt, G. T. Westermark, P. Westermark, M. B. Pepys, J. A. Rennie, S. G. Gilbey, and P. J. Watkins. Insulin as an amyloid-fibril protein at sites of repeated insulin injections in a diabetic patient. *Diabetologia*, 31(3):158–161, mar 1988.

[250] M. P. S. Sie, H. E. van der Wiel, F. M. M. Smedts, and A. C. de Boer. Human recombinant insulin and amyloidosis: an unexpected association. *Neth. J. Med.*, 68(3):138–40, mar 2010.

[251] S. Störkel, H. M. Schneider, H. Müntefering, and S. Kashiwagi. Iatrogenic, insulin-dependent, local amyloidosis. *Lab. Invest.*, 48(1):108–11, jan 1983.

[252] Saniye Yumlu, Robert Barany, Magdalena Eriksson, and Christoph Röcken. Localized insulin-derived amyloidosis in patients with diabetes mellitus: a case report. *Hum. Pathol.*, 40(11):1655–1660, nov 2009.

[253] Terumasa Nagase, Keiichi Iwaya, Yoshiki Iwaki, Fumio Kotake, Ryuji Uchida, Tsunao Oh-I, Hidenori Sekine, Kazuhiro Miwa, Satoshi Murakami, Tomotada Odaka, Masahiko Kure, Yoko Nemoto, Masayuki Noritake, and Yoshiya Katsura. Insulin-derived Amyloidosis and Poor Glycemic Control: A Case Series. *Am. J. Med.*, 127:450–454, 2014.

[254] Mario Bouchard, Jesús Zurdo, Ewan J. Nettleton, Christopher M. Dobson, and Carol V. Robinson. Formation of insulin amyloid fibrils followed by FTIR simultaneously with CD and electron microscopy. *Protein Sci.*, 9(10):1960–7, oct 2000.

[255] David F. Waugh. A Fibrous Modification of Insulin. I. The Heat Precipitate of Insulin. *J. Am. Chem. Soc.*, 68(2):247–250, feb 1946.

[256] Wei Wang. Protein aggregation and its inhibition in biopharmaceutics. *Int. J. Pharm.*, 289(1-2):1–30, jan 2005.

[257] R. T. Darrington and B. D. Anderson. Effects of insulin concentration and self-association on the partitioning of its A-21 cyclic anhydride intermediate to desamido insulin and covalent dimer. *Pharm. Res.*, 12(7):1077–84, jul 1995.

[258] Victoria Sluzky, J. A. Tamada, A. M. Klibanov, and R. Langer. Kinetics of insulin aggregation in aqueous solutions upon agitation in the presence of hydrophobic surfaces. *Proc. Natl. Acad. Sci. U. S. A.*, 88(21):9377–81, nov 1991.

[259] Victoria Sluzky, Alexander M. Klibanov, and Robert Langer. Mechanism of insulin aggregation and stabilization in agitated aqueous solutions. *Biotechnol. Bioeng.*, 40(8):895–903, oct 1992.

[260] José L. Jiménez, Ewan J. Nettleton, Mario Bouchard, Carol V. Robinson, Christopher M. Dobson, and Helen R. Saibil. The protofilament structure of insulin amyloid fibrils. *Proc. Natl. Acad. Sci. U. S. A.*, 99(14):9196–201, 2002.

[261] Liza Nielsen, S. Frokjaer, J. F. Carpenter, and J. Brange. Studies of the structure of insulin fibrils by Fourier transform infrared (FTIR) spectroscopy and electron microscopy. *J. Pharm. Sci.*, 90(1):29–37, jan 2001.

[262] E. D. Eanes and G.G. Glenner. X-ray diffraction studies on amyloid filaments. *J. Histochem. Cytochem.*, 16(11):673–677, nov 1968.

[263] Ewan J. Nettleton, P Tito, M. Sunde, Mario Bouchard, Christopher M. Dobson, and Carol V. Robinson. Characterization of the oligomeric states of insulin in self-assembly and amyloid fibril formation by mass spectrometry. *Biophys. J.*, 79(2):1053–65, aug 2000.

[264] Paula Tito, Ewan J. Nettleton, and Carol V. Robinson. Dissecting the hydrogen exchange properties of insulin under amyloid fibril forming conditions: a site-specific investigation by mass spectrometry. *J. Mol. Biol.*, 303(2):267–278, oct 2000.

[265] W. Colon and J. W. Kelly. Partial denaturation of transthyretin is sufficient for amyloid fibril formation in vitro. *Biochemistry*, 31(36):8654–60, sep 1992.

[266] J. R. Glover, A. S. Kowal, E. C. Schirmer, M. M. Patino, J. J. Liu, and S. L. Lindquist. Self-seeded fibers formed by Sup35, the protein determinant of [PSI+], a heritable prion-like factor of S. cerevisiae. *Cell*, 89(5):811–9, may 1997.

[267] Cristian Ionescu-Zanetti, R. Khurana, J. R. Gillespie, J. S. Petrick, L. C. Trabachino, L. J. Minert, S. A. Carter, and A. L. Fink. Monitoring the assembly of Ig light-chain amyloid fibrils by atomic force microscopy. *Proc. Natl. Acad. Sci. U. S. A.*, 96(23):13175–9, nov 1999.

[268] J. T. Jarrett and Peter T. Lansbury. Seeding "one-dimensional crystallization" of amyloid: a pathogenic mechanism in Alzheimer's disease and scrapie? *Cell*, 73(6):1055–8, jun 1993.

[269] Mahvish Muzaffar and Atta Ahmad. The mechanism of enhanced insulin amyloid fibril formation by NaCl is better explained by a conformational change model. *PLoS One*, 6(11):e27906, nov 2011.

[270] Tomas Sneideris, Domantas Darguzis, Akvile Botyriute, Martynas Grigaliunas, Roland Winter, and Vytautas Smirnovas. pH-Driven Polymorphism of Insulin Amyloid-Like Fibrils. *PLoS One*, 10(8):e0136602, aug 2015.

[271] Atta Ahmad, Ian S. Millett, Sebastian Doniach, Vladimir N. Uversky, and Anthony L. Fink. Partially Folded Intermediates in Insulin Fibrillation. *Biochemistry*, 42(39):11404–11416, oct 2003.

[272] H. Naiki and F. Gejyo. Kinetic analysis of amyloid fibril formation. *Methods Enzymol.*, 309:305–18, 1999.

[273] Evan T. Powers and David L. Powers. The kinetics of nucleated polymerizations at high concentrations: amyloid fibril formation near and above the "supercritical concentration". *Biophys. J.*, 91(1):122–132, jul 2006.

[274] David F. Waugh, Darthea F. Wilhelmson, Spencer L. Commerford, and Muriel L. Sackler. Studies of the Nucleation and Growth Reactions of Selected Types of Insulin Fibrils. *J. Am. Chem. Soc.*, 75(11):2592–2600, jun 1953.

[275] Bente Vestergaard, Minna Groenning, Manfred Roessle, Jette S. Kastrup, Marco van de Weert, James M. Flink, Sven Frokjaer, Michael Gajhede, and Dmitri I. Svergun. A helical structural nucleus is the primary elongating unit of insulin amyloid fibrils. *PLoS Biol.*, 5:1089–1097, may 2007.

[276] David F. Waugh. A mechanism for the formation of fibrils from protein molecules. *J. Cell. Comp. Physiol.*, 49(S1):145–164, may 1957.

[277] Fumio Oosawa and Sho Asakura. *Thermodynamics of the polymerization of protein*. Academic Press, 1975.

[278] Naveen Kumar, Ankit Bansal, G. S. Sarma, and Ravindra K. Rawal. Chemometrics tools used in analytical chemistry: An overview. *Talanta*, 123C:186–199, jun 2014.

[279] I. T. Joliffe. *Principal component analysis*. 2005.

[280] Raymond B. Cattell. The Scree Test For The Number Of Factors. *Multivariate Behav. Res.*, 1(2):245–276, apr 1966.

[281] Joaquim Jaumot, Raimundo Gargallo, Anna de Juan, and Romà Tauler. A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB. *Chemom. Intell. Lab. Syst.*, 76(1):101–110, mar 2005.

[282] Anna de Juan and Romà Tauler. Chemometrics applied to unravel multicomponent processes and mixtures. *Anal. Chim. Acta*, 500(1-2):195–210, dec 2003.

[283] Susana Navea, Romà Tauler, and Anna de Juan. Monitoring and Modeling of Protein Processes Using Mass Spectrometry, Circular Dichroism, and Multivariate Curve Resolution Methods. *Anal. Chem.*, 78(14):4768–4778, jul 2006.

[284] Romà Tauler. Multivariate curve resolution applied to second order data. *Chemom. Intell. Lab. Syst.*, 30(1):133–146, nov 1995.

[285] Romà Tauler, Age K. Smilde, and Bruce Kowalski. Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. *J. Chemom.*, 9:31–58, 1995.

[286] Marcel Maeder. Evolving factor analysis for the resolution of overlapping chromatographic peaks. *Anal. Chem.*, 59(3):527–530, feb 1987.

[287] Willem. Windig and Jean. Guilment. Interactive self-modeling mixture analysis. *Anal. Chem.*, 63(14):1425–1432, jul 1991.

[288] Edmund Malinowski. *Factor analysis in chemistry*. Wiley-Interscience, New York, NY, 3rd edition, 2002.

[289] Charles L. Lawson and Richard J. Hanson. *Solving Least Squares Problems*. Prentice Hall, Englewood Cliffs, NJ, 1974.

[290] Rasmus Bro and Sijmen de Jong. A fast non-negativity-constrained least squares algorithm. *J. Chemom.*, 11(5):393–401, sep 1997.

[291] Anna de Juan, Y. Vander Heyden, Romà Tauler, and D.L. Massart. Assessment of new constraints applied to the alternating least squares method. *Anal. Chim. Acta*, 346(3):307–318, 1997.

[292] Rasmus Bro and Nicholaos D. Sidiropoulos. Least squares algorithms under Unimodality and Non-negativity constraints. *J. Chemom.*, 1998.

[293] Romà Tauler and E. Casassas. Principal Component Analysis Applied to the Study of Successive Complex Formation Data in the Cu(II) ethanolamine Systems. *J. Chemom.*, 3:151–61, 1988.

[294] Rolf Manne. On the resolution problem in hyphenated chromatography. *Chemom. Intell. Lab. Syst.*, 27(1):89–94, jan 1995.

[295] Harald Gampp, Marcel Maeder, Charles J. Meyer, and Andreas D. Zuberbühler. Calculation of equilibrium constants from multiwavelength spectroscopic data-II. *Talanta*, 32(4):257–264, apr 1985.

[296] Anna de Juan, Susana Navea, Josef Diewok, and Romà Tauler. Local rank exploratory analysis of evolving rank-deficient systems. *Chemom. Intell. Lab. Syst.*, 70(1):11–21, jan 2004.

[297] Adriano Sebollela, Gina-Mirela Mustata, Kevin Luo, Pauline T. Velasco, Kirsten L. Viola, Erika N. Cline, Gajendra S. Shekhawat, Kyle C. Wilcox, Vinayak P. Dravid, and William L. Klein. Elucidating molecular mass and shape of a neurotoxic Abeta oligomer. *ACS Chem. Neurosci.*, 5(12):1238–1245, dec 2014.

[298] Cristiano Luis Pinto Oliveira, Manja Annette Behrens, Jesper Søndergaard Pedersen, Kurt Erlacher, Daniel Erik Otzen, and Jan Skov Pedersen. A SAXS Study of Glucagon Fibrillation. *J. Mol. Biol.*, 387(1):147–161, mar 2009.

[299] Ashley J. Pratt, David S. Shin, Gregory E. Merz, Robert P. Rambo, W. Andrew Lancaster, Kevin N. Dyer, Peter P. Borbat, Farris L. Poole, Michael W. W. Adams, Jack H. Freed, Brian R. Crane, John A. Tainer, and Elizabeth D. Getzoff. Aggregation propensities of superoxide dismutase G93 hotspot mutants mirror ALS clinical phenotypes. *Proc. Natl. Acad. Sci. U. S. A.*, 111(43):E4568–76, oct 2014.

[300] Annette Eva Langkilde and Bente Vestergaard. Structural Characterization of Prefibrillar Intermediates and Amyloid Fibrils by Small-Angle X-Ray Scattering. In Einar M. Sigurdsson, Miguel Calero, and María Gasset, editors, *Amyloid Proteins Methods Protoc.*, volume 849 of *Methods in Molecular Biology*, pages 137–155. Humana Press, Totowa, NJ, 2012.

[301] Guillaume Tresset, Clémence Le Coeur, Jean-François Bryche, Mouna Tatou, Mehdi Zeghal, Annie Charpilienne, Didier Poncet, Doru Constantin, and Stéphane Bressanelli. Norovirus Capsid Proteins Self-Assemble through Biphasic Kinetics via Long-Lived Stave-like Intermediates. *J. Am. Chem. Soc.*, 135(41):15373–15381, oct 2013.

[302] Johan Trygg, Elaine Holmes, and Torbjörn Lundstedt. Chemometrics in metabonomics. *J. Proteome Res.*, 6(2):469–79, feb 2007.

[303] Jascha Blobel, Pau Bernadó, Dmitri I. Svergun, Romà Tauler, and Miquel Pons. Low-resolution structures of transient protein-protein complexes using small-angle X-ray scattering. *J. Am. Chem. Soc.*, 131(12):4378–86, apr 2009.

[304] Tim E. Williamson, Bruce A. Craig, Elena Kondrashkina, Chris Bailey-Kellogg, and Alan M. Friedman. Analysis of self-associating proteins by singular value decomposition of solution scattering data. *Biophys. J.*, 94(12):4906–23, jun 2008.

[305] Shuji Akiyama, Satoshi Takahashi, Tetsunari Kimura, Koichiro Ishimori, Isao Morishima, Yukihiro Nishikawa, and Tetsuro Fujisawa. Conformational landscape of cytochrome c folding studied by microsecond-resolved small-angle x-ray scattering. *Proc. Natl. Acad. Sci. U. S. A.*, 99(3):1329–34, feb 2002.

[306] Sara Ayuso-Tejedor, Rebeca García-Fandiño, Modesto Orozco, Javier Sancho, and Pau Bernadó. Structural analysis of an equilibrium folding intermediate in the apoflavodoxin native ensemble by small-angle X-ray scattering. *J. Mol. Biol.*, 406(4):604–19, mar 2011.

[307] L. Chen, K. O. Hodgson, and Sebastian Doniach. A lysozyme folding intermediate revealed by solution X-ray scattering. *J. Mol. Biol.*, 261(5):658–71, sep 1996.

[308] Daniel J. Segel, Anthony L. Fink, K. O. Hodgson, and Sebastian Doniach. Protein denaturation: a small-angle X-ray scattering study of the ensemble of unfolded states of cytochrome c. *Biochemistry*, 37(36):12443–51, sep 1998.

[309] David D. L. Minh and Lee Makowski. Wide-angle X-ray solution scattering for protein-ligand binding: multivariate curve resolution with Bayesian confidence intervals. *Biophys. J.*, 104(4):873–83, feb 2013.

[310] Fátima Herranz-Trillo, Minna Groenning, Andreas van Maarschalkerweerd, Romà Tauler, Bente Vestergaard, and Pau Bernadó. Structural Analysis of Multi-component Amyloid Systems by Chemometric SAXS Data Decomposition. *Structure*, 25(1):5–15, 2017.

[311] Dmitri I. Svergun. Determination of the regularization parameter in indirect-transform methods using perceptual criteria. *J. Appl. Crystallogr.*, 25(4):495–503, aug 1992.

[312] Róbert Rajkó. Additional knowledge for determining and interpreting feasible band boundaries in self-modeling/multivariate curve resolution of two-component systems. *Anal. Chim. Acta*, 661(2):129–132, mar 2010.

[313] Romà Tauler. Calculation of maximum and minimum band boundaries of feasible solutions for species profiles obtained by multivariate curve resolution. *J. Chemom.*, 15(8):627–646, sep 2001.

[314] Robert P. Rambo and John A. Tainer. Accurate assessment of mass, models and resolution by small-angle scattering. *Nature*, 496(7446):477–81, apr 2013.

[315] Efstratios Mylonas and Dmitri I. Svergun. Accuracy of molecular mass determination of proteins in solution by small-angle X-ray scattering. *J. Appl. Crystallogr.*, 40(s1):245–249, feb 2007.

[316] Evan T. Powers and David L. Powers. Mechanisms of protein fibril formation: nucleated polymerization with competing off-pathway aggregation. *Biophys. J.*, 94(2):379–91, jan 2008.

[317] Ross A. Fredenburg, Carla Rospigliosi, Robin K. Meray, Jeffrey C. Kessler, Hilal A. Lashuel, David Eliezer, and Peter T. Lansbury. The impact of the E46K mutation on the properties of alpha-synuclein in its monomeric and oligomeric states. *Biochemistry*, 46(24):7107–18, jun 2007.

[318] Susana Navea, Anna de Juan, and Romà Tauler. Detection and resolution of intermediate species in protein folding processes using fluorescence and circular dichroism spectroscopies and multivariate curve resolution. *Anal. Chem.*, 74(23):6031–9, dec 2002.

[319] Yann Fichou, Giorgio Schirò, François-Xavier Gallat, Cedric Laguri, Martine Moulin, Jérôme Combet, Michaela Zamponi, Michael Härtlein, Catherine Picart, Estelle Mossou, Hugues Lortat-Jacob, Jacques-Philippe Colletier, Douglas J. Tobias, and Martin Weik. Hydration water mobility is enhanced around tau amyloid fibers. *Proc. Natl. Acad. Sci. U. S. A.*, 112(20):6365–70, may 2015.

[320] Søren Bang Nielsen, Francesca Macchi, Samuele Raccosta, Annette Eva Langkilde, Lise Giehm, Anders Kyrsting, Anna Sigrid Pii Svane, Mauro Manno, Gunna Christiansen, Niels Christian Nielsen, Lene Oddershede, Bente Vestergaard, and Daniel Erik Otzen. Wildtype and A30P mutant alpha-synuclein form different fibril structures. *PLoS One*, 8(7):e67713, jan 2013.

[321] Samuel I. A. Cohen, Michele Vendruscolo, Christopher M. Dobson, and Tuomas P. J. Knowles. From macroscopic measurements to microscopic mechanisms of protein aggregation. *J. Mol. Biol.*, 421(2-3):160–71, aug 2012.

[322] Adam R. Round, Daniel Franke, S. Moritz, R. Huchler, M. Fritsche, D. Malthan, R. Klaering, Dmitri I. Svergun, and Manfred Roessle. Automated sample-changing robot for solution scattering experiments at the EMBL Hamburg SAXS station X33. *J. Appl. Crystallogr.*, 41(Pt 5):913–917, oct 2008.

[323] Daniel Franke, Alexey G. Kikhney, and Dmitri I. Svergun. Automated acquisition and analysis of small angle X-ray scattering data. *Nucl. Instruments Methods Phys. Res. Sect. A Accel. Spectrometers, Detect. Assoc. Equip.*, 689:52–59, oct 2012.

[324] Georgii G. Krivov, Maxim V. Shapovalov, and Roland L. Dunbrack. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, 77(4):778–95, dec 2009.

[325] Dmitri I. Svergun and Jan Skov Pedersen. Propagating errors in small-angle scattering data treatment. *J. Appl. Crystallogr.*, 27(3):241–248, jun 1994.

[326] Eric F. Pettersen, Thomas D. Goddard, Conrad C. Huang, Gregory S. Couch, Daniel M. Greenblatt, Elaine C. Meng, and Thomas E. Ferrin. UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.*, 25(13):1605–12, oct 2004.

[327] Willem. Windig, Neal B. Gallagher, Jeremy M. Shaver, and Barry M. Wise. A new approach for interactive self-modeling mixture analysis. *Chemom. Intell. Lab. Syst.*, 77(1-2):85–96, may 2005.

[328] H. Pelletier and J. Kraut. Crystal structure of a complex between electron transfer partners, cytochrome c peroxidase and cytochrome c. *Science (80-.).*, 258:1748–1755, 1992.

[329] Stephanie Hutin, Martha E. Brennich, Benoit Maillot, and Adam R. Round. Online ion-exchange chromatography for small-angle X-ray scattering. *Acta Crystallogr. Sect. D Struct. Biol.*, 72:1090–1099, 2016.

[330] Daniel Franke, Maxim V. Petoukhov, Petr V. Konarev, A. Panjkovich, A. Tuukkanen, Haydyn D. T. Mertens, A. G. Kikhney, N. R. Hajizadeh, J. M. Franklin, Cy M. Jeffries, and Dmitri I. Svergun. ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *J. Appl. Cryst*, 2017.

[331] Abraham López, Fátima Herranz-Trillo, Martin Kotev, Margarida Gairí, Víctor Guallar, Pau Bernadó, Oscar Millet, Teresa Tarragó, and Ernest Giralt. Active-Site-Directed Inhibitors of Prolyl Oligopeptidase Abolish Its Conformational Dynamics. *ChemBioChem*, 17(10):913–917, 2016.

[332] D. F. Cunningham and B. O'Connor. A study of prolyl endopeptidase in bovine serum and its relevance to the tissue enzyme. *Int. J. Biochem. Cell Biol.*, 30(1):99–114, jan 1998.

[333] Vilmos Fülöp, Z. Böcskei, and L. Polgár. Prolyl oligopeptidase: an unusual beta-propeller domain regulates proteolysis. *Cell*, 94(2):161–70, jul 1998.

[334] Min Li, Changqing Chen, David R. Davies, and Thang K. Chiu. Induced-fit mechanism for prolyl endopeptidase. *J. Biol. Chem.*, 285(28):21487–95, jul 2010.

[335] L. Shan, I. I. Mathews, and C. Khosla. Structural and mechanistic analysis of two prolyl endopeptidases: Role of interdomain dynamics in catalysis and specificity. *Proc. Natl. Acad. Sci.*, 102(10):3599–3604, mar 2005.

[336] Peter Canning, Dean Rea, Rory E. Morty, Vilmos Fülöp, and L. C. Storoni. Crystal Structures of Trypanosoma brucei Oligopeptidase B Broaden the Paradigm of Catalytic Regulation in Prolyl Oligopeptidase Family Enzymes. *PLoS One*, 8(11):e79349, nov 2013.

[337] Nessim Kichik, Teresa Tarragó, Birgit Claasen, Margarida Gairí, Oscar Millet, and Ernest Giralt. 15N Relaxation NMR Studies of Prolyl Oligopeptidase, an 80 kDa Enzyme, Reveal a Pre-existing Equilibrium between Different Conformational States. *ChemBioChem*, 12(18):2737–2739, dec 2011.

[338] Karol Kaszuba, Tomasz Róg, Reinis Danne, Peter Canning, Vilmos Fülöp, Tünde Juhász, Zoltán Szeltner, J.-F. St. Pierre, Arturo García-Horsman, Pekka T. Männistö, Mikko Karttunen, Jyrki Hokkanen, and Alex Bunker. Molecular dynamics, crystallography and mutagenesis studies on the substrate gating mechanism of prolyl oligopeptidase. *Biochimie*, 94(6):1398–1411, jun 2012.

[339] Elena Di Daniel, Colin P. Glover, Emma Grot, Man K. Chan, Thirza H. Sanderson, Julia H. White, Catherine L. Ellis, Kathleen T. Gallagher, James Uney, Julia Thomas, Peter R. Maycox, and Anne W. Mudge. Prolyl oligopeptidase binds to GAP-43 and functions without its peptidase activity. *Mol. Cell. Neurosci.*, 41(3):373–382, jun 2009.

[340] Mari H. Savolainen, Xu Yan, Timo T. Myöhänen, and Henri J. Huttunen. Prolyl Oligopeptidase Enhances α-Synuclein Dimerization via Direct Protein-Protein Interaction. *J. Biol. Chem.*, 290(8):5117–5126, feb 2015.

[341] Haydyn D. T. Mertens and Dmitri I. Svergun. Structural characterization of proteins and complexes using small-angle X-ray solution scattering. *J. Struct. Biol.*, 172(1):128–141, oct 2010.

[342] David A. Jacques and Jill Trewhella. Small-angle scattering for structural biology-Expanding the frontier while avoiding the pitfalls. *Protein Sci.*, 19(4):642–657, apr 2010.

[343] T. Yoshimoto, K. Kawahara, F. Matsubara, K. Kado, and D. Tsuru. Comparison of inhibitory effects of prolinal-containing peptide derivatives on prolyl endopeptidases from bovine brain and Flavobacterium. *J. Biochem.*, 98(4):975–9, oct 1985.

[344] Abraham López. *Study of the conformational dynamics of prolyl oligopeptidase*. PhD thesis, Universitat de Barcelona, 2015.

[345] D.A. Case, T.A. Darden, T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, R.C. Walker, W. Zhang, K.M. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A.W. Goetz, I. Kolossvai, K.F. Wong, F. Paesani, J. Vanicek, R.M. Wolf, J. .

[346] Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins Struct. Funct. Bioinforma.*, 65(3):712–725, nov 2006.

[347] William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935, jul 1983.

[348] D. E. Shaw Research ("DESRES") - www.deshawresearch.com.

[349] Steve P. Meisburger, Alexander B. Taylor, Crystal A. Khan, Shengnan Zhang, Paul F. Fitzpatrick, and Nozomi Ando. Domain Movements upon Activation of Phenylalanine Hydroxylase Characterized by Crystallography and Chromatography-Coupled Small-Angle X-ray Scattering. *J. Am. Chem. Soc.*, 138(20):6506–6516, 2016.

[350] André J. Niestroj, Ulrich Heiser, Susanne Aust, and Hans-Ulrich Demuth. Inhibitors of Prolyl endopeptidase, 2006.

[351] Arístides Alberich, Cristina Ariño, José Manuel Díaz-Cruz, and Miquel Esteban. Multivariate curve resolution applied to the simultaneous analysis of electrochemical and spectroscopic data: Study of the Cd(II)/glutathione-fragment system by voltammetry and circular dichroism spectroscopy. *Anal. Chim. Acta*, 584(2):403–409, feb 2007.

[352] Clecio Dantas, Romà Tauler, and Márcia Miguel Castro Ferreira. Exploring in vivo violacein biosynthesis by application of multivariate curve resolution on fused UV–VIS absorption, fluorescence, and liquid chromatography–mass spectrometry data. *Anal. Bioanal. Chem.*, 405(4):1293–1302, feb 2013.

[353] L. Fotouhi, S. Yousefinejad, N. Salehi, A.A. Saboury, N. Sheibani, and A.A. Moosavi-Movahedi. Application of merged spectroscopic data combined with chemometric analysis for resolution of hemoglobin intermediates during chemical unfolding. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.*, 136:1974–1981, feb 2015.

[354] Mohammad-Bagher Gholivand, Ali R. Jalalvand, Hector C. Goicoechea, Raimundo Gargallo, and Thomas Skov. Chemometrics: An important tool for monitoring interactions of vitamin B7 with bovine serum albumin with the aim of developing an efficient biosensing system for the analysis of protein. *Talanta*, 132:354–365, jan 2015.

[355] Yingying Liu, Guowen Zhang, Ni Zeng, and Song Hu. Interaction between 8-methoxypsoralen and trypsin: Monitoring by spectroscopic, chemometrics and molecular docking approaches. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.*, 173:188–195, feb 2017.

[356] Brett A. Chromy, Richard J. Nowak, Mary P. Lambert, Kirsten L. Viola, Lei Chang, Pauline T. Velasco, Bryan W. Jones, Sara J. Fernandez, Pascale N. Lacor, Peleg Horowitz, Caleb E. Finch, Grant A. Krafft, and William L. Klein. Self-assembly of Abeta(1-42) into globular neurotoxins. *Biochemistry*, 42(44):12749–60, nov 2003.

[357] Greg M. Cole and Sally A. Frautschy. Alzheimer's amyloid story finds its star. *Trends Mol. Med.*, 12(9):395–6, sep 2006.

[358] Charles G. Glabe. Common mechanisms of amyloid oligomer pathogenesis in degenerative disease. *Neurobiol. Aging*, 27(4):570–575, apr 2006.

