# Methods for Hybrid Modeling of Solution Scattering Data and their Applications

Dissertation zur Erlangung
des naturwissenschaftlichen Doktorgrades
der Bayerischen Julius-Maximilians-Universität Würzburg

vorgelegt von
Alexander V. Shkumatov
aus
Minsk, Belarus

Würzburg 2011

Eingereicht am:

Mitglieder der Promotionskommission:
Vorsitzender:
- 1. Gutachter: Dr. habil. Matthias Wilmanns
- 2. Gutachter: Prof. Dr. Thomas Dandekar

Tag des Promotionskolloquiums:
Doktorurkunde ausgehändigt am:

# Contents

# Abstract

Small-angle X-ray scattering (SAXS) is a universal low-resolution method to study proteins in solution and to analyze structural changes in response to variations of conditions (pH, temperature, ionic strength etc). SAXS is hardly limited by the particle size, being applicable to the smallest proteins and to huge macromolecular machines like ribosomes and viruses. SAXS experiments are usually fast and require a moderate amount of purified material. Traditionally, SAXS is employed to study the size and shape of globular proteins, but recent developments have made it possible to quantitatively characterize the structure and structural transitions of metastable systems, e.g. partially or completely unfolded proteins.

In the absence of complementary information, low-resolution macromolecular shapes can be reconstructed *ab initio* and overall characteristics of the systems can be extracted. If a high- or low-resolution structure or a predicted model is available, it can be validated against the experimental SAXS data. If the measured sample is polydisperse, the oligomeric state and/or oligomeric composition in solution can be determined. One of the most important approaches for macromolecular complexes is a combined *ab initio*/rigid body modeling, when the structures (either complete or partial) of individual subunits are available and SAXS data is employed to build the entire complex. Moreover, this method can be effectively combined with information from other structural, computational and biochemical methods. All the above approaches are covered in a comprehensive program suite ATSAS for SAXS data analysis, which has been developed at the EMBL-Hamburg.

In order to meet the growing demands of the structural biology community, methods for SAXS data analysis must be further developed. This thesis describes the development of two new modules, RANLOGS and EM2DAM, which became part of ATSAS suite. The former program can be employed for constructing libraries of linkers and loops *de novo* and became a part of a combined *ab initio*/rigid body modeling program CORAL. EM2DAM can be employed to convert electron microscopy maps to bead models, which can be used for modeling or structure validation. Moreover, the programs CRYSOL and CRYSON, for computing X-ray and neutron scattering patterns from atomic models, respectively, were refurbished to work faster and new options were added to them.

Two programs, to be contributed to future releases of the ATSAS package, were also developed. The first program generates a large pool of possible models using rigid body

modeling program SASREF, selects and refines models with lowest discrepancy to experimental SAXS data using a docking program HADDOCK. The second program refines binary protein-protein complexes using the SAXS data and the high-resolution models of unbound subunits. Some results and conclusions from this work are presented here.

The developed approaches detailed in this thesis, together with existing ATSAS modules were additionally employed in a number of collaborative projects. New insights into the "structural memory" of natively unfolded tau protein were gained and supramodular structure of RhoA-specific guanidine nucleotide exchange factor was reconstructed. Moreover, high resolution structures of several hematopoietic cytokine-receptor complexes were validated and re-modeled using the SAXS data. Important information about the oligomeric state of yeast frataxin in solution was derived from the scattering patterns recorded under different conditions and its flexibility was quantitatively characterized using the Ensemble Optimization Method (EOM).

# Zusammenfassung

Röntgenkleinwinkelstreuung (small angle X-ray scattering, SAXS) ist eine fundamentale niedrigauflösende Methode zur Untersuchung von Proteinen in Lösung und Analyse von Strukturänderungen unter verschiedenen Bedingungen (pH, Temperatur, Ionenstärke, usw.). SAXS ist nicht durch die Teilchengröße begrenzt und die Anwendbarkeit reicht von kleinsten Proteinen bis hin zu großen makromolekularen Maschinen, wie Ribosomen und Viren. SAXS-Experimente sind normalerweise schnell durchzuführen und erfordern eine relativ geringe Menge gereinigten Materials. SAXS wird hauptsächlich eingesetzt, um Größe und Form der globulärer Proteine zu studieren. Die neuesten Entwicklungen ermöglichen jedoch auch die Untersuchung und quantitative Charakterisierung metastabiler Systeme, wie teilweise oder vollständig ungefaltete Proteine.

Für die SAXS-Datenanalyse existiert das umfassende Programmpaket ATSAS, welches am EMBL-Hamburg entwickelt wurde. Es ermöglicht die *de novo* Modellierung der Proteinform mit niedriger Auflösung, wenn keine ergänzende Information über die *dreidimensionale* Struktur vorhanden ist. Des weiteren können diverse Gesamteigenschaften des untersuchten Systems berechnet werden. Wenn ein hoch oder niedrig aufgelöstes strukturell bestimmtes oder vorgesagtes Modell vorhanden ist, kann es gegen experimentellen SAXS Daten validiert werden. Wenn die Probe polydispers ist, kann der oligomere Zustand und/oder der oligomere Zusammensetzung in Lösung bestimmt werden. Einer der wichtigsten Ansätze für SAXS Untersuchungen an makromolekularen Komplexen ist die kombinierte *ab initio*/Starrkörper-Modellierung, wenn entweder komplette oder partielle Strukturen der einzelnen Untereinheiten zusammen mit SAXS Daten benutzt werden, um daraus den gesamten Komplex zu konstruieren. Außerdem kann diese Methode mit Informationen von anderen strukturellen, rechnerischen und biochemischen Methoden effektiv kombiniert werden.

Um den Anwendungsbereich von SAXS in der Strukturbiologie zu erweitern, müssen Methoden für die SAXS-Datenanalyse weiter entwickelt werden. Im Rahmen dieser Arbeit wurden zwei neue Module, RANLOGS und EM2DAM, entwickelt und zur ATSAS Programmsuite hinzugefügt. Ersteres kann eingesetzt werden, um eine Bibliothek verknüpfender Polypeptidketten (*linkers*) und -schleifen (*loops*) *de novo* aufzubauen und ist bereits ein Teil des Programms CORAL zur kombinierten *ab initio*/Starrkörper-Modellierung. EM2DAM kann eingesetzt werden, um Elektronenmikroskopie-Dichtekarten

in Kugelmodelle umzuwandeln, welche für die Modellierung oder Struktur-Validierung benutzt werden können. Außerdem wurden die Programme CRYSOL und CRYSON zur Berechnung von Röntgenstrahl- beziehungsweise Neutronenstreumuster aus Atommodellen erweitert, um die Berechnung zu beschleunigen und neue Optionen einzubauen.

Zwei weitere Programme, die noch nicht Teil des ATSAS Pakets sind, wurden entwickelt. Das erste ist ein Programm, das mögliche Proteinmodelle von Komplexen unter Verwendung des SAXS Starrkörper-Modellierung-Programms SASREF erstellt. Dann werden Modelle zu experimentellen SAXS-Daten angepasst, ausgewählt und verfeinert unter Verwendung des Protein-Protein-Docking-Programms HADDOCK. Das zweite Programm verfeinert binäre Protein-Protein-Komplexe unter Verwendung von SAXS-Daten sowie hochaufgelöster Modelle der ungebundenen Untereinheiten. Im Folgenden werden die einige Ergebnisse dargestellt und diskutiert.

Die entwickelten Methoden wurden zusammen mit den vorhandenen ATSAS-Modulen im Rahmen von Kollaborationsprojekte eingesetzt. So war es möglich, neue Einblicke in das „strukturelle Gedächtnis“ des natürlicherweise ungefalteten Protein *tau* zu bekommen und die supramodulare Struktur eines RhoA-spezifischen Guanidinnukleotid-Austauschfaktors zu rekonstruieren. Außerdem wurden hoch aufgelöste Strukturen einiger blutbildender Cytokin-Empfänger-Komplexe unter Verwendung von SAXS Daten validiert und verfeinert. Wichtige Informationen über den oligomeren Zustand von Hefe-Frataxin in Lösung wurden aus den unter verschiedenen experimentelle Bedingungen gemessenen Streumustern abgeleitet, und seine Flexibilität wurde quantitativ unter Verwendung der Ensemble-Optimierungs-Methode (EOM) ermittelt.

# List of figures

# List of tables

# List of abbreviations

| | |
|---|---|
| SAXS | Small-Angle X-ray Scattering |
| EOM | Ensemble Optimization Method |
| NMR | Nuclear Magnetic Resonance |
| SANS | Small-Angle Neutron Scattering |
| SR | Synchrotron Radiation |
| SAS | Small-Angle Scattering |
| 3D | Three-Dimensional |
| ms | Millisecond |
| EM | Electron Microscopy |
| kDa | kilodalton |
| mDa | megadalton |
| MW | Molecular Weight |
| $R_g$ | Radius of gyration |
| PDB | Protein Data Bank |
| $D_{max}$ | Maximum particle diameter |
| $V_p$ | Hydrated particle volume |
| IDPs | Intrinsically Disordered Proteins |
| HADDOCK | High Ambiguity Driven DOCKing |
| NMADREFS | Normal Mode Analysis Driven REFinement with SAXS constraint |
| EN | Elastic Network |
| NM | Normal Modes |

| | |
|---|---|
| Å | Ångstroms |
| CD | Circular Dichroism |
| SEC | Size Exclusion Chromatography |
| AUC | Analytical Ultracentrifugation |
| DLS | Dynamic Light Scattering |
| min | minutes |
| SA | Simulated Annealing |
| NSD | Normalized Spatial Discrepancy |
| RANCH | RANdom CHain |
| GAJOE | Genetic Algorithm Judging Optimization of Ensembles |
| v. | version |
| RANLOGS | RANdom LOop Generator and Sorter |
| EM2DAM | Electron Microscopy density map To Dummy Atom Model |
| CASP | Critical Assessment in Structure Prediction |
| HPC | High Performace Computing |
| RMSD | Root Mean Square Deviation |
| i-RMSD | Interface Root Mean Square Deviation |
| NMA | Normal Mode Analysis |
| MD | Molecular Dynamics |
| ID | Identifier |
| LEA | Late Embryogenesis Abundant |
| EST | Expressed Sequence Tag |

| | |
|---|---|
| NRDB | Non-Redundant Data Base |
| COG | Cluster of Orthologous Groups |
| Figure | Fig. |
| aa | amino acid |
| GEF | Guanine Nucleotide Exchange Factor |
| RTKIII | The class-III receptor tyrosine kinase |
| AML | Acute Myeloid Leukemia |
| HSC | Hematopoietic Stem Cells |
| PDGFRα/β | Prototypic platelet-Derived Growth Factor Receptors |
| Flt3 | Fms-Like Tyrosine kinase receptor 3 |
| TM | Transmembrane Helix |
| JM | Juxtamembrane Domain |
| VEGFR | Vascular Endothelial Growth Factor Receptors |
| CSF-1R | Colony-Stimulating Factor 1 Receptor |
| Ig | Immunoglobulin |
| ITD | Internal Tandem Duplication |
| ITC | Isothermal Titration Calorimetry |
| SCF | Stem Cell Factor |
| FRDA | Friedreich's ataxia |
| ROS | reactive oxygen species |
| nm | nanometer |
| PRG | PDZRhoGEF |

| | |
|---|---|
| GEFs | Guanine Nucleotide Exchange Factors |
| LARG | Leukemia Associated RhoGEF |
| p115 | p115RhoGEF |
| CH | calponin-homology |
| SH3 | SRC-homology 3 |
| ASEC | analytical size exclusion chromatography |
| DH | Dbl-homology |
| DXMS | hydrogen/deuterium exchange mass spectrometry |
| PDZ | PSD-95/Disc-large/ZO-1 |
| PH | Pleckstrin-Homology |
| RGS | Regulators of G-protein signaling |
| RGSL | Regulators of G-protein signaling-like |
| rTEV | recombinant Tobacco Etch Virus |
| SLS | Static Light Scattering |
| PHFs | Paired Helical Filaments |
| CNS | Central Nervous System |
| FRET | Fluorescence Resonance Energy Transfer |
| SDS-PAGE | Sodium Dodecyl Sulfate-PolyAcrylamide Gel Electrophoresis |

Chapter 1

# Introduction

SAXS is becoming an increasingly important method in structural biology to study the solution structure of individual proteins and macromolecular assemblies under a variety of conditions, ranging from near physiological to highly denaturing. Unlike Nuclear Magnetic Resonance (NMR) or X-ray crystallography, SAXS is a low resolution method, providing information with respect to size and shape of particle systems (Petoukhov and Svergun 2007; Jacques and Trewhella 2010; Mertens and Svergun 2010). This technique allows rapid structural characterization of macromolecules practically without size limitations. Moreover, one can study the structure of partially or completely unfolded proteins, like tau protein involved in Alzheimer's disease (Mylonas, Hascher et al. 2008; Shkumatov, Chinnathambi et al. 2011). SAXS data can be complemented by experimental evidence from other methods or *in silico* predictions (Petoukhov and Svergun 2007). Especially useful is the application of SAXS for macromolecular complexes, where the quaternary structure of a complex can be reconstructed from the high resolution models of individual subunits.

## 1.1 SAXS history

The first X-ray applications date back to the late 1930s when the main principles of SAXS were developed in the fundamental work of Guinier (Guinier 1939), following his studies of metallic alloys. In the first monograph on SAXS by Guinier and Fournet (Guinier and Fournet 1955) it was already demonstrated that the method yields not only information on the sizes and shapes of particles but also on the internal structure of disordered and partially ordered systems.

In the 1960s, the method became increasingly important in the study of biological macromolecules in solution as it allowed to obtain low-resolution structural information on the overall shape and internal structure in the absence of crystals. A breakthrough in SAXS and small-angle neutron scattering (SANS) experiments came in the 1970s, thanks to the availability of synchrotron radiation (SR) and neutron sources, the latter paving the way for contrast variation by solvent exchange of $H_2O$ for $D_2O$ (Ibel and Stuhrmann 1975) and specific deuteration (Engelman, Moore et al. 1975) methods. It was realized that scattering studies of solutions provide, for a minimal investment in time and effort, useful insights into the structure of non-crystalline biochemical systems. Moreover, SAXS/SANS also made real-time investigations of intermolecular interactions possible, including the assembly and

large-scale conformational changes in macromolecular complexes (Akiyama, Takahashi et al. 2002; Oka, Inoue et al. 2005; Vestergaard, Groenning et al. 2007; Matsuo, Iwamoto et al. 2010)

The main challenge of small-angle scattering (SAS) as a structural method is to extract information about the three-dimensional (3D) structure of an object from the one-dimensional experimental data. In the past, only overall particle parameters (e.g. volume, mass and radius of gyration) of the macromolecules were directly determined from the experimental data, whereas the analysis in terms of 3D models was limited to simple geometrical bodies (e.g. ellipsoids, cylinders, etc) or was performed on an *ad hoc* trial-and-error basis (Glatter and Kratky 1982; Feigin and Svergun 1987). Electron microscopy (EM) was often used as a constraint in building consensus models (Pilz, Glatter et al. 1972; Tardieu and Vachette 1982). The disadvantage of a number of the earlier SAXS studies was that the structural conclusions were drawn from a number of overall parameters or trial-and-error models.

The 1990s brought a breakthrough in SAXS/SANS data analysis methods, allowing reliable *ab initio* shape and domain structure determination and detailed modeling of macromolecular complexes using rigid body refinement. This progress was accompanied by further advances in instrumentation, and time resolutions down to the sub-ms (millisecond) were achieved on third generation SR sources, in the study of protein and nucleic acid folding (Svergun and Koch 2003).

## 1.2 SAXS theory

SAXS is a general method for the structure analysis of materials, including biological macromolecules in solution (Feigin and Svergun 1987). SAXS is applicable to structures varying significantly in size: from small proteins and polypeptides to macromolecular complexes and machineries which can all be measured with modern instrumentation under near native conditions (Mertens and Svergun 2010). This method allows to not only study the low resolution structure but also to analyze structural changes in response to variation of external conditions (pH, temperature, light, addition of ligand, cofactors, denaturant, etc).

In a SAXS experiment, the macromolecular solution is exposed to a collimated beam of X-ray photons (either from a synchrotron or a laboratory source) and the scattering intensity of elastic scattering is recorded as a function of the scattering angle. Dilute aqueous solutions of proteins, nucleic acids or other macromolecules give rise to an isotropic scattering intensity which depends on the modulus of the momentum transfer s ($s = 4\pi\sin(\theta)/\lambda$, where $2\theta$ is the angle between the incident and scattered beam):

$$\mathbf{I(s) = \langle I(s)\rangle_{\Omega} = \langle A(s)A^{*}(s)\rangle_{\Omega}} \qquad \text{equation 1-1}$$

where the scattering amplitude A(s) is a Fourier transformation of the particle electron density. The scattering vector $s = (s, \Omega) = k_1 - k_0$, where $k_0$ and $k_1$ are the wave vectors of the incoming and the scattered waves, respectively, and the scattering intensity is averaged over all orientations $s = (s, \Omega)$. After subtraction of the solvent scattering, the intensity I(s) is proportional to the scattering of a single particle averaged over all orientations (Mertens and Svergun 2010).

Several overall parameters can be obtained directly from the scattering curves of macromolecular solutions enabling fast sample characterization. These parameters include the molecular weight (*MW*), radius of gyration ($R_g$), maximum particle diameter ($D_{max}$), and the hydrated particle volume ($V_p$) (Mertens and Svergun 2010). Furthermore, computational approaches to retrieve low resolution 3D structural models of proteins and complexes, either *ab initio* or by rigid body modeling, are well established and now widely used in structural biology (Svergun and Koch 2002; Petoukhov, Konarev et al. 2007; Putnam, Hammel et al. 2007).

Importantly, unlike most other structural methods, SAXS is applicable to flexible and metastable systems. One can characterize equilibrium and non-equilibrium mixtures and monitor kinetic processes such as (dis)assembly (Akiyama, Nohara et al. 2008) and (un)folding (Doniach 2001). In particular, SAXS can be employed to quantitatively characterize the overall structure and structural transitions of partially or completely unfolded proteins, including intrinsically disordered proteins (IDPs), an extremely interesting and important class of metastable objects.

## 1.3 Characterization of Intrinsically Disordered Proteins (IDPs) using SAXS

The scattering profile measured from a solution of a metastable system (e.g. a flexible system such as an IDP) reflects an average of the large number of conformations that the protein adopts in solution. Traditionally, Kratky plots ($I(s)\cdot s^2$ as a function of $s$) have been used to identify disordered states and distinguish them from globular ones (Doniach 2001). The scattering intensity of a globular protein at high angles behaves approximately as $1/s^4$, yielding a bell-shaped Kratky plot with a well defined maximum. Conversely, an ideal Gaussian chain has a $1/s^2$ dependence of $I(s)$, forming a plateau at large $s$ values. For unfolded proteins, the Kratky plot also presents a plateau instead of the maximum observed for the globular proteins, and the plateau is followed by a monotonic increase at larger $s$ (see **Fig. 1-1**).

When studying IDPs, SAXS patterns are normally analyzed in combination with other experimental techniques and bioinformatics tools to identify unstructured regions. Circular dichroism (CD), NMR, fluorescence spectroscopy, and hydrodynamic techniques such as size exclusion chromatography (SEC), analytical ultracentrifugation (AUC), or dynamic light scattering (DLS) have been used in combination with SAXS to identify proteins as IDPs (Gast, Damaschun et al. 1995; Gazi, Bastaki et al. 2008; Paz, Zeev-Ben-Mordehai et al. 2008).

IDPs are often involved in signaling processes and must change their global properties upon environmental modifications within the cell in order to bind to, or detach from their natural partners. SAXS is a suitable tool to rapidly monitor structural changes in proteins upon such environmental modifications. The changes, associated with varying pH (Konno 1997), ionic strength (Munishkina, Fink et al. 2004), temperature (Kjaergaard, Norholm et al. 2010; Shkumatov, Chinnathambi et al. 2011), presence of specific ions (He, Ramachandran et al. 2005), phosphorylation (He, Ramachandran et al. 2005), or additives (Hong, Fink et al. 2008), must induce global size variation in IDPs in order to be monitored by SAXS. These global alterations are reflected again in the changes of the apparent $R_g$, $D_{max}$, and the appearance of the Kratky plots.

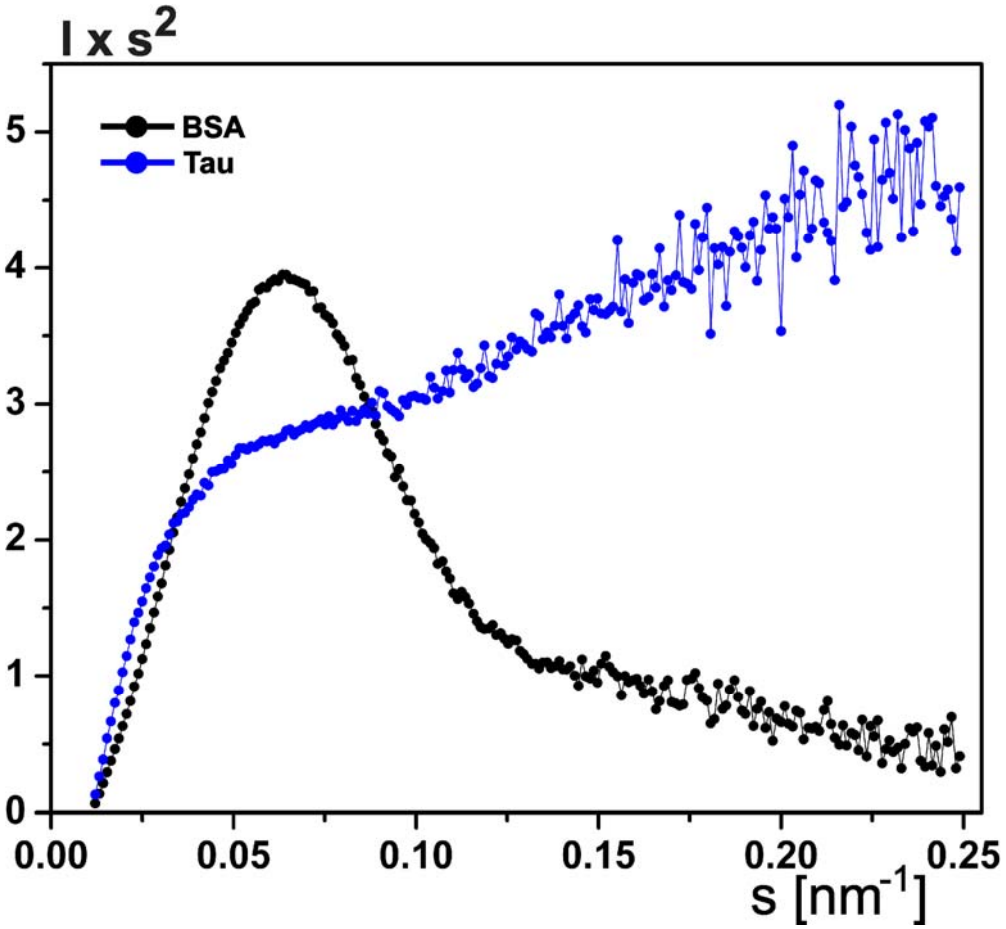



Figure 1-1. Comparison of Kratky plots for bovine serum albumin (*black circle;* MW 66kDa) and a natively unfolded Tau protein (*blue circle;* MW 45kDa).

CD combined with NMR studies of natively unfolded proteins have identified structural changes upon heating that result from the disordering of α-helices and polyproline II (PPII) structure, the combined effect of which is to promote (local or global) compaction (Kjaergaard, Norholm et al. 2010). The interpretation of the changes observed by CD spectroscopy is ambiguous. This is caused by the fact that structural changes in different segments may have spectroscopic contributions that cancel each other's signal. Specifically, folding of α-helices and unfolding of PPII structures give rise to a similar change in the CD spectrum (Kjaergaard, Norholm et al. 2010). Recently temperature-induced structural changes in IDPs have been reexamined using three different proteins: ACTR, NHE1 and Spd1 (Kjaergaard, Norholm et al. 2010). From a combined analysis using CD spectroscopy, SAXS, NMR chemical shift and peptide mimics, the bulk of the observed change in ellipticity with temperature is suggested to be due to a redistribution of the statistical coil ensemble, where PPII-like conformations are lost with increasing temperature. The

transiently formed α-helices, however, loose helical structures at increased temperatures (Kjaergaard, Norholm et al. 2010).

Recent novel data analysis methods make it possible to describe the flexibility of IDP ensembles in solution based on SAXS data (Bernado and Svergun 2008). Bernado et al (Bernado, Mylonas et al. 2007) proposed an approach allowing for the coexistence of different protein conformations contributing to the average experimental scattering pattern. The EOM approach has become very popular in the studies of metastable systems such as multidomain proteins with flexible linkers and IDPs, and a number of successful applications have already been reported by different groups (Petoukhov, Vicente et al. 2008; Bernado, Modig et al. 2010; Kjaergaard, Norholm et al. 2010).

## 1.4 Scope of the thesis

Main objectives of this work were to further develop methods for SAXS data analysis and create new tools utilizing information from bioinformatics predictors and experimental SAXS data. The enhanced methods were to be employed in the studies of solution structures of biological macromolecules with different levels of flexibility in near native conditions.

Chapter 2 describes currently available methods for SAXS data analysis and interpretation, ranging from data reduction to 3D modeling and assessment of flexibility. Applications of these methods to biological systems with different levels of flexibility are presented in Chapter 5.

Chapter 3 describes improvements and new developments in the ATSAS program suite (Petoukhov, Konarev et al. 2007). These include new faster versions of CRYSOL and CRYSON for evaluating the solution scattering from macromolecules with known atomic structure and fitting to experimental scattering from SAXS and SANS, respectively. The main improvements as well as new options are described.

RANLOGS and EM2DAM are the new tools developed by the author and recently included into the latest release of ATSAS package. Their functionality and applications are described in detail.

The last part of Chapter 3 includes results of simulated data analysis. Molecular weight (MW) estimation is crucial to confirm the oligomeric state of the samples studied by SAXS.

In order to find correlations between MW and the values of excluded volume obtained by SAXS, systematic calculations were performed on various different 3D structures from the Protein Data Bank (PDB) ranging in size and shape. These structures were also used to test the new program AUTOPOROD from the ATSAS package.

Chapter 4 describes novel developments utilizing bioinformatics predictors.

The first two subchapters describe post-processing, using CRYSOL and experimental SAXS data, of a large number of models generated either *de novo* or by simulating complexation.

The program ROSETTA (Bonneau, Strauss et al. 2002) was used to generate structural models of lysozyme. The latter were screened using CRYSOL and SAXS experimental data. ROSETTA is currently one of the most successful *de novo* structure prediction algorithms, which can accurately model 3D structures of small proteins.

The second subchapter describes the application of HADDOCK (High Ambiguity Driven DOCKing) to simulate the complexation of two or more proteins (Dominguez, Boelens et al. 2003). HADDOCK is a data-driven docking program that can be used together with experimental data or with bioinformatics interface prediction. Given that the high-resolution structures of the bound or free molecules are known, docking provides an alternative approach to distance constraint calculations or X-ray crystallography to predict the structure of a complex.

In the third subchapter an automatic method, combining SAXS rigid body modeling with subsequent docking refinement by HADDOCK is presented, as well as some results.

In the forth subchapter a novel method to perform rigid body modeling of protein complexes accounting for possible differences between the structures of free and bound subunits, NMADREFS (Normal Mode Analysis Driven REFinement with SAXS constraint), is presented. This method uses a linear combination of low-frequency normal modes (NM) from an elastic network (EN) description of the molecule in an iterative manner to deform the structure optimally to conform to the SAXS scattering curve. Unlike other available methods, NMADREFS samples conformational space of separate (unbound) subunits of a complex using NM while maintaining connectivity, avoiding steric clashes and preserving the interaction interface. The method is being tested on protein-protein docking benchmarks using simulated data. The preliminary results are presented.

The last subchapter covers the bioinformatics predictors, which were used to analyze the genome and proteome of tardigrades. Tardigrades represent an animal phylum with extraordinary resistance to environmental stress. Features of tardigrade specific adaption were identified by sequence and/or pattern search on the web-tool “tardigrade analyzer” co-developed by the author. Different protein clusters and regulatory elements implicated in tardigrade stress adaptations were scrutinized.

## 1.5 Further reading

For a more detailed introduction to SAXS and its application to biological systems, readers are referred to reviews on biological SAXS (Petoukhov and Svergun 2007; Putnam, Hammel et al. 2007; Jacques and Trewhella 2010; Mertens and Svergun 2010) as well as recent publications (Prischi, Konarev et al. 2010; Morgan, Schmidt et al. 2011; Shkumatov, Chinnathambi et al. 2011; Verstraete, Vandriessche et al. 2011).

Chapter 2

# Methods for SAXS data analysis

## 2.1 Data collection and reduction

The protein samples described in chapter 5 were measured at 10°C, except for tau protein, in a concentration range from 1 to 20mg/ml. All measurements were performed using the automated SAXS sample changer (Round, Franke et al. 2008), where the samples are kept in a temperature-controlled sample tray and injected into the independently temperature-controlled measuring cell. All measurements, except for tau protein, were performed under equilibrium temperature conditions, i.e. the temperature in the sample holder and measuring cell was set to 10°C. Tau protein was studied under equilibrium and non-equilibrium temperature conditions. During the non-equilibrium temperature experiments, the measurement cell was tempered to 50 and 10°C, whereas the temperature in the sample holder was set to 10 and 50°C, respectively. At equilibrium temperature conditions, the measurement cell and sample tray were held at the same temperature, either 10 or 50°C.

SAXS data were collected in several experimental sessions on the EMBL X33 beamline of the storage ring DORIS III (DESY, Hamburg). The data were recorded using either a counting Pilatus 1M pixel detector (DECTRIS) or MAR Image Plate detector (345 $mm^2$) at a sample-detector distance of 2.7 m and wavelength of 1.5 Angstroms (Å), covering the range of momentum transfer $0.012 < s < 0.6$ $Å^{-1}$ ( here, $s = 4\pi \sin\theta/\lambda$, where $2\theta$ is the scattering angle). A standard data collection time of 2 min was used for all samples split into time frames to assess and remove effects from radiation damage to the samples. The time frames were processed by the automatic pipeline (Petoukhov, Konarev et al. 2007), yielding radially averaged curves of normalized intensity versus the momentum transfer. The buffer scattering before and after each sample was averaged and used for background subtraction with PRIMUS (Konarev, Volkov et al. 2003).

The forward scattering I(0) and the radii of gyration $R_g$ were evaluated using the Guinier approximation (Guinier 1939) assuming that at very small angles ($s < 1.3/R_g$) the intensity is represented as $I(s) = I(0)\exp(-(sR_g)^2/3)$. These parameters were also computed from the entire scattering patterns using the program GNOM (Svergun 1992), which also provides the distance distribution functions p(r) and the maximum particle dimensions $D_{max}$. The solute MWexp was estimated by comparison of the forward scattering with that from reference solutions of bovine serum albumin (MW 66 kDa). The excluded volume of the hydrated particle (the Porod volume $V_p$) was computed using the Porod invariant (Porod 1982).

Evaluation of the theoretical scattering curves from high-resolution structures and fitting to the experimental scattering data was performed using the program CRYSOL (Svergun, Barberato et al. 1995).

In case of polydisperse systems, the form factors corresponding to individual high resolution structures were calculated using the FFMAKER tool from the ATSAS package (Petoukhov, Konarev et al. 2007). For fitting the observed scattering curves with the weighted combinations of the form-factors the program OLIGOMER (Konarev, Volkov et al. 2003) was used.

## 2.2 *Ab initio* shape reconstruction

Molecular envelopes were calculated using the DAMMIN (Svergun 1999) or DAMMIF (Franke and Svergun 2009) programs, where the overall shape and excluded volume were initially estimated by modeling with P1 symmetry. Where appropriate, symmetry and particle anisometry were imposed to refine the *ab initio* model. DAMMIN and DAMMIF represent the particle shape by an assembly of densely packed beads and employ simulated annealing (SA) to construct a compact interconnected model fitting the experimental data to minimize the discrepancy:

$$\chi = \sqrt{\frac{1}{(N-1)}\sum_{j}\left[\frac{\left(I_{\exp}(S_j) - cI_{calc}(S_j)\right)}{\left(\sigma(S_j)\right)}\right]^2} \qquad \text{equation 2-1}$$

where $N$ is the number of experimental points, $c$ is a scaling factor, $I_{\exp}(s)$, $I_{\text{calc}}(s)$ and $\sigma(s_j)$ are the experimental intensity, the calculated intensity and experimental error at the momentum transfer $s_j$, respectively. Normally, around ten bead models are averaged using the DAMAVER suite (Volkov and Svergun 2003). The mean normalized spatial discrepancy (NSD) can illustrate how similar reconstructed shapes are, as well as identify the most probable solution. NSD value below 1 indicates that shapes are quite similar, NSD from 1 to 1.5 - reasonable solution, NSD > 1.5 shows that two shapes are different.

Higher-resolution *ab initio* models were constructed by the program GASBOR (Svergun, Petoukhov et al. 2001) which models the particle in solution as a protein-like assembly of dummy residues, thus, representing more accurately the internal structure than shapes determined using DAMMIN or DAMMIF. Where appropriate, GASBOR models

were reconstructed with symmetry imposed. Subsequently, GASBOR models were averaged using the DAMAVER suite (Volkov and Svergun 2003), as described above.

## 2.3 Rigid body modeling

A rigid-body modeling was performed with the program SASREF (Petoukhov and Svergun 2005), which uses a SA protocol to search for the optimal positions and orientations of the rigid bodies fitting the experimental data. Where appropriate, symmetry constraints or contact restraints were imposed. If the relative positions of some of the subunits were known from high resolution models, these subunits were fixed in respect to each other as subcomplexes during modeling.

## 2.4 Combined *ab initio* and rigid body modeling

Some of the constructs described in chapter 5 were lacking high resolution structures of N- or C-termini, loops or large linkers due to flexibility. Combined *ab initio*/rigid-body modeling with BUNCH (Petoukhov and Svergun 2005) was employed to reconstruct the complete structures. Starting from a random domain arrangement, BUNCH uses SA to guide the translations and rotations of domains to minimize the discrepancy $\chi$ (equation 2.1) between experimental data and calculated data while maintaining chain connectivity without steric clashes. Missing loops, N- or C-termini or linkers between the individual subunits were modeled using dummy residues. The starting models were generated with PRE_BUNCH. BUNCH either in user or expert mode with default parameters except "DR formfactor multiplier" (individual amino acid specific form factors were used instead of averaged one) and "Cross penalty weight" (increased to 200 to avoid clashes between loops or domains) was used. Where appropriate, contact conditions were imposed. For BUNCH modeling, either single or multiple scattering curves were used. Multiple BUNCH runs were performed to ensure that stable and consistent solution was found.

## 2.5 Flexibility assessment

In cases when sample was expected to be flexible, the SAXS data were analyzed using an EOM, consisting of two separate programs – RANdom CHain (RANCH) and Genetic Algorithm Judging Optimization of Ensembles (GAJOE) (Bernado, Mylonas et al. 2007). EOM assumes coexistence of a number of conformations in solution for a given construct in order to fit the experimental SAXS data. It allows quantitative characterization of the flexibility and analysis of the size distributions of possible conformers. RANCH can be used to generate initial models covering the conformational space of multi-domain proteins with flexible linkers as well as natively unfolded proteins. In case of flexible single chain model, GAJOE with default parameters was used after generation of the random pool with RANCH. In case of symmetric model with flexible parts, starting models were generated with PRE_BUNCH, which consisted of rigid bodies from high resolution models. Each starting model was used to generate independently (with different random seed number) the BUNCH models. The resulting models were edited to contain a single chain only. Each model was then combined with the other single chain models in all possible ways, using COMBINEDPDBS.PL (unpublished tool developed by me), resulting in symmetric models with asymmetric missing parts. Theoretical scattering curves were then calculated for each model using CRYSOL (Svergun, Barberato et al. 1995). An intensities master file was generated using the program ONEFILE2 from the EOM package (Bernado, Mylonas et al. 2007). A Size_list file was created using GAJOE. After generation of the pool and additional files, GAJOE was used to select subsets of protein models (~20). The average experimental scattering was calculated for each subset and fitted to experimental SAXS data. Subsets were selected many times in order to minimize the discrepancy. Multiple runs of GAJOE (10 independent runs; for each GAJOE run, the genetic algorithm process was repeated 50 times using the default parameters of the genetic algorithm, i.e., 1000 generations, 50 ensembles of theoretical curves, 20 curves per ensemble, 10 mutations per ensemble, and 20 crossings per generation) were performed. In order to find the minimal set of models that can describe experimental data, only two or three curves were allowed to be selected per ensemble.

Chapter 3

# Improvements and new developments in the SAXS data analysis program suite (ATSAS)

**New developments in ATSAS program package for small angle scattering data analysis**

*D. Franke, M. Gajda, C. Gorba, P.V. Konarev, H. Mertens, M.V. Petoukhov, A.V. Shkumatov, G. Tria, D.I. Svergun*

Manuscript in preparation. The order of the author list and title are not final.

## 3.1 Improvements and new features in CRYSOL and CRYSON

CRYSOL and CRYSON are programs from ATSAS package, used routinely to validate high resolution structures when analysis of SAXS or SANS data is performed. There are two major steps in both programs, namely, (i) calculation of theoretical intensity and (ii) fitting the experimental curve. The time required for the calculation depends on the number of points, resolution of the theoretical curve (evaluation of theoretical intensity) and the number of experimental points (fitting step). Normally, it takes less than a minute or few minutes to do one calculation, but this also may take up to hours if one calculates the theoretical curve for large objects with highest possible resolution and accuracy. The methods covering the conformational space of macromolecules or simulating complexation can produce large pools of PDB models, which need to be processed and validated using SAXS experimental data. Moreover, with the new X-ray detectors, like PILATUS, for example, the number of experimental points collected increased dramatically (to thousands). The latter may slow down even more the fitting step of CRYSOL. Thus, the fast tools are essential to meet the growing demands in the field of SAS.

In the current subchapter improvements and new options of CRYSOL and CRYSON that were introduced by the author are described.

### 3.1.1 Implementation of novel minimization algorithm and new options in CRYSOL

#### Introduction

CRYSOL is a program for evaluating the solution scattering from macromolecules with known atomic structure and fitting it to experimental scattering curve from SAXS (Svergun, Barberato et al. 1995). As an input one can use PDB file (Berman, Henrick et al. 2003) with X-ray or NMR structure of a protein or a protein-DNA/RNA complex. In general CRYSOL calculates theoretical SAXS intensity, $I(s, p_0, \Delta p)$, from a PDB file using the equations described in (Svergun, Barberato et al. 1995) and, if experimental data is provided, adjusts a few parameters in order to fit experimental curve. A plain grid search is made to find the parameters, which will minimize the discrepancy to the experimental data (chi-square value). A bottleneck step is the curve fitting.

## Methods and results

The new version of CRYSOL (v.2.7) has several modifications compared to older one (v.2.6):

1). The experimental curve is divided into bins and the scattering intensities and errors are recalculated for each bin. The number of bins equals to the total number of points divided by ten if there are more than 1000 points. If number of points is less than 1000, number of bins equals to 200.

2). A constant subtraction is added for finding the best solution in terms of chi-square value. Constant is a sum of intensities over high-angle part of the experimental data divided by number of experimental points in this part. This operation accounts for possible systematic errors due to mismatched buffers in the experimental data. CRYSOL (v.2.7) also allows the user to provide a fixed value for the constant when "Another set of parameters" and "Minimize again with new limits" is chosen.

3). A linear least square minimization with boundaries (Stark and Parker 1995) is used for finding the scaling coefficient and the constant value when fitting the theoretical curve to the experimental data.

4). Additional options. By default CRYSOL discards fields starting with ATOM/HETATM, which have a string 'H' in the $13^{th}$ or $14^{th}$ column. A new option accounting explicitly for hydrogens has been added. Previous versions of CRYSOL were processing all ATOM fields in PDB file (Berman, Henrick et al. 2003). In case of NMR or Normal Mode Analysis (NMA) ensemble one had to preprocess PDB file such that only certain model is present. If the PDB file has several models separated by MODEL and ENDMDL fields, CRYSOL (v.2.7) provides the user a possibility to process only one certain model or all models independently as well as to calculate averaged scattering intensity for an ensemble. Other new options include: automatic constant subtraction, writing the experimental errors to the output .fit file, reading PDB file with an old atom naming conventions (before 2008), process only one chain ID and print version information.

New version of CRYSOL (v.2.7) is 5 to 10 times faster than CRYSOL (v.2.6).

### 3.1.2 Implementation of new minimization algorithm in CRYSON

**Background**

CRYSON is a program for evaluating the solution scattering from macromolecules with known atomic structure and fitting it to experimental scattering curve from SANS (Svergun, Richard et al. 1998). CRYSON calculates theoretical SANS intensity, I(s, p0, Δp), from a PDB file using the equations described in (Svergun, Richard et al. 1998) and, if experimental data is provided, adjusts few parameters in order to fit experimental curve. A plain grid search is made to find the parameters, which will minimize the discrepancy to the experimental data (chi-square value).

CRYSON is similar to CRYSOL. Differences between CRYSOL and CRYSON are (i) the neutron scattering lengths of atoms and atomic groups are used to evaluate scattering amplitude from the particle *in vacuo* (ii) in solution with $D_2O$ fraction $0 < Y < 1$ it is assumed that all hydrogens in hydrophilic groups are replaced by deuteriums with probability $Y$, and those belonging to the main chain *NH* groups with probability $0.9Y$ (iii) theoretical curve is smeared appropriately using resolution function to take the instrumental distortions into account.

**Methods and results**

The new version of CRYSON (v.2.7) has several modifications compared to older one (v.2.6):

(i) A linear least square minimization with boundaries (Stark and Parker 1995) was implemented in new version of CRYSON for finding the scaling coefficient and the constant value when fitting theoretical curve to experimental data.

(ii). Additional options, such as, (1) account for explicit hydrogen (as described for CRYSOL above), (2) writing the experimental errors to the output .fit file, (3) reading PDB files with the old atom naming conventions (before 2008) and (4) print version information were included.

## 3.2 RANLOGS – a tool to generate random loops and linkers *ab initio*

Unlike NMR or X-ray crystallography, SAXS is not restricted by size of the studied protein complex and allows one to conduct experiments in nearly physiological conditions. The rigid body modeling methods can be used to build protein complexes by placing high resolution structures of separate subunits such that scattering pattern from theoretical model fits experimental data. However, linkers between subunits or flexible loops can be missing in high-resolution crystallographic and/or NMR models. In this subchapter RANLOGS (RANdom LOop Generator and Sorter), an automatic tool for generating loops and domains' linkers *de novo* and their sorting, is presented. Unlike the other available tools, RANLOGS does not rely on existing structures in PDB (Bernstein, Koetzle et al. 1977) and can be used to generate loops up to 100 residues in size. RANLOGS can provide a pool of connectors for completing SAXS rigid body models. The linkers/loops library is a part of recently developed rigid body modeling program CORAL, included in the ATSAS package.

**State of the problem**

Inherent flexibility and conformational heterogeneity in proteins often result in the absence of loops and even entire domains in crystallographic or NMR models. Such missing fragments still contribute to the SAS intensity and their probable configuration can be found by fixing the known part of the structure and adding the missing parts to fit the SAS pattern for the entire particle (Svergun and Koch 2002).

Some programs in the ATSAS package (Petoukhov, Konarev et al. 2007) allow one to build missing fragments (Petoukhov, Eady et al. 2002) in X-ray or NMR models, provided SAXS experimental data is available. They use SA protocol (Petoukhov and Svergun 2005) and are time-consuming. Moreover they try to fit the data by a single conformation of the protein.

Available approaches to generate random linkers or loops fall into two main categories: knowledge based and *ab initio* (*de novo*) methods. Knowledge-based approaches try to find a segment of a protein with known 3D structure that fits the stem regions of a loop. The residues preceding and following the loop are called stem residues. Usually, a database search is followed by an evaluation of suitable candidates and an optimization by means of

an energy function (Michalsky, Goede et al. 2003). *Ab initio* methods search for or enumerate the conformations in common, usually based on potentials or scoring functions. Often knowledge-based parts are included, e.g. phi-, psi-maps of known loops (e.g. (Deane and Blundell 2001; Tosatto, Bindewald et al. 2002; Fiser and Sali 2003)).

SAXS-based rigid body modeling approaches rely only on the fraction of the experimental data describing overall shape of the particle. Atomic details in such models might not be correctly represented. The use of the atomic resolution models of the linker and/or loop regions would not improve the model, as the side chains are not contributing significantly to the changes in the shape. Moreover, the usage of the full atom linker/loop representation, including side chains, will slow even more the theoretical scattering calculation and hence rigid body modeling.

Here a program RANLOGS for automatic generation and sorting of loops up to 100 amino acid (aa) in size is reported. RANLOGS generates only C-alpha trace and uses two criteria for checking the models: (i) the distance between the C-alpha atoms must be 3.8 Å (ii) The bond and dihedral angles should fall within the map of allowed angle combinations (Kleywegt 1997). Essentially RANLOGS offers a pool of linkers and loops to select from. It is also possible to pre-calculate loops of different length, which can be inserted between anchor points in rigid-body models.

## Methods

### *Loop generation and sorting*

RANLOGS generates C-alpha traces of the loops. Each residue in the model is placed 3.8 Å from one of the two adjacent residues. The bond/dihedral angles combination is compared to a map of allowed angle combinations. If such combination is allowed, the combination is accepted. The generated loops are sorted by the number of aa into separate directories. The loops of the same length (same number of aa) are sorted according to the distances between the C- and N-termini into bins. Each loop/linker is stored in a separate file with a name starting with the ordial number of the bin and sequential number of the random loop.

### *Usage and output*

RANLOGS is implemented in object-oriented Fortran90. RANLOGS binaries are available for Linux, PC or Mac platforms as a part of ATSAS package. *"ranlogs -help"* command prints all available options. One can either provide all options in the command line or run it interactively by answering the questions, if RANLOGS is executed without any options. It is possible to specify minimum and maximum loop length, number of loops to generate per one loop length (multiplicand for loop length), size of the bin in Ångstroms, minimum number of models in one bin, as well as the maximum number of models to output per bin. The number of loops per one loop length is adjusted automatically for each loop length, i.e. multiplied by the loop length. If not all options are specified, RANLOGS will use default values (minimum loop length - 5, maximum loop length - 100, number of loops - 1000, bin size - 2, minimum number of models in one bin - 10, maximum number of models to output per bin - 20).

```
R A N L O G S    Version 1.1  15/02/2010

        RANdom LOop Generator and Sorter

AUTHOR: Alexander Shkumatov   EMBL HAMBURG

-- Program  started  at    3-Mar-2011   11:51:15      --

Minimum loop length ..................................... : 9
Maximum loop length ..................................... : 10
Number of models to generate ............................ : 1000
Bin size ................................................ : 2
Number of representatives in one bin .................... : 15
Maximum number of models written per bin ................ : 20
Total number of loops generated ......................... : 19000
Loop size -      9  Bin #      4  # of loops:       15
Loop size -      9  Bin #      6  # of loops:      232
Loop size -      9  Bin #      8  # of loops:      597
Loop size -      9  Bin #     10  # of loops:     1093
Loop size -      9  Bin #     12  # of loops:     1539
Loop size -      9  Bin #     14  # of loops:     1846
Loop size -      9  Bin #     16  # of loops:     1763
Loop size -      9  Bin #     18  # of loops:     1179
Loop size -      9  Bin #     20  # of loops:      512
Loop size -      9  Bin #     22  # of loops:      182
Loop size -      9  Bin #     24  # of loops:       40
Loop size -     10  Bin #      6  # of loops:      206
Loop size -     10  Bin #      8  # of loops:      510
Loop size -     10  Bin #     10  # of loops:      997
Loop size -     10  Bin #     12  # of loops:     1398
Loop size -     10  Bin #     14  # of loops:     1766
Loop size -     10  Bin #     16  # of loops:     1870
Loop size -     10  Bin #     18  # of loops:     1606
Loop size -     10  Bin #     20  # of loops:     1006
Loop size -     10  Bin #     22  # of loops:      458
Loop size -     10  Bin #     24  # of loops:      140
Loop size -     10  Bin #     26  # of loops:       30

-- Program  finished  at    3-Mar-2011   11:51:21      --
```

Figure 3-1. Log file created by RANLOGS. Loops with the length of 9 and 10 residues were generated *de novo* and sorted into bins of 2 Å. Each bin was filled only when least 15 models were generated with the distances between N- and C-termini corresponding to the bin size. Each bin was filled with maximum 20 models. For example, bin # 4 (loop size 9) contains models with distances between N- and C-termini ranging from 2 to 4 Å. There were 15 models generated and 15 PDB files written. The bin #4 (loop size 10) was not filled because there were less than 15 models generated with distances ranging from 2 to 4 Å between N- and C-termini. There were in total 232 models created for the bin # 6 (loop size 9) and only 20 PDB files were written.

*Example:*

```
$> ranlogs -m 9 -x 10 -b 2 -r 15 -w 20
```

RANLOGS will generate loops with the length of 9 and 10 residues and sort them into bins of 2 Å. Each bin is only filled when there were least 15 models generated with the distances between N- and C-termini corresponding to the bin size. Each bin is filled with a maximum of 20 models (see also **Fig. 3-1**).

**Conclusions**

RANLOGS is an automatic tool for *de novo* generation and sorting of C-alpha traces of the loops and domain linkers. Unlike other available tools RANLOGS neither relies on PDB (Bernstein, Koetzle et al. 1977) to generate loops, nor aims to generate loops in functional sites (e.g. Modloop (Fiser and Sali 2003)), nor imposes the strict restrictions on loop length (LIP (Michalsky, Goede et al. 2003) and Modloop (Fiser and Sali 2003)). RANLOGS offers a pool of linkers and loops to select from. It is possible to pre-calculate the loops/linkers of different length, which can be inserted between anchor points in rigid-body models. The process of inserting the loops, calculation of their scattering and fitting can be automated and used in the SAXS-based rigid-body modeling, as implemented in a recently developed program CORAL from the ATSAS package.

## 3.3 Validation of low-resolution models: EM2DAM tool

In the past, EM was used to provide helpful constraints when building consensus models based on SAXS data (Pilz, Glatter et al. 1972; Tardieu and Vachette 1982). Nowadays, with the increase in the popularity of hybrid methods, aiming to answer biological questions by applying a range of experimental and/or theoretical techniques, the situation has somewhat changed. The SAXS data is often used for validation and comparison with the EM reconstructions (Vestergaard, Sanyal et al. 2005; Andersen, Becker et al. 2006; Tidow, Melero et al. 2007). In order to validate the EM models, high resolution structures are normally docked into the EM density either manually or using the specific software. The resulting pseudo atomic (resolution wise) models of macromolecular assembly can be used

for calculation of the theoretical intensity and fitting the experimental curve with CRYSOL(Svergun, Barberato et al. 1995).

In this subchapter a tool EM2DAM for converting electron microscopy density maps to dummy atom models in PDB format (Berman, Henrick et al. 2003) is presented.

**The EM2DAM tool**

In order to rapidly validate the EM density maps, we developed a program EM2DAM (Electron Microscopy density map To Dummy Atom Model). EM2DAM can extract all required information from the EM density file in the MRC format (Crowther, Henderson et al. 1996). The user should only provide the threshold value, which defines the particle border and normally depends on the conditions of the EM experiment. One can use interactive mode and provide all the required information manually, if the MRC header is not available or the default values need to be changed. Basically, EM2DAM fills the EM density with dummy residues, which are located at the pixel size distance from each other. Output file has a PDB-like format and can be used, e.g. to compute the scattering pattern by CRYSOL (Svergun, Barberato et al. 1995) or for modeling with SASREF (Petoukhov and Svergun 2005). Moreover, it is also possible to refine the EM-based dummy atom model. If "--damform" option is selected, the resulting model can be used as an initial search volume in DAMMIN (Svergun 1999). It is possible to mark surface dummy residues within user-specified cut-off so that their phase can change from particle to solvent during the DAMMIN run.