[359] J. Hardy and Dennis J. Selkoe. The Amyloid Hypothesis of Alzheimer's Disease: Progress and Problems on the Road to Therapeutics. *Science (80-.).*, 297(5580):353–356, jul 2002.

[360] C. A. McLean, R. A. Cherny, F. W. Fraser, S. J. Fuller, M. J. Smith, K. Beyreuther, A. I. Bush, and C. L. Masters. Soluble pool of Abeta amyloid as a determinant of severity of neurodegeneration in Alzheimer's disease. *Ann. Neurol.*, 46(6):860–6, dec 1999.

[361] Andreas Aslund, Christina J. Sigurdson, Therése Klingstedt, Stefan Grathwohl, Tristan Bolmont, Dara L. Dickstein, Eirik Glimsdal, Stefan Prokop, Mikael Lindgren, Peter Konradsson, David M. Holtzman, Patrick R. Hof, Frank L. Heppner, Samuel Gandy, Mathias Jucker, Adriano Aguzzi, Per Hammarström, and K. Peter R. Nilsson. Novel pentameric thiophene derivatives for in vitro and in vivo optical imaging of a plethora of protein aggregates in cerebral amyloidoses. *ACS Chem. Biol.*, 4(8):673–84, aug 2009.

[362] Therése Klingstedt, Andreas Åslund, Rozalyn A. Simon, Leif B. G. Johansson, Jeffrey J. Mason, Sofie Nyström, Per Hammarström, and K. Peter R. Nilsson. Synthesis of a library of oligothiophenes and their utilization as fluorescent ligands for spectral assignment of protein aggregates. *Org. Biomol. Chem.*, 9(24):8356, dec 2011.

[363] K. Peter R. Nilsson, Andreas Åslund, Ina Berg, Sofie Nyström, Peter Konradsson, Anna Herland, Olle Inganäs, Frantz Stabo-Eeg, Mikael Lindgren, Gunilla T. Westermark, Lars Lannfelt, Lars N. G. Nilsson, and Per Hammarström. Imaging Distinct Conformational States of Amyloid-β Fibrils in Alzheimer's Disease Using Novel Luminescent Probes. *ACS Chem. Biol.*, 2(8):553–560, aug 2007.

[364] Per Hammarström, Rozalyn A. Simon, Sofie Nyström, Peter Konradsson, Andreas Åslund, and K. Peter R. Nilsson. A Fluorescent Pentameric Thiophene Derivative Detects in Vitro-Formed Prefibrillar Protein Aggregates. *Biochemistry*, 49(32):6838–6845, aug 2010.

[365] Karin Magnusson, Rozalyn A. Simon, Daniel Sjölander, Christina J. Sigurdson, Per Hammarström, and K. Peter R. Nilsson. Multimodal fluorescence microscopy of prion strain specific PrP deposits stained by thiophene-based amyloid ligands. *Prion*, 8(4):319–29, jul 2014.

[366] Christina J. Sigurdson, K. Peter R. Nilsson, Simone Hornemann, Giuseppe Manco, Magdalini Polymenidou, Petra Schwarz, Mario Leclerc, Per Hammarström, Kurt Wüthrich, and Adriano Aguzzi. Prion strain discrimination using luminescent conjugated polymers. *Nat. Methods*, 4(12):1023–30, dec 2007.

[367] Jonas Sjöqvist, Jerôme Maria, Rozalyn A. Simon, Mathieu Linares, Patrick Norman, K. Peter R. Nilsson, and Mikael Lindgren. Toward a molecular understanding of the detection of amyloid proteins with flexible conjugated oligothiophenes. *J. Phys. Chem. A*, 118(42):9820–9827, oct 2014.

[368] Martha E. Brennich, Adam R. Round, and Stephanie Hutin. Online Size-exclusion and Ion-exchange Chromatography on a SAXS Beamline. *J. Vis. Exp.*, (119):1–9, jan 2017.

[369] Francesc Puig-Castellví, Ignacio Alfonso, and Romà Tauler. Untargeted assignment and automatic integration of 1 H NMR metabolomic datasets using a multivariate curve resolution approach. *Anal. Chim. Acta*, 964:55–66, apr 2017.

[370] Yulia B. Monakhova, Alexey M. Tsikin, Thomas Kuballa, Dirk W. Lachenmeier, and Svetlana P. Mushtakova. Independent component analysis (ICA) algorithms for improved spectral deconvolution of overlapped signals in ¹ H NMR analysis: application to foods and related products. *Magn. Reson. Chem.*, 52(5):231–240, may 2014.

[371] Carla M. Teglia, Paola M. Peltzer, Silvia N. Seib, Rafael C. Lajmanovich, María J. Culzoni, and Héctor C. Goicoechea. Simultaneous multi-residue determination of twenty one veterinary drugs in poultry litter by modeling three-way liquid chromatography with fluorescence and absorption detection data. *Talanta*, 167:442–452, may 2017.

[372] Ana Carolina de Oliveira Neves, Romà Tauler, and Kássio Michell Gomes de Lima. Area correlation constraint for the MCRALS quantification of cholesterol using EEM fluorescence data: A new approach. *Anal. Chim. Acta*, 937:21–28, sep 2016.

[373] Alessandra Del Giudice, Cedric Dicko, Luciano Galantini, and Nicolae V. Pavel. Time-Dependent pH Scanning of the Acid-Induced Unfolding of Human Serum Albumin Reveals Stabilization of the Native Form by Palmitic Acid Binding. *J. Phys. Chem. B*, 121(17):4388–4399, may 2017.

[374] Masamune Morita, Mun'delanji Vestergaard, Tsutomu Hamada, and Masahiro Takagi. Real-time observation of model membrane dynamics induced by Alzheimer's amyloid beta. *Biophys. Chem.*, 147(1-2):81–86, mar 2010.

[375] Andreas van Maarschalkerweerd, Valeria Vetri, Annette Eva Langkilde, Vito Foderà, and Bente Vestergaard. Protein/lipid coaggregates are formed during alpha-synuclein-induced disruption of lipid bilayers. *Biomacromolecules*, 15(10):3643–3654, oct 2014.

Part IV

PUBLICATIONS

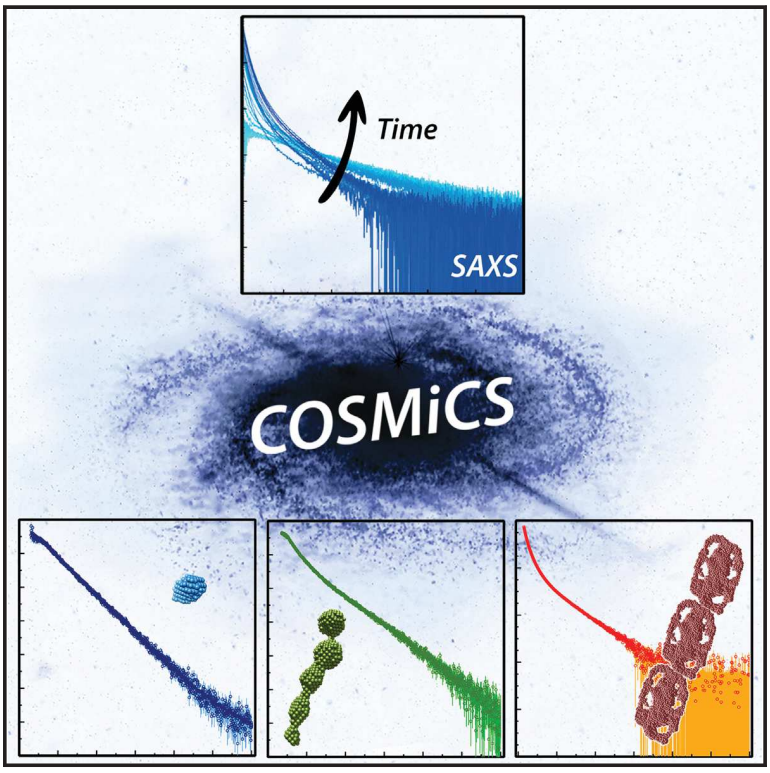
Paper I: Structural Analysis of Multi-component Amyloid Systems by Chemometric SAXS Data Decomposition

Article

Structure

Structural Analysis of Multi-component Amyloid Systems by Chemometric SAXS Data Decomposition

Graphical Abstract



Highlights

- SAXS is sensitive to the oligomeric forms appearing along an amyloidogenic process
- Mixtures can be chemometrically decomposed using multiple SAXS data representations
- Conversely to wild-type, α-synuclein E46K forms a large off-pathway disordered oligomer

Authors

Fátima Herranz-Trillo, Minna Groenning, Andreas van Maarschalkerweerd, Romà Tauler, Bente Vestergaard, Pau Bernadó

Correspondence

bente.vestergaard@sund.ku.dk (B.V.), pau.bernado@cbs.cnrs.fr (P.B.)

In Brief

Herranz-Trillo et al. have structurally characterized the species present along the fribrillation of insulin and α-synuclein E46K using SAXS data, identifying a new oligomeric form for the E46K mutant. The authors have disentangled data from amyloidogenic mixtures using a novel chemometric approach that simultaneously fits multiple SAXS representations.

Herranz-Trillo et al., 2017, Structure *25*, 5–15
January 3, 2017
http://dx.doi.org/10.1016/j.str.2016.10.013

CellPress

Structure
Article

Structural Analysis of Multi-component Amyloid Systems by Chemometric SAXS Data Decomposition

Fátima Herranz-Trillo,[1,2] Minna Groenning,[2,4] Andreas van Maarschalkerweerd,[2,4] Romà Tauler,[3] Bente Vestergaard,[2,*] and Pau Bernadó[1,5,*]

[1]Centre de Biochimie Structurale. INSERM U1054, CNRS UMR 5048, Université de Montpellier, 29, rue de Navacelles, 34090 Montpellier, France

[2]Department of Pharmacy and Department of Drug Design and Pharmacology, University of Copenhagen, Universitetsparken 2, 2100 Copenhagen, Denmark

[3]Environmental Chemometrics Group, Department of Environmental Chemistry, Institute of Environmental Assessment and Water Diagnostic (IDAEA-CSIC), 08034 Barcelona, Spain

[4]Present address: Novo Nordisk A/S, Novo Nordisk Park, 2760 Måløv, Denmark

[5]Lead Contact

*Correspondence: bente.vestergaard@sund.ku.dk (B.V.), pau.bernado@cbs.cnrs.fr (P.B.)

http://dx.doi.org/10.1016/j.str.2016.10.013

SUMMARY

Formation of amyloids is the hallmark of several neurodegenerative pathologies. Structural investigation of these complex transformation processes poses significant experimental challenges due to the co-existence of multiple species. The additive nature of small-angle X-ray scattering (SAXS) data allows for probing the evolution of these mixtures of oligomeric states, but the decomposition of SAXS data into species-specific spectra and relative concentrations is burdened by ambiguity. We present an objective SAXS data decomposition method by adapting the multivariate curve resolution alternating least squares (MCR-ALS) chemometric method. The approach enables rigorous and robust decomposition of synchrotron SAXS data by simultaneously introducing these data in different representations that emphasize molecular changes at different time and structural resolution ranges. The approach has allowed the study of fibrillogenic forms of insulin and the familial mutant E46K of α-synuclein, and is generally applicable to any macromolecular mixture that can be probed by SAXS.

INTRODUCTION

Some of the most important challenges in structural biology today concern processes involving highly dynamic large macromolecular complexes, the presence of concerted conformational fluctuations, and unstable, developing systems. The common feature of all these systems is their heterogeneity with multiple species or conformations co-existing in equilibrium. The application of traditional structural methods, such as macromolecular crystallography and high-resolution nuclear magnetic resonance is not straightforward for the structural analysis of changes associated with such highly complex systems.

Protein amyloid fibril formation is an example of such a challenging system. Protein fibrils are the hallmark of a number of severe diseases, notably the most common neurodegenerative pathologies Alzheimer's, Parkinson's, and Huntington's diseases (Chiti and Dobson, 2006; Moreno-Gonzalez and Soto, 2011). Protein fibrils are the final stage of the amyloid formation, while several soluble and transient oligomeric states are formed along the process. Increasing evidence places a central role on these transient species in the advancement of the disease, pinpointing the importance of their structural characterization (Bucciantini et al., 2002; Chiti and Dobson, 2006; Winner et al., 2011; Sebollela et al., 2014). This is, however, inherently difficult as the oligomers only exist in the context of other amyloidogenic species, and their physical isolation will perturb the equilibrium and potentially modify their structure.

Our group and others have shown that solution small-angle X-ray scattering (SAXS) is an ideal method for such investigations (Vestergaard et al., 2007; Oliveira et al., 2009; Giehm et al., 2011b; Pratt et al., 2014; Nors Pedersen et al., 2015). SAXS data are additive, thus measurements performed on a mixture correspond to a population-weighted average of the signal originating from all co-existing species. As a consequence, a dataset consisting of multiple curves obtained from developing mixtures with different relative populations of the same species is inherently very rich in information. In principle, it is possible to decompose such data series into the scattering profiles of the individual constituents (structures) as well as their relative populations (kinetics/thermodynamics) without physically isolating the co-existing species. The isolation of data originating from intermediately occurring species has previously been performed by different laborious and non-automated approaches for several amyloidogenic proteins (Vestergaard et al., 2007; Oliveira et al., 2009; Giehm et al., 2011b; Langkilde and Vestergaard, 2012; Pratt et al., 2014). Unfortunately, such strategies can only be applied to systems of limited complexity (Langkilde and Vestergaard, 2012), i.e., with a low number of co-existing species, and the imposition of initial and final data

curves as species pure. Therefore, it is clear that more robust and objective approaches are needed.

Objective decomposition of large datasets using chemometric approaches is routinely used in many research fields including analytical and organic chemistry and metabolomics (de Juan and Tauler, 2003; Trygg et al., 2007; Tresset et al., 2013). One of the most popular chemometric routines is multivariate curve resolution using alternate least squares (MCR-ALS) (Tauler, 1995; Jaumot et al., 2005). We and others have previously shown that chemometrics in general and MCR-ALS in particular are powerful tools in structural biology, allowing the study of transient biomolecular complexes (Williamson et al., 2008; Blobel et al., 2009) and folding processes (Chen et al., 1996; Segel et al., 1998; Akiyama et al., 2002; Ayuso-Tejedor et al., 2011) by SAXS or wide-angle X-ray scattering (Minh and Makowski, 2013). These systems were, however, significantly less complex than amyloidogenesis.

Degeneracy of mathematical solutions poses an intrinsic limitation of chemometric methods (Tauler et al., 1995; Tauler, 2001; Rajkó, 2010). Besides the use of constraints, the most efficient way to reduce ambiguity in MCR-ALS is the simultaneous analysis of multiple datasets measured under different experimental conditions and/or including additional data simultaneously measured using complementary techniques. However, in present SAXS beamlines the simultaneous measurement of complementary spectroscopic data is not possible. We have developed a new chemometric approach based on MCR-ALS with the capacity to decompose large SAXS datasets. A crucial aspect is that additional data from complementary sources are not required to solve potential solution degeneracies. In contrast, we show that zooming in on structural fingerprints at multiple length scales, by introducing SAXS data in different representations, guides the program to robust solutions. We have applied this approach, which we name COSMiCS (Complex Objective Structural analysis of Multi-Component Systems), to analyze SAXS data measured along the amyloid formation from two proteins with high biomedical interest: insulin and a familial mutant of α-synuclein (α-SN_{E46K}). The former case corroborates previously published results, and hence serves as proof of principle, while in the case of α-SN_{E46K} we reveal the presence of large transient oligomeric species of a significantly different nature than previously found for wild-type α-SN.

Our study describes a completely novel chemometrics-inspired strategy to derive structural and kinetic information of amyloidogenic processes that significantly complements the existing large toolbox of methods to address this biologically and medically relevant issue. In addition, it has a very broad range of applications in chemistry and biology to analyze complex data recorded along different reaction coordinates.

RESULTS

SAXS Data Collection and Primary Analysis on Fibrillating Insulin

Time-resolved synchrotron SAXS data were collected along a timeline of 11 hr from 28 fibrillating insulin samples, under experimental conditions comparable with those applied in a previous study (Vestergaard et al., 2007) while simultaneously monitoring the fibrillation kinetics with thioflavin T (ThT) fluorescence. The resulting profiles exhibit drastically increasing intensities in the small-angle region of the scattering patterns (Figures 1 and S1A) signifying a notable evolution from monomeric to fibril state. Using Guinier's approach, this evolution can be followed from the extrapolated forward scattering ($I(0)$) and the calculated radii of gyration (R_g) (Figure S2A). Principal component analysis (PCA) of the complete dataset indicated that three individual species significantly contributed to the time evolution of the data (Figures S3A and S3C).

COSMiCS Solves the Problem of Ambiguity When Decomposing the Insulin Fibrillation SAXS Dataset

The decomposition of the insulin SAXS intensities was performed using the MCR-ALS chemometric approach. As the principles behind MCR-ALS have been described elsewhere (Tauler, 1995), only a short description is included here. After estimating the number of components of the mixture using PCA, MCR-ALS uses the most significantly differing curves of the dataset as initial representatives of the co-existing species, and calculates the initial concentration profiles for the time points measured. Using an alternating least squares (ALS) algorithm, the solution of the species-pure profiles and relative populations is iteratively optimized. Multiple mathematical solutions can provide equivalent descriptions of the experimental data (ambiguity). Constraints based on either the physical nature of the investigated system or on prior knowledge can be applied to reduce the level of ambiguity of the solution. Here, mass conservation along the experiment (*closure*) and *non-negativity* values for both the resulting SAXS profiles and the time-dependent concentrations were introduced as constraints in the optimization procedure. Details on the implementation of these constraints can be found in Supplemental Experimental Procedures. However, MCR-ALS did not provide a good description of the data, as evidenced by an average χ_i^2 of the fit to the 28 data curves, $\langle\chi_i^2\rangle = 4.38$ (Table 1). In addition, MCR-ALS-derived SAXS curves had noticeable artifacts and large uncertainties that yielded non-physical pairwise distance distribution functions, $p(r)$ (Figures S4A and S4B). Concentration profiles also display non-coherent behavior with a higher initial concentration for the intermediate species compared with the native state (Figure S4C). The addition of a fourth species in the MCR-ALS analysis did not improve the quality of the fit ($\langle\chi_i^2\rangle = 5.08$) or the intelligibility of the results. The use of experimental errors to weight the agreement to the scattering intensities along the optimization process did not improve the quality of the resulting species-pure curves, which presented strong artifacts at low angles (Figure S7).

Most likely, the non-optimal solution is due to an ambiguity problem. To overcome this intrinsic limitation we have developed COSMiCS, which adapts MCR-ALS to the analysis of large SAXS datasets. COSMiCS simultaneously fits multiple representations of the same SAXS dataset, including, in addition to the absolute values ($I(s)$), the commonly used data representations introduced by Kratky ($I(s)*s^2$) (Glatter and Kratky, 1982), Porod ($I(s)*s^4$) (Porod, 1951), and Holtzer ($I(s)*s$) (Holtzer, 1955); the latter has been recently visited by Rambo and Tainer (2013). Although this does not enrich the information content in our input dataset, it emphasizes the structural changes at different time points

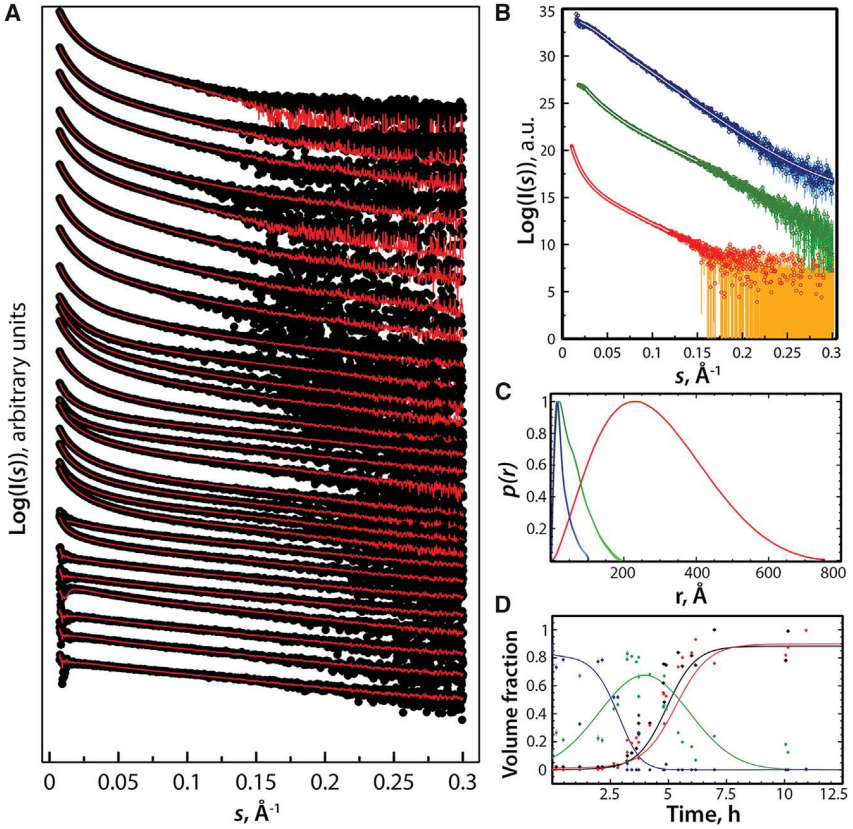


Figure 1. Analysis of the Fibrillation of Insulin

(A) SAXS profiles recorded during the evolution of fibrillation of insulin. Scattering intensity profiles (black dots) in logarithmic scale as a function of the momentum transfer ($s = 4\pi \sin(\theta)/\lambda$ [2θ, scattering angle; λ = 1.5 Å, X-ray wavelength]) measured from t = 0 (bottom curve) to t = 11 hr (top curve). COSMiCS fitted model combining Absolute and Holtzer (AH) data representations with three species are displayed as solid red lines. Curves are translated arbitrarily along the y axis for clarity.

(B) SAXS profiles in logarithmic scale from the decomposition of the insulin dataset using COSMiCS, displaying the monomer (blue), oligomer (green), and fibril (red) curves. Fits of the ab initio reconstructions are displayed as solid lines.

(C) Pairwise distribution functions derived from the individual curves for each species computed with the program GNOM (Svergun, 1992). The same color code as in (B) is used.

(D) Time-dependent concentration profiles derived from COSMiCS for each species with the same color code as in (B) (smoothing of the data in solid lines). The ThT fluorescence signal (black) is included to highlight its excellent correlation with the population of fibrils derived from the COSMiCS analysis.

along the fibrillation pathway. Species appearing along the fibrillation process present distinct structural features that emerge at specific momentum transfer ranges (resolution) and are captured differently by SAXS data representations (Figure S5). The consequent enhancement of data variability along the fibrillation process increases the discrimination power of the MCR-ALS optimization by reducing the ambiguity of the mathematical solutions. The simultaneous use of multiple SAXS data representations was tested first on simulated data (see Supplemental Experimental Procedures for details).

We used all combinations of dataset representations for COSMiCS analyses (Table 1). With the exception of Porod's representation, the inclusion of a second SAXS data representation yielded a systematic improvement in the quality of the fit to the 28 experimental profiles, while no improvement in the overall fit to the dataset was observed when increasing the number of data matrix representations to three. The best agreement to the dataset was obtained by combining the absolute value data representation with the Holtzer representation (AH; $\langle\chi_i^2\rangle = 3.15$). The fitting to the 28 curves (Figures 1A and S1A) demonstrates that the linear combination of three species and the population profile properly describes the complete experimental dataset with no systematic deviations along the time evolution. To confirm the PCA analysis we repeated the complete COSMiCS analysis using two and four species (Figures S4D–S4G). When fitting with two species the $\langle\chi_i^2\rangle$ is significantly increased to $\langle\chi_i^2\rangle = 7.66$, and the optimized spectra from the analysis with four species are not physically meaningful.

Structural Characterization of the Insulin Species and Their Kinetics

The pure SAXS curves of the insulin species and their relative population along the fibrillation process obtained from the COSMiCS analysis are shown in Figures 1B–1D. The capacity to derive species-specific information enables their detailed structural investigation (Table 2). An important piece of information that can be derived is the oligomeric state through the molecular weight (MW) estimation. We have applied several available strategies to derive these two parameters (Tables 2 and S1) (Mylonas and Svergun, 2007; Fischer et al., 2009; Petoukhov et al., 2012; Rambo and Tainer, 2013) providing coherent MW estimations despite their distinct approaches and inherent limitations. The use of an external standard depends on correct estimation of protein concentration and is very sensitive to the presence of small quantities of non-properly decomposed species, such as unspecific aggregates, while SAXS invariant volume of correlation, V_c, has not been calibrated for very large species (Rambo and Tainer, 2013). Consequently, we have chosen V_c for small species, and the external standard for large aggregates.

The $p(r)$ functions derived from the profiles indicate that the three species are a small particle, an intermediate oligomer, and a large aggregate (Figure 1C). This result is in very good agreement with previous manual decomposition applied to a similar fibrillation series of insulin (Vestergaard et al., 2007). The smallest species corresponds to a slightly elongated particle (Figure 2) with an R_g of 23.2 ± 0.2 Å, in full agreement with the need for a monomeric partially unfolded species to trigger insulin

Table 1. Fitting of the Insulin and α-SN$_{E46K}$ SAXS Datasets with COSMiCS Using Different Combinations of Data Matrices

Representations Included					Insulin[a]		α-SN$_{E46K}$[a]			
					Complete Dataset		Complete Dataset		Without Outlier[b]	
Code	Absolute $I(s)$	Holtzer $I(s)*s$	Kratky $I(s)*s^2$	Porod $I(s)*s^4$	$<\chi_i^2>$	χ_i^2 Range	$<\chi_i^2>$	χ_i^2 Range	$<\chi_i^2>$	χ_i^2 Range
A	+				4.38	1.57–11.85	2.62	9.46–0.54	2.38	8.96–0.73
AH	+	+			3.15[c]	1.37–12.90	1.22	3.42–0.54	1.08	2.32–0.55
AK	+		+		3.75	1.37–10.33	1.18	3.69–0.53	1.05	2.24–0.53
AP	+			+	5.61	1.49–17.02	1.33	4.82–0.57	1.18	3.05–0.58
AHK	+	+	+		3.72	1.37–13.07	1.16	3.55–0.56	1.04[c]	2.18–0.53
AHP	+	+		+	5.45	1.38–16.92	1.24	4.31–0.54	1.09	2.71–0.54
AKP	+		+	+	4.75	1.48–15.80	1.16[c]	3.98–0.53	1.05	2.08–0.53
AHKP	+	+	+	+	4.54	1.39–15.21	1.16	3.85–0.53	1.03	2.11–0.52

[a]Analysis performed using three species.
[b]COSMiCS analysis after extraction of the curve corresponding to t = 25.33 hr.
[c]Optimal solutions used for the structural analysis.

fibrillation (Ahmad et al., 2003). The partially unfolded nature of this species is also evident when plotting the isolated curve in the Kratky presentation and the low agreement with the crystallographic structure of the monomeric insulin (Figures S6A and S6B). The oligomer, which has an estimated mass of 4–8 protomers (Table 2), features an elongated shape with an R_g of 48.5 ± 0.1 Å and a D_{max} = 196 ± 10 Å according to the derived $p(r)$ (Figure 1C). The $p(r)$ function was used to derive a low-resolution structure of this elusive intermediate species. The resulting structure, which perfectly describes the experimental curve, shows that the oligomer is an elongated particle, with a bent/helical form (Figure 2) in excellent agreement with the structure derived by our group from a more concentrated fibrillation series (Vestergaard et al., 2007). The third species, which represents the repeating unit of insulin fibrils, is composed by ≈1730 insulin monomers with an R_g = 225.1 ± 0.7 Å and a D_{max} of 760 ± 10 Å. The low-resolution structure indicates that this fibril unit consists of several intertwining protofibrils, resulting in the relatively globular and flat appearance, in accordance with previous studies (Vestergaard et al., 2007).