The program is designed as a cross-platform console application and is used as follows:

```
$> em2dam [EMFILE] [OPTIONS]
```

where the options are:

- **-b, --bytes-to-skip=<NUMBER>,** where NUMBER is bytes to skip (header of EM file)
- **-n, --invert-bytes** if specified, inverted order of bytes is used
- **-i, --ni=<NUMBER>,** where NUMBER
- **-j, --nj=<NUMBER>,** is EM

- **-k, --nk=<NUMBER>,** map dimension
- **-p, --pixelsize =<NUMBER>,** where NUMBER is pixel size in Angstroms
- **-t, --threshold =<NUMBER>,** where NUMBER is contour level for EM map
- **-o, --output=<FILE>,** where FILE is output PDB file name; if not specified, the result is printed to STDOUT
- **-r, --reduced=<NUMBER>,** output only every <NUMBER>-th dummy atom recommended for large EM density maps
- **-q, --quiet** do not print log information to stdout
- **-d, --damform=<NUMBER>,** where NUMBER is a cut-off value in Angstroms to select surface atoms for DAMMIN refinement
- **-v, --version** print version information and exit
- **-h, --help** print this help text and exit

*Examples:*

```
$> em2dam emd_1654.map -o rubisco.pdb -t 3.0
```

Convert a density map in MRC format into a bead model using the contour level (also known as threshold) 3.0

```
$> em2dam emd_1654.map -o rubisco.pdb -t 3.0 -r 2 -d 5
```

Convert a density map in MRC format into a bead model with reduced number of dummy atoms. Contour level 3.0. The shape in terms of the surface dummy residues and neighbors within 5 Angstroms can be refined with DAMMIN.

## 3.4 MW estimation using excluded and Porod volumes

MW of the particle is one of the major overall parameters that can be directly deduced from the experimental solution scattering data. This parameter allows one to determine the oligomeric state of proteins or make conclusions on possible complex formation for mixtures of several distinct components. The MW is obtained from the forward scattering of the sample *I(0)* (by comparison with a reference, e.g. of bovine serum albumin), which is normalized by the sample concentration *c* so that the accuracy of the MW is limited by reliability of the measured *c* (Mylonas and Svergun 2007). In some cases its reliable estimate is difficult (e.g. for proteins containing few aromatic residues). Alternatively, MW may be estimated from the scattering data based on the excluded (i.e. hydrated) particle volume. The latter is computed without normalization of the intensity (Porod 1982) from the experimental data:

$$V_p \approx 2\pi^2 I(0) / \int_{s_{\min}}^{s_{\max}} s^2 [I(s) - A] ds \qquad \text{equation 3-1}$$

where $A$ is a constant, which normally has to be subtracted from each data point to force the $s^{-4}$ decay of the intensity at higher angles following the Porod's law (Porod 1982) for homogeneous particles. The exact relationship between MW and $V_p$ varies for different proteins depending on combination of several factors e.g. particle anisometry, flexibility etc.

In order to find correlation between the MW and the value of the excluded volume generated using either DAMMIN (Svergun 1999) or DAMMIF (Franke and Svergun 2009), different 3D structures from PDB (Bernstein, Koetzle et al. 1977) ranging from 14kDa to 500kDa in size were collected. High-resolution structures were solved by X-ray crystallography (34 models), electron crystallography (11 models) and NMR (8 models). Simulated data were generated using CRYSOL (Svergun, Barberato et al. 1995) and the atomic structures of biological assemblies either taken from PDB or generated using PISA (Krissinel and Henrick 2007) based on authors' assignment in PDB. Inverse Fourier transform and $R_g$ estimation were performed with GNOM (Svergun 1992) using the maximum diameter calculated by CRYSOL.

In this approach we utilized a concept from the information theorem stating that a scattering curve is characterized by its intensity at a discrete set of independent data points,

so-called Shannon channels, the number of these points being (Schannon and Weaver 1949; Moore 1980):

$$Ns = (s_{max} - s_{min}) D_{max} / \pi \qquad \text{equation 3-2}$$

For most SAXS curves, *Ns* value usually does not exceed 10–15.

The aim of the test calculations was to find the range of the scattering data for an adequate computation of the Porod volume such that the best correlation with the particle MW can be obtained.

To test whether the number of independent data points can influence the result, we considered three different scenarios: 8, 10 and 12 Shannon channels. Thus, for each PDB file tested, three separate inverse Fourier transforms corresponding to different number of independent data points were performed.

Table-3.1 summarizes the obtained results. We first tested if Porod volume estimation depends on MW. The simulated data was used to calculate Porod volume using the program AUTOPOROD. In **Figure 3-2**, Porod volume ($V_p$) divided by MW is plotted against the MW of the protein. $V_p$/MW ratio varies from 1.2 to 2 (excluding outliers) and neither depends on MW nor on particle anisometry (see **Figure 3-2**). The average $V_p$/MW ratio equals the 1.6±0.3

The correlation between the MW and the excluded volume calculated by DAMMIN and DAMMIF is shown on **Figure 3-3**. On the left side, average volume calculated by DAMMIN (top) or DAMMIF (bottom) divided by MW is plotted versus MW. On the right side, average volume calculated by DAMMIN (top) or DAMMIF (bottom) divided by MW is plotted against the maximum particle dimension divided by radius of gyration. For DAMMIN and DAMMIF, $V_a$/MW ratio varies from 1.4 to 2 and 1.1 to 2 (excluding outliers), respectively, and neither depends on MW nor on the particle anisometry (see **Figure 3-3**). The $V_a$/MW ratio equals to 1.7±0.3 for both DAMMIN and DAMMIF.

Thus, it was found that average Porod volume to MW ratio is about 1.7±0.3 independently of the size of the studied particle and its anisometry. These results helped us to design an algorithm (to be described elsewhere) for a reliable automated calculation of $V_p$. In case of DAMMIN or DAMMIF, one can get a rough estimation of the MW by dividing

excluded volume of *ab initio* model by 1.7 and thus confirm the oligomeric state of the particle in solution.

| Method | Vdammin/MW | | | Vdammif/MW | | | $V_a$/MW | | $V_p$/MW |
|---|---|---|---|---|---|---|---|---|---|
| | # Shannon Channels | | | # Shannon Channels | | | DAMMIN | DAMMIF | |
| | 8 | 10 | 12 | 8 | 10 | 12 | | | |
| X-ray | 1.73±0.27 | 1.67±0.22 | 1.64±0.23 | 1.76±0.21 | 1.61±0.28 | 1.56±0.33 | 1.70±0.19 | 1.66±0.22 | 1.65±0.23 |
| CEM | 1.75±0.46 | 1.75±0.36 | 1.65±0.37 | 1.80±0.43 | 1.65±0.41 | 1.59±0.41 | 1.72±0.39 | 1.68±0.40 | 1.62±0.42 |
| NMR | 1.67±0.24 | 1.65±0.27 | 1.56±0.23 | 1.70±0.21 | 1.55±0.29 | 1.56±0.30 | 1.62±0.23 | 1.61±0.23 | 1.55±0.23 |
| All | 1.73±0.31 | 1.68±0.26 | 1.63±0.26 | 1.76±0.26 | 1.61±0.30 | 1.56±0.34 | 1.69±0.25 | 1.65±0.26 | 1.62±0.28 |

*Table 3-1. Excluded volume calculation by DAMMIN and DAMMIF for simulated data. $V_a$ is an average volume calculated by DAMMIN or DAMMIF for different number of Shannon channels. $V_p$ is a Porod volume estimated by AUTOPOROD.*

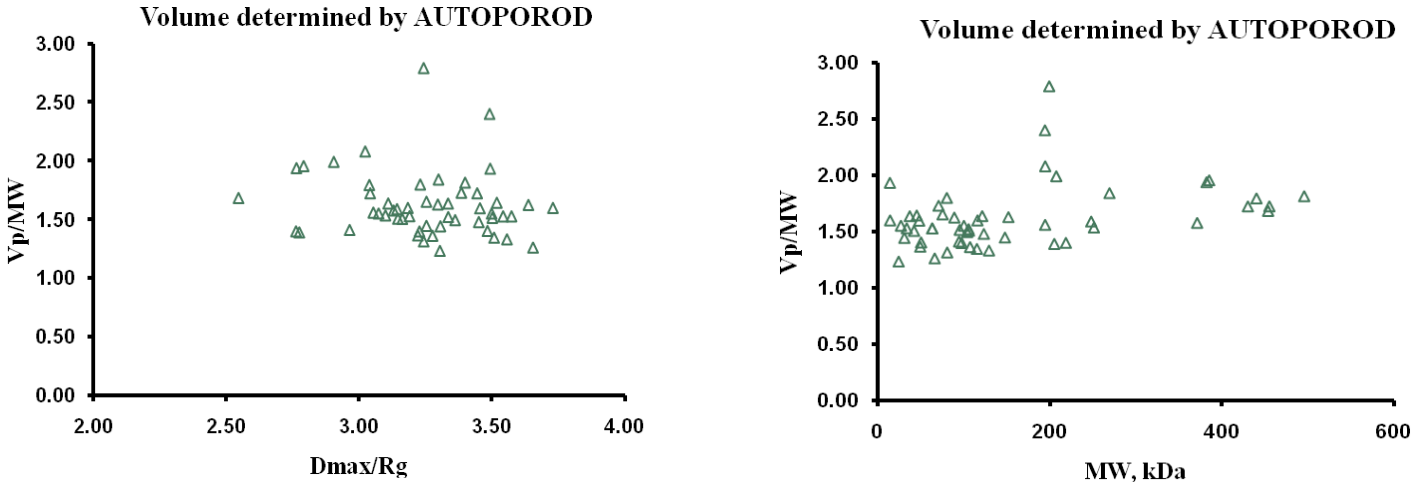



Figure 3-2. Correlation between MW and Porod volume calculated by AUTOPOROD. On the left panel, the volume calculated by AUTOPOROD divided by MW is plotted versus MW. On the right panel, the Porod volume divided by MW is plotted versus maximum particle dimension ($D_{max}$) divided by radius of gyration ($R_g$).

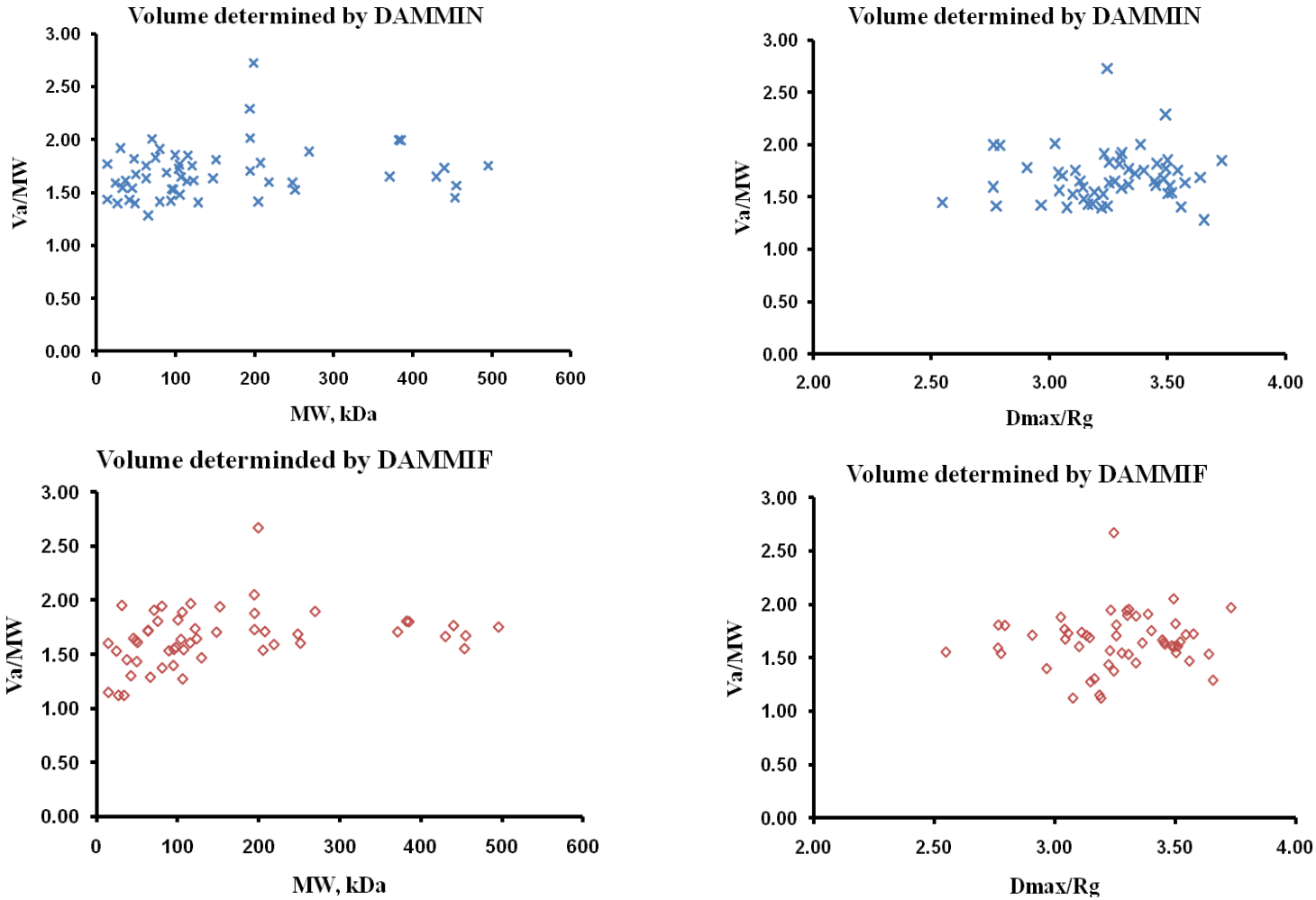



Figure 3-3. Correlation between MW and excluded volume calculated by DAMMIN (top figures) and DAMMIF (bottom figures). On the left, average volume divided by MW is plotted versus MW. On the right, average volume divided by MW is plotted against maximum particle dimension ($D_{max}$) divided by radius of gyration ($R_g$).

Chapter 4

# Novel developments utilizing bioinformatics predictors

## 4.1 Post processing of *ab initio* decoys generated by ROSETTA

### Background

Several papers have been published (Zheng and Doniach 2002; Zheng and Doniach 2005), describing the use of SAXS experimental data as a constraint for *de novo* protein modeling or threading. Both are still challenging problems in computational biology. *Ab initio* modeling and threading methods are used when no homologous sequences with 3D structures are available. Basically, SAXS experimental data is used as "a sieve" to filter out wrong models, leaving hundreds of plausible models. However, noticeable progress was observed in the development of algorithms for *ab initio* and threading structure prediction, resulting in quite accurate models generated for small proteins *de novo* (Bradley, Misura et al. 2005; Zhang 2007). The most successful *de novo* structure prediction algorithms – according to the biannual competition – CASP (Critical Assessment in Structure Prediction) are ROSETTA, TOUCHSTONE II and I-TASSER. For example, *ab initio* models for 85 amino acid protein with root mean square deviation (RMSD) < 1.5 Å have been reported (Bradley, Misura et al. 2005).

The main motivation for this work was to test the predictive power of the protein *ab initio* modeling module of ROSETTA combined with post processing using SAXS data as a constraint. Scoring functions used for *de novo* modeling may not suffice to discriminate the correct solutions and often additional experimental constraints are required. In order to see if information obtained from SAXS experiments can discriminate models close to the available high resolution X-ray structure, the following study was performed.

### Methods and results

1.000 decoys (structural models) were created for the lysozyme primary sequence using the ROSETTA++ software suite (Bonneau, Strauss et al. 2002) on the biosaxs cluster at the EMBL-Hamburg. All models were compared to the 3D structure of hen egg-white lysozyme (PDBID: 6LYZ), using SUPPDB and SUPCOMB (Kozin and Svergun 2001). SUPPDB searches for models of similar 3D structure and calculates RMSD. SUPCOMB, on the other hand, calculates a proximity measure between the overall shapes of the objects – Normalized

Spatial Discrepancy (NSD). It is normalized to be independent of size and geometric nature of models.

The top ten, bottom ten and 10 random models were selected, according to ROSETTA scores (see **Tables 4.1**, **4.2** and **4.3**). SAXS curves for these models were generated and fitted to the experimental data using CRYSOL (Svergun, Barberato et al. 1995). Given the atomic coordinates, CRYSOL minimizes the discrepancy by varying average displaced solvent volume per atomic group, contrast of the hydration shell and, if required, the value of relative background.

Some correlation between the ROSETTA score and RMSD, NSD or chi-square values was expected. Models with high ROSETTA scores (top ten models) should fit lysozyme experimental data and match available X-ray structure with lower chi-square and lower RMSD/NSD values, respectively.

| Top ten models | RMSD (Å) | NSD | $\chi$ | ROSETTA scores |
|---|---|---|---|---|
| S_0363_5508.pdb | 15.78 | 1.891 | 0.849 | -93.19 |
| S_0956_9405.pdb | 16.66 | 1.832 | 0.855 | -92.60 |
| S_0974_7687.pdb | 16.48 | 2.140 | 1.601 | -91.29 |
| S_0587_1065.pdb | 14.92 | 1.865 | 1.141 | -90.22 |
| S_0324_2902.pdb | 14.25 | 1.943 | 1.401 | -88.63 |
| S_0226_9656.pdb | 15.49 | 1.833 | 0.979 | -87.00 |
| S_0438_3550.pdb | 12.84 | 1.901 | 0.973 | -86.50 |
| S_0500_4067.pdb | 13.86 | 1.901 | 1.264 | -86.20 |
| S_0215_7155.pdb | 15.04 | 1.945 | 1.196 | -86.03 |
| S_0229_9446.pdb | 15.79 | 2.000 | 1.672 | -85.04 |

*Table 4-1. Top ten models (decoys) according to ROSETTA scores. RMSD is a root mean square deviation between the decoy and X-ray structure of lysozyme (PDBID: 6LYZ); NSD is a normalized spatial discrepancy between ROSETTA model and 6LYZ; $\chi$ is a discrepancy between decoy and experimental SAXS data.*

| Bottom ten models | RMSD (Å) | NSD | χ | ROSETTA scores |
|---|---|---|---|---|
| S_0739_2409.pdb | 13.54 | 1.835 | 1.032 | -6.36 |
| S_0020_8081.pdb | 13.18 | 2.118 | 1.703 | -7.71 |
| S_0330_6507.pdb | 15.36 | 1.998 | 1.529 | -13.84 |
| S_0721_6305.pdb | 13.27 | 1.957 | 1.088 | -14.77 |
| S_0468_2321.pdb | 14.38 | 1.893 | 1.325 | -15.51 |
| S_0277_9516.pdb | 17.83 | 1.897 | 1.440 | -16.22 |
| S_0516_9071.pdb | 13.87 | 2.004 | 1.592 | -17.31 |
| S_0142_5817.pdb | 13.69 | 1.809 | 0.822 | -17.73 |
| S_0154_3772.pdb | 14.68 | 1.849 | 0.705 | -18.36 |
| S_0791_2246.pdb | 14.47 | 2.102 | 1.455 | -19.23 |

*Table 4-2. Bottom ten models (decoys) according to ROSETTA scores. RMSD is a root mean square deviation between the decoy and X-ray structure of lysozyme (PDBID: 6LYZ); NSD is a normalized spatial discrepancy between ROSETTA model and 6LYZ; χ is a discrepancy between decoy and experimental SAXS data.*

Although some of the decoys have an overall shape very similar to that of lysozyme as judged by the chi-square value, RMSD and NSD values are rather independent of the ROSETTA scores. The latter indicates high conformational divergence of the generated decoys due to the ROSETTA algorithm which tries to cover immense conformational space of the protein.

**Conclusions**

Despite the recent progress made in *ab initio* modeling (Bradley, Misura et al. 2005; Raman, Vernon et al. 2009), the problem of 3D structure reconstruction of macromolecules purely based on physicochemical properties of single amino acids or small peptide stretches (3 or 9 residues) is still challenging and not solved. Some studies (e.g. (Bradley, Misura et al. 2005)) suggest that for certain cases it may be possible to reconstruct a 3D structure *de novo*, that resembles the high resolution structure well, however, this approach cannot be generalized and is only applicable to a very limited number of cases. Using the experimental information from SAXS (Zheng and Doniach 2002), NMR (Shen, Lange et al. 2008), EM (DiMaio, Tyka

et al. 2009) and other biophysical methods may impose additional restraints on the search space and provide close to native model(s) in some cases.

Current study shows that it is not possible to select close to native *de novo* model(s) of lysozyme from the pool of ROSETTA decoys using SAXS data as a constraint. Although the overall shape of some of the decoys was similar to that of lysozyme, the fold and atomic details could not be reconstructed and selected reliably.

| Ten random models | RMSD (Å) | NSD | $\chi$ | ROSETTA scores |
|---|---|---|---|---|
| S_0120_7696.pdb | 12.64 | 1.782 | 1.146 | -52.51 |
| S_0246_4812.pdb | 15.67 | 1.961 | 1.321 | -61.38 |
| S_0371_0049.pdb | 14.65 | 1.981 | 1.248 | -71.23 |
| S_0497_1901.pdb | 15.32 | 1.829 | 1.402 | -54.15 |
| S_0622_5954.pdb | 17.03 | 1.827 | 1.126 | -42.78 |
| S_0749_5966.pdb | 11.94 | 1.845 | 1.044 | -54.71 |
| S_0875_8164.pdb | 15.18 | 1.783 | 0.879 | -77.50 |
| S_0121_0194.pdb | 16.89 | 2.046 | 1.547 | -57.19 |
| S_0246_4812.pdb | 15.67 | 1.961 | 1.321 | -61.38 |
| S_0372_6109.pdb | 16.41 | 1.693 | 0.917 | -63.22 |

*Table 4-3. Ten random models (decoys) according to ROSETTA scores. RMSD is a root mean square deviation between the decoy and X-ray structure of lysozyme (PDBID: 6LYZ); NSD is a normalized spatial discrepancy between ROSETTA model and 6LYZ; $\chi$ is a discrepancy between decoy and experimental SAXS data.*

## 4.2 Selection and refinement of HADDOCK solutions

*Adapted from*

Shkumatov, A., Svergun, D.I. and Bonvin, A. M. J. J. (2008) Selection and Refinement of HADDOCK Solutions Using SAXS Data. Science *and Supercomputing in Europe (HPC-Europa report 2008)*, 274-278

Predicting protein-protein interactions is a challenging task of paramount importance. Theoretical methods for predicting protein-protein interactions, like the program HADDOCK (Dominguez, Boelens et al. 2003) for example, benefit significantly from experimental information available. This information can be derived from mutagenesis, cross-linking, mass spectrometry, NMR or SAXS. We used HADDOCK for rigid body modeling and experimental information from SAXS to select and refine subset of HADDOCK solutions. Interestingly, for one of the test complexes we found in a selected HADDOCK subset models with subunit orientation similar to one previously proposed (van den Heuvel, Svergun et al. 2003).

### State of the art

There is a growing amount of information about 3D protein structures. In the past the main focus of the structural studies was on the atomic details of protein structure, whereas nowadays we ask further challenging questions, namely, how proteins interact with each other, what are the molecular details of interaction and how does this relate to the function. HADDOCK (Dominguez, Boelens et al. 2003; van Dijk, Boelens et al. 2005) is an approach to predict protein-protein interactions. The HADDOCK algorithm consists of three steps: (i) proteins are represented as rigid bodies and a large number of plausible solutions are generated; (ii) flexibility is introduced first in the side chains and subsequently in both side-chains and backbone of pre-defined flexible segments encompassing the active and passive

residues; (iii) finally, the solutions are refined in explicit solvent. The final structures are clustered and scored using a combination of energy terms (mainly intermolecular van der Waals and electrostatic energies and restraint energies).

SAXS is a universal low-resolution method, providing high precision information in respect to size and shape of the particle in solution (Jacques and Trewhella 2010). Given a number of possible ways two or more proteins can interact with each other, SAXS experimental data can be used to help discriminating the correct solution(s).

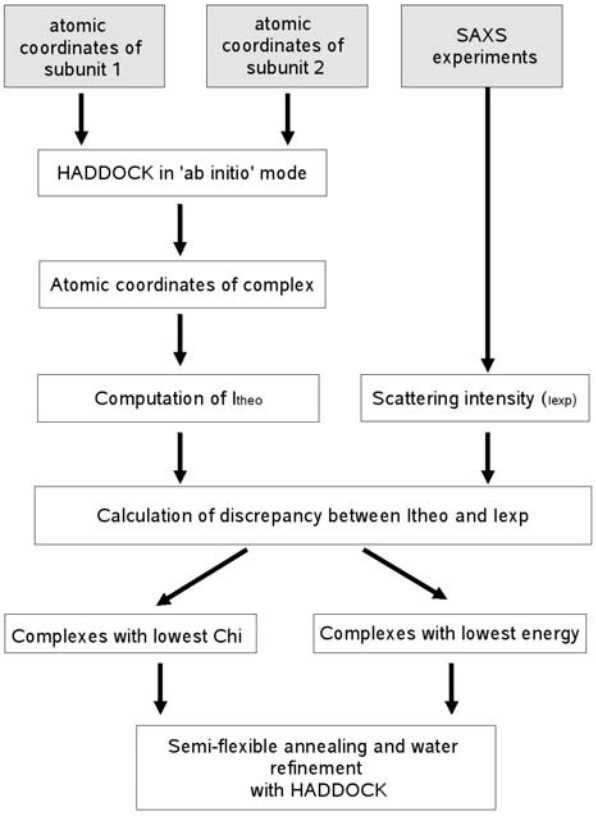



Figure 4-1. Scheme of the work

Docking is computationally very expensive. Usually, 5-10 docking runs should be performed for each target (protein complex), each generating 1.000 to 10.000 solutions. The new version of the program CRYSOL ((Svergun, Barberato et al. 1995), chapter 3.1) allows generation of theoretical SAXS curves and fast fitting to experimental data. This takes 4 to 10 seconds for a single HADDOCK solution, depending on the number of experimental points used. In this work we used the HPC (High Performace Computing)-Europa's systems in order to cover computational requirements needed for the large number of docking and CRYSOL runs.

**Methods**

HADDOCK is one of the approaches that makes use of experimental information to drive the docking while allowing for various degrees of flexibility. It is possible to first perform only plain rigid body docking, but at the further stages of refinement also employ additional information from experimental methods (see **Figure 4-1**). Usually structures are selected for flexible refinement based on the HADDOCK score. Here we used SAXS experimental data

instead to drive the selection process (see **Figure 4-2**). For this, a new faster version of CRYSOL (v.2.7) was employed.

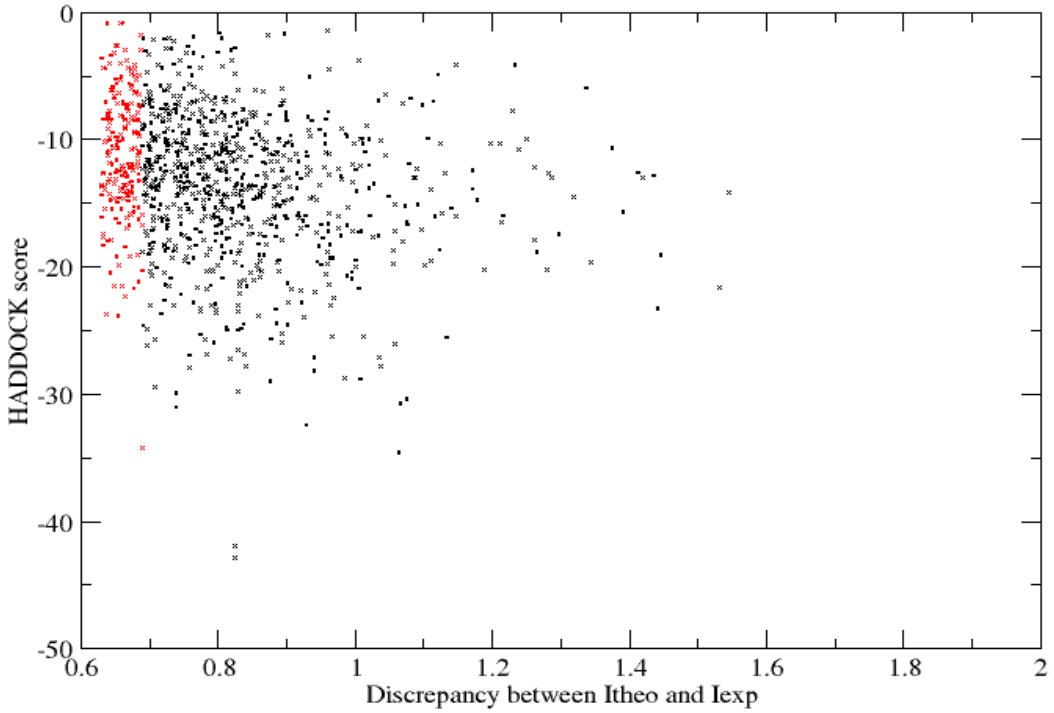



Figure 4-2. Selection based on discrepancy (not on score as in usual HADDOCK run)

We used a number of protein complexes to test the performance of HADDOCK together with SAXS-driven refinement. Among them are transketolase (Svergun, Petoukhov et al. 2000), which forms a homodimer; calmodulin with peptide (Niemann, Petoukhov et al. 2008), glutamate synthase and Ferredoxin, which form 1:1 complex according to SAXS data (van den Heuvel, Svergun et al. 2003); Met Receptor and protein InlB (Niemann, Petoukhov et al. 2008), which forms a multidomain complex; complex of Arc23 and actin (Boczkowska, Rebowski et al. 2008). For all of these complexes both SAXS experimental data and information about their interactions was available.

**Results**

For most of the test cases, following the HADDOCK and CRYSOL stages, either the fit of the theoretical curves to the experimental data did not improve compared to the target models

(transketolase) or the introduction of more restraints was required (Met-InlB, Arc23-actin, calmoduline-peptide complex) to keep the domains together during the refinement.

We have thus concentrated on the case of glutamate synthase and ferredoxin, where HADDOCK was able to find models with improved fits compared to previously published models (van den Heuvel, Svergun et al. 2003). Using our models, we compared HADDOCK scores and discrepancy values from CRYSOL and showed that (i) the subset of structures selected using SAXS experimental data shows a broad distribution of HADDOCK scores; (ii) selected models differ significantly in terms of their overall 3D structure, but not their overall shape; (iii) the selected subset contains models with subunit orientation close to the previously proposed model (van den Heuvel, Svergun et al. 2003).

## 4.3 Automatic selection and HADDOCK refinement of models

*Adapted from*

Shkumatov, A., Svergun, D. and Bonvin, A. M. J. J. (2009) Small-Angle X-ray Scattering Driven Docking. *HPC-Europa2: Science and Supercomputing in Europe* (research highlights 2009), 78

Protein-protein interactions are playing important roles in many different cellular processes. SAXS can give insights into the structure of large protein complexes. Especially interesting is the use of SAXS for macromolecular complexes, where the quaternary structure of the complex can be reconstructed from the high resolution models of individual subunits. Docking, on the other hand, is a theoretical approach to predict protein-protein interactions. Here we report an automatic method, combining SAXS rigid body modeling with subsequent docking refinement. First results are presented.

### State of the art

The SAXS-based rigid body modeling program SASREF uses a SA procedure to search for interconnected arrangements of subunits without clashes (Petoukhov and Svergun 2005). In order to establish a link between the docking and rigid body refinement methods the HADDOCK (Dominguez, Boelens et al. 2003; van Dijk, Boelens et al. 2005) approach was used, which is able to predict protein-protein interactions, generate tentative models and score the interfaces of the given models. This report details an automatic approach, combining rigid body modeling and docking which involves the joint use of HADDOCK and SASREF. Subunits are moved and rotated randomly by SASREF to fit the scattering data and the final refinement is done using HADDOCK, starting from the *N* best SASREF solutions, where *N* can be defined by the user. The first results of the combined approach are reported.

### Methods

During SASREF refinement (see subchapter 2.3) individual subunits are randomly moved and rotated at each step of SA in order to minimize the target function. Additional restraints

(contacting amino acids) or constrains can be used. Normally, one SASREF run gives only a single model which agrees well with SAXS experimental data. We modified SASREF to produce *N* best models (typically 1.000). Top 200 models according to the chi value are selected for two steps of HADDOCK refinement: flexible SA in torsion angle space (it1) and flexible water refinement (water).

HADDOCK is a data-driven docking program (see subchapter 4.2) that can be used with experimental data or with bioinformatics interface prediction. In essence HADDOCK is a collection of python scripts derived from ARIA, which uses CNS as structure calculation software. Due to its modular structure HADDOCK allows one to integrate additional programs. We have included the possibility to use SASREF rigid body refinement on the first step of HADDOCK, as well as added SASREF parameters into the HADDOCK configuration file.

**Dataset**

We used simulated data (PDBID: 2OMU) to test the performance of our predictions. 2OMU is a binary complex of the bacterial invasion protein internalin-A of Listeria monocytogenes and human receptor E-cadherin. A theoretical SAXS scattering curve with 256 data points was generated by CRYSOL (Svergun, Barberato et al. 1995) and used as a simulated data.

**Results and conclusions**

We performed a number of runs, using different initial positions of the subunits: (i) as in the X-ray structure, (ii) placing randomly one of the subunits (iii) and positioning randomly both subunits. In case of (ii) and (iii), the solutions found by SASREF were far from the native complex solved by X-ray crystallography. In case of (i), the models found by SASREF were close to the original model (see **Fig. 4-3 and Fig. 4-4**). We also analyzed the interface Root Mean Square Deviation (i-RMSD) which basically resembled the **Fig. 4-3**.

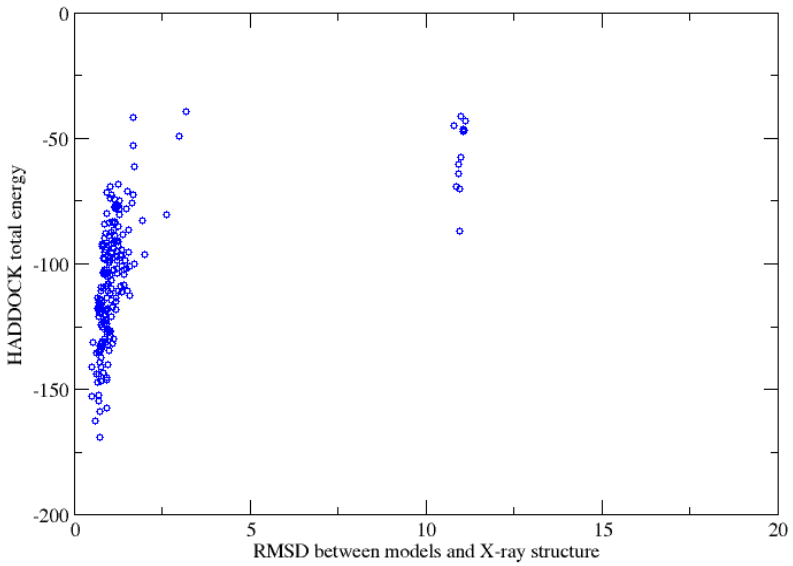



Figure 4-3. RMSD between X-ray structure (PDBID: 2OMU) and refined models plotted versus total HADDOCK scores.

There were two clusters of models predicted, one of which is the near-native with many structures having very good energies (see **Fig. 4-4A**); the other cluster has one of the subunits rotated approximately by 180 degrees (see **Fig. 4-4B**). The shape of the representative models of both clusters is quite similar, which makes it difficult to discriminate the correct orientation of the subunits with respect to each other by using only the SAXS data.

Thus, it was possible to predict close to native solutions when information about the binding interface was known. However, the models (see **Fig. 4-5**) deviated not significantly from the known X-ray structure of the complex in terms of RMSD, which may not be enough as larger overall conformational changes (> 1.5 Å) may occur upon complex formation. Given the results of our study, we put an objective to develop a protocol which can model larger overall conformational changes upon complexation. The protocol combining information about binding interface, available high resolution structures of unbound subunits and SAXS experimental data for the complex is described in detail in the next subchapter.

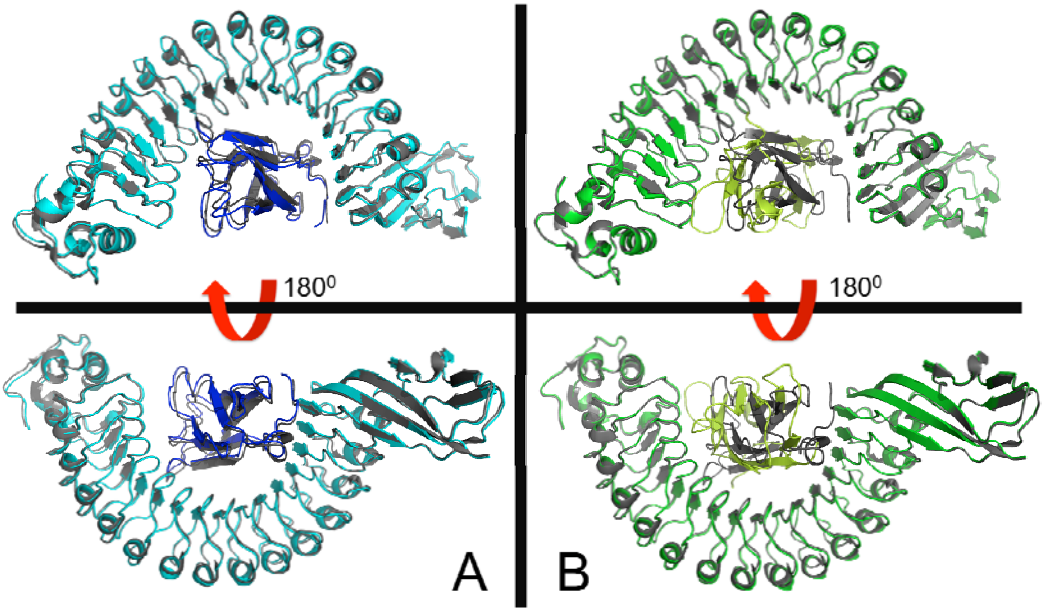



Figure 4-4. Two best clusters found by the combination of rigid body modeling based on SAXS experimental data and refinement by docking program HADDOCK. The smaller subunit in the second cluster (B) is rotated by approximately 180 degrees as compared to representative of the first cluster (A).

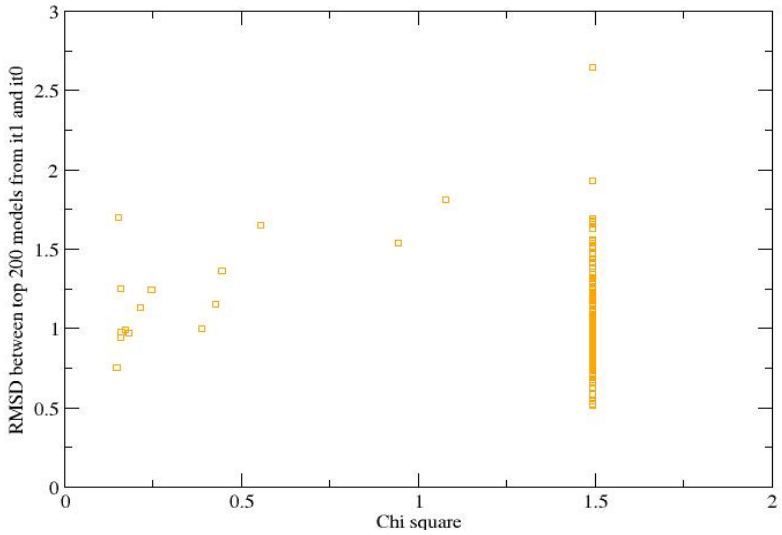



Figure 4-5. RMSD between SASREF rigid-body models before and after HADDOCK refinement.

## 4.4 **NMA-based refinement of binary complexes**

### Background

NMA is based on a physical description of the protein (harmonic approximation of the force field that describes the interaction between the atoms) and allows the flexibility to be decomposed into discrete deformations (modes) (Tama and Brooks 2006). It was previously reported that most protein conformational transitions documented in the PDB can be accounted for by a few low-frequency NM (Tama and Sanejouand 2001; Krebs, Alexandrov et al. 2002). NMA can be used to explore the collective motions in biomolecules (Lindahl, Azuara et al. 2006), refine the structures against experimental data, such as X-ray, SAXS (Gorba, Miyashita et al. 2008; Miyashita, Gorba et al. 2010), Cryo-EM (Tama, Miyashita et al. 2004), or to optimize the docked complexes by modeling receptor/ligand flexibility through NM motions (Lindahl and Delarue 2005). MD is the more accurate method for observing conformational changes. However, typical time scales accessible by MD simulations are few ms which makes MD extremely time consuming (it may take up to weeks) and there is a restriction on the size of studied system. NMA, on the other hand, can easily access ms time scales. It is much faster, as opposed to MD.

We present here a method (NMADREFS) for the refinement of binary protein-protein complexes using the 3D structures of unbound subunits and the SAXS data from the complex. This method uses a linear combination of low-frequency NM from an EN description of the molecule in an iterative manner to deform the structure optimally to conform to the SAXS scattering curve. Unlike other available methods, NMADREFS samples conformational space of the separate (unbound) subunits of a complex using NM while maintaining connectivity, avoiding steric clashes and preserving the interaction interface. Some results of our studies on simulated data from protein-protein docking benchmarks are presented.

### Methods

*NMA using an EN model*

NMA has proven to be useful for studies of collective motions of biological systems that may occur on time scales not easily accessible to conventional MD simulations. NMA consists of

the decomposition of motions into vibrational modes. Each normal-mode vector represents a specific motion of the molecule in terms of a 3-n dimensional vector $\vec{a}$, where n is the number of atoms in the molecule. The modes belonging to low frequencies represent preferential global (collective) motions (Tama and Sanejouand 2001).

Deformation of the molecule must be limited to displacements along the low-frequency NM. This ensures that the generated models are energetically accessible. It was previously shown that by using 2 or 3 out of the 10 lowest frequency NM, conformational changes could be described within 2-3 Å RMSD for a couple of test cases using an iterative fitting protocol (Gorba, Miyashita et al. 2008; Gorba, Miyashita et al. 2008).

NM are calculated using a simplified representation of the potential energy function, which was first introduced by Tirion (Tirion 1996). The EN potential was shown to be successful in reproducing large and collective motions of macromolecules (Bahar and Rader 2005)**.** For all-atom structures we use here an interaction cut-off of 8 Å between atomic pairs. This is the default value used for NM calculations performed by the ElNemo server (Suhre and Sanejouand 2004). The strength of the potential is a phenomenological constant assumed to be the same for all interacting pairs. Additionally, to speed up NM calculation, a building block approach is used. Three residues are put in one block. Every block is considered to be a rigid body. It was shown by Tama et al (Tama, Gadea et al. 2000) that using such an approach, low-frequency NM can be reproduced with very high accuracy. Both, the EN and the building-block methods were implemented by Sanejouand et. al and are freely available programs (http://ecole.modelisation.free.fr/modes.html), which also build the core of the ElNemo server. The refinement protocol used here (see *Algorithm*) was written in PERL and makes use of those two programs together with CRYSOL (Svergun, Barberato et al. 1995).

*Algorithm*

NMADREFS samples the conformational space of the separate subunits of an initial approximation by deforming their structures using NMA. NMADREFS uses an iterative refinement to obtain a better fit in terms of the $\chi$ value as calculated by CRYSOL (Svergun, Barberato et al. 1995), while preserving the interaction interface (see **Fig. 4-6**). The input for the refinement protocol includes the experimental SAXS data for the complex, interface contact file and an initial approximation. The latter is a PDBFILE with complex constituents

corresponding to 3D structures of unbound subunits positioned such that the predicted binding interface is an area where subunits contact each other. The 10 lowest-frequency modes are calculated as described above for each subunit independently (these are modes numbered 7 to 16 since the first 6 modes account for rigid body translation and rotation only and are therefore not needed). A random number between 0 and a maximum is computed for the actual amplitude of displacement $q_i$ along each mode $i$. This maximum is set such that every atom is only displaced by a maximum of 1.0 Å (or a user specified value). Using these amplitudes the atoms are then displaced along various (here 11) randomly picked different linear combinations of two modes out of modes 7 to 16 thereby creating new structures: when $\vec{x}^t = \left(x_{1,} y_{1,} z_{1,} .., x_n, y_n, z_n\right)$ is the 3n-dimensional vector of the atomic coordinates for iteration step $t$, the new coordinates $\vec{x}^{t+1}$ after displacement are given by:

$$\vec{x}^{t+1} = \sum_{i \in M} q_i \vec{a}_i + \vec{x}^t \qquad \text{equation 4-1}$$

Here $\vec{a}_i$ is the $i$-th NM unit vector and M is a subset of modes 7-16 used to describe the transition. It was found by (Gorba, Miyashita et al. 2008) that a linear combination of 2 (sometimes 3) out of 10 modes suffices to correctly describe the conformational changes. Typically 11 mode combinations are selected. In case the displacement is too large and breaks the structure, i.e. if the new C*a*-C*a* distances are larger than 4.0 Å, a smaller maximum displacement is applied successively until the new structure has proper C*a*-C*a* distances. We also consider a situation when only one of the subunits is changing. Thus, if there are two subunits, it results in (11+1) x (11+1) – 1 = 143 newly created complexes with the preserved interaction interface at each step of the refinement. To keep the interaction interface, as defined in the interface contacts file, the interface atoms from mode combinations are superposed on the unbound subunits' interface using SUPPDB (Kozin and Svergun 2001). Subsequently, the new complexes are checked for clashes. The complex with lowest discrepancy to experimental SAXS data is taken as the input structure for the next iteration of refinement until an improvement in the fit is obtained (see **Fig. 4-6**). The number of refinement steps can be defined by the user.

*Usage and output*

In order to run NMADREFS one has to provide the following arguments:

```
$> nmadrefs.pl PDBFILE AMPLITUDE DATAFILE STEPS INTERFACE
```

where arguments are:

- **PDBFILE** PDB file with initial approximation
- **AMPLITUDE** maximum amplitude for NMA displacement
- **DATAFILE** SAXS data for the complex
- **STEPS** number of steps for refinement
- **INTERFACE** file with interface information

PDBFILE should have an initial approximation of the complex (as described in *Algorithm*).

The interface file has the following format: 1st column should have residue number as in PDB file of the first unbound subunit in the initial approximation contacting the other residue in the second subunit (2nd column).

NMADREFS generates a LOG file (nmadrefs.log), which contains the header with some information about the program, when it was started and the command used to start the refinement process, "Step #" – Refinement step, "CHI" – Discrepancy to experimental SAXS data.

NMADREFS creates a directory called "intermediates", containing PDB models generated during each step of the refinement.

**Preliminary results**

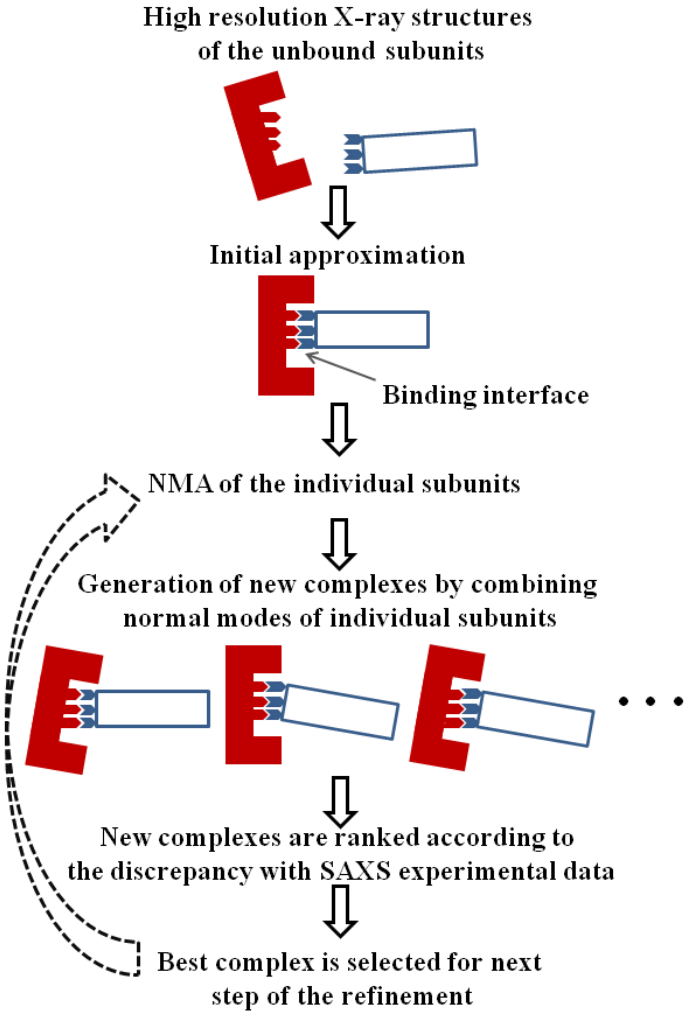



Figure 4-6. Workflow of NMADREFS

NMADREFS has been developed only recently and is now being validated on benchmarks used for testing the docking programs. One of the examples already gave interesting and encouraging results as shown on **Fig. 4-7**. Kallikrein-Hirustatin complex (PDB 1HIA) consists of two subunits (PDB IDs): 2PKA and 1BX8. When the high resolution structures of unbound subunits were merged into a single PDB file (initial approximation) with binding interface as in the known structure (1HIA), the discrepancy to simulated SAXS data was 2 (see **Fig. 4-7A**) and 0.46 nm RMSD to known X-ray structure of the complex (1HIA). The experimental and theoretical curves have a mismatch in the low angle as well as the higher angle (> 0.25 Å) regions. After 100 steps of NMADREFS refinement, the final model had 0.46 nm RMSD to known X-ray structure of the complex and discrepancy of 0.92 (see **Fig. 4-7A** and **B**). The curve computed from the refined structure fits better the simulated data both in low and high-angle regions (see **Fig. 4-7A**).

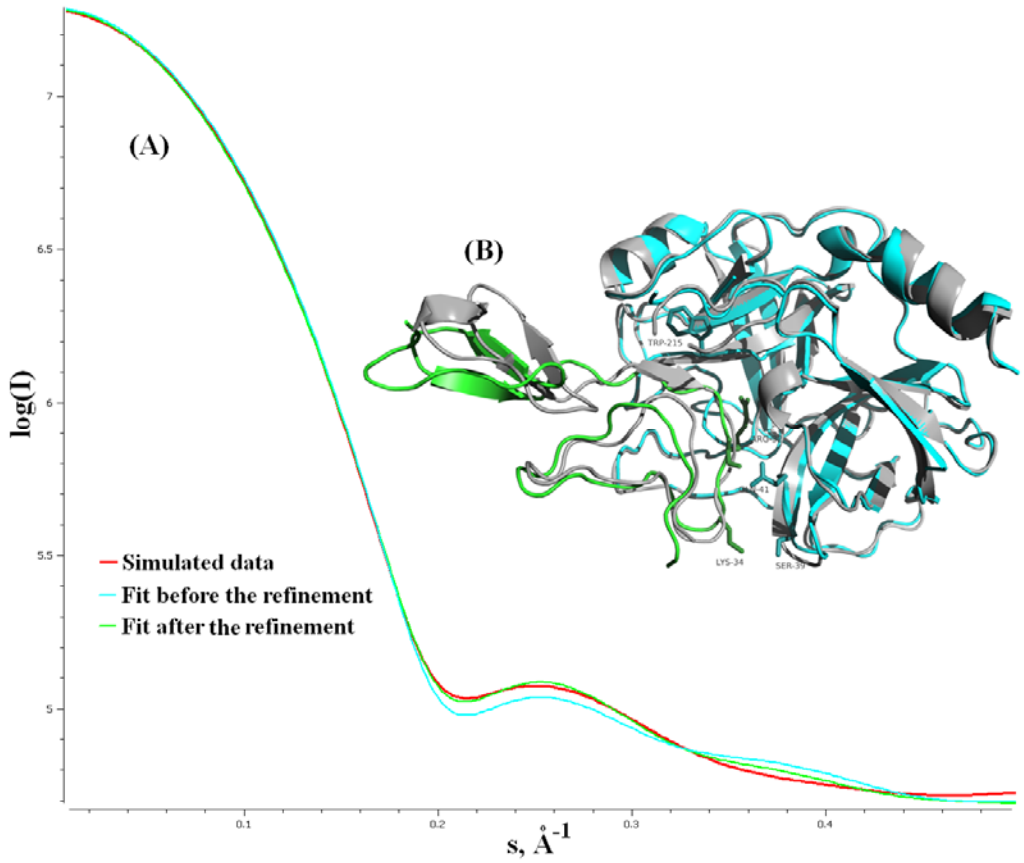



Figure 4-7. NMADREFS refinement of Kallikrein-Hirustatin complex. (A) Fit of the initial approximation (before the refinement) and refined model to experimental data. (B) Superposition (RMSD = 0.46) of refined model, shown in green and cyan) and a high resolution structure of Kallikrein-Hirustatin complex (PDB 1HIA), shown in grey. The interface residues are shown as sticks and labeled accordingly.

**Conclusions**

A new method for the refinement of binary protein-protein complexes using SAXS data and high resolution structures of unbound subunits is presented. Unlike other available methods, it samples conformational space of individual subunits rather than treating the complex as a single rigid body. The sampling is done by a linear combination of normal modes, shown to suffice in describing the conformational changes (Gorba, Miyashita et al. 2008). During the refinement the binding interface is preserved and a check for clashes is performed at each refinement step. The method has not yet been extensively tested; yet, first results are encouraging.

## 4.5 Application of bioinformatics tools to proteome analysis of tardigrades

*Adapted from*

**Tardigrade workbench: comparing stress-related proteins, sequence-similar and functional protein clusters as well as RNA elements in tardigrades.**

Foerster* F, Liang* C, Shkumatov* A, Beisser D, Engelmann JC, Schnoelzer M, Frohme M, Mueller T, Schill RO, Dandekar T. *BMC Genomics*. 2009 Oct 12; 10:469.

* Equal contributors

### Background

Tardigrades are small metazoans resembling microscopic bears ("water-bears", 0.05 mm to 1.5 mm in size) and live in marine, freshwater and terrestrial environments, especially in lichens and mosses (Marcus 1928; Marcus and Dahl 1928; Nelson 2002). They are a phylum of multi-cellular animals capable of reversible suspension of their metabolism and entering a state of cryptobiosis (Keilin 1959; Ramazzotti and Maucci 1983). A dehydrated tardigrade, known as anhydrobiotic tun-stage (Baumann 1922; Baumann 1927), can survive for years without water. Moreover, the tun is resistant to extreme pressures and temperatures (low/high), as well as radiation and vacuum (Wright 2001; Horikawa, Sakashita et al. 2006; Jonsson and Schill 2007; Jonsson, Rabbow et al. 2008; Hengherr, Worland et al. 2009; Hengherr, Worland et al. 2009).

Well known species include Hypsibius dujardini which is an obligatory parthenogenetic species (Ammermann 1967). The tardigrade H. dujardini can be cultured continuously for decades and can be cryopreserved. It has a compact genome, a little smaller than that of Caenorhabditis elegans or Drosophila melanogaster, and the rate of protein evolution in H. dujardini is similar to that of other metazoan taxa (Gabriel, McNuff et al. 2007). H. dujardini has a short generation time, 13-14 days at room temperature. Embryos of H. dujardini have a stereotyped cleavage pattern with asymmetric cell divisions, nuclear migrations, and cell migrations occurring in reproducible patterns (Gabriel, McNuff et al. 2007). Molecular data are sparse but include the purinergic receptor occuring in H. dujardini (Bavan, Straub et al. 2009).

Milnesium tardigradum is an abundant and ubiquitous terrestrial tardigrade species in Europe and possibly worldwide (Kinchin and Dennis 1994). It has unique anatomy and motion characteristics compared to other water bears. Most water bears prefer vegetarian food, M. tardigradum is more carnivorous, feeding on rotifers and nematodes. The animals are really tough and long-living, one of the reasons why M. tardigradum is one of the best-studied species so far.

Questions of general interest are: How related are tardigrade proteins to each other? Which protein families provide tardigrade-specific adaptations? Which regulatory elements influence the mRNA stability? Starting from all published tardigrade sequences as well as 607 unpublished new sequences from Milnesium tardigradum, we analysed tardigrade specific clusters of related proteins, functional protein clusters and conserved regulatory elements in mRNA mainly involved in mRNA stability. The different clusters and identified motifs are analysed and discussed, all data are also available as a first anchor to study specific adaptations of tardigrades in more detail (Tardigrade workbench). Furthermore, the tardigrade analyzer, a sequence server to analyse individual tardigrade specific sequences, is made available. It will be regularly updated to include new tardigrade sequences. It has a number of new features for tardigrade analysis not available from standard servers such as the NIH Entrez system (Baxevanis 2006): several new species-specific searches (Echiniscus testudo, Tulinus stephaniae), additional new sequence information (M. tardigradum) and pattern-searches for nucleotide sequences (including pattern search on nonredundant protein database, NRDB). An easy search for clusters of orthologous groups (COG, (Tatusov, Galperin et al. 2000)) different from the COGnitor tool (Tatusov, Fedorova et al. 2003) allowing tardigrade specific COG and eukaryotic COG (KOG) searches is also available.

Furthermore, a batch mode allows a rapid analysis of up to 100 sequences simultaneously when uploaded in a file in FASTA format (for tardigrade species or NRDB).

Two fifths of the tardigrade sequences cluster in longer protein families, and we hypothesise for a number of these that they are implicated in the unique stress adaptation potential of tardigrades. We find also ten tardigrade specific clusters. The unique tardigrade adaptions are furthermore indicated by a number of functional COGs and KOGs identified here, showing a particular emphasis on the protection of proteins and DNA. RNA read out is specifically regulated by several motifs for mRNA stability clearly overrepresented in tardigrades.

## Results and Discussion

We analysed all publicly available tardigrade sequences (status 9th of April 2009) as well as 607 unpublished *M. tardigradum* sequences from our ongoing transcriptome analysis.

### *Major tardigrade protein clusters of related sequence-similar protein*

All available tardigrade sequences were clustered by the CLANS algorithm (Frickey and Lupas 2004). Interestingly, 39.3% of the predicted proteins (mainly EST-based predictions) cluster in just 58 major families, each with at least 20 sequences [Appendix A: **Table A-1**]. These include 4,242 EST sequences from a total of 10,787.

Using these clusters, a number of tardigrade-specific adaptations become apparent (**Table 4-4** [and appendix A: **Table A-1**]): the clusters include elongation factors (cluster 12), ribosomal RNAs and proteins (cluster 1, 4, 32 and 56) which are part of the transcriptional or translational machinery. Cluster 5 (chitinase binding domain (Tjoelker, Gosting et al. 2000)) could provide membrane and structural reorganization or immune protection (e.g. fungi) according to homologous protein sequences characterized in other organisms. Other clusters show protein families related to the tardigrade stress adaptation potential, e.g. ubiquitin-related proteins (cluster 14; maybe stress-induced protein degradation) and cytochrome oxidase-related proteins (cluster 2, suggested to be involved in respiratory chain).

Moreover, proteins responsible for protein degradation (cluster 15) were found as well as proteins regulating peptidases (cluster 16). Cluster 23 consists of 53 heat shock proteins which are involved in many stress response reactions (Qian, McDonough et al. 2006). Few diapause specific proteins (cluster 24) are known from other animals. Diapause is a reversible state of developmental suspension. It is observed in diverse taxa, from plants to animals, including marsupials and some other mammals (Chen, Ge et al. 2009) as well as insects (associated molecular function varies but involves calcium channel inhibition (Kim, Nachman et al. 2008)) and should here support the tun formation or regulate other (e.g. developmental) metabolic inactive states. Furthermore, proteins involved in storage or transportation of fatty acids also seem to be important (cluster 31, (Alvarez-Ordonnez, Fernandez et al. 2009)). Late embryogenesis abundant (LEA) protein expression seems to be linked to desiccation stress and the acquisition of desiccation tolerance in organisms (Tunnacliffe and Wise 2007) e.g. nematodes (Goyal, Tisi et al. 2003; Browne, Dolan et al.

2004) and rotifers (Tunnacliffe, Lapinski et al. 2005). Thirty-one LEA type 1 family proteins were found in cluster 38.

Moreover, ten clusters (8, 18, 19, 30, 33, 35, 37, 42, 51, 55) consist of proteins which seem to be specific for tardigrades. These show no significant homology to known proteins.

***Functional clusters of stress-specific adaptations present in tardigrade***

To gain a systematic overview of involved tardigrade functions, all available tardigrade sequences were classified species-specific according to COG functional category (Tatusov, Galperin et al. 2000; Tatusov, Fedorova et al. 2003) as well as according to COG number and molecular function encoded. Note that in this section "protein" implies one type of protein.

A COG or KOG comprises often several sequences from different tardigrades. Prokaryotic (COG) and eukaryotic (KOG) gene clusters were compared (**Table 4-5**; details on the WEB http://waterbear.bioapps.biozentrum.uni-wuerzburg.de/). Again, several tardigrade-specific adaptations stand out, e.g. highly represented COGs regulate translation elongation factor and sulfate adenylate transferase and a strong ubiquitin system. There are many cysteine proteases (21 proteins). For redox protection there are 14 thioredoxin-domain containing proteins and 75 Heme/copper-type cytochrome/quinol-like proteins as well as ubiquinone oxidoreductase subunits (15 proteins). There are ten proteins involved in seleno-cysteine specific translation (Fagegaltier, Lescure et al. 2000; Lobanov, Hatfield et al. 2009). In eukaryotes, selenoproteins show a mosaic occurrence, with some organisms, such as vertebrates and algae, but notably also tardigrades, having dozens of these proteins, while other organisms, such as higher plants and fungi, having lost all selenoproteins during evolution (Lobanov, Hatfield et al. 2009).

Membrane GTPases (25 proteins) are often of Lep A (leader peptidase (March 1992)) type in tardigrades. In general, members of the GTPase superfamily regulate membrane signaling pathways in all cells. However, LepA, as well as NodO, are prokaryotic-type GTPases very similar to protein synthesis elongation factors but apparently have membrane-related functions (March 1992).

LEA proteins are wide-spread among plants and synthesized in response to certain stresses (Hong-Bo, Zong-Suo et al. 2005; Kobayashi, Maeta et al. 2008). The LEA type 1 family is well known in higher plants (rice, maize, carrots) to be synthesized during late embryogenesis and in ABA stress response. It includes desiccation-related protein PCC3-06 of Cratersostigma plantagineum. LEA type 1 family occurs in bacteria (e.g. Haemophilus influenzae, Dei inococcus radiodurans), but is atypical for animals. However, this is an animal example where LEA family type 1 is well represented and forms a full cluster.

We suggest that it will have similar function as known in other organisms and thus ensure protein translation (elongation factor) coupled to membrane integrity and possibly cytoskeletal rearrangement which would again boost the tardigrade resistance to stress. The KOGs show similar highly represented families and adaptations. Abberant proteins are rapidly recognized by ubiquitination-like proteins (220 proteins) and ubiquitin-ligase related enzymes (71 proteins) as well as proteasome regulatory subunits (85 proteins). For protein protection and refolding disulfide isomerases (26 proteins) and cyclophilin type peptidyl-prolyl cis-trans isomerases (43 proteins; KOG 0879-0885) are available. Connected to redox protection are also thirty AAA+type ATPases and three peroxisome assembly factor 2 containing proteins (KOG0736). This broad effort in protein protection is further supported by molecular chaperones (HSP70, mortalins and other; total of 50 proteins) and chaperonin complex components (32 proteins; KOG0356-0364). There are six superoxide dismutases and six copper chaperons for thioredoxins (37 proteins), glutaredoxin-like proteins (nine) and ten thiodisulfide isomerases as well as 52 glutathione-S-transferases. We found 22 hits to helicase. Tardigrade DNA protection is represented by 52 proteins of the molecular chaperone DNA J family: proteins of the DNA J family are classified into 3 types according to their structural domain decomposition. Type I J proteins compose of the J domain, a glyrich region connecting the J domain and a zinc finger domain, and possibly a C-terminal domain. Type II lacks the Zn-finger domain and type III only contains the J domain (Cheetham and Caplan 1998; Walsh, Bursac et al. 2004). The latter two are referred to as DnaJ-like proteins. Analysis of the domains present in tardigrade proteins by SMART (Letunic, Doerks et al. 2009) and Pfam (Finn, Tate et al. 2008) searches reveals only the J domain and in some cases a transmembrane region, identifying them as type III DnaJ-like proteins. For further information on these COGs/KOGs see **Table 4-6**.

*Table 4-4. CLANS clusters of sequence similar proteins in published tardigrade sequences*[1]

| Number/color | Cluster description | Sequences/percentage |
|---|---|---|
| 2 | Cytochrome c oxidase like (subunit 1, EC 1.9.3.1) | 425 (3.94%) |
| 3 | Uncharacterized protein U88/Glycosyltransferase 8 family | 302 (2.80%) |
| 5 | Proteins containing a Chitin binding domain | 191 (1.77%) |
| 6 | Proteins containing an IBR/Neuroparsin/DUF1096 domain | 189 (1.75%) |
| 7 | Fatty-acid binding protein (FABP) family | 127 (1.18%) |
| 8 | TSP2 (remote homology to Sericin 1) | 126 (1.17%) |
| 9 | Proteins containing a RNA polymerase Rpb3/Rpb11 dimerisation domain | 92 (0.85%) |
| 10 | Metallothionein superfamily (Type 15 family./Thioredoxin like) | 84 (0.78%) |
| 12 | GTP-binding elongation factor family. EF-Tu/EF-1A sub- family | 79 (0.72%) |
| 13 | GST superfamily. Sigma family | 78 (0.70%) |
| 14 | Ubiquitin family | 75 (0.69%) |
| 15 | Cathepsin family (EC 3.4.22.-) | 74 (0.67%) |
| 16 | Carboxypeptidase A inhibitor like | 72 (0.64%) |
| 17 | Trichohyalin/Translation initiation factor like | 69 (0.60%) |
| 18 | TSP2 | 65 (0.57%) |
| 19 | TSP2 | 61 (0.56%) |
| 20 | RNA/DNA-binding proteins | 60 (0.55%) |
| | ... | |
| 23 | Small Heat Shock Protein (HSP20) family | 53 (0.47%) |
| 24 | Diapause-specific proteins | 51 (0.44%) |
| | ... | |
| 38 | LEA type 1 family proteins | 31 (0.28%) |
| | ... | |

[1]Shown are the number of proteins found for the specified cluster, their percentages and the corresponding cluster number in **Figure 4-8.** The full Table with all clusters and their color code matching to **Figure 4-8** is given in [appendix A: **Table A-1** ] [2]Tardigrade specific protein clusters

*Table 4-5. Highly represented protein functions in Tardigrades (COGs and KOGs)*

| |
|---|
| *Information from COG clusters*[1]: |
| Information storage and processing |
| 75 Translation elongation factor EF-1 (COG5256) |
| 64 GTPases - translation elongation (COG0050) |
| 58 Peptide chain release factor RF-3 (COG4108) |
| Cellular processes and signaling |
| 31 Ubiquitin (COG5272) |
| 25 Membrane GTPase LepA (COG0481) |
| 21 Cysteine protease (COG4870) |
| Metabolism |
| 75 Heme/copper-type cytochrome/quinol oxidases (COG0843) |
| 67 GTPases - Sulfate adenylate transferase (COG2895) |
| Poorly characterized |
| 11 Dehydrogenases with different specificities (COG1028) |
| 11 Uncharacterized homolog of Blt101 (COG0401) |
| *Information from KOG clusters*[1]: |
| Information storage and processing |
| 77 Translation elongation factor EF-1 (KOG0052) |
| 71 Polypeptide release factor 3 (KOG0459) |
| 70 Elongation factor 1 alpha (KOG0458) |
| 53 Mitochondrial translation elongation factor Tu (KOG0460) |
| Cellular processes and signaling |
| 52 Glutathione S-transferase (KOG1695) |
| 46 Cysteine proteinase Cathepsin L (KOG1543) |
| 34 Apolipoprotein D/Lipocalin (KOG4824) |
| 31 Cysteine proteinase Cathepsin F (KOG1542) |
| Metabolism |
| 78 Cytochrome c oxidase (KOG4769) |
| 74 Fatty acid-binding protein FABP (KOG4015) |
| Poorly characterized |
| 31 Ubiquitin and ubiquitin-like proteins (KOG0001) |
| 16 GTPase Rab18, small G protein superfamily (KOG0080) |
| 15 Ras-related GTPase (KOG0394) |
| 15 GTPase Rab21, small G protein superfamily (KOG0088) |

[1]Detailed data and all COG/KOG numbers are given on the WEB page http://waterbear.bioapps.biozentrum-uni-wuerzburg.de/cgi-bin/cog_stat.pl. Summarized here are the functions of those clusters of orthologous groups (COGs) occurring particularly often or suggesting tardigrade specific adaptations.

Moreover, undesired proteins can be rapidly degraded by cathepsin F-like proteins (31 proteins) or L-like proteins (46 proteins). There are several calcium-dependent protein kinases (25 proteins; KOG0032-0034) and actin-bundling proteins. According to this observation calcium signaling should be implicated in adaptive rearrangement of the cytoskeleton during tardigrade rehydration. The cytoskeleton is a key element in the organisation of eukaryotic cells. It has been described in the literature that the properties of

actin are modulated by small heat-shock proteins including a direct actin-small heat-shock protein interaction to inhibit actin polymerization to protect the cytoskeleton (Mounier and Arrigo 2002; Sun and MacRae 2005) (compare with the CLANS cluster 24 (Diapause proteins) found in the above analysis).

Translation in tardigrades includes polypeptide release factors (71 proteins) and proteins for translation elongation (77 proteins). There are about 80 GTP-binding ADPribosylation factors. The secretion system and Rab/Ras GTPases are fully represented (183 proteins). Seventeen tubulin anchor proteins show that the cytoskeleton is well maintained. Finally, we find 14 TNF- associated factors and 34 apolipoprotein D/lipocalin proteins.

### *Typical motifs in tardigrade mRNAs*

The regulatory motif search showed a number of known regulatory RNA elements involved in tardigrade mRNA regulation (**Table 4-7** for *H. dujardini* and *M. tardigradum*). Certainly it cannot be formally ruled out that some of these elements work in a tardigrade modified way. Similarly, there are probably further patterns which are tardigrade specific, but not detected with the UTRscan software (Pesole and Liuni 1999) applied for analysis.

*Table 4-6. Identified DnaJ-family COGs/KOGs in Tardigrades and Milnesium tardigradum*[1].

| KOG/COG number | COG distribution | COG name | present in Tardigrades | present in *M. tardigradum** |
|---|---|---|---|---|
| COG0484 | | DnaJ-class molecular chaperone with C-terminal Zn finger domain | 5 | |
| COG2214 | | DnaJ-class molecular chaperone | 8 | |
| KOG0550 | A-DH-P- | Molecular chaperone (DnaJ superfamily) | 3 | 2 |
| KOG0691 | ACDHYP- | Molecular chaperone (DnaJ su perfamily) | 7 | 2 |
| KOG0712 | ACDHYPE | Molecular chaperone (DnaJ su perfamily) | 8 | 2 |
| KOG0713 | ACDH--- | Molecular chaperone (DnaJ su perfamily) | 5 | 1 |

[1]Shown are the number of proteins found for the specified COG/KOG number, the KOG distribution of the KOG in different eukaryotic species (see abbreviations) and the COG/KOG annotation.

Abbreviations: A *Arabidopsis thaliana*, C *Caenorhabditis elegans*, D *Drosophila melanogaster*, H *Homo sapiens*, Y *Saccharomyces cerevisiae*, P *Schizosaccharomyces pombe*, E *Encephalitozoon cuniculi*. *These include specific unpublished data from ongoing work on *M. tardigradum*

The RNA elements found include the lox-P DICE element (Ostareck-Lederer, Ostareck et al. 1998) in *H. dujardini* as top hit with as many as 1,269 ESTs (23.6% of all *H. dujardini* EST sequences). The cytidinerich 15-lipoxygenase differentiation control element (15-LOX DICE, (Ostareck-Lederer, Ostareck et al. 1994)) binds KH domain proteins of the type hnRNP E and K (stronger in multiple copies), mediating mRNA stabilization and translational control (Ostareck-Lederer, Ostareck et al. 1998).

Furthermore, a high number of mRNAs contains K-Boxes (cUGUGAUa, (Lai, Burks et al. 1998)) and brd Boxes (AGCUUUA, (Lai 2002)). All these elements are involved in mRNA storage and mRNA stability. These two elements are potential targets for miRNAs as shown in *Drosophila melanogaster* (Lai, Tam et al. 2005).

However, in the two tardigrade species compared, only 16 of 30 well known RNA elements are found, suggesting a clear bias in tardigrade mRNA regulation. For example, the widely used AU rich elements in higher organisms (Pesole and Liuni 1999) such as vertebrates are absent in tardigrades.

Regulatory elements in tardigrade mRNA are probably important for their adaptation, in particular to support transformation to tun stage and back to active stage again. The list of RNA elements found can be compared for instance to our data on regulatory elements in human anucleate platelets (Dittrich, Birschmann et al. 2006) where mRNAs have to be stockpiled for the whole life of the platelet. Due to this comparatively long life, a long mRNA untranslated region is important in these cells. The same should apply to tardigrade mRNAs, since their average UTR is predicted to be long. A different stock-piling scenario occurs in unfertilized eggs, but due to developmental constraints, here localization signals are often in addition important for developmental gradients. We tested for these in tardigrades but did not find a high representation of localization motifs.

***Web-tool tardigrade analyzer***

We created a convenient platform to allow rapid sequence comparisons of new protein sequences, in particular from new sequencing efforts in tardigrades, to our database by applying rapid heuristic local alignment using BLAST (Altschul, Madden et al. 1997) and allowing to search in selected species.

*Table 4-7. Regulatory elements in Hypsibius dujardini[1] and Milnesium tardigradum[2] mRNA sequences.*

| Motiv | *Hypsibius dujardini* | *Milnesium tardigradum* |
|---|---|---|
| 15-LOX-DICE | 1528 (1269)[3] | 46 (45)[3] |
| ADH DRE | 60 (58)[3] | 1 (1)[3] |
| BRE | 1(1)[3] | -- |
| Brd-Box | 152 (149)[3] | 28 (22)[3] |
| CPE | 21 (21)[3] | 15 (15)[3] |
| Elastin G3A | 1 (1)[3] | -- |
| GLUT1 | 1 (1)[3] | -- |
| GY-Box | 406 (372)[3] | 21 (21)[3] |
| IRE | 1 (1)[3] | -- |
| IRES | 1353 (1353)[3] | 90 (90)[3] |
| K-Box | 469 (447)[3] | 35 (33)[3] |
| SECIS-1 | 1 (1)[3] | -- |
| SECIS-2 | 6 (6)[3] | -- |
| TGE | 5 (5)[3] | 1 (1)[3] |
| TOP | 50 (50)[3] | 1 (1)[3] |

[1]We considered 5.378 ESTs in *H. dujardini*.[2] We considered 607 ESTs in *M. tardigradum*. [3] The number of hits is followed by the number of mRNAs with this hit in brackets to indicate multiple hits.

A batch mode allows the analysis of up to 100 sequences simultaneously when uploaded in a file in FASTA format. Output data are displayed according to an enhanced BLAST output format with graphical illustrations. Low expected E-values result for searches using the option of our tardigrade specific databases: a more specific smaller database reduces the probability of false positives. As an alternative for general sequence analysis, a search against the non-redundant database of GenBank can be performed. This takes more computational power and yields higher E-values, however, it identifies functions for most sequences. An additional useful feature is to scan all available data for peptide motifs or PROSITE signatures using a "pattern" module or assign potential functions by COGs (Tatusov, Galperin et al. 2000). The first is helpful to recognize tardigrade proteins in cases where the tardigrade sequence has diverged far, and only critical residues for function are still conserved as motif signatures. It can also be applied to search for regulatory RNA motifs such as polyadenylation sites (e.g. AAUAAA or AAUUAA) or recognize promotor modules such as the glucocorticoid receptor element (GRE; palindromic pattern: AGAA CAnnnTGTTCT). For this purpose, both, the tardigrade sequences and the non redundant database can be searched (e.g. to look for stress-specific regulatory RNA elements; appendix

B: fig. B-2). Interestingly, this nucleotide (RNA or DNA) specific option is not available on some common servers, e.g. the PHI-BLAST (Zhang, Schaffer et al. 1998) server at NIH. Further options include a user-defined database and interactively animated stress clusters (**Figure 4-8**). The tool http://waterbear.bioapps.biozentrum.uniwuerzburg.de/ allows rapid searches for tardigrade specific sequences, e.g. molecular adaptations against stress. For instance, a search for trehalase sequences shows no trehalase mRNA in the *H. dujardini* sequences. In contrast, there are several heat shock proteins in tardigrades, an example is HSP90 proteins (identified by sequence similarity as well as by a pattern hit based approach using the PROSITE entry PS00298 with the signature Y-x- [NQHD]- [KHR]- [DE]- [IVA]- F- [LM]-R- [ED]; **Table 4-8**). Specific COGs are also rapidly assigned for any desired sequence. This includes the option to map the query sequence of interest to any of the known tardigrade specific COGs. Furthermore, nucleotide patterns such as mRNA polyadenylation sites are rapidly identified e.g. in *H. dujardini* mRNAs. Similarly, other mRNA 3'UTR elements can be identified, e.g. AU rich sequences mediating mRNA instability or regulatory K-boxes (motif cUGUGAUa, (Lai, Burks et al. 1998)) in tardigrades.

*Table 4-8. HSP90 proteins identified in Hypsibius dujardini using the Tardigrade analyzer*[1].

| Hit | Predicted function/name (Tardigrade analyzer) | Pattern matched | Start position | End |
|---|---|---|---|---|
| gi:37213462 | hsp90[2] | YSNKEIFLRE | 68 | 77 |
| gi:37213713 | hsp90[2] | YSNKEIFLRE | 70 | 79 |

[1]These are hits using the pattern hit option and the heat shock protein PROSITE entry PS00298 for pattern generation and recognition. The pattern has the signature of Y – x – [NQHD] – [KHR] – [DE] – [IVA] – F- [LM] – R – [ED]. [2]Predicted similarity to Q7PT10 (HSP83 ANOGA) from Swissprot

### *Implications*

Tardigrades show a surprising large amount of related sequences. Certainly, one has to correct for a few genes sequenced from many lineages for phylogenetic studies in tardigrades (cytochrome c, rRNA etc.) However, despite this, a number of tardigrade-specific clusters still remain. Furthermore, **Table 4-1** shows that most of the annotated clusters are stress-related.

Looking at specific protein functions, both COG and KOG proteins show that tardigrades spend an extraordinary effort in protein protection, turnover and recycling as well as redox protection. Some other specific adaptations become apparent also from **Table 4-5**, but the complete extent of these adaptations is unclear given the limited sampling of available tardigrade sequences. Furthermore, protection of DNA is critical as it has been shown that tardigrade tuns accumulate DNA damage which first has to be repaired before resurrection occurs (Schill, Neumann et al. 2008; Neumann, Reuner et al. 2009). Taking this into consideration, DNA J proteins were investigated in more detail since proteins of this family are well represented in tardigrades, including several COGs and KOGs. Several data underline the extremely high resistance of tardigrades to temperature, pressure and radiation as well as a high repair potential regarding DNA (Jonsson, Rabbow et al. 2008; Neumann, Reuner et al. 2009). Thus, we suggest that the high repair potential is also mediated by this well represented protein family. Phylogenetic analysis (**Table 4-6**) shows that these proteins are represented by several KOGs as well as the classic COGs in tardigrades. In particular, the first three KOG families are also used in *M. tardigradum*, where extreme stress tolerance requires strong repair mechanisms (Kinchin and Dennis 1994). Furthermore, all these tardigrade proteins in **Table 4-6** are small, having neither zincfinger domains nor low complexity regions, but instead consisting of single DNA J domains which would always place them in type I (subfamily A) of DNA-J like proteins. This suggests that the direct interaction with DNA-J like proteins is the key molecular function.

Finally, we could show that there are 16 regulatory elements used in tardigrade mRNA, while a number of other elements known from higher eukaryotic organisms and vertebrates is not used. It is interesting to note that the elements often used in tardigrades are all involved in regulation of mRNA stability. Thus, they may be implicated in stage switching, as presumably in the initial phases of the tun awakening or tun formation, new supply of mRNA is turned off and instead regulation of synthesized mRNA becomes important.

In addition, and for further research we supply the web tool tardigrade analyzer. There are a number of alternative tools available, e.g. from NCBI http://www.ncbi.nlm.nih.gov/. However, we offer some species-specific searches not available from these sources as well as RNA and promotor pattern search (not only for tardigrades but also for NRDB; not available from NIH). Furthermore, there are functional COG predictions as well as new, unpublished tardigrade sequences from *M. tardigradum*, all above reported data including the reported

sequences and detailed functional clusterings as well as regular server updates. A better understanding of the survival mechanisms in these organisms will lead to the development of new methods in several areas of biotechnology. For example, preservation of biological materials *in situ*, macromolecules and cells from non-adapted organisms (Schill, Mali et al. 2009). This is, of course, only a first and very general overview on potential tardigrade specific adaptations, more species-specific data will be considered as more information becomes available.

### *Conclusion*

Tardigrade genomes invest in stress-specific adaptations, this includes major sequence related protein clusters, functional clusters for stress as well as specific regulatory elements in mRNA. For further tardigrade genome analysis we offer the tardigrade workbench as a flexible tool for rapid and efficient analysis of sequence similarity, protein function and clusters, COG membership and regulatory elements.

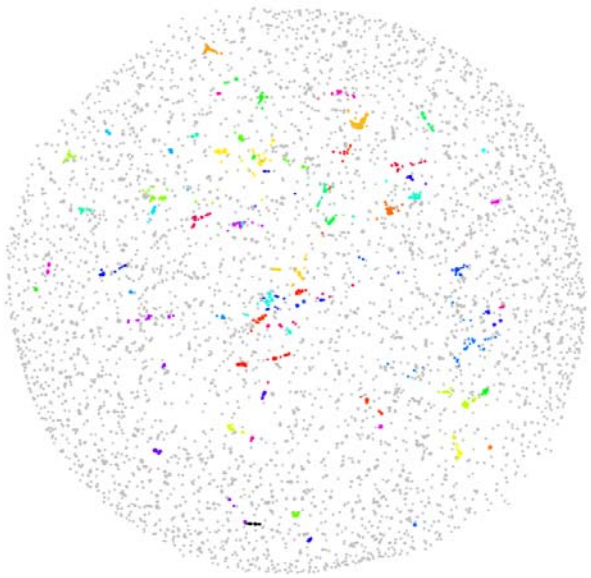


Figure 4-8. Functional clusters by CLANS of sequence related proteins in tardigrades. All available tardigrade protein sequences were clustered in a 3D sphere according to their sequence distance and were projected to the paper plane. Individual protein functions are colored [for color code see appendix A: **Table A-1**] and all listed in **Table 4-1**. Functional clusters appear as patches of an individual color. Color code and clusters can be interactively examined at the Tardigrade workbench http://waterbear.bioapps.biozentrum.uni-wuerzburg.de

Chapter 5

# Structural flexibility of biological macromolecules studied by SAXS

One of the major advantages of SAXS is that it can probe structure on a broad range of macromolecular sizes under near native conditions and different levels of flexibility, from rigid molecules with flexible inter-domain hinge regions (Verstraete, Vandriessche et al. 2011) to natively unfolded proteins, like e.g. tau (Mylonas, Hascher et al. 2008; Shkumatov, Chinnathambi et al. 2011).

Chapter 5 of this thesis is devoted to the SAXS projects, conducted in collaboration with external users at the EMBL X33 beamline. These projects involved analysis of the scattering data and model building.

The following subchapters are ordered according to the degree of flexibility in the studied biological macromolecules. The class-III receptor tyrosine kinase (RTKIII) Flt3 and its cytokine ligand (FL) play central roles in hematopoiesis and the immune system, by establishing signaling cascades crucial for the development and homeostasis of hematopoietic progenitors and antigen-presenting dendritic cells. Flt3 has two receptor arms, which consist of five Ig-like domains each. The flexible inter-domain hinge regions are comprised of 3-5 amino acids. In a collaborative effort, the structure of Flt3-FL complex was investigated using a variety of structural and biophysical methods. The results of this study are presented in subchapter 5.1

Using SAXS the solution structure of frataxin was studied, and the results of this study are presented in subchapter 5.2. Frataxin is a mitochondrial protein with a central role in iron homeostasis, including iron storage, detoxification, etc. Deficiency in frataxin function has been attributed to the progressive neurodegenerative disease Friedreich's ataxia. Frataxin is a relatively small molecule of 13.7 kDa, consisting of folded domain and flexible N-terminus of about 20 amino acids. We investigated the influence of different factors on oligomerization properties of frataxin in order to find which mechanisms control the oligomeric state of frataxin and may be important for frataxin oligomerization *in vivo*.

Subchapter 5.3 is dedicated to the SAXS analysis of PDZRhoGEF (PRG), which belongs to a small family of RhoA-specific RGS guanine nucleotide exchange factors (GEFs), mediating signaling through select 7-transmembrane receptors *via* $G\alpha_{12}$ and activating RhoA by catalyzing the exchange of GDP to GTP. PRG is a multidomain protein, composed of PDZ and RGSL domains followed by a catalytic tandem of DH and PH domains. The RGSL domain is preceded by a ~150 residue long linker (L1), which appears

to be disordered based on *in silico* sequence analysis. RGSL domain is followed by yet another linker (L2), ~200 residues in length, also presumed to be disordered. Using SAXS, we investigated the supramodular architecture of the PRG to propose a novel, two-tier model of PRG autoinhibition.

In the last subchapter the use of SAXS for investigation of the temperature dependent properties of tau protein are described. Tau is an IDP without a well-defined structure, which is involved in Alzheimer's disease. Physiologically, tau is a microtubule-associated protein, occurring mainly in the axons of neurons, where it stabilizes microtubules and the tau-microtubule interactions are known to be very temperature sensitive. Tau protein contains 4 semi-conserved sequences of 31 or 32 residues, so-called "repeats". Using different temperature protocols we investigated the structural properties of tau in solution. The results of this study contributed to the understating of IDPs structure and function.

## 5.1 Structural insights into the extracellular assembly of the hematopoietic Flt3 signaling complex

*Adapted from*

Verstraete, K., Vandriessche, G., Januar, M., Elegheert, J., Shkumatov, A. V., Desfosses, A., Van Craenenbroeck, K., Svergun, D. I., Gutsche, I., Vergauwen, B. and Savvides, S. N.

*Blood* (2011), blood-2011-2001-329532

Flt3 and its cytokine ligand (FL) play central roles in hematopoiesis and the immune system, by establishing signaling cascades crucial for the development and homeostasis of hematopoietic progenitors and antigen-presenting dendritic cells. However, Flt3 is also one of the most frequently mutated receptors in hematological malignancies and is currently a major prognostic factor and clinical target for acute myeloid leukemia (AML). Here, we report the structural basis for the Flt3 ligand-receptor complex and unveil an unanticipated extracellular assembly unlike any other RTKIII/V complex characterized to date. FL induces dimerization of Flt3 via a remarkably compact binding epitope localized at the tip of extracellular domain 3 of Flt3, and invokes a ternary complex devoid of homotypic receptor interactions. Comparisons of Flt3 with homologous receptors and available mutagenesis data for FL have allowed us to rationalize the unique features of the Flt3 extracellular assembly. Furthermore, thermodynamic dissection of complex formation points to a pronounced enthalpically-driven binding event coupled to an entropic penalty. Together, our data suggest that the high-affinity Flt3-FL complex is driven in part by a single preformed binding epitope on FL reminiscent of a 'lock-and-key' binding mode, thereby setting the stage for antagonist design.

**State of the problem**

Hematopoiesis is a finely regulated process during which diverse cell types originating from a limited and self-renewing population of hematopoietic stem cells (HSC), are stimulated to proliferate and differentiate to create the cellular repertoire that sustains the mammalian hematopoietic and immune systems (Metcalf 2008). The Fms-like tyrosine kinase receptor 3 (Flt3) is the most recent addition to the diverse family of hematopoietic receptors. Flt3 is activated on HSC and early myeloid and lymphoid progenitors by its cognate ligand (FL), to initiate downstream signaling via the PI3K/AKT and the RAS/RAF/MEK/ERK pathways (Stirewalt and Radich 2003; Parcells, Ikeda et al. 2006). Consistent with the narrow expression profile of Flt3 in the bone marrow environment, signaling via the Flt3 ligand/receptor complex primarily impacts early hematopoiesis, particularly the proliferation and development of HSC and B-cell progenitors (Stirewalt and Radich 2003; Kikushige, Yoshimoto et al. 2008). In recent years Flt3 and FL emerged as potent regulators of dendritic cell (DC) development and homeostasis (Onai, Obata-Onai et al. 2007; Waskow, Liu et al. 2008; Liu, Victora et al. 2009), and DC-mediated natural killer cell activation (Eidenschenk, Crozat et al. 2010), thereby gaining an important role at the interface of innate and acquired immunity and in cancer immunotherapy (Dong, McPherson et al. 2002; Wu and Liu 2007). Notably, Flt3/FL-driven DC generation yields both classical- and plasmacytoid DC from bone-marrow progenitors regardless of myeloid or lymphoid commitment, a property that is currently unmatched by any other receptor/cytokine system relevant for DC physiology (Liu and Nussenzweig 2010; Schmid, Kingston et al. 2010).

RTKIII together with the prototypic platelet-derived growth factor receptors (PDGFRα/β), colony-stimulating factor 1 receptor (CSF-1R), and KIT (Grassot, Gouy et al. 2006). Thus, Flt3 has been predicted to display a modular structure featuring an extracellular segment with 5 immunoglobulin (Ig)-like domains (residues 27-543), a single transmembrane helix (TM, residues 544-563), a cytoplasmic juxtamembrane domain (JM, residues 572-603) and a split intracellular kinase module (residues 604-958). The RTKIII family is closely related to the RTKV family of vascular endothelial growth factor receptors (VEGFR), which have 7 extracellular Ig-like domains. The hallmark of RTKIII/V signaling lies in the activation of the extracellular receptor segments upon binding of the cognate cytokines, followed by intermolecular autophosphorylation and activation of the intracellular kinase domains (Lemmon and Schlessinger 2010).

Besides the clear role of Flt3 signaling in hematopoiesis and immune system development, overexpression of wild type or oncogenic forms of Flt3 have been implicated in a number of hematopoietic malignancies (Stirewalt and Radich 2003; Sanz, Burnett et al. 2009), and inflammatory disorders (Dehlin, Bokarewa et al. 2008). In particular, internal tandem duplication (ITD) in the JM region or point mutations in the kinase activation loop occur in 35% of patients with AML resulting in constitutive activation of the receptor and uncontrolled proliferation of hematopoietic precursors (Kiyoi, Ohno et al. 2002; Stirewalt and Radich 2003; Parcells, Ikeda et al. 2006; Reindl, Bagrintseva et al. 2006; Frohling, Scholl et al. 2007). Such mutation fingerprints have established Flt3 as the predominant prognostic factor in AML cases (Eklund 2010), and have rationalized the targeting of Flt3 in a clinical setting (Stirewalt and Radich 2003; Parcells, Ikeda et al. 2006; Kindler, Lipka et al. 2010).

Although the cellular and physiological role of the Flt3 ligand-receptor interaction has been featured prominently in the biomedical literature over the last two decades, the Flt3 signaling complex has remained uncharacterized at the molecular and structural level. Such insights are the missing link to the structural and functional diversity of RTKIII/V extracellular complexes, and would help provide a nearly complete picture of the entire Flt3 signaling complex given the available structure of the Flt3 intracellular kinase domains (Griffith, Black et al. 2004). A recent flurry of studies of RTKIII/V extracellular complexes led to a structural paradigm for RTKIII/V activation, whereby the receptors bind via their N-terminal Ig-like domains to the activating dimeric cytokine and concomitantly make homotypic contacts between their membrane-proximal domains (Liu, Chen et al. 2007; Ruch, Skiniotis et al. 2007; Yuzawa, Opatowsky et al. 2007; Chen, Liu et al. 2008; Yang, Yuzawa et al. 2008; Shim, Liu et al. 2010; Leppanen, Prota et al. 2010).

A universal feature of all characterized RTKIII/V complexes thus far is that the cytokine-binding epitope is distributed equally between extracellular domains 2 and 3 covering ~2000 $Å^2$ of surface area, and that homotypic receptor-receptor interactions are mediated by well-conserved residues in the membrane-proximal domains (D4 in RTKIII and D7 in RTKV). Nonetheless, Flt3 appears to be an outlier among RTKIII/V receptors due to several unique features in its extracellular segment (Lyman, James et al. 1993; Maroc, Rottapel et al. 1993), thus raising the question whether the current structural paradigm could be extrapolated to Flt3. Notably, Flt3 exhibits intragenic homology relating extracellular domains 1 and 4, and

domains 2 and 5, indicative of an ancient internal duplication event during evolution. Furthermore, Flt3 contains 12 additional cysteines that are not present in any of the homologous receptors, and has a unique N-terminal sequence of 50 amino acids preceding Ig-like domain 1. Interestingly, a fully functional splice variant of murine Flt3 lacks extracellular domain 5 entirely, indicating that the domain most proximal to the cell-membrane is not critical for receptor activation (Lavagna, Marchetto et al. 1995) contrary to other RTKIII/V receptors (Broudy, Lin et al. 2001; Yang, Yuzawa et al. 2008; Yang, Xie et al. 2010). Here, we provide the structural basis of extracellular complex formation between Flt3 and its cognate cytokine. Our studies establish the uniqueness of Flt3 within the RTKIII/V family and provide a reference platform for further structure-function studies and antagonist design.

**Isolation of recombinant Flt3 ectodomain complexes and thermodynamic binding profile of complex formation**

High-affinity complexes of purified glycosylated $Flt3_{D1-D5}$, $Flt3_{D1-D4}$, and $Flt3_{D1-D3}$ with recombinant human FL produced in *E. coli* (Verstraete, Koch et al. 2009), consistent with bivalent binding of FL to each of the ectodomain constructs, were initially established by analytical size-exclusion chromatography (SEC). Subsequent batches for structural studies were obtained via preparative SEC in the presence of excess molar amounts of purified FL (**Fig. 5-1a-c**). The elution profiles for all three ectodomain complexes were indicative of ligand-induced receptor dimerization. In constrast to $Flt3_{D1-D5}$ and $Flt3_{D1-D4}$, preparations of recombinant $Flt3_{D1-D3}$ consistently contained a significant portion of receptor that was incapable of binding the ligand even in the presence of excess molar amounts of FL (**Fig. 5-1c**). Conversely, excess molar amounts of $Flt3_{D1-D3}$ did result in a complete titration of FL towards complex formation. On the other hand, we were not able to observe complex formation for $Flt3_{D1}$ and $Flt3_{D1-D2}$ via SEC providing direct evidence that these ectodomain constructs do not carry a high-affinity ligand binding site.

Characterization of Flt3 extracellular complexes by isothermal titration calorimetry (ITC) led to a number of consensus observations (**Fig. 5-1d-f**). Firstly, all three characterized

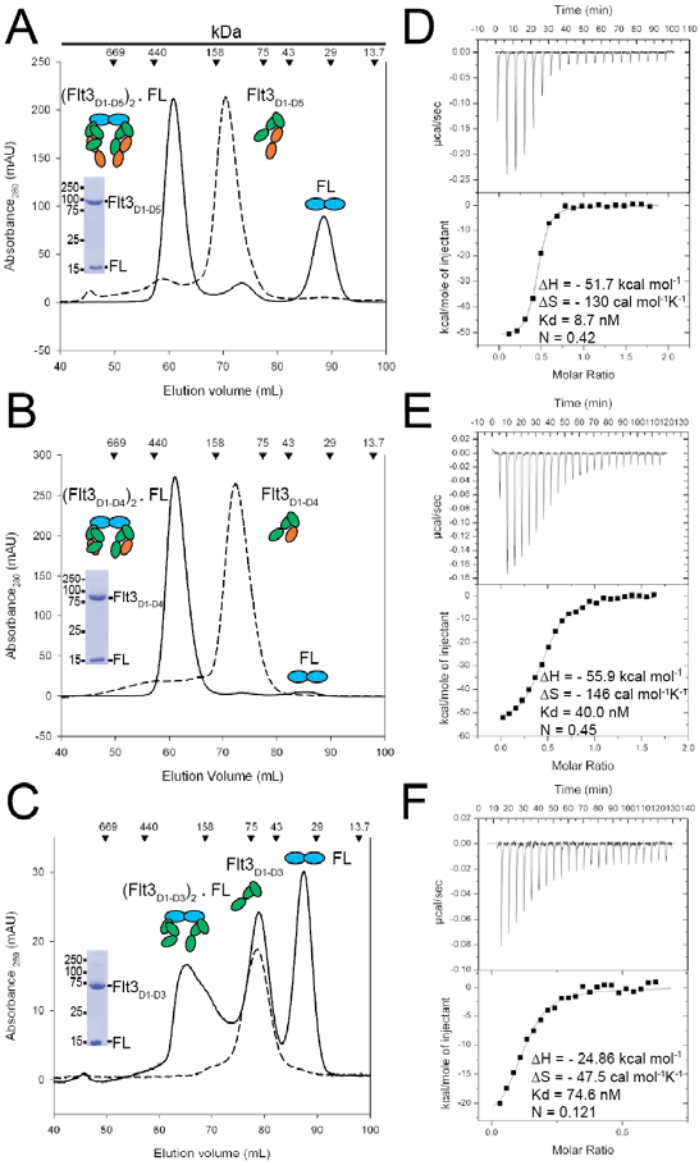



Figure 5-1. FL binds bivalently to Flt3 ectodomain variants to form high-affinity complexes. (A-B) Isolation of $Flt3_{D1-D5}$:FL and $Flt3_{D1-D4}$:FL by SEC. Also shown, are Coomassie-stained SDS-PAGE (sodium dodecyl sulfate-polyacrylamide gel) strips corresponding to the peak fraction of the isolated complexes. (C) Size-exclusion chromatography on the $Flt3_{D1-D3}$:FL mixture at the end of an ITC experiment, showing that a large amount of $Flt3_{D1-D3}$ remains in the unbound form. Identical elution profiles were obtained in standard SEC experiments as well, in the presence of a large molar excess of FL. (D-F) Binding isotherms and thermodynamic parameters of FL binding to Flt3 ectodomains obtained by ITC. All ITC experiments were carried out by titrating recombinant human Flt3 extracellular domains with FL.

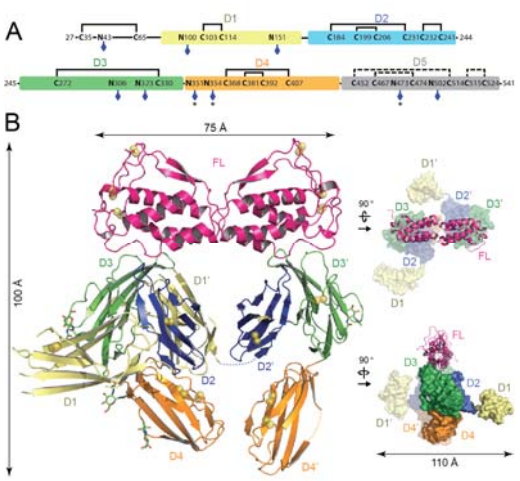



Figure 5-2. Crystal structure of the $Flt3_{D1-D4}$:FL complex. (A) Domain organization of the Flt3 extracellular segment. The five Ig-like domains of Flt3 (D1: residues 79-161, D2: residues 167-244, D3: residues 245-345, D4: residues 348-434 and D5: residues 435-533) are shown as colored boxes. N-linked glycosylation sites are indicated with blue diamonds. Partially occupied glycosylation sites are indicated with an asterisk. Also shown is the disulfide bond network in $Flt3_{D1-D4}$ as determined by mass-spectrometry. The putative disulfide bridges in $Flt3_{D5}$ are shown as dashed lines, based on homology with $Flt3_{D2}$ and $KIT_{D5}$. (B) Overall structure of the $Flt3_{D1-D4}$:FL complex. The crystal structure of the $Flt3_{D14}$:FL complex is shown in ribbon representation with the twofold symmetry axis of FL oriented along the vertical axis of the plane. Flt3 domains follow the coloring scheme in panel A. Disulfide bridges are shown as yellow spheres and N-linked glycans as green sticks. The structural panels to the right show two alternative views of the complex with FL in ribbon representation and the receptor in surface representation.

complexes exhibit high-affinity binding characterized by a strongly exothermic enthalpic term coupled to an entropic penalty. Secondly, FL exhibits bivalent binding to both $Flt3_{D1-D5}$ and $Flt3_{D1-D4}$ (N=0.5, 2 molecules of Flt3 to 1 molecule FL). We note that the observed stoichiometry for the FL: $Flt3_{D1-D3}$ interaction (N=0.12) is likely due to the inability of recombinant $Flt3_{D1-D3}$ to engage completely in a ternary complex in the presence of a molar excess of FL (**Fig. 5-1c**). Nonetheless, the ITC data clearly show that $Flt3_{D1-D3}$ is capable of a high-affinity ternary complex just like the larger ectodomain constructs, and that only about 25% of recombinant $Flt3_{D1-D3}$ may adopt an active conformation. Notably, the sequential exclusion of the membrane-proximal domains $Flt3_{D4}$ and $Flt3_{D5}$ leads to a very modest decrease in affinity ($K_d$ [$Flt3_{D1-D5}$:FL] = 8.7 nM ; $K_d$ [$Flt3_{D1-D4}$:FL] = 40 nM ; $K_d$ [$Flt3_{D1-D3}$:FL] = 74.6 nM) while the thermodynamic profiles remain similar (**Fig. 5-1d-f**). Taken together, our data suggest that the membrane-proximal module $Flt3_{D4-D5}$ does not contribute significantly to the overall stability of the complex.

### Overall structure of the $Flt3_{D1-D4}$:FL complex

The crystal structure of the $Flt3_{D1-D4}$:FL complex was determined to 4.3 Å resolution based on data obtained from a large-scale screening of crystals. Confronted with the recurring poor diffraction quality of $Flt3_{D1-D4}$:FL crystals derivatized with heavy-atoms and selenomethionine-labeled $Flt3_{D1-D4}$, we successfully combined molecular replacement strategies relying on the high resolution structure of human FL (Savvides, Boone et al. 2000) and a homology model for $Flt3_{D3}$ (Yuzawa, Opatowsky et al. 2007) (**Table 5-1**), with phase improvement protocols (Cowtan 2010) exploiting the non-crystallographic symmetry and high solvent content of the crystals. Such approaches combined with crystallographic refinement employing information from high resolution structures have recently emerged as a powerful option in macromolecular structure determination at low resolution (Blanc, Roversi et al. 2004; Huyton, Zhang et al. 2007; Schroder, Levitt et al. 2010). The ensuing electron density maps were exceptionally revealing and contained contiguous electron density for several unmodeled receptor domains, including direct crystallographic evidence for N-linked glycans (appendix B: **Fig. B-1a**). To facilitate chain tracing we determined the atypical disulfide-bond network of Flt3 as well as the actual number of N-linked glycosylation sites in extracellular Flt3 by mass-spectrometry. We could confirm that all nine

N-linked glycosylation sites are at least partially occupied and that all cysteine residues present in $Flt3_{D1-D4}$ are engaged in disulfide-bonds (**Fig. 5-2a**, appendix B: **Table B-1**).

The structure of the $Flt3_{D1-D4}$:FL complex is unlike any of the structurally characterized RTKIII/V complexes to date and is characterized by a number of surprising features (**Fig. 5-2b**, appendix B: **Fig. B-1** and **Fig. B-2a**). The $Flt3_{D1-D4}$:FL assembly can be described as a moderately open horseshoe ring structure measuring 100 Å x 75 Å x 110 Å, comprising FL, $Flt3_{D2}$, $Flt3_{D3}$ and $Flt3_{D4}$. FL binds to two receptor molecules bivalently and is accommodated by a binding epitope at the membrane-distal tip of $Flt3_{D3}$, while $Flt3_{D2}$ leans against the concave side of $Flt3_{D3}$ and is stowed underneath FL in the ring opening (**Fig. 5-2b**, appendix B: **Fig. B-1b**). Intriguingly, the apparent two-fold symmetry of the complex about the FL dimer interface only holds for the FL:$Flt3_{D2-D3}$ subcomplex, as both $Flt3_{D1}$ and $Flt3_{D4}$ adopt asymmetric orientations compared to their tandem modules in the complex (**Fig. 5-2b**). Remarkably, $Flt3_{D4}$ does not engage in any obvious homotypic interactions as seen in the KIT structure (Yuzawa, Opatowsky et al. 2007). The N-terminal $Flt3_{D1}$ exhibits significant disorder and domain plasticity manifested by at least two different orientations about the D1-D2 linker region (residues 162-166), and protrudes perpendicularly away from the plane of the ring assembly at the level of $Flt3_{D2}$ without making any interactions with the rest of the complex (**Fig. 5-2b**). Our electron density maps allowed us to reliably model only the core of the $Flt3_{D1}$ structure (residues 79-161), but residual positive difference electron density extending away from the N-terminus of our model suggested that the atypical 50 amino acid module preceding $Flt3_{D1}$ is likely associated with the core domain structure.