Figure 1D displays the time-dependent concentration profiles of the three species. The derived concentration profiles clearly identify the population behavior of the three species, but display spikes along the fibrillation due to the stochasticity of the fibrillation process in the individual sample wells. This deleterious effect could be overcome using SAXS laboratory sources whereby a single sample could be measured along the whole fibrillation process. Monomeric insulin, which is the most populated species at the beginning of the reaction, was not present in significant amounts after 3 hr of incubation. The intermediate oligomer is present during almost the complete observation time and becomes the major species around 4 hr, indicating that it is a relatively stable species. This species can hence not be considered as a thermodynamic nucleus, which per definition is the least stable species along the reaction coordinate. Rather, the species must be described as a structural nucleus (Powers and Powers, 2008), and most likely is a building block of protofibrils (Vestergaard et al., 2007). We have superimposed the ThT fluorescence profile from the same samples, which is sensitive to the fibrillar forms of the protein, and serves as an independent measure of the presence of the amyloidogenic fibrils. The SAXS-derived volume fractions of the large fibrillar species follows a sigmoidal growth after the lag phase that is in excellent agreement with the ThT fluorescence profile (Figures 1D and S2B) and substantiates the results derived from the chemometric SAXS decomposition. Our analysis indicates that all samples measured, with the exception of the last point, contain at least two co-existing species. This observation highlights the importance of chemometric approaches whereby no a priori assumptions on the composition of the individual SAXS curves are made.

SAXS Measurements during the Fibrillation Process of α-Synuclein$_{E46K}$

The COSMiCS strategy was subsequently applied to data from the fibrillation of the α-SN$_{E46K}$, associated with early-onset Parkinson's (Fredenburg et al., 2007). Although this mutant has been widely investigated, its fibrillation process has never been studied at structural level by SAXS. Time-resolved SAXS data were obtained following the protocols previously described (Langkilde and Vestergaard, 2012). A total of 51 SAXS curves were measured during 25 hr, starting from the monomeric protein at 12 mg/mL. Two of these curves presented severe radiation damage and were discarded. The remaining curves display a distinct evolution in scattering intensity along the fibrillation process (Figures 3 and S1). The initial analysis of the raw data revealed, as expected, the formation of very large species during the fibrillation process. Average R_g and the $I(0)$ of the individual curves were estimated using Guinier's approximation (Figure S2C). Interestingly, the steep increase in molecular mass occurs significantly later than the increase in average radii and the initiation of the elongation phase as indicated by ThT fluorescence.

A PCA of the complete α-SN$_{E46K}$ SAXS dataset indicates that it can be safely described with three co-existing species (Figures S3B and S3D). The α-SN$_{E46K}$ dataset was subsequently successfully decomposed with COSMiCS using the combinations of SAXS data representations (Table 1). As occurred for the case of insulin, a systematic improvement in the quality of the fit is observed when multiple data representations are introduced, with $<\chi_i^2>$ decreasing from 2.62 to a range of 1.16–1.33. This systematic amelioration of the data description is also observed in the range of individual χ_i^2 obtained among the

Table 2. Structural Information on the Pure Species Derived from COSMiCS Analysis of Insulin and α-SN$_{E46K}$ SAXS Datasets

Species		SAXS Curve-Derived Parameters[a]			MW, kDa (Oligomeric State)[b]	
		R_g, Å	D_{max}, Å	$I(0)$, a.u.	BSA[c]	Scatter[d]
Insulin	1	23.2	103.4	0.11	21.8 (3.8)	8 (1.4)
	2	48.5	196.0	0.22	45.7 (7.9)	23 (4.0)
	3	225.1	760.0	62.24	10,036 (1,730.3)	30,600 (5,275.9)
α-SN$_{E46K}$	1	47.1	209.5	2.56	47.6 (3.3)	13 (0.9)
	2	281.8	960.0	52.72	594.3 (41.0)	630 (43.4)
	3	251.7	920.0	111.22	2,565.5 (176.9)	1,470 (101.4)

[a]Parameters derived with GNOM (Svergun, 1992).
[b]Computed from molecular weight (MW) of the monomers of insulin (5.8 kDa) and α-SN$_{E46K}$ (14.5 kDa).
[c]MW estimation using BSA as external standard.
[d]Using V_c with the program Scatter under standard parameters (Rambo and Tainer, 2013).

49 curves (Table 1). However, not all data combinations guide the decomposition equivalently. Again, the inclusion of Porod's representation (P) only modestly decreases the $<\chi_i^2>$ compared with inclusion of Holtzer's or Kratky's representations (AH or AK). Four of the solutions present very similar $<\chi_i^2>$, between 1.16 and 1.18. Although quantitatively equivalent, a closer inspection of the decomposed curves shows that representations AHK and AHKP provide solutions yielding non-physical $p(r)$ functions with negative values. Therefore, the AKP solution, with a $<\chi_i^2> = 1.16$, was used for the subsequent analyses.

The inspection of the level of agreement of the individual curves indicates that the vast majority of χ_i^2 values are around 1.0 (Figure 4A), indicating that our optimization protocol is not overfitting the data. However, curve 45 (measurement at 25.33 hr) is not properly described by the model and presents a relatively large χ_i^2 (3.98). In fact, this curve presents the largest χ_i^2 in all data combinations. After removing this curve the complete COSMiCS decomposition was repeated, yielding reduced $<\chi_i^2>$ (1.03–1.08) for various data combinations (Table 1). Importantly, the improvement observed after discarding a single curve does not simply originate from the elimination of an outlier but corresponds to a systematic improvement of the individual χ_i^2 for the vast majority of the curves of the dataset (Figure 4B). A further refinement by removing an additional potential outlier curve was tested but no systematic improvement in the fit was observed (Figure 4C). These observations underline the robustness of the final solution and the ability of COSMiCS to detect outlier SAXS curves. The final fits from the COSMiCS analysis using the AHK combination ($<\chi_i^2> = 1.04$) are displayed in Figure 3, and the solution was subsequently analyzed in terms of structures and kinetics.

Structure and Transformation Kinetics of the Species Involved in the Fibrillation of α-Synuclein E46K Mutant

As can be seen from the COSMiCS-derived scattering curves and the $p(r)$ functions (Figure 5), two of the species are very large while the first species is a low MW particle. We estimate an MW of the first species of 13 kDa, in agreement with a monomeric state of the protein (14.6 kDa). Additionally the Kratky representation (not shown) of this species, along with the skewed $p(r)$ function and the relatively large values for R_g (47.1 ± 0.72 Å) and D_{max} (209 ± 5 Å), show the disordered nature of α-SN$_{E46K}$ (Bernadó and Svergun, 2012). The disordered nature of the monomeric species was substantiated by the ensemble optimization method (EOM) analysis of the SAXS profile (Bernadó et al., 2007) (see Experimental Procedures for details). The subensemble of conformations that collectively describe the SAXS curve display a broad range of R_g values indicating the large degree of flexibility of the protein in solution (Figures S6C and S6D).

The second oligomeric species is large (≈40 protomers) with an R_g of 282.4 ± 4.3 Å and a D_{max} of 960 ± 10 Å (i.e., at the resolution limit of our measurements), approaching values obtained for the final fibril species (D_{max} = 920 ± 10 Å and R_g = 256.8 ± 1.9 Å). However, when comparing the mass and overall dimensions of these two species (Tables 2 and S1) the mass density of the intermediate species is much lower than that of the fibrils. This is in agreement with the observation that the average R_g of the mixture increases significantly earlier than the average mass (Figure S2C), and suggests that the intermediate species is a large and disordered oligomer. It is evident also from the $p(r)$ functions (Figure 5) that the two species present distinct overall shapes. While the intermediate species is represented as an overall globular shape, the final species presents the typical elongated fibrillar shape (Oliveira et al., 2009; Giehm et al., 2011b). The COSMiCS curves for the large oligomer and the fibril were used to derive low-resolution structures. Whereas the fibril repetitive unit is a large and elongated particle (Figure S6E), attempts to derive a structure for the oligomer were unsuccessful. Indeed this is in agreement with the disordered nature of this species that precludes the determination of its ab initio structure. The distinct structural characteristics of both forms are corroborated when observing their time-dependent evolution, which is plotted together with the ThT fluorescence curves in Figures 5C and S2D. Clearly the evolution of the ThT signal coincides with the occurrence of the fibril-shaped species, whereas the second large species is not ThT active. Both large species co-exist after the lag phase, but the decrease of the second species at the final steps of the aggregation suggests a transformation from the large disordered aggregates to amyloidogenic fibrils. Interestingly the monomeric form is present throughout the whole experiment, suggesting that the disordered aggregate has to disassemble into monomers before forming amyloidogenic fibrils.

DISCUSSION

The interest in structural studies of molecular conversions, functional and structural heterogeneity, and time-evolving processes

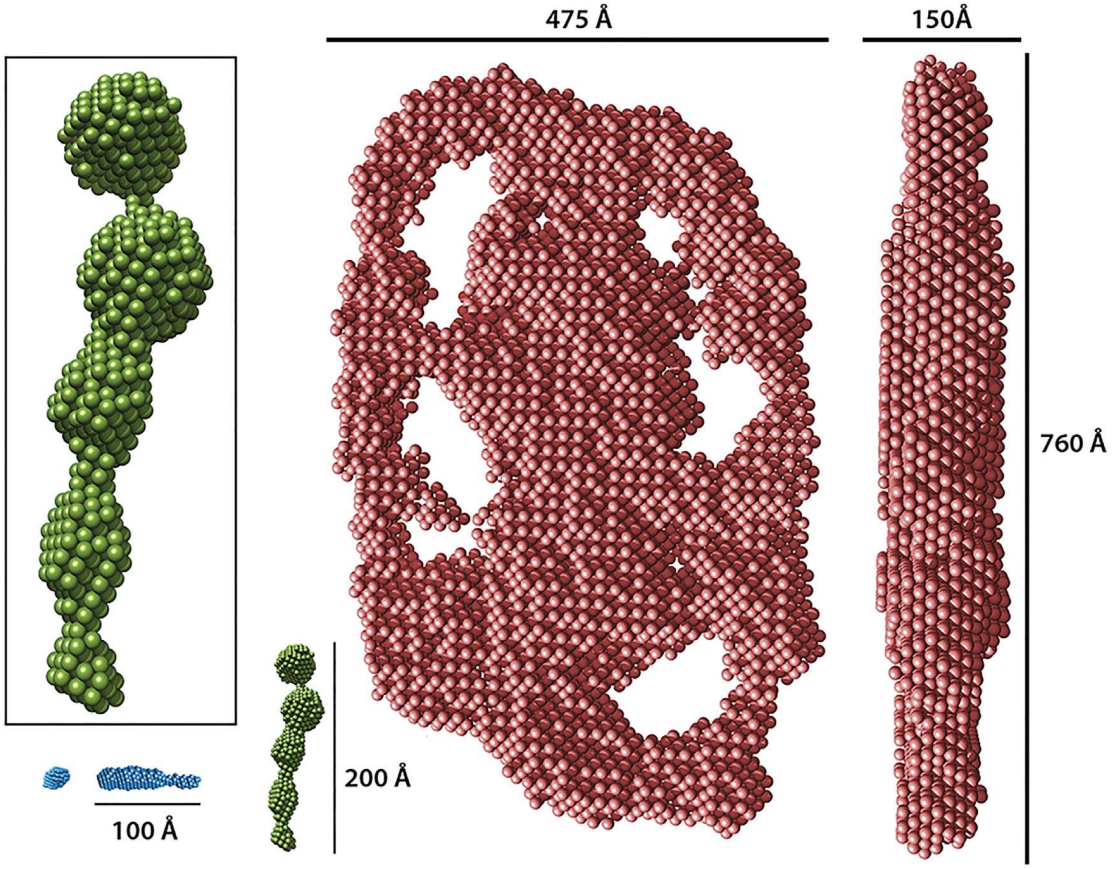


Figure 2. Structural Models of Insulin Fibrillar Species

Ab initio reconstructions of the three components obtained from the COSMiCS analysis of the SAXS data measured along the insulin fibrillation. Structures of the monomer (blue), oligomer (green), and the repeating unit of the fibril (red) are displayed in their relative sizes. The monomer is displayed in two orientations, one rotated 90°. The oligomer is displayed in more detail in the inset.

is significant but highly challenging. SAXS is an extremely well-suited technique to address the characterization of such complex mixed systems, but robust approaches are necessary to address the data decomposition process. Here, we present a method for the chemometric decomposition of multiple SAXS curves measured along a structurally developing process. This strategy, implemented as an extension of the popular MCR-ALS method (Tauler, 1995; Jaumot et al., 2005), has been applied to fibrillating proteins but could be extended to other macromolecular systems with tunable equilibrium phenomena such as the study of protein folding, transient interactions, intermediate structural states, viral capsid formation, and supramolecular assemblies.

Initial attempts to analyze SAXS data measured along insulin and α-SN_{E46K} fibrillation with MCR-ALS were unsuccessful, likely because the algorithm is trapped in a local minimum, yielding acceptable fits but non-physical solutions. The classical approach to resolve this ambiguity-related problem is to include orthogonal datasets (Navea et al., 2002, 2006). However, the simultaneous measurement of complementary data is not available in present SAXS beamlines. To overcome the ambiguity problem, we simultaneously analyze multiple representations of the same SAXS data. The empowerment of the decomposition when using multiple data representations reflects one of the central aspects of solution scattering data. The scattering curve arises from pairwise distances at both short and longer scales, thereby probing the shape of solutes, covering several orders of magnitude, from nanometer to micrometer sizes. In the case of fibril development this phenomenon is fully exploited as starting particles are of nanometer size, while fibrils are several micrometers in length. During the fibrillation process the structural features, which are coded in the scattering curves, dramatically change in a time-dependent manner. However, at different time periods of the fibrillation reaction these changes occur in different parts of the momentum transfer range measured, and can be highlighted depending on the representation of the data used.

The different sensitivity of SAXS data representations to structural features can be understood when considering their mathematical nature. Absolute scale curves have a large dynamic range and the most important intensity variations occur at the smallest angles, linked to the lowest-resolution structural information. In the other representations the intensity is multiplied by momentum transfer with increasing power functions: 1, 2, and 4 for Holtzer, Kratky, and Porod, respectively. This

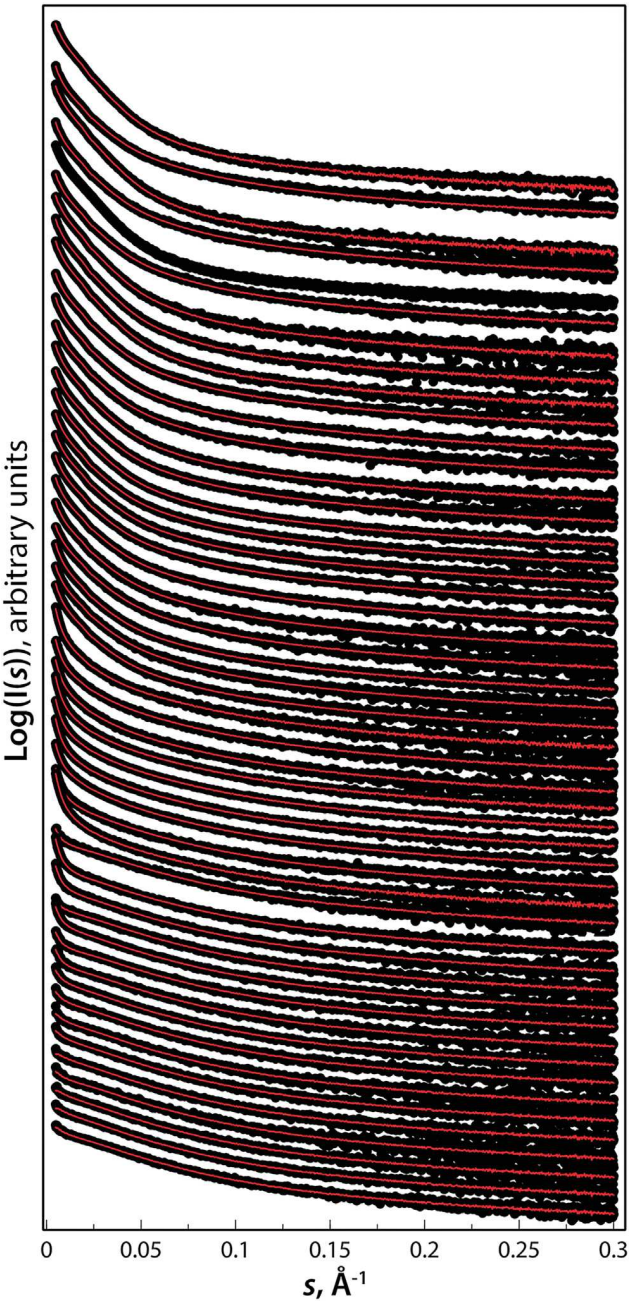


Figure 3. α-SN_{E46K} SAXS Data

SAXS profiles showing the evolution of fibrillation of α-SN_{E46K} (black dots) in logarithmic scale as a function of the momentum transfer ($s = 4\pi \sin(\theta)/\lambda$ [2θ, scattering angle; $\lambda = 1.24$ Å, X-ray wavelength]). Bottom curve corresponds to the first curve (time = 0 hr) and top curve corresponds to t = 25.6 hr. Red lines are the COSMiCS fits obtained from the three-component mixture using Absolute and Holtzer and Kratky representations (AHK) of the SAXS data. The curve corresponding to t = 25.33 hr has been identified as an outlier and is not used for the global fitting (and hence no COSMiCS fit is superposed). Curves are translated arbitrarily along the y axis for visualization purposes.

successively decreases the emphasis on intensities at low scattering angles (low resolution) and increases the emphasis at high scattering angles (high resolution). Therefore, the combination of the absolute scale, which enhances the low-resolution part of the SAXS curve, with the other representations, which enhance the high-resolution part of the SAXS profile, facilitates decomposition. Surprisingly, Porod's representation, in the examples tested, does not increase (or even decrease) the decomposition power of COSMiCS. Porod's law is valid for smooth interfaces between solutes and solvent. In the case of fibrillation, it is known that the fibril interface is disordered and poorly defined (Comellas et al., 2011), and presents high-entropy solvent around the interface (Nielsen et al., 2013; Fichou et al., 2015). This observation does not exclude, however, that Porod's representation could be an important one for the decomposition of other types of data.

A few examples from the present analysis are included here for clarity. The first example is the conversion of the intrinsically disordered protein, α-SN_{E46K}, into a large and more ordered intermediate that finally evolves toward elongated fibrils. The Kratky fingerprint of an intrinsically disordered protein is very characteristic with a lack of low-angle features, and a steadily rising profile at high angles (Bernadó and Svergun, 2012). These features contrast with those found in ordered systems where a pronounced peak is found at intermediate momentum transfer ranges. Therefore, the Kratky representation is extremely sensitive to the initial conversion from a disordered to a more ordered species. In contrast, the Holtzer representation is highly sensitive to overall changes in mass, which is very significant at later time points in the fibrillation process. Insulin, in contrast to α-SN_{E46K}, fibrillates from a (partially) folded-like species. Here, however, the subsequently formed species is elongated with a very distinct scattering curve that is easily discriminated from the starting species. The oligomeric intermediate is subsequently transformed into a much larger mature fibril. In this second transition, the difference in size, which induces a strong differentiation in the initial part of the SAXS curve and in the peak position in Kratky representation, facilitates their discrimination. Hence, simultaneous fitting of multiple data representations used in COSMiCS exploits complementary features that appear at different time points, enhancing the capacity of the chemometric approach to discriminate within the vast space of potential solutions.

It is worth noting that no specific shapes for the resolved scattering profiles were imposed as a constraint throughout the optimization procedure. We have observed that when multiple SAXS data representations were considered in the same analysis, the overall fit to the data is better and the possibility to obtain physically meaningful solutions increases. The presence of these local minima solutions with unreasonable SAXS profiles highlights the importance of identifying and discarding them. Here, all wrong solutions present strong artifacts and large uncertainties that identify them as unphysical. This observation points toward the need to find a proper mathematical description of SAXS curves to allow for the introduction of a new constraint to further decrease solution ambiguity.

Our approach represents a great advantage compared with other techniques that probe a single species. This is exemplified for the case of α-SN_{E46K}. In the decomposition of the SAXS data of this familial mutant, we reveal that two very large aggregated species co-exist for a long period of time. One of these species is ThT inactive and would remain invisible when using traditional fluorescence experiments, thereby dramatically biasing the

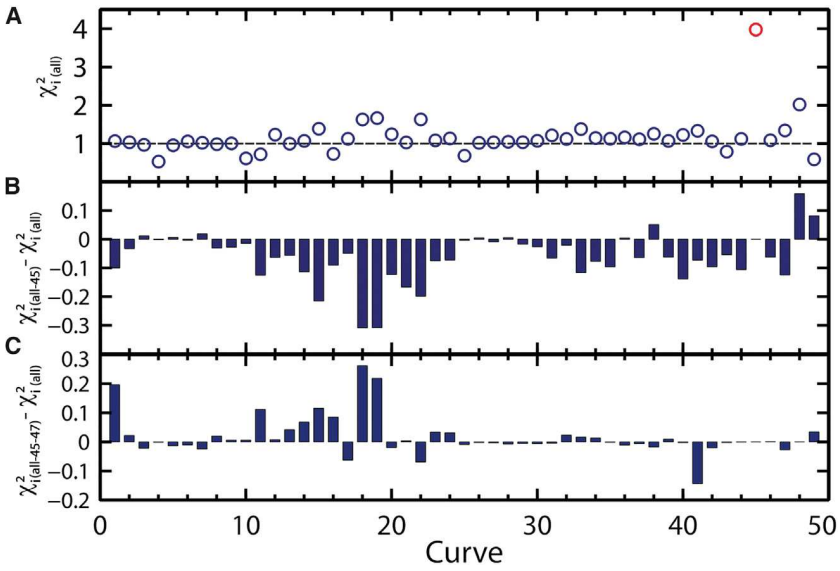


Figure 4. Assessment of Outlier Curves in the α-SN$_{E46K}$ SAXS Dataset

(A) Individual χ_i^2 values obtained from the COSMiCS analysis of the α-SN$_{E46K}$ complete dataset using the AKP combination of matrices. One single value appears as a potential outlier (curve 45, marked in red).

(B) The variation in the χ_i^2 values of the individual curves when extracting curve 45 (corresponding to a time of 25.33 hr) from the analysis using the AHK combination. A systematic improvement for the vast majority of curves of the dataset is observed upon extraction of the curve from the global fitting, and it is hence concluded that curve 45 is an outlier.

(C) Difference in individual χ_i^2 values derived from the COSMiCS analysis with the AHK combination using 48 curves of the α-SN$_{E46K}$ dataset and the AKP combination using 47 curves in the dataset after removing the SAXS curve corresponding to t = 25.6 hr from the analysis. Either no effect or an increase in the χ_i^2 values is observed. It can be concluded that the extraction of this second curve does not improve the derived model and it is therefore not justifiable.

interpretation of a protein fibrillation study solely based on ThT fluorescence monitoring (Cohen et al., 2012). Importantly, by applying our approach we uncover the existence of an intermediate species of a structural nature that has hitherto not been described in the context of fibrillation kinetics. This oligomer has a low density of mass, is predominantly disordered, and is of a reversible nature.

Here, we have demonstrated the power of SAXS data analysis to understand the nature of a very complex biological process such as protein amyloid formation. The general application of COSMiCS, however, spans much wider, to principally any phenomenon where multiple species or conformers co-exist in equilibrium and whose relative population can be changed by rationally perturbing experimental conditions such as temperature, pressure, pH, ionic strength, and presence of ligands or chemical denaturants. Protein and RNA folding processes, enzymatic reactions, transient biomolecular interactions, or the formation of large supramolecular assemblies are relevant examples of such complex equilibria. Many of these systems can be monitored by SAXS; therefore, the demonstrated decomposition power of the COSMiCS approach can be applied to resolve their intrinsic heterogeneity to characterize their structure and kinetics/thermodynamics.

EXPERIMENTAL PROCEDURES

Insulin Sample Preparation and Fluorescence Measurements

Human zinc insulin was obtained from Novo Nordisk. Zinc content was 0.37% (w/w) corresponding to approximately two Zn^{2+} ions per insulin hexamer. ThT was purchased as the chloride salt from Sigma-Aldrich. ThT was recrystallized three times in demineralized water before use. All other chemicals were of analytical grade. For ThT fluorescence assays, a Polarstar Optima platereader from BMG Labtechnologies was used with 96-well, black, polystyrene, non-sterile plates with optical bottom from Nalge Nunc International. The wells were covered with Polyolefin non-sterile sealing tape from Nalge Nunc International, and bottom/bottom measurements were performed. ThT fluorescence measurements were performed with λ_{ex} = 440 nm (10 nm bandpass) and λ_{em} = 480 nm (12 nm bandpass). A pellet of insulin was dissolved at 2.5 mg/mL insulin in 20% acetic acid (pH 2.0) with 0.1 M NaCl and 20 μM ThT. Afterward 100 μL of the solution was transferred to each well. The plate was placed in the platereader and ThT fluorescence measurements were conducted at 45°C without shaking. The measurement of the fluorescence intensity was performed every 300 s. The platereader was paused at appropriate time intervals and a sample of 80 μL was withdrawn from a well.

α-SN$_{E46K}$ Sample Preparation

α-SN$_{E46K}$ was produced in *Escherichia coli* BL21 using a pET-11a vector, and expressed and purified as described previously. Lyophilized α-SN was dissolved in 20 mM PBS buffer with 150 mM NaCl (pH 7.4). After filtration through 0.22-μm spin filters (Millipore) the concentration was determined by $A_{280\ nm}$ using a Nanodrop UV-Vis spectrophotometer (Thermo Scientific) with an extinction coefficient of 5,120 M^{-1} cm^{-1}. Solutions were prepared containing 12 mg/mL α-SN$_{E46K}$ and 20 μM ThT. Fibrillation of 150-μL aliquots was induced in 96-well optical bottom plates (Thermo Scientific) using a Fluostar Optima Platereader (BMG Labtech) under heating (37°C) and orbital shaking with 3-mm glass beads (Giehm et al., 2011a).

SAXS Data Collection and Primary Data Evaluation

Three aliquots of protein were extracted from each well and SAXS data were recorded immediately after extracting the sample. Scattering data of α-SN$_{E46K}$ were recorded at the P12 beamline at European Molecular Biology Laboratory (EMBL) on the Petra III storage ring (Deutsches Electronen Synchrotron [DESY]) using the automated loading system (Round et al., 2008). The data were collected on a PILATUS detector with a momentum transfer range of $0.0049 < s < 0.35$ $Å^{-1}$ by 20 individual exposures of 45 ms each. The data reduction and buffer subtraction was performed by the beamline automated procedure (Franke et al., 2012) followed by a subsequent manual control. After close inspection of the multiple frames recorded at each time point, two curves were discarded, which exhibited unsystematic features. The insulin samples were collected on the X33 beamline at the EMBL on DORIS III (DESY) at a wavelength of 1.5 Å, using a MAR345 Image Plate Detector, in the momentum transfer range $0.006 < s < 0.51$ $Å^{-1}$ with 2-min exposure time. No radiation damage was detected when performing repeated exposures. Zinc acetate was added to the background buffers corresponding to two Zn^{2+} ions per insulin hexamer in the protein sample, and buffer measurements were performed immediately before and after each protein sample measurement. An averaged buffer measurement used for background subtraction. When the previously reported buffer effect, typical for fibril scattering data, was observed, background correction was applied as previously reported (Nielsen et al., 2013).