**Flt3 employs a remarkably compact cytokine-binding epitope**

Perhaps the most unanticipated feature of the $Flt3_{D1-D4}$:FL complex is that the ligand-binding epitope is almost exclusively contributed by $Flt3_{D3}$ (**Fig. 5-3a**) for which electron density was exceptionally clear including information for some side-chains. This module is a member of the "I-set" Ig domains and is structurally homologous to extracellular domain 3 of KIT (Liu, Chen et al. 2007; Yuzawa, Opatowsky et al. 2007) and CSF-1R (Chen, Liu et al. 2008), featuring 8 β-strands making up the *ABED* and *A'FGC* β-sheets. However, the topology of $Flt3_{D3}$ is unusual such that the polypeptide chain extending from $Flt3_{D2}$ forms the N-terminal *A* strand in $Flt3_{D3}$ (residues 246 - 249) by complementing strand *B* in a parallel fashion,

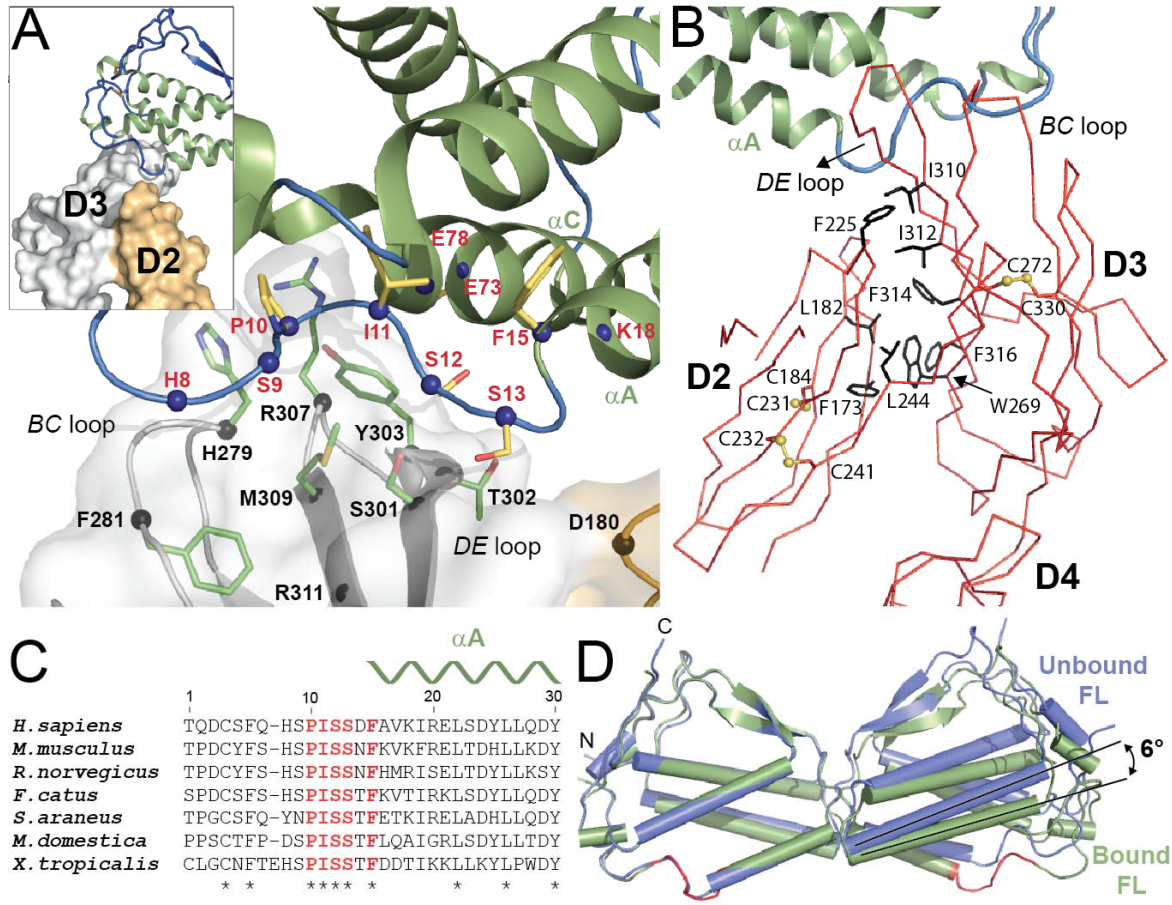



Figure 5-3. The Flt3-FL binding interface. (A) Close-up view of the Flt3-FL binding interface. FL is colored in green, $Flt3_{D3}$ in grey and $Flt3_{D2}$ in orange. Residues that constitute the cytokine-receptor interface are shown as sticks protruding from spheres centered at their Cα positions. FL residues are colored in yellow and Flt3 residues are colored in green. (B) The unusual $Flt3_{D2}$-$Flt3_{D3}$ interface. $Flt3_{D2-D4}$ (Cα trace in red) is shown together with FL in ribbon representation (green). Residues at the hydrophobic interface are shown as black sticks. Disulfide bonds in $Flt3_{D2-D3}$ are shown as ball and sticks (yellow). (C) Structure-based alignment of diverse FL sequences revealing strict conservation of the PISSXF-segment (residues 10 - 15) within the N-terminal loop (coloured in red). A complete alignment can be found in fig. B-3. (D) Structural comparison of bound versus the unbound FL.

while the *AA'* loop of $Flt3_{D3}$ (residues 250 - 258) adopts an extended conformation (**Fig. 5-3b**, appendix B: **Fig. B-1b**). $Flt3_{D2}$, which in all other RTKIII/V complexes contributes roughly half of the ligand-binding epitope, packs against the hydrophobic patch projected by the *ABED*-face of $Flt3_{D3}$ centered around Trp269 burying ~1000 $Å^2$ (**Fig. 5-3b**). $Flt3_{D2}$ is homologous to $KIT_{D5}$ and is a member of the C2 subset of the Ig-family (*ABED/CFG* topology), but contains an additional solvent-exposed disulfide (Cys232-Cys241) bridging

strands *F* and *G*. Although the *AB* and *EF* loops of $Flt3_{D2}$ point in the direction of the ligand they remain too far to engage in any interactions. The only point on $Flt3_{D2}$ approaching FL within a distance that could mediate any form of interaction is centered on Asp180 on the AB loop of $Flt3_{D2}$. However, it is not clear whether this interaction actually occurs because we were not able to model the side-chain of Asp180 (**Fig. 5-3a**).

The FL binding epitope on $Flt3_{D3}$ engages in extensive interactions with the N-terminal loop (residues 8-13) of FL leading to αA and Lys18 on αA, and is mainly contributed by the *BC* loop of $Flt3_{D3}$ (residues 279-280) and strand *D* (residues 301-303). Additional interactions are mediated by the *DE* loop of $Flt3_{D3}$ (residue 307) which contacts a small patch on the C-terminal region of helix αC of FL defined by residues 73 and 78. Therefore, the Flt3 ligand-receptor interaction results in a single contact site covering merely ~900 $Å^2$ of buried surface area. Structure-based alignments using diverse FL sequences revealed a remarkably strict conservation of the `PISSxF` cassette as well as Phe81 and Leu115, which help to lock the N-terminal loop in its observed conformation (**Fig. 5-3c**, appendix B: **Fig. B-3**).

**Plasticity of FL upon binding to Flt3**

Comparison of FL in its unbound (Savvides, Boone et al. 2000) and now in its receptor-bound form reveals that the cytokine ligand does not undergo any significant local structural changes at its receptor binding epitope (**Fig. 5-3d**). This is contrary to what has been observed in Stem Cell Factor (SCF) in complex with KIT, whereby the cytokine ligand undergoes a cascade of structural rearrangements (Liu, Chen et al. 2007; Yuzawa, Opatowsky et al. 2007). However, the two FL subunits display a hinge-like rigid-body rearrangement about the dimer interface, which increases the tilt angle between the two protomers by 5-6° (**Fig. 5-3d**). A similar motion was previously observed in the SCF-$KIT_{D1-D5}$ complex although SCF, unlike FL, already appears to have significant variability in the receptor-free form as shown by the range of its intersubunit tilt angles (2° to 6°) (Yuzawa, Opatowsky et al. 2007).

**The $Flt3_{D3}$-$Flt3_{D4}$ domain elbow and the absence of homotypic receptor interactions**

A second striking feature of the $Flt3_{D1-D4}$:FL complex is the absence of any obvious specific homotypic receptor interactions. Based on the current paradigm of RTKIII activation such interactions would be mediated by extracellular domain D4. While $Flt3_{D4}$ does point to its tandem $Flt3_{D4'}$ in the complex, the two receptor domains stay clearly away from each other and deviate from the two-fold symmetry of the complex. The inability of $Flt3_{D4}$ to engage in homotypic interactions may also explain the observed disorder for this part of the structure, as we could only reliably model and refine a complete $Flt3_{D4}$-$Flt3_{D4'}$ tandem in only one of the two complexes in the asymmetric unit of the crystal, whereas for the second we could only place one of the two domains.

Closer inspection of the $Flt3_{D4}$ topology and sequence reveals that $Flt3_{D4}$ does not possess the conserved structure-sequence fingerprints seen in all other RTKIII/V homologues for this domain. For instance, $Flt3_{D4}$ has two additional disulfide bridges, a solvent exposed cross-strand disulfide bridge (Cys368-Cys407) connecting strands *B* and *E*, and a second (Cys381-Cys392) connecting its unusual *C'E* loop with strand *C*. Most importantly, $Flt3_{D4}$ displays an *EF*-loop that drastically differs both in structure and sequence from all homologues (**Fig. 5-4-c**). The *EF*-loop constitutes the otherwise conserved 'tyrosine corner' motif in I-set Ig-domains (Harpaz and Chothia 1994), and has been shown to mediate homotypic interactions in the case of $KIT_{D4}$ (Yuzawa, Opatowsky et al. 2007) and $VEGFR_{D7}$ (**Fig. 5-4b**).

Structural comparisons of the two independent $Flt3_{D1-D4}$:FL complexes in the crystal asymmetric unit revealed slight orientational plasticity of $Flt3_{D4}$ about the $Flt3_{D3}$-$Flt3_{D4}$ linker region. This stretch of residues and strand *A* of $Flt3_{D4}$ are well conserved in Flt3 and KIT and other RTKIII suggesting a common functional role (**Fig. 5-4d**). A comparison of KIT in the cytokine-bound and -unbound forms, showed that the $KIT_{D3}$-$KIT_{D4}$ linker region acts as a hinge to reorient $KIT_{D4}$ for homotypic interactions upon ligand binding. Despite the absence of such homotypic contacts in Flt3, the domain elbow defined by $Flt3_{D3}$ and $Flt3_{D4}$ is similar to KIT, suggesting preservation of this interdomain relationship in both forms of the receptor. Thus, the orientation of $Flt3_{D4}$ appears to be restricted by a core of well-defined hydrophobic interactions mediated by Phe261 (*A'* strand of $Flt3_{D3}$), Val345 ($Flt3_{D3}$-$Flt3_{D4}$ linker), Phe349 (*A* strand of $Flt3_{D4}$) and Tyr376 (*BC* loop of $Flt3_{D4}$), as well as additional interactions between the *AA'* loop of $Flt3_{D3}$ and the *C'E* loop of $Flt3_{D4}$ (**Fig. 5-4a**).

**Architecture of the complete extracellular assembly of the Flt3 signaling complex**

Crystals of $Flt3_{D1-D5}$:FL grew reproducibly from a number of crystallization conditions but proved to be of low diffraction quality despite repeated attempts to improve crystal quality by various methods including glycan shaving. Nonetheless, a robust dataset to 7.8 Å resolution proved sufficient to elucidate the architecture of the complete extracellular Flt3 complex by molecular replacement based on the $Flt3_{D2-D3}$: FL subcomplex as refined in the $Flt3_{D1-D4}$:FL crystal structure (**Table 5-1**). We could subsequently place into electron density and optimize by rigid-body refinement protocols $Flt3_{D1}$ and $Flt3_{D4}$, extracted from the crystal structure of the $Flt3_{D1-D4}$:FL complex, as well as a conservative homology model of $Flt3_{D5}$ derived from the structure of human $KIT_{D5}$ (Yuzawa, Opatowsky et al. 2007). While the low resolution of the $Flt3_{D1-D5}$:FL structure does not allow discussion of structural details, it does provide a reliable and valuable depiction of the organization features of the complete extracellular Flt3 complex.

In the full-length ectodomain complex, the core structure observed in $Flt3_{D1-D4}$:FL is mounted onto two membrane-proximal $Flt3_{D5}$ facing each other to form an assembly resembling a hollow tennis racket (140x75x110 Å) (**Fig. 5-5**, appendix B: **Fig. B-2b**). Remarkably, the asymmetry exhibited by the tandem $Flt3_{D4}$ modules in $Flt3_{D1-D4}$:FL is not present in the complete extracellular complex. Instead, the two $Flt3_{D4}$ segments face each other nearly symmetrically according to the 2-fold symmetry of the $Flt3_{D2-D3}$:FL core structure and approach to about 20 Å from each other. While this inter-receptor separation is maintained at the ensuing $Flt3_{D5}$ modules, the apparent two-fold symmetry breaks down. Furthermore, the asymmetric projection of the N-terminal $Flt3_{D1}$ domains perpendicularly out of the plane of the racket head occurs in a manner analogous to what we observed in the $Flt3_{D1-D4}$:FL complex. Complementary studies of the full-length ectodomain complex by negative-stain negative-stain EM, SAXS in solution corroborated the overall structural features revealed by the crystal structure (Appendix B: **Fig. B-4 a, b**).

**Discussion**

Cytokine-mediated activation of hematopoietic cell-surface receptors is central to developing and sustaining hematopoiesis and the immune system. The RTKIII receptor Flt3 and its cognate cytokine ligand FL are arguably the most exciting new addition to the repertoire of

hematopoietic factors, due to their activity on hematopoietic progenitors and pronounced impact on the development and homeostasis of antigen-presenting DC. As the importance of Flt3 signaling in early and late hematopoiesis continues to mount, we sought to elucidate the structural basis of the extracellular Flt3 receptor-ligand complex. The structural studies we report here complemented by a thermodynamic dissection of complex formation, show that the Flt3-FL interaction is characterized by high-affinity bivalent binding of FL to Flt3 that does not invoke homotypic receptor interactions. The assembly showcases several unexpected features, which now establish Flt3 as a structural outlier within the RTKIII/V family (**Fig. 5-6**). The Flt3:FL interaction epitope is surprisingly a fraction of typical helical cytokine-receptor interaction, and is dominated by contacts between a preformed N-terminal segment of FL and $Flt3_{D3}$. Consistent with the polar receptor-cytokine interface, the thermodynamic blueprint of the interaction calls for an enthalpically-driven binding event. The interaction carries a concomitant significant entropic cost, which we now can rationalize in terms of the absence of a hydrophobic effect and the intrinsic entropy loss associated with bringing interaction partners together.

The Flt3-FL interaction interface covers a compact ~900 $Å^2$, which is at least 2 times less extensive than the buried surface area at the receptor-cytokine epitopes of all other RTKIII/V complexes, whereby the activating cytokine is harbored by a broad grapple defined by extracellular domains 2 and 3 (Chen, Liu et al. 2008); Leppanen, Prota et al. 2010) (**Fig. 5-6**). Comparison with diverse helical cytokine-receptor interactions (Stroud and Wells 2004; Wang, Lupardus et al. 2009) shows that Flt3 is the only receptor for a helical cytokine that uses a single interaction site to bind its cognate ligand. Heterodimeric protein-protein interactions bury on average ~1300 $Å^2$ of surface area, which is strongly correlated with high-affinity binding (Janin, Bahadur et al. 2008). Flt3 and FL are clearly able to establish a tight interaction via a much more compact binding interface. A plausible explanation could be drawn from the rigidity of the receptor epitope on FL as a preformed binding platform, reminiscent of a classical 'lock-and-key' binding mode observed in affinity-matured antibody-antigen interactions (Sundberg and Mariuzza 2002). Furthermore, a series of single amino acid substitutions (H8R, S9G, P10S, S13P/F, F15L) within the segment contributing almost the entire receptor binding epitope on FL, abolish receptor activation completely (Graddis, Brasel et al. 1998). This illustrates not only the possible individual contribution of each residue in this segment to binding but also the likely conformational stringency of the region. Indeed, a comparison of diverse FL sequences

showed that not only the receptor-binding epitope is exquisitely conserved but also residues that help to lock the N-terminal loop both in its observed receptor-bound and receptor-free forms (**Fig. 5-3c**, appendix B: **Fig. B-3**).

*Table 5-1. X-ray data collection and refinement statistics.*

*The values in parentheses refer to the highest resolution shell.*

| | $Flt3_{D1-D4}$:FL | $Flt3_{D1-D5}$:FL |
|---|---|---|
| **Data collection** | | |
| Source, Wavelength (Å) | ESRF ID23-1, 0.9762 | ESRF ID23-1, 1.0762 |
| Detector | ADSC-Q315R | ADSC-Q315R |
| Resolution (Å) | 40.00 - 4.30 (4.45 - 4.30) | 35.00 - 7.80 (8.00 - 7.80) |
| Space group | $P2_1$ | $P2_1$ |
| Unit cell parameters | a=103.89<br>b=146.26<br>c=105.95<br>α=γ=90°, β=109.7° | a=124.75<br>b=153.55<br>c=133.87<br>α=γ=90°, β=94.6° |
| Wilson B ($Å^2$) | 135 | 401 |
| Unique reflections | 20184 (1942) | 5656 (382) |
| Redundancy | 3.8 (3.8) | 3.4 (3.2) |
| Completeness (%) | 98.8 (98.9) | 96.6 (92.9) |
| $R_{meas}$ (%)[a] | 10.8 (75.9) | 12.8 (80.7) |
| Average $I/\sigma(I)$ | 12.04 (2.08) | 9.1 (1.9) |
| **Refinement** | | |
| Resolution (Å) | 40.00 - 4.30 (4.53 - 4.30) | 35.00 - 7.80 (8.72 - 7.80) |
| Reflections working set / test set | 19172 / 1010 (2763 / 141) | 5090 / 565 (1437 / 159) |
| $R_{work}$, $R_{free}$ | 0.260 / 0.281 (0.268 / 0.268) | 0.337 / 0.346 (0.334 / 0.313) |
| R.m.s. deviations | | |
| Bonds (Å) | 0.010 | n.a. |
| Angles (°) | 1.12 | n.a. |
| Average ADP ($Å^2$) | 169 | 364 |
| Ramachandran analysis (%) | | |
| Favorable | 89.8 | n.a. |
| Outliers | 2.0 | n.a. |
| Protein Data Bank access code | 3QS7 | 3QS9 |

[a] $R_{meas} = \sum_h \sqrt{n_h/(n_h-1)} \sum_h \sum_i |I(h,i)-\langle I(h)\rangle| / \sum_h \sum_i I(h,i)$, where $n_h$ is the multiplicity, $I(h,i)$ is the intensity of the $i^{th}$ measurement of reflection $h$, and $\langle I(h)\rangle$ is the average value over multiple measurements.

Comparison of the Flt3-FL interaction and representative receptor-cytokine epitopes for all other RTKIII/V, shows that engagement of Ig-like domain 3 (*BC* loop; *DE* loop and flanking residues) is the only common epitope feature of ligand binding (**Fig. 5-6**). It thus appears that binding of D3 of RTKIII/V to the activating cytokine satisfies a geometric requirement that allows receptor molecules to approach to a critical distance of ~60 Å from one another. This notion reinforces a fascinating aspect of RTKIII/V activation in that the cognate protein ligands are all dimeric with similar dimensions despite their grouping into two fundamentally different folds (4-helix bundles versus all-β cystine-knot scaffolds) (Savvides, Boone et al. 2000; Yuzawa, Opatowsky et al. 2007; Chen, Liu et al. 2008; Shim, Liu et al. 2010). Recently, interleukin-34 (IL-34) was identified as a second ligand to CSF-1R (Lin, Lee et al. 2008), thus adding a perplexing dimension to RTKIII signaling as IL-34 bears no sequence similarity to the currently known cytokine ligands for RTKIII/V or other proteins.

The uniqueness of the Flt3 extracellular complex is further highlighted by the absence of homotypic receptor interactions. Such interactions have recently emerged as an important aspect of RTKIII/V activation and are mediated by conserved structure-sequence fingerprints in the membrane-proximal domains (Ruch, Skiniotis et al. 2007; Yuzawa, Opatowsky et al. 2007; Chen, Liu et al. 2008; Yang, Yuzawa et al. 2008). In both of our Flt3-FL complexes, the membrane-proximal modules remain separated by ~20 Å at the D4-D5 junction, consistent with our comparative ITC data showing no significant contribution by the D4-D5 module. Flt3 is the only RTKIII/V family member that lacks the conserved set of residues involved in homotypic interactions in the homologous receptors, which now offers a strong rationale for the absence of such interactions in the extracellular Flt3 complex. Additional support comes from the existence of a fully active murine Flt3 isoform that lacks $Flt3_{D5}$ demonstrating the dispensability of the membrane-proximal domain for receptor activation (Lavagna, Marchetto et al. 1995), contrary to the apparent importance of $KIT_{D5}$ in signaling (Broudy, Lin et al. 2001). However, it is possible that homotypic interactions could be enhanced within the two-dimensional spatial confinement of the cell membrane, and as a result of additional interactions between TM and/or cytosolic segments of Flt3. In fact, the difference in the affinity for the complete ectodomain complex versus previously reported values for native Flt3 based on cell-assays (Turner, Lin et al. 1996; Graddis, Brasel et al. 1998) may reflect a combination of such factors. To this end, recent studies have highlighted the importance of homotypic interactions between TM and JM regions in RTKIII/V activation and pathology profiles (Finger, Escher et al. 2009; Oates, King et al. 2010).

Interestingly, oncogenic variants of Flt3 carrying ITD in the JM segment are constitutively active as homodimers or as heterodimers with wild-type receptor, indicating that enhanced intracellular receptor interactions can drive activation (Kiyoi, Ohno et al. 2002). Recently, a number of mutations in the extracellular segment of Flt3 were identified in AML patients (Schnittger, Kohl et al. 2006; Frohling, Scholl et al. 2007), which we now can map onto the Flt3 ectodomain (appendix B: **Fig. B-5**). However, the clinical relevance of these mutations will have to await further study.

In the absence of structural information for unbound Flt3 we are left to wonder about any possible domain rearrangements in Flt3 upon FL binding. Structural studies of KIT in the bound and unbound forms showed that KIT undergoes a large conformational switch at the $KIT_{D3\text{-}D4}$ junction leading to homotypic receptor interactions (Yuzawa, Opatowsky et al. 2007). However, the extensive and unique hydrophobic interface observed between $Flt3_{D2}$ and $Flt3_{D3}$ (1000 $Å^2$), the hydrophobic interface between $Flt3_{D3}$ and $Flt3_{D4,}$ and additional contacts between loops thereof, provide evidence that the $Flt3_{D2\text{-}D4}$ ectodomain segment would be too rigid to undergo significant domain rearrangements. This is further supported by our structural studies of the complete extracellular complex, which showed that the relative orientation of $Flt3_{D3}$ and $Flt3_{D4}$ only differs slightly from that observed in the $Flt3_{D1\text{-}D4}$:FL complex. Nonetheless, Flt3 does exhibit significant domain plasticity at the two extremities of the extracellular assembly. This is most pronounced for $Flt3_{D1}$, which emanates from the core of the assembly without making contacts with other complex components. We note that the stacking of $Flt3_{D2}$ against $Flt3_{D3}$ provides a fixed angle for projecting $Flt3_{D1}$ from a point approximately halfway down the height of the complex. $Flt3_{D1}$ is the largest and most atypical domain in Flt3 and the entire RTKIII/V family, but its role in Flt3 signaling is currently unknown. We are thus tempted to propose that $Flt3_{D1}$ could mediate intermolecular contacts at the cell surface and/or stabilize the unbound receptor.

The availability of structures for the complete extracellular Flt3 receptor-ligand complex including a delineation of the receptor-cytokine interface, will likely have a significant impact on renewed efforts to antagonize Flt3 activity in a clinical setting. This is because wild type and mutated forms of Flt3 as well as autocrine secretion of FL have been implicated in the development of myeloid leukemias (Stirewalt and Radich 2003; Zheng, Levis et al. 2004; Sanz, Burnett et al. 2009; Kindler, Lipka et al. 2010). Current strategies focus on inhibition of the intracellular kinase domains of Flt3, but are faced with drug

specificity issues and the emergence of primary and secondary resistance to treatment (Kindler, Lipka et al. 2010). More recently, an alternative therapeutic approach based on monoclonal antibodies directed against the extracellular domain of Flt3 (Piloto, Nguyen et al. 2006) indicated a possible momentum shift towards combined strategies in clinical targeting of Flt3. A daunting challenge in inhibiting protein-protein interactions is the extent of the interaction epitope which often covers >1500 $Å^2$ and the lack of prior knowledge of functional hotspots (Wells and McClendon 2007). In this regard, the compactness of the Flt3 ligand-receptor interface provides favorable perspectives for the druggability of the extracellular Flt3 binding epitope.

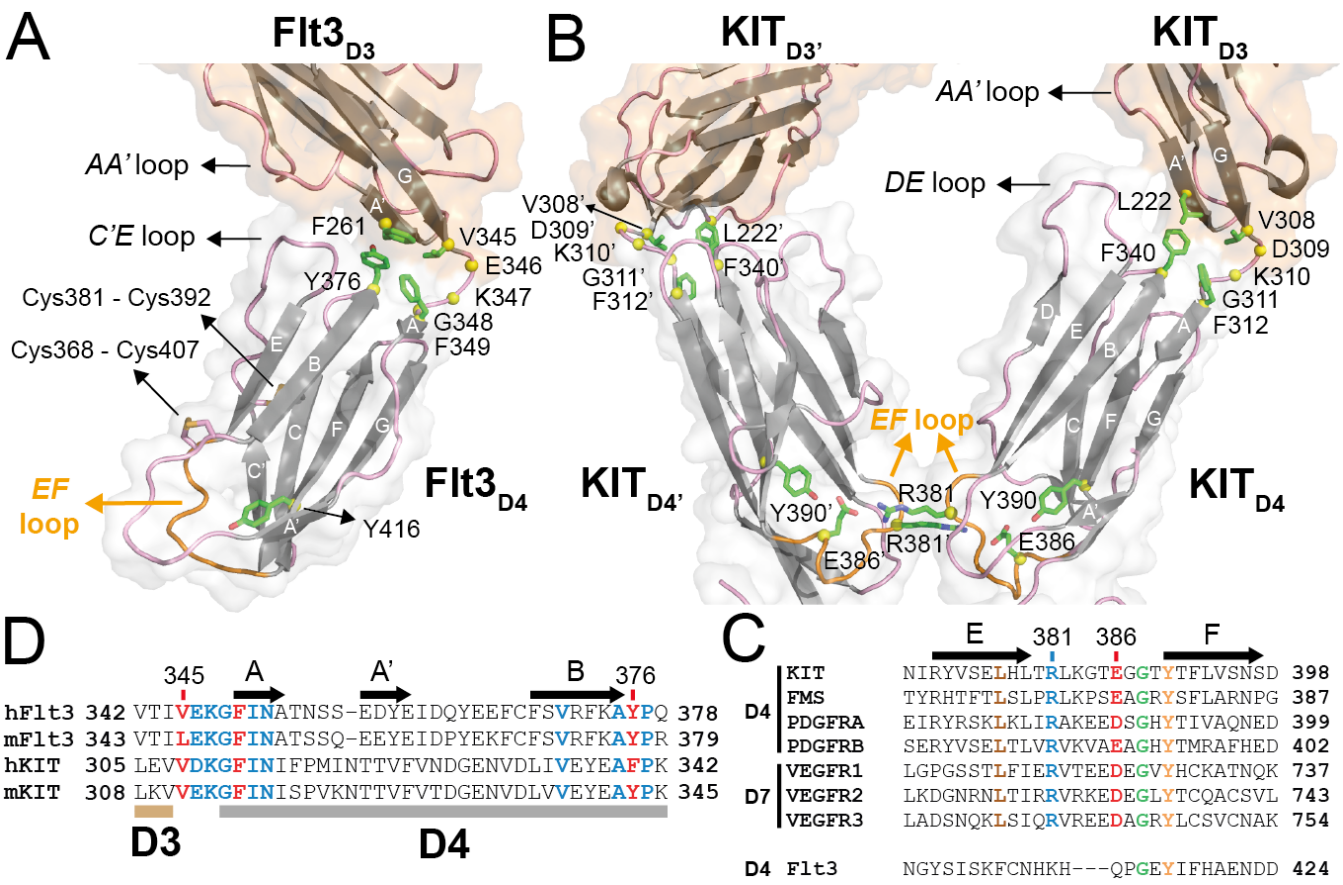



Figure 5-4. The $Flt3_{D3}$-$Flt3_{D4}$ elbow and the absence of homotypic receptor contacts in the Flt3:FL complex. (A) The $Flt3_{D3}$-$Flt3_{D4}$ elbow. $Flt3_{D3}$ (partially shown) and $Flt3_{D4}$ are shown in ribbon representation. For clarity purposes only the the locations of the atypical disulfide bridges in $Flt3_{D4}$ (Cys368-Cys407 and Cys381-Cys391) are indicated. Residues mediating hydrophobic interactions between $Flt3_{D3}$ and $Flt3_{D4}$ are shown as green sticks. Residues in the $Flt3_{D3}$-$Flt3_{D4}$ linker (346-348) are shown as yellow spheres centered at their Cα positions. The side-chains of residues mediating contacts between the *AA'* loop of $Flt3_{D3}$ and the *C'E* loop of $Flt3_{D4}$ could not be modelled due to the low resolution of our analysis. The *EF*-loop of $Flt3_{D4}$ constituting the 'tyrosine corner' around Y416 (green sticks) is shown in orange. (B) Sequence conservation of residues involved at the D3-D4 interface in KIT and Flt3 based on comparisons between human and murine Flt3 and KIT sequences. (C) $KIT_{D3}$-$KIT_{D4}$ orientation in the KIT:SCF complex. Homotypic receptor contacts between tandem ectodomain 4 modules in the KIT-SCF complex are mediated by salt-bridges via residues R381 and E386 residing on the *EF* loops (orange) (PDB entry 2E9W). Residues at the hydrophobic $KIT_{D3}$-$KIT_{D4}$ interface are shown as green sticks. Residues in the $KIT_{D3}$-$KIT_{D4}$ linker region (D309-G311) are shown as yellow spheres. (D) $Flt3_{D4}$ displays an atypical *EF*-loop within the RTKIII/V family. The pair of residues mediating the homotypic contacts in $KIT_{D4}$ and VEGFR-$2_{D7}$ is well conserved in the corresponding domains of all RTKIII/V members but not in $Flt3_{D4}$.

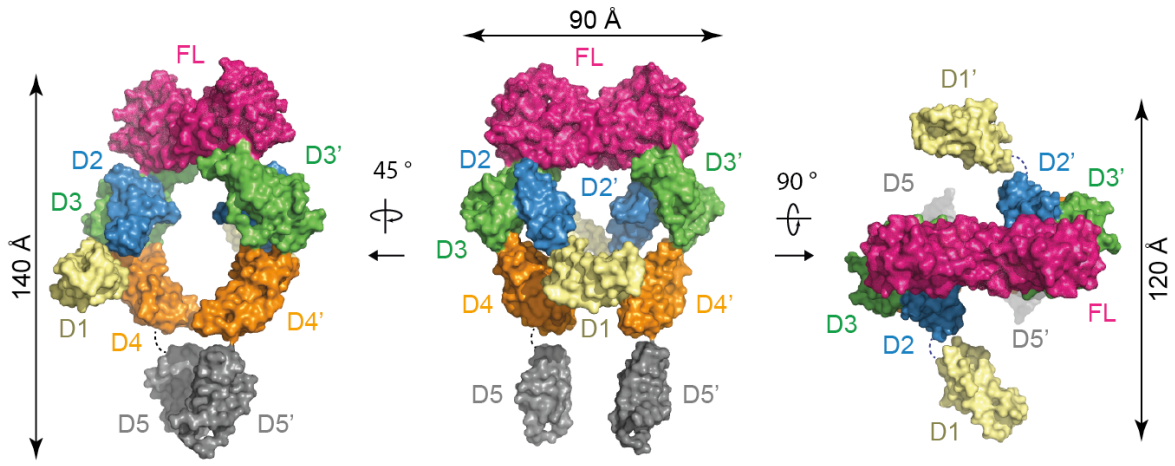



Figure 5-5. Assembly of the complete Flt3 ectodomain complex. Surface representations of the full-length Flt3 ectodomain complex. The central view shows the complex with the two-fold symmetry axis of FL oriented vertically in the plane of the paper.

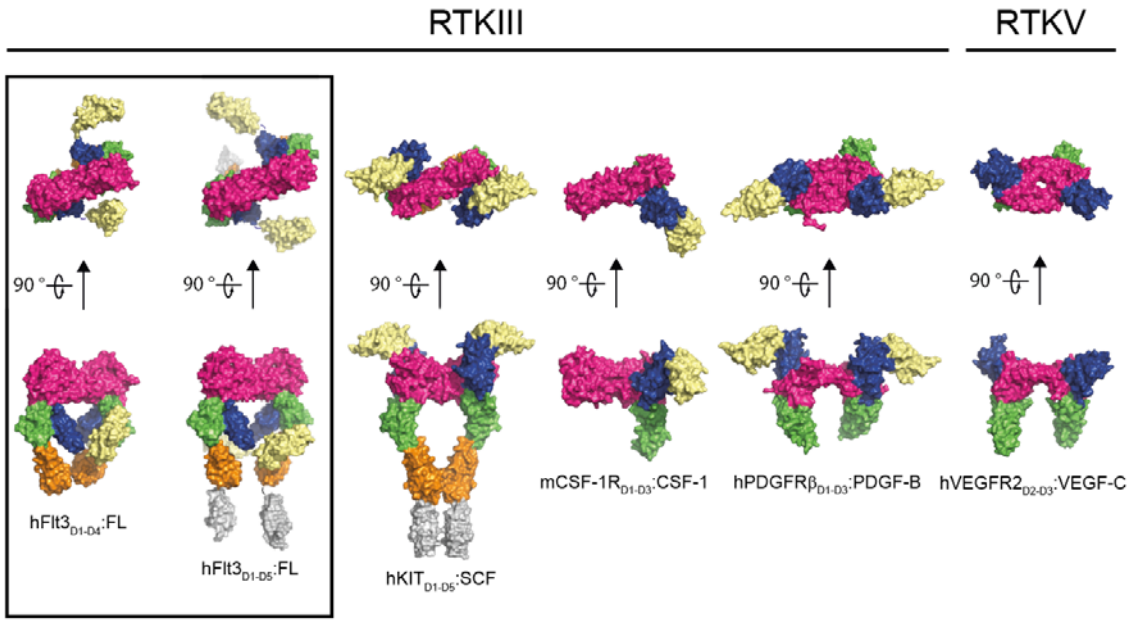



Figure 5-6. Comparison of representative RTKIII/V extracellular complexes. The structures shown represent the architecture of receptor-cytokine complexes for the different members of the RTKIII/V family: From left to right: human Flt3:FL (this study), human KIT:SCF (PDB 2E9W), mouse CSF-1R:CSF-1 (3EJJ), human PDGFR:PDGF (PDB 3MJG) and human VEGFR2:VEGF (2X1X). The dimeric ligands are colored in magenta. Receptor ectodomains are coloured as follows: D1 in pale yellow, D2 in blue, D3 in green, D4 in orange and D5 in grey.

## 5.2 Oligomerization propensity and flexibility of yeast frataxin studied by X-ray crystallography and SAXS

Christopher AG Söderberg, Alexander V. Shkumatov, Sreekanth Rajan, Oleksandr Gakh, Dmitri I. Svergun, Grazia Isaya and Salam Al-Karadaghi

*Manuscript in preparation.*

Frataxin is a mitochondrial protein with a central role in the iron homeostasis, including iron storage and detoxification, iron delivery to iron-sulphur cluster synthesis, heme synthesis and aconitase repair. Deficiency in frataxin function has been attributed to the progressive neurodegenerative disease Friedreich's ataxia, which is characterized by early onset between the ages of 5 and 15. Numerous biochemical studies have demonstrated that the function of frataxin is closely associated with its ability to form oligomeric species, however, the mechanisms of oligomerization, the types of oligomers present in solution and their possible function in iron homeostasis are poorly understood. In this work we study factors influencing yeast frataxin Yfh1 oligomerization. Using SAXS we demonstrate that the protein may exist in different oligomeric states in solution. These states may be induced by buffer content, amino acid composition of the N-terminus and presence of divalent metals in the solution. The presence of glycerol facilitates the dimerization of the protein, resulting in dynamic monomer-dimer equilibrium. SAXS data from $Co^{2+}$-induced oligomerization show that the oligomers are present as multiples of 3, suggesting that the trimer is the main building block of higher order oligomers. The X-ray structure of the Y73A trimer was determined in complex with $Co^{2+}$, revealing the position of the metal inside the channel at the 3-fold axis and suggesting that the metal contributes to the stabilization of the trimer. Together, our data suggest that there are several ways of controlling the oligomeric state of frataxin, thereby allowing us to speculate about multiple factors that can be important for frataxin function *in vivo*.

**State of the problem**

Iron is one of the most abundant elements on earth and it is essential for most living organisms, despite the toxic chemistry associated with it (Kaplan 2002). Thus, in an $O_2$ containing environment and at physiological pH, $Fe^{2+}$ ($6H_20$) is readily oxidised to the acidic $Fe^{3+}$ ($6H_20$), which rapidly produces insoluble $Fe(OH)_3$. In addition, the one-electron oxidation of iron by oxygen produces superoxide that in turn can react to form hydrogen peroxide. Hydrogen peroxide reaction with $Fe^{2+}$ produces the highly reactive and toxic hydroxyl radical through the Fenton reaction. Evidently, a tight control of iron chemistry in organisms is essential. Two proteins, ferritin and frataxin, are known to be central to iron storage and detoxification in cells. The importance of frataxin in iron homeostasis was first demonstrated by the observation that development of the hereditary autosomal disease Friedreich's ataxia (FRDA) is linked to the presence of extensive GAA repeats in the first intron of the frataxin-encoding gene, as well as to mutations in the gene sequence coding for the protein (Campuzano, Montermini et al. 1996). FRDA affects approximately one in 40.000 individuals, with early onset, often before the age of 15. Frataxin deficiency results in aberrations in cellular iron homeostasis and high activity of reactive oxygen species (ROS), leading to a progressive neurodegenerative disease characterized by neurological impairment, cardiomyopathy, and diabetes mellitus (Geoffroy, Barbeau et al. 1976; Harding and Hewer 1983; Finocchiaro, Baio et al. 1988).

Ferritin, as well as frataxin and its homologues have been implicated in the control of storage and detoxification of iron by catalyzing the oxidation of ferrous iron to the ferric form and its storage as a ferrihydrite biomineral within oligomeric species (Chasteen and Harrison 1999; Gakh, Adamec et al. 2002; Nichol, Gakh et al. 2003; Gakh, Park et al. 2006). In the case of ferritin these oligomers are spontaneously assembled in cells, while for frataxin and its homologues there appears to be a variation in the propensity of the protein from different sources to form oligomers and in the type of oligomers formed (Park, Gakh et al. 2003; O'Neill, Gakh et al. 2005; O'Neill, Gakh et al. 2005). Human and *Saccharomyces cerevisiae* (Yfh1) frataxin have been studied most extensively. They are expressed in the cytoplasm as 210 and 174 precursor polypeptides with 55 and 51 amino acid mitochondrial-targeting sequences, respectively. Processing by mitochondrial peptidases results in the 52-174 variant of the yeast frataxin, while the human protein appears to exist in several variants with a different length of the N-terminus (Koutnikova, Campuzano et al. 1998; Branda, Cavadini et al. 1999).

In addition to iron storage and detoxification, the role of frataxin in iron homeostasis also includes iron delivery to iron-sulphur cluster synthesis, heme biosynthesis and aconitase repair (Nuth, Yoon et al. 2002; Lesuisse, Santos et al. 2003; Bulteau, O'Neill et al. 2004). The function of frataxin has been strongly linked to its ability to form oligomeric species. Thus, in the presence of ferrous iron and oxygen yeast frataxin has been shown to undergo stepwise assembly from monomers to larger oligomers, which could contain up to 24, and even 48 subunits (Park, Gakh et al. 2003; O'Neill, Gakh et al. 2005; O'Neill, Gakh et al. 2005). Higher order oligomers could store up to ~50–75 iron atoms per subunit in 1–2 nm cores (Nichol, Gakh et al. 2003; Schagerlof, Elmlund et al. 2008). The oligomeric species of frataxin may be easily dissolved into monomers, e.g. by the addition of reducing agents (Nichol, Gakh et al. 2003; Park, Gakh et al. 2003) suggesting that iron plays an active role in oligomer stabilization. In humans, at least 4 variants of frataxin have been isolated. The variants lacking the mitochondrial-targeting sequence (56-210 and 42-210) may assemble into larger structures during expression in *E. coli* (Cavadini, Adamec et al. 2000; Cavadini, O'Neill et al. 2002). The large oligomers could be disassembled irreversibly into stable monomers by the addition of SDS (O'Neill, Gakh et al. 2005). However, in contrast to the yeast Yfh1 protein, purified human frataxin monomers did not form oligomers *in vitro*, even in the presence of iron (Cavadini, O'Neill et al. 2002). This could imply that the assembly of the human protein requires some additional assistance. Indeed, it has been shown that the Hsp70 type protein, SSQ1, is needed for functional Yfh1 in *S. cerevisiae* (Knight, Sepuri et al. 1998; Voisine, Schilke et al. 2000). On the other hand, the 81-210 and 78-210 variants could not form higher order oligomeric states (O'Neill, Gakh et al. 2005), implying that the residues 56-77 in the N-terminal tail of human frataxin are essential for the oligomerization and oligomer stabilization.

X-ray crystallographic studies of a variant of yeast frataxin in which amino acid Y73 was replaced by an alanine (Y73A) showed that this protein crystallized as a trimer, apparently stabilized by extensive interactions between the monomers around the three-fold axis as well as by the N-terminus, which bridged the monomers by forming additional interactions between them (Karlberg, Schagerlof et al. 2006). SEC also demonstrated that in contrast to the wild type Yfh1, which requires iron for assembly, the Y73A variant could assemble into larger oligomeric species, which contained up to 24 monomers (Karlberg, Schagerlof et al. 2006). In addition, single-particle reconstruction studies of both iron loaded and iron-free 24-meric particles clearly demonstrated that they were built up by trimeric units (Karlberg, Schagerlof et al. 2006), suggesting that the trimer could be the main building

block of larger frataxin oligomers. Interestingly, the arrangement of the functional features found in this structure, like the position of the ferroxidation and iron mineralization sites and charge distribution inside and around the channel at the three-fold axis, showed striking similarities to the arrangement of the corresponding features in ferritin, despite the absence of any evolutionary relationships at the amino acid sequence level (Karlberg, Schagerlof et al. 2006) (Schagerlof, Elmlund et al. 2008).

Given these data, the propensity of frataxin to form oligomeric species and in particular the relationships between the length and amino acid sequence of the N-terminus of the protein and the amino acid determinants of oligomer stabilization are far from understood. In the current work we attempted to gain an insight into the factors that control frataxin oligomerization and assembly of functional species. Using a combination of SAXS, X-ray crystallography and dynamic light scattering, we have studied the behaviour of monomeric yeast frataxin in solution, the Y73A variant, which is prone to oligomerization, $Co^{2+}$ binding to frataxin and $Co^{2+}$-induced oligomerization.

**Structure of yeast frataxin in solution**

The results of SAXS measurements on the monomeric yeast frataxin are summarized in **Fig. 5-7** and **Table 5-2**. Experimental scattering curves for samples with glycerol did not show concentration effects. Thus, data with the highest concentration were used for analysis and structural modeling (**Fig. 5-7**, iv-vi). The radius of gyration ($R_g$) estimated using Guinier approximation was 1.8±0.1 and 2.25±0.1 nm for $Yfh1_{wt}$ and $Yfh1_{wt}$ + glycerol, respectively. Porod analysis of the scattering data estimated a particle volume of ~24 and 44 $nm^3$ for frataxin samples without and with glycerol, respectively. Given that the ratio $V_p$/MW is expected to be approximately 1:1.6, the volume in the absence of glycerol is consistent with the molecular weight (13.7 kDa) of the monomeric frataxin. In the presence of glycerol the volume is significantly larger indicating the presence of higher oligomers. The theoretical scattering calculated for the X-ray structure of yeast frataxin yielded $R_g$ value of 2.07±0.1 nm (new X-ray model described here), while for the NMR structure (2GA5.PDB) the different conformers yielded $R_g$ values in the range from 1.54±0.1 to 1.72±0.1 nm. Although these values are close to the $R_g$ estimated from the SAXS data for the protein without glycerol, as shown on **Fig. 5-7**, neither the X-ray nor any of NMR models could fit well the SAXS experimental data (see **Fig. 5-7** (i), (ii), (iv), (v) and **Table 5-2**).

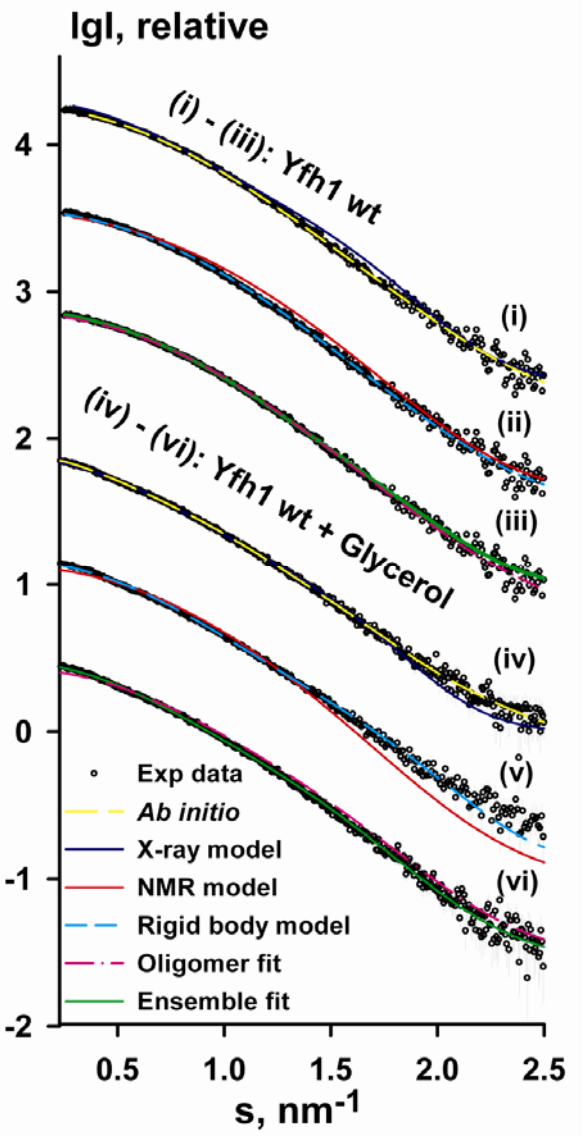



Figure 5-7. SAXS measurements of yeast frataxin. Experimental SAXS profiles (experimental data shown as black circles) for wild-type Yfh1 (i-iii) and wild type Yfh1 + glycerol (iv-vi) was appropriately displaced along the logarithmic axis for better visualization and overlaid with corresponding fits of (i, iv) X-ray structure (PDB ID: 2FQL) and *ab initio* model (ii, v) lowest energy NMR conformer (PDB ID: 2GA5) and rigid body model (iii, vi) *OLIGOMER* and EOM ensemble. Experimental SAXS data is shown to a maximal momentum transfer of $s=2.5$ nm$^{-1}$.

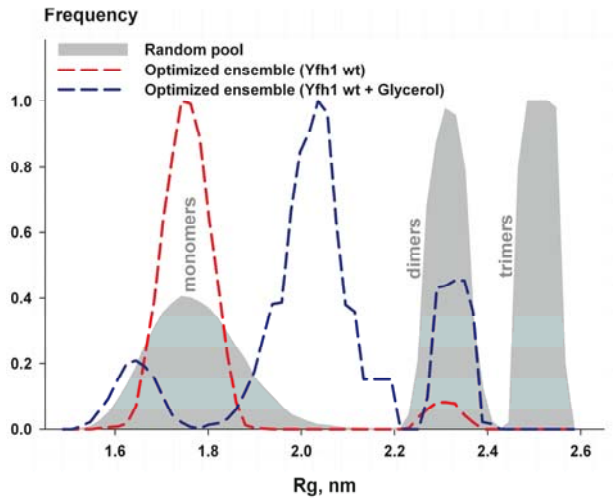



Figure 5-8. Flexibility of frataxin in solution. $R_g$ distribution for the random pool (6.000 models) containing equal fractions of monomers, dimers and trimers is shown by area filled with grey color. $R_g$ distribution of optimized ensemble corresponding to Yfh1 wt (red dashed line) and Yfh1 wt + glycerol (blue dashed line).

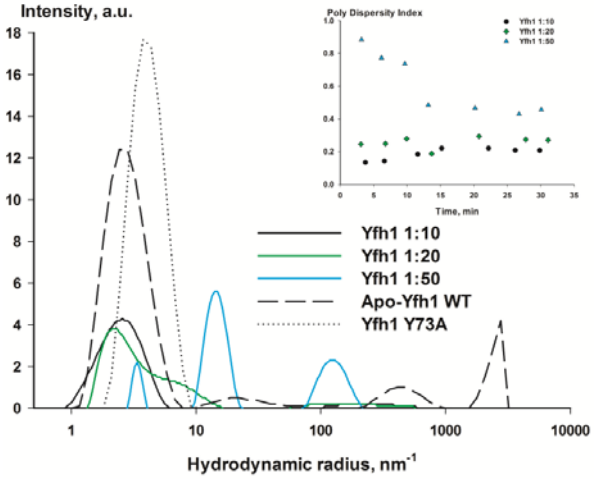



Figure 5-9. Oligomerisation of yeast frataxin homologue. **(a)** Sample size distribution (hydrodynamic radius) observed using DLS. Wild-type Yfh1 and the Y73A variant are shown in grey, Yfh1:$CoCl_2$ ratios 1:10, 1:20 and 1:50 are shown in blue, black and red, respectively. Insert shows the time-dependence of the polydispersity index after $CoCl_2$ addition.

Given that the monomeric frataxin contains a highly flexible N-terminus, as judged from the NMR model and the X-ray structure, a combined *ab initio* and rigid body modeling approach was employed to re-model the N-terminus, using the program BUNCH. During this process the inter-atomic distances and angles between 'dummy' atoms were constrained to mimic the Cα peptide chain. Ten BUNCH models were generated for the different data sets. The models showed reasonable agreement to the $Yfh1_{wt}$ scattering data (*s* range from 0.020 – 2.5 nm-1) with χ in the range 1.04-1.4, but rather poor agreement with the $Yfh1_{wt}$ + glycerol data, with χ in the range 1.4-2.4 (see **Fig. 5-7** (ii), (v) and **Table 5-2**). Since frataxin may represent a dynamic equilibrium of conformations in solution, we used OLIGOMER (Konarev, Volkov et al. 2003) to calculate the volume fractions contributing to the scattering profile for separate NMR conformers as well as for the X-ray structure. In this analysis, only two NMR conformers (# 8 and 16), contributing 32% and 68% volume fractions, respectively, were selected by OLIGOMER. These further improved discrepancy for the fitting to the $Yfh1_{wt}$ data, as compared to the lowest energy NMR conformer or the X-ray structure, but not for the data in the presence of glycerol (see **Fig. 5-7** (iii) and (vi) and **Table 5-2**).