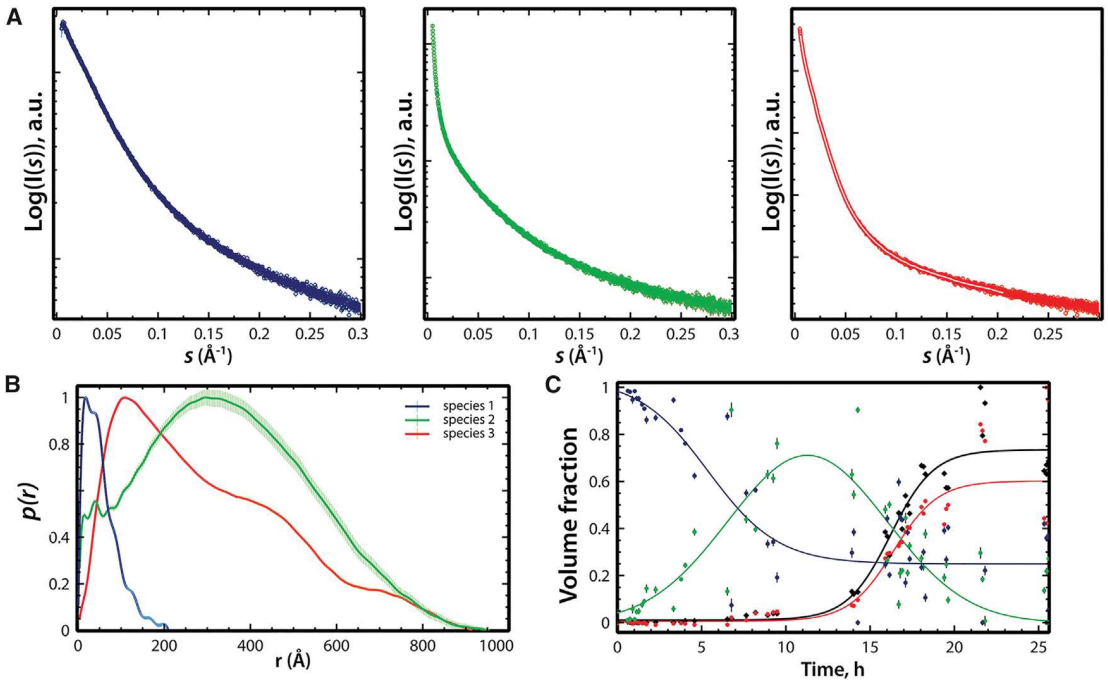


Figure 5. COSMiCS Analysis of α-SN$_{E46K}$ Fibrillation

Results from the decomposition of the α-SN$_{E46K}$ data with COSMiCS using the AHK combination of matrices.

(A) Decomposed SAXS profiles for the monomer (blue), oligomer (green), and fibril (red) species.

(B) $p(r)$ functions of the three species computed from the SAXS profiles of (A) using GNOM (Svergun, 1992).

(C) Concentration profiles with the same color code as before and with the ThT fluorescence signal superimposed (solid black line). The ThT profile is in excellent agreement with the population of fibrils.

Data analysis was performed using the software suite ATSAS (Petoukhov et al., 2012), and molecular masses were estimated relative to that of a standard reference solution of BSA. Guinier's approximation was applied to provide rough estimates of the extrapolated forward scattering ($I(0)$) and radii of gyration (R_g) for the evolving samples. The $p(r)$ functions were evaluated by the program GNOM (Svergun, 1992), providing the maximal dimension (D_{max}) within the particle, and a second estimate of $I(0)$ and R_g values.

The EOM was applied to structurally describe the COSMiCS-derived curve of monomeric α-SN$_{E46K}$ (Bernadó et al., 2007). A pool of 4,000 conformations of α-SN$_{E46K}$ was built with Flexible-Meccano (Bernadó et al., 2005). After addition of side chains with SCWRL4.0 (Krivov et al., 2009), the individual theoretical SAXS profiles were computed with CRYSOL (Svergun et al., 1995), and were used to select a subensemble of conformations that collectively described the experimental curve.

COSMiCS Analysis of Insulin Dataset

All SAXS representations used for the MCR-ALS optimization were used in a momentum transfer range of $0.0074 < s < 0.3$ Å^{-1}. Scattering curves numbers 3 (28 min), 4 (1 hr 11 min), and 28 (10 hr 56 min) were selected as starting points for the optimization. The number of maximum iterations was set to 50 and the convergence criteria were set to 0.1. ALS optimization is performed under the standard constraints for fibrillating system, *non-negativity* (both for spectra and concentration profiles) and *closure* (de Juan and Tauler, 2003) for concentration profiles.

COSMiCS Analysis of α-Synuclein$_{E46K}$ Dataset

The momentum transfer ranges used for the MCR-ALS analysis were $0.0074 < s < 0.3$ Å^{-1} for the absolute values, $0.0074 < s < 0.16$ Å^{-1} for Holtzer and Kratky representations, and $0.0074 < s < 0.07$ Å^{-1} for Porod's representation. The scattering curves selected as initial estimations were the curves 2 (37 min), 17 (6 hr 46 min), and 47 (16 hr 1 min). The maximum number of iterations, the convergence criteria, and the constraints were equivalent to those used in the insulin case. A Monte Carlo approach similar to the previously used by Svergun and Pedersen, (1994) was applied to estimate the standard deviations of the scattering intensities and the populations of the final solutions of the COSMiCS analyses of insulin and α-SN$_{E46K}$ datasets (see Supplemental Experimental Procedures).

Ab Initio Modeling

Ab Initio structures of the oligomer and fibril of insulin were obtained using the program DAMMIN (Svergun, 1999). The program employs a simulated annealing protocol to search for a complex bead model minimizing the discrepancy between the experimental and calculated curves at low resolution (up to s of about 0.15 Å^{-1}). The search volume, evaluated with the program BODIES (Petoukhov et al., 2012), for the fibril repeat was an ellipsoid with half-axes of 800, 500, and 150 Å using 31,985 spheres. Individual jobs were loaded, and 20 independent models were averaged using the program DAMAVER (Volkov and Svergun, 2003) and filtered with DAMFILT (Petoukhov et al., 2012). The oligomer was calculated inside a sphere with a diameter of 98 Å, obtaining a final averaged and filtered structure from 18 individual models. The structure of the monomer was obtained using DAMMIF (Franke and Svergun, 2009), starting from 20 arbitrary initial models to obtain the final model. Likewise, the structure of the repeating unit of α-SN$_{E46K}$ fibril was calculated with DAMMIN using as starting point an ellipsoid of 200, 500, and 100 Å in 20 individual runs that were averaged and filtered. The structures were rendered with the program CHIMERA (Pettersen et al., 2004).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, and one table and can be found with this article online at http://dx.doi.org/10.1016/j.str.2016.10.013.

AUTHOR CONTRIBUTIONS

F.H.-T., B.V., and P.B. designed research. F.H.-T., M.G., and A.v.M. conducted the experiments. R.T. contributed analytical tools. F.H.-T., B.V., and P.B. wrote the paper.

ACKNOWLEDGMENTS

This work was supported by SPIN-HD – Chaires d'Excellence 2011 from the Agence National de Recherche, ATIP-Avenir, SUDOE NeuroMed, and the French Infrastructure for Integrated Structural Biology (FRISBI – ANR-10-INSB-05-01) to P.B. We are grateful for funding from the Danish Council for Independent Research, Medical Sciences, postdoctoral grant (M.G.) and the Sapere Aude program SAFIR (0602-01377B; B.V., F.H.-T., A.v.M.) as well as support from INSERM (F.H.-T.). Assistance at the X33 and P12 EMBL beamlines (DESY, Hamburg, Germany) is greatly acknowledged, and the authors (F.H.-T., M.G., A.v.M, B.V.) are grateful for financial support from DANSCATT, covering traveling expenses for the SAXS data collection.

Received: April 4, 2016
Revised: August 23, 2016
Accepted: October 25, 2016
Published: November 23, 2016

REFERENCES

Ahmad, A., Millett, I.S., Doniach, S., Uversky, V.N., and Fink, A.L. (2003). Partially folded intermediates in insulin fibrillation. Biochemistry *42*, 11404–11416.

Akiyama, S., Takahashi, S., Kimura, T., Ishimori, K., Morishima, I., Nishikawa, Y., and Fujisawa, T. (2002). Conformational landscape of cytochrome c folding studied by microsecond-resolved small-angle x-ray scattering. Proc. Natl. Acad. Sci. USA *99*, 1329–1334.

Ayuso-Tejedor, S., García-Fandiño, R., Orozco, M., Sancho, J., and Bernadó, P. (2011). Structural analysis of an equilibrium folding intermediate in the apoflavodoxin native ensemble by small-angle X-ray scattering. J. Mol. Biol. *406*, 604–619.

Bernadó, P., and Svergun, D.I. (2012). Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. Mol. Biosyst. *8*, 151–167.

Bernadó, P., Blanchard, L., Timmins, P., Marion, D., Ruigrok, R.W.H., and Blackledge, M. (2005). A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. Proc. Natl. Acad. Sci. USA *102*, 17002–17007.

Bernadó, P., Mylonas, E., Petoukhov, M.V., Blackledge, M., and Svergun, D.I. (2007). Structural characterization of flexible proteins using small-angle X-ray scattering. J. Am. Chem. Soc. *129*, 5656–5664.

Blobel, J., Bernadó, P., Svergun, D.I., Tauler, R., and Pons, M. (2009). Low-resolution structures of transient protein-protein complexes using small-angle X-ray scattering. J. Am. Chem. Soc. *131*, 4378–4386.

Bucciantini, M., Giannoni, E., Chiti, F., Baroni, F., Formigli, L., Zurdo, J., Taddei, N., Ramponi, G., Dobson, C.M., and Stefani, M. (2002). Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. Nature *416*, 507–511.

Chen, L., Hodgson, K.O., and Doniach, S. (1996). A lysozyme folding intermediate revealed by solution X-ray scattering. J. Mol. Biol. *261*, 658–671.

Chiti, F., and Dobson, C.M. (2006). Protein misfolding, functional amyloid, and human disease. Annu. Rev. Biochem. *75*, 333–366.

Cohen, S.I.A., Vendruscolo, M., Dobson, C.M., and Knowles, T.P.J. (2012). From macroscopic measurements to microscopic mechanisms of protein aggregation. J. Mol. Biol. *421*, 160–171.

Comellas, G., Lemkau, L.R., Nieuwkoop, A.J., Kloepper, K.D., Ladror, D.T., Ebisu, R., Woods, W.S., Lipton, A.S., George, J.M., and Rienstra, C.M. (2011). Structured regions of α-synuclein fibrils include the early-onset Parkinson's disease mutation sites. J. Mol. Biol. *411*, 881–895.

de Juan, A., and Tauler, R. (2003). Chemometrics applied to unravel multicomponent processes and mixtures. Anal. Chim. Acta *500*, 195–210.

Fichou, Y., Schirò, G., Gallat, F.-X., Laguri, C., Moulin, M., Combet, J., Zamponi, M., Härtlein, M., Picart, C., Mossou, E., et al. (2015). Hydration water mobility is enhanced around tau amyloid fibers. Proc. Natl. Acad. Sci. USA *112*, 6365–6370.

Fischer, H., de Oliveira Neto, M., Napolitano, H.B., Polikarpov, I., and Craievich, A.F. (2009). Determination of the molecular weight of proteins in solution from a single small-angle X-ray scattering measurement on a relative scale. J. Appl. Crystallogr. *43*, 101–109.

Franke, D., and Svergun, D.I. (2009). DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. J. Appl. Crystallogr. *42*, 342–346.

Franke, D., Kikhney, A.G., and Svergun, D.I. (2012). Automated acquisition and analysis of small angle X-ray scattering data. Nucl. Instrum. Methods Phys. Res. A *689*, 52–59.

Fredenburg, R.A., Rospigliosi, C., Meray, R.K., Kessler, J.C., Lashuel, H.A., Eliezer, D., and Lansbury, P.T. (2007). The impact of the E46K mutation on the properties of alpha-synuclein in its monomeric and oligomeric states. Biochemistry *46*, 7107–7118.

Giehm, L., Lorenzen, N., and Otzen, D.E. (2011a). Assays for α-synuclein aggregation. Methods *53*, 295–305.

Giehm, L., Svergun, D.I., Otzen, D.E., and Vestergaard, B. (2011b). Low-resolution structure of a vesicle disrupting alpha-synuclein oligomer that accumulates during fibrillation. Proc. Natl. Acad. Sci. USA *108*, 3246–3251.

Glatter, O., and Kratky, O. (1982). Small Angle X-ray Scattering (Academic Press Inc).

Holtzer, A. (1955). Interpretation of the angular distribution of the light scattered by a polydisperse system of rods. J. Polym. Sci. *17*, 432–434.

Jaumot, J., Gargallo, R., de Juan, A., and Tauler, R. (2005). A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB. Chemom. Intell. Lab. Syst. *76*, 101–110.

Krivov, G.G., Shapovalov, M.V., and Dunbrack, R.L. (2009). Improved prediction of protein side-chain conformations with SCWRL4. Proteins *77*, 778–795.

Langkilde, A.E., and Vestergaard, B. (2012). Structural characterization of prefibrillar intermediates and amyloid fibrils by small-angle x-ray scattering. In Amyloid Proteins: Methods and Protocols, E.M. Sigurdsson, M. Calero, and M. Gasset, eds. (Humana Press), pp. 137–155.

Minh, D.D.L., and Makowski, L. (2013). Wide-angle X-ray solution scattering for protein-ligand binding: multivariate curve resolution with Bayesian confidence intervals. Biophys. J. *104*, 873–883.

Moreno-Gonzalez, I., and Soto, C. (2011). Misfolded protein aggregates: mechanisms, structures and potential for disease transmission. Semin. Cell Dev. Biol. *22*, 482–487.

Mylonas, E., and Svergun, D.I. (2007). Accuracy of molecular mass determination of proteins in solution by small-angle X-ray scattering. J. Appl. Crystallogr. *40*, 245–249.

Navea, S., de Juan, A., and Tauler, R. (2002). Detection and resolution of intermediate species in protein folding processes using fluorescence and circular dichroism spectroscopies and multivariate curve resolution. Anal. Chem. *74*, 6031–6039.

Navea, S., Tauler, R., and de Juan, A. (2006). Monitoring and modeling of protein processes using mass spectrometry, circular dichroism, and multivariate curve resolution methods. Anal. Chem. *78*, 4768–4778.

Nielsen, S.B., Macchi, F., Raccosta, S., Langkilde, A.E., Giehm, L., Kyrsting, A., Svane, A.S.P., Manno, M., Christiansen, G., Nielsen, N.C., et al. (2013). Wildtype and A30P mutant alpha-synuclein form different fibril structures. PLoS One *8*, e67713.

Nors Pedersen, M., Foderà, V., Horvath, I., van Maarschalkerweerd, A., Nørgaard Toft, K., Weise, C., Almqvist, F., Wolf-Watz, M., Wittung-Stafshede, P., and Vestergaard, B. (2015). Direct correlation between ligand-induced α-synuclein oligomers and amyloid-like fibril growth. Sci. Rep. *5*, 10422.

Oliveira, C.L.P., Behrens, M.A., Pedersen, J.S., Erlacher, K., Otzen, D.E., and Pedersen, J.S. (2009). A SAXS study of glucagon fibrillation. J. Mol. Biol. *387*, 147–161.

Petoukhov, M.V., Franke, D., Shkumatov, A.V., Tria, G., Kikhney, A.G., Gajda, M., Gorba, C., Mertens, H.D.T., Konarev, P.V., and Svergun, D.I. (2012). New developments in the ATSAS program package for small-angle scattering data analysis. J. Appl. Crystallogr. *45*, 342–350.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. J. Comput. Chem. *25*, 1605–1612.

Porod, G. (1951). Die Röntgenkleinwinkelstreuung von dichtgepackten kolloiden systemen—I. Teil. Kolloid-Z. *124*, 83–114.

Powers, E.T., and Powers, D.L. (2008). Mechanisms of protein fibril formation: nucleated polymerization with competing off-pathway aggregation. Biophys. J. *94*, 379–391.

Pratt, A.J., Shin, D.S., Merz, G.E., Rambo, R.P., Lancaster, W.A., Dyer, K.N., Borbat, P.P., Poole, F.L., Adams, M.W.W., Freed, J.H., et al. (2014). Aggregation propensities of superoxide dismutase G93 hotspot mutants mirror ALS clinical phenotypes. Proc. Natl. Acad. Sci. USA *111*, E4568–E4576.

Rajkó, R. (2010). Additional knowledge for determining and interpreting feasible band boundaries in self-modeling/multivariate curve resolution of two-component systems. Anal. Chim. Acta *661*, 129–132.

Rambo, R.P., and Tainer, J.A. (2013). Accurate assessment of mass, models and resolution by small-angle scattering. Nature *496*, 477–481.

Round, A.R., Franke, D., Moritz, S., Huchler, R., Fritsche, M., Malthan, D., Klaering, R., Svergun, D.I., and Roessle, M. (2008). Automated sample-changing robot for solution scattering experiments at the EMBL Hamburg SAXS station X33. J. Appl. Crystallogr. *41*, 913–917.

Sebollela, A., Mustata, G.-M., Luo, K., Velasco, P.T., Viola, K.L., Cline, E.N., Shekhawat, G.S., Wilcox, K.C., Dravid, V.P., and Klein, W.L. (2014). Elucidating molecular mass and shape of a neurotoxic Aβ oligomer. ACS Chem. Neurosci. *5*, 1238–1245.

Segel, D.J., Fink, A.L., Hodgson, K.O., and Doniach, S. (1998). Protein denaturation: a small-angle X-ray scattering study of the ensemble of unfolded states of cytochrome c. Biochemistry *37*, 12443–12451.

Svergun, D.I. (1992). Determination of the regularization parameter in indirect-transform methods using perceptual criteria. J. Appl. Crystallogr. *25*, 495–503.

Svergun, D.I. (1999). Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. Biophys. J. *76*, 2879–2886.

Svergun, D.I., and Pedersen, J.S. (1994). Propagating errors in small-angle scattering data treatment. J. Appl. Crystallogr. *27*, 241–248.

Svergun, D.I., Barberato, C., and Koch, M.H.J. (1995). CRYSOL—a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. J. Appl. Crystallogr. *28*, 768–773.

Tauler, R. (1995). Multivariate curve resolution applied to second order data. Chemom. Intell. Lab. Syst. *30*, 133–146.

Tauler, R. (2001). Calculation of maximum and minimum band boundaries of feasible solutions for species profiles obtained by multivariate curve resolution. J. Chemom. *15*, 627–646.

Tauler, R., Smilde, A.K., and Kowalski, B. (1995). Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. J. Chemom. *9*, 31–58.

Tresset, G., Le Coeur, C., Bryche, J.-F., Tatou, M., Zeghal, M., Charpilienne, A., Poncet, D., Constantin, D., and Bressanelli, S. (2013). Norovirus capsid proteins self-assemble through biphasic kinetics via long-lived stave-like intermediates. J. Am. Chem. Soc. *135*, 15373–15381.

Trygg, J., Holmes, E., and Lundstedt, T. (2007). Chemometrics in metabonomics. J. Proteome Res. *6*, 469–479.

Vestergaard, B., Groenning, M., Roessle, M., Kastrup, J.S., van de Weert, M., Flink, J.M., Frokjaer, S., Gajhede, M., and Svergun, D.I. (2007). A helical structural nucleus is the primary elongating unit of insulin amyloid fibrils. PLoS Biol. *5*, 1089–1097.

Volkov, V.V., and Svergun, D.I. (2003). Uniqueness of ab initio shape determination in small-angle scattering. J. Appl. Crystallogr. *36*, 860–864.

Williamson, T.E., Craig, B.A., Kondrashkina, E., Bailey-Kellogg, C., and Friedman, A.M. (2008). Analysis of self-associating proteins by singular value decomposition of solution scattering data. Biophys. J. *94*, 4906–4923.

Winner, B., Jappelli, R., Maji, S.K., Desplats, P.A., Boyer, L., Aigner, S., Hetzer, C., Loher, T., Vilar, M., Campioni, S., et al. (2011). In vivo demonstration that alpha-synuclein oligomers are toxic. Proc. Natl. Acad. Sci. USA *108*, 4194–4199.

Paper II: Active-Site-Directed Inhibitors of Prolyl Oligopeptidase Abolish Its Conformational Dynamics

ChemPubSoc Europe

DOI: 10.1002/cbic.201600102

CHEMBIOCHEM Communications

Active-Site-Directed Inhibitors of Prolyl Oligopeptidase Abolish Its Conformational Dynamics

Abraham López,[a, b] Fátima Herranz-Trillo,[c] Martin Kotev,[d] Margarida Gairí,[e] Víctor Guallar,[d, f] Pau Bernadó,[c] Oscar Millet,[g] Teresa Tarragó,[a, h] and Ernest Giralt*[a, b]

Deciphering conformational dynamics is crucial for understanding the biological functions of proteins and for designing compounds targeting them. In particular, providing an accurate description of microsecond–millisecond motions opens the opportunity for regulating protein–protein interactions (PPIs) by modulating the dynamics of one interacting partner. Here we analyzed the conformational dynamics of prolyl oligopeptidase (POP) and the effects of active-site-directed inhibitors on the dynamics. We used an integrated structural biology approach based on NMR spectroscopy and SAXS experiments complemented by MD simulations. We found that POP is in a slow equilibrium in solution between open and closed conformations, and that inhibitors effectively abolished this equilibrium by stabilizing the enzyme in the closed conformation.

Dynamics is essential for the biological functions of proteins.[1] Therefore, characterization of protein motions is fundamental for developing therapeutic compounds to modulate the conformational dynamics of their targets. Given the emerging therapeutic focus on protein–protein interactions (PPIs),[2] modulating microsecond–millisecond dynamics provides a valuable approach for regulating the affinity and specificity of recognition between flexible proteins.[3] Hence, the design of compounds that modify conformational dynamics stands as a promising strategy for controlling PPI networks involved in pathogenic mechanisms.

Prolyl oligopeptidase (POP) is an 81-kDa monomeric serine peptidase that hydrolyzes short peptides at the carboxyl side of proline.[4] POP has two domains, the α/β-hydrolase and the β-propeller, that are linked by a pair of hinge polypeptide chains. The X-ray structure of the mammalian enzyme shows the α/β-hydrolase and β-propeller domains packed together in a closed conformation (Figure 1 A).[5] However, the crystal structures of two bacterial POPs show a large hinge separation between domains[6] (Figure 1 B), thus suggesting that the enzyme undergoes interdomain flexibility.[7] Two studies, one based on ^{15}N line broadening NMR experiments[8] and the other on X-ray crystallography combined with MD simulations,[9] strongly support that POP is a highly flexible enzyme, but several fundamental aspects concerning the conformational landscape of POP in solution and the effects of inhibitors are largely unknown.

The in vivo role of POP is related to synaptic functions and neuronal development. It has been discovered that POP interacts with the intrinsically disordered proteins α-synuclein and GAP-43.[10] Recent studies have demonstrated that the direct

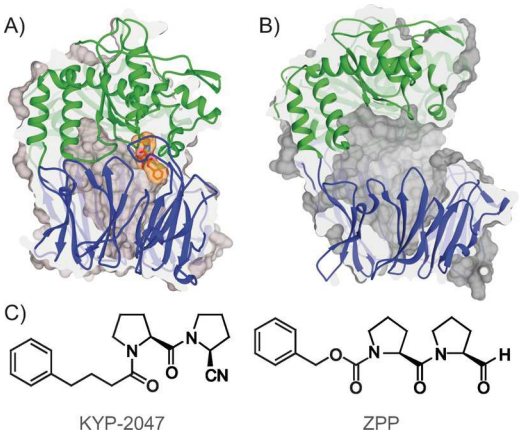


Figure 1. A) Porcine POP (PDB ID: 1QFS)[5] in the closed conformation covalently bound to the active-site-directed inhibitor ZPP (orange). The α/β-hydrolase domain is shown in green and the β-propeller is in blue. B) *Aeromonas punctata* POP in the open conformation (PDB ID: 3IUJ).[6a] C) Inhibitors KYP-2047 and ZPP.

[a] *Dr. A. López, Dr. T. Tarragó, Prof. E. Giralt*
Chemistry and Molecular Pharmacology Program
Institute for Research in Biomedicine
The Barcelona Institute of Science and Technology
Baldiri Reixac 10, 08028 Barcelona (Spain)
E-mail: ernest.giralt@irbbarcelona.org

[b] *Dr. A. López, Prof. E. Giralt*
Department of Organic Chemistry, University of Barcelona
Martí i Franquès, 1-11, 08028 Barcelona (Spain)

[c] *F. Herranz-Trillo, Dr. P. Bernadó*
Centre de Biochimie Structurale
INSERM U1054, CNRS UMR 5048, Université de Montpellier 1 and 2
29 rue de Navacelles, 34090 Montpellier (France)

[d] *Dr. M. Kotev, Dr. V. Guallar*
Joint BSC-CRG-IRB Research Program in Computational Biology
Barcelona Supercomputing Center
Jordi Girona 31, 08034 Barcelona (Spain)

[e] *Dr. M. Gairí*
NMR Facility
Scientific and Technological Centers University of Barcelona (CCiTUB)
Baldiri Reixac 10, 08028 Barcelona (Spain)

[f] *Dr. V. Guallar*
Institució Catalana de Recerca i Estudis Avançats (ICREA)
Passeig Lluís Companys 23, 08010 Barcelona (Spain)

[g] *Dr. O. Millet*
Structural Biology Unit, CIC bioGUNE
Parque Tecnológico de Vizcaya, Ed. 800, 48160 Derio (Spain)

[h] *Dr. T. Tarragó*
Iproteos, S L, Barcelona Science Park
Baldiri Reixac 10, 08028 Barcelona (Spain)

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/cbic.201600102.

interaction between POP and α-synuclein, the protein implicated in the development of Parkinson's disease, accelerates the aggregation of α-synuclein in vitro and in cells.[11] Interestingly, KYP-2047,[12] a covalent active-site-directed inhibitor of POP (Figure 1C) effectively reduced aggregation in vitro and in vivo.[11a,13] Further experiments have shown that this reduction is a consequence of increased clearance in vivo of aggregated forms of α-synuclein as a consequence of POP inhibition.[13b]

Nevertheless, the lack of knowledge on the conformational dynamics of POP and the effects of inhibitors represents a major impediment for exploring the mechanisms underlying POP-mediated aggregation of α-synuclein. Here, we analysed in detail the conformational equilibrium of POP in solution and how this is affected by the binding of active-site-directed inhibitors. We combined NMR spectroscopy to describe dynamic events at atomic resolution[14] with SAXS experiments complemented with MD simulations to probe large-scale structural fluctuations in solution.[15]

The conformational dynamics of POP in the microsecond–millisecond timescale was analyzed with methyl-TROSY $^{13}C,^{1}H$ multiple quantum relaxation dispersion (RD) experiments,[16] by using selective methyl labeling of methionine residues.[17] Methionine methyl groups are excellent reporters of structure and dynamics because of the simple spectra and the high sensitivity and resolution of TROSY.[18] In order to assign the methyl-TROSY spectrum, we used a site-directed mutagenesis approach (Supporting Information). In order to eliminate the contribution of dipolar relaxation from surrounding protons in the effective transverse relaxation rates ($R_{2,eff}$),[16,18] we produced a highly deuterated [methyl-^{13}C]methionine-labeled POP. Methionine auxotrophic *Escherichia coli* cells were supplemented with [methyl-^{13}C,2,3,3,4,4-D_5]-L-methionine in a highly deuterated expression medium. This methionine isotopomer was chemically synthesized in order to achieve high deuterium content at unlabeled positions, especially at the β and γ positions (Supporting Information). Interestingly, RD experiments of free POP showed pronounced decay curves for most methionine signals (Figure 2, red), thus reflecting the intense microsecond–millisecond dynamics of the free enzyme. Estimated values of the exchange parameters were extracted by fitting the RD data to a two-state model.[16] k_{ex} values obtained from fitting of RD data were between 38 and 167 s^{-1}, with a population between exchanging states around 50% (Table 1). The highest-amplitude motions were in the α/β-hydrolase domain (Figure 3A).

Next, the effects of binding of covalent active-site-directed inhibitors benzyloxycarbonyl-prolyl-prolinal (ZPP, Figure 1C)[19] and KYP-2047 on POP conformational dynamics were examined. Extensive changes in the methyl-TROSY spectra of inhibited POP indicated large-scale conformational rearrangement upon inhibitor binding, which predominantly affected the α/β-hydrolase domain (Figures 3B and S3). Remarkably, RD experiments of inhibited POP revealed that inhibition caused dramatic effects on the microsecond–millisecond dynamics. The decay in the RD profiles of methionine residues was effectively abolished (Figure 2, blue), thus indicating that the binding of inhibitors completely prevented the conformational dynamics of POP.