To quantitatively characterize the ensemble properties of frataxin in solution, EOM (Bernado, Mylonas et al. 2007) was applied to the measured data sets. The random pool contained monomers, dimers and trimers (6.000 models) of frataxins with random configurations of the N-termini. The sub-ensembles selected by the EOM analysis allowed us to substantially improve the fit to the experimental data sets (**Fig. 5-7** (iii, vi) and **Table 5-2** ($\chi_f$)). The average radius of gyration of the EOM ensemble was found to be close to the $R_g$ calculated by Guiner approximation (**Fig. 5-8**). Interestingly, $R_g$ distributions corresponding to $Yfh1_{wt}$ and $Yfh1_{wt}$ + glycerol ensembles differ significantly (**Fig. 5-8**). Thus, in the case of $Yfh1_{wt}$, only a negligible fraction of dimers was selected. If the random pool only contained monomers, the discrepancy did not get worse, as compared to the mixed full pool. On the other hand, the $Yfh1_{wt}$ + glycerol data could only be fitted with more extended monomers and a larger fraction of dimers, as compared to $Yfh1_{wt}$, but cannot be fitted using a random pool containing solely monomers. If only three curves per ensemble were allowed to be selected, dimers were always present in the selected structures for $Yfh1_{wt}$ + glycerol and the discrepancy was in the range 1.06-1.07. Based on these results, it can be concluded that (i) the N-terminus part of frataxin is indeed highly flexible in solution, (ii) the addition of glycerol facilitates formation of dimers and (iii) SAXS experimental data can be described

with an ensemble model, consisting of mainly monomers in the case of $Yfh1_{wt}$, and a dynamic monomer-dimer equilibrium in the case of $Yfh1_{wt}$ + glycerol.

**Characterisation of the Yfh1 Y73A variant oligomers using SAXS**

Gel filtration studies have demonstrated that the Yfh1 Y73A variant is eluted as two major fractions, one of lower molecular weight and the second of higher molecular weight. As noted above, the low molecular weight fraction was crystallized and revealed the trimeric frataxin structure, while the high molecular weight fraction was studied by single-particle EM reconstruction (Karlberg, Schagerlof et al. 2006; Schagerlof, Elmlund et al. 2008). For a better understanding of the role of the Y73A replacement on the oligomerization propensity of the protein, we have studied the low molecular weight fraction using DLS (**Fig. 5-9**) and SAXS (**Fig. 5-10** and **Table 5-3**). Analysis of the data and subsequent modeling, as well as DLS measurements indeed revealed the presence of several oligomeric species in solution. To fit the experimental data, several types of oligomeric species of different molecular weight were constructed. These models were screened using OLIGOMER and those providing best fits were selected. As seen from the table, while the major part of the protein in solution is in a monomeric form (around 60%), there is also a substantial proportion of trimers, hexamers and dodecamers (9%, 20% and 11%, respectively). With this pool of oligomers, OLIGOMER provided the fits to the SAXS data with $\chi^2$ =0.44. No signs of unspecific protein aggregation could be detected in the analysis of the Guinier region of the data. These results clearly show that the Y73A modification of the yeast frataxin results in a higher propensity to form higher order oligomers, independently of the presence of metals, perhaps as a result of reduced flexibility of the N-terminal region of the protein. Presumably, during crystallization the trimeric form is selected as the most stable in the formation of the crystal lattice.

**Dynamic light scattering studies of cobalt induced oligomerisation of yeast frataxin**

To further investigate the effects of various factors on the oligomerization of yeast frataxin, we studied the effect of divalent metals, since as mentioned above, biochemical data clearly indicate that $Fe^{2+}$ can induce the formation of Yfh1 oligomers. To avoid the formation of

metal oxidation products, we tested the effect of $Co^{2+}$. Initially we examined the effect of metal addition on the oligomeric state of the protein using DLS. The protein to metal ratios (Yfh1 : $CoCl_2$) were 1:10, 1:20 and 1:50 and the protein concentration was 3 mg/ml. We also used the Y73A Yfh1 variant, which as shown above, is prone to the formation of higher order oligomers, and the wild type metal-free Yfh1 as controls. For each concentration of $CoCl_2$, light scattering and the respective polydispersity indexes were measured several times over a period of 30 minutes. After approximately 30 minutes the polydispersity index was stable, indicating that equilibrium has been reached (**Fig. 5-9**, inset).

As shown in **Fig. 5-9**, at Yfh1 : $CoCl_2$ ratio of 1:10, no substantial changes in the DLS profile can be observed, as compared to the protein solution without added metal. Increasing the ratio to 1:20 results in tailing of the peak with a particle size distribution around 2.3 nm. At the Yfh1 : $CoCl_2$ ratio of 1:50 three separate peaks with particle size distributions around 3.5 nm, 15 nm and 125 nm were observed. This clearly demonstrates that Co(II) induces oligomerization of yeast frataxin into larger particles.

**Characterisation of cobalt induced yeast frataxin oligomers using SAXS**

To further characterize the type of oligomers induced by the addition of $Co^{2+}$, we examined the same protein : metal ratios from the DLS experiments (1:10, 1:20, 1:50) in SAXS measurements. Analysis of the Guinier region confirmed that addition of $Co^{2+}$ to the sample at these concentrations did not induce unspecific aggregation for most of the samples. Only the sample with 1:50 protein : metal ratio was showing signs of polymerization, given the very high average $R_g$ value of 5.45 nm (compared with 1.8 nm for the monomeric protein, as shown above). A comparison of the calculated average molecular weights, using both excluded volume and lysozyme as a standard, revealed an increase with higher concentrations of $Co^{2+}$. **Fig. 5-9** and **Table 5-3** show that the data for the 1:10, 1:20 and 1:50 ratios could be fitted with $\chi^2$ values of 0.55, 0.55 and 0.67, respectively, given the input we provided from our pool of oligomers, and using the program OLIGOMER. Analysis of the volume distributions of the different oligomer fitting to the respective SAXS profiles shows that as $Co^{2+}$ concentration increase, so does the volume fraction of frataxin trimer. In other words, $Co^{2+}$ strongly promotes formation of the trimers.

Analysis of the $Co^{2+}$-induced oligomeric states (**Table 5-3**) clearly shows a gradual increase in the percentage of trimers and hexamers and a concomitant decrease in the percentage of monomers, from 67% at 1:10 protein : metal ratio to 26% at the 1:50 ration. At the 1:50 ratio considerably higher order oligomers are formed, clearly supporting earlier biochemical evidence on metal-induced oligomerization of yeast frataxin.

**X-ray crystallographic characterisation of cobalt binding to yeast frataxin**

The crystals of Yfh1 (Y73A variant), which were produced by the modified crystallisation conditions, belonged to space group $I2_13$ with one monomer in the asymmetric subunit and diffracted to 2.9 Å resolution (table not shown). As discussed previously (Karlberg, Schagerlof et al. 2006), the monomers are involved in extensive interactions around the channel at the 3-fold axis. Their N-termini form a bridge between the monomers, which adds to the stability of the trimeric structure. The new data allowed the electron density for the complete N-terminal part of the structure to be traced, adding nine residues (V60-V52) to the new model. The side chains of the newly added S54 and N59 are at hydrogen bonding distance from the side chains of H106 and E103, respectively (**Fig. 5-11**), thus showing the possibility of additional interactions between the N-terminal sequence and the β-sheet. Analysis of the packing and interactions within the crystal lattice are of interest for understanding the interactions which stabilize the oligomers. The previous model of frataxin (2FQL.PDB) showed that the trimers within the crystal lattice were packed against each other with the N-termini from neighbouring trimers building an intersection (involving amino acid residues P69, L70, E71 and K72). In the new model, due to the extension of the N-terminal part, we could observe an additional intersection involving amino acid residues Q59, V60 and V61, (**Fig. 5-11b**). These interactions contribute to the stabilization of the interaction between trimers within the crystal lattice (**Fig. 5-11c**) and may also stabilize transient monomer-monomer interactions in thus driving the oligomerizaiton towards higher order oligomers.

To assess the ability of cobalt to bind to frataxin we first soaked the crystals of frataxin trimers with $CoCl_2$. X-ray data were collected and the presence of bound metal in the structure was validated by the examination of the Fo-Fc and the anomalous Fourier difference maps. Our previous crystallographic work demonstrated that the 3D structure of frataxin trimers pre-loaded with iron contained a metal bound in

*Table 5-2. The results of SAXS measurements on the monomeric yeast frataxin in the absence and presence of glycerol.*

| Sample | $R_g$(nm) | $D_{max}$(nm) | $V_e$(nm$^3$) | $V_p$ (nm$^3$) | $\chi_g$ | $\chi_x$ | $\chi_n$ | $\chi_b$ | $\chi_o$ | $\chi_f$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Yfh1 wt | 1.8 | 6.8 | 28 | 24 | 1.16 | 1.19 | 1.16 | 1.04 | 1.03 | 1.02 |
| Yfh1 wt + Glycerol | 2.25 | 7.5 | 32.2 | 44 | 1.36 | 1.2 | 3.26 | 1.4 | 2.53 | 1.01 |

$R_g$(nm), $D_{max}$, $V_e$(nm$^3$) and $V_p$(nm$^3$) are radius of gyration, maximum size, excluded volume and Porod volume, respectively. $\chi_g$ – discrepancy to *GASBOR ab initio* model, $\chi_x$ - discrepancy to new X-ray structure, $\chi_n$ - discrepancy to lowest energy NMR conformer (model #1 in 2GA5.pdb), $\chi_b$ - discrepancy to *BUNCH* model, $\chi_o$ - *OLIGOMER* discrepancy to twenty NMR conformers and X-ray model, $\chi_f$ - discrepancy to EOM ensemble model.

*Table 5-3. The combinations of the models with different oligomeric states used for fitting to the experimental data (figure 5-10) are shown together with their respective distributions in the mixture and the discrepancy ($\chi^2$). $R_g$(nm) and $D_{max}$ are radius of gyration and maximum size, respectively.*

| Sample | $R_g$ (nm) | $D_{max}$ (nm) | Oligomer Model | Distribution (%) | $\chi^2$ |
|---|---|---|---|---|---|
| Y73A Yfh1 | 3.27 | 11.5 | 1[a]/1[b]/3[c]/6[e]/12[f] | 36/24/9/20/11 | 0.44 |
| 1:10 Co[II]:Yfh1 | 2.35 | 8.06 | 1[a]/3[d]/6[e] | 67/24/9 | 0.55 |
| 1:20 Co[II]:Yfh1 | 2.53 | 8.87 | 1[a]/3[d]/6[e] | 54/33/13 | 0.55 |
| 1:50 Co[II]:Yfh1 | 5.45 | 19.1 | 1[a]/3[d]/6[e]/42[g] | 26/57/6/11 | 0.67 |

[a]NMR conformer(s), [b]Kinematic Loop Model(s), [c]Tilted trimer from Rosetta docking, [d]Trimer with rebuild N-terminus using Rosetta, [e]Hexamer from EM density docking, [f]Dodecamer from EM density docking, [g]42 mer from combining two EM docked 24 mers and removing two overlapping trimers

the channel at the three-fold axis of the trimer (Karlberg, Schagerlof et al. 2006), although the resolution of that structure was only 3.5 Å. With the current higher resolution data, soaking with cobalt clearly showed a peak both in the Fo-Fc and anomalous difference density maps at the levels of up to 6 σ and 5 σ, respectively, for three metal ions at 0.3 occupancy bound around the crystallographic 3-fold axis in the channel between the monomers (**Fig. 5-12**).

The metal is bound at a distance of approximately 5.4 Å from the invariant D143, closer to the wider opening of the channel, which earlier was suggested to function as the entrance of the channel. It cannot be excluded that some solvent molecules are bound to the metal, however, at the present resolution of the data it is not possible to resolve these. This mode of metal binding suggests a mechanism by which the bound metal contributes to trimer stabilization, thus driving the oligomerisation process towards higher order oligomers, as demonstrated by our SAXS results.

**Conclusions**

In this work we study the effect of three different factors on the oligomerization properties of yeast frataxin Yfh1 and demonstrate that depending on the conditions, it may exist in different oligomeric states in solution. The data obtained from $Yfh1_{wt}$ in the presence and absence of glycerol clearly show that the N-terminus is highly flexible and that the conformation observed in the X-ray structure of the trimer is a result of the stabilization of the N-terminus by interactions with neighbouring subunits. Apparently the presence of glycerol facilitates the dimerization of the protein, resulting in dynamic monomer-dimer equilibrium. Generally glycerol is known to contribute to protein stability, by inducing protein compaction and reducing flexibility (Vagenende, Yap et al. 2009). Taking into account the much higher viscosity of the native environment of proteins, and due to various crowding effects, it would be logical to suggest that the monomeric form of yeast frataxin is rare *in vivo*. It may also be that in solution the dimer structure serves as a seed for the growth of higher order oligomers, since the inter-twined interaction between the monomers may be formed already at this stage.

The second factor, affecting the oligomerization behaviour of frataxin appears to be the amino acid composition of the N-terminus. Thus, the higher propensity of the Y73A variant to form oligomers shows that by a single amino acid replacement in this part of

the protein, its properties shift to becoming more human frataxin-like. This suggests that due to the poor conservation of the amino acid sequence and the length of the N-terminus within the frataxin family, the oligomerization properties of the protein in different species may be different. This may in turn contribute to the variations found in the function of the protein. The third factor, affecting frataxin oligomerization is the presence of divalent metals. In this case $Co^{2+}$, by binding inside the channel at the 3-fold axis of the trimer, facilitates the formation of oligomers in a concentration-dependent fashion. Thus, starting from a trimer, higher metal concentrations lead to the formation of hexamers and higher order oligomers. However, from our data it would be difficult to decide whether the formation of higher order oligomers is directly stimulated by the metal or if it is merely a result of the increasing number of trimers, which through interactions spontaneously assemble into higher order structures. At least in the case of $Fe^{2+}$ it is known that the metal stabilizes higher order oligomers of frataxin, while the addition of reducing agents dissolves the protein complexes (Park, Gakh et al. 2003). In the case of the frataxin homologue, *E. coli* CyaY, it is also known the $Fe^{2+}$, when added anaerobically, stimulates the formation of tetramers, while in the presence of atmospheric oxygen larger oligomers are formed (Bou-Abdallah, Adinolfi et al. 2004; Layer, Ollagnier-de Choudens et al. 2006). However, sequence analysis shows that the N-terminus of CyaY is shorter than that of the yeast and human frataxin and lacks the part, which in the case of the yeast trimer is involved in oligomer stabilization. The role of the N-terminus in yeast frataxin oligomerization may have interesting implications for the human protein, which is known to exist in at least three different forms, 81-210, 56-210 and 42-210. It cannot be excluded, that these forms will behave in a different way, depending on the external conditions, the presence of iron, etc. However, this still awaits experimental verification.

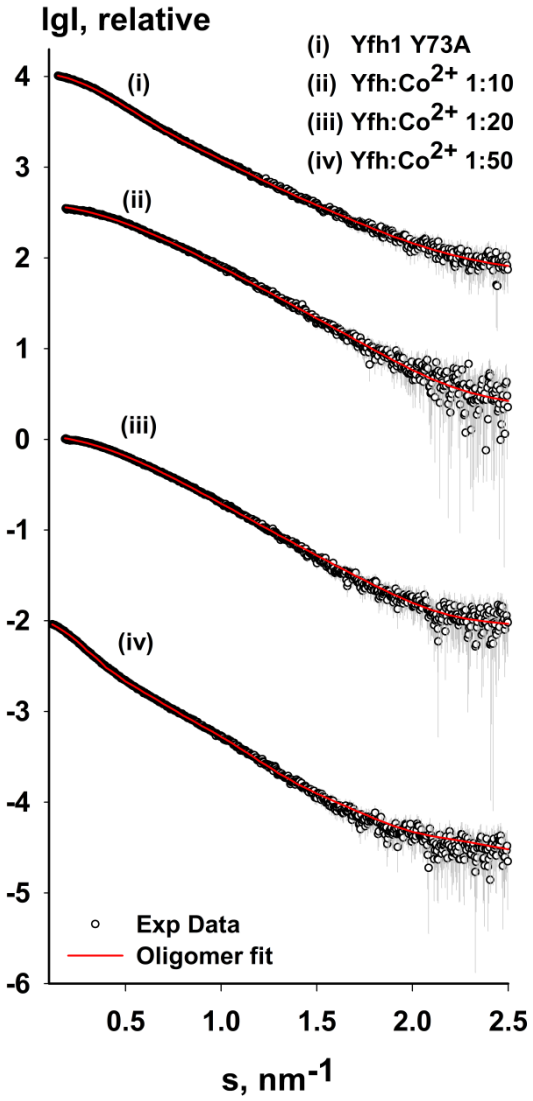



Figure 5-10. Experimental SAXS profiles (experimental data shown as black circles) for the Y73A variant of Yfh1 and wild-type Yfh1 in the presence of different amounts of metal, as shown on the figure. The corresponding fit curve is overlaid on each profile (red) (see table 5-3 for details). Data are shown to a maximum momentum transfer of s=2.5 $nm^{-1}$.

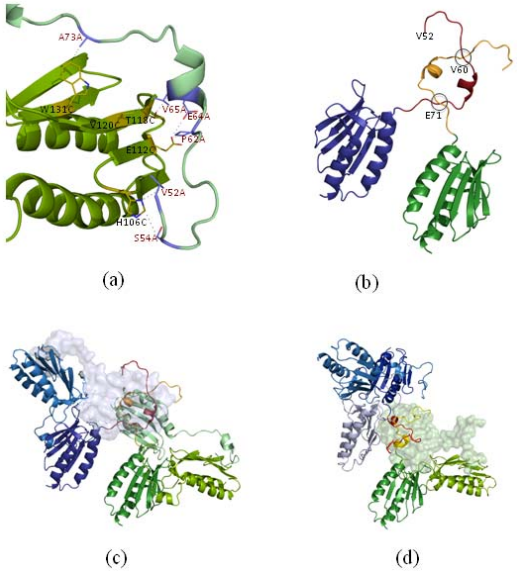



Figure 5-11. The X-ray structure of the Y73A frataxin variant. (a) The interaction of the N-terminal region of one of the subunits with the neighboring subunit is show. Amino acid side chains are shown as sticks. (b) The intersection of the N-termini of two monomers belonging to two different trimers within the crystal lattice is shown. (c) Packing of Y73A frataxin trimers, with each trimer colored uniquely. One of the monomers is shown in surface representation for clarity. The N-terminal residues 52-60 are coloured orange and brown.

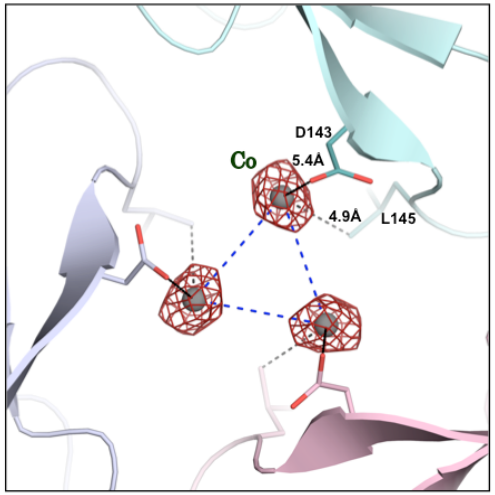



Figure 5-12. The 2Fo-Fc electron density map contoured at 1.0σ for the cobalt. Three cobalt atoms are located around the three-fold axis, with 30% occupancy each.

## 5.3 Insights into the molecular activation mechanism of the rhoa-specific guanine nucleotide exchange factor, PDZRHOGEF

Bielnicki, J.A.*, Shkumatov, A.V.*, Svergun, D.I. and Derewenda, Z.S.

*equal contributors

*Manuscript in preparation*

PDZRhoGEF (PRG) belongs to a small family of RhoA-specific RGS GEFs, which mediate signaling through selected 7-transmembrane receptors *via* $G\alpha_{12}$, and activate RhoA by catalyzing the exchange of GDP to GTP. PRG is a multidomain protein, composed of PDZ and RGSL domains followed by a catalytic tandem of DH and PH domains. PRG is autoinhibited in cytosol, and requires a conformational rearrangement and translocation to the membrane for full activation, although the molecular details of the regulation mechanism are not clear. It has been shown recently, that autoinhibition of PRG depends on the electrostatic interaction between the regulatory element ('activation box') located directly upstream of the catalytic DH domain and the RhoA-binding surface of the DH domain. Here, we show using a combination of biophysical, biochemical and structural methods that the mechanism of PRG regulation might be yet more complex, and involves an additional autoinhibitory element in the form of a molten globule region within the linker between RGSL and DH domains, which may sterically interfere with RhoA binding. We propose a novel, two-tier model of autoinhibition, where the 'activation box' and the molten globule region act concurrently to impair the ability of RhoA to bind to the catalytic DH-PH tandem. The molten globule region and the 'activation box' become less ordered in the PRG/RhoA complex, and dissociate from the RhoA binding site, which may constitute a critical step leading to PRG activation.

**State of the problem**

Dbl-homology guanine nucleotide exchange factors (GEFs) activate their cognate GTPases by catalyzing the exchange of GDP to GTP (Rossman, Der et al. 2005). PRG belongs to a small sub-family of Dbl-homology GEFs, which contain RGS domains (RGS GEFs) (Fukuhara, Chikumi et al. 2001). This family includes three homologous proteins – PRG, leukemia associated RhoGEF (LARG), and p115RhoGEF (p115). RGS GEFs link signaling through 7-transmembrane receptors acting *via* $G_{12/13}$ class of heterotrimeric G proteins, and activate their cognate GTPase – RhoA (Aittaleb, Boguth et al. 2010). Activated RhoA can consequently engage its downstream effectors, and exert a number of physiological responses leading to gene transcription, rearrangement of cytoskeleton, and transformation (Etienne-Manneville and Hall 2002; Wennerberg, Rossman et al. 2005). PRG is a multidomain protein composed of four modules – PDZ domain, responsible for association of the protein with C-terminal tails of selected 7-transmembrane receptors, RGSL domain, which acts as a GTPase-activating protein for $G\alpha_{12}$, and is believed to play a role in activating PRG, and DH domain closely followed by a PH domain, which act in tandem to catalyze exchange of nucleotide on RhoA (Fukuhara, Murga et al. 1999).

Many Dbl-homology GEFs in the basal state are autoinhibited by the supramodular arrangement of their domains (Erickson and Cerione 2004). Recently, it has been shown that Asef, a GEF for Rac1 and Cdc42, is autoinhibited by its SRC-homology 3 (SH3) domain, which binds intramolecularly to the DH domain and blocks RhoA binding site (Murayama, Shirouzu et al. 2007). Similar, supramodular architecture-based mechanism of autoinhibition was observed in Rec1-specific GEF, Vav1 (Yu, Martins et al. 2010). Here the catalytic site of the DH domain is obscured by an α-helix from an adjacent acidic domain. This interaction is strengthened by additional contacts of calponin-homology (CH) domain with acidic domain, and DH-PH tandem. Phosphorylation of Tyr 174 within the inhibitory helix relieves the autoinhibition. RhoA-specific p63RhoGEF is inhibited primarily by its PH domain, which has a certain degree of rotational flexibility, and can fluctuate between ‘closed’ and ‘opened’ conformations, therefore regulating the availability of DH domain for RhoA binding (Shankaranarayanan, Boguth et al. 2010).

Until recently, the molecular mechanism of RGS GEFs autoinhibition remained largely speculative, and it was believed that, similarly to other Dbl-homology GEFs, it

involves interactions between the individual domains (Erickson and Cerione 2004). Interestingly, new structural studies of PRG, p115, and LARG identified the regulatory element directly upstream of the DH domain, within the interdomain linker between RGSL and DH domains. This highly conserved region includes the 'GEF switch', which is responsible for direct intermolecular interactions with switch 1 of RhoA in LARG and p115, and is required to develop full catalytic potential (Kristelly, Gao et al. 2004; Chen, Guo et al. 2011). It has been proposed that the conformation and position of the GEF switch, and consequently GEF activity, is modulated by the flexible linker between RGSL and DH domain in p115 and its potential interactions with RGSL domain and regulatory $G\alpha_{13}$ subunit. A second regulatory component containing a stretch of four acidic residues (D706, E708, E710, and D712) – the 'activation box', which is located further upstream of the GEF switch, has been has been previously identified (Zheng, Cierpicki et al. 2009). It was showed that residues from the 'activation box' may interact electrostatically with RhoA-binding surface of DH domain, and hamper the nucleotide exchange reaction on RhoA. A charge reversal mutation to arginine (4R mutant) within the activation box causes release of autoinhibition and boosts the catalytic activity of PRG.

In the current study, a combination of biochemical, biophysical and structural techniques, including CD, DLS, SAXS, and others was used to structurally characterize PRG and its truncated variants (see fig. 5-13). It was shown that both linkers of PRG are able to form relatively compact, molten globule-like assemblies. A molten globule within the RGSL-DH linker of wild-type PRG may sterically interfere with RhoA binding, and therefore be a part of the autoinhibitory mechanism. Our results complement the current knowledge about regulation of PRG, and pave the way to fully understand the molecular basis of this complex process.

**Analysis of secondary structure and dynamics of PRG**

To gain insight into the putative supramodular architecture of PRG we experimentally analyzed the dynamics and the content of secondary structure elements within the linkers of PRG using a combination of far-UV CD and DXMS. The two linkers of PRG located between PDZ and RGSL (L1), and RGSL and DH domains (L2) are relatively long – ~180, and 220 residues, respectively – and largely unstructured according to secondary

structure predictions based on amino acid composition (**Fig. 5-13**). Furthermore, neither of the linkers possesses closely related homologs with known three-dimensional structure. This suggests that the protein might be dynamic and flexible. Unexpectedly,

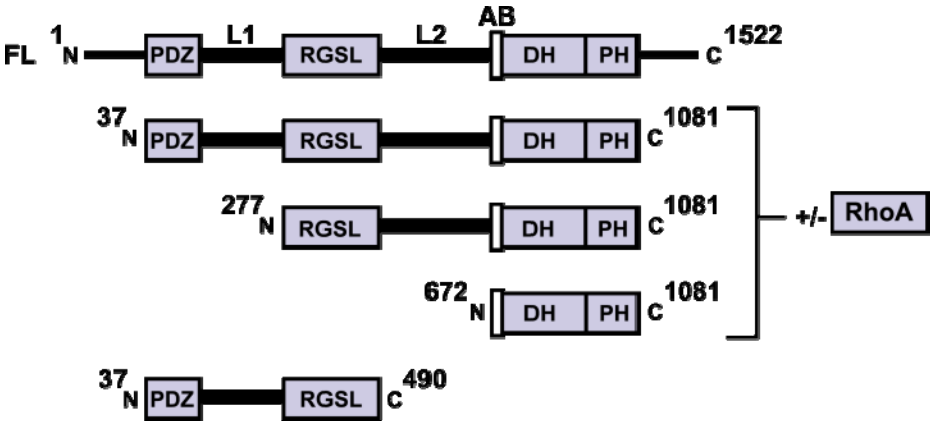



Figure 5-13. Schematic representation of multidomain PRG fragments used in this study. FL – full length, L1 – PDZ-RGSL linker, L2 – RGSL-DH linker, AB – 'activation box'. Some constructs were studied with and without RhoA, as indicated on the figure.

CD measurements of PRG 37-490 and PRG 277-1081 revealed that the content of secondary structure elements within L1 and L2 linkers is higher than predicted by bioinformatic tools (**Fig. 5-14**). If we assume that the linkers are completely unfolded, the content of α-helical elements in the PDZ-RGSL, and RGSL-DH fragments is on the order of 35 % and 45 %, respectively. These values increase by ~5 %, if we account for the presence of α-helices predicted using a program *Jpred3* (Cole, Barber et al. 2008*)*. The content of α-helices derived experimentally for the PDZ-RGSL fragment is 79 % (+/- 7 %), which is approx. 2-fold higher than for the construct with unfolded linker. The RGSL-DH fragment has 57 % (+/- 5 %) α-helices – ~ 10 % more than if the linker was unfolded. The percent content of β-strands in PRG 37-490 is negligible, whereas for the PRG 277-1081 it is roughly 13+/- 3 % (~8 % over the value for the completely unfolded linker). Those additional secondary structure elements within L1 and L2 linkers may influence an overall conformation of PRG, and aid in maintaining a relatively compact assembly. Furthermore, the stretch of ~ 40 residues upstream of the DH domain, which constitutes a fragment of the putative regulatory region is predicted to be hydrophobic (data not shown), which further increases the probability of folding into an ordered moiety.

The DXMS analysis of PRG 37-1081 revealed that both linkers are highly solvent accessible, suggesting low levels of protection from hydrogen/deuterium exchange, and consequently significant dynamics (appendix B: **Fig. B-1** *A*). However, it is noteworthy that the high level of solvent accessibility is not directly correlated with the level of protein folding. PDZ domain (residues 37-123), for instance, exchanges almost fully after even the shortest quenching times, yet it is a compact, globular domain composed of five β-strands and 2 α-helices (appendix B: **Fig. B-1** *B*).

**Characterization of PRG conformation in solution**

To further characterize the conformation, and oligomeric state of PRG in solution, a combination of SAXS, DLS, SLS, and ASEC was used. A set of truncation variants of PRG, including PRG 672-1081, PRG 37-490, PRG 277-1081, and PRG 37-1081, as well as their constitutively active mutants (D706R, E708R, E710R, and D712R – "4R"), and the respective complexes with RhoA (**Fig. 5-13**) was analyzed. To assess the compactness and flexibility of these proteins Kratky plot ($I(s) \times s^2$ as a function of $s$) was used (**Fig. 5-15** and appendix B: **Fig. B-3** *C*). Comparison of PRG Kratky plots with plots for globular bovine serum albumin and natively unfolded protein Tau (Shkumatov, Chinnathambi et al. 2011) suggest that with exception of the well-folded PRG 672-1081 (**Fig. B-3** *C*), all PRG variants are flexible in solution. PRG 277-1081 appears to be more compact than PRG 37-490, while PRG 37-1081, which combines the dynamic properties of linker L1 and L2, seems to be more dynamic than the shorter constructs. These findings were corroborated using DLS and analytical size SEC coupled with SLS. Comparison of hydrodynamic radii ($R_h$) of PRG variants with $R_h$ calculated for theoretical models of globular, branched, and linear polymers using a program *DYNAMICS* (Wyatt Technology), suggests that PRG is indeed a quite extended molecule (data not shown).

Interestingly, both DLS and SAXS show that the wild-type and the 4R mutant of PRG 277-1081 and PRG 37-1081 become more compact upon binding RhoA. These results are consistent with our hypothesis, stating that the formation of a RhoA complex may induce structural rearrangements in the supramodular architecture of PRG (**Table 5-4**).

To further support our results, we examined the ASEC elution profiles of all PRG variants (data not shown). Elution volumes for the three representative PRG constructs – PRG 37-1081, PRG 277-1081, and PRG 37-490 – show the apparent molecular weight of ~ 420 kDa, 250 kDa, and 145 kDa, respectively (**Table 5-4**). This may either suggest that the protein oligomerizes in solution, or that it adopts an extended, rather than globular conformation. The possibility of oligomerization was excluded by SLS measurements of isolated PRG 37-1081 and its RhoA complex, which may indicate that PRG conformation is responsible for the protein behavior in solution (**Table 5-4**). We limited SLS measurements to PRG 37-1081, because this fragment contains all the elements that could potentially be responsible for oligomerization.

**Identification of molten globule-like structures within the PRG linker regions**

Because of both the size and the intrinsic flexibility, PRG is difficult to study with high-resolution structural methods like X-ray crystallography, NMR or electron microscopy. To circumvent these issues SAXS was used. A comprehensive modeling of PRG variants, using *ab initio*, rigid body and combined *ab initio*/rigid body approaches was performed. Moreover, flexibility was independently assessed using EOM. Since it has been shown that modeling approaches are the most effective and informative when the scattering patterns from deletion mutants of the target protein are used concurrently (Petoukhov and Svergun 2005), we focused on four different variants of PRG, namely PRG 37-490, PRG 672-1081, PRG 277-1081, and PRG 37-1081. Moreover, we analyzed RhoA complexes of PRG 672-1081, PRG 277-1081, and PRG 37-1081, which are presumably locked in an active conformation (**Fig. 5-16** and appendix B: **Fig. B-3** *A*).

Analysis of models generated by the EOM suggests that PRG is quite dynamic, and that the linker conformations may range from extended to compact with flanking domains located in a close proximity to each other (appendix B: **Fig. B-2**). However, the experimental data cannot be well described by models with extreme compact or extended conformations, judging by the $\chi$ values of the fits. On the other hand, the discrepancies between the selected ensemble of models and the experimental data are much lower not only as compared to discrepancies of extreme (compact/extended) EOM models (appendix B: **Fig. B-2**), but also compared to discrepancies of models generated by *ab initio* and rigid body approaches (**Table 5-4** and **Fig. 5-16**).

The experimental data can be well described even if only two representative EOM models, one compact compact and one extended, are selected per ensemble. This strongly suggests that the conformation of PRG in solution may be a mixture of compact and extended states (**Table 5-4**). The distance distributions, *P(r)*, of PRG 37-490, and both the wild-type and 4R mutant of PRG 277-1081, are skewed functions with shoulders at higher distances (**Fig. 5-17**). These are typical distributions observed for extended multidomain proteins (Bernado 2010). The same is true for the *P(r)* functions of four-domain wild-type and 4R mutant of PRG 37-1081 in isolation and complexed with RhoA. In the latter case the *P(r)* plots suggest that these fragments have less distinct features, perhaps, due to the fact that they are more dynamic. Furthermore, all PRG fragments (except for PRG 37-490) seem to become more compact and ordered upon RhoA binding (**Fig. 5-17** and **Table 5-4**). The *P(r)* functions of PRG 672-1081 and PRG 672-1081 4R agree well with the model of a two-domain DH-PH tandem, and show the transition towards a globular conformation in the RhoA complexes (appendix B: **Fig. B-3** *B*). The rigid body models of all multidomain PRG fragments fit well into the *ab initio* molecular envelopes (**Fig. 5-17**). The differences between the models generated with the two methods, in particular regarding the putative position of the DH-PH tandem and conformation of the linkers, stem from the dynamic nature of the protein. The *ab initio* models often show a smearing of the features, like extendedness and loss of domain boundaries, in the reconstructions of flexible proteins (Bernado 2010).

The relatively high content of secondary structure elements within L1 and L2 linkers determined by the far-UV CD experiments suggests that PRG linkers may potentially form compact assemblies (fig. 5-14). Interestingly, inspection of the rigid body models generated by BUNCH reveals that both linker regions, indeed, seem to adopt a partially ordered structure, rather than being completely unfolded (fig. 5-17). This observation, combined with the aforementioned results suggesting highly dynamic nature of PRG, indicates that these relatively compact pseudo-domains may have the characteristics of molten globules.

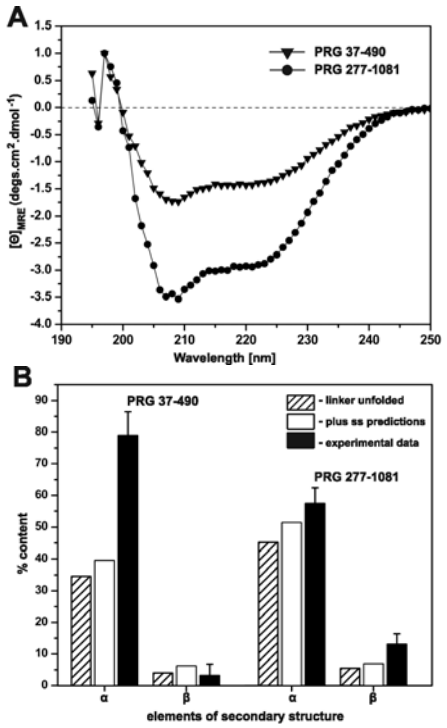
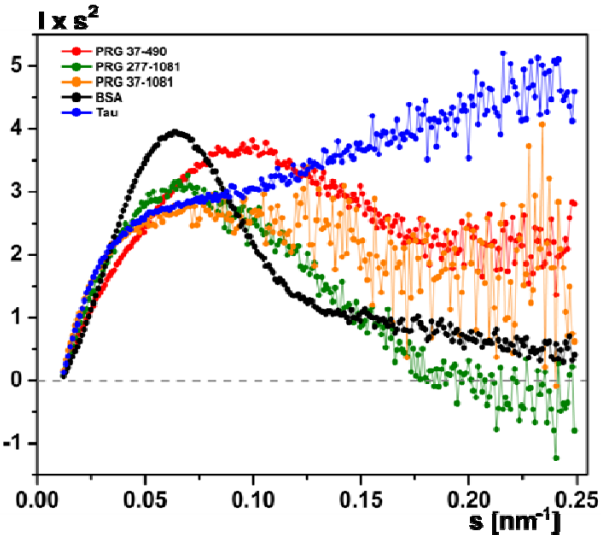
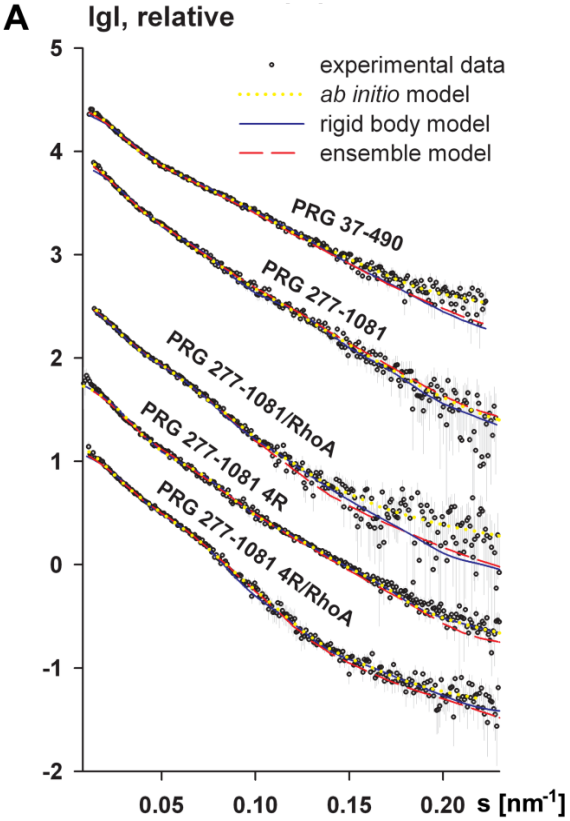



Figure 5-14. Results of far-UV CD measurements. (A) Spectra of PRG 37-490 (*solid triangle*) and PRG 277-1081 (*solid circle*) in normalized molar ellipticity units. (B) Percent content of secondary structure elements of PRG 37-490 and PRG 277-1081. Percent values assuming completely unfolded linker, including secondary structure predictions, and experimental data are shown as striped, white, and black bars, respectively.

Figure 5-15.Comparison of Kratky plots of PRG 37-490 (*red circle*), PRG 277-1081 (*green circle*), and PRG 37-1081 (*orange circle*), with bovine serum albumin (*black circle*), and protein Tau (*blue circle*).



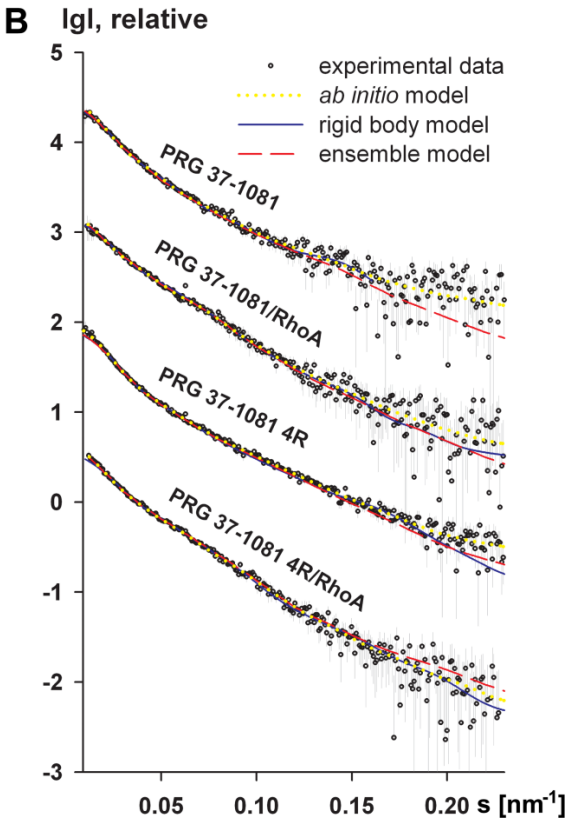




Figure 5-16. Scattering profiles for (A) truncated and (A) four-domain PRG fragments. Experimental data, fit to *ab initio*, rigid body, and EOM models are shown as open circles, yellow dotted line, blue solid line, and red dashed line, respectively.

*Table 5-4. Summary of analytical size exclusion, DLS and SLS measurements*

| | MW (kDa) | $MW_{ASEC}$ (kDa) | $MW_{SLS}$ (kDa) | $R_h$ (MAD) | %Pd (MAD)*) |
|---|---|---|---|---|---|
| **PRG 37-490** | 49 | 145 | - | 5.1 (0.23) | 11.5 (3.30) |
| **PRG 672-1081** | 48.7 | - | - | 4.1 (0.18) | 10.5 (3.50) |
| **PRG 672-1081/RhoA** | 69.3 | - | - | 4.3 (0.07) | 13.5 (2.35) |
| **PRG 672-1081 4R** | 49.3 | - | - | 4.4 (0.31) | 13.8 (5.99) |
| **PRG 672-1081 4R/RhoA** | 69.9 | - | - | 4.4 (0.18) | 12.6 (2.66) |
| **PRG 277-1081** | 93.2 | 250 | - | 6.3 (0.20) | 14.9 (3.84) |
| **PRG 277-1081/RhoA** | 114.6 | 260 | - | 6.0 (0.05) | 13.1 (7.55) |
| **PRG 277-1081 4R** | 92.8 | - | - | 7.0 (0.35) | 23.8 (4.95) |
| **PRG 277-1081 4R/RhoA** | 115.2 | - | - | 6.2 (0.20) | 12.6 (5.05) |
| **PRG 37-1081** | 118.2 | 420 | 115 | 7.7 (0.22) | 19.9 (2.78) |
| **PRG 37-1081/RhoA** | 139.6 | 450 | 137 | 7.2 (0.32) | 18.1 (4.35) |
| **PRG 37-1081 4R** | 118.8 | 390 | - | 7.8 (1.16) | 19.0 (6.05) |
| **PRG 37-1081 4R/RhoA** | 140.2 | 420 | - | 7.0 (0.45) | 9.74 (1.28) |

*) MW, $MW_{ASEC}$, and $MW_{SLS}$ are molecular weight, molecular weight determined with analytical size exclusion, and molecular weight determined by static light scattering, respectively. $R_h$ and %Pd are hydrodynamic radius and percent polydispersity, respectively. MAD is mean absolute deviation.

Similar, a partially ordered linker was also recently reported for p115 (Chen, Guo et al. 2011). The rigid body models of wild-type and 4R mutants of PRG 277-1081 and PRG 37-1081 show that the overall domain architecture is preserved. In all four PRG fragments the DH-PH domains tandem seems to fold onto the L2 linker, which suggests that PRG maintains a relatively ordered conformation despite its dynamic nature, with both linkers (if present) folding into molten globules (**Fig. 5-17**). Rigid body models of RhoA complexes also display significant conformational similarities. In all models the L2 linker is displaced from the RhoA-binding site on the DH domain, and the molten globule region upstream of the DH domain becomes less ordered than in an isolated protein. Despite the lack of any obvious structural variations, the $R_g$ values for the isolated 4R mutants and the RhoA complexes are larger than for the wild-type PRG (**Table 5-5**). This indicates that the 4R mutants of PRG might be more dynamic than wild-type proteins, and consequently suggests that the 4R mutants have a lower degree of order within the linker regions.

*Table 5-5. Overall structural parameters of PRG variants and their complexes with RhoA obtained by SAXS*

| | $R_g$ (nm) | Dmax (nm) | $v_p$ ($nm^3$) | χs | χg | χrb | χf *) |
|---|---|---|---|---|---|---|---|
| **PRG 37-490** | 5.8 | 21.5 | 118 | 0.932 | 1.1 | 1.36 | 0.80 |
| **PRG 672-1081** | 2.95 | 9.8 | 84 | 1.00 | 1.08 | 1.02 | 0.84 |
| **PRG 672-1081/RhoA** | 2.8 | 9.1 | 120 | 1.14 | 1.28 | 1.00 | 0.71 |
| **PRG 672-1081 4R** | 2.87 | 9.2 | 84 | 0.92 | 1.26 | 1.10 | 0.79 |
| **PRG 672-1081 4R/RhoA** | 2.8 | 9.2 | 124 | 0.92 | 1.13 | 0.95 | 0.80 |
| **PRG 277-1081** | 6.5 | 22 | 215 | 0.8 | 0.85 | 0.8 | 0.72 |
| **PRG 277-1081/RhoA** | 5.85 | 20 | 240 | 0.87 | 1.17 | 1.0 | 0.95 |
| **PRG 277-1081 4R** | 7.3 | 27.5 | 233 | 0.90 | 0.97 | 1.1 | 0.78 |
| **PRG 277-1081 4R/RhoA** | 6.44 | 25 | 270 | 0.77 | 0.8 | 0.83 | 0.74 |
| **PRG 37-1081** | 7.7 | 28 | 295 | 0.89 | 1.09 | 1.00 | 0.97 |
| **PRG 37-1081/RhoA** | 6.6 | 24.5 | 240 | 0.76 | 0.88 | 0.72 | 0.69 |
| **PRG 37-1081 4R** | 8.1 | 30 | 325 | 0.79 | 0.79 | 0.82 | 0.72 |
| **PRG 37-1081 4R/RhoA** | 7.6 | 27 | 330 | 0.79 | 0.80 | 0.71 | 0.684 |

*) $R_g$, Dmax, and $V_p$, are radius of gyration, maximum size, and excluded volume, respectively. χs, χg, χrb, and χf are discrepancies between the experimental data and computed scattering curves from *DAMMIN*, *GASBOR*, *BUNCH*, and *EOM* models, respectively.

We therefore propose that this difference in protein dynamics may alter the catalytic activity of PRG towards RhoA. Additionally, PRG variants complexed with RhoA have smaller $R_g$ values than isolated proteins, and this is consistent with the changes in the shape of their *P(r)* functions (**Table 5-5**). Therefore, we conclude that the degree of order within the linker regions, and consequently the compactness of PRG supramodular architecture depends on the functional state of the protein and may constitute the basis for PRG regulation.

We have shown previously, that the regulatory element of PRG ('activation box') is located upstream of the DH domain, within linker L2 (Zheng, Cierpicki et al. 2009). We used SAXS to investigate the potential conformational changes within this region between autoinhibited wild-type, and active 4R mutant of PRG 277-1081, and PRG 37-1081. The rigid body models show that the molten globule-like region upstream of the

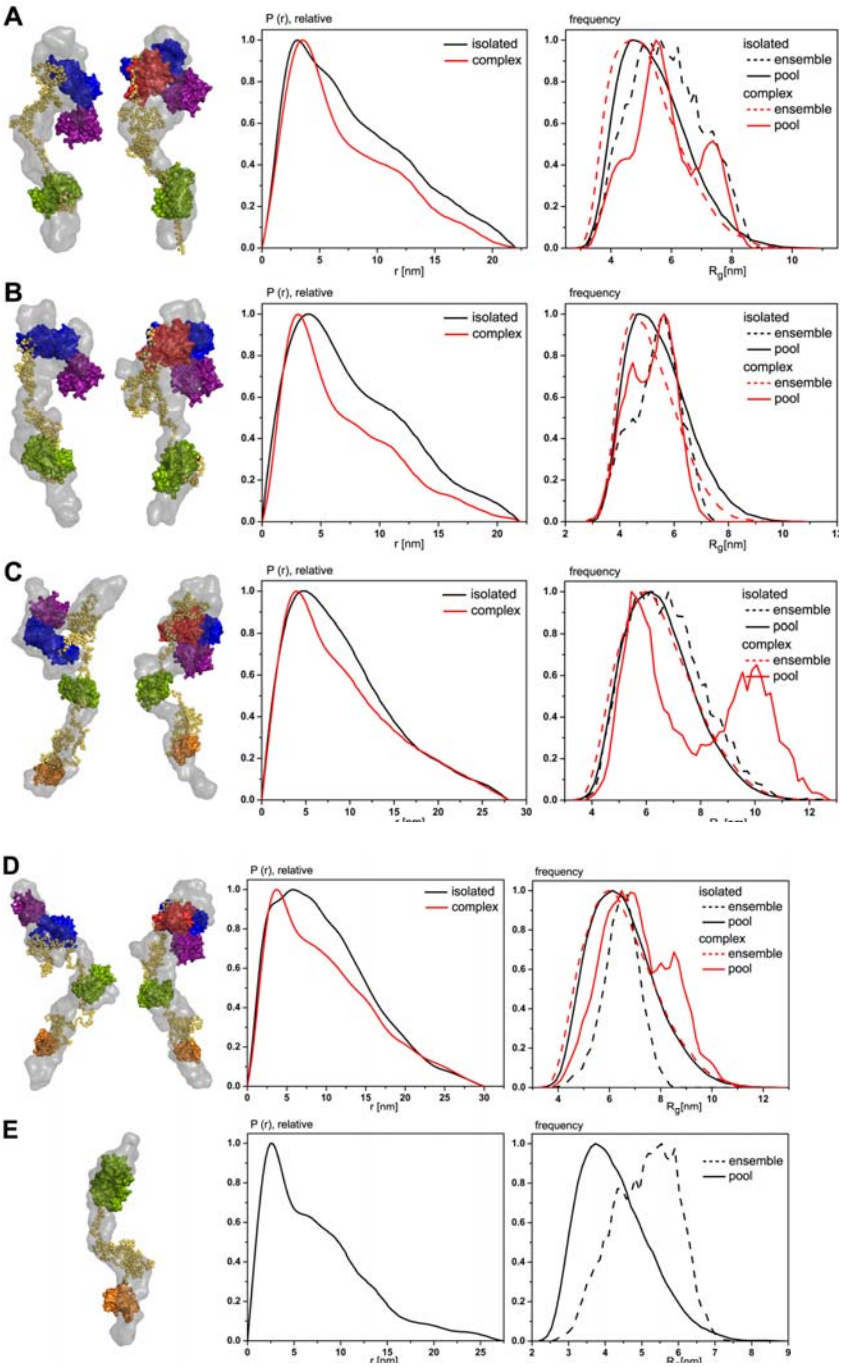



Figure 5-17. Rigid body models (BUNCH) superposed onto the *ab initio* models (GASBOR) of isolated PRG (*left*), and PRG/RhoA complex (*right*). (A) PRG 277-1081, (B) PRG 277-1081 4R, (C) PRG 37-1081, (D) PRG 37-1081 4R, and (E) PRG 37-490. *Ab initio* models, and folded domains of PRG and RhoA are depicted in surface representation (*ab initio* – gray, PDZ – orange, RGSL – green, DH – blue, PH – purple, RhoA - red), linker regions are represented as yellow spheres. Middle panel show distance distribution function, *P(r)*, of isolated PRG (*black line*) and RhoA complex (*red line*). Right panel shows the distribution of radius of gyration, $R_g$, for pools (*black* and *red solid lines* for the isolated PRG and the RhoA complex, respectively), and selected ensemble of models (*black* and *red dashed lines* for the isolated PRG and the RhoA complex, respectively).

DH domain is more compact in the wild-type protein than in the 4R mutant. Moreover, in the wild-type PRG the molten globule would overlap with RhoA molecule bound to a DH domain to a significantly larger degree than in the 4R mutant, which might explain the increased affinity towards RhoA in the latter (**Fig. 5-18**). Overall these results provide the structural insight into the role of the 'activation box' in the mechanism of autoinhibition, and suggest the presence of another, steric component influencing the activity of PRG.

**Model of regulation**

Based on the results of our study we propose a mechanism for the autoinhibition of PRG, and we hypothesize on the mechanism of activation. PRG, unlike other GEFs from the RGS family (e.g. Asef, Vav1, or p63RhoGEF) (Murayama, Shirouzu et al. 2007; Shankaranarayanan, Boguth et al. 2010; Yu, Martins et al. 2010), does not exhibit a well defined supramodular architecture where direct interactions between domains inhibit the catalytic activity. Instead, PRG seem to utilize a more subtle mechanism, governed by regulatory elements within the L2 linker, upstream of the DH domain. Using nucleotide exchange activity assay and NMR we have previously identified an autoinhibitory region, the 'activation box', which electrostatically interacts with positively charged patches on DH domain surface (most notably with R867 and R868), and may sterically interfere with binding of RhoA to DH (Zheng, Cierpicki et al. 2009). Two recent studies of other GEFs from the RGS family – p115 and LARG – show that the catalytic activity is also influenced by the very N-terminal extension of the DH domain, called the 'GEFswitch' (Kristelly, Gao et al. 2004; Chen, Guo et al. 2011), Sequence alignment (data not shown) indicates that this region is highly conserved among RGS GEFs, and therefore we suspect that it may also be present in PRG. GEF switch is important for increasing the rate of nucleotide exchange on RhoA, and perturbations of its conformation by the dynamic RGSL-DH linker inhibited the activity of p115. The formation and any potential conformational changes in a molten globule of PRG may add an additional steric inhibition to binding of RhoA to DH domain, and also affect the alignment of the activation box and the putative GEF switch. Better understanding of the autoinhibition mechanism allows us to hypothesize on the mechanism of PRG activation. We suggest that activation *in vivo* can only be achieved by protein-membrane and/or protein-protein interactions. Anchoring of PRG at the membrane *via* binding of PDZ domain to C-

terminal tails of relevant 7-TM receptors and hydrophobic interactions of PH domain with activated, membrane-bound RhoA, as well as interactions between PRG and Gα subunit of heterotrimeric G proteins might change the conformation and dynamics of the L2 linker and disfavor the formation of a molten globule. This may result in displacement of inhibitory elements from RhoA-binding site on DH domain, and upregulation of PRG catalytic activity (**Fig. 5-19**).

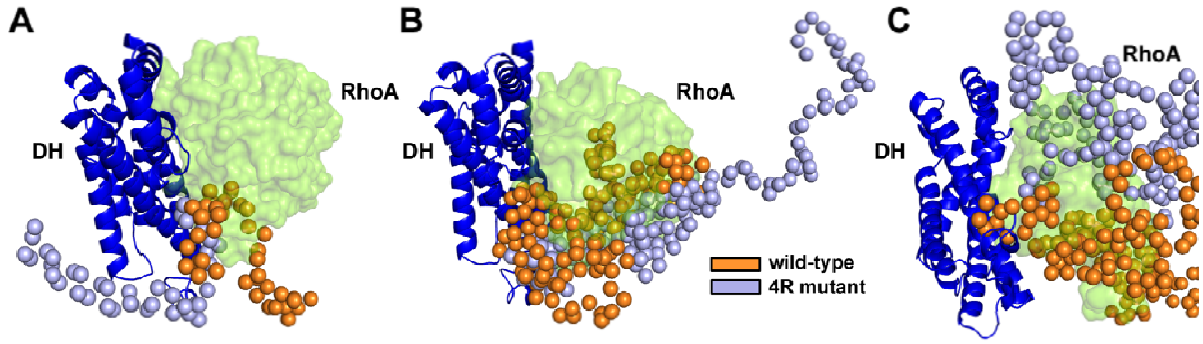



Figure 5-18. Conformational changes within the molten globule region of wild-type and 4R mutant of (A) PRG 672-1081, (B) PRG 277-1081, and (C) PRG 37-1081. Structures of wild-type and 4R mutant of PRG were superimposed using residues 714-932 of the DH domain. DH domain is shown as blue cartoon; wild-type and 4R mutant linkers are represented by orange and light blue spheres, respectively. Putative position of RhoA is shown in semi-transparent surface representation. PH domain was omitted for clarity.

## Conclusions

In this study we analyzed the structure of PRG and its truncated fragments, using a broad array of biophysical, biochemical and structural techniques. We discovered that PRG, despite being dynamic and relatively extended, maintains an ordered conformational arrangement due to the higher than predicted content of secondary structure elements, and the presence of molten globule structures within its linkers. We further show, that the molten globule located upstream of the DH domain is more ordered and compact in the wild-type PRG than in the 4R mutant, and that this region may constitute a critical element responsible for PRG autoinhibition by sterically hindering binding of RhoA to the catalytic face on the DH domain. Based on our results, we propose a new mechanism of PRG autoinhibition, where the 'activation box' and molten globule within L2 linker impair the catalytic activity of PRG due to electrostatic and steric interactions. Although these findings allow us to speculate on the PRG regulation, further structural studies

involving the active form of PRG are required to fully understand the molecular detail of this process.

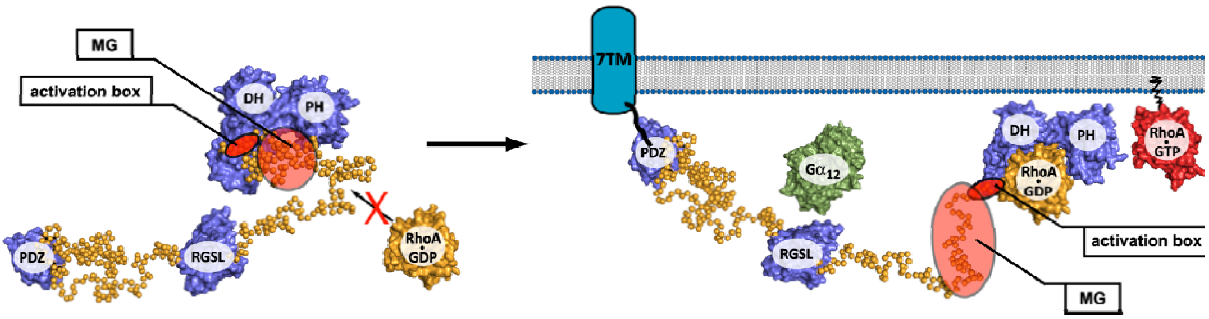



Figure 5-19. Model of regulation of PDZRhoGEF. In a resting state, in cytosol, activity of PRG is inhibited by steric and electrostatic interference with RhoA binding caused by the 'activation box' and the molten globule region within the linker. Anchoring at the membrane and interaction with Gα subunit might be one of the primary forces activating PRG, and results in conformational changes within the linker, which consequently displaces inhibitory elements from RhoA-binding site.

## 5.4 Structural memory of natively unfolded tau protein detected by Small-Angle X-ray Scattering

*Adapted from*

Alexander Shkumatov*, Subashchandrabose Chinnathambi*, Eckhard Mandelkow and Dmitri I. Svergun

**Proteins: Structure, Function and Bioinformatics. In press. doi: 10.1002/prot.23033**

*Contributed equally

SAXS is a universal low-resolution method to study size and shape of globular proteins in solution but recent developments facilitate the quantitative characterization of the structure and structural transitions of metastable systems like partially or completely unfolded proteins. We present here a study of temperature induced transitions in tau, a natively unfolded protein involved in Alzheimer's disease. Previous studies on full length tau and several disease-related mutants provided information about the residual structure in different domains revealing a specific role and extended conformations of the so-called repeat domains, which are considered to be responsible for the formation of amyloid-like fibrils ("paired helical filaments"). Here, we employ SAXS to investigate the temperature dependent properties of tau. Slow heating/cooling of the full length protein from 10$^{o}$C to 50$^{o}$C did not lead to detectable changes in the overall size. Surprisingly, quick heating/cooling caused tau to adopt a significantly more compact conformation, which was stable over up to 3 hours and represents a structural "memory" effect. This compaction is not observed for the shorter tau constructs containing largely the repeat domains. The structural and functional implications of the observed unusual behavior of tau under non-equilibrium conditions are discussed.

**State of the problem**

Alzheimer's disease is the most common form of dementia among elderly people (Haass and Selkoe 2007). Histological findings in patient brains include amyloid plaques and neurofibrillary tangles, composed of β-amyloid and tau protein, respectively. The neurofibrillary tangles are formed from β-sheet rich "Paired Helical Filaments" (PHFs) which in turn are aggregates of hyperphosphorylated tau protein (von Bergen, Friedhoff et al. 2000; Margittai and Langen 2004; Binder, Guillozet-Bongaarts et al. 2005; Mandelkow, von Bergen et al. 2007). Physiologically, tau is a microtubule-associated protein (Drubin and Kirschner 1986; Butner and Kirschner 1991) occurring mainly in the axons of neurons, where it stabilizes microtubules. In the human Central Nervous System (CNS) it is found in 6 alternatively spliced isoforms ranging from 352 to 441 residues depending on the presence or absence of exons 2, 3 and 10 (Wille, Mandelkow et al. 1992; Goedert, Jakes et al. 1996). In fetal brain the smallest isoform (ht23-lacking exons 2, 3 and 10) is the predominant one, whereas in adult brain all isoforms can be found in roughly equal amounts (Mukrasch, Bibow et al. 2009). Tau contains 4 semi-conserved sequences of 31 or 32 residues, so-called "repeats". The second repeat corresponds to exon 10 and may be absent in some of the isoforms. The repeat domain is essential both for the binding to microtubules as well as for the aggregation of tau into PHFs (Mylonas, Hascher et al. 2008). The domain structure of the full length tau is schematically illustrated in **Fig. 5-20**.

Overall, tau has a very low content of secondary structure as shown by sequence analysis (very high content of polar residues) and CD experiments (Jeganathan, von Bergen et al. 2006; Mylonas, Hascher et al. 2008). In vitro tau aggregation can be induced efficiently only by incubation with polyanions (e.g. heparin (Goedert, Jakes et al. 1996)). Moreover, phosphorylation negatively regulates both tau-microtubule as well as tau-tau interactions, with some of the sites being responsible for both (Margittai and Langen 2004). Interestingly, in Alzheimer's disease PHFs tau is hyperphosphorylated, a process that is poorly understood.

Structural investigations of tau are largely hampered by the unfolded nature of this protein. However, recent studies attempted to gain insights into the 3D structure of tau (Mylonas, Hascher et al. 2008; Mukrasch, Bibow et al. 2009). Mylonas et al (Mylonas, Hascher et al. 2008) studied the structures of various forms and deletion mutants of tau

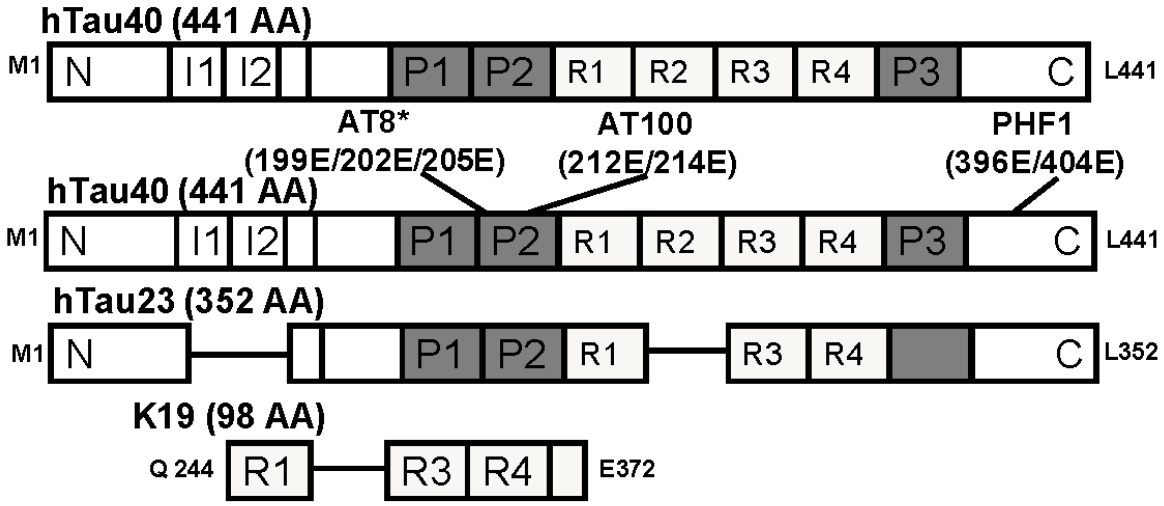



Figure 5-20. Studied tau constructs and their domain composition. Bar diagram showing the domains of tau (full length isoform hTau40wt, 441 residues, $hTau40_{AT8*+AT100+PHF1}$, hTau23wt and K19wt). To mimic the phosphorylation *in vitro,* pseudophosphorylation mutants with Glu substituted at the phosphorylable residues were used (Jicha, Bowser et al. 1997). Tau domains are broadly divided into the N-terminal 'projection domain' (amino acids M1-Y197) and C-terminal 'assembly domain' (amino acids Y198-L441). The C-terminal assembly domain includes three or four pseudo-repeats (~31 residues each, R1-R4), which together with their proline-rich flanking regions (P1and P2) constitute the microtubule binding region. Repeat R2 and the two near N-terminal inserts (I1and I2) may be absent due to alternative splicing. The N-terminal 120 residues have an overall acidic character, the rest of the molecule has a basic character.

by SAXS. From radii of gyration, Kratky plots, and the data fitting with the EOM, it was clearly shown that all constructs were unfolded, but structural differences were detected between the sequence domains. It was found that the repeat domain, which is considered the core constituent of the PHFs, is more extended than the terminal parts. Moreover, aggregation-promoting mutations have little effect on the overall shape of the protein in solution, even though they cause some rearrangements of the domains as seen by FRET (Fluorescence Resonance Energy Transfer) (Jeganathan, von Bergen et al. 2006) (Jeganathan, Hascher et al. 2008). These results favor a paper-clip model of tau structure (Jeganathan, von Bergen et al. 2006) and provide a clearer picture of the overall domain structure of tau and the contributions of different domains and phosphorylation states to the dynamic behavior of tau. Mukrasch et al. (Mukrasch, Bibow et al. 2009) analyzed the structural polymorphism of full length tau at high resolution using NMR. A novel methodology revealed that 441-residue tau is highly dynamic in solution with a distinct

domain character and an intricate network of transient long-range contacts important for pathogenic aggregation. Basically, the structural model of tau showed it in a much more compact form than previously expected from the EM images (Wille, Mandelkow et al. 1992). However, the molecule is still loosely packed, highly flexible, and exchanges between a large number of conformations, consistent with large number of conformations, consistent with large average values of the hydrodynamic radius.

**Results and discussion**

The full length tau (hTau40wt), its phosphorylation mutant ($hTau40_{AT8*+AT100+PHF1}$) and the two deletion mutants hTau23 and K19 were measured by SAXS and the scattering patterns are displayed in **Fig. 5-21**. The $R_g$ values at 10°C agree with those reported by Mylonas el al (Mylonas, Hascher et al. 2008), and the Kratky plots (**Fig. 5-21**, right panel) indicate that all constructs are unfolded. No major changes were observed between the wild type protein and the phosphorylation mutant. To investigate the effect of temperature on the overall dimensions of the constructs, we first carried out SAXS measurements at 50°C under equilibrium conditions. The samples were placed in the sample changer tray at 10°C and slowly (~30 min) warmed up to 50°C. Then the samples were injected into the measuring cell, which was also kept at 50°C. The $R_g$ values for all constructs summarized in Table 5-6 reveal no significant changes compared to the equilibrium data collected at 10°C (see also typical scattering data for hTau40wt in appendix D: **Fig. D-1**). To quantitatively characterize the ensemble properties of these constructs under equilibrium conditions, EOM was applied to all measured data sets. The EOM analysis allowed us to neatly fit the experimental data from all constructs (**Fig. 5-21**, left panel) and the obtained results depicted in **Fig. 5-22** demonstrate that no major differences are observed between the 10°C and 50°C samples. Both the average sizes and the widths of the $R_g$ distributions in the selected ensembles were only marginally affected by temperature, and all the differences were within the experimental errors. The distributions of the longer constructs were close to those of the random pools, whereas the shorter constructs appeared more extended than the random coils, in agreement with the results reported by Mylonas et al (Mylonas, Hascher et al. 2008).

Surprisingly, different results were obtained under non-equilibrium temperature conditions, when the samples were kept in the tray at one temperature and then

transferred to the measurement cell tempered to either higher or lower temperature. Two cases were explored, "quick heating" (sample tray at 10°C, measurement cell at 50°C) and "fast cooling" (sample tray at 50°C, measurement cell at 10°C). Each time when the measuring cell was filled the sample was held for half a minute prior to exposure in order to ensure that the desired temperature is reached (cell volume ~25ul). The experiments under non-equilibrium conditions (quick heating/cooling) showed that long tau constructs adopt a more compact and folded conformation, as judged by $R_g$ (**Table 5-6**) and the Kratky plots (**Fig. 5-21**, right panel). Also the ensembles selected by the EOM displayed the distributions shifted to smaller $R_g$ values compared to equilibrium temperature conditions (**Fig. 5-22**). At the same time, the short constructs demonstrated a behavior similar to the equilibrium temperature conditions, i.e. no dependence on the temperature history (**Table 5-6** and **Fig. 5-22**).

Table 5-6: Radii of Gyration

| **construct (no. of aa)** | ***Rg* (nm) at different temperature conditions. Temperature in sample holder/measurement cell, °C/°C** | | | |
|---|---|---|---|---|
| | **10/10** | **50/50** | **10/50** | **50/10** |
| hTau40wt (441) | 6.6±0.3 | 6.5±0.3 | 5.5±0.3 | 5.6±0.3 |
| $hTau40_{AT8*+AT100+PHF1}$ (441) | 6.6±0.3 | 6.7±0.3 | 5.9±0.3 | - |
| hTau23 (352) | 5.9±0.2 | 5.9±0.2 | 5.9±0.2 | 5.9±0.2 |
| K19 (98) | 3.5±0.2 | 3.5±0.2 | 3.5±0.2 | 3.7±0.2 |

Several independent experimental SAXS sessions with samples from different batches were performed to verify the observed unusual effect. They reproducibly revealed the compaction of the full length tau upon quick heating and fast cooling. The latter effect was measurable also after minutes and even hours of incubation of the protein at 10°C, indicating that the compaction has a memory effect. We performed additional measurement on incubated hTau40wt, which revealed that the compact state is preserved for at least 3 hours of incubation, but after 24 hours the protein is nearly reverted back to the native state (appendix D: **Fig. D-1** and **D-2**). The pseudo-

phosphorylated tau construct showed a tendency to aggregate over time and could thus not be measured under similar conditions.

In order to further confirm the obtained results, we employed other techniques providing information about the structure in solution. Our CD experiments did not show the memory effect when different temperature protocols were employed. Thus, the X-ray results reveal a global property, whereas CD measures average local properties. To test if elevated temperature could cause a structural transition, the secondary structure of soluble tau was determined by CD at various temperatures (5, 25, 45, 65, and 95°C). Upon stepwise elevation of temperature, the CD spectra of hTau40wt underwent a shift: the negative peak at 200 nm became less pronounced and the value at 217 nm became more negative with an isobestic point around 210 nm (**Fig. 5-23**), also see (Jeganathan, Hascher et al. 2008). By varying the time of incubation and the protocol of temperature shifts we found that the CD spectrum depends essentially on the temperature only, but not on the history. Thus, the "memory effect" seen by SAXS appears to be specific for the global changes in the protein seen by X-rays, not for the average local structure seen by CD.

To test the aggregation kinetics of hTau40wt measured by light scattering at 90° with increasing temperature, we kept the samples at 10°C for 20 min, 50°C for 20 min, and 50°C for 120 min and measured at room temperature. **Fig. 5-24A** confirms that temperature was not inducing any aggregation. The results obtained from the light scattering clearly suggest that there was no aggregation at various temperatures. The positive control of fully aggregated PHFs was confirmed by ThS fluorescence and electron microscopy. The same sets of samples were incubated, and after reaction all the samples were collected, pelleted by ultracentrifugation, and pellets and supernatants were analyzed by SDS gel electrophoresis (**Fig. 5-24B**, **C**). The results confirmed that aggregation in the tau solutions was negligible. This suggests that the memory effect revealed by X-ray scattering is due solely to changes in temperature.

To obtain independent information on the temperature dependence of the apparent size of tau DLS measurements were performed. The hydrodynamic radius $R_h$ of hTau40wt at 10°C was 5.6 nm (**Fig. 5-25a**), and this value did not change upon slow heating to 50°C (data not shown). However, when hTau40wt was quickly heated to 50°C, incubated and measured, the $R_h$ dropped to 5.3 nm (**Fig. 5-25b**). This compaction was in

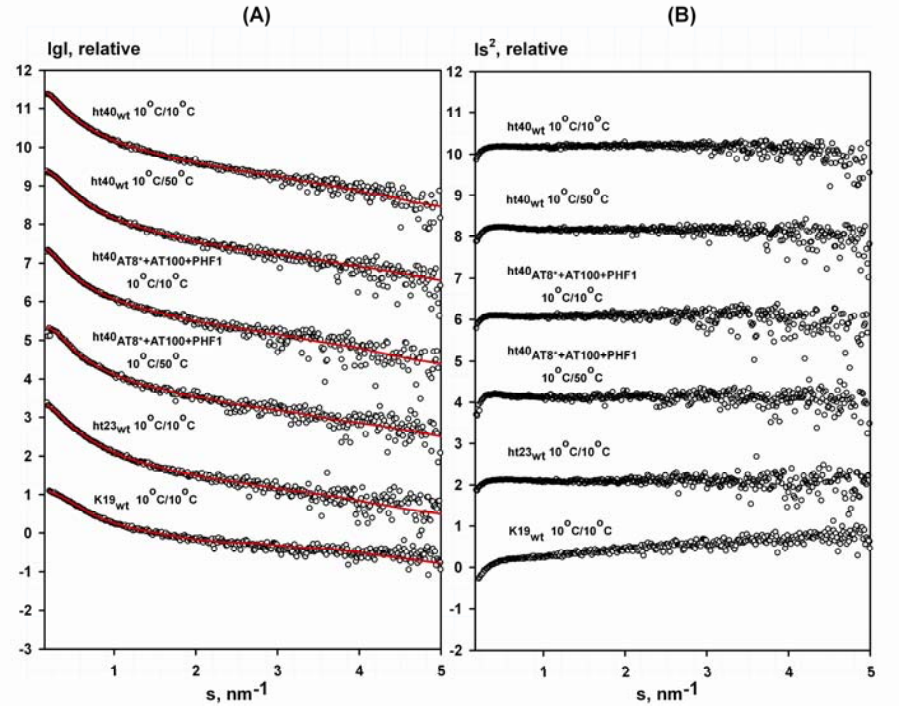



Figure 5-21. (A) Experimental SAXS data (o) with correspondingensemble fit (-) for full-length constructs (hTau40wt, hTau40$_{AT8*+AT100+PHF1}$) are shown at equilibrium ($10^oC/10^oC$) and non-equilibrium temperature conditions ($10^oC/50^oC$). Experimental data for short constructs (hTau23, K19) are shown at equilibrium temperature condition ($10^oC/10^oC$) Plots display the logarithm of the scattering intensity as a function of momentum transfer $s = 4\pi sin(\theta)/\lambda$, where $2\theta$ is the scattering angle and $\lambda$ is the X-ray wavelength. (B) Kratky plots corresponding to data in panel A. Experimental SAXS profiles were appropriately displaced along the logarithmic axis for better visualization and overlaid with corresponding ensemble fits.

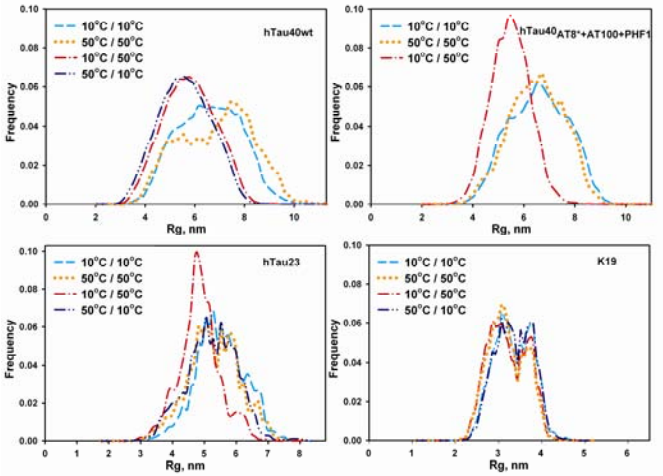



Figure 5-22. Temperature-jump induced changes in the ensemble dimensions studied by SAXS. Ensemble optimization analysis of the SAXS profile measured for full-length (hTau40wt, hTau40$_{AT8*+AT100+PHF1}$) and short (hTau23, K19) tau constructs at equilibrium ($10^oC/10^oC$, $50^oC/50^oC$) and non-equilibrium temperature conditions ($10^oC/50^oC$, $50^oC/10^oC$). Radius of gyration distributions ensembles under equilibrium conditions: $10^oC/10^oC$ (– –), $50^oC/50^oC$ (···) and non-equilibrium conditions: $10^oC/50^oC$ (– · –), $50^oC/10^oC$ (– ·· –).

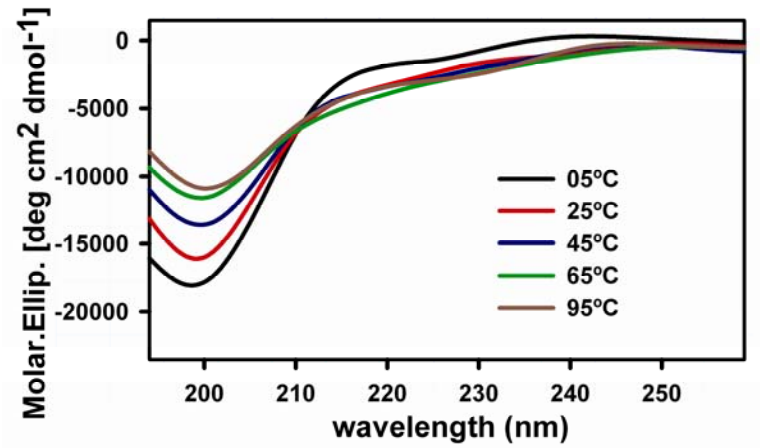



Figure 5-23. CD spectra of tau and temperature dependence. CD spectra of tau at different temperatures: 5°C, 25°C 45°C 65°C, and 95°C. Note that the negative peak at 200 nm becomes filled in as the temperature rises. Very similar traces are obtained by different protocols of varying the temperature change and time of incubation, indicating the presence, but not the extent of local conformational changes.

line with the SAXS results under non-equilibrium temperature conditions, although the decrease of the $R_h$ was smaller than that of the $R_g$ values from SAXS. It is interesting to note here that the ratio of $R_g/R_h$ depends on the flexibility of the macromolecule. For rigid structures, $R_g$ is close to $R_h$ ($R_g/R_h$ ranging from ~1.1 for anisometric polymers to ~0.8 for a solid sphere). For flexible systems $R_g$ may significantly exceed $R_h$, such that $R_g/R_h$ reaches 1.5 for a random coil (Rubinstein and Colby 2003). Our results therefore indicate that the quenched tau becomes not only more compact, but, expectedly, also less flexible than the native protein.

The polypeptide chain of tau has been shown to remain mostly natively disordered, loose and flexible under different conditions (Jeganathan, von Bergen et al. 2006; Mukrasch, Bibow et al. 2009). This structural plasticity is necessary for the unique functional repertoire of IDPs, which is complementary to the catalytic activities of ordered proteins (Uversky 2009). Changes occurring under different conditions have been reported for tau (Mukrasch, Bibow et al. 2009). There is some global folding (hairpin model, whereby N- and C-terminal domains approach the repeat domain), which could affect $R_g$. It is noteworthy that the hairpin model also shows compaction in some conditions (as determined by FRET), for example after hyperphosphorylation at several sites, consistent with the reaction with antibodies MC1 or Alz50, that are characteristic of incipient Alzheimer disease (Jeganathan, Hascher et al. 2008). It is therefore clear that the unfolded nature of tau protein allows it to adopt either more extended or compact conformations. It is intriguing to speculate that the compaction and memory effect observed after rapid heating/cooling may be related to the compaction observed upon hyperphosphorylation, characteristic of incipient neuronal degeneration in AD. The memory effect is observed with full-length tau, but not with the repeat domain alone, suggesting that the interplay between different domains in the whole protein might be responsible for the effect. One possible explanation could be the interplay between the acidic N-terminal domain (which varies among the tau isoforms due to alternative splicing) and the basic repeat domain (which also differs between isoforms). These issues will be addressed in future studies.

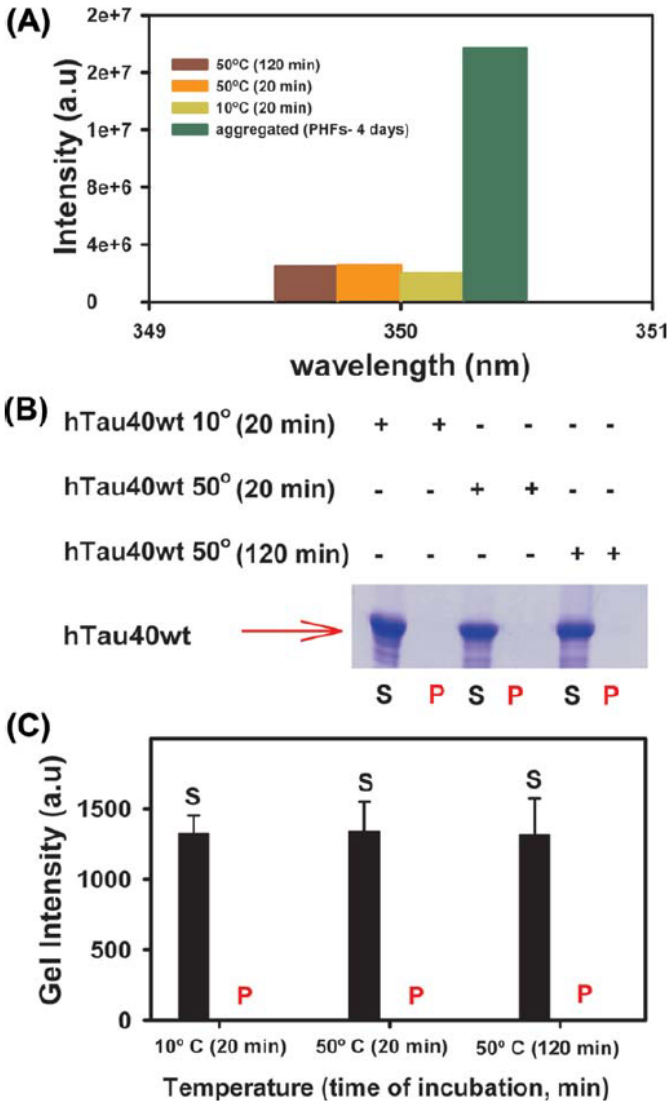



Figure 5-24. Light scattering and sedimentation analysis of tau. (A) Light scattering (90°, 350 nm) of soluble tau. Samples of soluble tau protein (concentration 25 µM, phosphate buffer pH 6.8) were incubated at various temperatures and time periods, 10°C for 20 min, 50°C for 20 min, 50°C for 120 min. Note that there was no increase in scattering intensity from the samples of soluble tau, indicating that there was no measurable aggregation during the duration of the experiment. As a control, aggregated PHFs (concentration 25 µM of tau) show a high scattering intensity. (B) Coomassie blue (R-250) stained with SDS-PAGE illustrating the sedimentation analysis of soluble tau at different temperatures and incubation periods. S and P represent supernatant and pellet. Note that there is no detectable aggregated protein in the pellet. (C) Quantification of (B).

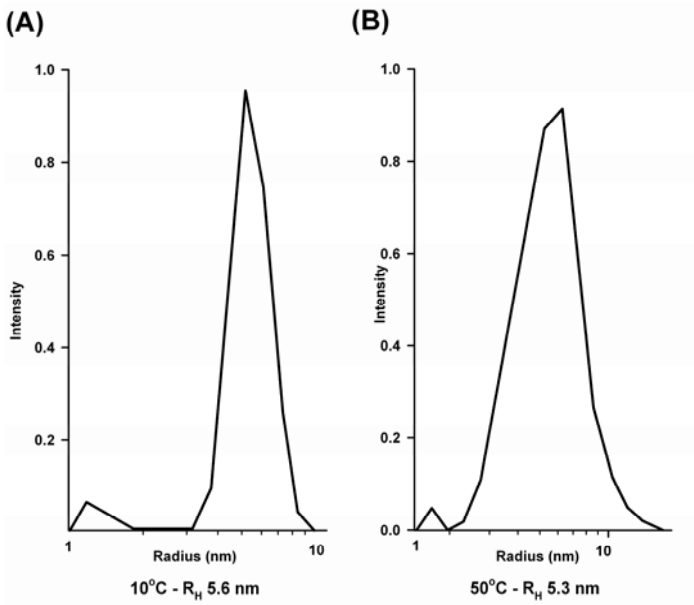



Figure 5-25. Dynamic light scattering measurement for hTau40wt at different temperatures. (A) Example of determination of $R_H$ by DLS (10°C for 20 min ($R_H$ 5.6 nm). (B) Comparison of $R_H$ values obtained at different temperatures, e.g. 10°C for 20 min ($R_H$ 5.6 nm) and 50°C for 120 min ($R_H$ 5.3 nm). Note that at increasing temperature (50°C for 120 min), $R_H$ becomes smaller. This indicates that the protein is somewhat more compact at higher temperature.

## Conclusions

IDPs are important representatives of metastable proteins, which are difficult to study by high resolution X-ray crystallography due to their inherent flexibility. SAXS is one of the most important methods allowing for quantitative structural descriptions of such objects. The possibility of modeling the SAXS data in terms of ensembles of conformations opens new perspectives in the study of unstructured and partially folded proteins. SAXS is easily applicable under changing conditions like temperature, pH, salinity etc, and even under non-equilibrium conditions, as demonstrated in the present study. The structural information that can be derived from SAS data applying the EOM strategy can be complemented by information collected with other structural techniques. Especially promising here is the combination with NMR that has been traditionally applied to the study of unfolded and partially folded proteins, but the techniques like CD, DLS, calorimetry also deliver important additional information. The potential of SAXS employed together with other methods is illustrated in the present work revealing an unusual effect of compaction of the intrinsically unfolded tau protein under non-equilibrium temperature conditions. An analogous compaction has also been observed as a result of posttranslational modifications of tau and is reminiscent of tau conformations during incipient Alzheimer disease, as seen by conformation-dependent antibodies (Jicha, Bowser et al. 1997; Jeganathan, Hascher et al. 2008).

# Concluding discussion

Over the last decades SAXS has gained increasing popularity in the structural biology community. The trend is reflected in the growing number of projects where SAXS can complement other structural biology methods. Thus, SAXS can either provide additional information or help to resolve some ambiguities. Moreover, SAXS can be employed as a standalone technique to get overall structure, study oligomeric properties and/or assess the flexibility of the biological macromolecules. Our group has been leading the development of novel computational methods for constructing 3D structural models from the scattering data. Special attention is given to the joint use of SAXS with structural, biophysical and biochemical techniques including macromolecular crystallography, NMR, electron microscopy, neutron scattering and bioinformatics (Petoukhov and Svergun 2007; Putnam, Hammel et al. 2007; Mertens and Svergun 2010).

The present study addresses two directions, namely, method development/improvement and practical application of SAXS to study the solution structure of biological macromolecules. In order to meet the growing demands of the structural biology community, existing tools must be fast, reliable and easy to use. A palette of methods, available as package suite ATSAS (Petoukhov, Konarev et al. 2007), offers useful tools for answering the major structural questions addressed by SAXS. In the scope of my thesis, I contributed to ATSAS by creating faster versions of CRYSOL and CRYSON for X-ray and neutron scattering calculation from atomic structures, respectively. Moreover, I developed two new programs, RANLOGS and EM2DAM, for constructing random linkers and validation of low resolution models from electron microscopy, respectively. RANLOGS became a part of a newly developed combined *ab initio*/rigid body modeling program CORAL, but can also be used as a stand-alone application. EM2DAM can rapidly convert a density map in MRC format to a bead model in PDB format. The latter can be used for calculation of the theoretical intensity and fitting to the experimental scattering curve using CRYSOL or for rigid body modeling with SASREF. Both, RANLOGS and EM2DAM were included in the recent release of the ATSAS package. Two programs, which are not part of ATSAS package, were developed by me. The first one is used to automatically select and refine models using HADDOCK (Shkumatov, Svergun et al. 2009) and experimental SAXS data. The second program (NMADREFS) is able to refine binary protein-protein complexes using

NMA and SAXS experimental data. The first method introduces relatively small changes to the structure of individual macromolecules when complexation is modeled. The second method is designed to account for possible changes upon complex formation.

Modularity is an important concept in biology. During the course of evolution common building blocks have been reused many times. Search for either structural or sequential patterns in the proteins to infer function is one of the bioinformatics tasks nowadays. I applied different bioinformatics predictors while trying to gain insight into Tardigrades' stress-specific adaptation potential. Tardigrades represent an animal phylum with extraordinary resistance to environmental stress. We found (Forster, Liang et al. 2009) that 39.3% of the total sequences clustered in 58 clusters of more than 20 proteins. Among these are ten tardigrade specific as well as a number of stress-specific protein clusters. Tardigrade-specific functional adaptations include strong protein, DNA- and redox protection, maintenance and protein recycling. Specific regulatory elements regulate tardigrade mRNA stability such as lox P DICE elements whereas 14 other RNA elements of higher eukaryotes are not found. I co-developed the web-tool "tardigrade analyzer". The work-bench offers nucleotide pattern analysis for promotor and regulatory element detection (tardigrade specific; nrdb) as well as rapid COG search for function assignments including species-specific repositories of all analysed data.

Although Tardigrade project is somewhat disconnected from the main direction of this thesis, it allowed me to learn bioinformatics concepts, which are generally quite useful and helpful when primary protein sequences need to be analyzed. This knowledge is advantageous when one performs rigid body modeling or/and analyzes the sequence properties of IDPs for instance.

In my work, I employed the newly developed methods together with the other interpretation and analysis techniques from the ATSAS suite in several collaborative projects. In all these projects, I participated in the SAXS experiments at the EMBL X33 beamline as a local contact for the external groups. Furthermore, I analyzed the scattering data and built the structural models in collaboration with the biological users of the beamline. I participated in about twenty of collaborative projects; in this thesis I included the projects which are either already published or the manuscripts are being submitted. Most of these projects are dealing with rather complicated objects, exhibiting intrinsic flexibility, and they are difficult to describe by other structural techniques than SAXS.

Using SAXS as a major experimental method, I managed to complement ongoing studies of biological macromolecules in solution and obtained new structural insights for a number of proteins with different degrees of flexibility which are summarized below.

The class-III receptor tyrosine kinase (RTKIII) Flt3 and its cytokine ligand (FL) play central roles in hematopoiesis and the immune system, by establishing signaling cascades crucial for the development and homeostasis of hematopoietic progenitors and antigen-presenting dendritic cells. However, Flt3 is also one of the most frequently mutated receptors in hematological malignancies and is currently a major prognostic factor and clinical target for AML. As a result of a collaborative effort, the structure of Flt3-FL complex was scrutinized using different structural biology and biophysical methods (Verstraete, Vandriessche et al. 2011). I contributed to the project by performing SAXS measurements as well as constraint rigid body modeling using SASREF with various possible scenarios. Together, data derived from different methods suggested that Flt3-FL complex is driven in part by a single preformed binding epitope on FL reminiscent of a 'lock-and-key' binding mode, thereby setting the stage for antagonist design.

Frataxin is a mitochondrial protein with a central role in iron homeostasis, including iron storage and detoxification, iron delivery to iron-sulphur cluster synthesis, heme synthesis and aconitase repair. Deficiency in frataxin function has been attributed to the progressive neurodegenerative disease Friedreich's ataxia. Numerous biochemical studies have demonstrated that the function of frataxin is closely associated with its ability to form oligomeric species, however, the mechanisms of oligomerization, the types of oligomers present in solution and their possible function in iron homeostasis are poorly understood. I investigated the influence of different factors on oligomerization properties of frataxin. I found that glycerol, although thought to stabilize proteins and make them more compact, induced oligomerization of frataxin, resulting in dynamic monomer-dimer equilibrium in solution. The presence of $Co^{2+}$ induced the formation of oligomers, which were present as multiples of 3, which confirmed the earlier suggestion that trimer is the main building block of higher order frataxin oligomers.

PRG belongs to a small family of RhoA-specific RGS GEFs, which mediate signaling through selected 7-transmembrane receptors *via* $G\alpha_{12}$, and activate RhoA by catalyzing the exchange of GDP to GTP. PRG is a multidomain protein, composed of PDZ and RGSL domains followed by a catalytic tandem of DH and PH domains. PRG is

autoinhibited in cytosol, and requires a conformational rearrangement and translocation to the membrane for full activation, although the molecular details of the regulation mechanism are not clear. It has been shown, that autoinhibition of PRG depends on the electrostatic interaction between the regulatory element ('activation box') located directly upstream of the catalytic DH domain and the RhoA-binding surface of the DH domain. Using SAXS, I showed that the mechanism of PRG regulation might be yet more complex, and involves an additional autoinhibitory element in the form of a molten globule region within the linker between RGSL and DH domains, which may sterically interfere with RhoA binding. We proposed a novel, two-tier model of autoinhibition, where the 'activation box' and the molten globule region act concurrently to impair the ability of RhoA to bind to the catalytic DH-PH tandem. The molten globule region and the 'activation box' become less ordered in the PRG/RhoA complex, and dissociate from the RhoA binding site, which may constitute a critical step leading to PRG activation.

The frataxin and Flt3-FL complex are relatively rigid molecules with flexible N-terminus and inter-domain hinge regions, respectively. The latter are only few residues long, the former is somewhat longer, but still relatively short. PDZRhoGEF, on the other hand, has rather long flexible linkers, which are recalcitrant to characterization by crystallographic methods. Moreover, one can neither find close nor remote homologues to linker regions with known crystal structures. There are high resolution structures available for the domains connected by the unstructured linkers providing useful information as well as restraining the SAXS model. A yet more challenging target, however, is a natively unfolded tau protein involved in Alzheimer's disease. Structural information about tau is very limited. Tau-microtubule interactions are known to be very temperature sensitive. Employing SAXS I investigated (Shkumatov, Chinnathambi et al. 2011) the temperature behaviour of tau. Slow heating/cooling of the full length protein from 10$^{o}$C to 50$^{o}$C did not lead to detectable changes in the overall size. Surprisingly, quick heating/cooling caused tau to adopt a significantly more compact conformation, which was stable over up to 3 hours and represents a structural "memory" effect. This compaction is not observed for the shorter tau constructs containing largely the repeat domains. We believe that the observed structural plasticity is necessary for the unique functional repertoire of tau, which is complementary to the catalytic activities of ordered proteins (Shkumatov, Chinnathambi et al. 2011). We further plan to investigate the time course of the observed event, minimal temperature difference required to observe compaction as well as pin down the region responsible for tau compaction. Moreover, we

want to investigate whether T-jump induced compaction is also characteristic for other IDPs of different length. Additionally we plan to get insights into functional role of a compact state of tau by performing kinetics experiments.

# Appendix A: supporting documents for subchapter 4.5

## Table A-1: CLANS clusters of sequence similar proteins in published tardigrade sequences

| Number/ color | Cluster description | Sequences/percentage[2] |
|---|---|---|
| 1 • | rRNA | 469 (4.35%) |
| 2 • | Cytochrome c oxidase like (subunit 1, EC 1.9.3.1) | 425 (3.94%) |
| 3 • | uncharacterized protein U88/glycosyltransferase 8 family | 302 (2.80%) |
| 4 • | rRNA | 282 (2.61%) |
| 5 • | Proteins containing a Chitin binding domain | 191 (1.77%) |
| 6 • | Proteins containing an IBR/Neuroparsin/DUF1096 domain | 189 (1.75%) |
| 7 • | Fatty-acid binding protein (FABP) family | 127 (1.18%) |
| 8 • | TSP[1] remote homology to Sericin 1 | 126 (1.17%) |
| 9 • | Proteins containing a RNA polymerase Rpb3/Rpb11 dimerisation domain | 92 (0.85%) |
| 10 • | Metallothionein superfamily (Type 15 family./Thioredoxin like) | 84 (0.78%) |
| 11 • | rRNA | 83 (0.73%) |
| 12 • | GTP-binding elongation factor family. EF-Tu/EF-1A subfamily | 79 (0.72%) |
| 13 • | GST superfamily. Sigma family | 78 (0.70%) |
| 14 • | Ubiquitin family | 75 (0.69%) |
| 15 • | Cathepsin family (EC 3.4.22.-) | 74 (0.67%) |
| 16 • | Carboxypeptidase A inhibitor like | 72 (0.64%) |
| 17 • | Trichohyalin/Translation initiation factor like | 69 (0.60%) |
| 18 • | TSP[1] | 65 (0.57%) |
| 19 • | TSP[1] | 61 (0.56%) |
| 20 • | RNA/DNA-binding proteins | 60 (0.55%) |
| 21 • | Apolipoprotein D like | 59 (0.50%) |
| 22 • | Histidine-rich glycoprotein like | 54 (0.49%) |
| 23 • | small heat shock protein (HSP20) family | 53 (0.47%) |
| 24 • | Diapause-specific proteins | 51 (0.44%) |
| 25 • | cGMP-specific 3′, 5′-cyclic phosphodieesterase / Putative surface protein bspA-like | 47 (0.42%) |
| 26 • | 26S proteasome (BOP1NT (NUC169) domain or 26S proteasome subunit RPN7) | 45 (0.42%) |
| 27 • | Sequestosome-1 like | 45 (0.41%) |
| 28 • | small GTPase superfamily | 44 (0.39%) |
| 29 • | Protein licA like | 42 (0.38%) |
| 30 • | TSP[1] | 41 (0.35%) |
| 31 • | Fatty-acid binding protein (FABP) family | 38 (0.34%) |
| 32 • | Ribosomal protein L41 like | 37 (0.33%) |
| 33 • | TSP[1] | 36 (0.33%) |
| 34 • | Protein IWS1 homolog/Neuraminidase like | 36 (0.32%) |
| 35 • | TSP[1] | 34 (0.30%) |
| 36 • | Histidine-rich glycoprotein like | 32 (0.30%) |
| 37 • | TSP[1] | 32 (0.29%) |
| 38 • | LEA type 1 family proteins | 31 (0.28%) |
| 39 • | Muscle LIM proteins | 30 (0.27%) |
| 40 • | Entericidin EcnA/B family | 29 (0.27%) |
| 41 • | Integrin, beta chain like | 29 (0.27%) |
| 42 • | TSP[1] | 29 (0.24%) |
| 43 • | ATP synthase subunit A like | 26 (0.24%) |
| 44 • | Plasma membrane proteolipid 3 like | 26 (0.23%) |
| 45 • | Actin family | 25 (0.23%) |
| 46 • | Proteins containing a CD80-like C2-set immunoglobulin domain | 29 (0.23%) |
| 47 • | Myosin light chain like proteins | 25 (0.23%) |
| 48 • | Zinc metalloproteinase nas-Family (EC 3.4.24.21) | 25 (0.22%) |
| 49 • | Protein Wnt-4 like | 24 (0.21%) |
| 50 • | GABA(A) receptor-associated protein-like 1/2 | 23 (0.21%) |
| 51 • | TSP[1] | 23 (0.20%) |
| 52 • | CUB and sushi domain-containing proteins | 22 (0.19%) |
| 53 • | NADH dehydrogenase subunit 2 like (EC 1.6.5.3) | 21 (0.19%) |
| 54 • | Eukaryotic translation initiation factor 4E-binding protein 2 like | 21 (0.19%) |
| 55 • | TSP[1] | 21 (0.19%) |
| 56 • | RNA polymerase II subunit B1 like (EC 2.7.7.6) | 21 (0.19%) |
| 57 • | short chain dehydrogenase like | 21 (0.19%) |
| 58 • | Niemann Pick type C2 protein homolog | 20 (0.19%) |

Tardigrade specific proteins; [2] There was a total of 10787 sequences, percentage of this total is given in ackets.

# Appendix B: supporting documents for subchapter 5.1

## Supplementary figures:

**Figure B-1. Quality of the electron density maps**

A. Unbiased electron density for Flt3D4. The electron density is contoured at 1σ and was obtained after implementation of phase improvement protocols based on a partial model of the complex consisting of only FL and Flt3D3.The final model for domains 3 and 4 in one of the receptor chains in the Flt3D1-D4-FL complex structure is shown in ribbon representation. N-linked glycans are shown in stick representation (magenta). This electron density map was obtained by applying NCS-averaging and solvent flattening protocols as implemented in PARROT1, and proved to be crucial early in the structure determination process providing the complete electron density trace for Flt3D4, including clear electron density for Nlinked glycans.

B. Quality of the electron density map to 4.3 Å resolution. Stereo diagram illustrating the quality of the final 2Fo-Fc electron density map to 4.3 Å resolution (contoured at 1σ) for the Flt3D1-D4:FL complex. The figure is centered on the Flt3D2-D3 interface and junction, with the final model for Flt3D2 (left) and Flt3D3 (right) displayed in ribbon representation (blue). The N-linked NAG glycan residue modeled at Asn306 is shown in sticks (magenta).

1. Cowtan, K. Recent developments in classical density modification. *Acta Crystallogr D Biol Crystallogr* 66, 470-478 (2010).

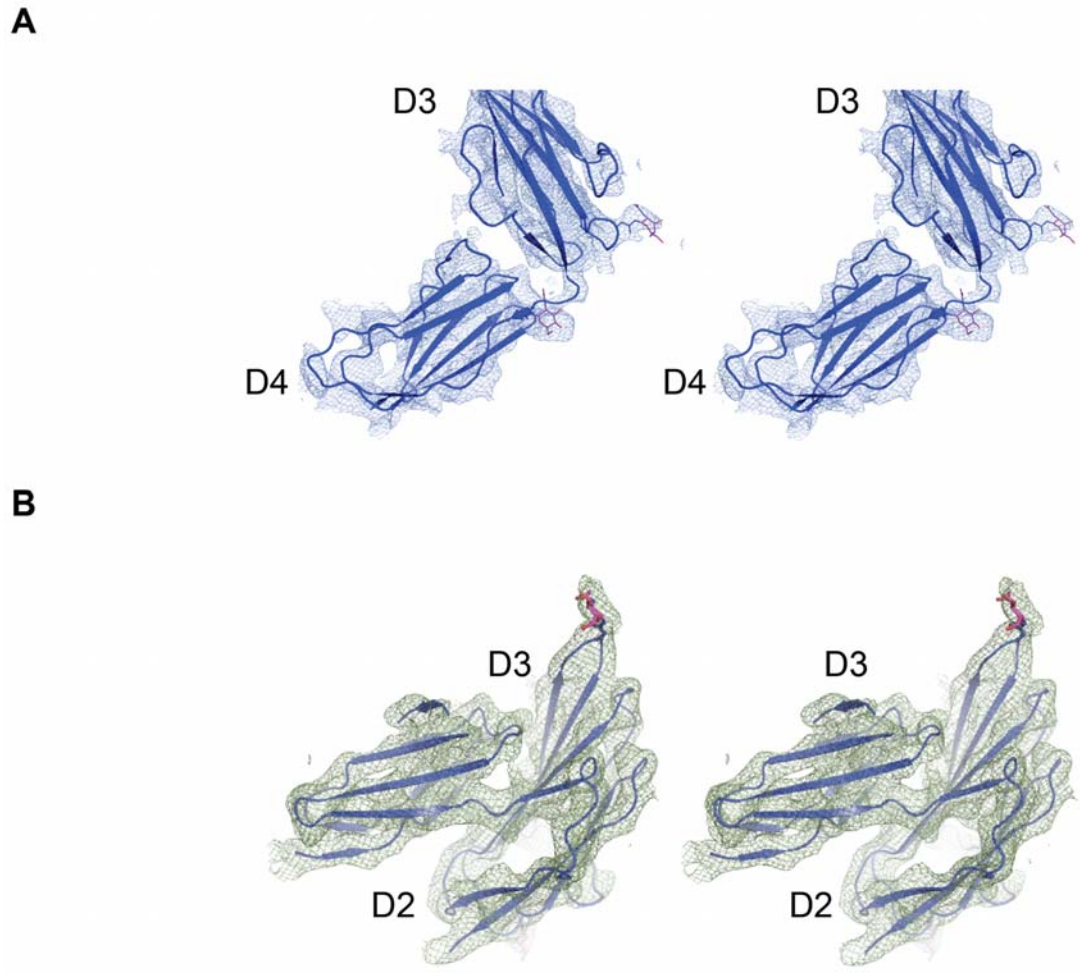
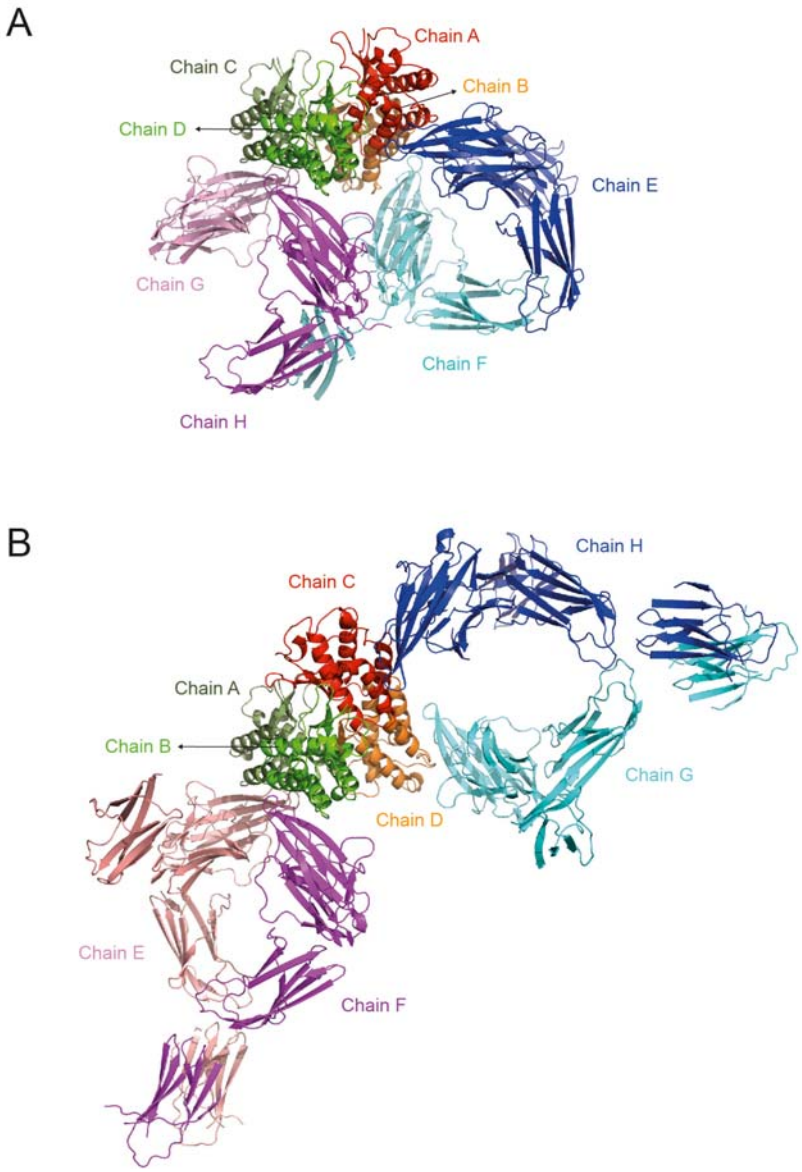


**Figure B-2. A. The asymmetric unit of Flt3D1-D4:FL complex crystals.** The Flt3D1-D4:FL complex crystallized in spacegroup P21 with two complexes in the asymmetric unit (asu). The two helical ligands in the different complexes (chains A-B and chains C-D) make extensive interactions in the asu. The receptor chains are labeled E, F, H and G. No density was visible for domains D1 of receptor chains G and H. D4 of chain G was also not modelled because of its weak density.