In order to unravel the structural aspects of the conformational dynamics of POP and the effects of inhibitor binding, we used SAXS, a highly versatile technique that probes molecular structure at low resolution in solution.[15] Free and ZPP-bound POP samples were analyzed by online gel filtration chromatog-

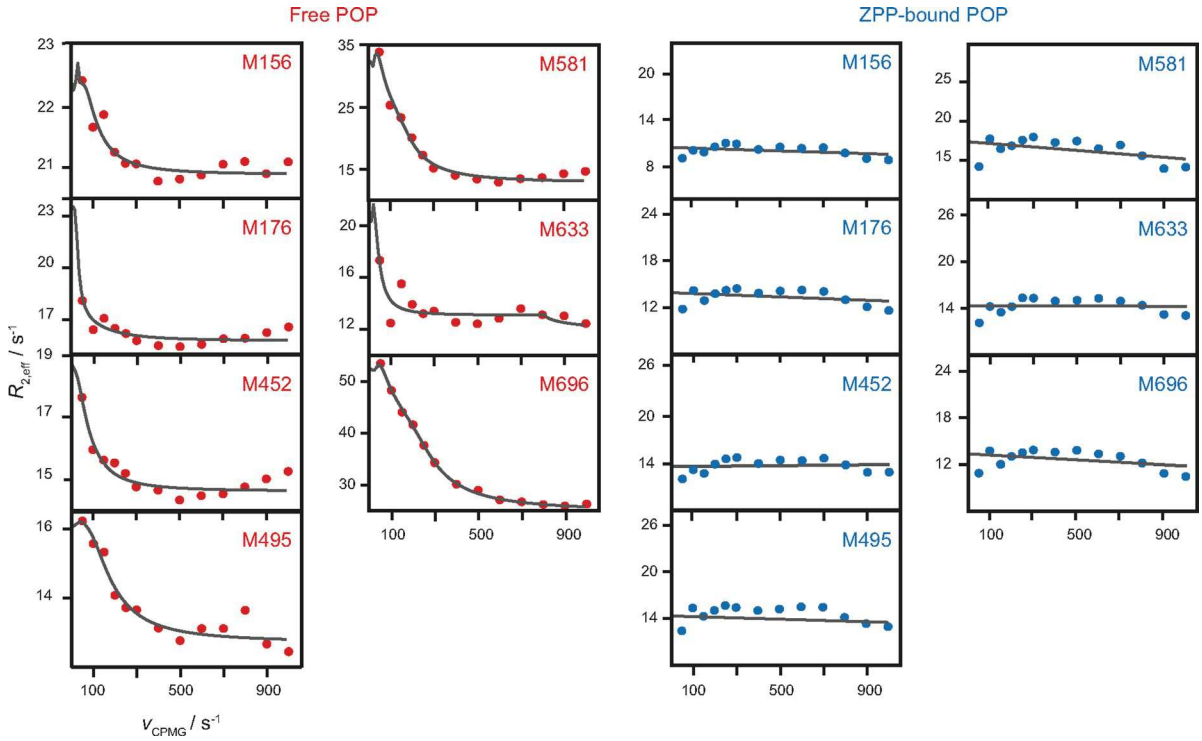


Figure 2. Multiple quantum RD experiments of highly deuterated [methyl-^{13}C]methionine-labeled POP. RD profiles of methionine residues of free POP and ZPP-bound POP are shown, with solid lines for the best fits of the data to a two-state model.

Table 1. Summary of exchange parameters of free POP extracted from the fitting of multiple quantum RD data.

Methionine	k_{ex} [s^{-1}]	p_B [%]	Methionine	k_{ex} [s^{-1}]	p_B [%]
156	45 ± 7	49	581	52 ± 18	46
176	50 ± 9	49	633	50 ± 13	50
452	38 ± 11	51	696	112 ± 14	47
495	167 ± 21	45			

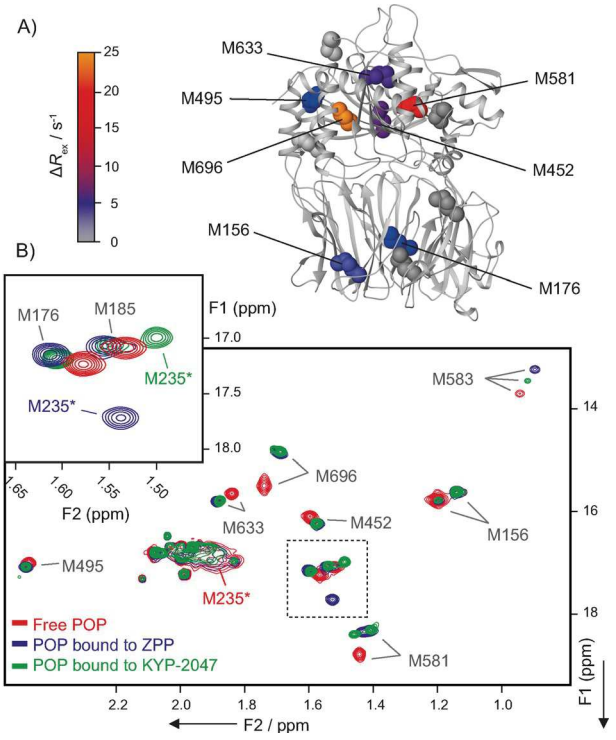


Figure 3. A) Structure distribution of ΔR_{ex} values obtained by multiple quantum RD experiments of free POP. B) Assigned methyl-TROSY spectrum of free POP (red) overlaid with the spectra of POP bound to ZPP (blue) and of POP bound to KYP-2047 (green). Enlarged view (boxed region) shows that the signal of Met235 (*) is strongly sensitive to the binding of active-site-directed inhibitors.

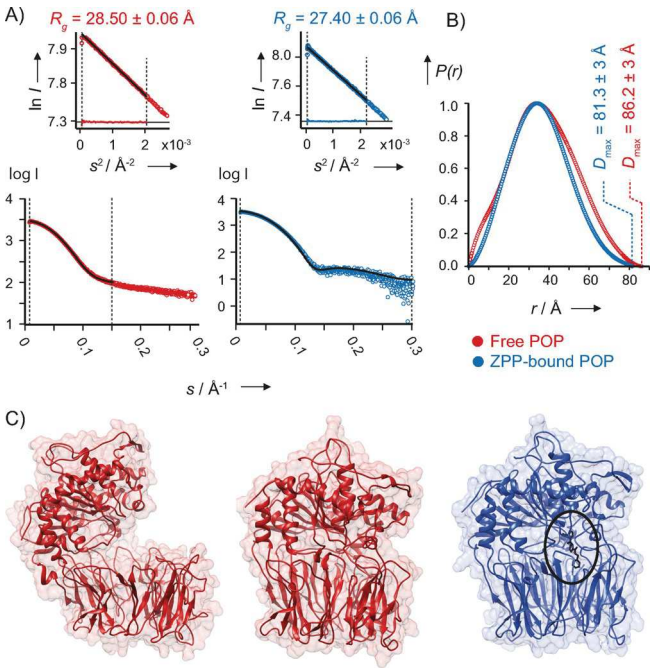


Figure 4. SAXS experiments of free and inhibited POP. A) Averaged scattering profiles of free POP (red) and ZPP-bound POP (blue). Nested Guinier plots and R_g values confirm the absence of interparticle interactions and radiation damage. The theoretical scattering profile obtained by the EOM is shown in black. B) $P(r)$ distribution of free POP (red) and ZPP-bound POP (blue). Dotted lines show the corresponding maximum dimensions (D_{max}), which reflects the bigger size of free POP. C) Structures of POP selected by the EOM. Representative open and closed structures of free POP are shown in red; representative closed conformation of POP covalently bound to ZPP is shown in blue; the inhibitor is marked by a black circle.

raphy coupled to SAXS, in order to eliminate the interference of protein aggregates (Figure S4 in the Supporting Information). In both cases, scattering profiles of eluted monomer species presented high spectral homogeneity, as determined by singular value decomposition (Supporting Information). These profiles were averaged to obtain the corresponding high-quality scattering profiles, which showed no signatures of interparticle interaction or radiation damage (Figures 4A and S5). The overall sizes of particles in solution were evaluated by extracting the radius of gyration (R_g). Comparison of R_g of free and inhibited POP revealed significant structural differences between the two forms (R_g 28.50 ± 0.06 and 27.40 ± 0.06 Å, respectively). In turn, the pair-distance distribution functions ($P(r)$; Figure 4B) of free and inhibited POP revealed significant differences in the global shape. The $P(r)$ function of free POP yielded a multimodal distribution with a maximum dimension (D_{max}) of 86 ± 3 Å, whereas ZPP-bound POP showed a Gaussian-like distribution with a smaller D_{max} (81 ± 3 Å).

In order to assess the conformational equilibrium of POP and the structural consequences of inhibitor binding, we used the ensemble optimization method (EOM),[20] a procedure that optimizes a sub-ensemble of structures from a large pool of model structures, by using the experimental scattering profile as a driving force. We obtained a large pool of models sampling a broad conformational space by performing MD simulations starting from the crystal structure of POP in the closed conformation and the porcine homology model of *Aeromonas punctata* POP in the open conformation (Supporting Information). The EOM of free and ZPP-bound POP provided sub-ensembles of conformations that collectively described the experimental scattering profiles with excellent fit (Figures 4A and S6). The EOM-selected structures of free POP consisted of 55% fully open and 45% closed conformation (Figure 4C, red). These conformational populations are in agreement with those extracted from the fitting of RD data, thus providing cross-validation between the two approaches. The theoretical averaged R_g of the selected open structures was 30.50 ± 0.08 Å, whereas for closed structures it was 27.35 ± 0.07 Å, fully compatible with the experimental R_g of free POP (28.50 ± 0.06 Å). In contrast, the selected structures of ZPP-bound POP consisted exclusively of closed conformations (Figure 4C, blue), as for the

X-ray structure of POP covalently bound to ZPP. The theoretical average R_g of selected inhibited POP structures also was in good agreement with the experimental value (27.23 ± 0.01 and 27.40 ± 0.06 Å, respectively).

In summary, our integrated approach combining NMR spectrometry and SAXS experiments complemented by MD simulations demonstrated that POP is a highly dynamic enzyme on the millisecond timescale, with equilibrium between open and closed conformations. We have shown that the binding of active-site-directed inhibitors effectively impedes the conformational exchange by stabilizing POP in a closed conformation. Therefore, it can be proposed that the microsecond–millisecond conformational dynamics of POP causes significant fluctuations in the configuration of the surface(s) involved in molecular recognition events. Hence, stabilizing POP in a closed conformation by inhibitors would cause substantial alterations to the affinity and specificity of the native PPIs of the enzyme. We speculate that this mechanism could represent a central feature for the reversibility of POP-mediated aggregation of α-synuclein induced by active-site-directed POP inhibitors, as has been reported.[11,13] Overall, the results open the way for designing novel POP inhibitors conceived as conformational modulators to regulate the native interactions of the enzyme.

Experimental Section

Cells were purchased from Novagen (Merck Millipore), chemicals were from Sigma–Aldrich, and deuterated chemicals were from Cambridge Isotope Laboratories (Tewksbury, MA). Affinity and size-exclusion chromatography columns were from GE Healthcare Life Sciences.

Expression of POP and [methyl-^{13}C]methionine-labeled POP: POP was expressed in *E. coli* BL21(DE3) cells by using pET-11 plasmid containing the human POP gene, His tag and TEV cleavage site, by following a standard protocol.[21] After Ni^{2+} affinity chromatography, the His tag was removed by digestion with TEV protease, and POP was purified in a Superdex 200 HiLoad column. For [methyl-^{13}C]methionine-labeled POP, auxotrophic *E. coli* B834 (DE3) cells were grown in minimal medium containing [methyl-^{13}C]-L-methionine (80 mg L^{-1}). Purification was performed as above. In the case of highly deuterated [methyl-^{13}C]methionine-labeled POP, auxotrophic *E. coli* B834(DE3) cells were transformed with pETM-10 plasmid (EMBL, Heidelberg, Germany) containing the human POP gene.[21] Cells were grown in deuterated minimal medium supplemented with [1,2,3,4,5,6,6-D_7]-D-glucose (2 g L^{-1}) and [methyl-^{13}C,2,3,3,4,4-D_5]-L-methionine (50 mg L^{-1}), synthesized as described in the Supporting Information. His tag cleavage was not performed in this case. The binding of ZPP and KYP-2047 inhibitors was carried out by drying an aliquot (10 equiv.) of inhibitor dissolved in 1,4-dioxane with a soft stream of N_2 in a small glass tube. Afterwards, POP sample in the buffer suitable for the experiment was added to the tubes containing the dried inhibitor and incubated for 20 min at room temperature.

NMR experiments and fitting of RD data: All NMR experiments were performed at 25 °C in an 800 MHz Avance III spectrometer (Bruker) equipped with a cryoprobe. POP samples were 200–250 μM in Tris[D_{11}]·HCl (50 mM, pH 8) containing NaCl (20 mM), [D_6]-DTT (1 mM), NaN_3 (0.03%), in D_2O. ^{1}H,^{13}C methyl-TROSY HMQC experiments used the pulse sequence described by Tugarinov et al.[18] Spectra comprised 128, 512 data points (F1, F2), with 240 scans per FID (interscan delay 1.5 s). ^{1}H,^{13}C methyl-TROSY HMQC RD experiments[16] used a CPMG element of 40 ms, with 0, 2, 4, 6, 8, 10, 12, 16, 20, 24, 28, 32, 36, and 40 randomly ordered inversion pulses. Spectra were recorded with 100, 512 data points (F1, F2), accumulating 24 scans (interval 1.5 s). Effective decay rates ($R_{2,eff}$) were extracted from the major resonances by using the following formula[22] [Eq. (1)]:

$$R_{2,\mathrm{eff}}(\nu_{\mathrm{CPMG}}) = \frac{-1}{T}\ln\left(\frac{I(\nu_{\mathrm{CPMG}})}{I_0}\right) \quad (1)$$

where T is total transversal relaxation time (40 ms), and I_0 and $I(\nu_{CPMG})$ are the peak intensities from spectra recorded without or with the CPMG element, respectively. The spectra of RD experiments were converted to NMRpipe format and processed with NMRpipe software.[23] This equation was used in the fitting of RD data to obtain the individual exchange contribution (k_{ex}) and the product $\Delta\omega{\cdot}p_B$ that accounts for the population and the chemical shift of the second state (B), which is ultimately related to the amplitude of the motion. The fitting was performed by considering each *i* methyl probe separately:

$$\Delta\omega_{Hi},\ \Delta\omega_{Ci},\ k_{exi},\ p_{Bi},\ R_{2MQ\infty i}\ (\chi_i^2)$$

Here $\Delta\omega$ is the frequency difference between the ^{1}H or ^{13}C resonances of exchanging signals, p_B is the population of the second state, and $R_{2MQ\infty}$ is the effective transverse relaxation at the fast pulsing limit. The global fitting, which considered the same k_{ex} and p_B for all methyl groups, was performed by adjusting the spectral density of RD data by least squares. The global fitting did not improve the results obtained by the independent one, as revealed by *F*-test statistical analysis.

The deconvolution of the product $\Delta\omega{\cdot}p_B$ from a dataset obtained in a single static field was not statistically reliable. For this reason, several replicate fittings were performed in order to evaluate the reliability of the independent fitting. All replicates yielded highly reproducible results, thus allowing extraction of estimated values for $\Delta\omega(^{13}C)$ and p_B separately. ΔR_{ex} values in Figure 3A were obtained as the difference between theoretical $R_{2,eff}$ at the low and the fast-pulsing limit ($R_{2\infty}$). Given the absence of measurable exchange in the RD profiles of POP bound to ZPP and to KYP-2047, fitting was not performed in the case of inhibitor-bound POP.

Online gel filtration coupled to SAXS: HisTag-cleaved samples of free and ZPP-bound POP were subjected to an online Superdex 200 10/300 SEC column coupled to EMBL beamline P12 of PETRA III (DESY, Hamburg, Germany) with a PILATUS2 M pixel detector (DECTRIS, Baden-Daettwil, Switzerland). The column was run at 0.35 mL min^{-1} to acquiring 1 frame per second (X-ray wavelength 1.24 Å; momentum transfer covered 0.007–0.444 Å^{-1}). The scattering profiles of all frames were inspected, and anomalous profiles were discarded. The scattering profiles corresponding to the pure buffer frames of free and ZPP-bound POP datasets were averaged and subtracted from all profiles in PRIMUS[24] (ATSAS data analysis software; EMBL, Hamburg, Germany). The same program was used to average the subtracted scattering profiles of monomer species of free and inhibited POP, and to derive forward scattering ($I(0)$) and R_g from the Guinier approximation. $P(r)$ distribution functions were obtained with the GNOM program.[25]

Computational methods and ensemble optimization method (EOM): All free-POP MD simulations were performed with AMBER12 software.[26] The ff99SB force field[27] for proteins was used, and explicit water molecules were incorporated as the TIP3P water model.[28] Protein structures were neutralized, and additional sodium and chloride ions were added to simulate physiological saline solution. Protein plus ions were then solvated in pre-equilibrated water molecules in a truncated octahedron box with a 15 Å layer. After energy minimization, the temperature was progressively raised to 300 K with constant pressure dynamics. All production runs were performed with 2.0 fs steps in NPT ensemble (1 bar, 298 K). The shorter MD simulation for ZPP-bound POP was computed with the Desmond molecular dynamics program.[29] The OPLS-AA force field and TIP3P water model were used.[28] The default relaxation protocol in Desmond was used, followed by the production run in the NPT ensemble. Prior to the EOM, theoretical scattering curves were calculated from the simulated PDB files in CRYSOL (ATSAS data analysis software).[30]

The EOM was performed over a sub-ensemble of N structures from a large pool of M model structures ($M \gg N$) by minimizing χ^2 between the experimental (I_{exp}) and theoretical (I_{theor}) profiles [Eq. (2)]:

$$\chi^2 = \frac{1}{K-1}\sum_{j=1}^{K}\left[\frac{\mu I_{theor}(s_j) - I_{exp}(s_j)}{\sigma(s_j)}\right]^2 \quad (2)$$

where K is the number of data points, $\sigma(s)$ is standard deviation, and μ is a scaling factor. $I_{theor,n}(s)$ is defined from the individual n profiles as follows [Eq. (3)]:

$$I_{theor}(s) = \frac{1}{N}\sum_{n=1}^{N} I_{theor,n}(s) \quad (3)$$

Experimental curves used in the EOM comprised data points from $s < 0.15$ Å^{-1} for POP, and from $s < 0.3$ Å^{-1} for POP bound to ZPP. Constant subtraction was applied in all cases. The EOM was carried out with 50 random initial sub-ensembles of $N = 20$ structures, as the use of more structures in this method ($N = 50$) did not improve the result.[20] A total of 1500 generations were performed. One hundred independent EOM runs were performed, and the most frequent result was taken as the solution with best fit (Figure 4 A).

Acknowledgements

This work was supported by the Institute for Research in Biomedicine, MINECO-FEDER (Bio2013–40716-R, CTQ2013–48287 and CTQ2012–32183/BQU), and the Generalitat de Catalunya (XRB and Grup Consolidat 2014SGR521). A.L. has received funding from the Instituto de Salud Carlos III. P.B. acknowledges the Agence Nationale de la Recherche (SPIN-HD-ANR-CHEX-2011) and the ATIP-Avenir program for financial support. F.H.T.'s fellowship is co-funded by the INSERM and the University of Copenhagen. Technical assistance from staff at the P12 beam line (EMBL/DESY) is acknowledged.

Keywords: NMR spectroscopy · prolyl oligopeptidase · protein dynamics · protein–protein interactions · SAXS

[1] a) P. Bernadó, M. Blackledge, *Nature* **2010**, *468*, 1046 – 1048; b) K. Henzler-Wildman, D. Kern, *Nature* **2007**, *450*, 964 – 972.

[2] a) L. Nevola, E. Giralt, *Chem. Commun.* **2015**, *51*, 3302 – 3315; b) S. Jaeger, P. Aloy, *IUBMB Life* **2012**, *64*, 529 – 537.

[3] D. D. Boehr, R. Nussinov, P. E. Wright, *Nat. Chem. Biol.* **2009**, *5*, 789 – 796.

[4] D. F. Cunningham, B. O'Connor, *Int. J. Biochem. Cell Biol.* **1998**, *30*, 99 – 114.

[5] V. Fülöp, Z. Böcskei, L. Polgár, *Cell* **1998**, *94*, 161 – 170.

[6] a) M. Li, C. Chen, D. R. Davies, T. K. Chiu, *J. Biol. Chem.* **2010**, *285*, 21487 – 21495; b) L. Shan, I. I. Mathews, C. Khosla, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 3599 – 3604.

[7] P. Canning, D. Rea, R. E. Morty, V. Fülöp, *PloS one* **2013**, *8*, e79349.

[8] N. Kichik, T. Tarragó, B. Claasen, M. Gairi, O. Millet, E. Giralt, *ChemBioChem* **2011**, *12*, 2737 – 2739.

[9] K. Kaszuba, T. Róg, R. Danne, P. Canning, V. Fülöp, T. Juhász, Z. Szeltner, J.-F. St. Pierre, A. Garcia-Horsman, P. T. Männistö, M. Karttunen, J. Hokkanen, A. Bunker, *Biochimie* **2012**, *94*, 1398 – 1411.

[10] E. Di Daniel, C. P. Glover, E. Grot, M. K. Chan, T. H. Sanderson, J. H. White, C. L. Ellis, K. T. Gallagher, J. Uney, J. Thomas, P. R. Maycox, A. W. Mudge, *Mol. Cell. Neurosci.* **2009**, *41*, 373 – 382.

[11] a) M. H. Savolainen, X. Yan, T. T. Myöhänen, H. J. Huttunen, *J. Biol. Chem.* **2015**, *290*, 5117 – 5126; b) I. Brandt, M. Gérard, K. Sergeant, B. Devreese, V. Baekelandt, K. Augustyns, S. Scharpé, Y. Engelborghs, A.-M. Lambeir, *Peptides* **2008**, *29*, 1472 – 1478.

[12] J. I. Venäläinen, J. A. Garcia-Horsman, M. M. Forsberg, A. Jalkanen, E. A. A. Wallén, E. M. Jarho, J. A. M. Christiaans, J. Gynther, P. T. Männistö, *Biochem. Pharmacol.* **2006**, *71*, 683 – 692.

[13] a) T. T. Myöhänen, M. J. Hannula, R. Van Elzen, M. Gerard, P. Van Der Veken, J. A. Garcia-Horsman, V. Baekelandt, P. T. Männistö, A. M. Lambeir, *British J. Pharmacol.* **2012**, *166*, 1097 – 1113; b) M. H. Savolainen, C. T. Richie, B. K. Harvey, P. T. Männistö, K. A. Maguire-Zeiss, T. T. Myöhänen, *Neurobiol. Dis.* **2014**, *68*, 1 – 15.

[14] A. G. Palmer III, *Chem. Rev.* **2004**, *104*, 3623 – 3640.

[15] a) H. D. T. Mertens, D. I. Svergun, *J. Struct. Biol.* **2010**, *172*, 128 – 141; b) D. A. Jacques, J. Trewhella, *Protein Sci.* **2010**, *19*, 642 – 657.

[16] D. M. Korzhnev, K. Kloiber, V. Kanelis, V. Tugarinov, L. E. Kay, *J. Am. Chem. Soc.* **2004**, *126*, 3964 – 3973.

[17] a) I. Gelis, A. M. J. J. Bonvin, D. Keramisanou, M. Koukaki, G. Gouridis, S. Karamanou, A. Economou, C. G. Kalodimos, *Cell* **2007**, *131*, 756 – 769; b) T. L. Religa, R. Sprangers, L. E. Kay, *Science* **2010**, *328*, 98 – 102.

[18] V. Tugarinov, P. M. Hwang, J. E. Ollerenshaw, L. E. Kay, *J. Am. Chem. Soc.* **2003**, *125*, 10420 – 10428.

[19] T. Yoshimoto, K. Kawahara, F. Matsubara, K. Kado, D. Tsuru, *J. Biochem.* **1985**, *98*, 975 – 979.

[20] P. Bernadó, E. Mylonas, M. V. Petoukhov, M. Blackledge, D. I. Svergun, *J. Am. Chem. Soc.* **2007**, *129*, 5656 – 5664.

[21] T. Tarragó, S. Frutos, R. A. Rodriguez-Mias, E. Giralt, *ChemBioChem* **2006**, *7*, 827 – 833.

[22] F. A. A. Mulder, N. R. Skrynnikov, B. Hon, F. W. Dahlquist, L. E. Kay, *J. Am. Chem. Soc.* **2001**, *123*, 967 – 975.

[23] F. Delaglio, S. Grzesiek, G. W. Vuister, G. Zhu, J. Pfeifer, A. Bax, *J. Biomol. NMR* **1995**, *6*, 277 – 293.

[24] P. V. Konarev, V. V. Volkov, A. V. Sokolova, M. H. J. Koch, D. I. Svergun, *J. Appl. Crystallogr.* **2003**, *36*, 1277 – 1282.

[25] D. Svergun, *J. Appl. Crystallogr.* **1992**, *25*, 495 – 503.

[26] D. A. Case, T. A. Darden, T. E. Cheatham III, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, R. C. Walker, W. Zhang, K. M. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A. W. Götz, I. Kolossváry, K. F. Wong, F. Paesani, J. Vanicek, et al., **2012**.

[27] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, C. Simmerling, *Proteins Struct. Funct. Bioinf.* **2006**, *65*, 712 – 725.

[28] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, *J. Chem. Phys.* **1983**, *79*, 926 – 935.

[29] D. E. Shaw Research www.deshawresearch.com.

[30] D. Svergun, C. Barberato, M. H. J. Koch, *J. Appl. Crystallogr.* **1995**, *28*, 768 – 773.