**B. The asymmetric unit of Flt3D1-D5:FL complex crystals.** Similar to the Flt3D1-D4:FL complex, the Flt3D1-D5:FL complex crystallized in spacegroup P21 with two complexes in the assymetric unit (asu). The contacts between the two complexes are entirely mediated by the two ligands (chains A-B and chains C-D). The Flt3 receptor chains are labeled E, F, H and G. The structure was refined by rigid-body refinement in autoBuster 2.8 using the FL protomers (residues 3-132), Flt3D1 (residues 79 - 161), Flt3D2-D3 (residues 167-345), Flt3D4 (residues 348-434) and Flt3D5 (residues 437-529) as rigid bodies. D1 of chain F was not modelled because of its weak density.

**Figure B-3. Interspecies comparison of FL sequences.** Sequence numbering and secondary structure assignment are according to the crystal structure of human Flt3 ligand (PDB 1ETE). Strictly conserved residues in the included FL sequences are highlighted in blue. The highly conserved N-terminal receptor-binding loop (residues 6-14) is highlighted in a red box. Other strictly conserved residues include the cysteines which form the disulfide bridges, residues Glu78 and Phe81 which stabilize the N-terminal loop, and residues Leu27 and Tyr30 at the dimer interface. The sequences were retrieved from the NCBI and Ensembl databases: *Homo sapiens* (NP_001450.2), *Mus musculus* (NP_038548.3), *Rattus norvegicus* (XP_002725623.1), *Papio cynocephalus* (AAO72538.1), *Felis catus* (NP_001009842.1), *Ailuropoda melanoleuca* (XP_002917887.1), *Canis lupus familiaris* (NP_001003350.1), *Pteropus vampyrus* (ENSPVAT00000010957), *Ovis aries* (NP_001072128.1), *Bos taurus* (NP_851373.1), *Sus scrofa* (ACZ63257.1), *Sorex araneus* (ENSSARP00000002887), *Cavia porcellus* (ENSCPOP00000020385), *Monodelphis domestica* (XP_001379894), *Xenopus tropicalis* (XP_002938571.1).

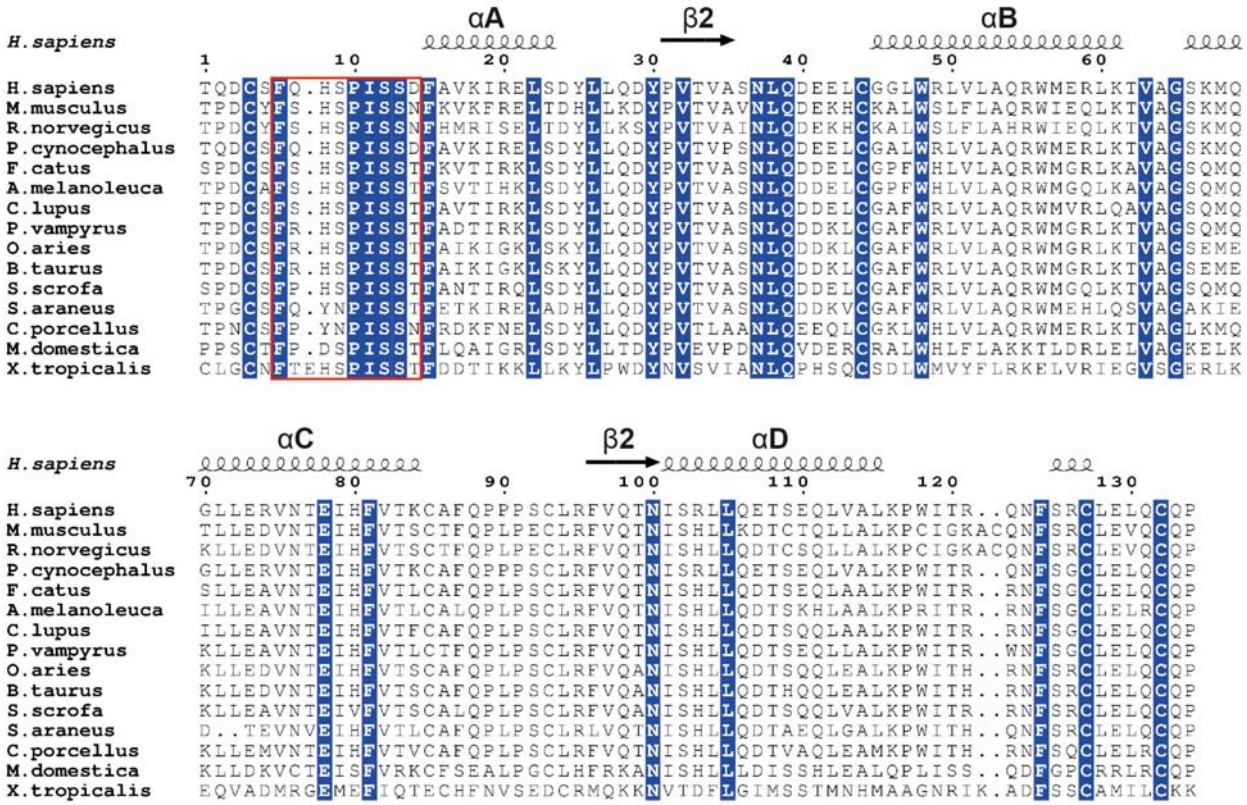
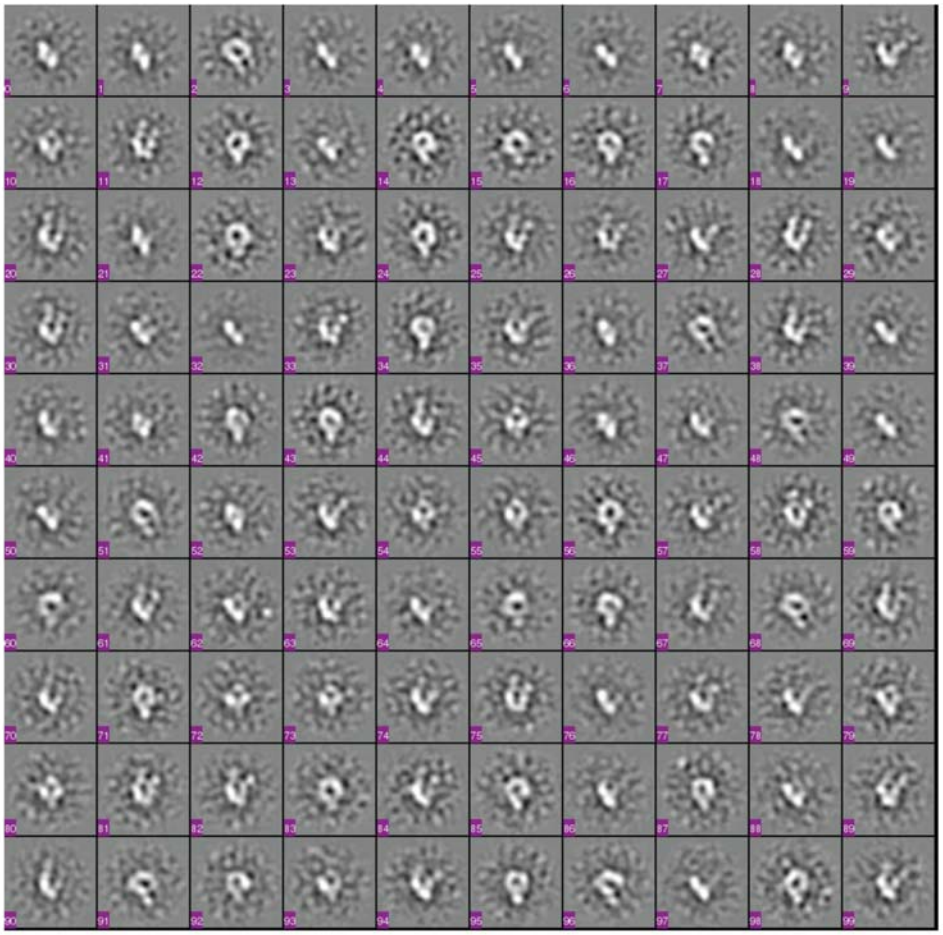


**Figure B-4. Negative-stain electron microscopy and SAXS analysis of the Flt3D1-D5-FL complex.** A. The displayed gallery of 100 class averages of the Flt3D1-D5:FL complex allows to recognize features corresponding to projections of the crystal structure at different orientations, notably the slightly open horseshoe ring structure with well-defined individual extracellular domains. B. The crystal structure of the Flt3D1-5-FL complex was refined as a rigid-body model against the experimental scattering curve obtained by SAXS. Fitting of the theoretical scattering curve calculated from the refined model (inset) to the experimental scattering curve shows a good agreement ($\chi 2=2.5$).

A



B

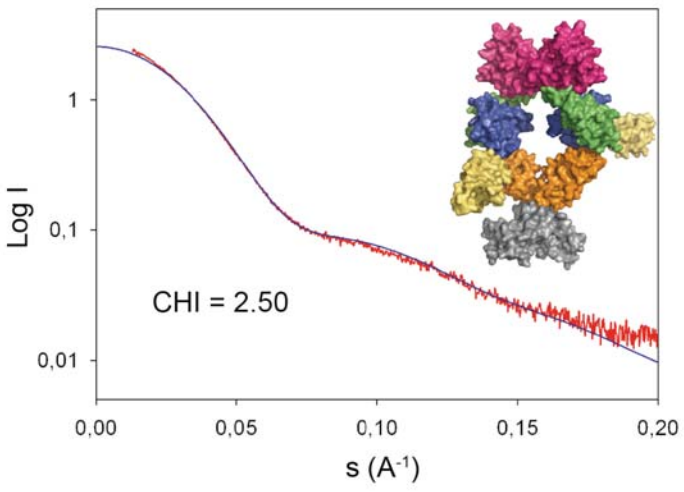
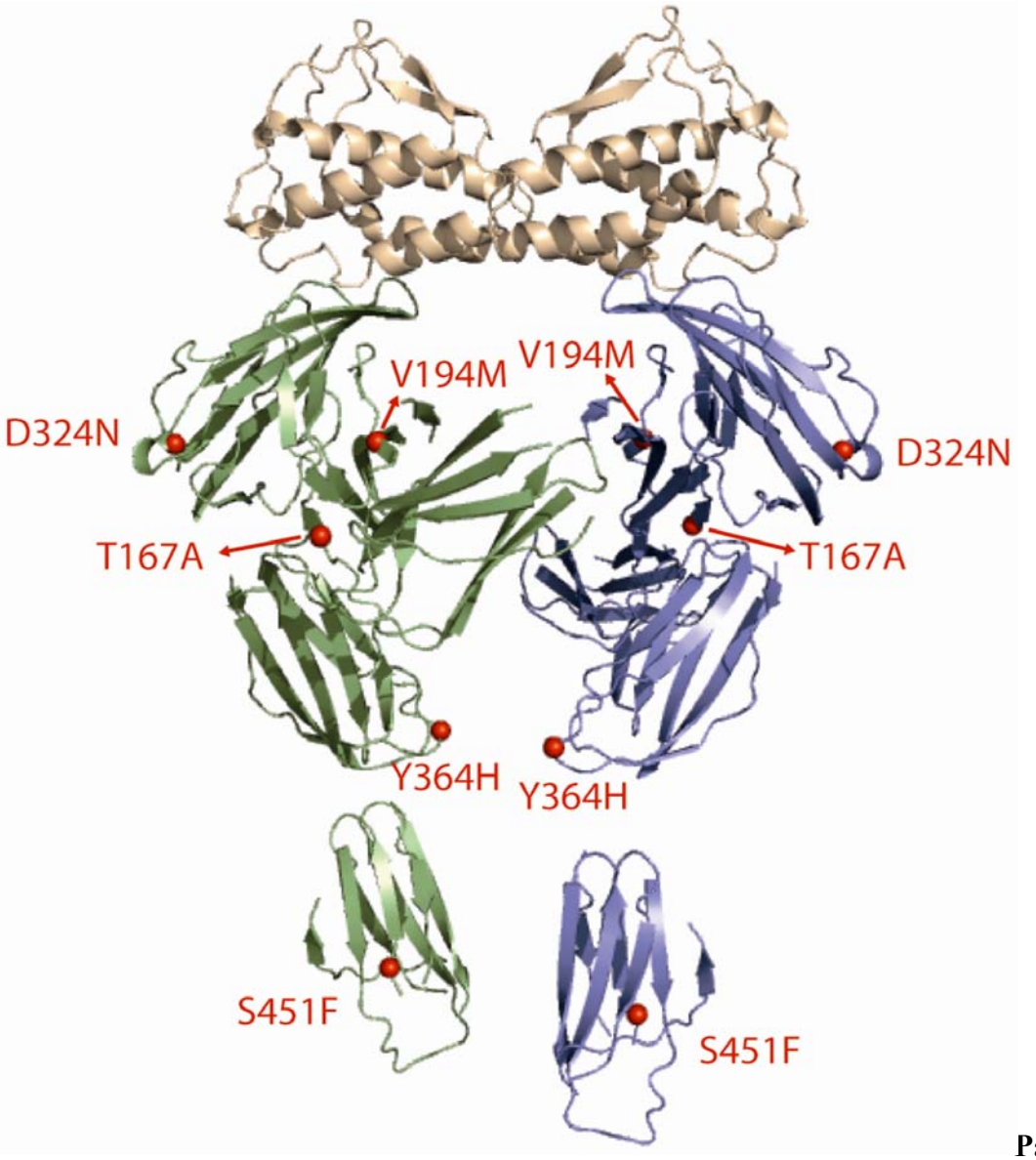


**Figure B-5. Mapping of non-synonymous sequence variants identified in the Flt3 ectodomain of AML patients.**

While the majority of oncogenic alterations in the *Flt3* gene are located in the JM and TKD regions, several mutations in the extracellular domains have recently been indentified in AML patients1,2. Expression of Flt3 carrying a mutation at position 451 (S451F) in BaF3 cells resulted in cytokine-independent proliferation and constitutive Flt3 autophosphorylation, demonstrating the oncogenic potential of this sequence variant. S451 is located at the solvent exposed site of strand *B* in the membrane proximal domain 5. Although the D324N variant did not result in ligand-independent activation it is associated with a higher risk of myeloid leukemias1,2. D324 is located in the EF-loop of domain 3. The possible role for all other sequence variants (T167A, V194M, Y364H) in leukemogenesis has not yet been demonstrated.

1. Frohling S*, et al.* (2007) Identification of driver and passenger mutations of FLT3 by highthroughput DNA sequence analysis and functional assessment of candidate alleles. Cancer Cell 12(6):501-513.

2. Schnittger S, et al. (2006) D324N single-nucleotide polymorphism in the FLT3 gene is associated with higher risk of myeloid leukemias. Genes Chromosomes Cancer 45(4):332-337.



## Supplementary table:

### Table B-1. Mapping of disulfide bridges and N-linked glycosylation sites in the Flt3 ectodomain by mass spectrometry.

***Asp-N peptides***

| Measured Mass | calculated mass | sequence positions | remarks |
|---|---|---|---|
| 2013.30 | 2013.09 | 248 – 265* | No glycosylation at #250 |
| 2081.56 | 2081.28 | 96 – 118 | Glycosylation at #100, disulfide bridge between #103 and #114 |
| 3486.61 | 3486.78 | 29 – 43<br>62 – 76* | Glycosylation at #43, disulfide bridge between #35 and #65 |
| 4156.83 | 4156.11 | 29 – 43<br>62 – 83 | Glycosylation at #43, disulfide bridge between #35 and #65 |

***Glu-C peptides***

| Measured Mass | calculated mass | sequence positions | remarks |
|---|---|---|---|
| 1499.56 | 1499.61 | 196 – 204<br>205 – 208 | Disulfide bridge between #199 and #206 |
| 1628.60 | 1628.65 | 196 – 204<br>205 – 209* | Disulfide bridge between #199 and #206 |
| 1573.53 | 1573.72 | 347 – 357 | Glycosylation at #351 and #354 |
| 1777.52 | 1777.78 | 347 – 360* | Glycosylation at #351 or #354 |
| 1980.58 | 1980.86 | 347 – 360* | Glycosylation at #351 and #354 |
| 4018.67 | 4018.32** | 25 – 44*<br>63 – 77* | Glycosylation at #43, disulfide bridge between #35 and #65 |
| (4134.96) | 4133.35** | 25 – 44*<br>62 – 77 | Glycosylation at #43, disulfide bridge between #35 and #65 |
| (4263.21) | 4262.39** | 24 – 44*<br>62 – 77 | Glycosylation at #43, disulfide bridge between #35 and #65 |
| 5559.91 | 5560.18** | 25 – 58<br>62 - 77 | Glycosylation at #43, disulfide bridge between #35 and #65 |

***Tryptic peptides***

| Measured Mass | calculated mass | sequence positions | remarks |
|---|---|---|---|
| 1735.62 | 1735.76 | 381 – 387<br>389 – 395 | Disulfide bridge between #381 and #392 |
| 1775.74 | 1775.71 | 323 – 334<br>272 - 273 | Glycosylation at #323, disulfide bridge between #330 and #272 |
| 2187,03 | 1983,98 | 491 – 508 | Glycosylation at #502 |
| 2214.78 | 2214.95 | 317 – 307 | Glycosylation at #306 *** |
| 2485,03 | 2281,98 | 467 – 485 | Glycosylation at #473 |
| 3542.46 | 3542.78 | 133 – 161 | Glycosylation at #151 |
| 3811.30 | 3811.62 | 348 – 372<br>406 – 410 | Glycosylation at #351 or #354, disulfide bridge between #368 and #407 |
| 4014.43 | 4014.70 | 348 – 372<br>406 – 410 | Glycosylation at #351 and #354, disulfide bridge between #368 and #407 |
| (5332.29) | 5333.06** | 176 – 215 *<br>231 – 234<br>240 – 243 | Disulfide bridges between #184 – #231 and #232 - #241 (from X-ray data) |
| 6614.06 | 6614.55** | 176 – 215*<br>220 – 234*<br>240 – 243 | Disulfide bridges between #184 – #231 and #232 - #241 (from X-ray data) |
| 8998.44 | 8998.13** | 35 – 41<br>50 – 108<br>109 – 123 | Glycosylation at 100 ***, disulfide bridges between #35 - #65 and #103 - #114 |

* Peptide containing in-complete or on-specific cleavage

** Averaged values

*** contains also very small amount of fucose (+ 146 Da)

# Appendix C: supporting documents for subchapter 5.3

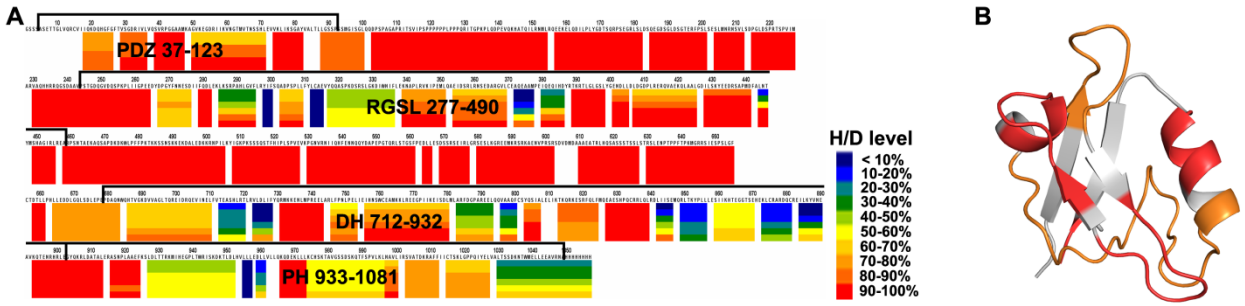



Figure C-1. (A) Map of solvent accessibility of PRG$^{37\text{-}1081}$. Each color block represents a deuteration level of a peptide with hydrogen/deuterium exchange quenched after 30, 100, 300, 1000, 3000, or 10,000 sec. Positions of folded domains are marked with brackets. (B) Crystal structure of a PDZ domain (PRG 37-123) color-coded according to the level of hydrogen/deuterium exchange.

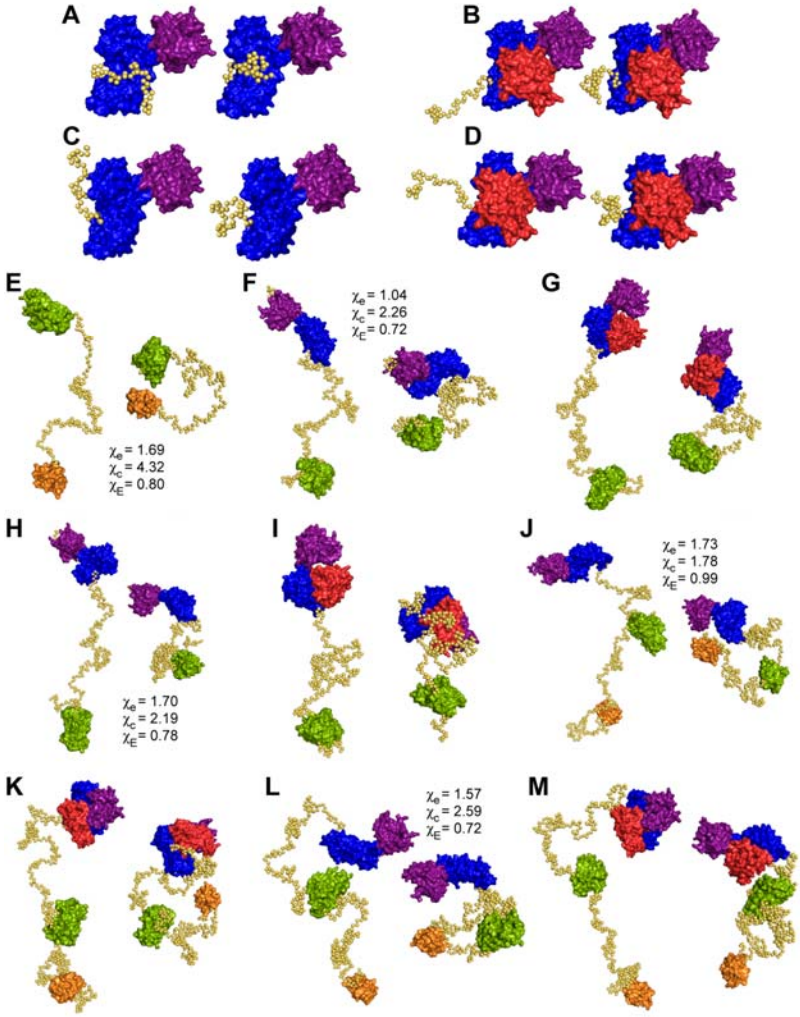



Figure C-2. The ensemble optimization method models of (A) PRG 672-1081, (B) PRG 672-1081/RhoA, (C) PRG 672-1081 4R, (D) PRG 672-1081 4R/RhoA, (E) PRG 37-490, (F) PRG 277-1081, (G) PRG 277-1081/RhoA, (H) PRG 277-1081 4R, (I) PRG 277-1081 4R/RhoA, (J) PRG 37-1081, (K) PRG 37-1081/RhoA, (L) PRG 37-1081 4R, and (M) PRG 37-1081 4R/RhoA. $\chi_e$, $\chi_c$, and $\chi_E$ are discrepancies to the extended, compact, and ensemble of models, respectively.

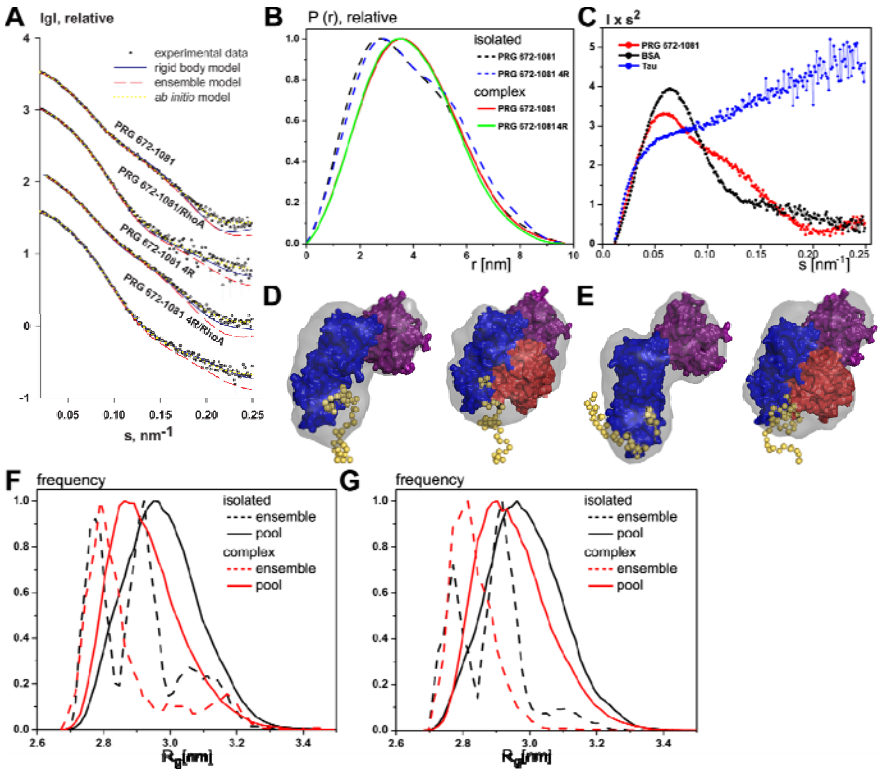
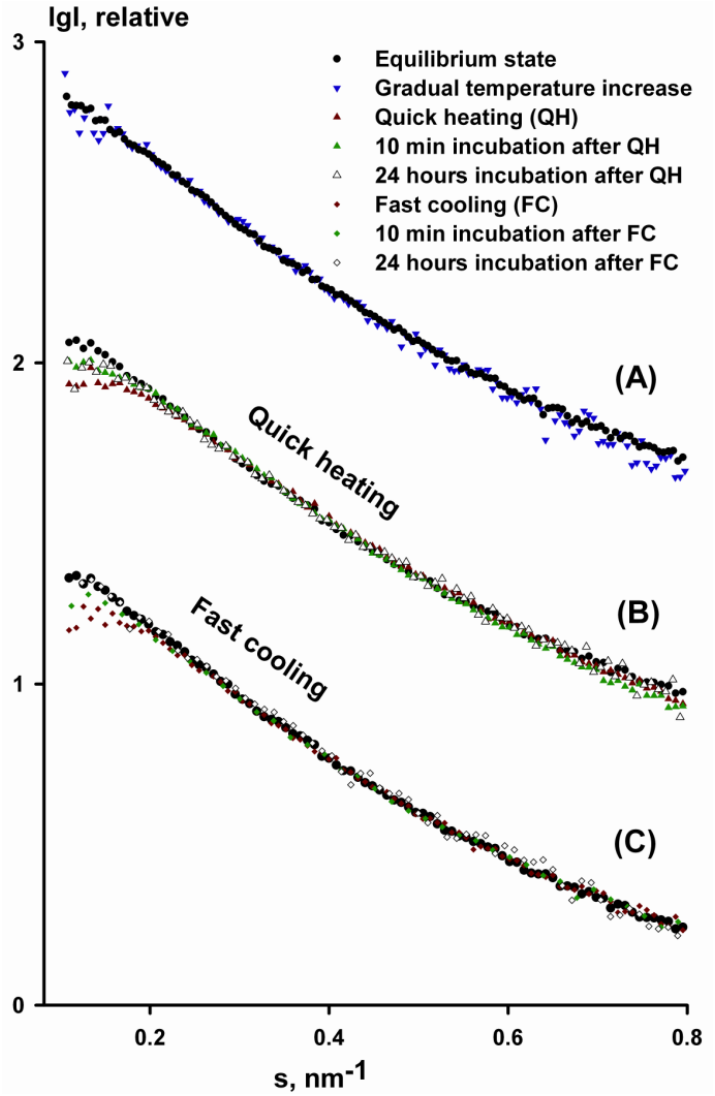
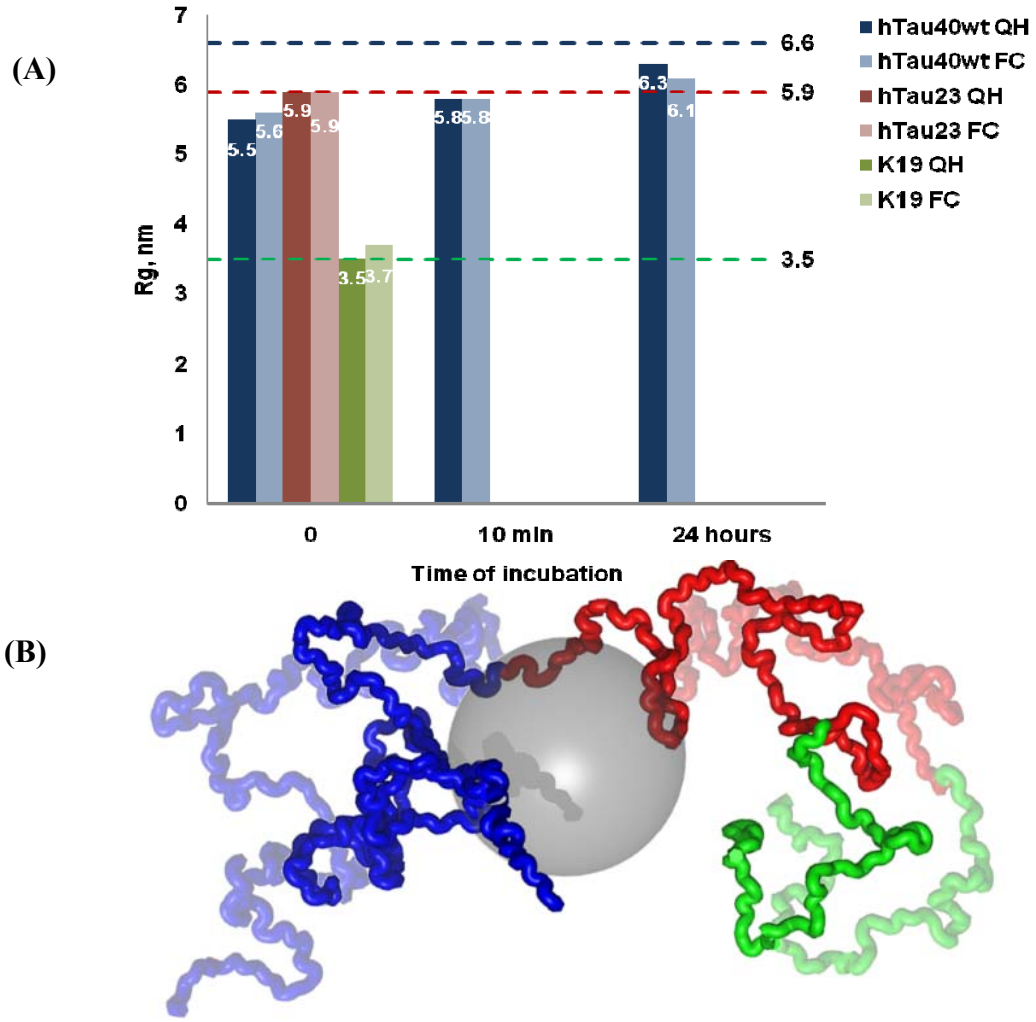



Figure C-3. Summary of SAXS data and models for PRG 672-1081 constructs. (A) Scattering profiles for PRG 672-1081 fragments. Experimental data, fit to *ab initio*, rigid body, and EOM models are shown as *open circles*, *yellow dotted line*, *blue solid line*, *and red dashed line*, respectively. Plots display the logarithm of the scattering intensity as a function of momentum transfer $s = 4\pi\sin(\theta)/\lambda$, where $2\theta$ is the scattering angle and $\lambda$ is the X-ray wavelength. Experimental SAXS profiles were appropriately displaced along the logarithmic axis for better visualization and overlaid with corresponding fits. (B) Distance distribution function, *P(r)*, of isolated PRG 672-1081 and PRG 672-1081 4R (*black and blue dashed lines*, respectively), and PRG 672-1081/RhoA complex and PRG 672-1081 4R/RhoA (*red and green solid lines*, respectively). (C) Comparison of Kratky plots of PRG 672-1081 (*red circle*), with bovine serum albumin (*black circle*), and protein Tau (*blue circle*). (D) Rigid body models (BUNCH) superposed onto the *ab initio* models (GASBOR) of isolated PRG 672-1081 (*left*), and PRG 672-1081/RhoA complex (*right*). (E) Rigid body models (BUNCH) superposed onto the *ab initio* models (GASBOR) of isolated PRG 672-1081 4R (*left*), and PRG 672-1081 4R/RhoA complex (*right*). (F) Radius of gyration distribution for PRG 672-1081. Distributions for pools of models are shown as *black* and *red solid lines*, for the isolated protein and the RhoA complex, respectively. Distributions for selected ensemble of models are shown as *black* and *red dashed lines*, for the isolated protein and the RhoA complex, respectively. (G) Radius of gyration distribution for PRG 672-1081 4R. Distributions for pools of models are shown as *black* and *red solid lines*, for the isolated protein and the RhoA complex, respectively. Distributions for selected ensemble of models are shown as *black* and *red dashed lines*, for the isolated protein and the RhoA complex, respectively.

# Appendix D: supporting documents for subchapter 5.4




**Supplementary figure D-1. Time course of the memory effect for hTau40wt after temperature-jump.** (A) Scattering pattern of hTau40wt after gradual temperature increase from 10 to 50°C (▼) shows no difference compared with that of the wild type (● ). (B) The scattering curves from the hTau40wt samples after quick heating incubated at 4 °C for 10 minutes (▲) and 24 hours (△), superimposed with the scattering from wild type (●) and hTau40wt after quick heating (▲). (C) Scattering curves from the samples after fast cooling incubated at 4 °C for 10 minutes (◆) and 24 hours (▽) superimposed with scattering from the wild type tau (●) and at fast cooling (◆) condition. All incubated samples were measured at 10°C. The compact state is preserved for least 3 hours of incubation, but after 24 hours the protein is nearly reverted back to the native state. Plots display the logarithm of the scattering intensity as a function of momentum transfer $s = 4\pi sin(\theta)/\lambda$, where $2\theta$ is the scattering angle and $\lambda$ is the X-ray wavelength. The scattering patterns for the three groups of samples were appropriately displaced along the logarithmic axis for better visualization. Every 2nd (A) and every 3rd (B, C) data point were merged for easier visualization.




**Supplementary figure D-2.** (A) Changes in radius of gyration for different tau samples over time. The dashed lines represent $R_g$ (nm) values for the tau constructs (dark blue – hTau40wt, red – hTau23, green – K19) measured at 10°C. Blue columns represent hTau40wt under non-equilibrium (dark blue - Quick Heating; light blue - Fast Cooling) conditions measured before and after incubation. The $R_g$ values of the samples incubated after temperature jump for 24 hours is similar to that of the wild type. Red columns represent annealed and quenched forms of hTau23 (dark red – Quick Heating, light red – Fast Cooling). Green columns represent annealed and quenched forms of K19 (green – Quick Heating; light green – Fast Cooling). The short constructs (hTau23 and K19), as opposed to the full length ones, do not change their overall size under non-equilibrium temperature conditions. (B) Hypothetical model of tau conformation before (pale colors) and after compaction induced by a temperature jump (bold colors). The N-terminus (residues 1-243) is colored blue, the repeat domain (residues 244-368) red, and the C-terminus (residues 369-441) green. The extent of compaction is exaggerated for better visibility. The gray sphere represents the Stokes radius of a well-folded protein (α-amylase) of the same chain length as htau40 (see Mylonas et al., 2007), illustrating that tau occupies a larger volume in space, both in the normal and compacted states.

# References

"http://ecole.modelisation.free.fr/modes.html."

"http://pilatus.web.psi.ch/pilatus.htm."

Aittaleb, M., C. A. Boguth, et al. (2010). "Structure and function of heterotrimeric G protein-regulated Rho guanine nucleotide exchange factors." Mol Pharmacol **77**(2): 111-125.

Akiyama, S., A. Nohara, et al. (2008). "Assembly and disassembly dynamics of the cyanobacterial periodosome." Mol Cell **29**(6): 703-716.

Akiyama, S., S. Takahashi, et al. (2002). "Conformational landscape of cytochrome c folding studied by microsecond-resolved small-angle x-ray scattering." Proc Natl Acad Sci U S A **99**(3): 1329-1334.

Altschul, S., T. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**: 3389 - 3402.

Alvarez-Ordonnez, A., A. Fernandez, et al. (2009). "Relationship between membrane fatty acid composition and heat resistance of acid and cold stressed Salmonella senftenberg CECT 4384." Food Microbiol **26**: 347 - 353.

Ammermann, D. (1967). "The cytology of parthenogenesis in the tardigrade Hypsibius dujardini." Chromosoma **23**(2): 203 - 213.

Andersen, C. B., T. Becker, et al. (2006). "Structure of eEF3 and the mechanism of transfer RNA release from the E-site." Nature **443**(7112): 663-668.

Bahar, I. and A. J. Rader (2005). "Coarse-grained normal mode analysis in structural biology." Curr Opin Struct Biol **15**(16143512): 586-592.

Baumann, H. (1922). "Die Anabiose der Tardigraden." Zool Jahrb **45**: 501 - 556.

Baumann, H. (1927). "Bemerkungen zur Anabiose von Tardigraden." Zool Anz **72**: 175 - 179.

Bavan, S., V. Straub, et al. (2009). "A P2X receptor from the tardigrade species Hypsibius dujardini with fast kinetics and sensitivity to zinc and copper." BMC Evol Biol **9**: 17.

Baxevanis, A. (2006). "Searching the NCBI databases using Entrez." Curr Protoc Hum Genet **Chapter 6**.

Berman, H., K. Henrick, et al. (2003). "Announcing the worldwide Protein Data Bank." Nat Struct Biol **10**(12): 980.

Bernado, P. (2010). "Effect of interdomain dynamics on the structure determination of modular proteins by small-angle scattering." Eur Biophys J **39**(5): 769-780.

Bernado, P., K. Modig, et al. (2010). "Structure and Dynamics of Ribosomal Protein L12: An Ensemble Model Based on SAXS and NMR Relaxation." Biophys J **98**(10): 2374-2382.

Bernado, P., E. Mylonas, et al. (2007). "Structural characterization of flexible proteins using small-angle X-ray scattering." J Am Chem Soc **129**(17): 5656-5664.

Bernado, P. and D. I. Svergun (2008). "New Perspectives in Small-Angle Scattering to Study Unstructured States of Proteins." In "Protein Misfolding", edited by Cian B. O'Doherty and Adam C. Byrne. Nova Publishers.

Bernstein, F. C., T. F. Koetzle, et al. (1977). "The Protein Data Bank: a computer-based archival file for macromolecular structures." J Mol Biol **112**(3): 535-542.

Binder, L. I., A. L. Guillozet-Bongaarts, et al. (2005). "Tau, tangles, and Alzheimer's disease." Biochim Biophys Acta **1739**(2-3): 216-223.

Blanc, E., P. Roversi, et al. (2004). "Refinement of severely incomplete structures with maximum likelihood in BUSTER-TNT." Acta Crystallogr D Biol Crystallogr **60**(Pt 12 Pt 1): 2210-2221.

Boczkowska, M., G. Rebowski, et al. (2008). "X-ray scattering study of activated Arp2/3 complex with bound actin-WCA." Structure **16**(5): 695-704.

Bonneau, R., C. E. Strauss, et al. (2002). "De novo prediction of three-dimensional structures for major protein families." J Mol Biol **322**(1): 65-78.

Bou-Abdallah, F., S. Adinolfi, et al. (2004). "Iron binding and oxidation kinetics in frataxin CyaY of Escherichia coli." J Mol Biol **341**(2): 605-615.

Bradley, P., K. M. Misura, et al. (2005). "Toward high-resolution de novo structure prediction for small proteins." Science **309**(5742): 1868-1871.

Branda, S. S., P. Cavadini, et al. (1999). "Yeast and human frataxin are processed to mature form in two sequential steps by the mitochondrial processing peptidase." J Biol Chem **274**(32): 22763-22769.

Broudy, V. C., N. L. Lin, et al. (2001). "The fifth immunoglobulin-like domain of the Kit receptor is required for proteolytic cleavage from the cell surface." Cytokine **15**(4): 188-195.

Browne, J., K. Dolan, et al. (2004). "Dehydration-specific induction of hydrophilic protein genes in the anhydrobiotic nematode Aphelenchus avenae." Eukaryot Cell **3**: 966 - 975.

Bulteau, A. L., H. A. O'Neill, et al. (2004). "Frataxin acts as an iron chaperone protein to modulate mitochondrial aconitase activity." Science **305**(5681): 242-245.

Butner, K. A. and M. W. Kirschner (1991). "Tau protein binds to microtubules through a flexible array of distributed weak sites." J Cell Biol **115**(3): 717-730.

Campuzano, V., L. Montermini, et al. (1996). "Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion." Science **271**(5254): 1423-1427.

Cavadini, P., J. Adamec, et al. (2000). "Two-step processing of human frataxin by mitochondrial processing peptidase. Precursor and intermediate forms are cleaved at different rates." J Biol Chem **275**(52): 41469-41475.

Cavadini, P., H. A. O'Neill, et al. (2002). "Assembly and iron-binding properties of human frataxin, the protein deficient in Friedreich ataxia." Human Molecular Genetics **11**(3): 217-227.

Chasteen, N. D. and P. M. Harrison (1999). "Mineralization in ferritin: an efficient means of iron storage." J Struct Biol **126**(3): 182-194.

Cheetham, M. and A. Caplan (1998). "Structure, function and evolution of DnaJ: conservation and adaptation of chaperone function." Cell Stress Chaperones **3**: 28 - 36.

Chen, W., X. Ge, et al. (2009). "A gene catalogue for post-diapause development of an anhydrobiotic arthropod Artemia franciscana." BMC Genomics **10**: 52.

Chen, X., H. Liu, et al. (2008). "Structure of macrophage colony stimulating factor bound to FMS: diverse signaling assemblies of class III receptor tyrosine kinases." Proc Natl Acad Sci U S A **105**(47): 18267-18272.

Chen, Z., L. Guo, et al. (2011). "Modulation of a GEF switch: autoinhibition of the intrinsic guanine nucleotide exchange activity of p115-RhoGEF." Protein Sci **20**(21064165): 107-117.

Cole, C., J. D. Barber, et al. (2008). "The Jpred 3 secondary structure prediction server." Nucleic Acids Res **36**(18463136): 197-201.

Cowtan, K. (2010). "Recent developments in classical density modification." Acta Crystallogr D Biol Crystallogr **66**(Pt 4): 470-478.

Crowther, R. A., R. Henderson, et al. (1996). "MRC image processing programs." J Struct Biol **116**(1): 9-16.

Deane, C. M. and T. L. Blundell (2001). "CODA: a combined algorithm for predicting the structurally variable regions of protein models." Protein Sci **10**(3): 599-612.

Dehlin, M., M. Bokarewa, et al. (2008). "Intra-articular fms-like tyrosine kinase 3 ligand expression is a driving force in induction and progression of arthritis." PLoS ONE **3**(11): e3633.

DiMaio, F., M. D. Tyka, et al. (2009). "Refinement of Protein Structures into Low-Resolution Density Maps Using Rosetta." Journal of Molecular Biology **392**(1): 181-190.

Dittrich, M., I. Birschmann, et al. (2006). "Analysis of SAGE data in human platelets: features of the transcriptome in an anucleate cell." Thromb Haemost **95**: 643 - 651.

Dominguez, C., R. Boelens, et al. (2003). "HADDOCK: a protein-protein docking approach based on biochemical or biophysical information." J Am Chem Soc **125**(7): 1731-1737.

Dong, J., C. M. McPherson, et al. (2002). "Flt-3 ligand: a potent dendritic cell stimulator and novel antitumor agent." Cancer Biol Ther **1**(5): 486-489.

Doniach, S. (2001). "Changes in biomolecular conformation seen by small angle X-ray scattering." Chem Rev **101**(6): 1763-1778.

Drubin, D. G. and M. W. Kirschner (1986). "Tau protein function in living cells." J Cell Biol **103**(6 Pt 2): 2739-2746.

Eidenschenk, C., K. Crozat, et al. (2010). "Flt3 permits survival during infection by rendering dendritic cells competent to activate NK cells." Proc Natl Acad Sci U S A **107**(20457904): 9759-9764.

Eklund, E. A. (2010). "Genomic analysis of acute myeloid leukemia: potential for new prognostic indicators." Curr Opin Hematol **17**(20075726): 75-78.

Engelman, D. M., P. B. Moore, et al. (1975). "Neutron scattering measurements of separation and shape of proteins in 30S ribosomal subunit of Escherichia coli: S2-S5, S5-S8, S3-S7." Proc Natl Acad Sci U S A **72**(10): 3888-3892.

Erickson, J. W. and R. A. Cerione (2004). "Structural elements, mechanism, and evolutionary convergence of Rho protein-guanine nucleotide exchange factor complexes." Biochemistry **43**(14744125): 837-842.

Etienne-Manneville, S. and A. Hall (2002). "Rho GTPases in cell biology." Nature **420**(12478284): 629-635.

Fagegaltier, D., A. Lescure, et al. (2000). "Structural analysis of new local features in SECIS RNA hairpins." Nucleic Acids Res **28**: 2679 - 2689.

Feigin, L. A. and D. I. Svergun (1987). Structure analysis by small-angle x-ray and neutron scattering. New York, Plenum Press.

Finger, C., C. Escher, et al. (2009). "The single transmembrane domains of human receptor tyrosine kinases encode self-interactions." Sci Signal **2**(89): ra56.

Finn, R., J. Tate, et al. (2008). "The Pfam protein families database." Nucleic Acids Res **36**: D281 - D288.

Finocchiaro, G., G. Baio, et al. (1988). "Glucose metabolism alterations in Friedreich's ataxia." Neurology **38**(8): 1292-1296.

Fiser, A. and A. Sali (2003). "ModLoop: automated modeling of loops in protein structures." Bioinformatics **19**(18): 2500-2501.

Forster, F., C. Liang, et al. (2009). "Tardigrade workbench: comparing stress-related proteins, sequence-similar and functional protein clusters as well as RNA elements in tardigrades." BMC Genomics **10**: 469.

Franke, D. and D. I. Svergun (2009). "DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering." J. Appl. Cryst. **42**: 342-346

Frickey, T. and A. Lupas (2004). "CLANS: a Java application for visualizing protein families based on pairwise similarity." Bioinformatics **20**: 3702 - 3704.

Frohling, S., C. Scholl, et al. (2007). "Identification of driver and passenger mutations of FLT3 by high-throughput DNA sequence analysis and functional assessment of candidate alleles." Cancer Cell **12**(6): 501-513.

Fukuhara, S., H. Chikumi, et al. (2001). "RGS-containing RhoGEFs: the missing link between transforming G proteins and Rho?" Oncogene **20**(13): 1661-1668.

Fukuhara, S., C. Murga, et al. (1999). "A novel PDZ domain containing guanine nucleotide exchange factor links heterotrimeric G proteins to Rho." J Biol Chem **274**(10026210): 5868-5879.

Gabriel, W., R. McNuff, et al. (2007). "The tardigrade Hypsibius dujardini, a new model for studying the evolution of development." Dev Biol **312**: 545 - 559.

Gakh, O., J. Adamec, et al. (2002). "Physical evidence that yeast frataxin is an iron storage protein." Biochemistry **41**(21): 6798-6804.

Gakh, O., S. Park, et al. (2006). "Mitochondrial iron detoxification is a primary function of frataxin that limits oxidative damage and preserves cell longevity." Hum Mol Genet **15**(3): 467-479.

Gast, K., H. Damaschun, et al. (1995). "Prothymosin alpha: a biologically active protein with random coil conformation." Biochemistry **34**(40): 13211-13218.

Gazi, A. D., M. Bastaki, et al. (2008). "Evidence for a coiled-coil interaction mode of disordered proteins from bacterial type III secretion systems." J Biol Chem **283**(49): 34062-34068.

Geoffroy, G., A. Barbeau, et al. (1976). "Clinical description and roentgenologic evaluation of patients with Friedreich's ataxia." Can J Neurol Sci **3**(4): 279-286.

Glatter, O. and O. Kratky (1982). Small Angle X-ray Scattering. London, Academic Press.

Goedert, M., R. Jakes, et al. (1996). "Assembly of microtubule-associated protein tau into Alzheimer-like filaments induced by sulphated glycosaminoglycans." Nature **383**(6600): 550-553.

Gorba, C., O. Miyashita, et al. (2008). "Normal-mode flexible fitting of high-resolution structure of biological molecules toward one-dimensional low-resolution data." Biophys J **94**(5): 1589-1599.

Gorba, C., O. Miyashita, et al. (2008). "Normal-mode flexible fitting of high-resolution structure of biological molecules toward one-dimensional low-resolution data." Biophys J **94**(17993489): 1589-1599.

Goyal, K., L. Tisi, et al. (2003). "Transition from natively unfolded to folded state induced by desiccation in an anhydrobiotic nematode protein." J Biol Chem **278**: 12977 - 12984.

Graddis, T. J., K. Brasel, et al. (1998). "Structure-function analysis of FLT3 ligand-FLT3 receptor interactions using a rapid functional screen." J Biol Chem **273**(28): 17626-17633.

Grassot, J., M. Gouy, et al. (2006). "Origin and molecular evolution of receptor tyrosine kinases with immunoglobulin-like domains." Mol Biol Evol **23**(6): 1232-1241.

Griffith, J., J. Black, et al. (2004). "The structural basis for autoinhibition of FLT3 by the juxtamembrane domain." Mol Cell **13**(2): 169-178.

Guinier, A. (1939). "La diffraction des rayons X aux tres petits angles; application a l'etude de phenomenes ultramicroscopiques." Ann. Phys. (Paris) **12**: 161-237.

Guinier, A. and G. Fournet (1955). Small Angle Scattering of X-Rays. New York, Wiley.

Haass, C. and D. J. Selkoe (2007). "Soluble protein oligomers in neurodegeneration: lessons from the Alzheimer's amyloid beta-peptide." Nat Rev Mol Cell Biol **8**(2): 101-112.