Manuscript received: February 19, 2016

Accepted article published: February 25, 2016

Final article published: March 30, 2016

Paper III: Multiple Stable Conformations Account for Reversible Concentration-Dependent Oligomerization and Autoinhibition of a Metamorphic Metallopeptidase

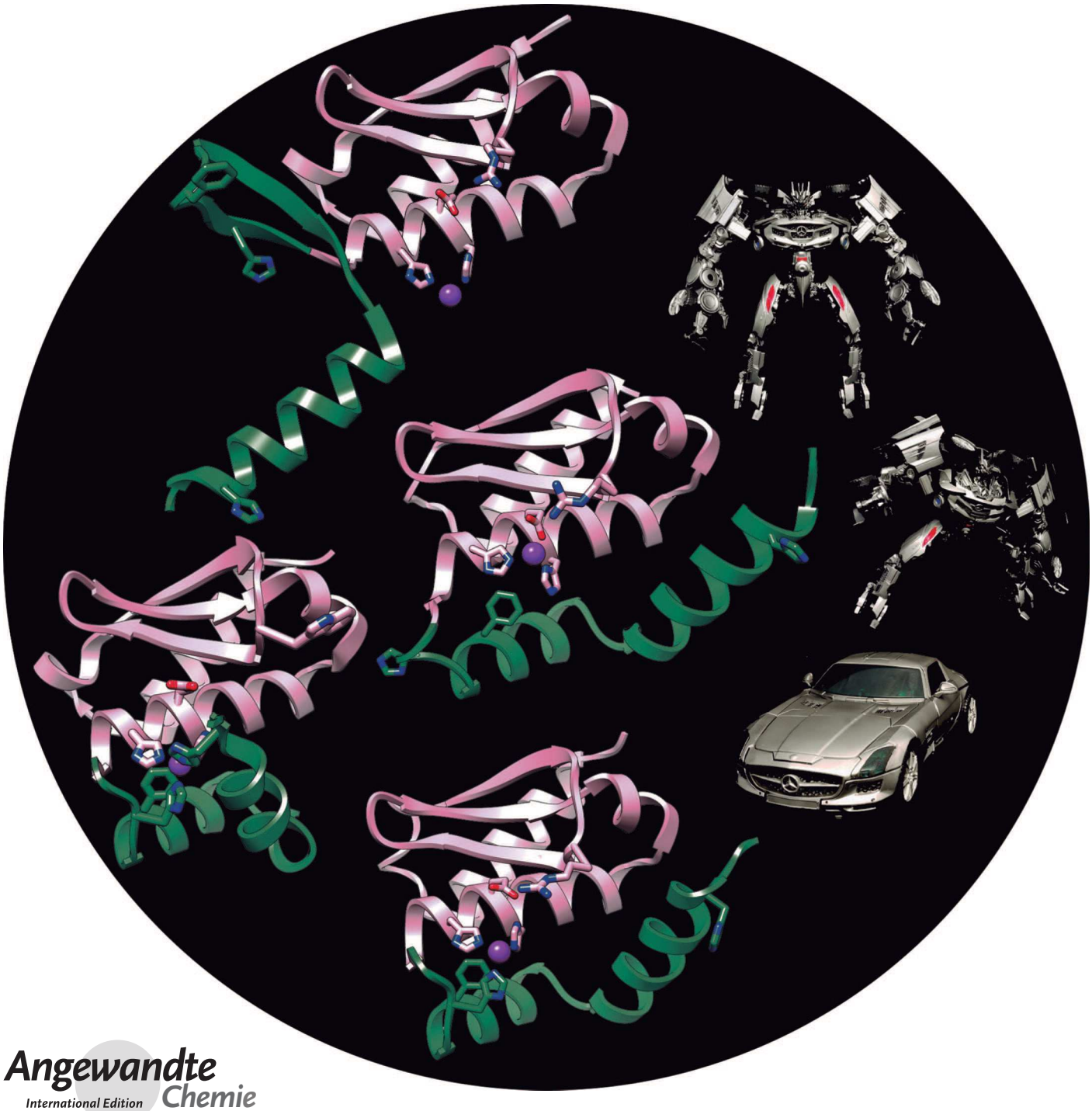
Angewandte Communications

Enzymatic Activity

DOI: 10.1002/anie.201405727

Multiple Stable Conformations Account for Reversible Concentration-Dependent Oligomerization and Autoinhibition of a Metamorphic Metallopeptidase

*Mar López-Pelegrín, Núria Cerdà-Costa, Anna Cintas-Pedrola, Fátima Herranz-Trillo, Pau Bernadó, Juan R. Peinado, Joan L. Arolas, and F. Xavier Gomis-Rüth**



Angewandte International Edition Chemie

Abstract: *Molecular plasticity controls enzymatic activity: the native fold of a protein in a given environment is normally unique and at a global free-energy minimum. Some proteins, however, spontaneously undergo substantial fold switching to reversibly transit between defined conformers, the "metamorphic" proteins. Here, we present a minimal metamorphic, selective, and specific caseinolytic metallopeptidase, selecase, which reversibly transits between several different states of defined three-dimensional structure, which are associated with loss of enzymatic activity due to autoinhibition. The latter is triggered by sequestering the competent conformation in incompetent but structured dimers, tetramers, and octamers. This system, which is compatible with a discrete multifunnel energy landscape, affords a switch that provides a reversible mechanism of control of catalytic activity unique in nature.*

In general, the native fold of a protein in a given environment is unique and at a global free-energy minimum.[1] However, some proteins spontaneously undergo substantial fold switching and reversibly transit between several conformers: "metamorphic" proteins.[2] Identifying and examining such proteins is a challenge because they are highly dynamic and impossible to identify a priori.[3] In contrast, minor rearrangement often occurs in single-domain enzymes upon binding of substrates, as shown for proteolytic enzymes of the metallopeptidase (MP) class.[4] As to enzymatic activity, an increase in enzyme concentration usually increases activity, as more substrate can be bound and turned over.[5] Here we describe a metamorphic minimal *sel*ective and specific *cas*einolytic metallopeptid*ase*, *selecase*, which shows a reversible and concentration-dependent equilibrium between different discrete states and an associated loss of enzymatic activity due to autoinhibition.

We recently discovered a family of soluble minimal MPs named minigluzincins and characterized two of them, proabylysin and projannalysin, but we only isolated them as inactive zymogens, each in a single conformation.[6] In the present study, we introduce selecase from *Methanocaldococcus jannaschii* as a novel family member. We recombinantly produced and purified selecase (see the Experimental Procedures [EP] and Supplemental Results and Discussion [SRD] in the Supporting Information for details). In contrast to the other minigluzincins, the 110-residue full-length selecase corresponded to a mature, fully active MP with narrow and selective—hitherto unreported—substrate specificity that cleaved bovine milk casein at a single site on its α_{s1} chain (Suppl. Figure 1 and Suppl. Tables 1 and 2).

Selecase was extremely soluble in aqueous buffer and did not precipitate at 130 mg mL^{-1}. Thus, we studied the concentration-dependent enzymatic activity of selecase on a peptide that mimics the casein cleavage site (peptide CCS). Normally, peptide-bond hydrolysis by MPs is an ordered single-displacement reaction that follows simple Michaelis–Menten kinetics.[7] This entails that higher enzyme concentrations enhance the initial rate of reaction in the pre-steady state following a hyperbolic curve until a plateau is reached upon saturation.[5] This is found for example, with tobacco-etch virus proteinase, which is widely used in biotechnology (Figure 1 a).

Surprisingly, although selecase activity did indeed increase with concentration between 0.025–0.25 mg mL^{-1}, it fell sharply thereafter to become only residual at 50 mg mL^{-1}. Most interestingly, this inactive concentrated selecase regained maximal activity following simple dilution with buffer. Accordingly, selecase showed reversible enzymatic autoinhibition due to changes in concentration—and not to inhibition by the substrate or any other reagent. This, to our knowledge, is novel for peptidases.

Subsequently, we explored the oligomerization of selecase in solution in the concentration range 0.15–65 mg mL^{-1} using several biophysical techniques (see EP and SRD for full details). Briefly, calibrated size-exclusion chromatography (SEC) revealed monomers, dimers, tetramers, and octamers in variable amounts depending on the concentration (Suppl. Figure 2 a). SEC-MALLS, which combines SEC with multi-angle laser light scattering (MALLS), revealed two average populations with molecular weights of 25 KDa and 80 KDa, possibly corresponding to dimeric and octameric selecase, respectively, along with additional species such as monomers and tetramers (Figure 1 b and Suppl. Figure 2 b). Sedimentation velocity analytical ultracentrifugation revealed the concentration-dependent presence of four oligomeric species, which would be consistent with monomers, dimers, tetramers, and octamers. This was backed by equilibrium velocity experiments showing concentration-dependent average masses ranging between monomers + dimers and octamers (Figure 1 c and Suppl. Table 3). Chemical crosslinking experiments followed by SDS-PAGE, in turn, showed monomers, dimers, monomer–dimer complexes, and tetramers. Higher oligomerization species were not detected due to intrinsic experimental limitations (Suppl. Figure 2 c). The circular dichroism spectra of selecase, with either zinc or nickel in the catalytic site, displayed the typical shape of well-folded mostly α-helical proteins (Suppl. Figure 2 d). Finally, SAXS revealed that the protein did not aggregate at concentrations of up to 65 mg mL^{-1} (Suppl. Table 4, Figure 1 d, Suppl. Figures 3 and 4). These results further showed that the relative population of the oligomeric species in solution was concentration dependent. In addition, single-value decomposition analysis of the SAXS dataset indicated that four

[*] M. López-Pelegrín,[+] Dr. N. Cerdà-Costa,[+] Dr. A. Cintas-Pedrola, Dr. J. R. Peinado,[$] Dr. J. L. Arolas, Prof. Dr. F. X. Gomis-Rüth
Proteolysis Lab, Molecular Biology Institute of Barcelona
CSIC, Barcelona Science Park
c/Baldiri Reixac, 15–21, 08028 Barcelona (Spain)
E-mail: fxgr@ibmb.csic.es
Homepage: http://www.ibmb.csic.es/home/xgomis
F. Herranz-Trillo, Dr. P. Bernadó
Centre de Biochimie Structurale, INSERM U1054
CNRS UMR 5048, Université Montpellier 1 and 2
29 rue de Navacelles, 34090 Montpellier (France)
[$] Present address: Department of Medical Sciences
University of Castilla-La Mancha, 13071 Ciudad Real (Spain)
[+] These authors contributed equally to this work.

Supporting information for this article (including experimental procedures, supplemental results and discussion, acknowledgments, supplemental references, tables, figures, and movies) is available on the WWW under http://dx.doi.org/10.1002/anie.201405727.

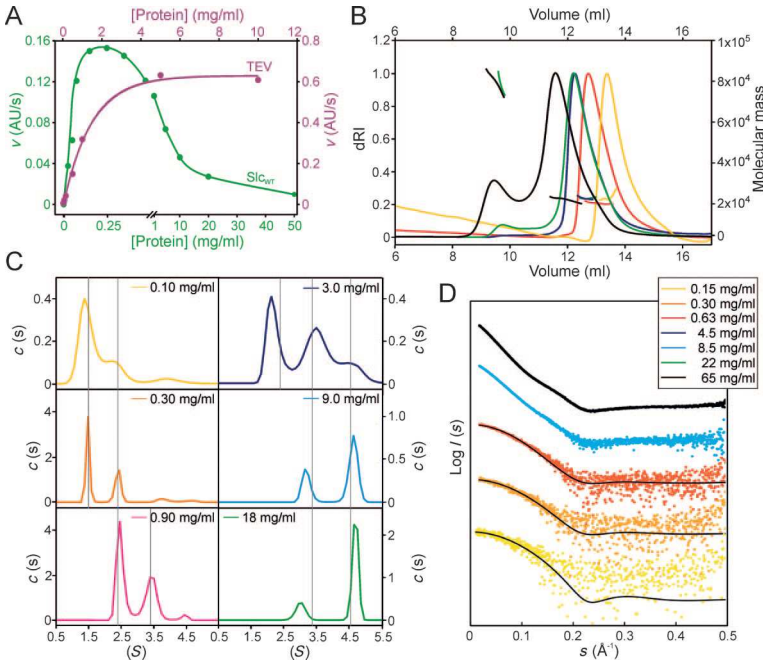


Figure 1. A polyoligomeric metallocaseinase with abnormal activity. A) Proteolytic activity of wild-type selecase on peptide CCS (green curve). Tobacco-etch virus proteinase mutant S219V, which shows comparable catalytic efficiency to selecase but normal concentration-dependent activity, is shown for comparison (purple curve). B) SEC-MALLS of selecase at selected initial concentrations (0.15–65 mg mL^{-1}; see also Suppl. Figure 2b). The peak pattern moves towards smaller elution volumes with increasing protein concentration, thus suggesting protein oligomerization. Curves are colored according to the inset in panel (D). dRI = differential refracting index. C) Analytical ultracentrifugation curves at six selected concentrations depicting the concentration-dependent oligomeric populations. Essentially, monomers are predominantly found at 0–0.3 mg mL^{-1}; dimers at 0.3–2 mg mL^{-1}; tetramers at 2–6 mg mL^{-1}; and octamers at >6 mg mL^{-1}. S = sedimentation coefficient, $c(S)$ = continuous sedimentation coefficient discribution. D) SAXS intensity profiles, $I(s)$, as a function of the momentum transfer, s, measured for wild-type selecase at selected concentrations (see Suppl. Figure 3 for all curves). Profiles have been displaced along the $I(s)$ axis for comparison. The experimental scattering curves at the three lowest concentrations studied indicate a mixed population of monomers and dimers based on the crystallographic structures of slc_1 and slc_2 (black curves).

species (monomers, dimers, tetramers, and octamers) were present.

Summarizing, biophysical analyses in solution indicated the presence of mixtures of monomers, dimers, tetramers, and octamers, with higher concentrations leading to greater oligomerization but not indiscriminate aggregation or precipitation. The concentrations at which monomeric selecase was predominant coincided with those of maximal enzymatic activity (0.2–0.3 mg mL^{-1}; Figure 1a), thus indicating that the monomer is the active species and that oligomers correspond to self-inhibiting species in all cases (see below). This would explain why higher enzyme concentrations yielded lower activity (Figure 1a) and is reminiscent of previous reports on oligomerization inhibiting the activity of phospholipase A2.[9] Notably, simple dilution with buffer reversed oligomerization to yield monomers and restore activity.

To identify the molecular determinants of this behavior, we crystallized and solved the structure of wild-type selecase (see EP, SRD, and Suppl. Table 5). We obtained three crystal forms—orthorhombic, tetragonal, and hexagonal—which serendipitously corresponded to monomeric (slc_1), dimeric (slc_2), and tetrameric (slc_4) forms of selecase, respectively. This indicated that at least three of the oligomerization states found in solution had a counterpart in the form of a stable, isolatable species, each one favored by particular crystallization conditions. The crystal structure of monomeric slc_1 reveals—by comparison with several functional but otherwise unrelated MPs—that the overall architecture, the metal-binding site, and the active-site environment are consistent with a competent and functional mature enzyme (see Refs. [7b,10]). This conclusion is supported by the enzymatic activity found for selecase in solution associated with a monomeric species (see above). It is also reinforced by SAXS for which the experimental scattering curves at the three lowest concentrations—covering the activity maximum of the enzyme—clearly indicated mixed populations of monomers and dimers based on the crystallographic coordinates of slc_1 and slc_2 (see below), with the monomeric fraction at the two lowest concentrations reaching 70% (see Figure 1d, SRD, and Suppl. Figures 3 and 4).

At 13.1 KDa, slc_1 is the smallest active peptidase structurally characterized to date and it has a compact globular shape 35–40 Å in diameter (Figure 2a). It consists of an upper N-terminal subdomain (NTS; residues M^1–Y^{76}) and a lower C-terminal subdomain (CTS; G^{77}–K^{109}), which are connected by a mostly hydrophobic interface (Suppl. Table 6) and separated by a horizontal central active-site cleft (Figure 2b). The NTS is an α/β-sandwich, with a three-stranded mixed β-sheet (β1–β3; Suppl. Table 7) that forms the roof of the selecase moiety (Figure 2a). Two roughly parallel α-helices ("backing helix" α1 and "active-site helix" α2) are attached to the convex surface of the sheet, which faces the central core of the protein. A short helical segment ("linking helix" αB) is inserted in the loop connecting strand β3 with helix α2 (Lβ3α2). Helix α2 roughly parallels the active-site cleft and ends with the last residue of the NTS at Y^{76}. It encompasses a metal-binding motif, H^{69}-E^{70}-X-X-H^{73}, which is characteristic of MPs and includes two metal-binding histidines and a general base/acid glutamate essential for catalysis.[11] Residue H^{80}, imbedded within Lα2α3 of the CTS, is the third metal ligand. The CTS mainly consists of two helices ("glutamate helix" α3 and "C-terminal helix" α4), whose axes intersect at roughly 90°. Helix α3 contains F^{84} at the center of the "Ser/Gly-turn",[6,11a] which creates a hydrophobic base for the metal-binding site and contributes to its stabilization. The active-site cleft of selecase is framed by helix α2; the "upper-rim" strand β2 of the NTS sheet and the preceding "bulge-edge segment" (L^{34}–I^{38}); helices α3 and α4; and Lα2α3, in particular through the side chains of K^{78} and Y^{79}. The catalytic metal ion resides at the bottom left of

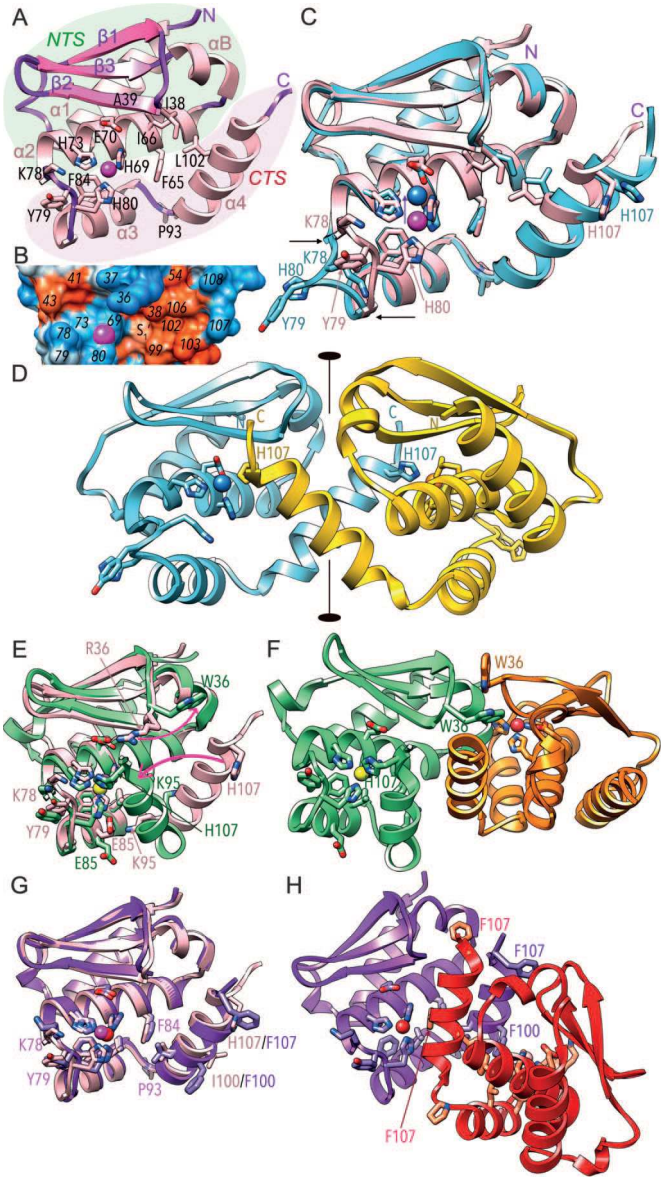


Figure 2. Competent monomer versus incompetent dimers. A) Ribbon representation of slc_1 in standard orientation.[8] Helices (α1, αB, and α2–α4) are shown in pink, β-strands (β1–β3) in magenta, and loops/coils in purple. For extent and nomenclature of regular secondary structure elements, see Suppl. Table 7. Selected residues are shown for their side chains, as is the catalytic metal ion (magenta sphere). The NTS and the CTS are shown over light green and light purple background, respectively. B) Surface representation of slc_1 colored according to Kyte–Doolittle hydrophobicity (blue = hydrophilic over white to orange = hydrophobic) in the same orientation as in (A) showing the active-site cleft with the hydrophobic S_1' specificity pocket. C) Superposition of slc_1 (in pink) and the slc_2 monomer (in cyan). Depicted are the respective metal ions, which are shifted relative to each other (purple arrow). Horizontal black arrows pinpoint the anchor points around which the conformational rearrangement occurs. D) Overall structure of symmetric dimeric slc_2 (chains in cyan and gold) depicted so that the crystallographic dyad (black horizontal ellipses joined by a line) is in the plane of the picture. E) Superposition of slc_1 (in pink) and molecule B of the $R^{36}W$ selecase dimeric mutant ($slc_{2'}$; in light green). Magenta arrows pinpoint the side-chain movement at position 36 owing to the mutation and the 50° rotation of C-terminal helix α4. F) Structure of the asymmetric dimer of $slc_{2'}$ consisting of helix-rotated molecule B (green) and close-to-native molecule A (orange). Both active-site clefts are blocked but following different mechanisms. Note the two W^{36} side chains at the interface. G) Superposition of slc_1 (in pink) and one of the two equivalent close-to-native monomers of selecase $I^{100}F + H^{107}F$ dimeric mutant ($slc_{2''}$; in purple). H) Inactive dimer of $slc_{2''}$ (in purple and red).

the cleft (Figure 2a,b). At its right, a deep hydrophobic S_1' pocket is shaped by I^{38}, A^{39}, F^{65}, I^{66}, L^{102}, and the solvent-accessible ring surface of H^{69}. This pocket optimally accommodates a phenylalanine in the P_1' position of substrates as found at the casein cleavage site. The slc_1 moiety is held together by a central hydrophobic core, which traverses the entire molecule, and several of the contributing residues also shape the NTS–CTS interface (Figure 2a and Suppl. Table 6).

The crystal structure of slc_2 shows a dimer (Suppl. Table 8), and superposition of slc_1 and slc_2 monomers reveals good overall fit, with only minor differences within the NTS (see Figure 2c). However, major metamorphic rearrangement is observed around the metal-binding site (see Figure 2c and Suppl. Movie 1). In slc_2, at the beginning of CTS, Lα2α3 folds outward between G^{77} and I^{81}, with a maximal displacement of 7 Å. This causes the third metal-binding protein residue in slc_1, H^{80}, to swing out and protrude from the molecular surface. This, in turn, leads to an upward shift of the catalytic metal towards the general base/acid E^{70} (Figure 2c). Two selecase monomers associate through C_2 symmetry under occlusion of a large surface (2130 $Å^2$; see Suppl. Table 8 and Figure 2d) and so the third metal-binding site is taken over by H^{107} from helix α4 of the symmetric molecule. Accordingly, this H^{80}/H^{107} ligand swap is an intermolecular event that yields a catalytically incompetent metal-binding site and a blocked active-site cleft in slc_2. This is consistent with oligomerization coinciding with inactive species in solution (see above).

As in slc_2, the protomer of tetrameric slc_4 shows good overall fit with slc_1 within the NTS, including the position and conformation of most side chains at the NTS–CTS interface. However, both major displacement and drastic conformational rearrangement are observed in the CTS (see Figure 3a and Suppl. Movie 2). The segment of the active-site helix with the first two metal ligands undergoes slight displacement (Figure 3a). Downstream loop Lα2α3 and glutamate helix α3—which is virtually unchanged in both slc_1 and slc_2—unfold and give rise to strands β4 and β5, which adopt a canonical β-ribbon structure (Figure 3a and Suppl. Table 7). Such long stretches of a protein only rarely undergo such dramatic transitions.[12] The β-ribbon protrudes away from the molecular moiety (Figure 3a,b), as a result of which metal-ligand H^{80} shifts roughly 16 Å from its position in slc_1 and no longer binds the metal. In contrast with α3, the C-terminal helix α4 keeps its helical structure but is displaced about 30 Å apart on average in slc_4 (Figure 3a). Overall, this metamorphic structural transition of selecase is stabilized by the association of four monomers in the crystal (Figure 3b–d, Suppl. Table 8, and Suppl. Movie 2), which would explain tetrameric oligomerization in solution (see above). The oligomer is a compact, almost spherical self-inhibitory particle 60–75 Å in diameter

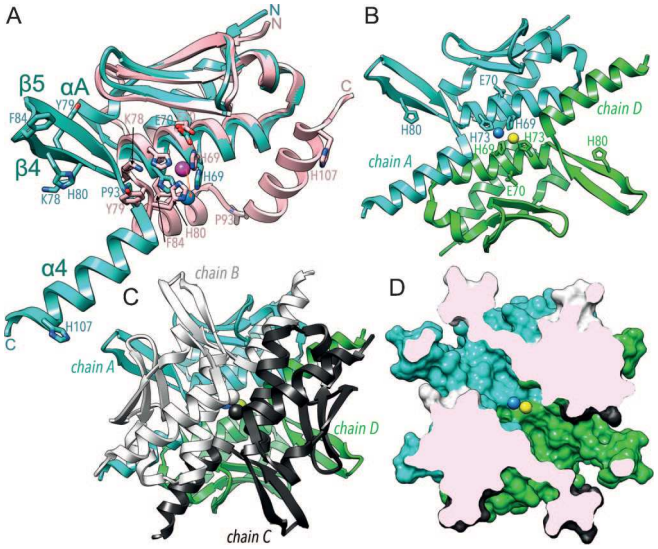


Figure 3. A compact autoinhibitory tetrameric particle. A) Superposition of slc_1 and slc_4 monomers in pink (magenta metal ion) and turquoise (blue metal ion), respectively, in the view of Figure 2a. Only the distinct secondary structure elements of slc_4 are labeled (see also Suppl. Table 7). Relevant residues undergoing major rearrangement are displayed for both structures and labeled. The metal is shifted downwards (red arrow). B) Within the slc_4 tetramer, two neighbor monomers as in (A), in turquoise (chain A; metal in blue) and light green (chain D; metal in yellow), bind over a crystallographic dyad perpendicular to the plane of the picture. This gives rise to a nonfunctional dimetallic zinc site bound by H^{69} and H^{73} of either monomer. C) Two dimers as in (B), in turquoise/light green (chains A and D) and white/dark gray (chains B and C), associate face to face under a relative 90° rotation to yield the overall tetrameric particle, with two dimetallic zinc sites in the particle lumen. D) Surface representation of (C) after clipping off the frontal part to delineate the central particle channel. Only the dimetallic site depicted in (B) is shown for clarity.

(Figure 3c,d). One monomer (chain A) interacts through D_2 symmetry—by hiding a total interface of 9850 $Å^2$—with two neighboring molecules (chains B and D) through mixed hydrophobic/hydrophilic contacts, and with one opposite monomer (chain C) through hydrophobic contacts (Suppl. Table 8). Two large elliptical openings (minor axis ≈16 Å, major axis ≈21 Å; Figure 3d and Suppl. Movie 2) on opposite faces of the particle are framed by upper-rim strands β2 and Lβ5α4 of two vicinal monomers (AB and CD). Access to the particle lumen through these entrances is limited by the respective β-ribbons, which protrude away from the particle surface and do not contact each other. The central lumen of the particle features a channel 50 Å in length and 15 Å in diameter and allocates two internal dimetallic zinc-binding sites. Each of them results from the fusion of two neighboring metal sites as originally found in slc_1 (chains AD and BC, respectively), with the two metal ions of each site roughly 3 Å apart (Figure 3b,d). Overall, this new conformation radically alters the structural segments that shape the S_1' pocket and the active-site cleft in competent slc_1 and, thus, indicates that, like slc_2, the tetrameric slc_4 structure corresponds to an inactive species. This, again, is consistent with tetramers coinciding with inactive species in solution.

Given the importance of the C-terminal helix α4 and loop Lβ1β2 in oligomerization, we selected residues R^{36}, I^{100}, I^{103}, and H^{107}, which had been observed to participate in dimerization in slc_2 and tetramerization in slc_4 (Suppl. Table 8), and generated a total of seven single, double, and triple point mutants in an attempt to ablate the interactions responsible for oligomerization and thus obtain monomeric forms. In addition, we constructed two deletion mutants targeting α4, lacking four (slcΔC4) and eight (slcΔC8) C-terminal residues. Moreover, we cloned two close orthologues from *Methanotorris igneus* and *Methanocaldococcus fervens*, which can be envisaged as natural fivefold and 19-fold point mutants of selecase (see EP and SRD for full details). All protein variants were produced, purified, and concentrated similarly to the wild-type except for slcΔC4, which was obtained with lower yields and could only be maximally concentrated to 5.0 $mg\,mL^{-1}$, and slcΔC8, which was insoluble and was discarded. This finding pointed to a stabilizing effect of helix α4 on the whole protein despite its overall flexibility in the various structures analyzed. Despite differences in the oligomer populations, all mutants displayed a concentration-dependent equilibrium between monomers, dimers, tetramers, and octamers and a reduction in activity as concentration increased, similar to the wild-type (Suppl. Figure 5). These results indicate that selecase is highly plastic, which allows it to adapt to potentially deleterious point mutations and retain its capacity to oligomerize.

This plasticity is backed by further structural studies. Out of all the aforementioned mutants and orthologues, we managed to crystallize variants $R^{36}W$ (hereafter $slc_{2'}$) and $I^{100}F + H^{107}F$ (hereafter $slc_{2''}$) and solved their crystal structures (see Figures 2e–h and Suppl. Movie 3). Most interestingly, $slc_{2'}$ showed a novel dimeric quaternary structure, distinct from slc_2, which displayed each protomer in a different conformation despite the chemical identity of the molecules. One molecule (A) essentially displays the conformation of functional monomeric slc_1, including the metal site and the active-site cleft. It only differs significantly from the latter at Lβ1β2, which, owing to the side-chain replacement at position 36, causes the entire loop and thus the latter side chain to undergo major rearrangement towards the molecular moiety. The other molecule (B) also essentially coincides with slc_1 but only until the glutamate helix. Thereafter, a 90° rotation around bond N–Cα of K^{95} results in C-terminal helix α4 being rotated as a rigid body by 50° so as to approach and thus sterically block its own active-site cleft on its primed side. This further causes H^{107} to bind the catalytic metal, as observed in slc_2, except that here this is an intramolecular rather than an intermolecular event (compare Figure 2d–f). This novel conformation of a selecase variant in molecule B is stabilized by an asymmetric interaction between C-terminal helices with molecule A triggered by an edge-to-face interaction of the W^{36} side chains (Figure 2f). This arrangement, in turn, causes the active-site cleft of molecule A to be blocked for substrate access by helix αB of molecule B, with Y^{57} of the latter interacting with the S_1' pocket of molecule A. The metal-binding site of the latter, in contrast, is unaffected. Accordingly, $slc_{2'}$ corresponds—like slc_2—to an inhibited conformation.

As to $slc_{2''}$, superposition of the two essentially identical monomers in the asymmetric unit onto slc_1 revealed a conformation that was close to that of the functional wild-type monomer, except that the end of the C-terminal helix was slightly unwound and more flexible owing to the two point mutations (Figure 2g). However, the two phenylalanine residues at positions 100 and 107 make two $slc_{2''}$ monomers symmetrically bind mainly through their respective C-terminal helices, which run roughly parallel to each other. As a result the nonprimed sides of the active-site clefts are occluded and the phenylalanine rings at position 100 penetrate the S_1' specificity pocket of the symmetric partner, as this residue matches the specificity of the enzyme. Further symmetric contacts are observed between the F^{107} side chain of one molecule and loop Lβ1β2 of the other, which enhance the overall flexibility of these regions. Accordingly, the structure of $slc_{2''}$ provides yet another mechanism of inhibition of selecase, in this case merely by the shielding of the cleft (Figure 2h). Thus, the two crystal structures of $slc_{2'}$ and $slc_{2''}$ may represent genuine dimeric conformations of the mutants triggered by the respective side-chain replacements, as none of the corresponding structures was trapped in crystals of the wild-type protein. This implies that replacement of just one and two residues leads to two new structures of selecase (thus totaling five), supporting the metamorphic character of this protein.

Summarizing, we have succeeded in identifying and probing for the first time the structural transitions of a natural metamorphic protein with a multifunnel folding energy landscape. Although metamorphic proteins may be encoded by a relevant fraction of all genomes, the lack of bioinformatics and structural approaches to identify them from the sequence restricts their discovery to serendipity. Consistently, to our knowledge 3D structural evidence for their existence has only been published for two natural proteins,[2,13] which just flip between two folds: ubiquitin protein ligase inhibitor Mad2[14] and the chemokine lymphotactin.[15] In selecase, the energy basins are occupied by distinct fully structured and stable states and not by unfolded species or molten globules (Figure 4 and Suppl. Movies 1–3). One conformer is catalytically competent and the others are incompetent but they coexist in equilibrium. These transitions between species are triggered by major rearrangement after residue G^{77} at the NTS–CTS interface, and they mainly affect the CTS. This is consistent with each subdomain corresponding to a distinct folding unit or foldon[16] and the subdomain interface acting as a reversible zipper. The high flexibility of CTS was further verified by computational analysis of local conformational frustration and assessment of interdomain flexibility based on the elastic network model (see SRD and Suppl. Figure 7). In addition, the thermodynamic consistency of interconversion was further backed by the calculated geometric and thermodynamic parameters of the solvation free energy of folding and of dissociation, as well as compactness, for wild-type selecase structures (see SRD and Suppl. Table 8). Owing to inherent flexibility of the CTS, it avoids kinetic trapping in an irreversible misfolded state during conversion between alternate conformers through the protein–protein interactions of oligomeric species as previously suggested for metamorphic proteins.[13] In our view, it is a striking observation that simple dilution/concentration of a sample at room temperature triggers fold switches that cause the repacking of a hydrophobic core and exposure of new binding surfaces, which in turn generate the spontaneous conversion between active monomers and inactive oligomers. This finding indicates that the energy barriers separating the minima are surmountable and that interconversion may proceed without passing through fully unfolded states,[2] as suggested by the finding of largely conserved NTS foldons. Finally, our results also provide the first evidence for a peptidase with a reversible, strictly concentration-dependent reduction of activity at higher concentrations, which is triggered by the sequestering of the competent conformation in incompetent but structured oligomers. This system affords a switch that provides a unique and reversible mechanism of control of catalytic activity in nature.

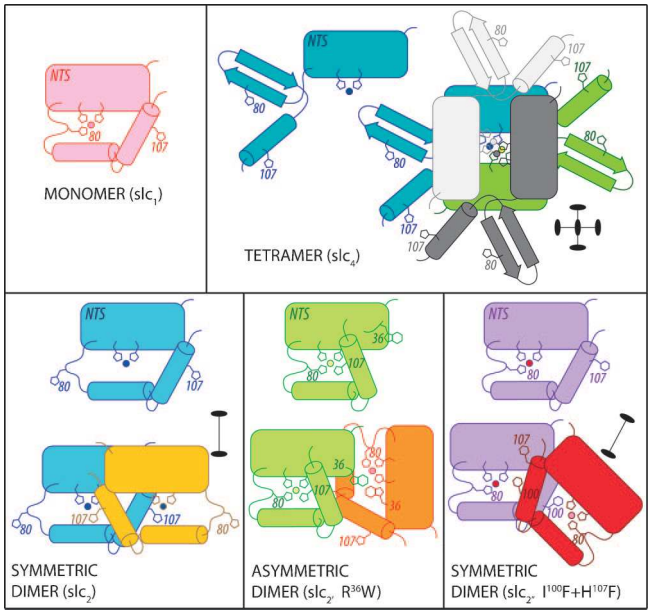


Figure 4. Scheme illustrating the topology of the distinct selecase structures reported. A black ellipse stands for a dyad vertical to the plane, two black ellipses connected by a line stand for a dyad in the plane. NTS, N-terminal subdomain; histidines H^{69}, H^{73}, H^{80}, and H^{107} are shown. In $slc_{2'}$, R^{36} is replaced by tryptophan; in $slc_{2''}$, I^{100} and H^{107} are replaced by phenylalanine.

Received: May 28, 2014
Published online: August 27, 2014

Keywords: metallopeptidases · metamorphic proteins · protein folding · protein structures

[1] C. B. Anfinsen, *Science* **1973**, *181*, 223–230.
[2] A. G. Murzin, *Science* **2008**, *320*, 1725–1726.
[3] S. C. Goodchild, P. M. G. Curmi, L. J. Brown, *Biophys. Rev. Lett.* **2011**, *3*, 143–153.
[4] P. Towler, B. Staker, S. G. Prasad, S. Menon, J. Tang, T. Parsons, D. Ryan, M. Fisher, D. Williams, N. A. Dales, M. A. Patane, M. W. Pantoliano, *J. Biol. Chem.* **2004**, *279*, 17996–18007.

[5] V. Henri, *C. R. Hebd. Seances Acad. Sci.* **1902**, *135*, 916 – 919.
[6] M. López-Pelegrín, N. Cerdà-Costa, F. Martínez-Jiménez, A. Cintas-Pedrola, A. Canals, J. R. Peinado, M. A. Martí-Renom, C. López-Otín, J. L. Arolas, F. X. Gomis-Rüth, *J. Biol. Chem.* **2013**, *288*, 21279 – 21294.
[7] a) L. Polgár in *Proteolytic enzymes—Tools and targets* (Eds.: E. E. Sterchi, W. Stöcker), Springer, Berlin, **1999**, pp. 148 – 166; b) B. W. Matthews, *Acc. Chem. Res.* **1988**, *21*, 333 – 340.
[8] F. X. Gomis-Rüth, T. O. Botelho, W. Bode, *Biochim. Biophys. Acta Proteins Proteomics* **2012**, *1824*, 157 – 163.
[9] a) T. L. Hazlett, E. A. Dennis, *Biochemistry* **1985**, *24*, 6152 – 6158; b) D. H. Fremont, D. H. Anderson, I. A. Wilson, E. A. Dennis, N. H. Xuong, *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 342 – 346.
[10] a) W. Bode, F. X. Gomis-Rüth, R. Huber, R. Zwilling, W. Stöcker, *Nature* **1992**, *358*, 164 – 167; b) F. X. Gomis-Rüth, *J. Biol. Chem.* **2009**, *284*, 15353 – 15357.
[11] a) F. X. Gomis-Rüth, *Crit. Rev. Biochem. Mol. Biol.* **2008**, *43*, 319 – 345; b) N. Cerdà-Costa, F. X. Gomis-Rüth, *Protein Sci.* **2014**, *23*, 123 – 144.
[12] X. Zhou, F. Alber, G. Folkers, G. H. Gonnet, G. Chelvanayagam, *Proteins Struct. Funct. Genet.* **2000**, *41*, 248 – 256.
[13] P. N. Bryan, J. Orban, *Curr. Opin. Struct. Biol.* **2010**, *20*, 482 – 488.
[14] M. Mapelli, L. Massimiliano, S. Santaguida, A. Musacchio, *Cell* **2007**, *131*, 730 – 743.
[15] R. L. Tuinstra, F. C. Peterson, S. Kutlesa, E. S. Elgin, M. A. Kron, B. F. Volkman, *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 5057 – 5562.
[16] A. R. Panchenko, Z. Luthey-Schulten, P. G. Wolynes, *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 2008 – 2013.

Review: Small-angle scattering studies of intrinsically disordered proteins and their complexes

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Current Opinion in Structural Biology

Small-angle scattering studies of intrinsically disordered proteins and their complexes

Tiago N Cordeiro[1], Fátima Herranz-Trillo[1,3], Annika Urbanek[1], Alejandro Estaña[1,2], Juan Cortés[2], Nathalie Sibille[1] and Pau Bernadó[1]

Intrinsically Disordered Proteins (IDPs) perform a broad range of biological functions. Their relevance has motivated intense research activity seeking to characterize their sequence/structure/function relationships. However, the conformational plasticity of these molecules hampers the application of traditional structural approaches, and new tools and concepts are being developed to address the challenges they pose. Small-Angle Scattering (SAS) is a structural biology technique that probes the size and shape of disordered proteins and their complexes with other biomolecules. The low-resolution nature of SAS can be compensated with specially designed computational tools and its combined interpretation with complementary structural information. In this review, we describe recent advances in the application of SAS to disordered proteins and highly flexible complexes and discuss current challenges.

Addresses
[1] Centre de Biochimie Structurale, INSERM U1054, CNRS UMR 5048, Université de Montpellier, 29, rue de Navacelles, 34090 Montpellier, France
[2] LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France
[3] Department of Drug Design and Pharmacology, University of Copenhagen, Universitetsparken 2, 2100 Copenhagen, Denmark

Corresponding author: Bernadó, Pau (pau.bernado@cbs.cnrs.fr)

Current Opinion in Structural Biology 2016, **42**:15–23

This review comes from a themed issue on **Proteins: bridging theory and experiment**

Edited by **Igor N Berezovsky** and **Ugo Bastolla**

http://dx.doi.org/10.1016/j.sbi.2016.10.011

Introduction

In the last two decades, Intrinsically Disordered Proteins or Regions (IDPs/IDRs) have emerged as fundamental molecules in a broad range of crucial biological functions such as cell signaling, regulation, and homeostasis [1,2,3••]. Due to their lack of a permanent secondary and tertiary structure, IDPs and IDRs are highly plastic and have the capacity to perform specialized functions that complement those of their globular (folded) counterparts [4]. Disordered regions, which can finely adapt to the structural and chemical features of their partners, are very well suited for protein–protein interactions and are thus abundant in hub positions of interactomes [5–7].

The importance of disordered proteins in a multitude of biological processes has fostered intense research efforts that seek to unravel the structural bases of their function. Nuclear Magnetic Resonance (NMR) has been the main structural biology technique used to characterize the conformational preferences at residue level, and, therefore, to localize partially structured elements [8,9]. However, a number of structural features related to the overall size and shape of IDPs or their complexes remain elusive to NMR. To study these properties, thereby complementing NMR residue-specific information, Small-Angle Scattering (SAS) of X-rays (SAXS) or Neutrons (SANS) is the most appropriate technique [10–12]. Although SAS is a low-resolution technique, the data obtained is sensitive to large-scale protein fluctuations and the presence of multiple species and/or conformations in solution [13–15]. However, the conversion of SAS properties into structural restraints is challenging due to the enormous conformational variability of IDPs and the ensemble-averaged nature of the experimental data [16]. The quantitative analysis of these data in terms of structure has prompted the development of computational approaches to both model disordered proteins and to use ensembles of conformations to describe the experimental data. Here we highlight the most relevant developments and applications of SAS to IDPs and IDRs, with a special emphasis on the computational strategies required to fully exploit the data in order to achieve biologically insightful information.

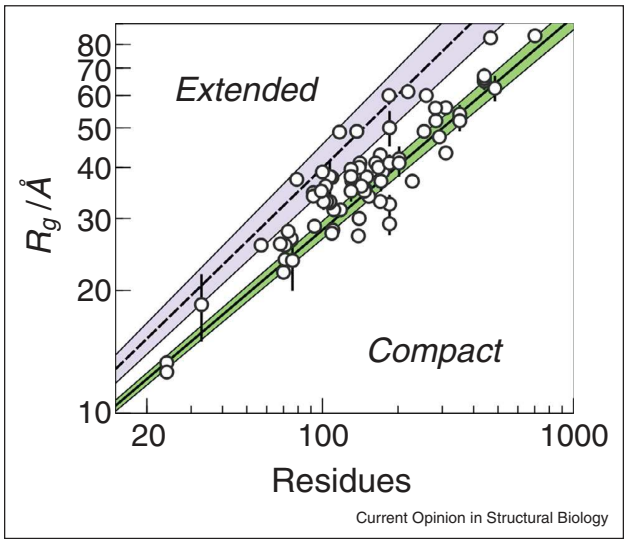
Structural models of IDPs and their experimental validation

For disordered proteins, the structural insights gained from overall SAS parameters, such as the radius of gyration, R_g, the pairwise intramolecular distance distribution, $p(r)$, and the maximum intramolecular distance, D_{max}, are limited. Neither these parameters nor the traditional Kratky representation, $I(s)s^2$ versus s where $I(s)$ represents the scattering intensity and s the momentum transfer, which qualitatively report on the compactness of biomolecules in solution, directly account for the ensemble

nature of disordered proteins. In order to fully exploit the structural and dynamic information encoded in SAS data, it is necessary to use realistic three-dimensional (3D) models. However, the generation of conformational ensembles of disordered proteins is extremely challenging, mainly because of the flat energy landscape and the large number of local minima separated by low-energy barriers [17]. The most popular methods to generate 3D models of IDPs are based on residue-specific conformational landscapes derived from large databases of crystallographic structures [18,19,20•]. However, the main limitation of these approaches is the absence of sequence context information, thereby precluding the prediction of transiently formed secondary structure elements or the presence of long-range interactions between distant regions of the protein. Accurate energy models (force-fields) accounting for the interactions within the chain and with the solvent are required to describe these features. The development of specific force-fields to study conformational fluctuations in disordered proteins is a very active field of research [21–24]. Molecular Dynamics (MD) or Monte-Carlo (MC) simulations, when an appropriate energy description is provided, are suitable methods to correctly sample the conformational space of IDPs. However, the high-dimensionality and the breadth of the energy landscape hamper exhaustive exploration of this space. Replica Exchange MD (REMD) [25,26], which exchanges conformations between parallel simulations running at multiple temperatures, or Multiscale Enhanced Sampling (MSES) [27], which couples temperature and Hamiltonian replica exchange, have been proposed to enhance the conformational exploration of MD methods. The performance of MD-based methods can also be improved by the inclusion of experimental data to delimit the exploration to the most relevant regions of the conformational space [28–30].

The quality of computational models of disordered proteins is normally validated using experimental data. The R_g derived from the low-angle region of SAXS curves or from the $p(r)$ function is an excellent probe of the overall size of a particle in solution. R_g compilations have been extensively used to validate models of denatured and natively disordered proteins through Flory's relationship, which correlates the R_g observed with the number residues of the chain [31,14]. The compilation of the R_gs from 76 IDPs (Figure 1) reveals that these proteins are more compact than chemically denatured ones. It has been shown that denatured proteins present an enhanced sampling of extended conformations, probably due to the interaction of the protein with chemical agents [32]. Importantly, deviations from the expected R_g values for canonical random-coil behavior, which is represented by the green line in Figure 1, indicate the presence of structural features that modify the overall size of the particle in solution towards more extended or more compact (Figure 1). The extendedness detected using this analysis for several Tau protein constructs has been linked to the presence of secondary structural elements probed by NMR [29]. These structural properties can be more thoroughly examined when the complete SAXS curve is used to validate the ensemble models of peptides [33] or proteins [19,34,35].

Figure 1



R_g values from 76 IDPs as a function of the number of residues of the protein are plotted in Log–Log scale. Only proteins lacking a permanent secondary or tertiary structure were considered for the compilation. Proteins with ordered domains, molten globules, or denatured proteins were not considered. Straight lines correspond to Flory's relationships parametrized for denatured proteins using experimental data (purple-dashed) [31] and IDPs using computational ensembles calculated with Flexible-Meccano (green-solid) [32]. Colored bands correspond to uncertainty of the parametrization for both models. Some IDPs contain local structural features and consequently they are globally more extended or more compact than expected for a random coil. These structural features, even if transient, can be manifested in the experimental R_g.

Ensemble approaches

In the last decade, ensemble methods have become highly popular to structurally characterize disordered proteins. Guided by experimental data, these methods aim to derive accurate ensemble models of flexible proteins. Several strategies that apply these methods to SAS data have been reported: Ensemble Optimization Method (EOM) [36,37]; Minimal Ensemble Search (MES) [38]; Basis-Set Supported SAXS (BSS-SAXS) [39]; Maximum Occurrence (MAX-Occ) [40]; Ensemble Refinement of SAXS (EROS) [41]; Broad Ensemble Generator with Re-weighting (BEGR) [42]; and Bayesian Ensemble SAXS (BE-SAXS) [43]. These methods share a common strategy that consists of the following three consecutive steps: (i) computational generation of a large ensemble that describes the conformational landscape of

the protein; (ii) calculation of the theoretical SAXS curves from the individual conformations; and (iii) use of a multiparametric optimization method to select a sub-ensemble of conformations that collectively describe the experimental profile. Despite the common strategy, these approaches present distinct features in the three steps. Readers are referred to the original articles for detailed descriptions. The availability of ensemble methods has transformed the study of flexible proteins by SAS. Ensemble methods provide a description in terms of the statistical distributions of structural parameters or conformations that is revolutionary with respect to traditional analyses based on averaged parameters extracted from raw data. Using this power, structural perturbations exerted by temperature [44•,45], buffer composition [46], or mutations [47] have been monitored in terms of ensembles of conformations.

Despite the popularity of ensemble methods, several aspects are still under debate. The most relevant ones are the use of discrete descriptions for entities that probe an astronomical number of conformations, and the statistical significance of ensembles derived from data containing a very limited amount of information. The strategies described use distinct philosophies to address these issues, including the search for the minimum number of conformations to describe the data [37,38], the representation of the optimal solution as a distribution of low-resolution structural parameters such as R_g or D_{max} [36], and the application of Bayesian statistics [39,43] or maximum entropy approaches [41]. Regardless of the strategy used to derive an ensemble of conformations compatible with the experimental data, one must be careful on the structural interpretation of the final solution. The optimized ensemble is a representation of the behavior of the protein in solution and not the exact enumeration of the conformations adopted by the protein. Consequently, the final ensemble can only be used to derive structural features that describe the protein. Importantly, the nature of these features depends on the experimental data used to derive the model. If only SAS data have been used, then an assessment of the degree of flexibility, and the size and shape distributions sampled by the protein can be obtained from the ensemble. Conversely, conformational preferences at residue level can be extracted if NMR information probing structure in a residue-specific manner is used along the refinement.

Enriching the definition of conformational ensembles of IDPs with complementary information

The definition of protein ensembles derived from SAS data using ensemble methods is limited to the overall structure and the space sampled by the protein in solution. Although this is an important improvement with respect to classical approaches, several crucial features, such as the localization of secondary structural elements or compact regions, remain elusive using this approach. Considerable research efforts have been channeled into enriching the resolution of the resulting ensemble with complementary information.

NMR is the only technique that can provide atomic-resolution information on IDPs and, consequently, it is the most common method applied in combination with SAS [48]. NMR is highly versatile and can measure multiple observables reporting on protein structure and dynamics [49]. Concretely, information reporting on the backbone conformational preferences at residue level can be probed by means of time-averaged and ensemble-averaged chemical-shifts (CSs), J-couplings and Residual-Dipolar Couplings (RDCs). NMR can also probe long-range interactions within a protein chain or in protein complexes through Paramagnetic Relaxation Enhancement (PRE) experiments. In these experiments, a stable radical or a paramagnetic metal is introduced in a specific position of the chain, and the spatially close atoms can be identified by a decrease in their signal intensity that is proportional to the distance.

The best manner to exploit the complementarity between NMR and SAS is to integrate the experimental data into the same refinement protocol. The programs ENSEMBLE [50,51] and ASTEROIDS [52] derive ensembles of disordered proteins by collectively describing SAXS curves, in addition to several NMR observables. These powerful approaches seek to find the appropriate way to combine data with very different information content while avoiding overfitting. In a pioneering study, ensembles of Tau and α-synuclein were determined by combining SAXS with multiple backbone CS, RDC, and PRE datasets [53••]. Those authors addressed the optimal combination of experimental data and the overfitting problem with extensive cross-validation tests that substantiated conformational bias in the aggregation-nucleation regions for both proteins.

Other structural techniques such as single molecule Fluorescence Resonance Energy Transfer (smFRET) [54] and Electron Paramagnetic Resonance (EPR) [55,56••] have been combined with SAXS to study large and flexible complexes. Recent developments in Mass Spectrometry (MS) offer novel sources of structural information [57]. Ion Mobility Spectrometry (IMS) can capture, in a similar way to SAS, the overall properties of conformational ensembles of disordered proteins. However, a recent study comparing IMS and SAXS data for some IDPs suggests that the conformations sampled in solution and in gas-phase are not equivalent [58]. Hydrogen/Deuterium Exchange MS (HDX/MS) probes structural elements in proteins by identifying regions that are protected from the exchange with solvent protons [57]. The availability of fast HDX/MS methods enables the

exploration of secondary structural elements in IDPs and localizing their interaction sites with globular partners [59]. In a recent study HDX/MS information was combined with SAXS to study the calcium-induced structure formation in RD, a protein hosting repeated regions able to bind this cation [60].

The structural definition of a SAXS derived ensemble model can also be enriched by the simultaneous analysis of curves measured for multiple deletion mutants of the same IDP [36]. When applied to two different isoforms of Tau protein, this approach identified the repeat region of the protein as the origin of distinct global rearrangements of its flanking regions [61].

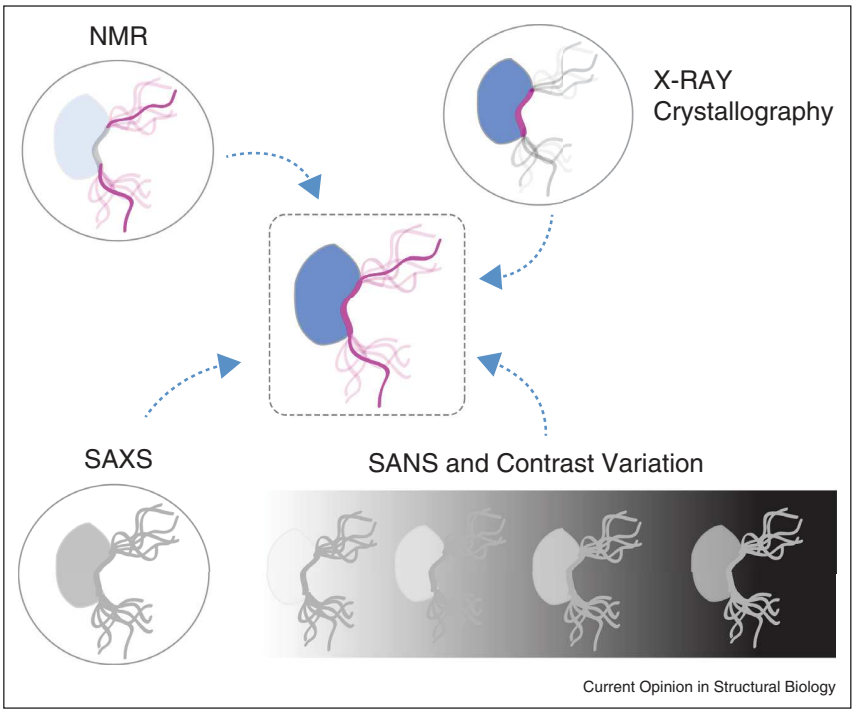
The large toolbox of structural techniques that can probe distinct structural features of IDPs will result in a better understanding on their structure–function relationship. In this regard, the future development of robust and reliable ways to integrate biophysical measurements in ensemble approaches is imperative when addressing complex biomolecular entities such as IDPs and their complexes.

Disordered proteins in complexes

The biological function of many IDPs is manifested when they recognize their biological folded partners [5]. This recognition frequently involves linear motifs of the disordered chain, which, upon binding, adopt relatively fixed conformations while the rest of the IDP remains flexible [62].