Harding, A. E. and R. L. Hewer (1983). "The heart disease of Friedreich's ataxia: a clinical and electrocardiographic study of 115 patients, with an analysis of serial electrocardiographic changes in 30 cases." Q J Med **52**(208): 489-502.

Harpaz, Y. and C. Chothia (1994). "Many of the immunoglobulin superfamily domains in cell adhesion molecules and surface receptors belong to a new structural set which is close to that containing variable domains." J Mol Biol **238**(4): 528-539.

He, G., A. Ramachandran, et al. (2005). "Phosphorylation of phosphophoryn is crucial for its function as a mediator of biomineralization." J Biol Chem **280**(39): 33109-33114.

Hengherr, S., M. Worland, et al. (2009). "Freeze tolerance, supercooling points and ice formation: comparative studies on the subzero temperature survival of limno-terrestrial tardigrades." J Exp Biol **212**: 802 - 807.

Hengherr, S., M. Worland, et al. (2009). "High-Temperature Tolerance in Anhydrobiotic Tardigrades Is Limited by Glass Transition." Physiol Biochem Zool **82**(6): 749 - 755.

Hong-Bo, S., L. Zong-Suo, et al. (2005). "LEA proteins in higher plants: structure, function, gene expression and regulation." Colloids Surf B Biointerfaces **45**: 131 - 135.

Hong, D. P., A. L. Fink, et al. (2008). "Structural characteristics of alpha-synuclein oligomers stabilized by the flavonoid baicalein." J Mol Biol **383**(1): 214-223.

Horikawa, D., T. Sakashita, et al. (2006). "Radiation tolerance in the tardigrade Milnesium tardigradum." Int J Radiat Biol **82**: 843 - 848.

Huyton, T., J. G. Zhang, et al. (2007). "An unusual cytokine:Ig-domain interaction revealed in the crystal structure of leukemia inhibitory factor (LIF) in complex with the LIF receptor." Proc Natl Acad Sci U S A **104**(31): 12737-12742.

Ibel, K. and H. B. Stuhrmann (1975). "Comparison of neutron and X-ray scattering of dilute myoglobin solutions." J Mol Biol **93**(2): 255-265.

Jacques, D. A. and J. Trewhella (2010). "Small-angle scattering for structural biology--expanding the frontier while avoiding the pitfalls." Protein Sci **19**(4): 642-657.

Janin, J., R. P. Bahadur, et al. (2008). "Protein-protein interaction and quaternary structure." Q Rev Biophys **41**(2): 133-180.

Jeganathan, S., A. Hascher, et al. (2008). "Proline-directed pseudo-phosphorylation at AT8 and PHF1 epitopes induces a compaction of the paperclip folding of Tau and generates a pathological (MC-1) conformation." J Biol Chem **283**(46): 32066-32076.

Jeganathan, S., M. von Bergen, et al. (2006). "Global hairpin folding of tau in solution." Biochemistry **45**(7): 2283-2293.

Jicha, G. A., R. Bowser, et al. (1997). "Alz-50 and MC-1, a new monoclonal antibody raised to paired helical filaments, recognize conformational epitopes on recombinant tau." J Neurosci Res **48**(2): 128-132.

Jonsson, K., E. Rabbow, et al. (2008). "Tardigrades survive exposure to space in low Earth orbit." Curr Biol **18**: R729 - R731.

Jonsson, K. and R. Schill (2007). "Induction of Hsp70 by desiccation, ionising radiation and heat-shock in the eutardigrade Richtersius coronifer." Comp Biochem Physiol B Biochem Mol Biol **146**: 456 - 460.

Kaplan, J. (2002). "Mechanisms of cellular iron acquisition: another iron in the fire." Cell **111**(5): 603-606.

Karlberg, T., U. Schagerlof, et al. (2006). "The structures of frataxin oligomers reveal the mechanism for the delivery and detoxification of iron." Structure **14**(10): 1535-1546.

Keilin, D. (1959). "The Leeuwenhoek Lecture: The problem of anabiosis or latent life: History and current concept." Proc R Soc Lond B Biol Sci **150**: 149 - 191.

Kikushige, Y., G. Yoshimoto, et al. (2008). "Human Flt3 is expressed at the hematopoietic stem cell and the granulocyte/macrophage progenitor stages to maintain cell survival." J Immunol **180**(11): 7358-7367.

Kim, Y., R. Nachman, et al. (2008). "The pheromone biosynthesis activating neuropeptide (PBAN) receptor of Heliothis virescens: identification, functional expression, and structure-activity relationships of ligand analogs." Peptides **29**: 268 - 275.

Kinchin, I. and R. Dennis (1994). Portland Press London.

Kindler, T., D. B. Lipka, et al. (2010). "FLT3 as a therapeutic target in AML: still challenging after all these years." Blood **116**(20705759): 5089-5102.

Kiyoi, H., R. Ohno, et al. (2002). "Mechanism of constitutive activation of FLT3 with internal tandem duplication in the juxtamembrane domain." Oncogene **21**(16): 2555-2563.

Kjaergaard, M., A. B. Norholm, et al. (2010). "Temperature-dependent structural changes in intrinsically disordered proteins: formation of alpha-helices or loss of polyproline II?" Protein Sci **19**(8): 1555-1564.

Kleywegt, G. J. (1997). "Validation of protein models from Calpha coordinates alone." J Mol Biol **273**(2): 371-376.

Knight, S. A., N. B. Sepuri, et al. (1998). "Mt-Hsp70 homolog, Ssc2p, required for maturation of yeast frataxin and mitochondrial iron homeostasis." J Biol Chem **273**(29): 18389-18393.

Kobayashi, F., E. Maeta, et al. (2008). "Positive role of a wheat HvABI5 ortholog in abiotic stress response of seedlings." Physiol Plant **134**: 74 - 86.

Konarev, P. V., V. V. Volkov, et al. (2003). "PRIMUS - a Windows-PC based system for small-angle scattering data analysis." J. Appl. Crystallogr. **36**: 1277-1282.

Konno, T., N. Tanaka, M. Kataoka, E. Takano, and M. Maki. (1997). "A circular dichroism study of preferential hydration and alcohol effects on a denatured protein, pig calpastatin domain I. ." Biochim Biophys Acta **1342**: 73-82.

Koutnikova, H., V. Campuzano, et al. (1998). "Maturation of wild-type and mutated frataxin by the mitochondrial processing peptidase." Hum Mol Genet **7**(9): 1485-1489.

Kozin, M. B. and D. I. Svergun (2001). "Automated matching of high- and low-resolution structural models." J. Appl. Crystallogr. **34**: 33-41.

Krebs, W. G., V. Alexandrov, et al. (2002). "Normal mode analysis of macromolecular motions in a database framework: Developing mode concentration as a useful classifying statistic." Proteins-Structure Function and Genetics **48**(4): 682-695.

Krissinel, E. and K. Henrick (2007). "Inference of macromolecular assemblies from crystalline state." J Mol Biol **372**(3): 774-797.

Kristelly, R., G. Gao, et al. (2004). "Structural determinants of RhoA binding and nucleotide exchange in leukemia-associated Rho guanine-nucleotide exchange factor." J Biol Chem **279**(15331592): 47352-47362.

Lai, E. (2002). "Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation." Nat Genet **30**: 363 - 364.

Lai, E., C. Burks, et al. (1998). "The K box, a conserved 3' UTR sequence motif, negatively regulates accumulation of enhancer of split complex transcripts." Development **125**: 4077 - 4088.

Lai, E., B. Tam, et al. (2005). "Pervasive regulation of Drosophila Notch target genes by GY-box-, Brd-box-, and K-box-class microRNAs." Genes Dev **19**: 1067 - 1080.

Lavagna, C., S. Marchetto, et al. (1995). "Identification and characterization of a functional murine FLT3 isoform produced by exon skipping." J Biol Chem **270**(7): 3165-3171.

Layer, G., S. Ollagnier-de Choudens, et al. (2006). "Iron-sulfur cluster biosynthesis: characterization of Escherichia coli CYaY as an iron donor for the assembly of [2Fe-2S] clusters in the scaffold IscU." J Biol Chem **281**(24): 16256-16263.

Lemmon, M. A. and J. Schlessinger (2010). "Cell signaling by receptor tyrosine kinases." Cell **141**(20602996): 1117-1134.

Lesuisse, E., R. Santos, et al. (2003). "Iron use for haeme synthesis is under control of the yeast frataxin homologue (Yfh1)." Hum Mol Genet **12**(8): 879-889.

Letunic, I., T. Doerks, et al. (2009). "SMART 6: recent updates and new developments." Nucleic Acids Res **37**: D229 - D232.

Lin, H., E. Lee, et al. (2008). "Discovery of a cytokine and its receptor by functional screening of the extracellular proteome." Science **320**(5877): 807-811.

Lindahl, E., C. Azuara, et al. (2006). "NOMAD-Ref: visualization, deformation and refinement of macromolecular structures based on all-atom normal mode analysis." Nucleic Acids Research **34**: W52-W56.

Lindahl, E. and M. Delarue (2005). "Refinement of docked protein-ligand and protein-DNA structures using low frequency normal mode amplitude optimization." Nucleic Acids Research **33**(14): 4496-4506.

Liu, H., X. Chen, et al. (2007). "Structural basis for stem cell factor-KIT signaling and activation of class III receptor tyrosine kinases." EMBO J **26**(3): 891-901.

Liu, K. and M. C. Nussenzweig (2010). "Origin and development of dendritic cells." Immunol Rev **234**(1): 45-54.

Liu, K., G. D. Victora, et al. (2009). "In vivo analysis of dendritic cell development and homeostasis." Science **324**(5925): 392-397.

Lobanov, A., D. Hatfield, et al. (2009). "Eukaryotic selenoproteins and selenoproteomes." Biochim Biophys Acta.

Lyman, S. D., L. James, et al. (1993). "Characterization of the protein encoded by the flt3 (flk2) receptor-like tyrosine kinase gene." Oncogene **8**(4): 815-822.

Mandelkow, E., M. von Bergen, et al. (2007). "Structural principles of tau and the paired helical filaments of Alzheimer's disease." Brain Pathol **17**(1): 83-90.

March, P. (1992). "Membrane-associated GTPases in bacteria." Mol Microbiol **6**: 1253 - 1257.

Marcus, E. (1928). "Zur Okologie und Physiologie der Tardigraden." Zool Jahrb Abt Phys **44**: 323 - 370.

Marcus, E. and F. Dahl (1928). Urban & Fischer Bei Elsevier.

Margittai, M. and R. Langen (2004). "Template-assisted filament growth by parallel stacking of tau." Proc Natl Acad Sci U S A **101**(28): 10278-10283.

Maroc, N., R. Rottapel, et al. (1993). "Biochemical characterization and analysis of the transforming potential of the FLT3/FLK2 receptor tyrosine kinase." Oncogene **8**(4): 909-918.

Matsuo, T., H. Iwamoto, et al. (2010). "Monitoring the structural behavior of troponin and myoplasmic free Ca2+ concentration during twitch of frog skeletal muscle." Biophys J **99**(1): 193-200.

Mertens, H. D. and D. I. Svergun (2010). "Structural characterization of proteins and complexes using small-angle X-ray solution scattering." J Struct Biol **172**(1): 128-141.

Metcalf, D. (2008). "Hematopoietic cytokines." Blood **111**(18182579): 485-491.

Michalsky, E., A. Goede, et al. (2003). "Loops In Proteins (LIP)--a comprehensive loop database for homology modelling." Protein Eng **16**(12): 979-985.

Miyashita, O., C. Gorba, et al. (2010). "Structure modeling from small angle X-ray scattering data with elastic network normal mode analysis." J Struct Biol.

Moore, P. B. (1980). "Small-angle scattering: Information content and error analysis." J. Appl. Cryst. **13**: 168-175.

Morgan, H. P., C. Q. Schmidt, et al. (2011). "Structural basis for engagement by complement factor H of C3b on a self surface." Nat Struct Mol Biol.

Mounier, N. and A. Arrigo (2002). "Actin cytoskeleton and small heat shock proteins: how do they interact?" Cell Stress Chaperones **7**: 167 - 176.

Mukrasch, M. D., S. Bibow, et al. (2009). "Structural polymorphism of 441-residue tau at single residue resolution." PLoS Biol **7**(2): e34.

Munishkina, L. A., A. L. Fink, et al. (2004). "Conformational prerequisites for formation of amyloid fibrils from histones." J Mol Biol **342**(4): 1305-1324.

Murayama, K., M. Shirouzu, et al. (2007). "Crystal structure of the rac activator, Asef, reveals its autoinhibitory mechanism." J Biol Chem **282**(17190834): 4238-4242.

Mylonas, E., A. Hascher, et al. (2008). "Domain conformation of tau protein studied by solution small-angle X-ray scattering." Biochemistry **47**(39): 10345-10353.

Mylonas, E. and D. I. Svergun (2007). "Accuracy of molecular mass determination of proteins in solution by small-angle X-ray scattering." J. Appl. Cryst. **40**: s245-s249.

Nelson, D. (2002). "Current Status of the Tardigrada: Evolution and Ecology." Integr Comp Biol **42**: 652 - 659.

Neumann, S., A. Reuner, et al. (2009). "DNA damage in storage cells of anhydrobiotic tardigrades." Comp Biochem Physiol A Mol Integr Physiol **153**: 425 - 429.

Nichol, H., O. Gakh, et al. (2003). "Structure of frataxin iron cores: an X-ray absorption spectroscopic study." Biochemistry **42**(20): 5971-5976.

Niemann, H. H., M. V. Petoukhov, et al. (2008). "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.

Nuth, M., T. Yoon, et al. (2002). "Iron-sulfur cluster biosynthesis: characterization of iron nucleation sites for assembly of the [2Fe-2S]2+ cluster core in IscU proteins." J Am Chem Soc **124**(30): 8774-8775.

O'Neill, H. A., O. Gakh, et al. (2005). "Supramolecular assemblies of human frataxin are formed via subunit-subunit interactions mediated by a non-conserved amino-terminal region." J Mol Biol **345**(3): 433-439.

O'Neill, H. A., O. Gakh, et al. (2005). "Assembly of human frataxin is a mechanism for detoxifying redox-active iron." Biochemistry **44**(2): 537-545.

Oates, J., G. King, et al. (2010). "Strong oligomerization behavior of PDGFbeta receptor transmembrane domain and its regulation by the juxtamembrane regions." Biochim Biophys Acta **1798**(3): 605-615.

Oka, T., K. Inoue, et al. (2005). "Structural transition of bacteriorhodopsin is preceded by deprotonation of Schiff base: Microsecond time-resolved X-ray diffraction study of purple membrane." Biophysical Journal **88**(1): 436-442.

Onai, N., A. Obata-Onai, et al. (2007). "Identification of clonogenic common Flt3+M-CSFR+ plasmacytoid and conventional dendritic cell progenitors in mouse bone marrow." Nat Immunol **8**(11): 1207-1216.

Ostareck-Lederer, A., D. Ostareck, et al. (1998). "Cytoplasmic regulatory functions of the KH-domain proteins hnRNPs K and E1/E2." Trends Biochem Sci **23**: 409 - 411.

Ostareck-Lederer, A., D. Ostareck, et al. (1994). "Translation of 15-lipoxygenase mRNA is inhibited by a protein that binds to a repeated sequence in the 3' untranslated region." Embo J **13**: 1476 - 1481.

Parcells, B. W., A. K. Ikeda, et al. (2006). "FMS-like tyrosine kinase 3 in normal hematopoiesis and acute myeloid leukemia." Stem Cells **24**(5): 1174-1184.

Parcells, B. W., A. K. Ikeda, et al. (2006). "FMS-like tyrosine kinase 3 in normal hematopoiesis and acute myeloid leukemia." Stem Cells **24**(16410383): 1174-1184.

Park, S., O. Gakh, et al. (2003). "Yeast frataxin sequentially chaperones and stores iron by coupling protein assembly with iron oxidation." J Biol Chem **278**(33): 31340-31351.

Paz, A., T. Zeev-Ben-Mordehai, et al. (2008). "Biophysical characterization of the unstructured cytoplasmic domain of the human neuronal adhesion protein neuroligin 3." Biophys J **95**(4): 1928-1944.

Pesole, G. and S. Liuni (1999). "Internet resources for the functional analysis of 5' and 3' untranslated regions of eukaryotic mRNAs." Trends Genet **15**: 378.

Petoukhov, M. V., N. A. Eady, et al. (2002). "Addition of missing loops and domains to protein models by x-ray solution scattering." Biophys J **83**(6): 3113-3125.

Petoukhov, M. V., P. V. Konarev, et al. (2007). "ATSAS 2.1 - towards automated and web-supported small-angle scattering data analysis." J. Appl. Cryst. **40**(s1): s223-s228.

Petoukhov, M. V. and D. I. Svergun (2005). "Global rigid body modelling of macromolecular complexes against small-angle scattering data." Biophys J **89**(2): 1237-1250.

Petoukhov, M. V. and D. I. Svergun (2007). "Analysis of X-ray and neutron scattering from biomacromolecular solutions." Curr Opin Struct Biol **17**(5): 562-571.

Petoukhov, M. V., J. B. Vicente, et al. (2008). "Quaternary structure of flavorubredoxin as revealed by synchrotron radiation small-angle X-ray scattering." Structure **16**(9): 1428-1436.

Piloto, O., B. Nguyen, et al. (2006). "IMC-EB10, an anti-FLT3 monoclonal antibody, prolongs survival and reduces nonobese diabetic/severe combined immunodeficient engraftment of some acute lymphoblastic leukemia cell lines and primary leukemic samples." Cancer Res **66**(9): 4843-4851.

Pilz, I., O. Glatter, et al. (1972). "[Small-angle x-ray-scattering studies on the substructure of Helix pomatia hemocyanin]." Z Naturforsch B **27**(5): 518-524.

Porod, G. (1982). General theory. Small-angle X-ray scattering. O. Kratky. London, Academic Press**:** 17-51.

Prischi, F., P. V. Konarev, et al. (2010). "Structural bases for the interaction of frataxin with the central components of iron-sulphur cluster assembly." Nat Commun **1**: 95.

Putnam, C. D., M. Hammel, et al. (2007). "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.

Qian, S., H. McDonough, et al. (2006). "CHIP-mediated stress recovery by sequential ubiquitination of substrates and Hsp70." Nature **440**: 551 - 555.

Raman, S., R. Vernon, et al. (2009). "Structure prediction for CASP8 with all-atom refinement using Rosetta." Proteins-Structure Function and Bioinformatics **77**: 89-99.

Ramazzotti, G. and W. Maucci (1983). "The Phylum Tardigrada." Memorie dell'Istituto Italiano di Idrobiologia, Pallanza **41**: 309 - 314.

Reindl, C., K. Bagrintseva, et al. (2006). "Point mutations in the juxtamembrane domain of FLT3 define a new class of activating mutations in AML." Blood **107**(9): 3700-3707.

Rossman, K. L., C. J. Der, et al. (2005). "GEF means go: turning on RHO GTPases with guanine nucleotide-exchange factors." Nat Rev Mol Cell Biol **6**(15688002): 167-180.

Round, A. R., D. Franke, et al. (2008). "Automated sample-changing robot for solution scattering experiments at the EMBL Hamburg SAXS station X33." J. Appl. Cryst. **10**: 913-917.

Rubinstein, M. and R. H. Colby (2003). Polymer physics. Oxford ; New York, Oxford University Press.

Sanz, M., A. Burnett, et al. (2009). "FLT3 inhibition as a targeted therapy for acute myeloid leukemia." Curr Opin Oncol **21**(6): 594-600.

Savvides, S. N., T. Boone, et al. (2000). "Flt3 ligand structure and unexpected commonalities of helical bundles and cystine knots." Nat Struct Biol **7**(6): 486-491.

Schagerlof, U., H. Elmlund, et al. (2008). "Structural basis of the iron storage function of frataxin from single-particle reconstruction of the iron-loaded oligomer." Biochemistry **47**(17): 4948-4954.

Schannon, C. E. and W. Weaver (1949). The Mathematical Theory of Communication, Urbana:University of Illinois Press.

Schill, R., B. Mali, et al. (2009). "Molecular mechanisms of tolerance in tardigrades: new perspectives for preservation and stabilization of biological material." Biotechnol Adv **27**: 348 - 352.

Schill, R., S. Neumann, et al. (2008). "Detection of DNA damage with single-cell gel electrophoresis in anhydrobiotic tardigrades." Comp Biochem Physiol A Mol Integr Physiol **151**: 32 - 32.

Schmid, M. A., D. Kingston, et al. (2010). "Instructive cytokine signals in dendritic cell lineage commitment." Immunol Rev **234**(1): 32-44.

Schnittger, S., T. M. Kohl, et al. (2006). "D324N single-nucleotide polymorphism in the FLT3 gene is associated with higher risk of myeloid leukemias." Genes Chromosomes Cancer **45**(4): 332-337.

Schroder, G. F., M. Levitt, et al. (2010). "Super-resolution biomolecular crystallography with low-resolution data." Nature **464**(7292): 1218-1222.

Shankaranarayanan, A., C. A. Boguth, et al. (2010). "Galpha q allosterically activates and relieves autoinhibition of p63RhoGEF." Cell Signal **22**(20214977): 1114-1123.

Shen, Y., O. Lange, et al. (2008). "Consistent blind protein structure generation from NMR chemical shift data." Proc Natl Acad Sci U S A **105**(12): 4685-4690.

Shkumatov, A., S. Chinnathambi, et al. (2011). "Structural memory of natively unfolded tau protein detected by small-angle x-ray scattering." Proteins: Structure, Function, and Bioinformatics: n/a-n/a.

Shkumatov, A., D. I. Svergun, et al. (2009). "Small-Angle X-ray Scattering Driven Docking." HPC-Europa2: Science and Supercomputing in Europe (research highlights 2009): 78.

Stark, P. B. and R. L. Parker (1995). "Bounded-Variable Least-Squares - an Algorithm and Applications." Computational Statistics **10**(2): 129-141.

Stirewalt, D. L. and J. P. Radich (2003). "The role of FLT3 in haematopoietic malignancies." Nat Rev Cancer **3**(9): 650-665.

Stirewalt, D. L. and J. P. Radich (2003). "The role of FLT3 in haematopoietic malignancies." Nat Rev Cancer **3**(12951584): 650-665.

Stroud, R. M. and J. A. Wells (2004). "Mechanistic diversity of cytokine receptor signaling across cell membranes." Sci STKE **2004**(231): re7.

Suhre, K. and Y. H. Sanejouand (2004). "ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement." Nucleic Acids Res **32**(Web Server issue): W610-614.

Sun, Y. and T. MacRae (2005). "Small heat shock proteins: molecular structure and chaperone function." Cell Mol Life Sci **62**: 2460 - 2476.

Sundberg, E. J. and R. A. Mariuzza (2002). "Molecular recognition in antibody-antigen complexes." Adv Protein Chem **61**: 119-160.

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**(6): 2879-2886.

Svergun, D. I., C. Barberato, et al. (1995). "CRYSOL - a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates." J. Appl. Crystallogr. **28**: 768-773.

Svergun, D. I. and M. H. J. Koch (2002). "Advances in structure analysis using small-angle scattering in solution." Curr Opin Struct Biol **12**(5): 654-660.

Svergun, D. I. and M. H. J. Koch (2003). "Small angle scattering studies of biological macromolecules in solution." Rep. Progr. Phys. **66**: 1735-1782.

Svergun, D. I., M. V. Petoukhov, et al. (2000). "Crystal versus solution structures of thiamine diphosphate-dependent enzymes." J Biol Chem **275**(1): 297-302.

Svergun, D. I., M. V. Petoukhov, et al. (2001). "Determination of domain structure of proteins from X-ray solution scattering." Biophys J **80**(6): 2946-2953.

Svergun, D. I., S. Richard, et al. (1998). "Protein hydration in solution: experimental observation by x-ray and neutron scattering." Proc Natl Acad Sci U S A **95**(5): 2267-2272.

Tama, F. and C. L. Brooks (2006). "Symmetry, form, and shape: guiding principles for robustness in macromolecular machines." Annu Rev Biophys Biomol Struct **35**: 115-133.

Tama, F., F. X. Gadea, et al. (2000). "Building-block approach for determining low-frequency normal modes of macromolecules." Proteins **41**(10944387): 1-7.

Tama, F., O. Miyashita, et al. (2004). "Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM." Journal of Structural Biology **147**(3): 315-326.

Tama, F. and Y. H. Sanejouand (2001). "Conformational change of proteins arising from normal mode calculations." Protein Engineering **14**(1): 1-6.

Tama, F. and Y. H. Sanejouand (2001). "Conformational change of proteins arising from normal mode calculations." Protein Eng **14**(11287673): 1-6.

Tardieu, A. and P. Vachette (1982). "Analysis of models of irregular shape by solution X-ray scattering: the case of the 50S ribosomal subunit from E. coli." EMBO J. **1**: 35-40.

Tatusov, R., N. Fedorova, et al. (2003). "The COG database: an updated version includes eukaryotes." BMC Bioinformatics **4**: 41.

Tatusov, R., M. Galperin, et al. (2000). "The COG database: a tool for genome-scale analysis of protein functions and evolution." Nucleic Acids Res **28**: 33 - 36.

Tidow, H., R. Melero, et al. (2007). "From the Cover: Quaternary structures of tumor suppressor p53 and a specific p53 DNA complex." Proc Natl Acad Sci U S A **104**(30): 12324-12329.

Tirion, M. M. (1996). "Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis." Phys Rev Lett **77**(10063201): 1905-1908.

Tjoelker, L., L. Gosting, et al. (2000). "Structural and functional definition of the human chitinase chitin-binding domain." J Biol Chem **275**: 514 - 520.

Tosatto, S. C., E. Bindewald, et al. (2002). "A divide and conquer approach to fast loop modeling." Protein Eng **15**(4): 279-286.

Tunnacliffe, A., J. Lapinski, et al. (2005). "A putative LEA protein, but no trehalose, is present in anhydrobiotic bdelloid rotifers." Hydrobiologia **546**: 315 - 321.

Tunnacliffe, A. and M. Wise (2007). "The continuing conundrum of the LEA proteins." Naturwissenschaften **94**: 791 - 812.

Turner, A. M., N. L. Lin, et al. (1996). "FLT3 receptor expression on the surface of normal and malignant human hematopoietic cells." Blood **88**(9): 3383-3390.

Uversky, V. N. (2009). "Intrinsically Disordered Proteins and Their Environment: Effects of Strong Denaturants, Temperature, pH, Counter Ions, Membranes, Binding Partners, Osmolytes, and Macromolecular Crowding." Protein Journal **28**(7-8): 305-325.

Vagenende, V., M. G. Yap, et al. (2009). "Mechanisms of protein stabilization and prevention of protein aggregation by glycerol." Biochemistry **48**(46): 11084-11096.

van den Heuvel, R. H., D. I. Svergun, et al. (2003). "The active conformation of glutamate synthase and its binding to ferredoxin." J Mol Biol **330**(1): 113-128.

van Dijk, A. D., R. Boelens, et al. (2005). "Data-driven docking for the study of biomolecular complexes." Febs J **272**(2): 293-312.

Verstraete, K., S. Koch, et al. (2009). "Efficient production of bioactive recombinant human Flt3 ligand in E. coli." Protein J **28**(2): 57-65.

Verstraete, K., G. Vandriessche, et al. (2011). "Structural insights into the extracellular assembly of the hematopoietic Flt3 signaling complex." Blood: blood-2011-2001-329532.

Vestergaard, B., M. Groenning, et al. (2007). "A helical structural nucleus is the primary elongating unit of insulin amyloid fibrils." Plos Biology **5**(5): 1089-1097.

Vestergaard, B., S. Sanyal, et al. (2005). "The SAXS solution structure of RF1 differs from its crystal structure and is similar to its ribosome bound cryo-EM structure." Mol Cell **20**(6): 929-938.

Voisine, C., B. Schilke, et al. (2000). "Role of the mitochondrial Hsp70s, Ssc1 and Ssq1, in the maturation of Yfh1." Mol Cell Biol **20**(10): 3677-3684.

Volkov, V. V. and D. I. Svergun (2003). "Uniqueness of ab initio shape determination in small-angle scattering." J. Appl. Cryst. **36**: 860-864

von Bergen, M., P. Friedhoff, et al. (2000). "Assembly of tau protein into Alzheimer paired helical filaments depends on a local sequence motif ((306)VQIVYK(311)) forming beta structure." Proc Natl Acad Sci U S A **97**(10): 5129-5134.

Walsh, P., D. Bursac, et al. (2004). "The J-protein family: modulating protein assembly, disassembly and translocation." EMBO Rep **5**: 567 - 571.

Wang, X., P. Lupardus, et al. (2009). "Structural biology of shared cytokine receptors." Annu Rev Immunol **27**: 29-60.

Waskow, C., K. Liu, et al. (2008). "The receptor tyrosine kinase Flt3 is required for dendritic cell development in peripheral lymphoid tissues." Nat Immunol **9**(6): 676-683.

Wells, J. A. and C. L. McClendon (2007). "Reaching for high-hanging fruit in drug discovery at protein-protein interfaces." Nature **450**(7172): 1001-1009.

Wennerberg, K., K. L. Rossman, et al. (2005). "The Ras superfamily at a glance." J Cell Sci **118**(15731001): 843-846.

Wille, H., E. M. Mandelkow, et al. (1992). "The juvenile microtubule-associated protein MAP2c is a rod-like molecule that forms antiparallel dimers." J Biol Chem **267**(15): 10737-10742.

Wright, J. (2001). "Cryptobiosis 300 Years on from van Leuwenhoek: What Have We Learned about Tardigrades?" Zoologischer Anzeiger - A Journal of Comparative Zoology **240**: 563 - 582.

Wu, L. and Y. J. Liu (2007). "Development of dendritic-cell lineages." Immunity **26**(6): 741-750.

Yu, B., I. R. S. Martins, et al. (2010). "Structural and energetic mechanisms of cooperative autoinhibition and activation of Vav1." Cell **140**(20141838): 246-256.

Yuzawa, S., Y. Opatowsky, et al. (2007). "Structural basis for activation of the receptor tyrosine kinase KIT by stem cell factor." Cell **130**(2): 323-334.

Zhang, Y. (2007). "Template-based modeling and free modeling by I-TASSER in CASP7." Proteins **69 Suppl 8**: 108-117.

Zhang, Z., A. Schaffer, et al. (1998). "Protein sequence similarity searches using patterns as seeds." Nucleic Acids Res **26**: 3986 - 3990.

Zheng, M., T. Cierpicki, et al. (2009). "On the mechanism of autoinhibition of the RhoA-specific nucleotide exchange factor PDZRhoGEF." BMC Struct Biol **9**(19460155): 36-36.

Zheng, W. and S. Doniach (2002). "Protein structure prediction constrained by solution X-ray scattering data and structural homology identification." J Mol Biol **316**(1): 173-187.

Zheng, W. and S. Doniach (2005). "Fold recognition aided by constraints from small angle X-ray scattering data." Protein Eng Des Sel **18**(5): 209-219.

# List of publications

## Publications related to this work

Forster, F., Liang, C., Shkumatov, A., Beisser, D., Engelmann, J. C., Schnolzer, M., Frohme, M., Muller, T., Schill, R. O., Dandekar, T. (2009). **Tardigrade workbench: comparing stress-related proteins, sequence-similar and functional protein clusters as well as RNA elements in tardigrades.** BMC Genomics **10**: 469.

Shkumatov, A., Svergun, D. and Bonvin A.M.J.J. (2008). **Selection and Refinement of HADDOCK Solutions Using SAXS Data.** Science and Supercomputing in Europe (report 2008): 274-278.

Shkumatov, A., Svergun, D. and Bonvin A.M.J.J. (2009). **Small-Angle X-ray Scattering Driven Docking.** HPC-Europa2: Science and Supercomputing in Europe (research highlights 2009): 78.

Verstraete, K., Vandriessche, G., Januar, M., Elegheert, J., Shkumatov, A.V., Desfosses, A., Van Craenenbroeck, K., Svergun, D.I., Gutsche, I., Vergauwen, B., Savvides, S.N. (2011). **Structural insights into the extracellular assembly of the hematopoietic Flt3 signaling complex.** Blood: blood-2011-2001-329532.

Shkumatov, A.V., Chinnathambi, S., Mandelkow, E., Svergun, D.I. (2011). **Structural memory of natively unfolded tau protein detected by small-angle x-ray scattering.** Proteins: Structure, Function, and Bioinformatics: DOI: 10.1002/prot.23033

Elegheert, J., Desfosses, A., Shkumatov, A., Wu, X., Bracke, N., Verstraete, K., Craenenbroeck, K., Brooks, B.R., Svergun, D.I., Vergauwen, B., Gutsche, I., Savvides,S.N. **Extracellular complexes of the hematopoietic human and mouse CSF-1 receptor are driven by common assembly principles.** Under review in Structure (12/05/2011).

Söderberg, C.A.G., Shkumatov, A.V., Rajan, S., Gakh, O., Svergun, D.I., Isaya, G. and Al-Karadaghi, S. **Oligomerization propensity and flexibility of yeast frataxin studied by x-ray crystallography and small angle x-ray scattering.** Manuscript in preparation.

Franke, D., Gajda, M., Gorba, C., Konarev, P.V., Mertens, H., Petoukhov, M.V., Shkumatov, A.V., Tria, G., Svergun D.I[&]. **New developments in ATSAS program package for small angle scattering data analysis.** Manuscript in preparation.

Bielnicki, J., Shkumatov, A.V., Svergun, D.I. and Derewenda, Z. **Supramodular structure of RhoA-specific Guanidine Nucleotide Exchange Factor.** Manuscript to be submitted.

Whelan, F., Stead, J.A., Shkumatov, A.V., Svergun, D.I., Antson, A.A., Sanders, C.M[&]. **Solution structure of E1 helicase domain.** Manuscript in preparation.

[&] Title and the order of the author list are not final.

# Publications resulting from my work before PhD

Lesnikovich, J.A., Adzerikho, I.E., Shkumatov, A.V., Cherniavsky, E.A., Shkumatov, V.M. **The effect of ultrasound on molecular transitions of antithrombin III: native monomer-latent-polymers.** Proceedings of 4TH Conference "Application of Power Ultrasound in Physical and Chemical Processing", Besancon, France, 22-23 may 2003, pp. 363-367.

Stark, A., A. Shkumatov and Russell, R. B. (2004). **Finding functional sites in structural genomics proteins.** Structure **12**(8): 1405-1412.

Shkumatov, A.V., Sentchouk, V.V. **3D structure modeling of human thyroglobulin terminal domain.** Biochemistry: collection of scientific articles. Minsk, 2005: pp. 33-38.

Shkumatov, A.V., Sentchouk, V.V. **3D structure modelling of human thyroid peroxidase.** Biochemistry: collection of scientific articles. Minsk, 2007: pp. 75-81

# Contributions

## Tardigrade workbench: comparing stress-related proteins, sequence-similar and functional protein clusters as well as RNA elements in tardigrades.

**Frank Förster**[*], **Chunguang Liang**[*], **Alexander Shkumatov**[*], **Daniela Beisser**, **Julia C Engelmann**, **Martina Schnölzer**, **Marcus Frohme**, **Tobias Müller**, **Ralph O Schill and Thomas Dandekar.** *BMC Genomics* **10**: 469. doi: 10.1186/1471-2164-10-469

In this paper all available published and unpublished tardigrades' sequences were compared.
Different protein clusters and regulatory elements implicated in tardigrade stress adaptations were analysed including unpublished tardigrade sequences.
I performed the initial setup of the tardigrade analyzer server, of the virtual ribosome and the CLANS clustering. This laid the groundwork for further development of the server and sequence comparisons.

## Selection and Refinement of HADDOCK Solutions Using SAXS Data.

**Alexander Shkumatov, Dmitri Svergun and Alexandre M.J.J. Bonvin.** *Science and Supercomputing in Europe (report 2008)*: 274-278.

In this report I describe the use of HADDOCK with selection and refinement of models from a large pool based on information from SAXS. Interestingly, for one of the test complexes I found in selected HADDOCK subset models with subunit orientation similar to previously proposed (van den Heuvel, Svergun et al. 2003).

## Small-Angle X-ray Scattering Driven Docking.

**Alexander Shkumatov, Dmitri Svergun and Alexandre M.J.J. Bonvin.** *HPC-Europa2: Science and Supercomputing in Europe (research highlights 2009)*: 78.

In this report I describe an automatic method, combining SAXS rigid body modeling with subsequent docking refinement as well as present first results.

## Structural insights into the extracellular assembly of the hematopoietic Flt3 signaling complex.

**Kenneth Verstraete, Gonzalez Vandriessche, Mariska Januar, Jonathan Elegheert, Alexander V. Shkumatov, Ambroise Desfosses, Kathleen Van Craenenbroeck, Dmitri I. Svergun, Irina Gutsche, Bjorn Vergauwen, Savvas N. Savvides.** *Blood*: blood-2011-2001-329532. doi: 10.1182/blood-2011-01-329532

This paper reports for the first time the structural basis for the Flt3 ligand-receptor complex and unveils an unanticipated extracellular assembly unlike any other RTKIII/V complexes characterized to date.
I participated in SAXS data collection, performed the data reduction and constrained rigid body modeling.

## Structural memory of natively unfolded tau protein detected by small-angle x-ray scattering.

**Alexander V. Shkumatov*, Subashchandrabose Chinnathambi*, Eckhard Mandelkow and Dmitri I. Svergun.** *Proteins: Structure, Function, and Bioinformatics.* doi: 10.1002/prot.23033

In this paper the influence of temperature on the dynamics of natively unfolded tau protein was investigated through SAXS, DLS, CD spectroscopy and light-scattering. The very surprising result is, that SAXS detects a compaction that occurs upon temperature jump from 10 to 50 and 50 to 10 degrees and that continues over several hours. We concluded that a memory effect was found. This is astounding since for IDPs, the time scales involved are assumed to be fast and not comparable with time scales of refolding of several hours.

I collected the SAXS data under different temperature conditions in multiple sessions at the SAXS beamline in Hamburg. I performed data analysis and wrote the manuscript.

## Extracellular complexes of the hematopoietic human and mouse CSF-1 receptor are driven by common assembly principles.

**Jonathan Elegheert, Ambroise Desfosses, Alexander V. Shkumatov, Xiongwu Wu, Nathalie Bracke, Kenneth Verstraete, Kathleen Van Craenenbroeck, Bernard R.**

**Brooks, Dmitri I. Svergun, Bjorn Vergauwen, Irina Gutsche, Savvas N. Savvides**. Under review in *Structure* (12/05/2011).

This paper reports a range of novel structural (SAXS, EM) and thermodynamic data (ITC) on the human CSF-1 ligand receptor complex. Cross-reactivity of human and mouse ligands and receptors are quantified for the first time. The aim of this work was to arrive to a clear structural and biophysical consensus for the assembly of extracellular CSF-1 complexes, and to position the ensuing consensus within the framework of the plethora of available structural data on extracellular RTKIII/V complexes.

I participated in SAXS data collection, performed the data reduction and constrained rigid body modeling. SAXS part in this paper is quite substantial, namely, (i) we showed via SAXS that the first three domains of hCSF-1R and mCSF-1R lacking the membrane-proximal domains make stable ternary complexes with dimeric hCSF-1L and mCSF-1L, respectively, thus establishing bivalent cytokine binding to receptor as a key common denominator to receptor activation; (ii) we provided direct evidence via SAXS that unliganded hCSF-1R undergoes large conformational changes to bind its ligand, and that it can exist in a predimerized form within a broad concentration range. This provides the first structural evidence for an earlier study showing that CSF-1R can exist in a dimeric form at the cell-surface.

## Oligomerization propensity and flexibility of yeast frataxin studied by x-ray crystallography and small angle x-ray scattering.

**Christopher AG Söderberg, Alexander V. Shkumatov, Sreekanth Rajan, Oleksandr Gakh, Dmitri I. Svergun, Grazia Isaya and Salam Al-Karadaghi**. Manuscript in preparation.

This paper describes different factors influencing oligomerization properties of frataxin in solution. We showed that different oligomeric states may be induced by buffer content, amino acid composition of the N-terminus and presence of divalent metals in the solution.

I collected the SAXS data for monomeric frataxin with and without glycerol, performed data reduction and modeling, wrote several parts of the manuscript.

## New developments in ATSAS program package for small angle scattering data analysis.

Daniel Franke, Michael Gajda, Christian Gorba, Peter V. Konarev, Haydyn Mertens, Maxim V. Petoukhov, Alexander V. Shkumatov, Giancarlo Tria, Dmitri I. Svergun [&]. Manuscript in preparation.

This paper describes methodological developments in ATSAS package. I contributed to this paper by developing faster version of CRYSOL and CRYSON. I created two new programs, RANLOGS and EM2DAM. Using simulated data I checked how MW can be estimated using excluded and Porod volumes obtained by DAMMI[N/F] and AUTOPOROD, respectively. I wrote the corresponding parts of the manuscript.

## Supramodular structure of RhoA-specific Guanidine Nucleotide Exchange Factor.

**Jakub Bielnicki*, Alexander V. Shkumatov*, Dmitri I. Svergun and Zygmunt Derewenda**. Manuscript in preparation.

This paper gives a new insight on mechanism of PRG regulation based on the data obtained by a combination of biophysical, biochemical and structural methods. SAXS is a major technique used in this project. Based on the obtained rigid body models, we proposed a model of PRG regulation.

I collected the SAXS data for all constructs described in the paper, performed data reduction and exhaustive modeling, wrote respective parts of the manuscript.

## Solution structure of E1 helicase domain.

**Fiona Whelan, Jonathan A. Stead, Alexander V. Shkumatov, Dmitri I. Svergun, Alfred A. Antson, Chris M. Sanders**. Manuscript in preparation.

This paper along with biochemical data describes the solution structure of hexameric E1 helicase domain obtained by SAXS.

I collected the SAXS data, performed data reduction and modeling, wrote respective parts of the manuscript.

***equal contributors**

# Poster contributions, participations and visits

1. Participation in 9th International EMBL PhD Student Symposium "Patterns in Biology – Organisation of Life in Space in Time", held in Heidelberg, Germany from 25th of October to the 27th of October 2007.
2. Participation in "Copenhagen workshop on biomacromolecules in solution studied with Small-Angle Scattering" at University of Copenhagen, Denmark, November 25-30, 2007.
3. Participation in the practical bioinformatics course "Multiple Sequence Alignment and Phylogenies" at EMBL Heidelberg, Germany 23rd – 24th of June, 2008
4. HPC-Europa Transnational Access visit. Title of the project "Selection and refinement of HADDOCK solutions using experimental Small-Angle X-ray Scattering Data." Host: Alexandre M.J.J. Bonvin, NMR Research Group, Bijvoet Center for Biomolecular Research, Utrecht University, Padulaan 8, 3584 CH Utrecht, The Netherlands; 30.08.2008 - 10.11.2008
5. Poster presentation "Selection and refinement of HADDOCK solutions using experimental Small-Angle X-ray Scattering Data" at Translational Access Meeting (TAM'08), Stuttgart, Germany, December 15-17, 2008
6. Poster presentation "Temperature effect on the structure and conformation of Tau protein in solution studied using Small-Angle X-ray Scattering" at HASYLAB user meeting at DESY, Hamburg, January 29, 2009
7. HPC-Europa2 Transnational Access visit. Title of the project "An automatic method for generation and refinement of SAXS-based models of protein-protein complexes." Host: Alexandre M.J.J. Bonvin, NMR Research Group, Bijvoet Center for Biomolecular Research, Utrecht University, Padulaan 8, 3584 CH Utrecht, The Netherlands; 01.09.2009 - 30.10.2009
8. Speaker at "Copenhagen workshop on BIO-macromolecules in solution studied with Small-Angle Scattering" at University of Copenhagen, Denmark, January 17-22, 2010.

   Title of the talk "Comparing Small-Angle Scattering data to protein crystallography and NMR using CRYSOL"
9. Tutor at EMBO practical courses on Solution Scattering from Biological Macromolecules on 19-26.11.2008 and 25.10-01.11.2010, EMBL Hamburg,

Germany.

10. Seminar speaker at Center for Molecular Protein Science at Lund University. 10th of November 2010. Title "Ways and means to study structural flexibility of proteins." Host: Salam al Karadaghi
11. Poster "Molten Globule Region of PDZRhoGEF Regulates Autoinhibition" by Jakub A. Bielnicki*, Alexander V. Shkumatov*, Dmitri I. Svergun, and Zygmunt S. Derewenda. Presented by Jakub Bielnicki on 55th Annual Biophysical Society Meeting, Baltimore, MD. March 5-9, 2011
12. Invited speaker at Ghent University 5th of April, 2011. Title "Structural flexibility studied by Small-Angle X-ray Scattering". Host: Savvas N Savvides
13. Speaker and tutor at the EMBO course (16 - 20 May 2011) in Grenoble, France. Title of the talk: Low and high-resolution structure validation (CRYSOL/CRYSON/EM2DAM) and binary complex refinement. Tutorial: "Use of structure validation tools".
14. Speaker at EMBL Hamburg. Title of the talk "Ways and means to study structural plasticity in macromolecules". 1st of July 2011

# Participation in other courses

1. "The Effective Team Leader"; 11th and 12th of June 2008, EMBL Hamburg. Trainer: Frances Scott.
2. "Advanced Presentation Skills"; 14th of January 2008, EMBL Hamburg. Trainer: Alison Sargent
3. "Managing your career after EMBL"; 24th of February 2009, EMBL Hamburg. Trainer: David Winter & Eric Evans
4. "How to boost your chances of getting a job"; 25th of February 2009, EMBL Hamburg. Trainer: David Winter & Eric Evans

# Acknowledgements

I would like to thank Dr. Dmitri I. Svergun for giving me the opportunity to do the PhD in his group. He provided excellent settings for work, helped establish a number of collaborations as well as gave an interesting research topic. I thank him for being a good and accessible supervisor in spite of his busy schedule. I acknowledge HFSP Research Grant Ref. RGP 55/2006 for the financial support.

Further I acknowledge all the member of the group for stimulating research atmosphere, especially Clement Blanchet for excellent technical support at the X33 beamline, Maxim Petoukhov and Daniel Franke for help and critical comments on programming issues, Haydyn Mertens for proofreading my manuscripts, Peter Konarev for helpful advices on SAXS data analysis. I am grateful to Christian Gorba and Haydyn Mertens for proofreading this Thesis.

In this work Christian Gorba contributed to development of NMADREFS by providing subroutines, performing linear combination of normal models. Maxim Petoukhov helped in the development of RANLOGS and EM2DAM as well as refurbishment of CRYSOL.

I am particularly grateful to people involved in collaborative projects with me. My thanks go to Subashchandrabose Chinnathambi from Eckhard Mandelkow group at MPUSMB for his contribution to tau project, Kenneth Verstraete and Jonathan Elegheert from Savvas Savvides group at University of Ghent, Jakub Bielnicki from Zygmunt Derewenda group at the University of Virginia, Fiona Whelan from Fred Antson group at the University of York, Christopher Söderberg from Salam Al-Karadaghi group at Lund University.

I further thank my university supervisor Thomas Dandekar and other members of my Thesis Advisory Committee, namely, Dmitri Svergun, Thomas Schneider and Anne-Claude Gavin for useful discussions and criticism.

I also thank Dr. habil. Matthias Wilmanns and Prof. Dr. Thomas Dandekar for agreeing to be referees for my Thesis.

I am grateful to EMBL administration for being helpful, supportive and fast.

I thank members of the “lunch club” for interesting scientific and mostly non scientific discussions, namely, Al Kikhney, Jacopo Negroni, Shao-Yang Ku, Fabio Dall’Antonia,

Jon Rapley, Nina Krueger, Matthew Dunne, Giancarlo Tria, Heidi Kaljunen and Anna Gieras.

Matthew Groves, Nathalie Thebaud and Doris Jan for their help with organization of staff association events. Especially Matt and Nathalie for a lot of fun during mini golf, ultimate frisbee and water skiing outings in Hamburg.

Viktor Lamzin and Lesleis Nagy for spending time playing chess with me.

I thank people, who shared office with me over the last couple of years at EMBL, for discussions, help and a nice atmosphere. My thanks go to Philipp Heuser, Annette Faust and Al-Kikhney, who shared office with me in Siberia in the first few month of my PhD at EMBL; Efstratious Mylonas, Maxim Petoukhov, Roberto Mosca, Fabio Dall'Antonia, Giancarlo Tria and Christian Gorba, who were my office mates during the last three years. I am very thankful to Fabio and Al for their help with preparing figures for posters and manuscripts. Fabio for regular 4pm coffee breaks and chats.

I am particularly grateful to my family Vladimir, Galina and Alena Shkumatova for their support, care, encouragement and practical advices they gave me throughout many years.

# Erklärung

I hereby declare that my thesis entitled “Methods for hybrid modeling of solution scattering data and their applications” is the result of my own work.

I did not receive any help or support from commercial consultants.

All sources and/or materials applied are listed and specified in the thesis.

Furthermore, I verify that this thesis has not been submitted as part of another examination process neither in identical nor similar form.

Alexander V. Shkumatov

# Curriculum Vitae

**Personal Data**

| | |
|---|---|
| Name | Alexander V. Shkumatov |
| Nationality | Belarus |
| Date of Birth | August 13th, 1983 |
| Place of Birth | Minsk, Belarus |

**Education**

| | |
|---|---|
| 15.03.2007 - present | **PhD student** at *EMBL-Hamburg*.Supervisor: Dr. Dmitri I. Svergun. Doktorvater: Prof. Dr. Thomas Dandekar |
| 08.08.2005 - 14.11.2006 | **PhD student** at *MPI for Developmental Biology*, Tübingen, Germany. Supervisor: Dr. Andrei Lupas |
| 01.09.2000 - 09.06.2005 | **Student** of *Belarusian State University*, Biology Department, specialization – biochemistry |
| 01.09.1996 - 25.05.2000 | Higher Secondary School, *school #65* (Minsk, Belarus) |
| 01.09.1989 - 25.05.1996 | Secondary School, *school #147* (Minsk, Belarus) |

**Work experience**

| | |
|---|---|
| 15.03.2007 - present | **PhD student** at *EMBL-Hamburg* |
| 15.11.2006 - 15.03.2007 | **Scientific assistant** at *Lehrstuhl für Bioinformatik* (Thomas Dandekar group), *Universität Würzburg* |
| 08.06.2004 - 30.09.2004 | **Summer trainee,** *Institute of Biotechnology*, Helsinki, Finland. Group of Prof. Liisa Holm. |
| 15.07.2003 - 15.09.2003 | **Summer trainee** at *EMBL-Heidelberg*, Germany. Group of Dr. Rob B. Russell. Supervisor: Alexander Stark |

# Lebenslauf

**Persönliche Daten**

| | |
|---|---|
| Name | Alexander V. Shkumatov |
| Nationalität | Belarus |
| Geburtsdatum | August 13th, 1983 |
| Geburtsort | Minsk, Belarus |

**Studium und Ausbildung**

| | |
|---|---|
| Seit 15.03.2007 | **Doktorand** am *EMBL-Hamburg.* Betreuer: Dr. Dmitri I. Svergun. Doktorvater: Prof. Dr. Thomas Dandekar |
| 08.08.2005 - 14.11.2006 | **Doktorand** am *MPI für Entwicklungsbiologie*, Tübingen, Germany. Betreuer: Dr. Andrei Lupas |
| 01.09.2000 - 09.06.2005 | **Student** der *Belarussischen staatlichen Universität*, biologische Fakultät, Spezialisierung – Biochemie |
| 01.09.1996 - 25.05.2000 | **Abitur** *Schule №65* (Minsk, Belarus) |
| 01.09.1989 - 25.05.1996 | **Abitur** *Schule №147* (Minsk, Belarus) |

**Beruflicher Werdegang**

| | |
|---|---|
| Seit 15.03.2007 | **Doktorand** am *EMBL-Hamburg.* |
| 15.11.2006 - 15.03.2007 | **Wissenschaftlicher Mitarbeiter** am *Lehrstuhl für Bioinformatik, Universität Würzburg.* Gruppenleiter: Thomas Dandekar) |
| 08.06.2004 - 30.09.2004 | **Sommer Praktikant,** *Institut der Biotechnologie*, Helsinki, Finnland. Gruppenleiter Prof. Liisa Holm |
| 15.07.2003 - 15.09.2003 | **Sommer Praktikant** am *EMBL-Heidelberg*, Deutschland. Gruppenleiter: Dr. Rob B. Russell. Betreuer: Alexander Stark |