The relevance of protein–protein complexes involving disordered partners has promoted growing interest in unraveling their structural characterization, with the aim to understand the bases of their biological activity. This structural characterization is complex and poses multiple challenges to traditional structural biology methods. SAXS has emerged as a valuable alternative. However, overall structural parameters or *ab initio* reconstructions derived from SAXS curves cannot capture the inherent plasticity of these complexes [63,64•,65]. Hybrid (or integrative) methods that combine information from multiple techniques, thus exploiting their individual strengths, are the most appropriate approaches to study highly flexible complexes [66]. In this context, it is important to describe how different structural biology

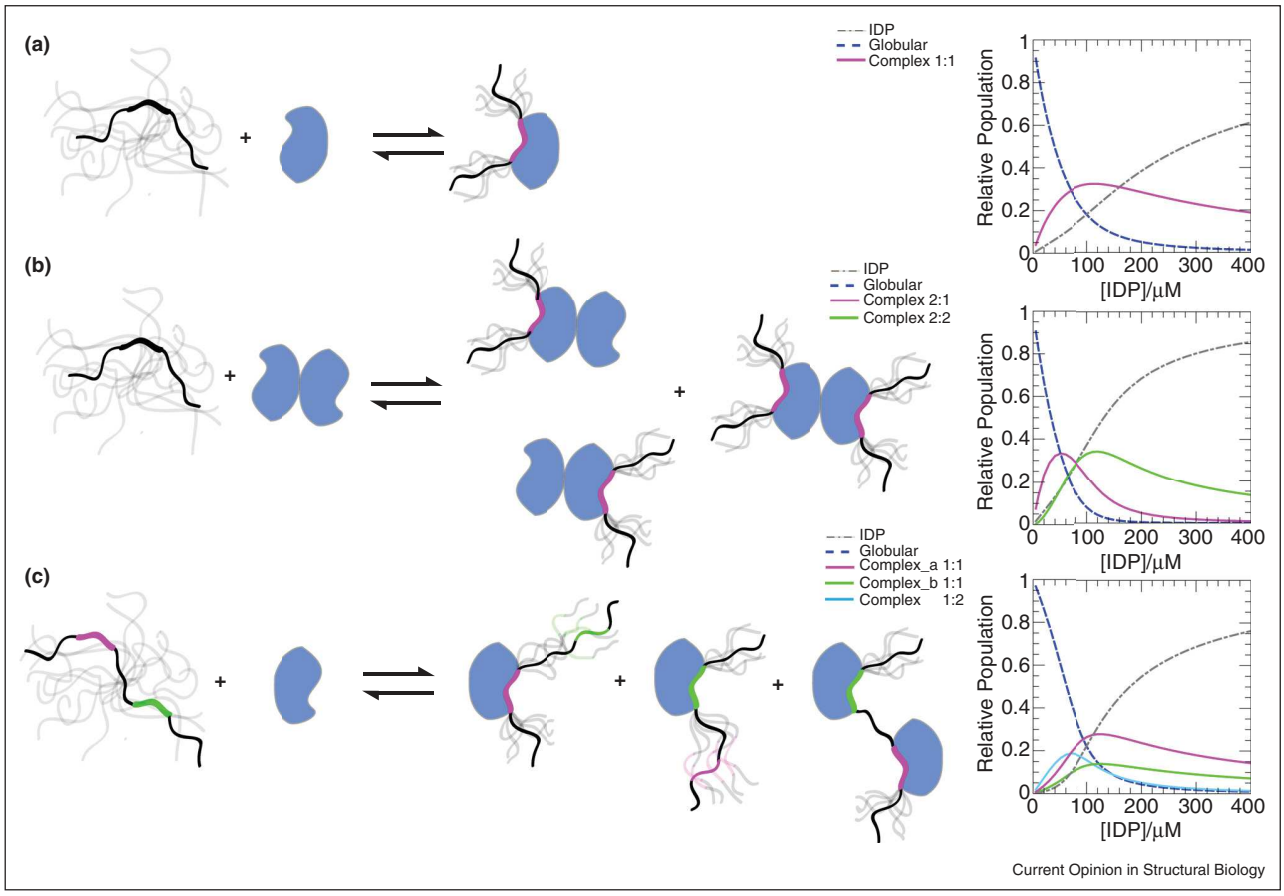
Figure 2



Cartoons representing the structural sensitivity of NMR, X-ray crystallography, and SAS for a complex involving a disordered protein (central cartoon). NMR normally probes the flexible regions of these complexes while the globular partner and the interacting region remain invisible. Crystallography provides detailed information of the interacting region of the complex but not for the flexible parts. SAXS probes the complete ensemble, although the details cannot be assessed due to its inherent low-resolution. SANS, through contrast variation experiments, can probe independently both partners in the context of the complex depending on the deuteration level of the partners and the D_2O/H_2O of the buffer. SAS is an ideal tool to integrate NMR and crystallographic information to build complete structural and dynamic models of disordered biomolecular complexes.

techniques probe complexes involving IDPs (Figure 2). Due to the dynamic nature of the interaction and the distinct hydrodynamic properties of the globular and disordered parts of the complex, NMR generally detects only those regions that remain flexible upon binding. Although not general, it is sometimes possible to crystallize the globular partner in the presence of a small peptide corresponding to the interacting region of the IDP. Therefore, X-ray crystallography provides an atomic resolution picture of the interacting regions that is complementary to NMR since the two techniques probe non-overlapping parts of the same entity [67]. Conversely, SAXS probes the complete assembly and can be used to integrate the information from both NMR and X-ray crystallography. If one of the partners is deuterated, contrast variation SANS experiments can be performed and the individual components of the assembly can be alternatively highlighted depending on the D_2O/H_2O ratio of the buffer. The power of combining multiple techniques is exemplified in the study of the interaction of the Vesicular Stomatitis Virus (VSV) nucleoprotein (N^0) and the dimeric phosphoprotein (P), a high-affinity complex that precludes the oligomerization of N^0 *in vivo* [68••]. Using EOM, the authors simultaneously fitted one SAXS curve and four SANS curves measured at different contrast levels for the complex of N^0 with deuterated P

Figure 3



Examples of polydisperse scenarios that can occur in low-affinity complexes involving an IDP and a globular partner. **(a)** Both proteins have a single binding site. The complex is in equilibrium with the free forms of both proteins. **(b)** The globular partner is a dimer and has two identical binding sites. The free forms are in equilibrium with three possible complexes recognizing one or two binding sites of the globular partner. Due to the symmetry of the dimer, the two singly bound complexes are however indistinguishable by SAS. **(c)** The IDP presents two similar binding sites (pink and green). The free forms are in equilibrium with two 1:1 complexes using a distinct IDP interacting site to bind the globular partner, and a complex where the IDP simultaneously interacts with two globular partners. On the right part of the figure, three panels are displayed representing the molar fraction of each species along a simulated titration experiment for each scenario. These populations were computed assuming a fixed concentration of the globular partner, [globular] = 100 μM, and increasing concentrations of IDP, [IDP], from 1 μM to 400 μM. A common dissociation constant K_d = 20 μM was used for scenarios A and B, in panel C the two IDP binding sites, pink and green, display a K_d = 20 μM and 40 μM, respectively. These panels exemplify the inherent polydispersity of moderate affinity complexes, and how multiple titration experiments will probe differently the species present and their relative populations.

protein. The additional information provided by the distinct contribution of the two proteins in the SANS experiments notably improved the description of the conformational properties of the complex.

In many cases, the conformational mobility of the interacting region of the IDP is reduced (or frozen) upon binding to the biological partner. There is an entropic cost associated with this rigidification that often leads to low-affinity to moderate-affinity complexes ($K_d > 1$ μM) [62]. The structural modulation of the affinity is key to achieving tunable responses to external signals, thereby explaining the prevalent role of disordered proteins in signaling processes [2,3]. In the concentration range normally used in SAXS experiments, the complex is in equilibrium with the free forms of the two partners, thereby giving rise to population-weighted averaged SAXS curves (Figure 3a). This scenario can be even more complex if one or both of the partners have multiple equivalent or similar binding sites (Figure 3b,c). In this case, the polydispersity of the mixture increases as a result of the presence of several complexes with distinct stoichiometries.

The interpretation of SAS data from polydisperse samples is challenging [69]. Although the coupling of SAXS to Size-Exclusion Chromatography (SEC-SAXS) can, in some instances, separate the components of the mixture, there are multiple examples where the coexistence of multiple species is unavoidable. In these circumstances and with the aim to isolate the contribution of the individual species within complex mixtures, analytical approaches have been developed to decompose large SAXS titration datasets [70,71]. This decomposition is easier when prior structural knowledge of the species is used for the analysis [69]. However, to apply this strategy to low-affinity flexible complexes, accurate conformational descriptions of all species in the free and bound forms are mandatory. The analysis of SAS data measured in samples with different relative concentrations of both partners seems the most appropriate strategy to enrich the information content in order to structurally characterize these extremely challenging scenarios (Figure 3).

Conclusions and outlook

During the last decade, SAS has been added to the toolbox of techniques used to study conformational fluctuations in proteins. This dynamic revolution of SAS is linked to the development of computational tools able to describe the conformational landscape of biomolecules and ensemble approaches with the capacity to interpret SAS data in terms of structural variability. These computational tools, which use chemical and structural knowledge of biomolecules, partially compensate for the limited amount of information coded in a SAS curve. Therefore, the capacity to fully exploit the structural information held in SAS data will necessarily be linked to the development of more advanced and precise computational approaches with specially developed force-fields. This notion is especially applicable to IDPs and IDRs, which populate a huge number of conformational states. For these proteins, SAS can be enriched with complementary information obtained by NMR, smFRET, EPR, or MS, and integrated into a common ensemble model embedding structure and dynamics. A particularly challenging subclass of IDPs is that containing Low-Complexity Regions (LCRs), which are involved in multitude of biological processes and are related to severe pathologies. LCRs are unusually simple protein sequences with a strong amino acid composition bias. The resulting similarity of chemical environments within their sequence hampers their structural characterization by NMR. SAS can be a valuable alternative through which to study this important but structurally neglected family of proteins [72–74].

The function of multitude of IDPs is determined by their interaction with biomolecular partners to form assemblies, which, in many cases, are of low to moderate affinity. The capacity of SAS to probe the size and shape of particles in solution places this technique in a unique position to address these polydisperse scenarios. A case in point is the fibrillation process that several IDPs undergo to form amyloids, which are linked to severe diseases. The decomposition of time-dependent SAXS datasets has been successfully used to characterize intermediate oligomeric forms [75•,76], thereby validating SAXS as a practical tool for this purpose.

The need to understand the mechanisms underlying complex cellular processes and recent technical and conceptual advances in structural biology techniques across the board have prompted researchers to tackle challenging systems that were inaccessible some years ago. Many of these systems are inherently dynamic and/or polydisperse and can be exquisitely probed by SAS. As a consequence, we anticipate that SAS will take on greater relevance in hybrid approaches where its unique information will be synergistically integrated with data from multiple sources to deliver accurate structural and dynamic models of disordered proteins and their complexes.

Conflict of interest statement

The authors declare no conflict of interest.

Acknowledgements

This work was supported by the ERC-CoG chemREPEAT, SPIN-HD–Chaires d'Excellence 2011 from the *Agence National de Recherche* (ANR), ATIP-Avenir, and the French Infrastructure for Integrated Structural Biology (FRISBI – ANR-10-INSB-05-01) to PB. FHT is supported by INSERM and the Sapere Aude Programme SAFIR of the University of Copenhagen. AU is supported by a grant from the *Fondation pour la Recherche Médicale*.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z: **Intrinsic disorder and protein function**. *Biochemistry* 2002, **41**:6573-6582.

2. Wright PE, Dyson HJ: **Intrinsically disordered proteins in cellular signalling and regulation**. *Nat Rev Mol Cell Biol* 2015, **16**:18-29.

3. •• Csizmok V, Follis AV, Kriwacki RW, Forman-Kay JD: **Dynamic protein interaction networks and new structural paradigms in signaling**. *Chem Rev* 2016, **116**:6424-6462.
Excellent review with a very complete list of references on the unique mechanisms used by disordered proteins to perform very specific functions in signaling processes. A description is provided of the present knowledge on the emerging field of phase separation induced by the interaction of IDRs with RNA.

4. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Uversky VN, Obradovic Z: **Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions**. *J Proteome Res* 2007, **6**:1882-1898.

5. Tompa P, Schad E, Tantos A, Kalmar L: **Intrinsically disordered proteins: emerging interaction specialists**. *Curr Opin Struct Biol* 2015, **35**:49-59.

6. Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN: **Flexible nets. The roles of intrinsic disorder in protein interaction networks**. *FEBS J* 2005, **272**:5129-5148.

7. Kim PM, Sboner A, Xia Y, Gerstein M: **The role of disorder in interaction networks: a structural analysis**. *Mol Systems Biol* 2008, **4**:179.

8. Dyson HJ, Wright PE: **Unfolded proteins and protein folding studied by NMR**. *Chem Rev* 2004, **104**:3607-3622.

9. Jensen MR, Ruigrok RWH, Blackledge M: **Describing intrinsically disordered proteins at atomic resolution by NMR**. *Curr Opin Struct Biol* 2013, **23**:426-435.

10. Feigin LA, Svergun DI: *Structure Analysis by Small-angle X-ray and Neutron Scattering*. Plenum Press; 1987.

11. Putnam CD, Hammel M, Hura GL, Tainer JA: **X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution**. *Quart Rev Biophys* 2007, **40**:191-285.

12. Jacques DA, Trewhella J: **Small-angle scattering for structural for structural biology-expanding the frontier while avoiding the pitfalls**. *Protein Sci* 2010, **19**:642-657.

13. Doniach S: **Changes in biomolecular conformation seen by small angle X-ray scattering**. *Chem Rev* 2001, **101**:1763-1778.

14. Bernadó P, Svergun DI: **Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering**. *Mol Biosyst* 2012, **8**:151-167.

15. Receveur-Brechot V, Durand D: **How random are intrinsically disordered proteins? A small angle scattering perspective**. *Curr Protein Pept Sci* 2012, **13**:55-75.

16. Bernadó P, Blackledge M: **Structural biology: proteins in dynamic equilibrium**. *Nature* 2010, **468**:1046-1048.

17. Zhou H-X: **Polymer models of protein stability, folding, and interactions**. *Biochemistry* 2004, **43**:2141-2154.

18. Jha AK, Colubri A, Freed KF, Sosnick TR: **Statistical coil model of the unfolded state: resolving the reconciliation problem**. *Proc Natl Acad Sci USA* 2005, **102**:13099-13104.

19. Bernadó P, Blanchard L, Timmins P, Marion D, Ruigrok RW, Blackledge M: **A structural model for unfolded proteins from residual dipolar couplings and small-angle X-ray scattering**. *Proc Natl Acad Sci USA* 2005, **102**:17002-17007.

20. • Ozenne V, Bauer F, Salmon L, Huang J, Jensen MR, Segard S, Bernadó P, Charavay C, Blackledge M: **Flexible-meccano: A tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables**. *Bioinformatics* 2012, **28**:1463-1470.
Description of Flexible-Meccano. This software computes ensembles of IDPs based on the conformational sampling found in coil regions of crystallographic structures. The program computes averaged RDCs and PREs from the ensembles, and provides scripts to add side chains, and to compute CSs and SAXS data.

21. Vitalis A, Pappu RV: **ABSINTH: a new continuum solvation model for simulations of polypeptides in aqueous solutions**. *J Comput Chem* 2009, **30**:673-699.

22. Best RB, Zheng W, Mittal J: **Balanced protein–water interactions improve properties of disordered proteins and non-specific protein association**. *J Chem Theory Comput* 2014, **10**:5113-5124.

23. Henriques J, Cragnell C, Skepö M: **Molecular dynamics simulations of intrinsically disordered proteins: force field evaluation and comparison with experiment**. *J Chem Theory Comput* 2015, **11**:3420-3431.

24. Mercadante D, Milles S, Fuertes G, Svergun DI, Lemke EA, Gräter F: **Kirkwood-buff approach rescues overcollapse of a disordered protein in canonical protein force fields**. *J Phys Chem B* 2015, **119**:7975-7984.

25. Chebaro Y, Ballard AJ, Chakraborty D, Wales DJ: **Intrinsically disordered energy landscapes**. *Sci Rep* 2015, **5**:10386.

26. Zerze GH, Miller CM, Granata D, Mittal J: **Free energy surface of an intrinsically disordered protein: comparison between temperature replica exchange molecular dynamics and bias-exchange metadynamics**. *J Chem Theory Comput* 2015, **11**:2776-2782.

27. Lee KH, Chen J: **Multiscale enhanced sampling of intrinsically disordered protein conformations**. *J Comput Chem* 2016, **37**:550-557.

28. Dedmon M, Lindorff-Larsen K, Christodoulou J, Vendruscolo M, Dobson CM: **Mapping long-range interactions in alpha-synuclein using spin-label NMR and ensemble molecular dynamics simulations**. *J Am Chem Soc* 2005, **127**:476-477.

29. Mukrasch MD, Markwick P, Biernat J, Bergen Mv, Bernadó P, Griesinger C, Mandelkow E, Zweckstetter M, Blackledge M: **Highly populated turn conformations in natively unfolded tau protein identified from residual dipolar couplings and molecular simulation**. *J Am Chem Soc* 2007, **129**:5235-5243.

30. Wu K-P, Weinstock DS, Narayanan C, Levy RM, Baum J: **Structural reorganization of α-synuclein at low pH observed by NMR and REMD simulations**. *J Mol Biol* 2009, **391**:784-796.

31. Kohn JE, Millett IS, Jacob J, Zagrovic B, Dillon TM, Cingel N, Dothager RS, Seifert S, Thiyagarajan P, Sosnick TR *et al.*: **Random-coil behavior and the dimensions of chemically unfolded proteins**. *Proc Natl Acad Sci USA* 2004, **101**:12491-12496.

32. Bernadó P, Blackledge M: **A self-consistent description of the conformational behavior of chemically denatured proteins from NMR and small angle scattering**. *Biophys J* 2009, **97**:2839-2845.

33. Zagrovic B, Lipfert J, Sorin EJ, Millett IS, van Gunsteren WF, Doniach S, Pande VS: **Unusual compactness of a polyproline type II structure**. *Proc Natl Acad Sci USA* 2005, **102**:11698-11703.

34. Wells M, Tidow H, Rutherford TJ, Markwick P, Jensen MR, Mylonas E, Svergun DI, Blackledge M, Fersht AR: **Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain**. *Proc Natl Acad Sci USA* 2008, **105**:5762-5767.

35. De Biasio A, Ibáñez de Opakua A, Cordeiro TN, Villate M, Merino N, Sibille N, Lelli M, Diercks T, Bernadó P, Blanco FJ: **p15PAF is an intrinsically disordered protein with nonrandom structural preferences at sites of interaction with other proteins**. *Biophys J* 2014, **106**:865-874.

36. Bernadó P, Mylonas E, Petoukhov MV, Blackledge M, Svergun DI: **Structural characterization of flexible proteins using small-angle X-ray scattering**. *J Am Chem Soc* 2007, **129**:5656-5664.

37. Tria G, Mertens HD, Kachala M, Svergun DI: **Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering**. *IUCrJ* 2015, **2**:207-217.

38. Pelikan M, Hura GL, Hammel M: **Structure and flexibility within proteins as identified through small angle X-ray scattering**. *Gen Physiol Biophys* 2009, **28**:174-189.

39. Yang S, Blachowicz L, Makowski L, Roux B: **Multidomain assembled states of Hck tyrosine kinase in solution**. *Proc Natl Acad Sci USA* 2010, **107**:15757-15762.

40. Bertini I, Giachetti A, Luchinat C, Parigi G, Petoukhov MV, Pierattelli R, Ravera E, Svergun DI: **Conformational space of flexible biological macromolecules from average data**. *J Am Chem Soc* 2010, **132**:13553-13558.

41. Rozycki B, Kim YC, Hummer G: **SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions**. *Structure* 2011, **19**:109-116.

42. Daughdrill GW, Kashtanov S, Stancik A, Hill SE, Helms G, Muschol M, Receveur-Bréchot V, Ytreberg FM: **Understanding the structural ensembles of a highly extended disordered protein**. *Mol Biosyst* 2012, **8**:308-319.

43. Antonov LD, Olsson S, Boomsma W, Hamelryck T: **Bayesian inference of protein ensembles from SAXS data**. *Phys Chem Chem Phys* 2016, **18**:5832-5838.

44. • Shkumatov AV, Chinnathambi S, Mandelkow E, Svergun DI: **Structural memory of natively unfolded tau protein detected by small-angle X-ray scattering**. *Proteins* 2011, **79**:2122-2131.
Interesting article reporting on the conformational changes experienced by Tau protein when submitted to temperature jumps. Although the authors do not provide a precise explanation on the origin of this phenomenon, it reflects that there are probably structural features in IDPs that have not been unveiled yet.

45. Kjaergaard M, Nørholm AB, Hendus-Altenburger R, Pedersen SF, Poulsen FM, Kragelund BB: **Temperature-dependent structural changes in intrinsically disordered proteins: formation of alpha-helices or loss of polyproline II?** *Protein Sci* 2010, **19**:1555-1564.

46. Leyrat C, Jensen MR, Ribeiro EA Jr, Gérard FC, Ruigrok RW, Blackledge M, Jamin M: **The N(0)-binding region of the vesicular stomatitis virus phosphoprotein is globally disordered but contains transient α-helices**. *Protein Sci* 2011, **20**:542-556.

47. Stott K, Watson M, Howe FS, Grossmann JG, Thomas JO: **Tail-mediated collapse of HMGB1 is dynamic and occurs via differential binding of the acidic tail to the A and B domains**. *J Mol Biol* 2010, **403**:706-722.

48. Sibille N, Bernadó P: **Structural characterization of intrinsically disordered proteins by the combined use of NMR and SAXS**. *Biochem Soc Trans* 2012, **40**:955-962.

49. Jensen MR, Zweckstetter M, Huang JR, Blackledge M: **Exploring free-energy landscapes of intrinsically disordered proteins at atomic resolution using NMR spectroscopy**. *Chem Rev* 2014, **114**:6632-6660.

50. Marsh JA, Neale C, Jack FE, Choy WY, Lee AY, Crowhurst KA, Forman-Kay JD: **Improved structural characterizations of the drkN SH3 domain unfolded state suggest a compact ensemble with native-like and non-native structure**. *J Mol Biol* 2007, **367**:1494-1510.

51. Krzeminski M, Marsh JA, Neale C, Choy WY, Forman-Kay JD: **Characterization of disordered proteins with ENSEMBLE**. *Bioinformatics* 2013, **29**:398-399.

52. Jensen MR, Houben K, Lescop E, Blanchard L, Ruigrok RW, Blackledge M: **Quantitative conformational analysis of partially folded proteins from residual dipolar couplings: application to the molecular recognition element of Sendai virus nucleoprotein**. *J Am Chem Soc* 2008, **130**:8055-8061.

53. •• Schwalbe M, Ozenne V, Bibow S, Jaremko M, Jaremko L, Gajda M, Jensen MR, Biernat J, Becker S, Mandelkow E *et al.*: **Predictive atomic resolution descriptions of intrinsically disordered hTau40 and α-synuclein in solution from NMR and small angle scattering**. *Structure* 2014, **22**:238-249.
Seminal study on the structural properties of α-synuclein and Tau. Using ASTEROIDS the authors interpreted complete CS, RDC and PRE datasets in combination with SAXS data to deliver conformational ensembles of both proteins. The most relevant part part of the article is the extensive cross-validation analyses that demonstrate the accuracy of the structural models.

54. Delaforge E, Milles S, Bouvignies G, Bouvier D, Boivin S, Salvi N, Maurin D, Martel A, Round A, Lemke EA *et al.*: **Large-scale conformational dynamics control H5N1 influenza polymerase PB2 binding to importin α**. *J Am Chem Soc* 2015, **137**:15122-15134.

55. Boura E, Rózycki B, Herrick DZ, Chung HS, Vecer J, Eaton WA, Cafiso DS, Hummer G, Hurley JH: **Solution structure of the ESCRT-I complex by small-angle X-ray scattering, EPR, and FRET spectroscopy**. *Proc Natl Acad Sci USA* 2011, **108**:9437-9442.

56. •• Boura E, Różycki B, Chung HS, Herrick DZ, Canagarajah B, Cafiso DS, Eaton WA, Hummer G, Hurley JH: **Solution structure of the ESCRT-I and -II supercomplex: implications for membrane budding and scission**. *Structure* 2012, **20**:874-886.
In this study the ensemble description of the flexible complex formed by ESCRT-I and II is obtained. The authors integrate SAXS, smFRET and EPR data to define the complex. A protocol is introduced in order to find the minimal number of components in the ensemble with the capacity to properly describe the three sources of data.

57. Konermann L, Vahidi S, Sowole MA: **Mass spectrometry methods for studying structure and dynamics of biological macromolecules**. *Anal Chem* 2014, **86**:213-232.

58. Borysik AJ, Kovacs D, Guharoy M, Tompa P: **Ensemble methods enable a new definition for the solution to gas-phase transfer of intrinsically disordered proteins**. *J Am Chem Soc* 2015, **137**:13807-13817.

59. Keppel TR, Weis DD: **Mapping residual structure in intrinsically disordered proteins at residue resolution using millisecond hydrogen/deuterium exchange and residue averaging**. *J Am Soc Mass Spectrom* 2015, **26**:547-554.

60. O'Brien DP, Hernandez B, Durand D, Hourdel V, Sotomayor-Pérez AC, Vachette P, Ghomi M, Chamot-Rooke J, Ladant D, Brier S, Chenal A: **Structural models of intrinsically disordered and calcium-bound folded states of a protein adapted for secretion**. *Sci Rep* 2015, **5**:14223.

61. Mylonas E, Hascher A, Bernadó P, Blackledge M, Mandelkow E, Svergun DI: **Domain conformation of tau protein studied by solution small-angle X-ray scattering**. *Biochemistry* 2008, **47**:10345-10353.

62. Sharma R, Raduly Z, Miskei M, Fuxreiter M: **Fuzzy complexes: specific binding without complete folding**. *FEBS Lett* 2015, **589**:2533-2542.

63. Shell SS, Putnam CD, Kolodner RD: **The N terminus of *Saccharomyces cerevisiae* Msh6 is an unstructured tether to PCNA**. *Mol Cell* 2007, **26**:565-578.

64. • Rochel N, Ciesielski F, Godet J, Moman E, Roessle M, Peluso-Iltis C, Moulin M, Haertlein M, Callow P, Mély Y *et al.*: **Common architecture of nuclear receptor heterodimers on DNA direct repeat elements with different spacings**. *Nat Struct Mol Biol* 2011, **18**:564-570.
Using a combination of SAXS, SANS and smFRET the authors studied the structure of three hormonal nuclear receptor heterodimers in complex with cognate dsDNA and intrinsically disordered co-activators. Although using *ab initio* reconstructions, the asymmetric singly-bound nature of the complex with the co-activator is demonstrated. Excellent study that highlights the power of SAXS/SANS to characterize complex biomolecular entities.

65. Devarakonda S, Gupta K, Chalmers MJ, Hunt JF, Griffin PR, Van Duyne GD, Spiegelman BM: **Disorder-to-order transition underlies the structural basis for the assembly of a transcriptionally active PGC-1α/ERRγ complex**. *Proc Natl Acad Sci USA* 2011, **108**:18678-18683.

66. Różycki B, Boura E: **Large, dynamic, multi-protein complexes: a challenge for structural biology**. *J Phys Condens Matter* 2014, **26**:463103.

67. De Biasio A, de Opakua AI, Mortuza GB, Molina R, Cordeiro TN, Castillo F, Villate M, Merino N, Delgado S, Gil-Cartón D *et al.*: **Structure of p15(PAF)-PCNA complex and implications for clamp sliding during DNA replication and repair**. *Nat Commun* 2015, **6**:6439.

68. •• Yabukarski F, Leyrat C, Martinez N, Communie G, Ivanov I, Ribeiro EA Jr, Buisson M, Gerard FC, Bourhis JM, Jensen MR *et al.*: **Ensemble structure of the highly flexible complex formed between vesicular stomatitis virus unassembled nucleoprotein and its phosphoprotein chaperone**. *J Mol Biol* 2016, **428**:2671-2694.
In this study the ensemble structure of the viral complex between the nucleocapsid protein N^0 and the phosphoprotein P is determined. The ensemble description is performed using the EOM approach. The novelty of the study is the simultaneous description of the SAXS data of the complex with SANS curves measured at four different contrast levels. This is the first study that profits from the rich information from contrast variation in the context of highly flexible biomolecular complexes using ensemble approaches.

69. Tuukkanen AT, Svergun DI: **Weak protein–ligand interactions studied by small-angle X-ray scattering**. *FEBS J* 2014, **281**:1974-1987.

70. Blobel J, Bernadó P, Svergun DI, Tauler R, Pons M: **Low-resolution structures of transient protein–protein complexes using small-angle X-ray scattering**. *J Am Chem Soc* 2009, **131**:4378-4386.

71. Chandola H, Williamson TE, Craig BA, Friedman AM, Bailey-Kellogg C: **Stoichiometries and affinities of interacting proteins from concentration series of solution scattering data: decomposition by least squares and quadratic optimization**. *J Appl Crystallogr* 2014, **47**:899-914.

72. Greving I, Dicko C, Terry A, Callow P, Vollrath F: **Small angle neutron scattering of native and reconstituted silk fibroin**. *Soft Matter* 2010, **6**:4389.

73. Boze H, Marlin T, Durand D, Pérez J, Vernhet A, Canon F, Sarni-Manchado P, Cheynier V, Cabane B: **Proline-rich salivary proteins have extended conformations**. *Biophys J* 2010, **99**:656-665.

74. Owens GE, New DM, West AP, Bjorkman PJ: **Anti-PolyQ antibodies recognize a short PolyQ stretch in both normal and mutant huntingtin exon 1**. *J Mol Biol* 2015, **427**:2507-2519.

75. • Vestergaard B, Groenning M, Roessle M, Kastrup JS, van de Weert M, Flink JM, Frokjaer S, Gajhede M, Svergun DI: **A helical structural nucleus is the primary elongating unit of insulin amyloid fibrils**. *PLoS Biol* 2007, **5**:1089-1097.
Pioneering study on the characterization of fibrillating proteins using SAXS. The fibrillation of insulin is monitored by SAXS in a time-dependent manner. The resulting curves are the population-weigted averages of all species co-existing in solution. In an arduous procedure, the species-pure curves for the three main components of the mixtures were decomposed allowing their structural characterization including their molecular weight, oligomerization state, and 3D arrangement.

76. Giehm L, Svergun DI, Otzen DE, Vestergaard B: **Low-resolution structure of a vesicle disrupting alpha-synuclein oligomer that accumulates during fibrillation**. *Proc Natl Acad Sci USA* 2011, **108**:3246-3251.